ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

UNIVERSITY OF PIRAEUS

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Πρόγραμμα Μεταπτυχιακών Σπουδών

«Πληροφοριακά Συστήματα & Υπηρεσίες»

Μεγάλα Δεδομένα και Αναλυτική

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Αναλυτική δεδομένων αγώνων καλαθοσφαίρισης για την πρόβλεψη αποτελεσμάτων και εξαγωγή γνώσης»

Γαρδέλης Σταύρος

ΜΕ1922

Επιβλέπων : Ηλίας Μαγκλογιάννης

Ιούνιος 2021

Περίληψη

Η παρούσα Διπλωματική εργασία ασχολείται με την Αναλυτική αθλητικών δεδομένων (Sports Analytics). Όπως σε πολλούς άλλους κλάδους, έτσι και στον αθλητισμό ο ρυθμός συλλογής δεδομένων αυξάνεται συνεχώς τα τελευταία χρόνια. Χρησιμοποιήθηκαν δεδομένα απο το άθλημα της καλαθοσφαίρησης (Basketball) .

Πιο συγκεκριμένα αναλύθηκαν 2 σύνολα δεδομένων (dataset) απο τα δύο μεγαλύτερα πρωταθλήματα του κόσμου, του NBA και της Euroleague. Κάθε dataset περιέχει στατιστικά δεδομένα ομάδων για τις χρονιές 2005-06 έως και 2018-19. Εκτός απο τα πολλά στατιστικά που αφορούν δεδομένα ενός αγώνα όπως πόντοι , ασιστ , ριμπάουντ, κλεψίματα κ.α. δίνεται και η επιτυχία εισαγωγής ή όχι στα playoff του αντίστοιχου πρωταθλήματος. Δεδομένο με το οποίο θα ασχοληθούμε, μιας και είναι η κλάση του dataset.

Χρησιμοποιήθηκαν οι παρακάτω αλγόριθμοι επιβλεπόμενης μηχανική μάθησης (Supervised Learning) , ώστε να παρουσιαστεί πιο μοντέλο μπορεί να κάνει αποδοτικότερη πρόβλεψη ανα πρωτάθλημα. Οι αλγόριθμοι ειναι Logistic Regression, k-nearest neighbors (KNN), Support Vector Machine (SVM), Random Forest, Νευρωνικό δίκτυο(Multi-layer Perceptron) . Χρησιμοποιήθηκαν σε 3 διαφορετικά σενάρια που αφόρούσαν επεξεργασμένα δεδομένα με διάφορες τεχνικές όπως standardization, Pearson correlation και στατιστικά μεσης τιμής .

Απότερος στόχος ήταν να κάνουμε πρόβλεψη της τελευταίας 5ετίας με τους παραπανω κατηγοριοποιητές , βάσει των προηγούμενων χρόνων για κάθε σύνολο δεδομένων.

Ως τελευταίο κομμάτι αυτής της διπλωματικής ήταν να χρησιμοποιήσουμε το καλύτερο μοντέλο ανα πρωτάθλημα και να το εφαρμόσουμε (fit) στο αντίθετο πρωτάθλημα.

Λέξεις κλειδιά : Euroleague , NBA , Μηχανική Μάθηση , Multi-layer Perceptron , k-nearest neighbors , Logistic Regression, Support Vector Machine , Pearson Correlation, Random Forest, κατηγοριοποίηση, πρόβλέψη μπάσκετ , Αναλυτική αθλητικών δεδομένων.

Abstract

The present thesis deals with Sports Analytics. As in many other domains, also in sports the rate of data collection has been steadily increasing in recent years. Basketball Data were used.

More specifically, two datasets from the two largest leagues in the world, the NBA and the Euroleague, were analyzed. Each dataset contains statistics for the seasons 2005-06 up to 2018-19. Apart from the many statistics that concern data of a match such as points, assists, rebounds, steals, etc. is given the success of admission or not in playoffs of the respective league. Attribute that we will deal with, since it is the class of each dataset.

The following Supervised Machine Learning algorithms were used to present a model that can make more efficient predictions per league. The algorithms are Logistic Regression, k-nearest neighbors (KNN), Support Vector Machine (SVM), Random Forest, Neural Network (Multi-layer Perceptron). They were used in 3 different scenarios involving processed data with different techniques such as standardization, Pearson correlation and average statistics.

Another goal was to make a forecast of the last 5 years with the above classifiers, based on previous years for each dataset.

As the last part of this thesis was to use the best model/classifier per league/dataset and fit it to the opposite league.

Keywords: Euroleague, NBA, machine learning, Multi-layer Perceptron, k-nearest neighbors, Logistic Regression, Support Vector Machine, Pearson Correlation, Random Forest, classification, basketball forecasting, sports analytics.

Λίστα Περιεχομένων

Λίστα Εικόνων

Λίστα Πινάκων

1.1. Ορισμός Προβλήματος – Θεματική περιοχή

Στην εισαγωγή αναφέρθηκε οτι το αντικείμενο της διπλωματικής εργασίας είναι τα sports analytics και η μηχανική μάθηση. Πιο συγκεκριμένα αναλύθηκαν 2 σύνολα δεδομένων και εφαρμόσθηκαν αλγόριθμοι επιβλεπόμενης μηχανικής μάθησης με απώτερο σκοπό να βρούμε ποιος αλγόριθμος είναι ο καταλληλότερος για τον τύπο των συνόλων δεδομένων που έχουμε. Επίσης πραγματοποιήθηκε η εφαρμογή των καλύτερων/καταλληλότερων αλγορίθμων τους ενός συνόλου στο άλλο, προσπαθώντας να κάνουμε μία σύγκριση μεταξύ των παρόμοιων αθλητικών συνόλων δεδομένων, αλλά από τελείως διαφορετικά πρωταθλήματα και ηπείρους.

1.2. Δομή Εργασίας

Η δομή της συγκεκριμένης μελέτης αποτελείται από τέσσερα (4) κεφάλαια. Στο 1ο κεφάλαιο παρουσιάζεται η εισαγωγή/περίληψη της εργασίας, μαζί με τον ορισμό του προβλήματος και τη θεματική περιοχή με στην οποία διαπραγματεύεται. Στο 2^{ο} κεφάλαιο παρουσιάζεται η μηχανική μάθηση γενικότερα με αναφορά στην μάθηση με ή χωρίς επίβλεψη. Έπειτα γίνεται αναφορά σχετικά με τα sports analytics και πιο συγκεκριμένα με τα basketball analytics που είναι και το θέμα της συγκεκριμένης εργασίας. Στο 3^{ο} κεφάλαιο αναλύεται στη θεωρία οι τεχνικές που χρησιμοποιήθηκαν όσο αναφορά τη μηχανική μάθηση με επίβλεψη. Επίσης παρουσιάζονται και οι μετρικές αξιολόγησης κατηγοριοποιητών, καθώς και οι τεχνικές προ επεξεργασίας δεδομένων. Στο 4^{ο} και τελευταίο κεφάλαιο παρουσιάζεται το πειραματικό κομμάτι της εργασίας με τα αντίστοιχα αποτελέσματα. Παρουσιάζεται η περιγραφή των συνόλων δεδομένων και τα σενάρια που χρησιμοποιήθηκαν, με αναφορές σε βασικές εντολές python που είναι και η γλώσσα προγραμματισμού που χρησιμοποιήθηκε.

2. Βιβλιογραφική επισκόπηση και Τεχνολογικό Υπόβαθρο

Σε αυτό το κεφάλαιο της διπλωματικής γίνεται η βιβλιογραφική επισκόπηση της εργασίας και αναλύονται σημαντικές έννοιες της θεωρίας.

2.1. Μέθοδοι και Αλγόριθμοι Γενικά Είδη Μηχανικής Μάθησης

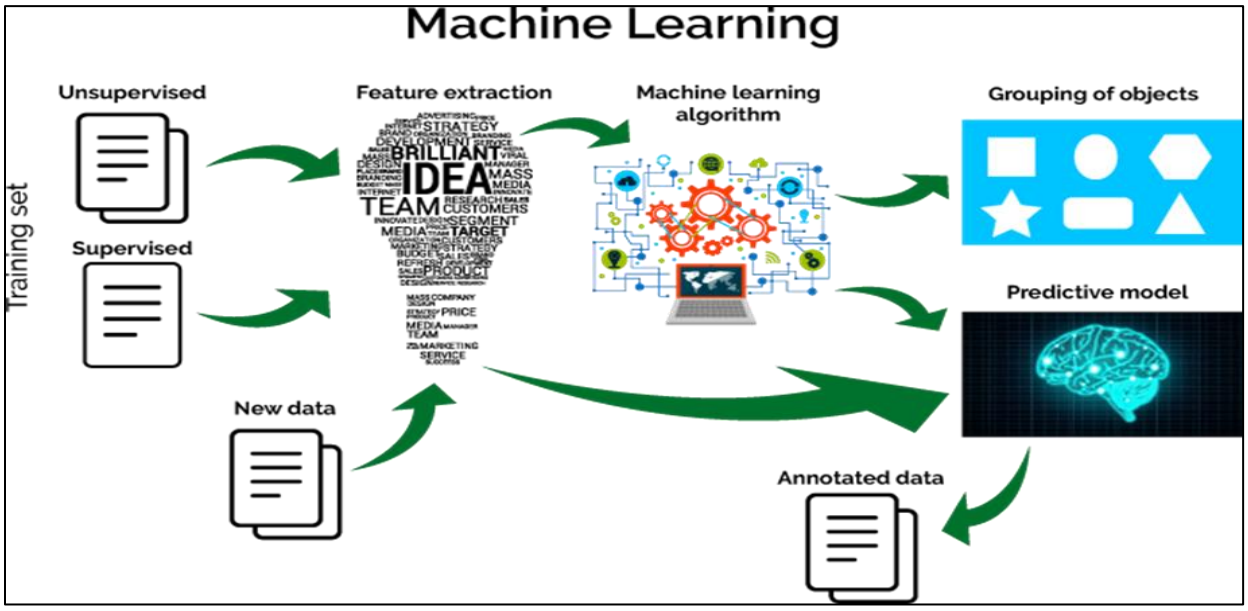
Η Μηχανική Μάθηση (Machine Learning) και η Τεχνητή Νοημοσύνη (Artificial Intelligence) συγκαταλέγονται στους κορυφαίους όρους για τις επιστήμες σήμερα. Η Μηχανική Μάθηση είναι μια εφαρμογή τεχνητής νοημοσύνης που χρησιμοποιείται ως εργαλείο για τη μετατροπή των πληροφοριών σε γνώση. (Burkov, 2020) Ουσιαστικά, είναι ένα σύνολο προτύπων, εργαλείων, διαδικασιών και μεθοδολογίας που στοχεύει να ελαχιστοποιήσει τις πιθανότητες εγκατάλειψης, λανθασμένης ή άσχετης εργασίας που γίνεται σε μια προσπάθεια επίλυσης ενός προβλήματος ή μίας ανάγκης. (Wilson, 2021) Σε αντίθεση όμως, με τις εφαρμογές τεχνητής νοημοσύνης, η μηχανική μάθηση περιλαμβάνει την εκμάθηση κρυφών μοτίβων στα δεδομένα (εξόρυξη δεδομένων) και στη συνέχεια τη χρήση των προτύπων για την ταξινόμηση ή την πρόβλεψη ενός συμβάντος που σχετίζεται με το πρόβλημα. (Berry, Mohamed & Yap, 2020) Η Μηχανική Μάθηση επιτρέπει στους υπολογιστές να μαθαίνουν και να λειτουργούν αυτόματα χωρίς ανθρώπινη βοήθεια, δημιουργώντας πιο αποδοτικές και αποτελεσματικές διαδικασίες. Ένα άτομο μπορεί να χρησιμοποιήσει τη Μηχανική Μάθηση για να δημιουργήσει λύσεις βάσει προγραμμάτων οδήγησης δεδομένων και να κάνει προβλέψεις που βοηθούν στην λήψη σύνθετων αποφάσεων. (Lakshmanan, Robinson & Munn, 2020)

Η Μηχανική Μάθηση έχει βρει εφαρμογή σε πολλούς τομείς, μερικοί από αυτούς περιλαμβάνουν τις επιχειρήσεις, τα χρηματοοικονομικά, την υγειονομική περίθαλψη, τον αθλητισμό, είτε ομαδικό, είτε ατομικό, κ.α.

Κατά τα τελευταία 20 χρόνια, παράλληλα με την αυξανόμενη ποσότητα δεδομένων που είναι διαθέσιμα για ανάλυση, αναπτύχθηκε μια ποικιλία διαφορετικών τεχνικών Μηχανικής Μάθησης. Η Μηχανική Μάθηση μπορεί να χωριστεί σε Μάθηση με Επίβλεψη (Supervised Learning) και σε Μάθηση Χωρίς Επίβλεψη (Unsupervised Learning) αν και ορισμένοι συγγραφείς ταξινομούν επίσης άλλους αλγόριθμους ως ενίσχυση, επειδή τέτοιες τεχνικές μαθαίνουν δεδομένα και προσδιορίζουν μοτίβα για σκοπούς αντίδρασης σε ένα περιβάλλον.

Ωστόσο, τα περισσότερα άρθρα αναγνωρίζουν αλγόριθμους μηχανικής μάθησης με επίβλεψη και χωρίς επίβλεψη. Η διαφορά μεταξύ αυτών των δύο κατηγοριών είναι η ύπαρξη ετικετών στο υποσύνολο των

δεδομένων εκπαίδευσης και αναλύονται με περισσότερη λεπτομέρεια στα επόμενα κεφάλαια. Στην Εικόνα 1 απεικονίζεται ένα μοντέλο Μηχανικής Μάθησης.



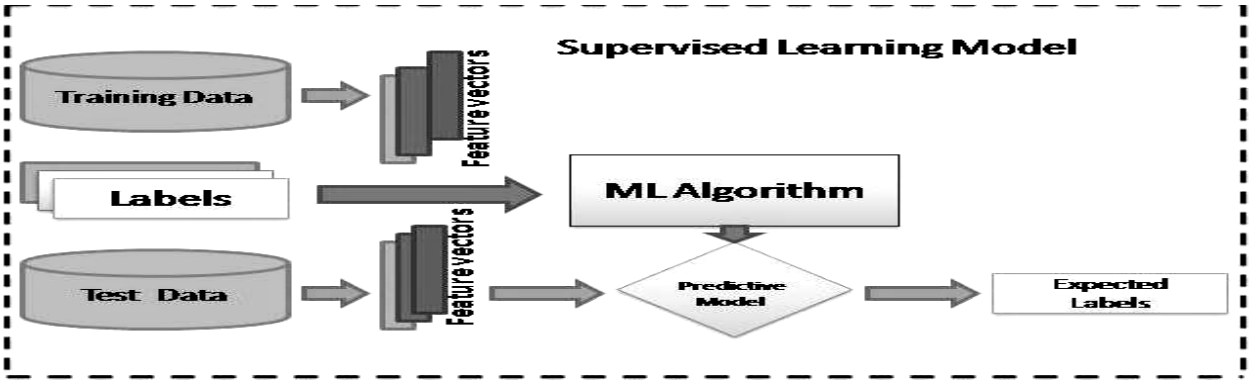
Εικόνα 1: Machine Learning (Πηγή: Khadka, 2017)

2.1.1. Μάθηση με Επίβλεψη (Supervised Learning)

Σύμφωνα με τον Κοτσιάντη, (2007) στη μάθηση με επίβλεψη ή αλλιώς μάθηση με παραδείγματα ένα σύστημα πρέπει να μάθει ένα χαρακτηριστικό, που μπορεί να είναι είτε μια έννοια, είτε μια συνάρτηση, μέσα από ένα σύνολο δεδομένων και η οποία περιγράφει ένα μοντέλο.

Στην μάθηση με επίβλεψη, ο αναλυτής δεδομένων δουλεύει με μια συλλογή με παραδείγματα με ετικέτες $\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$. Κάθε στοιχείο x_i του N ονομάζεται διάνυσμα χαρακτηριστικών. Στην επιστήμη των υπολογιστών, ένα διάνυσμα είναι ένας πίνακας μίας διάστασης. Ένας τέτοιος πίνακας, με τη σειρά του, είναι μια ακολουθία τιμών. Το μήκος αυτής της ακολουθίας τιμών, D, ονομάζεται διάσταση του διανύσματος. (Chourasiya & Jain, 2019)

Η διαδικασία εφαρμογής της μάθησης με επίβλεψη σε πραγματικό πρόβλημα περιγράφεται στο παρακάτω σχήμα.



Εικόνα 2 Supervised Machine Learning (Πηγή: Muhammad & Yan, 2015)

Στην εποπτευόμενη μάθηση, το πρώτο βήμα που πρέπει κάποιος να ασχοληθεί είναι με το σύνολο δεδομένων. Προκειμένου να επιτευχθεί καλύτερη εκπαίδευση σε σύνολο δεδομένων, ένας πιο έμπειρος αναλυτής θα μπορούσε να προτείνει καλύτερη επιλογή χαρακτηριστικών. Η άλλη προσέγγιση είναι η «ωμή δύναμη», που σημαίνει τη μέτρηση όλων των διαθέσιμων χαρακτηριστικών με την ελπίδα ότι μπορούν να απομονωθούν τα σωστά χαρακτηριστικά. Ωστόσο, ένα σύνολο δεδομένων που συλλέγεται με αυτή τη μέθοδο δεν είναι άμεσα κατάλληλο για επαγωγή γιατί στις περισσότερες περιπτώσεις περιέχει θόρυβο και λείπουν τιμές χαρακτηριστικών και συνεπώς απαιτεί σημαντική προ-επεξεργασία. (Kotsiantis, 2007) Στο επόμενο βήμα, η προετοιμασία δεδομένων και η προεπεξεργασία δεδομένων αποτελούν βασική λειτουργία του ερευνητή στην Μηχανική Μάθηση με Επίβλεψη (SML). Έχουν εισαχθεί αρκετές τεχνικές από διαφορετικούς ερευνητές για την αντιμετώπιση του ζητήματος των χαμένων δεδομένων. Για παράδειγμα, οι Hodge & Austin, (2004) διεξήγαγαν μια έρευνα σύγχρονων τεχνικών για την ανίχνευση εξωτερικών, ακραίων δεδομένων (θορύβου). Οι Karanjit & Shuchita, (2012) έχουν επίσης συζητήσει διαφορετικές μεθόδους ανίχνευσης εξωτερικών, ακραίων δεδομένων που χρησιμοποιούνται στη μηχανική μάθηση.

Η επιλογή ενός αλγορίθμου για την επίτευξη καλών αποτελεσμάτων είναι ένα σημαντικό βήμα. Η αξιολόγηση του αλγορίθμου κρίνεται ως επί το πλείστων από την ακρίβεια των προβλέψεων. Οι αλγόριθμοι προσπαθούν να προβλέψουν και να ταξινομήσουν το προκαθορισμένο χαρακτηριστικό, και η ακρίβεια τους και η εσφαλμένη ταξινόμηση τους μαζί με άλλα μέτρα απόδοσης εξαρτώνται από τις μετρήσεις του προκαθορισμένου χαρακτηριστικού που έχουν προβλεφθεί σωστά ή ταξινομηθεί ή άλλως. Είναι επίσης σημαντικό να σημειωθεί ότι η διαδικασία εκμάθησης σταματά όταν ο αλγόριθμος επιτυγχάνει ένα αποδεκτό επίπεδο απόδοσης. (Muhammad & Yan, 2015)

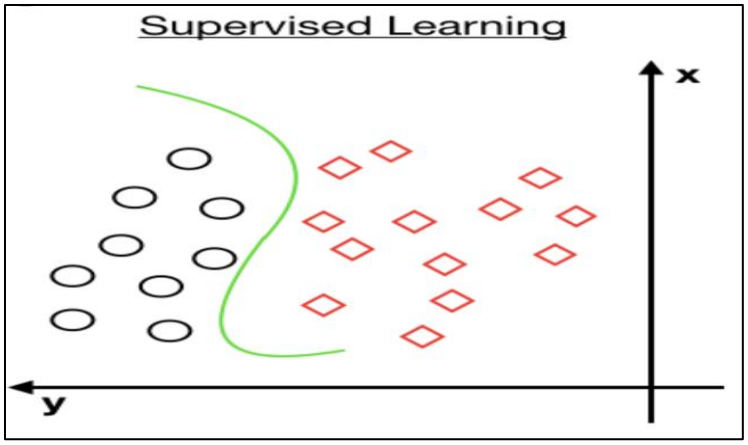
Ο στόχος ενός αλγορίθμου εποπτευόμενης μάθησης είναι η χρήση ενός συνόλου δεδομένων για την παραγωγή ενός μοντέλου που παίρνει ένα σύνολο χαρακτηριστικών X ως πληροφορίες εισόδου και εξόδου που επιτρέπει την εξαγωγή μιας ετικέτας για αυτό το σύνολο χαρακτηριστικών. Για παράδειγμα, ένα μοντέλο που δημιουργήθηκε χρησιμοποιώντας ένα σύνολο δεδομένων ασθενών θα μπορούσε να λάβει ως είσοδο έναν φορέα χαρακτηριστικών που περιγράφει έναν ασθενή και να αποδώσει μια πιθανότητα ότι ο ασθενής έχει καρκίνο.

Ακόμα κι αν το μοντέλο είναι συνήθως μια μαθηματική συνάρτηση, είναι βολικό κάποιος να σκεφτεί ότι το μοντέλο «κοιτάζει» τις τιμές ορισμένων χαρακτηριστικών στην είσοδο και, με βάση την εμπειρία με παρόμοια παραδείγματα, εξάγει μια τιμή. Αυτή η τιμή εξόδου είναι ένας αριθμός ή «κάτι παρόμοιο» με τις ετικέτες που έχουν δει στο παρελθόν στα παραδείγματα με παρόμοιες τιμές χαρακτηριστικών. (Burkov, 2020)

Λαμβάνοντας υπόψη την προσέγγιση που χρησιμοποιείται στη μηχανική μάθηση, έχει παρατηρηθεί ότι ένα υποσύνολο εκπαίδευσης είναι περίπου κατά 66% λογικό και βοηθά στην επίτευξη του επιθυμητού αποτελέσματος χωρίς να απαιτείται περισσότερος υπολογιστικός χρόνος. (Ng, 2012) Οι αλγόριθμοι μάθησης ταξινομούνται περαιτέρω σε αλγόριθμους ταξινόμησης και παλινδρόμησης. (Kotsiantis, 2007)

Στη μηχανική μάθηση με επίβλεψη, το πρόβλημα της πρόβλεψης μιας ομάδας (τάξης) ονομάζεται ταξινόμηση, ενώ το πρόβλημα της πρόβλεψης ενός πραγματικού αριθμού ονομάζεται παλινδρόμηση. Η τιμή που πρέπει να προβλεφθεί από ένα εποπτευόμενο μοντέλο ονομάζεται στόχος. Ένα παράδειγμα παλινδρόμησης είναι ένα πρόβλημα πρόβλεψης του μισθού ενός υπαλλήλου δεδομένης της εργασιακής εμπειρίας και των γνώσεων του. Ένα παράδειγμα ταξινόμησης είναι όταν ένας γιατρός εισάγει τα χαρακτηριστικά ενός ασθενούς σε μια εφαρμογή λογισμικού και η εφαρμογή επιστρέφει τη διάγνωση. (Burkov, 2020)

Οι πιο συχνά χρησιμοποιούμενοι αλγόριθμοι στην μηχανική μάθηση με επίβλεψη είναι οι k-Nearest Neighbor, τα Decision Trees, το Naïve Bayes, η Linear Regression, τα Support Vector Machine (SVM), και τα Neural Networks.



Εικόνα 3 Supervised Machine Learning με χρήση Linear Regression

2.1.2. Μάθηση Χωρίς Επίβλεψη (Unsupervised Learning)

Στην μάθηση χωρίς επίβλεψη η μηχανή λαμβάνει απλώς εισροές αλλά δεν λαμβάνει ούτε εποπτευόμενα αποτελέσματα-στόχους ούτε ανταμοιβές από το περιβάλλον της. Είναι περίεργο να φανταστεί κανείς τι θα μπορούσε να μάθει το μηχάνημα, δεδομένου ότι δεν λαμβάνει σχόλια από το περιβάλλον του. Ωστόσο, είναι δυνατό να αναπτυχθεί ένα μοντέλο με βάση την ιδέα ότι ο στόχος του μηχανήματος είναι να δημιουργήσει αναπαραστάσεις των δεδομένων εισόδου που μπορούν να χρησιμοποιηθούν για τη λήψη αποφάσεων, την πρόβλεψη μελλοντικών εισροών, κ.α. (Ghahramani, 2003)

Σύμφωνα με τον Hofmann (2001), οι αλγόριθμοι στη μάθηση χωρίς επίβλεψη είναι κατάλληλοι για τη δημιουργία ετικετών στα δεδομένα που στη συνέχεια χρησιμοποιούνται για την εφαρμογή εποπτευόμενων μαθησιακών εργασιών. Δηλαδή, οι αλγόριθμοι στη μάθηση χωρίς επίβλεψη προσδιορίζουν τις εγγενείς ομαδοποιήσεις εντός των δεδομένων χωρίς ετικέτα και στη συνέχεια εκχωρούν ετικέτα σε κάθε τιμή δεδομένων. (Marshland, 2015)

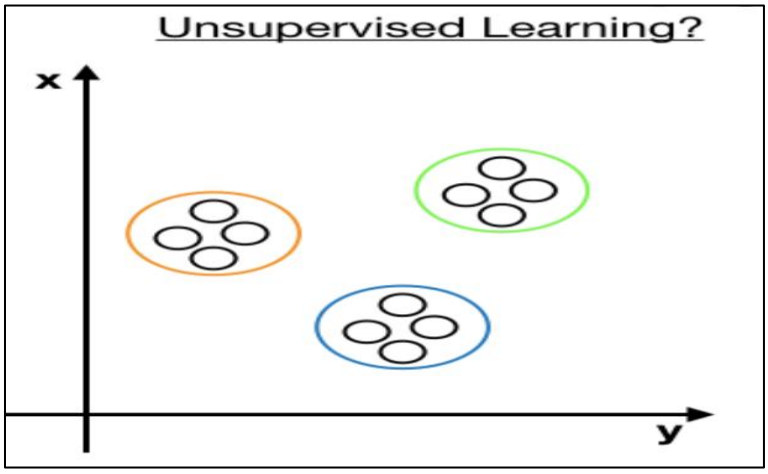
Στην μάθηση χωρίς επίβλεψη, το σύνολο δεδομένων είναι μια συλλογή από παρατηρήσεις χωρίς ετικέτα $\{x_1, x_2, \ldots, x_N\}$. Το x είναι ένα διάνυσμα χαρακτηριστικών, και ο αλγορίθμου μάθησης χωρίς επίβλεψη έχει ως στόχο να φτιάξει ένα μοντέλο που παίρνει ένα διάνυσμα χαρακτηριστικών x ως είσοδο και είτε μπορεί να το μετατρέψει σε άλλο φορέα είτε να το μετατρέψει σε μια τιμή που θα μπορούσε να χρησιμοποιηθεί με τέτοιο τρόπο που να συνεισφέρει στην λύση ενός πιο πρακτικού ζητήματος. (Chourasiya & Jain, 2019) Για παράδειγμα, στην ομαδοποίηση, το μοντέλο επιστρέφει το ID της ομάδας για κάθε διάνυσμα δυνατοτήτων στο σύνολο δεδομένων. Η ομαδοποίηση είναι χρήσιμη για την εύρεση ομάδων παρόμοιων αντικειμένων σε μια μεγάλη συλλογή αντικειμένων, όπως εικόνες ή έγγραφα

κειμένου. (Burkov, 2020) Με τη χρήση της ομαδοποίησης, για παράδειγμα, ο αναλυτής μπορεί να έχει ένα μικρό δείγμα αλλά αρκετά αντιπροσωπευτικό μη επισημασμένων παραδειγμάτων από μια μεγάλη συλλογή παραδειγμάτων και να τα επισημάνει μεμονωμένα. (Chourasiya & Jain, 2019)

Ένα τέτοιο μοντέλο μπορεί να χρησιμοποιηθεί για την ανίχνευση ή την παρακολούθηση των ακραίων τιμών. Για παράδειγμα, το x να αντιπροσωπεύει μοτίβα ανάγνωσης από τους αισθητήρες από μια μονάδα παραγωγής ενέργειας. Γίνεται η υπόθεση ότι το P(x) μαθαίνει από τα δεδομένα που συλλέγονται από μια κανονικά λειτουργούσα μονάδα. Αυτό το μοντέλο μπορεί να χρησιμοποιηθεί για την αξιολόγηση της πιθανότητας μιας νέας ανάγνωσης από τους αισθητήρες. Εάν αυτή η πιθανότητα είναι ασυνήθιστα χαμηλή, τότε είτε το μοντέλο είναι φτωχό είτε το εργοστάσιο συμπεριφέρεται ασυνήθιστα, οπότε κάποιος μπορεί να θέλει να το κλείσει. (Ghahramani, 2004)

Ένα μοντέλο πιθανοτήτων μπορεί επίσης να χρησιμοποιηθεί για ταξινόμηση. Ένα μοντέλο πιθανοτήτων μπορεί επίσης να χρησιμοποιηθεί για ταξινόμηση. Υποθέτοντας, ότι το P_1 (x) είναι ένα μοντέλο που απεικονίζει τα χαρακτηριστικά των κατόχων πιστωτικών καρτών που πληρώνουν εγκαίρως και το P_2 (x) είναι ένα μοντέλο που απεικονίζει τους κατόχους πιστωτικών καρτών που αθέτησαν τις πληρωμές τους, τότε αξιολογώντας τις σχετικές πιθανότητες P_1 (x') και P_2 (x') σε έναν νέο αιτούντα x', το μηχάνημα μπορεί να αποφασίσει να τον ταξινομήσει σε μία από αυτές τις δύο κατηγορίες. (Ghahramani, 2004)

Οι πιο συχνά χρησιμοποιούμενοι αλγόριθμοι στην μηχανική μάθηση χωρίς επίβλεψη είναι οι k-means clustering και τα Association Rules.



Εικόνα 4 Unsupervised Machine Learning

2.2. Sport Analytics – Σχετικές Εργασίες

Τα Sports Analytics είναι η διερεύνηση και μοντελοποίηση επαγγελματικών αθλητικών επιδόσεων και διαγωνισμών χρησιμοποιώντας επιστημονικές τεχνικές. Αυτό το πεδίο χρησιμοποιεί συχνά αρχές και τεχνικές από την στατιστική, την εξόρυξη δεδομένων, την θεωρία παιγνίων, κ.α. Το μπάσκετ είναι ένα από τα καλύτερα παραδείγματα ομαδικού αθλήματος. Παρόλο που η επιτυχία μιας ομάδας ορίζεται συνήθως ως προς τις νίκες και τις ήττες της. Αυτές οι νίκες και οι ήττες είναι τα αποτελέσματα των πολύπλοκων αλληλεπιδράσεων μεταξύ των επιδόσεων των μεμονωμένων παικτών των δύο ομάδων και των αβεβαιοτήτων που σχετίζονται με το εκάστοτε αθλητικό γεγονός. (Singh, 2020)

Η ανάλυση των αθλητικών επιδόσεων επιτρέπει στον προπονητή, τους παίκτες και τους προπονητές να αξιολογήσουν αντικειμενικά και να βελτιώσουν έτσι την αθλητική τους απόδοση. Η χρήση υπολογιστικών και στατιστικών προσεγγίσεων στα αθλητική ανάλυση κερδίζει δημοτικότητα ειδικά σε σχέση με τη σύγκριση απόδοσης, την οπτικοποίηση και την πρόβλεψη του αποτελέσματος του αγώνα.

Τα επίσημα παιχνίδια ταξινομούνται σε τέσσερις μεγάλες κατηγορίες. Η πρώτη είναι τα παιχνίδια που χαρακτηρίζονται από την Εισβολή (Invasion), η δεύτερη από αυτά που χαρακτηρίζονται από Δίχτυ/ Τείχος (Net/ Wall), η τρίτη κατηγορία είναι αυτά που έχουν κάποιου είδος Χτύπημα (Striking/ Fielding), και η τέταρτη κατηγορία αυτά που έχουν Στόχο (Target Games). (Ellis, 1983). Για παράδειγμα το Μπάσκετ είναι ένα παιχνίδι Invasion, ενώ το μπέιζμπολ είναι ένα παιχνίδι Striking/ Fielding. Σε αντίθεση με άλλες μορφές παιχνιδιών, τα παιχνίδια εισβολής χαρακτηρίζονται από τη χρήση ενός στόχου για σκορ, κοινών τακτικών εισβολής στην περιοχή του αμυνόμενου (στην περίπτωση του μπάσκετ από τα 7,5 μέτρα και προς το καλάθι του αντιπάλου) για να κάνει χώρο στην επίθεση, και τον περιορισμό του χώρου του όταν βρίσκεται σε άμυνα. Η συνεχής αντίθεση και η δυναμική δομή καθιστούν τα παιχνίδια εισβολής όπως το μπάσκετ και το ποδόσφαιρο, πιο περίπλοκα από άλλα είδη παιχνιδιών, και επομένως δυσκολότερο να αναλυθούν επαρκώς. Οποιαδήποτε ανάλυση απόδοσης στα παιχνίδια εισβολής θα πρέπει επομένως να δομηθεί με τη βοήθεια ενός συστήματος ανάλυσης με εξελιγμένα μηχανογραφημένα πληροφοριακά συστήματα.

2.2.1. Machine Learning in Sports

Σε αυτήν την ενότητα, εξετάζετε ένα μέρος της βιβλιογραφίας στον τομέα της Μηχανικής Μάθησης που σχετίζεται με τον αθλητισμό.

Τα Sports Analytics περιλαμβάνει μοντελοποίηση των αθλητικών γεγονότων βάσει δεδομένων, συμπεριλαμβανομένης της διαχείρισης της φυσικής απόδοσης των αθλητών, της κατανόησης των στρατηγικών παιχνιδιού και των τακτικών που αναπτύσσουν οι ομάδες. Σε πολλές περιπτώσεις για την συλλογή δεδομένων χρησιμοποιούνται φορητοί αισθητήρες που μετρούν τις κινήσεις των παικτών, το φυσικό φορτίο και την φυσική καταπόνηση κατά τη διάρκεια συγκρούσεων, σε συνδυασμό με κάμερες πολλαπλών προβολών που συλλαμβάνουν ολόκληρο το γήπεδο και χρησιμοποιούνται συνήθως για την παρακολούθηση των παικτών και των κινήσεων της μπάλας σε επαγγελματικά ομαδικά αθλήματα. Η ανάλυση των δεδομένων στις ομάδες μπάσκετ για την απόκτηση ανταγωνιστικού πλεονεκτήματος παρουσιάζει ενδιαφέρον για όλους τους συλλόγους και συνδέεται με την οικονομική τους επιτυχία. Η ποσοτική ανάλυση του αθλητισμού, ιδίως του μπάσκετ, είναι ένας κλάδος της επιστήμης, ο οποίος αναπτύχθηκε αρχικά μέσω μη ακαδημαϊκών εργασιών και έχει λάβει μεγάλο ακαδημαϊκό ενδιαφέρον την τελευταία δεκαετία. (Demenius & Kreivyte, 2017)

Ξεκινώντας από τους Albert, Bennett και Cochran, (2005) συνέλλεξαν άρθρα που είχαν ήδη δημοσιευτεί σχετικά με τη χρήση στατιστικών για την ανάλυση του αθλητισμού. Το βιβλίο τους περιέχει ξεχωριστές ενότητες αφιερωμένες στα μεγάλα ομαδικά αθλήματα, δηλαδή μπέιζμπολ, ποδόσφαιρο, μπάσκετ και χόκεϊ επί πάγου. Ο Winston (2009), που αφιερώνει ένα μεγάλο μέρος του βιβλίου του στο μπάσκετ, εισαγάγει τον αναγνώστη σε διάφορα μοντέλα που χρησιμοποιούνται για την ανάλυση του αθλητισμού. Παρατηρεί ότι ορισμένα από αυτά τα μοντέλα χρησιμοποιούνται στην πράξη από τη διοίκηση της ομάδας για να βοηθήσουν στη διαδικασία λήψης αποφάσεων. Ο Berri και ο Schmidt (2010) εμπνέονται από το γεγονός ότι οι άνθρωποι αντιμετωπίζουν γενικά προβλήματα στη λήψη σωστών αποφάσεων. Με βάση ερευνητικά στοιχεία, παρουσιάζουν διάφορες ιστορίες οι οποίες, όπως ισχυρίζονται, δεν πρέπει μόνο να αλλάξουν τον τρόπο με τον οποίο οι φίλαθλοι βλέπουν τις επιλογές των αγαπημένων τους ομάδων, αλλά επίσης επηρεάζουν τον τρόπο με τον οποίο οι οικονομολόγοι και άλλοι κοινωνικοί επιστήμονες σκέφτονται τη λήψη αποφάσεων από τον άνθρωπο.

2.2.2. Predictive in Sports

Σε αυτήν την υποενότητα παρουσιάζονται διάφορες μελέτες που έχουν γίνει τα τελευταία 20 χρόνια και αναδεικνύουν τον ρόλο της μηχανικής μάθησης στην πρόβλεψη των αποτελεσμάτων διαφόρων αγωνισμάτων, και εξετάζουν την σημαντικότητα διαφορετικών παραμέτρων κάθε φορά. Αυτές οι παράμετροι φαίνεται να παίζουν σημαντικό ρόλο στην έκβαση των αγώνων και έχουν σημασία για τις ομάδες αφού η βελτίωση αυτών οδηγεί σε καλύτερη απόδοση και θετικά αποτελέσματα.

Μια σημαντική, λοιπόν, παράμετρος για τα Basketball Analytics επικεντρώνεται στην πρόβλεψη του αποτελέσματος ενός παιχνιδιού μπάσκετ. Συχνά τέτοιες προβλέψεις εκτελούνται με την ανάπτυξη μοντέλων πιθανότητας που μπορούν και χρησιμοποιούν τις γνώσεις με βάση μελέτες ανάλυσης απόδοσης. Επιπλέον, τα δεδομένα play-by-play αποτελούν σημαντικό στοιχείο για την ανάπτυξη τέτοιων μοντέλων. Ωστόσο, οι μελέτες αξιολόγησης απόδοσης του μπάσκετ και τα σχετικά δεδομένα play-by-play που δημιουργούνται χρησιμοποιώντας δεδομένα παρακολούθησης παικτών επικεντρώνονται σε μεγάλο βαθμό σε μετρήσεις παιχνιδιού βασιζόμενες σε δεδομένα από την επίθεση και μέχρι πρόσφατα είχε καταβληθεί μέτρια προσπάθεια για τον ρόλο του αμυντικού παιχνιδιού. Αυτό οφείλεται κυρίως στην ευκολία καταγραφής των πόντων, των ασίστ και άλλων στατιστικών που σχετίζονται με την επίθεση. Η διαθεσιμότητα πληροφοριών για το αμυντικό παιχνίδι των ομάδων είναι μια σημαντική ανάγκη για την ανάπτυξη χρήσιμων αναλυτικών μοντέλων.

Ξεκινώντας από τον Purucker, (1996) στην έρευνα του χρησιμοποίησε ένα Τεχνητό Νευρωνικό Δίκτυο (Artificial Neural Network, ANN) καθώς και τεχνικές μάθησης χωρίς επίβλεψη για να προβλέψει τα αποτελέσματα των ποδοσφαιρικών αγώνων της National Football League (NFL) των ΗΠΑ το 1994, χρησιμοποιώντας δεδομένα από τις εβδομάδες 11 έως 16 του Πρωταθλήματος (90 αγώνες). Εξετάστηκαν έξι κατηγορίες: turnover margin, yardage differential, victories, time in possession, rushing yardage differential, και betting odds (η συμπερίληψη των αποδόσεων του στοιχήματος βρέθηκε να βελτιώνει τα αρχικά αποτελέσματα). Χρησιμοποιήθηκε ένα Τεχνητό Νευρωνικό Δίκτυο που εκπαιδεύτηκε με Back Propagation (BP). Οι αγώνες από τις εβδομάδες 12 έως 15 χρησιμοποιήθηκαν για την πρόβλεψη της εβδομάδας 16, με το Τεχνητό Νευρωνικό Δίκτυο να προβλέπει σωστά 11 από τους 14 αγώνες (78,6%) την εβδομάδα 16. Οι εβδομάδες 12 έως 14 χρησιμοποιήθηκαν επίσης για την πρόβλεψη της εβδομάδας 15, με το Τεχνητό Νευρωνικό Δίκτυο να προβλέπει σωστά 10 από τα 14 παιχνίδια (71,4%).

Ένας άλλος ερευνητής, ο Kahn (2003) χρησιμοποιώντας τεχνικές μηχανικής μάθησης αγώνες στο NFL χρησιμοποιώντας δεδομένα από 208 αγώνες κατά τη σεζόν του 2003. Χρησιμοποιήθηκε ένα Τεχνητό Νευρωνικό Δίκτυο με Back Propagation και τα χαρακτηριστικά που περιλαμβάνονται στο μοντέλο ήταν: rushing yardage differential, turnover differential, total yardage differential, home or away indicator, time in possession differential, home team outcome και away team outcome. Τα αριθμητικά χαρακτηριστικά υπολογίστηκαν ως ο ιστορικός μέσος όρος των 3 εβδομάδων (ο μέσος όρος της στατιστικής τις τελευταίες 3 εβδομάδες του Πρωταθλήματος), καθώς και ο μέσος όρος της σεζόν (ο μέσος όρος της στατιστικής για ολόκληρη τη σεζόν). Διαπιστώθηκε ότι η χρήση του μέσου όρου της σεζόν παρείχε καλύτερη ακρίβεια. Οι εβδομάδες 1 έως 13 του Πρωταθλήματος του 2003 χρησιμοποιήθηκαν ως δεδομένα εκπαίδευσης, με

τις εβδομάδες 14 και 15 ως τεστ. Επιτεύχθηκε ακρίβεια 75%, το οποίο ήταν ελαφρώς καλύτερο από τις προβλέψεις των ειδικών στα ίδια παιχνίδια.

Οι Joseph, Fenton, και Neil, (2006) διαπίστωσαν ότι μπορεί να είναι ωφέλιμο να ενσωματωθεί η γνώση εξειδικευμένων ατόμων στην πρόβλεψη αγώνων σε μια διαδικασία μοντελοποίησης. Η γνώση αυτών ενσωματώθηκε κατά την κατασκευή ενός μοντέλου Bayesian Network (BN) και διαπιστώθηκε ότι η συμπερίληψη αυτών των γνώσεων μπορεί να οδηγήσει σε μια ισχυρή απόδοση, ειδικά όταν το μέγεθος του δείγματος είναι μικρό. Ένα δέντρο αποφάσεων (MC4) και το μοντέλο k-Nearest Neighbor χρησιμοποιήθηκαν επίσης για την πρόβλεψη των αποτελεσμάτων των αγώνων ποδοσφαίρου που έπαιξε η αγγλική ομάδα Tottenham Hotspurs. Το σύνολο δεδομένων τους ήταν 76 αγώνες και τέσσερις μεταβλητές συμπεριλήφθηκαν στο μοντέλο αυτό, ενώ 30 μεταβλητές χρησιμοποιήθηκαν στο αρχικό γενικό τους μοντέλο. Το μοντέλο Bayesian Network βρέθηκε να παρέχει την καλύτερη απόδοση, επιτυγχάνοντας ακρίβεια 59,2% στην πρόβλεψη νίκη εντός έδρας, νίκη εκτός έδρας ή ισοπαλία. Για μελλοντικές έρευνες, οι συγγραφείς πρότειναν την ανάπτυξη ενός πιο συμμετρικού μοντέλου χρησιμοποιώντας παρόμοια δεδομένα, αλλά για όλες τις ομάδες του Αγγλικού Πρωταθλήματος, και την ενσωμάτωση χαρακτηριστικών ποιότητας του κάθε παίκτη (π.χ. παίκτες που έχουν παίξει σε διεθνές επίπεδο) στο μοντέλο. Επιπλέον, ανέφεραν ότι θα μπορούσαν να προστεθούν επιπλέον παράμετροι, όπως π.χ. ποιότητα επίθεσης με βάση τα σκορ, κ.α.

Οι McCabe & Trevathan (2008) εξέτασαν τέσσερα διαφορετικά αθλήματα: την Rugby League στο Εθνικό Πρωτάθλημα Ποδοσφαίρου της Αυστραλίας (NFL), το Australian Rules Football στο Πρωτάθλημα Ποδοσφαίρου Αυστραλίας (AFL), το Super Rugby και το EPL Soccer, χρησιμοποιώντας δεδομένα από το 2002 έως το 2007. Ένα Τεχνητό Νευρωνικό Δίκτυο εκπαιδεύτηκαν με Back Propagation και ένα Τεχνητό Νευρωνικό Δίκτυο που εκπαιδεύτηκε με τον αλγόριθμο Conjugative-Gradient Descent (CGD) εφαρμόστηκαν. Διαπιστώθηκε ότι το Τεχνητό Νευρωνικό Δίκτυο με Back Propagation ήταν ελαφρώς πιο ακριβής από τον αλγόριθμο CGD. Ωστόσο, ο χρόνος που χρειάστηκε με το Back Propagation ήταν μεγαλύτερος. Οι μεταβλητές που χρησιμοποιήθηκαν στη μελέτη ήταν οι ίδιες και στα τέσσερα αθλήματα. Η μέση ακρίβεια που επιτεύχθηκε από το Τεχνητό Νευρωνικό Δίκτυο εκπαιδεύτηκαν με Back Propagation ήταν 67,5%, υψηλότερη από τις προβλέψεις των ειδικών αναλυτών που κυμαίνονταν από 60% έως 65%.

Ο Soto Valero (2016) πραγματοποίησε μια συγκριτική μελέτη για την πρόβλεψη των αγώνων του μπέιζμπολ, χρησιμοποιώντας 10 χρόνια δεδομένων από τη Major League Baseball (MLB). Μερικά από τα υποψήφια μοντέλα για την έρευνα ήταν τα Lazy Learners, Τεχνητό Νευρωνικό Δίκτυο, Support Vector Machine (SVM) και Decision Trees. Εφαρμόστηκε το πλαίσιο Cross-Industry Standard Process for Data

Mining (CRISP-DM) και, επιπλέον, συγκρίθηκαν η απόδοση μιας προσέγγισης ταξινόμησης (νίκη ή ήττα για την γηπεδούχος ομάδα), καθώς και μια προσέγγιση αριθμητικής πρόβλεψης (πρόβλεψη της διαφοράς μεταξύ της γηπεδούχου και της εκτός έδρας ομάδας). Οι μέθοδοι επιλογής χαρακτηριστικών από το WEKA Software εφαρμόστηκαν για μείωση των διαστάσεων και κατάταξη του αρχικού συνόλου 60 χαρακτηριστικών. Το μοντέλο τους χρησιμοποίησε μόνο τις τρεις κορυφαίες μεταβλητές αφού η προσθήκη περισσότερων μεταβλητών δεν βρέθηκε να βελτιώνει τα αποτελέσματα. Το Support Vector Machine (SVM) βρέθηκε να παράγει αποτελέσματα με μεγαλύτερη ακρίβεια τόσο στην ταξινόμηση όσο και στις προσεγγίσεις αριθμητικής πρόβλεψης. Σύμφωνα με τους Delen, Cogdell, και Kasap (2012), η προσέγγιση ταξινόμησης είχε πολύ καλύτερη απόδοση από την προσέγγιση αριθμητικής πρόβλεψης. Το μοντέλο ταξινόμησης Support Vector Machine πέτυχε ακρίβεια περίπου 59%. Ωστόσο, όταν χρησιμοποίησαν τις σεζόν 2005-2013 ως δεδομένα προπόνησης και τη σεζόν 2014 ως δεδομένα δοκιμής, οι προβλέψεις του μοντέλου δεν ήταν σημαντικά πιο ακριβείς από τις προβλέψεις που προέρχονται από τις αποδόσεις στοιχημάτων αγώνα. Οι συγγραφείς υπογράμμισαν τη δυσκολία στην πρόβλεψη των αποτελεσμάτων στο Μπέιζμπολ χρησιμοποιώντας μόνο στατιστικά δεδομένα, αλλά πρότειναν ότι τα πειράματα που χρησιμοποιούν τα ιαπωνικά ή κορεατικά πρωταθλήματα μπέιζμπολ θα μπορούσαν να είναι χρήσιμα.

Οι Ivanković et al. (2010) χρησιμοποίησαν ένα Τεχνητό Νευρωνικό Δίκτυο με Back Propagation για την πρόβλεψη των αποτελεσμάτων της Σερβικής First B Basketball League χρησιμοποιώντας πέντε σεζόν δεδομένων από το 2005/2006 έως το 2009/2010 – στο σύνολο 890 αγώνες. Οι συγγραφείς εφάρμοσαν το πλαίσιο Cross-Industry Standard Process for Data Mining (CRISP-DM) ως την πειραματική τους προσέγγιση και διερεύνησαν πώς το επιτυχημένο ποσοστό στα σουτ σε έξι διαφορετικές περιοχές του γηπέδου επηρέασε τα αποτελέσματα των αγώνων. Το σύνολο δεδομένων διαιρέθηκε σε αναλογίες 75:25 για εκπαίδευση και δοκιμές και επιτεύχθηκε ακρίβεια 66,4% στο σύνολο δοκιμών. Οι συγγραφείς στη συνέχεια γύρισαν στη φάση προετοιμασίας δεδομένων του πλαισίου CRISP-DM για να δουν αν η προσθήκη επιπλέον μεταβλητών θα μπορούσε να βελτιώσει τα αποτελέσματα. Συγκεκριμένα, στη συνέχεια ενσωματώθηκαν μεταβλητές όπως τα επιθετικά ριμπάουντ, τα αμυντικά ριμπάουντ, οι ασίστ, τα κλεψίματα, τα λάθη και τα μπλοκ. Αυτό βελτίωσε την ακρίβεια που επιτεύχθηκε από το μοντέλο σε περίπου 81%. Από την μελέτη βγήκε το συμπέρασμα ότι οι ενέργειες στη ζώνη ακριβώς κάτω από το καλάθι – ριμπάουντ στην άμυνα, καθώς και σκοράρισμα σε αυτήν τη ζώνη ήταν κρίσιμες για τον καθορισμό του αποτελέσματος του παιχνιδιού.

Οι Miljković et al. (2010) χρησιμοποίησαν δεδομένα από 778 παιχνίδια στην κανονική σεζόν 2009/2010 του NBA για να προβλέψουν τα αποτελέσματα του αγώνων μπάσκετ. Τα χαρακτηριστικά πρόβλεψης χωρίστηκαν σε στατιστικά παιχνιδιών που σχετίζονται άμεσα με γεγονότα εντός του αγώνα, όπως π.χ. φάουλ ανά παιχνίδι και λάθη ανά παιχνίδι, και εκείνα που σχετίζονται με την βαθμολογία, π.χ. συνολικές νίκες και συνεχόμενες νίκες. Το Naive Bayes βρέθηκε να είναι το μοντέλο με την καλύτερη απόδοση σε σύγκριση με το μοντέλο k-Nearest Neighbor, το Decision Tree και το Support Vector Machine (SVM). Η ακρίβεια του 67% που επιτεύχθηκε με τον Naive Bayes, χρησιμοποιώντας 10πλάσια διασταυρούμενη επικύρωση. Τα μελλοντικά ερευνητικά τους σχέδια περιλάμβαναν την εφαρμογή του συστήματός τους σε άλλα αθλήματα και στον πειραματισμό με άλλα μοντέλα όπως τα ANN.

Ο Cao (2012) δημιούργησε ένα σύστημα που αυτοματοποίησε τη συλλογή δεδομένων και εφάρμοσε διάφορους αλγόριθμους εξόρυξης δεδομένων για την πρόβλεψη του αποτελέσματος ενός αγώνα μπάσκετ. Λήφθηκαν έξι χρόνια δεδομένων από αγώνες NBA, από τη σεζόν 2005/2006 έως τη σεζόν 2010/2011 (που περιλάμβαναν περίπου 4.000 αγώνες). Το σύνολο δεδομένων χωρίστηκε σε σύνολα εκπαίδευσης, σετ δοκιμών και σετ επικύρωσης. Χρησιμοποιήθηκαν τέσσερις αλγόριθμοι ταξινόμησης στα πειράματα: Simple Logistic Regression, Naive Bayes, Support Vector Machine (SVM) και Τεχνητό Νευρωνικό Δίκτυο με Back Propagation. Η διαδικασία επιλογής χαρακτηριστικών ήταν μη αυτόματη, με 46 διαφορετικές παραμέτρους να επιλέγονται με βάση τις γνώσεις του συγγραφέα. Όλα τα μοντέλα βρέθηκαν να παράγουν αρκετά παρόμοια ποσοστά ακρίβειας ταξινόμησης, με το απλό Logistic Regression να επιτυγχάνει την μεγαλύτερη ακρίβεια με 67,8%. Το μοντέλο Simple Logistic Regression παρείχε επίσης αυτόματη επιλογή χαρακτηριστικών. Οι προβλέψεις των ειδικών στο teamrankings.com ήταν ελαφρώς καλύτερες, επιτυγχάνοντας ακρίβεια 69,6%. Ο συγγραφέας πρότεινε ότι στο μέλλον οι τεχνικές ομαδοποίησης θα μπορούσαν να χρησιμοποιηθούν για την ομαδοποίηση παικτών ανά ομάδα θέσης ή για τον εντοπισμό εξαιρετικών παικτών. Η χρήση μεθόδων ανίχνευσης ακραίων τιμών για τον εντοπισμό πολύ καλών παικτών ή της κατάστασης της ομάδας, την διερεύνηση του αντίκτυπου της απόδοσης των παικτών στο αποτέλεσμα του αγώνα και η σύγκριση διαφορετικών χαρακτηριστικών που προέρχονται από τα συνολικά στατιστικά με τα στατιστικά της ομάδας αναφέρθηκαν επίσης ως σημαντικές προτάσεις για μελλοντική έρευνα.

Οι Thabtah et al. (2019) χρησιμοποίησαν το μοντέλο Naive Bayes, ένα Τεχνητό Νευρωνικό Δίκτυο και ένα Logistic Model Tree Decision Tree για να προβλέψει το αποτέλεσμα των αγώνων μπάσκετ του NBA. Επικεντρώθηκαν στη δοκιμή διαφορετικών συνόλων δεδομένων και χαρακτηριστικών προκειμένου να βρουν το βέλτιστο υποσύνολο προβλέψεων. Το σύνολο δεδομένων λήφθηκε από το Kaggle.com και

περιείχε 430 αγώνες τελικών του NBA από το 1980 έως το 2017 και 21 χαρακτηριστικά. Η μεταβλητή στόχος ήταν μια δυαδική μεταβλητή που εκφράζεται ως νίκη ή ήττα. Διαπιστώθηκε ότι τα αμυντικά ριμπάουντ ήταν ο πιο σημαντικός παράγοντας που επηρέαζε τα αποτελέσματα των αγώνων. Οι μέθοδοι επιλογής χαρακτηριστικών που χρησιμοποιήθηκαν βασίστηκαν στο Multiple Regression, στο Correlation Feature Subset – CFS και στον αλγόριθμο RIPPER. Τα αμυντικά ριμπάουντ επιλέχθηκαν ως σημαντικά χαρακτηριστικά και με τις τρεις μεθόδους επιλογής χαρακτηριστικών. Το μοντέλο με την καλύτερη απόδοση ως προς την ακρίβεια εκπαιδεύτηκε σε ένα σύνολο χαρακτηριστικών που περιέχει οκτώ μεταβλητές. Αυτά επιλέχθηκαν από τους κανόνες απόφασης RIPPER και στη συνέχεια εκπαιδεύτηκαν χρησιμοποιώντας το μοντέλο LMT, το οποίο έφτασε στο 83% ακρίβεια. Οι συγγραφείς πρότειναν ότι θα μπορούσαν να εξεταστούν μεγαλύτερα σύνολα δεδομένων και περισσότερα χαρακτηριστικά, όπως οι παίκτες και ο προπονητής της ομάδας. Η χρήση άλλων μοντέλων, όπως τεχνικές βασισμένες στη λειτουργία και μέθοδοι deep learning, αναφέρθηκαν επίσης ως δυνητικοί δρόμοι για περαιτέρω έρευνα.

2.2.3. Basketball Analytics

Ένα παιχνίδι μπάσκετ είναι μοντελοποιημένο ως μια διαδικασία με την οποία οι παίκτες που σχηματίζουν τις δυάδες προσελκύουν και αποκρούουν ο ένας τον άλλον για να παράγουν τα μοναδικά μοτίβα που χαρακτηρίζουν τις συμπεριφορές των παικτών. (Bourbousson, Sève & McGarry, 2010) Η ομαδική συμπεριφορά στο μπάσκετ μπορεί να χαρακτηριστεί από τη δυναμική δημιουργίας χώρου που σχετίζεται με την επιθετική συμπεριφορά και τη δυναμική προστασίας του χώρου που λειτουργούν για την εξουδετέρωση της δυναμικής δημιουργίας χώρου με αμυντικό παιχνίδι. Το Pick-and-Roll (PNR) έχει τη μεγαλύτερη συχνότητα εμφάνισης μεταξύ όλων των δυναμικών δημιουργίας χώρου στο μπάσκετ. (Lamas et al., 2011) Η κατανόηση των στρατηγικών παιχνιδιού από προηγούμενα γεγονότα (π.χ. ανάλυση του ιστορικού αγώνων) μπορεί να επιτρέψει στις ομάδες να αποκτήσουν ανταγωνιστικό πλεονέκτημα γνωρίζοντας τους αντιπάλους τους και καταλήγοντας σε νέες στρατηγικές για να μετριάσουν τα αντιληπτά δυνατά σημεία της αντίπαλης ομάδας. Στο παρελθόν, οι ερευνητές χρησιμοποίησαν κατά κύριο λόγο δύο τύπους δεδομένων για να αναλύσουν στρατηγικές παιχνιδιών μπάσκετ, όπως δεδομένα play-by-play που περιγράφουν διαφορετικά γεγονότα που συμβαίνουν στο γήπεδο, όπως shots pass, dribbles, και φάουλ και δεδομένα παρακολούθησης παικτών και μπάλας . Για παράδειγμα, τα δεδομένα play-by-play μπορούν να χρησιμοποιηθούν για να μάθουν την αποτελεσματικότητα διαφορετικών τύπων αναπαραγωγής Pick-and-Roll. (Marmarinos, Apostolidis, Kostopoulos & Apostolidis, 2016)

Οι Wu και Swartz, (2017) πρότειναν διορθώσεις για σφάλματα στις αλλαγές του αγώνα χρησιμοποιώντας μοντέλα ταξινόμησης που βασίζονται σε δεδομένα play-by-play που ελήφθησαν από το NBA. Οι Talukder και Vincent, (2016) αντιμετώπισαν το πρόβλημα της πρόβλεψης τραυματισμών στο μπάσκετ, όπου συνδύασαν δεδομένα play-by-play, με τον φόρτο εργασίας των παικτών και δεδομένα παρακολούθησης για να δημιουργήσουν έναν αλγόριθμο ταξινόμησης Random Forest ώστε να προβλέψουν τραυματισμούς για τους παίκτες του NBA.

Τα δεδομένα παρακολούθησης, οι κινήσεις των παικτών και της μπάλας χρησιμοποιούνται ευρέως για την ενημέρωση και τον ανασχεδιασμών των στρατηγικών του παιχνιδιού. Η ανάλυση των δεδομένων παρακολούθησης είναι πιο χρήσιμη σε ομαδικά αθλήματα όπου η χωρική οργάνωση των ομάδων σε σχέση με την μπάλα και η χρονική δυναμική της οργάνωσης της ομάδας παίζουν σημαντικό ρόλο στη στρατηγική του παιχνιδιού. Ως εκ τούτου, υπήρξε σημαντικό ενδιαφέρον για την εκμάθηση στρατηγικών παιχνιδιών από δεδομένα παρακολούθησης παικτών σε ομαδικά αθλήματα όπως το μπάσκετ και το ποδόσφαιρο. Προηγούμενες μελέτες έχουν δείξει ότι τα δεδομένα των παικτών για το κάθε παιχνίδι μπορούν να θεωρηθούν ως ένας από τους πιο αποτελεσματικούς παράγοντες που διατίθενται στις τεχνικές μηχανικής μάθησης: για παράδειγμα, χαρτογράφηση στρατηγικών παιχνιδιού NBA βάσει δεδομένων παρακολούθησης παικτών (Miller & Bornn, 2017), ανάλυση του στυλ στο σουτ των παικτών στο NBA χρησιμοποιώντας την τροχιά της μπάλας και την στάση του σώματος, και προβλέψεις αποτελεσμάτων των σουτ τριών πόντων χρησιμοποιώντας επαναλαμβανόμενα νευρωνικά δίκτυα. (Shah & Romijnders, 2016)

2.2.4. NBA and Euroleague Attributes

Με τα χρόνια, έχουν δημιουργηθεί πολλές μέθοδοι εγγραφής και ανάλυσης για τη μέτρηση των μεταβλητών του παιχνιδιού, από απλά φύλλα στατιστικών που συμπληρώνονται από βοηθούς προπονητές έως τις πλήρως ηλεκτρονικές διαδικασίες που καταγράφουν όλους τους σημαντικούς δείκτες απόδοσης παιχνιδιού. (Oliver, 2004) Προηγούμενες μελέτες που βασίζονται σε στατιστικά που σχετίζονται με το παιχνίδι έχουν δείξει ότι οι διαφορές μεταξύ των ομάδων που κερδίζουν και που χάνουν οφείλονται κυρίως στα αμυντικά ριμπάουντ, στα ποσοστά 2 πόντων και στις ασίστ. (Ibáñez et al., 2008) Πιο πρόσφατες έρευνες δείχνουν ότι τα στατιστικά στοιχεία που σχετίζονται με το παιχνίδι των επιδόσεων της ομάδας ποικίλλουν ανάλογα και παράγοντες, όπως η τοποθεσία του παιχνιδιού (εντός ή εκτός έδρας), ο τύπος του παιχνιδιού (κανονική σεζόν ή πλέι οφ) και οι διαφορές στο τελικό σκορ του παιχνιδιού παίζουν πολύ σημαντικό ρόλο. (Puente, Coso, Salinero & Abián-Vicén., 2015)

Γενικά, δεν έχουν γίνει πολλές μελέτες που συγκρίνουν το NBA και την Euroleague που είναι τα δύο καλύτερα πρωταθλήματα μπάσκετ στον κόσμο. Αρχικά, τα θεμελιώδη δομικά χαρακτηριστικά της επίθεσης μπάσκετ στο NBA και στην Euroleague δείχνουν πολλές ομοιότητες, ιδίως στον ρυθμό και στη δυναμική παιχνιδιού. (Selmanovic, Škegro & Milanović, 2021) Μια ανάλυση της τεχνικής του σουτ κατέληξε στο συμπέρασμα ότι δύο διακριτές διαφορές μεταξύ του NBA και του ευρωπαϊκού μπάσκετ είναι ότι τα καρφώματα είναι πιο συχνά και οι βολές είναι λιγότερο συχνές στο NBA σε σύγκριση με το ευρωπαϊκό μπάσκετ, το οποίο μπορεί να αποδοθεί στα καλύτερα αθλητικά προσόντα των παικτών του NBA. (Erčulj & Štrumbelj, 2015) Μία από τις λίγες μελέτες που συνέκριναν το NBA και το ευρωπαϊκό μπάσκετ αποκάλυψε ότι οι ομάδες του NBA χρησιμοποιούν το overhead pass ενώ οι ευρωπαϊκές ομάδες χρησιμοποιούν το bounce pass για να περάσουν την μπάλα στον παίκτη κοντά στο καλάθι. Σε αντίθεση με τις ευρωπαϊκές ομάδες, παίκτες που παίζουν σε διάφορες θέσεις βρέθηκαν σε post-up κατάσταση μέσα στο παιχνίδι στο NBA, όχι μόνο οι centers των ομάδων. Κατά συνέπεια, υπάρχει περισσότερο inside game στο NBA παρά στο ευρωπαϊκό μπάσκετ. (Mavridis, Tsamourtzis, Karipidis & Laios, 2009)

2.2.5. Διαθέσιμα Δημόσια Datasets

Τα dataset που θα χρησιμοποιηθούν σε αυτήν την μελέτη έχουν βρεθεί σε διάφορα site. Από το επίσημο site του NBA , στο αντίστοιχο μέρος των στατιστικών (www.nba.com/stats/) μπορούν να βρεθούν λίστες με τα στατιστικά των παιχτών του NBA από το 1996 και μετά. Επίσης, από το Kaggle.com μπορούν να αξιοποιηθούν datasets για την απόδοση των παιχτών και των ομάδων του NBA και της Euroleague. Τέλος, στο mysportsfeeds.com μπορούν να βρεθούν αξιόλογα datasets για το NBA και στο basketball-reference.com για την Euroleague, site το οποίο περιέχει στατιστικά απο το 2000.

3. Προτεινόμενη Μεθοδολογία

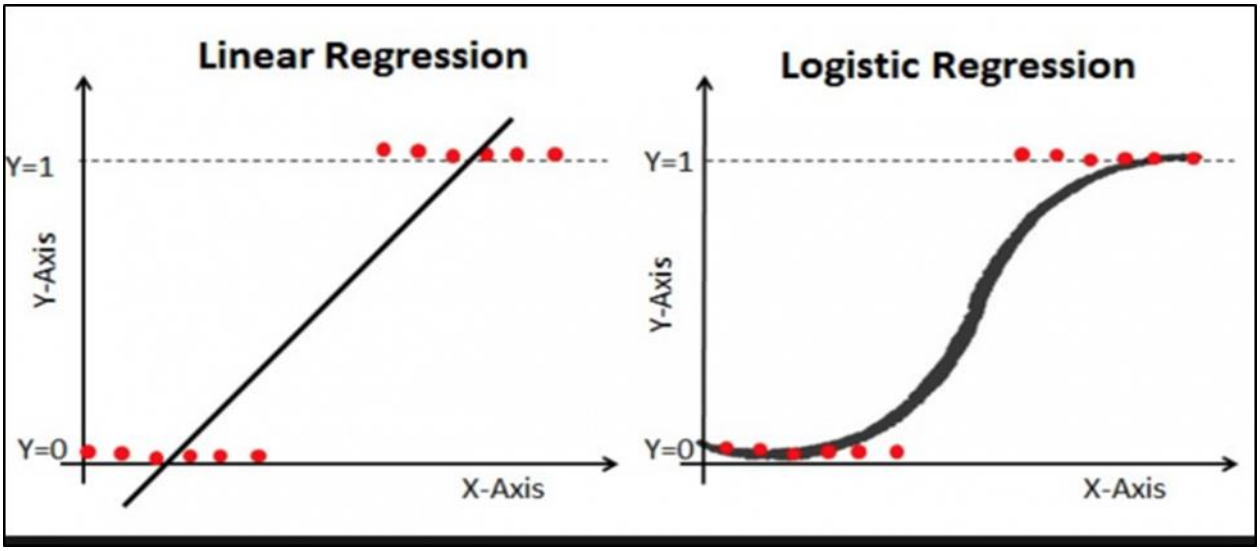
3.1. Τεχνικές που Χρησιμοποιήθηκαν στη Διπλωματική

Σε αυτό το κεφάλαιο της διπλωματικής αναλύονται οι μέθοδοι και οι τεχνικές που χρησιμοποιήθηκαν για την ανάλυση των δεδομένων της διπλωματικής εργασίας.

3.1.1. Logistic Regression

Η χρήση του μοντέλου Logistic Regression χρονολογείται από το 1845. Εμφανίστηκε για πρώτη φορά κατά τη διάρκεια των μαθηματικών μελετών για την αύξηση του πληθυσμού εκείνη την εποχή. Ο όρος προέρχεται από μετασχηματισμό logit, ο οποίος εφαρμόζεται στην εξαρτημένη μεταβλητή. Αυτή η περίπτωση, ταυτόχρονα, προκαλεί ορισμένες διαφορές τόσο στην εκτίμηση όσο και στην ερμηνεία.

Ο κύριος στόχος του μοντέλου Logistic Regression είναι η ταξινόμηση των δεδομένων σε διαφορετικές ομάδες. Μια τυπική εξίσωση παλινδρόμησης αποτελείται από πραγματικές τιμές μερικών ανεξάρτητων μεταβλητών και βαρών που παράγονται από το μοντέλο για την πρόβλεψη της τιμής της εξαρτημένης μεταβλητής. Από την άλλη πλευρά, στην Logistic Regression, η εκτιμώμενη τιμή κυμαίνεται από 0 έως 1. Με άλλα λόγια, αποκαλύπτει την πιθανότητα συγκεκριμένων συνεπειών για κάθε θέμα (για παράδειγμα "YES" ή "NO"). Η ανάλυση παράγει μια εξίσωση παλινδρόμησης που δίνει τη δυνατότητα να γίνει μια ακριβής εκτίμηση για την πιθανότητα ότι ένα άτομο εμπίπτει σε μία από τις κατηγορίες ("YES") ή ("NO"). Ως αποτέλεσμα, η κύρια διαφορά μεταξύ των δύο τεχνικών είναι ότι η τιμή της εξαρτημένης μεταβλητής εκτιμάται σε ανάλυση πολλαπλής γραμμικής παλινδρόμησης, ενώ η πιθανότητα εμφάνισης μίας από τις τιμές που μπορεί να έχει η εξαρτημένη μεταβλητή εκτιμάται στην ανάλυση με το μοντέλο Logistic Regression. (Peng, Lee and Ingersoll, 2002) Η Εικόνα 5 αποτυπώνει την διαφορά των δύο μοντέλων σε γραφικό επίπεδο.



Εικόνα 5 Linear Regression vs. Logistic Regression

Η Logistic Regression είναι μια ανάλυση που μας δίνει τη δυνατότητα να εκτιμήσουμε κατηγορηματικά αποτελέσματα όπως η συμμετοχή σε μια ομάδα με τη βοήθεια μιας ομάδας μεταβλητών. Η Logistic Regression, σε αντίθεση με την Ανάλυση Πολλαπλής Παλινδρόμησης (Multiple Regression Analysis), δεν απαιτεί παραδοχές για την κατανομή ανεξάρτητων μεταβλητών. Με άλλα λόγια, δεν πρέπει να ικανοποιούνται παραδοχές όπως η κανονική κατανομή ανεξάρτητων μεταβλητών, η γραμμικότητα και η ισότητα του πίνακα διακύμανσης – συνδιακύμανσης. Ουσιαστικά, η Logistic Regression είναι πολύ πιο ευέλικτη από άλλες τεχνικές. Τέλος, είναι λογικό να δηλώνεται ότι είναι πιο εύκολο να ερμηνευθεί το αποτέλεσμα της ανάλυσης με την Logistic Regression. (Leech, Barrett and Morgan, 2008)

3.1.2. k-nearest neighbors (KNN)

Μια τεχνική που περιγράφει τη διαδικασία επίλυσης προβλημάτων με βάση λύσεις για παρόμοια ήδη γνωστά προβλήματα, είναι η k-nearest neighbors (KNN). Ο αλγόριθμος KNN είναι ένας αλγόριθμος μηχανικής μάθησης που θεωρείται «τεμπέλης» αλγόριθμος εκμάθησης, με χαμηλό υπολογιστικό κόστος και πολύ απλή εφαρμογή. (Alkhatib, Najadat, Hmeidi and Shatnawi, 2013) Υποστηρίζει προβλήματα ταξινόμησης και παλινδρόμησης. Όταν κάνει μια πρόβλεψη, αποθηκεύει ολόκληρο το σύνολο δεδομένων εκπαίδευσης και το ρωτά για να εντοπίσει σημεία δεδομένων k στο σύνολο εκπαίδευσης που μοιάζουν περισσότερο με το σημείο δεδομένων που πρέπει να ταξινομηθεί. Επομένως, δεν υπάρχει κανένα άλλο μοντέλο εκτός από το ακατέργαστο σύνολο δεδομένων εκπαίδευσης και ο μόνος

υπολογισμός που εκτελείται είναι το ερώτημα του συνόλου δεδομένων κατάρτισης. Όταν η μέθοδος KNN χρησιμοποιείται για παλινδρόμηση, η τιμή απόκρισης υπολογίζεται ως το σταθμισμένο άθροισμα των αποκρίσεων όλων των k neigbors, όπου το βάρος είναι αντιστρόφως ανάλογο με την απόσταση από την εγγραφή εισόδου. (Wilson and Martinez, 2010)

Κάθε σύστημα εκμάθησης που βασίζεται σε αυτήν την μέθοδο απαιτεί ένα σύνολο παραμέτρων, που είναι:

- Μια συνάρτηση απόστασης που μετρά την ομοιότητα μεταξύ προβλημάτων ή καταχωρίσεων δεδομένων. Αυτό απαιτείται για να μετρηθεί ποιοι είναι οι πιο κοντινοί γείτονες σε ένα πρόβλημα.
- Ένας αριθμός γειτονικών παραμέτρων που λαμβάνονται υπόψη κατά την αντιμετώπιση ενός νέου προβλήματος.
- Μια συνάρτηση στάθμισης που επιτρέπει την περαιτέρω ποσοτικοποίηση των μεταβλητών που βρέθηκαν και να αυξήσει την πρόβλεψη και την ποιότητα μάθησης.
- Μια μέθοδος αξιολόγησης που περιγράφει μια συνάρτηση σχετικά με τον τρόπο χρήσης των κοντινών αριθμών που βρέθηκαν για την επίλυση του συγκεκριμένου προβλήματος.

Αυτές οι μέθοδοι εκμάθησης όπως αναφέρθηκε ήδη αποτελούν μέρος μεθόδων μάθησης που θεωρούνται αρκετά «τεμπέλικες», πράγμα που σημαίνει ότι δεν πραγματοποιείται υπολογισμός στα δεδομένα πριν δοθεί ερώτημα στο σύστημα. Αυτές οι μέθοδοι έρχονται σε αντίθεση με τις πρόθυμες μεθόδους μάθησης, όπως τα δέντρα απόφασης, που προσπαθούν να δομήσουν τα δεδομένα πριν λάβουν ερωτήματα. (Phung, Webb and Sammut, 2020)

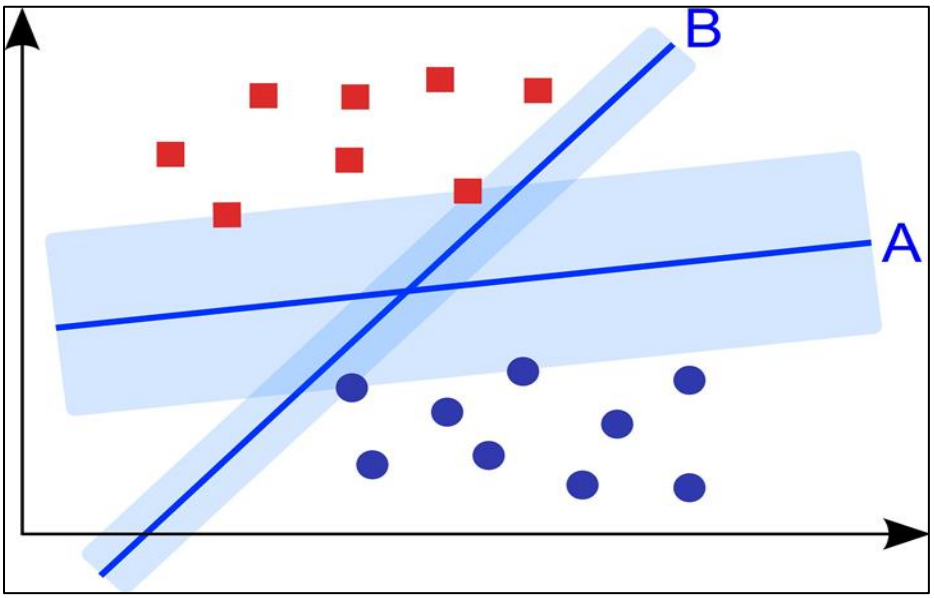
3.1.3. Support Vector Machine (SVM)

Η μέθοδος των Μηχανών Διανυσμάτων Υποστήριξης (Support Vector Machine) ανήκει στον τομέα των μεθόδων μάθησης με επίβλεψη και ως εκ τούτου χρειάζεται επισημασμένα, γνωστά δεδομένα για την ταξινόμηση νέων μη ορατών δεδομένων. Η βασική προσέγγιση για την ταξινόμηση των δεδομένων, ξεκινά προσπαθώντας να δημιουργήσει μια συνάρτηση που χωρίζει τα σημεία δεδομένων στις αντίστοιχες ετικέτες με:

- Τη μικρότερη δυνατή ποσότητα σφαλμάτων.
- Με το μεγαλύτερο δυνατό περιθώριο.

Αυτό οφείλεται στο γεγονός ότι μεγαλύτερες κενές περιοχές δίπλα στη λειτουργία διαχωρισμού οδηγούν σε λιγότερα σφάλματα, επειδή οι ετικέτες διακρίνονται καλύτερα μεταξύ τους.

Η Εικόνα 6 δείχνει ότι ένα σύνολο δεδομένων μπορεί κάλλιστα να διαχωρίζεται από πολλαπλές λειτουργίες χωρίς σφάλματα. Επομένως, το περιθώριο γύρω από μια συνάρτηση διαχωρισμού χρησιμοποιείται ως πρόσθετη παράμετρος για την αξιολόγηση της ποιότητας του διαχωρισμού. Στην περίπτωση αυτή, ο διαχωρισμός A είναι ο καλύτερος, καθώς διακρίνει τις δύο κατηγορίες με πιο ακριβή τρόπο.



Εικόνα 6 Support Vector Machine

Η στατιστική θεωρία μπορεί να προσδιορίσει με ακρίβεια τους παράγοντες που πρέπει να ληφθούν υπόψη για να μάθει επιτυχώς ορισμένους απλούς τύπους αλγορίθμων, ωστόσο, οι πραγματικές εφαρμογές συνήθως χρειάζονται πιο πολύπλοκα μοντέλα και αλγόριθμους (όπως τα νευρικά δίκτυα), που τους καθιστούν πολύ πιο δύσκολους σε ανάλυση. Η μέθοδος των Μηχανών Διανυσμάτων Υποστήριξης (Support Vector Machine) μπορεί να θεωρηθεί ότι βρίσκεται στη διασταύρωση της μαθησιακής θεωρίας και πρακτικής. Κατασκευάζει μοντέλα που είναι αρκετά περίπλοκα (που, για παράδειγμα, περιέχουν μια μεγάλη κατηγορία νευρωνικών δικτύων) και όμως είναι αρκετά απλά για να αναλυθούν μαθηματικά. Αυτό συμβαίνει επειδή ένα SVM μπορεί να θεωρηθεί ως ένας γραμμικός αλγόριθμος σε έναν χώρο υψηλών διαστάσεων. (Cristianini & Shawe-Taylor, 2014)

3.1.4. Random Forest

Η μέθοδος Random Forest είναι ένας εύχρηστος και ευέλικτος αλγόριθμος μηχανικής μάθησης που μπορεί να παράξει, ακόμη και χωρίς έχει ρυθμιστεί με υπερπαραμέτρους, ένα ακριβές αποτέλεσμα. Επίσης, ο αλγόριθμος αυτός είναι εξαιρετικά διαδεδομένος επειδή είναι απλός αλλά και επειδή έχει μεγάλη ποικιλομορφία. Αυτό γιατί μπορεί να χρησιμοποιηθεί όχι μόνο για παλινδρόμηση αλλά και για εργασίες ταξινόμησης, τα οποία αποτελούν την πλειοψηφία των υπαρχόντων συστημάτων machine learning. Ο Random Forest, λοιπόν, είναι ένας αλγόριθμος μάθησης με επίβλεψη. Το "δάσος" που χτίζει, είναι ένα σύνολο δέντρων αποφάσεων, συνήθως εκπαιδευμένο με τη μέθοδο bagging. (Breiman, 2001) Η κύρια ιδέα πίσω από τη μέθοδο αυτή είναι ότι ο συνδυασμός πολλών μοντέλων μάθησης μπορεί να αυξάνει το συνολικό αποτέλεσμα. Με άλλα λόγια, η μέθοδος Random Forest δημιουργεί δέντρα αποφάσεων, τα συγχωνεύει με στόχο να πάρει μια πιο σταθερή και πιο ακριβή πρόβλεψη. (Biau, 2012)

Η μέθοδος Random Forest έχει σχεδόν τις ίδιες υπερπαραμέτρους με ένα δέντρο αποφάσεων. Επίσης, κάνει το μοντέλο να έχει υψηλότερη τυχαιότητα, ενώ μεγαλώνει τα δέντρα αποφάσεων. Αντί να αναζητά το πιο σημαντικό χαρακτηριστικό ενώ διαχωρίζει έναν κόμβο, αναζητά το το πιο καλό χαρακτηριστικό μέσα από ένα τυχαίο υποσύνολο χαρακτηριστικών. Αυτό έχει ως αποτέλεσμα μεγαλύτερη ποικιλία που με τη σειρά της δίνει ένα καλύτερο μοντέλο. Επομένως, με τη μέθοδο αυτή, μόνο ένα τυχαίο υποσύνολο των χαρακτηριστικών λαμβάνεται υπόψη από τον αλγόριθμο. Μια άλλη εξαιρετική ιδιότητα της μεθόδου Random Forest είναι ότι είναι πολύ εύκολο να αξιολογηθεί και να μετρηθεί η σχετική σημασία κάθε χαρακτηριστικού στην πρόβλεψη. Εξετάζοντας τη σημασία της δυνατότητας αυτής, μπορεί ο ερευνητής να αποφασίσει ποια χαρακτηριστικά πιθανόν να απορριφθούν επειδή δεν προσφέρουν αρκετά (και σε μερικές φορές καθόλου) στην διαδικασία πρόβλεψης. Αυτό έχει ιδιαίτερη σημασία επειδή ένας γενικός κανόνας στη μηχανική μάθηση είναι ότι ο ερευνητής που έχει ιδιαίτερες και υψηλού επιπέδου ικανότητες, είναι πολύ πιθανό να κάνει το μοντέλο του να υποφέρει λόγω υπερβολικής εφαρμογής και αντίστροφα. (Biau, Lugosi & Devroye, 2008)

3.1.5. Νευρωνικό Δίκτυο (Multi-layer Perceptron)

Ένα Νευρικό Δίκτυο (Neural Network) είναι ένα διασυνδεδεμένο σύνολο απλών στοιχείων επεξεργασίας, μονάδων ή κόμβων, των οποίων η λειτουργικότητα βασίζεται στον νευρώνα. Η ικανότητα επεξεργασίας του δικτύου αποθηκεύεται στα δυνατά σημεία σύνδεσης, που λαμβάνονται με μια διαδικασία προσαρμογής ή μάθησης από ένα σύνολο προτύπων εκπαίδευσης. Τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Network, ANN) εμπνεύστηκαν από τον κλάδο των βιολογικών επιστημών που μελετά

πώς έχει αναπτυχθεί η νευροανατομία των ζωντανών ζώων για την επίλυση προβλημάτων. Οι πρώτες θεωρίες Τεχνητών Νευρωνικών Δικτύων αναπτύχθηκαν από ερευνητές που προσπαθούν να εξηγήσουν την ανθρώπινη συμπεριφορά και τη διαδικασία σκέψης με μοντελοποίηση του ανθρώπινου εγκεφάλου. (Boon, Kok & Beck, 1995)

Οι κόμβοι (νευρώνες) είναι απλά υπολογιστικά στοιχεία. Συγκεντρώνουν δεδομένα από άλλους νευρώνες μέσω ενός σταθμισμένου ποσού. Εάν επιτευχθεί ένα ορισμένο όριο, ο νευρώνας στέλνει πληροφορίες σε όλους τους άλλους συνδεδεμένους νευρώνες, διαφορετικά παραμένει σε ηρεμία. Μια σημαντική διαφορά σε σύγκριση με τα παραδοσιακά στατιστικά ή βασισμένα σε κανόνες συστήματα είναι η ικανότητα μάθησης ενός ANN. Στην αρχή μιας εκπαιδευτικής διαδικασίας, ένα ANN δεν περιέχει ρητές πληροφορίες. Στη συνέχεια, ένας μεγάλος αριθμός περιπτώσεων με γνωστό αποτέλεσμα παρουσιάζονται στο σύστημα και τα βάρη των ενδονευρωνικών συνδέσεων αλλάζουν με έναν αλγόριθμο εκπαίδευσης σχεδιασμένο για να ελαχιστοποιεί το συνολικό σφάλμα του συστήματος. (Traeger et al., 2003)

Ένα multi-layer perceptron αποτελείται από έναν αριθμό στρωμάτων που περιέχουν έναν ή περισσότερους νευρώνες. Ο ρόλος των εισερχόμενων νευρώνων (στρώμα εισόδου) είναι η τροφοδοσία μοτίβων εισόδου στο υπόλοιπο δίκτυο. Μετά από αυτό το επίπεδο, υπάρχουν ένα ή περισσότερα ενδιάμεσα στρώματα μονάδων, τα οποία ονομάζονται κρυφά στρώματα. Στη συνέχεια, τα κρυμμένα στρώματα ακολουθούνται από ένα τελικό επίπεδο εξόδου όπου διαβάζονται τα αποτελέσματα του υπολογισμού. Κάθε μονάδα συνδέεται με όλες τις μονάδες στο επόμενο επίπεδο και κάθε μονάδα λαμβάνει είσοδο από όλες τις μονάδες στο προηγούμενο επίπεδο. Κάθε σύνδεση έχει ένα συγκεκριμένο βάρος, και αυτό το βάρος απεικονίζει την επίδραση της μονάδας στην απόκριση της μονάδας στο επόμενο στρώμα. Η έξοδος ενός multi-layer perceptron εξαρτάται από την είσοδο και από την ισχύ των συνδέσεων των μονάδων. (Ramchoun et al., 2016) Όταν οι πληροφορίες προσφέρονται σε ένα multi-layer perceptron ενεργοποιώντας τους νευρώνες στο στρώμα εισόδου, αυτές οι πληροφορίες υποβάλλονται σε επεξεργασία στρώμα ανά στρώμα έως ότου τελικά ενεργοποιηθεί το στρώμα εξόδου. Δεδομένων αρκετών κρυφών μονάδων και αρκετών δεδομένων, έχει αποδειχθεί ότι τα multilayer perceptrons μπορούν να προσεγγίσουν σχεδόν οποιαδήποτε λειτουργία με οποιαδήποτε επιθυμητή ακρίβεια. Ωστόσο, αυτά τα αποτελέσματα ισχύουν εάν και μόνο εάν υπάρχει ένας αρκετά μεγάλος αριθμός εκπαιδευτικών δεδομένων στη σειρά. Εάν δεν υπάρχουν αρκετά δεδομένα ώστε να εκπαιδευτεί το νευρωνικό δίκτυο, το δίκτυο δεν θα μπορεί να μάθει με ακρίβεια την απαιτούμενη σχέση εισόδου-εξόδου. Ως εκ τούτου, τα multilayer perceptrons είναι πολύτιμα εργαλεία για την επίλυση πολύπλοκων

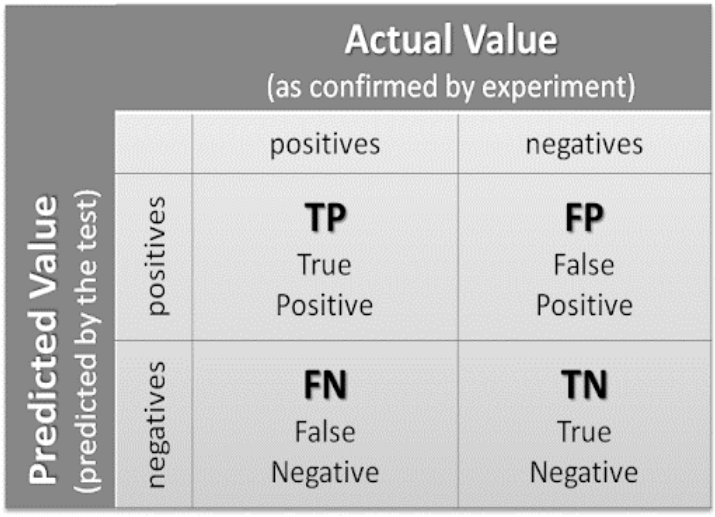
προβλημάτων όταν υπάρχουν επαρκή δεδομένα για την εκπαίδευσή τους. (Kompoliti & Verhagen Metman, 2010)

3.2. Μετρικές Αξιολόγησης Κατηγοριοποιητών

Ένα άλλο πολύ σημαντικό μέρος της μηχανικής μάθησης, είναι το πρόβλημα του πώς ένα πρόγραμμα υπολογιστή παρατηρεί ποια από τα αποτελέσματά του ήταν κατάλληλα και ποια περιείχαν λάθη. Πιο δύσκολα σενάρια εμφανίζονται σε ερευνητικούς τομείς με περιορισμένη ή καθόλου πρόσβαση σε δεδομένα πραγματικού κόσμου, όπως η αξιολόγηση μεταφράσεων εγγράφων. Αυτό απαιτεί μια επιπρόσθετη ανθρώπινη προσπάθεια για να ταξινομηθούν αυτές οι μεταφράσεις σε τάξεις ώστε να μπορούν να συγκριθούν με τα αποτελέσματα του προγράμματος υπολογιστή στο τέλος. Η αξιολόγηση των εργασιών ταξινόμησης γίνεται συνήθως διαχωρίζοντας το σύνολο δεδομένων σε ένα σύνολο δεδομένων εκπαίδευσης και ένα σύνολο δεδομένων δοκιμής. Στη συνέχεια, ο αλγόριθμος μηχανικής εκμάθησης εκπαιδεύεται στον πρώτο, ενώ το σύνολο δεδομένων δοκιμής χρησιμοποιείται για τον υπολογισμό των δεικτών απόδοσης προκειμένου να αξιολογηθεί η ποιότητα του αλγορίθμου. Ένα κοινό πρόβλημα με τους αλγόριθμους μηχανικής μάθησης έγκειται στην πρόσβαση σε περιορισμένα δεδομένα δοκιμών και εκπαίδευσης. Επομένως, η υπερβολική τοποθέτηση μπορεί να είναι ένα σοβαρό πρόβλημα κατά την αξιολόγηση αυτών των προγραμμάτων. Προκειμένου να αντιμετωπιστεί αυτό το πρόβλημα, μια κοινή προσέγγιση είναι, να χρησιμοποιήσετε μια επικύρωση X-Fold Cross Validation. Το Cross Validation περιγράφει τη διαδικασία διαχωρισμού ολόκληρου του συνόλου δεδομένων σε μέρη X και χρησιμοποίησης καθεμιάς από αυτές διαδοχικά ως σύνολο δεδομένων δοκιμής ενώ συνδυάζει τα άλλα με τα δεδομένα εκπαίδευσης. Στη συνέχεια, οι δείκτες απόδοσης υπολογίζονται κατά μέσο όρο σε όλες τις διαδικασίες επικύρωσης. Δεν υπάρχει τέλειος δείκτης για κάθε θέμα σχετικά με την αξιολόγηση των αλγορίθμων μηχανικής μάθησης, καθώς ο καθένας έχει τα ελαττώματα και τα πλεονεκτήματά του. (Singh & Chauhan, 2009) Οι πιο σημαντικοί παράγοντες για την αξιολόγηση της απόδοσης ενός προγράμματος μηχανικής μάθησης παρουσιάζονται στα παρακάτω υποκεφάλαια.

Πριν από αυτό, όμως, θα πρέπει να αναλυθεί το Confusion Matrix. Ένα Confusion Matrix είναι ένας πίνακας που χρησιμοποιείται για τον προσδιορισμό της απόδοσης των μοντέλων ταξινόμησης για ένα δεδομένο σύνολο δεδομένων δοκιμής. Μπορεί να προσδιοριστεί μόνο εάν είναι γνωστές οι πραγματικές τιμές για τα δεδομένα δοκιμής. Η ίδια η μήτρα μπορεί εύκολα να γίνει κατανοητή, αλλά οι σχετικές ορολογίες μπορεί να προκαλέσουν σύγχυση. Δεδομένου ότι δείχνει τα σφάλματα στην απόδοση του

μοντέλου με τη μορφή ενός πίνακα, είναι επίσης γνωστό και ως πίνακας σφάλματος. (Santra & Christy, 2012) Η Εικόνα 7 δείχνει ένα Confusion Matrix.



Εικόνα 7 Confusion Matrix

Ο παραπάνω πίνακας που απεικονίζεται στην Εικόνα 7 εξηγείται ως εξής:

- **True Negative:** Το μοντέλο έχει δώσει πρόβλεψη NO και η πραγματική τιμή ήταν επίσης FALSE.
- **True Positive:** Το μοντέλο έχει δώσει πρόβλεψη YES και η πραγματική τιμή ήταν επίσης TRUE.
- **False Negative:** Το μοντέλο έχει προβλέψει NO, αλλά η πραγματική τιμή ήταν TRUE (Type-II Error).
- **False Positive:** Το μοντέλο έχει δώσει πρόβλεψη YES και η πραγματική τιμή ήταν FALSE (Type-I Error). (Aggarwal, 2019)

3.2.1. Accuracy (Ακρίβεια)

Η Ακρίβεια (Accuracy) είναι ένα μέτρο που περιγράφει γενικά την απόδοση του μοντέλου σε όλες τις τάξεις. Είναι χρήσιμο όταν όλες οι τάξεις είναι εξίσου σημαντικές. Καθορίζει πόσο συχνά το μοντέλο προβλέπει το σωστό αποτέλεσμα. Τα κύρια μειονεκτήματα του Accuracy ως μέτρο αξιολόγησης είναι τα εξής:

- Παραμελεί τις διαφορές μεταξύ των τύπων σφαλμάτων.

- Εξαρτάται από την κατανομή της κλάσης στο σύνολο των δεδομένων (Novakovic et al., 2021)

Υπολογίζεται ως ο λόγος μεταξύ του αριθμού των σωστών προβλέψεων προς τον συνολικό αριθμό των προβλέψεων και δίνεται από τον παρακάτω τύπο.

$$Accuracy = \frac{True_{positive} + True_{negative}}{True_{positive} + True_{negative} + False_{positive} + False_{negative}}$$

3.2.2. Precision (Ακρίβεια)

Το μέτρο αξιολόγησης Precision (Ακρίβεια) μπορεί να οριστεί ως ο αριθμός των σωστών δεδομένων εξόδου που παρέχονται από το μοντέλο ή από όλες τις θετικές τάξεις που έχει προβλέψει σωστά το μοντέλο, πόσες από αυτές ήταν πραγματικά αληθείς. (Novakovic et al., 2021) Μπορεί να υπολογιστεί χρησιμοποιώντας τον παρακάτω τύπο:

$$Precision = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Positive(FP)}$$

3.2.3. Recall (Ανάκληση)

Η τιμή ανάκλησης που ονομάζεται επίσης ευαισθησία ορίζεται ως η σχετική ποσότητα πραγματικών ταξινομημένων περιπτώσεων μεταξύ όλων των πραγματικών περιπτώσεων. Η ανάκληση πρέπει να είναι όσο το δυνατόν μεγαλύτερη. (Powers, 2007) Μπορεί να υπολογιστεί χρησιμοποιώντας τον παρακάτω τύπο:

$$Recall = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Negative(FN)}$$

3.2.4. F1 – Score (F1 - measure)

Το μέτρο F1 – score ορίζεται ως ο αρμονικός μέσος των Precision και Recall με τιμές ανάμεσα σε μηδέν(0) και ένα(1) , με (1) για την καλύτερη ακρίβεια και (0) για την χειρότερη. (Powers, 2007) Μπορεί να υπολογιστεί χρησιμοποιώντας τον παρακάτω τύπο:

$$\text{F1-score} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

3.2.5. Καμπύλη ROC και AUC

Η Καμπύλη ROC είναι ένα γράφημα που εμφανίζει την απόδοση ενός ταξινομητή για όλα τα πιθανά όρια. Το γράφημα απεικονίζεται μεταξύ της Πραγματικής Θετικής Απόδοσης, TPR (στον άξονα Y) και της Λανθασμένης Θετικής Απόδοσης, FPR (στον άξονα x). (Novakovic et al., 2021) Οι τύποι που δίνουν τα TPR και FPR είναι οι παρακάτω: (Google Developers, 2020)

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

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

Το ζεύγος σημείων (0,1) είναι ο τέλειος ταξινομητής: ταξινομεί σωστά όλες τις θετικές και όλες τις αρνητικές περιπτώσεις. Αυτό επειδή το ψευδές θετικό ποσοστό είναι 0 (μηδέν) και το θετικό πραγματικό ποσοστό είναι 1 (όλα). Το σημείο (0, 0) είναι ένας ταξινομητής που προβλέπει ότι όλες οι περιπτώσεις είναι αρνητικές, ενώ το σημείο (1, 1) αντιστοιχεί στον ταξινομητή ο οποίος προβλέπει ότι κάθε περίπτωση είναι θετική. Το σημείο (1, 0) είναι ένας ταξινομητής που δεν είναι σωστός για όλες τις ταξινομήσεις. Σε πολλές περιπτώσεις, ο ταξινομητής έχει μια παράμετρο που μπορεί να προσαρμοστεί αυξάνοντας τα πραγματικά θετικά ποσοστά με το κόστος της αύξησης των ψευδώς θετικών τιμών ή τη μείωση του ψευδώς θετικού ρυθμού με βάση την πτώση της τιμής των πραγματικών θετικών τιμών. (Novakovic et al., 2021)

Τα χαρακτηριστικά της Καμπύλης ROC είναι:

- Η καμπύλη ROC είναι ανεξάρτητη από την κατανομή της κλάσης ή το κόστος των σφαλμάτων.
- Η καμπύλη ROC περιέχει όλες τις πληροφορίες που περιέχονται στον πίνακα σφαλμάτων. (Swets, 1988)
- Η καμπύλη ROC παρέχει ένα οπτικό εργαλείο για τη δοκιμή της ικανότητας του ταξινομητή να εντοπίζει σωστά θετικές και αρνητικές περιπτώσεις που ταξινομήθηκαν λανθασμένα.

Το AUC σημαίνει "Area Under ROC Curve". Δηλαδή, η περιοχή AUC μετρά ολόκληρη τη δισδιάστατη περιοχή κάτω από ολόκληρη την καμπύλη ROC από (0,0) έως (1,1). Η περιοχή AUC παρέχει ένα συνολικό μέτρο απόδοσης σε όλα τα πιθανά όρια ταξινόμησης. Ένας τρόπος ερμηνείας της AUC είναι η πιθανότητα ότι το μοντέλο κατατάσσει ένα τυχαίο θετικό παράδειγμα πιο ψηλά από ένα τυχαίο αρνητικό παράδειγμα. Η τιμή AUC κυμαίνεται από 0 έως 1. Ένα μοντέλο του οποίου οι προβλέψεις είναι 100% λάθος έχει AUC 0 και όταν οι προβλέψεις είναι 100% σωστές έχει AUC 1. (Bradley, 1997)

Η AUC είναι επιθυμητή για τους ακόλουθους δύο λόγους:

- Μετρά πόσο καλά ταξινομούνται οι προβλέψεις, παρά τις απόλυτες τιμές τους.
- Μετρά την ποιότητα των προβλέψεων του μοντέλου ανεξάρτητα από το ποιο όριο ταξινόμησης επιλέγεται. (Google Developers, 2020)

3.3. Τεχνικές Προεπεξεργασίας Δεδομένων

Η προεπεξεργασία δεδομένων είναι απαραίτητη για εργασίες μηχανικής μάθησης. Η προεπεξεργασία δεδομένων είναι το βήμα στο οποίο τα δεδομένα κωδικοποιούνται για να τα φέρουν σε αριθμητική κατάσταση με την οποία το μηχάνημα μπορεί εύκολα να τα διαβάσει. Οι τεχνικές προεπεξεργασίας δεδομένων αποτελούν μέρος της εξόρυξης δεδομένων, η οποία δημιουργεί τελικά προϊόντα από μη επεξεργασμένα δεδομένα τα οποία είναι τυποποιημένα / ομαλοποιημένα, δεν περιέχουν μηδενικές τιμές, κ.α. (Meel, 2021)

3.3.1. Standardization

Η Τυποποίηση (Standardization) είναι μια τεχνική κλιμάκωσης όπου οι τιμές επικεντρώνονται γύρω από τη μέση τιμή με τυπική απόκλιση μονάδας. Αυτό σημαίνει ότι ο μέσος όρος γίνεται μηδέν και η προκύπτουσα κατανομή έχει τυπική απόκλιση μονάδας. (Bhandari, 2020) Ο τύπος που δίνει την τυποποίηση είναι:

$$X' = \frac{X - \mu}{\sigma}$$

Η τυποποίηση μπορεί να είναι χρήσιμη σε περιπτώσεις όπου τα δεδομένα ακολουθούν μια κατανομή Gauss. Ωστόσο, αυτό δεν πρέπει απαραίτητα να ισχύει. Επίσης, η τυποποίηση δεν έχει οριακό εύρος. Έτσι, ακόμη και αν υπάρχουν ακραία δεδομένα μέσα στο σύνολο των δεδομένων δεν θα επηρεαστούν από την τυποποίηση. (Bhandari, 2020)

3.3.2. Pearson Correlation

Ο Συντελεστής Συσχέτισης του Pearson μετρά πόσο ισχυρή είναι η γραμμική συσχέτιση μεταξύ δύο συνεχών μεταβλητών και μπορεί να βρεθεί από τον παρακάτω τύπο:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Στον παραπάνω τύπο το r είναι ο Pearson Correlation, το $\mathbf{x}_i$ είναι η τιμή μιας μεταβλητής του X, το $\mathbf{y}_i$ είναι η τιμή μιας μεταβλητής του Y, το $\bar{\mathbf{x}}$ είναι η μέση τιμή του X και το $\bar{\mathbf{y}}$ είναι η μέση τιμή του Y.

Η τιμή της συσχέτισης βρίσκεται πάντα μεταξύ 1 και -1Η συσχέτιση θα μπορούσε να είναι θετική (r>0), που σημαίνει ότι και οι δύο μεταβλητές κινούνται προς την ίδια κατεύθυνση, ή αρνητική (r<0), που σημαίνει ότι όταν αυξάνεται η τιμή μιας μεταβλητής, οι τιμές των άλλων μεταβλητών μειώνονται. Η συσχέτιση μπορεί επίσης να είναι ουδέτερη (r=0) ή μηδενική, που σημαίνει ότι οι μεταβλητές δεν σχετίζονται. (Berman, 2018)

Η απόδοση ορισμένων αλγορίθμων μπορεί να επιδεινωθεί εάν δύο ή περισσότερες μεταβλητές σχετίζονται στενά, που ονομάζεται πολυγραμμικότητα (multicollinearity). Ένα παράδειγμα είναι η γραμμική παλινδρόμηση, όπου μία από τις συσχετιζόμενες μεταβλητές πρέπει να αφαιρεθεί προκειμένου να βελτιωθεί η ικανότητα του μοντέλου.

Μπορεί επίσης να έχει ενδιαφέρον η συσχέτιση μεταξύ των μεταβλητών εισόδου με τις μεταβλητές εξόδου προκειμένου να παραχθούν πληροφορίες για το ποιες μεταβλητές μπορεί ή όχι να είναι σχετικές ως είσοδος για την ανάπτυξη ενός μοντέλου. (Heumann & Schomaker, 2016)

Ο συντελεστής συσχέτισης Pearson μπορεί να χρησιμοποιηθεί για να συνοψίσει την ισχύ της γραμμικής σχέσης μεταξύ δύο δειγμάτων δεδομένων ή και περισσότερων μέσω ενός Pearson Correlation Matrix. (Heumann & Schomaker, 2016)

3.3.3. Missing Values

Στη μηχανική μάθηση, σπάνια λαμβάνονται σύνολα δεδομένων που είναι τέλεια από την αρχή. Στην πραγματικότητα, είναι ένα από τα τελικά βήματα. Τα δεδομένα πρέπει να καθαριστούν και να επεξεργαστούν από τον αναλυτή και να γίνει διαχείριση στις τιμές που λείπουν από το σύνολο δεδομένων.

Στην επιστήμη των δεδομένων, ο ακατάλληλος χειρισμός των missing values δεν επηρεάζει μόνο το συμπέρασμα αλλά και την πρόβλεψη της δημιουργίας μοντέλων. Ένας αλγόριθμος που εκπαιδεύεται σε προκατειλημμένα δεδομένα θα δημιουργήσει προκατειλημμένα αποτελέσματα.

Η βιβλιογραφία αναγνωρίζει τρεις βασικούς μηχανισμούς για τα missing values. Οι τύποι αυτοί είναι οι παρακάτω:

- **Missing Completely At Random (MCAR):** Τα δεδομένα λείπουν ανεξάρτητα τόσο από τα παρατηρούμενα όσο και από τα μη παρατηρημένα δεδομένα. Για παράδειγμα, σε μια έρευνα φοιτητών, εάν λείπουν τυχαία 5% από τις απαντήσεις, είναι MCAR.
- **Missing At Random (MAR):** Ο μηχανισμός MAR υπονοεί ότι η έλλειψη μιας δεδομένης μεταβλητής συσχετίζεται με τις τιμές τουλάχιστον μιας άλλης μεταβλητής. Για παράδειγμα, εάν λείπουν 10% απαντήσεις από την έρευνα των ανδρών φοιτητών και 5% από την έρευνα των γυναικών φοιτητών, τότε είναι MAR.
- **Missing Not At Random (MNAR):** Το MNAR υπονοεί ότι η πιθανότητα έλλειψης μιας τιμής μιας δεδομένης μεταβλητής σχετίζεται με αυτήν την τιμή. Για παράδειγμα, εάν χαμηλώσει ο ερευνητής τον ΜΟ ενός φοιτητή, τόσο υψηλότερα είναι τα missing values, τότε είναι το MNAR. (Cojocaru, 2019)

Οι τεχνικές για τον χειρισμό των Missing values που χρησιμοποιούνται στην πρακτική ανάλυση ποικίλλουν πολύ. Οι τεχνικές μεμονωμένου υπολογισμού (Single Imputation Techniques) δημιουργούν μια συγκεκριμένη τιμή για μια πραγματική τιμή που λείπει σε ένα σύνολο δεδομένων. Αυτή η τεχνική απαιτεί μικρότερο υπολογιστικό Κόστος. Οι πολλαπλές μέθοδοι υπολογισμού (Multiple Imputation Techniques) παράγουν πολλαπλές τιμές για τον υπολογισμό ενός missing value χρησιμοποιώντας

διαφορετικά μοντέλα προσομοίωσης. Αυτές οι μέθοδοι εισάγουν τη μεταβλητότητα των δεδομένων για να βρουν μια σειρά εύλογων αποκρίσεων. Οι πολλαπλές μέθοδοι υπολογισμού είναι πολύπλοκες στη φύση τους, αλλά δεν υποφέρουν από τιμές προκατάληψης όπως οι τεχνικές μεμονωμένου υπολογισμού. Στις πολλαπλές μεθόδους υπολογισμού, κάθε missing value αντικαθίσταται με τιμές m που λαμβάνονται από επαναλήψεις (όπου το m> 1 και το m κυμαίνεται συνήθως μεταξύ 3 και 10). (Khan & Hoque, 2020)

4. Υλοποίηση και Αποτελέσματα

4.1. Εισαγωγή στην Python

Η Python είναι μια γλώσσα γενικού σκοπού. Είναι εύκολο να την μάθει κάποιος και χρησιμοποιείται ευρέως στην επιστήμη στους τομείς του αριθμητικού προγραμματισμού, της τεχνητής νοημοσύνης, της επεξεργασίας εικόνας, της βιολογίας και άλλων. Είναι γνωστό ότι είναι εύκολη και μπορεί να εκφράσει με σαφήνεια πολύπλοκες ιδέες. Η Python είναι μέρος της Πρωτοβουλίας Ανοιχτού Κώδικα (Open Source Initiative).

Η Python έχει ένα περιβάλλον ανάπτυξης, που ονομάζεται IDLE, το οποίο είναι ιδιαίτερα κατάλληλο για επεξεργασία. Η γλώσσα δεν απαιτεί ούτε επιβάλλει τη χρήση της με οποιονδήποτε τρόπο. Το IDLE είναι μέρος της τυπικής διανομής της γλώσσας, προσφέρει επισήμανση σύνταξης, διαδικτυακή βοήθεια και διαδραστικές πλατφόρμες επεξεργασίας.

Οι ακόλουθες δύο λίστες συνοψίζουν τα πλεονεκτήματα και τα μειονεκτήματα του Python ως γλώσσα προγραμματισμού. Ξεκινώντας, τα πλεονεκτήματα είναι:

- **Φορητότητα:** Ευρύ φάσμα υλικού και εφαρμογής σε πλατφόρμες λογισμικού. Μπορεί να χρησιμοποιηθεί σε διάφορα λειτουργικά συστήματα, όπως τα Windows, τα Linux, το MacOS αλλά και σε smartphones. Επίσης, συνεργάζεται καλά με άλλες γλώσσες προγραμματισμού όπως η C, η C++, η Java, κ.α. (Jamsheer, 2020)
- **Ισχυρή Απλότητα:** Τα προγράμματα της Python είναι πολύ πιο γρήγορα να αναπτυχθούν από άλλες γλώσσες υψηλού επιπέδου. Η απλή, καθαρή σύνταξη όχι μόνο επιτρέπει στους αρχικούς προγραμματιστές να θυμούνται τι έκαναν, αλλά επίσης επιτρέπει σε άλλους προγραμματιστές να κατανοήσουν και να αλλάξουν προγράμματα. Αυτό επιτρέπει πολύ χαμηλότερο κόστος συντήρησης για προγράμματα που είναι γραμμένα Python. Η Python είναι μια εξαιρετικά ευέλικτη γλώσσα. Μπορεί να χρησιμοποιηθεί για τις απλούστερες εφαρμογές, καθώς και για την ανάπτυξη σύνθετων ιστότοπων ως και μέχρι την κατασκευή σύνθετων εφαρμογών. Τέλος, έχεις βολικές συμβολοσειρές, λίστες και λεξικά που αποθηκεύουν πολλά objects, χωρίς να χρειάζεται να γνωρίζει κάποιος εκ των προτέρων πόσα θέλει να αποθηκεύσει. (Parikh, 2018)
- **Εκτεταμένη Βιβλιοθήκη Λειτουργιών:** Στην Python καθιστά σχεδόν αδύνατο να γράψει κάποιος ασαφή κώδικα. Η σύνταξη είναι καθαρή, με σταθερή δομή και λειτουργίες. Η Python χρησιμοποιεί μικρά, καλά κατασκευασμένα συστατικά, που ονομάζονται modules. Οι ενότητες

είναι πολύ εύκολο να σχεδιαστούν και να χρησιμοποιηθούν, κάτι που ενθαρρύνει επίσημες και ανεπίσημες βιβλιοθήκες κώδικα.

- **Ελεύθερο Λογισμικό Ανοιχτού Κώδικα:** Χιλιάδες προγραμματιστές σε όλο τον κόσμο συμβάλλουν στην ανάπτυξη της Python. Αυτοί οι προγραμματιστές δεν πληρώνονται για την ανάπτυξη της Python, αλλά πληρώνονται για την ανάπτυξη εφαρμογών. Χρησιμοποιούν το Python σε πραγματικές συνθήκες. Αυτό διασφαλίζει ότι η Python είναι στιβαρή, ασφαλής, σχετικά αποτελεσματική, φορητή, κλιμακούμενη και διαθέτει ένα σύνολο χαρακτηριστικών που ικανοποιεί τις πραγματικές ανάγκες και όχι αυτό που οι προμηθευτές πιστεύουν ότι πρέπει να έχουν οι πελάτες. (Hladun, 2020)

Από την άλλη τα μειονεκτήματα της είναι:

- Για να επεξεργαστούν όλοι ένα πρόγραμμα θα πρέπει να εγκαταστήσουν ένα νέο ολόκληρο περιβάλλον ανάπτυξης.
- Δεν υπάρχει τυπικό graphical user interface στην Python. Αντ 'αυτού, δανείζεται από αλλού. Το πλησιέστερο πρότυπο είναι το Tkinter, το οποίο παρέχεται με τη Python.
- Η Python είναι ευαίσθητη στα κεφαλαία γράμματα. Τα ονόματα μεταβλητών DOD, Dod και dod είναι διαφορετικά και μπορεί να έχουν διαφορετικές τιμές. (Parikh, 2018)

4.2. Περιγραφή Dataset

Όπως αναφέρθηκε και στην εισαγωγή της εργασίας, χρησιμοποιήθηκαν δύο datasets απο τα δύο μεγαλύτερα πρωταθλήματα μπάσκετ του κόσμου. Αυτά ειναι του NBA (National Basketball Association) και της Euroleague. Κάθε dataset περιέχει μια πληθώρα στατιστικών στοιχείων ενός αγώνα μπάσκετ, εκ των οποίων πολλά είναι κοινά.

Τα στατιστικά αυτά αναλύονται σε χρονιές (*Season*) και ανα ομάδα (*Team*). Ένα απο τα πιο σημαντικά χαρακτηριστικά που ειναι και η κλάση μας ειναι το αν πέρασε η εκάστοτε ομάδα στα playoff (*Qualified*).

Η χρονιά έναρξης του συνόλου δεδομένων ειναι 2005-06 με τελική το 2015-16. Για τις ανάγκες της εργασίας συμπληρώθηκαν οι τελευταίες χρονιές (εως και 2018-19) που έλειπαν με στοιχεία απο τον παρακάτω σύνδεσμο. Λόγω του COVID-19 η διοργάνωση της Euroleague δεν ανέδειξε πρωταθλητή , οπότε προτιμήθηκε και για το NBA να είναι τελική χρονιά η 2018-2019.

https://www.basketball-reference.com/international/euroleague/2018.html#all_team_stats_totals

EUROLEAGUE		
Season	Season	object
Qualified	Qualified to playoffs	int64
Team	Team name	object
PLAYER COUNT	Players count of the tean	int64
G	Games played	int64
MP	Minutes Played	int64
PIR	Performance Index Rating	int64
AVG PIR	Performance Index Rating Average	float64
PTS	Points	int64
Avg PTS	Points Average	float64
TRB	Total Rebounds	int64
AVG TRB	Total Rebounds Average	float64
ORB	Offensive Rebounds	int64
AVG ORB	Offensive Rebounds Average	float64
DRB	Defensive Rebounds	int64
AVG DRB	Defensive Rebounds Average	float64
AST	Assists	int64
AVG AST	Assists Average	float64
STL	Steals	int64
AVG STL	Steals Average	float64
BLK	Blocks	int64
AVG BLK	Blocks Average	float64
BLK against	BLK against	int64
AVG BLK against	BLK against Average	float64
TOV	Turnovers	int64
AVG TOV	Turnovers Average	float64
Fouls Drawn	Fouls Drawn	int64
Fouls Drawn Ave.	Fouls Drawn Average	float64
Fouls Commited	Fouls Commited	int64
Fouls Commited Ave.	Fouls Commited Average	float64
FTA	Free Throw Attempts	int64
FT	Free Throws	int64
FT%	Free Throw Percentage	float64
2PA	2-Point Field Goal Attempts	int64
2P	2-Point Field Goals	int64

2P%	**2-Point Field Goal Percentage**	**float64**
3PA	**3-Point Field Goal Attempts**	**int64**
3P	**3-Point Field Goals**	**int64**
3P%	**3-Point Field Goal Percentage**	**float64**
TS%	**True Shooting Percentage**	**float64**
TOV.1	**Turnovers**	**int64**
AST.1	**Assists**	**int64**
AST-TOVratio(%)	**Assists - Turnovers ratio percentage**	**float64**
AVG AGE	**Average AGE**	**float64**
AVG HT	**Average Height**	**float64**

Πίνακας 1 Euroleague

NBA		
Season	**Season**	**object**
Qualified	**Qualified to playoffs**	**int64**
Team	**Team name**	**object**
G	**Games played**	**int64**
MP	**Minutes Played**	**int64**
FG	**Field Goals**	**int64**
FGA	**Field Goal Attempts**	**int64**
FG%	**Field Goal Percentage**	**float64**
3P	**3-Point Field Goals**	**int64**
3PA	**3-Point Field Goal Attempts**	**int64**
3P%	**3-Point Field Goal Percentage**	**float64**
2P	**2-Point Field Goals**	**int64**
2PA	**2-Point Field Goal Attempts**	**int64**
2P%	**2-Point Field Goal Percentage**	**float64**
FT	**Free Throws**	**int64**
FTA	**Free Throw Attempts**	**int64**
FT%	**Free Throw Percentage**	**float64**
ORB	**Offensive Rebounds**	**int64**
DRB	**Defensive Rebounds**	**int64**
TRB	**Total Rebounds**	**int64**
AST	**Assists**	**int64**
STL	**Steals**	**int64**

BLK	**Blocks**	**int64**
TOV	**Turnovers**	**int64**
PF	**Personal Fouls**	**int64**
PTS	**Points**	**int64**
PLAYER COUNT	**Players count of the team**	**int64**
AVG HT	**Average Height**	**float64**
AVG AGE	**Average AGE**	**float64**
SALARY ($)	**Salary in dollars**	**int64**

Πίνακας 2 NBA

4.3. Συμπλήρωση missing values

Στο προηγούμενο κεφάλαιο αναφέρθηκε ότι συμπληρώθηκαν τα στατιστικά για κάποιες χρονιές. Παρ' ολ'αυτά υπήρχαν συγκεκριμένα χαρακτηριστικά ανά σύνολο δεδομένων, τα οποία δε βρέθηκαν οι αντίστοιχες τιμές. Τέτοιες είναι για παράδειγμα το πεδίο «AVG AGE» της Euroleague και το πεδίο «SALARY ($)» του NBA.

Για τις ανάγκες συμπλήρωσης των τιμών αυτών, χρησιμοποιήθηκαν οι παρακάτω τεχνικές :

Αριθμητικό μέσος (mean)

Ο αριθμητικός μέσος που ονομάζεται αλλιώς και μέσος όρος, είναι το άθροισμα των τιμών μιας ακολουθίας αριθμών διαιρούμενο με το συνολικό πλήθος των αριθμών αυτών.

Η εντολή python για να υπολογιστεί ο αριθμητικός μέσος είναι: *.mean()*

Διάμεσος (median)

Η διάμεσος είναι η μέση τιμή μιας ομάδας αριθμών που έχουν ταξινομηθεί κατά μέγεθος. Είναι ουσιαστικά εκείνος ο αριθμός που βρίσκεται στη μέση, έτσι ώστε οι μισοί (σε πλήθος) ταξινομημένοι αριθμοί να είναι πάνω από τη διάμεσο και άλλοι μισοί κάτω από αυτή.

Η εντολή python για να υπολογιστεί ο διάμεσος είναι: *.median()*

Πρόβλεψη με γραμμική παλινδρόμηση (Prediction with Liner Regression)

Επιχειρήθηκε να προβλεφθούν οι κενές τιμές με τη μέθοδο της γραμμικής παλινδρόμησης.

Η εντολή python για την πρόβλεψη με γραμμική παλινδρόμηση είναι : *LinearRegression της sklearn*

Συμπέρασμα

Έπειτα από πολλές δοκιμές και με τους 3 τρόπους , προτιμήθηκε ο πρώτος, αυτός του αριθμητικού μέσου (mean) γιατί είχε τα καλύτερα αποτελέσματα. Παρατηρήθηκε ότι κατά τη συμπλήρωση κενών με πρόβλεψη μέσω γραμμικής παλινδρόμησης για το πεδίο AVG_AGE η πρόβλεψη έδειχνε πολλούς παίχτες να είναι σε ηλικία 22-23ετών που ήταν και λάθος μιας και η μέση τιμή ήταν γύρω στο 26. Από τη στιγμή που η απόκλιση του αριθμητικού μέσου με τη διάμεσο ήταν μικρή, προτιμήθηκε αυτή του αριθμητικού μέσου που είναι και πιο αποτελεσματικός τρόπος σε στατιστικά αγώνων γενικότερα.

4.4. Πρόβλεψη αποτελέσματος (Qualified) με 5 κατηγοριοποιητές

Για τις ανάγκες της πρόβλεψης του αποτελέσματος των χρησιμοποιήθηκαν οι παρακάτω κατηγοριοποιητές. Έπειτα από κάθε κατηγοριοποιητή παρουσιάζεται η αντίστοιχη python εντολή.

- **Logistic Regression**
    - *from sklearn.linear_model import LogisticRegression*
- **K nearest neighbors**
    - *from sklearn.neighbors import KNeighborsClassifier*
- **Support Vector Machine**
    - *from sklearn.svm import SVC*
- **Random Forest**
    - *from sklearn.ensemble import RandomForestClassifier*
- **Multi-layer Perceptron**
    - *from sklearn.neural_network import MLPClassifier*

Η κατηγοριοποίηση πραγματοποιήθηκε ανά σύνολο δεδομένων για τρία διαφορετικά σενάρια που αναλύονται παρακάτω.

- Τυποποίηση (Standardization)
- Pearson Συσχέτιση (Pearson correlation)
- Χαρακτηριστικών επι τοις εκατό (%) (Average characteristics)

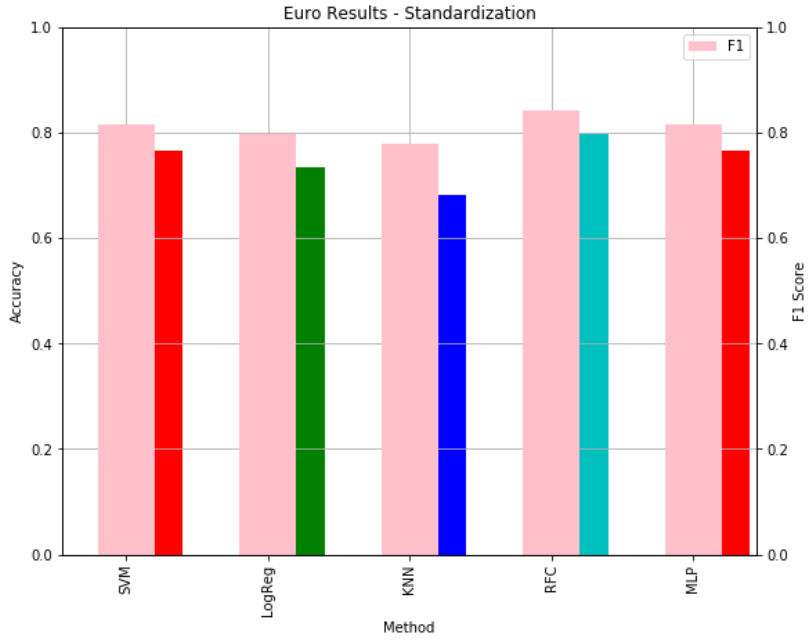
4.4.1 1ο σενάριο με τυποποίηση (standardization)

Σε αυτό το σενάριο μετατρέπουμε τα δεδομένα μας σε μια κοινή μορφή. Χρησιμοποιείται η python μέθοδος *StandardScaler* της *sklearn.preprocessing*.

Παρακάτω παρουσιάζονται οι πίνακες ανα σύνολο δεδομένων , με τα αντίστοιχα διαγράμματά τους για τους 5 κατηγοριοποιητές που έχουμε περιγράψει παραπάνω.

Classifier	Accuracy	F1-score	Precision	Recall
SVM	0.765957	0.816667	0.790323	0.844828
LogReg	0.734043	0.796748	0.753846	0.734043
KNN	0.680851	0.779412	0.679487	0.680851
RFC	0.797872	0.842975	0.809524	0.797872
MLP	0.765957	0.816667	0.790323	0.765957

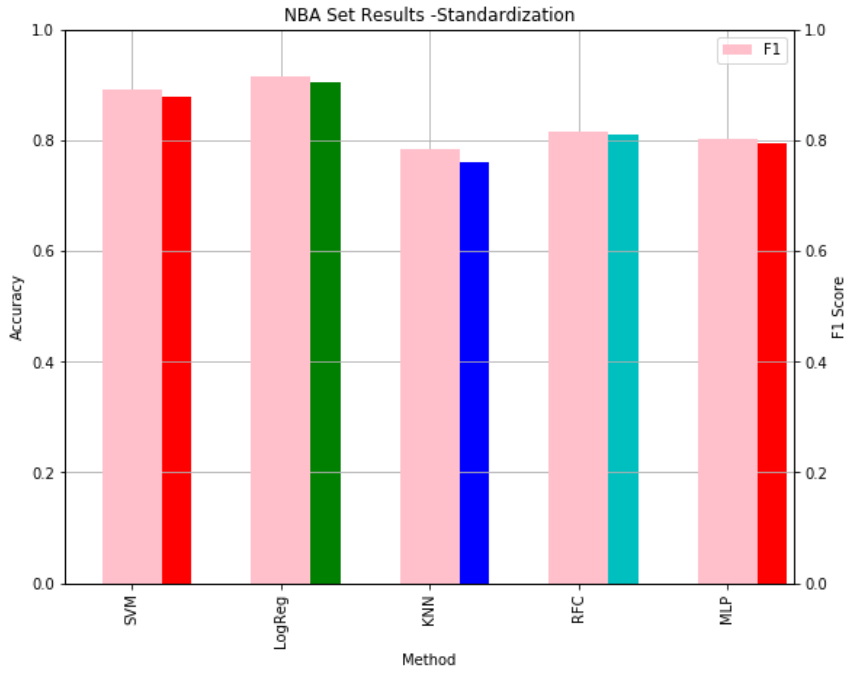
Πίνακας 3 Euroleague Standardization results



Εικόνα 8 Euroleague Standardization chart

Classifier	Accuracy	F1-score	Precision	Recall
SVM	0.880342	0.892308	0.865672	0.920635
LogReg	0.905983	0.914729	0.893939	0.905983
KNN	0.760684	0.784615	0.761194	0.760684
RFC	0.811966	0.816667	0.859649	0.811966
MLP	0.794872	0.803279	0.830508	0.794872

Πίνακας 4 NBA Standardization results



Εικόνα 9 NBA Standardization chart

Συμπέρασμα 1ου σεναρίου

Euroleague

Ο πιο αποδοτικός κατηγοριοποιητής είναι ο “Random Forest” που σε όλες τις μετρικές του «Πίνακας 3» δείχνει ότι έχει τη μεγαλύτερη τιμή με μια μικρή υπεροχή του “SVM” στη μετρική “recall”. Από την άλλη πλευρά ο χειρότερος κατηγοριοποιητής δείχνει να ειναι ο “K nearest neighbors”(KNN).

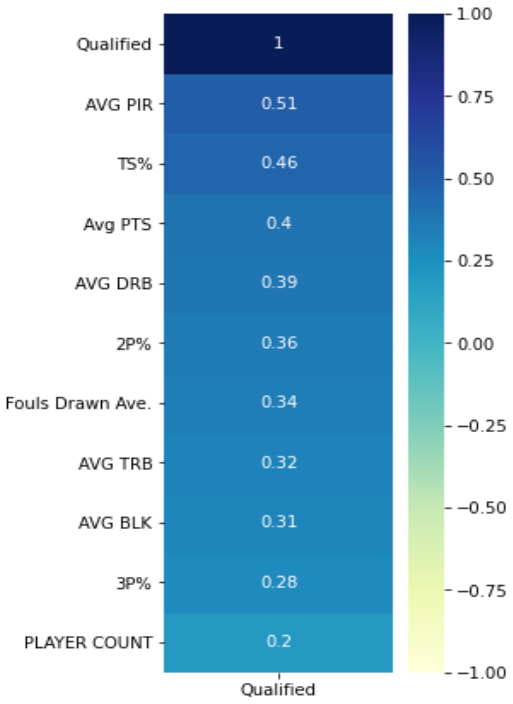
NBA

Ο πιο αποδοτικός κατηγοριοποιητής είναι ο “Logistic Regression” που σε όλες τις μετρικές του «Πίνακας 4» δείχνει ότι έχει τη μεγαλύτερη τιμή με μια μικρή υπεροχή και εδώ του “SVM” στη μετρική “recall”. Από την άλλη πλευρά ο χειρότερος κατηγοριοποιητής δείχνει πάλι να είναι ο “K nearest neighbors”(KNN).

4.4.2 2ο σενάριο με συσχέτιση Pearson (Pearson correlation)

Σε αυτό το σενάριο χρησιμοποιείται η συσχέτιση Pearson, ώστε να βρούμε τα 10 πιο συσχετιζόμενα χαρακτηριστικά με την κλάση του συνόλου δεδομένων που είναι το πεδίο «Qualified».

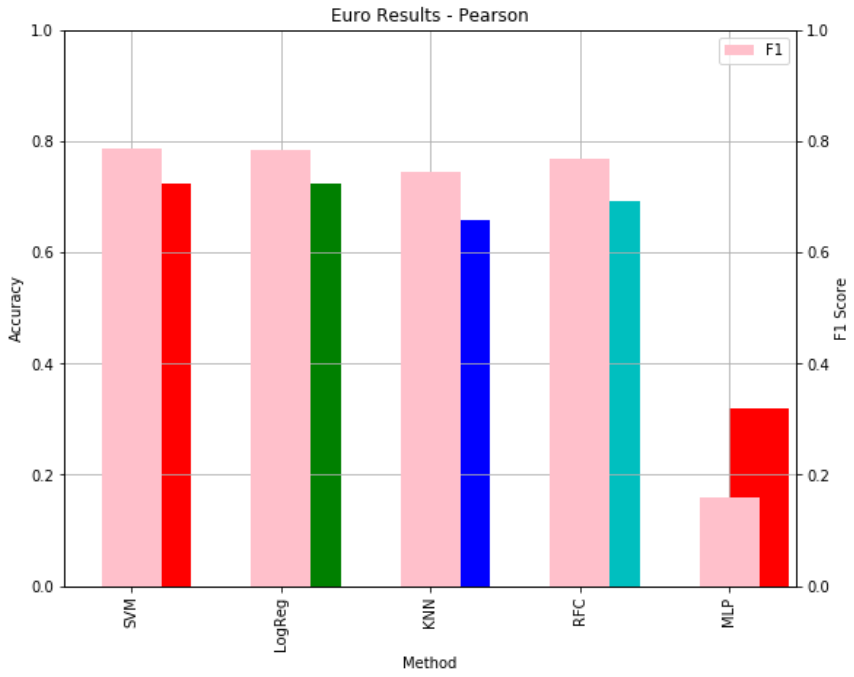
Παρακάτω παρουσιάζονται οι πίνακες ανά σύνολο δεδομένων , με τα αντίστοιχα διαγράμματά τους για τους 5 κατηγοριοποιητές:



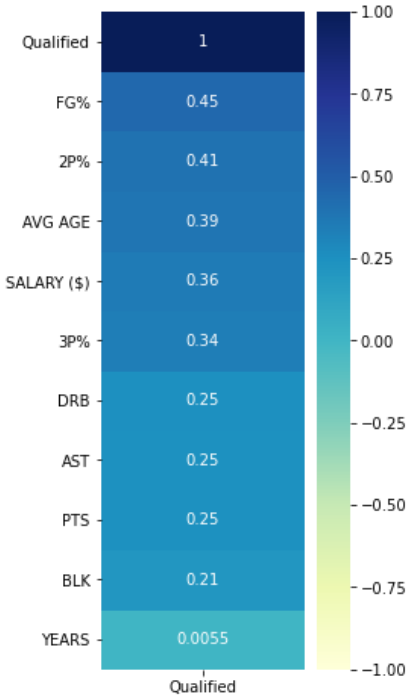
Εικόνα 10 Euroleague heatmap

Classifier	Accuracy	F1-score	Precision	Recall
SVM	0.723404	0.786885	0.872727	0.716418
LogReg	0.723404	0.783333	0.886792	0.723404
KNN	0.659574	0.746032	0.79661	0.659574
RFC	0.691489	0.768000	0.827586	0.691489
MLP	0.319149	0.157895	0.666667	0.319149

Πίνακας 5 Euroleague Pearson Correlation Results



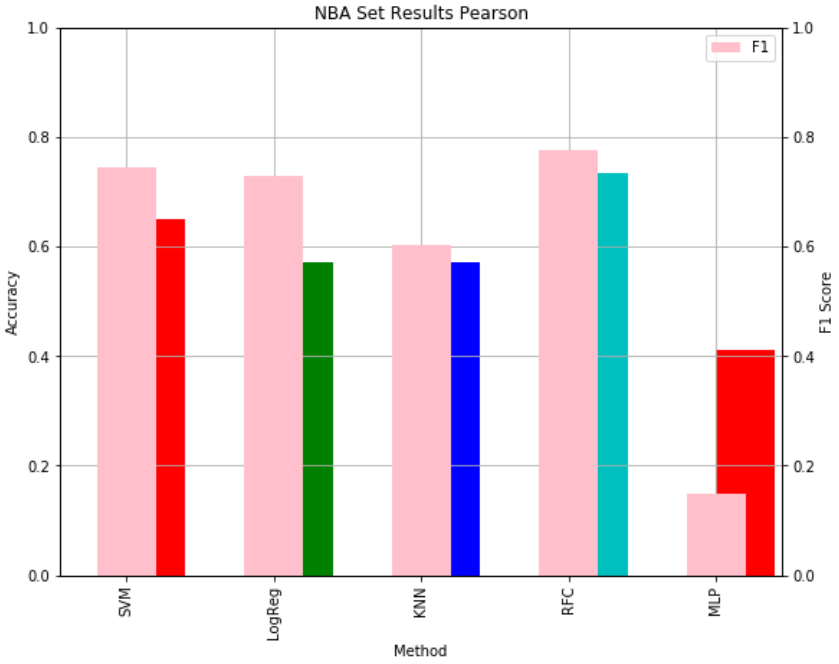
Εικόνα 11 Euroleague Pearson Correlation Chart



Εικόνα 12 NBA heatmap

Classifier	Accuracy	F1-score	Precision	Recall
SVM	0.649573	0.745342	0.638298	0.895522
LogReg	0.572650	0.728261	0.572650	0.572650
KNN	0.572650	0.603175	0.644068	0.572650
RFC	0.735043	0.776978	0.750000	0.735043
MLP	0.410256	0.148148	0.428571	0.410256

Πίνακας 6 NBA Pearson Correlation Results



Εικόνα 13 NBA Pearson Correlation Chart

Συμπέρασμα 2ου σεναρίου

Euroleague

Απο τις μετρικές του «Πίνακας 5» φαίνεται ότι η 4αδα κατηγοριοποιητών SVM,LogReg,KNN,RFC να έχουν την ίδια απόδοση με ελαφρά υπεροχή της Λογιστικής Παλινδρόμησης. Από την άλλη πλευρά ο χειρότερος κατηγοριοποιητής με διαφορά ειναι ο “Multi-layer Perceptron”(MLP).

NBA

Ο πιο αποδοτικός κατηγοριοποιητής είναι ο “Random Forest” που σε όλες τις μετρικές του «Πίνακας 6Πίνακας 4» δείχνει ότι έχει τη μεγαλύτερη τιμή με μια μικρή υπεροχή του “SVM” στη μετρική “recall”.

Από την άλλη πλευρά ο χειρότερος κατηγοριοποιητής πάλι με διαφορά ειναι ο “Multi-layer Perceptron”(MLP).

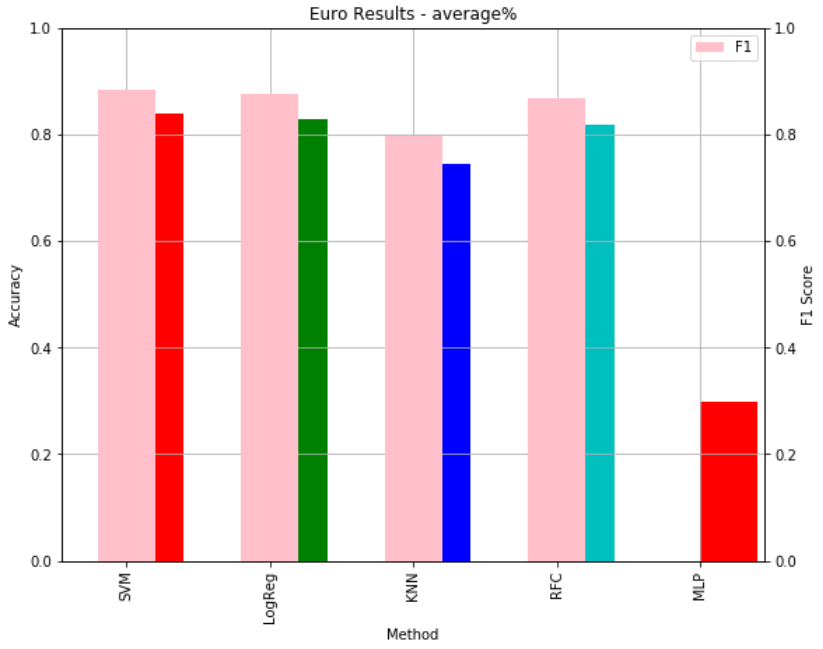
4.4.3 3ο σενάριο με χαρακτηριστικά επι τοις εκατό (%) (Average characteristics)

Σε αυτό το σενάριο χρησιμοποιούνται όλα εκείνα τα χαρακτηριστικά που είναι επι τις 100. Επιλέχθηκαν μόνο αυτά τα χαρακτηριστικά , ως τα πιο αντιπροσωπευτικά για δεδομένα αγώνων μπάσκετ.

Παρακάτω παρουσιάζονται οι πίνακες ανά σύνολο δεδομένων , με τα αντίστοιχα διαγράμματά τους για τους 5 κατηγοριοποιητές:

Classifier	Accuracy	F1-score	Precision	Recall
SVM	0.840426	0.885496	0.892308	0.878788
LogReg	0.829787	0.876923	0.890625	0.829787
KNN	0.744681	0.796610	0.903846	0.744681
RFC	0.819149	0.868217	0.888889	0.819149
MLP	0.297872	0.000000	0.000000	0.297872

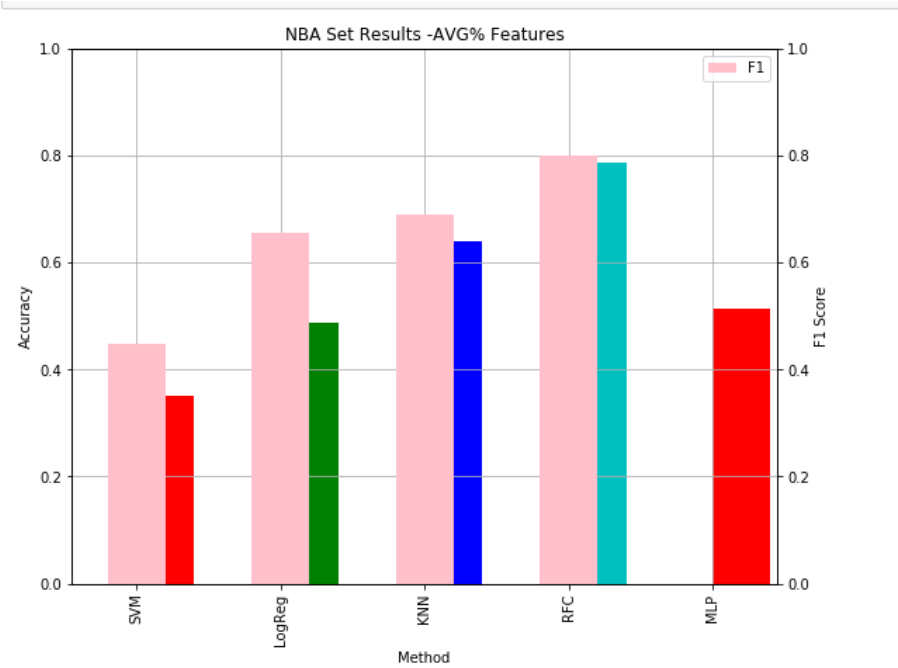
Πίνακας 7 Euroleague AVG Results



Εικόνα 14 Euroleague AVG chart

Classifier	Accuracy	F1-score	Precision	Recall
SVM	0.350427	0.449275	0.382716	0.543860
LogReg	0.487179	0.655172	0.487179	0.487179
KNN	0.641026	0.691176	0.594937	0.641026
RFC	0.786325	0.800000	0.735294	0.786325
MLP	0.512821	0.000000	0.000000	0.512821

Πίνακας 8 NBA AVG Results



Εικόνα 15 NBA AVG chart

Συμπέρασμα 3ου σεναρίου

Euroleague

Από τις μετρικές του «Πίνακας 7Πίνακας 5» φαίνεται ότι η 4αδα κατηγοριοποιητών SVM,LogReg,KNN,RFC να έχουν την ίδια απόδοση με ελαφρά υπεροχή του “Support Vector Machine”. Από την άλλη πλευρά ο χειρότερος κατηγοριοποιητής με διαφορά είναι ο “Multi-layer Perceptron”(MLP).

NBA

Ο πιο αποδοτικός κατηγοριοποιητής είναι ο “Random Forest” που σε όλες τις μετρικές του «Πίνακας 8Πίνακας 4» δείχνει ότι έχει τις μεγαλύτερες τιμές . Από την άλλη ο SVM και MLP είναι οι 2 χειρότεροι κατηγοριοποιητές.

4.4.4 Συμπέρασμα τριών παραπάνω σεναρίων

Σε αυτό το κεφάλαιο θα παρουσιάσουμε το πιο αποδοτικό σενάριο από τα παραπάνω ανά dataset.

Euroleague

Όσο αναφορά τη Euroleague το 3ο σενάριο (AVG characteristics) φαίνεται να είναι ελαφρά πιο αποδοτικό από το 1ο σενάριο (standardization). Εάν εξαιρεθεί ο “Multi-layer Perceptron”(MLP) που ειναι με διαφορά ο χειρότερος του 3ου σεναρίου, τότε οι υπόλοιποι 4 κατηγοριοποιητές (SVM,LogReg,KNN,RFC) είναι πιο αποδοτικοί σε σύγκριση με το 1ο σενάριο. Βέβαια η διαφορά τους είναι της τάξεως των 6 έως 8 μονάδων , που σημαίνει ότι η απόδοση τους είναι παρόμοια. ΄

Για παράδειγμα το accuracy του SVM για το 1ο σενάριο είναι 0.765957 , ενώ για το 3ο σενάριο είναι 0.840426. ή το precision του RFC για το 1ο σενάριο είναι 0.809524, ενώ για το 3ο είναι 0.888889.

Από την άλλη πλευρά το 2ο σενάριο δεν είναι και πολύ αποδοτικό μιας και οι μετρικές μας κυμαίνονται στο 0.65 με 0.70.

NBA

Όσο αναφορά το NBA το 1ο σενάριο (standardization) φαίνεται να είναι το πιο αποδοτικό με διαφορά από τα άλλα δύο σενάρια. Με μετρικές κοντά στο 0.8500 δείχνει να υπερτερεί του 2ου σεναρίου που οι μετρικές είναι στο 0.6080 και του 3ου σεναρίου στο 0,5272.

Ενδεικτικό παράδειγμα είναι ότι για τη λογαριθμική παλινδρόμηση(LogReg) το F1-Score είναι 0.914729 για το 1ο σενάριο , ενώ για το 2ο είναι στο 0,72 και για το 3ο στο 0,65. Επιπλέον το accuracy του SVM για το 1ο σενάριο είναι 0.880342 , ενώ για το 2ο είναι 0.649573 και για το 3ο 0.350427.

Σύνοψη

Το 1ο σενάριο με standardization δείχνει να είναι γενικά και στα 2 dataset αποδοτικό, με τη μόνη διαφορά ότι στο πρωτάθλημα της Euroleague το σενάριο με τα AVG χαρακτηριστικά να υπερέχει. Αυτό οφείλεται στο γεγονός ότι στο NBA τα ποσοστά γενικά είναι πιο υψηλά , μιας και οι άμυνες δεν είναι τόσο “σφιχτές” όσο στην Ευρώπη.

4.5. Πρόβλεψη αποτελέσματος για τις τελευταίες 5 χρονιές

Σε αυτό το κεφάλαιο επιχειρείται να προβλεφθεί το αποτέλεσμα των τελευταίων 5 ετών για κάθε σύνολο δεδομένων. Όπως και στο κεφάλαιο 0 , η πρόβλεψη πραγματοποιήθηκε με τους ίδιους 5 κατηγοριοποιητές (SVM,LogReg,KNN,RFC,MLP) για τις παρακάτω χρονιές :

- 2014-2015
- 2015-2016
- 2016-2017
- 2017-2018
- 2018-2019

Ουσιαστικά για τις ανάγκες της πρόβλεψης, οι παραπάνω χρονιές χαρακτηρίζονται ως το test set μας , ενώ όλες οι άλλες χρονιές ως το training set μας. Έγινε fit για τον κάθε αλγόριθμο ξεχωριστά και πραγματοποιήθηκε πρόβλεψη πρώτα για τη χρονιά 2014-2015. Μετά με τον ίδιο "fitαρισμένο" αλγόριθμο έγινε η πρόβλεψη για τη χρονιά 2015-2016 κ.ο.κ. Η πρόβλεψη πραγματοποιήθηκε για τα 3 ίδια σενάρια που αναφέρθηκαν στα κεφάλαια 4.4.1 , 4.4.2 , 4.4.3. Επιχειρήθηκε κάθε χρονιά για την οποία κάναμε πρόβλεψη, να την επισυνάπτουμε στο αρχικό μας training set ώστε να εκπαιδεύουμε τον αλγόριθμό μας, αλλά παρατηρήσαμε ότι τα αποτελέσματα δεν ήταν τόσο αποδοτικά, για αυτό και προτιμήθηκε και για τις 5 χρονιές να χρησιμοποιηθεί το ίδιο εκπαιδευμένο μοντέλο.

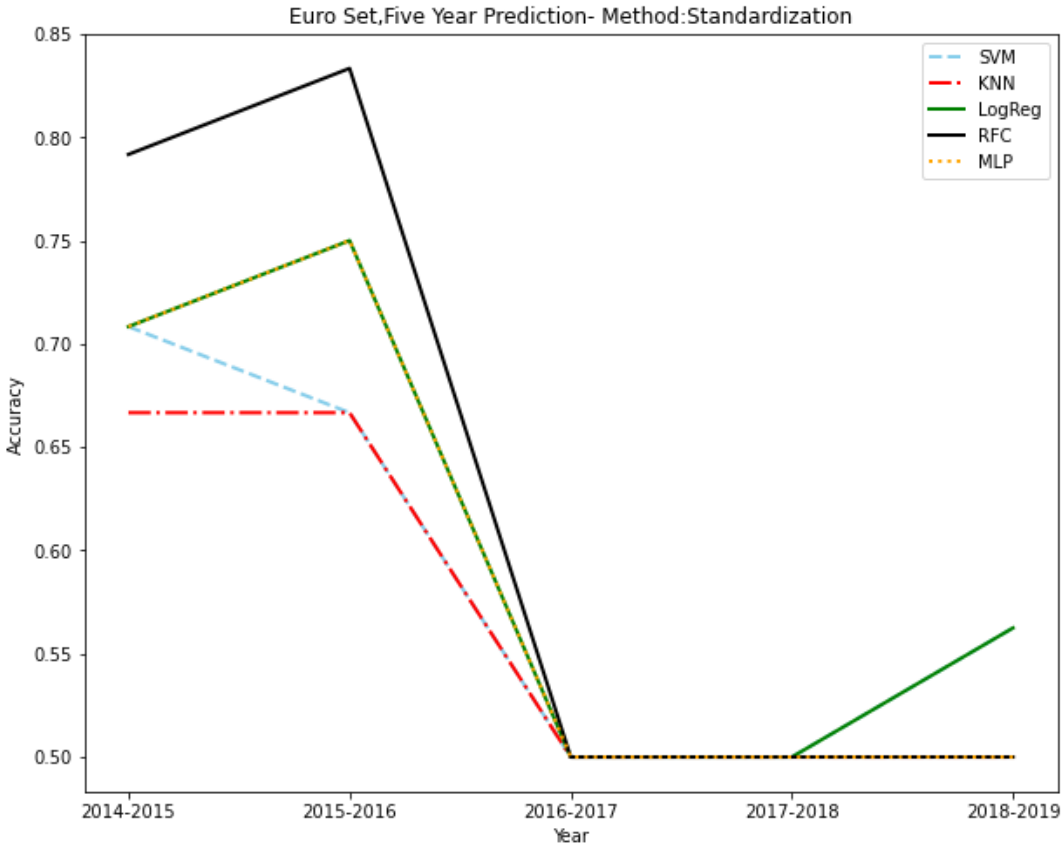
4.5.1 1ο σενάριο με τυποποίηση (standardization)

Σε αυτό το σενάριο κάνουμε πρόβλεψη των τελευταίων 5 σεζόν κάνοντας προεπεξεργασία με standardization.

Παρακάτω παρουσιάζονται οι πίνακες ανά σύνολο δεδομένων , με τα αντίστοιχα διαγράμματά τους για τους 5 κατηγοριοποιητές:

Classifier	Year	Accuracy	F1-score	Precision	Recall
SVM	2014-2015	0.708333333	0.8	0.736842	0.875
	2015-2016	0.666666667	0.75	0.75	0.75
	2016-2017	0.5	0.666667	0.5	1
	2017-2018	0.5	0.666667	0.5	1
	2018-2019	0.5625	0.695652	0.533333	1
LogReg	2014-2015	0.708333333	0.810811	0.714286	0.9375
	2015-2016	0.75	0.823529	0.777778	0.875
	2016-2017	0.5	0.666667	0.5	1
	2017-2018	0.5	0.666667	0.5	1
	2018-2019	0.5625	0.695652	0.533333	1
KNN	2014-2015	0.666666667	0.8	0.666667	1
	2015-2016	0.666666667	0.777778	0.7	0.875
	2016-2017	0.5	0.666667	0.5	1
	2017-2018	0.5	0.666667	0.5	1
	2018-2019	0.5	0.666667	0.5	1
RFC	2014-2015	0.791666667	0.864865	0.761905	1
	2015-2016	0.875	0.914286	0.842105	1
	2016-2017	0.5	0.666667	0.5	1
	2017-2018	0.5	0.666667	0.5	1
	2018-2019	0.5	0.666667	0.5	1
MLP	2014-2015	0.708333333	0.810811	0.714286	0.9375
	2015-2016	0.75	0.823529	0.777778	0.875
	2016-2017	0.5	0.666667	0.5	1
	2017-2018	0.5	0.666667	0.5	1
	2018-2019	0.5	0.666667	0.5	1

Πίνακας 9 Euroleague Standardization results 5 years prediction

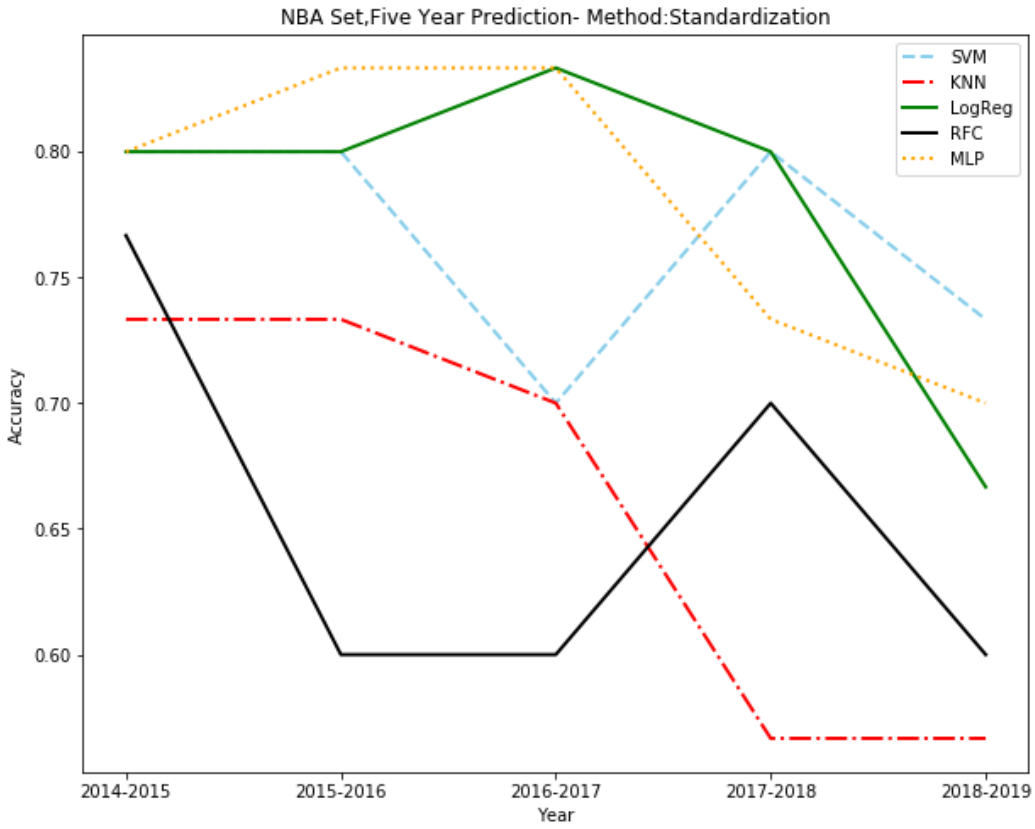


Εικόνα 16 Euroleague Standardization chart 5 years prediction

Classifier	Year	Accuracy	F1-score	Precision	Recall
SVM	2014-2015	0.8	0.82352941	0.777777778	0.875
	2015-2016	0.8	0.82352941	0.777777778	0.875
	2016-2017	0.7	0.68965517	0.769230769	0.625
	2017-2018	0.8	0.8125	0.8125	0.8125
	2018-2019	0.733333333	0.75	0.75	0.75
LogReg	2014-2015	0.8	0.82352941	0.777777778	0.875
	2015-2016	0.8	0.82352941	0.777777778	0.875
	2016-2017	0.833333333	0.85714286	0.789473684	0.9375
	2017-2018	0.8	0.83333333	0.75	0.9375
	2018-2019	0.666666667	0.73684211	0.636363636	0.875
KNN	2014-2015	0.733333333	0.76470588	0.722222222	0.8125
	2015-2016	0.733333333	0.75	0.75	0.75
	2016-2017	0.7	0.72727273	0.705882353	0.75
	2017-2018	0.566666667	0.62857143	0.578947368	0.6875
	2018-2019	0.566666667	0.66666667	0.565217391	0.8125

RFC	2014-2015	0.766666667	0.8	0.736842105	0.875
	2015-2016	0.6	0.66666667	0.6	0.75
	2016-2017	0.6	0.7	0.583333333	0.875
	2017-2018	0.7	0.78048781	0.64	1
	2018-2019	0.6	0.72727273	0.571428571	1
MLP	2014-2015	0.8	0.8125	0.8125	0.8125
	2015-2016	0.833333333	0.83870968	0.866666667	0.8125
	2016-2017	0.833333333	0.86486487	0.761904762	1
	2017-2018	0.733333333	0.77777778	0.7	0.875
	2018-2019	0.7	0.76923077	0.652173913	0.9375

Πίνακας 10 NBA Standardization results 5 years prediction



Εικόνα 17 NBA Standardization chart 5 years prediction

Συμπέρασμα 1ου σεναρίου

Euroleague

Ο πιο αποδοτικός κατηγοριοποιητής είναι ο “Random Forest” που σε όλες τις μετρικές του «Πίνακας 9» δείχνει ότι έχει τη μεγαλύτερη τιμή . Σαν 2ο αποδοτικότερο κατηγοριοποιητή παρατηρούμε τον SVM και

3ο τον LogReg με μικρή διαφορά. Παρατηρούμε επίσης μια ασυνήθιστη συμπεριφορά όλων των αλγορίθμων. Όσο περνούσαν τα χρόνια η ακρίβεια των προβλέψεων πέφτει στο μισό με εξαίρεση τον LogReg που ανέβηκε την τελευταία χρονιά. Αυτό συμβαίνει γιατί απομακρυνόμαστε από τις χρονιές που εκπαιδεύσαμε το μοντέλο μας και δείχνει ότι γενικά οι ομάδες αλλάζουν κατά πολύ , ως εκ τούτου αλλάζουν και τα στατιστικά τους.

NBA

Οι δύο πιο αποδοτικοί κατηγοριοποιητές με avg accuracy στο 0.78 είναι ο “Logistic Regression” και ο “Multi-Layer Perceptron” . Από τον «Πίνακας 10» φαίνεται ότι οι διαφορές τους είναι ελάχιστες. Από την άλλη πλευρά ο χειρότερος κατηγοριοποιητής δείχνει να είναι ο “K nearest neighbors”(KNN) με avg accuracy στο 0.66 αλλά με τις υπόλοιπες μετρικές να υστερούν.

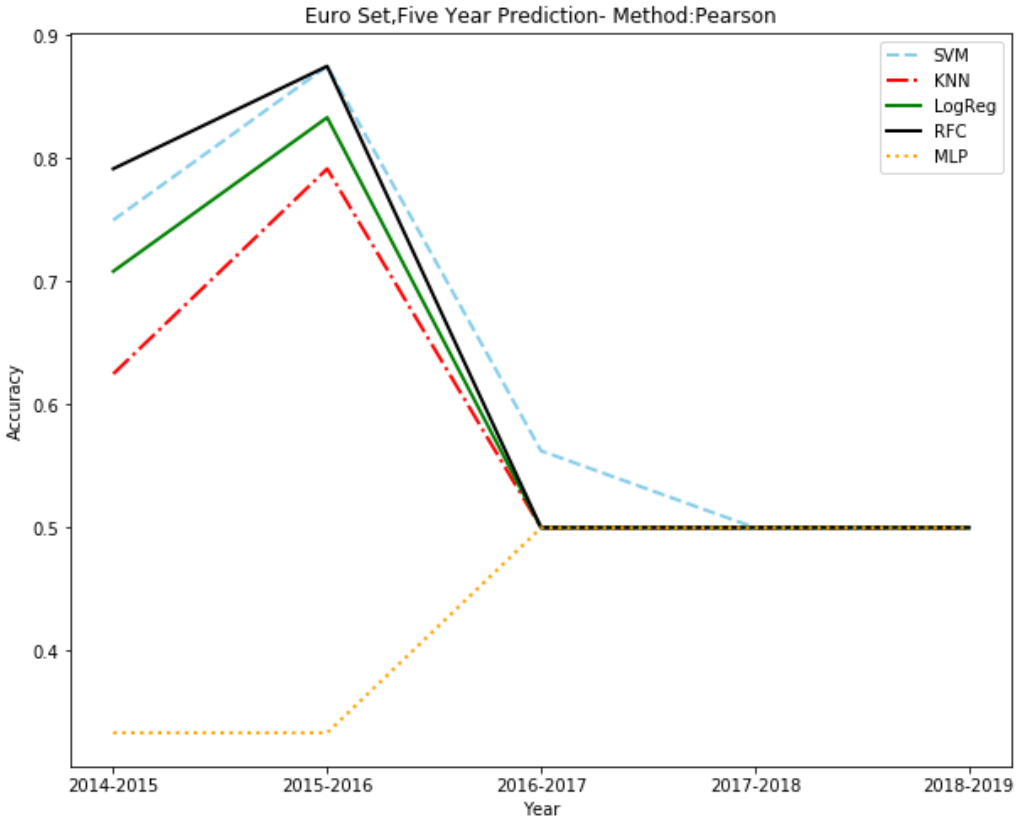
4.5.2 2ο σενάριο με συσχέτιση Pearson (Pearson correlation)

Σε αυτό το σενάριο χρησιμοποιείται η συσχέτιση Pearson, ώστε να βρούμε τα 10 πιο σχετιζόμενα χαρακτηριστικά (*Εικόνα 10 Euroleague heatmap &* Εικόνα 12 NBA heatmap) με την κλάση του συνόλου δεδομένων που είναι το πεδίο «Qualified». Έπειτα πραγματοποιείται πρόβλεψη των τελευταίων 5 σεζόν βάσει αυτών των 10 χαρακτηριστικών.

Παρακάτω παρουσιάζονται οι πίνακες ανά σύνολο δεδομένων , με τα αντίστοιχα διαγράμματά τους για τους 5 κατηγοριοποιητές:

Classifier	Year	Accuracy	F1-score	Precision	Recall
SVM	2014-2015	0.75	0.8125	0.8125	0.8125
	2015-2016	0.875	0.909090909	0.882353	0.9375
	2016-2017	0.5625	0.695652174	0.533333	1
	2017-2018	0.5	0.666666667	0.5	1
	2018-2019	0.5	0.666666667	0.5	1
LogReg	2014-2015	0.708333333	0.787878788	0.764706	0.8125
	2015-2016	0.833333333	0.875	0.875	0.875
	2016-2017	0.5	0.666666667	0.5	1
	2017-2018	0.5	0.666666667	0.5	1
	2018-2019	0.5	0.666666667	0.5	1
KNN	2014-2015	0.625	0.756756757	0.666667	0.875
	2015-2016	0.791666667	0.857142857	0.789474	0.9375
	2016-2017	0.5	0.666666667	0.5	1
	2017-2018	0.5	0.666666667	0.5	1
	2018-2019	0.5	0.666666667	0.5	1
RFC	2014-2015	0.791666667	0.864864865	0.761905	1
	2015-2016	0.875	0.914285714	0.842105	1
	2016-2017	0.5	0.666666667	0.5	1
	2017-2018	0.5	0.666666667	0.5	1
	2018-2019	0.5	0.666666667	0.5	1
MLP	2014-2015	0.333333333	0	0	0
	2015-2016	0.333333333	0	0	0
	2016-2017	0.5	0	0	0
	2017-2018	0.5	0	0	0
	2018-2019	0.5	0	0	0

Πίνακας 11 Euroleague Pearson Correlation Results 5 years prediction

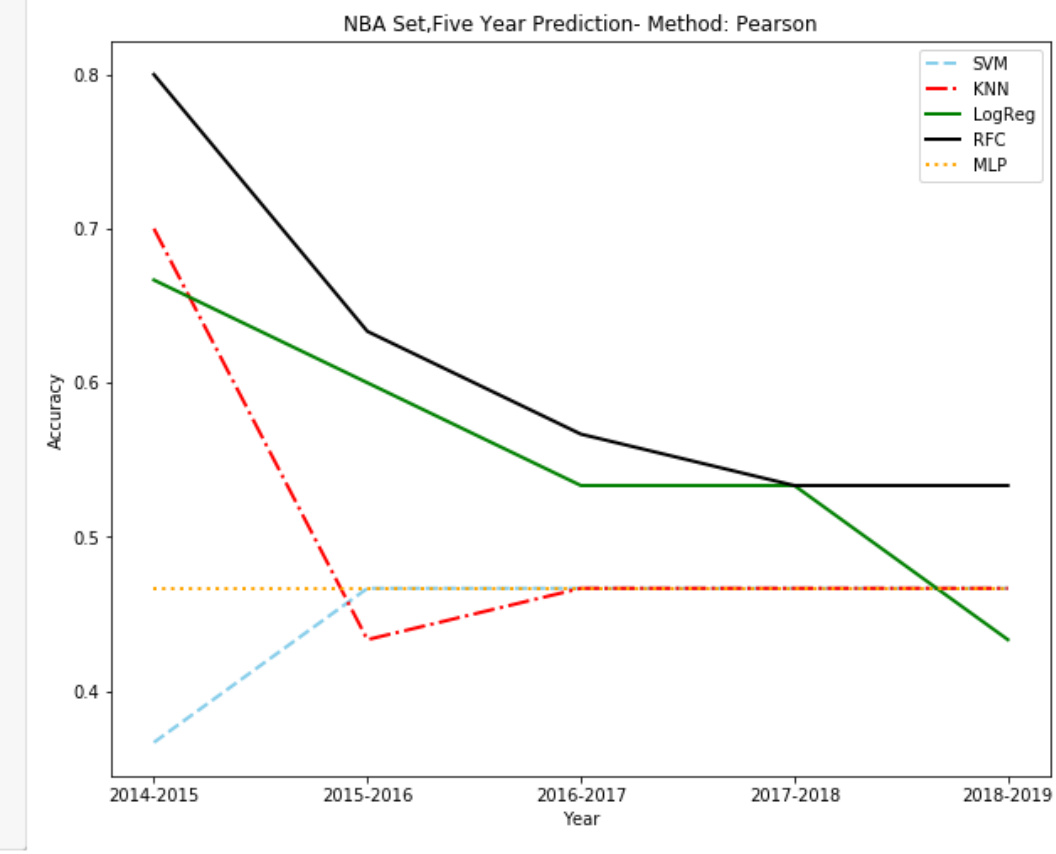


Εικόνα 18 Euroleague Pearson Correlation Chart 5 years prediction

Classifier	Year	Accuracy	F1-score	Precision	Recall
SVM	2014-2015	0.366666667	0.173913043	0.285714286	0.125
	2015-2016	0.466666667	0.2	0.5	0.125
	2016-2017	0.466666667	0	0	0
	2017-2018	0.466666667	0	0	0
	2018-2019	0.466666667	0	0	0
LogReg	2014-2015	0.666666667	0.736842105	0.636363636	0.875
	2015-2016	0.6	0.7	0.583333333	0.875
	2016-2017	0.533333333	0.695652174	0.533333333	1
	2017-2018	0.533333333	0.695652174	0.533333333	1
	2018-2019	0.433333333	0.564102564	0.47826087	0.6875
KNN	2014-2015	0.7	0.742857143	0.684210526	0.8125
	2015-2016	0.433333333	0.514285714	0.473684211	0.5625
	2016-2017	0.466666667	0	0	0
	2017-2018	0.466666667	0	0	0
	2018-2019	0.466666667	0	0	0

RFC	2014-2015	0.8	0.842105263	0.727272727	1
	2015-2016	0.633333333	0.731707317	0.6	0.9375
	2016-2017	0.566666667	0.711111111	0.551724138	1
	2017-2018	0.533333333	0.695652174	0.533333333	1
	2018-2019	0.533333333	0.695652174	0.533333333	1
MLP	2014-2015	0.466666667	0	0	0
	2015-2016	0.466666667	0	0	0
	2016-2017	0.466666667	0	0	0
	2017-2018	0.466666667	0	0	0
	2018-2019	0.466666667	0	0	0

Πίνακας 12 NBA Pearson Correlation Results 5 years prediction



Εικόνα 19 NBA Pearson Correlation Chart 5 years prediction

Συμπέρασμα 2ου σεναρίου

Euroleague

Ο πιο αποδοτικός κατηγοριοποιητής , όπως και στο 1ο σενάριο, είναι ο “Random Forest”, που σε όλες τις μετρικές του «Πίνακας 11Πίνακας 9» δείχνει ότι έχει τη μεγαλύτερη τιμή . Σαν 2ο αποδοτικότερο

κατηγοριοποιητή πάλι συναντάμε τον SVM και 3ο τον LogReg με μικρή διαφορά. Η ίδια ασυνήθιστη συμπεριφορά όλων των αλγορίθμων παρατηρείται και σε αυτό το σενάριο. Όσο περνούσαν τα χρόνια η ακρίβεια των προβλέψεων πέφτει στο μισό με εξαίρεση τον MLP που ουσιαστικά ήταν πιο κάτω από 0.5 και ανέβηκε την χρονιά 2016-2017.

NBA

Ο πιο αποδοτικός κατηγοριοποιητής αυτού του σεναρίου είναι “Random Forest”. Από τον «Πίνακας 12» φαίνεται ότι ελαφρά πιο πίσω βρίσκεται ο LogReg. Από την άλλη πλευρά οι χειρότεροι κατηγοριοποιητές είναι ο SVM και MLP με πολύ μεγάλη διαφορά. Για την ακρίβεια τα συγκεκριμένα 2 μοντέλα δεν αποδίδουν καθόλου καλά.

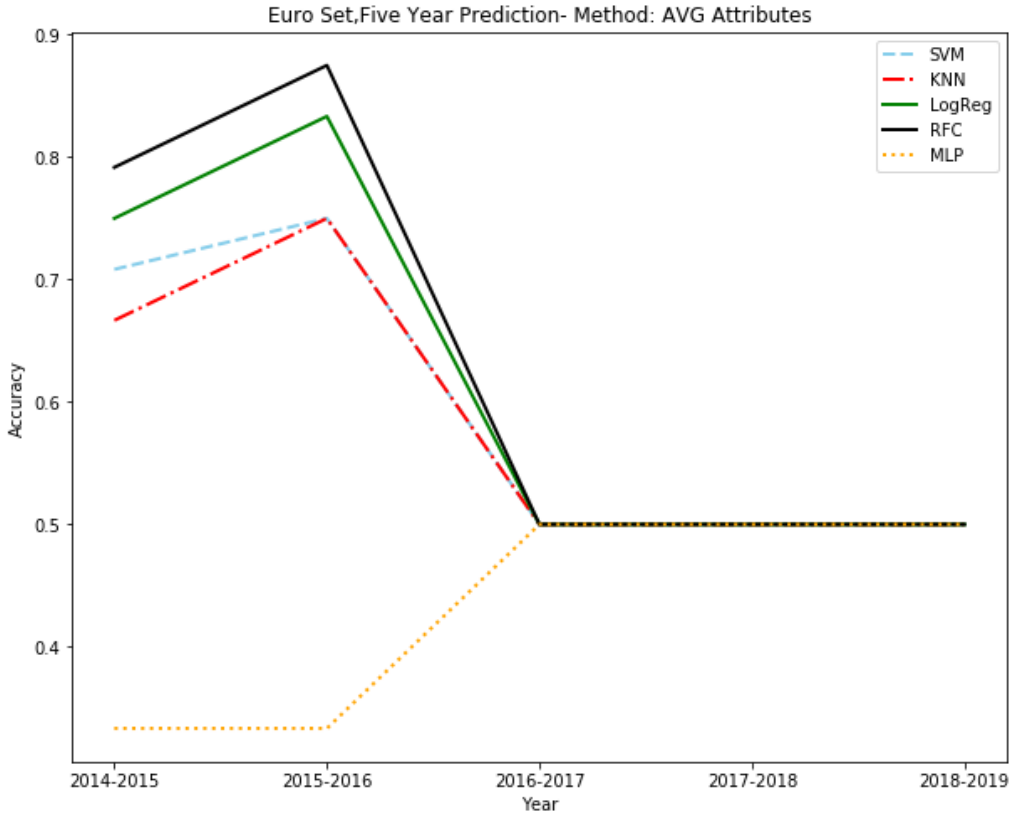
4.5.3 3ο σενάριο με χαρακτηριστικά επι τοις εκατό (%) (Average characteristics)

Σε αυτό το σενάριο χρησιμοποιούνται όλα εκείνα τα χαρακτηριστικά που είναι επι τις 100. Επιλέχθηκαν μόνο αυτά τα χαρακτηριστικά , ως τα πιο αντιπροσωπευτικά για δεδομένα αγώνων μπάσκετ. Έπειτα πραγματοποιείται πρόβλεψη των τελευταίων 5 σεζόν βάσει αυτών των χαρακτηριστικών.

Παρακάτω παρουσιάζονται οι πίνακες ανά σύνολο δεδομένων , με τα αντίστοιχα διαγράμματά τους για τους 5 κατηγοριοποιητές:

Classifier	Year	Accuracy	F1-score	Precision	Recall
SVM	2014-2015	0.708333333	0.787878788	0.7647059	0.8125
	2015-2016	0.75	0.8125	0.8125	0.8125
	2016-2017	0.5	0	0	0
	2017-2018	0.5	0	0	0
	2018-2019	0.5	0	0	0
LogReg	2014-2015	0.75	0.8125	0.8125	0.8125
	2015-2016	0.833333333	0.875	0.875	0.875
	2016-2017	0.5	0	0	0
	2017-2018	0.5	0	0	0
	2018-2019	0.5	0	0	0
KNN	2014-2015	0.666666667	0.8	0.6666667	1
	2015-2016	0.75	0.842105263	0.7272727	1
	2016-2017	0.5	0.666666667	0.5	1
	2017-2018	0.5	0.666666667	0.5	1
	2018-2019	0.5	0.666666667	0.5	1
RFC	2014-2015	0.791666667	0.864864865	0.7619048	1
	2015-2016	0.875	0.914285714	0.8421053	1
	2016-2017	0.5	0.666666667	0.5	1
	2017-2018	0.5	0.666666667	0.5	1
	2018-2019	0.5	0.666666667	0.5	1
MLP	2014-2015	0.333333333	0	0	0
	2015-2016	0.333333333	0	0	0
	2016-2017	0.5	0	0	0
	2017-2018	0.5	0	0	0
	2018-2019	0.5	0	0	0

Πίνακας 13 Euroleague AVG Results 5 years prediction

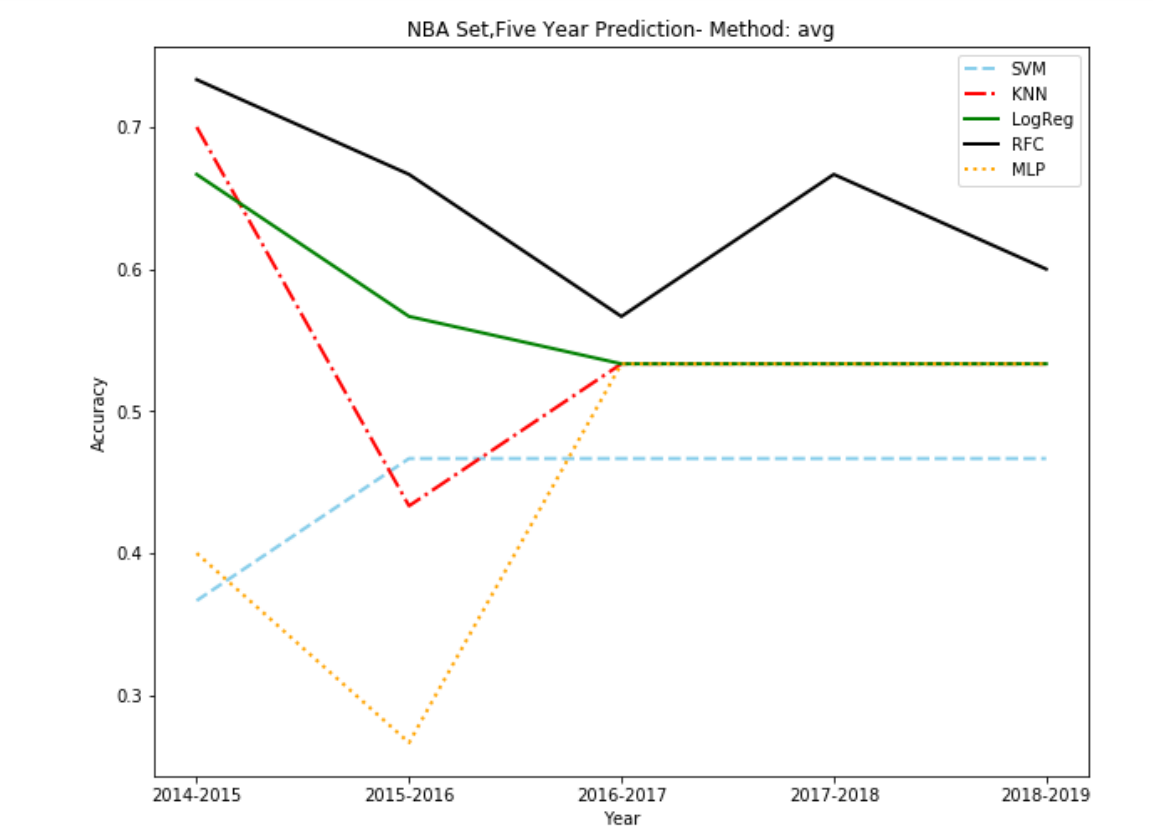


Εικόνα 20 Euroleague AVG chart 5 years prediction

Classifier	Year	Accuracy	F1-score	Precision	Recall
SVM	2014-2015	0.366666667	0.173913043	0.285714286	0.125
	2015-2016	0.466666667	0.2	0.5	0.125
	2016-2017	0.466666667	0	0	0
	2017-2018	0.466666667	0	0	0
	2018-2019	0.466666667	0	0	0
LogReg	2014-2015	0.666666667	0.736842105	0.636363636	0.875
	2015-2016	0.566666667	0.682926829	0.56	0.875
	2016-2017	0.533333333	0.695652174	0.533333333	1
	2017-2018	0.533333333	0.695652174	0.533333333	1
	2018-2019	0.533333333	0.695652174	0.533333333	1
KNN	2014-2015	0.7	0.742857143	0.684210526	0.8125
	2015-2016	0.433333333	0.514285714	0.473684211	0.5625
	2016-2017	0.533333333	0.695652174	0.533333333	1

	2017-2018	0.533333333	0.695652174	0.533333333	1
	2018-2019	0.533333333	0.695652174	0.533333333	1
RFC	2014-2015	0.733333333	0.777777778	0.7	0.875
	2015-2016	0.666666667	0.75	0.625	0.9375
	2016-2017	0.566666667	0.697674419	0.555555556	0.9375
	2017-2018	0.666666667	0.761904762	0.615384615	1
	2018-2019	0.6	0.727272727	0.571428571	1
MLP	2014-2015	0.4	0.526315789	0.454545455	0.625
	2015-2016	0.266666667	0.3125	0.3125	0.3125
	2016-2017	0.533333333	0.695652174	0.533333333	1
	2017-2018	0.533333333	0.695652174	0.533333333	1
	2018-2019	0.533333333	0.695652174	0.533333333	1

Πίνακας 14 NBA AVG Results 5 years prediction



Εικόνα 21 NBA AVG chart 5 years prediction

Συμπέρασμα 3ου σεναρίου

Euroleague

Ο πιο αποδοτικός κατηγοριοποιητής , όπως στο 1ο και 2ο σενάριο, είναι ο “Random Forest”, που σε όλες τις μετρικές του «Πίνακας 13Πίνακας 9» δείχνει ότι έχει τη μεγαλύτερη τιμή . Όσο αναφορά το accuracy ο LogReg φαίνεται να είναι ο 2ος πιο αποδοτικός αλγόριθμος , αλλά υστερεί αρκετά στις άλλες μετρικές σε σχέση με τον KNN. Όπως και στα άλλα 2 σενάρια μετά τη χρονιά 2016-2017 το accuracy πέφτει στο 50%

NBA

Ο πιο αποδοτικός κατηγοριοποιητής αυτού του σεναρίου είναι “Random Forest”. Από τον «Πίνακας 14 » φαίνεται ότι υπερτερεί πολύ έναντι των άλλων αλγορίθμων. Ο χειρότερος όμως όλων είναι ο SVM.

4.5.4 Συμπεράσματα τριών παραπάνω σεναρίων

Σε αυτό το κεφάλαιο θα παρουσιάσουμε το πιο αποδοτικό σενάριο από τα παραπάνω ανά dataset.

Euroleague

Όσο αναφορά τη Euroleague όλα τα σενάρια δείχνουν ότι το accuracy είναι παρόμοιο, κοντά στο 0.58. Από εκεί και πέρα , υπολογίζοντας όλες τις άλλες μετρικές μας, το 1ο σενάριο (standardization) φαίνεται να υπερτερεί των άλλων 2 σεναρίων κατά πολύ. Το 2ο σενάριο (pearson) βέβαια, φαίνεται να μην είναι στο σύνολο πολύ αποδοτικό λόγω του αλγορίθμου MLP. Αν εξαιρεθεί δηλαδή ο MLP το 2ο σενάριο είναι το ίδιο αποδοτικό με το πρώτο.

Για παράδειγμα το accuracy του MLP στο 2ο σενάριο έχει avg accuracy 0.43 με τις υπόλοιπες μετρικές στο μηδέν. Το υπόλοιπο avg accuracy των υπολοίπων 4 αλγορίθμων είναι στο 0.62, με αντίστοιχες τιμές στο precision. Το recall από την άλλη πλευρά είναι πολύ υψηλό στο 1ο σενάριο. Από την άλλη πλευρά στο 3ο σενάριο το recall είναι καλό μόνο στους RFC και KNN.

NBA

Όσο αναφορά το NBA το 1ο σενάριο (standardization) είναι το πιο αποδοτικό σε σχέση με το accuracy αλλά και στο σύνολο των μετρικών. Με avg accuracy όλων των αλγορίθμων στο 0.73 , υπερτερεί πολύ των άλλων 2 σεναρίων που έχουν 0.51 και 0.53 το 2ο και 3ο σενάριο αντίστοιχα. Οι υπόλοιπες μετρικές δείχνουν ακριβώς την ίδια κατάσταση για το 2ο και 3ο σενάριο.

Για παράδειγμα το f1-score κυμαίνεται στο 0.77 στο 1ο σενάριο, σε αντίθεση με το 3ο που είναι στο 0.55. Εκτός «ανταγωνισμού» για το f1-score είναι το 2ο σενάριο που η τιμή κυμαίνεται στο 0.34

Σύνοψη

Όπως και στο κεφάλαιο «4.4 Πρόβλεψη αποτελέσματος (Qualified) με 5 κατηγοριοποιητές» , το 1ο σενάριο με standardization δείχνει να είναι το πιο αποδοτικό και στα 2 dataset. Είναι και η πιο σωστή μέθοδος προ επεξεργασίας στατιστικών δεδομένων αθλητικών αγώνων.

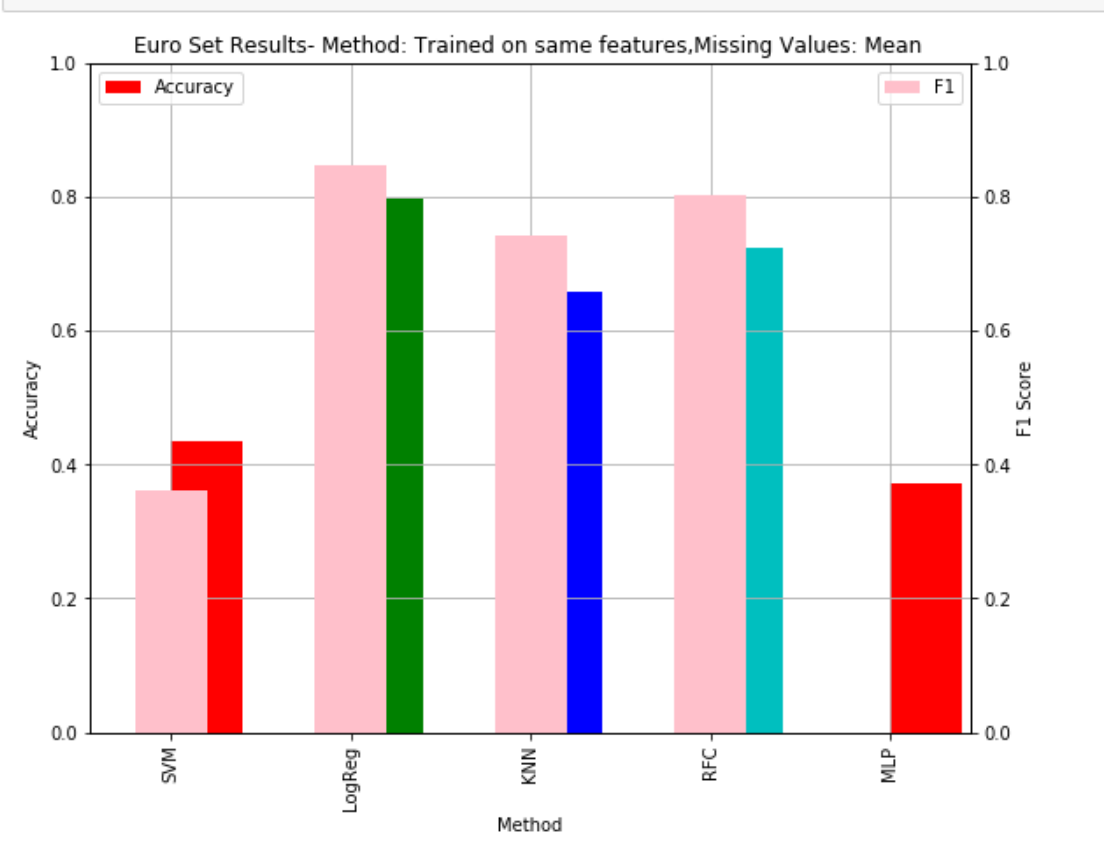
Μια παρατήρηση σχετικά με τις τιμές της Euroleague είναι ότι από την 3η χρονιά πρόβλεψης (2016-2017) το accuracy έπεφτε στο 0.5 σε αντίθεση με το NBA που δεν ήταν παρόμοια η κατάσταση. Αυτό συνέβη γιατί η εκπαίδευση των αλγορίθμων μας πραγματοποιήθηκε για τα πρώτα χρόνια και όσο αποκρινόμασταν από την 1η χρονιά (2014-2015) έπεφτε η ακρίβεια. Από την άλλη πλευρά αυτό δεν παρατηρήθηκε στο dataset του NBA. Αυτό οφείλεται γιατί η λογική του NBA είναι οι ομάδες να μην αλλάζουν budget και γενικά δυναμική κάθε χρόνο σε σχέση με τη Euroleague. Επίσης, όπως αναφέραμε και παραπάνω οι άμυνες στη Euroleague είναι πιο «σφιχτές», σε αντίθεση με το NBA που το μπάσκετ είναι πιο ελεύθερο. Αυτό σημαίνει ότι στο NBA τα ποσοστά ουσιαστικά είναι παρόμοια σε κάθε τους χρονιά.

4.6. Καλύτερο μοντέλο ανα dataset και fit στο άλλο.

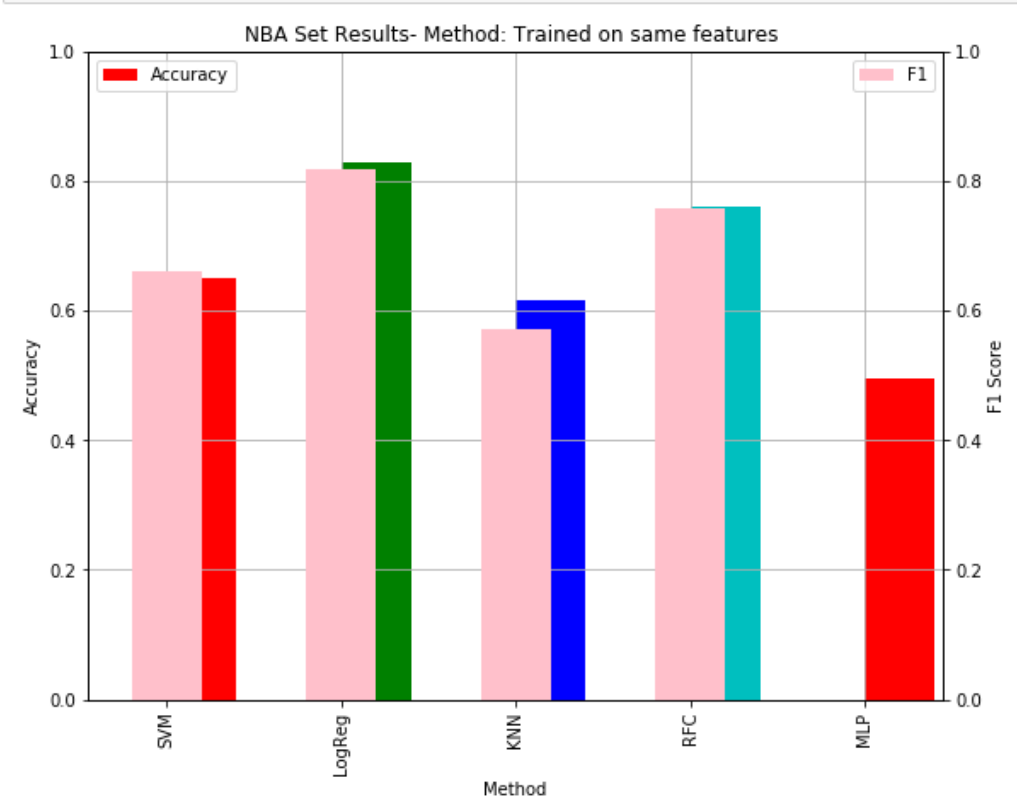
Σε αυτό το κεφάλαιο πραγματοποιήθηκε η εύρεση του καλύτερου αλγορίθμου ανα dataset και η εφαρμογή του εκπαιδευμένου μοντέλου αλγορίθμου στη χρονιά 2018-2019 του άλλου dataset για την εύρεση αποτελέσματος.

Πιο συγκεκριμένα βρήκαμε τα κοινά πεδία του κάθε dataset, ώστε το αποτέλεσμά μας να είναι πιο ακριβές και αξιόπιστο. Αυτό πραγματοποιήθηκε με την εντολή .intersection της python. Όπως σε κάθε προεπεξεργασία γεμίσαμε τις κενές τιμές μας με τη μέση τιμή (.mean). Στο συγκεκριμένο κομμάτι της εργασίας προτιμήσαμε την κανονικοποίηση (standardization) σαν πρόσθετο κομμάτι προεπεξεργασίας.

Εκπαιδεύσαμε εκ νέου και τους 5 αλγορίθμους που έχουμε επιλέξει και παραπάνω και βρήκαμε ότι καλύτερα αποτελέσματα έχει η λογαριθμική παλινδρόμηση και στα 2 dataset. Έτσι με τα εκπαιδευμένα μοντέλα LogReg προβλέψαμε τη χρονιά 2018-2019 του αντίθετου dataset.



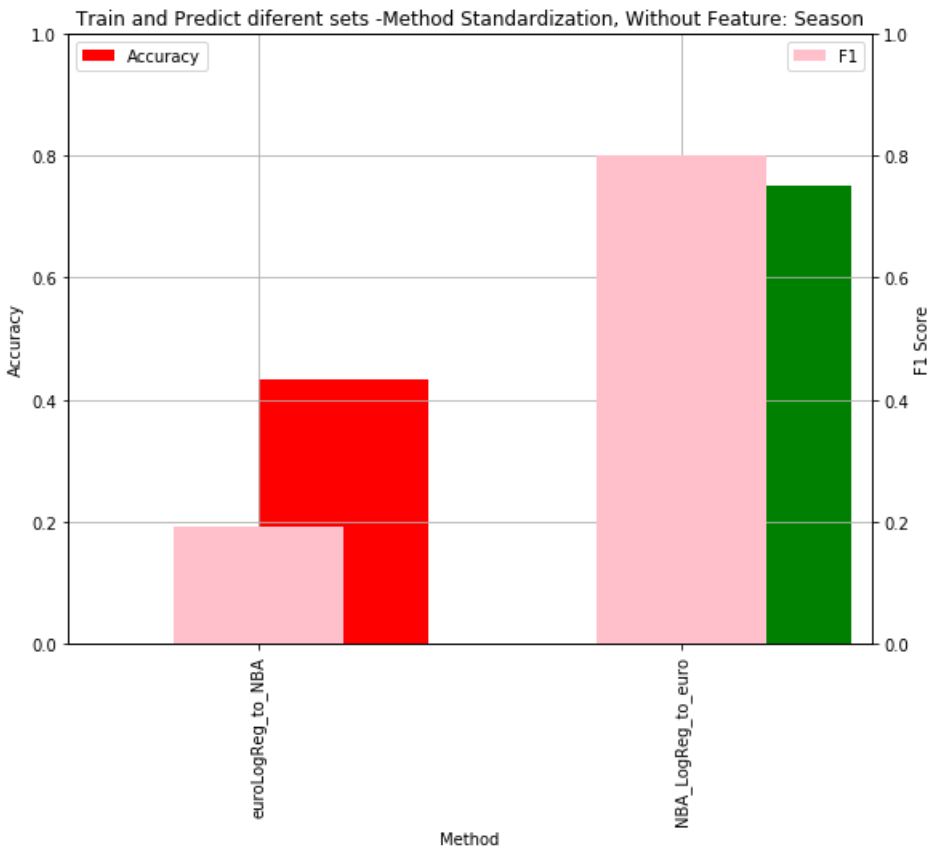
Εικόνα 22 Euroleague chart on same features



Εικόνα 23 NBA chart on same features

Name	Season	Accuracy	F1-score	Precision	Recall
Euroleague LogReg to NBA	2018 - 2019	0.42875	0.190526	0.666667	0.50000
NBA LogReg to Euroleague		0.76875	0.80000	0.636464	0.6875

Πίνακας 15 Log Regression fit one to another results



Εικόνα 24 Log Regression fit one to another chart

4 Συμπεράσματα

Μετά την εφαρμογή των αλγορίθμων που εκπαιδεύσαμε στο αντίθετο dataset παρατηρήσαμε ότι το μοντέλο της Euroleague δεν εφαρμόζει πολύ καλά στο NBA. Για παράδειγμα η ακρίβεια (accuracy) πρόβλεψης είναι στο 0.42875 που δε θεωρείται καλό. Αντίστοιχα και οι άλλες μετρικές δεν έχουν καλή απόδοση, όπως το F1 που είναι λίγο κάτω από το 0.2 . Precision και recall κυμαίνονται στο 0.6 και 0.5 αντίστοιχα.

Από την άλλη πλευρά, το μοντέλο που εκπαιδεύσαμε για το NBA θα λέγαμε ότι έχει πολύ καλύτερα αποτελέσματα όταν το εφαρμόζουμε στο dataset της Euroleague. Με accuracy στο 0.76875 και F1 στο 0.8 η απόδοση του συγκεκριμένου μοντέλου κρίνεται αποδοτική. Precision και recall κυμαίνονται στο 0.636464 και 0.6875 αντίστοιχα.

Αυτή η διαφορά οφείλεται στον αριθμό των αγώνων ανά πρωτάθλημα. Η Euroleague με Μ.Ο. παιχνιδιών στα 14 με 15 δε μπορεί να φέρει την ίδια απόδοση σε σχέση με το NBA που έχει αριθμό αγώνων 82. Επίσης τα στατιστικά είναι σε άλλο επίπεδο. Για παράδειγμα ο Μ.Ο. 3πόντων μιας ομάδας στο NBA είναι 635 τη χρονιά, όταν στη Euroleague είναι 107. Ως εκ τούτου η πληροφορία που δίνεται από το dataset του NBA είναι μεγαλύτερη για αυτό και χαρακτηρίζεται η εκπαίδευση του μοντέλου μας ως καλύτερο και έχει αντίστοιχα καλύτερα αποτελέσματα όταν το εφαρμόζουμε στο dataset της Euroleague.

5 Συμπεράσματα και μελλοντική εργασία

Σε αυτή την εργασία μελετήθηκαν και αναλύθηκαν 2 σύνολα στατιστικών δεδομένων από τα 2 μεγαλύτερα πρωταθλήματα μπάσκετ το κόσμου, αυτά της Euroleague και του NBA. Τα στατιστικά αφορούσαν την κανονική περίοδο και σαν κλάση υπήρχε η πληροφορία του αν πέρασε ή όχι στα playoff. Επιλέχθηκαν 3 διαφορετικά σενάρια για ανάλυση. Πρώτα επεξεργασία με standardization , έπειτα τα 10 πιο σχετιζόμενα χαρακτηριστικά του κάθε dataset με Pearson correlation και τέλος την επιλογή όλων των avg χαρακτηριστικών του κάθε συνόλου.

Πραγματοποιήθηκαν δύο ειδών προβλέψεις για καθένα από τα παραπάνω σενάρια με 5 κατηγοριοποιητές (SVM,LogReg,KNN,RFC,MLP). Αρχικά για όλο το σύνολο των στατιστικών του κάθε dataset (4.4) και μετά πρόβλεψη των τελευταίων 5 χρόνων για κάθε dataset (4.5). Σε κάθε μια πρόβλεψη ο κάθε κατηγοριοποιητής είχε τα δικά του θετικά και αρνητικά αποτελέσματα. Σε σχέση όμως με τα 3 σενάρια που επιλέχθηκα, το πρώτο (standardization) δείχνει να υπερέχει στο γενικό σύνολο των άλλων 2 σεναρίων.

Στο τέλος αφού εκπαιδεύσαμε 5 διαφορετικά μοντέλα, ένα για κάθε κατηγοριοποιητή, για όλες τις χρονιές εκτός από την τελευταία, επιλέξαμε τον αποδοτικότερο αλγόριθμο και τον εφαρμόσαμε στο αντίθετο dataset. Αυτός ήταν της λογαριθμικής παλινδρόμησης και φαίνεται ότι το μοντέλο του NBA εφαρμόστηκε πολύ καλύτερα σε αυτό της Euroleague και όχι το αντίθετο.

Μια μελλοντική εργασία θα μπορούσε να συμπεριλάβει συνέχεια των στατιστικών, δηλαδή αυτά των playoff και να αναδείξει ένα τελικό αποτέλεσμα του νικητή όλης της χρονιάς. Ως γνωστόν και στα δύο πρωταθλήματα, αλλά κυρίως σε αυτό του NBA τα στατιστικά και ο τρόπος παιχνιδιού είναι τελείως διαφορετικός στα playoff. Θα μπορούσαν να χρησιμοποιηθούν μόνο στατιστικά playoff και μετά όλο το σύνολο δεδομένων και να γίνει μια σύγκριση αποτελέσματος. Ενδιαφέρον θα είχε και εδώ η εφαρμογή του πιο αποδοτικού αλγορίθμου στο αντίθετο dataset.

Βιβλιογραφία (Harvard Reference System)

1. Aggarwal, C., 2019. *Neural Networks and Deep Learning*. New York: Springer.
2. Alkhatib, K., Najadat, H., Hmeidi, I. and Shatnawi, M., 2013. Stock Price Prediction Using k-nearest neighbor (kNN) Algorithm. *International Journal of Business, Humanities and Technology*, 3(3), pp.32-44.
3. Albert, J., Bennett, J. and Cochran, J.J., 2005. *Anthology of Statistics in Sports*. Philadelphia: Society for Industrial and Applied Mathematics.
4. Bengio, Y., Goodfellow, I. and Courville, A., 2017. *Deep Learning*. Massachusetts: MIT Press.
5. Berri, D.J. and Schmidt, M.B., 2010. *Stumbling on Wins: Two Economists Explore the Pitfalls on the Road to Victory in Professional Sports*. New Jersey: Financial Times Press.
6. Berry, M., Mohamed, A. and Yap, B., 2020. *Supervised and Unsupervised Learning for Data Science*. Cham: Springer Publications.
7. Biau, G., 2012. Analysis of a Random Forests Model. *Journal of Machine Learning Research*, 12, pp.1063-1095.
8. Biau, G., Lugosi, G. and Devroye, L., 2008. Consistency of Random Forests and Other Averaging Classifiers. *Journal of Machine Learning Research,*, 9, pp.2015-2033.
9. Boon, M., Kok, L. and Beck, S., 1995. Histological Validation of Neural-Network Assisted Cervical Screening: Comparison with the Conventional Approach. *Cell Vision*, 2, pp.23-27.
10. Bhandari, A., 2020. *Feature Scaling | Standardization Vs Normalization*. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/> [Accessed 17 March 2021].
11. Bourbousson, J., Sève, C. and McGarry, T., 2010. Space–time Coordination Dynamics in Basketball: Part 1. Intra- and Inter-couplings Among Player Dyads. *Journal of Sports Sciences*, 28(3), pp.339-347.
12. Bradley, A., 1997. The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognition*, 30(7), pp.1145-1159.
13. Breiman, L. 2001. Random Forests. *Machine Learning*, 45, pp.5–32.
14. Burkov, A., 2020. *Machine Learning Engineering*. Québek: True Positive Inc.
15. Cao, C., 2012. Sports Data Mining Technology Used in Basketball Outcome Prediction. Masters Dissertation. Technological University Dublin.
16. Chourasiya, S. and Jain, S., 2019. A Study Review on Supervised Machine Learning Algorithms. *International Journal of Computer Science and Engineering*, 6(8), pp.16-20.
17. Cojocaru, A., 2019. Handling Missing Data: Traditional Techniques Versus Machine Learning. *University of Twente*,.
18. Cristianini, N. and Shawe-Taylor, J., 2014. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge: Cambridge University Press.
19. Delen, D., Cogdell, D. and Kasap, N., 2012. A Comparative Analysis of Data Mining Methods in Predicting NCAA Bowl Outcomes. *International Journal of Forecasting*, 28(2), pp.543-552.

20. Demenius, J. and Kreivyte, R., 2017. The Benefits of Advanced Data Analytics in Basketball: Approach of Managers and Coaches of Lithuanian Basketball League Teams. *Baltic Journal of Sport and Health Sciences, 1(104)*, 8–13.

21. Dy, J. and Brodley, C., 2000. Feature Selection for Unsupervised Learning. *Journal of Machine Learning Research*, 5, pp.845–889.

22. Ellis, M., 1983. *Similarities and Differences in Games: A System for Classification*. AEISEP Conference.

23. Erčulj, F. and Štrumbelj, E., 2015. Basketball Shot Types and Shot Success in Different Levels of Competitive Basketball. *PLOS ONE*, 10(6).

24. Ghahramani, Z. 2004. *Unsupervised Learning*. In: Bousquet O., von Luxburg U., Rätsch G. (eds) Advanced Lectures on Machine Learning. ML 2003. Lecture Notes in Computer Science, vol. 3176. Berlin, Heidelberg: Springer Publications.

25. Google Developers, 2020. *Classification: ROC Curve and AUC | Machine Learning Crash Course*. [online] Google Developers. Available at: <https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc> [Accessed 16 March 2021].

26. Haykin, S., 2009. Neural Networks and Learning Machines. New York: Pearson Education.

27. Heumann, C. and Schomaker, M., 2016. *Introduction to Statistics and Data Analysis*. Zurich: Springer International Publishing.

28. Hladun, I., 2020. *The Benefits of Python for Software Projects | Waverley*. [online] Waverley. Available at: <https://waverleysoftware.com/blog/the-benefits-of-python/> [Accessed 16 March 2021].

29. Hodge, V. and Austin, J., 2004. A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, 22(2), pp.85-126.

30. Hofmann, T. 2001. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning, 42*, 177–196.

31. Ibáñez, S., Sampaio, J., Feu, S., Lorenzo, A., Gómez, M. and Ortega, E., 2008. Basketball Game-related Statistics that Discriminate Between Teams' Season-long Success. *European Journal of Sport Science*, 8(6), pp.369-372.

32. Ivanković, Z., Racković, M., Markoski, B., Radosav, D., and Ivković, M. (2010). Analysis of Basketball Games Using Neural Networks, 2010. In: *11th International Symposium on Computational Intelligence and Informatics (CINTI)*, pp. 251–256.

33. Jamsheer, K., 2020. *Top Advantages and Disadvantages of Python: A 2021 Guide*. [online] Web Solutions Blog. Available at: <https://acodez.in/advantages-and-disadvantages-of-python/> [Accessed 17 March 2021].

34. Joseph, A., Fenton, N. and Neil, M., 2006. Predicting Football Results Using Bayesian Nets and Other Machine Learning Techniques. *Knowledge-Based Systems*, 19(7), pp.544-553.

35. Karanjit, S. and Shuchita, U., 2012. Outlier Detection: Applications and Techniques. *International Journal of Computer Science Issues*, 9(1), pp.307-323.

36. Khadka, R., 2017. *Introduction to Machine Learning #1*. [online] Medium. Available at: <https://towardsdatascience.com/machine-learning-65dbd95f1603> [Accessed 18 February 2021].

37. Kahn, J., 2003. Neural Network Prediction of NFL Football Games. World Wide Web Electronic Publication, pp. 9–15.

38. Khan, S. and Hoque, A., 2020. SICE: An Improved Missing Data Imputation Technique. *Journal of Big Data*, 7(1).

39. Khanum, M., Mahboob, T., Imtiaz, W., Abdul Ghafoor, H. and Sehar, R., 2015. A Survey on Unsupervised Machine Learning Algorithms for Automation, Classification and Maintenance. *International Journal of Computer Applications*, 119(13), pp.34-39.

40. Kompoliti, K. and Verhagen Metman, L., 2010. Neural Networks. In: *Encyclopedia of Movement Disorders*. Amsterdam, Netherlands: Elsevier Ltd.

41. Kotsiantis, S., 2007. Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, 31, pp.249-268.

42. Lakshmanan, V., Robinson, S. and Munn, M., 2020. *Machine Learning Design Patterns: Solutions to Common Challenges in Data Preparation, Model Building, and MLOps*. Newton, Massachusetts: O'Reilly Media, Inc.

43. Lazzeri, F., 2021. *Machine Learning for Time Series Forecasting with Python*. Indianapolis: Wiley.

44. Leech, N., Barrett, K. and Morgan, G., 2008. *SPSS for intermediate statistics*. New York: Lawrence Erlbaum Associates.

45. Marshland, S. 2015. *Machine Learning: An Algorithm Perspective*. Boca Raton, FL: CRC Press.

46. Marmarinos, C., Apostolidis, N., Kostopoulos, N. and Apostolidis, A., 2016. Efficacy of the "Pick and Roll" Offense in Top Level European Basketball Teams. *Journal of Human Kinetics*, 51(1), pp.121-129.

47. Mavridis, G., Tsamourtzis, E., Karipidis, A. and Laios, A., 2009. The Inside Game in World Basketball. Comparison between European and NBA Teams. *International Journal of Performance Analysis in Sport*, 9(2), pp.157-164.

48. Mccabe, A. and Trevathan, J., 2008. Artificial Intelligence in Sports Prediction. In: *Fifth International Conference on Information Technology: New Generations*.

49. Meel, V., 2021. *Data Preprocessing Techniques for Machine Learning with Python | viso.ai*. [online] viso.ai. Available at: <https://viso.ai/deep-learning/data-preprocessing-techniques-for-machine-learning-with-python/> [Accessed 16 March 2021].

50. Miljković, D., Gajić, L., Kovačević, A., and Konjović, Z., 2010. The use of data mining for basketball matches outcomes prediction. In: *IEEE 8th International Symposium on Intelligent Systems and Informat*ics, pp. 309– 312.

51. Miller, A.C. and Bornn, L., 2017. Possession Sketches: Mapping NBA Strategies. *In: Proceedings of the MIT Sloan Sports Analytics Conference*, Boston, MA, USA.

52. Muhammad, I. and Yan, Z., 2015. Supervised Machine Learning Approaches: A Survey. *International Journal of Soft Computing*, 05(03), pp.946-952.

53. Ng, A., 2012. 1. Supervised learning. *In CS229: Machine Learning, Stanford University,* 1, pp. 1–30.

54. Novakovic, J., Veljovic, A., Ilic, S., Papic, Z. and Tomovic, M., 2021. Evaluation of Classification Models in Machine Learning. *Theory and Applications of Mathematics & Computer Science*, 7(1), pp.39-46.

55. Oliver, D., 2004. *Basketball on Paper: Rules and Tools for Performance Analysis*. Washington DC: Potomac Books.

56. Parikh, M., 2018. *Advantages Of Python Over Other Programming Languages - eLearning Industry*. [online] eLearning Industry. Available at: <https://elearningindustry.com/advantages-of-python-programming-languages> [Accessed 16 March 2021].

57. Peng, C., Lee, K. and Ingersoll, G., 2002. An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research*, 96(1), pp.3-14.

58. Phung, D., Webb, G. and Sammut, C., 2020. *Encyclopedia of Machine Learning and Data Science*. New York, NY: Springer US.

59. Powers, D., 2007. *Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation*. Technical Report SIE-07-001. Adelaide: School of Informatics and Engineering of Flinders University of South Australia.

60. Puente, C., Coso, J., Salinero, J. and Abián-Vicén., J., 2015. Basketball Performance Indicators During the ACB Regular Season from 2003 to 2013. *International Journal of Performance Analysis in Sport*, 15(3), pp.935-948.

61. Purucker, M. C., 1996. Neural Network Quarterbacking. *IEEE Potentials*, 15(3), pp. 9-15.

62. Ramchoun, H., Amine, M., Idrissi, J., Ghanou, Y. and Ettaouil, M., 2016. Multilayer Perceptron: Architecture Optimization and Training. *International Journal of Interactive Multimedia and Artificial Intelligence*, 4(1), p.26.

63. Santra, A. and Christy, J., 2012. Genetic Algorithm and Confusion Matrix for Document Clustering. *International Journal of Computer Science Issues*, 9(1), pp.322-328.

64. Selmanovic, A., Škegro, D. and Milanović, D., 2021. Basic Characteristics of Offensive Modalities in the Euroleague and the NBA. *Acta Kinesiologica*, 9(2), pp.83-87.

65. Shah, R. and Romijnders, R., 2016. Applying Deep Learning to Basketball Trajectories. *In: Proceedings of the Knowledge Discovery and Data Mining*, San Francisco, CA, USA.

66. Shea, S., 2013. *Basketball Analytics: Objective and Efficient Strategies for Understanding How Teams Win*. Lake, St. Louis: CreateSpace Independent Publishing Platform.

67. Singh, N., 2020. Sport Analytics: A Review. *The International Technology Management Review*, 9(1), pp.64-69.

68. Singh, D. and Chauhan, A., 2009. Neural Networks in Data Mining. *Journal of Theoretical and Applied Information Technology*, 5(1), pp.37-42.

69. Soto Valero, C., 2016. Predicting Win-Loss Outcomes in MLB Regular Season Games – A Comparative Study Using Data Mining Methods. *International Journal of Computer Science in Sport*, 15(2), pp.91-112.

70. Talukder, H., Vincent, T., Foster, G., Hu, C.,Huerta, J., 2016. Preventing in-game Injuries for NBA Players. In: *Proceedings of the MIT Sloan Sports Analytics Conference*, Boston, MA, USA.

71. Thabtah, F., Zhang, L. and Abdelhamid, N., 2019. NBA Game Result Prediction Using Feature Analysis and Machine Learning. *Annals of Data Science*, 6(1), pp.103-116.

72. Traeger, M., Eberhart, A., Geldner, G., Morin, A., Putzke, C., Wulf, H. and Eberhart, L., 2003. Artificial Neural Networks. Theory and Applications in Anesthesia, Intensive Care and Emergency Medicine. *Anaesthesist*, 52(11), pp.1055-61.

73. Veal, A. and Darcy, S., 2014. *Research Methods in Sport Studies and Sport Management*. Abingdon, Oxon: Routledge.

74. Wilson, B., 2021. *Machine Learning Engineering*. New York: Manning Publications.

75. Wilson, D. and Martinez, T., 2010. Reduction Techniques for Instance-based Learning Algorithms. *Machine Learning*, 38, pp.257-286.

76. Winston, W.L., 2009. *Mathletics: How Gamblers, Managers, and Sports Enthusiasts Use Mathematics in Baseball, Basketball, and Football*. Oxfordshire: Princeton University Press.

77. Wu, S. and Swartz, T., 2017. Using AI to Correct Play-by-play Substitution Errors. In: *Proceedings of the MIT Sloan Sports Analytics Conference*, Boston, MA, USA.

78. Zuccolotto, P. and Manisera, M., 2020. *Basketball Data Science*. London: CRC Press.