ΔΙΕΘΝΕΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΗΣ ΕΛΛΑΔΟΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Εφαρμογή Μηχανικής Μάθησης στον Προσδιορισμό Βέλτιστων Στρατηγικών και Σύνθεσης Ομάδας Μπάσκετ»

«Εικόνα»

Του φοιτητή
Χρυσόστομου Αργυριάδη
Αρ. Μητρώου: 2020016

Επιβλέπων
Ονοματεπώνυμο: Χρήστος
Ηλιούδης

Ημερομηνία 22/01/2025

Τίτλος Δ.Ε.: Εφαρμογή Μηχανικής Μάθησης στον Προσδιορισμό Βέλτιστων Στρατηγικών και Σύνθεσης Ομάδας Μπάσκετ

Κωδικός Δ.Ε.: 25116

Ονοματεπώνυμο φοιτητή: Χρυσόστομος Αργυριάδης

Ονοματεπώνυμο εισηγητή Χρήστος Ηλιούδης

Ημερομηνία ανάληψης Δ.Ε: 14/02/2025

Ημερομηνία περάτωσης Δ.Ε.: 22/01/2026

Βεβαιώνω ότι είμαι ο συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως διπλωματική εργασία, στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

Πρόλογος

Η επιλογή του θέματος της παρούσας διπλωματικής εργασίας προέκυψε από το προσωπικό μου ενδιαφέρον για το μπάσκετ και τη διαχρονική μου περιέργεια σχετικά με τους παράγοντες που επηρεάζουν την αγωνιστική απόδοση των παικτών και τη συνολική λειτουργία μιας ομάδας. Το μπάσκετ, ως ένα δυναμικό και πολυπαραγοντικό άθλημα, προσφέρει ένα ιδιαίτερα ενδιαφέρον πεδίο μελέτης, στο οποίο η ατομική απόδοση, οι ρόλοι των παικτών και οι μεταξύ τους αλληλεπιδράσεις διαμορφώνουν το τελικό αγωνιστικό αποτέλεσμα.

Παράλληλα, η ενασχόληση με το αντικείμενο των sports analytics αποτέλεσε μια ευκαιρία για την ουσιαστική εξοικείωσή μου με τις τεχνικές μηχανικής μάθησης και τη στατιστική ανάλυση δεδομένων. Μέσα από την εκπόνηση της εργασίας, απέκτησα πρακτική εμπειρία στην εφαρμογή μεθόδων όπως η ομαδοποίηση και η παλινδρόμηση, καθώς και στην ερμηνεία των αποτελεσμάτων τους σε πραγματικά δεδομένα. Η διαδικασία αυτή με βοήθησε να κατανοήσω πώς η μηχανική μάθηση μπορεί να αξιοποιηθεί όχι μόνο στο μπάσκετ, αλλά και γενικότερα στον χώρο του αθλητισμού, ως εργαλείο ανάλυσης και υποστήριξης αποφάσεων.

Περίληψη

Τα τελευταία χρόνια, τα sports analytics έχουν εξελιχθεί σε βασικό πυλώνα λήψης αποφάσεων στον επαγγελματικό αθλητισμό, ως αποτέλεσμα της αυξημένης διαθεσιμότητας προηγμένων δεδομένων απόδοσης και της ανάπτυξης τεχνικών μηχανικής μάθησης. Στόχος της παρούσας διπλωματικής εργασίας είναι η διερεύνηση του τρόπου με τον οποίο δεδομενοκεντρικές μέθοδοι μπορούν να αξιοποιηθούν για την αναγνώριση αγωνιστικών ρόλων παικτών, την αξιολόγηση της ατομικής συνεισφοράς και την υποστήριξη στρατηγικών αποφάσεων που σχετίζονται με τη σύνθεση ομάδας στο μπάσκετ. Η μελέτη βασίζεται στη δημιουργία ενός δομημένου συνόλου δεδομένων προηγμένων στατιστικών χαρακτηριστικών παικτών, το οποίο συλλέχθηκε, καθαρίστηκε και προεπεξεργάστηκε με στόχο τη διασφάλιση της συνέπειας και της αξιοπιστίας του. Στη συνέχεια εφαρμόζονται τεχνικές μη επιβλεπόμενης μάθησης, και συγκεκριμένα οι αλγόριθμοι K-Means και Gaussian Mixture Models, για την ομαδοποίηση των παικτών βάσει των στατιστικών τους προφίλ και την ανάδειξη διακριτών αγωνιστικών ρόλων. Οι ομάδες που προκύπτουν αναλύονται περαιτέρω μέσω μοντέλων γραμμικής παλινδρόμησης, με μεταβλητή στόχο το Net Rating, το οποίο αποτυπώνει τη συνολική αγωνιστική επίδραση των παικτών στο παρκέ. Η ανάλυση πραγματοποιείται τόσο στο σύνολο των παικτών όσο και ξεχωριστά εντός κάθε αναγνωρισμένου ρόλου. Τα αποτελέσματα δείχνουν ότι οι δείκτες αποτελεσματικότητας και οι σύνθετοι δείκτες απόδοσης παρουσιάζουν ισχυρότερη και πιο σταθερή σχέση με το Net Rating σε σύγκριση με δείκτες όγκου συμμετοχής. Παράλληλα, αναδεικνύεται ότι η σημασία των επιμέρους χαρακτηριστικών διαφοροποιείται σημαντικά μεταξύ διαφορετικών αγωνιστικών ρόλων. Τέλος, παρουσιάζεται μία ενδεικτική εφαρμογή βελτιστοποίησης σύνθεσης πεντάδας παικτών, η οποία καταδεικνύει πώς τα αποτελέσματα της ανάλυσης μπορούν να αξιοποιηθούν για την υποστήριξη αποφάσεων στη δημιουργία αποτελεσματικών αγωνιστικών συνθέσεων.

.

«Application of Machine Learning for Identifying Optimal Strategies and Basketball Team Composition»

«Chrysostomos Argyriadis»

Abstract

In recent years, sports analytics has become a fundamental component of decision-making in professional basketball, driven by the increasing availability of advanced performance data and the development of machine learning techniques. The objective of this thesis is to explore how data-driven methods can be used to identify player roles, evaluate individual contribution, and support strategic decisions related to team composition. The study is based on a structured dataset of advanced basketball statistics, which was collected, cleaned, and preprocessed to ensure consistency and reliability. Unsupervised learning techniques, including K-Means clustering and Gaussian Mixture Models, are applied in order to group players according to their statistical profiles and reveal distinct playing roles. These clusters are subsequently analyzed using linear regression models, with the Net Rating metric serving as the target variable that captures the overall on-court impact of players. The analysis is conducted both on the entire player population and separately within each identified role. The results demonstrate that efficiency-based metrics and composite performance indicators have a stronger and more consistent relationship with Net Rating compared to volume-based statistics. Moreover, the influence of individual features varies significantly across different player roles, highlighting the limitations of using a single global evaluation model. Finally, a proof-of-concept lineup optimization framework is presented, illustrating how the derived role-based regression results can be used to support the selection of effective player combinations under predefined constraints. Overall, the thesis demonstrates that the integration of machine learning and advanced basketball data can provide interpretable insights and practical tools for performance analysis and strategic decision support.

Ευχαριστίες

Ευχαριστώ τον επιβλέποντα καθηγητή Χρήστο Ηλιούδη για την βοήθεια και τις χρήσιμες συμβουλές του.

Περιεχόμενα

Συντομογραφίες

Δ.Ε.	Διπλωματική Εργασία
ΔΙΠΑΕ	Διεθνές Πανεπιστήμιο Ελλάδος
Π.Ε.	Πτυχιακή Εργασία

Κεφάλαιο 1ο: Εισαγωγή

1.1 Το αντικείμενο

Ο τομέας των **sports analytics** έχει εξελιχθεί εντυπωσιακά τις τελευταίες δύο δεκαετίες, μετατρέποντας τον επαγγελματικό αθλητισμό σε ένα πεδίο όπου τα δεδομένα αποτελούν κεντρικό πυλώνα στρατηγικής και αξιολόγησης. Η πρόοδος στην υπολογιστική ισχύ, οι εξελιγμένες μέθοδοι συλλογής δεδομένων (όπως optical tracking, player-tracking sensors και advanced play-by-play καταγραφές) και η ωρίμανση των τεχνικών μηχανικής μάθησης έχουν επιτρέψει στους αναλυτές να ερμηνεύουν τον αθλητικό χώρο με τρόπους που παλαιότερα θεωρούνταν αδύνατοι. Η αξιοποίηση τέτοιων δεδομένων επιτρέπει μια πιο συστηματική κατανόηση των παραγόντων που επηρεάζουν την απόδοση τόσο σε ατομικό όσο και σε ομαδικό επίπεδο, υποστηρίζοντας διαδικασίες όπως η αξιολόγηση παικτών , η ανάλυση τακτικών επιλογών και η διαμόρφωση αγωνιστικών στρατηγικών [1].

Στο μπάσκετ ειδικότερα, η άνθηση των analytics έχει συνδεθεί άμεσα με τη μεταβολή του τρόπου που παίζεται και αναλύεται το παιχνίδι. Οι σύγχρονες τεχνολογίες καταγραφής δεδομένων —όπως οι πλατφόρμες SportVU και Second Spectrum [2]— έχουν φέρει σε πραγματικό χρόνο εκατομμύρια σημεία δεδομένων ανά αγώνα. σε πραγματικό χρόνο εκατομμύρια σημεία δεδομένων ανά αγώνα. Αυτές οι πληροφορίες επιτρέπουν όχι μόνο την καλύτερη αξιολόγηση των παικτών, αλλά και τη χαρτογράφηση της αγωνιστικής ταυτότητας μιας ομάδας, τον εντοπισμό προτύπων κίνησης, την αναγνώριση στρατηγικών επιλογών και την κατανόηση των αλληλεπιδράσεων μεταξύ πεντάδων.

Η εξέλιξη του παιχνιδιού —με την αύξηση του ρυθμού, τη μετάβαση σε spacing-oriented offenses, τη μεγαλύτερη έμφαση στα τρίποντα και τη διαφοροποίηση των ρόλων— έχει καταστήσει αναγκαία την ύπαρξη εργαλείων που μπορούν να αναλύσουν λεπτομερώς τις ανάγκες και τις αλληλεπιδράσεις μιας ομάδας. Στην πράξη, οι προπονητές καλούνται να λάβουν αποφάσεις που αφορούν την κατανομή λεπτών συμμετοχής, τη δημιουργία αποτελεσματικών lineups, την ανάθεση ρόλων, καθώς και την επιλογή στρατηγικών που μεγιστοποιούν τις πιθανότητες νίκης. Δεν μπορούν να βασίζονται αποκλειστικά στην προσωπική τους εμπειρία ή τη διαισθητική αντίληψη του παιχνιδιού· η πολυπλοκότητα του σύγχρονου μπάσκετ απαιτεί τη συστηματική αξιοποίηση δεδομένων και αναλυτικών εργαλείων ώστε οι αποφάσεις που λαμβάνονται να είναι τεκμηριωμένες και όσο το δυνατόν ακριβέστερες.

Σε αυτό το πλαίσιο, η μηχανική μάθηση προσφέρει μεθόδους που μπορούν να εντοπίσουν μοτίβα τα οποία δεν είναι άμεσα ορατά στον ανθρώπινο παρατηρητή, να ομαδοποιήσουν παίκτες με βάση κοινά χαρακτηριστικά, να προβλέψουν επιδόσεις και να υποστηρίξουν διαδικασίες όπως η κατασκευή βέλτιστου ρόστερ [3] και η ανάπτυξη στρατηγικών παιχνιδιού. Η παρούσα διπλωματική εργασία εντάσσεται ακριβώς σε αυτό το αναδυόμενο πεδίο, προτείνοντας μια προσέγγιση που αξιοποιεί δεδομένα αγώνων και τεχνικές μηχανικής μάθησης για την ανακάλυψη ρόλων παικτών, την αξιολόγηση συνθέσεων ομάδας και την υποστήριξη λήψης στρατηγικών αποφάσεων.

1.2 Οι στόχοι της διπλωματικής

Οι στόχοι της παρούσας διπλωματικής εργασίας διαμορφώνονται έτσι ώστε να καλύπτουν σφαιρικά όλα τα στάδια που απαιτούνται για μια ολοκληρωμένη, ανάλυση στον χώρο του μπάσκετ με βάση τα δεδομένα. Αφορούν τόσο το θεωρητικό υπόβαθρο και τη μεθοδολογία όσο και τις πρακτικές εφαρμογές που μπορούν να αξιοποιηθούν στην προπονητική διαδικασία και στη λήψη διοικητικών ή τεχνικών αποφάσεων.

Πιο αναλυτικά, η εργασία επιδιώκει:

- **Τη συστηματική συλλογή, καθαρισμό και προεπεξεργασία δεδομένων αγώνων μπάσκετ.** Η διαδικασία αυτή περιλαμβάνει την ανάλυση των διαθέσιμων πηγών, την επιλογή των καταλληλότερων στατιστικών δεικτών, την αντιστοίχιση δεδομένων παικτών και ομάδων, καθώς και την αντιμετώπιση πιθανών ασυνεπειών. Έμφαση δίνεται στη δημιουργία ενός αξιόπιστου και συνεκτικού dataset που αντανακλά με ακρίβεια την αγωνιστική πραγματικότητα.
- **Την εφαρμογή τεχνικών μηχανικής μάθησης, όπως clustering, classification και dimensionality reduction**, για την αναγνώριση ρόλων παικτών και την αποκάλυψη υποκείμενων μοτίβων απόδοσης. Στο πλαίσιο αυτό εξετάζεται ο τρόπος με τον οποίο χαρακτηριστικά όπως η αποτελεσματικότητα, η συμμετοχή στη δημιουργία παιχνιδιού και η αμυντική συνεισφορά διαμορφώνουν διαφορετικά προφίλ παικτών.
- **Την αξιολόγηση και σύγκριση διαφορετικών αλγορίθμων και μετρικών απόδοσης**, με στόχο να εντοπιστούν οι μέθοδοι που προσφέρουν τα πιο σταθερά και χρήσιμα αποτελέσματα. Η διαδικασία αυτή περιλαμβάνει ανάλυση της ακρίβειας, της ευρωστίας και της ερμηνευσιμότητας των μοντέλων, ώστε τα παραγόμενα συμπεράσματα να μπορούν να αξιοποιηθούν στην πράξη.
- **Την ανάπτυξη ενός μοντέλου υποστήριξης απόφασης για τη βέλτιστη σύνθεση ομάδας**, το οποίο λαμβάνει υπόψη στατιστικά χαρακτηριστικά παικτών, μοτίβα συνεργασίας και τις μεταξύ τους αλληλεπιδράσεις. Στόχος είναι η δημιουργία ενός συστήματος που βοηθά στην επιλογή πεντάδων ή συνολικών ρόστερ με βάση αντικειμενικά κριτήρια και όχι μόνο εμπειρική εκτίμηση.
- **Τη διερεύνηση του ρόλου της μηχανικής μάθησης στη χάραξη αγωνιστικών στρατηγικών**, όπως στη διαχείριση χρόνου συμμετοχής, στην επιλογή αμυντικών προσεγγίσεων, στην πρόβλεψη της απόδοσης υπό διαφορετικά σενάρια και στη μακροχρόνια πορεία της ομάδας μέσα σε μια σεζόν.

1.3 Τα επιτεύγματα της διπλωματικής

Η παρούσα διπλωματική εργασία κατέληξε σε μια ολοκληρωμένη αναλυτική προσέγγιση του σύγχρονου μπάσκετ, συνδυάζοντας πραγματικά αγωνιστικά δεδομένα με τεχνικές μηχανικής μάθησης και στατιστικής ανάλυσης. Τα βασικά επιτεύγματα της εργασίας συνοψίζονται στα ακόλουθα σημεία.

Αρχικά, επιτεύχθηκε η δημιουργία ενός αξιόπιστου και συνεκτικού dataset παικτών, το οποίο προέκυψε έπειτα από συστηματική συλλογή, καθαρισμό και προεπεξεργασία δεδομένων αγώνων. Η επιλογή advanced στατιστικών δεικτών, όπως δείκτες αποτελεσματικότητας, συμμετοχής και αμυντικής

συνεισφοράς, επέτρεψε την αποτύπωση της αγωνιστικής συμπεριφοράς των παικτών με τρόπο που υπερβαίνει τα παραδοσιακά box-score στατιστικά.

Στη συνέχεια, εφαρμόστηκαν τεχνικές μη επιβλεπόμενης μάθησης για την αναγνώριση υποκείμενων προτύπων απόδοσης και τη χαρτογράφηση διακριτών ρόλων παικτών. Μέσω αλγορίθμων clustering (K-Means και Gaussian Mixture Models), η εργασία ανέδειξε ομάδες παικτών με σαφώς διαφοροποιημένα αγωνιστικά χαρακτηριστικά, τα οποία μπορούν να ερμηνευθούν ως λειτουργικοί ρόλοι εντός του σύγχρονου παιχνιδιού. Η ανάλυση αυτή προσέφερε μια πιο δομημένη και αντικειμενική κατηγοριοποίηση των παικτών, σε αντίθεση με εμπειρικές ή καθαρά θέσει-κεντρικές προσεγγίσεις.

Επιπλέον, υλοποιήθηκαν μοντέλα γραμμικής παλινδρόμησης με στόχο την κατανόηση των παραγόντων που επηρεάζουν τη συνολική αγωνιστική επίδραση των παικτών, όπως αυτή αποτυπώνεται μέσω του NET Rating. Η ανάλυση πραγματοποιήθηκε τόσο σε συνολικό επίπεδο όσο και ανά cluster, επιτρέποντας τη σύγκριση της σημασίας των στατιστικών χαρακτηριστικών μεταξύ διαφορετικών τύπων παικτών. Τα αποτελέσματα κατέδειξαν ότι οι δείκτες αποτελεσματικότητας διαδραματίζουν καθοριστικό ρόλο στην ερμηνεία της συνολικής συνεισφοράς, ενώ η σημασία επιμέρους χαρακτηριστικών διαφοροποιείται ανάλογα με τον αγωνιστικό ρόλο.

Ιδιαίτερη συνεισφορά της εργασίας αποτελεί η συγκριτική αξιολόγηση διαφορετικών μεθόδων clustering. Η σύγκριση μεταξύ K-Means και Gaussian Mixture Models ανέδειξε πλεονεκτήματα και περιορισμούς κάθε προσέγγισης, τόσο ως προς την ερμηνευσιμότητα όσο και ως προς τη στατιστική τους απόδοση στα επακόλουθα μοντέλα παλινδρόμησης. Με τον τρόπο αυτό, η εργασία δεν περιορίστηκε στην εφαρμογή αλγορίθμων, αλλά προχώρησε σε κριτική αξιολόγηση της καταλληλότητάς τους για το συγκεκριμένο πρόβλημα.

Τέλος, αναπτύχθηκε ένα απλοποιημένο μοντέλο υποστήριξης απόφασης για τη βελτιστοποίηση σύνθεσης πεντάδας παικτών (lineup optimization). Αξιοποιώντας τα αποτελέσματα της παλινδρόμησης ανά cluster, υπολογίστηκε ένας αναμενόμενος δείκτης αγωνιστικής επίδρασης για κάθε παίκτη, ο οποίος χρησιμοποιήθηκε ως αντικειμενική συνάρτηση σε διαδικασία επιλογής πεντάδας με προκαθορισμένη κατανομή ρόλων. Η προσέγγιση αυτή καταδεικνύει πώς τα αποτελέσματα της μηχανικής μάθησης μπορούν να μεταφραστούν σε πρακτικά εργαλεία υποστήριξης προπονητικών και στρατηγικών αποφάσεων.

1.4 Η διάρθρωση της διπλωματικής

Η δομή της διπλωματικής εργασίας έχει σχεδιαστεί με τρόπο που να επιτρέπει στον αναγνώστη να ακολουθήσει βήμα προς βήμα τη διαδικασία ανάλυσης, από το θεωρητικό υπόβαθρο έως την πειραματική υλοποίηση και τα τελικά συμπεράσματα.

- **Κεφάλαιο 2 — Ανάλυση δεδομένων στον χώρο του μπάσκετ & τεχνολογίες μηχανικής μάθησης.** Το κεφάλαιο αυτό παρουσιάζει το θεωρητικό πλαίσιο της εργασίας, περιγράφοντας βασικές έννοιες των sports analytics, των αλγορίθμων μηχανικής μάθησης και των μεθόδων που χρησιμοποιούνται για την αναγνώριση προτύπων στο μπάσκετ. Γίνεται επίσης συγκριτική αναφορά σε εφαρμογές που έχουν παρουσιαστεί στη βιβλιογραφία, αναδεικνύοντας τα πλεονεκτήματα και τις προκλήσεις των διαφορετικών προσεγγίσεων.

- **Κεφάλαιο 3 — Πηγές δεδομένων & διαδικασία προεπεξεργασίας.** Εδώ περιγράφονται αναλυτικά οι πηγές άντλησης δεδομένων, η δομή των datasets, τα χαρακτηριστικά που χρησιμοποιούνται στην ανάλυση, καθώς και τα βήματα καθαρισμού, κανονικοποίησης και μετασχηματισμού των δεδομένων. Δίνεται έμφαση στην ανάγκη επαρκούς ποιότητας δεδομένων, ώστε τα αποτελέσματα των μοντέλων να είναι αξιόπιστα.
- **Κεφάλαιο 4 — Πειραματική υλοποίηση & αξιολόγηση μοντέλων.** Το κεφάλαιο αυτό αποτελεί τον πυρήνα της εργασίας. Περιγράφεται η εφαρμογή των μεθόδων μηχανικής μάθησης, τα πειράματα που πραγματοποιήθηκαν, οι παράμετροι των αλγορίθμων, καθώς και τα αποτελέσματα που προέκυψαν. Παρουσιάζεται επίσης η διαδικασία αξιολόγησης και σύγκρισης των μοντέλων, με στόχο την ανάδειξη των μεθόδων που προσφέρουν την υψηλότερη απόδοση και τη μεγαλύτερη πρακτική αξία.
- **Κεφάλαιο 5 — Συμπεράσματα & μελλοντικές επεκτάσεις.** Στο τελευταίο κεφάλαιο συνοψίζονται τα συμπεράσματα που προέκυψαν από την πειραματική διαδικασία και προτείνονται κατευθύνσεις για μελλοντική έρευνα και ανάπτυξη συστημάτων ανάλυσης στο μπάσκετ. Η ενότητα αυτή αναδεικνύει τη συνεισφορά της εργασίας και θέτει τις βάσεις για περαιτέρω εμβάθυνση στο πεδίο.

Κεφάλαιο 2ο: Αναλυτική δεδομένων στον χώρο του αθητισμού και της καλαθοσφαίρισης

2.1 Μηχανική Μάθηση στα σπορ

Τα τελευταία χρόνια, τα sports analytics και η μηχανική μάθηση έχουν εξελιχθεί σε βασικά εργαλεία για την αξιολόγηση και κατανόηση της αθλητικής απόδοσης. Η πρόοδος στις μεθόδους συλλογής δεδομένων και η αυξανόμενη υπολογιστική ισχύς έχουν επιτρέψει την αξιοποίηση τεχνικών μηχανικής μάθησης σε μια ευρεία γκάμα αθλημάτων, πολύ πέρα από τον χώρο του μπάσκετ [23].

Αλγόριθμοι ταξινόμησης, παλινδρόμησης, ομαδοποίησης, Bayesian μοντέλα και σύγχρονες deep learning προσεγγίσεις έχουν εφαρμοστεί σε διάφορα σπορ, προσφέροντας λύσεις σε ποικίλους τομείς: από την πρόβλεψη αποτελεσμάτων και την ανάλυση τακτικών, μέχρι την αξιολόγηση παικτών, την βελτιστοποίηση στρατηγικών, και τη μελέτη των spatio-temporal δεδομένων. Η διείσδυση αυτή δείχνει ότι η μηχανική μάθηση αποτελεί πλέον κεντρικό εργαλείο στην αθλητική ανάλυση συνολικά, και όχι μια τεχνολογία που αφορά αποκλειστικά το μπάσκετ [24].

2.1.1 Sensor Technology και Μηχανική Μάθηση

Η αξιοποίηση φορετών αισθητήρων (wearable sensor technology) αποτελεί πλέον μία από τις βασικότερες και πιο αξιόπιστες τεχνικές συλλογής δεδομένων στον χώρο του αθλητισμού. Συστήματα όπως accelerometers, gyroscopes, magnetometers και γενικότερα inertial measurement units (IMUs) επιτρέπουν την καταγραφή λεπτομερών κινηματικών πληροφοριών σε πραγματικό χρόνο, προσφέροντας μια άμεση και χαμηλού κόστους εναλλακτική λύση σε σχέση με τα παραδοσιακά video- ή motion-capture συστήματα. Οι αισθητήρες αυτοί μπορούν να τοποθετηθούν με ευελιξία σε διάφορα σημεία του σώματος, παρέχοντας δεδομένα που αφορούν τη στάση, τη γωνιακή ταχύτητα, τις επιταχύνσεις, τις μετατοπίσεις και συνολικά το κινητικό προφίλ του αθλητή [22].

Η ανάπτυξη της μηχανικής μάθησης έχει ενισχύσει ακόμη περισσότερο τη χρησιμότητα των sensor-based δεδομένων, καθώς η επεξεργασία των σημάτων από τους αισθητήρες επιτρέπει την αυτόματη εξαγωγή χαρακτηριστικών, την αναγνώριση μοτίβων και τη δημιουργία μοντέλων ικανά να ταξινομήσουν, να προβλέψουν ή να αξιολογήσουν αθλητικές συμπεριφορές. Με τον συνδυασμό αυτών των τεχνολογιών, καθίσταται εφικτή η ανάλυση εξειδικευμένων τεχνικών κινήσεων, η κατηγοριοποίηση χτυπημάτων ή στάσεων, καθώς και η αντικειμενική αξιολόγηση της απόδοσης, χωρίς την ανάγκη βιντεοσκόπησης ή χειροκίνητης ερμηνείας [21].

Παρακάτω παρουσιάζονται χαρακτηριστικά παραδείγματα μελετών που αξιοποιούν αισθητήρες και μοντέλα μηχανικής μάθησης σε διαφορετικά αθλήματα, αναδεικνύοντας πώς η sensor-based προσέγγιση έχει συμβάλει σημαντικά στην κατανόηση, ανάλυση και αξιολόγηση της τεχνικής και της απόδοσης των αθλητών.

Ο Connaghan et al. [4] ανέπτυξαν ένα σύστημα αναγνώρισης χτυπημάτων στο τένις χρησιμοποιώντας φορετούς αισθητήρες (IMUs) για τη συλλογή δεδομένων επιτάχυνσης και γυροσκοπίου, και στη συνέχεια εφάρμοσαν τεχνικές μηχανικής μάθησης ώστε να ταξινομήσουν αυτόματα διαφορετικούς τύπους χτυπημάτων (όπως forehand, backhand, serve και volley). Η μελέτη τους δείχνει ότι η ανάλυση τεχνικής στο τένις μπορεί να πραγματοποιηθεί αξιόπιστα χωρίς ακριβά συστήματα καταγραφής κίνησης, βασιζόμενη αποκλειστικά σε wearable τεχνολογία.

Ο Ghasemzadeh et al. [5] πρότειναν ένα σύστημα εκπαίδευσης για το γκολφ που βασίζεται σε φορετούς motion sensors (inertial και body-worn sensors) ώστε να συλλέξουν δεδομένα από την κίνηση του σώματος, με έμφαση στη μέτρηση της περιστροφής του καρπού (wrist rotation) κατά την εκτέλεση του swing. Αξιοποιώντας στατιστικές μεθόδους για την επεξεργασία των raw σημάτων, οι συγγραφείς έδειξαν πως μπορούν να εξαχθούν ποσοτικές μετρήσεις που αξιολογούν την ποιότητα και την τεχνική της κίνησης. Η συμβολή της μελέτης είναι ότι αποδεικνύει πως, με σχετικά απλή φορετή τεχνολογία και χωρίς ακριβά motion-capture συστήματα, μπορεί να πραγματοποιηθεί αξιόπιστη και λεπτομερής τεχνική ανάλυση σε αθλήματα όπως το γκολφ, καθιστώντας τη sensor-based αξιολόγηση πρακτική για πραγματικές συνθήκες προπόνησης.

Ghosh, Ramamurthy et al. [6] ανέπτυξαν συστήματα αναγνώρισης και αξιολόγησης κινήσεων στο badminton αξιοποιώντας φορετούς αισθητήρες, όπως accelerometers και motion sensors, για την ακριβή καταγραφή της στάσης, των μετατοπίσεων και των μοτίβων κίνησης των παικτών. Τα δεδομένα των αισθητήρων υποβλήθηκαν σε επεξεργασία και τροφοδοτήθηκαν σε μοντέλα μηχανικής μάθησης και deep learning, επιτρέποντας την αυτόματη κατηγοριοποίηση στάσεων, χτυπημάτων και άλλων τεχνικών στοιχείων του παιχνιδιού. Η μελέτη δείχνει ότι μια τέτοια sensor-based προσέγγιση μπορεί να αναγνωρίζει με υψηλή ακρίβεια τις κινήσεις του badminton, ενώ παράλληλα προσφέρει τη δυνατότητα ποσοτικής αξιολόγησης της τεχνικής και της συνολικής απόδοσης των παικτών, καθιστώντας τη κατάλληλη για σύγχρονες εφαρμογές coaching και performance analysis.

2.1.2 Ανάλυση Κίνησης με Βιντεοσκοπικά Δεδομένα

Η εξαγωγή δεδομένων από βίντεο αποτελεί έναν από τους σημαντικότερους και πιο ευέλικτους τρόπους συλλογής πληροφορίας στον σύγχρονο αθλητισμό. Σε αντίθεση με τις τεχνικές που βασίζονται σε φορετούς αισθητήρες, η video-based ανάλυση δεν απαιτεί εξοπλισμό πάνω στον αθλητή, γεγονός που την καθιστά ιδιαίτερα κατάλληλη για αγωνιστικές συνθήκες και για αθλήματα όπου η άμεση παρέμβαση στον εξοπλισμό του παίκτη δεν είναι εφικτή. Παράλληλα, η επεξεργασία βίντεο επιτρέπει τη συλλογή μεγάλου όγκου δεδομένων χωρίς την ανάγκη προηγούμενης εγκατάστασης αισθητήρων, αξιοποιώντας υλικό που παράγεται ούτως ή άλλως σε κάθε επίσημο αγώνα.

Με τη χρήση τεχνικών υπολογιστικής όρασης, pose estimation και deep learning, τα βίντεο μπορούν να μετατραπούν σε δομημένα δεδομένα που περιγράφουν λεπτομερώς την κίνηση, τη στάση του σώματος, την ταχύτητα, τους τύπους ενεργειών και τα κινηματικά μοτίβα. Επιπλέον, πολλές σύγχρονες πλατφόρμες ανάλυσης εξάγουν από τα βίντεο tracking data – δηλαδή τις ακριβείς θέσεις και τροχιές παικτών και μπάλας σε κάθε χρονικό σημείο. Με αυτόν τον τρόπο, το βίντεο λειτουργεί ως η «πρώτη ύλη» από την οποία παράγονται τα ακριβή δεδομένα κίνησης που χρησιμοποιούνται σε ποσοτικές αναλύσεις, τακτικά μοντέλα και αλγοριθμικές εκτιμήσεις αγωνιστικών καταστάσεων.

Πέρα από την αξιολόγηση της ατομικής τεχνικής, η video-based προσέγγιση έχει αποδειχθεί εξαιρετικά αποτελεσματική και στην ανάλυση τακτικής και στρατηγικής. Επιτρέπει την παρακολούθηση των θέσεων των παικτών στο γήπεδο, την καταγραφή των συστημάτων παιχνιδιού, την αναγνώριση μοτίβων συνεργασίας και την κατανόηση της ομαδικής συμπεριφοράς. Παράλληλα, συμβάλλει στην αποτύπωση χαρακτηριστικών παικτών, όπως στυλ παιχνιδιού, προτιμώμενες κινήσεις, ρυθμός λήψης αποφάσεων και δείκτες τεχνικής απόδοσης. Έτσι, η video-based ανάλυση συνδυάζει ευελιξία, πλούσια πληροφορία και δυνατότητα παραγωγής μεγάλων ποσοτήτων δεδομένων, υποστηρίζοντας τόσο την αξιολόγηση μεμονωμένων αθλητών όσο και τη μελέτη της συνολικής δυναμικής μιας ομάδας.

Ο Cai et al. [7] ανέπτυξαν ένα σύστημα αναγνώρισης αθλητικών ενεργειών στο χόκεϊ βασισμένο σε δεδομένα βίντεο, αξιοποιώντας δύο συμπληρωματικές πηγές πληροφορίας: την εκτίμηση στάσης (pose estimation) και το optical flow. Χρησιμοποιώντας το Part Affinity Fields μοντέλο εξάγουν λεπτομερή spatial χαρακτηριστικά από κάθε εικόνα — συμπεριλαμβανομένης της θέσης του hockey stick ως επιπλέον άρθρωση — ενώ με το LiteFlowNet υπολογίζουν optical flows που αποτυπώνουν τη χρονική και κινηματική πληροφορία της κίνησης. Οι δύο αυτές ροές πληροφορίας συνδυάζονται σε μία unified two-stream deep learning αρχιτεκτονική για την ταξινόμηση τεσσάρων ενεργειών: skating forward, skating backward, passing και shooting. Η μελέτη δείχνει ότι ο συνδυασμός spatial και temporal χαρακτηριστικών βελτιώνει σημαντικά την απόδοση του συστήματος, επιτυγχάνοντας ακρίβεια έως 85%, ενώ η ενσωμάτωση του hockey stick στο pose estimation ενισχύει περαιτέρω την αναγνώριση των ενεργειών. Επιπλέον, οι συγγραφείς εισάγουν το HARPET, ένα νέο dataset με sequences εικόνων και πλήρη annotation αρθρώσεων και ενεργειών, το οποίο επιτρέπει την αξιόπιστη εκπαίδευση και αξιολόγηση μοντέλων σε ρεαλιστικές συνθήκες αγωνιστικού χόκεϊ.

Piergiovanni και Ryoo [8] ανέπτυξαν ένα σύστημα αναγνώρισης λεπτομερών αθλητικών ενεργειών στο baseball χρησιμοποιώντας δεδομένα που εξάγονται αποκλειστικά από βίντεο. Δημιούργησαν το MLB-YouTube dataset, ένα μεγάλο σύνολο broadcast βίντεο από αγώνες baseball, όπου οι διαφορές μεταξύ των ενεργειών (όπως swing, bunt, hit, strike, ball) είναι εξαιρετικά μικρές, κάτι που καθιστά το πρόβλημα ιδιαίτερα απαιτητικό. Το σύστημα αξιοποιεί σύγχρονες τεχνικές spatio-temporal deep learning και μεθόδους χρονικής ομαδοποίησης (temporal pooling) για να αναλύσει τη λεπτή κινηματική πληροφορία και να ταξινομήσει με ακρίβεια τις ενέργειες μέσα σε βίντεο

Merckx et al. [9] ανέπτυξαν μια μέθοδο για την αυτόματη ανίχνευση και αξιολόγηση της αποτελεσματικότητας του pressing στο ποδόσφαιρο, βασισμένη σε δεδομένα θέσεων παικτών και συμβάντων αγώνα. Αρχικά, ορίζουν ένα σύνολο κανόνων που εντοπίζει στιγμές όπου η αμυντική ομάδα ασκεί πραγματική πίεση στον παίκτη με την μπάλα, λαμβάνοντας υπόψη την απόσταση των αμυντικών, την ένταση και την κατεύθυνση της κίνησής τους, καθώς και το κατά πόσο περιορίζουν τον διαθέσιμο χώρο του αντιπάλου. Με αυτόν τον μηχανισμό αναγνωρίζονται αυτόματα οι φάσεις πίεσης και ομαδοποιούνται σε πλήρεις pressing καταστάσεις.Στη συνέχεια, οι συγγραφείς προτείνουν ένα μοντέλο που αξιολογεί την αποτελεσματικότητα κάθε φάσης, αντιμετωπίζοντας το pressing ως μια στρατηγική επιλογή με πιθανό όφελος και πιθανό κόστος. Το όφελος αφορά την πιθανότητα ανάκτησης της μπάλας σε ευνοϊκή θέση, ενώ το κόστος συνδέεται με τον κίνδυνο να εκτεθεί η άμυνα και να δημιουργήσει ο αντίπαλος επικίνδυνη επίθεση. Για να εκτιμήσουν αυτές τις πιθανότητες, χρησιμοποιούν επιμέρους μοντέλα που προβλέπουν την επόμενη πάσα του αντιπάλου, την πιθανότητα να κερδηθεί η κατοχή και την πιθανότητα να οδηγήσει η φάση σε απειλή προς την εστία. Ο συνδυασμός αυτών επιτρέπει μια

ολοκληρωμένη και αντικειμενική μέτρηση της αποτελεσματικότητας του pressing σε διαφορετικά αγωνιστικά περιβάλλοντα.

Almujahed et al. [10] ανέπτυξαν ένα σύστημα υποστήριξης αποφάσεων για την ανάλυση αγώνων βόλεϊ, βασισμένο στην επεξεργασία μεγάλου όγκου δεδομένων και στη χρήση βιντεοσκοπημένου υλικού ως κύρια πηγή πληροφορίας. Το σύστημα συλλέγει δεδομένα από βίντεο αγώνων και τα μετατρέπει σε δομημένη μορφή μέσω διαδικασιών καταγραφής και ανάλυσης ενεργειών, με στόχο να αξιολογήσει κρίσιμα στοιχεία του παιχνιδιού όπως οι πάσες, τα σέρβις, τα block και οι επιθέσεις. Η προσέγγιση αυτή επιτρέπει την αντικειμενική ανάλυση αγωνιστικών μοτίβων και την εξαγωγή στατιστικών δεικτών που υποστηρίζουν τόσο την τεχνική αξιολόγηση όσο και τη διαδικασία λήψης αποφάσεων. Παράλληλα, η αξιοποίηση τεχνικών big data για την αποθήκευση και επεξεργασία των πληροφοριών δίνει τη δυνατότητα εντοπισμού τάσεων και σύγκρισης αγωνιστικών προφίλ, καθιστώντας το σύστημα χρήσιμο εργαλείο για προπονητές και αναλυτές.

2.2 Ανάλυση Δεδομένων και Μηχανική Μάθηση στο Μπάσκετ

Στις προηγούμενες ενότητες παρουσιάστηκαν ενδεικτικές εφαρμογές της μηχανικής μάθησης και της ανάλυσης δεδομένων σε διαφορετικά αθλήματα, αναδεικνύοντας την ευρεία υιοθέτηση των σύγχρονων αναλυτικών μεθόδων στον αθλητισμό. Οι προσεγγίσεις αυτές κατέδειξαν πώς τεχνικές βασισμένες σε αισθητήρες, βίντεο και δεδομένα κίνησης μπορούν να χρησιμοποιηθούν για την ανάλυση απόδοσης, τακτικής και στρατηγικής σε ποικίλα αγωνιστικά περιβάλλοντα.

Στη συνέχεια, η βιβλιογραφική ανασκόπηση επικεντρώνεται αποκλειστικά στο μπάσκετ, ένα άθλημα που αποτελεί ένα από τα πλέον μελετημένα πεδία στα sports analytics. Η υψηλή συχνότητα γεγονότων, η διαθεσιμότητα λεπτομερών στατιστικών δεδομένων και η εκτεταμένη χρήση προηγμένων τεχνολογιών καταγραφής έχουν καταστήσει το μπάσκετ ιδανικό αντικείμενο για την εφαρμογή και εξέλιξη μεθόδων μηχανικής μάθησης. Στις επόμενες ενότητες παρουσιάζονται βασικές έρευνες και μεθοδολογίες που αφορούν την ανάλυση παικτών, ομάδων και αγωνιστικών καταστάσεων στο μπάσκετ, θέτοντας το πλαίσιο για τη μελέτη που ακολουθεί.

Στις παρακάτω ενότητες παρουσιάζονται επιλεγμένες ερευνητικές μελέτες και επιστημονικά έργα που έχουν συμβάλει ουσιαστικά στην ανάπτυξη της ανάλυσης δεδομένων και της μηχανικής μάθησης στο μπάσκετ. Τα έργα αυτά καλύπτουν ένα ευρύ φάσμα εφαρμογών, όπως η αξιολόγηση της απόδοσης παικτών, η ανάλυση αγωνιστικών καταστάσεων, η μοντελοποίηση τακτικών και η εξαγωγή προχωρημένων δεικτών απόδοσης. Μέσα από την παρουσίαση και την ανάλυσή τους, αναδεικνύονται οι βασικές μεθοδολογίες που έχουν χρησιμοποιηθεί στη βιβλιογραφία, καθώς και οι τρόποι με τους οποίους η μηχανική μάθηση έχει συμβάλει στη βαθύτερη κατανόηση και ανάλυση του παιχνιδιού του μπάσκετ

2.2.1 Αξιολόγηση παικτών και Roster Building

Η αξιολόγηση παικτών και η σύνθεση ρόστερ αποτελούν βασικές εφαρμογές της ανάλυσης δεδομένων στο μπάσκετ. Η χρήση προχωρημένων στατιστικών και μεθόδων μηχανικής μάθησης έχει επιτρέψει την πιο αντικειμενική αποτίμηση της αγωνιστικής συνεισφοράς των παικτών, καθώς και τη σύγκριση αγωνιστικών προφίλ με στόχο τη βελτιστοποίηση των αποφάσεων που αφορούν μεταγραφές, ρόλους

και δομή ομάδων. Έχουν πραγματοποιηθεί πολλές μελέτες οι οποίες ππαρουσιάζουν ενδεικτικές προσεγγίσεις οι οποίες αξιοποιούν δεδομένα αγώνων για την υποστήριξη αποφάσεων σχετικών με το player evaluation και το roster building.

Zhang et al. [11] εφάρμοσαν τον αλγόριθμο K-means clustering με στόχο την κατηγοριοποίηση παικτών του NBA που αγωνίζονται στη θέση του guard, βασιζόμενοι αποκλειστικά στα αγωνιστικά τους στατιστικά. Χρησιμοποιώντας σύνολα χαρακτηριστικών που περιλαμβάνουν δείκτες σκοραρίσματος, δημιουργίας παιχνιδιού, ευστοχίας και συμμετοχής στο παιχνίδι, οι συγγραφείς επιδίωξαν να ομαδοποιήσουν τους guards σε διακριτές κατηγορίες με παρόμοια αγωνιστικά προφίλ. Η προσέγγιση αυτή δεν στηρίζεται σε προκαθορισμένους ρόλους ή υποκειμενικές ετικέτες, αλλά επιτρέπει στα δεδομένα να αποκαλύψουν φυσικές ομάδες παικτών με κοινά χαρακτηριστικά. Τα αποτελέσματα δείχνουν ότι το clustering μπορεί να διαχωρίσει αποτελεσματικά διαφορετικούς τύπους guards, όπως scorers, playmakers ή πιο ισορροπημένους παίκτες, προσφέροντας ένα χρήσιμο εργαλείο για το χτίσιμο προφίλ και τη σύγκριση των παικτών.

Patton et al. [12] παρουσίασαν μια μελέτη στην οποία χρησιμοποιούν tracking data από χιλιάδες αγώνες κολεγιακού μπάσκετ — παραγωγή δεδομένων από broadcast βίντεο μέσω computer vision τεχνικών — για να προβλέψουν τη μετατροπή ενός παίκτη σε μελλοντικό NBA παίκτη. Η προσέγγιση βασίζεται στην εξαγωγή πλούσιων χαρακτηριστικών για κάθε παίκτη (π.χ. κινηματικά μοτίβα, defensive match-ups, ball-screens, post-ups και άλλα play actions) από τα tracking δεδομένα που δημιουργήθηκαν από περισσότερα από 650.000 possessions και εκατοντάδες εκατομμύρια καρέ βίντεο, και στη συνέχεια εκπαίδευση μοντέλων μηχανικής μάθησης για να εκτιμήσουν την πιθανότητα ένας παίκτης να γίνει NBA και σε ποια θέση του NBA Draft μπορεί να βρίσκεται. Η χρήση των tracking δεδομένων αποδεικνύεται πιο αποτελεσματική για την πρόβλεψη ταλέντου σε σχέση με τα παραδοσιακά play-by-play στατιστικά, καθώς το tracking log-loss βελτιώνεται σημαντικά, και επιπλέον οι συγγραφείς ενσωματώνουν επεξηγηματικές τεχνικές Μηχανικής Μάθησης όπως οι Shapley values ώστε να εντοπίζουν τα δυνατά και αδύνατα σημεία κάθε παίκτη. Με αυτόν τον τρόπο, η μελέτη όχι μόνο παράγει πιο ακριβείς προβλέψεις για τις μελλοντικές επιδόσεις παικτών, αλλά προσφέρει και βαθύτερη εικόνα για τα χαρακτηριστικά που διαφοροποιούν τους πιο επιτυχημένους μελλοντικούς NBA παίκτες.

Riccardi et al. [3] μελέτησαν το πρόβλημα της σύνθεσης ρόστερ στο NBA, εστιάζοντας στο πώς οι διαφορετικοί τύποι παικτών συνδυάζονται ώστε να μεγιστοποιηθεί η αγωνιστική επιτυχία μιας ομάδας. Αντί να χρησιμοποιήσουν τις παραδοσιακές θέσεις των παικτών, εφάρμοσαν μεθόδους ομαδοποίησης (clustering) βασισμένες στα αγωνιστικά στατιστικά, με σκοπό να προσδιορίσουν λειτουργικούς ρόλους και αγωνιστικά προφίλ παικτών. Στη συνέχεια, ανέλυσαν τη σύνθεση των ρόστερ εξετάζοντας πώς οι συνδυασμοί διαφορετικών τύπων παικτών σχετίζονται με την ομαδική απόδοση, χρησιμοποιώντας στατιστικά μοντέλα παλινδρόμησης για να εκτιμήσουν τη συμβολή της συμπληρωματικότητας των παικτών. Τα αποτελέσματα δείχνουν ότι η επιτυχία μιας ομάδας δεν εξαρτάται μόνο από την παρουσία κορυφαίων παικτών, αλλά και από τον τρόπο με τον οποίο αυτοί πλαισιώνονται από παίκτες με κατάλληλα και συμπληρωματικά χαρακτηριστικά, αναδεικνύοντας τη σημασία της ανάλυσης δεδομένων στη λήψη αποφάσεων για το χτίσιμο του ρόστερ..

Ο Soliman et al. [14] παρουσίασαν μια προσέγγιση αξιολόγησης παικτών στο NBA με στόχο την πρόβλεψη της επιλογής ενός παίκτη ως All-Star, χρησιμοποιώντας αλγόριθμο Random Forest. Η μελέτη βασίζεται σε αγωνιστικά στατιστικά παικτών, όπως πόντοι, ριμπάουντ, ασίστ και δείκτες απόδοσης προσαρμοσμένους στον χρόνο συμμετοχής, τα οποία χρησιμοποιούνται ως χαρακτηριστικά εισόδου στο μοντέλο ταξινόμησης. Το πρόβλημα διατυπώνεται ως δυαδική ταξινόμηση (All-Star / μη All-Star), με στόχο να αποτυπωθεί ποιοι παίκτες ξεχωρίζουν βάσει της συνολικής τους συνεισφοράς στο παιχνίδι. Τα αποτελέσματα δείχνουν ότι το Random Forest επιτυγχάνει υψηλή ακρίβεια πρόβλεψης, γεγονός που αναδεικνύει τη χρησιμότητα των μεθόδων μηχανικής μάθησης ως εργαλείων αξιολόγησης παικτών. Παράλληλα, η ανάλυση των χαρακτηριστικών που συμβάλλουν περισσότερο στις προβλέψεις προσφέρει χρήσιμες ενδείξεις για το ποια στατιστικά στοιχεία σχετίζονται περισσότερο με την αναγνώριση κορυφαίων παικτών, καθιστώντας την προσέγγιση σχετική και με διαδικασίες scouting.

2.2.2 Εφαρμογές Μηχανικής Μάθησης στην Πρόβλεψη Αγώνων Μπάσκετ

Η πρόβλεψη του αποτελέσματος αγώνων αποτελεί μία από τις βασικές εφαρμογές της μηχανικής μάθησης στο μπάσκετ, με έμφαση στη μοντελοποίηση της απόδοσης των ομάδων και στην ανάλυση ιστορικών και αγωνιστικών δεδομένων. Στη βιβλιογραφία έχουν προταθεί διάφορες προσεγγίσεις που αξιοποιούν ομαδικά στατιστικά, προηγμένα μετρικά και πληροφορίες αγωνιστικού πλαισίου, με στόχο τη βελτίωση της ακρίβειας των προβλέψεων. Στις μελέτες που ακολουθούν παρουσιάζονται ενδεικτικά παραδείγματα εφαρμογής τέτοιων μοντέλων στην πρόβλεψη αποτελεσμάτων αγώνων μπάσκετ.

Caliwag et al. [14] πρότειναν μια μέθοδο πρόβλεψης αποτελεσμάτων αγώνων NBA βασισμένη σε έναν συνδυασμό πολλαπλών αλγορίθμων μηχανικής μάθησης, την οποία ονόμασαν cascading algorithm. Η προσέγγισή τους ενσωματώνει διαδοχικά τρεις τεχνικές: Naive Bayes για την εκτίμηση της πιθανότητας νίκης βάσει ιστορικού αποτελεσμάτων, Four Factor Analysis για την αποτύπωση βασικών αγωνιστικών δεικτών που σχετίζονται με τη νίκη (σουτ, λάθη, ριμπάουντ και ελεύθερες βολές) και Fuzzy Logic για τη συνδυαστική αξιολόγηση των επιμέρους αποτελεσμάτων. Το μοντέλο εκπαιδεύτηκε και αξιολογήθηκε σε δεδομένα της κανονικής περιόδου NBA 2015–2016, χρησιμοποιώντας στατιστικά από τους τρεις πιο πρόσφατους αγώνες κάθε ομάδας, επιλογή που αποδείχθηκε πιο αποτελεσματική σε σχέση με μεγαλύτερα χρονικά παράθυρα. Τα αποτελέσματα δείχνουν ότι η προτεινόμενη cascading προσέγγιση επιτυγχάνει ακρίβεια πρόβλεψης περίπου 70%, υπερέχοντας σε σύγκριση με μεμονωμένα μοντέλα όπως SVM, Logistic Regression και Naive Bayes. Η μελέτη αναδεικνύει ότι ο συνδυασμός διαφορετικών μεθόδων μπορεί να βελτιώσει την αξιοπιστία της πρόβλεψης αποτελεσμάτων και να λειτουργήσει ως εργαλείο υποστήριξης αποφάσεων για προπονητές και αναλυτές.

Leicht et al. [15] μελέτησαν τους παράγοντες που εξηγούν το αποτέλεσμα αγώνων στο ανδρικό τουρνουά μπάσκετ των Ολυμπιακών Αγώνων, αναλύοντας επίσημα στατιστικά αγώνων από διαφορετικές διοργανώσεις. Η προσέγγισή τους βασίζεται στη σύγκριση νικητών και ηττημένων ομάδων μέσω ομαδικών αγωνιστικών δεικτών, όπως η ευστοχία στα σουτ, τα ριμπάουντ, οι ασίστ, τα λάθη και οι βολές, με στόχο να εντοπιστούν τα χαρακτηριστικά που σχετίζονται περισσότερο με τη νίκη. Τα αποτελέσματα δείχνουν ότι η αποτελεσματικότητα στην επίθεση, και ιδιαίτερα η ευστοχία στα σουτ εντός πεδιάς, αποτελεί τον σημαντικότερο παράγοντα διαφοροποίησης μεταξύ νικητών και ηττημένων, ενώ δευτερεύοντα ρόλο παίζουν τα αμυντικά ριμπάουντ και ο έλεγχος των λαθών. Η μελέτη προσφέρει μια ερμηνευτική προσέγγιση στην ανάλυση αποτελέσματος αγώνων, αναδεικνύοντας ποιοι

ομαδικοί δείκτες συνδέονται περισσότερο με την επιτυχία σε αγώνες υψηλού επιπέδου, και λειτουργεί συμπληρωματικά προς πιο καθαρά προβλεπτικά μοντέλα μηχανικής μάθησης.

Ozkan et al. [16] πρότειναν ένα υβριδικό μοντέλο πρόβλεψης αποτελεσμάτων αγώνων μπάσκετ, συνδυάζοντας τεχνητά νευρωνικά δίκτυα με ασαφή λογική σε ένα Concurrent Neuro-Fuzzy System (CNFS). Η μελέτη βασίζεται σε δεδομένα από τη σεζόν 2015–2016 του Τουρκικού Πρωταθλήματος Μπάσκετ και αξιοποιεί ομαδικά αγωνιστικά χαρακτηριστικά, όπως η πρόσφατη απόδοση των ομάδων, η θέση στη βαθμολογία και η ποιότητα των αντιπάλων, με στόχο την πρόβλεψη του νικητή κάθε αγώνα. Αρχικά αναπτύσσεται ένα μοντέλο τεχνητού νευρωνικού δικτύου, το οποίο επιτυγχάνει ακρίβεια περίπου 70,8%, ενώ στη συνέχεια ενσωματώνεται ένα σύστημα ασαφούς λογικής που εκτιμά ποια ομάδα θεωρείται φαβορί, βασισμένο σε κανόνες που προσομοιώνουν ανθρώπινη κρίση. Ο συνδυασμός των δύο προσεγγίσεων στο CNFS οδηγεί σε σημαντική βελτίωση της ακρίβειας πρόβλεψης, η οποία φτάνει το 79,2%, αναδεικνύοντας ότι τα υβριδικά μοντέλα μπορούν να υπερέχουν έναντι μεμονωμένων μεθόδων στην πρόβλεψη αποτελεσμάτων αγώνων μπάσκετ.

Συγγραφείς	**Άθλημα**	**Πεδίο Εφαρμογής**	**Δεδομένα**	**Μεθοδολογία / Τεχνικές**
Cai et al. [7]	Χόκεϊ	Αναγνώριση αθλητικών ενεργειών	Βίντεο αγώνων	Pose Estimation (Part Affinity Fields), Optical Flow (LiteFlowNet), Two-stream Deep Learning, Action Classification
Piergiovanni & Ryoo [8]	Baseball	Αναγνώριση λεπτομερών αγωνιστικών ενεργειών	Broadcast βίντεο (MLB-YouTube dataset)	Spatio-temporal Deep Learning, Temporal Pooling
Merckx et al. [9]	Ποδόσφαιρο	Ανίχνευση και αξιολόγηση pressing	Tracking δεδομένα & event data	Rule-based detection, Probabilistic models, Passing & Possession Prediction
Almujahed et al. [10]	Βόλεϊ	Υποστήριξη αποφάσεων και ανάλυση αγώνα	Βίντεο αγώνων	Video Analysis, Big Data Processing, Εξαγωγή στατιστικών δεικτών
Zhang et al. [11]	NBA	Κατηγοριοποίηση παικτών (guards)	Αγωνιστικά στατιστικά	K-means Clustering
Patton et al. [12]	Μπάσκετ (NCAA/NBA)	Πρόβλεψη μελλοντικής καριέρας NBA	Tracking δεδομένα από broadcast βίντεο	Machine Learning, Feature Extraction, Explainable ML (Shapley Values)
Riccardi et al. [3]	NBA	Σύνθεση ρόστερ και συμπληρωματικότητα παικτών	Αγωνιστικά στατιστικά	Clustering, Regression Analysis
Soliman et al. [14]	NBA	Αξιολόγηση παικτών (All-Star prediction)	Αγωνιστικά στατιστικά	Random Forest Classification

Πίνακας 2.1: Σύνοψη σχετικών ερευνών που εφαρμόζουν τεχνικές μηχανικής μάθησης στον αθλητισμό.

2.2.3 Βασικοί αλγόριθμοι μάθησης στα Sports Analytics

Η ραγδαία εξέλιξη της ανάλυσης αθλητικών δεδομένων τα τελευταία χρόνια έχει καταστήσει τη μηχανική μάθηση αναπόσπαστο μέρος των σύγχρονων συστημάτων sports analytics. Η αυξανόμενη διαθεσιμότητα αγωνιστικών στατιστικών, tracking δεδομένων και οπτικοακουστικού υλικού δημιούργησε την ανάγκη για μεθόδους που μπορούν να επεξεργαστούν μεγάλο όγκο πληροφοριών και να αποκαλύψουν πρότυπα τα οποία δεν είναι άμεσα εμφανή μέσω παραδοσιακών στατιστικών προσεγγίσεων. Στο πλαίσιο αυτό, έχει διαμορφωθεί ένα σύνολο αλγορίθμων μηχανικής μάθησης που χρησιμοποιούνται κατά κόρον στον αθλητισμό, τόσο για την αξιολόγηση παικτών και ομάδων όσο και για την υποστήριξη στρατηγικών και αγωνιστικών αποφάσεων. Στην παρούσα ενότητα παρουσιάζονται και αναλύονται πέντε βασικοί αλγόριθμοι που αποτελούν τη βάση των σύγχρονων εφαρμογών sports analytics, αναδεικνύοντας τον ρόλο, τα πλεονεκτήματα και τις τυπικές περιπτώσεις χρήσης τους στον χώρο της αθλητικής ανάλυσης.

Γραμμική Παλινδρόμηση (Linear Regression)

Η γραμμική παλινδρόμηση αποτελεί έναν από τους πλέον θεμελιώδεις και ευρέως χρησιμοποιούμενους αλγορίθμους στα sports analytics, κυρίως λόγω της απλότητάς της και της υψηλής ερμηνευσιμότητας που προσφέρει. Χρησιμοποιείται για την ποσοτικοποίηση της σχέσης μεταξύ ενός συνόλου αγωνιστικών χαρακτηριστικών και μιας μεταβλητής-στόχου, όπως η συνολική αγωνιστική επίδραση ενός παίκτη, το Net Rating ή το Plus-Minus. Μέσω της εκτίμησης των συντελεστών του μοντέλου, είναι δυνατή η κατανόηση του τρόπου με τον οποίο κάθε στατιστικό χαρακτηριστικό συμβάλλει στην τελική απόδοση, γεγονός που καθιστά τη γραμμική παλινδρόμηση ιδιαίτερα χρήσιμη σε περιβάλλοντα όπου η ερμηνεία των αποτελεσμάτων είναι εξίσου σημαντική με την ακρίβεια των προβλέψεων. Παρά τους περιορισμούς της σε περιπτώσεις μη γραμμικών σχέσεων, η γραμμική παλινδρόμηση παραμένει βασικό εργαλείο ανάλυσης και συχνά χρησιμοποιείται ως baseline μοντέλο σε μελέτες αθλητικής απόδοσης.

K-Means Clustering

Ο αλγόριθμος K-Means clustering χρησιμοποιείται εκτενώς στα sports analytics για την ομαδοποίηση παικτών ή ομάδων σε διακριτές κατηγορίες με βάση την ομοιότητα των στατιστικών τους χαρακτηριστικών. Πρόκειται για μέθοδο μη επιβλεπόμενης μάθησης, η οποία δεν βασίζεται σε προκαθορισμένες ετικέτες ή αγωνιστικούς ρόλους, αλλά επιτρέπει στα ίδια τα δεδομένα να αποκαλύψουν φυσικές δομές και πρότυπα. Στην πράξη, ο K-Means αξιοποιείται για την αναγνώριση αγωνιστικών προφίλ, όπως scorers, playmakers ή αμυντικών ειδικών, προσφέροντας μια πιο αντικειμενική και δεδομενοκεντρική προσέγγιση σε σχέση με τις παραδοσιακές θέσεις. Η απλότητα και η υπολογιστική αποδοτικότητα του αλγορίθμου τον καθιστούν ιδιαίτερα δημοφιλή, ειδικά σε περιβάλλοντα με μεγάλο όγκο δεδομένων, αν και η ανάγκη προκαθορισμού του αριθμού των clusters αποτελεί ένα από τα βασικά του μειονεκτήματα.

Gaussian Mixture Models (GMM)

Τα Gaussian Mixture Models αποτελούν μια πιο ευέλικτη και πιθανοτική προσέγγιση στην ομαδοποίηση δεδομένων και χρησιμοποιούνται ολοένα και περισσότερο στα sports analytics για την

ανάλυση αγωνιστικών προφίλ. Σε αντίθεση με τον K-Means, ο οποίος αναθέτει κάθε παρατήρηση σε ένα και μόνο cluster, τα GMM επιτρέπουν τη μερική συμμετοχή ενός παίκτη σε περισσότερες από μία ομάδες, εκφράζοντας την αβεβαιότητα μέσω πιθανοτήτων. Αυτή η ιδιότητα είναι ιδιαίτερα χρήσιμη στον αθλητισμό, όπου οι αγωνιστικοί ρόλοι συχνά αλληλεπικαλύπτονται και δεν είναι αυστηρά διακριτοί. Τα GMM χρησιμοποιούνται για την αναγνώριση σύνθετων αγωνιστικών προφίλ και για την αποτύπωση της ποικιλομορφίας της απόδοσης των παικτών, προσφέροντας βαθύτερη κατανόηση της δομής των δεδομένων σε σύγκριση με απλούστερες μεθόδους clustering.

Random Forest

Ο αλγόριθμος Random Forest ανήκει στην κατηγορία των ensemble μεθόδων και χρησιμοποιείται ευρέως στα sports analytics για προβλήματα ταξινόμησης και πρόβλεψης. Βασίζεται στη δημιουργία πολλαπλών δέντρων αποφάσεων, τα οποία συνδυάζονται ώστε να παραχθεί ένα πιο σταθερό και αξιόπιστο τελικό αποτέλεσμα. Η μέθοδος αυτή έχει αποδειχθεί ιδιαίτερα αποτελεσματική σε δεδομένα υψηλής διάστασης και με θόρυβο, χαρακτηριστικό που συναντάται συχνά σε αθλητικά δεδομένα. Στην πράξη, το Random Forest χρησιμοποιείται για την πρόβλεψη τραυματισμών, την αξιολόγηση παικτών, την εκτίμηση πιθανότητας νίκης ή την πρόβλεψη μελλοντικής απόδοσης. Επιπλέον, παρέχει μετρικές σημασίας χαρακτηριστικών, οι οποίες προσφέρουν χρήσιμες ενδείξεις για το ποιοι παράγοντες επηρεάζουν περισσότερο τα αποτελέσματα, ενισχύοντας έτσι τη χρησιμότητά του σε διαδικασίες ανάλυσης και scouting.

Νευρωνικά Δίκτυα και Deep Learning

Τα νευρωνικά δίκτυα και ειδικότερα οι μέθοδοι βαθιάς μάθησης (Deep Learning) έχουν αποκτήσει κεντρικό ρόλο στα σύγχρονα sports analytics, ιδιαίτερα σε εφαρμογές που βασίζονται σε δεδομένα υψηλής πολυπλοκότητας, όπως tracking data και βίντεο αγώνων. Τα μοντέλα αυτά είναι ικανά να μαθαίνουν πολύπλοκες μη γραμμικές σχέσεις και να εξάγουν αυτόματα χωρικά και χρονικά χαρακτηριστικά χωρίς την ανάγκη χειροκίνητου feature engineering. Χρησιμοποιούνται σε εφαρμογές όπως η ανάλυση κινήσεων παικτών, η αναγνώριση αγωνιστικών ενεργειών και η πρόβλεψη αγωνιστικών καταστάσεων σε πραγματικό χρόνο. Παρότι απαιτούν σημαντικούς υπολογιστικούς πόρους και μεγάλα σύνολα δεδομένων, τα νευρωνικά δίκτυα προσφέρουν υψηλή ακρίβεια και ανοίγουν νέες δυνατότητες στην κατανόηση και ανάλυση του αθλητικού παιχνιδιού.

Κεφάλαιο 3ο: Πηγές Δεδομένων

3.1 Εισαγωγή

Οι μέθοδοι ανάλυσης δεδομένων και μηχανικής μάθησης που παρουσιάστηκαν στις προηγούμενες ενότητες βασίζονται σε διαφορετικά είδη δεδομένων, τα οποία αποτελούν τον πυρήνα κάθε αναλυτικής προσέγγισης στον αθλητισμό. Στον χώρο του μπάσκετ, η διαθεσιμότητα πλούσιων και ετερογενών δεδομένων — από παραδοσιακά στατιστικά αγώνων έως δεδομένα κίνησης και tracking — έχει επιτρέψει την ανάπτυξη και την εφαρμογή προηγμένων μοντέλων ανάλυσης και πρόβλεψης. Στην παρούσα ενότητα παρουσιάζονται τα βασικά είδη δεδομένων που χρησιμοποιούνται στη βιβλιογραφία, οι κύριες πηγές από τις οποίες αντλούνται, καθώς και οι πηγές δεδομένων που θα αξιοποιηθούν στη συγκεκριμένη διπλωματική εργασία.

3.2 Κατηγορίες και Μορφές Δεδομένων στα Basketball Analytics

Τα δεδομένα που προκύπτουν από την καταγραφή των αγώνων αποτελούν μία από τις βασικότερες κατηγορίες πληροφορίας στο μπάσκετ και περιλαμβάνουν τόσο τα παραδοσιακά στατιστικά (box score), όσο και τα play-by-play δεδομένα και τους προηγμένους δείκτες απόδοσης. Τα παραδοσιακά στατιστικά, όπως πόντοι, ριμπάουντ, ασίστ, λάθη και ποσοστά ευστοχίας, αποτελούν τη συνοπτική αποτύπωση της αγωνιστικής εικόνας ενός αγώνα ή ενός παίκτη και είναι άμεσα διαθέσιμα για κάθε διοργάνωση. Η απλότητα και η τυποποιημένη μορφή τους τα καθιστούν ιδιαίτερα εύχρηστα, τόσο για βασικές αναλύσεις όσο και ως πρώτη ύλη για πιο σύνθετες αναλυτικές προσεγγίσεις.

Πέρα από τη συνοπτική αυτή καταγραφή, τα play-by-play δεδομένα περιγράφουν τη χρονική αλληλουχία όλων των αγωνιστικών γεγονότων που λαμβάνουν χώρα κατά τη διάρκεια ενός αγώνα, όπως σουτ, φάουλ, λάθη και αλλαγές κατοχής. Η μορφή αυτή δεδομένων επιτρέπει την ανάλυση της ροής του παιχνιδιού σε επίπεδο κατοχής, τη μελέτη κρίσιμων χρονικών διαστημάτων και την κατανόηση της εξέλιξης του σκορ μέσα στον χρόνο. Σε αντίθεση με τα box score στατιστικά, τα play-by-play δεδομένα διατηρούν το αγωνιστικό πλαίσιο κάθε ενέργειας, γεγονός που τα καθιστά ιδιαίτερα χρήσιμα σε προβλεπτικά μοντέλα και σε αναλύσεις στρατηγικής.

Chen et al. [17] παρουσίασαν μια ολοκληρωμένη πλατφόρμα διαχείρισης και ανάλυσης μεγάλου όγκου δεδομένων μπάσκετ, εστιάζοντας στη συλλογή, αποθήκευση και επεξεργασία τόσο box score όσο και play-by-play δεδομένων. Η προτεινόμενη πλατφόρμα έχει σχεδιαστεί ώστε να υποστηρίζει την ανάλυση δεδομένων σε μεγάλη κλίμακα, αντιμετωπίζοντας προκλήσεις που σχετίζονται με τον όγκο, την ταχύτητα και την ετερογένεια των αγωνιστικών δεδομένων. Μέσω της ενοποίησης διαφορετικών τύπων δεδομένων αγώνα, η πλατφόρμα επιτρέπει τη δημιουργία πιο σύνθετων αναλυτικών ερωτημάτων, την εξαγωγή προηγμένων στατιστικών και τη στήριξη εφαρμογών ανάλυσης απόδοσης και πρόβλεψης αποτελεσμάτων. Η εργασία αναδεικνύει τη σημασία της κατάλληλης υποδομής δεδομένων στο basketball analytics, δείχνοντας ότι η αποτελεσματική αξιοποίηση box score και play-by-play δεδομένων δεν εξαρτάται μόνο από τους αλγορίθμους, αλλά και από τον τρόπο οργάνωσης και διαχείρισης των δεδομένων.

Οι προηγμένοι δείκτες απόδοσης (advanced statistics) αποτελούν εξέλιξη των παραπάνω μορφών δεδομένων, καθώς συνδυάζουν πληροφορία από τα box score και τα play-by-play δεδομένα με σκοπό

την πληρέστερη αποτίμηση της αγωνιστικής συνεισφοράς παικτών και ομάδων. Μετρικές όπως το Player Efficiency Rating (PER), το Plus-Minus και οι προσαρμοσμένες εκδοχές του επιχειρούν να διορθώσουν περιορισμούς των παραδοσιακών στατιστικών, λαμβάνοντας υπόψη τον χρόνο συμμετοχής, το πλαίσιο των κατοχών και την επίδραση ενός παίκτη στο συνολικό αποτέλεσμα. Μέσω αυτών των δεικτών, η ανάλυση μεταβαίνει από την απλή καταγραφή γεγονότων στη σύνθετη ερμηνεία της απόδοσης.

Olivo et al. [18] μελέτησαν την εξέλιξη και τη χρησιμότητα των προηγμένων δεικτών απόδοσης στο μπάσκετ, εστιάζοντας στο πώς τα box score και τα play-by-play δεδομένα μπορούν να αξιοποιηθούν για τη δημιουργία πιο σύνθετων και ερμηνευτικών μετρικών. Η εργασία αναλύει τους περιορισμούς των παραδοσιακών στατιστικών, τα οποία συχνά αποτυπώνουν μόνο μέρος της αγωνιστικής συνεισφοράς ενός παίκτη, και παρουσιάζει πώς δείκτες όπως το Plus-Minus, τα on/off στατιστικά και οι προσαρμοσμένες εκδοχές τους επιχειρούν να εκτιμήσουν την πραγματική επίδραση ενός παίκτη στο παιχνίδι. Μέσα από τη χρήση play-by-play δεδομένων, οι advanced metrics λαμβάνουν υπόψη το αγωνιστικό πλαίσιο, τις κατοχές και τη σύνθεση των πεντάδων, προσφέροντας μια πιο ολοκληρωμένη εικόνα της απόδοσης. Η μελέτη αναδεικνύει ότι οι προηγμένοι δείκτες δεν αντικαθιστούν τα βασικά στατιστικά, αλλά τα συμπληρώνουν, αποτελώντας κρίσιμο εργαλείο για αξιολόγηση παικτών, ανάλυση ομάδων και εφαρμογές μηχανικής μάθησης στο basketball analytics.

Συνολικά, τα δεδομένα καταγραφής αγώνων συνιστούν ένα ενιαίο οικοσύστημα πληροφορίας, όπου τα box score παρέχουν τη βασική εικόνα, τα play-by-play προσθέτουν χρονική και αγωνιστική λεπτομέρεια και οι advanced metrics επιχειρούν να συνθέσουν μια πιο ολοκληρωμένη εκτίμηση της απόδοσης. Η συνδυαστική αξιοποίησή τους αποτελεί κοινή πρακτική στη βιβλιογραφία του basketball analytics και προσφέρει ένα ισχυρό υπόβαθρο για εφαρμογές μηχανικής μάθησης, όπως η αξιολόγηση παικτών, η ανάλυση ομάδων και η πρόβλεψη αγωνιστικών αποτελεσμάτων.

3.3 Tracking και video-based Δεδομένα

Πέρα από τα δεδομένα που προκύπτουν από την καταγραφή αγωνιστικών γεγονότων, η ανάλυση του μπάσκετ έχει εμπλουτιστεί σημαντικά τα τελευταία χρόνια με δεδομένα που αποτυπώνουν άμεσα την κίνηση των παικτών και της μπάλας στον αγωνιστικό χώρο. Όπως παρουσιάστηκε σε προηγούμενες ενότητες μέσα από παραδείγματα από άλλα αθλήματα, τα δεδομένα αυτά προέρχονται είτε από συστήματα tracking είτε από την επεξεργασία βιντεοσκοπημένου υλικού και παρέχουν λεπτομερή χωροχρονική πληροφορία σχετικά με τη θέση, την ταχύτητα και τα μοτίβα κίνησης κατά τη διάρκεια του αγώνα.

Στην παρούσα ενότητα, η ανάλυση επικεντρώνεται πλέον στο μπάσκετ, εξετάζοντας πώς τα tracking και video-based δεδομένα εφαρμόζονται σε αυτό το άθλημα και πώς συμβάλλουν στην κατανόηση πτυχών του παιχνιδιού που δεν μπορούν να αποτυπωθούν από τα παραδοσιακά στατιστικά ή τα play-by-play δεδομένα. Μέσω αυτών των δεδομένων καθίσταται δυνατή η μελέτη στοιχείων όπως το spacing, η αμυντική τοποθέτηση, οι κινήσεις χωρίς την μπάλα και οι αλληλεπιδράσεις μεταξύ παικτών, ανοίγοντας τον δρόμο για πιο προηγμένες αναλυτικές και προβλεπτικές εφαρμογές.

Τα δεδομένα tracking και video-based αποτελούν μία από τις πιο προηγμένες και πλούσιες μορφές πληροφορίας στο σύγχρονο basketball analytics, καθώς επιτρέπουν την άμεση καταγραφή της χωροχρονικής εξέλιξης του παιχνιδιού. Τα tracking δεδομένα περιλαμβάνουν τις συντεταγμένες θέσης των παικτών και της μπάλας σε διαδοχικά χρονικά στιγμιότυπα, παρέχοντας λεπτομερή πληροφορία σχετικά με την κίνηση, την ταχύτητα και τις αλληλεπιδράσεις εντός του αγωνιστικού χώρου. Στο μπάσκετ, τα δεδομένα αυτά συλλέγονται είτε μέσω εξειδικευμένων συστημάτων πολλαπλών καμερών είτε μέσω τεχνικών υπολογιστικής όρασης που εφαρμόζονται σε βιντεοσκοπημένο υλικό αγώνων.

Η ανάλυση tracking δεδομένων επιτρέπει τη μελέτη πτυχών του παιχνιδιού που δεν μπορούν να αποτυπωθούν από τα παραδοσιακά στατιστικά ή τα play-by-play δεδομένα. Μέσω αυτών καθίσταται δυνατή η ποσοτική αξιολόγηση εννοιών όπως το spacing, η αμυντική τοποθέτηση, η κάλυψη χώρου, οι off-ball κινήσεις και η δυναμική των πεντάδων. Επιπλέον, τα δεδομένα κίνησης χρησιμοποιούνται για την ανάλυση τακτικών επιλογών, όπως τα pick-and-rolls, οι αμυντικές περιστροφές και η δημιουργία πλεονεκτημάτων μέσα από τη συνεργασία των παικτών, προσφέροντας μια πιο λεπτομερή εικόνα της ομαδικής συμπεριφοράς.

Miller et al. [19] παρουσίασαν μια μέθοδο αξιοποίησης δεδομένων player tracking στο μπάσκετ με στόχο την εκμάθηση ατομικών δεξιοτήτων παικτών και την πρόβλεψη της απόδοσης των ομάδων. Χρησιμοποιώντας χωροχρονικά δεδομένα θέσης παικτών και μπάλας, οι συγγραφείς ανέπτυξαν ένα μοντέλο που αποσυνθέτει την αγωνιστική συμπεριφορά σε επιμέρους δεξιότητες, όπως η ικανότητα δημιουργίας χώρου, η αμυντική τοποθέτηση και η αποτελεσματικότητα σε καταστάσεις κατοχής. Οι δεξιότητες αυτές εκτιμώνται μέσω στατιστικής μοντελοποίησης και στη συνέχεια συνδυάζονται για την πρόβλεψη της συνολικής απόδοσης της ομάδας. Τα αποτελέσματα δείχνουν ότι τα μοντέλα που βασίζονται σε tracking δεδομένα υπερέχουν σε σχέση με προσεγγίσεις που χρησιμοποιούν μόνο παραδοσιακά στατιστικά, αναδεικνύοντας τη δυνατότητα των δεδομένων κίνησης να αποτυπώνουν λεπτές πτυχές της αγωνιστικής συνεισφοράς των παικτών. Η μελέτη καταδεικνύει πώς τα player tracking δεδομένα μπορούν να χρησιμοποιηθούν όχι μόνο για περιγραφική ανάλυση, αλλά και για την εκμάθηση δεξιοτήτων και την πρόβλεψη ομαδικής απόδοσης στο μπάσκετ.

Τα video-based δεδομένα αποτελούν την πρωτογενή πηγή από την οποία συχνά εξάγονται τα tracking δεδομένα. Μέσω τεχνικών υπολογιστικής όρασης, όπως ανίχνευση αντικειμένων, pose estimation και αναγνώριση ενεργειών, το βίντεο μετατρέπεται σε δομημένα δεδομένα που περιγράφουν τόσο τις κινήσεις όσο και τις ενέργειες των παικτών. Στο μπάσκετ, η video-based ανάλυση έχει χρησιμοποιηθεί για την αναγνώριση τύπων σουτ, τη μελέτη αμυντικών συμπεριφορών, καθώς και για την ανάλυση αγωνιστικών καταστάσεων σε πραγματικές συνθήκες αγώνα. Η συνδυαστική χρήση video-based και tracking δεδομένων επιτρέπει τη σύνδεση της χωρικής πληροφορίας με το αγωνιστικό πλαίσιο, προσφέροντας πιο ολοκληρωμένες αναλυτικές δυνατότητες.

Li et al. [20] πρότειναν ένα σύστημα ανάλυσης αγώνων μπάσκετ βασισμένο σε video-based δεδομένα και ελαφριά μοντέλα βαθιάς μάθησης, σχεδιασμένο ώστε να λειτουργεί αποδοτικά σε περιβάλλοντα Internet of Things (IoT). Η προσέγγισή τους αξιοποιεί βιντεοσκοπημένο υλικό αγώνων για την αυτόματη αναγνώριση αγωνιστικών ενεργειών, κινήσεων παικτών και βασικών γεγονότων του παιχνιδιού, χρησιμοποιώντας lightweight deep learning αρχιτεκτονικές που μειώνουν το υπολογιστικό κόστος χωρίς σημαντική απώλεια ακρίβειας. Το σύστημα επιτρέπει την εξαγωγή δομημένων δεδομένων από βίντεο σε σχεδόν πραγματικό χρόνο, υποστηρίζοντας εφαρμογές ανάλυσης απόδοσης και τακτικής χωρίς την ανάγκη εξειδικευμένων tracking υποδομών. Μέσω της προτεινόμενης αρχιτεκτονικής, το

βίντεο αναδεικνύεται ως αυτόνομη πηγή δομημένων δεδομένων, ικανή να υποστηρίξει εφαρμογές ανάλυσης απόδοσης στο μπάσκετ.

Παρά τα πλεονεκτήματά τους, τα tracking και video-based δεδομένα συνοδεύονται από σημαντικούς περιορισμούς. Η συλλογή και επεξεργασία τους απαιτεί εξειδικευμένο εξοπλισμό, αυξημένη υπολογιστική ισχύ και, στις περισσότερες περιπτώσεις, πρόσβαση σε εμπορικές πλατφόρμες δεδομένων, γεγονός που περιορίζει τη διαθεσιμότητά τους σε ερευνητικά και ακαδημαϊκά περιβάλλοντα. Επιπλέον, η πολυπλοκότητα των δεδομένων αυτών καθιστά απαραίτητη τη χρήση προηγμένων μεθόδων ανάλυσης και μηχανικής μάθησης, αυξάνοντας τις απαιτήσεις σε χρόνο και πόρους.

Συνολικά, τα tracking και video-based δεδομένα προσφέρουν μια εξαιρετικά λεπτομερή και δυναμική αναπαράσταση του παιχνιδιού του μπάσκετ, επιτρέποντας την ανάλυση τόσο της ατομικής όσο και της ομαδικής συμπεριφοράς σε επίπεδο που δεν είναι εφικτό με τα παραδοσιακά δεδομένα αγώνα. Παρότι η χρήση τους δεν είναι πάντοτε εφικτή λόγω περιορισμών πρόσβασης, αποτελούν σημαντικό σημείο αναφοράς στη σύγχρονη βιβλιογραφία και θέτουν το πλαίσιο για την εξέλιξη των αναλυτικών μεθόδων στο μπάσκετ.

3.4 Πηγές Δεδομένων στο Μπασκετ

Τα διαφορετικά είδη δεδομένων που παρουσιάστηκαν στις προηγούμενες ενότητες αντλούνται από ποικίλες πηγές, οι οποίες διαφοροποιούνται σημαντικά ως προς το επίπεδο λεπτομέρειας, τη διαθεσιμότητα, το κόστος πρόσβασης και τις δυνατότητες αξιοποίησής τους. Στον χώρο του μπάσκετ, οι πηγές συλλογής δεδομένων εκτείνονται από επίσημους αθλητικούς οργανισμούς και εμπορικές πλατφόρμες υψηλής εξειδίκευσης έως ανοιχτές βάσεις δεδομένων και προγραμματιστικές διεπαφές (APIs). Η κατανόηση των χαρακτηριστικών, των πλεονεκτημάτων και των περιορισμών κάθε κατηγορίας πηγών είναι απαραίτητη τόσο για την αξιολόγηση της σχετικής βιβλιογραφίας όσο και για τη σωστή ερμηνεία των εφαρμογών ανάλυσης δεδομένων στο μπάσκετ.

Οι επίσημες λίγκες και αθλητικοί οργανισμοί, όπως το NBA, η FIBA και το NCAA, αποτελούν θεμελιώδεις πηγές συλλογής αγωνιστικών δεδομένων. Μέσω των επίσημων ιστοσελίδων τους ή μέσω συνεργαζόμενων παρόχων, παρέχουν παραδοσιακά στατιστικά στοιχεία, αναλυτικά box scores και λεπτομερή play-by-play δεδομένα για κάθε αγώνα. Τα δεδομένα αυτά χαρακτηρίζονται από υψηλή αξιοπιστία, συνέπεια στη δομή τους και τυποποιημένη καταγραφή, γεγονός που τα καθιστά ιδιαίτερα κατάλληλα για συγκριτικές αναλύσεις και μακροχρόνιες μελέτες. Ωστόσο, η πρόσβαση σε πιο προηγμένες μορφές πληροφορίας, όπως δεδομένα tracking ή ανάλυση βίντεο, είναι συνήθως περιορισμένη και συχνά διατίθεται μόνο μέσω εμπορικών συμφωνιών, γεγονός που περιορίζει τη χρήση τους σε ακαδημαϊκά και ανεξάρτητα ερευνητικά περιβάλλοντα.

Ιδιαίτερα σημαντικό ρόλο στη σύγχρονη ανάλυση μπάσκετ διαδραματίζουν οι εμπορικές πλατφόρμες δεδομένων, όπως οι Stats Perform, Second Spectrum και Synergy. Οι πλατφόρμες αυτές προσφέρουν εξαιρετικά λεπτομερή και πλούσια δεδομένα, τα οποία περιλαμβάνουν player και ball tracking, προηγμένες στατιστικές μετρικές, ανάλυση τακτικών μοτίβων και εκτεταμένο βιντεοσκοπημένο υλικό με ετικέτες αγωνιστικών ενεργειών. Τα δεδομένα αυτά αποτελούν τη βάση πολλών σύγχρονων εφαρμογών στον επαγγελματικό αθλητισμό, όπως scouting, ανάλυση απόδοσης και υποστήριξη λήψης

αποφάσεων. Παρόλα αυτά, το υψηλό οικονομικό κόστος και οι περιορισμοί πρόσβασης καθιστούν τις πλατφόρμες αυτές δύσκολα αξιοποιήσιμες στο πλαίσιο ακαδημαϊκών εργασιών, όπου η αναπαραγωγιμότητα και η ελεύθερη πρόσβαση στα δεδομένα αποτελούν βασικές προϋποθέσεις.

Παράλληλα, ιδιαίτερη σημασία για την έρευνα και την εκπαίδευση έχουν οι ανοιχτές πηγές δεδομένων, όπως το Basketball-Reference, καθώς και σύνολα δεδομένων που διατίθενται μέσω πλατφορμών όπως το Kaggle. Οι πηγές αυτές προσφέρουν ελεύθερη πρόσβαση σε box score, advanced statistics και ιστορικά δεδομένα αγώνων και παικτών, επιτρέποντας τη διεξαγωγή αναλύσεων χωρίς οικονομικούς περιορισμούς. Αν και τα δεδομένα αυτά δεν διαθέτουν το επίπεδο λεπτομέρειας των εμπορικών πλατφορμών, αποτελούν αξιόπιστη και ευρέως χρησιμοποιούμενη βάση για εφαρμογές μηχανικής μάθησης, ανάλυση αγωνιστικής απόδοσης και ακαδημαϊκή έρευνα, καθώς διευκολύνουν την αναπαραγωγή και τη σύγκριση ερευνητικών αποτελεσμάτων.

Τέλος, οι προγραμματιστικές διεπαφές (APIs), όπως το NBA API και άλλες sports APIs, παρέχουν έναν ιδιαίτερα ευέλικτο και δυναμικό τρόπο συλλογής δεδομένων σε δομημένη μορφή. Μέσω αυτών καθίσταται δυνατή η αυτοματοποιημένη άντληση στατιστικών παικτών και ομάδων, δεδομένων αγώνων και play-by-play καταγραφών, επιτρέποντας τη δημιουργία προσαρμοσμένων συνόλων δεδομένων ανάλογα με τις ανάγκες της ανάλυσης. Η χρήση APIs προσφέρει μεγαλύτερο έλεγχο στη διαδικασία συλλογής και προεπεξεργασίας των δεδομένων, ωστόσο απαιτεί τεχνική εξοικείωση και προσεκτικό χειρισμό των περιορισμών πρόσβασης και χρήσης. Παρά τους περιορισμούς αυτούς, οι προγραμματιστικές διεπαφές αποτελούν σημαντικό εργαλείο για σύγχρονες ερευνητικές προσεγγίσεις, γεφυρώνοντας το χάσμα μεταξύ ανοιχτών δεδομένων και πιο προηγμένων αναλυτικών εφαρμογών.

Κεφάλαιο 4ο: Πειραματική Υλοποίηση

4.1 Επιλογή και Περιγραφή Δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν στην παρούσα εργασία αντλήθηκαν από τη δημόσια διαθέσιμη βάση στατιστικών του NBA, μέσω της βιβλιοθήκης nba_api της γλώσσας προγραμματισμού Python. Η συγκεκριμένη βιβλιοθήκη παρέχει προγραμματιστική πρόσβαση στα επίσημα endpoints της λίγκας, επιτρέποντας την αξιόπιστη και δομημένη εξαγωγή αγωνιστικών δεδομένων που αφορούν παίκτες, ομάδες και αγωνιστικές περιόδους.

Η συλλογή των δεδομένων πραγματοποιήθηκε για πέντε διαδοχικές αγωνιστικές περιόδους της κανονικής διάρκειας (Regular Season), από τη σεζόν 2020–21 έως και τη σεζόν 2024–25. Η επιλογή ενός πολυετούς χρονικού ορίζοντα κρίθηκε απαραίτητη, ώστε να εξασφαλιστεί ένα επαρκές και στατιστικά αντιπροσωπευτικό δείγμα παικτών, μειώνοντας την επίδραση τυχαίων διακυμάνσεων που ενδέχεται να εμφανίζονται σε μεμονωμένες σεζόν. Παράλληλα, η περίοδος αυτή αντανακλά τη σύγχρονη μορφή του παιχνιδιού, όπως αυτή έχει διαμορφωθεί τα τελευταία χρόνια στο NBA.

Για την υλοποίηση της παρούσας εργασίας χρησιμοποιήθηκαν αποκλειστικά δημόσια διαθέσιμα αγωνιστικά δεδομένα καλαθοσφαίρισης, τα οποία προέρχονται από ανοικτές πηγές και δεν απαιτούν ειδική άδεια χρήσης. Η συλλογή των δεδομένων πραγματοποιήθηκε μέσω προγραμματιστικής διαδικασίας, επιτρέποντας την αυτοματοποιημένη εξαγωγή τόσο βασικών όσο και advanced στατιστικών δεικτών, όπως δείκτες αποδοτικότητας, συμμετοχής και συνολικής αγωνιστικής επίδρασης των παικτών. Η χρήση προγραμματιστικής πρόσβασης διασφαλίζει τη συνέπεια στη δομή των δεδομένων και περιορίζει τον κίνδυνο σφαλμάτων που συχνά προκύπτουν από χειροκίνητες διαδικασίες συλλογής.

Η επιλογή της βιβλιοθήκης `nba_api` έγινε για συγκεκριμένους επιστημονικούς και πρακτικούς λόγους. Αρχικά, η βιβλιοθήκη βασίζεται σε επίσημα και αξιόπιστα δεδομένα, γεγονός που ενισχύει την εγκυρότητα των αποτελεσμάτων της ανάλυσης. Επιπλέον, προσφέρει άμεση πρόσβαση σε εκτενή σύνολα advanced στατιστικών, τα οποία είναι απαραίτητα για την εφαρμογή τεχνικών μηχανικής μάθησης και την εξαγωγή ουσιαστικών συμπερασμάτων. Παράλληλα, η χρήση ενός προγραμματιστικού API επιτρέπει την αναπαραγωγιμότητα της ερευνητικής διαδικασίας, καθώς η συλλογή των δεδομένων μπορεί να επαναληφθεί με τον ίδιο ακριβώς τρόπο σε μελλοντικές μελέτες ή επεκτάσεις της εργασίας.

Τέλος, η επιλογή της συγκεκριμένης προσέγγισης επέτρεψε την πλήρη υλοποίηση της διαδικασίας δημιουργίας του dataset από τον συγγραφέα, χωρίς τη χρήση έτοιμων εμπορικών ή κλειστών πλατφορμών ανάλυσης. Με τον τρόπο αυτό εξασφαλίστηκε πλήρης έλεγχος της ποιότητας, της δομής και της προεπεξεργασίας των δεδομένων, στοιχείο που είναι ιδιαίτερα σημαντικό σε μελέτες που βασίζονται σε στατιστική ανάλυση και αλγορίθμους μηχανικής μάθησης.

4.1.1 Τεχνολογικό Περιβάλλον, Διαμόρφωση Dataset και Επιλογή Αλγορίθμων

Η υλοποίηση της παρούσας εργασίας πραγματοποιήθηκε σε τοπικό υπολογιστικό περιβάλλον, βασισμένο στη γλώσσα προγραμματισμού Python, η οποία αποτελεί ένα από τα πλέον διαδεδομένα εργαλεία για ανάλυση δεδομένων και εφαρμογές μηχανικής μάθησης. Η επιλογή της Python έγινε λόγω της ευελιξίας της, της εκτεταμένης υποστήριξης βιβλιοθηκών για επιστημονικούς υπολογισμούς και της ευρείας χρήσης της στη σύγχρονη ερευνητική και επαγγελματική πρακτική. Η ανάπτυξη και η εκτέλεση των πειραμάτων πραγματοποιήθηκαν σε περιβάλλον Jupyter Notebook, το οποίο επιτρέπει τη διαδραστική συγγραφή κώδικα, την άμεση παρουσίαση ενδιάμεσων αποτελεσμάτων και τη σταδιακή τεκμηρίωση της ερευνητικής διαδικασίας, ενισχύοντας τη διαφάνεια και την αναπαραγωγιμότητα της ανάλυσης.
Για τη διαχείριση, τον καθαρισμό και τον μετασχηματισμό των δεδομένων χρησιμοποιήθηκαν κυρίως οι βιβλιοθήκες pandas και NumPy, οι οποίες προσφέρουν αποδοτικά εργαλεία για τον χειρισμό δομημένων δεδομένων και αριθμητικών υπολογισμών. Η επιλογή αυτών των βιβλιοθηκών επέτρεψε την ομογενοποίηση των δεδομένων που προέρχονταν από διαφορετικές αγωνιστικές περιόδους, καθώς και την εφαρμογή φίλτρων, μετασχηματισμών και ελέγχων ποιότητας. Για την υλοποίηση των αλγορίθμων μηχανικής μάθησης και των στατιστικών μοντέλων χρησιμοποιήθηκε η βιβλιοθήκη scikit-learn, η οποία παρέχει ανοικτού κώδικα και ευρέως αποδεκτές υλοποιήσεις βασικών αλγορίθμων, χωρίς τη χρήση έτοιμων εμπορικών ή κλειστών πλατφορμών ανάλυσης δεδομένων.

Η διαδικασία διαμόρφωσης του τελικού dataset περιλάμβανε συστηματικό καθαρισμό και επιλεκτικό φιλτράρισμα των δεδομένων, με στόχο τη βελτίωση της αξιοπιστίας της ανάλυσης. Συγκεκριμένα, αποκλείστηκαν παίκτες με περιορισμένο χρόνο συμμετοχής ή μικρό αριθμό αγώνων, καθώς τέτοιες περιπτώσεις ενδέχεται να παρουσιάζουν ακραίες ή μη σταθερές τιμές σε ορισμένα στατιστικά μεγέθη. Η απόφαση αυτή περιορίζει τον στατιστικό θόρυβο και μειώνει την πιθανότητα οι αλγόριθμοι να επηρεαστούν από μη αντιπροσωπευτικές παρατηρήσεις, οδηγώντας σε πιο αξιόπιστα και ερμηνεύσιμα αποτελέσματα.

Παράλληλα, δόθηκε έμφαση στη χρήση advanced στατιστικών χαρακτηριστικών, τα οποία αποτυπώνουν την αποδοτικότητα, τη συμμετοχή και τη συνολική αγωνιστική επίδραση των παικτών, σε αντίθεση με απλά αθροιστικά μεγέθη. Η επιλογή των συγκεκριμένων χαρακτηριστικών βασίστηκε στη σχετική βιβλιογραφία και στη διαπίστωση ότι οι δείκτες αποδοτικότητας παρέχουν πιο συγκρίσιμη και ουσιαστική εικόνα της συνεισφοράς ενός παίκτη, ανεξάρτητα από τον χρόνο συμμετοχής ή τον αγωνιστικό του ρόλο. Με τον τρόπο αυτό, το dataset διαμορφώθηκε έτσι ώστε να είναι κατάλληλο για την εφαρμογή μεθόδων μηχανικής μάθησης και στατιστικής ανάλυσης.

Όσον αφορά τη μεθοδολογία ανάλυσης, στην παρούσα εργασία εφαρμόστηκαν αλγόριθμοι μη επιβλεπόμενης μάθησης, συγκεκριμένα οι K-Means clustering και Gaussian Mixture Models, με στόχο την ομαδοποίηση των παικτών σε διακριτά αγωνιστικά προφίλ. Η χρήση μη επιβλεπόμενων μεθόδων επιτρέπει στα δεδομένα να αποκαλύψουν φυσικές δομές και πρότυπα, χωρίς την επιβολή προκαθορισμένων ρόλων ή υποκειμενικών κατηγοριοποιήσεων.

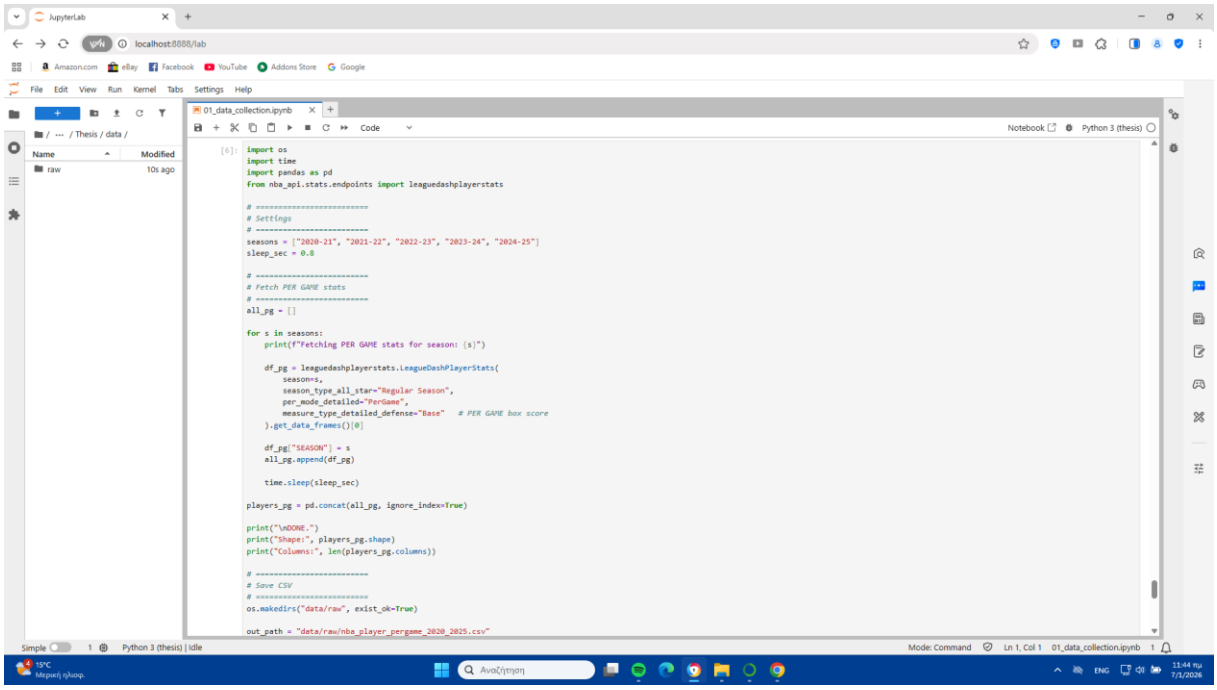
Στη συνέχεια, εφαρμόστηκαν στατιστικά μοντέλα παλινδρόμησης για τη διερεύνηση της σχέσης μεταξύ των επιλεγμένων αγωνιστικών χαρακτηριστικών και του δείκτη Net Rating, ο οποίος χρησιμοποιήθηκε ως μέτρο συνολικής αγωνιστικής επίδρασης. Συγκεκριμένα, χρησιμοποιήθηκε γραμμική παλινδρόμηση μέσω της μεθόδου Ordinary Least Squares (OLS), τόσο στο σύνολο των παικτών όσο και ξεχωριστά

εντός κάθε προκύπτοντος cluster. Η χρήση της OLS επιτρέπει την ποσοτικοποίηση της επίδρασης κάθε χαρακτηριστικού και προσφέρει υψηλό βαθμό ερμηνευσιμότητας, στοιχείο ιδιαίτερα σημαντικό στο πλαίσιο αθλητικής ανάλυσης.

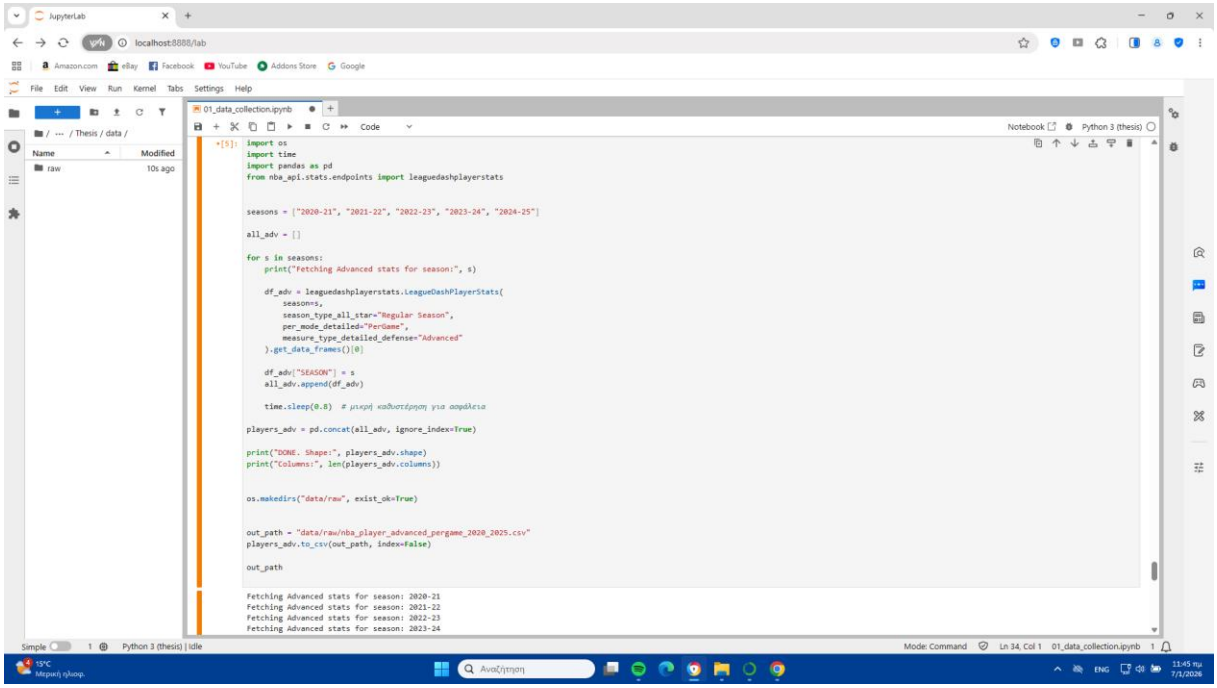
Όλοι οι αλγόριθμοι και τα στατιστικά μοντέλα που χρησιμοποιήθηκαν είναι δημόσια διαθέσιμα και υλοποιημένα σε βιβλιοθήκες ανοικτού κώδικα, ενώ η εκπαίδευση και αξιολόγησή τους πραγματοποιήθηκαν εξ ολοκλήρου στο τοπικό υπολογιστικό περιβάλλον του συγγραφέα. Η επιλογή αυτή ενισχύει τη διαφάνεια, την αναπαραγωγιμότητα και την επιστημονική αξιοπιστία της μεθοδολογίας, και επιτρέπει τη μελλοντική επέκταση της ανάλυσης σε διαφορετικά σύνολα δεδομένων ή αγωνιστικά πλαίσια.

4.1.2 Δημιουργία Dataset

Για κάθε σεζόν συλλέχθηκαν τόσο παραδοσιακά στατιστικά ανά αγώνα (per game box score statistics) όσο και προχωρημένα στατιστικά απόδοσης (advanced statistics). Τα δεδομένα ενοποιήθηκαν σε επίπεδο παίκτη–σεζόν μέσω του μοναδικού αναγνωριστικού παίκτη (PLAYER_ID) και της αντίστοιχης αγωνιστικής περιόδου, δημιουργώντας ένα ενιαίο σύνολο δεδομένων.

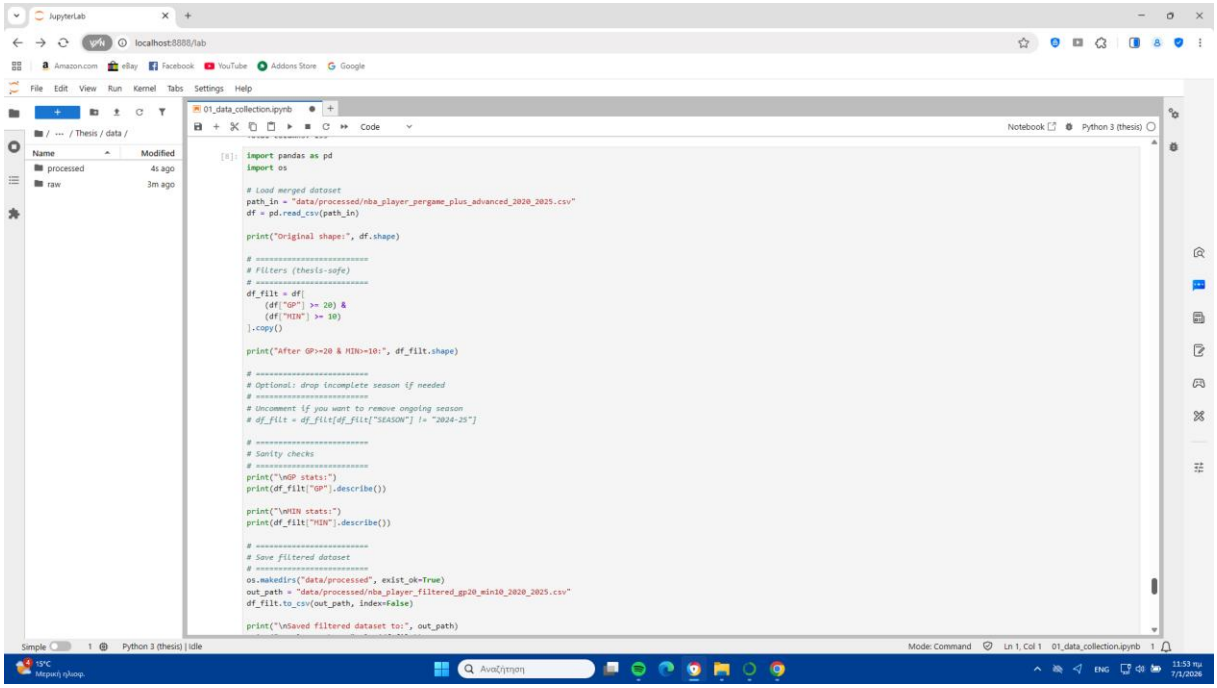


Εικόνα 4.2: Συλλογή Per Game Στατιστικών

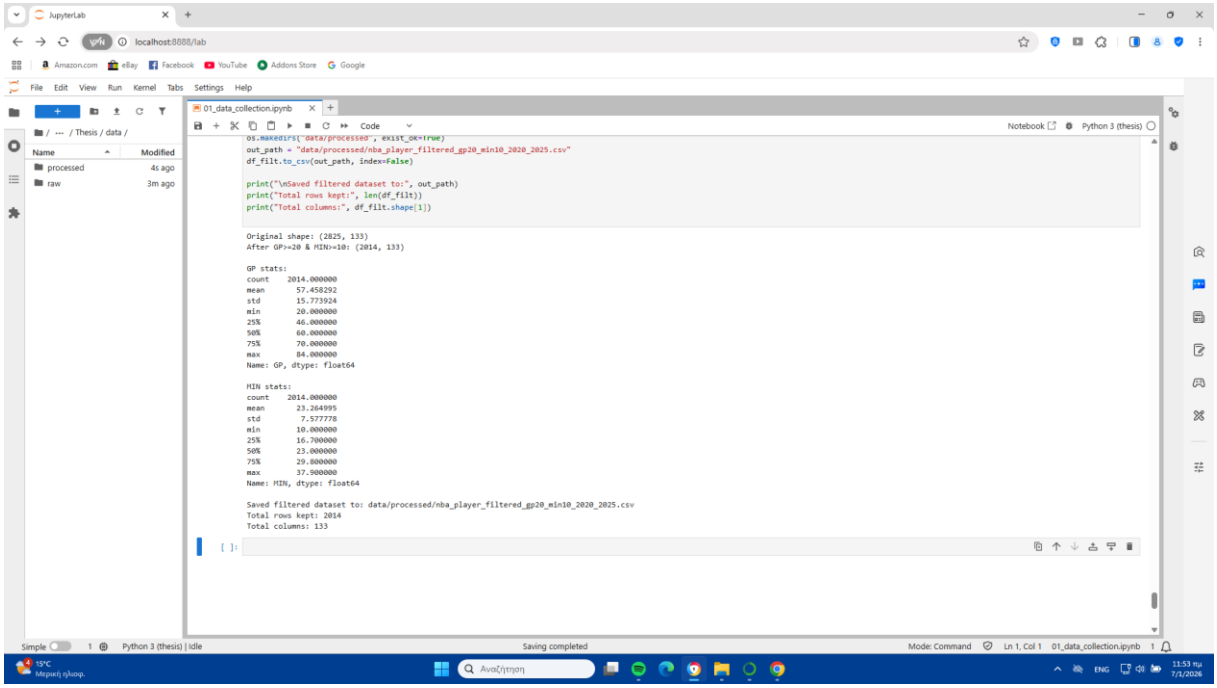


Εικόνα 4.3: Συλλογή Advanced στατιστικών

Στη συνέχεια εφαρμόστηκε φιλτράρισμα με σκοπό τη μείωση του θορύβου στα δεδομένα. Συγκεκριμένα, διατηρήθηκαν μόνο οι παίκτες που συμμετείχαν σε τουλάχιστον 20 αγώνες ανά σεζόν και είχαν μέσο χρόνο συμμετοχής τουλάχιστον 10 λεπτά ανά αγώνα. Το φιλτράρισμα αυτό αποκλείει περιπτώσεις περιστασιακής συμμετοχής, διατηρώντας τους βασικούς παίκτες και τους παίκτες rotation της λίγκας.

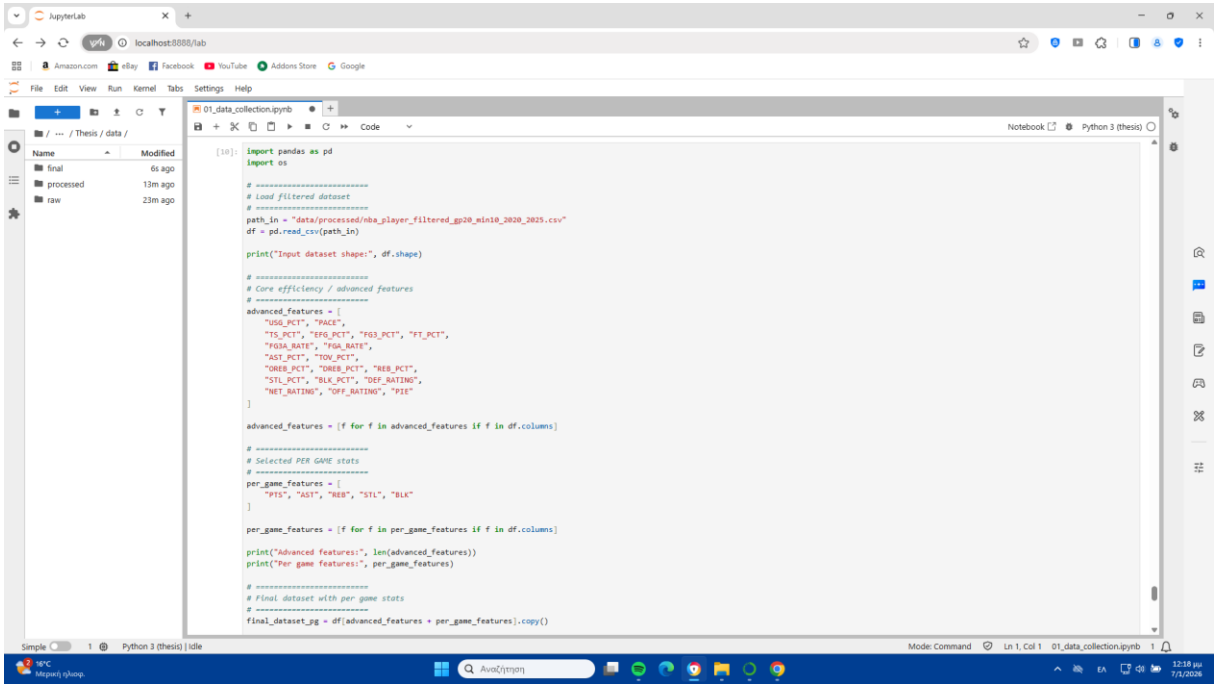


Εικόνα 4.4: Filtering για GP > 20 και MIN > 10

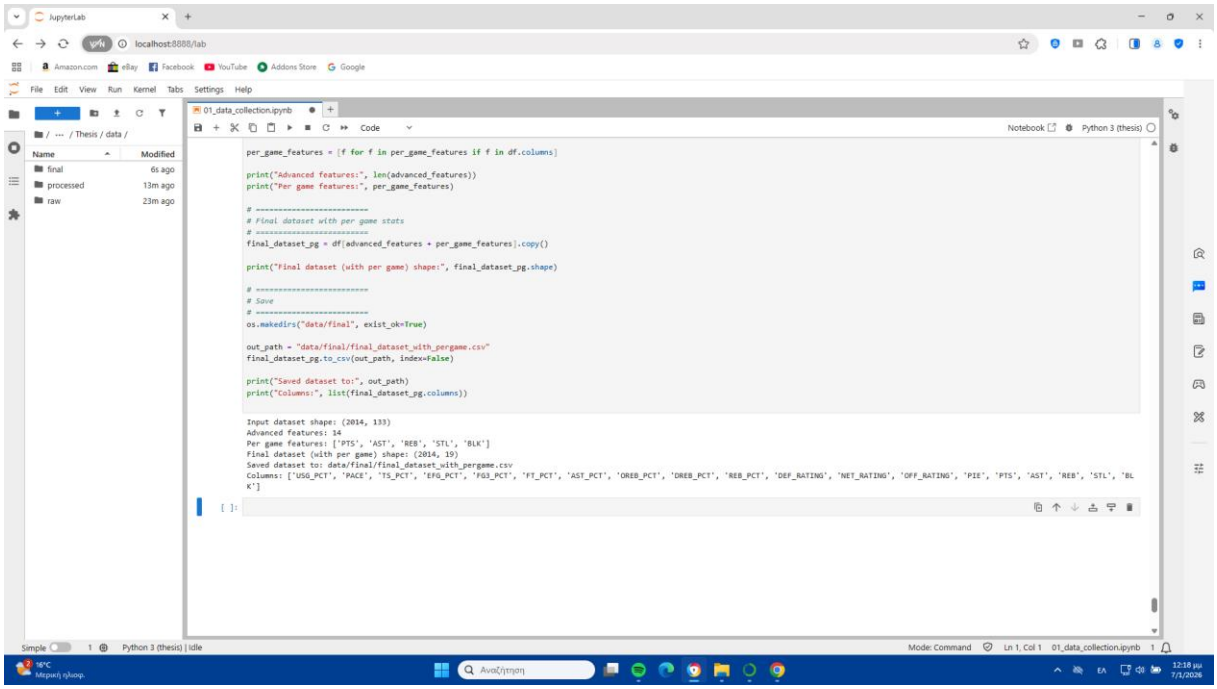


Εικόνα 4.5: Filtering για GP > 20 και MIN > 10

Ακολούθησε διαδικασία επιλογής χαρακτηριστικών (feature selection), με στόχο τη δημιουργία dataset κατάλληλου για εφαρμογή μεθόδων μη επιβλεπόμενης μάθησης. Από το σύνολο των μεταβλητών αφαιρέθηκαν αναγνωριστικά, μεταδεδομένα και μετρήσεις όγκου που εξαρτώνται άμεσα από τον χρόνο συμμετοχής. Αντ' αυτών, επιλέχθηκαν δείκτες αποδοτικότητας, ποσοστιαίας συμμετοχής και συνολικής επίδρασης στο παιχνίδι, ώστε η σύγκριση των παικτών να βασίζεται στον αγωνιστικό τους ρόλο και την αποτελεσματικότητά τους.



Εικόνα 4.6: Feature Selection



Εικόνα 4.7: Feature Selection

Τέλος, δημιουργήθηκαν δύο τελικά σύνολα δεδομένων: ένα βασισμένο αποκλειστικά σε προχωρημένα και ποσοστιαία χαρακτηριστικά, το οποίο χρησιμοποιείται για PCA και clustering, και ένα εναλλακτικό σύνολο που περιλαμβάνει επιλεγμένα per game στατιστικά, το οποίο αξιοποιείται για ελέγχους ευρωστίας και αναλύσεις παλινδρόμησης.

4.2 Ανάλυση Διαστάσεων και Ομαδοποίηση Παικτών

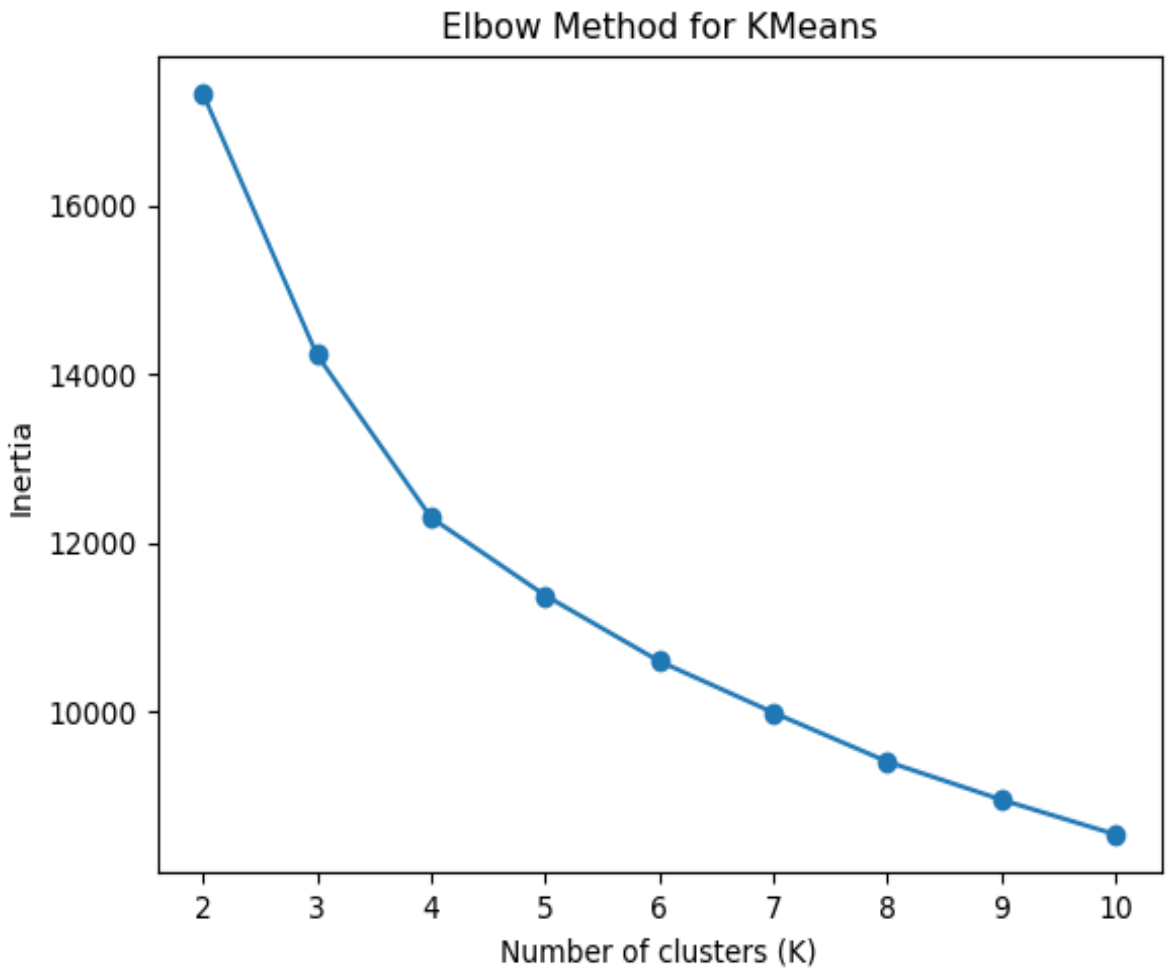
Στο πλαίσιο της πειραματικής υλοποίησης, μετά την κατασκευή του τελικού dataset, εφαρμόστηκαν τεχνικές ανάλυσης διαστάσεων και μη επιβλεπόμενης μάθησης με στόχο τον εντοπισμό διακριτών τύπων παικτών (player archetypes).

Αρχικά, τα επιλεγμένα χαρακτηριστικά του τελικού dataset κανονικοποιήθηκαν μέσω της μεθόδου standardization, ώστε όλα τα μεγέθη να έχουν μηδενικό μέσο και μοναδιαία διασπορά. Η κανονικοποίηση κρίθηκε απαραίτητη, καθώς τα χαρακτηριστικά προέρχονται από διαφορετικές κλίμακες μέτρησης και επηρεάζουν διαφορετικά τη διαδικασία ομαδοποίησης.

Στη συνέχεια εφαρμόστηκε Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis – PCA) με σκοπό τη μείωση της διαστασιμότητας του προβλήματος, διατηρώντας παράλληλα το μεγαλύτερο δυνατό ποσοστό της πληροφορίας. Από την ανάλυση της αθροιστικής εξηγούμενης διακύμανσης προέκυψε ότι οι πρώτες πέντε κύριες συνιστώσες εξηγούν πάνω από το 80% της συνολικής διακύμανσης των δεδομένων, γεγονός που οδήγησε στην επιλογή τους για τα επόμενα στάδια της ανάλυσης.

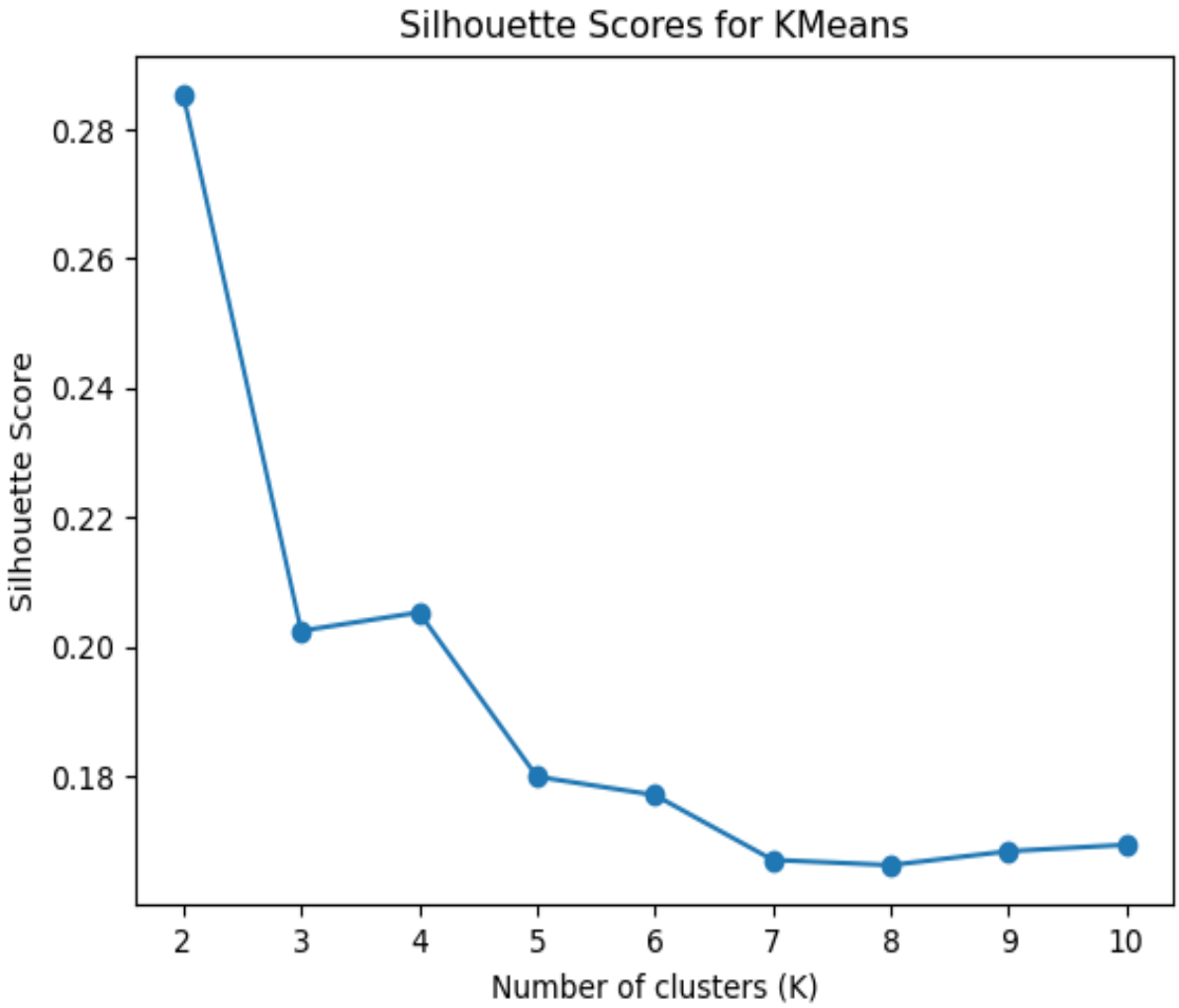
Πάνω στον μειωμένο χώρο χαρακτηριστικών εφαρμόστηκε ο αλγόριθμος K-Means για την ομαδοποίηση των παικτών. Η επιλογή του αριθμού των clusters πραγματοποιήθηκε με συνδυαστική αξιολόγηση της μεθόδου elbow και του δείκτη silhouette, προκειμένου να επιτευχθεί ισορροπία μεταξύ απλότητας του μοντέλου και ποιότητας της ομαδοποίησης.

Η μέθοδος elbow χρησιμοποιήθηκε για την εξέταση της μεταβολής της ενδο-ομαδικής διακύμανσης ως συνάρτηση του αριθμού των clusters. Τα αποτελέσματα έδειξαν ότι μετά την επιλογή τεσσάρων clusters η περαιτέρω μείωση της διακύμανσης είναι περιορισμένη, γεγονός που υποδηλώνει φθίνουσα απόδοση από την αύξηση του αριθμού των ομάδων. Παράλληλα, ο δείκτης silhouette παρουσίασε σχετικά υψηλότερες τιμές για μικρό αριθμό clusters, με την επιλογή K=4K=4K=4 να προσφέρει ικανοποιητικό επίπεδο συνοχής εντός των ομάδων χωρίς να οδηγεί σε υπερβολικό κατακερματισμό των δεδομένων.



Σχήμα 4.1: Elbow Method for K-Means

Η επιλογή μικρότερου αριθμού clusters, όπως *K=2* ή *K=3*, οδηγούσε σε υπεραπλουστευμένη ομαδοποίηση, στην οποία διαφορετικοί αγωνιστικοί ρόλοι συγχωνεύονταν σε ευρύτερες ομάδες. Αντίθετα, η επιλογή μεγαλύτερου αριθμού clusters παρότι αύξανε οριακά τον δείκτη silhouette, παρήγαγε ομάδες με μικρότερη ερμηνευσιμότητα και αυξημένο κίνδυνο υπερπροσαρμογής. Συνεπώς, η επιλογή τεσσάρων clusters κρίθηκε ως ο πλέον κατάλληλος συμβιβασμός μεταξύ στατιστικής ποιότητας και ερμηνευσιμότητας των αποτελεσμάτων.



Σχήμα 4.2: Silhouette score for K-Means

Η ανάλυση των μέσων τιμών των στατιστικών ανά cluster ανέδειξε τέσσερα σαφώς διαφοροποιημένα αγωνιστικά προφίλ, τα οποία μπορούν να ερμηνευθούν ως διακριτοί ρόλοι παικτών εντός του σύγχρονου παιχνιδιού. Ένα από τα προκύπτοντα clusters χαρακτηρίζεται από ιδιαίτερα αυξημένες τιμές σε δείκτες επιθετικής χρήσης και δημιουργίας, όπως το ποσοστό χρήσης (USG%) και το ποσοστό ασίστ (AST%). Το προφίλ αυτό αντιστοιχεί σε παίκτες που αναλαμβάνουν κεντρικό ρόλο στην επιθετική λειτουργία της ομάδας, συμμετέχοντας ενεργά τόσο στη δημιουργία ευκαιριών για τους συμπαίκτες τους όσο και στην τελική εκτέλεση των επιθέσεων. Οι παίκτες αυτής της ομάδας εμφανίζουν αυξημένη επιρροή στη ροή του παιχνιδιού και φέρουν αυξημένο βάρος στις επιθετικές αποφάσεις, γεγονός που

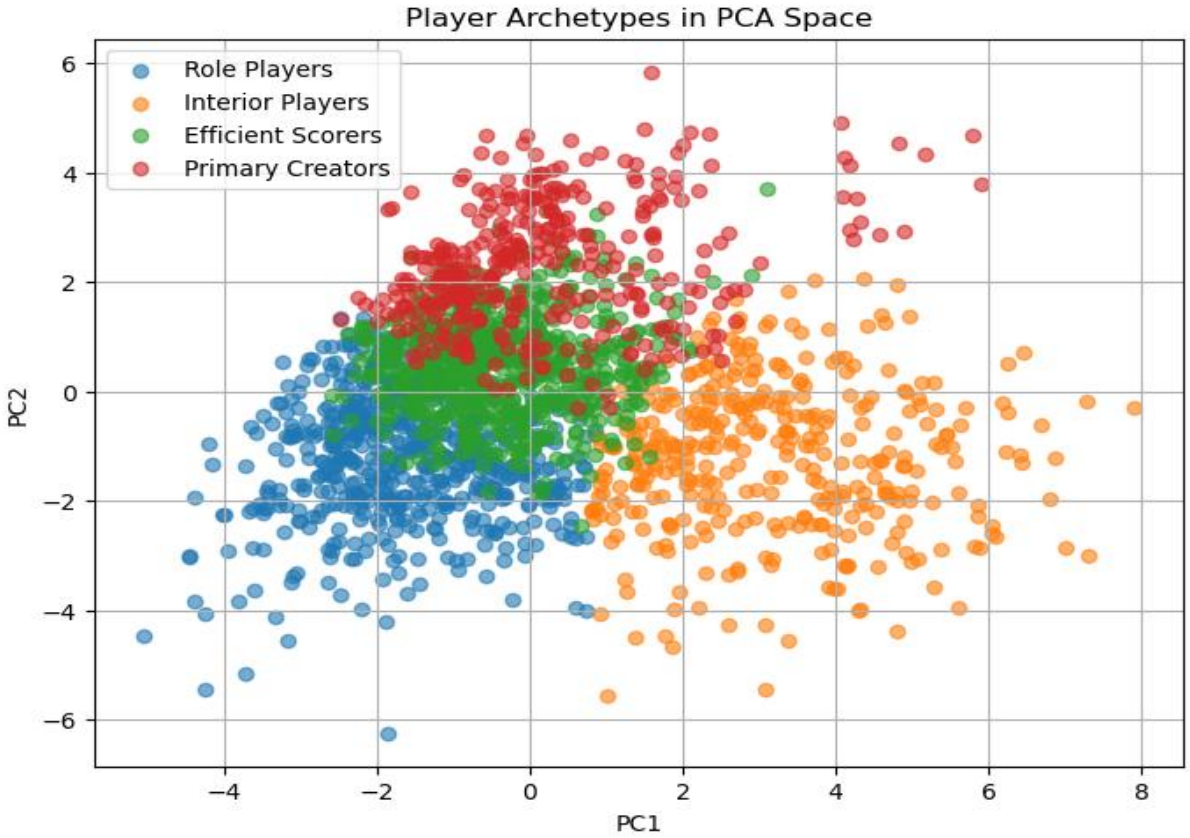
καθιστά τον ρόλο τους κομβικό για τη συνολική απόδοση της ομάδας. Ενδεικτικά παραδείγματα παικτών που ανήκουν σε αυτή την κατηγορία είναι ο Jaylen Brown και ο Jayson Tatum των Boston Celtics.

Παράλληλα, προκύπτει ομάδα παικτών με αυξημένη αποδοτικότητα στην εκτέλεση, αλλά χωρίς αντίστοιχα υψηλά επίπεδα επιθετικής χρήσης ή δημιουργίας. Οι παίκτες αυτοί εμφανίζουν ευνοϊκούς δείκτες ευστοχίας και συνολικής επιθετικής αποτελεσματικότητας, λειτουργώντας κυρίως ως τελικοί αποδέκτες των επιθετικών ενεργειών μέσα σε οργανωμένα επιθετικά σχήματα. Το συγκεκριμένο προφίλ αντιστοιχεί σε παίκτες που συνεισφέρουν σημαντικά στο σκοράρισμα χωρίς να αποτελούν τον βασικό άξονα δημιουργίας, συμπληρώνοντας αποτελεσματικά τους κεντρικούς δημιουργούς της ομάδας. Και σε αυτή την περίπτωση, η αναφορά σε αντιπροσωπευτικούς παίκτες μπορεί να συμβάλει στη σαφέστερη ερμηνεία των αποτελεσμάτων. Χαρακτηριστικά παραδείγματα αποτελούν ο Buddy Hield και ο Jalen Suggs.

Ένα ακόμη cluster χαρακτηρίζεται από αυξημένες τιμές σε στατιστικά που σχετίζονται με το εσωτερικό παιχνίδι και την αμυντική παρουσία, όπως τα ποσοστά ριμπάουντ και τα μπλοκ, σε συνδυασμό με χαμηλότερη επιθετική χρήση και περιορισμένη δημιουργία. Οι παίκτες αυτής της ομάδας συμβάλλουν κυρίως μέσω φυσικής παρουσίας στη ρακέτα, προστασίας του καλαθιού και ελέγχου των κατοχών, διαδραματίζοντας ρόλο σταθεροποιητικό για την αμυντική λειτουργία της ομάδας. Το αγωνιστικό τους προφίλ υποδηλώνει έμφαση στη συλλογική ισορροπία και λιγότερο στην ατομική επιθετική πρωτοβουλία, γεγονός που αντικατοπτρίζει τον ρόλο των εσωτερικών παικτών στο σύγχρονο παιχνίδι. Ενδεικτικά παραδείγματα παικτών μπορούν να ενταχθούν και εδώ και συγκεκριμένα είναι παίκτες όπως ο Mason Plumlee και ο Montrezl Harell.

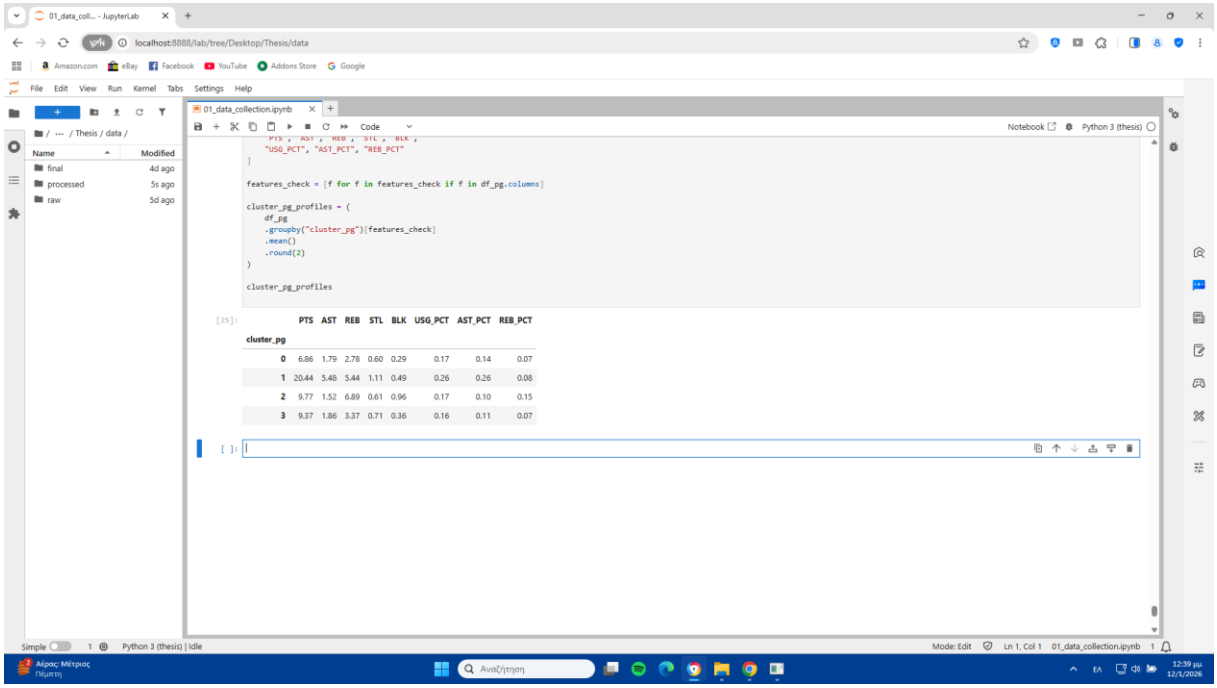
Τέλος, εντοπίζεται ομάδα παικτών με χαμηλότερες τιμές τόσο σε επιθετική χρήση όσο και σε δημιουργικά ή αμυντικά στατιστικά, η οποία μπορεί να ερμηνευθεί ως ομάδα παικτών συμπληρωματικών ρόλων. Οι παίκτες αυτοί δεν κυριαρχούν σε κάποιον επιμέρους τομέα του παιχνιδιού, ωστόσο συμβάλλουν στη συνολική λειτουργία της ομάδας μέσω εξειδικευμένων ή υποστηρικτικών καθηκόντων, τα οποία δεν αποτυπώνονται απαραίτητα μέσω υψηλών στατιστικών τιμών. Η παρουσία τους αναδεικνύει τη σημασία των ρόλων που δεν σχετίζονται άμεσα με την παραγωγή στατιστικών, αλλά είναι κρίσιμοι για τη συνοχή και την προσαρμοστικότητα της ομάδας. Εδώ ανήκουν παίκτες όπως Chima Okeke και ο Svi Mykhailiuik.

Συνολικά, η ύπαρξη αυτών των σαφώς διακριτών αγωνιστικών προφίλ υποδηλώνει ότι η διαδικασία ομαδοποίησης δεν βασίζεται σε τυχαίους αριθμητικούς διαχωρισμούς, αλλά συλλαμβάνει ουσιαστικές αγωνιστικές συμπεριφορές που αντανακλούν πραγματικούς ρόλους εντός του παιχνιδιού. Ως εκ τούτου, τα αποτελέσματα του clustering παρέχουν ένα συνεκτικό και ερμηνεύσιμο πλαίσιο, το οποίο μπορεί να αξιοποιηθεί ως βάση για περαιτέρω επιβλεπόμενες αναλύσεις, όπως η παλινδρόμηση που ακολουθεί στην επόμενη ενότητα.



Σχήμα 4.3: Player Archetypes

Τέλος, πραγματοποιήθηκε έλεγχος ευρωστίας (robustness check) επαναλαμβάνοντας τη διαδικασία PCA και clustering σε εναλλακτικό dataset που περιλάμβανε επιπλέον παραδοσιακά per-game στατιστικά. Η σύγκριση των αποτελεσμάτων έδειξε ότι οι βασικοί τύποι παικτών παραμένουν ποιοτικά παρόμοιοι, γεγονός που υποδηλώνει ότι τα ευρήματα της ανάλυσης δεν εξαρτώνται αποκλειστικά από την επιλογή συγκεκριμένων χαρακτηριστικών, αλλά αντικατοπτρίζουν σταθερές αγωνιστικές συμπεριφορές.



Εικόνα 4.8: Robustness Check

4.3 Ανάλυση Παλινδρόμησης και Ρόλων Παικτών

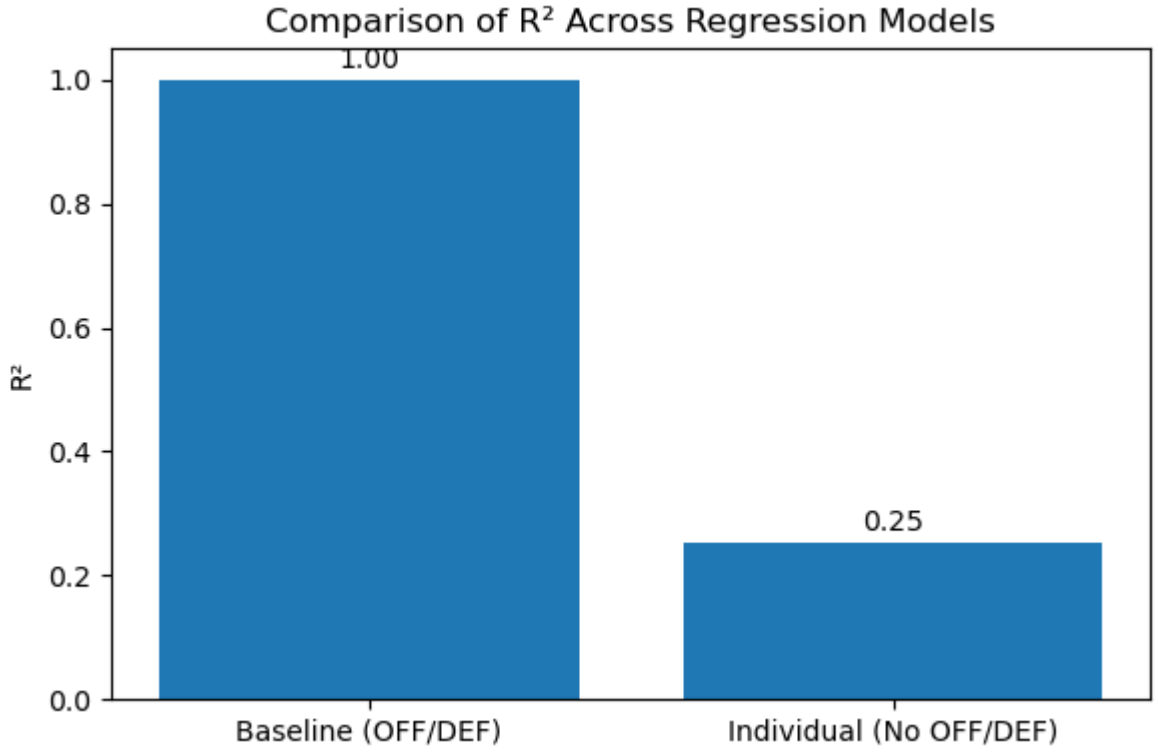
Στην παρούσα ενότητα εξετάζεται η σχέση μεταξύ ατομικών στατιστικών απόδοσης παικτών και του δείκτη *NET Rating*, με στόχο την κατανόηση των παραγόντων που σχετίζονται με τη συνολική αγωνιστική επίδραση ενός παίκτη στο ομαδικό αποτέλεσμα. Η ανάλυση πραγματοποιείται σε δύο στάδια: αρχικά μέσω συνολικών μοντέλων παλινδρόμησης και στη συνέχεια μέσω μοντέλων προσαρμοσμένων σε διαφορετικούς αγωνιστικούς ρόλους, όπως αυτοί προκύπτουν από μεθόδους ομαδοποίησης.

Επιλέχθηκαν στην παρούσα εργασία οι αλγόριθμοι K-Means και Gaussian Mixture Models (GMM) λόγω της ευρείας χρήσης τους στη βιβλιογραφία της ανάλυσης αθλητικών δεδομένων και της συμπληρωματικής φιλοσοφίας ομαδοποίησης που υιοθετούν. Ο K-Means προσφέρει μια απλή και υπολογιστικά αποδοτική προσέγγιση, επιτρέποντας τον διαχωρισμό των παικτών σε διακριτές ομάδες βάσει ομοιότητας των αγωνιστικών χαρακτηριστικών τους. Αντίθετα, τα Gaussian Mixture Models υιοθετούν μια πιθανοκρατική προσέγγιση, επιτρέποντας την αποτύπωση πιο σύνθετων και επικαλυπτόμενων αγωνιστικών προφίλ. Η παράλληλη χρήση των δύο μεθόδων καθιστά δυνατή τη σύγκριση διαφορετικών υποθέσεων ομαδοποίησης και την αξιολόγηση της σταθερότητας και της ερμηνευσιμότητας των αποτελεσμάτων, ενισχύοντας την αξιοπιστία της ανάλυσης.

4.3.1 Συνολική παλινδρόμηση (Baseline και Ατομικό Μοντέλο)

Αρχικά εκτιμήθηκε ένα baseline μοντέλο παλινδρόμησης, στο οποίο το NET Rating χρησιμοποιήθηκε ως μεταβλητή-στόχος και ως επεξηγηματικές μεταβλητές συμπεριλήφθηκαν οι δείκτες *OFF Rating* και *DEF Rating*. Το συγκεκριμένο μοντέλο παρουσίασε εξαιρετικά υψηλή επεξηγηματική ικανότητα,

γεγονός που ήταν αναμενόμενο, καθώς το NET Rating ορίζεται άμεσα ως η διαφορά μεταξύ των δύο αυτών μεγεθών. Ωστόσο, το αποτέλεσμα αυτό δεν θεωρείται ερμηνευτικά χρήσιμο, καθώς οι συγκεκριμένοι δείκτες ενσωματώνουν πληροφορία που σχετίζεται άμεσα με τον ορισμό της μεταβλητής-στόχου, δημιουργώντας φαινόμενο *data leakage*. Για τον λόγο αυτό, το baseline μοντέλο χρησιμοποιείται αποκλειστικά ως σημείο αναφοράς (upper bound) και όχι ως τελικό ερμηνευτικό εργαλείο.



Σχήμα 4.3.1: Σύγκριση του συντελεστή προσδιορισμού (R^2) μεταξύ του baseline και του ατομικού μοντέλου. Το υψηλό R^2 του baseline μοντέλου αποδίδεται σε πληροφοριακή διαρροή (data leakage).

Στη συνέχεια εκτιμήθηκε ένα ατομικό μοντέλο παλινδρόμησης, στο οποίο αφαιρέθηκαν οι δείκτες ομαδικής απόδοσης και διατηρήθηκαν αποκλειστικά ατομικά στατιστικά χαρακτηριστικά, όπως οι δείκτες χρήσης (USG%), αποδοτικότητας (TS%, EFG%), δημιουργίας (AST%), ριμπάουντ (OREB%, DREB%, REB%) και ο σύνθετος δείκτης συνολικής επίδρασης *Player Impact Estimate (PIE),* ο οποίος αποτελεί σύνθετο μέτρο συνολικής αγωνιστικής συνεισφοράς. Ο δείκτης αυτός ενσωματώνει θετικές και αρνητικές ενέργειες του παίκτη (όπως πόντους, ριμπάουντ, ασίστ, κλεψίματα, λάθη και άστοχα σουτ), αποτυπώνοντας τη συνολική του επίδραση στο παιχνίδι σε σχέση με τους υπόλοιπους παίκτες του αγώνα. Σε αντίθεση με μεμονωμένα στατιστικά μεγέθη, το PIE παρέχει μια πιο ολιστική εικόνα της απόδοσης, γεγονός που εξηγεί τη σταθερά θετική και ισχυρή συσχέτισή του με το NET Rating στα αποτελέσματα της παλινδρόμησης. Το εύρημα αυτό υποδηλώνει ότι η συνολική αγωνιστική επίδραση ενός παίκτη, και όχι η υπεροχή σε έναν επιμέρους τομέα, αποτελεί κρίσιμο παράγοντα για τη θετική ομαδική απόδοση. Το μοντέλο αυτό παρουσιάζει ικανοποιητική επεξηγηματική ικανότητα ($R^2 \approx 0.38$), δεδομένου ότι βασίζεται αποκλειστικά σε ατομικά δεδομένα και δεν λαμβάνει υπόψη το πλήρες ομαδικό και τακτικό πλαίσιο.

Τα αποτελέσματα δείχνουν ότι οι δείκτες αποδοτικότητας και συνολικής αγωνιστικής επίδρασης (ιδίως το PIE) εμφανίζουν σταθερά θετική και στατιστικά σημαντική σχέση με το NET Rating. Αντίθετα, δείκτες που σχετίζονται με τον όγκο συμμετοχής, όπως το USG%, εμφανίζουν συχνά αρνητική συσχέτιση, υποδηλώνοντας ότι η αυξημένη χρήση δεν μεταφράζεται απαραίτητα σε θετικό ομαδικό αποτέλεσμα εάν δεν συνοδεύεται από υψηλή αποτελεσματικότητα. Τα ευρήματα αυτά ενισχύουν την άποψη ότι στο σύγχρονο μπάσκετ η ποιότητα των αποφάσεων και η συνολική συνεισφορά υπερέχουν της απλής συσσώρευσης στατιστικών.

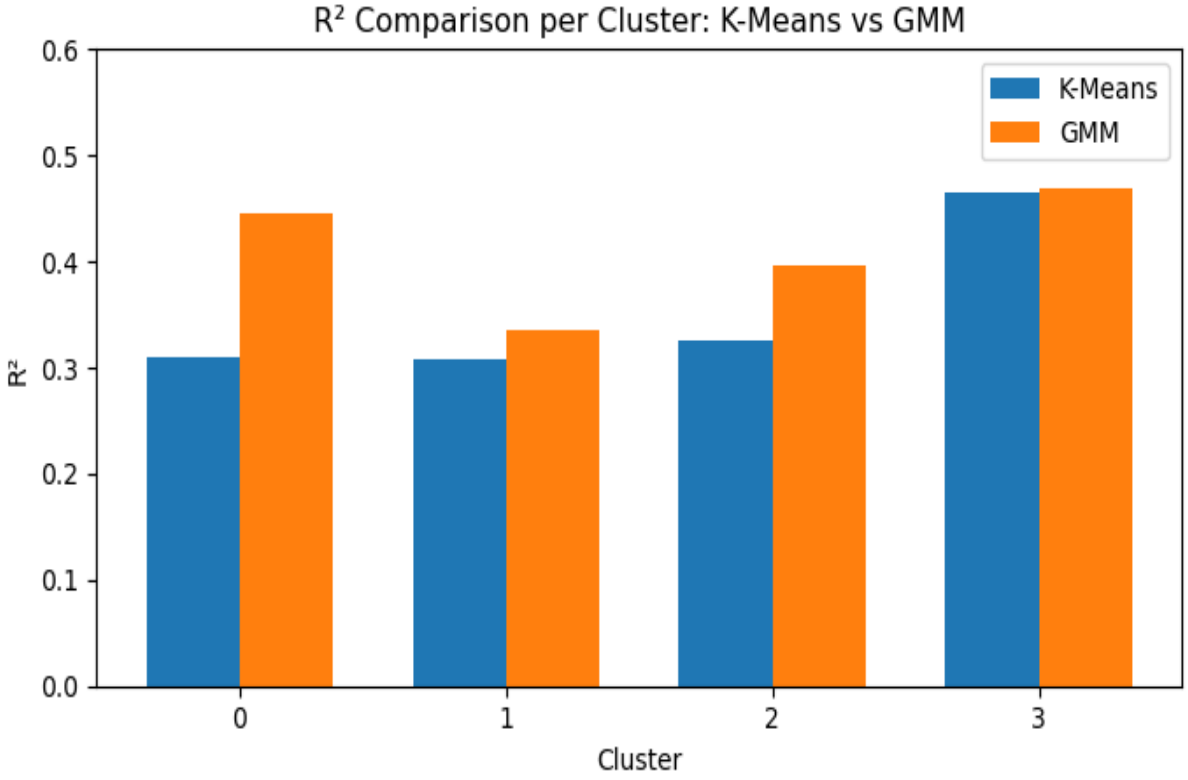
4.3.2 Παλινδρόμηση βάσει Αγωνιστικών Ρόλων (Clustering και GMM)

Παρότι το συνολικό μοντέλο παλινδρόμησης παρέχει σημαντικές ενδείξεις για τη σχέση μεταξύ ατομικών στατιστικών και NET Rating, προϋποθέτει ότι η επίδραση των μεταβλητών είναι ομοιόμορφη για όλους τους παίκτες. Ωστόσο, στο πλαίσιο του σύγχρονου μπάσκετ, οι παίκτες επιτελούν διακριτούς αγωνιστικούς ρόλους, γεγονός που καθιστά πιθανή τη διαφοροποίηση των παραγόντων που επηρεάζουν τη συνολική τους επίδραση στο παιχνίδι. Για τον λόγο αυτό, εφαρμόστηκαν μέθοδοι μη επιβλεπόμενης μάθησης με στόχο την ομαδοποίηση παικτών σε ρόλους και τη μετέπειτα εκτίμηση ξεχωριστών μοντέλων παλινδρόμησης.

Αρχικά χρησιμοποιήθηκε ο αλγόριθμος K-Means, ο οποίος ομαδοποιεί τους παίκτες βάσει της απόστασής τους στον μειωμένο χώρο χαρακτηριστικών που προέκυψε από την Ανάλυση Κύριων Συνιστωσών. Η επιλογή του αριθμού των clusters βασίστηκε σε κριτήρια συνοχής και διαχωρισιμότητας (elbow method και silhouette score), οδηγώντας στη λύση των τεσσάρων ομάδων. Η εφαρμογή ξεχωριστών παλινδρομήσεων για κάθε cluster αποκάλυψε σημαντικές διαφοροποιήσεις τόσο στην επεξηγηματική ικανότητα των μοντέλων όσο και στη στατιστική σημασία των επιμέρους μεταβλητών. Το αποτέλεσμα αυτό υποδηλώνει ότι η σχέση μεταξύ ατομικών στατιστικών και NET Rating εξαρτάται ουσιαστικά από τον αγωνιστικό ρόλο του παίκτη και δεν μπορεί να αποδοθεί μέσω ενός ενιαίου μοντέλου.

Στη συνέχεια εφαρμόστηκαν Gaussian Mixture Models (GMM), τα οποία επεκτείνουν την προσέγγιση του K-Means επιτρέποντας την πιθανοκρατική ανάθεση των παικτών σε ομάδες. Σε αντίθεση με την «σκληρή» ομαδοποίηση του K-Means, τα GMM επιτρέπουν σε έναν παίκτη να ανήκει ταυτόχρονα σε περισσότερους από έναν ρόλους με διαφορετικές πιθανότητες, αποτυπώνοντας με μεγαλύτερη ακρίβεια τη ρευστότητα και την πολυδιάστατη φύση των αγωνιστικών προφίλ. Η ανάλυση παλινδρόμησης ανά GMM cluster οδήγησε σε μοντέλα με συγκρίσιμη ή και υψηλότερη επεξηγηματική ικανότητα σε σχέση με τον K-Means, ενώ παράλληλα παρείχε πιο καθαρή ερμηνεία των ρόλων, ιδίως για παίκτες υψηλής συνολικής επίδρασης.

Ιδιαίτερα στα clusters που αντιστοιχούν σε παίκτες υψηλού αντίκτυπου, τα μοντέλα GMM παρουσίασαν τα υψηλότερα επίπεδα R^2, με δείκτες αποδοτικότητας και τον δείκτη PIE να αναδεικνύονται ως οι κυρίαρχοι παράγοντες που σχετίζονται με το NET Rating. Αντίθετα, σε ομάδες παικτών χαμηλής χρήσης ή καθαρά συμπληρωματικών ρόλων, η επεξηγηματική ικανότητα των μοντέλων ήταν χαμηλότερη, γεγονός που υποδηλώνει ότι το ομαδικό και τακτικό πλαίσιο επηρεάζει περισσότερο τη συνολική απόδοση από τα μεμονωμένα ατομικά χαρακτηριστικά.



Σχήμα 4.3.2: Σύγκριση του R^2 των μοντέλων παλινδρόμησης ανά cluster για τις μεθόδους K-Means και Gaussian Mixture Models.

Η σύγκριση της επεξηγηματικής ικανότητας των μοντέλων παλινδρόμησης πραγματοποιήθηκε μέσω του δείκτη R^2, εξετάζοντας ξεχωριστά κάθε cluster που προέκυψε από τις μεθόδους K-Means και Gaussian Mixture Models. Τα αποτελέσματα δείχνουν ότι η ομαδοποίηση των παικτών οδηγεί σε διαφοροποιημένη ερμηνευτική ισχύ των μοντέλων, γεγονός που επιβεβαιώνει ότι οι σχέσεις μεταξύ των αγωνιστικών χαρακτηριστικών και του NET Rating δεν είναι ομοιόμορφες για όλους τους παίκτες. Σε αρκετές περιπτώσεις, τα μοντέλα που βασίζονται στα clusters των GMM εμφανίζουν συγκρίσιμες ή ελαφρώς υψηλότερες τιμές R^2 σε σχέση με τα αντίστοιχα του K-Means, υποδηλώνοντας ότι η πιθανοκρατική φύση της ομαδοποίησης επιτρέπει μια πιο εύκαμπτη αποτύπωση της ετερογένειας των παικτών.

Κεφάλαιο 5ο: Συμπεράσματα και επεκτάσεις

5.1 Συμπεράσματα

Στην παρούσα διπλωματική εργασία μελετήθηκε η σχέση μεταξύ ατομικών στατιστικών απόδοσης παικτών καλαθοσφαίρισης και της συνολικής ομαδικής επίδρασης, όπως αυτή αποτυπώνεται μέσω του δείκτη NET Rating. Μέσω της αξιοποίησης τεχνικών στατιστικής ανάλυσης και μηχανικής μάθησης, επιχειρήθηκε η κατανόηση των παραγόντων που συνδέονται με τη θετική ή αρνητική συνεισφορά ενός παίκτη στο αγωνιστικό αποτέλεσμα.

Αρχικά, η ανάλυση παλινδρόμησης στο σύνολο των παικτών ανέδειξε ότι τα ατομικά στατιστικά είναι σε θέση να εξηγήσουν σε ικανοποιητικό βαθμό τη διακύμανση του NET Rating, ακόμη και χωρίς τη χρήση δεικτών άμεσης ομαδικής απόδοσης, όπως τα OFF και DEF Rating. Το εύρημα αυτό επιβεβαιώνει ότι η ατομική απόδοση αποτελεί σημαντικό παράγοντα της συνολικής αγωνιστικής επίδρασης, αν και δεν επαρκεί από μόνη της για την πλήρη περιγραφή της πολυπλοκότητας του παιχνιδιού.

Ιδιαίτερη σημασία παρουσιάζει το γεγονός ότι οι δείκτες αποδοτικότητας (όπως TS% και EFG%) και ο σύνθετος δείκτης Player Impact Estimate (PIE) εμφανίζονται συστηματικά πιο σημαντικοί από δείκτες όγκου συμμετοχής, όπως το Usage Rate. Το αποτέλεσμα αυτό υποδηλώνει ότι η ποιότητα των αποφάσεων και η συνολική συνεισφορά ενός παίκτη υπερισχύουν της απλής αύξησης της επιθετικής συμμετοχής, επιβεβαιώνοντας σύγχρονες αντιλήψεις της ανάλυσης καλαθοσφαίρισης.

Στη συνέχεια, μέσω της εφαρμογής τεχνικών ομαδοποίησης (K-Means και Gaussian Mixture Models), αναδείχθηκε ότι οι παίκτες δεν αποτελούν ένα ομοιογενές σύνολο, αλλά μπορούν να ταξινομηθούν σε διακριτούς αγωνιστικούς ρόλους. Η εκτίμηση ξεχωριστών μοντέλων παλινδρόμησης για κάθε ομάδα έδειξε ότι η επίδραση των στατιστικών χαρακτηριστικών στο NET Rating διαφοροποιείται ουσιαστικά ανάλογα με τον ρόλο του παίκτη. Τα αποτελέσματα αυτά καταδεικνύουν ότι η χρήση ενός ενιαίου μοντέλου για όλους τους παίκτες ενδέχεται να αποκρύπτει σημαντικές δομικές διαφορές.

Η σύγκριση μεταξύ K-Means και Gaussian Mixture Models έδειξε ότι, παρότι και οι δύο μέθοδοι οδηγούν σε παρόμοια ποιοτικά συμπεράσματα, τα Gaussian Mixture Models προσφέρουν αυξημένη ευελιξία και πιο ρεαλιστική αποτύπωση των αγωνιστικών ρόλων, επιτρέποντας την ύπαρξη υβριδικών προφίλ παικτών. Η παρατήρηση αυτή ενισχύει τη χρησιμότητα των πιθανοκρατικών προσεγγίσεων σε προβλήματα αθλητικής ανάλυσης, όπου οι ρόλοι δεν είναι αυστηρά διακριτοί.

Συνολικά, η εργασία καταδεικνύει ότι ο συνδυασμός ατομικών στατιστικών, τεχνικών ομαδοποίησης και παλινδρόμησης αποτελεί ένα ισχυρό πλαίσιο για την ανάλυση της αγωνιστικής επίδρασης των παικτών, προσφέροντας ερμηνεύσιμα και πρακτικά χρήσιμα συμπεράσματα.

5.2 Μελλοντικές Επεκτάσεις

Παρότι τα αποτελέσματα της παρούσας εργασίας είναι ενθαρρυντικά και αναδεικνύουν τη χρησιμότητα των ατομικών στατιστικών και των αγωνιστικών ρόλων στην κατανόηση της συνολικής απόδοσης των παικτών, υπάρχουν αρκετές κατευθύνσεις για μελλοντική επέκταση και περαιτέρω βελτίωση του προτεινόμενου πλαισίου.

Μία πρώτη κατεύθυνση αφορά την ενσωμάτωση πλουσιότερων δεδομένων απόδοσης, όπως χωρικά και χρονικά δεδομένα (player tracking data), τα οποία θα μπορούσαν να προσφέρουν λεπτομερέστερη εικόνα της αμυντικής συνεισφοράς, της κίνησης χωρίς την μπάλα και της συνεργασίας μεταξύ παικτών. Η αξιοποίηση τέτοιων δεδομένων θα επέτρεπε την υπέρβαση των περιορισμών των παραδοσιακών box-score στατιστικών.

Επιπλέον, μελλοντική εργασία θα μπορούσε να εξετάσει τη χρήση μη γραμμικών και πιο σύνθετων μοντέλων μηχανικής μάθησης, όπως ensemble μέθοδοι ή νευρωνικά δίκτυα, με στόχο τη σύλληψη πιο πολύπλοκων αλληλεπιδράσεων μεταξύ των χαρακτηριστικών. Παράλληλα, η σύγκριση των αποτελεσμάτων αυτών με τα γραμμικά και ερμηνεύσιμα μοντέλα που χρησιμοποιήθηκαν στην παρούσα εργασία θα μπορούσε να προσφέρει σημαντικά συμπεράσματα ως προς το συμβιβασμό μεταξύ ακρίβειας και ερμηνευσιμότητας.

Μία ακόμη ενδιαφέρουσα προοπτική αφορά τη δυναμική ανάλυση αγωνιστικών ρόλων, εξετάζοντας πώς οι ρόλοι και η επίδραση των στατιστικών χαρακτηριστικών μεταβάλλονται διαχρονικά, είτε εντός μιας αγωνιστικής περιόδου είτε κατά την εξέλιξη της καριέρας ενός παίκτη. Μια τέτοια προσέγγιση θα μπορούσε να αναδείξει μοτίβα εξέλιξης και προσαρμογής των παικτών σε διαφορετικά αγωνιστικά πλαίσια.

Τέλος, ένα ιδιαίτερα σημαντικό επόμενο βήμα αφορά τη μετάβαση από την ανάλυση στην υποστήριξη αποφάσεων, μέσω της σύνθεσης βέλτιστων πεντάδων παικτών (lineup optimization). Αξιοποιώντας τα αποτελέσματα της παρούσας εργασίας, και ειδικότερα τις εκτιμήσεις της επίδρασης των ατομικών στατιστικών στο NET Rating καθώς και τη διάκριση των παικτών σε αγωνιστικούς ρόλους, θα μπορούσε να διαμορφωθεί ένα πρόβλημα βελτιστοποίησης με στόχο τη μεγιστοποίηση της αναμενόμενης συνολικής απόδοσης μιας πεντάδας. Η προσέγγιση αυτή επιτρέπει την ενσωμάτωση περιορισμών που σχετίζονται με την ισορροπία ρόλων, τη θέση των παικτών, τον χρόνο συμμετοχής ή ακόμη και οικονομικούς περιορισμούς, μετατρέποντας το προτεινόμενο πλαίσιο σε εργαλείο στρατηγικού σχεδιασμού.

Η κατεύθυνση αυτή αποτελεί φυσική συνέχεια της παρούσας εργασίας και διερευνάται ενδεικτικά στην επόμενη υποενότητα, με στόχο την επίδειξη της πρακτικής αξιοποίησης των αποτελεσμάτων μέσω μιας απλής υλοποίησης βελτιστοποίησης πεντάδας.

5.2.1 Lineup Optimization – Πειραματική Υλοποίηση και Μελλοντική Εξέλιξη

Στο πλαίσιο των μελλοντικών επεκτάσεων της παρούσας εργασίας, πραγματοποιήθηκε μία πειραματική υλοποίηση βελτιστοποίησης πεντάδας παικτών (lineup optimization), με στόχο την ανάδειξη του τρόπου με τον οποίο τα αποτελέσματα των προηγούμενων σταδίων (clustering και regression) μπορούν να αξιοποιηθούν σε εφαρμογές υποστήριξης αγωνιστικών αποφάσεων.

Η προσέγγιση βασίστηκε στη χρήση των συντελεστών παλινδρόμησης (regression coefficients) που προέκυψαν από τα μοντέλα OLS ανά cluster του Gaussian Mixture Model (GMM). Για κάθε παίκτη, υπολογίστηκε ένας συνολικός δείκτης αναμενόμενης επίδρασης (expected impact score), ο οποίος προκύπτει ως γραμμικός συνδυασμός των ατομικών του στατιστικών χαρακτηριστικών (π.χ. USG%, TS%, AST%, REB%, PIE), σταθμισμένων με τους αντίστοιχους συντελεστές του cluster στο οποίο ανήκει.

Στη συνέχεια, οι παίκτες κατηγοριοποιήθηκαν σε λειτουργικούς ρόλους (roles), όπως *Creator*, *Shooter*, *Big* και *Role Player*, με βάση το προφίλ απόδοσής τους. Η διαδικασία επιλογής της πεντάδας πραγματοποιήθηκε με έναν απλό greedy αλγόριθμο, ο οποίος επέλεξε τους παίκτες με το υψηλότερο expected impact score, τηρώντας έναν προκαθορισμένο περιορισμό σύνθεσης (1 Creator, 1 Shooter, 1 Big και 2 Role Players). Το αποτέλεσμα είναι μία πεντάδα που μεγιστοποιεί το άθροισμα του expected impact score υπό τους δεδομένους περιορισμούς.

Η συγκεκριμένη υλοποίηση έχει πειραματικό χαρακτήρα και δεν στοχεύει στην πλήρη προσομοίωση αγωνιστικών συνθηκών, ωστόσο καταδεικνύει με σαφήνεια πώς τα αποτελέσματα της μηχανικής μάθησης μπορούν να μεταφραστούν σε λειτουργικά εργαλεία λήψης αποφάσεων. Στην παρούσα μορφή της, η μέθοδος δεν λαμβάνει υπόψη παράγοντες όπως λεπτά συμμετοχής, αλληλεπιδράσεις μεταξύ παικτών (synergy), αγωνιστικό ρυθμό ή αντίπαλα matchups.

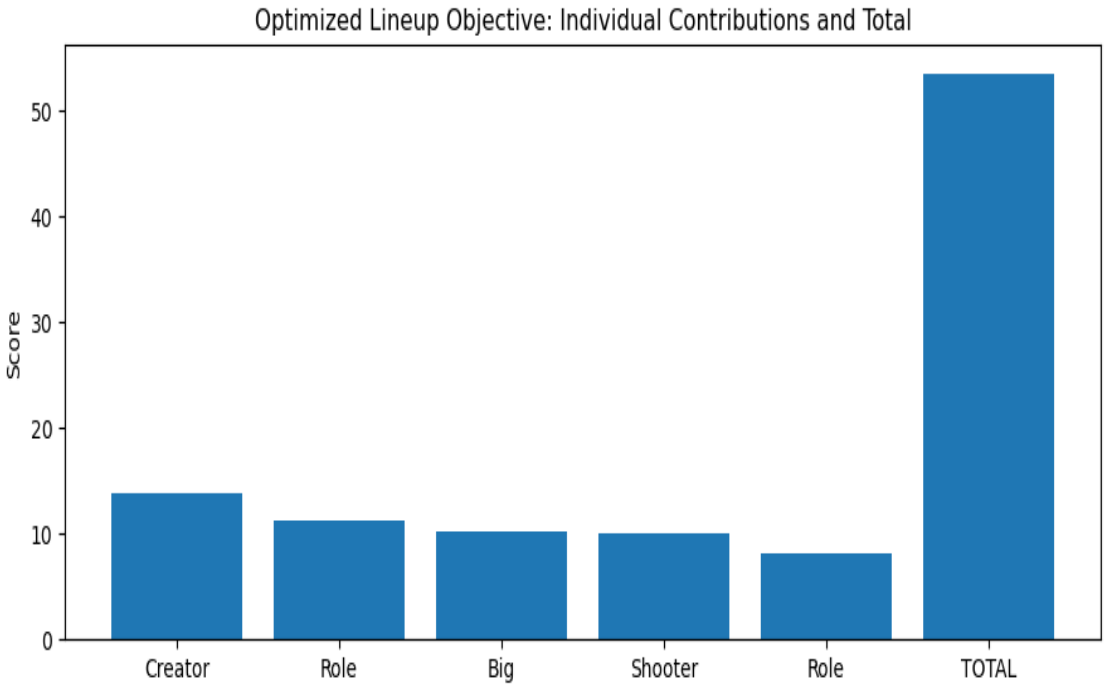
Η παραπάνω προσέγγιση μπορεί να επεκταθεί σημαντικά σε μελλοντική έρευνα. Ενδεικτικά:

- Η διαδικασία επιλογής πεντάδας μπορεί να διατυπωθεί ως πρόβλημα ακέραιου γραμμικού προγραμματισμού (Integer Linear Programming – ILP), επιτρέποντας ταυτόχρονα περισσότερους περιορισμούς, όπως salary cap, ηλικία, εμπειρία ή μέγιστο αριθμό παικτών ανά ομάδα.

- Ο δείκτης expected impact score μπορεί να εξελιχθεί σε προβλεπτικό μοντέλο lineups, λαμβάνοντας υπόψη όχι μόνο ατομικά χαρακτηριστικά αλλά και συνδυαστικά χαρακτηριστικά πεντάδων.

- Μπορούν να ενσωματωθούν δεδομένα play-by-play ή on/off metrics, ώστε να αποτυπώνεται πιο ρεαλιστικά η επίδραση συγκεκριμένων συνδυασμών παικτών στο NET Rating.

- Τέλος, η μέθοδος μπορεί να αποτελέσει τη βάση για ένα decision-support system για προπονητές ή αναλυτές, το οποίο θα προτείνει βέλτιστες πεντάδες ανάλογα με τον αντίπαλο, το αγωνιστικό στυλ ή το σενάριο αγώνα.



Σχήμα 5.2.1: Συνεισφορά των επιλεγμένων παικτών και συνολική τιμή του αντικειμενικού στόχου (expected impact score) για τη βέλτιστη πεντάδα.

Συνολικά, η πειραματική υλοποίηση του lineup optimization λειτουργεί ως γέφυρα μεταξύ θεωρητικής ανάλυσης και πρακτικής εφαρμογής, επιβεβαιώνοντας ότι τα αποτελέσματα της παρούσας εργασίας μπορούν να αποτελέσουν τη βάση για πιο σύνθετα και ρεαλιστικά συστήματα αθλητικής ανάλυσης στο μέλλον.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] E. Mukhopadhyay, "Optimizing Offensive Gameplan in the National Basketball Association with Machine Learning," *2025 IEEE World AI IoT Congress (AIIoT)*, Seattle, WA, USA, 2025, pp. 0385-0391, doi: 10.1109/AIIoT65859.2025.11105326.

[2] L. Cai, C. Zhao and X. Wang, "Situation and lessons of application of NBA big data technology," *2021 International Conference on Information Technology and Contemporary Sports (TCS)*, Guangzhou, China, 2021, pp. 228-231, doi: 10.1109/TCS52929.2021.00054.

[3] Riccardi, N. (2023). Optimizing NBA Roster Construction. Academy of Economics and Finance Journal, 14, 28-38.

[4] D. Connaghan, P. Kelly, N. E. O'Connor, M. Gaffney, M. Walsh and C. O'Mathuna, "Multi-sensor classification of tennis strokes," *SENSORS, 2011 IEEE*, Limerick, Ireland, 2011, pp. 1437-1440, doi: 10.1109/ICSENS.2011.6127084.

[5] Hassan Ghasemzadeh, Vitali Loseu, Eric Guenterberg, and Roozbeh Jafari. 2009. Sport training using body sensor networks: a statistical approach to measure wrist rotation for golf swing. In Proceedings of the Fourth International Conference on Body Area Networks (BodyNets '09). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Brussels, BEL, Article 2, 1–8. https://doi.org/10.4108/ICST.BODYNETS2009.6035.

[6] Ghosh, I., Ramamurthy, S. R., Chakma, A., Dey, E., Hasan, Z., & Roy, N. (2020). Badminton activity recognition (bar). IEEE Dataport.

[7] Cai, Z., Neher, H., Vats, K., Clausi, D. A., & Zelek, J. (2019). Temporal hockey action recognition via pose and optical flows. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. IEEE.

[8] Piergiovanni, A., & Ryoo, M. S. (2018). Fine-grained activity recognition in baseball videos. In Proceedings of the IEEE Conference on Com-puter Vision and Pattern Recognition (CVPR) Workshops. IEEE.

[9] S. Merckx, P. Robberechts, Y. Euvrard, and J. Davis, “Measuring the effectiveness of pressing in soccer,” in Proc. 8th Workshop on Machine Learning and Data Mining for Sports Analytics, 2021.

[10] S. Almujahed, N. Ongor, J. Tigmo and N. Sagoo, "Sports analytics: Designing a volleyball game analysis decision-support tool using big data," *2013 IEEE Systems and Information Engineering Design Symposium*, Charlottesville, VA, USA, 2013, pp. 19-24, doi: 10.1109/SIEDS.2013.6549487.

[11] Zhang, Libao, et al. "Application of K-means clustering algorithm for classification of NBA guards." *International Journal of Science and Engineering Applications* 5.1 (2016): 1-6.

[12] Patton, A.N. Scott, M., Walker, N., Ottenwess, A., Power, P., Cherukumudi, A. and Lucey, P. (2020). Predicting NBA Talent from Enormous Amounts of College Basketball Tracking Data, , pp. 1-14.

[13] G. Soliman, A. El-Nabawy, A. Misbah and S. Eldawlatly, "Predicting all star player in the national basketball association using random forest," *2017 Intelligent Systems Conference (IntelliSys)*, London, UK, 2017, pp. 706-713, doi: 10.1109/IntelliSys.2017.8324371

[14] Jasmin A. Caliwag, Maria Christina R. Aragon, Reynaldo E. Castillo, and Ellizer Mikko S. Colantes. 2018. Predicting Basketball Results Using Cascading Algorithm. In Proceedings of the 1st International Conference on Information Science and Systems (ICISS '18). Association for Computing Machinery, New York, NY, USA, 64–68. https://doi.org/10.1145/3209914.3209921.

[15] Leicht, A. S., Gómez, M. A., & Woods, C. T. (2017). Explaining Match Outcome During The Men's Basketball Tournament at The Olympic Games. *Journal of sports science & medicine*, *16*(4), 468–473.

[16] Ozkan, I. A. (2020) 'A Novel Basketball Result Prediction Model Using a Concurrent Neuro-Fuzzy System', *Applied Artificial Intelligence*, 34(13), pp. 1038–1054. doi: 10.1080/08839514.2020.1804229.

[17] Vinué G. A Basketball Big Data Platform for Box Score and Play-by-Play Data. *Big Data*. 2025;13(4):285-303. doi:10.1089/big.2023.0177.

[18] Olivo, F. (2023). Advanced basketball analytics. Doctoral dissertation, University of Bologna.

[19] Skinner B, Guy SJ (2015) A Method for Using Player Tracking Data in Basketball to Learn Player Skills and Predict Team Performance. PLoS ONE 10(9): e0136393. https://doi.org/10.1371/journal.pone.0136393.

[20] Yang, Tianyu, Jiang, Congmeng, Li, Pengfei, Video Analysis and System Construction of Basketball Game by Lightweight Deep Learning under the Internet of Things, *Computational Intelligence and Neuroscience*, 2022, 6118798, 14 pages, 2022. doi: https://doi.org/10.1155/2022/6118798.

[21] Yuan, Bin, Kamruzzaman, M. M., Shan, Shaonan, Application of Motion Sensor Based on Neural Network in Basketball Technology and Physical Fitness Evaluation System, *Wireless Communications and Mobile Computing*, 2021, 5562954, 11 pages, 2021. https://doi.org/10.1155/2021/5562954.

[22] Seçkin, A.Ç.; Ateş, B.; Seçkin, M. Review on Wearable Technology in Sports: Concepts, Challenges and Opportunities. *Appl. Sci.* **2023**, *13*, 10399. https://doi.org/10.3390/app131810399.

[23] K. Apostolou and C. Tjortjis, "Sports Analytics algorithms for performance prediction," *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, Patras, Greece, 2019, pp. 1-4, doi: 10.1109/IISA.2019.8900754.

[24] Davis, J., Bransen, L., Devos, L. *et al.* Methodology and evaluation in sports analytics: challenges, approaches, and lessons learned. *Mach Learn* **113**, 6977–7010 (2024). https://doi.org/10.1007/s10994-024-06585-0.

[25] R. Ji, "Research on Basketball Shooting Action Based on Image Feature Extraction and Machine Learning," in *IEEE Access*, vol. 8, pp. 138743-138751, 2020, doi: 10.1109/ACCESS.2020.3012456.

[26] Horvat, T., Havaš, L., & Srpak, D. (2020). The Impact of Selecting a Validation Method in Machine Learning on Predicting Basketball Game Outcomes. *Symmetry*, *12*(3), 431. https://doi.org/10.3390/sym12030431.

[27] Alves, J. M., & Barbosa, R. S. (2025). Machine Learning for Basketball Game Outcomes: NBA and WNBA Leagues. *Computation*, *13*(10), 230. https://doi.org/10.3390/computation13100230.

[28] Kannan, Adarsh; Kolovich, Brian; Lawrence, Brandon; and Rafiqi, Sohail (2018) "Predicting National Basketball Association Success: A Machine Learning Approach," *SMU Data Science Review*: Vol. 1: No. 3, Article 7.

[29] Foteinakis, P. F., Kokkotis, C., Karamousalidis, G., Avloniti, A., Pavlidou, S., Zaras, N., Stampoulis, T., Pantazis, D., Aggelakis, P., Balampanos, D., Liu, J., Laparidis, K., & Chatzinikolaou, A. (2025). From Data to Decisions: Using Explainable Machine Learning to Predict EuroLeague Basketball Outcomes. *Applied Sciences*, *15*(23), 12401. https://doi.org/10.3390/app152312401.

[30] O. Farghaly and P. Deshpande, "Leveraging Machine Learning to Predict National Basketball Association Player Injuries," *2024 IEEE International Workshop on Sport, Technology and Research (STAR)*, Lecco, Italy, 2024, pp. 216-221, doi: 10.1109/STAR62027.2024.10636005.

[31] de Araújo Costa, D., Fechine, J. M., da Silva Brito, J. R., Ferro, J. V. R., de Barros Costa, E., & Lopes, R. V. V. (2024). A Machine Learning Approach Using Interpretable Models for Predicting Success of NCAA Basketball Players to Reach NBA. In *ICAART (3)* (pp. 758-765).

[32] C. Witheesawas and S. Phimoltares, "Winner Prediction Model for Basketball Matches Based on Statistical Data of Teams, Players, and Line-ups," *2024 IEEE 3rd Conference on Information Technology and Data Science (CITDS)*, Debrecen, Hungary, 2024, pp. 1-6, doi: 10.1109/CITDS62610.2024.10791348.

[33] D. Sikka and R. D, "Basketball Win Percentage Prediction using Ensemble-based Machine Learning," *2022 6th International Conference on Electronics, Communication and Aerospace Technology*, Coimbatore, India, 2022, pp. 885-890, doi: 10.1109/ICECA55336.2022.10009313.

[34] Z. Ivanković, M. Racković, B. Markoski, D. Radosav and M. Ivković, "Analysis of basketball games using neural networks," *2010 11th International Symposium on Computational Intelligence and Informatics (CINTI)*, Budapest, Hungary, 2010, pp. 251-256, doi: 10.1109/CINTI.2010.5672237.

[35] P. -H. Hsu, S. Galsanbadam, J. -S. Yang and C. -Y. Yang, "Evaluating Machine Learning Varieties for NBA Players' Winning Contribution," *2018 International Conference on System Science and Engineering (ICSSE)*, New Taipei, Taiwan, 2018, pp. 1-6, doi: 10.1109/ICSSE.2018.8520017.

[36] N. Ma, "NBA Playoff Prediction Using Several Machine Learning Methods," *2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, Taiyuan, China, 2021, pp. 113-116, doi: 10.1109/MLBDBI54094.2021.00030.

[37] Mohamad El-Hajj, Benjamin Kwon, Craeg Jethro Infante, Jackson Steed, Victor Gore, Nhi Phan, Mohammed Elmorsy, "Leveraging machine learning to predict factors that drive successful basketball team formation," Proc. SPIE 13540, Fifth Symposium on Pattern Recognition and Applications (SPRA 2024), 135400B (10 February 2025); https://doi.org/10.1117/12.3056366.

[38] Iatropoulos, D., Sarlis, V., & Tjortjis, C. (2025). A Data Mining Approach to Identify NBA Player Quarter-by-Quarter Performance Patterns. *Big Data and Cognitive Computing*, *9*(4), 74. https://doi.org/10.3390/bdcc9040074

[39] A. Biró, L. Kovács and L. Szilágyi, "Game-Based Learning and Gamified Efficiency for Team Performance Estimation in Professional Basketball," *2025 International Conference on Emerging eLearning Technologies and Applications (ICETA)*, Stary Smokovec, Slovakia, 2025, pp. 95-100, doi: 10.1109/ICETA67772.2025.11278907.

[40] L. Penner, “Player archetypes within basketball: optimizing roster composition to create a championship team”, *Frontiers in Sports and Active Living*, 7, 2025. https://doi.org/10.3389/fspor.2025.1639431

[41] S. Siddique, L. Li and Y. Wang, "Finding Optimal Teams: An Analysis of NBA Statistics and Constraints," *2025 IEEE 11th International Conference on Big Data Computing Service and Machine Learning Applications (BigDataService)*, Tucson, AZ, USA, 2025, pp. 26-34, doi: 10.1109/BigDataService65758.2025.00010.

[42] H. Yongliang and C. Haibin, "A Basketball Player Technical Analysis Framework Based on Decision Tree Optimized Recursive Neural Network," *2023 IEEE International Conference on Image Processing and Computer Applications (ICIPCA)*, Changchun, China, 2023, pp. 1214-1219, doi: 10.1109/ICIPCA59209.2023.10257696.

[43] Tian, C., De Silva, V., Caine, M., & Swanson, S. (2020). Use of Machine Learning to Automate the Identification of Basketball Strategies Using Whole Team Player Tracking Data. *Applied Sciences*, *10*(1), 24. https://doi.org/10.3390/app10010024

[44] Siddique, S. (2024). Teaming Strategy Optimization: An Analysis Of NBA Statistics, Shot Charts, And Constraints. Retrieved from https://digitalcommons.pvamu.edu/pvamu-theses/1535

[45] Kim, P., & Lee, S. (2025). How to assess leader capabilities: Applying AI algorithms to evaluate NBA head coaches. *Journal of Sports Analytics*, *11*. https://doi.org/10.1177/22150218251357538.

[46] Hamdad, L., Benatchba, K., Belkham, F., Cherairi, N. (2018). Basketball Analytics. Data Mining for Acquiring Performances. In: Amine, A., Mouhoub, M., Ait Mohamed, O., Djebbar, B. (eds) Computational Intelligence and Its Applications. CIIA 2018. IFIP Advances in Information and Communication Technology, vol 522. Springer, Cham. https://doi.org/10.1007/978-3-319-89743-1_2

[47] Casals, M., & Martinez, A. J. (2013). Modelling player performance in basketball through mixed models. *International Journal of Performance Analysis in Sport*, *13*(1), 64–82. https://doi.org/10.1080/24748668.2013.11868632

[48] Gonzalez, Adam M.; Hoffman, Jay R.; Rogowski, Joseph P.; Burgos, William; Manalo, Edwin; Weise, Keon; Fragala, Maren S.; Stout, Jeffrey R. Performance Changes in NBA Basketball Players Vary in Starters vs. Nonstarters Over a Competitive Season. Journal of Strength and Conditioning Research 27(3):p 611-615, March 2013. | DOI: 10.1519/JSC.0b013e31825dd2d9

[49] Zuccolotto, P., Sandri, M. & Manisera, M. Spatial performance analysis in basketball with CART, random forest and extremely randomized trees. *Ann Oper Res* **325**, 495–519 (2023). https://doi.org/10.1007/s10479-022-04784-3

[50] T. Zhang, G. Hu and Q. Liao, "Analysis of offense tactics of basketball games using link prediction," *2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS)*, Niigata, Japan, 2013, pp. 207-212, doi: 10.1109/ICIS.2013.6607842.

ΠΑΡΑΡΤΗΜΑ Α : ΚΩΔΙΚΑΣ

```
"""
Πλήρης Κώδικας Υλοποίησης (με σχόλια)
==================================================

Σκοπός:
- Συλλογή NBA δεδομένων (2020-21 έως 2024-25) μέσω nba_api
- Καθαρισμός / φιλτράρισμα (ξεσκαρτάρισμα) παικτών
- Δημιουργία τελικού dataset (features + target Net Rating)
- Clustering: K-Means και Gaussian Mixture Models (GMM)
- Παλινδρόμηση: OLS (συνολικά και ανά cluster)
- Αποθήκευση αποτελεσμάτων σε αρχεία (CSV) για αναπαραγωγιμότητα

Απαιτούμενα πακέτα:
pip install nba_api pandas numpy scikit-learn statsmodels matplotlib

Σημείωση:
Τα ακριβή ονόματα στηλών μπορεί να διαφέρουν ανά endpoint/έκδοση.
Ο κώδικας περιλαμβάνει "ασφαλιστικές δικλείδες" και σχόλια για εύκολη προσαρμογή.
"""

import time
import warnings
from typing import List, Tuple, Optional, Dict

import numpy as np
import pandas as pd
```

```
from nba_api.stats.endpoints import leaguedashplayerstats
from nba_api.stats.library.parameters import SeasonTypeAllStar, PerModeSimple

from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.mixture import GaussianMixture

import statsmodels.api as sm

warnings.filterwarnings("ignore")



# ------------------------------------------------------------
# A.1 Συλλογή δεδομένων (nba_api)
# ------------------------------------------------------------

def fetch_player_stats_for_season(
    season: str,
    season_type: str = SeasonTypeAllStar.regular,   # "Regular Season"
    per_mode: str = PerModeSimple.per_game,          # PerGame (μπορείς να το αλλάξεις αν θες)
    measure_type: str = "Advanced",                 # Advanced stats
    timeout_sleep: float = 0.8
) -> pd.DataFrame:
    """
    Φέρνει player stats για μία σεζόν από το endpoint leaguedashplayerstats.

    - season: π.χ. "2020-21"
    - measure_type: "Advanced" για advanced stats
      (εναλλακτικά: "Base", "Misc", "Scoring", "Usage", "Defense", κ.ά.)
```

```
    Επιστρέφει DataFrame με τις στήλες που δίνει το NBA endpoint.
    """
    # Μικρό sleep για να μειώσουμε πιθανότητα throttling
    time.sleep(timeout_sleep)

    resp = leaguedashplayerstats.LeagueDashPlayerStats(
        season=season,
        season_type_all_star=season_type,
        per_mode_detailed=per_mode,
        measure_type_detailed_defense=measure_type
    )
    df = resp.get_data_frames()[0].copy()
    df["SEASON"] = season
    df["MEASURE_TYPE"] = measure_type
    return df


def fetch_multi_season_advanced_stats(
    seasons: List[str],
    measure_types: List[str] = ["Advanced"],  # μπορείς να προσθέσεις π.χ. ["Advanced","Usage","Scoring"]
    timeout_sleep: float = 0.8
) -> pd.DataFrame:
    """
    Μαζεύει δεδομένα για πολλές σεζόν και (προαιρετικά) πολλά measure_types.
    Ενώνει όλα τα αποτελέσματα σε ένα DataFrame.
    """
    all_parts = []
    for s in seasons:
        for mt in measure_types:
```

```
            try:
                part = fetch_player_stats_for_season(
                    season=s,
                    measure_type=mt,
                    timeout_sleep=timeout_sleep
                )
                all_parts.append(part)
            except Exception as e:
                print(f"[WARN]  Αποτυχία  λήψης  για  season={s},
measure_type={mt}: {e}")
                # περιμένουμε λίγο περισσότερο και συνεχίζουμε
                time.sleep(2.0)

    if not all_parts:
        raise  RuntimeError("Δεν  συλλέχθηκαν  δεδομένα.  Έλεγξε
σύνδεση/endpoint/παραμέτρους.")
    return pd.concat(all_parts, ignore_index=True)



# ------------------------------------------------------------
# Α.2 Καθαρισμός / Φιλτράρισμα (ξεσκαρτάρισμα) & Dataset build
# ------------------------------------------------------------

def coerce_numeric(df: pd.DataFrame, cols: List[str]) -> pd.DataFrame:
    """Μετατρέπει στήλες σε numeric όπου γίνεται."""
    for c in cols:
        if c in df.columns:
            df[c] = pd.to_numeric(df[c], errors="coerce")
    return df



def build_model_dataset_from_advanced(
```

```
    df_adv: pd.DataFrame,
    min_gp: int = 15,
    min_min: float = 250.0,
    dropna_threshold: float = 0.20
) -> pd.DataFrame:
    """
    Δημιουργεί το τελικό dataset:
    - Επιλέγει/κρατά τις απαραίτητες στήλες
    - Κάνει φιλτράρισμα παικτών (min games, min minutes)
    - Διαχειρίζεται missing values

    Σημαντικό:
    Το endpoint Advanced συνήθως περιέχει (ανάμεσα σε άλλα):
    GP, MIN, NET_RATING, OFF_RATING, DEF_RATING, TS_PCT, EFG_PCT,
USG_PCT, AST_PCT, REB_PCT, PIE κ.λπ.
    Αν κάποια στήλη λείπει, απλώς δεν θα χρησιμοποιηθεί (βλ. παρακάτω).
    """

    df = df_adv.copy()

    # Βασικές στήλες αναγνώρισης
    id_cols = [c for c in ["PLAYER_ID", "PLAYER_NAME", "TEAM_ID",
"TEAM_ABBREVIATION", "SEASON"] if c in df.columns]

    # Βασικά φίλτρα συμμετοχής
    for c in ["GP", "MIN"]:
        if c not in df.columns:
            raise ValueError(f"Λείπει η αναμενόμενη στήλη '{c}' από τα
δεδομένα.")

    df = coerce_numeric(df, ["GP", "MIN"])
    df = df[df["GP"] >= min_gp].copy()
```

```
    df = df[df["MIN"] >= min_min].copy()

    # Επιλέγουμε candidate features (κυρίως advanced/efficiency)
    candidate_features = [
        "USG_PCT", "TS_PCT", "EFG_PCT",
        "AST_PCT", "REB_PCT", "OREB_PCT", "DREB_PCT",
        "TOV_PCT",
        "PACE", "PIE",
        "OFF_RATING", "DEF_RATING"  # θα τα χρησιμοποιήσεις ή όχι ανά μοντέλο
    ]

    # Target
    target_col = "NET_RATING"
    if target_col not in df.columns:
        raise ValueError("Λείπει η στήλη target 'NET_RATING' από τα Advanced stats.")

    keep_features = [c for c in candidate_features if c in df.columns]
    keep_cols = id_cols + keep_features + [target_col]

    df = df[keep_cols].copy()

    # Μετατροπή σε numeric για features+target
    df = coerce_numeric(df, keep_features + [target_col])

    # Διαχείριση missing values:
    # 1) Αν μια feature έχει υπερβολικά πολλά NaN, την αφαιρούμε
    feature_na_rate = df[keep_features].isna().mean()
    features_ok = feature_na_rate[feature_na_rate <= dropna_threshold].index.tolist()
```

```
    # 2) Κρατάμε μόνο αυτές τις features
    final_cols = id_cols + features_ok + [target_col]
    df = df[final_cols].copy()

    # 3) Αφαιρούμε γραμμές με NaN στο target
    df = df.dropna(subset=[target_col]).copy()

    # 4) Για τα υπόλοιπα NaN στις features, κάνουμε απλό imputation με median (ανά στήλη)
    for c in features_ok:
        med = df[c].median()
        df[c] = df[c].fillna(med)

    # Reset index
    df = df.reset_index(drop=True)
    return df


# ------------------------------------------------------------
# A.3 Προεπεξεργασία για ML (Scaling) & Feature sets
# ------------------------------------------------------------

def make_feature_sets(df: pd.DataFrame) -> Tuple[List[str], List[str]]:
    """
    Φτιάχνει δύο feature sets:
    - X_full: όλα τα διαθέσιμα features
    - X_no_offdef: όλα εκτός από OFF_RATING & DEF_RATING (αν υπάρχουν)
    """
    ignore_cols = {"PLAYER_ID", "PLAYER_NAME", "TEAM_ID", "TEAM_ABBREVIATION", "SEASON", "NET_RATING"}
    features = [c for c in df.columns if c not in ignore_cols]
```

```
    x_full = features.copy()
    x_no_offdef = [c for c in features if c not in ["OFF_RATING", "DEF_RATING"]]

    # Αν για κάποιο λόγο αδειάσει, κρατάμε το full
    if len(x_no_offdef) == 0:
        x_no_offdef = x_full.copy()

    return x_full, x_no_offdef


def scale_features(df: pd.DataFrame, feature_cols: List[str]) -> Tuple[np.ndarray, StandardScaler]:
    """
    Standardization (z-score) για να δουλεύουν σωστά KMeans/GMM.
    """
    scaler = StandardScaler()
    X = scaler.fit_transform(df[feature_cols].values)
    return X, scaler


# ------------------------------------------------------------
# A.4 Clustering: K-Means & GMM
# ------------------------------------------------------------

def run_kmeans(X: np.ndarray, n_clusters: int = 5, random_state: int = 42) -> np.ndarray:
    km = KMeans(n_clusters=n_clusters, n_init=25, random_state=random_state)
    labels = km.fit_predict(X)
    return labels
```

```
def run_gmm(X: np.ndarray, n_components: int = 5, random_state: int = 42)
-> np.ndarray:
    gmm           =           GaussianMixture(n_components=n_components,
covariance_type="full", random_state=random_state)
    labels = gmm.fit_predict(X)
    return labels



# ------------------------------------------------------------
# Α.5 Παλινδρόμηση OLS (συνολικά & ανά cluster)
# ------------------------------------------------------------

def fit_ols(
    df: pd.DataFrame,
    feature_cols: List[str],
    target_col: str = "NET_RATING"
) -> sm.regression.linear_model.RegressionResultsWrapper:
    """
    OLS με statsmodels, ώστε να έχουμε πλήρες summary (t-stats, p-values,
R^2).
    """
    X = df[feature_cols].copy()
    y = df[target_col].copy()

    # Προσθέτουμε σταθερό όρο
    X = sm.add_constant(X, has_constant="add")
    model = sm.OLS(y, X).fit()
    return model



def                                          ols_report_to_df(model:
sm.regression.linear_model.RegressionResultsWrapper) -> pd.DataFrame:
    """
```

```
    Μετατρέπει βασικά αποτελέσματα OLS σε DataFrame (για export).
    """
    out = pd.DataFrame({
        "coef": model.params,
        "std_err": model.bse,
        "t": model.tvalues,
        "p_value": model.pvalues
    })
    out.index.name = "term"
    return out.reset_index()


def run_ols_global_and_by_cluster(
    df: pd.DataFrame,
    feature_cols: List[str],
    cluster_col: Optional[str] = None,
    target_col: str = "NET_RATING",
    min_cluster_size: int = 50
) -> Dict[str, pd.DataFrame]:
    """
    Τρέχει:
    - Global OLS στο σύνολο
    - OLS ανά cluster (αν δοθεί cluster_col), με φίλτρο ελάχιστου μεγέθους
    Επιστρέφει λεξικό με πίνακες αποτελεσμάτων.
    """
    results = {}

    # Global
    model_global = fit_ols(df, feature_cols, target_col=target_col)
    results["GLOBAL_COEFS"] = ols_report_to_df(model_global)
    results["GLOBAL_META"] = pd.DataFrame([{
```

```
    "n_obs": int(model_global.nobs),
    "r2": float(model_global.rsquared),
    "adj_r2": float(model_global.rsquared_adj),
    "aic": float(model_global.aic),
    "bic": float(model_global.bic)
}])

# By cluster
if cluster_col and cluster_col in df.columns:
    cluster_tables = []
    meta_tables = []

    for cl in sorted(df[cluster_col].unique()):
        dfi = df[df[cluster_col] == cl].copy()
        if len(dfi) < min_cluster_size:
            continue

        m = fit_ols(dfi, feature_cols, target_col=target_col)
        coefs = ols_report_to_df(m)
        coefs.insert(0, "cluster", cl)
        cluster_tables.append(coefs)

        meta_tables.append({
            "cluster": cl,
            "n_obs": int(m.nobs),
            "r2": float(m.rsquared),
            "adj_r2": float(m.rsquared_adj),
            "aic": float(m.aic),
            "bic": float(m.bic)
        })
```

```
        if cluster_tables:
            results[f"{cluster_col}_COEFS"] = pd.concat(cluster_tables,
ignore_index=True)
            results[f"{cluster_col}_META"] = pd.DataFrame(meta_tables)

    return results



# ------------------------------------------------------------
# A.6 Main pipeline (όλα μαζί)
# ------------------------------------------------------------

def main():
    # 1) Σεζόν (Regular Season) 2020-21 έως 2024-25
    seasons = ["2020-21", "2021-22", "2022-23", "2023-24", "2024-25"]

    # 2) Συλλογή advanced stats (μπορείς να προσθέσεις κι άλλα
measure_types αν τα χρειάζεσαι)
    print("Συλλογή δεδομένων από nba_api...")
    df_adv = fetch_multi_season_advanced_stats(
        seasons=seasons,
        measure_types=["Advanced"],  # advanced stats
        timeout_sleep=0.8
    )
    print(f"Raw rows: {len(df_adv):,} | Raw cols: {df_adv.shape[1]}")

    # 3) Καθαρισμός / ξεσκαρτάρισμα και τελικό dataset
    print("Καθαρισμός / φιλτράρισμα & δημιουργία dataset...")
    df = build_model_dataset_from_advanced(
        df_adv=df_adv,
        min_gp=15,       # μπορείς να το αλλάξεις (π.χ. 20)
        min_min=250.0,   # μπορείς να το αλλάξεις (π.χ. 400)
```

```
        dropna_threshold=0.20
    )
    print(f"Final dataset rows: {len(df):,} | cols: {df.shape[1]}")

    # 4) Feature sets
    x_full, x_no_offdef = make_feature_sets(df)
    print("Features (full):", x_full)
    print("Features (no OFF/DEF):", x_no_offdef)

    # 5) Scaling (χρειάζεται για clustering)
    X_full, scaler_full = scale_features(df, x_full)
    X_no, scaler_no = scale_features(df, x_no_offdef)

    # 6) Clustering (KMeans + GMM) πάνω στο full feature set (ή επέλεξε X_no)
    #    Εσύ μπορείς να αποφασίσεις n_clusters / n_components όπως έκανες στην εργασία σου.
    n_clusters = 5
    print(f"Clustering με KMeans (k={n_clusters})...")
    df["CL_KMEANS"]    =    run_kmeans(X_full,    n_clusters=n_clusters, random_state=42)

    print(f"Clustering με GMM (k={n_clusters})...")
    df["CL_GMM"]     =     run_gmm(X_full,     n_components=n_clusters, random_state=42)

    # 7) Παλινδρόμηση OLS:
    #    (α) Global OLS με full features
    #    (β) Global OLS χωρίς OFF/DEF (αν υπάρχουν)
    #    (γ) OLS ανά cluster (KMeans & GMM)
    print("OLS (Global + ανά cluster)...")

    # (α) Full features
```

```
res_full_global = run_ols_global_and_by_cluster(
    df=df,
    feature_cols=x_full,
    cluster_col=None,
    target_col="NET_RATING"
)

# (β) Χωρίς OFF/DEF
res_no_global = run_ols_global_and_by_cluster(
    df=df,
    feature_cols=x_no_offdef,
    cluster_col=None,
    target_col="NET_RATING"
)

# (γ) Ανά cluster KMeans
res_full_km = run_ols_global_and_by_cluster(
    df=df,
    feature_cols=x_full,
    cluster_col="CL_KMEANS",
    target_col="NET_RATING",
    min_cluster_size=50
)

# (δ) Ανά cluster GMM
res_full_gmm = run_ols_global_and_by_cluster(
    df=df,
    feature_cols=x_full,
    cluster_col="CL_GMM",
    target_col="NET_RATING",
    min_cluster_size=50
```

```
    )

    # 8) Αποθήκευση αποτελεσμάτων
    #    (Καλό για παράρτημα/αναπαραγωγιμότητα: dataset + outputs)
    out_dir = "./outputs_appendix"
    print(f"Αποθήκευση αποτελεσμάτων στον φάκελο: {out_dir}")

    import os
    os.makedirs(out_dir, exist_ok=True)

    # Dataset
    df.to_csv(os.path.join(out_dir,  "final_dataset_with_clusters.csv"),
index=False)

    # OLS outputs
    # Full global
    res_full_global["GLOBAL_COEFS"].to_csv(os.path.join(out_dir,
"ols_full_global_coefs.csv"), index=False)
    res_full_global["GLOBAL_META"].to_csv(os.path.join(out_dir,
"ols_full_global_meta.csv"), index=False)

    # No OFF/DEF global
    res_no_global["GLOBAL_COEFS"].to_csv(os.path.join(out_dir,
"ols_no_offdef_global_coefs.csv"), index=False)
    res_no_global["GLOBAL_META"].to_csv(os.path.join(out_dir,
"ols_no_offdef_global_meta.csv"), index=False)

    # By clusters
    if "CL_KMEANS_COEFS" in res_full_km:
        res_full_km["CL_KMEANS_COEFS"].to_csv(os.path.join(out_dir,
"ols_full_by_kmeans_coefs.csv"), index=False)
        res_full_km["CL_KMEANS_META"].to_csv(os.path.join(out_dir,
"ols_full_by_kmeans_meta.csv"), index=False)
```

```
    if "CL_GMM_COEFS" in res_full_gmm:
        res_full_gmm["CL_GMM_COEFS"].to_csv(os.path.join(out_dir,
"ols_full_by_gmm_coefs.csv"), index=False)
        res_full_gmm["CL_GMM_META"].to_csv(os.path.join(out_dir,
"ols_full_by_gmm_meta.csv"), index=False)

    print("Ολοκληρώθηκε επιτυχώς.")


if __name__ == "__main__":
    main()
```