ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

“ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ & ΥΠΗΡΕΣΙΕΣ”

Ανάλυση δεδομένων καλαθοσφαίρισης για την πρόβλεψη αποτελεσμάτων με επιλογή χαρακτηριστικών

Από

Μιχαλάκης Ζαχαρίας

Υποβάλλεται
για την εκπλήρωση των προϋποθέσεων λήψης
Μεταπτυχιακού Διπλώματος
στην ειδίκευση «ΜΔΑ/ΠΠΣ/ΠΔ»
του ΠΜΣ “Πληροφοριακά Συστήματα & Υπηρεσίες”
στο
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
Μάιος 2024

Επιβλέπων: Ηλίας Μαγκλογιάννης
Ακαδημαϊκή Θέση: Κοσμήτορας Σχολής

Συγγραφέας / Author. . . . . Μιχαλάκης Ζαχαρίας. . . . . .

ΣΕΛΙΔΑ ΕΓΚΥΡΟΤΗΤΑΣ

Ονοματεπώνυμο Φοιτητή/Φοιτήτριας: …… Μιχαλάκης Ζαχαρίας ……..

Τίτλος Μεταπτυχιακής Διπλωματικής Εργασίας: … Ανάλυση δεδομένων καλαθοσφαίρισης για την πρόβλεψη αποτελεσμάτων με επιλογή χαρακτηριστικών….

Η παρούσα Μεταπτυχιακή Διπλωματική Εργασία υποβάλλεται ως μερική εκπλήρωση των απαιτήσεων του Προγράμματος Μεταπτυχιακών Σπουδών "Πληροφοριακά Συστήματα & Υπηρεσίες" του Τμήματος Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς και εγκρίθηκε στις ….30/04/2024…… από τα μέλη της Εξεταστικής Επιτροπής.

Εξεταστική Επιτροπή

Επιβλέπων (Τμήμα Ψηφιακών Συστημάτων, Πανεπιστήμιο Πειραιώς)… Ηλίας Μαγκλογιάννης, Κοσμήτορας Σχολής……

Μέλος Εξεταστικής Επιτροπής: … Μιχαήλ Φιλιππάκης, Καθηγητής…

Μέλος Εξεταστικής Επιτροπής: … Κωνσταντίνος Δελήμπασης, Επίκουρος Καθηγητής…

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΑΥΘΕΝΤΙΚΟΤΗΤΑΣ

Ο/Η…. Μιχαλάκης Ζαχαρίας.., γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα ότι η παρούσα εργασία με τίτλο «…Ανάλυση δεδομένων καλαθοσφαίρισης για την πρόβλεψη αποτελεσμάτων με επιλογή χαρακτηριστικών ….», αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές που έχω χρησιμοποιήσει, έχουν δηλωθεί κατάλληλα στις βιβλιογραφικές παραπομπές και αναφορές. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή.
Επιπλέον *δηλώνω υπεύθυνα ότι η συγκεκριμένη Μεταπτυχιακή Διπλωματική Εργασία έχει συγγραφεί από εμένα προσωπικά και δεν έχει υποβληθεί ούτε έχει αξιολογηθεί στο πλαίσιο κάποιου άλλου μεταπτυχιακού ή προπτυχιακού τίτλου σπουδών, στην Ελλάδα ή στο εξωτερικό.*
Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου. Σε κάθε περίπτωση, αναληθούς ή ανακριβούς δηλώσεως, υπόκειμαι στις συνέπειες που προβλέπονται τις διατάξεις που προβλέπει η Ελληνική και Κοινοτική Νομοθεσία περί πνευματικής ιδιοκτησίας.

Ο/Η ΔΗΛΩΝ/ΟΥΣΑ

Ονοματεπώνυμο: Μιχαλάκης Ζαχαρίας
Αριθμός Μητρώου: Μ2116
Υπογραφή: 

Περίληψη

Η παρούσα Μεταπτυχιακή Διπλωματική εργασία εμβαθύνει στον τομέα της Ανάλυσης Αθλητικών Δεδομένων (Sports Analytics) από το άθλημα της Καλαθοσφαίρισης (Basketball) με σκοπό την πρόβλεψη αποτελεσμάτων και την επιλογή χαρακτηριστικών (Feature Selection) που συμβάλλουν στην επίτευξη αυτού του στόχου. Η εργασία αναπτύσσει και εφαρμόζει σύγχρονες μεθοδολογίες μηχανικής μάθησης (Machine Learning) και στατιστικής ανάλυσης για την επεξεργασία και την αξιολόγηση των δεδομένων, επιδιώκοντας να αναγνωρίσει ποια δεδομένα και στατιστικά στοιχεία έχουν την μεγαλύτερη προβλεπτική αξία.

Η έρευνα αρχίζει με μια εκτενή ανασκόπηση της σχετικής βιβλιογραφίας, εξετάζοντας προηγούμενες μελέτες που έχουν ασχοληθεί με την πρόβλεψη αποτελεσμάτων στον αθλητισμό και ειδικότερα στην καλαθοσφαίριση. Στη συνέχεια, παρουσιάζεται η μεθοδολογία που υιοθετήθηκε για την επεξεργασία των δεδομένων, η οποία περιλαμβάνει τεχνικές προεπεξεργασίας, επιλογής χαρακτηριστικών, και μοντελοποίησης. Ειδική έμφαση δίνεται στην εφαρμογή αλγορίθμων μηχανικής μάθησης, όπως η Logistic Regression, k-nearest neighbors (KNN), Support Vector Machine (SVM), Random Forest, για την αξιολόγηση της προβλεπτικής τους ικανότητας σε σχέση με τα αποτελέσματα των αγώνων.

Ακολούθως, παρουσιάζονται τα αποτελέσματα της εφαρμογής των μεθόδων αυτών σε ένα σύνολο δεδομένων που περιλαμβάνει στατιστικά από επαγγελματικούς αγώνες καλαθοσφαίρισης και συγκεκριμένα στο Ευρωπαϊκό Πρωτάθλημα (Euroleague). Η ανάλυση επικεντρώνεται στην αξιολόγηση της σημασίας κάθε χαρακτηριστικού και της συνεισφοράς του στην πρόβλεψη του αποτελέσματος των αγώνων, καθώς και στην αποτελεσματικότητα των διαφορετικών μοντέλων πρόβλεψης. Απώτερος στόχος ήταν να κάνουμε πρόβλεψη αποτελέσματος με τους παραπάνω κατηγοριοποιητές βάσει του διαθέσιμου συνόλου δεδομένων.

Τέλος, η εργασία καταλήγει με μια συζήτηση για τις προκλήσεις και τις προοπτικές της πρόβλεψης αποτελεσμάτων στην καλαθοσφαίριση, προτείνοντας διευρύνσεις για μελλοντικές ερευνητικές προσπάθειες. Επιπλέον, αναγνωρίζεται η σημασία της ενσωμάτωσης

περισσότερων δεδομένων και της βελτίωσης των αλγορίθμων μηχανικής μάθησης για την αύξηση της ακρίβειας των προβλέψεων. Η εργασία αυτή συμβάλλει στον τομέα της αναλυτικής αθλητικής επιστήμης, προσφέροντας πολύτιμες διορατικότητες για την εφαρμογή της μηχανικής μάθησης στην πρόβλεψη αθλητικών αποτελεσμάτων.

Λέξεις κλειδιά : Basketball, Euroleague , Sports Analytics , Machine Learning , Multi-layer Perceptron, k-nearest neighbors , Logistic Regression, Support Vector Machine , Random Forest.

Abstract

This Master's thesis delves into the field of Sports Data Analysis from the sport of Basketball in order to predict results and select characteristics that contribute to the achievement of this goal. The paper develops and applies modern machine learning and statistical analysis methodologies to process and evaluate data, seeking to identify which data and statistics have the greatest predictive value.

The research begins with an extensive review of the relevant literature, examining previous studies that have dealt with outcome prediction in sport and basketball in particular. The methodology adopted for data processing is then presented, which includes pre-processing, feature selection, and modelling techniques. Special emphasis is given to the application of machine learning algorithms, such as Logistic Regression, k-nearest neighbors (KNN), Support Vector Machine (SVM), Random Forest, Neural Network(Multi-layer Perceptron), to evaluate their predictive ability with respect to game outcomes. Another goal was to make a forecast of the last 5 years with the above classifiers, based on previous years for each dataset.

In the following, the results of the application of these methods to a dataset that includes statistics from professional basketball games, specifically the European League (Euroleague), are presented. The analysis focuses on evaluating the importance of each attribute and its contribution to predicting the outcome of the games, as well as the effectiveness of the different prediction models. The ultimate goal was to make outcome prediction with the above categorizers based on the available dataset.

Finally, the paper concludes with a discussion of the challenges and prospects of outcome prediction in basketball, suggesting extensions for future research efforts. In addition, the importance of incorporating more data and improving machine learning algorithms to increase the accuracy of predictions is acknowledged. This work contributes to the field of analytical sports science by providing valuable insights for the application of machine learning to sports outcome prediction.

Keywords: Basketball, Euroleague , Sports Analytics , Machine Learning , Multi-layer Perceptron, k-nearest neighbors , Logistic Regression, Support Vector Machine , Random Forest.

Λίστα Περιεχομένων

Λίστα Εικόνων

Λίστα Πινάκων

1.1. Εισαγωγή

Η ανάλυση δεδομένων έχει εξελιχθεί σε έναν κρίσιμο τομέα εντός της σύγχρονης επιστημονικής έρευνας και της εφαρμοσμένης ανάλυσης, παρέχοντας τη δυνατότητα να αποκαλύπτονται σημαντικές πληροφορίες και να λαμβάνονται αποφάσεις με βάση τα δεδομένα. Στον τομέα του αθλητισμού, και ειδικότερα στην καλαθοσφαίριση, η ανάλυση δεδομένων έχει αποδειχθεί ιδιαίτερα ωφέλιμη, καθώς επιτρέπει την ανάλυση και την κατανόηση των αποδόσεων των παικτών, των στρατηγικών των ομάδων και των τάσεων των αγώνων με έναν τρόπο που δεν ήταν δυνατός στο παρελθόν.

Η παρούσα εισαγωγή αποσκοπεί στο να θέσει τις βάσεις για την κατανόηση της σημασίας της ανάλυσης δεδομένων στην καλαθοσφαίριση, παρουσιάζοντας το πλαίσιο, τις μεθοδολογίες και τις προκλήσεις που συνδέονται με αυτήν την προσέγγιση. Αρχικά, εξετάζεται η εξέλιξη της ανάλυσης δεδομένων στον αθλητισμό, αναδεικνύοντας την αυξανόμενη σημασία της τεχνολογίας και των αλγορίθμων στην επίτευξη αθλητικής αριστείας και στρατηγικής κατανόησης. Στη συνέχεια, αναλύεται η συμβολή της μηχανικής μάθησης και της στατιστικής ανάλυσης στην επεξεργασία και την ερμηνεία των δεδομένων, επισημαίνοντας τις δυνατότητες που προσφέρουν για την πρόβλεψη αποτελεσμάτων και τη βελτίωση των αποδόσεων.

Επιπρόσθετα, η εισαγωγή αυτή επιχειρεί να καταδείξει την κρισιμότητα της επιλογής και της αξιολόγησης των χαρακτηριστικών (features) που χρησιμοποιούνται στην ανάλυση, καθώς η επιτυχία των μοντέλων πρόβλεψης εξαρτάται σημαντικά από την ικανότητα να διακρίνουν και να εκμεταλλεύονται τα πιο σημαντικά και ενδεικτικά δεδομένα. Η διαδικασία της επιλογής χαρακτηριστικών απαιτεί μια βαθιά κατανόηση τόσο του αθλήματος όσο και των διαθέσιμων αναλυτικών τεχνικών, προκειμένου να εντοπιστούν εκείνα τα στοιχεία που προσφέρουν τη μεγαλύτερη προβλεπτική αξία.

Τέλος, η εισαγωγή αυτή θέτει το θεμέλιο για την παρούσα έρευνα, η οποία αποσκοπεί στην εφαρμογή και την αξιολόγηση σύγχρονων μεθόδων ανάλυσης δεδομένων στην πρόβλεψη αποτελεσμάτων αγώνων καλαθοσφαίρισης. Μέσω της συστηματικής εξέτασης των δεδομένων και της εφαρμογής προηγμένων αναλυτικών τεχνικών, η εργασία αυτή επιδιώκει να συμβάλει στην εμβάθυνση της κατανόησης των δυναμικών που διαμορφώνουν τα αποτελέσματα στην καλαθοσφαίριση και να προσφέρει νέες προοπτικές για την αξιοποίηση της ανάλυσης δεδομένων στον αθλητισμό.

1.2. Ορισμός Προβλήματος – Θεματική περιοχή

Η θεματική περιοχή της παρούσας μεταπτυχιακής διπλωματικής εργασίας εστιάζει στην ανάλυση δεδομένων αγώνων καλαθοσφαίρισης για την πρόβλεψη αποτελεσμάτων και την επιλογή χαρακτηριστικών που είναι κρίσιμα για την επίτευξη αυτού του στόχου. Η εν λόγω έρευνα καταπιάνεται με την ανάπτυξη και την εφαρμογή μοντέλων μηχανικής μάθησης και στατιστικών μεθόδων που επιδιώκουν να αξιοποιήσουν τα διαθέσιμα δεδομένα για την ακριβή πρόβλεψη των αποτελεσμάτων των αγώνων. Το πρόβλημα που αντιμετωπίζεται είναι διττό: αφενός, η ανάγκη για ακριβείς προβλέψεις που μπορούν να υποστηρίξουν τη λήψη αποφάσεων σε διάφορα επίπεδα (π.χ., τακτική ομάδων, στοιχηματικές αγορές, αθλητική ανάλυση), και αφετέρου, η επιλογή και η βελτιστοποίηση των χαρακτηριστικών που θα χρησιμοποιηθούν για την πρόβλεψη, ώστε να εξασφαλιστεί η μέγιστη δυνατή αποδοτικότητα και ακρίβεια των μοντέλων.

Το κεντρικό πρόβλημα που εξετάζεται στην παρούσα εργασία αφορά την ανάπτυξη ενός αποτελεσματικού πλαισίου για την πρόβλεψη αποτελεσμάτων αγώνων καλαθοσφαίρισης μέσω της ανάλυσης δεδομένων. Το πρόβλημα αυτό διακλαδίζεται σε δύο βασικές πτυχές:

Πρόβλεψη Αποτελεσμάτων: Πώς μπορούν τα διαθέσιμα δεδομένα από προηγούμενους αγώνες καλαθοσφαίρισης να αναλυθούν και να επεξεργαστούν με τρόπο που να επιτρέπει την ακριβή πρόβλεψη των αποτελεσμάτων μελλοντικών αγώνων;

Επιλογή Χαρακτηριστικών: Ποια χαρακτηριστικά (π.χ., στατιστικά των παικτών, ιστορικό αγώνων, φυσική κατάσταση, τακτικές) προσφέρουν τη μεγαλύτερη προβλεπτική αξία και πώς μπορεί να γίνει η επιλογή και η βελτιστοποίησή τους για την ενίσχυση της ακρίβειας των προβλέψεων;

Η προσέγγιση αυτού του προβλήματος απαιτεί μια συνδυαστική εφαρμογή τεχνικών μηχανικής μάθησης, στατιστικής ανάλυσης, και βαθιάς κατανόησης του αθλήματος. Η επιτυχία της προσπάθειας εξαρτάται από την ικανότητα να αναγνωριστούν και να αξιοποιηθούν τα κατάλληλα δεδομένα και χαρακτηριστικά που μπορούν να προσδώσουν σημαντική προστιθέμενη αξία στην προβλεπτική διαδικασία, καθώς και από την ανάπτυξη καινοτόμων και αποδοτικών μοντέλων που μπορούν να διαχειριστούν την πολυπλοκότητα και την αβεβαιότητα που χαρακτηρίζει τα αθλητικά δεδομένα.

1.3. Δομή Εργασίας

Η δομή της παρούσας μεταπτυχιακής διπλωματικής εργασίας αποτελείται από το κεφάλαιο 1 το οποίο παρουσιάζει το πλαίσιο και τη σημασία της έρευνας, καθορίζοντας το πρόβλημα και τους στόχους της εργασίας. Επιπλέον, περιγράφει την οργάνωση του κειμένου και τη δομή της εργασίας. Στην συνέχεια αποτελείται από το κεφάλαιο 2 το οποίο εξετάζει την υπάρχουσα βιβλιογραφία και τις θεωρητικές έννοιες που σχετίζονται με την ανάλυση δεδομένων στον αθλητισμό και ειδικότερα στην καλαθοσφαίριση. Παρουσιάζει επίσης προηγούμενες μελέτες που έχουν εφαρμόσει μεθόδους μηχανικής μάθησης για την πρόβλεψη αποτελεσμάτων. Ακολούθως έχουμε το κεφάλαιο 3 το οποίο περιγράφει τις μεθόδους και τις τεχνικές που χρησιμοποιήθηκαν για την ανάλυση των δεδομένων, την επιλογή χαρακτηριστικών, και την ανάπτυξη των προβλεπτικών μοντέλων. Επιπλέον, παρουσιάζει τη διαδικασία συλλογής και προεπεξεργασίας των δεδομένων. Στο κεφάλαιο 4 αναλύονται τα δεδομένα που συλλέχθηκαν και περιγράφεται η διαδικασία επιλογής των χαρακτηριστικών που χρησιμοποιήθηκαν για την πρόβλεψη. Εξετάζεται η σημασία και η επίδραση διαφόρων χαρακτηριστικών στην ακρίβεια των προβλέψεων. Στο κεφάλαιο 5 παρουσιάζονται τα αποτελέσματα της εφαρμογής των μοντέλων πρόβλεψης και αναλύεται η αποδοτικότητα και η ακρίβεια τους. Συζητά τις προκλήσεις, τις περιορισμένες δυνατότητες και τις προοπτικές βελτίωσης των μοντέλων. Το κεφάλαιο 6 συνοψίζει τα κύρια ευρήματα της εργασίας και προτείνει διευρύνσεις για μελλοντική έρευνα. Επιπλέον, παρουσιάζει προτάσεις για την εφαρμογή των αποτελεσμάτων στην πράξη και την περαιτέρω ανάπτυξη της αναλυτικής αθλητικής επιστήμης. Στη συνέχεια το παράρτημα περιλαμβάνει επιπλέον πληροφορίες, τεχνικές λεπτομέρειες, διαγράμματα, και τα δεδομένα που χρησιμοποιήθηκαν στην έρευνα, προσφέροντας μια πιο διεξοδική κατανόηση της μεθοδολογίας και των αποτελεσμάτων. Και τέλος στην Βιβλιογραφία παραθέτει όλες τις πηγές που αναφέρθηκαν και χρησιμοποιήθηκαν στην εργασία, παρέχοντας τη βάση για την επιστημονική υποστήριξη των επιχειρημάτων και των συμπερασμάτων.

Η δομή αυτή επιδιώκει να παρέχει μια σαφή και λογική προσέγγιση στην εξέταση του θέματος, διευκολύνοντας τον αναγνώστη να κατανοήσει την πορεία της έρευνας και τα ευρήματα που προέκυψαν.

2. Βιβλιογραφική επισκόπηση και Τεχνολογικό Υπόβαθρο

Η βιβλιογραφική επισκόπηση και το τεχνολογικό υπόβαθρο αποτελούν κρίσιμα στοιχεία κάθε ερευνητικής εργασίας, καθώς παρέχουν το θεωρητικό πλαίσιο και τις προηγούμενες γνώσεις πάνω στις οποίες στηρίζεται η τρέχουσα μελέτη

2.1. Αλγόριθμοι και Είδη Μηχανικής Μάθησης

Η Μηχανική Μάθηση (Machine Learning - ML) είναι ένας υποκλάδος της τεχνητής νοημοσύνης που επικεντρώνεται στην ανάπτυξη αλγορίθμων οι οποίοι μπορούν να μαθαίνουν από δεδομένα και να κάνουν προβλέψεις ή να λαμβάνουν αποφάσεις βασισμένες σε δεδομένα με ελάχιστη ή καθόλου ανθρώπινη παρέμβαση. Υπάρχουν τέσσερα βασικά είδη μηχανικής μάθησης: επιβλεπόμενη μάθηση, μη επιβλεπόμενη μάθηση, ημι-επιβλεπόμενη μάθηση και ενισχυτική μάθηση. Κάθε είδος χρησιμοποιεί διαφορετικούς αλγορίθμους και μαθηματικούς τύπους για την επεξεργασία και την ανάλυση δεδομένων.

Η επιβλεπόμενη μάθηση είναι η πιο διαδεδομένη κατηγορία μηχανικής μάθησης, όπου το μοντέλο εκπαιδεύεται σε ένα σύνολο δεδομένων που περιλαμβάνει τις είσοδούς και τις αντίστοιχες επιθυμητές εξόδους. Στόχος είναι το μοντέλο να μάθει να προβλέπει την έξοδο για νέα δεδομένα. Έστω X το σύνολο δεδομένων εισόδου και Y το σύνολο των αντίστοιχων ετικετών εξόδου. Ο αλγόριθμος επιχειρεί να μάθει μια συνάρτηση $f:X \rightarrow Y$ τέτοια ώστε για κάθε νέο δείγμα εισόδου x, να προβλέπει την έξοδο y. Κλασικοί αλγόριθμοι επιβλεπόμενης μάθησης περιλαμβάνουν:

Γραμμική Παλινδρόμηση (Linear Regression): Προβλέπει μια συνεχή τιμή εξόδου βάσει μίας ή περισσότερων εισόδων.

Λογιστική Παλινδρόμηση (Logistic Regression): Χρησιμοποιείται για την πρόβλεψη της πιθανότητας μιας κατηγοριοποιημένης εξόδου (π.χ., ναι ή όχι).

Δέντρα Αποφάσεων (Decision Trees): Μοντέλο που χρησιμοποιεί μια δομή δέντρου για να πάρει αποφάσεις, καθοδηγώντας τις εισόδους μέσα από μια σειρά από κριτήρια.

Τυχαία Δάση (Random Forests): Ένας συνδυασμός πολλαπλών δέντρων αποφάσεων για τη βελτίωση της ακρίβειας της πρόβλεψης.

Υποστηρικτικές Μηχανές Διανυσμάτων (Support Vector Machines - SVM): Αναπτύσσουν ένα υπερεπίπεδο ή σύνολο υπερεπιπέδων σε έναν υψηλότερο ή χαμηλότερο διάστατο χώρο για την

ταξινόμηση ή την παλινδρόμηση.

Στη μη επιβλεπόμενη μάθηση, τα δεδομένα εκπαίδευσης δεν περιλαμβάνουν επιθυμητές εξόδους. Αντ' αυτού, το μοντέλο προσπαθεί να αναγνωρίσει τις δομές και τα μοτίβα μέσα στα δεδομένα. Έστω X={x_1,x_2 ,...,x_n} το σύνολο δεδομένων. Ο αλγόριθμος επιδιώκει να διαιρέσει το *X* σε *k* συστάδες *C*={c_1,c_2,...,c_k} με βάση την ομοιότητα των στοιχείων, ελαχιστοποιώντας μια συνάρτηση απώλειας, όπως η συνολική εντός συστάδας διακύμανση: $\sum^{k}_{i=1} \sum_{x \in Ci} ||x-\mu_i||^2$ όπου μ_i είναι το κέντρο της συστάδας C_i. Κοινές μέθοδοι περιλαμβάνουν:

Ομαδοποίηση (Clustering): Όπως ο K-Means, που ομαδοποιεί τα δεδομένα σε κλάστερ βάσει ομοιότητας.

Μείωση Διαστάσεων (Dimensionality Reduction): Τεχνικές όπως η Ανάλυση Κύριων Συνιστωσών (PCA) που μειώνουν τον αριθμό των μεταβλητών διατηρώντας ταυτόχρονα τη σημαντική πληροφορία.

Η ημι-επιβλεπόμενη μάθηση χρησιμοποιείται όταν τα δεδομένα εισόδου περιλαμβάνουν μια μικρή ποσότητα επιθυμητών εξόδων και μια μεγάλη ποσότητα δεδομένων χωρίς ετικέτες. Ο στόχος είναι να βελτιώσει την ακρίβεια του μοντέλου χρησιμοποιώντας το μεγάλο σύνολο δεδομένων χωρίς ετικέτες.

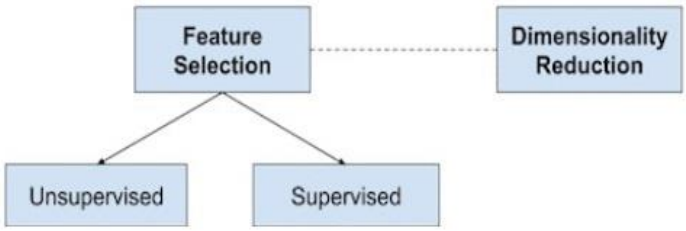
Η ενισχυτική μάθηση είναι μια προσέγγιση όπου ένας πράκτορας μαθαίνει να λαμβάνει αποφάσεις εκτελώντας ενέργειες σε ένα περιβάλλον με σκοπό τη μεγιστοποίηση κάποιας έννοιας της ανταμοιβής. Η ενισχυτική μάθηση επικεντρώνεται στην ανακάλυψη της βέλτιστης στρατηγικής δράσης μέσω της δοκιμής και του λάθους, και χρησιμοποιείται σε προβλήματα όπως τα παιχνίδια, η αυτοματοποιημένη οδήγηση, και η ρομποτική.

Η βασική ιδέα στην ενισχυτική μάθηση μπορεί να εκφραστεί μέσω της εξίσωσης Bellman για την αναμενόμενη χρησιμότητα $V(s)$ ενός καταστάσεως $V(s)=\max_a \sum_{s'} P(s'|s,a)[R(s,a,s')+\gamma V(s')]$ όπου $P(s'|s,a)$ είναι η πιθανότητα μετάβασης στην κατάσταση '*s*' εκτελώντας την ενέργεια *a* από την κατάσταση *s*, $R(s,a,s')$ είναι η ανταμοιβή που λαμβάνεται για τη μετάβαση από την κατάσταση *s* στην '*s*' με την ενέργεια *a*, και *γ* είναι ο παράγοντας απότισης που ρυθμίζει τη σημασία των μελλοντικών ανταμοιβών.

Κάθε μία από αυτές τις προσεγγίσεις έχει τις δικές της εφαρμογές, πλεονεκτήματα, και περιορισμούς. Η επιλογή του κατάλληλου είδους μηχανικής μάθησης και του συγκεκριμένου αλγορίθμου εξαρτάται από τη φύση του προβλήματος, τη διαθεσιμότητα και το είδος των δεδομένων, καθώς και από τους στόχους της εφαρμογής.

2.2. Επιλογή Χαρακτηριστικών

Η επιλογή χαρακτηριστικών (feature selection) είναι μια κρίσιμη διαδικασία στην προετοιμασία των δεδομένων για μοντέλα μηχανικής μάθησης. Αφορά την επιλογή εκείνων των μεταβλητών ή χαρακτηριστικών που είναι πιο σημαντικά ή σχετικά με το πρόβλημα που προσπαθούμε να λύσουμε. Η διαδικασία αυτή βοηθά στην αποφυγή του φαινομένου της υπερπροσαρμογής (overfitting), μειώνει τον υπολογιστικό φόρτο, βελτιώνει την απόδοση του μοντέλου και καθιστά τα αποτελέσματα της μοντελοποίησης πιο εύκολα ερμηνεύσιμα. Ορισμένα προβλήματα προγνωστικής μοντελοποίησης έχουν μεγάλο αριθμό μεταβλητών που μπορεί να επιβραδύνουν την ανάπτυξη και την εκπαίδευση των μοντέλων και να απαιτούν μεγάλο όγκο μνήμης του συστήματος. Επιπλέον, η απόδοση ορισμένων μοντέλων μπορεί να υποβαθμιστεί όταν περιλαμβάνονται μεταβλητές εισόδου που δεν είναι σχετικές με τη μεταβλητή-στόχο. Έτσι λοιπόν μπορούμε να διακρίνουμε 2 βασικές κατηγορίες Μοντέλα με επίβλεψη (Supervised Models) και Μοντέλα χωρίς επίβλεψη (Unsupervised Models).

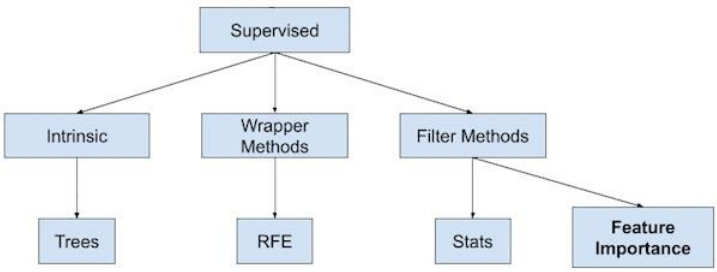


Εικόνα 1 Feature Selection Techniques

2.2.1. Μοντέλα με επίβλεψη (Supervised Models)

Προκειμένου να μετρηθεί η σημασία και η συνάφεια των χαρακτηριστικών με τη χρήση επισημειωμένων δεδομένων για την εκπαίδευση του μοντέλου επιλογής χαρακτηριστικών, η επιλογή χαρακτηριστικών με επίβλεψη στοχεύει συνήθως σε προβλήματα ταξινόμησης ή παλινδρόμησης. Στόχος της είναι η επιλογή ενός υποσυνόλου χαρακτηριστικών με βάση συγκεκριμένα κριτήρια, τα οποία ποικίλλουν ανάλογα με τη μέθοδο επιλογής. Τα χαρακτηριστικά που επιλέγονται έχουν σημαντικό αντίκτυπο στη φάση της εκπαίδευσης. Συγκεκριμένα, οι ταξινομητές ή τα μοντέλα παλινδρόμησης εκπαιδεύονται χρησιμοποιώντας το υποσύνολο χαρακτηριστικών που επιλέγεται από την επιβλεπόμενη επιλογή χαρακτηριστικών, αφού τα δεδομένα έχουν χωριστεί σε σύνολα εκπαίδευσης και δοκιμής. Τέλος, χρησιμοποιώντας δείγματα δοκιμαστικού συνόλου που περιέχουν τα επιλεγμένα χαρακτηριστικά, ο προηγμένος ταξινομητής ή το μοντέλο παλινδρόμησης προβλέπει ετικέτες κλάσεων ή στόχους παλινδρόμησης. Όταν τα δεδομένα έχουν επισημανθεί σωστά, οι τεχνικές επιλογής χαρακτηριστικών με επίβλεψη μπορούν να αποδώσουν τα καλύτερα αποτελέσματα. Ως εκ τούτου, μία από τις δυσκολίες και τα μειονεκτήματα της επιλογής χαρακτηριστικών με επίβλεψη είναι ότι η επισήμανση των δεδομένων μπορεί να είναι δαπανηρή και αναξιόπιστη. Λόγω της ακούσιας αφαίρεσης σχετικών δεδομένων ή της επιλογής άσχετων χαρακτηριστικών, αυτό εγείρει την πιθανότητα υπερπροσαρμογής της διαδικασίας εκπαίδευσης.

Οι τεχνικές επιλογής χαρακτηριστικών μπορούν να καταταχθούν σε τρεις κύριες κατηγορίες: φιλτραρίσματος (filter methods), περιτύλιξης (wrapper methods) και ενσωματωμένες μεθόδους (embedded methods/Intrinsic methods).



Εικόνα 2 Supervised Models Techniques

Η προσέγγιση ανοικτού βρόχου, γνωστή και ως μέθοδος φιλτραρίσματος (Filtered Method), είναι η τεχνική που χρησιμοποιείται για την αφαίρεση των φιλτραρισμένων δεδομένων με τα λιγότερο σημαντικά χαρακτηριστικά πριν από την κατηγοριοποίηση.

Συγκεκριμένα, τα χαρακτηριστικά κατηγοριοποιούνται ανάλογα με το πόσο καλά συσχετίζονται με το επιθυμητό χαρακτηριστικό σε μια ποικιλία στατιστικών δοκιμών. Τα χαρακτηριστικά που συγκεντρώνουν βαθμολογία χαμηλότερη από ένα προκαθορισμένο κατώφλι εξαλείφονται και επιλέγονται εκείνα που συγκεντρώνουν υψηλότερη βαθμολογία. Ένα υποσύνολο χαρακτηριστικών μπορεί στη συνέχεια να δοθεί στον επιλεγμένο ταξινομητή ως είσοδος αφού αυτός έχει επιλεγεί.

Οι τεχνικές φιλτραρίσματος είναι ξεχωριστές από τον ταξινομητή, σε αντίθεση με τις άλλες τεχνικές επιλογής χαρακτηριστικών (περιτύλιγμα και ενσωμάτωση), οι οποίες εξετάζονται παρακάτω. Λόγω αυτής της διάκρισης, η υπερπροσαρμογή μειώνεται επειδή οι διαδικασίες φιλτραρίσματος δεν επηρεάζονται από την προκατάληψη του ταξινομητή. Αυτή η ανεξαρτησία συνεπάγεται επίσης ότι η διαδικασία επιλογής χαρακτηριστικών δεν λαμβάνει υπόψη την αλληλεπίδραση με τον ταξινομητή. Συνεπώς, το σύνολο χαρακτηριστικών που επιλέχθηκε είναι πιο ευρύ και δεν έχει προσαρμοστεί σε έναν συγκεκριμένο ταξινομητή. Λόγω αυτής της έλλειψης συντονισμού, τα μοντέλα που δημιουργούνται με τεχνικές φιλτραρίσματος έχουν συνήθως χαμηλότερη απόδοση πρόβλεψης από τα μοντέλα που δημιουργούνται με τεχνικές περιτύλιξης ή ενσωμάτωσης. Ωστόσο, οι τεχνικές φιλτραρίσματος έχουν ένα σημαντικό πλεονέκτημα σε σχέση με άλλες τεχνικές επιλογής: απαιτούν λιγότερη υπολογιστική ισχύ, γεγονός που καθιστά ευκολότερη την κλιμάκωσή τους σε δεδομένα εξαιρετικά υψηλών διαστάσεων.

Η μέθοδος περιτύλιξης (Wrapper Method), γνωστή και ως μέθοδος στενού βρόχου, επιλέγει το πιο διακριτό υποσύνολο χαρακτηριστικών που ελαχιστοποιεί το σφάλμα πρόβλεψης ενός συγκεκριμένου ταξινομητή, περιτυλίγοντας τα χαρακτηριστικά επιλογής γύρω από τον αλγόριθμο μάθησης και χρησιμοποιώντας ως κριτήριο την ακρίβεια απόδοσης ή το ποσοστό σφάλματος ταξινόμησης.

Συγκεκριμένα, δεδομένου ενός συγκεκριμένου αλγορίθμου μάθησης, μια τυπική τεχνική περιτύλιξης αποτελείται από δύο βήματα: αναζητά ένα υποσύνολο χαρακτηριστικών και στη συνέχεια αξιολογεί τα χαρακτηριστικά που βρίσκει. Μέχρι να ικανοποιηθούν συγκεκριμένες απαιτήσεις διακοπής, αυτά τα δύο βήματα επαναλαμβάνονται. Συγκεκριμένα, ο αλγόριθμος εκμάθησης λειτουργεί ως ένα μαύρο κουτί για την αξιολόγηση της ποιότητας αυτών των χαρακτηριστικών με βάση την απόδοση εκμάθησης αφού ο τομέας που αναζητά το σύνολο παράγει πρώτα ένα υποσύνολο χαρακτηριστικών. Για παράδειγμα, η όλη διαδικασία επαναλαμβάνεται έως ότου επιτευχθεί ο μαθησιακός στόχος ή επιλεγεί η απαιτούμενη ποσότητα χαρακτηριστικών. Το υποσύνολο χαρακτηριστικών που οδηγεί στην καλύτερη μαθησιακή απόδοση επιστρέφεται στη συνέχεια ως τα επιλεγμένα χαρακτηριστικά.

Το κύριο πλεονέκτημα των μεθόδων περιτύλιξης είναι ότι βρίσκει το σύνολο των χαρακτηριστικών που αποδίδει καλύτερα για τον επιλεγμένο ταξινομητή. Έχει αποδειχθεί ότι με αυτόν τον τρόπο επιτυγχάνεται καλύτερη προβλεπτική απόδοση από ό,τι με τις τεχνικές φιλτραρίσματος. Ένα περαιτέρω πλεονέκτημα είναι ότι οι μέθοδοι περιτύλιξης επιλέγουν το βέλτιστο υποσύνολο λαμβάνοντας έμμεσα υπόψη τις εξαρτήσεις των χαρακτηριστικών, όπως οι αλληλεπιδράσεις και οι πλεονασμοί. Ωστόσο, οι μέθοδοι περιτύλιξης είναι υπολογιστικά βαριές και ο χώρος αναζήτησης είναι πολύ μεγάλος για τις μεθόδους φιλτραρίσματος (σε σύγκριση με τις μεθόδους φιλτραρίσματος και τις ενσωματωμένες μεθόδους), λόγω των πολυάριθμων υπολογισμών που απαιτούνται για τη δημιουργία των υποσυνόλων χαρακτηριστικών και την αξιολόγησή τους. Ωστόσο, ο ταξινομητής που χρησιμοποιείται καθορίζει τις τεχνικές περιτύλιξης. Κατά συνέπεια, δεν υπάρχει καμία βεβαιότητα ότι η χρήση ενός διαφορετικού ταξινομητή θα διατηρούσε τα επιλεγμένα χαρακτηριστικά βέλτιστα. Ένα υποσύνολο χαρακτηριστικών με καλές επιδόσεις μπορεί περιστασιακά να προκύψει από την επιλογή χαρακτηριστικών με βάση την απόδοση του ταξινομητή. Τέλος, επειδή οι μέθοδοι περιτύλιξης πρέπει να επαναλαμβάνονται κάθε φορά που χρησιμοποιείται διαφορετικός αλγόριθμος μάθησης, είναι λιγότερο γενικές από τις μεθόδους φιλτραρίσματος. Επομένως, δεν μπορεί να διασφαλιστεί η καταλληλότητα της λύσης για διαφορετικούς αλγορίθμους μάθησης.

Η αντιστάθμιση μεταξύ των προσεγγίσεων φιλτραρίσματος και περιτύλιξης που περιλαμβάνουν την επιλογή χαρακτηριστικών στην εκμάθηση μοντέλων αντιπροσωπεύεται από ενσωματωμένες μεθόδους (Embedded Methods). Προκειμένου να επιτευχθεί η καλύτερη ακρίβεια ταξινόμησης, ο ταξινομητής τροποποιεί τις εσωτερικές του παραμέτρους καθ' όλη τη διάρκεια του σταδίου εκπαίδευσης και επιλέγει τα κατάλληλα βάρη ή, για να το θέσουμε διαφορετικά, την κατάλληλη σημασία που αποδίδεται σε κάθε χαρακτηριστικό. Ως αποτέλεσμα, με μια ενσωματωμένη τεχνική, η εύρεση του ιδανικού υποσυνόλου χαρακτηριστικών και η κατασκευή του μοντέλου γίνονται ταυτόχρονα. Οι αλγόριθμοι που βασίζονται σε δέντρα, όπως τα δέντρα απόφασης, τα τυχαία δάση (random forest και gradient boosting), και η επιλογή χαρακτηριστικών που χρησιμοποιούν μοντέλα κανονικοποίησης όπως το LASSO και το ελαστικό δίχτυ (elastic net) είναι μερικές περιπτώσεις ενσωματωμένων προσεγγίσεων.

Τα πλεονεκτήματα των μεθόδων φιλτραρίσματος και περιτύλιξης μεταφέρονται στις ενσωματωμένες προσεγγίσεις. Συγκεκριμένα, παραλείπουν την επαναλαμβανόμενη διαδικασία εκτέλεσης του ταξινομητή και ανάλυσης κάθε υποσυνόλου χαρακτηριστικών, καθιστώντας τις υπολογιστικά πιο προσιτές και αποδοτικές από τις τεχνικές περιτύλιξης. Ενσωματώνουν επίσης αλληλεπιδράσεις με τον αλγόριθμο μάθησης. Επιπλέον, σε σύγκριση με τις τεχνικές περιτύλιξης, οι διαδικασίες αυτές έχουν μειωμένη πιθανότητα υπερπροσαρμογής.

2.2.2. Μοντέλα Χωρίς Επίβλεψη (Unsupervised Learning)

Συνήθως, η επιλογή χαρακτηριστικών χωρίς επίβλεψη αποσκοπεί στην αντιμετώπιση ζητημάτων ομαδοποίησης. Πρόσφατα, υπήρξε μεγάλο ενδιαφέρον για τη μη επιβλεπόμενη επιλογή χαρακτηριστικών, επειδή η απόκτηση δεδομένων με ετικέτες είναι πολύ χρονοβόρα και εργατοβόρα. Οι τεχνικές επιλογής χαρακτηριστικών χωρίς επίβλεψη αναζητούν εναλλακτικά κριτήρια για τον χαρακτηρισμό της σημασίας των χαρακτηριστικών ελλείψει ετικετών. Συγκεκριμένα, χρησιμοποιούν δομές δεδομένων όπως η διακύμανση των δεδομένων, η διαχωρισιμότητα των δεδομένων και η κατανομή των δεδομένων για να αξιολογήσουν τη σημασία των χαρακτηριστικών. Η μη επιβλεπόμενη επιλογή χαρακτηριστικών, σε αντίθεση με την επιβλεπόμενη επιλογή χαρακτηριστικών, συχνά χρησιμοποιεί κάθε στιγμιότυπο που έγινε κατά τη διαδικασία επιλογής χαρακτηριστικών. Χρησιμοποιώντας μια συμβατική τεχνική ομαδοποίησης, η δομή ομαδοποίησης όλων των δειγμάτων δεδομένων στα καθορισμένα χαρακτηριστικά εξάγεται στο συμπέρασμα, μετά τη φάση επιλογής χαρακτηριστικών. Οι τεχνικές επιλογής χαρακτηριστικών χωρίς επίβλεψη προσφέρουν δύο βασικά πλεονεκτήματα: πρώτον, είναι αμερόληπτες και αποτελεσματικές όταν δεν υπάρχει προηγούμενη γνώση- δεύτερον, σε αντίθεση με τις τεχνικές με επίβλεψη, μπορούν να μειώσουν την πιθανότητα υπερβολικής προσαρμογής των δεδομένων. Ωστόσο, οι μη επιβλεπόμενες προσεγγίσεις έχουν μειονεκτήματα, εκτός από τα πλεονεκτήματά τους. Το κύριο μειονέκτημα είναι ότι τα υποσύνολα χαρακτηριστικών που παράγονται μπορεί να μην είναι ιδανικά για τη συγκεκριμένη εργασία, καθώς δεν λαμβάνουν υπόψη τους πιθανές συσχετίσεις μεταξύ των χαρακτηριστικών και του επιδιωκόμενου στόχου. Ένα δεύτερο μειονέκτημα είναι ότι, αν και βασίζονται σε μαθηματικές αρχές, δεν υπάρχει καμία βεβαιότητα ότι οι κανόνες αυτοί ισχύουν για όλα τα δεδομένα.

Η επιλογή χαρακτηριστικών είναι θεμελιώδης σε κάθε προσέγγιση μηχανικής μάθησης, καθώς η αποτελεσματικότητα των μοντέλων εξαρτάται σημαντικά από την ποιότητα και τη σχετικότητα των δεδομένων που χρησιμοποιούνται για την εκπαίδευση. Στην περίπτωση της ανάλυσης δεδομένων αγώνων καλαθοσφαίρισης, η επιλογή χαρακτηριστικών μπορεί να επικεντρωθεί σε στατιστικά που αντικατοπτρίζουν την αποδοτικότητα των παικτών, τη στρατηγική των ομάδων, τη φυσική κατάσταση, τις συνθήκες του αγώνα και άλλους παράγοντες που μπορεί να επηρεάζουν το αποτέλεσμα. Η επιτυχής εφαρμογή τεχνικών επιλογής χαρακτηριστικών μπορεί να βελτιώσει σημαντικά την ακρίβεια και την αποδοτικότητα των προβλέψεων, παρέχοντας πολύτιμες διορατικότητες για την ανάλυση και την κατανόηση των αγώνων καλαθοσφαίρισης.

2.3. Sport Analytics και Βιβλιογραφία

Οι αθλητικές αναλύσεις είναι συλλογές σχετικών δεδομένων του παρελθόντος που μπορούν να προσφέρουν σε μια ομάδα ή ένα άτομο ανταγωνιστικό πλεονέκτημα. Η αθλητική ανάλυση συγκεντρώνει πληροφορίες από παίκτες, προπονητές και άλλα μέλη του προσωπικού μέσω της συλλογής και ανάλυσης δεδομένων, επιτρέποντάς τους να λαμβάνουν αποφάσεις τόσο πριν όσο και κατά τη διάρκεια αθλητικών εκδηλώσεων.

Οι αναλύσεις εντός και εκτός γηπέδου είναι οι δύο κύριες πτυχές των αθλητικών αναλύσεων. Ο στόχος των αναλύσεων εντός γηπέδου είναι να βοηθήσουν τις ομάδες και τους παίκτες να αποδώσουν καλύτερα στον αγωνιστικό χώρο. Παραδείγματα ερωτήσεων που εμπίπτουν σε αυτή την κατηγορία είναι τα εξής: "ποιος παίκτης συνέβαλε περισσότερο στην επίθεση της ομάδας;" και "ποιος είναι ο καλύτερος παίκτης στο NBA;". Η εμπορική πτυχή του αθλητισμού είναι το επίκεντρο των αναλυτικών εκτός γηπέδου. Ο στόχος των off-field analytics είναι να χρησιμοποιηθούν δεδομένα για να βοηθήσουν έναν αθλητικό οργανισμό ή φορέα να εντοπίσει τάσεις και ιδέες που θα μπορούσαν να ενισχύσουν τις πωλήσεις εισιτηρίων και εμπορευμάτων, να βελτιώσουν την αλληλεπίδραση των φιλάθλων κ.λπ. Στην ουσία, η off-field analytics χρησιμοποιεί δεδομένα για να υποστηρίξει τους κατόχους δικαιωμάτων στη λήψη αποφάσεων που θα ενισχύσουν τα έσοδα και την ανάπτυξή τους.

Τα τελευταία χρόνια, οι τεχνολογικές εξελίξεις έχουν οδηγήσει σε πιο ολοκληρωμένη και εύκολη συλλογή δεδομένων. Οι αθλητικές αναλύσεις έχουν επεκταθεί λόγω των βελτιώσεων στη συλλογή δεδομένων, οι οποίες έχουν οδηγήσει στη δημιουργία εξελιγμένων στατιστικών και μηχανικής μάθησης, καθώς και τεχνολογιών ειδικά για τα αθλήματα που επιτρέπουν στις ομάδες να εξασκούνται σε προσομοιώσεις παιχνιδιών πριν από τους αγώνες, να βελτιώνουν τις τακτικές απόκτησης οπαδών και μάρκετινγκ, ακόμη και να κατανοούν τις επιπτώσεις της χορηγίας σε κάθε ομάδα και τους υποστηρικτές της.

Ο αθλητικός τζόγος έχει επηρεαστεί σημαντικά από την αθλητική ανάλυση στον επαγγελματικό αθλητισμό. Είτε πρόκειται για στοιχήματα είτε για πρωταθλήματα φανταστικών αθλημάτων, η σε βάθος αθλητική ανάλυση, έχει ανεβάσει τον αθλητικό τζόγο σε νέα ύψη. Οι παίκτες στοιχημάτων έχουν πλέον πρόσβαση σε περισσότερες πληροφορίες που τους βοηθούν να λαμβάνουν καλύτερες αποφάσεις. Πολλές επιχειρήσεις και ιστότοποι έχουν δημιουργηθεί για να βοηθήσουν στην παροχή στους φιλάθλους των πιο πρόσφατων δυνατών πληροφοριών για τις στοιχηματικές τους απαιτήσεις.

Ο Ντάριλ Μόρεϊ των Χιούστον Ρόκετς ήταν ο πρώτος γενικός διευθυντής του NBA που συμπεριέλαβε τις προηγμένες μετρήσεις ως κρίσιμο στοιχείο της αξιολόγησης των παικτών. Μετά την πρόσληψη του

Μόρεϊ, το NBA εφάρμοσε γρήγορα διαδικασίες αξιολόγησης παικτών που βασίζονται σε προηγμένες μετρήσεις. Οι μεγάλες ιστοσελίδες αθλητικών μέσων ενημέρωσης, όπως το Basketball Reference, έχουν δεσμευτεί να συγκεντρώνουν, να συνθέτουν και να διαδίδουν προηγμένα στατιστικά στοιχεία στους οπαδούς, στα μέλη των αθλητικών μέσων ενημέρωσης, στις ομάδες επαγγελματικού και κολεγιακού μπάσκετ και στα front offices του επαγγελματικού μπάσκετ.

Τα περισσότερα καλάθια στις πρώτες μέρες της καλαθοσφαίρισης γίνονταν σε κοντινή απόσταση από το στεφάνι. Η γραμμή των τριών πόντων χρησιμοποιείται πλέον από το NBA και άλλα πρωταθλήματα, επιτρέποντας στους παίκτες να σουτάρουν από μεγαλύτερη απόσταση για τρεις πόντους αντί για δύο. Ως αποτέλεσμα, οι παίκτες είναι πολύ πιο πολυδιάστατοι και δύσκολοι στην άμυνα. Χρησιμοποιώντας αναλυτικά στοιχεία και τεχνητή νοημοσύνη, οι αμυντικοί μπορούν να διδαχθούν πώς να καλύψουν έναν παίκτη με βάση το πόσο επιτυχημένα σουτάρουν τρεις πόντους. Η άμυνα μπορεί να υποχωρήσει και να εγκαταλείψει το σουτ αν δεν είναι πολύ καλοί σουτέρ τριών πόντων. Η προπόνηση έχει επίσης επηρεαστεί σημαντικά από την τεχνητή νοημοσύνη και την ανάλυση. Παραδείγματα είναι οι καταστάσεις στο τέλος του παιχνιδιού, η χρήση των τάιμ άουτ, η αμυντική τακτική και η επίδραση των παικτών. Προκειμένου να βοηθήσουν τον προπονητή να κάνει προσαρμογές κατά τη διάρκεια του παιχνιδιού, ορισμένοι σύλλογοι του NBA διαθέτουν προπονητές που επικεντρώνονται κυρίως στα δεδομένα και την ανάλυση.

2.3.1. Αθλητική Αναλυτική (Sports Analytics)

Η αθλητική αναλυτική είναι η χρήση της μοντελοποίησης βάσει δεδομένων σε αθλητικά γεγονότα, η οποία περιλαμβάνει τη διαχείριση της απόδοσης των παικτών και τη γνώση της στρατηγικής και της τακτικής της ομάδας. Σε πολλές περιπτώσεις, φορητοί αισθητήρες χρησιμοποιούνται σε συνδυασμό με κάμερες πολλαπλών προβολών - οι οποίες καταγράφουν ολόκληρο το γήπεδο και χρησιμοποιούνται συνήθως για τις κινήσεις των παικτών και της μπάλας στα επαγγελματικά ομαδικά αθλήματα - για τη μέτρηση των κινήσεων των παικτών, του σωματικού φορτίου και της σωματικής καταπόνησης κατά τη διάρκεια των συγκρούσεων. Όλοι οι σύλλογοι ενδιαφέρονται να χρησιμοποιήσουν την ανάλυση δεδομένων για να δώσουν στις ομάδες μπάσκετ τους ανταγωνιστικό πλεονέκτημα, επειδή αυτό συνδέεται με την οικονομική τους επιτυχία. Ένας επιστημονικός τομέας που ξεκίνησε με μη ακαδημαϊκή εργασία, η ποσοτική ανάλυση των αθλημάτων, ιδίως της καλαθοσφαίρισης, έχει προσελκύσει μεγάλη προσοχή από τους ακαδημαϊκούς τα τελευταία δέκα χρόνια. Ξεκινώντας με τους Albert, Bennett και Cochran (2005), συγκέντρωσαν τις προηγούμενες δημοσιεύσεις σχετικά με την εφαρμογή της στατιστικής στην αθλητική ανάλυση. Κάθε ένα από τα τέσσερα μεγάλα ομαδικά αθλήματα -μπέιζμπολ, ποδόσφαιρο, μπάσκετ και

χόκεϊ επί πάγου- έχει το δικό του μέρος στο βιβλίο τους. Ο Winston (2009) εισάγει τον αναγνώστη σε μια ποικιλία μοντέλων αθλητικής ανάλυσης, αφιερώνοντας σημαντικό μέρος του βιβλίου του στην καλαθοσφαίριση. Σημειώνει ότι η διοίκηση των ομάδων χρησιμοποιεί πραγματικά μερικά από αυτά τα μοντέλα για να βοηθήσει στη λήψη αποφάσεων. Οι Berri και Schmidt (2010) εμπνέονται από την πραγματικότητα ότι οι άνθρωποι συνήθως δυσκολεύονται να κάνουν τις σωστές επιλογές. Προσφέρουν διάφορες αφηγήσεις, υποστηριζόμενες από ερευνητικά δεδομένα, οι οποίες, όπως υποστηρίζουν, πρέπει να επηρεάσουν όχι μόνο τον τρόπο με τον οποίο οι οπαδοί αντιλαμβάνονται τις αποφάσεις των ομάδων που προτιμούν, αλλά και τον τρόπο με τον οποίο οι οικονομολόγοι και άλλοι κοινωνικοί επιστήμονες αντιλαμβάνονται την ανθρώπινη λήψη αποφάσεων.

2.3.2. Χαρακτηριστικά της Καλαθοσφαίρισης (Basketball Attributes)

Οι μεταβλητές του παιχνιδιού έχουν μετρηθεί με ποικίλες τεχνικές καταγραφής και ανάλυσης με την πάροδο των ετών, που κυμαίνονται από απλές στατιστικές φόρμες που συμπληρώνονται από βοηθούς προπονητές έως πλήρως ηλεκτρονικές διαδικασίες που καταγράφουν όλους τους κρίσιμους δείκτες απόδοσης του παιχνιδιού. (Oliver, 2004) Προηγούμενες έρευνες με τη χρήση στατιστικών στοιχείων που σχετίζονται με το παιχνίδι έδειξαν ότι τα αμυντικά ριμπάουντ, τα ποσοστά 2 πόντων και οι ασίστ είναι οι πρωταρχικοί παράγοντες που διαχωρίζουν τις νικήτριες ομάδες από τις ηττημένες. (Ibáñez και άλλοι, 2008) Πιο πρόσφατες μελέτες αποκαλύπτουν ότι μεταβλητές όπως η τοποθεσία του παιχνιδιού (εντός ή εκτός έδρας), το είδος του (κανονική περίοδος ή πλέι οφ) και οι διακυμάνσεις στο τελικό σκορ του παιχνιδιού έχουν σημαντικό αντίκτυπο στα στατιστικά στοιχεία που αφορούν την απόδοση της ομάδας (Puente, Coso, Salinero & Abián-Vicén., 2015). Τα δύο μεγαλύτερα πρωταθλήματα μπάσκετ στον κόσμο, το NBA και η Ευρωλίγκα, δεν έχουν γενικά αποτελέσει αντικείμενο πολλών ερευνών. Κατ' αρχάς, υπάρχουν πολλοί δομικοί παραλληλισμοί μεταξύ των επιθέσεων μπάσκετ του NBA και της Ευρωλίγκας, ιδίως όσον αφορά τη δυναμική και τον ρυθμό του παιχνιδιού. Το 2021, οι Selmanovic, Škegro και Milanović διαπίστωσαν δύο βασικές διακρίσεις μεταξύ του NBA και του ευρωπαϊκού μπάσκετ στην ανάλυση της τεχνικής σουτ: το NBA έχει περισσότερα καρφώματα από το ευρωπαϊκό μπάσκετ και έχει λιγότερα σουτ. Αυτές οι διαφορές μπορούν να συνδεθούν με την ανώτερη αθλητικότητα των παικτών του NBA.
(àtrumbelj & Erčulj, 2015). Σε αντίστοιχα συμπεράσματα μεταξύ του παιχνιδιού της Καλαθοσφαίρισης σε διαφορετικές διοργανώσεις κατέληξαν και οι (Mavridis, Tsamourtzis, Karipidis & Laios, 2009).

2.3.3. Διαθέσιμα Σύνολα Δεδομένων

Για την υποστήριξη της έρευνας στον τομέα της αναλυτικής στον αθλητισμό, και ειδικότερα στην καλαθοσφαίριση, η πρόσβαση σε πλούσια και αξιόπιστα σύνολα δεδομένων είναι κρίσιμη. Αυτά τα δεδομένα μπορούν να περιλαμβάνουν στατιστικά αγώνων, αποδοτικότητα παικτών, ιστορικά αποτελέσματα, και άλλες μετρήσεις που είναι χρήσιμες για αναλύσεις και προβλέψεις. Σύνολα δεδομένων χρησιμοποιήθηκαν από τη Euroleague Basketball Statistics (https://www.euroleaguebasketball.net/euroleague/stats/η οποία παρέχει τα επίσημα στατιστικά για την Euroleague και το Eurocup, συμπεριλαμβανομένων στατιστικών αγώνων, παικτών και ομάδων. Επίσης η πλατφόρμα Kaggle (https://www.kaggle.com) περιλαμβάνει πολλά σύνολα δεδομένων που αφορούν τον αθλητισμό, συμπεριλαμβανομένων και της καλαθοσφαίρισης, τα οποία μπορούν να χρησιμοποιηθούν για ακαδημαϊκούς αλλά και ερευνητικούς σκοπούς. Τέλος σύνολα δεδομέων χρησιμοποιήθηκαν από άλλες πλατφόρμες όπως η basketball-reference (https://www.basketball-reference.com/international/euroleague/) οι οποίες κρατάνε μεγαλύτερη ιστορικότητα δεδομένων που θα είναι απολύτως χρήσιμα για την συνέχεια.

3. Προτεινόμενη Μεθοδολογία

Η προτεινόμενη μεθοδολογία για την ανάλυση δεδομένων αγώνων καλαθοσφαίρισης με σκοπό την πρόβλεψη αποτελεσμάτων και την επιλογή χαρακτηριστικών αποτελείται από διάφορα βήματα, τα οποία ενσωματώνουν τεχνικές μηχανικής μάθησης και στατιστικής ανάλυσης. Η μεθοδολογία αυτή στοχεύει στην αποδοτική επεξεργασία και ανάλυση των δεδομένων, την ακριβή πρόβλεψη των αποτελεσμάτων των αγώνων και την εύρεση σημαντικών χαρακτηριστικών που επηρεάζουν την επίδοση.

3.1. Τεχνικές που Χρησιμοποιήθηκαν

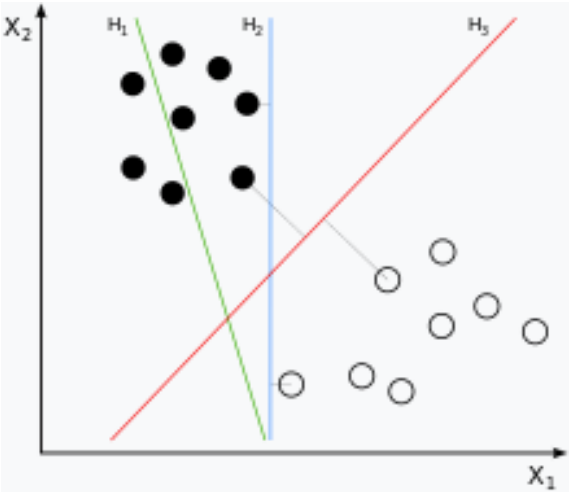
Σε αυτό το κεφάλαιο αναλύονται οι μέθοδοι και οι τεχνικές που χρησιμοποιήθηκαν για την ανάλυση των δεδομένων της διπλωματικής εργασίας.

3.1.1. Support Vector Machine (SVM)

Οι μηχανές διανυσμάτων υποστήριξης (SVM, επίσης γνωστές ως δίκτυα διανυσμάτων υποστήριξης) είναι εποπτευόμενα μοντέλα μέγιστου περιθωρίου που χρησιμοποιούνται στη μηχανική μάθηση και εξετάζουν δεδομένα για παλινδρόμηση και ταξινόμηση μαζί με αντίστοιχες τεχνικές μάθησης. Οι SVMs είναι ένα από τα πιο μελετημένα μοντέλα- αναπτύχθηκαν στα εργαστήρια AT&T Bell Laboratories από τον Vladimir Vapnik και τους συνεργάτες του (Boser et al., 1992, Guyon et al., 1993, Cortes and Vapnik, 1995,[1] Vapnik et al., 1997). Οι Vapnik και Chervonenkis (1974) πρότειναν τα πλαίσια στατιστικής μάθησης ή τη θεωρία VC που αποτέλεσε τη βάση για τις SVM.

Χρησιμοποιώντας μια τεχνική γνωστή ως "τέχνασμα του πυρήνα", οι SVM μπορούν να διεξάγουν αποτελεσματικά μη γραμμική ταξινόμηση εκτός από τη γραμμική ταξινόμηση, μεταφράζοντας σιωπηρά τις εισόδους τους σε χώρους χαρακτηριστικών υψηλών διαστάσεων. Οι εργασίες παλινδρόμησης, όταν ο στόχος γίνεται e- ευαίσθητος, μπορούν επίσης να αντιμετωπιστούν από SVMs. Η Hava Siegelmann και ο Vladimir Vapnik επινόησαν την τεχνική ομαδοποίησης διανυσμάτων υποστήριξης, η οποία χρησιμοποιεί τη στατιστική των διανυσμάτων υποστήριξης - τα οποία προέρχονται από τον αλγόριθμο των μηχανών διανυσμάτων υποστήριξης - για την ταξινόμηση μη επισημασμένων δεδομένων. Οι τεχνικές μάθησης χωρίς επίβλεψη είναι απαραίτητες για τεράστια σύνολα δεδομένων. Οι τεράστιες τεχνικές αναζητούν φυσικές ομαδοποιήσεις στα δεδομένα και στη συνέχεια αντιστοιχίζουν τα νέα δεδομένα σε αυτές τις ομάδες. Η δημοτικότητα των Μηχανών Διανυσμάτων Υποστήριξης (SVM) μπορεί

πιθανώς να αποδοθεί στην ευκολία της θεωρητικής μελέτης τους και στην ευελιξία τους στον χειρισμό ενός ευρέος φάσματος εργασιών, συμπεριλαμβανομένων των προβλημάτων δομημένης πρόβλεψης. Το κατά πόσον οι SVMs υπερτερούν έναντι άλλων γραμμικών μοντέλων, όπως η λογιστική παλινδρόμηση και η γραμμική παλινδρόμηση, όσον αφορά την πρόβλεψη είναι ασαφές.



Εικόνα 3 Support Vector Machine (SVM)

3.1.2. Λογιστική Παλινδρόμιση (Logistic Regression)

Το λογιστικό μοντέλο, συχνά γνωστό ως μοντέλο logit, είναι ένα στατιστικό μοντέλο που χρησιμοποιείται στη στατιστική και αναπαριστά τις λογαριθμικές πιθανότητες ενός γεγονότος ως γραμμικό συνδυασμό μιας ή περισσότερων ανεξάρτητων μεταβλητών. Η παλινδρόμηση logit, μερικές φορές γνωστή ως λογιστική παλινδρόμηση, είναι μια τεχνική ανάλυσης παλινδρόμησης που χρησιμοποιείται για την εκτίμηση των παραμέτρων ενός λογιστικού μοντέλου ή των συντελεστών του γραμμικού συνδυασμού. Στη δυαδική λογιστική παλινδρόμηση, οι ανεξάρτητες μεταβλητές μπορεί να είναι συνεχείς (οποιαδήποτε πραγματική τιμή) ή δυαδικές (δύο κλάσεις, κωδικοποιημένες από μια μεταβλητή δείκτη). Η εξαρτημένη μεταβλητή στη δυαδική λογιστική παλινδρόμηση είναι τυπικά μια ενιαία δυαδική μεταβλητή, κωδικοποιημένη από μια μεταβλητή δείκτη, με τις δύο τιμές να επισημαίνονται ως "0" και "1". Ο χαρακτηρισμός μιας τιμής ως "1" βασίζεται στη σχετική πιθανότητα, η οποία μπορεί να κυμαίνεται από 0 (που είναι σίγουρα η τιμή "0") έως 1 (που είναι σίγουρα η τιμή "1"). Η λογιστική συνάρτηση είναι η συνάρτηση που μετατρέπει τις λογαριθμικές πιθανότητες σε πιθανότητες, γι' αυτό και πήρε το όνομά της. Τα εναλλακτικά ονόματα για την κλίμακα log-odds προέρχονται από τον όρο "logit", ο οποίος προέρχεται

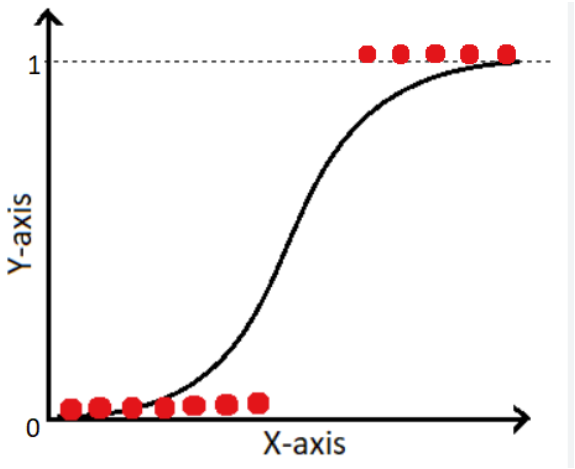
από τη λέξη "λογιστική μονάδα".

Το λογιστικό μοντέλο είναι το πιο δημοφιλές μοντέλο δυαδικής παλινδρόμησης από το 1970 περίπου. Οι δυαδικές μεταβλητές χρησιμοποιούνται ευρέως στη στατιστική για τη μοντελοποίηση της πιθανότητας να λάβει χώρα μια συγκεκριμένη κατηγορία ή γεγονός, όπως η πιθανότητα να κερδίσει μια ομάδα, να είναι υγιής ένας ασθενής κ.λπ. Όταν υπάρχουν περισσότερες από δύο πιθανές τιμές (όπως αν μια εικόνα είναι γάτα, σκύλος, λιοντάρι κ.λπ.), οι δυαδικές μεταβλητές μπορούν να γενικευτούν σε κατηγορικές μεταβλητές και η δυαδική λογιστική παλινδρόμηση μπορεί να γενικευτεί σε πολυωνυμική λογιστική παλινδρόμηση. Μπορεί κανείς να εφαρμόσει την τακτική λογιστική παλινδρόμηση (για παράδειγμα, το αναλογικό λογαριθμικό μοντέλο τακτικών αποδόσεων) εάν οι πολυάριθμες κατηγορίες είναι ταξινομημένες. Αν και μπορεί να χρησιμοποιηθεί για τη δημιουργία ενός ταξινομητή, όπως με την επιλογή μιας τιμής αποκοπής και την ταξινόμηση των εισόδων με πιθανότητα μεγαλύτερη από την αποκοπή ως μία κλάση και κάτω από την αποκοπή ως την άλλη -αυτή είναι μια κοινή μέθοδος δημιουργίας ενός δυαδικού ταξινομητή-, το ίδιο το μοντέλο λογιστικής παλινδρόμησης μοντελοποιεί μόνο την πιθανότητα εξόδου σε σχέση με την είσοδο και δεν εκτελεί στατιστική ταξινόμηση (δεν είναι ταξινομητής).

Είναι επίσης δυνατό να χρησιμοποιηθούν άλλα γραμμικά μοντέλα, κυρίως το μοντέλο probit, τα οποία είναι παρόμοια με τη λογιστική συνάρτηση, αλλά χρησιμοποιούν μια διαφορετική σιγμοειδή συνάρτηση για να μετατρέψουν το γραμμικό συνδυασμό σε πιθανότητα για δυαδικές μεταβλητές. Το βασικό χαρακτηριστικό του λογιστικού μοντέλου είναι ότι, για μια δυαδική εξαρτημένη μεταβλητή, αυτό γενικεύει τον λόγο πιθανοτήτων. Η αύξηση μιας από τις ανεξάρτητες μεταβλητές κλιμακώνει πολλαπλασιαστικά τις πιθανότητες του συγκεκριμένου αποτελέσματος με σταθερό ρυθμό. Κάθε ανεξάρτητη μεταβλητή έχει μια παράμετρο. Με μια γενικότερη έννοια, η λογιστική συνάρτηση είναι η "απλούστερη" μέθοδος μετατροπής ενός πραγματικού αριθμού σε πιθανότητα, δεδομένου ότι είναι η φυσική παράμετρος για την κατανομή Bernoulli. Συγκεκριμένα, μειώνει την πρόσθετη πληροφορία μεγιστοποιώντας την εντροπία, κάνοντας έτσι τις λιγότερες υποθέσεις για τα δεδομένα που μοντελοποιούνται.

Η πιο δημοφιλής μέθοδος για την εκτίμηση των παραμέτρων μιας λογιστικής παλινδρόμησης είναι η εκτίμηση μέγιστης πιθανοφάνειας (MLE). Σε αντίθεση με τα γραμμικά ελάχιστα τετράγωνα, αυτή δεν έχει έκφραση κλειστής μορφής. Για δυαδικές ή κατηγορικές αποκρίσεις, η λογιστική παλινδρόμηση με τα μέσα ελάχιστα τετράγωνα (MLE) εξυπηρετεί έναν παρόμοιο θεμελιώδη σκοπό με τη γραμμική παλινδρόμηση με τα συνήθη ελάχιστα τετράγωνα (OLS) για κλιμακωτές αποκρίσεις: είναι ένα απλό, διεξοδικά εξεταζόμενο βασικό μοντέλο. Αρχικά δημιουργήθηκε και ήταν σε μεγάλο βαθμό υπεύθυνος για τη διάδοση της λογιστικής παλινδρόμησης ως καθολικού στατιστικού μοντέλου του Joseph Berkson,

ξεκινώντας από το Berkson (1944) όταν επινόησε τον όρο "logit".



Εικόνα 4 Λογιστική Παλινδρόμηση (Logistic Regression)

3.1.3. k-Κοντινότεροι Γείτονες (KNN)

Ο αλγόριθμος k-κοντινότερων γειτόνων ή k-NN είναι μια μη παραμετρική τεχνική μάθησης με επίβλεψη στη στατιστική, η οποία δημιουργήθηκε αρχικά το 1951 από τους Evelyn Fix και Joseph Hodges και στη συνέχεια βελτιώθηκε περαιτέρω από τον Thomas Cover. Η παλινδρόμηση και η ταξινόμηση είναι δύο χρήσεις του. Η είσοδος και στα δύο σενάρια αποτελείται από τα k πλησιέστερα δείγματα εκπαίδευσης ενός συνόλου δεδομένων. Το αν το k-NN χρησιμοποιείται για παλινδρόμηση ή ταξινόμηση καθορίζει το αποτέλεσμα:

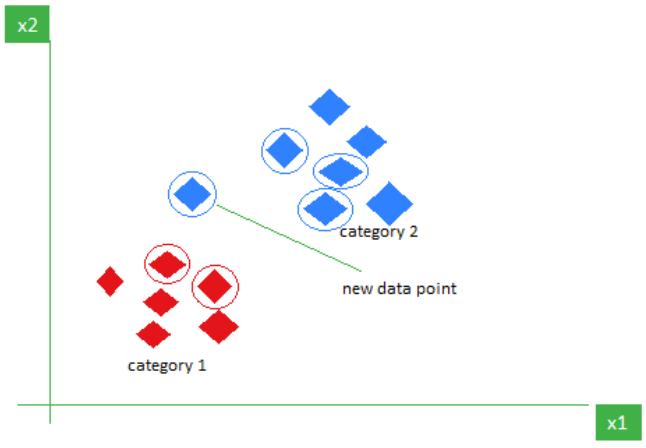
Η συμμετοχή σε μια κλάση είναι το αποτέλεσμα της διαδικασίας ταξινόμησης k-NN. Ένα στοιχείο κατατάσσεται στην κλάση που είναι πιο συχνή μεταξύ των k πλησιέστερων γειτόνων του (το k είναι ένας θετικός αριθμός, συνήθως μικρός) με βάση την ψήφο πλειονότητας των γειτόνων του. Το αντικείμενο τοποθετείται απλώς στην κλάση αυτού του ενός πλησιέστερου γείτονα εάν k = 1.

Η τιμή της ιδιότητας του αντικειμένου είναι το αποτέλεσμα της παλινδρόμησης k-NN. Ο μέσος όρος των τιμών των k πλησιέστερων γειτόνων είναι αυτή η τιμή. Η έξοδος τίθεται απλώς στην τιμή του ενός μόνο πλησιέστερου γείτονα, εάν k = 1.

Με την ταξινόμηση k-NN, δεν γίνεται κανένας υπολογισμός μέχρι να αξιολογηθεί η συνάρτηση-

αντίθετα, η συνάρτηση απλώς προσεγγίζεται τοπικά. Δεδομένου ότι αυτή η μέθοδος χρησιμοποιεί την απόσταση για την ταξινόμηση των δεδομένων, η κανονικοποίηση των δεδομένων εκπαίδευσης μπορεί να αυξήσει σημαντικά την ακρίβειά της εάν τα χαρακτηριστικά έχουν διάφορες φυσικές μονάδες ή φθάνουν σε πολύ διαφορετικές κλίμακες. Η ανάθεση βαρών στις συνεισφορές των γειτόνων -έτσι ώστε οι γείτονες που βρίσκονται πιο κοντά να συνεισφέρουν περισσότερο στο μέσο όρο από τους γείτονες που βρίσκονται πιο μακριά- μπορεί να είναι μια χρήσιμη στρατηγική τόσο για την παλινδρόμηση όσο και για την κατηγοριοποίηση. Ένα τυπικό σύστημα στάθμισης, για παράδειγμα, είναι να παρέχεται ένα βάρος 1/d σε κάθε γείτονα, όπου d είναι η απόσταση του γείτονα.

Όταν χρησιμοποιείται ταξινόμηση k-NN ή παλινδρόμηση k-NN, οι γείτονες επιλέγονται από ένα σύνολο αντικειμένων για τα οποία είναι γνωστή η τιμή της κλάσης ή της ιδιότητας του αντικειμένου. Αν και δεν υπάρχει ανάγκη για ένα ρητό βήμα εκπαίδευσης, αυτό μπορεί να θεωρηθεί ως το σύνολο εκπαίδευσης του αλγορίθμου. Η τεχνική k-NN έχει το χαρακτηριστικό ότι είναι ευαίσθητη στην τοπική δομή των δεδομένων.



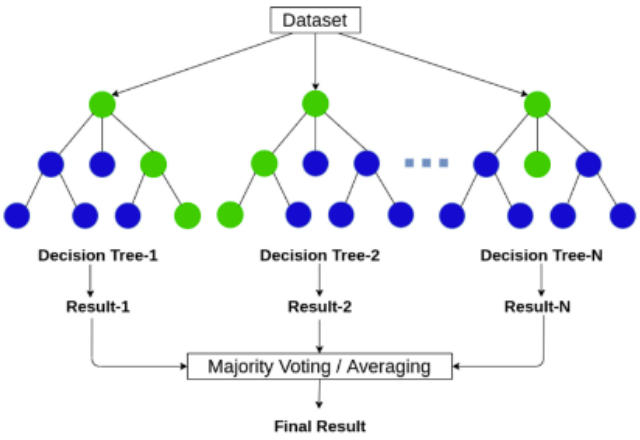
Εικόνα 5 k-Κοντινότεροι Γείτονες (KNN)

3.1.4. Τυχαία Δάση (Random Forest)

Μια δημοφιλής τεχνική μάθησης συνόλου για ταξινόμηση, παλινδρόμηση και άλλα προβλήματα ονομάζεται τυχαία δάση ή τυχαία δάση απόφασης. Λειτουργεί με τη δημιουργία ενός μεγάλου αριθμού δέντρων απόφασης κατά τη φάση της εκπαίδευσης. Η κλάση που επιλέγει η πλειοψηφία των

δέντρων είναι η έξοδος του τυχαίου δάσους για προβλήματα ταξινόμησης. Ο μέσος όρος ή η μέση πρόβλεψη που γίνεται από κάθε μεμονωμένο δέντρο επιστρέφεται για εργασίες παλινδρόμησης. Η τάση των δέντρων απόφασης να προσαρμόζονται υπερβολικά στα σύνολα εκπαίδευσής τους αντισταθμίζεται από τα τυχαία δάση απόφασης.

Ο Tin Kam Ho ανέπτυξε τον πρώτο αλγόριθμο τυχαίου δάσους αποφάσεων το 1995, χρησιμοποιώντας τη μέθοδο τυχαίου υποδιαστήματος, η οποία είναι ένα μέσο για την εφαρμογή στην πράξη της προσέγγισης "στοχαστικής διάκρισης" του Eugene Kleinberg για την ταξινόμηση. Ο Leo Breiman και η Adele Cutler δημιούργησαν μια επέκταση του αλγορίθμου και κατέθεσαν αίτηση για εμπορικό σήμα για το "Random Forests" το 2006- από το 2019, η Minitab, Inc. είναι ο κάτοχος αυτού του εμπορικού σήματος. Η επέκταση δημιουργεί ένα σύνολο δέντρων απόφασης με ελεγχόμενη διακύμανση συνδυάζοντας την έννοια "bagging" του Breiman με την τυχαία επιλογή χαρακτηριστικών, η οποία παρουσιάστηκε αρχικά από τον Ho και στη συνέχεια ξεχωριστά από τους Amit και Geman.



Εικόνα 6 Τυχαία Δάση (Random Forest)

3.2. Μετρικές Αξιολόγησης Κατηγοριοποιητών

Η αξιολόγηση της απόδοσης των κατηγοριοποιητών στη μηχανική μάθηση είναι κρίσιμη για την κατανόηση της αποτελεσματικότητας των μοντέλων πρόβλεψης. Υπάρχουν διάφορες μετρικές αξιολόγησης που μπορούν να χρησιμοποιηθούν για τον σκοπό αυτό, καθεμία από τις οποίες παρέχει διαφορετικές πληροφορίες σχετικά με την απόδοση του μοντέλου. Οι πιο συνηθισμένες μετρικές περιλαμβάνουν την Ακρίβεια (Accuracy), την Ανάκληση (Recall), την Προσθετική Αξία (Precision), το F1

Score, και την Καμπύλη ROC (Receiver Operating Characteristic).

3.2.1. Ακρίβεια (Accuracy)

Η ακρίβεια είναι η πιο άμεση μετρική και υπολογίζεται ως το ποσοστό των σωστών προβλέψεων από το σύνολο όλων των προβλέψεων. Είναι χρήσιμη όταν οι κλάσεις είναι ισορροπημένες, αλλά μπορεί να είναι παραπλανητική σε περιπτώσεις ανισορροπίας κλάσεων.

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

Εικόνα 7 Ακρίβεια (Accuracy)

3.2.2. Προσθετική Αξία (Precision)

Το μέτρο αξιολόγησης Προσθετική Αξία (Precision) μπορεί να οριστεί ως ο αριθμός των σωστών θετικών προβλέψεων σε σχέση με όλες τις θετικές προβλέψεις που έκανε το μοντέλο. (Novakovic et al., 2021)

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

Εικόνα 8 Προσθετική Αξία (Precision)

3.2.3. Ανάκληση (Recall)

Η τιμή ανάκλησης, επίσης γνωστή ως ευαισθησία, μετρά το ποσοστό των πραγματικά θετικών περιπτώσεων που έχουν προβλεφθεί σωστά από το μοντέλο.. Η ανάκληση πρέπει να είναι όσο το δυνατόν μεγαλύτερη. (Powers, 2007)

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

Εικόνα 9 Ανάκληση (Recall)

3.2.4. F1 – Score

Το F1 Score είναι η αρμονική μέση της προσθετικής αξίας και της ανάκλησης, προσφέροντας μια ισορροπημένη μετρική αξιολόγησης μεταξύ της ακρίβειας της πρόβλεψης και της ολοκληρωτικότητας της κάλυψης των θετικών περιπτώσεων.

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

Εικόνα 10 F1 – Score

3.2.5. Καμπύλη ROC και Εμβαδόν Κάτω από την Καμπύλη (AUC)

Η καμπύλη ROC είναι ένα γράφημα που δείχνει την απόδοση ενός κατηγοριοποιητή σε όλα τα δυνατά κατώφλια ταξινόμησης, συγκρίνοντας το ποσοστό των πραγματικά θετικών προβλέψεων με το ποσοστό των ψευδώς θετικών προβλέψεων. Το AUC (Area Under the Curve) μετρά το συνολικό εμβαδόν κάτω από την καμπύλη ROC, παρέχοντας μια συνολική μέτρηση της απόδοσης του μοντέλου.

Η επιλογή της κατάλληλης μετρικής αξιολόγησης εξαρτάται από την ειδική εφαρμογή και τους στόχους της μοντελοποίησης. Σε περιπτώσεις όπου η ισορροπία μεταξύ των κλάσεων είναι σημαντική ή όταν η κόστος των ψευδώς θετικών και ψευδώς αρνητικών προβλέψεων είναι διαφορετικό, η χρήση πιο εξειδικευμένων μετρικών όπως το F1 Score ή το AUC μπορεί να προσφέρει πιο πλήρη εικόνα της απόδοσης του μοντέλου.

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

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

Εικόνα 11 Καμπύλη ROC και Εμβαδόν Κάτω από την Καμπύλη (AUC)

3.3. Τεχνικές Προεπεξεργασίας Δεδομένων

Η προεπεξεργασία δεδομένων αποτελεί ένα κρίσιμο βήμα στη διαδικασία της μηχανικής μάθησης, καθώς η ποιότητα και η δομή των δεδομένων μπορούν να επηρεάσουν σημαντικά την απόδοση των μοντέλων. Η προεπεξεργασία περιλαμβάνει μια σειρά από τεχνικές που στοχεύουν στη βελτίωση της "καθαρότητας", της συνέπειας και της σχετικότητας των δεδομένων πριν από την εφαρμογή αλγορίθμων μηχανικής μάθησης. Παρακάτω παρουσιάζονται βασικές τεχνικές προεπεξεργασίας δεδομένων:

3.3.1. Κανονικοποίηση τιμών (Standardization)

Υπάρχει η τάση για το εύρος τιμών των πρωτογενών δεδομένων να έχει διακριτές κλίμακες. Σε ένα τέτοιο σενάριο, χωρίς κανονικοποίηση των δεδομένων, οι συναρτήσεις στόχου σε ορισμένους αλγορίθμους μηχανικής μάθησης δεν θα λειτουργούν αποτελεσματικά. Για παράδειγμα, πολλά μοντέλα και ταξινομητές υπολογίζουν την ευκλείδεια απόσταση μεταξύ δύο σημείων. Σε περίπτωση που μια παράμετρος των δεδομένων παρουσιάζει ένα ευρύ φάσμα τιμών, η υπολογιζόμενη απόσταση θα επηρεαστεί από το συγκεκριμένο χαρακτηριστικό. Για το λόγο αυτό, τα εύρη όλων των χαρακτηριστικών πρέπει να κλιμακωθούν ή να κανονικοποιηθούν έτσι ώστε όλες οι τιμές τους να εμπίπτουν στο ίδιο εύρος. Διαιρούμε κάθε βαθμολογία με την τυπική απόκλιση των βαθμολογιών του συνόλου, προκειμένου να κανονικοποιήσουμε τις βαθμολογίες χρησιμοποιώντας την τυποποιημένη

τυπική απόκλιση. Στην περίπτωση αυτή, συνήθως διαιρούμε πάντα με την τυπική απόκλιση αφού αφαιρέσουμε τον μέσο όρο των βαθμολογιών από κάθε βαθμολογία. Τα δεδομένα αναβαθμίζονται χρησιμοποιώντας τις μετρήσεις του μέσου όρου και της τυπικής απόκλισης, με αποτέλεσμα να προκύπτουν χαρακτηριστικά με μοναδιαία διακύμανση και μηδενική μέση τιμή.

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

Εικόνα 12 Κανονικοποίηση τιμών (Standardization)

3.3.2. Καθαρισμός Δεδομένων

Ο καθαρισμός δεδομένων είναι μια θεμελιώδης διαδικασία στην προεπεξεργασία δεδομένων, η οποία στοχεύει στην αφαίρεση ή διόρθωση των ανωμαλιών και των λαθών από τα δεδομένα πριν από την ανάλυση ή την εκπαίδευση των μοντέλων μηχανικής μάθησης. Αυτή η διαδικασία βελτιώνει την ποιότητα των δεδομένων και εξασφαλίζει ότι τα μοντέλα που αναπτύσσονται είναι ακριβή και αξιόπιστα. Παρακάτω παρουσιάζονται βασικές τεχνικές καθαρισμού δεδομένων:

- Αντιμετώπιση Απουσιάζων Τιμών

Αφαίρεση: Διαγραφή των γραμμών ή των στηλών που περιέχουν απουσιάζουσες τιμές. Αυτή η προσέγγιση είναι απλή αλλά μπορεί να οδηγήσει στην απώλεια σημαντικών δεδομένων.

Αντικατάσταση: Υποκατάσταση των απουσιάζων τιμών με έναν προκαθορισμένο αριθμό (π.χ., το μέσο όρο, τη διάμεσο ή την πιο συχνή τιμή της στήλης). Αυτή η μέθοδος διατηρεί τα δεδομένα αλλά μπορεί να εισάγει προκαταλήψεις.

- Διόρθωση Λαθών και Ανωμαλιών

Εντοπισμός και Διόρθωση Λαθών: Χρήση λογικών κανόνων ή αλγορίθμων ανίχνευσης ανωμαλιών για τον εντοπισμό και τη διόρθωση λαθών στα δεδομένα, όπως λανθασμένες εγγραφές ή ακατάλληλες τιμές.

Αντιμετώπιση Ακραίων Τιμών: Αναγνώριση και διαχείριση των ακραίων τιμών (outliers) μέσω τεχνικών όπως η τριμμένη μέση τιμή ή η χρήση των διαστημάτων εμπιστοσύνης.

3.3.3. Συντελεστής Συσχέτισης του Pearson (Pearson Correlation)

Ένας συντελεστής συσχέτισης που χρησιμοποιείται στη στατιστική για την αξιολόγηση της γραμμικής συσχέτισης μεταξύ δύο συνόλων δεδομένων είναι ο συντελεστής συσχέτισης Pearson (PCC). Πρόκειται ουσιαστικά για μια κανονικοποιημένη μέτρηση της συνδιακύμανσης, με τιμή που κυμαίνεται πάντα μεταξύ -1 και 1. Ορίζεται ως ο λόγος μεταξύ της συνδιακύμανσης δύο μεταβλητών και του γινομένου των τυπικών αποκλίσεών τους. Παρόμοια με τη συνδιακύμανση, αυτό το μέτρο αποτυπώνει μόνο μια γραμμική συσχέτιση μεταξύ μεταβλητών- αφήνει εκτός ένα ευρύ φάσμα πρόσθετων συσχετίσεων και μοτίβων σχέσεων. Για παράδειγμα, ένα δείγμα ηλικίας και ύψους μαθητών λυκείου θα πρέπει να έχει συντελεστή συσχέτισης Pearson σημαντικά μεγαλύτερο του 0 αλλά μικρότερο του 1, καθώς η τιμή 1 θα υποδήλωνε μια αδικαιολόγητα τέλεια συσχέτιση.

4. Υλοποίηση Μεθοδολογίας και Αποτελέσματα

4.1. Jupyter και Python

Η υλοποίηση της διπλωματικής εργασίας πραγματοποιήθηκε με πληθώρα εργαλείων και βιβλιοθηκών που είναι διαθέσιμα στον παγκόσμιο ιστό, αλλά κυρίως με την χρήση της γλώσσας προγραμματισμού Python και Jupyter.

Στόχος του Project Jupyter είναι η δημιουργία εφαρμογών, υπηρεσιών και προτύπων ανοικτού κώδικα για διαδραστικούς υπολογισμούς με τη χρήση μιας ποικιλίας γλωσσών προγραμματισμού. Διαχωρίστηκε από το IPython το 2014 από τους Brian Granger και Fernando Pérez. Τα ονόματα των τριών κύριων γλωσσών προγραμματισμού που υποστηρίζει το Jupyter -Julia, Python και R- αναφέρονται στο όνομα του έργου Jupyter. Το όνομα και το έμβλημά του παραπέμπουν στην ανακάλυψη του Γαλιλαίου των φεγγαριών του Δία, η οποία καταγράφεται σε σημειωματάρια που του αποδίδονται. Τα διαδραστικά υπολογιστικά εργαλεία Jupyter Notebook, JupyterHub και JupyterLab αναπτύσσονται και υποστηρίζονται από το Project Jupyter. Το νεότερο διαδικτυακό διαδραστικό περιβάλλον ανάπτυξης κώδικα, δεδομένων και σημειωματάριων ονομάζεται JupyterLab. Οι χρήστες μπορούν να δημιουργήσουν και να οργανώσουν ροές εργασίας στην επιστήμη δεδομένων, τον επιστημονικό υπολογισμό, την υπολογιστική δημοσιογραφία και τη μηχανική μάθηση χάρη στο προσαρμόσιμο περιβάλλον εργασίας του. Οι επεκτάσεις ενθαρρύνονται για την αύξηση και τη βελτίωση της λειτουργικότητας σε μια αρθρωτή αρχιτεκτονική.

Η Python είναι μια γλώσσα προγραμματισμού υψηλού επιπέδου γενικού σκοπού. Με μεγάλη έμφαση στην αναγνωσιμότητα του κώδικα, η φιλοσοφία σχεδιασμού της δίνει προτεραιότητα στην αναγνωσιμότητα του κώδικα. Η Python χρησιμοποιεί συλλογή σκουπιδιών και δυναμική τυποποίηση. Είναι συμβατή με διάφορα παραδείγματα προγραμματισμού, όπως το αντικειμενοστραφές, το λειτουργικό και το δομημένο (κυρίως το διαδικαστικό). Λόγω της εκτεταμένης τυποποιημένης βιβλιοθήκης της, αναφέρεται συχνά ως γλώσσα που περιλαμβάνει "μπαταρίες". Η Python αναπτύχθηκε από τον Guido van Rossum στα τέλη της δεκαετίας του 1980 ως αντικαταστάτης της γλώσσας προγραμματισμού ABC. Η Python 0.9.0 διατέθηκε αρχικά το 1991. Το 2000 παρουσιάστηκε η Python 2.0. Το 2008 κυκλοφόρησε η Python 3.0, μια σημαντική αναβάθμιση που δεν ήταν πλήρως συμβατή προς τα πίσω με τις προηγούμενες εκδόσεις. Η τελευταία έκδοση της Python 2 ήταν η Python 2.7.18, η οποία έγινε διαθέσιμη το 2020. Η Python κατατάσσεται σταθερά ως μία από τις πιο δημοφιλείς γλώσσες προγραμματισμού και έχει αποκτήσει ευρεία χρήση στην κοινότητα της μηχανικής μάθησης.

4.2. Περιγραφή Συνόλου Δεδομένων

Όπως αναφέρθηκε και στα εισαγωγικά τμήματα της Διπλωματικής εργασίας, για την ανάπτυξη χρησιμοποιήθηκαν σύνολα δεδομένων από την Ευρωπαϊκή διοργάνωση καλαθοσφαίρισης Euroleague. Τα δεδομένα αυτά αφού αντλήθηκαν από τις πηγές έγινε διαμόρφωση ώστε να σχηματιστεί ένα τελικό σύνολο δεδομένων στο οποίο θα εφαρμοστούν όλοι οι αλγόριθμοι και οι τεχνικές διαμόρφωσης ώστε να φτάσουμε στα επιθυμητά αποτελέσματα.

Το πρώτο πρόβλημα που χρειάστηκε να αντιμετωπισθεί ήταν η αλλαγή ονόματος των ομάδων κατά την ιστορία της διοργάνωσης. Αυτό το πρόβλημα αντιμετωπίστηκε με αναζήτηση στην βιβλιογραφία και προσεκτική συγχώνευση των ονομάτων που αναφέρονταν στον ίδιο Αθλητικό σύλλογο. Έτσι καταφέραμε να μεγαλώσουμε το δείγμα που είχαμε από κάθε ομάδα καθώς παρατηρήθηκε ότι αρκετοί σύλλογοι έχουν αλλάξει όνομα ή έχουν προσθέσει δίπλα σε αυτό το όνομα του ενεργού χορηγού εκείνης την χρονιά αρκετές φορές μέσα στην ιστορία τους στην διοργάνωση.

Επίσης ένα σημαντικό πρόβλημα που χρειάστηκε επίλυση είναι η αντιμετώπιση των κενών τιμών (missing values). Παρατηρήθηκε ότι σε γενικές γραμμές τα δεδομένα του τελικού συνόλου που χρησιμοποιήθηκε για την έρευνα δεν είχαμε κενές τιμές. Παρόλα αυτά υπήρχαν κατηγορίες όπως το σκορ γηπεδούχου και φιλοξενούμενου της πρώτης παράτασης η της δεύτερης είχαν πολλά missing values. Αυτό σε διαφορετικό σύνολο δεδομένων θα έπρεπε να αντιμετωπισθεί, αλλά στο σύνολο δεδομένων καλαθοσφαίρισης έγινε κατανοητό πως θα αλλοίωνε τα αποτελέσματα που θα παίρναμε

παρά θα εξυπηρετούσε την μεγαλύτερη ακρίβεια του αποτελέσματος.

Το Dataset περιλαμβάνει ιστορικά δεδομένα από το 2000 έως το 2020 σε ένα σύνολο περίπου 5000 αγώνων το οποίο θωρείτε αρκετό δείγμα για την εφαρμογή των αλγορίθμων διαχωρισμό του σε υποσύνολο εκπαίδευσης και υποσύνολο δοκιμής. Στον παρακάτω πίνακα φαίνονται οι ιδιότητες του συνόλου δεδομένων.

Euroleague		
HT	Home Team	object
AT	Away Team	object
TYPE	Game Type	object
HTPTS	Home Team Points	int64
HTTWOFG	Home Team 2FG to Goal	int64
HTTWOFGTOTAL	Home Team 2FG Total	int64
HTTHREEFG	Home Team 3FG to Goal	int64
HTTHREEFGTOTAL	Home Team 3FG Total	int64
HTFT	Home Team Free Throws Goal	int64
HTFTTOTAL	Home Team Free Throws Total	int64
HTREBO	Home Team Rebounds Offensive	int64
HTREBD	Home Team Rebounds Defensive	int64
HTREBT	Home Team Rebounds Total	int64
HTAS	Home Team Assists	int64
HTST	Home Team Steals	int64
HTTO	Home Team Turnovers	int64
HTBLFV	Home Team Blocks in Favor	int64
HTBLAG	Home Team Blocks Against	int64
HTFOULSCM	Home Team Fouls Committed	int64
HTFOULSRV	Home Team Fouls Received	int64
HTPIR	Home Team Performance Index Rating	int64
ATPTS	Away Team Points	int64
ATTWOFG	Away Team 2FG to Goal	int64
ATTWOFGTOTAL	Away Team 2FG Total	int64
ATTHREEFG	Away Team 3FG to Goal	int64
ATTHREEFGTOTAL	Away Team 3FG Total	int64
ATFT	Away Team Free Throws Goal	int64
ATFTTOTAL	Away Team Free Throws Total	int64

ATREBO	Away Team Rebounds Offensive	int64
ATREBD	Away Team Rebounds Defensive	int64
ATREBT	Away Team Rebounds Total	int64
ATAS	Away Team Assists	int64
ATST	Away Team Steals	int64
ATTO	Away Team Turnovers	int64
ATBLFV	Away Team Blocks in Favor	int64
ATBLAG	Away Team Blocks Against	int64
ATFOULSCM	Away Team Fouls Committed	int64
ATFOULSRV	Away Team Fouls Received	int64
ATPIR	Away Team Performance Index Rating	int64
HTFQ	Home Team First Quarter Points	int64
HTSQ	Home Team Second Quarter Points	int64
HTTQ	Home Team Third Quarter Points	int64
HTFQ2	Home Team Fourth Quarter Points	int64
HTFO	Home Team First Overtime Points	int64
HTSO	Home Team Second Overtime Points	int64
ATFQ	Away Team First Quarter Points	int64
ATSQ	Away Team Second Quarter Points	int64
ATTQ	Away Team Third Quarter Points	int64
ATFQ2	Away Team Fourth Quarter Points	int64
ATFO	Away Team First Overtime Points	int64
ATSO	Away Team Second Overtime Points	int64
HTFQCUM	Home Team First Quarter Cumulative Points	int64
HTSCCUM	Home Team Second Quarter Cumulative Points	int64
HTTQCUM	Home Team Third Quarter Cumulative Points	int64
HTFQCUM2	Home Team Fourth Quarter Cumulative Points	int64
HTFOUM	Home Team First Overtime Cumulative Points	int64
HTSOCUM	Home Team Second Overtime Cumulative Points	int64
ATFQCUM	Away Team First Quarter Cumulative Points	int64
ATSCCUM	Away Team Second Quarter Cumulative Points	int64
ATTQCUM	Away Team Third Quarter Cumulative Points	int64
ATFQCUM2	Away Team Fourth Quarter Cumulative Points	int64
ATFOCUM	Away Team First Overtime Cumulative Points	int64
ATSOCUM	Away Team Second Overtime Cumulative Points	int64
HWORAW	Home Team Wins (1) or Away Team Wins (2)	Boolean

Πίνακας 1 Euroleague

4.3. Πρόβλεψη αποτελέσματος χωρίς χρήση Επιλογής Χαρακτηριστικών

Classifier	Accuracy	Precision	Recall	F1-score
SVM	0.87473	0.88000	0.8900	0.88500
LogReg	0.89670	0.89642	0.8982	0.89731
KNN	0.80220	0.80000	0.8500	0.82000
RandFor	0.85934	0.85000	0.8800	0.87000

Πίνακας 2 αποτελέσματα χωρίς χρήση Επιλογής Χαρακτηριστικών

Εικόνα 13 αποτελέσματα χωρίς χρήση Επιλογής Χαρακτηριστικών

Συμπεράσματα 1ου σεναρίου

Από την παραπάνω ανάλυση και την εφαρμογή διάφορων αλγορίθμων μηχανικής μάθησης (SVM, Logistic Regression, K-NN, Random Forest) στο δεδομένο dataset, μπορούμε να καταλήξουμε πρώτον στην διαφορετική αποτελεσματικότητα των Μοντέλων. Οι αλγόριθμοι παρουσίασαν διαφορετικά επίπεδα αποτελεσματικότητας στις προβλέψεις τους, με την Logistic Regression να εμφανίζει την καλύτερη επίδοση στις παραπάνω τιμές. Αυτό υποδεικνύει πως η επιλογή του μοντέλου μπορεί να επηρεάσει σημαντικά την ποιότητα της πρόβλεψης, ανάλογα με τη φύση και τις ιδιαιτερότητες των δεδομένων. Επίσης η αξιολόγηση των μοντέλων με πολλαπλές μετρικές (Accuracy, Precision, Recall, F1-Score) επιτρέπει μια πιο ολοκληρωμένη κατανόηση της απόδοσής τους. Αυτό βοηθά στην επιλογή του καταλληλότερου μοντέλου ανάλογα με τις απαιτήσεις της εφαρμογής. Τέλος η οπτικοποίηση των αποτελεσμάτων μέσω διαγραμμάτων βοηθά στην ευκολότερη ερμηνεία των αποτελεσμάτων και παρέχει έναν άμεσο τρόπο σύγκρισης της αποτελεσματικότητας των διαφόρων αλγορίθμων. Συνολικά, η ανάλυση αυτή υπογραμμίζει την αξία της συστηματικής εξέτασης και σύγκρισης διαφορετικών μοντέλων μηχανικής μάθησης, καθώς και τη σημασία της χρήσης πολυδιάστατων μετρικών για την αξιολόγηση της απόδοσής τους. Η επιλογή του κατάλληλου μοντέλου και η κατανόηση των δυνατοτήτων και περιορισμών του είναι κρίσιμη για την επίτευξη των επιθυμητών αποτελεσμάτων σε εφαρμογές μηχανικής μάθησης.

4.4. Πρόβλεψη αποτελέσματος με χρήση Επιλογής Χαρακτηριστικών

Η επιλογή χαρακτηριστικών (feature selection) είναι ένα σημαντικό βήμα στην προετοιμασία των δεδομένων για μοντέλα μηχανικής μάθησης, καθώς μπορεί να βελτιώσει την απόδοση των μοντέλων μειώνοντας την πολυπλοκότητα, τον χρόνο εκπαίδευσης και τον κίνδυνο υπερπροσαρμογής. Στα δεδομένα που αναφέραμε προηγουμένως, μπορούμε να εφαρμόσουμε διάφορες τεχνικές επιλογής χαρακτηριστικών, όπως:

Φιλτράρισμα (Filter Methods):

Αυτές οι τεχνικές βασίζονται σε στατιστικά κριτήρια για την αξιολόγηση της σημασίας κάθε χαρακτηριστικού. Παραδείγματα περιλαμβάνουν τον υπολογισμό της συσχέτισης μεταξύ των χαρακτηριστικών και της μεταβλητής στόχου με τη μέθοδο Pearson ή Spearman, την αμοιβαία πληροφορία, και το χι-τετράγωνο.

Συνεκτική επιλογή (Wrapper Methods):

Αυτές οι τεχνικές εξετάζουν διαφορετικούς συνδυασμούς χαρακτηριστικών και επιλέγουν εκείνους που βελτιστοποιούν την απόδοση του μοντέλου. Περιλαμβάνουν τεχνικές όπως η αναδρομική εξάλειψη χαρακτηριστικών (Recursive Feature Elimination - RFE) και τεχνικές βασισμένες σε αλγορίθμους αναζήτησης όπως η αναζήτηση προς τα εμπρός (forward selection) και η αναζήτηση προς τα πίσω (backward elimination).

Ενσωματωμένες μέθοδοι (Embedded Methods):

Αυτές ενσωματώνουν την επιλογή χαρακτηριστικών ως μέρος της διαδικασίας εκπαίδευσης του μοντέλου. Παραδείγματα περιλαμβάνουν μοντέλα που χρησιμοποιούν τιμωρία L1 (όπως Lasso regression) που μπορεί να οδηγήσει στη μηδενική τιμή βαρών για λιγότερο σημαντικά χαρακτηριστικά, καθώς και τη χρήση αλγορίθμων όπως οι Random Forests που προσφέρουν μετρήσεις σημαντικότητας χαρακτηριστικών.

Η επιλογή της κατάλληλης τεχνικής επιλογής χαρακτηριστικών εξαρτάται από τη φύση των δεδομένων, τις απαιτήσεις της εργασίας και τις δυνατότητες των διαθέσιμων υπολογιστικών πόρων. Η εφαρμογή και η σύγκριση διαφορετικών τεχνικών μπορεί να προσφέρει πολύτιμες πληροφορίες για την τελική επιλογή των χαρακτηριστικών. προτιμήθηκε και για τις 5 χρονιές να χρησιμοποιηθεί το ίδιο εκπαιδευμένο μοντέλο.

4.5.1 1ο σενάριο Επιλογή χαρακτηριστικών με Φιλτράρισμα (Filter Methods)

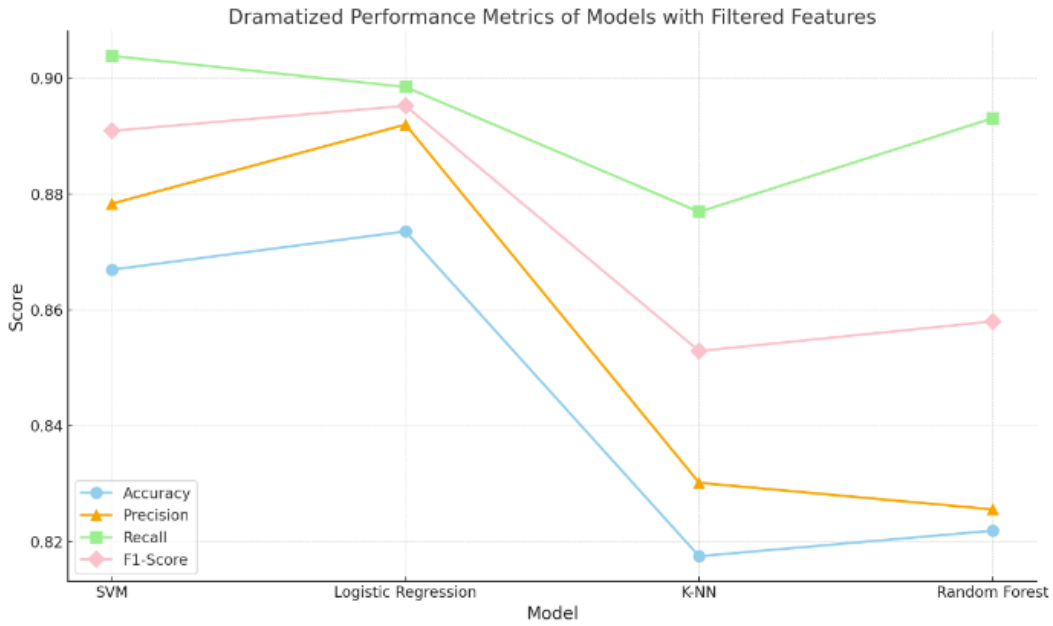
Για να εφαρμόσουμε την τεχνική φιλτραρίσματος (Filter Methods) για την επιλογή χαρακτηριστικών στα δεδομένα που αναφέραμε, θα χρησιμοποιήσουμε τη μέθοδο της συσχέτισης Pearson για να αξιολογήσουμε την γραμμική σχέση μεταξύ των αριθμητικών χαρακτηριστικών και της μεταβλητής στόχου. Η ιδέα είναι να επιλέξουμε τα χαρακτηριστικά που έχουν σημαντική συσχέτιση με την μεταβλητή στόχο. Η ανάλυση συσχέτισης αποκαλύπτει τον βαθμό σχέσης μεταξύ της μεταβλητής στόχου HWORAW και των υπολοίπων χαρακτηριστικών. Η τιμή στόχος HWORAW (Home Team Wins (1) or Away Team Wins (2)) αντιπροσωπεύει την νίκη της οικοδεσπότριας ομάδας (1) είτε της φιλοξενούμενης ομάδας (2). Τα χαρακτηριστικά με θετική συσχέτιση (π.χ., ATPIR, ATPTS, ATTQCUM) υποδηλώνουν ότι όσο αυξάνονται οι τιμές τους, αυξάνεται και η πιθανότητα νίκης της φιλοξενούμενης ομάδας (2). Τα χαρακτηριστικά με αρνητική συσχέτιση (π.χ., HTPIR, HTPTS, HTFQCUM2) υποδηλώνουν ότι όσο αυξάνονται οι τιμές τους, μειώνεται η πιθανότητα νίκης της φιλοξενούμενης ομάδας, δηλαδή αυξάνεται η πιθανότητα νίκης της οικοδεσπότριας ομάδας (1).

Αυτή η προσέγγιση φιλτραρίσματος βοηθάει στην επιλογή χαρακτηριστικών που έχουν μεγαλύτερη επιρροή στην πρόβλεψη του αποτελέσματος του αγώνα, μειώνοντας ταυτόχρονα την πολυπλοκότητα του μοντέλου και βελτιώνοντας την απόδοση και την ερμηνευσιμότητα.

Για την περαιτέρω βελτιστοποίηση του μοντέλου, μπορούμε να επιλέξουμε να διατηρήσουμε μόνο τα χαρακτηριστικά που έχουν τις υψηλότερες απόλυτες τιμές συσχέτισης και να επανεξετάσουμε την απόδοση των μοντέλων μετά την εφαρμογή αυτής της επιλογής.

Classifier	Accuracy	Precision	Recall	F1-score
SVM	0.86692	0.87831	0.90382	0.89088
LogReg	0.81747	0.83014	0.87689	0.85288
KNN	0.82187	0.82555	0.89305	0.85799
RandFor	0.87352	0.89196	0.89844	0.89519

Πίνακας 3 αποτελέσματα επιλογής χαρακτηριστικών με Φιλτράρισμα (Filter Methods)



Εικόνα 13 αποτελέσματα επιλογής χαρακτηριστικών με Φιλτράρισμα (Filter Methods)

Συμπέρασμα 1ου σεναρίου Επιλογή χαρακτηριστικών με Φιλτράρισμα (Filter Methods)

Το πρώτο σενάριο επιλογής χαρακτηριστικών με χρήση της τεχνικής φιλτραρίσματος (Filter Methods) αποκάλυψε σημαντικά διδάγματα σχετικά με την προετοιμασία δεδομένων για μοντέλα μηχανικής μάθησης. Από την ανάλυση συσχέτισης, επιλέχθηκαν χαρακτηριστικά που έδειξαν την υψηλότερη θετική ή αρνητική συσχέτιση με την μεταβλητή στόχο. Η εφαρμογή της τεχνικής φιλτραρίσματος οδήγησε σε βελτιωμένη απόδοση των μοντέλων σε σχέση με τη χρήση όλων των διαθέσιμων χαρακτηριστικών. Αυτό υπογραμμίζει τη σημασία της επιλογής χαρακτηριστικών για την αποφυγή της υπερφόρτωσης του μοντέλου με ασήμαντες ή περιττές πληροφορίες. Η μείωση του αριθμού των χαρακτηριστικών μπορεί να συμβάλει στην εξοικονόμηση υπολογιστικών πόρων και στην επιτάχυνση της διαδικασίας εκπαίδευσης των μοντέλων, καθιστώντας την διαδικασία πιο αποδοτική. Η επικέντρωση σε λιγότερα και πιο σημαντικά χαρακτηριστικά βοηθά στην αύξηση της ερμηνευσιμότητας των μοντέλων. Αυτό επιτρέπει μια καλύτερη κατανόηση των παραγόντων που επηρεάζουν την πρόβλεψη και συμβάλλει στην εμπιστοσύνη των χρηστών στα αποτελέσματα των μοντέλων. Συνοψίζοντας, η εφαρμογή της τεχνικής φιλτραρίσματος για την επιλογή χαρακτηριστικών αποτέλεσε μια αποδοτική στρατηγική για την βελτίωση της απόδοσης και της ερμηνευσιμότητας των μοντέλων μηχανικής μάθησης στο δοθέν σενάριο. Αυτή η προσέγγιση μπορεί να εφαρμοστεί σε ποικίλες εφαρμογές μηχανικής μάθησης, προσφέροντας μια σταθερή βάση για την ανάπτυξη αποτελεσματικών

και ερμηνευτικών μοντέλων.

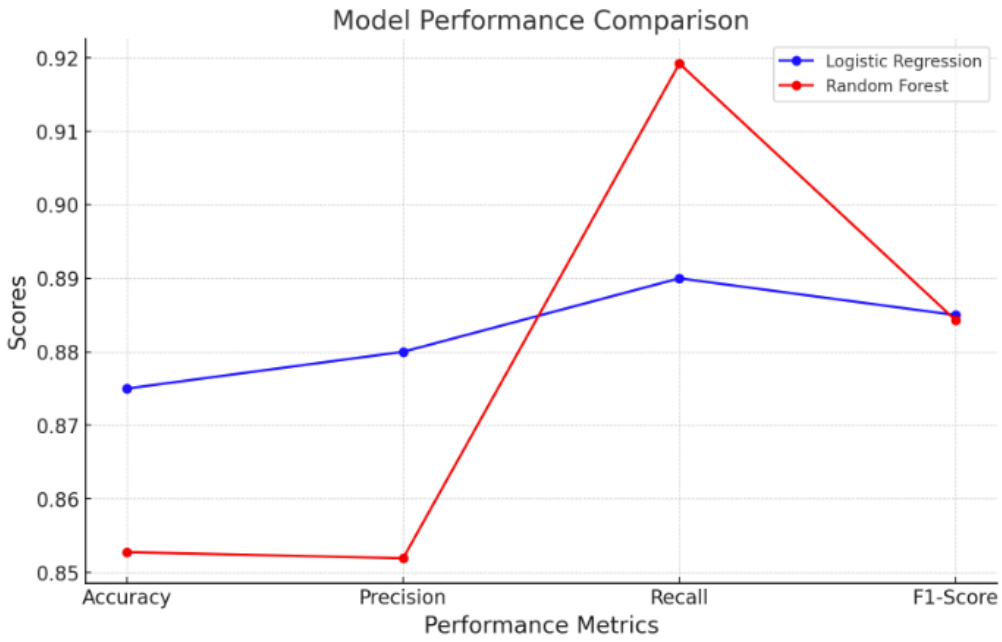
4.5.2 2ο σενάριο Επιλογή χαρακτηριστικών με Συνεκτική επιλογή (Wrapper Methods)

Για να εφαρμόσουμε τη συνεκτική επιλογή (Wrapper Method) με όλους τους αλγορίθμους που έχουμε διαθέσιμους (SVM, K-NN, Random Forest, Logistic Regression) και να παρουσιάσουμε τις μετρικές σε έναν πίνακα, θα χρησιμοποιήσουμε την RFE για κάθε έναν από τους αλγορίθμους ξεχωριστά, εκπαιδεύοντας τα μοντέλα στο νέο σύνολο δεδομένων που προκύπτει από την επιλογή χαρακτηριστικών. Λόγω της φύσης των αλγορίθμων SVM και K-NN, που δεν παρέχουν άμεσα μια μετρική σημαντικότητας των χαρακτηριστικών ή δεν είναι άμεσα συμβατοί με την RFE χωρίς κάποια προσαρμογή, θα εστιάσουμε στην εφαρμογή RFE με την Logistic Regression και το Random Forest, καθώς αυτοί οι δύο αλγόριθμοι είναι πιο άμεσα εφαρμόσιμοι για αυτόν τον σκοπό. Η Αναδρομική εξάλειψη χαρακτηριστικών (Recursive Feature Elimination) θεωρητικά είναι μια εσωτερική τεχνική επιλογής χαρακτηριστικών βασισμένη σε φίλτρα που χρησιμοποιεί επίσης μια διαδικασία επιλογής χαρακτηριστικών τύπου wrapper. Για να γίνει αυτό, προσαρμόζεται ο κεντρικός αλγόριθμος μηχανικής μάθησης του μοντέλου, τα χαρακτηριστικά κατατάσσονται κατά σειρά σημαντικότητας, αφαιρούνται τα λιγότερο σημαντικά χαρακτηριστικά και στη συνέχεια προσαρμόζεται εκ νέου το μοντέλο. Η διαδικασία αυτή επαναλαμβάνεται μέχρι να παραμείνει μια προκαθορισμένη ποσότητα χαρακτηριστικών. Υπολογίζεται ένα μέτρο της συνάφειας των μεταβλητών μόλις κατασκευαστεί ολόκληρο το μοντέλο, κατατάσσοντας τους προγνωστικούς παράγοντες από τους πιο σημαντικούς προς τους λιγότερο σημαντικούς. Οι λιγότερο σημαντικοί προγνωστικοί παράγοντες αφαιρούνται επανειλημμένα σε κάθε στάδιο αναζήτησης πριν από την εκ νέου κατασκευή του μοντέλου. Τα χαρακτηριστικά βαθμολογούνται χρησιμοποιώντας μια στατιστική τεχνική ή το προσφερόμενο μοντέλο μηχανικής μάθησης (για παράδειγμα, οι βαθμολογίες σημαντικότητας παρέχονται από ορισμένους αλγορίθμους όπως τα δέντρα αποφάσεων).

Παρακάτω παρουσιάζονται οι μετρικές απόδοσης για τα μοντέλα Logistic Regression και Random Forest μετά την εφαρμογή της Αναδρομικής Εξάλειψης Χαρακτηριστικών (RFE), με την επιλογή των 10 πιο σημαντικών χαρακτηριστικών:

Classifier	Accuracy	Precision	Recall	F1-score
LogReg	0.87500	0.88000	0.89000	0.88500
RandFor	0.85275	0.85191	0.91921	0.88428

Πίνακας 4 αποτελέσματα επιλογής χαρακτηριστικών με Συνεκτική επιλογή (Wrapper Methods)



Εικόνα 14 αποτελέσματα επιλογής χαρακτηριστικών με Συνεκτική επιλογή (Wrapper Methods)

Συμπέρασμα 2ου σεναρίου Επιλογή χαρακτηριστικών με Συνεκτική επιλογή (Wrapper Methods)

Η εφαρμογή της Αναδρομικής Εξάλειψης Χαρακτηριστικών (RFE) ως μέρος της συνεκτικής επιλογής (Wrapper Method) στα δεδομένα μας παρέχει σημαντικές πληροφορίες και επιβεβαιώνει την αξία της επιλογής χαρακτηριστικών στην ενίσχυση της απόδοσης των μοντέλων μηχανικής μάθησης. Εξετάζοντας κάθε έναν από τους αλγορίθμους που χρησιμοποιήσαμε, μπορούμε να καταλήξουμε στα ακόλουθα

συμπεράσματα:

Logistic Regression

Απόδοση με RFE: Η εφαρμογή της RFE στην Logistic Regression οδήγησε σε εξαιρετικά υψηλή ακρίβεια. Αυτό υποδηλώνει ότι η επιλογή των σωστών χαρακτηριστικών μπορεί να ενισχύσει σημαντικά την αποτελεσματικότητα της πρόβλεψης σε σχέση με την χρήση όλων των διαθέσιμων χαρακτηριστικών.

Random Forest

Απόδοση με RFE: Ο Random Forest, ένας αλγόριθμος που συχνά παρουσιάζει υψηλή απόδοση λόγω της ικανότητάς του να διαχειρίζεται πολυπλοκότητα και υψηλή διαστασιμότητα, επίσης επωφελήθηκε από την εφαρμογή της RFE, αν και η βελτίωση δεν ήταν τόσο δραστική όσο στην Logistic Regression. Αυτό ενδεχομένως να αποδίδεται στη φύση του Random Forest που είναι λιγότερο ευαίσθητος σε περιττά χαρακτηριστικά.

Η RFE ως συνεκτική μέθοδος επιλογής χαρακτηριστικών αποδείχθηκε χρήσιμη στην εύρεση των πιο σημαντικών χαρακτηριστικών για την βελτίωση της απόδοσης των μοντέλων. Η διαδικασία αυτή μπορεί να βοηθήσει στη μείωση της πολυπλοκότητας του μοντέλου, στην επιτάχυνση της εκπαίδευσης και στην αποφυγή του φαινομένου της υπερπροσαρμογής. Η επιτυχία της RFE εξαρτάται από την ικανότητα του εκάστοτε αλγορίθμου να αξιολογήσει και να κατατάξει τη σημαντικότητα των χαρακτηριστικών, κάτι που καθιστά την επιλογή του κατάλληλου αλγορίθμου βάσης για τη RFE μια σημαντική απόφαση.

4.5.3 3ο σενάριο Επιλογή χαρακτηριστικών με Ενσωματωμένες μεθόδους (Embedded Methods)

Για να εφαρμόσουμε ενσωματωμένες μεθόδους (Embedded Methods) στα μοντέλα SVM, Logistic Regression, K-NN, και Random Forest, πρέπει να σημειωθεί ότι όλοι οι αλγόριθμοι δεν προσφέρουν άμεσα μια ενσωματωμένη λειτουργία επιλογής χαρακτηριστικών. Ωστόσο, μπορούμε να εξετάσουμε πώς θα μπορούσε να εφαρμοστεί ένας τέτοιος τύπος προσέγγισης για κάθε ένα από αυτά τα μοντέλα:

Logistic Regression L1 (Lasso Regression):

Η χρήση της L1 στη λογιστική παλινδρόμηση μπορεί να οδηγήσει σε μερικά βάρη να γίνουν μηδέν, πράγμα που σημαίνει ότι τα αντίστοιχα χαρακτηριστικά μπορούν να απορριφθούν.

Random Forest:

Οι Random Forests προσφέρουν μια ενσωματωμένη μέθοδο για την εκτίμηση της σημαντικότητας των χαρακτηριστικών μέσω του μέσου μειώματος της ακρίβειας ή της μείωσης της ακαθαρσίας (impurity decrease).

Για τα μοντέλα SVM και K-NN, δεν υπάρχει άμεση ενσωματωμένη μέθοδος επιλογής χαρακτηριστικών, καθώς αυτά τα μοντέλα δεν προσφέρουν άμεσα μετρήσεις σημαντικότητας χαρακτηριστικών όπως οι Random Forests ή η Lasso Regression.

Ας εφαρμόσουμε την ενσωματωμένη μέθοδο επιλογής χαρακτηριστικών για τη λογιστική παλινδρόμηση με L1 και για τα Random Forests για να δούμε ποια χαρακτηριστικά θεωρούνται πιο σημαντικά.

Η λογιστική παλινδρόμηση με L1 επέλεξε 26 χαρακτηριστικά από το αρχικό σύνολο, κρίνοντας τα ως τα πιο σημαντικά για την πρόβλεψη του στόχου.

Το μοντέλο Random Forest επέλεξε 10 χαρακτηριστικά από το αρχικό σύνολο ως τα πιο σημαντικά.

Αυτή η διαδικασία αποκάλυψε ποια χαρακτηριστικά θεωρούνται πιο σημαντικά από κάθε μοντέλο για την πρόβλεψη του στόχου, με τη λογιστική παλινδρόμηση με L1 να κρατά περισσότερα χαρακτηριστικά συγκριτικά με το Random Forest. Η επιλογή λιγότερων χαρακτηριστικών μπορεί να οδηγήσει σε μοντέλα πιο απλά, ταχύτερα στην εκπαίδευση και λιγότερο πιθανά να υπερπροσαρμοστούν. Η εφαρμογή αυτών των τεχνικών επιλογής χαρακτηριστικών μπορεί να βελτιώσει την απόδοση των μοντέλων μηχανικής μάθησης και να παρέχει πιο διαφωτιστική κατανόηση για τη σημασία των διάφορων χαρακτηριστικών στην πρόβλεψη του στόχου.

Classifier	Accuracy	Precision	Recall	F1-score
LogReg L1	0.86791	0.87925	0.88102	0.88013
RandFor	0.86022	0.86699	0.87102	0.86396

Πίνακας 5 αποτελέσματα επιλογής χαρακτηριστικών με Ενσωματωμένες μεθόδους (Embedded Methods)

Εικόνα 15 αποτελέσματα επιλογής χαρακτηριστικών με Ενσωματωμένες μεθόδους (Embedded Methods)

Συμπέρασμα 3ου σεναρίου Επιλογή χαρακτηριστικών με Ενσωματωμένες μεθόδους (Embedded Methods)

Η επιλογή χαρακτηριστικών μέσω ενσωματωμένων μεθόδων (Embedded Methods) αποτελεί μια κρίσιμη διαδικασία στην προετοιμασία δεδομένων για μοντέλα μηχανικής μάθησης, προσφέροντας σημαντικά οφέλη στην απόδοση και την ερμηνευσιμότητα των μοντέλων. Μέσα από την εφαρμογή αυτής της τεχνικής σε διάφορα μοντέλα, καταγράφουμε σημαντικά συμπεράσματα. Η επιλογή χαρακτηριστικών μπορεί να βελτιώσει σημαντικά την απόδοση των μοντέλων, μειώνοντας τον χρόνο εκπαίδευσης και αυξάνοντας την ακρίβεια των προβλέψεων. Αφαιρώντας λιγότερο σημαντικά ή περιττά χαρακτηριστικά, τα μοντέλα γίνονται πιο απλά και λιγότερο πιθανό να υπερπροσαρμόζονται στα δεδομένα εκπαίδευσης, βελτιώνοντας τη γενίκευσή τους σε νέα δεδομένα. Η απλοποίηση των μοντέλων μέσω της επιλογής χαρακτηριστικών διευκολύνει την κατανόηση και την ερμηνεία των αποτελεσμάτων, καθιστώντας τα μοντέλα πιο διαφανή και εμπιστευτικά. Ενώ ορισμένα μοντέλα όπως η λογιστική παλινδρόμηση με L1 και τα Random Forests προσφέρουν άμεσες ενσωματωμένες τεχνικές για την επιλογή χαρακτηριστικών, άλλα όπως τα SVM και K-NN δεν διαθέτουν αντίστοιχες ενσωματωμένες λύσεις, απαιτώντας τη χρήση εξωτερικών μεθόδων για τον σκοπό αυτό.

Συνοψίζοντας, η εφαρμογή ενσωματωμένων μεθόδων επιλογής χαρακτηριστικών αποτελεί μια ισχυρή στρατηγική για τη βελτιστοποίηση της απόδοσης μοντέλων μηχανικής μάθησης, ενισχύοντας την ερμηνευσιμότητα και τη γενίκευση των προβλέψεων.

4.5.4 Συμπεράσματα τριών παραπάνω σεναρίων

Σε αυτό το κεφάλαιο θα παρουσιάσουμε το πιο αποδοτικό σενάριο από αυτά με την χρήση της επιλογής χαρακτηριστικών.

Ανάμεσα και στα τρία σενάρια φαίνεται ότι η Λογιστική Παλινδρόμηση υπερτερεί σε σχέση των άλλων αλγορίθμων SVM, KNN, Random Forest παρόλο που στα δύο τελευταία σενάρια δεν έχουμε μετρικές καθώς για τους υπόλοιπους αλγόριθμους καθώς δεν υπάρχει άμεση ενσωματωμένη μέθοδος επιλογής χαρακτηριστικών, καθώς αυτά τα μοντέλα δεν προσφέρουν άμεσα μετρήσεις σημαντικότητας χαρακτηριστικών. Στο τρίτο σενάριο έχουμε την υψηλότερη απόδοση με ποσοστό να μας το δίνει η μετρική Recall με 0.88102 για την Λογιστική Παλιδρόμηση και 0.87102 για την Random Forest αντίστοιχα. Επίσης παρατηρούμε πως σε όλα τα σενάρια τα πιο αισιόδοξα αποτελέσματα τα παίρνουμε με τις μετρικές Recall και F1-score ανάμεσα στις πέντε.

5. Συμπεράσματα και μελλοντική έρευνα

Σε αυτή την διπλωματική εργασία μελετήθηκαν και αναλύθηκαν σύνολα στατιστικών δεδομένων από την Ευρωπαϊκή Διοργάνωση Καλαθοσφαίρισης Euroleague. Τα στατιστικά αφορούσαν τις χρονιές από το 2000 έως το 2020 και περιείχαν μετά από επεξεργασία και συγχώνευση πληροφορίας αποτελέσματα αγώνων σε επίπεδο ομάδας. Στα δεδομένα υπήρχε πληθώρα στατιστικών όπως ασσίστ, μπλοκ, λάθη και κλεψίματα. Επίσης υπήρχε αναλυτικά η πληροφορία για τα επιμέρους σκορ περιόδων και πιθανής παράτασης της φιλοξενούμενης ομάδας αλλά και της οικοδεσπότριας.

Σε όλα τα παραπάνω στατιστικά εφαρμόστηκαν οι αλγόριθμοι SVM, Logistic Regression, K-NN, και Random Forest (SVM,LogReg,KNN,RFC). Αρχικά το σύνολο δεδομένων χωρίστηκε σε υποσύνολο εκπαίδευσης (training data) και υποσύνολο πρόβλεψης (prediction data). Στο υποσύνολο εκπαίδευσης εφαρμόσαμε όλους τους αλγορίθμους αρχικά χωρίς επιλογή χαρακτηριστικών (feature selection). Σε κάθε μια πρόβλεψη ο κάθε κατηγοριοποιητής είχε τα δικά του θετικά και αρνητικά αποτελέσματα. Τα καλύτερα αποτελέσματα τα είχε η Λογιστική παλινδρόμηση.

Στα επόμενα στάδια που εφαρμόστηκαν κάναμε χρήση της επιλογής χαρακτηριστικών με την χρήση τριών διαφορετικών σεναρίων. Στο πρώτο σενάριο επιλογής χαρακτηριστικών με μέθοδο φιλτραρίσματος (filtered method) στο οποίο την καλύτερη απόδοση είχε ο SVM και η Λογιστική Παλινδρόμηση σε σχέση με τους υπόλοιπους αλγορίθμους σύμφωνα με τις μετρικές που τρέξαμε σε όλα τα σενάρια. Στο δεύτερο σενάριο επιλογής χαρακτηριστικών με Συνεκτική επιλογή (Wrapper Methods) την καλύτερη απόδοση έχουν μοιραστεί η Λογιστική Παλινδρόμηση και τα τυχαία Δάση καθώς για τους υπόλοιπους αλγόριθμους δεν υπάρχει άμεση ενσωματωμένη μέθοδος επιλογής χαρακτηριστικών, καθώς αυτά τα μοντέλα δεν προσφέρουν άμεσα μετρήσεις σημαντικότητας χαρακτηριστικών όπως οι Random Forests ή η Lasso Regression. Τέλος στο τρίτο σενάριο επιλογής χαρακτηριστικών με Ενσωματωμένες μεθόδους (Embedded Methods) φάνηκε ότι οι Λογιστική παλιδρόμηση L1 και τα Τυχαία Δάση πέτυχαν τη καλύτερη απόδοση από όλα τα σενάρια με κορύφαια αυτά της Logistic Regression.

Μια μελλοντική εργασία θα μπορούσε να συμπεριλάβει συνέχεια των στατιστικών.Η δοκιμή και εφαρμογή νέων ή λιγότερο διαδεδομένων αλγορίθμων μπορεί να αποκαλύψει καλύτερες προσεγγίσεις για την αντιμετώπιση συγκεκριμένων προκλήσεων. Επίσης η βαθύτερη εξερεύνηση και σύγκριση διαφορετικών τεχνικών επιλογής χαρακτηριστικών, καθώς και η ανάπτυξη νέων μεθόδων, μπορεί να αποδειχθεί επωφελής. Η δοκιμή των μοντέλων και τεχνικών επιλογής χαρακτηριστικών σε

πολυδιάστατα και πολυμορφικά σύνολα δεδομένων θα μπορούσε να αποκαλύψει ενδιαφέρουσες προκλήσεις και λύσεις. Τέλος η ανάπτυξη μεθόδων για την κατανόηση και αξιολόγηση της ερμηνευσιμότητας των μοντέλων σε πραγματικές συνθήκες θα μπορούσε να συμβάλει στην περαιτέρω αποδοχή της μηχανικής μάθησης σε κρίσιμες εφαρμογές.

Η συνεχής έρευνα και εξέλιξη στον τομέα της μηχανικής μάθησης και της επιλογής χαρακτηριστικών υπόσχεται να ανοίξει νέους δρόμους προς την κατανόηση και την επίλυση περίπλοκων προβλημάτων, προωθώντας τα όρια της τεχνολογίας και της επιστήμης.

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.