ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΑΓΡΟΝΟΜΩΝ ΚΑΙ ΤΟΠΟΓΡΑΦΩΝ ΜΗΧΑΝΙΚΩΝ
ΔΠΜΣ «ΓΕΩΠΛΗΡΟΦΟΡΙΚΗ»

Μεταπτυχιακή Εργασία

Διερεύνηση της αποτελεσματικότητας των αλγορίθμων «Δέντρα Απόφασης» και «Τυχαία Δάση» στην ανίχνευση κτιρίων μέσω αντικειμενοστρεφούς ανάλυσης υψηλής χωρικής διακριτικής ικανότητας πολυφασματικών εικόνων

Ταυλάκη Χρυσάνθη
Διπλωματούχος Αγρονόμος και τοπογράφος Μηχανικός ΕΜΠ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΑΡΓΙΑΛΑΣ Δ.

Μεταπτυχιακή Εργασία

Διερεύνηση της αποτελεσματικότητας των αλγορίθμων «Δέντρα Απόφασης» και «Τυχαία Δάση» στην ανίχνευση κτιρίων μέσω αντικειμενοστρεφούς ανάλυσης υψηλής χωρικής διακριτικής ικανότητας πολυφασματικών εικόνων

Εξεταστική επιτροπή:

Αργιαλάς Δημήτριος – Καθηγητής, ΣΑΤΜ Ε.Μ.Π. (Επιβλέπων)

Καράντζαλος Κωνσταντίνος – Επίκουρος Καθηγητής, ΣΑΤΜ Ε.Μ.Π. (Μέλος)

Καραθανάση Βασιλεία – Αναπληρώτρια Καθηγήτρια, ΣΑΤΜ ΕΜΠ (Μέλος)

1

Περιεχόμενα

Περιεχόμενα Εικόνων

Περιεχόμενα Πινάκων

Περίληψη

Η παρούσα μεταπτυχιακή εργασία αποσκοπεί στη διερεύνηση της αποτελεσματικότητας των αλγορίθμων επιβλεπόμενης ταξινόμησης «Δέντρα απόφασης» και «Τυχαία δάση» στην ανίχνευση κτιρίων μέσω αντικειμενοστρεφούς ανάλυσης εικόνας. Πιο αναλυτικά, σε πρώτο επίπεδο γίνεται ενδελεχής βιβλιογραφική έρευνα σχετικά με το θεωρητικό υπόβαθρο των παραπάνω αλγορίθμων και σε δεύτερο επίπεδο γίνεται υλοποίηση των τελευταίων στην υψηλής ευκρίνειας δορυφορική εικόνα Pleiades. Αναλυτικά, η διαδικασία της υλοποίησης διαρθρώνεται ως εξής: Αρχικά γίνεται εξομάλυνση και κατάτμηση της εικόνας εισόδου μέσω της μεθόδου της πολυκλιμακωτής κατάτμησης και εν συνεχεία μέσω του αλγορίθμου της φασματικής διαφοράς. Ακολούθως, υλοποιείται ένα σύνολο δοκιμών οι οποίες στοχεύουν στην εύρεση των τιμών των παραμέτρων οι οποίες θα δώσουν τα βέλτιστα δυνατά αποτελέσματα, εστιάζοντας συγκεκριμένα στην ανίχνευση κτιρίων. Τέλος, γίνεται μία προσπάθεια εύρεσης των χαρακτηριστικών των αντικειμένων τα οποία θα δώσουν την ακριβέστερη ταξινόμηση με το χαμηλότερο δυνατό υπολογιστικό κόστος μέσω του αλγορίθμου των τυχαίων δασών. Τα συμπεράσματα τα οποία εξήχθησαν μέσω των παραπάνω δοκιμών αξιολογήθηκαν σε διαφορετικά τμήματα τα εικόνας εισόδου. Τα αποτελέσματα είναι ενθαρρυντικά σε ό,τι αφορά την ικανοποίηση των κριτηρίων τόσο της πληρότητας όσο και της ορθότητας καθώς τα συγκεκριμένα συγκεντρώνουν ποσοστά μεγαλύτερα του 75%.

Λέξεις κλειδιά: Ανίχνευση κτιρίων, δορυφορικές εικόνες υψηλής ευκρίνειας, αντικειμενοστρεφής ανάλυση εικόνας, Δέντρα Απόφασης, Τυχαία Δάση

Abstract

The aim of this thesis is to investigate the effectiveness of two supervised classification techniques in automated building detection through object- based image analysis. More specifically we investigated the theoretical background of the Decision Trees and Random Forest algorithms and implemented them on a Pleiades image. We initially smoothed the image by applying the bilateral filter to it and then segmented the pre processed image using the multiresolution segmentation algorithm and the Spectral difference algorithm. Subsequently we performed a number of different tests in order to determine the parameter values that give the most accurate classification results for both algorithms. Finally, we determined the object metrics that aid in distinguishing buildings from other image objects. The results that were obtained from the aforementioned tests were evaluated on different parts of the Pleiades image and indicated good user and producer accuracies for the classification of buildings.

Keywords: Building detection, VHR satellite images, Object based image analysis, Decision trees, Random Forests

1 Εισαγωγή

Στόχος της παρούσας μεταπτυχιακής εργασίας είναι η διερεύνηση της αποτελεσματικότητας των αλγορίθμων δέντρα απόφασης και τυχαία δάση στην ταξινόμηση τηλεπισκοπικών δεδομένων. Διευκρινίζεται πως η συγκεκριμένη μελέτη εστιάζει περισσότερο στην αποτελεσματικότητα των αλγορίθμων αυτών σε ό,τι αφορά την ανίχνευση κτιρίων στα δεδομένα εισόδου.

Η εικόνα, η οποία μελετήθηκε προέρχεται από τους δορυφόρους Pléiades και είναι υψηλής ευκρίνειας. Η περιοχή που εμφανίζεται στα συγκεκριμένα δεδομένα είναι η πόλη Commerce της πολιτείας Colorado των Ηνωμένων Πολιτειών.

Αναλυτικά, η παρούσα εργασία διαρθρώνεται ως εξής:

- Στο **πρώτο κεφάλαιο «Εισαγωγή»** γίνεται μία σύντομη εισαγωγή στα θέματα τα οποία πραγματεύεται η παρούσα εργασία. Αναλυτικά, γίνεται επεξήγηση της έννοιας αντικειμενοστρεφής ανάλυση εικόνας καθώς και περιγραφή της διαδικασίας ταξινόμησης. Τέλος, γίνεται σύντομη αναφορά στα χαρακτηριστικά των δορυφόρων Pléiades.
- Αντικείμενο του **δεύτερου κεφαλαίου** «**Βιβλιογραφική ανασκόπηση**» είναι οι αλγόριθμοι δέντρα απόφασης και τυχαία δάση. Πιο συγκεκριμένα, γίνεται αναζήτηση των βιβλιογραφικών πηγών που αναφέρονται στις υπό μελέτη τεχνικές. Στόχος της διαδικασίας αυτής είναι αφενός η ενδελεχής μελέτη του θεωρητικού υπόβαθρου των συγκεκριμένων αλγορίθμων και αφετέρου η διερεύνηση της αποτελεσματικότητας των τελευταίων σε εφαρμογές της Ψηφιακής Τηλεπισκόπησης.
- Στο **τρίτο κεφάλαιο «Μεθοδολογία των αλγορίθμων «Δέντρα απόφασης» και «Τυχαία δάση» στο περιβάλλον του eCognition»** περιγράφεται η διαδικασία εφαρμογής των δέντρων απόφασης και των τυχαίων δασών στο περιβάλλον του eCognition. Συγκεκριμένα, περιγράφονται αναλυτικά όλες οι ενέργειες από την προεπεξεργασία των δεδομένων εισόδου έως και την ταξινόμηση τους. Επιπροσθέτως, γίνεται μελέτη της επιρροής των τιμών των παραμέτρων των συγκεκριμένων αλγορίθμων στην ποιότητα του παραγόμενου αποτελέσματος. Στη συνέχεια, επιχειρείται να γίνει βελτίωση των επιδόσεων του αλγορίθμου των τυχαίων δασών αξιοποιώντας τη δυνατότητα που παρέχει ο συγκεκριμένος για υπολογισμό της σημαντικότητας των μεταβλητών (Variable importance). Μέσω της ιδιότητας αυτής γίνεται επιλογή των χαρακτηριστικών των αντικειμένων τα οποία θα δώσουν τα βέλτιστα δυνατά αποτελέσματα σε ό,τι αφορά την ποιότητα της ταξινόμησης. Τέλος, γίνεται εφαρμογή της προτεινόμενης μεθοδολογίας σε διαφορετικό τμήμα της εικόνας- εισόδου. Στόχος της συγκεκριμένης διαδικασίας ήταν η επιβεβαίωση των συμπερασμάτων των κεφαλαίων 3 κα 4.
- Στο **τέταρτο** και τελευταίο **κεφάλαιο «Συμπεράσματα - Προοπτικές»** γίνεται παράθεση των συμπερασμάτων που αντλήθηκαν βάσει των εφαρμογών της προηγούμενης ενότητας

1.1 Αντικειμενοστρεφής ανάλυση εικόνας

Η Ψηφιακή Τηλεπισκόπηση ξεκίνησε ουσιαστικά το 1972, όταν εκτοξεύθηκε ο πρώτος δορυφόρος με πολυφασματικό αισθητήρα, Landsat 1. Σύντομα αναπτύχθηκαν μέθοδοι ταξινόμησης, οι οποίες είχαν σαν μοναδιαίο στοιχείο τα pixel της εικόνας. Οι συγκεκριμένες

τεχνικές, ωστόσο είναι περισσότερο αποδοτικές σε περιπτώσεις όπου τα εικονοστοιχεία απεικονίζουν αντικείμενα με διαστάσεις μικρότερες ή ίσες με τη χωρική ανάλυση των δεδομένων (Blaschke et al., 2014).

Η επιστήμη της Ψηφιακής Τηλεπισκόπησης έχει σημειώσει ραγδαία εξέλιξη από τη στιγμή της εκτόξευσης του Landsat 1. Οι νέοι δορυφόροι έχουν πλέον εξοπλιστεί με αισθητήρες πολυφασματικούς, υπέρφασματικούς και ραντάρ, ενώ η χωρική καθώς και η φασματική τους ανάλυση αυξάνεται συνεχώς (Arvor et al., 2013).

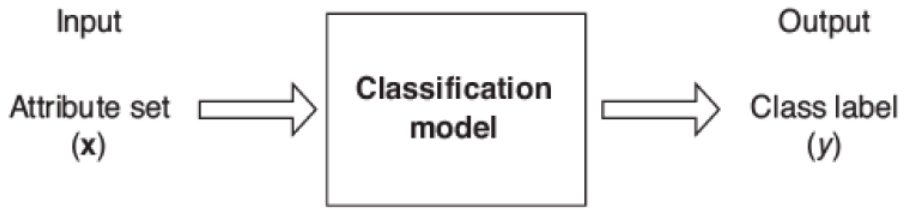
Το παραπάνω είχε σαν επακόλουθο τη δημιουργία στις αρχές του 2000 του πρώτου εμπορικού λογισμικού, μέσω του οποίου ήταν δυνατή η δημιουργία και η ανάλυση αντικειμένων σε μία εικόνα. Αναπτύχθηκε με άλλα λόγια η αντικειμενοστρεφής ανάλυση εικόνας (Object Based Image Analysis – OBIA). Ο όρος GEOBIA (Geographic Object Image Analysis) υιοθετήθηκε διότι στις οντότητες που μελετώνται μέσω της Ψηφιακής Τηλεπισκόπησης υπεισέρχεται η έννοια του χώρου (Blaschke et al., 2014).

Η GEOBIA αποτελεί ουσιαστικά μία σύνδεση του κόσμου των εικονοστοιχείων με εκείνου των διανυσμάτων. Η συγκεκριμένη τεχνική έχει γίνει ιδιαίτερα δημοφιλής την τελευταία δεκαετία καθώς παρέχει τη δυνατότητα ενσωμάτωσης σημασιολογίας, η οποία βασίζεται στην περιγραφική αξιολόγηση καθώς και στη γνώση. Συνεπώς, η παρούσα προσέγγιση ενσωματώνει στη διαδικασία της αυτόματης ταξινόμησης τη «σοφία» του χρήστη. Η επιστήμη της ανάλυσης εικόνας στο σύνολό της αποτελεί ουσιαστικά ένα εγχείρημα στο οποίο η γνώση μεταπλάθεται, καθώς οι χρήστες αξιοποιούν την υπάρχουσα εμπειρία τους προκειμένου να δημιουργήσουν αξιόπιστους, ποιοτικούς χάρτες (Arvor et al., 2013).

Τέλος αξίζει να σημειωθεί πως η αντικειμενοστρεφής ανάλυση εικόνας έχει συνδεθεί με την έννοια του παραδείγματος (paradigm) όπως διατυπώθηκε από τον Kuhn. Συμφώνα με τον Kuhn παράδειγμα είναι «ό,τι μοιράζονται τα μέλη μίας επιστημονικής κοινότητας. Αυτό περιλαμβάνει όχι μόνο τους νόμους και τα αποτελέσματα αυτής της επιστημονικής κοινότητας αλλά και τις μεθοδολογίες, τους στόχους, τις συμβάσεις, τις ερευνητικές ερωτήσεις και άλυτα προβλήματα τους» (Blaschke et al., 2014).

1.2 Ταξινόμηση

Η ταξινόμηση είναι μία διαδικασία ανάθεσης αντικειμένων σε προκαθορισμένες κλάσεις. Τα δεδομένα εισόδου x ενός μοντέλου ταξινόμησης είναι ένα σύνολο εγγραφών (ή οντοτήτων ή παραδειγμάτων) που κάθε μία από αυτές αποτελείται από μία πλειάδα χαρακτηριστικών (x,y). Ως x συμβολίζεται το σύνολο γνωρισμάτων μίας εγγραφής και ως y η «κλάση» στην οποία αυτή ανήκει (ή κατηγορία ή χαρακτηριστικό- στόχος) (Εικόνα 1.1). Η κλάση είναι ουσιαστικά ένα επιπλέον γνώρισμα κάθε εγγραφής με τη διαφορά πως το σύνολο x παίρνει τιμές τόσο συνεχείς όσο και διακριτές, ενώ εκείνη αποκλειστικά διακριτές. Το παραπάνω χαρακτηριστικό είναι εκείνο που διαφοροποιεί τη διαδικασία της ταξινόμησης από εκείνη της παλινδρόμησης καθώς στην τελευταία το χαρακτηριστικό στόχος παίρνει συνεχείς τιμές(Tan et al., 2005).



ΕΙΚΟΝΑ 1.1: ΓΡΑΦΙΚΗ ΑΝΑΠΑΡΑΣΤΑΣΗ ΕΝΟΣ ΜΟΝΤΕΛΟΥ ΤΑΞΙΝΟΜΗΣΗΣ (TAN ET AL., 2005).

Η ταξινόμηση είναι πρακτικά μία διαδικασία εκμάθησης μίας συνάρτησης – στόχου f (target function). Στόχος της f είναι ο καθορισμός των εγγραφών βάσει των χαρακτηριστικών τους x σε μία κλάση y. Η συνάρτηση στόχος είναι ανεπίσημα γνωστή ως μοντέλο ταξινόμησης και μπορεί να χρησιμοποιηθεί ως εξής:

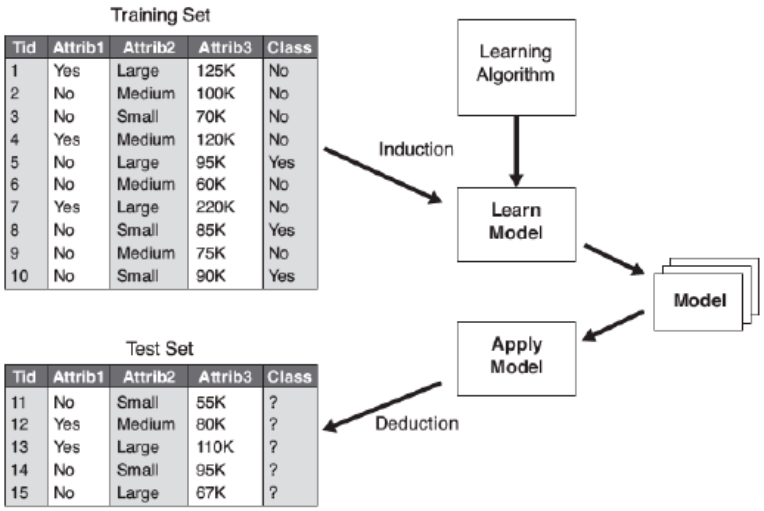
- **Περιγραφικό μοντέλο:** Μέσω του συγκεκριμένου γίνεται διάκριση των χαρακτηριστικών των αντικειμένων τα οποία ανήκουν σε διαφορετικές κατηγορίες.
- **Μοντέλο πρόβλεψης:** Στην προκειμένη περίπτωση γίνεται χρήση του μοντέλου προκειμένου να γίνει πρόβλεψη των κλάσεων στις οποίες ανήκουν ένα πλήθος εγγραφών.

Μία τεχνική ταξινόμησης (ή ταξινομητής – classifier) είναι ουσιαστικά μία συστηματική προσπάθεια δημιουργίας μοντέλων ταξινόμησης από ένα σύνολο δεδομένων εισόδου. Στη βιβλιογραφία αναφέρεται πλήθος διαφορετικών ταξινομητών όπως τα δέντρα απόφασης, τα νευρωνικά δίκτυα καθώς και τα Support Vector Machines. Κάθε μία από τις παραπάνω τεχνικές εφαρμόζει ένα διαφορετικό αλγόριθμο εκμάθησης προκειμένου να κατασκευαστεί το βέλτιστο μοντέλο ταξινόμησης των δεδομένων εισόδου. Το τελευταίο παράγει ακριβή αποτελέσματα και σε περιπτώσεις άγνωστων εγγραφών(Tan et al., 2005).

Το σύνολο των δεδομένων εισόδου διακρίνεται σε:

- Σύνολο εκπαίδευσης (training set): Το συγκεκριμένο χρησιμοποιείται για την κατασκευή του μοντέλου
- Σύνολο ελέγχου (test set): Το συγκεκριμένο χρησιμοποιείται για την επικύρωση του μοντέλου

Στην Εικόνα 1.2 εμφανίζεται μία γενική προσέγγιση για την επίλυση προβλημάτων ταξινόμησης. Το σύνολο των δεδομένων εκπαίδευσης αποτελείται από εγγραφές με γνωστή κλάση y και οι συγκεκριμένες χρησιμοποιούνται προκειμένου να γίνει η κατασκευή του μοντέλου (επαγωγή- induction). Στη συνέχεια, μέσω των δεδομένων ελέγχου γίνεται εφαρμογή καθώς και αξιολόγηση του μοντέλου που προέκυψε (συμπέρασμα -deduction) (Tan et al., 2005).



ΕΙΚΟΝΑ 1.2: ΚΑΤΑΣΚΕΥΗ ΜΟΝΤΕΛΟΥ ΤΑΞΙΝΟΜΗΣΗΣ (TAN ET AL., 2005).

Η αξιολόγηση των επιδόσεων ενός μοντέλου ταξινόμησης γίνεται μέσω της καταμέτρησης των οντοτήτων που ταξινομήθηκαν σωστά από το τελευταίο. Οι παραπάνω μετρήσεις καταγράφονται σε ένα πίνακα σύγχυσης μέσω του οποίου παρέχεται όλη η απαραίτητη πληροφορία αναφορικά με την ποιότητα του αποτελέσματος. Ωστόσο, σε περιπτώσεις σύγκρισης διαφορετικών μοντέλων είναι ιδιαίτερα εύχρηστη η ύπαρξη ενός μοναδικού αριθμού, μέσω του οποίου θα περιγράφεται η αποτελεσματικότητα αυτών. Ένα παράδειγμα τέτοιου μετρητή είναι η ακρίβεια, η οποία υπολογίζεται βάσει του ακόλουθου τύπου:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

Επίσης, ιδιαίτερα χρήσιμος είναι ο υπολογισμός του ακόλουθου δείκτη:

$$Error\ rate = \frac{Number\ of\ wrong\ predictions}{Total\ number\ of\ predictions}$$

Είναι εμφανές πως το μοντέλο εκείνο το οποίο επιλέγεται τελικά είναι εκείνο το οποίο πετυχαίνει τα υψηλότερα ποσοστά ακρίβειας καθώς και τα μικρότερα ποσοστά σφάλματος (Tan et al., 2005).

1.3 Αλγόριθμοι «δέντρα απόφασης» και «τυχαία δάση»

Μέσω της παρούσας εργασίας επιχειρείται η ενδελεχής ανάλυση δύο συγγενών μεταξύ τους αλγορίθμων ταξινόμησης εκείνων των δέντρων απόφασης και των τυχαίων δασών καθώς και η διευρεύνηση της αποτελεσματικότητας των εν λόγω αλγορίθμων σε εφαρμογές τηλεπισκόπησης και πιο συγκεκριμένα στην ανίχνευση κτιρίων. Τα δέντρα απόφασης αφενός αποτελούν μία από τις πλέον διαδεδομένες τεχνικές ταξινόμησης καθώς παρουσιάζουν πληθώρα πλεονεκτημάτων στα οποία συμπεριλαμβάνονται η αυτό-επεξηγηματικότητα καθώς η δυνατότητα που παρέχουν να διαχειρίζονται σύνολα δεδομένων με ελλιπείς τιμές. Το βασικό μειονέκτημα, ωστόσο, των δέντρων απόφασης έγκειται στο γεγονός πως σε πολλές περιπτώσεις εμφανίζουν υπερπροσαρμογή στα δεδομένα εκπαίδευσης και το παραπάνω επηρεάζει αρνητικά την ακρίβεια του μοντέλου. Το παραπάνω δημιούργησε την ιδέα για την κατασκευή των τυχαίων δασών δηλαδή συνόλων από ασυσχέτιστα μεταξύ τους δέντρα απόφασης. Τα τυχαία δάση διατηρούν τα

πλεονεκτήματα των δέντρων απόφασης και παράλληλα διαχειρίζονται με αποτελεσματικό τρόπο τα μειονεκτήματά τους και για το λόγο αυτό, τα συγκεκριμένα αποτελούν μία ιδιαίτερα δημοφιλή τεχνική ταξινόμησης. Στόχος της παρούσας εργασίας είναι η εφαρμογή των παραπάνω αλγορίθμων σε υψηλής χωρικής ανάλυσης τηλεπισκοπικά δεδομένα και πιο συγκεκριμένα η διερεύνηση της αποτελεσματικότητας τους ειδικά σε ό,τι αφορά την ανίχνευση κτιρίων.

1.4 Λίγα λόγια για το eCognition

Το eCognition αποτελεί ένα λογισμικό ανάλυσης εικόνων το οποίο διευκολύνει την γρήγορη και ακριβή εξαγωγή γεω-πληροφορίων από τηλεπισκοπικά δεδομένα. Αποτελεί το πρωτότυπο λογισμικό αντικειμενοστρεφούς ανάλυσης εικόνας (OBIA) και έχει χρησιμοποιηθεί ευρέως από κορυφαίους παρόχους δεδομένων, επαγγελματίες τηλεπισκόπησης καθώς και πανεπιστημιακούς ερευνητές σε εφαρμογές όπως ο αστικός σχεδιασμός, η δασοκομία, η γεωργία για περισσότερο από δέκα χρόνια. Το eCognition προσφέρει μια ολοκληρωμένη σειρά εργαλείων για την ανάπτυξη ισχυρών εφαρμογών ανάλυσης εικόνας και είναι σε θέση να χειριστεί όλες τα είδη δεδομένων όπως χαμηλής καθώς και υψηλής ανάλυσης δορυφορικά δεδομένα, αεροφωτογραφίες, δεδομένα ραντάρ, καθώς και υπερφασματικά δεδομένα[1].

1.5 Δορυφόροι Pléiades

Οι Pléiades είναι ένα σύστημα δορυφόρων, οι οποίοι παρέχουν στο κοινό εικόνες υψηλής ευκρίνειας. Αποτελείται από δύο δέκτες, οι οποίοι προσφέρουν χωρική ανάλυση στο ναδίρ της τάξης των 0,5 μέτρων και οπτικό πεδίο 20 χιλιομέτρων. Ο πρώτος από αυτούς, ο Pléiades 1-A εκτοξεύτηκε στις 17 Δεκεμβρίου 2011 και ακολούθησε ο δεύτερος στα τέλη του έτους 2012.

Οι δορυφόροι αυτοί παρέχουν τη δυνατότητα καθημερινής πρόσβασης σε πληροφορία αντλούμενη από όλο τον κόσμο, γεγονός που ικανοποιεί ανάγκες που αφορούν στην άμυνα καθώς και την προστασία του πολίτη. Επιπροσθέτως, οι απεικονίσεις Pléiades μπορούν να χρησιμοποιηθούν σε εφαρμογές χαρτογραφίας, καθώς η κλίμακα απεικόνισης τους είναι καλύτερη από εκείνη των δορυφόρων SPOT. Τέλος, αξίζει να σημειωθεί πως η υψηλή χωρική ανάλυση των συγκεκριμένων δορυφορικών εικόνων παρέχει τη δυνατότητα χαρτογράφησης του αστικού περιβάλλοντος και προσφέρει συμπληρωματική πληροφορία σε δεδομένα αντλούμενα από αεροφωτογραφίες (Tinel et al., 2012).

Στον πίνακα που ακολουθεί (Πίνακας 1.1) παρουσιάζονται κάποια από τα χαρακτηριστικά του συγκεκριμένου συστήματος δορυφόρων.

ΠΙΝΑΚΑΣ 1.1: ΚΥΡΙΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΤΩΝ ΔΟΡΥΦΟΡΩΝ PLEIADES (PLEIADES IMAGERY USERS GUIDE, 2012)

Πλήθος δορυφόρων	2: Pléiades 1-A, Pléiades 1-B
Υψόμετρο	694 km
Τροχιά	Ηλιοσύχρονη, 10:30 AM descending node
Περίοδος	98.79 min
Γωνία κλίσης	98.2''
Φασματική κλίση	Pan: 0.47-0.83 μm, Blue: 0.43-0.55 μm,

[1] http://www.eCognition.com/sites/default/files/Trimble%20eCognition.pdf

	Green: 0.50-0.62 μm, Near Infrared: 0.74-0.94 μm (NIR)
Χωρική διακριτική ικανότητα	Panchromatic: 0.5m, Multisprectral: 2.0m

1.6 Στόχοι

Μέσω της παρούσας μεταπτυχιακής εργασίας επιχειρούνται τα ακόλουθα:

- Διερεύνηση της υπάρχουσας βιβλιογραφίας για τον αλγόριθμο των δέντρων απόφασης
- Διερεύνηση της υπάρχουσας βιβλιογραφίας που αφορά στην εφαρμογή του αλγορίθμου των δέντρων απόφασης στην επιστήμη της Ψηφιακής Τηλεπισκόπησης
- Διερεύνηση της υπάρχουσας βιβλιογραφίας για τον αλγόριθμο των τυχαίων δασών
- Διερεύνηση της υπάρχουσας βιβλιογραφίας που αφορά στην εφαρμογή του αλγορίθμου των τυχαίων δασών στην επιστήμη της Ψηφιακής Τηλεπισκόπησης
- Διερεύνηση της αποτελεσματικότητας της εφαρμογής των παραπάνω αλγορίθμων σε τηλεπισκοπικά δεδομένα σε ό,τι αφορά την ανίχνευση κτιρίων
- Επιλογή τιμών παραμέτρων
- Επιλογή χαρακτηριστικών
- Σύγκριση των επιδόσεων των παραπάνω αλγορίθμων με εκείνες του Εγγύτερου Γείτονα σε ό,τι αφορά την ανίχνευση κτιρίων

2 Βιβλιογραφική ανασκόπηση

2.1 Δέντρα απόφασης

Στην παρούσα ενότητα γίνεται αναλυτική περιγραφή μίας τεχνικής ταξινόμησης εκείνης των δέντρων απόφασης. Τα συγκεκριμένα μοντέλα είναι ιδανικά για την επίλυση προβλημάτων που περιγράφονται από μία αλληλουχία αποφάσεων, καθώς αποδίδουν γραφικά τις ενέργειες που πρέπει να γίνουν, τα γεγονότα που ενδέχεται να προκύψουν καθώς και τα αποτελέσματα που συνδέονται με ένα σύνολο των παραπάνω αποφάσεων και γεγονότων. Συνεπώς, βασικός στόχος ενός δέντρου απόφασης είναι ο προσδιορισμός των βέλτιστων λύσεων σε μία κατάσταση.[2]

Τα δέντρα απόφασης αποτελούν, ουσιαστικά, ένα είδος ταξινομητή, ο οποίος μπορεί να εκφραστεί ως μία επαναληπτική διαδικασία διαχωρισμού των δεδομένων σε θεματικές κατηγορίες. Τα παραπάνω συντίθενται από μία αλληλουχία πιθανών ερωτήσεων προκειμένου να καταστεί δυνατή η επίλυση ενός προβλήματος (Tan et al., 2005) (Rokach and Maimon, 2005).

Στο σημείο αυτό κρίνεται σκόπιμο να γίνει επιγραμματική περιγραφή των όρων που συνδέονται με ένα δέντρο απόφασης, καθώς η διαδικασία αυτή θα συντελέσει ουσιαστικά στην κατανόηση της δομής και της λειτουργίας του εν λόγω ταξινομητή. Αρχικά, σημειώνεται πως το συγκεκριμένο μοντέλο ταξινόμησης αποτελεί ουσιαστικά ένα είδος γράφου. Ένας γράφος $G = (V, E)$ αποτελείται από πεπερασμένο αριθμό κόμβων (η κορυφών) V καθώς και από ένα σύνολο ακμών E και ονομάζεται κατευθυνόμενος στην περίπτωση που οι ακμές είναι διατεταγμένα ζεύγη (v, w) των κόμβων. Ένα μονοπάτι είναι μία αλληλουχία ακμών της μορφής $(v_1, v_2), (v_2, v_3), \dots, (v_{n-1}, v_n)$. Το συγκεκριμένο ονομάζεται μονοπάτι από τη v_1 στη v_n και το μήκος του είναι ίσο με n. Στην περίπτωση που ο γράφος δεν περιέχει κύκλους (δηλαδή οι διαδρομές δύο ή περισσοτέρων κόμβων που καταλήγουν στον κόμβο αρχής) ονομάζεται κατευθυνόμενος μη κυκλικός. Ένα κατευθυνόμενο (ή ριζωμένο) δέντρο απόφασης είναι ουσιαστικά ένα είδος κατευθυνόμενου μη κυκλικού γράφου το οποίο πληροί τις ακόλουθες ιδιότητες:

- Υπάρχει ακριβώς ένας κόμβος ο οποίος δεν έχει εισερχόμενες ακμές. Ο συγκεκριμένος ονομάζεται ρίζα του δέντρου και περιλαμβάνει εγγραφές από όλα τα είδη κλάσεων
- Κάθε κόμβος του δέντρου πλην της ρίζας έχει ακριβώς μία εισερχόμενη ακμή
- Το μονοπάτι από τη ρίζα στον εκάστοτε κόμβο του δέντρου είναι μοναδικό (SafanianS. R., LandgrebeD., 1990)

Βάσει των παραπάνω προκύπτει πως τα δέντρα απόφασης αποτελούνται από κόμβους, οι οποίοι σχηματίζουν «ριζωμένα» δέντρα (rooted tree). Αυτό σημαίνει πως τα τελευταία είναι κατευθυνόμενα και αφετηρία για το κάθε ένα από αυτά είναι ο κόμβος ρίζα (root). Αναλυτικά, στα δέντρα απόφασης διακρίνονται τα ακόλουθα είδη κόμβων:

- Ρίζα: ο κόμβος αυτός έχει μόνο εξερχόμενα κλαδιά.
- Εσωτερικοί κόμβοι: οι συγκεκριμένοι έχουν εισερχόμενα και εξερχόμενα κλαδιά
- Φύλλα: Οι κόμβοι αυτοί έχουν μόνο εισερχόμενα κλαδιά (Rokach and Maimon, 2005)

[2]http://treeplan.com/

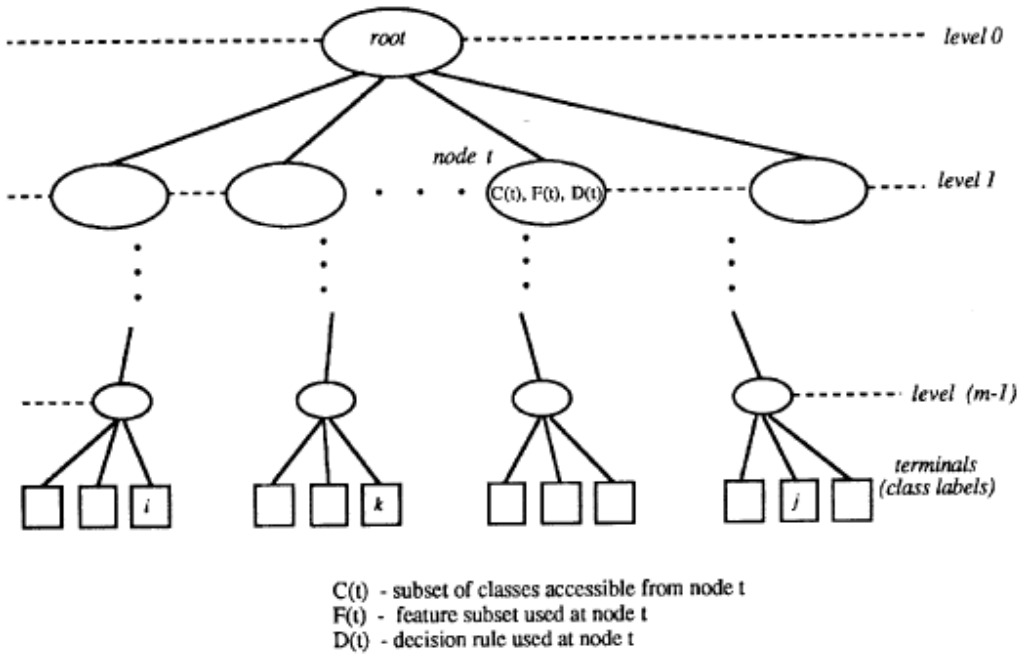
Κάθε φύλλο του δέντρου αντιστοιχίζεται σε μία κλάση. Τα υπόλοιπα είδη των κόμβων αντιστοιχίζονται σε κάποιο γνώρισμα των αντικειμένων βάσει του οποίου γίνεται ο διαχωρισμός του κόμβου σε παιδιά. Στο σημείο αυτό, διευκρινίζεται πως, ο κόμβος v σε μία ακμή (v, w) ονομάζεται πατέρας του v και αντίστοιχα ο w γιος του v (Tan et al., 2005) (Safanian and Landgrebe , 1990).

Ορισμένα βασικά χαρακτηριστικά των δέντρων απόφασης είναι τα ακόλουθα:

- Βάθος ενός κόμβου: ο αριθμός των ακμών από τη ρίζα του δέντρου
- Ύψος ενός κόμβου είναι το πλήθος των ακμών από τον κόμβο έως το φύλλο του δέντρου[3]

Βασικά χαρακτηριστικά τα οποία περιγράφουν ένα κόμβο είναι το βάθος και το ύψος του. Αναλυτικά, βάθος ενός κόμβου v είναι το μήκος της διαδρομής από τη ρίζα έως το v, ενώ ύψος του v είναι το μήκος της μεγαλύτερης διαδρομής από τον v στη ρίζα. Βάσει του παραπάνω προκύπτει πως το ύψος του δέντρου είναι ουσιαστικά το ύψος της ρίζας του. Τέλος, σημειώνεται πως το επίπεδο ενός κόμβου vείναι ίσο με τη διαφορά του βάθους του δέντρου από το ύψος του δέντρου (Safanian and Landgrebe , 1990).

Στην Εικόνα 2.1 εμφανίζονται γραφικά τα προαναφερθέντα τμήματα και χαρακτηριστικά ενός δέντρου απόφασης.



ΕΙΚΟΝΑ 2.1: ΠΑΡΑΔΕΙΓΜΑ ΕΝΟΣ ΔΕΝΤΡΟ ΑΠΟΦΑΣΗΣ

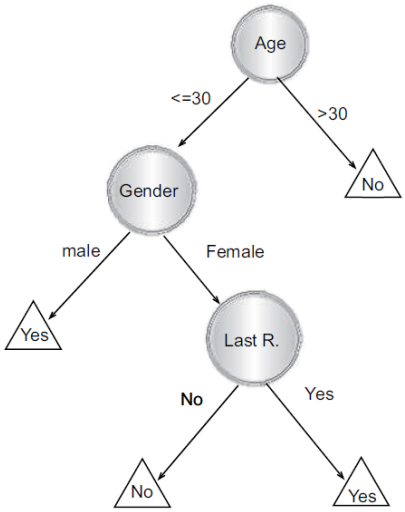
Η ταξινόμηση ενός αντικειμένου είναι ιδιαίτερα απλή, εφόσον ολοκληρωθεί η δημιουργία του δέντρου. Οι εγγραφές κατατάσσονται σε κλάσεις μέσω πλοήγησης από τη ρίζα έως και τα φύλλα του δέντρου. Αφετηρία της διαδικασίας είναι ο κόμβος ρίζα όπου εκεί γίνεται ένας πρώτος έλεγχος βάσει των χαρακτηριστικών της εγγραφής. Στη συνέχεια, βάσει του παραπάνω γίνεται μετάβαση σε εσωτερικό κόμβο στον οποίο πραγματοποιείται νέος έλεγχος. Η διαδικασία αυτή επαναλαμβάνεται έως ότου η διαδρομή φτάσει στα φύλλα του

[3] https://www.cs.cmu.edu/~adamchik/15-121/lectures/Trees/trees.html

δέντρου (Tan et al., 2005). Αξίζει να σημειωθεί, μάλιστα, πως κάθε διαδρομή από την ρίζα έως και τα φύλλα μπορεί εύκολα να μετατραπεί σε κανόνα συνδέοντας τους επιμέρους ελέγχους που εμφανίζονται στους κόμβους αυτής. Στην Εικόνα 2.2 εμφανίζεται ένα δέντρο απόφασης στο οποίο απεικονίζεται η ανταπόκριση σε Direct Mailing βάσει του φύλου και της ηλικίας (Rokach and Maimon, 2005)(Tan et al., 2005).

Πλέον, οι κατασκευαστές δέντρων απόφασης προτιμούν λιγότερο περίπλοκά δέντρα καθώς τα συγκεκριμένα είναι πιο εύκολο να ερμηνευθούν. Η πολυπλοκότητα του δέντρου ελέγχεται αποκλειστικά από το κριτήριο τερματισμού (stopping criteria) καθώς και τη μέθοδο κλαδέματος (pruning method) που εφαρμόζεται και η συγκεκριμένη ιδιότητα μετριέται μέσω των ακόλουθων μεγεθών:

- το συνολικό αριθμό των κόμβων
- το συνολικό αριθμό των φύλλων
- το βάθος του δέντρου (δηλαδή ο αριθμός των κόμβων της μεγαλύτερης διαδρομής από τη ρίζα έως τα φύλλα)
- τον αριθμό των χαρακτηριστικών που χρησιμοποιήθηκαν (Rokach and Maimon, 2005)



ΕΙΚΟΝΑ 2.2: ΥΠΟΔΕΙΓΜΑ ΤΑΞΙΝΟΜΗΣΗΣ ΜΕΣΩ ΔΕΝΤΡΟΥ ΑΠΟΦΑΣΗΣ (ROKACH AND MAIMON, 2005)

2.1.1 Κατασκευή δέντρου απόφασης

Αρχικά, διευκρινίζεται πως από κάθε σύνολο εκπαίδευσης μπορεί να κατασκευαστεί μεγάλος αριθμός από διαφορετικά δέντρα απόφασης. Η εύρεση του βέλτιστου μοντέλου από το παραπάνω σύνολο είναι ένα εγχείρημα ακατόρθωτο λόγω του πεπερασμένου χώρου αποθήκευσης και για το σκοπό αυτό, έχουν επινοηθεί αλγόριθμοι κατασκευής δέντρων απόφασης. Τα μοντέλα που προκύπτουν είναι όσο το δυνατόν ακριβέστερα και αποδοτικότερα σε ό,τι αφορά το χρόνο (Tan et al., 2005).

Ένας αλγόριθμος κατασκευής δέντρων απόφασης καλείται να αντιμετωπίσει τα ακόλουθα δύο ζητήματα:

- Πώς γίνεται διαχωρισμός των δεδομένων εκπαίδευσης;
- Πότε τερματίζεται η διαδικασία διαχωρισμού;

Αλγόριθμοι κατασκευής δέντρων απόφασης

Οι inducers των δέντρων απόφασης είναι αλγόριθμοι που αυτόματα κατασκευάζουν τα συγκεκριμένα μοντέλα βάσει ενός δοθέντος σετ δεδομένων. Συνήθως, ο στόχος είναι να βρεθεί το βέλτιστο δέντρο με το μικρότερο σφάλμα γενίκευσης (αναλύεται διεξοδικά παρακάτω). Ωστόσο υπάρχει δυνατότητα να ληφθούν υπόψη περαιτέρω κριτήρια όπως η ελαχιστοποίηση του αριθμού των κόμβων ή του μέσου βάθους του δέντρου. Γενικά, η εύρεση του βέλτιστου δέντρου απόφασης από ένα δοθέν σετ δεδομένων αποτελεί ένα εγχείρημα ιδιαίτερα απαιτητικό (Rokach and Maimon, 2005).

Οι αλγόριθμοι υιοθετούν δύο βασικές προσεγγίσεις διαχωρισμού των δεδομένων, την από πάνω προς τα κάτω (top down) και την από κάτω προς τα πάνω (bottom up). Η βιβλιογραφία εστιάζει περισσότερο στην πρώτη κατηγορία στην οποία εντάσσονται πληθώρα αλγορίθμων όπως ο ID3, C4.5 και ο CART. Ορισμένοι από αυτούς όπως οι C4.5 και CART διαρθρώνονται σε δύο στάδια: το «μεγάλωμα»(growing) και το κλάδεμα (pruning) του δέντρου, ενώ υπάρχουν και εκείνοι οι οποίοι περιλαμβάνουν αποκλειστικά το στάδιο του μεγαλώματος (Rokach and Maimon, 2005).

Στην Εικόνα 2.3 εμφανίζεται ένα υπόδειγμα αλγορίθμου κατασκευής δέντρων απόφασης. Το στοιχείο εισόδου του αλγορίθμου είναι ένα σύνολο δεδομένων εκπαίδευσης E καθώς και ένα σύνολο χαρακτηριστικών F. Ο αλγόριθμος μέσω του βήματος 7 βρίσκει επαναληπτικά το βέλτιστο κριτήριο διαχωρισμού των δεδομένων και στη συνέχεια επεκτείνει τους κόμβους φύλλα (βήμα 11 και 12) έως ότου ικανοποιηθεί κάποιο κριτήριο τερματισμού. Στη συνέχεια, παρατίθενται λεπτομέρειες σχετικά με τις επιμέρους συναρτήσεις του αλγορίθμου:

I. Η συνάρτηση **createNode()** επεκτείνει το δέντρο απόφασης δημιουργώντας ένα καινούργιο κόμβο. Ο συγκεκριμένος ενδέχεται να είναι είτε εσωτερικός (*node.test_cond*), είτε φύλλο (*node.label*). Διευκρινίζεται πως ο εσωτερικός περιέχει κάποιο έλεγχο συνθήκης ενώ το φύλλο την ετικέτα κάποιας κλάσης
II. Η συνάρτηση **find_best_split()** προσδιορίζει το χαρακτηριστικό εκείνο βάσει του οποίου θα γίνει ο έλεγχος συνθήκης. Η επιλογή του ελέγχου αυτού βασίζεται σε μέτρα μη καθαρότητας όπως η εντροπία, ο δείκτης Gini και τα στατιστικά χ^2 (αναλύονται παρακάτω στο χωρίο Μέτρα επιλογής του βέλτιστου δυνατού διαχωρισμού).
III. Η συνάρτηση **Classify()** προσδιορίζει την ετικέτα της κλάσης η οποία θα ανατεθεί σε κάποιον κόμβο - φύλλο. Στις περισσότερες περιπτώσεις η θεματική κατηγόρια που επιλέγεται είναι εκείνη με το μεγαλύτερο αριθμό εγγραφών στο συγκεκριμένο κόμβο δηλαδή:

$$leaf.label = \arg\max_i p(i|t)$$

Όπου ο τελεστής $argmax$ επιστρέφει το όρισμα i, το οποίο μεγιστοποεί τη συνθήκη $p(i|t)$

IV. Η συνάρτηση **stopping_cond()** χρησιμοποιείται προκειμένου να τερματιστεί η διαδικασία διαχωρισμού των κόμβων. Αυτό πρακτικά υλοποιείται μέσω ενός ελέγχου στα χαρακτηριστικά ή στις κλάσεις των εγγραφών. Στην περίπτωση που οι εγγραφές ενός κόμβου ανήκουν στην ίδια θεματική κατηγορία ή έχουν παρόμοια γνωρίσματα ο διαχωρισμός τερματίζεται. Ένας ακόμα τρόπος διακοπής της διαδικασίας είναι ο έλεγχος κατά πόσον το πλήθος των εγγραφών στον κόμβο είναι μικρότερο από ένα ορισμένο κατώφλι (Tan et al., 2005).

Οι συνθήκες τερματισμού συνοψίζονται ως εξής:

- Όλα τα αντικείμενα του δείγματος εκπαίδευσης αντιστοιχίζονται σε μία μοναδική τιμή χαρακτηριστικού y
- Το βάθος του δέντρου είναι το μέγιστο δυνατό
- Ο αριθμός των εγγραφών στους τερματικούς κόμβους είναι μικρότερος από τον ελάχιστο αριθμό εγγράφων στους κόμβους –γονείς
- Σε περίπτωση διαχωρισμού ο αριθμός των εγγραφών στους κόμβους παιδιά είναι μικρότερος από ένα ορισμένο κατώφλι
- Το βέλτιστο δυνατό κριτήριο διαχωρισμού δεν είναι μεγαλύτερο από ένα συγκεκριμένο κατώφλι (Rokach and Maimon, 2005)

Εφόσον ολοκληρωθεί η διαδικασία σχηματισμού του δέντρου σε ορισμένους αλγορίθμους ακολουθεί η φάση του κλαδέματος του δέντρου. Τα μεγάλα σε μέγεθος δέντρα απόφασης ενδέχεται να εμφανίζουν υπερπροσαρμογή (overfitting) στα δεδομένα εκπαίδευσης (Οι έννοιες αυτές αναλύονται διεξοδικά στις ακόλουθες ενότητες). Στις περιπτώσεις αυτές η αφαίρεση ορισμένων από τα κλαδιά του αρχικού δέντρου βελτιώνει τις επιδόσεις του τελευταίου (Εικόνα 2.4).

Algorithm 4.1 A skeleton decision tree induction algorithm.

```
TreeGrowth (E, F)
 1: if stopping_cond(E,F) = true then
 2:    leaf = createNode().
 3:    leaf.label = Classify(E).
 4:    return leaf.
 5: else
 6:    root = createNode().
 7:    root.test_cond = find_best_split(E, F).
 8:    let V = {v|v is a possible outcome of root.test_cond }.
 9:    for each v ∈ V do
10:       E_v = {e | root.test_cond(e) = v and e ∈ E}.
11:       child = TreeGrowth(E_v, F).
12:       add child as descendent of root and label the edge (root → child) as v.
13:    end for
14: end if
15: return root.
```

ΕΙΚΟΝΑ 2.3: ΥΠΟΔΕΙΓΜΑ ΑΛΓΟΡΙΘΜΟΥ ΔΗΜΙΟΥΡΓΙΑΣ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΜΕ ΠΡΟΚΛΑΔΕΜΑ (ROKACH AND MAIMON, 2005)

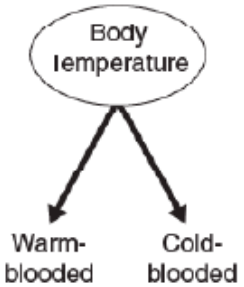
```
TreeGrowing (S,A,y)
Where:
S - Training Set
A - Input Feature Set
y - Target Feature
Create a new tree T with a single root node.
IF One of the Stopping Criteria is fulfilled THEN
    Mark the root node in T as a leaf with the most
    common value of y in S as a label.
ELSE
    Find a discrete function f(A) of the input
        attributes values such that splitting S
        according to f(A)'s outcomes (v1,...,vn) gains
        the best splitting metric.
    IF best splitting metric > treshold THEN
        Label t with f(A)
        FOR each outcome vi of f(A):
            Set Subtreei= TreeGrowing (σf(A)=vi S,A,y).
            Connect the root node of tT to Subtreei with
                    an edge that is labelled as vi
        END FOR
    ELSE
        Mark the root node in T as a leaf with the most
            common value of y in S as a label.
    END IF
END IF
RETURN T
TreePruning (S,T,y)
Where:
S - Training Set
y - Target Feature
T - The tree to be pruned
DO
    Select a node t in T such that pruning it
        maximally improve some evaluation criteria
    IF t≠∅ THEN T=pruned(T,t)
UNTIL t=∅
RETURN T
```

ΕΙΚΟΝΑ 2.4: ΥΠΟΔΕΙΓΜΑ ΑΛΓΟΡΙΘΜΟΥ ΔΗΜΙΟΥΡΓΙΑ ΔΕΝΤΡΟΥ ΑΠΟΦΑΣΗΣ ΜΕ ΜΕΤΑΚΛΑΔΕΜΑ

Έλεγχος χαρακτηριστικών

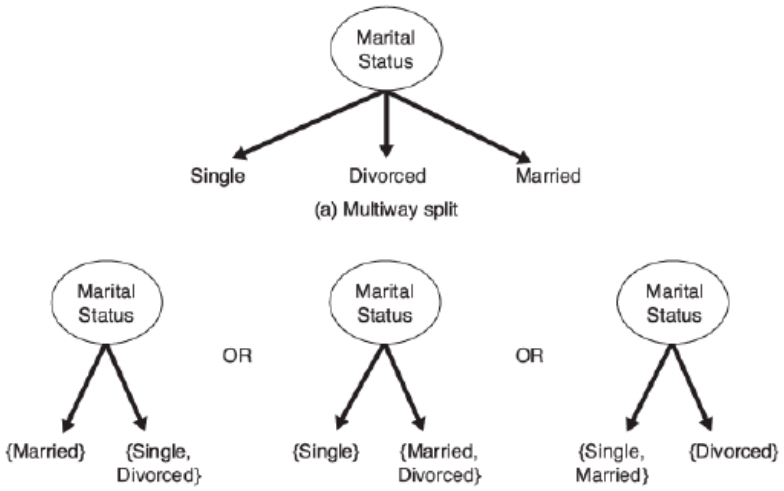
Οι αλγόριθμοι κατασκευής δέντρων απόφασης παρέχουν μεθόδους τέτοιες ώστε να είναι δυνατή η έκφραση ελέγχου συνθήκης για διαφορετικά είδη γνωρισμάτων (Tan et al., 2005).

- Δυαδικά χαρακτηριστικά: Η συγκεκριμένη συνθήκη έχει δύο πιθανά αποτελέσματα (Εικόνα 2.5)



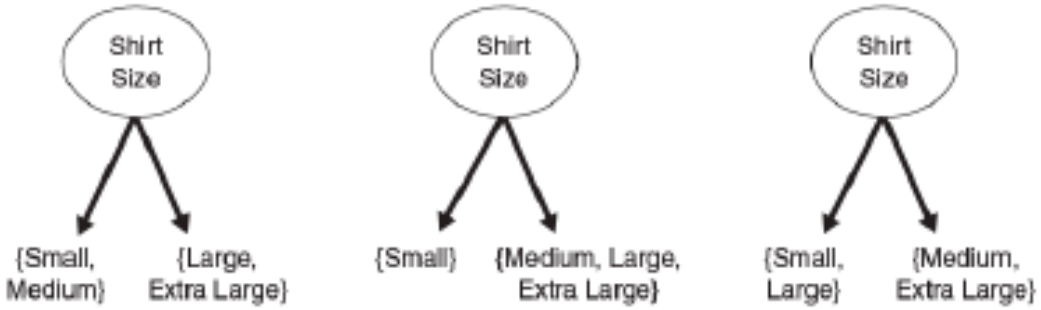
ΕΙΚΟΝΑ 2.5: ΠΑΡΑΔΕΙΓΜΑ ΔΥΑΔΙΚΟΥ ΧΑΡΑΚΤΗΡΙΣΤΙΚΟΥ (TAN ET AL., 2005).

- Διακριτά/ονομαστικά χαρακτηριστικά (nominal): Τα διακριτά χαρακτηριστικά ενδέχεται να παίρνουν πολλές διαφορετικές τιμές και η συνθήκη ελέγχου μπορεί να εκφραστεί με δύο τρόπους (Εικόνα 2.6). Μέσω του πρώτου το πλήθος των εξερχόμενων κόμβων είναι ίσο με τον αριθμό των διαφορετικών χαρακτηριστικών (πολλαπλός διαχωρισμός). Σε ορισμένους αλγορίθμους τα δέντρα απόφασης που κατασκευάζονται έχουν μόνο δυαδικούς κόμβους. Στις περιπτώσεις αυτές υπάρχουν $2^{κ-1}$ διαφορετικοί τρόποι δυαδικού διαμερισμού των κ χαρακτηριστικών (Tan et al., 2005).



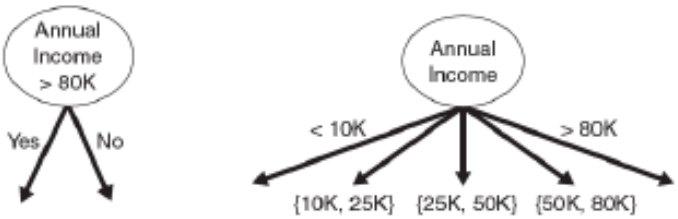
ΕΙΚΟΝΑ 2.6: ΔΙΑΚΡΙΤΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ (TAN ET AL., 2005)

- Διατεταγμένα χαρακτηριστικά: Ομοίως με τα διακριτά τα διατεταγμένα χαρακτηριστικά αποτυπώνονται μέσω δυαδικών και πολλαπλών διαχωρισμών. Τα διατεταγμένα χαρακτηριστικά είναι δυνατόν να ομαδοποιηθούν σε δυαδικούς κόμβους εφόσον δεν αλλάζει η σωστή σειρά των τιμών (Εικόνα 2.7)(Tan et al., 2005).



ΕΙΚΟΝΑ 2.7: ΔΙΑΤΕΤΑΓΜΕΝΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ (TAN ET AL., 2005).

- Συνεχή χαρακτηριστικά: Η συνθήκη ελέγχου στην περίπτωση των συνεχών χαρακτηριστικών μπορούν να εκφραστούν μέσω μίας ή περισσότερων συγκριτικών δοκιμών (Εικόνα 2.8) (Tan et al., 2005).



ΕΙΚΟΝΑ 2.8: ΣΥΝΕΧΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ (TAN ET AL., 2005)

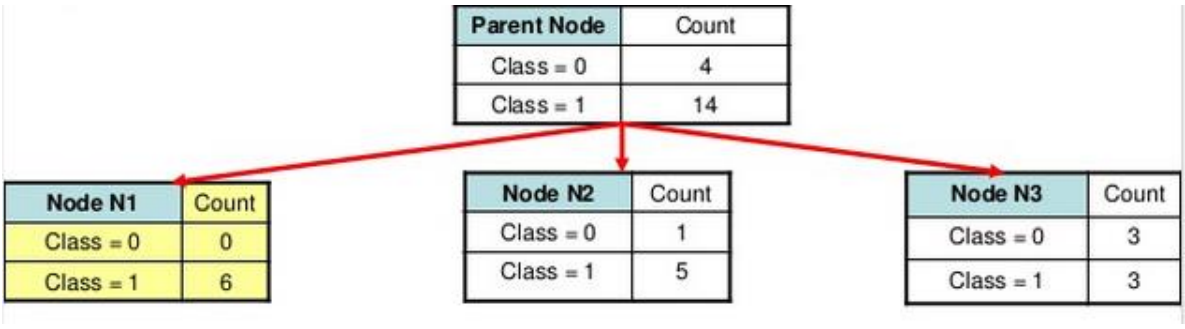
Μέτρα επιλογής του βέλτιστου δυνατού διαχωρισμού

ΜΕΤΡΑ ΜΗ ΚΑΘΑΡΟΤΗΤΑΣ

Στη βιβλιογραφία καταγράφεται πλήθος διαφορετικών μέτρων διαχωρισμού των δεδομένων. Τα συγκεκριμένα συγκρίνουν την κατανομή των κλάσεων των εγγραφών πριν και μετά το διαχωρισμό του κόμβου στα επιμέρους παιδιά. Διαισθητικά προτιμώνται εκείνοι με ομοιογενείς κατανομές κλάσεων και συνεπώς σε μία ιδανική περίπτωση όλες οι εγγραφές ανήκουν σε μία και μόνο θεματική κατηγορία. Το παραπάνω εκφράζεται στη βιβλιογραφία ως μέτρο μη καθαρότητας του κόμβου (impurity measure). Οι κόμβοι με τη μεγαλύτερη ομοιογένεια ως προς την κατανομή των κλάσεων έχουν μικρότερο βαθμό μη καθαρότητας και ο τελευταίος παίρνει την ελάχιστη δυνατή τιμή (impurity measure=0) στην περίπτωση που όλες οι εγγραφές ανήκουν στην ίδια κλάση. Αντιθέτως, η μέγιστη εμφανίζεται όταν οι εγγραφές είναι ομοιόμορφα κατανεμημένες στις υπάρχουσες θεματικές κατηγορίες (Tan et al., 2005).

Στην Εικόνα 2.9 ο κόμβος γονέας (Parent Node) έχει διαχωριστεί σε τρεις διαφορετικούς κόμβους παιδιά, οι οποίοι διαφέρουν ως προς τις κατανομές των εγγραφών τους στις κλάσεις 0 και 1. Αν με I συμβολιστεί το μέτρο μη καθαρότητας κάθε κόμβου τότε σύμφωνα με τα παραπάνω ισχύει πως:

$$I(N1) < I(N2) < I(N3)$$



ΕΙΚΟΝΑ 2.9: ΠΑΡΑΔΕΙΓΜΑΤΑ ΟΜΟΙΟΓΕΝΕΙΑΣ ΚΟΜΒΟΥ[4]

Στη βιβλιογραφία καταγράφονται διάφορα μέτρα σχετικά με τη μη καθαρότητα ενός κόμβου. Τα πλέον διαδεδομένα εξ αυτών είναι η εντροπία (Entropy), ο δείκτης GINI (GINI Index) και οι λάθος ταξινομήσεις (Misclassification Error):

[4](http://www.slideshare.net/)

$$Entropy\ (t) = -\sum_{i=0}^{c-1} p(i|t) \log_2 p(i|t)$$

$$Gini(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2$$

$$Classification\ error(t) = 1 - \max_i [p(i|t)]$$

Όπου:

- Με $p(i|t)$ συμβολίζεται η σχετική συχνότητα της κλάσης i στον κόμβο t. Μέσω του συγκεκριμένου μεγέθους υπολογίζεται ουσιαστικά το ποσοστό εγγραφών της κλάσης i στον κόμβο t
- c είναι το πλήθος των κλάσεων

Σημειώνεται πως στους παραπάνω υπολογισμούς θεωρείται πως $0 \log_2 0 = 0$.

Στο γράφημα της Εικόνα 2.10 γίνεται σύγκριση των μέτρων μη καθαρότητας στην περίπτωση μίας δυαδικής ταξινόμησης. Παρατηρείται πως τα τρία μέτρα μεγιστοποιούνται όταν η κατανομή των εγγραφών στις κλάσεις είναι ομοιόμορφη, συνεπώς όταν η σχετική συχνότητα των κλάσεων είναι $p = 0{,}5$. Επίσης, σημειώνεται πως η ελάχιστη τιμή των παραπάνω μέτρων είναι ίση με μηδέν και αυτό συμβαίνει όταν $p = 0$ ή 1. Στην Εικόνα 2.11 παρουσιάζονται τα αποτελέσματα υπολογισμού των τριών μέτρων καθαρότητας για τρεις διαφορετικούς κόμβους. Μέσω των δύο Εικόνων(Εικόνα 2.10, Εικόνα 2.11) αποτυπώνεται η συνέπεια των τιμών στα τρία διαφορετικά μέτρα ομοιότητας (Tan et al., 2005).

Ένας κόμβος γονέας μπορεί να διασπαστεί σε πλήθος διαφορετικών συνδυασμών από κόμβους παιδιά. Το ερώτημα που εύλογα γεννάται είναι πώς επιλέγεται η βέλτιστη δυνατή διάσπαση ενός κόμβου από ένα μεγάλο αριθμό διαφορετικών εναλλακτικών. Για το σκοπό αυτό, γίνεται σύγκριση του μέτρου μη καθαρότητας του κόμβου γονέα (πριν τη διάσπαση) και των αντίστοιχων για τα παιδιά (όπως προκύπτουν μετά τη διάσπαση). Υπολογίζεται με άλλα λόγια το κέρδος Δ της διάσπασης βάσει της ακόλουθης σχέσης:

$$\Delta = I(parent) - \sum_{j=1}^{k} \frac{N(u_j)}{N} I(u_j)$$

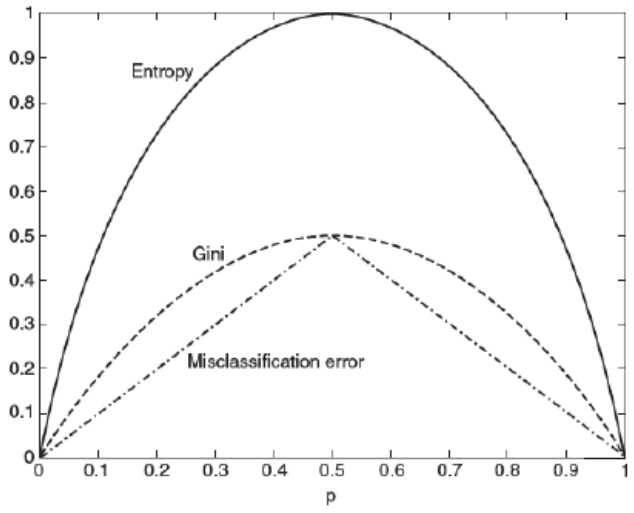
Όπου:

- $I()$ το μέτρο καθαρότητας του εκάστοτε κόμβου
- N ο συνολικός αριθμός των εγγραφών στον κόμβο γονέα
- $N(u_j)$ ο αριθμός των εγγραφών στον κόμβο παιδί
- k ο αριθμός των κόμβων παιδιών

Συνεπώς, μέσω του κέρδους Δ υπολογίζεται η διαφορά του σταθμισμένου μέσου όρου των μέτρων μη καθαρότητας των κόμβων παιδιών από το μέτρο μη καθαρότητας του πατέρα. Η εναλλακτική, η οποία επιλέγεται είναι εκείνη για την οποία το Δ παίρνει τη μέγιστη δυνατή τιμή. Διατηρείται δηλαδή η περίπτωση εκείνη στην οποία ο σταθμισμένος μέσος όρος της μη καθαρότητας των παιδιών ελαχιστοποιείται, καθώς η αντίστοιχη του γονέα είναι συγκεκριμένη και συνεπώς δε μεταβάλλεται για το πλήθος των εναλλακτικών διαχωρισμών. Τέλος, σημειώνεται πως το κέρδος Δ ονομάζεται και κέρδος πληροφορίας (information

gain) στην περίπτωση που στην παραπάνω εξίσωση ο υπολογισμός της μη καθαρότητας γίνεται μέσω της σχέσης της εντροπίας (Tan et al., 2005).

Σημειώνεται πως η παραπάνω διαδικασία επαναλαμβάνεται για το σύνολο των χαρακτηριστικών και τελικά επιλέγεται ο διαχωρισμός εκείνος που οδηγεί στη μέγιστη δυνατή αύξηση του κέρδους Δ.



ΕΙΚΟΝΑ 2.10: ΓΡΑΦΗΜΑ ΤΩΝ ΤΡΙΩΝ ΔΙΑΦΟΟΡΕΤΙΚΩΝ ΜΕΤΡΩΝ ΜΗ ΚΑΘΑΡΟΤΗΤΑΣ (TAN ET AL., 2005)

Node N_1	Count
Class=0	0
Class=1	6

$\text{Gini} = 1 - (0/6)^2 - (6/6)^2 = 0$
$\text{Entropy} = -(0/6)\log_2(0/6) - (6/6)\log_2(6/6) = 0$
$\text{Error} = 1 - \max[0/6, 6/6] = 0$

Node N_2	Count
Class=0	1
Class=1	5

$\text{Gini} = 1 - (1/6)^2 - (5/6)^2 = 0.278$
$\text{Entropy} = -(1/6)\log_2(1/6) - (5/6)\log_2(5/6) = 0.650$
$\text{Error} = 1 - \max[1/6, 5/6] = 0.167$

Node N_3	Count
Class=0	3
Class=1	3

$\text{Gini} = 1 - (3/6)^2 - (3/6)^2 = 0.5$
$\text{Entropy} = -(3/6)\log_2(3/6) - (3/6)\log_2(3/6) = 1$
$\text{Error} = 1 - \max[3/6, 3/6] = 0.5$

ΕΙΚΟΝΑ 2.11: ΥΠΟΛΟΓΙΣΜΟΣ ΜΕΤΡΩΝ ΜΗ ΚΑΘΑΡΟΤΗΤΑΣ ΓΙΑ ΤΡΕΙΣ ΔΙΑΦΟΡΕΤΙΚΟΥΣ ΚΟΜΒΟΥΣ (TAN ET AL., 2005)

Στις παραπάνω παραγράφους αναλύθηκαν διεξοδικά τα πλέον διαδεδομένα κριτήρια διαχωρισμού των δεδομένων σε επιμέρους κόμβους. Το DKM αποτελεί ακόμα ένα κριτήριο διαχωρισμού, το οποίο ομοίως με τα παραπάνω βασίζεται στη μη καθαρότητα των κόμβων. Είναι χρήσιμο στις περιπτώσεις χαρακτηριστικών δυαδικών κλάσεων και διατυπώνεται ως εξής:

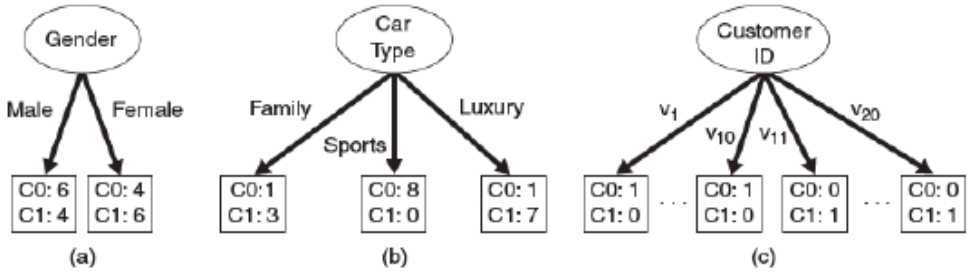
$$DKM(y,S) = 2\sqrt{\left(\frac{|\sigma_{y=c_1}S|}{|S|}\right)\left(\frac{|\sigma_{y=c_2}S|}{|S|}\right)}$$

Όπου:

- y είναι το εκάστοτε γνώρισμα
- S είναι το σύνολο των δεδομένων
- $\sigma_{y=c_i}$ είναι η πιθανότητα το γνώρισμα να πάρει την τιμή c_i

ΜΕΤΡΑ ΚΑΝΟΝΙΚΟΠΟΙΗΜΕΝΗΣ ΜΗ ΚΑΘΑΡΟΤΗΤΑΣ

Τα μέτρα μη καθαρότητας όπως για παράδειγμα η εντροπία και ο δείκτης GINI τείνουν να ευνοούν χαρακτηριστικά τα οποία λαμβάνουν μεγάλο αριθμό διαφορετικών διακριτών τιμών. Για παράδειγμα στην εφαρμογή της Εικόνα 2.12 εμφανίζονται τρεις διαφορετικές περιπτώσεις διαχωρισμού δεδομένων. Τα συμπεράσματα τα οποία προκύπτουν έπειτα από σύγκριση του (a), (b) και (c) είναι πως η αύξηση του αριθμού των κόμβων παιδιών αυξάνει πράγματι την ομοιογένεια των τελευταίων. Σημειώνεται, ωστόσο, πως στην τρίτη περίπτωση ο κωδικός πελάτη δεν αποτελεί προγνωστικό χαρακτηριστικό, καθώς η τιμή που παίρνει είναι διαφορετική για κάθε εγγραφή. Ακόμα και σε μία λιγότερο ακραία περίπτωση ο μεγάλος αριθμός παιδιών κόμβων δεν παράγει αξιόπιστα αποτελέσματα καθώς το πλήθος των εγγραφών σε κάθε ένα από αυτούς δεν είναι αρκετά μεγάλο ώστε να δώσει αξιόπιστα αποτελέσματα.



ΕΙΚΟΝΑ 2.12: ΠΕΡΙΠΤΩΣΕΙΣ ΔΙΑΦΟΡΕΤΙΚΩΝ ΔΙΑΧΩΡΙΣΜΩΝ (TAN ET AL., 2005)

Για την αντιμετώπιση του παραπάνω προβλήματος στη βιβλιογραφία καταγράφονται οι δύο ακόλουθες στρατηγικές. Η πρώτη από αυτές είναι η δημιουργία αυστηρά δύο κόμβων παιδιών. Η συγκεκριμένη εφαρμόζεται σε αλγορίθμους κατασκευής δέντρων απόφασης, όπως ο CART. Η δεύτερη στρατηγική προτείνει την τροποποίηση του κριτηρίου διαχωρισμού ώστε να λαμβάνεται υπόψη ο αριθμός των αποτελεσμάτων που προκύπτει από τον έλεγχο συνθήκης. Τέτοια κριτήρια είναι τα ακόλουθα (Tan et al., 2005):

- Λόγος Κέρδους

Το συγκεκριμένο αποτελεί το κριτήριο διαχωρισμού στον αλγόριθμο C4.5 και προσδιορίζει την ποιότητα του παραγόμενου αποτελέσματος. Το συγκεκριμένο υπολογίζεται βάσει της ακόλουθης σχέσης:

$$Gain\ ratio = \frac{\Delta_{info}}{Split\ Info}$$

Όπου το $Split\ Info = -\sum_{i=1}^{k} P(u_i) \log_2 P(u_i)$ με k το συνολικό αριθμό των διαχωρισμών. Για παράδειγμα, αν κάθε τιμή του χαρακτηριστικού έχει τον ίδιο αριθμό εγγραφών, τότε $\forall\, i: P(u_i) = 1/k$ και η πληροφορία αναφορικά με το διαχωρισμό είναι ίση με $\log_2 k$. Στην περίπτωση που ένα χαρακτηριστικό παράγει μεγάλο αριθμό διαχωρισμών η πληροφορία αναφορικά με το διαχωρισμό παίρνει μεγάλες τιμές και συνεπώς ο λόγος κέρδους μειώνεται (Tan et al., 2005).

- Μετρητής απόστασης (Distance Measure)

Ο μετρητής απόστασης όπως και ο λόγος κέρδους κανονικοποιεί το μέτρο καθαρότητας $\Delta\varphi$. Η σχέση που εφαρμόζεται είναι η ακόλουθη:

$$\frac{\Delta\varphi(a_i, S)}{-\sum_{u_{i,j}\in dom(a_i)}\sum_{c_k\in dom(y)}(\frac{\left|\sigma_{a_i=u_i\ AND\ y=c_k}\ S\right|}{|S|}\log_2\frac{\left|\sigma_{a_i=u_i\ AND\ y=c_k}\ S\right|}{|S|}}$$

(Rokach and Maimon, 2005)

- Κριτήριο Twoing

Ο δείκτης Gini ενδέχεται να εμφανίσει προβλήματα στις περιπτώσεις όπου το πεδίο ορισμού (domain) του χαρακτηριστικού στόχου είναι ευρύ. Στην περίπτωση αυτή, είναι πιθανόν να εφαρμοστεί ένα δυαδικό κριτήριο Twoing. Το συγκεκριμένο ορίζεται ως εξής:

$$twoing(ai, dom_1(a_i), dom_2(a_i), S)$$
$$= 0.25\frac{\left|\sigma_{a_i}\in\ dom_1(a_i)S\right|}{|S|}\frac{\left|\sigma_{a_i}\in\ dom_2(a_i)S\right|}{|S|}(\sum_{c_i\in dom(y)}\left|\frac{\left|\sigma_{a_i}\in\ dom_1(a_i)AND_y\in c_i\ S\right|}{\left|\sigma_{a_i}\in\ dom_1(a_i)S\right|}\right.$$
$$\left.-\frac{\left|\sigma_{a_i}\in\ dom_2(a_i)AND_y\in c_i\ S\right|}{\left|\sigma_{a_i}\in\ dom_2(a_i)S\right|}\right|)$$

Σημειώνεται πως σε δυαδικά χαρακτηριστικά οι τιμές των κριτηρίων Gini και Twoing είναι ίσες. Επιπροσθέτως, παρατηρείται πως σε προβλήματα πολλών κατηγοριών το κριτήριο Twoing «προτιμάει» χαρακτηριστικά ίσων διαστημάτων.

(Rokach and Maimon, 2005)

- Ορθογώνιο κριτήριο (Orthogonal (ORT) Criterion)

Το παρόν δυαδικό κριτήριο προτάθηκε από τους Fayyad και Irani (1992) και ορίζεται ως εξής:

$$ORT(a_i, dom_1(a_i), dom_2(a_i), S) = 1 - cos\theta(P_{y,1}, P_{y,2})$$

Όπου θ η γωνία των διανυσμάτων $P_{y,1}, P_{y,2}$. Τα διανύσματα αυτά αναπαριστούν την κατανομή πιθανότητας του χαρακτηριστικού στόχου στα διαστήματα $\sigma_{a_i}\in\ dom_1(a_i)S$ και $\sigma_{a_i}\in\ dom_2(a_i)S$ αντίστοιχα. Έχει αποδειχθεί πως το παρόν κριτήριο δίνει καλύτερα αποτελέσματα από εκείνο του πληροφοριακού κέρδους και του Gini σε συγκεκριμένες κατηγορίες προβλημάτων (Rokach and Maimon, 2005).

- Κριτήριο Kolmogorov, Smirnov

Το συγκεκριμένο κριτήριο χρησιμοποιεί την απόσταση Kolmogorov, Smirnov. Προτάθηκε από τους Friedman (1977) και Rounds (1980). Έχοντας ένα υποθετικό δυαδικό χαρακτηριστικό και πιο συγκεκριμένα το $\mathrm{dom(y)} = \{\mathrm{c}_1, \mathrm{c}_2\}$ το κριτήριο διατυπώνεται ως εξής:

$$KS(a_i, dom_1(a_i)dom_2(a_i), S)$$
$$= \sum_{c_i\in dom(y)}\left|\frac{\left|\sigma_{a_i}\in\ dom_1(a_i)AND_y\in c_1\ S\right|}{\left|\sigma_y = c_1 S\right|}\right.$$
$$\left.-\frac{\left|\sigma_{a_i}\in\ dom_2(a_i)AND_y\in c_2\ S\right|}{\left|\sigma_y = c_2 S\right|}\right|$$

Το παρόν μέτρο εμπλουτίστηκε από τους (Utgoff και Clouse, 1996) ούτως ώστε να διαχειρίζεται χαρακτηριστικά πολλαπλών κλάσεων. Τα αποτελέσματα μαρτυρούν πώς η προτεινόμενη μεθοδολογία ξεπερνάει σε αποτελεσματικότητα το κριτήριο αναλογία κέρδους (Rokach and Maimon, 2005).

- Λοιπά μονομεταβλητά κριτήρια

Στη βιβλιογραφία γίνεται αναφορά σε πρόσθετα μονομεταβλητά κριτήρια διαχωρισμού όπως σε στατιστικά στοιχεία μετάθεσης (permutation criteria) (Li and Dubes, 1986), μέσες ύστερες βελτιώσεις (mean posterior improvements) (Taylor and Silverman, 1993) και τα μέτρα υπεργεωμετρικής κατανομής (hypergeometric distribution measures) (Martin, 1997) (Rokach and Maimon, 2005).

ΣΥΓΚΡΙΣΗ ΤΩΝ ΜΟΝΟΜΕΤΑΒΛΗΤΩΝ ΚΡΙΤΗΡΙΩΝ ΔΙΑΧΩΡΙΣΜΟΥ

Τα παραπάνω κριτήρια έχουν αποτελέσει αντικείμενο έρευνας τα τελευταία τριάντα χρόνια. Τα συγκριτικά αποτελέσματα βασίζονται σε εμπειρικά καθώς και θεωρητικά συμπεράσματα.

Οι ερευνητές υποστηρίζουν πως η επιλογή του κριτηρίου διαχωρισμού δεν επηρεάζει ιδιαίτερα την απόδοση του δέντρου. Μάλιστα επισημαίνουν πως κάθε ένα από τα προαναφερθέντα κριτήρια επιτυγχάνει σε ορισμένες περιπτώσεις μεγαλύτερη και σε άλλες μικρότερη ακρίβεια συγκριτικά με τα υπόλοιπα (No-Free-Lunch θεώρημα) (Rokach and Maimon, 2005).

ΠΟΛΥΜΕΤΑΒΛΗΤΑ ΚΡΙΤΗΡΙΑ ΔΙΑΧΩΡΙΣΜΟΥ

Τα πολυμεταβλητά κριτήρια διαχωρισμού χρησιμοποιούν περισσότερα του ενός γνωρίσματα σε κάθε κόμβο του δέντρου. Συνεπώς, η εύρεση του βέλτιστου δυνατού διαχωρισμού αποτελεί μία διαδικασία περισσότερο πολύπλοκη συγκριτικά με εκείνη των μονομεταβλητών. Για τον παραπάνω λόγο, τα κριτήρια αυτά είναι λιγότερο διαδεδομένα σε σχέση με τα μονομεταβλητά παρά το γεγονός πως τα συγκεκριμένα μπορούν να βελτιώσουν δραματικά τις επιδόσεις του δέντρου απόφασης.

Τα περισσότερα από τα πολυμεταβλητά κριτήρια βασίζονται στο γραμμικό συνδυασμό των χαρακτηριστικών εισόδου. Η εύρεση του βέλτιστου δυνατού συνδυασμού επιτυγχάνεται με χρήση άπληστης αναζήτησης (greedy search), γραμμικού προγραμματισμού (linear programming), γραμμική διαχωριστική ανάλυση (linear discriminant analysis) καθώς και από άλλες μεθόδους (Rokach and Maimon, 2005).

2.1.2 Υπερπροσαρμογή του μοντέλου

Τα σφάλματα τα οποία εμφανίζονται σε ένα μοντέλο ταξινόμησης κατηγοριοποιούνται ως εξής:

- Σφάλματα εκπαίδευσης (training, resubstitution, apparent): Περιλαμβάνει τις λανθασμένες ταξινομήσεις των δεδομένων κατά τη διαδικασία της εκπαίδευσης
- Σφάλματα γενίκευσης (generalization): το αναμενόμενο σφάλμα του μοντέλου σε άγνωστα σε αυτό δεδομένα

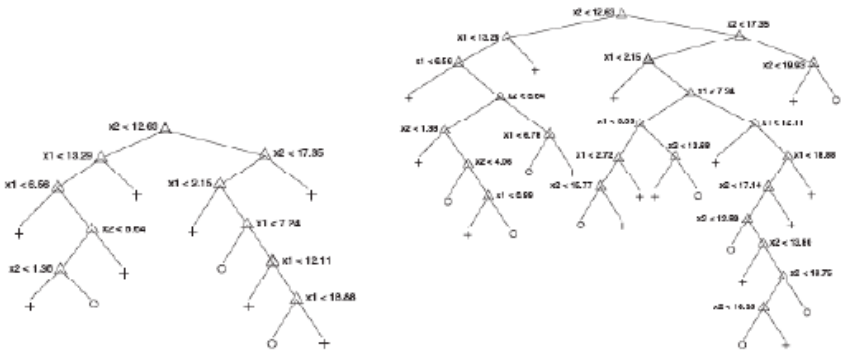
Ένα μοντέλο ταξινόμησης είναι αξιόπιστο όταν τα αποτελέσματα που παράγει είναι ορθά σε εγγραφές οι οποίες δεν εμφανίζονται στα δεδομένα εκπαίδευσης. Με άλλα λόγια έχει χαμηλό σφάλμα τόσο εκπαίδευσης όσο και γενίκευσης. Διευκρινίζεται, μάλιστα, πως ένα μοντέλο το οποίο εμφανίζει μεγάλη προσαρμογή στα δεδομένα εκπαίδευσης (συνεπώς μικρό σφάλμα εκπαίδευσης) ενδέχεται να έχει μεγαλύτερο λάθος γενίκευσης συγκριτικά με κάποιο που ταιριάζει λιγότερο στα τελευταία. Το παραπάνω αναφέρεται στη βιβλιογραφία ως υπερπροσαρμογή μοντέλου (model overfitting) (Tan et al., 2005).

Σε μικρά σε μέγεθος δέντρα ενδέχεται να εμφανιστεί η υποπροσαρμογή μοντέλου (model underfitting). Στις περιπτώσεις αυτές, το μοντέλο δεν έχει μάθει την πραγματική δομή των δεδομένων και συνεπώς εμφανίζει υψηλά σφάλματα τόσο γενίκευσης όσο και εκπαίδευσης. Για το λόγο αυτό, είναι απαραίτητη η αύξηση του πλήθους των κόμβων που περιέχονται στο δέντρο απόφασης (Tan et al., 2005).

Συνεπώς, βάσει των παραπάνω προκύπτει πως:

- Μεγάλο σε μέγεθος δέντρο: Ενδέχεται να εμφανίζει υπερπροσαρμογή του μοντέλου και συνεπώς μεγάλο σφάλμα γενίκευσης
- Μικρό σε μέγεθος δέντρο: Ενδέχεται να εμφανίζει υποπροσαρμογή του μοντέλου και επομένως μεγάλο σφάλμα εκπαίδευσης και γενίκευσης(Tan et al., 2005)

Στην ακόλουθη Εικόνα (Εικόνα 2.13) εμφανίζονται δύο μοντέλα διαφορετικής πολυπλοκότητας.



(a) Decision tree with 11 leaf nodes.

(b) Decision tree with 24 leaf nodes.

ΕΙΚΟΝΑ 2.13: ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ ΤΑ ΟΠΟΙΑ ΕΜΦΑΝΙΖΟΥΝ ΔΙΑΦΟΡΕΤΙΚΗ ΠΟΛΥΠΛΟΚΟΤΗΤΑ (TAN ET AL., 2005)

Η υπερπροσαρμογή ενός μοντέλου μπορεί να οφείλεται σε πλήθος διαφορετικών παραγόντων. Οι συγκεκριμένοι αναλύονται στις ακόλουθες ενότητες

Υπερπροσαρμογή λόγω θορύβου

Έστω ότι ένα σύνολο δεδομένων αποτελείται από ορισμένες ψευδείς εγγραφές (θόρυβος). Οι εγγραφές αυτές επηρεάζουν τη συνθήκη ελέγχου(Tan et al., 2005).

Υπερπροσαρμογή εξαιτίας μη επαρκών δειγμάτων

Τα μοντέλα που βασίζονται σε μικρό αριθμό δειγμάτων εκπαίδευσης είναι επίσης επιρρεπή σε υπερπροσαρμογή. Το πρόβλημα προκύπτει από το γεγονός πως η απουσία επαρκούς αριθμού αντιπροσωπευτικών δειγμάτων, κάνει δύσκολη τη δημιουργία ενός αποτελεσματικού μοντέλου (Tan et al., 2005).

Υπερπροσαρμογή λόγω πολλαπλών επιλογών

Η υπερπροσαμογή του μοντέλου μπορεί να προκύψει πολλές φορές σε αλγορίθμους ταξινόμησης, οι οποίοι εφαρμόζουν μία τεχνική που είναι γνωστή ως διαδικασία πολλαπλών επιλογών. Αναλυτικά, πολλοί αλγόριθμοι διερευνούν ένα σύνολο εναλλακτικών $\{\gamma_i\}$ και από αυτές επιλέγουν τη γ_{max} που μεγιστοποιεί μία δοθείσα συνάρτηση κριτηρίου (criterion function). Ο αλγόριθμός προσθέτει γ_{max} στο παρόν μοντέλο, προκειμένου να βελτιώσει τη γενική απόδοσή του και η διαδικασία επαναλαμβάνεται έως ότου δεν παρατηρείται περαιτέρω αύξηση της ακρίβειάς του. Για παράδειγμα, κατά τη διαδικασία σχηματισμού του μοντέλου, εφαρμόζονται πολλαπλοί έλεγχοι προκειμένου να προσδιοριστεί το χαρακτηριστικό εκείνο που θα επιτύχει τον καλύτερο δυνατό διαχωρισμό στα δεδομένα εκπαίδευσης. Βάσει αυτού το δέντρο επεκτείνεται όσο η παρατηρούμενη βελτίωση είναι στατιστικά σημαντική (Tan et al., 2005).

Έστω ότι T_0 είναι το αρχικό δέντρο και T_χ είναι το νέο δέντρο έπειτα από τη δημιουργία του νέο εσωτερικού κόμβου που αφορά στο χαρακτηριστικό χ. Κατ' αρχήν το χ προστίθεται στο δέντρο εφόσον το παρατηρούμενο κέρδος $\Delta(T_0, T_\chi)$ είναι μεγαλύτερο από ένα προκαθορισμένο κατώφλι α. Στην περίπτωση που αξιολογείται μόνο μία συνθήκη ελέγχου η εισαγωγή πλαστών κόμβων αποφεύγεται εφόσον η τιμή του α επιλέγεται ώστε να είναι αρκετά μεγάλη. Πρακτικά, ωστόσο, οι συνθήκες ελέγχου είναι περισσότερες από μία και ο αλγόριθμος καλείται να επιλέξει το καλύτερο δυνατό χαρακτηριστικό $x_{\max}$από ένα σύνολο $\{x_1, x_2, \dots, x_k\}$ προκειμένου να γίνει ο διαχωρισμός των δεδομένων. Στην περίπτωση αυτή, ο εφαρμόζεται στην πραγματικότητα μία διαδικασία πολλαπλών συγκρίσεων προκειμένου να αποφασιστεί κατά πόσον το δέντρο πρέπει να επεκταθεί. Πιο συγκεκριμένα υλοποιείται ο έλεγχος $\Delta(T_0, T_{\chi_{max}}) > a$ αντί του $\Delta(T_0, T_\chi) > a$. Όσο ο αριθμός των εναλλακτικών k αυξάνεται, τόσο συμβαίνει το αντίστοιχο για την πιθανότητα να βρεθεί $\Delta(T_0, T_\chi) > a$. Ο αλγόριθμός ενδέχεται να προσθέτει ακούσια περιττούς κόμβους στο δέντρο (υπερπροσαρμογή), εκτός αν η συνάρτηση κέρδους ή το κατώφλι της συνάρτησης τροποποιείται ώστε να το k να ληφθεί υπόψη(Tan et al., 2005).

Η προαναφερθείσα επίδραση είναι περισσότερο σαφής όταν ο αριθμός των εγγραφών εκπαίδευσης, εκ των οποίων επιλέγεται το $x_{\max}$, είναι μικρό. Αυτό συμβαίνει διότι η διακύμανση $\Delta(T_0, T_{\chi_{max}})$ είναι μεγαλύτερη σε περιπτώσεις όπου το σύνολο εκπαίδευσης είναι μικρό και αυτό οδηγεί σε αύξηση της πιθανότητας εύρεσης του $\Delta(T_0, T_{\chi_{max}}) > a$. Αυτό συμβαίνει συχνά σε περιπτώσεις, όπου το δέντρο απόφασης αυξάνεται σε μέγεθος, το οποίο με τη σειρά του έχει σαν αποτέλεσμα τη μείωση του αριθμού των εγγραφών που ευρίσκονται στους κόμβους. Το παραπάνω αυξάνει την πιθανότητα να προστεθούν στο δέντρο απόφασης περιττοί κόμβοι. Συνεπώς, η αδυναμία προσαρμογής του μοντέλου σε απαιτήσεις πολλών εναλλακτικών ή μικρού δείγματος εκπαίδευσης οδηγεί σε υπερπροσαρμογή(Tan et al., 2005).

2.1.3 Διαχείριση υπερπροσαρμογής

Στη βιβλιογραφία αναφέρονται δύο τρόποι διαχείρισης της υπερπροσαρμογής:

Προ- κλάδεμα (Pre pruning, Early Stopping Rule)

Μέσω της προσέγγισης αυτής ο αλγόριθμος ανάπτυξης του δέντρου σταματάει τη διαδικασία προτού το μοντέλο προσαρμοστεί πλήρως στα δεδομένα εκπαίδευσης. Αυτό πρακτικά υλοποιείται μέσω της επιβολής μίας αυστηρότερης συνθήκης τερματισμού της διαδικασίας ανάπτυξης. Για παράδειγμα, ένα φύλλο κόμβος δεν επεκτείνεται στην περίπτωση που το κέρδος αναφορικά με την καθαρότητα είναι μικρότερο από ένα κατώφλι. Το πλεονέκτημα του προ- κλαδέματος είναι το γεγονός πως δε δημιουργούνται υπερβολικά περίπλοκα υποδέντρα τα οποία υπερπροσαρμόζονται στα δεδομένα εκπαίδευσης. Δυσκολίες, ωστόσο εντοπίζονται σε ό,τι αφορά την επιλογή του κατάλληλου κατωφλίου, καθώς ένα αυστηρό όριο οδηγεί σε υποπροσαρμοσμένα μοντέλα, ενώ στην αντίθετη περίπτωση σε υπέπροσαρμοσμένα. Επιπροσθέτως, σε πολλές περιπτώσεις παρατηρείται πως ο διαχωρισμός ενός κόμβου σε επιμέρους αυξάνει την αποτελεσματικότητα του μοντέλου ταξινόμησης ακόμα και αν δεν υπάρχει σημαντικό κέρδος στην καθαρότητα των κόμβων παιδιών (Tan et al., 2005).

Μετά- κλάδεμα (Post- pruning)

Μέσω της προσέγγισης αυτής το δέντρο απόφασης κατασκευάζεται αρχικά έως το μεγαλύτερο δυνατό μέγεθος. Στη συνέχεια εφαρμόζεται σε αυτό η διαδικασία του

κλαδέματος, η οποία επεξεργάζεται το μοντέλο ταξινόμησης από κάτω προς τα πάνω (bottom up). Κατά τους (Tan et al., 2005) το κλάδεμα έχει τις εξής εναλλακτικές:

- Το υποδέντρο αντικαθίσταται από ένα κόμβο φύλλο. Η ετικέτα που δίνεται στον τελευταίο ορίζεται από την κλάση που έχει η πλειοψηφία των εγγραφών που βρίσκονται σε αυτόν.
- Το υποδέντρο αντικαθίσταται από το πιο συχνό κλαδί του υποδέντρου.

Η διαδικασία του κλαδέματος τερματίζεται εφόσον δεν παρατηρείται περαιτέρω βελτίωση στις επιδόσεις του δέντρου. Η συγκεκριμένη μέθοδος δίνει καλύτερα αποτελέσματα συγκριτικά με την προαναφερθείσα καθώς οι αποφάσεις λαμβάνονται σε ένα πλήρως αναπτυγμένο δέντρο ενώ το προ- κλάδεμα μπορεί να οδηγήσει σε πολλές περιπτώσεις σε πρόωρο τερματισμό του μοντέλου ταξινόμησης. Ωστόσο, το μετακλάδεμα εμφανίζει το μειονέκτημα πως στη διαδικασία της κατασκευής του δέντρου γίνονται υπολογισμοί, οι οποίοι είναι περιττοί σε περίπτωση που κόμβοι του δέντρου αφαιρούνται (Tan et al., 2005).

Στις ακόλουθες Ενότητες περιγράφονται τα διαφορετικά είδη κλαδέματος που καταγράφονται στην υπάρχουσα βιβλιογραφία.

ΚΛΑΔΕΜΑ ΚΟΣΤΟΥΣ ΠΟΛΥΠΛΟΚΟΤΗΤΑΣ (COST- COMPLEXITY PRUNING)

Το κλάδεμα κόστους πολυπλοκότητας (γνωστό και ως κλάδεμα της πιο αδύναμης σύνδεσης ή κλάδεμα σφάλματος – πολυπλοκότητας) περιλαμβάνει δύο στάδια. Αρχικά, κατασκευάζεται βάσει των δεδομένων εκπαίδευσης μία αλληλουχία δέντρων $T_0, T_1, \dots, T_k$ όπου T_0 είναι το αρχικό δέντρο πριν το κλάδεμα και το T_kτο δέντρο, το οποίο περιλαμβάνει μόνο τον κόμβο ρίζα.

Στο δεύτερο στάδιο γίνεται επιλογή του τελικού μοντέλου ταξινόμησης από το παραπάνω σύνολο. Το κριτήριο στο οποίο βασίζεται η παραπάνω διαδικασία είναι η εκτίμηση του σφάλματος γενίκευσης.

Το T_{i+1} δέντρο διατηρείται αντικαθιστώντας ένα ή περισσότερα υπό- δέντρα του προκατόχου του T_i με τα κατάλληλα φύλλα. Τα υπό δέντρα τα οποία παραμένουν είναι εκείνα τα οποία διατηρούν τη μικρότερη αύξηση στο φαινομενικό ποσοστό σφάλματος (apparent error rate) ανά κάθε φύλλο που κλαδεύτηκε:

$$a = \frac{\varepsilon\, pruned(T,t), S - \varepsilon(T,S)}{|leaves(T)| - |leaves(pruned(T,t))|}$$

Όπου:

- $\varepsilon(T,S)$ συμβολίζει το ποσοστό σφάλματος δέντρου Τ για το δείγμα S
- $|leaves(T)|$ συμβολίζει τον αριθμό των φύλλων στο Τ
- $pruned(T,t)$ συμβολίζει το δέντρο που προκύπτει έπειτα από την αντικατάσταση του κόμβου t στο Τ με κάποια κατάλληλο φύλλο

Στη δεύτερη φάση υπολογίζεται το σφάλμα γενίκευσης κάθε κλαδεμένου δέντρου $T_0, T_1, \dots, T_k$ και επιλέγεται το βέλτιστο από αυτά. Στην περίπτωση που το δοθέν σύνολο δεδομένων είναι μεγάλο, οι συγγραφείς προτείνουν το διαχωρισμό του σε σύνολο εκπαίδευσης και σύνολο κλαδέματος. Τα δέντρα κατασκευάζονται αξιοποιώντας το σύνολο εκπαίδευσης και αξιολογούνται μέσω του συνόλου κλαδέματος. Στην αντίθετη περίπτωση του μικρού συνόλου δεδομένων προτείνεται η χρήση της μεθοδολογίας Cross Validation (περιγράφεται διεξοδικά στις ακόλουθες ενότητες) (Rokach and Maimon, 2005).

ΚΛΑΔΕΜΑ ΜΕΙΩΜΕΝΟΥ ΣΦΑΛΜΑΤΟΣ (REDUCED ERROR PRUNING)

Η παρούσα μεθοδολογία προτάθηκε από τον Quinlan (1987) και αποτελεί μία απλή διαδικασία κλαδέματος δέντρων απόφασης. Η συγκεκριμένη ελέγχει κατά πόσον η αντικατάσταση κάθε εσωτερικού κόμβου με την πιο συχνή κλάση επηρεάζει την ακρίβεια του δέντρου. Στην περίπτωση που το συγκεκριμένο μέτρο δεν ελαττώνεται ο κόμβος αφαιρείται από το μοντέλο. Το παραπάνω υλοποιείται από τη βάση έως την κορυφή του δέντρου και η διαδικασία επαναλαμβάνεται έως ότου η περαιτέρω αφαίρεση κόμβων έχει σαν αποτέλεσμα τη μείωση της ακρίβειας του δέντρου.

Ο Quinlan προτείνει τη χρήση ενός συνολικού κλαδέματος προκειμένου να καταστεί δυνατή η εκτίμηση της ακρίβειας του μοντέλου. Μάλιστα μπορεί να αποδειχθεί πως η διαδικασία αυτή τερματίζεται με το μικρότερο δυνατό σε μέγεθος υποδέντρο σε σχέση με ένα δοθέν σύνολο κλαδέματος (Rokach and Maimon, 2005).

ΚΛΑΔΕΜΑ ΕΛΑΧΙΣΤΟΥ ΣΦΑΛΜΑΤΟΣ (MINIMUM ERROR PRUNING)

Η παρούσα διαδικασία προτάθηκε από τους (Olaru και Wehenkel, 2003). Η συγκεκριμένη διασχίζει τους εσωτερικούς κόμβους από τη βάση έως την κορυφή του δέντρου. Η μέθοδος αυτή υπολογίζει για κάθε κόμβο το ποσοστό σφάλματος της l-πιθανότητας με και χωρίς το κλάδεμα.

Το ποσοστό σφάλματος της l-πιθανότητας αποτελεί μία διόρθωση της απλής εκτίμησης πιθανότητας με χρήση συχνοτήτων. Αν η S_t συμβολίζει τις οντότητες που έχουν «φτάσει» ένα φύλλο t, τότε το ποσοστό σφάλματος της l-πιθανότητας υπολογίζεται ως εξής:

$$\varepsilon'(t) = 1 - \max_{c_i \in dom(y)} \frac{\left|\sigma_{y=c_i} S_t\right| + l\, p_{apr}(y = c_i)}{S_t + l}$$

Όπου: $p_{apr}(y = c_i)$ είναι η a- priori πιθανότητα το y να πάρει την τιμή c_i ενώ το l συμβολίζει το βάρος που δίνεται στην παραπάνω πιθανότητα.

Το ποσοστό σφάλματος κάθε εσωτερικού κόμβου είναι ο σταθμισμένος μέσος όρος των ποσοστών σφάλματος των κλαδιών του. Το βάρος υπολογίζεται ανάλογα με το πλήθος των οντοτήτων (instances) που ευρίσκονται σε κάθε κλαδί και ο υπολογισμός εκτελείται αναδρομικά έως τα φύλλα.

Στην περίπτωση που αφαιρεθεί ένας εσωτερικός κόμβος, μετατρέπεται σε φύλλο και το ποσοστό σφάλματος υπολογίζεται αυτόματα μέσω της τελευταίας εξίσωσης. Συνεπώς, γίνεται υπολογισμός του εν λόγω μεγέθους πριν και μετά το κλάδεμα ενός συγκεκριμένου εσωτερικού κόμβου. Στην περίπτωση που το κλάδεμα δεν αυξάνει το ποσοστό σφάλματος, η ενέργεια αυτή γίνεται αποδεκτή και το μοντέλο αλλάζει (Rokach and Maimon, 2005).

ΔΥΣΟΙΩΝΟ ΚΛΑΔΕΜΑ (PESSIMISTIC PRUNING)

Το παρόν κλάδεμα αποφεύγει τη χρήση συνόλου εκπαίδευσης καθώς και της μεθόδου cross validation και αντ' αυτού χρησιμοποιεί έλεγχο πεσιμιστικής στατιστικής συσχέτισης (pessimistic statistical correlation test). Η βασική ιδέα της συγκεκριμένης διαδικασίας είναι η ακόλουθη: το ποσοστό σφάλματος το οποίο υπολογίστηκε χρησιμοποιώντας το σύνολο εκπαίδευσης δεν είναι αρκετά αξιόπιστο. Αντιθέτως, γίνεται χρήση ενός περισσότερου ρεαλιστικού μέτρου, το οποίο είναι γνωστό ως διόρθωση συνέχειας για διωνυμική κατανομή (continuity correction for binomial distribution):

$$\varepsilon'(T, S) = \varepsilon(T, S) + \frac{|leaves(T)|}{2\,|S|}$$

Η παρούσα διόρθωση παράγει ωστόσο μία αισιόδοξη αναλογία σφάλματος (an optimistic error rate). Συνεπώς, θεωρείται πως το κλάδεμα σε έναν εσωτερικό κόμβο t εφόσον το ποσοστό σφάλματος είναι στο εσωτερικό ενός ορισμένου σφάλματος

$$\varepsilon'(pruned(T,t),S) \leq \varepsilon'(T,S) + \sqrt{\frac{\varepsilon'(T,S)\big(1-\varepsilon'(T,S)\big)}{|S|}}$$

Συνήθως, η συνθήκη αυτή το T αναφέρεται το υποδέντρο του οποίο ρίζα είναι ο εσωτερικός κόμβος t. Με S συμβολίζεται η αναλογία των δεδομένων εκπαίδευσης του κόμβου t.

Το απαισιόδοξο κλάδεμα διασχίζει τους εσωτερικούς κόμβους από πάνω προς τα κάτω. Στην περίπτωση που κάποιος από τους κόμβους κλαδευτεί αφαιρούνται από τη διαδικασία του κλαδέματος όλοι οι απόγονοι του τελευταίου. Συνεπώς το δυσοίωνο κλάδεμα είναι μία πολύ γρήγορη διαδικασία.

ΚΛΑΔΕΜΑ ΒΑΣΙΖΟΜΕΝΟ ΣΤΟ ΣΦΑΛΜΑ (ERROR BASED PRUNING - EBP)

Το συγκεκριμένο κλάδεμα αποτελεί εξέλιξη του προαναφερθέντος. Έχει εφαρμοστεί στο γνωστό αλγόριθμο C4.5.

Κατά αντιστοιχία με το δυσοίωνο κλάδεμα το ποσοστό σφάλματος υπολογίζεται χρησιμοποιώντας το πάνω όριο του διαστήματος της στατιστικής εμπιστοσύνης των αναλογιών.

$$\varepsilon_{UB} = \varepsilon(T,S) + Z_a \sqrt{\frac{\varepsilon(T,S)\ (1-\varepsilon(T,S))}{|S|}}$$

Όπου το $\varepsilon(T,S)$ συμβολίζει το ποσοστό εσφαλμένης ταξινόμησης του δέντρου T στο διάστημα εκπαίδευσης S. Το Z είναι ο αντίστροφος της κανονικής τυπικής αθροιστικής κατανομής (standard normal cumulative distribution) και α είναι επιθυμητό επίπεδο σημαντικότητας (desired significance level).

Έστω ότι το $subtree(T,t)$ συμβολίζει το ριζωμένο υποδέντρο που ξεκινάει από τον κόμβο t και το $maxchild(T,t)$ τον πιο συχνό κόμβο – παιδί (child node) του t (δηλαδή τον κόμβο εκείνο στον οποίο καταλήγουν οι περισσότερες οντότητες του S_t όλες οι οντότητες του S που φτάνουν στον κόμβο t).

Η συγκεκριμένη διαδικασία διασχίζει τους κόμβους του δέντρου με κατεύθυνση από τη ρίζα προς τα φύλλα και συγκρίνει για κάθε έναν από αυτούς τις ακόλουθες τιμές:

- $\varepsilon_{UB}(subtree(T,t), S_t)$
- $\varepsilon_{UB}(pruned(subtree(T,t),t), S_t)$
- $\varepsilon_{UB}\big(subtree\big(T, maxchild(T,t)\big)\big), S_{maxchild(T,t)}\big)$

Στη συνέχεια, βάσει των παραπάνω τιμών το δέντρο παραμένει ως έχει ή γίνεται κλάδεμα του κόμβου t ή γίνεται αντικατάσταση του κόμβου t με το ριζωμένο υποδέντρο από το $maxchild(T,t)$.

ΒΕΛΤΙΣΤΟ ΚΛΑΔΕΜΑ

Το θέμα του βέλτιστου κλαδέματος μελετήθηκε ενδελεχώς από τους (Bratko and Brohanec, 1994) και (Almuallim, 1996). Στην πρώτη προσπάθεια προτάθηκε ένας αλγόριθμος, ο OPT, ο οποίος αξιοποιεί τεχνικές του δυναμικού προγραμματισμού και χρησιμοποιεί την πολυπλοκότητα των $\Theta(|leveas(T)|^2)$, όπου T είναι το αρχικό δέντρο απόφασης. Η δεύτερη έρευνα αφορά σε μία βελτίωση του OPT, τον OPT-2 και ο συγκεκριμένος ομοίως με τον OPT αξιοποιεί τεχνικές του δυναμικού προγραμματισμού. Ωστόσο, η πολυπλοκότητα του χρόνου και του χώρου για τον OPT-2 υπολογίζεται βάσει της σχέσης$\Theta(|leveas(T^*)|)|internal\ (T)|$, όπου T^* είναι το κλαδεμένο δέντρο- στόχος και Tείναι το αρχικό δέντρο απόφασης (Rokach and Maimon, 2005).

ΚΛΑΔΕΜΑ ΕΛΑΧΙΣΤΟΥ ΜΗΚΟΥΣ ΠΕΡΙΓΡΑΦΗΣ (MINIMUM DESCRIPTION LENGTH – MDL)

Η παρούσα διαδικασία μπορεί να χρησιμοποιηθεί για την αξιολόγηση της γενικευμένης ακρίβειας ενός κόμβου (Rissanen, 1989; Quinlan and Rivest, 1989; Mehtaet al., 1995). Η μεθοδολογία αυτή υπολογίζει την ακρίβεια του δέντρου βάσει του αριθμού των bits που απαιτούνται για την κωδικοποίηση του δέντρου και προτιμά τα μοντέλα εκείνα με το μικρότερο μέγεθος αποθήκευσης. Το κόστος της διάσπασης στο φύλλο t υπολογίζεται βάσει της ακόλουθης σχέσης:

$$Cost(t) = \sum_{c_i \in dom(y)} \left|\sigma_{y=c_i} S_t\right| \ln \frac{|S_t|}{\left|\sigma_{y=c_i} S_t\right|} + \frac{|dom(y)-1|}{2} \ln \frac{|S_t|}{2} + ln \frac{\pi^{\frac{|dom(y)|}{2}}}{\Gamma(\frac{|dom(y)|}{2})}$$

Όπου το S_t συμβολίζει τις οντότητες που έχουν φθάσει τον κόμβο t. Το κόστος διαχωρισμού (splitting cost) ενός εσωτερικού κόμβου υπολογίζεται αθροίζοντας το επιμέρους κόστος (του κόστους συνάθροισης (cost aggregation)) των παιδιών του (Rokach and Maimon, 2005).

ΣΥΓΚΡΙΣΗ ΤΩΝ ΜΕΘΟΔΩΝ ΚΛΑΔΕΜΑΤΟΣ

Στη βιβλιογραφία καταγράφεται πλήθος ερευνών οι οποίες συγκρίνουν τις επιδόσεις διαφόρων τεχνικών κλαδέματος (Quinlan, 1987; Mingers, 1989; Esposito et al., 1997). Τα αποτελέσματα των παραπάνω δείχνουν πως ορισμένες από τις μεθόδους (όπως η κόστους πολυπλοκότητας και η μειωμένου σφάλματος) τείνουν να κλαδεύουν τα δέντρα απόφασης περισσότερο από το επιθυμητό δημιουργώντας με τον τρόπο αυτό μικρότερα αλλά και λιγότερο ακριβή μοντέλα. Οι λοιπές τεχνικές (όπως το κλάδεμα βασιζόμενο στο σφάλμα, το απαισιόδοξο κλάδεμα και το κλάδεμα ελάχιστου σφάλματος) ρέπουν προς τη δημιουργία μεγάλων σε μέγεθος δέντρων και συνεπώς υπερπροσαρμοσμένων.

Τέλος, οι μελέτες έδειξαν πως το θεώρημα «no free lunch» ισχύει και σε αυτήν την περίπτωση και επομένως δεν υπάρχει κάποια μεθοδολογία η οποία να επιτυγχάνει σε όλες τις περιπτώσεις εμφανώς υψηλότερες επιδόσεις συγκριτικά με τις υπόλοιπες (Rokach and Maimon, 2005).

2.1.4 Υπολογισμός του σφάλματος γενίκευσης

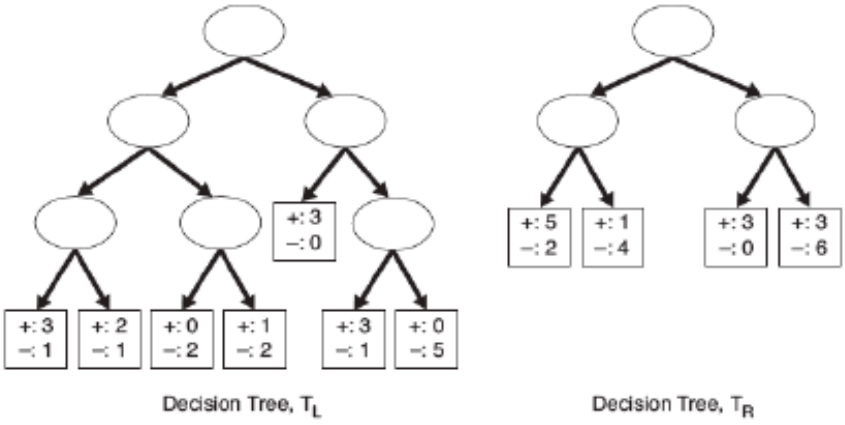
Σε προηγούμενη ενότητα αναφέρθηκε πως η πολυπλοκότητα ενός δέντρου απόφασης επηρεάζει άμεσα την ακρίβειά του. Συγκεκριμένα, αναφέρθηκε πως ένα μικρό σε μέγεθος και απλό δέντρο απόφασης εμφανίζει υποπροσαρμογή, ενώ αντίθετα ένα ιδιαίτερα περίπλοκο υπερπροσαρμογή στα δεδομένα εκπαίδευσης. Το ερώτημα που εύλογα γεννάται είναι πώς μπορεί να προσδιοριστεί η ιδανική πολυπλοκότητα ενός μοντέλου. Ένας αλγόριθμος ταξινόμησης έχει πρόσβαση αποκλειστικά στα δεδομένα εκπαίδευσης και όχι σε εκείνα του ελέγχου. Συνεπώς, δεν υπάρχει γνώση σχετικά με την απόδοση ενός δέντρου απόφασης σε άγνωστες σε αυτό εγγραφές. Μία λύση για τη διαχείριση του συγκεκριμένου προβλήματος είναι ο υπολογισμός του σφάλματος γενίκευσης του δέντρου απόφασης. Στα ακόλουθα χωρία αναλύονται διεξοδικά οι προτεινόμενες μέθοδοι για το συγκεκριμένο εγχείρημα (Tan et al., 2005).

Εκτίμηση εκπαίδευσης- επαναντικατάστασης (resubstitution)

Η συγκεκριμένη βασίζεται στην υπόθεση πως το σύνολο δεδομένων εκπαίδευσης είναι αντιπροσωπευτικό του συνόλου των δεδομένων. Συνεπώς το σφάλμα εκπαίδευσης το οποίο είναι γνωστό και ως σφάλμα επαναντικατάστασης μπορεί να αποτελέσει μία αισιόδοξη εκτίμηση του σφάλματος γενίκευσης. Με την παραδοχή αυτή ο αλγόριθμος δημιουργίας δέντρων απόφασης επιλέγει το μοντέλο εκείνο με την ελάχιστη αναλογία

σφάλματος εκπαίδευσης. Σε πολλές περιπτώσεις ωστόσο το συγκεκριμένο αποτελεί μία κακή εκτίμηση εκείνου της γενίκευσης.

Στην Εικόνα 2.14 παρουσιάζονται δύο δέντρα απόφασης τα οποία έχουν κατασκευαστεί από το ίδιο σύνολο εκπαίδευσης. Το αριστερό δέντρο έχει αναπτυχθεί περισσότερο συγκριτικά με το δεξί. Το σύνολο εκπαίδευσης αποτελείται από συνολικά 24 εγγραφές και το σφάλμα εκπαίδευσης υπολογίζεται στα δέντρα ως εξής. Σε κάθε φύλλο γίνεται αναζήτηση της κλάσης (+ ή -) με τις περισσότερες εγγραφές. Οι εγγραφές της αντίθετης κλάσης υπολογίζονται ως σφάλμα. Ο συνολικός αριθμός των σφαλμάτων σε όλα τα φύλλα διαιρείται με το συνολικό αριθμό των εγγραφών του συνόλου εκπαίδευσης και η τιμή που προκύπτει είναι το σφάλμα εκπαίδευσης. Συνεπώς το σφάλμα του αριστερού δέντρου είναι ίσο με $e(T_L) = \frac{4}{24} = 0{,}167$ ενώ στο δεξί $e(T_R) = \frac{6}{24} = 0{,}25$. Άρα, επιλέγεται τελικά το αριστερό δέντρο(Tan et al., 2005).



ΕΙΚΟΝΑ 2.14: ΠΑΡΑΔΕΙΓΜΑΤΑ ΔΥΟ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ (TAN ET AL., 2005)

Ενσωμάτωση της πολυπλοκότητας του μοντέλου

Όπως έχει ήδη προαναφερθεί η πιθανότητα υπερπροσαρμογής του μοντέλου μεγαλώνει όσο αυξάνεται η πολυπλοκότητα του δέντρου. Για τον παραπάνω λόγο προτιμώνται τα απλοϊκότερα μοντέλα και στρατηγική αυτή είναι σύμφωνη με την ακόλουθη αρχή:

Occam's Razor: Δοθέντων δύο μοντέλων με παρόμοια λάθη γενίκευσης, πρέπει να προτιμάται το απλούστερο από το πιο περίπλοκο.

Η παραπάνω αρχή βασίζεται στη λογική πως οι πρόσθετοι κόμβοι σε ένα περίπλοκο δέντρο απόφασης είναι πολύ πιθανόν να έχουν τοποθετηθεί σε αυτό καθαρά από θέμα τύχης. «Οτιδήποτε πρέπει να φτιαχτεί όσο το δυνατόν απλό, αλλά όχι απλούστερο», όπως έχει διατυπωθεί και από τον Einstein. Στη συνέχεια, αναλύονται μέθοδοι ενσωμάτωσης της πολυπλοκότητας του μοντέλου (Tan et al., 2005).

Απαισιόδοξη προσέγγιση εκτίμησης

Η απαισιόδοξη προσέγγιση υπολογίζει το σφάλμα γενίκευσης ως άθροισμα του σφάλματος γενίκευσης, καθώς και ενός όρου ποινής σχετικά με την πολυπλοκότητα του μοντέλου. Το σφάλμα γενίκευσης το οποίο προκύπτει μπορεί να θεωρηθεί ως μία απαισιόδοξη εκτίμηση του σφάλματος του μοντέλου. Η απαισιόδοξη εκτίμηση του σφάλματος $e_g(T)$ ενός δέντρου T υπολογίζεται βάσει της ακόλουθης σχέσης:

$$e_g(T) = \frac{\sum_{i=1}^{k}[e(t_i) + \Omega(t_i)]}{\sum_{i=1}^{k} n(t_i)} = \frac{e(T) + \Omega(T)}{N_t}$$

Όπου:

- k το πλήθος των κόμβων φύλλων
- $n(t)$ είναι ο αριθμός των εγγραφών του συνόλου εκπαίδευσης τα οποία έχουν ταξινομηθεί στον κόμβο t
- $e(t)$ είναι ο αριθμός των λάθος ταξινομημένων εγγραφών
- $e(T)$ είναι το συνολικό σφάλμα εκπαίδευσης
- $\Omega(t_i)$ είναι η ποινή η οποία συνδέεται με τον κόμβο t

N_t είναι ο συνολικός αριθμός των δεδομένων εκπαίδευσης (Tan et al., 2005)

Ο τρόπος υπολογισμού της απαισιόδοξης εκτίμησης του σφάλματος για τα δέντρα απόφασης της Εικόνα 2.14 είναι ο ακόλουθος. Έστω ότι ο όρος ποινή είναι ίσος με 0,5 τότε:

$$e(T_L) = \frac{4 + 7\ 0{,}5}{24} = \frac{7{,}5}{24} = 0{,}3125$$

$$e(T_R) = \frac{6 + 4\ 0{,}5}{24} = \frac{8}{24} = 0{,}3333$$

Συνεπώς, το αριστερό δέντρο έχει μικρότερο σφάλμα συγκριτικά με το δεξί(Tan et al., 2005).

2.1.5 Αξιολόγηση των επιδόσεων ενός ταξινόμητη

Σε προηγούμενη ενότητα περιεγράφηκαν αναλυτικά ορισμένες μέθοδοι υπολογισμού του σφάλματος γενίκευσης. Μέσω του συγκεκριμένου είναι δυνατή η εύρεση ενός μοντέλου το οποίο έχει αφενός τη σωστή πολυπλοκότητα και αφετέρου δεν είναι επιρρεπές σε υπερπροσαρμογή. Εφόσον ολοκληρωθεί η κατασκευή του μοντέλου εφαρμόζεται σε αυτό ένα σύνολο ελέγχων μέσω των οποίων γίνεται ταξινόμηση άγνωστων για το μοντέλο εγγραφών.

Σε πολλές περιπτώσεις κρίνεται απαραίτητη η αξιολόγηση των επιδόσεων ενός μοντέλου σε ένα σύνολο ελέγχου, καθώς μέσω της συγκεκριμένης προκύπτει μία αντικειμενική εκτίμηση του σφάλματος γενίκευσης. Η ακρίβεια ή η αναλογία σφάλματος που προκύπτει βάσει της διαδικασίας αυτής μπορεί επίσης να χρησιμοποιηθεί προκειμένου να γίνει σύγκριση των επιδόσεων διαφορετικών ταξινομητών οι οποίοι ωστόσο ανήκουν στον ίδιο τομέα.

Στις ακόλουθες ενότητες αναλύονται διεξοδικά οι μεθοδολογίες που σχετίζονται με την αξιολόγηση των επιδόσεων ενός ταξινομητή (Tan et al., 2005).

Μέθοδος προπύργιο (hold out)

Μέσω της συγκεκριμένης μεθόδου τα δεδομένα εισόδου ενός ταξινομητή χωρίζονται σε δύο σύνολα, εκείνο της εκπαίδευσης και εκείνο του ελέγχου. Στη συνέχεια, γίνεται κατασκευή του μοντέλου ταξινόμησης μέσω των εγγραφών που καταχωρήθηκαν στο πρώτο σύνολο και οι επιδόσεις του τελευταίου αξιολογούνται μέσω του συνόλου ελέγχου. Η αναλογία των εγγραφών που καταχωρούνται σε κάθε μία από τις παραπάνω κατηγορίες είναι στην ευχέρεια της ανάλυσης (για παράδειγμα 50-50) και η ακρίβεια του μοντέλου ταξινόμησης υπολογίζεται βάσει των επιδόσεων του τελευταίου στο σύνολο ελέγχου.

Η μέθοδος προπύργιο είναι ιδιαίτερα απλή στην εφαρμογή της, εμφανίζει ωστόσο τα ακόλουθα μειονεκτήματα. Αρχικά, οι εγγραφές που προορίζονται για την εκπαίδευση του μοντέλου μειώνονται καθώς ορισμένες από αυτές καταχωρούνται στο σύνολο ελέγχου και το παραπάνω επιδρά αρνητικά στην ποιότητα αυτού. Επιπροσθέτως, το δέντρο απόφασης εξαρτάται σε μεγάλο βαθμό από τη σύνθεση των συνόλων εκπαίδευσης και ελέγχου και σημειώνεται επίσης πως όσο μικρότερο είναι το μέγεθος του συνόλου εκπαίδευσης τόσο αυξάνεται η απόκλιση του μοντέλου. Αυτό έχει σαν αποτέλεσμα η εκτιμώμενη ακρίβεια που προκύπτει από το σύνολο ελέγχου να είναι λιγότερο αξιόπιστη σε περιπτώσεις που το συγκεκριμένο είναι μικρό σε μέγεθος. Τέλος, τα σύνολα εκπαίδευσης και ελέγχου δεν είναι μεταξύ τους ανεξάρτητα. Αυτό οφείλεται στο γεγονός πως τα παραπάνω είναι υποσύνολα του αρχικού και συνεπώς κλάσεις που εμφανίζονται συχνά σε ένα από αυτά ενδέχεται να απουσιάζουν στο άλλο και αντιστρόφως (Tan et al., 2005).

Τυχαία υποδειγματοληψία (Random subsampling)

Η επανάληψη της προαναφερθείσας τεχνικής βελτιώνει τις επιδόσεις του ταξινομητή και η μέθοδος αυτή είναι γνωστή ως τυχαία υποδειγματοληψία. Αν acc_i είναι η ακρίβεια του μοντέλου στην i επανάληψη, τότε η συνολική ακρίβεια υπολογίζεται από τη σχέση $acc_{sub} = \sum_{i=1}^{k} acc_i/k$. Η συγκεκριμένη μεθοδολογία εμφανίζει ορισμένα από τα μειονεκτήματα της μεθόδου προπύργιο καθώς ομοίως με την τελευταία δε χρησιμοποιεί τόσα δεδομένα όσα είναι δυνατόν προκειμένου να εκπαιδευτεί το μοντέλο. Επιπροσθέτως, δεν έχει έλεγχο των φορών που κάθε εγγραφή χρησιμοποιείται για έλεγχο ή εκπαίδευση και συνεπώς, ορισμένες ενδέχεται να εμφανίζονται στις παραπάνω διαδικασίες περισσότερες από μία φορές (Tan et al., 2005).

Διασταυρωμένη επικύρωση (Cross Validation)

Η συγκεκριμένη μέθοδος αποτελεί μία εναλλακτική της τυχαίας δειγματοληψίας με τη διαφορά πως η προσέγγιση αυτή χρησιμοποιεί κάθε εγγραφή ίδιο αριθμό φορών για εκπαίδευση και μία φορά για έλεγχο. Η διαδικασία λειτουργεί ως εξής: Αρχικά, γίνεται διαχωρισμός των δεδομένων των δεδομένων σε δύο ίσα σύνολα, ένα εκπαίδευσης καθώς και ένα ελέγχου. Στη συνέχεια, γίνεται ανταλλαγή των ρόλων των δύο υποσυνόλων. Η προσέγγιση αυτή ονομάζεται διπλή cross validation (two fold cross validation).Το συνολικό σφάλμα της διαδικασίας είναι ίσο με το άθροισμα των σφαλμάτων που προκύπτει για τις παραπάνω δύο επαναλήψεις.

Στο προαναφερθέν παράδειγμα κάθε εγγραφή χρησιμοποιείται ακριβώς μία φορά για εκπαίδευση και για έλεγχο. Στην πολλαπλή cross validation (kfold cross validation) τα δεδομένα διασπώνται σε k ίσων διαστημάτων σύνολα. Κατά τη διάρκεια κάθε γύρου ένα από τα παραπάνω σύνολα χρησιμοποιείται για έλεγχο και τα υπόλοιπα ενοποιούνται σε ένα προκειμένου να είναι δυνατή η εκπαίδευση του μοντέλου. Η διαδικασία επαναλαμβάνεται k φορές ούτως ώστε κάθε σύνολο να χρησιμοποιηθεί για την τελευταία διαδικασία ακριβώς μία φορά. Το συνολικό σφάλμα της πολλαπλής cross validationείναι ίσο με το άθροισμα των σφαλμάτων για τις παραπάνω k επαναλήψεις.

Σε μία ειδική περίπτωση της πολλαπλής cross validation ορίζεται k=N., δηλαδή ίσο με το πλήθος των δεδομένων εισόδου. Η προσέγγιση αυτή ονομάζεται αφήνω- ένα- έξω (leave-one- out) καθώς κάθε σύνολο ελέγχου αποτελείται από ακριβώς μία εγγραφή. Η συγκεκριμένη διατηρεί το πλεονέκτημα πως αξιοποιεί το μέγιστο αριθμό δεδομένων για εκπαίδευση. Επιπροσθέτως, τα σύνολα ελέγχου είναι αμοιβαίως αποκλειόμενα και συνεπώς καλύπτουν αποτελεσματικά ολόκληρο το σύνολο δεδομένων. Το μειονέκτημα,

ωστόσο, της συγκεκριμένης μεθόδου είναι πως η επανάληψη της διαδικασίας N φορές είναι υπολογιστικά ακριβή. Τέλος, σημειώνεται πως η απόκλιση της εκτιμώμενης ακρίβειας είναι υψηλή, καθώς κάθε έλεγχος περιέχει ακριβώς μία εγγραφή (Tan et al., 2005).

Bootstrap

Οι προαναφερθείσες μέθοδοι βασίζονται στην υπόθεση πως οι εγγραφές του συνόλου εκπαίδευσης αποτελούν δείγματα χωρίς αντικατάσταση και συνεπώς δεν υπάρχουν διπλές καταγραφές στα σύνολα. Στην παρούσα μέθοδο τα δείγματα επανατοποθετούνται στην αρχική «δεξαμενή» εγγραφών και συνεπώς είναι εξίσου πιθανό να επιλεχθούν ξανά. Αν το αρχικό σύνολο δεδομένων αποτελείται από N εγγραφές έχει αποδειχθεί ότι κατά μέσο όρο ότι το δείγμα boostrap περιέχει το 63,2% των αρχικών δεδομένων. Η προσέγγιση αυτή προκύπτει από το γεγονός πως η πιθανότητα επιλογής μία εγγραφής στην προκειμένη περίπτωση είναι ίση με $1 - \left(1 - \frac{1}{N}\right)^N$ και όταν το N είναι επαρκώς μεγάλο η πιθανότητα προσεγγίζει συμπτωματικά το $1 - e^{-1}$. Οι εγγραφές οι οποίες δε συμπεριλαμβάνονται στο δείγμα boostrap συμπεριλαμβάνονται στο σύνολο ελέγχου. Το μοντέλο συνεπώς εκπαιδεύεται από το δείγμα boostrap και στη συνέχεια αξιολογείται από το σύνολο ελέγχου, προκειμένου να υπολογιστεί η ακρίβεια του πρώτου, e_i. Η παραπάνω διαδικασία επαναλαμβάνεται b φορές και συνεπώς παράγονται boostrap δείγματα.

Αξίζει να σημειωθεί πως στη βιβλιογραφία αναγράφεται πλήθος παραλλαγών σε ό,τι αφορά τον υπολογισμό της ακρίβειας του μοντέλου που προκύπτει μέσω της συγκεκριμένης μεθοδολογίας. Μία από τις πιο ευρέως διαδεδομένες προσεγγίσεις είναι η .632 bootstrap, η οποία υπολογίζει τη συνολική ακρίβεια συνδυάζοντας τις επιμέρους ακρίβειες κάθε δείγματος bootstrap ως εξής:

$$Accuracy, acc_{boot} = \frac{1}{b}\sum_{i=1}^{b}(0{,}632\ e_i + 0{,}368\ acc_s)$$

Όπου:

- e_i τα δείγματα bootstrap
- acc_sη ακρίβεια που υπολογίστηκε από το σύνολο εκπαίδευσης το οποίο περιέχει όλες τις εγγραφές των αρχικών δεδομένων(Tan et al., 2005)

2.1.6 Μέθοδοι σύγκρισης των μοντέλων ταξινόμησης

Η σύγκριση της απόδοσης μεταξύ διαφορετικών μοντέλων ταξινόμησης είναι χρήσιμη σε πολλές περιπτώσεις. Μέσω αυτής επιλέγεται ο ταξινομητής εκείνος ο οποίος δίνει καλύτερα αποτελέσματα σε ένα δοθέν σύνολο δεδομένων. Ωστόσο, η παρατηρούμενη διαφορά στην ακρίβεια δύο ταξινομητών ενδέχεται να μην είναι στατιστικά σημαντική και αυτό εξαρτάται από το μέγεθος του συνόλου των δεδομένων. Στόχος της παρούσας ενότητας είναι η διερεύνηση ορισμένων στατιστικών ελέγχων, μέσω των οποίων η διευκολύνεται η επιλογή του ακριβέστερου μοντέλου ταξινόμησης.

Στο σημείο αυτό, θα δοθεί ένα παράδειγμα για επεξηγηματικούς λόγους. Έστω πως υπάρχουν διαθέσιμα δύο μοντέλα ταξινόμησης,M_A, M_B. Το πρώτο εκ των δύο έχει ακρίβεια 85% σε ένα σύνολο δεδομένων ελέγχου το οποίο αποτελείται από 30 εγγραφές. Ομοίως το M_B έχει ακρίβεια 75% σε 5000 εγγραφές. Βάσει αυτών των πληροφοριών το M_A έχει καλύτερες επιδόσεις από το M_B;

Το παραπάνω παράδειγμα εγείρει δύο βασικά ερωτήματα, αναφορικά με τη στατιστική σημαντικότητα των μετρήσεων απόδοσης:

1. Το μοντέλο M_A έχει υψηλότερη ακρίβεια συγκριτικά με το M_B. Ωστόσο, το σύνολο ελέγχου του δεύτερου μοντέλου είναι σαφώς μεγαλύτερο από εκείν0ο του πρώτου. Βάσει αυτή της πληροφορίας μπορεί να θεωρηθεί πως οι επιδόσεις του πρώτου μοντέλου είναι καλύτερες συγκριτικά με εκείνες του δεύτερου;
2. Είναι δυνατόν η διαφορά στην ακρίβεια να αποδοθεί σε αποκλίσεις στη σύνθεση του συνόλου ελέγχου.

Το πρώτο ερώτημα σχετίζεται με το θέμα του υπολογισμού του διαστήματος εμπιστοσύνης της δοθείσας ακρίβειας του μοντέλου. Το δεύτερο σχετίζεται με το θέμα ελέγχου της στατιστικής σημαντικότητας μίας παρατηρούμενης απόκλισης. Τα θέματα αυτά διερευνώνται στα ακόλουθα χωρία του κεφαλαίου αυτού.

Υπολογισμός του διαστήματος εμπιστοσύνης για την ακρίβεια

Προκειμένου να επιτευχθεί το συγκεκριμένο εγχείρημα είναι απαραίτητο να προσδιοριστεί η κατανομή πιθανοτήτων που διέπει το μέτρο ακρίβειας. Η παρούσα ενότητα περιγράφει μία προσέγγιση για την εξαγωγή ενός διαστήματος εμπιστοσύνης μοντελοποιώντας τη διαδικασία της ταξινόμησης ως ένα διωνυμικό πείραμα. Ορισμένα από τα χαρακτηριστικά ενός διωνυμικού πειράματος είναι τα εξής:

- Το πείραμα αποτελείται από N ανεξάρτητες δοκιμές, όπου κάθε δοκιμή έχει δύο πιθανά αποτελέσματα: επιτυχία ή αποτυχία
- Η πιθανότητα της επιτυχίας, p, σε κάθε δοκιμή είναι σταθερή

Ένα παράδειγμα ενός διωνυμικού πειράματος είναι η καταμέτρηση του πλήθους των κεφαλών στις N ρήψεις ενός νομίσματος. Αν X 0είναι το πλήθος των «επιτυχιών» στις N ρήψεις, τότε η πιθανότητα το X να πάρει μία συγκεκριμένη τιμή υπολογίζεται μέσω διωνυμικής κατανομής με μέσο Np και απόκλιση $Np\ (1-p)$:

$$P(X = u) = \binom{N}{p} p^u (1-p)^{N-u}$$

Στο παράδειγμα του νομίσματος η πιθανότητα να έρθει κεφαλή σε μία ρίψη είναι $p = 0{,}5$. Η πιθανότητα το πλήθος των κεφαλών να είναι ίσο με 20 σε συνολικά 50 ρίψεις υπολογίζεται βάσει της παραπάνω σχέσης ως εξής:

$$P(X = 20) = \binom{50}{20} 0{,}5^{20}(1-0{,}5)^{30} = 0.0419$$

Στην περίπτωση που το πείραμα επαναλαμβάνεται πολλές φορές ο μέσος αριθμός των κεφαλών ο οποίος αναμένεται να εμφανιστεί είναι ίσος με $50\ 0{,}5 = 25$ και τυπική απόκλιση είναι ίση με $50\ 0{,}5\ 0{,}5 = 12{,}5$.

Η διαδικασία πρόβλεψης της κλάσης ενός συνόλου εγγραφών μπορεί επίσης να θεωρηθεί ένα διωνυμικό πείραμα. Δοθέντος ενός συνόλου ελέγχου το οποίο περιέχει N εγγραφές, έστω ότι Xείναι ο αριθμός των εγγραφών που έχουν ταξινομηθεί σωστά από μοντέλο και p η πραγματική ακρίβεια του τελευταίου. Αν θεωρηθεί πως η διαδικασία πρόβλεψης είναι διωνυμικό πείραμα τότε το X κολουθεί τη διωνυμική κατανομή με μέσο Np και απόκλιση $Np(1-p)$. Έχει αποδειχτεί πως η εμπειρική ακρίβεια, $acc = X/N$ ακολουθεί επίσης τη διωνυμική κατανομή με μέσο N και απόκλιση $p(1-p)/N$. Παρά το γεγονός πως η διωνυμική κατανομή μπορεί να αξιοποιηθεί για την πρόβλεψη του διαστήματος

εμπιστοσύνης της accσυχνά υπολογίζεται μέσω μίας κανονικής κατανομής όταν το N είναι επαρκώς μεγάλο. Το διάστημα εμπιστοσύνης της acc στην περίπτωση αυτή είναι ίσο με:

$$P(-Z_{\alpha/2} \leq \frac{acc - p}{\sqrt{\frac{p(1-p)}{N}}} \leq Z_{1-\alpha/2} = 1 - a$$

Όπου: $Z_{\alpha/2}$ και $Z_{1-\alpha/2}$ το μέγιστο και ελάχιστο όριο όπως προκύπτει σε μία κανονική κατανομή με επίπεδο εμπιστοσύνης $1 - a$. Γενικά, προκύπτει πως$Z_{\frac{a}{2}} = Z_{1-\frac{a}{2}}$ εφόσον η κανονική κατανομή είναι συμμετρική γύρω από το σημείο $Z = 0$. Όσον αφορά την ανισότητα το διάστημα μπιστοσύνης για το p προκύπτει ως εξής:

$$\frac{xNxacc + Z_{\alpha/2}^2 \pm Z_{\alpha/2}\sqrt{Z_{\alpha/2}^2 + 4Nacc - 4Nacc^2}}{2(N + Z_{\alpha/2}^2)}$$

Στον ακόλουθο Πίνακα (Πίνακας 2.1) αναγράφονται οι τιμές του $Z_{\alpha/2}$σε διαφορετικά διαστήματα εμπιστοσύνης:

ΠΙΝΑΚΑΣ 2.1: ΤΙΜΕΣ ΤΟΥ $Z_{\alpha/2}$ ΣΕ ΔΙΑΦΟΡΕΤΙΚΑ ΔΙΑΣΤΗΜΑΤΑ ΕΜΠΙΣΤΟΣΥΝΗΣ

$1 - \alpha$	0.99	0.98	0.95	0.9	0.8	0.7	0.5
$Z_{\alpha/2}$	2.58	2.33	1.96	1.65	1.28	1.04	0.67

Σύγκριση της απόδοσης δύο μοντέλων

Έστω δύο μοντέλα M_1 και M_2. Οι επιδόσεις των συγκεκριμένων έχουν αξιολογηθεί βάσει δύο ανεξάρτητων συνόλων ελέγχου D_1 και D_2. Έστω ότι με n_1συμβολίζεται ο αριθμός των εγγραφών στο D_1και αντίστοιχα με n_2στο δεύτερο σύνολο. Επιπροσθέτως, γίνεται υπόθεση πως το ποσοστό σφάλματος για το M_1είναι e_1και e_2για το M_2. Στόχος της ακόλουθης διαδικασίας είναι να εξετάσει κατά πόσον η παρατηρούμενη διαφορά των e_1και e_2είναι στατιστικά σημαντική.

Θεωρώντας πως το πλήθος των δειγμάτωνn_1και n_2είναι επαρκώς μεγάλο, τα ποσοστά σφάλματος e_1και e_2μπορούν να υπολογιστούν μέσω κανονικής κατανομής. Το αντίστοιχο ισχύει και για τη διαφορά $d = e_1 - e_2$η οποία έχει μέση τιμή d_t(πραγματική διαφορά)και απόκλιση σ_d^2. Ησ_d^2μπορεί να υπολογιστεί βάσει της ακόλουθης σχέσης:

$$\sigma_d^2 \cong \hat{\sigma}_d^2 = \frac{e_1(1 - e_1)}{n_1} + \frac{e_2(1 - e_2)}{n_2}$$

Όπου $\frac{e_1(1-e_1)}{n_1}$, $\frac{e_2(1-e_2)}{n_2}$ οι αποκλίσεις των ποσοστών σφαλμάτων. Τέλος στο επίπεδο εμπιστοσύνης $(1 - \alpha)\%$, μπορεί να αποδειχτεί πως το διάστημα εμπιστοσύνης για την πραγματική διαφορά d_tυπολογίζεται βάσει της ακόλουθης εξίσωσης:

$$d_t = d \pm z_{a/2}\hat{\sigma}_d$$

Σύγκριση της απόδοσης δύο ταξινομητών

Έστω ότι απαιτείται η σύγκριση της απόδοσης δύο ταξινομητών, χρησιμοποιώντας την προσέγγιση διασταυρωμένη επικύρωση (cross validation). Αρχικά, το σύνολο δεδομένων D χωρίζεται σε k ίσου μεγέθους τμήματα. Στη συνέχεια, κάθε ταξινομητής κατασκευάζει ένα μοντέλο από τα $k - 1$τμήματα των δεδομένων και το ελέγχει μέσω του μοναδικού

τμήματος το οποίο παραμένει. Το βήμα αυτό επαναλαμβάνεται k φορές και ο ταξινομητής χρησιμοποιεί σε κάθε επανάληψη διαφορετικό τμήμα των δεδομένων ως σύνολο ελέγχου.

Έστω ότι με M_{ij} συμβολίζεται το μοντέλο το οποίο παράχθηκε από την τεχνική ταξινόμησης L_iκατά την j^{th}επανάληψη. Είναι φανερό πως τα ζεύγη μοντέλων M_{1j}και M_{2j} δοκιμάστηκαν στο ίδιο τμήμα j. Έστω ότι e_{1j}και e_{2j} είναι τα ποσοστά σφάλματος των M_{1j}και M_{2j} αντίστοιχα, η διαφορά τους είναι ίση με $d_j = e_{1,j} - e_{2,j}$. Στην περίπτωση που το k είναι επαρκώς μεγάλο τότε η προαναφερθείσα διαφορά ακολουθεί την κανονική κατανομή με μέση τιμή d_t^{cu}και (πραγματική διαφορά των ποσοστών σφάλματος) και απόκλιση σ^{cu}. Η συνολική απόκλιση των παρατηρούμενων διαφορών υπολογίζεται βάσει της ακόλουθης σχέσης:

$$\hat{\sigma}^2_{d^{cu}} = \frac{\sum_{j=1}^{k}(d_j - \bar{d})^2}{k(k-1)}$$

Όπου: $\bar{d}$ είναι η μέση διαφορά.

Στην προσέγγιση αυτή απαιτείται η χρήση της κατανομής t προκειμένου να υπολογιστεί το διάστημα εμπιστοσύνης για το d_t^{cu}:

$$d_t^{cu} = \bar{d} \pm t_{(1-a),k-1}\hat{\sigma}_{d^{cu}}$$

Η τιμή του $t_{(1-a),k-1}$ αντλείται μέσω ενός πίνακα πιθανοτήτων με δύο στοιχεία εισόδου: το επίπεδο εμπιστοσύνης $1-\alpha$και το πλήθος των βαθμών ελευθερίας $k-1$.

2.1.7 Λοιπά θέματα

Απόδοση βαρών σε οντότητες

Ορισμένοι αλγόριθμοι δέντρων απόφασης δίνουν μεγαλύτερη έμφαση σε ορισμένες οντότητες και αυτό πρακτικά υλοποιείται με την απόδοση βάρους (με τιμές 0-1) στις συγκεκριμένες. Με τον τρόπο αυτό η συνεισφορά των οντοτήτων αυτών στην κατασκευή του δέντρου διαφέρει από τις υπόλοιπες (Rokach and Maimon, 2005).

Κόστος λανθασμένης ταξινόμησης

Ορισμένοι αλγόριθμοι έχουν αριθμητικές κυρώσεις σε περιπτώσεις όπου μία οντότητα ταξινομείται σε λανθασμένη θεματική κατηγορία (Rokach and Maimon, 2005).

Διαχείριση ελλιπών τιμών

Η απουσία τιμών αποτελεί συνηθισμένο φαινόμενο στα (πραγματικά) σύνολα δεδομένων. Το παραπάνω δημιουργεί προβλήματα σε ό,τι αφορά την εξαγωγή συμπερασμάτων (induction) (για ένα σύνολο εκπαίδευσης όπου ορισμένες από τις τιμές απουσιάζουν) καθώς και στη διαδικασία της ταξινόμησης (μία καινούργια οντότητα στην οποία απουσιάζουν ορισμένες τιμές) (Rokach and Maimon, 2005).

Το παρόν πρόβλημα έχει αποτελέσει αντικείμενο μελέτης πλήθους ερευνητών (Friedman, 1977; Breiman et al., 1989; Quinlan, 1989). Ένας αποτελεσματικός τρόπος για τη διαχείριση του συγκεκριμένου σε ό,τι αφορά το σύνολο των δεδομένων εκπαίδευσης είναι ο ακόλουθος. Έστω ότι $\sigma_{a_i=?}S$ οι τιμές του υποσυνόλου S των οποίων οι τιμές a_iαπουσιάζουν. Στην περίπτωση αυτή οι οντότητες με άγνωστες τιμές a_i παραλείπονται κατά τη διαδικασία υπολογισμού του κριτηρίου διαχωρισμού. Συνεπώς το κριτήριο διαχωρισμού έχει ως εξής: $\Delta\Phi(a_i, S - \sigma_{a_i=?}S)$.

Από την άλλη πλευρά το κριτήριο διαχωρισμού μειώνεται αναλογικά στην περίπτωση ελλιπών τιμών, καθώς οι συγκεκριμένες οντότητες δε συνεισφέρουν στον υπολογισμό του συγκεκριμένου. Συνεπώς, γίνεται η ακόλουθη διόρθωση:

$$\frac{\left|S - \sigma_{a_i=?}S\right|}{|S|} \Delta\Phi\left(a_i, S - \sigma_{a_i=?}S\right)$$

Σε περιπτώσεις όπου γίνεται κανονικοποίηση του κριτηρίου (όπως για παράδειγμα στην αναλογία κέρδους) ο παρονομαστής υπολογίζεται με τρόπο τέτοιο ώστε οι ελλιπείς τιμές να αναπαριστούν μία πρόσθετη τιμή στον τομέα των ιδιοτήτων (attribute domain). Για παράδειγμα, η αναλογία κέρδους στην περίπτωση ελλιπών τιμών υπολογίζεται ως εξής:

$$GainRatio(a_i, S) = \frac{\frac{\left|S-\sigma_{a_i=?}S\right|}{|S|} InformationGain(a_i, S - \sigma_{a_i=?}S)}{-\frac{\left|\sigma_{a_i=?}S\right|}{|S|}\log\left(\frac{\left|\sigma_{a_i=?}S\right|}{|S|}\right) - \sum_{u_{i,j}\in dom(a_i)} \frac{\left|\sigma_{a_i=u_{i,j}}S\right|}{|S|}\log\left(\frac{\left|\sigma_{a_i=u_{i,j}}S\right|}{|S|}\right)}$$

Στη συνέχεια, εφόσον γίνει ο διαχωρισμός του κόμβου, είναι απαραίτητο να προστεθεί ο όρος $\sigma_{a_i=?}S$ σε κάθε μία από τις εξερχόμενες ακμές μέσω του ακόλουθου βάρους:

$$|\sigma_{a_i=u_{i,j}}S|/\left|S - \sigma_{a_i=?}S\right|$$

Στην περίπτωση ταξινόμησης μίας νέας οντότητας, της οποίας ορισμένες τιμές απουσιάζουν, γίνεται χρήση μίας παρόμοιας με την παραπάνω ιδέας. Όταν η οντότητα συναντά κάποιον κόμβο.

Οι Breiman et al. (1984) παρουσίασαν μία ακόμα προσέγγιση του εν λόγω προβλήματος. Η συγκεκριμένη ονομάζεται αναπληρωματικός διαχωρισμός (surrogate split) και εφαρμόζεται στον αλγόριθμο CART. Η ιδέα στην οποία βασίζεται η συγκεκριμένη μεθοδολογία είναι η εξής: Για κάθε κόμβο/ διαχωρισμό (split) γίνεται προσπάθεια να βρεθεί ένας αντίστοιχος αναπληρωματικός ο οποίος χρησιμοποιεί μεν διαφορετικό χαρακτηριστικό εισόδου, ωστόσο είναι παρόμοιος με τον αρχικό. Στην περίπτωση που η τιμή του χαρακτηριστικού του αρχικού κόμβου απουσιάζει για μία οντότητα γίνεται χρήση του αναπληρωματικού. Η ομοιότητα ανάμεσα στις δυαδικούς κόμβους για το χαρακτηριστικό S υπολογίζεται ως εξής:

$$res\left(a_i, dom_1(a_i), dom_2(a_i), a_j, dom_{1(a_j)}, dom_2\left(a_j\right), S\right) = \frac{|\sigma_{a_i\in dom_1(a_i)}\ AND_{a_j\in dom_1(a_j)}S|}{|S|} + \frac{|\sigma_{a_i\in dom_2(a_i)}\ AND_{a_j\in dom_2(a_j)}S|}{|S|}$$

Στην περίπτωση που ο πρώτος κόμβος αναφέρεται στο χαρακτηριστικό a_i ο $dom(a_i)$χωρίζεται σε $dom_1(a_i)$ και $dom_2(a_i)$. Ομοίως για το χαρακτηριστικό a_j.

Οι ελλιπείς τιμές μίας οντότητας μπορούν να υπολογιστούν βάσει άλλων οντοτήτων (Loh and Shih, 1997).Αναλυτικά, στην περίπτωση που η τιμή μίας ονομαστικής μεταβλητής a_iστην πλειάδα qαπουσιάζει κατά τη διαδικασία της εκπαίδευσης, υπολογίζεται βάσει των οντοτήτων που έχουν την ίδια τιμή ενός χαρακτηριστικού στόχου. Αναλυτικά

$$estimate\left(a_i, y_q, S\right) = argmax_{u_{i,j}\in dom(a_i)}|\sigma_{a_i=u_{i,j}}AND_{y=y_q}S|$$

Όπου το y_q συμβολίζει την τιμή του χαρακτηριστικού στόχου στην πλειάδα q. Στην περίπτωση που το χαρακτηριστικό a_i είναι αριθμητικό χρησιμοποιείται η μέση τιμή αυτού (Rokach and Maimon, 2005).

2.1.8 Παραδείγματα αλγόριθμων δέντρων απόφασης (Decision Trees Inducers)

Στο παρόν κεφάλαιο περιγράφονται διεξοδικά τα χαρακτηριστικά των δημοφιλέστερων αλγορίθμων δημιουργίας δέντρων απόφασης.

ID3

Ο ID3 είναι ένας απλός αλγόριθμος δημιουργίας δέντρων απόφασης ο οποίος προτάθηκε (Quinlan, 1986). Ο συγκεκριμένος χρησιμοποιεί για κριτήριο διαχωρισμού εκείνο του πληροφοριακού κέρδους και η κατασκευή του δέντρου διακόπτεται όταν όλες οι οντότητες ανήκουν σε μία μόνο κατηγορία ή το βέλτιστο δυνατό πληροφοριακό κέρδος δεν είναι μεγαλύτερο από την τιμή 0. Ο ID3 δεν ενσωματώνει κάποια διαδικασία κλαδέματος και επιπροσθέτως δεν είναι σε θέση να διαχειριστεί καταστάσεις ελλιπών τιμών και αριθμητικών χαρακτηριστικών (Rokach and Maimon, 2005).

C4.5

Ο C4.5 αποτελεί εξέλιξη του ID3 και κατασκευάστηκε από τον ίδιο δημιουργό (Quinlan, 1993). Κατά τη διαδικασία σχηματισμού του δέντρου χρησιμοποιεί την αναλογία κέρδους για κριτήριο διαχωρισμού και ο διαχωρισμός διακόπτεται εφόσον ο αριθμός των οντοτήτων που απομένουν να διαχωριστούν είναι κάτω από ένα ορισμένο κατώφλι. Στη συνέχεια, εφαρμόζεται στο παραχθέν δέντρο το κλάδεμα βασιζόμενο στο σφάλμα (Error based pruning) Τέλος, ο C4.5 είναι σε θέση να διαχειριστεί καταστάσεις ελλιπών τιμών χρησιμοποιώντας την αναθεωρημένη σχέση της αναλογίας κέρδους (Rokach and Maimon, 2005).

CART

Ο CART (Classification and Regression Trees) κατασκευάστηκε από τους(Breiman et al., 1984). Το χαρακτηριστικό του συγκεκριμένου αλγορίθμου είναι το γεγονός πως κατασκευάζει δυαδικά δέντρα. Συνεπώς, κάθε κόμβος ενός δέντρου έχει δύο ακριβώς εξερχόμενες ακμές. Οι διαχωρισμοί επιλέγονται βάσει του κριτηρίου towing και το δέντρο που προκύπτει κλαδεύεται εφαρμόζοντας τη μεθοδολογία κόστους πολυπλοκότητας. Ο CART είναι σε θέση να υπολογίσει τα κόστη λανθασμένης ταξινόμησης καθώς και την εκ των προτέρων κατανομή πιθανοτήτων.

Ένα ακόμα σημαντικό χαρακτηριστικό του συγκεκριμένου αλγορίθμου είναι η ικανότητά του να παράγει δέντρα παλινδρόμησης. Τα φύλλα των τελευταίων προβλέπουν πραγματικούς αριθμούς και όχι κλάσεις. Στη συγκεκριμένη περίπτωση ο αλγόριθμος αναζητά διαχωρισμούς οι οποίοι ελαχιστοποιούν το προβλεπόμενο τετραγωνικό σφάλμα και η πρόβλεψη για κάθε φύλλο υπολογίζεται βάσει του σταθμισμένου μέσου όρου του κόμβου (Rokach and Maimon, 2005).

CHAID

Στις αρχές της δεκαετίας του 70 οι ερευνητές στην εφαρμοσμένη στατιστική ανέπτυξαν διαδικασίες για τη δημιουργία δέντρων απόφασης όπως οι: AID (Sonquist et al., 1971), MAID (Gillo, 1972), THAID (Morgan and Messanger, 1973) και CHAID (Kass, 1980).OCHAID (Chi-square – Automatic – Interaction –Detection) σχεδιάστηκε αρχικά για τη διαχείριση αποκλειστικά ονομαστικών χαρακτηριστικών. Για κάθε χαρακτηριστικό εισόδου a_i, ο CHAID βρίσκει το ζεύγος των τιμών στο V_iτα οποία εμφανίζουν μικρή σημαντική διαφορά σε ό,τι αφορά ένα χαρακτηριστικό στόχο. Η στατιστική σημαντική διαφορά υπολογίζεται βάσει της τιμής p η οποία προέρχεται από στατιστικούς ελέγχους. Το στατιστικό τεστ βασίζεται στο

είδος του χαρακτηριστικού στόχου. Στην περίπτωση που το τελευταίο είναι συνεχές γίνεται χρήση του Ftest, ενώ αν είναι ονομαστικό γίνεται χρήση του Pearson chi squared test.

Ο αλγόριθμος ελέγχει για κάθε ζεύγος που επιλέγεται κατά πόσον η τιμή p υπερβαίνει ένα ορισμένο κατώφλι. Στην περίπτωση που η παραπάνω πρόταση ισχύει, οι τιμές ενώνονται και στη συνέχεια γίνεται έλεγχος για πιθανές συνενώσεις με άλλα ζεύγη. Η διαδικασία επαναλαμβάνεται έως ότου δεν υπάρξουν πρόσθετες συνενώσεις.

Στη συνέχεια, γίνεται επιλογή του καλύτερου χαρακτηριστικού εισόδου το οποίο θα χρησιμοποιηθεί για το διαχωρισμό του υπό εξέταση κόμβου. Η παραπάνω διαδικασία γίνεται με τρόπο τέτοιο ώστε κάθε παιδί κόμβος αποτελείται από ένα σύνολο τιμών οι οποίες είναι ομοιογενείς για τα συγκεκριμένα χαρακτηριστικά. Η διαδικασία τερματίζεται εφόσον ικανοποιηθούν κάποια από τα ακόλουθα κριτήρια:

- Επιτυγχάνεται το μέγιστο δυνατό βάθος
- Επιτυγχάνεται ο ελάχιστος αριθμός περιπτώσεων σε έναν κόμβο γονέα. Συνεπώς, ο κόμβος δε μπορεί να διαχωριστεί περαιτέρω
- Επιτυγχάνεται ο ελάχιστος αριθμός περιπτώσεων σε έναν κόμβο παιδί

Ο συγκεκριμένος αλγόριθμος διαχειρίζεται περιπτώσεις ελλιπών τιμών, επεξεργάζοντάς τες στο σύνολό τους σαν μία μοναδική κατηγορία. Τέλος, σημειώνεται πως ο συγκεκριμένος αλγόριθμος δεν περιλαμβάνει διαδικασία κλαδέματος (Rokach and Maimon, 2005).

QUEST

Ο αλγόριθμος QUEST (Quick, Unbiased, Efficient Statistical Tree) προτάθηκε από τους (Loh and Shih, 1997) και υποστηρίζει μονομεταβλητούς και γραμμικών συνδυασμών διαχωρισμούς. Η συσχέτιση ανάμεσα στα χαρακτηριστικά εισόδου και το χαρακτηριστικό στόχο υπολογίζεται μέσω του ANOVAF-test ή μέσω του Levene (για συνεχή χαρακτηριστικά) ή μέσω του Pearson chisquare (για ονομαστικά χαρακτηριστικά). Στην περίπτωση που το χαρακτηριστικό στόχος είναι πολυονομαστικό (multinomial) γίνεται ομαδοποίηση των τιμών προκειμένου να κατασκευαστούν δύο υπερκλάσεις. Στον QUEST γίνεται ανάλυση τετραγωνικής διακρίνουσας (quadratic discriminant) προκειμένου να βρεθεί το βέλτιστο σημείο διαχωρισμού για το χαρακτηριστικό εισόδου και η μέθοδος κλαδέματος που εφαρμόζεται είναι η Ten fold διασταυρωμένη επικύρωση (Rokach and Maimon, 2005).

Αναφορά σε άλλους αλγορίθμους

Στον ακόλουθο Πίνακα (Πίνακας 2.2) περιγράφονται συνοπτικά πρόσθετοι αλγόριθμοι δέντρων απόφασης, οι οποίοι δεν αναλύθηκαν παραπάνω και μάλιστα υπάρχει ακόμα πλήθος τέτοιων αλγορίθμων που δε συμπεριλαμβάνονται στην ακόλουθη λίστα. Αξίζει, ωστόσο να επισημανθεί πως οι περισσότεροι από αυτούς αποτελούν παραλλαγή των προαναφερθέντων (Rokach and Maimon, 2005).

ΠΙΝΑΚΑΣ 2.2: ΠΡΟΣΘΕΤΟΙ ΑΛΓΟΡΙΘΜΟΙ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ (ROKACH AND MAIMON, 2005)

Αλγόριθμος	Περιγραφή	Αναφορά
CAL5	Κατασκευάστηκε αποκλειστικά για αριθμητικά χαρακτηριστικά	Muller and Wisotzki (1994)
FACT	Αποτελεί μία πρότερη εκδοχή του QUEST. Το συγκεκριμένο χρησιμοποιεί στατιστικούς ελέγχους προκειμένου να γίνει επιλογή του χαρακτηριστικού εκείνου βάσει του οποίου θα γίνει ο διαχωρισμός σε κάποιον κόμβο και στη συνέχεια χρησιμοποιεί διακρίνουσα ανάλυση προκειμένου να βρεθεί το σημείο διαχωρισμού	Loh and Vanichsetakul (1988)
LMDT	Κατασκευάζει δέντρα απόφασης τα οποία βασίζονται σε μονομεταβλητούς ελέγχους. Οι συγκεκριμένοι είναι γραμμικοί συνδυασμοί των χαρακτηριστικών	Brodley and Utgoff (1995)
TI	Δημιουργεί ένα δέντρο απόφασης ενός επιπέδου, το οποίο ταξινομεί οντότητες χρησιμοποιώντας ένα μόνο χαρακτηριστικό. Οι ελλιπείς τιμές αντιμετωπίζονται ως ειδικές τιμές. Υποστηρίζει τόσο ονομαστικές όσο και συνεχείς τιμές	Holte (1993)
PUBLIC	Ενσωματώνει το μεγάλωμα και το κλάδεμα του δέντρου χρησιμοποιώντας το κόστος MDL. Με τον τρόπο αυτό γίνεται μείωση της υπολογιστικής περιπλοκότητας	Ragotsi and Shim (2000)
MARS	Υπολογίζεται μια πολλαπλή συνάρτηση regression χρησιμοποιώντας γραμμικές συναρτήσεις παρεμβολής καθώς και τα tensor παράγωγά τους	Friedman (1991)

2.1.9 Χαρακτηριστικά των αλγορίθμων κατασκευής δέντρων απόφασης

i. Οι αλγόριθμοι κατασκευής δέντρων απόφασης είναι μη παραμετρικοί. Με άλλα λόγια δεν απαιτούν κάποια εκ των προτέρων υπόθεση αναφορικά με το είδος των κατανομών πιθανότητας των κλάσεων καθώς και άλλων χαρακτηριστικών

ii. Η εύρεση του βέλτιστου δέντρου απόφασης είναι ένα *NP complete* πρόβλημα. (τα προβλήματα για τα οποία δεν έχει βρεθεί πολυωνυμική λύση- δηλαδή λύση η οποία να βρεθεί σε εύλογο χρονικό διάστημα) Πολλοί αλγόριθμοι δέντρων απόφασης

εφαρμόζουν μία ευρετική προσέγγιση (heuretic approach) προκείμενου να καθοδηγήσουν την έρευνά τους στον αχανή χώρο των υποθέσεων.

iii. Οι τεχνικές δημιουργίας δέντρων απόφασης έχουν χαμηλό υπολογιστικό κόστος. Συνεπώς, η διαδικασία κατασκευής των συγκεκριμένων ταξινομητών είναι ελάχιστα απαιτητική ακόμα και σε περιπτώσεις που το σύνολο των δεδομένων εκπαίδευσης είναι μεγάλο σε όγκο. Σημειώνεται επίσης πως εφόσον ολοκληρωθεί η κατασκευή του δέντρου η κατηγοριοποίηση είναι μία διαδικασία σύντομη.

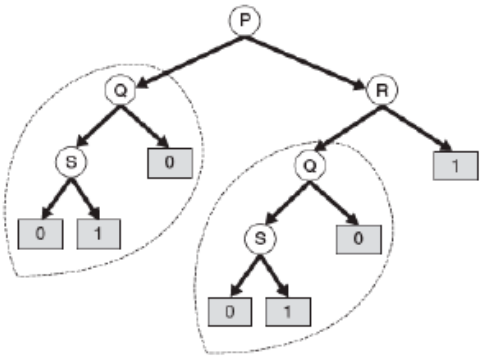
iv. Τα δέντρα απόφασης (και ιδιαίτερα εκείνα που είναι μικρά σε μέγεθος) είναι εύκολο να ερμηνευτούν

v. Οι αλγόριθμοι δέντρων απόφασης δεν είναι επιρρεπείς στο θόρυβο, ιδιαίτερα σε περιπτώσεις όπου οι μέθοδοι αποφεύγουν την υπερπροσαρμογή του μοντέλου

vi. Η παρουσία πλεοναζόντων χαρακτηριστικών (χαρακτηριστικών που εμφανίζουν υψηλή συσχέτιση) δεν επηρεάζει δυσμενώς την ακρίβεια του δέντρου. Στην περίπτωση αυτή το ένα εκ των δύο γνωρισμάτων αποκλείεται από τη διαδικασία διαχωρισμού. Ωστόσο, προβλήματα δημιουργούνται όταν το σύνολο των δεδομένων περιέχει πολλά περιττά γνωρίσματα (γνωρίσματα τα οποία δε συνεισφέρουν ιδιαίτερα στη διαδικασία της ταξινόμησης). Στις περιπτώσεις αυτές, τα τελευταία ενδέχεται να επιλεχθούν στο σχηματισμό των κόμβων και αυτό να οδηγήσει σε μεγαλύτερα δέντρα απόφασης. Οι τεχνικές επιλογής χαρακτηριστικών συντελούν στην αύξηση της ακρίβειας των μοντέλων παραλείποντας τα περιττά γνωρίσματα κατά τη διαδικασία της προεπεξεργασίας των δεδομένων.

vii. Οι περισσότεροι αλγόριθμοι κατασκευής δέντρων απόφασης υιοθετούν μία από πάνω προς τα κάτω (top-down) επαναληπτική διαδικασία διαχωρισμού των δεδομένων. Αυτό έχει σαν αποτέλεσμα τη μείωση του αριθμού των εγγραφών κατά την πορεία από τη ρίζα στα φύλλα του δέντρου. Ενδέχεται μάλιστα ο αριθμός των εγγραφών στα φύλλα να είναι τόσο μικρός ώστε το αποτέλεσμα της ταξινόμησης να μην είναι στατιστικά σημαντικό. Αυτό αναφέρεται στη βιβλιογραφία ως διάσπαση των δεδομένων (data fragmentation). Στις περιπτώσεις αυτές προτείνεται παρεμπόδιση της περαιτέρω διάσπασης των κόμβων, όταν ο αριθμός των εγγραφών σε αυτούς είναι μικρότερος από ένα κατώφλι.

viii. Σε ορισμένες περιπτώσεις εμφανίζονται στο δέντρο απόφασης ορισμένα υποδέντρα περισσότερες από μία φορές (Εικόνα 2.15). Αυτό οδηγεί σε πολυπλοκότερα μοντέλα, τα οποία είναι πιθανόν δύσκολα κατανοητά. Το παραπάνω συμβαίνει σε περιπτώσεις όπου η δημιουργία ενός δέντρου απόφασης βασίζεται σε ένα μόνο χαρακτηριστικό σε κάθε κόμβο. Εφόσον οι αλγόριθμοι δημιουργίας δέντρων βασίζονται στην τεχνική του διαίρει και βασίλευε, είναι πολύ πιθανόν να χρησιμοποιηθεί η ίδια συνθήκη ελέγχου σε διαφορετικούς κόμβους και συνεπώς να εμφανιστεί το εν λόγω πρόβλημα.



ΕΙΚΟΝΑ 2.15: ΠΑΡΑΔΕΙΓΜΑ ΕΠΑΝΑΛΗΠΤΙΚΟΤΗΤΑΣ ΚΟΜΒΩΝ

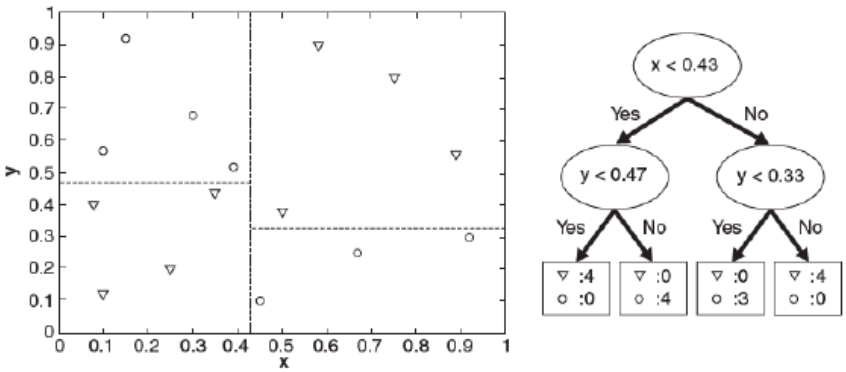
ix. Οι έλεγχοι συνθήκης που αναλύθηκαν εκτενώς στο παρόν κεφάλαιο χρησιμοποιούν μόνο ένα χαρακτηριστικό τη φορά. Συνεπώς, η δημιουργία του δέντρου αποτελεί μία διαδικασία διαχωρισμού των τιμών των χαρακτηριστικών σε επιμέρους περιοχές. Αυτό επαναλαμβάνεται έως ότου κάθε μία από τις τελευταίες να περιέχει εγγραφές οι οποίες ανήκουν σε μία κλάση. Η οριακή γραμμή μεταξύ δύο γειτονικών περιοχών που ανήκουν σε διαφορετικές κλάσεις ονομάζεται σύνορο απόφασης (decision boundary). Εφόσον οι συνθήκες ελέγχου είναι μονομεταβλητές, τα σύνορα απόφασης μπορούν να αναπαρασταθούν μέσω ορθογωνίων (οι τιμές που παίρνουν είναι παράλληλες στους άξονες συντεταγμένων) (Εικόνα 2.16). Αυτό δυσχεραίνει την αναπαράσταση σύνθετων σχέσεων ανάμεσα σε συνεχή γνωρίσματα (Εικόνα 2.17).

Στις περιπτώσεις αυτές προτείνεται η χρήση ενός πλάγιου/ λοξού (obligue) δέντρου απόφασης καθώς μέσω του συγκεκριμένου είναι δυνατή η αξιοποίηση περισσότερων του ενός γνωρισμάτων σε συνθήκες ελέγχου. Το σύνολο δεδομένων της Εικόνα 2.17 μπορεί εύκολα να αναπαρασταθεί μέσω ενός πλάγιου δέντρου απόφασης το οποίο περιλαμβάνει την ακόλουθη συνθήκη ελέγχου:

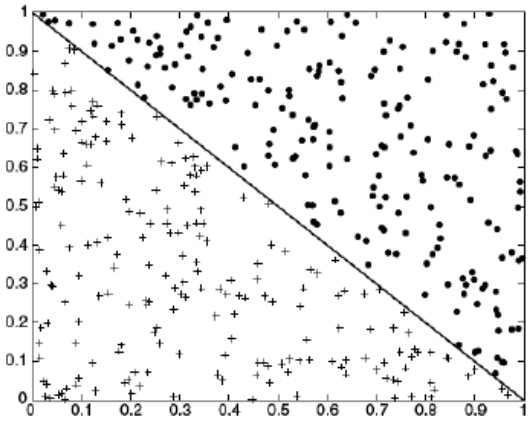
$$x+y<1$$

Οι τεχνικές αυτές παρέχουν μεγαλύτερη εκφραστικότητα συγκριτικά με τις προαναφερθείσες. Ωστόσο, οι συγκεκριμένες έχουν μεγάλο υπολογιστικό κόστος.

Ένας ακόμα τρόπος διαχείρισης τέτοιων καταστάσεων είναι η τεχνική της εποικοδομητικής επαγωγής (constructive induction). Μέσω αυτής είναι δυνατή η αναπαράσταση σύνθετων γνωρισμάτων μέσω αριθμητικών ή λογικών συνδυασμών άλλων χαρακτηριστικών.



ΕΙΚΟΝΑ 2.16: ΑΝΑΠΑΡΑΣΤΑΣΗ ΣΥΝΟΡΩΝ ΑΠΟΦΑΣΗΣ (TAN ET AL., 2005



ΕΙΚΟΝΑ 2.17: ΠΑΡΑΔΕΙΓΜΑ ΣΥΝΘΕΤΩΝ ΣΥΝΕΧΩΝ ΓΝΩΡΙΣΜΑΤΩΝ (TAN ET AL., 2005

x. Τα συμπεράσματα τα οποία προκύπτουν από έρευνες στο συγκεκριμένο μοντέλο δείχνουν πως η επιλογή του μέτρου καθαρότητας δεν επηρεάζει ιδιαίτερα την ακρίβεια του δέντρου απόφασης. Αυτό συμβαίνει διότι τα μέτρα εμφανίζουν παρόμοια συμπεριφορά, γεγονός που επιβεβαιώνεται το γράφημα της Εικόνα 2.10 (Tan et al., 2005).

2.1.10Πλεονεκτήματα και Μειονεκτήματα των Δέντρων Απόφασης

Στόχος της παρούσας ενότητας είναι η παράθεση των πλεονεκτημάτων και μειονεκτημάτων των δέντρων απόφασης. Αναλυτικά, στη βιβλιογραφία αναφέρονται τα ακόλουθα πλεονεκτήματα:

- Τα δέντρα απόφασης είναι αυτό – επεξηγηματικά. Με άλλα λόγια ένα δέντρο απόφασης το οποίο έχει έναν λογικό αριθμό φύλλων, μπορεί εύκολα να χρησιμοποιηθεί από μη επαγγελματίες χρήστες. Επιπροσθέτως, τα δέντρα απόφασης μπορούν εύκολα να μετατραπούν σε ένα σύνολο κανόνων. Βάσει των παραπάνω προκύπτει πως τα συγκεκριμένα αποτελούν κατανοητή μορφή αναπαράστασης.
- Τα δέντρα απόφασης μπορούν να χειριστούν τόσο αριθμητικά όσο και ονομαστικά χαρακτηριστικά εισόδου
- Η αναπαράσταση των δέντρων απόφασης είναι τόσο πλούσια ώστε τα συγκεκριμένα μπορούν να αναπαραστήσουν οποιοδήποτε ταξινομητή διακριτών τιμών
- Τα δέντρα απόφασης είναι σε θέση να διαχειριστούν σύνολα δεδομένων τα οποία εμπεριέχουν σφάλματα
- Τα δέντρα απόφασης είναι σε θέση να διαχειριστούν σύνολα δεδομένων με ελλιπείς τιμές
- Τα δέντρα απόφασης είναι μία μη παραμετρική μέθοδος. Αυτό σημαίνει πως τα συγκεκριμένα δεν εμπεριέχουν υποθέσεις αναφορικά με την κατανομή του χώρου καθώς και τη δομή του ταξινομητή (Rokach and Maimon, 2005)

Από την άλλη πλευρά τα δέντρα απόφασης έχουν τα ακόλουθα μειονεκτήματα:

- Πολλοί από τους αλγορίθμους όπως οι ID3 καιC4.5 απαιτούν τα χαρακτηριστικά στόχοι να έχουν μόνο διακριτές τιμές
- Εμφανίζουν αυξημένη ευαισθησία σε ό,τι αφορά τα δεδομένα εκπαίδευσης, τα μη συσχετισμένα χαρακτηριστικά καθώς και το θόρυβο (Rokach and Maimon, 2005)

2.1.11 Βιβλιογραφικές Σημειώσεις

Στη βιβλιογραφία αναφέρεται πως τα πρώτα συστήματα ταξινόμησης αναπτύχθηκαν προκειμένου να είναι δυνατή η οργάνωση μεγάλων συλλογών από αντικείμενα. Για παράδειγμα τα Dewey Decimal και Library of Congress σχεδιάστηκαν προκειμένου να γίνει καταγραφή και κατηγοριοποίηση του μεγάλου αριθμού των βιβλίων.

Η αυτοματοποιημένη ταξινόμηση αποτέλεσε αντικείμενο έρευνας για πολλά χρόνια. Η μελέτη της συγκεκριμένης διαδικασίας μέσω της στατιστικής είναι γνωστή ως διακριτική ανάλυση (discriminant analysis). Στόχος της τελευταίας είναι η πρόβλεψη της συμμετοχής ενός αντικειμένου σε μία κατηγορία βάσει ορισμένων μεταβλητών πρόβλεψης. Μία από τις πλέον διαδεδομένες μεθόδους διακριτικής ανάλυσης είναι εκείνη του Fisher, μέσω της οποίας γίνεται αναζήτηση της γραμμικής προβολής των δεδομένων. Με τον τρόπο αυτό, εξάγεται πληροφορία σχετικά με τη μεγαλύτερη τιμή της απόκλισης ανάμεσα σε αντικείμενα διαφορετικών κλάσεων.

Επιπροσθέτως, πολλά προβλήματα αναγνώρισης προτύπων απαιτούν πληροφορίες σχετικά με τις τιμές απόκλισης ανάμεσα σε διαφορετικά αντικείμενα. Σε εφαρμογές αναγνώρισης προτύπων περιλαμβάνονται διαδικασίες όπως η αναγνώριση ομιλιών και γραφικών χαρακτήρων καθώς και η ταξινόμηση εικόνας.

Ορισμένοι από τους πλέον διαδεδομένους αλγορίθμους κατασκευής δέντρων απόφασης είναι οι CART, ID3, C.4.5 και CHAID. Οι ID3 καιC.4.5 εφαρμόζουν τη συνάρτησης της εντροπίας προκειμένου να γίνει διαχωρισμός των δεδομένων. Ο Quinlan δημοσίευσε μία εις βάθος μελέτη του αλγορίθμου C4.5, στην οποία διατυπώνεται μεταξύ άλλων πως ο C4.5 τροποποιείται ώστε να μπορεί να διαχειριστεί σύνολα δεδομένων στα οποία ορισμένες τιμές γνωρισμάτων απουσιάζουν. Ο αλγόριθμος CART αναπτύχθηκε από τους Breiman et al. και χρησιμοποιεί για συνάρτηση διαχωρισμού το δείκτη GINI. Τέλος ο αλγόριθμος CHAID χρησιμοποιεί το στατιστικό έλεγχο χ^2προκειμένου να προσδιορίσει το βέλτιστο διαχωρισμό των δεδομένων κατά τη διαδικασία δημιουργίας του δέντρου.

Οι αλγόριθμοι κατασκευής δέντρων απόφασης που αναλύθηκαν εκτενώς στο παρόν κεφάλαιο βασίζονται στην υπόθεση πως ο διαχωρισμός των δεδομένων σε υποσύνολα μέσω του μοντέλου γίνεται βάσει μίας μόνο μεταβλητής σε κάθε κόμβο. Ένα πλάγιο δέντρο ταξινόμησης, ωστόσο, αξιοποιεί πολλαπλά γνωρίσματα προκειμένου να διαμορφώσει τις συνθήκες ελέγχου σε κάθε εσωτερικό κόμβο. Οι Breiman et al. μέσω του αλγορίθμου τους CART παρέχουν παρέχουν τη δυνατότητα γραμμικού συνδυασμού των επιμέρους γνωρισμάτων σε συνθήκες ελέγχου. Επιπροσθέτως οι Heath et al., Murthy et al., Cantupaz και Kamath και Utgoff και Brodley πρότειναν αλγορίθμους κατασκευής πλάγιων δέντρων απόφασης. Οι εν λόγω αλγόριθμοι βελτιώνουν την εκφραστικότητα της αναπαράστασης ενός δέντρου απόφασης. Εμφανίζουν, ωστόσο, το μειονέκτημα πως η εφαρμογή τους είναι ιδιαίτερα απαιτητική υπολογιστικά. Ένας ακόμα τρόπος βελτίωσης της εκφραστικότητας ενός δέντρου είναι μέσω της μεθόδου κατασκευαστικής επαγωγής (constructive induction). Η μέθοδος αυτή απλοποιεί τη διαδικασία της εκπαίδευσης μέσω της δημιουργίας σύνθετων γνωρισμάτων από τα απλά χαρακτηριστικά.

Οι προαναφερθέντες αλγόριθμοι βασίζονται στη στρατηγική από πάνω προς τα κάτω (top down) προκειμένου να κατασκευάσουν το δέντρο απόφασης. Εναλλακτικές προσέγγισες είναι η από κάτω προς τα πάνω (bottom- up) των Landeweerd et al., Pattipi καθώς και η αμφίδρομη των Alexandridis, Kim και Landgrebe, Schuermann και Doster και Wang και Suen. Στην τελευταία χρησιμοποιείται ένα «επιεικές» κριτήριο διαχωρισμού (soft splitting

criterion) και μέσω της προσέγγισης αυτής κάθε εγγραφή εκχωρείται σε διαφορετικά κλαδιά των δέντρων απόφασης με διαφορετικές πιθανότητες.

Η υπερπροσαρμογή του μοντέλου αποτελεί ένα σημαντικό θέμα στην κατασκευή των δέντρων απόφασης καθώς η αντιμετώπισή του είναι απαραίτητη ώστε να διασφαλιστεί πως το εκάστοτε μοντέλο έχει τις ίδιες επιδόσεις σε διαφορετικές άγνωστες σε αυτό καταγραφές. Το εν λόγω πρόβλημα έχει διερευνηθεί από πλήθος διαφορετικών συγγραφέων όπως οι Breiman et al., Schaffer, Mingers και Jensen και Cohen. Οι περισσότεροι βασίζονται στη υπόθεση πως το συγκεκριμένο πρόβλημα είναι αποτέλεσμα της παρουσίας θορύβου. Ωστόσο, οι Jensen και Cohen επισημαίνουν πως η υπερπροσαρμογή είναι αποτέλεσμα της χρήσης διαφορετικών ελέγχων υποθέσεων σε διαδικασίες πολλαπλών συγκρίσεων.

Στη βιβλιογραφία ορίζεται πως το σφάλμα γενίκευσης είναι η πιθανότητα της λανθασμένης ταξινόμησης ενός παραδείγματος ενώ το σφάλμα ελέγχου η παρουσία (fraction) σφαλμάτων σε ένα καινούργιο σύνολο ελέγχου. Συνεπώς, αντλείται το συμπέρασμα πως το σφάλμα γενίκευσης είναι το αναμενόμενο σφάλμα ελέγχου ενός ταξινομητή.

Τέλος, αξίζει να αναφερθεί πως ο Kohavi έκανε μία εκτενή εμπειρική μελέτη προκειμένου να συγκρίνει τις διαφορετικές μεθόδους εκτίμησης της απόδοσης διαφορετικών αλγορίθμων απόφασης, όπως η τυχαία υποδειγματοληψία, το bootstrapping και η πολλαπλή cross validation. Τα αποτελέσματα έδειξαν πως η καλύτερη μέθοδος είναι εκείνη της πολλαπλής cross validation και πιο συγκεκριμένα ο διαχωρισμός των δεδομένων σε 10 τμήματα. Οι Efron και Tibshirani παρείχαν μία θεωρητική και εμπειρική σύγκριση ανάμεσα στις μεθόδους bootstrapping και η πολλαπλή cross validation, η οποία είναι γνωστή με το όνομα 632+ κανόνας(Tan et al., 2005).

2.1.12 Τα δέντρα απόφασης στην επιστήμη της Ψηφιακής Τηλεπισκόπησης

Τα δέντρα απόφασης αποτελούν ένα απλό αλλά απαιτητικό χρονικά αλγόριθμο ταξινόμησης. Για την κατασκευή του συγκεκριμένου είναι απαραίτητη η μελέτη και λεπτομερής ανάλυση των φασματικών διαγραμμάτων διασποράς των διαφορετικών θεματικών κατηγοριών. Αφετηρία της συγκεκριμένης διαδικασίας είναι η ρίζα του δέντρου απόφασης στην οποία περιλαμβάνονται όλες οι θεματικές κατηγορίες. Στη συνέχεια, υλοποιούνται επιμέρους ταξινομήσεις στους εσωτερικούς κόμβους, έως το τερματισμό της διαδικασίας στο φύλλωμα του δέντρου, στο οποίο επιτυγχάνεται η ταξινόμηση της κάθε κατηγορίας (Αργιαλάς, 1998).

Διευκρινίζεται πως η παραπάνω διαδικασία αναφέρεται σε περιπτώσεις ταξινόμησης στις οποίες μοναδιαίο στοιχείο είναι το εικονοστοιχείο. Στην αντικειμενοστραφή ανάλυση εικόνας τα γνωρίσματα, στα οποία βασίζεται η ταξινόμηση της εικόνας είναι περισσότερα από εκείνο της φασματικής ανακλαστικότητας.

Τα δέντρα απόφασης έχουν χρησιμοποιηθεί επανειλημμένως στην Επιστήμη της Ψηφιακής Τηλεπισκόπησης. Ενδεικτικά, αναφέρονται οι ακόλουθες εφαρμογές:

- Οι (Friedl and Brodley , 1997) στο άρθρο τους “Decision tree classification of land cover from remotely sensed data” παρουσίασαν τα αποτελέσματα εφαρμογής διαφορετικών αλγορίθμων κατασκευής δέντρων απόφασης για την ταξινόμηση τριών τηλεπισκοπικών δεδομένων. Αναλυτικά, έγινε χρήση ενός μονομεταβλητού

(univariate), ενός πολυμεταβλητού (multivariate) και ενός υβριδικού δέντρου ταξινόμησης. Μέσω του τελευταίου είναι δυνατή η συμπερίληψη διαφορετικού είδους αλγορίθμων σε ένα μοναδικό δέντρο απόφασης.

- Οι (Pal and Mather , 2003) στο άρθρο τους "An assessment of the effectiveness of decision tree methods for land cover classification" αξιοποιούν αλγορίθμους δημιουργίας δέντρων απόφασης για την ταξινόμηση πολυφασματικών δεδομένων από το δέκτη Landsat ETM+ καθώς και υπερφασματικών από το DAIS. Η ταξινόμηση έγινε μέσω μονομεταβλητών (univariate) καθώς και πολυμεταβλητών (multivariate) δέντρων απόφασης. Στα πλαίσια της παρούσας εφαρμογής έγινε πλήθος πειραμάτων σε ό,τι αφορά το μέγεθος του συνόλου εκπαίδευσης, τις διαστάσεις του χώρου των διαστάσεων και τη μέθοδο κλαδέματος. Τα συμπεράσματα τα οποία προέκυψαν είναι πως το επίπεδο της ακρίβειας του αποτελέσματος ταξινόμησης μέσω των μονομεταβλητών δέντρων απόφασης είναι συγκρίσιμο με εκείνο άλλων δημοφιλών μεθόδων ταξινόμησης. Επιπροσθέτως δε σημειώνονται σημαντικές διαφορές σε ό,τι αφορά την ακρίβεια σε μονομεταβλητά και πολυμεταβλητά δέντρα απόφασης. Τέλος, αξίζει να σημειωθεί πως οι επιδόσεις των δέντρων απόφασης ήταν χαμηλότερες συγκριτικά με άλλες μεθόδους ταξινόμησης σε μεγάλων διαστάσεων δεδομένα.
- Οι (Xu et al., 2005) στο άρθρο τους "Decision tree regression for soft classification of remote sensing data" αναφέρουν πως τα δέντρα απόφασης έχουν εφαρμοστεί σε πολλές περιπτώσεις σε τηλεπισκοπικά δεδομένα. Το αποτέλεσμα αυτής της διαδικασίας είναι μία «σκληρή» ταξινόμηση (hard and crisp classification).Βασικό χαρακτηριστικό των δορυφορικών εικόνων ιδιαίτερα σε εκείνες με χαμηλή χωρική ανάλυση είναι η σύνθεση τους από εικονοστοιχεία τα οποία αντιστοιχίζονται στο έδαφος σε περισσότερες από μία θεματικές κατηγορίες. Συνεπώς, η εφαρμογή σε αυτές διαδικασιών ταξινόμησης εικονοστοιχείου οδηγεί σε χαμηλής ποιότητας αποτελέσματα και για το λόγο αυτό, οι συγγραφείς προτείνουν την αξιοποίηση μεθόδων «απαλής» ταξινόμησης (soft classification). Οι συγκεκριμένες βασίζονται στην ιδέα πως κάθε εικονοστοιχείο ενδέχεται να ανήκει σε περισσότερες από μία κλάσεις και συνεπώς έχει κάποιο βαθμό συμμετοχής σε κάθε εμφανιζόμενη θεματική κατηγορία της εικόνας. Στην παρούσα εφαρμογή το παραπάνω υλοποιείται μέσω αναδρομικών(regression)δέντρων απόφασης.
- Οι (Laliberte et al., 2007) στο άρθρο τους "Combining Decision Trees with Hierarchical Object- oriented Image Analysis for Mapping Arid Rangelands" εντόπισαν μέσω δέντρων απόφασης άγονες περιοχές σε μία υψηλής ευκρίνειας δορυφορική εικόνα Quickbird. Η ταξινόμηση έγινε σε αντικείμενα της εικόνας και μάλιστα σημειώνεται πως η κατάτμηση έγινε σε τέσσερις διαφορετικές κλίμακες.
- Ο (Heumann, 2011) στο άρθρο του "An Object-Based Classification of Mangroves Using a Hybrid Decision Tree- Support Vector Machine Approach" γίνεται ανίχνευση ριζοφόρων από τηλεπισκοπικά δεδομένα υψηλής ευκρίνειας μέσω αντικειμενοστρεφούς ανάλυσης εικόνας. Οι εικόνες προέρχονται από το δέκτη Worldview-2 και η υλοποίηση της εφαρμογής έγινε μέσω μίας συνδυαστικής μεθόδου ταξινόμησης. Αναλυτικά, έγινε χρήση των δέντρων ταξινόμησης και SVM. Η ακρίβεια των αποτελεσμάτων είναι εντυπωσιακή καθώς είναι μεγαλύτερη από 94%.
- Οι (Powers et al., 2015) στο άρθρο τους "Remote sensing and object- based techniques for mapping fine- scale industrial disturbances" εστίασαν τις προσπάθειες τους στην αναγνώριση περιβαλλοντικών διαταραχών από

βιομηχανικές δραστηριότητες και πιο συγκεκριμένα στη χαρτογράφηση πετρελαϊκής άμμου στο βορειοανατολικό αρκτικό δάσος της Αλβέρτας. Συγκεκριμένα, έγινε αντικειμενοστρεφής ανάλυση μίας δορυφορικής εικόνας προερχόμενης από το δέκτη SPOT-5. Η ταξινόμηση έγινε μέσω δέντρων απόφασης και η συνολική ακρίβεια της διαδικασίας αυτής έφτασε ποσοστό υψηλότερο από 88%.

2.2 Τυχαία δάση

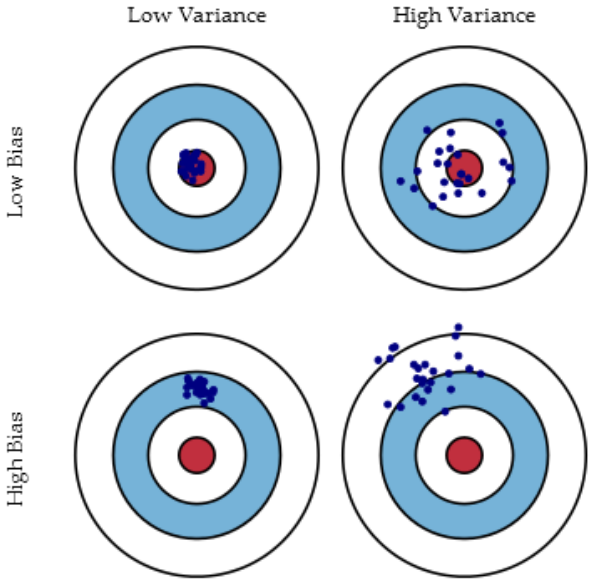
2.2.1 Εισαγωγικά στοιχεία

Διακύμανση και συστηματικό σφάλμα

Τα σφάλματα στη διαδικασία της ταξινόμησης κατηγοριοποιούνται ως εξής:

- Συστηματικά σφάλματα (bias): Η συγκεκριμένη κατηγορία σφαλμάτων αποτυπώνει ουσιαστικά τη διαφορά ανάμεσα στην τιμή του μοντέλου και στην πραγματική. Είναι φυσικό πως στην περίπτωση που υπάρχει μόνο ένα μοντέλο ο όρος αναμενόμενη ή μέση τιμή δεν είναι ιδιαίτερα χρήσιμος. Το παραπάνω έχει σημασία όταν η διαδικασία δημιουργίας μοντέλων επαναλαμβάνεται περισσότερες από μία φορές (όπως για παράδειγμα στο bagging και στα τυχαία δάση) καθώς τα τελευταία δίνουν ένα εύρος διαφορετικών προβλέψεων. Μέσω του συγκεκριμένου μεγέθους υπολογίζεται η διαφορά των προβλέψεων αυτών από τις πραγματικές τιμές.
- Διακύμανση (Variance): Τα σφάλματα που οφείλονται στη διακύμανση σχετίζονται με την μεταβλητότητα της πρόβλεψης του μοντέλου ως προς ένα δοθέν σημείο (μία εγγραφή). Όπως και στην προηγούμενη περίπτωση ο όρος αυτός αποκτά σημασία στην περίπτωση που γίνεται χρήση πολλών διαφορετικών μοντέλων. Μέσω του συγκεκριμένου μεγέθους υπολογίζεται πόσο διαφέρουν οι προβλέψεις για ένα δοθέν σημείο ανάμεσα στα διαφορετικά μοντέλα.

Στην ακόλουθη Εικόνα (Εικόνα 2.18) εμφανίζεται γραφικά η διαφορά ανάμεσα στα συστηματικά σφάλματα και στη διακύμανση[5].



ΕΙΚΟΝΑ 2.18:ΓΡΑΦΙΚΗ ΑΠΕΙΚΟΝΙΣΗ ΤΥΠΙΚΟΥ ΣΦΑΛΜΑΤΟΣ ΚΑΙ ΔΙΑΚΥΜΑΝΣΗΣ[6]

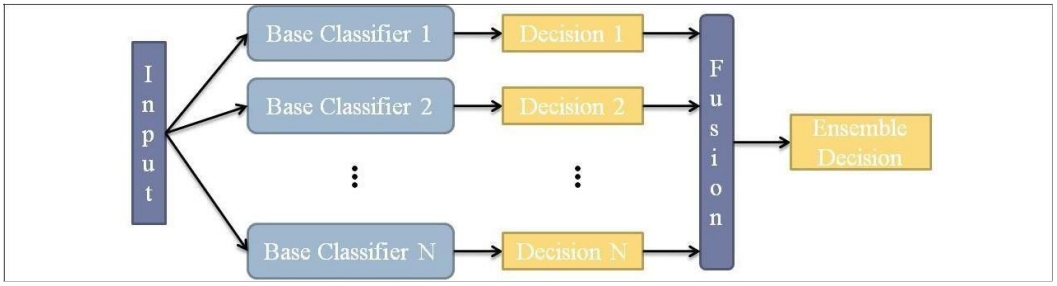
[5] http://scott.fortmann-roe.com/docs/BiasVariance.html

Ομάδες ταξινομητών ή συνδυαστικός ταξινομητής (Ensemble classifier)

Σε προηγούμενη ενότητα έγινε εκτενής αναφορά στην επιβλεπόμενη ταξινόμηση. Συνοπτικά, αναφέρεται πως η συγκεκριμένη αποτελεί μία διαδικασία εκπαίδευσης ενός αλγορίθμου από ένα σύνολο εγγραφών, οι οποίες έχουν γνωστές τιμές χαρακτηριστικών. Στόχος της διαδικασίας αυτής είναι ταξινόμηση των εγγραφών σε θεματικές κατηγορίες βάσει των γνωρισμάτων τους.

Οι επιδόσεις ενός ταξινομητή εξαρτώνται μεταξύ άλλων από την κατανομή των τιμών των γνωρισμάτων. Αξίζει να σημειωθεί πως σε πολύπλοκες περιπτώσεις η εκπαίδευση ενός μόνο μοντέλου δεν αρκεί καθώς οι πιθανότητες χαμηλής προσαρμογής του συγκεκριμένου στα δεδομένα εισόδου είναι υψηλές (Tasnim and Rahman, 2014).

Η λύση στο παραπάνω πρόβλημα δόθηκε μέσω της ιδέας των συνδυαστικών ταξινομητών (ensemble classifier) (Εικόνα 2.19). Μέσω του συγκεκριμένου όρου γίνεται αναφορά σε μία ομάδα μεμονωμένων ταξινομητών οι οποίοι έχουν εκπαιδευτεί συνεργατικά και ανεξάρτητα μέσω ενός συνόλου δεδομένων προκειμένου να καταστεί δυνατή η επίλυση ενός προβλήματος επιβλεπόμενης ταξινόμησης. Κάθε ένας από τους παραπάνω ταξινομητές - βάση (base classifiers) παρέχει κατά τη διαδικασία της πρόβλεψης μία συγκεκριμένη απόφαση και το σύνολο των παραπάνω ενοποιούνται σε μια μέσω μίας μεθόδου συγχώνευσης. Στη βιβλιογραφία γίνεται αναφορά σε πλήθος τέτοιων τεχνικών. Ενδεικτικά αναφέρονται οι ψήφισμα πλειοψηφίας (majority voting), η καταμέτρηση Borda και οι αλγεβρικοί συνδυασμοί (algebraic combiners) (Tasnim and Rahman, 2014).



ΕΙΚΟΝΑ 2.19: ΣΥΝΟΛΟ ΤΑΞΙΝΟΜΗΤΩΝ (TASNIM AND RAHMAN, 2014)

Τα σύνολα ταξινομητών βασίζονται στην ιδέα πως η ύπαρξη πολλών μοντέλων αντισταθμίζει κατά κάποιο τρόπο το σφάλμα στο αποτέλεσμα της ταξινόμησης του ενός. Οι συγκεκριμένοι αλγόριθμοι χρησιμοποιούν ίδια ή και διαφορετικά μοντέλα προκειμένου να επιτύχουν πολλές επαναλαμβανόμενες ταξινομήσεις συγκεκριμένων δεδομένων. Η αντιμετώπιση, ωστόσο, της εκπαίδευσης των ταξινομητών- βάση με ένα απλοϊκό τρόπο δεν συνεισφέρει στην επίλυση του προβλήματος. Έχει αποδειχτεί πως ο συγκεκριμένος αλγόριθμος επιτυγχάνει καλύτερα αποτελέσματα όταν τα επιμέρους μοντέλα που εντάσσονται σε αυτό είναι ακριβή και μεταξύ τους ασυσχέτιστα. Στη βιβλιογραφία οι μέθοδοι συσχέτισης δεδομένων διακρίνονται σε μέτρα συσχέτισης ανά ζεύγη (τα στατιστικά Q, συντελεστής συσχέτισης, μέτρο διαφοράς, μέτρο διπλού λάθους) και στα υπόλοιπα (εντροπία, απόκλιση Kohavi -Wolpert) (Tasnim and Rahman, 2014).
Γενικά αποδεικνύεται πως οι τεχνικές συνδυαστικών ταξινομητών επιτυγχάνουν υψηλότερες επιδόσεις συγκριτικά με τους απλούς ταξινομητές. Το παραπάνω οφείλεται

[6] http://scott.fortmann-roe.com/docs/BiasVariance.html

στο γεγονός πως οι συγκεκριμένες εκμεταλλεύονται τα πλεονεκτήματα των επιμέρους ταξινομητών και παράλληλα καταστρατηγούν τις αδυναμίες τους (Rodriguez- Galiano et al., 2011).

Ορισμένες από τις πλέον διαδεδομένες τεχνικές συνδυαστικών ταξινομητών είναι οι ακόλουθες:

- Bagging: βασίζεται στην εκπαίδευση πολλών ταξινομητών από δείγματα που έχουν ληφθεί από το σύνολο εκπαίδευσης μέσω της τεχνικής booststrap. Το παραπάνω έχει αποδειχτεί πως μειώνει την απόκλιση της ταξινόμησης.
- Boosting: μέσω της συγκεκριμένης τεχνικής γίνεται επανάληψη της διαδικασίας εκπαίδευσης σε περιπτώσεις που ορισμένα δείγματα είναι λάθος ταξινομημένα. Μάλιστα στα τελευταία δίνεται βάρος. Σημειώνεται πως κατά τη διαδικασία της εκπαίδευσης των επιμέρους ταξινομητών λαμβάνονται υπόψη όλα τα δείγματα εκπαίδευσης. Το παραπάνω καθιστά τον αλγόριθμο αργό (ιδιαίτερα συγκριτικά με το Bagging), ωστόσο ο συγκεκριμένος έχει εμφανώς μεγαλύτερη ακρίβεια σε σχέση με την προαναφερθείσα τεχνική. Το boosting μειώνει σε γενικές γραμμές τόσο την απόκλιση όσο και το συστηματικό σφάλμα της ταξινόμησης και επιπροσθέτως έχει αποδειχθεί πως είναι μία ιδιαίτερα ακριβής μέθοδος ταξινόμησης. Η συγκεκριμένη εμφανίζει ωστόσο ορισμένα μειονεκτήματα όπως το γεγονός πως είναι αργή και ευαίσθητη στο θόρυβο (Gislason et al., 2004).
- Τα τυχαία δάση είναι ουσιαστικά μία τροποποίηση του bagging τα οποία αποτελούνται από ένα σύνολο μη συσχετισμένων δέντρων απόφασης. Οι επιδόσεις των τυχαίων δασών είναι παρόμοιες με εκείνες του boosting και παράλληλα τα συγκεκριμένα είναι απλούστερα στη διαδικασία της εκπαίδευσης. Η συγκεκριμένη μέθοδος είναι ιδιαίτερα δημοφιλής και εφαρμόζεται σε πλήθος διαφορετικών εφαρμογών (Hastie et al., 2008).

ΤΕΧΝΙΚΗ BAGGING (BOOTSTRAP AGGREGATING)

Σε προηγούμενη ενότητα έγινε εκτενής παρουσίαση της τεχνικής bootstrap μέσω της οποίας γίνεται εκτίμηση της ακρίβειας μίας πρόβλεψης. Μέσω του Bagging γίνεται χρήση της προαναφερθείσας τεχνικής ούτως ώστε να γίνει βελτίωση της πρόβλεψης αυτής καθαυτής.

Αρχικά, εξετάζεται ένα πρόβλημα παλινδρόμησης (regression). Βάσει των δεδομένων εκπαίδευσης $\boldsymbol{Z} = \{(x_!, y_1), (x_2, y_2), \dots, (x_n, y_n)$ γίνεται προσαρμογή μίας συνάρτησης $\hat{f}(x)$ του χαρακτηριστικού x. Η μέθοδος bagging λαμβάνει το μέσο όρο των προβλέψεων αυτών για ένα σύνολο δεδομένων bootstrap και με τον τρόπο αυτό γίνεται μείωση των αποκλίσεων ανάμεσα σε αυτές. Αναλυτικά, για κάθε δείγμα bootstrap $\boldsymbol{Z}^{*\boldsymbol{b}}, b = 1, 2, \dots B$ γίνεται πρόβλεψη ενός μοντέλου $f^{*b}(x)$ και βάσει αυτού ο υπολογισμός του bagging γίνεται βάσει της ακόλουθης σχέσης:

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x)$$

Στην περίπτωση της ταξινόμησης δημιουργείται ένα σύνολο μοντέλων όπου το κάθε ένα από αυτά ψηφίζει μία κλάση βάσει της δικής του πρόβλεψης.

Η τεχνική bagging έχει ιδιαίτερα εντυπωσιακά αποτελέσματα σε περιπτώσεις υψηλής απόκλισης καθώς και σε συστηματικά σφάλματα (bias) (Hastie et al., 2008).

ΤΕΧΝΙΚΗ BOOSTING

Η τεχνική boosting αποτελεί όπως και το bagging μία ομαδική μέθοδος. Μέσω της συγκεκριμένης προσέγγισης, ωστόσο, οι αδύναμοι ταξινομητές εξελίσσονται στη διάρκεια του χρόνου και κάθε ένα από τα μέλη ψηφίζει για το τελικό αποτέλεσμα βάσει ενός διαφορετικού βάρους. Η συγκεκριμένη μέθοδος είναι ισχυρότερη και ως εκ τούτου δημοφιλέστερη σε σχέση με εκείνη του Bagging.

ΤΕΧΝΙΚΗ ΤΟΥ HO

Τα δέντρα απόφασης αποτελούν ένα ευρέως διαδεδομένο είδος ταξινομητή καθώς η εφαρμογή τους παρουσιάζει πλήθος πλεονεκτημάτων. Η ιδέα στην οποία βασίζεται ο συγκεκριμένος αλγόριθμος είναι ιδιαίτερα ελκυστική και η διαδικασία της ταξινόμησης μέσω των δέντρων απόφασης είναι εντυπωσιακά γρήγορη. Κατά καιρούς έχει υλοποιηθεί πλήθος μελετών οι οποίες αποσκοπούν στην αύξηση της ακρίβειας της ταξινόμησης και στη μείωση του μεγέθους των μοντέλων. Το βασικό μειονέκτημα των δέντρων απόφασης έγκειται στο γεγονός πως σε πολλές περιπτώσεις εμφανίζουν υπερπροσαρμογή στα δεδομένα εκπαίδευσης και το παραπάνω επηρεάζει αρνητικά την ακρίβεια του μοντέλου. Οι προτεινόμενες μεθοδολογίες κλαδέματος των δέντρων αυξάνουν το σφάλμα γενίκευσης ενός μοντέλου και μειώνουν την ακρίβεια προσαρμογής του δέντρου στα δεδομένα εκπαίδευσης (Ho, 1995).

Βάσει της υπάρχουσας βιβλιογραφίας προκύπτει πως υπάρχει ένας βασικός περιορισμός σχετικά με την πολυπλοκότητα των δέντρων καθώς δεν έχει αναπτυχθεί κάποια μεθοδολογία η οποία να είναι σε θέση να αυξάνει την ακρίβεια τόσο τα δεδομένα εκπαίδευσης όσο και σε εκείνα του ελέγχου Ο (Ho, 1995) προτείνει τη χρήση πλάγιων δέντρων (oblique decision trees) (όχι μονομεταβλητά αλλά πολυμεταβλητά δέντρα απόφασης). Η συγκεκριμένη κατηγορία αναλύθηκε εκτενώς στο κεφάλαιο των δέντρων απόφασης. Επιγραμματικά, επισημαίνεται πως μέσω του συγκεκριμένου παρέχεται η δυνατότητα αξιοποίησης περισσότερων του ενός γνωρισμάτων σε συνθήκες ελέγχου. Η συνθήκη ελέγχου στις περιπτώσεις αυτές είναι δυνατόν να εκφραστεί μέσω μίας γραμμικής συνάρτησης των επιμέρους χαρακτηριστικών (Ho, 1995).

Οι προτεινόμενες μεθοδολογίες δημιουργούν πολύπλοκα δέντρα τα οποία προσαρμόζονται στα δεδομένα εκπαίδευσης. Προβλήματα, ωστόσο εντοπίζονται καθώς η επιλογή της εκάστοτε συνθήκης ελέγχου εμφανίζει προτίμηση στα δεδομένα εκπαίδευσης και δεν είναι απαραίτητα αντικειμενική. Επιπροσθέτως, η μείωση του μεγέθους του δέντρου και συνεπώς του βαθμού προσαρμογής του στα δεδομένα εισόδου επιδρά αρνητικά στο σφάλμα εκπαίδευσης. Το παραπάνω δεν αποτελεί θετική ένδειξη ως προς την απόδοση του δέντρου στα δεδομένα ελέγχου. Η αξιοποίηση τεχνικών συνόλων αποτελεί ένα αποτελεσματικό τρόπο διαχείρισης του συγκεκριμένου μειονεκτήματος. Για το σκοπό αυτό ο (Ho, 1995) προτείνει τη χρήση πολλαπλών δέντρων απόφασης, την ανάπτυξη με άλλα λόγια ενός δάσους.

Για το σκοπό αυτό, είναι απαραίτητη η εύρεση ενός τρόπου μέσω του οποίου τα δέντρα απόφασης θα αναπτύσσονται ανεξάρτητα. Παράλληλα, κατασκευάζεται μία διακριτή συνάρτηση, μέσω της οποίας γίνεται συνδυασμός του αποτελέσματος των ταξινομήσεων των μεμονωμένων δέντρων απόφασης σε μία.

Στο σημείο αυτό τίθεται το εξής ερώτημα: «Πώς είναι δυνατή η δημιουργία μεγάλου αριθμού δέντρων απόφασης από ένα μόνο σύνολο δεδομένων;» Στη βιβλιογραφία αναγράφεται πλήθος τεχνικών δημιουργίας διαφορετικών δέντρων απόφασης, μέσω των

οποίων οι διαφορές στα μοντέλα που κατασκευάζονται είναι αυθαίρετες. Οι συγκεκριμένες δεν δίνουν τα επιθυμητά αποτελέσματα, δηλαδή δέντρα απόφασης τα οποία είναι 100% ακριβή στα δεδομένα ελέγχου και παράλληλα διαφέρουν μεταξύ τους ως προς το σφάλμα γενίκευσης. Παράδειγμα τέτοιας τεχνικής είναι η δημιουργία μοντέλων από διαφορετικά υποσύνολα των δεδομένων. Είναι εμφανές πως τα δέντρα αυτά δεν είναι σε θέση να ταξινομήσουν με τη μεγαλύτερη δυνατή ακρίβεια το σύνολο των δεδομένων εκπαίδευσης.

Ο (Ho, 1995) αναφέρει πως η τυχαιοποίηση (randomization) αποτελεί ένα ισχυρό εργαλείο δημιουργίας διαφορετικών ταξινομητών. Μέσω της συγκεκριμένης είναι δυνατή η εκπαίδευση αλγορίθμων διαφορετικής σύνθεσης οι οποίοι κατασκευάζουν διαφορετικούς ταξινομητές.

Ο (Ho, 1995) προτείνει μία νέα μεθοδολογία δημιουργίας δέντρων απόφασης, τα οποία διαφέρουν μεταξύ τους ως προς τα χαρακτηριστικά τα οποία χρησιμοποιούν για τη σύνθεση των συνθηκών ελέγχου. Με άλλα λόγια κάθε ένα από τα επιμέρους μοντέλα χρησιμοποιεί ένα διαφορετικό υποσύνολο των γνωρισμάτων των εγγραφών και η επιλογή του τελευταίου γίνεται μέσω της τυχαιοποίησης. Για ένα σύνολο m διαφορετικών χαρακτηριστικών υπάρχουν 2^mδιαφορετικοί συνδυασμοί γνωρισμάτων. Επιπροσθέτως σημειώνεται πως τα δέντρα εκπαιδεύονται αξιοποιώντας το σύνολο των δεδομένων εκπαίδευσης.

2.2.2 Τυχαία δάση (Random Forest)

Η βασική ιδέα στην οποία βασίζεται το bagging (αναλύθηκε εκτενώς παραπάνω) είναι η “εξομάλυνση” μοντέλων που περιέχουν θόρυβο (άρα μεγάλη διακύμανση) και παράλληλα εμφανίζουν μικρό συστηματικό σφάλμα. Με τον τρόπο αυτό επιτυγχάνεται μείωση της διακύμανσης των προβλεπόμενων τιμών.

Τα δέντρα απόφασης είναι ιδανικοί ταξινομητές- βάση για την τεχνική bagging καθώς παρέχουν τη δυνατότητα αναγνώρισης και αναπαράστασης σύνθετων συσχετίσεων ανάμεσα στα διαφορετικά χαρακτηριστικά των δεδομένων. Επιπροσθέτως, τα συγκεκριμένα εμφανίζουν μικρό συστηματικό σφάλμα (bias), υπό την προϋπόθεση πως έχουν αναπτυχθεί επαρκώς. Το συγκεκριμένο είδος ταξινομητή εμφανίζει, ωστόσο, το μειονέκτημα του θορύβου, συνεπώς η τεχνική bagging λειτουργεί βοηθητικά προς την κατεύθυνση της διαχείρισης της συγκεκριμένης αδυναμίας (Hastie et al., 2008).

Τα δέντρα απόφασης που δημιουργούνται μέσω της τεχνικής bagging είναι όμοια κατανεμημένα (identically distributed) και συνεπώς το μέσο συστηματικό σφάλμα που εμφανίζεται για το σύνολο B των μοντέλων είναι ίσο με εκείνο που εμφανίζει κάθε ένα από αυτά. Συνεπώς, ο μόνος τρόπος βελτίωσης των επιδόσεων των δέντρων είναι μέσω της μείωσης της διακύμανσης. Το παραπάνω στοιχείο είναι εκείνο που διαφοροποιεί την τεχνική bagging από εκείνη του boosting καθώς τα δέντρα αναπτύσσονται με τρόπο τέτοιο ώστε να μειώσουν το συστηματικό σφάλμα (Hastie et al., 2008).

Το μέσο συστηματικό σφάλμα των B ανεξάρτητων (δηλαδή $\rho = 0$) όμοια κατανεμημένων τυχαίων μεταβλητών (δηλαδή οι τιμές των γνωρισμάτων των εγγραφών), εκ των οποίων κάθε μία από τις οποίες εμφανίζει απόκλιση/ μεταβλητότητα (variance) σ^2, έχει διακύμανση $\frac{1}{B}\sigma^2$. Στην περίπτωση που οι μεταβλητές είναι απλά ομοιόμορφα κατανεμημένες (όχι απαραίτητα ανεξάρτητες) και έχουν ανά ζεύγη θετική συσχέτιση ρ, η διακύμανση του συνόλου των ταξινομητών υπολογίζεται βάσει της ακόλουθης σχέσης:

$$\rho\,\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

Εξισωση 1

Όσο το πλήθος B αυξάνεται ο δεύτερος όρος της παραπάνω σχέσης μειώνεται και συνεπώς παραμένει μόνο ο πρώτος. Επομένως, η συσχέτιση ρ ανά ζεύγη δέντρων περιορίζει τα οφέλη της τεχνικής του συνόλου των ταξινομητών. Μέσω των τυχαίων δασών γίνεται προσπάθεια μείωσης της συνολικής διακύμανσης του bagging μέσω της μείωσης της συσχέτισης ρ . Το συγκεκριμένο επιτυγχάνεται μέσω της τυχαίας επιλογής των χαρακτηριστικών εισόδου. Η ιδέα αυτή είναι εμπνευσμένη από τον αλγόριθμο του Ho που αναλύθηκε παραπάνω (Hastie et al., 2008).

Ιδιαίτερα σε περιπτώσεις που η ανάπτυξη ενός δέντρου γίνεται μέσω ενός συνόλου δεδομένων που προέκυψε μέσω bootstrap διαδικασίας:

Προτού γίνει διαχωρισμός βάσει μίας συνθήκης ελέγχου γίνεται επιλογή $m \leq p$ *γνωρισμάτων*

Συνήθως, η τιμή που επιλέγονται για το m είναι ίση με $\sqrt{p}$. Εφόσον, ολοκληρωθεί η δημιουργία B δέντρων της μορφής $\{T; \Theta_b\}_B^1$ η συνάρτηση τυχαίων δασών που προκύπτει είναι η ακόλουθη:

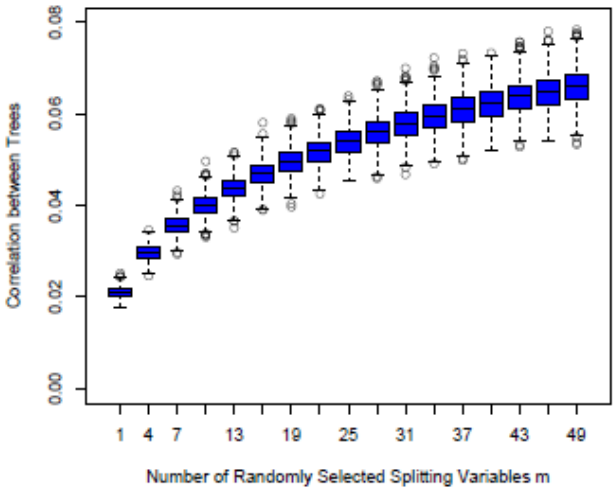
$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x; \Theta_b)$$

όπου μέσω του Θ_b συμβολίζεται το bth δέντρο του τυχαίου δάσους (Hastie et al., 2008).

Διευκρινίζεται πως η παραπάνω συνάρτηση ισχύει στην περίπτωση της παλινδρόμησης (regression). Διαισθητικά η μείωση του m οδηγεί σε μείωση της συσχέτισης ανάμεσα στα μοντέλα του δάσους και συνεπώς προκύπτει πως βάσει της Εξίσωση 1 η διακύμανση μειώνεται (Εικόνα 2.20).

Στην περίπτωση δέντρων που έχουν αναπτυχθεί μέσω της μεθόδου bootstrap η τιμή του ρ είναι συνήθως μικρή (0.05 ή και μικρότερη - εικονα) ενώ η τιμή σ^2 δεν είναι πολύ μεγαλύτερη συγκριτικά με του αρχικού δέντρου (Hastie et al., 2008).

Στην Εικόνα 2.21 εμφανίζεται ο αλγόριθμος των τυχαίων δασών για την περίπτωση της παλινδρόμησης καθώς και της ταξινόμησης.



ΕΙΚΟΝΑ 2.20: ΣΥΣΧΕΤΙΣΗ ΤΩΝ ΔΕΝΤΡΩΝ Ρ ΣΥΝΑΡΤΗΣΕΙ ΤΗΣ ΜΕΤΑΒΛΗΤΗΣ Μ ΤΑ ΤΕΤΡΑΓΩΝΑ ΑΝΑΠΑΡΙΣΤΟΥΝ ΤΗ ΣΥΣΧΕΤΙΣΕΙΣ ΣΤΑ 600 ΤΥΧΑΙΑ ΕΠΙΛΕΓΜΕΝΑ ΣΗΜΕΙΑ ΠΡΟΒΛΕΨΗΣ Χ (HASTIE ET AL., 2008).

Algorithm 17.1 *Random Forest for Regression or Classification.*

1. For $b = 1$ to B:
   (a) Draw a bootstrap sample $\mathbf{Z}^*$ of size N from the training data.
   (b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
      i. Select m variables at random from the p variables.
      ii. Pick the best variable/split-point among the m.
      iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.

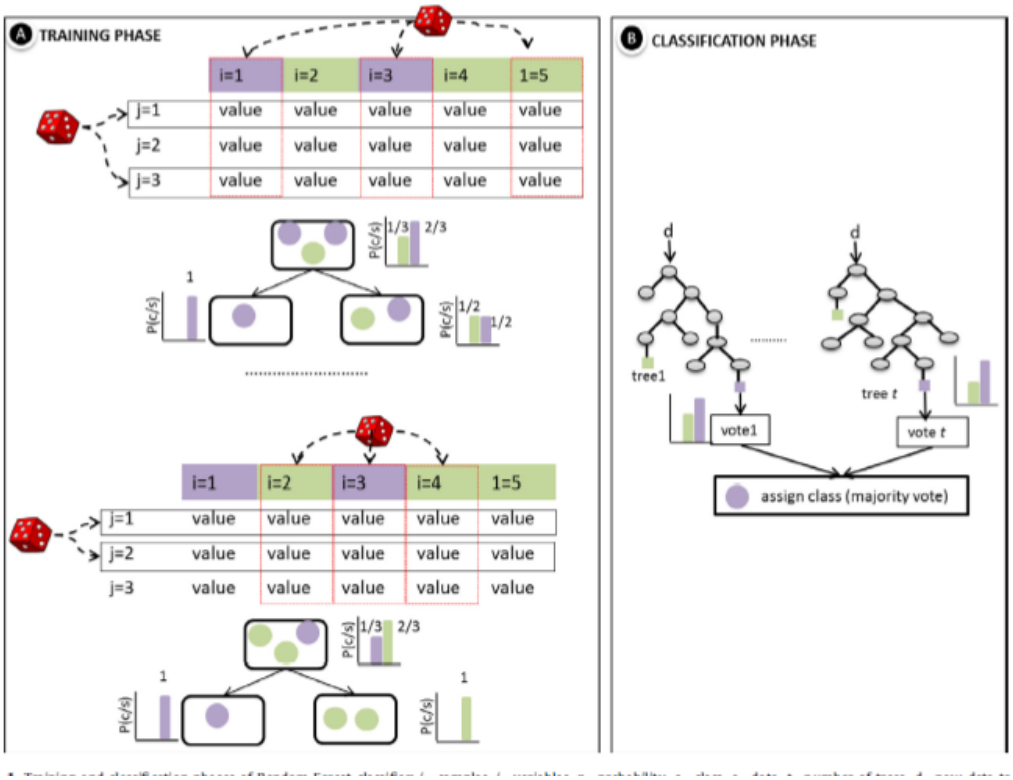
To make a prediction at a new point x:

Regression: $\hat{f}_{\text{rf}}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$.

Classification: Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{\text{rf}}^B(x) = \textit{majority vote}\ \{\hat{C}_b(x)\}_1^B$.

ΕΙΚΟΝΑ 2.21: ΑΛΓΟΡΙΘΜΟΣ ΤΥΧΑΙΩΝ ΔΑΣΩΝ

Στην Εικόνα 2.22 εμφανίζεται η διαδικασία εφαρμογής του αλγορίθμου των τυχαίων δασών σε ένα σύνολο δεδομένων.



ΕΙΚΟΝΑ 2.22: ΟΠΤΙΚΟΠΟΙΗΣΗ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ

Ο υπό μελέτη αλγόριθμος είναι ιδιαίτερα δημοφιλής και προτάθηκε για πρώτη φορά από τον Breiman. Μάλιστα, η συνεργάτης του τελευταίου Adele Cutler διατηρεί ιστοσελίδα[7] μέσω της οποίας ο κώδικας των τυχαίων δασών διανέμεται ελεύθερα. Είναι εντυπωσιακό πως ο αριθμός των μεταφορτώσεων το έτος 2002 έφτασε τις 3000. Οι συγγραφείς των τυχαίων δασών αναφέρονται στην επιτυχία του αλγορίθμου και τον χαρακτηρίζουν ως τον πιο ακριβή και τον απλούστερο/ πιο εύληπτο. Ένα ακόμα πλεονέκτημα του τελευταίου είναι η απουσία ρύθμισης πολλών παραμέτρων (Hastie et al., 2008).

Δημιουργία του τυχαίου δάσους

Η διαδικασία δημιουργίας του δάσους είναι η ακόλουθη:

I. Αρχικά δημιουργούνται n_{tree}bootstrap δείγματα
II. Βάσει κάθε bootstrap δείγματος δημιουργείται ένα δέντρο απόφασης το οποίο δεν υφίσταται κάποια διαδικασία κλαδέματος. Η διαδικασία διαχωρισμού των δεδομένων στους επιμέρους κόμβους είναι διαφοροποιημένη στην περίπτωση των τυχαίων δασών συγκριτικά με εκείνη των συμβατικών δέντρων απόφασης. Αν με M συμβολίζεται ο συνολικός αριθμός γνωρισμάτων κάθε εγγραφής τότε το πλήθος των χαρακτηριστικών που επιλέγονται τυχαία σε κάθε κόμβο είναι ίσο με m, τέτοιο ώστε m<<M. Για το διαχωρισμό του κόμβου χρησιμοποιείται η καλύτερη διάσπαση σε αυτά τα m (είναι μονομεταβλητό). Σημειώνεται πως το m διατηρείται σταθερό κατά τη διαδικασία δημιουργίας του δέντρου. (Η τεχνική bagging είναι ουσιαστικά μία υποπερίπτωση των τυχαίων δασών όπου m=M)
III. Η ταξινόμηση των νέων δεδομένων έπειτα από τη συγκέντρωση των αποφάσεων των n_{tree} δέντρων απόφασης (Liaw and Wienner, 2002).

[7] http://www.math.usu.edu/~adele/forests/

Λεπτομέρειες των τυχαίων δασών

Όπως έχει ήδη αναφερθεί ο συγκεκριμένος αλγόριθμος χρησιμοποιείται σε περιπτώσεις τόσο ταξινόμησης όσο και παλινδρόμησης. Στην πρώτη περίπτωση κάθε ένα από τα δέντρα του δάσους ψηφίζουν σχετικά με το αποτέλεσμα της ταξινόμησης και η τελική απόφαση λαμβάνεται βάσει της πλειοψηφίας των δέντρων. Στην παλινδρόμηση υπολογίζεται ο μέσος όρος των προβλέψεων αναφορικά με κάποιο σημείο (target point x).

Ως προς τη ρύθμιση των παραμέτρων του αλγορίθμου οι δημιουργοί του έργου προτείνουν τα ακόλουθα:

- Στην ταξινόμηση η προεπιλεγμένη τιμή του m είναι $\sqrt{p}$ και το ελάχιστο μέγεθος του κόμβου είναι ίσο με 1
- Στην παλινδρόμηση η προεπιλεγμένη τιμή του m είναι $p/3$ και το ελάχιστο μέγεθος του κόμβου είναι ίσο με 1

Πρακτικά, ωστόσο, η παραπάνω διαδικασία εξαρτάται σε μεγάλο βαθμό από το είδος του προβλήματος (Hastie et al., 2008).

Δείγματα εκτός της σακούλας (Out of bag samples)

Ένα ιδιαίτερα σημαντικό χαρακτηριστικό των τυχαίων δασών είναι ο άμεσος τρόπος υπολογισμού των επιδόσεων του εκάστοτε μοντέλου. Πιο συγκεκριμένα, στον παρόν αλγόριθμο παρέχεται η δυνατότητα υπολογισμού ενός αμερόληπτου σφάλματος κατά τη διαδικασία δημιουργίας του δάσους χωρίς να είναι απαραίτητη η διενέργεια διασταυρωμένης επικύρωσης ή οποιουδήποτε άλλου ξεχωριστού ελέγχου.

$$Ταξινόμησε\ μέσω\ του\ τυχαίου\ δάσους\ κάθε\ εγγραφή$$

$$z_i = (x_i, y_i)\ χρησιμοποιώντας\ μόνο\ τα\ δέντρα\ εκείνα$$

$$τα\ οποία\ δε\ χρησιμοποίησαν\ τη\ συγκεκριμένη\ ως\ δεδομένο\ εκπαίδευσης$$

Κάθε δέντρο απόφασης σε ένα τυχαίο δάσος κατασκευάζεται βάσει ενός διαφορετικού bootstrap δείγματος από τα αρχικά δεδομένα. Κατά τη διαδικασία αυτή περίπου το ένα τρίτο των συνολικών εγγραφών παραμένουν εκτός κατά τη σύνθεση του εκάστοτε δέντρου απόφασης και συνεπώς δε συνεισφέρουν στην εκπαίδευσή του (δείγματα εκτός σακούλας, out of bag samples- oob).

Οι εγγραφές αυτές αποτελούν το σύνολο ελέγχου των επιδόσεων κάθε δέντρου. Με τον τρόπο αυτό, γίνεται έλεγχος κάθε εγγραφής που περιέχεται στο σύνολο δεδομένων από το ένα τρίτο περίπου των δέντρων απόφασης που συνθέτουν το δάσος. Η εκτίμηση του σφάλματος oob υλοποιείται βάσει της ακόλουθης διαδικασίας: Η εκάστοτε jth εγγραφή ταξινομείται από το τυχαίο δάσος χρησιμοποιώντας αποκλειστικά το υποσύνολο των δέντρων απόφασης τα οποία δεν εκπαιδεύτηκαν βάσει της εγγραφής αυτής. Μέσω της διαδικασίας αυτής προκύπτουν $B/3$ προβλέψεις για το jth δείγμα όπου με B συμβολίζεται το συνολικό πλήθος των δέντρων απόφασης στο τυχαίο δάσος. Οι τελευταίες ενοποιούνται σε μία λαμβάνοντας το μέσο όρο αυτών (στην περίπτωση της παλινδρόμησης) ή την πλειοψηφία (στην περίπτωση της ταξινόμησης). Συνεπώς, μέσω αυτού προκύπτει μία OOB πρόβλεψη για την jth εγγραφή. Η παραπάνω διαδικασία επαναλαμβάνεται για το σύνολο των δειγμάτων τα οποία συμπεριλαμβάνονται στο σύνολο εκπαίδευσης και με τον τρόπο αυτό προκύπτει ένα συνολικό OOB MSE (για προβλήματα ταξινόμησης) και ένα σφάλμα

ταξινόμησης (για σφάλματα ταξινόμησης- δηλαδή διαιρώ τα σφάλματα με το συνολικό πλήθος των εγγραφών) (James et al., 2013).

Η διαδικασία της εκπαίδευσης είναι δυνατόν να τερματιστεί εφόσον το σφάλμα oob σταθεροποιηθεί (Hastie et al., 2008).

Σημαντικότητα μεταβλητών (Variable Importance)

Σε προηγούμενη ενότητα έγινε αναφορά στα πλεονεκτήματα της χρήσης των συνδυαστικών ταξινομητών μεταξύ των οποίων συμπεριλαμβάνεται η αύξηση στην ακρίβεια του αποτελέσματος. Η χρήση όμως των συγκεκριμένων αλγορίθμων συνοδεύεται από ορισμένα μειονεκτήματα όπως είναι για παράδειγμα η δυσκολία στην ερμηνεία του εκάστοτε μοντέλου. Το τελευταίο δεν ισχύει για τα δέντρα απόφασης καθώς τα συγκεκριμένα είναι ιδιαίτερα ελκυστικά και εύκολα στην ερμηνεία τους. Στην περίπτωση, ωστόσο, χρήσης μεγάλου αριθμού αυτών η αναπαράσταση της τελικής διαδικασίας εκπαίδευσης από ένα μόνο δέντρο είναι αδύνατη και επιπροσθέτως δεν είναι πια ξεκάθαρο ποιες από τις μεταβλητές είναι οι πλέον απαραίτητες στη διαδικασία της ταξινόμησης. Βάσει αυτών προκύπτει πως η χρήση των συνδυαστικών ταξινομητών αυξάνει μεν την ακρίβεια του αποτελέσματος αλλά αυτό συμβαίνει σε βάρος της επεξηγηματικότητας του μοντέλου (James et al., 2013).

Η λύση στο παραπάνω πρόβλημα δίνεται στον αλγόριθμο των τυχαίων δασών ως εξής: Σε κάθε δέντρο που έχει αναπτυχθεί πλήρως σε ένα δάσος τοποθετούνται για ταξινόμηση οι περιπτώσεις oob και για τις συγκεκριμένες γίνεται καταμέτρηση των δέντρων που έδωσαν ορθό αποτέλεσμα. Στη συνέχεια, γίνεται μετάθεση των τιμών της μεταβλητής m τυχαία και η παραπάνω διαδικασία επαναλαμβάνεται. Ο αριθμός των ψήφων που τοποθετούν τις εγγραφές στη σωστή κλάση στα νέα δεδομένα αφαιρείται από τον αντίστοιχο για τα αρχικά. Η διαφορά που προκύπτει διαιρείται με το συνολικό πλήθος των δέντρων που περιλαμβάνονται στο δάσος και η τιμή αυτή ονομάζεται βαθμολογία ακατέργαστης σημασίας για τη μεταβλητή m.

Σε περιπτώσεις που ο αριθμός των μεταβλητών είναι αρκετά μεγάλος, η εκπαίδευση του δάσους με αξιοποίηση του συνόλου των χαρακτηριστικών γίνεται μόνο μία φορά. Στη συνέχεια, η διαδικασία επαναλαμβάνεται με χρήση των πιο σημαντικών μεταβλητών όπως αυτές προέκυψαν μέσω της πρώτης εκτέλεσης[8].

Σημαντικότητα GINI (GINI importance)

Κριτήριο για την επιλογή ενός διαχωρισμού βάσει ενός γνωρίσματος m σε κάθε δέντρο απόφασης είναι ο δείκτης μη καθαρότητας GINI (αναλύθηκε διεξοδικά σε παραπάνω ενότητα). Αναλυτικά, σε κάθε δέντρο απόφασης του τυχαίου δάσους ο δείκτης GINI των κόμβων –παιδιά είναι μικρότερος σε σχέση με τον αντίστοιχο του γονέα. Επιπροσθέτως, η μείωση του συγκεκριμένου μέτρου για κάθε ανεξάρτητη μεταβλητή στο σύνολο των δέντρων του δάσους δίνει αξιόπιστη πληροφορία σχετικά με τη σημαντικότητα των μεταβλητών [9].

Αλληλεπιδράσεις/ Συσχετίσεις (Interactions)

Η έννοια της αλληλεπίδρασης ορίζεται στο συγκεκριμένο αλγόριθμο ως εξής: Έστω m,k δύο χαρακτηριστικά των εγγραφών. Τα παραπάνω αλληλοεπιδρούν όταν ο διαχωρισμός βάσει της μίας εκ των παραπάνω μεταβλητών, λόγου χάρη της m καθιστά τον αντίστοιχο για την k

[8] https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.html
[9] https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.html

περισσότερο ή λιγότερο πιθανό. Αυτό πρακτικά υλοποιείται μέσω των δεικτών GINI για κάθε ένα από τα δέντρα του δάσους. Οι τιμές που προκύπτουν κατατάσσονται για κάθε δέντρο καθώς και για κάθε δύο γνωρίσματα και στη συνέχεια λαμβάνεται ο μέσος όρος της απόλυτης διαφοράς των παραπάνω κατατάξεων για όλα τα δέντρα.

Το παρόν μέγεθος υπολογίζεται επίσης, βάσει της υπόθεσης πως τα δύο γνωρίσματα είναι ανεξάρτητα και συνεπώς γίνεται αφαίρεση της τιμής του ενός από την αντίστοιχη του άλλου. Στην περίπτωση που η διαφορά που προκύπτει είναι ένα ένας μεγάλος θετικός αριθμός σημαίνει ότι ο διαχωρισμός βάσει του ενός χαρακτηριστικού αναστέλλει τον αντίστοιχο για τον άλλο και αντίστροφα. Πρόκειται για μια πειραματική διαδικασία τα συμπεράσματα της οποίας θα πρέπει να αντιμετωπίζονται με προσοχή καθώς η συγκεκριμένη έχει δοκιμαστεί σε λίγα σύνολα δεδομένων[10].

Γράφημα ομοιότητας

Ένα από τα στοιχεία εξόδου του αλγορίθμου των τυχαίων δασών είναι τα διαγράμματα ομοιότητας. Κατά τη δημιουργία του συγκεκριμένου αλγορίθμου κατασκευάζεται ένας πίνακας ομοιότητας διαστάσεων N×N τα στοιχεία του οποίου συμπληρώνονται μέσω της ακόλουθης διαδικασίας: Εφόσον ο σχηματισμός του δέντρου ολοκληρωθεί, όλα τα δεδομένα εκπαίδευσης και τα oob τοποθετούνται σαν στοιχεία εισόδου στο μοντέλο. Στην περίπτωση που οι εγγραφές k και n καταλήγουν στον ίδιο τερματικό κόμβο το μέτρο της εγγύτητας τους αυξάνεται κατά ένα βαθμό[11] (Hastie et al., 2008).

Οι χρήστες των δασών παρατήρησαν πως σε περιπτώσεις μεγάλου σε όγκο δεδομένων ένας πίνακας διαστάσεων NχN δεν ήταν δυνατόν να χωρέσει στη γρήγορη μνήμη και για το λόγο αυτό ο συγκεκριμένος αντικαταστάθηκε από έναν αντίστοιχο διαστάσεων NxT όπου T ο συνολικός αριθμός των δέντρων του δάσους

Ο πίνακας αναπαρίσταται γραφικά στο χώρο μέσω μίας πολυδιάστατης κλιμάκωσης. Μέσω της συγκεκριμένης γίνεται αναζήτηση ενός χώρου χαμηλών διαστάσεων, συνήθως ευκλείδειου στον οποίο κάθε ένα από τα σημεία αναπαριστούν ένα αντικείμενο (εν προκειμένω μία εγγραφή) με τρόπο τέτοιο ώστε η απόσταση ανάμεσα στα σημεία του χώρου {d_rs} αναπαριστά με τον καλύτερο δυνατό τρόπο τις αρχικές ανομοιότητες ανάμεσα στα αντικείμενα (Hastie et al., 2008) (Cox and Cox, 2001).

Η διαδικασία της κλιμάκωσης είναι η ακόλουθη: Οι ομοιότητες ανάμεσα σε δύο εγγραφές n και k διαμορφώνουν ένα πίνακα $\{prox(n,k)\}$. Εξ ορισμού ο συγκεκριμένος είναι συμμετρικός και το σύνολο των στοιχείων του είναι θετικοί αριθμοί μεγαλύτεροι της μονάδας. Σημειώνεται επίσης πως τα στοιχεία της διαγώνιου του Πίνακα είναι ίσα με τη μονάδα. Όπως προκύπτει οι τιμές $1 - prox(n,k)$ είναι οι τετραγωνικές αποστάσεις στον Ευκλείδειο χώρο με διαστάσεις όχι μεγαλύτερες από τον αριθμό των συνολικών εγγραφών.

Έστω ότι με $prox(-,k)$ συμβολίζεται ο μέσος όρος των στοιχείων $prox(n,k)$ για την πρώτη συντεταγμένη, με $prox(n,-)$ ο μέσος όρος για τη δεύτερη συντεταγμένη και τέλος με $prox(-,-)$ ο μέσος όρος του συνόλου των στοιχείων. Ο πίνακας

$$cv(n,k) = 5\,(prox(n,k) - prox(n,-) - prox(-,k) + prox(-,-))$$

[10] https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.html
[11] https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.html

είναι ο πίνακας των εσωτερικών γινομένων των αποστάσεων και επιπροσθέτως είναι συμμετρικός. Έστω ότι οι ιδιοτιμές του cv είναι $\lambda(j)$ και το ιδιοάνυσμα είναι $v_j(n)$. Τότε τα διανύσματα:

$$x(n) = (O\lambda(1)v(n), O\lambda(2)v(n), .)$$

έχουν μεταξύ τους τετραγωνικές αποστάσεις ίσες με $1 - prox(n,k)$. Η τιμή του $O\lambda(j)v(n)$ αναφέρεται ως η jth συντεταγμένη κλίμακας. Στη μετρική κλίμακα, η ιδέα βασίζεται στον υπολογισμό των διανυσμάτων μέσω των ελάχιστων πρώτων συντεταγμένων κλιμάκωσης. Αυτό στα τυχαία δάση υλοποιείται μέσω εξαγωγής των ελάχιστων ιδιοτιμών του πίνακα cv και των σχετικών ιδιοδιανυσμάτων. Η δισδιάστατη απεικόνιση των συντεταγμένων της ith ως προς της jth δίνει συχνά χρήσιμες πληροφορίες για τα δεδομένα.

Εφόσον, οι ιδιοσυναρτήσεις είναι οι ελάχιστες τιμές ενός Πίνακα NχN, ο υπολογιστικός φόρτος ενδέχεται σε πολλές περιπτώσεις να είναι ιδιαίτερα χρονοβόρος. Για το λόγο αυτό, προτείνεται η μείωση των διαστάσεων του Πίνακα, ώστε το μέγεθός του να είναι μικρότερο από εκείνο του πλήθους των δειγμάτων.

Στη βιβλιογραφία αναφέρεται πλήθος μεθόδων, οι οποίες μειώνουν τις αποστάσεις σε μικρότερες διαστάσεις, όπως για παράδειγμα ο αλγόριθμος Rowels και Saul. Ωστόσο, οι καλές επιδόσεις της μετρητικής κλιμάκωσης λειτουργεί αποθαρρυντικά στην ενσωμάτωση στη διαδικασία των τυχαίων δασών μεγαλύτερων σε ακρίβεια αλγορίθμων. Ένα ακόμα πλεονέκτημα της συγκεκριμένης μεθόδου είναι οι μεγάλες ταχύτητες επεξεργασίας δεδομένων[12].

Τα διαγράμματα ομοιότητας των τυχαίων δασών είναι σε πολλές περιπτώσεις παρόμοια μεταξύ τους, γεγονός που δημιουργεί αμφιβολίες ως προς τη χρησιμότητά τους. Τα συγκεκριμένα έχουν σχήμα αστεριού συνήθως, ένας βραχίονας/ ένα τμήμα ανά κλάση (one arm per class), το οποίο είναι περισσότερο εμφανές όσο καλύτερη είναι η απόδοση της ταξινόμησης (Hastie et al., 2008).

Οι ομοιότητες χρησιμοποιούνται:

- Για οπτικοποίηση του τυχαίου δάσους
- Αντικατάσταση των ελλιπών τιμών
- Αναγνώριση των εσφαλμένα καταχωρημένων δεδομένων, των ακραίων τιμών και των διαφορετικών εγγραφών
- Υπολογισμό των προτύπων

Πρότυπα (Prototypes)

Τα πρότυπα αποτελούν ένα αποτελεσματικό τρόπο πρόχειρης εκτίμησης του τρόπου με τον όποιο τα γνωρίσματα σχετίζονται με τη διαδικασία της ταξινόμησης. Αναλυτικά, για κάθε κλάση γίνεται αναζήτηση ενός μικρού αριθμού εγγραφών οι οποίες αναπαριστούν τη συγκεκριμένη. Οι μικρές και ομοιογενείς κλάσεις μπορούν να αναπαρασταθούν από ένα πρότυπο. Αντιθέτως, οι ετερογενείς κλάσεις αναπαρίστανται από περισσότερα του ενός πρότυπα. Η διαδικασία υπολογισμού του πρώτου πρωτότυπου είναι η ακόλουθη:

- Έστω η κλάση j
- Για κάθε εγγραφή της κλάσης j κάνε εύρεση των κ εγγύτερων γειτόνων (μέσω του γραφήματος ομοιότητας)

[12] https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

- Επίλεξε την εγγραφή η οποία έχει το μεγαλύτερο πλήθος των γειτόνων οι οποίοι ανήκουν στην κλάση j.
- Η μεσαία τιμή των γειτόνων αυτών είναι το πρότυπο και 25η και 75η εκατοστιαία τιμή δίνουν πληροφορίες αναφορικά με τη σταθερότητα
- Με παρόμοιο τρόπο γίνεται αναζήτηση μίας ακόμα εγγραφής (του δεύτερου πρότυπου) η οποία δεν ανήκει στους κ εγγύτερους γείτονες της προηγούμενης εγγραφής και η οποία έχει το μεγαλύτερο πλήθος γειτόνων οι οποίοι ανήκουν στην κλάση j[13]

Στην περίπτωση που τα πρωτότυπα αποτελούν στοιχείο εξόδου στην οθόνη ή σε κάποιο αποθηκευμένο αρχείο η διαδικασία που ακολουθείται διαφοροποιείται για τις συνεχείς και τις ονομαστικές εγγραφές. Αναλυτικά, στην πρώτη περίπτωση γίνεται κανονικοποίηση της τιμής με τον ακόλουθο τρόπο: αφαιρείται η πέμπτη εκατοστιαία τιμή από την τιμή του συνεχούς γνωρίσματος και το αποτέλεσμα που προκύπτει διαιρείται με τη διαφορά της πέμπτης από την ενενηκοστή πέμπτη τιμή. Για τα ονομαστικά χαρακτηριστικά, η τιμή που επιλέγεται ως πρωτότυπο είναι η πιο συχνή [14].

Ελλιπείς τιμές με αντικατάσταση (Missing values with replacemen)t

- Σύνολο εκπαίδευσης (Training Set)

Τα τυχαία δάση έχουν δύο τρόπους μέσω των οποίων διαχειρίζονται τις ελλιπείς τιμές. Ο πρώτος εξ αυτών είναι γρήγορος. Στην περίπτωση που το mth γνώρισμα δεν είναι ονομαστικό (categorical) η μέθοδος υπολογίζει τη διάμεσο όλων των τιμών του συγκεκριμένου γνωρίσματος για την κλάση j. Στην περίπτωση που το εν λόγω χαρακτηριστικό είναι ονομαστικό η τιμή που παίρνει το συγκεκριμένο είναι η πιο συχνή μη ελλιπής τιμή της κλάσης j. Οι συγκεκριμένες τιμές ονομάζονται «γεμίσματα».

Ο δεύτερος τρόπος αντικατάστασης των ελλιπών τιμών επιτυγχάνει καλύτερα αποτελέσματα σε σχέση με τον πρώτο ακόμα και σε περιπτώσεις όπου απουσιάζουν πολλά στοιχεία, ωστόσο εμφανίζει το μειονέκτημα πως είναι υπολογιστικά απαιτητικότερος. Ο συγκεκριμένος αντικαθιστά ελλιπείς τιμές μόνο στο σύνολο εκπαίδευσης. Η διαδικασία ξεκινάει κάνοντας μία πρόχειρη και ανακριβή συμπλήρωση των στοιχείων που απουσιάζουν. Στη συνέχεια, γίνεται εφαρμογή του αλγορίθμου των τυχαίων δασών και υπολογίζονται οι ομοιότητες (proximities).

Στην περίπτωση που η x(m,n) είναι μία συνεχής ελλιπής τιμή, η εκτίμηση της συγκεκριμένης είναι ο σταθμισμένος μέσος όρος των mth γνωρισμάτων, με βάρος τις ομοιότητες ανάμεσα στην nth εγγραφή και εκείνες που δεν εμφανίζουν ελλείψεις σε ό,τι αφορά τις τιμές των γνωρισμάτων τους. Στην περίπτωση ονομαστικής τιμής γίνεται αντικατάσταση με την πλέον συχνή μη ελλιπή τιμή. Σημειώνεται πως η συχνότητα σταθμίζεται με την ομοιότητα[15].

- Σύνολο ελέγχου (Test Set)

Οι Breiman προτείνουν τις ακόλουθες δύο μεθοδολογίες για την περίπτωση που απουσιάζουν τιμές γνωρισμάτων των εγγραφών του συνόλου ελέγχου:

Η επιλογή μίας εξ αυτών εξαρτάται από την ύπαρξη ή μη των κλάσεων (labels) των γνωρισμάτων.

[13]http://www.math.usu.edu/adele/RandomForests/ENAR.pdf
[14] https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.html
[15] https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.html

Στην περίπτωση της ύπαρξης γίνεται χρήση των γεμισμάτων όπως προέκυψαν από το σύνολο εκπαίδευσης. Στην αντίθετη κάθε μία από τις εγγραφές αναπαράγεται τόσες φορές όσες και το πλήθος των κλάσεων. Αναλυτικά, γίνεται η υπόθεση πως το πρώτο αντίγραφο ανήκει στην πρώτη κλάση και βάσει αυτής γίνεται συμπλήρωση των τιμών που απουσιάζουν. Η διαδικασία αυτή επαναλαμβάνεται για όλες τις υπάρχουσες κλάσεις. Το παραπάνω επαυξημένο σύνολο ελέγχου αποτελεί στοιχείο εισόδου κάθε δέντρου. Σε κάθε σύνολο αντιγράφων εκείνο με τις περισσότερες ψήφους προσδιορίζει την κλάση της αρχικής εγγραφής[16].

Λάθος ταξινομημένες εγγραφές (Mislabeled cases)

Σε πολλές περιπτώσεις το σύνολο των δεδομένων εκπαίδευσης ταξινομούνται σε κάποια θεματική κατηγορία με βάση την ανθρώπινη κρίση. Η παραπάνω διαδικασία έχει σε πολλές περιπτώσεις αρνητικά αποτελέσματα καθώς οι εγγραφές αυτές τοποθετούνται σε λάθος κατηγορίες. Οι συγκεκριμένες εντοπίζονται ως ακραίες τιμές (αναλύεται διεξοδικά παρακάτω).

Ακραίες τιμές (Outliers)

Ως ακραίες τιμές ορίζονται οι εγγραφές εκείνες που αφαιρούνται από το κύριο σώμα των δεδομένων, καθώς εμφανίζουν χαμηλή ομοιότητα με τις υπόλοιπες του συνόλου εκπαίδευσης. Μία χρήσιμη αναθεώρηση της παραπάνω πρότασης είναι ο προσδιορισμός των ακραίων τιμών βάσει της κλάσης στην οποία ανήκουν. Συνεπώς, οι λανθασμένες εγγραφές είναι εκείνες οι οποίες εμφανίζουν μικρή ομοιότητα με τις υπόλοιπες που εντάσσονται στην ίδια κλάση.

Η μέση ομοιότητα της εγγραφής n στην κλάση j με το σύνολο των υπόλοιπων που εντάσσονται στην τελευταία υπολογίζεται βάσει της ακόλουθης σχέσης:

$$\bar{P} = \sum_{cl(k)=j} prox^2(n,k)$$

Το μέτρο ακραίας τιμής για την εγγραφή n ορίζεται ως εξής:

$$nsample/\bar{P}$$

Το παραπάνω μέγεθος παίρνει μεγάλες τιμές στην περίπτωση που η μέση ομοιότητα είναι μικρή. Στη συνέχεια υπολογίζεται η μεσαία τιμή των παραπάνω μέτρων για κάθε θεματική κατηγορία, καθώς και της απόκλισης κάθε ενός εξ αυτών από την τελευταία. Το τελικό μέτρο ακραίων τιμών για κάθε μία εγγραφή n υπολογίζεται ως εξής: η μεσαία τιμή των μέτρων αφαιρείται από κάθε ένα εξ αυτών και στη συνέχεια το αποτέλεσμα διαιρείται με την απόλυτη απόκλιση.

Εναλλακτικά, είναι δυνατόν να υπολογιστεί το μέτρο της μη προσαρμοστικότητας κάθε εγγραφής ως εξής:

$$outlyingness\ of\ case\ n = 1/\sum_{cl(k)=j} prox^2(n,k)$$

Στην περίπτωση που το παραπάνω μέτρο παίρνει ασυνήθιστα μεγάλες τιμές (δηλαδή τιμές μεγαλύτερες του 10) τότε η εγγραφή αυτή είναι υποψήφια ακραία τιμή.

[16] https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.html

Μη επιβλεπόμενη ταξινόμηση (Unsupervised learning)

Στη μη επιβλεπόμενη ταξινόμηση τα δεδομένα εισόδου αποτελούνται από ένα σύνολο x-διανυσμάτων ίδιων διαστάσεων τα οποία δεν περιέχουν στοιχεία αναφορικά με τις θεματικές κατηγορίες στις οποίες ανήκουν τα πρώτα. Επιπροσθέτως, δεν υπάρχει κάποιος συντελεστής κέρδους ώστε να είναι δυνατή η βελτιστοποίηση και συνεπώς το πεδίο της κλάσης είναι ανοιχτό σε αυθαίρετα συμπεράσματα. Ο συνήθης στόχος της παρούσας διαδικασίας είναι η ομαδοποίηση των δεδομένων βάσει των τιμών τους σε κατηγορίες όπου κάθε μία από αυτές συνδέεται με μία διαφορετική έννοια.

Η προσέγγιση η οποία υιοθετείται στα τυχαία δάση είναι η ακόλουθη: Αρχικά, όλα τα δεδομένα καταχωρούνται σε μία αυθαίρετη θεματική κατηγορία, έστω την 1. Στη συνέχεια, κατασκευάζεται μία δεύτερη συνθετική κλάση η οποία ορίζεται ως κλάση 2. Η τελευταία δημιουργείται μέσω τυχαίας δειγματοληψίας τιμών από τα αρχικά δεδομένα. Συνεπώς, ένα μέλος της δεύτερης κλάσης δημιουργείται ως εξής: η πρώτη συντεταγμένη παίρνει μία τυχαία τιμή από τις N τιμές {x(1,n)}. Η διαδικασία επαναλαμβάνεται για τις υπόλοιπες συντεταγμένες.

Συνεπώς, η δεύτερη θεματική κατηγορία έχει κατανομή τυχαίων ανεξάρτητων μεταβλητών κάθε μία από τις οποίες έχει την ίδια μονομεταβλητή κατανομή όπως της αντίστοιχης (μεταβλητής) στα αρχικά δεδομένα. Από το παραπάνω προκύπτει πως η δεύτερη κλάση καταργεί την εξαρτημένη δομή των αρχικών δεδομένων. Στην παρούσα περίπτωση υπάρχουν, λοιπόν, δύο κλάσεις, οι οποίες αποτελούν το αντικείμενο ταξινόμησης για τον αλγόριθμο των τυχαίων δασών.

Στην περίπτωση που η αναλογία λανθασμένων ταξινομήσεων oob παίρνει τιμή 40% ή και μεγαλύτερη, το παραπάνω συνεπάγεται πως οι x μεταβλητές φαίνεται να είναι ανεξάρτητες για τα τυχαία δάση. Στην αντίθετη περίπτωση ο διαχωρισμός είναι καλός και συνεπώς είναι δυνατόν να γίνει χρήση όλων των εργαλείων των τυχαίων δασών προκειμένου να γίνει κατανοητή η δομή των δεδομένων.

Τα πλεονεκτήματα της προαναφερθείσας μεθόδου των δύο κλάσεων είναι τα ακόλουθα:

- Οι ελλιπείς τιμές μπορούν να αντικατασταθούν αποτελεσματικά
- Εντοπίζονται εύκολα οι ακραίες τιμές
- Είναι δυνατόν να υπολογιστεί η σημαντικότητα των γνωρισμάτων
- Είναι δυνατόν να γίνει κλιμάκωση (Αν ήταν γνωστή η τιμή της κλάσης για τα αρχικά δεδομένα, η μη επιβλεπόμενη ταξινόμηση συχνά διατηρεί τη δομή της αρχικής κλιμάκωσης)
- Παρέχεται η δυνατότητα ομαδοποίησης (clustering) [17].

Αξιολόγηση του τυχαίου δάσους

Από έρευνες έχει αποδειχθεί πως η αναλογία σφάλματος (error rate) του δάσους εξαρτάται από τα ακόλουθα σημεία:

- Από τη συσχέτιση δύο οποιωνδήποτε δέντρων στο εσωτερικό του δάσους. Πιο συγκεκριμένα, η αύξηση στη συσχέτιση οδηγεί σε αύξηση του ποσοστού σφάλματος στο δάσος.

[17] https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.html

- Από τη δύναμη κάθε δέντρου στο δάσος. Ένα δέντρο με χαμηλό ποσοστό σφάλματος είναι ένας «δυνατός» ταξινομητής. Συνεπώς, η αύξηση στη δύναμη των δέντρων βάση οδηγεί σε μείωση του ποσοστού σφάλματος του δάσους.

Τέλος, διευκρινίζεται πως η μείωση του πλήθους χαρακτηριστικών m επιφέρει τα αντίστοιχα αποτελέσματα τόσο στη συσχέτιση όσο και στη δύναμη των δέντρων. Συνεπώς, στόχος είναι εύρεση του βέλτιστου δυνατού εύρους τιμών του συγκεκριμένου μεγέθους. Η χρήση του σφάλματος oob διευκολύνει τη διαδικασία εύρεσης των παραπάνω ορίων. Το παραπάνω μέγεθος είναι η μοναδική προσαρμοστική (adjustable) παράμετρος, στην οποία εμφανίζει ευαισθησία ο αλγόριθμος των τυχαίων δασών.

Χαρακτηριστικά

Ο Breiman συνοψίζει τα χαρακτηριστικά των τυχαίων δασών ως εξής:

- Ο συγκεκριμένος αλγόριθμος εμφανίζει εντυπωσιακά μεγαλύτερη ακρίβεια σε σχέση με τους υπόλοιπους αλγόριθμος πρόβλεψης
- Μπορεί να εκτελεστεί σε μεγάλες σε μέγεθος βάσεις δεδομένων
- Είναι σε θέση να επεξεργαστεί χιλιάδες γνωρίσματα χωρίς να κάνει κάποια εκκαθάριση
- Είναι σε θέση να προσδιορίσει τα γνωρίσματα εκείνα τα οποία είναι σημαντικά στη διαδικασία της ταξινόμησης
- Τα τυχαία δάση εξάγουν μία εκτίμηση του σφάλματος γενίκευσης κατά τη διαδικασία δημιουργία του τυχαίου δάσους
- Τα συγκεκριμένα παρέχουν έναν αποτελεσματικό τρόπο εκτίμησης γνωρισμάτων που απουσιάζουν από τις εγγραφές των δεδομένων εκπαίδευσης. Παράλληλα είναι σε θέση να διατηρεί την ακρίβεια σε σταθερά επίπεδα στην περίπτωση που ο όγκος των ελλιπών τιμών είναι μεγάλος
- Τα παραγόμενα μοντέλα των τυχαίων δασών μπορούν να χρησιμοποιηθούν μελλοντικά σε άλλα δεδομένα
- Τα πρωτότυπα που υπολογίζονται παρέχουν πληροφορίες για τη σχέση ανάμεσα στα γνωρίσματα των εγγραφών και τη διαδικασία της ταξινόμησης
- Μέσω αυτών υπολογίζεται το μέτρο της ομοιότητας ανάμεσα σε ζεύγη των εγγραφών του δείγματος εκπαίδευσης. Μέσω αυτού είναι δυνατόν να εντοπιστούν οι ακραίες τιμές στα γνωρίσματα των εγγραφών και επιπροσθέτως παρέχεται μία εναλλακτική οπτική των δεδομένων εισόδου (έπειτα από την πολυδιάστατη κλίμακα)
- Το προαναφερθέν χαρακτηριστικό είναι δυνατόν να επεκταθεί σε δεδομένα με άγνωστο το στοιχειό της κλάσης. Συνεπώς, μέσω των τυχαίων δασών είναι δυνατόν να γίνει μη επιβλεπόμενη ταξινόμηση αντικειμένων
- Τα τυχαία δάση παρέχουν μία πειραματική μέθοδο για την ανίχνευση της συσχέτισης / αλληλεπίδρασης ανάμεσα στα γνωρίσματα[18].

2.2.3 Τα τυχαία δάση στην επιστήμη της Ψηφιακής Τηλεπισκόπησης

Η χρήση των τυχαίων δασών στην επιστήμη της Ψηφιακής Τηλεπισκόπησης είναι ευρέως διαδεδομένη τα τελευταία χρόνια. Το παραπάνω οφείλεται στο γεγονός πως ο συγκεκριμένος ταξινομητής δίνει εξαιρετικά αποτελέσματα ως προς την ποιότητα της ταξινόμησης και παράλληλα είναι πολύ γρήγορος. Επιπροσθέτως, ο συγκεκριμένος αλγόριθμος είναι σε θέση να επιλέξει και να κατατάξει τα χαρακτηριστικά εκείνα βάσει των επιτυγχάνεται η καλύτερη δυνατή διάκριση μεταξύ των διαφορετικών κλάσεων- στόχος. Το παραπάνω «προσόν» είναι ιδιαίτερα σημαντικό για την επιστήμη της Ψηφιακής

[18] https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.html

Τηλεπισκόπησης καθώς ο μεγάλος αριθμός διαστάσεων των δορυφορικών δεδομένων καθιστά τη διαδικασία της επιλογής των κατάλληλων γνωρισμάτων ιδιαίτερα χρονοβόρα, επιρρεπή στα σφάλματα και υποκειμενική (Belgiu and Dragut, 2016).

Οι συνδυαστικοί ταξινομητές στην επιστήμη της Ψηφιακής Τηλεπισκόπησης

Οι επιβλεπόμενοι παραμετρικοί ταξινομητές όπως η ταξινόμηση μέγιστης πιθανοφάνειας (maximum likelihood classification) δίνουν εξαιρετικά αποτελέσματα στην περίπτωση ανάλυσης δεδομένων τα οποία ακολουθούν την κανονική κατανομή (unimodal data). Αντιθέτως, οι μη παραμέτρικοί επιβλεπόμενοι ταξινομητές όπως τα δέντρα απόφασης, τα διανύσματα υποστήριξης μηχανής (Support Vector Machines) καθώς και τα τεχνητά νευρωνικά δίκτυα δε διαμορφώνουν κάποια υπόθεση αναφορικά με την κατανομή των δεδομένων. Ως εκ τούτου οι τελευταίοι είναι ιδιαίτερα δημοφιλείς σε διαδικασίες ταξινόμησης τηλεπισκοπικών δεδομένων, καθώς τα συγκεκριμένα σπάνια ακολουθούν την κανονική κατανομή (Belgiu and Dragut, 2016).

Η ανάλυση των δορυφορικών δεδομένων περιορίζεται σε μία διαδικασία εμπειρικής συσχέτισης ανάμεσα στα χωρικά φαινόμενα και στα εμφανιζόμενα πρότυπα (των εικόνων). Το παραπάνω οφείλεται στο γεγονός πως η φύση καθώς και οι αιτίες της χωρικής διακύμανσης δεν είναι πλήρως κατανοητές και για το λόγο αυτό διαμορφώνεται η υπόθεση πως η πραγματικότητα έχει μία σταθερή φασματική απόκριση στις εικόνες. Η συγκεκριμένη, ωστόσο, συχνά παραβιάζεται λόγω της σύνθετης αλληλεπίδρασης παραγόντων όπως η πολυπλοκότητα του τοπίου και η κλίμακα. Βάσει των παραπάνω προκύπτει πως οι επιδόσεις των απλών ταξινομητών δεν επαρκούν σε πολλές περιπτώσεις σε τηλεπισκοπικές εφαρμογές (Belgiu and Dragut, 2016).

Τα τελευταία χρόνια το ενδιαφέρον της κοινότητας της τηλεπισκόπησης έχει στραφεί στους συνδυαστικούς ταξινομητές. Μάλιστα, οι έρευνες έχουν δείξει πως η χρήση των συνδυαστικών ταξινομητών επιτυγχάνει καλύτερα αποτελέσματα συγκριτικά με εκείνα των μεμονωμένων και επιπροσθέτως οι συγκεκριμένοι είναι περισσότερο ανθεκτικοί στο θόρυβο (Belgiu and Dragut, 2016).

Ο (Diettrich, 2000) μέσω έρευνας σε δεδομένα προερχόμενα από διαφορετικού τομείς απέδειξε πως το boosting δίνει ακριβέστερα αποτελέσματα συγκριτικά με το bagging. Ωστόσο, η τεχνική αυτή εμφανίζει πλήθος μειονεκτημάτων όπως

- το υψηλό υπολογιστικό κόστος το οποίο απαιτείται για την εφαρμογή του
- την υπερπροσαρμογή στην περίπτωση που τα δεδομένα εκπαίδευσης δεν επαρκούν καθώς και
- την ευαισθησία τους σε περίπτωση ύπαρξης ακραίων τιμών (outliers) στα δεδομένα εκπαίδευσης (Belgiu and Dragut, 2016)

Παραδείγματα boosting μεθόδων τα οποία χρησιμοποιήθηκαν στην επιστήμη της Ψηφιακής Τηλεπισκόπησης είναι οι AdaBoost (Chan and Paelickx, 2008; Mia et al., 2012) και οι JointBoost (Guo et al., 2015) (Belgiu and Dragut, 2016).

Οι τεχνικές bagging από τη άλλη πλευρά μειώνουν την απόκλιση της ταξινόμησης (δηλαδή το πόσο διαφέρουν τα αποτελέσματα των διαφορετικών μοντέλων) αλλά παράλληλα επηρεάζουν το τυπικό σφάλμα της ταξινόμησης (Belgiu and Dragut, 2016).

Ο ταξινομητής τυχαία δάση

Τα τυχαία δάση είναι ένα είδος συνδυαστικού ταξινομητή ο οποίος αποτελείται από διαφορετικά δέντρα απόφασης. Τα τελευταία δημιουργούνται μέσω τυχαίας επιλογής ενός υποσυνόλου των δειγμάτων εκπαίδευσης. Τα δείγματα τα οποία επιλέγονται επανατοποθετούνται στο σύνολο και ως εκ τούτου έχουν την ίδια πιθανότητα να επιλεγούν ξανά στη διαδικασία της εκπαίδευσης (διαδικασία Bootstrap). Βάσει του παραπάνω προκύπτει πως ορισμένα δείγματα ενδέχεται να επιλεχθούν περισσότερες από μία φορές και ορισμένα να μην επιλεχθούν ποτέ (Belgiu and Dragut, 2016).

Περίπου τα 2/3 των δειγμάτων (τα οποία είναι γνωστά και ως δείγματα εντός της τσάντας- in- bag samples), ενώ το υπόλοιπο 1/3 των δειγμάτων (τα οποία είναι γνωστά ως δείγματα εκτός τσάντας out- of- bag samples) χρησιμοποιούνται από το μοντέλο σαν δεδομένα ελέγχου (Belgiu and Dragut, 2016).

Το σφάλμα το οποίο προκύπτει μέσω των τελευταίων είναι γνωστό ως σφάλμα εκτός σακούλας (out- of- bag error). Κάθε ένα από τα δέντρα απόφασης κατασκευάζεται ανεξάρτητα από τα υπόλοιπα και δεν υφίσταται κάποια διαδικασία κλαδέματος. Κάθε ένας από τους κόμβους διαχωρίζεται χρησιμοποιώντας ένα προσδιοριζόμενο από το χρήστη αριθμό χαρακτηριστικών (Mtry) τα οποία επιλέγονται τυχαία. Ο αλγόριθμος αναπτύσσει το δάσος έως ένα αριθμό δέντρων (Ntree) ο οποίος προσδιορίζεται από το χρήστη), τα οποία έχουν υψηλή απόκλιση. Η τελική απόφαση σχετικά με την ταξινόμηση λαμβάνεται υπολογίζοντας το μέσο όρο των πιθανοτήτων εκχώρησης κάθε αντικειμένου στην εκάστοτε κλάση για όλα τα δέντρα (Belgiu and Dragut, 2016).

Οι παράμετροι που προσδιορίζονται από το χρήστη για την κατασκευή του τυχαίου δάσους είναι οι ακόλουθες:

- το πλήθος των δέντρων απόφασης (Ntree)
- το πλήθος των παραμέτρων που θα επιλεχθούν για την κατασκευή του εκάστοτε κόμβου

Τόσο θεωρητικές όσο και εμπειρικές έρευνες έχουν δείξει πως ο ταξινομητής τυχαία δάση είναι περισσότερο ευαίσθητος στον αριθμό των δέντρων σε σχέση με το πλήθος των παραμέτρων. Εφόσον ο αλγόριθμός τυχαία δάση είναι υπολογιστικά αποδοτικός (efficient) και δεν υπερπροσαρμόζεται το πλήθος των δέντρων απόφασης είναι δυνατόν να πάρει μεγάλες τιμές. Η πλειοψηφία των ερευνών ορίζει την εν λόγω παράμετρο σε 500 καθώς τα σφάλματα σταθεροποιούνται προτού το πλήθος των δέντρων φτάσει τη συγκεκριμένη τιμή. Ορισμένες έρευνες έχουν χρησιμοποιήσει διαφορετικές τιμές όπως 5000, 1000 και 100. Μία μελέτη η οποία είναι αφιερωμένη σε ταξινόμηση polarimetric synthetic aperture radar δεδομένων (Du et al., 2015) εξήγαγε το συμπέρασμα πως η παράμετρος του πλήθους των δέντρων δεν επηρέασε το τελικό αποτέλεσμα. Παρόμοια ήταν τα αποτελέσματα στην έρευνα που υλοποιήθηκε από τους (Topouzelis and Psyllos, 2012) οι οποίοι χρησιμοποιήσαν τον αλγόριθμο των τυχαίων δασών για την ανίχνευση πετρελαιοκηλίδων από SAR δεδομένα. Οι συγκεκριμένοι κατέληξαν στο συμπέρασμα πως η αύξηση του πλήθους των δέντρων πέρα από 70 δεν επηρέαζε το αποτέλεσμα της ταξινόμησης. Βάσει των συμπερασμάτων που έχουν προκύψει από το πλήθος των ερευνών έως τώρα οι προτείνουν τη ρύθμιση του πλήθους των δέντρων απόφασης σε 500 στην περίπτωση αξιοποίησης του εν λόγω αλγορίθμου σε δορυφορικά δεδομένα (Belgiu and Dragut, 2016).

Η παράμετρος Mtry παίρνει συνήθως τιμή ίση με την τετραγωνική ρίζα του πλήθους των μεταβλητών εισόδου. Οι (Gosh et al., 2014) όρισαν την τιμή της συγκεκριμένης παραμέτρου ίση με το συνολικό πλήθος των μεταβλητών. Η προσέγγιση αυτή αυξάνει τον υπολογιστικό χρόνο καθώς ο αλγόριθμος πρέπει να υπολογίσει το πληροφοριακό κέρδος βάσει όλων των μεταβλητών προκειμένου να γίνει ο διαχωρισμός των κόμβων (Belgiu and Dragut, 2016).

Τα τελευταία χρόνια υπάρχει αυξανόμενο ενδιαφέρον σε ό,τι αφορά πρόσθετες συναρτήσεις των τυχαίων δασών. Οι (Belgiu et al., 2014b, Corcoran et al., 2013, Pedergnana et al., 2013) χρησιμοποίησαν τη σημαντικότητα μεταβλητών προκειμένου να κάνουν βελτιστοποίηση του χώρου των μεταβλητών. Οι (Peerbhay et al., 2015) μέτρησαν τη συσχέτιση ανάμεσα σε μεγάλα σε μέγεθος δεδομένα στη βάση (on the basis of) των μετρήσεων πίνακα εσωτερικών συσχετίσεων (internal proximities matrix measurements). Τέλος, οι (Corcoran et al., 2013) αναγνώρισαν τις ακραίες τιμές στα δεδομένα εκπαίδευσης μέσω της διερευνητικής ανάλυσης (explorative analysis) της ομοιότητας των δειγμάτων (Belgiu and Dragut, 2016).

Η ομοιότητα μεταβλητών είναι δυνατόν να υπολογιστεί εσωτερικά μέσω πολλών διαφορετικών τρόπων όπως μέσω της μέση μείωσης του δείκτη GINI (mean decrease in GINI) ή μέσω της μέσης μείωσης στην ακρίβεια (mean decrease in accuracy).Μέσω του πρώτου μεγέθους υπολογίζεται κατά πόσον μία μεταβλητή μειώνει το μέτρο καθαρότητας σε μία συγκεκριμένη κλάση. Από την άλλη πλευρά η μέση μείωση της ακρίβειας λαμβάνει υπόψη τη διαφορά ανάμεσα στο σφάλμα OOB το οποίο προκύπτει από ένα σύνολο δεδομένων τα οποία έχουν προκύψει μέσω τυχαίων ανταλλαγών στις τιμές των διαφορετικών μεταβλητών από το σφάλμα OOB το οποίο προέκυψε από το αρχικό σύνολο δεδομένων. Η πλειοψηφία των ερευνών οι οποίες μελετήθηκαν από τους (Belgiu and Dragut, 2016) χρησιμοποίησαν τη μέση μείωση στην ακρίβεια προκειμένου να υπολογίσουν τη σημαντικότητα των μεταβλητών (Belgiu and Dragut, 2016).

Ο υπολογιστικός χρόνος ο οποίος απαιτείται για την εφαρμογή του μοντέλου ταξινόμησης των τυχαίων δασών υπολογίζεται βάσει της σχέσης:

$$T\sqrt{MNlog(N)}$$

Όπου Τ ο αριθμός των δέντρων, Μ ο αριθμός των μεταβλητών που χρησιμοποιούνται σε κάθε κόμβο και Ν το πλήθος των δεδομένων εκπαίδευσης.

Μέχρι στιγμής ο αλγόριθμος των τυχαίων δασών έχει εφαρμοστεί σε πλήθος διαφορετικών λογισμικών όπως το eCognition, το R software, Weka, το πακέτο skit- learn, το imager, το Ranger (το οποίο είναι γνωστό και ως Τυχαία Ζούγκλα - Random Jungle), το STATISTICA, το Willows και το Matlab (Belgiu and Dragut, 2016).

Τα τυχαία δάση έχουν εφαρμοστεί ευρέως στην επιστήμη της Ψηφιακής Τηλεπισκόπησης. Ενδεικτικά, αναφέρονται τα ακόλουθα επιστημονικά άρθρα:

- Οι (Bosch et al., 2007) στο άρθρο τους «Image Classification using Random Forests and Ferns» εξετάζουν το πρόβλημα της ταξινόμησης εικόνων βάσει κατηγοριών αντικειμένων. Για το σκοπό αυτό, γίνεται χρήση των ακόλουθων στοιχείων: αυτόματη επιλογή των περιοχών ενδιαφέροντος προκειμένου να γίνει εκπαίδευση του αλγορίθμου, χρήση του αλγορίθμου των τυχαίων δασών. Το πλεονέκτημα του εν λόγω αλγορίθμου είναι η ευκολία στη διαδικασία της εκπαίδευσης και του

ελέγχου. Ο συγκεκριμένος αλγόριθμος εφαρμόστηκε σε δεδομένα Caltech-101 και Caltech-256.

- Οι (Chehata, 2009) στο άρθρο τους "Airborne lidar Feature Selection for urban classification using random Forests" χρησιμοποίησαν τα τυχαία δάση σε δεδομένα lidar. Οι συγκεκριμένοι αναφέρουν πως ο αλγόριθμος αυτός παρέχει ακριβή αποτελέσματα και παράλληλα είναι σε θέση να επεξεργαστεί μεγάλα σε όγκο δεδομένα σε μικρό χρονικό διάστημα. Η ακρίβεια ταξινόμησης των αστικών περιοχών φτάνει το 94,35%
- Οι (Rodriguez- Galiano et al., 2011) στο άρθρο τους «An assessment of effectiveness of a random forest classifier for land- cover classification» αναφέρουν πως τα τυχαία δάση αποτελούν ένα ισχυρό ταξινομητή, ο οποίος ωστόσο δεν έχει εφαρμοστεί ευρέως σε θέματα εδαφοκάλυψης. Μάλιστα επισημαίνουν πως ο συγκριμένος αλγόριθμος έχει πλεονεκτήματα όπως: τη μη παραμετρική φύση τους, την υψηλή τους ακρίβεια, τη δυνατότητα να παρέχουν στοιχεία σχετικά με τη σημαντικότητα των μεταβλητών, την ικανότητά τους να υπολογίζουν ελλιπείς τιμές και τέλος την προσαρμοστικότητα τους διαφορετικούς τύπους δεδομένων, καθώς είναι σε θέση να κάνουν ταξινόμηση, παλινδρόμηση, ανάλυση επιβίωσης καθώς και μη επιβλεπόμενη ταξινόμηση. Ωστόσο, αναφέρουν πως τα κριτήρια διαχωρισμού στη διαδικασία της ταξινόμησης είναι άγνωστα και συνεπώς τα τυχαία δάση αντιμετωπίζονται σαν μαύρο κουτί. Ο αλγόριθμος των τυχαίων δασών εφαρμόστηκε σε τηλεπισκοπικά δεδομένα προερχόμενα από το δορυφόρο Landsat-5 από την περίοδο άνοιξη καλοκαίρι και τα συγκεκριμένα χρησιμοποιήθηκαν προκειμένου να γίνει ταξινόμηση 14 διαφορετικών ειδών εδαφοκάλυψης. Τα αποτελέσματα του υπό μελέτη αλγορίθμου έδωσαν ακρίβεια που φτάνει το 92%.

Το πλήθος των δημοσιεύσεων που αφορούν στην εφαρμογή των τυχαίων δασών συνδυαστικά με αντικειμενοστραφή ανάλυση εικόνων είναι περιορισμένο. Το παραπάνω οφείλεται στο γεγονός πως ο συγκεκριμένος αλγόριθμος είναι σχετικά καινούργιος και ως εκ τούτου δεν έχει διερευνηθεί επαρκώς η συνεισφορά του στην αντικειμενοστραφή ταξινόμηση τηλεπισκοπικών δεδομένων.

Οι (Strumpf and Kerle , 2011) στο άρθρο τους "Object- oriented mapping of landlslides using Random Forests" χρησιμοποίησαν δεδομένα υψηλής ευκρίνειας των δορυφόρων Geoeye-1, IKONOS, Quickbird καθώς και μία τρικάναλη αεροφωτογραφία προκειμένου να αναγνωρίσουν σε αυτές περιοχές με κατολισθήσεις. Συμπληρωματικά έγινε χρήση δεδομένων DEM. Μέσω της πολυκλιμακωτής κατάτμησης δημιούργησαν στο περιβάλλον του eCognition 15 διαφορετικά επίπεδα αντικειμένων. Ιδιαίτερο ενδιαφέρον παρουσιάζει το γεγονός πως η διαδικασία δημιουργίας των αντικειμένων βασίστηκε αποκλειστικά σε φασματικά κριτήρια καθώς στην παράμετρο του σχήματος δόθηκε βάρος 0. Στη συνέχεια έγινε υπολογισμός των χαρακτηριστικών (φασματικών και υφής) των αντικειμένων βάσει των οποίων θα γίνει ο διαχωρισμός των τελευταίων στις κλάσεις «κατολισθήσεις» (O_{LS}) και «λοιπές περιοχές» (O_{NLS}). Ακολούθως, έγινε εκπαίδευση και εφαρμογή των τυχαίων δασών στα δεδομένα εισόδου. Οι Strumpf A., Kerle N δεν περιορίστηκαν στην απλή εφαρμογή του συγκεκριμένου αλγορίθμου ταξινόμησης στις εικόνες αλλά εστίασαν σε δύο βασικά ζητήματα. Αρχικά έγινε αναζήτηση των γνωρισμάτων των αντικειμένων τα οποία είναι πράγματι χρήσιμα για το διαχωρισμό των τελευταίων στις δύο θεματικές κατηγορίες. Η διαδικασία αυτή υλοποιήθηκε βάσει της μεθόδου που προτείνουν οι (Diaz- Uriate and Alvarez de Andres, 2006), η οποία θα αναλυθεί διεξοδικά στην ενότητα 3.12. Παράλληλα έγινε μία προσπάθεια διερεύνησης της επιρροής της κλίμακας στη διαδικασία επιλογής των

γνωρισμάτων. Στη συνέχεια οι έρευνες εστίασαν στη διαχείριση του φαινομένου της «ανισορροπίας των κλάσεων» (class imbalance). Πιο συγκεκριμένα, στις περιοχές μελέτης οι κατολισθήσεις καλύπτουν μικρό μέρος των ευρύτερων πριοχών. Το παραπάνω οδηγεί σε μία ανισορροπία ανάμεσα στις δύο θεματικές κατηγορίες (O_{NLS} και O_{LS}) και ενδεχομένως προκαλεί μία «προτίμηση» του μοντέλου ταξινόμησης στις μη πληγείσες περιοχές. Στη βιβλιογραφία αναφέρεται πλήθος μεθόδων διαχείρισης τους συγκεκριμένου ζητήματός, καμία ωστοσο από αυτές δε δίνει λύση σε όλες περιπτώσεις. Στα πλαίσια του συγκεκριμένου άρθρου έγινε εφαρμογή της ακόλουθης διαδικασίας: Αρχικά, έγινε διαχωρισμός των δεδομένων εκπαίδευσης σε εκπαίδευσης Tr_{20} (το 20% των δεδομένων) και ελέγχου Te_{80} (το 80% των δεδομένων) και στη συνέχεια, το σύνολο Tr_{20} διασπάστηκε επαναληπτικά σε επιμέρους υπόσύνολα εκπαίδευσης ($train_{sub}$) και ελέγχου ($test_{sub}$). Στόχος της διαδικασίας αυτής ήταν ο προσδιορισμός της τιμής της παραμέτρου β_i η οποία οδηγεί σε ισορροπία των τιμών της ακρίβειας και της ορθότητας των αποτελεσμάτων της ταξινόμησης. Ως β_i συμβολίζεται η αναλογία του πλήθους των αντικειμένων O_{NLS} προς O_{LS} σε κάθε ένα από τα διαφορετικά $train_{sub}$. Η διαδικασία ξεκίνησε με τιμή $\beta_i = 1$ και σε κάθε επανάληψη αυξήθηκε κατά 0,1. Για κάθε τιμή β_i δημιουργήθηκαν δέκα διαφορετικά υποσύνολα $train_{sub}$, $test_{sub}$ από το Tr_{20} και βάσει αυτών εκπαιδεύτηκε και αξιολογήθηκε διαφορετικό μοντέλο των τυχαίων δασών και βάσει των δέκα αυτών μοντέλων υπολογίστηκε ένα μέσο σφάλμα και η τυπική απόκλιση . Βάσει των τιμών β_i. δημιουργήθηκαν διαφορετικά υποσυνολά των Tr_{20} , κατασκευάστηκαν μοντέλα των τυχαίων δασών τα οποία και αξιολογήθηκαν μέσω του Te_{80} . Τα μονέλα αυτά εφαρμόστηκαν στα διαφορετικά επίπεδα κλίμακας και με τον τροπο αυτό διερευνήθηκε η επιρροή της κατάτμησης στα ποσοστά της ποιότητας της ταξινόμησης.Η ακρίβεια των θεματικών χαρτών που προέκυψαν κυμαίνονται από 73% με 87%.

Οι (Smith A.et al., 2012) στο άρθρο τους “Updating the national wetland inventory in Minnesota by intergrating air photo-interpretation, object- oriented image analysis and mulitisourve data fusion” ενημέρωσαν την υπάρχουσα βάση δεδομένων της Minnesota που αφορά στην καταγραφή των υγρότοπων της περιοχής. Για το σκοπό αυτό χρησιμοποίησαν αεροφωτογραφίες SURDEX (4 κανάλια) υψηλής χωρικής ανάλυσης (της τάξης των 50 cm), δεδομένα RADAR, δεδομένα DEM καθώς και δεδομένα επιτόπιου ελέγχου. Αρχικά, τα δεδομένα υπόκεινται σε διαδικασία κατάτμησης στο eCognition αξιοποιώντας αρχικά τον αλγορίθμο του τετραδικού δέντρου (quadtree) και στη συνέχεια στα τμήματα που προκύπτουν την πολυκλιμακωτή κατάτμηση. Ακολούθως έγινε ταξινόμηση των αντικειμένων σε θεματικές κατηγορίες βάσει των φασματικών, των DEM καθώς και των RADAR δεδομένων. Οι ακρίβειες της ταξινόμησης έφτασε το 92,2% σε ό,τι αφορά την αναγνώριση των υγρότοπων και το 66,87% στην περίπτωση της ταξινόμσηης των τελευταίων σε κατηγορίες. Τα αποτελέσματα της ταξινόμησης μέσω του αλγορίθμου των τυχαίων δασών δεν ήταν τα παραδοτέα αλλά χρησιμοποιήθηκαν βοηθητικά στη διαδικασία της φωτοερμηνείας των περιοχών από έμπειρους φωτοερμηνευτές.

Οι (Belgiu and Dragut, 2014) στο άρθρο τους ‘Comparing supervised and unsupervised multiresolution segmentation approaches for extracting buildings form very high resolution images” διερεύνησαν τις υπάρχουσες επιβλεπόμενες και μη μεθόδους ρύθμισης των παραμέτρων της πολυκλιμακωτής κατάτμησης. Όπως αναφέρουν η διαδικασία της κατάτμησης αποτελεί άλυτο πρόβλημα στην αντικειμενοστραφή ανάλυση εικόνων και τα αποτελέσματα της επηρεάζουν την ποιότητα της ταξινόμησης. Για το λόγο αυτό η σωστή ρύθμιση των παραμέτρων είναι κρίσιμο θέμα για την κοινότητα της Ψηφιακής Τηλεπισκόπησης. Στα πλαίσια της παρούσας εφαρμογής έγινε διερεύνηση μίας

επιβλεπόμενης και δύο μη επιβλεπόμενων μεθόδων ρύθμισης παραμέτρων για τον πλέον διαδεδομένο αλγόριθμο κατάτμησης εκείνον της πολυκλιμακωτής. Για το σκοπό αυτό, μελετήθηκαν δεδομένα υψηλής ευκρίνειας προερχόμενα από τους δορυφόρους Quickbird και Worlview-2 της πόλης Salzburg. Σε κάθε περίπτωση δημιουργήθηκε ένα επίπεδο κατάτμησης. Η αξιολόγηση των αποτελεσμάτων των παραπάνω διαδικασιών έγινε μέσω της ταξινόμησης των αντικειμένων που προέκυψαν σε θεματικές κατηγορίες με τον αλγόριθμο των τυχαίων δασών. Αναλυτικά, το μοντέλο που προέκυψε ταξινόμησε τα κτίρια της περιοχής που προέκυψαν σε έξι θεματικές κατηγορίες αξιοποιώντας φασματικά, γεωμετρικά καθώς και χαρακτηριστικά υφής. Εκτύπωση προκαλεί πως ο αλγόριθμος των τυχαίων δασών έδωσε πολύ παρόμοιους θεματικούς χάρτες παρά το γεγονός πως οι μέθοδοι κατάτμησης έδωσαν διαφορετικά αποτελέσματα. Η ακρίβεια της ταξινόμησης φτάνει το 86%.

Οι (Rougier and Puissant, 2014) στο άρθρο τους “improvements of urban vegetation segmentation and classification using multi- temporal Pleiades Images” διερεύνησαν την αποτελεσματικότητα του αλγορίθμου ταξινόμησης τυχαία δάση στην ανίχνευση περιοχών αστικής βλάστησης. Για το σκοπό αυτό χρησιμοποίησαν τρία δεδομένα υψηλής ευκρίνειας προερχόμενα από τους δορυφόρους Peiades, τα οποία λήφθηκαν τον Αύγουστο και το Σεπτέμβριο του έτους 2012 καθώς και τον Απρίλιο του 2013. . Πρώτο στάδιο της διαδικασίας ήταν αυτό της κατάτμησης. Οι αλγόριθμοι που εφαρμόστηκαν ήταν η μεσαία μετατόπιση (mean shift) καθώς και η πολύ κλιμακωτή κατάτμηση. Για κάθε ένα από τους παραπάνω αλγορίθμους κατασκευάστηκαν 25 διαφορετικά επίπεδα. Τα χαρακτηριστικά τα οποία εξήχθησαν από τα παραγόμενα αντικείμενα είναι φασματικά, σχήματος, υφής καθώς και διαχρονικά (multitemporal). Στη συνέχεια έγινε εκπαίδευση του αλγορίθμου των τυχαίων δασών βάσει των παραπάνω χαρακτηριστικών και παράλληλα έγινε αναγνώριση των χρησιμότερων ων εξ αυτών βάσει της μεθόδου των (Diaz- Uriate and Alvarez de Andres, 2006). Οι κλάσεις στις οποίες τοποθετήθηκαν τα αντικείμενα της βλάστησης είναι οι «Γρασίδι» και «Δέντρα» Τα F-measures της διαδικασίας αυτής φτάνουν το 0,875 για τη θεματική κατηγορία των δέντρων και το 0,618 για εκείνη του γρασιδιού.

Οι (Puissant et al., 2014) στο άρθρο τους “Object- oriented mapping of urban trees using Random Forest classifiers” αναγνώρισαν τις περιοχές αστικής βλάστησης στην πόλη του Στρασβούργου. Τα δεδομένα που μελετήθηκαν είναι υψηλής ευκρίνειας και προέρχονται από τους δορυφόρους Quickbird. Αρχικά, έγινε κατάκτηση της εικόνας εισόδου στο περιβάλλον του eCognition μέσω του αλγορίθμου της πολυκλιμακωτής κατάτμησης Multiresolution Segmentation και εν συνεχεία εκείνου της φασματικής διαφοράς. Για κάθε ένα από τα αντικείμενα που δημιουργήθηκαν εξήχθησαν 100 χαρακτηριστικά, τα οποία μπορούν ομαδοποιηθούν σε φασματικά, γεωμετρικά και υφής. Εν συνεχεία έγινε εκπαίδευση και εφαρμογή του αλγορίθμου των τυχαίων δασών ώστε να αναγνωριστούν οι περιοχές βλάστησης. Προκείμενου να γίνει μείωση του υπολογιστικού φόρτου κάθε ένα από τα δέντρα απόφασης του τυχαίου δάσους κατασκευάστηκε μέσω bootstrap δείγματος στο οποίο περιλαμβάνεται μόλις το 5% των δεδομένων εκπαίδευσης. Οι (Puissant et al., 2014) διαχειρίζονται το πρόβλημα της ανισορροπίας των κλάσεων με τρόπο παρόμοιο- πλην μικρών εξαιρέσεων- με εκείνο των (Strumpf and Kerle , 2011). Η διαδικασία επιλογής των χαρακτηριστικών είναι εκείνη που προτείνουν οι (Diaz- Uriate and Alvarez de Andres, 2006). Τέλος, η συγκεκριμένη έρευνα εστιάζει στην εύρεση της βέλτιστης τιμής της παραμέτρου πλήθος δειγμάτων ανά κόμβο. Οι (Puissant et al., 2014) αναφέρουν πως η μείωση της τιμής της τελευταίας οδηγεί σε λιγότερο συσχετισμένα δέντρα. Οι (Gislason et al., 2006) βρήκαν εμπειρικά πως ο μικρός αριθμός χαρακτηριστικών ανά κόμβο σε

συνδυασμό με ένα μεγάλο πλήθος δέντρων οδηγεί σε υψηλά επίπεδα ακρίβειας στα τηλεπισκοπικά δεδομένα. Για το λόγο αυτό στα πλαίσια της συγκεκριμένης δημοσίευσης έγινε πειραματισμός σε ό,τι αφορά τον αριθμότων χαρακτηριστικών ξεκινώντας από την προτεινόμενη τιμή (τετραγωνική ρίζα του συνολικού πλήθους των γνωρισμάτων) και συνέχεια μειώνοντας την.Το Fmeasure της παραπάνω διαδικασίας φτάνει το 64,84%.

Τέλος, οι (Juel et al., 2015) στο άρθρο τους "Spatial application of Random Forest models for fine- scale coastal vegetation classification using object- based analysis of aerial orthophoto and DEM data"εστίασαν τις προσπάθειες τους στην αναγνώριση βιότοπων. Η περιοχή μελέτης στην προκειμένη περίπτωση είναι η ακτογραμμή της Δανίας μήκους 7300km. Για τις ανάγκες των ερευνών έγινε χρήση μωσαϊκού ορθοφωτογραφιών για δύο περιόδους. Οι εικόνες λήφθηκαν μέσω ψηφιακών φωτογραφικών μηχανών Vexcel Utracam, διαφορετικής χωρικής και φασματικής ανάλυσης. Παράλληλα αξιοποιήθηκαν DEM δεδομένα ποικίλων χωρικών αναλύσεων. Η διαδικασία επεξεργασίας των δεδομένων ήταν η ακόλουθη: Αρχικά, έγινε εισαγωγή των δεδομένων στο λογισμικό του eCognition όπου και έγινε επεξεργασία των εικόνων μέσω της πολυκλιμακωτής κατάτμησης και στη συνέχεια εκείνης της φασματικής διαφοράς. Για κάθε ένα από τα παραγόμενα αντικείμενα έγινε υπολογισμός 409 χαρακτηριστικών τα οποία προήλθαν από το μωσαϊκό των ορθοφωτογραφιών, τα υψομετρικά δεδομένα καθώς και τα μεταδεδομένα. Εν συνεχεία έγινε εφαρμογή του αλγορίθμου τω ν τυχαίων δασών προκειμένου να γίνει ταξινόμηση των πρωτογενών αντικειμένων σε θεματικές κατηγορίες. Τα μοντέλα που προέκυψαν αξιολογήθηκαν στα δύο επίπεδα της κατάτμησης βάσει τριών σεναρίων: Κατασκευή και αξιολόγηση του μοντέλου χωρίς να γίνει χωρικός διαχωρισμός των δεδομένων εκπαίδευσης και αξιολόγησης, κατασκευή και αξιολόγηση σε χωρικά διαχωρισμένα δεδομένα εκπαίδευσης και αξιολόγησης στο εσωτερικό των περιοχών εκπαίδευσης, κατασκευή και αξιολόγηση του μοντέλου σε χωρικά διαχωρισμένα δεδομένα εκπαίδευσης και αξιολόγησης διαφορετικών περιοχών. Για τις ανάγκες της συγκεκριμένης μελέτης οι (Juel et al., 2015) προτείνουν τη δημιουργία μεγάλων σε μέγεθος δέντρων και αξιοποίηση μεγάλου πλήθους χαρκαητριστικών. Σε ό,τι αφορά τα διαφορετικά σεναρία κατέληξαν πως ο διαχωρισμός των δεδομένων εκπαίδευσης και αξιολόγησης οδήγησε σε μείωση της ακρίβειας των παραγόμενων μοντέλων. Ανλαυτικά στην περίπτωση που δε γίνει διαχωρισμός των δεδομένων η συνολική ακρίβεια φτάνει το 92,1%.

3 Μεθοδολογία των αλγορίθμων «δέντρα απόφασης» και «τυχαία δάση» στο περιβάλλον του eCognition

3.1 Εισαγωγικά στοιχεία

Στόχος της παρούσας εργασίας είναι η αντικειμενοστρεφής ταξινόμηση δορυφορικών εικόνων υψηλής χωρικής ανάλυσης μέσω των αλγορίθμων «δέντρα απόφασης» και «τυχαία δάση». Αναλυτικά, γίνεται διερεύνηση της αποτελεσματικότητας των τεχνικών αυτών σε ό,τι αφορά την ανίχνευση κτιρίων. Η διαδικασία αυτή διαρθρώνεται στα ακόλουθα στάδια:

- Προεπεξεργασία εικόνας εισόδου: Αποκοπή τμήματος της αρχικής εικόνας εισόδου και εφαρμογή αμφίπλευρου φίλτρου.
- Κατάτμηση: Δημιουργία αντικειμένων στο επεξεργασμένο τμήμα της εικόνας εισόδου
- Ταξινόμηση: Κατηγοριοποίηση των αντικειμένων σε κλάσεις

Στα ακόλουθα κεφάλαια περιγράφεται αναλυτικά η διαδικασία εφαρμογής των υπό εξέταση αλγορίθμων στα δεδομένα εισόδου.

3.2 Δορυφορική Εικόνα

Η δορυφορική εικόνα που αναλύθηκε μέσω της παρούσας εργασίας προέρχεται από τους δορυφόρους Pléiades και απεικονίζει τμήμα της πόλης Commerce City της πολιτείας Colorado των Ηνωμένων Πολιτειών (Εικόνα 3.1). Αποτελείται από 4 κανάλια και το χρωματικό βάθος της είναι 8 bit.

ΕΙΚΟΝΑ 3.1: ΔΟΡΥΦΟΡΙΚΗ ΕΙΚΟΝΑ ΤΗΣ ΠΟΛΗΣ COMMERCE CITY ΣΤΗΝ ΕΙΚΟΝΑ ΣΗΜΕΙΩΝΟΝΤΑΙ ΟΙ ΔΥΟ ΠΕΡΙΟΧΕΣ ΣΤΙΣ ΟΠΟΙΕΣ ΘΑ ΓΙΝΕΙ ΕΦΑΡΜΟΓΗ ΤΩΝ ΥΠΟ ΜΕΛΕΤΗ ΑΛΓΟΡΙΘΜΩΝ ΤΑΞΙΝΟΜΗΣΗΣ

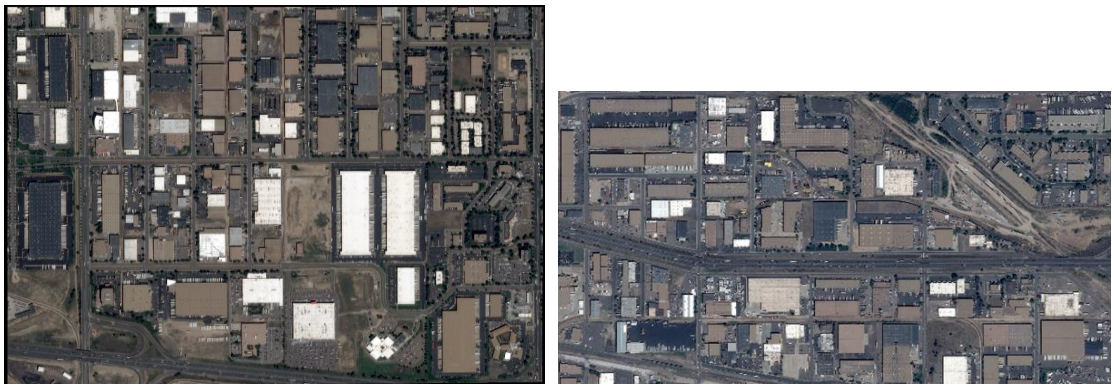
Η χωρική ανάλυση της δορυφορικής εικόνας είναι 0,5m και το μέγεθος της 14029x14012 εικονοστοιχεία.

3.3 Προεπεξεργασία εικόνας εισόδου

3.3.1 Αποκοπή των δύο τμημάτων

Αρχικά, έγινε η εισαγωγή της εικόνας εισόδου στο περιβάλλον του ECognition προκειμένου να γίνει μία πρώτη προσπάθεια κατάτμησης αυτής. Η εκτιμώμενη από το λογισμικό χρονική διάρκεια εφαρμογής της διαδικασίας αυτής στα δεδομένα εισόδου έφτανε τη μισή ώρα και για το λόγο αυτό η διεργασία ματαιώθηκε. Το παραπάνω οφείλεται στο γεγονός πως τα υπό μελέτη δεδομένα εισόδου είναι μεγάλου μεγέθους και ως εκ τούτου η επεξεργασία αυτών απαιτεί αρκετό χρόνο.

Για το λόγο αυτό έγινε εισαγωγή της εικόνας στο περιβάλλον του προγράμματος ελεύθερου λογισμικού QGIS. Μέσω του εργαλείου Raster> Extraction> Clipper έγινε αποκοπή και αποθήκευση του τμήματος που εμφανίζεται στην Εικόνα 3.2.



ΕΙΚΟΝΑ 3.2: ΑΡΙΣΤΕΡΑ: 1ΟΤΜΗΜΑ ΤΗΣ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ ΔΕΞΙΑ: 2^{Ο} ΤΜΗΜΑ ΤΗΣ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ

3.3.2 Φιλτράρισμα της εικόνας εισόδου

Εν συνεχεία, έγινε εισαγωγή του τμήματος της εικόνας στο λογισμικό του eCogntition και έγινε μία πρώτη προσπάθεια κατάτμησης αυτής. Στην Εικόνα 3.3 εμφανίζεται το αποτέλεσμα εφαρμογής του αλγορίθμου της πολυκλιμακωτής κατάτμησης στα δεδομένα εισόδου. Είναι εμφανές πως το παραγόμενο αποτέλεσμα δεν είναι ικανοποιητικό, καθώς πολλά από τα εμφανιζόμενα κτίρια αποδόθηκαν από τον αλγόριθμο μέσω περισσότερων του ενός αντικείμενου (σημειώνονται με κόκκινο). Το παραπάνω οφείλεται στο γεγονός πως η χωρική ανάλυση της υπό μελέτη εικόνας είναι ιδιαίτερα υψηλή και συνεπώς εμφανίζονται σε αυτήν πολλές ανεπιθύμητες λεπτομέρειες. Οι τελευταίες αποτελούν εμπόδιο στη διαδικασία της κατάτμησης καθώς ο αλγόριθμος αντιλαμβάνεται τις περιοχές αυτές σαν ξεχωριστά αντικείμενα.

ΕΙΚΟΝΑ 3.3: ΕΦΑΡΜΟΓΗ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΗΣ ΠΟΛΥΚΛΙΜΑΚΩΤΗΣ ΚΑΤΑΤΜΗΣΗΣ ΣΤΗΝ ΕΙΚΟΝΑ ΕΙΣΟΔΟΥ (ΚΛΙΜΑΚΑ (SCALE): 150, ΣΧΗΜΑ (SHAPE): 0.4, ΣΥΜΠΑΓΟΤΗΤΑ (COMPACTNESS): 0.3)

Για το λόγο αυτό, στην εικόνα εισόδου εμφανίστηκε χαμηλοπερατό φίλτρο με στόχο την ομαλοποίηση των περιοχών υψηλής συχνότητας. Για τις ανάγκες της παρούσας μελέτης επιλέχθηκε το αμφίπλευρο φίλτρο (bilateral) το οποίο εξομαλύνει την εικόνα εισόδου διατηρώντας παράλληλα τις εμφανιζόμενες ακμές. Το συγκεκριμένο αποτελεί ένα μη γραμμικό χαμηλοπερατό φίλτρο, το οποίο παρουσιάστηκε για πρώτη φορά από τους Aurich and Weule το έτος 1995 με το όνομα "μη γραμμικό φίλτρο Gauss" (nonlinear Gaussian filter). Ακολούθως, ασχολήθηκαν με αυτό οι Tomasi και Manduchi το έτος 1998, οι οποίοι του έδωσαν και το όνομα αμφίπλευρο φίλτρο (Bilateral filter). Η έξοδος του φίλτρου σε κάθε εικονοστοιχείο είναι ένας σταθμισμένος μέσος όρος των γειτονικών σε αυτό εικονοστοιχείων. Το βάρος που αποδίδεται σε κάθε «γείτονα» εξαρτάται αφενός από την απόσταση του από το κεντρικό εικονοστοιχείο, αφετέρου από τη διαφορά της έντασης του γκρίζου τόνου του από την αντίστοιχη του κεντρικού. Συγκεκριμένα, το αμφίπλευρο φίλτρο ορίζεται ως εξής:

$$I_p^b = \frac{1}{W_b^b} \sum_{q \in S} G_{\sigma_S}(\|p-q\|)\, G_{\sigma_r}(|I_p - I_q|) I_q$$

Όπου: $W_b^b = \sum_{q \in S} G_{\sigma_S}(\|p-q\|)\, G_{\sigma_r}(|I_p - I_q|)$

Στην παραπάνω σχέση:

- Η παράμετρος σ_Sρυθμίζει το μέγεθος του πυρήνα του φίλτρου.
- Η παράμετρος σ_rρυθμίζει τα βάρη σε ό,τι αφορά τη διαφορά έντασης των εικονοστοιχείων
- Οι μεταβλητές p,q συμβολίζουν τις θέσεις του κεντρικού και του γειτονικού εικονοστοιχείου αντίστοιχα
- Οι μεταβλητές I_p, I_qσυμβολίζουν την ένταση του γκρίζου τόνου στη θέση του κεντρικού και του γειτονικού εικονοστοιχείου αντίστοιχα (Paris and Durand, 2009)

Το παρόν φίλτρο εφαρμόστηκε στη γλώσσα προγραμματισμού Python μέσω των ακόλουθων εντολών:

```
#Προσθήκη βιβλιοθηκων
import numpy as np
from numpy import array
import scipy as sp
from scipy import ndimage

#Προσδιορισμος εικονας-εισόδου
```

```
file1 = 'clipper.tif'

#Η εικόνα είναι γεωαναφερμενη συνεπώς κάνω χρηση των συναρτησεων
geoimread και geoimwrite
execfile('geoimread.py')
execfile('geoimwrite.py')

#Καταχώρηση των στοιχείων της εικόνας στα img, geoTransform, proj, drv_name
(img, geoTransform, proj, drv_name) = geoimread(file1)

#Αποθήκευση των καναλιών της εικόνας σε 4 πίνακες
img_0=img[:, :, 0]
img_1=img[:, :, 1]
img_2=img[:, :, 2]
img_3=img[:, :, 3]

#Εφαρμογή του αμφίπλευρου φίλτρου σε κάθε ένα από τα κανάλια της εικόνας
from skimage.restoration import denoise_tv_chambolle, denoise_bilateral
img_0_bil=denoise_bilateral(img_0, sigma_range=0.05, sigma_spatial=15)
img_1_bil=denoise_bilateral(img_1, sigma_range=0.05, sigma_spatial=15)
img_2_bil=denoise_bilateral(img_2, sigma_range=0.05, sigma_spatial=15)
img_3_bil=denoise_bilateral(img_3, sigma_range=0.05, sigma_spatial=15)

#Αποθήκευση των καναλιών όπως πρόεκυψαν μετά την εφαρμογή του
αμφίπλευρου φίλτρου
sp.misc.imsave('img_0_bil.jpeg',img_0_bil)
sp.misc.imsave('img_1_bil.jpeg',img_1_bil)
sp.misc.imsave('img_2_bil.jpeg',img_2_bil)
sp.misc.imsave('img_3_bil.jpeg',img_3_bil)

img_0_bil1=denoise_bilateral(img_0, sigma_range=0.5, sigma_spatial=20)
img_1_bil1=denoise_bilateral(img_1, sigma_range=0.5, sigma_spatial=20)
img_2_bil1=denoise_bilateral(img_2, sigma_range=0.5, sigma_spatial=20)
img_3_bil1=denoise_bilateral(img_3, sigma_range=0.5, sigma_spatial=20)

sp.misc.imsave('img_0_bil1.jpeg',img_0_bil1)
sp.misc.imsave('img_1_bil1.jpeg',img_1_bil1)
sp.misc.imsave('img_2_bil1.jpeg',img_2_bil1)
sp.misc.imsave('img_3_bil1.jpeg',img_3_bil1)

img_0_bil2=denoise_bilateral(img_0, sigma_range=0.5, sigma_spatial=30)
img_1_bil2=denoise_bilateral(img_1, sigma_range=0.5, sigma_spatial=30)
img_2_bil2=denoise_bilateral(img_2, sigma_range=0.5, sigma_spatial=30)
img_3_bil2=denoise_bilateral(img_3, sigma_range=0.5, sigma_spatial=30)

sp.misc.imsave('img_0_bil2.jpeg',img_0_bil2)
sp.misc.imsave('img_1_bil2.jpeg',img_1_bil2)
sp.misc.imsave('img_2_bil2.jpeg',img_2_bil2)
sp.misc.imsave('img_3_bil2.jpeg',img_3_bil2)
```

Η εφαρμογή του συγκεκριμένου φίλτρου στην εικόνα εισόδου απαιτεί από το χρήστη τη ρύθμιση των ακόλουθων παραμέτρων:

- sigma_range (χωρική διακύμανση)
- sigma_spatial (φασματική διακύμανση)

Για τις ανάγκες της παρούσας εφαρμογής έγινε εφαρμογή του αμφίπλευρου φίλτρου στην εικόνα εισόδου για διαφορετικούς συνδυασμούς τιμών των παραπάνω παραμέτρων. Ο συνδυασμός ο οποίος επιλέχθηκε είναι ο:

- sigma_range=0,5
- sigma_spatial=30

Στην ακόλουθη εικόνα (Εικόνα 3.4) εμφανίζεται το αποτέλεσμα της εφαρμογής του αμφίπλερου φίλτρου στα δύο τμήματα της εικόνας εισόδου

ΕΙΚΟΝΑ 3.4: ΕΦΑΡΜΟΓΗ ΤΟΥ ΑΜΦΙΠΛΕΥΡΟΥ ΦΙΛΤΡΟΥ ΣΤΑ ΔΥΟ ΤΜΗΜΑΤΑ ΤΗΣ ΕΙΚΟΝΑΣ ΕΙΣΟΔΟΥ

3.4 Κατάτμηση πρώτου τμήματος της εικόνας εισόδου

3.4.1 Κατάτμηση πολλαπλής ανάλυσης (Multiresolution Segmentation)

Στόχος της παρούσας διαδικασίας είναι η δημιουργία πρωτογενών αντικειμένων (dummy objects) τα οποία μέσω της διαδικασίας της ταξινόμησης θα μετατραπούν σε σημασιολογικά αντικείμενα. Τα όρια των πρωτογενών αντικειμένων παίζουν καθοριστικό ρόλο στην ποιότητα του τελικού αποτελέσματος.

Η μέθοδος κατάτμησης, η οποία ενσωματώνεται στο πρόγραμμα eCognition είναι εκείνη της πολλαπλής ανάλυσης (Multiresolution Segmentation). Η συγκεκριμένη υλοποιείται εφόσον οριστούν αρχικά από το χρήστη οι ακόλουθες παράμετροι:

- Κλίμακα: Μέσω της συγκεκριμένης ορίζεται ο βαθμός ετερογένειας των εικονοστοιχείων που συνθέτουν τα παραγόμενα αντικείμενα. Συνεπώς, η αύξηση της τιμής της κλίμακας οδηγεί σε αύξηση του μεγέθους των πρωτογενών αντικειμένων. Σημειώνεται πως ο προσδιορισμός της τιμής της κλίμακας είναι άμεσα εξαρτημένος από τη χωρική ανάλυση της υπό μελέτη εικόνας.
- Κριτήριο της ομοιογένειας: Η συγκεκριμένη παράμετρος καθορίζεται από:
  - ✓ Το χρώμα: μέσω του συγκεκριμένου κριτηρίου δίνεται βαρύτητα στη φασματική ομοιότητα των εικονοστοιχείων
  - ✓ Το σχήμα: μέσω του συγκεκριμένου κριτηρίου δίνεται έμφαση στο σχήμα των παραγόμενων αντικειμένων. Συγκεκριμένα, ορίζεται κατά πόσον τα αντικείμενα τα οποία θα προκύψουν θα έχουν ομαλά όρια (smoothness) ή αν θα προσεγγίζουν κανονικά σχήματα (compactness).

Στα πλαίσια της παρούσας εφαρμογής έγιναν δοκιμές με διάφορές τιμές των παραμέτρων. Ο συνδυασμός εκείνος ο οποίος έδωσε τα βέλτιστα δυνατά αποτελέσματα είναι ο ακόλουθος:

- Κλίμακα (Scale): 150
- Σχήμα (Shape): 0.4
- Συμπαγότητα (Compactness): 0.3 (Εικόνα 3.5)

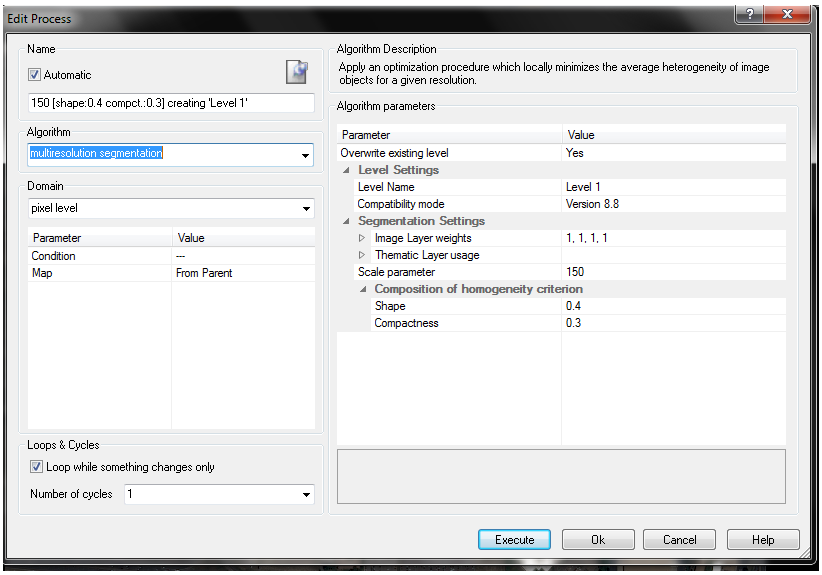
Παρατηρείται πως η τιμή της παραμέτρου της κλίμακας είναι μεγάλη και το παραπάνω συνδέεται με το γεγονός πως η υπό μελέτη δορυφορική εικόνα είναι ιδιαίτερα ευκρινής.

Επιπροσθέτως, στη συγκεκριμένη κατάτμηση δόθηκε μία σχετικά μεγάλη τιμή στην παράμετρο του σχήματος καθώς μέσω της παρούσας εφαρμογής επιδιώκεται όπως επισημάνθηκε προηγουμένως η αυτόματη ανίχνευση κτιρίων. Τα αντικείμενα της συγκεκριμένης θεματικής κατηγορία έχουν χαρακτηριστικό σχήμα το οποίο προσεγγίζει σε μεγάλο βαθμό ένα κανονικό σχήμα, δηλαδή ένα ορθογώνιο παραλληλόγραμμο. Για το λόγο αυτό δόθηκε βάρος 0,4 στην παράμετρο του σχήματος και 0,3 σε εκείνη της προσέγγισης της συπαγότητας. Τέλος, αξίζει να σημειωθεί πως τα κτίρια έχουν ομαλά όρια και ως εκ τούτου δόθηκε βάρος 0,7 στο εν λόγω κριτήριο.

Δεδομένου ότι η περιοχή μελέτης περιέχει μεγάλα σε μεγέθος ομοιόμορφα κτίρια, τα οποία εμφανίζουν έντονες αντιθέσεις στις φασματικές τιμές σε σχέση με τα γειτονικά αντικείμενα τα αποτελέσματα της κατάτμησης είναι ικανοποιητικά σε ό,τι αφορά την ανίχνευση κτιρίων παρόλο που δε δόθηκε υπερβολικά μεγάλη τιμή στις παραμέτρους του σχήματος και της συμπαγότητας.

Τέλος, παρατηρείται ότι το αποτέλεσμα της κατάτμησης δεν είναι τόσο ικανοποιητικό σε ό,τι αφορά τη θεματική κατηγορία των δρόμων καθώς ορισμένοι δρόμοι περιλαμβάνουν και τμήματα μη δρόμων, το οποίο οφείλεται στο γεγονός πως στην παρούσα εφαρμογή έγινε ένα επίπεδο κατάτμησης. Ωστόσο, μέσω της συγκεκριμένης εργασίας δίνεται έμφαση στα κτίρια και για το λόγο αυτό δε δόθηκε βάρος στην ποιότητα της κατάτμησης που αφορά σε αντικείμενα των δρόμων.

Στην Εικόνα 3.6 εμφανίζεται το πρώτο επίπεδο κατάτμησης (Level 1). Το αποτέλεσμα της παρούσας διαδικασίας είναι ικανοποιητικό καθώς τα όρια των πρωτογενών αντικειμένων ταυτίζονται σε μεγάλο βαθμό με εκείνα των πραγματικών. Ειδικότερα σε ό,τι αφορά τα κτίρια το αποτέλεσμα της κατάτμησης σχεδόν ταυτίζεται με το επιθυμητό. Είναι φανερό πως το συγκεκριμένο είναι εμφανώς καλύτερο σε σχέση με εκείνο της Εικόνα 3.3. Προβλήματα, ωστόσο εξακολουθούν να εντοπίζονται στα μεγάλα σε μέγεθος εμφανιζόμενα κτίρια τα οποία στο συγκεκριμένο επίπεδο αναπαρίστανται από δύο ή τρία πρωτογενή αντικείμενα.



ΕΙΚΟΝΑ 3.5: ΡΥΘΜΙΣΕΙΣ ΠΑΡΑΜΕΤΡΩΝ ΠΡΩΤΟΥ ΕΠΙΠΕΔΟΥ

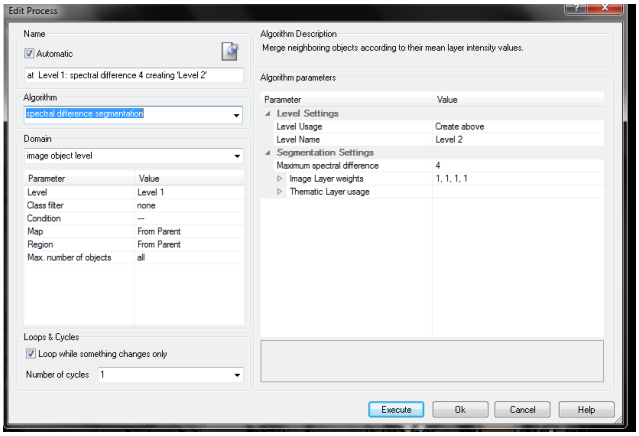
ΕΙΚΟΝΑ 3.6: ΠΡΩΤΟ ΕΠΙΠΕΔΟ ΚΑΤΑΤΜΗΣΗΣ

3.4.2 Κατάτμηση φασματικής διαφοράς (Spectral difference Segmentation)

Ο αλγόριθμος της φασματικής διαφοράς συγχωνεύει γειτονικά αντικείμενα της εικόνας με κριτήριο την ομοιότητα των μέσων φασματικών τους τιμών. Αρχικά, ορίζεται από το χρήστη ένα ανώτερο κατώφλι φασματικής διαφοράς. Τα γειτονικά αντικείμενα της εικόνας συγχωνεύονται στις περιπτώσεις όπου η τιμή του συγκεκριμένου μεγέθους είναι μικρότερη από το παραπάνω όριο. Ο παρών αλγόριθμος έχει σχεδιαστεί με στόχο να βελτιώνει τα αποτελέσματα μίας υπάρχουσας κατάτμησης και συνεπώς δεν είναι σε θέση να κατασκευάσει αντικείμενα από το επίπεδο του εικονοστοιχείου (Trimble, eCognition Developer, 2016).

Ο συγκεκριμένος αλγόριθμος ενσωματώθηκε στην παρούσα εφαρμογή προκειμένου να γίνει βελτίωση του αποτελέσματος της προαναφερθείσας διαδικασίας. Το κατώφλι της μέγιστης φασματικής διαφοράς(maximum spectral difference) ορίστηκε έπειτα από δοκιμές ίσο με 4. Επιπροσθέτως, έγινε ρύθμιση ώστε το αποτέλεσμα του συγκεκριμένου αλγορίθμου να αποθηκευτεί σε νέο επίπεδο (Level 2) το οποίο δημιουργήθηκε πάνω(Create above) από το επίπεδο 1 (Level 1) (Εικόνα 3.7).

Στην Εικόνα 3.8 εμφανίζεται το αποτέλεσμα εφαρμογής του αλγορίθμου της φασματικής διαφοράς στο πρώτο επίπεδο κατάτμησης. Είναι εμφανές πως τα αντικείμενα του δεύτερου επιπέδου ταυτίζονται σε μεγαλύτερο βαθμό με τα πραγματικά συγκριτικά με εκείνα του πρώτου. Αναλυτικά τα μεγάλα σε μέγεθος κτίρια αναπαρίστανται στο επίπεδο 2 από ένα αντικείμενο. Η βελτίωση του αποτελέσματος είναι αισθητή και σε ό,τι αφορά τα τμήματα του οδικού δικτύου στο οποίο παρατηρείται πως ο αριθμός των αντικειμένων της συγκεκριμένης θεματικής κατηγορίας στο νέο επίπεδο έχει μειωθεί.



ΕΙΚΟΝΑ 3.7: ΡΥΘΜΙΣΕΙΣ ΚΑΤΑΤΜΗΣΗΣ ΔΕΥΤΕΡΟΥ ΕΠΙΠΕΔΟΥ

ΕΙΚΟΝΑ 3.8: ΔΕΥΤΕΡΟ ΕΠΙΠΕΔΟ ΚΑΤΑΤΜΗΣΗΣ

3.5 Ταξινόμηση αντικειμένων

Στόχος στο παρόν στάδιο της εφαρμογής είναι η ταξινόμηση των αντικειμένων σε θεματικές κατηγορίες. Για το σκοπό αυτό, έγινε αρχικά φωτοερμηνεία της εικόνας, προκειμένου να αναγνωριστούν σε αυτήν οι υπάρχουσες θεματικές κατηγορίες. Βάσει αυτών δημιουργήθηκαν οι κλάσεις στις οποίες θα γίνει η ταξινόμηση των πρωτογενών αντικειμένων. Συγκεκριμένα, δομήθηκαν οι ακόλουθες κατηγορίες:

- Κτίρια (Buildings)
- Δρόμοι (Roads)
- Αστική βλάστηση (Urban Vegetation)
- Άγονο Έδαφος (Barren)
- Χώρος στάθμευσης (Parking)

ΕΙΚΟΝΑ 3.9: ΘΕΜΑΤΙΚΕΣ ΚΑΤΗΓΟΡΙΕΣ

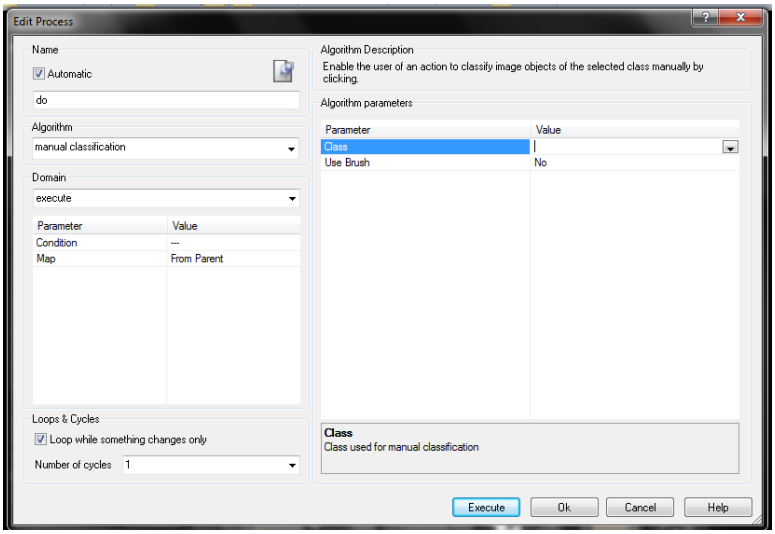
Τόσο τα δέντρα απόφασης όσο και τα τυχαία δάση αποτελούν αλγορίθμους επιβλεπόμενης ταξινόμησης. Τα βασικά στάδια τα οποία περιλαμβάνονται σε μία τυπική διαδικασία επιβλεπόμενης ταξινόμησης είναι τα ακόλουθα:

- Το στάδιο της επίβλεψης (εκπαίδευσης) στο οποίο ο Φωτοερμηνευτής αναγνωρίζει αντιπροσωπευτικές περιοχές εκπαίδευσης. Βάσει αυτών αναπτύσσεται μία αριθμητική περιγραφή των ιδιοτήτων κάθε θεματικής κατηγορίας κάλυψης γης.
- Το στάδιο της ταξινόμησης. Μέσω του συγκεκριμένου τα αντικείμενα της εικόνας ταξινομούνται βάσει των ιδιοτήτων τους στην κλάση εκείνη στην οποία μοιάζει. (Αργιαλάς, 1998)

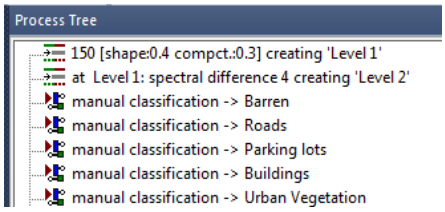
«Ο γενικός στόχος του σταδίου της επίβλεψης είναι η συλλογή ενός συνόλου στατιστικών στοιχείων τα οποία περιγράφουν τα πρότυπα φασματικής απόκρισης για την κάθε θεματική κατηγορία που θα ταξινομηθεί στην εικόνα.» (Αργιαλάς, 1998). Η ποιότητα των αποτελεσμάτων της ταξινόμησης συνδέεται άμεσα με την ποιότητα των δεδομένων εκπαίδευσης. Αναλυτικά, τα δείγματα θα πρέπει να είναι αντιπροσωπευτικά και πλήρη. Το παραπάνω συνεπάγεται πως τα δεδομένα εκπαίδευσης πρέπει να καλύπτουν όλες τις φασματικές κατηγορίες που συνιστούν την κάθε θεματική κατηγορία (Αργιαλάς, 1998).

Η διαδικασία εκπαίδευσης των υπό μελέτη αλγορίθμων ταξινόμησης στο ECognition είναι κοινή έως ένα σημείο και πρακτικά υλοποιείται ως εξής:

- Μέσω χειροκίνητης ταξινόμησης (manual classification) γίνεται καταχώρηση των δειγμάτων στις θεματικές κατηγορίες. Αναλυτικά, στο πεδίο Class επιλέγεται η αντίστοιχη θεματική κατηγορία (Εικόνα 3.10). Η συγκεκριμένη διαδικασία επαναλαμβάνεται για όλες τις κλάσεις της εικόνας (Εικόνα 3.9).

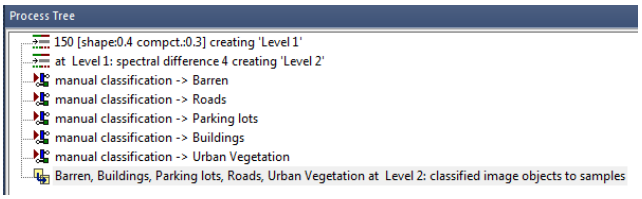


ΕΙΚΟΝΑ 3.10: ΠΑΡΑΘΥΡΟ ΔΙΑΛΟΓΟΥ ΓΙΑ ΧΕΙΡΟΚΙΝΗΤΗ ΤΑΞΙΝΟΜΗΣΗ

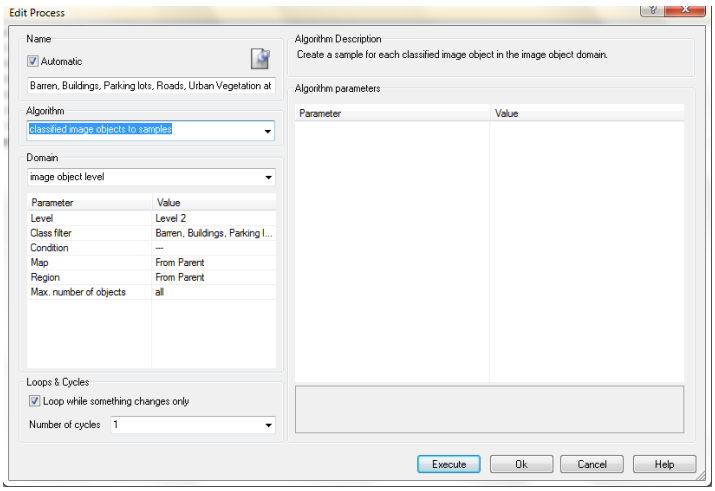


ΕΙΚΟΝΑ 3.11: ΔΕΝΤΡΟ ΔΙΑΔΙΚΑΣΙΩΝ (PROCESS TREE) ΓΙΑ ΤΗ ΧΕΙΡΟΚΙΝΗΤΗ ΤΑΞΙΝΟΜΗΣΗ

- Στη συνέχεια μέσω της εντολής Classified image objects to samples γίνεται καταχώρηση των ταξινομημένων αντικειμένων στα δείγματα (Εικόνα 3.12 και Εικόνα 3.13).

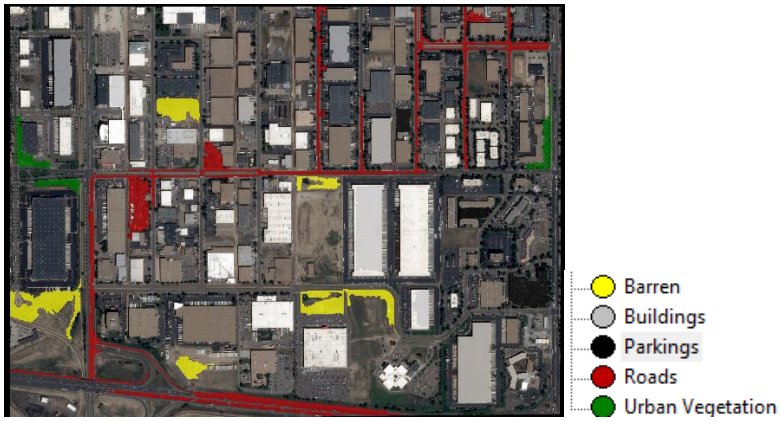


ΕΙΚΟΝΑ 3.12: ΔΕΝΤΡΟ ΔΙΑΔΙΚΑΣΙΩΝ (PROCESS TREE) ΕΩΣ ΚΑΙ ΤΗΝ ΕΝΤΟΛΗ CLASSIFIED IMAGE OBJECTS TO SAMPLES



ΕΙΚΟΝΑ 3.13: ΠΑΡΑΘΥΡΟ ΔΙΑΛΟΓΟΥ ΓΙΑ ΤΗΝ ΕΝΤΟΛΗ CLASSIFIED IMAGE OBJECTS TO SAMPLES

Στην Εικόνα 3.14 εμφανίζονται τα δεδομένα εκπαίδευσης των αλγορίθμων ταξινόμησης. Τα δείγματα επιλέχθηκαν ώστε να είναι επαρκή και αντιπροσωπευτικά των θεματικών κατηγοριών.



ΕΙΚΟΝΑ 3.14: ΔΕΙΓΜΑΤΑ

3.5.1 Εφαρμογή αλγορίθμου δέντρων απόφασης

Η εφαρμογή του αλγορίθμου των δέντρων απόφασης υλοποιήθηκε στο περιβάλλον του ECognition σε δύο στάδια:

- Εκπαίδευση του αλγορίθμου
- Εφαρμογή του αλγορίθμου

Εκπαίδευση του αλγορίθμου

Η εκπαίδευση του αλγορίθμου των δέντρων απόφασης έγινε μέσω της επιλογής Advanced Classification>Classifier των διεργασιών του ECognition. Στο παράθυρο διαλόγου επιλέγεται το επίπεδο των αντικειμένων (image object level) στο πεδίο τομέας (domain) (Εικόνα 3.15). Στις παραμέτρους του συγκεκριμένου πεδίου ορίζεται το επίπεδο στο οποίο θα γίνει η ταξινόμηση (Level 2) καθώς και οι υποψήφιες κλάσεις στις οποίες θα ταξινομηθούν τα αντικείμενα (Εικόνα 3.16).

Στη συνέχεια γίνεται ρύθμιση των παραμέτρων του αλγορίθμου. Πιο συγκεκριμένα, στο πεδίο διεργασία (Operation) επιλέχθηκε Εκπαίδευση (Train) και ορίστηκε μέσω των παραμέτρων του συστήματος (Configuration) η αλφαριθμητική μεταβλητή (string) στην οποία θα αποθηκευτεί ο «εκπαιδευμένος» αλγόριθμος (Εικόνα 3.17).Τα δέντρα απόφασης εκπαιδεύτηκαν αποκλειστικά βάσει των δειγμάτων που δόθηκαν από το χρήστη (Use samples only : Yes) (Εικόνα 3.15).

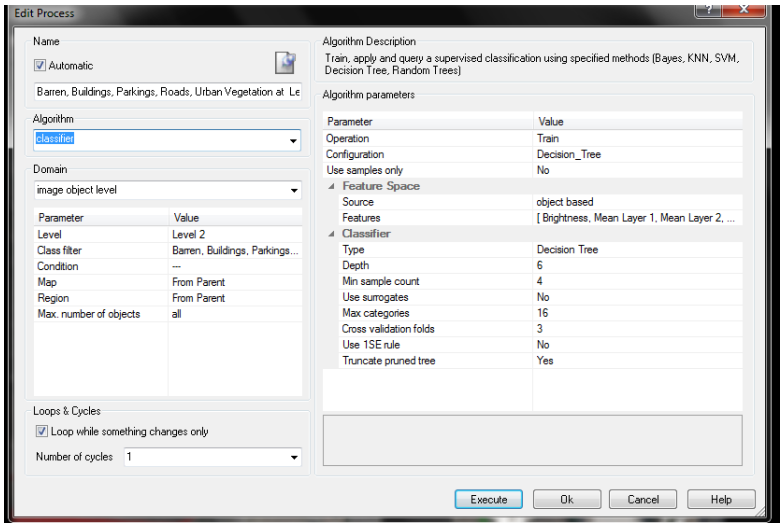
Ακολούθως έγινε ρύθμιση των χαρακτηριστικών βάσει των οποίων θα γίνει η εκπαίδευση των αλγορίθμων με σειρά προτεραιότητα. Στόχος της συγκεκριμένης διαδικασίας είναι κατά κύριο λόγο η ανίχνευση κτιρίων. Για το λόγο αυτό δόθηκε βαρύτητα στο κόκκινο κανάλι καθώς στο συγκεκριμένο τα ανθρωπογενή αντικείμενα εμφανίζουν υψηλές τιμές ανακλαστικότητας (Εικόνα 3.19). Σημειώνεται επιπροσθέτως πως μέσω της επιλογής Προσαρμοσμένα Χαρακτηριστικά (Customized Features) κατασκευάστηκε ο δείκτης βλάστησης βάσει του οποίου διευκολύνεται η διάκριση των αντικειμένων του αστικού πράσινου από τις υπόλοιπες θεματικές κατηγορίες (Εικόνα 3.18).

Έμφαση δόθηκε επίσης στο σχήμα και πιο συγκεκριμένα ορίστηκαν τα χαρακτηριστικά:

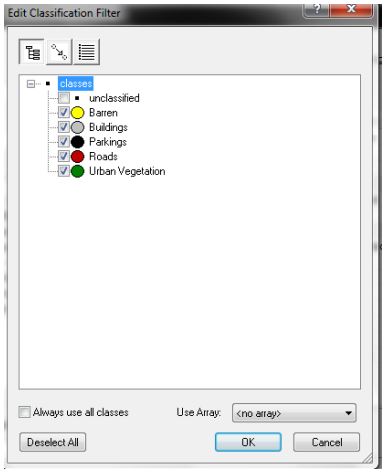
- Λόγος μήκους προς πλάτος του εκάστοτε αντικειμένου (Length/Width)
- Συμπαγότητα του αντικειμένου (Compactness)
- Ομοιότητα του σχήματος του αντικειμένου με το σχήμα του ορθογωνίου(Rectangular fit)
- Εμβαδόν αντικειμένου (Area)(Εικόνα 3.19).

Τέλος στο πεδίο Ταξινομητής ορίστηκε το είδος (Type) του ταξινομητή, τα δέντρα απόφασης (Decision Tree) εν προκειμένω καθώς και ένα σύνολο παραμέτρων τα οποία είναι διαφορετικά για κάθε αλγόριθμο. Συγκεκριμένα, σε ό,τι αφορά τα Δέντρα απόφασης ο χρήστης είχε τη δυνατότητα να ορίσει τις ακόλουθες μεταβλητές:

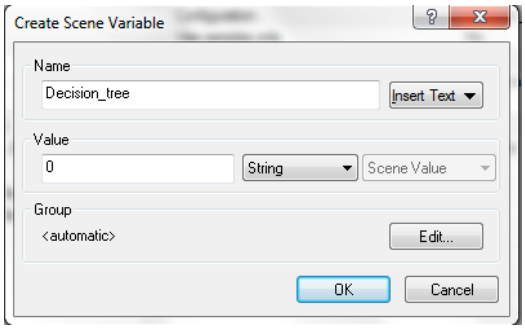
- Βάθος δέντρου (Depth): Μέγιστο βάθος δέντρου (προεπιλεγμένη τιμή: 0)
- Ελάχιστος αριθμός δειγμάτων (Min sample count): Ο ελάχιστος αριθμός δειγμάτων σε κάθε κόμβο (προεπιλεγμένη τιμή: 0)
- Χρήση αντικαταστατών (Use surrogates): Χρήση αντικαταστατών για την ελλιπείς τιμές. Στην περίπτωση που επιλέγεται η τιμή Ναι (yes) κατασκευάζονται κόμβοι αντικαταστατών, οι οποίοι δημιουργούνται προκειμένου να είναι δυνατή η διαχείριση των ελλιπών τιμών.
- Μέγιστος αριθμός κατηγοριών (Max categories): Ταξινόμηση των πιθανών τιμών μίας ονομαστικής τιμής σε K (μικρότερος αριθμός από εκείνον του μέγιστου αριθμού κατηγοριών) κατηγορίες (προεπιλεγμένη τιμή: 16)
- Cross Validation folds: Ο αριθμός των Cross Validation που υλοποιούνται (προεπιλεγμένη τιμή: 3)
- Use 1 SE rule: Χρήση ενός κανόνα SE για το κλάδεμα του δέντρου (προεπιλεγμένη τιμή: No)
- Αφαίρεση των κλαδεμένων κλαδιών (Truncate pruned trees): Στην περίπτωση που επιλεχθεί η τιμή Ναι (Yes) τα κλαδεμένα κλαδιά αφαιρούνται φυσικά από το δέντρο (προεπιλεγμένη τιμή: Yes) (Trimble, eCognition Developer, 2016)(Εικόνα 3.15).



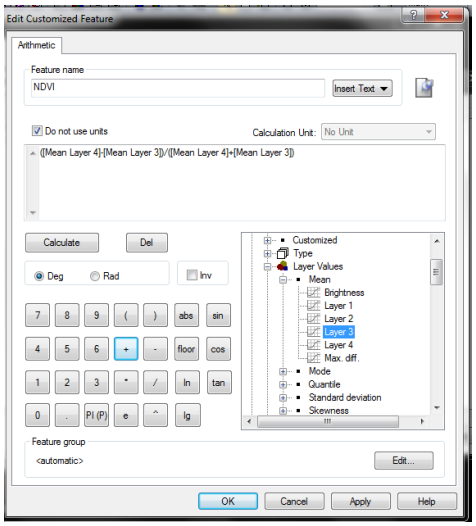
ΕΙΚΟΝΑ 3.15: ΠΑΡΑΘΥΡΟ ΔΙΑΛΟΓΟΥ ΓΙΑ ΤΗΝ ΕΚΠΑΙΔΕΥΣΗ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ



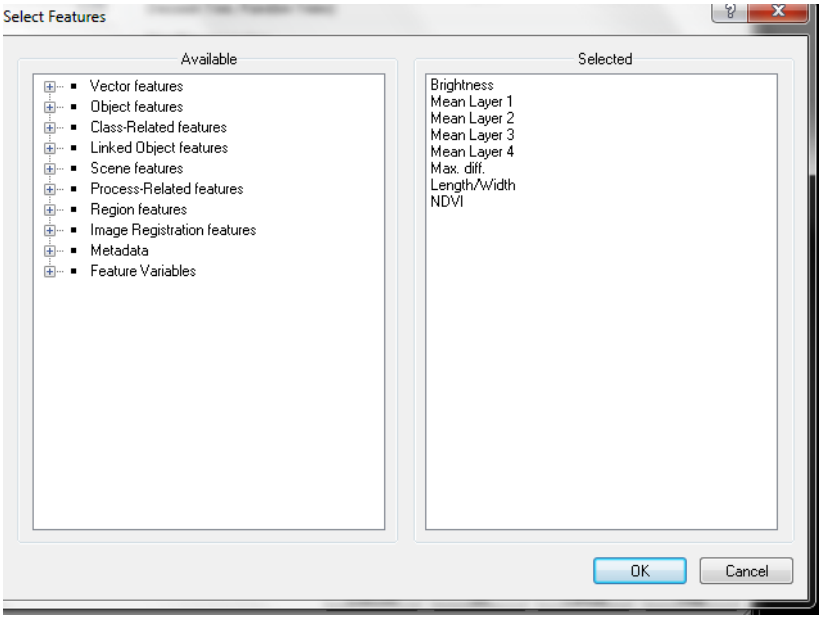
ΕΙΚΟΝΑ 3.16: ΕΠΙΛΟΓΗ ΤΩΝ ΚΛΑΣΕΩΝ ΣΤΙΣ ΟΠΟΙΕΣ ΘΑ ΕΦΑΡΜΟΣΤΕΙ Ο ΑΛΓΟΡΙΘΜΟΣ



ΕΙΚΟΝΑ 3.17: ΔΗΜΙΟΥΡΓΙΑ ΜΕΤΑΒΛΗΤΗΣ ΣΤΗΝ ΟΠΟΙΑ ΘΑ ΑΠΟΘΗΚΕΥΤΕΙ Ο ΕΚΠΑΙΔΕΥΜΕΝΟΣ ΑΛΓΟΡΙΘΜΟΣ



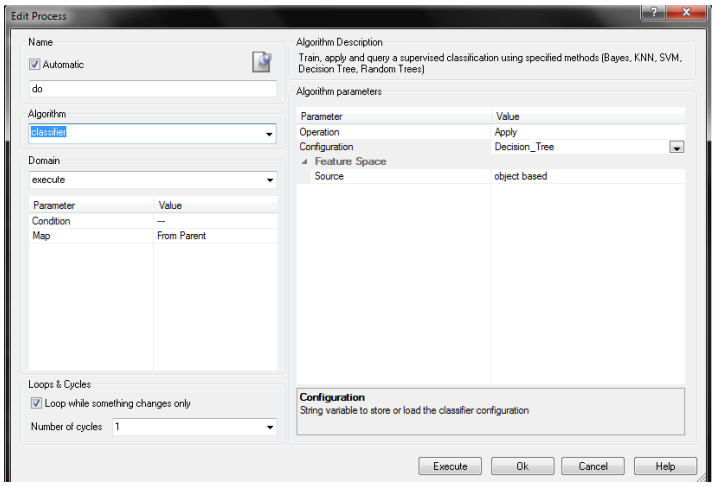
ΕΙΚΟΝΑ 3.18: ΚΑΝΟΝΙΚΟΠΟΙΗΜΕΝΟΣ ΔΕΙΚΤΗΣ ΒΛΑΣΤΗΣΗΣ



ΕΙΚΟΝΑ 3.19: ΛΙΣΤΑ ΤΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΗΘΗΚΑΝ ΓΙΑ ΤΗΝ ΕΚΠΑΙΔΕΥΣΗ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ

Εφαρμογή του αλγορίθμου

Τέλος, έγινε εφαρμογή του αλγορίθμου μέσω της εντολής Advanced Classification> Classifier. Στο πεδίο Διεργασία (Operation) επιλέχθηκε στην περίπτωση αυτή Εφαρμογή (Apply), δηλαδή την εφαρμογή του εκπαιδευμένου αλγορίθμου στα αντικείμενα του δεύτερου επιπέδου. Τέλος, στο Παράμετροι συστήματος (configuration) ορίστηκε η αλφαριθμητική μεταβλητή που δημιουργήθηκε στο στάδιο της εκπαίδευσης του αλγορίθμου (Εικόνα 3.20).



ΕΙΚΟΝΑ 3.20: ΕΦΑΡΜΟΓΗ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΣΤΑ ΑΝΤΙΚΕΙΜΕΝΑ ΤΟΥ ΕΠΙΠΕΔΟΥ 2

3.5.2 Εφαρμογή του αλγορίθμου των τυχαίων δασών

Τα τυχαία δάση όπως και τα δέντρα απόφασης αποτελούν αλγόριθμο επιβλεπόμενης ταξινόμησης. Συνεπώς, η εφαρμογή της συγκεκριμένης μεθόδου υλοποιήθηκε στο περιβάλλον του ECognition σε δύο στάδια:

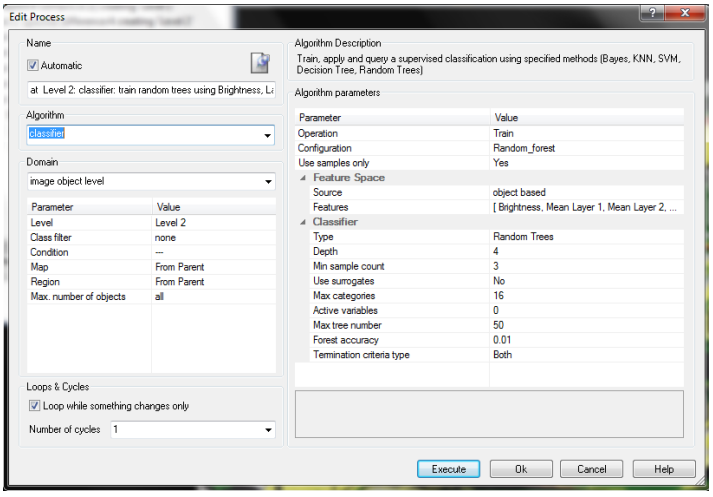
- Εκπαίδευση του αλγορίθμου

- Εφαρμογή του αλγορίθμου

Εκπαίδευση του αλγορίθμου

Η διαδικασία της εκπαίδευσης είναι πανομοιότυπη με εκείνη των δέντρων απόφασης με τη διαφορά ότι στο πεδίο Είδος (Type) του ταξινομητή επιλέγεται ο αλγόριθμος Τυχαία Δάση (Random Trees). Στα πλαίσια εφαρμογής του συγκεκριμένου αλγορίθμου δίνεται η δυνατότητα στο χρήση να ρυθμίσει τις ακόλουθες παραμέτρους:

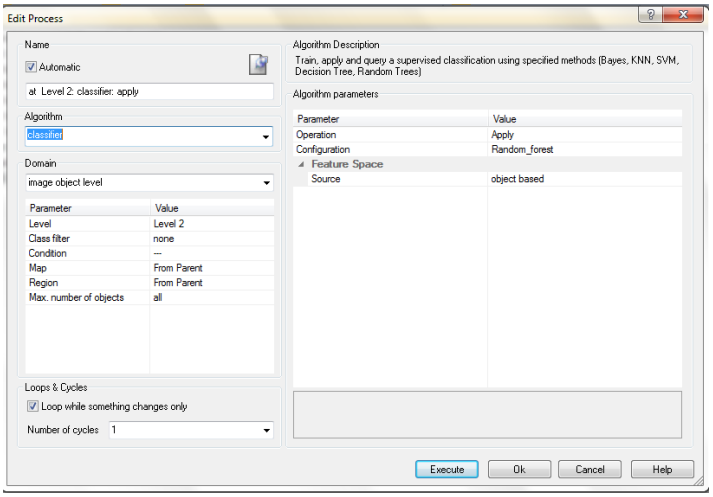
- Βάθος δέντρου (Depth): Μέγιστο βάθος δέντρου (προεπιλεγμένη τιμή: 0)
- Ελάχιστος αριθμός δειγμάτων (Min sample count): Ο ελάχιστος αριθμός δειγμάτων σε κάθε κόμβο (προεπιλεγμένη τιμή: 0)
- Χρήση αντικαταστατών (Use surrogates): Χρήση αντικαταστατών για τις ελλιπείς τιμές. Στην περίπτωση που επιλέγεται η τιμή Ναι (yes) κατασκευάζονται κόμβοι αντικαταστατών, οι οποίοι κατασκευάζονται προκειμένου να είναι δυνατή η διαχείριση των ελλιπών τιμών.
- Μέγιστος αριθμός κατηγοριών (Max categories): Ταξινόμηση των πιθανών τιμών μίας ονομαστικής τιμής σε K (μικρότερος αριθμός από εκείνον του μέγιστου αριθμού κατηγοριών) κατηγορίες (προεπιλεγμένη τιμή: 16)
- Ενεργές μεταβλητές (Active Variables): Το μέγεθος του τυχαία επιλεγμένου υποσυνόλου χαρακτηριστικών σε κάθε κόμβο του δέντρου. Το συγκεκριμένο χρησιμοποιείται ώστε να επιλεχθεί ο καλύτερος δυνατός κόμβος. (προεπιλεγμένη τιμή: 0. Στην περίπτωση που το μέγεθος τίθεται σε 0 η τιμή της μεταβλητής ορίζεται στην τετραγωνική τιμή του συνολικού αριθμού των χαρακτηριστικών)
- Μέγιστος αριθμός δέντρων (Max tree number): Ο μέγιστος αριθμός των δέντρων στο δάσος (προεπιλεγμένη τιμή: 50)
- Ακρίβεια δάσους (Forest accuracy): Επαρκής ακρίβεια του εκπαιδευμένου δάσους σε ποσοστό επί τις εκατό. (προεπιλεγμένη τιμή: 0,01)
- Τύπος κριτηρίου τερματισμού (Termination criteria type): Βάσει της συγκεκριμένης επιλογής ορίζεται κατά πόσον η διαδικασία εκπαίδευσης του δέντρου τερματίζεται όταν:
    - Συμπληρώνεται ο μέγιστος αριθμός δέντρων
    - Το τυχαίο δάσος έχει την απαιτούμενη ακρίβεια
    - Ικανοποιούνται και οι δύο παραπάνω συνθήκες (Εικόνα 3.21)



ΕΙΚΟΝΑ 3.21: ΠΑΡΑΘΥΡΟ ΔΙΑΛΟΓΟΥ ΓΙΑ ΤΗΝ ΕΦΑΡΜΟΓΗ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ

Εφαρμογή του αλγορίθμου

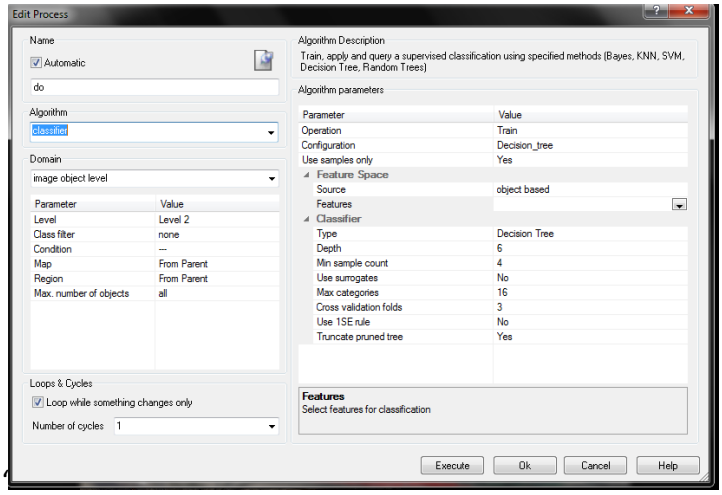
Ο αλγόριθμος των τυχαίων δασών εφαρμόστηκε μέσω της εντολής Advanced Classification> Classifier. Στο πεδίο Διεργασία (Operation) επιλέχθηκε στην περίπτωση αυτή Εφαρμογή (Apply), δηλαδή την εφαρμογή του εκπαιδευμένου αλγορίθμου στα αντικείμενα του δεύτερου επιπέδου. Τέλος, στο Παράμετροι συστήματος (configuration) ορίστηκε η αλφαριθμητική μεταβλητή που δημιουργήθηκε στο στάδιο της εκπαίδευσης του αλγορίθμου (Εικόνα 3.22).



ΕΙΚΟΝΑ 3.22: ΕΦΑΡΜΟΓΗ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΑ ΑΝΤΙΚΕΙΜΕΝΑ ΤΟΥ ΕΠΙΠΕΔΟΥ 2

Τα χαρακτηριστικά των αντικειμένων στα οποία βασίστηκε η διαδικασία της ταξινόμησης για τα δέντρα απόφασης είναι με σειρά προτεραιότητας τα ακόλουθα:

- Μέση τιμή φωτεινότητας για το κανάλι 3 (Mean Layer 3)
- Μέση τιμή φωτεινότητας για το κανάλι 4 (Mean Layer 4)
- Μέση τιμή φωτεινότητας για το κανάλι 1 (Mean Layer 1)
- Μέση τιμή φωτεινότητας για το κανάλι 2 (Mean Layer)
- Δείκτης βλάστησης (NDVI)
- Λόγος μήκος προς πλάτος του εκάστοτε αντικειμένου (Length/Width)
- Συμπαγότητα του αντικειμένου (Compactness)
- Ομοιότητα του σχήματος του αντικειμένου με το σχήμα του ορθογωνίου(Rectangular fit)
- Εμβαδόν αντικειμένου (Area)
- Μέγιστη διαφορά (Max. Diff.)
- Μέση φωτεινότητα του αντικειμένου (Brightness) (Εικόνα 3.23)



ΕΙΚΟΝΑ 3.23: ΡΥΘΜΙΣΗ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΣΤΗΝ ΠΡΩΤΗ ΔΟΚΙΜΗ

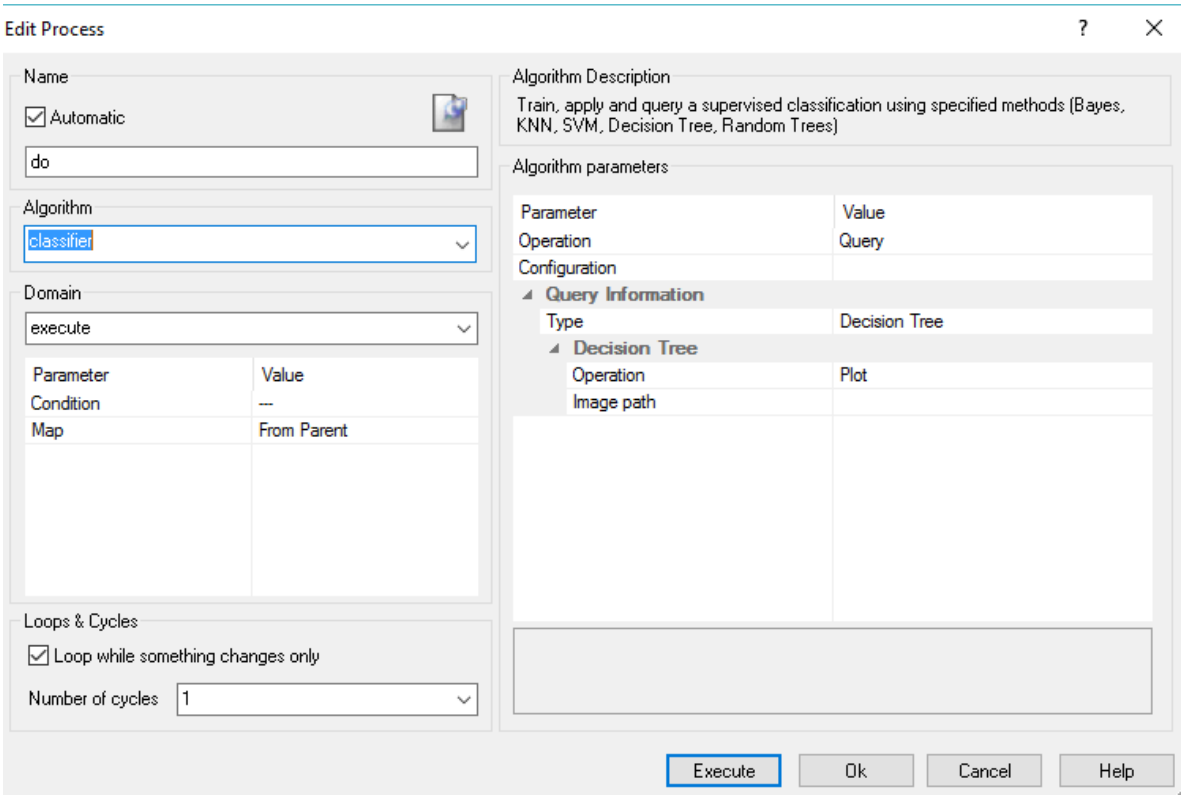
Στη συνέχεια έγιναν διάφορες δοκιμές στις παραμέτρους των δέντρων απόφασης και των τυχαίων δασών προκειμένου να καταστεί δυνατή η εύρεση του συνδυασμού των τιμών που θα δώσει τα βέλτιστα δυνατά αποτελέσματα ταξινόμησης. Στις ακόλουθες Ενότητες γίνεται σχολιασμός της ποιότητας των αποτελεσμάτων για όλες τις θεματικές κατηγορίες που περιέχονται στην εικόνα εισόδου. Σημειώνεται, ωστόσο, πως θα δοθεί μεγαλύτερη έμφαση σε εκείνη των κτιρίων.

Εκτύπωση των μοντέλων

Τέλος, αξίζει να σημειωθεί πως το λογισμικό του eCognition παρέχει τη δυνατότητα οπτικοποίησης των διαφορετικών μοντέλων ταξινόμησης. Το παραπάνω επιτυγχάνεται μέσω της εντολής plot.

Αναλυτικά, η διαδικασία που ακολουθείται έχει ως εξής:

- Στο πεδίο διεργασία του αλγορίθμου ταξινομητής (Classifier) επιλέγεται η εντολή ερώτημα (Query)
- Στο πεδίο των παραμέτρων του συστήματος επιλέγεται η αλφαριθμητική μεταβλητή στην οποία έχει αποθηκευτεί το μοντέλο που εκπαιδεύτηκε
- Στο πεδίο της διεργασίας ορίζεται εκείνη της εκτύπωσης (plot)
- Τέλος στο image path δίνεται από το χρήστη ο φάκελος στον οποίο θα αποθηκευτεί η εικόνα του μοντέλου (Εικόνα 3.24)



ΕΙΚΟΝΑ 3.24: ΠΑΡΑΘΥΡΟ ΔΙΑΛΟΓΟΥ ΓΙΑ ΤΗΝ ΕΚΤΥΠΩΣΗ ΤΟΥ ΜΟΝΤΕΛΟΥ

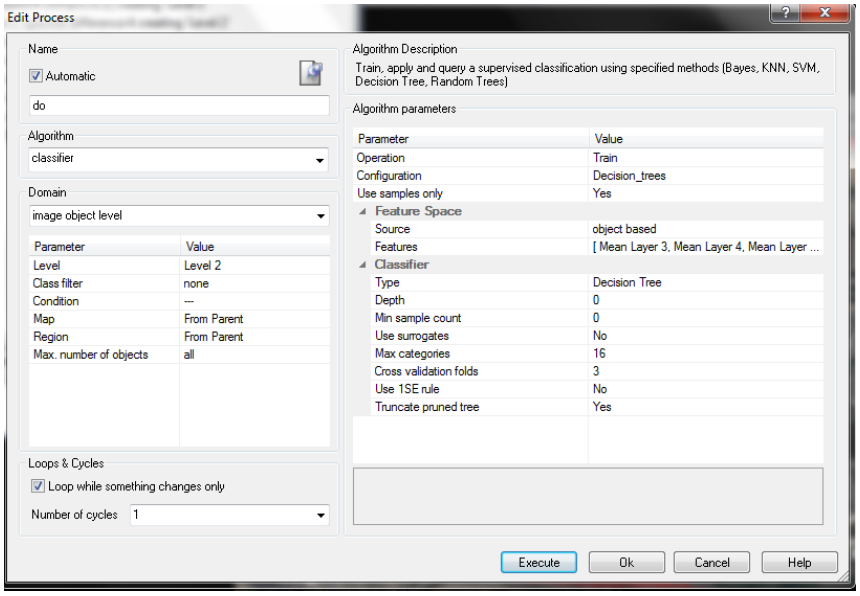
3.6 Υλοποίηση του αλγορίθμου των δέντρων απόφασης στο περιβάλλον του ECognition στο πρώτο τμήμα της εικόνας

Στα πλαίσια του συγκεκριμένου κεφαλαίου έγινε πλήθος δοκιμών σε ό,τι αφορά τις τιμές των παραμέτρων που στόχο είχαν τον προσδιορισμό εκείνων που δίνουν τα βέλτιστα δυνατά αποτελέσματα.

3.6.1 Δοκιμή 1 (Προκαθορισμένες παράμετροι)

Στα πλαίσια της πρώτης δοκιμής έγινε χρήση των προκαθορισμένων τιμών των παραμέτρων. Αναλυτικά, ορίστηκαν οι ακόλουθες τιμές:

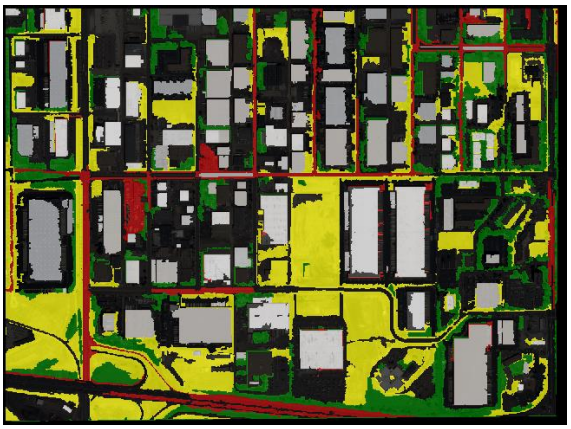
- Βάθος δέντρου (Depth): 0
- Ελάχιστος αριθμός δειγμάτων (Min sample count): 0
- Χρήση αντικαταστατών (Use surrogates): Όχι (No)
- Μέγιστος αριθμός κατηγοριών (Max categories): 16
- Cross Validation folds: 3
- Use 1 SE rule: Όχι (No)
- Αφαίρεση των κλαδεμένων κλαδιών (Truncate pruned trees): Ναι (Yes) (Εικόνα 3.25).



ΕΙΚΟΝΑ 3.25: ΡΥΘΜΙΣΗ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΓΙΑ ΤΗΝ ΠΡΩΤΗ ΔΟΚΙΜΗ

Σχολιασμός αποτελεσμάτων

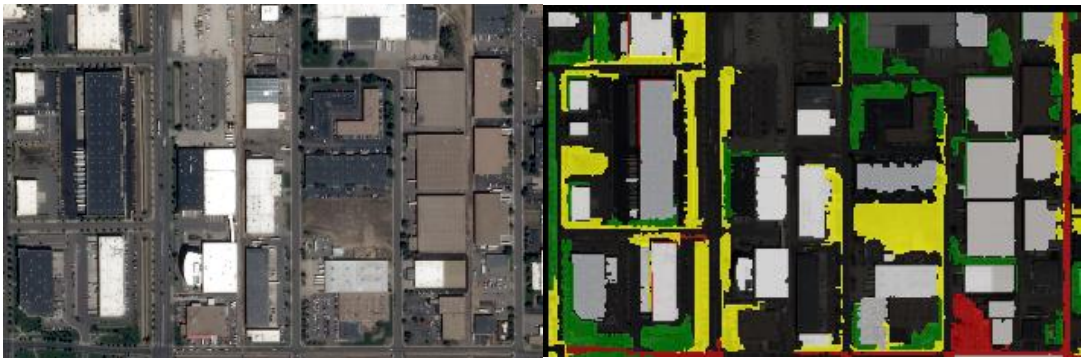
Στην Εικόνα 3.26 εμφανίζεται το αποτέλεσμα της ταξινόμησης των δέντρων απόφασης στα πλαίσια της πρώτης δοκιμής.



ΕΙΚΟΝΑ 3.26: ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΓΙΑ ΤΗΝ ΠΡΩΤΗ ΔΟΚΙΜΗ

ΚΤΙΡΙΑ

Σε ό,τι αφορά τη θεματική κατηγορία των κτιρίων τα αποτελέσματα είναι ικανοποιητικά. Πιο συγκεκριμένα, έχει εντοπιστεί μεγάλος αριθμός αντικειμένων που πράγματι ανήκουν στην κλάση αυτή (True Positives), ενώ παράλληλα δεν έχουν καταχωρηθεί σε αυτήν αντικείμενα από άλλες θεματικές κατηγορίες (False Positives). Προβλήματα ωστόσο εμφανίζονται σε ό,τι αφορά τα αντικείμενα τα οποία ανήκουν στην εν λόγω κλάση αλλά έχουν ταξινομηθεί εσφαλμένα από τον αλγόριθμο σε άλλες όπως για παράδειγμα στην κατηγορία των χώρων στάθμευσης (False Negatives). Βάσει των παραπάνω προκύπτει πως το ποσοστό της ορθότητας για τη θεματική κατηγορία των κτιρίων είναι ιδιαίτερα υψηλό, ενώ τα αποτελέσματα για το κριτήριο της πληρότητας δεν είναι τόσο ενθαρρυντικά (Εικόνα 3.27).

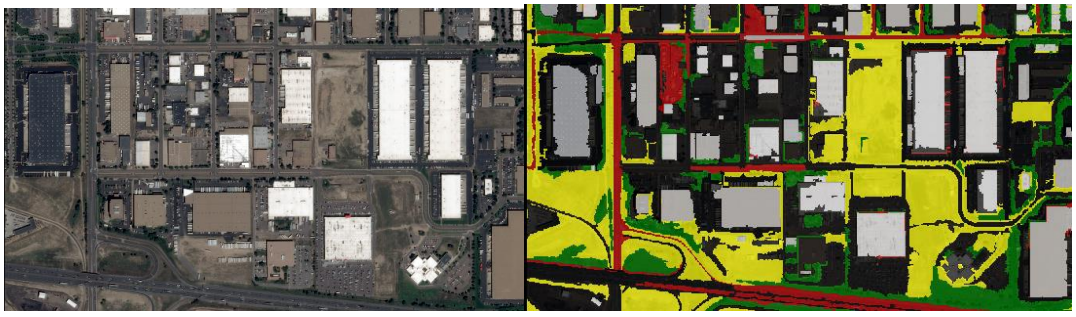


ΕΙΚΟΝΑ 3.27: ΑΡΙΣΤΕΡΑ: 1Ο ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ ΔΕΞΙΑ: 1Ο ΑΠΟΣΠΑΣΜΑ ΤΑΞΙΝΟΜΗΜΕΝΗΣ ΕΙΚΟΝΑΣ ΓΙΑ ΤΗΝ ΠΡΩΤΗ ΔΟΚΙΜΗ

Δρόμοι

Τα αποτελέσματα για τη θεματική κατηγορία των δρόμων είναι απογοητευτικά σε ό,τι αφορά το κριτήριο της πληρότητας. Το παραπάνω επιβεβαιώνεται από το γεγονός πως ο αριθμός των αντικειμένων που ανήκουν στην εν λόγω κλάση αλλά έχουν ταξινομηθεί από τον αλγόριθμο σε εκείνη των χώρων στάθμευσης είναι πολύ μεγάλος. Πιο συγκεκριμένα, τα αντικείμενα εκείνα που έχουν ορθώς ταξινομηθεί από τον αλγόριθμο στους δρόμους είναι μόνο όσα δόθηκαν στη φάση της εκπαίδευσης του αλγορίθμου.

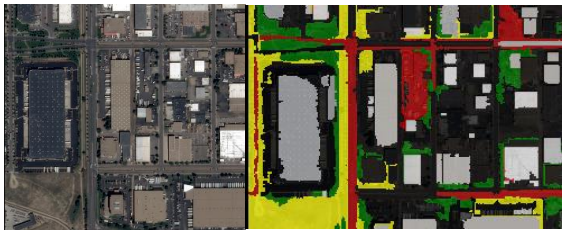
Τα αποτελέσματα σχετικά με το κριτήριο της ορθότητας είναι περισσότερο ενθαρρυντικά καθώς ο αριθμός των αντικειμένων που έχουν λανθασμένα ταξινομηθεί από το μοντέλο στη συγκεκριμένη θεματική κατηγορία είναι περιορισμένος (Εικόνα 3.28).



ΕΙΚΟΝΑ 3.28: ΑΡΙΣΤΕΡΑ:2Ο ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ ΔΕΞΙΑ: 2Ο ΑΠΟΣΠΑΣΜΑ ΤΑΞΙΝΟΜΗΜΕΝΗΣ ΕΙΚΟΝΑΣ ΓΙΑ ΤΗΝ ΠΡΩΤΗ ΔΟΚΙΜΗ

Χώροι Στάθμευσης

Η θεματική κατηγορία των χώρων στάθμευσης είναι σε μεγάλο βαθμό πλήρης καθώς ο αριθμός των αντικειμένων που έχουν ορθώς ταξινομηθεί από τον αλγόριθμο σε αυτή είναι αρκετά μεγάλος και πλησιάζει τον πραγματικό. Τα αποτελέσματα ωστόσο δεν είναι τόσο ενθαρρυντικά σε ό,τι αφορά το κριτήριο της ορθότητας, καθώς στη κατηγορία αυτή έχει καταχωρηθεί πλήθος αντικειμένων τα οποία στην πραγματικότητα ανήκουν σε εκείνη των κτιρίων και των δρόμων.

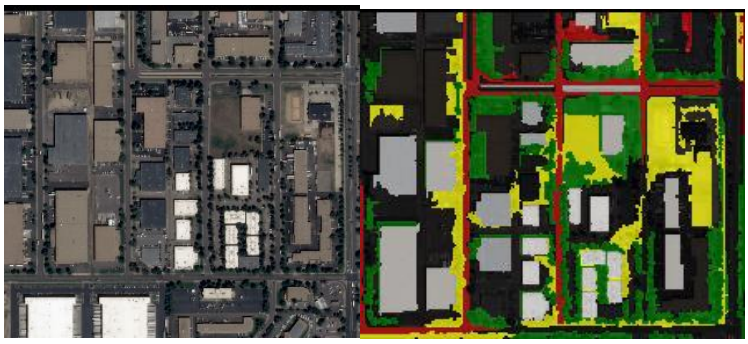


ΕΙΚΟΝΑ 3.29: ΑΡΙΣΤΕΡΑ:3^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ ΔΕΞΙΑ: 3^{Ο} ΑΠΟΣΠΑΣΜΑ ΤΑΞΙΝΟΜΗΜΕΝΗΣ ΕΙΚΟΝΑΣ ΓΙΑ ΤΗΝ ΠΡΩΤΗ ΔΟΚΙΜΗ

Αστικό πράσινο

Τα αποτελέσματα σε ό,τι αφορά τη θεματική κατηγορία του αστικού πρασίνου είναι ικανοποιητικά. Πιο συγκεκριμένα, ο αριθμός των αντικειμένων που έχουν εσφαλμένα ταξινομηθεί από τον αλγόριθμο στην εν λόγω κλάση είναι περιορισμένος. Βάσει του παραπάνω προκύπτει πως το κριτήριο της ορθότητας ικανοποιείται στα πλαίσια της παρούσας εφαρμογής για τη συγκεκριμένη κλάση.

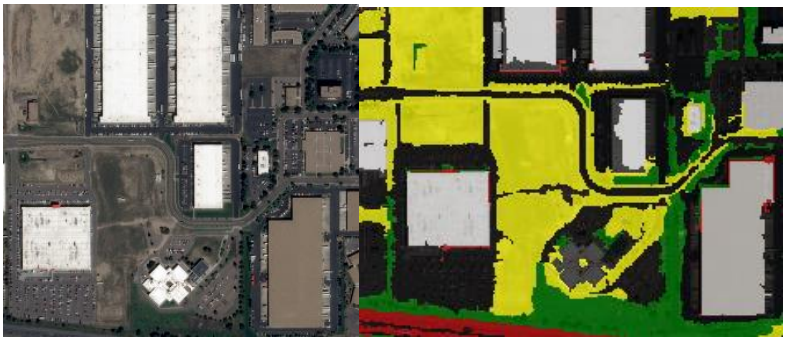
Αρκετά περιθώρια βελτίωσης του αποτελέσματος υπάρχουν, ωστόσο σε ό,τι αφορά την ικανοποίηση του κριτήριου της πληρότητας. Το παραπάνω απορρέει από το γεγονός πως ο αριθμός των αντικειμένων που ανήκουν στην κατηγορία του αστικού πρασίνου και έχουν εσφαλμένα ταξινομηθεί από τον αλγόριθμο στην κλάση του άγονου εδάφους και των χώρων στάθμευσης είναι αρκετά μεγάλος (Εικόνα 3.30).



ΕΙΚΟΝΑ 3.30: ΑΡΙΣΤΕΡΑ: 4^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ ΔΕΞΙΑ: 4^{Ο} ΑΠΟΣΠΑΣΜΑ ΤΑΞΙΝΟΜΗΜΕΝΗΣ ΕΙΚΟΝΑΣ ΓΙΑ ΤΗΝ ΠΡΩΤΗ ΔΟΚΙΜΗ

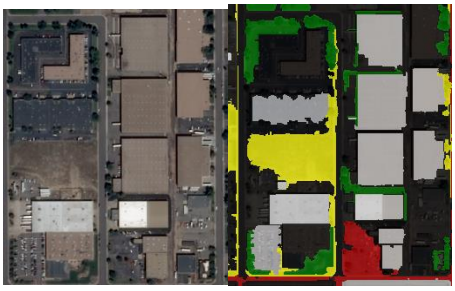
Άγονο Έδαφος

Το κριτήριο τόσο της πληρότητας όσο και της ορθότητας είναι αρκετά υψηλό για τα αντικείμενα του Άγονου Εδάφους (Εικόνα 3.31).



ΕΙΚΟΝΑ 3.31: ΑΡΙΣΤΕΡΑ: 5^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ ΔΕΞΙΑ: 5^{Ο} ΑΠΟΣΠΑΣΜΑ ΤΑΞΙΝΟΜΗΜΕΝΗΣ ΕΙΚΟΝΑΣ ΓΙΑ ΤΗΝ ΠΡΩΤΗ ΔΟΚΙΜΗ

Ποσοτική αξιολόγηση αποτελεσμάτων



ΕΙΚΟΝΑ 3.32: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΗΝ ΠΡΩΤΗ ΔΟΚΙΜΗ

Όπως διευκρινίστηκε παραπάνω βασικός στόχος της παρούσας μεταπτυχιακής εργασίας είναι η ανίχνευση κτιρίων από τη δοθείσα δορυφορική εικόνα. Τυπικά, η συγκεκριμένη αξιολόγηση, θα έπρεπε να πραγματοποιηθεί βάσει των πραγματικών διανυσματικών δεδομένων που περιγράφουν την περιοχή μελέτης. Εφόσον, τα εν λόγω δεδομένα δε διατίθενται, η διαδικασία που ακολουθείται είναι η εξής: Επιλέγεται αντιπροσωπευτική περιοχή της εικόνας- εισόδου (Εικόνα 3.32). Για την περιοχή αυτή εντοπίζεται μέσω φωτοερμηνείας ο αριθμός των True Positives, False Positives και False Negatives δεδομένων. Βάσει των παραπάνω στοιχείων κατασκευάζονται οι ακόλουθοι Πίνακες (Πίνακας 3.1, Πίνακας 3.2):

ΠΙΝΑΚΑΣ 3.1: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (1Η ΔΟΚΙΜΗ).

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	11	2	5

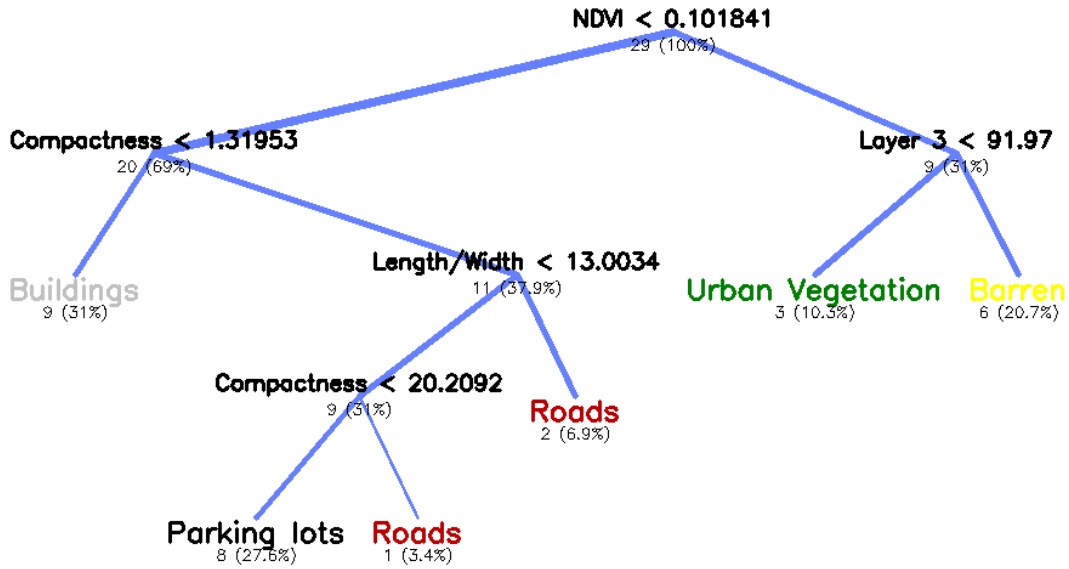
ΠΙΝΑΚΑΣ 3.2: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (1Η ΔΟΚΙΜΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	68,75%	84,62%	61,11%	31,25%	12,50%

Στην ακόλουθη Εικόνα (Εικόνα 3.33) εμφανίζεται το δέντρο απόφασης το οποίο κατασκευάστηκε βάσει των παραπάνω τιμών των παραμέτρων. Το συγκεκριμένο δέντρο απόφασης είναι δυαδικό (εφόσον το λογισμικό του eCognition χρησιμοποιεί τον αλγόριθμο CART) επίσης τα κριτήρια βάσει των οποίων γίνεται ο διαχωρισμός των δεδομένων είναι μονομεταβλητά. Τα χαρακτηριστικά εκείνα τα οποία χρησιμοποιήθηκαν για να γίνει ο διαχωρισμός των δεδομένων είναι τα ακόλουθα:

- Συμπαγότητα (Compactness)
- Η τιμή φωτεινότητας για το κανάλι 3
- Ο λόγος μήκους προς πλάτος

Το βάθος του συγκεκριμένου μοντέλου δεν είναι ίσο με 0 όπως ήταν αναμενόμενο. Η τιμή που δόθηκε στην παράμετρο αυτή ήταν η βέλτιστη δυνατή κατά τον αλγόριθμο και πιο συγκεκριμένα δόθηκε η τιμή 4.



ΕΙΚΟΝΑ 3.33: ΔΕΝΤΡΟ ΑΠΟΦΑΣΗΣ ΓΙΑ ΠΡΟΚΑΘΟΡΙΣΜΕΝΕΣ ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ

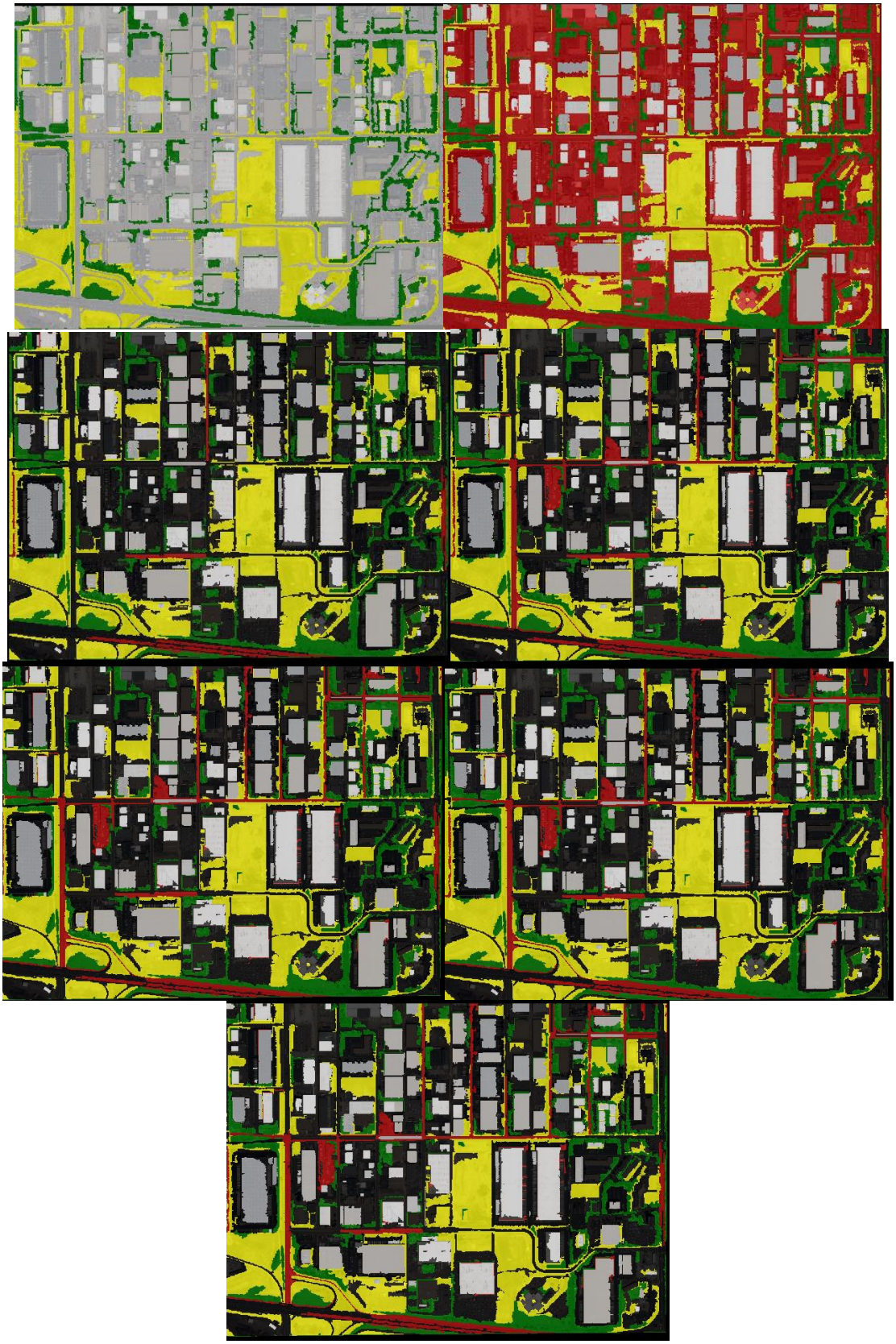
3.6.2 Δοκιμή 2 (Βάθος δέντρου)

Στόχος της συγκεκριμένης δοκιμής ήταν η διερεύνηση της επιρροής της παραμέτρου Βάθος του δέντρου στο αποτέλεσμα του αλγορίθμου των δέντρων απόφασης. Το εν λόγω μέγεθος σχετίζεται άμεσα με την υποπροσαρμογή ή υπερπροσαρμογή του μοντέλου στα δεδομένα εκπαίδευσης. Μέσω της διαδικασίας αυτής επιδιώκεται η εύρεση της τιμής εκείνης η οποία θα δώσει τα βέλτιστα δυνατά αποτελέσματα. Αναλυτικά, δημιουργήθηκε ένα σύνολο 7 διαφορετικών μοντέλων βάσει των ακόλουθων τιμών:

- **Βάθος δέντρου (Depth): 2, 3, 4, 5, 10, 25, 50**
- Ελάχιστος αριθμός δειγμάτων (Min sample count): 0
- Χρήση αντικαταστατών (Use surrogates): Όχι (No)
- Μέγιστος αριθμός κατηγοριών (Max categories): 16
- Cross Validation folds: 3
- Use 1 SE rule: Όχι (No)
- Αφαίρεση των κλαδεμένων κλαδιών (Truncate pruned trees): Ναι (Yes)

Σχολιασμός αποτελεσμάτων

Στην Εικόνα 3.34 εμφανίζεται το αποτέλεσμα εφαρμογής του αλγορίθμου των δέντρων απόφασης για διαφορετικές τιμές βάθους. Είναι εμφανές πως για τιμές άνω των 5 η ρύθμιση της συγκεκριμένης μεταβλητής δεν επηρέασε το αποτέλεσμα της ταξινόμησης. Αντιθέτως τα δέντρα απόφασης με μικρότερο βάθος, δηλαδή τιμές 2, 3, 4 εμφανίζουν διαφορετικά αποτελέσματα. Το παραπάνω οφείλεται στο γεγονός πως τα δέντρα απόφασης στις περιπτώσεις αυτές εμφανίζουν υποπροσαρμογή στα δεδομένα εισόδου.

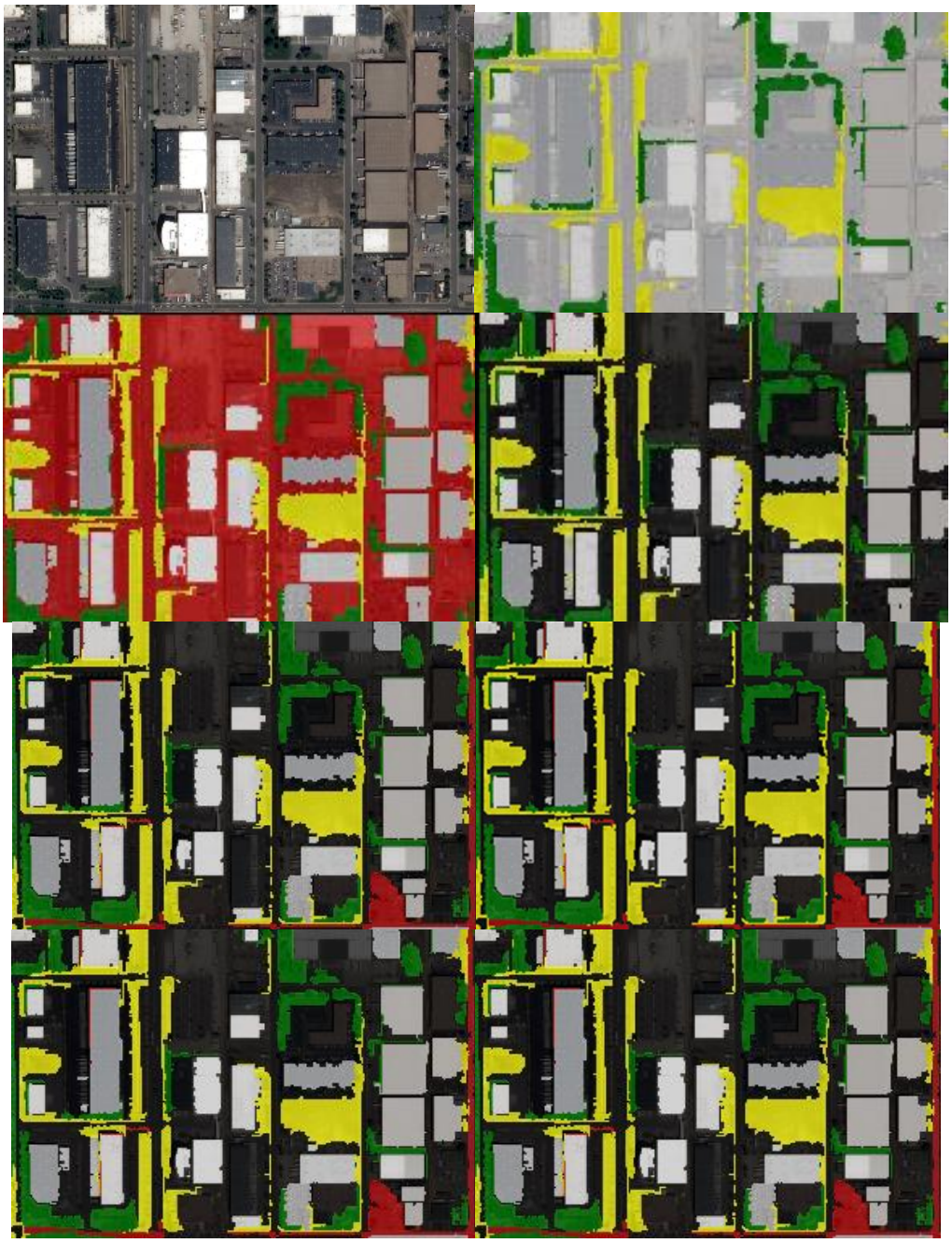


ΕΙΚΟΝΑ 3.34: ΑΠΟΤΕΛΕΣΜΑ ΕΦΑΡΜΟΓΗΣ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΓΙΑ ΤΙΜΕΣ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΒΑΘΟΣ ΔΕΝΤΡΩΝ ΑΠΟ ΠΑΝΩ ΑΡΙΣΤΕΡΑ 2, 3, 4, 5, 10, 25, 50

ΚΤΙΡΙΑ

Η συγκεκριμένη θεματική κατηγορία εμφανίζει υψηλά ποσοστά πληρότητας στην περίπτωση ρύθμισης της τιμής του βάθους σε 2, καθώς στον παραγόμενο θεματικό χάρτη έχει αναγνωριστεί το σύνολο των εμφανιζόμενων κτιρίων. Στην κλάση αυτή, ωστόσο, έχουν καταχωρηθεί όλα τα αντικείμενα εκείνα τα οποία στην πραγματικότητα ανήκουν σε εκείνη των δρόμων και των χώρων στάθμευσης. Βάσει αυτού προκύπτει ότι το κριτήριο της ορθότητας εμφανίζει πολύ χαμηλά ποσοστά στην προκειμένη περίπτωση.

Η ρύθμιση της παραμέτρου σε 3 εμφανίζει ακριβώς τα αντίστοιχα αποτελέσματα με εκείνα των προκαθορισμένων τιμών (Εικόνα 3.35).



ΕΙΚΟΝΑ 3.35: ***ΑΠΟ ΤΗΝ ΑΡΧΗ: 1^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ, ΓΙΑ ΤΗΝ ΤΙΜΗ 2 ΤΟΥ ΒΑΘΟΥΣ ΔΕΝΤΡΩΝ, ΓΙΑ ΤΗΝ ΤΙΜΗ 3, ΓΙΑ ΤΗΝ ΤΙΜΗ 4, ΓΙΑ ΤΗΝ ΤΙΜΗ 5, ΓΙΑ ΤΗΝ ΤΙΜΗ 10, ΓΙΑ ΤΗΝ ΤΙΜΗ 25, ΓΙΑ ΤΗΝ ΤΙΜΗ 50***

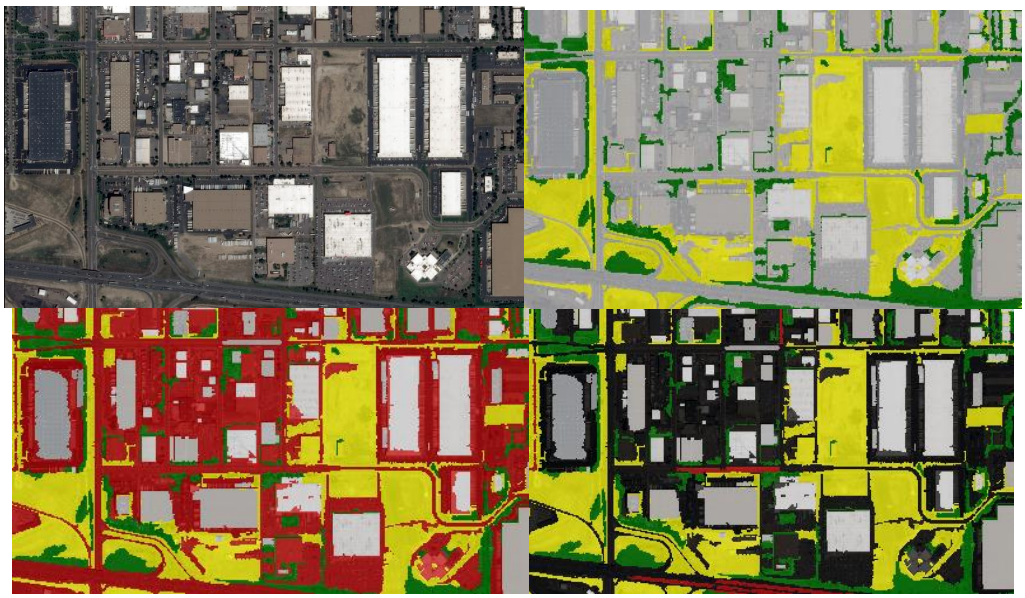
ΔΡΟΜΟΙ

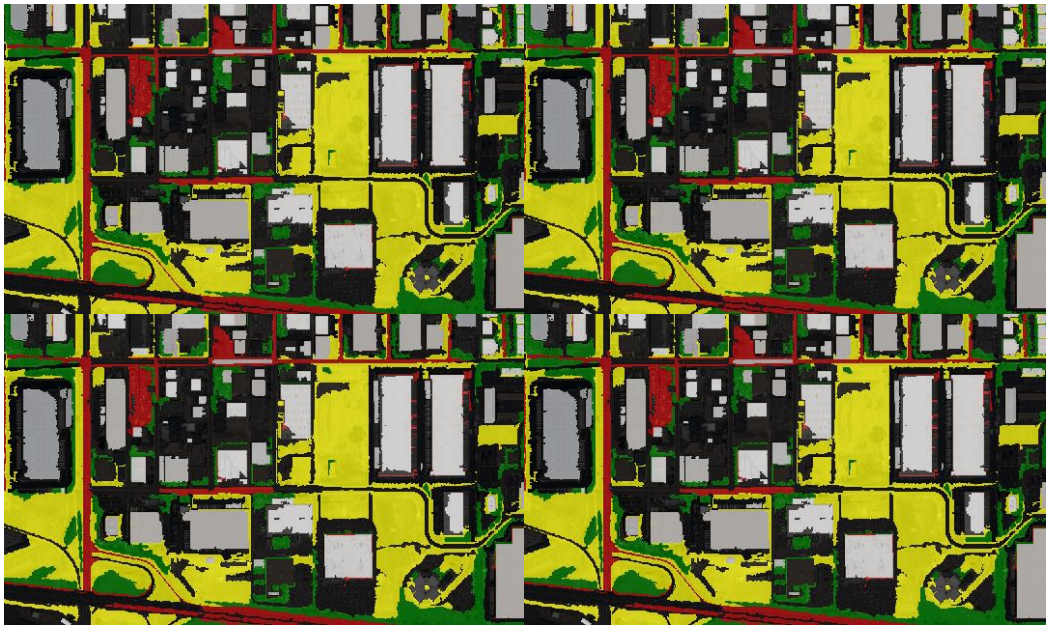
Στην περίπτωση ρύθμισης της παραμέτρου του βάθους σε 2 τα ποσοστά πληρότητας και ορθότητας είναι 0%, καθώς στον παραγόμενο θεματικό χάρτη δεν εμφανίζονται αντικείμενα τα κατηγορίας αυτής. Το παραπάνω οφείλεται στο γεγονός πως το βάθος του δέντρου είναι πολύ μικρό και ως εκ τούτου το μοντέλο δεν έχει προσαρμοστεί στα δεδομένα εκπαίδευσης.

Η αύξηση του βάθους του δέντρου σε 3 αύξησε τον αριθμό των αντικειμένων της κατηγορίας αυτής. Η πληρότητα του παραγόμενου θεματικού χάρτη στην περίπτωση αυτή είναι 100% σε ό,τι αφορά τη συγκεκριμένη θεματική κατηγορία. Το ποσοστό της ορθότητας, ωστόσο, είναι απογοητευτικό καθώς πολλά από τα αντικείμενα τα οποία έχουν καταχωρηθεί στην κλάση των δρόμων ανήκουν στην πραγματικότητα σε εκείνη των κτιρίων και των χώρων στάθμευσης.

Η ρύθμιση της τιμής σε 4 μείωσε τον αριθμό των αντικειμένων της κατηγορίας αυτής. Το τελευταίο αύξησε τα ποσοστά της ορθότητας καθώς απομακρύνθηκαν από την κλάση αυτή πολλά αντικείμενα τα οποία δεν ανήκουν στην πραγματικότητα στη συγκεκριμένη. Παράλληλα, μειώθηκε το ποσοστό πληρότητας καθώς ο αριθμός των οδικών αξόνων οι οποίοι εντοπίζονται στον τελικό θεματικό χάρτη είναι πολύ μικρός.

Τέλος, το αποτέλεσμα της ταξινόμησης είναι πανομοιότυπο για τις τιμές βάθους 5, 10, 25, 50, 100. Τα αντικείμενα της εν λόγω κλάσης αυξήθηκαν και το παραπάνω οδήγησε σε αύξηση της πληρότητας αυτής (Εικόνα 3.36).





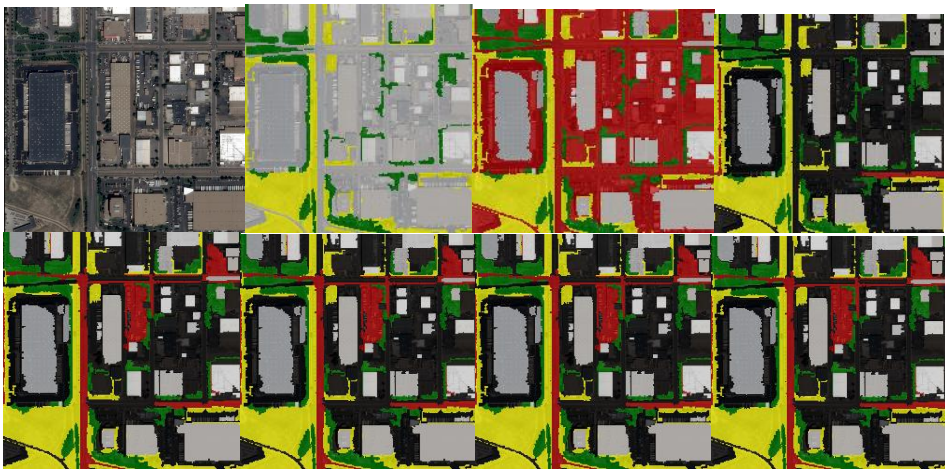
ΕΙΚΟΝΑ 3.36: ***ΑΠΟ ΤΗΝ ΑΡΧΗ: 2ο ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ, ΓΙΑ ΤΗΝ ΤΙΜΗ 2 ΤΟΥ ΒΑΘΟΥΣ ΔΕΝΤΡΩΝ, ΓΙΑ ΤΗΝ ΤΙΜΗ 3, ΓΙΑ ΤΗΝ ΤΙΜΗ 4, ΓΙΑ ΤΗΝ ΤΙΜΗ 5, ΓΙΑ ΤΗΝ ΤΙΜΗ 10, ΓΙΑ ΤΗΝ ΤΙΜΗ 25, ΓΙΑ ΤΗΝ ΤΙΜΗ 50***

ΧΩΡΟΙ ΣΤΑΘΜΕΥΣΗΣ

Η κλάση των χώρων στάθμευσης δεν εμφανίζεται στους θεματικούς χάρτες που προέκυψαν για τιμές βάθους 2 και 3. Ως εκ τούτου τα ποσοστά ορθότητας και πληρότητας είναι μηδενικά για την κατηγορία αύτη.

Η ρύθμιση της τιμής της συγκεκριμένης παραμέτρου σε 4 αύξησε τον αριθμό των αντικειμένων της κατηγορίας αυτής. Το ποσοστό της πληρότητας αγγίζει το 100% καθώς έχει ανιχνευτεί από τον αλγόριθμο όλοι οι εμφανιζόμενοι χώροι στάθμευσης. Παράλληλα, στην κλάση αυτή προστέθηκαν κτίρια καθώς και τμήματα των οδικών αξόνων. Συνεπώς εμφανίζονται προβλήματα σχετικά με την ικανοποίηση του κριτηρίου της ορθότητας.

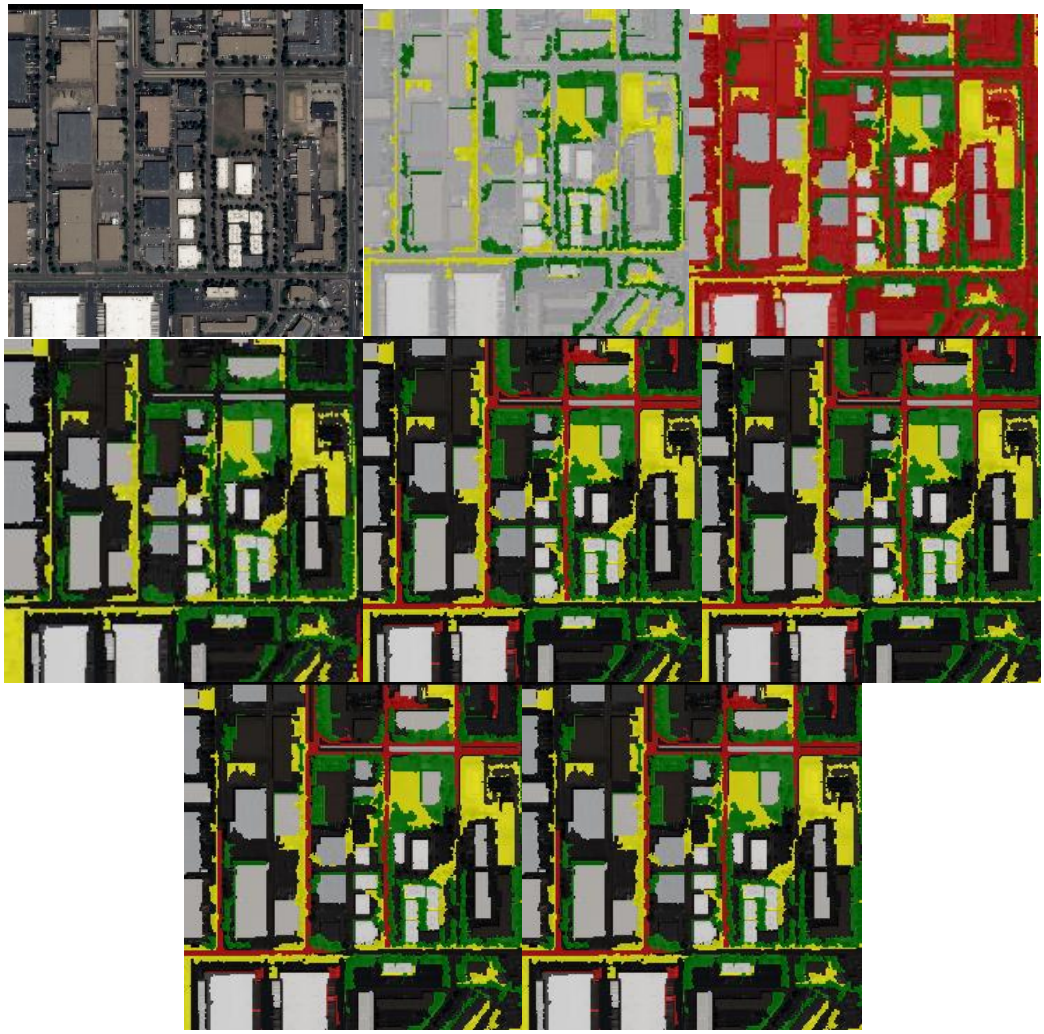
Τέλος, το αποτέλεσμα της ταξινόμησης είναι πανομοιότυπο για τις τιμές βάθους 5, 10, 25, 50, 100. Η αύξηση της τιμής του βάθους οδήγησε σε αύξηση των ποσοστών ορθότητας της κλάσης των χώρων στάθμευσης καθώς από την κατηγορία αυτή απομακρύνθηκαν τμήματα των εμφανιζόμενων οδικών αξόνων (Εικόνα 3.37).



ΕΙΚΟΝΑ 3.37: ***ΑΠΟ ΤΗΝ ΑΡΧΗ: 3ο ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ, ΓΙΑ ΤΗΝ ΤΙΜΗ 2 ΤΟΥ ΒΑΘΟΥΣ ΔΕΝΤΡΩΝ, ΓΙΑ ΤΗΝ ΤΙΜΗ 3, ΓΙΑ ΤΗΝ ΤΙΜΗ 4, ΓΙΑ ΤΗΝ ΤΙΜΗ 5, ΓΙΑ ΤΗΝ ΤΙΜΗ 10, ΓΙΑ ΤΗΝ ΤΙΜΗ 25, ΓΙΑ ΤΗΝ ΤΙΜΗ 50***

ΑΣΤΙΚΟ ΠΡΑΣΙΝΟ

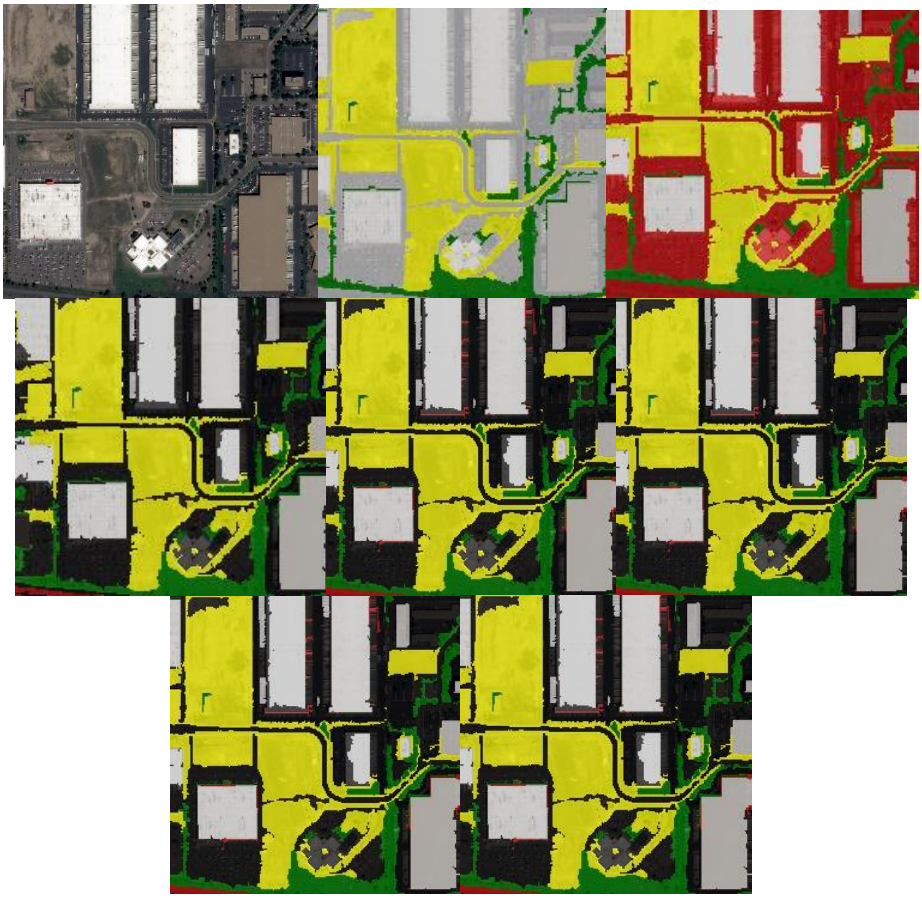
Το αποτέλεσμα της ταξινόμησης για την κλάση του αστικού πρασίνου είναι πανομοιότυπο για τις διαφορετικές τιμές του βάθους των δέντρων απόφασης (Εικόνα 3.38).



ΕΙΚΟΝΑ 3.38: ***ΑΠΟ ΤΗΝ ΑΡΧΗ: 4^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ, ΓΙΑ ΤΗΝ ΤΙΜΗ 2 ΤΟΥ ΒΑΘΟΥΣ ΔΕΝΤΡΩΝ, ΓΙΑ ΤΗΝ ΤΙΜΗ 3, ΓΙΑ ΤΗΝ ΤΙΜΗ 4, ΓΙΑ ΤΗΝ ΤΙΜΗ 5, ΓΙΑ ΤΗΝ ΤΙΜΗ 10, ΓΙΑ ΤΗΝ ΤΙΜΗ 25, ΓΙΑ ΤΗΝ ΤΙΜΗ 50***

ΆΓΟΝΟ ΈΔΑΦΟΣ

Το αποτέλεσμα της ταξινόμησης για την κλάση του άγονου εδάφους είναι ακριβώς το ίδιο για τις διαφορετικές τιμές του βάθους των δέντρων απόφασης (Εικόνα 3.39).



ΕΙΚΟΝΑ 3.39: ***ΑΠΟ ΤΗΝ ΑΡΧΗ: 5^{ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ, ΓΙΑ ΤΗΝ ΤΙΜΗ 2 ΤΟΥ ΒΑΘΟΥΣ ΔΕΝΤΡΩΝ, ΓΙΑ ΤΗΝ ΤΙΜΗ 3, ΓΙΑ ΤΗΝ ΤΙΜΗ 4, ΓΙΑ ΤΗΝ ΤΙΜΗ 5, ΓΙΑ ΤΗΝ ΤΙΜΗ 10, ΓΙΑ ΤΗΝ ΤΙΜΗ 25, ΓΙΑ ΤΗΝ ΤΙΜΗ 50***

Ποσοτική αξιολόγηση αποτελεσμάτων

ΒΑΘΟΣ ΔΕΝΤΡΩΝ: 2

ΕΙΚΟΝΑ 3.40: ***ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΙΜΗ ΒΑΘΟΥΣ ΔΕΝΤΡΟΥ 2***

Βάσει της Εικόνα 3.40 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.3, Πίνακας 3.4)

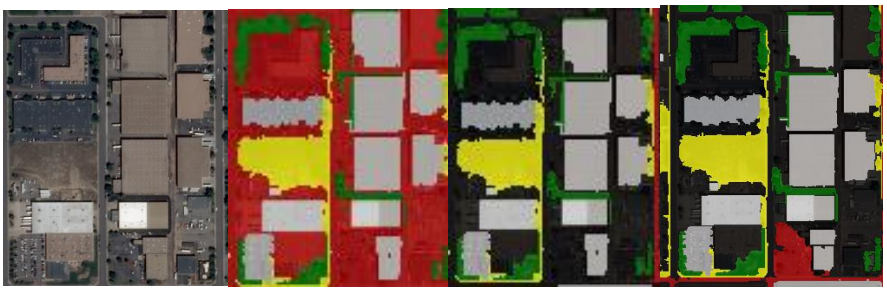
ΠΙΝΑΚΑΣ 3.3: ***ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (2^{Η} ΔΟΚΙΜΗ).***

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	16	9	0

ΠΙΝΑΚΑΣ 3.4: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (2^{Η} ΔΟΚΙΜΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	100,00%	64,00%	64,00%	0,00%	56,25%

ΒΑΘΟΣ ΔΕΝΤΡΩΝ: 3, 4, 5, 10, 25, 50



ΕΙΚΟΝΑ 3.41: ΑΠΟ ΑΡΙΣΤΕΡΑ: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΙΜΗ ΒΑΘΟΥΣ 3, ΓΙΑ ΤΙΜΗ ΒΑΘΟΥΣ 4, ΓΙΑ ΤΙΜΕΣ ΒΑΘΟΥΣ ΔΕΝΤΡΟΥ 5, 10, 50, 100

Βάσει της Εικόνα 3.41 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.5, Πίνακας 3.6)

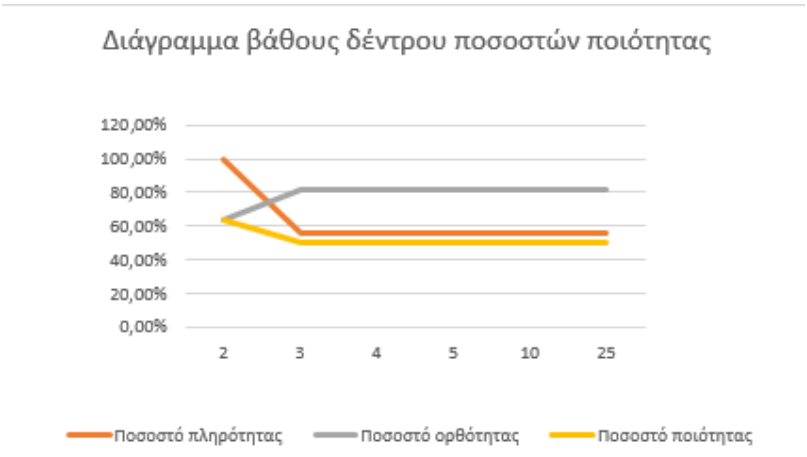
ΠΙΝΑΚΑΣ 3.5: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (2Η ΔΟΚΙΜΗ).

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	11	2	5

ΠΙΝΑΚΑΣ 3.6: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (2Η ΔΟΚΙΜΗ).

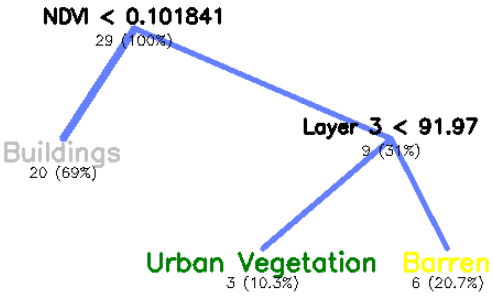
	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	68,75%	84,62%	61,11%	31,25%	12,50%

Στην Εικόνα 3.42 εμφανίζεται το Διάγραμμα βάθους δέντρων ποσοστών ποιότητας.

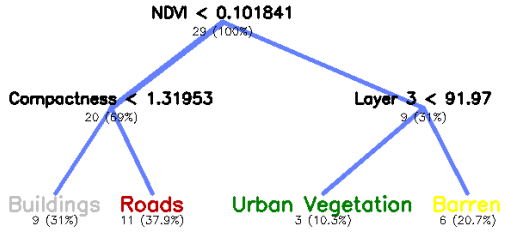


ΕΙΚΟΝΑ 3.42: ΔΙΑΓΡΑΜΜΑ ΒΑΘΟΥΣ ΔΕΝΤΡΩΝ ΠΟΣΟΣΤΩΝ ΠΟΙΟΤΗΤΑΣ

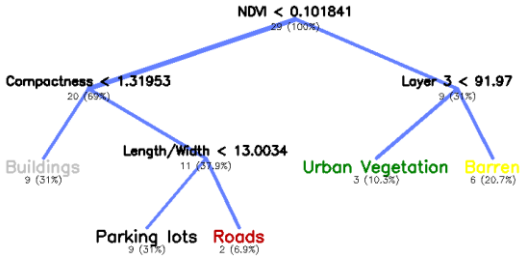
Στις ακόλουθες Εικόνες (Εικόνα 3.43 - Εικόνα 3.46) εμφανίζονται τα δέντρα απόφασης για τις διαφορετικές τιμές βάθους.



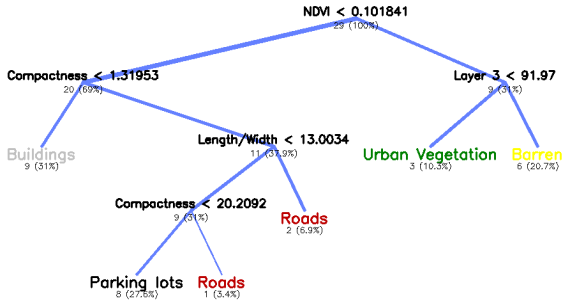
ΕΙΚΟΝΑ 3.43: ΔΕΝΤΡΟ ΑΠΟΦΑΣΗΣ ΒΑΘΟΥΣ 2



ΕΙΚΟΝΑ 3.44: ΔΕΝΤΡΟ ΑΠΟΦΑΣΗΣ ΓΙΑ ΒΑΘΟΣ 3



ΕΙΚΟΝΑ 3.45: ΔΕΝΤΡΟ ΑΠΟΦΑΣΗΣ ΓΙΑ ΒΑΘΟΣ 4



ΕΙΚΟΝΑ 3.46: ΔΕΝΤΡΟ ΑΠΟΦΑΣΗΣ ΓΙΑ ΒΑΘΟΣ ΜΕΓΑΛΥΤΕΡΟ Η ΙΣΟ ΤΟΥ 5

Στην Εικόνα 3.46 παρατηρείται πως τα δέντρα τα οποία κατασκευάζονται για τιμές βάθους μεγαλύτερες ή ίσες του 5 είναι πανομοιότυπα. Το παραπάνω οφείλεται στο γεγονός πως η παράμετρος του αλγορίθμου σχετικά με το κλάδεμα του δέντρου είναι ενεργοποιημένη (Truncate pruned trees: Yes).

3.6.3 Δοκιμή 3 (Ελάχιστός αριθμός δειγμάτων)

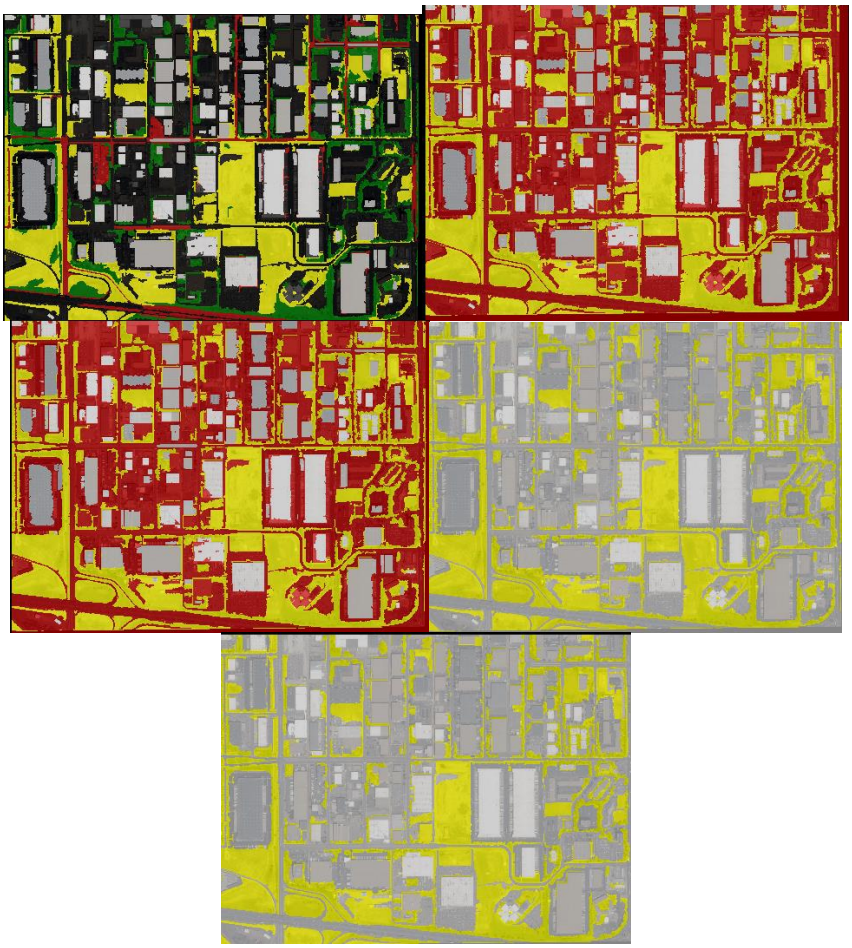
Στα πλαίσια της τρίτης δοκιμής διερευνήθηκε η επιρροή του ελάχιστου αριθμού δειγμάτων στην αποτελεσματικότητα της ταξινόμησης. Αναλυτικά ορίστηκαν οι ακόλουθες τιμές:

- Βάθος δέντρου (Depth): 0
- **Ελάχιστος αριθμός δειγμάτων (Min sample count): 5, 10, 15, 20, 25**

- Χρήση αντικαταστατών (Use surrogates): Όχι (No),
- Μέγιστος αριθμός κατηγοριών (Max categories): 16
- Cross Validation folds: 3
- Use 1 SE rule: Όχι (No)
- Αφαίρεση των κλαδεμένων κλαδιών (Truncate pruned trees): Ναι (Yes)

Σχολιασμός αποτελεσμάτων

Στην Εικόνα 3.47 εμφανίζεται το αποτέλεσμα της ταξινόμησης με τον αλγόριθμο των δέντρων απόφασης για διαφορετικές τιμές της παραμέτρου των ελάχιστων δειγμάτων. Η ρύθμιση της συγκεκριμένης σε 5 έδωσε αποτέλεσμα ταξινόμησης ίδιο με εκείνο των προκαθορισμένων παραμέτρων. Επιπροσθέτως, ο θεματικός χάρτης ο οποίος προέκυψε για την τιμή 10 είναι ακριβώς ο ίδιος με εκείνον της 15. Το αντίστοιχο ισχύει για τις τιμές 20 και 25.



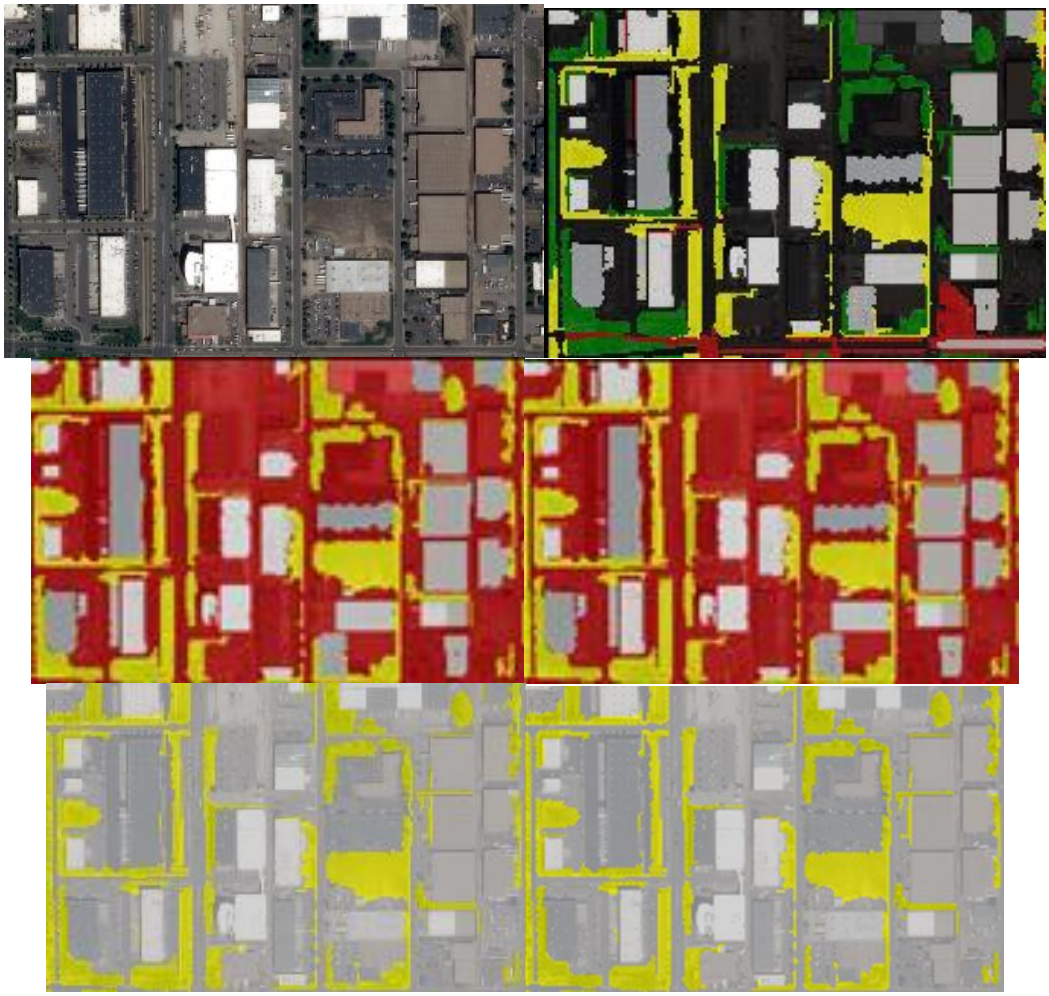
ΕΙΚΟΝΑ 3.47: ***ΑΠΟΤΕΛΕΣΜΑ ΕΦΑΡΜΟΓΗΣ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΓΙΑ ΤΙΜΕΣ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΕΛΑΧΙΣΤΟΣ ΑΡΙΘΜΟΣ ΔΕΙΓΜΑΤΩΝ ΑΠΟ ΠΑΝΩ ΑΡΙΣΤΕΡΑ 5, 10, 15, 20, 25***

ΚΤΙΡΙΑ

Το αποτέλεσμα της ταξινόμησης είναι πανομοιότυπο όταν ο ελάχιστος αριθμός δειγμάτων είναι ίσος με 5, 10, 15.

Η ρύθμιση της εν λόγω παραμέτρου σε 20 και 25 οδήγησε σε αύξηση του πλήθους των κτιρίων στον τελικό θεματικό χάρτη. Το παραπάνω είχε θετικά αποτελέσματα σε ό,τι αφορά την ικανοποίηση του κριτηρίου της πληρότητας. Στην κλάση αυτή προστέθηκαν, ωστόσο

όλα τα αντικείμενα των κλάσεων του αστικού πρασίνου, των χώρων στάθμευσης και των δρόμων. Ως εκ τούτου το κριτήριο της ορθότητας δεν ικανοποιείται στην προκειμένη περίπτωση (Εικόνα 3.48).

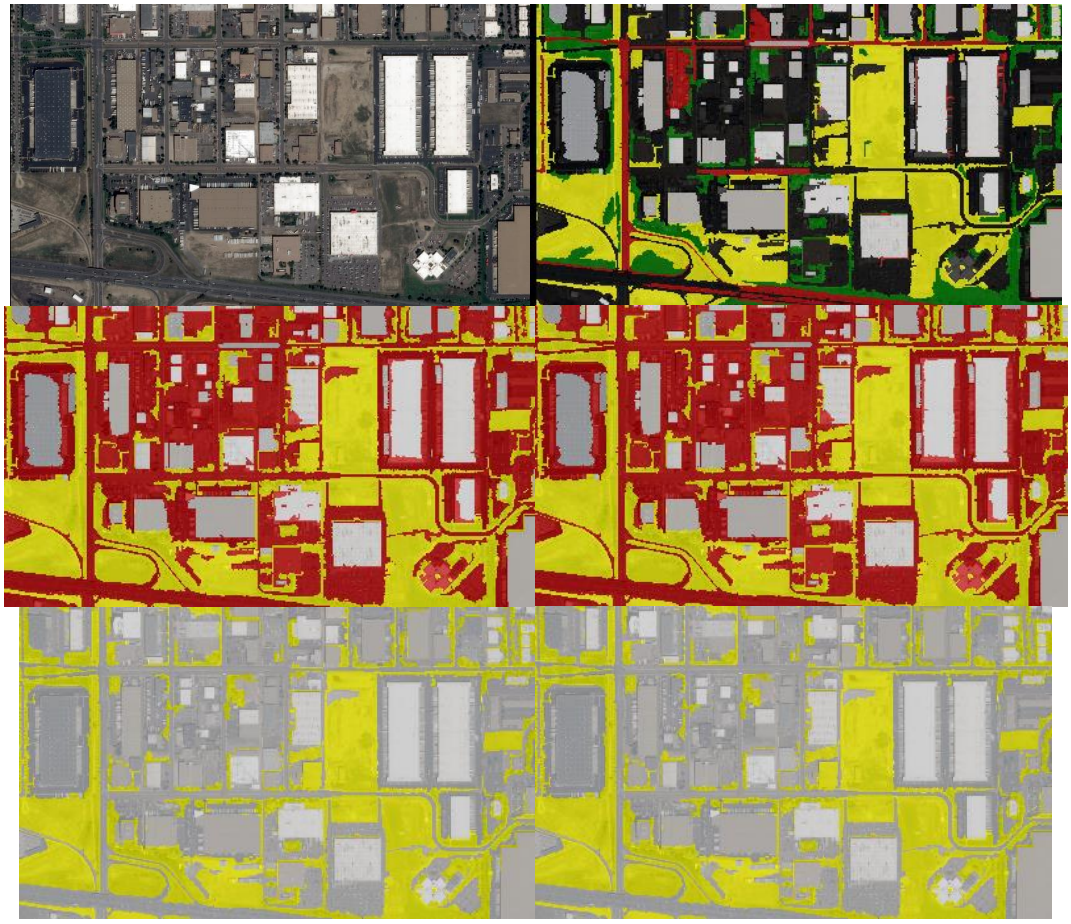


ΕΙΚΟΝΑ 3.48: ***ΑΠΟ ΤΗΝ ΑΡΧΗ: 1^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ, ΓΙΑ ΤΗΝ ΤΙΜΗ 5 ΤΟΥ ΕΛΑΧΙΣΤΟΥ ΑΡΙΘΜΟΥ ΔΕΙΓΜΑΤΩΝ, ΓΙΑ ΤΗΝ ΤΙΜΗ 10, ΓΙΑ ΤΗΝ ΤΙΜΗ 15, ΓΙΑ ΤΗΝ ΤΙΜΗ 20, ΓΙΑ ΤΗΝ ΤΙΜΗ 25***

ΔΡΟΜΟΙ

Η ρύθμιση της παραμέτρου ελάχιστος αριθμός δειγμάτων σε 10 και 15 οδήγησε σε αύξηση του πλήθους των αντικειμένων της κλάσης των δρόμων. Το παραπάνω επέφερε θετικά αποτελέσματα σε ό,τι αφορά το κριτήριο της πληρότητας. Παράλληλα, στην κλάση αυτή προστέθηκαν πολλά αντικείμενα τα οποία στην πραγματικότητα ανήκουν σε εκείνη των κτιρίων, των χώρων στάθμευσης και του αστικού πρασίνου. Συνεπώς, το κριτήριο της ορθότητας δεν ικανοποιείται στην περίπτωση αυτή.

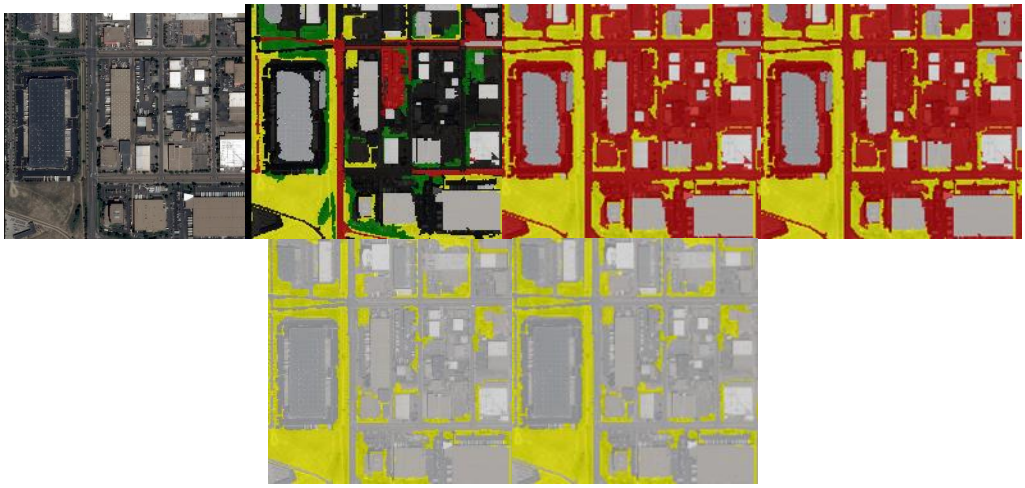
Οι θεματικοί χάρτες οι οποίοι προέκυψαν για τις τιμές 20 και 25 της συγκεκριμένης παραμέτρου δεν περιέχουν την κλάση των δρόμων (Εικόνα 3.49).



ΕΙΚΟΝΑ 3.49: ***ΑΠΟ ΤΗΝ ΑΡΧΗ: 2^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ, ΓΙΑ ΤΗΝ ΤΙΜΗ 5 ΤΟΥ ΕΛΑΧΙΣΤΟΥ ΑΡΙΘΜΟΥ ΔΕΙΓΜΑΤΩΝ, ΓΙΑ ΤΗΝ ΤΙΜΗ 10, ΓΙΑ ΤΗΝ ΤΙΜΗ 15, ΓΙΑ ΤΗΝ ΤΙΜΗ 20, ΓΙΑ ΤΗΝ ΤΙΜΗ 25***

ΧΩΡΟΙ ΣΤΑΘΜΕΥΣΗΣ

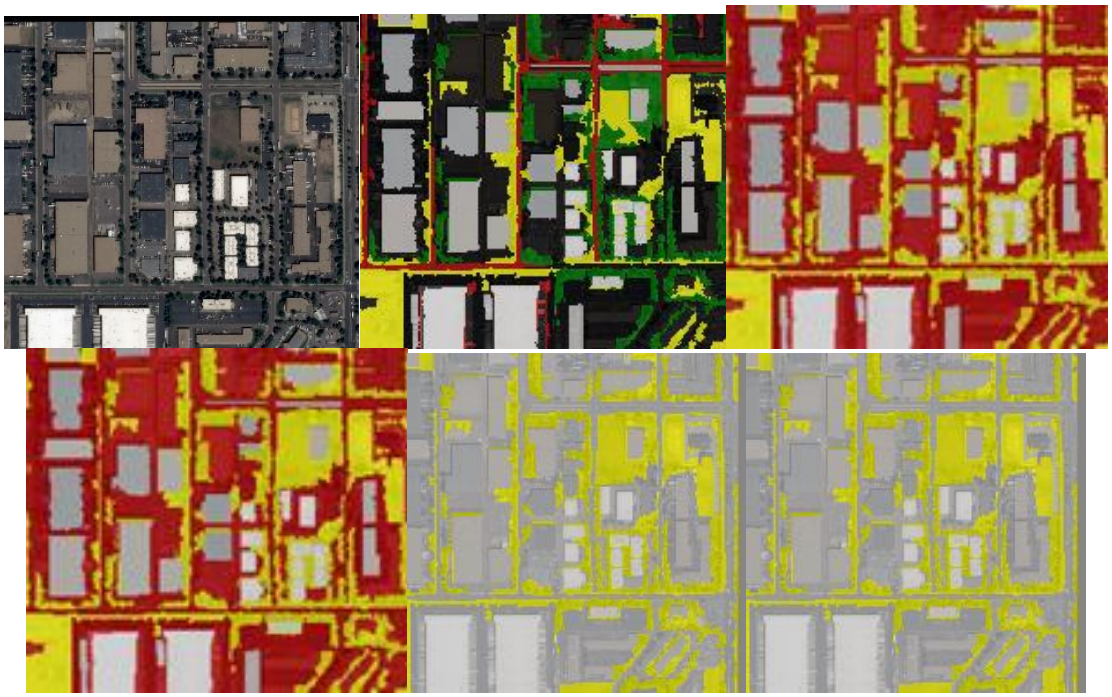
Οι χώροι στάθμευσης δεν εμφανίζονται στους θεματικούς χάρτες που προέκυψαν για τις τιμές 10, 15, 20 και 25 της παραμέτρου του ελάχιστου αριθμού παραμέτρων (Εικόνα 3.50).



ΕΙΚΟΝΑ 3.50: ***ΑΠΟ ΤΗΝ ΑΡΧΗ: 3^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ, ΓΙΑ ΤΗΝ ΤΙΜΗ 5 ΤΟΥ ΕΛΑΧΙΣΤΟΥ ΑΡΙΘΜΟΥ ΔΕΙΓΜΑΤΩΝ, ΓΙΑ ΤΗΝ ΤΙΜΗ 10, ΓΙΑ ΤΗΝ ΤΙΜΗ 15, ΓΙΑ ΤΗΝ ΤΙΜΗ 20, ΓΙΑ ΤΗΝ ΤΙΜΗ 25***

ΑΣΤΙΚΟ ΠΡΑΣΙΝΟ

Η θεματική κατηγορία του αστικού πρασίνου δεν εμφανίζεται στους θεματικούς χάρτες που προέκυψαν για τις τιμές 10, 15, 20 και 25 της παραμέτρου του ελάχιστου αριθμού παραμέτρων (Εικόνα 3.51).



ΕΙΚΟΝΑ 3.51: ***ΑΠΟ ΤΗΝ ΑΡΧΗ: 4^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ, ΓΙΑ ΤΗΝ ΤΙΜΗ 5 ΤΟΥ ΕΛΑΧΙΣΤΟΥ ΑΡΙΘΜΟΥ ΔΕΙΓΜΑΤΩΝ, ΓΙΑ ΤΗΝ ΤΙΜΗ 10, ΓΙΑ ΤΗΝ ΤΙΜΗ 15, ΓΙΑ ΤΗΝ ΤΙΜΗ 20, ΓΙΑ ΤΗΝ ΤΙΜΗ 25***

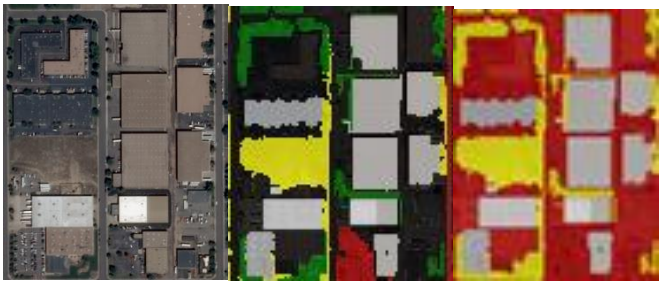
ΆΓΟΝΟ ΈΔΑΦΟΣ

Η ρύθμιση της παραμέτρου ελάχιστος αριθμός δειγμάτων δεν επέφερε αλλαγές στη θεματική κατηγορία του Άγονου Εδάφους (Εικόνα 3.52).

ΕΙΚΟΝΑ 3.52: ***ΑΠΟ ΤΗΝ ΑΡΧΗ: 5^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ, ΓΙΑ ΤΗΝ ΤΙΜΗ 5 ΤΟΥ ΕΛΑΧΙΣΤΟΥ ΑΡΙΘΜΟΥ ΔΕΙΓΜΑΤΩΝ, ΓΙΑ ΤΗΝ ΤΙΜΗ 10, ΓΙΑ ΤΗΝ ΤΙΜΗ 15, ΓΙΑ ΤΗΝ ΤΙΜΗ 20, ΓΙΑ ΤΗΝ ΤΙΜΗ 25***

ΠΟΣΟΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ

ΕΛΑΧΙΣΤΟΣ ΑΡΙΘΜΟΣ ΔΕΙΓΜΑΤΩΝ: 5,10, 15



ΕΙΚΟΝΑ 3.53: ΑΠΟ ΑΡΙΣΤΕΡΑ: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΙΜΗ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ 5 ΚΑΙ 10, 15

Βάσει της Εικόνα 3.53 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.7, Πίνακας 3.8)

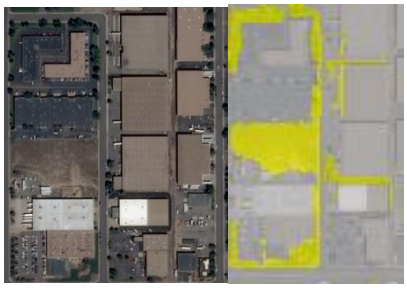
ΠΙΝΑΚΑΣ 3.7*: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (3Η ΔΟΚΙΜΗ).*

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	11	2	5

ΠΙΝΑΚΑΣ 3.8: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (3Η ΔΟΚΙΜΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	68,75%	84,62%	61,11%	31,25%	12,50%

ΕΛΑΧΙΣΤΟΣ ΑΡΙΘΜΟΣ ΔΕΙΓΜΑΤΩΝ: 20, 25



ΕΙΚΟΝΑ 3.54: ΑΠΟ ΑΡΙΣΤΕΡΑ: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΙΜΗ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ 20 ΚΑΙ 25

Βάσει της Εικόνα 3.54 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.9, Πίνακας 3.10)

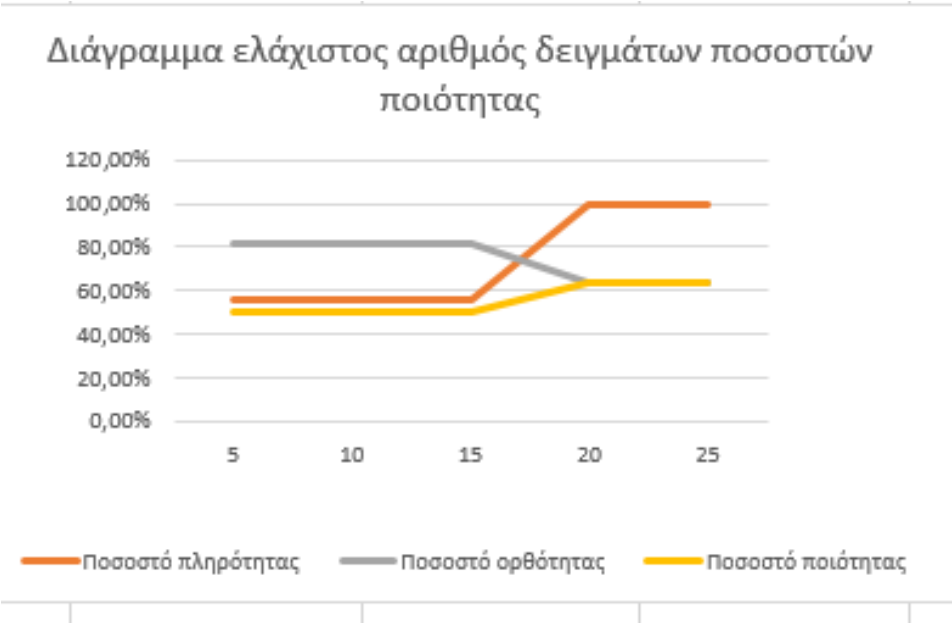
ΠΙΝΑΚΑΣ 3.9*: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (3Η ΔΟΚΙΜΗ).*

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	16	9	0

ΠΙΝΑΚΑΣ 3.10: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (3Η ΔΟΚΙΜΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	100,00%	64,00%	64,00%	0,00%	56,25%

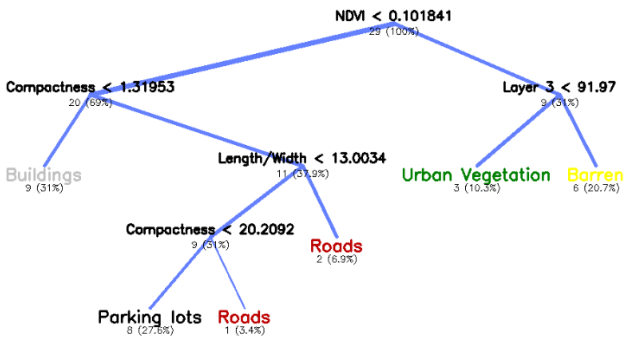
Στην Εικόνα 3.55 εμφανίζεται το Διάγραμμα ελάχιστου αριθμού δειγμάτων ποσοστών ποιότητας.



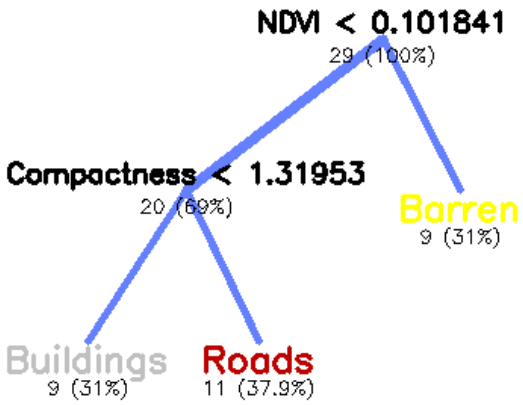
ΕΙΚΟΝΑ 3.55: ΔΙΑΓΡΑΜΜΑ ΕΛΑΧΙΣΤΟΥ ΑΡΙΘΜΟΥ ΔΕΙΓΜΑΤΩΝ ΠΟΣΟΣΤΩΝ ΠΟΙΟΤΗΤΑΣ

Στις ακόλουθες εικόνες (Εικόνα 3.56- Εικόνα 3.58) εμφανίζονται τα δέντρα απόφασης τα οποία κατασκευάστηκαν για τις διαφορετικές τιμές της παραμέτρου ελάχιστος αριθμός δειγμάτων. Τα συμπεράσματα τα οποία αντλούνται έπειτα από προσεκτική παρατήρηση των μοντέλων είναι τα ακόλουθα:

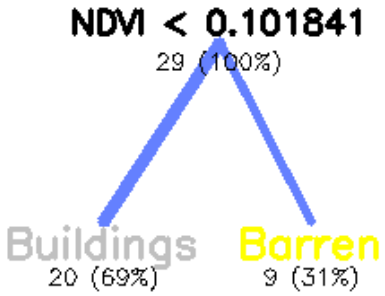
- Το μοντέλο το οποίο κατασκευάστηκε ελάχιστο αριθμό δειγμάτων ίσο με 5 είναι εκείνο των προκαθορισμένων παραμέτρων
- Η αύξηση της τιμής της συγκεκριμένης παραμέτρου οδήγησε σε μικρότερα σε μέγεθος και συνεπώς υποπροσαρμοσμένα δέντρα απόφασης



ΕΙΚΟΝΑ 3.56: ΔΕΝΤΡΟ ΑΠΟΦΑΣΗΣ ΓΙΑ ΕΛΑΧΙΣΤΟ ΑΡΙΘΜΟ ΔΕΙΓΜΑΤΩΝ 5



ΕΙΚΟΝΑ 3.57: ΔΕΝΤΡΟ ΑΠΟΦΑΣΗΣ ΓΙΑ ΕΛΑΧΙΣΤΟ ΑΡΙΘΜΟ ΔΕΙΓΜΑΤΩΝ 10 ΚΑΙ 15



ΕΙΚΟΝΑ 3.58: ΔΕΝΤΡΟ ΑΠΟΦΑΣΗΣ ΓΙΑ ΕΛΑΧΙΣΤΟ ΑΡΙΘΜΟ ΔΕΙΓΜΑΤΩΝ 20 ΚΑΙ 2 5

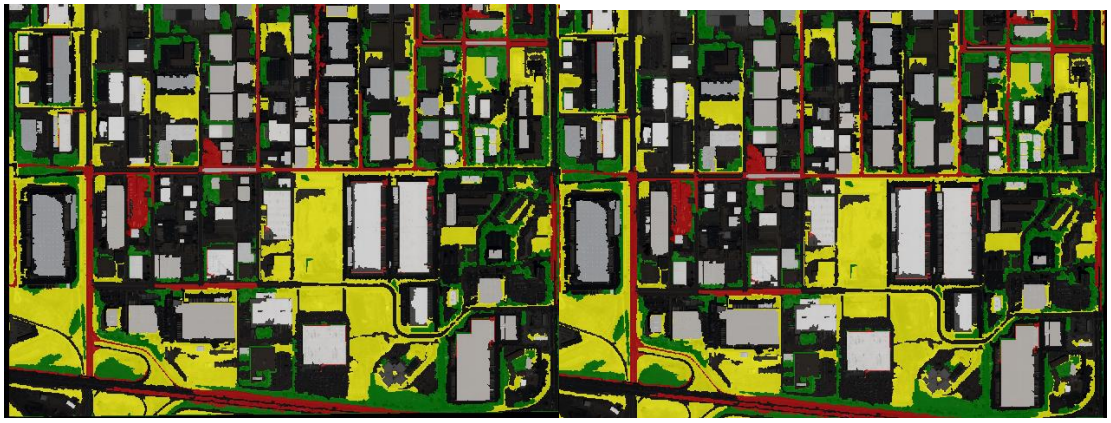
3.6.4 Δοκιμή 4 (Χρήση αντικαταστατών)

Μέσω της τρίτης δοκιμής διερευνήθηκε κατά πόσον η χρήση αντικαταστατών επηρεάζει την ποιότητα της ταξινόμησης:

- Βάθος δέντρου (Depth): 0
- Ελάχιστος αριθμός δειγμάτων (Min sample count): 0
- **Χρήση αντικαταστατών (Use surrogates): Όχι (No), Ναι (Yes)**
- Μέγιστος αριθμός κατηγοριών (Max categories): 16
- Cross Validation folds: 3
- Use 1 SE rule: Όχι (No)
- Αφαίρεση των κλαδεμένων κλαδιών (Truncate pruned trees): Ναι (Yes)

Σχολιασμός αποτελεσμάτων

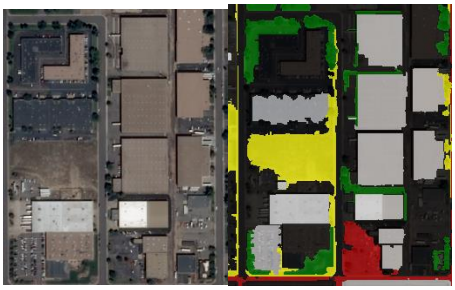
Στην Εικόνα 3.59 εμφανίζεται το αποτέλεσμα εφαρμογής του αλγορίθμου των δέντρων απόφασης για τη χρήση και μη αντικαταστατών. Είναι εμφανές πως η ρύθμιση της συγκεκριμένης μεταβλητής δεν επηρέασε το αποτέλεσμα της ταξινόμησης.



ΕΙΚΟΝΑ 3.59: ΑΠΟΤΕΛΕΣΜΑ ΕΦΑΡΜΟΓΗΣ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΓΙΑ ΤΙΜΕΣ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΧΡΗΣΗ ΑΝΤΙΚΑΤΑΣΤΑΤΩΝ ΟΧΙ (ΑΡΙΣΤΕΡΑ), ΝΑΙ (ΔΕΞΙΑ)

Ποσοτική αξιολόγηση αποτελεσμάτων

ΧΡΗΣΗ ΑΝΤΙΚΑΤΑΣΤΑΤΩΝ: ΝΑΙ



ΕΙΚΟΝΑ 3.60: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΧΡΗΣΗ ΑΝΤΙΚΑΤΑΣΤΑΤΩΝ

Βάσει της Εικόνα 3.60 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.11, Πίνακας 3.12)

ΠΙΝΑΚΑΣ 3.11: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (4Η ΔΟΚΙΜΗ).

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	11	2	5

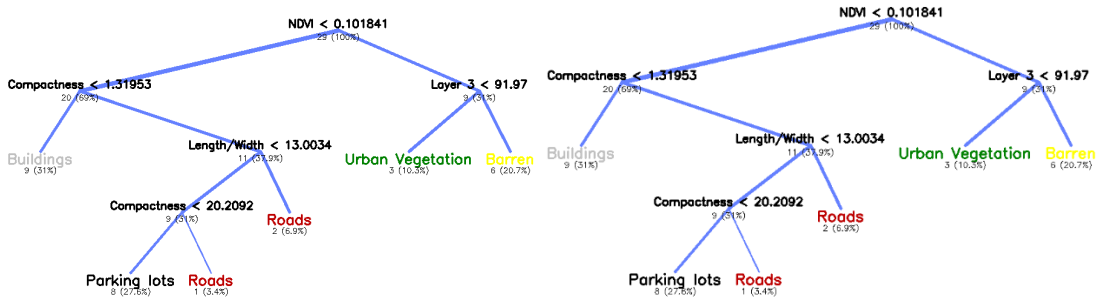
ΠΙΝΑΚΑΣ 3.12: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (4Η ΔΟΚΙΜΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	68,75%	84,62%	61,11%	31,25%	12,50%

Στην Εικόνα 3.61 εμφανίζεται το Διάγραμμα χρήσης αντικαταστατών ποσοστών ποιότητας

ΕΙΚΟΝΑ 3.61: ΔΙΑΓΡΑΜΜΑ ΧΡΗΣΗΣ ΑΝΤΙΚΑΤΑΣΤΑΤΩΝ ΠΟΣΟΣΤΩΝ ΠΟΙΟΤΗΤΑΣ

Στην Εικόνα 3.62 εμφανίζονται τα δέντρα απόφασης στη περίπτωση χρήσης και μη αντικαταστατών. Παρατηρείται πως τα μοντέλα αυτά είναι πανομοιότυπα και ως εκ τούτου είναι αναμενόμενο πως τα αποτελέσματα ταξινόμησης των τελευταίων δε διαφέρουν μεταξύ τους.



ΕΙΚΟΝΑ 3.62: ΑΡΙΣΤΕΡΑ: ΔΕΝΤΡΟ ΑΠΟΦΑΣΗΣ ΓΙΑ ΠΡΟΚΑΘΟΡΙΣΜΕΝΕΣ ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΔΕΞΙΑ: ΔΕΝΤΡΟ ΑΠΟΦΑΣΗΣ ΓΙΑ ΧΡΗΣΗ ΑΝΤΙΚΑΤΑΣΤΑΤΩΝ

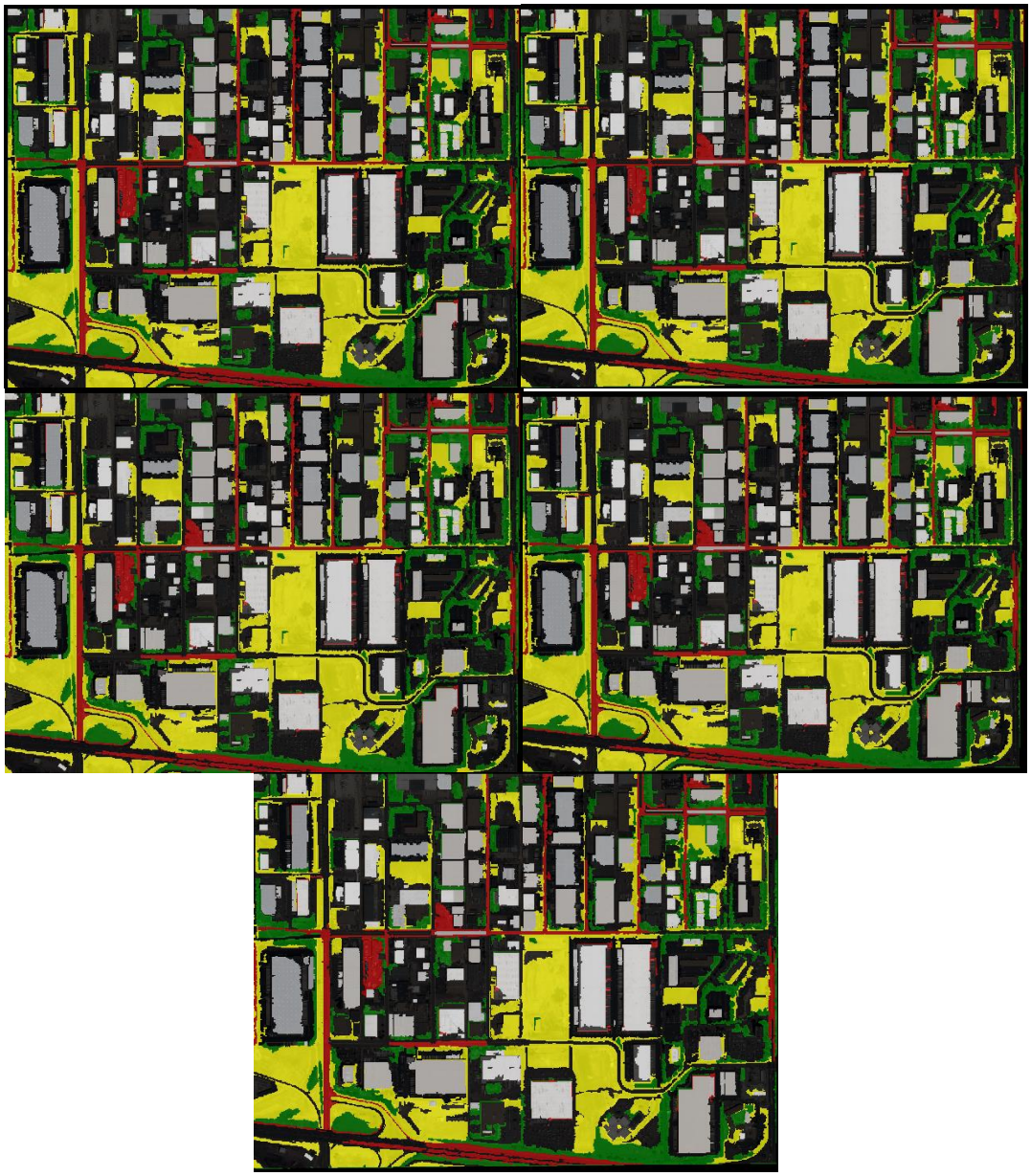
3.6.5 Δοκιμή 5 (Μέγιστος αριθμός κατηγοριών)

Στα πλαίσια της τέταρτης δοκιμής διερευνήθηκε η επιρροή της μεταβλητής του μέγιστου αριθμού κατηγοριών στην ποιότητα της ταξινόμησης. Αναλυτικά, οι τιμές των παραμέτρων ήταν οι ακόλουθες:

- Βάθος δέντρου (Depth): 0
- Ελάχιστος αριθμός δειγμάτων (Min sample count): 0
- Χρήση αντικαταστατών (Use surrogates): Όχι (No)
- **Μέγιστος αριθμός κατηγοριών (Max categories): 2, 8, 16, 30, 100**
- Cross Validation folds: 3
- Use 1 SE rule: Όχι (No)
- Αφαίρεση των κλαδεμένων κλαδιών (Truncate pruned trees): Ναι (Yes)

Σχολιασμός αποτελεσμάτων

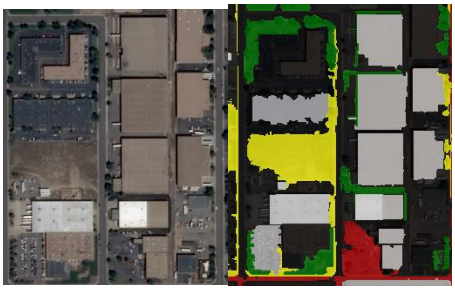
Στην Εικόνα 3.63 εμφανίζεται το αποτέλεσμα εφαρμογής του αλγορίθμου των δέντρων απόφασης για διαφορετικές τιμές του μέγιστου αριθμού κατηγοριών. Ομοίως με τις προηγούμενες δοκιμές είναι εμφανές πως η ρύθμιση της συγκεκριμένης μεταβλητής δεν επηρέασε το αποτέλεσμα της ταξινόμησης.



Εικονα 3.63: Αποτελεσμα εφαρμογης του αλγοριθμου των δεντρων αποφασης για τιμες της παραμετρου των μεγιστων κατηγοριων 2 (πανω αριστερα), 8 (πανω δεξια), 16 (αριστερα), 32 (δεξια), 100 (κατω)

Ποσοτική αξιολόγηση αποτελεσμάτων

ΜΕΓΙΣΤΟΣ ΑΡΙΘΜΟΣ ΚΑΤΗΓΟΡΙΩΝ: 2, 8, 16, 30, 100



Εικονα 3.64: Χαρακτηριστικο αποσπασμα αστικης δομησης απο την περιοχη μελετης για μεγιστο αριθμο κατηγοριων 2, 8, 16, 30, 100

Βάσει της Εικόνα 3.64 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.13, Πίνακας 3.14)

ΠΙΝΑΚΑΣ 3.13: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (5Η ΔΟΚΙΜΗ).

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	11	2	5

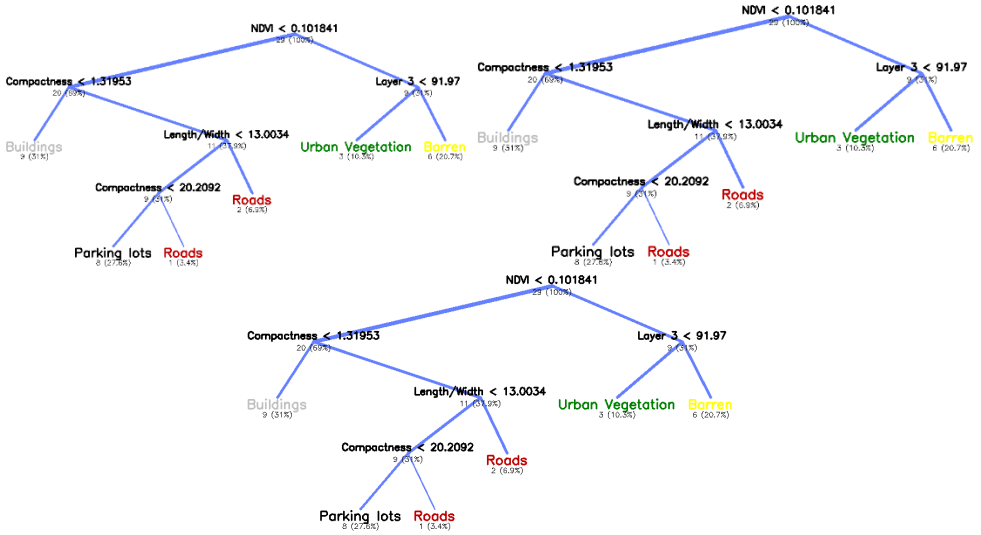
ΠΙΝΑΚΑΣ 3.14: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (5^{Η} ΔΟΚΙΜΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	68,75%	84,62%	61,11%	31,25%	12,50%

Στην Εικόνα 3.65 εμφανίζεται το Διάγραμμα μεγίστου αριθμού κατηγοριών ποσοστών ποιότητας

ΕΙΚΟΝΑ 3.65: ΔΙΑΓΡΑΜΜΑ ΜΕΓΙΣΤΟΥ ΑΡΙΘΜΟΥ ΚΑΤΗΓΟΡΙΩΝ ΠΟΣΟΣΤΩΝ ΠΟΙΟΤΗΤΑΣ

Στην Εικόνα 3.66 εμφανίζονται ενδεικτικά τα δέντρα απόφασης για τα πλήθη κατηγοριών 2, 8 και 16. Τα συγκεκριμένα είναι πανομοιότυπα και ως εκ τούτου επιβεβαιώνεται πως η ρύθμιση της παραμέτρου αυτής δε μετέβαλε το αποτέλεσμα της ταξινόμησης.



ΕΙΚΟΝΑ 3.66: ΔΕΝΤΡΟ ΑΠΟΦΑΣΗΣ ΓΙΑ ΜΕΓΙΣΤΟ ΑΡΙΘΜΟ ΚΑΤΗΓΟΡΙΩΝ ΙΣΟ ΜΕ 2 (ΠΑΝΩ ΑΡΙΣΤΕΡΑ), 8 (ΠΑΝΩ ΔΕΞΙΑ), 16 (ΚΑΤΩ)

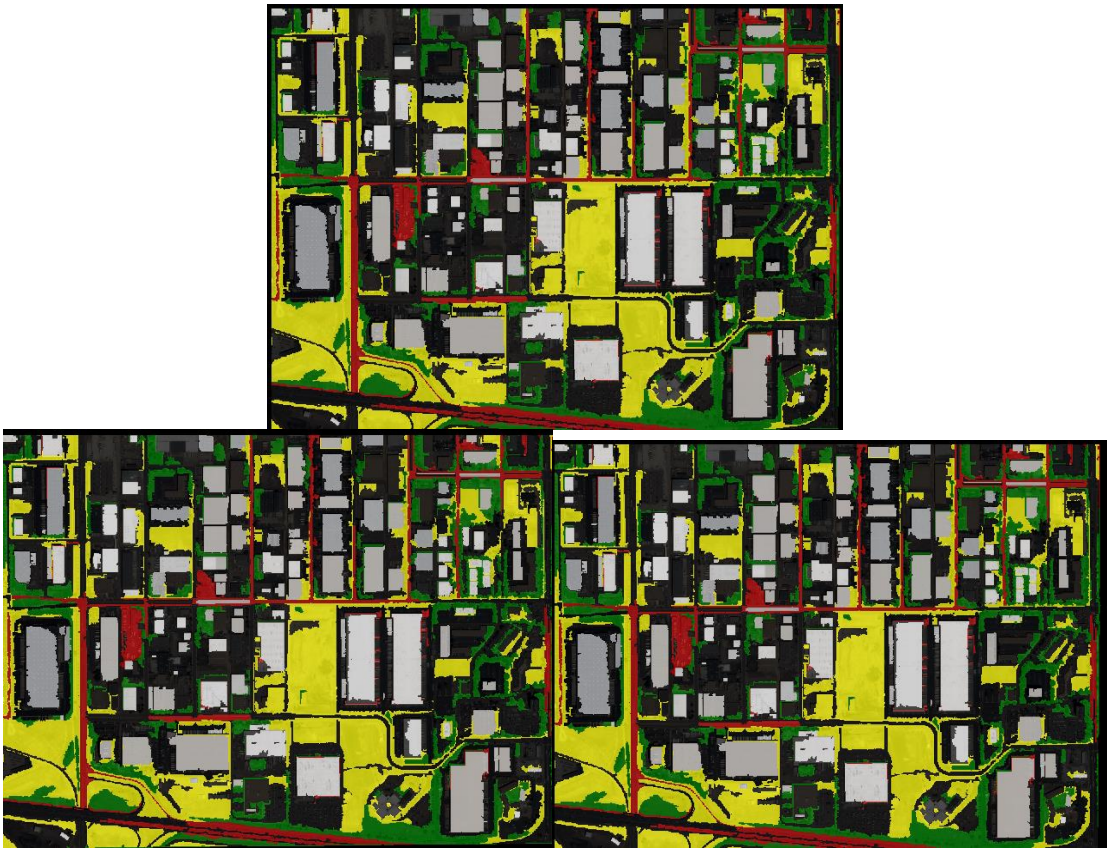
3.6.6 Δοκιμή 6 (Cross validation folds)

Στη δοκιμή αυτή διερευνήθηκε η επιρροή του πλήθους των cross validations. Αναλυτικά, οι τιμές των παραμέτρων ορίστηκαν ως εξής:

- Βάθος δέντρου (Depth): 0
- Ελάχιστος αριθμός δειγμάτων (Min sample count):0
- Χρήση αντικαταστατών (Use surrogates): Όχι (No)
- Μέγιστος αριθμός κατηγοριών (Max categories): 2, 8, 16, 30, 100
- **Cross Validation folds: 3, 6, 9**
- Use 1 SE rule: Όχι (No)
- Αφαίρεση των κλαδεμένων κλαδιών (Truncate pruned trees): Ναι (Yes)

Σχολιασμός αποτελεσμάτων

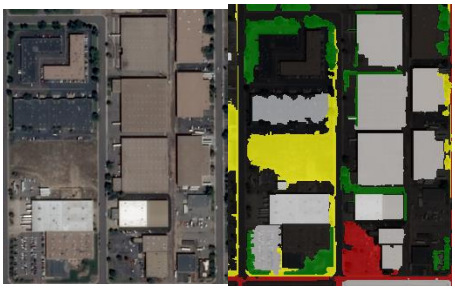
Στην Εικόνα 3.67 εμφανίζεται το αποτέλεσμα για διαφορετικές τιμές των cross validation folds. Ομοίως με τις προηγούμενες δοκιμές η ρύθμιση της εν λόγω παραμέτρου δεν επηρέασε το αποτέλεσμα της ταξινόμησης.



ΕΙΚΟΝΑ 3.67: ΑΠΟΤΕΛΕΣΜΑ ΕΦΑΡΜΟΓΗΣ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΓΙΑ ΤΙΜΕΣ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΤΩΝ CROSS VALIDATIONS 3, 6, 9

Ποσοτική αξιολόγηση αποτελεσμάτων

ΑΡΙΘΜΟΣ CROSS VALIDATION FOLDS: 6, 9



ΕΙΚΟΝΑ 3.68: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΗΣ ΠΑΡΑΜΕΤΡΟΥ ΤΩΝ CROSS VALIDATIONS 6, 9

Βάσει της Εικόνα 3.68 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.15, Πίνακας 3.16)

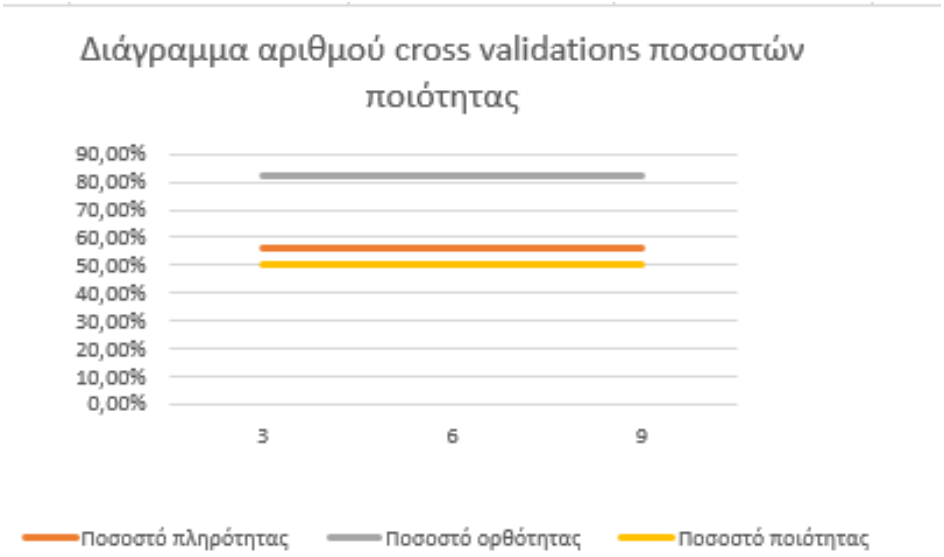
ΠΙΝΑΚΑΣ 3.15: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (6Η ΔΟΚΙΜΗ).

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	11	2	

ΠΙΝΑΚΑΣ 3.16: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (6Η ΔΟΚΙΜΗ).

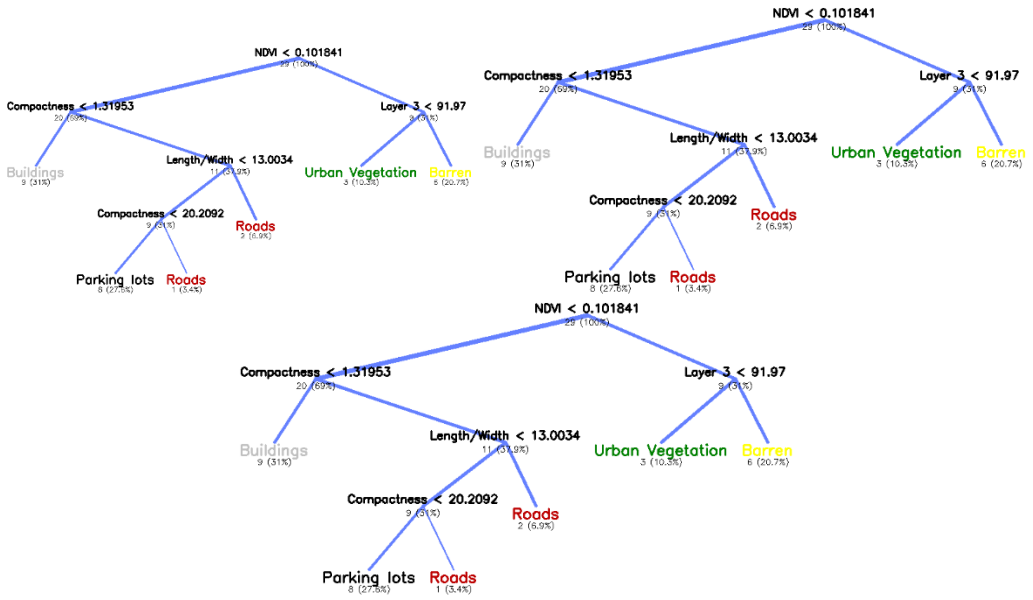
	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	68,75%	84,62%	61,11%	31,25%	12,50%

Στην Εικόνα 3.69 εμφανίζεται το Διάγραμμα αριθμού cross validations ποσοστών ποιότητας



ΕΙΚΟΝΑ 3.69: ΔΙΑΓΡΑΜΜΑ ΜΕΓΙΣΤΟΥ ΑΡΙΘΜΟΥ CROSS VALIDATIONS ΠΟΣΟΣΤΩΝ ΠΟΙΟΤΗΤΑΣ

Στην Εικόνα 3.70 εμφανίζονται τα δέντρα απόφασης για τις διαφορετικές τιμές του αριθμού των cross validations. Τα μοντέλα αυτά δε διαφέρουν μεταξύ τους και για το λόγο αυτό έδωσαν ακριβώς τα ίδια αποτελέσματα ταξινόμησης.



ΕΙΚΟΝΑ 3.70: ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ ΓΙΑ ΜΕΓΙΣΤΟ ΑΡΙΘΜΟ CROSS VALIDATIONS ΙΣΟ ΜΕ 3 (ΠΑΝΩ ΑΡΙΣΤΕΡΑ), 6 (ΠΑΝΩ ΔΕΞΙΑ), 9 (ΚΑΤΩ)

3.6.7 Δοκιμή 7 (Χρήση 1 SE κανόνα)

Στην παρούσα δοκιμή έγινε ρύθμιση των τιμών των παραμέτρων ως εξής:

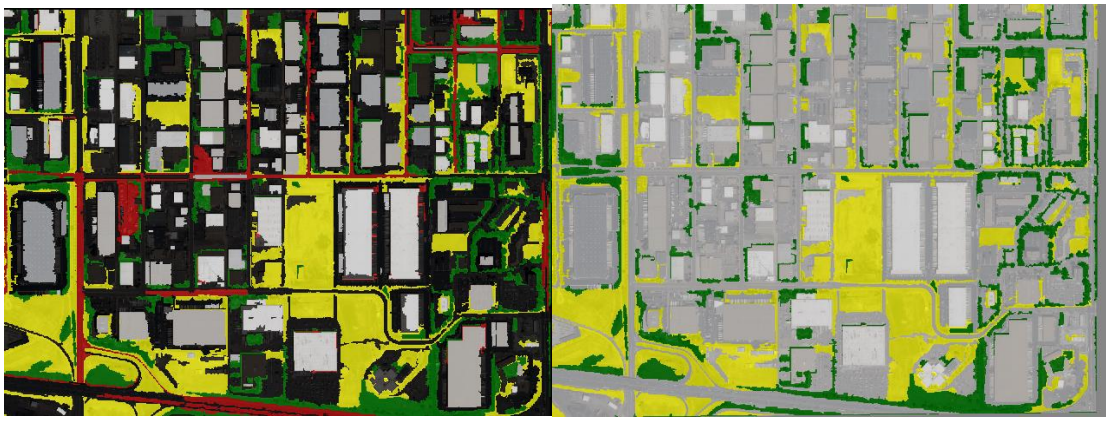
- Βάθος δέντρου (Depth): 0
- Ελάχιστος αριθμός δειγμάτων (Min sample count): 0
- Χρήση αντικαταστατών (Use surrogates): Όχι (No)
- Μέγιστος αριθμός κατηγοριών (Max categories): 16
- Cross Validation folds: 3
- **Use 1 SE rule: Όχι, (No), Ναι (Yes)**
- Αφαίρεση των κλαδεμένων κλαδιών (Truncate pruned trees):Ναι (Yes).

Σχολιασμός αποτελεσμάτων

Στην Εικόνα 3.71 εμφανίζεται το αποτέλεσμα της ταξινόμησης για την ένατη δοκιμή. Είναι εμφανές πως στην προκειμένη περίπτωση ο θεματικός χάρτης είναι σε σχέση με τις προηγούμενες δοκιμές. Οι εμφανιζόμενες κλάσεις στην προκειμένη περίπτωση είναι οι ακόλουθες:

- Κτίρια
- Αστικό Πράσινο
- Άγονο Έδαφος

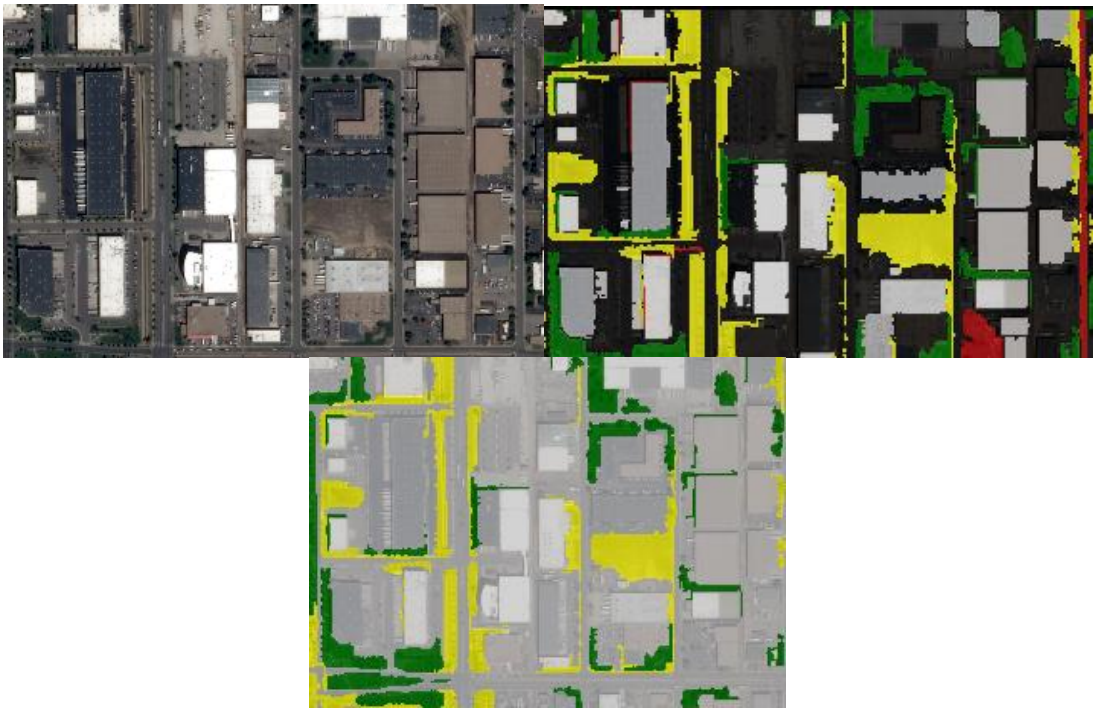
Οι υπόλοιπες θεματικές κατηγορίες δε συμπεριλαμβάνονται στο χάρτη που προέκυψε.



ΕΙΚΟΝΑ 3.71: ΑΠΟΤΕΛΕΣΜΑ ΕΦΑΡΜΟΓΗΣ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΓΙΑ ΤΙΜΕΣ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΧΡΗΣΗ ΚΑΝΟΝΑ SE ΟΧΙ (ΑΡΙΣΤΕΡΑ), ΝΑΙ (ΔΕΞΙΑ)

Κτίρια

Η ρύθμιση της παραμέτρου της χρήσης κανόνα SE σε ναι αύξησε το πλήθος των κτιρίων. Πιο συγκεκριμένα, η κλάση αυτή εμφανίζει πληρότητα 100% καθώς έχουν ταξινομηθεί σε αυτή όλα τα εμφανιζόμενα κτίρια. Στην κατηγορία αυτή, ωστόσο, έχει προστεθεί από τον αλγόριθμο μεγάλος αριθμός αντικειμένων τα οποία στην πραγματικότητα ανήκουν σε εκείνες των δρόμων και των χώρων στάθμευσης. Βάσει αυτού προκύπτει πως το κριτήριο της ορθότητας δεν ικανοποιείται στην προκειμένη περίπτωση (Εικόνα 3.72).



ΕΙΚΟΝΑ 3.72: ΑΠΟ ΤΗΝ ΑΡΧΗ: 1^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ, ΓΙΑ ΤΗΝ ΤΙΜΗ ΟΧΙ ΣΤΗ ΧΡΗΣΗ ΕΝΟΣ ΚΑΝΟΝΑ SE, ΓΙΑ ΤΗΝ ΤΙΜΗ ΝΑΙ ΣΤΗ ΧΡΗΣΗ ΕΝΟΣ ΚΑΝΟΝΑ SE

Δρόμοι

Στο θεματικό χάρτη της Εικόνα 3.71 δε συμπεριλαμβάνονται αντικείμενα τα οποία ανήκουν στη συγκεκριμένη κλάση. Ως εκ τούτου η πληρότητα, η ορθότητα και η ποιότητα των δρόμων είναι μηδενική.

Χώροι στάθμευσης

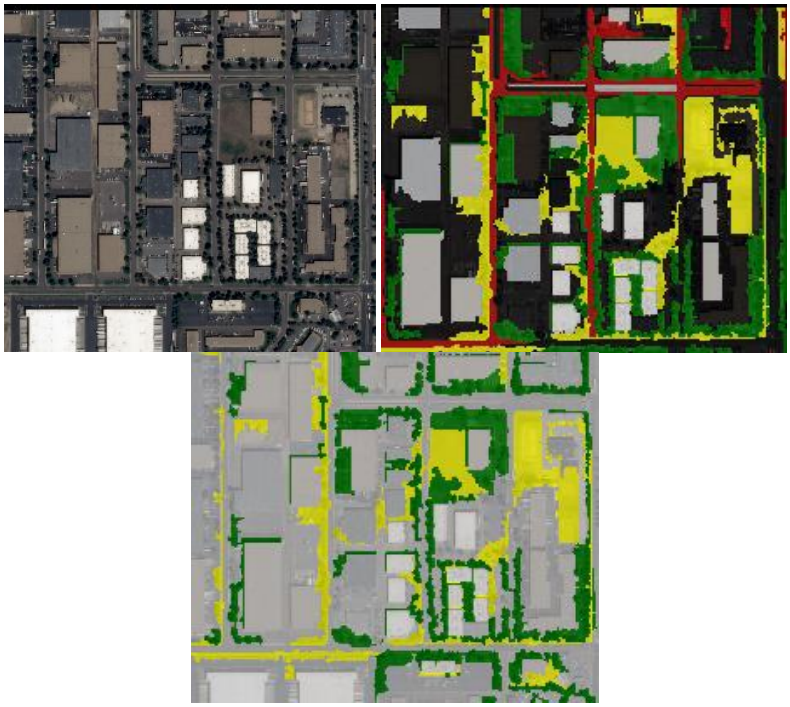
Η συγκεκριμένη κλάση δε συμπεριλαμβάνεται στο αποτέλεσμα της ταξινόμησης του αλγορίθμου των δέντρων απόφασης στην περίπτωση της χρήσης ενός SE κανόνα. Συνεπώς, δεν ικανοποιείται κανένα από τα κριτήρια ποιότητας για τη θεματική κατηγορία των δρόμων.

Αστικό πράσινο

Η συγκεκριμένη θεματική κατηγορία εμφανίζει υψηλά ποσοστά ορθότητας στην περίπτωση χρήσης ενός SE κανόνα. Το παραπάνω προκύπτει από το γεγονός πως το σύνολο των αντικειμένων που έχουν καταταχθεί από τον αλγόριθμο στην κλάση αυτή είναι πράγματι χώροι αστικού πρασίνου.

Προβλήματα, ωστόσο, εντοπίζονται σε ό,τι αφορά το κριτήριο της πληρότητας. Το παραπάνω προκύπτει από το γεγονός πως ο αριθμός των αντικειμένων που ανήκουν στην κατηγορία του αστικού πρασίνου και έχουν εσφαλμένα ταξινομηθεί από τον αλγόριθμο στις κλάσεις του άγονου εδάφους και των κτιρίων είναι αρκετά μεγάλος.

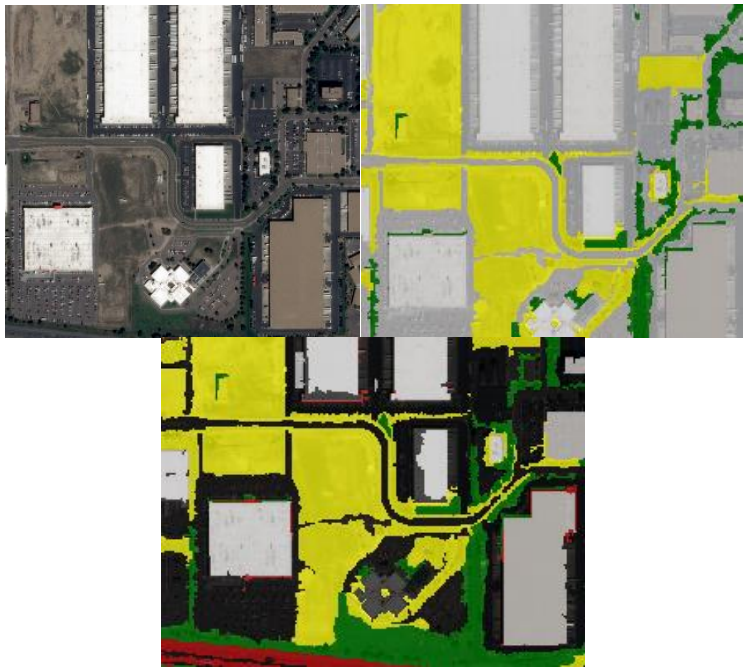
Γενικά, παρατηρείται πως το αποτέλεσμα της ταξινόμησης για τη θεματική κατηγορία του Αστικού Πρασίνου είναι ακριβώς το ίδιο με εκείνο των προηγούμενων δοκιμών (Εικόνα 3.73).



ΕΙΚΟΝΑ 3.73: ΑΠΟ ΤΗΝ ΑΡΧΗ: 2^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ, ΓΙΑ ΤΗΝ ΤΙΜΗ ΟΧΙ ΣΤΗ ΧΡΗΣΗ ΕΝΟΣ ΚΑΝΟΝΑ SE, ΓΙΑ ΤΗΝ ΤΙΜΗ ΝΑΙ ΣΤΗ ΧΡΗΣΗ ΕΝΟΣ ΚΑΝΟΝΑ SE

Άγονο Έδαφος

Η θεματική κατηγορία του Άγονου Εδάφους ικανοποιεί τα κριτήρια τόσο της πληρότητας όσο και της ορθότητας. Το αποτέλεσμα της ταξινόμησης για τη συγκεκριμένη θεματική κατηγορία είναι σχεδόν πανομοιότυπο με εκείνο των προηγούμενων δοκιμών (Εικόνα 3.74).



ΕΙΚΟΝΑ 3.74: ΑΠΟ ΤΗΝ ΑΡΧΗ: 3^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ, ΓΙΑ ΤΗΝ ΤΙΜΗ ΟΧΙ ΣΤΗ ΧΡΗΣΗ ΕΝΟΣ ΚΑΝΟΝΑ SE, ΓΙΑ ΤΗΝ ΤΙΜΗ ΝΑΙ ΣΤΗ ΧΡΗΣΗ ΕΝΟΣ ΚΑΝΟΝΑ SE

ΕΙΚΟΝΑ 3.75: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΗΝ ΤΙΜΗ ΝΑΙ ΣΤΗ ΧΡΗΣΗ ΕΝΟΣ ΚΑΝΟΝΑ SE

Βάσει του χαρακτηριστικού αποσπάσματος της Εικόνα 3.75 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.17, Πίνακας 3.18).

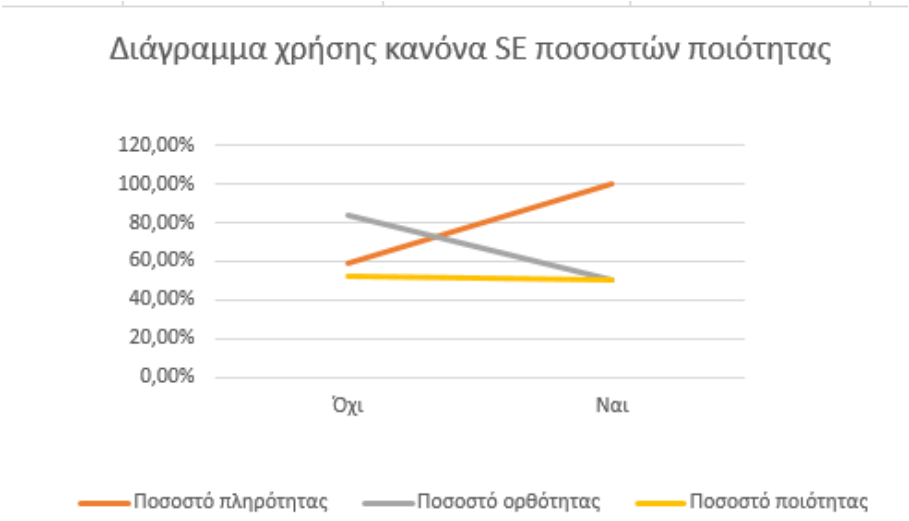
ΠΙΝΑΚΑΣ 3.17: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (7Η ΔΟΚΙΜΗ).

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	16	16	0

ΠΙΝΑΚΑΣ 3.18: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (7Η ΔΟΚΙΜΗ).

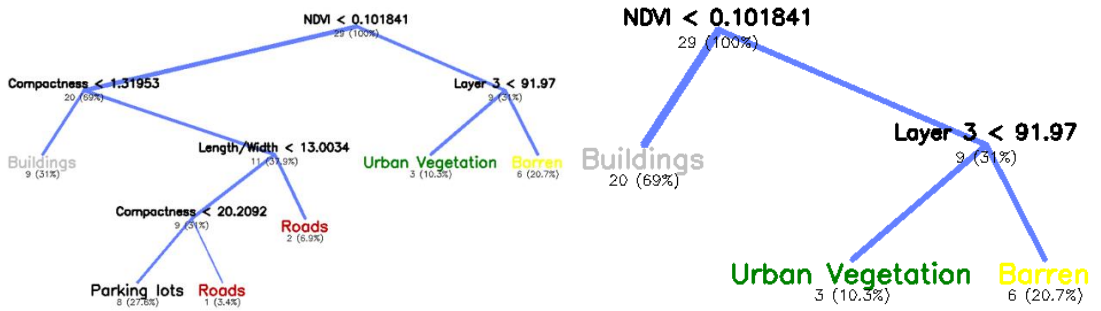
	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	100,00%	50,00%	50,00%	0,00%	100,00%

Στην Εικόνα 3.76 εμφανίζεται Διάγραμμα χρήσης ενός κανόνα SE ποσοστών ποιότητας.



ΕΙΚΟΝΑ 3.76: ΔΙΑΓΡΑΜΜΑ ΧΡΗΣΗΣ ΚΑΝΟΝΑ 1-SE ΠΟΣΟΣΤΩΝ ΠΟΙΟΤΗΤΑΣ

Στην Εικόνα 3.77 εμφανίζεται το δέντρο απόφασης για τις διαφορετικές τιμές της παραμέτρου χρήση κανόνα 1 - SE.



ΕΙΚΟΝΑ 3.77: ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ ΓΙΑ ΤΙΜΕΣ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΧΡΗΣΗ 1-SE ΌΧΙ (ΑΡΙΣΤΕΡΑ) ΝΑΙ (ΔΕΞΙΑ)

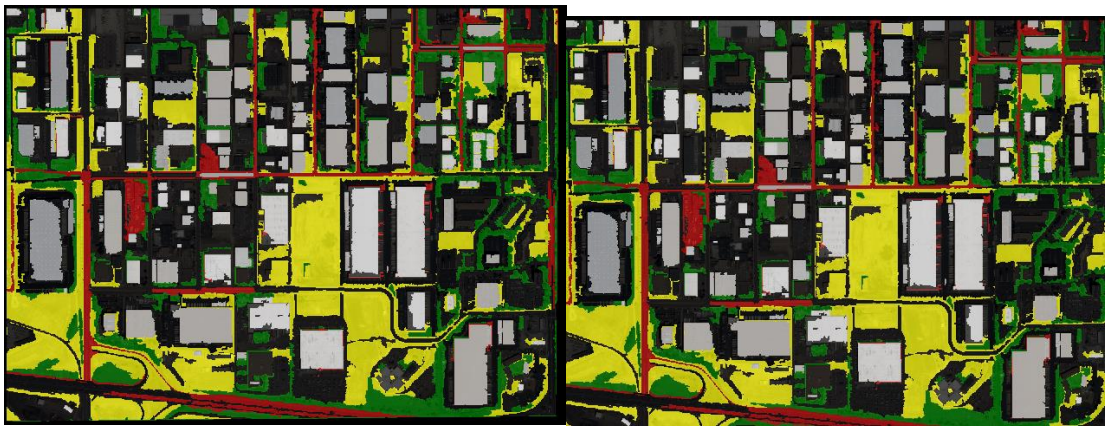
3.6.8 Δοκιμή 8 (Αφαίρεση των κλαδεμένων κλαδιών)

Στη δοκιμή αυτή διερευνήθηκε η επιρροή του κοψίματος των κλαδεμένων κλαδιών στο αποτέλεσμα της ταξινόμησης. Αναλυτικά, οι τιμές των παραμέτρων ορίστηκαν ως εξής:

- Βάθος δέντρου (Depth): 0
- Ελάχιστος αριθμός δειγμάτων (Min sample count): 0
- Χρήση αντικαταστατών (Use surrogates): Όχι (No)
- Μέγιστος αριθμός κατηγοριών (Max categories): 2, 8, 16, 30, 100
- Cross Validation folds: 3, 6, 9
- Use 1 SE rule: Όχι (No)
- **Αφαίρεση των κλαδεμένων κλαδιών (Truncate pruned trees): Όχι (No), Ναι (Yes)**

Σχολιασμός αποτελεσμάτων

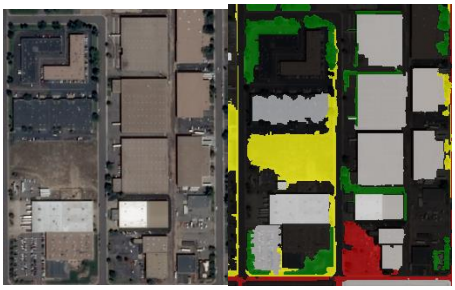
Στην Εικόνα 3.67 εμφανίζεται το αποτέλεσμα για το αφαίρεση και μη των κλαδεμένων κλαδιών. Η ρύθμιση της συγκεκριμένης παραμέτρου δεν επηρέασε το αποτέλεσμα της ταξινόμησης.



ΕΙΚΟΝΑ 3.78: ΑΠΟΤΕΛΕΣΜΑ ΕΦΑΡΜΟΓΗΣ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΓΙΑ ΑΦΑΙΡΕΣΗ ΚΑΙ ΜΗ ΤΩΝ ΚΛΑΔΕΜΕΝΩΝ ΚΛΑΔΙΩΝ

Ποσοτική αξιολόγηση αποτελεσμάτων

ΑΦΑΙΡΕΣΗ ΚΛΑΔΕΜΕΝΩΝ ΚΛΑΔΙΩΝ: ΌΧΙ



ΕΙΚΟΝΑ 3.79: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΑΦΑΙΡΕΣΗ ΚΛΑΔΕΜΕΝΩΝ ΚΛΑΔΙΩΝ ΣΕ ΌΧΙ

Βάσει της Εικόνα 3.79 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.19, Πίνακας 3.20)

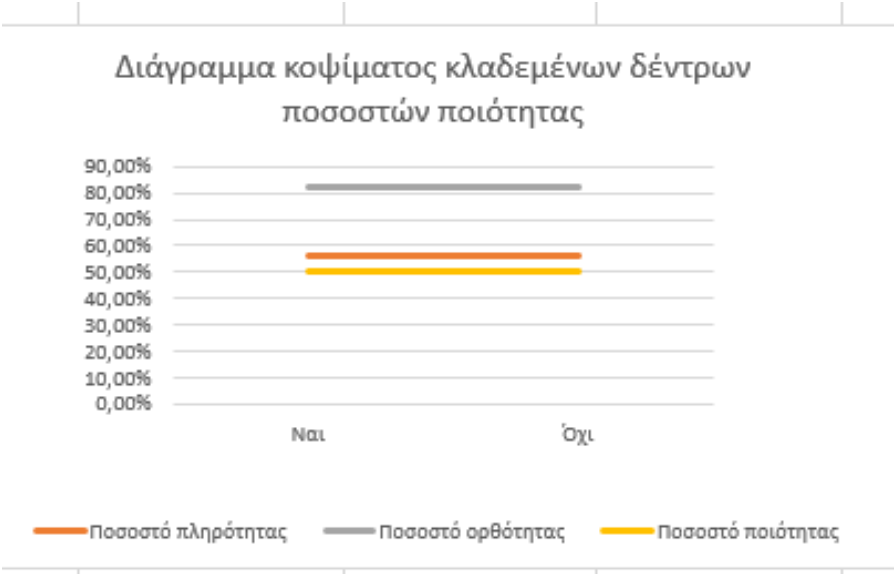
ΠΙΝΑΚΑΣ 3.19: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (8Η ΔΟΚΙΜΗ).

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	11	2	5

ΠΙΝΑΚΑΣ 3.20: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (8Η ΔΟΚΙΜΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	68,75%	84,62%	61,11%	31,25%	12,50%

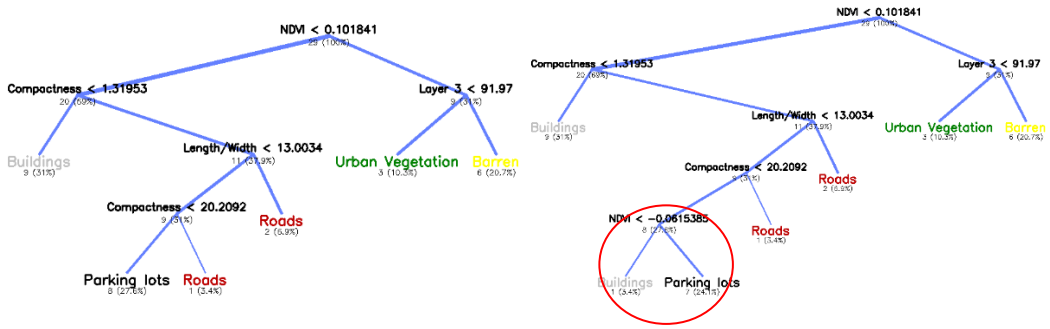
Στην Εικόνα 3.80 εμφανίζεται το διάγραμμα για την αφαίρεση και μη των κλαδεμένων κλαδιών.



ΕΙΚΟΝΑ 3.80: *ΔΙΑΓΡΑΜΜΑ ΓΙΑ ΑΦΑΙΡΕΣΗ ΚΑΙ ΜΗ ΤΩΝ ΚΛΑΔΕΜΕΝΩΝ ΚΛΑΔΙΩΝ ΠΟΣΟΣΤΩΝ ΠΟΙΟΤΗΤΑΣ*

Στην Εικόνα 3.81 εμφανίζονται τα δέντρα απόφασης στην περίπτωση κλαδέματος και μη. Είναι εμφανές πως όταν η τιμή της συγκεκριμένης παραμέτρου ρυθμίστηκε σε Ναι, ένα υποδέντρο (το οποίο σημειώνεται με κόκκινο στην εικόνα) του αρχικού δέντρου δεν

εμφανίζεται στο τελικό. Εντύπωση, ωστόσο, προκαλεί πως το γεγονός πως τα δύο παραπάνω μοντέλα ταξινόμησης δίνουν πανομοιότυπα αποτελέσματα. Το παραπάνω επιβεβαιώθηκε από πλήθος δοκιμών για διαφορετικές τιμές της παραμέτρου βάθος.

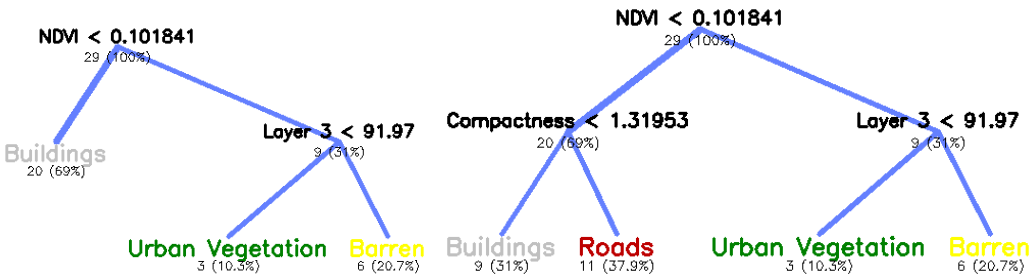


ΕΙΚΟΝΑ 3.81: ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ ΣΕ ΠΕΡΙΠΤΩΣΗ ΚΛΑΔΕΜΑΤΟΣ (ΔΕΞΙΑ) ΚΑΙ ΜΗ (ΑΡΙΣΤΕΡΑ)

Στις ακόλουθες εικόνες (Εικόνα 3.82- Εικόνα 3.88) εμφανίζονται τα δέντρα απόφασης καθώς και τα αποτελέσματα ταξινόμησης για διαφορετικές τιμές της παραμέτρου Βάθος των δέντρων.

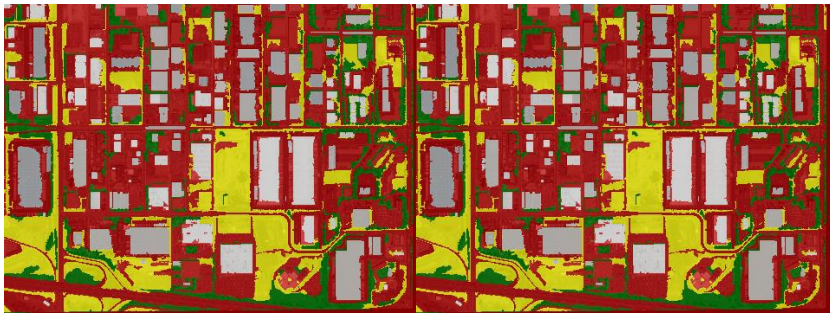
- Για βάθος δέντρου 2:

ΕΙΚΟΝΑ 3.82: *ΑΠΟΤΕΛΕΣΜΑ ΕΦΑΡΜΟΓΗΣ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΓΙΑ ΑΦΑΙΡΕΣΗ ΚΑΙ ΜΗ ΤΩΝ ΚΛΑΔΕΜΕΝΩΝ ΚΛΑΔΙΩΝ (ΤΙΜΗ ΒΑΘΟΥΣ: 2)*

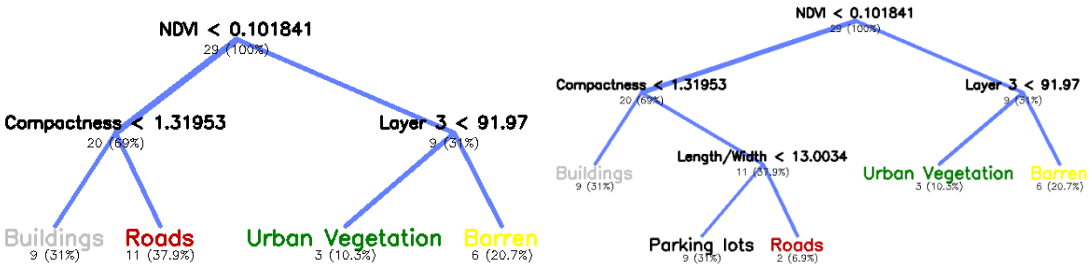


ΕΙΚΟΝΑ 3.83: ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ ΣΕ ΠΕΡΙΠΤΩΣΗ ΚΛΑΔΕΜΑΤΟΣ (ΔΕΞΙΑ) ΚΑΙ ΜΗ (ΑΡΙΣΤΕΡΑ) *(ΤΙΜΗ ΒΑΘΟΥΣ: 2)*

- Τιμή βάθους: 3

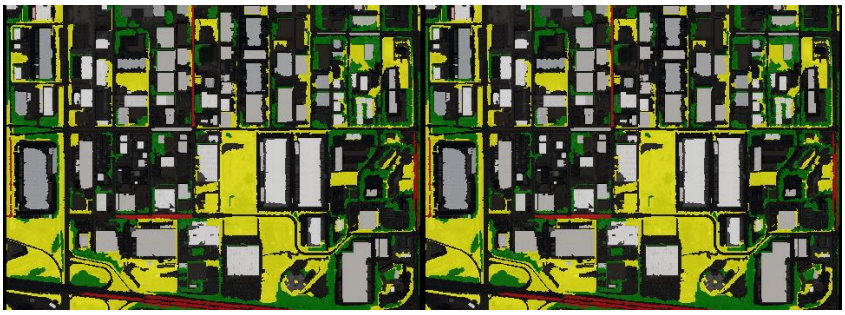


ΕΙΚΟΝΑ 3.84: ***ΑΠΟΤΕΛΕΣΜΑ ΕΦΑΡΜΟΓΗΣ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΓΙΑ ΑΦΑΙΡΕΣΗ ΚΑΙ ΜΗ ΤΩΝ ΚΛΑΔΕΜΕΝΩΝ ΚΛΑΔΙΩΝ (ΤΙΜΗ ΒΑΘΟΥΣ: 3)***

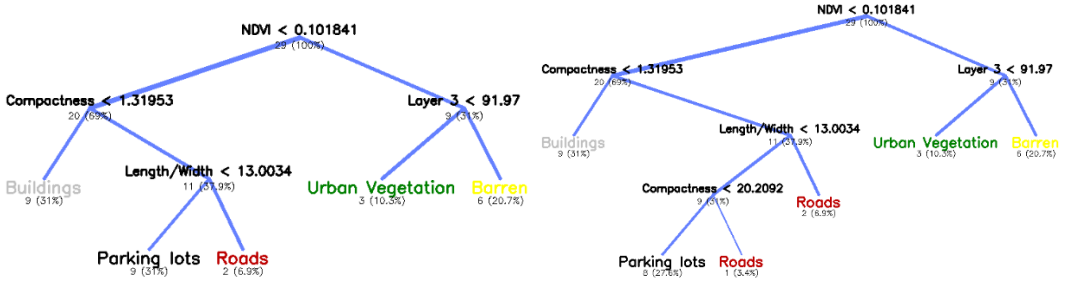


ΕΙΚΟΝΑ 3.85: ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ ΣΕ ΠΕΡΙΠΤΩΣΗ ΚΛΑΔΕΜΑΤΟΣ (ΔΕΞΙΑ) ΚΑΙ ΜΗ (ΑΡΙΣΤΕΡΑ) ***(ΤΙΜΗ ΒΑΘΟΥΣ: 3)***

- Τιμή βάθους: 4

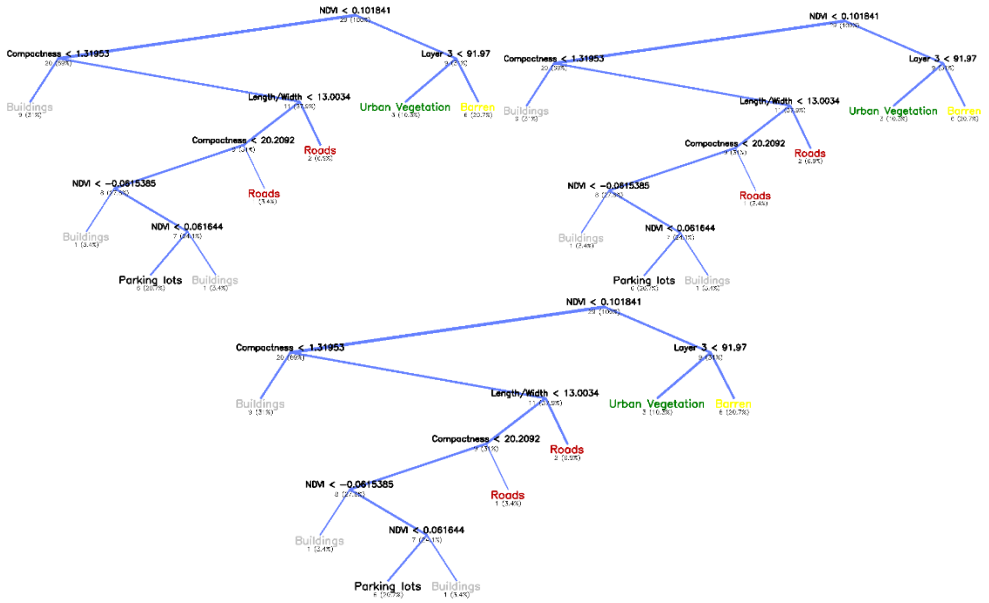


ΕΙΚΟΝΑ 3.86: ***ΑΠΟΤΕΛΕΣΜΑ ΕΦΑΡΜΟΓΗΣ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΓΙΑ ΑΦΑΙΡΕΣΗ ΚΑΙ ΜΗ ΤΩΝ ΚΛΑΔΕΜΕΝΩΝ ΚΛΑΔΙΩΝ (ΤΙΜΗ ΒΑΘΟΥΣ: 4)***



ΕΙΚΟΝΑ 3.87: ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ ΣΕ ΠΕΡΙΠΤΩΣΗ ΚΛΑΔΕΜΑΤΟΣ (ΔΕΞΙΑ) ΚΑΙ ΜΗ (ΑΡΙΣΤΕΡΑ) ***(ΤΙΜΗ ΒΑΘΟΥΣ: 4)***

- Λοιπές τιμές βάθους



ΕΙΚΟΝΑ 3.88: ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ ΤΑ ΟΠΟΙΑ ΔΕΝ ΕΧΟΥΝ ΥΠΟΣΤΕΙ ΔΙΑΔΙΚΑΣΙΑ ΚΛΑΔΕΜΑΤΟΣ (ΤΙΜΕΣ ΒΑΘΟΥΣ ΑΠΟ ΑΡΙΣΤΕΡΑ: 10, 25, 50)

Τα συμπεράσματα, τα οποία προκύπτουν βάσει των παραπάνω εικόνων είναι τα ακόλουθα:

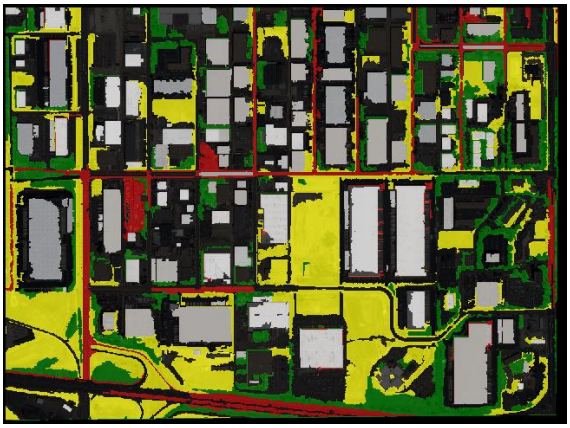
- Η ρύθμιση της παραμέτρου Κλάδεμα των δέντρων δεν επηρεάζει το αποτέλεσμα της ταξινόμησης
- Το δέντρο το οποίο κατασκευάζεται για τιμή βάθους έστω d χωρίς κλάδεμα είναι ίδιο με εκείνο για τιμή βάθους d+1 με κλάδεμα
- Τα δέντρα τα οποία κατασκευάζονται για τιμές βάθους μεγαλύτερες της τιμής 6 είναι πανομοιότυπα. Βάσει των παραπάνω εικόνων προκύπτει πως το μεγαλύτερο δέντρο έχει βάθος 7.

3.6.9 Τελική επιλογή των παραμέτρων των δέντρων απόφασης

Βάσει των παραπάνω επαναλήψεων προκύπτει πως οι τιμές εκείνες των παραμέτρων οι οποίες δίνουν τα υψηλότερα δυνατά ποσοστά ποιότητας είναι οι ακόλουθες:

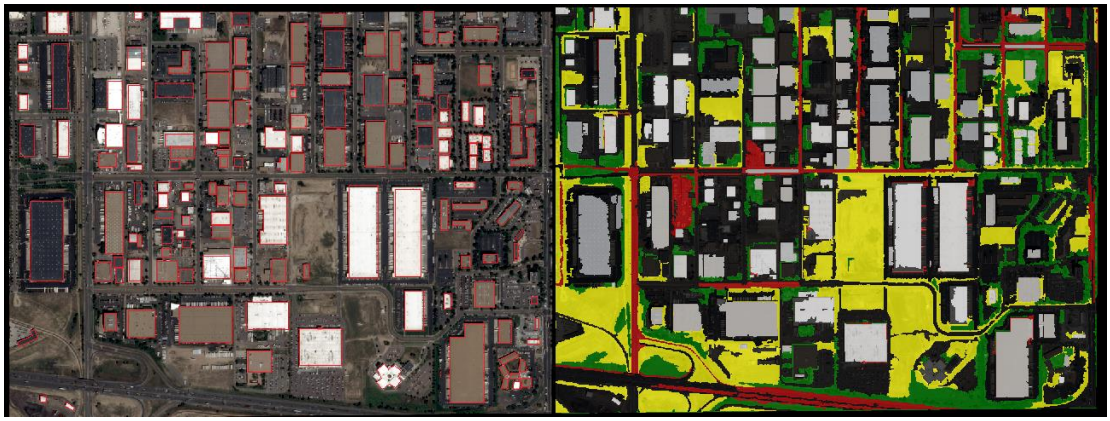
- Βάθος δέντρου (Depth): 5
- Ελάχιστος αριθμός δειγμάτων (Min sample count): 0
- Χρήση αντικαταστατών (Use surrogates): Όχι (No)
- Μέγιστος αριθμός κατηγοριών (Max categories): 16
- Cross Validation folds: 3
- Use 1 SE rule: Όχι (No)
- Αφαίρεση των κλαδεμένων κλαδιών (Truncate pruned trees): Ναι (Yes)

Στην Εικόνα 3.89 εμφανίζεται το αποτέλεσμα της ταξινόμησης μέσω του αλγορίθμου των δέντρων απόφασης βάσει των παραπάνω παραμέτρων.



ΕΙΚΟΝΑ 3.89: ΤΕΛΙΚΟ ΑΠΟΤΕΛΕΣΜΑ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΒΑΣΕΙ ΤΩΝ ΔΟΚΙΜΩΝ ΑΝΑΦΟΡΙΚΑ ΜΕ ΤΙΣ ΤΙΜΕΣ ΤΩΝ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ

Η αξιολόγηση του τελικού αποτελέσματος έγινε έπειτα από αξιολόγηση του αποτελέσματος της ταξινόμησης στο σύνολο της εικόνας. Για το σκοπό αυτό έγινε ψηφιοποίηση όλων των εμφανιζόμενων κτιρίων στο υπό μελέτη τμήμα της εικόνας (Εικόνα 3.90Εικόνα 3.194). Οι δείκτες ποιότητας όπως προέκυψαν έπειτα από φωτοερμηνεία εμφανίζονται στους Πίνακας 3.21 και Πίνακας 3.22.



ΕΙΚΟΝΑ 3.90: ΑΞΙΟΛΟΓΗΣΗ ΤΩΝ ΕΠΙΔΟΣΕΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΣΕ Ο,ΤΙ ΑΦΟΡΑ ΤΗΝ ΑΝΙΧΝΕΥΣΗ ΚΤΙΡΙΩΝ

ΠΙΝΑΚΑΣ 3.21*: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (ΤΕΛΙΚΗ ΕΠΙΛΟΓΗ ΤΩΝ ΤΙΜΩΝ ΤΩΝ ΠΑΡΑΜΕΤΡΩΝ)*

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Σύνολο εικόνας	94	22	63

ΠΙΝΑΚΑΣ 3.22: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (*ΤΕΛΙΚΗ ΕΠΙΛΟΓΗ ΤΩΝ ΤΙΜΩΝ ΤΩΝ ΠΑΡΑΜΕΤΡΩΝ*).

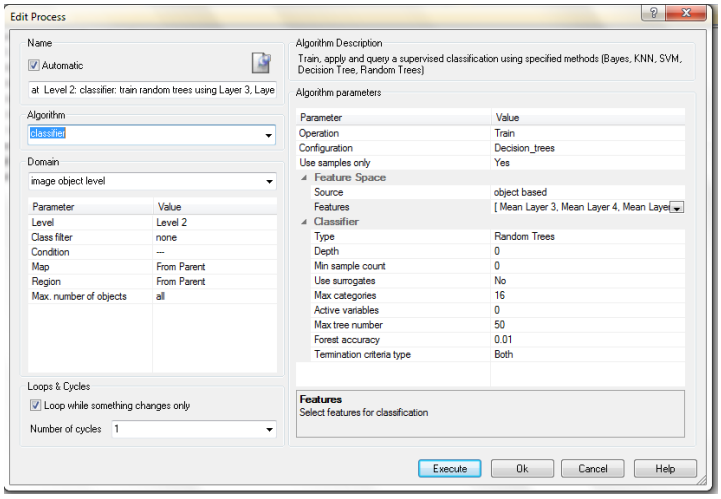
	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Σύνολο εικόνας	59,87%	81,03%	52,51%	40,13%	14,01%

3.7 Υλοποίηση του αλγορίθμου των τυχαίων δασών στο περιβάλλον του eCognition στο πρώτο τμήμα της εικόνας

3.7.1 Δοκιμή 1 (Προκαθορισμένες παράμετροι)

Στα πλαίσια της πρώτης δοκιμής έγινε χρήση των προκαθορισμένων τιμών των παραμέτρων. Αναλυτικά, ορίστηκαν οι ακόλουθες τιμές:

- Βάθος δέντρου (Depth): 0
- Ελάχιστος αριθμός δειγμάτων (Min sample count): 0
- Χρήση αντικαταστατών (Use surrogates): Όχι (No)
- Μέγιστος αριθμός κατηγοριών (Max categories): 16
- Ενεργές μεταβλητές: 0 (δηλαδή ουσιαστικά για $\sqrt{9} = 3$)
- Μέγιστος αριθμός δέντρων (Max tree number): 50
- Ακρίβεια δάσους (Forest accuracy): 0.01
- Τύπος κριτηρίου τερματισμού (termination criteria type): Και τα δύο (Both)(Εικόνα 3.91).



ΕΙΚΟΝΑ 3.91: *ΡΥΘΜΙΣΗ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΓΙΑ ΤΗΝ ΠΡΩΤΗ ΔΟΚΙΜΗ*

Σχολιασμός αποτελεσμάτων

Στην Εικόνα 3.92 εμφανίζεται το αποτέλεσμα της ταξινόμησης των δέντρων απόφασης στα πλαίσια της πρώτης δοκιμής.

ΕΙΚΟΝΑ 3.92: ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΓΙΑ ΤΗΝ ΠΡΩΤΗ ΔΟΚΙΜΗ

Κτίρια

Η θεματική κατηγορία κτίρια εμφανίζει σχετικά υψηλά ποσοστά ορθότητας. Πιο συγκεκριμένα, τα αντικείμενα που έχουν ταξινομηθεί στην κατηγορία αυτή ανήκουν στην πλειοψηφία τους πράγματι στη συγκεκριμένη κλάση.

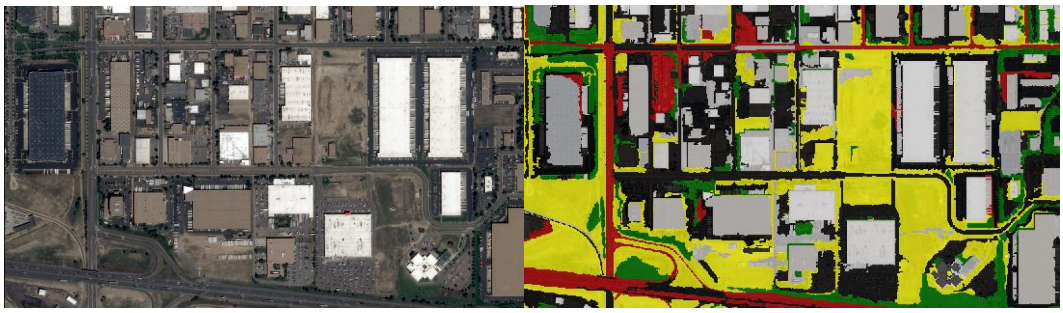
Υπάρχουν, ωστόσο, προβλήματα σε ό,τι αφορά την ικανοποίηση του κριτηρίου της πληρότητας καθώς πολλά αντικείμενα της κατηγορίας αυτής έχουν ταξινομηθεί σε εκείνη των χώρων στάθμευσης (Εικόνα 3.93).

ΕΙΚΟΝΑ 3.93: ΑΡΙΣΤΕΡΑ: 1^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ ΔΕΞΙΑ: 1^{Ο} ΑΠΟΣΠΑΣΜΑ ΤΑΞΙΝΟΜΗΜΕΝΗΣ ΕΙΚΟΝΑΣ ΓΙΑ ΤΗΝ ΠΡΩΤΗ ΔΟΚΙΜΗ (ΤΥΧΑΙΑ ΔΑΣΗ)

Δρόμοι

Η θεματική κατηγορία των δρόμων εμφανίζει απογοητευτικά αποτελέσματα σε ό,τι αφορά το κριτήριο της πληρότητας, καθώς, τα αντικείμενα εκείνα που έχουν ορθώς ταξινομηθεί από τον αλγόριθμο στους δρόμους είναι μόνο όσα δόθηκαν στη φάση της εκπαίδευσης του αλγορίθμου.

Τα αποτελέσματα σχετικά με το κριτήριο της ορθότητας είναι περισσότερο ενθαρρυντικά καθώς ο αριθμός των αντικειμένων που έχουν λανθασμένα ταξινομηθεί από το μοντέλο στη συγκεκριμένη θεματική κατηγορία είναι μικρός (Εικόνα 3.94).



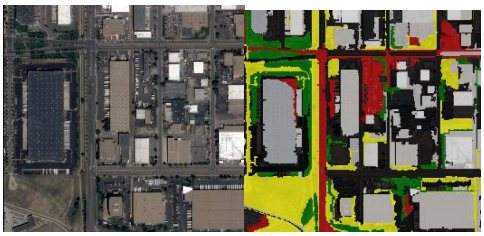
ΕΙΚΟΝΑ 3.94: ΑΡΙΣΤΕΡΑ:2^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ ΔΕΞΙΑ: 2^{Ο} ΑΠΟΣΠΑΣΜΑ ΤΑΞΙΝΟΜΗΜΕΝΗΣ ΕΙΚΟΝΑΣ ΓΙΑ ΤΗΝ ΠΡΩΤΗ ΔΟΚΙΜΗ

Χώροι Στάθμευσης

Σε ό,τι αφορά τη θεματική κατηγορία των δρόμων εμφανίζεται ικανοποιητικό ποσοστό πληρότητας, καθώς έχει ανιχνευτεί μεγάλο ποσοστό των εμφανιζόμενων χώρων στάθμευσης.

Πολλά αντικείμενα, ωστόσο, τα οποία έχουν καταχωρηθεί στη συγκεκριμένη θεματική κατηγορία ανήκουν στην πραγματικότητα σε εκείνη των κτιρίων και των δρόμων. Συνεπώς,

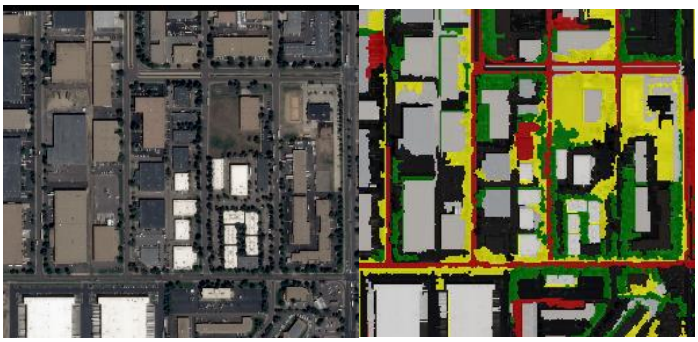
προκύπτει πως για την κλάση αυτή εμφανίζονται προβλήματα σχετικά με το κριτήριο της ορθότητας(Εικόνα 3.95).



ΕΙΚΟΝΑ 3.95: ΑΡΙΣΤΕΡΑ:3^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ ΔΕΞΙΑ: 3^{Ο} ΑΠΟΣΠΑΣΜΑ ΤΑΞΙΝΟΜΗΜΕΝΗΣ ΕΙΚΟΝΑΣ ΓΙΑ ΤΗΝ ΠΡΩΤΗ ΔΟΚΙΜΗ

Αστικό πράσινο

Στην Εικόνα 3.96 εμφανίζεται το αποτέλεσμα εφαρμογής του αλγορίθμου σε ένα αντιπροσωπευτικό σημείο αστικού πρασίνου. Είναι εμφανές πως το μοντέλο έχει εντοπίσει μεγάλο μέρος των εμφανιζόμενων αντικειμένων και επομένως εμφανίζεται ικανοποιητικό ποσοστό πληρότητας. Ενθαρρυντικό είναι επίσης το αποτέλεσμα σε ό,τι αφορά την ικανοποίηση του κριτηρίου της ορθότητας καθώς στην κλάση αυτή έχει καταχωρηθεί μικρός αριθμός αντικειμένων τα οποία στην πραγματικότητα ανήκουν σε εκείνη των χώρων στάθμευσης.

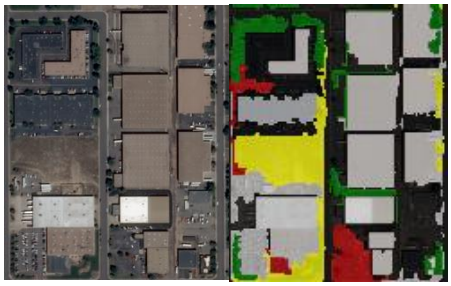


ΕΙΚΟΝΑ 3.96: ΑΡΙΣΤΕΡΑ: 4^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ ΔΕΞΙΑ: 4^{Ο} ΑΠΟΣΠΑΣΜΑ ΤΑΞΙΝΟΜΗΜΕΝΗΣ ΕΙΚΟΝΑΣ ΓΙΑ ΤΗΝ ΠΡΩΤΗ ΔΟΚΙΜΗ

Άγονο Έδαφος

Το αποτέλεσμα σχετικά με την κατηγορία του Άγονου Εδάφους είναι ικανοποιητικό τόσο σε ό,τι αφορά κριτήριο τόσο της πληρότητας όσο και της ορθότητας(Εικόνα 3.97).

ΕΙΚΟΝΑ 3.97: ΑΡΙΣΤΕΡΑ: 5^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ ΔΕΞΙΑ: 5^{Ο} ΑΠΟΣΠΑΣΜΑ ΤΑΞΙΝΟΜΗΜΕΝΗΣ ΕΙΚΟΝΑΣ ΓΙΑ ΤΗΝ ΠΡΩΤΗ ΔΟΚΙΜΗ



ΕΙΚΟΝΑ 3.98: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΗΝ ΠΡΩΤΗ ΔΟΚΙΜΗ

Η διαδικασία ποσοτικής αξιολόγησης του αλγορίθμου των τυχαίων δασών είναι ακριβώς η ίδια με εκείνη που εφαρμόστηκε στην περίπτωση των δέντρων απόφασης. Πιο συγκεκριμένα, επιλέχθηκε η ίδια αντιπροσωπευτική περιοχή της εικόνας- εισόδου (Εικόνα 3.98) και στη συνέχεια για την τελευταία εντοπίστηκε ο αριθμός των True Positives, False Positives και False Negatives δεδομένων. Βάσει των παραπάνω στοιχείων κατασκευάζονται οι ακόλουθοι Πίνακες (Πίνακας 3.23, Πίνακας 3.24):

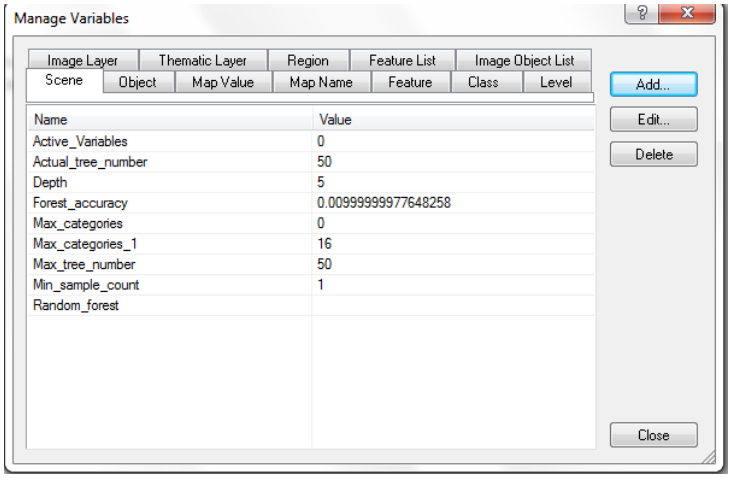
ΠΙΝΑΚΑΣ 3.23: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (1Η ΔΟΚΙΜΗ).

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	14	3	2

ΠΙΝΑΚΑΣ 3.24: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (1Η ΔΟΚΙΜΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	87,50%	82,35%	73,68%	12,50%	18,75%

Στην Εικόνα 3.99 εμφανίζονται οι πραγματικές τιμές των παραμέτρων έπειτα από τη δημιουργία του μοντέλου των τυχαίων δασών για τη δοκιμή 1. Η τιμή του βάθους όπως είναι αναμενόμενο δεν είναι μηδενική αλλά έχει πάρει βάθος ίσο με 5. Το ίδιο συμβαίνει και σε ό,τι αφορά τις παραμέτρους ελάχιστο πλήθος κόμβων και ενεργές μεταβλητές. Τέλος, αξίζει να σημειωθεί πως το πλήθος των δέντρων που εκπαιδεύτηκαν από τον αλγόριθμο είναι 46.



ΕΙΚΟΝΑ 3.99: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 1Η ΔΟΚΙΜΗ

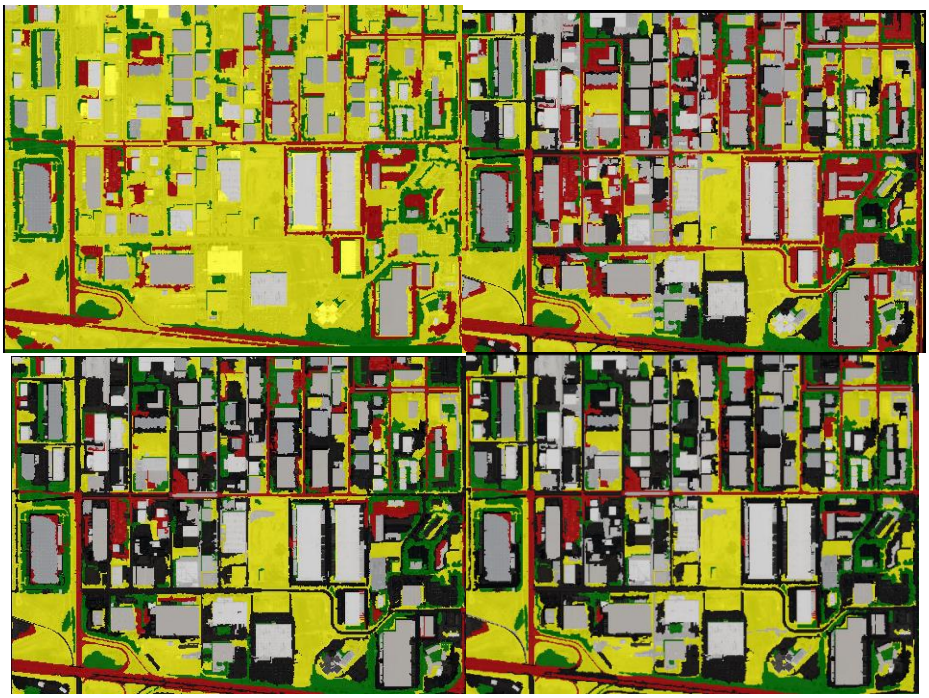
3.7.2 Δοκιμή 2 (Βάθος δέντρων)

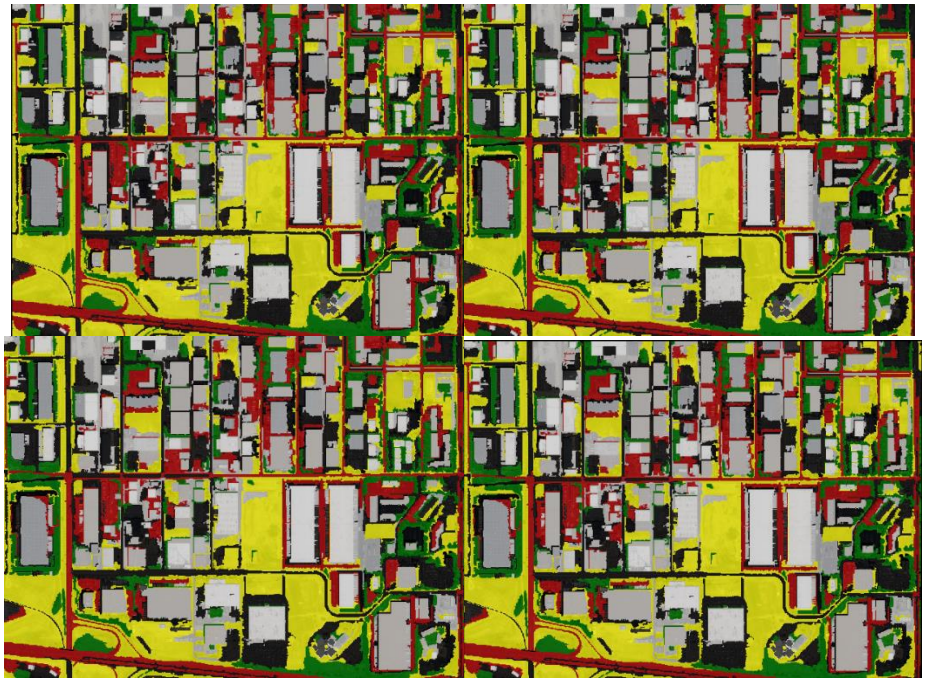
Στόχος της συγκεκριμένης δοκιμής ήταν η διερεύνηση της επιρροής της παραμέτρου Βάθος Δέντρου στην αποτελεσματικότητα του αλγορίθμου των τυχαίων δασών. Αναλυτικά, δημιουργήθηκε ένα σύνολο 8 διαφορετικών μοντέλων βάσει των ακόλουθων τιμών:

- **Βάθος δέντρου (Depth): 2, 3, 4, 5, 10, 25, 50, 100**
- Ελάχιστος αριθμός δείγμα, των (Min sample count): 0
- Χρήση αντικαταστατών (Use surrogates): Όχι (No)
- Μέγιστος αριθμός κατηγοριών (Max categories): 16
- Ενεργές μεταβλητές: 0 (δηλαδή ουσιαστικά για $\sqrt{9} = 3$)
- Μέγιστος αριθμός δέντρων (Max tree number): 50
- Ακρίβεια δάσους (Forest accuracy): 0.01
- Τύπος κριτηρίου τερματισμού (termination criteria type): Both

Σχολιασμός αποτελεσμάτων

Στην Εικόνα 3.100 εμφανίζεται το αποτέλεσμα εφαρμογής των τυχαίων δασών για τις 8 διαφορετικές τιμές της παραμέτρου βάθος δέντρων.





ΕΙΚΟΝΑ 3.100: ΑΠΟ ΤΗΝ ΑΡΧΗ: ΑΠΟΤΕΛΕΣΜΑ ΕΦΑΡΜΟΓΗΣ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΓΙΑ ΤΙΜΕΣ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΒΑΘΟΣ ΔΕΝΤΡΩΝ 2, 3, 4, 5, 10, 25, 50, 100

Κτίρια

Η αύξηση της παραμέτρου του βάθους σε 2 οδήγησε σε μείωση του πλήθους των εμφανιζόμενων κτιρίων συγκριτικά με εκείνο των προκαθορισμένων παραμέτρων. Η πληρότητα της συγκεκριμένης θεματικής κατηγορίας εμφανίζει πολύ χαμηλά ποσοστά, καθώς πολλά από τα κτίρια δεν εμφανίζονται στον παραγόμενο θεματικό χάρτη.

Η μείωση της παραμέτρου του βάθους σε 3 αύξησε τον αριθμό των αντικειμένων της συγκεκριμένης κλάσης. Το παραπάνω οδήγησε σε αύξηση του κριτηρίου της πληρότητας, αλλά και σε ελάττωση της ορθότητας καθώς στην κλάση αυτή καταχωρήθηκαν μεταξύ άλλων πολλοί από τους εμφανιζόμενους χώρους στάθμευσης.

Τα αποτελέσματα σε ό,τι αφορά τη θεματική κατηγορία των κτιρίων είναι παρόμοια για τις τιμές 4 και 5. Μάλιστα, ο θεματικός χάρτης ο οποίος προέκυψε για τιμή βάθους 5 είναι ακριβώς ο ίδιος με εκείνον των προκαθορισμένων τιμών.

Στη συνέχεια, η ρύθμιση της εν λόγω παραμέτρου σε 10 αύξησε ξανά τον αριθμό των αντικειμένων της κατηγορίας αυτής. Ορισμένα, από τα τελευταία ήταν όντως κτίρια και κάποια ανήκαν στην κλάση των χώρων στάθμευσης. Βάσει αυτού προκύπτει πως η αύξηση της τιμής σε 10 είχε θετικά αποτελέσματα ως προς το κριτήριο της πληρότητας και αρνητικά ως προς εκείνο της ορθότητας.

Η αύξηση του βάθους των δέντρων σε 25, 50 και 100 έδωσε αποτελέσματα πανομοιότυπα με εκείνα της τιμής 10. Συνεπώς, η ρύθμιση του βάθους σε τιμή μεγαλύτερη των 10 δε μετέβαλλε το αποτέλεσμα της εν λόγω ταξινόμησης (Εικόνα 3.101).

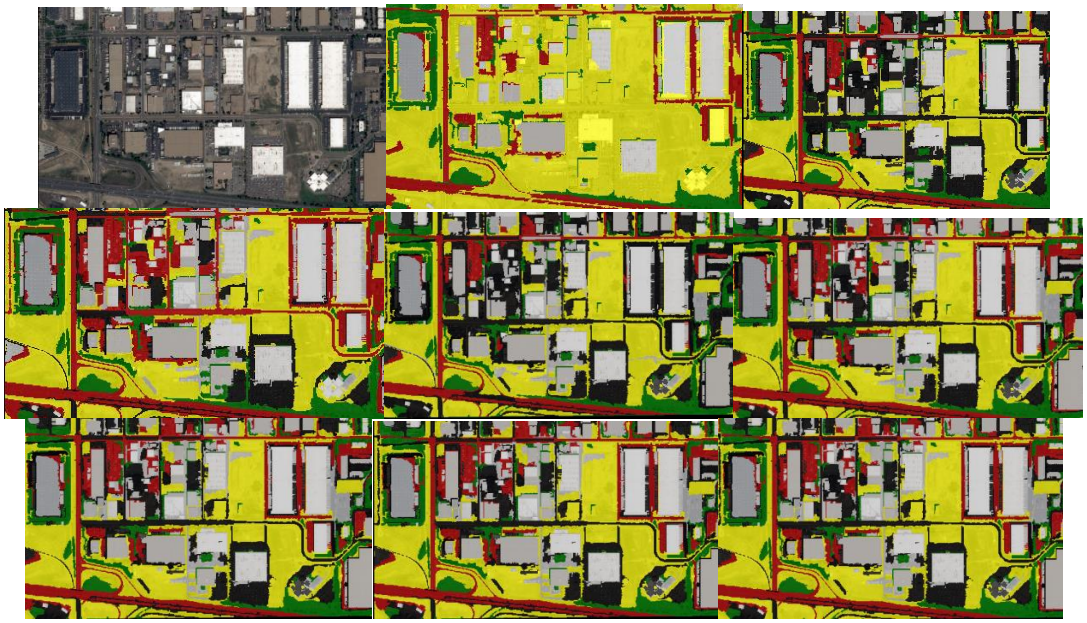
ΕΙΚΟΝΑ 3.101: : ΑΠΟ ΤΗΝ ΑΡΧΗ: 1^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ, ΓΙΑ ΤΗΝ ΤΙΜΗ 2 ΤΟΥ ΒΑΘΟΥΣ ΔΕΝΤΡΩΝ, ΓΙΑ ΤΗΝ ΤΙΜΗ 3, ΓΙΑ ΤΗΝ ΤΙΜΗ 4, ΓΙΑ ΤΗΝ ΤΙΜΗ 5, ΓΙΑ ΤΗΝ ΤΙΜΗ 25, ΓΙΑ ΤΗΝ ΤΙΜΗ 50, ΓΙΑ ΤΗΝ ΤΙΜΗ 100

Δρόμοι

Η ρύθμιση του βάθους των δέντρων σε 2, 4 και 5 εμφάνισε πανομοιότυπα με εκείνα των προκαθορισμένων παραμέτρων σε ό,τι αφορά την κλάση των δρόμων οδήγησε σε μείωση του πλήθους των εμφανιζόμενων αξόνων του οδικού δικτύου.

Ο θεματικός χάρτης ο οποίος προέκυψε για τιμή βάθους 3 εμφανίζει περισσότερα αντικείμενα της κλάσης των δρόμων συγκριτικά με εκείνον για τις τιμές 2, 4 και 5. Το κριτήριο της ορθότητας εμφανίζει χαμηλά ποσοστά στην παρούσα περίπτωση καθώς στη συγκεκριμένη θεματική κατηγορία έχουν καταχωρηθεί πολλά αντικείμενα τα οποία στην πραγματικότητα ανήκουν σε εκείνη των χώρων στάθμευσης.

Η μεταβολή της τιμής της συγκεκριμένης παραμέτρου σε 10 αύξησε τον αριθμό των αντικειμένων της κλάσης αυτής. Το παραπάνω μείωσε το ποσοστό της ορθότητας της ταξινόμησης καθώς στην κατηγορία αυτή προστέθηκαν αντικείμενα τα οποία ανήκουν σε εκείνη των χώρων στάθμευσης. Τέλος, σημειώνεται πως η ταξινόμηση έπειτα από την αύξηση της τιμής του βάθους σε 25, 50 και 100 δε εμφανίζει καμία απολύτως μεταβολή σε σχέση με την προαναφερθείσα (Εικόνα 3.102).



ΕΙΚΟΝΑ 3.102: ΑΠΟ ΤΗΝ ΑΡΧΗ: 2^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ, ΓΙΑ ΤΗΝ ΤΙΜΗ 2 ΤΟΥ ΒΑΘΟΥΣ ΔΕΝΤΡΩΝ, ΓΙΑ ΤΗΝ ΤΙΜΗ 3, ΓΙΑ ΤΗΝ ΤΙΜΗ 4, ΓΙΑ ΤΗΝ ΤΙΜΗ 5, ΓΙΑ ΗΝ ΤΙΜΗ 25, ΓΙΑ ΤΗΝ ΤΙΜΗ 50, ΓΙΑ ΤΗΝ ΤΙΜΗ 100

Χώροι στάθμευσης

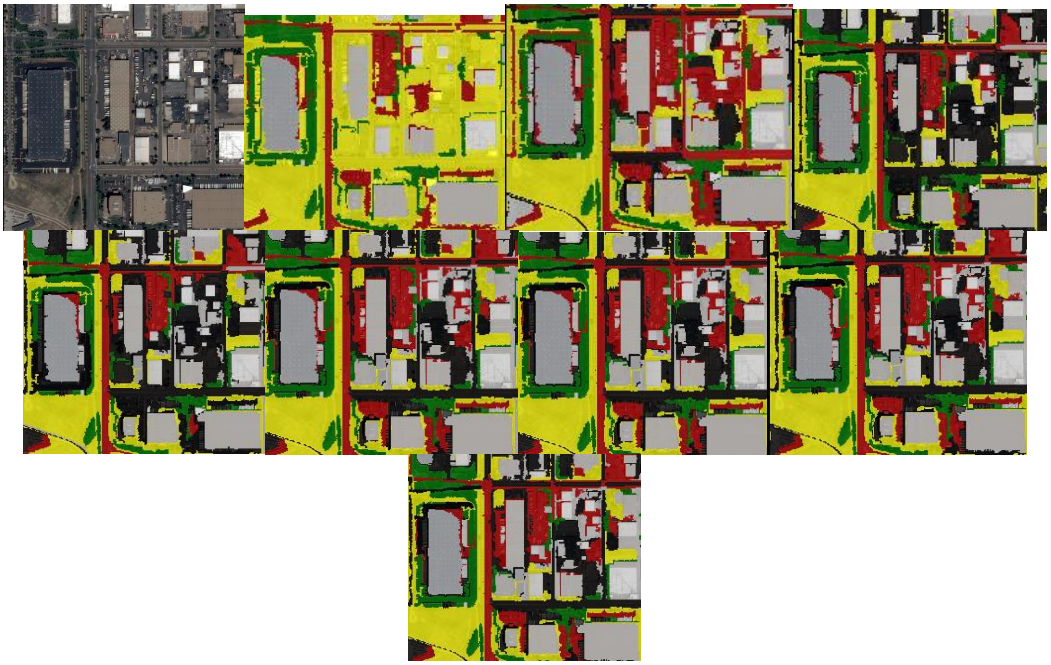
Η θεματική κατηγορία των χώρων στάθμευσης δεν εμφανίζεται στο θεματικό χάρτη ο οποίος προέκυψε για τιμή βάθους 2. Συνεπώς, τα κριτήρια ορθότητας καθώς και πληρότητας είναι μηδενικά.

Η ρύθμιση της τιμής του βάθους σε 3 αύξησε σε πολύ μικρό βαθμό το πλήθος των εμφανιζόμενων χώρων στάθμευσης. Το ποσοστό πληρότητας εξακολουθεί, ωστόσο να είναι εξαιρετικά χαμηλό καθώς το μεγαλύτερο μέρος των τελευταίων έχουν ταξινομηθεί από το μοντέλο στο άγονο έδαφος.

Η ρύθμιση της τιμής της συγκεκριμένης παραμέτρου σε 4 είχε θετικά αποτελέσματα σε ό,τι αφορά την ικανοποίηση του κριτηρίου της ορθότητας καθώς και της πληρότητας καθώς ο αριθμός των ορθώς καταχωρημένων στην κλάση αυτή αντικειμένων αυξήθηκε κατακόρυφα.

Τα αποτελέσματα δεν ήταν ίδια όταν η τιμή του βάθους αυξήθηκε σε 10. Στην περίπτωση αυτή ωστόσο υπήρξε μείωση του ποσοστού της πληρότητας καθώς από την κατηγορία αυτή εξαιρέθηκαν πολλά αντικείμενα. Τα τελευταία καταχωρήθηκαν εσφαλμένα από το μοντέλο στην κλάση των δρόμων.

Η ρύθμιση του βάθους σε 25, 50 και 100 εμφάνισε πανομοιότυπα αποτελέσματα με εκείνη της τιμής 10 (Εικόνα 3.103).



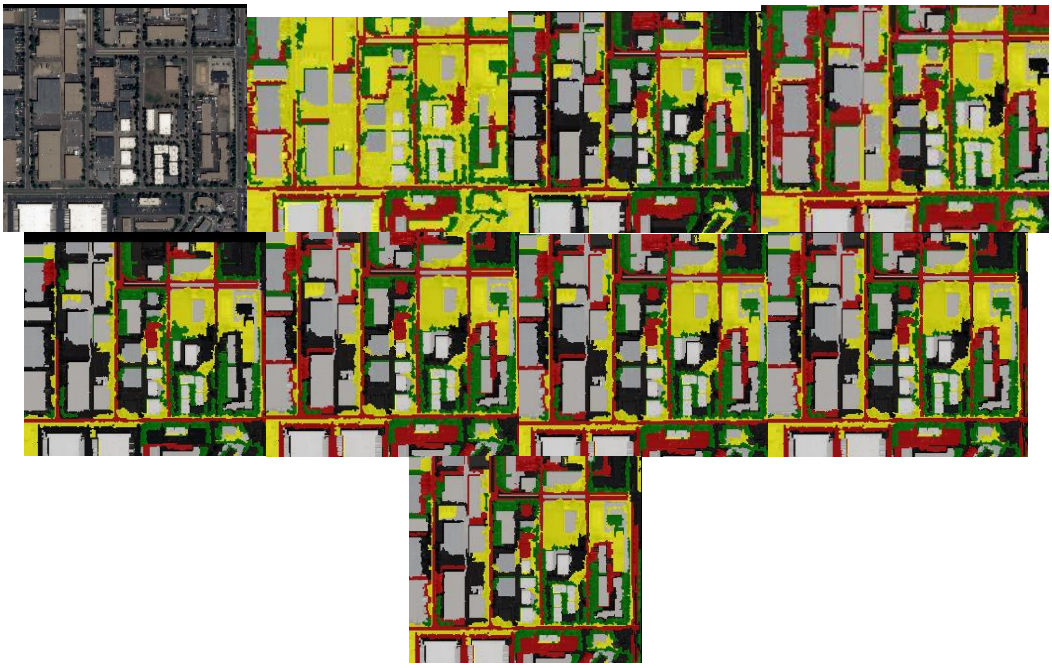
ΕΙΚΟΝΑ 3.103: ΑΠΟ ΤΗΝ ΑΡΧΗ: 3^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ, ΓΙΑ ΤΗΝ ΤΙΜΗ 2 ΤΟΥ ΒΑΘΟΥΣ ΔΕΝΤΡΩΝ, ΓΙΑ ΤΗΝ ΤΙΜΗ 3, ΓΙΑ ΤΗΝ ΤΙΜΗ 4, ΓΙΑ ΤΗΝ ΤΙΜΗ 5, ΓΙΑ ΗΝ ΤΙΜΗ 25, ΓΙΑ ΤΗΝ ΤΙΜΗ 50, ΓΙΑ ΤΗΝ ΤΙΜΗ 100

Αστικό πράσινο

Η θεματική κατηγορία του αστικού πρασίνου δεν εμφανίζει διαφορές για τις τιμές 2, 3, 4 και 5 του βάθους των δέντρων.

Αντιθέτως, η ρύθμιση της τιμής της παραμέτρου σε 10 μείωσε τον αριθμό των αντικειμένων και αυτό οδήγησε σε αύξηση της τιμής της ορθότητας. Αυτό συνέβη καθώς τα αντικείμενα τα οποία αφαιρέθηκαν από το νέο θεματικό χάρτη ανήκουν στην πραγματικότητα σε εκείνη των κτιρίων.

Τέλος, σημειώνεται πως η αύξηση του βάθους σε 25, 50 και 100 δε μετέβαλε το αποτέλεσμα της ταξινόμησης (Εικόνα 3.104).

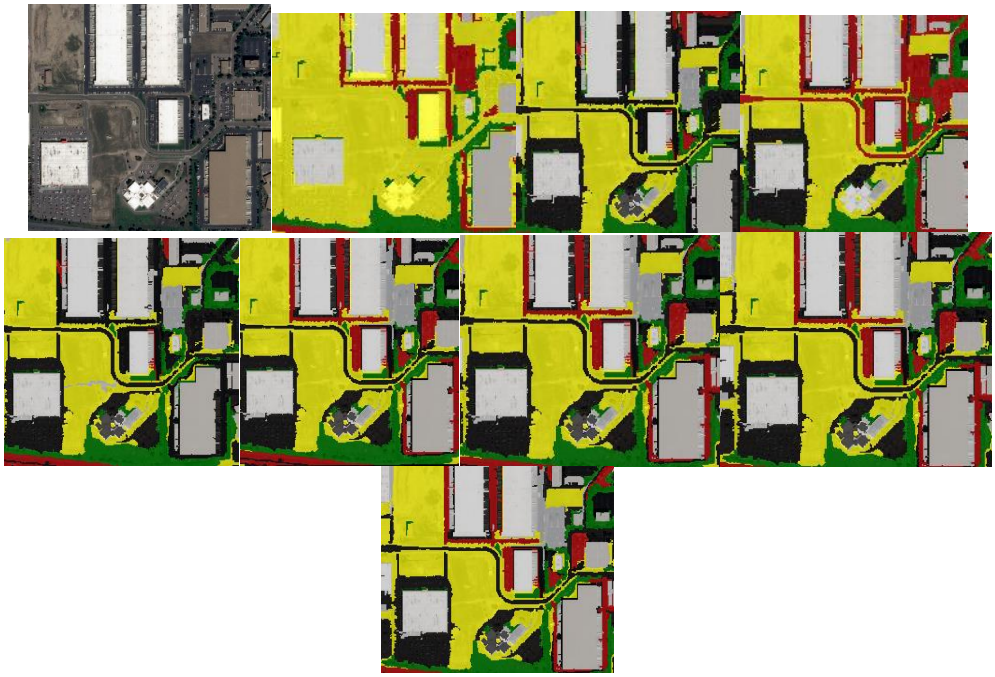


ΕΙΚΟΝΑ 3.104: ΑΠΟ ΤΗΝ ΑΡΧΗ: 4^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ, ΓΙΑ ΤΗΝ ΤΙΜΗ 2 ΤΟΥ ΒΑΘΟΥΣ ΔΕΝΤΡΩΝ, ΓΙΑ ΤΗΝ ΤΙΜΗ 3, ΓΙΑ ΤΗΝ ΤΙΜΗ 4, ΓΙΑ ΤΗΝ ΤΙΜΗ 5, ΓΙΑ ΗΝ ΤΙΜΗ 25, ΓΙΑ ΤΗΝ ΤΙΜΗ 50, ΓΙΑ ΤΗΝ ΤΙΜΗ 100

Άγονο Έδαφος

Η ρύθμιση της παραμέτρου του βάθους σε 2 οδήγησε σε αύξηση του αριθμού των αντικειμένων της θεματικής κατηγορίας του άγονου εδάφους. Το παραπάνω έγινε εις βάρος του κριτηρίου της ορθότητας καθώς πολλά από τα αντικείμενα τα οποία προστέθηκαν στη συγκεκριμένη κλάση ανήκουν στην πραγματικότητα σε εκείνη των κτιρίων, των χώρων στάθμευσης και των δρόμων.

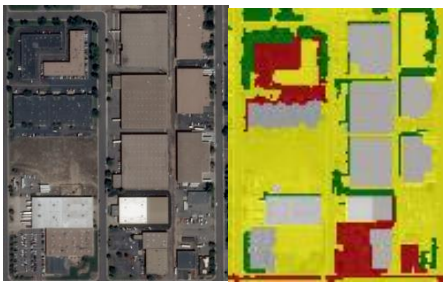
Στη συνέχεια, οι αλλαγές στην τιμή της παραμέτρου του βάθους σε 3, 4, 5, 25, 50 και 100 δεν επέφερε καμία απολύτως αλλαγή σε ό,τι αφορά την ταξινόμηση των αντικειμένων του άγονου εδάφους συγκριτικά με εκείνη των προκαθορισμένων παραμέτρων (Εικόνα 3.105).



ΕΙΚΟΝΑ 3.105: ΑΠΟ ΤΗΝ ΑΡΧΗ: 5^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ ΓΙΑ ΤΗΝ ΤΙΜΗ 2 ΤΟΥ ΒΑΘΟΥΣ ΔΕΝΤΡΩΝ, ΓΙΑ ΤΗΝ ΤΙΜΗ 3, ΓΙΑ ΤΗΝ ΤΙΜΗ 4, ΓΙΑ ΤΗΝ ΤΙΜΗ 5, ΓΙΑ ΗΝ ΤΙΜΗ 25, ΓΙΑ ΤΗΝ ΤΙΜΗ 50, ΓΙΑ ΤΗΝ ΤΙΜΗ 100

ΠΟΣΟΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

ΒΑΘΟΣ ΔΕΝΤΡΩΝ: 2



ΕΙΚΟΝΑ 3.106: *ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΗ ΔΕΥΤΕΡΗ ΔΟΚΙΜΗ*

Βάσει της Εικόνα 3.106 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.25, Πίνακας 3.26)

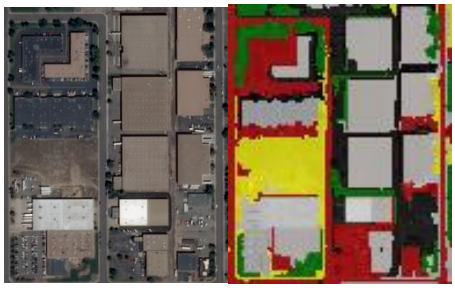
ΠΙΝΑΚΑΣ 3.25: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (2^{Η} ΔΟΚΙΜΗ)

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	14	1	2

ΠΙΝΑΚΑΣ 3.26: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (2^{Η} ΔΟΚΙΜΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	87,50%	93,33%	82,35%	12,50%	6,25%

ΒΑΘΟΣ ΔΕΝΤΡΩΝ: 3



ΕΙΚΟΝΑ 3.107: ***ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΙΜΗ ΒΑΘΟΥΣ ΙΣΗ ΜΕ 3***

Βάσει της υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.27, Πίνακας 3.28)

ΠΙΝΑΚΑΣ 3.27:*ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (2Η ΔΟΚΙΜΗ)*

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	14	3	2

ΠΙΝΑΚΑΣ 3.28: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (2Η ΔΟΚΙΜΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	87,50%	82,35%	73,68%	12,50%	18,75%

ΒΑΘΟΣ ΔΕΝΤΡΩΝ: 4, 5

ΕΙΚΟΝΑ 3.108: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΙΜΗ ΒΑΘΟΥΣ ΙΣΗ ΜΕ 4

Βάσει της Εικόνα 3.108 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.29, Πίνακας 3.30)

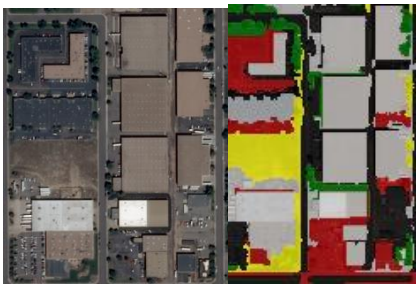
ΠΙΝΑΚΑΣ 3.29*: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (2Η ΔΟΚΙΜΗ)*

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	14	3	2

ΠΙΝΑΚΑΣ 3.30: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (2Η ΔΟΚΙΜΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	87,50%	82,35%	73,68%	12,50%	18,75%

ΒΑΘΟΣ ΔΕΝΤΡΩΝ: 10, 25, 50, 100



ΕΙΚΟΝΑ 3.109: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΙΜΗ ΒΑΘΟΥΣ ΙΣΗ ΜΕ 10, 25, 50, 100

Βάσει της Εικόνα 3.109 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.31, Πίνακας 3.32)

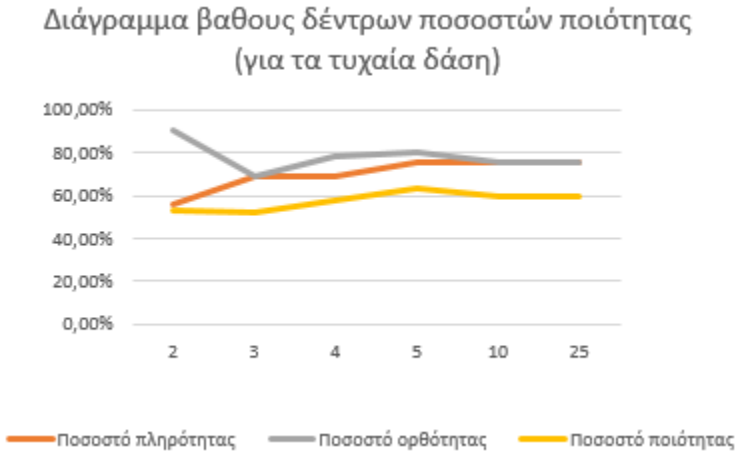
ΠΙΝΑΚΑΣ 3.31: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (2Η ΔΟΚΙΜΗ)

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	14	4	2

ΠΙΝΑΚΑΣ 3.32: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (2Η ΔΟΚΙΜΗ)

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	87,50%	77,78%	70,00%	12,50%	25,00%

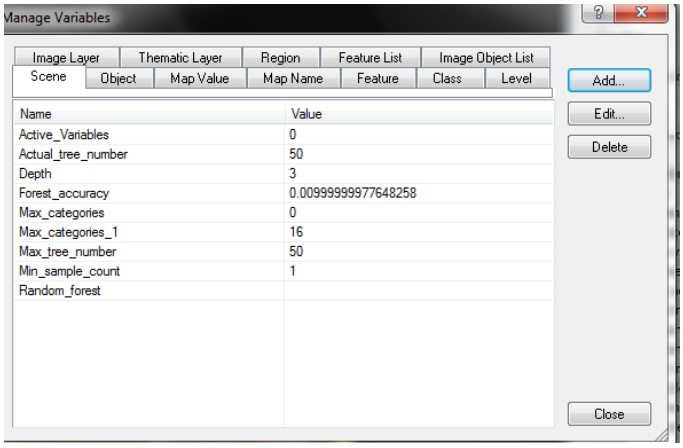
Στο ακόλουθο διάγραμμα (Εικόνα 3.110) εμφανίζονται τα ποσοστά πληρότητας συναρτήσει του βάθους δέντρων



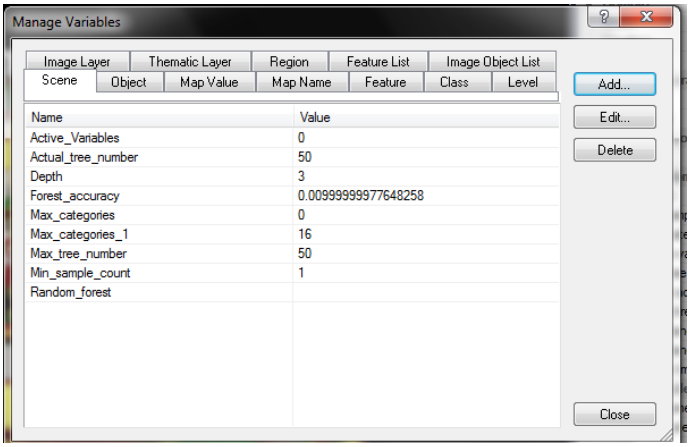
ΕΙΚΟΝΑ 3.110: ΔΙΑΓΡΑΜΜΑ ΒΑΘΟΥΣ ΔΕΝΤΡΩΝ ΠΟΣΟΣΤΩΝ ΠΟΙΟΤΗΤΑΣ

Στις ακόλουθες Εικόνες (Εικόνα 3.111- Εικόνα 3.118) εμφανίζονται οι πραγματικές τιμές των παραμέτρων έπειτα από τη δημιουργία των μοντέλων των τυχαίων δασών για τη δοκιμή 2. Παρατηρείται πως η μεταβολή στο βάθος των δέντρων επιφέρει αλλαγές στο πλήθος των δέντρων. Το παραπάνω οφείλεται στο γεγονός πως στην παράμετρο αναφορικά με τον τερματισμό της κατασκευής του τυχαίου δάσους έχει οριστεί η τιμή Και τα δύο (Both). Συνεπώς, η δημιουργία του δάσους τερματίζεται εφόσον ικανοποιηθεί το κριτήριο σχετικά με το σφάλμα του μοντέλου και ο αλγόριθμος κατασκευάζει τόσα δέντρα απόφασης όσα

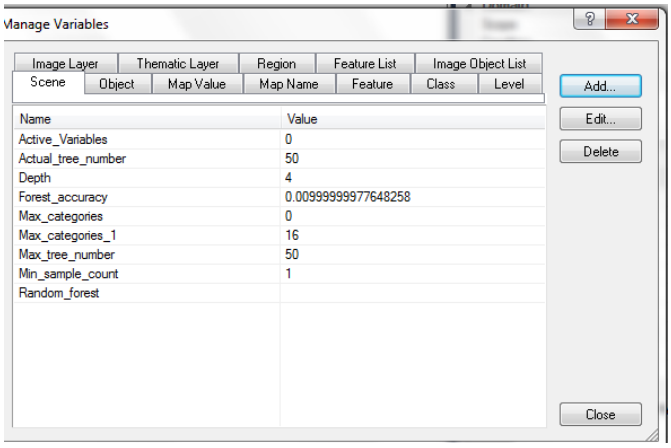
είναι απαραίτητα προκειμένου να ικανοποιηθεί ο παραπάνω περιορισμός. Σημειώνεται, ωστόσο, πως το πλήθος των τελευταίων δεν πρέπει να υπερβαίνουν τα 50 (max tree number).



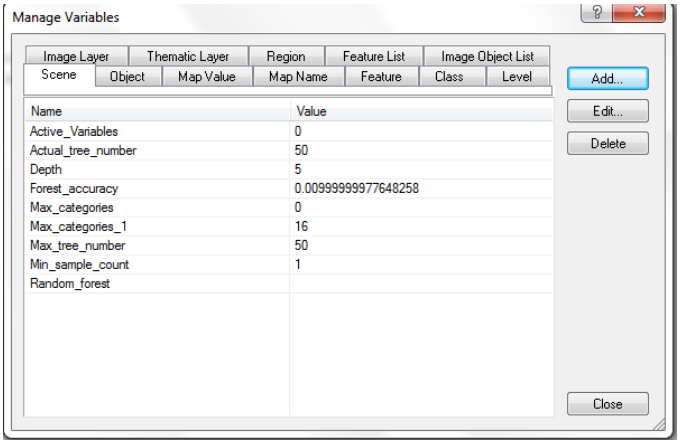
ΕΙΚΟΝΑ 3.111: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 2Η ΔΟΚΙΜΗ (ΒΑΘΟΣ: 2)



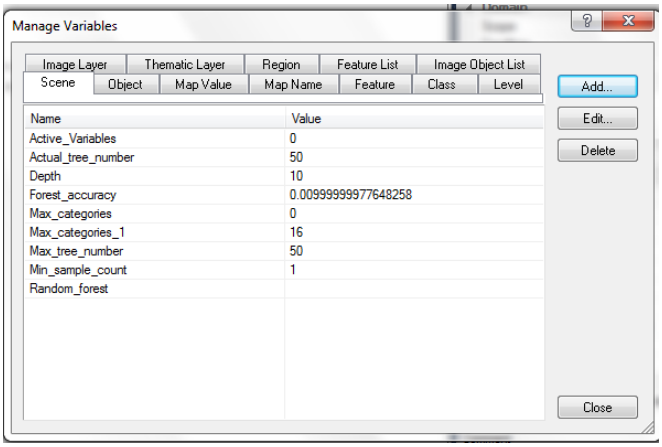
ΕΙΚΟΝΑ 3.112: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 2Η ΔΟΚΙΜΗ (ΒΑΘΟΣ: 3)



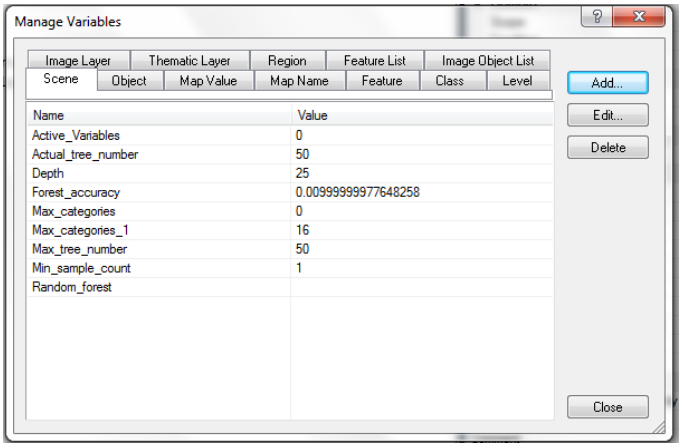
ΕΙΚΟΝΑ 3.113: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 2Η ΔΟΚΙΜΗ (ΒΑΘΟΣ: 4)



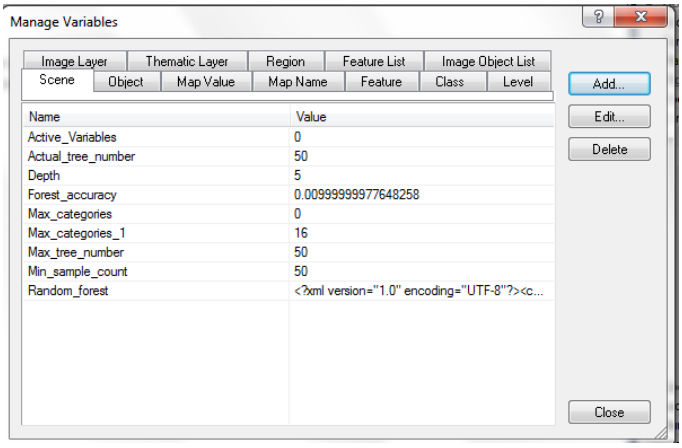
ΕΙΚΟΝΑ 3.114: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 2Η ΔΟΚΙΜΗ (ΒΑΘΟΣ: 5)



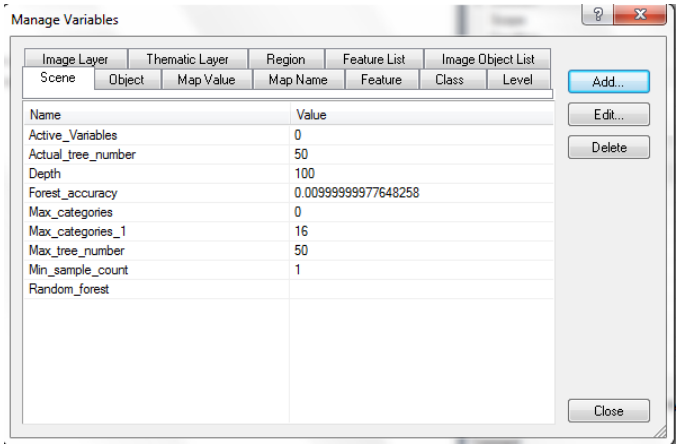
ΕΙΚΟΝΑ 3.115: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 2Η ΔΟΚΙΜΗ (ΒΑΘΟΣ: 10)



ΕΙΚΟΝΑ 3.116: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 2Η ΔΟΚΙΜΗ (ΒΑΘΟΣ: 25)



ΕΙΚΟΝΑ 3.117: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 2Η ΔΟΚΙΜΗ (ΒΑΘΟΣ: 50)



ΕΙΚΟΝΑ 3.118: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 2Η ΔΟΚΙΜΗ (ΒΑΘΟΣ: 100)

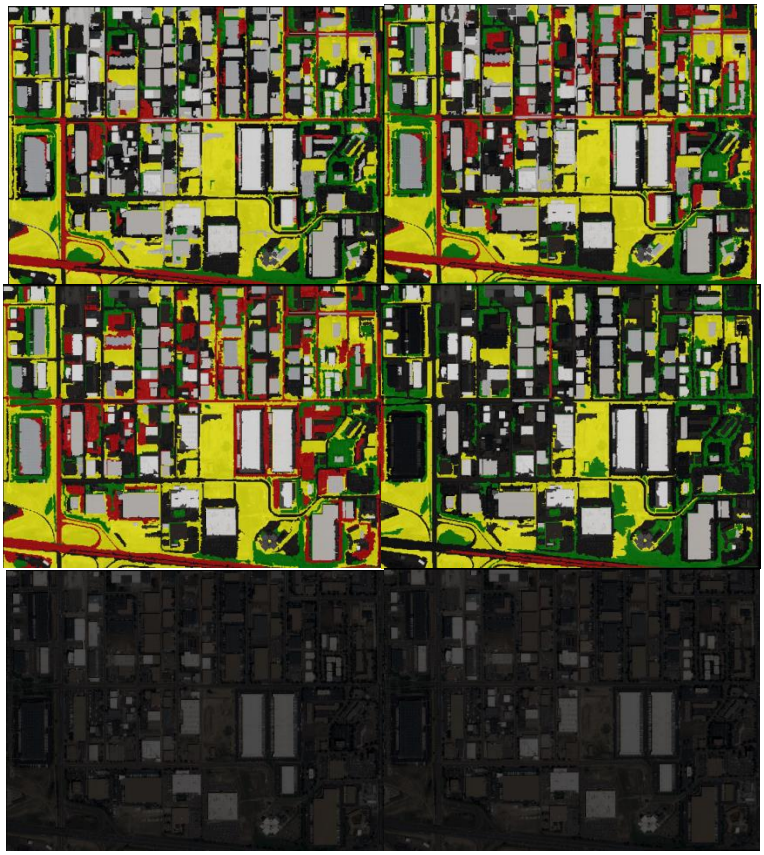
3.7.3 Δοκιμή 3 (Ελάχιστος αριθμός δειγμάτων)

Στα πλαίσια της δοκιμής αυτής έγινε πειραματισμός με τις τιμές της παραμέτρου του ελάχιστου αριθμού δειγμάτων. Πιο συγκεκριμένα ορίστηκαν οι: 5, 10, 25, 50 100. Αναλυτικά, οι τιμές των παραμέτρων ήταν οι ακόλουθες:

- Βάθος δέντρου (Depth): 0
- **Ελάχιστος αριθμός δειγμάτων (Min sample count): 5, 10, 25, 50, 100**
- Χρήση αντικαταστατών (Use surrogates): Όχι (No)
- Μέγιστος αριθμός κατηγοριών (Max categories): 16
- Ενεργές μεταβλητές (Active Variables): 0 (δηλαδή ουσιαστικά για $\sqrt{9} = 3$)
- Μέγιστος αριθμός δέντρων (Max tree number): 50
- Ακρίβεια δάσους (Forest accuracy): 0.01
- Τύπος κριτηρίου τερματισμού (termination criteria type): Both

Σχολιασμός αποτελεσμάτων

Στην ακόλουθη Εικόνα (Εικόνα 3.119) εμφανίζεται το αποτέλεσμα εφαρμογής των συγκεκριμένων μοντέλων στην εικόνα εισόδου. Παρατηρείται πως η αύξηση στην τιμή της συγκεκριμένης παραμέτρου οδηγεί σε μείωση της ποιότητας του τελικού αποτελέσματος, καθώς είναι εμφανές στους θεματικούς χάρτες των τιμών 50 και 100 εμφανίζεται αποκλειστικά η κλάση των χώρων στάθμευσης.

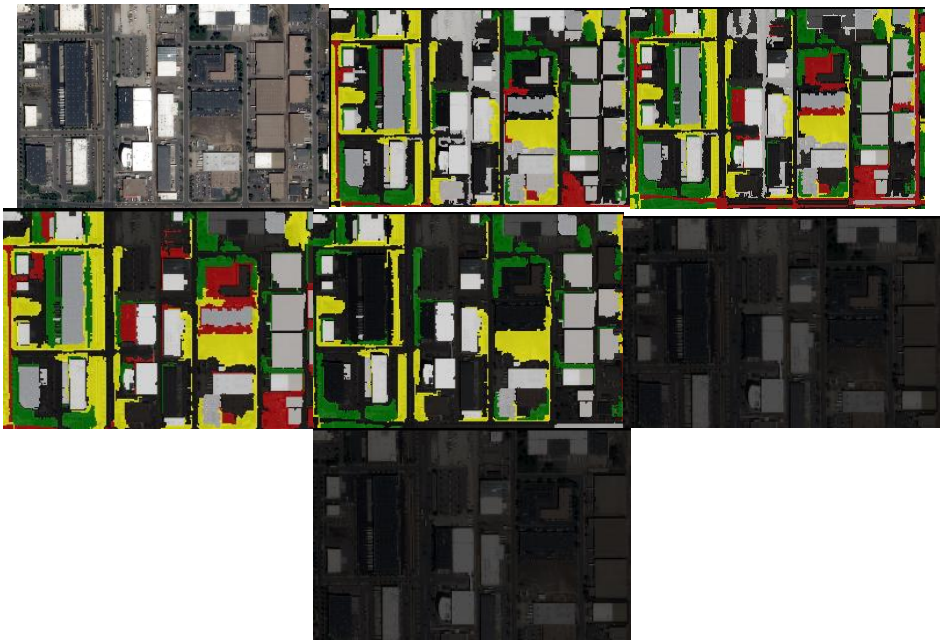


ΕΙΚΟΝΑ 3.119: ΑΠΟΤΕΛΕΣΜΑ ΕΦΑΡΜΟΓΗΣ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΓΙΑ ΤΙΜΕΣ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΤΩΝ ΕΛΑΧΙΣΤΩΝ ΔΕΙΓΜΑΤΩΝ 0 (ΠΑΝΩ ΑΡΙΣΤΕΡΑ), 5 (ΠΑΝΩ ΔΕΞΙΑ), 10 (ΑΡΙΣΤΕΡΑ), 25 (ΔΕΞΙΑ), 50 (ΚΑΤΩ ΑΡΙΣΤΕΡΑ), 100 (ΚΑΤΩ ΔΕΞΙΑ)

Κτίρια

Η αύξηση της τιμής του ελάχιστου αριθμού των δειγμάτων από 0 σε 100 οδήγησε σε σταδιακή μείωση των εμφανιζόμενων κτιρίων. Πιο συγκεκριμένα, η ρύθμιση της τιμής από 0 σε 5, 10 και 25 είχε σαν αποτέλεσμα την ελάττωση του αριθμού των αντικειμένων της κλάσης αυτής στους παραγόμενους θεματικούς χάρτες. Το παραπάνω είχε αρνητικές συνέπειες σε ό,τι αφορά την ικανοποίηση του κριτηρίου της πληρότητας καθώς πολλά από τα αντικείμενα που απομακρύνθηκαν ανήκαν πράγματι στην κλάση των κτιρίων. Παράλληλα, το παραπάνω είχε θετικά αποτελέσματα για την ορθότητα της συγκεκριμένης θεματικής κατηγορίας καθώς πολλά αντικείμενα τα οποία ανήκαν στους χώρους στάθμευσης και είχαν αρχικά ταξινομηθεί στην κατηγορία των κτιρίων κατατάχθηκαν έπειτα από την αύξηση της τιμής της συγκεκριμένης παραμέτρου στη σωστή κλάση

Το αποτέλεσμα είναι απογοητευτικό για τις τιμές 50 και 100 καθώς η κλάση των κτιρίων έχει εξαφανιστεί στους παραγόμενους χάρτες. Βάσει αυτού προκύπτει πως οι παράμετροι της πληρότητας, της ορθότητας και συνεπώς της ποιότητας συγκεντρώνουν ποσοστό 0% (Εικόνα 3.120).

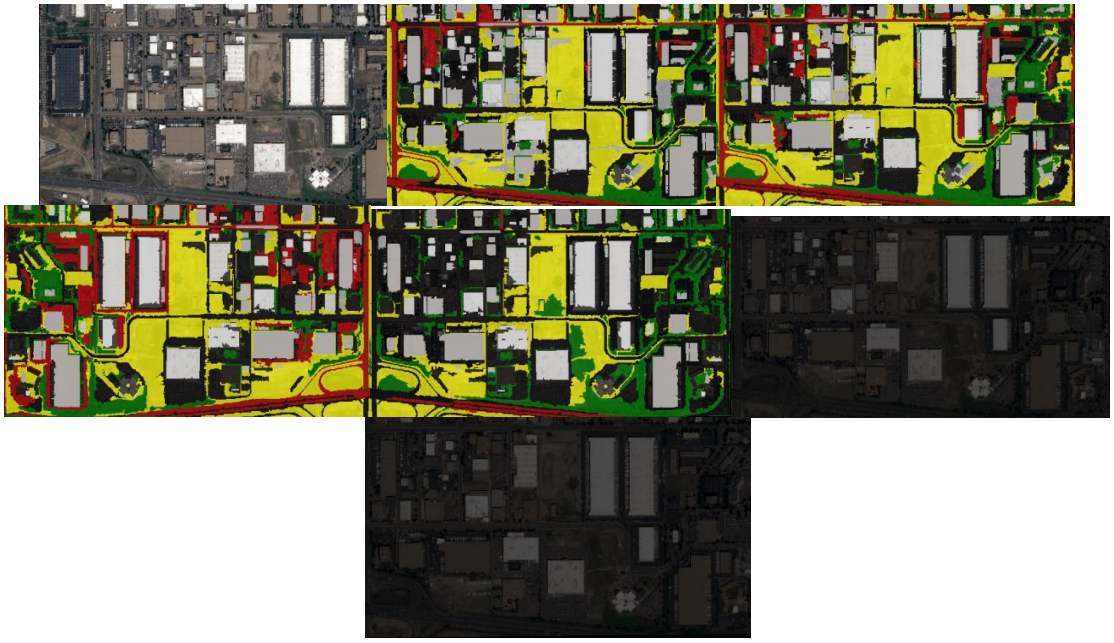


ΕΙΚΟΝΑ 3.120: ΑΠΟ ΤΗΝ ΑΡΧΗ: 1^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ, ΓΙΑ ΤΗΝ ΤΙΜΗ 0 ΤΩΝ ΕΛΑΧΙΣΤΩΝ ΔΕΙΓΜΑΤΩΝ, ΓΙΑ ΤΗΝ ΤΙΜΗ 5, ΓΙΑ ΤΗΝ ΤΙΜΗ 10, ΓΙΑ ΤΗΝ ΤΙΜΗ 25, ΓΙΑ ΤΗΝ ΤΙΜΗ 50, ΓΙΑ ΤΗΝ ΤΙΜΗ 100

Δρόμοι

Το αποτέλεσμα για την περίπτωση του οδικού δικτύου είναι διαφορετικό με εκείνο των κτιρίων. Πιο συγκεκριμένα, η ρύθμιση της τιμής των ελάχιστων δειγμάτων 0 σε 5 και στη συνέχεια σε 10 οδήγησε σε αύξηση του αριθμού των αντικειμένων της εν λόγω κλάσης. Το παραπάνω επέφερε μικρή αύξηση του ποσοστού της πληρότητας καθώς μικρό τμήμα οδικού άξονα το οποίο είχε ταξινομηθεί στην κατηγορία των χώρων στάθμευσης στο θεματικό χάρτη της τιμής 0 κατατάχθηκε σε εκείνη των δρόμων για τους 5 και 10. Το αποτέλεσμα, ωστόσο δεν είναι ικανοποιητικό σε ό,τι αφορά την ικανοποίηση του κριτηρίου της ορθότητας καθώς πολλά από τα αντικείμενα τα οποία προστέθηκαν στους χάρτες των 5 και 10 ανήκουν στην πραγματικότητα στην κλάση των χώρων στάθμευσης.

Το αποτέλεσμα είναι απογοητευτικό για τους χάρτες των 25, 50 και 100. Αναλυτικά, στην περίπτωση του χάρτη της τιμής 25 ο αριθμός των οδικών αξόνων έχει μειωθεί και το παραπάνω έχει σα συνέπεια τη μείωση του ποσοστού πληρότητας του παραγόμενου αποτελέσματος. Στα αποτελέσματα των 50 και 100 οι δείκτες ποιότητας είναι μηδενικοί καθώς η κλάση των δρόμων δεν εμφανίζεται καθόλου στον παραγόμενο θεματικό χάρτη (Εικόνα 3.121).

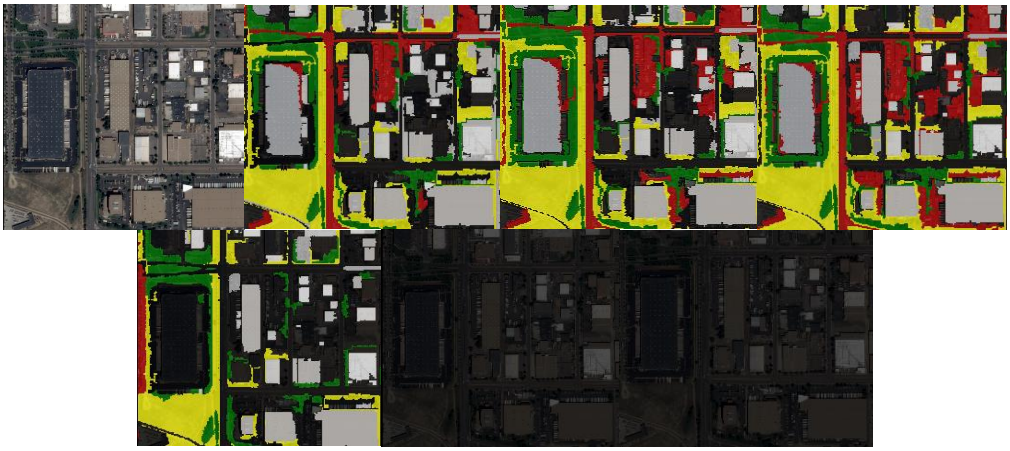


ΕΙΚΟΝΑ 3.121: ΑΠΟ ΤΗΝ ΑΡΧΗ: 2^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ, ΓΙΑ ΤΗΝ ΤΙΜΗ 0 ΤΩΝ ΕΛΑΧΙΣΤΩΝ ΔΕΙΓΜΑΤΩΝ, ΓΙΑ ΤΗΝ ΤΙΜΗ 5, ΓΙΑ ΤΗΝ ΤΙΜΗ 10, ΓΙΑ ΤΗΝ ΤΙΜΗ 25, ΓΙΑ ΤΗΝ ΤΙΜΗ 50, ΓΙΑ ΤΗΝ ΤΙΜΗ 100

Χώροι στάθμευσης

Η αύξηση της τιμής της παραμέτρου των ελάχιστων δειγμάτων οδήγησε σε αύξηση του αριθμού των αντικειμένων της κλάσης των χώρων στάθμευσης. Αναλυτικά, η ρύθμιση της τιμής της συγκεκριμένης παραμέτρου σε 5, 10 και 25 αύξησε σταδιακά τον αριθμό των αντικειμένων της κλάσης αυτής. Το παραπάνω είχε θετικά αποτελέσματα σε ό,τι αφορά την ικανοποίηση του κριτηρίου της πληρότητας της συγκεκριμένης κατηγορίας, καθώς πολλοί χώροι στάθμευσης οι οποίοι αρχικά είχαν ταξινομηθεί στην κατηγορία των κτιρίων κατατάχθηκαν έπειτα από την αύξηση της τιμής των ελάχιστων δειγμάτων στη σωστή κλάση. Προβλήματα, ωστόσο, δημιουργήθηκαν στην ορθότητα των νέων θεματικών χαρτών καθώς μέρος των αντικειμένων που προστέθηκαν στη συγκεκριμένη κατηγορία ανήκουν στην πραγματικότητα σε εκείνη των κτιρίων.

Η ρύθμιση της τιμής των ελάχιστων δειγμάτων σε 50 και 100 οδήγησε σε πληρότητα 100%. Το αποτέλεσμα, ωστόσο είναι απογοητευτικό σε ό,τι αφορά την ορθότητα και συνεπώς την ποιότητα των θεματικών χαρτών καθώς στην κλάση αυτή ταξινομήθηκαν όλα τα αντικείμενα της εικόνας (Εικόνα 3.122).



ΕΙΚΟΝΑ 3.122: ΑΠΟ ΤΗΝ ΑΡΧΗ: 3^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ, ΓΙΑ ΤΗΝ ΤΙΜΗ 0 ΤΩΝ ΕΛΑΧΙΣΤΩΝ ΔΕΙΓΜΑΤΩΝ, ΓΙΑ ΤΗΝ ΤΙΜΗ 5, ΓΙΑ ΤΗΝ ΤΙΜΗ 10, ΓΙΑ ΤΗΝ ΤΙΜΗ 25, ΓΙΑ ΤΗΝ ΤΙΜΗ 50, ΓΙΑ ΤΗΝ ΤΙΜΗ 100

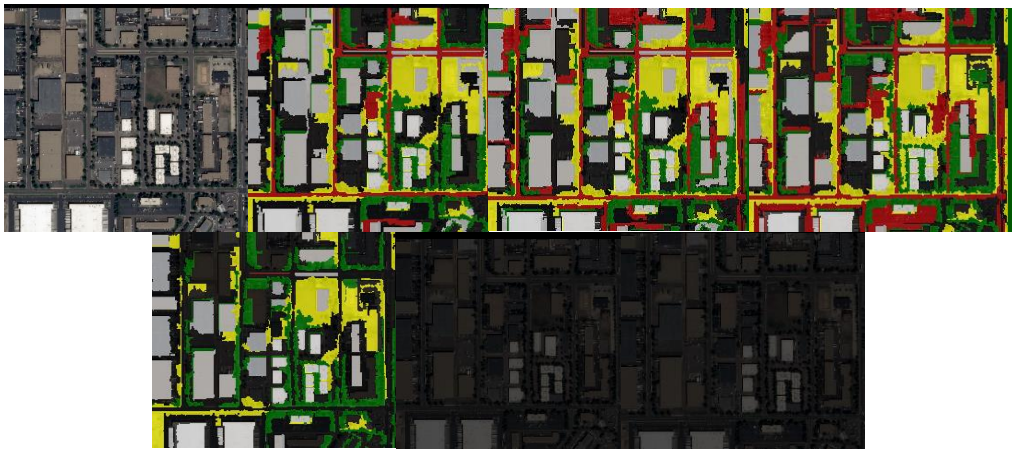
Αστικό πράσινο

Τα αποτελέσματα είναι περισσότερο περίπλοκα για τη θεματική κατηγορία του Αστικού Πρασίνου. Πιο συγκεκριμένα η μεταβολή της τιμής της παραμέτρου από 0 σε 5 οδήγησε σε μικρή αύξηση του αριθμού των αντικειμένων της κλάσης αυτής. Το παραπάνω δεν επηρέασε το ποσοστό πληρότητας, αλλά είχε αρνητικές συνέπειες για το κριτήριο της ορθότητας καθώς τα αντικείμενα που προστέθηκαν στην κλάση αυτή ανήκουν στην πραγματικότητα σε εκείνη των χώρων στάθμευσης.

Η ρύθμιση της τιμής της εν λόγω μεταβλητής από 5 σε 10 είχε σαν αποτέλεσμα τη μείωση των αντικειμένων της κλάσης αυτής. Αυτό επέδρασσε αρνητικά στο κριτήριο της πληρότητας καθώς από το θεματικό χάρτη της τιμής 10 απομακρύνθηκαν αντικείμενα τα οποία ανήκαν πράγματι στη συγκεκριμένη κλάση.

Η μεταβολή της τιμής του ελάχιστου αριθμού δειγμάτων από 10 σε 25 οδήγησε σε αύξηση του αριθμού των αντικειμένων που ανήκουν πράγματι στην κλάση του αστικού πρασίνου. Παράλληλα, ωστόσο, στην κατηγορία αυτή προστέθηκαν κάποια αντικείμενα χώρων στάθμευσης.

Τέλος, στους θεματικούς χάρτες των τιμών 50 και 100 δεν εμφανίζονται αντικείμενα αστικού πρασίνου. Βάσει αυτού προκύπτει πως τα ποσοστά πληρότητας και ορθότητας είναι μηδενικά (Εικόνα 3.123).



ΕΙΚΟΝΑ 3.123: ΑΠΟ ΤΗΝ ΑΡΧΗ: 4^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ, ΓΙΑ ΤΗΝ ΤΙΜΗ 0 ΤΩΝ ΕΛΑΧΙΣΤΩΝ ΔΕΙΓΜΑΤΩΝ, ΓΙΑ ΤΗΝ ΤΙΜΗ 5, ΓΙΑ ΤΗΝ ΤΙΜΗ 10, ΓΙΑ ΤΗΝ ΤΙΜΗ 25, ΓΙΑ ΤΗΝ ΤΙΜΗ 50, ΓΙΑ ΤΗΝ ΤΙΜΗ 100

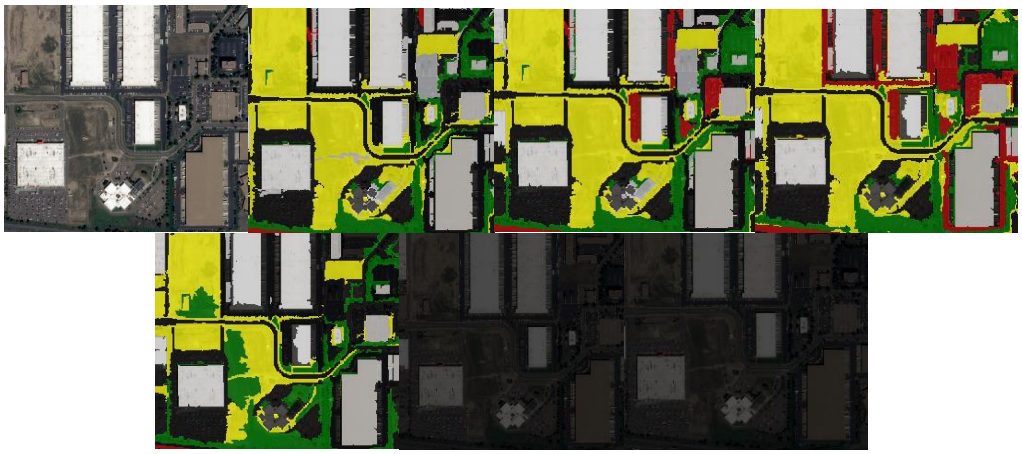
Άγονο Έδαφος

Η ρύθμιση της παραμέτρου του ελάχιστου αριθμού δειγμάτων από 0 σε 5 δεν επέφερε αλλαγή στα αντικείμενα της θεματικής κατηγορίας του Άγονου Εδάφους. Συνεπώς, τα κριτήρια της ορθότητας και της πληρότητας για τη συγκεκριμένη κλάση έχουν ακριβώς τα ίδια ποσοστά για τις δύο παραπάνω διαφορετικές τιμές.

Η αύξηση της τιμής της εν λόγω παραμέτρου από 5 σε 10 είχε σαν αποτέλεσμα μικρές μεταβολές στα αντικείμενα της κλάσης του Άγονου Εδάφους. Πιο συγκεκριμένα, στην τελευταία προστέθηκαν κάποιες περιοχές Αστικού Πρασίνου, γεγονός που είχε αρνητικές συνέπειες σε ό,τι αφορά την ορθότητα του παραγόμενου αποτελέσματος.

Η μεταβολή της τιμής του ελάχιστου αριθμού δειγμάτων από 10 σε 25 οδήγησε σε μείωση του αριθμού των αντικείμενων της παρούσας κλάσης. Το παραπάνω οδήγησε σε αύξηση της ορθότητας του τελικού αποτελέσματος καθώς από το θεματικό χάρτη της 20 αφαιρέθηκαν κάποιες περιοχές Άγονου Εδάφους.

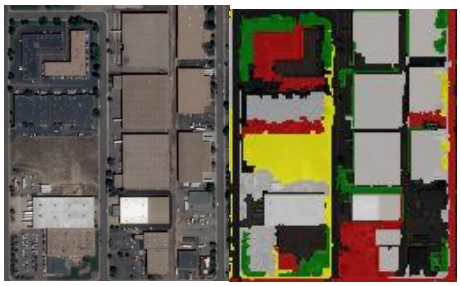
Τέλος, στους θεματικούς χάρτες των τιμών 50 και 100 οι περιοχές άγονου εδάφους έχουν ταξινομηθεί στο σύνολό τους στην κλάση των χώρων στάθμευσης. Συνεπώς, τα κριτήρια της ορθότητας και πληρότητας συμπληρώνουν ποσοστό 0% (Εικόνα 3.124).



ΕΙΚΟΝΑ 3.124: ΑΠΟ ΤΗΝ ΑΡΧΗ: 5^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ, ΓΙΑ ΤΗΝ ΤΙΜΗ 0 ΤΩΝ ΕΛΑΧΙΣΤΩΝ ΔΕΙΓΜΑΤΩΝ, ΓΙΑ ΤΗΝ ΤΙΜΗ 5, ΓΙΑ ΤΗΝ ΤΙΜΗ 10, ΓΙΑ ΤΗΝ ΤΙΜΗ 25, ΓΙΑ ΤΗΝ ΤΙΜΗ 50, ΓΙΑ ΤΗΝ ΤΙΜΗ 100

Ποσοτική αξιολόγηση αποτελεσμάτων

ΕΛΑΧΙΣΤΟΣ ΑΡΙΘΜΟΣ ΔΕΙΓΜΑΤΩΝ: 5



ΕΙΚΟΝΑ 3.125: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΙΜΗ ΕΛΑΧΙΣΤΩΝ ΔΕΙΓΜΑΤΩΝ ΙΣΗ ΜΕ 5

Βάσει της Εικόνα 3.125 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.33, Πίνακας 3.34)

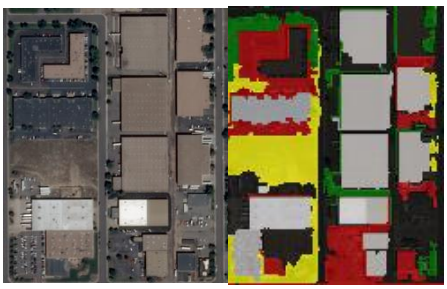
ΠΙΝΑΚΑΣ 3.33: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (3Η ΔΟΚΙΜΗ)

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	12	3	4

ΠΙΝΑΚΑΣ 3.34: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (3Η ΔΟΚΙΜΗ)

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	75,00%	80,00%	63,16%	25,00%	18,75%

ΕΛΑΧΙΣΤΟΣ ΑΡΙΘΜΟΣ ΔΕΙΓΜΑΤΩΝ: 10



ΕΙΚΟΝΑ 3.126: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΙΜΗ ΕΛΑΧΙΣΤΩΝ ΔΕΙΓΜΑΤΩΝ ΙΣΗ ΜΕ 10

Βάσει της Εικόνα 3.126 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.35, Πίνακας 3.36)

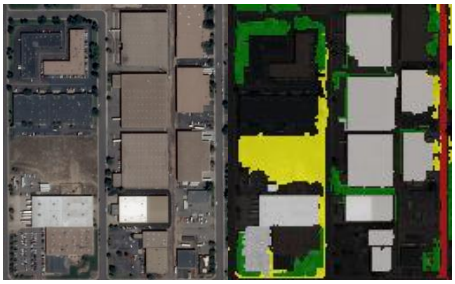
ΠΙΝΑΚΑΣ 3.35: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (3Η ΔΟΚΙΜΗ)

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	11	3	5

ΠΙΝΑΚΑΣ 3.36: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (3Η ΔΟΚΙΜΗ)

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	68,75%	78,57%	57,89%	31,25%	18,75%

ΕΛΑΧΙΣΤΟΣ ΑΡΙΘΜΟΣ ΔΕΙΓΜΑΤΩΝ: 25



ΕΙΚΟΝΑ 3.127: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΙΜΗ ΕΛΑΧΙΣΤΩΝ ΔΕΙΓΜΑΤΩΝ ΙΣΗ ΜΕ 25

Βάσει της Εικόνα 3.127 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.37, Πίνακας 3.38)

ΠΙΝΑΚΑΣ 3.37: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (3Η ΔΟΚΙΜΗ)

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	10	1	6

ΠΙΝΑΚΑΣ 3.38: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (3Η ΔΟΚΙΜΗ)

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	62,50%	90,91%	58,82%	37,50%	6,25%

ΕΛΑΧΙΣΤΟΣ ΑΡΙΘΜΟΣ ΔΕΙΓΜΑΤΩΝ: 50, 100

ΕΙΚΟΝΑ 3.128: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΙΜΗ ΕΛΑΧΙΣΤΩΝ ΔΕΙΓΜΑΤΩΝ ΙΣΗ ΜΕ 10

Βάσει της Εικόνα 3.128 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.39, Πίνακας 3.40)

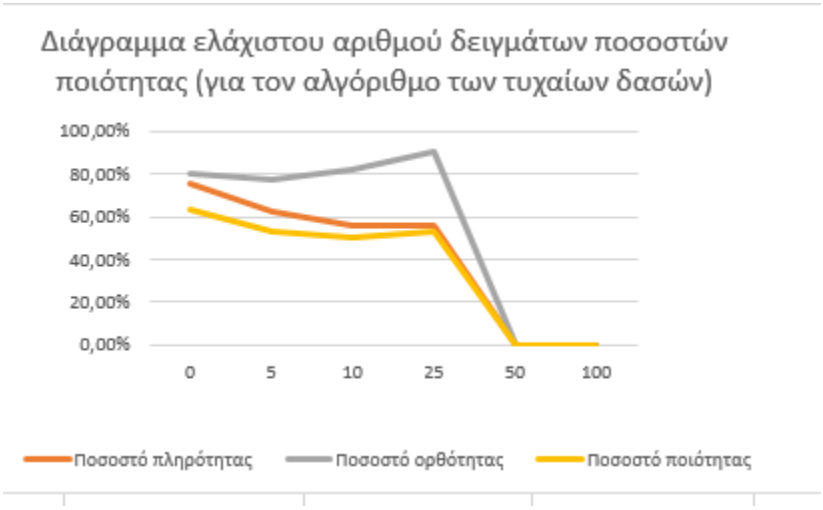
ΠΙΝΑΚΑΣ 3.39: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (3Η ΔΟΚΙΜΗ)

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	0	0	0

ΠΙΝΑΚΑΣ 3.40: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (3Η ΔΟΚΙΜΗ)

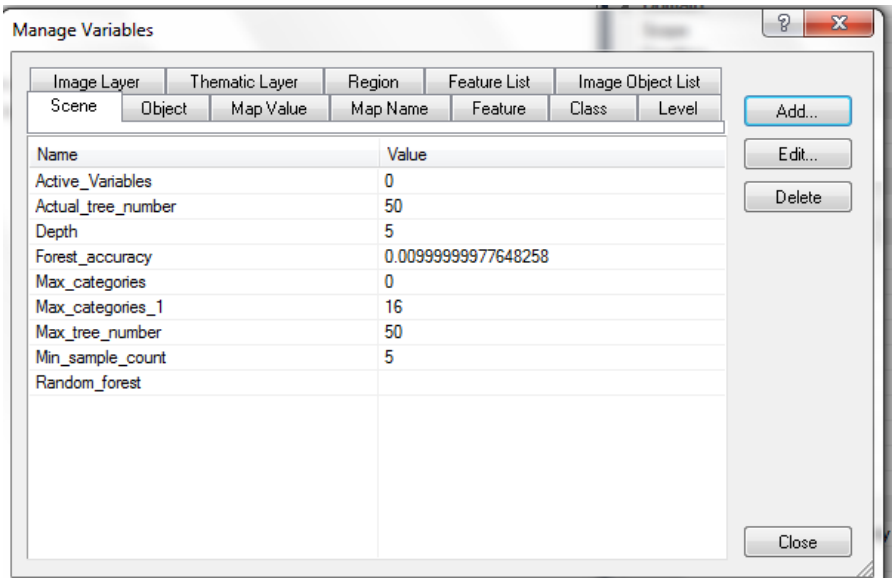
	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	0%	0%	0%	0%	0%

Στο ακόλουθο διάγραμμα (Εικόνα 3.129) εμφανίζονται τα ποσοστά πληρότητας συναρτήσει του αριθμού δειγμάτων

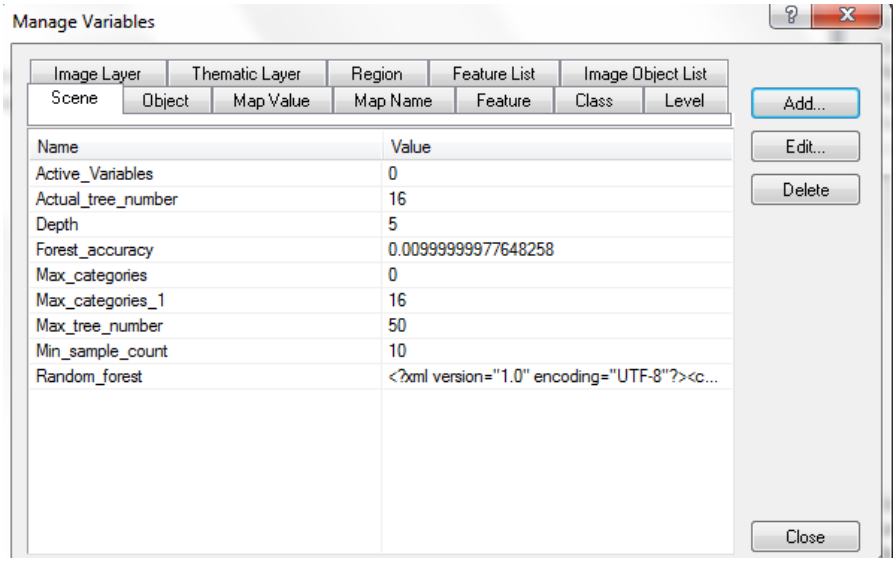


ΕΙΚΟΝΑ 3.129: ΔΙΑΓΡΑΜΜΑ ΑΡΙΘΜΟΥ ΔΕΙΓΜΑΤΩΝ ΠΟΣΟΣΤΩΝ ΠΟΙΟΤΗΤΑΣ

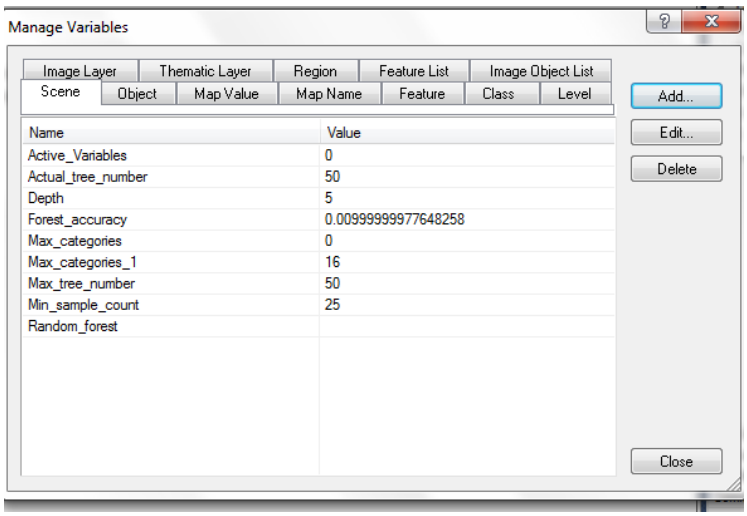
Στις ακόλουθες Εικόνες (Εικόνα 3.130- Εικόνα 3.134) εμφανίζονται οι πραγματικές τιμές των παραμέτρων έπειτα από τη δημιουργία των μοντέλων των τυχαίων δασών για τη δοκιμή 3. Η μεταβολή της μεταβλητής ελάχιστος αριθμός δειγμάτων επηρεάζει, όπως και στη δεύτερη δοκιμή, τη τιμή της μεταβλητής πλήθος δέντρων.



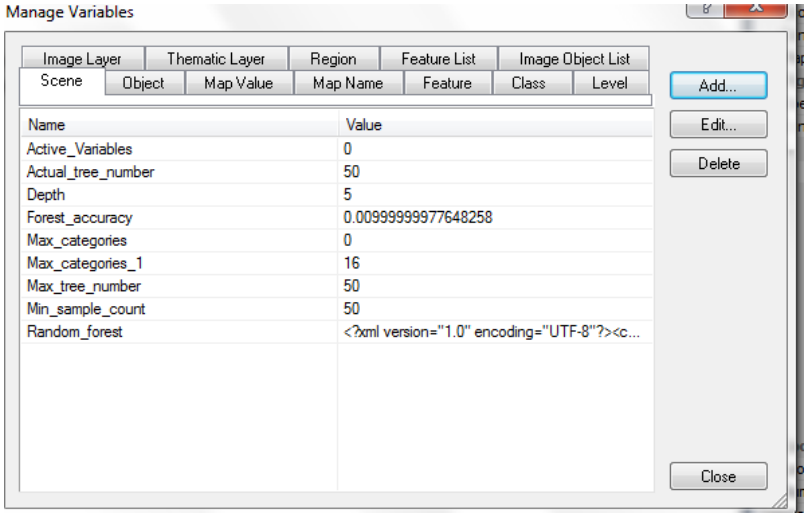
ΕΙΚΟΝΑ 3.130: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 3Η ΔΟΚΙΜΗ (ΕΛΑΧΙΣΤΟΣ ΑΡΙΘΜΟΣ ΔΕΙΓΜΑΤΩΝ: 5)



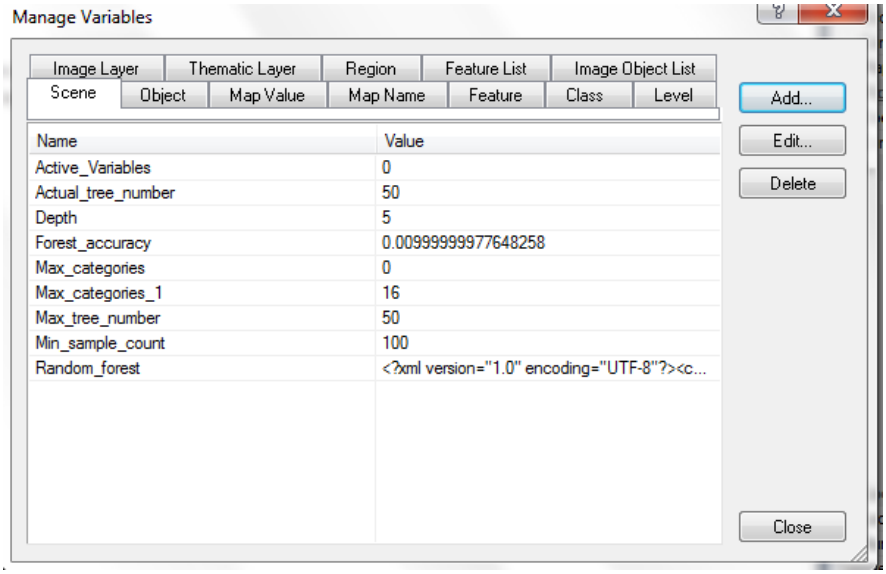
ΕΙΚΟΝΑ 3.131: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 3Η ΔΟΚΙΜΗ (ΕΛΑΧΙΣΤΟΣ ΑΡΙΘΜΟΣ ΔΕΙΓΜΑΤΩΝ: 10)



ΕΙΚΟΝΑ 3.132: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 3Η ΔΟΚΙΜΗ (ΕΛΑΧΙΣΤΟΣ ΑΡΙΘΜΟΣ ΔΕΙΓΜΑΤΩΝ: 25)



ΕΙΚΟΝΑ 3.133: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 3Η ΔΟΚΙΜΗ (ΕΛΑΧΙΣΤΟΣ ΑΡΙΘΜΟΣ ΔΕΙΓΜΑΤΩΝ: 50)



ΕΙΚΟΝΑ 3.134: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 3Η ΔΟΚΙΜΗ (ΕΛΑΧΙΣΤΟΣ ΑΡΙΘΜΟΣ ΔΕΙΓΜΑΤΩΝ: 100)

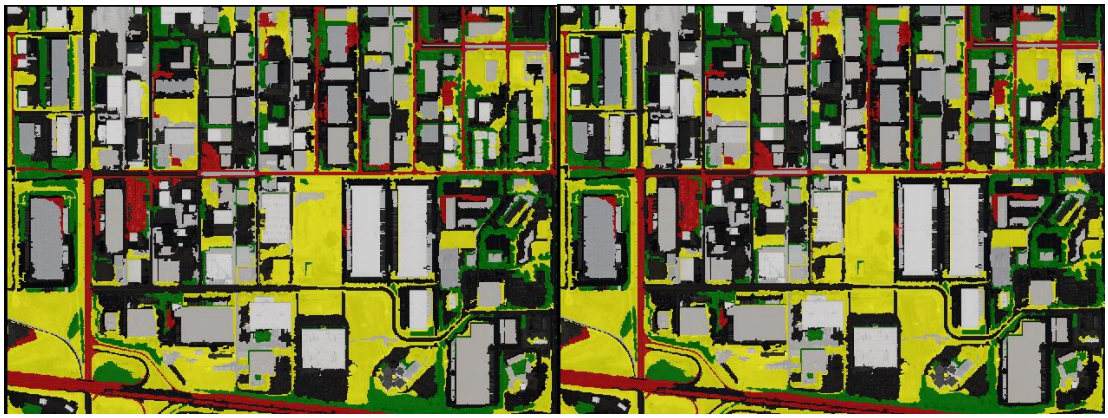
3.7.4 Δοκιμή 4 (Χρήση αντικαταστατών)

Στα πλαίσια της τέταρτης δοκιμής έγινε διερεύνηση της επιρροής της παραμέτρου που αφορά στη χρήση αντικαταστατών στην ποιότητα της ταξινόμησης. Πιο συγκεκριμένα ορίστηκαν οι ακόλουθες τιμές:

- Βάθος δέντρου (Depth): 0
- Ελάχιστος αριθμός δειγμάτων (Min sample count): 0
- **Χρήση αντικαταστατών (Use surrogates): Όχι (No), Ναι (Yes)**
- Μέγιστος αριθμός κατηγοριών (Max categories): 16
- Ενεργές μεταβλητές (Active Variables): 0 (δηλαδή ουσιαστικά για $\sqrt{9} = 3$)
- Μέγιστος αριθμός δέντρων (Max tree number): 50
- Ακρίβεια δάσους (Forest accuracy): 0.01
- Τύπος κριτηρίου τερματισμού (termination criteria type): Και τα δύο (Both)

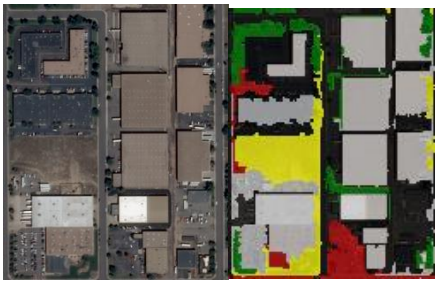
Σχολιασμός αποτελεσμάτων

Στην ακόλουθη Εικόνα (Εικόνα 3.135) εμφανίζεται το αποτέλεσμα εφαρμογής των συγκεκριμένων μοντέλων στην εικόνα εισόδου. Παρατηρείται πως ρύθμιση της παραμέτρου «χρήση αντικαταστατών» δεν επηρέασε αποτέλεσμα και συνεπώς την ποιότητα της ταξινόμησης.



ΕΙΚΟΝΑ 3.135:ΑΠΟΤΕΛΕΣΜΑ ΕΦΑΡΜΟΓΗΣ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΓΙΑ ΤΙΜΕΣ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΧΡΗΣΗ ΑΝΤΙΚΑΤΑΣΤΑΤΩΝ ΌΧΙ (ΔΕΞΙΑ), ΝΑΙ (ΑΡΙΣΤΕΡΑ)

Ποσοτική αξιολόγηση αποτελεσμάτων



ΕΙΚΟΝΑ 3.136: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΧΡΗΣΗ ΑΝΤΙΚΑΤΑΣΤΑΤΩΝ

Βάσει της Εικόνα 3.136 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.41, Πίνακας 3.42)

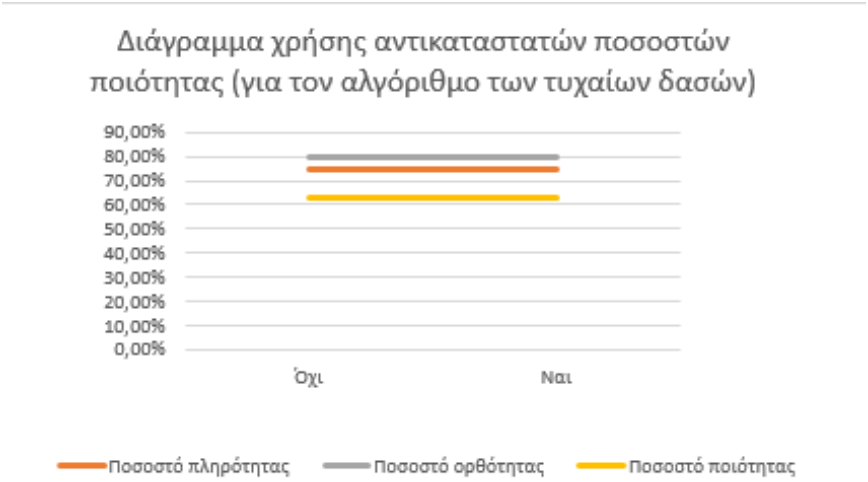
ΠΙΝΑΚΑΣ 3.41:ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (4Η ΔΟΚΙΜΗ).

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	14	3	2

ΠΙΝΑΚΑΣ 3.42: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (4Η ΔΟΚΙΜΗ).

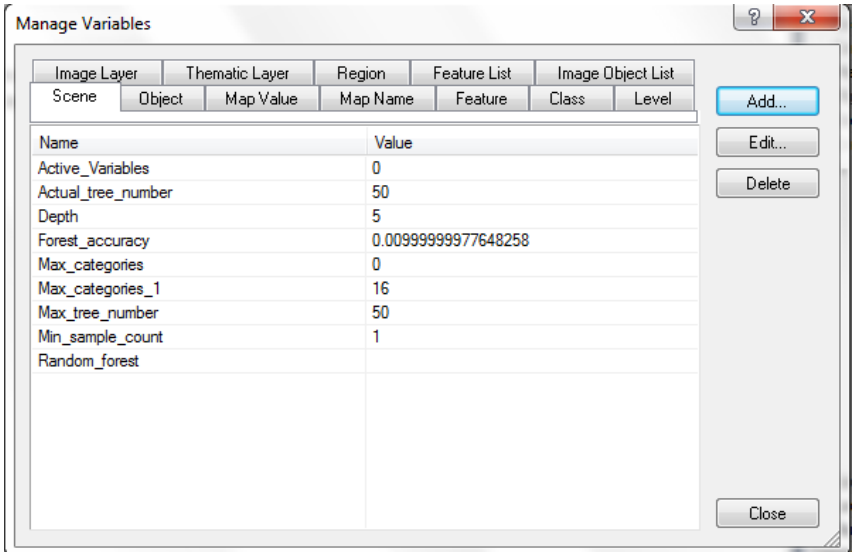
	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	87,50%	82,35%	73,68%	12,50%	18,75%

Στο ακόλουθο διάγραμμα (Εικόνα 3.137) εμφανίζονται τα ποσοστά πληρότητας συναρτήσει της χρήσης αντικαταστατών.



ΕΙΚΟΝΑ 3.137: ΔΙΑΓΡΑΜΜΑ ΧΡΗΣΗΣ ΑΝΤΙΚΑΤΑΣΤΑΤΩΝ ΠΟΣΟΣΤΩΝ ΠΟΙΟΤΗΤΑΣ

Στην Εικόνα 3.138 εμφανίζονται οι πραγματικές τιμές των παραμέτρων έπειτα από τη δημιουργία των μοντέλων των τυχαίων δασών για τη δοκιμή 4.



ΕΙΚΟΝΑ 3.138: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 4Η ΔΟΚΙΜΗ (ΧΡΗΣΗ ΑΝΤΙΚΑΤΑΣΤΑΤΩΝ: ΝΑΙ)

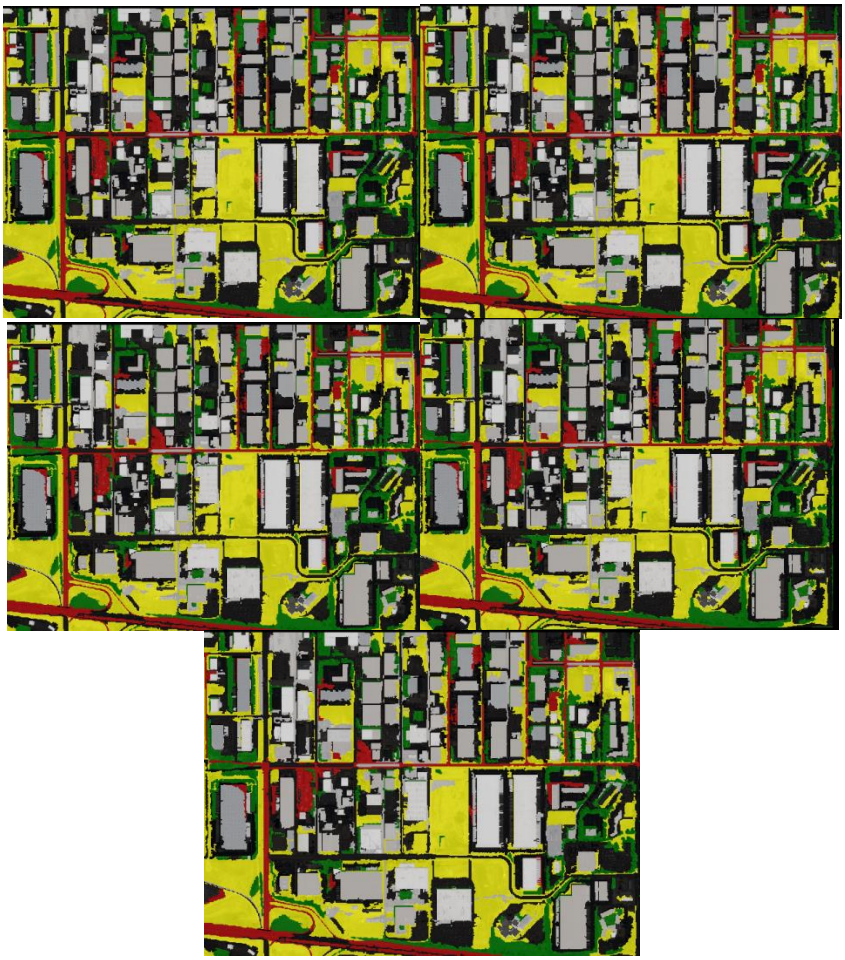
3.7.5 Δοκιμή 5 (Μέγιστος αριθμός κατηγοριών)

Στα πλαίσια της πέμπτης δοκιμής έγινε διερεύνηση της επιρροής της μεταβλητής των μέγιστων κατηγοριών στο αποτέλεσμα της ταξινόμησης. Αναλυτικά, δόθηκαν οι ακόλουθες τιμές:

- Βάθος δέντρου (Depth): 0
- Ελάχιστος αριθμός δειγμάτων (Min sample count): 0
- Χρήση αντικαταστατών (Use surrogates): Όχι (No)
- **Μέγιστος αριθμός κατηγοριών (Max categories): 2, 8, 16, 30, 100**
- Ενεργές μεταβλητές (Active Variables): 0 (δηλαδή ουσιαστικά για $\sqrt{9} = 3$)
- Μέγιστος αριθμός δέντρων (Max tree number): 50
- Ακρίβεια δάσους (Forest accuracy): 0.01
- Τύπος κριτηρίου τερματισμού (termination criteria type): Και τα δύο (Both)

Σχολιασμός αποτελεσμάτων

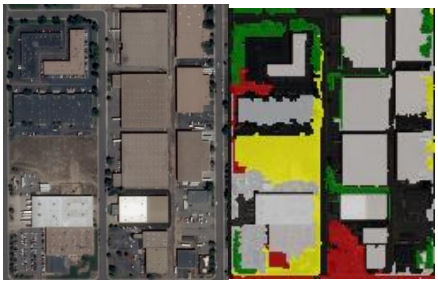
Στην Εικόνα 3.139 εμφανίζεται το αποτέλεσμα εφαρμογής του αλγορίθμου των τυχαίων δασών για τις διαφορετικές τιμές της παραμέτρου του μέγιστου αριθμού κατηγοριών. Είναι εμφανές πως η ρύθμιση της εν λόγω μεταβλητής δε μετέβαλε το αποτέλεσμα της ταξινόμησης.



ΕΙΚΟΝΑ 3.139: ΑΠΟΤΕΛΕΣΜΑ ΕΦΑΡΜΟΓΗΣ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΓΙΑ ΤΙΜΕΣ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΤΩΝ ΜΕΓΙΣΤΩΝ ΚΑΤΗΓΟΡΙΩΝ 2 (ΠΑΝΩ ΑΡΙΣΤΕΡΑ), 8 (ΠΑΝΩ ΔΕΞΙΑ), 16 (ΑΡΙΣΤΕΡΑ), 32 (ΔΕΞΙΑ), 100 (ΚΑΤΩ)

Ποσοτική αξιολόγηση αποτελεσμάτων

ΜΕΓΙΣΤΟΣ ΑΡΙΘΜΟΣ ΚΑΤΗΓΟΡΙΩΝ: 2, 8, 32, 100



ΕΙΚΟΝΑ 3.140: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΜΕΓΙΣΤΟ ΑΡΙΘΜΟ ΚΑΤΗΓΟΡΙΩΝ 2,8,32 100

Βάσει της Εικόνα 3.140 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.43, Πίνακας 3.44)

ΠΙΝΑΚΑΣ 3.43:ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (5Η ΔΟΚΙΜΗ).

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	1	3	2

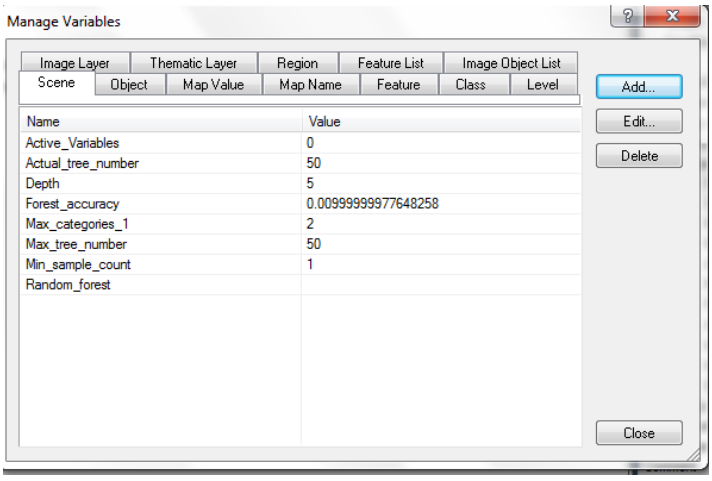
ΠΙΝΑΚΑΣ 3.44: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (5Η ΔΟΚΙΜΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	87,50%	82,35%	73,68%	12,50%	18,75%

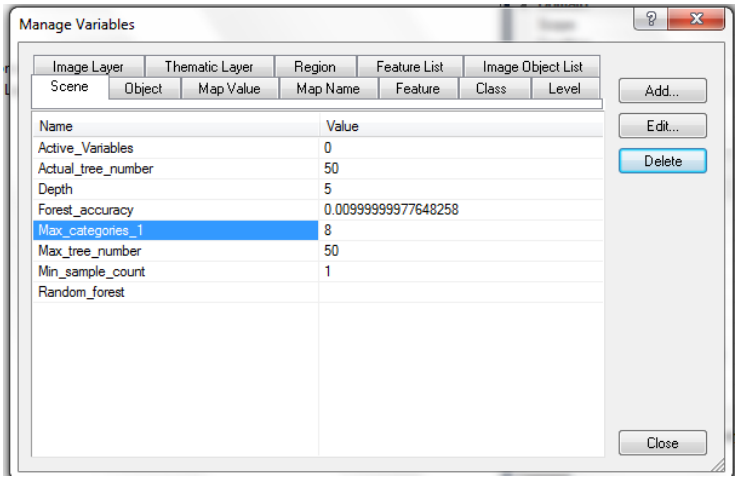
Στο ακόλουθο διάγραμμα (Εικόνα 3.141) εμφανίζονται τα ποσοστά πληρότητας συναρτήσει της χρήσης αντικαταστατών.

ΕΙΚΟΝΑ 3.141: ΔΙΑΓΡΑΜΜΑ ΧΡΗΣΗΣ ΑΝΤΙΚΑΤΑΣΤΑΤΩΝ ΠΟΣΟΣΤΩΝ ΠΟΙΟΤΗΤΑΣ

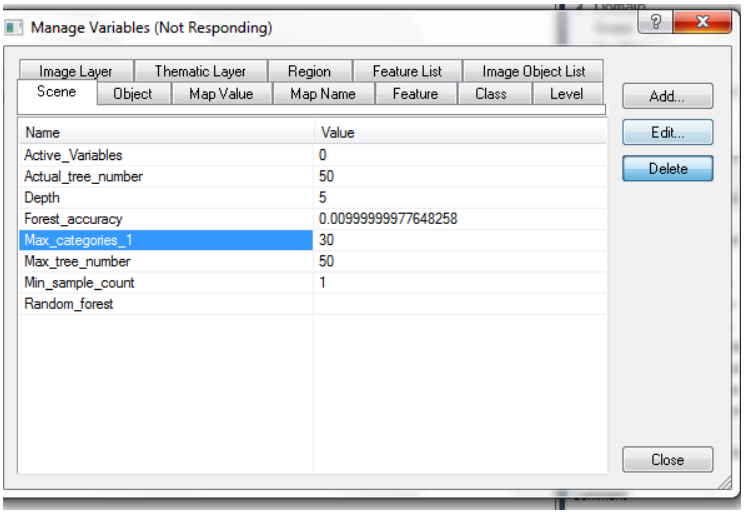
Στις ακόλουθες Εικόνες (Εικόνα 3.142- Εικόνα 3.144) εμφανίζονται οι πραγματικές τιμές των παραμέτρων έπειτα από τη δημιουργία των μοντέλων των τυχαίων δασών για τη δοκιμή 5.



ΕΙΚΟΝΑ 3.142: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 6Η ΔΟΚΙΜΗ (ΜΕΓΙΣΤΟΣ ΑΡΙΘΜΟΣ ΚΑΤΗΓΟΡΙΩΝ: 2)



ΕΙΚΟΝΑ 3.143: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 6Η ΔΟΚΙΜΗ (ΜΕΓΙΣΤΟΣ ΑΡΙΘΜΟΣ ΚΑΤΗΓΟΡΙΩΝ: 8)



ΕΙΚΟΝΑ 3.144: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 6Η ΔΟΚΙΜΗ (ΜΕΓΙΣΤΟΣ ΑΡΙΘΜΟΣ ΚΑΤΗΓΟΡΙΩΝ: 30)

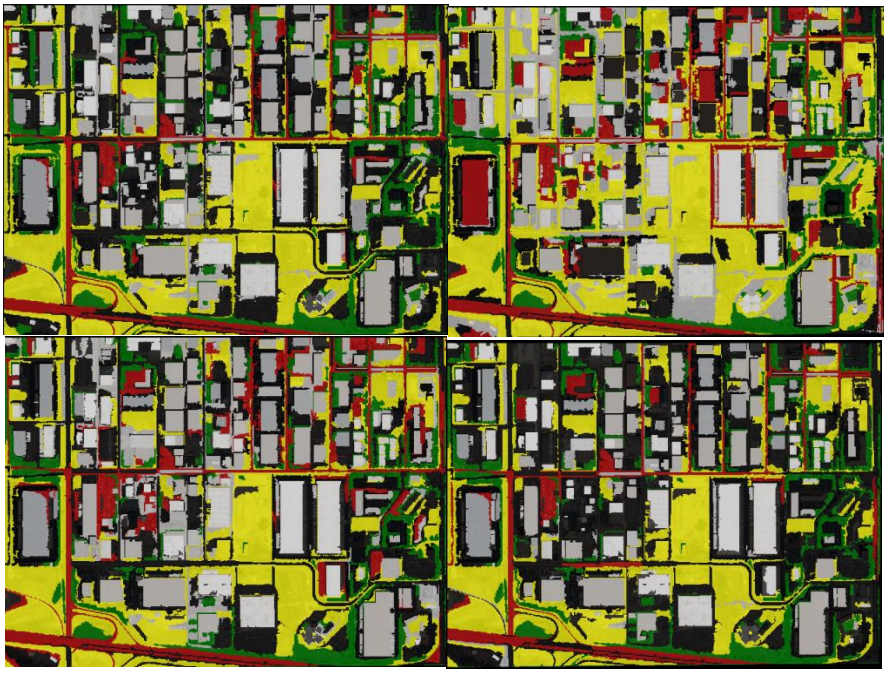
3.7.6 Δοκιμή 6 (Πλήθος ενεργών μεταβλητών)

Μέσω της έκτης δοκιμής έγινε διερεύνηση της επιρροής του πλήθους των ενεργών μεταβλητών στην ποιότητα της ταξινόμησης. Μέσω της συγκεκριμένης παραμέτρου προσδιορίζεται το μέγεθος του υποσυνόλου των χαρακτηριστικών τα οποία χρησιμοποιούνται για την εύρεση του βέλτιστου διαχωρισμού. Αναλυτικά, οι τιμές των μεταβλητών είναι οι ακόλουθες:

- Βάθος δέντρου (Depth): 0
- Ελάχιστος αριθμός δειγμάτων (Min sample count): 0
- Χρήση αντικαταστατών (Use surrogates): Όχι (No)
- Μέγιστος αριθμός κατηγοριών (Max categories): 16
- **Ενεργές μεταβλητές (Active Variables): 0 (δηλαδή ουσιαστικά για $\sqrt{9} = 3$), 2, 5, 9 (δηλαδή ο μέγιστος αριθμός χαρακτηριστικών)**
- Μέγιστος αριθμός δέντρων (Max tree number): 50ς
- Ακρίβεια δάσους (Forest accuracy): 0.01
- Τύπος κριτηρίου τερματισμού (termination criteria type): Και τα δύο (Both)

Σχολιασμός αποτελεσμάτων

Στην Εικόνα 3.145 εμφανίζεται το αποτέλεσμα εφαρμογής του αλγορίθμου των τυχαίων δασών για τις διαφορετικές τιμές της παραμέτρου των ενεργών μεταβλητών.



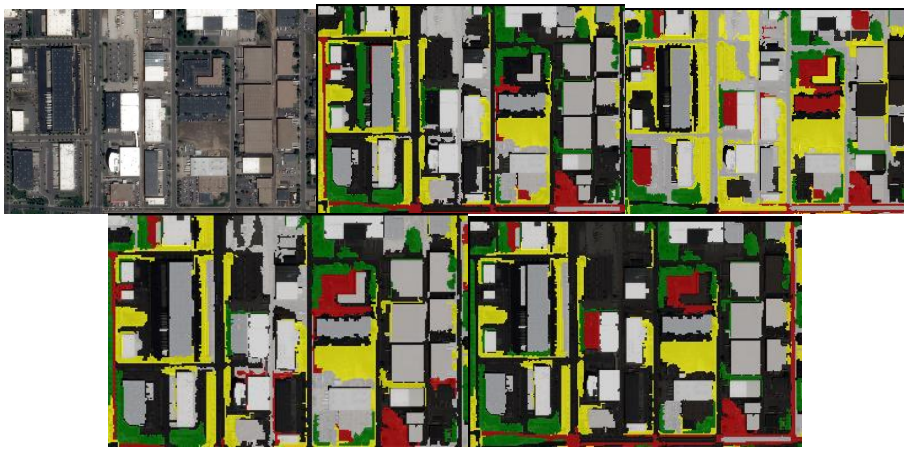
ΕΙΚΟΝΑ 3.145: ΑΠΟΤΕΛΕΣΜΑ ΕΦΑΡΜΟΓΗΣ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΓΙΑ ΤΙΜΕΣ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΕΝΕΡΓΕΣ ΜΕΤΑΒΛΗΤΕΣ 0 – ΔΗΛΑΔΗ $\sqrt{9} = 3$(ΠΑΝΩ ΑΡΙΣΤΕΡΑ), 2 (ΠΑΝΩ ΔΕΞΙΑ), 5 (ΚΑΤΩ ΑΡΙΣΤΕΡΑ), 9 (ΚΑΤΩ ΔΕΞΙΑ)

Κτίρια

Η μείωση του αριθμού των ενεργών μεταβλητών από 3 σε 2 επέφερε αρνητικά αποτελέσματα σε ό,τι αφορά τόσο την πληρότητα όσο και την ορθότητα της κλάσης των κτιρίων. Πιο συγκεκριμένα, στην κατηγορία αυτή προστέθηκαν πολλά αντικείμενα τα οποία ανήκουν σε εκείνη των δρόμων, των χώρων στάθμευσης και παράλληλα αφαιρέθηκαν από αυτήν κτίρια τα οποία καταχωρήθηκαν στους δρόμους, στους χώρους στάθμευσης και στο άγονο έδαφος.

Στη συνέχεια, η ρύθμιση του αριθμού των μεταβλητών σε 5 δε μετέβαλλε σε μεγάλο βαθμό την ποιότητα της ταξινόμησης σε σχέση με εκείνη της τιμής 3. Πιο συγκεκριμένα, το πλήθος των αντικειμένων της κλάσης αυτής σημείωσε μικρή αύξηση, η οποία ήταν επιβαρυντική σε ό,τι αφορά το κριτήριο της ορθότητας. Το παραπάνω οφείλεται στο γεγονός πως στην κλάση αυτή προστέθηκαν χώροι στάθμευσης.

Η μεταβολή του πλήθους των μεταβλητών σε 9 μείωσε τον αριθμό των αντικειμένων της κλάσης των κτιρίων. Το παραπάνω επηρέασε αρνητικά το κριτήριο της πληρότητας καθώς στο νέο θεματικό χάρτη αφαιρέθηκαν από την κλάση αυτή πολλά κτίρια. Παράλληλα, το κριτήριο της ορθότητας αυξήθηκε καθώς τα αντικείμενα της κατηγορίας αυτής ανήκουν πράγματι στο σύνολο τους σχεδόν στη συγκεκριμένη (Εικόνα 3.146).



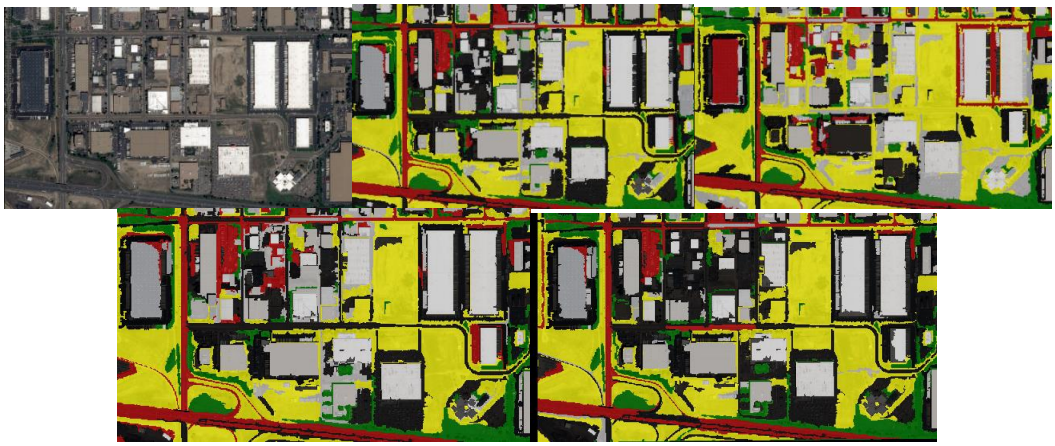
ΕΙΚΟΝΑ 3.146: ΑΠΟ ΤΗΝ ΑΡΧΗ: 1^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ ΓΙΑ ΤΙΜΕΣ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΕΝΕΡΓΕΣ ΜΕΤΑΒΛΗΤΕΣ 0 – ΔΗΛΑΔΗ $\sqrt{9} = 3$(ΠΑΝΩ ΑΡΙΣΤΕΡΑ), 2 (ΠΑΝΩ ΔΕΞΙΑ), 5 (ΚΑΤΩ ΑΡΙΣΤΕΡΑ), 9 (ΚΑΤΩ ΔΕΞΙΑ)

Δρόμοι

Η ρύθμιση της παραμέτρου των ενεργών μεταβλητών από 3 σε 2 αύξησε τον αριθμό των αντικειμένων του οδικού δικτύου. Το παραπάνω, ωστόσο, μείωσε το ποσοστό της ορθότητας του παραγόμενου αποτελέσματος, καθώς τα αντικείμενα τα οποία προστέθηκαν ανήκουν στην πραγματικότητα σε εκείνη των κτιρίων και των χώρων στάθμευσης.

Ακριβώς το αντίστοιχο συνέβη κατά τη ρύθμιση των ενεργών μεταβλητών από 3 σε 5.

Τέλος, η αύξηση των παραμέτρων από 5 σε 9 μείωσε τον αριθμό των εμφανιζόμενων δρόμων. Το παραπάνω επέδρασε θετικά στην ορθότητα της ταξινόμησης καθώς από την κλάση αυτή απομακρύνθηκαν οι χώροι στάθμευσης. Παράλληλα, αυξήθηκε η πληρότητα της εν λόγω θεματικής κατηγορίας, διότι προστέθηκαν σε αυτήν τμήματα του οδικού δικτύου τα οποία δεν εμφανίζονταν στην περίπτωση των τιμών 2,3 και 5 (Εικόνα 3.147).



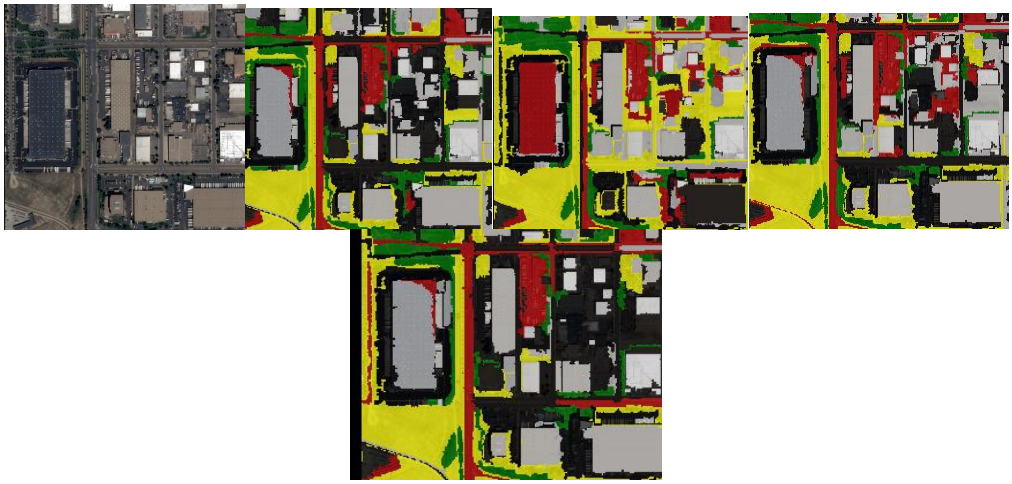
ΕΙΚΟΝΑ 3.147: ΑΠΟ ΤΗΝ ΑΡΧΗ: 2^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ ΓΙΑ ΤΙΜΕΣ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΕΝΕΡΓΕΣ ΜΕΤΑΒΛΗΤΕΣ 0 – ΔΗΛΑΔΗ $\sqrt{9} = 3$(ΠΑΝΩ ΑΡΙΣΤΕΡΑ), 2 (ΠΑΝΩ ΔΕΞΙΑ), 5 (ΚΑΤΩ ΑΡΙΣΤΕΡΑ), 9 (ΚΑΤΩ ΔΕΞΙΑ)

Χώροι στάθμευσης

Η μείωση των ενεργών μεταβλητών από 3 σε 2 επηρέασε αρνητικά τόσο την ορθότητα όσο και την ποιότητα της θεματικής κατηγορίας των Χώρων στάθμευσης. Πιο συγκεκριμένα, στην κλάση αυτή προστέθηκαν πολλά αντικείμενα τα οποία στην πραγματικότητα ανήκουν σε εκείνη των κτιρίων και παράλληλα αφαιρέθηκαν από τον παραγόμενο θεματικό χάρτη χώροι στάθμευσης.

Η αύξηση των ενεργών μεταβλητών από 3 σε 5 δεν επηρέασε σε μεγάλο βαθμό την ποιότητα της ταξινόμησης αναφορικά με τους χώρους στάθμευσης.

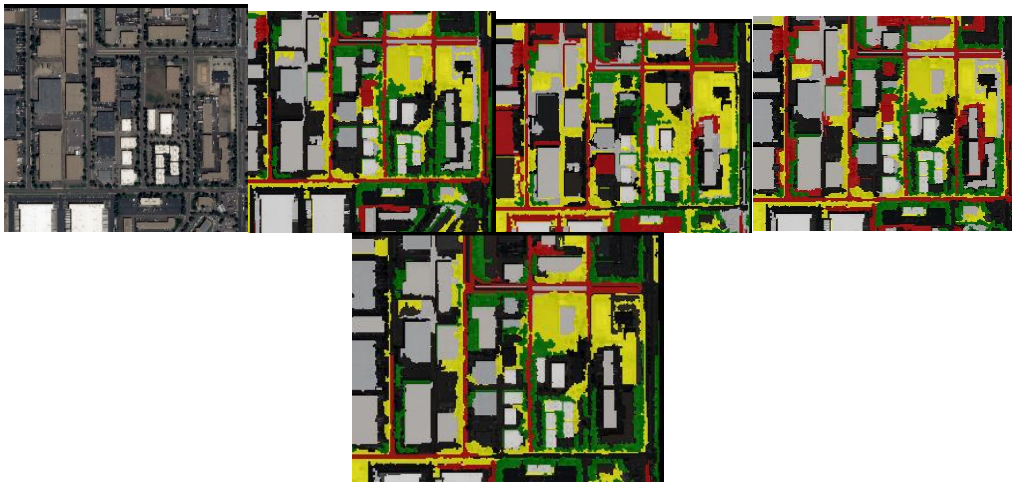
Τέλος, η ρύθμιση της παραμέτρου σε 9 αύξησε το πλήθος των αντικειμένων της υπό μελέτη κλάσης. Το παραπάνω αύξησε την πληρότητα της συγκεκριμένης θεματικής κατηγορίας και παράλληλα μείωσε το ποσοστό της ορθότητας (Εικόνα 3.148).



ΕΙΚΟΝΑ 3.148: ΑΠΟ ΤΗΝ ΑΡΧΗ: 3ο ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ ΓΙΑ ΤΙΜΕΣ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΕΝΕΡΓΕΣ ΜΕΤΑΒΛΗΤΕΣ 0 – ΔΗΛΑΔΗ $\sqrt{9} = 3$(ΠΑΝΩ ΑΡΙΣΤΕΡΑ), 2 (ΠΑΝΩ ΔΕΞΙΑ), 5 (ΚΑΤΩ ΑΡΙΣΤΕΡΑ), 9 (ΚΑΤΩ ΔΕΞΙΑ)

Αστικό πράσινο

Η ρύθμιση της παραμέτρου των ενεργών μεταβλητών επηρέασε σε πολύ μικρό βαθμό το αποτέλεσμα της ταξινόμησης για την κατηγορία του αστικού πρασίνου (Εικόνα 3.149).

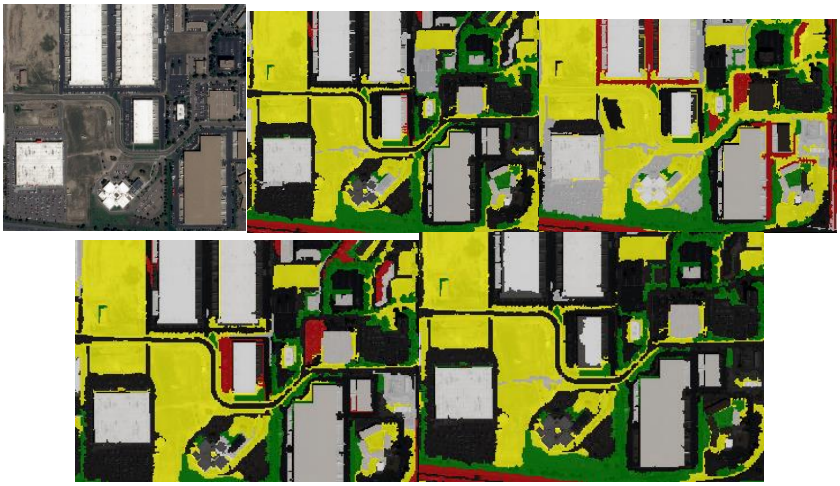


ΕΙΚΟΝΑ 3.149: ΑΠΟ ΤΗΝ ΑΡΧΗ: 4ο ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ ΓΙΑ ΤΙΜΕΣ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΕΝΕΡΓΕΣ ΜΕΤΑΒΛΗΤΕΣ 0 – ΔΗΛΑΔΗ $\sqrt{9} = 3$(ΠΑΝΩ ΑΡΙΣΤΕΡΑ), 2 (ΠΑΝΩ ΔΕΞΙΑ), 5 (ΚΑΤΩ ΑΡΙΣΤΕΡΑ), 9 (ΚΑΤΩ ΔΕΞΙΑ)

Άγονο Έδαφος

Η μείωση των ενεργών μεταβλητών από 3 σε 2 αύξησε το πλήθος των αντικειμένων του άγονου εδάφους. Το παραπάνω ελάττωσε σε μεγάλο βαθμό την ορθότητα του αποτελέσματος καθώς στην κλάση αυτή προστέθηκαν πολλοί άξονες του οδικού δικτύου, χώροι στάθμευσης και κτίρια.

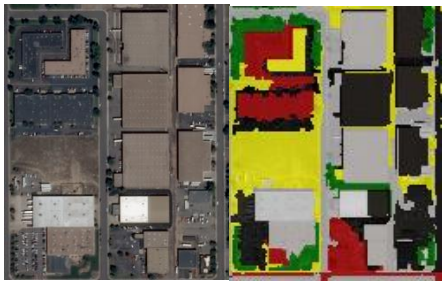
Η ρύθμιση της παραμέτρου σε 5 και 9 εμφάνισε ακριβώς τα ίδια αποτελέσματα για την κατηγορία του άγονου εδάφους με εκείνη της τιμής 3 (Εικόνα 3.150).



ΕΙΚΟΝΑ 3.150: ΑΠΟ ΤΗΝ ΑΡΧΗ: 5^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ ΓΙΑ ΤΙΜΕΣ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΕΝΕΡΓΕΣ ΜΕΤΑΒΛΗΤΕΣ 0 – ΔΗΛΑΔΗ $\sqrt{9} = 3$(ΠΑΝΩ ΑΡΙΣΤΕΡΑ), 2 (ΠΑΝΩ ΔΕΞΙΑ), 5 (ΚΑΤΩ ΑΡΙΣΤΕΡΑ), 9 (ΚΑΤΩ ΔΕΞΙΑ)

Ποσοτική αξιολόγηση αποτελεσμάτων

ΠΛΗΘΟΣ ΕΝΕΡΓΩΝ ΜΕΤΑΒΛΗΤΩΝ: 2



ΕΙΚΟΝΑ 3.151: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΠΛΗΘΟΣ ΕΝΕΡΓΩΝ ΜΕΤΑΒΛΗΤΩΝ ΙΣΟ ΜΕ 2

Βάσει της Εικόνα 3.151 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.45, Πίνακας 3.46)

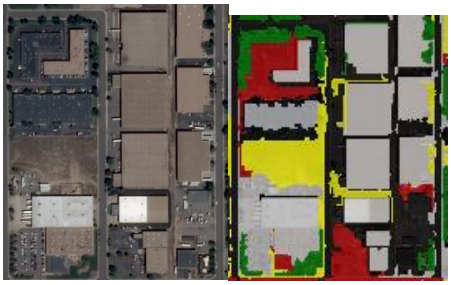
ΠΙΝΑΚΑΣ 3.45:ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (6^{Η} ΔΟΚΙΜΗ).

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	7	4	9

ΠΙΝΑΚΑΣ 3.46: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (6^{Η} ΔΟΚΙΜΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	43,75%	63,64%	35,00%	56,25%	25,00%

ΠΛΗΘΟΣ ΕΝΕΡΓΩΝ ΜΕΤΑΒΛΗΤΩΝ: 5



ΕΙΚΟΝΑ 3.152: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΠΛΗΘΟΣ ΕΝΕΡΓΩΝ ΜΕΤΑΒΛΗΤΩΝ ΙΣΟ ΜΕ 5

Βάσει της Εικόνα 3.152 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.47, Πίνακας 3.48)

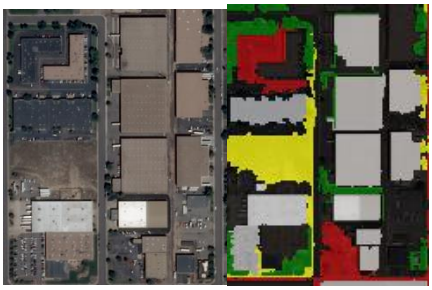
ΠΙΝΑΚΑΣ 3.47:ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (6Η ΔΟΚΙΜΗ).

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	14	3	2

ΠΙΝΑΚΑΣ 3.48: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (6Η ΔΟΚΙΜΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	87,50%	82,35%	73,68%	12,50%	18,75%

ΠΛΗΘΟΣ ΕΝΕΡΓΩΝ ΜΕΤΑΒΛΗΤΩΝ: 9



ΕΙΚΟΝΑ 3.153: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΠΛΗΘΟΣ ΕΝΕΡΓΩΝ ΜΕΤΑΒΛΗΤΩΝ ΙΣΟ ΜΕ 9

Βάσει της Εικόνα 3.153 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.49, Πίνακας 3.50)

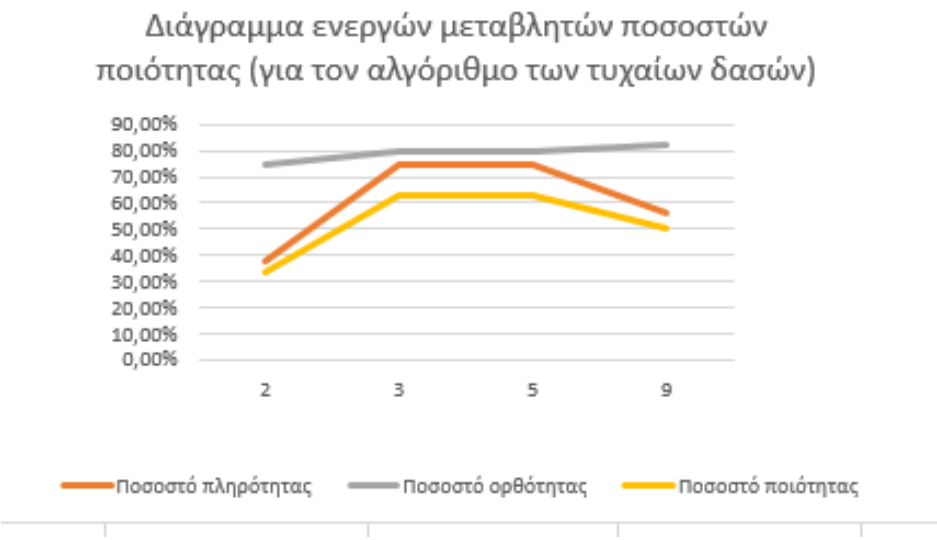
ΠΙΝΑΚΑΣ 3.49:ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (6Η ΔΟΚΙΜΗ).

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	11	2	5

ΠΙΝΑΚΑΣ 3.50: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (6Η ΔΟΚΙΜΗ).

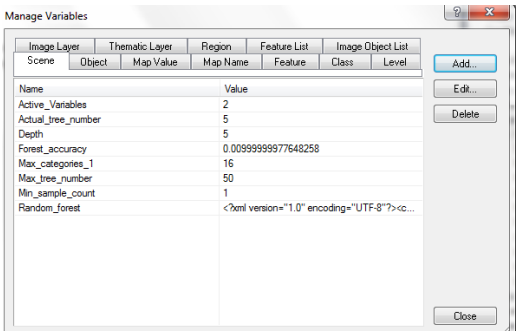
	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	68,75%	84,62%	61,11%	31,25%	12,50%

Στην ακόλουθη Εικόνα (Εικόνα 3.154) εμφανίζεται το διάγραμμα πλήθους κατηγοριών ποσοστών πληρότητας.

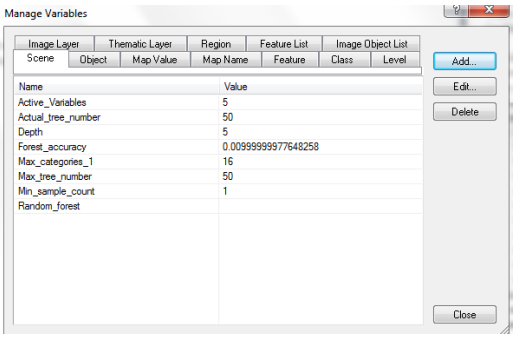


ΕΙΚΟΝΑ 3.154: ΔΙΑΓΡΑΜΜΑ ΠΛΗΘΟΣ ΚΑΤΗΓΟΡΙΩΝ ΠΟΣΟΣΤΩΝ ΠΟΙΟΤΗΤΑΣ

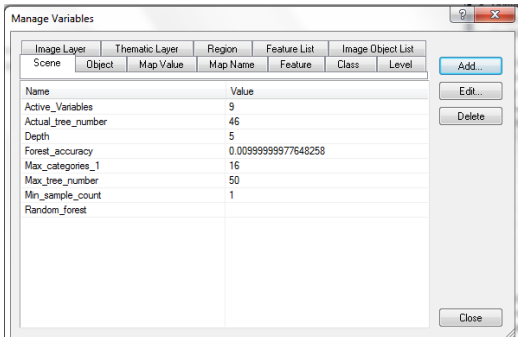
Στις ακόλουθες Εικόνες (Εικόνα 3.155- Εικόνα 3.157) εμφανίζονται οι πραγματικές τιμές των παραμέτρων έπειτα από τη δημιουργία των μοντέλων των τυχαίων δασών για τη δοκιμή 5.



ΕΙΚΟΝΑ 3.155: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 5Η ΔΟΚΙΜΗ (ΕΝΕΡΓΕΣ ΜΕΤΑΒΛΗΤΕΣ: 2)



ΕΙΚΟΝΑ 3.156: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 5Η ΔΟΚΙΜΗ (ΕΝΕΡΓΕΣ ΜΕΤΑΒΛΗΤΕΣ: 5)



ΕΙΚΟΝΑ 3.157: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 5Η ΔΟΚΙΜΗ (ΕΝΕΡΓΕΣ ΜΕΤΑΒΛΗΤΕΣ: 9)

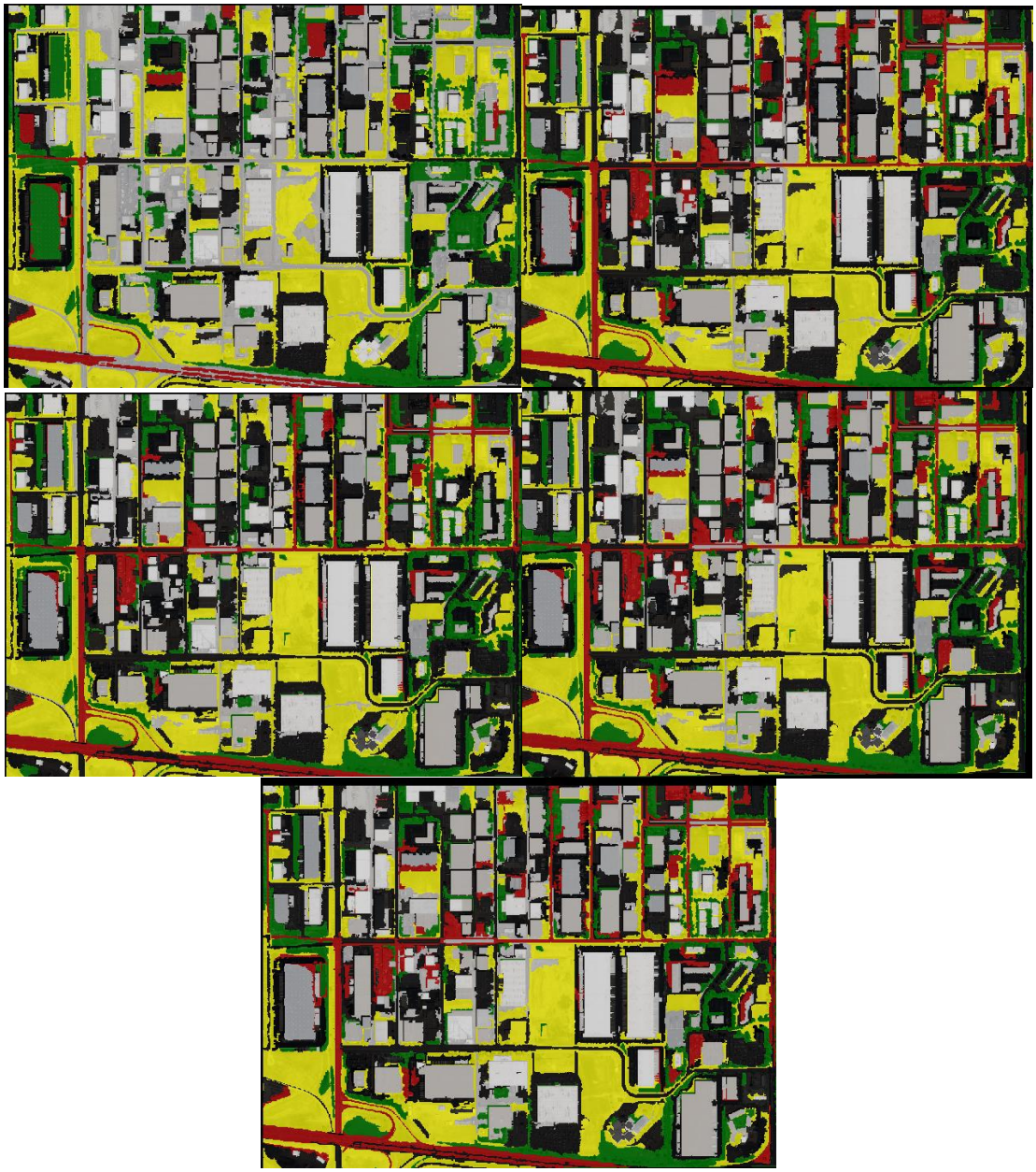
3.7.7 Δοκιμή 7 (Πλήθος δέντρων απόφασης)

Στα πλαίσια της δοκιμής αυτής μελετήθηκε η επιρροή του πλήθους των δέντρων απόφασης ενός τυχαίου δάσους στην ποιότητα της ταξινόμησης. Πιο συγκεκριμένα, ορίστηκαν οι ακόλουθες τιμές:

- Βάθος δέντρου (Depth): 0
- Ελάχιστος αριθμός δειγμάτων (Min sample count): 0
- Χρήση αντικαταστατών (Use surrogates): Όχι (No)
- Μέγιστος αριθμός κατηγοριών (Max categories): 16
- Ενεργές μεταβλητές (Active Variables): 0 (δηλαδή ουσιαστικά για $\sqrt{9} = 3$)
- **Μέγιστος αριθμός δέντρων (Max tree number): 5, 20, 50, 100, 1000**
- Ακρίβεια δάσους (Forest accuracy): 0.01
- Τύπος κριτηρίου τερματισμού (termination criteria type): Και τα δύο (Both)

Σχολιασμός αποτελεσμάτων

Στην Εικόνα 3.158ν εμφανίζεται το αποτέλεσμα εφαρμογής του αλγορίθμου των τυχαίων δασών για διαφορετικές τιμές της παραμέτρου πλήθος των δέντρων.



ΕΙΚΟΝΑ 3.158: ΑΠΟΤΕΛΕΣΜΑ ΕΦΑΡΜΟΓΗΣ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΓΙΑ ΤΙΜΕΣ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΠΛΗΘΟΣ ΔΕΝΤΡΩΝ 5 (ΠΑΝΩ ΑΡΙΣΤΕΡΑ), 20 (ΠΑΝΩ ΔΕΞΙΑ), 50 (ΚΑΤΩ ΑΡΙΣΤΕΡΑ- Η ΠΡΟΚΑΘΟΡΙΣΜΕΝΗ), 100 (ΚΑΤΩ ΔΕΞΙΑ), 1000 (Η ΤΕΛΕΥΤΑΙΑ)

Κτίρια

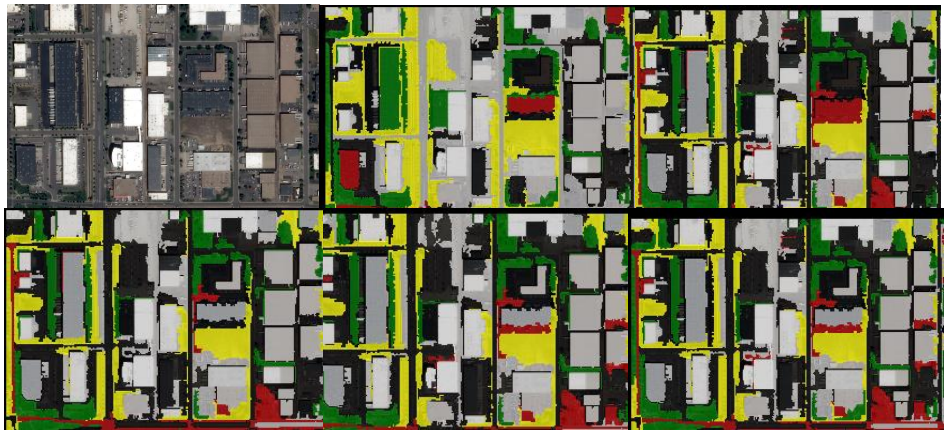
Η ρύθμιση της παραμέτρου «πλήθος των δέντρων» από 50 σε 5 αύξησε τον αριθμό των κτιρίων στον παραγόμενο θεματικό χάρτη. Το παραπάνω επιβάρυνε την ορθότητα της ταξινόμησης για τη συγκεκριμένη κλάση καθώς τα αντικείμενα τα οποία προστέθηκαν ανήκουν στην πραγματικότητα στους χώρους στάθμευσης, στους δρόμους, στο αστικό πράσινο και στο άγονο έδαφος. Παράλληλα, από την κλάση αυτή αφαιρέθηκαν κτίρια, γεγονός που λειτούργησε επιβαρυντικά στην ικανοποίηση του κριτηρίου της πληρότητας.

Η αύξηση της παραμέτρου από 5 σε 20 μείωσε τον αριθμό των εμφανιζόμενων κτιρίων. Το παραπάνω αύξησε το ποσοστό της ορθότητας καθώς τα αντικείμενα τα οποία αφαιρέθηκαν από την κατηγορία αυτή δεν ανήκαν στην πραγματικότητα σε αυτήν (χώροι στάθμευσης, άγονο έδαφος, δρόμοι). Παράλληλα, ωστόσο, αφαιρέθηκε ένα μικρός αριθμός κτιρίων,

γεγονός που λειτούργησε επιβαρυντικά σε ό,τι αφορά την ικανοποίηση του κριτηρίου της πληρότητας.

Η αύξηση της παραμέτρου πλήθος δέντρων από 50 σε 100 μείωσε ελάχιστα το πλήθος των αντικειμένων της κλάσης αυτής, γεγονός που λειτούργησε βοηθητικά σε ό,τι αφορά την ικανοποίηση του κριτηρίου της ορθότητας.

Τέλος, σημειώνεται πως η αύξηση του πλήθους των δέντρων από 100 σε 1000 έδωσε πανομοιότυπο αποτέλεσμα με εκείνο των 50 (Εικόνα 3.159).



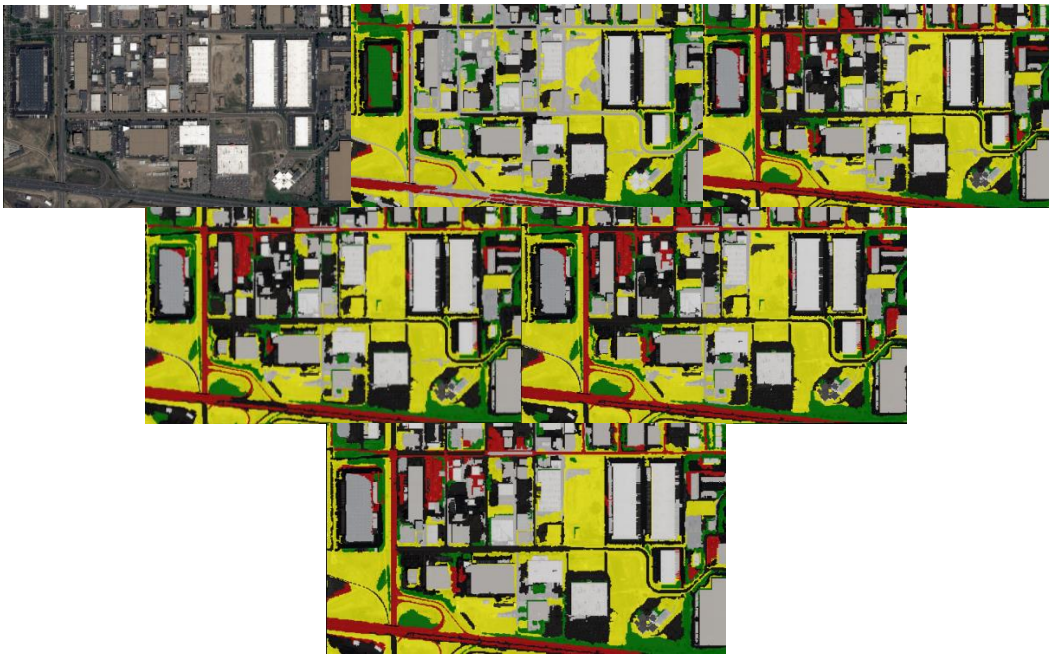
ΕΙΚΟΝΑ 3.159: ΑΠΟ ΤΗΝ ΑΡΧΗ: 1^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ ΓΙΑ ΤΙΜΕΣ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΠΛΗΘΟΣ ΔΕΝΤΡΩΝ 5, 20, 50, 100, 1000

Δρόμοι

Η θεματική κατηγορία των δρόμων εμφάνισε μείωση τόσο στην ορθότητα όσο και στην πληρότητα της ταξινόμησης όταν το πλήθος των δέντρων μειώθηκε από 50 σε 5. Το παραπάνω οφείλεται στο γεγονός πως πολλά αντικείμενα της κατηγορίας αυτής καταχωρήθηκαν σε εκείνη των κτιρίων και το αντίστροφο.

Η αύξηση του πλήθος από 5 σε 20 μείωσε τον αριθμό των εμφανιζόμενων δρόμων στο νέο θεματικό χάρτη. Το παραπάνω επέδρασσε θετικά στην ικανοποίηση του κριτηρίου της πληρότητας καθώς στην κλάση αυτή προστέθηκαν άξονες του οδικού δικτύου. Παράλληλα, ωστόσο, εξακολουθούν να εμφανίζονται προβλήματα σχετικά με την ορθότητα του παραγόμενου αποτελέσματος καθώς στην κατηγορία αυτή έχουν ενταχθεί πολλά από τα απεικονιζόμενα κτίρια.

Τέλος, σημειώνεται πως η αύξηση του πλήθους των δέντρων από 50 σε 100 και 1000 λειτούργησε επιβαρυντικά για το κριτήριο της ορθότητας. Το παραπάνω οφείλεται στο γεγονός πως στους δύο νέους θεματικούς χάρτες προστέθηκαν αντικείμενα τα οποία ανήκουν στην πραγματικότητα στα κτίρια (Εικόνα 3.160).

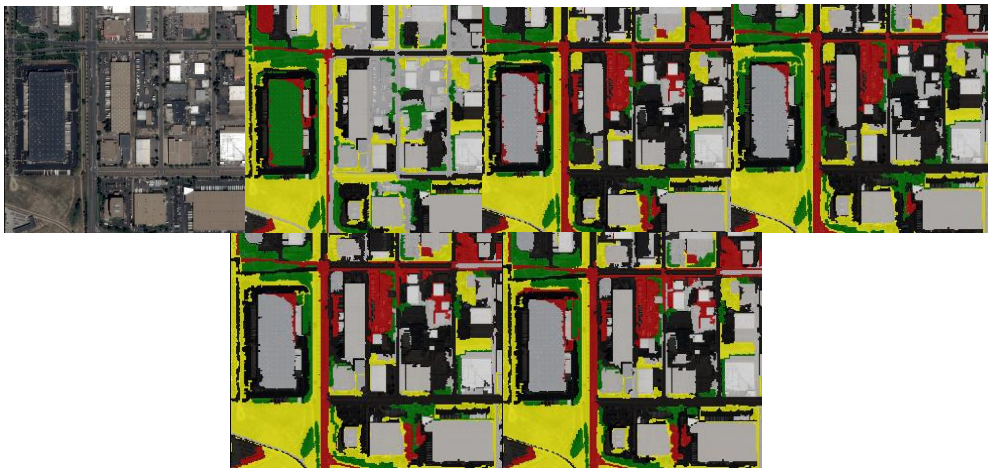


ΕΙΚΟΝΑ 3.160: ΑΠΟ ΤΗΝ ΑΡΧΗ: 2ο ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ ΓΙΑ ΤΙΜΕΣ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΠΛΗΘΟΣ ΔΕΝΤΡΩΝ 5, 20, 50, 100, 1000

Χώροι στάθμευσης

Η μείωση του πλήθους των δέντρων από 50 σε 5 οδήγησε σε μείωση του πλήθους των αντικειμένων που έχουν καταχωρηθεί στην κλάση των χώρων στάθμευσης. Τα αποτελέσματα στην περίπτωση αυτή είναι απογοητευτικά τόσο σε ό,τι αφορά το κριτήριο της ορθότητας όσο και εκείνο της πληρότητας. Αναλυτικά, στην κλάση αυτή ταξινομήθηκαν πολλά αντικείμενα τα οποία ανήκουν σε εκείνη των κτιρίων και αντιστρόφως.

Η αύξηση του πλήθους των δέντρων από 5 σε 20 αύξησε τόσο την πληρότητα όσο και την ορθότητα της κλάσης αυτής. Το ίδιο συνέβη και στην αύξηση από 20 σε 50. Τέλος, σημειώνεται πως η ρύθμιση της εν λόγω μεταβλητής σε 100 και 1000 δεν επηρέασε την ποιότητα του αποτελέσματος της συγκεκριμένης κλάσης (Εικόνα 3.161).



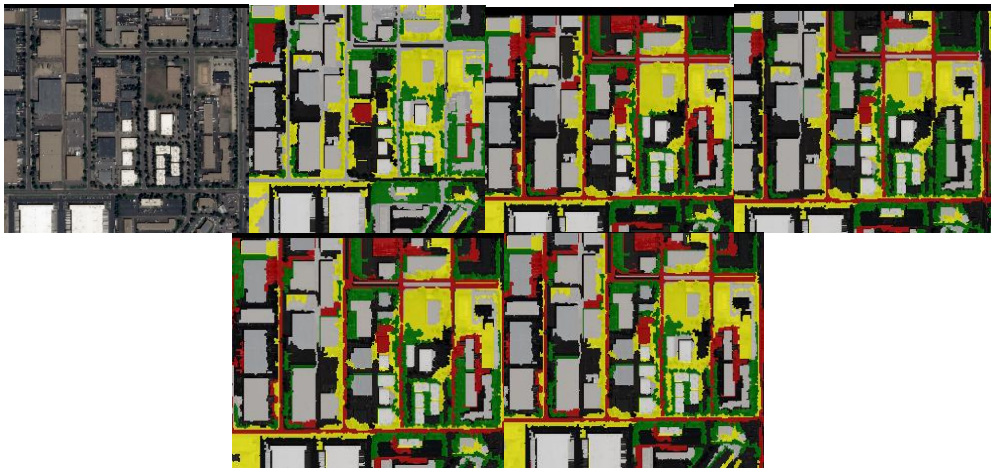
ΕΙΚΟΝΑ 3.161: ΑΠΟ ΤΗΝ ΑΡΧΗ: 3ο ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ ΓΙΑ ΤΙΜΕΣ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΠΛΗΘΟΣ ΔΕΝΤΡΩΝ 5, 20, 50, 100, 1000

Αστικό πράσινο

Η ρύθμιση της τιμής της παραμέτρου από 50 σε 5 αύξησε τον αριθμό των αντικειμένων της κλάσης αυτής. Το παραπάνω οδήγησε σε μείωση του ποσοστού της ορθότητας καθώς στη θεματική κατηγορία του αστικού πρασίνου προστέθηκαν πολλά κτίρια.

Η αύξηση της παραμέτρου από 5 σε 20 μείωσε το πλήθος των χώρων αστικού πρασίνου και παράλληλα αύξησε το ποσοστό ορθότητας της συγκεκριμένης θεματικής κατηγορίας.

Τέλος, σημειώνεται πως η αύξηση του πλήθους των δέντρων σε 50, 100 και 1000 επηρέασε ελάχιστα το παραγόμενο αποτέλεσμα για την κλάση του αστικού πρασίνου (Εικόνα 3.162).

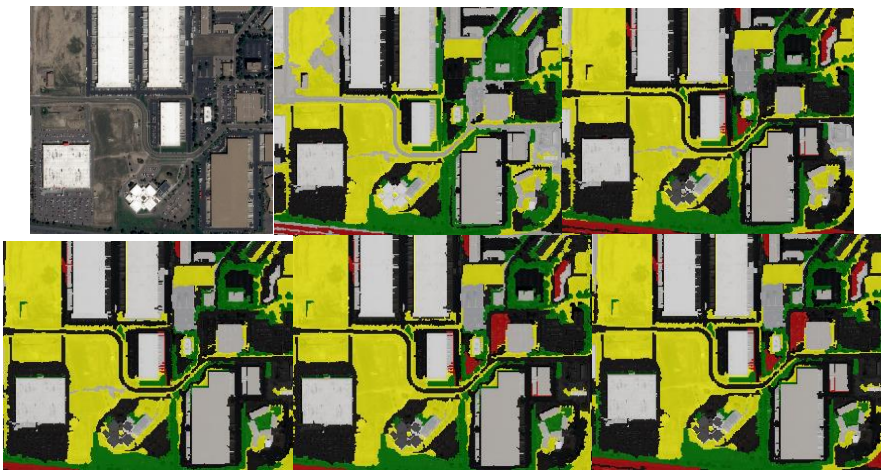


ΕΙΚΟΝΑ 3.162: ΑΠΟ ΤΗΝ ΑΡΧΗ: 4 ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ ΓΙΑ ΤΙΜΕΣ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΠΛΗΘΟΣ ΔΕΝΤΡΩΝ 5, 20, 50, 100, 1000

Άγονο Έδαφος

Η μείωση του πλήθους των δέντρων από 50 σε 5 μείωσε τον αριθμό των αντικειμένων του άγονου εδάφους. Το παραπάνω οδήγησε σε μείωση του ποσοστού πληρότητας της εν λόγω κλάσης.

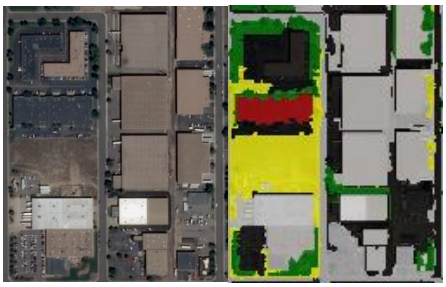
Τα αποτελέσματα για τις τιμές 20, 50, 100 και 1000 της συγκεκριμένης παραμέτρου είναι πανομοιότυπα για τη συγκεκριμένη θεματική κατηγορία (Εικόνα 3.163).



ΕΙΚΟΝΑ 3.163: ΑΠΟ ΤΗΝ ΑΡΧΗ: 5^{ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ ΓΙΑ ΤΙΜΕΣ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΠΛΗΘΟΣ ΔΕΝΤΡΩΝ 5, 20, 50, 100, 1000

Ποσοτική αξιολόγηση αποτελεσμάτων

ΠΛΗΘΟΣ ΔΕΝΤΡΩΝ: 5



ΕΙΚΟΝΑ 3.164: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΠΛΗΘΟΣ ΕΝΕΡΓΩΝ ΜΕΤΑΒΛΗΤΩΝ ΙΣΟ ΜΕ 2

Βάσει της Εικόνα 3.164 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.51, Πίνακας 3.52)

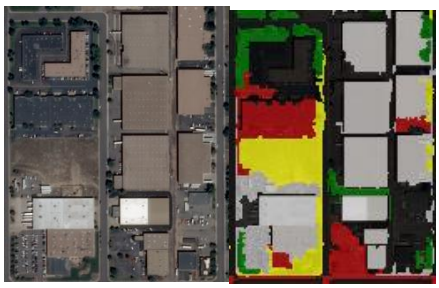
ΠΙΝΑΚΑΣ 3.51:ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (7Η ΔΟΚΙΜΗ).

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	12	4	4

ΠΙΝΑΚΑΣ 3.52: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (7Η ΔΟΚΙΜΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	75,00%	75,00%	60,00%	25,00%	25,00%

ΠΛΗΘΟΣ ΔΕΝΤΡΩΝ: 20



ΕΙΚΟΝΑ 3.165: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΠΛΗΘΟΣ ΕΝΕΡΓΩΝ ΜΕΤΑΒΛΗΤΩΝ ΙΣΟ ΜΕ 2

Βάσει της Εικόνα 3.165 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.53, Πίνακας 3.54)

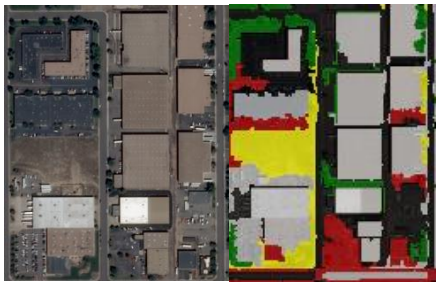
ΠΙΝΑΚΑΣ 3.53:ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (7Η ΔΟΚΙΜΗ).

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	12	2	4

ΠΙΝΑΚΑΣ 3.54: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (7Η ΔΟΚΙΜΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	75,00%	85,71%	66,67%	25,00%	12,50%

ΠΛΗΘΟΣ ΔΕΝΤΡΩΝ: 50, 100, 1000



ΕΙΚΟΝΑ 3.166: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΠΛΗΘΟΣ ΕΝΕΡΓΩΝ ΜΕΤΑΒΛΗΤΩΝ ΙΣΟ ΜΕ 2

Βάσει της Εικόνα 3.166 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.55, Πίνακας 3.56)

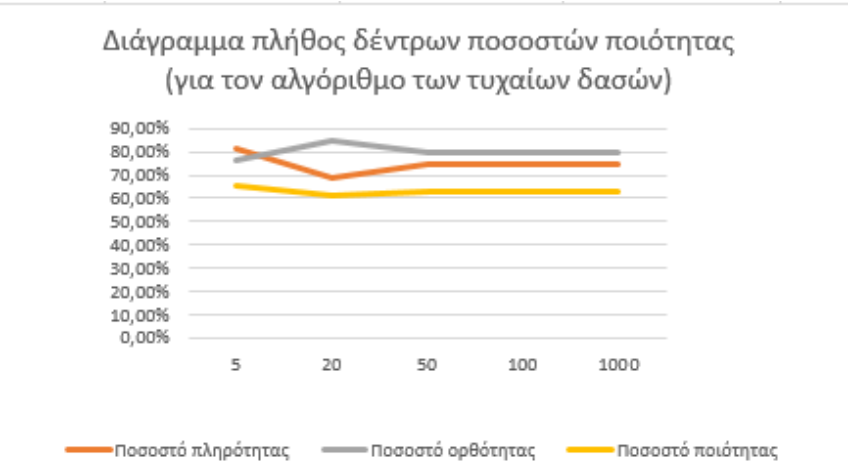
ΠΙΝΑΚΑΣ 3.55:ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (7Η ΔΟΚΙΜΗ).

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	14	3	2

ΠΙΝΑΚΑΣ 3.56: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (7Η ΔΟΚΙΜΗ).

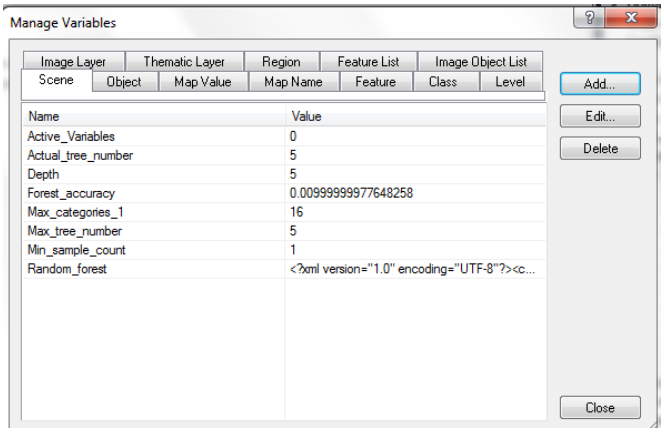
	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	87,50%	82,35%	73,68%	12,50%	18,75%

Στην ακόλουθη Εικόνα (Εικόνα 3.167) εμφανίζεται το διάγραμμα πλήθους δέντρων ποσοστών πληρότητας.

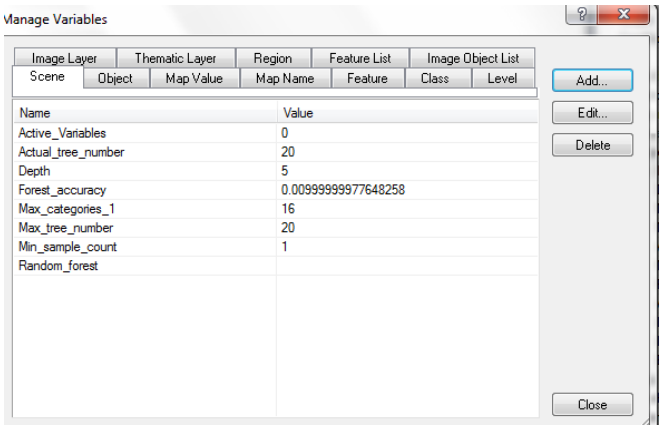


ΕΙΚΟΝΑ 3.167: ΔΙΑΓΡΑΜΜΑ ΠΛΗΘΟΣ ΔΕΝΤΡΩΝ ΠΟΣΟΣΤΩΝ ΠΟΙΟΤΗΤΑΣ

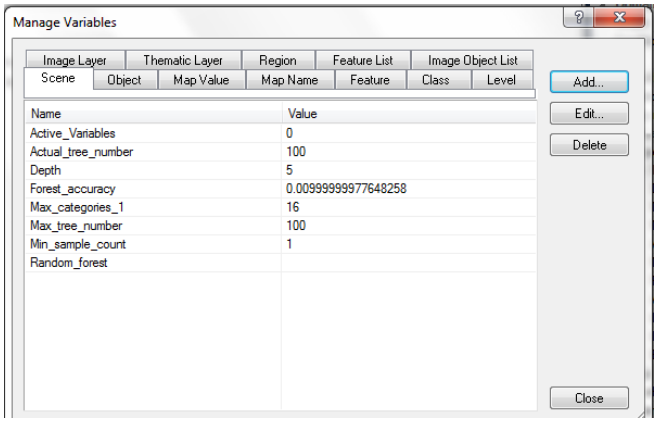
Στις ακόλουθες Εικόνες (Εικόνα 3.168- Εικόνα 3.171) εμφανίζονται οι πραγματικές τιμές των παραμέτρων έπειτα από τη δημιουργία των μοντέλων των τυχαίων δασών για τη δοκιμή 7.



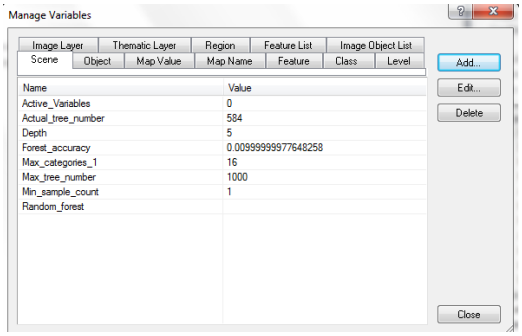
ΕΙΚΟΝΑ 3.168: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 6Η ΔΟΚΙΜΗ (ΜΕΓΙΣΤΟΣ ΑΡΙΘΜΟΣ ΔΕΝΤΡΩΝ: 5)



ΕΙΚΟΝΑ 3.169: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 6Η ΔΟΚΙΜΗ (ΜΕΓΙΣΤΟΣ ΑΡΙΘΜΟΣ ΔΕΝΤΡΩΝ: 20)



ΕΙΚΟΝΑ 3.170: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 6Η ΔΟΚΙΜΗ (ΜΕΓΙΣΤΟΣ ΑΡΙΘΜΟΣ ΔΕΝΤΡΩΝ: 100)



ΕΙΚΟΝΑ 3.171: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 6Η ΔΟΚΙΜΗ (ΜΕΓΙΣΤΟΣ ΑΡΙΘΜΟΣ ΔΕΝΤΡΩΝ: 1000)

3.7.8 Δοκιμή 8 (Ακρίβεια τυχαίου δάσους)

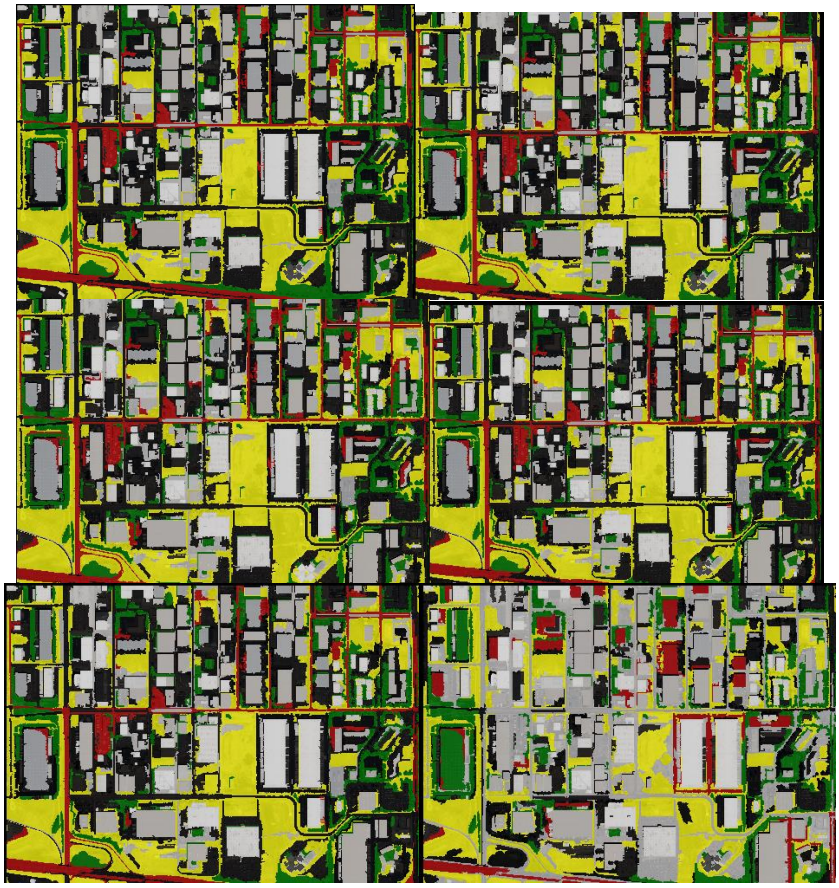
Μέσω της όγδοης δοκιμής εξετάστηκε η επιρροή της παραμέτρου Ακρίβεια του δάσους στην ποιότητα του παραγόμενου αποτελέσματος. Αναλυτικά, κατασκευάστηκαν έξι διαφορετικά μοντέλα βάσει των ακόλουθων τιμών των χαρακτηριστικών:

- Βάθος δέντρου (Depth): 0
- Ελάχιστος αριθμός δειγμάτων (Min sample count): 0
- Χρήση αντικαταστατών (Use surrogates): Όχι (No)
- Μέγιστος αριθμός κατηγοριών (Max categories): 16
- Ενεργές μεταβλητές (Activee Variables): 0 (δηλαδή ουσιαστικά για $\sqrt{9} = 3$)
- Μέγιστος αριθμός δέντρων (Max tree number): 50
- **Ακρίβεια δάσους (Forest accuracy): 0.01, 0,02, 0,05, 0,1, 0,5, 1**
- Τύπος κριτηρίου τερματισμού (termination criteria type): Και τα δύο (Both)

Σχολιασμός αποτελεσμάτων

Στην Εικόνα 3.172 εμφανίζονται τα αποτελέσματα για διαφορετικές τιμές της παραμέτρου ακρίβεια του δάσους. Τα συμπεράσματα τα οποία αντλούνται έπειτα από προσεκτική παρατήρηση των αποτελεσμάτων είναι πως:

- Οι θεματικοί χάρτες για τις ακρίβειες 0,01 και 0,02 είναι πανομοιότυποι
- Οι θεματικοί χάρτες για τις ακρίβειες 0,05, 0,1 και 0,5 είναι πανομοιότυποι



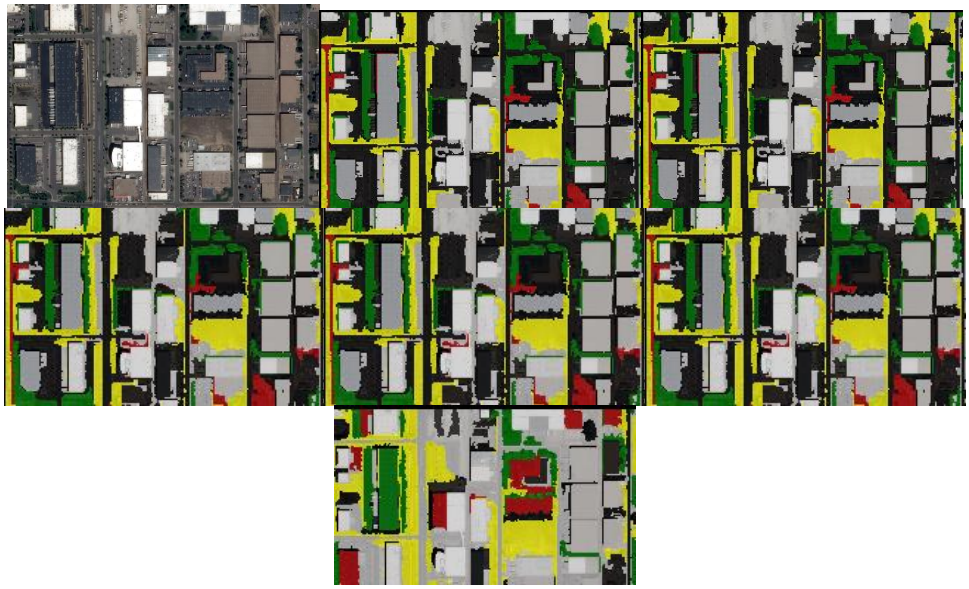
ΕΙΚΟΝΑ 3.172: ΑΠΟ ΠΑΝΩ ΑΡΙΣΤΕΡΑ: ΑΠΟΤΕΛΕΣΜΑ ΤΑΞΙΝΟΜΗΣΗΣ ΜΕ ΤΗ ΜΕΘΟΔΟ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΓΙΑ ΤΙΜΗ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΑΚΡΙΒΕΙΑ ΔΑΣΟΥΣ: 0,01, 0,02, 0,05, 0,1, 0,5, 1

ΚΤΙΡΙΑ

Η ρύθμιση της ακρίβειας του τυχαίου δάσους από 0,01 σε 0,02 δε μετέβαλε το αποτέλεσμα της ταξινόμησης.

Η αύξηση της εν λόγω παραμέτρου σε 0,05 μείωσε ελάχιστα το πλήθος των αντικειμένων της συγκεκριμένης θεματικής κατηγορίας και το παραπάνω επέφερε μικρή ελάττωση στο κριτήριο της πληρότητας. Οι θεματικοί χάρτες που παράχθηκαν για τις τιμές 0,1 και 0,5 ήταν πανομοιότυποι με εκείνον της 0,05

Τέλος, η ρύθμιση της ακρίβειας σε 1 αύξησε τον αριθμό των κτιρίων στο αποτέλεσμα της ταξινόμησης. Το κριτήριο της ορθότητας στο νέο θεματικό χάρτη για την κλάση των κτιρίων εμφάνισε χαμηλότερα ποσοστά καθώς τα αντικείμενα τα οποία προστέθηκαν στην τελευταία ανήκουν στην πραγματικότητα σε εκείνη των χώρων στάθμευσης και των δρόμων. Παράλληλα, από την κλάση αυτή απομακρύνθηκαν κάποια αντικείμενα τα οποία στην πραγματικότητα ανήκουν στη συγκεκριμένη. Ως εκ τούτου η αύξηση της συγκεκριμένης σε 1 μείωσε τους δείκτες ποιότητας για την κατηγορία των κτιρίων (Εικόνα 3.173)

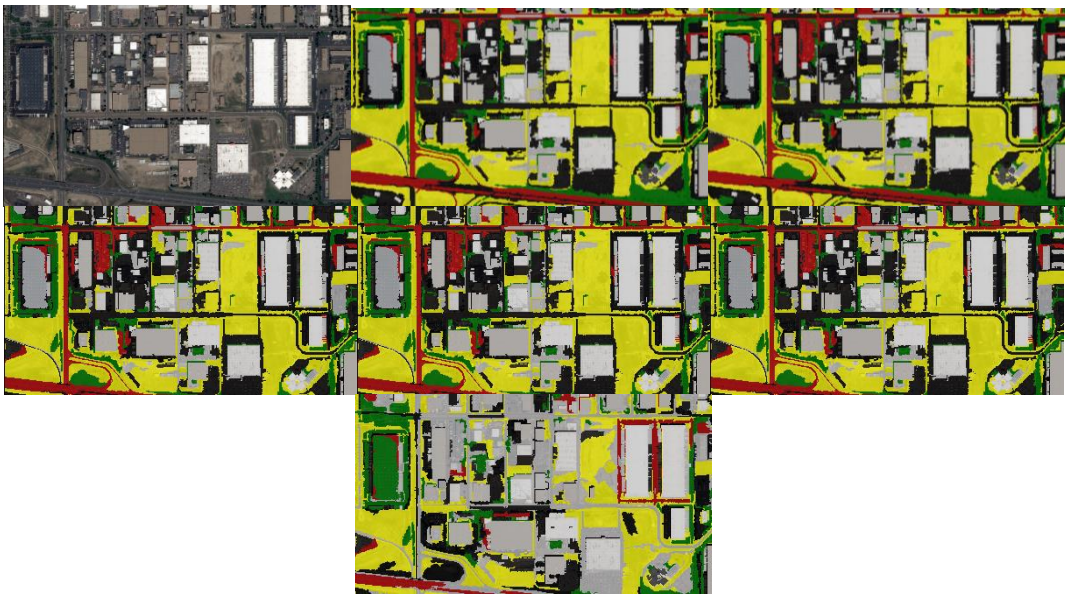


ΕΙΚΟΝΑ 3.173: ***ΑΠΟ ΤΗΝ ΑΡΧΗ: 1^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ ΓΙΑ ΤΙΜΕΣ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΑΚΡΙΒΕΙΑ ΔΑΣΟΥΣ:*** **0,01, 0,02, 0,05, 0,1, 0,5, 1**

ΔΡΟΜΟΙ

Η θεματική κατηγορία των δρόμων εμφανίζει πανομοιότυπα αποτελέσματα με εκείνα των προκαθορισμένων παραμέτρων για τις τιμές 0,02, 0.05, 0,1 και 0,5.

Η ρύθμιση της ακρίβειας του τυχαίου δάσους είχε αρνητικά αποτελέσματα τόσο σε ό,τι αφορά το κριτήριο της ορθότητας όσο και της πληρότητας. Το παραπάνω οφείλεται στο γεγονός πως στο νέο θεματικό χάρτη απομακρύνθηκαν από την κλάση των δρόμων αντικείμενα τα οποία ανήκουν στην πραγματικότητα στη συγκεκριμένη και παράλληλα προστέθηκαν σε αυτή ορισμένα από τα εμφανιζόμενα κτίρια (Εικόνα 3.174).



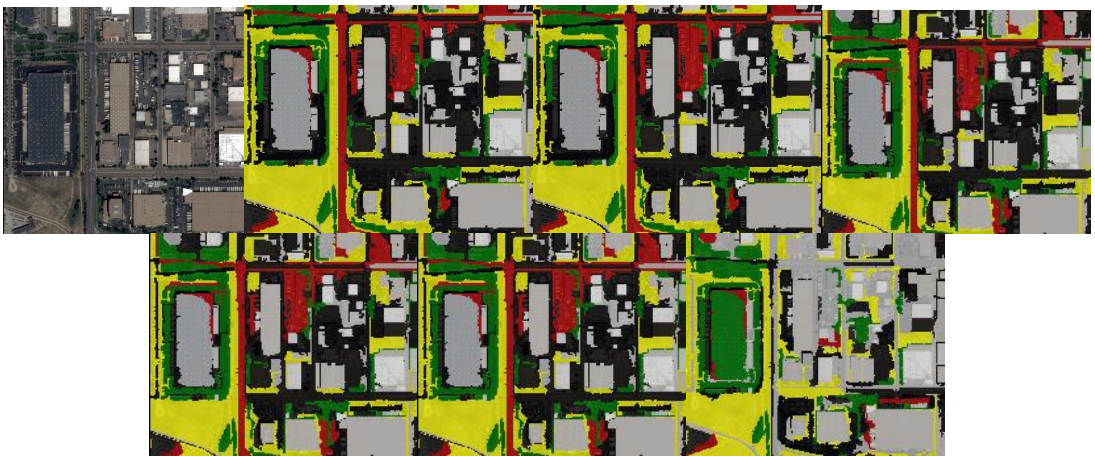
ΕΙΚΟΝΑ 3.174: ***ΑΠΟ ΤΗΝ ΑΡΧΗ: 2^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ ΓΙΑ ΤΙΜΕΣ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΑΚΡΙΒΕΙΑ ΔΑΣΟΥΣ:*** **0,01, 0,02, 0,05, 0,1, 0,5, 1**

ΧΩΡΟΙ ΣΤΑΘΜΕΥΣΗΣ

Το αποτέλεσμα για το θεματικό χάρτη της τιμής 0,02 είναι πανομοιότυπο με εκείνον της 0,01.

Η ρύθμιση της παραμέτρου της ακρίβειας σε 0,05, 0,1 και 0,5 αύξησε ελάχιστα το πλήθος των αντικειμένων της κατηγορίας των χώρων στάθμευσης. Το παραπάνω μείωσε ελάχιστα το ποσοστό της ορθότητας στους νέους θεματικού χάρτες.

Τέλος, η ρύθμιση της τιμής σε 1 μείωσε το πλήθος των αντικειμένων της κλάσης των χώρων στάθμευσης και το παραπάνω επέδρασε για τα κριτήρια της ποιότητας του αποτελέσματος (Εικόνα 3.175).

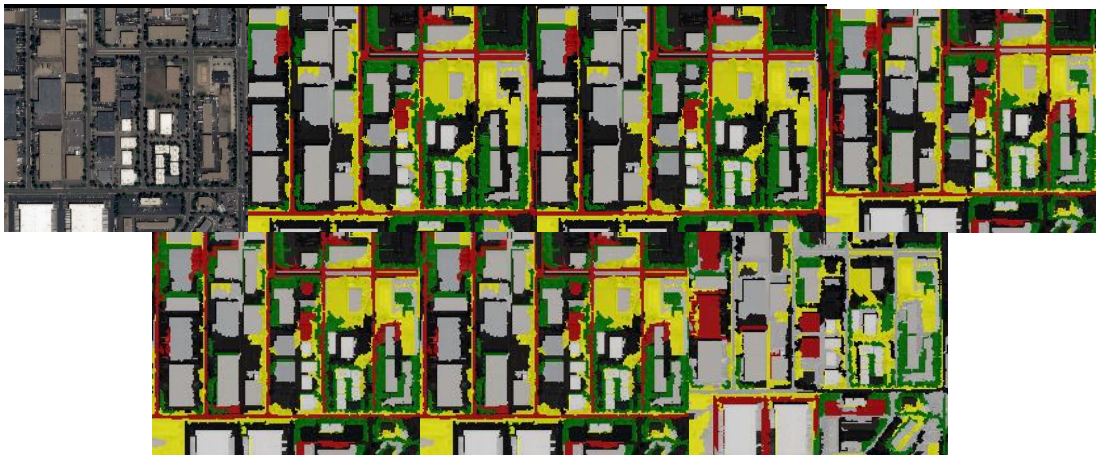


ΕΙΚΟΝΑ 3.175: ΑΠΟ ΤΗΝ ΑΡΧΗ: *3^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ ΓΙΑ ΤΙΜΕΣ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΑΚΡΙΒΕΙΑ ΔΑΣΟΥΣ:* 0,01, 0,02, 0,05, 0,1, 0,5, 1

ΑΣΤΙΚΟ ΠΡΑΣΙΝΟ

Οι θεματικοί χάρτες που κατασκευάστηκαν για τις τιμές 0,01, 0,02, 0,05, 0,1 και 0,5 της παραμέτρου της ακρίβειας των τυχαίων δασών εμφάνισαν πανομοιότυπα αποτελέσματα σε ό,τι αφορά την κλάση του αστικού πρασίνου.

Η αύξηση της εν λόγω παραμέτρου σε 1 επέδρασε για τα κριτήρια της ποιότητας του αποτελέσματος καθώς από την κλάση του αστικού πρασίνου απομακρύνθηκαν αντικείμενα τα οποία ανήκουν στη συγκεκριμένη και παράλληλα προστέθηκαν σε αυτήν αντικείμενα τα οποία δεν ανήκουν στην πραγματικότητα σε εκείνη (Εικόνα 3.176).

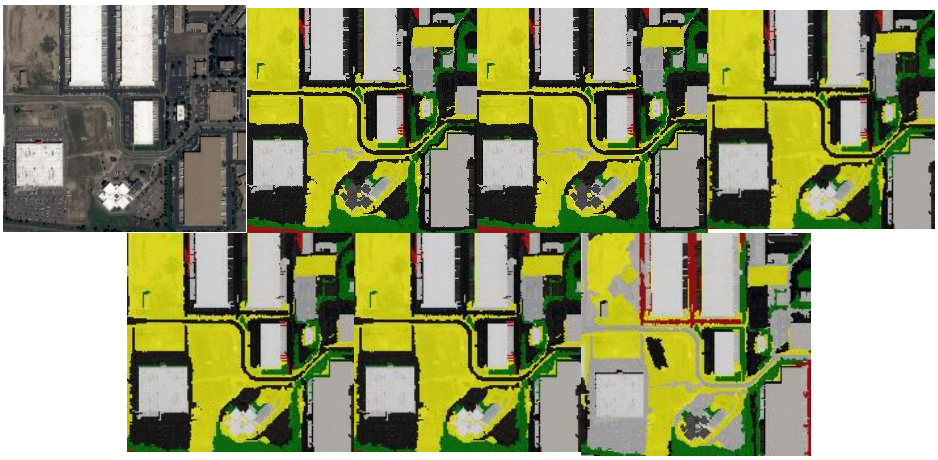


ΕΙΚΟΝΑ 3.176: ΑΠΟ ΤΗΝ ΑΡΧΗ: *4^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ ΓΙΑ ΤΙΜΕΣ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΑΚΡΙΒΕΙΑ ΔΑΣΟΥΣ:* 0,01, 0,02, 0,05, 0,1, 0,5, 1

ΆΓΟΝΟ ΈΔΑΦΟΣ

Ομοίως με την προαναφερθείσα κλάση οι θεματικοί χάρτες που κατασκευάστηκαν για τις τιμές 0,01, 0,02, 0,05, 0,1 και 0,5 της παραμέτρου της ακρίβειας των τυχαίων δασών εμφάνισαν πανομοιότυπα αποτελέσματα σε ό,τι αφορά την κατηγορία του άγονου εδάφους.

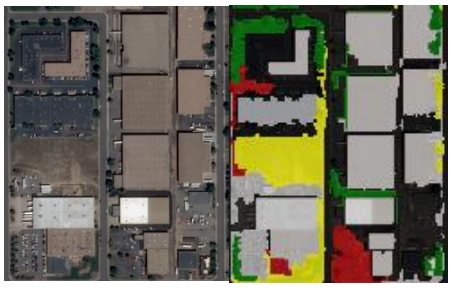
Η αύξηση της ακρίβειας 1 μείωσε την πληρότητα της κατηγορίας του άγονου εδάφους καθώς απομακρύνθηκαν από αυτήν αντικείμενα τα οποία ανήκουν στη συγκεκριμένη (Εικόνα 3.177).



ΕΙΚΟΝΑ 3.177: ΑΠΟ ΤΗΝ ΑΡΧΗ: *5^{Ο} ΑΠΟΣΠΑΣΜΑ ΑΡΧΙΚΗΣ ΕΙΚΟΝΑΣ ΓΙΑ ΤΙΜΕΣ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΑΚΡΙΒΕΙΑ ΔΑΣΟΥΣ:* 0,01, 0,02, 0,05, 0,1, 0,5, 1

ΠΟΣΟΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

ΑΚΡΙΒΕΙΑ ΤΥΧΑΙΟΥ ΔΑΣΟΥΣ: 0,1, 0,2



ΕΙΚΟΝΑ 3.178*: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΗΝ ΠΡΩΤΗ ΔΟΚΙΜΗ*

Βάσει της Εικόνα 3.178 κατασκευάστηκαν οι ακόλουθοι Πίνακες (Πίνακας 3.57, Πίνακας 3.58):

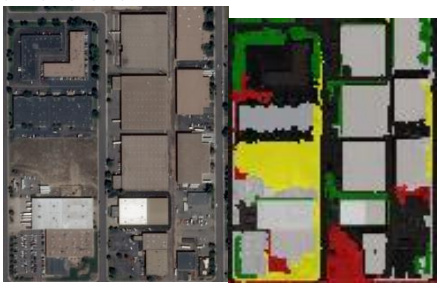
ΠΙΝΑΚΑΣ 3.57*: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (8Η ΔΟΚΙΜΗ).*

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	14	3	2

ΠΙΝΑΚΑΣ 3.58: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (8Η ΔΟΚΙΜΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	87,50%	82,35%	73,68%	12,50%	18,75%

ΑΚΡΙΒΕΙΑ ΤΥΧΑΙΟΥ ΔΑΣΟΥΣ: 0,05, 0,1, 0,5



ΕΙΚΟΝΑ 3.179*: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΗΝ ΠΡΩΤΗ ΔΟΚΙΜΗ*

Βάσει της Εικόνα 3.179 κατασκευάστηκαν οι ακόλουθοι Πίνακες (Πίνακας 3.59, Πίνακας 3.60):

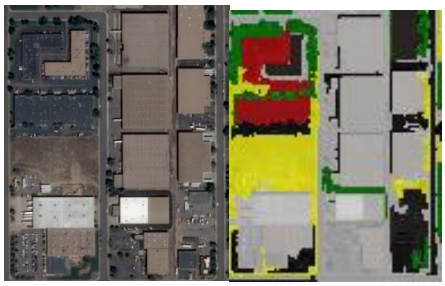
ΠΙΝΑΚΑΣ 3.59*: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (8Η ΔΟΚΙΜΗ).*

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	13	3	3

ΠΙΝΑΚΑΣ 3.60: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (8Η ΔΟΚΙΜΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	81,25%	81,25%	68,42%	18,75%	18,75%

ΑΚΡΙΒΕΙΑ ΤΥΧΑΙΟΥ ΔΑΣΟΥΣ: 1



ΕΙΚΟΝΑ 3.180*: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΗΝ ΠΡΩΤΗ ΔΟΚΙΜΗ*

Βάσει της Εικόνα 3.180 κατασκευάστηκαν οι ακόλουθοι Πίνακες (Πίνακας 3.61, Πίνακας 3.62):

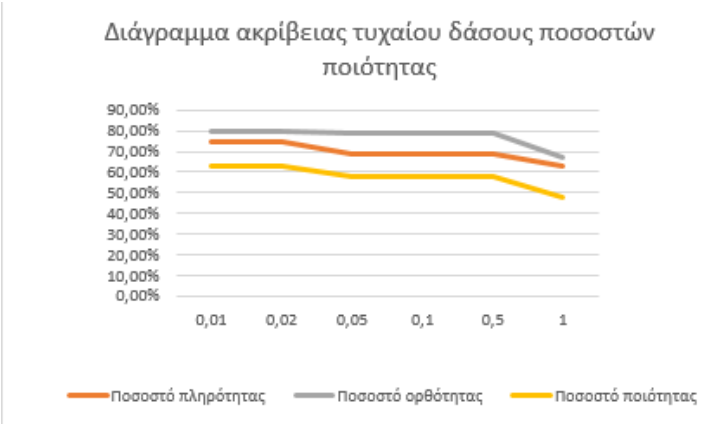
ΠΙΝΑΚΑΣ 3.61*: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (8Η ΔΟΚΙΜΗ).*

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	11	5	5

ΠΙΝΑΚΑΣ 3.62: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (8Η ΔΟΚΙΜΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	68,75%	68,75%	52,38%	31,25%	31,25%

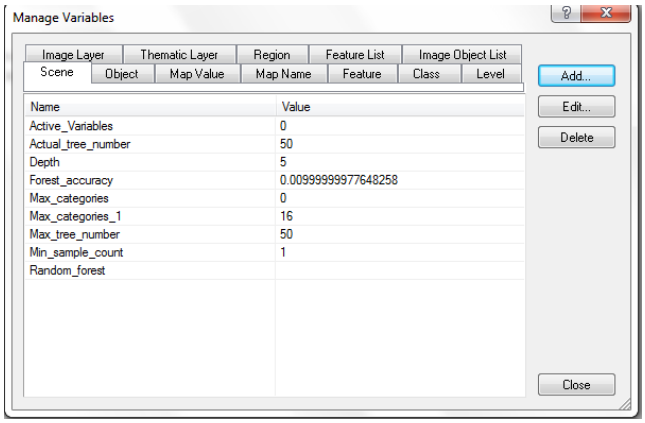
Στην Εικόνα 3.181 εμφανίζεται το διάγραμμα ακρίβειας τυχαίου δάσους ποσοστών ποιότητας. Είναι εμφανές πως η αύξηση της τιμής της συγκεκριμένης παραμέτρου οδήγησε σε μείωση των ποσοστών ποιότητας του παραγόμενου αποτελέσματος.



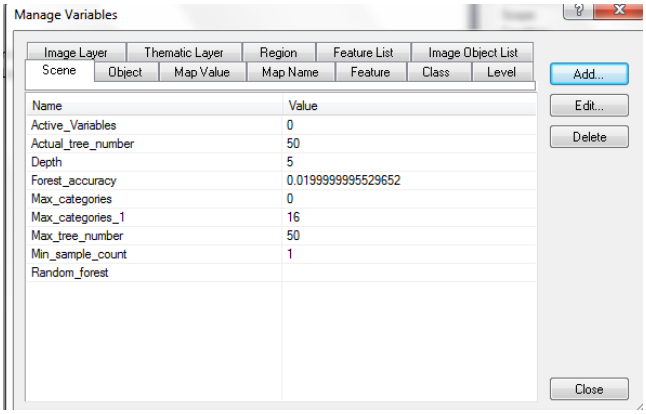
ΕΙΚΟΝΑ 3.181: ΔΙΑΓΡΑΜΜΑ ΑΚΡΙΒΕΙΑΣ ΤΥΧΑΙΟΥ ΔΑΣΟΥΣ ΠΟΣΟΣΤΩΝ ΠΟΙΟΤΗΤΑΣ

Στις ακόλουθες Εικόνες (Εικόνα 3.182 - Εικόνα 3.187) εμφανίζονται οι πραγματικές τιμές των παραμέτρων έπειτα από τη δημιουργία των μοντέλων των τυχαίων δασών για τη δοκιμή 8. Στις συγκεκριμένες παρατηρείται πως η αύξηση της ακρίβειας του τυχαίου δάσους οδηγεί σε μείωση του πλήθους των δέντρων. Μάλιστα στην περίπτωση των τιμών 0,5 και 1 κατασκευάστηκε μόλις ένα δέντρο. Το παραπάνω οφείλεται στο γεγονός πως η μείωση στις απαιτήσεις του αλγορίθμου οδήγησε σε πρόωρη διακοπή του σχηματισμού του τυχαίου δάσους μέσω του κριτηρίου τερματισμού. Αυτό οφείλεται στο γεγονός πως το

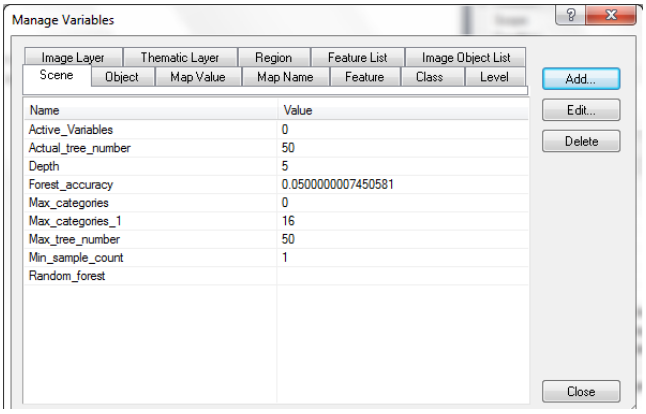
μοντέλο έφτασε στην απαιτούμενη ακρίβεια προτού συμπληρωθεί το απαιτούμενο πλήθος δέντρων



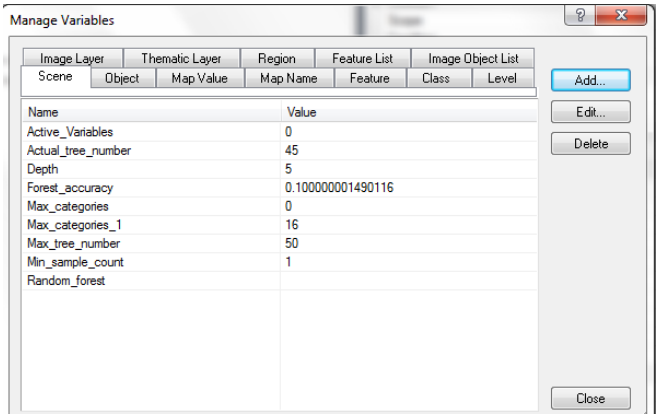
ΕΙΚΟΝΑ 3.182: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 8Η ΔΟΚΙΜΗ (ΑΚΡΙΒΕΙΑ ΔΑΣΟΥΣ: 0,01)



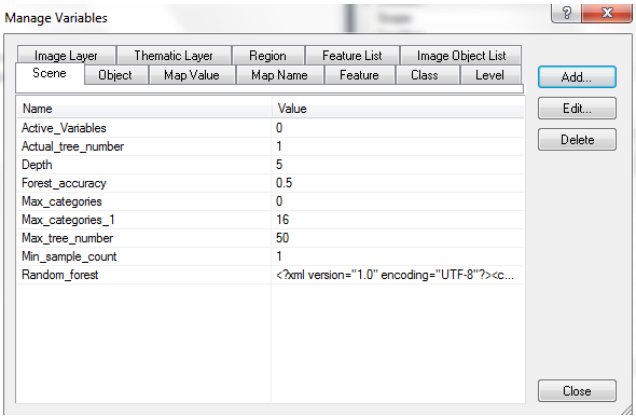
ΕΙΚΟΝΑ 3.183: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 8Η ΔΟΚΙΜΗ (ΑΚΡΙΒΕΙΑ ΔΑΣΟΥΣ: 0,02)



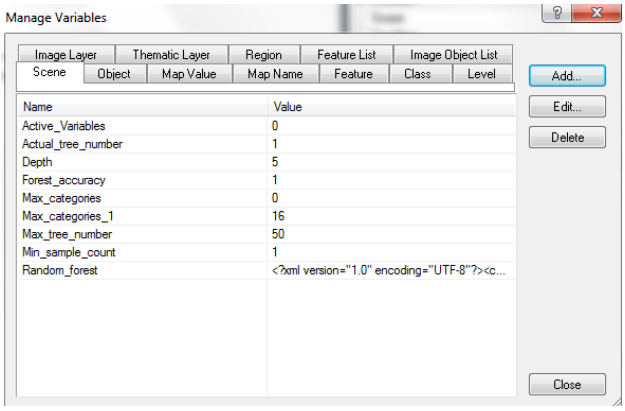
ΕΙΚΟΝΑ 3.184: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 8Η ΔΟΚΙΜΗ (ΑΚΡΙΒΕΙΑ ΔΑΣΟΥΣ: 0,05)



ΕΙΚΟΝΑ 3.185: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 8Η ΔΟΚΙΜΗ (ΑΚΡΙΒΕΙΑ ΔΑΣΟΥΣ: 0,1)



ΕΙΚΟΝΑ 3.186: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 8Η ΔΟΚΙΜΗ (ΑΚΡΙΒΕΙΑ ΔΑΣΟΥΣ: 0,5)



ΕΙΚΟΝΑ 3.187: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 8Η ΔΟΚΙΜΗ (ΑΚΡΙΒΕΙΑ ΔΑΣΟΥΣ: 1)

3.7.9 Δοκιμή 9 (Κριτήριο τερματισμού)

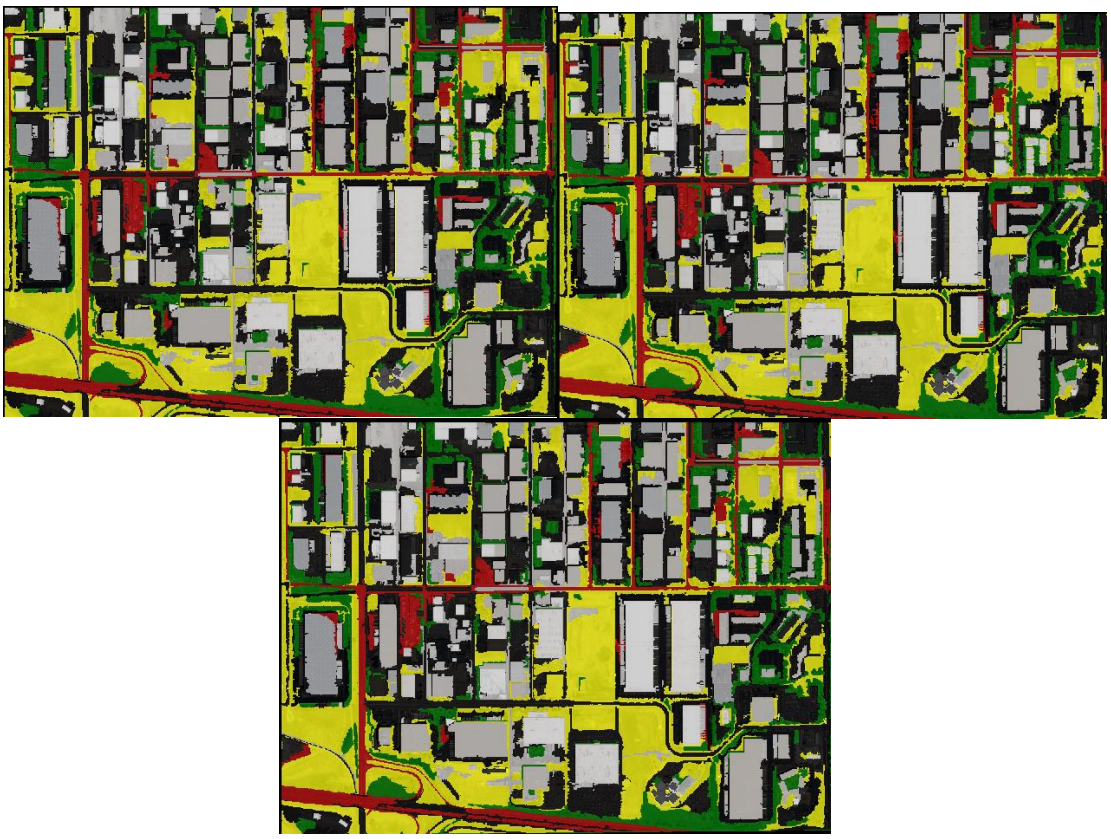
Στα πλαίσια της συγκεκριμένης δοκιμής έγινε πειραματισμός ως προς τις τιμές της παραμέτρου «κριτήριο τερματισμού». Αναλυτικά, οι τιμές που δόθηκαν είναι οι ακόλουθες:

- Βάθος δέντρου (Depth): 0
- Ελάχιστος αριθμός δειγμάτων (Min sample count): 0
- Χρήση αντικαταστατών (Use surrogates): Όχι (No)

- Μέγιστος αριθμός κατηγοριών (Max categories): 16
- Ενεργές μεταβλητές (Activee Variables): 0 (δηλαδή ουσιαστικά για $\sqrt{9} = 3$)
- Μέγιστος αριθμός δέντρων (Max tree number): 50
- Ακρίβεια δάσους (Forest accuracy): 0.01
- **Τύπος κριτηρίου τερματισμού (termination criteria type): Και τα δύο (Both), Τερματισμός εκπαίδευσης στην περίπτωση που συμπληρωθεί ο μέγιστος αριθμός δέντρων (Terminate learning by max tree number), Τερματισμός εκπαίδευσης όταν ικανοποιηθεί το κριτήριο της ακρίβειας του τυχαίου δάσους (Terminate learning by forest accuracy)**

Σχολιασμός αποτελεσμάτων

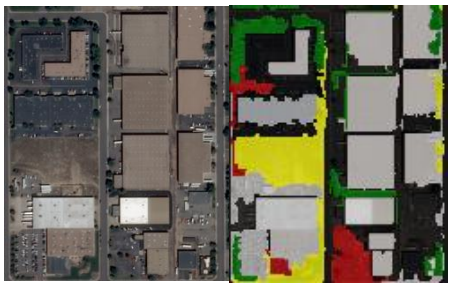
Στην Εικόνα 3.188 εμφανίζεται το αποτέλεσμα της ταξινόμησης για τα τρία διαφορετικά κριτήρια τερματισμού. Είναι εμφανές πως η ρύθμιση της συγκεκριμένης μεταβλητής δεν επηρέασε το αποτέλεσμα της ταξινόμησης του αλγορίθμου των τυχαίων δασών.



Εικόνα 3.188: Αποτέλεσμα εφαρμογής του αλγορίθμου των τυχαίων δασών για τιμές της παραμέτρου τύπος τερματισμού και τα δύο (πάνω αριστερά), αριθμός δέντρων (πάνω δεξιά), ακρίβεια (κάτω)

Ποσοτική αξιολόγηση

ΤΥΠΟΣ ΚΡΙΤΗΡΙΟΥ: ΑΡΙΘΜΟΣ ΔΕΝΤΡΩΝ, ΑΚΡΙΒΕΙΑ ΤΥΧΑΙΟΥ ΔΑΣΟΥΣ



ΕΙΚΟΝΑ 3.189: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΑ ΚΡΙΤΗΡΙΑ ΤΕΡΜΑΤΙΣΜΟΥ ΑΡΙΘΜΟΣ ΔΕΝΤΡΩΝ ΚΑΙ ΑΚΡΙΒΕΙΑ ΤΥΧΑΙΟΥ ΔΑΣΟΥΣ

Βάσει της Εικόνα 3.189 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.63, Πίνακας 3.64)

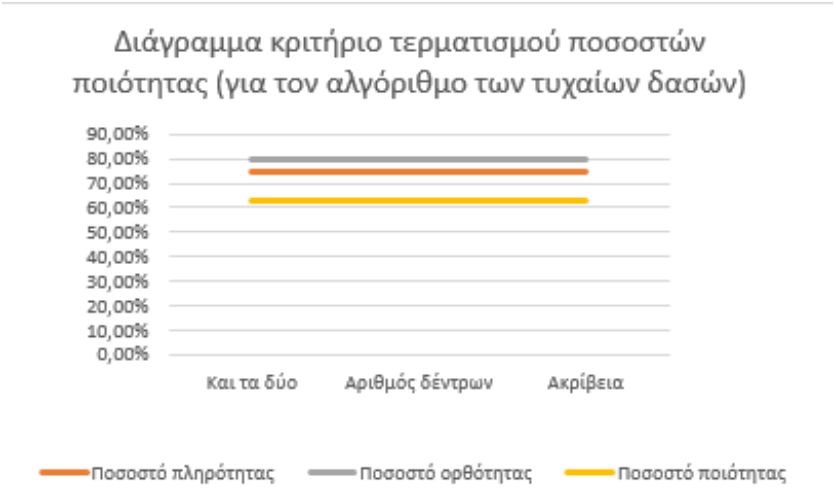
ΠΙΝΑΚΑΣ 3.63: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (9Η ΔΟΚΙΜΗ).

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	14	3	2

ΠΙΝΑΚΑΣ 3.64: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (9Η ΔΟΚΙΜΗ).

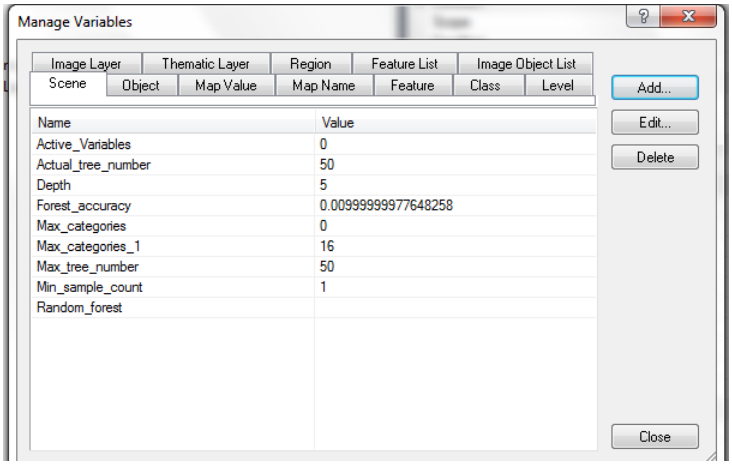
	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	87,50%	82,35%	73,68%	12,50%	18,75%

Στην Εικόνα 3.190 εμφανίζεται το διάγραμμα κριτηρίου τερματισμού – ποσοστών ποιότητας.

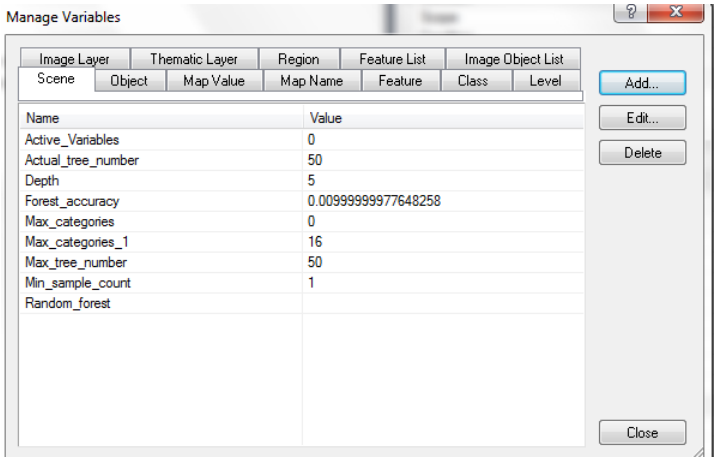


ΕΙΚΟΝΑ 3.190: ΔΙΑΓΡΑΜΜΑ ΚΡΙΤΗΡΙΟΥ ΤΕΡΜΑΤΙΣΜΟΥ ΠΟΣΟΣΤΩΝ ΠΟΙΟΤΗΤΑΣ

Στις ακόλουθες Εικόνες () εμφανίζονται οι πραγματικές τιμές των παραμέτρων έπειτα από τη δημιουργία των μοντέλων των τυχαίων δασών για τη δοκιμή 9



ΕΙΚΟΝΑ 3.191: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 9Η ΔΟΚΙΜΗ (ΤΕΡΜΑΤΙΣΜΟΣ: ΠΛΗΘΟΣ ΔΕΝΤΡΩΝ)



ΕΙΚΟΝΑ 3.192: ΤΙΜΕΣ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΑ ΔΑΣΗ ΓΙΑ ΤΗΝ 9Η ΔΟΚΙΜΗ (ΤΕΡΜΑΤΙΣΜΟΣ: ΑΚΡΙΒΕΙΑ ΔΑΣΟΥΣ)

3.7.10 Τελική επιλογή παραμέτρων του αλγορίθμου των τυχαίων δασών

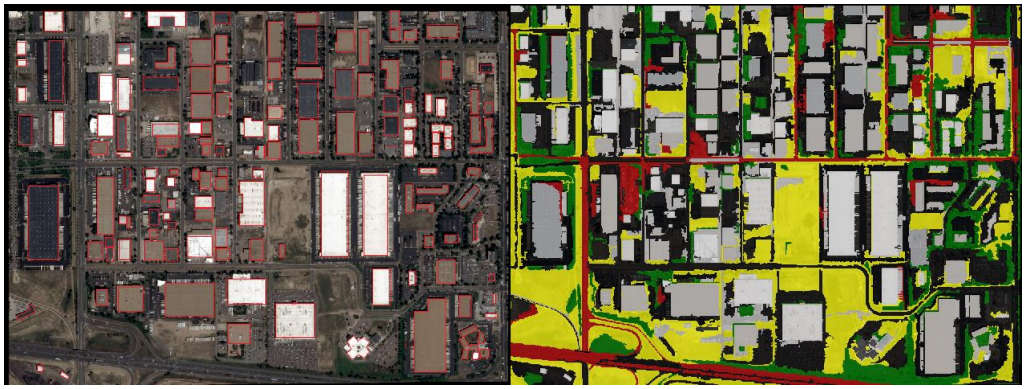
Βάσει των παραπάνω επαναλήψεων προκύπτει πως οι τιμές εκείνες των παραμέτρων οι οποίες δίνουν τα υψηλότερα δυνατά ποσοστά ποιότητας είναι οι ακόλουθες:

- Βάθος δέντρων: 10
- Ελάχιστος αριθμός δειγμάτων ανά κόμβο: 0
- Χρήση αντικαταστατών: Όχι
- Μέγιστος αριθμός κατηγοριών: 16
- Πλήθος ενεργών μεταβλητών: 3
- Πλήθος δέντρων: 50
- Ακρίβεια τυχαίου δάσους:0,01
- Κριτήριο τερματισμού: Και τα δύο

Στην Εικόνα 3.193 εμφανίζεται το αποτέλεσμα της ταξινόμησης μέσω του αλγορίθμου των τυχαίων δασών βάσει των παραπάνω παραμέτρων.

ΕΙΚΟΝΑ 3.193: ΤΕΛΙΚΟ ΑΠΟΤΕΛΕΣΜΑ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΒΑΣΕΙ ΤΩΝ ΔΟΚΙΜΩΝ ΑΝΑΦΟΡΙΚΑ ΜΕ ΤΙΣ ΤΙΜΕΣ ΤΩΝ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ

Η αξιολόγηση του τελικού αποτελέσματος έγινε έπειτα από αξιολόγηση του αποτελέσματος της ταξινόμησης στο σύνολο της εικόνας. Για το σκοπό αυτό έγινε ψηφιοποίηση όλων των εμφανιζόμενων κτιρίων στο υπό μελέτη τμήμα της εικόνας (Εικόνα 3.194). Οι δείκτες ποιότητας όπως προέκυψαν έπειτα από φωτοερμηνεία εμφανίζονται στους Πίνακας 3.109 και Πίνακας 3.110.



ΕΙΚΟΝΑ 3.194: ΑΞΙΟΛΟΓΗΣΗ ΤΩΝ ΕΠΙΔΟΣΕΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΕ Ο,ΤΙ ΑΦΟΡΑ ΤΗΝ ΑΝΙΧΝΕΥΣΗ ΚΤΙΡΙΩΝ

ΠΙΝΑΚΑΣ 3.65*: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (ΤΕΛΙΚΗ ΕΠΙΛΟΓΗ ΤΩΝ ΤΙΜΩΝ ΤΩΝ ΠΑΡΑΜΕΤΡΩΝ)*

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Σύνολο εικόνας	129	52	28

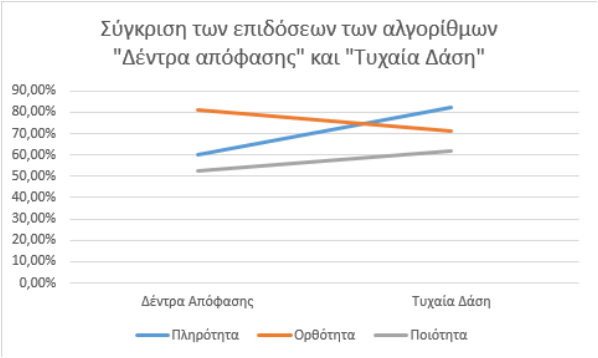
ΠΙΝΑΚΑΣ 3.66: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (*ΤΕΛΙΚΗ ΕΠΙΛΟΓΗ ΤΩΝ ΤΙΜΩΝ ΤΩΝ ΠΑΡΑΜΕΤΡΩΝ*).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Σύνολο εικόνας	82,17%	71,27%	61,72%	17,83%	33,12%

3.8 Σύγκριση των επιδόσεων των αλγορίθμων «Δέντρα απόφασης» και «Τυχαία Δάση»

Στην Εικόνα 3.195 εμφανίζεται το διάγραμμα επιδόσεων των αλγορίθμων δέντρα απόφασης και τυχαία δάση. Συγκεκριμένα, γίνεται σύγκριση των δεικτών ποιότητας των μοντέλων που κατασκευάστηκαν για τις τελικά επιλεγμένες τιμές των παραμέτρων και η

εκπαίδευση των οποίων έγινε βάσει των ίδιων δειγμάτων και γνωρισμάτων. Παρατηρείται πως οι επιδόσεις του συνδυαστικού ταξινομητή (τυχαία δάση) είναι εμφανώς βελτιωμένες συγκριτικά με εκείνες του μεμονωμένου (δέντρα απόφασης) καθώς οι δείκτες ποιότητας και πληρότητας στην περίπτωση των τυχαίων δασών έχουν αυξηθεί. Συνεπώς η ιδέα πως η ύπαρξη πολλών ταξινομητών αντισταθμίζει κατά κάποιο τρόπο το σφάλμα στο αποτέλεσμα της ταξινόμησης του ενός επιβεβαιώνεται.



ΕΙΚΟΝΑ 3.195: ΔΙΑΓΡΑΜΜΑ ΕΠΙΔΟΣΕΩΝ ΤΩΝ ΑΛΓΟΡΙΘΜΩΝ «ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ» ΚΑΙ «ΤΥΧΑΙΑ ΔΑΣΗ» ΣΕ Ο,ΤΙ ΑΦΟΡΑ ΤΗΝ ΑΝΙΧΝΕΥΣΗ ΚΤΙΡΙΩΝ

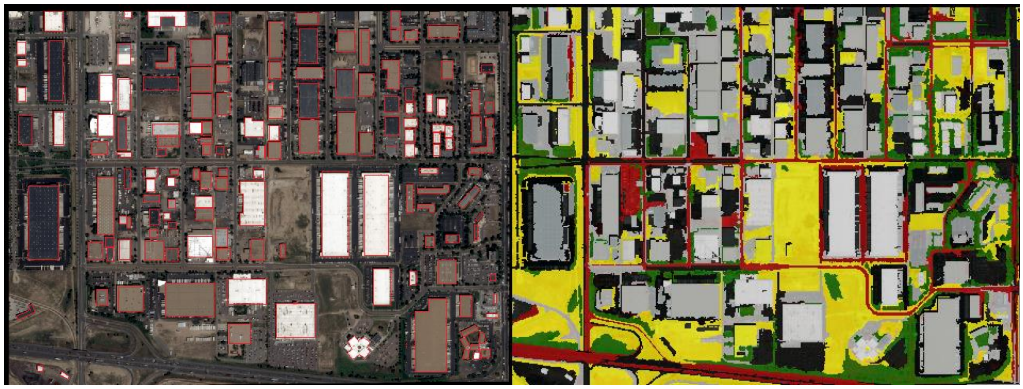
3.9 Σύγκριση των επιδόσεων των αλγορίθμων ταξινόμησης «Δέντρα απόφασης» και «Τυχαία Δάση» με εκείνες του «Εγγύτερου Γείτονα»

Στα πλαίσια της παρούσας ενότητας έγινε εφαρμογή του πλέον διαδεδομένου αλγορίθμου επιβλεπόμενης ταξινόμησης, εκείνου του εγγύτερου γειτονα, στα δεδομένα της παρούσας μελέτης. Στόχος της παρούσας διαδικασίας είναι η σύγκριση των επιδόσεων του τελευταίου με τους υπό μελέτη αλγορίθμους «Δέντρα Απόφασης» και «Τυχαία Δάση» σε ό,τι αφορά την ανίχνευση κτιρίων.

Τα χαρακτηριστικά στα οποία βασίστηκε η ταξινόμηση είναι τα αντίστοιχα με εκείνα των «Δέντρων Απόφασης» και των «Τυχαίων Δασών»:

- Μέση τιμή φωτεινότητας για το κανάλι 3 (Mean Layer 3)
- Μέση τιμή φωτεινότητας για το κανάλι 4 (Mean Layer 4)
- Μέση τιμή φωτεινότητας για το κανάλι 1 (Mean Layer 1)
- Μέση τιμή φωτεινότητας για το κανάλι 2 (Mean Layer 2)
- Δείκτης βλάστησης (NDVI)
- Λόγος μήκος προς πλάτος του εκάστοτε αντικειμένου (Length/Width)
- Συμπαγότητα του αντικειμένου (Compactness)
- Ομοιότητα του σχήματος του αντικειμένου με το σχήμα του ορθογωνίου(Rectangular fit)
- Εμβαδόν αντικειμένου (Area)
- Μέγιστη διαφορά (Max. Diff.)
- Μέση φωτεινότητα του αντικειμένου (Brightness)

Επιπροσθέτως, έγινε χρήση των ίδιων δεδομένων εκπαίδευσης με εκείνα των προηγούμενων δοκιμών.
Στην Εικόνα 3.196 εμφανίζεται το αποτέλεσμα της εφαρμογής του Εγγύτερου Γείτονα στην εικόνα της πόλης Commerce.



ΕΙΚΟΝΑ 3.196: ΑΞΙΟΛΟΓΗΣΗ ΤΩΝ ΕΠΙΔΟΣΕΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ «ΕΓΓΥΤΕΡΟΣ ΓΕΙΤΟΝΑΣ» ΣΕ Ο,ΤΙ ΑΦΟΡΑ ΤΗΝ ΑΝΙΧΝΕΥΣΗ ΤΩΝ ΚΤΙΡΙΩΝ

Στους Πίνακας 3.67, Πίνακας 3.68 εμφανίζονται οι δείκτες ποιότητας για το αποτέλεσμα της ταξινόμησης του Εγγύτερου Γείτονα

ΠΙΝΑΚΑΣ 3.67*: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (ΕΓΓΥΤΕΡΟΣ ΓΕΙΤΟΝΑΣ).*

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Σύνολο εικόνας	135	70	22

ΠΙΝΑΚΑΣ 3.68*: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (ΕΓΓΥΤΕΡΟΣ ΓΕΙΤΟΝΑΣ)*

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Σύνολο εικόνας	85,99%	65,85%	59,47%	14,01%	44,59%

Στον Πίνακας 3.69 εμφανίζονται οι δείκτες ποιότητας των αλγορίθμων ταξινόμησης «Δέντρα Απόφασης», «Τυχαία Δάση» και «Εγγύτερος Γείτονας» σε ό,τι αφορά την ανίχνευση κτιρίων. Είναι εμφανές πως ο αλγόριθμος των Τυχαίων Δασών εμφανίζει το υψηλότερο ποσοστό ποιότητας το οποίο φτάνει το 62%. Ακολουθεί εκείνος του Εγγύτερου Γείτονα με ποσοστό 59%, ενώ τις χαμηλότερες επιδόσεις εμφανίζουν τα δέντρα απόφασης - μόλις 50%.

Σε ό,τι αφορά την ικανοποίηση του κριτηρίου της πληρότητας, ο αλγόριθμος του Εγγύτερου Γείτονα εμφανίζει το υψηλότερο ποσοστό (86%). Ο συγκεκριμένος, ωστόσο, υστερεί σε ό,τι αφορά την ορθότητα του παραγόμενου αποτελέσματος καθώς στη θεματική κατηγορία των κτιρίων έχουν συμπεριληφθεί πολλά αντικείμενα της κατηγορίας «άγονο έδαφος» και «δρόμοι». Το ποσοστό του εν λόγω κριτηρίου φτάνει μόλις το 66% σε αντίθεση με το αντίστοιχο των Τυχαίων Δασών που συγκεντρώνει το ποσοστό του 71%. Το αποτέλεσμα της πληρότητας του αλγορίθμου των δέντρων απόφασης υστερεί κατά πολύ σε σχέση με εκείνο των Τυχαίων Δασών και του Εγγύτερου Γείτονα καθώς παίρνει τιμή μόλις 60%.

Βάσει των παραπάνω προκύπτει πως ο αλγόριθμος ταξινόμησης που έχει τα βέλτιστα αποτελέσματα σε ό,τι αφορά την ανίχνευση κτιρίων είναι εκείνος των Τυχαίων Δασών. Ο συγκεκριμένος υστερεί σε σχέση με τον Εγγύτερου Γείτονα σε ό,τι αφορά το κριτήριο της πληρότητας, αλλά το αποτέλεσμα του είναι εμφανώς πιο αξιόπιστο συγκριτικά του τελευταίου.

ΠΙΝΑΚΑΣ 3.69: ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ ΓΙΑ ΤΟΥΣ ΑΛΓΟΡΙΘΜΟΥΣ ΤΑΞΙΝΟΜΗΣΗΣ «ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ», «ΤΥΧΑΙΑ ΔΑΣΗ» ΚΑΙ «ΕΓΓΥΤΕΡΟΣ ΓΕΙΤΟΝΑΣ»

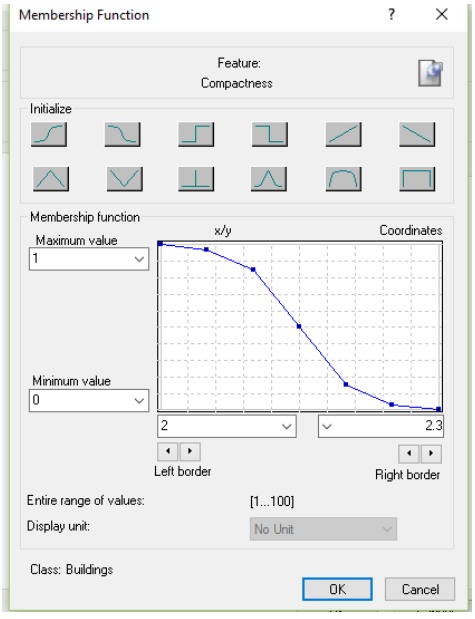
	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Δέντρα απόφασης	59,87%	**81,03%**	52,51%	40,13%	14,01%
Τυχαία Δάση	82,17%	71,27%	**61,72%**	17,83%	33,12%
Εγγύτερος Γείτονας	**85,99%**	65,85%	59,47%	14,01%	44,59%

3.10 Αξιολόγηση των επιδόσεων του αλγορίθμου «Εγγύτερος γείτονας» σε συνδυασμό με fuzzy κανόνες

Στον Πίνακας 3.69 παρατηρείται πως οι αλγόριθμοι δέντρα απόφασης και Εγγύτερος Γείτονας εμφανίζουν τα υψηλότερα ποσοστά ορθότητας και πληρότητας αντίστοιχα. Στα πλαίσια της παρούσας ενότητας επιχειρείται μία προσπάθεια βελτίωσης των επιδόσεων του αλγορίθμου «Εγγύτερος Γείτονας» προσθέτοντας fuzzy κανόνες. Για τη σύνταξη των τελευταίων αξιοποιήθηκαν ορισμένοι από τους κανόνες βάσει των οποίων έγινε ο διαχωρισμός των δεδομένων μέσω των δέντρων απόφασης.

Πιο συγκεκριμένα, στο αποτέλεσμα της ταξινόμησης μέσω του αλγορίθμου του «Εγγύτερου Γείτονα» παρατηρείται πως πολλά από τα αντικείμενα τα οποία έχουν καταχωρηθεί στη θεματική κατηγορία των κτιρίων ανήκουν στην πραγματικότητα σε εκείνη των χώρων στάθμευσης και των δρόμων. Για το λόγο αυτό στη θεματική κατηγορία «Κτίρια» διατυπώνονται οι ακόλουθοι κανόνες:

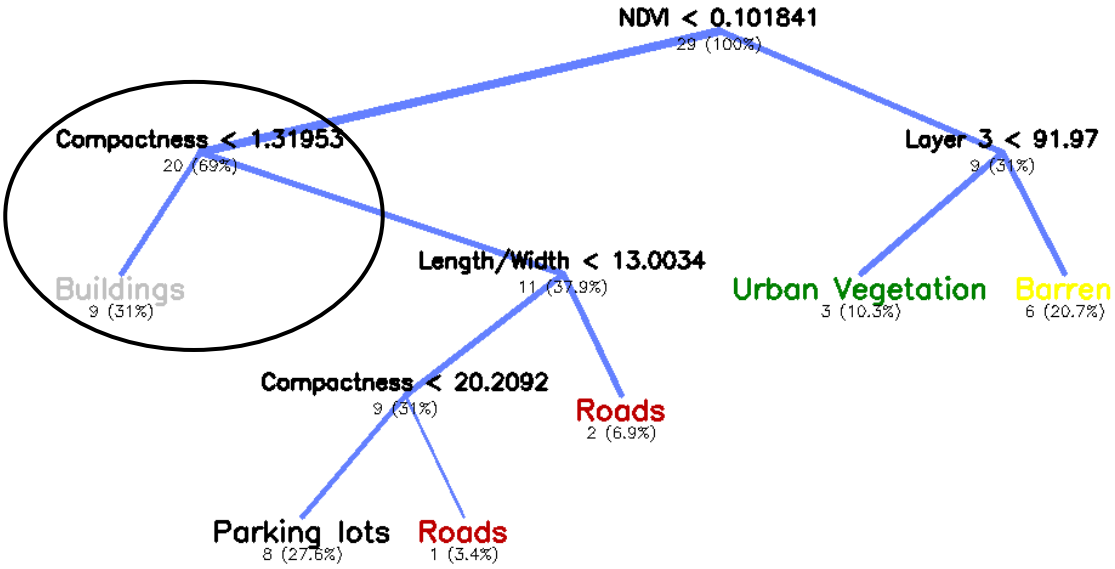
- Συμπαγότητα αντικειμένου (Εικόνα 3.197)



ΕΙΚΟΝΑ 3.197: FUZZY ΚΑΝΟΝΑΣ ΣΤΗ ΘΕΜΑΤΙΚΗ ΚΑΤΗΓΟΡΙΑ "ΚΤΙΡΙΑ" ΣΧΕΤΙΚΑ ΜΕ ΤΗ ΣΥΜΠΑΓΟΤΗΤΑ ΤΟΥ ΑΝΤΙΚΕΙΜΕΝΟΥ

Ο παραπάνω κανόνας αποτελεί κριτήριο στο τελικό μοντέλο του αλγορίθμου «Δέντρα απόφασης» στον κόμβο που αφορά στην καταχώρηση των αντικείμενων στη θεματική κατηγορία «Κτίρια» (Εικόνα 3.198). Σημειώνεται, ωστόσο πως η τιμή του κατωφλίου είναι μεγαλύτερη συγκριτικά με εκείνη που ορίζεται στο δέντρο απόφασης (2,3 και όχι 1,3 όπως ορίζεται στο δέντρο απόφασης). Το παραπάνω οφείλεται στο γεγονός πως όταν επιλέγεται

η τιμή 1,3 πολλά από τα αντικείμενα της θεματικής κατηγορίας των κτιρίων ταξινομούνται εσφαλμένα στην κατηγορία των χώρων στάθμευσης.



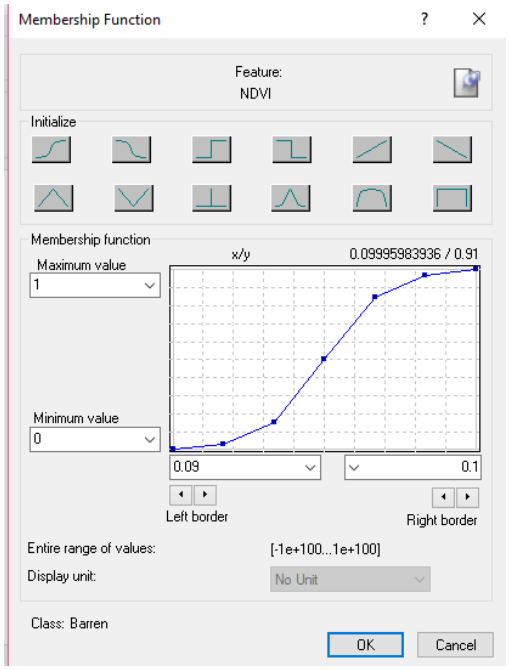
ΕΙΚΟΝΑ 3.198: ΤΕΛΙΚΟ ΜΟΝΤΕΛΟ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ «ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ»

Στην Εικόνα 3.199 εμφανίζεται το αποτέλεσμα της ταξινόμησης ττου αλγορίθμου «Εγγύτερος Γείτονας», έπειτα από την εφαρμογή του προαναφερθέντος fuzzy κανόνα. Το αποτέλεσμα είναι εμφανώς βελτιωμένο σε σχέση με το αρχικό αποτέλεσμα σε ό,τι αφορά την ικανοποίηση του κριτηρίου της ορθότητας. Ωστόσο πολλά αντικείμενα της συγκεκριμένης θεματικής κατηγορίας έχουν προστεθεί εσφαλμένα στην κλάση του άγονου εδάφους.

ΕΙΚΟΝΑ 3.199: ΑΠΟΤΕΛΕΣΜΑ ΤΑΞΙΝΟΜΗΣΗΣ ΜΕ ΤΟΝ ΑΛΓΟΡΙΘΜΟ ΕΓΓΥΤΕΡΟΣ ΓΕΙΤΟΝΑΣ ΕΠΕΙΤΑ ΑΠΟ ΕΦΑΡΜΟΓΗ ΤΟΥ ΚΑΝΟΝΑ ΤΗΣ ΣΥΜΠΑΓΟΤΗΤΑΣ ΣΤΗ ΘΕΜΑΤΙΚΗ ΚΑΤΗΓΟΡΙΑ ΤΩΝ ΚΤΙΡΙΩΝ

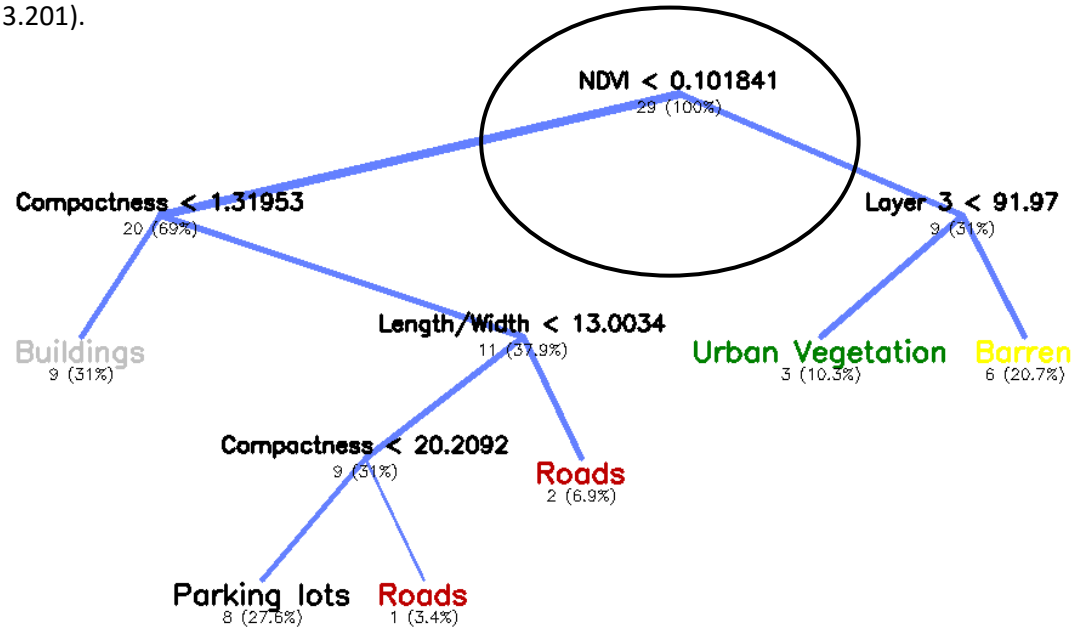
Για το σκοπό στην κλάση «Άγονο Έδαφος» προστίθεται ο ακόλουθος κανόνας:

- Κανονικοποιημένος δείκτης βλάστησης (Εικόνα 3.200)



ΕΙΚΟΝΑ 3.200: FUZZY ΚΑΝΟΝΑΣ ΣΤΗ ΘΕΜΑΤΙΚΗ ΚΑΤΗΓΟΡΙΑ "ΑΓΟΝΟ ΈΔΑΦΟΣ" ΣΧΕΤΙΚΑ ΜΕ ΤΟΝ ΚΑΝΟΝΙΚΟΠΟΙΗΜΕΝΟ ΔΕΙΚΤΗ ΒΛΑΣΤΗΣΗΣ

Ο παραπάνω κανόνας αποτελεί κριτήριο διαχωρισμού στο τελικό μοντέλο του αλγορίθμου «Δέντρα απόφασης» στον κόμβο που αφορά στο διαχωρισμό των αντικείμενων σε «Κτίρια» «Δρόμοι» και «Χώροι στάθμευσης» και σε «Αστική βλάστηση» και «Άγονο Έδαφος»(Εικόνα 3.201).



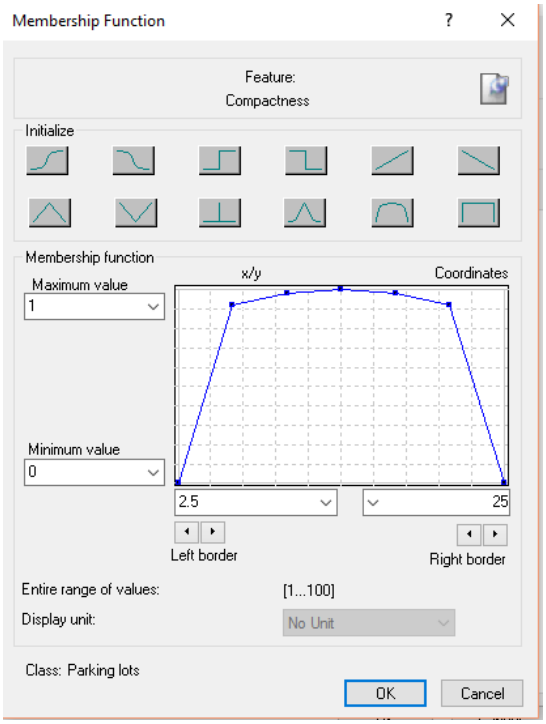
ΕΙΚΟΝΑ 3.201: ΤΕΛΙΚΟ ΜΟΝΤΕΛΟ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ «ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ»

Στην Εικόνα 3.202 εμφανίζεται το αποτέλεσμα της ταξινόμησης έπειτα από εφαρμογή του ως άνω κανόνα στη θεματική κατηγορία του άγονο έδαφος. Παρατηρείται πως πολλά από τα αντικείμενα τα οποία ανήκουν στη θεματική κατηγορία των κτιρίων έχουν εσφαλμένα ταξινομηθεί στην κλάση των χώρων στάθμευσης.

ΕΙΚΟΝΑ 3.202: ΑΠΟΤΕΛΕΣΜΑ ΤΑΞΙΝΟΜΗΣΗΣ ΜΕ ΤΟΝ ΑΛΓΟΡΙΘΜΟ ΕΓΓΥΤΕΡΟΣ ΓΕΙΤΟΝΑΣ ΕΠΕΙΤΑ ΑΠΟ ΕΦΑΡΜΟΓΗ ΤΟΥ ΚΑΝΟΝΑ ΓΙΑ ΤΟΝ ΚΑΝΟΝΙΚΟΠΟΙΗΜΕΝΟ ΔΕΙΚΤΗ ΒΛΑΣΤΗΣΗΣ ΣΤΗ ΘΕΜΑΤΙΚΗ ΚΑΤΗΓΟΡΙΑ ΤΟΥ ΆΓΟΝΟΥ ΕΔΑΦΟΥΣ

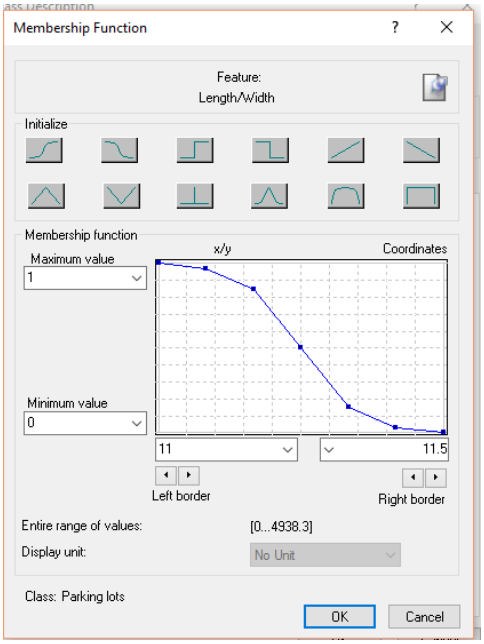
Για το λόγο αυτό, στη θεματική κατηγορία «Χώροι στάθμευσης» προστέθηκαν οι ακόλουθοι κανόνες:

- Συμπαγότητα αντικειμένου (Εικόνα 3.203)



ΕΙΚΟΝΑ 3.203: FUZZY ΚΑΝΟΝΑΣ ΣΧΕΤΙΚΑ ΜΕ ΤΗ ΣΥΜΠΑΓΟΤΗΤΑ ΣΤΗ ΘΕΜΑΤΙΚΗ ΚΑΤΗΓΟΡΙΑ «ΧΩΡΟΙ ΣΤΑΘΜΕΥΣΗΣ»

- Λόγος μήκους προς πλάτος (Εικόνα 3.204)



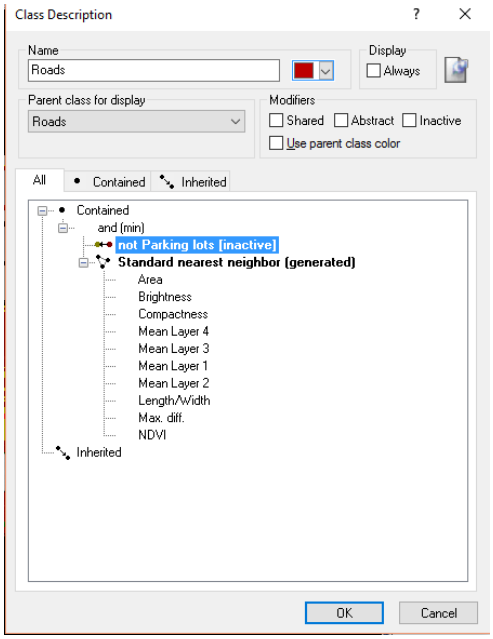
ΕΙΚΟΝΑ 3.204: FUZZY ΚΑΝΟΝΑΣ ΣΧΕΤΙΚΑ ΜΕ ΤΟ ΛΟΓΟ ΜΗΚΟΥΣ ΠΡΟΣ ΠΛΑΤΟΣ ΣΤΗ ΘΕΜΑΤΙΚΗ ΚΑΤΗΓΟΡΙΑ «ΧΩΡΟΙ ΣΤΑΘΜΕΥΣΗΣ»

Στην Εικόνα 3.205 εμφανίζεται το αποτέλεσμα της ταξινόμησης έπειτα από την εφαρμογή των παραπάνω κανόνων. Στη συγκεκριμένη παρατηρείται πως πολλά από τα αντικείμενα των χώρων στάθμευσης έχουν τοποθετηθεί στη θεματική κατηγορία των δρόμων.

ΕΙΚΟΝΑ 3.205: ΑΠΟΤΕΛΕΣΜΑ ΤΑΞΙΝΟΜΗΣΗΣ ΜΕ ΤΟΝ ΑΛΓΟΡΙΘΜΟ ΕΓΓΥΤΕΡΟΣ ΓΕΙΤΟΝΑΣ ΕΠΕΙΤΑ ΑΠΟ ΕΦΑΡΜΟΓΗ ΤΩΝ ΚΑΝΟΝΩΝ ΛΟΓΟΣ ΜΗΚΟΥΣ ΠΡΟΣ ΠΛΑΤΟΣ ΚΑΙ ΣΥΜΠΑΓΟΤΗΤΑ ΣΤΗ ΘΕΜΑΤΙΚΗ ΚΑΤΗΓΟΡΙΑ «ΧΩΡΟΙ ΣΤΑΘΜΕΥΣΗΣ»

Για το σκοπό αυτό στη θεματική κατηγορία των δρόμων προστίθεται ο ακόλουθος κανόνας:

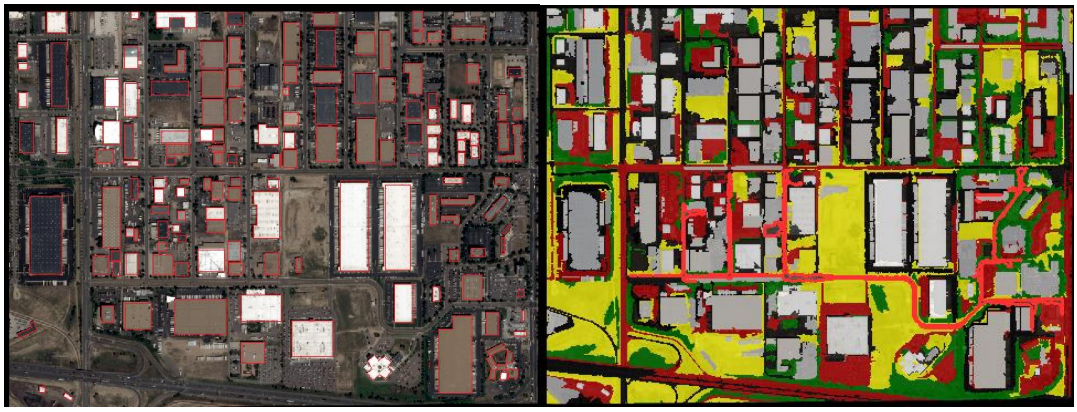
- Όχι χώροι στάθμευσης (Εικόνα 3.206)



ΕΙΚΟΝΑ 3.206: ΚΑΝΟΝΑΣ ΟΧΙ ΧΩΡΟΙ ΣΤΑΘΜΕΥΣΗΣ

Στην Εικόνα 3.207 εμφανίζεται το αποτέλεσμα της ταξινόμησης έπειτα από την εφαρμογή του παραπάνω κανόνα. Παρατηρείται πως τα αντικείμενα της θεματικής κατηγορίας των δρόμων έχουν μειωθεί σε αυτήν, ωστόσο εξακολουθούν να υπάρχουν προβλήματα. Η συγκεκριμένη εργασία, ωστόσο, εστιάζει σε θέματα ανίχνευσης κτιρίων και ως εκ τούτου δε θα δοθεί μεγαλύτερη προσοχή στο εν λόγω εμφανιζόμενο πρόβλημα.

ΕΙΚΟΝΑ 3.207: ΑΠΟΤΕΛΕΣΜΑ ΤΑΞΙΝΟΜΗΣΗΣ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ "ΕΓΓΥΤΕΡΟΣ ΓΕΙΤΟΝΑΣ" ΕΠΕΙΤΑ ΑΠΟ ΕΦΑΡΜΟΓΗ ΤΟΥ ΚΑΝΟΝΑ ΟΧΙ ΧΩΡΟΙ ΣΤΑΘΜΕΥΣΗΣ ΣΤΗ ΘΕΜΑΤΙΚΗ ΚΑΤΗΓΟΡΙΑ ΤΩΝ ΔΡΟΜΩΝ



ΕΙΚΟΝΑ 3.208: ΤΕΛΙΚΟ ΑΠΟΤΕΛΕΣΜΑ ΕΦΑΡΜΟΓΗΣ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ «ΕΓΓΥΤΕΡΟΣ ΓΕΙΤΟΝΑΣ» ΜΕ FUZZY ΚΑΝΟΝΕΣ

Στους Πίνακας 3.70, Πίνακας 3.71 εμφανίζονται οι δείκτες ποιότητας για το αποτέλεσμα της ταξινόμησης του Εγγύτερου Γείτονα βάσει της

ΠΙΝΑΚΑΣ 3.70*: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (ΕΓΓΥΤΕΡΟΣ ΓΕΙΤΟΝΑΣ ΜΕ FUZZY ΚΑΝΟΝΕΣ).*

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Σύνολο εικόνας	129	50	28

ΠΙΝΑΚΑΣ 3.71*: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (ΕΓΓΥΤΕΡΟΣ ΓΕΙΤΟΝΑΣ ΜΕ FUZZY ΚΑΝΟΝΕΣ)*

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Σύνολο εικόνας	82,17%	72,07%	62,32%	17,83%	31,85%

Στα παραπάνω αποτελέσματα παρατηρείται πως η εφαρμογή fuzzy κανόνων βελτίωσε το αποτέλεσμα της ταξινόμησης του αλγορίθμου «εγγύτερος γείτονας» και οι επιδόσεις του είναι αντίστοιχες με εκείνες των τυχαίων δασών.

3.12 Επιλογή των χαρακτηριστικών των αντικειμένων για τη διαδικασία της ταξινόμησης μέσω του αλγορίθμου των τυχαίων δασών

Στα πλαίσια της συγκεκριμένης ενότητας έγινε επιλογή των χαρακτηριστικών των αντικειμένων που θα δώσουν το βέλτιστο αποτέλεσμα ταξινόμησης με το χαμηλότερο δυνατό υπολογιστικό κόστος. Με άλλα λόγια επιδιώκεται ο προσδιορισμός των γνωρισμάτων εκείνων τα οποία θα οδηγήσουν σε ένα αποτελεσματικό διαχωρισμό των αντικειμένων που ανήκουν σε διαφορετικές θεματικές κατηγορίες.

Η διαδικασία που εφαρμόστηκε βασίστηκε σε εκείνη των (Diaz- Uriate and Alvarez de Andres, 2006). Οι συγκεκριμένοι προτείνουν τον υπολογισμό της σημαντικότητας των μεταβλητών (Variable Importance) μέσω της κατασκευής ενός αρχικού μοντέλου των τυχαίων δασών με πλήθος δέντρων 5000. Το 20% των λιγότερο σημαντικών χαρακτηριστικών αφαιρείται από τη λίστα και επαναλαμβάνεται η διαδικασία δημιουργίας ενός μοντέλου το οποίο αποτελείται αυτή τη φορά από 2000 δέντρα. Τα 20% των λιγότερο σημαντικών χαρακτηριστικών που προέκυψαν βάσει της σημαντικότητας των μεταβλητών του τελευταίου μοντέλου αφαιρείται από τη διαδικασία της δημιουργίας ενός νέου μοντέλου 2000 δέντρων. Το παραπάνω βήμα επαναλαμβάνεται πολλές φορές και το μοντέλο το οποίο τελικά επιλέγεται είναι εκείνο με το μικρότερο σφάλμα oob.

Στα πλαίσια της παρούσας εφαρμογής έγινε εφαρμογή της παραπάνω διαδικασίας με ορισμένες διαφοροποιήσεις. Αρχικά διευκρινίζεται πως το πλήθος των δέντρων που κατασκευάστηκαν σε όλα τα μοντέλα είναι ίσο με 50. Ο λόγος για τον οποίο επιλέχθηκε η συγκεκριμένη τιμή προκύπτει από τα συμπεράσματα της 7^{ης} δοκιμής της ενότητας 3.7. Μέσω αυτής διερευνήθηκε η επιρροή της παραμέτρου πλήθος δέντρων στην ποιότητα της ταξινόμησης. Η δημιουργία δάσους με περισσότερα των 50 δέντρα αυξάνει το υπολογιστικό κόστος χωρίς να συνεισφέρει στη βελτίωση της ποιότητας της ταξινόμησης. Αναλυτικά, οι τιμές των παραμέτρων του αλγορίθμου των τυχαίων δασών ορίστηκαν ως εξής:

- Βάθος δέντρων: 10
- Ελάχιστος αριθμός δειγμάτων ανά κόμβο: 0
- Χρήση αντικαταστατών: Όχι
- Μέγιστος αριθμός κατηγοριών: 16
- Πλήθος ενεργών μεταβλητών: 0 (δηλαδή ίσο με τη ρίζα του πλήθους των χαρακτηριστικών)
- Πλήθος δέντρων: 50
- Ακρίβεια τυχαίου δάσους:
- Κριτήριο τερματισμού: Και τα δύο

Οι συγκεκριμένες δίνουν τα βέλτιστα δυνατά αποτελέσματα σε ό,τι αφορά την ανίχνευση κτιρίων βάσει των συμπερασμάτων της ενότητας 3.7.

1^{η} επανάληψη (όλα τα χαρακτηριστικά – 17’ η εφαρμογή του μοντέλου στα δεδομένα)

Στα πλαίσια της πρώτης επανάληψης έγινε δημιουργία ενός μοντέλου τυχαίων δασών στο οποίο έγινε αξιοποίηση των ακόλουθων γνωρισμάτων του Πίνακας 3.72. Πολλά από τα ακόλουθα χαρακτηριστικά είναι εκείνα που προτείνουν οι (Du, Zhang, Zhang, 2015)

ΠΙΝΑΚΑΣ 3.72: ΓΝΩΡΙΣΜΑΤΑ ΣΤΑ ΠΛΑΙΣΙΑ ΤΗΣ 1ΗΣ ΕΠΑΝΑΛΗΨΗΣ

Φασματικά χαρακτηριστικά	Mean
	Brightness
	Max. Diff.
	Std. Dev.
	Skewness
	HSI
	NDVI
	MSAVI2
Υφής	GLCM std. dev
	GLCM Mean
	GLCM Homegeneity
	GLCM Contrast
	GLCM Dissimilarity
	GLCM Entropy
	GLDV Ang. 2nd Mom.
	GLCM Correlation
	GLDV
Γεωμετρίας	Area
	Border length
	Length
	Length/ width
	Border Index
	Compactness
	Elliptic fit
	Main direction
	Radius Ellipse
	Rectangular fit
	Roundness
	Shape index
	Rel. border to image border
	Number of pixels
	Width
	Assymetry
	Density

Η διαδικασία της εκπαίδευσης του αλγορίθμου ήταν ιδιαίτερα χρονοβόρα στην προκειμένη περίπτωση και έφτασε τα 6 λεπτά, ενώ εκείνη της εφαρμογής του μοντέλου στα δεδομένα εκπαίδευσης έφτασε τα 30 λεπτά.

Το αποτέλεσμα της ταξινόμησης εμφανίζεται στην ακόλουθη εικόνα (Εικόνα 3.209).

ΕΙΚΟΝΑ 3.209: ΑΠΟΤΕΛΕΣΜΑ ΕΦΑΡΜΟΓΗΣ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΓΙΑ ΤΗΝ ΠΡΩΤΗ ΕΠΑΝΑΛΗΨΗ (ΟΛΑ ΤΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ)

ΕΙΚΟΝΑ 3.210: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΗΝ ΠΡΩΤΗ ΔΟΚΙΜΗ

Η διαδικασία που ακολουθείται προκειμένου να γίνει ποσοτική αξιολόγηση είναι η εξής: Επιλέγεται αντιπροσωπευτική περιοχή της εικόνας- εισόδου (Εικόνα 3.210). Για την περιοχή αυτή εντοπίζεται μέσω φωτοερμηνείας ο αριθμός των True Positives, False Positives και False Negatives δεδομένων. Βάσει των παραπάνω στοιχείων κατασκευάζονται οι ακόλουθοι Πίνακες (Πίνακας 3.73, Πίνακας 3.74):

ΠΙΝΑΚΑΣ 3.73: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (1Η ΕΠΑΝΑΛΗΨΗ).

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	13	3	3

ΠΙΝΑΚΑΣ 3.74: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (1Η ΕΠΑΝΑΛΗΨΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	81,25%	81,25%	68,42%	18,75%	18,75%

Στον ακόλουθο Πίνακα (Πίνακας 3.75) εμφανίζονται οι τιμές της σημαντικότητας για κάθε μία από τις μεταβλητές του Πίνακας 3.72.

ΠΙΝΑΚΑΣ 3.75: ΤΙΜΕΣ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ ΤΩΝ ΜΕΤΑΒΛΗΤΩΝ ΓΙΑ ΤΗΝ 1Η ΕΠΑΝΑΛΗΨΗ

Feature	Importance
Brightness	0
Mean Layer 1	0
Mean Layer 3	0
Mean Layer 4	0
Standard deviation Layer 3	0
Standard deviation Layer 4	0
Skewness Layer 1	0
Skewness Layer 2	0
Skewness Layer 3	0
Skewness Layer 4	0
HSI Transformation Hue(R='Layer 3',G='Layer 2',B='Layer 1')	0
HSI Transformation Intensity(R=Layer 3,G=Layer 2,B=Layer 1)	0
HSI Transformation Saturation(R=Layer 3,G=Layer 2,B=Layer 1)	0
Border length	0
GLCM StdDev (all dir.)	0
GLCM StdDev (0ϱ)	0
GLCM StdDev (45ϱ)	0
GLCM StdDev (90ϱ)	0
GLCM Homogeneity (45ϱ)	0
GLCM Homogeneity (90ϱ)	0
GLCM Homogeneity (135ϱ)	0
GLCM Contrast (all dir.)	0
GLCM Contrast (45ϱ)	0
GLCM Contrast (90ϱ)	0
GLCM Dissimilarity (all dir.)	0
GLCM Dissimilarity (0ϱ)	0
GLCM Dissimilarity (45ϱ)	0
GLCM Dissimilarity (90ϱ)	0
GLDV Ang. 2nd moment (all dir.)	0
GLDV Ang. 2nd moment (0ϱ)	0
GLDV Ang. 2nd moment (45ϱ)	0
GLDV Ang. 2nd moment (90ϱ)	0
GLDV Ang. 2nd moment (135ϱ)	0
GLCM Correlation (0ϱ)	0
GLCM Correlation (45ϱ)	0
GLCM Correlation (90ϱ)	0
Area	0
Length/Width	0
Rel. border to image border	0
Compactness	0
Main direction	0

Radius of smallest enclosing ellipse	0
Roundness	0
Shape index	0
MSAVI2	0
Number of pixels	0
Width	0
Asymmetry	0
Border index	0
Density	0
GLCM Entropy (all dir.)	0
GLCM Entropy (0ฐ)	0
GLCM Entropy (45ฐ)	0
GLCM Entropy (90ฐ)	0
GLCM Entropy (135ฐ)	0
GLCM Ang. 2nd moment (all dir.)	0
GLCM Ang. 2nd moment (0ฐ)	0
GLCM Ang. 2nd moment (90ฐ)	0
GLCM Ang. 2nd moment (135ฐ)	0
GLCM Mean (all dir.)	0
GLCM Mean (0ฐ)	0
GLCM Mean (45ฐ)	0
GLCM Mean (90ฐ)	0
GLCM Mean (135ฐ)	0
Mean Layer 2	0.0303030312
Standard deviation Layer 2	0.0303030312
GLCM Homogeneity (all dir.)	0.0303030312
GLCM Homogeneity (0ฐ)	0.0303030312
GLCM Contrast (135ฐ)	0.0303030312
GLCM Dissimilarity (135ฐ)	0.0303030312
GLCM Correlation (135ฐ)	0.0303030312
Elliptic fit	0.0303030312
Radius of largest enclosed ellipse	0.0303030312
GLCM Ang. 2nd moment (45ฐ)	0.0303030312
GLCM StdDev (135ฐ)	0.0606060624
GLCM Correlation (all dir.)	0.0606060624
Rectangular fit	0.0606060624
Length	0.0606060624
NDVI	0.0909090936
GLCM Contrast (0ฐ)	0.0909090936
Standard deviation Layer 1	0.1212121248
Max. diff.	0.151515156

2η επανάληψη(15’ η εφαρμογή του μοντέλου στα δεδομένα)

Στα πλαίσια της δεύτερης επανάληψης έγινε αφαίρεση των 20% λιγότερο σημαντικών χαρακτηριστικών βάσει των τιμών σημαντικότητας του Πίνακας 3.75.Τα γνωρίσματα βάσει

του οποίου έγινε εκπαίδευση του νέου μοντέλου των τυχαίων δασών αναγράφονται στον Πίνακας 3.76.

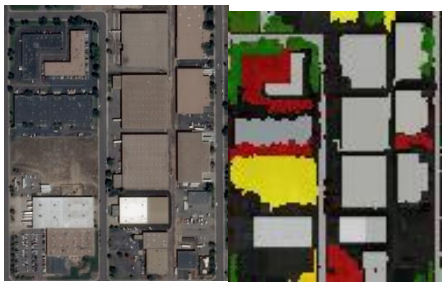
ΠΙΝΑΚΑΣ 3.76: ΓΝΩΡΙΣΜΑΤΑ ΣΤΑ ΠΛΑΙΣΙΑ ΤΗΣ 2ΗΣ ΕΠΑΝΑΛΗΨΗΣ

Feature
Brightness
Mean Layer 3
GLCM StdDev (135₰)
GLCM Contrast (90₰)
GLCM Correlation (0₰)
Width
Mean Layer 1
Mean Layer 2
Skewness Layer 4
HSI Transformation Saturation(R=Layer 3,G=Layer 2,B=Layer 1)
GLCM StdDev (all dir.)
GLCM Homogeneity (0₰)
GLCM Dissimilarity (90₰)
GLDV Ang. 2nd moment (all dir.)
GLDV Ang. 2nd moment (45₰)
GLDV Ang. 2nd moment (135₰)
Area
Rel. border to image border
Elliptic fit
Radius of largest enclosed ellipse
Shape index
Length
Asymmetry
Border index
GLCM Entropy (all dir.)
GLCM Entropy (0₰)
GLCM Entropy (45₰)
GLCM Entropy (90₰)
GLCM Entropy (135₰)
GLCM Mean (all dir.)
GLCM Mean (0₰)
GLCM Mean (45₰)
GLCM Mean (90₰)
GLCM Mean (135₰)
HSI Transformation Hue(R='Layer 3',G='Layer 2',B='Layer 1')
GLCM Homogeneity (90₰)
GLCM Contrast (45₰)
GLCM Correlation (all dir.)
GLCM Correlation (45₰)
GLCM Correlation (135₰)
Roundness
Number of pixels
Density
Mean Layer 4
Standard deviation Layer 4
GLCM Contrast (all dir.)

GLCM Dissimilarity (all dir.)
GLCM Dissimilarity (0ฐ)
GLCM Dissimilarity (135ฐ)
GLCM Correlation (90ฐ)
Compactness
MSAVI2
GLCM Ang. 2nd moment (all dir.)
GLCM Dissimilarity (45ฐ)
Length/Width
GLCM Homogeneity (45ฐ)
HSI Transformation Intensity(R=Layer 3,G=Layer 2,B=Layer 1)
NDVI
GLCM Homogeneity (135ฐ)
Standard deviation Layer 1
Standard deviation Layer 2
GLCM Homogeneity (all dir.)
GLCM Contrast (135ฐ)
Max. diff.
GLCM Contrast (0ฐ)
Standard deviation Layer 3

Στην ακόλουθη Εικόνα (Εικόνα 3.211) εμφανίζεται το αποτέλεσμα εφαρμογής του μοντέλου των τυχαίων δασών στα πλαίσια της δεύτερης επανάληψης.

ΕΙΚΟΝΑ 3.211: ΑΠΟΤΕΛΕΣΜΑ ΕΦΑΡΜΟΓΗΣ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΓΙΑ ΤΗ ΔΕΥΤΕΡΗ ΕΠΑΝΑΛΗΨΗ



ΕΙΚΟΝΑ 3.212: *ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΗ ΔΕΥΤΕΡΗ ΕΠΑΝΑΛΗΨΗ*

Βάσει της Εικόνα 3.212 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.77, Πίνακας 3.78)

ΠΙΝΑΚΑΣ 3.77*: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (2Η ΕΠΑΝΑΛΗΨΗ)*

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	13	2	3

ΠΙΝΑΚΑΣ 3.78: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (2Η ΕΠΑΝΑΛΗΨΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	81,25%	86,67%	72,22%	18,75%	12,50%

Στον ακόλουθο Πίνακα (Πίνακας 3.79) εμφανίζονται οι τιμές της σημαντικότητας για κάθε μία από τις μεταβλητές της δεύτερης επανάληψης

ΠΙΝΑΚΑΣ 3.79: ΤΙΜΕΣ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ ΤΩΝ ΜΕΤΑΒΛΗΤΩΝ ΓΙΑ ΤΗΝ 2Η ΕΠΑΝΑΛΗΨΗ

Feature	**Importance**
Brightness	0.0095238099
Mean Layer 3	0.0063492064
GLCM StdDev (135₪)	0.0063492064
GLCM Contrast (90₪)	0.0063492064
GLCM Correlation (0₪)	0.0063492064
Width	0.0063492064
Mean Layer 1	0.0095238099
Mean Layer 2	0.0095238099
Skewness Layer 4	0.0095238099
HSI Transformation Saturation(R=Layer 3,G=Layer 2,B=Layer 1)	0.0095238099
GLCM StdDev (all dir.)	0.0095238099
GLCM Homogeneity (0₪)	0.0095238099
GLCM Dissimilarity (90₪)	0.0095238099
GLDV Ang. 2nd moment (all dir.)	0.0095238099
GLDV Ang. 2nd moment (45₪)	0.0095238099
GLDV Ang. 2nd moment (135₪)	0.0095238099
Area	0.0095238099
Rel. border to image border	0.0095238099
Elliptic fit	0.0095238099
Radius of largest enclosed ellipse	0.0095238099
Shape index	0.0095238099
Length	0.0095238099
Asymmetry	0.0095238099
Border index	0.0095238099
GLCM Entropy (all dir.)	0.0095238099
GLCM Entropy (0₪)	0.0095238099
GLCM Entropy (45₪)	0.0095238099
GLCM Entropy (90₪)	0.0095238099

GLCM Entropy (135฀)	0.0095238099
GLCM Mean (all dir.)	0.0095238099
GLCM Mean (0฀)	0.0095238099
GLCM Mean (45฀)	0.0095238099
GLCM Mean (90฀)	0.0095238099
GLCM Mean (135฀)	0.0095238099
HSI Transformation Hue(R='Layer 3',G='Layer 2',B='Layer 1')	0.0126984129
GLCM Homogeneity (90฀)	0.0126984129
GLCM Contrast (45฀)	0.0126984129
GLCM Correlation (all dir.)	0.0126984129
GLCM Correlation (45฀)	0.0126984129
GLCM Correlation (135฀)	0.0126984129
Roundness	0.0126984129
Number of pixels	0.0126984129
Density	0.0126984129
Mean Layer 4	0.0158730168
Standard deviation Layer 4	0.0158730168
GLCM Contrast (all dir.)	0.0158730168
GLCM Dissimilarity (all dir.)	0.0158730168
GLCM Dissimilarity (0฀)	0.0158730168
GLCM Dissimilarity (135฀)	0.0158730168
GLCM Correlation (90฀)	0.0158730168
Compactness	0.0158730168
MSAVI2	0.0158730168
GLCM Ang. 2nd moment (all dir.)	0.0158730168
GLCM Dissimilarity (45฀)	0.0190476198
Length/Width	0.0190476198
GLCM Homogeneity (45฀)	0.0222222228
HSI Transformation Intensity(R=Layer 3,G=Layer 2,B=Layer 1)	0.0253968257
NDVI	0.0253968257
GLCM Homogeneity (135฀)	0.0253968257
Standard deviation Layer 1	0.0285714287
Standard deviation Layer 2	0.0285714287
GLCM Homogeneity (all dir.)	0.0380952395
GLCM Contrast (135฀)	0.0380952395
Max. diff.	0.0444444455
GLCM Contrast (0฀)	0.0507936515
Standard deviation Layer 3	0.0539682545

3η επανάληψη (13’ η εφαρμογή του μοντέλου στα δεδομένα)

Στα πλαίσια της τρίτης επανάληψης έγινε αφαίρεση των 20% λιγότερο σημαντικών χαρακτηριστικών βάσει των τιμών σημαντικότητας του Πίνακας 3.79.Τα γνωρίσματα βάσει του οποίου έγινε εκπαίδευση του νέου μοντέλου των τυχαίων δασών αναγράφονται στον Πίνακας 3.80.

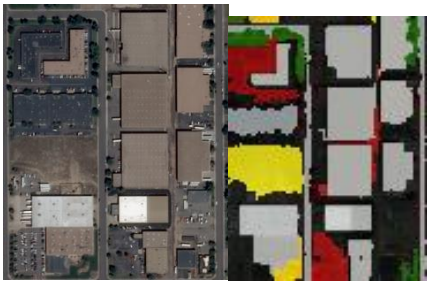
ΠΙΝΑΚΑΣ 3.80: ΓΝΩΡΙΣΜΑΤΑ ΣΤΑ ΠΛΑΙΣΙΑ ΤΗΣ 3ΗΣ ΕΠΑΝΑΛΗΨΗΣ

Feature
Mean Layer 4
Standard deviation Layer 3
HSI Transformation Hue(R='Layer 3',G='Layer 2',B='Layer 1')
GLCM Contrast (0ฐ)
GLCM Dissimilarity (0ฐ)
GLCM Dissimilarity (45ฐ)
GLCM Dissimilarity (135ฐ)
GLDV Ang. 2nd moment (all dir.)
GLCM Correlation (all dir.)
Area
Rel. border to image border
Compactness
Elliptic fit
Radius of largest enclosed ellipse
Roundness
Shape index
MSAVI2
Number of pixels
Asymmetry
Density
GLCM Entropy (0ฐ)
GLCM Entropy (45ฐ)
GLCM Entropy (135ฐ)
GLCM Ang. 2nd moment (all dir.)
GLCM Mean (all dir.)
GLCM Mean (45ฐ)
GLCM Mean (90ฐ)
GLCM Mean (135ฐ)
Standard deviation Layer 4
GLCM Homogeneity (90ฐ)
GLCM Contrast (45ฐ)
GLCM Correlation (90ฐ)
Length/Width
GLCM Entropy (all dir.)
NDVI
GLCM Dissimilarity (all dir.)
GLDV Ang. 2nd moment (45ฐ)
GLDV Ang. 2nd moment (135ฐ)
Length
GLCM Entropy (90ฐ)
GLCM Mean (0ฐ)
Max. diff.

Standard deviation Layer 1
HSI Transformation Intensity(R=Layer 3,G=Layer 2,B=Layer 1)
GLCM Homogeneity (all dir.)
GLCM Correlation (135ϱ)
GLCM Homogeneity (45ϱ)
Border index
Standard deviation Layer 2
GLCM Homogeneity (135ϱ)
GLCM Correlation (45ϱ)
GLCM Contrast (all dir.)
GLCM Contrast (135ϱ)

Στην ακόλουθη Εικόνα (Εικόνα 3.213) εμφανίζεται το αποτέλεσμα εφαρμογής του μοντέλου των τυχαίων δασών στα πλαίσια της τρίτης επανάληψης.

ΕΙΚΟΝΑ 3.213: ΑΠΟΤΕΛΕΣΜΑ ΕΦΑΡΜΟΓΗΣ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΓΙΑ ΤΗΝ ΤΡΙΤΗ ΕΠΑΝΑΛΗΨΗ



ΕΙΚΟΝΑ 3.214: ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΗΝ ΤΡΙΤΗ ΕΠΑΝΑΛΗΨΗ

Βάσει της Εικόνα 3.212 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.81, Πίνακας 3.82)

ΠΙΝΑΚΑΣ 3.81: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (3Η ΕΠΑΝΑΛΗΨΗ)

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	14	2	2

ΠΙΝΑΚΑΣ 3.82: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (3Η ΕΠΑΝΑΛΗΨΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	87.50%	87.50%	77.78%	12.50%	12.50%

Στον ακόλουθο Πίνακα (Πίνακας 3.83) εμφανίζονται οι τιμές της σημαντικότητας για κάθε μία από τις μεταβλητές της τρίτης επανάληψης.

ΠΙΝΑΚΑΣ 3.83: ΤΙΜΕΣ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ ΜΕΤΑΒΛΗΤΩΝ ΓΙΑ ΤΗΝ ΤΡΙΤΗ ΕΠΑΝΑΛΗΨΗ

Feature	Importance
Mean Layer 4	0
Standard deviation Layer 3	0
HSI Transformation Hue(R='Layer 3',G='Layer 2',B='Layer 1')	0
GLCM Contrast (0₪)	0
GLCM Dissimilarity (0₪)	0
GLCM Dissimilarity (45₪)	0
GLCM Dissimilarity (135₪)	0
GLDV Ang. 2nd moment (all dir.)	0
GLCM Correlation (all dir.)	0
Area	0
Rel. border to image border	0
Compactness	0
Elliptic fit	0
Radius of largest enclosed ellipse	0
Roundness	0
Shape index	0
MSAVI2	0
Number of pixels	0
Asymmetry	0
Density	0
GLCM Entropy (0₪)	0
GLCM Entropy (45₪)	0
GLCM Entropy (135₪)	0
GLCM Ang. 2nd moment (all dir.)	0
GLCM Mean (all dir.)	0
GLCM Mean (45₪)	0
GLCM Mean (90₪)	0
GLCM Mean (135₪)	0
Standard deviation Layer 4	0.0123456791
GLCM Homogeneity (90₪)	0.0123456791
GLCM Contrast (45₪)	0.0123456791
GLCM Correlation (90₪)	0.0123456791
Length/Width	0.0123456791
GLCM Entropy (all dir.)	0.0123456791
NDVI	0.0246913582

GLCM Dissimilarity (all dir.)	0.0246913582
GLDV Ang. 2nd moment (45ฐ)	0.0246913582
GLDV Ang. 2nd moment (135ฐ)	0.0246913582
Length	0.0246913582
GLCM Entropy (90ฐ)	0.0246913582
GLCM Mean (0ฐ)	0.0246913582
Max. diff.	0.0370370373
Standard deviation Layer 1	0.0370370373
HSI Transformation Intensity(R=Layer 3,G=Layer 2,B=Layer 1)	0.0370370373
GLCM Homogeneity (all dir.)	0.0370370373
GLCM Correlation (135ฐ)	0.0370370373
GLCM Homogeneity (45ฐ)	0.0493827164
Border index	0.0493827164
Standard deviation Layer 2	0.0740740746
GLCM Homogeneity (135ฐ)	0.0740740746
GLCM Correlation (45ฐ)	0.0740740746
GLCM Contrast (all dir.)	0.0987654328
GLCM Contrast (135ฐ)	0.1481481493

4η επανάληψη (9’ η εφαρμογή του μοντέλου στα δεδομένα)

Στον Πίνακας 3.84 εμφανίζονται τα γνωρίσματα στα οποία βασίστηκε το μοντέλο της τέταρτης επανάληψης.

ΠΙΝΑΚΑΣ 3.84: ΓΝΩΡΙΣΜΑΤΑ ΣΤΑ ΠΛΑΙΣΙΑ ΤΗΣ 4ΗΣ ΕΠΑΝΑΛΗΨΗΣ

Feature
Mean Layer 4
GLCM Contrast (135ฐ)
GLDV Ang. 2nd moment (all dir.)
GLDV Ang. 2nd moment (135ฐ)
GLCM Correlation (90ฐ)
GLCM Correlation (135ฐ)
Area
Length/Width
Compactness
Roundness
Shape index
MSAVI2
GLCM Entropy (all dir.)
GLCM Entropy (0ฐ)
GLCM Ang. 2nd moment (all dir.)
Standard deviation Layer 4
GLCM Dissimilarity (0ฐ)
Length
Number of pixels

Asymmetry
GLCM Entropy (90ฐ)
GLCM Mean (all dir.)
GLCM Mean (0ฐ)
GLCM Dissimilarity (all dir.)
GLCM Correlation (all dir.)
GLCM Correlation (45ฐ)
Elliptic fit
Border index
Standard deviation Layer 2
HSI Transformation Intensity(R=Layer 3,G=Layer 2,B=Layer 1)
GLDV Ang. 2nd moment (45ฐ)
Standard deviation Layer 1
GLCM Homogeneity (45ฐ)
Standard deviation Layer 3
GLCM Homogeneity (all dir.)
GLCM Homogeneity (90ฐ)
GLCM Homogeneity (135ฐ)
GLCM Contrast (all dir.)
GLCM Contrast (0ฐ)
GLCM Contrast (45ฐ)
Max. diff.
NDVI

Στην ακόλουθη Εικόνα (Εικόνα 3.215) εμφανίζεται το αποτέλεσμα εφαρμογής του μοντέλου των τυχαίων δασών στα πλαίσια της τέταρτης επανάληψης.

ΕΙΚΟΝΑ 3.215: ΑΠΟΤΕΛΕΣΜΑ ΕΦΑΡΜΟΓΗΣ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΓΙΑ ΤΗΝ ΤΕΤΑΡΤΗ ΕΠΑΝΑΛΗΨΗ

ΕΙΚΟΝΑ 3.216: ***ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΗΝ ΤΕΤΑΡΤΗ ΕΠΑΝΑΛΗΨΗ***

Βάσει της Εικόνα 3.216 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.85, Πίνακας 3.86)

ΠΙΝΑΚΑΣ 3.85: ***ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (4Η ΕΠΑΝΑΛΗΨΗ)***

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	13	2	3

ΠΙΝΑΚΑΣ 3.86: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (4Η ΕΠΑΝΑΛΗΨΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	81,25%	86,67%	72,22%	18,75%	12,50%

Στον ακόλουθο Πίνακα (Πίνακας 3.87) εμφανίζονται οι τιμές της σημαντικότητας για κάθε μία από τις μεταβλητές της τέταρτης επανάληψης.

ΠΙΝΑΚΑΣ 3.87: ΤΙΜΕΣ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ ΜΕΤΑΒΛΗΤΩΝ ΓΙΑ ΤΗΝ ΤΕΤΑΡΤΗ ΕΠΑΝΑΛΗΨΗ

Feature	**Importance**
Mean Layer 4	0
GLCM Contrast (135₴)	0
GLDV Ang. 2nd moment (all dir.)	0
GLDV Ang. 2nd moment (135₴)	0
GLCM Correlation (90₴)	0
GLCM Correlation (135₴)	0
Area	0
Length/Width	0
Compactness	0
Roundness	0
Shape index	0
MSAVI2	0
GLCM Entropy (all dir.)	0
GLCM Entropy (0₴)	0
GLCM Ang. 2nd moment (all dir.)	0
Standard deviation Layer 4	0.0069930069
GLCM Dissimilarity (0₴)	0.0069930069
Length	0.0069930069
Number of pixels	0.0069930069

Asymmetry	0.0069930069
GLCM Entropy (90ฐ)	0.0069930069
GLCM Mean (all dir.)	0.0069930069
GLCM Mean (0ฐ)	0.0069930069
GLCM Dissimilarity (all dir.)	0.0139860138
GLCM Correlation (all dir.)	0.0139860138
GLCM Correlation (45ฐ)	0.0139860138
Elliptic fit	0.0139860138
Border index	0.0209790207
Standard deviation Layer 2	0.0279720277
HSI Transformation Intensity(R=Layer 3,G=Layer 2,B=Layer 1)	0.0279720277
GLDV Ang. 2nd moment (45ฐ)	0.0419580415
Standard deviation Layer 1	0.0489510484
GLCM Homogeneity (45ฐ)	0.0489510484
Standard deviation Layer 3	0.0559440553
GLCM Homogeneity (all dir.)	0.0559440553
GLCM Homogeneity (90ฐ)	0.062937066
GLCM Homogeneity (135ฐ)	0.062937066
GLCM Contrast (all dir.)	0.062937066
GLCM Contrast (0ฐ)	0.062937066
GLCM Contrast (45ฐ)	0.062937066
Max. diff.	0.0909090936
NDVI	0.1538461447

5η επανάληψη (7’ η εφαρμογή του μοντέλου στα δεδομένα)

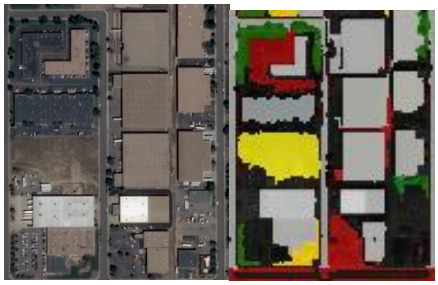
Στον Πίνακας 3.88 εμφανίζονται τα γνωρίσματα στα οποία βασίστηκε το μοντέλο της πέμπτης επανάληψης.

ΠΙΝΑΚΑΣ 3.88: ΓΝΩΡΙΣΜΑΤΑ ΣΤΑ ΠΛΑΙΣΙΑ ΤΗΣ 5ΗΣ ΕΠΑΝΑΛΗΨΗΣ

Feature
Mean Layer 4
HSI Transformation Intensity(R=Layer 3,G=Layer 2,B=Layer 1)
Number of pixels
GLCM Contrast (0ฐ)
GLCM Dissimilarity (all dir.)
GLCM Dissimilarity (0ฐ)
GLDV Ang. 2nd moment (45ฐ)
GLCM Correlation (all dir.)
Area
Length/Width
Elliptic fit
Shape index
MSAVI2
Length
Asymmetry
Border index
GLCM Entropy (90ฐ)
GLCM Mean (all dir.)
GLCM Mean (0ฐ)
Standard deviation Layer 3
Standard deviation Layer 4
GLCM Contrast (all dir.)
GLCM Contrast (45ฐ)
Compactness
GLCM Homogeneity (all dir.)
GLCM Correlation (45ฐ)
GLCM Homogeneity (90ฐ)
GLCM Entropy (all dir.)
Standard deviation Layer 1
Standard deviation Layer 2
GLCM Homogeneity (135ฐ)
Max. diff.
GLCM Homogeneity (45ฐ)
NDVI
GLCM Entropy (0ฐ)

Στην ακόλουθη Εικόνα (Εικόνα 3.217) εμφανίζεται το αποτέλεσμα εφαρμογής του μοντέλου των τυχαίων δασών στα πλαίσια της πέμπτης επανάληψης.

ΕΙΚΟΝΑ 3.217: ΑΠΟΤΕΛΕΣΜΑ ΕΦΑΡΜΟΓΗΣ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΓΙΑ ΤΗΝ ΠΕΜΠΤΗ ΕΠΑΝΑΛΗΨΗ



ΕΙΚΟΝΑ 3.218: *ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΗΝ ΠΕΜΠΤΗ ΕΠΑΝΑΛΗΨΗ*

Βάσει της Εικόνα 3.218 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.89, Πίνακας 3.90)

ΠΙΝΑΚΑΣ 3.89*: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (5Η ΕΠΑΝΑΛΗΨΗ)*

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	14	2	2

Πίνακας 3.90: Αξιολόγηση αποτελέσματος ταξινόμησης των τυχαίων δασών στην εικόνα του Colorado (δείκτες ποιότητας) (5η επανάληψη).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	87.50%	87.50%	77.78%	12.50%	12.50%

Στον ακόλουθο Πίνακα (Πίνακας 3.91) εμφανίζονται οι τιμές της σημαντικότητας για κάθε μία από τις μεταβλητές της πέμπτης επανάληψης.

ΠΙΝΑΚΑΣ 3.91: ΤΙΜΕΣ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ ΜΕΤΑΒΛΗΤΩΝ ΓΙΑ ΤΗΝ ΠΕΜΠΤΗ ΕΠΑΝΑΛΗΨΗ

Feature	**Importance**
Mean Layer 4	0.0172413792
HSI Transformation Intensity(R=Layer 3,G=Layer 2,B=Layer 1)	0
Number of pixels	0.0086206896
GLCM Contrast (0ฐ)	0.0172413792
GLCM Dissimilarity (all dir.)	0.0172413792
GLCM Dissimilarity (0ฐ)	0.0172413792
GLDV Ang. 2nd moment (45ฐ)	0.0172413792

GLCM Correlation (all dir.)	0.0172413792
Area	0.0172413792
Length/Width	0.0172413792
Elliptic fit	0.0172413792
Shape index	0.0172413792
MSAVI2	0.0172413792
Length	0.0172413792
Asymmetry	0.0172413792
Border index	0.0172413792
GLCM Entropy (90ş)	0.0172413792
GLCM Mean (all dir.)	0.0172413792
GLCM Mean (0ş)	0.0172413792
Standard deviation Layer 3	0.0258620679
Standard deviation Layer 4	0.0258620679
GLCM Contrast (all dir.)	0.0258620679
GLCM Contrast (45ş)	0.0258620679
Compactness	0.0258620679
GLCM Homogeneity (all dir.)	0.0344827585
GLCM Correlation (45ş)	0.0344827585
GLCM Homogeneity (90ş)	0.043103449
GLCM Entropy (all dir.)	0.043103449
Standard deviation Layer 1	0.0517241359
Standard deviation Layer 2	0.0517241359
GLCM Homogeneity (135ş)	0.0517241359
Max. diff.	0.0603448264
GLCM Homogeneity (45ş)	0.0603448264
NDVI	0.068965517
GLCM Entropy (0ş)	0.068965517

6η επανάληψη (6’ η εφαρμογή του μοντέλου στα δεδομένα)

Στον Πίνακας 3.92 εμφανίζονται τα γνωρίσματα στα οποία βασίστηκε το μοντέλο της έκτης επανάληψης.

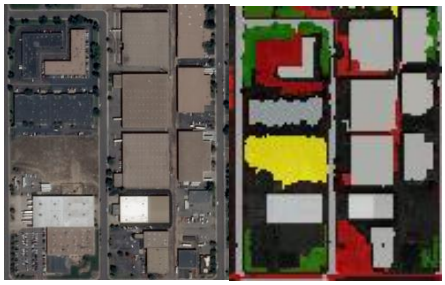
ΠΙΝΑΚΑΣ 3.92: ΓΝΩΡΙΣΜΑΤΑ ΣΤΑ ΠΛΑΙΣΙΑ ΤΗΣ 6ΗΣ ΕΠΑΝΑΛΗΨΗΣ

Feature
Max. diff.
Area
Length
GLCM Correlation (45ş)
Compactness
MSAVI2
GLCM Entropy (all dir.)
GLCM Entropy (90ş)
GLCM Mean (all dir.)
GLCM Mean (0ş)
GLCM Homogeneity (90ş)

Shape index
GLCM Correlation (all dir.)
Asymmetry
Standard deviation Layer 4
Elliptic fit
Length/Width
GLCM Entropy (0ฐ)
GLCM Homogeneity (135ฐ)
Border index
Standard deviation Layer 2
Standard deviation Layer 1
GLCM Contrast (all dir.)
GLCM Contrast (45ฐ)
Standard deviation Layer 3
GLCM Homogeneity (all dir.)
GLCM Homogeneity (45ฐ)
NDVI

Στην ακόλουθη Εικόνα (Εικόνα 3.219) εμφανίζεται το αποτέλεσμα εφαρμογής του μοντέλου των τυχαίων δασών στα πλαίσια της έκτης επανάληψης.

ΕΙΚΟΝΑ 3.219: ΑΠΟΤΕΛΕΣΜΑ ΕΦΑΡΜΟΓΗΣ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΓΙΑ ΤΗΝ ΕΚΤΗ ΕΠΑΝΑΛΗΨΗ



ΕΙΚΟΝΑ 3.220: *ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΗΝ ΕΚΤΗ ΕΠΑΝΑΛΗΨΗ*

Βάσει της Εικόνα 3.218 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.93, Πίνακας 3.94)

ΠΙΝΑΚΑΣ 3.93*: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (6Η ΕΠΑΝΑΛΗΨΗ)*

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	13	2	3

ΠΙΝΑΚΑΣ 3.94: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (6Η ΕΠΑΝΑΛΗΨΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	81.25%	86.67%	72.22%	18.75%	12.50%

Στον ακόλουθο Πίνακα (Πίνακας 3.95) εμφανίζονται οι τιμές της σημαντικότητας για κάθε μία από τις μεταβλητές της έκτης επανάληψης.

ΠΙΝΑΚΑΣ 3.95: ΤΙΜΕΣ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ ΜΕΤΑΒΛΗΤΩΝ ΓΙΑ ΤΗΝ ΕΚΤΗ ΕΠΑΝΑΛΗΨΗ

Feature	Importance
Max. diff.	0.0275650844
Area	0.0245022979
Length	0.0260336921
GLCM Correlation (45ฐ)	0.0275650844
Compactness	0.0275650844
MSAVI2	0.0275650844
GLCM Entropy (all dir.)	0.0275650844
GLCM Entropy (90ฐ)	0.0275650844
GLCM Mean (all dir.)	0.0275650844
GLCM Mean (0ฐ)	0.0275650844
GLCM Homogeneity (90ฐ)	0.0290964786
Shape index	0.0290964786
GLCM Correlation (all dir.)	0.0306278728
Asymmetry	0.0306278728
Standard deviation Layer 4	0.0321592651
Elliptic fit	0.0336906612
Length/Width	0.0352220535
GLCM Entropy (0ฐ)	0.0382848419
GLCM Homogeneity (135ฐ)	0.0398162343
Border index	0.0413476266
Standard deviation Layer 2	0.0428790227
Standard deviation Layer 1	0.044410415
GLCM Contrast (all dir.)	0.044410415
GLCM Contrast (45ฐ)	0.044410415
Standard deviation Layer 3	0.0474732034
GLCM Homogeneity (all dir.)	0.0474732034
GLCM Homogeneity (45ฐ)	0.0474732034
NDVI	0.0704441071

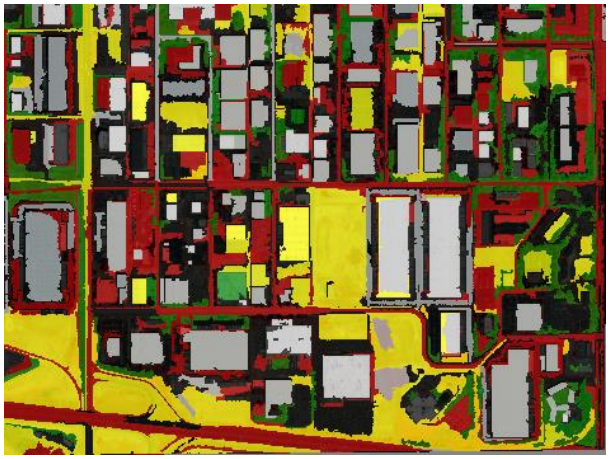
7η επανάληψη (5’ η εφαρμογή του μοντέλου στα δεδομένα)

Στον Πίνακας 3.96εμφανίζονται τα γνωρίσματα στα οποία βασίστηκε το μοντέλο της έβδομης επανάληψης.

ΠΙΝΑΚΑΣ 3.96: ΓΝΩΡΙΣΜΑΤΑ ΣΤΑ ΠΛΑΙΣΙΑ ΤΗΣ 7ΗΣ ΕΠΑΝΑΛΗΨΗΣ

Feature
Standard deviation Layer 1
Border index
Length/Width
Elliptic fit
MSAVI2
GLCM Mean (all dir.)
GLCM Entropy (0ฐ)
Shape index
Asymmetry
GLCM Contrast (all dir.)
GLCM Correlation (all dir.)
Compactness
GLCM Contrast (45ฐ)
GLCM Homogeneity (90ฐ)
GLCM Homogeneity (135ฐ)
Standard deviation Layer 4
NDVI
GLCM Correlation (45ฐ)
Standard deviation Layer 2
Standard deviation Layer 3
GLCM Homogeneity (45ฐ)
GLCM Homogeneity (all dir.)

Στην ακόλουθη Εικόνα (Εικόνα 3.221) εμφανίζεται το αποτέλεσμα εφαρμογής του μοντέλου των τυχαίων δασών στα πλαίσια της έβδομης επανάληψης.



ΕΙΚΟΝΑ 3.221: ΑΠΟΤΕΛΕΣΜΑ ΕΦΑΡΜΟΓΗΣ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΓΙΑ ΤΗΝ ΕΒΔΟΜΗ ΕΠΑΝΑΛΗΨΗ

ΕΙΚΟΝΑ 3.222: ***ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΗΝ ΕΒΔΟΜΗ ΕΠΑΝΑΛΗΨΗ***

Βάσει της Εικόνα 3.222 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.97, Πίνακας 3.98)

ΠΙΝΑΚΑΣ 3.97: ***ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (7Η ΕΠΑΝΑΛΗΨΗ)***

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	12	2	4

ΠΙΝΑΚΑΣ 3.98: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (7Η ΕΠΑΝΑΛΗΨΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	75.00%	85.71%	66.67%	25.00%	12.50%

Στον ακόλουθο Πίνακα (Πίνακας 3.99) εμφανίζονται οι τιμές της σημαντικότητας για κάθε μία από τις μεταβλητές της έβδομης επανάληψης.

ΠΙΝΑΚΑΣ 3.99: ΤΙΜΕΣ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ ΜΕΤΑΒΛΗΤΩΝ ΓΙΑ ΤΗΝ ΕΒΔΟΜΗ ΕΠΑΝΑΛΗΨΗ

Feature	**Importance**
Standard deviation Layer 1	0.0256410278
Border index	0.021367522
Length/Width	0.0256410278
Elliptic fit	0.0256410278
MSAVI2	0.0256410278
GLCM Mean (all dir.)	0.0256410278
GLCM Entropy (0ş)	0.0299145319
Shape index	0.0341880359
Asymmetry	0.0341880359
GLCM Contrast (all dir.)	0.0384615399
GLCM Correlation (all dir.)	0.0384615399
Compactness	0.0384615399
GLCM Contrast (45ş)	0.0427350439
GLCM Homogeneity (90ş)	0.0470085479
GLCM Homogeneity (135ş)	0.0470085479
Standard deviation Layer 4	0.0512820557
NDVI	0.0598290637
GLCM Correlation (45ş)	0.0598290637
Standard deviation Layer 2	0.0683760718

Standard deviation Layer 3	0.0769230798
GLCM Homogeneity (45ฐ)	0.0769230798
GLCM Homogeneity (all dir.)	0.1068376154

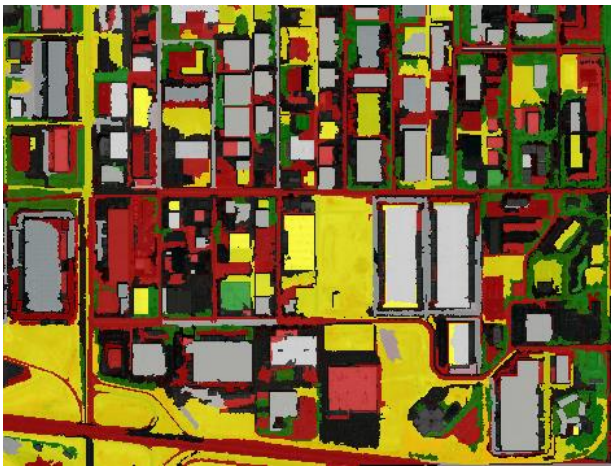
8η επανάληψη (5’ η εφαρμογή του μοντέλου στα δεδομένα)

Στον Πίνακας 3.100 εμφανίζονται τα γνωρίσματα στα οποία βασίστηκε το μοντέλο της όγδοης επανάληψης.

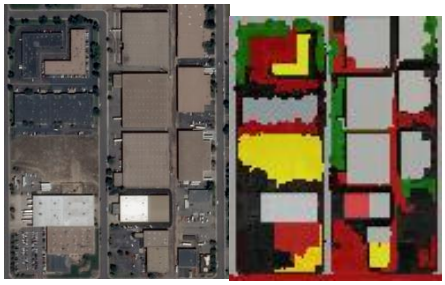
ΠΙΝΑΚΑΣ 3.100: ΓΝΩΡΙΣΜΑΤΑ ΣΤΑ ΠΛΑΙΣΙΑ ΤΗΣ 8ΗΣ ΕΠΑΝΑΛΗΨΗΣ

Feature
Standard deviation Layer 2
GLCM Contrast (45ฐ)
Compactness
MSAVI2
Asymmetry
GLCM Mean (all dir.)
GLCM Homogeneity (135ฐ)
GLCM Correlation (all dir.)
GLCM Correlation (45ฐ)
Shape index
Standard deviation Layer 3
Standard deviation Layer 4
GLCM Homogeneity (all dir.)
GLCM Entropy (0ฐ)
GLCM Homogeneity (45ฐ)
GLCM Contrast (all dir.)
NDVI
GLCM Homogeneity (90ฐ)

Στην ακόλουθη Εικόνα (Εικόνα 3.223) εμφανίζεται το αποτέλεσμα εφαρμογής του μοντέλου των τυχαίων δασών στα πλαίσια της όγδοης επανάληψης.



ΕΙΚΟΝΑ 3.223: ΑΠΟΤΕΛΕΣΜΑ ΕΦΑΡΜΟΓΗΣ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΓΙΑ ΤΗΝ ΟΓΔΟΗ ΕΠΑΝΑΛΗΨΗ



ΕΙΚΟΝΑ 3.224: ***ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΗΝ ΟΓΔΟΗ ΕΠΑΝΑΛΗΨΗ***

Βάσει της Εικόνα 3.224 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.101, Πίνακας 3.102)

ΠΙΝΑΚΑΣ 3.101*: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (8Η ΕΠΑΝΑΛΗΨΗ)*

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	10	2	6

ΠΙΝΑΚΑΣ 3.102: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (8Η ΕΠΑΝΑΛΗΨΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	62.50%	83.33%	55.56%	37.50%	12.50%

Στον ακόλουθο Πίνακα (Πίνακας 3.103) εμφανίζονται οι τιμές της σημαντικότητας για κάθε μία από τις μεταβλητές της όγδοης επανάληψης.

ΠΙΝΑΚΑΣ 3.103: ΤΙΜΕΣ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ ΜΕΤΑΒΛΗΤΩΝ ΓΙΑ ΤΗΝ ΟΓΔΟΗ ΕΠΑΝΑΛΗΨΗ

Feature	**Importance**
Standard deviation Layer 2	0.0273972601
GLCM Contrast (45ş)	0.01369863
Compactness	0.0273972601
MSAVI2	0.0273972601
Asymmetry	0.0273972601
GLCM Mean (all dir.)	0.0273972601
GLCM Homogeneity (135ş)	0.0410958901
GLCM Correlation (all dir.)	0.0410958901
GLCM Correlation (45ş)	0.0410958901
Shape index	0.0410958901
Standard deviation Layer 3	0.0547945201
Standard deviation Layer 4	0.0547945201
GLCM Homogeneity (all dir.)	0.0547945201
GLCM Entropy (0ş)	0.0547945201
GLCM Homogeneity (45ş)	0.0684931502
GLCM Contrast (all dir.)	0.1232876703
NDVI	0.1369863003
GLCM Homogeneity (90ş)	0.1369863003

9η επανάληψη (5’ η εφαρμογή του μοντέλου στα δεδομένα)

Στον Πίνακας 3.104 εμφανίζονται τα γνωρίσματα στα οποία βασίστηκε το μοντέλο της ένατης επανάληψης.

ΠΙΝΑΚΑΣ 3.104: ΓΝΩΡΙΣΜΑΤΑ ΣΤΑ ΠΛΑΙΣΙΑ ΤΗΣ 9ΗΣ ΕΠΑΝΑΛΗΨΗΣ

Feature
Standard deviation Layer 3
Standard deviation Layer 4
NDVI
GLCM Homogeneity (all dir.)
GLCM Homogeneity (45ᵹ)
GLCM Homogeneity (90ᵹ)
GLCM Homogeneity (135ᵹ)
GLCM Contrast (all dir.)
GLCM Correlation (all dir.)
GLCM Correlation (45ᵹ)
Asymmetry
GLCM Entropy (0ᵹ)
GLCM Mean (all dir.)
Shape index

Στην ακόλουθη Εικόνα (Εικόνα 3.225) εμφανίζεται το αποτέλεσμα εφαρμογής του μοντέλου των τυχαίων δασών στα πλαίσια της ένατης επανάληψης.

ΕΙΚΟΝΑ 3.225: ΑΠΟΤΕΛΕΣΜΑ ΕΦΑΡΜΟΓΗΣ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΓΙΑ ΤΗΝ ΕΝΑΤΗ ΕΠΑΝΑΛΗΨΗ

ΕΙΚΟΝΑ 3.226: *ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΗΝ ΕΝΑΤΗ ΕΠΑΝΑΛΗΨΗ*

Βάσει της Εικόνα 3.226 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.101, Πίνακας 3.102)

ΠΙΝΑΚΑΣ 3.105*: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (9Η ΕΠΑΝΑΛΗΨΗ)*

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	12	2	4

ΠΙΝΑΚΑΣ 3.106: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (9Η ΕΠΑΝΑΛΗΨΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	75.00%	85.71%	66.67%	25.00%	12.50%

Στον ακόλουθο Πίνακα (Πίνακας 3.107)εμφανίζονται οι τιμές της σημαντικότητας για κάθε μία από τις μεταβλητές της ένατης επανάληψης.

ΠΙΝΑΚΑΣ 3.107: ΤΙΜΕΣ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ ΜΕΤΑΒΛΗΤΩΝ ΓΙΑ ΤΗΝ ΕΝΑΤΗ ΕΠΑΝΑΛΗΨΗ

Feature	Importance
Standard deviation Layer 3	0
Standard deviation Layer 4	0
NDVI	0.2666666806
GLCM Homogeneity (all dir.)	0.4000000358
GLCM Homogeneity (45ฐ)	0
GLCM Homogeneity (90ฐ)	0.0666666701
GLCM Homogeneity (135ฐ)	0.2666666806
GLCM Contrast (all dir.)	0
GLCM Correlation (all dir.)	0
GLCM Correlation (45ฐ)	0
Asymmetry	0
GLCM Entropy (0ฐ)	0
GLCM Mean (all dir.)	0
Shape index	0

10η επανάληψη (5’ η εφαρμογή του μοντέλου στα δεδομένα)

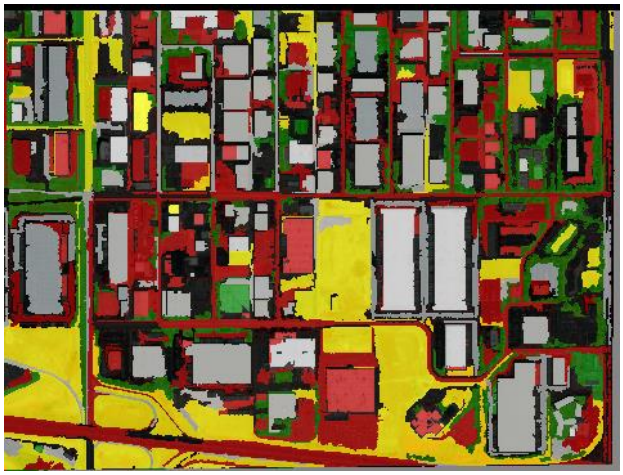
Στον Πίνακας 3.108 εμφανίζονται τα γνωρίσματα στα οποία βασίστηκε το μοντέλο της δέκατης επανάληψης.

ΠΙΝΑΚΑΣ 3.108: ΓΝΩΡΙΣΜΑΤΑ ΣΤΑ ΠΛΑΙΣΙΑ ΤΗΣ 10ΗΣ ΕΠΑΝΑΛΗΨΗΣ

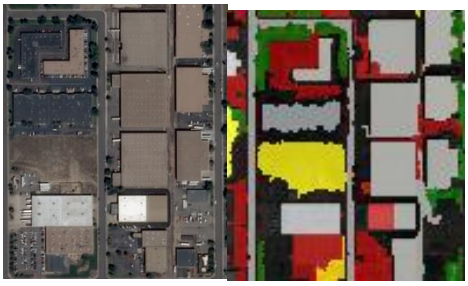
Feature
Standard deviation Layer 3
GLCM Correlation (all dir.)
GLCM Correlation (45ฐ)
GLCM Mean (all dir.)
GLCM Homogeneity (90ฐ)
Standard deviation Layer 4
GLCM Entropy (0ฐ)
GLCM Homogeneity (all dir.)
GLCM Homogeneity (135ฐ)

GLCM Contrast (all dir.)
NDVI

Στην ακόλουθη Εικόνα (Εικόνα 3.227) εμφανίζεται το αποτέλεσμα εφαρμογής του μοντέλου των τυχαίων δασών στα πλαίσια της δέκατης επανάληψης.



ΕΙΚΟΝΑ 3.227: ΑΠΟΤΕΛΕΣΜΑ ΕΦΑΡΜΟΓΗΣ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΓΙΑ ΤΗ ΔΕΚΑΤΗ ΕΠΑΝΑΛΗΨΗ



ΕΙΚΟΝΑ 3.228: *ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ ΑΠΟΣΠΑΣΜΑ ΑΣΤΙΚΗΣ ΔΟΜΗΣΗΣ ΑΠΟ ΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ ΓΙΑ ΤΗΝ ΔΕΚΑΤΗ ΕΠΑΝΑΛΗΨΗ*

Βάσει της Εικόνα 3.228 υπολογίστηκαν οι δείκτες ποιότητας που εμφανίζονται στους ακόλουθους Πίνακες (Πίνακας 3.109, Πίνακας 3.110)

ΠΙΝΑΚΑΣ 3.109*: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (10Η ΕΠΑΝΑΛΗΨΗ)*

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	11	2	5

ΠΙΝΑΚΑΣ 3.110: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (10Η ΕΠΑΝΑΛΗΨΗ).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	68.75%	84.62%	61.11%	31.25%	12.50%

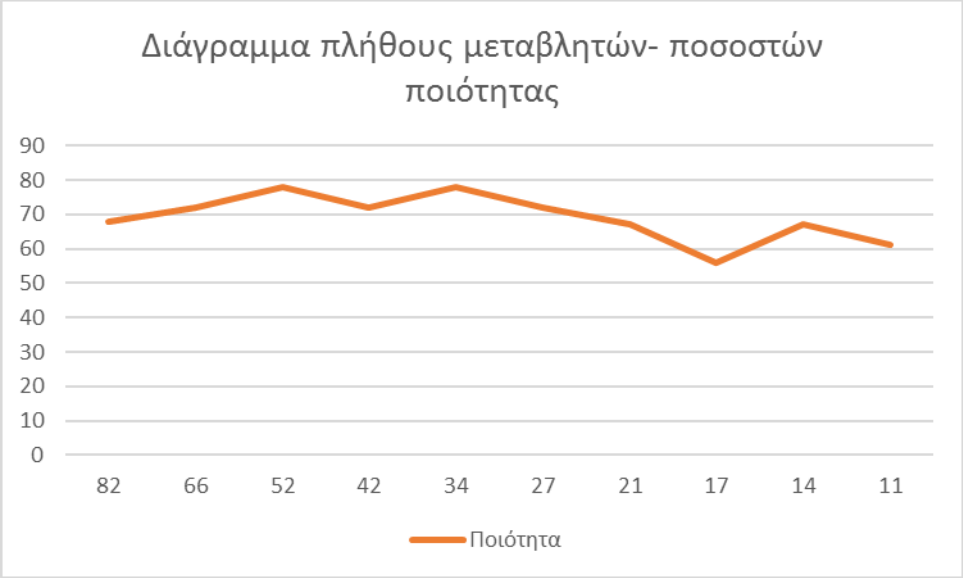
Στον ακόλουθο Πίνακα (Πίνακας 3.111)εμφανίζονται οι τιμές της σημαντικότητας για κάθε μία από τις μεταβλητές της δέκατης επανάληψης.

ΠΙΝΑΚΑΣ 3.111: ΤΙΜΕΣ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ ΜΕΤΑΒΛΗΤΩΝ ΓΙΑ ΤΗ ΔΕΚΑΤΗ ΕΠΑΝΑΛΗΨΗ

Feature	Importance
Standard deviation Layer 3	0.0632411093
GLCM Correlation (all dir.)	0.0671936795
GLCM Correlation (45ฐ)	0.0711462498
GLCM Mean (all dir.)	0.0711462498
GLCM Homogeneity (90ฐ)	0.07509882
Standard deviation Layer 4	0.0790513903
GLCM Entropy (0ฐ)	0.0830039531
GLCM Homogeneity (all dir.)	0.0869565234
GLCM Homogeneity (135ฐ)	0.0988142341
GLCM Contrast (all dir.)	0.1264822185
NDVI	0.1778656244

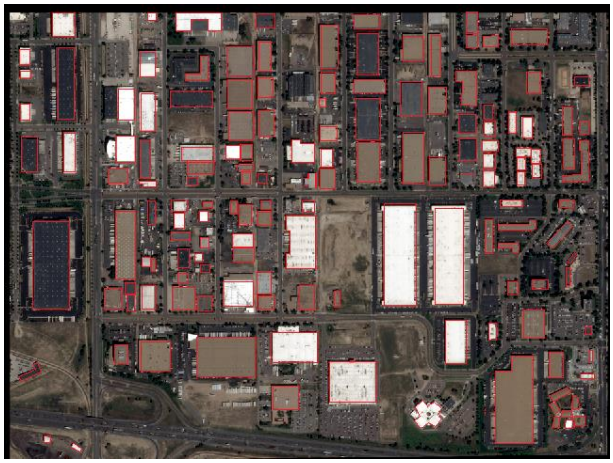
3.12.1Τελική επιλογή των χαρακτηριστικών

Στο ακόλουθο διάγραμμα (Εικόνα 4.21) εμφανίζονται τα ποσοστά ποιότητας συναρτήσει του πλήθους των γνωρισμάτων. Παρατηρείται πως το ποσοστό ποιότητας μεγιστοποιείται όταν το πλήθος των γνωρισμάτων παίρνει τις τιμές 52 και 34. Συνεπώς, επιλέγεται η τιμή 34 για λόγους υπολογιστικούς κόστους.



ΕΙΚΟΝΑ 3.229: ΔΙΑΓΡΑΜΜΑ ΠΛΗΘΟΥΣ ΓΝΩΡΙΣΜΑΤΩΝ - ΠΟΣΟΣΤΩΝ ΠΟΙΟΤΗΤΑΣ

Στόχος των παραπάνω επαναλήψεων ήταν η επιλογή των χαρακτηριστικών εκείνων οι οποίες θα δώσουν τα υψηλότερα δυνατά ποσοστά ποιότητας. Για το σκοπό αυτό έγινε ψηφιοποίηση του συνόλου των κτιρίων στην περιοχή μελέτης (Εικόνα 3.230) και στη συνέχεια έγινε υπολογισμός των δεικτών ποιότητας για το μοντέλο με τα τελικά επιλεχθέντα χαρακτηριστικά (5η επανάληψη) (Πίνακας 3.112, Πίνακας 3.113).



ΕΙΚΟΝΑ 3.230: ΦΩΤΟΕΡΜΗΝΕΙΑ ΤΩΝ ΚΤΙΡΙΩΝ ΣΤΗΝ ΠΕΡΙΟΧΗ ΜΕΛΕΤΗΣ

ΠΙΝΑΚΑΣ 3.112*: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (ΤΕΛΙΚΗ ΕΠΙΛΟΓΗ ΤΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ)*

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Σύνολο εικόνας	110	30	47

ΠΙΝΑΚΑΣ 3.113: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (*ΤΕΛΙΚΗ ΕΠΙΛΟΓΗ ΤΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ*).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Σύνολο εικόνας	70,06%	78,57%	58,82%	29,94%	19,11%

Στους ακόλουθους Πίνακες (Πίνακας 3.114, Πίνακας 3.115) εμφανίζεται το αντίστοιχο αποτέλεσμα κατά τη διαδικασία επιλογής των τιμών των παραμέτρων (Ενότητα 3.7.10), όπου η δημιουργία του μοντέλου έγινε βάσει των ακόλουθων χαρακτηριστικών:

- Μέση τιμή φωτεινότητας για το κανάλι 3 (Mean Layer 3)
- Μέση τιμή φωτεινότητας για το κανάλι 4 (Mean Layer 4)
- Μέση τιμή φωτεινότητας για το κανάλι 1 (Mean Layer 1)
- Μέση τιμή φωτεινότητας για το κανάλι 2 (Mean Layer)
- Δείκτης βλάστησης (NDVI)
- Λόγος μήκος προς πλάτος του εκάστοτε αντικειμένου (Length/Width)
- Συμπαγότητα του αντικειμένου (Compactness)
- Ομοιότητα του σχήματος του αντικειμένου με το σχήμα του ορθογωνίου(Rectangular fit)
- Εμβαδόν αντικειμένου (Area)
- Μέγιστη διαφορά (Max. Diff.)
- Μέση φωτεινότητα του αντικειμένου (Brightness)

ΠΙΝΑΚΑΣ 3.114*: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (ΤΕΛΙΚΗ ΕΠΙΛΟΓΗ ΤΩΝ ΤΙΜΩΝ ΤΩΝ ΠΑΡΑΜΕΤΡΩΝ)*

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Σύνολο εικόνας	129	52	28

ΠΙΝΑΚΑΣ 3.115: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (*ΤΕΛΙΚΗ ΕΠΙΛΟΓΗ ΤΩΝ ΤΙΜΩΝ ΤΩΝ ΠΑΡΑΜΕΤΡΩΝ*).

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Σύνολο εικόνας	82,17%	71,27%	61,72%	17,83%	33,12%

Παρατηρείται πως η μεταβολή στα χαρακτηριστικά οδήγησε σε μείωση του ποσοστού της πληρότητας της ταξινόμησης σε ό,τι αφορά την ανίχνευση των κτιρίων. Παράλληλα, ωστόσο, παρατηρήθηκε πως το ποσοστό της ορθότητας αυξήθηκε κατά 7%.

Στα πλαίσια της συγκεκριμένης ενότητας καταβλήθηκαν προσπάθειες ώστε τα χαρακτηριστικά τα οποία τελικά θα επιλεχθούν να οδηγούν σε υψηλά ποσοστά τόσο πληρότητας όσο και ορθότητας. Για το σκοπό αυτό, έγινε προσεκτική παρατήρηση των γνωρίσματων όπως προέκυψαν στα πλαίσια της 5^{ης} επανάληψης και διερευνήθηκε η συσχέτιση αυτών ανά ζεύγη μέσω του εργαλείου Tools>2D Feature Space Plot. Μέσω της διαδικασίας αυτής έγινε αποκλεισμός από τη διαδικασία της ταξινόμησης ενός από τα χαρακτηριστικά κάθε ζεύγους στην περίπτωση που εκείνα εμφάνιζαν συσχέτιση μεγαλύτερη από 0,98. Επιπροσθέτως, έγινε προσθήκη των χαρακτηριστικών της ενότητας 3.12.1. Η τελική λίστα των χαρακτηριστικών όπως προέκυψαν έπειτα από δοκιμές εμφανίζονται στον Πίνακας 3.116. Το πλήθος των γνωρισμάτων που τελικά επιλέχθηκαν ήταν 32.

ΠΙΝΑΚΑΣ 3.116: ΤΕΛΙΚΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ

Features
NDVI
Mean Layer 3
Mean Layer 4
Elliptic fit
Length
Standard deviation Layer 3
Standard deviation Layer 1
Length/ Width
Density
Standard deviation Layer 4
Asymmetry
Mean Layer 2
Area
Mean Layer 1
Standard deviation Layer 2
GLCM Correlation (all dir)
Border Index
GLCM Homogeneity (90)
Compactness
GLCM Dissimilarity (all dir.)
GLCM Entropy (all dir)
GLCM Contrast (0)
Max diff
Shape Index
HIS Transformation Intensity
GLCM Homogeneity (135)

GLCM Contrast (all dir)
Rectangular fit
MSAVI2
GLDV Ang. 2nd moment (45)
GLCM Mean (all dir)

Στη συνέχεια, έγινε εκπαίδευση του αλγορίθμου των τυχαίων δασών βάσει των γνωρισμάτων του Πίνακας 3.116. Το αποτέλεσμα της εφαρμογής του μοντέλου που προέκυψε στα υπό μελέτη δορυφορικά δεδομένα εμφανίζεται στην Εικόνα 3.231.

ΕΙΚΟΝΑ 3.231: ΑΠΟΤΕΛΕΣΜΑ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΓΙΑ ΤΗΝ ΤΕΛΙΚΗ ΕΠΙΛΟΓΗ ΤΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

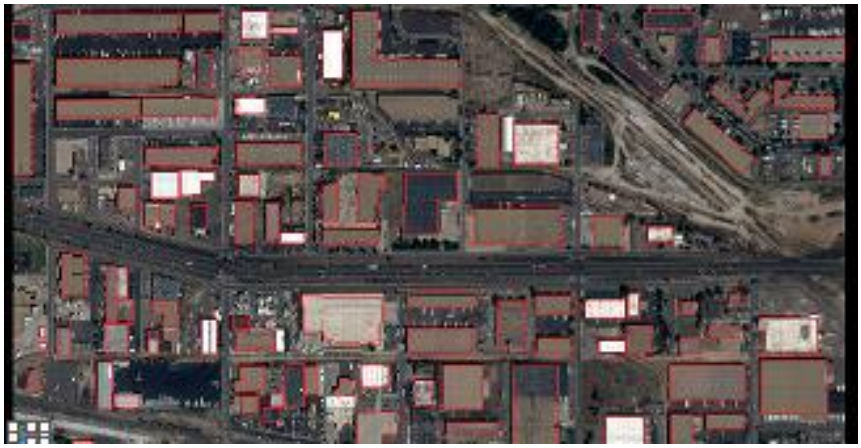
Εν συνεχεία, έγινε αξιολόγηση των επιδόσεων του τελικού μοντέλου για το σύνολο της εικόνας εισόδου βάσει της ψηφιοποιημένης εικόνας (Εικόνα 3.230). Στον ακόλουθο Πίνακα Πίνακας 3.117()εμφανίζονται τα αποτελέσματα της αξιολόγησης της ποιότητας του αλγορίθμου για την αρχικά επιλεγμένη λίστα των χαρακτηριστικών (ενότητα 3.7.10), για την 5η επανάληψη καθώς και για τα χαρακτηριστικά του Πίνακας 3.116. Βάσει του κριτηρίου της ποιότητας παρατηρείται πως η τελική επιλογή των γνωρισμάτων δίνει τα υψηλότερα ποσοστά και μάλιστα με διαφορά περίπου 6% και 9% από την αντίστοιχα για την αρχικά επιλεγμένη λίστα καθώς και εκείνη της 5ης επανάληψης. Το ίδιο ισχύει και σε ό,τι αφορά την πληρότητα του παραγόμενου αποτελέσματος όπου το ποσοστό της τελικής λίστας των χαρακτηριστικών είναι 84% ενώ εκείνο των άλλων δύο δοκιμών 82% και 70%. Το υψηλότερο ποσοστό ορθότητας εμφανίζεται στην περίπτωση της πέμπτης επανάληψης (79%), ωστόσο η διαφορά του από την τελική επιλογή των χαρακτηριστικών είναι αμελητέα (περίπου 1%). Βάσει των παραπάνω προκύπτει πως τα γνωρίσματα του Πίνακας 3.116 εμφανίζουν τις καλύτερες επιδόσεις και συνεπώς είναι εκείνα που τελικά επιλέγονται.

ΠΙΝΑΚΑΣ 3.117: ΣΥΓΚΡΙΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ ΤΩΝ ΕΠΙΔΟΣΕΩΝ ΤΩΝ ΤΡΙΩΝ ΜΟΝΤΕΛΩΝ

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Αρχική λίστα χαρακτηριστικών (ενότητα 3.7.10)	82,17%	71,27%	61,72%	17,83%	33,12%
Χαρακτηριστικά 5ης επανάληψης (ενότητα 0)	70,06%	**78,57%**	58,82%	29,94%	19,11%
Τελική επιλογή χαρακτηριστικών	**84,08%**	77,65%	**67,69%**	15,92%	24,20%

Ποσοτική αξιολόγηση αποτελεσμάτων

Στα πλαίσια της παρούσας αξιολόγησης έγινε ψηφιοποίηση των κτιρίων στο δεύτερο τμήμα της εικόνας (Εικόνα 3.232). Όπως έχει ήδη διευκρινιστεί στόχος της παρούσας εργασίας είναι αξιολόγηση των επιδόσεων του αλγορίθμου των τυχαίων δασών σε ό,τι αφορά την ανίχνευση κτιρίων. Παρατηρείται πως οι επιδόσεις του μοντέλου είναι υψηλές και μάλιστα παρόμοιες με εκείνες της εφαρμογής του αντίστοιχου στο πρώτο τμήμα της εικόνας (Πίνακας 3.118, Πίνακας 3.119). Αναλυτικά, το ποσοστό της πληρότητας είναι υψηλό και στο συγκεκριμένο τμήμα και φτάνει το 82,24%. Παρόμοια είναι τα αποτελέσματα και σε ό,τι αφορά την ικανοποίηση του κριτηρίου της ορθότητας, η οποία φτάνει το 80% Βάσει αυτών προκύπτει πως η προτεινόμενη μεθοδολογία δίνει ικανοποιητικά αποτελέσματα στην ανίχνευση κτιρίων και συνεπώς είναι και εκείνη η οποία τελικά επιλέγεται και προτείνεται σε παρόμοιες εφαρμογές.



ΕΙΚΟΝΑ 3.232: ΦΩΤΟΕΡΜΗΝΕΙΑ ΤΟΥ ΔΕΥΤΕΡΟΥ ΤΜΗΜΑΤΟΣ ΤΗΣ ΕΙΚΟΝΑΣ ΕΙΣΟΔΟΥ

ΠΙΝΑΚΑΣ 3.118*: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΑΡΙΘΜΟΣ ΔΙΑΝΥΣΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ) (2^{Ο} ΤΜΗΜΑ ΤΗΣ ΕΙΚΟΝΑΣ)*

	Πλήθος TP	Πλήθος FP	Πλήθος FN
Απόσπασμα εικόνας	88	22	19

ΠΙΝΑΚΑΣ 3.119*: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (2^{Ο} ΤΜΗΜΑ ΤΗΣ ΕΙΚΟΝΑΣ)*

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Απόσπασμα εικόνας	82,24%	80,00%	68,22%	17,76%	20,56%

3.13 Εφαρμογή σε δεύτερο τμήμα της εικόνας εισόδου

Στα πλαίσια της παρούσας ενότητας έγινε αξιολόγηση των συμπερασμάτων που εξήχθησαν στα προηγούμενα κεφάλαια. Αναλυτικά έγινε εφαρμογή των παραπάνω διαδικασιών σε νέο τμήμα της εικόνας- εισόδου ούτως ώστε να διευκρινιστεί κατά πόσον οι προτεινόμενες λύσεις που διατυπώθηκαν στα προαναφερθέντα χωρία είναι αποδοτικές σε διαφορετική περιοχή μελέτης. Οι εργασίες που εφαρμόστηκαν είναι πανομοιότυπες με εκείνες του τμήματος 1.

3.13.1 Προεπεξεργασία της εικόνας εισόδου

Αποκοπή τμήματος

Αρχικά, έγινε εισαγωγή της εικόνας της Commerce πόλης στο περιβάλλον του QGIS. Στη συνέχεια μέσω του εργαλείου έγινε αποκοπή και αποθήκευση του τμήματος που εμφανίζεται στην Εικόνα 3.233.

ΕΙΚΟΝΑ 3.233: ΤΜΗΜΑ 2 ΤΗΣ ΔΟΡΥΦΟΡΙΚΗΣ ΕΙΚΟΝΑΣ ΤΗΣ ΠΟΛΗΣ COMMERCE ΤΗΣ ΠΟΛΙΤΕΙΑΣ ΤΟΥ COLORADO

Φιλτράρισμα της εικόνας εισόδου

Εν συνεχεία έγινε εφαρμογή στο παραπάνω τμήμα του αμφίπλευρου φίλτρου (bilateral filter). Το παραπάνω υλοποιήθηκε μέσω εντολών στη γλώσσα προγραμματισμού Python. Στην Εικόνα 3.234 εμφανίζεται το αποτέλεσμα της εν λόγω διαδικασίας

ΕΙΚΟΝΑ 3.234: ΤΜΗΜΑ 2 ΤΗΣ ΔΟΡΥΦΟΡΙΚΗΣ ΕΙΚΟΝΑΣ ΕΠΕΙΤΑ ΑΠΟ ΕΦΑΡΜΟΓΗ ΣΕ ΑΥΤΗΝ ΤΟΥ ΑΜΦΙΠΛΕΥΡΟΥ ΦΙΛΤΡΟΥ

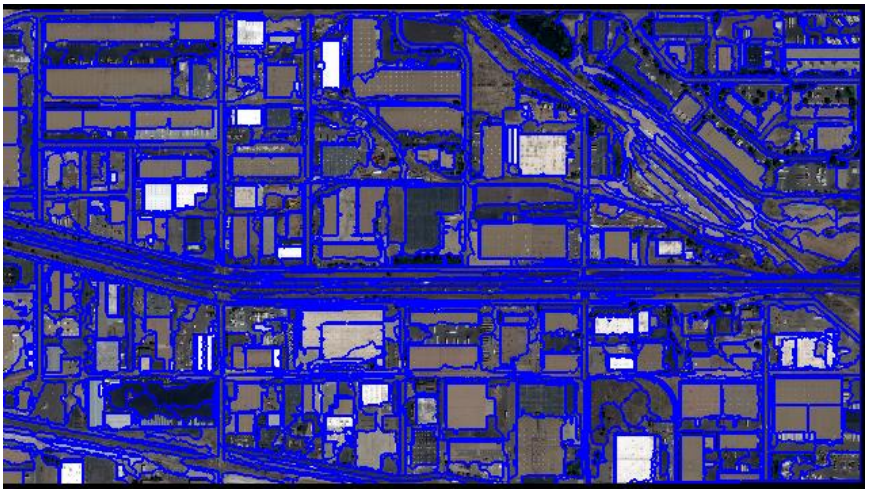
Κατάτμηση του δεύτερου τμήματος

Ομοίως με το προηγούμενο τμήμα της εικόνας έγινε κατάτμηση αυτού μέσω της πολυκλιμακωτής κατάτμησης. Οι τιμές των παραμέτρων της διαδικασίας αυτής είναι οι ακόλουθες:

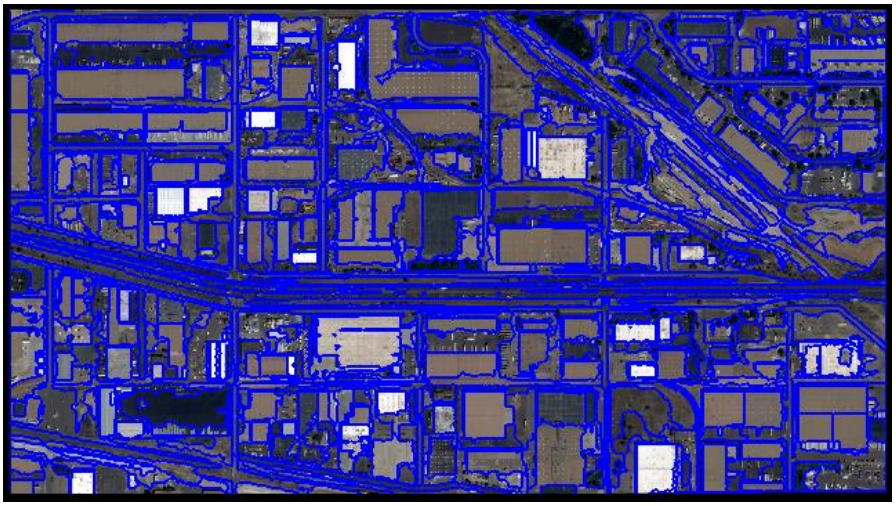
- Κλίμακα (Scale): 150
- Σχήμα (Shape): 0.4
- Προσέγγιση κανονικού σχήματος (Compactness): 0.3

Στη συνέχεια εφαρμόστηκε στο παραγόμενο αποτέλεσμα ο αλγόριθμος της φασματικής διαφοράς με τιμή για τη μέγιστή φασματική διαφορά ίση με 4.

Τα αποτελέσματα των παραπάνω διαδικασιών εμφανίζονται στις ακόλουθες εικόνες (Εικόνα 3.235, Εικόνα 3.236).



ΕΙΚΟΝΑ 3.235: ΕΠΙΠΕΔΟ 1 ΚΑΤΑΤΜΗΣΗΣ (ΠΟΛΙΚΛΙΜΑΚΩΤΗ ΚΑΤΑΤΜΗΣΗ)



ΕΙΚΟΝΑ 3.236: ΕΠΙΠΕΔΟ 2 ΚΑΤΑΤΜΗΣΗΣ (ΑΛΓΟΡΙΘΜΟΣ ΦΑΣΜΑΤΙΚΗΣ ΔΙΑΦΟΡΑΣ)

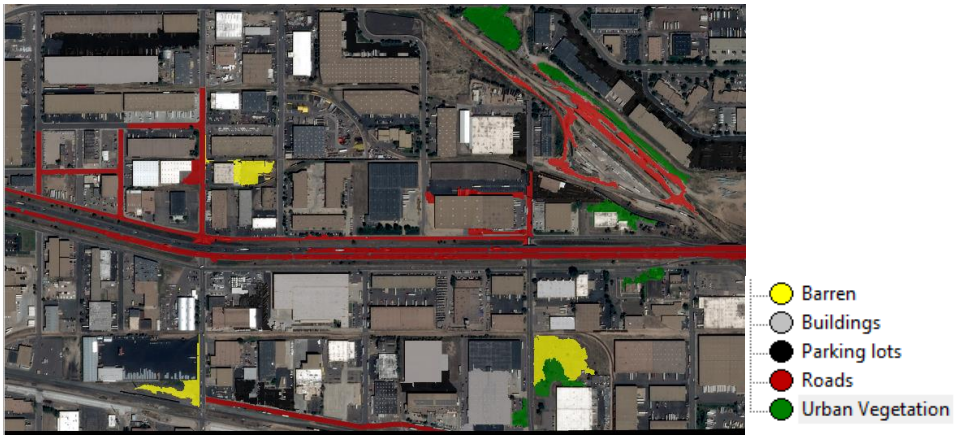
Ταξινόμηση της εικόνας εισόδου

Στόχος της παρούσας ενότητας είναι η ταξινόμηση των αντικειμένων της διαδικασίας 0 σε θεματικές κατηγορίες, μέσω του αλγορίθμου των τυχαίων δασών. Αναλυτικά, ορίστηκαν οι ακόλουθες κλάσεις:

- Κτίρια
- Δρόμοι

- Χώροι στάθμευσης
- Άγονο Έδαφος
- Αστικό πράσινο

Στην Εικόνα 3.237 εμφανίζονται τα δείγματα τα οποία δόθηκαν στο μοντέλο. Εν συνεχεία, ορίστηκαν τα γνωρίσματα των αντικειμένων βάσει των οποίων θα γίνει ο διαχωρισμός τους στις διαφορετικές κλάσεις. Ο καθορισμός των χαρακτηριστικών βασίστηκε στα συμπεράσματα τα οποία εξήχθησαν βάσει της μεθοδολογίας των (Diaz- Uriate and Alvarez de Andres, 2006). Αναλυτικά, τα χαρακτηριστικά τα οποία διατηρήθηκαν ήταν εκείνα της 5^{ης} επανάληψης (ενότητα 0). Τα συγκεκριμένα δίνουν το βέλτιστο δυνατό αποτέλεσμα με το μικρότερο υπολογιστικό κόστος (Πίνακας 3.120).



ΕΙΚΟΝΑ 3.237: ΔΕΙΓΜΑΤΑ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ

ΠΙΝΑΚΑΣ 3.120: ΓΝΩΡΙΣΜΑΤΑ ΤΕΛΙΚΗΣ ΕΠΙΛΟΓΗΣ

Features
NDVI
Mean Layer 3
Mean Layer 4
Elliptic fit
Length
Standard deviation Layer 3
Standard deviation Layer 1
Length/ Width
Density
Standard deviation Layer 4
Asymmetry
Mean Layer 2
Area
Mean Layer 1
Standard deviation Layer 2
GLCM Correlation (all dir)
Border Index
GLCM Homogeneity (90)
Compactness
GLCM Dissimilarity (all dir.)
GLCM Entropy (all dir)
GLCM Contrast (0)

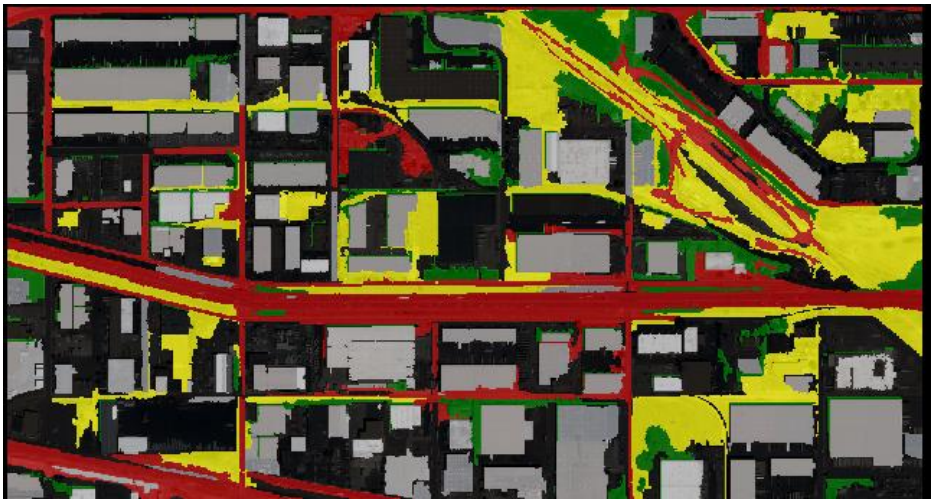
Max diff
Shape Index
HIS Transformation Intensity
GLCM Homogeneity (135)
GLCM Contrast (all dir)
Rectangular fit
MSAVI2
GLDV Ang. 2nd moment (45)
GLCM Mean (all dir)

Οι παράμετροι του αλγορίθμου των τυχαίων δασών ορίστηκαν ως εξής:

- Βάθος δέντρων: 10
- Ελάχιστος αριθμός δειγμάτων ανά κόμβο: 0
- Χρήση αντικαταστατών: Όχι
- Μέγιστος αριθμός κατηγοριών: 16
- Πλήθος ενεργών μεταβλητών: 3
- Πλήθος δέντρων: 50
- Ακρίβεια τυχαίου δάσους:
- Κριτήριο τερματισμού: Και τα δύο

Οι συγκεκριμένες δίνουν τα βέλτιστα αποτελέσματα βάσει των συμπερασμάτων της ενότητας 3.7.

Στην Εικόνα 3.238 εμφανίζονται το αποτέλεσμα εφαρμογής του αλγορίθμου των τυχαίων δασών σε τμήμα της πόλης Commerce.



ΕΙΚΟΝΑ 3.238: ΤΑΞΙΝΟΜΗΣΗ ΤΟΥ ΔΕΥΤΕΡΟΥ ΤΜΗΜΑΤΟΣ ΤΗΣ ΕΙΚΟΝΑΣ ΤΗΣ ΠΟΛΗΣ COMMERCE ΜΕΣΩ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ

4 Συμπεράσματα- Προοπτικές

4.1 Συμπεράσματα

Στόχος της παρούσας μεταπτυχιακής εργασίας ήταν η διερεύνηση της αποτελεσματικότητας των αλγορίθμων δέντρα απόφασης και τυχαία δάση στην ανίχνευση κτιρίων μέσω αντικειμενοστρεφούς ανάλυσης εικόνας. Το παραπάνω εγχείρημα υλοποιήθηκε πρακτικά στο λογισμικό του eCognition και τα τηλεπισκοπικά δεδομένα τα οποία αναλύθηκαν είναι υψηλής ευκρίνειας και προέρχονται από τους δορυφόρους Pléiades. Διευκρινίζεται πως στο λογισμικό του eCognition έχουν ενσωματωθεί οι αλγόριθμοι CART καθώς και Random Forest και πως στο συγκεκριμένο περιβάλλον δίνεται η δυνατότητα στο χρήστη να ρυθμίσει ένα σύνολο παραμέτρων σχετικά με τα παραγόμενα μοντέλα. Μέσω της παρούσας εργασίας εξετάστηκε κατά πόσον οι τελευταίες επηρεάζουν την ποιότητα του αποτελέσματος της ταξινόμησης. Για το λόγο αυτό, υλοποιήθηκε ένα σύνολο δοκιμών οι οποίες στόχο είχαν τον εντοπισμό των τιμών εκείνων οι οποίες θα δώσουν το βέλτιστο δυνατό αποτέλεσμα.

Ο αλγόριθμος των δέντρων απόφασης απαιτούσε από το χρήστη τη ρύθμιση των ακόλουθων παραμέτρων:

- Το βάθος του δέντρου
- Ο ελάχιστος αριθμός δειγμάτων ανά κόμβο
- Χρήση αντικαταστατών
- Μέγιστος αριθμός κατηγοριών
- Cross Validation folds
- Χρήση κανόνα 1-SE
- Αφαίρεση των κλαδεμένων κλαδιών

Τα συμπεράσματα τα οποία αντλήθηκαν βάσει των παραπάνω δοκιμών είναι πως η μεταβολή στην τιμή του βάθους επηρεάζει σε μεγάλο βαθμό το παραγόμενο αποτέλεσμα. Πιο συγκεκριμένα, η δημιουργία μικρών σε μέγεθος δέντρων επιδρά αρνητικά στην ποιότητα του παραγόμενου αποτελέσματος καθώς τα συγκεκριμένα δεν έχουν προσαρμοστεί στα δεδομένα εκπαίδευσης και ως εκ τούτου το μοντέλο το οποίο προκύπτει είναι περισσότερο απλοϊκό από το επιθυμητό. Από την άλλη πλευρά ο καθορισμός ενός μεγάλου σε μέγεθος βάθους οδηγεί σε μεγάλα σε μέγεθος δέντρα τα οποία είναι υπέρπροσαρμοσμένα στα δεδομένα εκπαίδευσης. Το βάθος του δέντρου το οποίο επιλέχθηκε έπειτα από ένα πλήθος δοκιμών είναι 5.

Ο ελάχιστος αριθμός δειγμάτων ανά κόμβο είναι άμεσα συσχετισμένος με το βάθος του δέντρου. Πιο συγκεκριμένα, η αύξηση του μεγέθους της συγκεκριμένης μεταβλητής οδήγησε σε μείωση του βάθους των δέντρων. Το παραπάνω οδηγεί σε υπόπροσαρμογή των δέντρων απόφασης στα δεδομένα εκπαίδευσης. Για το λόγο αυτό η τιμή, η οποία επιλέγεται είναι η μηδενική.

Η χρήση αντικαταστατών είναι μία πολύ σημαντική παράμετρος στην περίπτωση που ορισμένα από τα δεδομένα εκπαίδευσης δεν έχουν τιμές για κάποια χαρακτηριστικά τους. Στα πλαίσια της συγκεκριμένης δοκιμής, ωστόσο, δεν εμφανίζονται προβλήματα ελλιπών τιμών και ως εκ τούτου η ρύθμιση της συγκεκριμένης μεταβλητής δεν επέφερε μεταβολές στον παραγόμενο θεματικό χάρτη.

Η ρύθμιση του πλήθους του μέγιστου αριθμού κατηγοριών δεν άλλαξε το αποτέλεσμα της ταξινόμησης. Το παραπάνω οφείλεται στο γεγονός πως το πλήθος των θεματικών κατηγοριών ήταν εκ των προτέρων καθορισμένο από το χρήστη.

Το πλήθος των cross validations διαμορφώνει καθοριστικό ρόλο στον υπολογισμό του σφάλματος γενίκευσης καθώς και στη διαδικασία του κλαδέματος κόστους πολυπλοκότητας. Η ρύθμιση, ωστόσο, της εν λόγω παραμέτρου στα πλαίσια της παρούσας δοκιμής δε μετέβαλε το αποτέλεσμα της ταξινόμησης.

Το κλάδεμα μέσω του κανόνα 1-SE οδηγεί σε σταθερότητα του αλγορίθμου σε ό,τι αφορά την επιλογή του τελικού δέντρου απόφασης βάσει του κλαδέματος κόστους πολυπλοκότητας. Στην προκειμένη περίπτωση, ωστόσο, οδήγησε σε χαμηλής ποιότητας αποτελέσματα.

Τέλος, η παράμετρος αφαίρεση των κλαδεμένων κλαδιών παρέχει στο χρήστη τη δυνατότητα οπτικοποίησης των δέντρων απόφασης όπως έχουν διαμορφωθεί πριν εφαρμοστεί σε αυτά η διαδικασία του κλαδέματος. Η μεταβολή, ωστόσο, της τιμής της συγκεκριμένης σε Όχι δε μετέβαλε τον παραγόμενο θεματικό χάρτη.

Συνεπώς, οι τιμές των παραμέτρων του αλγορίθμου των δέντρων απόφασης, οι οποίες δίνουν τα βέλτιστα δυνατά αποτελέσματα σε ό,τι αφορά την ανίχνευση κτιρίων είναι οι ακόλουθες:

- Βάθος δέντρου (Depth): 5
- Ελάχιστος αριθμός δειγμάτων (Min sample count): 0
- Χρήση αντικαταστατών (Use surrogates): Όχι (No)
- Μέγιστος αριθμός κατηγοριών (Max categories): 16
- Cross Validation folds: 3
- Use 1 SE rule: Όχι (No)
- Αφαίρεση των κλαδεμένων κλαδιών (Truncate pruned trees): Ναι (Yes)

Οι δείκτες ποιότητας όπως προέκυψαν έπειτα από εφαρμογή του συγκεκριμένου μοντέλου εκπαίδευσης σε ό,τι αφορά την ανίχνευση κτιρίων εμφανίζονται στον Πίνακας 4.1

ΠΙΝΑΚΑΣ 4.1: ***ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ)***

Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
59,87%	81,03%	52,51%	40,13%	14,01%

Παρατηρείται πως τα παραπάνω ποσοστά είναι ικανοποιητικά σε ό,τι αφορά την ικανοποίηση του κριτηρίου της ορθότητας, η οποία συγκεντρώνει ποσοστό 81%. Ωστόσο τα αποτελέσματα έχουν αρκετά περιθώρια βελτίωσης σε ό,τι αφορά την πληρότητα καθώς η συγκεκριμένη συγκεντρώνει ποσοστό 60%.

Εφόσον ολοκληρώθηκε το σύνολο των δοκιμών για τον αλγόριθμο των δέντρων απόφασης ακολούθησε η εφαρμογή των τυχαίων δασών στα δεδομένα εκπαίδευσης:

- Βάθος δέντρων
- Ελάχιστος αριθμός δειγμάτων ανά κόμβο
- Χρήση αντικαταστατών
- Μέγιστος αριθμός κατηγοριών
- Πλήθος ενεργών μεταβλητών

- Πλήθος δέντρων
- Ακρίβεια τυχαίου δάσους
- Κριτήριο τερματισμού

Ομοίως με τον αλγόριθμο των δέντρων απόφασης η ρύθμιση της τιμής του βάθους διαδραματίζει καθοριστικό ρόλο στην ποιότητα του παραγόμενου θεματικού χάρτη. Πιο συγκεκριμένα, η επιλογή μικρής τιμής για τη συγκεκριμένη παράμετρο οδηγεί σε υποπροσαρμογή του αλγορίθμου στα δεδομένα. Το παραπάνω επιβεβαιώνεται από το γεγονός πως ο θεματικός χάρτης που προέκυψε για την τιμή 2 δεν περιέχει ορισμένες από τις εμφανιζόμενες κλάσεις. Το βάθος εκείνο που έδωσε τα βέλτιστα δυνατά αποτελέσματα είναι εκείνο που προέκυψε για την τιμή 4 για το λόγο αυτό επιλέγεται η συγκεκριμένη. Αξίζει να σημειωθεί πως ο θεματικός χάρτης που προέκυψε για την τιμή 5 είναι πανομοιότυπος με εκείνον της 4. Το εν λόγω μοντέλο, ωστόσο απορρίπτεται καθώς έχει μεγαλύτερο υπολογιστικό κόστος συγκριτικά με εκείνου της τιμής 4 και παράλληλα δε συνεισφέρει στη βελτίωση των ποσοστών ποιότητας.

Η αύξηση της τιμής της μεταβλητής «ελάχιστο πλήθος δειγμάτων ανά κόμβο» οδήγησε σε μείωση, όπως ήταν αναμενόμενο των ποσοστών ποιότητας των παραγόμενων αποτελεσμάτων ταξινόμησης. Για το λόγο αυτό, η τιμή που τελικά επιλέγεται είναι η μηδενική.

Η ρύθμιση της τιμής της παραμέτρου χρήση αντικαταστατών σε Ναι δε μετέβαλε την ποιότητα του παραγόμενου θεματικού χάρτη καθώς δεν υπήρχαν στα δεδομένα εκπαίδευσης ελλιπείς τιμές.

Η μεταβολή της τιμής του μέγιστου αριθμού κατηγοριών δε μετέβαλε το αποτέλεσμα της ταξινόμησης μέσω του αλγορίθμου των τυχαίων δασών.

Η παράμετρος του πλήθους των ενεργών μεταβλητών επηρεάζει την ποιότητα του αποτελέσματος της ταξινόμησης. Η προτεινόμενη από τους δημιουργούς του αλγορίθμου τιμή είναι η $\sqrt{p}$ όπου p το πλήθος των γνωρισμάτων. Στα πλαίσια της παρούσας διπλωματικής έγιναν πειράματα για διάφορες τιμές της συγκεκριμένης παραμέτρου. Οι τιμές εκείνες οι οποίες έδωσαν τα ποιοτικότερα αποτελέσματα είναι οι 3 (δηλαδή η προτεινόμενη) και η 5. Για λόγους οικονομίας υπολογιστικής μνήμης επιλέγεται η μικρότερη τιμή, δηλαδή η 3.

Το πλήθος των δέντρων σχετίζεται άμεσα με την ποιότητα του παραγόμενου αποτελέσματος. Αναλυτικά, ένας μικρός αριθμός οδηγεί σε ένα λιγότερο ποιοτικό θεματικό χάρτη και η ταξινόμηση στην περίπτωση αυτή είναι παρόμοια με εκείνη ενός μεμονωμένου ταξινόμητη. Από την άλλη πλευρά η δημιουργία πολλών δέντρων αυξάνει το υπολογιστικό κόστος και δε συνεισφέρει στην αύξηση των κριτηρίων ποιότητας. Η τιμή η οποία επιλέχθηκε ως η βέλτιστη δυνατή στην προκειμένη περίπτωση είναι η 50.

Τέλος, το κριτήριο τερματισμού δεν επηρέασε την ποιότητα του παραγόμενου αποτελέσματος.

Βάσει των παραπάνω προκύπτει πως οι τιμές των παραμέτρων του αλγορίθμου των τυχαίων δασών, οι οποίες δίνουν τα βέλτιστα δυνατά αποτελέσματα σε ό,τι αφορά την ανίχνευση κτιρίων είναι οι ακόλουθες:

- Βάθος δέντρων: 10

- Ελάχιστος αριθμός δειγμάτων ανά κόμβο: 0
- Χρήση αντικαταστατών: Όχι
- Μέγιστος αριθμός κατηγοριών: 16
- Πλήθος ενεργών μεταβλητών: 3
- Πλήθος δέντρων: 50
- Ακρίβεια τυχαίου δάσους:0,01
- Κριτήριο τερματισμού: Και τα δύο

Οι δείκτες ποιότητας όπως προέκυψαν έπειτα από εφαρμογή του συγκεκριμένου μοντέλου εκπαίδευσης σε ό,τι αφορά την ανίχνευση κτιρίων εμφανίζονται στον Πίνακας 4.2.

ΠΙΝΑΚΑΣ 4.2: ***ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ)***

Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
82,17%	71,27%	61,72%	17,83%	33,12%

Παρατηρείται πως τα παραπάνω ποσοστά είναι υψηλότερα συγκριτικά με εκείνα των δέντρων απόφασης. Το παραπάνω επιβεβαιώνει την υπόθεση πως η χρήση των συνδυαστικών ταξινομητών δίνουν εμφανώς ποιοτικότερα αποτελέσματα συγκριτικά με εκείνα των μεμονωμένων.

Στο δεύτερο μέρος της μελέτης έγινε διερεύνηση της χρησιμότητας των χαρακτηριστικών των αντικειμένων βάσει των οποίων θα γίνει η ταξινόμηση των αντικειμένων. Αναλυτικά, ο αλγόριθμος των τυχαίων δασών παρέχει τη δυνατότητα στο χρήστη να έχει εικόνα της σημαντικότητας των γνωρισμάτων στη διαδικασία της ταξινόμησης. Στα πλαίσια της παρούσας εφαρμογής έγινε αξιοποίηση της παραπάνω ιδιότητας μέσω της μεθόδου που πρότειναν οι (Diaz- Uriate and Alvarez de Andres, 2006). Προς την κατεύθυνση αυτή εφαρμόστηκε μία επαναληπτική διαδικασία σχηματισμού μοντέλων του αλγορίθμου των τυχαίων δασών όπου σε κάθε επανάληψη γινόταν παράλειψη του 20% των λιγότερο σημαντικών αντικειμένων. Το μοντέλο το οποίο τελικά επιλέχθηκε ήταν εκείνο το οποίο έδωσε τα βέλτιστα ποσοστά ποιότητας. Αναλυτικά, διατηρήθηκε εκείνο το οποίο βασιζόταν στα χαρακτηριστικά του Πίνακας 4.3. Παρατηρείται πως η πλειοψηφία των γνωρισμάτων ανήκουν στις κατηγορίες της υφής και του σχήματος και όχι του φάσματος. Τα ποσοστά της ποιότητας στην περίπτωση αυτή εμφανίζονται στον Πίνακας 4.4. Η μεταβολή στα χαρακτηριστικά οδήγησε σε βελτίωση των επιδόσεων του αλγορίθμου σε ό,τι αφορά την ικανοποίηση του κριτηρίου της ορθότητας (Πίνακας 4.3). Το παραπάνω, ωστόσο, δεν ισχύει για το κριτήριο της πληρότητας.

ΠΙΝΑΚΑΣ 4.3: ΓΝΩΡΙΣΜΑΤΑ ΣΤΑ ΠΛΑΙΣΙΑ ΤΗΣ 5ΗΣ ΕΠΑΝΑΛΗΨΗΣ

Feature
Mean Layer 4
HSI Transformation Intensity(R=Layer 3,G=Layer 2,B=Layer 1)
Number of pixels
GLCM Contrast (0ฐ)
GLCM Dissimilarity (all dir.)
GLCM Dissimilarity (0ฐ)
GLDV Ang. 2nd moment (45ฐ)
GLCM Correlation (all dir.)

Area
Length/Width
Elliptic fit
Shape index
MSAVI2
Length
Asymmetry
Border index
GLCM Entropy (90ฐ)
GLCM Mean (all dir.)
GLCM Mean (0ฐ)
Standard deviation Layer 3
Standard deviation Layer 4
GLCM Contrast (all dir.)
GLCM Contrast (45ฐ)
Compactness
GLCM Homogeneity (all dir.)
GLCM Correlation (45ฐ)
GLCM Homogeneity (90ฐ)
GLCM Entropy (all dir.)
Standard deviation Layer 1
Standard deviation Layer 2
GLCM Homogeneity (135ฐ)
Max. diff.
GLCM Homogeneity (45ฐ)
NDVI
GLCM Entropy (0ฐ)

ΠΙΝΑΚΑΣ 4.4*: ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΗΝ ΕΙΚΟΝΑ ΤΟΥ COLORADO (ΔΕΙΚΤΕΣ ΠΟΙΟΤΗΤΑΣ) (5Η ΕΠΑΝΑΛΗΨΗ).*

Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
70,06%	78,57%	58,82%	29,94%	19,11%

Στη συνέχεια, έγινε διερεύνηση της συσχέτισης μεταξύ των γνωρισμάτων ανά ζεύγη. Μέσω της διαδικασίας αυτής απομακρύνθηκαν από τη διαδικασία της ταξινόμησης όσο χαρακτηριστικά εμφάνιζαν συσχέτιση με κάποιο από τα εναπομείναντα μεγαλύτερη από 0.95.

Ακολούθως, έγινε μία προσπάθεια ώστε το τελικό αποτέλεσμα της ταξινόμησης να εμφανίζει αφενός τα υψηλά ποσοστά ποιότητας της αρχικής επιλογής των γνωρισμάτων (Πίνακας 4.2) και αφετέρου την υψηλή ορθότητα εκείνων του Πίνακας 4.3. Για το σκοπό αυτό, δημιουργήθηκε μία νέα λίστα στην οποία περιλαμβάνονται τα μη συσχετισμένα χαρακτηριστικά του Πίνακας 4.3 καθώς και εκείνα της αρχικής επιλογής (Πίνακας 4.5).

ΠΙΝΑΚΑΣ 4.5: ΤΕΛΙΚΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ

Features
NDVI

Mean Layer 3
Mean Layer 4
Elliptic fit
Length
Standard deviation Layer 3
Standard deviation Layer 1
Length/ Width
Density
Standard deviation Layer 4
Asymmetry
Mean Layer 2
Area
Mean Layer 1
Standard deviation Layer 2
GLCM Correlation (all dir)
Border Index
GLCM Homogeneity (90)
Compactness
GLCM Dissimilarity (all dir.)
GLCM Entropy (all dir)
GLCM Contrast (0)
Max diff
Shape Index
HIS Transformation Intensity
GLCM Homogeneity (135)
GLCM Contrast (all dir)
Rectangular fit
MSAVI2
GLDV Ang. 2nd moment (45)
GLCM Mean (all dir)

Τα ποσοστά ποιότητας όπως προέκυψαν έπειτα από εφαρμογή του μοντέλου στο υπό μελέτη τμήμα της εικόνας εισόδου εμφανίζονται στον Πίνακας 4.6. Παρατηρείται πως η τελική λίστα των επιλεγμένων χαρακτηριστικών είναι εκείνη στην οποία εμφανίζονται τα υψηλότερα ποσοστά ποιότητας καθώς και πληρότητας. Επιπροσθέτως, το ποσοστό της ορθότητας συγκεντρώνει ικανοποιητικά ποσοστά (78%). Συνεπώς, ο επιθυμητός στόχος του συνδυασμού καλών επιδόσεων σε ό,τι αφορά τόσο την ορθότητα όσο και την πληρότητα ικανοποιήθηκε μέσω της παραπάνω λίστας.

ΠΙΝΑΚΑΣ 4.6: ΣΥΓΚΡΙΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ ΤΩΝ ΕΠΙΔΟΣΕΩΝ ΤΩΝ ΤΡΙΩΝ ΜΟΝΤΕΛΩΝ

	Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
Αρχική λίστα χαρακτηριστικών (ενότητα 3.7.10)	82,17%	71,27%	61,72%	17,83%	33,12%
Χαρακτηριστικά 5ης επανάληψης (ενότητα 0)	70,06%	**78,57%**	58,82%	29,94%	19,11%
Τελική επιλογή χαρακτηριστικών	**84,08%**	77,65%	**67,69%**	15,92%	24,20%

Τέλος, τα παραπάνω συμπεράσματα επιβεβαιώθηκαν έπειτα από εφαρμογή της παραπάνω διαδικασίας σε διαφορετικό τμήμα της εικόνας εισόδου. Το μοντέλο του τυχαίου δάσους το οποίο κατασκευάστηκε είχε τις τελικά επιλεγμένες τιμές των παραμέτρων και η

εκπαίδευση του έγινε βάσει των τελικά επιλεγμένων γνωρισμάτων της λίστα του Πίνακας 4.5. Στον Πίνακας 4.7 εμφανίζονται οι δείκτες ποιότητες της ταξινόμησης μέσω της παραπάνω διαδικασίας. Παρατηρείται πως τα ποσοστά πληρότητας και της ορθότητας είναι ενθαρρυντικά καθώς συγκεντρώνουν ποσοστά μεγαλύτερα του 80%.

ΠΙΝΑΚΑΣ 4.7: ***ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΤΑΞΙΝΟΜΗΣΗΣ ΤΩΝ ΤΥΧΑΙΩΝ ΔΑΣΩΝ ΣΤΟ 2^{Ο} ΤΜΗΜΑ ΤΗΣ ΕΙΚΟΝΑ ΤΟΥ COLORADO***

Πληρότητα	Ορθότητα	Ποιότητα	Σφάλμα Παράλειψης	Σφάλμα Συμπερίληψης
82,24%	80,00%	68,22%	17,76%	20,56%

4.2 Προοπτικές

Στόχος της παρούσας μεταπτυχιακής εργασίας ήταν διερεύνηση της αποτελεσματικότητας των αλγορίθμων δέντρα απόφασης και τυχαία δάση σε ό,τι αφορά την ανίχνευση κτιρίων. Παράλληλα καταβλήθηκαν προσπάθειες ώστε οι δείκτες ποιότητας των παραγόμενων αποτελεσμάτων να αυξηθούν. Το παραπάνω υλοποιήθηκε μέσω της διερεύνησης της επιρροής των τιμών των παραμέτρων των αλγορίθμων καθώς και των χαρακτηριστικών των αντικειμένων (για την περίπτωση του τυχαίου δάσους) στην ποιότητα των θεματικών χαρτών. Τα αποτελέσματα ήταν ενθαρρυντικά, ιδιαίτερα στην περίπτωση των τυχαίων δασών καθώς οι δείκτης πληρότητας και ορθότητας συγκεντρώνουν ποσοστά μεγαλύτερα του 75%.

Μελλοντικά θα είχε ιδιαίτερο ενδιαφέρον να εξεταστεί η αποτελεσματικότητα των παραπάνω αλγορίθμων σε ό,τι αφορά την ταξινόμηση των κτιρίων σε θεματικές κατηγορίες. Το παραπάνω απαιτεί εκ νέου διερεύνηση της επιρροής των παραμέτρων του αλγορίθμου καθώς και των χαρακτηριστικών των αντικειμένων στην ποιότητα του παραγόμενου θεματικού χάρτη.

Παράλληλα, θα ήταν χρήσιμο να γίνει διαχείριση του ζητήματος της ανισορροπίας των κλάσεων μέσω των μεθοδολογιών που προτείνουν οι (Strumpf and Kerle , 2011) και (Puissant et al., 2014). Μέσω των συγκεκριμένων είναι δυνατή η μείωση του σφάλματος που προκαλείται σε βάρος των θεματικών κατηγοριών που καλύπτουν μικρότερο μέρος της εικόνας συγκριτικά με άλλες.

Η επίλυση των παραπάνω ζητημάτων θα αποτελέσει σημαντικό βήμα προς την κατεύθυνση της αυτόματης αναγνώρισης κτιρίων μέσω δορυφορικών εικόνων. Οι αλγόριθμοι που μελετήθηκαν στα πλαίσια της παρούσας εργασίας και ιδιαίτερα εκείνος των τυχαίων δασών δίνουν εντυπωσιακά αποτελέσματα σε ό,τι αφορά τη συγκεκριμένη εφαρμογή. Για το λόγο αυτό κρίνεται σκόπιμη η περαιτέρω διερεύνηση των δυνατοτήτων τους καθώς και η διαχείριση των υπάρχουσων αδυναμιών ούτως ώστε τα αποτελέσματα που προκύπτουν μέσω των συγκεκριμένων να είναι τα βέλτιστα δυνατά.

Βιβλιογραφία

Arvor, Durieux, Andrés, Laporte M. (2013). Advances in Geographic ObjectBased Image Analysis with ontologies: A review of main contributions and limitations from a remote sensing perspective. *ISPRS Journal of Photogrammetry and Remote Sensing* .

Belgiu and Dragut. (2014). Comparing supervised and unsupervised multiresolution segmentation approaches for extracting buildings form very high resolution images. *ISPRS Journal of Photogrammetry and Remote Sensing*.

Belgiu and Dragut. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing 114*, pp. 24-31.

Blaschke, Hay, Kelly, Lang, Hofmann, Addink, Feitosa, Meer, Werff, Coillie, Tiede. (2014). Geographic Object - Based Image Analysis - Towards a new paradigm. *ISPRS Journal of Photogrammetry and Remote Sensing*.

Bosch, Zisserman, Munoz. (2007). Image Classification using Random Forests and Ferns. *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on.* Rio de Janeiro: IEEE.

Chehata, Guo, Mallet. (2009). Airborne lidar Feature Selection for urban classification using random Forests. *Bretar F, Pierrot-Deseilligny M, Vosselman G (Eds) Laser scanning 2009, IAPRS, Vol. XXXVIII, Part 3/W8 – .* Paris, France, September 1-2.

Cox and Cox. (2001). *Multiresolution Scaling.* Chapman & Hall/CRC.

Díaz-Uriarte & De Andres (2006). Gene selection and classification of microarray data using random forest. BMC bioinformatics, 7(1), 1..

Du, Zhang, Zhang. (2015). Semantic classification of urban building vombining VHR Image and GIS data: An improved Random Forest approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, pp. 107-119.

Friedl and Brodley . (1997). Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment Vol 61, Issue 3* (pp. 319-440). Elsevier.

Gislason, Benediktsson, Sveinsson. (2004). Random Forest Classification of Multisource Remote Sensing and Geographic Data. *IEEE*, pp. 1049-1052.

Gislason, Benediktsson, & Sveinsson (2006). Random forests for land cover classification. Pattern Recognition Letters, 27(4),pp 294-300.

Ham, Chen, Crawford, Ghosh . (2005, March). Investigation of the Random Forest Framework for classification of hyperspectral data. *Geoscience and Remote Sensing, IEEE Transactions on (Volume:43 , Issue: 3*, σσ. 492 - 501.

Friedman, Hastie, & Tibshirani (2001). The elements of statistical learning (Vol. 1). Springer, Berlin: Springer series in statistics.

Heumann (2011). An object-based classification of mangroves using a hybrid decision tree—Support vector machine approach. Remote Sensing, 3(11), 2440-2460.James, W. H. (2013). *An Introduction to Statistical Learning.* Springer.

Juel, Groom, Svenning, Ejirnaes. (2015). Spatial application of Random Forest models for fine- scale coastal vegetation classification using object- based analysis of aerial orthophoto and DEM data. *International Journal of Applied Earth Observation and Geoinformation*.

Laliberte, Fredickson, Rango. (2007). Combining Decision Trees with Hierchical Object-oriented Image Analysis for Mapping Arid Rangelands. *Photogrammetric Engineering & Remote Sensing*, 197-207.

Liaw & Wiener(2002). Classification and regression by randomForest. R news, 2(3), 18-22.

Pal and Mather (2003). An assessment of the effectiveness of decision tree methods for land cover classification. *Remote Sensing Environment Vol. 86, Issue 4* (pp. 554-565). Elsevier.

Paris & Durand (2009). A fast approximation of the bilateral filter using a signal processing approach. International journal of computer vision, 81(1), 24-52.Pléiades imagery users guide. (2012, October).

Powers, Hermosilla, Coops, Chen. (2015). Remote sensing and object- based techniques for mapping fine- scale industrial disturbances. *International Journal of Applied Earth Observation and Geoinformation*, 51-57.

Puissant, Rougier, Stumpf (2014). Object-oriented mapping of urban trees using Random Forest classifiers. International Journal of Applied Earth Observation and Geoinformation, 26, 235-245.

Rodriguez- Galiano, Ghimire, Rogan, Chica- Olmo,Rigol- Sanchez. (2011, December). An assessment of effectiveness of a random forest classifier for land- cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, pp. 93-104.

Rokach (2009). A survey of clustering algorithms. In Data mining and knowledge discovery handbook (pp. 269-298). Springer US.Rougier and Puissant. (2014). improvements of urban vegetation segmentation and classification using multi- temporal Pleiades Images. *South- Eastern European Journal of Earth Observation and Geomatics*.

Safanian and Landgrebe . (1990). Ανάκτηση από NASA Technical Reports Server (NTRS): http://ntrs.nasa.gov/

Smith, Macleod, Kloiber. (2012). Updating the national wetland inventory in Minnesota by intergrating air photo-interpretation, object- oriented image analysis and mulitisourve data fusion. *ASPRS 2012 Annual Conference.* Sacramento, California.

Strumpf and Kerle . (2011). Object- oriented mapping of landslides using Random Forests. Στο *Remote Sensing of Environenment* (pp. 2564-2572).

Tan, Steinbach, Kumar. (2005). *Introduction to Data Mining.* Addison-Wesley Longman Publishing Co.

Tasnim and Rahman. (2014, April). Ensemble Classifiers and Their Applications: A Review. *International Journal of Computer Trends and Technology (IJCTT), vol.10 number 1*, pp. 31-35.

Tinel, Grizonnet, Fontannaz, Boissezon, Giros.(2012).Orfeo the Pléiades accompaniment pro gramm and its users thematic commissioning.I*nternational Archives of the Photogra mmetry, Remote Sensing and Spatial Information Science.* XXII ISPRS Congress.

Trimble, eCognition Developer. (2016). *Reference Book.* Munich, Germany: Trimble.

Xu, Watanachaturaporn, Varshney, Arora. (2005). Decision tree regression for soft classification of remote sensing data. Στο *Remote Sensing of Environment Vol 97, Issue 3* (σσ. 332-336). Elsevier.

Αργιαλάς. (1998). *Ψηφιακή Τηλεπισκόπηση.* Αθήνα: Εθνικό Μετσόβιο Πολυτεχνείο.

Ho, (1995, August). Random decision forests. In Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on (Vol. 1, pp. 278-282). IEEE..

Ηλεκτρονικές Πηγές

TreePlan = The Decision Tree Add-in (http://treeplan.com/) (προσβάστηκε στις 5/5/2016)

Trimble http://www.eCognition.com/ (πρόσβαστηκε στις 4/5/2016)

Uta State University - Mathematics and Statistics http://www.math.usu.edu/ (προσβάστηκε στις 2/3/2016)

Carnegie Mellon - University School of Computer Science https://www.cs.cmu.edu/ (προσβάστηκε στις 5/6/2016)

University of California Berkley – Department of Statistics https://www.stat.berkeley.edu/ (προσβάστηκε στις 28/3/2016)