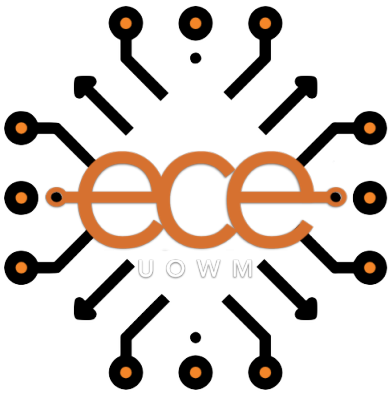


Πανεπιστήμιο Δυτικής Μακεδονίας

Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

Αλγόριθμοι μηχανικής μάθησης για sports analytics

Μάριος Κράλλης (ΑΜ: 1345)

Επιβλέπων Καθηγητής: Νικόλαος Πλόσκας

Εργαστήριο Ευφυών Συστημάτων & Βελτιστοποίησης

23 Σεπτεμβρίου 2024

Περίληψη

Η μελέτη διερευνά τη χρήση αλγορίθμων μηχανικής μάθησης για την πρόβλεψη της αποτελεσματικότητας των μπασκετμπολίστων στην EuroLeague, με ειδική εστίαση σε εφαρμογές για φανταστική επιλογή ομάδων της EuroLeague. Αναλύοντας εκτεταμένα δεδομένα παικτών από πολλές χρονιές, χρησιμοποιήθηκαν διάφορα μοντέλα μηχανικής μάθησης, συμπεριλαμβανομένης της γραμμικής παλινδρόμησης, των δέντρων αποφάσεων και των νευρωνικών δικτύων, για την πρόβλεψη της βαθμολογία απόδοσης παίκτη (PER). Η μελέτη μας δείχνει ότι αυτά τα μοντέλα μπορούν να βελτιώσουν σημαντικά την ακρίβεια των προβλέψεων απόδοσης των παικτών σε σύγκριση με τις παραδοσιακές μεθόδους. Οι πληροφορίες που αποκτήθηκαν από αυτές τις προβλέψεις μπορούν να εφαρμοστούν απευθείας για τη βελτιστοποίηση της επιλογής παικτών στο διαδικτυακό παιχνίδι Fantasy EuroLeague, δίνοντας τη δυνατότητα στους χρήστες να δημιουργήσουν πιο ανταγωνιστικές ομάδες που θα αποφέρουν περισσότερους πόντους. Η μελέτη τονίζει επίσης τη σημασία της ρύθμισης των μοντέλων και της επιλογής χαρακτηριστικών στη βελτίωση της αξιοπιστίας της πρόβλεψης. Συνολικά, αυτή η εργασία παρέχει ένα πολύτιμο εργαλείο τόσο για τους αναλυτές όσο και για τους λάτρεις των παιχνιδιών Fantasy, γεφυρώνοντας το χάσμα μεταξύ προηγμένων sports analytics και πρακτικών εφαρμογών στη διαχείριση παιχνιδιών Fantasy.

Λέξεις κλειδιά: Μηχανική Μάθηση, Sports Analytics, Fantasy EuroLeague, Ρύθμιση Μοντέλων, Επιλογή Χαρακτηριστικών, Βελτιστοποίηση

Abstract

The study explores the use of machine learning algorithms to predict the effectiveness of EuroLeague basketball players, with a special focus on applications for fantasy EuroLeague team selection. Analyzing extensive player data from many years, various machine learning models, including linear regression, decision trees and neural networks, were used to predict Player Performance Rating (PER). Our study shows that these models can significantly improve the accuracy of player performance predictions compared to traditional methods. The information gained from these predictions can be applied directly to optimize player selection in the online Fantasy EuroLeague game, enabling users to create more competitive teams that will yield more points. The study also highlights the importance of model tuning and feature selection in improving prediction reliability. Overall, this work provides a valuable tool for both analysts and fans of Fantasy games, bridging the gap between advanced sports analytics and practical applications in Fantasy game management.

Keywords: Machine Learning, Sports Analytics, Fantasy EuroLeague, Model Tuning, Feature Selection, Optimization

Δήλωση Πνευματικών Δικαιωμάτων

Δήλωση Πνευματικών Δικαιωμάτων Δηλώνω ρητά ότι, σύμφωνα με το άρθρο 8 του Ν. 1599/1986 και τα άρθρα 2,4,6 παρ. 3 του Ν. 1256/1982, η παρούσα Διπλωματική Εργασία με τίτλο ”Αλγόριθμοι μηχανικής μάθησης για sports analytics” καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας και αναφέρονται ρητώς μέσα στο κείμενο που συνοδεύουν, και η οποία έχει εκπονηθεί στο Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών του Πανεπιστημίου Δυτικής Μακεδονίας, υπό την επίβλεψη του μέλους του Τμήματος κ. Νικόλαου Πλόσκα αποτελεί αποκλειστικά προϊόν προσωπικής εργασίας και δεν προσβάλλει κάθε μορφής πνευματικά δικαιώματα τρίτων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία 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.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα. Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και μόνο.

Υπογραφή Φοιτητή

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Ορισμός του προβλήματος

Η EuroLeague είναι η κορυφαία διοργάνωση μπάσκετ της Ευρώπης, που συχνά θεωρείται ως η ισοδύναμη της ηπείρου με το NBA όσον αφορά το κύρος. Με κορυφαίες ομάδες από διάφορες ευρωπαϊκές χώρες, είναι γνωστό για τον έντονο ανταγωνισμό, το στρατηγικό παιχνίδι και τους παθιασμένους θαυμαστές του. Ένας τρόπος για τους οπαδούς του μπάσκετ να ασχοληθούν με τη σεζόν της EuroLeague σε βαθύτερο επίπεδο είναι η συμμετοχή τους στο διαδικτυακό παιχνίδι Fantasy EuroLeague. Σε αυτό το παιχνίδ, οι συμμετέχοντες δημιουργούν εικονικές ομάδες που αποτελούνται από πραγματικούς παίκτες της EuroLeague, κερδίζοντας πόντους με βάση τις επιδόσεις τους σε πραγματικούς αγώνες. Η πρόκληση έγκειται στην επιλογή μιας ισορροπημένης ομάδας εντός ενός προϋπολογισμού, λαμβάνοντας υπόψη παράγοντες όπως η φόρμα του παίκτη, οι αγώνες και οι τραυματισμοί.

Η μηχανική μάθηση έχει γίνει ένα μετασχηματιστικό εργαλείο στα sports analytics, φέρνοντας επανάσταση στον τρόπο με τον οποίο οι ομάδες, οι προπονητές και οι αναλυτές προσεγγίζουν τη στρατηγική, την απόδοση των παικτών και τα αποτελέσματα των αγώνων. Με την αξιοποίηση τεράστιων ποσοτήτων δεδομένων από στατιστικά στοιχεία παικτών και βιομετρικές μετρήσεις έως γεγονότα εντός του παιχνιδιού και ιστορικές επιδόσεις, τα μοντέλα μηχανικής μάθησης μπορούν να εντοπίσουν μοτίβα και συσχετισμούς που μπορεί να μην είναι άμεσα εμφανείς στους αναλυτές. Στον αθλητισμό, η μηχανική μάθηση χρησιμοποιείται για την πρόβλεψη της απόδοσης των παικτών, την αξιολόγηση των κινδύνων τραυματισμών, τη βελτιστοποίηση των στρατηγικών της ομάδας και ακόμη και για την αναζήτηση ταλέντων. Για πα-

ράδειγμα, τα μοντέλα πρόβλεψης μπορούν να προβλέψουν τη μελλοντική απόδοση ενός παίκτη με βάση τα ιστορικά του δεδομένα, βοηθώντας τις ομάδες να λαμβάνουν τεκμηριωμένες αποφάσεις για αποκτήσεις παικτών ή αλλαγές στη σύνθεση.

Στα παιχνίδια Fantasy, η μηχανική μάθηση δίνει τη δυνατότητα πρόβλεψης των παικτών που είναι πιθανό να έχουν καλή απόδοση στα επερχόμενα παιχνίδια, βοηθώντας τους συμμετέχοντες να λαμβάνουν καλύτερες αποφάσεις όταν επιλέγουν τις ομάδες τους. Συνολικά, η μηχανική μάθηση ανεβάζει τα sports analytics σε νέα επίπεδα, προσφέροντας βαθύτερες πληροφορίες και πιο ακριβείς προβλέψεις από ποτέ.

1.2 Κίνητρα και στόχοι υλοποίησης

Σκοπός αυτής της μελέτης είναι να αναπτύξει ένα μοντέλο πρόβλεψης που προβλέπει με ακρίβεια την αποτελεσματικότητα των παικτών μπάσκετ στην EuroLeague, με στόχο τη βελτιστοποίηση της επιλογής παικτών για το διαδικτυακό παιχνίδι Fantasy EuroLeague. Αναλύοντας ιστορικά δεδομένα απόδοσης και στατιστικά στοιχεία παικτών, η μελέτη στοχεύει στον εντοπισμό βασικών δεικτών της απόδοσης των παικτών, όπως πόντοι, ασίστ, ριμπάουντ και άλλες σχετικές μετρικές. Οι γνώσεις που αποκτήθηκαν από αυτό το μοντέλο πρόβλεψης όχι μόνο θα ενισχύσουν την ακρίβεια των προβλέψεων απόδοσης των παικτών, αλλά θα παρέχουν επίσης στρατηγικές συστάσεις για την επιλογή παικτών στο Fantasy EuroLeague, βελτιώνοντας έτσι το συνολικό ποσοστό επιτυχίας των ομάδων. Η μελέτη επιδιώκει να γεφυρώσει το χάσμα μεταξύ της στατιστικής ανάλυσης και της στρατηγικής των παιχνιδιών Fantasy, προσφέροντας πολύτιμα εργαλεία τόσο για τους ερευνητές όσο και για τους λάτρεις των παιχνιδιών Fantasy.

1.3 Διάρθρωση κειμένου

Τα υπόλοιπα κεφάλαια οργανώνονται ως εξής: Το Κεφάλαιο 2 παρουσιάζει το θεωρητικό υπόβαθρο και τους αλγορίθμους που χρησιμοποιήθηκαν, το Κεφάλαιο 3 παρουσιάζει τη σχετική βιβλιογραφική ανασκόπηση, το Κεφάλαιο 4 παρουσιάζει την υλοποίηση των αλγορίθμων της βιβλιογραφικής ανασκόπησης, το Κεφάλαιο 5 παρουσιάζει την πειραματική διαδικασία και τα αποτελέσματα, ενώ τα συμπερά-

σματα παρουσιάζονται στο Κεφάλαιο 6.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

Μηχανική Μάθηση

Η μηχανική μάθηση είναι ένα υποσύνολο της τεχνητής νοημοσύνης που περιλαμβάνει την ανάπτυξη αλγορίθμων και στατιστικών μοντέλων που επιτρέπουν στους υπολογιστές να εκτελούν εργασίες χωρίς ρητές οδηγίες. Αντίθετα, αυτά τα συστήματα εκτελούν προβλέψεις ή λαμβάνουν αποφάσεις καθώς και μαθαίνουν από αυτές με βάση τα δεδομένα.

Μορφές Μηχανικής Μάθησης

1. **Εποπτευόμενη μάθηση**: Το μοντέλο εκπαιδεύεται σε ένα επισημασμένο σύνολο δεδομένων, που σημαίνει ότι κάθε παράδειγμα εκπαίδευσης συνδυάζεται με μια ετικέτα εξόδου. Ο στόχος είναι να μάθουμε μια αντιστοίχιση από τις εισόδους στις εξόδους. Τα παραδείγματα περιλαμβάνουν εργασίες ταξινόμησης και παλινδρόμησης.

2. **Εκμάθηση χωρίς επίβλεψη**: Το μοντέλο εκπαιδεύεται σε δεδομένα χωρίς αποκρίσεις με ετικέτα. Ο στόχος είναι να συμπεράνουμε τη φυσική δομή που υπάρχει μέσα σε ένα σύνολο σημείων δεδομένων. Τα παραδείγματα περιλαμβάνουν εργασίες ομαδοποίησης και συσχέτισης.

3. **Ημι-εποπτευόμενη μάθηση**: Ένας συνδυασμός εποπτευόμενης και μη εποπτευόμενης μάθησης, αυτή η προσέγγιση χρησιμοποιεί μια μικρή ποσότητα δεδομένων με ετικέτα και μια μεγάλη ποσότητα δεδομένων χωρίς ετικέτα κατά τη διάρκεια της εκπαίδευσης.

4. **Ενισχυτική μάθηση**: Το μοντέλο μαθαίνει αλληλεπιδρώντας με ένα περιβάλλον και λαμβάνοντας αναπληροφόρηση με τη μορφή ανταμοιβών ή ποινών. Ο

στόχος είναι να μάθετε μια στρατηγική που μεγιστοποιεί τη σωρευτική ανταμοιβή.

Αλγόριθμοι που χρησιμοποιήθηκαν:

- **Γραμμική παλινδρόμηση**:

  Η γραμμική παλινδρόμηση περιλαμβάνει τη χρήση μιας γραμμικής προσέγγισης για τη μοντελοποίηση της σχέσης μεταξύ μιας εξαρτημένης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών. Ο πρωταρχικός στόχος είναι να δημιουργηθεί μια γραμμική εξίσωση που να προβλέπει καλύτερα την εξαρτημένη μεταβλητή με βάση τις δεδομένες ανεξάρτητες μεταβλητές. Το μοντέλο γραμμικής παλινδρόμησης μπορεί να εκφραστεί ως εξής:

  $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

  όπου:

  Το Y είναι η εξαρτημένη μεταβλητή

  β_0 είναι η αναχαίτιση

  $\beta_1, \beta_2, ..., \beta_n$ είναι οι συντελεστές που αντιστοιχούν σε κάθε ανεξάρτητη μεταβλητή

  $X_1, X_2, ..., X_n$ είναι οι ανεξάρτητες μεταβλητές

  ϵ είναι ο όρος σφάλματος, που αντιπροσωπεύει την απόκλιση των πραγματικών τιμών από τις προβλεπόμενες τιμές.

- **Διαβαθμιζόμενη ενίσχυση (Gradient Boosting)**

  Η ενίσχυση βαθμίδας είναι μια τεχνική μηχανικής μάθησης που χρησιμοποιείται για εργασίες παλινδρόμησης και ταξινόμησης. Δημιουργεί ένα σύνολο αδύναμων μαθητών, συνήθως δέντρα, με διαδοχικό τρόπο. Κάθε νέο δέντρο εκπαιδεύεται να διορθώνει τα λάθη που έγιναν από τα προηγούμενα δέντρα, με

αποτέλεσμα ένα ισχυρό μοντέλο πρόβλεψης.

- **Γραμμική Μηχανή Διανυσμάτων Στήριξης SVM**

  Μια μηχανή διανυσμάτων στήριξης είναι ένας εποπτευόμενος αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται για εργασίες ταξινόμησης, παλινδρόμησης και ανίχνευσης ακραίων στοιχείων. Έχει ως στόχο να βρει το βέλτιστο υπερεπίπεδο που διαχωρίζει καλύτερα τις κλάσεις στον χώρο χαρακτηριστικών. Η γραμμική μηχανή διανυσμάτων στήριξης βρίσκει ένα γραμμικό υπερεπίπεδο που διαχωρίζει καλύτερα τις κλάσεις στο χώρο χαρακτηριστικών. Ο στόχος είναι να βρεθεί το υπερεπίπεδο που μεγιστοποιεί το περιθώριο μεταξύ των κλάσεων. Μαθηματικά, αυτό μπορεί να αναπαρασταθεί ως:

  $$f(x) = w * x + b$$

  όπου:

  w είναι το διάνυσμα βάρους κάθετο στο υπερεπίπεδο

  x είναι το διάνυσμα χαρακτηριστικών εισόδου

  b είναι ο όρος μεροληψίας

- **Τυχαίο Δάσος**

  Το τυχαίο δάσος είναι μια μέθοδος εκμάθησης συνόλου που χρησιμοποιείται τόσο για εργασίες ταξινόμησης όσο και για εργασίες παλινδρόμησης. Δημιουργεί πολλαπλά δέντρα απόφασης κατά τη διάρκεια της εκπαίδευσης και εξάγει τη λειτουργία των κλάσεων για ταξινόμηση ή τη μέση πρόβλεψη για παλινδρόμηση. Αυτή η προσέγγιση βοηθά στη βελτίωση της ακρίβειας του μοντέλου και στον έλεγχο της υπερπροσαρμογής. Κατά την εκπαίδευση δημιουργεί πολλαπλά σύνολα δεδομένων από τα αρχικά δεδομένα εκπαίδευσης, έπειτα εκπαιδεύει ένα δέντρο αποφάσεων σε κάθε σύνολο δεδομένων. Κατά τη διάρκεια

της εκπαίδευσης κάθε δέντρου, επιλέγεται ένα τυχαίο υποσύνολο χαρακτηριστικών σε κάθε διαχωρισμό. Για προβλήματα παλινδρόμησης, όπως είναι και το δικό μας, κάθε δέντρο προβλέπει μια τιμή. Η τελική πρόβλεψη είναι ο μέσος όρος όλων των προβλέψεων από τα δέντρα.

- **Νευρωνικά Δίκτυα (Multilayer Perceptron)**

  Ένα νευρωνικό δίκτυο είναι ένα υπολογιστικό μοντέλο εμπνευσμένο από τον ανθρώπινο εγκέφαλο, που αποτελείται από διασυνδεδεμένα στρώματα κόμβων (νευρώνες) που μπορούν να μάθουν να εκτελούν εργασίες προσαρμόζοντας τις συνδέσεις (βάρη) μεταξύ κόμβων με βάση δεδομένα. Πιο συγκεκριμένα, ο Multilayer Perceptron (MLP), τον οποίο και χρησιμοποιούμε, είναι μια κατηγορία τεχνητών νευρωνικών δικτύων τροφοδοσίας που αποτελείται από τουλάχιστον τρία στρώματα κόμβων: ένα στρώμα εισόδου, ένα ή περισσότερα κρυφά στρώματα και ένα στρώμα εξόδου. Κάθε κόμβος (νευρώνας) σε ένα στρώμα συνδέεται με κάθε κόμβο στο επόμενο στρώμα, καθιστώντας τα MLP πλήρως συνδεδεμένα δίκτυα. Ένα νευρωνικό δίκτυο λειτουργεί ως εξής:

  1. Στρώμα εισόδου: Λαμβάνει τις δυνατότητες εισαγωγής για είσοδο X:

  $$X = [x_1, x_2, ..., x_n]$$

  2. Κρυφά στρώματα: Κάθε νευρώνας κρυφού στρώματος υπολογίζει:

  $$z_j = \sum_{i=1}^{n} w_{ij} x_i + b_j$$

  $$a_j = f(z_j)$$

  όπου:

  z_j είναι το σταθμισμένο άθροισμα

  w_{ij} τα βάρη

  b_j είναι η "προκατάληψη"

f είναι η συνάρτηση ενεργοποίησης

3. Στρώμα εξόδου: Οι νευρώνες του τελικού στρώματος υπολογίζουν:

$$z_k = \sum_{j=1}^{n} w_{jk}\alpha_j + b_k$$

$$\hat{y} = g(z_k)$$

όπου g είναι η συνάρτηση ενεργοποίησης για το στρώμα εξόδου.

- **Χειριστής Ελάχιστης Απόλυτης Συρρίκνωσης και Επιλογής (Lasso)** Ο αλγόριθμος Lasso είναι ένας τύπος μεθόδου ανάλυσης παλινδρόμησης που εκτελεί τόσο επιλογή μεταβλητών όσο και κανονικοποίηση για να βελτιώσει την ακρίβεια πρόβλεψης και την ερμηνευτικότητα του στατιστικού μοντέλου που παράγει. Είναι ιδιαίτερα χρήσιμος όταν ασχολούμαστε με δεδομένα υψηλών διαστάσεων, όπου ο αριθμός των χαρακτηριστικών είναι μεγάλος σε σύγκριση με τον αριθμό των παρατηρήσεων. Ο Lasso βασίζεται στις αρχές της γραμμικής παλινδρόμησης, όπου η σχέση μεταξύ της εξαρτημένης μεταβλητής y και οι ανεξάρτητες μεταβλητές X μοντελοποιούνται ως:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$$

όπου, β_p είναι οι συντελεστές και ϵ είναι ο όρος σφάλματος.

Ο Lasso προσθέτει μια ποινή στη συνάρτηση απώλειας γραμμικής παλινδρόμησης για να αποτρέψει την υπερπροσαρμογή και να χειριστεί την πολυσυγγραμμικότητα συρρικνώνοντας ορισμένους από τους συντελεστές στο μηδέν, επιλέγοντας έτσι ένα απλούστερο μοντέλο. Ως ποινή στον Lasso ορίζεται το

άθροισμα των απόλυτων τιμών των συντελεστών (L1 norm):

$$LassoPenalty = \lambda \sum_{j=1}^{p} |\beta_j|$$

Η αντικειμενική συνάρτηση στον Lasso συνδυάζει το σφάλμα ελαχίστων τετραγώνων με την ποινή L1:

$$\overset{minimize}{\beta} \{\frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|\}$$

όπου:

n είναι ο αριθμός των παρατηρήσεων

p είναι ο αριθμός των χαρακτηριστικών

λ είναι η παράμετρος κανονικοποίησης

- **Παλινδρόμηση Κορυφογραμμής (Ridge)**

  Η παλινδρόμηση κορυφογραμμής είναι μια τεχνική που χρησιμοποιείται στη μηχανική μάθηση για την ανάλυση πολλαπλών δεδομένων παλινδρόμησης που πάσχουν από πολυσυγγραμμικότητα. Όταν εμφανίζεται πολυσυγγραμμικότητα, οι εκτιμήσεις ελαχίστων τετραγώνων είναι αμερόληπτες, αλλά οι διακυμάνσεις τους είναι μεγάλες, γεγονός που οδηγεί σε μεγάλο μέσο τετράγωνο σφάλμα. Η παλινδρόμηση Ridge αντιμετωπίζει αυτό το ζήτημα προσθέτοντας έναν βαθμό "προκατάληψης" στις εκτιμήσεις παλινδρόμησης, γεγονός που μειώνει τα τυπικά σφάλματα. Η παλινδρόμηση κορυφογραμμής στοχεύει στην ελαχιστοποίηση του υπολειπόμενου αθροίσματος τετραγώνων μεταξύ των παρατηρούμενων αποκρίσεων και των αποκρίσεων που προβλέπονται από τη γραμμική προσέγγιση.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$$

Η παλινδρόμηση κορυφογραμμής προσθέτει μια ποινή στο μέγεθος των συντελεστών. Ο στόχος είναι να συρρικνωθούν οι συντελεστές και, στη συνέχεια, να μειωθεί η διακύμανσή τους. Αυτό γίνεται με την προσθήκη ενός όρου τακτοποίησης στη συνάρτηση απώλειας. Ως ποινή στην παλινδρόμηση κορυφογραμμής είναι το άθροισμα των τετραγώνων των συντελεστών (L2 norm):

$$RidgePenalty = \lambda \sum_{j=1}^{p} \beta_j^2$$

όπου λ είναι η παράμετρος κανονικοποίηση που ελέγχει την ένταση της ποινής.

Η αντικειμενική συνάρτηση στην παλινδρόμηση κορυφογραμμής συνδυάζει το σφάλμα ελαχίστων τετραγώνων με την ποινή L2:

$$\overset{minimize}{\beta} \{\frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2\}$$

όπου: n είναι ο αριθμός των παρατηρήσεων

p είναι ο αριθμός των χαρακτηριστικών

λ είναι η παράμετρος κανονικοποίησης

- **Δέντρα Αποφάσεων (Decision Trees)**

  Τα δέντρα αποφάσεων είναι ένας αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται τόσο για εργασίες ταξινόμησης όσο και για εργασίες παλινδρόμησης. Λειτουργούν με αναδρομικό διαχωρισμό των δεδομένων σε υποσύνολα με βάση την τιμή των χαρακτηριστικών εισόδου, δημιουργώντας ένα μοντέλο αποφάσεων που μοιάζει με δέντρο. Ένα δέντρο αποτελείται από κόμβους που αντιπροσωπεύουν ένα χαρακτηριστικό και ακμές, βέλη που συνδέουν κόμβους, που αντιπροσωπεύουν τους κανόνες απόφασης. Οι κόμβοι χωρίζονται σε:

  – Κόμβος ρίζας: Ο κορυφαίος κόμβος σε ένα δέντρο αποφάσεων, που αντιπροσωπεύει το καλύτερο χαρακτηριστικό για τον διαχωρισμό των δεδομένων

– Εσωτερικοί κόμβοι: αντιπροσωπεύουν χαρακτηριστικά που χρησιμοποιούνται για τον διαχωρισμό των δεδομένων σε κάθε βήμα

– Κόμβοι φύλλων: τερματικοί κόμβοι που αντιπροσωπεύουν την τελική έξοδο ή απόφαση

Κατά την κατασκευή ενός δέντρου αποφάσεων υπολογίζεται το μέτρο ακαθαρσίας για πιθανές διασπάσεις και έπειτα επιλέγεται το χαρακτηριστικό και το διαχωρισμό που μειώνει καλύτερα την ακαθαρσία. Γίνεται ένας διαχωρισμός του σύνολου δεδομένων σε υποσύνολα με βάση την τιμή του επιλεγμένου χαρακτηριστικού. Αυτή η διαδικασία εφαρμόζεται αναδρομικά σε κάθε υποσύνολο, δημιουργώντας υπο-κόμβους εώς ότου πληρείται ένα κριτήριο διακοπής ή όλα τα σημεία δεδομένων σε έναν κόμβο ανήκουν στην ίδια κλάση.

Διασταυρούμενη Επικύρωση Κ-πλαισίων (K-Fold Cross Validation)

Η διασταυρούμενη επικύρωση Κ-πλαισίων είναι μια στατιστική μέθοδος που χρησιμοποιείται στη μηχανική μάθηση για την αξιολόγηση της απόδοσης ενός μοντέλου. Περιλαμβάνει τη διαίρεση του συνόλου δεδομένων σε κ ισομεγέθη πλαίσια (ή υποσύνολα) και στη συνέχεια τη χρήση κάθε πλαισίου ως σύνολο επικύρωσης ενώ τα υπόλοιπα κ–1 πλαίσια χρησιμοποιούνται ως το σύνολο εκπαίδευσης. Αυτή η διαδικασία επαναλαμβάνεται κ φορές, με κάθε πλαίσιο να χρησιμοποιείται ακριβώς μία φορά ως σύνολο επικύρωσης. Το μέτρο απόδοσης που αναφέρεται από τη διασταυρούμενη επικύρωση είναι τότε ο μέσος όρος των τιμών που υπολογίζονται στον βρόχο.

Το πρόβλημα της επιλογής των παικτών στο Fantasy Euroleague αποτελεί πρόβλημα βελτιστοποίησης με περιορισμούς.

Αλγόριθμοι Βελτιστοποίησης

Η βελτιστοποίηση στη μηχανική μάθηση αναφέρεται στη διαδικασία προσαρμογής των παραμέτρων ενός μοντέλου για την ελαχιστοποίηση (ή τη μεγιστοποίηση) μιας συγκεκριμένης αντικειμενικής συνάρτησης. Αυτή η αντικειμενική συνάρτηση, που συχνά ονομάζεται συνάρτηση απώλειας ή κόστους, μετρά την απόκλιση μεταξύ των προβλεπόμενων εκροών του μοντέλου και των πραγματικών εκροών. Ο στόχος της βελτιστοποίησης είναι να βρεθεί το σύνολο των παραμέτρων που έχει ως αποτέλεσμα την καλύτερη απόδοση του μοντέλου στα δεδομένα. Συγκεκριμένα ο

γραμμικός προγραμματισμός, που χρησιμοποιήσαμε, είναι μια μαθηματική τεχνική βελτιστοποίησης όπου ο στόχος είναι η μεγιστοποίηση ή η ελαχιστοποίηση μιας γραμμικής συνάρτησης που υπόκειται σε ένα σύνολο γραμμικών περιορισμών. Στη μηχανική μάθηση μπορεί να χρησιμοποιηθεί σε ορισμένα προβλήματα βελτιστοποίησης, όπως μηχανές διανυσμάτων στήριξης (SVM) για ταξινόμηση και κατανομή πόρων στην εκπαίδευση μοντέλων.

Προγραμματισμός Περιορισμών

Ο προγραμματισμός περιορισμών χρησιμοποιείται για την επίλυση συνδυαστικών προβλημάτων που περιλαμβάνουν περιορισμούς σε μεταβλητές. Επικεντρώνεται στην εύρεση εφικτών λύσεων που ικανοποιούν όλους τους δεδομένους περιορισμούς. Αυτή η προσέγγιση είναι ιδιαίτερα χρήσιμη για την επίλυση προβλημάτων όπου οι περιορισμοί είναι περίπλοκοι και δύσκολο να κωδικοποιηθούν σε παραδοσιακές μαθηματικές μορφές.

Sports Analytics

Τα Sports Analytics αναφέρονται στην εφαρμογή ανάλυσης δεδομένων, στατιστικών μεθόδων και τεχνολογίας για την απόκτηση γνώσεων και τη βελτίωση της απόδοσης στον αθλητισμό. Περιλαμβάνει τη συλλογή, την επεξεργασία και την ανάλυση δεδομένων που σχετίζονται με διάφορες πτυχές των αθλημάτων, όπως η απόδοση των παικτών, οι στρατηγικές της ομάδας και τα αποτελέσματα του παιχνιδιού. Τα Sports Analytics μπορούν να χρησιμοποιηθούν από ομάδες, προπονητές, παίκτες και οργανισμούς για να λάβουν τεκμηριωμένες αποφάσεις και να αποκτήσουν ανταγωνιστικό πλεονέκτημα.

Κεφάλαιο 3

Βιβλιογραφική ανασκόπηση

Ο Silva [11] παρέχει μια επισκόπηση του τομέα των sports analytics και των εφαρμογών του σε διάφορα αθλήματα. Ο συγγραφέας εξηγεί πώς τα sports analytics περιλαμβάνουν τη χρήση ανάλυσης δεδομένων και στατιστικών για τη λήψη τεκμηριωμένων αποφάσεων σχετικά με την απόδοση των παικτών, τη στρατηγική αγώνων και τη διαχείριση της ομάδας. Ο Silva καλύπτει μια σειρά θεμάτων, συμπεριλαμβανομένης της χρήσης δεδομένων στην αναζήτητση και τη στρατολόγηση, του ρόλου των στατιστικών στην ανάπτυξη και εκπαίδευση παικτών και τη χρήση προηγμένων αναλυτικών στοιχείων για τη βελτίωση της στρατηγικής και τη νίκη περισσότερων αγώνων. Επιπλέον, διερευνά τον αντίκτυπο των sports analytics στις αθλητικές εκπομπές και τη συμμετοχή των οπαδών, τονίζοντας πώς οι πληροφορίες που βασίζονται σε δεδομένα μπορούν να βελτιώσουν την εμπειρία θέασης για τους οπαδούς. Συνολικά, η μελέτη παρέχει μια ολοκληρωμένη εισαγωγή στα sports analytics και την αυξανόμενη σημασία τους στον κόσμο των αθλημάτων.

Οι Apostoloy και Tzortzis [1] πραγματεύονται τη χρήση των sports analytics στην πρόβλεψη της απόδοσης των αθλητών. Ο αρθρογράφος υπογραμμίζει τη σημασία της συλλογής και ανάλυσης δεδομένων στα σύγχρονα αθλήματα, συμπεριλαμβανομένης της χρήσης μηχανικής μάθησης και τεχνικών προγνωστικής μοντελοποίησης για τον εντοπισμό προτύπων και την πραγματοποίηση ακριβών προβλέψεων. Το άρθρο παρέχει παραδείγματα για το πώς χρησιμοποιούνται τα sports analytics σε διάφορα αθλήματα, όπως η πρόβλεψη τραυματισμών παικτών, η βελτιστοποίηση προγραμμάτων προπόνησης και η αξιολόγηση της απόδοσης των παικτών. Ο συγγραφέας εξετάζει επίσης ορισμένες από τις προκλήσεις που σχετίζονται με τα sports analytics, συμπεριλαμβανομένης της ποιότητας των δεδομένων και των ηθι-

κών ανησυχιών. Συνολικά, το άρθρο τονίζει τα πιθανά οφέλη των sports analytics για τη βελτίωση της απόδοσης των αθλητών, καθώς και τη σημασία της συνεργασίας μεταξύ αθλητικών οργανισμών και επιστημόνων δεδομένων για την ανάπτυξη αποτελεσματικών αλγορίθμων και μοντέλων.

Οι Sarlis και Tzortzis [10] διερευνούν τη χρήση sports analytics για την αξιολόγηση της απόδοσης παικτών και ομάδων μπάσκετ. Συζητά τη σημασία της συλλογής, ανάλυσης και οπτικοποίησης δεδομένων και υπογραμμίζει τη χρήση στατιστικών μοντέλων και αλγορίθμων μηχανικής μάθησης για τον εντοπισμό προτύπων και τάσεων. Το άρθρο εξετάζει επίσης πώς τα sports analytics μπορούν να χρησιμοποιηθούν για την ενημέρωση στρατηγικών αποφάσεων, όπως η ανάπτυξη παικτών, το χτίσιμο ενός ρόστερ και οι τακτικές εντός του παιχνιδιού. Οι αρθρογράφοι καταλήγουν στο συμπέρασμα ότι τα sports analytics είναι ένα ισχυρό εργαλείο για τη βελτίωση της απόδοσης των παικτών και των ομάδων και ότι θα συνεχίσει να παίζει έναν ολοένα και πιο σημαντικό ρόλο στον κόσμο του μπάσκετ.

Οι Horvat κ.α. [5] περιγράφουν τη χρήση του αλγόριθμου k-Nearest Neighbors (k-NN) για την πρόβλεψη των αποτελεσμάτων αγώνων μπάσκετ της Euroleague. Οι αρθρογράφοι χρησιμοποίησαν ένα σύνολο δεδομένων που περιείχε πληροφορίες σχετικά με τα στατιστικά της ομάδας και την απόδοση των παικτών και δοκίμασαν το μοντέλο τους σε αρκετές σεζόν της Euroleague. Βρήκαν ότι ο αλγόριθμος k-NN ήταν σε θέση να προβλέψει τα αποτελέσματα των αγώνων με υψηλό βαθμό ακρίβειας, υποδηλώνοντας ότι αυτή η προσέγγιση θα μπορούσε να είναι χρήσιμη για αθλητικά στοιχήματα ή άλλες εφαρμογές όπου οι ακριβείς προβλέψεις είναι σημαντικές. Συγκεκριμένα, χρησιμοποιήθηκαν δύο παραλλαγές για την προετοιμασία δεδομένων. Η πρώτη με όνομα DefenseOffense, η οποία θέλοντας να διαφοροποιήσει την απόδοση μιας ομάδας στην άμυνα και στην επίθεση έκανε χρήση των παρακάτω προηγμένων στατιστικών στοιχείων: δίποντα επιτυχημένα/προσπάθειες, τρίποντα επιτυχημένα/προσπάθειες, βολές επιτυχημένες/προσπάθειες, αμυντικά/επιθετικά ριμπάουντ, ασίστ, κλεψίματα, λάθη, μπλοκ εναντίον/υπέρ, φάουλ που έκαναν/τους έκαναν και η δεύτερη με όνομα Components η οποία είναι μια πιο γενική προσέγγιση, έκανε χρήση των παρακάτω γενικευμένων στατιστικών στοιχείων: πόντοι, ριμπάουντ, κλεψίματα/λάθη, ασίστ, μπλοκ, φάουλ. Από τα αποτελέσματα παρατηρήθηκε ότι για τις πέντε χρονιές που χρησιμοποιήθηκαν οι δύο αλγόριθμοι

και για k=3,5,7,9 η παραλλαγή DefenseOffense είχε καλύτερη ευστοχία με 70-84%, ενώ η παραλλαγή Components 57-77%.

Οι Osken και Onay [8] εξετάζουν μια νέα προσέγγιση για την πρόβλεψη της νικήτριας ομάδας σε αγώνες μπάσκετ. Η προσέγγιση χρησιμοποιεί έναν συνδυασμό τεχνικών μηχανικής μάθησης, συμπεριλαμβανομένης της ομαδοποίησης και νευρωνικών δικτύων (GA-ANN: Genetic Algorithm-Artificial Neural Networks), για να αναλύσει διάφορα στατιστικά στοιχεία, όπως πόντοι ομάδας, ριμπάουντ και λάθη. Συγκεκριμένα, έγινε χρήση μιας διπλής προσέγγισης στη μηχανή πρόβλεψης: πρώτα εκτελείται μια μέθοδος ομαδοποίησης (clustering), συγκεκριμένα η c-means, για τον προσδιορισμό του τύπου παικτών και στη συνέχεια γίνεται χρήση αυτών των τύπων παικτών σε κάθε ομάδα για την πρόβλεψη των αποτελεσμάτων των αγώνων, χρησιμοποιώντας τεχνητό νευρωνικό δίκτυο (Artificial Neural Network - ANN). Πάρθηκαν 49 στατιστικά στοιχεία στα οποία εφαρμόστηκε ανάλυση κύριων συνιστωσών με μια αποκοπή που ορίζεται στο 90% της αθροιστικής διακύμανσης και προσδιορίστηκαν 15 συνιστώσες. Επίσης, προστέθηκαν άλλα τρία χαρακτηριστικά: Αριθμός ημερών μεταξύ της ημερομηνίας του αγώνα και του τελευταίου αγώνα που έπαιξαν οι γηπεδούχοι και οι φιλοξενούμενοι, ψευδομεταβλητές (dummy) για τον μήνα της σεζόν και ποσοστό νίκης κάθε ομάδας μέχρι την ημερομηνία του αγώνα σε εκείνη τη σεζόν. Η μελέτη διαπίστωσε ότι αυτά τα μοντέλα ήταν σε θέση να προβλέψουν με ακρίβεια τη νικήτρια ομάδα με υψηλό βαθμό ακρίβειας (ακρίβεια πρόβλεψης 76% σε μια περίοδο πέντε σεζόν NBA και ακρίβεια πρόβλεψης 71% σε μια σεζόν που δεν χρησιμοποιείται για προπόνηση μοντέλων), ξεπερνώντας τις παραδοσιακές μεθόδους πρόβλεψης. Το άρθρο καταλήγει στο συμπέρασμα ότι αυτή η προσέγγιση θα μπορούσε να είναι χρήσιμη για αθλητικούς αναλυτές και στοιχηματικές που θέλουν να κάνουν πιο ακριβείς προβλέψεις σε αγώνες μπάσκετ.

Οι Chen κ.α. [3] προτείνουν ένα υβριδικό μοντέλο πρόβλεψης αποτελεσμάτων αγώνων μπάσκετ που ενσωματώνει μεθόδους μηχανικής μάθησης και στατιστικής ανάλυσης για την Εθνική Ένωση Μπάσκετ (NBA). Οι συγγραφείς συνέλεξαν δεδομένα από 1.230 αγώνες NBA τη σεζόν 2019-2020 και χρησιμοποίησαν τεχνικές επιλογής χαρακτηριστικών για να προσδιορίσουν τα πιο σημαντικά στατιστικά στοιχεία. Χρησιμοποιήθηκαν πέντε τεχνικές εξόρυξης δεδομένων, συγκεκριμένα: ακραία μηχανή εκμάθησης ELM (extreme learning machine), πολυμεταβλητές προσαρμο-

στικές στροφές παλινδρόμησης MARS (multivariate adaptive regression splines), k-πλησιέστεροι γείτονες KNN (k-nearest neighbors), ακραία ενίσχυση κλίσης XGBoost (eXtreme gradient boosting) και στοχαστική ενίσχυση κλίσης SGB (stochastic gradient boosting) για να αναπτύξουν ένα νέο σχήμα το οποίο έχει δύο παραλλαγές, το Ενιαίο Μοντέλο (Single Model) και το Μοντέλο Δύο Σταδίων (Two Stage). Αυτή η μελέτη χρησιμοποίησε το μέσο απόλυτο ποσοστό σφάλματος (MAPE) ως δείκτη για την αξιολόγηση της απόδοσης των μοντέλων πρόβλεψης. Το μοντέλο T-XGBoost χρησιμοποιώντας game-lag = 4 πέτυχε την καλύτερη απόδοση πρόβλεψης μεταξύ των 10 ανταγωνιστικών μοντέλων, με MAPE = 0.0818. Τα έξι σημαντικά στατιστικά (χαρακτηριστικά) που προσδιορίστηκαν με βάση game-lag = 4 είναι, μέσος όρος αμυντικών ριμπάουντ, μέσος όρος ποσοστού δύο πόντων, μέσος όρος ποσοστού ελεύθερων βολών, μέσος όρος επιθετικών ριμπάουντ, μέσου όρος ασίστ και μέσος όρος προσπαθειών τριών πόντων. Τα ευρήματα αυτής της μελέτης μπορούν να εφαρμοστούν στην ανάπτυξη αρκετών εφαρμογών για άλλες ομάδες ή ακόμα και μεμονωμένα αθλήματα.

Οι Vinué και Epifanio [12] παρουσιάζουν μια μέθοδο για την πρόβλεψη της μελλοντικής απόδοσης των παικτών μπάσκετ χρησιμοποιώντας αραιά λειτουργικά δεδομένα. Οι συγγραφείς χρησιμοποιούν δεδομένα από την Εθνική Ομοσπονδία Καλαθοσφαίρισης (NBA) και εφαρμόζουν μια ανάλυση λειτουργικού κύριου στοιχείου (FPCA) για να εξαγάγουν χαρακτηριστικά από τα δεδομένα. Στη συνέχεια, χρησιμοποιούν μια τεχνική αραιής παλινδρόμησης για να επιλέξουν τα πιο σχετικά χαρακτηριστικά και να προσαρμόσουν ένα μοντέλο για να προβλέψουν τη μελλοντική απόδοση των παικτών. Τέλος, κάνουν χρήση του αλγόριθμου ROPES [4] για την πρόβλεψη της απόδοσης των παικτών και τον συγκρίνουν με τον αλγόριθμο CARMELO [6]. Η μελέτη χρησιμοποίησε έναν μεγάλο αριθμό χαρακτηριστικών, συμπεριλαμβανομένων των βασικών στατιστικών βαθμολογίας κουτιού (όπως πόντοι, ριμπάουντ και ασίστ), προηγμένα στατιστικά (όπως βαθμολογία απόδοσης παικτών και ποσοστό πραγματικών σουτ) και μεταβλητές με βάση τα συμφραζόμενα (όπως η απόδοση της ομάδας και η κατάσταση του παιχνιδιού). Το μέσο τετραγωνικό σφάλμα πρόβλεψης (MSPE) χρησιμοποιήθηκε ως μέτρο ακρίβειας της πρόβλεψης αναφέρθηκε ότι ήταν 0.0275 για τη μεταβλητή των πόντων και 0.0132 για τη μεταβλητή των ριμπάουντ. Αυτές οι τιμές υποδεικνύουν ένα σχετικά χαμηλό σφάλμα

πρόβλεψης, υποδηλώνοντας ότι το μοντέλο έχει κάποια προγνωστική ισχύ για αυτές τις μεταβλητές. Τα αποτελέσματα δείχνουν ότι η προτεινόμενη μέθοδος υπερέχει από πολλές άλλες μεθόδους που χρησιμοποιούνται συνήθως. Οι συγγραφείς καταλήγουν στο συμπέρασμα ότι η μέθοδός τους παρέχει μια πολλά υποσχόμενη προσέγγιση για την πρόβλεψη της απόδοσης των μπασκετμπολίστων.

Οι Piette κ.α. [9] προτείνουν ένα στατιστικό μοντέλο δικτύου για την αξιολόγηση της απόδοσης των παικτών μπάσκετ. Το μοντέλο χρησιμοποιεί μια αναπαράσταση που βασίζεται σε γράφημα για να καταγράψει τις αλληλεπιδράσεις και τις εξαρτήσεις μεταξύ των παικτών σε μια ομάδα, με τους παίκτες να αναπαριστώνται ως κόμβοι και οι αλληλεπιδράσεις τους ως ακμές. Στη συνέχεια χρησιμοποιούν στατιστικά μοντέλα δικτύου, όπως μοντέλα εκθετικών τυχαίων γραφημάτων, για να αναλύσουν τα δεδομένα και να αξιολογήσουν την απόδοση μεμονωμένων παικτών. Η απόδοση ενός παίκτη αξιολογείται με βάση τη συμβολή του στη συνολική απόδοση της ομάδας. Οι συγγραφείς χρησιμοποιούν τα μοντέλα δικτύου για να εντοπίσουν σημαντικά χαρακτηριστικά που συμβάλλουν στην επιτυχία μιας ομάδας, όπως η κίνηση της μπάλας και η αμυντική πίεση. Επίσης, γίνεται χρήση τριών μέτρων κεντρικότητας για τον εντοπισμό των πιο σημαντικών κόμβων σε ένα δίκτυο, οι οποίοι μπορούν να ερμηνευθούν ως οι παίκτες με τη μεγαλύτερη επιρροή ή καλά συνδεδεμένους παίκτες σε έναν αγώνα μπάσκετ, αυτά είναι: κεντρικότητα βαθμού, κεντρικότητα ιδιοδιανύσματος και κεντρικότητα μεταξύ. Η κεντρικότητα βαθμού μετρά τον αριθμό των άμεσων συνδέσεων που έχει ένας παίκτης με άλλους παίκτες, ενώ η κεντρικότητα ιδιοδιανύσματος λαμβάνει υπόψη τόσο τον αριθμό όσο και τη σημασία των συνδέσεων ενός παίκτη. Η κεντρικότητα μεταξύ μετράει τον βαθμό στον οποίο ένας παίκτης λειτουργεί ως γέφυρα μεταξύ άλλων παικτών στο δίκτυο. Αναλύοντας αυτά τα μέτρα κεντρικότητας, οι συγγραφείς είναι σε θέση να εντοπίσουν τους παίκτες με τη μεγαλύτερη επιρροή σε ένα παιχνίδι και να αξιολογήσουν την απόδοσή τους. Διαπιστώνουν ότι οι παίκτες με υψηλές βαθμολογίες κεντρικότητας είναι πιο πιθανό να έχουν θετικό αντίκτυπο στο αποτέλεσμα του παιχνιδιού και ότι αυτά τα μέτρα μπορούν να χρησιμοποιηθούν για την πρόβλεψη της απόδοσης μεμονωμένων παικτών σε μελλοντικά παιχνίδια.

Διαπιστώνουν ότι αυτά τα χαρακτηριστικά μπορούν να χρησιμοποιηθούν για την πρόβλεψη της έκβασης των αγώνων μπάσκετ με υψηλό βαθμό ακρίβειας. Η

προτεινόμενη μέθοδος συγκρίνεται με τις υπάρχουσες μεθόδους που χρησιμοποιούν δεδομένα από το NBA και τα αποτελέσματα δείχνουν ότι μπορεί να εντοπίσει αποτελεσματικά παίκτες κλειδιά και να παρέχει χρήσιμες πληροφορίες για την απόδοση της ομάδας. Οι αρθρογράφοι υποδηλώνουν ότι η προσέγγισή τους θα μπορούσε να χρησιμοποιηθεί από προπονητές και αναλυτές για να λάβουν πιο ενημερωμένες αποφάσεις σχετικά με την επιλογή παικτών και τη στρατηγική εντός του αγώνα.

Οι Balli και Özdemir [2] παρουσιάζουν μια νέα προσέγγιση για την πρόβλεψη των αποτελεσμάτων των αγώνων μπάσκετ της EuroLeague χρησιμοποιώντας έναν συνδυασμό τεχνικών μηχανικής εκμάθησης και μεθόδων εξαγωγής χαρακτηριστικών. Συγκεκριμένα, για την εξαγωγή χαρακτηριστικών έγινε χρήση του μοντέλου DefenseOfense από το [5], του μοντέλου Four Factors, που είναι παρόμοιο με το DefenseOfense, με τη διαφορά ότι έχει υπολογιστεί η επίδραση των παρακάτω τεσσάρων στατιστικών στοιχείων στη νίκη μιας ομάδας: σουτ 40%, χαμένες κατοχές 25%, ριμπάουντ 20% και ελεύθερες βολές 15%. Επίσης, έγινε χρήση του μοντέλου DefenseOfense detailed που περιέχει 16 υπό-χαρακτηριστικά που χρησιμοποιούνται στο μοντέλο DefenseOfense και του μοντέλου Four Factors detailed που περιέχει 8 υπό-χαρακτηριστικά που χρησιμοποιούνται στο μοντέλο Four Factors. Ακόμα, δημιουργήθηκαν άλλα τέσσερα υβριδικά μοντέλα από τον συνδυασμό των κύριων μοντέλων. Οι τεχνικές μηχανικής μάθησης που χρησιμοποιήθηκαν για να γίνουν οι προβλέψεις είναι οι εξής: k-Nearest Neighbors (k-NN), Λογιστική παλινδρόμηση, Πολυστρωματικό Αντίληπτρο (Multilayer perceptron), Naive Bayes, j48 και Voting. Για την αξιολόγηση της απόδοσης των μοντέλων έγινε χρήση των F-measure, ROC, Accuracy και RMSE. Ως αποτέλεσμα της αξιολόγησης των μοντέλων γίνεται αντιληπτό ότι το μοντέλο που συνδυάζει DefenseOfense και Four Factors detailed με τον αλγόριθμο Multilayer perceptron φαίνεται ότι επιτυγχάνει τη μεγαλύτερη επιτυχία με ποσοστό 98.9%.

Οι Nguyen κ.α. [7] παρουσιάζουν μια μέθοδο για την πρόβλεψη της απόδοσης των παικτών μπάσκετ στο NBA και τη δημοτικότητά τους. Για την πρόβλεψη της απόδοσης των παικτών δοκιμάστηκαν διάφοροι αλγόριθμοι μηχανικής μάθησης, όπως γραμμική παλινδρόμηση (Linear Regression), Gradient boosting machine, μηχανές διανυσματικής υποστήριξης (Support vector machine), "τυχαία δέντρα" (Random forest) και νευρονικά δίκτυα (Neural network). Οι συγγραφείς χρησιμοποιούν 19

προγνωστικές μεταβλητές οι οποίες πάρθηκαν από την Εθνική Ομοσπονδία Καλαθοσφαίρισης (NBA) και σαν μετρικές απόδοσης επελέγησαν, η ρίζα του μέσου τετραγωνικού σφάλματος (Root Mean Square Error - RMSE) και το μέσο απόλυτο σφάλμα (Mean Absolute Error - MAE). Αφού έγινε χρήση δεκάπτυχης διασταυρωμένης επικύρωσης συμπέραναν ότι ο αλγόριθμος Gradient boosting machine δίνει σταθερά τα καλύτερα αποτελέσματα με RMSE = 2.19 και MAE = 1.64. Δοκιμάστηκε και η τεχνική βαθιά μάθηση (Deep Learning), αλλά τα αποτελέσματα δεν ήταν αποδεκτά σύμφωνα με τους συγγραφείς.

Κεφάλαιο 4

Υλοποίηση

Σε αυτό το κεφάλαιο, παρουσιάζονται τα αποτελέσματα της αναπαραγωγής των πειραμάτων από την υπάρχουσα βιβλιογραφία. Επικεντρώνονται στην πρόβλεψη της απόδοσης των παικτών μπάσκετ και της επιτυχίας της ομάδας στην Ευρωλίγκα. Πρωταρχικός στόχος αυτής της προσπάθειας είναι η ανεξάρτητη επαλήθευση των αποτελεσμάτων που παρουσιάζονται από τους αρχικούς συγγραφείς, διασφαλίζοντας την αξιοπιστία και την αναπαραγωγιμότητα των ευρημάτων τους. Με την ανακατασκευή των πειραματικών ρυθμίσεων και την εφαρμογή των ίδιων μεθοδολογιών, μπορεί να αξιολογηθεί η ακρίβεια των μοντέλων που χρησιμοποιούνται και η δυνατότητα εφαρμογής τους σε πραγματικά σενάρια.

Πίνακας 4.1: Tomislav Horvat

Variant	**Seasons**	**Accuracy (%)**			
		k=3	**k=5**	**k=7**	**k=9**
DefenseOfense	1	75.25%	76.24%	78.22%	78.22%
	3	80.86%	82.51%	82.18%	82.84%
	5	82.02%	82.41%	82.81%	83.00%
Components	1	72.28%	73.27%	78.22%	77.23%
	3	75.25%	77.89%	79.21%	79.54%
	5	76.48%	76.88%	75.49%	76.48%

Ο Πίνακας 4.1 εμπεριέχει τα αποτελέσματα από τη δική μας αναπαραγωγή του πειράματος [5] και παρατηρούμε ότι το μοντέλο DefenseOfense όντως δίνει την καλύτερη ακρίβεια και συγκεκριμένα η χρήση 5 σεζόν με 9 γείτονες με 83%.

Ο Πίνακας 4.2 εμπεριέχει τα αποτελέσματα από τη δική μας αναπαραγωγή του

Πίνακας 4.2: 10-Fold Cross-Validation Accuracy (%) 5 Seasons

10-Fold Cross-Validation Accuracy (%) 5 Seasons																
Algorithm	**DefenseOfense**		**Four Factors**		**DefenseOfense Detailed**		**Four Factors Detailed**		**Model 5**		**Model 6**		**Model 7**		**Model 8**	
	Average	**Best**	**Average**	**Best**	**Average**	**Best**	**Average**	**Best**	**Average**	**Best**	**Average**	**Best**	**Average**	**Best**	**Average**	**Best**
J48	79.49%	85.77%	64.23%	72.73%	65.81%	71.54%	79.33%	83.40%	78.46%	83.79%	83.72%	88.14%	72.17%	78.26%	78.50%	83.79%
Naive Bayes	84.31%	88.93%	74.03%	79.05%	77.19%	83.00%	88.26%	91.30%	83.48%	88.93%	90.87%	94.86%	79.17%	84.58%	85.06%	89.72%
Multilayer Perceptron	84.58%	89.72%	74.07%	79.05%	84.23%	87.35%	93.32%	96.84%	84.66%	89.72%	92.06%	95.26%	84.07%	87.75%	87.00%	90.91%
Logistic Regression	84.58%	89.33%	74.07%	79.05%	83.79%	86.96%	91.94%	94.47%	84.74%	89.33%	89.45%	93.28%	84.03%	86.96%	89.05%	93.28%
XGBoost	83.99%	88.14%	71.74%	77.08%	74.51%	78.26%	87.94%	90.51%	83.52%	88.14%	83.87%	90.12%	74.23%	78.26%	74.39%	78.26%
Support Vector Machine	84.55%	88.93%	74.07%	79.05%	82.85%	86.96%	93.04%	96.05%	84.55%	88.93%	84.58%	88.93%	82.57%	87.35%	82.49%	87.35%
Stochastic Gradient Descent	84.62%	88.93%	72.57%	80.24%	78.66%	86.96%	92.06%	94.86%	83.28%	91.30%	83.16%	92.09%	79.13%	87.35%	81.54%	86.17%
Random Forest	82.02%	89.33%	71.42%	79.05%	75.53%	84.58%	92.92%	95.65%	83.87%	89.72%	84.98%	92.09%	76.48%	84.58%	80.79%	87.35%
Neighbors	k=11		k=15		k=15		k=15		k=7		k=11		k=15		k=9	
K-Nearest-Neighbors	83.79%	88.93%	71.74%	77.08%	74.51%	78.26%	87.94%	90.51%	83.24%	88.93%	83.75%	90.51%	74.23%	78.26%	73.28%	78.66%

πειράματος [2] με δεκάπτυχη διασταυρωμένη επικύρωση για τις χρονιές 2012-2017 και παρατηρούμε ότι το μοντέλο Four Factors detailed με τον αλγόριθμο Support Vector Machine δίνει την καλύτερη ακρίβεια με 96.05%.

Πίνακας 4.3: 10-Fold Cross-Validation Accuracy (%) All Seasons

10-Fold Cross-Validation Accuracy (%) All Seasons																
Algorithm	**DefenseOfense**		**Four Factors**		**DefenseOfense Detailed**		**Four Factors Detailed**		**Model 5**		**Model 6**		**Model 7**		**Model 8**	
	Average	**Best**	**Average**	**Best**	**Average**	**Best**	**Average**	**Best**	**Average**	**Best**	**Average**	**Best**	**Average**	**Best**	**Average**	**Best**
J48	78.61%	81.69%	62.98%	68.33%	67.46%	74.70%	80.66%	84.27%	77.43%	80.33%	83.53%	87.75%	73.77%	78.64%	79.91%	83.79%
Naive Bayes	83.52%	87.88%	73.21%	77.58%	75.90%	79.09%	89.70%	91.97%	82.58%	86.06%	90.62%	93.79%	78.32%	81.82%	84.44%	88.18%
Multilayer Perceptron	83.96%	89.09%	73.36%	77.58%	83.88%	88.18%	93.38%	95.01%	83.22%	87.12%	92.80%	95.76%	84.05%	89.09%	91.11%	93.64%
Logistic Regression	83.81%	88.64%	73.34%	77.58%	84.16%	89.24%	92.77%	95.31%	84.08%	88.48%	91.14%	94.10%	84.58%	89.55%	91.71%	94.24%
XGBoost	83.20%	88.94%	70.34%	73.37%	73.77%	76.21%	87.47%	90.76%	82.46%	87.58%	83.07%	88.79%	73.84%	76.06%	73.96%	76.21%
Support Vector Machine	83.84%	88.94%	73.27%	77.42%	83.81%	88.33%	93.27%	95.61%	83.81%	88.94%	83.93%	89.24%	83.76%	88.03%	83.64%	88.18%
Stochastic Gradient Descent	83.13%	88.20%	73.17%	76.67%	78.88%	85.33%	92.71%	95.31%	82.91%	87.73%	85.19%	90.32%	76.46%	85.91%	80.11%	87.42%
Random Forest	82.55%	86.84%	72.55%	76.82%	79.36%	86.06%	91.52%	94.24%	82.25%	86.54%	84.49%	88.05%	74.93%	86.69%	77.61%	88.79%
Neighbors	k=15		k=13		k=13		k=13		k=15		k=15		k=11		k=11	
K-Nearest-Neighbors	82.57%	87.88%	70.36%	74.55%	73.92%	76.21%	87.32%	90.91%	82.46%	87.58%	83.07%	88.79%	73.66%	76.52%	73.84%	76.52%

Ο Πίνακας 4.3 εμπεριέχει τα αποτελέσματα από τη δική μας αναπαραγωγή του πειράματος [2] με δεκάπτυχη διασταυρωμένη επικύρωση για τις χρονιές 2012-2023 και παρατηρούμε ότι το μοντέλο 6 που συνδιάζει DefenseOfense και Four Factors detailed με τον αλγόριθμο Multilayer perceptron δίνει την καλύτερη ακρίβεια με 95.76%. Επίσης, παρατηρούμε μια μικρή πτώση στην ακρίβεια των αλγορίθμων, κάτι που περιμέναμε λόγω τις αύξησης του όγκου δεδομένων.

Πίνακας 4.4: 66% Train Set Accuracy (%) 5 Seasons

66% Train Set Accuracy (%) 5 Seasons								
Algorithm	**DefenseOfense**	**Four Factors**	**DefenseOfense Detailed**	**Four Factors Detailed**	**Model 5**	**Model 6**	**Model 7**	**Model 8**
J48	79.91%	65.51%	66.43%	80.49%	78.63%	85.95%	71.89%	78.98%
Naive Bayes	85.48%	73.87%	75.96%	89.43%	83.04%	90.59%	80.37%	87.22%
Multilayer Perceptron	85.60%	74.10%	83.16%	94.08%	85.25%	90.59%	83.74%	86.06%
Logistic Regression	85.60%	73.87%	84.09%	91.87%	84.90%	88.04%	83.74%	89.90%
XGBoost	83.74%	74.22%	78.05%	89.66%	84.09%	93.38%	83.39%	90.13%
Support Vector Machine	84.55%	74.80%	81.77%	92.45%	84.09%	85.83%	82.35%	83.74%
Stochastic Gradient Descent	84.20%	73.40%	83.16%	93.03%	83.97%	83.39%	82.81%	69.57%
Random Forest	80.49%	70.03%	77.82%	89.08%	82.69%	92.80%	81.77%	88.04%
Neighbors	k=13	k=15	k=13	k=15	k=11	k=9	k=15	k=13
K-Nearest-Neighbors	84.32%	72.36%	74.33%	88.04%	84.79%	83.62%	77.24%	78.28%

Ο Πίνακας 4.4 εμπεριέχει τα αποτελέσματα από τη δική μας αναπαραγωγή του πειράματος [2] με το 66% τον δεδομένων στο σετ εκπαίδευσης για τις χρονιές 2012-2017 και παρατηρούμε ότι το μοντέλο Four Factors detailed με τον αλγόριθμο Multilayer perceptron δίνει την καλύτερη ακρίβεια με 94.08%.

Πίνακας 4.5: 66% Train Set Accuracy (%) All Seasons

66% Train Set Accuracy (%) All Seasons								
Algorithm	**DefenseOfense**	**Four Factors**	**DefenseOfense Detailed**	**Four Factors Detailed**	**Model 5**	**Model 6**	**Model 7**	**Model 8**
J48	79.55%	62.76%	68.64%	81.02%	78.49%	84.05%	74.52%	81.60%
Naive Bayes	84.01%	72.78%	77.64%	88.78%	83.07%	90.11%	77.46%	85.12%
Multilayer Perceptron	83.83%	72.92%	83.74%	93.36%	84.10%	91.76%	84.86%	88.33%
Logistic Regression	84.32%	72.74%	84.45%	92.07%	83.74%	90.29%	84.90%	91.27%
XGBoost	83.61%	72.20%	79.20%	89.58%	83.96%	91.98%	83.47%	91.31%
Support Vector Machine	84.37%	73.18%	84.05%	92.96%	83.70%	83.61%	83.47%	83.92%
Stochastic Gradient Descent	82.27%	72.47%	77.51%	92.25%	83.03%	84.01%	66.15%	71.40%
Random Forest	81.07%	68.06%	77.95%	89.40%	83.25%	90.38%	81.60%	89.40%
Neighbors	k=7	k=15	k=15	k=15	k=11	k=7	k=15	k=15
K-Nearest-Neighbors	83.07%	70.51%	75.32%	88.11%	83.16%	83.74%	76.17%	75.59%

Ο Πίνακας 4.5 εμπεριέχει τα αποτελέσματα από τη δική μας αναπαραγωγή του πειράματος [2] με το 66% τον δεδομένων στο σετ εκπαίδευσης για τις χρονιές 2012-2023 και παρατηρούμε ότι το μοντέλο Four Factors detailed με τον αλγόριθμο Multilayer perceptron δίνει την καλύτερη ακρίβεια με 93.36%.

Πίνακας 4.6: 80% Train Set Accuracy (%) 5 Seasons

80% Train Set Accuracy (%) 5 Seasons								
Algorithm	**DefenseOfense**	**Four Factors**	**DefenseOfense Detailed**	**Four Factors Detailed**	**Model 5**	**Model 6**	**Model 7**	**Model 8**
J48	79.84%	65.22%	70.75%	79.64%	81.23%	86.76%	74.31%	77.47%
Naive Bayes	84.58%	76.28%	77.67%	92.09%	81.03%	91.50%	80.63%	83.79%
Multilayer Perceptron	84.98%	76.28%	81.82%	93.48%	83.40%	89.53%	82.02%	87.35%
Logistic Regression	84.78%	75.89%	82.21%	92.49%	85.18%	88.93%	81.62%	89.92%
XGBoost	86.36%	73.12%	83.00%	88.93%	84.39%	93.08%	84.98%	89.33%
Support Vector Machine	86.76%	73.32%	85.18%	92.49%	84.19%	84.39%	82.41%	81.23%
Stochastic Gradient Descent	84.39%	73.52%	84.58%	90.12%	81.23%	86.76%	82.21%	83.99%
Random Forest	80.83%	69.37%	79.84%	88.34%	83.99%	91.50%	82.81%	87.15%
Neighbors	k=15	k=11	k=11	k=11	k=15	k=11	k=13	k=13
K-Nearest-Neighbors	83.79%	75.49%	77.27%	89.92%	83.60%	83.79%	75.69%	74.70%

Ο Πίνακας 4.6 εμπεριέχει τα αποτελέσματα από τη δική μας αναπαραγωγή του πειράματος [2] με το 80% τον δεδομένων στο σετ εκπαίδευσης για τις χρονιές 2012-2017 και παρατηρούμε ότι το μοντέλο Four Factors detailed με τον αλγόριθμο Multilayer perceptron δίνει την καλύτερη ακρίβεια για ακόμη μια φορά με 93.48%.

Πίνακας 4.7: 80% Train Set Accuracy (%) All Seasons

80% Train Set Accuracy (%) All Seasons								
Algorithm	**DefenseOfense**	**Four Factors**	**DefenseOfense Detailed**	**Four Factors Detailed**	**Model 5**	**Model 6**	**Model 7**	**Model 8**
J48	76.99%	63.56%	66.41%	81.11%	77.64%	83.03%	73.81%	80.76%
Naive Bayes	82.59%	72.52%	77.82%	90.11%	82.85%	91.18%	78.04%	84.01%
Multilayer Perceptron	82.59%	72.61%	83.96%	93.05%	84.86%	93.05%	83.83%	89.04%
Logistic Regression	82.89%	72.56%	84.45%	92.12%	84.59%	91.14%	84.59%	91.40%
XGBoost	82.21%	72.78%	79.87%	90.69%	84.10%	92.25%	83.03%	90.82%
Support Vector Machine	83.04%	72.52%	84.10%	93.27%	84.94%	84.50%	83.65%	82.98%
Stochastic Gradient Descent	81.53%	72.69%	60.40%	92.61%	83.83%	84.01%	83.30%	53.36%
Random Forest	78.73%	68.37%	79.29%	89.98%	83.21%	91.27%	81.78%	88.73%
Neighbors	k=15	k=15	k=13	k=13	k=11	k=7	k=15	k=15
K-Nearest-Neighbors	81.30%	71.22%	76.39%	87.57%	84.28%	84.59%	74.88%	74.61%

Ο Πίνακας 4.7 εμπεριέχει τα αποτελέσματα από τη δική μας αναπαραγωγή του πειράματος [2] με το 80% τον δεδομένων στο σετ εκπαίδευσης για τις χρονιές 2012-2023 και παρατηρούμε ότι το μοντέλο Four Factors detailed με τον αλγόριθμο Support Vector Machine δίνει την καλύτερη ακρίβεια για ακόμη μια φορά με 93.27%.

Πίνακας 4.8: Nguyen Hoang Nguyen (Per Season)

Player Efficiency Rating (Per Season)				
Algorithm	**RMSE**		**MAE**	
	Average	**Best**	**Average**	**Best**
Linear Regression	25.53	23.88	19.63	4.24
Gradient Boosting Machine	30.86	28.37	22.66	4.57
Linear Support Vector Machine	26.10	24.27	19.73	4.29
Polynomial Support Vector Machine	80.34	74.55	64.72	7.95
Random Forest	35.12	34.24	25.10	4.90
Neural Network	28.44	27.35	20.75	4.49
Lasso	25.54	24.56	19.54	4.42
Ridge	25.54	24.70	19.56	4.43
Decision Trees	51.62	49.82	38.11	6.19

Ο Πίνακας 4.8 εμπεριέχει τα αποτελέσματα από τη δική μας αναπαραγωγή του πειράματος [7]. Παρατηρούμε ότι τα αποτελέσματα είναι πολύ διαφορετικά από αυτά που των συγγραφέων, αυτό συμβαίνει διότι δεν μπορέσαμε να υπολογίσουμε τα win shares (ws) για τον κάθε παίκτη. Γι'αυτό το λόγο αποφασίσαμε να τα αντικαταστήσουμε με το συνολικό player efficiency rating (per) για κάθε χρονιά που συμμετείχαν οι παίκτες στην Euroleague που έχουμε διαθέσιμο. Λαμβάνοντας υπόψιν τη μεγάλη αριθμητική διαφορά μεταξύ ws και per, τα αποτελέσματα είναι αποδεκτά. Αυτό που παρατηρούμε όμως είναι πως η γραμμική παλινδρόμιση δίνει ελαφρός καλύτερα αποτελέσματα από τον προτεινόμενο αλγόριθμο Gradient boosting machine.

Πίνακας 4.9: Nguyen Hoang Nguyen (Per Game)

Player Efficiency Rating (Per Game)				
Algorithm	**RMSE**		**MAE**	
	Average	**Best**	**Average**	**Best**
Linear Regression	2.35	2.30	1.81	1.34
Gradient Boosting Machine	2.04	2.04	1.52	1.22
Linear Support Vector Machine	2.36	2.32	1.80	1.33
Random Forest	3.52	3.54	2.70	1.64
Neural Network	1.81	1.73	1.34	1.14
Lasso	2.38	2.33	1.84	1.34
Ridge	2.35	2.30	1.81	1.34
Decision Trees	3.60	3.52	2.68	1.61

Ο Πίνακας 4.9 εμπεριέχει τα αποτελέσματα από τη δική μας αναπαραγωγή του πειράματος [7], αντικαθιστώντας τα win shares (ws) του κάθε παίκτη με το player efficiency rating (per) για κάθε αγώνα. Αυτό που παρατηρούμε είναι πως η χρήση νευρονικών δικτύων, συγκεκριμένα του Multilayer Perceptron, δίνει καλύτερα αποτελέσματα από τον προτεινόμενο αλγόριθμο Gradient boosting machine.

Με βάση τα αποτελέσματα των παραπάνω αλγορίθμων μπορούμε να συμπεράνουμε ότι μπορούμε να προβλέψουμε με μεγάλη ακρίβεια τη νικητήρια ομάδα σε αγωνές μπάσκετ στην Ευρωλίγκα, καθώς και την απόδοση των παικτών αυτών των ομάδων. Οι παραπάνω αλγόριθμοι θα μπορούσαν να χρησιμοποιηθούν για στοιχηματικούς σκοπούς, για την καλύτερη επιλογή παικτών στο ιντερνετικό παιχνίδι Fantasy Euroleague με σκοπό τη μεγαλύτερη συγκομιδή βαθμών, από μπασκετικούς συλλόγους για την κατάλληλη επιλογή παικτών σύμφωνα με τον αντίπαλο καθώς και για εύρεση παικτών για μελλοντικές μεταγραφές.

Κεφάλαιο 5

Πειραματική Διαδικασία

Για τη συλλογή των δεδομένων αναπτύχθηκε ένα script σε Python, για τη σάρωση δεδομένων ιστοσελίδων (web crawler) για την ιστοσελίδα basketnews.com με την βοήθεια της βιβλιοθήκης BeautifulSoup, ο οποίος βρίσκει όλους τους παίκτες που συμμετείχαν στην Euroleague για τη χρονιά 2023/2024, διαβάζει την καρτέλα του κάθε παίκτη και συλλέγει όλα τα στατιστικά στοιχεία, την θέση αλλά και την ομάδα για την οποία έπαιζε για όλες τις χρονιές που συμμετείχε στην Euroleague. Έπειτα, αναπτύχθηκε ένα script για τη σάρωση δεδομένων ιστοσελίδων για την ιστοσελίδα basketstories.net, ο οποίος συλλέγει την αξία όλων των παικτών για το Fantasy Euroleague και την αντιστοιχεί με κάθε παίκτη. Τα δεδομένα που συλλέξαμε είναι διαθέσιμα και στην επίσημη ιστοσελίδα της Euroleague, όμως έχει δημιουργηθεί με τέτοιο τρόπο που δεν επιτρέπει την ανίχνευσή της, γι' αυτό τον λόγο χρειάστηκε να βρεθούν άλλες εναλλακτικές. Με τα στατιστικά στοιχεία που συλλέχθηκαν μπορέσαμε να υπολογίσουμε άλλα, πιο προηγμένα στατιστικά στοιχεία, όπως το Αληθινό Ποσοστό Σουτ (True Shooting %), το Ποσοστό της Πραγματικής Ευστοχίας Εντός Πεδιάς (Effective Field Goal %) και το Ποσοστό Χρήσης Παίκτη (Player Usage %). Αφότου υπολογίσαμε όλα τα προηγμένα στατιστικά που χρειάστηκαν, δημιουργήσαμε πλαίσια δεδομένων χρησιμοποιώντας την βιβλιοθήκη Pandas, ένα που έχει όλα τα δεδομένα για τους παίκτες για όλες τις χρονιές που συμμετείχαν στην Euroleague, ένα που έχει όλα τα δεδομένα για τους παίκτες μόνο για τη χρονιά 2023/2024, ένα που έχει όλα τα δεδομένα για τους παίκτες για όλες τις χρονιές που συμμετείχαν στην Euroleague εκτός από τη χρονιά 2023/2024, ένα που έχει όλα τα δεδομένα για τους παίκτες μόνο για τη χρονιά 2022/2023, ένα που έχει όλα τα δεδομένα για τους παίκτες μόνο για την τελευταία τους αγωνι-

στική στην Euroleague και ένα που έχει όλα τα δεδομένα για τους παίκτες μόνο για τη χρονιά 2023/2024 εκτός από την τελευταία τους αγωνιστική στην Euroleague. Έπειτα, με την βοήθεια των πακέτων που προσφέρει το Scikit-learn, εκπαιδεύσαμε τα μοντέλα μας χρησιμοποιώντας ως δεδομένα εκπαίδευσης τα εξής: Λεπτά Συμμετοχής (MP), Πόντοι (PTS), Ποσοστό Εύστοχων Σουτ (FG%), Αληθινό Ποσοστό Σουτ (TS%), Ποσοστό Πραγματικής Ευστοχίας Εντός Πεδιάς (EFG%), Ποσοστό Εύστοχων Δίποντων (2P%), Εύστοχα Τρίποντα (3PM), Ποσοστό Εύστοχων Τριπόντων (3PM%), Ποσοστό Τρίποντων που Αποπειράθηκαν (3PA%), Ποσοστό Εύστοχων Ελεύθερων Βολών (FT%), Ασίστ (AST), Ποσοστό Ασίστ (AST%), Μπλοκ (BLK), Ποσοστό Μπλοκ (BLK%), Μπλοκ που Δέχθηκε (RBLK), Αμυντικά Ριμπάουντ (DRB), Ποσοστό Αμυντικών Ριμπάουντ (DRB%), Επιθετικά Ριμπάουντ (ORB), Κλεψίματα (STL), Ποσοστό Κλεψίματων (STL%), Ποσοστό Λαθών (TOV%), Προσωπικά Φάουλ (PF), Κερδισμένα Φάουλ (RF) και Ποσοστό Χρήσης Παίκτη (USG%), ενώ ως δεδομένο πρόβλεψης χρησιμοποιήθηκε η αποδοτικότητα (EFF). Οι αλγόριθμοι πρόβλεψης που χρησιμοποιήθηκαν είναι οι εξής: Γραμμική Παλινδρόμηση, Διαβαθμιζόμενη Ενίσχυση, Γραμμική Μηχανή Διανυσμάτων Στήριξης, Τυχαίο Δάσος, Νευρωνικά Δίκτυα (Multilayer Perceptron), Χειριστής Ελάχιστης Απόλυτης Συρρίκνωσης και Επιλογής (Lasso), Παλινδρόμηση Κορυφογραμμής (Ridge) και Δέντρα Αποφάσεων. Για την αξιολόγηση των μοντέλων κάναμε χρήση της Διασταυρούμενης Επικύρωσης με δέκα πλαίσια και ως μετρική αξιολόγησης το μέσο απόλυτο σφάλμα (mean absolute error) και το μέσο τετραγωνικό σφάλμα (mean squared error). Τέλος, για την επιλογή των δέκα παικτών για το διαδικτυακό παιχνίδι Fantasy Euroleague κάναμε χρήση του γραμμικού προγραμματισμού με στόχο τη μεγιστοποίηση της αποδοτικότητας με περιορισμούς οι αξία της ομάδας να μην ξεπερνάει το εκατό δεκαπέντε και να επιλεγούν αυστηρά δύο παίκτες που η θέση τους είναι Center, τέσσερεις παίκτες που η θέση τους είναι Guard και τέσσερεις παίκτες που η θέση τους είναι Forward.

Για το πρώτο πείραμα που διεξήγαμε χρησιμοποιήσαμε το πλαίσιο δεδομένων που έχει όλα τα δεδομένα για τους παίκτες για όλες τις χρονιές που συμμετείχαν στην Euroleague και χρησιμοποιήσαμε το 80% για την εκπαίδευση του μοντέλου, ενώ το υπολειπόμενο 20% το χρησιμοποιήσαμε για την αξιολόγηση του.

Για το δεύτερο πείραμα που διεξήγαμε χρησιμοποιήσαμε το πλαίσιο δεδομένων

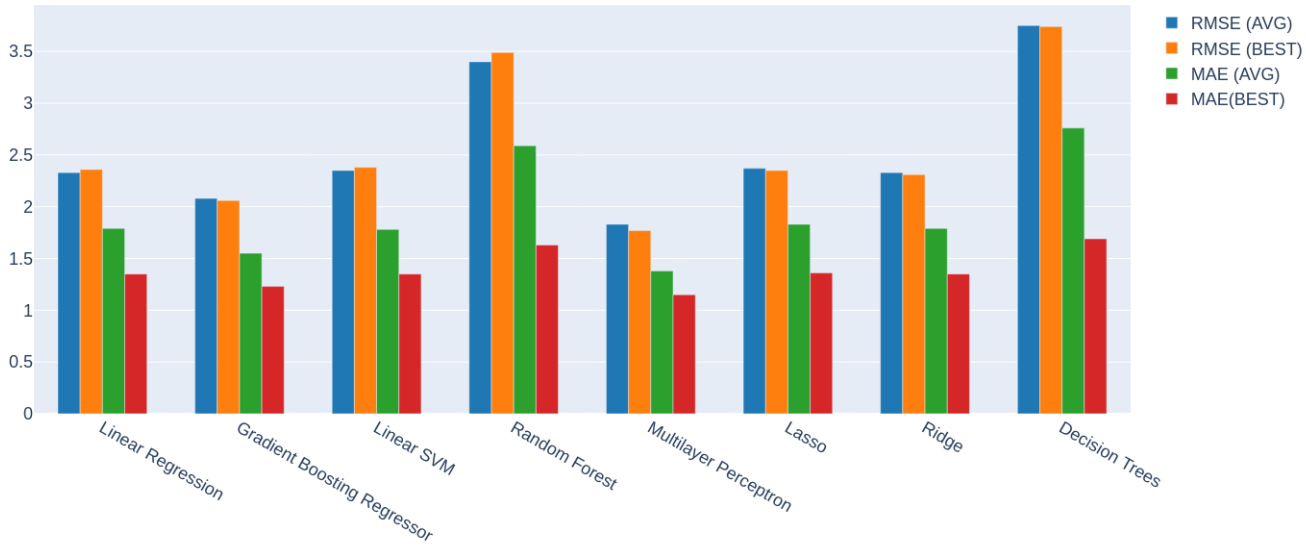
που έχει όλα τα δεδομένα για τους παίκτες για όλες τις χρονιές που συμμετείχαν στην Euroleague εκτός από τη χρονιά 2023/2024 για την εκπαίδευση του μοντέλου, ενώ για την αξιολόγηση χρησιμοποιήσαμε το πλαίσιο δεδομένων που έχει τα δεδομένα για τους παίκτες μόνο για τη χρονιά 2023/2024.

Για το τρίτο πείραμα που διεξήγαμε χρησιμοποιήσαμε το πλαίσιο δεδομένων που έχει όλα τα δεδομένα για τους παίκτες μόνο για τη χρονιά 2022/2023 για την εκπαίδευση του μοντέλου, ενώ για την αξιολόγηση χρησιμοποιήσαμε το πλαίσιο δεδομένων που έχει τα δεδομένα για τους παίκτες μόνο για τη χρονιά 2023/2024.

Για το τέταρτο πείραμα που διεξήγαμε χρησιμοποιήσαμε το πλαίσιο δεδομένων που έχει όλα τα δεδομένα για τους παίκτες μόνο για τη χρονιά 2023/2024 εκτός από την τελευταία τους αγωνιστική στην Euroleague για την εκπαίδευση του μοντέλου, ενώ για την αξιολόγηση χρησιμοποιήσαμε το πλαίσιο δεδομένων που έχει όλα τα δεδομένα για τους παίκτες μόνο για την τελευταία τους αγωνιστική.

5.1 Αποτελέσματα

Σχήμα 5.1: 80-20 split (όσο μικρότερο τόσο καλύτερα αποτελέσματα)



Το Σχήμα 5.1 απεικονίζει το πείραμα που διεξήγαμε για την πρόβλεψη της απόδοσης των παικτών χρησιμοποιώντας όλα τα διαθέσιμα δεδομένα των παικτών για όλες τις χρονιές που συμμετείχαν στην Euroleague και χρησιμοποιήσαμε το 80%

για την εκπαίδευση του μοντέλου, ενώ το υπολειπόμενο 20% το χρησιμοποιήσαμε για την αξιολόγηση του. Παρατηρούμε ότι τα νευρωνικά δίκτυα και συγκεκριμένα ο Multilayer Perceptron, προσφέρει τα καλύτερα αποτελέσματα, με μέσο τετραγωνικό σφάλμα ίσο με 1.83 και μέσο απόλυτο σφάλμα ίσο με 1.38, ενώ τα αμέσως καλύτερα αποτελέσματα τα δίνει η διαβαθμιζόμενη ενίσχυση (Gradient Boosting) με 2.08 και 1.55 αντίστοιχα. Παρατηρούμε επίσης ότι όλοι οι αλγόριθμοι που χρησιμοποιήθηκαν δίνουν αρκετά καλά αποτελέσματα και είναι πολύ κοντά μεταξύ τους, με εξαίρεση το τυχαίο δάσος και τα δέντρα αποφάσεων που συγκριτικά με τους υπολοίπους αλγορίθμους, βρίσκονται αρκετά εκτός των επιθυμητών αποτελεσμάτων. Ένας περιορισμός που μπορεί να επηρεάζει τα αποτελέσματα είναι το γεγονός πως κάποιοι παίκτες έχουν συμμετάσχει σε πολύ λίγους αγώνες ή η χρονιά 2023/2024 είναι η πρώτη τους χρονιά στην Euroleague και κατά τον διαχωρισμό των δεδομένων μπορεί τα δεδομένα των συγκεκριμένων παικτών να μην είναι επαρκή ή να μην υπάρχουν και καθόλου στην εκπαίδευση του μοντέλου.

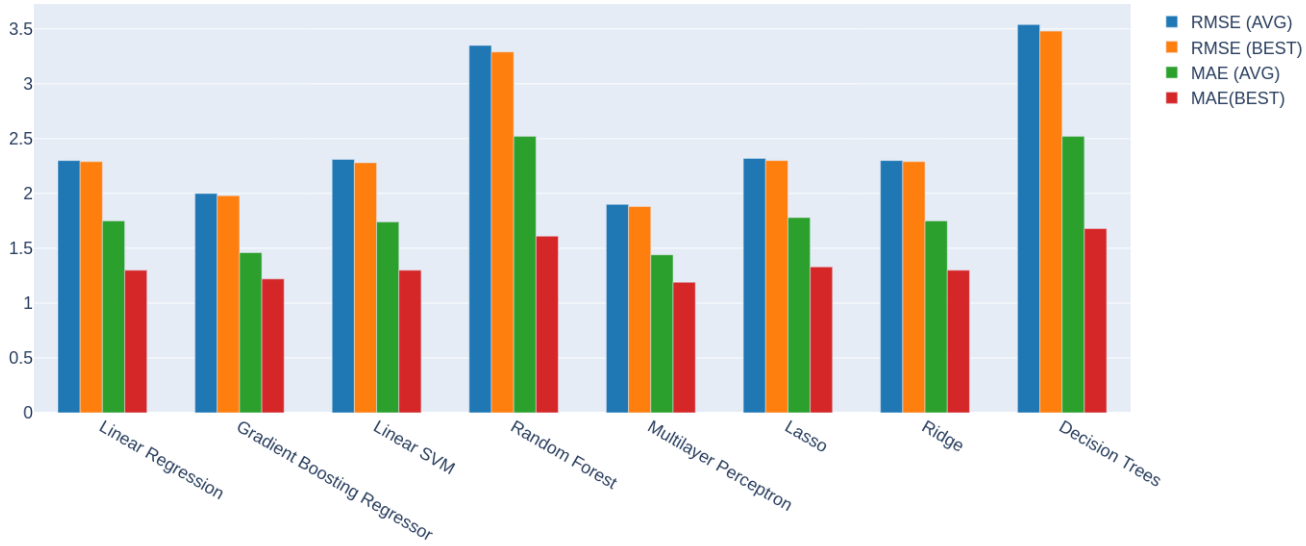
Πίνακας 5.1: Επιλογή ομάδας μοντέλου 80-20

Name	**Position**	**Price**	**Predicted EFF**	**Total Points**
Tibor Pleiss	C	7.3	34.351605	194.00
Brandon Davies	C	11.5	36.041970	357.70
Bonzie Colson	F	13.2	36.783651	601.80
Konstantinos Mitoglou	F	10.8	38.382628	548.80
Achille Polonara	F	6.6	35.519613	115.50
Nikola Mirotic	F	15.7	38.153593	413.00
Shane Larkin	G	17.8	40.443593	735.30
Nick Calathes	G	10.5	36.650189	428.00
Nemanja Nedovic	G	12.3	37.827331	404.80
Kevin Pangos	G	5.3	39.912778	55.90

Ο Πίνακας 5.1 απεικονίζει τους δέκα παίκτες που επιλέχθηκαν με βέλτιστο τρόπο για τη μεγιστοποίηση της συνολικής αποδοτικότητας των παικτών με τη χρήση γραμμικού προγραμματισμού με περιορισμούς. Η προβλεπόμενη αποτελεσματικότητα των παικτών από το μοντέλο του πειράματος χρησιμοποιήθηκε για την επιλογή τους, έχοντας και ως περιορισμό η συνολική αξία της ομάδας να μην ξεπερνάει το 115. Οι συνολικοί πόντοι που αναγράφονται στον Πίνακα 5.1, είναι οι συνολικοί πόντοι που συγκέντρωσαν οι επιλεχθέντες παίκτες για τη χρονιά 2023/2024 στο διαδικτυακό παιχνίδι Fantasy Euroleague οι οποίοι πάρθηκαν από την επίσημη

ιστοσελίδα της Euroleague. Με την προϋπόθεση ότι δεν θα αλλάξει καθ' όλη τη διάρκεια της χρονιάς, οι συνολικοί πόντοι που συγκεντρώθηκαν από την επιλεγμένη ομάδα ισούνται με 3854.80 και συνολική αξία 111, πράγμα που σημαίνει ότι δεν αξιοποιήθηκε όλος ο διαθέσιμος προϋπολογισμός των 115 που προσφέρει το Fantasy Euroleague. Με το περισσευούμενο μπάτζετ θα μπορούσε να γίνει μια καλύτερη επιλογή στους Guard και αντί του Kevin Pangos να διαλέγαμε τον Thomas Walkup που θα απέφερε 396.30 πόντους, δηλαδή 340.40 πόντους παραπάνω. Συγκρίνοντας τους συγκεντρωμένους πόντους από την ομάδα που επιλέχθηκε με τους συνολικούς πόντους του νικητή του Fantasy Euroleague για τη χρονιά 2023/2024 που συγκέντρωσε 7633 πόντους, παρατηρούμε ότι είμαστε αρκετά μακριά από τις πρώτες θέσεις. Οφείλουμε να λάβουμε υπόψιν ότι προβλέψαμε μια ομάδα για όλη τη χρονιά χωρίς αλλαγές, ένας περιορισμός ο οποίος προφανώς επηρεάζει αρνητικά την προσπάθεια μας να μεγιστοποιήσουμε τη συγκέντρωση βαθμών. Με την προϋπόθεση ότι δεν θα αλλάξει καθ' όλη τη διάρκεια της χρονιάς, οι συνολικοί πόντοι που συγκεντρώθηκαν από την επιλεγμένη ομάδα ισούνται με 3854.80 και συνολική αξία 111. Συγκρίνοντας τους συγκεντρωμένους πόντους από την ομάδα που επιλέχθηκε με τους συνολικούς πόντους του νικητή του Fantasy Euroleague για τη χρονιά 2023/2024 που συγκέντρωσε 7633 πόντους, παρατηρούμε ότι είμαστε αρκετά μακριά από τις πρώτες θέσεις. Οφείλουμε να λάβουμε υπόψιν ότι προβλέψαμε μια ομάδα για όλη τη χρονιά χωρίς αλλαγές.

Σχήμα 5.2: ALL-23/24 (όσο μικρότερο τόσο καλύτερα αποτελέσματα)



Το Σχήμα 5.2 απεικονίζει το πείραμα που διεξήγαμε για την πρόβλεψη της απόδοσης των παικτών για τη χρονιά 2023/2024, χρησιμοποιώντας τα δεδομένα των παικτών από όλες τις χρονιές που συμμετείχαν στην Euroleague για την εκπαίδευση του μοντέλου και τη χρονιά 2023/2024 για την αξιολόγηση του μοντέλου. Παρατηρούμε ότι ο Multilayer Perceptron, προσφέρει τα καλύτερα αποτελέσματα, με μέσο τετραγωνικό σφάλμα ίσο με 1.87 και μέσο απόλυτο σφάλμα ίσο με 1.41, ενώ τα αμέσως καλύτερα αποτελέσματα τα δίνει η διαβαθμιζόμενη ενίσχυση (Gradient Boosting) με 1.96 και 1.44 αντίστοιχα. Οι υπόλοιποι αλγόριθμοι κυμαίνονται στο 1.90-2.00 για το μέσο τετραγωνικό σφάλμα και 2.30-2.40 για το μέσο απόλυτο σφάλμα, με εξαίρεση το τυχαίο δάσος και τα δέντρα αποφάσεων που για ακόμα ένα πείραμα δίνουν με διαφορά τα χειρότερα αποτελέσματα με το μέσο τετραγωνικό σφάλμα να είναι ίσο με 3.40 και 3.57 και το μέσο απόλυτο σφάλμα να είναι ίσο με 2.51 και 2.50. Ένας περιορισμός που πιθανόν να επηρεάζει τα αποτελέσματα είναι η πιθανότητα ένας παίκτης να μεταγράφηκε τη χρονιά 2023/2024 ή στα μέσα της χρονιάς, με αποτέλεσμα τα δεδομένα των συγκεκριμένων παικτών να μην είναι επαρκή ή να μην υπάρχουν και καθόλου στην εκπαίδευση του μοντέλου.

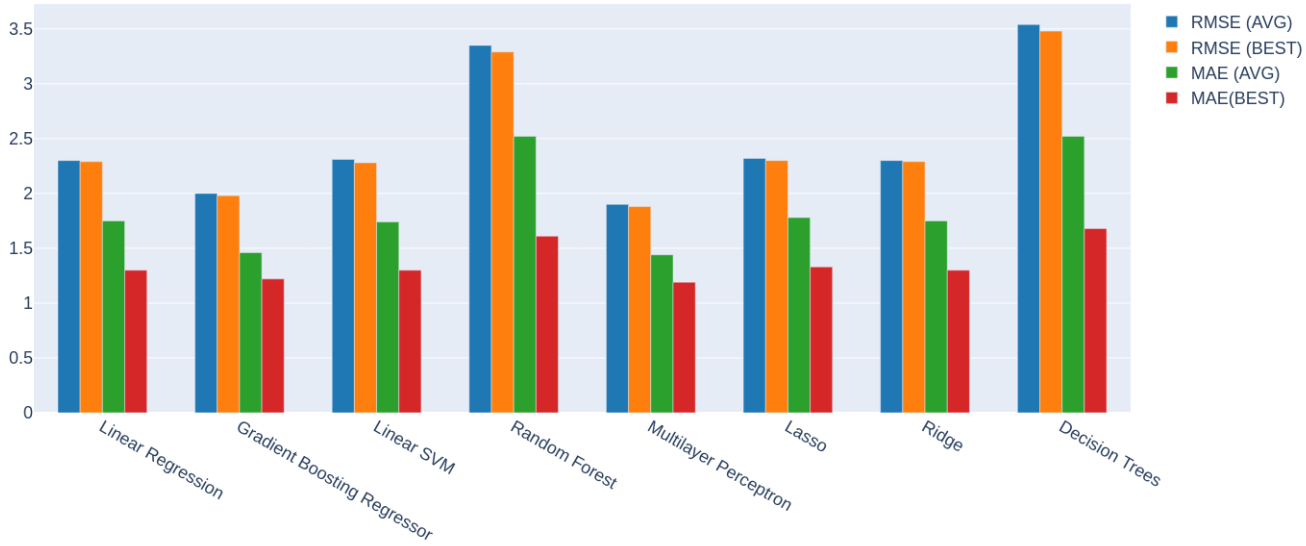
Πίνακας 5.2: Επιλογή ομάδας μοντέλου ALL-23/24

Name	Position	Price	Predicted EFF	Total Points
Joffrey Lauvergne	C	11.0	13.927813	414.00
Johannes Thiemann	C	12.5	15.515147	485.90
Chima Moneke	F	14.8	17.609621	661.80
Konstantinos Mitoglou	F	10.8	13.867940	548.80
Alec Peters	F	12.0	14.771176	612.60
Louis Olinde	F	7.4	10.620621	186.70
Markus Howard	G	10.7	13.412159	457.60
Nicolas Laprovittola	G	10.3	12.164187	421.00
Mike James	G	17.1	19.753527	811.10
Marco Belinelli	G	8.1	10.412891	340.90

Ο Πίνακας 5.2 απεικονίζει τους δέκα παίκτες που επιλέχθηκαν με σκοπό τη μεγιστοποίηση της συνολικής αποδοτικότητας των παικτών κάνοντας χρήση γραμμικού προγραμματισμού με περιορισμούς. Η προβλεπόμενη αποτελεσματικότητα των παικτών από το μοντέλο του πειράματος χρησιμοποιήθηκε για την επιλογή τους, έχοντας και ως περιορισμό η συνολική αξία της ομάδας να μην ξεπερνάει το 115. Οι συνολικοί πόντοι που αναγράφονται στον Πίνακα 5.2, είναι οι συνολικοί πόντοι που συγκέντρωσαν οι επιλεχθέντες παίκτες για τη χρονιά 2023/2024 στο διαδικτυακό παιχνίδι Fantasy Euroleague οι οποίοι πάρθηκαν από την επίσημη ιστοσελίδα της Euroleague. Με την προϋπόθεση ότι δεν θα αλλάξει καθ' όλη τη διάρκεια της χρονιάς, οι συνολικοί πόντοι που συγκεντρώθηκαν από την επιλεγμένη ομάδα ισούνται με 4940.4 και συνολική αξία 114.7. Με το υπολειπόμενο μπάτζετ να ισούται με 0.3, δεν μας δίνονται πολλά περιθώρια για βελτίωση της ομάδας, ωστόσο μια αλλαγή που θα μπορούσαμε να κάνουμε είναι αυτή του Louis Olinde που απέφερε μόλις 186.70 πόντους, με τον Marius Grigonis του Παναθηναϊκού που συγκέντρωσε 384.90 πόντους με αξία 7.7. Με αυτή την αλλαγή θα συγκεντρώναμε 198.20 παραπάνω πόντους και θα χρησιμοποιούσαμε όλο τον διαθέσιμο προϋπολογισμό που προσφέρει το Fantasy Euroleague. Συγκρίνοντας τους συγκεντρωμένους πόντους από την ομάδα που επιλέχθηκε με τους συνολικούς πόντους του νικητή του Fantasy Euroleague για τη χρονιά 2023/2024 που συγκέντρωσε 7633 πόντους, παρατηρούμε ότι είμαστε αρκετά μακριά από τις πρώτες θέσεις, ωστόσο σε σύγκριση με το πρώτο πείραμα που διεξήγαμε καταφέραμε να συγκεντρώσουμε 1085.6 περισσότερους πόντους, κάτι που περιμέναμε λόγο των περιορισμών του πρώτου πει-

ράματος. Επίσης, οφείλουμε να λάβουμε υπόψιν ότι προβλέψαμε μια ομάδα για όλη τη χρονιά χωρίς αλλαγές, ένας περιορισμός ο οποίος προφανώς επηρεάζει αρνητικά την προσπάθεια μας να μεγιστοποιήσουμε τη συγκέντρωση βαθμών.

Σχήμα 5.3: 22/23-23/24 (όσο μικρότερο τόσο καλύτερα αποτελέσματα)



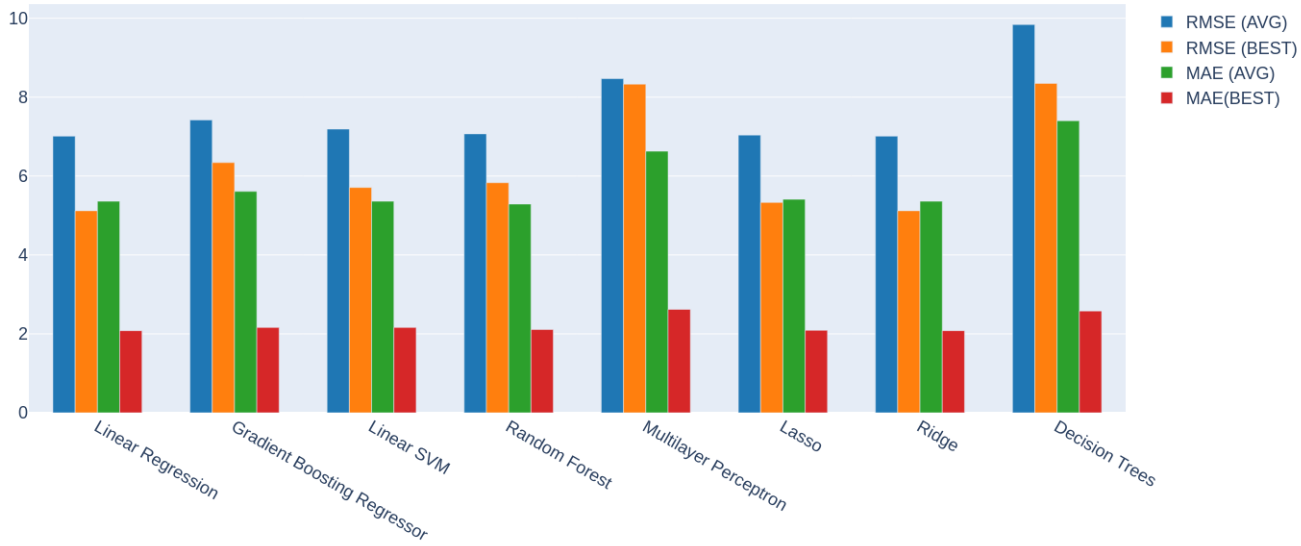
Το Σχήμα 5.3 απεικονίζει το πείραμα που διεξήγαμε για την πρόβλεψη της απόδοσης των παικτών για τη χρονιά 2023/2024, χρησιμοποιώντας τα δεδομένα των παικτών από τη χρονιά 2022/2023 για την εκπαίδευση του μοντέλου και τη χρονιά 2023/2024 για την αξιολόγηση του μοντέλου. Παρατηρούμε ότι για ακόμη μια φορά ο Multilayer Perceptron, προσφέρει τα καλύτερα αποτελέσματα, με μέσο τετραγωνικό σφάλμα ίσο με 1.90 και μέσο απόλυτο σφάλμα ίσο με 1.44, ενώ τα αμέσως καλύτερα αποτελέσματα τα δίνει η διαβαθμιζόμενη ενίσχυση (Gradient Boosting) με 2.00 και 1.46 αντίστοιχα. Οι υπόλοιποι αλγόριθμοι κυμαίνονται στο 1.90-2.00 για το μέσο τετραγωνικό σφάλμα και 2.30-2.40 για το μέσο απόλυτο σφάλμα, με εξαίρεση το τυχαίο δάσος και τα δέντρα αποφάσεων που για ακόμα ένα πείραμα δίνουν με διαφορά τα χειρότερα αποτελέσματα με το μέσο τετραγωνικό σφάλμα να είναι ίσο με 3.35 και 3.59 και το μέσο απόλυτο σφάλμα να είναι ίσο με 2.52 και 2.50. Ένας περιορισμός που πιθανόν να επηρεάζει τα αποτελέσματα είναι η πιθανότητα ένας παίκτης να μην πάρει συμμετοχές στην Euroleague τη χρονιά 2022/2023 ή να μεταγράφηκε τη χρονιά 2023/2024, με αποτέλεσμα αυτοί οι παίκτες να μην υπάρχουν καθόλου στην εκπαίδευση του μοντέλου.

Πίνακας 5.3: Επιλογή ομάδας μοντέλου 22/23-23/24

Name	Position	Price	Predicted EFF	Total Points
Joffrey Lauvergne	C	11.0	13.927813	414.00
Johannes Thiemann	C	12.5	15.515147	485.90
Chima Moneke	F	14.8	17.609621	661.80
Konstantinos Mitoglou	F	10.8	13.867940	548.80
Alec Peters	F	12.0	14.771176	612.60
Louis Olinde	F	7.4	10.620621	186.70
Markus Howard	G	10.7	13.412159	457.60
Nicolas Laprovittola	G	10.3	12.164187	421.00
Mike James	G	17.1	19.753527	811.10
Marco Belinelli	G	8.1	10.412891	340.90

Ο Πίνακας 5.3 απεικονίζει τους δέκα παίκτες που επιλέχθηκαν με σκοπό τη μεγιστοποίηση της συνολικής αποδοτικότητας των παικτών κάνοντας χρήση γραμμικού προγραμματισμού με περιορισμούς. Η προβλεπόμενη αποτελεσματικότητα των παικτών από το μοντέλο του πειράματος χρησιμοποιήθηκε για την επιλογή τους, έχοντας και ως περιορισμό η συνολική αξία της ομάδας να μην ξεπερνάει το 115. Οι συνολικοί πόντοι που αναγράφονται στον Πίνακα 5.3, είναι οι συνολικοί πόντοι που συγκέντρωσαν οι επιλεχθέντες παίκτες για τη χρονιά 2023/2024 στο διαδικτυακό παιχνίδι Fantasy Euroleague οι οποίοι πάρθηκαν από την επίσημη ιστοσελίδα της Euroleague. Με την προϋπόθεση ότι δεν θα αλλάξει καθ' όλη τη διάρκεια της χρονιάς, οι συνολικοί πόντοι που συγκεντρώθηκαν από την επιλεγμένη ομάδα ισούνται με 4940.4 και συνολική αξία 114.7. Συγκρίνοντας τους συγκεντρωμένους πόντους από την ομάδα που επιλέχθηκε με τους συνολικούς πόντους του νικητή του Fantasy Euroleague για τη χρονιά 2023/2024 που συγκέντρωσε 7633 πόντους, παρατηρούμε ότι είμαστε αρκετά μακριά από τις πρώτες θέσεις, ωστόσο σε σύγκριση με το πρώτο πείραμα που διεξήγαμε καταφέραμε να συγκεντρώσουμε 1085.6 περισσότερους πόντους, κάτι που περιμέναμε λόγο των περιορισμών του πρώτου πειράματος. Επίσης, οφείλουμε να λάβουμε υπόψιν ότι προβλέψαμε μια ομάδα για όλη τη χρονιά χωρίς αλλαγές, ένας περιορισμός ο οποίος προφανώς επηρεάζει αρνητικά την προσπάθεια μας να μεγιστοποιήσουμε τη συγκέντρωση βαθμών. Παρατηρούμε επίσης, ότι η ομάδα που προβλέψαμε είναι ίδια με αυτή του δεύτερου πειράματος, αυτό προκύπτει από το γεγονός ότι το μέσο τετραγωνικό σφάλμα και το μέσο απόλυτο σφάλμα των δύο πειραμάτων είναι πολύ κοντά μεταξύ τους.

Σχήμα 5.4: Τελευταία αγωνιστική (όσο μικρότερο τόσο καλύτερα αποτελέσματα)



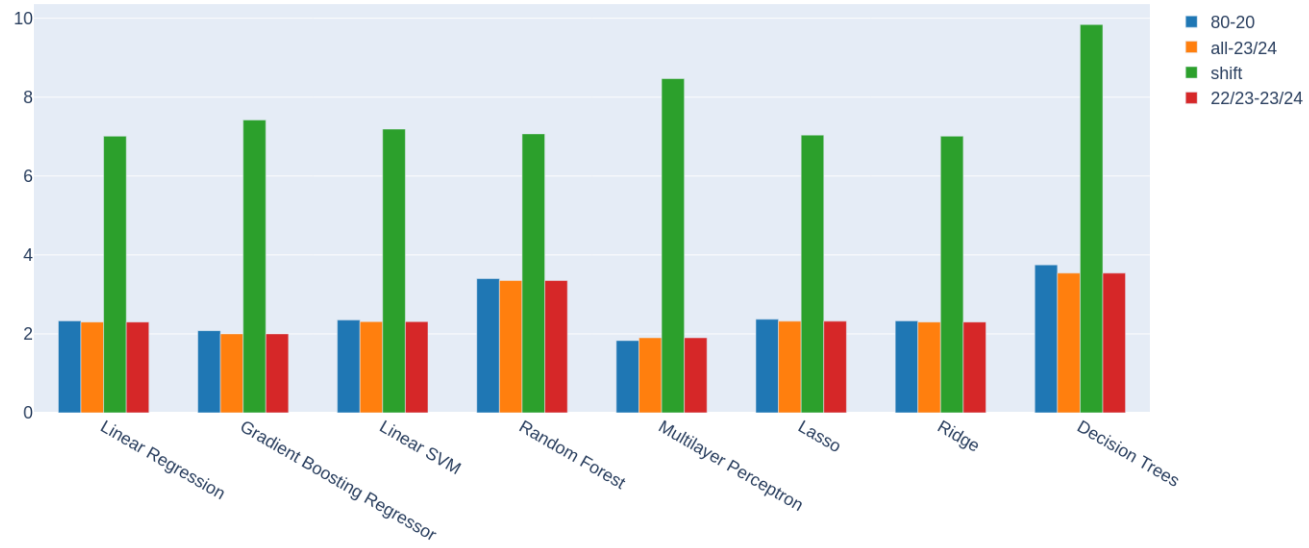
Το Σχήμα 5.4 απεικονίζει το πείραμα που διεξήγαμε για την πρόβλεψη της απόδοσης των παικτών για την τελευταία τους αγωνιστική στην Euroleague, χρησιμοποιώντας όλα τα δεδομένα για τη χρονιά 2023/2024 εκτός από την τελευταία αγωνιστική για την εκπαίδευση του μοντέλου και την τελευταία αγωνιστική για την αξιολόγηση του μοντέλου. Παρατηρούμε ότι η χρήση της γραμμικής παλινδρόμησης μας δίνει το καλύτερο μέσο τετραγωνικό σφάλμα ίσο με 7.01, ενώ το τυχαίο δάσος μας δίνει το καλύτερο μέσο απόλυτο σφάλμα ίσο με 5.29. Παρατηρούμε επίσης ότι ο Multilayer Perceptron και τα δέντρα αποφάσεων μας δίνουν με διαφορά τα χειρότερα αποτελέσματα με το μέσο τετραγωνικό σφάλμα ίσο με 8.47 και 9.84 αντίστοιχα, καθώς και μέσο απόλυτο σφάλμα ίσο με 6.63 και 7.40 αντίστοιχα. Εξαιτίας των ελλιπών δεδομένων, το μοντέλο δεν εκπαιδεύτηκε σωστά, αυτό γίνεται εμφανές από τη χαμηλή ευστοχία κατά την εκπαίδευση και πόσο μάλλον από την ακόμα χαμηλότερη γενίκευση, αυτό είναι γνωστό και ως φαινόμενο υποπροσαρμογής (underfitting).

Πίνακας 5.4: Επιλογή ομάδας μοντέλου Shift

Name	Position	Price	Predicted EFF	Total Points
Mathias Lessort	C	18.8	14.805906	24.2
Joel Bolomboy	C	12.6	15.160535	10.0
Deshaun Thomas	F	9.0	14.614326	4.4
Timothe Luwawu-Cabarrot	F	9.6	14.306993	4.4
Nikola Mirotic	F	15.7	15.033494	9.0
Edgaras Ulanovas	F	10.2	13.993053	4.0
Lorenzo Brown	G	12.7	14.448255	24.0
Tamir Blatt	G	7.5	13.576383	5.0
Shaquielle McKissic	G	7.9	13.710501	13.2
P.J. Dozier	G	9.9	14.383299	17.6

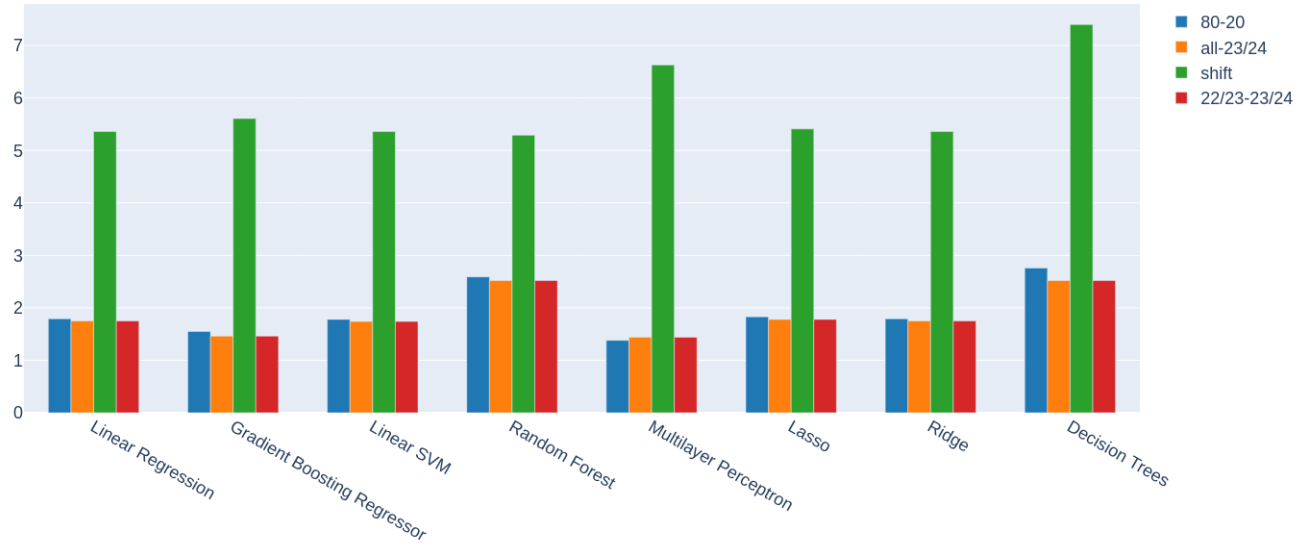
Ο Πίνακας 5.4 απεικονίζει τους δέκα παίκτες που επιλέχθηκαν με βέλτιστο τρόπο για τη μεγιστοποίηση της συνολικής αποδοτικότητας των παικτών με τη χρήση γραμμικού προγραμματισμού με περιορισμούς. Η προβλεπόμενη αποτελεσματικότητα των παικτών από το μοντέλο του πειράματος χρησιμοποιήθηκε για την επιλογή τους, έχοντας και ως περιορισμό η συνολική αξία της ομάδας να μην ξεπερνάει το 115. Οι συνολικοί πόντοι που αναγράφονται στον Πίνακα 5.4, είναι οι συνολικοί πόντοι που συγκέντρωσαν οι επιλεχθέντες παίκτες για τη χρονιά 2023/2024 στο διαδικτυακό παιχνίδι Fantasy Euroleague οι οποίοι πάρθηκαν από την επίσημη ιστοσελίδα της Euroleague. Οι επιλογή των παικτών έγινε για τον τελευταίο αγώνα στον οποίο συμμετείχε ο κάθε παίκτης, οι συνολικοί πόντοι που συγκεντρώθηκαν από την επιλεγμένη ομάδα ισούνται με 115.8 και συνολική αξία 113.9. Συγκρίνοντας τους συγκεντρωμένους πόντους από την ομάδα που επιλέχθηκε με την θεωρητικά καλύτερη ομάδα της αγωνιστικής που συγκέντρωσε 212 πόντους, παρατηρούμε ότι είμαστε μόλις 100 πόντους μακριά. Ωστόσο, η θεωρητικά καλύτερη ομάδα της αγωνιστικής ξεπερνάει κατά πολύ το όριο των 115 του διαθέσιμου μπάτζετ που έχουμε. Παρ' όλ' αυτά, με μια καλύτερη διαχείριση του μπάτζετ θα μπορούσαν να γίνουν καλύτερες επιλογές που θα μας επέφεραν περισσότερους πόντους, φέρνοντάς μας πιο κοντά στην θεωρητικά καλύτερη ομάδα. Αλλάζοντας τους Edgaras Ulanovas και Timothe Luwawu-Cabarrot με τους Dyshawn Pierre και Shaquielle Mckissic θα είχαμε αύξηση στους πόντους κατά 27.4, φέρνοντας μας στους 143.2 πόντους, και μείωση στη συνολική αξία της ομάδας κατά 5.2, δίνοντας μας έτσι τη δυνατότητα για επιπλέον βελτίωση της ομάδας.

Σχήμα 5.5: Μέσο τετραγωνικό σφάλμα μοντέλων (όσο μικρότερο τόσο καλύτερα αποτελέσματα)



Το Σχήμα 5.5 απεικονίζει το μέσο τετραγωνικό σφάλμα όλων των μοντέλων μαζί. Όπως παρατηρούμε τα αποτελέσματα των μοντέλων για τους επιλεγμένους αλγορίθμους είναι πολύ κοντά, με εξαίρεση αυτά του τέταρτου μοντέλου (SHIFT) που δίνει με διαφορά τα χειρότερα αποτελέσματα για όλους τους αλγορίθμους που χρησιμοποιήθηκαν με το χειρότερο μέσο τετραγωνικό σφάλμα να δίνουν τα δέντρα αποφάσεων και να ισούται με 9.84. Ωστόσο το περιμέναμε, διότι το τέταρτο μοντέλο δέχεται με διαφορά τα λιγότερα δεδομένα στο στάδιο της εκπαίδευσης και του ζητείται να προβλέψει έναν αγώνα για τον κάθε παίκτη. Το καλύτερο μέσο τετραγωνικό σφάλμα το προσφέρει ο Multilayer Perceptron του πρώτου μοντέλου (80-20) που ισούται με 1.83, αυτό οφείλεται στον μεγάλο όγκο δεδομένων που είχε στη διάθεση του σε σχέση με τα υπόλοιπα μοντέλα. Παρατηρούμε επίσης, ότι τα δέντρα αποφάσεων δίνουν με διαφορά το χειρότερο μέσο τετραγωνικό σφάλμα για όλα τα μοντέλα, ενώ ο Multilayer Perceptron το καλύτερο, με εξαίρεση το τέταρτο μοντέλο που τα καλύτερα αποτελέσματα τα παίρνουμε από τη γραμμική παλινδρόμηση.

Σχήμα 5.6: Μέσο απόλυτο σφάλμα μοντέλων (όσο μικρότερο τόσο καλύτερα αποτελέσματα)



Το Σχήμα 5.6 απεικονίζει το μέσο απόλυτο σφάλμα όλων των μοντέλων μαζί. Όπως παρατηρούμε τα αποτελέσματα των μοντέλων για τους επιλεγμένους αλγορίθμους είναι πολύ κοντά, με εξαίρεση για ακόμα μια φορά, αυτά του τέταρτου μοντέλου (SHIFT) που όπως και για το μέσο τετραγωνικό σφάλμα, δίνει με διαφορά τα χειρότερα αποτελέσματα για όλους τους αλγορίθμους που χρησιμοποιήθηκαν με το χειρότερο μέσο απόλυτο σφάλμα να το δίνουν τα δέντρα αποφάσεων και να ισούται με 7.40. Το καλύτερο μέσο απόλυτο σφάλμα το προσφέρει ο Multilayer Perceptron του πρώτου μοντέλου (80-20) που ισούται με 1.38. Παρατηρούμε επίσης, ότι τα δέντρα αποφάσεων δίνουν με διαφορά το χειρότερο μέσο τετραγωνικό σφάλμα για όλα τα μοντέλα, ενώ ο Multilayer Perceptron τα καλύτερα, με εξαίρεση το τέταρτο μοντέλο που τα καλύτερα αποτελέσματα τα παίρνουμε από τη γραμμική παλινδρόμηση.

Κεφάλαιο 6

Συμπεράσματα

Η μηχανική μάθηση μπορεί να έχει βαθύ αντίκτυπο στα sports analytics με την ικανότητά της να προβλέπει τα μελλοντικά αποτελέσματα αγώνων μπάσκετ καθώς και τη μελλοντική απόδοση των παικτών, αυτό μπορεί να φανεί μέσα από τις μελέτες που έχουν διεξαχθεί διαχρονικά, μια εκ των οποίων είναι και η δική μας μελέτη, με αποτελέσματα ίσα με 1.83 για το μέσο τετραγωνικό σφάλμα και 1.38 για το μέσο απόλυτο σφάλμα με τη χρήση νευρωνικών δικτύων, για την πρόβλεψη της αποδοτικότητας των παικτών για τη χρονιά 2023/2024 της Euroleague. Διεξήγαμε τέσσερα πειράματα, εκ των οποίων, το πρώτο είχε μέσο τετραγωνικό σφάλμα ίσο με 1.83, μέσο απόλυτο σφάλμα ίσο με 1.38 και η ομάδα που η επιλέχθηκε συγκέντρωσε 3854.80 πόντους, το δεύτερο είχε μέσο τετραγωνικό σφάλμα ίσο με 1.87, μέσο απόλυτο σφάλμα ίσο με 1.41 και η ομάδα που η επιλέχθηκε συγκέντρωσε 4940.4 πόντους, το τρίτο είχε μέσο τετραγωνικό σφάλμα ίσο με 1.90, μέσο απόλυτο σφάλμα ίσο με 1.44 και η ομάδα που η επιλέχθηκε επίσης συγκέντρωσε 4940.4 πόντους και το τέταρτο είχε μέσο τετραγωνικό σφάλμα ίσο με 7.01, μέσο απόλυτο σφάλμα ίσο με 5.29 και η ομάδα που η επιλέχθηκε συγκέντρωσε 115.8 πόντους. Επιπλέον, συγκρίνοντας τα αποτελέσματα της μελέτης μας, με αυτά της μελέτης στην οποία βασιστήκαμε [7], καταφέραμε να έχουμε βελτίωση κατά 16.7% για το μέσο τετραγωνικό σφάλμα και 16.2% για το μέσο απόλυτο σφάλμα. Τα αποτελέσματα θα μπορούσαν να ήταν ακόμα καλύτερα αν δεν υπήρχαν παίκτες που η χρονιά 2023/2024 ήταν η πρώτη τους χρονιά στην Euroleague, με αποτέλεσμα τα δεδομένα που υπάρχουν γι' αυτούς να είναι ελλιπή. Η επιλογή της ομάδας για το διαδικτυακό παιχνίδι Fantasy Euroleague έγινε με τη χρήση γραμμικού προγραμματισμού, αξιοποιώντας τις προβλεπόμενες αποδόσεις των παικτών από τα μοντέλα

που αναπτύχθηκαν από τη μελέτη μας και καταφέραμε να συγκεντρώσουμε στην καλύτερη 4940.4 πόντους για όλη τη χρονιά και 115.8 για μια αγωνιστική. Επειδή η μελέτη μας χρησιμοποίησε εξ' ολοκλήρου στατιστικά του μπάσκετ για την ανάπτυξη των μοντέλων, πιθανώς παραμελεί την κρίσιμη επίδραση εξωτερικών παραγόντων στην απόδοση των παικτών. Έτσι, περαιτέρω μελέτες σε αυτόν τον τομέα που θα βελτιώσουν την ικανότητα πρόβλεψης και θα παρέχουν πιο ολοκληρωμένη κατανόηση του βαθμού σημασίας των διαφορετικών παραγόντων.

Βιβλιογραφία

[1] Konstantinos Apostolou and Christos Tjortjis. Sports analytics algorithms for performance prediction. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–4. IEEE, 2019.

[2] Serkan Ballı and Engin Özdemir. A novel method for prediction of euroleague game results using hybrid feature extraction and machine learning techniques. *Chaos, Solitons & Fractals*, 150:111119, 2021.

[3] Wei-Jen Chen, Mao-Jhen Jhou, Tian-Shyug Lee, and Chi-Jie Lu. Hybrid basketball game outcome prediction model by integrating data mining methods for the national basketball association. *Entropy*, 23(4):477, 2021.

[4] Alexander Dokumentov, Rob J Hyndman, et al. Low-dimensional decomposition, smoothing and forecasting of sparse functional data. *Technical report, Monash University*, 2014.

[5] Tomislav Horvat, Josip Job, and Vladimir Medved. Prediction of euroleague games based on supervised classification algorithm k-nearest neighbours. In *6th International Congress on Support Sciences Research and Technology Support*, volume 20, page 21, 2018.

[6] Albert Y. Kim, Chester Ismay, and Jennifer Chunn. The fivethirtyeight r package: 'tame data' principles for introductory statistics and data science courses. *Technology Innovations in Statistics Education*, 11, 2018.

[7] Nguyen Hoang Nguyen, Duy Thien An Nguyen, Bingkun Ma, and Jiang Hu. The application of machine learning and deep learning in sport: predicting nba players' performance and popularity. *Journal of Information and Telecommunication*, 6(2):217–235, 2022.

[8] Cem Osken and Ceylan Onay. Predicting the winning team in basketball: A novel approach. *Heliyon*, 8(12):e12189, 2022.

[9] James Piette, Lisa Pham, and Sathyanarayan Anand. Evaluating basketball player performance via statistical network modeling. In *The 5th MIT Sloan sports analytics conference*, pages 4–5, 2011.

[10] Vangelis Sarlis and Christos Tjortjis. Sports analytics—evaluation of basketball players and team performance. *Information Systems*, 93:101562, 2020.

[11] Rajitha Minusha Silva. *Sports analytics*. PhD thesis, Simon Fraser University, 2016.

[12] Guillermo Vinué Visús and Irene Epifanio. Forecasting basketball players' performance using sparse functional data. 2019.