**ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ**

Σχολή Αγρονόμων & Τοπογράφων Μηχανικών – Μηχανικών Γεωπληροφορικής
Τομέας Τοπογραφίας
Εργαστήριο Τηλεπισκόπησης

**Διπλωματική Εργασία**

# Αξιολόγηση μεθόδων ταξινόμησης καλλιεργειών με χρήση χρονοσειρών Sentinel δεδομένων

Ελευθέριος Θεοδωρόπουλος

Επιβλέπουσα καθηγήτρια: Δρ. Βασιλεία Καραθανάση

Αθήνα, Ιούλιος 2023

# Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω την καθηγήτρια Βασιλεία Καραθανάση για την καθοδήγηση και τις συζητήσεις πάνω σε προβλήματα που προέκυψαν κατά τη διάρκεια της διπλωματικής, καθώς με έκανε να ψάξω και να μάθω νέες τεχνικές, αλλά και τις γνώσεις που αποκόμισα από τη διδασκαλία της στα μαθήματα.

Επίσης, θα ήθελα να ευχαριστήσω ειλικρινά όλο το εργαστήριο τηλεπισκόπησης, τους καθηγητές Δημήτρη Αργιαλά και Κωνσταντίνο Καράντζαλο και τους κ. Κολοκούση, τον κ. Ανδρώνη, την κα. Βασιλείου, των οποίων η συμβολή υπήρξε καθοριστική για τις μελλοντικές επιδιώξεις μου. Ιδιαίτερες ευχαριστίες στον κ. Χρήστο Ιωσηφίδη για τις αμέτρητες ώρες συζητήσεων και συμβουλών!

Παράλληλα, ευχαριστίες θα ήθελα να δώσω και στον Δρ. Χάρη Κοντοέ, Διευθυντή Ερευνών του Εθνικού Αστεροσκοπείου Αθηνών και Υπεύθυνο της ομάδας BEYOND και τους ερευνητές Ιάσονα Τσαρδανίδη και Άλκη Κούκο για τη διάθεση των δεδομένων και τη βοήθειά τους.

Τέλος, βαθιές ευχαριστίες στην οικογένειά μου, τους φίλους και τις φίλες μου για την υποστήριξη όλα αυτά τα χρόνια. Ένα μεγάλο στον Μιχάλη και στην Χριστίνα για την υπομονή τους και την στήριξη.

# Περιεχόμενα

# Περίληψη

Σκοπός της παρούσας διπλωματικής εργασίας είναι να αξιολογηθούν διάφοροι αλγόριθμοι επιβλεπόμενης και μη-επιβλεπόμενης μάθησης για την ταξινόμηση καλλιεργειών με χρήση χρονοσειρών από δορυφορικά δεδομένα Sentinel-1 και Sentinel-2.

Η παρούσα εργασία εστιάζει σε τρεις περιοχές της Λιθουανίας. Το σύνολο των δεδομένων αποτελείται από 27073 αγροτεμάχια με 7 κατηγορίες καλλιεργειών και 462 δορυφορικά χαρακτηριστικά για κάθε αγροτεμάχιο.

Οι αλγόριθμοι μη-επιβλεπόμενης μάθησης που χρησιμοποιήθηκαν είναι οι k-means, Gaussian Mixture Model και εφαρμόστηκαν τεχνικές μείωσης διαστάσεων όπως PCA και t-SNE. Οι παραπάνω αλγόριθμοι αξιολογήθηκαν με βάση τη βαθμολογία silhouette.

Σχετικά με την επιβλεπόμενη ταξινόμηση, χρησιμοποιήθηκαν οι αλγόριθμοι Random Forest, Linear SVM και Gradient Boosting και αναδείχθηκε ο καταλληλότερος για τη συγκεκριμένη εργασία. Στη συνέχεια, εφαρμόστηκαν τεχνικές εξαγωγής των σημαντικότερων χαρακτηριστικών με σκοπό να μειωθεί ο όγκος των δεδομένων και να επιτευχθεί υψηλή ακρίβεια. Τέλος, εξετάστηκαν τεχνικές για τη γενίκευση των μοντέλων εκπαίδευσης και της μεταφοράς τους σε άλλες περιοχές, είτε με εκπαίδευση σε μεγαλύτερο αριθμό δειγμάτων ή εφαρμόζοντας τεχνικές oversampling και undersampling για να παραχθεί ένα πιο καλά ισορροπημένο σύνολο δεδομένων.

Τα αποτελέσματα παρέχουν πληροφορίες για την καταλληλότητα και την απόδοση διαφορετικών αλγορίθμων στο πλαίσιο παρακολούθησης και ταξινόμησης καλλιεργειών με χρήση δορυφορικών δεδομένων.

# Abstract

The purpose of this thesis is to evaluate various supervised and unsupervised learning algorithms for crop classification using time series from Sentinel-1 and Sentinel-2 satellite data.

The thesis focuses on three regions in Lithuania and utilizes a dataset comprising 27073 crops. with 7 labels and 462 features derived from satellite bands and indices.

To explore the dataset through unsupervised learning two algorithms, k-means and Gaussian Mixture Model algorithms were employed. Additionaly, dimensionality reduction techniques such as PCA and t-SNE were applied. The evaluation of the unsupervised algorithms was conducted using the silhouette score.

As far as the supervised classification is concerned, Random Forest, Linear SVM and Gradient Boosting algorithms were used. Furthermore, feature importance techniques were applied in order to reduce the amount of data and achieve high accuracy. Finally, inference techniques for generalizing the training models and transferring them to other regions were considered, either by training on a larger number of samples or applying oversampling and undersampling techniques to produce a more well-balanced data set.

The results provide valuable insights into the suitability and performance of different algorithms in the domain of crop monitoring and classification utilizing satellite data.

# Κεφάλαιο 1 – Εισαγωγή

Η ταξινόμηση των καλλιεργειών παίζει σημαντικό ρόλο στη γεωργία ακρίβειας, στην παρακολούθηση και την διαχείριση των καλλιεργειών, καθώς και της Κοινής Αγροτικής Πολιτικής, παρέχοντας γνώσεις για την υγεία των καλλιεργειών, τα διάφορα φαινολογικά στάδια που ακολουθούν, καθώς και της πρόβλεψης της σοδειάς. Η ζήτηση που υπάρχει παγκοσμίως για πιο βιώσιμες γεωργικές πρακτικές και παραγωγή τροφίμων, απαιτεί αποτελεσματικότερη παρακολούθηση και κατανόηση των καλλιεργειών. Η βελτίωση στις ταξινομήσεις των καλλιεργειών επιτρέπει στους υπεύθυνους χάραξης πολιτικής, τους αγρότες και όσους διαχειρίζονται γη να βελτιστοποιήσουν την κατανομή των πόρων και να λαμβάνουν τεκμηριωμένες αποφάσεις σχετικά με τις γεωργικές πρακτικές.

Με τις εξελίξεις στην τηλεπισκόπηση, τα δορυφορικά δεδομένα έχουν γίνει σημαντική πηγή πληροφοριών για την ταξινόμηση των καλλιεργειών. Συγκεκριμένα, τα υπερφασματικά και τα ραντάρ δεδομένα μπορούν να

Στόχος της παρούσας διπλωματικής εργασίας είναι η αξιολόγηση διάφορων αλγορίθμων επιβλεπόμενης και μη-επιβλεπόμενης μάθησης σε χρονοσειρές δορυφορικών δεδομένων Sentinel-1 και Sentinel-2, καθώς και η διερεύνηση τεχνικών εξαγωγής σημαντικών χαρακτηριστικών και γενίκευσης μοντέλων με σκοπό της μεταφοράς τους σε άλλες περιοχές.

Η διπλωματική εργασία ξεκινάει με το θεωρητικό υπόβαθρο, το οποίο θέτει τις βάσεις για την κατανόηση εννοιών όπως η τηλεπισκόπηση, η μηχανική μάθηση, οι αλγόριθμοι επιβλεπόμενης και μη-επιβλεπόμενης ταξινόμησης, μετρικές αξιολόγησης και τεχνικές επεξεργασίας δεδομένων.

Η βιβλιογραφική ανασκόπηση έχει ως στόχο να εμβαθύνει σε εφαρμογές διάφορων αλγορίθμων (όπως ο k-means, ο random forest κτλπ) στην ταξινόμηση καλλιεργειών και να επισημάνει διαφορετικές μεθοδολογίες, αλγορίθμους και πηγές δεδομένων που έχουν χρησιμοποιηθεί σε άλλες μελέτες.

Στη συνέχεια, το κεφάλαιο 4 της μεθοδολογίας θα περιγράψει αναλυτικά τα βήματα που ακολουθήθηκαν, συμπεριλαμβανομένων της περιοχής μελέτης, των δεδομένων (επιτόπιων και δορυφορικών), των τεχνικών επεξεργασίας των δεδομένων, της επιλογής των αλγορίθμων επιβλεπόμενης και μη-επιβλεπόμενης μάθησης, των τεχνικών εξαγωγής σημαντικών χαρακτηριστικών και γενίκευσης των μοντελων, καθώς και την αξιολόγησή τους.

Στο κεφάλαιο 5, θα παρουσιαστούν τα αποτελέσματα της μελέτης, μεταξύ των οποίων οι αξιολογήσεις των μοντέλων επιβλεπόμενης και μη-επιβλεπόμενης μάθησης, συγκρίσεις διαφορετικών αλγορίθμων επιβλεπόμενης μάθησης, καθώς και τεχνικές για την βελτίωση των αποτελεσμάτων.

Τέλος, στο κεφάλαιο 6, θα συζητηθούν τα συμπεράσματα της εργασίας και θα αναλυθούν μελλοντικές κατευθύνσεις για την έρευνα.

## 2.1. Τηλεπισκόπηση

Ο όρος τηλεπισκόπηση (τηλέ + επισκοπώ = εξετάζω κάτι από μακρυά) αναφέρεται στην επιστήμη σχετικά με την απόκτηση πληροφοριών για τα χαρακτηριστικά της επιφάνεια της Γης από απόσταση, χωρίς να απαιτείται φυσική επαφή. Οι πληροφορίες για την μελέτη του εκάστοτε χαρακτηριστικού προκύπτουν από την αλληλεπίδραση του αντικειμένου με την ηλεκτρομαγνητική ακτινοβολία.

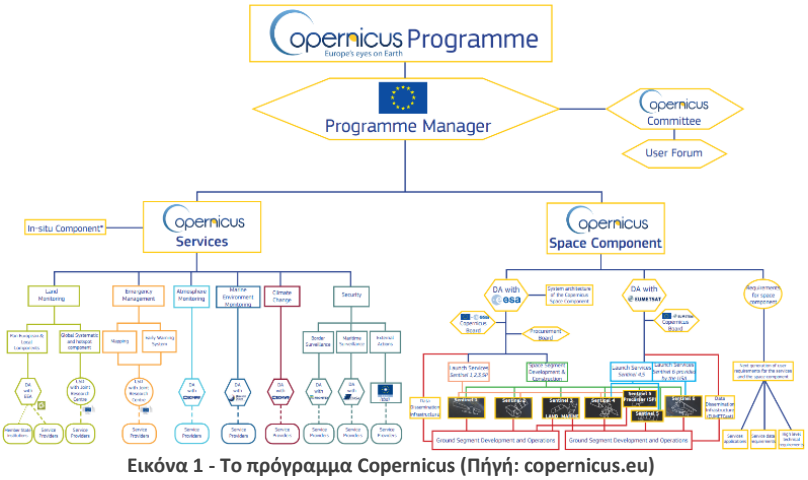
Όσον αφορά την τηλεπισκόπηση, χρειάζεται ένα αντικείμενο προς παρατήρηση, ένας αισθητήρας που θα πραγματοποιήσει τις μετρήσεις και μία πλατφόρμα, η οποία θα μεταφέρει τον αισθητήρα. Οι αισθητήρες περιλαμβάνουν φωτογραφικές μηχανές, συστήματα ραντάρ, πολυφασατικούς δέκτες, lidar κ.λ.π. Οι πλατφόρμες, οι οποίες μεταφέρουν αυτούς τους αισθητήρες μπορεί να είναι ένα drone, ένα μετεωρολογικό μπαλόνι, ένα μη-επανδρωμένο αεροσκάφος ή ένας δορυφόρος.

### 2.1.2. Πρόγραμμα Copernicus

Το Copernicus είναι το ευρωπαϊκό πρόγραμμα Παρατήρησης της Γης, το οποίο παρατηρεί τον πλανήτη Γη και παρέχει πληροφορίες βασιζόμενες σε επιτόπια δεδομένα (in-situ) και διαστημικά δεδομένα, από τις αποστολές Sentinels και τις συνεργαζόμενες διαστημικές αποστολές.

Συντονιστής του προγράμματος είναι η Ευρωπαϊκή Επιτροπή και η υλοποίηση του γίνεται σε συνεργασία με τα κράτη μέλη, τον Ευρωπαϊκό Οργανισμό Διαστήματος (ESA), τον Ευρωπαϊκό Οργανισμό Εκμετάλλευσης Μετεωρολογικών Δορυφόρων (EUMETSAT), το Ευρωπαϊκό Κέντρο Μεσοπρόθεσμων Μετεωρολογικών Προβλέψεων (ECMWF), οργανισμούς της ΕΕ και την εταιρεία Mercator Océan. Οι θεματικές που καλύπτει το πρόγραμμα Copernicus είναι οι εξής:

- Ατμόσφαιρα
- Θάλασσα
- Ξηρά
- Κλιματική Αλλαγή
- Ασφάλεια
- Έκτακτες Ανάγκες




**Εικόνα 1 - Το πρόγραμμα Copernicus (Πήγή: copernicus.eu)**

## 2.1.3. Δορυφόρος Sentinel-1

Η ευρωπαϊκή αποστολή Sentinel-1 του προγράμματος Copernicus της ESA αποτελείται από δύο πανομοιότυπους δορυφόρους ,τον Sentinel-1A και Sentinel-1B που εκτοξεύτηκαν το 2014 και το 2016 αντίστοιχα. Οι δύο δορυφόροι φέρουν ένα όργανο ραντάρ συνθετικού ανοίγματος που λειτουργεί στη C μπάντα, χάρη στο οποίο υπάρχουν δεδομένα ανεξαρτήτως των καιρικών φαινομένων που επικρατούν στη γήινη επιφάνεια και των συνθηκών φωτισμού. Μερικές από τις κύριε εφαρμογές της αποστολής Sentinel-1 είναι η χερσαία παρακολούθηση στο κομμάτι των καλλιεργειών, των δασικών εκτάσεων και του αστικού ιστού, η θαλάσσια παρακολούθηση όσον αφορά την παρακολούθηση πάγων, πετρελαιοκηλίδων και πλοίων και ,τέλος, η διαχείριση κρίσεων όπως πλημμύρες, σεισμοί, ηφαίστεια και κατολισθήσεις. Ωστόσο, ο δορυφόρος Sentinel-1B βρίσκεται εκτός λειτουργείας από τον Δεκέμβριο του 2021, οπότε υπάρχουν δεδομένα διαθέσιμα ανά 12 ημέρες. Σύντομα, ο Sentinel-1B θα αποκατασταθεί από τον δορυφόρο Sentinel-1C.

**Εικόνα 2 - Δορυφόρος Sentinel-1 (Πηγή: esa.int)**

## 2.1.3. Δορυφόρος Sentinel-2

Η ευρωπαϊκή αποστολή Sentinel-2 του προγράμματος Copernicus της ESA αποτελείται από δύο πανομοιότυπους δορυφόρους ,τον Sentinel-2A και Sentinel-2B που εκτοξεύτηκαν το 2015 και το 2017 αντίστοιχα. Οι δύο αυτοί δορυφόροι φέρουν ένα υπερφασματικό όργανο (MSI) ,το οποίο λειτουργεί στο οπτικό φάσμα και «φωτογραφίζουν» τη Γη κάθε 5 ημέρες στον Ισημερινό και κάθε 2-3 μέρες σε μεσαία γεωγραφική πλάτη. Σκοπός της αποστολής Sentinel-2 είναι η παρακολούθηση Γης, τόσο σε χερσαίο όσο και σε υδάτινο περιβάλλον, η κλιματική αλλαγή, καθώς και η διαχείριση κρίσεων. Στον παρακάτω πίνακα παρουσιάζονται οι μπάντες του Sentinel-2, το κεντρικό μήκος κύματος κάθε μπάντας.

| Sentinel-2 Μπάντες | Μήκος Κύματος (μm) | Χωρική Ανάλυση (m) |
|---|---|---|
| B01 – Coastal Aerosol | 0.443 | 60 |
| B02 – Blue | 0.490 | 10 |
| B03 – Green | 0.560 | 10 |
| B04 – Red | 0.665 | 10 |
| B05 – Vegetation Red Edge | 0.705 | 20 |
| B06 – Vegetation Red Edge | 0.740 | 20 |
| B07 – Vegetation Red Edge | 0.783 | 20 |
| B08 – NIR | 0.842 | 10 |
| B08A – Vegetation Red Edge | 0.865 | 20 |
| B09 – Water Vapour | 0.945 | 60 |
| B10 – SWIR – Cirrus | 1.374 | 60 |
| B11 – SWIR | 1.610 | 20 |
| B12 – SWIR | 2.190 | 20 |

**Εικόνα 3 - Δορυφόρος Sentinel-2 (Πηγή: esa.int)**

## 2.2. Μηχανική Μάθηση

Η μηχανική μάθηση είναι ένας κλάδος της τεχνητής νοημοσύνης που εστιάζει στην ανάπτυξη αλγορίθμων και μοντέλων, τα οποία δίνουν στα συστήματα υπολογιστών τη δυνατότητα να μαθαίνουν από μόνα τους και να αποφασίζουν ή να προβλέπουν χωρίς κάποιος να τα προγραμματίζει. Μέσω μιας επαναληπτικής διαδικασίας εκπαίδευσης, αυτοί οι αλγόριθμοι βελτιώνουν την απόδοσή τους, προσαρμόζονται σε νέες πληροφορίες και βελτιστοποιούν τις προγνωστικές τους ικανότητες. Η μηχανική μάθηση βρίσκει εφαρμογές σε διάφορους τομείς, όπως η αναγνώριση εικόνας, η επεξεργασία φυσικής γλώσσας, τα αυτόνομα οχήματα, καθώς και η τηλεπισκόπηση.

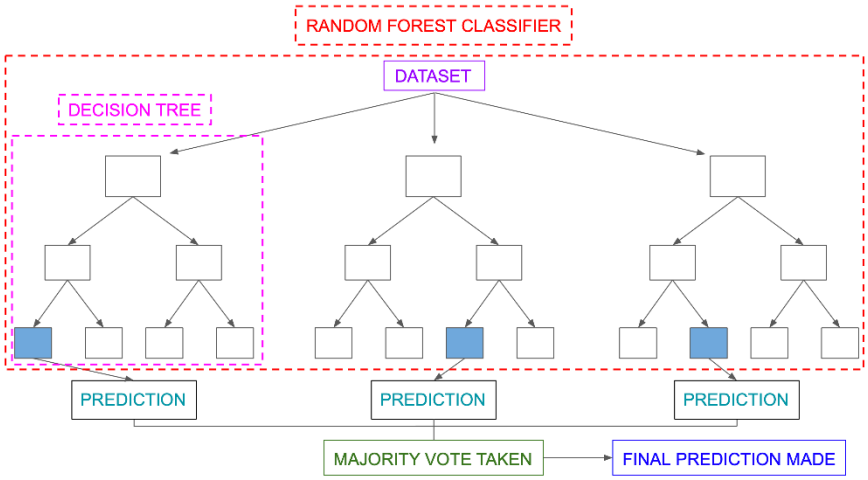
### 2.2.1. Επιβλεπόμενη Μάθηση

Η επιβλεπόμενη μάθηση είναι μία διαδικασία, στην οποία ένας αλγόριθμος μαθαίνει να ταξινομεί ή να κάνει προβλέψεις με τη χρήση δεδομένων εκπαίδευσης που έχουν ετικέτες (labels). Τα δεδομένα εκπαίδευσης αποτελούνται από τις μεταβλητές εισόδου (χαρακτηριστικά) και τις αντίστοιχες μεταβλητές εξόδου (ετικέτες). Ο αλγόριθμος μαθαίνει τη σχέση μεταξύ των δεδομένων εσόδου και εξόδου και χρησιμοποιεί αυτή τη γνώση για να κάνει προβλέψεις ή να ταξινομήσει νέα δεδομένα.

#### *2.2.1.1. Random Forest*

Ο Random Forest είναι ένας ταξινομητής συνόλου δεδομένων και αποτελείται από πολλά δέντρα αποφάσεων (decision trees). Ένα δέντρο απόφασης είναι ένας αλγόριθμος που οδηγεί στη δημιουργία μιας δενδροειδούς μορφής, όπου κάθε εσωτερικός κόμβος ανταποκρίνεται σε κάποιο χαρακτηριστικό και κάθε φύλλο αποτελεί μία κατηγορία ταξινόμησης (κλάση). Τα βήματα του Random Forest είναι τα εξής:

1. Τυχαία επιλογή υποσυνόλου των δεδομένων εκπαίδευσης, με σκοπό τη δημιουργία πολλαπλών δέντρων απόφασης.
2. Κάθε δέντρο απόφασης εκπαιδεύεται σε ένα υποσύνολο χαρακτηριστικών που επιλέγεται τυχαία από το πλήρες σύνολο χαρακτηριστικών.
3. Για κάθε δέντρο απόφασης, ο αλγόριθμος χωρίζει τα δεδομένα με βάση τα επιλεγμένα χαρακτηριστικά.
4. Κατά τη διαδικασία της πρόβλεψης, κάθε δέντρο απόφασης ταξινομεί ανεξάρτητα τα δεδομένα εισόδου του.
5. Η τελική πρόβλεψη καθορίζεται με βάση τις προβλέψεις των δέντρων απόφασης μέσω της πλειοψηφίας (για ταξινόμηση) ή του μέσου όρου (για παλινδρόμηση).



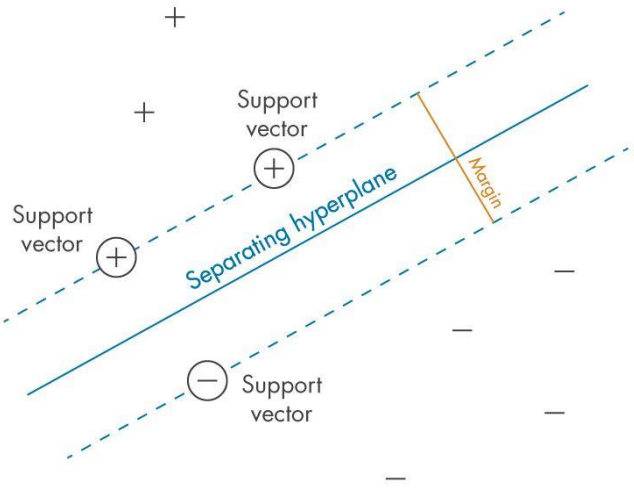

**Εικόνα 4 - Random Forest (Πηγή: section.io)**

### *2.2.1.2. Linear SVM*

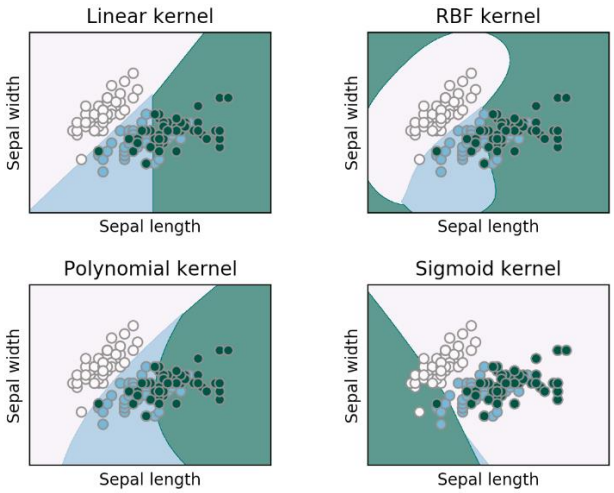
Ο Support Vector Machine (SVM) είναι ένας δημοφιλής αλγόριθμος επιβλεπόμενης μάθησης που χρησιμοποιείται για ταξινόμηση και παλινδρόμηση. Στοχεύει στην εύρεση ενός βέλτιστου υπερεπίπεδου που διαχωρίζει τα δεδομένων διαφορετικών κλάσεων.

1. Δεδομένου ενός συνόλου επισημασμένων παραδειγμάτων εκπαίδευσης, ο αλγόριθμος SVM βρίσκει το βέλτιστο υπερεπίπεδο (hyperplane) που διαχωρίζει στο μέγιστο τα σημεία δεδομένων που ανήκουν σε διαφορετικές κλάσεις.
2. Το υπερεπίπεδο καθορίζεται από διανύσματα υποστήριξης (support vectors), τα οποία είναι τα σημεία δεδομένων που βρίσκονται πιο κοντά στο όριο απόφασης.
3. Κατά τη διάρκεια της εκπαίδευσης, σκοπός του SVM είναι να ελαχιστοποιήσει τη συνάρτηση απώλειας, η οποία με τη σειρά της ελαχιστοποιεί τις λάθος ταξινομήσεις και μεγιστοποιεί το περιθώριο μεταξύ του ορίου απόφασης και των διανυσμάτων υποστήριξης.
4. Το τελικό υπερεπίπεδο καθορίζεται με βάση τα διανύσματα στήριξης και τα αντίστοιχα βάρη τους, τα οποία αντιπροσωπεύουν τη σημαντικότητά τους στο όριο απόφασης. Το υπερεπίπεδο είναι τοποθετημένο έτσι ώστε να μεγιστοποιεί το περιθώριο, εξασφαλίζοντας σαφή διαχωρισμό μεταξύ των κατηγοριών.

Υπάρχουν διαφορετικά ήδη SVM, όπως ο Linear, ο Radial-Basis-Function(RBG), ο Polynomial και ο Sigmoid. Η διαφορά κάθε SVM έγκειται στη διαφορά του σχήματος του ορίου απόφασης, το οποίο καθορίζει και το υπερεπίπεδο. Για τον Linear svm το όριο απόφασης είναι γραμμικό, ο RBF χρησιμοποιεί ένα μη-γραμμικό, καμπύλο όριο απόφασης, ο Sigmoid χρησιμοποιεί ένα όριο απόφασης σε σχήμα S και ο Polynomial χρησιμοποιεί ένα καμπύλο όριο απόφασης μεγαλύτερης πολυωνυμικής ταξης.




**Εικόνα 5 - Υπερεπίπεδο SVM (Πηγή: mathworks.com)**

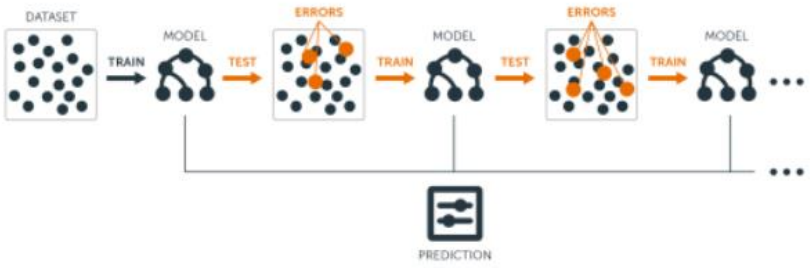



**Εικόνα 6 - Τύποι SVM (Πηγή: towardsdatascience.com)**

### *2.2.1.3. Gradient Boosting*

Ο ταξινομητής Gradient Boosting είναι μια τεχνική μηχανικής μάθησης που δημιουργεί ένα μοντέλο πρόβλεψης συνδυάζοντας πολλούς weak learners, συνήθως δέντρα αποφάσεων. Παρακάτω παρουσιάζονται τα βήματα του Gradient Boosting:

1. Η διαδικασία ξεκινάει με την αρχικοποίηση του μοντέλου με έναν weak learner.
2. Ο weak learner εκπαιδεύεται στα δεδομένα εκπαίδευσης και το μοντέλο κάνει προβλέψεις. Έπειτα, χρησιμοποιείται μία συνάρτηση απώλειας για να υπολογίσει το υπόλοιπο, δηλαδή τη διαφορά μεταξύ των προβλέψεων και των πραγματικών τιμών.
3. Στη συνέχεια ένας νέος weak learner εκπαιδεύεται στα υπόλοιπα, δίνοντας έμφαση στα δείγματα που δεν προβλέφθηκαν σωστά από τον προηγούμενο weak learner.
4. Τέλος, οι προβλέψεις του νέου weak learner συνδυάζονται με τις προηγούμενες προβλέψεις και το μοντέλο ανανεώνεται αναδιαμορφώνοντας τα βάρη σε κάθε weak learner βάση την απόδοσή τους να ελαχιστοποιήσουν τη συνάρτηση απώλειας.
5. Επανάληψη των βημάτων 2-4 για προκαθορισμένο αριθμό επαναλήψεων ή μέχρι να ικανοποιηθεί ένα κριτήριο διακοπής.
6. Κατά τη διάρκεια της πρόβλεψης, το μοντέλο συγκεντρώνει τις προβλέψεις όλων των αδύναμων μαθητών, σταθμισμένες με τη συνεισφορά τους, για να κάνει την τελική πρόβλεψη. Η συνάρτηση απώλειας δεν εμπλέκεται άμεσα στο βήμα πρόβλεψης, αλλά παίζει καθοριστικό ρόλο στη διαδικασίας εκπαίδευσης.




**Εικόνα 5 - Ταξινομητής Gradient Boosting (Πηγή: machine-learning.paperspace.com)**

### *2.2.1.4. Μέθοδος Αξιολόγησης*

Για την αξιολόγηση των αποτελεσμάτων των αλγορίθμων επιβλεπόμενης μάθησης χρησιμοποιήθηκε η ακρίβεια, η ανάκληση και η βαθμολογία F1.

True Positives (TP): Ο αριθμός των περιπτώσεων μίας κατηγορίας που ταξινομήθηκαν σωστά για την συγκεκριμένη κατηγορία.

True Negatives (TN): Ο αριθμός των περιπτώσεων μίας κατηγορίας που ταξινομήθηκαν λάθος σε άλλη κατηγορία.

False Positives (FP): Ο αριθμός των περιπτώσεων μίας κατηγορίας που ταξινομήθηκαν ως σωστές για μια διαφορετική κατηγορία, όταν στην πραγματικότητα είναι λανθασμένες.

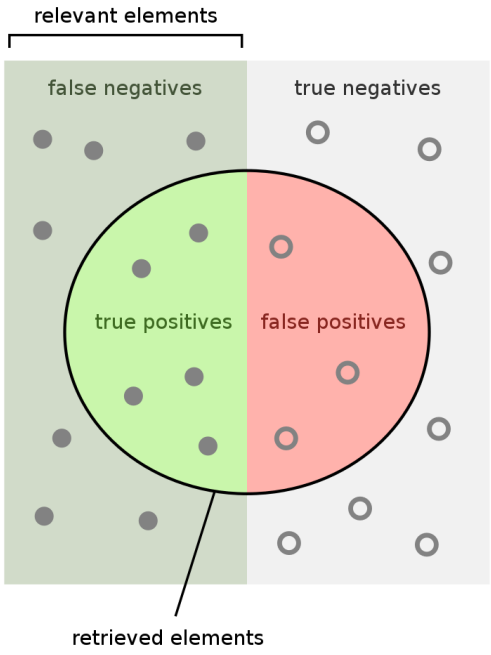
False Negative (FN): Ο αριθμός των περιπτώσεων μίας κατηγορίας που ταξινομήθηκαν ως λανθασμένες για μια διαφορετική κατηγορία, όταν στην πραγματικότητα είναι σωστές.

Χρησιμοποιώντας τους παραπάνω όρους, δύναται να ορίσουμε τις τιμές της ακρίβειας, την ανάκλησης και της βαθμολογίας F1 ως:

$Ακρίβεια(precision): \frac{TP}{TP+FP}$ , ορίζεται ως το κλάσμα των σωστά ταξινομημένων περιπτώσεων για μία δεδομένη κλάση ως προς τον αριθμό των περιπτώσεων που ταξινομήθηκαν στην ίδια κλάση.

$Ανάκληση(recall): \frac{TP}{TP+FN}$ , ορίζεται ως το κλάσμα των σωστά ταξινομημένων περιπτώσεων για μία δεδομένη κλάση ως προς τον πραγματικό αριθμό των περιπτώσεων της κλάσης.

$F1-score: \frac{2*Ακρίβεια*Ανάκληση}{Ακρίβεια+Ανάκληση}$, ορίζεται ως ο αρμονικός μέσος της ακρίβειας και της ανάκλησης.




**Εικόνα 6 - TP/TN/FP/FN (Πηγή: wikipedia)**

### *2.2.1.5 Η τεχνική SMOTE για oversampling*

Το SMOTE είναι μία τεχνική oversampling που χρησιμοποιείται συνήθως για την αντιμετώπιση των μη-ισορροπημένων κλάσεων σε δεδομένα μηχανικής μάθησης. Δημιουργεί συνθετικά δείγματα για τις κατηγορίες με λίγα δείγματα, παρεμβάλλοντας νέα συνθετικά δείγματα ανάμεσα από υπάρχοντα δείγματα. Τα βήματα που ακολουθεί η SMOTE είναι τα εξής:

1. Για κάθε δείγμα της κλάση με λίγα δείγματα υπολογίζει το k-κοντινότερο γείτονα (k-nearest neighbor) με βάση μία εξίσωση απόστασης (πχ. Ευκλείδια)
2. Τυχαία επιλογή ενός k-nearest neighbor
3. Δημιουργία ενός συνθετικού δείγματος, το οποίο προκύπτει ύστερα από τον πολλαπλασιασμό με έναν τυχαίο αριθμό μεταξύ 0 και 1 της διαφοράς του δείγματος με τον k-nearest neighbor
4. Επανάληψη των βημάτων 2,3

#### *2.2.1.5 Τεχνική RandomUndersampler για undersampling*

Το RandomUnderSampler είναι μια τεχνική που χρησιμοποιείται για το undersampling της των τάξεων με μεγάλο δείγμα. Επιλέγει τυχαία ένα υποσύνολο δειγμάτων από την κλάση με τα πολλά δείγματα για να δημιουργήσει ένα πιο ισορροπημένο σύνολο δεδομένων. Ακολουθούν τα βήματα που εμπλέκονται στο RandomUnderSampler:

1. Υπολογίζει τη διαφορά που προκύπτει αφαιρώντας τον επιθυμητό αριθμό δειγμάτων από τα αρχικά δείγματα
2. Με βάση τον αριθμό που προέκυψε από τη διαφορά, διαλέγει τυχαία ένα δείγμα από το αρχικό δείγμα δεδομένων

### 2.2.2. Μη Επιβλεπόμενη Μάθηση

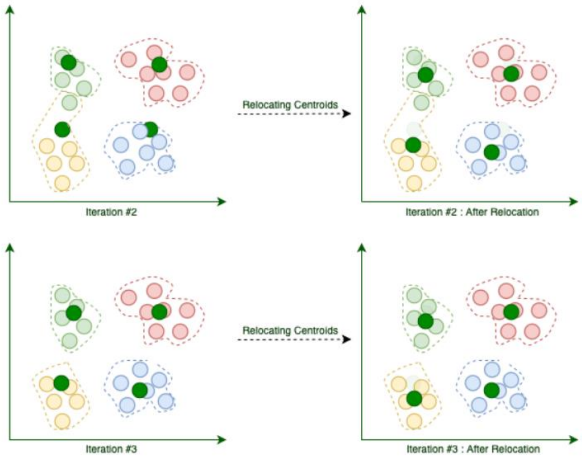
Η μη-επιβλεπόμενη μάθηση είναι ένας κλάδος της μηχανικής μάθησης, όπου το μοντέλο εκπαιδεύεται σε δεδομένα χωρίς ετικέτα (label). Σε αντίθεση με την επιβλεπόμενη μάθηση που αναπτύχθηκε στο παραπάνω κεφάλαιο, στην μη-επιβλεπόμενη μάθηση δεν υπάρχει γνώση για τις ετικέτες και τις κατηγορίες των δεδομένων εισόδου. Στόχος μέσα από τους αλγόριθμους μη-επιβλεπόμενης μάθησης είναι να ανακαλυφθούν μοτίβα και σχέσεις μεταξύ των δεδομένων, χωρίς προηγούμενη γνώση.

#### *2.2.2.1. K-means*

Ο k-means είναι ένας αλγόριθμος μη-επιβλεπόμενης ταξινόμησης που χρησιμοποιείται για ομαδοποίηση δεδομένων. Στόχος είναι να χωρίσει ένα σύνολο δεδομένων σε k διακριτές κλάσεις, όπου κάθε κλάση αντιπροσωπεύει μια ομάδα παρόμοιων σημείων. Ο αλγόριθμος λειτουργεί επαναληπτικά, ξεκινώντας να αναθέτει σημεία στο κοντινότερο κεντροειδές κάθε κλάσης και στη συνέχεια να ανανεώνει τα κεντροειδή με βάση τις νέες κλάσεις.

Συγκεκριμένα, ο αλγόριθμος ακολουθεί τα παρακάτω βήματα:

1. Τυχαία αρχικοποίηση κεντροειδών των k κλάσεων
2. Ανάθεση κάθε σημείου στο κοντινότερο κεντροειδές με βάση την Ευκλείδεια απόσταση
3. Ανανέωση των κεντροειδών, υπολογίζοντας τη νέα μέση απόσταση των σημείων σε κάθε κλάση
4. Επανάληψη των βημάτων 2 και 3 μέχρι να συγκλίνουν τα ανανεωμένα κεντροειδή κάθε κλάσης

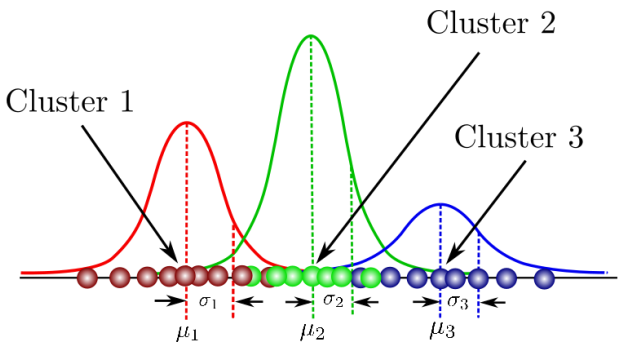



**Εικόνα 7 Τα βήματα του αλγορίθμου k-means (πηγή: www.baeldung.com)**

### 2.2.2.2. Gaussian Mixture Models

Ο αλγόριθμος Μοντέλων Μίξης Γκαουσιανών Κατανομών είναι ένας πιθανοτικός αλγόριθμος ομαδοποίησης που στοχεύει στον εντοπισμό κλάσεων σε ένα σύνολο δεδομένων. Ο αλγόριθμος αποτελείται από πολλά βήματα:

1. Καθορισμός των επιθυμητών κλάσεων
2. Αρχικοποίηση των παραμέτρων μέσος (m), συνδιακύμνασης (σ) και συντελεστής ανάμειξης (π) για κάθε κλάση. Οι κλάσεις είναι οι Γκαουσιανές συνιστώσες.
3. Βήμα E (Expectation), υπολογισμός της πιθανότητας για κάθε σημείο να ανήκει σε κάποια κλάση
4. Βήμα M (Maximizing), εκ νέου υπολογισμός των παραμέτρων m, σ και π για κάθε Γκαουσιανή συνιστώσα.
    4.1 Η ανανεωμένη παράμετρος m υπολογίζεται ο σταθμισμένος μέσος μεταξύ των σημείων κάθε συνιστώσας. Το βάρος για κάθε σημείο υπολογίζεται από την πιθανότητα του να ανήκει στην εκάστοτε συνιστώσα. Όσο μεγαλύτερη η πιθανότητα, τόσο μεγαλύτερο το βάρος.
    4.2 Η ανανεωμένη συνδιακύμανση σ υπολογίζεται ως ο σταθισμένος μέσος όρος του τετραγώνου της διαφοράς της απόστασης κάθε σημείου της συνιστώσας από το μέσο m της συνιστώσας. Και σε αυτή την περίπτωση, τα βάρη προκύπτουν από την πιθανότητα κάθε σημείου να ανήκει στη γκαουσιανή συνιστώσα.
    4.3 Ο νέος συντελεστής ανάμειξης υπολογίζεται από τον μέσο όρο των πιθανοτήτων κάθε σημείου για την εκάστοτε συνιστώσα.
5. Επανάληψη βήματος 3,4 μέχρι την σύγκλισή τους.




**Εικόνα 8 - Μοντέλα Μίξης Γκαουσιανών Κατανομών (Πηγή: towardsdatascience.com)**

### 2.2.2.3. PCA

Η PCA είναι μια γραμμική τεχνική μείωσης διαστάσεων που χρησιμοποιείται για τη μετατροπή δεδομένων υψηλών διαστάσεων σε χώρο χαμηλότερης διάστασης, διατηρώντας παράλληλα τις πιο σημαντικές πληροφορίες. Η PCA ακολουθεί τα παρακάτω βήματα:

1. Αφαίρεση του μέσου όρου κάθε διάστασης από τις τιμές που απαρτίζουν τις διαστάσεις
2. Υπολογισμός του πίνακα συνδιακύμανσης
3. Υπολογισμός ιδιοτιμών και ιδιοδιανυσμάτων
4. Τα ιδιοδιανύσματα με την ψηλότερη τιμή ιδιοτιμών αποτελούν τις κύριες συνιστώσες. Για παράδειγμα αν θέλουμε να πάρουμε δύο κύριες συνιστώσες, τότε θα επιλέξουμε τα ιδιοδιανύσματα που έχουν τις δύο υψηλότερες ιδιοτιμές.
5. Τα νέα δεδομένα προκύπτουν από τον πολλαπλασιασμό των αρχικών δεδομένων με τα ιδιοδιανύσματα

### 2.2.2.3. t-SNE

Ο t-SNE είναι ένας αλγόριθμος που χρησιμοποιείται για τη μείωση των διαστάσεων ενός συνόλου δεδομένων. Σε αντίθεση με τον PCA, ο t-SNE είναι μη γραμμικός αλγόριθμος. Τα βήματα που ακολουθεί για να μείωση τις διαστάσεις είναι:

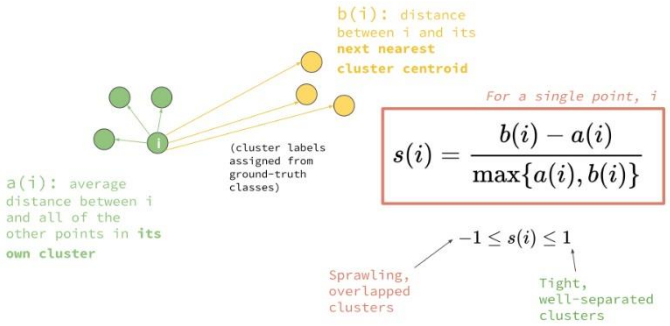
1. Υπολογισμός ομοιοτήτων: ο t-SNE αρχικά υπολογίζει τις ομοιότητες ανά ζεύγη μεταξύ δύο σημείων στον χώρο υψηλών διαστάσεων. Οι ομοιότητες βασίζονται σε εξισώσεις απόστασης, όπως η Ευκλείδεια απόσταση
2. Κατασκευή κατανομών πιθανοτήτων: έπειτα, ο t-SNE κατασκευάζει κατανομές πιθανοτήτων τόσο στον χώρο υψηλών διαστάσεων, όσο και στον χώρο χαμηλών διαστάσεων. Ορίζει μια κατανομή πιθανότητας στον χαμηλό χώρο, ώστε να ακολουθεί τις ομοιότητες ανά ζεύγη των σημείων των υψηλότερων διαστάσεων
3. Βελτιστοποίηση: το τελικό βήμα είναι η επαναληπτική προσαρμογή των θέσεων των σημείων στον χαμηλό χώρο διαστάσεων, με σκοπό να μειωθεί η διαφορά των ομοιοτήτων στον χώρο υψηλό διαστάσεων. Η διαδικασία της βελτιστοποίησης δημιουργεί μια αναπαράσταση χαμηλότερων διαστάσεων όπου παρόμοια σημεία βρίσκονται κοντά μεταξύ τους.

### 2.2.2.4. Μέθοδος Αξιολόγησης - Βαθμολογία Silhouette

Η βαθμολογία Silhouette είναι μια μέτρηση που χρησιμοποιείται για την αξιολόγηση των αποτελεσμάτων στους αλγόριθμους ομαδοποίησης. Μετρά τον βαθμό στον οποίο κάθε σημείο είναι καλά κατανεμημένο στην κλάση που του έχει ανατεθεί, ενώ ταυτόχρονα διαχωρίζεται από γειτονικές κλάσεις. Για κάθε σημείο υπολογίζονται δύο τιμές:

- η τιμή a, η οποία είναι η μέση απόσταση κάθε σημείου μίας κλάσης με τα υπόλοιπα σημεία που βρίσκονται στην ίδια κλάση
- η τιμή b, η οποία είναι η μέση απόσταση κάθε σημείου μίας κλάσης με τα υπόλοιπα σημεία που βρίσκονται στην πιο κοντινή κλάση

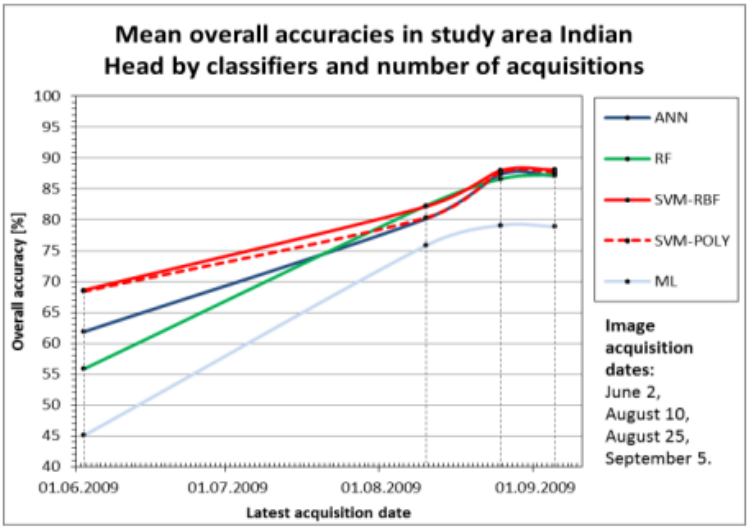
Στη συνέχεια προκύπτει η βαθμολογία silhouette με χρήση της εξίσωσης $\frac{b-a}{\max(a,b)}$, δηλαδή αφαιρώντας το a από το b και διαιρώντας το με την μέγιστη τιμή από τα a,b. Η βαθμολογία κυμαίνεται από -1 έως 1, όπου τιμές κοντά στο 1 υποδηλώνουν πως το σημείο είναι καλά διαχωρισμένο στην κλάση, ενώ τιμές κοντά στο -1 αναδεικνύουν σημεία που έχουν ταξινομηθεί λάθος. Η συνολική βαθμολογία silhouette είναι ο μέσος όρος όλων των βαθμολογιών για όλα τα σημεία, παρέχοντας μια συνολική αξιολόγηση της ταξινόμησης.




**Εικόνα 9 - Βαθμολογία Silhouette (Πηγή: platform.ai)**

Στο άρθρο [1], χρησιμοποιήθηκε ο αλγόριθμος ομαδοποίησης K-means για την ανάλυση κάλυψης γης στην Ινδία. Συγκεκριμένα, εφαρμόστηκε ο αλγόριθμος K-means σε δορυφορικές εικόνες για να ταξινομήσουν τις δορυφορικές εικόνες Resourcesat-1 LISS III και Landsat-8, προσδιορίζοντας δασικές και μη δασικές περιοχές. Η μέθοδος ομαδοποίησης K-means βοήθησε στην ομαδοποίηση των εικονοστοιχείων με βάση την ομοιότητά τους. Τα καλύτερα αποτελέσματα τα πέτυχαν συνδυάζοντας την ομαδοποίηση K-means με μια εποπτευόμενη μέθοδο ταξινόμησης Maximum Likelihood.

Στο συγκεκριμένο άρθρο[2], συγκρίνονται παραδοσιακές και μη-παραδοσιακές μέθοδοι ταξινόμησης καλλιεργειών. Συγκεκριμένα, χρησιμοποιούνται υπερφασματικά δεδομένα από τον δορυφόρο RapidEye για μία περιοχή στον Καναδά. Οι αλγόριθμοι που χρησιμοποιούνται είναι ο Maximum Likelihood(ML) ως παραδοσιακή τεχνική και οι Random Forest (RF), Support Vector Machine Radial-Basic-Function (SVM-RBF), ο SVM Polyonimal Kernel (SVM-POLY) και Artificial Neural Network (ANN) ως πιο σύγχρονες τεχνικές. Τα αποτελέσματα δείχνουν πως οι μη-παραδοσιακές μέθοδοι έχουν καλύτερα αποτελέσματα, καθώς μπορούν να χειριστούν με μεγαλύτερη ευκολία μη-γραμμικές σχέσεις και να παρέχουν καλύτερες γενικεύσεις των μοντέλων.




**Εικόνα 10 - Σύγκριση παραδοσιακών και μη-παραδοσιακών τεχνικών ταξινόμησης (Πηγή: Nitze, Ingmar & Schulthess, Urs & Asche, H. (2012). Comparison of machine learning algorithms Random Forest, Artificial Neural Network and Support Vector Machine to Maximum Likelihood for supervised crop type classification**

Στο άρθρο [3] τονίζεται η βελτίωση των αποτελεσμάτων ταξινόμησης, όταν συνδυάζονται δεδομένα από τους δορυφόρους Sentinel-1 και Sentinel-2. Αναλυτικότερα, ο συνδυασμός των δεδομένων βελτίωσε τα αποτελέσματα στον Hierarchical Random Forest. Η ενσωμάτωση, τόσο δεδομένων Sentinel-1 όσο και Sentinel-2, επέτρεψε μια πιο ολοκληρωμένη κατανόηση της δομής των καλλιεργειών και των διαφορών στη φαινολογία. Αξιοποιώντας τα πλεονεκτήματα και των δύο δορυφόρων, συμπεριλαμβανομένης της ικανότητας του ραντάρ να διεισδύει μέσα από τα σύννεφα και των φασματικών

πληροφοριών του Sentinel-2, η μελέτη πέτυχε υψηλότερη ακρίβεια ταξινόμησης σε σύγκριση με την χρήση δεδομένων από Sentinel-1 ή Sentinel-2 δεδομένα.

| | Sentinel 2 | | | | | Sentinel 1 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # | March | April | May June | July | August | March | April | May June | July | August | OA | κ |
| 1 | | | | | | X | | | | | 0.44 | 0.33 |
| 2 | | | | | | X | X | | | | 0.60 | 0.52 |
| 3 | | | | | | X | X | X | | | 0.72 | 0.61 |
| 4 | | | | | | X | X | X X | | | 0.73 | 0.70 |
| 5 | | | | | | X | X | X X | X | | 0.72 | 0.64 |
| 6 | | | | | | X | X | X X | X | X | 0.78 | 0.70 |
| 7 | X | | | | | | | | | | 0.42 | 0.30 |
| 8 | X | X | | | | | | | | | 0.56 | 0.42 |
| 9 | X | X | X | | | | | | | | 0.70 | 0.57 |
| 10 | X | X | X X | | | | | | | | 0.73 | 0.65 |
| 11 | X | X | X X | X | | | | | | | 0.74 | 0.72 |
| 12 | X | X | X X | X | X | | | | | | 0.80 | 0.73 |
| 13 | X | | | | | X | | | | | 0.55 | 0.42 |
| 14 | X | X | | | | X | X | | | | 0.67 | 0.57 |
| 15 | X | X | X | | | X | X | X | | | 0.74 | 0.67 |
| 16 | X | X | X X | | | X | X | X X | | | 0.81 | 0.74 |
| 17 | X | X | X X | X | | X | X | X X | X | | 0.83 | 0.80 |
| 18 | X | X | X X | X | X | X | X | X X | X | X | 0.84 | 0.79 |

**Εικόνα 11 - Ακρίβεια (OA) για διαφορετικό συνδυασμό Sentinel-1 και Sentinel 2 δεδομένων (Πηγή: Felegari, S.; Sharifi, A.; Moravej, K.; Amin, M.; Golchin, A.; Muzirafuti, A.; Tariq, A.; Zhao, N. Integration of Sentinel 1 and Sentinel 2 Satellite Images for Crop Mapping**

Όπως τονίζεται και στο άρθρο [4], ο Random Forest έχει σημαντικό πλεονέκτημα έναντι των κλασσικών μεθόδων, όπως ο Maximum Likelihood, καθώς αλγόριθμοι σαν τον ML κάνουν υποθέσεις κανονικών κατανομών στα δεδομένα, με αποτέλεσμα να μην λαμβάνεται υπόψιν αν τα δεδομένα έχουν ισορροπία ως προς την κατανομή τους.

Η συγκεκριμένη μελέτη [5] εστιάζει στη βελτίωση της ακρίβειας χαρτογράφησης των καλλιεργειών μέσω της χρήσης τεχνικών oversampling σε σύνολα δεδομένων που δεν είναι ισορροπημένα. Αρχικά, επιλέχθηκαν 15 χαρακτηριστικά ύστερα από τεχνικές σημαντικότητας χαρακτηριστικών και στη συνέχεια εφαρμόστηκαν οι τεχνικές oversampling SMOTE, Borderline-SMOTE, SMOTE-ENN και Distance-SMOTE. Τέλος, στο αρχικό σύνολο δεδομένων και στα καινούρια σύνολα δεδομένων με χρήση τεχνικών oversampling εφαρμόστηκε ο αλγόριθμος Gray-Wolf-Optimizer SVM και έδειξε βελτιωμένα αποτελέσματα.

| Oversampling Technology | | Raw Data | Smote | Smote-enn | Borderline-smote1 | Borderline-smote2 | Distance-smote |
|---|---|---|---|---|---|---|---|
| PA | Wheat | 0.96 | 0.98 | 0.97 | 0.98 | 0.98 | 0.99 |
| | Rape | 0.79 | 0.90 | 0.85 | 0.93 | 0.92 | 0.91 |
| | Woodland | 0.76 | 0.81 | 0.75 | 0.75 | 0.78 | 0.82 |
| UA | Wheat | 0.95 | 0.91 | 0.96 | 0.93 | 0.95 | 0.98 |
| | Rape | 0.93 | 0.99 | 0.93 | 1.00 | 0.97 | 0.98 |
| | Woodland | 0.59 | 0.77 | 0.68 | 0.71 | 0.70 | 0.93 |
| F1 score | Wheat | 0.96 | 0.95 | 0.96 | 0.97 | 0.95 | 0.96 |
| | Rape | 0.85 | 0.86 | 0.81 | 0.92 | 0.91 | 0.90 |
| | Woodland | 0.67 | 0.86 | 0.71 | 0.83 | 0.78 | 0.84 |
| Accuracy (%) | | 0.8940 | 0.9224 | 0.9206 | 0.9334 | 0.9358 | 0.9636 |

**Εικόνα 12 - Ακρίβεια (User Accuracy), Ανάκληση (Producer Accuracy PA) και F1-score για δεδομένα ύστερα από oversampling τεχνικές (Πηγή: Zhang, H.; Gao, M.; Ren, C. Feature-Ensemble-Based Crop Mapping for Multi-Temporal Sentinel-2 Data Using Oversampling Algorithms and Gray Wolf Optimizer Support Vector Machine)**

# Κεφάλαιο 4 – Μεθοδολογία

Στο συγκεκριμένο κεφάλαιο παρουσιάζεται η μεθοδολογία που ακολουθήθηκε για την αξιολόγηση διαφορετικών τεχνικών με σκοπό την ταξινόμηση καλλιεργειών με χρήση χρονοσειρών Sentinel-1 και Sentinel-2 ακολουθήθηκαν.

Αρχικά, παρουσιάζεται η περιοχή μελέτης με γενικές πληροφορίες για το κλίμα και τη βλάστηση. Στη συνέχεια, αναλύονται τα δεδομένα που θα χρησιμοποιηθούν, τόσο τα επιτόπια με τις κατηγορίες των καλλιεργειών, όσο και τα δορυφορικά δεδομένα από τους δορυφόρους Sentinel-1 και Sentinel-2 και περιγράφονται τα βήματα προ-επεξεργασίας των δορυφορικών δεδομένων. Τέλος, παρουσιάζονται οι αλγόριθμοι μη-επιβλεπόμενης και επιβλεπόμενης μάθησης που θα χρησιμοποιηθούν και οι τεχνικές εξαγωγής σημαντικών χαρακτηριστικών και μεταφοράς γνώσης.

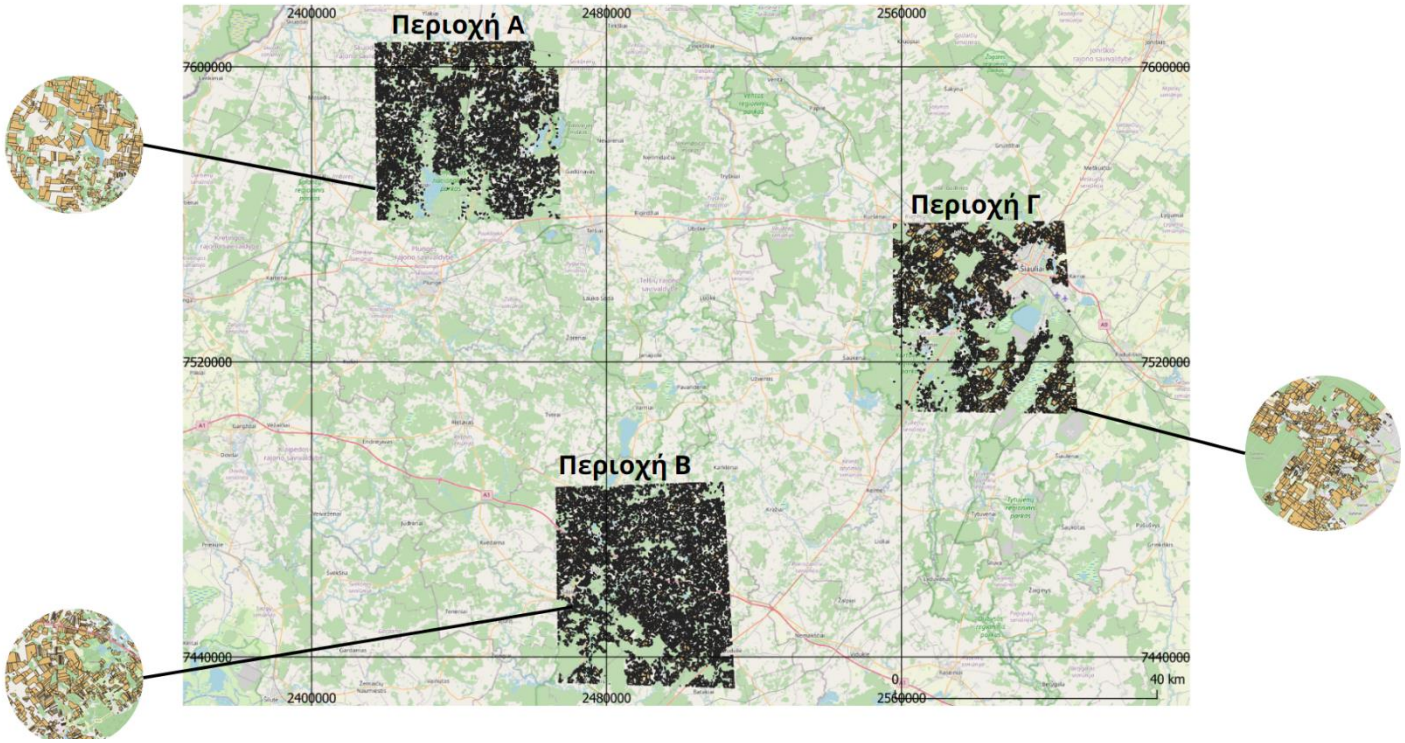
Όλη η επεξεργασία των δεδομένων και των τεχνικών μη-επιβλεπόμενων και επιβλεπόμενων αλγορίθμων πραγματοποιήθηκε σε περιβάλλον Jupyter Notebook με χρήση της γλώσσας προγραμματισμού Python.

## 4.1 Περιοχή Μελέτης

Η μελέτη επικεντρώνεται στη Λιθουανία, μια χώρα που βρίσκεται στη Βορειοανατολική Ευρώπη και καλύπτει μια έκταση 65300 km². Αποτελεί ένα από τα τρία Βαλτικά κράτη και έχει πληθυσμό 2.8 εκατομμύρια σύμφωνα με τα δεδομένα από την εθνική στατιστική υπηρεσία (stat.gov.lt) για τον Ιούλιο του 2022. Η Λιθουανία διαθέτει ένα εύκρατο κλίμα που επηρεάζεται τόσο από θαλάσσιους όσο και από ηπειρωτικούς παράγοντες. Σύμφωνα με την κατηγοριοποίηση του Koppen, το κλίμα της χώρας χαρακτηρίζεται κυρίως ως υγρό ηπειρωτικό. Ωστόσο, η παράκτια ζώνη που βρίσκεται δίπλα στη Βαλτική Θάλασσα παρουσιάζει χαρακτηριστικά που μπορούν να χαρακτηριστούν ωκεάνια, με χαμηλές θερμοκρασίες, συχνές βροχοπτώσεις και χιονισμένους χειμώνες.

Οι δασώδεις εκτάσεις καταλαμβάνουν το 1/3 της χώρας, ενώ τα λιβάδια και τα σιτηρά αποτελούν τους κυρίαρχους τύπους καλλιέργειας στη Λιθουανία. Τα λιβάδια καταλαμβάνουν σχεδόν το 50% των αγρών, με διαχωρισμό σε μόνιμα και προσωρινά λιβάδια. Τα μόνιμα λιβάδια περιλαμβάνουν πολυετή βοσκοτόπια και άλλα φυσικά ή ημι-φυσικά λιβάδια που χρησιμοποιούνται για πάνω από πέντε χρόνια, ενώ τα προσωρινά λιβάδια αναφέρονται σε πιο πρόσφατες εκμεταλλεύσεις γης με διάρκεια μικρότερη από πέντε χρόνια.

Συγκεκριμένα, τα δεδομένα που χρησιμοποιήθηκαν αφορούν τρεις περιοχές. Η Περιοχή Α βρίσκεται στο γεωγραφικό διαμέρισμα Τελσιάι, η Περιοχή Β βρίσκεται στο γεωγραφικό διαμέρισμα Σιλάλε και η Περιοχή Γ βρίσκεται στο γεωγραφικό διαμέρισμα Σιαουλιάϊ.



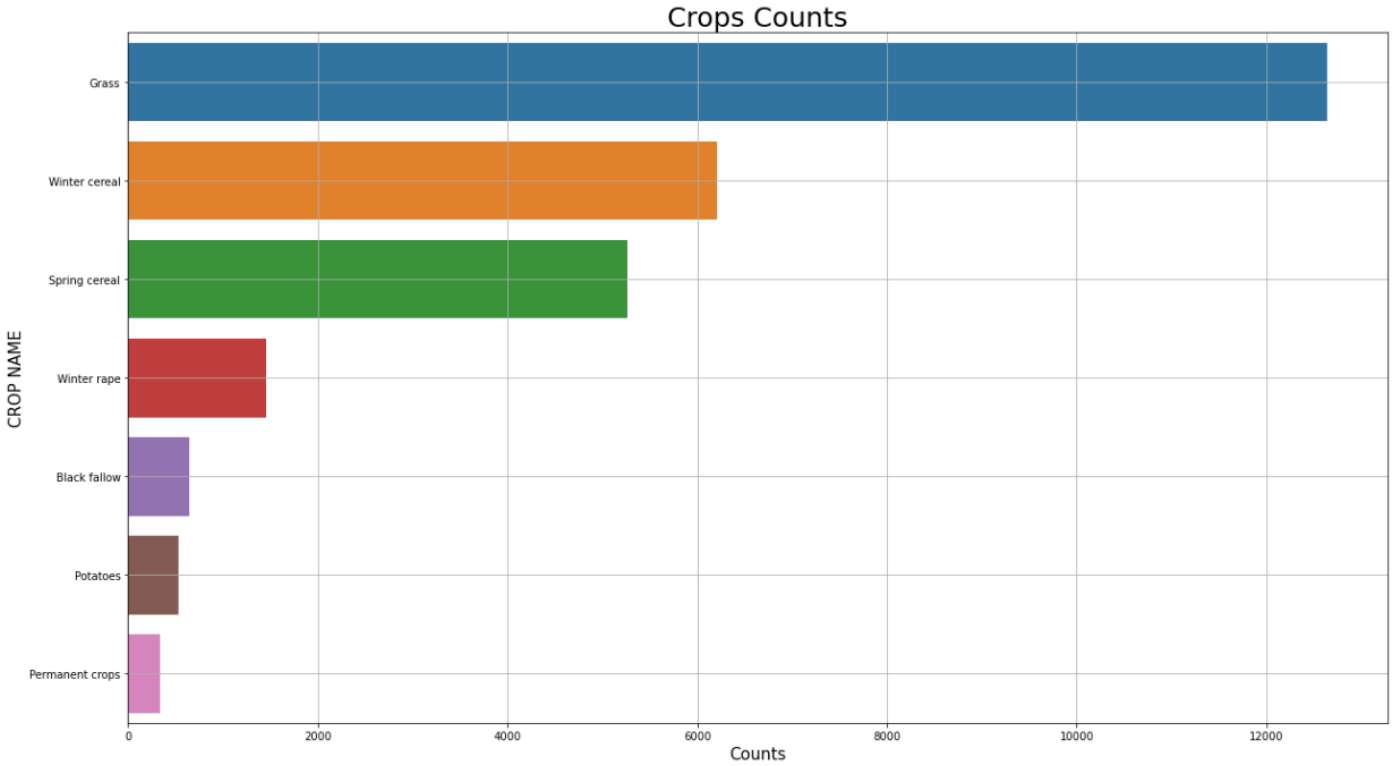

**Εικόνα 13 - Περιοχή Μελέτης: Λιθουανία**

## 4.2 Επιτόπια Δεδομένα

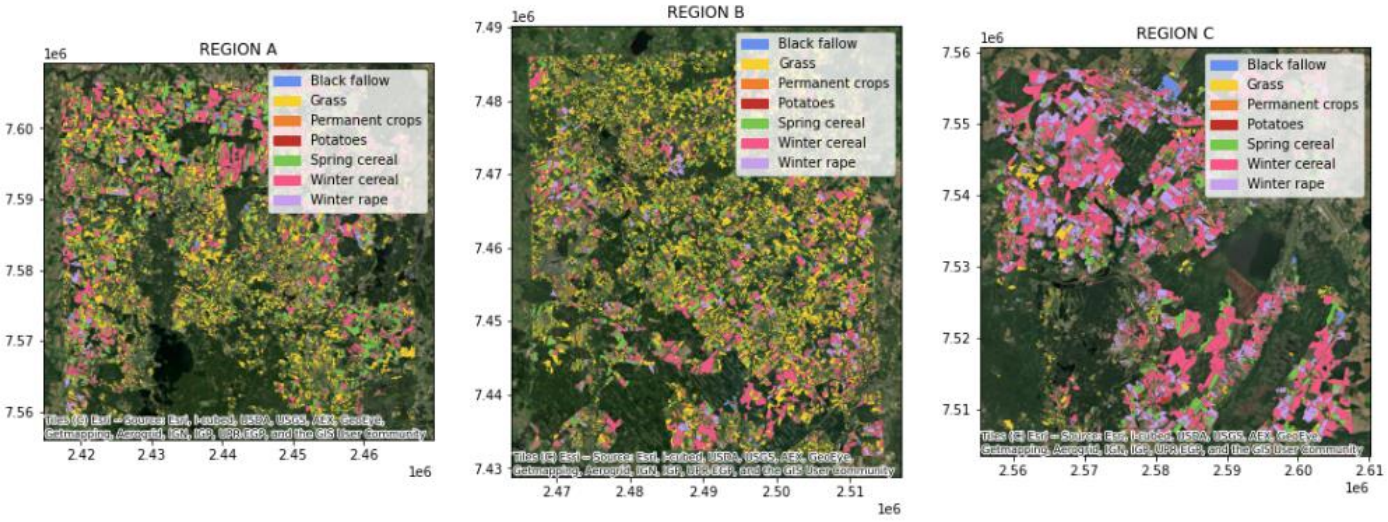
Στο πλαίσιο της συγκεκριμένης διπλωματικής εργασίας, χρησιμοποιήθηκε ένα σύνολο δεδομένων που αποτελείται από 27073 αγροτεμάχια, τα οποία προέρχονται από τον Οργανισμό Γεωργικών Πληρωμών της Λιθουανίας (NPA) και ,συγκεκριμένα, έχουν προκύψει ύστερα από δηλώσεις αγροτών, στο πλαίσιο του ερευνητικού προγράμματος ENVISION H2020. Τα επιτόπια δεδομένα είναι καταμερισμένα στις τρεις περιοχές που αναφέρθηκαν στην παραπάνω παράγραφο. Μέσω της δήλωσης των αγροτών παρέχονται πληροφορίες για την κατηγορία κάθε καλλιέργειας. Συνολικά υπάρχουν 7 κατηγορίες καλλιεργειών στα δεδομένα:

- Αγρανάπαυση,
- Γρασίδι,
- Μόνιμες Καλλιέργειες,
- Πατάτες,
- Ανοιξιάτικες Καλλιέργειες Σιτηρών,
- Χειμερινές Καλλιέργειες Σιτηρών και
- Ελαιοκράμβη

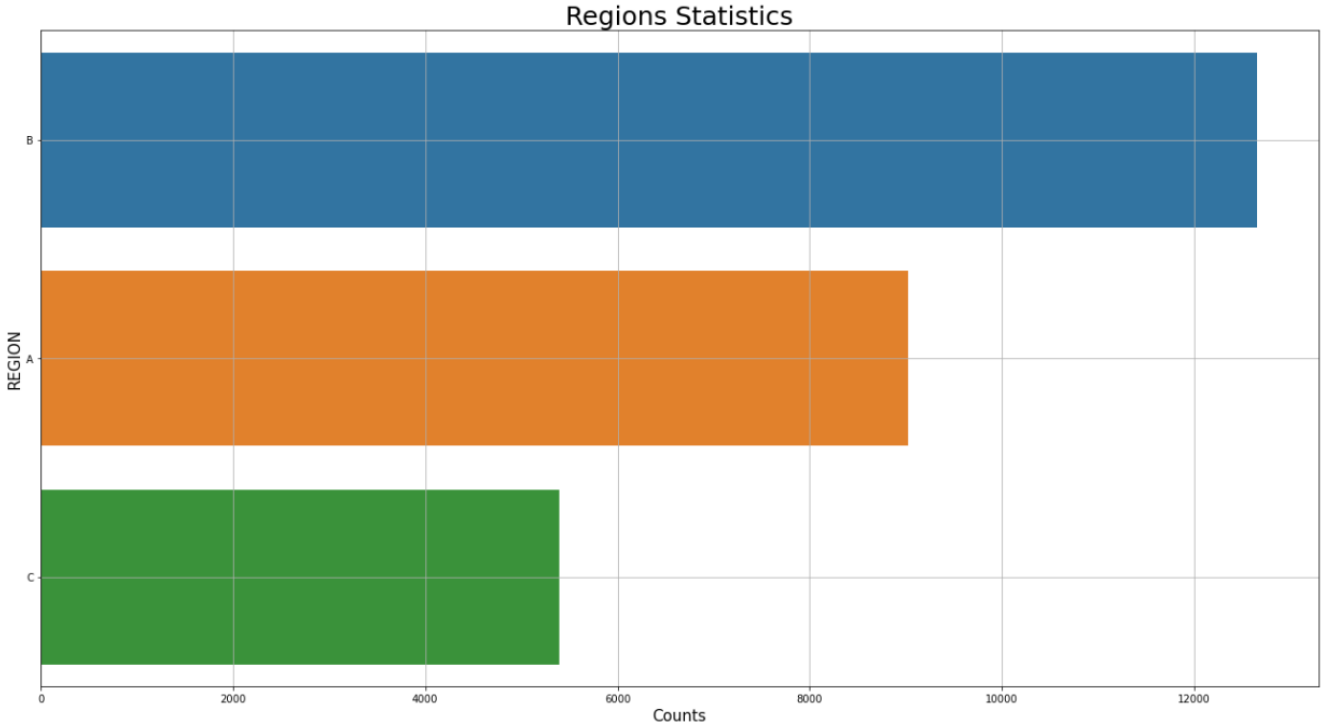
Για να γίνει πιο αντιληπτή η κατανομή του συνόλου δεδομένων, πραγματοποιήθηκε ανάλυση που περιλαμβάνει στοιχεία, τα οποία απεικονίζουν το συνολικό άθροισμα ανά καλλιέργεια, στατιστικά ανά περιοχή που δείχνουν τον συνολικό αριθμό των καλλιεργειών, καθώς και χάρτες που αντιπροσωπεύουν τις καλλιέργειες που υπάρχουν σε κάθε περιοχή.



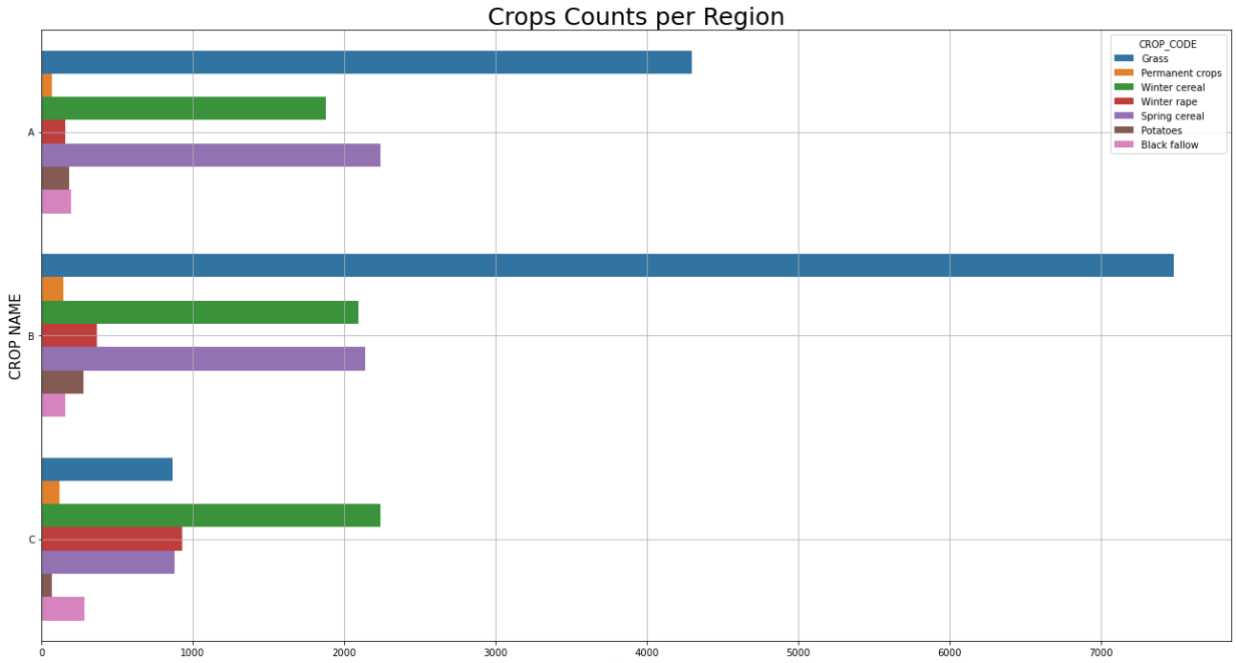

**Εικόνα 14 - Κατανομή δειγμάτων καλλιεργειών**




**Εικόνα 17 - Χάρτες καλλιεργειών ανά περιοχή**




**Εικόνα 15 - Αριθμός δειγμάτων ανά περιοχή**




**Εικόνα 16 - Αριθμός δειγμάτων ανά καλλιέργεια σε κάθε περιοχή**

## 4.3 Δορυφορικά δεδομένα και προ-επεξεργασία

Για την εφαρμογή και αξιολόγηση μεθόδων ταξινόμησης των καλλιεργειών χρησιμοποιήθηκαν δορυφορικά δεδομένα από τους δορυφόρους Sentinel-1 και Sentinel-2 για την περίοδο Μάρτιος 2022 – Σεπτέμβρης 2022. Η προ-επεξεργασία των δεδομένων αυτών προέκυψε από το ερευνητικό πρόγραμμα ENVISION H2020 της Επιχειρησιακής Μονάδα BEYOND του Εθνικού Αστεροσκοπείου Αθηνών, [6][7].

Από τον δορυφόρο Sentinel-2 χρησιμοποιήθηκαν τα παρακάτω tiles:

- 34VEH: 91 Εικόνες
- 34VFH: 61 Εικόνες
- 34UEG: 62 Εικόνες
- 34UFG: 92 Εικόνες


**Εικόνα 17 - Sentinel-2 Tiles**

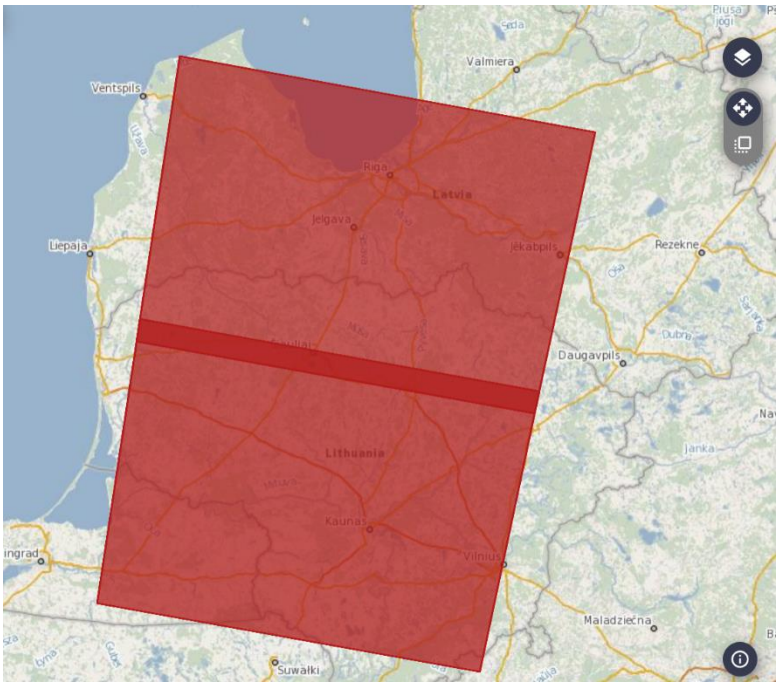
Στα δεδομένα Sentinel-2 πραγματοποιήθηκε ατμοσφαιρική διόρθωση και από Top-of-Atmosphere L1C προϊόντα έγιναν Bottom-of-Atmosphere L2A προϊόντα. Στη συνέχεια δημιουργήθηκε ένας κύβος δεδομένων (data cube), όπου όλα τα δεδομένα (S1 και S2) αναφέρονταν στις ημερομηνίες των απεικονίσεων Sentinel-1 τροχιάς 131, με τεχνικές παρεμβολής, έγινε αφαίρεση των περιοχών με νεφοκάλυψη και παρεμβολή για την ανάκτηση των τιμών και συμπροσαρμογή των εικόνων [6][7]. Ετσι υπήρξε κοινό χρονικό βήμα 10 ημερών για όλες τις απεικονίσεις. Η ημερομηνία του συνόλου των δεδομένων ξεκινάει από τις 2 Μαρτίου, συνεχίζει 12 Μαρτίου, 22 Μαρτίου κοκ.

Έπειτα έγινε η επιλογή των μπαντών και ο υπολογισμός των δεικτών που θα βρίσκονται στο σύνολο των δεδομένων. Στον παρακάτω πίνακα φαίνεται ποια χαρακτηριστικά του Sentinel-2 λήφθηκαν υπόψιν στη δημιουργία των δεδομένων.

| Χαρακτηριστικά που χρησιμοποιήθηκαν | Εξισώσεις |
|---|---|
| B02 | |
| B03 | |
| B04 | |
| B05 | |
| B06 | |
| B07 | |
| B08 | |
| B08A | |
| B11 | |
| B12 | |
| NDVI | $\frac{B08 - B04}{B08 + B04}$ |
| NDWI | $\frac{B04 - B11}{B04 + B11}$ |
| NDMI | $\frac{B08 - B11}{B08 + B11}$ |
| TVI | $\sqrt{NDVI + 0.5}$ |
| EVI | $2.5 * \frac{B08 - B04}{B08 + 6 * B04 - 7.5 * B02 + 1}$ |
| SAVI | $1.5 * \frac{B08 - B04}{B08 + B04 + 0.5}$ |
| BSI | $2.5 * \frac{(B11 + B04) - (B08 + B02)}{(B11 + B04) + (B08 + B02)}$ |
| PSRI | $\frac{B04 - B02}{B06}$ |

Σχετικά με τον δορυφόρο Sentinel-1, χρησιμοποιήθηκε η σχετική τροχιά 153 και έγινε λήψη:

- Περιοχή Α+Γ: 17 εικόνες
- Περιοχή Β: 17 εικόνες




**Εικόνα 181 - Sentinel-1 Relative Orbit**

Καθώς από τον Δεκέμβρη του 2021 ο Sentinel-1B έχει τεθεί εκτός λειτουργίας, τα δεδομένα για την περιοχή μελέτης είναι διαθέσιμα ανά 12 μέρες.

Τα δεδομένα Sentinel-1 που λήφθηκαν είναι Level-1 Ground-Range-Detected (GRD), τα οποία επεξεργάστηκαν με τη βιβλιοθήκη snappy. Αρχικά έγινε περικοπή των εικόνων στην περιοχή μελέτης για τη μείωση του όγκου των δεδομένων και ακολούθησε η ραδιομετρική διόρθωση. Έπειτα, εφαρμόστηκε το φίλτρο Refined-Lee για τη μείωση της κηλίδωσης και πραγματοποιήθηκε γεωαναφορά των εικόνων με βάση την αποστολή Shuttle Radar Topography (SRTM). Τέλος, οι τιμές οπισθοσκέδασης που υπολογίστηκαν, μετατράπηκαν σε decibels (dB).

Στο σύνολο των δεδομένων χρησιμοποιήθηκαν οι παρακάτω τιμές και δείκτες από τον δορυφόρο Sentinel-1.

| **Χαρακτηριστικά που χρησιμοποιήθηκαν** | **Εξισώσεις** |
|---|---|
| Οπισθοσκέδαση VV – $\sigma_{o\,VV}$ (dB) | |
| Οπισθοσκέδαση VH – $\sigma_{o\,VH}$ (dB) | |
| Radar Vegetation Index - RVI | $\frac{(4 * \sigma_{0VH})}{\sigma_{0VH} + \sigma_{0VV}}$ |

Αφού έγινε η προ-επεξεργασία των δορυφορικών δεδομένων και υπολογίστηκαν τα 21 χαρακτηριστικά που θα χρησιμοποιηθούν για τις τεχνικές μη-επιβλεπόμενης και επιβλεπόμενης μάθησης, δημιουργήθηκε ένας κύβος δεδομένων με διαστάσεις 27073 X 471. Κάθε γραμμή απεικονίζει το εκάστοτε αγροτεμάχιο και κάθε στήλη απεικονίζει την κατηγορία κάθε αγροτεμαχίου (1 στήλη), κάποιες πληροφορίες για τα αγροτεμάχια (έκταση, περιοχή, γεωμετρία κλπ.) (8 στήλες) και τις τιμές των μπαντών και των υπολογισμένων δεικτών (21 τιμές * 22 ημερομηνίες = 462 στήλες).

## 4.4 Μη Επιβλεπόμενη Ταξινόμηση

Σε αυτήν την ενότητα, το σύνολο δεδομένων καλλιεργειών υποβλήθηκε σε τεχνικές μη επιβλεπόμενης ταξινόμησης με σκοπό να γίνει εύρεση φασματικών συσσωρεύσεων χωρίς τη χρήση γνώσης για τον πραγματικό αριθμό σχετικά με τις κατηγορίες των καλλιεργειών. Πέντε διαφορετικά σύνολα χαρακτηριστικών εξετάστηκαν για ανάλυση: ο Κανονικοποιημένος Δείκτης Βλάστησης (NDVI), ο Δείκτης Βλάστησης Ραντάρ (RVI) και δεδομένα ύστερα από τεχνικές μείωσης διαστάσεων με τους αλγόριθμους PCA και t-SNE στους δείκτες NDVI, RVI και σε όλο το σύνολο των δεδομένων. Ο NDVI και ο RVI περιέχουν 22 τιμές για κάθε αγροτεμάχιο, και όλο το σύνολο των δεδομένων περιέχει 462 τιμές (22 τιμές X 21 χαρακτηριστικά).

Για να διερευνηθούν οι ομάδες και οι ομοιότητες στα δείγματα των καλλιεργειών, εφαρμόστηκαν διάφοροι αλγόριθμοι μη επιβλεπόμενης ταξινόμησης, όπως ο k-means, ο Gaussian Mixture Model (GMM) και εφαρμόστηκαν οι τεχνικές μείωσης διαστάσεων PCA και t-SNE.

Για κάθε σύνολο χαρακτηριστικών, πραγματοποιήθηκαν οι παραπάνω αλγόριθμοι με διαφορετικές τιμές n (αριθμός κλάσεων) που κυμαίνονται από 2 έως 10. Οι πραγματικές κατηγορίες καλλιέργειας είναι 7, ωστόσο για τη διερεύνηση και την αξιολόγηση διαφορετικών αλγορίθμων επιλέχθηκε ένα μεγαλύτερο εύρος κλάσεων. Η βαθμολογία silhouette χρησιμοποιήθηκε ως μέτρο αξιολόγησης, υποδεικνύοντας τη συνοχή και τον διαχωρισμό των κλάσεων. Αναλύοντας τις βαθμολογίες silhouette και τις υπολογισμένες κλάσεις, στόχος είναι να προσδιοριστεί ο βέλτιστος αριθμό κλάσεων, χωρίς να λαμβάνεται υπόψιν ο πραγματικός αριθμός των διαφορετικών κατηγοριών καλλιεργειών.

## 4.5 Επιβλεπόμενη Ταξινόμηση

Σε αυτήν την ενότητα, δόθηκε έμφαση σε τεχνικές επιβλεπόμενης ταξινόμησης για τις καλλιέργειες, λαμβάνοντας υπόψιν τις ετικέτες κάθε αγροτεμαχίου και για τις τρεις περιοχές. Το σύνολο δεδομένων χωρίστηκε σε ένα σύνολο εκπαίδευσης (70%) και ένα σύνολο δοκιμών (30%) για την αξιολόγηση της απόδοσης των εκπαιδευμένων μοντέλων. Λαμβάνοντας υπόψιν την υψηλή ανισορροπία της κατανομής των κατηγοριών, χρησιμοποιήθηκαν τεχνικές weights_balanced, ώστε να δοθούν μεγαλύτερα βάρη στις κατηγορίες με λίγα δείγματα. Από τα 27063 αγροτεμάχια, τα 18951 χρησιμοποιήθηκαν για την εκπαίδευση των μοντέλων και τα υπόλοιπα 8122 χρησιμοποιήθηκαν για την αξιολόγηση των μοντέλων. Για κάθε αγροτεμάχιο χρησιμοποιήθηκε όλο το σύνολο των δεδομένων, 21 δείκτες και μπάντες για 22 ημερομηνίες, δηλαδή 462 τιμές συνολικά.

Χρησιμοποιήθηκαν τρεις δημοφιλείς αλγόριθμοι επιβλεπόμενης μάθησης: Random Forest, Linear Support Vector Machine (Linear SVM) και Gradient Boosting. Ο λόγος που επιλέχθηκαν οι τρεις αυτοί αλγόριθμοι είναι διότι ο Random Forest προσφέρει καλά αποτελέσματα στον χειρισμό δεδομένων υψηλών διαστάσεων, ο Linear SVM για την αποτελεσματικότητά τους σε δεδομένα με γραμμική σχέση και ο αλγόριθμος Gradient Boosting για την αποτελεσματικότητά του σε δεδομένα που οι συναρτήσεις απόφασης για τον διαχωρισμό τους σε κατηγορίες δεν είναι γραμμικές.

Για την αξιολόγηση κάθε αλγορίθμου αναλύεται η αξιολόγηση της ταξινόμησης, στην οποία υπάρχουν τιμές ακρίβειας, ανάκλησης και f1-score και ο πίνακας σύγχυσης (confusion matrix), στον οποίο θα φαίνονται πόσες περιπτώσεις ταξινομήθηκαν σωστά, αλλά και σε πόσες περιπτώσεις υπήρξε λάθος ταξινόμηση-σύγχυση με άλλες κατηγορίες.

Σχετικά με τον αλγόριθμο Random Forest έγινε παραμετροποίηση, όπως παρουσιάζεται παρακάτω:

- **n_estimators**: Χρησιμοποιήθηκαν τρεις διαφορετικές τιμές - 10, 50 και 100. Αυτή η παράμετρος ελέγχει τον αριθμό των δέντρων απόφασης στο σύνολο τυχαίων δασών. Μεταβάλλοντας τον αριθμό των εκτιμητών, μπορεί να εκτιμηθεί η αλλαγή της απόδοση του μοντέλου ανάλογα με την πολυπλοκότητα του συνόλου.
- **min_samples_split**: Για να προσδιορίσουμε τη βέλτιστη τιμή για αυτήν την παράμετρο, δοκιμάζουμε δύο τιμές - 2 και 5. Η παράμετρος καθορίζει τον ελάχιστο αριθμό δειγμάτων που απαιτούνται για τη διαίρεση ενός εσωτερικού κόμβου κατά τη διαδικασία δημιουργίας δέντρων. Προσαρμόζοντας αυτήν την τιμή, μπορούμε να επηρεάσουμε το βάθος του δέντρου και να αποτρέψουμε την υπερπροσαρμογή.
- **max_depth**: Το μέγιστο βάθος κάθε δέντρου απόφασης ορίστηκε στο 20. Αυτή η παράμετρος περιορίζει το βάθος των δέντρων στο σύνολο. Ελέγχοντας το βάθος του δέντρου, επιτυγχάνουμε μια ισορροπία μεταξύ της αποτύπωσης περίπλοκων σχέσεων και της αποφυγής της υπερπροσαρμογής.
- **min_samples_leaf**: Η παράμετρος ορίστηκε σε 5. Αυτή η παράμετρος καθορίζει τον ελάχιστο αριθμό δειγμάτων που απαιτείται να βρίσκονται σε έναν κόμβο φύλλου. Η ρύθμιση μιας υψηλότερης τιμής βοηθά στην αποφυγή υπερπροσαρμογής διασφαλίζοντας ότι κάθε φύλλο περιέχει επαρκή αριθμό δειγμάτων.

Σχετικά με τον αλγόριθμος Linear SVM έγινε παραμετροποίηση στα::

- **C**: Οι τιμές που χρησιμοποιήθηκαν είναι 0.2, 0.5, 1, 2, 5 και 10. Η τιμή C καθορίζει την ποινή για εσφαλμένη ταξινόμηση των σημείων. Οι μικρότερες τιμές του C δημιουργούν ένα μεγαλύτερο περιθώριο ταξινόμησης, επιτρέποντας περισσότερα εσφαλμένα αποτελέσματα, ενώ οι μεγαλύτερες τιμές του C στοχεύουν σε αυστηρότερη ταξινόμηση.

Η παραμετροποίηση στον αλγόριθμο Gradient Boosting, πραγματοποιήθηκε με τις παρακάτω επιλογές.

- **n_estimators**: Επιλέχθηκαν τρεις διαφορετικές τιμές: 10, 50 και 100. Αυτή η παράμετρος καθορίζει τον αριθμό των ασθενών μαθητών (weak learner), δηλαδή των δέντρων απόφασης (decision trees). Οι χαμηλότερες τιμές των εκτιμητών στην ενίσχυση της κλίσης συνεπάγονται ένα απλούστερο μοντέλο με μειωμένη πολυπλοκότητα, ενώ υψηλότερες τιμές υποδεικνύουν ένα πιο σύνθετο μοντέλο, αλλά υψηλότερο κίνδυνο υπερπροσαρμογής των δεδομένων.
- **learning_rate**: Για τη συγκεκριμένη παράμετρο έγινε επιλογή τριών ποσοστών μάθησης: 0.1, 0.5 και 1.0. Ο ρυθμός εκμάθησης, ελέγχει το μέγεθος βήματος στο οποίο ο αλγόριθμος προσαρμόζει τις προβλέψεις του μοντέλου κατά τη διάρκεια κάθε επανάληψης. Ένας μικρότερος ρυθμός εκμάθησης οδηγεί σε πιο συντηρητικές προσαρμογές, ενώ ένας μεγαλύτερος ρυθμός εκμάθησης επιτρέπει πιο έντονες προσαρμογές, επηρεάζοντας την ταχύτητα σύγκλισης και την πιθανότητα υπερπροσαρμογής.

Το καλύτερο μοντέλο επιλέγεται με βάση το ανάλυση απόδοσης του μοντέλου και τον πίνακα σύγχυσης. Οι μετρήσεις απόδοσης όπως η ακρίβεια, η ανάκληση και τοF1-score χρησιμοποιούνται για τον προσδιορισμό του καλύτερου μοντέλου για την ταξινόμηση των καλλιεργειών.

## 4.6 Εξαγωγή Σημαντικών Χαρακτηριστικών

Αφού επιλεχθεί το μοντέλο με τα καλύτερα αποτελέσματα, στη συνέχεια θα δοθεί έμφαση στον εντοπισμό των πιο σημαντικών χαρακτηριστικών για την ταξινόμηση των καλλιεργειών. Αυτό το βήμα έχει στόχο να αποκτηθούν γνώσεις σχετικά με τη σημασία των διαφορετικών χαρακτηριστικών για την ακριβή πρόβλεψη των καλλιεργειών.

Η σημαντικότητα των χαρακτηριστικών στα τυχαία δάση υπολογίζεται με την αξιολόγηση της μείωσης του impurity που επιτυγχάνεται σε κάθε κόμβο όταν χρησιμοποιείται ένα συγκεκριμένο χαρακτηριστικό για διαχωρισμό των δεδομένων. Έτσι, για κάθε χαρακτηριστικό υπολογίζεται το άθροισμα των μειώσεων του impurity για όλους τους κόμβους σε όλα τα δέντρα αποφάσεων. Όσο μεγαλύτερο είναι αυτό το άθροισμα, τόσο πιο σημαντικό είναι το χαρακτηριστικό.

Με αυτό τον τρόπο, υπολογίστηκαν τα τρία πιο σημαντικά χαρακτηριστικά που συνέβαλαν περισσότερο στην ταξινόμησης των καλλιεργειών. Στόχος της συγκεκριμένης τεχνικής είναι να μειωθεί δραστικά ο όγκος των δεδομένων που λαμβάνονται υπόψιν στην ταξινόμηση, όταν αυτή πραγματοποιείται στο σύνολο των δεδομένων, και να παραχθούν αποτελέσματα με εξίσου υψηλή ακρίβεια.

## 4.7 Μεταφορά Γνώσης

Το τελευταίο βήμα της ανάλυσης είναι η διερεύνηση της δυνατότητας μεταφοράς των εκπαιδευμένων μοντέλων σε διαφορετικές περιοχές. Ο στόχος είναι να αξιολογηθεί εάν ένα μοντέλο εκπαιδευμένο σε δεδομένα από μια περιοχή θα μπορούσε να εφαρμοστεί αποτελεσματικά για την ταξινόμηση των καλλιεργειών σε άλλες περιοχές. Αυτή η προσέγγιση αξιοποιεί τη γνώση και τα χαρακτηριστικά που αποκτήθηκαν από ένα σύνολο δεδομένων για τη βελτίωση της απόδοσής του σε ένα διαφορετικό, αλλά σημασιολογικά ίδιο σύνολο δεδομένων.

Στο πρώτο μέρος, διεξήχθησαν δύο πειράματα για να αξιολογηθεί η ικανότητα μεταφοράς των μοντέλων. Το κριτήριο επιλογής των περιοχών ήταν ο αριθμός των δειγμάτων ανά περιοχή.

Στο πρώτο πείραμα, το μοντέλο που εκπαιδεύτηκε σε δεδομένα από την Περιοχή Β μεταφέρθηκε τόσο στην Περιοχή Α όσο και στην Περιοχή Γ. Με την εφαρμογή του προ-εκπαιδευμένου μοντέλου στα σύνολα δεδομένων δοκιμής των άλλων περιοχών, αξιολογήθηκε η απόδοση της ταξινόμησης. Αυτό το πείραμα είχε ως στόχο να διερευνήσει τη δυνατότητα μεταφοράς του μοντέλου σε νέες περιοχές με δυνητικά διαφορετικά χαρακτηριστικά και κατανομές των ίδιων καλλιεργειών.

Στο δεύτερο πείραμα, ένα νέο μοντέλο εκπαιδεύτηκε χρησιμοποιώντας δεδομένα τόσο από την Περιοχή Α όσο και από την Περιοχή Β, έτσι ώστε να ενδυναμώσει τον αριθμό δειγμάτων της εκπαίδευσης και στη συνέχεια μεταφέρθηκε στην Περιοχή Γ. Αυτό το πείραμα είχε στόχο να διερευνήσει τα οφέλη της εκπαίδευσης ενός μοντέλου σε συνδυασμένα σύνολα δεδομένων από πολλές περιοχές και να αξιολογήσει την απόδοσή του σε χωριστή περιοχή.

Η απόδοση των μεταφερόμενων μοντέλων αναλύθηκε χρησιμοποιώντας διάφορες μετρήσεις όπως την αξιολόγηση της ταξινόμησης με τον υπολογισμό της ακρίβεια (precision), της ανάκληση (recall) και του δείκτη F1-score. Συγκρίνοντας την απόδοση των μεταφερόμενων μοντέλων με τα βασικά μοντέλα που εκπαιδεύτηκαν σε μεμονωμένες περιοχές, αποκτήθηκαν πολύτιμες γνώσεις σχετικά με την αποτελεσματικότητα της μάθησης μεταφοράς στην ταξινόμηση των καλλιεργειών.

Στο δεύτερο μέρος, αφού αξιολογήθηκαν τα αποτελέσματα του πρώτου μέρους, ακολουθήθηκε διαφορετική τεχνική επιλογής δεδομένων. Με βάση τον δείκτη NDVI για κάθε κατηγορία και στις τρεις περιοχές, αναδείχθηκαν κάποιες σχέσεις μεταξύ του δείκτη NDVI. Συγκεκριμένα, για πολλές κατηγορίες φάνηκε πως ο NDVI στην περιοχή Α και ο NDVI στην περιοχή Γ, έχουν το μεγαλύτερο εύρος μεταξύ τους. Για αυτό το λόγο αποφασίστηκε να εκπαιδευτεί ένα μοντέλο σε αυτές τις δύο περιοχές Α+Γ με σκοπό να έχει ως input τις τιμές του NDVI που έχουν μεγάλο εύρος, με την προοπτική πως το μοντέλο θα γενικεύσει αυτές τις ακραίες τιμές και θα έχει καλύτερη προσαρμογή στην περιοχή Β. Ταυτόχρονα, η περιοχή Γ είχε και μεγάλο αριθμό δειγμάτων για κατηγορίες που γενικά δεν αντιπροσωπεύονται επαρκώς στο σύνολο των δεδομένων όπως η αγρανάπαυση, οι μόνιμες καλλιέργειες και η ελαιοκράμβη. Γι αυτό το λόγο αρχικά χρησιμοποιήθηκαν τεχνικές oversampling και undersampling για τη δημιουργία ενός πιο ισορροπημένου σύνολο δεδομένων. Στη συνέχεια, το μοντέλο των περιοχών Α+Γ μεταφέρθηκε στην περιοχή Β, όπου και αξιολογήθηκε. Για την βελτίωση των αποτελεσμάτων, εφαρμόστηκαν εκ νέου τεχνικές oversampling και undersampling στην περιοχή Β για την πιο ισορροπημένη κατανομή των δεδομένων. Τέλος, το μοντέλο των περιοχών Α+Γ μεταφέρθηκε ξανά στην περιοχή Β και αξιολογήθηκε.

# Κεφάλαιο 5 – Αποτελέσματα

Παρακάτω παρουσιάζονται τα αποτελέσματα όπως προέκυψαν ύστερα από την εφαρμογή αλγορίθμων μη-επιβλεπόμενης και επιβλεπόμενης ταξινόμησης στο dataset, καθώς και οι τεχνικές εξαγωγής των σημαντικότερων χαρακτηριστικών, η μεταφορά γνώσης σε άλλες περιοχές και η αξιοποίηση μεθόδων oversampling και undersampling.

## 5.1 Μη Επιβλεπόμενη ταξινόμηση

Από το σύνολο των δεδομένων και των τριών περιοχών, αρχικά επιλέχθηκαν οι δείκτες NDVI και RVI για την εφαρμογή τεχνικών μη επιβλεπόμενης ταξινόμησης. Παράλληλα, ενδιαφέρον παρουσίασε να εξεταστεί και όλο το σύνολο των δεδομένων και να κριθεί ως προ την επίδοσή του στην μη επιβλεπόμενη ταξινόμηση.

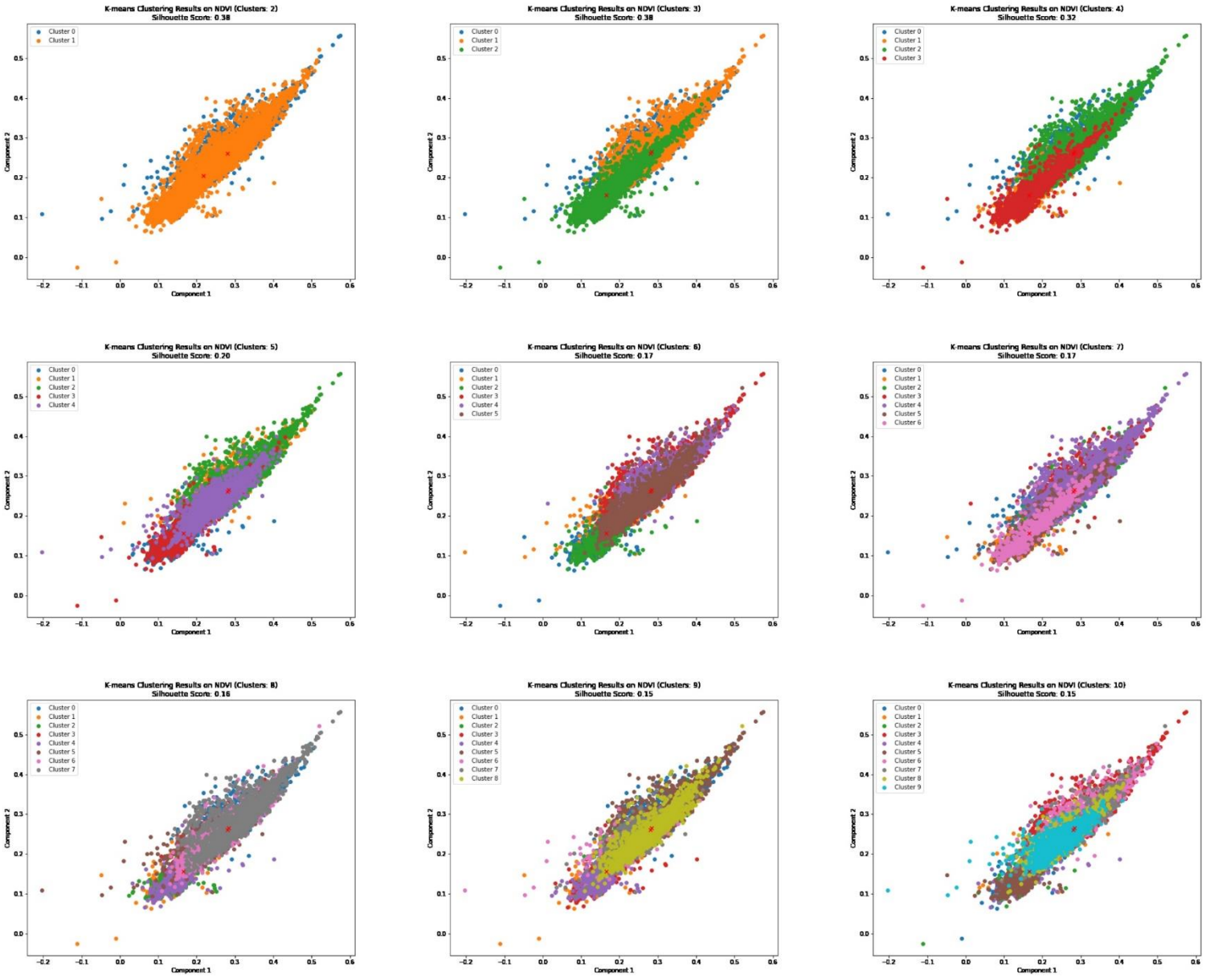
Το σύνολο το δεδομένων απαρτίζεται από 27073 αγροτεμάχια, 22 τιμές για τον δείκτη NDVI, 22 τιμές για τον δείκτη RVI και 411 τιμές για το σύνολο των δεδομένων. Λόγω των υψηλών διαστάσεων εφαρμόστηκαν τεχνικές μείωσης διαστάσεων, όπως η PCA (Ανάλυση Κυρίων Συνιστωσών) και η t-SNE ( τ-Κατανεμημένη Στοχαστική Ενσωμάτωση Γείτονα). Οι διαστάσεις στις οποίες μειώθηκαν τα δεδομένα και με τις δύο τεχνικές είναι δύο.

Έτσι, για τους αλγόριθμους k-means και Gaussian Mixture Model έχουμε πέντε αποτελέσματα (ένα για τον NDVI στις 22 τιμές, ένα για τον NDVI ύστερα από PCA ένα για τον RVI στις 22 τιμές, ένα για τον RVI PCA και ένα για το σύνολο των δεδομένων, αφού εφαρμόστηκε PCA στις 411 τιμές). Όσον αφορά τον t-SNE, επειδή κάνει μείωση των διαστάσεων, παρέχονται τρία αποτελέσματα, ένα για τον NDVI, ένα για τον RVI και ένα για το σύνολο των δεδομένων.

Παρακάτω δίνονται τα αποτελέσματα ύστερα από την εφαρμογή των αλγορίθμων που αναφέρθηκαν. Σε κάθε διάγραμμα φαίνεται η απόδοση κάθε αλγορίθμου για κάθε διαφορετικό αριθμό κλάσεων και επίσης, αναγράφεται η βαθμολογία silhouette, η οποία αφορά την αξιολόγηση κάθε αλγορίθμου.

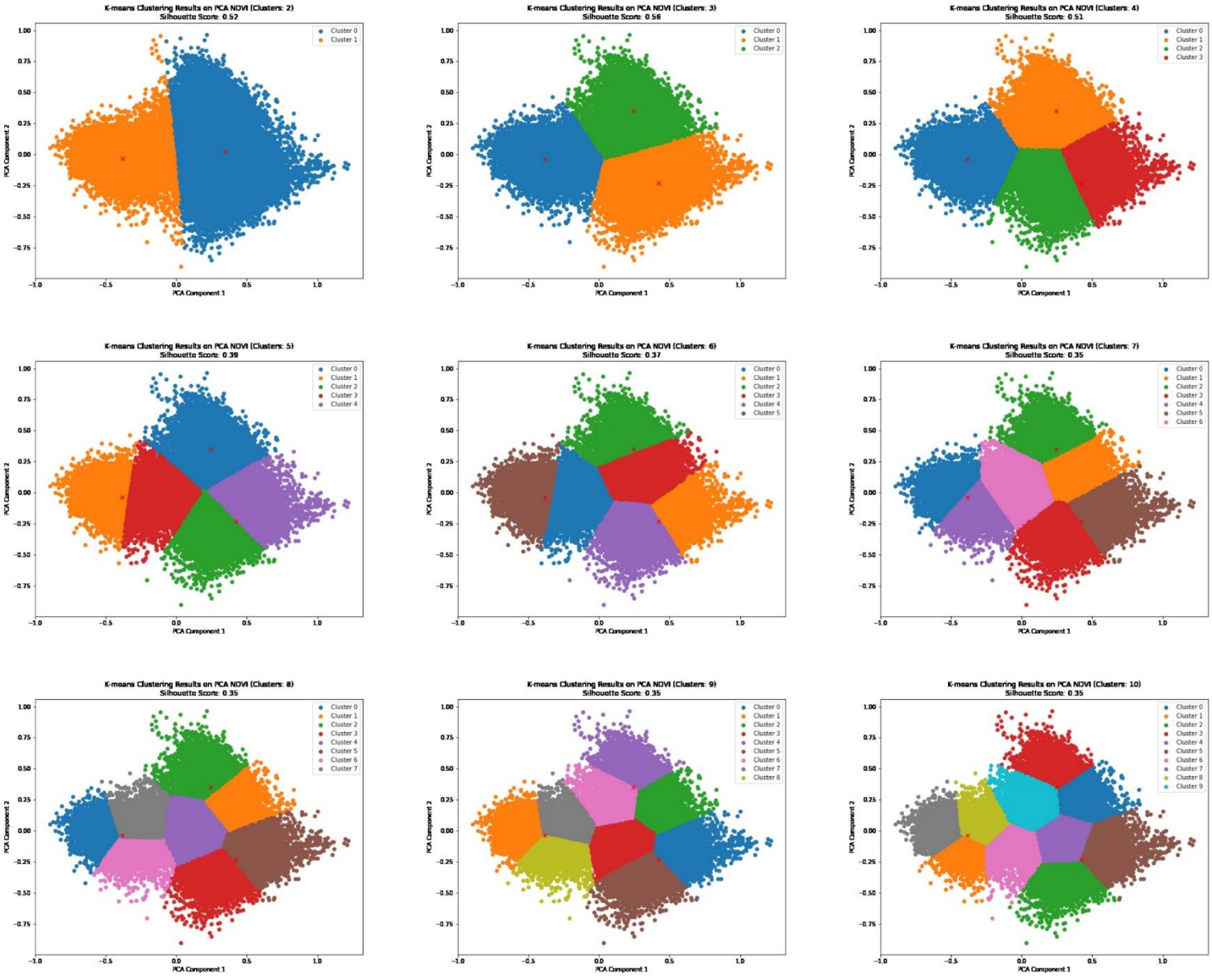
## 5.1.1 K-means

### *5.1.1.1 K-means στον NDVI*



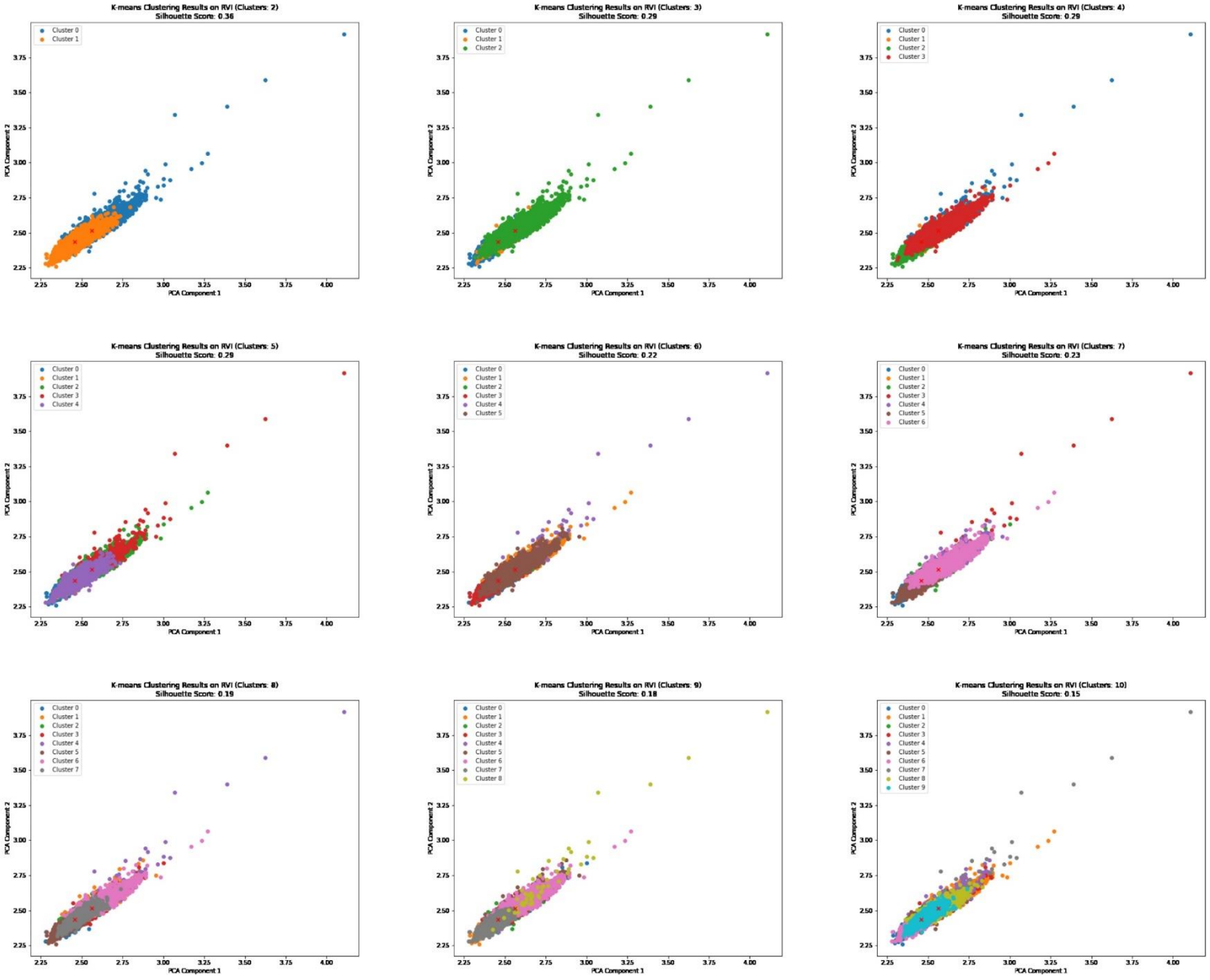

**Εικόνα 22 - K-means NDVI**

### *5.1.1.2 K-means σε PCA NDVI*



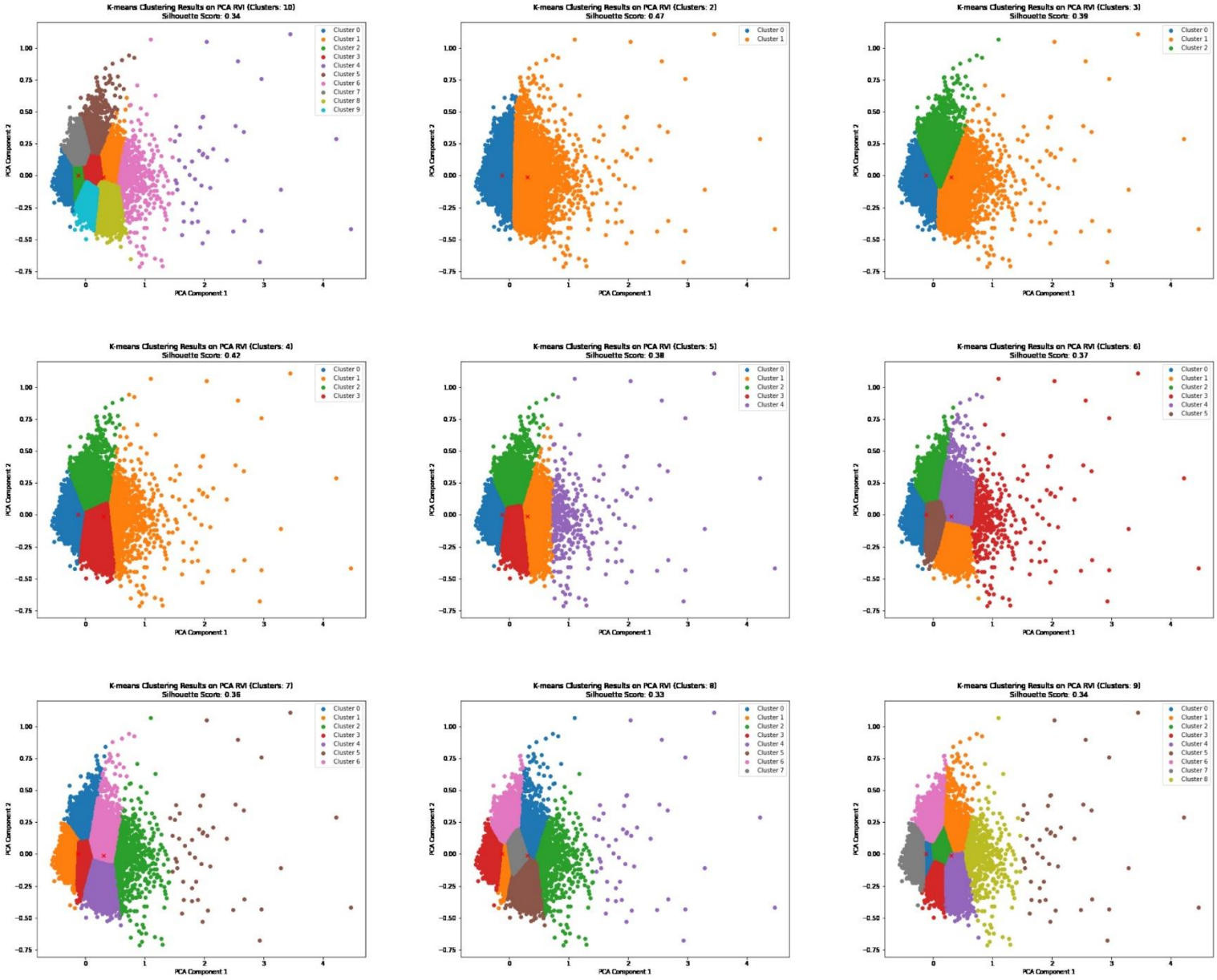

**Εικόνα 2319 - K-means NDVI PCA**

### 5.1.1.3 K-means στον RVI



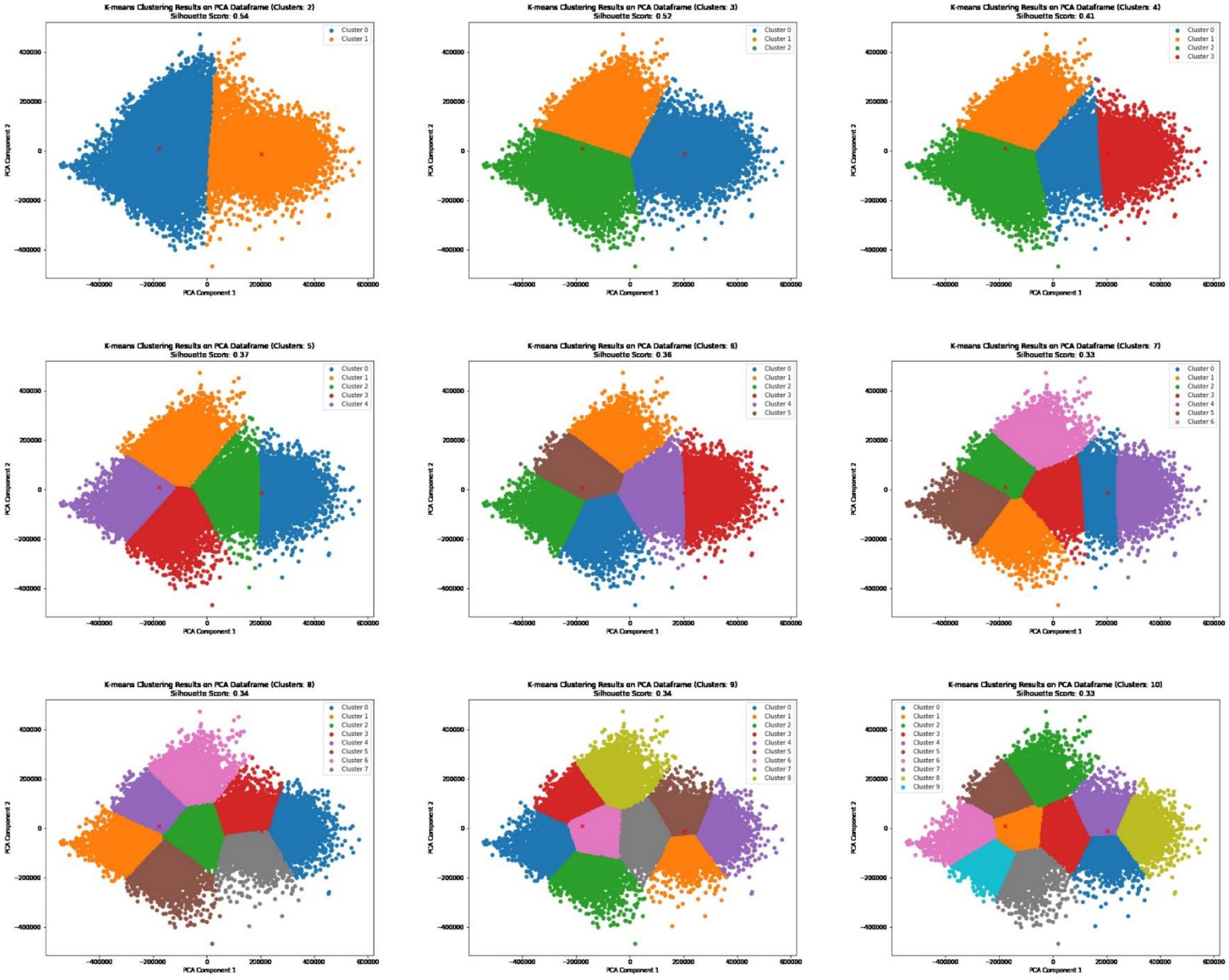

**Εικόνα 24 - K-means RVI**

### 5.1.1.4 K-means σε PCA RVI




**Εικόνα 25 - K-means RVI PCA**

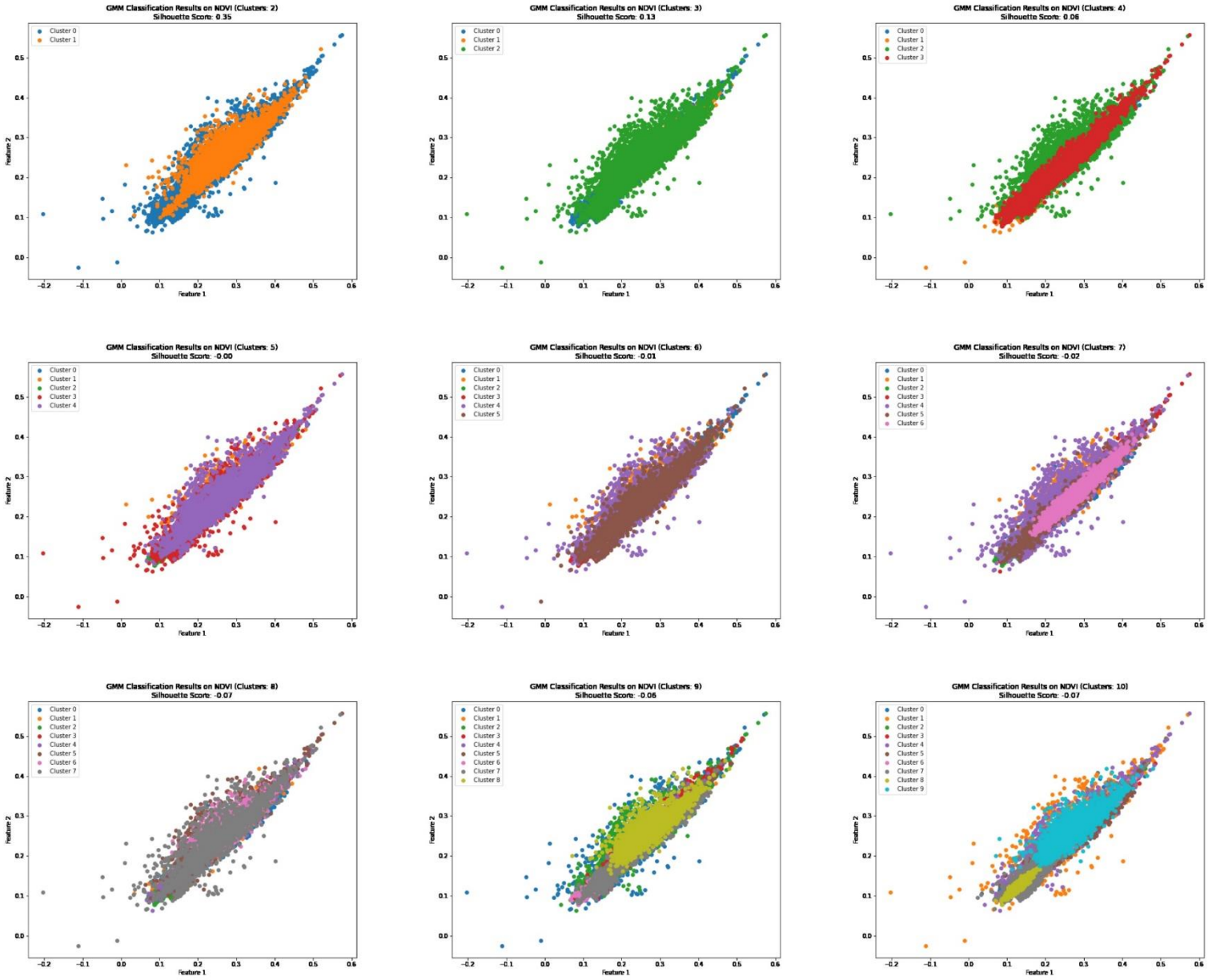
### 5.1.1.5 K-means σε PCA του Dataset




**Εικόνα 26 - K-means Dataframe**

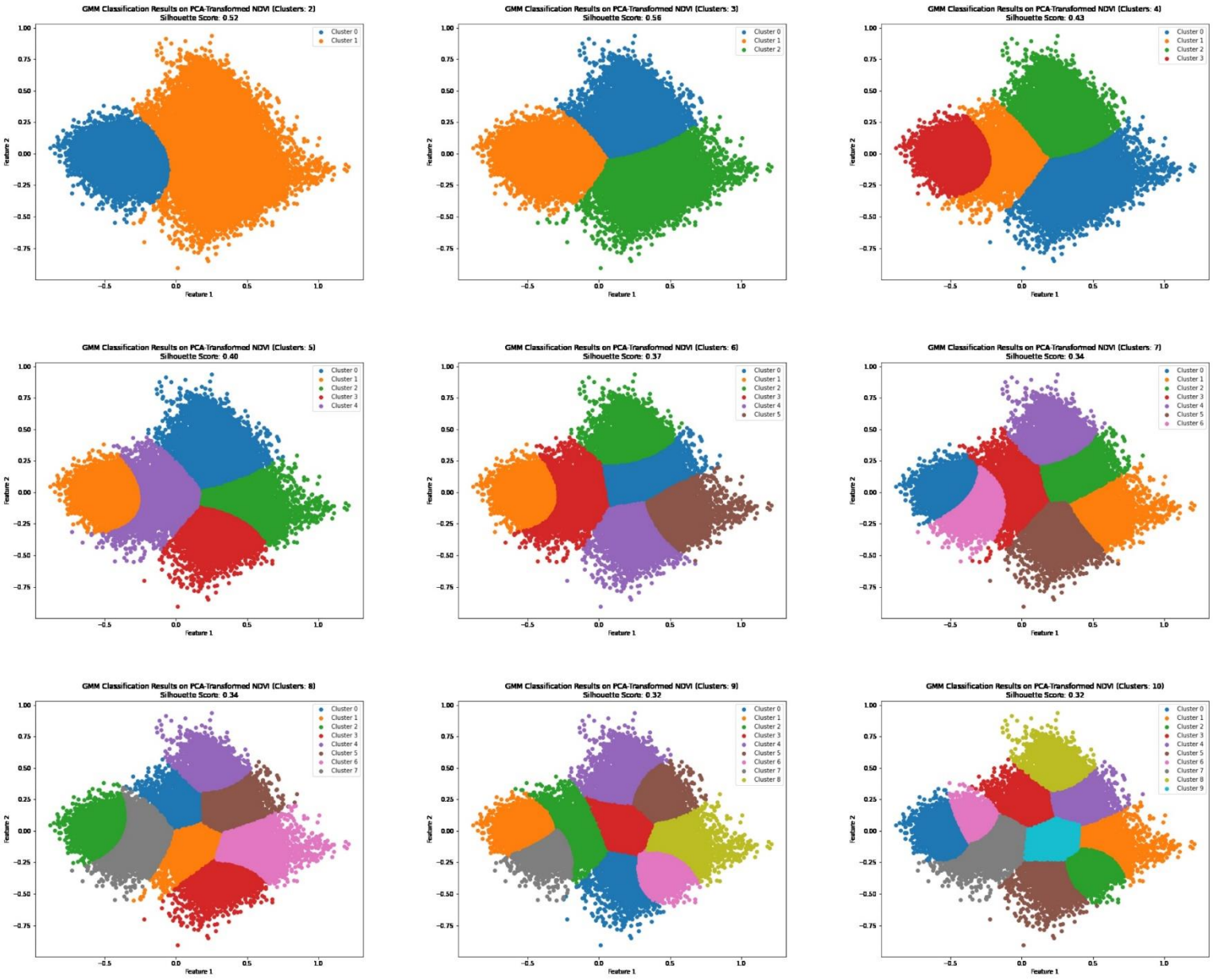
## 5.1.2 Gaussian Mixture Model (GMM)

### *5.1.2.1 GMM στον NDVI*



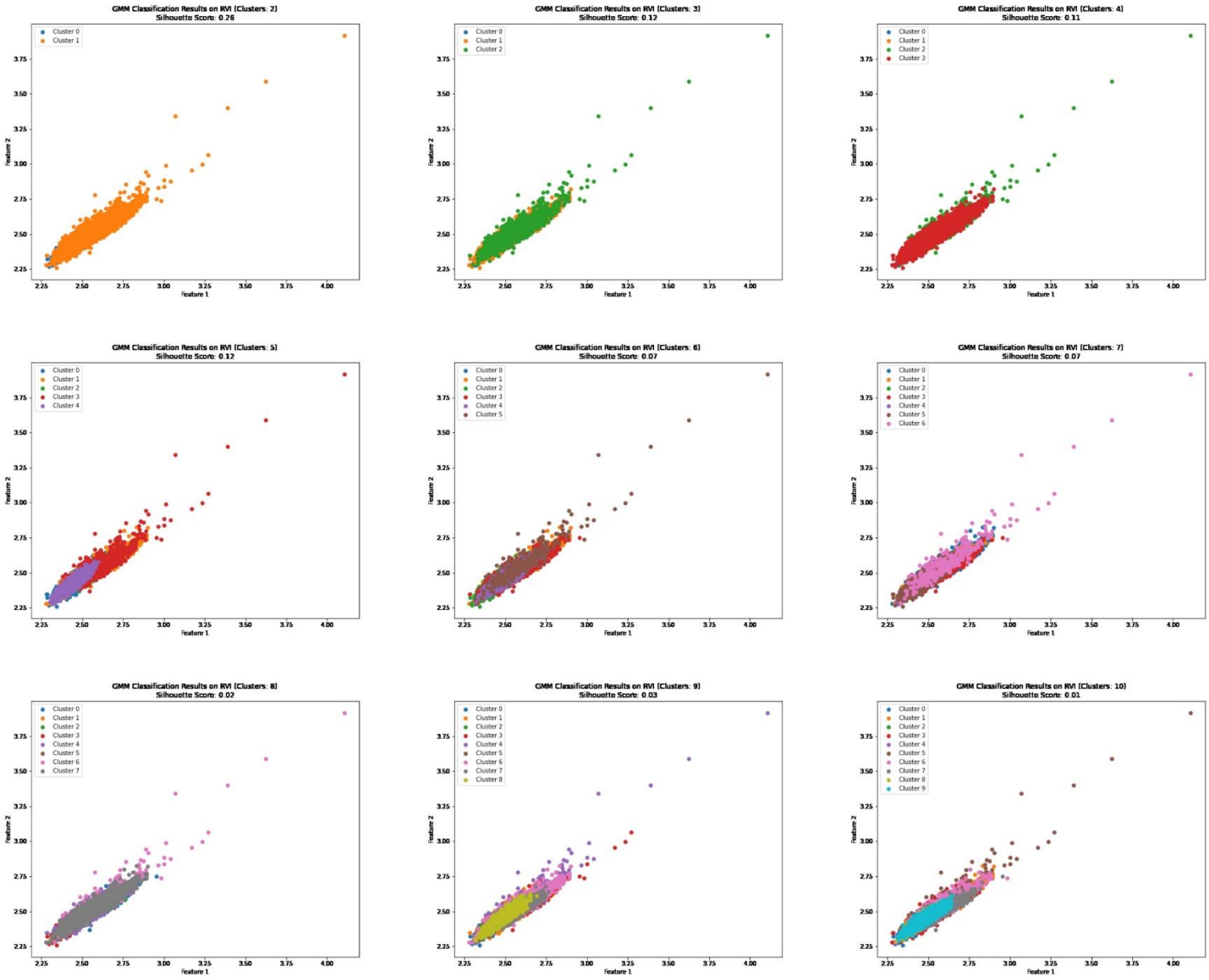

**Εικόνα 20 - GMM NDVI**

### 5.1.2.2 GMM - PCA NDVI



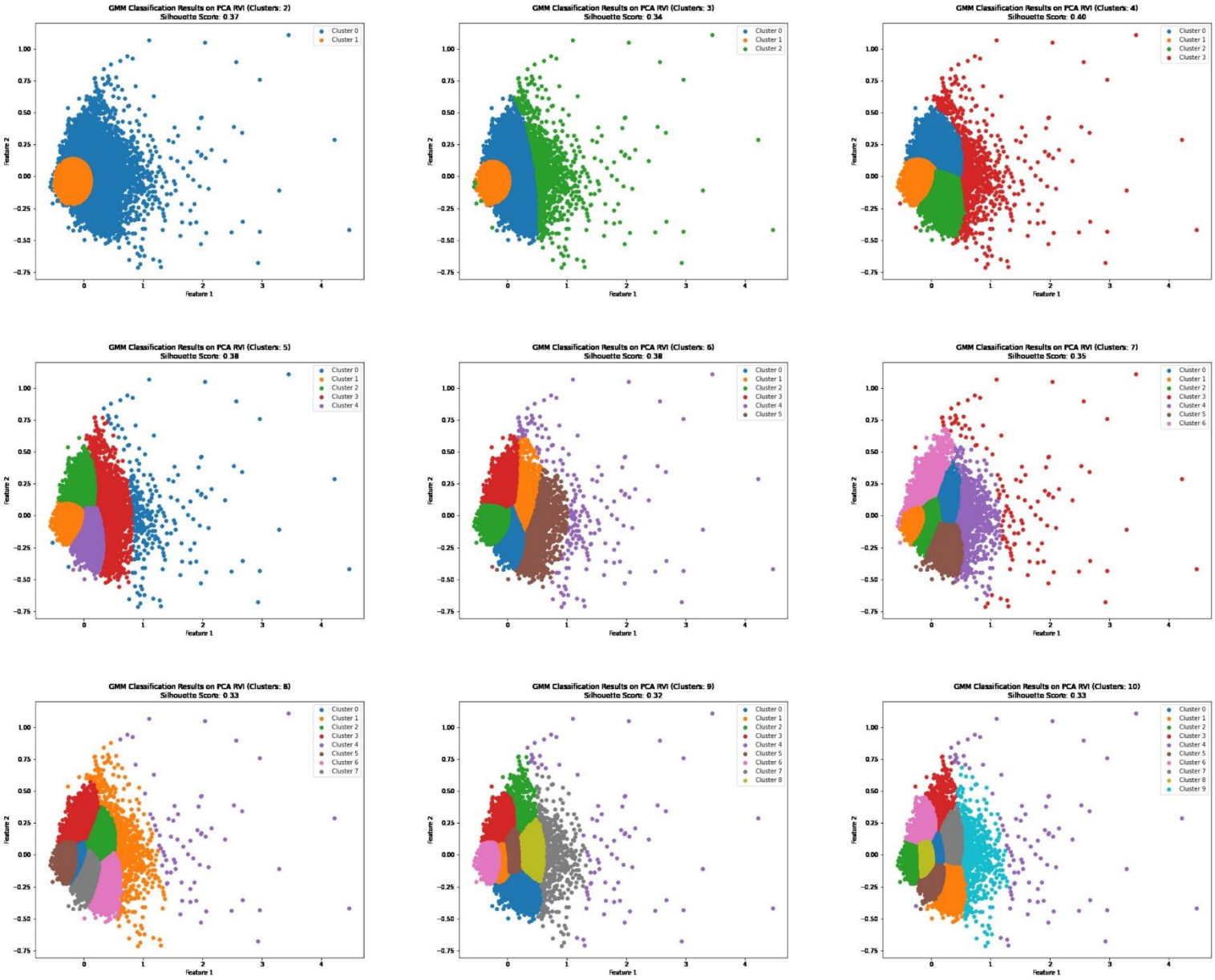

**Εικόνα 28 - GMM NDVI PCA**

### 5.1.2.3 GMM – RVI



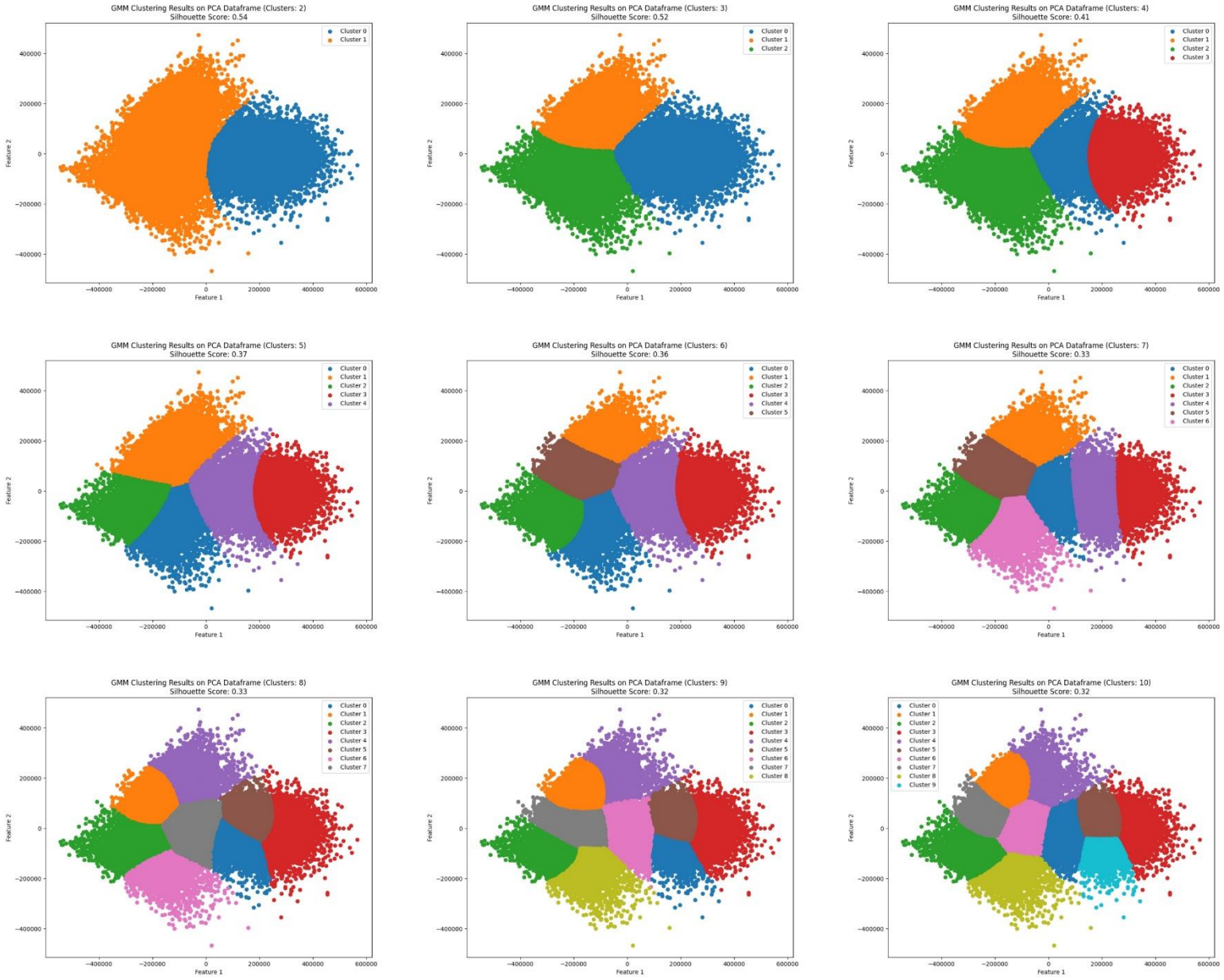

**Εικόνα 29- GMM RVI**

### 5.1.2.4 GMM – PCA RVI




**Εικόνα 30- GMM RVI PCA**

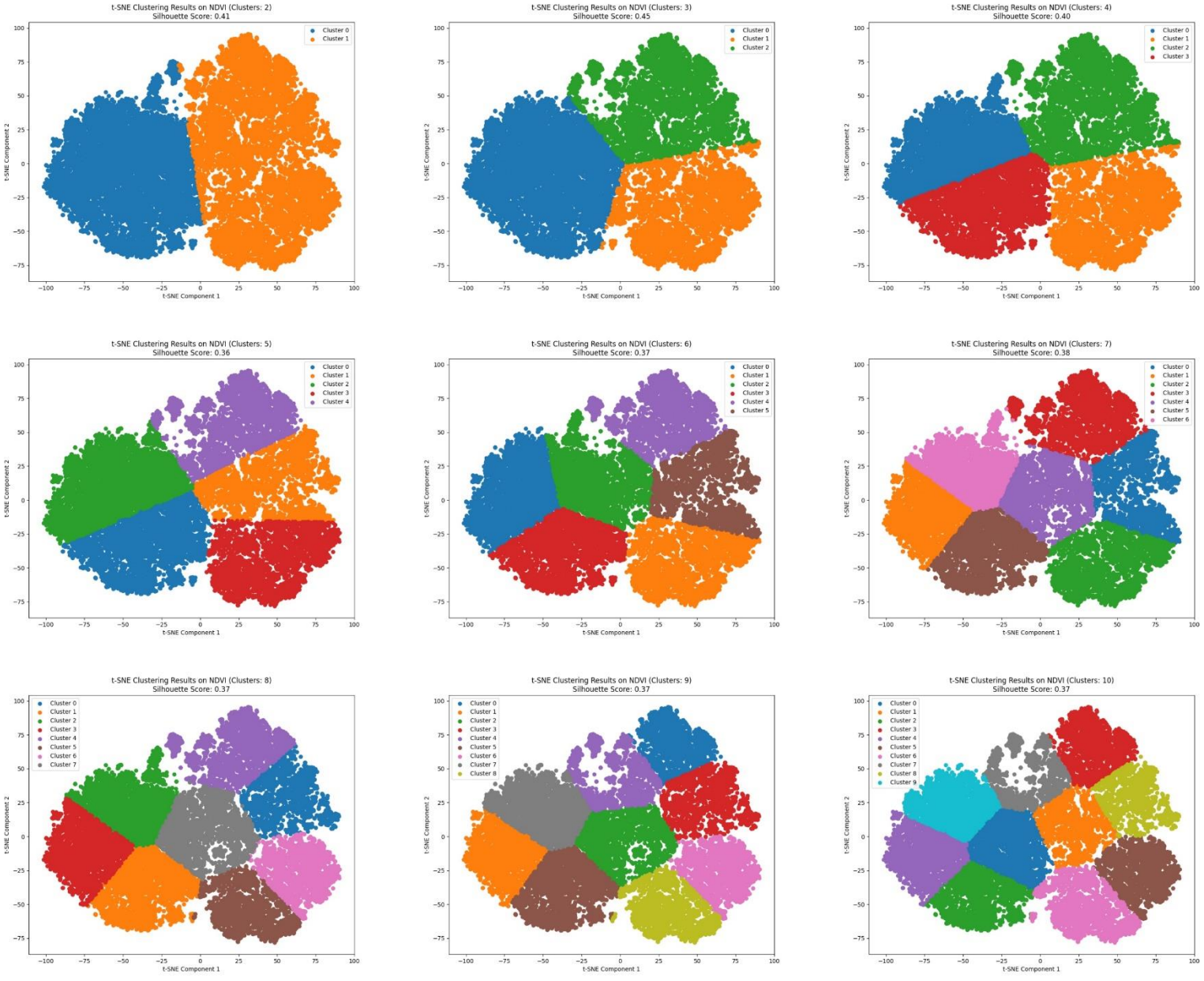
### 5.1.2.5 GMM – PCA Dataset




**Εικόνα 31 – GMM Dataframe**

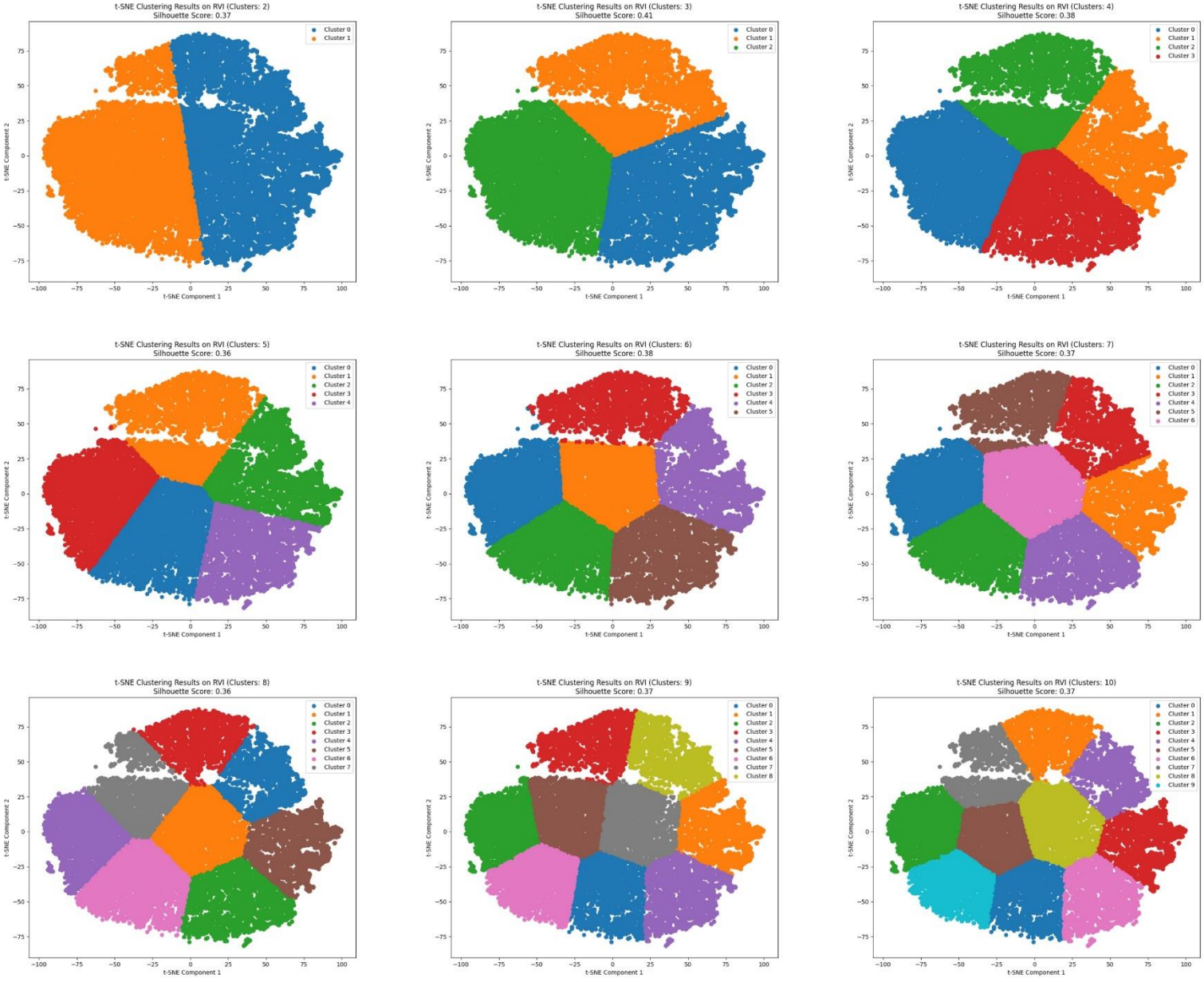
## 5.1.3 t-SNE

### *5.1.3.1 t-SNE NDVI*



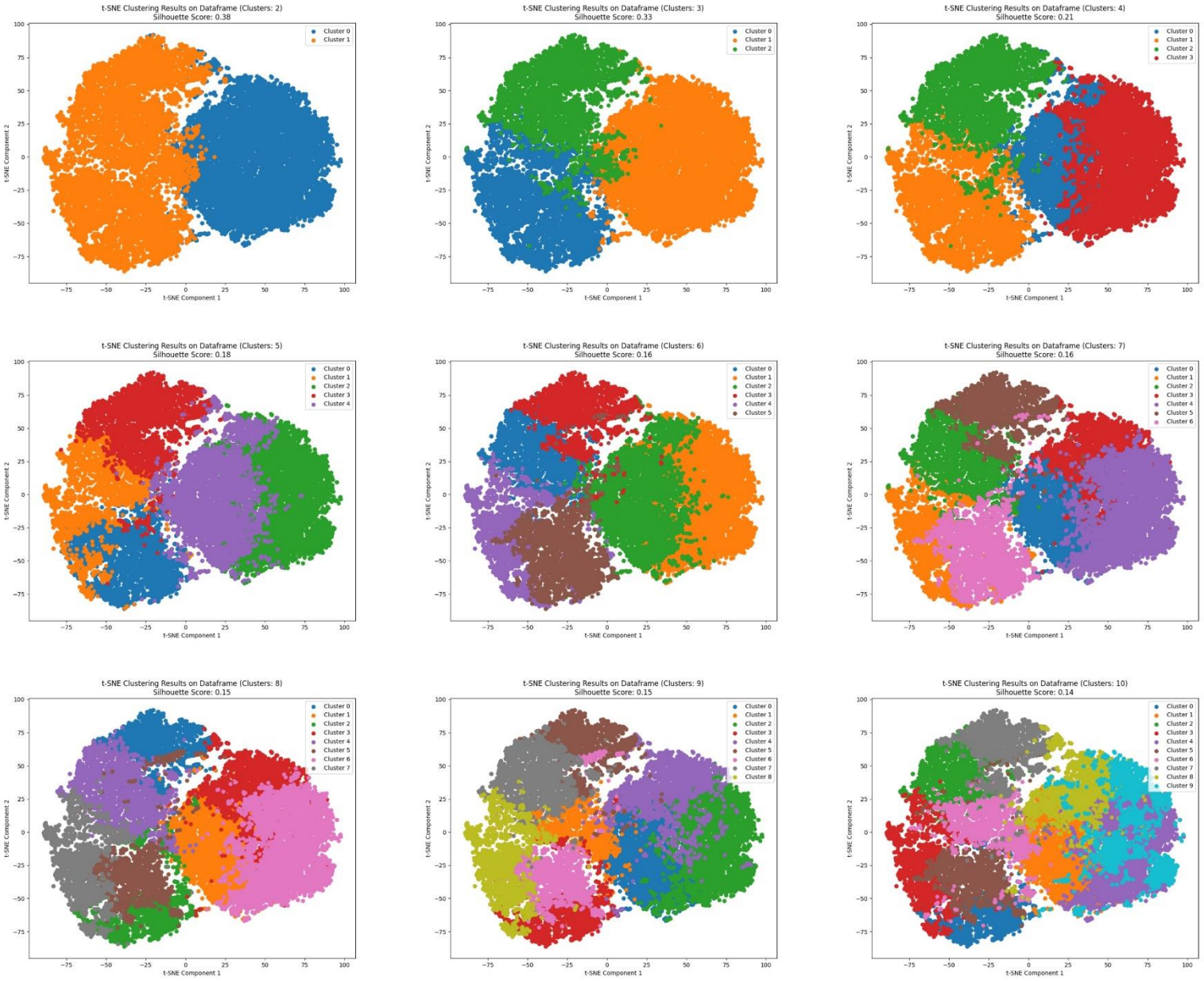

**Εικόνα 32- t-SNE NDVI**

### 5.1.3.2 t-SNE RVI




**Εικόνα 21- t-SNE RVI**

### 5.1.3.2 t-SNE Dataset




**Εικόνα 3422- t-SNE Dataframe**

### 5.1.4 Αξιολόγηση αποτελεσμάτων μη επιβλεπόμενης ταξινόμησης

Αφού πραγματοποιήθηκαν οι αλγόριθμοι στους δείκτες NDVI, RVI και στο σύνολο των δεδομένων, είτε με μείωση διαστάσεων είτε όχι, δημιουργήθηκε ο παρακάτω πίνακας, στον οποίο παρουσιάζονται οι βαθμολογίες silhouette για κάθε αλγόριθμο και κάθε αριθμό κλάσεων. Με μπλε χρώμα φαίνονται οι μέγιστες βαθμολογίες silhouette, ενώ με κόκκινο χρώμα παρουσιάζονται οι ελάχιστες βαθμολογίες silhouette. Επιπρόσθετα, δημιουργήθηκε μία στήλη εύρους «Range», η οποία είναι η αφαίρεση της μικρότερης βαθμολογίας silhouette από την μεγαλύτερη βαθμολογία για κάθε αλγόριθμος.

Υψηλότερες βαθμολογίες silhouette υποδηλώνουν καλά καθορισμένα και διακριτά αποτελέσματα, ενώ χαμηλότερες τιμές υποδηλώνουν επικαλυπτόμενες ή κακώς διαχωρισμένες κλάσεις.

| | n=2 | n=3 | n=4 | n=5 | n=6 | n=7 | n=8 | n=9 | n=10 | Range |
|---|---|---|---|---|---|---|---|---|---|---|
| K-means NDVI | 0.38 | 0.38 | 0.32 | 0.2 | 0.17 | 0.17 | 0.16 | 0.15 | 0.14 | 0.24 |
| K-means NDVI PCA | 0.52 | 0.56 | 0.51 | 0.39 | 0.37 | 0.35 | 0.35 | 0.35 | 0.35 | 0.21 |
| K-means RVI | 0.36 | 0.29 | 0.29 | 0.29 | 0.22 | 0.23 | 0.19 | 0.18 | 0.15 | 0.21 |
| K-means RVI PCA | 0.49 | 0.37 | 0.42 | 0.38 | 0.37 | 0.36 | 0.33 | 0.34 | 0.34 | 0.16 |
| K-means Dataframe | 0.54 | 0.52 | 0.41 | 0.37 | 0.36 | 0.33 | 0.34 | 0.34 | 0.33 | 0.21 |
| GMM NDVI | 0.35 | 0.13 | 0.06 | 0 | -0.01 | -0.02 | -0.07 | -0.06 | -0.07 | 0.42 |
| GMM NDVI PCA | 0.52 | 0.56 | 0.48 | 0.4 | 0.37 | 0.35 | 0.35 | 0.34 | 0.33 | 0.23 |
| GMM RVI | 0.26 | 0.12 | 0.11 | 0.12 | 0.07 | 0.05 | 0.06 | 0.03 | 0.03 | 0.23 |
| GMM RVI PCA | 0.37 | 0.35 | 0.4 | 0.38 | 0.38 | 0.35 | 0.32 | 0.33 | 0.33 | 0.08 |
| GMM Dataframe | 0.54 | 0.52 | 0.41 | 0.37 | 0.36 | 0.33 | 0.33 | 0.32 | 0.32 | 0.22 |
| t-SNE NDVI | 0.4 | 0.44 | 0.39 | 0.38 | 0.38 | 0.37 | 0.37 | 0.37 | 0.37 | 0.07 |
| t-SNE RVI | 0.37 | 0.4 | 0.38 | 0.35 | 0.37 | 0.37 | 0.35 | 0.36 | 0.37 | 0.05 |
| t-SNE Dataframe | 0.38 | 0.33 | 0.21 | 0.18 | 0.16 | 0.16 | 0.15 | 0.15 | 0.14 | 0.24 |

**Πίνακας 1 - Βαθμολογίες Silhouette**

Αναλύοντας τα αποτελέσματα, παρατηρείται πως οι αλγόριθμοι k-means και GMM στους δείκτες NDVI και RVI δεν έχουν πετύχει καλά αποτελέσματα, η μέγιστη τιμή βρίσκεται στον k-means NDVI με 0.38 και η ελάχιστη τιμή βρίσκεται στον GMM RVI με 0.26. Επίσης, οι παραπάνω αλγόριθμοι στον NDVI και RVI πέτυχαν τις μέγιστες βαθμολογίες silhouette στις κλάσεις με αριθμό 2 (n=2), εκτός από τον k-means NDVI που είχε και στις κλάσεις με αριθμό 3 μέγιστη τιμή, αρκετά λιγότερες από τις πραγματικές κλάσεις που είναι 7.

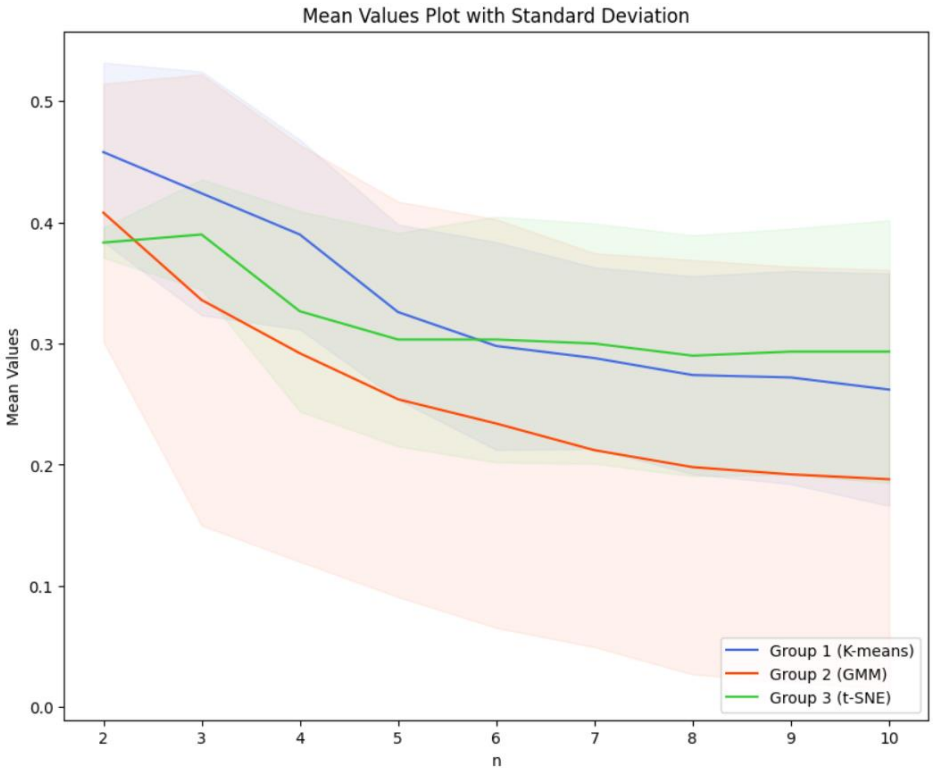
Αναφορικά με τα μετασχηματισμένα δεδομένα NDVI και RVI με τεχνικές PCA ή t-SNE, τα αποτελέσματα είναι αρκετά βελτιωμένα. Συγκεκριμένα, ο k-means NDVI PCA πέτυχε υψηλή βαθμολογία silhouette, 0.56, στις 3 κλάσεις. Επιπρόσθετα, ο k-means RVI PCA πέτυχε υψηλότερη τιμή 0.49 στις 2 κλάσεις, ωστόσο στις 3 κλάσεις μειώθηκε στο 0.37 και στις 4 κλάσεις ανέβηκε ξανά στο 0.42. Αναφορικά με τον GMM NDVI PCA, πέτυχε υψηλότερη τιμή

στις 2 κλάσεις με 0.56 και ο GMM RVI PCA πέτυχε την υψηλότερη τιμή του στις 4 κλάσεις με βαθμολογία 0.4. Τέλος, ο t-SNE NDVI πέτυχε υψηλότερη τιμή 0.44 στις 3 κλάσεις και αντίστοιχα ο t-SNE RVI πέτυχε υψηλότερη τιμή 0.4 στις 3 κλάσεις.

Όσον αφορά τον k-means στο Dataframe και τον GMM στο Dataframe πέτυχαν υψηλότερες τιμές στις κλάσεις με αριθμό 2 με τιμή 0.54, ενώ ο t-SNE Dataframe πέτυχε υψηλότερη τιμή 0.38 στις κλάσεις με αριθμό 2.

Σημαντικό είναι να εξεταστεί επίσης πως τα πήγαν οι αλγόριθμοι στις πραγματικές κλάσεις των δεδομένων που είναι 7. Από τον πίνακα παρατηρείται πως τις μεγαλύτερες βαθμολογίες silhouette τις πέτυχαν οι t-SNE σε NDVI και RVI με τιμή 0.37, και αμέσως μετά ο k-means RVI PCA με 0.36. Ωστόσο, είναι λογικό να μην επιτευχθούν οι καλύτερες τιμές silhouette στις κλάσεις με αριθμό 7, καθώς οι αλγόριθμοι μη επιβλεπόμενης ταξινόμησης δεν έχουν γνώση για τον πραγματικό αριθμό των κλάσεων από πριν και για αυτό το λόγο καθίσταται δύσκολη η επιτυχής ανάθεση των καλύτερων αποτελεσμάτων στις πραγματικές κλάσεις.

Συνολικά, ο GMM RVI PCA πέτυχε το κοντινότερο στο επιθυμητό αποτέλεσμα με την υψηλότερη βαθμολογία του στις κλάσεις με αριθμό 4.




**Εικόνα 23 - Μέσος όρος βαθμολογιών silhouette**

Για την καλύτερη οπτικοποίηση της στήλης εύρους, δημιουργήθηκε το παραπάνω γράφημα που δείχνει τον μέσο όρο των βαθμολογιών silhouette για κάθε κατηγορία αλγόριθμου στον διαφορετικό αριθμό κλάσεων, καθώς και την τυπική απόκλιση.

Παρατηρείται πως το μεγαλύτερο εύρος βαθμολογιών silhouette το καταλαμβάνει ο αλγόριθμος GMM, ενώ ο t-SNE έχει το μικρότερο εύρος μεταξύ των τριών αλγορίθμων.

## 5.2 Επιβλεπόμενη ταξινόμηση

Για την επιβλεπόμενη ταξινόμηση χρησιμοποιήθηκε το σύνολο των δεδομένων και στις τρεις περιοχές, λαμβάνοντας υπόψιν και τα 21 χαρακτηριστικά που διαθέτει, δηλαδή τις μπάντες και τους δείκτες των δορυφόρων Sentinel-1 και Sentinel-2.

Αρχικά, πραγματοποιήθηκε επιβλεπόμενη ταξινόμηση με τη χρήση των αλγορίθμων Random Forest (RF), Linear Support Vector Machine (Linear SVM) και Gradient Boosting (GB).

Στη συνέχεια, αφού επιλέχθηκε ο καταλληλότερος αλγόριθμος μεταξύ των τριών λαμβάνοντας υπόψιν τις μετρικές precision, recall και F1 Score, πραγματοποιήθηκε τεχνική εξαγωγής των σημαντικότερων χαρακτηριστικών (feature importance) και το μοντέλο επανεκπαιδεύτηκε μόνο με τη συνεισφορά των τριών πιο σημαντικών χαρακτηριστικών.

Στο τέλος, εφαρμόστηκαν διάφορες τεχνικές για τη αξιολόγηση μεθόδων μεταφοράς γνώσης. Ουσιαστικά, εκπαιδεύτηκαν από την αρχή μοντέλα σε διάφορες περιοχές και σε συνδυασμό περιοχών και αυτά τα μοντέλα μεταφέρθηκαν σε άλλες περιοχές με σκοπό να εξεταστεί η μεταφερσιμότητα του εκάστοτε μοντέλου.

### 5.2.1 Ταξινομητές επιβλεπόμενης ταξινόμησης

#### *5.2.1.1. Ταξινομητής Random Forest*

| Κατηγορία | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **Αγρανάπαυση** | 0.59 | 0.66 | 0.62 | 192 |
| **Γρασίδι** | 0.96 | 0.96 | 0.96 | 3793 |
| **Μόνιμες Καλλιέργειες** | 0.46 | 0.28 | 0.35 | 100 |
| **Πατάτες** | 0.76 | 0.75 | 0.76 | 159 |
| **Ανοιξιάτικες καλλιέργειες σιτηρών** | 0.91 | 0.93 | 0.92 | 1578 |
| **Χειμερινές καλλιέργειες σιτηρών** | 0.93 | 0.94 | 0.93 | 1863 |
| **Ελαιοκράμβη** | 0.96 | 0.93 | 0.94 | 437 |
| **Ακρίβεια** | | | 0.93 | 8122 |
| **Μέσος όρος** | 0.80 | 0.78 | 0.78 | 8122 |
| **Σταθμισμένος Μέσος Όρος** | 0.93 | 0.93 | 0.93 | 8122 |
| **Χρόνο υλοποίησης: 4.5sec** | | | | |

Ο πίνακας παρουσιάζει τα αποτελέσματα ταξινόμησης που προέκυψαν από τον Random Forest που εφαρμόστηκε στο σύνολο των δεδομένων.

**Αγρανάπαυση**: Η ακρίβεια για αυτήν την κατηγορία είναι 0.59, υποδεικνύοντας ότι το 59% των δειγμάτων που ταξινομήθηκαν ως αγρανάπαυση ήταν στην πραγματικότητα αγρανάπαυση. Η ανάκληση είναι 0.66, υποδηλώνοντας ότι το 66% των πραγματικών δειγμάτων ταξινομήθηκαν σωστά. Το F1-Score, το οποίο λαμβάνει υπόψη τόσο την ακρίβεια όσο και την ανάκληση, είναι 0.62, αντιπροσωπεύοντας τον αρμονικό μέσο όρο ακρίβειας και ανάκλησης.

**Γρασίδι**: Ο ταξινομητής είχε εξαιρετική απόδοση για αυτή την κατηγορία με υψηλή ακρίβεια, ανάκληση και F1 score 0.96. Αυτό υποδεικνύει ακριβείς και αξιόπιστες προβλέψεις για την κατηγορία Γρασίδι.

**Μόνιμες Καλλιέργειες**: Η ακρίβεια για αυτήν την κατηγορία είναι σχετικά χαμηλή στο 0.46, υποδηλώνοντας μεγαλύτερο αριθμό ψευδώς θετικών. Η ανάκληση είναι 0.28, υποδηλώνοντας ότι μόνο το 28% από το σύνολο των Μόνιμων Καλλιεργειών αναγνωρίστηκε σωστά. Το F1-Score είναι 0.35, υποδεικνύοντας ότι ο ταξινομητής δυσκολεύτηκε να ταξινομήσει αποτελεσματικά αυτήν την κατηγορία.

**Πατάτες:** Η ακρίβεια για τις Πατάτες είναι 0.76, υποδεικνύοντας ότι το 76% των προβλεπόμενων δειγμάτων ήταν σωστά. Η ανάκληση είναι 0.75, υποδηλώνοντας ότι το 75% των πραγματικών δειγμάτων στην κατηγορία Πατάτες ταξινομήθηκε με ακρίβεια. Το F1-Score είναι 0.76, υποδηλώνοντας μία ικανοποιητική ταξινόμηση.

**Ανοιξιάτικες καλλιέργειες σιτηρών**: Ο ταξινομητής πέτυχε τιμές υψηλής ακρίβειας, ανάκλησης και F1-Score, 0.91, 0.93 και 0.92, αντίστοιχα, επιδεικνύοντας ακριβείς προβλέψεις.

**Χειμερινές καλλιέργειες σιτηρών**: Παρόμοια με την κλάση «Ανοιξιάτικες καλλιέργειες σιτηρών», η κλάση εμφάνισε τιμές υψηλής ακρίβειας, ανάκλησης και F1, 0.93, 0.94 και 0.93, αντίστοιχα, υποδεικνύοντας εξαιρετικά αποτελέσματα.

**Ελαιοκράμβη**: Ο ταξινομητής είχε καλή απόδοση για τη συγκεκριμένη κατηγορία με ακρίβεια 0.96, υποδηλώνοντας υψηλό ποσοστό σωστών προβλέψεων. Η ανάκληση είναι 0.93, υποδεικνύοντας ότι το 93% των πραγματικών δειγμάτων ταξινομήθηκαν σωστά. Το F1-Score είναι 0.94, αντιπροσωπεύοντας μια καλή ισορροπία μεταξύ ακρίβειας και ανάκλησης.

**Ακρίβεια:** Η συνολική ακρίβεια του Random Forest είναι 0.93, υποδεικνύοντας ότι το 93% των δειγμάτων στο σύνολο δεδομένων ταξινομήθηκαν σωστά.

**Μέσος όρος:** Ο μέσος όρος ακρίβειας, ανάκλησης και F1-score είναι 0.80, 0.78 και 0.78, αντίστοιχα. Ο μέσος όρος δίνει το ίδιο βάρος σε κάθε κλάση, χωρίς να λαμβάνει υπόψιν τον αριθμό των δειγμάτων. για κάθε τάξη και παρέχει ένα συνολικό μέτρο απόδοσης.

**Σταθμισμένος μέσος όρος:** Η σταθμισμένη μέση ακρίβεια, ανάκληση και F1-score είναι 0.93, 0.93 και 0.93, αντίστοιχα, το οποίο υποδηλώνει εξαιρετικά αποτελέσματα. Ο σταθμισμένος μέσος όρος δίνει μεγαλύτερο βάρος σε κλάσεις με βάση των αριθμό των δειγμάτων τους.

Συνολικά, ο Random Forest επέδειξε υψηλές επιδόσεις στην ταξινόμηση των περισσότερων κατηγοριών καλλιεργειών, ιδιαίτερα του Γρασιδιού, των Ανοιξιάτικων και Χειμερινών καλλιεργειών σιτηρών και της Ελαιοκράμβης. Ωστόσο, δεν παρείχε ικανοποιητικά αποτελέσματα με την κατηγορία Μόνιμων Καλλιεργειών, επιδεικνύοντας χαμηλότερη ακρίβεια και ανάκληση.

Στη συνέχεια παρουσιάζεται και ο πίνακας σύγχυσης, ο οποίος χρησιμοποιείται για την αξιολόγηση του μοντέλου. Κάθε σειρά αφορά τις πραγματικές κλάσεις και κάθε στήλη αφορά τις κλάσεις που ταξινόμησε το μοντέλο. Το ιδανικότερο για τον πίνακα σύγχυσης είναι να έχει τις υψηλότερες τιμές στη διαγώνιό του, το οποίο σημαίνει πως το μοντέλο ταξινόμησε τις περισσότερες περιπτώσεις στις πραγματικές κλάσεις.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Αγρανάπαυση | 127 | 18 | 2 | 8 | 14 | 19 | 4 |
| Γρασίδι | 24 | 3648 | 29 | 8 | 49 | 31 | 4 |
| Μόνιμες Καλλιέργειες | 5 | 59 | 28 | 1 | 1 | 6 | 0 |
| Πατάτες | 12 | 4 | 0 | 120 | 18 | 5 | 0 |
| Ανοιξιάτικες Καλλιέργειες Σιτηρών | 24 | 28 | 0 | 15 | 1460 | 49 | 2 |
| Χειμερινές Καλλιέργειες Σιτηρών | 22 | 28 | 2 | 4 | 56 | 1743 | 8 |
| Ελαιοκράμβη | 3 | 3 | 0 | 2 | 6 | 18 | 405 |
| | Αγρανάπαυση | Γρασίδι | Μόνιμες Καλλιέργειες | Πατάτες | Ανοιξιάτικες Καλλιέργειες Σιτηρών | Χειμερινές Καλλιέργειες Σιτηρών | Ελαιοκράμβη |

**Αγρανάπαυση:** Ο Random Forest ταξινόμησε σωστά 127 περιπτώσεις αγρανάπαυσης. Ωστόσο, υπήρξαν ορισμένες εσφαλμένες ταξινομήσεις, με 18 περιπτώσεις να ταξινομούνται ως γρασίδι, 2 περιπτώσεις ως μόνιμες καλλιέργειες, 8 περιπτώσεις ως πατάτες, 14 περιπτώσεις ως ανοιξιάτικες καλλιέργειες σιτηρών, 19 περιπτώσεις ως χειμερινές καλλιέργειες σιτηρών και 4 περιπτώσεις ως ελαιοκράμβη. Συνολικά, ο ταξινομητής απέδωσε αρκετά καλά στον προσδιορισμό των αγρανάπαυσης.

**Γρασίδι:** Το μοντέλο πέτυχε υψηλή ακρίβεια στην ταξινόμηση των συγκεκριμένων καλλιεργειών, εντοπίζοντας σωστά 3648 περιπτώσεις. Ωστόσο, υπήρξαν ορισμένες εσφαλμένες ταξινομήσεις, με 24 περιπτώσεις να ταξινομούνται ως αγρανάπαυση, 29 περιπτώσεις ως μόνιμες καλλιέργειες, 8 περιπτώσεις ως πατάτες, 49 περιπτώσεις ως ανοιξιάτικες καλλιέργειες σιτηρών, 31 περιπτώσεις ως χειμερινές καλλιέργειες σιτηρών και 4 περιπτώσεις ως ελαιοκράμβη.

**Μόνιμες καλλιέργειες:** Ο ταξινομητής δυσκολεύτηκε να αναγνωρίσει σωστά περιπτώσεις μόνιμων καλλιεργειών, με μόνο 28 περιπτώσεις να ταξινομούνται σωστά. Υπήρξαν εσφαλμένες ταξινομήσεις, με 5 περιπτώσεις να ταξινομούνται ως αγρανάπαυση, 59 περιπτώσεις ως γρασίδι, 1 περίπτωση ως πατάτες, 1 περίπτωση ως ανοιξιάτικες καλλιέργειες σιτηρών, 6 περιπτώσεις ως χειμερινές καλλιέργειες σιτηρών και καμία περίπτωση ως ελαιοκράμβη.

**Πατάτες:** Ο Random Forest είχε καλή απόδοση στην ταξινόμηση των καλλιεργειών πατάτας, προσδιορίζοντας σωστά 120 περιπτώσεις. Ωστόσο, υπήρξαν ορισμένες εσφαλμένες ταξινομήσεις, με 12 περιπτώσεις να ταξινομούνται ως αγρανάπαυση, 4 περιπτώσεις ως χόρτο, 18 περιπτώσεις ως ανοιξιάτικες καλλιέργειες σιτηρών, 5 περιπτώσεις ως χειμερινές καλλιέργειες σιτηρών και καμία περίπτωση ως ελαιοκράμβη.

**Ανοιξιάτικες καλλιέργειες σιτηρών:** Ο ταξινομητής πέτυχε υψηλή ακρίβεια στην ταξινόμηση των συγκεκριμένων καλλιεργειών, εντοπίζοντας σωστά 1460 περιπτώσεις. Υπήρξαν ορισμένες εσφαλμένες ταξινομήσεις, με 24 περιπτώσεις να ταξινομούνται ως αγρανάπαυση, 28 περιπτώσεις ως γρασίδι, 15 περιπτώσεις ως πατάτες, 49 περιπτώσεις ως χειμερινές καλλιέργειες σιτηρών, 2 περιπτώσεις ως ελαιοκράμβη και καμία περίπτωση ως μόνιμη καλλιέργεια.

**Χειμερινές καλλιέργειες σιτηρών:** Ο ταξινομητής είχε καλή απόδοση στην ταξινόμηση των χειμερινών καλλιεργειών σιτηρών, προσδιορίζοντας σωστά 1743 περιπτώσεις. Ωστόσο, υπήρξαν εσφαλμένες ταξινομήσεις, με 22 περιπτώσεις να ταξινομούνται ως αγρανάπαυση, 28 περιπτώσεις ως γρασίδι, 2 περιπτώσεις ως μόνιμες καλλιέργειες, 4 περιπτώσεις ως πατάτες, 56 περιπτώσεις ως ανοιξιάτικες καλλιέργειες σιτηρών και 8 περιπτώσεις ως ελαιοκράμβη.

**Ελαιοκράμβη:** Ο ταξινομητής πέτυχε υψηλή ακρίβεια στην ταξινόμηση της ελαιοκράμβης εντοπίζοντας σωστά 405 περιπτώσεις. Υπήρξαν ορισμένες εσφαλμένες ταξινομήσεις, με 3 περιπτώσεις να ταξινομούνται ως αγρανάπαυση, 3 περιπτώσεις ως γρασίδι, 2 περιπτώσεις ως πατάτες, 6 περιπτώσεις ως ανοιξιάτικες καλλιέργειες σιτηρών, 18 περιπτώσεις ως χειμερινές καλλιέργειες σιτηρών και καμία περίπτωση ως μόνιμη καλλιέργεια.

Συνολικά, ο Random Forest έδειξε εξαιρετική απόδοση στην ταξινόμηση ορισμένων καλλιεργειών όπως το γρασίδι, οι ανοιξιάτικες καλλιέργειες σιτηρών, οι χειμερινές καλλιέργειες σιτηρών και η ελαιοκράμβη, ενώ επέδειξε ικανοποιητικά αποτελέσματα στις καλλιέργειες της αγρανάπαυσης και της πατάτας. Ωστόσο, δυσκολεύτηκε με τον εντοπισμό μόνιμων καλλιεργειών.

#### *5.2.1.2. Ταξινομητής Linear SVM*

| Κατηγορία | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **Αγρανάπαυση** | 0.51 | 0.73 | 0.60 | 192 |
| **Γρασίδι** | 0.97 | 0.92 | 0.95 | 3793 |
| **Μόνιμες Καλλιέργειες** | 0.26 | 0.53 | 0.35 | 100 |
| **Πατάτες** | 0.73 | 0.80 | 0.76 | 159 |
| **Ανοιξιάτικες καλλιέργειες σιτηρών** | 0.93 | 0.92 | 0.92 | 1578 |
| **Χειμερινές καλλιέργειες σιτηρών** | 0.94 | 0.96 | 0.95 | 1863 |
| **Ελαιοκράμβη** | 0.95 | 0.92 | 0.94 | 437 |
| **Ακρίβεια** | | | 0.92 | 8122 |
| **Μέσος όρος** | 0.76 | 0.83 | 0.78 | 8122 |
| **Σταθμισμένος Μέσος Όρος** | 0.93 | 0.92 | 0.92 | 8122 |
| **Χρόνο υλοποίησης: 20 min.** | | | | |

**Αγρανάπαυση:** Ο ταξινομητής πέτυχε σχετικά χαμηλή ακρίβεια και F1-score, υποδεικνύοντας ότι υπήρξε δυσκολία στον σωστό προσδιορισμό αυτής της κατηγορίας. Ωστόσο, είχε μια μέτρια ανάκληση 73%, υποδηλώνοντας ότι αποτύπωσε ένα ικανοποιητικό μέρος των πραγματικών καλλιεργειών αγρανάπαυσης.

**Γρασίδι:** Το μοντέλο απέδωσε εξαιρετικά, επιτυγχάνοντας υψηλή ακρίβεια, ανάκληση και F1-score. Αυτό δείχνει ότι εντόπισε με ακρίβεια περιπτώσεις Γρασιδιού στο σύνολο δεδομένων.

**Μόνιμες καλλιέργειες:** Ο Random Forest δυσκολεύτηκε να ταξινομήσει περιπτώσεις Μόνιμων καλλιεργειών και πέτυχε χαμηλές τιμές σε ακρίβεια, ανάκληση και F1-score.

**Πατάτες:** Το μοντέλο πέτυχε σχετικά καλές τιμές σε ακρίβεια, ανάκληση και F1-score ,υποδεικνύοντας ότι είχε ικανοποιητική απόδοση στον εντοπισμό περιπτώσεων καλλιεργειών πατάτας.

**Ανοιξιάτικες καλλιέργειες σιτηρών:** Το μοντέλο πέτυχε υψηλή ακρίβεια, ανάκληση και F1-score για την κατηγορία, υποδηλώνοντας ότι ταξινόμησε με μεγάλη ακρίβεια περιπτώσεις αυτής της κατηγορίας.

**Χειμερινές καλλιέργειες σιτηρών:** Στη συγκεκριμένη κατηγορία, ο ταξινομητής απέδωσε πολύ καλά, επιτυγχάνοντας υψηλή ακρίβεια, ανάκληση και F1-score.

**Ελαιοκράμβη:** Όσον αφορά τις καλλιέργειες της ελαιοκράμβης, ο Random Forest πέτυχε υψηλή ακρίβεια, ανάκληση και F1-score, υποδεικνύοντας ότι εντόπισε με υψηλή ακρίβεια περιπτώσεις καλλιεργειών ελαιοκράμβης.

Ο ταξινομητής Linear SVM, πέτυχε εξαιρετικά αποτελέσματα στις κατηγορίες γρασίδι, ελαιοκράμβη, χειμερινές και ανοιξιάτικες καλλιέργειες σιτηρών και ,παράλληλα, πέτυχε ικανοποιητικά αποτελέσματα για τις κατηγορίες πατάτες, μέτρια αποτελέσματα για την κατηγορία αγρανάπαυση και πολύ χαμηλά αποτελέσματα για την κατηγορία μόνιμες καλλιέργειες. Το F1-score του σταθμισμένου μέσου όρου (0.92) είναι υψηλότερη από την F1-score του μέσου όρου (0.78), υποδηλώνοντας ότι το μοντέλο απέδωσε καλύτερα όταν λαμβάνεται υπόψιν κατανομή των περιπτώσεων κάθε κατηγορίας. Αυτό σημαίνει πως το μοντέλο είναι ιδιαίτερα αποτελεσματικό στην ταξινόμηση των κατηγοριών με μεγάλο αριθμό δειγμάτων υποστήριξης όπως το γρασίδι, οι ανοιξιάτικες καλλιέργειες σιτηρών, οι χειμερινές καλλιέργειες σιτηρών και η ελαιοκράμβη.

Στη συνέχεια παρουσιάζεται ο πίνακας σύγχυσης για την περαιτέρω αξιολόγηση του Linear SVM.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Αγρανάπαυση | 140 | 19 | 5 | 7 | 9 | 9 | 3 |
| Γρασίδι | 64 | 3503 | 136 | 16 | 44 | 22 | 8 |
| Μόνιμες Καλλιέργειες | 13 | 34 | 53 | 0 | 0 | 0 | 0 |
| Πατάτες | 13 | 3 | 1 | 127 | 8 | 6 | 1 |
| Ανοιξιάτικες Καλλιέργειες Σιτηρών | 27 | 26 | 2 | 19 | 1446 | 56 | 2 |
| Χειμερινές Καλλιέργειες Σιτηρών | 12 | 14 | 4 | 3 | 39 | 1783 | 8 |
| Ελαιοκράμβη | 6 | 2 | 5 | 2 | 3 | 15 | 404 |
| | Αγρανάπαυση | Γρασίδι | Μόνιμες Καλλιέργειες | Πατάτες | Ανοιξιάτικες Καλλιέργειες Σιτηρών | Χειμερινές Καλλιέργειες Σιτηρών | Ελαιοκράμβη |

**Αγρανάπαυση:** Ο ταξινομητής Linear SVM ταξινόμησε σωστά 140 περιπτώσεις αγρανάπαυσης, ενώ απέτυχε σε 52 περιπτώσεις που απέδωσε διαφορετικές κατηγορίες.

**Γρασίδι:** Στη συγκεκριμένη καλλιέργεια το μοντέλο πέτυχε εξαιρετικά αποτελέσματα, ταξινομώντας σωστά 3503 περιπτώσεις, ενώ απέτυχε σε 290 περιπτώσεις.

**Μόνιμες καλλιέργειες:** Όσον αφορά τις μόνιμες καλλιέργειες, ο ταξινομητής δεν πέτυχε ικανοποιητικά αποτελέσματα. Ταξινόμησε σωστά 53 περιπτώσεις, ενώ απέτυχε σε 47 περιπτώσεις.

**Πατάτες:** Αναφορικά με την κατηγορία πατάτες, ταξινόμησε σωστά 127 περιπτώσεις και απέτυχε σε 32 περιπτώσεις.

**Ανοιξιάτικες καλλιέργειες σιτηρών:** Στην κατηγορία ανοιξιάτικες καλλιέργειες σιτηρών, ο ταξινομητής πέτυχε πολύ καλά αποτελέσματα, ταξινομώντας 1446 περιπτώσεις σωστά, ενώ απέτυχε σε 132 περιπτώσεις.

**Χειμερινές καλλιέργειες σιτηρών:** Ο Linear SVM στη συγκεκριμένη κατηγορία ταξινόμησε σωστά 1783 περιπτώσεις, ενώ 80 περιπτώσεις δεν ταξινομήθηκαν σωστά και επέδειξε εξαιρετικά αποτελέσματα.

**Ελαιοκράμβη:** Στη συγκεκριμένη κατηγορία το μοντέλο τα πήγε πολύ καλά, καθώς ταξινομήθηκαν σωστά 404 περιπτώσεις, ενώ 33 περιπτώσεις ταξινομήθηκαν λάθος.

Γενικά, ο ταξινομητής Linear SVM πέτυχε ικανοποιητικά αποτελέσματα, ειδικά στις κατηγορίες γρασίδι, ελαιοκράμβη, ανοιξιάτικες και χειμερινές καλλιέργειες σιτηρών. Ωστόσο, στην κατηγορία μόνιμες καλλιέργειες δυσκολεύτηκε και πέτυχε χαμηλά αποτελέσματα.

#### *5.2.1.3 Ταξινομητής Gradient Boosting*

| Κατηγορία | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **Αγρανάπαυση** | 0.73 | 0.59 | 0.66 | 192 |
| **Γρασίδι** | 0.95 | 0.97 | 0.96 | 3793 |
| **Μόνιμες Καλλιέργειες** | 0.30 | 0.11 | 0.16 | 100 |
| **Πατάτες** | 0.76 | 0.68 | 0.72 | 159 |
| **Ανοιξιάτικες καλλιέργειες σιτηρών** | 0.91 | 0.92 | 0.92 | 1578 |
| **Χειμερινές καλλιέργειες σιτηρών** | 0.93 | 0.95 | 0.94 | 1863 |
| **Ελαιοκράμβη** | 0.93 | 0.93 | 0.93 | 437 |
| **Ακρίβεια** | | | 0.93 | 8122 |
| **Μέσος όρος** | 0.79 | 0.74 | 0.75 | 8122 |
| **Σταθμισμένος Μέσος Όρος** | 0.92 | 0.93 | 0.92 | 8122 |
| **Χρόνο υλοποίησης: 51 min.** | | | | |

**Αγρανάπαυση:** Το μοντέλο πέτυχε σχετικά ικανοποιητική ακρίβεια 0.73, αλλά χαμηλή ανάκληση 0.59, υποδηλώνοντας πως δυσκολεύτηκε στον σωστό προσδιορισμό αυτής της κατηγορίας.

**Γρασίδι:** Το μοντέλο απέδωσε αρκετά καλά για τη συγκεκριμένη κατηγορία, επιτυγχάνοντας υψηλή ακρίβεια, ανάκληση και F1-score. Αυτό σημαίνει ότι εντόπισε με ακρίβεια περιπτώσεις Γρασιδιού στο σύνολο των δεδομένων.

**Μόνιμες καλλιέργειες:** Ο Gradient Boosting απέτυχε σε αυτή την κατηγορία, καθώς παρατηρούνται αρκετά χαμηλές τιμές σε ακρίβεια και ανάκληση, υποδηλώνοντας πως το μοντέλο πραγματοποίησε αρκετές εσφαλμένες ταξινομήσεις και δεν εντόπισε περιπτώσεις μόνιμων καλλιεργειών.

**Πατάτες:** Στη συγκεκριμένη κατηγορία, ο ταξινομητής πέτυχε τιμές 0.76 και 0.68 για την ακρίβεια και ανάκληση , υποδηλώνοντας μία ικανοποιητική διάκριση των περιπτώσεων αυτής της κατηγορίας.

**Ανοιξιάτικες καλλιέργειες σιτηρών:** Το μοντέλο πέτυχε υψηλή ακρίβεια, ανάκληση και F1-score για την κατηγορία, υποδηλώνοντας ότι ταξινόμησε με μεγάλη ακρίβεια περιπτώσεις αυτής της κατηγορίας.

**Χειμερινές καλλιέργειες σιτηρών:** Στη συγκεκριμένη καλλιέργεια ο Gradient Boostingαπέδωσε εξαιρετικά, επιτυγχάνοντας υψηλή ακρίβεια, ανάκληση και F1-score.

**Ελαιοκράμβη:** Το μοντέλο πέτυχε υψηλή ακρίβεια, ανάκληση και F1-score, υποδεικνύοντας ότι εντόπισε με ακρίβεια περιπτώσεις καλλιεργειών ελαιοκράμβης.

Συνολικά, ο Gradient Boosting πέτυχα αρκετά υψηλές τιμές σε ακρίβεια και σταθμισμένο μέσο όρο. Το F1-score του σταθμισμένου μέσου όρου (0.92) είναι υψηλότερη από το F1-score του μέσος όρος. (0.78), υποδηλώνοντας ότι το μοντέλο απέδωσε καλύτερα όταν λαμβάνεται υπόψη η κατανομή της κατηγορίας. Αυτό σημαίνει πως το μοντέλο είναι ιδιαίτερα αποτελεσματικό στην ταξινόμηση των κατηγοριών με μεγάλο αριθμό δειγμάτων υποστήριξης όπως το γρασίδι, οι ανοιξιάτικες καλλιέργειες σιτηρών, οι χειμερινές καλλιέργειες σιτηρών και η ελαιοκράμβη.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Αγρανάπαυση | 114 | 31 | 2 | 7 | 15 | 20 | 3 |
| Γρασίδι | 12 | 3608 | 17 | 9 | 48 | 23 | 4 |
| Μόνιμες Καλλιέργειες | 2 | 73 | 11 | 1 | 3 | 8 | 2 |
| Πατάτες | 10 | 6 | 1 | 108 | 25 | 7 | 2 |
| Ανοιξιάτικες Καλλιέργειες Σιτηρών | 10 | 36 | 4 | 10 | 1456 | 56 | 6 |
| Χειμερινές Καλλιέργειες Σιτηρών | 8 | 27 | 2 | 4 | 47 | 1761 | 14 |
| Ελαιοκράμβη | 0 | 2 | 0 | 3 | 2 | 22 | 408 |
| | Αγρανάπαυση | Γρασίδι | Μόνιμες Καλλιέργειες | Πατάτες | Ανοιξιάτικες Καλλιέργειες Σιτηρών | Χειμερινές Καλλιέργειες Σιτηρών | Ελαιοκράμβη |

**Αγρανάπαυση:** Ο ταξινομητής Gradient Boosting ταξινόμησε σωστά 114 περιπτώσεις αγρανάπαυσης, ενώ απέτυχε σε 78 περιπτώσεις που απέδωσε διαφορετικές κατηγορίες.

**Γρασίδι:** Στη συγκεκριμένη καλλιέργεια ο ταξινομητής πέτυχε εξαιρετικά αποτελέσματα, ταξινομώντας σωστά 3608 περιπτώσεις, και ταξινομώντας λανθασμένα σε 113 περιπτώσεις.

**Μόνιμες καλλιέργειες:** Όσον αφορά τις μόνιμες καλλιέργειες, ο ταξινομητής δεν πέτυχε καθόλου καλά αποτελέσματα. Ταξινόμησε σωστά 11 περιπτώσεις, ενώ απέτυχε σε 89 περιπτώσεις.

**Πατάτες:** Για τη συγκεκριμένη κατηγορία, το μοντέλο ταξινόμησε σωστά 108 περιπτώσεις και απέτυχε σε 51 περιπτώσεις.

**Ανοιξιάτικες καλλιέργειες σιτηρών:** Στην κατηγορία ανοιξιάτικες καλλιέργειες σιτηρών, ο Gradient Boosting πέτυχε πολύ καλά αποτελέσματα, ταξινομώντας 1456 περιπτώσεις σωστά, ενώ απέτυχε σε 122 περιπτώσεις.

**Χειμερινές καλλιέργειες σιτηρών:** Ο ταξινομητής στη συγκεκριμένη κατηγορία ταξινόμησε σωστά 1761 περιπτώσεις, ενώ 102 περιπτώσεις δεν ταξινομήθηκαν σωστά, πετυχαίνοντας αρκετά καλά αποτελέσματα.

**Ελαιοκράμβη:** Στη συγκεκριμένη κατηγορία ταξινομήθηκαν σωστά 408 περιπτώσεις, ενώ 29 περιπτώσεις ταξινομήθηκαν λάθος.

### *5.2.1.4 Σύγκριση Ταξινομητών*

**Ταξινομητής Randon  Forest**

Ακρίβεια: 0.93

Μέσος Όρος: 0.78

F1-score (σταθμισμένος μέσος όρος): 0.93

Ο Random Forest πέτυχε ακρίβεια 0.93, υποδεικνύοντας ότι προέβλεψε σωστά τις κατηγορίες των καλλιεργειών για το 93% των περιπτώσεων. Το F1-score του σταθμισμένου μέσου όρου, 0.93, υποδηλώνει μια καλή ισορροπία μεταξύ ακρίβειας και ανάκλησης σε όλες τις κατηγορίες. Ο Random Forest είχε καλή απόδοση σε τέσσερις κατηγορίες, πετυχαίνοντας F1-score μεγαλύτερο του 0.92 στις κατηγορίες γρασίδι, ανοιξιάτικες καλλιέργειες σιτηρών, χειμερινές καλλιέργειες σιτηρών και ελαιοκράμβη. Ωστόσο, στις υπόλοιπες τρεις κατηγορίες πέτυχε F1-score μεταξύ 0.38-0.76. Συνολικά, ο Random Forest τα πήγε αρκετά καλά για την ταξινόμηση των καλλιεργειών.

**Ταξινομητής Linear SVM**

Ακρίβεια: 0.92

Μέσος Όρος: 0.78

F1-score(σταθμισμένος μέσος όρος): 0.92

Ο Linear SVM πέτυχε ακρίβεια 0.92. Το F1-score του σταθμισμένου μέσου όρου, 0.92, δείχνει μια καλή συνολική απόδοση του μοντέλου. Σε σύγκριση με τον Random Forest, ο Linear SVM έδειξε καλύτερη απόδοση στην κατηγορία χειμερινές καλλιέργειες σιτηρών και χαμηλότερη απόδοση στις κατηγορίες  αγρανάπαυση, γρασίδι και ελαιοκράμβη.

**Ταξινομητής Gradient Boosting**

Ακρίβεια: 0.93

Μέσος Όρος: 0.75

F1-score (σταθμισμένος μέσος όρος): 0.92

Ο αλγόριθμος Gradient Boosting πέτυχε ακρίβεια 0.93, ίδια με του Random Forest. Το F1-score του σταθμισμένου μέσου όρου, 0.92, υποδηλώνει σταθερή απόδοση σε όλες τις κατηγορίες, παρόμοια με του ταξινομητή Linear SVM. Σε σχέση με τον Linear SVM, ο Gradient Boosting επέδειξε ελαφρώς καλύτερα αποτελέσματα στις κατηγορίες αγρανάπαυση και γρασίδι, ενώ επέδειξε χειρότερα αποτελέσματα στις κατηγορίες μόνιμες καλλιέργειες και πατάτες.

Σε σχέση με τον Random Forest, ο Gradient Boosting επέδειξε ελαφρώς καλύτερα αποτελέσματα στις κατηγορίες αγρανάπαυση και χειμερινές καλλιέργειες σιτηρών, ενώ στις κατηγορίες μόνιμες καλλιέργειες, πατάτες και ελαιοκράμβη είχε χειρότερα αποτελέσματα.

Λαμβάνοντας υπόψιν τις μετρήσεις της ακρίβειας, του μέσου όρου, του σταθμισμένου όρου και του χρόνου ολοκλήρωσης κάθε αλγορίθμου, ο Random Forest κρίνεται ο πιο κατάλληλος για την ταξινόμηση του συγκεκριμένου συνόλου δεδομένων. Είναι επίσης σημαντικό πως ο

Random Forest επέδειξε τα καλύτερα αποτελέσματα συνολικά για τις κλάσεις που δεν έχουν μεγάλο αριθμό δειγμάτων, όπως η αγρανάπαυση, οι πατάτες και οι μόνιμες καλλιέργειες.

Στους παρακάτω πίνακες συνοψίζονται οι τιμές ακρίβειας, ανάκλησης και F1-score και για τους τρεις αλγόριθμους, Random Forest, Linear SVM και Gradient Boosting.

#### *5.2.1.5 Ανάλυση Random Forest*

Καθώς επιλέχθηκε ο Random Forest ως ο καλύτερος, έγινε μια περαιτέρω ανάλυση και υπολογίστηκε ο πίνακας producer accuracy, ο οποίος αναλύει το recall, δηλαδή το ποσοστό των περιπτώσεων που ταξινομήθηκαν σωστά από τον Random Forest για κάθε προβλεπόμενη κατηγορία.

| Crop Code | Crop Name | Declared parcels | Well Classified | Producer Accuracy | Confusion class 1 | 1% | Confusion class 2 | 2% | Confusion class 3 | 3% |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Black fallow | 192 | 127 | 0.66 | Winter cereals | 0.1 | Grass | 0.09 | Spring cereals | 0.07 |
| 1 | Grass | 3793 | 3648 | 0.96 | Spring cereals | 0.01 | Winter cereals | 0.01 | Permanent crops | 0.01 |
| 2 | Permanent crops | 100 | 28 | 0.28 | Grass | 0.59 | Winter cereals | 0.06 | Black fallow | 0.05 |
| 3 | Potatoes | 159 | 120 | 0.76 | Spring cereals | 0.11 | Black fallow | 0.08 | Winter cereals | 0.03 |
| 4 | Spring cereals | 1578 | 1460 | 0.93 | Winter cereals | 0.03 | Grass | 0.02 | Black fallow | 0.02 |
| 5 | Winter cereals | 1863 | 1743 | 0.94 | Spring cereals | 0.03 | Grass | 0.02 | Black fallow | 0.01 |
| 6 | Winter rape | 437 | 405 | 0.93 | Winter cereals | 0.04 | Spring cereals | 0.01 | Black fallow | 0.01 |

**Αγρανάπαυση:** Με ακρίβεια 66%, η κατηγορία αγρανάπαυση είχε σημαντική σύγχυση με τις κατηγορίες χειμερινών καλλιεργειών σιτηρών (10%) και το γρασίδι (9%).

**Γρασίδι:** Η κατηγορία πέτυχε υψηλή ακρίβεια 96%, αλλά υπήρξαν μικρές περιπτώσεις σύγχυσης με τις ανοιξιάτικες καλλιέργειες σιτηρών, τις χειμερινές καλλιέργειες σιτηρών και τις μόνιμες καλλιέργειες με 1% για κάθε κατηγορία.

**Μόνιμες Καλλιέργειες:** Αυτή η κατηγορία είχε χαμηλή ακρίβεια 28%, με σημαντική σύγχυση 59% με το γρασίδι, τις χειμερινές καλλιέργειες σιτηρών (6%) και την αγρανάπαυση (5%).

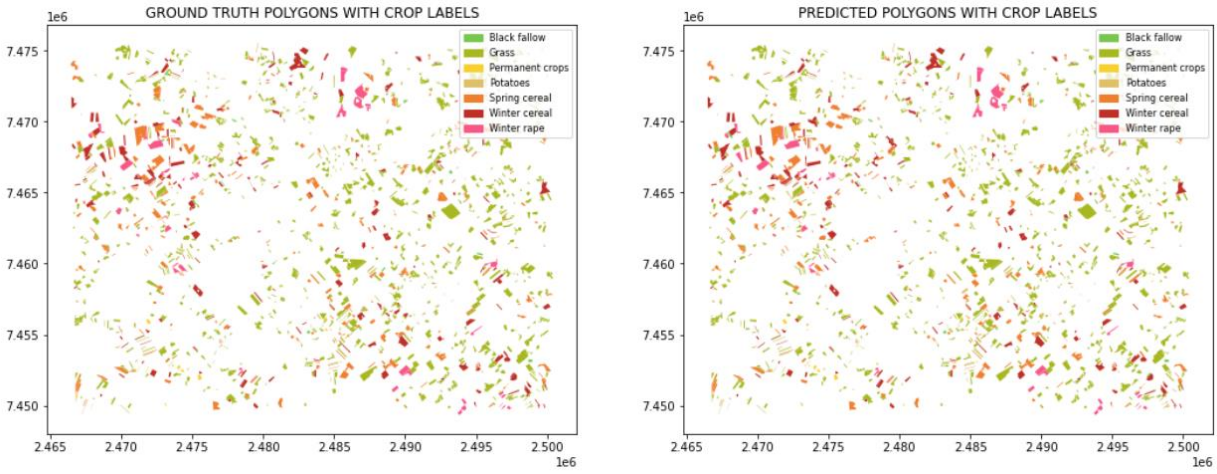
**Πατάτες:** Οι πατάτες παρουσίασαν ακρίβεια 76%, αλλά υπήρξαν συγχύσεις με τις ανοιξιάτικες καλλιέργειες σιτηρών (11%),την κατηγορία αγρανάπαυση (8%) και τις χειμερινές καλλιέργειες σιτηρών (3%).

**Ανοιξιάτικες καλλιέργειες σιτηρών:** Τα ανοιξιάτικα δημητριακά είχαν μεγάλη ακρίβεια 93% και μικρή σύγχυση με τις χειμερινές καλλιέργειες σιτηρών δημητριακά (3%),το γρασίδι (2%) και την ελαιοκράμβη (2%).

**Χειμερινές καλλιέργειες σιτηρών:** Με ακρίβεια 94%, οι χειμερινές καλλιέργειες σιτηρών είχαν καλή απόδοση, αλλά υπήρξαν περιπτώσεις μικρής σύγχυσης με τις ανοιξιάτικες καλλιέργειες σιτηρών (3%),το γρασίδι (2%) και την αγρανάπαυση (1%).

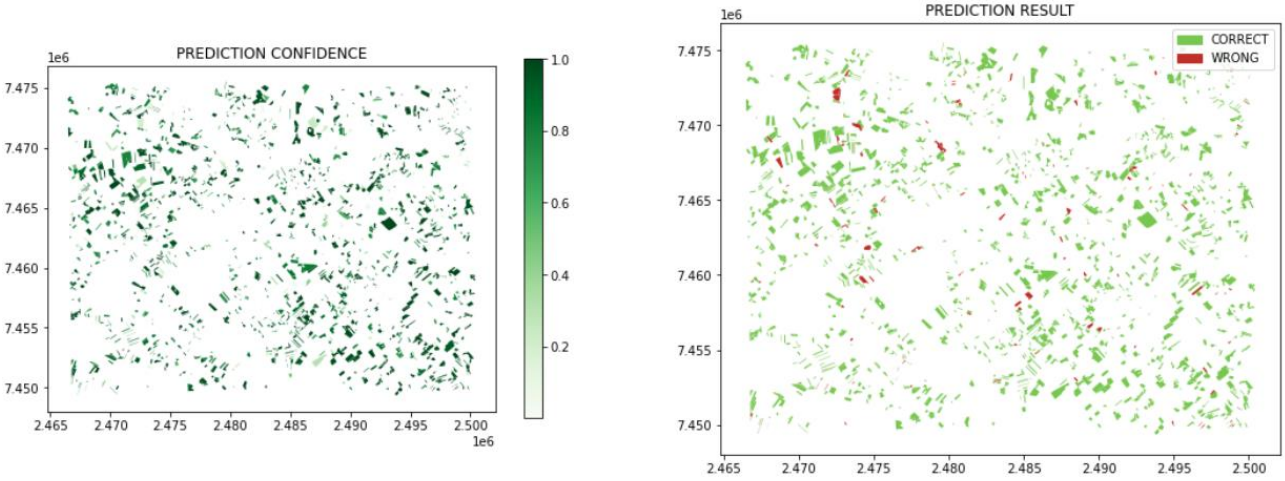
**Ελαιοκράμβη:** Η κατηγορία πέτυχε ακρίβεια παραγωγού 93% και παρουσίασε μικρή σύγχυση με τις χειμερινές καλλιέργειες σιτηρών (4%),τις ανοιξιάτικες καλλιέργειες σιτηρών (1%) και την αγρανάπαυση (1%).

Στις παρακάτω εικόνες, παρατηρούμε δύο σύνολα πολυγώνων που αντιπροσωπεύουν τα πραγματικά δεδομένα εδάφους και τα προβλεπόμενα δεδομένα εδάφους όπως προέκυψαν από τον Random Forest. Τα χρώματα αντιπροσωπεύουν διαφορετικές κατηγορίες καλλιεργειών.




**Εικόνα 246 - Δεδομένα εδάφους και προβλεπόμενες ταξινομήσεις**

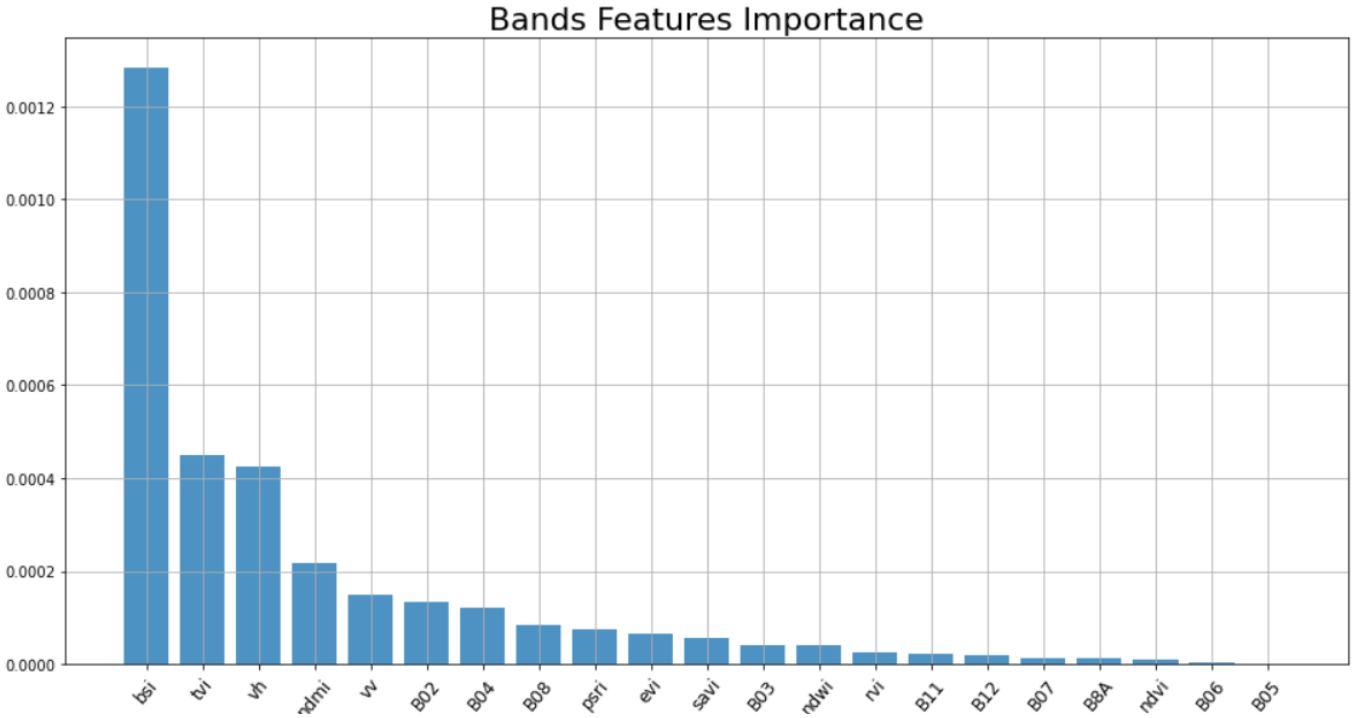
Στις παρακάτω εικόνες παρουσιάζεται αριστερά η εμπιστοσύνη των προβλέψεων για τις καλλιέργειες, με χρήση μιας παλέτας χρωμάτων που αντιπροσωπεύει την κλίμακα της εμπιστοσύνης. Οι περιοχές με υψηλή εμπιστοσύνη εμφανίζονται με πιο έντονο πράσινο χρώμα. Στη δεξιά εικόνα, παρουσιάζεται το αποτέλεσμα της πρόβλεψης για τις καλλιέργειες, χρησιμοποιώντας δύο χρώματα για τις δύο πιθανές κατηγορίες, σωστή και λάθος ταξινόμηση. Οι περιοχές με το πράσινο χρώμα υποδεικνύουν την σωστή κατηγορία, ενώ οι περιοχές με το κόκκινο χρώμα υποδεικνύουν τα αγροτεμάχια που ταξινομήθηκαν σε λάθος κατηγορία.




**Εικόνα 37 - Επίπεδα εμπιστοσύνης και αποτελέσματα ταξινομήσεων**

### 5.2.2 Σημαντικότητα Χαρακτηριστικών

Σκοπός της συγκεκριμένης τεχνικής είναι να αναδειχθούν τα τρία χαρακτηριστικά που συμβάλουν περισσότερο στην ταξινόμηση του Random Forest στο σύνολο των δεδομένων.




**Εικόνα 38 - Σημαντικότητα χαρακτηριστικών**

Από το παραπάνω διάγραμμα φαίνεται πως τα τρία πιο σημαντικά χαρακτηριστικά που συμμετείχαν στη διαδικασία της επιβλεπόμενης ταξινόμησης με τον Random Forest είναι οι δείκτες Bare Soil Index, Transformed Vegetation Index από τον Sentinel-2 και η πόλωση VH του Sentinel-1.

Έτσι, δημιουργήθηκε ένα καινούριο σύνολο δεδομένων, το οποίο αποτελείται από τους δείκτες BSI, TVI και τις τιμές VH και πραγματοποιήθηκε ταξινόμηση στα αγροτεμάχια με τον Random Forest. Το μέγεθος του καινούριου συνόλου δεδομένων αποτελείται από 27073 αγροτεμάχια και 22 τιμές για κάθε σημαντικό χαρακτηριστικό, δηλαδή 66 τιμές στο σύνολο.

| Κατηγορία | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **Αγρανάπαυση** | 0.65 | 0.68 | 0.66 | 192 |
| **Γρασίδι** | 0.96 | 0.96 | 0.96 | 3793 |
| **Μόνιμες Καλλιέργειες** | 0.39 | 0.19 | 0.26 | 100 |
| **Πατάτες** | 0.74 | 0.76 | 0.74 | 159 |
| **Ανοιξιάτικες καλλιέργειες σιτηρών** | 0.91 | 0.92 | 0.92 | 1578 |
| **Χειμερινές καλλιέργειες σιτηρών** | 0.93 | 0.93 | 0.93 | 1863 |
| **Ελαιοκράμβη** | 0.94 | 0.94 | 0.94 | 437 |
| **Ακρίβεια** | | | 0.93 | 8122 |
| **Μέσος όρος** | 0.79 | 0.77 | 0.77 | 8122 |
| **Σταθμισμένος Μέσος Όρος** | 0.92 | 0.93 | 0.92 | 8122 |
| **Χρόνο υλοποίησης: 3 sec.** | | | | |

**Αγρανάπαυση:** Το μοντέλο πέτυχε σχετικά ικανοποιητική ακρίβεια και F1-score, υποδεικνύοντας ότι υπήρξε δυσκολία στον σωστό προσδιορισμό αυτής της κατηγορίας.

**Γρασίδι:** Ο Random Forest απέδωσε εξαιρετικά καλά, επιτυγχάνοντας υψηλή ακρίβεια, ανάκληση και F1-score. Αυτό σημαίνει ότι εντόπισε με ακρίβεια περιπτώσεις Γρασιδιού στο σύνολο των δεδομένων.

**Μόνιμες καλλιέργειες:** Το μοντέλο απέτυχε σε αυτή την κατηγορία, καθώς παρατηρούνται αρκετά χαμηλές τιμές σε ακρίβεια και ανάκληση, υποδηλώνοντας πως το μοντέλο πραγματοποίησε αρκετές εσφαλμένες ταξινομήσεις και δεν εντόπισε περιπτώσεις μόνιμων καλλιεργειών.

**Πατάτες:** Το μοντέλο επέδειξε αξιοπρεπή ακρίβεια και ανάκληση, 0.74 και 0.76 αντίστοιχα, τα οποία δείχνουν μια λογική ικανότητα διάκρισης περιπτώσεων αυτής της κατηγορίας.

**Ανοιξιάτικες καλλιέργειες σιτηρών:** Το μοντέλο πέτυχε υψηλή ακρίβεια, ανάκληση και F1-score για την κατηγορία, υποδηλώνοντας ότι ταξινόμησε με μεγάλη ακρίβεια περιπτώσεις αυτής της κατηγορίας.

**Χειμερινές καλλιέργειες σιτηρών:** Το μοντέλο απέδωσε πολύ καλά για αυτή την κατηγορία, επιτυγχάνοντας υψηλή ακρίβεια, ανάκληση και F1-score.

**Ελαιοκράμβη:** Στη συγκεκριμένη κατηγορία, ο Random Forest πέτυχε υψηλή ακρίβεια, ανάκληση και F1-score, υποδεικνύοντας ότι εντόπισε με ακρίβεια περιπτώσεις καλλιεργειών ελαιοκράμβης.

Ο νέος Random Forest που λαμβάνει υπόψιν τα τρία πιο σημαντικά χαρακτηριστικά πέτυχε υψηλό σταθμισμένο F1-score, το οποίο σημαίνει πως με λιγότερα χαρακτηριστικά πέτυχε υψηλές ακρίβειες. Ωστόσο, σε σχέση με τον Random Forest που εκπαιδεύτηκε σε όλο το σύνολο των δεδομένων, ο νέος Random Forest πέτυχε χαμηλότερα F1-score στις τρεις κατηγορίες με τον μικρό αριθμό δειγμάτων, δηλαδή στην αγρανάπαυση, τις μόνιμες καλλιέργειες και τις πατάτες. Αυτό αναδεικνύει πως η σημαντικότητα χαρακτηριστικών είναι αρκετά καλή μέθοδος για να μειώσεις τον όγκο των δεδομένων και να κρατήσεις υψηλές ακρίβειες στο μοντέλο, αλλά πρώτα πρέπει να ελέγχεται ο αριθμός των δειγμάτων, ώστε να μην υπάρχει μεγάλη διαφορά στον αριθμό κάθε κλάσης.

Παρακάτω παρουσιάζεται ο πίνακας σύγχυσης για κάθε κατηγορία, όπως προέκυψε από την εφαρμογή του Random Forest στα τρία πιο σημαντικά χαρακτηριστικά.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Αγρανάπαυση | 130 | 23 | 0 | 9 | 8 | 18 | 4 |
| Γρασίδι | 18 | 3648 | 27 | 12 | 46 | 37 | 5 |
| Μόνιμες Καλλιέργειες | 5 | 66 | 19 | 3 | 0 | 6 | 1 |
| Πατάτες | 8 | 4 | 0 | 121 | 21 | 5 | 0 |
| Ανοιξιάτικες Καλλιέργειες Σιτηρών | 17 | 32 | 0 | 15 | 1458 | 50 | 6 |
| Χειμερινές Καλλιέργειες Σιτηρών | 18 | 31 | 3 | 6 | 63 | 1734 | 8 |
| Ελαιοκράμβη | 3 | 3 | 0 | 0 | 4 | 12 | 411 |
| | Αγρανάπαυση | Γρασίδι | Μόνιμες Καλλιέργειες | Πατάτες | Ανοιξιάτικες Καλλιέργειες Σιτηρών | Χειμερινές Καλλιέργειες Σιτηρών | Ελαιοκράμβη |

**Αγρανάπαυση:** Ο ταξινομητής Linear SVM ταξινόμησε σωστά 130 περιπτώσεις αγρανάπαυσης, ενώ απέτυχε σε 62 περιπτώσεις που απέδωσε διαφορετικές κατηγορίες.

**Γρασίδι:** Στη συγκεκριμένη καλλιέργεια ο ταξινομητής πέτυχε εξαιρετικά αποτελέσματα, ταξινομώντας σωστά 3648 περιπτώσεις, ενώ απέτυχε σε 145 περιπτώσεις.

**Μόνιμες καλλιέργειες:** Όσον αφορά τις μόνιμες καλλιέργειες, ο ταξινομητής απέτυχε. Ταξινόμησε σωστά 19 περιπτώσεις, ενώ απέτυχε σε 81 περιπτώσεις.

**Πατάτες:** Αναφορικά με την κατηγορία πατάτες, ταξινόμησε σωστά 121 περιπτώσεις και απέτυχε σε 38 περιπτώσεις.

**Ανοιξιάτικες καλλιέργειες σιτηρών:** Στην κατηγορία ανοιξιάτικες καλλιέργειες σιτηρών, ο Linear SVM πέτυχε πολύ καλά αποτελέσματα, ταξινομώντας 1458 περιπτώσεις σωστά, ενώ απέτυχε σε 120 περιπτώσεις.

**Χειμερινές καλλιέργειες σιτηρών:** Το μοντέλο στη συγκεκριμένη κατηγορία ταξινόμησε σωστά 1734 περιπτώσεις, ενώ 129 περιπτώσεις δεν ταξινομήθηκαν σωστά.

**Ελαιοκράμβη:** Στη συγκεκριμένη κατηγορία ταξινομήθηκαν σωστά 411 περιπτώσεις, ενώ 22 περιπτώσεις ταξινομήθηκαν λάθος.

Συγκρίνοντας τα αποτελέσματα των μοντέλων που προέκυψαν από τους ταξινομητές Random Forest, παρατηρείται ότι και τα δύο μοντέλα πέτυχαν παρόμοια απόδοση όσον αφορά το F1-score του μέσου και του σταθμισμένου μέσου όρου. Το δεύτερο μοντέλο, το οποίο βασίστηκε σε τρία χαρακτηριστικά του συνόλου δεδομένων και ολοκλήρωσε την ταξινόμηση σε λιγότερο χρόνο κατά 30%, παρείχε πετυχημένα αποτελέσματα. Αυτό υποδηλώνει ότι τα επιλεγμένα χαρακτηριστικά περιέχουν σημαντικές πληροφορίες που είναι απαραίτητες για την ταξινόμηση. Ωστόσο, πρέπει να σημειωθεί ότι το πρώτο μοντέλο που εκπαιδεύτηκε στο πλήρες σύνολο δεδομένων, έδειξε ελαφρώς βελτιωμένη συνολική απόδοση, ειδικότερα στις κατηγορίες με χαμηλό αριθμό δειγμάτων.

### 5.2.3. Μεταφορά Γνώσης

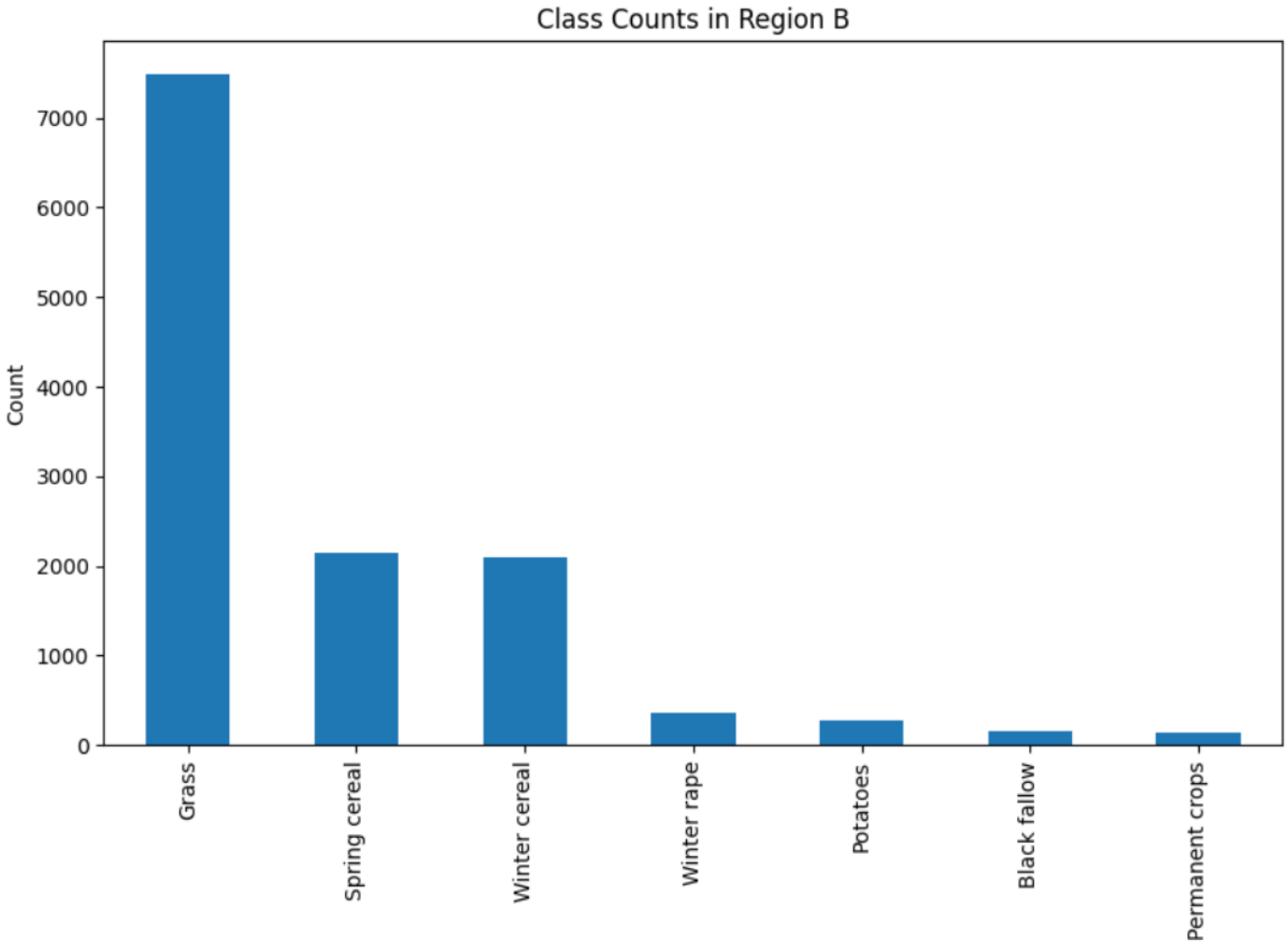
Σκοπός της συγκεκριμένης τεχνικής είναι να αξιολογηθεί η μεταφερσιμότητα μοντέλων Random Forest που εκπαιδεύτηκαν σε μία περιοχή ή σε συνδυασμό περιοχών και εφαρμόζονται σε άλλες περιοχές.

Τα κριτήρια επιλογής των περιοχών αφορούν τον συνολικό αριθμό των δειγμάτων ανά καλλιέργεια, έτσι ώστε να υπάρχει ένας ικανός αριθμός για να εκπαιδευτεί και να αξιολογηθεί το μοντέλο.

Για όλες τις εκπαιδεύσεις των μοντέλων χρησιμοποιήθηκε 70% για δεδομένα εκπαίδευσης και 30% για δεδομένα ελέγχου.

#### *5.2.3.1. Εκπαίδευση στην Περιοχή B*

Αρχικά, επιλέχθηκε η περιοχή B λόγω του μεγάλου αριθμού των δειγμάτων της. Η περιοχή B αποτελείται από 12658 αγροτεμάχια, δηλαδή το 47% του συνόλου των δεδομένων. Φαίνεται από το γράφημα πως τον μεγαλύτερο αριθμό αγροτεμαχίων καταλαμβάνει η καλλιέργεια γρασίδι, ενώ η αγρανάπαυση και οι μόνιμες καλλιέργειες καταλαμβάνουν το μικρότερο αριθμό δειγμάτων.




**Εικόνα 39 - Μέγεθος δείγματος ανά καλλιέργεια στην Περιοχή B**

Στη συνέχεια, το μοντέλο του Random Forest εκπαιδεύτηκε στην περιοχή Β με τα εξής αποτελέσματα.

| Κατηγορία | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **Αγρανάπαυση** | 0.54 | 0.58 | 0.56 | 48 |
| **Γρασίδι** | 0.97 | 0.96 | 0.96 | 2245 |
| **Μόνιμες Καλλιέργειες** | 0.38 | 0.28 | 0.32 | 43 |
| **Πατάτες** | 0.70 | 0.72 | 0.71 | 83 |
| **Ανοιξιάτικες καλλιέργειες σιτηρών** | 0.88 | 0.91 | 0.89 | 642 |
| **Χειμερινές καλλιέργειες σιτηρών** | 0.92 | 0.93 | 0.93 | 627 |
| **Ελαιοκράμβη** | 0.93 | 0.94 | 0.93 | 110 |
| **Ακρίβεια** | | | 0.92 | 3798 |
| **Μέσος όρος** | 0.76 | 0.76 | 0.76 | 3798 |
| **Σταθμισμένος Μέσος Όρος** | 0.92 | 0.92 | 0.92 | 3798 |
| **Χρόνο υλοποίησης: 3.5 sec.** | | | | |

Όπως είναι λογικό, το μοντέλο πέτυχε αρκετά ικανοποιητικά αποτελέσματα με υψηλή ακρίβεια 93% και F1-score 76% για το μέσο όρο και 92% για τον σταθμισμένο μέσο όρο. Ωστόσο, η κατηγορία των μόνιμων καλλιεργειών και της αγρανάπαυσης δεν παρουσίασαν καλά αποτελέσματα, όπως ήταν αναμενόμενο

### *5.2.3.2. Μεταφορά Μοντέλου Β στην Περιοχή Α*

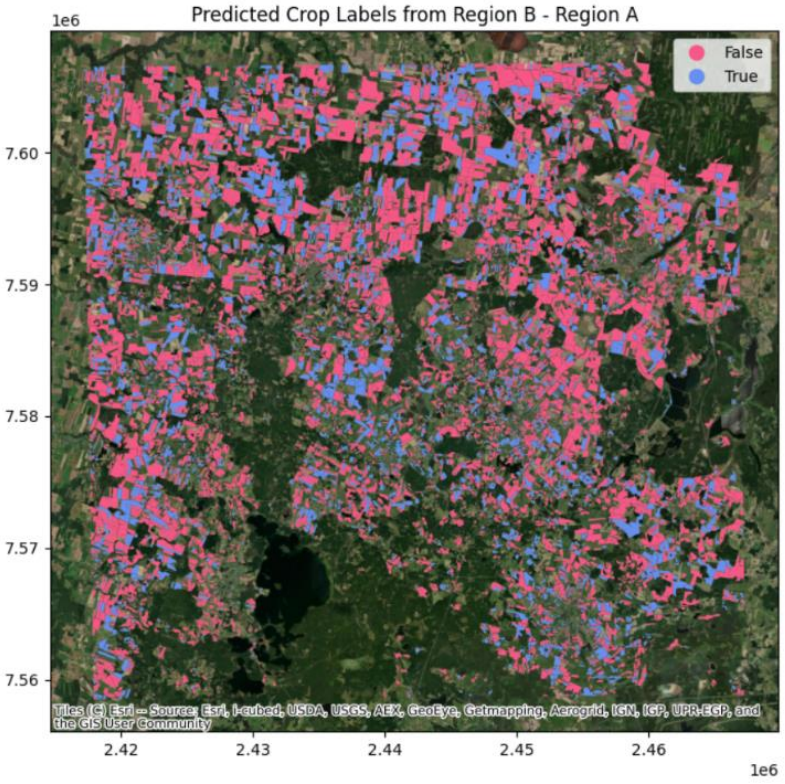
Αφού εκπαιδεύτηκε το μοντέλο στην Περιοχή Β, στη συνέχεια μεταφέρθηκε στην περιοχή Α και παρουσίασε τα εξής αποτελέσματα.

| Κατηγορία | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **Αγρανάπαυση** | 0.03 | 0.03 | 0.03 | 138 |
| **Γρασίδι** | 1.00 | 0 | 0 | 3005 |
| **Μόνιμες Καλλιέργειες** | 0.01 | 0.16 | 0.02 | 49 |
| **Πατάτες** | 0.25 | 0.30 | 0.27 | 128 |
| **Ανοιξιάτικες καλλιέργειες σιτηρών** | 0.51 | 0.96 | 0.66 | 1568 |
| **Χειμερινές καλλιέργειες σιτηρών** | 0.16 | 0.28 | 0.21 | 1317 |
| **Ελαιοκράμβη** | 0.00 | 0.00 | 0.00 | 111 |
| **Ακρίβεια** | | | 0.31 | 6316 |
| **Μέσος όρος** | 0.28 | 0.25 | 0.17 | 6316 |
| **Σταθμισμένος Μέσος Όρος** | 0.64 | 0.31 | 0.21 | 6316 |

Φαίνεται πως η ταξινόμηση μετά τη μεταφορά μάθησης στην περιοχή Α χρησιμοποιώντας το μοντέλο που εκπαιδεύτηκε στην περιοχή Β δείχνει αρκετά χαμηλή απόδοση στις περισσότερες κατηγορίες. Η ακρίβεια, η ανάκληση και το F1-score για τις περισσότερες καλλιέργειες είναι χαμηλές, υποδηλώνοντας χαμηλή ακρίβεια ταξινόμησης. Συγκεκριμένα, η κατηγορία γρασίδι έχει 1.0 ακρίβεια, αλλά ανάκληση 0. Αυτό υποδηλώνει ότι το μοντέλο έχει κάνει overfitting και προβλέπει εσφαλμένα την συγκεκριμένη κατηγορία για πολλές περιπτώσεις.

Επίσης, η κατηγορία ανοιξιάτικες καλλιέργειες σιτηρών έχουν 51% ακρίβεια και 96% ανάκληση, το οποίο σημαίνει  ότι το μοντέλο εντόπισε ικανοποιητικά ένα μεγάλο ποσοστό των πραγματικών περιπτώσεων της συγκεκριμένης καλλιέργειας. Η συνολική ακρίβεια είναι επίσης χαμηλή, στο 31%, υποδηλώνοντας σημαντικό ποσοστό εσφαλμένης ταξινόμησης. Το F1-score του μέσου και του σταθμισμένου μέσου όρου είναι χαμηλό, υποδηλώνοντας κακή συνολική απόδοση. Συμπερασματικά, η προσέγγιση μεταφοράς του μοντέλου από την περιοχή Β στην περιοχή Α χρησιμοποιώντας το μοντέλο τυχαίου δάσους δεν απέδωσε ικανοποιητικά αποτελέσματα.

Στον παρακάτω χάρτη φαίνονται οι σωστά και λανθασμένα ταξινομημένες κατηγορίες για την περιοχή Α.




**Εικόνα 40 - Χάρτης ταξινόμησης περιοχής Α**

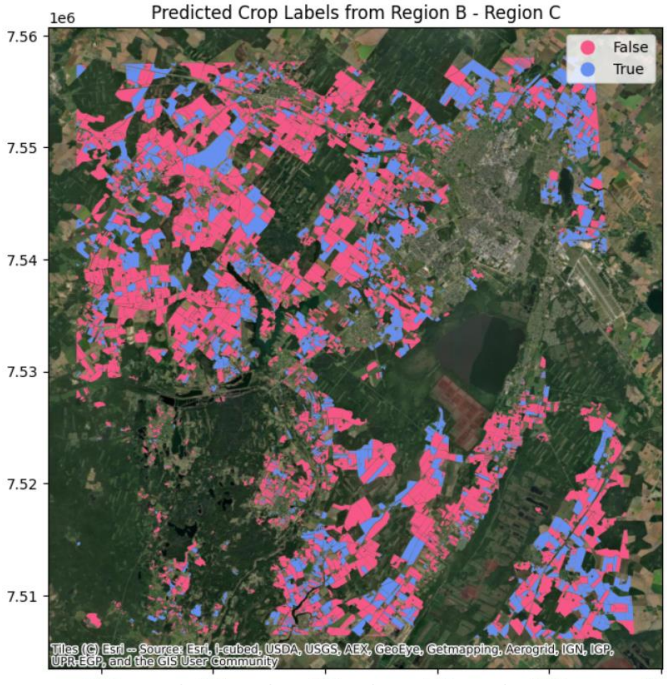
### *5.2.3.3. Μεταφορά Μοντέλου Β στην Περιοχή Γ*

Ύστερα από την Περιοχή Α, το μοντέλο της περιοχής Β εφαρμόστηκε και στην περιοχή Γ με τα παρακάτω αποτελέσματα.

| Κατηγορία | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **Αγρανάπαυση** | 0.02 | 0.01 | 0.01 | 199 |
| **Γρασίδι** | 0.82 | 0.01 | 0.03 | 608 |
| **Μόνιμες Καλλιέργειες** | 0.09 | 0.34 | 0.14 | 83 |
| **Πατάτες** | 0.01 | 0.04 | 0.02 | 49 |
| **Ανοιξιάτικες καλλιέργειες σιτηρών** | 0.34 | 0.94 | 0.49 | 616 |
| **Χειμερινές καλλιέργειες σιτηρών** | 0.59 | 0.56 | 0.57 | 1566 |
| **Ελαιοκράμβη** | 0.00 | 0.00 | 0.00 | 652 |
| **Ακρίβεια** | | | 0.39 | 3773 |
| **Μέσος όρος** | 0.27 | 0.27 | 0.18 | 3773 |
| **Σταθμισμένος Μέσος Όρος** | 0.43 | 0.39 | 0.33 | 3773 |

Τα αποτελέσματα της μεταφοράς γνώσης στην περιοχή Γ είναι εξίσου χαμηλά, αλλά το F1-score του σταθμισμένου μέσου όρου παρουσιάζει βελτίωση σε σχέση με τα αποτελέσματα της περιοχής Α κατά 12% . Η ακρίβεια, η ανάκληση, καθώς και η F1-score ποικίλουν για τις διαφορετικές κατηγορίες, αλλά είναι χαμηλές. Αναλυτικότερα, η κατηγορία γρασίδι έχει υψηλή ακρίβεια, αλλά σχεδόν μηδενική ανάκληση, το οποίο σημαίνει ότι το μοντέλο είχε κακή απόδοση στον προσδιορισμό της κατηγορίας. Επιπρόσθετα, το μοντέλο στην κατηγορία χειμερινές καλλιέργειες σιτηρών πέτυχε σχετικά ικανοποιητικά αποτελέσματα σε ακρίβεια και στην ανάκληση με 59% και 56% αντίστοιχα. Η συνολική ακρίβεια είναι 39%, υποδηλώνοντας ένα χαμηλό επίπεδο ταξινόμησης.

Στον παρακάτω χάρτη φαίνονται οι σωστά και λανθασμένα ταξινομημένες κατηγορίες για την περιοχή Γ.

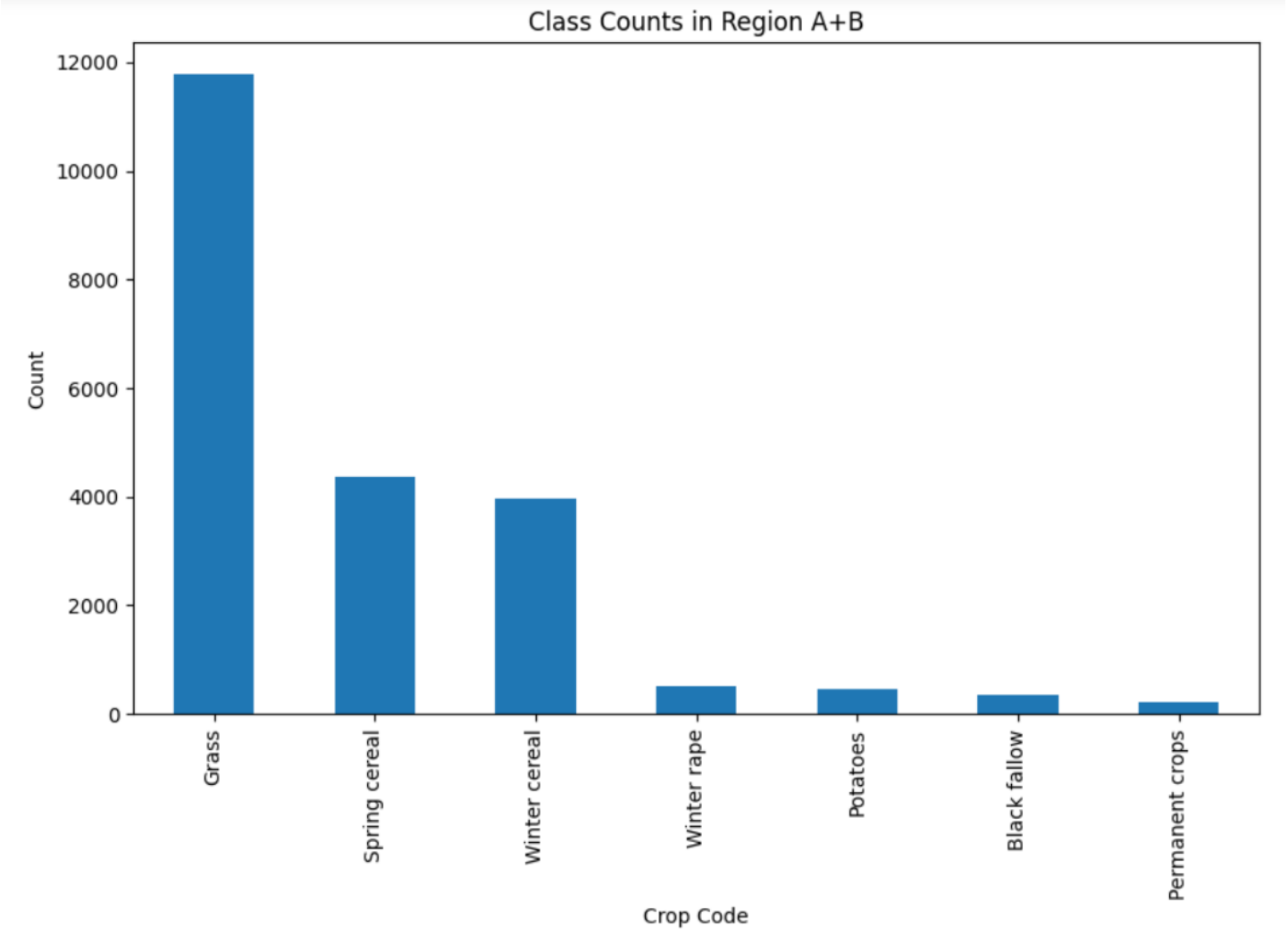



**Εικόνα 41 - Χάρτης ταξινόμησης περιοχής Γ**

### *5.2.3.4. Εκπαίδευση δεδομένων στις Περιοχές A+B*

Λόγω των κακών αποτελεσμάτων που παρείχε το μοντέλο που εκπαιδεύτηκε στην περιοχή B, επιλέχθηκε η ενίσχυση του με τα δεδομένα της περιοχής A, ώστε να αυξηθεί το δείγμα των καλλιεργειών.

Το νέο σύνολο δεδομένων που δημιουργήθηκε αποτελείται από 21682 αγροτεμάχια, δηλαδή το 80% των αγροτεμαχίων όλου του συνόλου δεδομένων . Όπως και στην προηγούμενη ενότητα που εκπαιδεύτηκε το μοντέλο στην περιοχή, έτσι και στο συνδυασμό των περιοχών A+B παρατηρείται μεγάλη ανισορροπία μεταξύ των κλάσεων. Για την κατηγορία γρασίδι υπάρχουν περίπου 12000 αγροτεμάχια διαθέσιμα, ενώ για τις κατηγορίες αγρανάπαυση, ελαιοκράμβη, μόνιμες καλλιέργειες και πατάτες υπάρχουν περίπου 500 αγροτεμάχια διαθέσιμα για κάθε κατηγορία.




**Εικόνα 42 - Μέγεθος δείγματος ανά καλλιέργεια στις περιοχές A και B**

Αρχικά, το νέο μοντέλο A+B εκπαιδεύτηκε με τον Random Forest στις περιοχές A+B και παρουσίασε τα παρακάτω αποτελέσματα.

| Κατηγορία | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **Αγρανάπαυση** | 0.64 | 0.64 | 0.64 | 107 |
| **Γρασίδι** | 0.96 | 0.96 | 0.96 | 3553 |
| **Μόνιμες Καλλιέργειες** | 0.38 | 0.20 | 0.27 | 64 |
| **Πατάτες** | 0.73 | 0.80 | 0.77 | 138 |
| **Ανοιξιάτικες καλλιέργειες σιτηρών** | 0.92 | 0.93 | 0.93 | 1314 |
| **Χειμερινές καλλιέργειες σιτηρών** | 0.92 | 0.92 | 0.92 | 1192 |
| **Ελαιοκράμβη** | 0.93 | 0.89 | 0.91 | 157 |
| **Ακρίβεια** | | | 0.93 | 6505 |
| **Μέσος όρος** | 0.78 | 0.77 | 0.77 | 6505 |
| **Σταθμισμένος Μέσος Όρος** | 0.93 | 0.93 | 0.93 | 6505 |

Το μοντέλο που εκπαιδεύτηκε στις περιοχές A+B είχε αρκετά υψηλή απόδοση στην ταξινόμηση των περισσότερων κατηγοριών καλλιεργειών, όπως υποδεικνύεται από τις τιμές υψηλής ακρίβειας, ανάκλησης και F1-score. Οι κατηγορίες γρασίδι, ανοιξιάτικες καλλιέργειες σιτηρών, χειμερινές καλλιέργειες σιτηρών και η ελαιοκράμβη πέτυχαν υψηλές βαθμολογίες σε όλες τις μετρήσεις αξιολόγησης, υποδεικνύοντας ακριβείς προβλέψεις. Σχετικά με τις μόνιμες καλλιέργειες, η κατηγορία επέδειξε αρκετά χαμηλή τιμή F1-score 0.27, κάτι που υποδηλώνει πως ο Random Forest απέτυχε να ταξινομήσει τη συγκεκριμένη κατηγορία. Συνολικά, το μοντέλο επέδειξε υψηλή ακρίβεια 93%, υποδεικνύοντας την αποτελεσματικότητά του στην ταξινόμηση των καλλιεργειών στις περιοχές A+B.

#### *5.2.3.5. Μεταφορά Μοντέλου A+B στην Περιοχή Γ*

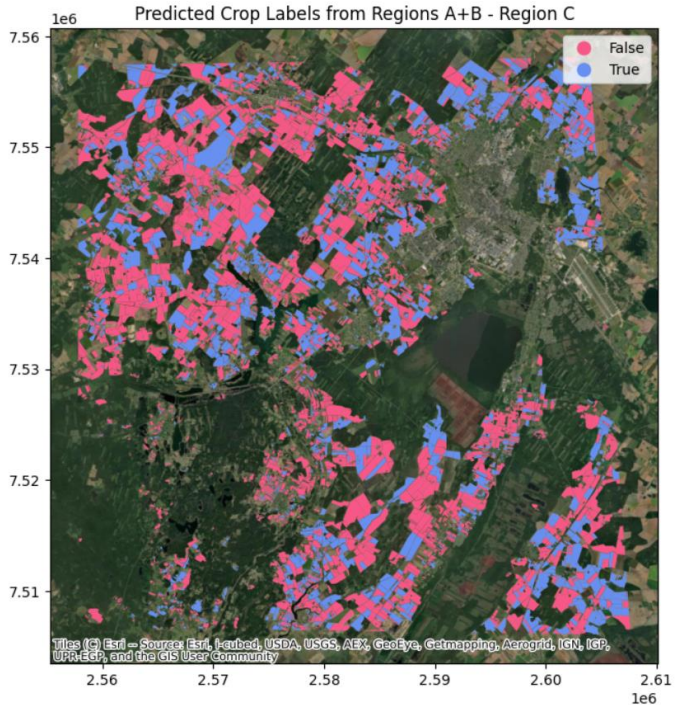
Αφού έγινε η εκπαίδευση των δεδομένων στις περιοχές A+B, το μοντέλο μεταφέρθηκε στην περιοχή Γ και παρακάτω παρουσιάζονται τα αποτελέσματα.

| Κατηγορία | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **Αγρανάπαυση** | 0.12 | 0.14 | 0.13 | 199 |
| **Γρασίδι** | 0.82 | 0.57 | 0.67 | 608 |
| **Μόνιμες Καλλιέργειες** | 0.12 | 0.01 | 0.02 | 83 |
| **Πατάτες** | 0.03 | 0.10 | 0.05 | 49 |
| **Ανοιξιάτικες καλλιέργειες σιτηρών** | 0.27 | 0.86 | 0.42 | 616 |
| **Χειμερινές καλλιέργειες σιτηρών** | 0.72 | 0.35 | 0.47 | 1566 |
| **Ελαιοκράμβη** | 0.44 | 0.17 | 0.25 | 652 |
| **Ακρίβεια** | | | 0.42 | 3773 |
| **Μέσος όρος** | 0.36 | 0.31 | 0.29 | 3773 |
| **Σταθμισμένος Μέσος Όρος** | 0.56 | 0.42 | 0.42 | 3773 |

Το μοντέλο που εκπαιδεύτηκε στις περιοχές A+B και εφαρμόστηκε στην περιοχή Γ εμφάνισε ποικίλες επιδόσεις σε διαφορετικές κατηγορίες καλλιεργειών. Το γρασίδι είχε σχετικά υψηλή

ακρίβεια 82%, αλλά χαμηλότερη ανάκληση 57%, υποδεικνύοντας ότι ενώ το μοντέλο ταξινόμησε σωστά ένα σημαντικό μέρος των περιπτώσεων, έχασε επίσης έναν σημαντικό αριθμό. Η κατηγορία ανοιξιάτικες καλλιέργειες σιτηρών πέτυχε υψηλότερη ανάκληση 86%, υποδηλώνοντας ότι το μοντέλο ήταν αποτελεσματικό στον εντοπισμό περιπτώσεων αυτής της κατηγορίας καλλιεργειών, αλλά η ακρίβεια ήταν αρκετά χαμηλή στο 27%. Η κατηγορία χειμερινές καλλιέργειες σιτηρών είχε αρκετά υψηλότερα ακρίβεια από την ανάκληση.

Ωστόσο, άλλες κατηγορίες όπως η αγρανάπαυση, οι μόνιμες καλλιέργειες, οι πατάτες και η ελαιοκράμβη είχαν χαμηλότερες τιμές ακρίβειας και ανάκλησης, υποδεικνύοντας πως το μοντέλο δεν κατάφερε να προσδιορίσει με ακρίβεια αυτές τις κατηγορίες στην περιοχή Γ. Συνολικά, η απόδοση του μοντέλου στην περιοχή Γ ήταν μέτρια, με ακρίβεια 42%,F1-score 29% για το μέσο όρο και 42% για τον σταθμισμένο μέσο όρο.




**Εικόνα 25 - Χάρτης ταξινόμησης περιοχής Γ**

### *5.2.3.6 Σύγκριση των δύο μοντέλων*

**Μοντέλο που εκπαιδεύτηκε στην Περιοχή Β και μεταφέρθηκε στην Περιοχή Γ:**

**Συνολική ακρίβεια**: 39%

**F1-score του μέσου όρου:** 18%

**F1-score του σταθμισμένου μέσου όρου**: 33%

Η απόδοση αυτού του μοντέλου είναι σχετικά χαμηλή, με χαμηλή ακρίβεια και F1-score στις περισσότερες κατηγορίες. Το μοντέλο απέτυχε να ταξινομήσει σωστά τις κατηγορίες, ειδικότερα την αγρανάπαυση, το γρασίδι, τις πατάτες και την ελαιοκράμβη, όπου ακρίβεια ή ανάκληση είναι σχεδόν μηδενικές.

Ωστόσο, στις ανοιξιάτικες καλλιέργειες σιτηρών πέτυχε αρκετά υψηλή τιμή ανάκλησης, 0.94, που σημαίνει πως μπορεί να ταξινομήσει το 96% των πραγματικών ανοιξιάτικων καλλιεργειών σιτηρών. Ταυτόχρονα, η μεταφορά του μοντέλου πέτυχε σχετικά ικανοποιητικές τιμές ακρίβειας κα ανάκλησης 0.59 και 0.56 αντίστοιχα για την κατηγορία χειμερινών καλλιεργειών σιτηρών.

**Μοντέλο που εκπαιδεύτηκε στις Περιοχές Α+Β και μεταφέρθηκε στην Περιοχή Γ:**

**Συνολική ακρίβεια**: 42%

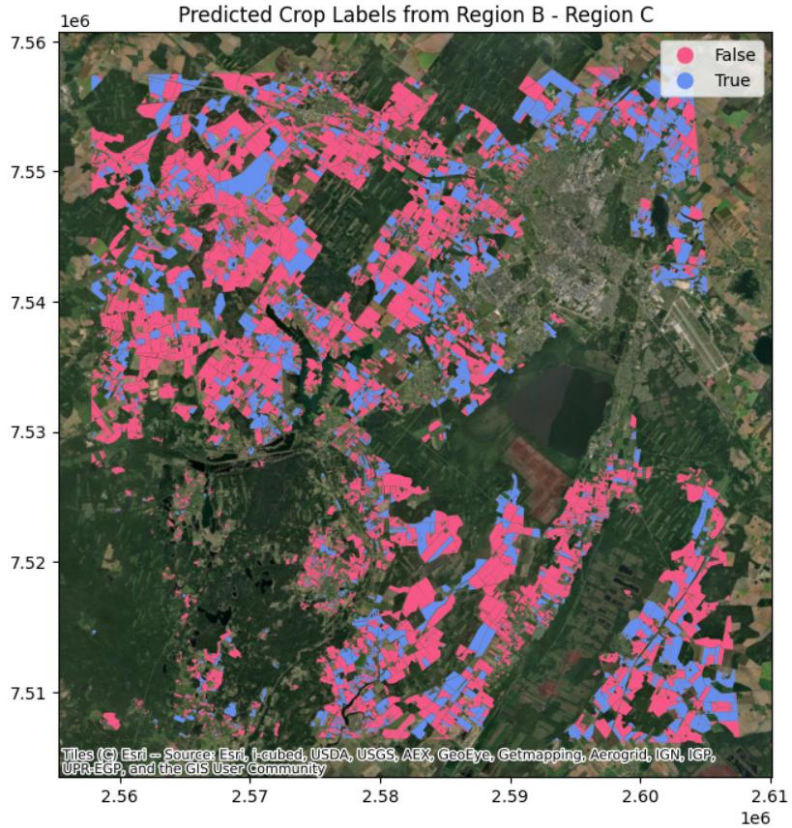
**F1-score του μέσου όρου**: 29%

**F1-score του σταθμισμένου μέσου όρου**: 42%

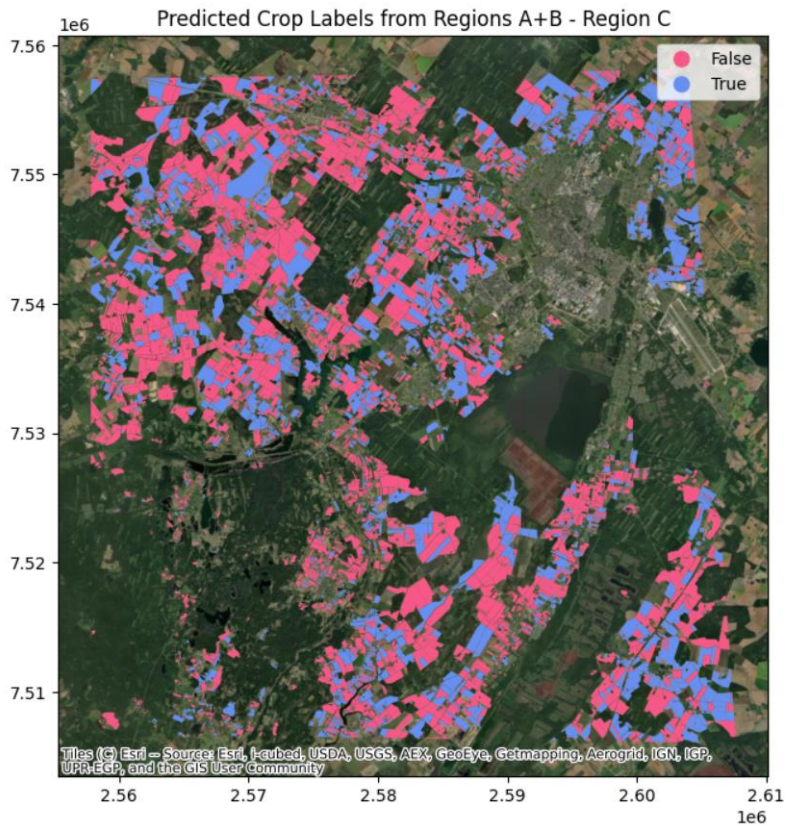
Το μοντέλο που εκπαιδεύτηκε και στις δύο περιοχές και μεταφέρθηκε στην Περιοχή Γ παρουσιάζει βελτίωση σε σύγκριση με το προηγούμενο μοντέλο. Η συνολική ακρίβεια, τα F1-score του μέσου και του σταθμισμένου όρου είναι υψηλότερες κατά 4%, 11% και 9% αντίστοιχα.

Με βάση τα παραπάνω αποτελέσματα φαίνεται πως η εισαγωγή περισσότερων δειγμάτων, στην προκειμένη ο συνδυασμός των δύο περιοχών, παρέχει βελτιωμένα αποτελέσματα. Ωστόσο, είναι σημαντικό να αναδειχθεί ότι η εισαγωγή περισσότερων δεδομένων οδηγεί σε μεγαλύτερη ανισορροπία σχετικά με την ακρίβεια των κατηγοριών με πολλά δείγματα και των κατηγοριών με αρκετά μικρότερο αριθμό δειγμάτων. Για αυτό το λόγο παρατηρήθηκαν και μικρότερα ποσοστά ανάκλησης στην περιοχή Γ από το μοντέλο Α+Β σε σχέση με το μοντέλο της περιοχής Β.

Παρακάτω φαίνονται οι δύο χάρτες που εμφανίζουν τις σωστές και τις λανθασμένες ταξινομήσεις για την Περιοχή Γ από το μοντέλο που εκπαιδεύτηκε στην περιοχή Β και στις περιοχές Α+Β αντίστοιχα.




Εικόνα 26 - Χάρτης μοντέλου που εκπαιδεύτηκε στην περιοχή Β και μεταφέρθηκε στην περιοχή Γ (Ακρίβεια 39%)

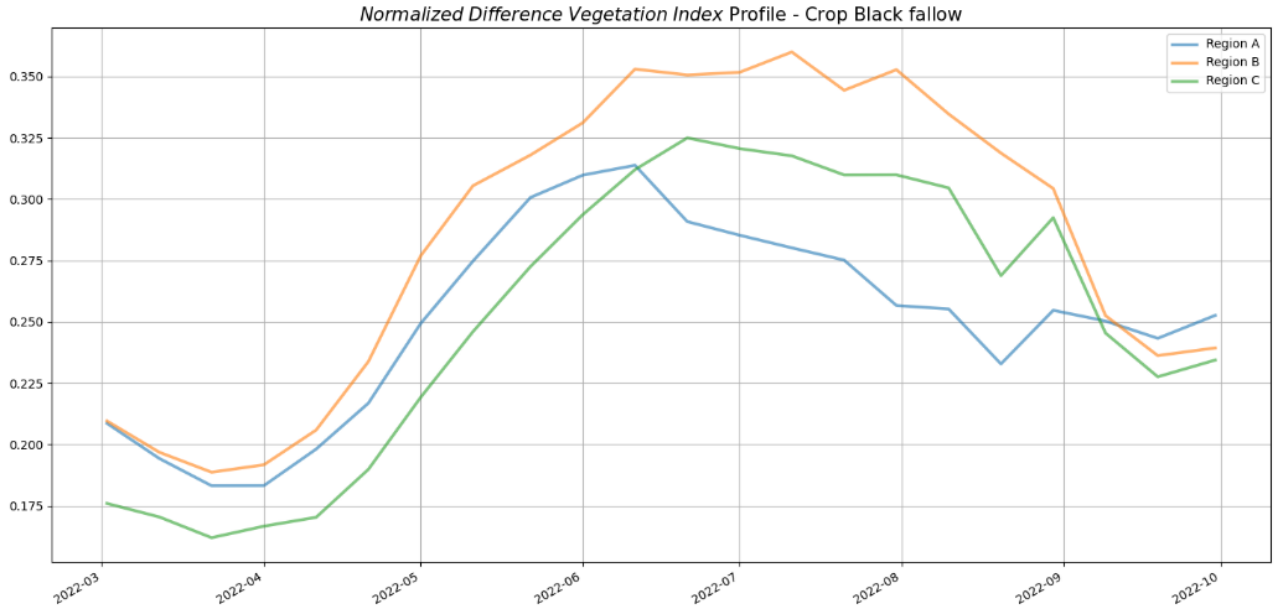


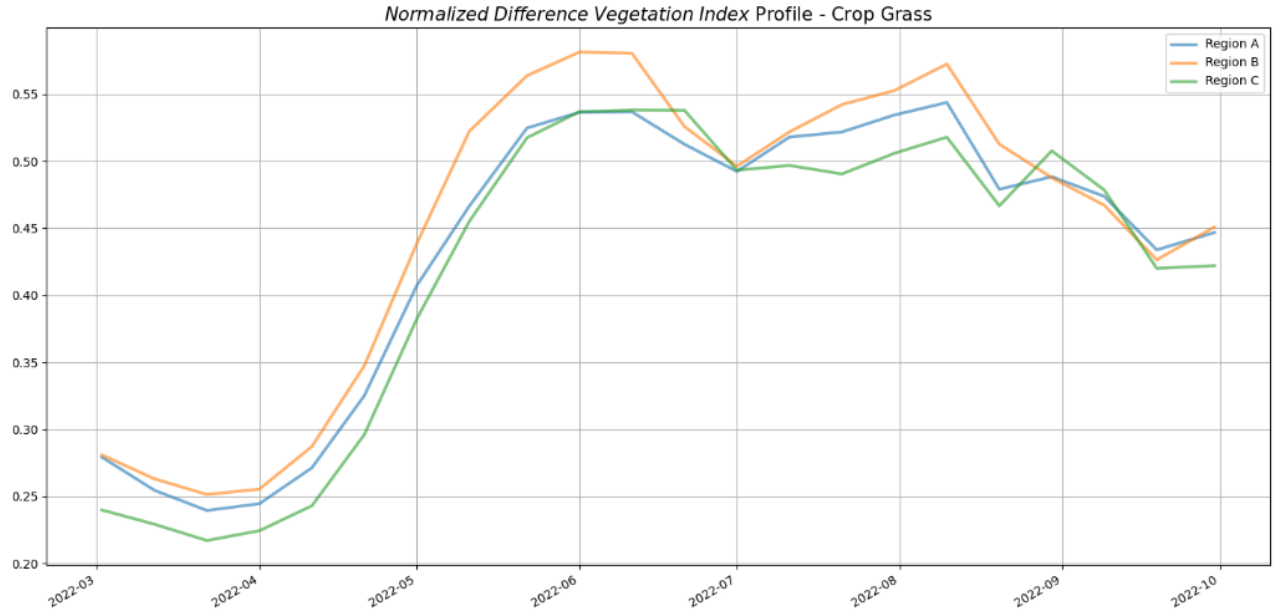

Εικόνα 45 - Χάρτης μοντέλου που εκπαιδεύτηκε στις περιοχές Α+Β και μεταφέρθηκε στην περιοχή Γ (Ακρίβεια 42%)

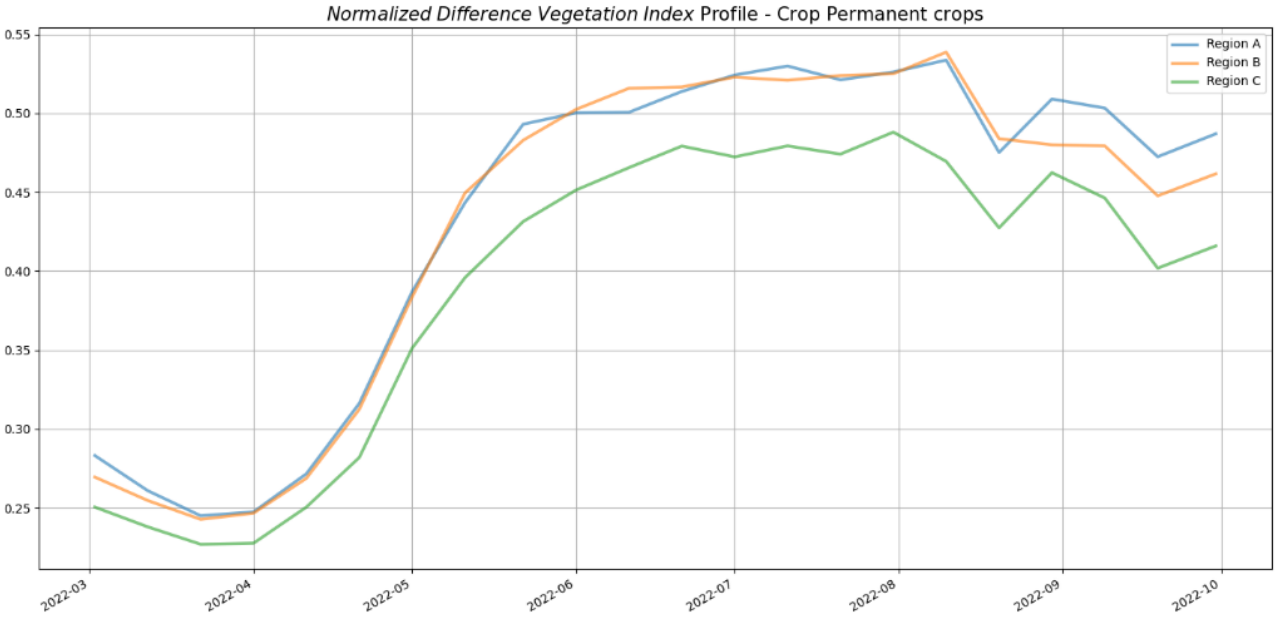
### *5.2.3.7 Μέθοδοι Oversampling και Undersampling στις περιοχές Α+Γ*

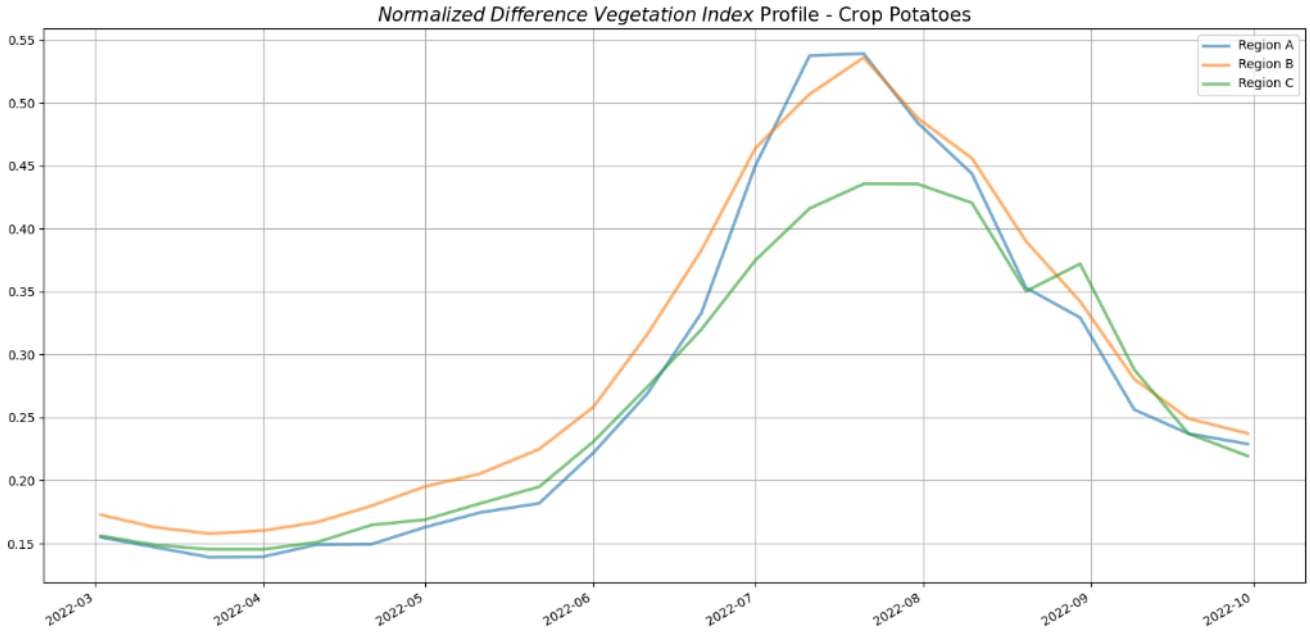
Με την προοπτική να γίνει μια περαιτέρω ανάλυση για να αξιοποιηθούν νέες τεχνικές και να βελτιωθούν τα αποτελέσματα, πραγματοποιήθηκαν οι τεχνικές του oversampling και του undersampling στα δείγματα.

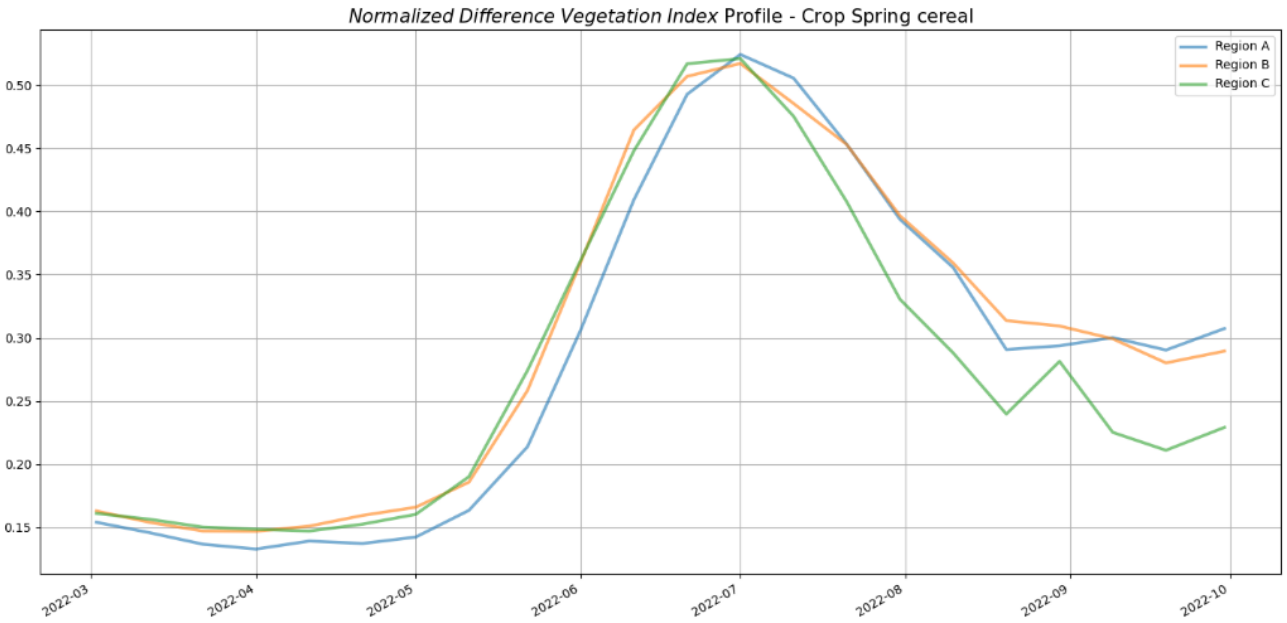
Αρχικά, πραγματοποιήθηκαν αναλύσεις των προφίλ του δείκτη NDVI για κάθε κατηγορία και στις τρεις περιοχές.

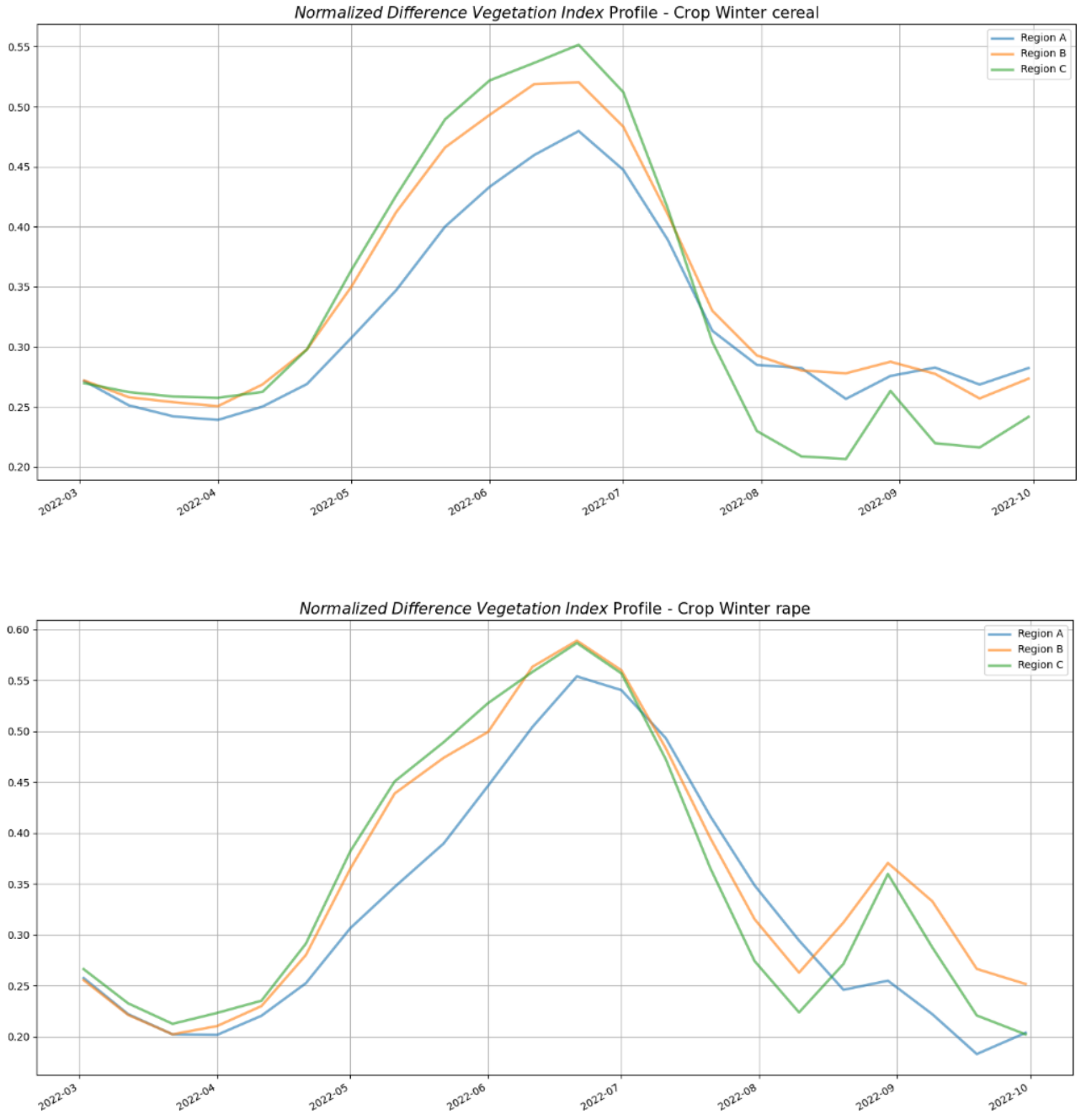



*Normalized Difference Vegetation Index* Profile - Crop Permanent crops
Region A
Region B
Region C
0.55
0.50
0.45
0.40
0.35
0.30
0.25
2022-03
2022-04
2022-05
2022-06
2022-07
2022-08
2022-09
2022-10


*Normalized Difference Vegetation Index* Profile - Crop Potatoes
Region A
Region B
Region C
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
2022-03
2022-04
2022-05
2022-06
2022-07
2022-08
2022-09
2022-10

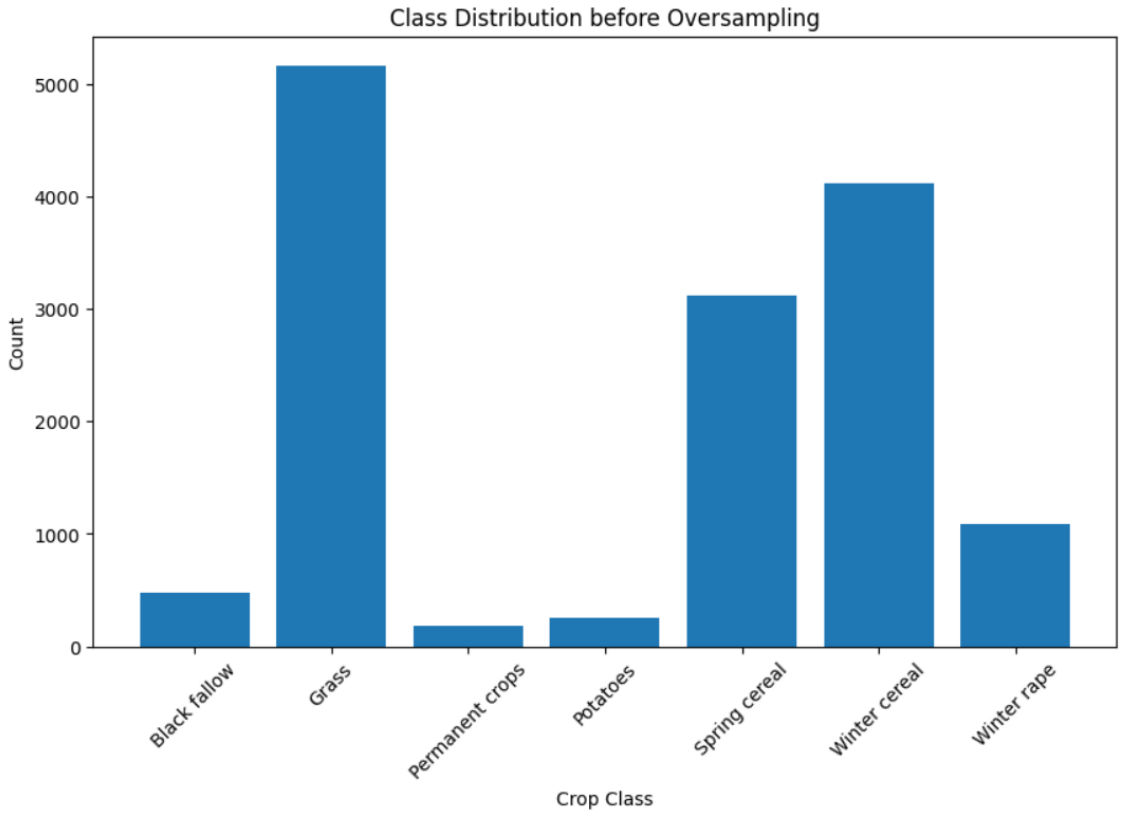

*Normalized Difference Vegetation Index* Profile - Crop Spring cereal
Region A
Region B
Region C
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
2022-03
2022-04
2022-05
2022-06
2022-07
2022-08
2022-09
2022-10




**Εικόνα 46 - Δείκτης NDVI για κάθε καλλιέργεια στις τρεις περιοχές**

Παρατηρώντας τις κατηγορίες, φαίνεται πως σε πολλές από αυτές υπάρχει μεγάλο εύρος στις τιμές του NDVI μεταξύ των περιοχών Α και Γ. Για αυτό το λόγο, επιλέχθηκε ο συνδυασμός των δύο περιοχών, με την προοπτική ότι υπάρχει ένα εύρος μεταξύ των τιμών NDVI στο οποίο βρίσκονται οι τιμές της περιοχής Β.

Τα δεδομένα των περιοχών Α+Γ αποτελούνται από 14134 αγροτεμάχια, δηλαδή το 52% του συνόλου των δεδομένων και των τριών περιοχών.

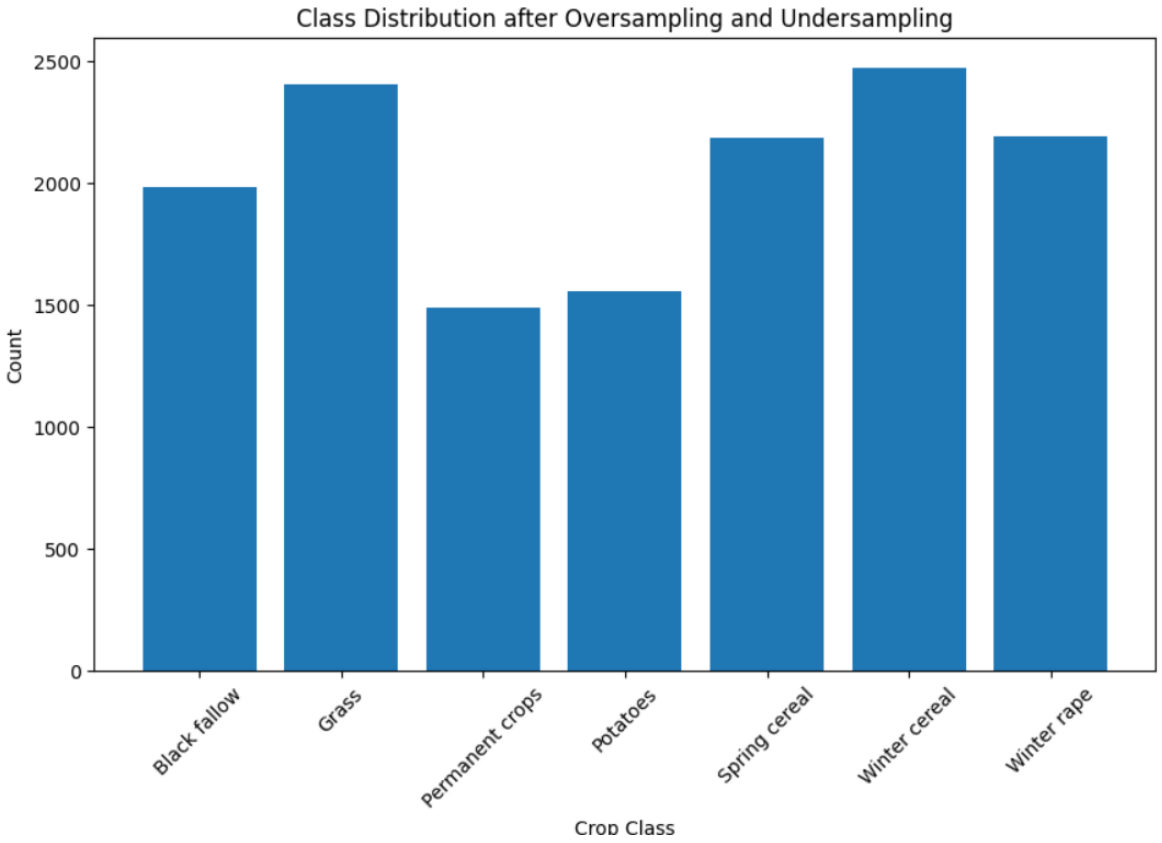



**Εικόνα 47 - Μέγεθος δειγμάτων ανά καλλιέργεια στις περιοχές Α και Γ**

Η παραπάνω εικόνα αναδεικνύει αυτό που είχε φανεί και από τα προηγούμενα παραδείγματα, πως υπάρχει μεγάλη ανισορροπία μεταξύ των κλάσεων. Για να λυθεί το πρόβλημα αυτό εξετάστηκαν μέθοδοι oversampling και undersampling, ώστε να παραχθεί ένα πιο ισορροπημένο αποτέλεσμα. Για τις κατηγορίες γρασίδι, χειμερινές και ανοιξιάτικες καλλιέργειες σιτηρών πραγματοποιήθηκε undersampling με την μέθοδο Random Undersampler και στις κατηγορίες αγρανάπαυση, μόνιμες καλλιέργειες και πατάτες πραγματοποιήθηκαν μέθοδοι oversampling με το μοντέλο SMOTE.

Για την κατηγορία αγρανάπαυση χρησιμοποιήθηκε ο συντελεστή 4Χ, για την κατηγορία μόνιμες καλλιέργειες χρησιμοποιήθηκε ο συντελεστής 8Χ, για την κατηγορία πατάτες αξιοποιήθηκε ο συντελεστής 6Χ και για την κατηγορία ελαιοκράμβη χρησιμοποιήθηκε ο συντελεστής 2Χ. Όσον αφορά τις χειμερινές καλλιέργειες σιτηρών, αξιοποιήθηκε ο συντελεστής undersampling 0.6Χ, για τις ανοιξιάτικες καλλιέργειες σιτηρών ο συντελεστής ήταν 0.6Χ, ενώ για το γρασίδι αξιοποιήθηκε ο συντελεστής 0.5Χ.

Σκοπός της τεχνικής είναι να δημιουργήσει ένα εύρος μεταξύ των μέγιστων και ελάχιστων τιμών για κάθε κατηγορία. Πριν τις τεχνικές Oversampling και Undersampling οι μόνιμες καλλιέργειες αποτελούνταν από 188 δείγματα και το γρασίδι από 5162 δείγματα, ενώ μετά τις τεχνικές, η κατηγορία με τα λιγότερα δείγματα είναι οι μόνιμες καλλιέργειες με 1488 δείγματα και η κατηγορία με τα περισσότερα δείγματα είναι το γρασίδι και οι χειμερινές καλλιέργειες σιτηρών με 2400 και 2471 δείγματα αντίστοιχα. Έγινε προσπάθεια να παραμείνουν οι τάσεις των κατηγοριών, αλλά να μειωθεί το εύρος διαφοράς των δειγμάτων τους.




**Εικόνα 48 - Μέγεθος δειγμάτων ανά καλλιέργεια ύστερα από τεχνικές oversampling και undersampling στις περιοχές Α και Γ**

**Εκπαίδευση στην Περιοχή Α+Γ**

Αφού μειώθηκε η ανισορροπία μεταξύ των καλλιεργειών, εκπαιδεύτηκε ένα νέο μοντέλο με τον Random Forest στα καινούρια δεδομένα των περιοχών Α+Γ. Τα αποτελέσματα του Random Forest φαίνονται παρακάτω.

Στη συγκεκριμένη ενότητα με τις τεχνικές oversampling και undersampling το ποσοστό εκπαίδευσης και ελέγχου ορίστηκε σε 60-40. Στα προηγούμενα παραδείγματα είχε χρησιμοποιηθεί το ποσοστό 70-30, αλλά λόγω του πιο ισορροπημένου συνόλου δεδομένων , αποφασίστηκε να διερευνηθεί και ένα νέο ποσοστό εκπαίδευσης/ελέγχου.

| Κατηγορία | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **Αγρανάπαυση** | 0.78 | 0.98 | 0.87 | 611 |
| **Γρασίδι** | 0.92 | 0.92 | 0.92 | 714 |
| **Μόνιμες Καλλιέργειες** | 0.92 | 0.85 | 0.88 | 423 |
| **Πατάτες** | 0.96 | 0.92 | 0.94 | 459 |
| **Ανοιξιάτικες καλλιέργειες σιτηρών** | 0.94 | 0.89 | 0.94 | 648 |
| **Χειμερινές καλλιέργειες σιτηρών** | 0.94 | 0.90 | 0.92 | 744 |
| **Ελαιοκράμβη** | 0.98 | 0.93 | 0.95 | 682 |
| **Ακρίβεια** | | | 0.92 | 4281 |
| **Μέσος όρος** | 0.92 | 0.91 | 0.91 | 4281 |
| **Σταθμισμένος Μέσος Όρος** | 0.92 | 0.91 | 0.92 | 4281 |

Όπως φαίνεται από τον παραπάνω πίνακα, η μέθοδος του oversampling δείχνει σημαντική διαφορά στις κλάσεις που στο πραγματικό σύνολο των δεδομένων έχουν μικρό αριθμό δειγμάτων, όπως η αγρανάπαυση, οι μόνιμες καλλιέργειες, οι πατάτες και η ελαιοκράμβη.

Σε σχέση με την απόδοση του Random Forest που είχε εκπαιδευτεί στο σύνολο των δεδομένων, υπάρχει 14% αύξηση στο F1-score του μέσου όρου, πράγμα που υποδηλώνει πως η ανισορροπία ήταν ένα ζήτημα που στοίχιζε σε ακρίβεια, ανάκληση και F1-score για αυτές τις κατηγορίες.

**Μεταφορά του μοντέλου Α+Γ στην περιοχή Β**

Αφού εκπαιδεύτηκε το μοντέλο του Random Forest στις περιοχές Α+Γ, επόμενο βήμα ήταν η μεταφορά του στην περιοχή Β. Η περιοχή Β αποτελείται από 12658 αγροτεμάχια, περίπου το 47% του συνόλου των δεδομένων των τριών περιοχών.

ΗΗ μεταφορά του μοντέλου στην περιοχή Β παρήγαγε τα παρακάτω αποτελέσματα.

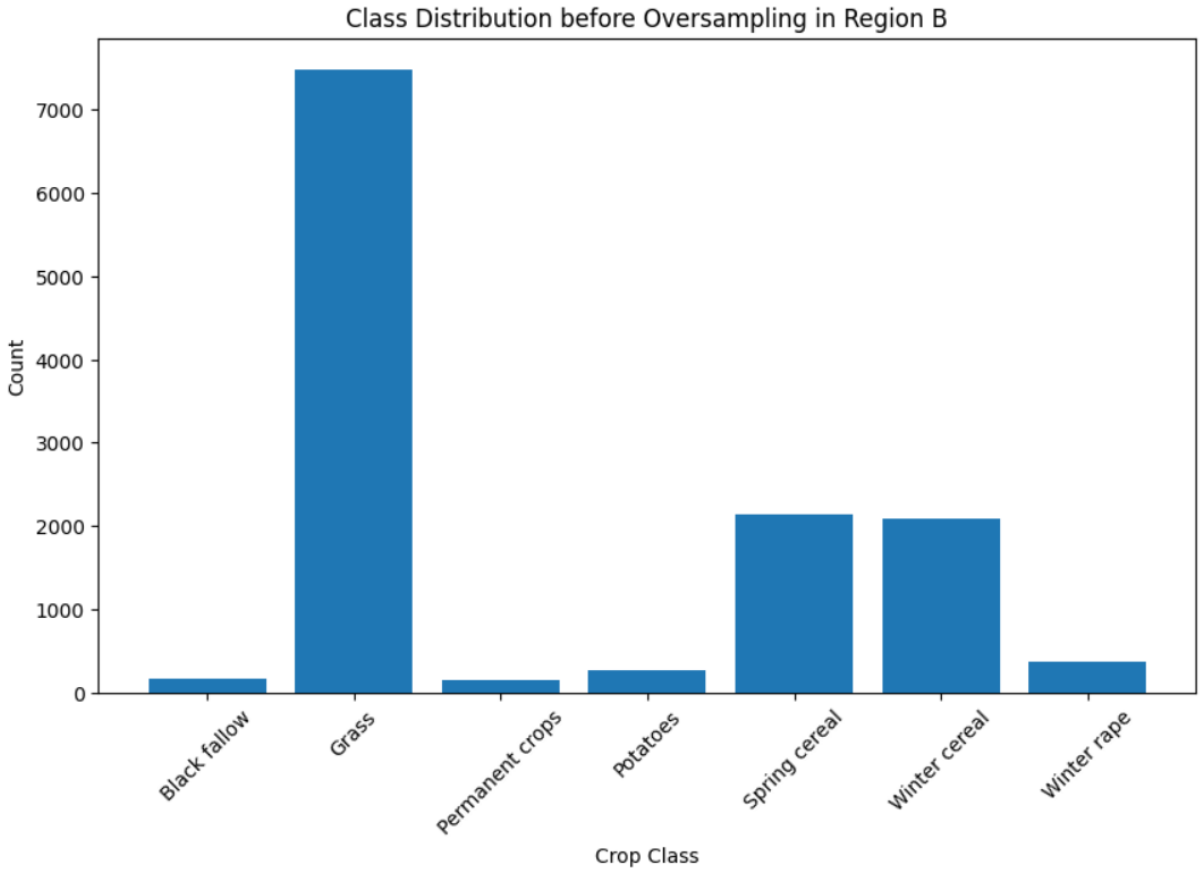
| Κατηγορία | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **Αγρανάπαυση** | 0.00 | 0.01 | 0.00 | 159 |
| **Γρασίδι** | 0.99 | 0.07 | 0.13 | 7482 |
| **Μόνιμες Καλλιέργειες** | 0.01 | 0.12 | 0.02 | 144 |
| **Πατάτες** | 0.23 | 0.36 | 0.28 | 276 |
| **Ανοιξιάτικες καλλιέργειες σιτηρών** | 0.37 | 0.85 | 0.51 | 2139 |
| **Χειμερινές καλλιέργειες σιτηρών** | 0.21 | 0.42 | 0.28 | 2092 |
| **Ελαιοκράμβη** | 0.00 | 0.00 | 0.00 | 366 |
| **Ακρίβεια** | | | 0.26 | 12658 |
| **Μέσος όρος** | 0.26 | 0.26 | 0.18 | 12658 |
| **Σταθμισμένος Μέσος Όρος** | 0.69 | 0.26 | 0.22 | 12658 |

Με βάση τον παραπάνω πίνακα γίνεται αντιληπτό, πως παρόλο που το μοντέλο έδειξε βελτιωμένα αποτελέσματα στις περιοχές Α+Γ, στην περιοχή Β δεν παρήγαγε καθόλου καλά αποτελέσματα. Αξίζει να σχολιαστεί πως στις καλλιέργειες αγρανάπαυσης και ελαιοκράμβης απέτυχε τελείως με F1-score 0 και στις κατηγορίες χειμερινές και ανοιξιάτικες καλλιέργειες σιτηρών, καθώς και στις πατάτες πέτυχε μία ανάκληση 0.42, 0.85 και 0.36 αντίστοιχα. Επίσης, η κατηγορία γρασίδι φαίνεται να πέτυχε Overfitting, καθώς η ακρίβεια είναι σχεδόν 1, ενώ η ανάκληση 7%, που σημαίνει πως το μοντέλο ταξινομεί σωστά το 99% της κατηγορίας αυτής, αλλά αναγνωρίζει μόνο το 7% των αγροτεμαχίων της συγκεκριμένης κατηγορίας.

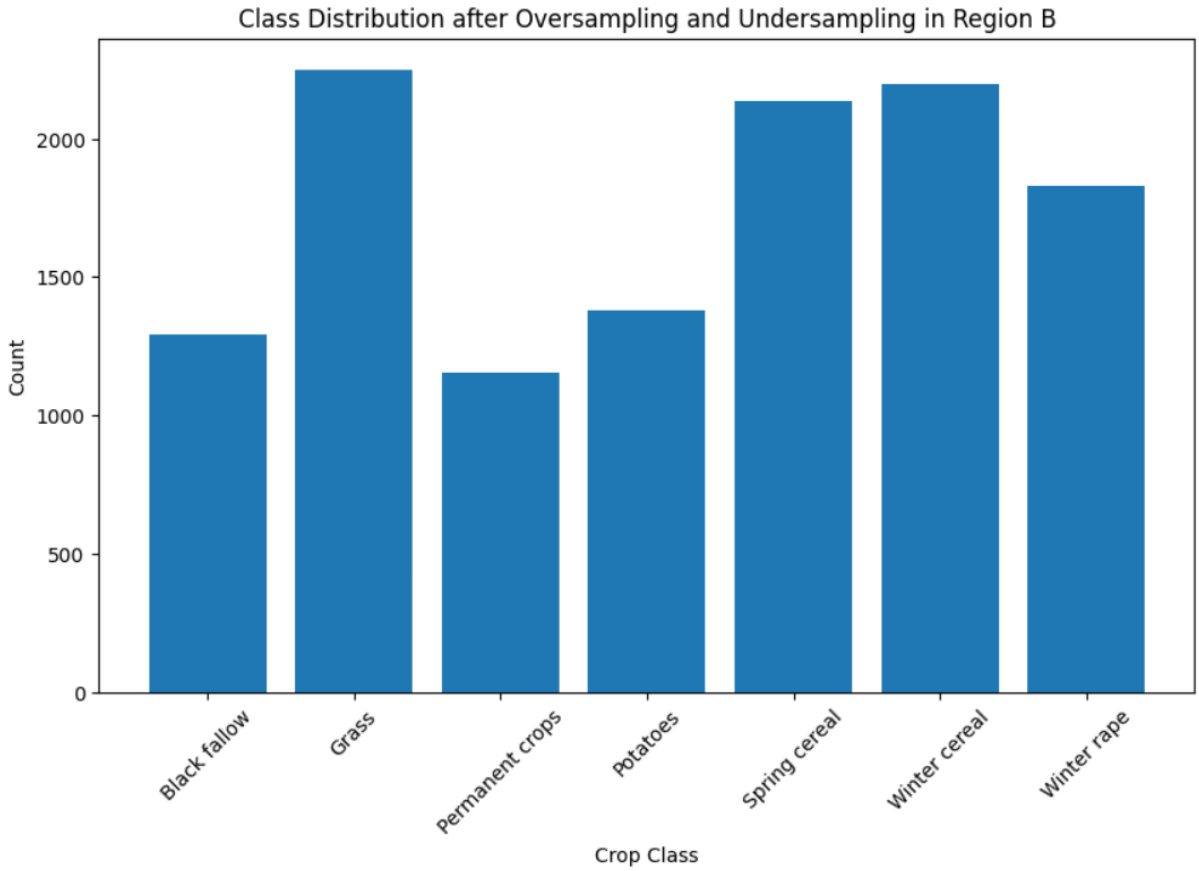
Το F1-score για τον μέσο όρο ήταν 0.18, ενώ για τον σταθμισμένο 0.22, υποδηλώνοντας αποτυχία του μοντέλου στην περιοχή Β.

Ωστόσο, παρατηρείται και στην Περιοχή Β η ανισορροπία ειδικά στην αγρανάπαυση που αποτελείται από 159 δείγματα, και οι μόνιμες καλλιέργειες που αποτελούνται από 144 δείγματα σε σχέση με το γρασίδι που αποτελείται από 7482 δείγματα.

Για αυτό το λόγο πραγματοποιήθηκαν μέθοδοι oversampling και undersampling στην περιοχή Β με τα παρακάτω αποτελέσματα.




**Εικόνα 49 - Μέγεθος δειγμάτων ανά καλλιέργεια στην περιοχή B**




**Εικόνα 50 - Μέγεθος δειγμάτων ανά καλλιέργεια ύστερα από τεχνικές oversampling και undersampling στην περιοχή B**

Τα νέα δεδομένα για την περιοχή Β είναι 12242, ενώ πριν ήταν 12658, σχεδόν ίδια σε αριθμό. Ωστόσο, γίνεται αντιληπτό πως με μεθόδους oversampling και undersampling αναδιαμορφώθηκαν τα δεδομένα και είναι αρκετά πιο ισορροπημένα.

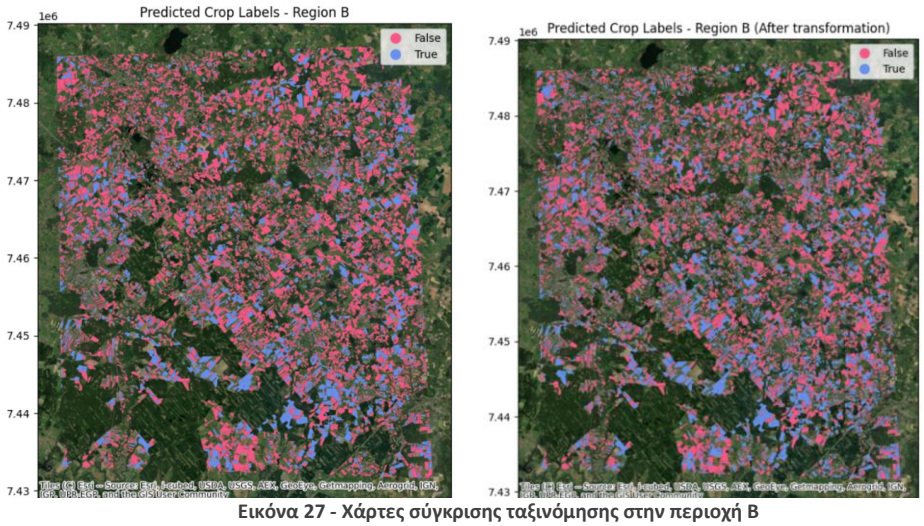
Στη συνέχεια πραγματοποιήθηκε ξανά η μεταφορά του μοντέλου Α+Γ στην περιοχή Β, στην οποία πραγματοποιήθηκαν οι τεχνικές για την ισορροπία των δειγμάτων ανά κατηγορία.

| Κατηγορία | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **Αγρανάπαυση** | 0.02 | 0.01 | 0.02 | 1292 |
| **Γρασίδι** | 0.86 | 0.14 | 0.25 | 2250 |
| **Μόνιμες Καλλιέργειες** | 0.12 | 0.14 | 0.13 | 1152 |
| **Πατάτες** | 0.67 | 0.57 | 0.62 | 1380 |
| **Ανοιξιάτικες καλλιέργειες σιτηρών** | 0.41 | 0.75 | 0.53 | 2138 |
| **Χειμερινές καλλιέργειες σιτηρών** | 0.31 | 0.61 | 0.41 | 2200 |
| **Ελαιοκράμβη** | 0.60 | 0.04 | 0.07 | 1830 |
| **Ακρίβεια** | | | 0.35 | 12242 |
| **Μέσος όρος** | 0.43 | 0.32 | 0.29 | 12242 |
| **Σταθμισμένος Μέσος Όρος** | 0.46 | 0.35 | 0.31 | 12658 |

Είναι απολύτως κατανοητό πως μετά την αναδιαμόρφωση των δεδομένων με τις μεθόδους oversampling και undersampling, το F1-score σε μέσο και σταθμισμένο μέσο όρο αυξήθηκε κατά 9%. Σημαντικές βελτιώσεις σε σχέση με την εφαρμογή του μοντέλου των περιοχών Α+Γ που εφαρμόστηκε στην περιοχή Β χωρίς τις τεχνικές αυτές είναι η αύξηση του F1-score όλων των κατηγοριών από 2% μέχρι και 34%. Την μεγαλύτερη αύξηση από τις τεχνικές oversampling και undersampling την πέτυχε η κατηγορία πατάτες όπου ακρίβεια και ανάκληση αυξήθηκαν 44% και 21% αντίστοιχα.

Ωστόσο, η ταξινόμηση της καλλιέργειας αγρανάπαυσης απέτυχε με σχεδόν μηδενική ακρίβεια και ανάκληση, ενώ η ελαιοκράμβη πέτυχε 60% ακρίβεια, αλλά πολύ χαμηλό ποσοστό ανάκλησης 4%.

Παρακάτω φαίνονται οι χάρτες ταξινόμησης της περιοχής Β, πριν και ύστερα από τεχνικές oversampling και undersampling.




**Εικόνα 27 - Χάρτες σύγκρισης ταξινόμησης στην περιοχή Β**

# Κεφάλαιο 6 – Συμπεράσματα

Η συγκεκριμένη εργασία είχε ως σκοπό να αξιολογήσει τη χρήση αλγορίθμων επιβλεπόμενης και μη-επιβλεπόμενης ταξινομήσης καλλιεργειών χρησιμοποιώντας χρονοσειρές δορυφορικών δεδομένων Sentinel-1 και Sentinel-2. Μέσω της ανάλυσης και των πειραμάτων, αποκτήθηκαν πολύτιμες γνώσεις σχετικά με την απόδοση και τις δυνατότητες διαφορετικών αλγορίθμων για την ακριβή ταξινόμηση διάφορων τύπων καλλιεργειών.

Συγκεκριμένα, για την μη-επιβλεπόμενη ταξινόμηση τις υψηλότερες βαθμολογίες silhouette για τις πραγματικές κλάσεις τις πέτυχε ο t-SNE στον δείκτη NDVI και RVI. Ο t-SNE αναδείχθηκε σε έναν αρκετά καλό αλγόριθμο για την μείωση διαστάσεων. Τις υψηλότερες βαθμολογίες silhouette τις πέτυχε ο k-means και ο GMM στον NDVI PCA. Ωστόσο, τις πιο σταθερές τιμές στις βαθμολογίες silhouette τις πέτυχε ο t-SNE σε NDVI και RVI.

Όσον αφορά την επιβλεπόμενη ταξινόμηση, όλοι οι αλγόριθμοι έδωσαν ικανοποιητικά αποτελέσματα, αλλά ο Random Forest ξεπέρασε για λίγο τους άλλους δύο αλγόριθμους.

Σχετικά με την εξαγωγή σημαντικών χαρακτηριστικών και την επανεκπαίδευση του μοντέλου, επισημάνθηκε πως παρείχε σχεδόν ίδια αποτελέσματα με το αρχικό μοντέλο. Ωστόσο, επειδή από τα 22 χαρακτηριστικά κρατήθηκαν τα 3, χάθηκαν φασματικές πληροφορίες και αυτό είχε ως αποτέλεσμα χαμηλότερη ακρίβεια σε κατηγορίες καλλιεργειών που δεν είχαν μεγάλο δείγμα.

Η γενίκευση των μοντέλων και η μεταφορά τους σε άλλες περιοχές δεν παρείχε ικανοποιητικά αποτελέσματα. Για τη βελτίωση των αποτελεσμάτων αξιοποιήθηκαν δύο τεχνικές.

Η πρώτη τεχνική ήταν η ενσωμάτωση περισσότερων δειγμάτων και η μεταφορά σε μία περιοχή. Αρχικά, έγινε εκπαίδευση στην περιοχή Β και μεταφορά του μοντέλου στις περιοχές Α και Γ. Στη συνέχεια έγινε εκπαίδευση του μοντέλου στις περιοχές Α+Β και η μεταφορά του στην περιοχή Γ. Η τεχνική αυτή είχε ως αποτέλεσμα την αύξηση της ακρίβειας κατά 3%.

Η δεύτερη τεχνική εστίασε σε μεθόδους oversampling και undersampling, ώστε να βρεθεί λύση στο πρόβλημα των μη-ισορροπημένων δεδομένων. Αρχικά, πραγματοποιήθηκαν μέθοδοι oversampling και undersampling στις περιοχές Α+Γ, όπου εκπαιδεύτηκε το μοντέλο και μεταφέρθηκε στην περιοχή Β, πετυχαίνοντας ακρίβεια 26%. Στη συνέχεια, εφαρμόστηκαν μέθοδοι oversampling και undersampling στην περιοχή Β και έγινε εκ νέου μεταφορά του μοντέλου στην περιοχή, πετυχαίνοντας 35% ακρίβεια. Η τεχνική αυτή οδήγησε σε αύξηση της ακρίβειας 9%.

Μελλοντικά, κρίνεται ενδιαφέρον να αξιολογηθούν και να ερευνηθούν:

- τεχνικές που αφορούν το oversampling και undersampling, καθώς το πρόβλημα των μη-ισορροπημένων δεδομένων είναι αρκετά σύνηθες και σημαντικό
- τεχνικές ημι-επιβλεπόμενης μάθησης, όπου γίνεται συνδυασμός μη-επιβλεπόμενων και επιβλεπόμενων μαθήσεων. Χαρακτηριστικό παράδειγμα αποτελεί το pseudolabeling, από το οποίο μπορούν να δημιουργηθούν ετικέτες για δείγματα, γνωρίζοντας μόνο ένα μικρό δείγμα δειγμάτων.
- νευρωνικά δίκτυα για τις γενικεύσεις των μοντέλων κα τη μεταφορά τους σε άλλες περιοχές
- μέθοδοι συνδυασμού μη-επιβλεπόμενων αλγορίθμων (k-means) και στατιστικών μεθόδων

# Κεφάλαιο 7 – Βιβλιογραφία


[1]: Navin, M. Sam & Loganathan, Agilandeeswari. (2019). Land use Land Cover Change Detection using K-means Clustering and Maximum Likelihood Classification Method in the Javadi Hills, Tamil Nadu, India. 10.35940/ijeat.A1011.1291S319.

[2]: Nitze, Ingmar & Schulthess, Urs & Asche, H. (2012). Comparison of machine learning algorithms Random Forest, Artificial Neural Network and Support Vector Machine to Maximum Likelihood for supervised crop type classification

[3]: Felegari, S.; Sharifi, A.; Moravej, K.; Amin, M.; Golchin, A.; Muzirafuti, A.; Tariq, A.; Zhao, N. Integration of Sentinel 1 and Sentinel 2 Satellite Images for Crop Mapping. *Appl. Sci.* **2021**, *11*, 10104. https://doi.org/10.3390/app112110104

[4]: Mariana Belgiu, Lucian Drăguţ, Random forest in remote sensing: A review of applications and future directions, ISPRS Journal of Photogrammetry and Remote Sensing, Volume 114, 2016. https://doi.org/10.1016/j.isprsjprs.2016.01.011.

[5] Zhang, H.; Gao, M.; Ren, C. Feature-Ensemble-Based Crop Mapping for Multi-Temporal Sentinel-2 Data Using Oversampling Algorithms and Gray Wolf Optimizer Support Vector Machine. Remote Sens. 2022, 14, 5259. https://doi.org/10.3390/rs14205259

[6] T. Drivas, V. Sitokonstantinou, I. Tsardanidis, A. Koukos, C. Kontoes and V. Karathanassi, "A Data Cube of Big Satellite Image Time-Series for Agriculture Monitoring," 2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), Nafplio, Greece, 2022, pp. 1-5, doi: 10.1109/IVMSP54334.2022.9816291.

[7] Sitokonstantinou, V.; Koukos, A.; Drivas, T.; Kontoes, C.; Papoutsis, I.; Karathanassi, V. A Scalable Machine Learning Pipeline for Paddy Rice Classification Using Multi-Temporal Sentinel Data. Remote Sens. 2021, 13, 1769. https://doi.org/10.3390/rs13091769

[8] MacQueen J. B. Some methods for classification and analysis of multivariate observations Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability 1967 University of California Press 281-297

[9] L.J.P. van der Maaten and G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9(Nov):2579-2605, 2008

[10]Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[11] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics, 1189–1232.

[12] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.

[11] Karl Pearson F.R.S. . (1901). LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11), 559–572. https://doi.org/10.1080/14786440109462720

[13] Rei Sonobe, Yuki Yamaya, Hiroshi Tani, Xiufeng Wang, Nobuyuki Kobayashi, and Kan-ichiro Mochizuki "Crop classification from Sentinel-2-derived vegetation indices using

ensemble learning," Journal of Applied Remote Sensing 12(2), 026019 (18 May 2018). https://doi.org/10.1117/1.JRS.12.026019

[14] Nobuyuki Kobayashi, Hiroshi Tani, Xiufeng Wang & Rei Sonobe (2020) Crop classification using spectral indices derived from Sentinel-2A imagery, Journal of Information and Telecommunication, 4:1, 67-90, DOI: 10.1080/24751839.2019.1694765