ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΤΜΗΜΑ ΜΗΧΑΝΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΑΚΑΔΗΜΑΪΚΟ ΕΤΟΣ 2015-2016

Τεχνικές Εξόρυξης Δεδομένων

Μελέτη Εφαρμογής της Εξόρυξης Δεδομένων στον Αθλητισμό με Χρήση του Λογισμικού Weka

---

Σωφρονάς Ηλίας

Οκτώβριος, 2015

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον Καθηγητή μου κ. Πόνη Σταύρο για την άψογη συνεργασία που είχαμε καθ' όλη τη διάρκεια εκπόνησης της διπλωματικής εργασίας, καθώς επίσης και την οικογένειά μου για τη στήριξή τους όλα αυτά τα χρόνια.

ΠΕΡΙΕΧΟΜΕΝΑ

Περίληψη

Η εξόρυξη δεδομένων είναι η ανάλυση μεγάλων συνόλων δεδομένων με σκοπό να βρούμε σχέσεις που δεν υποψιαζόμαστε και να συνοψίσουμε τα δεδομένα με καινοτόμους τρόπους που είναι κατανοητοί και χρήσιμοι στον κάτοχο των δεδομένων. Οι σχέσεις και οι συνοψίσεις που παράγονται από μία διαδικασία data mining αναφέρονται ως μοντέλα ή πρότυπα. Το πρώτο βήμα μιας διαδικασίας data mining είναι η προεπεξεργασία των δεδομένων και πιθανά η μείωση των διαστάσεων της βάσης δεδομένων. Αυτό μπορεί να γίνει με τον καθαρισμό των δεδομένων, την εφαρμογή στατιστικών μεθόδων, αλλά και άλλων τεχνικών που παρέχουν λογισμικά όπως το WEKA.

Επιπλέον, η αξιολόγηση των μεθόδων μάθησης είναι πολύ σημαντική στην πράξη, καθώς καθοδηγεί τη μέθοδο μάθησης ή το μοντέλο και δίνει ένα μέτρο της ποιότητας του επιλεγμένου μοντέλου. Τέλος, το data mining αποτελεί μία πολύ χρήσιμη εφαρμογή για οργανισμούς και επιχειρήσεις, καθώς βοηθά στην ανάλυση των δεδομένων με ευέλικτα λογισμικά όπως το Weka, και τη λήψη αποφάσεων. Πρόσφατα αναπτύχθηκαν και νέα πεδία του data mining,όπως το text mining, το web mining και το sports mining.

Η παρούσα μελέτη ασχολείται με το sports data mining, δηλαδή τη διαδικασία εξόρυξης δεδομένων στα σπορ. Στην παρούσα εργασία, αρχικά παρουσιάζονται βασικές έννοιες, τα διάφορα στάδια, οι εφαρμογές, οι κατηγορίες, και οι διάφορες τεχνικές μοντελοποίησης του data mining, και γίνεται μία εισαγωγή στο υπολογιστικό πακέτο Weka που χρησιμοποιείται για την εκτέλεση των μεθόδων και των αλγορίθμων του data mining. Στη συνέχεια, παρουσιάζονται αναλυτικά διάφορα είδη μάθησης, όπως αυτό της επιλογής χαρακτηριστικών, της ταξινόμησης και της συσταδοποίησης, και υιοθετούνται γνωστά μέτρα αξιολόγησης καθώς και η τεχνική της ανάλυσης καμπυλών λειτουργικού χαρακτηριστικού δέκτη, τα οποία εφαρμόζονται για την εκτίμηση των αποτελεσμάτων και την αξιολόγηση της απόδοσης των αλγορίθμων.

Abstract

Data mining is the analysis of large data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. The relationships and summaries derived through a data mining exercise are often referred to as models or patterns. The first step in a data mining procedure is the preprocessing of the data and probably the dimension reduction of the database. This can be done through data cleaning and the application of statistical methods, as well as through other techniques which WEKA provides.

Moreover, assessment of a learning method is extremely important in practice, since it guides the choice of a learning method, and gives us a measure of the quality of the ultimately chosen model. In conclusion, data mining provides a very useful application for organizations and businesses as it helps in analyzing data via flexible software, such as Weka, and in making decisions. Recently, new fields of data mining, such as text mining, web mining and sports mining have been developed.

This study deals with sports data mining, namely data mining process in sports. In this paper, we initially present the basic concepts, various steps, applications, categories, and various modelling techniques of data mining, and an introduction to computational Weka package, used to perform the methods and algorithms of data mining. Then, we present in detail the kinds of learning, such as feature selection, classification and clustering, and we adopt well known evaluation measures and the technique of receiver operating characteristic curves analysis, which are all applied to assess results and evaluate the performance of the applied algorithms.

1. Εισαγωγή

1.1 Εισαγωγή στην εξόρυξη δεδομένων

Από τα πρώτα χρόνια ύπαρξής του, ο άνθρωπος ζούσε παρατηρώντας τον κόσμο γύρω του και συλλέγοντας συνεχώς πληροφορίες οι οποίες τον βοηθούσαν στην επιβίωση αλλά και την εξέλιξή του. Σήμερα, η ραγδαία ανάπτυξη της τεχνολογίας σε συνδυασμό με τη σύγκλιση της χρήσης ηλεκτρονικών υπολογιστών και επικοινωνιών έχει δημιουργήσει μια κοινωνία που τρέφεται από πληροφορίες. Οι περισσότερες από αυτές τις πληροφορίες βρίσκονται στην ακατέργαστη μορφή τους: τα δεδομένα. Αν μπορούμε να χαρακτηρίσουμε τα δεδομένα ως καταγεγραμμένα γεγονότα, τότε οι πληροφορίες είναι μια σειρά από πρότυπα ή προσδοκίες που αποτελούν τη βάση των δεδομένων. Υπάρχει μια τεράστια ποσότητα πληροφοριών «κλειδωμένη» στις βάσεις δεδομένων, πληροφορίες δυνητικά σημαντικές που ακόμη δεν έχουν ανακαλυφθεί. Αποστολή μας είναι να τις βγάλουμε προς τα έξω.

Η κινητήριος δύναμη πίσω από την ανάπτυξη αρχικά των στατιστικών τεχνικών, και μετέπειτα των αλγορίθμων της εξόρυξης δεδομένων είναι τα προβλήματα ανάλυσης δεδομένων που χρειάζεται να επιλυθούν. Οι χώροι από τους οποίους αναδύονται τέτοιου είδους προβλήματα είναι τόσο ποικίλοι, όσες και οι προσεγγίσεις των λύσεων. Τα προβλήματα που διευθετούνται διαμέσου των τεχνικών της εξόρυξης δεδομένων μπορεί να έχουν πλατιά επίδραση και να επηρεάζουν την καθημερινή ζωή πολλών ανθρώπων, όπως για παράδειγμα στην περίπτωση της σωστής αναγνώρισης σε μια δόλια συναλλαγή μίας πιστωτικής κάρτας ή στην επιστροφή σχετικών αποτελεσμάτων σε μια μηχανή αναζήτησης. Αυτά είναι μερικά από τα παραδείγματα όπου η ανάλυση μεγάλεων βάσεων δεδεομένων γίνεται ακέραιο κομμάτι της καθημερινής μας ζωής.

Η ιδέα στην οποία στηρίζεται η εξόρυξη δεδομένων είναι η κατασκευή υπολογιστικών προγραμμάτων που χρησιμοποιούν στατιστικά αποτελέσματα για το κρησάρισμα βάσεων δεδομένων με στόχο την εξαγωγή προτύπων και άλλων πληροφοριών.

1.2 Ορισμός και βασικές αρχές του data mining

Ο όρος Data Mining (Εξόρυξη Δεδομένων), περιγράφει την αυτόματη ή ημιαυτόματη διαδικασία, μέσω χρήσης ηλεκτρονικού υπολογιστή, εύρεσης ανωμαλιών, προτύπων και συσχετίσεων από μεγάλες βάσεις δεδομένων. Με άλλα λόγια είναι η εξαγωγή «υπονοούμενων», προηγουμένως άγνωστων και ενδεχομένως χρήσιμων πληροφοριών από τα δεδομένα.

Σύμφωνα με τον ορισμό του Fayaad η ανακάλυψη γνώσης από βάσεις δεδομένων είναι μια ντετερμινιστική διαδικασία αναγνώρισης έγκυρων, καινοτόμων, ενδεχομένως χρήσιμων και τελικά κατανοητών προτύπων στα δεδομένα (Fayyad, 1996).

Εξόρυξη Δεδομένων καλείται η εξεύρεση (σημαντικών, αυτονόητων, άγνωστων και πιθανόν χρήσιμων) πληροφοριών ή επαναλαμβανόμενων προτύπων (patterns) σε τεράστιες βάσεις δεδομένων.

Ένας συνοπτικός ορισμός που περιέχει την ουσία όμως του Data Mining είναι ο ακόλουθος : Εξόρυξη χρήσιμων πληροφοριών από μεγάλα σύνολα δεδομένων (Hand et al. 2001).

Παρόλο που είναι δύσκολο να καθοριστούν με ακρίβεια το εύρος και τα όρια μελέτης αυτού του κλάδου, παραβλέπουμε τις λεπτομέρειες και αποδεχόμαστε ως ορισμό του Data Mining τον παρακάτω:

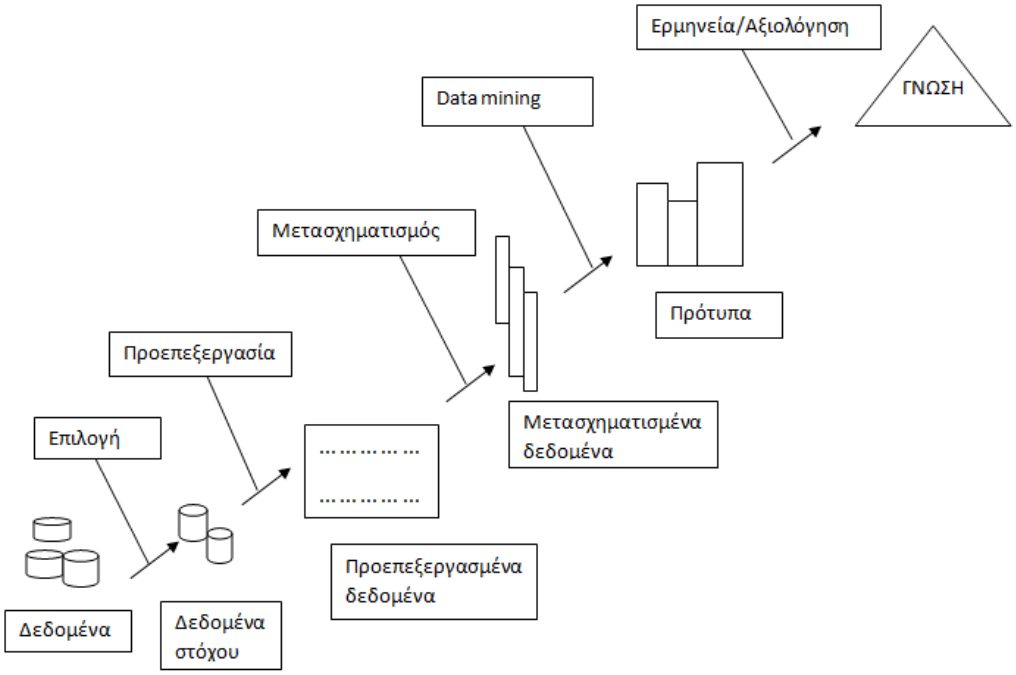
Data Mining είναι η ανάλυση συχνά μεγάλων παρατηρούμενων συνόλων δεδομένων με σκοπό να βρούμε σχέσεις που δεν υποψιαζόμαστε, και να συνοψίσουμε τα δεδομένα με καινοτόμους τρόπους, κατανοητούς και χρήσιμους για τον κάτοχο των δεδομένων.

Εικόνα 1: Πληροφορίες πίσω από τα δεδομένα

Γιατί όμως είναι σημαντικό το Data Mining; Ζώντας στον αιώνα της τεχνολογίας, αναπόφευκτα κατακλυζόμαστε από πλήθος δεδομένων τα οποία χρόνο με το χρόνο συνεχώς αυξάνονται. Τέτοιου είδους μη δομημένες πληροφορίες αποτελούν το 90% του ψηφιακού κόσμου. Περισσότερες όμως πληροφορίες δε σημαίνει απαραίτητα και περισσότερη γνώση. Το Data Mining μας επιτρέπει να κοσκινίσουμε όλον αυτό το χαοτικό και επαναλαμβανόμενο «θόρυβο», να καταλάβουμε τι είναι σχετικό και τότε να χρησιμοποιήσουμε κατάλληλα τα δεδομένα με σκοπό να αξιολογήσουμε σωστά μελλοντικά αποτελέσματα. Το Data Mining, ως μια σύνθετη αρχή, αντιπροσωπεύει μια ποικιλία μεθόδων και τεχνικών, οι οποίες με τη χρήση κατάλληλων κανόνων

επεξεργάζονται τον όγκο των διαθέσιμων δεδομένων με στόχο τη βέλτιστη λήψη αποφάσεων. Η εύρεση ισχυρών προτύπων, αν υπάρχουν, είναι ένα πολύ χρήσιμο εργαλείο για την ακριβή πρόβλεψη μελλοντικών δεδομένων, για τη γενίκευση από ένα δείγμα του συνόλου στο πλήρες σύνολο καθώς και για τη συμπίεση μεγάλων δεδομένων σε μικρότερα με σκοπό να γίνουν πιο εύχρηστα και κατανοητά.

Ο όρος Data Mining δεν ήταν ευρέως διαδεδομένος μέχρι τη δεκετία του 1980 και είχε εκφραστεί με διάφορους τρόπους. Στη βιβλιογραφία το data mining συναντάται και με τον όρο «knowledge discovery in databases» (KDD – ανακάλυψη γνώσης από βάσεις δεδομένων), όρο δανεισμένο από την τεχνητή νοημοσύνη, μια διαδικασία που περιγράφεται σχηματικά ως ακολούθως.



Εικόνα 2: Τα βήματα του Data Mining

Συνοψίζοντας, σε ένα τυπικό πρόβλημα data mining υπάρχει μία λίστα βημάτων που πρέπει να πραγματοποιηθούν:

1. Ανάπτυξη και κατανόηση του στόχου της περιοχής της εφαρμογής, της σχετικά προγενέστερης γνώσης του προς εξέταση τομέα, και τους στόχους του τελικού χρήστη.

2. Απόκτηση του συνόλου δεδομένων που θα χρησιμοποιηθούν στην ανάλυση. Υπάρχουν διαφορετικά είδη αποθηκών πληροφοριών που μπορούν να χρησιμοποιηθούν στη διαδικασία εξόρυξης γνώσης. Κατά συνέπεια, οι

πολλαπλές πηγές δεδομένων μπορούν να συνδυαστούν καθορίζοντας το σύνολο στο οποίο πρόκειται τελικά να εφαρμοστεί η διαδικασία εξόρυξης. Έτσι δημιουργείται ο στόχος –σύνολο δεδομένων (dataset) και επιλέγεται το σύνολο δεδομένων στο οποίο πρόκειται να εκτελεστεί η διαδικασία εξόρυξης.

3. Εξερεύνηση, καθαρισμός και προεπεξεργασία των δεδομένων.
   Αυτό το βήμα περιλαμβάνει βασικές διαδικασίες όπως είναι η αφαίρεση του θορύβου ή των άτυπων τιμών των outliers, η συλλογή των απαραίτητων πληροφοριών για τη διαμόρφωση ή τη μέτρηση του θορύβου, η απόφαση σχετικά με τις στρατηγικές διαχείρισης των ελλειπόντων πεδίων δεδομένων.

4. Μείωση των δεδομένων, αν είναι απαραίτητο, και (όταν εμπλέκεται μάθηση με επίβλεψη) , διαχωρισμός των δεδομένων σε δεδομένα εκπαίδευσης (training set), δεδομένα επαλήθευσης (quiz set-validation set) και δεδομένα ελέγχου (test dataset). Τα δεδομένα μετασχηματίζονται ή παγιώνονται σε μορφές κατάλληλες για εξόρυξη. Γίνεται χρήση μεθόδων μείωσης διαστάσεων ή μετασχηματισμού, για τη μείωση του αριθμού των υπό εξέταση μεταβλητών ή την εύρεση κατάλληλης αντιπροσώπευσης των δεδομένων.

5. Προσδιορισμός του είδους της μάθησης του data mining (ταξινόμηση, πρόβλεψη,ομαδοποίηση, συσχέτιση). Εκτέλεση της εξόρυξης δεδομένων.

6. Επιλογή των τεχνικών του data mining που θα χρησιμοποιηθούν.

7. Χρησιμοποίηση αλγορίθμων για να διεξάγουν το έργο της μάθησης (αλγόριθμοι εκμάθησης).

8. Ερμηνεία των αποτελεσμάτων των αλγορίθμων-αξιολόγηση των προτύπων.
   Τα εξαγόμενα πρότυπα ή μοτίβα αξιολογούνται με κάποια μέτρα, προκειμένου να προσδιοριστούν εκείνα τα οποία αντιπροσωπεύουν τη γνώση καλύτερα, δηλαδή τα μοτίβα για τα οποία ενδιαφερόμαστε περισσότερο.

9. Ανάπτυξη του μοντέλου - σταθεροποίηση και παρουσίαση της γνώσης.
   Σε αυτό το βήμα, η εξορυγμένη γνώση ενσωματώνεται στο σύστημα και χρησιμοποιούνται κάποιες τεχνικές αντιπροσώπευσης αυτής, προκειμένου να παρουσιαστεί ευκρινώς στο χρήστη.

1.3 Το data mining στη συμβολή άλλων επιστημονικών πεδίων

Το data mining αναγνωρίζεται ως ένα πολύ δυνατό εργαλείο, αλλά για να είναι τα αποτελέσματα της εφαρμογής του σε πρακτικά προβλήματα ασφαλή, δεν θα πρέπει να στηρίζονται μόνο στην εφαρμογή των αλγορίθμων του σε υποδείγματα, δηλαδή στη

μηχανική μάθηση μέσω υποδειγμάτων, αλλά να συνδυάζονται και με στατιστική ανάλυση. Μπορούμε επομένως να πούμε ότι η στατιστική ανάλυση και οι αλγόριθμοι εξόρυξης πληροφορίας από βάσεις δεδομένων αποτελούν δύο βασικά συστατικά του data mining για την ανάλυση δεδομένων πρακτικών προβλημάτων. Στην πραγματικότητα όμως, η «καταγωγή» του data mining έχει τις ρίζες της σε τρεις βασικούς επιστημονικούς κλάδους: τη στατιστική (statistics), την τεχνητή νοημοσύνη (artificial intelligence) και τη μηχανική μάθηση (machine learning).

Εικόνα 3: Οι «ρίζες» της εξόρυξης δεδομένων

Στη στατιστική έρευνα, το Data Mining αναπτύχθηκε ως μια μέθοδος για να ευρίσκει μοντέλα ή πρότυπα «πίσω» από τις σχέσεις. Το πλεονέκτημα της στατιστικής είναι η ικανότητά της να ερμηνεύει τα δεδομένα με μαθηματικό τρόπο, ενώ το μειονέκτημά της εντοπίζεται στη δυσκολία της να γενικεύσει σε πολύ μεγάλα σύνολα δεδομένων. Από τη στατιστική μπορούμε να βρούμε και να μετρήσουμε τη δύναμη μιας σχέσης ανάμεσα σε δύο ή περισσότερες μεταβλητές από τα δεδομένα. Αλλά αυτή η στατιστική μέτρηση από μόνη της δεν είναι ικανή να εξηγήσει γιατί υπάρχει αυτή η σχέση ή τί πιθανή επίδραση μπορεί να έχει στο μέλλον. Το Data Mining μας εφοδιάζει με εκείνα τα εργαλεία ούτως ώστε να ερευνήσουμε σε βάθος τα δεδομένα και να αποκτήσουμε επιπλέον γνώση για τις σχέσεις εξάρτησης. Αυτό το επιτυγχάνει μέσω μιας διαδραστικής, επαναληπτικής ή/και διερευνητικής ανάλυσης των δεδομένων. Οι σχέσεις που προκύπτουν από το data mining συχνά αναφέρονται ως μοντέλα ή πρότυπα και περιλαμβάνουν γραμμικές εξισώσεις, κανόνες, συστάδες, γραφήματα, δέντρα, επαναλαμβανόμενα πρότυπα σε χρονοσειρές κ.α.

Σε μία σύγκριση μεταξύ του data mining και της στατιστικής θα μπορούσαμε να παρουσιάσουμε συνοπτικά τα παρακάτω:

Data Mining και Στατιστική

- Data Mining

- Δεν χρειάζονται υποθέσεις.
- Μπορεί να βρει πρότυπα σε τεράστιες ποσότητες δεδομένων.
- Χρησιμοποιεί όλα τα διαθέσιμα δεδομένα.
- Ορολογία που χρησιμοποιείται: πεδίο, καταχώρηση, μάθηση με επίβλεψη, μάθηση χωρίς επίβλεψη.

- Στατιστική

- Χρησιμοποιεί τεστ υποθέσεων (hypothesis testing).
- Οι τεχνικές δεν είναι κατάλληλες για μεγάλα σύνολα δεδομένων.
- Βασίζεται στη δειγματοληψία.
- Ορολογία που χρησιμοποιείται: μεταβλητή, παρατήρηση, ανάλυση της εξάρτησης,ανάλυση της αλληλεξάρτησης.

Η εξόρυξη δεδομένων σε γενικά πλαίσια καλείται να αντιμετωπίσει

- Το τεράστιο μέγεθος των δεδομένων
- Το μεγάλο αριθμό διαστάσεων
- Τη μη ομοιογενή και την κατανεμημένη φύση των δεδομένων

1.4 Κατηγορίες του data mining

Το data mining διακρίνεται σε δύο βασικές κατηγορίες, το κατευθυνόμενο data mining (directed data mining) και το μη κατευθυνόμενο data mining (undirected data mining). Το directed data mining προσπαθεί να εξηγήσει ή να ταξινομήσει κάποια πεδία στόχους. Το undirected data mining προσπαθεί να βρει ομοιότητες σε ομάδες εγγραφών χωρίς τη χρήση ενός συγκεκριμένου πεδίου στόχου ή μιας συλλογής προκαθορισμένων κλάσεων.

Μια άλλη βασική κατηγοριοποίηση αφορά στις τεχνικές του data mining ,οι οποίες διακρίνονται σε «supervised» και «unsupervised» μεθόδους.

1. Αλγόριθμοι εκμάθησης με επίβλεψη (supervised learning algorithms) είναι εκείνοι που χρησιμοποιούνται στην ταξινόμηση και στην πρόβλεψη. Στην περίπτωση αυτή, πρέπει να έχουμε διαθέσιμα δεδομένα, των οποίων η τιμή του αποτελέσματος που μας ενδιαφέρει είναι γνωστή. Πιο συγκεκριμένα, ο όρος supervised learning αναφέρεται στη διαδικασία του να τροφοδοτήσεις έναν αλγόριθμο με εγγραφές στις οποίες μια μεταβλητή απόκρισης (output variable)

που μας ενδιαφέρει είναι γνωστή και ο αλγόριθμος να «μάθει» πώς να προβλέψει την τιμή με νέες εγγραφές όπου το αποτέλεσμα είναι άγνωστο. Δηλαδή, μοντελοποιούν μια μεταβλητή απόκρισης βασιζόμενοι σε μία ή περισσότερες επεξηγηματικές μεταβλητές (input variable).

2. Αλγόριθμοι εκμάθησης χωρίς επίβλεψη (unsupervised learning algorithms) είναι εκείνοι που χρησιμοποιούνται όταν δεν υπάρχει μία μεταβλητή απόκρισης να προβλεφθεί ή να ταξινομηθεί. Πιο συγκεκριμένα, ο όρος unsupervised learning αναφέρεται στην ανάλυση που κάποιος επιχειρεί, για να μάθει κάτι άλλο για τα δεδομένα πέρα από την πρόβλεψη της τιμής μίας μεταβλητής που τον ενδιαφέρει, αν ανήκει σε κάποιο cluster για παράδειγμα. Δηλαδή οι unsupervised τεχνικές χρησιμοποιούνται όταν δεν υπάρχει κάποιο πεδίο να προβλεφθεί αλλά οι σχέσεις των δεδομένων εξευρευνούνται ώστε να ανακαλυφθεί η γενική δομή τους.

Πολύ συχνά supervised και unsupervised τεχνικές χρισιμοποιούνται μαζί. Η επιλογή ενός κατάλληλου συνδυασμού από τεχνικές που πρέπει να εφαρμοστούν σε ένα συγκεκριμένο πρόβλημα εξαρτάται από τη φύση του σκοπού του data mining ,από τη φύση των διαθέσιμων δεδομένων, καθώς επίσης και από τις δυνατότητες και τις προτιμήσεις του data miner.

1.5 Παραδοσιακές εφαρμογές

Η τεχνική του Data Mining σε ένα επιχειρηματικό περιβάλλον αποτελεί έναν από τους συνηθέστερους τρόπους απόκτησης οργανωτικής οξυδέρκειας. Παραδοσιακά, το Data Mining όσον αφορά τις επιχειρήσεις επικεντρώνεται στην κατάκτηση νέας γνώσης από τα δεδομένα. Τα δεδομένα αυτά μπορεί να είναι είτε δομημένα, ακατέργαστα μεν, όπως για παράδειγμα οι λιανικές αγορές, είτε μη δομημένα και δύσκολα ως προς την ανάλυσή τους. Είτε έτσι είτε αλλιώς, μόλις τα δεδομένα μας «καθαριστούν», αναλυθούν, και οργανωθούν καταλλήλως τότε το Data Mining ξεκινάει να διαδραματίζει το ρόλο του.

Τράπεζες, επιχειρήσεις, κατασκευαστικές εταιρίες, επιχειρήσεις τηλεπικοινωνιών, ασφαλιστικές, οργανισμοί υγείας και αθλητικοί σύλλογοι, μεταξύ άλλων, χρησιμοποιούν το Data Mining για να ανακαλύψουν σχέσεις, πρότυπα και συσχετίσεις, από την τιμολόγηση και την προώθηση προϊόντων μέχρι το πως η οικονομία, ο ανταγωνισμός και τα μέσα κοινωνικής δικτύωσης επηρεάζουν τα επιχειρηματικά τους πλάνα και τις σχέσεις μεταξύ πελατών.

Τηλεπικοινωνίες

Σε μια υπερφορτωμένη αγορά, όπου ο ανταγωνισμός είναι έντονος, οι απαντήσεις συχνά βρίσκονται στα δεδομένα των καταναλωτών. Για παράδειγμα, η εταιρία πολυμέσων Sanoma χρησιμοποιεί αναλυτικά μοντέλα προκειμένου να εξάγει συμπεράσματα σχετικά με τις συναλλαγές μιας εβδομάδας και να προβλέπει τη συμπεριφορά των πελατών έτσι ώστε να προτείνει στοχευμένες και σχετικές καμπάνιες.

Βιομηχανία

Ο συντονισμός των σχεδίων προμήθειας ανάλογα με τις προβλέψεις της ζήτησης είναι απαραίτητος, έτσι ώστε να καθίσταται δυνατή η έγκαιρη διάγνωση προβλημάτων και η διασφάλιση της ποιότητας. Για παράδειγμα, η αυτοκινητοβιομηχανία Volvo αναλύει περισσότερες από 100 παραμέτρους στα οχήματά της με σκοπό να προβλέψει τη φθορά και την αποφυγή ενδεχόμενων απρόβλεπτων συμβάντων, μειώνοντας ταυτόχρονα το χρόνο απόκρισης.

Εμπόριο

Τεράστιες βάσεις δεδομένων πελατών «κρύβουν» πληροφορίες που μας βοηθούν να βελτιώσουμε τις σχέσεις μεταξύ επιχειρήσεων και πελατών, να βελτιστοποιήσουμε τις καμπάνιες μάρκετινγκ και να προβλέψουμε τον αριθμό των πωλήσεων. Η εταιρία Staples ερευνά περίπου 1500 διακαναλικές καμπάνιες ετησίως βασισμένες σε 25 εκατομμύρια μητρώα πελατών.

Οργανισμοί Υγείας

Με τη χρήση του Data Mining, οι ασφαλιστές υγείας μπορούν να ελλατώσουν τις περιπτώσεις απάτης, τα διαγνωστικά κέντρα να βελτιώσουν τις παροχές τους και οι ασθενείς να δέχονται ασφαλέστερες και οικονομικότερες υπηρεσίες υγείας.

1.6 Συμπεράσματα

Παρατηρούμε λοιπόν, πως αν και ο όρος Data Mining δεν είναι ευρέως γνωστός, παρόλα αυτά εφαρμόζεται ευρέως σε πολλούς και διαφορετικούς τομείς, βελτιώνοντας την καθημερινότητά μας σε μεγάλο βαθμό. Ίσως κι εμείς οι ίδιοι να έχουμε κάνει χρήση του, εμπειρικά βέβαια, σε διάφορες εκφάνσεις της ζωής μας χωρίς να το έχουμε καταλάβει. Το μόνο σίγουρο είναι ότι ότι με την πάροδο των χρόνων και τη διαρκή εξέλιξη των τεχνολογικών μέσων η εφαρμογή του Data Mining θα αποτελεί επιτακτική ανάγκη και αυτό θα έχει ως αποτέλεσμα να διεισδύσει σε ακόμη περισσότερους επιστημονικούς τομείς.

2. Sports Data Mining

2.1 Εισαγωγή

Τεράστιες ποσότητες δεδομένων υπάρχουν σε όλους τους τομείς των σπορ. Τα δεδομένα αυτά μπορεί να εκφράζουν την ατομική επίδοση ενός παίκτη, τις αποφάσεις που πρέπει να πάρει ο προπονητής ή ο μάνατζερ του συλλόγου μέχρι το πως λειτουργεί η ομάδα σαν σύνολο. Για να έχουν σημασία τα παραπάνω δε μας ενδιαφέρει το πως θα συλλέξουμε τα δεδομένα, αλλά ποια δεδομένα πρέπει να συλλέξουμε έτσι ώστε να τα χρησιμοποιήσουμε με τον καλύτερο δυνατό τρόπο. Βρίσκοντας, λοιπόν, τους σωστούς τρόπους συλλογής δεδομένων με πρακτικό νόημα, και στη συνέχεια μέσω της εξόρυξης δεδομένων και της μετατροπής του όγκου των δεδομένων σε πρακτική γνώση, οι αθλητικοί σύλλογοι έχουν τη δυνατότητα να εξασφαλίζουν ανταγωνιστικό πλεονέκτημα απέναντι στους αντιπάλους τους. Η αναζήτηση γνώσης μπορεί να εφαρμοστεί σε όλα τα επίπεδα μιας αθλητικής οργάνωσης. Από τους παίκτες, βελτιώνοντας την επίδοσή τους χρησιμοποιώντας αναλυτικές τεχνικές βίντεο μέχρι τους αθλητικούς κατασκόπους (scouts) κάνοντας χρήση στατιστικών αναλύσεων και μεθόδων προβολής με στόχο να αναγνωρίσουν ποια παράμετρος θα έχει τη μεγαλύτερη επίδραση. Με αυτόν τον τρόπο το Data Mining γρήγορα μετετρέπεται σε αναπόσπαστο κομμάτι στη διαδικασία λήψης αποφάσεων όπου οι προπονητές χρησιμοποιώντας τεχνικές μηχανικής μάθησης και προσομοίωσης μπορούν να ανακαλύψουν βέλτιστες στρατηγικές για ολόκληρη την επόμενη αγωνιστική περίοδο.

Το πρώτο μέρος του προβλήματος είναι ο εντοπισμός των μετρήσεων της επίδοσης. Αρκετές ήδη υπάρχουσες αθλητικές μετρήσεις μπορεί εύκολα να «κακομεταχειριστούν» ή χειρότερα να μη δίνουν ως αποτέλεσμα την επίδοση στο πλαίσιο της επίτευξης περισσότερων πόντων από τους αντιπάλους, που είναι και ο ουσιαστικός στόχος όλων των αθλητικών οργανώσεων.

Το δεύτερο μέρος του προβλήματος εντοπίζεται στην ανακάλυψη αξιόλογων και ενδιαφερόντων προτύπων στα δεδομένα. Τα πρότυπα αυτά μπορεί να περιλαμβάνουν τις τάσεις των αντίπαλων παικτών ή ομάδων, τον προσδιορισμό εμφάνισης τραυματισμών ή τη δημιουργία προβλέψεων βασισμένων σε ιστορικά δεδομένα.

Επαγγελματικοί αθλητικοί σύλλογοι, συμπεριφέρονται σήμερα, ως επιχειρήσεις πολλών εκατομυρίων ξοδεύοντας αρκετά χρήματα σε μία μόνο απόφαση. Με τέτοια ποσότητα κεφαλαίου να διακυβεύεται, είναι κατανοητό πως μια κακή ή εσφαλμένη απόφαση είναι ικανή να γυρίσει ένα σύλλογο χρόνια πίσω. Έτσι λοιπόν, με αρκετές περιπτώσεις ρίσκου στη διαδικασία λήψης αποφάσεων, η βιομηχανία των σπορ αποτελεί ένα ελκυστικό περιβάλλον για εφαρμογές του Data Mining.

2.2 Εφαρμογή του data mining στα σπορ

Υποθέτουμε ότι υπάρχουν εκατοντάδες σχέσεις μεταξύ των σπορ και των τεχνικών που έχουν σχεδιαστεί για να κάνουν χρήση των δεδομένων που σχετίζονται με αυτά. Οι σχέσεις αυτές εξαρτώνται από τη φύση των οργανώσεων και της ανάγκης τους για ιστορικά δεδομένα. Στη συνέχεια παρουσιάζουμε μια προτεινόμενη ιεραρχία των επιπέδων εξάρτησης και των χαρακτηριστικών τους.

Στην πρώτη περίπτωση δεν υπάρχει καμία σχέση μεταξύ των σπορ και των δεδομένων που προκύπτουν από κάθε παιχνίδι. Αυτοί είναι επαγγελματικοί σύλλογοι οι οποίοι παίζουν το παιχνίδι τους, καταγράφουν στοιχεία που προκύπτουν από κάθε παίκτη και δεν πραγματοποιούν τίποτα περαιτέρω με αυτά τα δεδομένα. Γι' αυτούς τους οργανισμούς, η συγκέντρωση των δεδομένων αποτελεί είτε απλή καταφραφή των γεγονότων ενός αγώνα είτε είναι αποτέλεσμα παράδοσης. Σε αυτήν την κατηγορία ανήκουν ερασιτεχνικοί αθλητικοί σύλλογοι οι οποίοι δίνουν έμφαση στη ψυχαγωγία ή στην εκμάθηση των βασικών αρχών ενός σπορ (π.χ. οργανισμοί στους οποίους συμμετέχουν παιδιά μικρής ηλικίας ή ενήλικες που κάνουν το χόμπι τους για διασκέδαση).

Η επόμενη κατηγορία χαρακτηρίζεται από τη χρησιμοποίηση ανθρώπων ειδικών στο χώρο του αθλητισμού προκειμένου να κάνουν προβλέψεις βασιζόμενοι στην εμπειρία τους. Αρχικά, υπήρχε η πεποίθηση πως αυτοί οι ειδικοί (μάνατζερς, προπονητές, scouts) μπορούσαν να συνδυάσουν αποτελεσματικά τις παρατηρήσεις με την εμπειρία τους, με σκοπό να παίρνουν ικανοποιητικές αποφάσεις. Οι αποφάσεις που προκύπτουν από αυτού του είδους τις σχέσεις βασίζονται σε εικασίες ή στο ένστικτο και όχι σε χειροπιαστά στοιχεία. Τέτοιου είδους αποφάσεις μπορεί να περιλαμβάνουν την εκτέλεση ορισμένων συστημάτων (plays) ή μιας αλλαγής ενός παίκτη (substitution) κατά τη διάρκεια του αγώνα μόνο και μόνο επειδή αυτός που παίρνει την απόφαση αισθάνεται ότι είναι σωστή, χωρίς να λαμβάνει υπόψη του προηγούμενα δεδομένα.

Η τρίτη κατηγορία είναι εκείνη στην οποία οι ειδικοί ξεκινούν να κάνουν χρήση των ιστορικών δεδομένων. Αποφάσεις από αυτήν την κατηγορία περιλαμβάνουν τη δημιουργία ευνοϊκών μαρκαρισμάτων μεταξύ των παικτών ή την εκτέλεση εκείνων των plays που ιστορικά αποφέρουν το καλύτερο αποτέλεσμα.

Η τέταρτη κατηγορία αρχίζει να ενσωματώνει τη στατιστική στη διαδικασία λήψης αποφάσεων. Αυτές οι στατιστικές μετρήσεις μπορεί να είναι απλά αριθμητικές συχνότητες συγκεκριμένων γεγονότων ή πιο σύνθετα, μέθοδοι που αναλύουν τη συνολική επίδοση της ομάδας, και προσδιορίζουν την επίδοση κάθε παίκτη ξεχωριστά ανάλογα με τη συνεισφορά του στην προσπάθεια της να πετύχει το στόχο της. Η τελευταία κατηγορία που περιγράφει τη σχέση μεταξύ των σπορ και των δεδομένων που σχετίζονται με αυτά, χαρακτηρίζεται από τη χρήση της μεθόδων του Data Mining. Οι τεχνικές αυτές διαφέρουν απ' όλες τις προηγούμενες καθώς έχουν τη δυνατότητα να προσαρμόζονται σε νέες καταστάσεις και προβλέψεις τις οποίες οι ίδιες

δημιουργούν. Καθώς οι στατιστικές τεχνικές αποτελούν ακόμα την καρδιά του Data Mining, τα στατιστικά στοιχεία χρησιμοποιούνται για να ξεχωρίσουν εδιαφέροντα πρότυπα, όπως για παράδειγμα τις τάσεις ενός παίκτη, επιτρέποντας στους αναλυτές και κατ' επέκταση στις αθλητικές οργανώσεις να κάνουν προβλέψεις ανάλογα με το αποτέλεσμα. Τα στατιστικά στοιχεία από μόνα τους δεν μπορούν να εξηγήσουν τη συσχέτιση των παραπάνω. Αυτός είναι ο σκοπός του Data Mining. Αυτού του είδους η συσχέτιση έχει τη δυναμική να αυξήσει τις αποφάσεις των ειδικών ή να χρησιμοποιηθεί ανεξάρτητα στη λήψη αποφάσεων χωρίς την ενέργεια των παραπάνω.

Η τελευταία αυτή χρήση των μεθόδων Data Mining, χωρίς δηλαδή την επίδραση της ανθρώπινης ενέργειας, πρέπει να είναι ανεξάρτητη και από τις διάφορες ανθρώπινες προκαταλήψεις, οι οποίες παίζουν καθοριστικό ρόλο στη διαδικασία λήψης αποφάσεων. Ένα χαρακτηριστικό παράδειγμα αυτού είναι όταν ένας αθλητικός scout επικεντρώνεται μόνο στα θετικά στοιχεία ενός παίκτη, αγνοώντας τις αδυναμίες του. Αφαιρώντας, λοιπόν, τις ανθρώπινες προκαταλήψεις από τη διαδικασία λήψης αποφάσεων, προπονητές και μάνατζερς μπορούν να διαχειρίζονται καλύτερα τους παίκτες τους και να παίρνουν αντικειμενικές αποφάσεις προς όφελος της ομάδας.

Αρκετοί αθλητικοί σύλλογοι ανήκουν σε ορισμένες από τις παραπάνω κατηγορίες, όπου τα δεδομένα χρησιμοποιούνται σε κάποιο βαθμό στη διαδικασία λήψης αποφάσεων. Παίρνοντας ως παράδειγμα το επαγγελματικό baseball, αρκετές ομάδες ανήκουν στην τρίτη ή τέταρτη από τις παραπάνω κατηγορίες. Οι σύλλογοι αυτοί χρησιμοποιούν τα δεδομένα έτσι ώστε να φτάσουν στους κατάλληλους συνδυασμούς παικτών (batters versus pitchers) με τους οποίους τα συστήματά τους θα έχουν περισσότερες πιθανότητες επιτυχίας. Υψηλότερου επιπέδου παραδείγματα περιλαμβάνουν μέτρηση της αξίας απόδοσης ενός παίκτη (π.χ. πόσο συχνά ένας παίκτης φτάνει στη βάση) μέχρι το πόσο καλά αποδίδει σε σύγκριση με τους υπόλοιπους παίκτες του πρωταθλήματος. Πολλοί λίγοι σύλλογοι φτάνουν στο τελικό στάδιο και υϊοθετούν τις τεχνικές του Data Mining. Παρότι η εισαγωγή του Data Mining στο χώρο του αθλητισμού είναι σχετικά πρόσφατη, η επίδρασή του στις ομάδες που χρησιμοποιούν αυτές τις τεχνικές είναι καθοριστική.

Γυρνώντας στο παράδειγμα του baseball, η ομάδα Oakland Athletics (A's), συνήθως τερμάτιζε στις τελευταίες θέσεις του πρωταθλήματος εξαιτίας των χαμηλών μισθών που πλήρωνε στους παίκτες της. Υπήρχε η πεποίθηση πως το ύψος των χρημάτων που ξόδευε η ομάδα για τους παίκτες είχε άμεση σχέση με την επιτυχία της, καθώς παραδοσιακά όσο περισσότερα χρήματα ξοδεύονται για το ρόστερ τόσες περισσότερες νίκες πετυχαίνει η ομάδα. Στις αρχές του 2000 οι A's ξεκίνησαν να χρησιμοποιούν το Data Mining ως ένα τρόπο να παραμένουν ανταγωνιστικοί. Από τότε ξεκίνησε μια περίοδος επιτυχιών, είτε τερματίζοντας κάθε χρόνο σε όλο και υψηλότερες θέσεις στη βαθμολογία είτε συμμετέχοντας στα playoffs παρά το μικρό τους προϋπολογισμό (budget).

Καθώς η χρησιμοποίηση των στατιστικών στοιχείων στη διαδικασία λήψης αποφάσεων αποτελεί σίγουρα βελτίωση ενάντια μόνο στο ένστικτο, τα στατιστικά από

μόνα τους μπορεί να είναι παραπλανητικά χωρίς την κατανόηση του βασικού νοήματός τους. Αυτή η παραπλανητική τάση μπορεί να προέρχεται είτε από την ανακριβή μέτρηση της επίδοσης είτε από την υπερβολική έμφαση μερικών στατιστικών στοιχείων από την αθλητική κοινότητα. Σαν απόδειξη σκεφτείτε το γεγονός ότι ορισμένοι παίκτες μπορεί να έχουν εντυπωσιακά ατομικά στατιστικά, η επίδρασή τους όμως στην απόδοση της ομάδας να είναι μικρή. Τα διάφορα στατιστικά στοιχεία στο χώρο των σπορ «υποφέρουν» από ανακρίβεια, καθώς από μόνα τους δεν μπορούν να απεικονίσουν την ατομική συνεισφορά στη συνολική απόδοση της ομάδας.

Το άθλημα του baseball, όπως προαναφέραμε δεν είναι το μοναδικό σπορ που χαρακτηρίζεται από ανακρίβεια των στατιστικών του στοιχείων. Ας περάσουμε στο **άθλημα της καλαθοσφαίρησης (basketball), το οποίο μελετά η παρούσα εργασία**, κι ας μελετήσουμε ένα παράδειγμα. Στο παιχνίδι του basketball, το αμυντικό rebound απεικονίζει τις φορές τις οποίες ένας παίκτης που αμύνεται αποκτά κατοχή της μπάλας ύστερα από μια άστοχη προσπάθεια (missed shot). Για να «μαζέψει» όμως ένας παίκτης το rebound, σημαίνει πως οι συμπαίκτες του «μπλόκαραν» τους αντιπάλους τους, με αποτέλεσμα οι ίδιοι να μην μπορούν να πιάσουν τη μπάλα.

Εικόνα 4: Block Out

Παρόλα αυτά, η πράξη τους να μπλοκάρουν τους αντιπάλους τους, τους καθιστά το ίδιο σημαντικούς με εκείνον τον παίκτη που τελικά έπιασε τη μπάλα. Με τον τρόπο όμως που μετριέται ένα rebound, μόνο ο παίκτης που παίρνει στην κατοχή του τελικά τη μπάλα πιστώνεται με αυτό.

Εικόνα 5: Rebound

Για όλους τους παραπάνω λόγους, δε θα πρέπει να αποτελεί έκπληξη το γεγονός πως οι αθλητικοί σύλλογοι που έχουν υϊοθετήσει αυτές τις τεχνικές απολαμβάνουν και την επιτυχία. Οι παραδοσιακές προσεγγίσεις λήψης αποφάσεων από διαίσθηση ή ένστικτο σταδιακά ξεπερνώνται. Αντιθέτως, πλεόν γίνονται εκτιμήσεις στη βάση βαθιάς ανάλυσης και επιστημονικής έρευνας. Με ολοένα και περισσότερες ομάδες να αγκαλιάζουν τη ψηφιακή εποχή, σύντομα θα είναι μια μάχη εύρεσης του καλύτερου αλγορίθμου ή της καλύτερης φόρμας μέτρησης της απόδοσης, με τους ειδικούς αναλυτές να είναι το ίδιο σημαντικοί με τους παίκτες που αγωνίζονται στο γήπεδο.

2.3 Ιστορία

Η υϊοθέτηση του Data Mining από τα οργανωμένα αθλητικά σωματεία δεν ήταν φαινόμενο μιας νύχτας. Αντίθετα, η εφαρμογή του πέρασε από πολλά στάδια και διήρκησε αρκετές δεκαετίες. Στο άθλημα του baseball, η βασική αλλαγή από τη χρήση μόνο των στατιστικών στοιχείων πιστώνεται στον Bill James. Το 1977, ο James ξεκίνησε να εκδίδει το ετήσιο περιοδικό του που ονομαζόταν «Bill James Baseball Abstracts». Τα κείμενα αυτά χρησιμοποιήθηκαν ως το προσωπικό του forum, όπου σχολίαζε τους παραδοσιακούς τρόπους μέτρησης της απόδοσης και διατύπωνε ερμηνευτικά σχόλια σχετικά με τα προβλήματα που παρουσίαζαν οι μέχρι τότε υπάρχουσες μέθοδοι για τη μέτρησή της. Παρά το γεγονός ότι αρχικά πωλούσε μόνο 50 αντίγραφα, ο James συνέχισε να εκδίδει το ετήσιο περιοδικό του με απόψεις, ανορθόδοξους τύπους κατάταξης και νέους τρόπους μέτρησης της απόδοσης, τους

οποίους ονόμασε sabermetrics. Οι αναγνώστες άρχισαν να αυξάνονται και οι εγγραφόμενοι στο forum του James ξεκίνησαν να ενδιαφέρονται για καινούριους τρόπους περιγραφής της απόδοσης. Ο ίδιος ο James το 1982 έσπευσε να περιγράψει τη φιλοσοφία του γύρω από τα sabermetrics, λέγοντας πως τα νούμερα και τα στατιστικά στοιχεία δεν είναι το θέμα συζήτησης. Το θέμα είναι το παιχνίδι του baseball. Η σχέση που έχουν οι αριθμοί με το παιχνίδι και με εμάς τους ίδιους που τα μελετούμε, είναι ακριβώς η σχέση που έχουν τα εργαλεία μιας μηχανής με το μηχανικό που τη χρησιμοποιεί για να κάνει τη δουλειά του.

Τα επόμενα χρόνια, οι παραπάνω τεχνικές δεν είχαν μεγάλη απήχηση στους αθλητικούς συλλόγους, καθώς ήταν διστακτικοί στην εφαρμογή τους, μιας και θεωρήθηκαν φιλοσοφικές και μη τεκμηριωμένες επιστημονικά. Καθώς τις μετέπειτα δεκαετίες τα sabermetrics εξελίχθηκαν κι έγιναν ευρέως πια γνωστά σε αρκετά σπορ, το 2002 ενσωματώθηκαν στο άθλημα του baseball. Ο General Manager των Oakland A's, Billy Beane, στην προσπάθειά του να φτιάξει μια ανταγωνιστική ομάδα χρησιμοποίησε τα sabermetrics για την επιλογή των παικτών χωρίς να συβμουλευτεί τους scouts της ομάδας του. Η τακτική αυτή βρήκε την ομάδα του τα επόμενα 5 χρόνια είτε στα playoffs είτε στη διεκδίκηση εισαγωγής τους σε αυτά. Μέσω των sabermetrics, ο Beane είχε το προνόμιο να αξιολογεί τους παίκτες με βάση δείκτες οι οποίοι μετρούσαν την απόδοσή τους κι όχι σύμφωνα με τα παραδοσιακά στατιστικά. Με αυτό τον τρόπο επέλεγε παίκτες τους οποίους άλλες ομάδες προσπερνούσαν. Ερχόταν σε συμφωνία μαζί τους με συμβόλαια μεγάλης διάρκειας με μικρό κόστος και ουσιστικά επένδυε σε ένα πλάνο για αρκετά χρόνια.

Το baseball, όμως δεν ήταν το μοναδικό άθλημα που υφίσταται στατιστική αναμόρφωση. Τη δεκαετία του 1980, ο Dean Oliver ξεκίνησε να αναζητά παρόμοιες τεχνικές για το παιχνίδι του basketball. Ο Oliver ενδιαφερόταν περισσότερο στη δημιουργία στατιστικών που αφορούν συνολικά την ομάδα κι όχι τον κάθε παίκτη ξεχωριστά. Όπως ο James, έτσι και ο Oliver το 2005 εκδίδει τις σκέψεις του και μεθόδους μέτρησης απόδοσης στη μπασκετική κοινότητα. Μέσα από τη δουλειά του διευκόλυνε τους αναλυτές να αναγνωρίσουν τη συνεισφορά κάθε παίκτη κι ακόμα περισσότερο να διερευνήσουν τη χημεία της ομάδας μετρώντας το πόσο καλά ορισμένοι παίκτες συνεργάζονται μεταξύ τους. Τόσο ο James όσο και ο Oliver είναι παγκοσμίως αναγνωρισμένοι από τους επαγγελματικούς αθλητικούς συλλόγους ως θεμελιώδεις σπορ αναλυτές, επιβεβαιώνοντας έτσι το ρόλο της στατιστικής ανάλυσης στο χώρο του αθλητισμού.

Ακολουθώντας τις επαναστατικές μεθόδους μέτρησης της απόδοσης, το Data Mining γρήγορα έκανε την εμφάνισή του και άρχισε να χρησιμοποιείται στη διαδικασία λήψης αποφάσεων. Αθλητικοί σύλλογοι από διαφορετικά σπορ ξεκίνησαν να παίρνουν στα σοβαρά τις τεχνικές εξόρυξης δεδομένων προκειμένου να αποκτήσουν ανταγωνιστικό πλεονέκτημα. Ενώ κάποια αθλητικά σωματεία χρησιμοποιούν κατά κόρον το Data Mining ως προέκταση της άσκησης, υπάρχουν και μερικά που αξιοποιούν τις μεθόδους αυτές με τρόπο κάπως ανορθόδοξο μεν, αλλά ιδιαιτέρως ενδιαφέρον δε. Ένα τέτοιο παράδειγμα αποτελεί το εργαλείο πρόβλεψης βιοϊατρικών τραυματισμών που

χρησιμοποιεί η πολλή γνωστή ομάδα του ιταλικού πρωταθλήματος ποδοσφαίρου AC Milan. Το συγκεκριμένο εργαλείο χρησιμοποιεί λογισμικό το οποίο απεικονίζει την ποιότητα προπόνησης των παικτών και συγκρίνει τα αποτελέσματα με κάποιες αρχικές τιμές. Οποιαδήποτε τιμή βρίσκεται κάτω από την αναμενόμενη, μπορεί να σημαίνει πως είτε ο αθλητής υποφέρει από έναν τραυματισμό τον οποίο δεν έχει αποκαλύψει είτε ένας ήδη υπάρχον έχει επιδεινωθεί. Ένα δεύτερο ενδιαφέρον παράδειγμα χρήσης του Data Mining είναι η δημιουργία ενός λογισμικού από τους Αθλητικούς Συμβούλους του Las Vegas το οποίο ερευνά τη διαδικασία των στοιχημάτων στα σπορ προσπαθώντας να ανακαλύψει ύποπτα στοιχήματα τα οποία οδηγούν σε «στημένους» αγώνες. Ένα τελευταίο παράδειγμα έρχεται από το χώρο του επαγγελματικού αμερικανικού football (NFL), όπου οι ειδικοί αναλυτές προσπαθούν να ανακαλύψουν τη σχέση που έχει η σωματική ικανότητα με την απόδοση του παίκτη, περνώντας τους κάθε χρόνο από ειδικές εξετάσεις. Πέρα όμως από τη σωματική ικανότητα οι αθλητές αξιολογούνται και για την πνευματική τους επάρκεια μέσω εξειδικευμένων μετρήσεων. Αξίζει να σημειωθεί πως ανάλογα με τη θέση στην οποία αγωνίζεται ένας παίκτης η βαθμολογία αλλάζει αφού αλλάζει και ο βαθμός δυσκολίας λήψης αποφάσεων τη στιγμή του αγώνα.

Τα επαγγελματικά αθλήματα αποτελούν ουσιαστικά μεγάλες επιχειρήσεις. Καθώς τα έσοδα από την παρουσία των φιλάθλων αποτελούν βασικό στοιχείο, ομάδες που αγωνίζονται στα playoffs και/ή κερδίζουν πρωταθλήματα, όχι μόνο αυξάνουν την παρουσία του κόσμου στο γήπεδο αλλά επίσης βρίσκουν επιπρόσθετα έσοδα από επικερδείς τηλεοπτικές εκπομπές. Το «κλειδί» είναι απλό, η νίκη. Σε έναν τόσο ανταγωνιστικό χώρο με μεγάλο αριθμό αποφάσεων που πρέπει να λαμβάνονται συνεχώς, γίνεται κατανοητό πως μόνο οι σωστές ενέργειες θα επιτρέψουν σε ένα σύλλογο να παραμένει ανταγωνιστικός στο υψηλότερο επίπεδο. Οι αποφάσεις αυτές προέρχονται από τα γεγονότα και από τα δεδομένα που συλλέγουμε. Πρέπει απλά να βρούμε τρόπους να ξεκλειδώσουμε τη γνώση που βρίσκεται παγιδευμένη στα δεδομένα και να τη χρησιμοποιήσουμε.

2.4 Κοινωνικές διαστάσεις

Οι επαγγελματικοί αθλητικοί οργανισμοί δεν μπορούν από μόνοι τους να οδηγηθούν στην επιτυχία. Οι φίλαθλοι (fans), οι πολίτες ενός κράτους, οι απλοί λάτρεις των σπορ και οι ειδικοί αναλυτές, όλοι έχουν μερίδιο στην επιτυχημένη πορεία μιας ομάδας.

Οι φίλαθλοι θεωρούνται ένα από τα σημαντικότερα κομμάτια για την επιτυχία μιας ομάδας. Συμβάλλουν στο να αυξήσει ο αθλητικός σύλλογος τα έσοδά του με τη συνεχή παρουσία τους στο γήπεδο, αγοράζοντας προϊόντα της ομάδας και παρακολουθώντας σε απευθείας σύνδεση μεταδόσεις, προωθώντας με αυτόν τον τρόπο την ομάδα σε ακόμα περισσότερους ανθρώπους. Πέρα όμως, από αυτές τις πολύ σημαντικές συνεισφορές στην οικονομική κατάσταση ενός συλλόγου, οι φίλαθλοι

διαδραματίζουν ακόμη ένα σπουδαίο ρόλο. Την ικανότητα να παρακινούν την ομάδα τους και να τη βοηθούν να κερδίζει. Αυτό φαίνεται ξεκάθαρα από αυτό που ονομάζουμε «εντός έδρας πλεονέκτημα», όπου οι φίλαθλοι μπορούν να ασκήσουν ένα ορισμένο βαθμό επιρροής στο αποτέλεσμα των αγώνων. Το να κρατήσει μια ομάδα τους φιλάθλους της στις κερκίδες δεν είναι και τόσο εύκολη διαδικασία. Αν η ομάδα κερδίζει, οι φίλαθλοι αυξάνονται. Αυτοί με τη σειρά τους παρακινούν την ομάδα να κερδίζει, με αποτέλεσμα ακόμη περισσότεροι να έρχονται κοντά στο σύλλογο. Το μυστικό είναι να κερδίζει η ομάδα.

Τα σπορ πέρα από το κομμάτι των φιλάθλων, μπορούν να αποτελέσουν και πηγή εθνικής υπερηφάνιας. Μέσα από αγώνες σε παγκόσμιες διοργανώσεις όπως είναι το Παγκόσμιο Κύπελο και οι Ολυμπιακοί Αγώνες, οι εθνικές ομάδες μπορούν να «αιχμαλωτίσουν» τη ψυχή ενός κράτους και να δημιουργήσουν ένα αίσθημα εθνικής υπερηφάνιας. Συνέπεια αυτού είναι η διαφήμιση της χώρας στο εξωτερικό με αποτελέσμα την αύξηση του τουρισμού και κατ' επέκταση τη στήριξη της οικονομίας μιας χώρας. Βροντερό παράδειγμα αποτελεί το άθλημα του ποδοσφαίρου (soccer). Το Παγκόσμιο Κύπελο διοργανώνεται κάθε τέσσερα χρόνια και προσελκύει αναμφισβήτητα τους πιο παθιασμένους φιλάθλους. Είναι αυτοί που θα κυματίσουν την εθνική τους σημαία, θα τραγουδήσουν τον εθνικό ύμνο, θα βάψουν τα πρόσωπά τους με τα χρώματα της χώρας τους, δημιουργώντας ένα κλίμα εθνικού ενθουσιασμού. Ωστόσο, αυτός ο ενθουσιασμός μερικές φορές μπορεί να οδηγήσει σε φαινόμενα βίας. Καβγάδες μεταξύ αντίπαλων οπαδών γίνονται όλο και συνιθέστεροι. Σε ένα δυσάρεστο γεγονός κατά τη διάρκεια αγώνα του Παγκοσμίου Κυπέλου το 1994 μεταξύ Κολομβίας και Ηνωμένων Πολιτειών, ο Κολομβιανός ποδοσφαιριστής Andres Escobar, έβαλε κατά λάθος αυτογκόλ με αποτέλεσμα η ομάδα του να χάσει τον αγώνα. Ο Escobar, ένας από τους δημοφιλέστερους παίκτες της ομάδας του, δολοφονήθηκε ως αντίποινα για το αυτογκόλ που πέτυχε και «καταδίκασε» τη χώρα του σε αποκλεισμό από τη συνέχεια του τουρνουά. Ωστόσο, η βία είναι περισσότερο η εξαίρεση παρά ο κανόνας. Άλλα παραδείγματα εθνικής υπερηφάνιας έχουν να κάνουν με ασυνήθιστες συμμετοχές και αουτσάιντερ. Η εθνική ομάδα ελκήθρου της Τζαμάικα είναι ίσως το αντιπροσωπευτικότερο παράδειγμα. Ενώ το θερμό κλίμα της χώρας δεν είναι και το ιδανικότερο για ένα χειμερινό άθλημα όπως είναι οι αγώνες ελκήθρου, η εθνική ομάδα κατάφερε να προκριθεί στους χειμερινούς Ολυμπιακούς Αγώνες του 1988 στο Calgary. Η ομάδα βρισκόταν στη ζώνη των μεταλλίων, μέχρι τη στιγμή που το έλκηθρό τους έσπασε, αναγκάζοντας τους τέσσερις αθλητές να το κουβαλήσουν στα χέρια τους μέχρι τον τερματισμό. Η αποφασιστικότητά τους κέρδισε το σεβασμό των αντίπαλων ομάδων και κέντρισε την προσοχή της παγκόσμιας αθλητικής κοινότητας κι όχι μόνο. Αργότερα τερμάτισαν στη 14^{η} θέση στους Ολυμπιακούς το 1992 ενώ κέρδισαν χρυσό μετάλλιο σε παγκόσμιο πρωτάθλημα ελκήθρου το 2000.

Ένα άλλο κομμάτι που έχει σημαντικό μερίδιο στην επιτυχία του franchise των σπορ είναι η ενασχόληση των φιλάθλων με τις «φανταστικές» ομάδες. Τα «φανταστικά» σπορ είναι στην ουσία ιδιωτικά παιχνίδια προσωμοίωσης τα οποία χρησιμοποιούν πραγματικά στατιστικά παικτών. Οι «φανταστικές» ομάδες οργανώνονται σε

πρωταθλήματα με τους ιδιοκτήτες να προσπαθούν να μιμηθούν τη λειτουργία τους σαν να ήταν πραγματικότητα. Οι ιδιοκτήτες επιλέγουν ή ανταλλάσσουν παίκτες ανάλογα με τις προσδοκίες τους για το πόσο καλά ένας παίκτης θα αποδώσει. Στη συνέχεια συλλέγονται τα στατιστικά των παικτών από κάθε αγώνα ξεχωριστά και μεταφράζονται σε πόντους. Οι ομάδες στο τέλος κατατάσσονται ανάλογα με τους συνολικούς πόντους που έχουν συγκεντρώσει. Το κλειδί για να κερδίσει κάποιος στα «φανταστικά» σπορ είναι να διατηρεί στην ομάδα του εκείνους τους παίκτες που με την απόδοσή τους μεγιστοποιούν τους πόντους. Αυτό εξελίχθηκε στο ιδανικότερο περιβάλλον για την εφαρμογή των sabermetrics. Αρκετοί ιδιοκτήτες υϊοθέτησαν τα sabermetrics ως μέθοδο για την επιλογή των παικτών στην ομάδα τους και γνώρισαν την επιτυχία σε σχέση με αυτούς που συμβουλεύονταν μόνο τα στατιστικά. Θα χρειαστούν αρκετά χρόνια προτού τα sabermetrics θα μεταπηδήσουν στα επαγγελματικά αθλήματα.

2.5 Ανίχνευση της απάτης

Η απάτη στο χώρο των σπορ πάντοτε αποτελούσε πρόβλημα. Αρκετά σκάνδαλα κατά καιρούς έχουν λάβει μέρος όπως για παράδειγμα το 1919, όταν οκτώ παίκτες της ομάδας baseball Chicago White Sox έχασαν τη σειρά των τελικών με αποτέλεσμα η ομοσπονδία να αποφασίσει τον ισόβιο αποκλεισμό τους από το παιχνίδι. Άλλο παράδειγμα αποτελεί η περίπτωση του Pete Rose, ο οποίος πιάστηκε να στοιχηματίζει στους Cincinnati Reds, όντας προπονητής της ομάδας τη συγκεκριμένη περίοδο. Ένα πρόσφατο παράδειγμα απάτης έρχεται από τη γειτονική Ιταλία, όπου μια από τις σημαντικότερες και πιο δημοφιλείς ομάδες ποδοσφαίρου της, η θρυλική Juventus, ενεπλάκη σε σκάνδαλο παράνομων αγώνων με αποτέλεσμα τον υποβιβασμό της στη δεύτερη κατηγορία. Αλλά ας μη πηγαίνουμε μακριά. Στη χώρα μας είναι ουκ ολίγα τα σκάνδαλα που έχουν ξεσπάσει τα τελευταία χρόνια τόσο στο άθλημα του ποδοσφαίρου όσο και του μπάσκετ, με τα εμπλεκόμενα πρόσωπα να είναι από παίκτες, παράγοντες, διαιτητές και προέδρους συλλόγων μέχρι ακόμα και πολιτικούς. Εκτός όμως από τα παραπάνω, ακόμα πιο διαδεδομένες είναι οι περιπτώσεις χρήσης απαγορευμένων ουσιών από τους παίκτες με στόχο τη βελτίωση των επιδόσεών τους. Όταν η δόλια δραστηριότητα εμφανίζεται στα σπορ, γενικά εμπίπτει σε μια από τις ακόλουθες τρεις κατηγορίες: κακή απόδοση των παικτών, ασυνήθιστα σφυρίγματα και αποφάσεις από τους διαιτητές και μονόπλευρος στοιχηματισμός.

Η κακή απόδοση των παικτών αποτελεί έναν τρόπο με τον οποίο μπορεί να τεθεί σε κίνδυνο η ακεραιότητα ενός αγώνα. Αυτό περιλαμβάνει έναν ή περισσότερους παίκτες οι οποίοι σκόπιμα δεν αποδίδουν τα μέγιστα με στόχο να επηρεάσουν το στοιχηματικό όριο του αγώνα. Πριν από τη διεξαγωγή του αγώνα, οι στοιχηματκοί πράκτορες καθορίζουν ένα όριο σύμφωνα με το οποίο συγκεντρώνεται ίσος αριθμός χρημάτων από τα στοιχήματα και για τις δύο ομάδες, με τη χαμένη πλευρά να πληρώνει τους νικητές εκτός από την προμήθεια των πρακτόρων. Εάν τα όρια είναι ανισόρροπα, οι

πράκτορες θεωρούνται υπεύθυνοι για τη διαφορά με αποτέλεσμα να χάσουν χρήματα, την επιχείρησή τους ή και τα δύο. Αν μια ομάδα τώρα είναι το απόλυτο φαβορί, τότε το όριο του στοιχήματος γίνεται πιο σαφές, με τη συγκεκριμένη ομάδα να πρέπει να πετύχει μεγαλύτερη νίκη για να κερδίσει το στοίχημα. Η μη επίτευξη πόντων σε έναν αγώνα είναι η προσπάθεια ενός παίκτη να επηρρεάσει την έκβαση του παιχνιδιού αποφεύγοντας να πλησιάσει τα στοιχηματικά όρια (παράδειγμα αυτής της περίπτωσης είναι τα γνωστά στο χώρο του στοιχήματος handicaps που η απλή νίκη ή ήττα μιας ομάδας δεν κερδίζει το στοίχημα). Μια πρόσφατη έρευνα στους αγώνες κολλεγιακού πρωταθλήματος καλαθοσφαίρησης των Ηνωμένων Πολιτειών (NCAA), έδειξε πως στο 1% των παιχνιδιών υπήρξε κάποια μορφή σκόπιμης μη επίτευξης πόντων. Η ανακάλυψη τέτοιων περιστατικών είναι εξαιρετικά δύσκολη, από τη στιγμή που τυπικά δεν υπάρχει αύξουσα συσχέτιση μεταξύ των στοιχηματικών αγορών από αγώνα σε αγώνα.

Οι διαιτητές (referees), είναι ένα εξίσου σημαντικό κομμάτι, καθώς με τα σφυρίγματα και τις αποφάσεις τους μπορούν να χειραγωγήσουν την έκβαση ενός αγώνα. Όμοια με την προηγούμενη περίπτωση, οι διαιτητές με τον τρόπο τους έχουν την ικανότητα να επηρρεάζουν τα στοιχηματικά όρια, κάνοντας εύκολο ή δύσκολο το παιχνίδι για μια εκ των δύο ομάδων. Ένα πρόσφατο παράδειγμα ήταν το καλοκαίρι του 2007, όταν ο διαιτητής του NBA (National Basketball Association) Tim Donaghy, ερευνήθηκε και καταδικάστηκε ότι έθετε σε κίνδυνο παιχνίδια με σκοπό να πληρώσει προσωπικά χρέη προερχόμενα από το τζόγο.

Τόσο η σκόπιμη μη επίτευξη πόντων από τους παίκτες όσο και οι αμφισβητήσιμες αποφάσεις των διαιτητών έχουν ένα και μοναδικό σκοπό. Το χρήμα. Έτσι ο μονόπλευρος στοιχηματισμός μπορεί να αποτελέσει ένα δείκτη για τους «ύποπτους» αγώνες. Η περίπτωση αυτή μπορεί να περιλαμβάνει είτε υπερβολική ποσότητα στοιχημάτων σε σχέση με το αναμενόμενο είτε μεγάλες ποσότητες χρημάτων εναντίον του φαβορί. Σε ένα συγκεκριμένο παράδειγμα, ένας εθισμένος με το στοίχημα παίκτης από το Detroit των Ηνωμένων Πολιτειών, πόνταρε επαναλαμβανόμενα στοιχήματα εναντίον του Πανεπιστημίου του Τολέδο στον αγώνα του κολεγιακού πρωταθλήματος φουτμπολ με αντίπαλο την ομάδα Temple. Ένα από τα στοιχήματα ήταν 20.000$, αριθμός περίπου τέσσερις φορές μεγαλύτερος από τα συνηθισμένα τυπικά ποσά του συγκεκριμένου αγώνα. Ο παίκτης εν τέλει σωστά είχε επιλέξει να στοιχηματίσει εναντίον του Τολέδο. Αυτό όμως κίνησε τις υποψίες των στοιχηματικών πρακτόρων. Στον επόμενο αγώνα περισσότερα παράτυπα στοιχήματα έκαναν την εμφάνισή τους εναντίον της ομάδας του Τολέδο με αποτέλεσμα ορισμένοι πράκτορες να απαγορεύσουν το ποντάρισμα στο συγκεκριμένο παιχνίδι απο την επιχείρησή τους. Οι πράκτορες βγάζουν τα χρήματά τους από την ομοιόμορφη κατανομή των στοιχημάτων και αφού πληρώσουν πρώτα τους κερδισμένους κρατούν μια προμήθεια. Όταν οι αγώνες είναι «ύποπτοι» και τα στοιχήματα άνισα τότε προφανώς χάνουν χρήματα. Επομένως, η ακεραιότητα ενός αγώνα εξαρτάται από το κομμάτι των στοιχημάτων και το κατά πόσο αυτά είναι νόμιμα.

Μία από τις οργανώσεις που δραστηριοποιείται ενεργά στην καταπολέμηση της απάτης (νοθείας) στο χώρο των επαγγελματικών σπορ είναι όπως έχουμε προαναφέρει η Las Vegas Sports Consultants Inc (LVSC). Ο συγκεκριμένος οργανισμός καθορίζει τα στοιχηματικά όρια σε ποσοστό 90% στα καζίνο του Las Vegas. Η LVSC στατιστικά αναλύει τόσο τα όρια των στοιχημάτων όσο και την αποδόση των παικτών, αναζητώντας οποιαδήποτε ασυνήθιστη δραστηριότητα. Όσον αφορά την απόδοση των αθλητών ερευνά μεταβλητές που την επηρρεάζουν, όπως οι επιδόσεις τους σε προηγούμενους αγώνες, η τύχη καθώς και η υγεία τους. Συμψηφίζοντας τα παραπάνω θέτει τα όρια των στοιχημάτων έτσι ώστε να μην υπάρχουν παρατυπίες. Τα διάφορα επαγγελματικά πρωταθλήματα είναι πρόθυμα να χρησιμοποιήσουν τις υπηρεσίες της LVSC προκειμένου να εξασφαλίσουν την εντιμότητα των αγώνων. Βασικοί πελάτες της LVSC είναι οι αμερικανικές ομοσπονδίες καλαθοσφαίρησης (NBA), φουτμπολ (NFL) και χοκεϊ (NHL)

2.6 Συμπεράσματα

Τα εργαλεία Data Mining για παίκτες, αθλητικούς κατασκόπους, επαγγελματικούς συλλόγους κι όχι μόνο γίνονται συνηθέστερα και μερικές φορές θεωρούνται αναγκαία για την επίτευξη των στόχων τους. Τα εργαλεία αυτά ποικίλλουν σε μέγεθος, πεδίο δράσης και επίπεδο λεπτομέρειας. Από απλές περιγραφές της μη φυσιολογικής έκβασης ενός αγώνα μέχρι οπτικοποιήσεις που επιτρέπουν στους ειδικούς να επεξεργάζονται εύκολα και γρήγορα πολύπλοκα δεδομένα, τα εργαλεία αυτά προάγουν νέους τρόπους και ιδέες κατανόησης των επαγγελματικών αθλημάτων. Εκτός από την ικανότητά τους να παρέχουν ανταγωνιστικό πλεονέκτημα, οι τεχνικές αυτές συμβάλλουν στη διατήρηση της ακεραιότητας των αγώνων ελέγχοντας συνεχώς τα δεδομένα τους για «ύποπτα» πρότυπα.

3. Πηγές Δεδομένων

3.1 Εισαγωγή

Τα δεδομένα, η «πηγή ζωής» της αθλητικής ανάλυσης, είχαν τη δική τους εξέλιξη στη διάρκεια των χρόνων. Αρχικά τα δεδομένα εξετάζονταν ως μια απλή καταγραφή των γεγονότων ενός αγώνα τα οποία τα κρατούσαν στο αρχείο τους οι αθλητικοί σύλλογοι ή οι αντίστοιχες ομοσπονδίες των πρωταθλημάτων για ιστορικούς λόγους. Τα δεδομένα αυτά σύντομα τροποποιήθηκαν σε συνοπτικές φόρμες έτσι ώστε να προσφέρουν μια σύντομη ανακεφαλαίωση των γεγονότων του αγώνα, διαβάζοντας απλά μια εφημερίδα. Με την έκδοση των δεδομένων τα επόμενα χρόνια, ολοένα και περισσότεροι ενδιαφερόμενοι μπορούσαν να έχουν πρόσβαση στα γεγονότα των αγώνων. Τότε τα δεδομένα ξεκίνησαν να διευρύνονται, με αρκετές συγκρίσεις να γίνονται σε διάφορες κατηγορίες στατιστικών. Η δραστηριότητα αυτή οδήγησε σε ξεκαθάρισμα, καθώς οι νέες ιδέες που συνεχώς παρουσιάζονταν υπαγόρευαν ποια δεδομένα πρέπει να κρατήσουμε και ποια όχι. Στη συνέχεια με την έλευση του διαδικτύου η προσβασιμότητα στα δεδομένα είναι θέμα μερικών δευτερολέπτων, όπου τα δεδομένα που σχετίζονται με τα σπορ μπορούν να βρεθούν εύκολα και γρήγορα, τις περισσότερες φορές σε ερευνήσιμη μορφή.

Τα αθλητικά δεδομένα μπορεί να προέρχονται από πολλές διαφορετικές πηγές. Η πιο τυπική από αυτές είναι από ένα στατιστικό αναλυτή ο οποίος προσλαμβάνεται από έναν επαγγελματικό σύλλογο προκειμένου να καταγράφουν τόσο τα συνολικά όσο και τα ατομικά στατιστικά της ομάδας. Επειδή αρκετοί αθλητικοί οργανισμοί κρατούν τα δεδομένα για τον εαυτό τους, έχουν αναπτυχθεί επαγγελματικές κοινότητες και επιχειρήσεις ειδικών εφαρμογών, οι οποίες καλύπτουν το κενό, παρέχοντας πηγές δεδομένων στους λάτρεις των σπορ και μερικές φορές ακόμα και στους ίδιους τους συλλόγους. Αυτό με τη σειρά του οδήγησε στην εξέλιξη αρκετών αθλητικών παραγώγων, όπως για παράδειγμα τα συστήματα παρακολούθησης της απόδοσης, τα οποία βασίζονται σε πραγματικά sports data. Επαγγελματικές κοινότητες, οργανισμοί σχετικοί με τα σπορ και ειδικού ενδιαφέροντος πηγές συμπληρώνουν τα κενά από τα αθλητικά δεδομένα και παρέχουν σημαντικές πληροφορίες που αλλιώς θα ήταν πολύ δύσκολο να ανακαλύψουμε.

3.2 Εύρεση πηγών για το άθλημα της καλαθοσφαίρησης

Οι επαγγελματικές κοινότητες προσφέρουν δεδομένα σχετικά με τα σπορ και λειτουργούν ως forum, όπου οι ενδιαφερόμενοι μοιράζονται και εξερευνούν τις γνώσεις τους. Πολλές από αυτές τις οργανώσεις συλλέγουν, αξιολογούν, αρχειοθετούν και διασπείρουν αθλητικά δεδομένα για τα μέλη τους διατηρώντας παράλληλα ενημερωτικά δελτία και περιοδικά. Ωστόσο, η κυριότερη δραστηριότητά τους περιστρέφεται γύρω από την ανακάλυψη και τη διάδοση γνώσεων στο χώρο των σπορ.

3.3 Οργάνωση για την έρευνα της επαγγελματικής καλαθοσφαίρησης (association for professional basketball research)

Όσον αφορά την καλαθοσφαίρηση, η συγκεκριμένη οργάνωση (APBR) ιδρύθηκε το 1997, με σκοπό να πρωωθήσει την ιστορία του αθλήματος καθώς και να αναλύσει αντικειμενικά τα στατιστικά στοιχεία του παιχνιδιού. Ενώ η έρευνα της APBR επικεντρώνεται κυρίως σε στατιστικά που προέρχονται από το χώρο του NBA (National Basketball Association), περιλαμβάνει επίσης και δεδομένα και από άλλα ανταγωνιστικά πρωταθλήματα μερικά από τα οποία αυτή τη στιγμή δεν υφίστανται. Όμοια με τα sabermetrics, που όπως έχουμε προαναφέρει εφαρμόζονται στο άθλημα του baseball, η APBR ανέπτυξε τα APBRmetrics τα οποία χρησιμοποιούνται για να δημιουργήσουν καλύτερες μετρήσεις και να παραθέσουν στατιστικά μέτρα σύγκρισης για ανάλογους σκοπούς. Τα APBRmetrics ουσιαστικά βασίστηκαν στις βασικές αρχές των sabermetrics. Αρχικά η APBR χρησιμοποιούσε γραμμικές μεθόδους μετρήσεων, μελετώντας κρίσιμες στατιστικές παραμέτρους που επηρρέαζαν την απόδοση. Ωστόσο προς το τέλος της δεκαετίας του 1990 η APBR και ο Dean Oliver κυρίως, ξεκίνησαν την έρευνα γύρω από μια σημαντική παράμετρο στο άθλημα της καλαθοσφαίρησης, όπως είναι η κατοχή, καθώς επίσης και για τα συνολικά στατιστικά της ομάδας. Με τον τρόπο αυτό η APBR εξελίχθηκε ακόμη περισσότερο και από τότε ορίστηκε ως η πρώτη πηγή εύρεσης στατιστικών δεδομένων για το άθλημα του basketball.

3.4 Οργανισμοί που σχετίζονται με τα σπορ (sport-related associations)

Επιπρόσθετα με τις επαγγελματικές κοινότητες, οι συγκεκριμένοι οργανισμοί επίσης συλλέγουν και διασπείρουν πληροφορίες στα μέλη τους. Διαφέρουν όμως από τις παραπάνω, καθώς δεν ασχολούνται με κάποιο συγκεκριμένο άθλημα, έχοντας ως μοναδικό στόχο να βελτιώνουν συνεχώς τις ήδη υπάρχουσες τεχνικές καθώς και να αρχειοθετούν τα δεδομένα για τις επόμενες γενιές.

3.4.1 Παγκόσμιος Οργανισμός της Επιστήμης των Υπολογιστών στα Σπορ (International Association on Computer Science in Sports)

Ο συγκεκριμένος οργανισμός, γνωστός κι ως IACSS, ιδρύθηκε το 1997 με σκοπό να βελτιώσει τη συνεργασία μεταξύ ερευνητών διεθνώς οι οποίοι ενδιαφέρονται στην εφαρμογή μεθόδων και τεχνολογιών της Επιστήμης των Υπολογιστών σε προκλήσεις προερχόμενες από το χώρο των σπορ. Ο IACSS επικεντρώνεται στη διάδοση της έρευνας των μελών του μέσα από ενημερωτικά δελτία, περιοδικά και εξαμηνιαία συνέδρια.

3.4.2 Παγκόσμιος Οργανισμός Πληροφοριών για τα Σπορ (International Association for Sports Information)

Ο Παγκόσμιος Οργανισμός Πληροφοριών για τα Σπορ (IASI) ιδρύθηκε το 1960 με στόχο την τυποποίηση και την αρχειοθέτηση των παγκόσμιων βιβλιοθηκών πληροφορίας γύρω από τα σπορ. Αποτελείται από ειδικούς στο χώρο των σπορ καθώς και από ανθρώπους που γνωρίζουν πως να αρχειοθετούν και να κατανέμουν τη γνώση. Η διάδοση των πληροφοριών του Οργανισμού γίνεται με την ανά τριετή έκδοση ενός ενημερωτικού δελτίου καθώς και με τη σύγκλιση ενός Παγκόσμιου Συνεδρίου κάθε τέσσερα χρόνια.

3.5 Πηγές ειδικού ενδιαφέροντος

Επιπρόσθετα με τους οργανισμούς και τις κοινότητες που σχετίζονται με το χώρο των σπορ, υπάρχουν κι άλλοι συνήθως εμπορικοί, οι οποίοι συλλέγουν και αναλύουν ειδικά στατιστικά στοιχεία. Συχνά οι πηγές αυτές προσφέρουν στους ενδιαφερόμενους, παραδοσιακά στατιστικά και επαυξημένα δεδομένα στη μορφή βιογραφικών στοιχείων, ρεκόρ και βραβείων.

Καλαθοσφαίρηση (Basketball)

Η τεχνολογία Synergy Sports επαναπροσδιόρισε την εύκολη πρόσβαση σε δεδομένα σχετικά με το άθλημα του basketball προερχόμενα από ζωντανές τηλεοπτικές μεταδόσεις. Το προϊόν της Synergy, Synergy Online, επιτρέπει στους χρήστες να εξετάζουν κάθε φάση του αγώνα, με συνεχείς ενημερώσεις των στατιστικών των παικτών και απευθείας μεταδόσεις στους υπολογιστές των χρηστών ή ακόμα και σε φορητές συσκευές (smartphones, tablets). Η Synergy τρέχει επίσης τα δεδομένα πίσω από δημοφιλή ηλεκτρονικά παιχνίδια όπως είναι το NBA Live 09 και NBA Live 10.

Το χαρακτηριστικό προϊόν της εταιρίας, Digital DNA, επεξεργάζεται περίπου χίλια χαρακτηριστικά και τάσεις ενός παίκτη, με σκοπό την καλύτερη μοντελοποίηση της αναμενόμενης απόδοσής τους, φέρνοντας με αυτό τον τρόπο το ρεαλισμό μέσα από την εμπειρία του παιχνιδιού. Η τεχνολογία αυτή μπορεί να χρησιμοποιηθεί και από τους προπονητές του NBA έτσι ώστε να αναγνωρίσουν τα ευνοϊκότερα μαρκαρίσματα καθώς επίσης και να βελτιώσουν τον τρόπο με τον οποίο διαχειρίζονται τι αλλαγές τους κατά τη διάρκεια του αγώνα. Το 82games.com είναι επίσης μια παρόμοια πηγή που χρησιμοποιούν φίλαθλοι, προπονητές και μέσα ενημέρωσης για την εύρεση δεδομένων που σχετίζονται με το basketball.

3.6 Ειδικές ενδιαφέρουσες πηγές (web sports data extraction and visualization)

Όπως έχουμε προαναφέρει, καθώς τα sabermetrics ξύπνησαν την επιθυμία του κόσμου, που ασχολείται με τα σπορ, για ακόμη περισσότερα δεδομένα και κατά συνέπεια νέους τρόπους ανάλυσής τους, τα ίδια τα δεδομένα άρχισαν να εξελίσσονται. Αρχικά μεταφέρθηκαν από γραπτές σελίδες σε διαδικτυακές πηγές (online). Ενώ αυτό το βήμα θεωρήθηκε απλά μια αλλαγή, τα δεδομένα εξακολουθούσαν να είναι δεδομένα, σύντομα όμως μετατράπηκαν σε κάτι παραπάνω. Διαδικτυακές εφαρμογές ξεκίνησαν να ταξινομούν τα δεδομένα στις αντίστοιχες στατιστικές κατηγορίες παρέχοντας με αυτόν τον τρόπο χρήσιμες πληροφορίες προς τους ενδιαφερόμενους. Από εκεί κι έπειτα, οι εφαρμογές αυτές εξελίχθηκαν ακόμη περισσότερο δημιουργώντας γραφικές απεικονίσεις, μετατρέποντας έτσι τις πληροφορίες σε πραγματική γνώση.

Τα δεδομένα είναι άρρηκτα συνδεδεμένα με τη διαδικασία εξαγωγής συμπερασμάτων ως προς την απόδοση στο χώρο των αθλημάτων. Όσο πιο διαθέσιμα είναι τα δεδομένα τόσο ευκολότερα μπορούμε να μετράμε και να συγκρίνουμε αποδόσεις. Οι διαδικτυακές πηγές δεδομένων έχουν γίνει ακόμα πιο άφθονες λόγω της αύξησης χρήσης του Διαδικτύου (Internet) και της ανάγκης για άμεσα, ακριβή και εύκολα εργαλεία προς χρήση. Τα δεδομένα, οι εφαρμογές τους και οι ερωτήσεις που θέλουμε να απαντήσουμε συνεχώς εξελίσσονται.

3.6.1 Διαδικτυακές Πηγές

Οι διαδικτυακές πηγές εύρεσης δεδομένων σχετικών με τα σπορ έχουν αυξηθεί σε σχέση με τα προηγούμενα χρόνια. Πολλές από αυτές τις πηγές προέρχονται από τις αντίστοιχες ομοσπονδίες των αθλημάτων, υπάρχουν όμως και κάποιες ανεξάρτητες οι οποίες προσφέρουν κι αυτές με τη σειρά τους αξιόπιστες πληροφορίες. Η πρόσβαση στα δεδομένα μερικών από αυτές τις πηγές γίνεται μέσω εγγραφής, ενώ κάποιες άλλες

προσφέρουν τα δεδομένα τους δωρεάν και βγάζουν τα λεφτά τους από διαδικτυακές διαφημίσεις ή άλλες πηγές εσόδων. Με την εμφάνιση των πηγών αυτών, ο παλλαπλασιασμός των δεδομένων σε σχέση με την προηγούμενη δεκαετία ήταν ιλιγγιώδης και η ευκολία πρόσβασης σε αυτά θεαματική. Ακόμα καλύτερα, μερικά websites απεικονίζουν τα συνολικά δεδομένα σε διαγράμματα ή γραφήματα δίνοντας τη δυνατότητα στους χρήστες να τα ερευνούν εις βάθος και να ανακαλύπτουν νέα πρότυπα.

Καλαθοσφαίρηση (Basketball)

Το άθλημα του basketball διαθέτει πλούσιες σε δεδομένα διαδικτυακές πηγές. Ξεκινώντας από την ιστοσελίδα NBA.com, η οποία παρέχει ιστορικά στατιστικά στοιχεία αγώνων, διαγράμματα και αρκετό σχετικό υλικό και φτάνοντας μέχρι την ιστοσελίδα Basketball-reference.com με περιγραφές για ιδανικά μαρκαρίσματα μεταξύ παικτών καθώς κι άλλες διορατικές αναλύσεις. Όσον αφορά το ευρωπαϊκό basketball, κάθε χώρα διαθέτει εταιρίες οι οποίες καταγράφουν τα στατιστικά από τους αγώνες των εγχώριων πρωταθλημάτων και είναι διαθέσιμα προς το κοινό. Στην Ελλάδα υπάρχουν κάποιες εταιρίες που ασχολούνται με το αντικείμενο αυτό με την πιο γνωστή να είναι η GalanisSportsData. Η συγκεκριμένη εταιρία μέσα από την ιστοσελίδα της παρέχει ιστορικά στατιστικά των δέκα τελευταίων χρόνων για τις δύο πρώτες εθνικές κατηγορίες. Επίσης η ιστοσελίδα της Ευρωλίγκα (Euroleague) διαθέτει αντίστοιχες πληροφορίες για τις ομάδες που αγωνίζονται στη διοργάνωση από διάφορες χώρες της Ευρώπης.

NBA.com

Όπως προδίδει και το όνομά της, η συγκεκριμένη ιστοσελίδα αναφέρεται στο επαγγελματικό αμερικάνικο πρωτάθλημα basketball (NBA), παρέχοντας μια εκτενή σειρά από δεδομένα στους επισκέπτες της. Τα δεδομένα αυτά μπορεί να είναι από απλές στατιστικές κατηγορίες, τόσο ατομικές (κάθε παίκτη ξεχωριστά), όσο και ολόκληρης της ομάδας, μέχρι εξελιγμένοι δείκτες όπως το plus/minus και διαδραστικά γραφήματα που απεικονίζουν τον τρόπο με τον οποίο ένας παίκτης σουτάρει μέσα στο γήπεδο. Σαν παράδειγμα της στατιστικής κάλυψης, καθημερινά παραθέτει τους παίκτες οι οποίοι προπορεύονται σε κατηγορίες όπως είναι οι πόντοι (points), τα ριμπάουντ (rebounds), οι ασσίστς (assists), τα κλεψίματα (steals), τα κοψίματα (blocks), τα λάθη (turnovers) και τα τρίποντα (three pointers).

POINTS PER GAME

RUSSELL WESTBROOK

0 G / OKC

28.1

Regular Season
Points Per Game

Εικόνα 6: Πρώτος στην κατηγορία πόντων στο NBA τη σεζόν 2014-2015

ASSISTS PER GAME

CHRIS PAUL

3 G / LAC

10.2

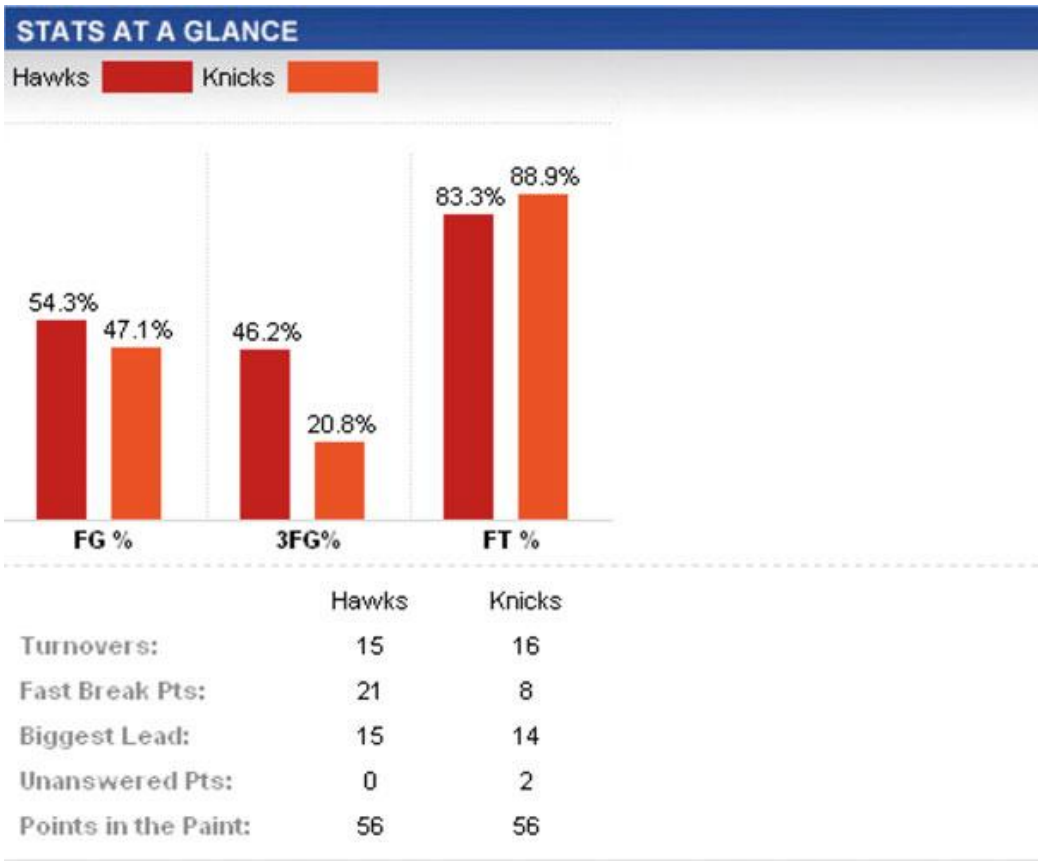
Regular Season
Assists Per Game

Εικόνα 7: Πρώτος στην κατηγορία ασσιστ στο NBA τη σεζόν 2014-2015

Συγκεκριμένες στατιστικές κατηγορίες που βασίζονται τόσο σε παίκτες όσο και σε ομάδες, παρέχουν χρήσιμες πληροφορίες ως προς την απόδοση των παικτών. Ένας τέτοιος δείκτης είναι ο plus/minus, ο οποίος αναγνωρίζει τον καλύτερο συνδυασμό πεντάδας που πετυχαίνει τους περισσότερους πόντους, ενώ συγχρόνως δέχεται από

τους αντιπάλους της όσο το δυνατόν λιγότερους. Ενδεικτικά αναφέρουμε για την αγωνιστική περίοδο 2009-2010 από την ομάδα των Dallas Mavericks η πεντάδα από τους Kidd, Dampier, Nowitzki, Marion και Terry σκόραρε 172 πόντους ενώ παράλληλα κράτησε τους αντιπάλους της στους 105. Ο δείκτης plus/minus γι' αυτή την πεντάδα ήταν +67.

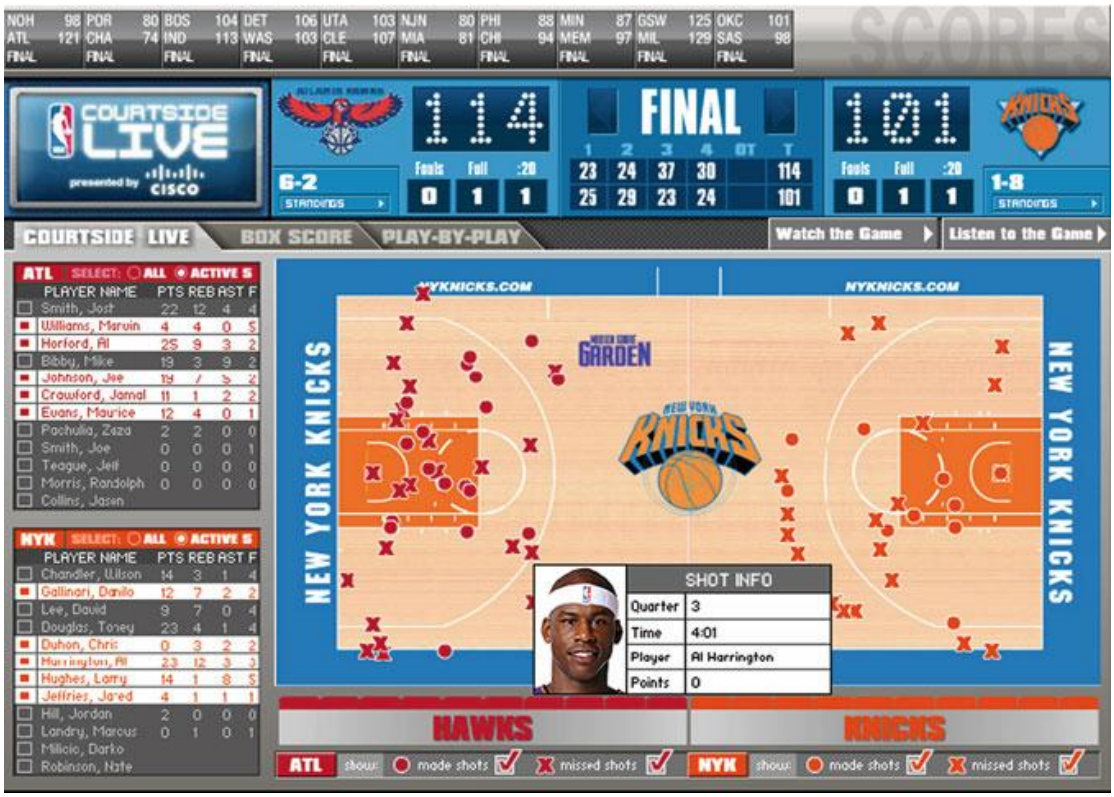
Οι γραφικές απεικονίσεις των δεδομένων είναι μακράν η πιο ενδιαφέρουσα εφαρμογή του NBA.com, καθώς ποικίλλουν σε μεγάλο βαθμό ως προς το περιεχόμενο και τη διαδραστικότητα. Είναι εύκολες στη χρήση και παρέχουν στους χρήστες γρήγορη και ξεκάθαρη πρόσβαση στις πληροφορίες. Για παράδειγμα, στη συγκεκριμένη ιστοσελίδα μπορούμε να βρούμε για κάθε ολοκληρωμένο αγώνα NBA έναν αναλυτικό πίνακα (boxscore) με όλες τις στατιστικές κατηγορίες (ατομικά και ομαδικά στατιστικά), ένα άρθρο που περιγράφει τον αγώνα από την αρχή μέχρι το τέλος (play by play) και ένα γράφημα στο οποίο απεικονίζονται τα ποσοστά ευστοχίας στα σουτ εντός παιδιάς (field goals), τις ελεύθερες βολές (free throws) και τα τρίποντα (three pointers).



Εικόνα 8: Πίνακας στατιστικών στοιχείων ενός αγώνα

Όπως φαίνεται οι Atlanta Hawks κυριάρχησαν των New York Knicks τόσο στα σουτ εντός παιδιάς με ποσοστό 54.3 έναντι 47.1%, όσο και στα σουτ τριών πόντων με ποσοστό 46.2 έναντι 20.8% των αντιπάλων τους. ωστόσο η ομάδα της Νέας Υόρκης είχε μεγαλύτερο ποσοστό στις ελεύθερες βολές.

Εκτός από γραφήματα όπως το παραπάνω, το NBA.com μας παρέχει και κάποιες περισσότερο διαδραστικές εφαρμογές όπως είναι το Courtside live.



Εικόνα 9: Courtside live

Μέσα από αυτό το γραφικό περιβάλλον ο χρήστης πληροφορείται τόσο για τα ατομικά όσο και για τα ομαδικά στατιστικά (επάνω και δεξια πλευρά), ενώ στο κέντρο της εφαρμογής απεικονίζεται ένα γήπεδο μπάσκετ δείχνοντας όλες τις προσπάθειες (attempts) για σουτ και από τις δύο ομάδες. Με κόκκινο χρώμα είναι οι προσπάθειες των Atlanta Hawks ενώ με πορτοκαλί των New York Knicks. Οι κουκίδες αντιπροσωπεύουν τις εύστοχες προσπάθειες ενώ το Χ τις άστοχες. Τα σημεία αυτά είναι διαδραστικά καθώς με ένα πάτημα πάνω τους εμφανίζεται ένα μικρότερο παράθυρο στο κάτω μέρος της οθόνης με περισσότερες πληροφορίες. Στο συγκεκριμένο σημείο που πατήσαμε βλέπουμε ότι ο παίκτης των New York Knicks All Harrington επιχείρησε ένα σουτ από εκεί, δεξιά μεριά του γηπέδου πίσω από τη γραμμή των τριών πόντων, κατά τη διάρκεια του τρίτου δεκαλέπτου με 4:01 να απομένουν για τη λήξη του και αστόχησε. Αν συνεχίσουμε με τον ίδιο τρόπο να μελετάμε κι άλλα σουτ που επιχείρησαν οι παίκτες της Νέας Υόρκης θα δούμε ότι ο All Harrington δοκίμασε αρκετές φορές έξω από τη γραμμή των τριών πόντων και αστόχησε σε όλες του τις προσπάθειες. Ένα εργαλείο σαν κι αυτό μπορεί εύκολα και γρήγορα να αναγνωρίζει από ποιες θέσεις του γηπέδου οι παίκτες είναι περισσότερο και λιγότερο εύστοχοι.

Αυτές ήταν μερικές από τις εφαρμογές που προσφέρει η ιστοσελίδα NBA.com. Οι ενδιαφερόμενοι μπορούν να την επισκεφθούν και να ανακαλύψουν τις δυνατότητές της μέσα από το τεράστιο πλήθος στατιστικών στοιχείων που διαθέτει κι όχι μόνο.

3.7 Συμπεράσματα

Η εισαγωγή των sabermetrics στην ανάλυση των δεδομένων πριν από μερικές δεκαετίες, οδήγησε σε μια έκρηξη εφαρμογής παρόμοιων τεχνικών και εργαλείων σχεδόν σε όλα τα σπορ. Τα περισσότερα αμερικάνικα αθλήματα άρχισαν να βλέπουν μια σταθερότητα στις μετρήσεις που χρησιμοποιούσαν, αποδεχόμενοι ότι τα εργαλεία αυτά αποτελούν απαραίτητο βοήθημα για τους προπονητές και επιπρόσθετο πλεονέκτημα για τους φιλάθλους. Είναι συνηθισμένο για τους αθλητικούς οργανισμούς και για τους φιλάθλους τους να ανατρέχουν σε αυτές τις πηγές σαν κομμάτι της παιδείας τους.

4. Έρευνα της Στατιστικής των Αθλημάτων

4.1 Εισαγωγή

Αφού συγκεντρώσουμε τα δεδομένα μας, τα επομένα βήματα περιλαμβάνουν μια διαδικασία ανακάλυψης γνώσης από αυτά. Πολλές διαφορετικές μορφές στατιστικών αναλύσεων μπορούν να εφαρμοσθούν τόσο σε αθλήματα πλούσια από δεδομένα, όπως είναι το basketball και το baseball, όσο και σε αθλήματα λιγότερο διαδεδομένα όπως είναι το curling. Καθώς οι τεχνικές και οι μετρήσεις αλλάζουν από άθλημα σε άθλημα, η καρδιά του ζητήματος, τα στατιστικά στοιχεία, είναι αναγνωρίσιμα σε όλα τα σπορ, ακόμα κι αν τα μέτρα τους δεν είναι συγκρίσιμα. Η μεθοδολογία πίσω από τα νούμερα παραμένει η ίδια σε όλα τα αθλήματα. Κάποιες αναλύσεις χρησιμοποιούνται για να μετρήσουν την απόδοση των παικτών, την ισορροπία μιας ομάδας, τις αδυναμίες των αντιπάλων ακόμα και για να προβλέψουν έναν ενδεχόμενο μελλοντικό τραυματισμό.

4.2 Στατιστικές των αθλημάτων (sports statistics)

Χιλιάδες στατιστικά στοιχεία από αγώνες όλων των αθλημάτων κρατώνται ως αρχεία τον τελευταίο αιώνα. Τα στατιστικά αυτά θεωρούνταν δεδομένα και πολύ σπάνια κάποιος προσπάθησε να τα αμφισβητήσει ή να ρωτήσει περαιτέρω γι αυτά. Ωστόσο, μερκοί στο χώρο των σπορ άρχισαν να αναρωτώνται αν όντως μετράμε αυτά που νομίζουμε ότι μετράμε. Πρωτοπόροι της στατιστικής ανάλυσης όπως ο Bill James (baseball) και ο Dean Oliver (basketball) όχι μόνο αναρωτήθηκαν για τα παραπάνω αλλά επίσης ξεκίνησαν να προτείνουν και νέες μεθόδους στατιστικών αναλύσεων.

4.2.1 Έμφυτα προβλήματα στη στατιστική των αθλημάτων

Το άθλημα του basketball, όπως και τα υπόλοιπα αθλήματα, αντιμετωπίζει κι αυτό με τη σειρά του κάποια προβλήματα ανακρίβειας σε μερικές στατιστικές κατηγορίες όπως για παράδειγμα το ποσοστό των σουτ εντός παιδιάς (field goals percentage) και τα ριμπάουντ (rebounds). Το ποσοστό των σουτ εντός παιδιάς είναι ο αριθμός των εύστοχων σουτ διαιρεμένος με το συνολικό αριθμό προσπαθειών (εύστοχων και άστοχων). Ένας παίκτης που έχει σκοράρει πολλούς πόντους ενώ συγχρόνως έχει χαμηλό ποσοστό στα σουτ εντός παιδιάς μπορεί να θεωρηθεί μη αποδοτικός. Ομοίως, τα ριμπάουντ, δηλαδή το πόσες φορές ένας παίκτης παίρνει στην κατοχή του τη μπάλα

ύστερα από μια άστοχη προσπάθεια, δε μας καταδεικνύουν αν η ομάδα θα σκοράρει κιόλας.

Το πρόβλημα με τις παραδοσιακές στατιστικές κατηγορίες εντοπίζεται στο τι σκοπεύει να μετρήσει κάθε κατηγορία. Συχνά τα δεδομένα συλλέγονται σε μορφές που δύσκολα μπορούν να μεταφραστούν σε πραγματική γνώση. Τα δεδομένα από μόνα τους δεν είναι λανθασμένα, όσο οι μέθοδοι που χρησιμοποιούνται για να συγκρίνουν τις αποδόσεις των παικτών. Αυτό μας οδηγεί στο να καταλάβουμε ότι πολλά προβλήματα δε λύνονται μόνο μέσα από τη μελέτη της στατιστικής. Η αμφισβήτηση της στατιστικής, ο πυλώνας των σύγχρονων αθλημάτων, έφερε στην επιφάνεια νέες τεχνικές και μετρήσεις οι οποίες αποτελούν πια, αναπόσπαστο κομμάτι όλων των σύγχρονων σπορ.

Dean Oliver

Όπως έχουμε προαναφέρει, ο Dean Oliver στην προσπάθειά του να ποσοτικοποιήσει τη συνεισφορά των παικτών στην έκβαση ενός αγώνα basketball, πρότεινε την εφαρμογή των APBRmetrics, κάτι αντίστοιχο των sabermetrics για το άθλημα του basketball. Επικεντρώθηκε στη σωστή χρήση της στατιστικής κατηγορίας που έχει να κάνει με την κατοχή, όπου η κατοχή ορίζεται ως ο συνολικός χρόνος (κατά τη διάρκεια ενός αγώνα) που μια ομάδα έχει τη μπάλα στην κατοχή της. Κομμάτι της έρευνας του Dean Oliver ήταν να αξιολογήσει την απόδοση μιας ομάδας ως προς το πόσους πόντους σκοράρει ή δέχεται από τους αντιπάλους της ανά 100 κατοχές (επιθέσεις). Το 2004, ο Dean Oliver προσλήφθηκε ως σύμβουλος από την ομάδα των Seatle Supersonics, καταφέρνοντας την επόμενη χρονιά η ομάδα του Seatle να πανηγυρίσει τον τίτλο στην περιφέρειά της.

Καλαθοσφαίρηση (Basketball)

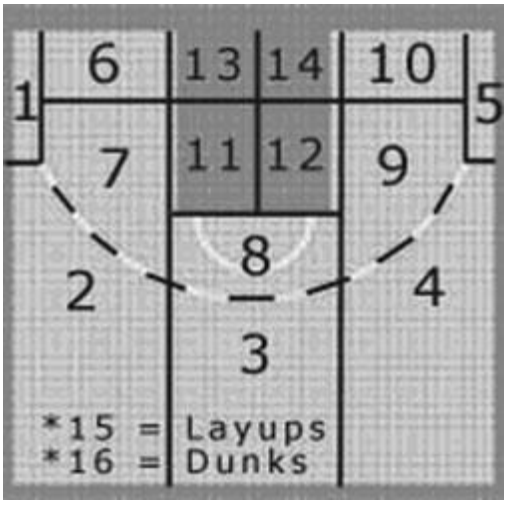
Το basketball είναι ένα άθλημα που αναμφισβήτητα διέπεται από πληθώρα δεδομένων και βάθος ως προς το κομμάτι της στατιστικής. Γι' αυτό άλλωστε και η εφαρμογή των APBRmetrics έγινε με τρόπο εντυπωσιακό. Τα APBRmetrics έχουν την ιδιαιτερότητα να εξετάζουν τα στατιστικά δεδομένα από την άποψη της ομάδας ως σύνολο, κι όχι μελετώντας την ατομική απόδοση. Ένα τέτοιο παράδειγμα αποτελεί η κατοχή, κατά τη διάρκεια του αγώνα, μιας ομάδας και κατά πόσο αποτελεσματική είναι η ομάδα αυτή στο να σκοράρει πόντους. Η κεντρική ιδέα είναι ότι αφού οι ομάδες πρέπει να λειτουργούν ως συνεκτικές μονάδες, πρέπει να αναλύονται κι ως ολότητες, χωρίς να προσπαθούμε να ποσοτικοποιήσουμε τη χημεία μιας ομάδας, ή το πόσο καλά αποδίδουν οι παίκτες σε ατομικό επίπεδο.

Για να εξηγήσουμε τα παραπάνω καλύτερα, οι παίκτες μπορεί να έχουν είτε θετική είτε αρνητική επίδραση στην απόδοση της ομάδας τους. Για παράδειγμα, την αγωνιστική περίοδο 2004-2005 ο Stephon Marbury (παίκτης των New York Knicks τότε), όσο αγωνιζόταν είχε αρνητική επίδραση στην ομάδα του -0,4 του πόντου (όσο

έπαιζε δηλαδή κατά τη διάρκεια του αγώνα η ομάδα του βρισκόταν πίσω στο σκορ με 0,4 πόντους κατά μέσο όρο). Με μια πρώτη ματιά το ατομικό αυτό στατιστικό στοιχείο υποδεικνύει ότι ο Marbury είχε αρνητική επίδραση στο παιχνίδι της ομάδας του. Ωστόσο, όταν ο συγκεκριμένος αθλητής βρισκόταν στον πάγκο, ανήμπορος να βοηθήσει, η ομάδα βρισκόταν πίσω στο σκορ με -12 πόντους κατά μέσο όρο. Οι 11,6 αυτοί πόντοι διαφοράς, όταν ο Marbury έπαιζε κι όταν ήταν εκτός παιχνιδιού, αποδεικνύουν ότι ο παίκτης βελτιώνει την επίδοση της ομάδας όταν βρίσκεται εντός παρκέ.

Shot Zones

Ένα γήπεδο basketball μπορεί να διαιρεθεί σε 16 ζώνες (περιοχές) από τις οποίες ένας παίκτης μπορεί να σουτάρει κατά τη διάρκεια του αγώνα. Αναλύοντας το ποσοστό ευστοχίας των παικτών από όλες τις θέσεις μπορούμε να αναπτύξουμε και τις αντίστοιχες στρατηγικές. Αναλυτικά, ως προς τον αμυντικό προσανατολισμό, οι προπονητές να προσπαθήσουν να περιορίσουν τους αντίπαλους παίκτες να σουτάρουν από τις θέσεις που έχουν μεγαλύτερα ποσοστά ευστοχίας, ενώ επιθετικά να προσπαθήσουν να κάνουν τους παίκτες τους, μέσα από κατάλληλα plays, να σουτάρουν από αυτές.



Εικόνα 10: Shot zones

Από μια shot zones ανάλυση την αγωνιστική περίοδο 2004-2005, στο NBA, βρέθηκε ότι από τις περιοχές των τριών πόντων 1 και 5 (κόρνερ), ο τότε παίκτης των Golden State Warriors Michael Dunleavy είχε ποστό ευστοχίας 57% από την αριστερή γωνία ενώ ο Cuttino Mobley των Sacramento Kings 60% από τη δεξιά. Την ίδια σεζόν επίσης, ο Shaquille O' Neal των Miami Heat επιχείρησε τις περισσότερες προσπάθειες στις περιοχές 13 και 14 με ποσοστά ευστοχίας 41 και 42% αντίστοιχα. Γνωρίζοντας,

λοιπόν, από ποια σημεία του γηπέδου οι παίκτες είναι περισσότερο εύστοχοι μπορούμε να εφαρμόσουμε και ανάλογες επιθετικές ή αμυντικές στρατηγικές.

Player Efficiency Rating (PER)

Το PER είναι μια σύνθετη εκτίμηση της αποτελεσματικότητας ενός παίκτη ανάλογα με το χρόνο τον οποίο αγωνίζεται. Ο τύπος σύμφωνα με τον οποίο υπολογίζεται περιλαμβάνει τόσο τις θετικές στατιστικές κατηγορίες, όπως για παράδειγμα οι πόντοι, τα ριμπάουντ, οι ασσίστ, τα κλεψίματα, τα εύστοχα σουτ, οι ελεύθερες βολές όσο και τις αρνητικές όπως τα άστοχα σουτ και τα λάθη. Με αυτόν τον τρόπο μπορούμε να έχουμε μια εικόνα της απόδοσης των παικτών τόσο σε έναν αγώνα όσο και για ολόκληρη την αγωνιστική περίοδο. Αναφορικά, αν και έχει αποσυρθεί αρκετά χρόνια από την ενεργό δράση, ο μοναδικός Michael Jordan παραμένει πρωτοπόρος σε αυτήν την κατηγορία μέχρι σήμερα.

Plus/Minus Rating

Μια άλλη μέθοδος υπολογισμού της απόδοσης των παικτών είναι μέσω του δείκτη Plus/Minus, όπου κάθε παίκτης αξιολογείται υπολογίζοντας τους πόντους που σκοράρει η ομάδα του μείον αυτούς που δέχεται όσο αυτός παίζει στον αγώνα. Ο υπολογισμός αυτός γίνεται για κάθε παίκτη όσο αυτός παίζει κι όσο βρίσκεται στον πάγκο. Έτσι η συνεισφορά του κάθε παίκτη μετριέται ως η διαφορά στους πόντους όσο παίζει και όσο όχι.

Για παράδειγμα, έστω ότι ο Kobe Bryant ξεκινάει βασικός στον αγώνα με το σκορ στο 0-0, και βγαίνει από τον αγώνα με το σκορ στο 90-80, υπέρ της ομάδας του, αυτοί οι +10 πόντοι διαφορά αποτελούν το plus/minus για τον Kobe στο συγκεκριμένο παιχνίδι. Αντίστροφα, αν ο ίδιος αθλητής εισέλθει στον αγώνα με το σκορ να βρίσκεται στο 85-80, υπέρ της ομάδας του, και βγει από το παιχνίδι με το σκορ 85-95, κατά της ομάδας του, αυτοί οι -15 πόντοι αποτελούν αντίστοιχα το δείκτη plus/minus.

Θετικές τιμές του δείκτη υποδεικνύουν θετική συνεισφορά του παίκτη στην παραγωγή πόντων για την ομάδα του, ενώ αρνητικές υποδηλώνουν επιβλαβή δραστηριότητα.

Μετρώντας τη συνεισφορά του παίκτη στις νίκες

Μια περαιτέρω μέτρηση ως προς την αξιολόγηση της συνεισφοράς ενός παίκτη που παίζει σε σχέση με έναν αναπληρωματικό, έρχεται να προσαρμόσει το δείκτη plus/minus στο ταλέντο των συμπαικτών. Ο λόγος είναι γιατί η συνεισφορά ενός παίκτη στη συνολική απόδοση της ομάδας του εξαρτάται σε μεγάλο βαθμό από την προσπάθεια όλων των παικτών.

Εκτιμώντας τις clutch αποδόσεις

Προκειμένου να απεικονίσουμε τη συνεισφορά ενός παίκτη, σε έναν αγώνα NBA, μελετούμε την απόδοσή του στις διάφορες στατιστικές κατηγορίες για χρόνο 40 λεπτών κι όχι 48 που είναι η διάρκεια του κανονικού αγώνα. Αυτό γίνεται διότι τα τελευταία λεπτά διαφέρουν από την υπόλοιπη διάρκεια του παιχνιδιού. Για παράδειγμα σε περιπτώσεις όπου μια ομάδα κερδίζει με αρκετούς πόντους διαφορά, η αντίπαλός της μπορεί να ξεκινήσει τη διαδικασία των φάουλ έτσι ώστε να σταματήσει το χρόνο και να κάνει κατοχή η ίδια. Η συμπεριφορά αυτή τείνει να σκεβρώσει τα στατιστικά στοιχεία που έχουν συλλεγεί μέχρι εκείνη τη στιγμή. Τα τελευταία 8 λεπτά και η όποια παράταση, αν χρειαστεί, με την προϋπόθεση ότι η διαφορά στο σκορ μεταξύ των ομάδων είναι περίπου στους 5 πόντους, αναφέρεται και ως clutch. Μερικοί παίκτες τείνουν να υπερτερούν σε σχέση με κάποιους άλλους κατ' αυτή τη διάρκεια του αγώνα, κάνοντας τους ειδικούς να μελετούν τις clutch αποδόσεις τους. Κάποιοι υποστηρίζουν ότι η συνεισφορά ενός παίκτη κατά τη διάρκεια των τελευταίων λεπτών είναι σημαντικότερη από τη συνεισφορά του στην κανονική διάρκεια γιατί στο τέλος κρίνεται το αποτέλεσμα. Κάποιοι θρύλοι του αθλήματος, όπως ο Michael Jordan, έχουν συνδέσει το όνομά τους με τις αποδόσεις τους στα τελευταία λεπτά. Η χρήση του δείκτη PER για τα τελευταία λεπτά βοηθά τους προπονητές να εξελίξουν το επιθετικό κομμάτι της ομάδας τους, αλλά και να μελετήσουν τρόπους αντιμετώπισης αντίπαλων παικτών που έχουν την ικανότητα να κρίνουν ένα αποτέλεσμα.

PIE (Player Impact Estimate)

Ο συγκεκριμένος δείκτης εκφράζει την ποσοστιαία επίδραση ενός παίκτη στη συνολική απόδοση της ομάδας του. Η φόρμουλα υπολογισμού του περιλαμβάνει τόσο τα ατομικά όσο και τα ομάδικά στατιστικά στοιχεία. Μπορεί να υπολογιστεί για κάθε αγώνα ξεχωριστά, προσφέροντας στους ειδικούς χρήσιμες πληροφορίες μετά το πέρας της αγωνιστικής περιόδου.

Αν και χρησιμοποιείται ευρέως από τους ειδικούς για παίκτες των ομάδων του NBA, παρέχει μια σαφή εικόνα, και γι' αυτό το λόγο χρησιμοποιήθηκε και στο πρακτικό κομμάτι της παρούσας εργασίας.

Υπάρχουν αρκέτες μελέτες γύρω από την εύρεση δεικτών οι οποίοι να απεικονίζουν με ακρίβεια την απόδοση των παικτών και τη συνεισφορά τους στη συνολική επίδοση της ομάδας στην οποία αγωνίζονται. Οι παραπάνω δείκτες αποτελούν μερικούς από τους πιο διαδεδομένους στο χώρο του αμερικάνικου basketball. Οι ειδικοί στις μέρες μας με τη βοήθεια της τεχνολογίας έχουν φτάσει σε σημείο να υπολογίζουν πολύ πιο εξειδικευμένους δείκτες μέτρησης της απόδοσης κι όχι μόνο, κάνοντας τη δουλειά των προπονητών και των συνεργατών τους ευκολότερη. Αρκεί κανείς να επισκεφθεί την ιστοσελίδα NBA.com και θα διαπιστώσει την πληθώρα των πληροφοριών που παρέχει.

4.2.2 Δείκτες Απόδοσης

Όσον αφορά την ελληνική πραγματικότητα, οι ειδικοί στη χώρας μας που ασχολούνται με το κομμάτι της στατιστικής προτείνουν κάποιους άλλους δείκτες απόδοσης. Ενδεικτικά, παρακάτω αναφέρουμε μερικούς από τους δείκτες που χρησιμοποιήθηκαν και στην παρούσα εργασία.

Tendex

Αποτελεί δείκτη που εκφράζει την απόδοση ενός παίκτη αναλογικά με το χρόνο συμμετοχής του στον αγώνα. Η φόρμουλα υπολογισμού του περιλαμβάνει τόσο θετικές στατιστικές κατηγορίες (πόντοι, ριμπάουντ, ασσίστ), όσο και αρνητικές (λάθη, χαμένα σουτ). Υπολογίζεται για παίκτες οι οποίοι έχουν τουλάχιστον 7 λεπτά συμμετοχής στον αγώνα. Όσο κοντινότερη στη μονάδα είναι η τιμή του tendex τόσο καλύτερη απόδοση είχε ο παίκτης. Αποτελεί ένα χρήσιμο δείκτη αξιολόγησης που χρησιμοποιείται εδώ και αρκετά χρόνια από τους ειδικούς για την εξαγωγή συμπερασμάτων.

Economy

Ο συγκεκριμένος δείκτης υπολογίζεται λαμβάνοντας υπόψη τρεις θετικές στατιστικές κατηγορίες (ασσίστ, κλεψίματα και κοψίματα) και μια αρνητική (λάθη). Χρησιμοποιείται διότι προσφέρει μια εικόνα του παίκτη χωρίς να συνυπολογίζει τους πόντους που σκοράρει, υποδηλώνοντας με αυτόν τον τρόπο πως ένας παίκτης μπορεί να είναι σημαντικός για την ομάδα του προσφέροντας σε κάποιους άλλους τομείς πέρα των πόντων.

Assist/Turnover

Δείκτης που αποτυπώνει με τη σειρά του πόσες τελικές πάσες (assists) «προσφέρει» ή καλύτερα στη μπασκετική γλώσσα «σερβίρει» ένας παίκτης για κάθε ένα λάθος (turnover) στο οποίο υποπίπτει. Δίνει χρήσιμα συμπεράσματα για όλους τους παίκτες, αλλά ίσως να χαρακτηρίζει λίγο περισσότερο τους play makers (point guards) μιας ομάδας από τους οποίους οι προπονητές περιμένουν όσο το δυνατόν περισσότερες τελικές πάσες και λιγότερα λάθη.

4.3 Συμπεράσματα

Με αρκετά σπορ να χρησιμοποιούν διάφορες μορφές στατιστικών στοιχείων με σκοπό να μετρήσουν την απόδοση των παικτών, τη συνεκτικότητα της ομάδας, την πρόληψη τραυματισμών αλλά και τις τάσεις των αντιπάλων τους, η στατιστική έχει αρχίσει να αποτελεί ένα πολύτιμο εργαλείο στα χέρια των ειδικών. Η χρήση αυτών των τεχνικών και η εξαγωγή νέων προτύπων από αυτά είναι το επόμενο βήμα για την πραγματική κατάκτηση της γνώσης.

5. Εργαλεία και Συστήματα για την Ανάλυση των Δεδομένων

5.1 Εισαγωγή

Οι αθλητικοί σύλλογοι στην εποχή μας επενδύουν μεγάλα ποσά χρημάτων επιδιώκοντας το καλύτερο δυνατό αποτελέσμα, έναντι των αντιπάλων τους, ανάλογα με τους στόχους που θέτουν. Έτσι λοιπόν, χρησιμοποιούν όλα τα μέσα που διαθέτουν, με την τεχνολογία να παίζει καθοριστικό ρόλο προς αυτή την κατεύθυνση. Τα τελευταία χρόνια έχουν αναπτυχθεί τόσο συστήματα Data Mining όσο και εργαλεία κατασκοπείας, τα οποία οι αθλητικές οργανώσεις χρησιμοποιούν προς όφελός τους. Η αγορά αναπτύσσεται συνεχώς, με τις εταιρίες κατασκευής να προμηθεύουν τις ομάδες με τα προϊόντα τους. Οι υπηρεσίες που προσφέρουν αυτά τα εργαλεία καλύπτουν ένα ευρύ φάσμα, όπως μελέτη της συμπεριφοράς και των τάσεων των παικτών, εις βάθος αναφορές κατασκοπείας (scouting reports) καθώς και αποκάλυψη περιπτώσεων απάτης στο περιβάλλον του αθλητισμού όπως έχει αναφερθεί σε προηγούμενο κεφάλαιο.

5.2 Εργαλεία sports data mining

Αρκετά είναι τα βοηθήματα που χρησιμοποιούνται προς αυτό το σκοπό με τη διαφορά πως δε βασίζονται στη βασική ιδέα του Data Mining. Τα συστήματα αυτά επικεντρώνονται σε μια εξειδικευμένη ανάλυση του βίντεο των αγώνων και την εξαγωγή συμπερασμάτων μέσα από κατάλληλες διαδικασίες. Η Virtual Gold αποτελεί παράδειγμα εταιρίας που παρέχει αυτή τη μοναδική υπηρεσία. Άλλες ξεχωριστές μέθοδοι περιλαμβάνουν απλές γραφικές αναλύσεις των στατιστικών στοιχείων, επιτρέποντας στους ειδικούς να ανακαλύπτουν πρότυπα μέσα από τα δεδομένα. Η οπτικοποίηση των πληροφοριών αποτελεί έναν αποτελεσματικό τρόπο διαχείρησης της γνώσης.

Advanced Scout

Το Advanced Scout αναπτύχθηκε από την εταιρία IBM στα μέσα της δεκαετίας του 1990 ως ένα ηλεκτρονικό πρόγραμμα εξόρυξης δεδομένων και διαχείρησης γνώσης. Σκοπός του είναι να ανακαλύψει καινούρια πρότυπα, αναλύοντας τα στατιστικά δεδομένα από αγώνες του NBA προσφέροντας επιπρόσθετη διορατικότητα στους προπονητές αλλά και σε άλλους επίσημους οργανισμούς στο χώρο της καλαθοσφαίρησης. Το εργαλείο αυτό συλλέγει εκτός από δομημένα στατιστικά στοιχεία κατά τη διάρκεια του αγώνα και μη δομημένο υλικό από τα πολυμέσα. Με

όλες τις ομάδες του NBA να έχουν πρόσβαση στο Advanced Scout, προπονητές και παίκτες μπορούν να το χρησιμοποιήσουν έτσι ώστε να προετοιμαστούν κατάλληλα απέναντι στους αντιπάλους τους για τους επόμενους αγώνες, μελετώντας τους μέσα από τα ιστορικά δεδομένα.

Από την άποψη των πολυμέσων το Advanced Scout συλλέγει ακατέργαστο υλικό από τους αγώνες, ελέγχει για τυχόν λάθη στο περιεχόμενο και τελικά κατανέμει το υλικό αυτό στις αντίστοιχες κατηγορίες όπως οι πόντοι, τα ριμπάουντ, τα κλεψίματα κ.λ.π. Η επεξεργασία και το στάδιο του ελέγχου για τυχόν λάθη περιλαμβάνουν μια σειρά διαδικασιών βασισμένες σε κανόνες με στόχο την επαλήθευση ως προς τη συνοχή και την ακρίβεια των δεδομένων.

Το Advanced Scout περιλαμβάνει ένα επιπλέον συστατικό διαχείρησης γνώσης, το Attribute Focusing, όπου ένα συγκεκριμένο χαρακτηριστικό αξιολογείται σε σχέση με τη συνολική κατανομή των δεδομένων με τα αποτελέσματα να παρατίθενται τόσο σε μορφή κειμένου όσο και σε γραφική περιγραφή των «ανώμαλων» υποσυνόλων. Τα υποσύνολα αυτά που παρουσιάζουν μια διαφορετική στατιστική κατανομή χρήζουν περισσότερης μελέτης από τους προπονητές ή τους παίκτες. Ας πάρουμε για παράδειγμα μια περιγραφή σε μορφή κειμένου που μας παρέχει το Advanced Scout.

Όταν ο παίκτης Mark Price αγωνιζόταν στη θέση του Point Guard, ο παίκτης Williams αστόχησε σε 0% (0) σε προσπάθειες για σουτ εντός παιδιάς και ευστόχησε σε 100% (4) προσπάθειες. Ο συνολικός αριθμός προσπαθειών ήταν 4. Αυτό είναι ένα διαφορετικό πρότυπο από το κανονικό που δείχνει ότι οι παίκτες της ομάδας συνολικά αστόχησαν σε προσπάθειες για σουτ εντός παιδιάς με ποσοστό 50,7% και ευστόχησαν με ποσοστό 49,3%.

Η περιγραφή αυτή απεικονίζει μια κατανοητή ανάλυση της «ανώμαλης» συμπεριφοράς του παίκτη Williams όσο ο παίκτης Mark Price αγωνιζόταν στη θέση του Point Guard. Όταν οι προπονητές ή οι παίκτες λάβουν αυτή την πληροφορία, είναι στην κρίση τους να την ερμηνεύσουν κατάλληλα και να την αξιοποιήσουν αναλόγως. Η ερμηνεία του συγκεκριμένου παραδείγματος ήταν πως όποτε οι αντίπαλοι προσπαθούσαν να παγιδεύσουν (trap) ή να δυσκολέψουν με διπλό μαρκάρισμα (double team) τον Mark Price, εκείνος πάσαρε τη μπάλα στον ελεύθερο Williams για ένα ελεύθερο σουτ.

Εκτός από την ανίχνευση «ανωμαλιών» μέσω του Attribute Focusing, το Advanced Scout έχει τη δυνατότητα να ανακαλύπτει εκείνα τα χαρακτηριστικά που επηρεάζουν περισσότερο τη έκβαση του αγώνα. Η συγκεκριμένη υπηρεσία που προσφέρει το εργαλείο αυτό είναι εξίσου σημαντική γιατί δίνει την ευκαιρία στους προπονητές να μελετούν τα χαρακτηριστικά και τις τάσεις τόσο των δικών τους όσο και των αντίπαλων παικτών.

Synergy Online

Εργαλείο παρόμοιο με το Advanced Scout είναι και το Synergy Online από τη Synergy Sports Technology. Το προϊόν αυτό είναι αφιερωμένο σε πολυμέσα που βασίζονται στο άθλημα του basketball και περιέχει ένα ευρετήριο από ζωντανές μεταδόσεις ως αναζητήσιμα μέσα ενημέρωσης. Με το σύστημα αυτό προπονητές, παίκτες αλλά και φίλαθλοι μπορούν να εξετάζουν και να μελετούν το παιχνίδι της ομάδας τους καθώς και να λαμβάνουν συνεχώς ανανεώσεις σχετικά με τη στατιστική των παικτών. Μέσω του Synergy Online οι φίλαθλοι μπορούν να παρακολοθούν ζωντανές μεταδόσεις αγώνων από τον ηλεκτρονικό τους υπολογιστή ή ακόμα και από φορητές συσκευές όπως τα tablets και τα κινητά τηλέφωνα.

Εικόνα 11: Live streaming analysis σε tablet από το synergy online

Η ίδια εταιρία βρίσκεται πίσω από την κατασκευή ηλεκτρονικών παιχνιδιών, όπως είναι η σειρά NBA Live προσφέροντας πραγματικά δεδομένα αγώνων στους ενδιαφερόμενους. Το βασικό προϊόν της, το Digital DNA, μοντελοποιεί περίπου 1000 χαρακτηριστικά και τάσεις παικτών με σκοπό να προβλέψει τις μελλοντικές τους

αποδόσεις, προσφέροντας με αυτό τον τρόπο μια πραγματική εικόνα μέσα από την εμπειρία του παιχνιδιού.

Εικόνα 12: Synergy Online στο ηλεκτρονικό παιχνίδι NBA LIVE 15

5.3 Scouting tools (εργαλεία κατασκοπείας)

Οι αθλητικοί κατάσκοποι συνήθιζαν παλαιότερα να βασίζονται στους παραδοσιακούς τρόπους παρακαλούθησης της απόδοσης των παικτών χωρίς να χρησιμοποιούν τεχνολογικά μέσα. Αυτό ήρθε να αλλάξει με την κατασκευή ηλεκτρονικών προγραμμάτων με τη βοήθεια των οποίων οι scouts μπορούν να παρακολουθούν τόσο τη συνολική εξέλιξη του αγώνα όσο και εκείνα τα χαρακτηριστικά που επηρεάζουν την ατομική απόδοση των παικτών.

Digital Scout

Το Digital Scout είναι ένα ηλεκτρονικό εργαλείο με τη βοήθεια του οποίου μπορούμε να επεξεργαστούμε τη στατιστική που συλλέξαμε από ένα παιχνίδι. Φίλαθλοι και αθλητικοί οργανισμοί μπορούν να χρησιμοποιήσουν αυτό το λογισμικό με βάση ιστορικά στοιχεία που έχουν συλλέξει. Με το πρόγραμμα αυτό οι scouts κι όχι μόνο έχουν τη δυνατότητα να παρακολουθούν όλες τις στατιστικές κατηγορίες που τους ενδιαφέρουν τόσο συνολικά για την ομάδα όσο και ξεχωριστά για κάθε παίκτη. Μπορούν ακόμη να παρακολουθούν και να εξετάζουν κάθε σουτ, ριμπάουντ ή ασσίστ

καθώς επίσης και να μελετούν γραφικές απεικονίσεις συγκεκριμένων χαρακτηριστικών, όπως για παράδειγμα οι προσπάθειες για σουτ.

Εικόνα 13: Digital Scout σε mobile συσκευές

5.4 Συμπεράσματα

Τα Data Mining εργαλεία για τους προπονητές, τους παίκτες, τους φιλάθλους και άλλους ενδιαφερόμενους γίνονται όλο και συνηθέστερα αποτελώντας μέχρι και απαίτηση από μερικούς στο χώρο του επαγγελματικού basketball για την καλύτερη εξαγωγή συμπερασμάτων. Τα εργαλεία αυτά διαφέρουν σε μέγεθος, πεδίο δράσης και επίπεδο λεπτομέρειας. Από απλές αναφορές σε μορφή κειμένου ανίχνευσης μιας μη φυσιολογικής εξέλιξης κατά τη διάρκεια του αγώνα, μέχρι εξειδικευμένες οπτικοποιήσεις που επιτρέπουν στους ειδικούς να ερευνούν πολύπλοκα δεδομένα, τα συστήματα αυτά έχουν αναπτύξει νέες ιδέες και προοπτικές στον τρόπο που βλέπουμε τα σπορ σήμερα. Εκτός από τη δυνατότητά τους να παρέχουν ανταγωνιστικό πλεονέκτημα, με κατάλληλη χρήση κρατούν τους αγώνες ασφαλείς από κερδοσκόπους.

6. Μοντέλα Πρόβλεψης (Predictive Modeling)

6.1 Εισαγωγή

Η προγνωστική μοντελοποίηση ανέκαθεν αποτελούσε πρωταρχικό στόχο αρκετών ατόμων και οργανισμών. Η επιστήμη αυτή έχει αρκετές τεχνικές, με την προσωμοίωση και τη μηχανική μάθηση να είναι οι βασικότερες. Προσομοιώσεις στο άθλημα του basketball, που προσφέρουν κάποια προγράμματα, μπορούν να μοντελοποιήσουν μια ολόκληρη αγωνιστική περίοδο και να βγάλουν χρήσιμα συμπεράσματα μέσα από την ανακάλυψη προτύπων ως προς την ιδανική χρησιμοποιήση των παικτών κατά τη διάρκεια του αγώνα ή τις ενδεχόμενες τάσεις τους ως προς το σκοράρισμα. Επιπρόσθετες προσομοιώσεις μπορούν να αναπτυχθούν έτσι ώστε απρόβλεπτα γεγονότα που μπορεί για παράδειγμα να οδηγήσουν σε έναν απρόσμενο μακροχρόνιο τραυματισμό να μπορούν να προβλεφθούν. Εκτός από τη δυναμική των προσομοιώσεων, οι τεχνικές μηχανικής μάθησης μπορούν επίσης να ανακαλύψουν νέα πρότυπα και τάσεις.

Το κίνητρο πρόβλεψης αποτελεσμάτων βασισμένων σε ιστορικά δεδομένα, οδήγησε στην ανάπτυξη αρκετών εφαρμογών σχετικών με τα σπορ, όπως οι στατιστικές προσομοιώσεις και οι τεχνικές μηχανικής μάθησης. Χρησιμοποιώντας αυτά τα εργαλεία, τα πρότυπα στα δεδομένα μπορούν να χρησιμοποιηθούν και να χειραγωγηθούν για προσωπικούς, ανταγωνιστικούς και οικονομικούς λόγους.

Μέσα από την προβλεπτική μοντελοποίηση ένα μέρος της αθλητικής δραστηριότητας κέρδισε την προσοχή, κι όχι μόνο από την προοπτική της στατιστικής. Η έννοια της μέγιστης απόδοσης ενός παίκτη, γνωστή και ως hot hand, όταν δηλαδή ένας παίκτης ανεβάζει την απόδοσή του πάνω από το μέσο όρο για παρατεταμένο χρονικό διάστημα. Σε έρευνα επ' αυτού, στο άθλημα της καλαθοσφαίρησης, οι ερευνητές υποστηρίζουν πως αν υπήρχε τέτοια κατάσταση, τότε η ευστοχία σε ένα σουτ θα αύξανε την πιθανότητα το επόμενο σουτ να ήταν κι αυτό εύστοχο. Ωστοσο, από εμπειρικές μελέτες έχει βρεθεί πως η επιτυχία σε ένα σουτ είναι ανεξάρτητη από το αποτέλεσμα προηγούμενων προσπαθειών.

Από την άποψη της στατιστικής και των αριθμών μπορεί κάτι τέτοιο να ισχύει, χωρίς όμως να είναι τελεσίδικο, καθώς είναι αδύνατον να προεξοφλήσουμε και να παρομοιάσουμε τη μελλοντική απόδοση των αθλητών με τη ρίψη ενός νομίσματος που έχει τις ίδιες πιθανότητες το αποτέλεσμα να είναι κορώνα ή γράμματα. Κι αυτό διότι στον αθλητισμό γενικά και συγκεκριμένα στο basketball η ψυχολογία του αθλητή

παίζει καθοριστικό ρόλο στην απόδοσή του, έχοντας τη δυνατότητα να την εκτοξεύσει αλλά και να τη μειώσει παράλληλα.

Έτσι οι προσωμοιώσεις και οι τεχνικές μηχανικής μάθησης αποτελούν σημαντικά εργαλεία μελέτης τέτοιου είδους συμπεριφορών με στόχο την έγκαιρη πρόβλεψη.

6.2 Στατιστικές προσομοιώσεις

Οι στατιστικές προσομοιώσεις περιλαμβάνουν την απομίμηση καινούριων δεδομένων αγώνων έχοντας ως αναφορά ιστορικά στοιχεία. Όταν αυτή η «απομίμηση» των δεδομένων κατασκευαστεί, μπορεί να συγκριθεί με πραγματικά στοιχεία αγώνων έτσι ώστε να εκτιμηθεί η ακρίβεια των προβλέψεων.

Basketball

Το άθλημα της καλαθοσφαίρισης αποτελεί το πλέον ιδανικό σπορ για την εφαρμογή προσομοιώσεων, με το BBall να είναι το κατεξοχήν δημοφιλές εργαλείο προς αυτό το στόχο. Αναπτύχθηκε από τον ειδικό ερευνητή Bob Chaikin, σύμβουλο της ομάδας του NBA Miami Heat. Το λογισμικό αυτό χρησιμοποιεί ιστορικά δεδομένα και APBRmetrics έχοντας τη δυναμική να προσομοιάζει τα δεδομένα από έναν αγώνα μέχρι μιας ολόκληρης αγωνιστικής περιόδου. Σχεδιασμένο για προπονητές, scouts και τεχνικούς διευθυντές επαγγελματικών συλλόγων, το BBall μπορεί να προσδιορίσει το βέλτιστο πρότυπο ως προς τη διαχείριση των παικτών (substitution pattern) μιας ολόκληρης αγωνιστικής περιόδου (το πρότυπο δηλαδή εκείνο που αποφέρει τις περισσότερες νίκες), την επίπτωση που θα έχει στη συνολική απόδοση μιας ομάδας η μεταγραφή ενός παίκτη, την επίδραση επικείμενων τραυματισμών ενός ή περισσότερων παικτών καθώς και να αναγνωρίσει τους παράγοντες που είναι απαραίτητοι για τη βελτίωση της απόδοσης της ομάδας (πόντοι, ριμπάουντ, ασσίστ). Χιλιάδες τέτοιες προσομοιώσεις μπορούν να εκτελεστούν έτσι ώστε να μοντελοποιήσουν ένα μεγάλο εύρος μεταβλητών που συνεχώς αλλάζουν.

6.3 Μηχανική μάθηση (machine learning)

Εκτός από τη στατιστική πρόβλεψη, οι τεχνικές μηχανικής μάθησης είναι μια άλλη μέθοδος με την οποία μπορούμε να προβούμε σε προβλέψεις που αφορούν τα σπορ. Τα Νευρωνικά Δίκτυα (Neural Networks) είναι ένα από τα επικρατέστερα συστήματα μηχανικής μάθησης που εφαρμόζονται στο χώρο των αθλημάτων. Στη μέθοδο αυτή τα σύνολα δεδομένων αφομοιώνονται από το σύστημα, έτσι ώστε κρυμμένες τάσεις σε αυτά να εκμεταλλεύονται για ανταγωνιστικούς και οικονομικούς σκοπούς. Άλλες

τεχνικές μηχανικής μάθησης που χρησιμοποιούνται ευρέως στο χώρο του αθλητισμού είναι οι γενετικοί αλγόριθμοι (genetic algorithms), ο αλγόριθμος ID3 δένδρου αποφάσεων και μια παραλλαγή βασισμένη στην παλινδρόμηση του ταξινομητή Support Vector Machine, που ονομάζεται Support Vector Regression.

6.4 Εμπορικά προϊόντα

Ξεχωριστά από τα προγράμματα προσομοίωσης και μηχανικής μάθησης, υπάρχουν και διάφορα περισσότερο εμπορικά βοηθήματα. Αναφέρουμε ορισμένα από αυτά τα οποία μας προσφέρουν ποικίλες δυνατότητες: Synergy Online, Dr.Z System, Front Office Football και Visual Sports. Το Synergy Online, όπως έχει αναφερθεί και παραπάνω, είναι ένα προϊόν που επιτρέπει στους πάσης φύσης ενδιαφερόμενους να έχουν πρόσβαση σε πραγματικά δεδομένα αγώνων. Αποτελεί τη βάση δημοφιλών ηλεκτρονικών παιχνιδιών που σχετίζονται με το άθλημα του basketball φέρνοντας το ρεαλισμό μέσα από την εμπειρία του παιχνιδιού. Τα υπόλοιπα εργαλεία βρίσκουν καλύτερη εφαρμογή σε άλλα αθλήματα όπως είναι το ποδόσφαιρο, το baseball, το γκολφ και το χόκεϋ.

6.5 Συμπεράσματα

Τα συστήματα προσομοίωσης και μηχανικής μάθησης αποτελούν σημαντικό κομμάτι στην ανάπτυξη των μοντέρνων αθλημάτων. Η δυνατότητα να εφαρμόζουμε στατιστικές μεθόδους και μαθηματικά μοντέλα με στόχο να φτάσουμε σε άμεσα αποτελέσματα μετατρέπεται σε ανεκτίμητο αγαθό. Τα συστήματα αυτά ποικίλουν, από προσομοιώσεις που μοντελοποιούν την αξία των δεδομένων μιας ολόκληρης αγωνιστικής περιόδου αναγνωρίζοντας τις μεγαλύτερες πιθανότητες νίκης, μέχρι προσομοιώσεις που ανιχνεύουν αδυναμίες στο παιχνίδι της ομάδας προσφέροντας συμβουλές διόρθωσής τους. Έχουν τη δυνατότητα να προσφέρουν άμεσα αποτελέσματα αν κάποιος αλλάξει τις ενέργειες που προτείνουν. Με αυτό τον τρόπο ο χρήστης μπορεί να δει τα αποτελέσματα της δουλειάς του σε πραγματικό χρόνο εισπράττοντας ένα αίσθημα εκπλήρωσης καθώς τα συστήματα αυτά βελτιώνουν με τον καλύτερο τρόπο τις ιδέες του. Αν και ακόμα βρίσκονται σε πρώιμο στάδιο, είναι πολύ ενδιαφέρον να παρακολουθήσουμε την πορεία τους τα επόμενα χρόνια.

7. Το Λογισμικό Weka

7.1 Εισαγωγή στο Weka

Το εργαλείο WEKA (Waikato Environment for Knowledge Analysis) είναι ένα δημοφιλές λογισμικό με σκοπό τη μηχανική μάθηση, αναπτυγμένο σε Java, το οποίο αναπτύχθηκε στο Πανεπιστήμιο του Waikato, στη Νέα Ζηλανδία. Το WEKA περιλαμβάνει μία συλλογή από εργαλεία οπτικοποίησης και αλγορίθμους για ανάλυση δεδομένων, εξόρυξη γνώσης και ανάπτυξη προγνωστικών μοντέλων (predictive modelling).

Από το 1997, οπότε και δημιουργήθηκε η πρώτη έκδοσή του πλήρως βασισμένη σε Java, μέχρι σήμερα, το Weka χρησιμοποιείται ευρέως, και ιδιαίτερα για εκπαιδευτικούς αλλά και ερευνητικούς σκοπούς. Το Weka αποτελεί μία ιδιαιτέρως δημοφιλής εφαρμογή παγκοσμίως δεδομένου ότι πρόκειται για ελεύθερο λογισμικό, μπορεί να αναπτυχθεί σε όλες τις σύγχρονες πλατφόρμες λογισμικού εξαιτίας του ότι είναι αναπτυγμένο σε Java, είναι ιδιαιτέρως εύχρηστο χάρη στο γραφικό περιβάλλον που διαθέτει, και φυσικά ότι περιέχει μία εκτενή συλλογή τεχνικών ανάλυσης δεδομένων και εξόρυξης γνώσης. Ειδικότερα, με τη βοήθεια του Weka πέρα από την προεπεξεργασία και οπτικοποίηση των δεδομένων, μπορούμε να εκτελέσουμε ποικίλες τεχνικές εξόρυξης γνώσης από δεδομένα όπως είναι η κατηγοριοποίηση (classification), η συσταδοποίηση (clustering), η εύρεση κανόνων συσχέτισης (association rules mining) και η πρόβλεψη (prediction). Για κάθε μία από τις παραπάνω τεχνικές το εργαλείο παρέχει ένα σύνολο αλγορίθμων που μπορούν να χρησιμοποιηθούν ανάλογα με τις ανάγκες της εκάστοτε μελέτης, παρέχοντας πλήρη δυνατότητα παραμετροποίησης τους από το χρήστη.

7.2 Δομή αρχείων στο Weka

7.2.1 Αρχεία arff στο Weka

Τα βασικά αρχεία τα οποία δέχεται ως είσοδο το Weka έχουν την κατάληξη ARFF (Attribute-Relation File Format) και πρόκειται για ένα αρχείο κειμένου χαρακτήρων ASCII (ASCII text file) το οποίο περιγράφει/περιέχει μια σειρά από υποδείγματα (instances) τα οποία περιγράφονται από κάποια χαρακτηριστικά (attributes). Ένα παράδειγμα ενός τέτοιου αρχείου φαίνεται παρακάτω στην Εικόνα 14.

```
% 1. Title: Iris Plants Database
%
% 2. Sources:
%      (a) Creator: R.A. Fisher
%      (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
%      (c) Date: July, 1988
%
@RELATION iris

@ATTRIBUTE sepallength  NUMERIC
@ATTRIBUTE sepalwidth   NUMERIC
@ATTRIBUTE petallength  NUMERIC
@ATTRIBUTE petalwidth   NUMERIC
@ATTRIBUTE class        {Iris-setosa,Iris-versicolor,Iris-virginica}



@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
```

Εικόνα 14: Παράδειγμα αρχείου arff

Οι γραμμές οι οποίες ξεκινάνε με "%" είναι σχόλια, δηλ., δε λαμβάνονται υπόψη όταν «φορτώνεται» το αρχείο, έτσι ώστε να καθιστάται σαφές το τι ακριβώς περιλαμβάνεται τελικά στο αρχείο.

@relation

Μετά από τα εισαγωγικά σχόλια ακολουθεί το όνομα που περιγράφει κατά κάποιο τρόπο το αρχείο, στη γραμμή που αρχίζει με το @relation. Μάλιστα, η γραμμή αυτή είναι απαραίτητη και δεν μπορεί να παραλειφθεί.

Η δήλωση γίνεται χρησιμοποιώντας την παρακάτω εντολή:

@relation <relation-name>

Στο παράδειγμά μας στην Εικόνα 14 το <relation-name> είναι το iris.

@attributes

Μετά από τη γραμμή @relation, ακολουθεί η δήλωση όλων των χαρακτηριστικών που περιγράφουν το συγκεκριμένο σύνολο παραδειγμάτων. Η δήλωση γίνεται χρησιμοποιώντας την παρακάτω εντολή:

@attribute <attribute-name> <datatype>

όπου <attribute-name> είναι το όνομα του χαρακτηριστικού, (στο παράδειγμά μας στην Εικόνα 14 το <attribute-name> αντιστοιχεί στα sepallength, sepalwidth, petallength, petalwidth, class), και το οποίο πρέπει να ξεκινά με γράμμα. Σε περίπτωση που ένα χαρακτηριστικό περιγράφεται με δύο ή περισσότερες λέξεις που χωρίζονται με κενό, τότε θα πρέπει όλες αυτές να περικλείονται σε εισαγωγικά (“ .”).

Το όρισμα <datatype> καθορίζει τον τύπο του χαρακτηριστικού, με το Weka να υποστηρίζει τέσσερις διαφορετικούς τύπους

1. Αριθμητικά (numeric)
2. Ονομαστικά (<nominal-specification>)
3. Αλφαριθμητικά (string)
4. Ημερομηνίες (date [<date-format>])

@data

Μετά από τη δήλωση των χαρακτηριστικών ακολουθεί η δήλωση ότι θα ακολουθήσουν τα δεδομένα. Η γραμμή @data δηλώνει ότι θα ακολουθήσουν τα δεδομένα (υποδείγματα). Κάθε υπόδειγμα αναπαριστάται από μία γραμμή, με το τέλος του υποδείγματος να σηματοδοτείται με χαρακτήρες επαναφοράς (carriage returns). Οι τιμές των χαρακτηριστικών για κάθε παράδειγμα διαχωρίζονται μεταξύ τους με κόμμα. Θα πρέπει να εμφανίζονται με τη σειρά με την οποία έχουν δηλωθεί στην επικεφαλίδα του αρχείου. Στην περίπτωση που κάποιες τιμές λείπουν για κάποιο λόγο, τότε στη θέση τους αναγράφεται ένα λατινικό ερωτηματικό όπως φαίνεται παρακάτω

@data 4.4,?,1.5,?,Iris-setosa

Η τυποποίηση arff αποτελεί τη φυσική μέθοδο αποθήκευσης δεδομένων του Weka. Τα arff αρχεία έχουν δύο ξεχωριστά τμήματα. Το πρώτο τμήμα είναι το Header, το οποίο ακολουθείται από το τμήμα Data. Το Header του αρχείου ARFF περιέχει το όνομα της σχέσης, μια λίστα των μεταβλητών (οι στήλες στα δεδομένα), και τους τύπους τους. Υποστηρίζει τόσο αριθμητικά (numeric) όσο και ονομαστικά (nominal) χαρακτηριστικά (attributes). Τα δεδομένα πολύ συχνά βρίσκονται σε υπολογιστικά φύλλα ή σε βάσεις δεδομένων. Τα προγράμματα που τα χειρίζονται συνήθως επιτρέπουν την εξαγωγή των δεδομένων σε τυποποίηση csv. Το Weka έχει ενσωματωμένο μετατροπέα αρχείων από csv σε arff, με τη διαδικασία μετατροπής να παρουσιάζεται παρακάτω.

Πρώτο βήμα:

Ανοίγουμε το αρχείο με κατάληξη .csv που θέλουμε να μετατρέψουμε, με έναν επεξεργαστή κειμένου (text editor).

Δεύτερο βήμα:

Προσθέτουμε το όνομα του σετ δεδομένων σε μια γραμμή στην αρχή, η οποία θα αρχίζει με την έκφραση @relation, τις πληροφορίες για τα κάθε χαρακτηριστικό χρησιμοποιώντας την έκφραση @attribute (κάθε χαρακτηριστικό σε νέα γραμμή) και μια σειρά με την έκφραση @data, που δείχνει ότι από εκεί και κάτω βρίσκονται τα δεδομένα.

Συνοψίζοντας:

Η δήλωση @relation

Το όνομα της σχέσης ορίζεται ως η πρώτη γραμμή στο αρχείο arff. Η μορφή είναι:
@relation <relation-name>

όπου το <relation-name> είναι ένα string και αποτελεί το όνομα του συνόλου δεδομένων.

Οι δηλώσεις @attribute

Με αυτές προσθέτουμε τις πληροφορίες των μεταβλητών. Η μορφή τους είναι:
@attribute <attribute-name> <data-type>

όπου το <attribute-name πρέπει να ξεκινά με έναν αλφαβητικό χαρακτήρα και το <data-type> μπορεί να είναι numeric, nominal, string και date.

Το τμήμα Data περιέχει τη γραμμή δήλωσης των δεδομένων και τις γραμμές των στιγμιοτύπων.

Η Δήλωση @data

Η δήλωση @data είναι μια απλή γραμμή που δηλώνει την έναρξη του τμήματος των δεδομένων στο αρχείο. Η μορφή είναι @data.

Τα δεδομένα των στιγμιότυπων

Κάθε στιγμιότυπο αναπαριστάται με μια απλή γραμμή. Ένα σύμβολο ποσοστού (%) εισάγει ένα σχόλιο, το οποίο συνεχίζει στο τέλος της γραμμής.

Οι τιμές των μεταβλητών για κάθε στιγμιότυπο χωρίζονται με κόμματα. Πρέπει να εμφανίζονται με τη σειρά που δηλώθηκαν στο τμήμα header (δηλ. τα δεδομένα που αντιστοιχούν στην ν-οστή δήλωση @attribute είναι πάντα το ν-οστό πεδίο της μεταβλητής). Οι ελλιπείς τιμές αναπαριστώνται με ένα ερωτηματικό.

Τρίτο βήμα:

Κάνουμε αποθήκευση του αρχείου στην μορφή arff, και το αρχείο είναι έτοιμο για να εισαχθεί στο Weka. Ακόμα και για ένα δυσνόητο αρχείο arff λόγω του μεγάλου όγκου των δεδομένων, είναι εμφανή τα σημεία τα οποία ξεκινούν με @relation , @attribute και @data.

7.2.2 Τύποι χαρακτηριστικών του Weka

Αριθμητικά χαρακτηριστικά (numeric attributes)

Τα αριθμητικά χαρακτηριστικά μπορεί να είναι είτε πραγματικοί είτε ακέραιοι αριθμοί, στο παράδειγμά μας στην Εικόνα 14 τέτοια είναι τα χαρακτηριστικά sepallength, sepalwidth, petallength, petalwidth.

Ονομαστικά χαρακτηριστικά (nominal attributes)

Τα χαρακτηριστικά που παίρνουν ονομαστικές τιμές ορίζονται χρησιμοποιώντας αγκύλες εντός των οποίων γράφονται όλες οι δυνατές τιμές:{<nominal-name1>, <nominal-name2>, <nominal-name3>, ...}. Στο παράδειγμά μας στην Εικόνα15, τέτοιο είναι το χαρακτηριστικό class {Iris-setosa,Iris-versicolor,Iris-virginica}. Όπως και προηγουμένως για την περίπτωση κενών θα πρέπει να χρησιμοποιούνται (“ .”).

Αλφαριθμητικά χαρακτηριστικά (string attributes)

Τα αλφαριθμητικά χαρακτηριστικά επιτρέπουν τη δημιουργία αυθαίρετων αλφαριθμητικών δομών κάτι το οποίο αφορά κυρίως εφαρμογές text-mining. Ο ορισμός ενός τέτοιου χαρακτηριστικού έχει την ακόλουθη μορφή

@attribute LCC string

Ημερομηνίες (date attributes)

Ο καθορισμός χαρακτηριστικών που παίρνουν ως τιμή ημερομηνίες γίνεται με την παρακάτω εντολή

@attribute <name> date [<date-format>]

όπου <name> είναι το όνομα του χαρακτηριστικού και date είναι η ημερομηνία σύμφωνα με το παρακάτω format

"yyyy-MM-dd'T'HH:mm:ss"(ISO-8601).

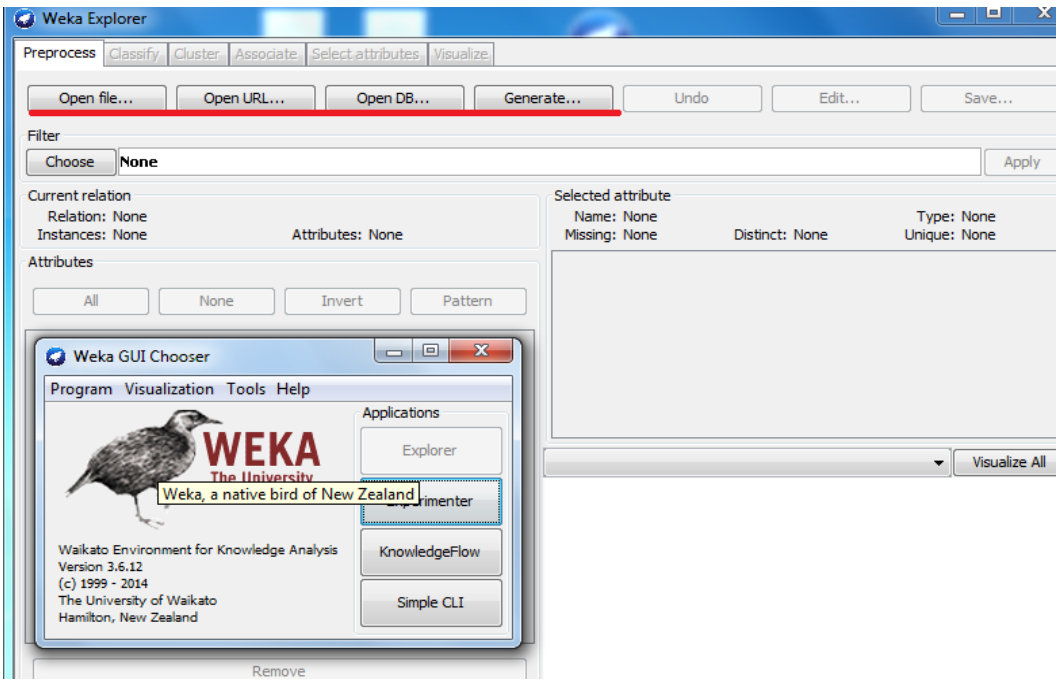
Π.χ. 2015-07-10-T10:43:22

Πηγή: **http://www.cs.waikato.ac.nz/ml/weka/arff.html**

7.3 Φόρτωση δεδομένων στο Weka

Η φόρτωση δεδομένων στο Weka μπορεί να γίνει με ποικίλλους τρόπους. Στο περιβάλλον εργασίας Explorer, το οποίο χρησιμοποιούμε και στην παρούσα εργασία, παρατηρούμε ότι υπάρχουν στην καρτέλα preprocess οι επιλογές:

- Open file – για την εύρεση του αρχείου στον υπολογιστή μας, το οποίο θέλουμε να φορτώσουμε
- Open URL – για τη φόρτωση δεδομένων κατευθείαν από κάποια ιστοσελίδα του διαδικτύου
- Open DB – για να φορτώσουμε όποια δεδομένα θέλουμε από μια βάση δεδομένων (database)
- Generate – για τη δημιουργία τυχαίων δεδομένων μέσα από διάφορους αλγορίθμους, σε περίπτωση που δεν έχουμε δεδομένα διαθέσιμα και θέλουμε να πειραματιστούμε με το Weka



Εικόνα 15: Φόρτωση αρχείου/δεδομένων στο Weka

Τα δεδομένα αυτά βέβαια μπορεί να βρίσκονται σε διάφορες μορφές και τύπους αρχείων. Το Weka περιέχει ενσωματωμένους μετατροπείς για τους πιο κοινούς τύπους αρχείων για να μετατρέψει στην προαναφερθείσα τυποποίηση, δηλαδή την arff με την οποία μπορεί να τα χειριστεί.

7.4 Περιγραφή περιβάλλοντος του Weka

Εικόνα 16: Το εργαλείο Weka

Όταν ξεκινά η εκτέλεση του Weka, ο χρήστης καλείται να διαλέξει ένα από τα τέσσερα πιθανά περιβάλλοντα εργασίας που το Weka του παρέχει με τις ονομασίες:

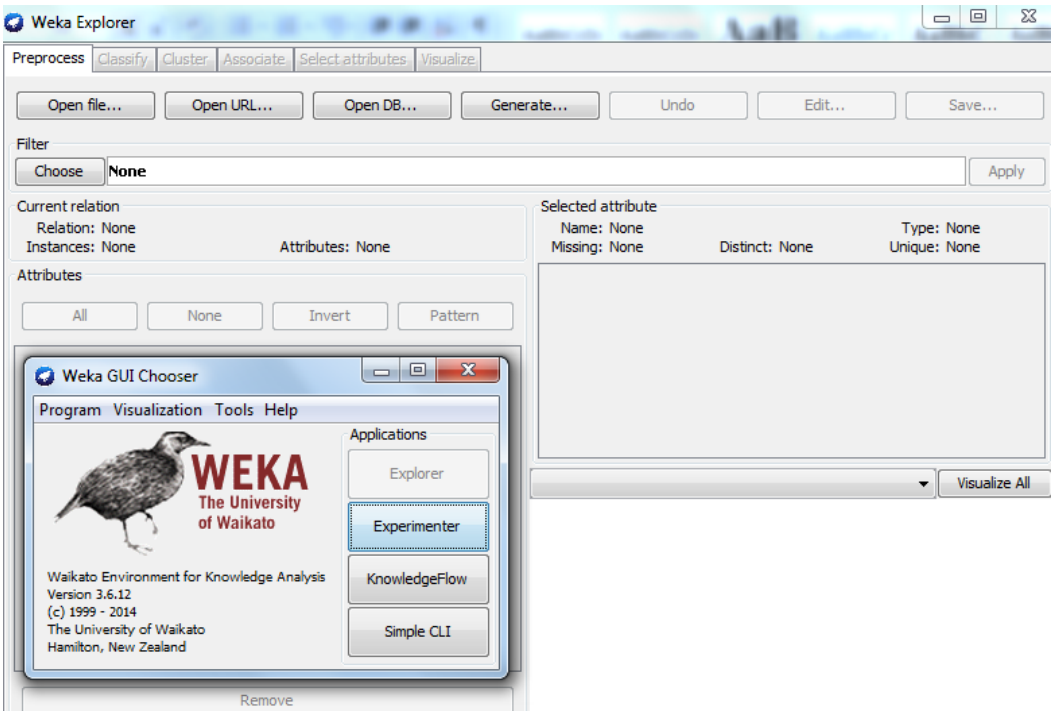
- Explorer
- Experimenter
- Knowledge Flow
- Simple CLI (Command Line Interface)

Ακολουθεί μια συνοπτική παρουσίαση των κυρίων στοιχείων του κάθε περιβάλλοντος.

7.3.1 Το περιβάλλον εργασίας Explorer

Ο πιο εύχρηστος τρόπος για να χρησιμοποιήσει κανείς το Weka είναι μέσω αυτού του γραφικού περιβάλλοντος. Παρέχει πρόσβαση σε όλες τις δυνατότητες που έχει το Weka, παρουσιάζοντάς του τις επιλογές του μέσα από κατάλληλα οργανωμένες λίστες. Επιλογές που δεν είναι συμβατές με την εκάστοτε διαδικασία που ακολουθεί ο

χρήστης, παρουσιάζονται γκριζαρισμένες. Ακόμα χρησιμοποιούνται λογικές προεπιλεγμένες τιμές σε διάφορες επιλογές ώστε ο χρήστης να είναι σε θέση να έχει κάποια αποτελέσματα με την ελάχιστη δυνατή προσπάθεια. Μέσα από τον Explorer μπορεί κανείς να εξετάσει τα αποτελέσματα τα οποία έχουν προκύψει από την εφαρμογή των διάφορων αλγορίθμων, να αξιολογήσει και να συγκρίνει διαφορετικά μοντέλα που έχει δημιουργήσει από διάφορα σύνολα δεδομένων, και να οπτικοποιήσει τόσο τα μοντέλα όσο και το σύνολα δεδομένων καθαυτά.



Εικόνα 17: Το περιβάλλον εργασίας Explorer στο Weka

Ο Explorer είναι οργανωμένος σε έξι μεγάλες κατηγορίες λειτουργιών με τις αντίστοιχες ετικέτες:

- Preprocess

Αποτελεί κατηγορία που μπορεί κανείς να βρει εργαλεία και αλγορίθμους που αφορούν την επιλογή ή την τροποποίηση του συνόλου δεδομένων που τυγχάνει επεξεργασίας. Αναλυτικότερα, μας επιτρέπει να εισάγουμε στοιχεία από μια βάση δεδομένων, ένα csv ή ένα arff αρχείο κ.λπ., και να προεπεξεργαστούμε αυτά τα δεδομένα χρησιμοποιώντας αλγόριθμους φιλτραρίσματος. Αυτά τα φίλτρα μπορούν να χρησιμοποιηθούν για να μετασχηματίσουν τα δεδομένα (π.χ. να μετατρέψουν

αριθμητικές τιμές σε αντίστοιχες διακριτές) και δύναται να διαγράψουν στιγμιότυπα και μεταβλητές σύμφωνα με συγκεκριμένα κριτήρια.

- Classify

Αποτελεί κατηγορία που περιέχει αλγορίθμους κατάλληλους για προβλήματα ταξινόμησης ή παλινδρόμησης. Αναλυτικότερα, επιτρέπει στο χρήστη να εφαρμόσει αλγόριθμους ταξινόμησης και παλινδρόμησης (που καλούνται ταξινομητές στο WEKA) σε ένα σύνολο δεδομένων, να υπολογίσει την ακρίβεια του μοντέλου πρόβλεψης που προκύπτει καθώς και άλλα μέτρα αξιολόγησης της απόδοσης, όπως οι πίνακες συνάφειας, οι καμπύλες ROC κ.λπ. ή να οπτικοποιεί τις εσφαλμένες προβλέψεις ή το ίδιο το μοντέλο (εάν το μοντέλο είναι δυνατόν να οπτικοποιηθεί, όπως π.χ. σε ένα δέντρο απόφασης).

- Cluster

Αποτελεί κατηγορία όπου περιέχονται αλγόριθμοι για την εύρεση υποομάδων μέσα από το σύνολο δεδομένων. Αναλυτικότερα, δίνει πρόσβαση στις τεχνικές συσταδοποίησης στο WEKA που στόχο έχουν την εύρεση ομάδων «όμοιων» δεδομένων ή με άλλα λόγια την εύρεση συστάδων χαρακτηριστικών με «πανομοιότυπη» συμπεριφορά.

- Associate

Αποτελεί κατηγορία όπου περιέχονται αλγόριθμοι κατάλληλοι για την εύρεση κανόνων συσχέτισης μέσα στο σύνολο δεδομένων και την αξιολόγησή τους. Αναλυτικότερα, παρέχει πρόσβαση στην εκμάθηση κανόνων συσχετίσεων που επιχειρούν να προσδιορίσουν όλες τις σημαντικές αλληλεξαρτήσεις μεταξύ των μεταβλητών στα δεδομένα.

- Select attributes

Αποτελεί κατηγορία στην οποία περιέχονται αλγόριθμοι για την επιλογή των πιο σχετικών χαρακτηριστικών μέσα από το σύνολο δεδομένων. Αναλυτικότερα, παρέχει αλγόριθμους για τον προσδιορισμό των μεταβλητών με το μεγαλύτερο βαθμό επιρροής στην πρόβλεψη σε ένα σύνολο δεδομένων.

- Visualize

Αποτελεί κατηγορία που μπορεί κανείς να βρει εργαλεία για την οπτικοποίηση των δεδομένων ή των μοντέλων μέσω γραφημάτων, πινάκων διαγραμμάτων τα οποία μπορούν να επιλεχθούν, να διερευνηθούν και να αναλυθούν περαιτέρω χρησιμοποιώντας διάφορες εφαρμογές.

7.3.2 Το περιβάλλον εργασίας Experimenter

Το περιβάλλον Experimenter εξυπηρετεί ερευνητικές εργασίες στις οποίες εκτελούνται πειράματα μεγαλύτερου εύρους, δηλαδή εφαρμόζονται διάφορα μαθησιακά σχήματα, σε πολλά διαφορετικά σετ δεδομένων και συχνά με διαφορετικές παραμέτρους. Ο Experimenter αυτοματοποιεί την πειραματική διαδικασία. Οι πληροφορίες που προκύπτουν για τα διάφορα μαθησιακά σχήματα και τα διάφορα σετ δεδομένων μπορούν να αποθηκευθούν και να αποτελέσουν με τη σειρά τους αντικείμενο περαιτέρω μελέτης και εφαρμογής data mining.

Το περιβάλλον Experimenter υπερέχει σε τέτοιου είδους απαιτητικές εργασίες έναντι των περιβάλλοντων Explorer και Knowledge Flow (βλ. παρακάτω). Τα περιβάλλοντα Explorer και Knowledge Flow εξυπηρετούν τους χρήστες που θέλουν να διαπιστώσουν πόσο καλή απόδοση έχουν τα μαθησιακά σχήματα σε συγκεκριμένα σετ δεδομένων. Επιπλέον, ο Experimenter έχει ένα χαρακτηριστικό υπεροχής ανάλογο αυτού που έχει το περιβάλλον Knowledge Flow. Ενώ το Knowledge Flow ξεπερνά τους περιορισμούς σχετικά με το μέγεθος του αρχείου που μπορεί να επεξεργαστεί, εξετάζοντας κάθε υπόδειγμα από το σετ δεδομένων χωριστά χωρίς να χρειάζεται να φορτώσει ολόκληρο το σετ δεδομένων, ο Experimenter ξεπερνά τους χρονικούς περιορισμούς. Περιέχει υποδομές για προχωρημένους χρήστες, ώστε να διαμοιράσουν το υπολογιστικό φορτίο που απαιτείται από μεγάλου εύρους πειράματα, σε διάφορους υπολογιστές. Ευνόητο είναι ότι το συγκεκριμένο χαρακτηριστικό απαιτεί ένα επίπεδο γνώσεων από τη μεριά του χρήστη και απευθύνεται σε αρκετά προχωρημένους χρήστες.

7.3.3 Το περιβάλλον εργασίας Knowledge Flow

Λειτουργικά, το Knowledge Flow περιβάλλον μοιάζει πολύ με τον Explorer. Μπορεί κανείς να εκτελέσει αντίστοιχες εργασίες και στα δύο. Το περιβάλλον Knowledge Flow απευθύνεται σε πιο προχωρημένους χρήστες του Weka. Πιο συγκεκριμένα απευθύνεται σε όσους θέλουν να έχουν επίγνωση του πως τα δεδομένα και οι πληροφορίες που παράγονται από αυτά «κυλούν» μέσα στο σύστημα. Το Knowledge Flow παρέχει περισσότερη ευελιξία από την άποψη ότι ο πειραματιστής μπορεί να εξετάσει όλη τη διαδικασία λεπτομερώς και όχι μόνο το αποτέλεσμα που προκύπτει από αυτή. Ωστόσο το στοιχείο που το ξεχωρίζει από τον Explorer και το κάνει να υπερέχει είναι η δυνατότητα που παρέχει στο χρήστη για αυξητική λειτουργία (incremental operation).

Ο χρήστης επιλέγει τα διάφορα συστατικά κομμάτια του Weka, από μια μπάρα εργαλείων, τα τοποθετεί σε έναν πίνακα και τα συνδέει σε ένα κατευθυνόμενο γράφημα που υποδεικνύει πως γίνεται η ανάλυση και η επεξεργασία των δεδομένων. Αν όλα τα στοιχεία που έχουν συνδεθεί στον πίνακα έχουν τη δυνατότητα να

λειτουργήσουν αυξητικά, τότε έτσι λειτουργεί ολόκληρο το μαθησιακό σχήμα. Δεν διαβάζει ολόκληρο το σετ δεδομένων που του δίνεται σαν input πριν αρχίσει η «μάθηση», όπως θα έκανε ο Explorer, αλλά διαβάζει κάθε υπόδειγμα ξεχωριστά και το προωθεί τη διαδικασία που έχει σχηματιστεί στο Knowledge Flow πριν πάει στο επόμενο. Μια τέτοια διάταξη μπορεί επομένως να επεξεργαστεί αρχεία οποιουδήποτε μεγέθους, ακόμα και μεγαλύτερου της κύριας μνήμης του συστήματος, καθώς δεν χρειάζεται να τα αποθηκεύσει εσωτερικά για να ξεκινήσει τη διαδικασία.

7.3.4 Το περιβάλλον εργασίας Command Line Interface

Το Command Line Interface αποτελεί το τελευταίο περιβάλλον εργασίας που εμφανίζεται στο Weka. Επιλέγοντας το, έρχεται στην επιφάνεια ένας κενός χώρος με μια γραμμή εισαγωγής εντολών στο κάτω μέρος. Πρόκειται για το πιο απλό και χωρίς γραφικά βοηθήματα περιβάλλον εργασίας και απευθύνεται σε χρήστες που γνωρίζουν εις βάθος το Weka και τις εντολές του.

8. Βασικές Έννοιες και Εξερεύνηση Δεδομένων στο Weka

Πριν αρχίσουμε να μελετάμε τις στατιστικές μεθόδους και τους αλγορίθμους που θα εφαρμόσουμε για την εξόρυξη γνώσης από τα δεδομένα μας, θα ασχοληθούμε σε αυτό το κεφάλαιο με τις βασικές έννοιες που συναντούμε στο data mining. Επίσης, θα αναφερθούμε στην οπτικοποίηση των δεδομένων, τη φάση της προεπεξεργασίας τους καθώς και στα είδη γνώσης που παράγονται.

8.1 Είδη μάθησης

Στόχος του data mining είναι η εύρεση μιας κατανοητής και λειτουργικής περιγραφής μιας «αντίληψης» με εφαρμογή κατάλληλων και αποτελεσματικών αλγορίθμων. Ο όρος «αντίληψη» (concept) αναφέρεται στο αντικείμενο της μάθησης, και η περιγραφή της ζητούμενης αντίληψης, δηλαδή η απεικόνισή της, μπορεί να γίνει με ποικίλλους τρόπους, ανάλογα με τις ανάγκες του προβλήματός μας.

Σε ένα κλασικό πρόβλημα εξόρυξης δεδομένων, έχουμε ένα σύνολο δεδομένων εκπαίδευσης (training set) στο οποίο γνωρίζουμε την τιμή του αποτελέσματος και τις τιμές των χαρακτηριστικών που μας ενδιαφέρουν, και προσπαθούμε με βάση αυτά τα δεδομένα να κατασκευάσουμε ένα μοντέλο πρόβλεψης. Εν συνεχεία, το μοντέλο αυτό θα το χρησιμοποιήσουμε για να προβλέψουμε το αποτέλεσμα νέων συνόλων δεδομένων εξέτασης (test set), στα οποία σύνολα είναι γνωστές οι τιμές των χαρακτηριστικών αλλά δεν είναι γνωστή η τιμή του αποτελέσματος, δηλαδή η τιμή της τάξης (class).

Διακρίνουμε δύο κύριες μορφές μάθησης, τη μάθηση με επίβλεψη ή εποπτευομένη μάθηση (supervised learning) και τη μάθηση χωρίς επίβλεψη ή μη εποπτεύομενη μάθηση (unsupervised learning). Στην εποπτεύομενη μάθηση τα δεδομένα εκπαίδευσης συνοδεύονται από ετικέτες για την κλάση (class labels) στην οποία ανήκει το καθένα. Ουσιαστικά, στην εποπτεύομενη μάθηση η διαδικασία μάθησης οδηγείται από την παρουσία των αποτελεσμάτων της κλάσης. Στη μη εποπτεύομενη μάθηση δεν είναι γνωστό σε ποια κλάση ανήκουν τα δεδομένα εκπαίδευσης, δηλαδή ο πειραματιστής γνωρίζει μόνο τις τιμές των χαρακτηριστικών και όχι την τιμή του αποτελέσματος.

Από μία διαδικασία εξόρυξης πληροφοριών, η γνώση η οποία προκύπτει μπορεί να κατηγοριοποιηθεί με διάφορους τρόπους ανάλογα με το στόχο του προβλήματος που εξετάζουμε.

Τα κύρια είδη μάθησης κατηγοριοποιούνται ως εξής:

- Ταξινόμηση (classification)→ταξινόμηση των υποδειγμάτων σε μια προκαθορισμένη τάξη
- Συσχέτιση (association)→ανακάλυψη συσχετίσεων διαφόρων χαρακτηριστικών του συνόλου των δεδομένων
- Συσταδοποίηση ή Ομαδοποίηση δεδομένων (clustering)→εύρεση ομάδων αντικείμενων με υψηλό βαθμό ομοιότητας και εκχώρηση υποδειγμάτων στις ομάδες αυτές
- Αριθμητική πρόβλεψη→πρόβλεψη μίας αριθμητικής ποσότητας. Η εν λόγω διαδικασία είναι όμοια με την ταξινόμηση μόνο που η τάξη σε αυτή την περίπτωση είναι αριθμητική

8.2 Οπτικοποίηση και εξερεύνηση δεδομένων

Απλά εργαλεία οπτικοποίησης συχνά χρησιμοποιούνται για την κατανόηση των δεδομένων, όπως για παράδειγμα τα ιστογράμματα που απεικονίζουν την κατανομή των τιμών των ονομαστικών χαρακτηριστικών ή γραφήματα για τις τιμές των αριθμητικών χαρακτηριστικών. Τα εργαλεία αυτά μας βοηθούν να δούμε εύκολα αν η κατανομή είναι συμβατή με τη γνώση του πεδίου ή αν υπάρχουν άτυπες τιμές. Ένα ακόμη εργαλείο οπτικοποίησης είναι τα διαγράμματα δύο ή τριών διαστάσεων τα οποία μας υποδεικνύουν εξαρτήσεις και αλληλοσυσχετίσεις μεταξύ των χαρακτηριστικών. Ένα τέτοιο εργαλείο είναι απαραίτητο για την κατανόηση τεράστιων βάσεων, λόγω του αχανούς όγκου των δεδομένων τους.

Το λογισμικό Weka που χρησιμοποιούμε σε αυτή την εργασία παρέχει τη δυνατότητα στο χρήστη να οπτικοποιεί εύκολα το σύνολο των δεδομένων καθώς διαθέτει ξεχωριστό πλαίσιο (panel) οπτικοποίησης (Visualize). Ο χρήστης ανοίγοντας το σύνολο των δεδομένων του στον Explorer του Weka μπορεί αρχικά να δει διάφορα ιστογράμματα των χαρακτηριστικών, επιλέγοντας κάθε φορά το επιθυμητό χαρακτηριστικό, καθώς και το χαρακτηριστικό εκείνο το οποίο επιθυμεί να θεωρήσει ως τάξη (class). Επιπρόσθετα, κάνοντας χρήση του tab Visualize, μπορούμε να δούμε ένα πίνακα γραφημάτων για όλα τα χαρακτηριστικά. Αν θέλουμε μπορούμε να δούμε σε ξεχωριστό παράθυρο καθένα από αυτά τα γραφήματα επιλέγοντάς το, ή να κάνουμε zoom σε μία συγκεκριμένη περιοχή του (επιλογή Rectangle), να αλλάξουμε το μέγεθός του, το μέγεθος των σημείων, αλλά και το τρεμούλιασμα της εικόνας (jitter) το οποίο μας βοηθά να ξεχωρίσουμε τα σημεία και να αποκτήσουμε μία πιο καθαρή εικόνα.

8.3 Προεπεξεργασία δεδομένων

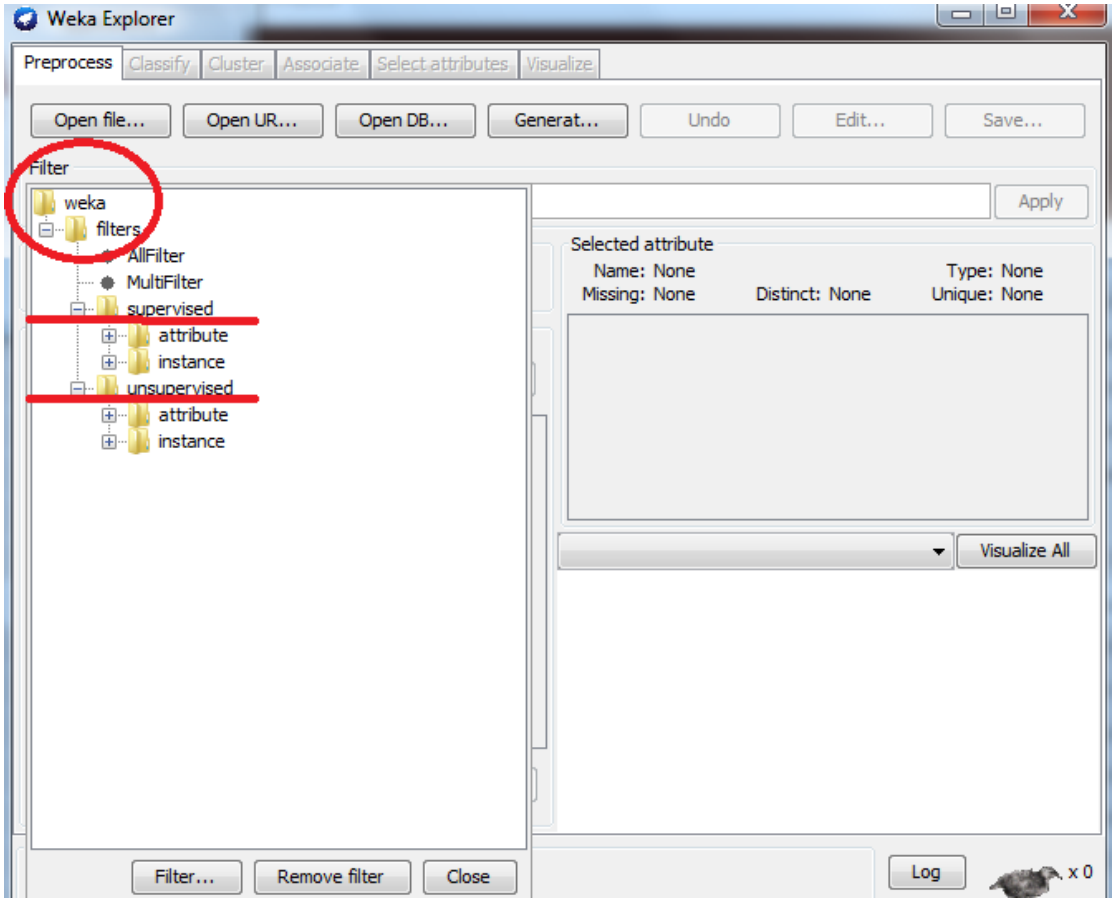
Η προεπεξεργασία των δεδομένων αποτελεί το πρώτο και ίσως σημαντικότερο βήμα σε μία διαδικασία εξόρυξης γνώσης. Πολύ συχνά απαιτείται ο μετασχηματισμός των δεδομένων σε μορφή κατάλληλη αλλά και αποδοτική για τους αλγορίθμους που σκοπεύουμε να εφαρμόσουμε στη συνέχεια της ανάλυσής μας. Ο μετασχηματισμός των δεδομένων είναι συχνά απαραίτητος για την ανάδειξη νέων ιδιαιτεροτήτων ή και διαφορετικών οπτικών γωνιών θέασης του συνόλου των δεδομένων μας. Η φύση των αλγορίθμων εκμάθησης ή η μορφή των δεδομένων μας είναι τα δύο στοιχεία που κυρίως υποδεικνύουν διάφορους μετασχηματισμούς οι οποίοι μπορεί να είναι μαθηματικοί ή λογικοί ή ακόμα και μετασχηματισμοί βασισμένοι στη γνώση του πεδίου.

Στις περισσότερες εφαρμογές του data mining, ο αριθμός των χαρακτηριστικών είναι υπερβολικά μεγάλος για την αποτελεσματική διαχείρισή τους από τους αλγορίθμους που εφαρμόζονται. Τα περισσότερα χαρακτηριστικά είναι ευκρινώς μη συσχετιζόμενα ή περιττά και κατά συνέπεια κρίνεται απαραίτητο να επιλεγεί ένα υποσύνολο δεδομένων προς χρήση στη διαδικασία εκμάθησης. Βέβαια, οι μέθοδοι εκμάθησης εκ φύσεως επιτελούν τη διαδικασία επιλογής των κατάλληλων χαρακτηριστικών και απόρριψης των υπολοίπων, όμως η αποδοτικότητά τους μπορεί να βελτιωθεί κατά πολύ μέσω μιας προεπιλογής που πραγματοποιείται από τον πειραματιστή χειροκίνητα ή μέσω αυτοματοποιημένων μεθόδων.

Η επιλογή του υποσυνόλου των χαρακτηριστικών μπορεί να υλοποιηθεί μέσω δύο βασικών κατηγοριών μεθόδων, αυτών της διήθησης (filter) και της ενσωμάτωσης (wrapper).

Η πρώτη κατηγορία μεθόδων είναι αυτή της διήθησης, η οποία πραγματοποιεί ανεξάρτητη αποτίμηση του υποσυνόλου βασισμένη στα γενικά χαρακτηριστικά των δεδομένων. Η δεύτερη κατηγορία μεθόδων είναι αυτή της ενσωμάτωσης, στην οποία η διαδικασία επιλογής χαρακτηριστικών προσκολλάται στη διαδικασία εκπαίδευσης και η αποτίμηση του υποσυνόλου γίνεται βάσει της τελικής απόδοσης του αλγορίθμου εκμάθησης με τελικό κριτή αυτής της διαδικασίας να είναι το ύψος του σφάλματος σε νέα δεδομένα. Στην παρούσα εργασία θα αναφερθούμε διεξοδικά και θα εφαρμόσουμε αρκετές τεχνικές διήθησης. Για το λόγο αυτό δεν θα αναφερθούμε με περισσότερες λεπτομέρειες στις τεχνικές ενσωμάτωσης.

Στο πάνελ “Filter”, κλικάρουμε την επιλογή “Choose” και ανοίγει ένα παράθυρο με μια λίστα διαθέσιμων φίλτρων, όπως φαίνεται παρακάτω στην Εικόνα 21.



Εικόνα 18: Φίλτρα στο Weka

8.4 Διακριτοποίηση των δεδομένων

Μία άλλη διαδικασία που πολύ συχνά απαιτείται να εφαρμόσουμε στο στάδιο της προεπεξεργασίας των δεδομένων, και εφαρμόζουμε στην παρούσα εργασία είναι η διακριτοποίηση των χαρακτηριστικών. Η διαδικασία αυτή είναι απολύτως αναγκαία στην περίπτωση που κάποια χαρακτηριστικά είναι αριθμητικά αλλά η επιλεγμένη μέθοδος μάθησης μπορεί να διαχειριστεί μόνο ρητά χαρακτηριστικά. **Ακόμη και οι μέθοδοι που μπορούν να χειριστούν αριθμητικά χαρακτηριστικά συχνά παράγουν καλύτερα αποτελέσματα ή λειτουργούν ταχύτερα αν τα χαρακτηριστικά που δέχονται ως είσοδο έχουν διακριτοποιηθεί.**

Για το πρόβλημα της διακριτοποίησης υπάρχουν δύο βασικές προσεγγίσεις. Η μία είναι να ποσοτικοποιήσουμε κάθε χαρακτηριστικό χωρίς να γνωρίζουμε την τάξη στην οποία ανήκει κάθε υπόδειγμα του συνόλου εκπαίδευσης. Σε αυτήν την περίπτωση έχουμε διακριτοποίηση χωρίς επίβλεψη. Η δεύτερη προσέγγιση, αυτή της διακριτοποίησης με επίβλεψη, είναι να λάβουμε υπόψη μας και τη γνωστή τάξη των χαρακτηριστικών όταν διακριτοποιούμε.

Για να διακριτοποιήσουμε χωρίς επίβλεψη ένα αριθμητικό χαρακτηριστικό, ένας εύκολος τρόπος είναι να διαιρέσουμε το εύρος του σε προκαθορισμένο αριθμό ίσων διαστημάτων. Αυτή είναι μία διαδικασία που συνήθως γίνεται κατά τη συλλογή των δεδομένων. Όμως, όπως συμβαίνει σε κάθε μέθοδο διακριτοποίησης χωρίς επίβλεψη, υπάρχει ο κίνδυνος των λανθασμένων διακρίσεων χρησιμοποιώντας πολύ μεγάλες διαμερίσεις ή ατυχείς επιλογές των φραγμάτων που ομαδοποιούν μαζί πολλά υποδείγματα διαφορετικών κλάσεων.

Στην περίπτωση που εφαρμόζουμε διαμέριση σε διαστήματα σταθερού εύρους (equal-interval binning) κάποια από αυτά μπορεί να περιέχουν πολλά υποδείγματα και κάποια να μην περιέχουν κανένα, κάτι το οποίο πιθανόν να έχει ως συνέπεια να μην μπορούμε να κατασκευάσουμε ένα καλό μοντέλο για τα δεδομένα μας. Γι' αυτό συχνά είναι προτιμητέο να χρησιμοποιήσουμε διαστήματα διαφορετικού μεγέθους, επιλέγοντάς τα έτσι ώστε σε κάθε ένα από αυτά να περιέχεται ο ίδιος αριθμός υποδειγμάτων του συνόλου εκπαίδευσης. Αυτή η μέθοδος λέγεται διακριτοποίηση σε διαστήματα σταθερής συχνότητας (equal frequency binning) ή αντιστάθμιση ιστογράμματος (histogram equalization) και διαιρεί το εύρος των χαρακτηριστικών σε προκαθορισμένο αριθμό διαστημάτων με βάση την κατανομή των υποδειγμάτων.

Μία από τις καλύτερες μεθόδους διακριτοποίησης με επίβλεψη είναι αυτή στην οποία εφαρμόζουμε τη μέθοδο της εντροπίας και η οποία θεωρείται ως η πλέον αποδοτική και αξιόπιστη. Σύμφωνα με αυτή τη μέθοδο κατασκευάζουμε ένα δέντρο απόφασης για τη διακριτοποίηση του χαρακτηριστικού χρησιμοποιώντας ως κριτήριο διαχωρισμού των τιμών του αρχικού συνόλου το μέτρο της εντροπίας. Ως κριτήριο διακοπής της διαδικασίας διακριτοποίησης χρησιμοποιούμε την αρχή ελαχίστου τετραγώνου μήκους περιγραφής (minimum description length-mdl).

Σε άλλες μεθόδους διακριτοποίησης με επίβλεψη μπορούμε να αντικαταστήσουμε τη διαδικασία της από πάνω προς τα κάτω (top-down) διαμέρισης των διαστημάτων μέχρι κάποιο κριτήριο διακοπής να ικανοποιηθεί, με μια προσέγγιση από κάτω προς τα πάνω (bottom-up), όπου πρώτα τοποθετούμε κάθε υπόδειγμα στο διάστημα και μετά εξετάζουμε αν θα συγχωνεύσουμε γειτονικά διαστήματα. Μπορούμε να χρησιμοποιήσουμε κάποιο στατιστικό κριτήριο για να δούμε ποια δύο διαστήματα είναι τα καλύτερα για να συγχωνευθούν και να τα συγχωνεύσουμε αν το στατιστικό υπερβαίνει ένα συγκεκριμένο επίπεδο εμπιστοσύνης, επαναλαμβάνοντας τη διαδικασία μέχρι καμία πιθανή συγχώνευση να μην υπερβαίνει τον έλεγχο. Ένα τεστ που χρησιμοποιείται σε τέτοιες περιπτώσεις είναι το test X^2.

Μία άλλη προσέγγιση είναι να μετρήσουμε τον αριθμό των λαθών που γίνονται σε μια διαδικασία διακριτοποίησης, όταν προβλέπουμε την τάξη κάθε υποδείγματος του συνόλου εκπαίδευσης υποθέτοντας ότι κάθε διάστημα λαμβάνει την πλειοψηφούσα τάξη.

Τέλος, το αντίστροφο πρόβλημα, δηλαδή αυτό στο οποίο επιθυμούμε να μετατρέψουμε ρητά χαρακτηριστικά σε αριθμητικά μπορεί ναι μεν να προκύψει όμως είναι λιγότερο συχνό και επιτυγχάνεται κωδικοποιώντας υποσύνολο των ονομαστικών χαρακτηριστικών ως δυαδικά.

Αξίζει να σημειωθεί ότι στο στάδιο της προεπεξεργασίας των δεδομένων πρέπει να εφαρμόζεται όπου χρειάζεται και ο καθαρισμός τους από άτυπες τιμές ή άλλες ανωμαλίες της βάσης δεδομένων, διαδικασία η οποία επιτυγχάνεται συνήθως με μεθόδους οπτικοποίησης.

8.5 Φίλτρα και προεπεξεργασία των χαρακτηριστικών

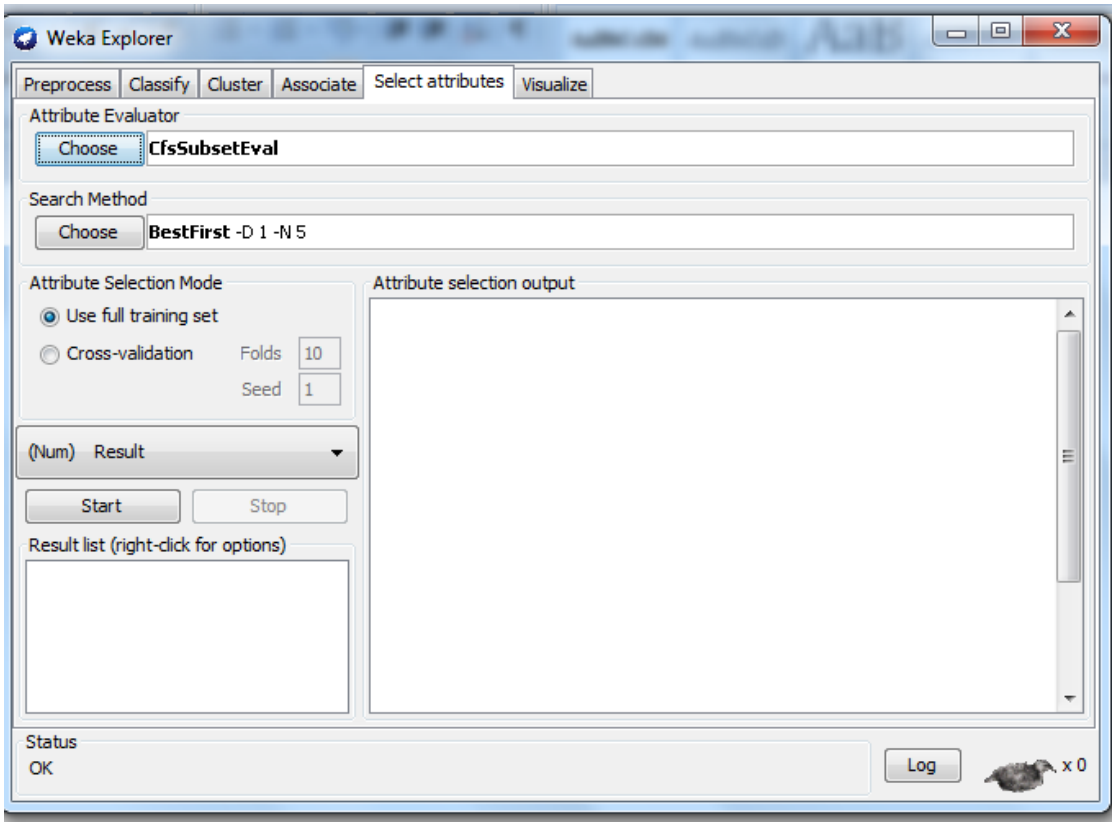
Το Weka δίνει τη δυνατότητα στο χρήστη να εφαρμόσει όσα προαναφέραμε σχετικά με την προεπεξεργασία των δεδομένων με τα φίλτρα που διαθέτει αλλά και με την ξεχωριστή επιλογή "attribute selection". Ανοίγοντας το σύνολο δεδομένων μας, μπορούμε να δούμε τα φίλτρα στο πεδίο "filter" επιλέγοντας το tab "choose". Έτσι προκύπτει η λίστα φίλτρων η οποία αποτελείται από φίλτρα χαρακτηριστικών, με και χωρίς επίβλεψη, καθώς και από φίλτρα υποδειγμάτων, δηλαδή φίλτρα που εφαρμόζονται σε υποδείγματα, με και χωρίς επίβλεψη.

Ενδεικτικά, αναφέρουμε κάποια ευρέως χρησιμοποιούμενα όπως το φίλτρο "Remove" στο φάκελο "unsupervised attribute filters", το οποίο αφαιρεί τα χαρακτηριστικά που επιλέγουμε, το φίλτρο "Nominal to Binary", στον ίδιο φάκελο, το οποίο μετατρέπει τα ονομαστικά χαρακτηριστικά σε δυαδικά, αντικαθιστώντας κάθε χαρακτηριστικό k τιμών με k δυαδικά χαρακτηριστικά χρησιμοποιώντας μία απλή κωδικοποίηση, καθώς και το φίλτρο "Discretize", το οποίο διακριτοποιεί τα χαρακτηριστικά χρησιμοποιώντας τη μέθοδο "equal frequency binning" μέθοδο που περιγράψαμε παραπάνω. Επιλέγοντας κάποιο από τα φίλτρα και κάνοντας κλικ πάνω σε αυτό ανοίγει ένα νέο παράθυρο με πληροφορίες για το φίλτρο και με διάφορες επιλογές για την εφαρμογή του.

Επίσης, μπορούμε να επιλέξουμε το tab "select attributes" και να αποκτήσουμε πρόσβαση σε πολλές μεθόδους επιλογής χαρακτηριστικών. Στο panel αυτό μπορούμε να επιλέξουμε την επιθυμητή μέθοδο εκτίμησης των χαρακτηριστικών από μία λίστα τέτοιων μεθόδων που προκύπτει επιλέγοντας το tab "choose" στο πεδίο "attribute evaluator", καθώς και τη μέθοδο αναζήτησης από μία λίστα που προκύπτει επιλέγοντας "choose" στο πεδίο "search method".

9. Επιλογή χαρακτηριστικών στο Weka

Το λογισμικό Weka διαθέτει ξεχωριστό tab με τίτλο "select attributes" το οποίο εφαρμόζει διάφορους αλγόριθμους επιλογής χαρακτηριστικών, και πρωταρχικό στόχο έχει τη μείωση της διάστασης της βάσης δεδομένων. Πιο συγκεκριμένα, όπως βλέπουμε και στην παρακάτω εικόνα, το παράθυρο που ανοίγει από το tab "select attributes", διαθέτει δύο υποπεδία επιλογών. Το πρώτο είναι το υποπεδίο "attribute evaluator" και το δεύτερο υποπεδίο το "search method".



Εικόνα 19: Επιλογή χαρακτηριστικών στο Weka

Για την επιλογή των χαρακτηριστικών θα πρέπει να καθορίσουμε τη μέθοδο εκτίμησης χαρακτηριστικών, από τη λίστα μεθόδων που παρέχει το πεδίο "attribute evaluator", καθώς και τη μέθοδο αναζήτησης, από το πεδίο "search method".

9.1 Μέθοδοι εκτίμησης χαρακτηριστικών

Για την εκτίμηση των χαρακτηριστικών (attribute evaluator), μπορούμε να επιλέξουμε ανάμεσα από τις παρακάτω τέσσερις μεθόδους

1. CfsSubsetEval

2. ClassifierSubsetEval

3. ConsistencySubsetEval

4. WrapperSubsetEval

οι οποίες εκτιμούν ένα υποσύνολο χαρακτηριστικών. Οι μέθοδοι αυτές παίρνουν ένα υποσύνολο χαρακτηριστικών και επιστρέφουν ένα αριθμητικό μέτρο που καθοδηγεί την αναζήτηση.

9.1.1 CfsSubsetEval

Η πρώτη μέθοδος "CfsSubsetEval", θεωρεί την τιμή πρόβλεψης για κάθε χαρακτηριστικό ξεχωριστά, μαζί με το βαθμό πλεονασμού μεταξύ των χαρακτηριστικών. Ουσιαστικά, η μέθοδος "CfsSubsetEval" αξιολογεί την προβλεπτική ικανότητα κάθε χαρακτηριστικού ξεχωριστά και το βαθμό πλεονασμού μεταξύ αυτών, προτιμώντας τα σύνολα χαρακτηριστικών που έχουν υψηλή συσχέτιση με την τάξη, αλλά έχουν χαμηλή ενδοσυσχέτιση (intercorrelation). Η μέθοδος αυτή ανήκει στην κατηγορία των μεθόδων διήθησης (filter) στις οποίες αναφερθήκαμε παραπάνω.

Μία επιλογή του αλγορίθμου (locally predictive) προσθέτει επαναληπτικά χαρακτηριστικά που έχουν την υψηλότερη συσχέτιση με την τάξη, εφόσον το σύνολο δεν περιέχει ήδη ένα χαρακτηριστικό του οποίου η συσχέτιση με το ζητούμενο χαρακτηριστικό να είναι ακόμα υψηλότερη. Η δεύτερη επιλογή του αλγορίθμου (missingSeparate) δίνει τη δυνατότητα να αντιμετωπιστεί μία ελλείπουσα τιμή ως ξεχωριστή τιμή ή εναλλακτικά ο αριθμός των ελλειπουσών τιμών να κατανεμηθεί μεταξύ άλλων τιμών ανάλογα με τη συχνότητά τους.

9.1.2 ClassifierSubsetEval

Η μέθοδος "ClassifierSubsetEval" χρησιμοποιεί έναν ταξινομητή, δηλαδή μία μέθοδο ταξινόμησης με σκοπό να εκτιμήσει σύνολα χαρακτηριστικών στα δεδομένα εκπαίδευσης ή σε ξεχωριστά σύνολα ελέγχου παρακράτησης. Η μέθοδος αυτή ανήκει

στην κατηγορία των μεθόδων ενσωμάτωσης (wrapper) στις οποίες αναφερθήκαμε παραπάνω.

9.1.3 ConsistencySubsetEval

Η μέθοδος "ConsistencySubsetEval" προβάλλει ένα σύνολο εκπαίδευσης πάνω σε σύνολο χαρακτηριστικών και μέτρο συνέπειας στις τιμές της τάξης. Συγκεκριμένα, η μέθοδος αυτή εκτιμά σύνολα χαρακτηριστικών από το βαθμό συνέπειας στις τιμές της τάξης όταν τα υποδείγματα εκπαίδευσης προβάλλονται πάνω σε αυτό το σύνολο. Η συνέπεια οποιουδήποτε συνόλου δεν μπορεί ποτέ να είναι χαμηλότερη από ότι η συνέπεια ολόκληρου του συνόλου των χαρακτηριστικών, και επομένως η συνηθισμένη πρακτική είναι να εφαρμόσουμε αυτό τον εκτιμητή υποσυνόλων από κοινού με την εξαντλητική (exhaustive search) ή τυχαία μέθοδο αναζήτησης (random search), οι οποίες αναζητούν το μικρότερο υποσύνολο με συνέπεια ίδια με αυτή που έχει ολόκληρο το σύνολο των χαρακτηριστικών. Η μέθοδος αυτή, όπως και η μέθοδος "CfsSubsetEval" ανήκει στην κατηγορία των μεθόδων διήθησης (filter).

9.1.4 WrapperSubsetEval

Η μέθοδος "WrapperSubsetEval" ανήκει στις μεθόδους ενσωμάτωσης, όπως φανερώνει και ο τίτλος της, και χρησιμοποιεί έναν ταξινομητή για την εκτίμηση υποσυνόλων χαρακτηριστικών σε συνδυασμό με τη μέθοδο διασταυρωμένης επικύρωσης (cross validation) έτσι ώστε να εκτιμήσει την ακρίβεια του «σχήματος» εκμάθησης για κάθε σύνολο.

9.2 Μέθοδοι αναζήτησης

Για την εκτίμηση των χαρακτηριστικών, οι προαναφερθείσες τέσσερις μέθοδοι που περιγράψαμε παραπάνω, θα πρέπει να συνδυαστούν με μία από τις επτά παρακάτω μεθόδους αναζήτησης (search method):

1. BestFirst
2. ExhaustiveSearch
3. GeneticSearch
4. GreedyStepwise
5. RaceSearch
6. RandomSearch
7. RankSearch

Οι μέθοδοι αναζήτησης διασχίζουν το χώρο των χαρακτηριστικών και βρίσκουν ένα καλό υποσύνολο. Η ποιότητα του υποσυνόλου μετράται σύμφωνα με τη μέθοδο επιλογής χαρακτηριστικών που καθορίσαμε νωρίτερα.

9.2.1 BestFirst

Η πρώτη μέθοδος "BestFirst" εκτελεί μία «άπληστη αναρρίχηση λόφου» (greedy hill climbing) με οπισθοδρόμηση και μπορούμε να ορίσουμε τον αριθμό των διαδοχικών, μη βελτιωμένων κόμβων μέχρι την οπισθοδρόμηση. Η διαδικασία αυτή μπορεί να αρχίσει προς τα εμπρός από το άδειο σύνολο των χαρακτηριστικών, δηλαδή το κενό σύνολο, προς τα πίσω από ολόκληρο το σύνολο των χαρακτηριστικών ή εναλλακτικά να ξεκινήσει από ένα ενδιάμεσο σημείο και να ψάχνει και προς τις δύο κατευθύνσεις θεωρώντας όλες τις πιθανές προσθαφαιρέσεις των χαρακτηριστικών.

9.2.2 GreedyStepwise

Η μέθοδος "GreedyStepwise" αναζητά και αυτή άπληστα στο χώρο των υποσυνόλων των χαρακτηριστικών, και όπως η "BestFirst" μπορεί να εξελιχθεί προς τα εμπρός ή προς τα πίσω. Όμως , σε αντίθεση με την "BestFirst", δεν εφαρμόζει οπισθοδρόμηση αλλά τερματίζεται μόλις η πρόσθεση ή η αφαίρεση οποιουδήποτε εναπομείναντος χαρακτηριστικού μειώνει την εκτίμηση. Επιπλέον, μπορεί να ταξινομήσει τα χαρακτηριστικά διασχίζοντας το χώρο και εγγράφοντας τη σειρά με την οποία επιλέγονται τα χαρακτηριστικά.

9.2.3 RaceSearch

Η μέθοδος "RaceSearch" χρησιμοποιείται μαζί με τη μέθοδο εκτίμησης χαρακτηριστικών "ClassifierSubsetEval" και υπολογίζει το σφάλμα διασταυρωμένης επικύρωσης συγκρίνοντας τα υποσύνολα χαρακτηριστικών με τη χρήση της αναζήτησης ανταγωνισμού (race search). Μπορούμε να χρησιμοποιήσουμε τέσσερα σχήματα αναζήτησης, την προς τα εμπρός επιλογή (forward selection), την προς τα πίσω απαλοιφή (backward elimination), την αναζήτηση σχημάτων (schemata search) και τον ανταγωνισμό κατάταξης (rank racing). Στην τελευταία περίπτωση, ένας ξεχωριστός εκτιμητής χαρακτηριστικών, ο οποίος μπορεί και αυτός να επιλεχθεί, χρησιμοποιείται για να παραχθεί μία αρχική κατάταξη. Χρησιμοποιώντας την προς τα εμπρός επιλογή, μπορούμε να δημιουργήσουμε μία λίστα καταταγμένων χαρακτηριστικών συνεχίζοντας τις συγκρίσεις, τον ανταγωνισμό μεταξύ των χαρακτηριστικών, μέχρι να επιλεγούν όλα τα χαρακτηριστικά και κατατάσσοντάς τα με τη σειρά που αυτά προστίθενται.

9.2.4 GeneticSearch

Η μέθοδος "GeneticSearch" χρησιμοποιεί για την αναζήτηση έναν απλό γενετικό αλγόριθμο. Οι παράμετροι που μπορούμε να καθορίσουμε αφορούν το μέγεθος του

πληθυσμού, τον αριθμό των γενεών, και τις πιθανότητες διασταύρωσης και μετάλλαξης. Επίσης, μπορούμε να καθορίσουμε μία λίστα με δείκτες χαρακτηριστικών ως αρχικό σημείο η οποία γίνεται μέλος του αρχικού πληθυσμού.

9.2.5 RandomSearch

Η μέθοδος "RandomSearch" ψάχνει τυχαία το χώρο των υποσυνόλων των χαρακτηριστικών. Αν ένα αρχικό σύνολο παρέχεται, τότε αναζητούνται υποσύνολα που να βελτιώνουν ή να ισούνται με το αρχικό σημείο και να έχουν λιγότερα, ή τον ίδιο αριθμό χαρακτηριστικών. Διαφορετικά, η μέθοδος ξεκινά από ένα τυχαίο σημείο και αναφέρει τα καλύτερα σύνολα υποσύνολα που βρίσκει.

9.2.6 ExhaustiveSearch

Η μέθοδος "ExhaustiveSearch" ψάχνει το χώρο των υποσυνόλων των χαρακτηριστικών ξεκινώντας από ένα άδειο σύνολο και αναφέροντας τα καλύτερα υποσύνολα που βρίσκει. Αν δίνεται ένα αρχικό σύνολο, η μέθοδος ψάχνει προς τα πίσω από αυτό το αρχικό σημείο και αναφέρει το μικρότερο υποσύνολο με την καλύτερη εκτίμηση.

9.3 Συνδυασμός μεθόδων αναζήτησης/εκτίμησης χαρακτηριστικών

Στις παραπάνω ενότητες 9.1 και 9.2, αναφερθήκαμε σε τέσσερις μεθόδους εκτίμησης χαρακτηριστικών που συνδυάζονται με επτά μεθόδους αναζήτησης με σκοπό την εκτίμηση των χαρακτηριστικών.

Μία γρηγορότερη ενδεχομένως, αλλά λιγότερο ακριβής προσέγγιση είναι να εκτιμήσουμε τα χαρακτηριστικά ξεχωριστά και να τα ταξινομήσουμε απορρίπτοντας τα χαρακτηριστικά που βρίσκονται κάτω από ένα καθορισμένο σημείο απόρριψης. Αυτή η προσέγγιση επιτυγχάνεται επιλέγοντας μία από τις οχτώ μεθόδους εκτίμησης «ξεχωριστών» χαρακτηριστικών που παρέχονται στο πεδίο "Attribute subset evaluator", σε συνδυασμό με τη μέθοδο κατάταξης "Ranker", του πεδίου "search method".

Η μέθοδος "Ranker" είναι ένα σχήμα κατάταξης των χαρακτηριστικών για ξεχωριστά χαρακτηριστικά που τα ταξινομεί με βάση τις χωριστές εκτιμήσεις τους και πρέπει να χρησιμοποιηθεί μαζί με έναν από τους εκτιμητές των ξεχωριστών χαρακτηριστικών, και όχι των εκτιμητών των υποσυνόλων των χαρακτηριστικών. Η μέθοδος όχι μόνο κατατάσσει τα χαρακτηριστικά αλλά εφαρμόζει και επιλογή χαρακτηριστικών αφαιρώντας εκείνα τα χαρακτηριστικά που είναι χαμηλά στην κατάταξη.

10. Ταξινόμηση στο Weka

10.1 Εισαγωγή

Η ταξινόμηση αποτελεί μία από τις βασικές τεχνικές εξόρυξης δεδομένων. Βασίζεται στην εξέταση των χαρακτηριστικών ενός νέου αντικειμένου (μη κατηγοριοποιημένου) το οποίο με βάση τα χαρακτηριστικά αυτά αντιστοιχίζεται σε ένα προκαθορισμένο σύνολο κλάσεων. Η διαδικασία της κατηγοριοποίησης χαρακτηρίζεται από ένα σαφή καθορισμό των κατηγοριών, και το σύνολο που χρησιμοποιείται για την εκπαίδευση του μοντέλου αποτελείται από προκαθορισμένα υποδείγματα. Η ταξινόμηση δεδομένων είναι μια διαδικασία η οποία βρίσκει τις κοινές ιδιότητες μεταξύ ενός συνόλου αντικειμένων σε μια βάση δεδομένων και ταξινομεί τα αντικείμενα αυτά σε διαφορετικές κλάσεις σύμφωνα με ένα μοντέλο ταξινόμησης.

Η ταξινόμηση βρίσκει ποικίλες εφαρμογές και αποτελεί αντικείμενο μελέτης για τη στατιστική, τη μηχανική μάθηση και φυσικά την εξόρυξη δεδομένων. Πρόκειται για μάθηση με επίβλεψη καθώς οι ομάδες ταξινόμησης είναι εκ των προτέρων γνωστές και το πραγματικό αποτέλεσμα κάθε υποδείγματος είναι επίσης γνωστό. Επομένως, είναι δυνατόν να μετράμε το βαθμό αξιοπιστίας σε μη χρησιμοποιούμενα για τη διαμόρφωση της αντίληψης δεδομένα.

Η τυπική προσέγγιση που χρησιμοποιούν οι τεχνικές ταξινόμησης είναι η δημιουργία ενός μοντέλου μέσω της αξιολόγησης του συνόλου δεδομένων εκπαίδευσης και η εφαρμογή του μοντέλου σε νέα δεδομένα. Οι πιο κοινές τεχνικές είναι τα δέντρα αποφάσεων, τεχνικές βασισμένες στην απόσταση, κλασικές στατιστικές τεχνικές, νευρωνικά δίκτυα κ.α.

10.2 Δέντρα αποφάσεων

Στην υποενότητα αυτή θα εξετάσουμε την ταξινόμηση με τη βοήθεια των δέντρων αποφάσεων. Τα δέντρα απόφασης ή ταξινόμησης (decision ή classification trees) είναι μια μέθοδος που προσφέρει ακρίβεια, σαφήνεια, ταχύτητα και η οποία έχει το πλεονέκτημα ότι η γνώση που προκύπτει γίνεται εύκολα κατανοητή. Πρόκειται για μια τεχνική στην οποία η συνάρτηση εκπαίδευσης αναπαρίσταται μέσω ενός δέντρου απόφασης. Πιο συγκεκριμένα η τεχνική εξελίσσεται ικανοποιώντας τη συνθήκη: διαίρεσε το σύνολο των υποδειγμάτων σε υποσύνολα, με βάση κάποιο χαρακτηριστικό, με κριτήριο το κέρδος πληροφορίας, εξασφαλίζοντας ότι κάθε υποσύνολο θα έχει κατά το δυνατό την ίδια τιμή για αυτό το χαρακτηριστικό.

Τα δέντρα απόφασης ταξινομούν στιγμιότυπα διατάσσοντάς τα από τη ρίζα σε κάποιο κόμβο-φύλλο. Κάθε κόμβος αναφέρεται στην εξέταση κάποιου χαρακτηριστικού

(attribute) του στιγμιότυπου και κάθε αμέσως επόμενος κλάδος αντιστοιχεί σε μια από τις πιθανές τιμές του στιγμιότυπου. Γενικά, κάθε διαδρομή από τη ρίζα του δέντρου σε κάποιο φύλλο αντιστοιχεί σε μια ένωση των χαρακτηριστικών, ενώ κάθε διακλάδωση σε μια διάζευξη αυτών των συζεύξεων. Οι κόμβοι σε ένα δέντρο απόφασης υλοποιούν έναν έλεγχο για την τιμή ενός συγκεκριμένου χαρακτηριστικού. Συνήθως η τιμή του χαρακτηριστικού συγκρίνεται με μία σταθερά, όμως υπάρχουν περιπτώσεις όπου συγκρίνονται οι τιμές δύο χαρακτηριστικών ή χρησιμοποιείται μία συνάρτηση με ένα ή περισσότερα χαρακτηριστικά. Τα φύλλα εκχωρούν σε όλα τα υποδείγματα που καταλήγουν σε αυτά ταξινόμηση, ή σύνολο ταξινομήσεων ή κατανομές πιθανότητας. Για να ταξινομήσουμε υποδείγματα άγνωστης τάξης ακολουθούμε την πορεία από την αρχή (ρίζα) του δέντρου σύμφωνα με τις τιμές των ελέγχων στους κόμβους και όταν φτάσουμε σε κάποιο φύλλο το υπόδειγμα ταξινομείται σύμφωνα με την τάξη του φύλλου.

Στην περίπτωση που έχουμε ονομαστικά χαρακτηριστικά ο αριθμός των κλάδων μετά από ένα κόμβο ισούται συνήθως με τον αριθμό των διακριτών τιμών του χαρακτηριστικού και κατά συνέπεια κάθε χαρακτηριστικό δεν ελέγχεται περισσότερες από μία φορές. Εναλλακτικά, οι τιμές των χαρακτηριστικών διαιρούνται συνήθως σε δύο υποσύνολα και τα κλαδιά του δέντρου ελέγχουν σε ποιο από τα δύο υποσύνολα ανήκει η συγκεκριμένη τιμή, οπότε σε αυτή την περίπτωση μπορεί το χαρακτηριστικό να ελεγχθεί περισσότερες από μία φορές.

Στην περίπτωση αριθμητικών χαρακτηριστικών, σε κάθε κόμβο εκτελείται η σύγκριση της τιμής του χαρακτηριστικού με μία σταθερά, όταν πρόκειται για ακέραιο αριθμό, ή με ένα σύνολο τιμών, όταν πρόκειται για πραγματικό αριθμό. Επομένως, κάθε χαρακτηριστικό μπορεί να ελεγχθεί περισσότερες από μία φορές. Μια παραλλαγή της παραπάνω διαδικασίας προκύπτει αν διαιρέσουμε σε τρία ή περισσότερα υποσύνολα και συγκρίνουμε την τιμή του χαρακτηριστικού με αυτά τα υποσύνολα.

Οι γνωστότεροι αλγόριθμοι ταξινόμησης (ID3, C4.5,CART) χρησιμοποιούν μια top-down, εξαντλητική αναζήτηση στο χώρο των πιθανών δέντρων απόφασης. Αρχίζουν με ένα κενό δέντρο και προοδευτικά θέτουν πιο περίπλοκες υποθέσεις με στόχο την εύρεση ενός δέντρου που ταξινομεί σωστά τα δεδομένα εκπαίδευσης.

Έτσι η διαδικασία κατασκευής δέντρου απόφασης έχει ως εξής:

1. Επιλογή χαρακτηριστικού για τη θέση του αρχικού κόμβου (ρίζας) και δημιουργία κλάδων για κάθε πιθανή τιμή του χαρακτηριστικού.
2. Διάσπαση υποδειγμάτων σε υποσύνολα, ένα για κάθε κλάδο που επεκτείνεται από τη ρίζα.
3. Επανάληψη των παραπάνω για κάθε κλάδο με χρήση μόνο του υποσυνόλου των υποδειγμάτων κάθε κλάδου.
4. Ολοκλήρωση της διαδικασίας όταν όλα τα υποδείγματα σε ένα κόμβο ανήκουν στην ίδια τάξη.

Όταν έχει ολοκληρωθεί η διαδικασία ανακάλυψης γνώσης με χρήση του αλγόριθμου, τότε το δένδρο μπορεί να αναπαρασταθεί ως σύνολο κανόνων της μορφής:
«Εάν <ΣΥΝΟΛΟ ΣΥΝΘΗΚΩΝ> τότε <ΣΥΜΠΕΡΑΣΜΑ> .

10.3 Οι αλγόριθμοι ID-3, J48 και CART

10. 3. 1 Ο αλγόριθμος ID-3

Η μέθοδος ID-3 αναπτύχθηκε το 1975 από τον Ross Quinlan και βασική της ιδέα είναι η κατασκευή δέντρου απόφασης μέσω μίας άπληστης, από πάνω προς τα κάτω αναζήτησης του δοθέντος συνόλου έτσι ώστε να ελεγχθεί κάθε χαρακτηριστικό σε κάθε κόμβο του δέντρου. Για να επιλέξουμε το χαρακτηριστικό που είναι το καταλληλότερο για την ταξινόμηση του δοθέντος συνόλου θα πρέπει να ελαχιστοποιήσουμε την έννοια του κέρδους πληροφορίας. Για να ποσοτικοποιήσουμε την επίδραση που προκαλεί ο διαχωρισμός του συνόλου δεδομένων χρησιμοποιώντας ένα συγκεκριμένο χαρακτηριστικό, χρησιμοποιούμε το κέρδος πληροφορίας (information gain–IG) το οποίο υπολογίζει την αναμενόμενη μείωση της εντροπίας που θα προκύψει από το διαχωρισμό αυτό. Η εντροπία αποτελεί ένα μέγεθος το οποίο μετρά την αβεβαιότητα των δεδομένων και η μονάδα μέτρησής του είναι τα bits. Το κέρδος πληροφορίας ισούται με τη συνολική εντροπία για ένα χαρακτηριστικό αν για κάθε τιμή του χαρακτηριστικού υπάρχει μοναδική ταξινόμηση ως προς την τάξη.

10. 3. 2 Ο αλγόριθμος J48

Μία βελτιωμένη εκδοχή του ID-3 είναι ο αλγόριθμος C4.5, ο οποίος εφαρμόζεται στο Weka από τον J48. Ο αλγόριθμος J48 επεκτείνει την ταξινόμηση από ρητά σε αριθμητικά δεδομένα και διαχειρίζεται καλύτερα τις ελλείπουσες τιμές και τα δεδομένα με θόρυβο. Ένα μειονέκτημα του αλγορίθμου ID-3 είναι το γεγονός ότι μεροληπτεί υπέρ των χαρακτηριστικών με μεγάλο αριθμό τιμών με πιθανή συνέπεια την υπερπροσαρμογή (overfitting), δηλαδή την επιλογή υποβέλτιστου χαρακτηριστικού. Ο αλγόριθμος J48 χρησιμοποιεί το λόγο κέρδους (gain ratio) αντί για το κέρδος πληροφορίας, και αυτή η τροποποίηση της συνάρτησης καταλληλότητας αντιμετωπίζει το πρόβλημα της μεροληψίας που προαναφέραμε.

10. 3. 3 Ο αλγόριθμος CART

Ο CART αλγόριθμος βασίζεται στη θεωρία των δέντρων ταξινόμησης και παλινρόμησης που διατυπώθηκε από τους Breiman et al. (1984). Ο CART αλγόριθμος διαχωρίζει τα δεδομένα σε δύο υποσύνολα έτσι ώστε οι καταχωρήσεις μέσα σε κάθε υποσύνολο να είναι περισσότερο ομοιογενείς από ότι μέσα στο προηγούμενο υποσύνολο. Είναι μια επαναληπτική διαδικασία – κάθε ένα από αυτά τα δύο υποσύνολα διαχωρίζεται ξανά, και η διαδικασία επαναλαμβανεται μέχρι να επιτευχθεί το κριτήριο της ομοιογένειας ή μέχρι κάποιο άλλο κριτήριο διακοπής διαδικασίας να ικανοποιείται (όπως συμβαίνει σε όλες τις μέθοδους κατασκευής δέντρων). Το ίδιο πεδίο πρόβλεψης μπορεί να χρησιμοποιηθεί αρκετές φορές σε διαφορετικά επίπεδα στο δέντρο. Ο CART αλγόριθμος είναι αρκετά ευέλικτος. Ο CART αλγόριθμος

δουλεύει επιλέγοντας ένα διαχωρισμό σε κάθε κόμβο τέτοιο ώστε κάθε θυγατρικός κόμβος που δημιουργείται από τον διαχωρισμό να είναι πιο καθαρός από τον γεννήτορα-μητρικό κόμβο του.

10.4 Αλγόριθμοι αυτόματης παραγωγής κανόνων ταξινόμησης

Η τάση για δημιουργία κανόνων με αυτοματοποιημένο τρόπο αναπτύσσεται ταχύτατα. Οι κανόνες ταξινόμησης είναι ένας άλλος τρόπος απεικόνισης της ταξινόμησης και αποτελούν μία δημοφιλή εναλλακτική στα δέντρα αποφάσεων. Η προϋπόθεση (antecedent) ενός κανόνα είναι ένα σύνολο ελέγχων, όμοιοι με τους ελέγχους που γίνονται στους κόμβους ενός δέντρου απόφασης, και οι οποίοι έλεγχοι συμπλέκονται μεταξύ τους σε λογικές πράξεις, με πιο συνηθισμένη τη λογική σύζευξη (και). Το συμπέρασμα (consequent) ενός κανόνα είναι η εκχώρηση ταξινόμησης, ή συνόλου ταξινομήσεων ή κατανομής πιθανότητας, στο υπόδειγμα που καλύπτεται από αυτό τον κανόνα. Το πρόβλημα που προκύπτει από τους κανόνες ταξινόμησης είναι ότι κάποιες φορές οι υποδείξεις των κανόνων είναι διαφορετικές για το ίδιο υπόδειγμα.

Η μετατροπή ενός δέντρου απόφασης σε κανόνες ταξινόμησης είναι μία εύκολη διαδικασία καθώς αντιστοιχούμε ένα κανόνα σε κάθε φύλλο, έτσι ώστε η προϋπόθεση του κανόνα να περιέχει μία συνθήκη για κάθε κόμβο που συναντάται από τη ρίζα ως το φύλλο και το συμπέρασμα να ορίζεται ως η τάξη εκχώρησης. Οι παραγόμενοι κανόνες είναι σαφείς και ορίζονται μονοσήμαντα ενώ η σειρά εκτέλεσής τους δεν παίζει ρόλο στο αποτέλεσμα. Ωστόσο, υπάρχουν περιπτώσεις που οι κανόνες οι οποίοι προκύπτουν είναι υπερβολικά περίπλοκοι και γι' αυτό απαιτείται «κλάδεμα» για την απομάκρυνση των περιττών κανόνων και ελέγχων.

Θα περιγράψουμε παρακάτω συνοπτικά μερικούς από τους πιο γνωστούς και ευρέως χρησιμοποιούμενος αλγόριθμους κανόνων ταξινόμησης.

10.4.1 Ο αλγόριθμος Conjuctive Rule

Ένας ευρέως γνωστός αλγόριθμος που παρέχεται από το Weka για τη δημιουργία κανόνων ταξινόμησης είναι ο «Conjuctive Rule». Ο αλγόριθμος αυτός υλοποιεί ένα μόνο conjunctive rule learner που μπορεί να προβλέψει για αριθμητικές και κατηγορικές κλάσεις. Ένας κανόνας αποτελείται από υποθέσεις που έχουν AND μεταξύ τους και το συμπέρασμα (τιμή κλάση) για την ταξινόμηση. Σε αυτή την περίπτωση, το συμπέρασμα του κανόνα είναι η κατανομή των διαθέσιμων κλάσεων (ή μέση τιμή για μια αριθμητική τιμή) στο σύνολο δεδομένων. Αν το στιγμιότυπο ελέγχου δεν καλύπτεται από αυτό τον κανόνα, τότε προβλέπεται χρησιμοποιώντας τη τιμή της προεπιλεγμένης κλάσης που δεν καλύπτεται από τον κανόνα στα δεδομένα εκπαίδευσης. Αυτός ο learner επιλέγει μια υπόθεση υπολογίζοντας το κέρδος

πληροφορίας κάθε υπόθεσης και κλαδεύει τον παραγόμενο κανόνα χρησιμοποιώντας το Reduced Error Pruning (REP) στον αριθμό των υποθέσεων. Για την ταξινόμηση, η πληροφορία μιας υπόθεσης είναι ο σταθμισμένος μέσος όρος των εντροπιών και των δεδομένων που καλύπτονται, και αυτών που δεν καλύπτονται από τον κανόνα.

10.4.2 Ο αλγόριθμος Decision Table

Ο αλγόριθμος αυτός προτάθηκε από τον Kohavi. Συνοψίζει το σύνολο δεδομένων με έναν "πίνακα απόφασης" ο οποίος περιέχει τον ίδιο αριθμό μεταβλητών με το αρχικό σύνολο δεδομένων. Κατόπιν, ένα νέο στοιχείο δεδομένων ανατίθεται σε μια κατηγορία βρίσκοντας τη γραμμή στον πίνακα απόφασης που να ταιριάζει με τις τιμές της μη κλάσης του στοιχείου δεδομένων. Ο Decision Table υιοθετεί τη μέθοδο wrapper για να βρει ένα καλό υποσύνολο μεταβλητών για να το εντάξει στον πίνακα. Με τη διαγραφή μεταβλητών που συμβάλλουν ελάχιστα ή καθόλου σε ένα μοντέλο του συνόλου δεδομένων, ο αλγόριθμος μειώνει την πιθανότητα του overfitting, δηλαδή της υπερπροσαρμογής, και δημιουργεί ένα μικρότερο και συμπυκνωμένο πίνακα απόφαση.

10.4.3 Ο αλγόριθμος OneR

Ο OneR ή "One Rule" είναι ένας απλός αλγόριθμός που προτάθηκε από τον Holt. Ο OneR κατασκευάζει έναν κανόνα για κάθε μεταβλητή στα δεδομένα εκπαίδευσης και μετά επιλέγει τον κανόνα με το μικρότερο ποσοστό σφάλματος. Για να δημιουργηθεί ένας κανόνας για μια μεταβλητή, η πιο συχνή κλάση για κάθε τιμή της μεταβλητής πρέπει να προσδιοριστεί. Η πιο συχνή κλάση είναι απλά η κλάση που εμφανίζεται πιο συχνά για αυτή την τιμή της μεταβλητής. Ένας κανόνας είναι απλά ένα σύνολο από τιμές μεταβλητών που δεσμεύεται στην κλάση πλειοψηφίας τους. Ο OneR επιλέγει τον κανόνα με το χαμηλότερο ποσοστό σφάλματος. Σε περίπτωση που δύο ή περισσότεροι κανόνες έχουν το ίδιο ποσοστό σφάλματος, ο κανόνας επιλέγεται τυχαία.

10.4.4 Ο αλγόριθμος PART

Ο PART είναι ένας αλγόριθμος που ακολουθεί την τεχνική «separate and conquer» και προτάθηκε από τους Eibe και Witten. Ο αλγόριθμος παράγει σύνολα κανόνων που καλούνται "λίστες απόφασης" οι οποίες είναι διατεταγμένα σύνολα κανόνων. Ένα νέο δεδομένο συγκρίνεται με κάθε κανόνα στη λίστα με τη σειρά, και το στοιχείο ανατίθεται σε μια κατηγορία του πρώτου κανόνα ταιριάσματος (μια προεπιλογή εφαρμόζεται αν δεν ταιριάζει με επιτυχία κανένας κανόνας). Ο PART κατασκευάζει ένα μερικό δέντρο απόφασης σε κάθε επανάληψη και αντιστοιχεί το "καλύτερο φύλλο" σε έναν κανόνα. Ο αλγόριθμός είναι ένας συνδυασμός του C4.5 και του RIPPER.

10.4.5 Ο αλγόριθμος Prism

Ο στόχος μας εδώ είναι η μεγιστοποίηση της ακρίβειας. Υποθέτουμε τώρα ότι t είναι ο συνολικός αριθμός των υποδειγμάτων που καλύπτονται από τον κανόνα, p είναι τα υποδείγματα της τάξης που εκχωρεί ο κανόνας, δηλαδή οι σωστές προβλέψεις, και άρα t-p ο αριθμός σφαλμάτων του κανόνα. Τότε, επιλέγουμε το νέο όρο που μεγιστοποιεί το λόγο p/t. Η διαδικασία ολοκληρώνεται όταν p/t=1 ή το σύνολο των υποδειγμάτων δε μπορεί να διασπαστεί περαιτέρω. Αυτά που περιγράψαμε είναι ο αλγόριθμος PRISM για την κατασκευή κανόνων. Ο αλγόριθμος PRISM, χωρίς τον εξωτερικό βρόγχο, παράγει λίστα κανόνων απόφασης για μία τάξη και οι επόμενοι κανόνες σχεδιάζονται για υποδείγματα που δεν καλύπτονται από τους προηγούμενους κανόνες. Η σειρά δεν έχει σημασία, καθώς όλοι οι κανόνες προβλέπουν την ίδια τάξη. Ο εξωτερικός βρόγχος λαμβάνει υπόψη κάθε τάξη χωριστά και δεν υποδηλώνεται εξάρτηση από τη διαδοχή. Από αυτή τη διαδικασία μπορεί να προκύψουν επικαλυπτόμενοι κανόνες και άρα είναι αναγκαίος ο ορισμός προκαθορισμένου κανόνα.

10.4.6 Ο αλγόριθμος RIDOR

Ο αλγόριθμος Ridor είναι μια υλοποίηση ενός Ripple-Down learner που προτάθηκε από τους Gaines και Compton. Παράγει πρώτα ένα προεπιλεγμένο κανόνα και στη συνέχεια τις εξαιρέσεις για τον προεπιλεγμένο κανόνα με το ελάχιστο (σταθμισμένο) ποσοστό σφάλματος. Κατόπιν παράγει τις "καλύτερες" εξαιρέσεις (για κάθε εξαίρεση) και επαναλαμβάνει τη διαδικασία μέχρι να είναι καθαρό. Έτσι εκτελεί μια επέκταση εξαιρέσεων που μοιάζει με δέντρο. Οι εξαιρέσεις είναι ένα σύνολο κανόνων που προβλέπει κλάσεις διαφορετικές από την προεπιλεγμένη. Ο IREP χρησιμοποιείται για να παράγει τις εξαιρέσεις.

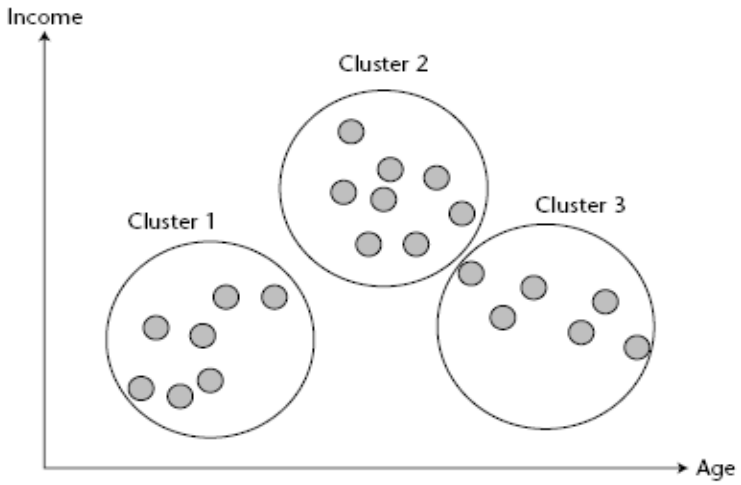
10.4.7 Ο αλγόριθμος JRip (RIPPER)

Προτάθηκε από τον William W. Cohen ως μια βελτιστοποιημένη έκδοση του IREP. Ο RIPPER κατασκευάζει ένα σύνολο κανόνων προσθέτοντας επαναληπτικά κανόνες σε ένα κενό σύνολο κανόνων μέχρι να καλυφθούν όλα τα θετικά παραδείγματα. Οι κανόνες σχηματίζονται προσθέτοντας συνθήκες στην υπόθεση ενός κανόνα (ξεκινώντας με κενή υπόθεση) μέχρι να μην καλύπτονται αρνητικά παραδείγματα. Ένας συνδυασμός με τεχνικές cross-validation και μήκους ελάχιστης περιγραφής χρησιμοποιείται για να αποτρέψει την υπερπροσαρμογή (overfitting). Μπορεί να εφαρμοστεί επίσης και σε συνεχή δεδομένα. Έχει πολύ καλή απόδοση από άποψη χρόνου εκτέλεσης ακόμα και για μεγάλες βάσεις δεδομένων με θόρυβο.

11. Συσταδοποίηση

11.1 Εισαγωγή

Η συσταδοποίηση (clustering) ή ανάλυση κατά συστάδες (Cluster analysis), όπως διαφορετικά συναντάμε αυτό τον όρο στη βιβλιογραφία, είναι μία διερευνητική τεχνική ανάλυσης δεδομένων, σχεδιασμένη ώστε να αποκαλύπτει (φυσικές) ομάδες μέσα σε σύνολα δεδομένων. Ο όρος χρησιμοποιήθηκε πρώτη φορά από τον Tryon το 1939, και πλαισιώνει ένα αριθμό αλγορίθμων και μεθόδων για την ομαδοποίηση αντικειμένων όμοιου είδους σε αντίστοιχες κατηγορίες. Η ομαδοποίηση-συσταδοποίηση (clustering) είναι η εργασία του καταμερισμού ενός ετερογενούς πληθυσμού σε ένα σύνολο λιγότερο ετερογενών υποομάδων (clusters). Η συσταδοποίηση είναι ένα εργαλείο διερευνητικής ανάλυσης δεδομένων το οποίο στοχεύει στο να ταξινομήσει διαφορετικά αντικείμενα σε ομάδες (συστάδες) με ένα τέτοιο τρόπο ώστε ο βαθμός συσχέτισης μεταξύ δύο αντικειμένων να είναι μέγιστος αν τα αντικείμενα αυτά ανήκουν στην ίδια ομάδα, και ελάχιστος σε διαφορετική περίπτωση. Η συσταδοποίηση μπορεί λοιπόν να χρησιμοποιηθεί για την ανακάλυψη δομών σε δεδομένα με βάση την ομοιότητα των υπό εξέταση αντικειμένων. Η συσταδοποίηση διαφοροποιείται από την ταξινόμηση (κατηγοριοποίηση) ως προς το ότι ανήκει στην κατηγορία της μάθησης χωρίς επίβλεψη (unsupervised learning) και δεν προϋποθέτει καμία εκ των προτέρων γνώση του αριθμού των συστάδων ή της σημασίας τους. Επίσης, αυτό που διαφοροποιεί την ομαδοποίηση από την ταξινόμηση είναι ότι η ομαδοποίηση δε βασίζεται σε προκαθορισμένες κλάσεις. Στην ταξινόμηση, ο πληθυσμός διαιρείται σε κλάσεις αναθέτοντας κάθε στοιχείο ή εγγραφή σε μία προκαθορισμένη κλάση με βάση ένα μοντέλο που αναπτύσσεται μέσω της εκπαίδευσής του, με υποδείγματα που έχουν κατηγοριοποιηθεί εκ των προτέρων, ενώ στη συσταδοποίηση οι εγγραφές ομαδοποιούνται σε σύνολα με βάση την ομοιότητα που παρουσιάζουν μεταξύ τους ,όπως φαίνεται για παράδειγμα στην παρακάτω εικόνα.



Εικόνα 20: Ομαδοποίηση εγγραφών σε σύνολα

Οι τεχνικές συσταδοποίησης θα πρέπει να μπορούν να διαχειριστούν τις άτυπες παρατηρήσεις (outliers) των βάσεων δεδομένων και το θόρυβο, να μπορούν να κλιμακώνονται και να διαχειρίζονται δυναμικά μεταβαλλόμενα δεδομένα αλλάζοντας τις συστάδες στην πορεία του χρόνου, να βρίσκουν ποια δεδομένα πρέπει να χρησιμοποιηθούν στη διαδικασία καθώς και να ερμηνεύουν και να αξιολογούν τα αποτελέσματα που προκύπτουν.

Η συσταδοποίηση δεν είναι τόσο μία τυπική στατιστική μέθοδος όσο μία συλλογή διαφορετικών αλγορίθμων που «τοποθετούν αντικείμενα σε συστάδες, σύμφωνα με καλά ορισμένους κανόνες ομοιότητας». Σε αντίθεση με πολλές άλλες στατιστικές διαδικασίες, οι μέθοδοι της συσταδοποίησης χρησιμοποιούνται όταν δεν έχουμε καμία εκ των προτέρων υπόθεση αλλά είμαστε ακόμα στη διερευνητική φάση της έρευνάς μας. Κατά μία έννοια η συσταδοποίηση βρίσκει την πιο σημαντική πιθανή λύση και επομένως, οι έλεγχοι στατιστικής σημαντικότητας δεν είναι κατάλληλοι για τη συσταδοποίηση.

Συμπερασματικά, η συσταδοποίηση μπορεί να χρησιμοποιηθεί για να ανακαλύψουμε δομές που παράγονται από τα δεδομένα μας, χωρίς να εφαρμόσουμε πρώτα κάποια διαδικασία εξήγησης ή ερμηνείας τους. Αυτό αποτελεί πολύ σημαντικό πλεονέκτημα της συσταδοποίησης και για αυτό η διαδικασία αυτή κυριαρχεί σε όλα τα επιστημονικά πεδία που ασχολούνται με την ανάλυση πολυμεταβλητών δεδομένων. Το 1975 ο Hartigan πρόβαλε μία άριστη περίληψη των πολλών δημοσιευμένων μελετών που αναφερόντουσαν στα αποτελέσματά της.

11.2 Ο K-Means αλγόριθμος

Ο k-means είναι ένας κλασικός αλγόριθμος συσταδοποίησης της κατηγορίας των αλγορίθμων διαχωριστικής συσταδοποίησης. Οι διαχωριστικοί αλγόριθμοι συσταδοποίησης, στοχεύουν στη δημιουργία ενός διαχωριστικού «επιπέδου» του συνόλου αντικειμένων, δηλαδή, στο διαμερισμό των αντικειμένων σε μη επικαλυπτόμενα υποσύνολα (συστάδες) έτσι ώστε κάθε αντικείμενο να ανήκει ακριβώς σε ένα υποσύνολο. Ο αλγόριθμος αυτός προτάθηκε το 1956 από τον H.Steinhaus και στην πιο κοινή μορφή του χρησιμοποιεί μία επαναληπτική ευρετική (heuristic) προσέγγιση γνωστή ως Lloyd’s algorithm (1957).

Η μέθοδος θεωρεί ότι ο αριθμός των συστάδων που θα προκύψουν είναι εκ των προτέρων γνωστός. Αυτό αποτελεί ένα περιορισμό της μεθόδου, καθώς είτε πρέπει να τρέξουμε τον αλγόριθμο με διαφορετικές επιλογές ως προς το πλήθος των συστάδων είτε πρέπει με κάποιον άλλο τρόπο να έχουμε καταλήξει στον αριθμό των συστάδων. Η μέθοδος δουλεύει επαναληπτικά. Χρησιμοποιεί την έννοια του κέντρου της ομάδας και στη συνέχεια κατατάσσει τις παρατηρήσεις ανάλογα με την απόστασή τους από τα

κέντρα όλων των ομάδων. Το κέντρο της ομάδας δεν είναι τίποτα άλλο από τη μέση τιμή για κάθε μεταβλητή όλων των παρατηρήσεων της ομάδας, δηλαδή αντιστοιχεί στο διάνυσμα των μέσων. Στη συνέχεια, για κάθε παρατήρηση υπολογίζουμε την ευκλείδεια απόστασή της από τα κέντρα των ομάδων που έχουμε και κατατάσσουμε κάθε παρατήρηση στην ομάδα που είναι πιο κοντά (για την ακρίβεια στην ομάδα με κέντρο πιο κοντά στην παρατήρηση). Αφού κατατάξουμε όλες τις παρατηρήσεις, τότε υπολογίζουμε εκ νέου τα κέντρα, απλώς ως τα διανύσματα των μέσων για τις παρατηρήσεις που ανήκουν στην κάθε ομάδα. Η διαδικασία επαναλαμβάνεται μέχρις ότου δεν υπάρχουν διαφορές ανάμεσα σε δύο διαδοχικές επαναλήψεις.

Συνήθως, η απόσταση που χρησιμοποιείται για την κατάταξη των παρατηρήσεων είναι η Ευκλείδεια. Αν θέλουμε να χρησιμοποιήσουμε άλλη απόσταση θα πρέπει να κάνουμε ειδικούς μετασχηματισμούς στα δεδομένα, πριν τη χρησιμοποιήσουμε. Ο αλγόριθμος αυτός δουλεύει ικανοποιητικά για μεγάλα σύνολα δεδομένων επειδή σε αυτή την περίπτωση δουλεύει πολύ πιο γρήγορα από την ιεραρχική συσταδοποίηση.

Αλγοριθμικά έχουμε τα εξής βήματα:

1. Αρχικά καθορίζουμε σε πόσες ομάδες(συστάδες) k θα γίνει η ομαδοποίηση.
2. Στο επόμενο βήμα επιλέγονται k σημεία ως κέντρα των συστάδων.
3. Εκχωρούμε κάθε υπόδειγμα στη συστάδα της οποίας το κέντρο έχει τη μικρότερη Ευκλείδεια απόσταση από την παρατήρηση
4. Στη συνέχεια, γίνεται ο υπολογισμός των κεντροειδών των ομάδων από τις παρατηρήσεις που είναι μέσα στην ομάδα, τα οποία θεωρούνται ως οι νέες κεντρικές τιμές των συστάδων που αντιστοιχούν.
5. Η παραπάνω διαδικασία επαναλαμβάνεται με τα νέα κέντρα των συστάδων μέχρι τη σύγκλιση, δηλαδή μέχρι τα κεντρικά σημεία να μην αλλάζουν.

Τα κριτήρια τερματισμού μπορούν να οριστούν από το χρήστη, καθώς σε μεγάλα σύνολα δεδομένων με πολύπλοκη δομή ο αλγόριθμος μπορεί να καθυστερεί πολύ, αν το κριτήριο τερματισμού είναι αυστηρό.

Τα κύρια πλεονεκτήματα του k-means αλγόριθμου είναι η απλότητα του και η ταχύτητά του κάτι που επιτρέπει την εφαρμογή του αλγορίθμου και σε μεγάλα σύνολα δεδομένων.

12. Μέτρα Αξιολόγησης

12.1 Εισαγωγή

Η αξιολόγηση της αξιοπιστίας ενός ταξινομητή είναι απαραίτητη για τη διασφάλιση της ποιότητας των δεδομένων. Το πλέον συνηθισμένο κριτήριο για την αξιολόγηση της ποιότητας ενός μοντέλου ταξινόμησης είναι το κριτήριο της διάκρισης, το οποίο μετρά το κατά πόσο καλά διαχωρίζονται οι δύο κλάσεις στο σύνολο δεδομένων. Στη μελέτη μας θεωρούμε τα πιο συχνά χρησιμοποιούμενα μέτρα διάκρισης για την αξιολόγηση της απόδοσης των μεθόδων ταξινόμησης που εφαρμόζουμε, και υιοθετούμε τους κλασικούς ορισμούς που χρησιμοποιούνται στη δυαδική ταξινόμηση.

12.2 Μέτρα ακρίβειας

Στην παρούσα ενότητα παρουσιάζονται τα μέτρα αξιολόγησης τα οποία υιοθετήσαμε για την αξιολόγηση της απόδοσης των αλγορίθμων ταξινόμησης καθώς και της προβλεπτικής τους ικανότητας.

Δοθέντος ενός ταξινομητή και μίας καταχώρησης, υπάρχουν τέσσερις πιθανές εκβάσεις. Οι θετικές καταχωρήσεις οι οποίες προβλέπονται σωστά ως θετικές (True Positive-TP), οι θετικές καταχωρήσεις οι οποίες προβλέπονται εσφαλμένα ως αρνητικές (False Negative-FN), οι αρνητικές καταχωρήσεις οι οποίες προβλέπονται εσφαλμένα ως θετικές (False Positive-FP), και τέλος οι αρνητικές καταχωρήσεις οι οποίες προβλέπονται σωστά ως αρνητικές (True Negative-TN).

Η ακρίβεια ταξινόμησης (**accuracy**) χρησιμοπιείται ως πρώτο κριτήριο. Η ακρίβεια ορίζεται ως το ποσοστό των σωστά ταξινομημένων καταχωρήσεων στο σύνολο εξέτασης. Τα άλλα δύο κριτήρια που χρησιμοποιούνται είναι η ευαισθησία και η ειδικότητα που είναι δύο στατιστικά μέτρα της απόδοσης ενός τεστ, δυαδικής ταξινόμησης. Η ευαισθησία (**sensitivity**) μετρά το ποσοστό των πραγματικά θετικών που έχουν αναγνωριστεί σωστά ως θετικά, ενώ η ειδικότητα (**specificity**) μετρά το ποσοστό των πραγματικά αρνητικών που έχουν αναγνωριστεί σωστά ως αρνητικά.

Η ευασθησία και η ειδικότητα μπορούν εναλλακτικά να περιγραφούν ως εξής:

- Ευαισθησία=1-Σφάλμα Τύπου II
- Ειδικότητα=1-Σφάλμα Τύπου I

Τα άλλα κριτήρια που χρησιμοποιούνται στην ανάκτηση πληροφοριών είναι η ανάκληση (**recall**) η οποία αντιστοιχεί στην ευαισθησία και η ακρίβεια (**precision**) που είναι το ποσοστό των αληθώς θετικών μεταξύ όλων των προβλεπόμενων θετικών. Στα προβλήματα ταξινόμησης ένα άλλο ευρέως διαδεδομένο μέτρο απόδοσης είναι ο γεωμετρικός μέσος των τιμών ακρίβειας των κλάσεων (**G-mean**), ο οποίος θέτει όλες τις κατηγορίες επί ίσοις όροις και δεν δίνει μεγαλύτερη προτεραιότητα στις σπάνιες

θετικές κλάσεις. Ένα μέτρο απόδοσης που επιτρέπει κάτι τέτοιο είναι το F-μέτρο (**F-measure**) το οποίο δεν λαμβάνει υπόψη του την απόδοση των αρνητικών κλάσεων.

Σε προβλήματα δυαδικής ταξινόμησης τα προαναφερθέντα μέτρα ακρίβειας ορίζονται ως ακολούθως:

- $\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$,
- $\text{Sensitivity (Recall)} = \frac{TP}{TP+FN}$,
- $\text{Specificity} = \frac{TN}{TN+FP}$,
- $\text{Precision} = \frac{TP}{TP+FP}$,
- $\text{G-mean} = \sqrt{\text{sensitivity} * \text{specificity}}$,
- $\text{F-measure} = \frac{\text{precision} * \text{recall}}{\beta\text{precision} + (1-\beta)\text{recall}}$,

όπου η πράμετρος β με 0<β<1 επιτρέπει στον πειραματιστή να εκχωρήσει τα σχετικά βάρη στην ακρίβεια και στην ανάκληση, με το 0.5 να τους αποδίδει την ίδια σημασία.

Ένα άλλο δημοφιλές στατιστικό εργαλείο είναι οι καμπύλες λειτουργικού χαρακτηριστικού δέκτη (Receiver Operating Characteristic Curves – ROC curves), οι οποίες εξ ορισμού χρησιμοποιούνται για την αξιολόγηση της απόδοσης ενός συστήματος με διχοτομικά εξαγόμενα αποτελέσματα. Μία καμπύλη **ROC** παρουσιάζεται ως το γράφημα της ευαισθησίας έναντι του 1-ειδικότητα για όλες τις πιθανές τιμές αποκοπής. Παραδοσιακά, το εμβαδόν κάτω από την καμπύλη ROC (Area Under the Curve-AUC) χρησιμοποιείται ως συνοπτικός δείκτης ακρίβειας ενός τεστ, και είναι χρήσιμο ως περιγραφικό μέτρο της συνολικής απόδοσης ενός τεστ. Στατιστικά μιλώντας, το εμβαδόν κάτω από την καμπύλη ROC (**AUC**) ενός ταξινομητή είναι η πιθανότητα ένας ταξινομητής να κατατάξει ένα τυχαία επιλεγμένο θετικό υπόδειγμα υψηλότερα από ένα τυχαίο επιλεγμένο αρνητικό υπόδιεγμα.

12. 3 Πίνακες συνάφειας

Στη θεωρία ανίχνευσης σημάτων, μία καμπύλη λειτουργικού χαρακτηριστικού δέκτη (receiver operating characteristic curve) ή απλά μία ROC καμπύλη, είναι μία γραφική παράσταση της ευαισθησίας (sensitivity), ή των αληθώς θετικών, έναντι του 1 – ειδικότητα (specificity), ή ψευδώς θετικών, για ένα σύστημα δυαδικής ταξινόμησης, καθώς το όριο ταξινόμησης ποικίλει. Ο δέκτης λειτουργικού χαρακτηριστικού μπορεί ισοδύναμα να εκπροσωπηθεί με τη γραφική παράσταση του ποσοστού των αληθώς θετικών (TPR = true positive rate) έναντι του ποσοστού των ψευδών θετικών (FPR =

false positive rate). Είναι επίσης γνωστή ως καμπύλη σχετικού λειτουργικού χαρακτηριστικού, αφού αποτελεί τη σύγκριση δύο λειτουργικών χαρακτηριστικών (TPR, FPR) καθώς το κριτήριο αλλάζει.

Ας εξετάσουμε ένα πρόβλημα πρόβλεψης διπλής κλάσης (δυαδική ταξινόμηση), στο οποίο το αποτέλεσμα χαρακτηρίζεται ως θετική *(p)* ή αρνητική *(n)* κλάση. Υπάρχουν τέσσερις πιθανές εκβάσεις για ένα δυαδικό ταξινομητή. Αν το αποτέλεσμα της πρόβλεψη είναι *p* και η πραγματική τιμή είναι επίσης *p*, αυτό ονομάζεται *αληθώς θετικό*. Ωστόσο, εάν η πραγματική τιμή είναι *n*, λέγεται *ψευδώς θετικό*. Αντίθετα, ένα *αληθώς αρνητικό* έχει προκύψει όταν τόσο το αποτελέσματα της πρόβλεψης όσο και η πραγματικής τιμή είναι *n*, και ένα *ψευδώς αρνητικό* όταν το αποτέλεσμα πρόβλεψης είναι *n*, ενώ η πραγματική τιμή είναι *p*.

Ας ορίσουμε ένα πείραμα με P θετικές και N αρνητικές περιπτώσεις. Τα τέσσερα αποτελέσματα μπορούν να παρουσιαστούν με ένα 2 x 2 πίνακα συνάφειας ως εξής:

Πίνακας 1: 2x2 πίνακας συνάφειας αποτελεσμάτων δοκιμασίας

		Πραγματική Τιμή		
		p	n	Σύνολο
Αποτέλεσμα Πρόβλεψης	p´	Αληθώς Θετικό	Ψευδώς Θετικό	P´
	n´	Ψευδώς Αρνητικό	Αληθώς Αρνητικό	N´
	Σύνολο	P	N	

Πίνακας 2: Βασική ορολογία

Ευαισθησία ή ποσοστό αληθώς θετικών (TPR)

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$$

Ποσοστό ψευδώς θετικών (FPR)

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN}$$

Ειδικότητα ή ποσοστό αληθώς αρνητικών (SPC)

$$SPC = \frac{TN}{N} = \frac{TN}{FP + TN} = 1\text{-}FPR$$

Ακρίβεια (Accuracy)

$$ACC = \frac{TP + TN}{P + N}$$

13. Διασταυρωμένη Επικύρωση

13.1 Ορισμός

Η διασταυρωμένη επικύρωση (cross- validation) είναι μία στατιστική μέθοδος αξιολόγησης και σύγκρισης αλγορίθμων μάθησης με διαίρεση των δεδομένων σε δύο τμήματα: το ένα χρησιμοποιείται για τη μάθηση και το άλλο για την επικύρωση του μοντέλου. Στην τυπική διασταυρωμένη επικύρωση, τα σύνολα εκπαίδευσης και επικύρωσης πρέπει να διασταυρωθούν σε διαδοχικούς γύρους έτσι ώστε κάθε δεδομένο να έχει μία ευκαιρία να επικυρωθεί. Η βασική μορφή διασταυρωμένης επικύρωσης είναι η k-πλάσια διασταυρωμένη επικύρωση (k-fold cross- validation). Άλλες μορφές της διασταυρωμένης επικύρωσης είναι ειδικές περιπτώσεις της k-πλάσιας διασταυρωμένης επικύρωσης ή περιλαμβάνουν επαναλαμβανόμενους γύρους της k-πλάσιας διασταυρωμένης επικύρωσης .

Στην k-πλάσια διασταυρωμένη επικύρωση τα δεδομένα πρώτα κατανέμονται σε k τμήματα ή πτυχές ίσου ή σχεδόν ίσου μεγέθους. Στη συνέχεια k επαναλήψεις εκπαίδευσης και επικύρωσης εκτελούνται έτσι ώστε σε κάθε επανάληψη μια διαφορετική πτυχή των δεδομένων να επαρκεί για επικύρωση ενώ οι εναπομείνασες k-1 να χρησιμοποιούνται για μάθηση.

Η διασταυρωμένη επικύρωση χρησιμοποιείται για την αξιολόγηση ή τη σύγκριση των αλγορίθμων εκμάθησης ως εξής: σε κάθε επανάληψη ένας ή και περισσότεροι αλγόριθμοι εκμάθησης χρησιμοποιούν k-1 πτυχές των δεδομένων για να μάθουν ένα ή και περισσότερα μοντέλα και στη συνέχεια τα εξακριβωμένα μοντέλα καλούνται να κάνουν προβλέψεις σχετικά με τα δεδομένα στην πτυχή επικύρωσης. Η απόδοση κάθε αλγορίθμου μάθησης για κάθε πτυχή μπορεί να παρακολουθηθεί με τη χρήση κάποιων προκαθορισμένων μετρικών ακρίβειας. Μέχρι την ολοκλήρωση, k δείγματα της μετρικής απόδοσης θα είναι διαθέσιμα για κάθε αλγόριθμο. Διαφορετικές μεθοδολογίες, όπως αυτές της μέσης τιμής μπορούν να χρησιμοποιηθούν για να ληφθεί ένα συνολικό μέτρο από αυτά τα δείγματα, ή αυτά τα δείγματα μπορούν να χρησιμοποιηθούν σε έναν έλεγχο στατιστικής υπόθεσης για να δειχθεί ότι ένας αλγόριθμος είναι ανώτερος ενός άλλου.

Υπάρχουν δύο πιθανοί στόχοι στη διασταυρωμένη επικύρωση:

(i) Η εκτίμηση της απόδοσης των μοντέλων μάθησης από τα διαθέσιμα δεδομένα με τη χρήση ενός αλγορίθμου. Με άλλα λόγια η μέτρηση της (δυνατότητας) γενίκευσης ενός αλγορίθμου.

(ii) Η σύγκριση της επίδοσης ενός ή και περισσότερων διαφορετικών αλγορίθμων και η εύρεση του καλύτερου αλγορίθμου για τα διαθέσιμα δεδομένα, ή εναλλακτικά η

σύγκριση της απόδοσης δύο ή και περισσότερων παραλλαγών ενός παραμετρικοποιημένου μοντέλου.

Οι δύο παραπάνω στόχοι είναι ιδιαιτέρως συσχετισμένοι, καθώς ο δεύτερος στόχος επιτυγχάνεται αυτόματα, εάν κάποιος γνωρίζει τις ακριβείς εκτιμήσεις απόδοσης. Δεδομένων ενός δείγματος N περιπτώσεων και ενός αλγορίθμου εκμάθησης Α, η μέση ακρίβεια διασταυρωμένης επικύρωσης του Α, σε αυτές τις N περιπτώσεις, μπορεί να ληφθεί σαν μια εκτίμηση για την ακρίβεια του Α σε αφανή δεδομένα, όταν ο Α εκπαιδεύεται όλες τις N περιπτώσεις. Εναλλακτικά, εάν ο τελικός στόχος είναι η σύγκριση δύο αλγορίθμων μάθησης, τα δείγματα απόδοσης που λαμβάνονται μέσω της διασταυρωμένης επικύρωσης μπορούν να χρησιμοποιηθούν για τη διεξαγωγή ελέγχων στατιστικών υποθέσεων, συγκρίνοντας ένα ζεύγος αλγορίθμων μάθησης.

13.2 Κ-πλάσια διασταυρωμένη επικύρωση

Στην k-πλάσια διασταυρωμένη επικύρωση (k-fold cross- validation) τα δεδομένα πρώτα κατανέμονται σε k πτυχές ίσου (ή σχεδόν ίσου) μεγέθους. Στη συνέχεια διεξάγονται k επαναλήψεις εκπαίδευσης και επικύρωσης, έτσι ώστε σε κάθε επανάληψη μια διαφορετική πτυχή των δεδομένων να επαρκεί για επικύρωση ενώ οι εναπομείνασες k-1 πτυχές χρησιμοποιούνται για μάθηση. Τα δεδομένα συνήθως στρωματοποιούνται πριν χωριστούν σε k πτυχές. Η διαστρωμάτωση είναι η διαδικασία της ανακατανομής των δεδομένων έτσι ώστε να εξασφαλιστεί ότι κάθε πτυχή είναι ένας καλός αντιπρόσωπος του συνόλου. Για παράδειγμα σε ένα δυαδικό πρόβλημα ταξινόμησης, όπου κάθε κλάση αποτελείται από το 50% των δεδομένων είναι καλύτερο να κατανεμηθούν τα δεδομένα έτσι ώστε, σε κάθε πτυχή, κάθε κλάση να αποτελείται από περίπου τις μισές περιπτώσεις.

13.3 10-πλάσια διασταυρωμένη επικύρωση

Ένας ιδανικός ή στατιστικά άριστος στατιστικός σχεδιασμός πρέπει να παρέχει έναν επαρκή αριθμό ανεξάρτητων μετρήσεων της απόδοσης του αλγορίθμου. Για να γίνουν ανεξάρτητες μετρήσεις της απόδοσης του αλγορίθμου πρέπει να διασφαλιστεί ότι α παράγοντες που επηρεάζουν τη μέτρηση είναι ανεξάρτητοι από το ένα τρέξιμο στο επόμενο. Αυτοί οι παράγοντες είναι: (i) τα δεδομένα εκπαίδευσης από τα οποία μαθαίνει ο αλγόριθμος, (ii) τα δεδομένα ελέγχου που χρησιμοποιούνται για τη μέτρηση της απόδοσης του αλγορίθμου. Εάν κάποια δεδομένα χρησιμοποιούνται για δοκιμή σε περισσότερους από ένα γύρους τα αποτελέσματα που λαμβάνονται θα είναι εξαρτημένα και μια στατιστική σύγκριση δε θα είναι έγκυρη.

Όχι μόνο τα σύνολα δεδομένων θα πρέπει να ελέγχονται ανεξάρτητα στα διάφορα τρεξίματα, αλλά δε θα πρέπει να υπάρχει καμιά επικάλυψη μεταξύ των δεδομένων που χρησιμοποιούνται για μάθηση και των δεδομένων που χρησιμοποιούνται για επικύρωση στο ίδιο τρέξιμο. Τυπικά ένας αλγόριθμος μάθησης μπορεί να κάνει πιο ακριβείς προβλέψεις στα δεδομένα που έχουν ιδωθεί κατά τη διάρκεια της φάσης

μάθησης από ότι σε αυτά που δεν έχουν. Για αυτό το λόγο, μια επικάλυψη μεταξύ του συνόλου εκπαίδευσης και επικύρωσης μπορεί να οδηγήσει σε μια υπερεκτίμηση της μετρικής απόδοσης και απαγορεύεται. Για να ικανοποιηθεί η άλλη απαίτηση, συγκεκριμένα ένα επαρκώς μεγάλο δείγμα, οι περισσότεροι στατιστικοί κάνουν έκκληση για 30+ δείγματα.

Για ένα πραγματικά καλό στατιστικό σχεδιασμό, τα διαθέσιμα δεδομένα θα πρέπει να διαιρεθούν σε $30 \times 2 = 60$ διχοτομήσεις για να εκτελεστούν 30 πραγματικά ανεξάρτητα τρεξίματα εκπαίδευσης-δοκιμής. Ωστόσο, αυτό δεν είναι πρακτικό, διότι η απόδοση των αλγορίθμων μάθησης και κατάταξή τους δεν είναι δεν είναι αναλλοίωτες σε σχέση με τον αριθμό των δειγμάτων που είναι διαθέσιμα για μάθηση. Με άλλα λόγια, μια εκτίμηση ακριβείας σε μια τέτοια περίπτωση, θα μπορούσε να αντιστοιχεί στην ακρίβεια του αλγορίθμου μάθησης όταν μαθαίνει από μόλις το $1/60$ των διαθέσιμων δεδομένων με την προϋπόθεση ότι τα σύνολα εκπαίδευσης και επικύρωσης είναι του ιδίου μεγέθους. Εντούτοις, η ακρίβεια του αλγορίθμου μάθησης στα αφανή δεδομένα, όταν ο αλγόριθμος εκπαιδεύεται σε όλα τα επί του παρόντος διαθέσιμα δεδομένα, είναι πολύ μεγαλύτερη, δεδομένου ότι οι αλγόριθμοι μάθησης εν γένει βελτιώνονται όσον αφορά την ακρίβεια καθώς περισσότερα δεδομένα καθίστανται διαθέσιμα για μάθηση. Ομοίως συγκρίνοντας δύο αλγορίθμους Α και Β, ακόμα και αν ο Α έχει αποδειχτεί ότι είναι ο καλύτερος αλγόριθμος όταν χρησιμοποιείται το $1/60$ των διαθέσιμων δεδομένων δεν υπάρχει εγγύηση ότι θα είναι επίσης ο καλύτερος αλγόριθμος σε σχέση με όταν χρησιμοποιούνται όλα τα διαθέσιμα δεδομένα για μάθηση. Πολλοί υψηλής απόδοσης αλγόριθμοι μάθησης χρησιμοποιούν σύνθετα μοντέλα με πολλές παραμέτρους και απλά δεν αποδίδουν καλά με ένα πολύ μικρό πλήθος δεδομένων. Αλλά μπορεί να είναι εξαιρετικοί όταν επαρκή δεδομένα είναι διαθέσιμα για να μάθουν από αυτά.

Υπενθυμίζεται ότι δύο παράγοντες επηρεάζουν το μέτρο απόδοσης: το σύνολο εκπαίδευσης και το σύνολο ελέγχου. Το σύνολο εκπαίδευσης επηρεάζει τη μέτρηση έμμεσα μέσω του αλγορίθμου μάθησης, ενώ η σύνθεση του συνόλου δοκιμής έχει άμεσο αντίκτυπο στο μέτρο απόδοσης .Ένας λογικός πειραματικός συμβιβασμός μπορεί να είναι να επιτρέπονται τα επικαλυπτόμενα σύνολα εκπαίδευσης διατηρώντας τα σύνολα δοκιμής ανεξάρτητα. Η k-πλάσια διασταυρωμένη επικύρωση κάνει ακριβώς αυτό.

Τώρα το ζητούμενο γίνεται η επιλογή μιας κατάλληλης τιμής για το k. Ένα μεγάλο k είναι φαινομενικά επιθυμητό καθώς με ένα μεγάλο k (i) υπάρχουν περιπτώσεις εκτίμησης απόδοσης και (ii) το μέγεθος του συνόλου εκπαίδευσης είναι πιο κοντά στο μέγεθος του πλήρους συνόλου των δεδομένων. Έτσι αυξάνοντας την πιθανότητα ότι οποιοδήποτε συμπέρασμα που βγαίνει σχετικά με τους αλγορίθμους μάθησης στα πλαίσια της δοκιμής, θα γενικεύεται στην περίπτωση που όλα τα δεδομένα χρησιμοποιούνται για την εκπαίδευση του μοντέλου μάθησης. Καθώς το k αυξάνεται, ωστόσο, η επικάλυψη μεταξύ των συνόλων εκπαίδευσης επίσης αυξάνεται. Για παράδειγμα, με 5-πλάσια διασταυρωμένη επικύρωση, κάθε σύνολο εκπαίδευσης μοιράζεται μόνο τα $3/4$ των περιπτώσεών του με κάθε ένα από τα άλλα 4 σύνολα

εκπαίδευσης, ενώ με τη 10-πλάσια διασταυρωμένη επικύρωση, κάθε σύνολο εκπαίδευσης μοιράζεται τα 8/9 των περιπτώσεών του με κάθε ένα από τα υπόλοιπα 9 σύνολα εκπαίδευσης. Επιπλέον, η αύξηση του k συρρικνώνει το μέγεθος του συνόλου δοκιμής, οδηγώντας σε λιγότερο ακριβείς και λιγότερο λεπτομερείς μετρήσεις της μετρικής απόδοσης. Για παράδειγμα, με ένα σύνολο δοκιμής μεγέθους 10 περιπτώσεων, μπορεί κανείς να μετρήσει την ακρίβεια με προσέγγιση 10%, ενώ με 20 περιπτώσεις η ακρίβεια μπορεί να μετρηθεί με προσέγγιση 5%. **Αυτοί οι ανταγωνιστικοί παράγοντες ελήφθησαν όλοι υπόψη και η γενική συναίνεση φαίνεται να είναι ότι ο k=10 είναι ένας καλός συμβιβασμός. Αυτή η τιμή του k είναι ιδιαιτέρως ελκυστική διότι κάνει προβλέψεις με τη χρήση του 90% των δεδομένων, καθιστώντας πιο πιθανή τη γενίκευση στο σύνολο των δεδομένων.**

14. Εφαρμογή στο Weka

14.1 Επιλογή χαρακτηριστικών

14.1.1 Επιλογή χαρακτηριστικών για την ομάδα του Αρκαδικού

Μεταβλητή-στόχος: Result

1. Ακολουθούμε τα βήματα

Explorer→Preprocess→Open File

και «φορτώνουμε» το αρχείο "Arkadikos 4 seasons.csv", το οποίο αφορά στην ομάδα του Αρκαδικού και περιλαμβάνει τις υπό εξέταση μεταβλητές για τέσσερις συνολικά χρονικές περιόδους 2011-2012, 2012-2013, 2013-2014 και 2014-2015.

2. Έπειτα, αφαιρούμε τη μεταβλητή «Arkadikos» (τικάρουμε και Remove), η οποία αποτελεί τη μεταβλητή- Index η οποία απαριθμεί τα παιχνίδια ($1^{ο}$ παιχνίδι, $2^{ο}$ παιχνίδι κ.λπ.) των τεσσάρων χρονικών περιόδων, και επιλέγουμε το χαρακτηριστικό «result» ως αυτό το οποίο δείχνει σε ποια κλάση ανήκει κάθε φορά το υπόδειγμα. Το χαρακτηριστικό «result» αποτελεί μία δίτιμη μεταβλητή με τη τιμή **0 να δηλώνει την «ήττα» και την τιμή 1 τη «νίκη».**

3. Προχωρούμε στη διαδικασία διακριτοποίησης των δεδομένων μας, ακολουθώντας τα παρακάτω βήματα:

«Choose→weka→filters→unsupervised→attribute»

Έπειτα, επιλέγουμε από τη λίστα διαθέσιμων επιλογών " το "Numeric to Nominal" κλικάρουμε "Filter→Filtering Capabilities"

- ✓ Numeric Attributes
- ✓ Numeric class

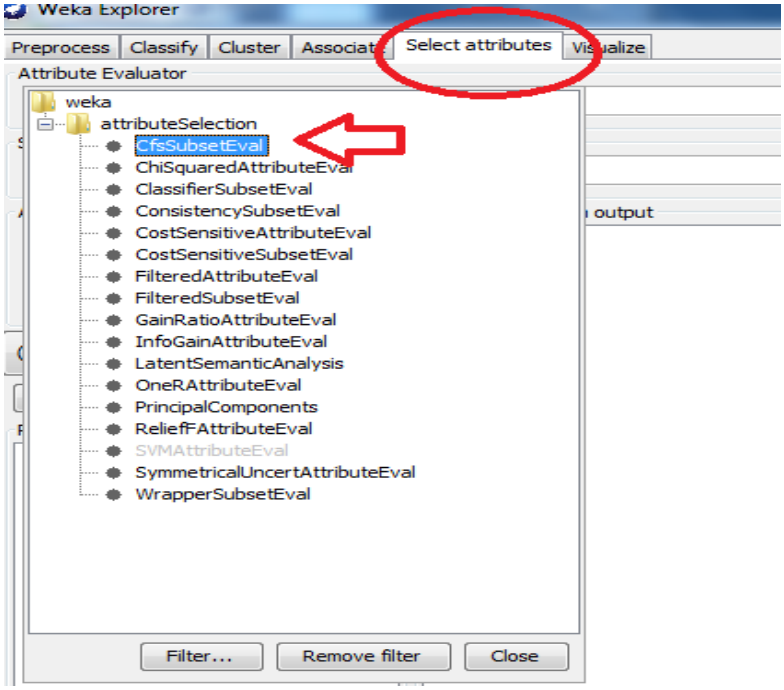
Ok

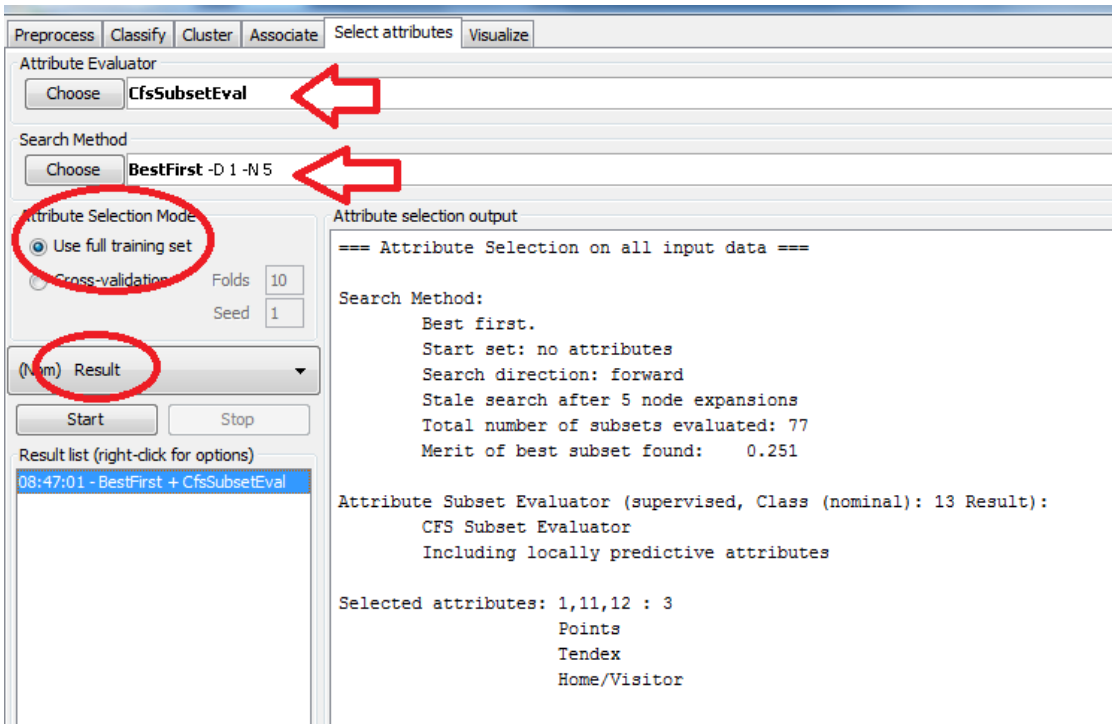
και →Apply.

Επιλέγουμε στο πάνελ την καρτέλα «Select attributes» όπου περιλαμβάνονται ποικίλες τεχνικές επιλογής χαρακτηριστικών με διαφορετικές δυνατότητες συνδυασμών για το χρήστη η κάθε μία.

Ακολουθούμε την παρακάτω διαδικασία»:

«Choose→Weka→attributeSelection→CfsSubsetEval»





Έπειτα, επιλέγουμε ως «Search method», τη μέθοδο BestFirst, ως «Attribute Selection Mode» το «Use Full Training Set» και ως μεταβλητή-στόχο τη μεταβλητή «Result» και →Start.

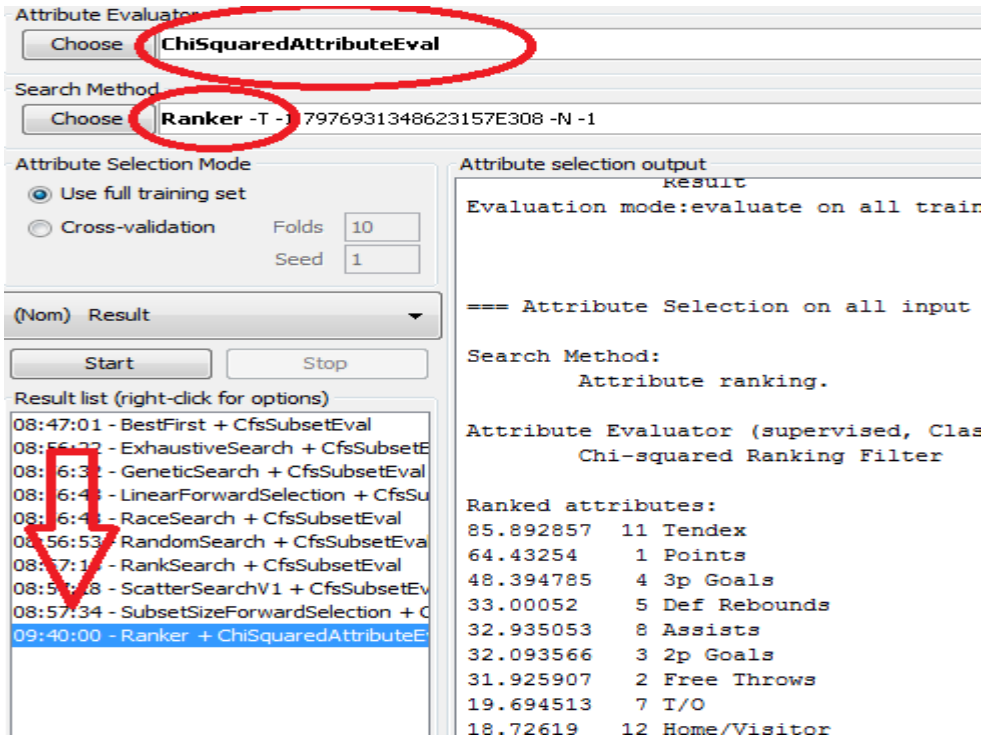
Επιλέγονται **3** μεταβλητές (από τις 12 συνολικά) ως σημαντικές, δηλαδή μόνο οι μεταβλητές Points, Tendex και Home/Visitor επηρρεάζουν σημαντικά τη μεταβλητή Result.

Selected attributes: 1,11,12: **3**
Points
Tendex
Home/Visitor

<u>Παρατήρηση</u>

Επιλέγοντας οποιαδήποτε μέθοδο από τη λίστα διαθέσιμων επιλογών «Search method» (εκτός της Ranker που δεν μπορεί να υλοποιηθεί σε συνδυασμό με τη CfsSubsetEval) προκύπτουν οι ίδιες τρεις σημαντικές μεταβλητές, οι οποίες ιεραρχικά ταξινομούνται ως 1. Points 2. Tendex 3. Home/Visitor. Διαφορετικά αποτελέσματα έδωσε μόνο η μέθοδος Scatter Search που επέλεξε μόνο μία μεταβλητή ως σημαντική, τη μεταβλητή «Tendex».

Έπειτα, για να αξιολογήσουμε τη σημαντικότητα κάθε μεταβλητής σε σχέση με τη μεταβλητή στόχο «Result», επιλέγουμε ως «Attribute Evaluator» τη μέθοδο «ChiSquaredAttributeEval» και ως «Search method» τη μέθοδο «Ranker», η οποία αποτελεί και την εξ ορισμού επιλογή του Weka και → Start.



Οι **12** μεταβλητές κατατάσσονται (ως ακολούθως) ιεραρχικά ανάλογα με το πόσο επηρεάζουν τη μεταβλητή-στόχο Result, ή με άλλα λόγια ανάλογα με το βαθμό της προβλεπτικής τους ικανότητας.

Selected attributes: 11,1,4,5,8,3,2,7,12,6,9,10 : **12**

Ranked attributes:

85.89	11 Tendex
64.43	1 Points
48.39	4 3p Goals
33.00	5 Def Rebounds
32.93	8 Assists
32.09	3 2p Goals
31.92	2 Free Throws
19.69	7 T/O
18.72	12 Home/Visitor
17.72	6 Off Rebounds
9.63	9 Steals
8.95	10 Blocks

Μεταβλητή-στόχος: Tendex

Τώρα θα εφαρμόσουμε την αντίστοιχη διαδικασία (όπως αυτή παρουσιάστηκε παραπάνω) θεωρώντας ως μεταβλητή στόχο τη μεταβλητή «Tendex».

1. Ακολουθούμε τα βήματα

Explorer→Preprocess→Open File

και «φορτώνουμε» το αρχείο "Arkadikos 4 seasons.csv", το οποίο αφορά στην ομάδα του Αρκαδικού και περιλαμβάνει τις υπό εξέταση μεταβλητές για τέσσερις συνολικά χρονικές περιόδους 2011-2012, 2012-2013, 2013-2014 και 2014-2015.

2. Έπειτα, αφαιρούμε τη μεταβλητή «Arkadikos» (τικάρουμε και Remove), η οποία αποτελεί τη μεταβλητή- Index η οποία απαριθμεί τα παιχνίδια ($1^{ο}$ παιχνίδι, $2^{ο}$ παιχνίδι

κ.λπ.) των τεσσάρων χρονικών περιόδων, και επιλέγουμε το χαρακτηριστικό «Tendex» ως αυτό το οποίο δείχνει σε ποια κλάση ανήκει κάθε φορά το υπόδειγμα.

3. Προχωρούμε στη διαδικασία διακριτοποίησης των δεδομένων μας, ακολουθώντας τα παρακάτω βήματα:

«Choose→weka→filters→unsupervised→attribute»

Έπειτα, επιλέγουμε από τη λίστα διαθέσιμων επιλογών ” το “Numeric to Nominal” κλικάρουμε “Filter→Filtering Capabilities”

- ✓ Numeric Attributes
- ✓ Numeric class

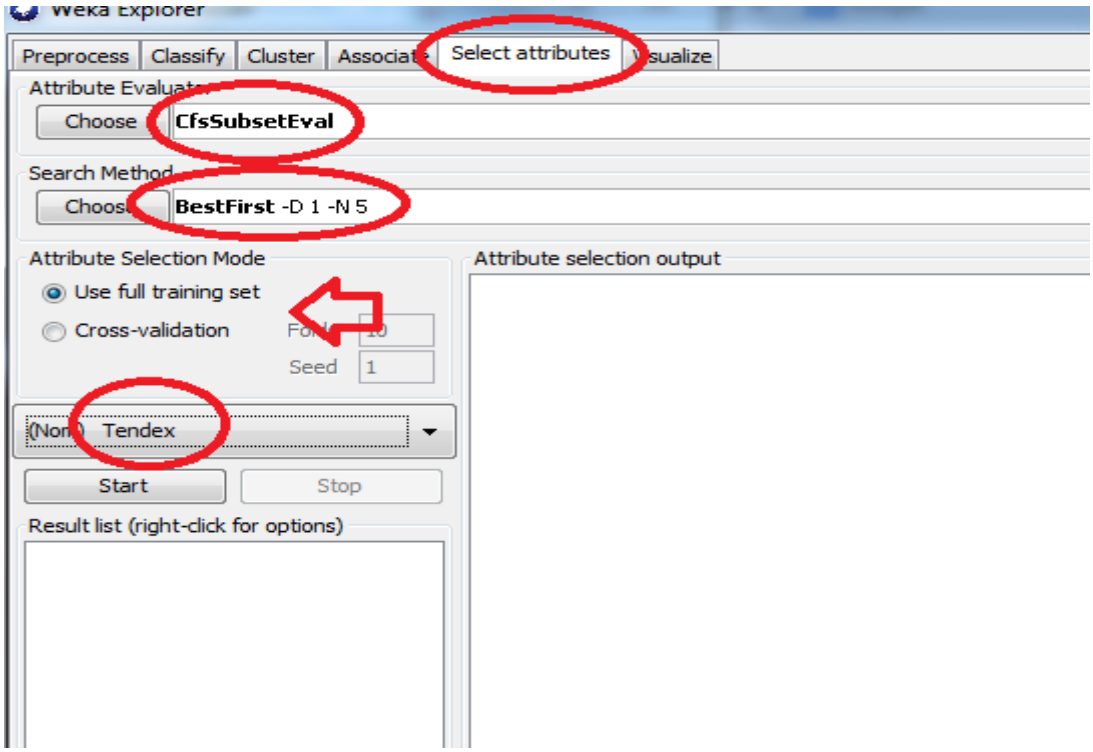
Ok

και →Apply.

Επιλέγουμε στο πάνελ την καρτέλα «Select attributes».

Ακολουθούμε την παρακάτω διαδικασία»:

«Choose→Weka→attributeSelection→CfsSubsetEval»



Έπειτα, επιλέγουμε ως «Search method», τη μέθοδο BestFirst, ως «Attribute Selection Mode» το «Use Full Training Set» και ως μεταβλητή-στόχο τη μεταβλητή «Tendex» και →Start.

Επιλέγονται και οι **12** μεταβλητές ως σημαντικές, δηλαδή και οι 12 επηρρεάζουν σημαντικά τη μεταβλητή Tendex.

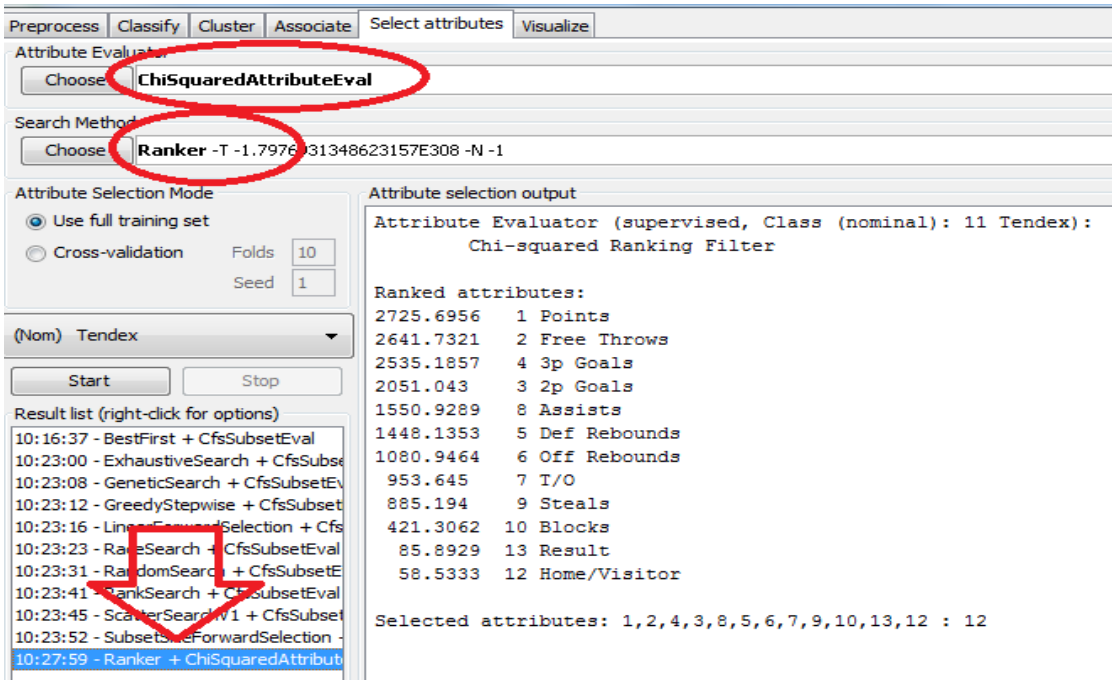
Selected attributes:
1,2,3,4,5,6,7,8,9,10,12,13 : **12**

Points, Free Throws, 2p Goals, 3p Goals, Def Rebounds, Off Rebounds, T/O, Assists, Steals, Blocks, Home/Visitor, Result

<u>Παρατήρηση</u>

Επιλέγοντας οποιαδήποτε μέθοδο από τη λίστα διαθέσιμων επιλογών «Search method» (εκτός της Ranker που δεν μπορεί να υλοποιηθεί σε συνδυασμό με τη CfsSubsetEval) προκύπτουν τα ίδια ακριβώς αποτελέσματα με παραπάνω.

Έπειτα, για να αξιολογήσουμε τη σημαντικότητα κάθε μεταβλητής σε σχέση με τη μεταβλητή στόχο «Tendex», επιλέγουμε ως «Attribute Evaluator» τη μέθοδο «ChiSquaredAttributeEval» και ως «Search method» τη μέθοδο «Ranker», η οποία αποτελεί και την εξ ορισμού επιλογή του Weka και → Start.



Οι **12** μεταβλητές κατατάσσονται (ως ακολούθως) ιεραρχικά ανάλογα με το πόσο επηρεάζουν τη μεταβλητή-στόχο Tendex.

Selected attributes: 1,2,4,3,8,5,6,7,9,10,13,12 : **12**

Ranked attributes:

2725.69	1 Points
2641.73	2 Free Throws
2535.18	4 3p Goals
2051.04	3 2p Goals
1550.92	8 Assists
1448.13	5 Def Rebounds
1080.94	6 Off Rebounds
953.64	7 T/O
885.19	9 Steals
421.30	10 Blocks
85.89	13 Result
58.53	12 Home/Visitor

14.1.2 Επιλογή χαρακτηριστικών για την ομάδα του Λαυρίου

Μεταβλητή-στόχος: Result

1. Ακολουθούμε τα βήματα

Explorer→Preprocess→Open File

και «φορτώνουμε» το αρχείο "Lavrio 4 seasons.csv", το οποίο αφορά στην ομάδα του Λαυρίου και περιλαμβάνει τις υπό εξέταση μεταβλητές για τέσσερις συνολικά χρονικές περιόδους 2011-2012, 2012-2013, 2013-2014 και 2014-2015.

2. Έπειτα, αφαιρούμε τη μεταβλητή «Lavrio» (τικάρουμε και Remove), η οποία αποτελεί τη μεταβλητή- Index η οποία απαριθμεί τα παιχνίδια ($1^{ο}$ παιχνίδι, $2^{ο}$ παιχνίδι κ.λπ.) των τεσσάρων χρονικών περιόδων, και επιλέγουμε το χαρακτηριστικό «result» ως αυτό το οποίο δείχνει σε ποια κλάση ανήκει κάθε φορά το υπόδειγμα. Το

χαρακτηριστικό «result» αποτελεί μία δίτιμη μεταβλητή με τη τιμή **0 να δηλώνει την «ήττα» και την τιμή 1 τη «νίκη».**

3. Προχωρούμε στη διαδικασία διακριτοποίησης των δεδομένων μας, ακολουθώντας τα παρακάτω βήματα:

«Choose→weka→filters→unsupervised→attribute»

Έπειτα, επιλέγουμε από τη λίστα διαθέσιμων επιλογών ” το “Numeric to Nominal” κλικάρουμε “Filter→Filtering Capabilities”

- ✓ Numeric Attributes
- ✓ Numeric class

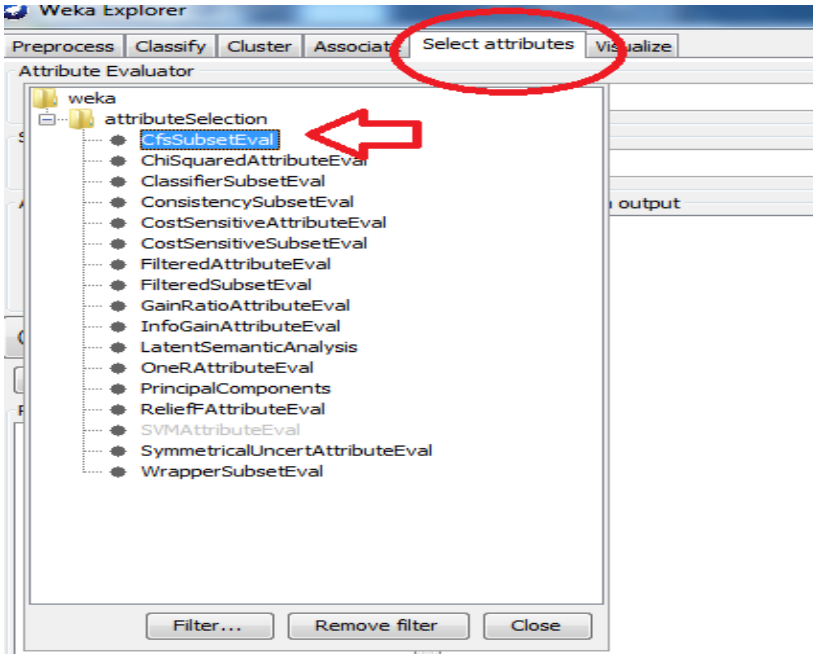
Ok

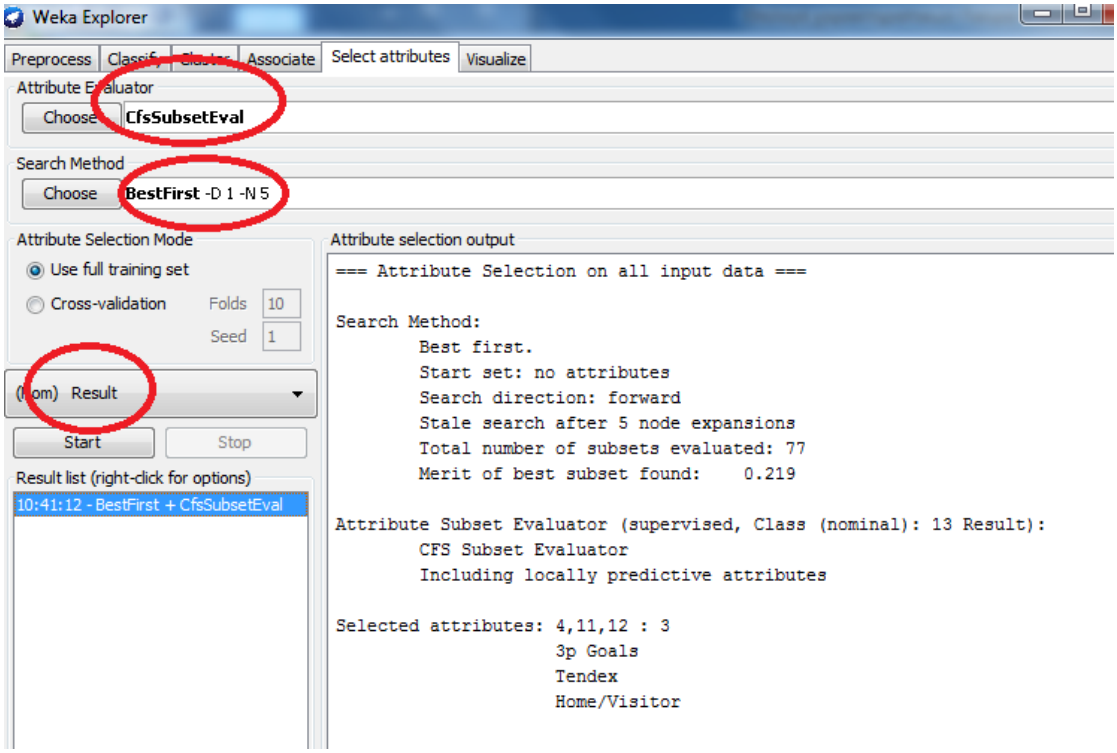
και →Apply.

Επιλέγουμε στο πάνελ την καρτέλα «Select attributes» όπου περιλαμβάνονται ποικίλες τεχνικές επιλογής χαρακτηριστικών με διαφορετικές δυνατότητες συνδυασμών για το χρήστη η κάθε μία.

Ακολουθούμε την παρακάτω διαδικασία»:

«Choose→Weka→attributeSelection→CfsSubsetEval»





Έπειτα, επιλέγουμε ως «Search method», τη μέθοδο BestFirst, ως «Attribute Selection Mode» το «Use Full Training Set» και ως μεταβλητή-στόχο τη μεταβλητή «Result» και →Start.

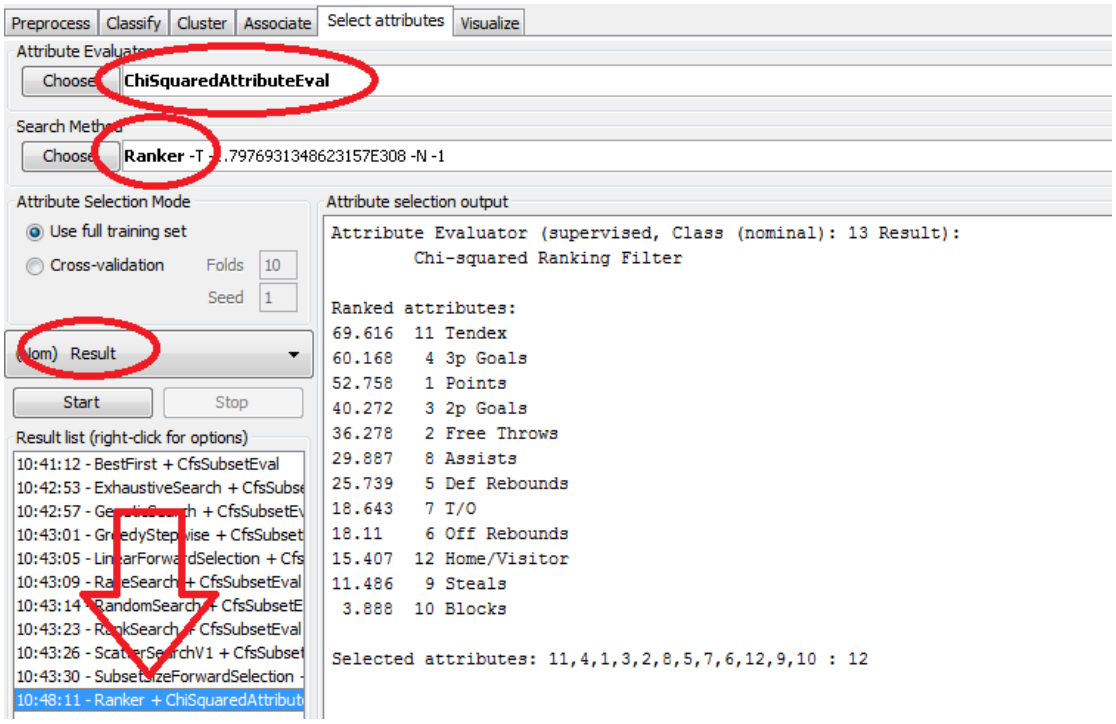
Επιλέγονται **3** μεταβλητές (από τις 12 συνολικά) ως σημαντικές, δηλαδή μόνο οι μεταβλητές Points, Tendex και Home/Visitor επηρρεάζουν σημαντικά τη μεταβλητή Result.

Selected attributes: 4,11,12: **3**
3p Goals
Tendex
Home/Visitor

<u>Παρατήρηση</u>

Επιλέγοντας οποιαδήποτε μέθοδο από τη λίστα διαθέσιμων επιλογών «Search method» (εκτός της Ranker που δεν μπορεί να υλοποιηθεί σε συνδυασμό με τη CfsSubsetEval) προκύπτουν οι ίδιες τρεις σημαντικές μεταβλητές, οι οποίες ιεραρχικά ταξινομούνται ως 1. 3p Goals 2. Tendex 3. Home/Visitor.

Έπειτα, για να αξιολογήσουμε τη σημαντικότητα κάθε μεταβλητής σε σχέση με τη μεταβλητή στόχο «Result», επιλέγουμε ως «Attribute Evaluator» τη μέθοδο «ChiSquaredAttributeEval» και ως «Search method» τη μέθοδο «Ranker», η οποία αποτελεί και την εξ ορισμού επιλογή του Weka και → Start.



Οι **12** μεταβλητές κατατάσσονται (ως ακολούθως) ιεραρχικά ανάλογα με το πόσο επηρεάζουν τη μεταβλητή-στόχο Result, ή με άλλα λόγια ανάλογα με το βαθμό της προβλεπτικής τους ικανότητας.

Selected attributes: 11,4,1,3,2,8,5,7,6,12,9,10 : **12**

Ranked attributes:

69.61	11 Tendex
60.16	4 3p Goals
52.75	1 Points
40.27	3 2p Goals
36.27	2 Free Throws
29.88	8 Assists
25.73	5 Def Rebounds
18.64	7 T/O
18.11	6 Off Rebounds
15.40	12 Home/Visitor
11.48	9 Steals
3.88	10 Blocks

Μεταβλητή-στόχος: Tendex

Τώρα θα εφαρμόσουμε την αντίστοιχη διαδικασία (όπως αυτή παρουσιάστηκε παραπάνω) θεωρώντας ως μεταβλητή στόχο τη μεταβλητή «Tendex».

1. Ακολουθούμε τα βήματα

Explorer→Preprocess→Open File

και «φορτώνουμε» το αρχείο "Lavrio 4 seasons.csv", το οποίο αφορά στην ομάδα του Λαυρίου και περιλαμβάνει τις υπό εξέταση μεταβλητές για τέσσερις συνολικά χρονικές περιόδους 2011-2012, 2012-2013, 2013-2014 και 2014-2015.

2. Έπειτα, αφαιρούμε τη μεταβλητή «Lavrio» (τικάρουμε και Remove), η οποία αποτελεί τη μεταβλητή- Index η οποία απαριθμεί τα παιχνίδια ($1^{ο}$ παιχνίδι, $2^{ο}$ παιχνίδι κ.λπ.) των τεσσάρων χρονικών περιόδων, και επιλέγουμε το χαρακτηριστικό «Tendex» ως αυτό το οποίο δείχνει σε ποια κλάση ανήκει κάθε φορά το υπόδειγμα.

3. Προχωρούμε στη διαδικασία διακριτοποίησης των δεδομένων μας, ακολουθώντας τα παρακάτω βήματα:

«Choose→weka→filters→unsupervised→attribute»

Έπειτα, επιλέγουμε από τη λίστα διαθέσιμων επιλογών " το "Numeric to Nominal" κλικάρουμε "Filter→Filtering Capabilities"

- ✓ Numeric Attributes
- ✓ Numeric class

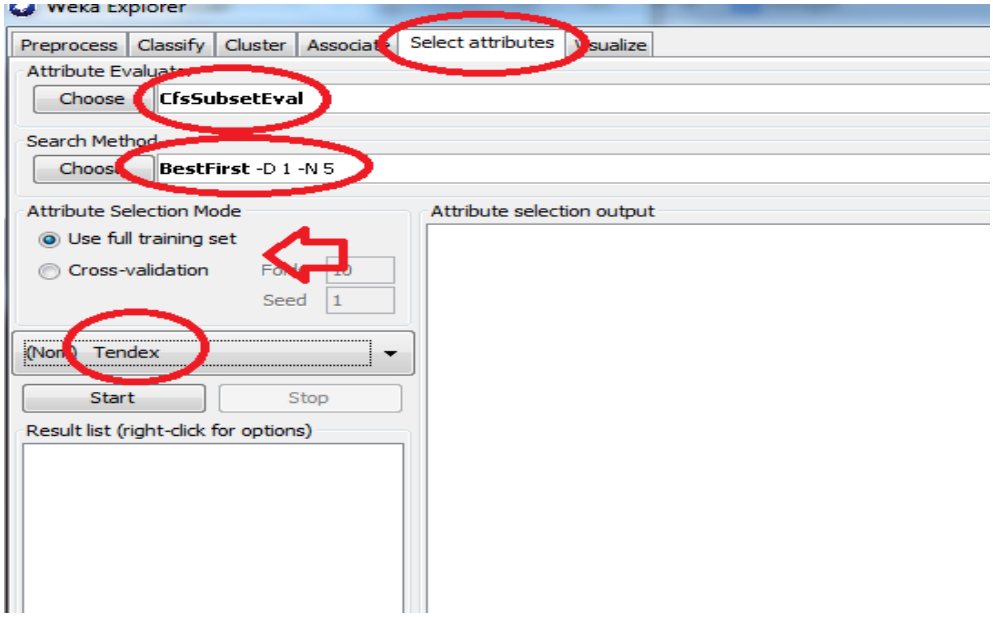
Ok

και →Apply.

Επιλέγουμε στο πάνελ την καρτέλα «Select attributes».

Ακολουθούμε την παρακάτω διαδικασία»:

«Choose→Weka→attributeSelection→CfsSubsetEval»



Έπειτα, επιλέγουμε ως «Search method», τη μέθοδο BestFirst, ως «Attribute Selection Mode» το «Use Full Training Set» και ως μεταβλητή-στόχο τη μεταβλητή «Tendex» και →Start.

Επιλέγονται και οι **12** μεταβλητές ως σημαντικές,

δηλαδή και οι 12 επηρεάζουν σημαντικά τη

μεταβλητή Tendex.

Selected attributes:
1,2,3,4,5,6,7,8,9,10,12,13 : **12**

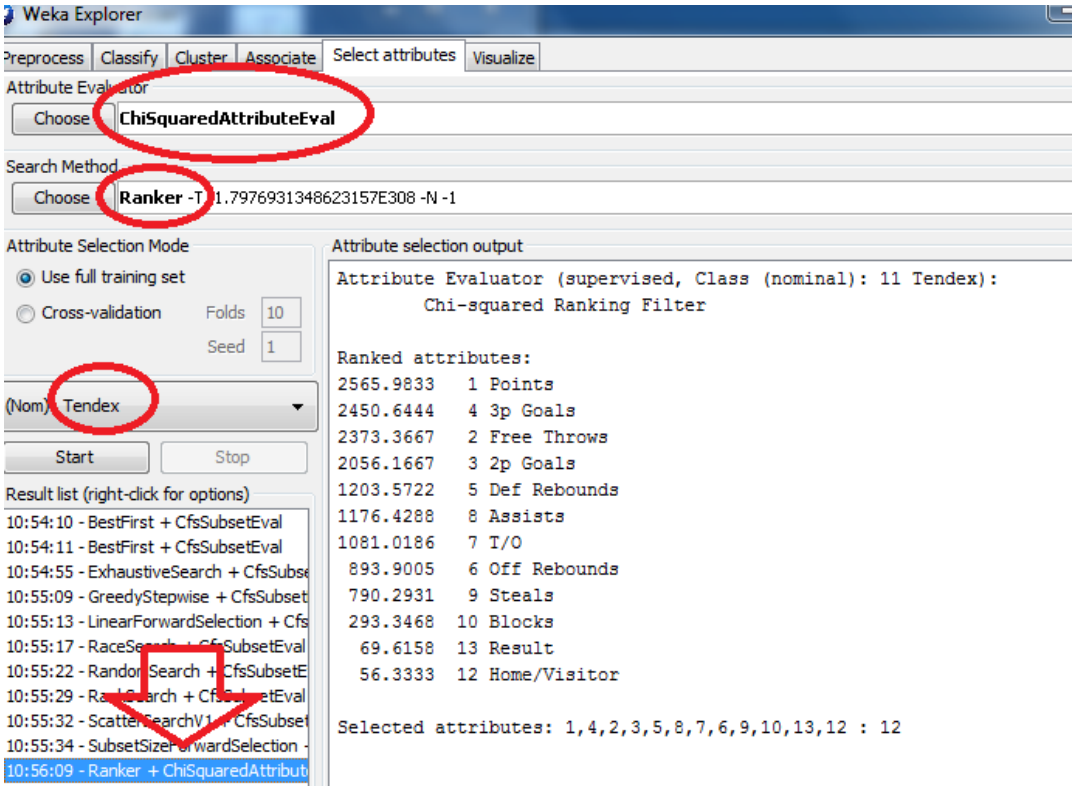
Points, Free Throws, 2p Goals, 3p Goals, Def Rebounds, Off Rebounds, T/O, Assists, Steals, Blocks, Home/Visitor, Result

<u>Παρατήρηση</u>

Επιλέγοντας οποιαδήποτε μέθοδο από τη λίστα διαθέσιμων επιλογών «Search method» (εκτός της Ranker που δεν μπορεί να υλοποιηθεί σε συνδυασμό με τη CfsSubsetEval) προκύπτουν τα ίδια ακριβώς αποτελέσματα με παραπάνω.

Έπειτα, για να αξιολογήσουμε τη σημαντικότητα κάθε μεταβλητής σε σχέση με τη μεταβλητή στόχο «Tendex», επιλέγουμε ως «Attribute Evaluator» τη μέθοδο

«ChiSquaredAttributeEval» και ως «Search method» τη μέθοδο «Ranker», η οποία αποτελεί και την εξ'ορισμού επιλογή του Weka και → Start.



Οι **12** μεταβλητές κατατάσσονται (ως ακολούθως) ιεραρχικά ανάλογα με το πόσο επηρεάζουν τη μεταβλητή-στόχο Tendex.

Selected attributes: 1,4,2,3,5,8,7,6,9,10,13,12 : **12**

Ranked attributes:

2565.98 1 Points

2450.64 4 3p Goals

2373.36 2 Free Throws

2056.16 3 2p Goals

1203.57 5 Def Rebounds

1176.42 8 Assists

1081.01 7 T/O

893.90 6 Off Rebounds

790.29 9 Steals

293.34 10 Blocks

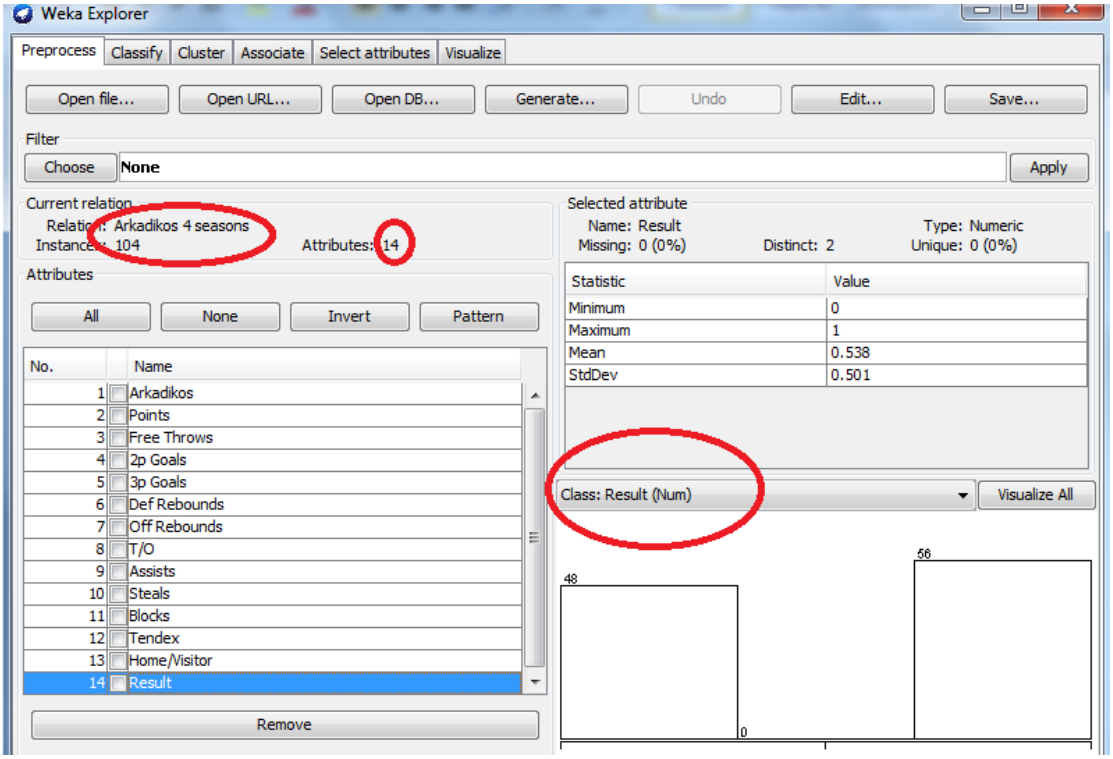
69.61 13 Result

56.33 12 Home/Visitor

14. 2 Ταξινόμηση μέσω δέντρων απόφασης

14.2.1 Δέντρα απόφασης για την ομάδα του Αρκαδικού

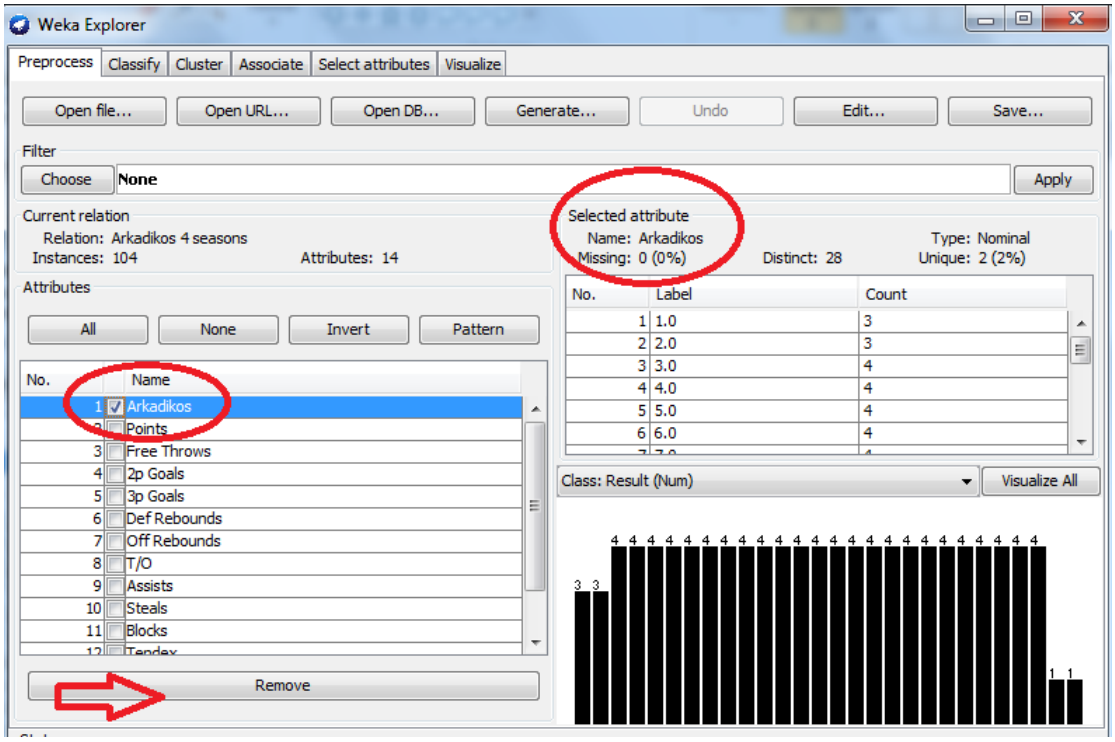
Ξεκινώντας το Weka, ακολουθούμε τα βήματα «Explorer→Preprocess→Open File» και «φορτώνοντας» το αρχείο "Arkadikos 4 seasons.csv", στην οθόνη θα εμφανιστεί το παράθυρο το οποίο απεικονίζεται παρακάτω. Το αρχείο "Arkadikos 4 seasons.csv" αφορά στην ομάδα του Αρκαδικού και περιλαμβάνει τις υπό εξέταση μεταβλητές για τέσσερις συνολικά χρονικές περιόδους 2011-2012, 2012-2013, 2013-2014 και 2014-2015.



Εικόνα 21: Το πάνελ Preprocess- Φόρτωση αρχείου "Arkadikos 4 seasons.csv

Στο παράθυρο αυτό, στο μέσο και αριστερά, αναγράφεται ότι το συγκεκριμένο σύνολο δεδομένων "Arkadikos 4 seasons.csv περιλαμβάνει 104 υποδείγματα (instances) και κάθε υπόδειγμα απαρτίζεται από 14 χαρακτηριστικά – μεταβλητές. Στο συγκεκριμένο παράδειγμα το χαρακτηριστικό «result» έχει επιλεγεί ως αυτό το οποίο δείχνει σε ποια κλάση ανήκει κάθε φορά το υπόδειγμα. Συνήθως πρόκειται (χωρίς να είναι δεσμευτικό) για το τελευταίο χαρακτηριστικό που καταχωρούμε, και στην περίπτωση που έχουμε πρόβλημα κατηγοριοποίησης δείχνει την κατηγορία στην οποία ανήκει το υπόδειγμα, ενώ για προβλήματα παλινδρόμησης δείχνει την τιμή της παραμέτρου που μας ενδιαφέρει. Το χαρακτηριστικό «result» αποτελεί μία δίτιμη μεταβλητή με τη τιμή **0 να δηλώνει την «ήττα» και την τιμή 1 τη «νίκη».**

Πριν προχωρήσουμε στη διαδικασία διακριτοποίησης των δεδομένων μας, αρχικά αφαιρούμε τη μεταβλητή «Arkadikos» (τικάρουμε και Remove), η οποία έρχεται πρώτη στη λίστα των διαθέσιμων χαρακτηριστικών μας, και αποτελεί τη μεταβλητή-Index η οποία απαριθμεί τα παιχνίδια ($1^{ο}$ παιχνίδι, $2^{ο}$ παιχνίδι κ.λπ.) των τεσσάρων χρονικών περιόδων.



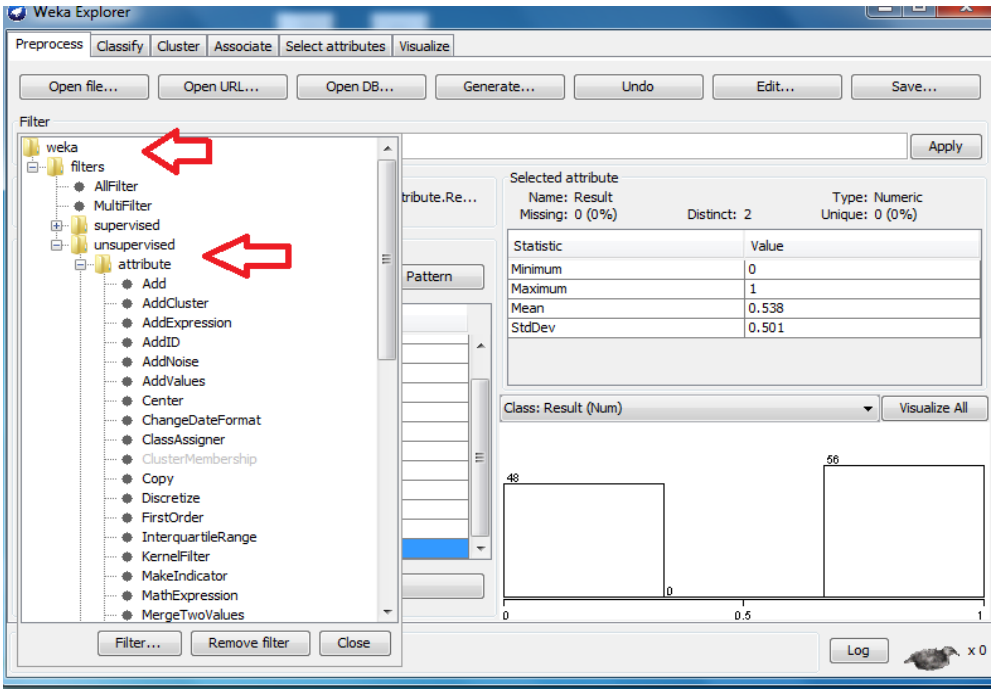
Εικόνα 22: Αφαίρεση μεταβλητής "Arkadikos" Index

Το επόμενο και σημαντικό βήμα αποτελεί η διακριτοποίηση των δεδομένων μας. Ένας γενικός τρόπος που μπορούμε να κάνουμε διακριτοποίηση των αριθμητικών ή συνεχών μεταβλητών είναι επιλέγοντας από τη λίστα το φίλτρο «weka→ filters→ unsupervised→ attribute→ discretize».

Εναλλακτικά, δεδομένου ότι οι μεταβλητές του συνόλου δεδομένων μας είναι αριθμητικές ή συνεχείς (εκτός της μεταβλητής "home / visitor" και της μεταβλητής class "result", οι οποίες είναι δίτιμες 0/1) συνιστάται συνήθως μετά από εισαγωγή csv αρχείου να μετατρέπουμε αυτές τις αριθμητικές μεταβλητές αντίστοιχα σε ονομαστικές.

Ακολουθούμε τα παρακάτω βήματα:

«Choose→weka→filters→unsupervised→attribute»



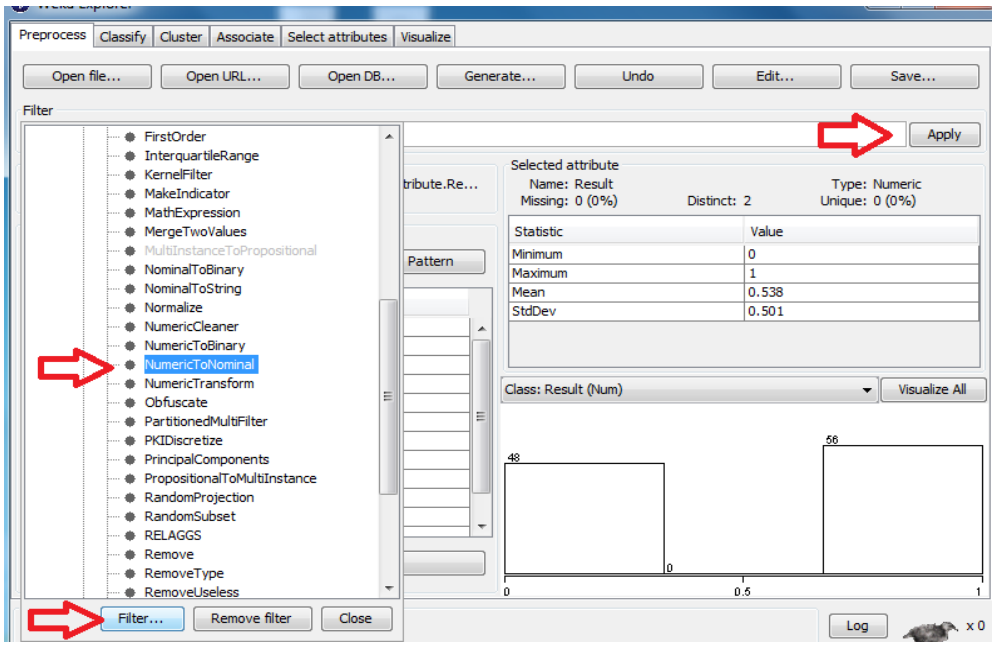
Εικόνα 23: Επιλογή φίλτρου

Έπειτα, επιλέγουμε από τη λίστα διαθέσιμων επιλογών ” το “Numeric to Nominal” κλικάρουμε “Filter→Filtering Capabilities”

- ✓ Numeric Attributes
- ✓ Numeric class

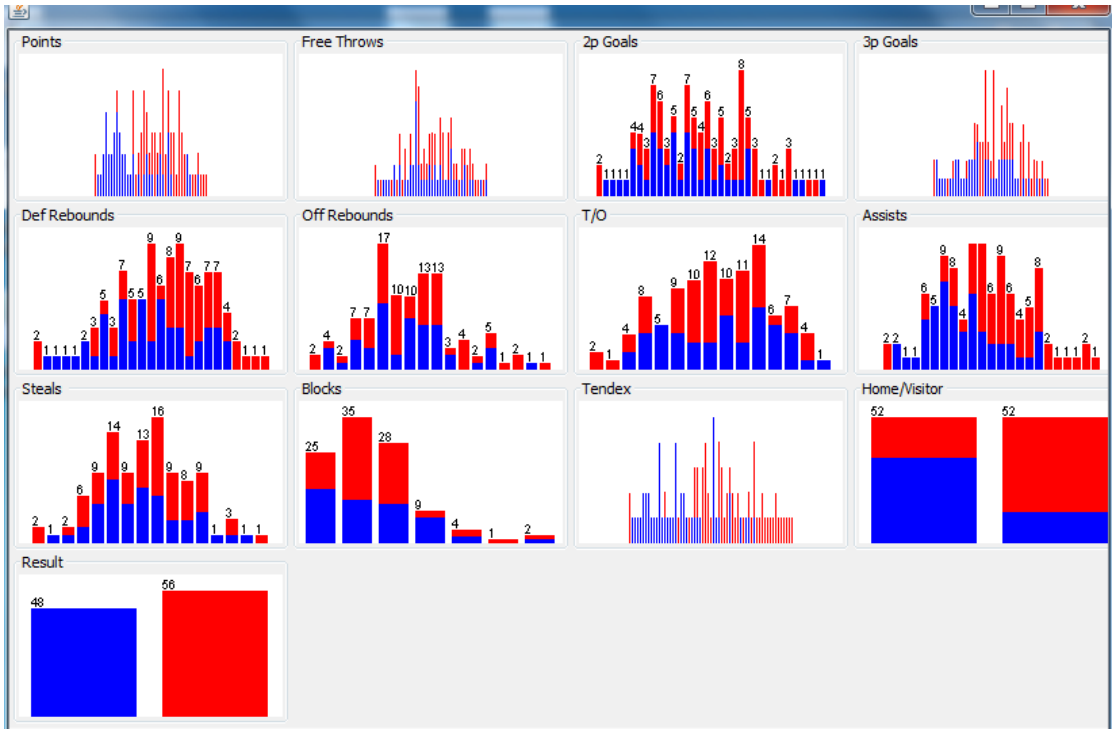
Ok

και →Apply.



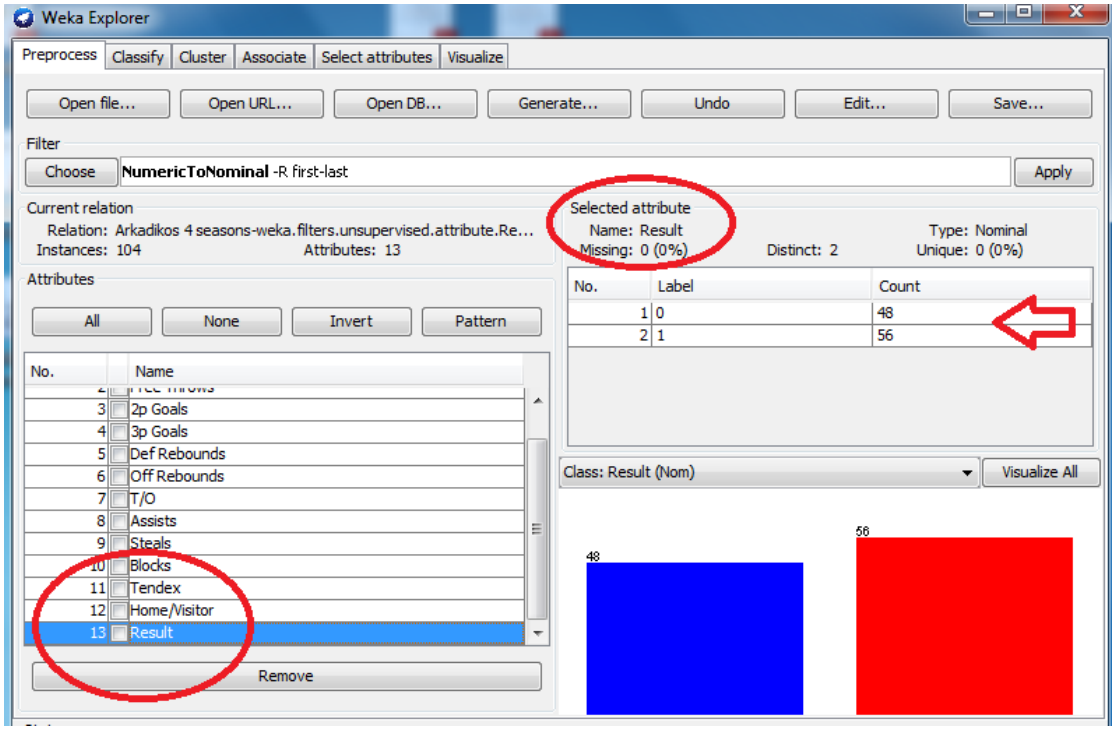
Εικόνα 24: Διαδικασία διακριτοποίησης

Έπειτα, επιλέγοντας το "Visualize all" λαμβάνουμε σε νέο παράθυρο την οπτικοποίηση και την απεικόνιση της κατανομής των 11 συνολικά μεταβλητών που μετασχηματίστηκαν. Σημειώνεται ότι στο παράθυρο οπτικοποίησης συμπεριλαμβάνονται και οι μη μετασχηματισμένες δίτιμες εξ αρχής μεταβλητές, η μεταβλητή home/visitor και η μεταβλητή-στόχος "result".



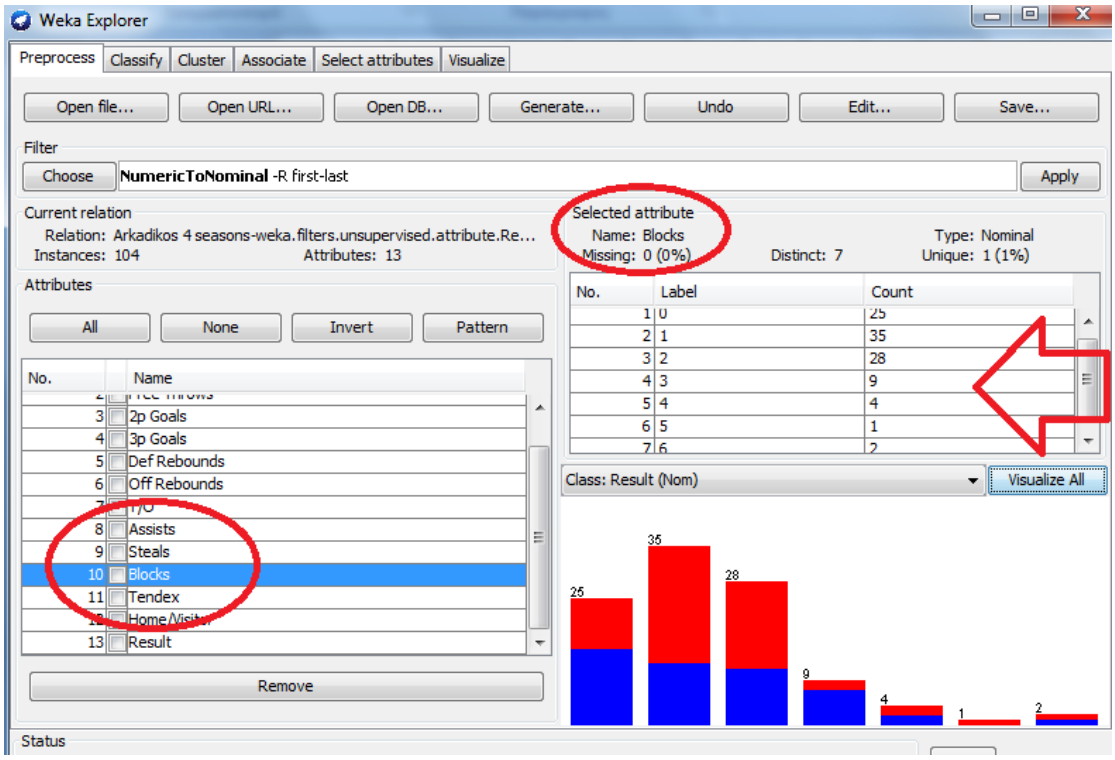
Εικόνα 25 : Οπτικοποίηση των μεταβλητών

Τα παραπάνω ιστογράμματα μας βοηθούν στην κατανόηση των δεδομένων, δεδομένου ότι υποδεικνύουν την κατανομή των τιμών των ονομαστικών χαρακτηριστικών. Για παράδειγμα, για τη μεταβλητή στόχο "Result" παρατηρούμε ότι έχουμε 48 «ήττες» και 56 «νίκες».



Εικόνα 26: Η κατανομή των τιμών της μεταβλητής "Result"

Για παράδειγμα, μεμονωμένα για τη μεταβλητή "Blocks" παρατηρούμε ότι έχουμε 0 Blocks σε 25 παιχνίδια, 1 Block σε 35 παιχνίδια , 2 Blocks σε 28 παιχνίδια, 3 Blocks σε 9 παιχνίδια, 4 Blocks σε 4 παιχνίδια, 5 Blocks σε 1 παιχνίδι, 6 Blocks σε 2 παιχνίδια, όπως άλλωστε φαίνεται παρακάτω στην Εικόνα 27.



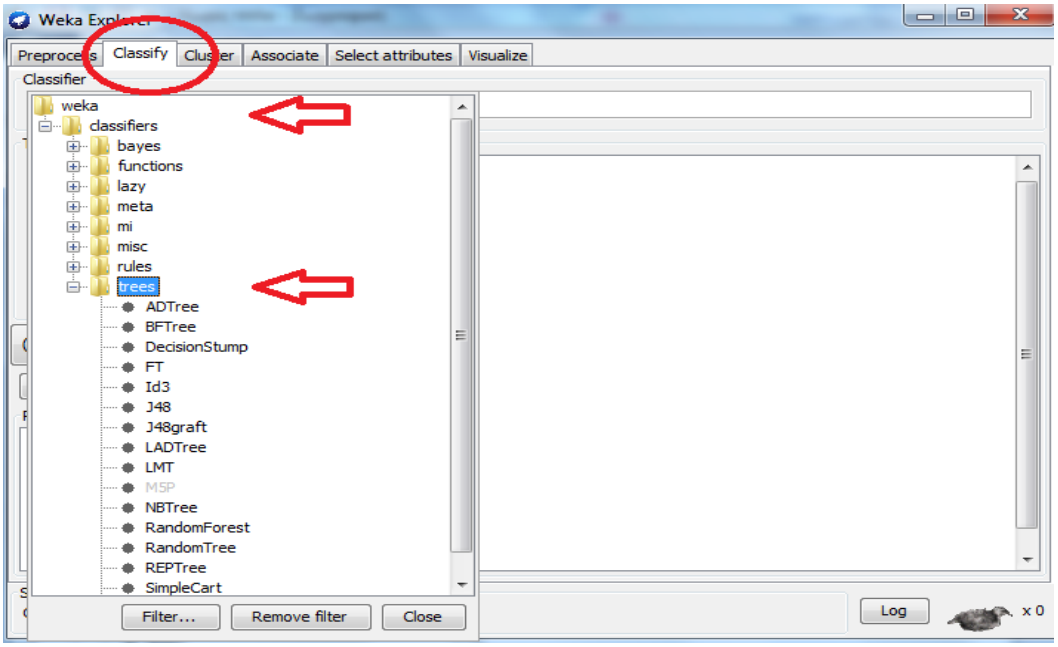
Εικόνα 27: Η κατανομή των τιμών της μεταβλητής "Blocks"

Επιλέγουμε στο πάνελ την καρτέλα «Classify» όπου περιλαμβάνονται ποικίλες τεχνικές ταξινόμησης με διαφορετικές δυνατότητες για το χρήστη η κάθε μία.

Αρχικά θα εφαρμόσουμε τεχνικές στις οποίες η κατηγοριοποίηση των δεδομένων υλοποιείται μέσω της κατασκευής των δέντρων ταξινόμησης.

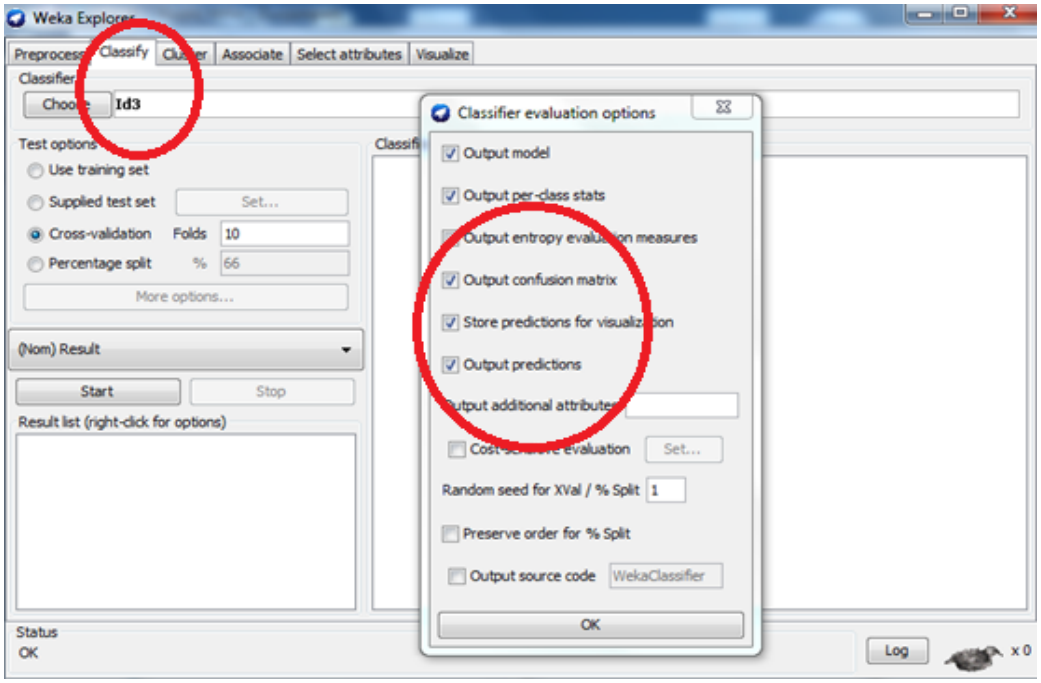
Ακολουθούμε την παρακάτω διαδικασία»:

«Choose→Weka→Classifiers→Trees»



- **Αλγόριθμος ID 3**

Μέσα από τη λίστα των διαθέσιμων τεχνικών των δέντρων ταξινόμησης «trees» επιλέγουμε τον αλγόριθμο ID-3, έπειτα επιλέγουμε στο "Test options" το "Cross-Validation" αφήνοντας την επιλογή «10», έτσι ώστε να υλοποιηθεί η διαδικασία της διασταυρωμένης επικύρωσης με 10 πεδία για την τελική αξιολόγηση της απόδοσης της μεθόδου, και κλικάρουμε το tab "More Options" έτσι ώστε να ανοίξει ένα νέο παράθυρο διαλόγου, το "Classifier evaluation options", το οποίο μας παρέχει τη δυνατότητα να επιλέξουμε τα επιθυμητά εξαγόμενα αποτελέσματα τα οποία θα εμφανιστούν μετά την εκτέλεση του αλγορίθμου.

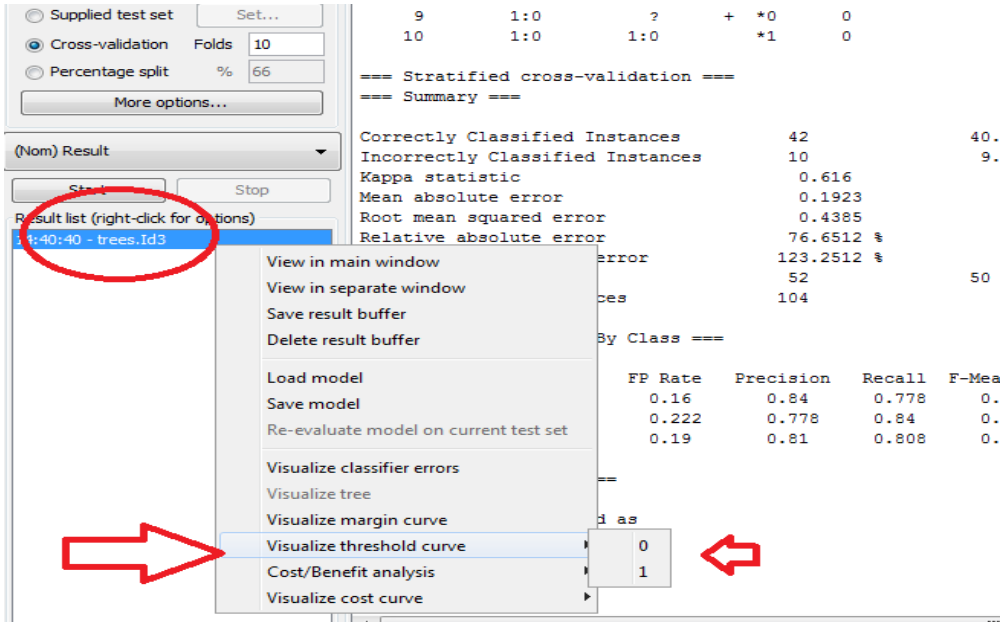


Εικόνα 28: Υλοποίηση αλγορίθμου ID-3

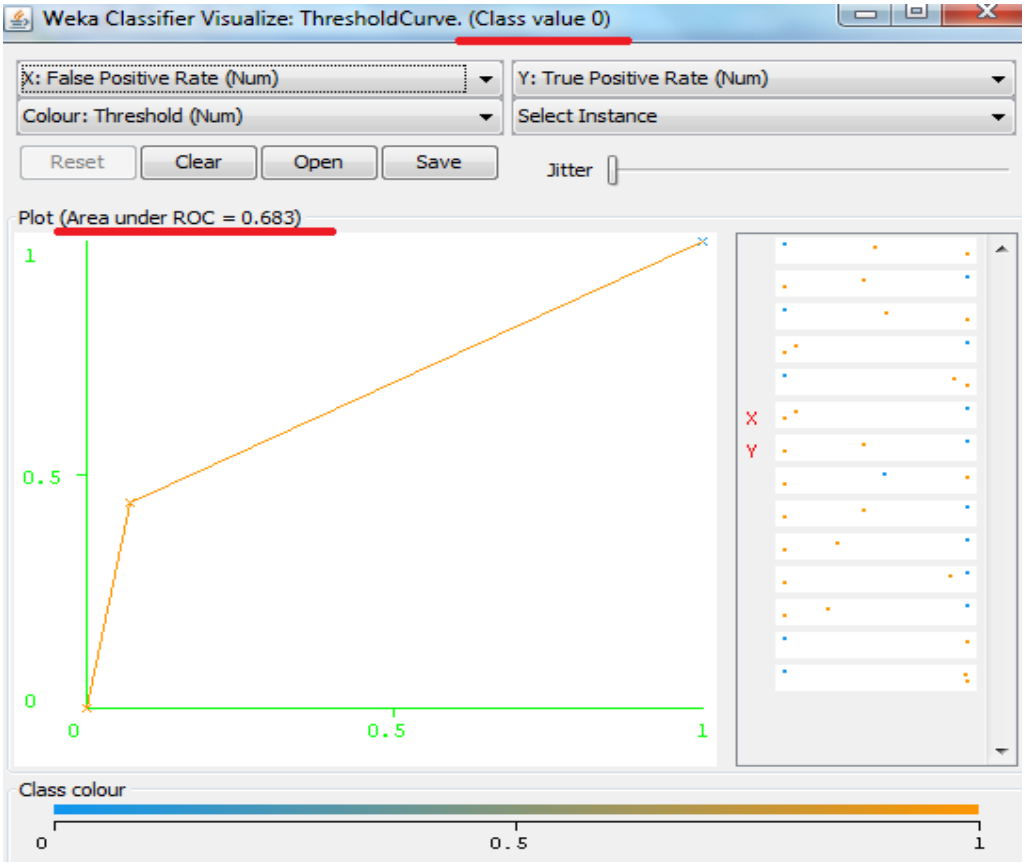
```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          42               40.3846 %
Incorrectly Classified Instances        10                9.6154 %
Kappa statistic                          0.616
Mean absolute error                      0.1923
Root mean squared error                  0.4385
Relative absolute error                 76.6512 %
Root relative squared error            123.2512 %
UnClassified Instances                  52               50      %
Total Number of Instances              104

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.778     0.16       0.84      0.778     0.808      0.683     0
                   0.84      0.222      0.778     0.84      0.808      0.625     1
Weighted Avg.      0.808     0.19       0.81      0.808     0.808      0.655

=== Confusion Matrix ===

  a  b   <-- classified as
 21  6 |  a = 0
  4 21 |  b = 1
```

Εικόνα 29: Εξαγόμενα αποτελέσματα ID-3 για το σύνολο δεδομένων "Arkadikos 4 Seasons.csv

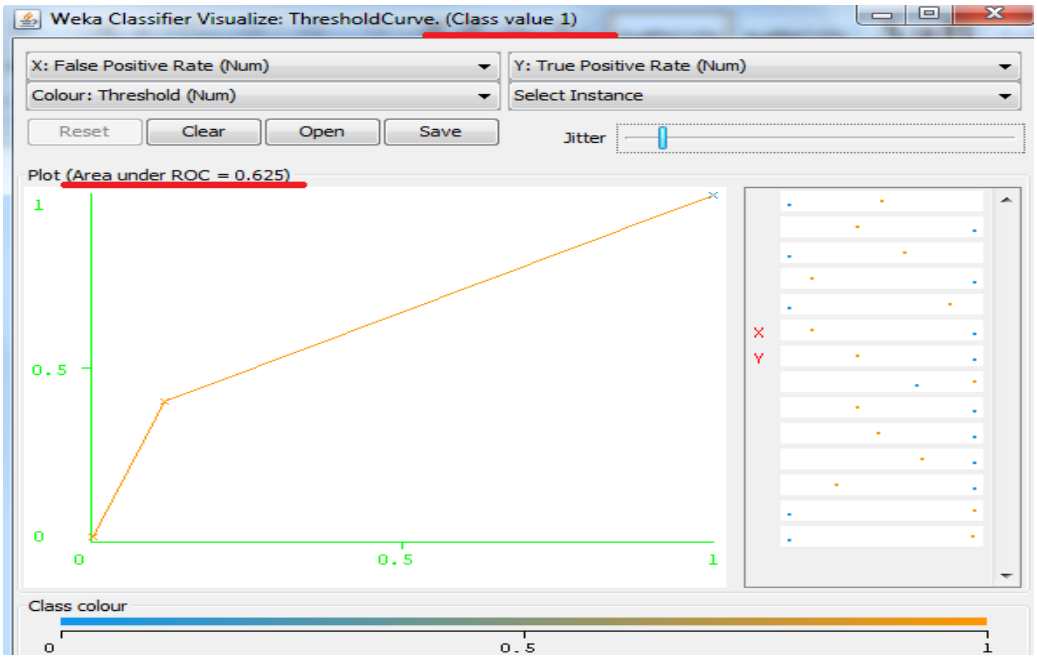
Έπειτα, για την γραφική αναπαράσταση της καμπύλης λειτουργικού χαρακτηριστικού δέκτη (ROC) και τον υπολογισμό του εμβαδού κάτω από την καμπύλη ROC (Area Under the Curve-AUC) , κάνουμε δεξί κλικ στο "trees.Id3" και επιλέγουμε στο νέο παράθυρο διαλόγου που ανοίγει το "Visualize threshold curve" και για τις δύο κλάσεις 0 και 1, αντίστοιχα.



Εικόνα 30: Αναπαράσταση της ROC καμπύλης και υπολογισμός του εμβαδού κάτω από την καμπύλη AUC



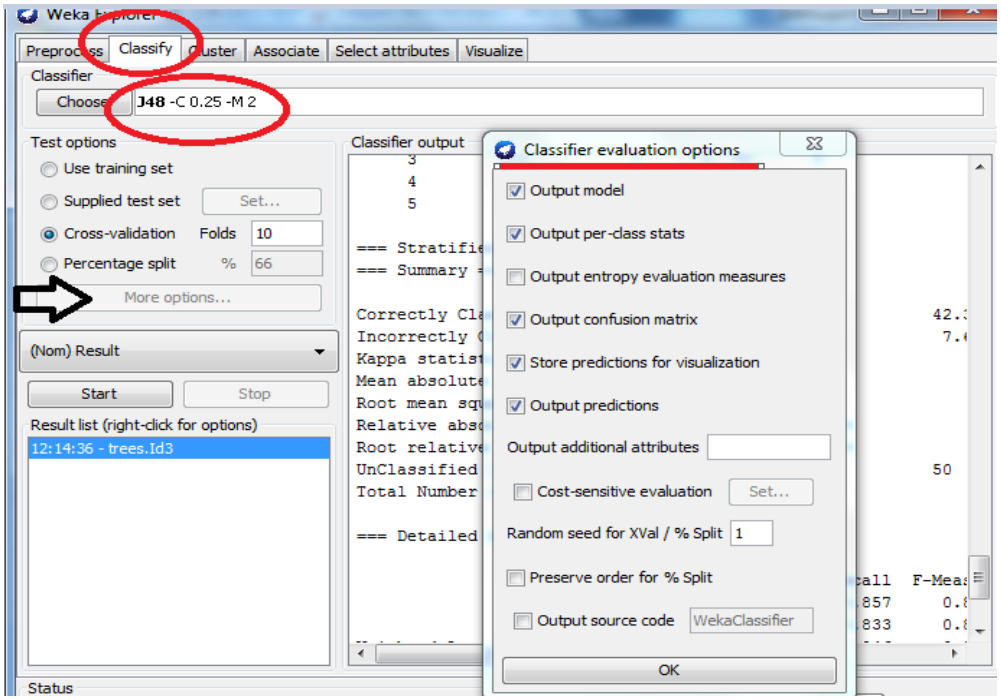
Εικόνα 31: ROC καμπύλη για την κλάση 0 – AUC=0.683



Εικόνα 32: ROC καμπύλη για την κλάση 1 – AUC=0.625

Αλγόριθμος J48

Μέσα από τη λίστα των διαθέσιμων τεχνικών των δέντρων ταξινόμησης «trees» επιλέγουμε τον αλγόριθμο J48, έπειτα επιλέγουμε στο "Test options" το "Cross-Validation" αφήνοντας την επιλογή «10», έτσι ώστε να υλοποιηθεί η διαδικασία της διασταυρωμένης επικύρωσης με 10 πεδία για την τελική αξιολόγηση της απόδοση της μεθόδου, και κλικάρουμε το tab "More Options" έτσι ώστε να ανοίξει ένα νέο παράθυρο διαλόγου, το "Classifier evaluation options", το οποίο μας παρέχει τη δυνατότητα να επιλέξουμε τα επιθυμητά εξαγόμενα αποτελέσματα τα οποία θα εμφανιστούν μετά την εκτέλεση του αλγορίθμου.

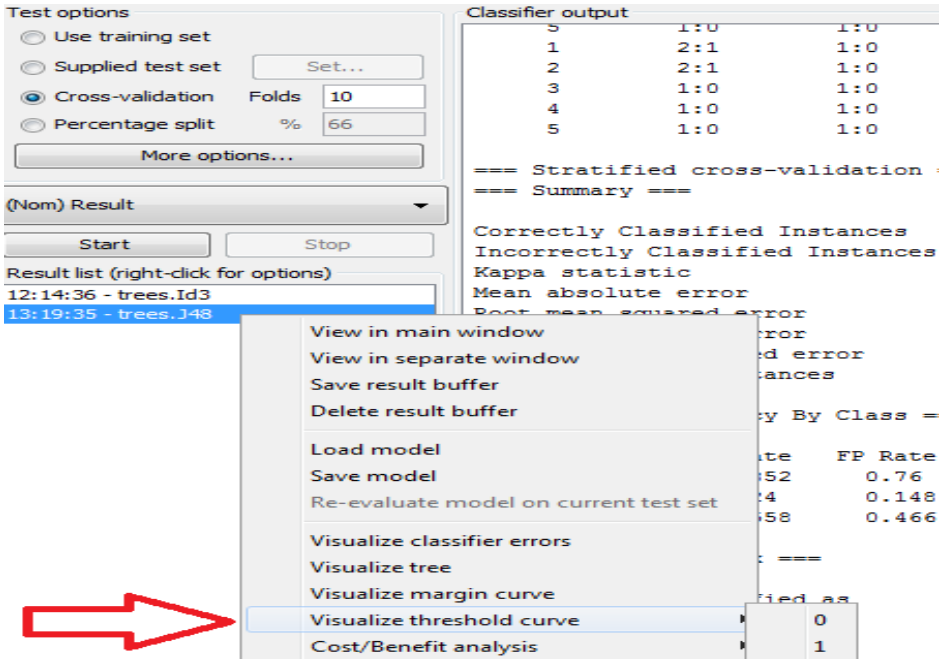


Εικόνα 33: Υλοποίηση αλγορίθμου J48

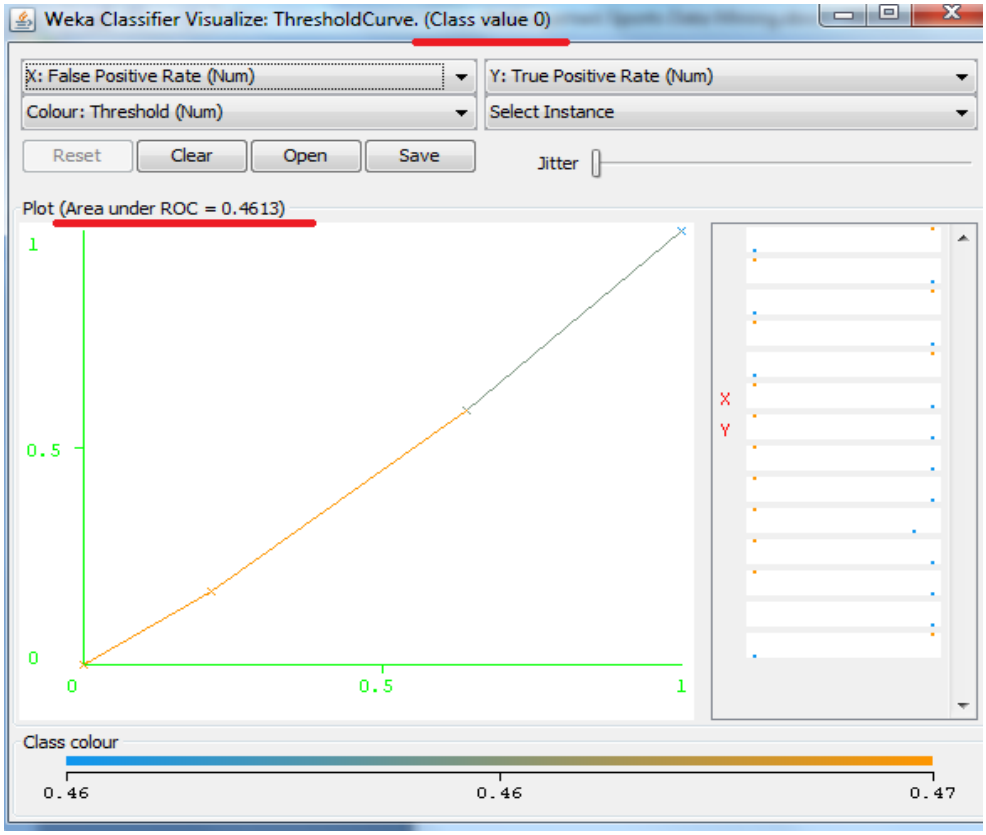
```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          56               53.8462 %
Incorrectly Classified Instances        48               46.1538 %
Kappa statistic                          0
Mean absolute error                      0.4973
Root mean squared error                  0.4988
Relative absolute error                 99.9887 %
Root relative squared error            100.0011 %
Total Number of Instances              104

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.461    0
                 1         1          0.538     1         0.7        0.461    1
Weighted Avg.    0.538     0.538      0.29      0.538     0.377      0.461

=== Confusion Matrix ===

  a  b   <-- classified as
  0 48 |  a = 0
  0 56 |  b = 1
```

Εικόνα 34: Εξαγόμενα αποτελέσματα J48 για το σύνολο δεδομένων "Arkadikos 4 Seasons.csv

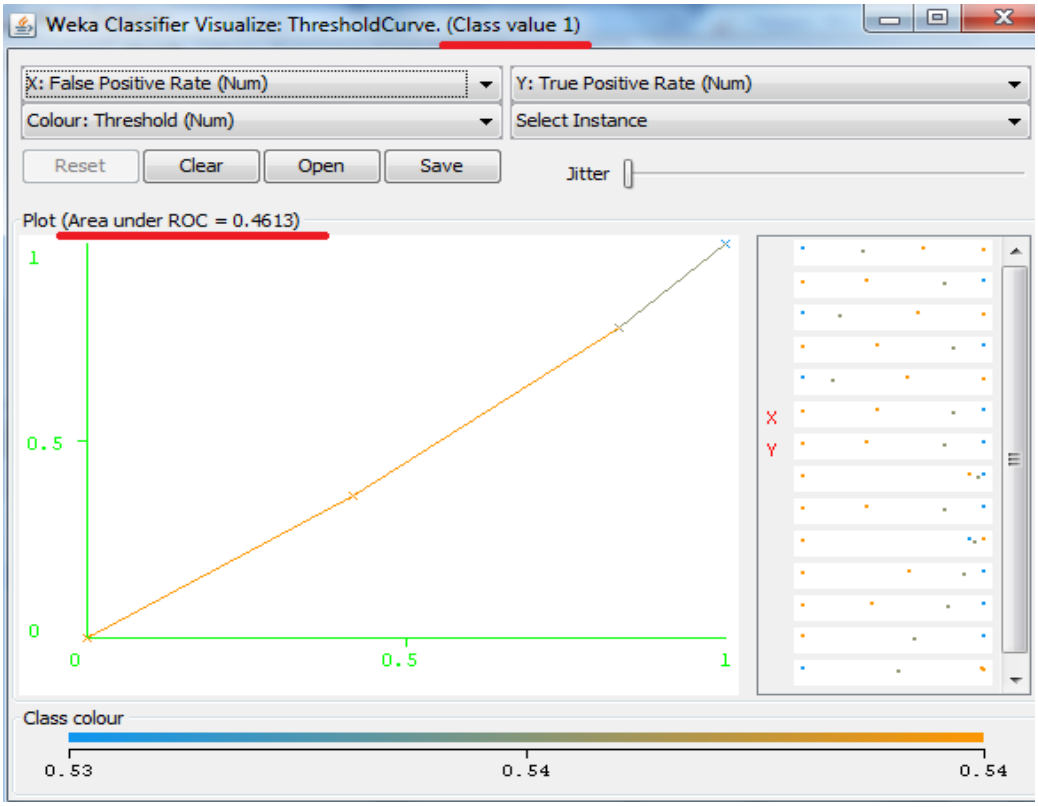
Έπειτα, για την γραφική αναπαράσταση της καμπύλης λειτουργικού χαρακτηριστικού δέκτη (ROC) και τον υπολογισμό του εμβαδού κάτω από την καμπύλη ROC (Area Under the Curve-AUC) , κάνουμε δεξί κλικ στο "trees.J48" και επιλέγουμε στο νέο παράθυρο διαλόγου που ανοίγει το "Visualize threshold curve" και για τις κλάσεις 0 και 1, αντίστοιχα.



Εικόνα 35: Αναπαράσταση της ROC καμπύλης και υπολογισμός του εμβαδού κάτω από την καμπύλη AUC



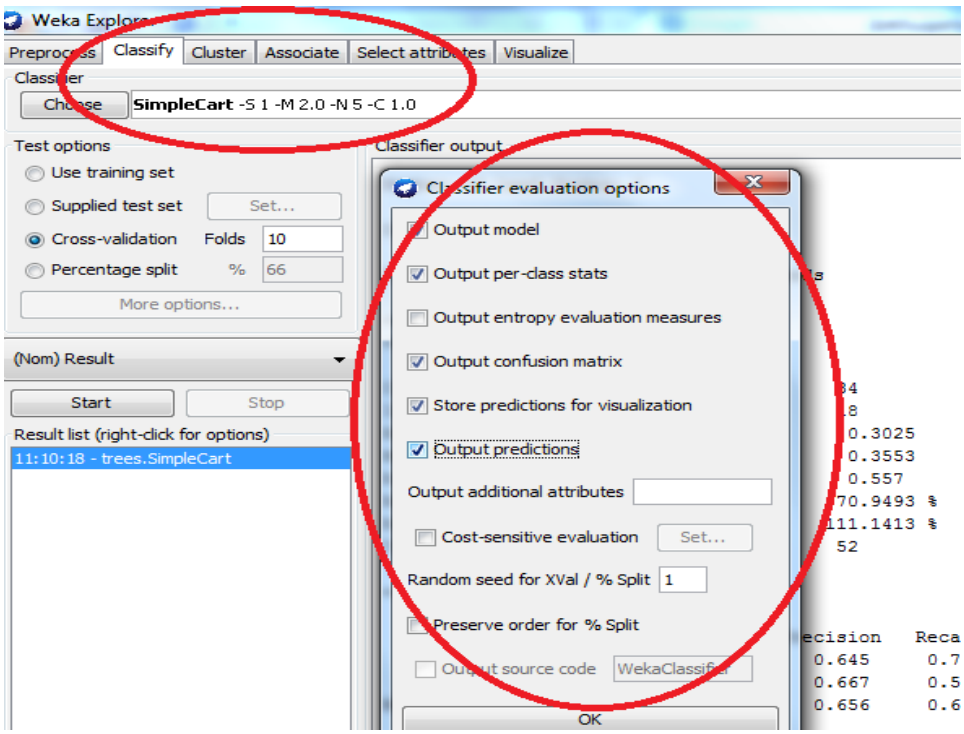
Εικόνα 36: ROC καμπύλη για την κλάση 0 – AUC=0.4613



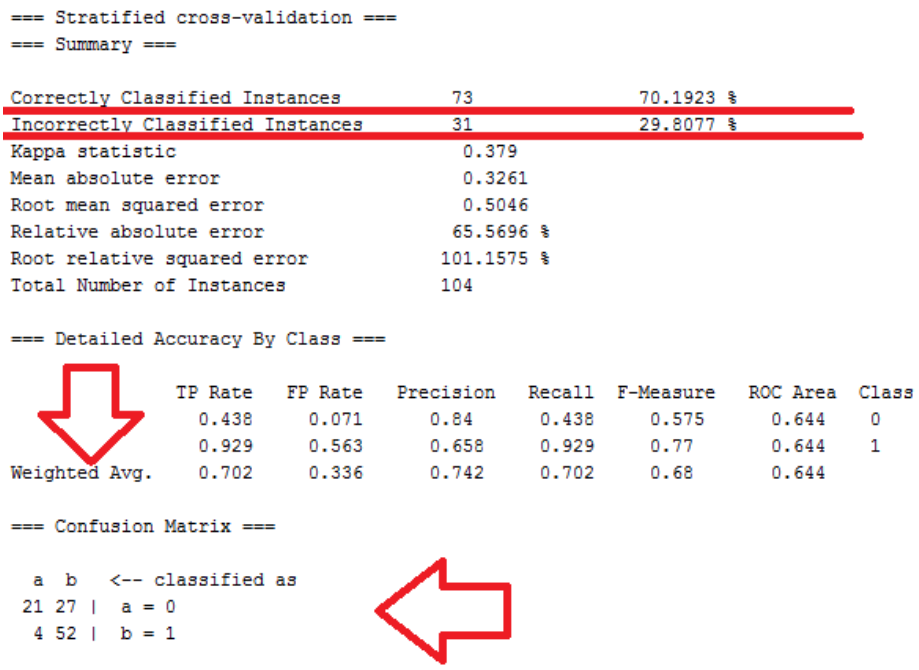
Εικόνα 37: ROC καμπύλη για την κλάση 1 – AUC=0.4613

Αλγόριθμος CART

Μέσα από τη λίστα των διαθέσιμων τεχνικών των δέντρων ταξινόμησης «trees» επιλέγουμε τον αλγόριθμο SimpleCart, έπειτα επιλέγουμε στο "Test options" το "Cross- Validation" αφήνοντας την επιλογή «10», έτσι ώστε να υλοποιηθεί η διαδικασία της διασταυρωμένης επικύρωσης με 10 πεδία για την τελική αξιολόγηση της απόδοση της μεθόδου, και κλικάρουμε το tab "More Options" έτσι ώστε να ανοίξει ένα νέο παράθυρο διαλόγου, το "Classifier evaluation options", το οποίο μας παρέχει τη δυνατότητα να επιλέξουμε τα επιθυμητά εξαγόμενα αποτελέσματα τα οποία θα εμφανιστούν μετά την εκτέλεση του αλγορίθμου.

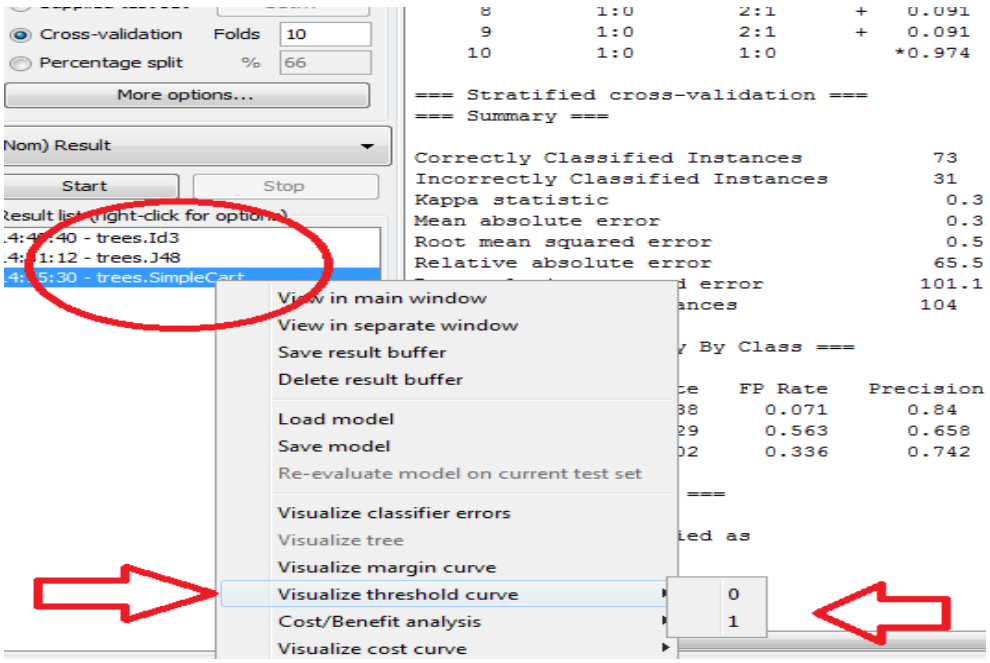


Εικόνα 38: Υλοποίηση αλγορίθμου SimpleCart

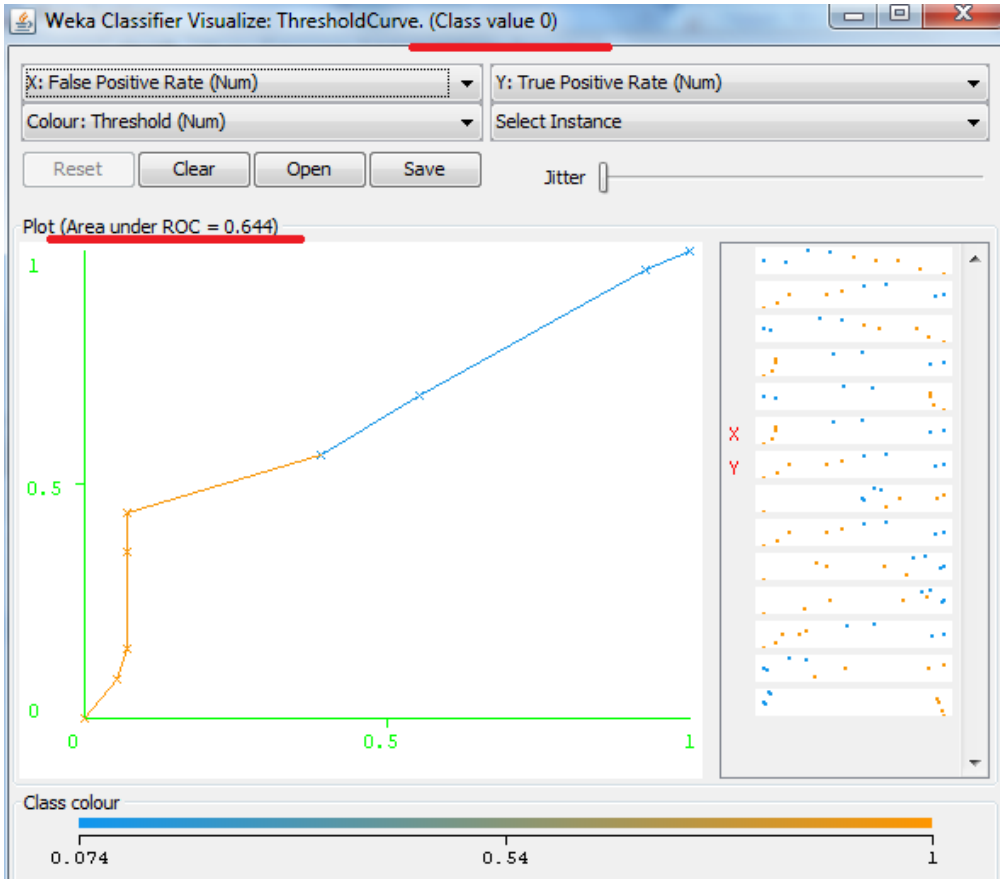


Εικόνα 39: Εξαγόμενα αποτελέσματα SimpleCart για το σύνολο δεδομένων "Arkadikos 4 Seasons.csv"

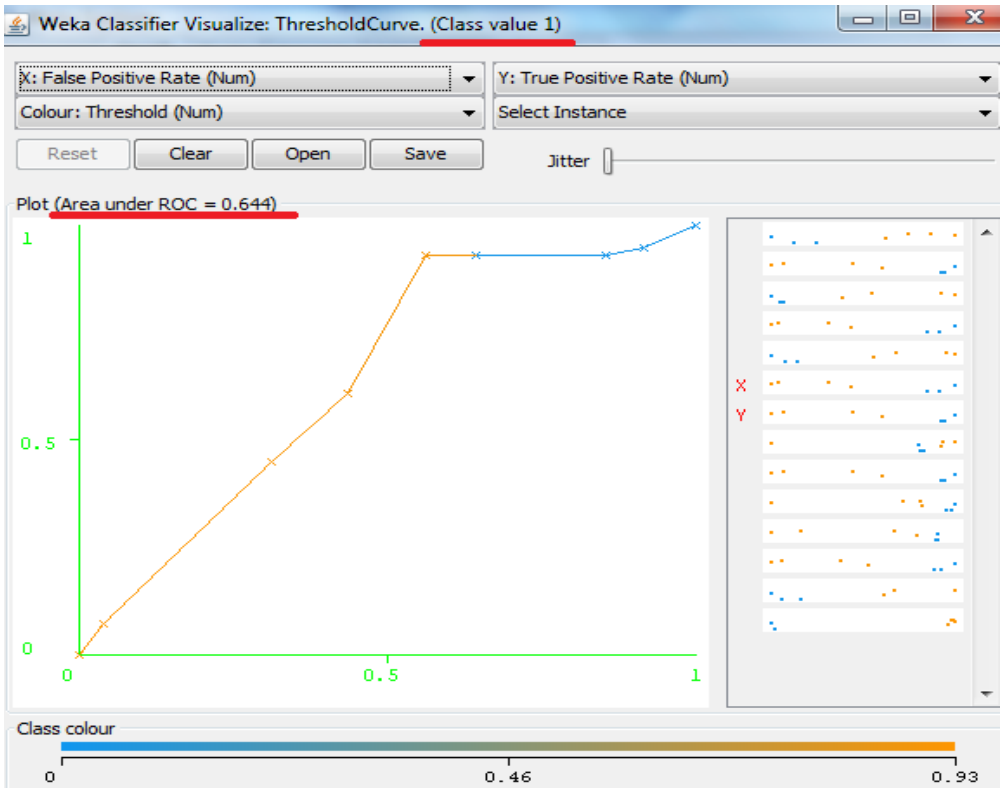
Έπειτα, για την γραφική αναπαράσταση της καμπύλης λειτουργικού χαρακτηριστικού δέκτη (ROC) και τον υπολογισμό του εμβαδού κάτω από την καμπύλη ROC (Area Under the Curve-AUC) , κάνουμε δεξί κλικ στο "trees.SimpleCart" και επιλέγουμε στο νέο παράθυρο διαλόγου που ανοίγει το "Visualize threshold curve" και για τις κλάσεις 0 και 1, αντίστοιχα.



Εικόνα 40: Αναπαράσταση της ROC καμπύλης και υπολογισμός του εμβαδού κάτω από την καμπύλη AUC



Εικόνα 41: ROC καμπύλη για την κλάση 0 – AUC=0.644



Εικόνα 42: ROC καμπύλη για την κλάση 1 – AUC=0.644

Πίνακας 3: Συγκριτικά αποτελέσματα δέντρων ταξινόμησης για το σύνολο δεδομένων "Arkadikos 4 Seasons.csv"

	Correctly Classified Instances - Accuracy	Incorrectly Classified Instances	Unclassified Instances	Avg. TP Rate	Avg. FP Rate	Avg. Precision	Avg. Recall	Avg. F-Measure	Avg. ROC Area
ID-3	40.38 %	9.62 %	50 %	0.808	0.19	0.81	0.808	0.808	0.655
J48	53.85 %	46.15 %	0 %	0.538	0.538	0.29	0.538	0.377	0.461
Simple CART	70.20 %	29.80 %	0 %	0.702	0.336	0.742	0.702	0.680	0.644

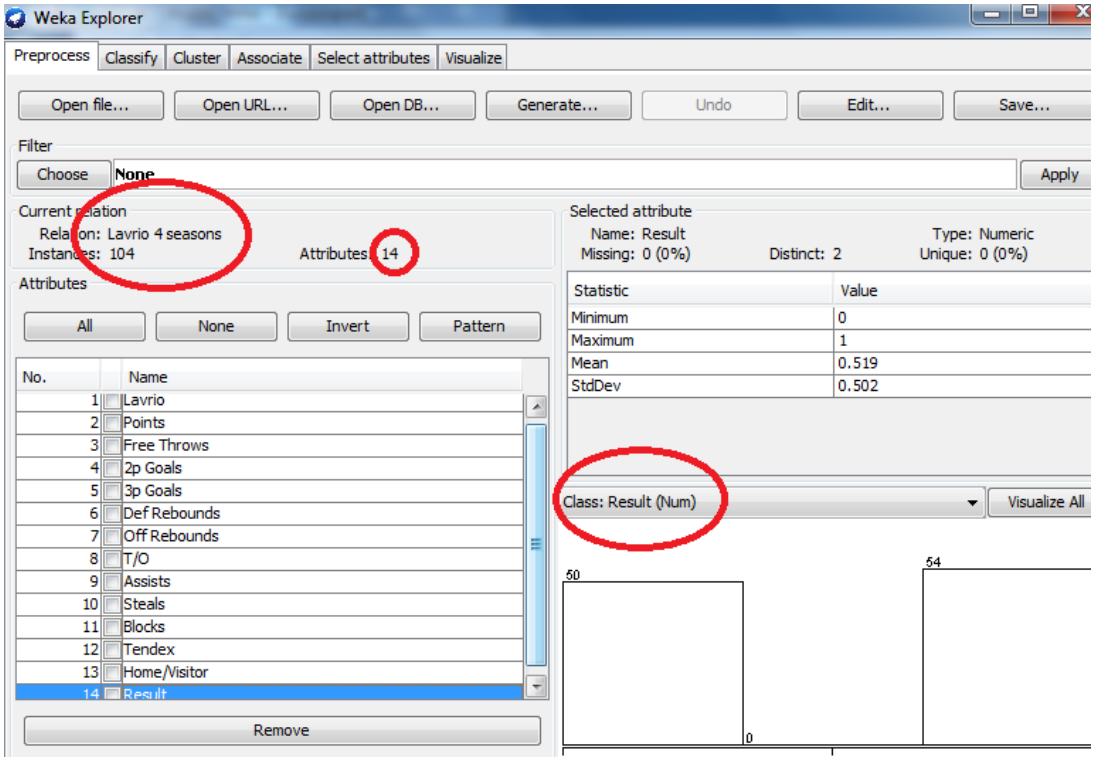
Από τα αποτελέσματα τα οποία προέκυψαν, τα οποία παρουσιάζονται συγκεντρωτικά στον παραπάνω πίνακα, διαπιστώνουμε ότι

1. ο αλγόριθμος με τη χειρότερη απόδοση είναι ο αλγόριθμος ID-3, του οποίου τα αποτελέσματα δεν λαμβάνουμε καθόλου υπόψη, δεδομένου ότι δεν είναι αξιόπιστα (50 % των υποδειγμάτων δεν ταξινομήθηκαν καθόλου)

2. ο αλγόριθμος με την καλύτερη απόδοση και μέγιστη προβλεπτική ικανότητα είναι ο αλγόριθμος CART (σε σύγκριση με τον αλγόριθμο J48), για τον οποίο παρατηρείται το μεγαλύτερο ποσοστό ακρίβειας, μεγαλύτερο ποσοστό σωστά ταξινομημένων υποδειγμάτων, μηδενικό ποσοστό αταξινόμητων υποδειγμάτων, το μεγαλύτερο ποσοστό των θετικών υποδειγμάτων που έχουν αναγνωριστεί σωστά ως θετικά (μεγαλύτερη ευαισθησία) και το μικρότερο ποσοστό των αρνητικών υποδειγμάτων που έχουν αναγνωριστεί λανθασμένα ως θετικά (μεγαλύτερη ειδικότητα), υψηλότερες τιμές για όλα τα υπό εξέταση μέτρα ακρίβειας (precision, recall, f-measure), υψηλότερη τιμή εμβαδού κάτω από την ROC καμπύλη.

14.2.2 Δέντρα απόφασης για την ομάδα του Λαυρίου

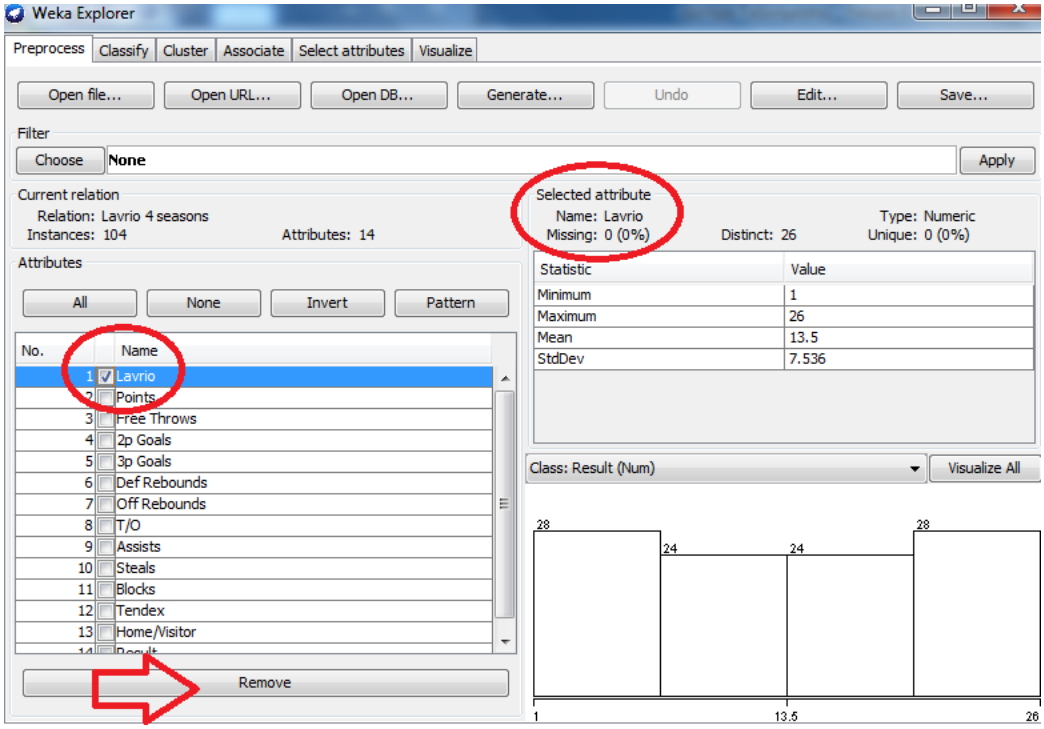
Ξεκινώντας το Weka, ακολουθούμε τα βήματα «Explorer→Preprocess→Open File» και «φορτώνοντας» το αρχείο "Lavrio 4 seasons.csv", στην οθόνη θα εμφανιστεί το παράθυρο το οποίο απεικονίζεται παρακάτω. Το αρχείο "Lavrio 4 seasons.csv" αφορά στην ομάδα του Λαυρίου και περιλαμβάνει τις υπό εξέταση μεταβλητές για τέσσερις συνολικά χρονικές περιόδους 2011-2012, 2012-2013, 2013-2014 και 2014-2015.



Εικόνα 43 : Το πάνελ Preprocess- Φόρτωση αρχείου "Lavrio 4 seasons.csv

Στο παράθυρο αυτό, στο μέσο και αριστερά, αναγράφεται ότι το συγκεκριμένο σύνολο δεδομένων "Lavrio 4 seasons.csv περιλαμβάνει 104 υποδείγματα (instances) και κάθε υπόδειγμα απαρτίζεται από 14 χαρακτηριστικά – μεταβλητές. Στο συγκεκριμένο παράδειγμα το χαρακτηριστικό «result» έχει επιλεγεί ως αυτό το οποίο δείχνει σε ποια κλάση ανήκει κάθε φορά το υπόδειγμα. Πρόκειται για το τελευταίο χαρακτηριστικό που καταχωρούμε, και στην περίπτωση μας που αντιμετωπίζουμε ένα πρόβλημα κατηγοριοποίησης δείχνει την κατηγορία στην οποία ανήκει το υπόδειγμα,. Το χαρακτηριστικό «result» αποτελεί μία δίτιμη μεταβλητή **με την τιμή 0 να δηλώνει την «ήττα», και την τιμή 1 τη «νίκη».**

Πριν προχωρήσουμε στη διαδικασία διακριτοποίησης των δεδομένων μας, αρχικά αφαιρούμε τη μεταβλητή «Lavrio» (τικάρουμε και Remove), η οποία έρχεται πρώτη στη λίστα των διαθέσιμων χαρακτηριστικών μας, και αποτελεί τη μεταβλητή- Index η οποία απαριθμεί τα παιχνίδια ($1^{ο}$ παιχνίδι, $2^{ο}$ παιχνίδι κ.λπ.) των τεσσάρων χρονικών περιόδων.

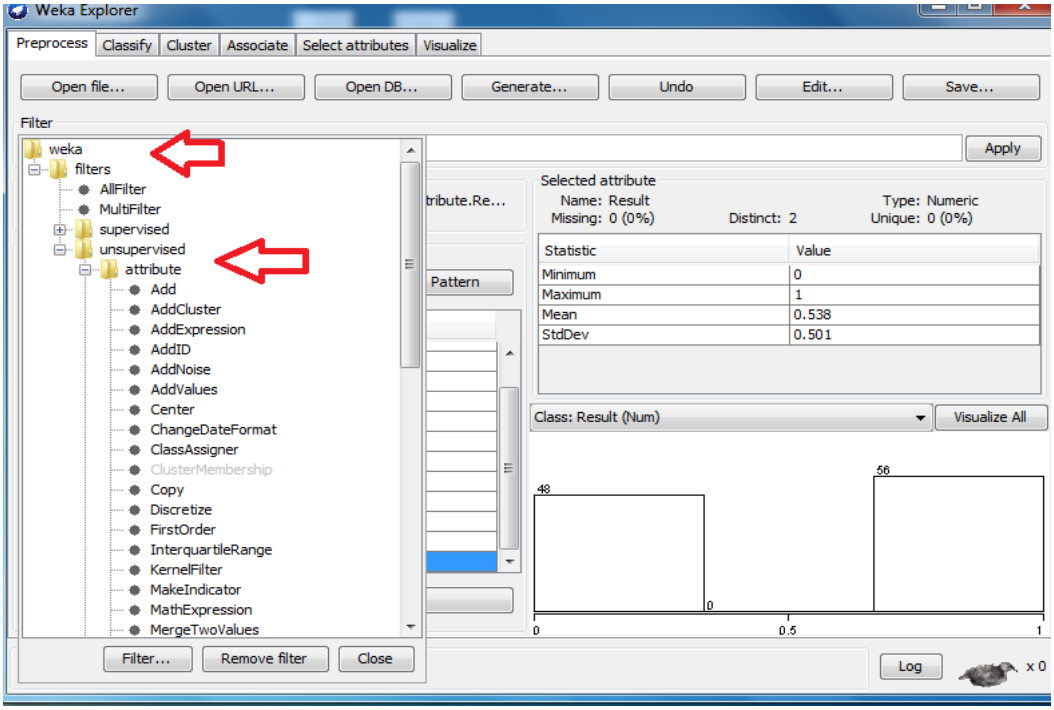


Εικόνα 44: Αφαίρεση μεταβλητής "Lavrio" Index

Το επόμενο και σημαντικό βήμα αποτελεί η διακριτοποίηση των δεδομένων μας. Δεδομένου ότι οι μεταβλητές του συνόλου δεδομένων μας είναι αριθμητικές ή συνεχείς (εκτός της μεταβλητής "home / vistitor" και της μεταβλητής class "result", οι οποίες είναι δίτιμες 0/1) συνιστάται συνήθως μετά από εισαγωγή csv αρχείου να μετατρέπουμε αυτές τις αριθμητικές μεταβλητές αντίστοιχα σε ονομαστικές.

Ακολουθούμε τα παρακάτω βήματα:

«Choose→weka→filters→unsupervised→attribute»

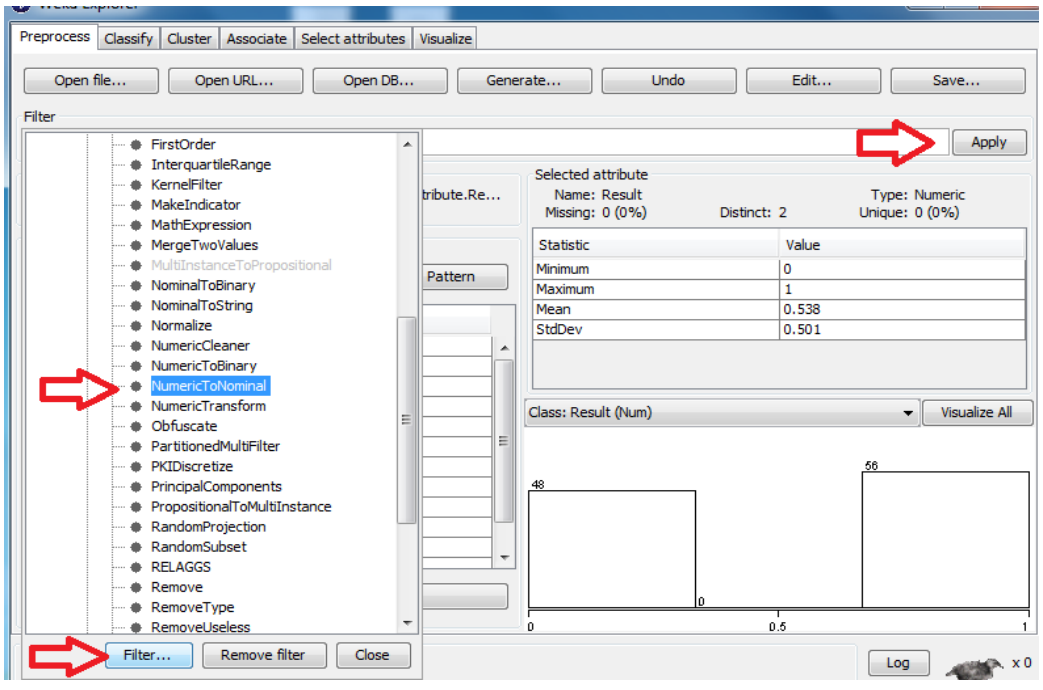


Εικόνα 45: Επιλογή φίλτρου

Έπειτα, επιλέγουμε από τη λίστα διαθέσιμων επιλογών ” το “Numeric to Nominal” κλικάρουμε “Filter→Filtering Capabilities”

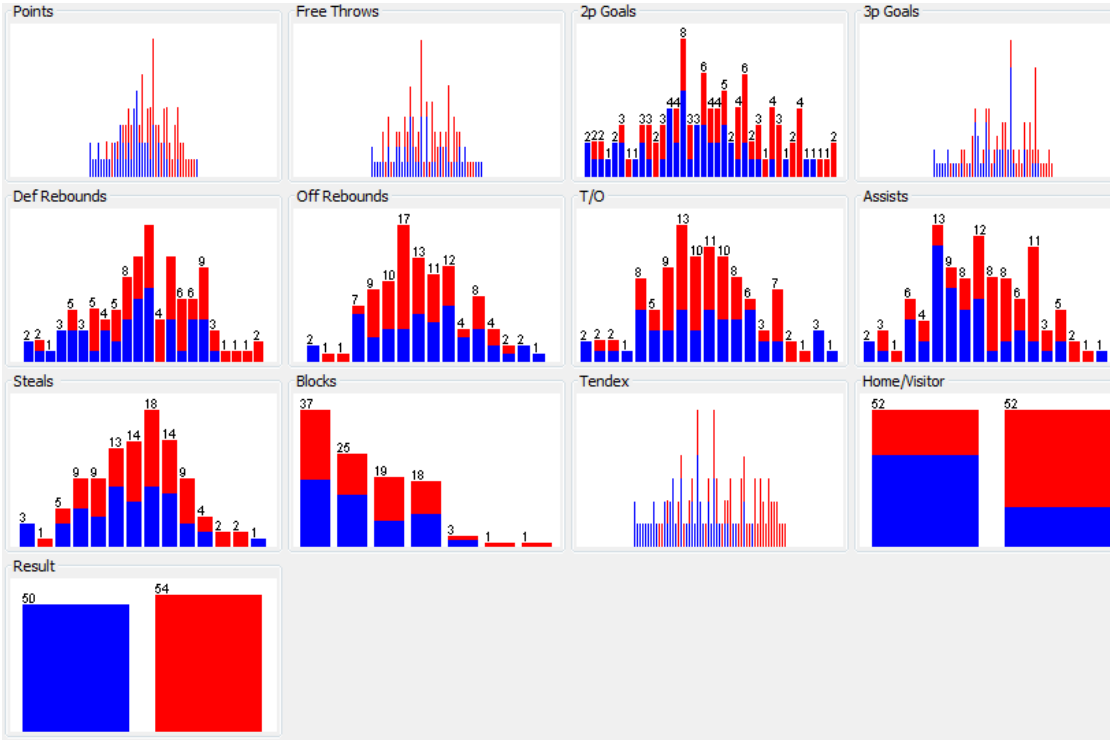
- ✓ Numeric Attributes
- ✓ Numeric class

Ok και →Apply.



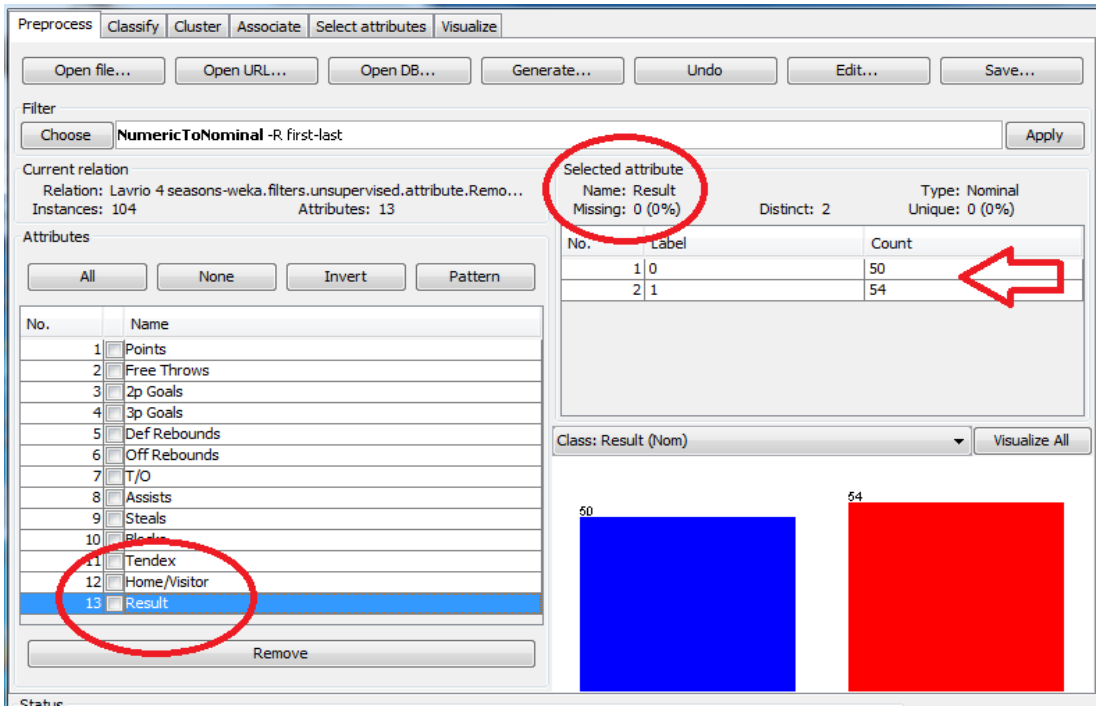
Εικόνα 46: Διαδικασία διακριτοποίησης

Έπειτα, επιλέγοντας το "Visualize all" λαμβάνουμε σε νέο παράθυρο την οπτικοποίηση και την απεικόνιση της κατανομής των 11 συνολικά μεταβλητών που μετασχηματίστηκαν. Σημειώνεται ότι στο παράθυρο οπτικοποίησης συμπεριλαμβάνονται και οι μη μετασχηματισμένες δίτιμες εξ'αρχής μεταβλητές, η μεταβλητή home/visitor και η μεταβλητή-στόχος "result".



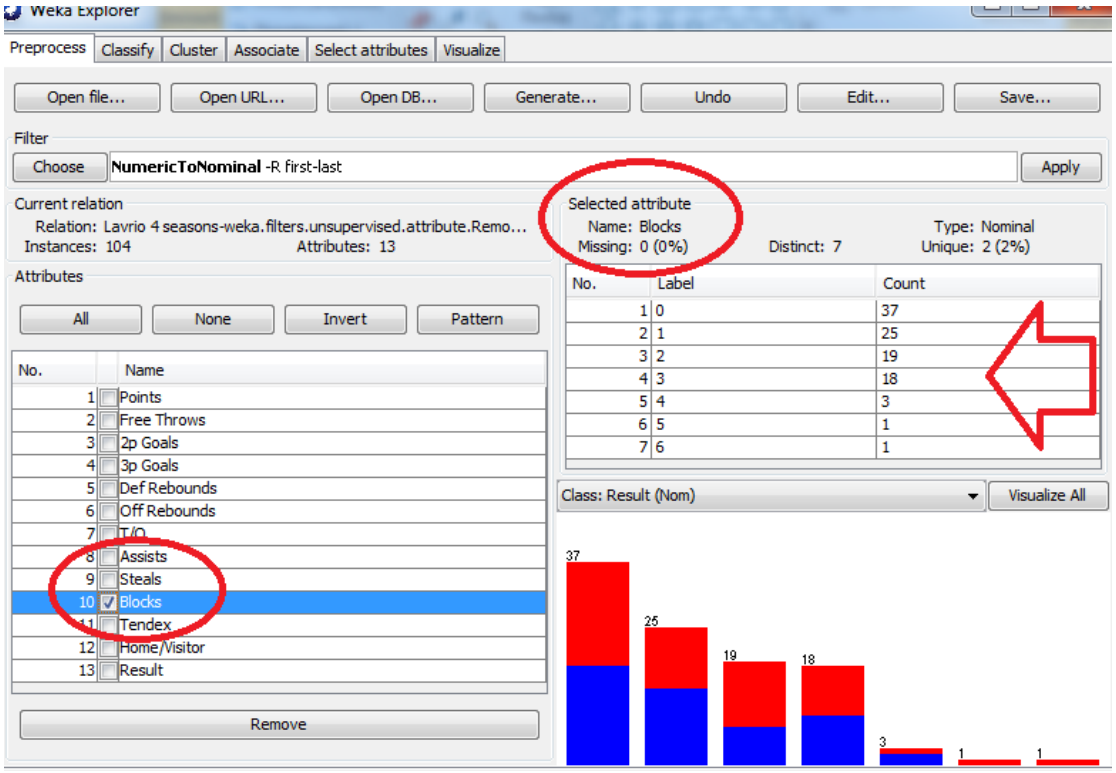
Εικόνα 47 : Οπτικοποίηση των μεταβλητών

Τα παραπάνω ιστογράμματα μας βοηθούν στην κατανόηση των δεδομένων, δεδομένου ότι υποδεικνύουν την κατανομή των τιμών των ονομαστικών χαρακτηριστικών. Για παράδειγμα, για τη μεταβλητή στόχο "Result" παρατηρούμε ότι έχουμε 50 «ήττες» και 54 «νίκες».



Εικόνα 48: Η κατανομή των τιμών της μεταβλητής "Result"

Για παράδειγμα, μεμονωμένα για τη μεταβλητή "Blocks" παρατηρούμε ότι έχουμε 0 Blocks σε 37 παιχνίδια, 1 Block σε 25 παιχνίδια , 2 Blocks σε 19 παιχνίδια, 3 Blocks σε 18 παιχνίδια, 4 Blocks σε 3 παιχνίδια, 5 Blocks σε 1 παιχνίδι, 6 Blocks σε 1 παιχνίδι, όπως άλλωστε φαίνεται στην παρακάτω Εικόνα 49.



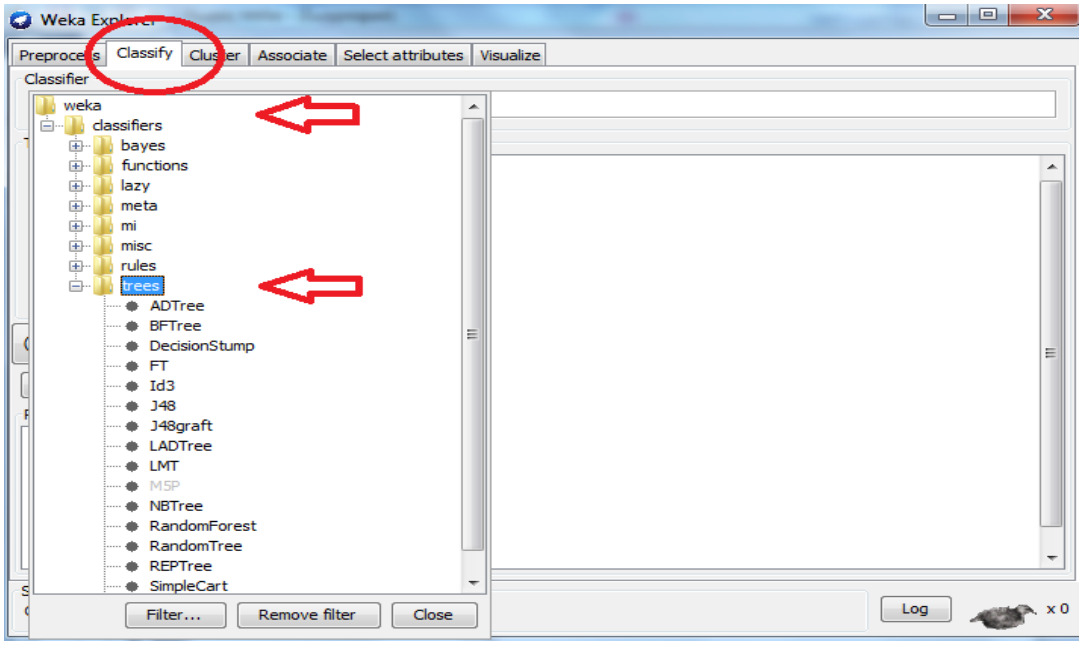
Εικόνα 49: Η κατανομή των τιμών της μεταβλητής "Blocks"

Επιλέγουμε στο πάνελ την καρτέλα «Classify» όπου περιλαμβάνονται ποικίλες τεχνικές ταξινόμησης με διαφορετικές δυνατότητες για το χρήστη η κάθε μία.

Αρχικά θα εφαρμόσουμε τεχνικές στις οποίες η κατηγοριοποίηση των δεδομένων υλοποιείται μέσω της κατασκευής των δέντρων ταξινόμησης.

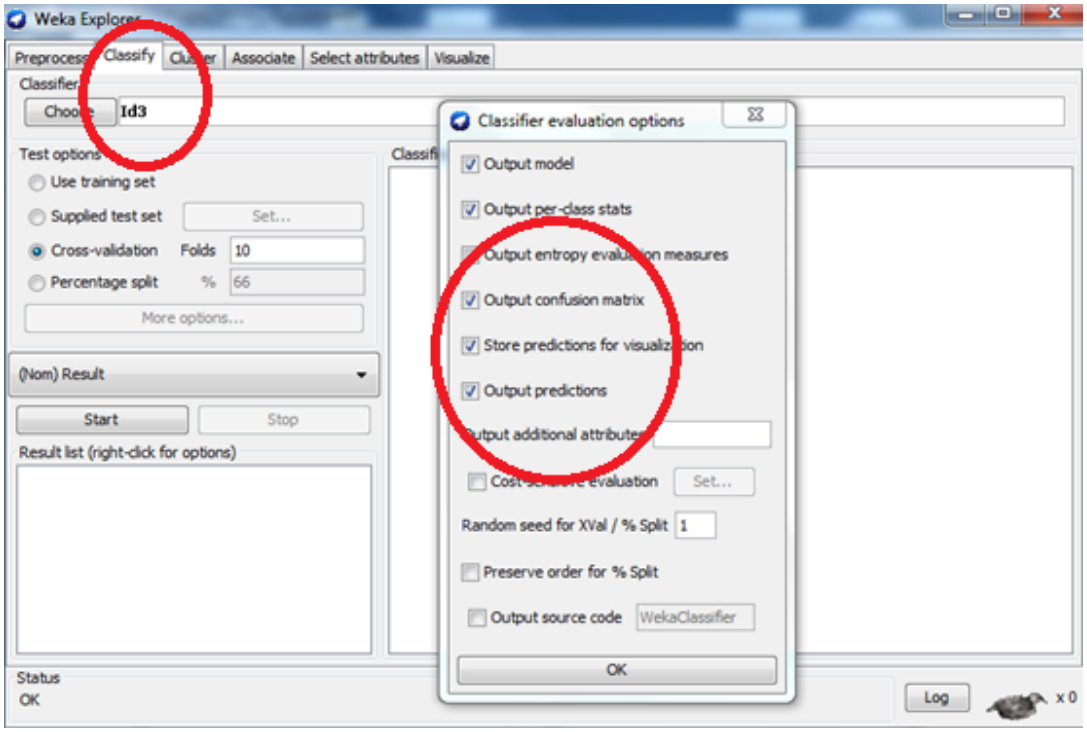
Ακολουθούμε την παρακάτω διαδικασία»:

«Choose→Weka→Classifiers→Trees»



- **Αλγόριθμος ID 3**

Μέσα από τη λίστα των διαθέσιμων τεχνικών των δέντρων ταξινόμησης «trees» επιλέγουμε τον αλγόριθμο ID-3, έπειτα επιλέγουμε στο "Test options" το "Cross-Validation" αφήνοντας την επιλογή «10», έτσι ώστε να υλοποιηθεί η διαδικασία της διασταυρωμένης επικύρωσης με 10 πεδία για την τελική αξιολόγηση της απόδοσης της μεθόδου, και κλικάρουμε το tab "More Options" έτσι ώστε να ανοίξει ένα νέο παράθυρο διαλόγου, το "Classifier evaluation options", το οποίο μας παρέχει τη δυνατότητα να επιλέξουμε τα επιθυμητά εξαγόμενα αποτελέσματα τα οποία θα εμφανιστούν μετά την εκτέλεση του αλγορίθμου.



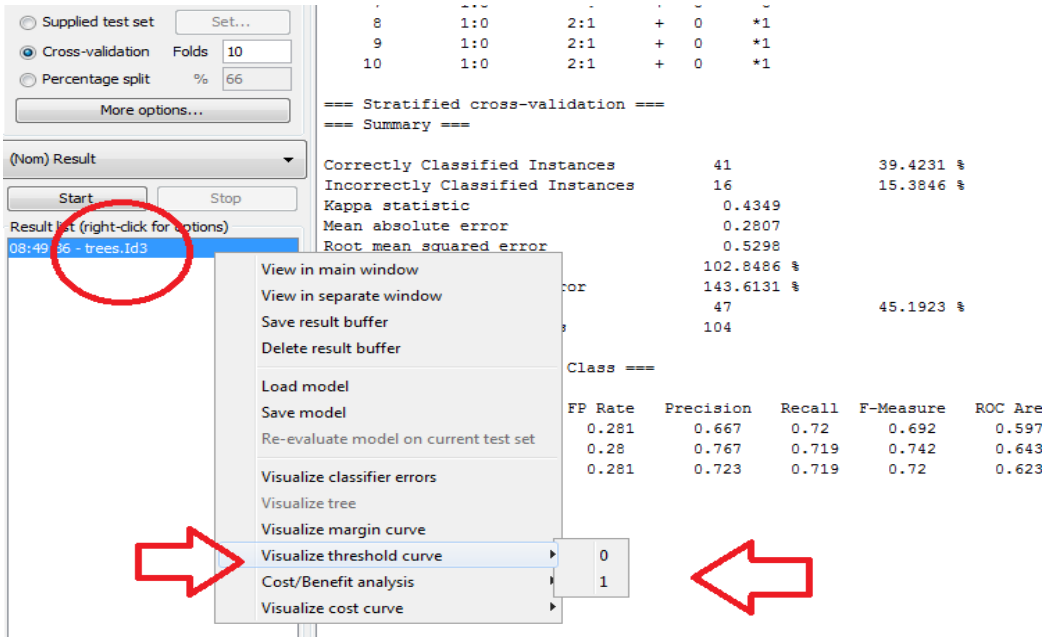
Εικόνα 50: Υλοποίηση αλγορίθμου ID-3

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          41               39.4231 %
Incorrectly Classified Instances        16               15.3846 %
Kappa statistic                          0.4349
Mean absolute error                      0.2807
Root mean squared error                  0.5298
Relative absolute error                102.8486 %
Root relative squared error            143.6131 %
UnClassified Instances                  47               45.1923 %
Total Number of Instances              104

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.72      0.281       0.667     0.72      0.692      0.597    0
                   0.719     0.28        0.767     0.719     0.742      0.643    1
Weighted Avg.      0.719     0.281       0.723     0.719     0.72       0.623

=== Confusion Matrix ===

  a  b   <-- classified as
 18  7 |  a = 0
  9 23 |  b = 1
```

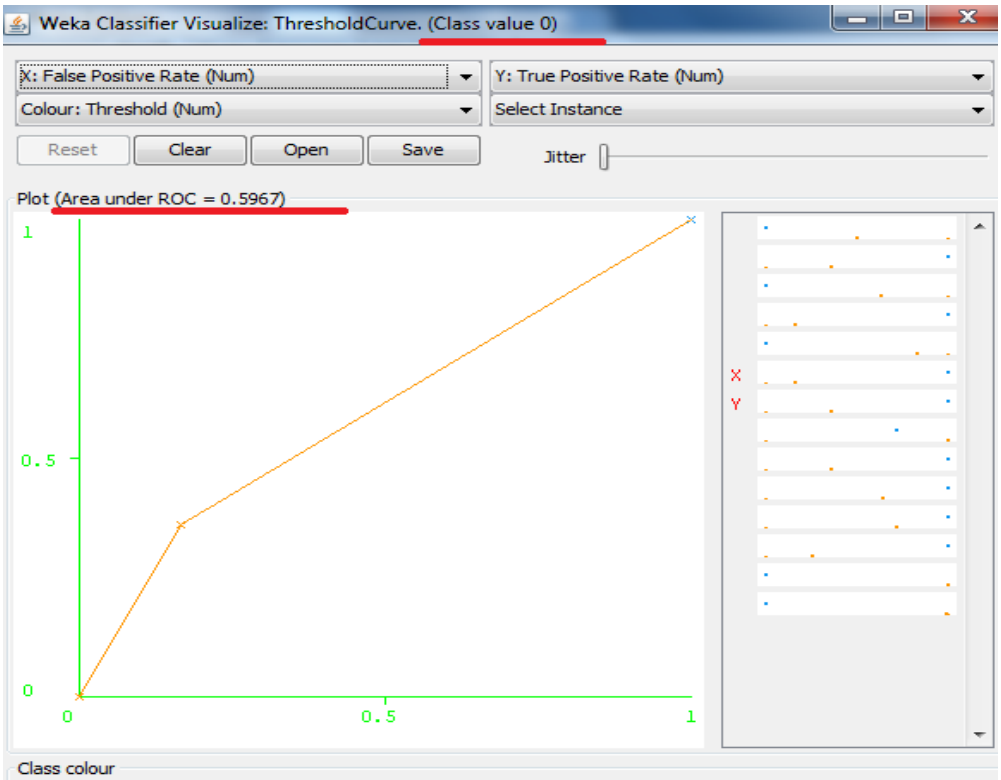
Εικόνα 51: Εξαγόμενα αποτελέσματα ID-3 για το σύνολο δεδομένων "Lavrio 4 Seasons.csv

Έπειτα, για την γραφική αναπαράσταση της καμπύλης λειτουργικού χαρακτηριστικού δέκτη (ROC) και τον υπολογισμό του εμβαδού κάτω από την καμπύλη ROC (Area Under the Curve-AUC) , κάνουμε δεξί κλικ στο "trees.Id3" και επιλέγουμε στο νέο

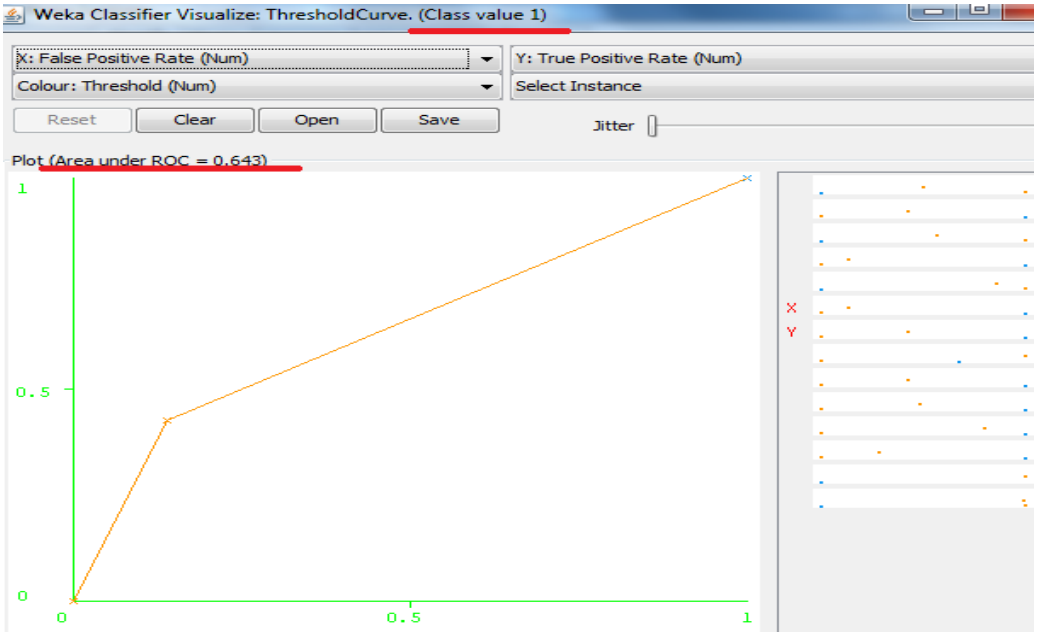
παράθυρο διαλόγου που ανοίγει το “Visualize threshold curve” και για τις δύο κλάσεις 0 και 1, αντίστοιχα.



Εικόνα 52: Αναπαράσταση της ROC καμπύλης και υπολογισμός του εμβαδού κάτω από την καμπύλη AUC



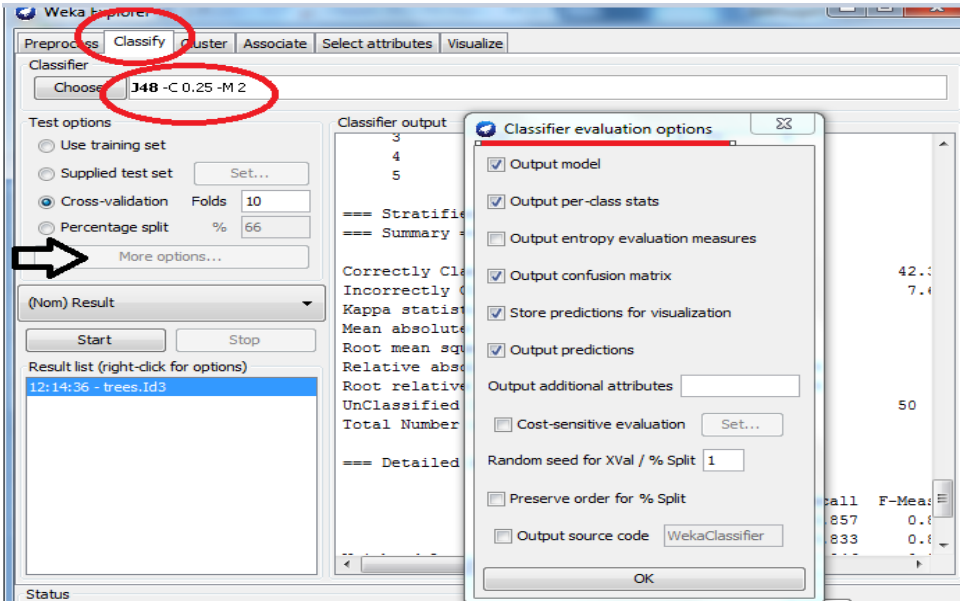
Εικόνα 53: ROC καμπύλη για την κλάση 0 – AUC=0.5967



Εικόνα 54: ROC καμπύλη για την κλάση 1 – AUC=0.643

- **Αλγόριθμος J48**

Μέσα από τη λίστα των διαθέσιμων τεχνικών των δέντρων ταξινόμησης «trees» επιλέγουμε τον αλγόριθμο J48, έπειτα επιλέγουμε στο "Test options" το "Cross-Validation" αφήνοντας την επιλογή «10», έτσι ώστε να υλοποιηθεί η διαδικασία της διασταυρωμένης επικύρωσης με 10 πεδία για την τελική αξιολόγηση της απόδοση της μεθόδου, και κλικάρουμε το tab "More Options" έτσι ώστε να ανοίξει ένα νέο παράθυρο διαλόγου, το "Classifier evaluation options", το οποίο μας παρέχει τη δυνατότητα να επιλέξουμε τα επιθυμητά εξαγόμενα αποτελέσματα τα οποία θα εμφανιστούν μετά την εκτέλεση του αλγορίθμου.

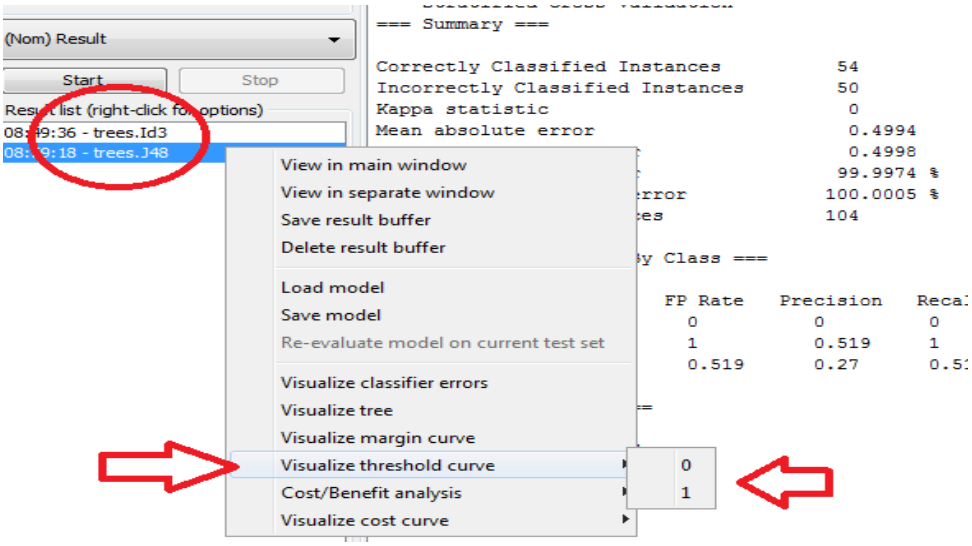


Εικόνα 55: Υλοποίηση αλγορίθμου J48

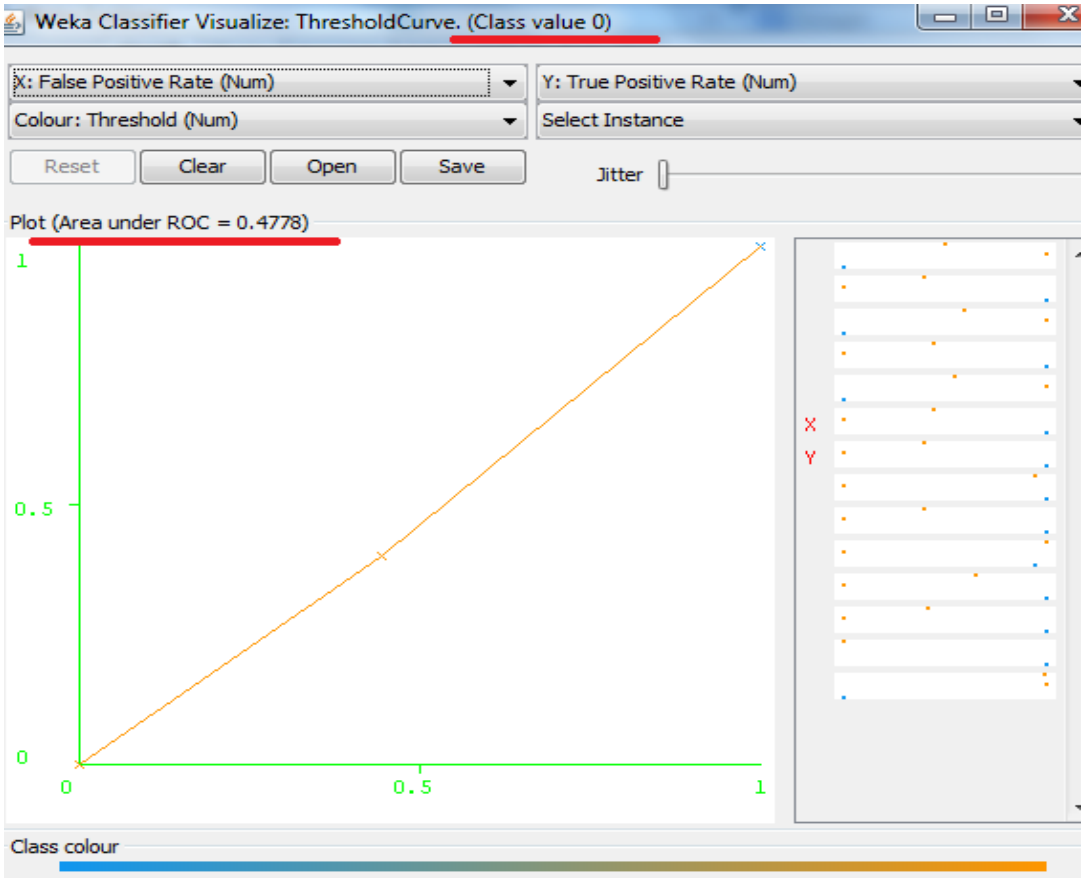
```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          54               51.9231 %
Incorrectly Classified Instances        50               48.0769 %
Kappa statistic                          0
Mean absolute error                      0.4994
Root mean squared error                  0.4998
Relative absolute error                 99.9974 %
Root relative squared error            100.0005 %
Total Number of Instances              104

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.478    0
                 1         1          0.519     1         0.684      0.478    1
Weighted Avg.    0.519     0.519      0.27      0.519     0.355      0.478

=== Confusion Matrix ===

  a  b   <-- classified as
  0 50 |  a = 0
  0 54 |  b = 1
```

Εικόνα 56: Εξαγόμενα αποτελέσματα J48 για το σύνολο δεδομένων "Lavrio 4 Seasons.csv

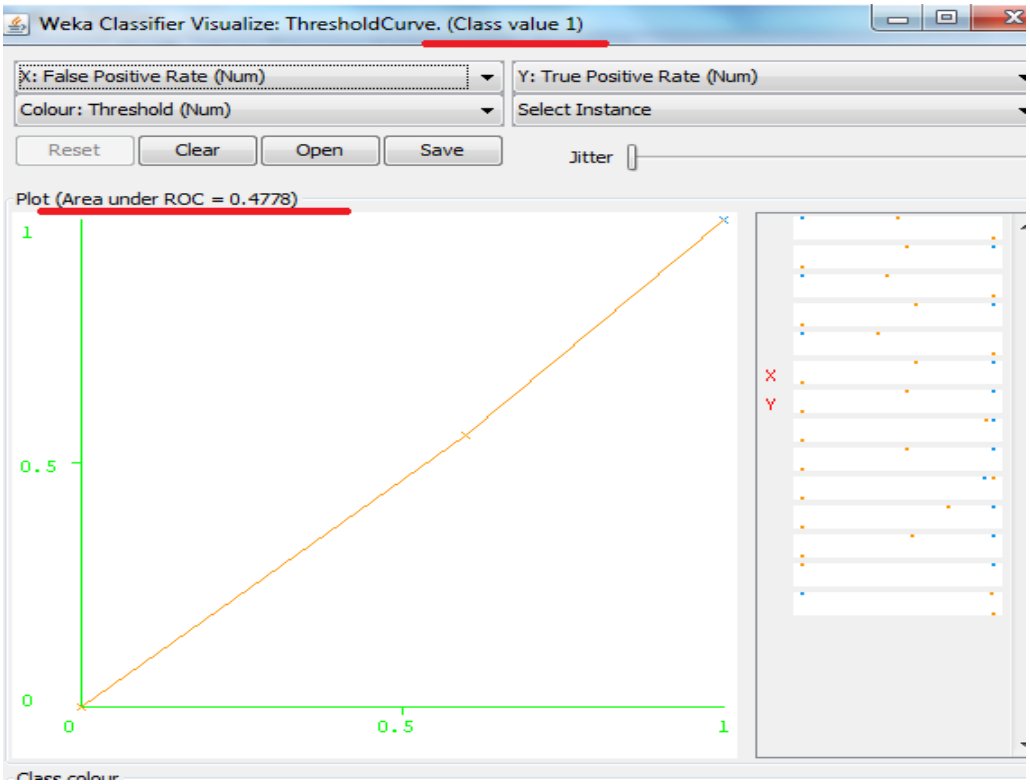
Έπειτα, για την γραφική αναπαράσταση της καμπύλης λειτουργικού χαρακτηριστικού δέκτη (ROC) και τον υπολογισμό του εμβαδού κάτω από την καμπύλη ROC (Area Under the Curve-AUC) , κάνουμε δεξί κλικ στο "trees.J48" και επιλέγουμε στο νέο παράθυρο διαλόγου που ανοίγει το "Visualize threshold curve" και για τις κλάσεις 0 και 1, αντίστοιχα.



Εικόνα 57: Αναπαράσταση της ROC καμπύλης και υπολογισμός του εμβαδού κάτω από την καμπύλη AUC



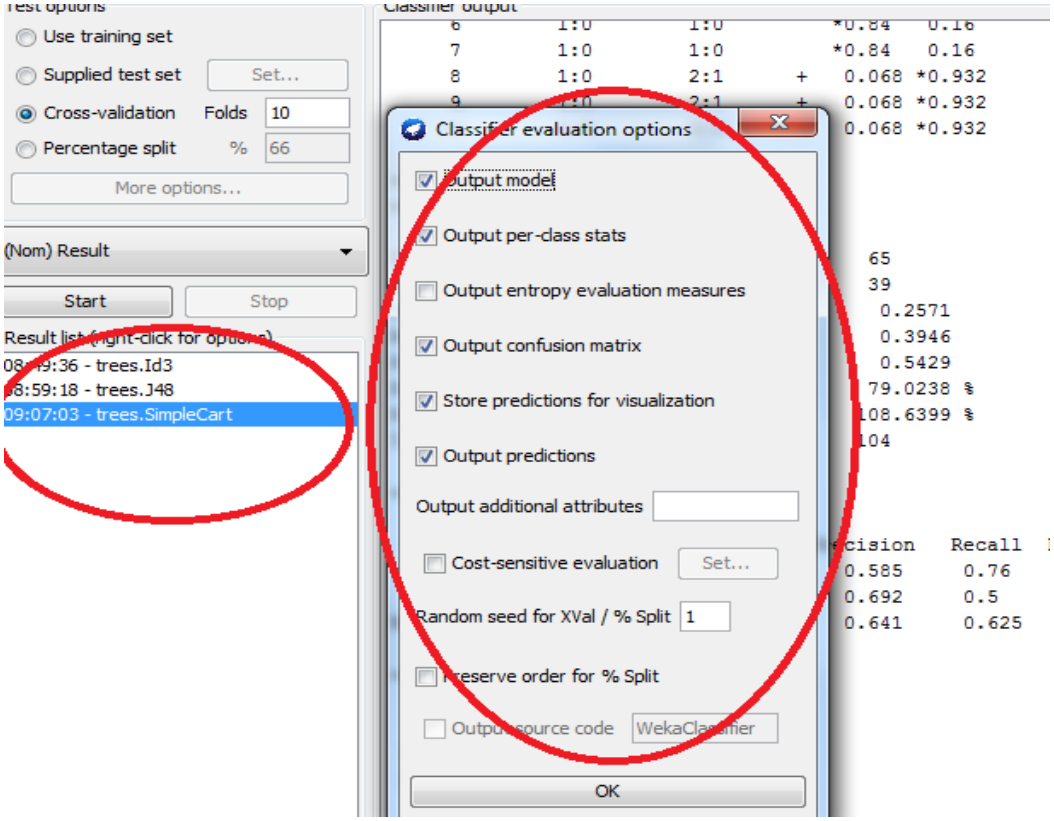
Εικόνα 58: ROC καμπύλη για την κλάση 0 – AUC=0.4778



Εικόνα 59: ROC καμπύλη για την κλάση 1 – AUC=0.4778

Αλγόριθμος CART

Μέσα από τη λίστα των διαθέσιμων τεχνικών των δέντρων ταξινόμησης «trees» επιλέγουμε τον αλγόριθμο SimpleCart, έπειτα επιλέγουμε στο "Test options" το "Cross- Validation" αφήνοντας την επιλογή «10», έτσι ώστε να υλοποιηθεί η διαδικασία της διασταυρωμένης επικύρωσης με 10 πεδία για την τελική αξιολόγηση της απόδοση της μεθόδου, και κλικάρουμε το tab "More Options" έτσι ώστε να ανοίξει ένα νέο παράθυρο διαλόγου, το "Classifier evaluation options", το οποίο μας παρέχει τη δυνατότητα να επιλέξουμε τα επιθυμητά εξαγόμενα αποτελέσματα τα οποία θα εμφανιστούν μετά την εκτέλεση του αλγορίθμου.

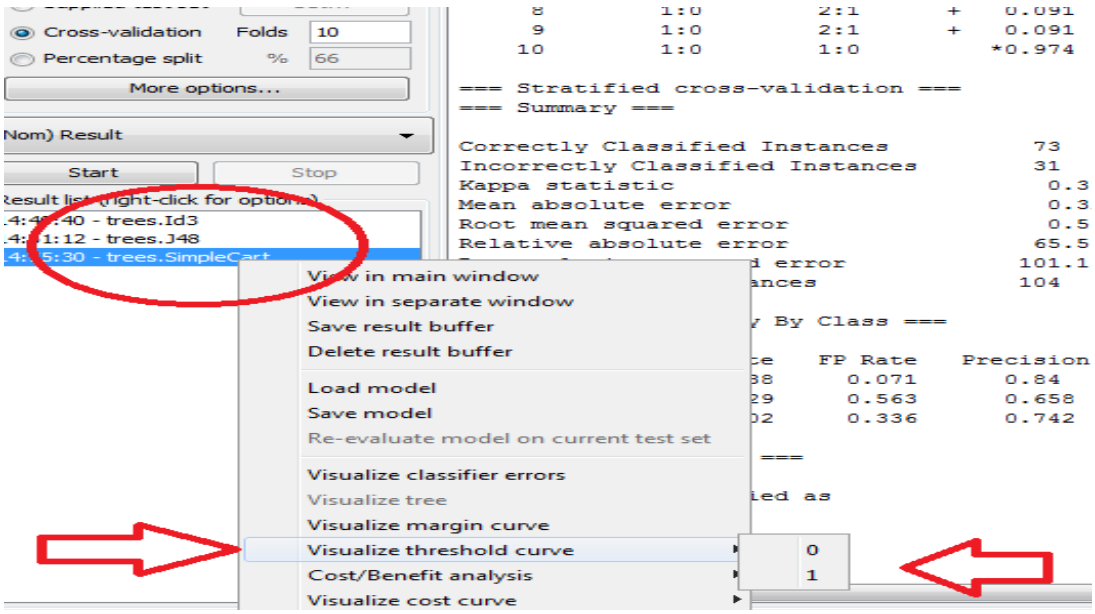


Εικόνα 60: Υλοποίηση αλγορίθμου SimpleCart

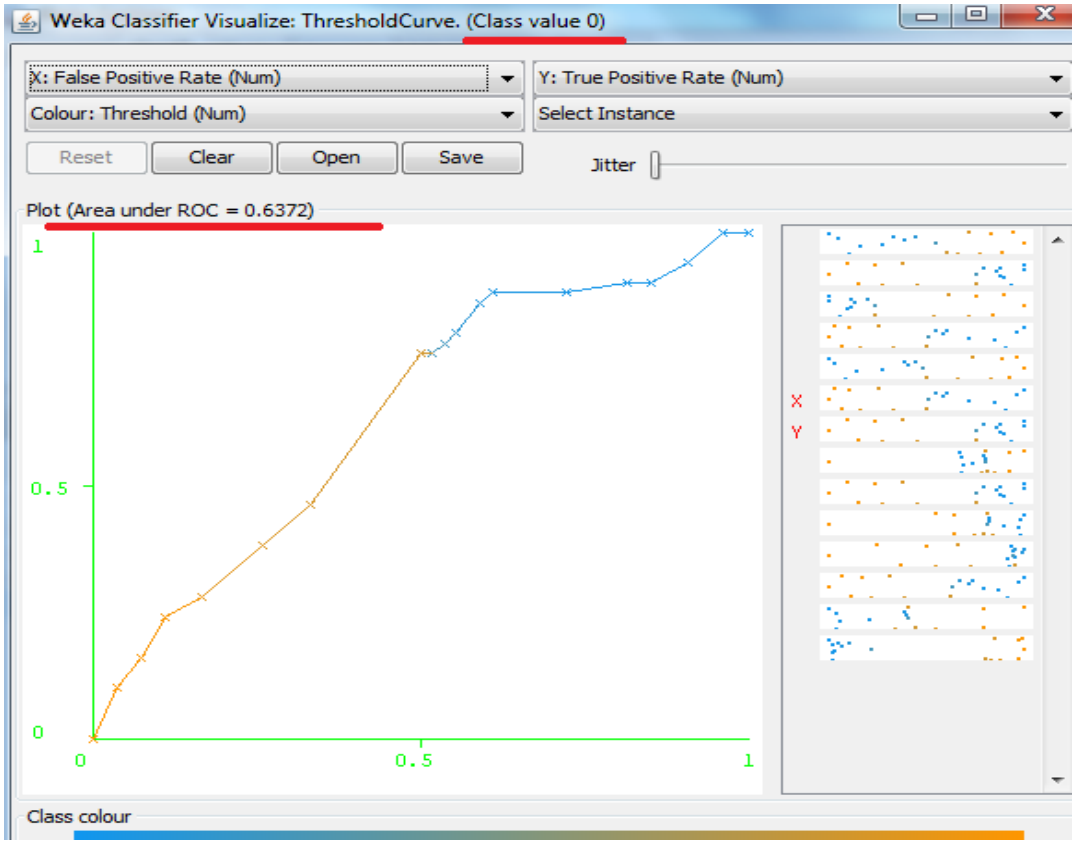
```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          65               62.5    %
Incorrectly Classified Instances        39               37.5    %
Kappa statistic                          0.2571
Mean absolute error                      0.3946
Root mean squared error                  0.5429
Relative absolute error                 79.0238 %
Root relative squared error            108.6399 %
Total Number of Instances              104

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.76      0.5        0.585     0.76      0.661      0.637    0
                 0.5       0.24       0.692     0.5       0.581      0.637    1
Weighted Avg.    0.625     0.365      0.641     0.625     0.619      0.637

=== Confusion Matrix ===

  a  b   <-- classified as
 38 12 |  a = 0
 27 27 |  b = 1
```

Εικόνα 61: Εξαγόμενα αποτελέσματα SimpleCart για το σύνολο δεδομένων "Lavrio 4 Seasons.csv"

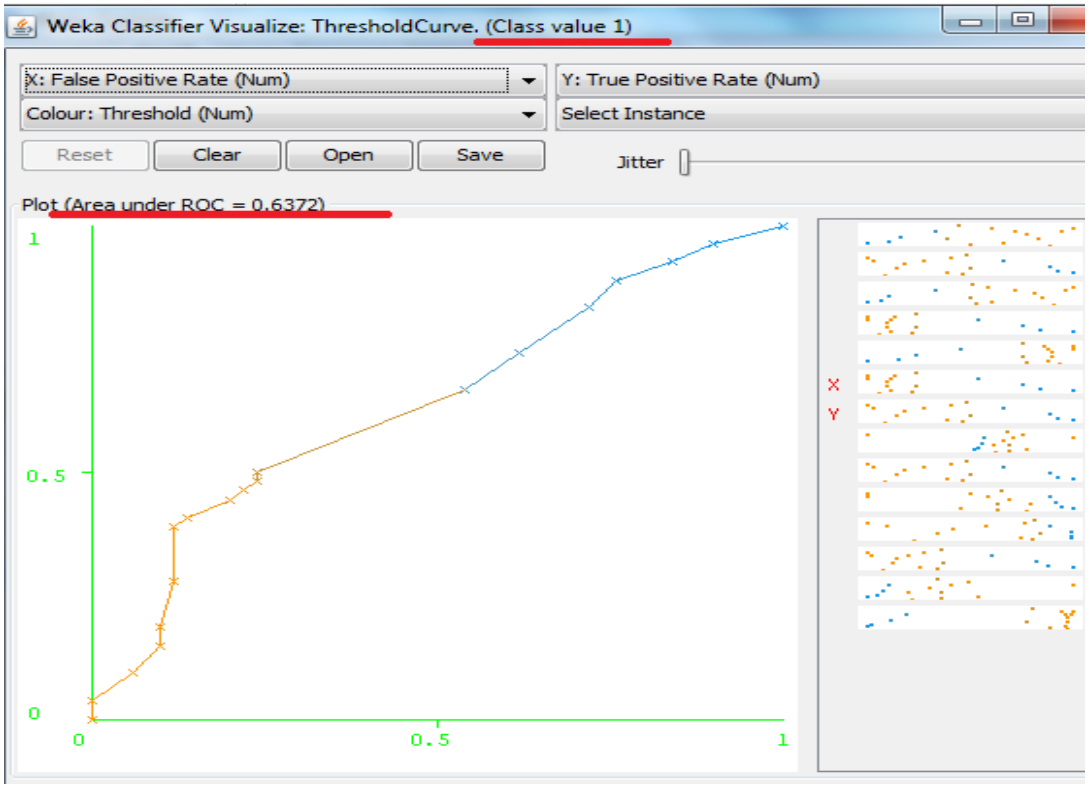
Έπειτα, για την γραφική αναπαράσταση της καμπύλης λειτουργικού χαρακτηριστικού δέκτη (ROC) και τον υπολογισμό του εμβαδού κάτω από την καμπύλη ROC (Area Under the Curve-AUC) , κάνουμε δεξί κλικ στο "trees.SimpleCart" και επιλέγουμε στο νέο παράθυρο διαλόγου που ανοίγει το "Visualize threshold curve" και για τις κλάσεις 0 και 1, αντίστοιχα.



Εικόνα 62: Αναπαράσταση της ROC καμπύλης και υπολογισμός του εμβαδού κάτω από την καμπύλη AUC



Εικόνα 63: ROC καμπύλη για την κλάση 0 – AUC=0.6372



Εικόνα 64: ROC καμπύλη για την κλάση 1 – AUC=0.6372

Πίνακας 4: Συγκριτικά αποτελέσματα δέντρων ταξινόμησης για το σύνολο δεδομένων "Lavrio 4 Seasons.csv"

	Correctly Classified Instances - Accuracy	Incorrectly Classified Instances	Unclassified Instances	Avg. TP Rate	Avg. FP Rate	Avg. Precision	Avg. Recall	Avg. F-Measure	Avg. ROC Area
ID-3	39.42 %	15.38 %	45.20 %	0.719	0.281	0.723	0.719	0.72	0.623
J48	51.92 %	48.08 %	0 %	0.519	0.519	0.27	0.519	0.355	0.478
Simple CART	62.50 %	37.50 %	0 %	0.625	0.365	0.641	0.625	0.619	0.637

Από τα αποτελέσματα τα οποία προέκυψαν, τα οποία παρουσιάζονται συγκεντρωτικά στον παραπάνω πίνακα, διαπιστώνουμε ότι

1. ο αλγόριθμος με τη χειρότερη απόδοση είναι ο αλγόριθμος ID-3, του οποίου τα αποτελέσματα δεν λαμβάνουμε καθόλου υπόψη, δεδομένου ότι δεν είναι αξιόπιστα (45.20 % των υποδειγμάτων δεν ταξινομήθηκαν καθόλου)

2. ο αλγόριθμος με την καλύτερη απόδοση και μέγιστη προβλεπτική ικανότητα είναι ο αλγόριθμος CART (σε σύγκριση με τον αλγόριθμο J48), για τον οποίο παρατηρείται το μεγαλύτερο ποσοστό ακρίβειας, μεγαλύτερο ποσοστό σωστά ταξινομημένων υποδειγμάτων, μηδενικό ποσοστό αταξινόμητων υποδειγμάτων, το μεγαλύτερο ποσοστό των θετικών υποδειγμάτων που έχουν αναγνωριστεί σωστά ως θετικά (μεγαλύτερη ευαισθησία) και το μικρότερο ποσοστό των αρνητικών υποδειγμάτων που έχουν αναγνωριστεί λανθασμένα ως θετικά (μεγαλύτερη ειδικότητα), υψηλότερες τιμές για όλα τα υπό εξέταση μέτρα ακρίβειας (precision, recall, f-measure), υψηλότερη τιμή εμβαδού κάτω από την ROC καμπύλη.

14.3 Κανόνες ταξινόμησης

14.3.1 Κανόνες ταξινόμησης για την ομάδα του Αρκαδικού

Αρχικά ακολουθούμε τα παρακάτω βήματα (τα οποία ακολουθήσαμε και πριν την υλοποίηση των αλγορίθμων δέντρων απόφασης, ID-3, J48 και CART).

1. Ακολουθούμε τα βήματα

Explorer→Preprocess→Open File

και «φορτώνοντας» το αρχείο "Arkadikos 4 seasons.csv", το οποίο αφορά στην ομάδα του Αρκαδικού και περιλαμβάνει τις υπό εξέταση μεταβλητές για τέσσερις συνολικά χρονικές περιόδους 2011-2012, 2012-2013, 2013-2014 και 2014-2015.

2. Αφαιρούμε τη μεταβλητή «Arkadikos» (τικάρουμε και Remove), η οποία αποτελεί τη μεταβλητή- Index η οποία απαριθμεί τα παιχνίδια ($1^{ο}$ παιχνίδι, $2^{ο}$ παιχνίδι κ.λπ.) των τεσσάρων χρονικών περιόδων, και επιλέγουμε το χαρακτηριστικό «result» ως αυτό το οποίο δείχνει σε ποια κλάση ανήκει κάθε φορά το υπόδειγμα. Το χαρακτηριστικό «result» αποτελεί μία δίτιμη μεταβλητή με τη τιμή **0 να δηλώνει την «ήττα» και την τιμή 1 τη «νίκη».**

3. Προχωρούμε στη διαδικασία διακριτοποίησης των δεδομένων μας, ακολουθώντας τα παρακάτω βήματα:

«Choose→weka→filters→unsupervised→attribute»

Έπειτα, επιλέγουμε από τη λίστα διαθέσιμων επιλογών " το "Numeric to Nominal" κλικάρουμε "Filter→Filtering Capabilities"

- ✓ Numeric Attributes
- ✓ Numeric class

Ok

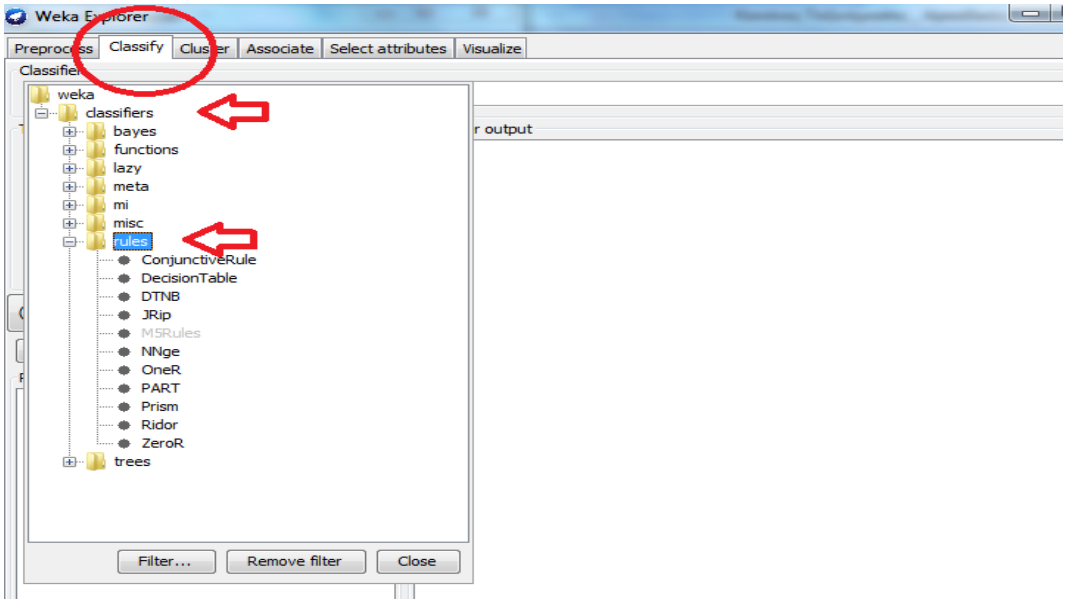
και →Apply.

Επιλέγουμε στο πάνελ την καρτέλα «Classify» όπου περιλαμβάνονται ποικίλες τεχνικές ταξινόμησης με διαφορετικές δυνατότητες για το χρήστη η κάθε μία.

Θα εφαρμόσουμε τεχνικές στις οποίες η κατηγοριοποίηση των δεδομένων υλοποιείται μέσω της εύρεσης κανόνων απόφασης/ταξινόμησης.

Ακολουθούμε την παρακάτω διαδικασία»:

«Choose→Weka→Classifiers→Rules»



Αλγόριθμος Conjuctive Rule

Μέσα από τη λίστα των διαθέσιμων τεχνικών των κανόνων ταξινόμησης «Rules» επιλέγουμε τον αλγόριθμο ConjuctiveRule, έπειτα επιλέγουμε στο "Test options" το "Cross- Validation" αφήνοντας την επιλογή «10», έτσι ώστε να υλοποιηθεί η διαδικασία της διασταυρωμένης επικύρωσης με 10 πεδία για την τελική αξιολόγηση της απόδοσης της μεθόδου, και κλικάρουμε το tab "More Options" έτσι ώστε να ανοίξει ένα νέο παράθυρο διαλόγου, το "Classifier evaluation options", το οποίο μας παρέχει τη δυνατότητα να επιλέξουμε τα επιθυμητά εξαγόμενα αποτελέσματα τα οποία θα εμφανιστούν μετά την εκτέλεση του αλγορίθμου.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          73               70.1923 %
Incorrectly Classified Instances        31               29.8077 %
Kappa statistic                          0.3994
Mean absolute error                      0.4143
Root mean squared error                  0.4543
Relative absolute error                 83.2881 %
Root relative squared error             91.0847 %
Total Number of Instances              104

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.667    0.268      0.681     0.667     0.674      0.703    0
                 0.732    0.333      0.719     0.732     0.726      0.703    1
Weighted Avg.    0.702    0.303      0.702     0.702     0.702      0.703

=== Confusion Matrix ===

  a  b   <-- classified as
 32 16 |  a = 0
 15 41 |  b = 1
```

Εικόνα 65: Εξαγόμενα αποτελέσματα ConjuctiveRule για το σύνολο δεδομένων "Arkadikos 4 Seasons.csv

- **Αλγόριθμος Decision Table**

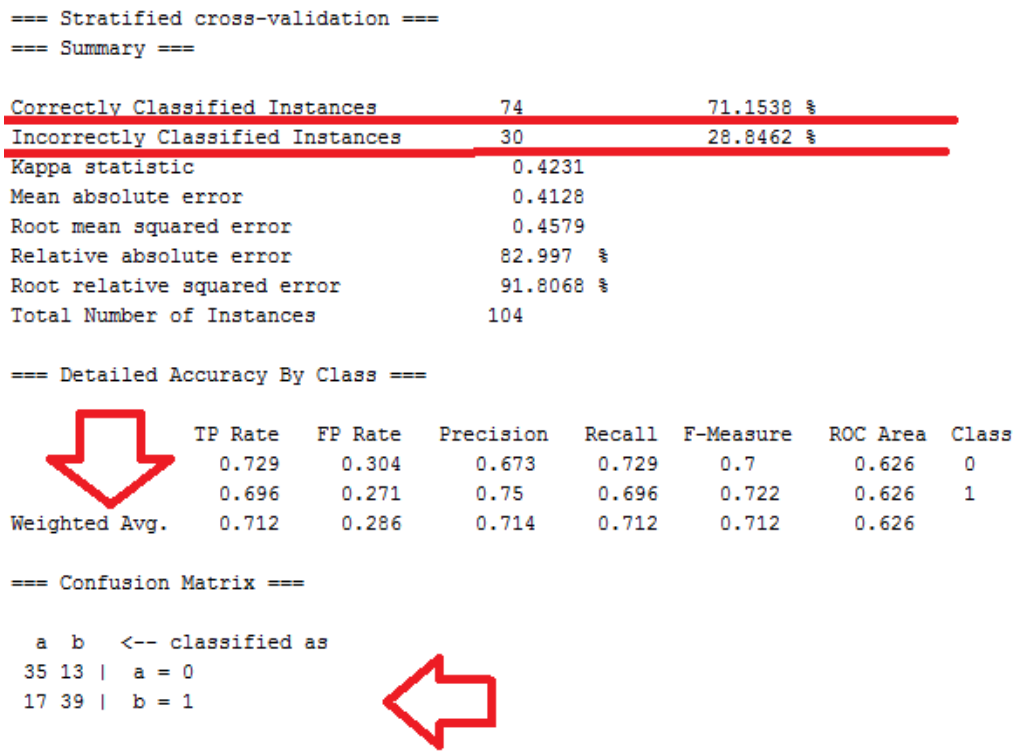
Μέσα από τη λίστα των διαθέσιμων τεχνικών των κανόνων ταξινόμησης «Rules» επιλέγουμε τον αλγόριθμο DecisionTable, έπειτα επιλέγουμε στο "Test options" το "Cross- Validation" αφήνοντας την επιλογή «10», έτσι ώστε να υλοποιηθεί η διαδικασία της διασταυρωμένης επικύρωσης με 10 πεδία για την τελική αξιολόγηση της απόδοσης της μεθόδου, και κλικάρουμε το tab "More Options" έτσι ώστε να ανοίξει ένα νέο παράθυρο διαλόγου, το "Classifier evaluation options", το οποίο μας παρέχει τη δυνατότητα να επιλέξουμε τα επιθυμητά εξαγόμενα αποτελέσματα τα οποία θα εμφανιστούν μετά την εκτέλεση του αλγορίθμου.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          69               66.3462 %
Incorrectly Classified Instances        35               33.6538 %
Kappa statistic                          0.3219
Mean absolute error                      0.4297
Root mean squared error                  0.4762
Relative absolute error                 86.3949 %
Root relative squared error             95.4615 %
Total Number of Instances              104

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.625     0.304      0.638     0.625      0.632      0.646    0
                 0.696     0.375      0.684     0.696      0.69       0.646    1
Weighted Avg.    0.663     0.342      0.663     0.663      0.663      0.646

=== Confusion Matrix ===

  a  b   <-- classified as
 30 18 |  a = 0
 17 39 |  b = 1
```

Εικόνα 66: Εξαγόμενα αποτελέσματα DecisionTable για το σύνολο δεδομένων "Arkadikos 4 Seasons.csv

- **Αλγόριθμος JRip**

Μέσα από τη λίστα των διαθέσιμων τεχνικών των κανόνων ταξινόμησης «Rules» επιλέγουμε τον αλγόριθμο JRip, έπειτα επιλέγουμε στο "Test options" το "Cross-Validation" αφήνοντας την επιλογή «10», έτσι ώστε να υλοποιηθεί η διαδικασία της διασταυρωμένης επικύρωσης με 10 πεδία για την τελική αξιολόγηση της απόδοσης της μεθόδου, και κλικάρουμε το tab "More Options" έτσι ώστε να ανοίξει ένα νέο παράθυρο διαλόγου, το "Classifier evaluation options", το οποίο μας παρέχει τη δυνατότητα να επιλέξουμε τα επιθυμητά εξαγόμενα αποτελέσματα τα οποία θα εμφανιστούν μετά την εκτέλεση του αλγορίθμου.

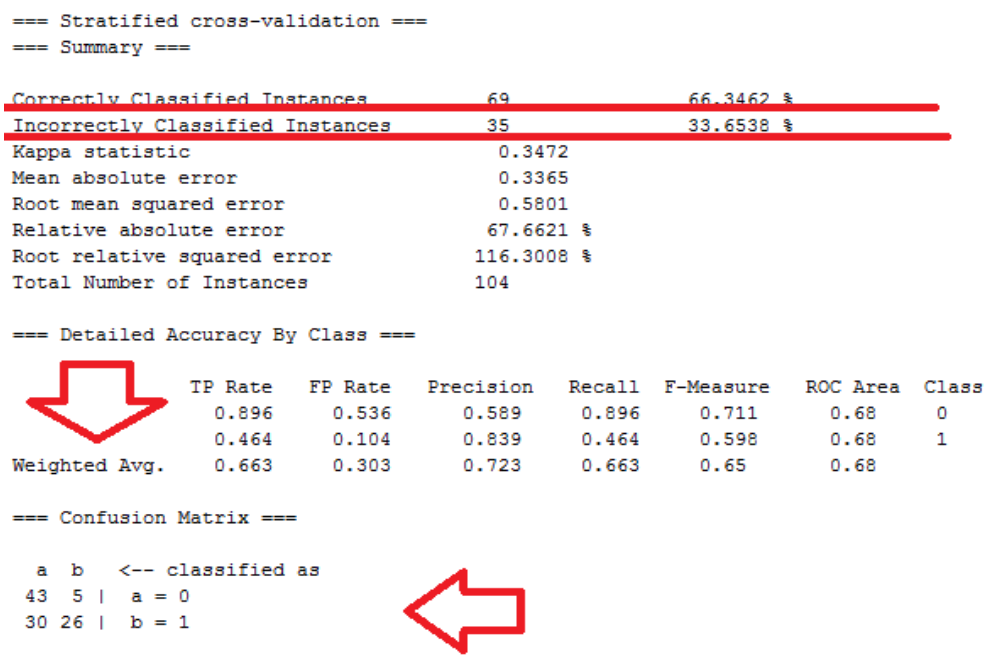


```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          74               71.1538 %
Incorrectly Classified Instances        30               28.8462 %
Kappa statistic                          0.4231
Mean absolute error                      0.4128
Root mean squared error                  0.4579
Relative absolute error                 82.997  %
Root relative squared error             91.8068 %
Total Number of Instances              104

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.729     0.304       0.673     0.729       0.7       0.626     0
                   0.696     0.271       0.75      0.696       0.722     0.626     1
Weighted Avg.      0.712     0.286       0.714     0.712       0.712     0.626

=== Confusion Matrix ===

  a  b   <-- classified as
 35 13 |  a = 0
 17 39 |  b = 1
```

Εικόνα 67: Εξαγόμενα αποτελέσματα JRip για το σύνολο δεδομένων "Arkadikos 4 Seasons.csv

- **Αλγόριθμος OneR**

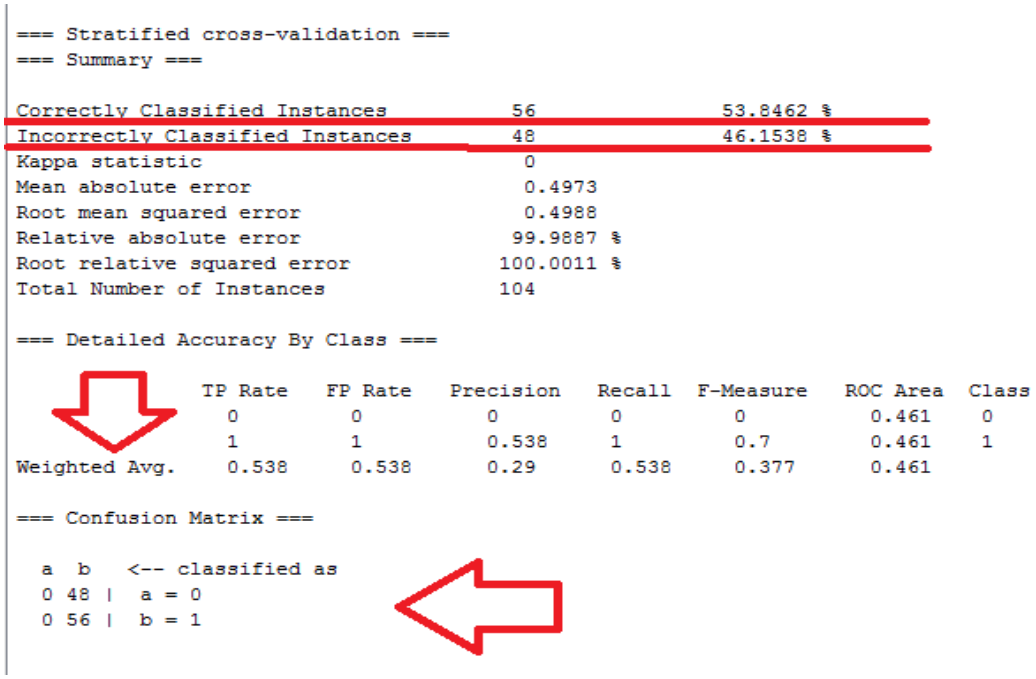
Μέσα από τη λίστα των διαθέσιμων τεχνικών των κανόνων ταξινόμησης «Rules» επιλέγουμε τον αλγόριθμο OneR, έπειτα επιλέγουμε στο "Test options" το "Cross-Validation" αφήνοντας την επιλογή «10», έτσι ώστε να υλοποιηθεί η διαδικασία της διασταυρωμένης επικύρωσης με 10 πεδία για την τελική αξιολόγηση της απόδοσης της μεθόδου, και κλικάρουμε το tab "More Options" έτσι ώστε να ανοίξει ένα νέο παράθυρο διαλόγου, το "Classifier evaluation options", το οποίο μας παρέχει τη δυνατότητα να επιλέξουμε τα επιθυμητά εξαγόμενα αποτελέσματα τα οποία θα εμφανιστούν μετά την εκτέλεση του αλγορίθμου.



Εικόνα 68: Εξαγόμενα αποτελέσματα OneR για το σύνολο δεδομένων "Arkadikos 4 Seasons.csv

- **Αλγόριθμος PART**

Μέσα από τη λίστα των διαθέσιμων τεχνικών των κανόνων ταξινόμησης «Rules» επιλέγουμε τον αλγόριθμο PART, έπειτα επιλέγουμε στο "Test options" το "Cross-Validation" αφήνοντας την επιλογή «10», έτσι ώστε να υλοποιηθεί η διαδικασία της διασταυρωμένης επικύρωσης με 10 πεδία για την τελική αξιολόγηση της απόδοσης της μεθόδου, και κλικάρουμε το tab "More Options" έτσι ώστε να ανοίξει ένα νέο παράθυρο διαλόγου, το "Classifier evaluation options", το οποίο μας παρέχει τη δυνατότητα να επιλέξουμε τα επιθυμητά εξαγόμενα αποτελέσματα τα οποία θα εμφανιστούν μετά την εκτέλεση του αλγορίθμου.

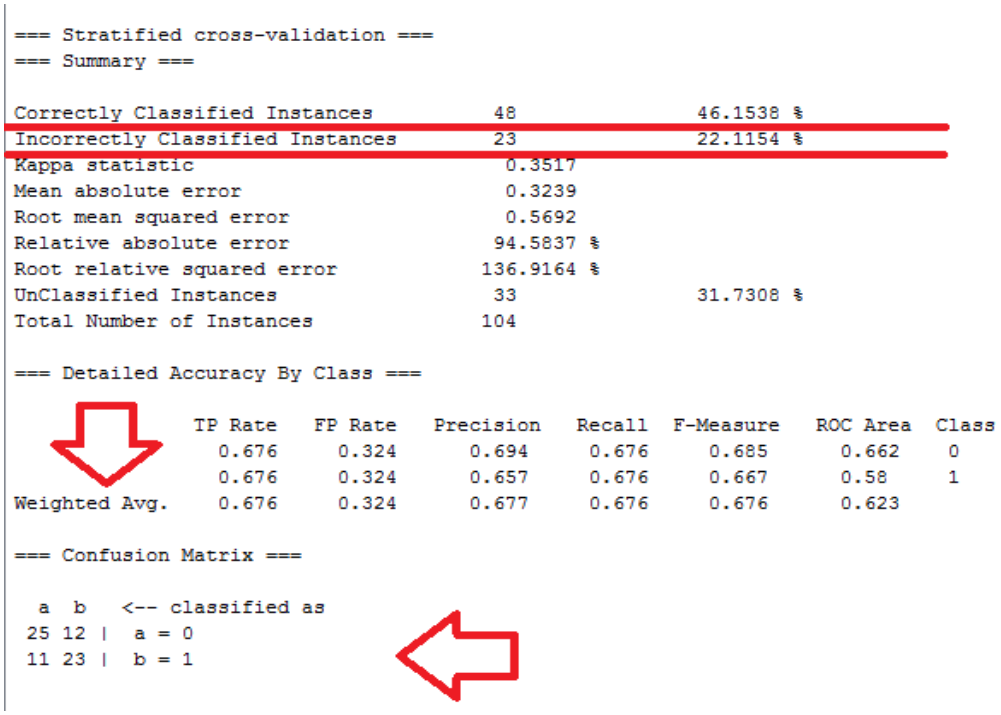


```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          56               53.8462 %
Incorrectly Classified Instances        48               46.1538 %
Kappa statistic                          0
Mean absolute error                      0.4973
Root mean squared error                  0.4988
Relative absolute error                 99.9887 %
Root relative squared error            100.0011 %
Total Number of Instances              104

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.461    0
                 1         1          0.538     1         0.7        0.461    1
Weighted Avg.    0.538     0.538      0.29      0.538     0.377      0.461

=== Confusion Matrix ===

  a  b   <-- classified as
  0 48 |  a = 0
  0 56 |  b = 1
```

Εικόνα 69: Εξαγόμενα αποτελέσματα PART για το σύνολο δεδομένων "Arkadikos 4 Seasons.csv

- **Αλγόριθμος Prism**

Μέσα από τη λίστα των διαθέσιμων τεχνικών των κανόνων ταξινόμησης «Rules» επιλέγουμε τον αλγόριθμο Prism, έπειτα επιλέγουμε στο "Test options" το "Cross-Validation" αφήνοντας την επιλογή «10», έτσι ώστε να υλοποιηθεί η διαδικασία της διασταυρωμένης επικύρωσης με 10 πεδία για την τελική αξιολόγηση της απόδοσης της μεθόδου, και κλικάρουμε το tab "More Options" έτσι ώστε να ανοίξει ένα νέο παράθυρο διαλόγου, το "Classifier evaluation options", το οποίο μας παρέχει τη δυνατότητα να επιλέξουμε τα επιθυμητά εξαγόμενα αποτελέσματα τα οποία θα εμφανιστούν μετά την εκτέλεση του αλγορίθμου.

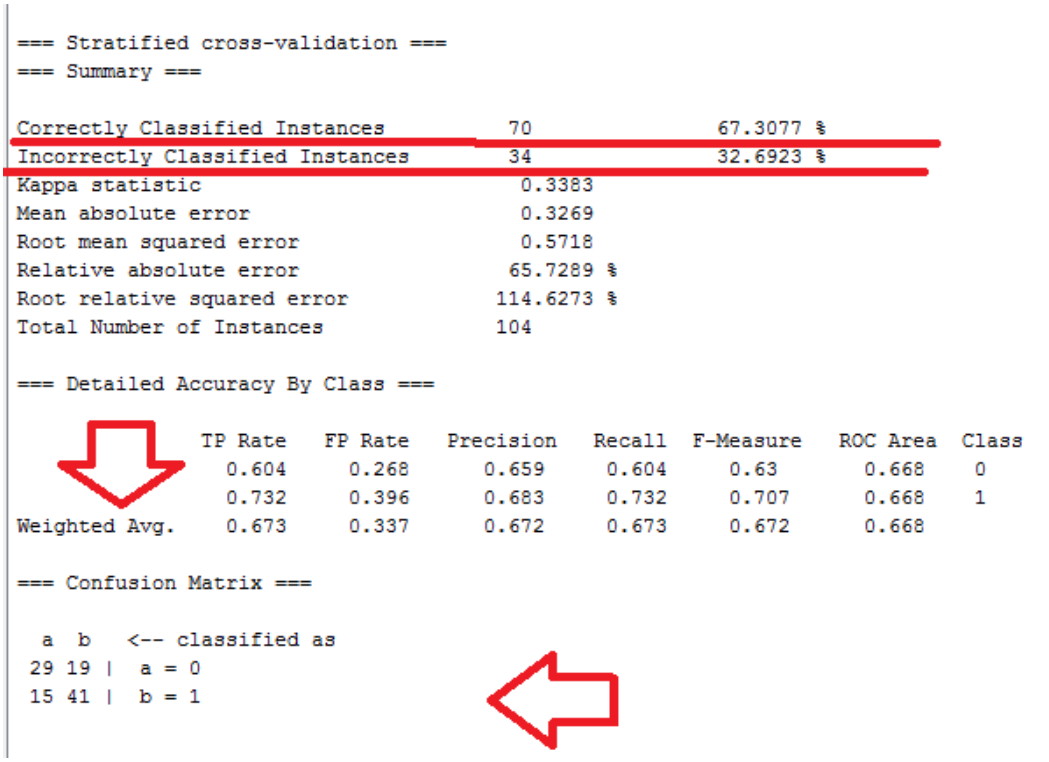


```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          48               46.1538 %
Incorrectly Classified Instances        23               22.1154 %
Kappa statistic                          0.3517
Mean absolute error                      0.3239
Root mean squared error                  0.5692
Relative absolute error                 94.5837 %
Root relative squared error            136.9164 %
UnClassified Instances                  33               31.7308 %
Total Number of Instances              104

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.676     0.324      0.694     0.676     0.685      0.662    0
                 0.676     0.324      0.657     0.676     0.667      0.58     1
Weighted Avg.    0.676     0.324      0.677     0.676     0.676      0.623

=== Confusion Matrix ===

  a  b   <-- classified as
 25 12 |  a = 0
 11 23 |  b = 1
```

Εικόνα 70: Εξαγόμενα αποτελέσματα Prism για το σύνολο δεδομένων "Arkadikos 4 Seasons.csv

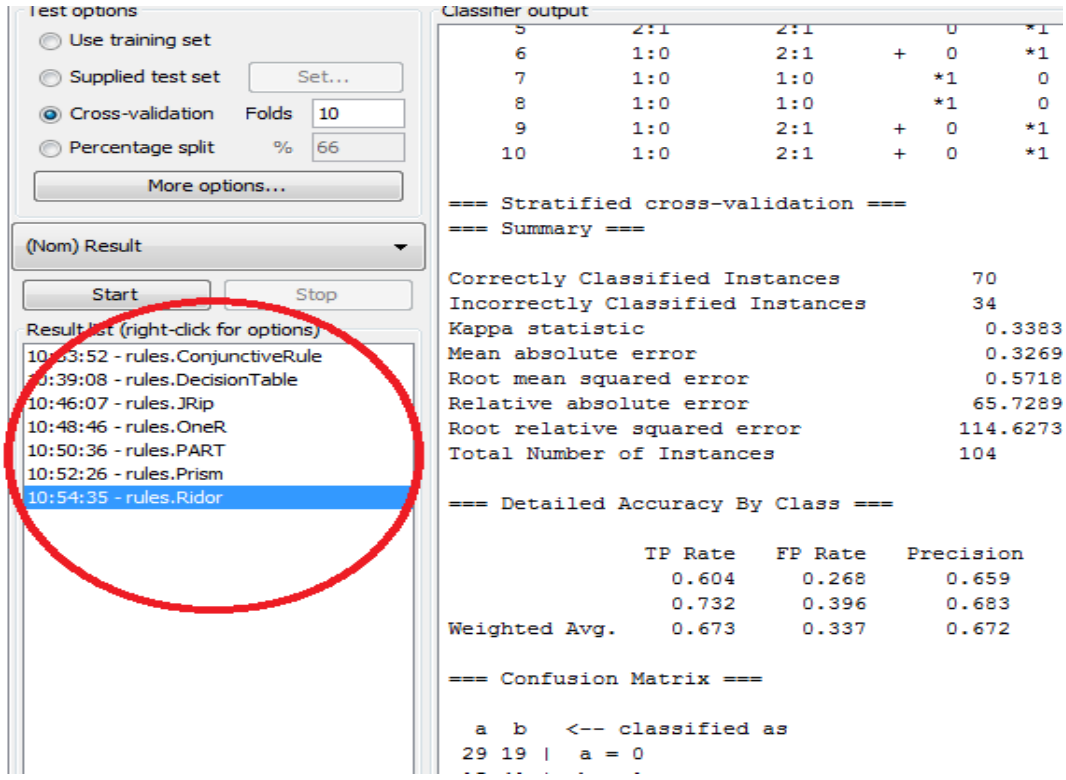
- **Αλγόριθμος Ridor**

Μέσα από τη λίστα των διαθέσιμων τεχνικών των κανόνων ταξινόμησης «Rules» επιλέγουμε τον αλγόριθμο Ridor, έπειτα επιλέγουμε στο "Test options" το "Cross-Validation" αφήνοντας την επιλογή «10», έτσι ώστε να υλοποιηθεί η διαδικασία της διασταυρωμένης επικύρωσης με 10 πεδία για την τελική αξιολόγηση της απόδοσης της μεθόδου, και κλικάρουμε το tab "More Options" έτσι ώστε να ανοίξει ένα νέο παράθυρο διαλόγου, το "Classifier evaluation options", το οποίο μας παρέχει τη δυνατότητα να επιλέξουμε τα επιθυμητά εξαγόμενα αποτελέσματα τα οποία θα εμφανιστούν μετά την εκτέλεση του αλγορίθμου.



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          70               67.3077 %
Incorrectly Classified Instances        34               32.6923 %
Kappa statistic                          0.3383
Mean absolute error                      0.3269
Root mean squared error                  0.5718
Relative absolute error                 65.7289 %
Root relative squared error            114.6273 %
Total Number of Instances              104

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.604     0.268      0.659     0.604      0.63       0.668    0
                 0.732     0.396      0.683     0.732      0.707      0.668    1
Weighted Avg.    0.673     0.337      0.672     0.673      0.672      0.668

=== Confusion Matrix ===

  a  b   <-- classified as
 29 19 |  a = 0
 15 41 |  b = 1
```

Εικόνα 71: Εξαγόμενα αποτελέσματα Ridor για το σύνολο δεδομένων "Arkadikos 4 Seasons.csv



Εικόνα 72: Περιβάλλον Weka μετά την υλοποίηση των αλγορίθμων κανόνων ταξινόμησης 1) ConjuctiveRule, 2) DecisionTable, 3) JRip, 4) OneR, 5) PART, 6) Prism, 7) Ridor

Ακολουθούν οι παραγόμενοι κανόνες ταξινόμησης:

Single rule learner:

(Home/Visitor = 1) => Result = 1

If Points >= 70 then 1

If Def Rebounds < = 21 then 0,

If T/O> = 22 then 0

If Tendex >= 0.38 then 1

If Off Rebounds >= 13 then 1,

If Free Throws> = 0.88 then 1

If Assists >= 21 then 1

If Blocks> = 5 then 1

If 2p Goals >= 0.6 then 1

If 3p Goals >= 0.28 then 1,

If 3p Goals = 0.32 and Points = 75 then 0

If 3p Goals = 0.29 and Assists = 14 then 1

Result = 0

1. Except (Home/Visitor = 1) and (Off Rebounds = 8) => Result = 1

2. Except (Home/Visitor = 1) => Result = 1

Από τα αποτελέσματα τα οποία προέκυψαν, τα οποία παρουσιάζονται συγκεντρωτικά στον παρακάτω πίνακα, διαπιστώνουμε ότι

1. Ο αλγόριθμος με την καλύτερη απόδοση και μέγιστη προβλεπτική ικανότητα είναι ο αλγόριθμος JRip για τον οποίο παρατηρείται το μεγαλύτερο ποσοστό ακρίβειας, μεγαλύτερο ποσοστό σωστά ταξινομημένων υποδειγμάτων, μηδενικό ποσοστό αταξινόμητων υποδειγμάτων, το μεγαλύτερο ποσοστό των θετικών υποδειγμάτων που έχουν αναγνωριστεί σωστά ως θετικά (μεγαλύτερη ευαισθησία) και το μικρότερο ποσοστό των αρνητικών υποδειγμάτων που έχουν αναγνωριστεί λανθασμένα ως θετικά (μεγαλύτερη ειδικότητα), υψηλότερες τιμές για όλα τα υπό εξέταση μέτρα ακρίβειας (precision, recall, f-measure), και τιμή εμβαδού κάτω από την ROC καμπύλη ίση με 0.626.

2. Παρόμοια απόδοση με τον αλγόριθμο JRip παρατηρείται από τον αλγόριθμο ConjuctiveRule. Ο ConjuctiveRule παρουσιάζει τη δεύτερη καλύτερη απόδοση δεδομένου ότι παρατηρείται το 2^{ο} μεγαλύτερο ποσοστό ακρίβειας, 2^{ο} μεγαλύτερο ποσοστό σωστά ταξινομημένων υποδειγμάτων, μηδενικό ποσοστό αταξινόμητων υποδειγμάτων, το 2^{ο} μεγαλύτερο ποσοστό των θετικών υποδειγμάτων που έχουν αναγνωριστεί σωστά ως θετικά (μεγαλύτερη ευαισθησία) και το 2^{ο} μικρότερο ποσοστό των αρνητικών υποδειγμάτων που έχουν αναγνωριστεί λανθασμένα ως θετικά (μεγαλύτερη ειδικότητα), 2^{ες} υψηλότερες τιμές για όλα τα υπό εξέταση μέτρα ακρίβειας (precision, recall, f-measure), και η υψηλότερη τιμή εμβαδού κάτω από την ROC καμπύλη ίση με 0.703.

3. Ο αλγόριθμος Ridor έχει την τρίτη καλύτερη απόδοση.

4. Ακολουθεί ο OneR με την τέταρτη καλύτερη απόδοση, και ο Decision Table με την πέμπτη καλύτερη απόδοση. Οι αλγόριθμοι OneR και DecisionTable εμφανίζουν παρόμοια απόδοση (πανομοιότυπα αποτελέσματα).

5. Ακολουθεί ο PART με την έκτη καλύτερη απόδοση.

6. Ο αλγόριθμος με τη χειρότερη απόδοση είναι ο αλγόριθμος Prism, του οποίου τα αποτελέσματα δεν λαμβάνουμε καθόλου υπόψη, δεδομένου ότι δεν είναι αξιόπιστα (31.73 % των υποδειγμάτων δεν ταξινομήθηκαν καθόλου).

Πίνακας 5: Συγκριτικά αποτελέσματα κανόνων ταξινόμησης για το σύνολο δεδομένων "Arkadikos 4 Seasons.csv"

	Correctly Classified Instances - Accuracy	Incorrectly Classified Instances	Unclassified Instances	Avg. TP Rate	Avg. FP Rate	Avg. Precision	Avg. Recall	Avg. F-Measure	Avg. ROC Area
ConjuctiveRule	**70.20%**	**29.80%**	**0%**	**0.702**	**0.303**	**0.702**	**0.702**	**0.702**	**0.703**
DecisionTable	**66.34%**	**33.66%**	**0%**	**0.663**	**0.342**	**0.663**	**0.663**	**0.663**	**0.646**
JRip	**71.15%**	**28.85**	**0%**	**0.712**	**0.286**	**0.714**	**0.712**	**0.712**	**0.626**

OneR	**66.35%**	**33.65%**	**0%**	**0.663**	**0.303**	**0.723**	**0.663**	**0.650**	**0.680**
PART	**53.85%**	**46.15%**	**0%**	**0.538**	**0.538**	**0.290**	**0.538**	**0.377**	**0.461**
Prism	**46.15%**	**22.12%**	**31.73%**	**0.676**	**0.324**	**0.677**	**0.676**	**0.676**	**0.623**
Ridor	**67.30%**	**32.70%**	**0%**	**0.673**	**0.337**	**0.672**	**0.673**	**0.672**	**0.668**

14.3.2 Κανόνες ταξινόμησης για την ομάδα του Λαυρίου

Αρχικά ακολουθούμε τα παρακάτω βήματα (τα οποία ακολουθήσαμε πριν την υλοποίηση των αλγορίθμων δέντρων απόφασης, ID-3, J48 και CART).

1. Ακολουθούμε τα βήματα

Explorer→Preprocess→Open File

και «φορτώνοντας» το αρχείο "Lavrio 4 seasons.csv", το οποίο αφορά στην ομάδα του Λαυρίου και περιλαμβάνει τις υπό εξέταση μεταβλητές για τέσσερις συνολικά χρονικές περιόδους 2011-2012, 2012-2013, 2013-2014 και 2014-2015,

2. Αφαιρούμε τη μεταβλητή «Lavrio» (τικάρουμε και Remove), η οποία αποτελεί τη μεταβλητή- Index η οποία απαριθμεί τα παιχνίδια ($1^{ο}$ παιχνίδι, $2^{ο}$ παιχνίδι κ.λπ.) των τεσσάρων χρονικών περιόδων, και επιλέγουμε το χαρακτηριστικό «result» ως αυτό το οποίο δείχνει σε ποια κλάση ανήκει κάθε φορά το υπόδειγμα. Το χαρακτηριστικό «result» αποτελεί μία δίτιμη μεταβλητή με τη τιμή **0 να δηλώνει την «ήττα» και την τιμή 1 τη «νίκη».**

3. Προχωρούμε στη διαδικασία διακριτοποίησης των δεδομένων μας, ακολουθώντας τα παρακάτω βήματα:

«Choose→weka→filters→unsupervised→attribute»

Έπειτα, επιλέγουμε από τη λίστα διαθέσιμων επιλογών " το "Numeric to Nominal" κλικάρουμε "Filter→Filtering Capabilities"

- ✓ Numeric Attributes
- ✓ Numeric class

Ok

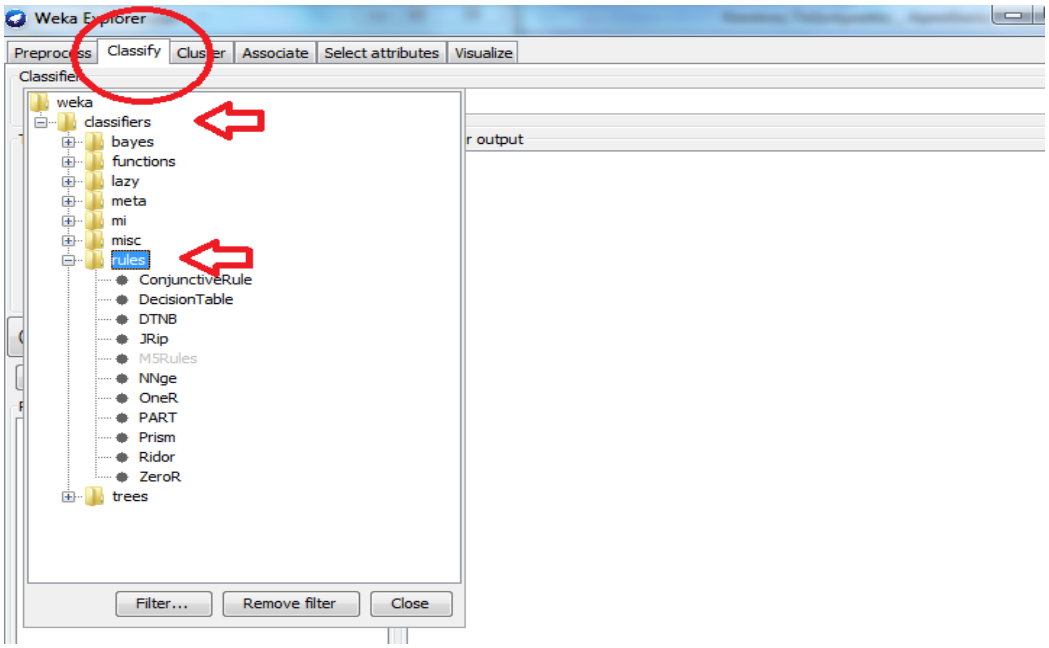
και →Apply.

Επιλέγουμε στο πάνελ την καρτέλα «Classify» όπου περιλαμβάνονται ποικίλες τεχνικές ταξινόμησης με διαφορετικές δυνατότητες για το χρήστη η κάθε μία.

Θα εφαρμόσουμε τεχνικές στις οποίες η κατηγοριοποίηση των δεδομένων υλοποιείται μέσω της εύρεσης κανόνων απόφασης/ταξινόμησης.

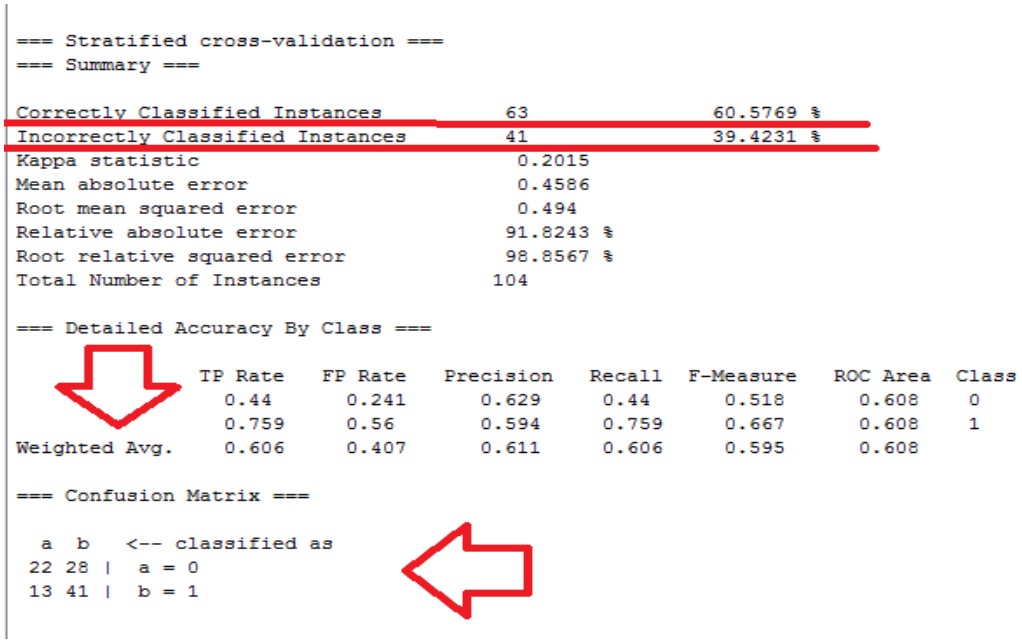
Ακολουθούμε την παρακάτω διαδικασία»:

«Choose→Weka→Classifiers→Rules»



- **Αλγόριθμος Conjuctive Rule**

Μέσα από τη λίστα των διαθέσιμων τεχνικών των κανόνων ταξινόμησης «Rules» επιλέγουμε τον αλγόριθμο ConjuctiveRule, έπειτα επιλέγουμε στο "Test options" το "Cross- Validation" αφήνοντας την επιλογή «10», έτσι ώστε να υλοποιηθεί η διαδικασία της διασταυρωμένης επικύρωσης με 10 πεδία για την τελική αξιολόγηση της απόδοσης της μεθόδου, και κλικάρουμε το tab "More Options" έτσι ώστε να ανοίξει ένα νέο παράθυρο διαλόγου, το "Classifier evaluation options", το οποίο μας παρέχει τη δυνατότητα να επιλέξουμε τα επιθυμητά εξαγόμενα αποτελέσματα τα οποία θα εμφανιστούν μετά την εκτέλεση του αλγορίθμου.

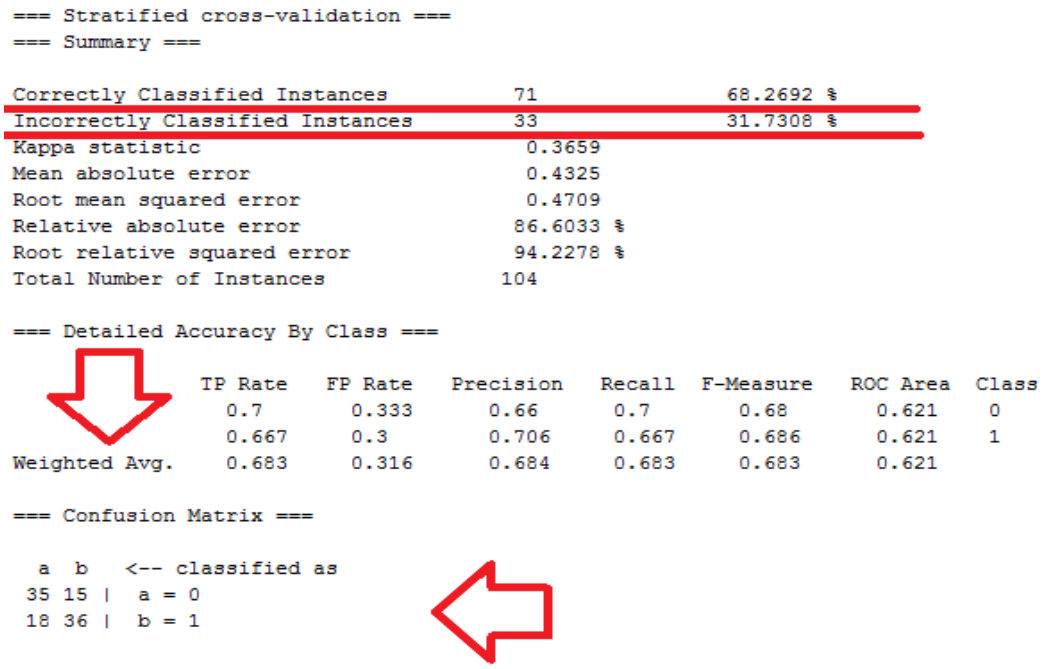


```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          63               60.5769 %
Incorrectly Classified Instances        41               39.4231 %
Kappa statistic                          0.2015
Mean absolute error                      0.4586
Root mean squared error                  0.494
Relative absolute error                 91.8243 %
Root relative squared error             98.8567 %
Total Number of Instances              104

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.44      0.241      0.629     0.44      0.518      0.608    0
                 0.759     0.56       0.594     0.759     0.667      0.608    1
Weighted Avg.    0.606     0.407      0.611     0.606     0.595      0.608

=== Confusion Matrix ===

  a  b   <-- classified as
 22 28 |  a = 0
 13 41 |  b = 1
```

Εικόνα 73: Εξαγόμενα αποτελέσματα ConjuctiveRule για το σύνολο δεδομένων "Lavrio 4 Seasons.csv

Αλγόριθμος Decision Table

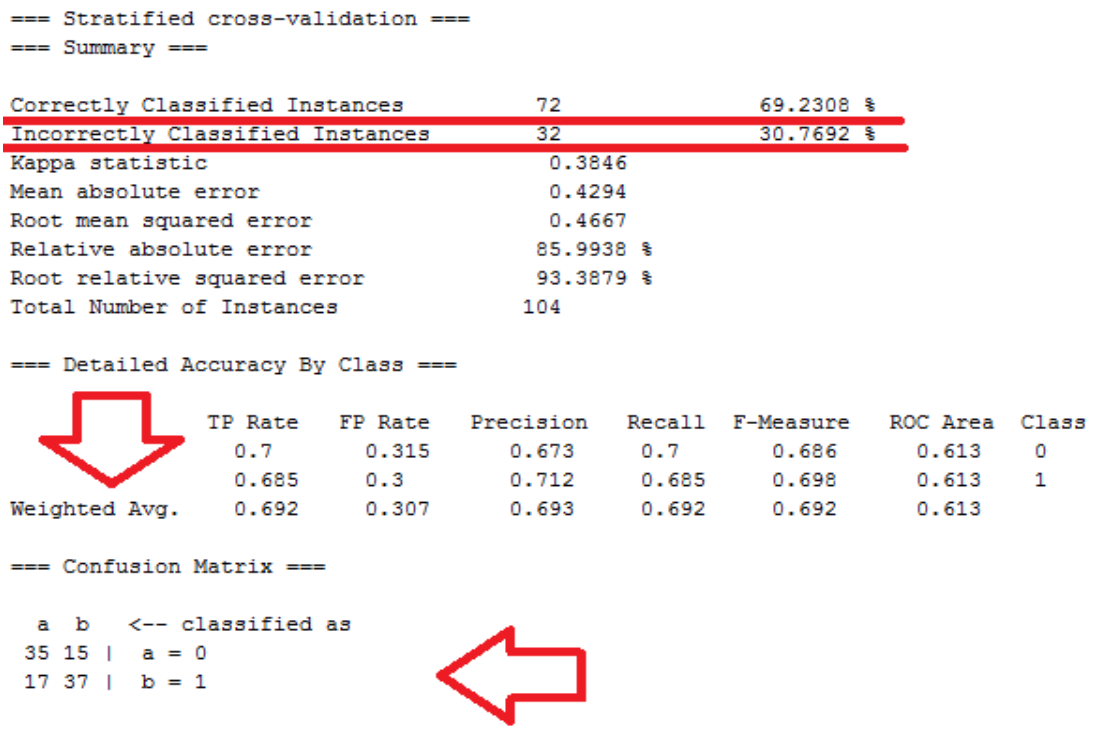
Μέσα από τη λίστα των διαθέσιμων τεχνικών των κανόνων ταξινόμησης «Rules» επιλέγουμε τον αλγόριθμο DecisionTable, έπειτα επιλέγουμε στο "Test options" το "Cross- Validation" αφήνοντας την επιλογή «10», έτσι ώστε να υλοποιηθεί η διαδικασία της διασταυρωμένης επικύρωσης με 10 πεδία για την τελική αξιολόγηση της απόδοσης της μεθόδου, και κλικάρουμε το tab "More Options" έτσι ώστε να ανοίξει ένα νέο παράθυρο διαλόγου, το "Classifier evaluation options", το οποίο μας παρέχει τη δυνατότητα να επιλέξουμε τα επιθυμητά εξαγόμενα αποτελέσματα τα οποία θα εμφανιστούν μετά την εκτέλεση του αλγορίθμου.



Εικόνα 74: Εξαγόμενα αποτελέσματα DecisionTable για το σύνολο δεδομένων "Lavrio 4 Seasons.csv

- **Αλγόριθμος JRip**

Μέσα από τη λίστα των διαθέσιμων τεχνικών των κανόνων ταξινόμησης «Rules» επιλέγουμε τον αλγόριθμο JRip, έπειτα επιλέγουμε στο "Test options" το "Cross-Validation" αφήνοντας την επιλογή «10», έτσι ώστε να υλοποιηθεί η διαδικασία της διασταυρωμένης επικύρωσης με 10 πεδία για την τελική αξιολόγηση της απόδοσης της μεθόδου, και κλικάρουμε το tab "More Options" έτσι ώστε να ανοίξει ένα νέο παράθυρο διαλόγου, το "Classifier evaluation options", το οποίο μας παρέχει τη δυνατότητα να επιλέξουμε τα επιθυμητά εξαγόμενα αποτελέσματα τα οποία θα εμφανιστούν μετά την εκτέλεση του αλγορίθμου.

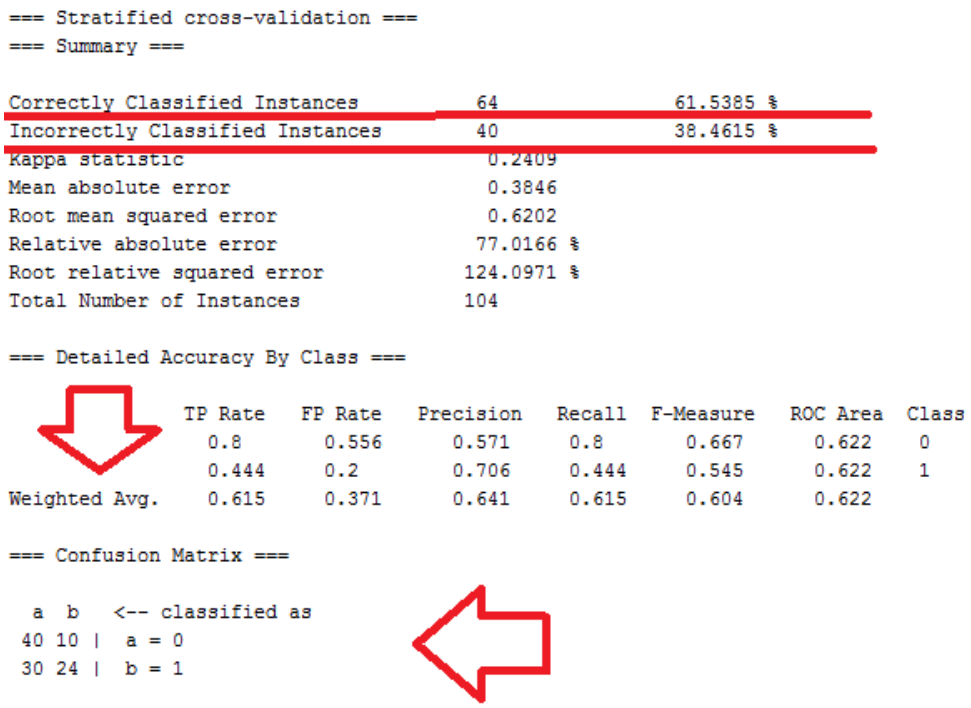


```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          72               69.2308 %
Incorrectly Classified Instances        32               30.7692 %
Kappa statistic                          0.3846
Mean absolute error                      0.4294
Root mean squared error                  0.4667
Relative absolute error                 85.9938 %
Root relative squared error             93.3879 %
Total Number of Instances              104

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.7       0.315       0.673     0.7       0.686      0.613    0
                 0.685     0.3         0.712     0.685     0.698      0.613    1
Weighted Avg.    0.692     0.307       0.693     0.692     0.692      0.613

=== Confusion Matrix ===

  a  b   <-- classified as
 35 15 |  a = 0
 17 37 |  b = 1
```

Εικόνα 75: Εξαγόμενα αποτελέσματα JRip για το σύνολο δεδομένων "Lavrio 4 Seasons.csv

- **Αλγόριθμος OneR**

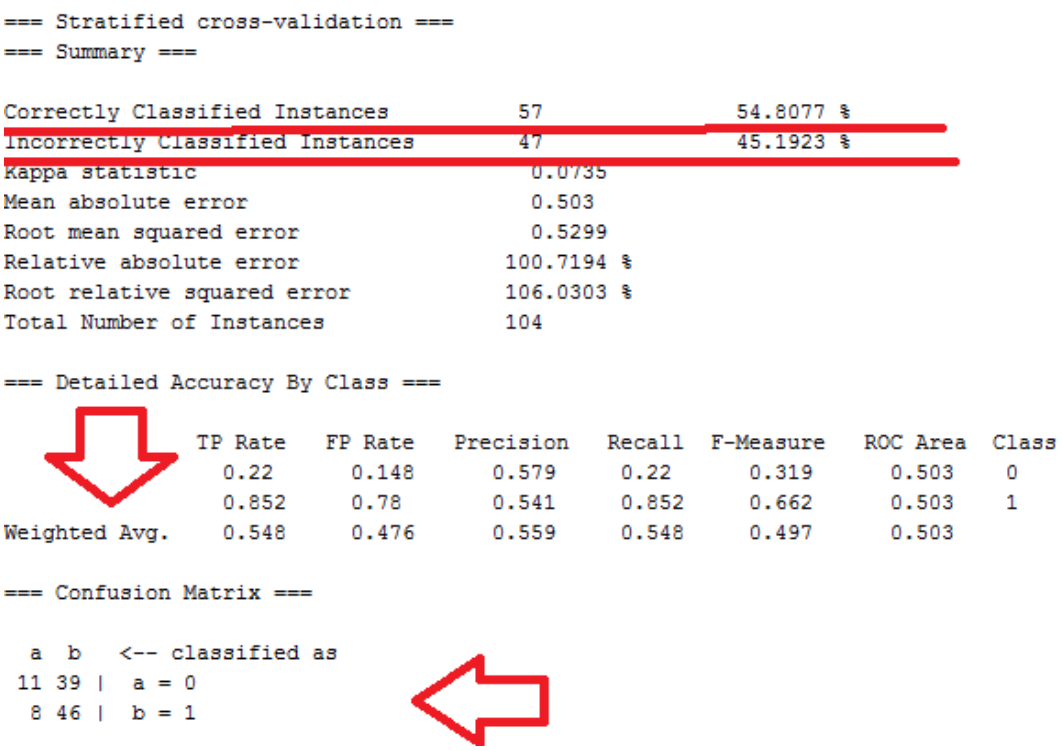
Μέσα από τη λίστα των διαθέσιμων τεχνικών των κανόνων ταξινόμησης «Rules» επιλέγουμε τον αλγόριθμο OneR, έπειτα επιλέγουμε στο "Test options" το "Cross-Validation" αφήνοντας την επιλογή «10», έτσι ώστε να υλοποιηθεί η διαδικασία της διασταυρωμένης επικύρωσης με 10 πεδία για την τελική αξιολόγηση της απόδοσης της μεθόδου, και κλικάρουμε το tab "More Options" έτσι ώστε να ανοίξει ένα νέο παράθυρο διαλόγου, το "Classifier evaluation options", το οποίο μας παρέχει τη δυνατότητα να επιλέξουμε τα επιθυμητά εξαγόμενα αποτελέσματα τα οποία θα εμφανιστούν μετά την εκτέλεση του αλγορίθμου.



Εικόνα 76: Εξαγόμενα αποτελέσματα OneR για το σύνολο δεδομένων "Lavrio 4 Seasons.csv

- **Αλγόριθμος PART**

Μέσα από τη λίστα των διαθέσιμων τεχνικών των κανόνων ταξινόμησης «Rules» επιλέγουμε τον αλγόριθμο PART, έπειτα επιλέγουμε στο "Test options" το "Cross-Validation" αφήνοντας την επιλογή «10», έτσι ώστε να υλοποιηθεί η διαδικασία της διασταυρωμένης επικύρωσης με 10 πεδία για την τελική αξιολόγηση της απόδοσης της μεθόδου, και κλικάρουμε το tab "More Options" έτσι ώστε να ανοίξει ένα νέο παράθυρο διαλόγου, το "Classifier evaluation options", το οποίο μας παρέχει τη δυνατότητα να επιλέξουμε τα επιθυμητά εξαγόμενα αποτελέσματα τα οποία θα εμφανιστούν μετά την εκτέλεση του αλγορίθμου.



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          57               54.8077 %
Incorrectly Classified Instances        47               45.1923 %
Kappa statistic                          0.0735
Mean absolute error                      0.503
Root mean squared error                  0.5299
Relative absolute error                100.7194 %
Root relative squared error            106.0303 %
Total Number of Instances              104

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.22      0.148       0.579     0.22      0.319      0.503    0
                 0.852     0.78        0.541     0.852     0.662      0.503    1
Weighted Avg.    0.548     0.476       0.559     0.548     0.497      0.503

=== Confusion Matrix ===

  a  b   <-- classified as
 11 39 |  a = 0
  8 46 |  b = 1
```

Εικόνα 77: Εξαγόμενα αποτελέσματα PART για το σύνολο δεδομένων "Lavrio 4 Seasons.csv

- **Αλγόριθμος Prism**

Μέσα από τη λίστα των διαθέσιμων τεχνικών των κανόνων ταξινόμησης «Rules» επιλέγουμε τον αλγόριθμο Prism, έπειτα επιλέγουμε στο "Test options" το "Cross-Validation" αφήνοντας την επιλογή «10», έτσι ώστε να υλοποιηθεί η διαδικασία της διασταυρωμένης επικύρωσης με 10 πεδία για την τελική αξιολόγηση της απόδοσης της μεθόδου, και κλικάρουμε το tab "More Options" έτσι ώστε να ανοίξει ένα νέο παράθυρο διαλόγου, το "Classifier evaluation options", το οποίο μας παρέχει τη δυνατότητα να επιλέξουμε τα επιθυμητά εξαγόμενα αποτελέσματα τα οποία θα εμφανιστούν μετά την εκτέλεση του αλγορίθμου.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          53               50.9615 %
Incorrectly Classified Instances        30               28.8462 %
Kappa statistic                          0.2828
Mean absolute error                      0.3614
Root mean squared error                  0.6012
Relative absolute error                 90.7518 %
Root relative squared error            134.7543 %
UnClassified Instances                  21               20.1923 %
Total Number of Instances              104

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.718     0.432       0.596     0.718      0.651      0.604    0
                 0.568     0.282       0.694     0.568      0.625      0.621    1
Weighted Avg.    0.639     0.352       0.648     0.639      0.637      0.613

=== Confusion Matrix ===

  a  b   <-- classified as
 28 11 |  a = 0
 19 25 |  b = 1
```

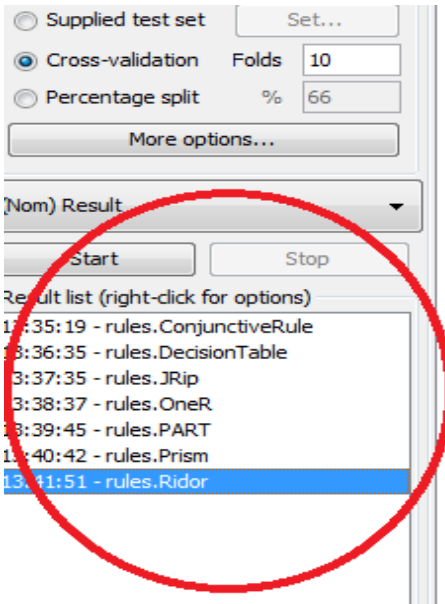
Εικόνα 78: Εξαγόμενα αποτελέσματα Prism για το σύνολο δεδομένων "Lavrio 4 Seasons.csv

- **Αλγόριθμος Ridor**

Μέσα από τη λίστα των διαθέσιμων τεχνικών των κανόνων ταξινόμησης «Rules» επιλέγουμε τον αλγόριθμο Ridor, έπειτα επιλέγουμε στο "Test options" το "Cross-Validation" αφήνοντας την επιλογή «10», έτσι ώστε να υλοποιηθεί η διαδικασία της διασταυρωμένης επικύρωσης με 10 πεδία για την τελική αξιολόγηση της απόδοσης της μεθόδου, και κλικάρουμε το tab "More Options" έτσι ώστε να ανοίξει ένα νέο παράθυρο διαλόγου, το "Classifier evaluation options", το οποίο μας παρέχει τη δυνατότητα να επιλέξουμε τα επιθυμητά εξαγόμενα αποτελέσματα τα οποία θα εμφανιστούν μετά την εκτέλεση του αλγορίθμου.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          69               66.3462 %
Incorrectly Classified Instances        35               33.6538 %
Kappa statistic                          0.3254
Mean absolute error                      0.3365
Root mean squared error                  0.5801
Relative absolute error                 67.3895 %
Root relative squared error            116.0823 %
Total Number of Instances              104

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.64      0.315      0.653     0.64      0.646      0.663    0
                   0.685     0.36       0.673     0.685     0.679      0.663    1
Weighted Avg.      0.663     0.338      0.663     0.663     0.663      0.663

=== Confusion Matrix ===

  a  b   <-- classified as
 32 18 |  a = 0
 17 37 |  b = 1
```

Εικόνα 79: Εξαγόμενα αποτελέσματα Ridor για το σύνολο δεδομένων "Lavrio 4 Seasons.csv




```
     7        1:0        2:1        +    0
     8        1:0        2:1        +    0
     9        1:0        2:1        +    0
    10        1:0        1:0           *1

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          69
Incorrectly Classified Instances        35
Kappa statistic                          0
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                 67
Root relative squared error            116
Total Number of Instances              104

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precisi
                   0.64      0.315      0.65
                   0.685     0.36       0.67
Weighted Avg.      0.663     0.338      0.66

=== Confusion Matrix ===

  a  b   <-- classified as
 32 18 |  a = 0
 17 37 |  b = 1
```

Εικόνα 80: Περιβάλλον Weka μετά την υλοποίηση των αλγορίθμων κανόνων ταξινόμησης 1) ConjuctiveRule, 2) DecisionTable, 3) JRip, 4) OneR, 5) PART, 6) Prism, 7) Ridor

Ακολουθούν οι παραγόμενοι κανόνες ταξινόμησης:

Single rule learner:

(Home/Visitor = 1) => Result = 1

If Tendex >=0.385 then 1

If Tendex = 0.385 AND Blocks = 0 then 1

If Tendex = 0.44 AND Home/Visitor = 0 then 0

If Points >= 78 then 1

If 2p Goals> = 0.43 then 1

If Def Rebounds = 26 then 1

If T/O > = 22 then 0

If 3p Goals >= 0.15 then 1

If Free Throws> = 0.59 then 1

If Points = 67 and Free Throws = 0.68 then 1

If Points = 69 and Free Throws = 0.75 then 1

Result = 0

1. Except (Home/Visitor = 1) and (Assists = 18) => Result = 1

2. Except (Home/Visitor = 1) => Result = 1

Από τα αποτελέσματα τα οποία προέκυψαν, τα οποία παρουσιάζονται συγκεντρωτικά στον παρακάτω πίνακα, διαπιστώνουμε ότι

1. Ο αλγόριθμος με την καλύτερη απόδοση και μέγιστη προβλεπτική ικανότητα είναι ο αλγόριθμος JRip για τον οποίο παρατηρείται το μεγαλύτερο ποσοστό ακρίβειας,

μεγαλύτερο ποσοστό σωστά ταξινομημένων υποδειγμάτων, μηδενικό ποσοστό αταξινόμητων υποδειγμάτων, το μεγαλύτερο ποσοστό των θετικών υποδειγμάτων που έχουν αναγνωριστεί σωστά ως θετικά (μεγαλύτερη ευαισθησία) και το μικρότερο ποσοστό των αρνητικών υποδειγμάτων που έχουν αναγνωριστεί λανθασμένα ως θετικά (μεγαλύτερη ειδικότητα), υψηλότερες τιμές για όλα τα υπό εξέταση μέτρα ακρίβειας (precision, recall, f-measure), και τιμή εμβαδού κάτω από την ROC καμπύλη ίση με 0.613.

2. Παρόμοια απόδοση με τον αλγόριθμο JRip παρατηρείται από τον αλγόριθμο DecisionTable. Ο DecisionTable παρουσιάζει τη δεύτερη καλύτερη απόδοση δεδομένου ότι παρατηρείται το 2ο μεγαλύτερο ποσοστό ακρίβειας, 2ο μεγαλύτερο ποσοστό σωστά ταξινομημένων υποδειγμάτων, μηδενικό ποσοστό αταξινόμητων υποδειγμάτων, το 2ο μεγαλύτερο ποσοστό των θετικών υποδειγμάτων που έχουν αναγνωριστεί σωστά ως θετικά (μεγαλύτερη ευαισθησία) και το 2ο μικρότερο ποσοστό των αρνητικών υποδειγμάτων που έχουν αναγνωριστεί λανθασμένα ως θετικά (μεγαλύτερη ειδικότητα), 2ες υψηλότερες τιμές για όλα τα υπό εξέταση μέτρα ακρίβειας (precision, recall, f-measure), και τιμή εμβαδού κάτω από την ROC καμπύλη ίση με 0.621.

3. Ο αλγόριθμος Ridor έχει την τρίτη καλύτερη απόδοση και την υψηλότερη τιμή εμβαδού κάτω από την ROC καμπύλη.

4. Ακολουθεί ο OneR με την τέταρτη καλύτερη απόδοση, και ο ConjuctiveRule έχει την πέμπτη καλύτερη απόδοση.

5. Ακολουθεί ο PART με την έκτη καλύτερη απόδοση.

6. Ο αλγόριθμος με τη χειρότερη απόδοση είναι ο αλγόριθμος Prism, του οποίου τα αποτελέσματα δεν λαμβάνουμε καθόλου υπόψη, δεδομένου ότι δεν είναι αξιόπιστα (20.19 % των υποδειγμάτων δεν ταξινομήθηκαν καθόλου).

Πίνακας 6: Συγκριτικά αποτελέσματα κανόνων ταξινόμησης για το σύνολο δεδομένων "Lavrio 4 Seasons.csv"

	Correctly Classified Instances - Accuracy	Incorrectly Classified Instances	Unclassified Instances	Avg. TP Rate	Avg. FP Rate	Avg. Precision	Avg. Recall	Avg. F-Measure	Avg. ROC Area
ConjuctiveRule	**60.58%**	**39.42%**	**0%**	**0.606**	**0.407**	**0.611**	**0.606**	**0.595**	**0.608**
DecisionTable	**68.27%**	**31.73%**	**0%**	**0.683**	**0.316**	**0.684**	**0.683**	**0.683**	**0.621**
JRip	**69.23%**	**30.77%**	**0%**	**0.692**	**0.307**	**0.693**	**0.692**	**0.692**	**0.613**
OneR	**61.54%**	**38.46%**	**0%**	**0.615**	**0.371**	**0.641**	**0.615**	**0.604**	**0.622**

PART	**54.80%**	**45.20%**	**0%**	**0.548**	**0.476**	**0.559**	**0.548**	**0.497**	**0.503**
Prism	**50.96%**	**28.85%**	**20.19%**	**0.639**	**0.352**	**0.648**	**0.639**	**0.637**	**0.613**
Ridor	**66.35%**	**33.65%**	**0%**	**0.663**	**0.338**	**0.663**	**0.663**	**0.663**	**0.663**

14.4 Συσταδοποίηση

14.4.1 Εύρεση συστάδων για τον Αρκαδικό 2012-2013

1. Ακολουθούμε τα βήματα

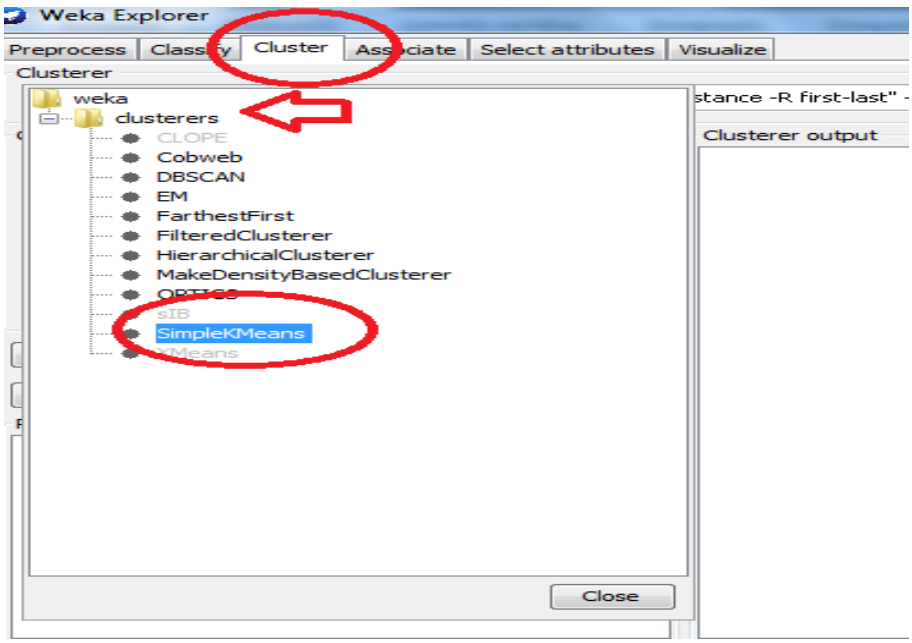
Explorer→Preprocess→Open File

και «φορτώνουμε» το αρχείο "Arkadikos Total Stats 12-13.csv", το οποίο αφορά στην ομάδα του Αρκαδικού και περιλαμβάνει τις υπό εξέταση μεταβλητές για τους παίκτες του Αρκαδικού για τη χρονική περιόδο 2012-2013.

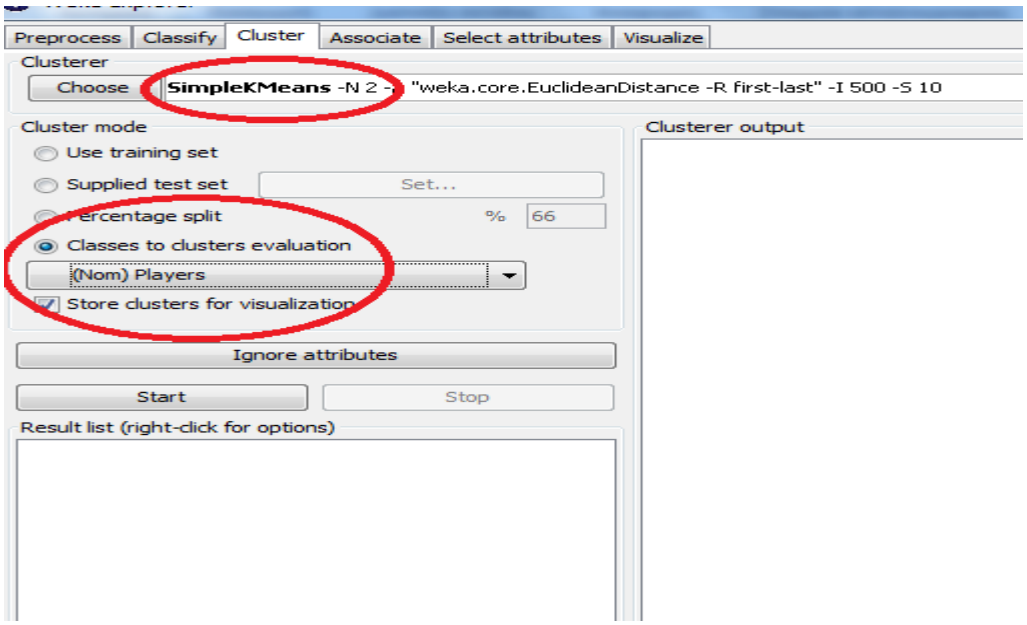
2. Αφαιρούμε τη μεταβλητή – γραμμή «Total Stats»

3. Προχωρούμε στη διαδικασία συσταδοποίησης των δεδομένων μας, ακολουθώντας τα παρακάτω βήματα:

Επιλέγουμε την καρτέλα «Cluster» και weka→clusterers→Simple KMeans



Έπειτα, επιλέγουμε «Classes to clusters evaluation» και τη μεταβλητή-στόχο **Players** ως προς την οποία θα γίνει η συσταδοποίηση, δεδομένου ότι στόχος μας είναι να τοποθετήσουμε τους παίκτες σε ομάδες, δηλαδή να βρούμε τους συνδυασμούς παικτών με τη βέλτιστη απόδοση.




```
Class attribute: Players
Classes to Clusters:

 0 1  <-- assigned to cluster
 0 1 | DJURDJEVIC Nenad
 0 1 | TSIOTRAS Thodoros
 1 0 | TSUBRILO Miroslav
 1 0 | GIZOGIANNIS Sokratis
 0 1 | GOVAS Theodosis
 0 1 | PANOU Spyros
 1 0 | SIGOUNAS Alexandros

Cluster 0 <-- TSUBRILO Miroslav
Cluster 1 <-- DJURDJEVIC Nenad
```

Εικόνα 81: Συστάδες για την ομάδα του Αρκαδικού για τη χρονική περίοδο 2012-2013

Παρατηρούμε ότι δημιουργήθηκαν δύο συστάδες, το Cluster 0 και το Cluster 1.

Το κέντρο της συστάδας «Cluster 0» αποτελεί ο παίκτης TSUBRILO Miroslav (Cluster 0 <-- TSUBRILO Miroslav).

Το κέντρο της συστάδας «Cluster 1» αποτελεί ο παίκτης DJURDJEVIC Nenad (Cluster 1 <-- DJURDJEVIC Nenad).

Η 1η ομάδα (Cluster 0) περιλαμβάνει τους παρακάτω 3 παίκτες

0 1 <-- assigned to cluster

---

1 0 | TSUBRILO Miroslav

1 0 | GIZOGIANNIS Sokratis

1 0 | SIGOUNAS Alexandros

Η 2η ομάδα (Cluster 1) περιλαμβάνει τους παρακάτω 4 παίκτες

0 **1** <-- assigned to cluster

---

0 **1** | DJURDJEVIC Nenad

0 **1** | TSIOTRAS Thodoros

0 **1** | GOVAS Theodosis

0 **1** | PANOU Spyros

Συμπερασματικά, διαπιστώνουμε ότι η βέλτιστη απόδοση της ομάδας του Αρκαδικού για τη χρονική περίοδο 2012-2013 επιτυγχάνεται με τους εξής συνδυασμούς παικτών

1. Τριάδα παικτών με τον παίκτη TSUBRILO Miroslav να αποτελεί το κέντρο της ομάδας

2. Τετράδα παικτών με τον παίκτη DJURDJEVIC Nenad να αποτελεί το κέντρο της ομάδας

Η 1η συστάδα

TSUBRILO Miroslav

GIZOGIANNIS Sokratis

SIGOUNAS Alexandros

Η 2η συστάδα

DJURDJEVIC Nenad

TSIOTRAS Thodoros

GOVAS Theodosis

PANOU Spyros

14.4.2 Εύρεση συστάδων για τον Αρκαδικό 2013-2014

1. Ακολουθούμε τα βήματα

Explorer→Preprocess→Open File

και «φορτώνουμε» το αρχείο "Arkadikos Total Stats 13-14.csv", το οποίο αφορά στην ομάδα του Αρκαδικού και περιλαμβάνει τις υπό εξέταση μεταβλητές για τους παίκτες του Αρκαδικού για τη χρονική περιόδο 2013-2014.

2. Αφαιρούμε τη μεταβλητή – γραμμή «Total Stats»

3. Προχωρούμε στη διαδικασία συσταδοποίησης των δεδομένων μας, ακολουθώντας τα παρακάτω βήματα:

Επιλέγουμε την καρτέλα «Cluster» και weka→clusterers→Simple KMeans

Έπειτα, επιλέγουμε «Classes to clusters evaluation» και τη μεταβλητή-στόχο **Players** ως προς την οποία θα γίνει η συσταδοποίηση, δεδομένου ότι στόχος μας είναι να τοποθετήσουμε τους παίκτες σε ομάδες, δηλαδή να βρούμε τους συνδυασμούς παικτών με τη βέλτιστη απόδοση.

```
Class attribute: Players
Classes to Clusters:

 0 1  <-- assigned to cluster
 1 0 | KOPSAFTIS Giorgos
 0 1 | PRODROMOU Giannis
 1 0 | POPADIC Marko
 0 1 | SAKOTA Milos
 1 0 | GOVAS Theodosis
 0 1 | GIANNOULAKOS Vasilis
 0 1 | PANOU Spyros

Cluster 0 <-- KOPSAFTIS Giorgos
Cluster 1 <-- PRODROMOU Giannis
```

Εικόνα 82 :Συστάδες για την ομάδα του Αρκαδικού για τη χρονική περίοδο 2013-2014

Παρατηρούμε ότι δημιουργήθηκαν δύο συστάδες, το Cluster 0 και το Cluster 1.

Το κέντρο της συστάδας «Cluster 0» αποτελεί ο παίκτης KOPSAFTIS Giorgos (Cluster 0 <-- KOPSAFTIS Giorgos).

Το κέντρο της συστάδας «Cluster 1» αποτελεί ο παίκτης PRODROMOU Giannis (Cluster 1 <-- PRODROMOU Giannis).

Η 1η ομάδα (Cluster 0) περιλαμβάνει τους παρακάτω 3 παίκτες

0 1 <-- assigned to cluster

---

1 0 | KOPSAFTIS Giorgos

1 0 | POPADIC Marko

1 0 | GOVAS Theodosis

Η 2η ομάδα (Cluster 1) περιλαμβάνει τους παρακάτω 4 παίκτες

0 **1** <-- assigned to cluster

---

0 **1** | PRODROMOU Giannis

0 **1** | SAKOTA Milos

0 **1** | GIANNOULAKOS Vasilis

0 **1** | PANOU Spyros

Συμπερασματικά, διαπιστώνουμε ότι η βέλτιστη απόδοση της ομάδας του Αρκαδικού για τη χρονική περίοδο 2013-2014 επιτυγχάνεται με τους εξής συνδυασμούς παικτών

1. Τριάδα παικτών με τον παίκτη KOPSAFTIS Giorgos να αποτελεί το κέντρο της ομάδας

2. Τετράδα παικτών με τον παίκτη PRODROMOU Giannis να αποτελεί το κέντρο της ομάδας

Η 1η συστάδα

KOPSAFTIS Giorgos

POPADIC Marko

GOVAS Theodosis

Η 2η συστάδα

PRODROMOU Giannis

SAKOTA Milos

GIANNOULAKOS Vasilis

PANOU Spyros

Η φιλοσοφία του αλγορίθμου K-means βασίζεται στη διάσπαση. Συγκεκριμένα, χρησιμοποιώντας μία παράμετρο *K* διαιρεί *n* στοιχεία σε *K* συστάδες με μικρή ομοιότητα μεταξύ των συστάδων, και ελαχιστοποιεί τη συνολική απόσταση της κάθε συστάδας από το κέντρο της. Ο υπολογισμός της ομοιότητας υλοποιείται από τη μέση τιμή του συνόλου κάθε συστάδας. Το μέτρο ομοιότητας που χρησιμοποιεί ο αλγόριθμος είναι η Ευκλείδεια απόσταση μεταξύ των αντικειμένων.

Οι παραπάνω συστάδες για την ομάδα του Αρκαδικού, είτε για τη χρονική περίοδο 2012-2013 είτε για τη χρονική περίοδο 2013-2014, αποτελούν συστάδες βασισμένες σε κάποιο κέντρο-παίκτη. **Επιτυγχάνεται ο βασικός σκοπός της ομαδοποίησης, ο οποίος είναι η μεγιστοποίηση της ομοιότητας μέσα στις συστάδες-ομάδες παικτών, και η ελαχιστοποίηση της ομοιότητας ανάμεσα στις συστάδες-ομάδες παικτών.** Με άλλα λόγια, μια συστάδα είναι ένα σύνολο από αντικείμενα (παίκτες) τέτοιο ώστε ένα αντικείμενό της (παίκτης) να είναι κοντινότερο στο (ή πιο όμοιο με το) «κέντρο» της συστάδας αυτής από ό,τι είναι από το κέντρο οποιασδήποτε άλλης συστάδας. Ως κέντρο της ομάδας συχνά θεωρούμε το κεντροειδές (centroid) των σημείων της συστάδας, δηλαδή το μέσο όρο ή το μεσοειδές (medoid), δηλαδή το πιο «αντιπροσωπευτικό» σημείο της συστάδας, το πιο πρωτότυπο.

15. Σύνοψη

Στην παρούσα διπλωματική εργασία, εξετάζονται ποικίλες μεθόδοι εξόρυξης δεδομένων, όπως αυτές της επιλογής χαρακτηριστικών, της ταξινόμησης, συγκεκριμένα αλγόριθμοι δέντρων απόφασης, και αλγόριθμοι ικανοί να παράγουν κανόνες ταξινόμησης, καθώς και η τεχνική της συσταδοποίησης. Οι μέθοδοι αυτές εφαρμόζονται σε πραγματικά αθλητικά δεδομένα καλαθοσφαίρισης, και αναδεικνύουν πολύπλοκες σχέσεις μεταξύ παραγόντων που μπορούν να οδηγήσουν σε βελτιωμένη παροχή υπηρεσιών στα σπορ. Γίνεται έτσι φανερή η αλληλεπίδραση των παρακάτω περιοχών:

- Στατιστική ανάλυση
- Εξόρυξη δεδομένων
- Πρακτικές εφαρμογές στα σπορ (προβλεπτική ικανότητα)

Αξίζει να σημειωθεί εδώ, ότι η εφαρμογή των τεχνικών εξόρυξης δεδομένων στις ομάδες του Αρκαδικού και του Λαυρίου, η αξιολόγηση της απόδοσης των μεθόδων, καθώς και η ανάλυση των αποτελεσμάτων έγινε κάτω από την παραδοχή ότι οι υπό εξέταση μεταβλητές αποτελούν αποκλειστικά ποσοτικά χαρακτηριστικά των παικτών, “αριθμοί”, και δεν εξετάστηκαν καθόλου ποιοτικά χαρακτηριστικά, όπως οι τραυματισμοί, η ψυχολογία, η χημεία της ομάδας κ.α., παράγοντες που διαδραματίζουν επίσης καθοριστικό ρόλο στην απόδοση των ομάδων, αλλά είναι εξαιρετικά δύσκολο να “μετρηθούν” και να αναλυθούν περαιτέρω. Επίσης, τα πραγματικά αθλητικά δεδομένα για τις ομάδες αλλά και για τους παίκτες του Αρκαδικού και του Λαυρίου ελήφθησαν από την εταιρεία Galanis Sports Data.

http://www.galanissportsdata.com/

Βιβλιογραφία

[1] Abdennadher S., Olama A., Salem N. and Thaber A., *ARM: Automatic Rule Miner*, Lecture Notes in Computer Science, Vol. 4407, 17-25, 2007.

[2] Bhandari, Inderpal, et al., *Advanced scout: Data mining and knowledge discovery in NBA data*, Data Mining and Knowledge Discovery 1.1, 121-125, 1997.

[3] Bozdogan H., *Statistical Data Mining and Knowledge Discovery*, Chapman & Hall/CRC, 2004.

[4] Breiman L., Friedman J. H., Olshen R. A. and Stone C. J., *Classification and Regression Trees*, New York: Chapman & Hall/CRC, 1984.

[5] Cao C., *Sports data mining technology used in basketball outcome prediction*, Masters Dissertation, Dublin Institute of Technology, 2012.

[6] Chen M. S., Han J. and Yu P., *Data mining: an overview from database perspective*, IEEE Transactions on Knowledge and Data Engineering, 1996.

[7] Clifton C., *Introduction to Data Mining*, University of Purdue, 2004.

[8] Cohen W.W., *Fast effective rule induction*. In Machine Learning: Proceedings of the Twelfth International Conference, Lake Tahoe, California, 1995.

[9] Dunham M.H., *Data mining, introductory and advanced Topics*, Prentice Hall, 2002.

[10] Fayyad U., *From Data Mining to Knowledge Discovery in Databases*, 1996.

[11] FIBA - Official Basketball Rules, 2008.

[12] FIBA - Basketball Statisticians' Manual, 2008.

[13] Fieltz L. and Scott D., *Prediction of Physical Performance Using Data Mining*, Research Quarterly for Exercise and Sport, 74(1), 1-25, 2003.

[14] Frank E. and Witten I.H., *Generating Accurate Rule Sets Without Global Optimization*, In: Fifteenth International Conference on Machine Learning, 144-151, 1998.

[15] Gaines B.R. and Compton P., *Induction of Ripple-Down Rules Applied to Modeling Large Databases*. J. Intell. Inf. Syst., 5(3), 211-228, 1995.

[16] Guyon I. and Elisseeff A., *An introduction to variable and feature selection*, Journal of machine learning research, 3, 1157-1182, 2003.

[17] Han J. and Kamber M., *Data mining, concepts and techniques*, second edition, Morgan Kaufmann, 2006.

[18] Hand D.J., *Data Mining: Statistics and More, Source: The American Statistician*, Vol. 52, No. 2 pp. 112-118, Published by: American Statistical Association, 1998.

[19] Hand D., Mannila H. and Smyth P., *Principles of data minin*g, The MIT press, 2001.

[20] Hanley J.A. and McNeil B.J., *The meaning and use of the area under a receiver operating characteristic (ROC) curve*, Radiology, 143, 29-36, 1982.
[21] Hartigan J.A., *Clustering Algorithms*, 1975.

[22] Hastie T., Tibshirani R. and Friedman J., *The elements of statistical earning, Data mining, Inference and Prediction*, Springer, 2001.

[23] Holte R.C., *Very simple classification rules perform well on most commonly used datasets*, Machine Learning, 11, 63-91, 1993.

[24] Hsu W.H., *Overview of Data Mining and Knowledge Discovery in Databases (KDD)*, Department of Computing and Information Sciences, Kansas State University, 2003.

[25] Kettenring J.R., *A perspective on cluster analysis*, Wiley InterScience 2008.

[26] Kohavi R., *The Power of Decision Tables*, In: 8th European Conference on Machine Learning, 174-189, 1995.

[27] Larose D., *Discovering Knowledge in Data: An Introduction to Data Mining*, Wiley, 2005.

[28] Larose D., *Data mining, methods and models*, Wiley, 2006.

[29] Lyons K., *Data Mining and Knowledge Discovery*, Australian Sports Commission Journals, 2(4), 2005.

[30] Marchi L.D., *Data mining of sports performance datα*, Erasmus computing, 2010/2011.

[31] Miljkovic, Dejan, et al., *The use of data mining for basketball matches outcomes prediction*, Intelligent Systems and Informatics (SISY), 2010 8th International Symposium on. IEEE, 2010.

[32] Mitchell T.M., *Machine Learning*, McGrawn-Hill, Science/Engineering/Math, 1997.

[33] Oliver D., *Basketball on Paper - Rules and Tools for Performance Analysis*, Washington DC, 2004.

[34] O'Reilly N. and Knight P., *Knowledge Management Best Practices in National Sport Organizations*, International Journal of Sport Management and Marketing, 2(3), 264-280, 2007.

[35] Pepe M.S., *Receiver operating characteristic methodology*, Journal of the American Statistical Association, 95, 308-311, 2000.

[36] Quinlan J.R., *Induction of Decision Trees*, Machine Learning, 1, 81-106, 1986.

[37] Quinlan J.R., *C4.5:Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[38] Remco R., Bouckaert R.R., Frank E., Hall M., Kirkby R., Reutemann P., Seewald A. and Scuse D., *WEKA Manual for Version 3-6-2*, 2010.

[39] Schumaker R., Solieman O. and Chen H., *Sports Data Mining*, Springer, 2010.

[40] Solieman O.K., *Data Mining in Sports:A Research Overview*, MIS Masters Project, August 2006.

[41] Soukup T. and Davidson I., *Visual data mining- techniques and tools for data visualization and mining*, Wiley J., 2002.

[42] Stefani R., *A Taxonomy of Sports Rating Systems*, IEEE Transactions on Systems, Man, and Cybernetics - Part A 29(1), 116-120, 1999.

[43] Tryon R.C., *Cluster analysis*, New York, McGraw-Hill, 1939.

[44] Witten I.H. and Frank E., *Data Mining: Practical Machine learning Tools and Techniques with Java Implementations*, 2nd edn., Morgan Kaufmann Publishers, San Francisco, 2005.