ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

Αξιολόγηση Ομάδων Καλαθοσφαίρισης με Τεχνικές Μηχανικής Μάθησης

Λεωνίδας Βασιλείου

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Ιούνιος 2024

Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

Αξιολόγηση Ομάδων Καλαθοσφαίρισης με Τεχνικές Μηχανικής Μάθησης

Λεωνίδας Βασιλείου

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Ιούνιος 2024

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ..... συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Σ. Μπερσίμης, Καθηγητής (Επιβλέπων)
- Κ. Πολίτης, Αναπληρωτής Καθηγητής
- Σ. Τασουλής, Επίκουρος Καθηγητής

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

EVALUATING BASKETBALL TEAMS WITH MACHINE LEARNING TECHNIQUES

By

Leonidas Vasileiou

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
June 2024

Ευχαριστίες

Το τέλος της συγκεκριμένης εργασίας συμπίπτει με το τέλος των μεταπτυχιακών μου σπουδών και σηματοδοτεί το κλείσιμο ενός μεγάλου κεφάλαιου για εμένα, διότι έθεσε τις βάσεις για την επαγγελματική μου σταδιοδρομία.

Θα ήθελα να ευχαριστήσω όλους τους καθηγητές του πανεπιστημίου, που μου πρόσφεραν ο καθένας ξεχωριστά ένα σημαντικό κομμάτι γνώσεων. Θα ήθελα επίσης να ευχαριστήσω ξεχωριστά τον επιβλέποντα καθηγητή μου κ. Σωτήριο Μπερσίμη, που μου έδωσε την ευκαιρία να ασχοληθώ με ένα αντικείμενο που με ενδιαφέρει στενά και θα ήθελα να ερευνήσω περαιτέρω μελλοντικά.

Τέλος θα ήθελα να ευχαριστηθώ όλους τους συγγράφεις των βιβλιογραφικών πηγών, που μου έδωσαν μια ενδότερη και πιο καθαρή ματιά στον κόσμο των Basketball analytics.

Περίληψη

Στις μέρες μας η ραγδαία ανάπτυξη της μηχανικής μάθησης αποτελεί ένα χρήσιμο εργαλείο σε διάφορους κλάδους της καθημερινότητας. Ένας από αυτούς είναι και η καλαθοσφαίριση.

Στα πλαίσια της διπλωματικής εργασίας εξερευνάται η σχέση της καλαθοσφαίρισης με την μηχανική μάθηση. Πιο αναλυτικά θα μελετηθούν οι παράγοντες που επηρεάζουν την απόδοση μιας ομάδας, με την ανίχνευση του σύγχρονου τρόπου παιχνιδιού μέσω συσταδοποίησης. Επιπλέον ερευνώνται τα χαρακτηριστικά που δίνουν μεγαλύτερη προβλεπτική ικανότητα, αναμεσά στα κλασικά στατιστικά και σε πιο αναλυτικούς δείκτες. Για την επίτευξη των προαναφερθέντων, συλλέχθηκαν δεδομένα από την επίσημη ιστοσελίδα του NBA και εφαρμόστηκαν πολλαπλές μέθοδοι μηχανικής μάθησης, για πιο ακριβή αποτελέσματα.

Ακόμη δίνεται μια εκτενής ματιά στην ιστορία της καλαθοσφαίρισης, από τα πρώιμα χρονιά μέχρι και την εισαγωγή των δεδομένων. Γίνεται αναλυτική βιβλιογραφική ανασκόπηση και ερευνάται επιπλέον το θεωρητικό υπόβαθρο των μεθόδων.

Από τις εφαρμογές προέκυψαν 12 είδη παικτών, καθώς και η σημασία των παικτών που είναι αποτελεσματικοί πίσω από την γραμμή τριών πόντων και είναι κορυφαίοι αμυντικοί, η ρευστότητα των σύγχρονων forward και η σημασία των ψηλών. Για τα προβλεπτικά μοντέλα προέκυψε ότι η καλύτερη μέθοδος διαφέρει ανάλογα με την επιλογή των χαρακτηριστικών. Τα συνελικτικά νευρωνικά δίκτυα και η λογιστική παλινδρόμηση παρουσίασαν την καλύτερη απόδοση για διαφορετικά σύνολα μεταβλητών.

Abstract

Nowadays, the rapid development of machine learning is a useful tool in various branches of everyday life. One of them is basketball.

In the context of the thesis, the relationship between basketball and machine learning is explored. More specifically the factors that affect a team's performance will be studied, by identifying the modern playstyle through clustering. In addition, the characteristics that give greater predictive ability are investigated, among classic statistics and more analytical metrics. To achieve the above, data was collected from the official NBA website and multiple machine learning methods were applied for more accurate results.

Moreover, it is given an extensive look at the history of basketball, from the earlier years to the introduction of analytics in basketball. An analytical literature review is carried out and the theoretical background of the methods is additionally investigated.

From the applications emerged the importance of players who are effective behind the three-point line and are elite defenders, the fluidity of modern forwards and the importance of big men. For predictive models it was found that the best method differs depending on the choice of features. Convolutional neural networks and logistic regression gave the best performance for different sets of variables.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

1.1 Σημασία της ανάλυσης δεδομένων στην καλαθοσφαίριση

Η καλαθοσφαίριση είναι ένα άθλημα που διαχρονικά εξελίσσεται. Παρόλα αυτά η εξέλιξη του αθλήματος δεν έχει επηρεαστεί περισσότερο από ότι τα τελευταία χρονιά. Ο λόγος αυτής της εξέλιξης οφείλεται στην εισαγωγή της ανάλυσης δεδομένων στον χώρο του αθλητισμού. Η ανάλυση δεδομένων έχει γίνει αναπόσπαστο κομμάτι της καλαθοσφαίρισης και αποτελεί πλέον βασικό κριτήριο για την επιτυχία μιας ομάδας. Γίνεται λοιπόν σαφές ότι η μελέτη της απόδοσης των ομάδων μέσω της μηχανικής μάθησης είναι ένα θέμα που μπορεί να οδηγήσει σε μια πιο βαθιά γνώση του αθλήματος.

1.2 Περιγραφή και στόχοι των εφαρμογών

Η ανάλυση των δεδομένων στον χώρο της καλαθοσφαίρισης αποτελείται κατά κύριο λόγο από δυο βασικά πεδία. Την πρόβλεψη αποτελεσμάτων και την ανάλυση της απόδοσης μιας ομάδας ή ενός παίκτη. Σε αυτή την εργασία θα αναλυθούν και τα δύο πεδία.

Στο πρώτο κομμάτι της εργασίας θα μελετηθεί η εξέλιξη του αθλήματος πέρα από το συμβατικό πρότυπο των πέντε κλασικών θέσεων (Point Guard-Shooting guard-Small forward-Power forward-Center), με σκοπό την ανίχνευση του σύγχρονου τρόπου παιχνιδιού και τα διαφορετικά είδη παικτών που επικρατούν. Έπειτα η παραπάνω ανάλυση θα χωριστεί σε 3 κατηγορίες ανάλογα με το ρεκόρ στην κανονική περίοδο και θα επαναληφθεί η ανάλυση με σκοπό να βρεθούν διαφορές στα πρότυπα παικτών ανάμεσα στις καλύτερες και στις χειρότερες ομάδες. Για τα παραπάνω χρησιμοποιήθηκε η μέθοδος συσταδοποίησης K-means. Από την ανάλυση προέκυψαν 12 διαφορετικά είδη παικτών που αναφέρονται λεπτομερώς στην συνέχεια. Επίσης μέσα από τις εφαρμογές αναδεικνύεται η σημασία των ψηλών ως σημείο αναφορά στις ομάδες με τα καλύτερα ρεκόρ, η ρευστότητα των σύγχρονων forward και η μεγαλύτερη εστίαση που δείχνουν οι καλύτερες ομάδες στο κομμάτι της άμυνας.

Στο δεύτερο κομμάτι της εργασίας εφαρμόζεται πρόβλεψη αποτελεσμάτων. Για την πρόβλεψη αποτελεσμάτων θα χρησιμοποιηθούν δυο διαφορετικά σύνολα δεδομένων με διαφορετικά χαρακτηριστικά και πολλαπλές μεθόδους μηχανικής μάθησης, για την πρόβλεψη νίκης ή ήττας μιας ομάδας. Στόχος είναι να μελετηθούν οι διαφορές που προκύπτουν από τα δυο σύνολα δεδομένων. Τα αποτελέσματα που δοθήκαν ήταν ικανοποιητικά με ακρίβεια που ξεπερνούσε το 70%. Τα συνελικτικά νευρωνικά δίκτυα και η λογιστική παλινδρόμηση ξεχώρισαν στην αποτελεσματικότητα τους στα δυο σύνολα δεδομένων.

1.3 Δομή εργασίας

Προχωρώντας στην δομή της εργασίας, αρχικά γίνεται μια εκτενής ανάλυση της καλαθοσφαίρισης και της εξέλιξης της ανά τα χρόνια μέχρι την εισαγωγή των δεδομένων.

Αναφέρονται επίσης οι κυριότεροι παράγοντες που επηρεάζουν το άθλημα συνολικά, όπως και οι διάφορες ανάμεσα στα κορυφαία πρωταθλήματα.

Στο 3ο κεφάλαιο αρχικά παρουσιάζεται μια έρευνα που δείχνει τον αντίκτυπο των δεδομένων στο σύγχρονο NBA και στην συνέχεια γίνεται βιβλιογραφική ανασκόπηση χωρισμένη στα δυο βασικά πεδία που αναφέρθηκαν και προσθέτοντας ένα τρίτο με διαφορά πεδία που έχουν αναλυθεί τα τελευταία χρόνια πάνω στην καλαθοσφαίριση.

Το κεφάλαιο 4 περιέχει μια ανασκόπηση της θεωρίας όλων των μεθόδων μηχανικής μάθησης που θα χρησιμοποιηθούν. Στο κεφάλαιο 5 γίνονται οι εφαρμογές και στο τελευταίο κεφάλαιο καταγράφονται τα συμπεράσματα.

2ο ΚΕΦΑΛΑΙΟ
Γενική συζήτηση για την καλαθοσφαίριση και την μηχανική μάθηση

2.1 Εισαγωγή

Ως Sports Analytics ορίζεται η διερεύνηση και η μοντελοποίηση των αθλητικών επιδόσεων με επιστημονικές τεχνικές. Ένας δημοφιλής τρόπος διερεύνησης γίνεται με την χρήση της μηχανικής μάθησης (Machine Learning(ML)), αναλύοντας ιστορικά δεδομένα, με στόχο την εξαγωγή χρήσιμων συμπερασμάτων. Σκοπός της διαδικασίας είναι η μεγιστοποίηση της απόδοσης ενός αθλητή ή μιας ομάδας. Η εκθετική αύξηση των δεδομένων στον αθλητισμό, συνέβαλε σε πολύ μεγάλο βαθμό στην άνθηση τους. Αποτέλεσμα είναι να αποτελούν πλέον έναν καθοριστικής σημασίας παράγοντα για την επίτευξη των στόχων των οργανισμών.

Η συλλογή και ανάλυση δεδομένων στον χώρο του αθλητισμού δεν είναι κάτι καινούργιο. Πρώιμα δείγματα μπορούν να βρεθούν σε όλα τα αθλήματα. Στην καλαθοσφαίριση η συλλογή δεδομένων, υπάρχει μέσω του παραδοσιακού Box score (συλλογή των βασικότερων κατηγοριών όπως πόντοι, ριμπάουντ, ασίστ), από την σεζόν 1937-38 (Basketball-Reference.com). Οι κατηγορίες που είχαν καλυφθεί ήταν αρκετά περιορισμένες και κατά βάση απευθύνεται στα ομαδικά στατιστικά (πόντοι που σκοράρει ανά αγώνα, πόντοι που δέχεται ανά αγώνα). Είναι προφανές ότι αυτά από μόνα τους δεν μπορούν να δώσουν, παρά μια επιφανειακή απεικόνιση, αλλά ήταν ένα πρώτο δείγμα για τους προπονητές.

Έπειτα στον χώρο του ποδοσφαίρου, μια πρώιμη προσπάθεια από τον Charles Reep ήταν να εφαρμόσει στατιστική ανάλυση για την βελτίωση της απόδοσης της ομάδας του. Από το 1950 είχε ξεκινήσει να καταγραφεί τα δεδομένα της ομάδας της πόλης του και εξήγαγε συμπεράσματα τα οποία βελτίωσαν αισθητά την απόδοση της (Arastey, 2019, Apostolou, 2019).

Στον χώρο του baseball έγινε η πιο οργανωμένη προσπάθεια για στατιστική ανάλυση αγώνων μέχρι την δεκαετία του 70 από την Society of American Baseball Research (SABR) το 1971. Ο Bill James, ενώ δούλευε ως νυχτερινός φύλακας, έγραψε τα θεμελιώδη για το άθλημα Baseball Abstract βιβλία (James B., 1981). Ακόμη ήταν αυτός που εισήγαγε τον όρο ‘’sabermetrics’’ (από τα ακρωνύμια της SABR), δηλαδή της ανάλυσης ενός αγώνα baseball, με την χρήση αναλυτικών δεδομένων.

Η τεχνολογική πρόοδος οδήγησε στην ανάπτυξη της μηχανικής μάθησης. Σύμφωνα με την IBM, μηχανική μάθηση είναι ένας κλάδος της τεχνητής νοημοσύνης (AI), που χρησιμοποιεί δεδομένα και αλγορίθμους, για να μιμηθεί τον τρόπο που σκέφτεται ένας άνθρωπος και να βελτιώσει την ακρίβεια των προβλέψεων (ibm.com/topics/machine-learning). Η εισαγωγή του ορού προήλθε από τον Arthur Samuel, έναν πρωτοπόρο στον χώρο της τεχνητής νοημοσύνης. Η δουλειά του είχε ξεκινήσει από τα τέλη της δεκαετίας του ’50 και μια από τις γνωστότερες δουλειές του είναι πάνω στο παιχνίδι ‘’checkers’’. Το 1962 ο Robert Nealey, ένας από τους καλύτερους παίκτες στον κόσμο, έχασε από τον υπολογιστή που περιείχε το μοντέλο του

Samuel (IBM.com). Η δουλειά του αποτελεί αφετηρία για την ανάπτυξη της μηχανικής μάθησης.

Ξεκινώντας από το ‘’checkers’’, η μηχανική μάθηση απλώθηκε σε όλους τους κλάδους που έχουν πληθώρα δεδομένων. Ένας τέτοιος κλάδος είναι του αθλητισμού και ειδικότερα αυτός της καλαθοσφαίρισης. Στην καλαθοσφαίριση καταγράφονται κατά την διάρκεια κάθε αγώνα τα στατιστικά κάθε παίκτη και της ομάδας συνολικά. Αυτά αφορούν είτε την επίθεση (πόντοι, ασίστ, ποσοστό 3-πόντων), είτε την άμυνα (κλεψίματα, κοψίματα). Παρόλα αυτά, οι δείκτες αυτοί δεν επαρκούν, για να δώσουν μια πλήρης εικόνα του αντίκτυπου ενός παίκτη και της απόδοσης μιας ομάδας (Daly-Grafstein & Bornn, 2019). Χαρακτηριστικά η άμυνα μιας ομάδας ή ενός παίκτη δεν μπορεί να κριθεί αποκλειστικά από τον αριθμό των κλεψιμάτων ή των κοψίματων, διότι παρεμβάλλονται πολλοί παράγοντες. Δεν αποτυπώνεται η πίεση που ασκεί η άμυνα, τα deflections (όταν παρεμβάλλεται ο αμυνόμενος στην πορεία της μπάλας) ή οι προϋποθέσεις που σουτάρει η αντίπαλη άμυνα. Πολλές φορές ένα κλέψιμο μπορεί να έρθει απέναντι σε έναν κακό χειριστή της μπάλας ή ένα κόψιμο μπορεί να προκύψει από την υψομετρική διαφορά επιτιθέμενου- αμυνόμενου. Οι Vaz De Melo et al. (2012) είχαν αφιερώσει ένα ξεχωριστό κομμάτι της έρευνας τους, πάνω σε αυτόν τον ισχυρισμό.

Με δεδομένο αυτές τις παρατηρήσεις προπονητές, αναλυτές και ειδικοί του αθλήματος έκριναν αναγκαία την εισαγωγή νέων δεικτών. Έτσι εισήχθησαν κατηγορίες με πιο ειδικό χαρακτήρα. Αντί να βλέπει ένας προπονητής μόνο πόσους πόντους πέτυχε ένας παίκτης, πλέον έβλεπε και σε πόσες προσπάθειες το κατάφερε. Με την ραγδαία άνοδο των σουτ 3 πόντων έγινε αρκετά χρήσιμο το effective Field goal Percentage (eFG%), που έδινε ένα παραπάνω βάρος στα σουτ 3 πόντων (basketball-reference.com/about/glossary.html).

Για πιο ομαδικές κατηγορίες, οι ομάδες σταμάτησαν να κοιτάνε πόσους πόντους πέτυχαν ή δέχτηκαν και άρχισαν να στρέφουν το βλέμμα τους στον αριθμό κατοχών, καθώς κάποιες είχαν πιο γρήγορο στυλ παιχνιδιού, με απότοκο περισσότερους πόντους σε άμυνα και επίθεση. Αυτό οδήγησε στο Offensive/Defensive Rating, δείκτες που μετρούν τους πόντους που σκοράρει/δέχεται ανά 100 κατοχές. Με αυτό τον τρόπο ήταν πολύ πιο αξιόπιστο να θεωρείται μια ομάδα με χαμηλότερο από τον μέσο ορό DRtg καλή αμυντικά, σε σχέση με μια που είχε περισσότερα κλεψίματα (basketball-reference.com/about/glossary.html).

Όπως γίνεται φανερό η άνθηση των Basketball Analytics και του Machine Learning είναι ένα αναπόφευκτο φαινόμενο και συγχρόνως ένα πολύ χρήσιμο εργαλείο. Η διαρκής εξέλιξη σε πιο ειδικούς δείκτες σε συνδυασμό με την μηχανική μάθηση, μπορεί να οδηγήσει σε μια καλύτερη ‘’ανάγνωση’’ του αθλήματος και σε πιο ασφαλή συμπεράσματα, που οδηγούν στην μεγιστοποίηση της απόδοσης. Όσο ο κλάδος των Analytics αναπτύσσεται, τόσο πιο ανταγωνιστικός γίνεται. Μέρα με την μέρα, οργανισμοί εισέρχονται στον χώρο βλέποντας τις δυνατότητες. Η μηχανική μάθηση έχει εισχωρήσει σε όλες τις πτυχές του μπάσκετ καθαρά αγωνιστικές και μη. Αναλυτικότερα θα μελετηθούν στο 3ο κεφάλαιο, αλλά αναφέρεται ότι κατηγορίες όπως τους στοιχήματος έχουν δει μεγάλη αύξηση τα τελευταία χρονιά.

Η εισαγωγή των Big Data και του AI στον αθλητισμό, έχει κατά γενική παραδοχή επηρεάσει το άθλημα και έχει βοηθήσει πολλές ομάδες να αναπτυχθούν. Η τάση αυτή ωστόσο, δεν έχει την ίδια αποδοχή από όλους, ενώ ταυτόχρονα υπάρχει και σκεπτικισμός, όσον αφορά την αισθητική του παιχνιδιού. Ακόμα και μερικοί διακεκριμένοι αναλυτές όπως ο Kirk Goldsberry, έχουν αναφερθεί στην πιθανότητα οι ομάδες να βασιστούν τόσο πολύ στα δεδομένα, που στο τέλος το άθλημα να πληγεί περισσότερο από όσο ωφελείται (Goldsberry, 2019). Οι Rockets του Daryl Morey, στον οποίο θα γίνει αναφορά αργότερα, παρά την αδιαμφισβήτητη επιτυχία τους, προκάλεσαν ένα κύμα προβληματισμού από την μονοτονία που είχε το παιχνίδι τους. Μέρος του κόσμου της καλαθοσφαίρισης θεωρεί, ότι αν υιοθετηθεί ένα

παρόμοιο σύστημα και από άλλες ομάδες, το άθλημα θα χάσει μεγάλο μέρος των υποστηρικτών του, ακόμα και αν η αποτελεσματικότητα των ομάδων αυξηθεί.

2.2 Ανασκόπηση των Basketball Analytics

Το σημείο καμπής των sport analytics ήρθε στις αρχές του 21ου αιώνα, όταν ο Billy Bean με βοηθό τον Paul DePodesta, πρωτοστάτησαν στον κόσμο του baseball με την ομάδα Oakland Athletics. Ο στόχος των δυο ήταν να παρουσιάσουν μια ανταγωνιστική εικόνα της ομάδας με πολύ μικρότερο προυπολογισμό από τις υπόλοιπες κορυφαίες ομάδες της λίγκας και το κατάφεραν χρησιμοποιώντας στατιστική ανάλυση, για την εύρεση παιχτών χαμηλής αξίας που μπορούσαν να αποδώσουν το ίδιο με έναν πιο ακριβό παίχτη. Η προσέγγιση αυτή επέφερε αναπάντεχα αποτελέσματα, και έγινε ο προάγγελος για την ευρεία διάγνωση των δεδομένων στον χώρο του αθλητισμού. (Lewis, 2004).

Για το μπάσκετ αναλυτικότερα τα δεδομένα δεν ήταν κάτι άγνωστο, καθώς οι ομάδες χρησιμοποιούσαν δεδομένα πολύ πριν την ευρεία χρήση που υπάρχει πλέον. Τα δεδομένα που χρησιμοποιούνταν ήταν πολύ βασικά και επιπλέον οι ομάδες δεν χρησιμοποιούσαν κάποια στατιστική ανάλυση για να εξάγουν συμπεράσματα, αλλά βασίζεται κυρίως σε οπτικές εκτιμήσεις. Η πρώτη ομάδα που βασίστηκε σε στατιστικούς δείκτες και άρχισε να δείχνει τον δρόμο για την χρήση των analytics ήταν οι San Antonio Spurs με General Manager των R.C. Buford. Ο Buford με γνώμονα τα δεδομένα έχτισε μέσω του draft (*διαδικασία επιλογής παικτών που έχουν δηλώσει διαθεσιμότητα για το NBA*) μια δυναστεία, καθώς οι Spurs έγιναν μια ομάδα που σταθερά έφτανε πολύ μπροστά στα play offs για παραπάνω από 15 χρόνια. Χαρακτηριστικό είναι ότι το πρώτο τους πρωτάθλημα ήρθε το 1999 και το πέμπτο το 2014, με όλες τις υπόλοιπες χρονιές να χαρακτηρίζονται από αξιοσημείωτες πορείες. (Alamar, 2018).

Παρόλα αυτά ο άνθρωπος που έφερε την επανάσταση των analytics θεωρείται ο Daryl Morey. Ο Daryl Morey χρησιμοποίησε στατιστικά μοντέλα σε όλες τις πτυχές του παιχνιδιού, με πιο χαρακτηριστική κίνηση την ραγδαία μείωση των σουτ από μέση απόσταση (*η περιοχή λίγο μέσα από την γραμμή των 3 πόντων*) και την κατακόρυφη άνοδο των σουτ πίσω από την γραμμή του τρίποντου (Chen, 2018). Για την επίτευξη του στόχου αυτού έφερε τον Mike D'Antoni, έναν επαναστατικό τακτικά προπονητή με έφεση στο σουτ πίσω από τα 6.75 μέτρα (εκεί βρίσκεται η γραμμή του τρίποντου) και τον James Harden, που από 6ος παίκτης (*ο πρώτος παίκτης που μπαίνει μετά την βασική πεντάδα*) στους Oklahoma City Thunder, έγινε το επίκεντρο για μια από τις πιο πρωτοποριακές επιθέσεις που θα άλλαζαν μέχρι σήμερα την προσέγγιση των ομάδων. Η μέθοδος του Morey οδήγησε σταθερά τους Rockets και μετέπειτα τους Philadelphia 76ers, σε πολύ υψηλά ποσοστά νικών στην κανονική περίοδο (*τα 82 παιχνίδια πριν τα play offs*) και σε μακρινές πορείες στα playoffs (*η τελική φάση του πρωταθλήματος*), όμως στην θητεία του μέχρι σήμερα δεν έχει καταφέρει να φτάσει σε κάποιο πρωτάθλημα ή τελικό. Το παράδειγμα των Spurs και κυρίως του Morey ακολουθούν σταδιακά και οι υπόλοιπες ομάδες, με τα analytics να αποτελούν πλέον αναπόσπαστο κομμάτι για όλες τις ομάδες που θέλουν να ανέβουν επίπεδο.

2.3 Παράγοντες που επηρεάζουν την απόδοση μιας ομάδας

Σκοπός κάθε ομάδας στην καλαθοσφαίριση είναι πάντοτε η κατάκτηση της κορυφής. Για την επίτευξη αυτού του στόχου όμως παρεμβάλλονται πολλαπλοί παράγοντες που μπορούν να επηρεάσουν καθοριστικά την πορεία μιας ομάδας. Σε αυτό το κομμάτι θα αναλυθούν μερικοί από αυτούς τους παράγοντες που παίζουν κομβικό ρολό στην απόδοση της ομάδας.

2.3.1 Παίκτες

''I think the players win the championship, and the organization has something to do with it, don't get me wrong. But don't try to put the organization above the players.''
Michael Jordan

Ο πρώτος και ο κατά γενική ομολογία πιο σημαντικός παράγοντας της απόδοσης μιας ομάδας είναι οι ίδιοι οι παίχτες. Οι παίχτες είναι αυτοί που αντικατοπτρίζουν το τελικό αποτέλεσμα, αυτοί που θα πάρουν τις δύσκολες αποφάσεις, που θα πάρουν τα δύσκολα σουτ και που έχουν την μεγαλύτερη πίεση, καθώς η δουλειά τους εκτίθεται δημόσια. Όπως γίνεται λογικό η κάθε ομάδα έχει μεγαλύτερες πιθανότητες, για μια καλύτερη πορεία αν διαθέτει στο δυναμικό της ποιοτικούς παίχτες που μπορούν να κάνουν πολλαπλές δουλειές στο γήπεδο. Μερικές φορές το μόνο που χρειάζεται μια ομάδα είναι έναν παίχτη υψηλού επιπέδου για να μπορέσει να ''τρυπήσει'' το ταβάνι της και να φτάσει στην κορυφή. Μια τέτοια είναι η περίπτωση των Toronto Raptors το 2019, όπου αν και είχαν αξιοπρεπείς πορείες για αρκετά χρονιά κατάφεραν να κατακτήσουν το πρωτάθλημα μόλις έφεραν τον Kawhi Leonard, έναν από τους κορυφαίους παίχτες εκείνη την περίοδο.

Μια κοινή παρανόηση, είναι ότι η μεγαλύτερη ποιότητα του έμψυχου δυναμικού συνεπάγεται και σε καλύτερα αποτελέσματα. Αν και υπάρχει λογική πίσω από αυτό το επιχείρημα, έχει αποδειχθεί σε πολλαπλές περιπτώσεις ότι δεν είναι τόσο απλό όσο φαίνεται. Οι παίχτες δεν μπορούν απλά να τοποθετηθούν στο γήπεδο και να κερδίζουν. Για να δουλέψει το κάθε εγχείρημα, οι παίχτες θα πρέπει να δουλέψουν εντατικά για μήνες, με σκοπό να αποκτήσουν την απαραίτητη ''χημεία''.

Ένα άλλο εμπόδιο που συναντάνε οι ομάδες με πολλούς κορυφαίους παίχτες, είναι εξωαγωνιστικοί παράγοντες όπως ο εγωισμός. Πολλοί stars έχοντας συνηθίσει να είναι ο κύριος πόλος της ομάδας, δυσκολεύονται να προσαρμοστούν δίπλα σε παίχτες ίδιου επιπέδου. Χαρακτηριστική είναι η περίπτωση των Brooklyn Nets. Οι Nets ήταν μια ομάδα με κυρίως νεαρούς παίχτες που δουλεύαν επί χρονιά μαζί, έχοντας χτίσει μια ελκυστική εικόνα για τον θεατή, χωρίς όμως κάποια ουσιώδη επιτυχία. Για αυτό ανταλλάξανε τον νεανικό κορμό για 2 από τους καλυτέρους παίχτες στο πρωτάθλημα (Irving, Durant) και ανταλλάξανε μερικούς ακόμα για τον James Harden την επόμενη χρονιά, άλλον έναν σπουδαίο γκαρντ (basketball-reference.com/leagues/NBA_2020_transactions.html). Στα χαρτιά το ταλέντο των τριών ήταν υπεραρκετό για να μπορέσει να έχει πολύ υψηλές βλέψεις ο οργανισμός. Παρόλα αυτά οι προσωπικότητες των τριών δεν επέτρεψαν να διατηρηθούν για παραπάνω από 1.5 χρόνο, λόγω προβλημάτων που είχαν αναπτυχθεί στις σχέσεις των παιχτών.

Οι ομάδες που συνήθως έχουν την μεγαλύτερη επιτυχία, είναι όσες έχουν έναν σταθερό κορμό παιχτών για χρονιά. Αυτές οι ομάδες δεν είναι απαραίτητο να έχουν στην διάθεση τους τους πιο ακριβούς παίχτες, αλλά παίχτες που έχουν αναπτύξει ομοιογένεια και αναβαθμίζουν το ταλέντο τους μέσα από την εύρυθμη λειτουργία της ομάδας. Μια τέτοια είναι η περίπτωση των Golden State Warriors. Οι Warriors που διαχρονικά ήταν μια ομάδα χωρίς πολλές επιτυχίες, διατηρούσε επί χρονιά έναν σταθερό κορμό παιχτών και αναπτυσσόταν σταδιακά. Αποτέλεσμα αυτής της διαδικασίας ήταν να έρθει το πρώτο πρωτάθλημα το 2015 και έπειτα με την προσθήκη του Kevin Durant έγιναν μια από τις καλύτερες ομάδες όλων των εποχών.

2.3.2 Προπονητές και τεχνικό επιτελείο

Πέρα από τους παίχτες ιδιαίτερα σημαντική είναι η παρουσία ενός ικανού προπονητή. Οι προπονητές αναλαμβάνουν να διαχειριστούν το εκάστοτε σύνολο με σκοπό την μεγιστοποίηση

της απόδοσης . Καλούνται να διαχειριστούν τα rotations (η σύνθεση που βρίσκεται στο γήπεδο), όταν υπάρχει ενδεχόμενη κούραση, προβλήματα με φάουλ, κακές αποδόσεις παικτών, ενώ πρέπει να είναι έτοιμοι να κάνουν τακτικές προσαρμογές στις κρίσιμες στιγμές (Zhang, 2018). Παραπάνω ειπώθηκε ότι οι παίκτες αποτελούν τον σημαντικότερο παράγοντα για την τελική απόδοση της ομάδας. Αυτό βέβαια δεν θα πρέπει να υποβαθμίσει την σημασία του ρολού του προπονητή.

Ένας αξιόλογος προπονητής μπορεί να εκτοξεύσει σύνολα που έχουν χαμηλότερης αξίας παίκτες. Ο Larry Brown είναι ένα από τα γνωστότερα παραδείγματα. Την σεζόν 2004-2005 κατάφερε να πάρει το πρωτάθλημα απέναντι στους πρωταθλητές τα τελευταία 3 χρόνια Lakers, χωρίς να έχει στο roster του κάποιον All-Star (*Διαδικασία επιλογής των κορυφαίων παικτών μέχρι τα μέσα της χρονιάς. Η ψηφοφορία γίνεται από φίλαθλους- δημοσιογράφους- προπονητές*). Μέχρι και σήμερα δεν έχει καταφέρει κάποια ομάδα κάτι παρόμοιο.

Η ικανότητα ενός προπονητή να πάρει από έναν παίκτη το μέγιστο των δυνατοτήτων του, είναι ίσως το πιο υποτιμημένο κομμάτι της δουλείας τους. Μια τέτοια περίπτωση είναι ο Erik Spoelstra. Ο Spoelstra είναι προπονητής των Miami Heat και ένα από τα γνωρίσματα του, είναι να βγάζει το μέγιστο των δυνατοτήτων από παίκτες χαμηλότερης δυναμικότητας, με μεγάλο μέρος του roster να απαρτίζεται από undrafted παίκτες ή από παίκτες που άλλες ομάδες δεν πιστεύανε ότι είχαν τις απαραίτητες ικανότητες.

2.3.3 Διοικητικά στελέχη

Ανεβαίνοντας πλέον στο διοικητικό σκέλος γίνεται εμφανές ότι η πραγμάτωση των παραπάνω γίνεται εφικτή κυρίως μέσω των ανωτέρων στελεχών. Αυτοί είναι που μέσω του κεφαλαίου θέτουν σε κίνηση τον οργανισμό. Καθορίζουν τον προϋπολογισμό και έχουν τον τελικό λόγο για κάθε κίνηση αγωνιστική ή μη. Οι δυο βασικότεροι ρόλοι είναι συνήθως του ιδιοκτήτη και του γενικού διευθυντή. Ο κάθε ιδιοκτήτης έχει την απολυτή ελευθέρια να παρεμβεί σε οποιοδήποτε κομμάτι κρίνει αυτός απαραίτητο. Όμως στις πιο πετυχημένες ομάδες, η πιο κοινή τακτική είναι να μην παρεμβαίνουν σε υπερβολικό βαθμό και να αφήνουν κυρίως τις αποφάσεις τους σε άτομα που έχουν μεγαλύτερη εμπειρία στον χώρο. Ο πρόεδρος των New York Knicks (NYK) αντιπροσωπεύει τις δυο τακτικές και τις επιπτώσεις τους. Αν και έχει στην κατοχή του έναν από τους πιο ελκυστικούς οργανισμούς στον κόσμο, η ομάδα του επί χρονιά κυμαίνονταν από πολύ μέτριες μέχρι απογοητευτικές πορείες. Εκείνα τα χρονιά χαρακτηρίζονταν από βιαστικές και εκ του αποτελέσματος λανθασμένες αποφάσεις που έπαιρνε ο ίδιος. Την τελευταία τριετία όμως, αποφάσισε να δώσει μεγαλύτερη ελευθερία στον καινούργιο γενικό διευθυντή Leon Rose. Το αποτέλεσμα είναι μια κατακόρυφη άνοδος, με την ομάδα να είναι στα play off 2 από τα 3 τελευταία χρονιά και μάλιστα φέτος να φτάνει στον 2ο γύρο, κάτι που είχε να συμβεί από το 2013.

Ο γενικός διευθυντής είναι αυτός που επιβλέπει όλες τις πτυχές του οργανισμού. Αναλαμβάνει να διαχειριστεί από τα οικονομικά για τις μετάγραφες μέχρι το scouting. Έχει διαρκή επικοινωνία από παίχτες και προπονητές μέχρι την διοίκηση. Ο ρόλος του γενικού διευθυντή είναι κομβικής σημασίας, καθώς θέτει την κατεύθυνση που θα λάβει η ομάδα. Το έργο τους είναι απαιτητικό, διότι θα πρέπει συχνά να δουλέψουν με περιορισμένο προϋπολογισμό και λαμβάνουν μεγάλο μερίδιο ευθύνης σε περίπτωση αποτυχίας. Για αυτό το λόγο και έχουν απόλυτη ελευθερία κινήσεων (Juravich, 2017). Οι Juravich et al.(2017) μέσω της θεωρίας upper echelon κατέληξαν ότι το επίπεδο του γενικού διευθυντή και η απόδοση της ομάδας έχουν θετική συσχέτιση. Γενικοί διευθυντές σαν τον Daryl Morey, ο οποίος έχει ήδη αναφερθεί, στην ανάγκη να βρουν νέους τρόπους να επιτύχουν, κάνανε τα analytics να είναι ένας από τους κύριους μοχλούς στην σύγχρονη καλαθοσφαίριση. (Juravich et al., 2017)

2.4 Διαφορές ανάμεσα στα δύο κορυφαία πρωταθλήματα

Η καλαθοσφαίριση είναι ένα από τα δημοφιλέστερα αθλήματα παγκοσμίως. Αυτό που το κάνει να ξεχωρίζει σε σχέση με τα αλλά, είναι η ποικιλομορφία που παρουσιάζει στους κανόνες και αυτό φαίνεται όταν παρατηρούμε τις διάφορες που υπάρχουν στο ευρωπαϊκό και στο αμερικανικό μπάσκετ. Αν και ο σκοπός παραμένει ο ίδιος, πολλά πράγματα φαίνεται να αλλάζουν όταν παρακολουθεί κάποιος ένα παιχνίδι στην Ευρώπη και ένα στην Αμερική και συνεπώς αλλάζουν τα κριτήρια αξιολόγησης. Στην Αμερική μάλιστα υπάρχουν ακόμα μεγαλύτερες αλλαγές όταν μιλάμε για κολεγιακό πρωτάθλημα και για το επαγγελματικό. Αρχικά θα επικεντρωθούμε στις διάφορες που παρουσιάζουν οι δυο πιο γνωστές και κορυφαίες βάσει ταλέντου λίγκες, το NBA και η Euroleague. Οι διαφορές δεν είναι απλά τεχνικές, αλλά τόσο βασικές που επηρεάζουν σε μεγάλο βαθμό την απόδοση παικτών και ομάδων του εκάστοτε πρωταθλήματος.

Πιο συγκεκριμένα η Euroleague έχει το κλασικό χρονόμετρο που έχει επικρατήσει στην μεγάλη πλειοψηφία των διοργανώσεων και είναι στα 10 λεπτά ανά περίοδο. Στο NBA αυτός ο χρόνος είναι αυξημένος κατά συνολικά 8 λεπτά, δηλαδή 2 ανά περίοδο (official.nba.com/rule-no-5-scoring-and-timing/). Αν και τα 2 λεπτά μπορεί να μην φαίνονται τόσα πολλά, αλλάζουν την τακτική προσέγγιση, διότι δίνεται η ευκαιρία σε παίκτες να αναδείξουν το ταλέντο τους, κάτι που σε μικρότερο χρόνο ενδεχομένως να μην μπορούσε να γίνει. Η αυξημένη διάρκεια ενέχει και κίνδυνους, αφού λόγω της έντασης του αθλήματος, θα πρέπει να διαλέγονται πιο προσεκτικά τα διαστήματα που θα ξεκουράζονται οι παίκτες, για να αποφεύγονται οι τραυματισμοί.

Η άλλη βασική διαφορά στους κανόνες αφορά τις διαστάσεις του γηπέδου. Τα ευρωπαϊκά γήπεδα έχουν διαστάσεις 28x15 μέτρα, ενώ τα γήπεδα του NBA 28.65x15.24 μέτρα. Επιπλέον η γραμμή του τρίποντου στην Ευρώπη είναι στα 6.75μ και στο NBA στα 7.25μ (jr.nba.com/3-point-shot/). Αυτή η διαφορά δείχνει να ευνοεί τους σουτέρ που παίζουν στην Ευρώπη περισσότερο, αλλά δεν είναι τόσο απλό όσο φαίνεται. Η διαφορά αυτή είναι σημαντική αλλά πρέπει να επισημάνουμε μια ακόμα διαφοροποίηση για να δούμε αν όντως αυτό το μισό μετρό είναι τόσο σημαντικό. Τα γήπεδα στις δυο ηπείρους διαφέρουν στις διαστάσεις τους, με τους αγωνιστικούς χώρους να είναι μεγαλύτεροι σε μήκος και πλάτος στο NBA. Αυτό προκαλεί διάφορες στις προϋποθέσεις του σουτ και στους χώρους που δημιουργούνται για ‘’slashers’’ (*παίκτες με έφεση να διεισδύουν κοντά στο καλάθι*).

Το NBA ειδικότερα τα τελευταία χρονιά, φαίνεται να μην βασίζεται τόσο στους παραδοσιακούς ‘’ψηλούς’’(*Nba.com*.) και να βασίζεται κυρίως σε κοντύτερους και πιο ευκίνητους παίκτες. Αντίθετα στην Ευρώπη οι παραδοσιακοί ‘’ψηλοί’’ ευδοκιμούν και παραμένουν βασικό κομμάτι των πετυχημένων ομάδων, σαν τον Edy Tavares της πρωταθλήτριας Ευρώπης Ρεάλ Μαδρίτης. Ένας λόγος που παρατηρείται αυτό, είναι ο επιτρεπόμενος χρόνος που μπορεί να μείνει κάποιος στο ‘’ζωγραφιστό’’(*Η περιοχή ανάμεσα στο καλάθι μέχρι την γραμμή ελεύθερων βολών*). Στην Euroleague δεν υπάρχει κάποιο χρονικό περιθώριο, για την παραμονή στο ‘’ζωγραφιστό’’, σε αντίθεση με το NBA που δεν επιτρέπει σε κανέναν να μένει για παραπάνω από 3 δευτερόλεπτα. Η παράβαση αυτή επιφέρει μια βολή για την αντίπαλη ομάδα. Αυτός ο κανόνας έχει δώσει μια ροπή στις ομάδες να ψάχνουν για πιο αθλητικές περιπτώσεις παικτών θυσιάζοντας το ύψος σε πολλές περιπτώσεις, επειδή παραδοσιακά οι πιο ψηλοί είναι πιο αργοί σε ταχύτητα και κοστίζουν αμυντικά για την ομάδα τους με αυτόν τον κανόνα.

Οι Milanović et al. (2014) μετά από στατιστική έρευνα για τις διαφορές σε ευρωπαϊκό και αμερικανικό μπάσκετ, κατέληξαν ότι το NBA βασίζεται στις επιθέσεις στο transition και ότι το

επιθετικό παιχνίδι είναι πιο αποτελεσματικό. Για το ευρωπαϊκό μπάσκετ υπάρχουν περισσότερες οργανωμένες επιθέσεις και μεγαλύτερος αριθμός πασών σε κάθε κατοχή (Milanović et al., 2014).

Όλες οι παραπάνω διάφορες συνδράμουν σε πολύ διαφορετικό στυλ παιχνιδιού. Οι Selmanovic et al. (2015) συμπέραναν ότι οι διάφορες στους κανόνες έχουν ως αποτέλεσμα στην Ευρώπη το παιχνίδι να εστιάζει περισσότερο σε set επιθέσεις (*επιθέσεις που δεν είναι αιφνιδιασμοί και βασίζονται στο σύστημα της ομάδας*) και σε περισσότερες δράσεις pick and roll (*Η δράση κατά την οποία ένας παίκτης κάνει ένα screen στον παίκτη που μαρκάρει τον χειριστή της μπάλας, έτσι ώστε να δημιουργηθεί αμυντική ανισορροπία. Ταυτόχρονα ο παίκτης που κάνει το screen διεισδύει στο καλάθι*). Στο NBA δίνεται περισσότερη έμφαση στο transition offense (*επιθέσεις που βασίζονται σε ταχύτερες αποφάσεις για να δημιουργήσουν ευκαιρίες σκοραρίσματος*) και στο pick and pop (*Παρόμοια δράση με το pick and roll, με την διαφορά ότι ο παίκτης που κάνει το screen παραμένει μακριά από το καλάθι για να βρει σουτ με ευνοϊκές συνθήκες*).

2.5 Η εξέλιξη του αθλήματος

Το μπάσκετ είναι κατά παραδοχή ένα από τα συναρπαστικότερα αθλήματα παγκοσμίως. Ένας λόγος είναι ότι σε αντίθεση με αλλά αθλήματα, διαρκώς εξελίσσεται, με αποτέλεσμα να ανανεώνεται το ενδιαφέρον του θεατή. Η εξέλιξη αυτή οφείλεται σε διάφορους παράγοντες όπως η αλλαγή κανόνων, ή πιο πρόσφατα η εισαγωγή των analytics. Η αλλαγή του τρόπου του παιχνιδιού διαχρονικά ξεκίναγε από το NBA και σταδιακά απλωνόταν και στον υπόλοιπο κόσμο. Για αυτό και η ανάδρομη θα βασιστεί στο αμερικανικό πρωτάθλημα (NBA). Στα πρώτα χρονιά του επαγγελματικού μπάσκετ, οι πιο ψηλοί παίχτες ήταν αυτοί που ξεχώριζαν, λόγω της σωματικής διάπλασης τους και της έλλειψης τεχνογνωσίας που δεν ήταν ανεπτυγμένη ακόμα. Χαρακτηριστικά το βραβείο του πολυτιμότερου παίχτη κατακτήθηκε, από την στιγμή που ιδρύθηκε(1955), 20 φορές στα 24 χρονιά από έναν Center (nba.com/news/history-mvp-award-winners). Τέτοιοι παίχτες ήταν οι George Mikan, Bill Russell και Wilt Chamberlain (Goldsberry, 2019).

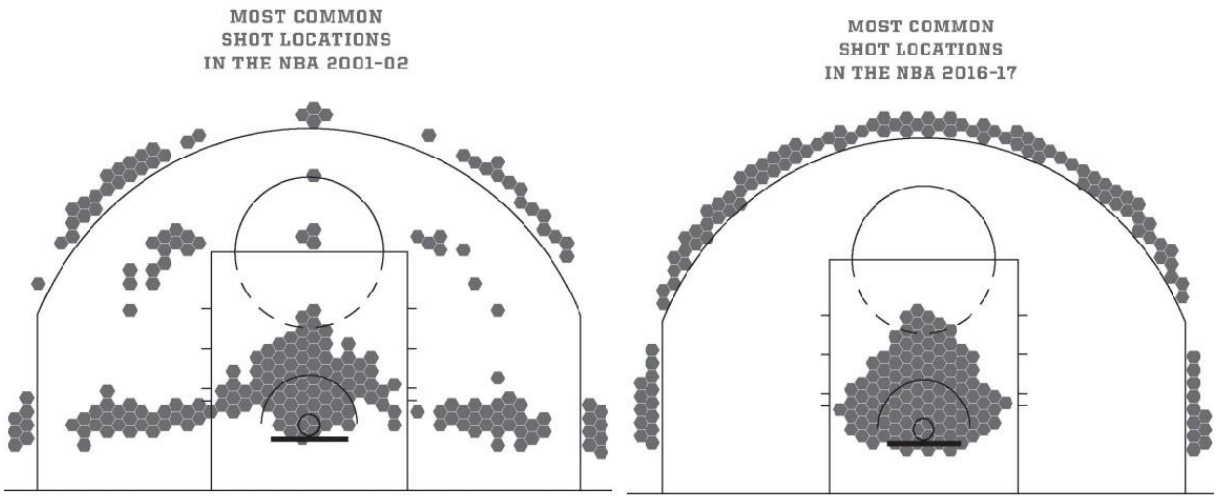
Η κυριαρχία τέτοιων παικτών ήταν τέτοια, που η ομοσπονδία αποφάσισε να εισάγει κανόνες, με σκοπό οι πιο μικρόσωμοι παίκτες να έχουν μεγαλύτερη συνεισφορά και οι πιο ογκώδεις να μην υπερέχουν με τόση ευκολία. Τέτοιοι κανόνες ήταν η μεγαλύτερη απόσταση που εκτελούσαν ‘’ελεύθερες βολές’’, η διεύρυνση του ζωγραφιστού (Goldsberry, 2019).

Η αλλαγή που έδωσε πραγματικά μια άλλη διάσταση στο παιχνίδι όμως, ήταν η προσθήκη της γραμμής των 3-ποντων. Οι παίχτες μέχρι την εισαγωγή της γραμμής δεν είχαν λόγο να σουτάρουνε από μακρινές αποστάσεις, εφόσον δεν είχε όφελος να επιχειρήσει μια πιο δύσκολη προσπάθεια για την ίδια παραγωγή πόντων. Αυτό οδηγούσε σε ένα παιχνίδι, στο οποίο όλοι οι παίχτες ήταν πολύ κοντά μεταξύ τους, εξού και η κυριαρχία των ψηλότερων παικτών. Ο Geroge Mikan στην μετέπειτα καριέρα του έγινε ο πρώτος κομισάριος του ABA και εισήγαγε την γραμμή 3 πόντων (Goldsberry K., 2019) Έτσι οι παίχτες είχαν λόγο να σουτάρουν από πιο μακριά, αφού μπορούσαν να δώσουν περισσοτέρους πόντους, άρα και μεγαλύτερη πιθανότητα νίκης.

Η κίνηση αυτή οδήγησε τους ψηλούς παίχτες να φεύγουν πιο μακριά από το καλάθι και έδωσε την ευκαιρία σε πιο μικρόσωμους παίκτες να διεισδύουν στο καλάθι και να αναδειχθούν ανά τα χρονιά. Με την πάροδο του χρόνου η λίγκα συνέχισε προς αυτή την κατεύθυνση, δηλαδή να διαμορφώνει το παιχνίδι με τρόπο τέτοιο που η ισχύς των ψηλών να μειώνεται. Η κατάργηση του hand checking (η δυνατότητα ενός αμυντικού να χρησιμοποιεί το χέρι του για να σταματήσει τον επιτιθέμενο), άνοιξε περισσότερο την ‘’ψαλίδα’’ μεταξύ guard και center.

Πλέον τα βραβεία του πολυτιμότερου παίκτη κατέληγαν σε παίκτες όπως ο Steve Nash (nba.com/news/history-mvp-award-winners), που βασιζόντουσαν στα ευστοχά σουτ, τις θεαματικές πάσες και την ευφυΐα τους. Αξίζει να αναφερθεί ότι ο Steve Nash έχει ύψος 191 εκ., ενώ παλαιότεροι νικητές του MVP βραβείου είχαν ύψη κοντά η και παραπάνω από 210 εκ. (nba.com/news/history-mvp-award-winners).

Από τις αρχές του 21ου αιώνα οι guards πλέον ήταν αυτοί καθορίζουν σε μεγαλύτερο βαθμό την απόδοση μιας ομάδας. Μοναδική εξαίρεση αποτελούσε ο Shaquille O'Neal, ένας θηριωδής center, που θύμιζε πιο πολύ τους κυριαρχικούς παίκτες του παρελθόντος. Η εισαγωγή όμως των analytics στον χώρο της καλαθοσφαίρισης, ήταν αυτή που έφερε ακόμα πιο σαρωτικές αλλαγές και έθεσε τους κλασικούς center, ουσιαστικά εκτός πρωταθλήματος. Η μεγαλύτερη αλλαγή που ήρθε με την άφιξη τους, ήταν η μεγαλύτερη συχνότητα σουτ από την περιφέρεια και η εξάλειψη σουτ από το mid-range (*η περιοχή έξω από το ζωγραφιστό μέχρι την γραμμή του τρίποντου*). Ενδεικτικά παρουσιάζονται από τον Kirk Goldsberry οι πιο συχνές περιοχές που σουτάραν οι παίχτες την σεζόν 2001-2002 και την σεζόν 2016-2017.



Σχήμα 2.1. Γράφημα από τον Kirk Goldsberry για τα σημεία που επιχειρούνται πιο συχνά τα σουτ τις σεζόν 2001-2002 και 2016-2017 (Πηγή:Goldsberry (2019))

Παράλληλα οι παίκτες που δεν σουτάραν αποτελεσματικά και που δεν ήταν ιδιαίτερα αθλητικοί, σταδιακά παραγκωνίστηκαν. Ο Roy Hibbert αν και ήρθε δεύτερος στην ψηφοφορία του αμυντικού της χρονιάς την σεζόν 2013-14, δεν ήταν αποτελεσματικός σουτέρ και λόγω του ύψους του, ήταν αρκετά δυσκίνητος (basketball- reference.com/awards/awards_2014.html). Τέσσερα χρόνια μετά, βρισκόταν εκτός πρωταθλήματος.

Κατά πλειοψηφία πλέον οι κορυφαίες ομάδες είχαν στο ενεργητικό τους έναν κορυφαίο guard είτε με playmaking ικανότητες (Kyle Lowry, Chris Paul) είτε με μεγάλη έφεση στο σκοράρισμα (Stephen Curry, Jamal Murray). Το παιχνίδι πλέον χρειάζεται αλτικότητα και

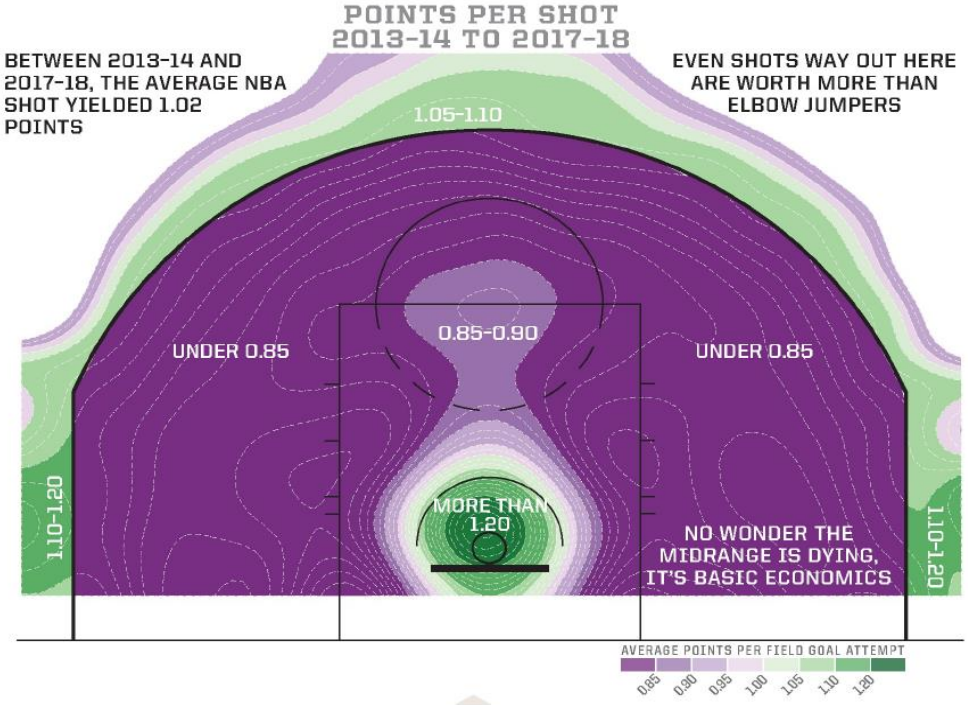
αθλητές που μπορούν να ασκήσουν επιρροή σε όλες τις πτυχές του. Αυτό για πολλά χρονιά είχε μεγάλο αρνητικό αντίκτυπο για τους πολύ ψηλούς αθλητές.

Για αυτό και τα τελευταία χρόνια παρατηρείται μια εξέλιξη στο παιχνίδι τους. Ο Nikola Jokic αν και αθλητικά υπολείπεται των συναθλητών του, εκμεταλλεύεται στο έπακρο την ικανότητα του να λαμβάνει γρήγορες αποφάσεις και να σουτάρει αποτελεσματικά. Σε αντίθεση με τους προκάτοχους του, χρησιμοποιεί το ύψος του (2.11 μ.), για να λειτουργήσει πιο πολύ ως ένας ογκώδης guard. Η τακτική του αυτή απέφερε, 2 σερί βραβεία MVP (nba.com/news/history-mvp-award-winners) και το πρωτάθλημα στην ομάδα του ως ο κορυφαίος παίκτης της και έχει επηρεάσει το παιχνίδι των περισσοτέρων ψηλών, που βλέπουν στο στυλ του, έναν πολύ αποτελεσματικό τρόπο για να ‘’επιβιώσουν’’ στο πολύ ανταγωνιστικό περιβάλλον.

2.5.1 Η επιρροή των δεδομένων στην εξέλιξη της καλαθοσφαίρισης

Η εισαγωγή των δεδομένων στην καλαθοσφαίριση έχει ως σκοπό την μεγιστοποίηση της απόδοσης μιας ομάδας. Ένας βασικός τρόπος χρήσης τους, είναι η ανάλυση των προσπαθειών που παίρνουν οι παίκτες σε όλα τα σημεία του γηπέδου. Για να γίνει κατανοητή η εξάλειψη του mid-range και η συμβολή των δεδομένων σε αυτή την τάση, παρατίθεται παρακάτω ένα γράφημα από τον Kirk Goldsberry, με τους πόντους ανά σουτ (*ο λόγος του αριθμού των πόντων από κάποια περιοχή προς των αριθμό προσπαθειών*) για την περίοδο 2013-2018. Παρατηρούμε ότι οι πόντοι ανά σουτ στην περιοχή του mid-range είναι εμφανώς λιγότεροι από τα σημεία πίσω από την γραμμή τριών πόντων.

Αυτή η διαπίστωση οδήγησε το μπάσκετ σε μια νέα εποχή και σε έναν νέο τρόπο προσέγγισης της νίκης. Πλέον οι ομάδες ψάχνανε κατά πλειοψηφία τρόπους για να σουτάρουν είτε τριποντά είτε να προσπαθούν να φτάσουν κοντά στο καλάθι, όπου οι πόντοι ανά σουτ βρίσκονται στην υψηλότερη τιμή τους.



Σχήμα 2.2. Πόντοι ανά σουτ από το 2013-14 μέχρι 2017-18 (Πηγή:Goldsberry (2019))

3ο ΚΕΦΑΛΑΙΟ
Βιβλιογραφική ανασκόπηση

3.1 Εισαγωγή

Παραπάνω έγινε αναφορά στην σημασία των δεδομένων για την εξέλιξη του αθλήματος. Σε αυτό το κεφάλαιο θα εντρυφήσουμε, στο πως τα δεδομένα εξέλιξαν το άθλημα, ποιοι ήταν αυτοί που βοήθησαν στην διάδοση των δεδομένων, καθώς και στις τεχνικές που χρησιμοποιήθηκαν για την εξαγωγή συμπερασμάτων.

3.2 Ο αντίκτυπος των δεδομένων

Το 2015 το ESPN και ο Kevin Pelton διεξήγαγαν μια τεράστια έρευνα αποτελούμενη από ερευνητές και ειδικούς για να αξιολογήσουν τα τμήματα analytics όλων των ομάδων καλαθοσφαίρισης, rugby, baseball, hockey στην Αμερική. Εστιάζοντας αποκλειστικά στην καλαθοσφαίριση, τα αποτελέσματα είναι ενδεικτικά της σημασίας των δεδομένων, για την επιτυχία μιας ομάδας. Στην έρευνα ο Pelton χώρισε τις ομάδες σε 5 κατηγορίες:

- All-In: Οι οργανισμοί που έχουν τα analytics ως πρωταρχικό εργαλείο για την λήψη αποφάσεων. Αυτές είναι οι:
    1. Dallas Mavericks
    2. Houston Rockets,
    3. Philadelphia 76ers
    4. San Antonio Spurs.
- Believers: Σε αυτή την κατηγορία εντάσσονται οι ομάδες που έχουν σε περίοπτη θέση τα analytics και που βασίζονται σε μεγάλο βαθμό σε αυτά. Οι ομάδες αυτής της κατηγορίας είναι οι:
    1. Atlanta Hawks
    2. Boston Celtics
    3. Cleveland Cavaliers
    4. Detroit Pistons
    5. Golden State Warriors
    6. Memphis Grizzlies
    7. Oklahoma City Thunder
    8. Portland Trail Blazers
- One Foot In: Ομάδες αυτής της κατηγορίας έχουν ως εργαλείο τα analytics, αλλά δεν βασίζονται σε μεγάλο βαθμό και χρησιμοποιούν και άλλους παράγοντες για την λήψη αποφάσεων.
    1. Miami Heat
    2. Milwaukee Bucks
    3. Indiana Pacers
    4. Charlotte Hornets
    5. Orlando Magic
    6. Phoenix Suns
    7. Sacramento Kings
    8. Toronto Raptors

9. Utah Jazz

- Skeptics: Εδώ κατατάσσονται οργανισμοί που έχουν μια πιο συντηρητική προσέγγιση των analytics και που θεωρούν πιο σημαντικές κλασικές τεχνικές.
    1. Chicago Bulls
    2. Denver Nuggets
    3. Los Angeles Clippers
    4. Minnesota Timberwolves
    5. New Orleans Pelicans
    6. Washington Wizards
- Nonbelievers: Τελευταία κατηγορία αποτελείται από οργανισμούς που έχουν κάνει ελάχιστα ή και καθόλου βήματα για την ανάπτυξη του τμήματος των analytics:
    1. Los Angeles Lakers
    2. Brooklyn Nets
    3. New York Knicks

Αυτό που αξίζει να σημειωθεί, από αυτή την έρευνα είναι η πορεία που είχαν εκείνη την σεζόν. Την σεζόν 2014/2015 στην κανονική περίοδο οι πρώτες 2 ομάδες κάθε περιφέρειας απαρτιζόταν από ομάδες που ανήκουν στις 2 πρώτες κατηγορίες (East:Hawks-Cavaliers, West:Rockets-Warriors).

Στα playoffs επίσης οι ίδιες ομάδες έφτασαν μέχρι τους τελικούς περιφέρειας. Από τις ομάδες των πρώτων 2 κατηγοριών στα playoff δεν πέρασαν οριακά μόνο οι Thunder. Από τις 3 ομάδες που δεν χρησιμοποιούσαν καθόλου τα δεδομένα, μόνο οι Nets πέρασαν 8^{οι} (η ελαχίστη απαιτούμενη κατάταξη για να μπει κάποιος) με ισοβαθμία, ενώ οι Knicks και οι Lakers είχαν τα 2 από τα 3 χειρότερα ποσοστά νικών σε όλο το πρωτάθλημα (Pelton, 2015).

3.3 Βιβλιογραφική ανασκόπηση

Η εξέλιξη των analytics στην καλαθοσφαίριση είναι μια διαδικασία που άρχισε αργότερα σε σχέση με αλλά αθλήματα. Στο baseball η μεγάλη άνοδος είχε έρθει από τις αρχές της δεκαετίας του ’80 με τον προαναφερθέν Bill James. (Papageorgiou, 2022).

Αρχικά γίνεται ειδική αναφορά σε 2 πήγες που έχουν επηρεάσει τις περισσότερες δουλείες που θα επεκταθούμε παρακάτω. Η επακόλουθη δομή που θα ακολουθηθεί θα είναι παρόμοια με αυτή που πρότειναν οι López et al.(2013). Πιο αναλυτικά τα 2 πεδία που έχουν αναλυθεί περισσότερο (Outcome Prediction, Team/Player Performance), θα παρουσιαστούν ξεχωριστά, ενώ μια 3η κατηγορία θα αποτελείται από έρευνες που έχουν αναλυθεί λιγότερο, όπως το shot-chart analysis. Σημειώνεται επίσης ότι η σειρά που θα παρουσιάζονται θα είναι με βάση την χρονολογία έκδοσης.

Για την καλαθοσφαίριση η αφετηρία έγινε με τον Dean Oliver και το βιβλίο του ‘’Basketball on Paper’’(2004). Το ‘’Basketball on Paper’’ ήταν το πρώτο βιβλίο που γράφτηκε για προπονητές και εισήγαγε νέους τρόπους να αντιλαμβάνεται κάνεις το άθλημα. Η ανάλυση του παιχνιδιού με βάση τις κατοχές, είναι μια τεχνική που χρησιμοποιείται μέχρι σήμερα και είναι ο πρώτος που την σύστησε επισημά, αν και όπως γραφεί ο ίδιος υπήρχαν προπονητές που είχαν χρησιμοποιήσει ήδη την ίδια προσέγγιση. Η πιο γνωστή συνεισφορά του θεωρείται η εισαγωγή των ‘’4 factors’’. Οι 4 παράγοντες που κατέληξε, πως καθορίζουν μια επιτυχημένη ομάδα είναι το αποτελεσματικό ποσοστό εντός παιδίας (eFG%), το ποσοστό λαθών (TOV%), ποσοστό επιθετικών ριμπάουντ (ORB%) και ο ρυθμός των επιθέσεων που καταλήγουν σε ελεύθερες βολές (FTR). Σκοπός του Oliver, ήταν η εισαγωγή στατιστικών μεθόδων, για την αξιολόγηση ομάδων και την συνεισφορά παικτών (Oliver, 2011).

Με γνώμονα την ανάλυση του παιχνιδιού μέσω των κάτοχων, οι Kubatko et al. (2007) έκαναν μια αναλυτική καταγραφή όλων των δεικτών, που χρειάζεται για την αξιολόγηση ομάδων και παικτών. Ουσιαστικά η πρόθεση τους είναι να χαράξουν ένα θεμέλιο για μελλοντικές έρευνες. Για τα κυριότερα κομμάτια της έρευνας τους έχουν εντάξει στατιστικά που βασίζονται στους ‘’4 Factors’’ του Oliver, κάνουν εκτενή αναφορά στα plus/minus statistics (*σύγκριση της προσφοράς ενός παίκτη όταν βρίσκεται στο γήπεδο και όταν πηγαίνει στον πάγκο*) και στα individual possessions (*μέτρο για το πόσο έντονα οι παίκτες χρησιμοποιούν τις κατοχές, μέσω των προσπαθειών στο καλάθι, των assist, των λαθών και των επιθετικών rebound*). Ακόμη χρησιμοποιούν συντελεστές για να μετρήσουν την σημασία των στατιστικών των παικτών και κάνουν μια εισαγωγή στην Bell Curve Method, που σχετίζεται με την κανονική κατανομή και τους πόντους ανά παιχνίδι. Η ερευνά έχει επηρεαστεί από παλαιοτέρα άρθρα του Oliver(2004).

3.3.1 Outcome Prediction

Στο προηγούμενο κεφάλαιο ειπώθηκε ότι η άνοδος της μηχανικής μάθησης οδήγησε σε ανάπτυξη όλων των πτυχών της καλαθοσφαίρισης. Μια από αυτές είναι η πρόβλεψη αποτελεσμάτων. Οι Sarlis και Tjortjis (2020) χρησιμοποίησαν τους διαθέσιμους δείκτες του μπάσκετ και μέσω διαφορετικών μοντέλων ML πρόβλεψαν τον πολυτιμότερο (MVP) και καλύτερο (DPOY) αμυντικό παίκτη της χρονιάς. Συγκεκριμένα χρησιμοποίησαν διαφορετικά στατιστικά, που θα εντόπιζαν τους πιο αποτελεσματικούς παίκτες. Αυτά ήταν: Plus/Minus, Adjusted Plus Minus, Real Plus Minus, Player Impact Plus Minus, Player Impact Estimate, CARMELO, Expected Possession Value, Wins Above Replacement, Performance Index Rating, Game Score, Net Rating, Pythagorean Win Percentage, Player Efficiency Rating (PER), Value over Replacement Player, Win Shares και Tendex. Μέσω του δείκτη Aggregated Performance Indicator (API), προέβλεψαν σωστά τον νικητή του βραβείου MVP για 5 συνεχόμενα χρόνια. Επιπλέον αναφέρουν ότι οι πιο αποτελεσματικοί δείκτες είναι τα Adjusted Plus Minus (APM) και Plus/ Minus.

Οι Loeffelholz et al. (2009) χρησιμοποίησαν νευρωνικά δίκτυα για να προβλέψουν τα αποτελέσματα αγώνων. Ως μέτρο σύγκρισης χρησιμοποίησαν τις γνώμες ειδικών του αθλήματος, πάνω στο ποια ομάδα είναι φαβορί. Στην έρευνα χρησιμοποιήθηκαν 4 διαφορετικά νευρωνικά δίκτυα:

1. Feed Forward Neural Networks
2. Radial Basis Functions
3. Probabilistic Neural Networks
4. Generalized Neural Networks

Έπειτα χρησιμοποίησαν δυο διαφορετικές τεχνικές fusion (Bayesian Belief Network-Conditional Probability Distribution), για να αποκτήσουν μεγαλύτερη ακρίβεια. Τα αποτελέσματα τους ήταν αρκετά ικανοποιητικά, με όλα τα νευρωνικά δίκτυα να είναι ακριβέστερα από τους ειδικούς και μάλιστα να φτάνουν σε ακρίβεια μέχρι 74.3% ακρίβεια με τα Feed Forward Neural Networks.

Ο Sill (2010) πρότεινε μια βελτίωση πάνω στο APM με την χρήση Ridge παλινδρόμησης. Σύμφωνα με τον Sill η χρήση αυτής της Μπευζιανής τεχνικής μπορεί να ξεπεράσει τα θέματα πολυσυγγραμικότητας και υπερπροσαρμογής (overfitting) που δημιουργούνται με το απλό APM.

Οι Štrumbelj & Vračar (2012) χρησιμοποίησαν ένα possession based Markov μοντέλο, για να προσομοιώσουν τα πιθανά αποτελέσματα μετά από διάφορες δράσεις κατά την διάρκεια του

παιχνιδιού. Το μοντέλο πετυχαίνει ικανοποιητικές προβλέψεις. Επιπλέον σε προβλέψεις και με άλλα μοντέλα προέκυψε ότι οι αποδόσεις των bookmakers είναι οι πιο ακριβείς.

Οι Shi et al. (2013) χρησιμοποίησαν 4 διαφορετικά μοντέλα μηχανικής μάθησης (Naïve-Bayes, Rule learners, Artificial Neural Networks, Decision Trees, Ensemble Learners) για την πρόβλεψη αποτελεσμάτων. Ένα ενδιαφέρον συμπέρασμα της έρευνας του, ήταν ότι η προβλεπτική ικανότητα βελτιώνεται με ορθότερη επιλογή χαρακτηριστικών και όχι πολυπλοκότερων μοντέλων.

Οι Chen et al. (2016) εφάρμοσαν την αρχή Maximum Entropy, για να κατασκευάσουν το μοντέλο NBA Maximum Entropy (NBAME), το οποίο προβλέπει τα αποτελέσματα αγώνων playoff στο NBA. Το μοντέλο αντιμετωπίζει ζητήματα γνωστών αλγορίθμων, όπως την υπόθεση ανεξαρτησίας για το Naïve Bayes και ταυτόχρονα παρουσιάζει πολύ υποσχόμενα αποτελέσματα. Για την κατασκευή χρησιμοποιήθηκε ο μέσος από τα τελευταία 6 παιχνίδια της εκάστοτε ομάδας και επιλέχθηκαν 14 ξεχωριστά χαρακτηριστικά. Σε σύγκριση που έγινε με άλλους ML αλγορίθμους, το NBAME είχε στις περισσότερες σεζόν την υψηλότερη προβλεπτική ικανότητα, φτάνοντας ακόμα μέχρι 74.4%. Ένα ενδιαφέρον συμπέρασμα της έρευνας είναι ότι αν αυξήσουν το threshold για την πιθανότητα νίκης, προβλέπουν λιγότερα παιχνίδια, αλλά με μεγαλύτερη ακρίβεια.

Οι Shah και Romijnders (2016) με την χρήση Recurrent Neural Networks (RNN), προέβλεψαν τα αποτελέσματα των προσπαθειών πίσω από την γραμμή του τρίποντου. Πιο συγκεκριμένα χρησιμοποιούν μια παραλλαγή των RNN, το Long Short Term Memory (LSTM), για sequential data και Gradient Boosted Machines για non sequential data. Για την σύγκριση τους χρησιμοποιήθηκε η Area Under Curve (AUC), όπου τα RNN παρουσίασαν πολύ καλύτερα αποτελέσματα.

Ο Zhao et al. (2017) με την χρήση νευρωνικών δικτύων πέτυχε ικανοποιητική πρόβλεψη για την πρόβλεψη της πορείας που ακολουθούν τα σουτ τριών πόντων. Για την έρευνα χρησιμοποίησε την προσέγγιση bidirectional long short-term memory (BLSTM) και mixture density network και χωρικά δεδομένα.

Οι Horvat et al. (2018) εφάρμοσαν τον αλγόριθμο KNN, για πρόβλεψη αγώνων της Euroleague. Πιο αναλυτικά κατασκεύασαν 2 παραλλαγές δεδομένων, μια που θα ομαδοποιεί τα αμυντικά και επιθετικά στοιχεία (variant DefenseOfense) των ομάδων και άλλη μια που θα ομαδοποιεί το λογικά στοιχεία του παιχνιδιού (variant Components). Στην συνέχεια εφάρμοσαν feature selection με κριτήριο το μέτρο information gain. Τα αποτελέσματα που αποκόμισε ήταν ιδιαίτερα ικανοποιητικά, με προβλέψεις να ξεπερνούν το 80%. Η παραλλαγή DefenseOfense αποδείχτηκε καλύτερη, ενώ επίσης από την έρευνα συμπέραναν ότι ο συντελεστής k για το KNN δεν παίζει σημαντικό ρόλο, σε αντίθεση με το feature selection.

Ο Migliorati (2020) χρησιμοποίησε τις τεχνικές CART και Random Forest, για να προβλέψει αγώνες των Golden State Warriors, για να εξάγει παράγοντες που οδηγούν στην νίκη. Η προβλεπτική ικανότητα του μοντέλου ξεπέρασε το 70%, ενώ αξίζει να σημειωθεί ότι θέλοντας να ελέγξει τους 4 factors του Oliver(2004), κατέληξε ότι τα αμυντικά rebound είναι σημαντικότερα από τα λάθη για την έκβαση του αγώνα.

Οι Thakur S. & Karthik R. (2022) με κριτήριο το shot selection του Kobe Bryant στα τελευταία λεπτά ενός αγώνα, χρησιμοποίησαν μηχανική μάθηση, για να βρουν ποιος αλγόριθμος προβλέπει πιο αποτελεσματικά τα αποτελέσματα των προσπαθειών του. Μέσω λογιστικής παλινδρόμησης πέτυχαν ακρίβεια 62.49%, ενώ ακόμη κατέληξαν ότι η προσθήκη παραπάνω training data βελτιώνουν την απόδοση του μοντέλου.

Οι Alonso & Babac (2022) χρησιμοποίησαν 3 διαφορετικούς αλγορίθμους μηχανικής μάθησης (K-Nearest Neighbors, Decision trees, Naive Bayes), για την πρόβλεψη

αποτελεσμάτων στο NBA. Ο καλύτερος αλγόριθμος ήταν ο Naïve Bayes classifier με ακρίβεια λίγο παραπάνω από 71%.

Οι Osken et al. (2022) ομαδοποιούν τους παίκτες ανάλογα με το στυλ παιχνιδιού τους και με κριτήριο αυτό επιχειρούν να κάνουν πρόβλεψη αποτελεσμάτων. Με την χρήση συσταδοποίησης και συγκεκριμένα της μεθόδου k-means, χωρίζουν τους παίκτες σε 25 συστάδες κα στην συνέχεια με την χρήση νευρωνικών δικτυών επιτυγχάνουν πολύ ικανοποιητική πρόβλεψη γύρω στο 76%. Ένα ενδιαφέρον συμπέρασμα της έρευνας είναι, ότι η προβλεπτική ικανότητα είναι ακριβέστερη για ομάδες με υψηλοτέρα ποσοστά νικών σε εντός και εκτός έδρας παιχνίδια.

3.3.2 Team/Player Performance

Μια από τις συχνότερες χρήσεις των basketball analytics είναι η εκτίμηση της απόδοσης μιας ομάδας ή ενός παίκτη. Σε αυτό το κομμάτι θα παρουσιαστούν προηγούμενες δουλείες που επιχείρησαν να βρουν τρόπους βελτίωσης της λειτουργίας της ομάδας στο γήπεδο, όπως η ερευνά που έκανε ο Skinner (2011), που έδειξε ότι αν μια ομάδα έχει πλήρη γνώση των ατομικών ικανοτήτων των παικτών, μπορεί να ‘’ρυθμίσει’’ τον τρόπο που θα δομηθεί η επίθεση για την μέγιστη αποτελεσματικότητα.

Οι López et al. (2013) θέλησαν να ερευνήσουν τον αντίκτυπο που είχε η προσθήκη του Pau Gasol στο roster των Los Angeles Lakers. Συγκεκριμένα χρησιμοποίησαν μια χωρική τεχνική clustering για να αναλύσουν τα δεδομένα. Η τεχνική clustering ήταν το Kulldorf Test και μέσω αυτής φάνηκε πως η έλευση του Gasol επηρέασε τις θέσεις που σουτάρανε οι υπόλοιποι παίκτες. Ακόμη για την σύγκριση των shooting maps χρησιμοποιείται το V-test.

Οι Maymin et al. (2013) χρησιμοποίησαν ως βάση των δείκτη APM του (Rosenbaum, 2004), και κατασκεύασαν το Skills Plus Minus (SPM), με την χρήση λογιστικής παλινδρόμησης και συνάρτηση σύνδεσης probit. Σκοπός τους ήταν η ανίχνευση των καλυτέρων συνθέσεων μέσω 3 κατηγοριών: Ball Handling, Scoring, Rebounding. Βασική διαφορά των δυο δεικτών είναι, ότι το SPM λαμβάνει υπόψιν το πως αρχίζει η κάθε κατοχή, κάτι που μπορεί να έχει μεγάλη διαφορά στην παραγωγή πόντων. Ένα ακόμα προτέρημα, είναι ότι μπορεί να προβλέψει ποια χαρακτηριστικά μπορούν να επηρεάσουν θετικά την ομάδα και συνεπώς ποιοι παίκτες ταιριάζουν καλυτέρα μεταξύ τους. Για την τελική αξιολόγηση παικτών χρησιμοποιεί το Points over Replacement Player (PORP). Επιπλέον το SPM αξιοποιήθηκε, για την εύρεση ανταλλαγών παικτών που θα ήταν ωφέλιμες και για τις δυο ομάδες.

Οι Skinner & Guy (2015) κατασκεύασαν ένα network based μοντέλο, με στόχο να ποσοτικοποιήσουν την ικανότητα κάθε παίκτη στην επίθεση. Το μοντέλο τους έδειξε ικανοποιητική προβλεπτική ικανότητα ακόμα και όταν οι παίκτες βρισκόντουσαν σε διαφορετικές συνθέσεις.

Οι Zhang L. et al. (2016) χρησιμοποίησαν συσταδοποίηση K means, έτσι ώστε να ταξινομήσουν τους guards στο NBA. Από την ανάλυση τους προέκυψαν 6 κατηγορίες διαφορετικών guard. Σκοπός της έρευνας τους ήταν να μια αντικειμενική απεικόνιση των παικτών με βάση τα στατιστικά τους.

Πάνω στην αξιολόγηση συνθέσεων παικτών δούλεψε ο Goldberg (2017). Συγκεκριμένα ήθελε να κατηγοριοποιήσει παίκτες ανάλογα με το σύγχρονο play-style και με βάση αυτό να ερευνήσει ποιες είναι οι αποτελεσματικότερες πεντάδες παικτών. Για την έρευνα του χρησιμοποίησε κατά κύριο λόγο, γραμμική παλινδρόμηση. Για να μπορέσει να έχει αξιόπιστο clustering, χρησιμοποίησε δείκτες που απευθύνονται σε επίθεση, άμυνα και ριμπάουντ, ενώ επιπλέον ανάθεσε επιπλέον βάρη στα χαρακτηριστικά της άμυνας και του ριμπάουντ, έτσι ώστε να μην επισκιαστούν από την πληθώρα των χαρακτηριστικών της επίθεσης. Η μέθοδος

clustering που χρησιμοποίησε ήταν η k-means και με το silhouette score κατέληξε σε 8 διαφορετικά clusters, που καθόριζαν έναν διαφορετικό τρόπο παιχνιδιού. Στην συνέχεια χρησιμοποίησε ως κριτήριο κατάταξης των παικτών το Points Above Replacement Rating (*Η διαφορά πόντων ανάμεσα στην πεντάδα που βρίσκεται ο υπό μελέτη παίκτης ενάντια σε μια πεντάδα αναπληρωματικών*). Παρόμοια τακτική ακολουθήσε και για την αξιολόγηση των καλύτερων συνθέσεων (*Διαφορά πόντων μια σύνθεσης ενάντια σε μια σύνθεση αναπληρωματικών*).Τα αποτελέσματα του ανταποκρινόταν ικανοποιητικά στην πραγματικότητα. Στα συμπεράσματα του, παρατήρησε ότι οι περισσότερες αποτελεσματικές συνθέσεις, είχαν στην διάθεση τους τουλάχιστον έναν καλό σουτέρ 3-ποντων. Κατέληξε επίσης, ότι ο δείκτης Regularized Adjusted Plus Minus (RAPM), πάρα την χρησιμότητα, δεν μπορεί κανείς να βασιστεί αποκλειστικά σε αυτόν για να χτίσει την ιδανική σύνθεση.

Οι Zhang et al. (2017) ερεύνησαν την διακύμανση της απόδοσης των παικτών με μεταβλητές: τον συντελεστή μεταβλητότητας, τα ματς που κερδίζουν και χάνουν, δυνατές και αδύναμες ομάδες, έντονος ανταγωνισμός και αδύναμος ανταγωνισμός.

Οι Bianchi et al. (2017) επιχείρησαν με την βοήθεια της μηχανικής μάθησης να ερμηνεύσουν την εξέλιξη των θέσεων στην καλαθοσφαίριση, πέρα των πέντε κλασικών(PG,SG,SF,PF,C). Ορμώμενοι από μια έρευνα που κατέληξε σε 13 διαφορετικές θέσεις μέσα από Topological Data Analysis (TDA), χρησιμοποίησαν clustering και νευρωνικά δίκτυα (Neural Networks-NN), για να επιβεβαιώσουν ότι οι 5 κλασικές θέσεις δεν ανταποκρίνονται στους σημερινούς αθλητές. Για να καταλήξουν σε έναν τελικό αριθμό clusters χρησιμοποίησαν 4 διαφορετικά κριτήρια (Partition Coefficient, the Partition Entropy, the Silhouette, the Fuzzy Silhouette), με το 3 να είναι ο τελικός αριθμός. Έπειτα έγινε μείωση διαστάσεων μέσω του Multidimensional Scaling (MDS). Το συμπέρασμα που έβγαλε από την έρευνα κατέληξε σε 5 διαφορετικές θέσεις(διαχωρίστηκαν τα πρώτα 2 κολάστρες σε δυο και δυο), που ξεφεύγουν από τα κλασικά πλαίσια και ερμηνεύονται ως:

1. Αποτελεσματικοί παίκτες σε πολλές κατηγορίες (All Around All Star-AAS).
2. Παίκτες με έφεση στο σκοράρισμα και στο passing game (Scoring Backcourt-SB).
3. Παίκτες που συλλέγουν πολλά ριμπάουντ και σκοράρουν ικανοποιητικά(Scoring Rebounder-SC).
4. Παίκτες με υψηλά νούμερα σε ριμπάουντ, κλεψίματα, κοψίματα (Paint Protector-PP)
5. Παίκτες με αποτελεσματικότητά σε μια κατηγορία. Απευθύνεται σε αναπληρωματικούς (Role Player-RP)

Οι Zhang et al. (2018) χρησιμοποίησαν μια παρόμοια προσέγγιση, δίνοντας έμφαση και σε ανθρωποκεντρικούς παράγοντες όπως η εμπειρία, το ύψος και το βάρος. Από την cluster analysis κατέληξαν σε 5 διαφορετικά clusters, που ποίκιλαν σε ύψος, βάρος και εμπειρία.

Οι Chen et al. (2018) χρησιμοποίησαν οπτικά δεδομένα για να προσομοιώσουν την συμπεριφορά της άμυνας, σε συνάρτηση με την κίνηση της μπάλας και το επιθετικό παιχνίδι. Χρησιμοποίησαν το μοντέλο Generative Adversarial Network (GAN), για να παράξουν τις κινήσεις των αμυνόμενων. Μερικά ενδιαφέροντα συμπεράσματα της έρευνας τους, είναι ότι οι αμυνόμενοι στην πραγματικότητα τρέχουν πιο αργά από το αναμενόμενο και επίσης η χρήση ενός επιπλέον όρου που θα λαμβάνει υπόψιν τις κατοχές που οι αμυντικοί αφήνουν ελευθέρους τους παίκτες που μαρκάρουν.

Οι Zhang S. et el. (2019) χρησιμοποίησαν την μέθοδο non-Metric Multidimensional Scaling (nMDS), για να ερευνήσουν τις διάφορες στους τρόπους παιχνιδιού των ομάδων κατά την διάρκεια μιας ολόκληρης σεζόν. Από την έρευνα τους συμπέραναν, ότι οι ομάδες γίνονται πιο αποτελεσματικές, όσο πλησιάζουν στο τέλος της σεζόν.

Στο τελευταίο κομμάτι της βιβλιογραφικής ανασκόπησης, θα αναφερθούν έρευνες σε πεδία της καλαθοσφαίρισής που έχουν αναλυθεί σε μικρότερο βαθμό. Συγκεκριμένα οι δυο κατηγορίες αφορούν την ανάλυση της αποτελεσματικότητας των σουτ, με την χρήση tracking data και την αξιολόγηση των στρατηγικών που χρησιμοποιούν οι ομάδες, με την βοήθεια χωρικής ανάλυσης.

3.3.3 Shot Chart- Spatial Analysis

Για την ανάλυση της απόδοσης στο γήπεδο, μια χρήσιμη τακτική είναι η ανάλυση του Shot Chart (*Απεικόνιση των σημείων που σουτάρει ένας παίκτης μέσα στο γήπεδο*) ενός παίκτη ή μιας ομάδας. Οι Reich et al. (2006) ανέλυσαν με ιεραρχικά χωρικά μοντέλα και μεθόδους Markov Chain Monte Carlo τα σουτ που έπαιρνε ο Sam Cassell. Αρχικά χρησιμοποίησε λογιστική παλινδρόμηση για να βρει τις εκ των υστέρων διαμέσους και ποιες μεταβλητές ήταν σημαντικές. Έπειτα χρησιμοποιεί την κατανομή Conditionally Autoregressive (CAR) και την two neighbor relation CAR, έτσι ώστε οι συντελεστές να είναι παρόμοιοι σε κοντινά σημεία του γηπέδου και συγκρίνει τα μοντέλα μέσω του κριτηρίου Deviance Information (DIC). Αποτέλεσμα της ανάλυσης ήταν μια αποτελεσματική αντιμετώπιση του υπό μελέτη παίκτη σε μελλοντικά παιχνίδια.

Οι Miller et al. (2014) μοντελοποίησαν τα δεδομένα με προσπάθειες σουτ ως μια point process, για να δημιουργήσουν μια αναπαράσταση με χαμηλές διαστάσεις για τα διάφορα είδη σουτέρ στο NBA. Για να το κάνουν αυτό χρησιμοποιούν ως μέθοδο μείωσης διαστάσεων την non-negative matrix factorization (NMF).

Οι Cervone et al. (2014) θέλησαν να εκμεταλλευτούν χωρικά δεδομένα, από την εισαγωγή ειδικών καμερών στο NBA, για να κατασκευάσουν έναν δείκτη, που θα μετράει τους αναμενομένους πόντους κάθε κατοχής, με κριτήριο τις επιλογές που γίνονται από τον χειριστή της μπάλας. Η ονομασία ήταν Expected Possession Value (EPV). Μέσω του EPV, δίνεται η δυνατότητα να υπολογιστεί και η αξία των προσπαθειών στο σουτ σε σχέση με την επιλογή πάσας (Shot Satisfaction).

Οι Csapo & Raab (2014) έκαναν έρευνα για να ελέγξουν τον ισχυρισμό του ‘’hot hand’’(*η τάση ενός παίκτη να είναι πιο εύστοχος αν έχει ευστοχήσει σε διαδοχικές προσπάθειες*) και πως οι άμυνες συμπεριφέρονται απέναντι σε αυτές τις καταστάσεις. Το συμπέρασμα τους ήταν ότι οι παίκτες που έχουν το ‘’hot hand’’, τείνουν να είναι πιο άστοχοι από παίκτες που έχουν διαδοχικές άστοχες προσπάθειες, ενώ λαμβάνουν και σουτ υπό δύσκολες προϋποθέσεις πιο συχνά. Επίσης διαπιστώθηκε ότι οι αμυντικοί συχνότερα μαρκάρουν πιο έντονα αυτούς τους παίκτες.

Οι Franks et al. (2015) χρησιμοποιούν χωρική ανάλυση, για να μπορέσουν να ανιχνεύσουν πως οι αμυντικοί επηρεάζουν την συχνότητα και την αποτελεσματικότητα των σουτ των επιτιθέμενων. Με χρήση του NMF βρίσκουν τις κύριες περιοχές των σουτ. Για την εύρεση των matchups χρησιμοποιούνται ο αλγόριθμος Expectation–Maximization και τα generalized least squares. Τα συμπεράσματα που εξάγουν δίνουν μια πιο βαθιά ματιά στην ατομική αμυντική συμπεριφορά.

Οι Erčulj & Štrumbelj (2015) θέλησαν να ερευνήσουν τα σουτ που επιλέγουν οι παίκτες ανάλογα με το πρωτάθλημα που βρίσκονται. Τα πρωταθλήματα που ερευνήθηκαν ήταν του NBA, της Euroleague, το πρωτάθλημα Σλοβενίας και η U16 και U14 ομάδες της Σλοβενίας. Στην έρευνα τους μελετήθηκε η σχετική συχνότητα διαφόρων τύπων shooting. Το μοντέλο που χρησιμοποιήθηκε ήταν ένα Μπεϋζιανό ιεραρχικό πολυώνυμο μοντέλο λογιστικής παλινδρόμησης. Κατασκευάστηκαν επίσης 3 μοντέλα και για τον έλεγχο της προβλεπτικής τους ικανότητας χρησιμοποιήθηκε το κριτήριο Widely Applicable Information Criterion

(WAIC). Στα αποτελέσματα φάνηκε να μην παρατηρείται στατιστικά σημαντική διαφορά για NBA, Euroleague και πρωταθλήματος Σλοβενίας. Μόνη διαφορά ήταν η μεγαλύτερη συχνότητα καρφωμάτων στο NBA και η μικρότερη χρήση hook shots (σουτ με το ένα χέρι, οπού το χέρι εκτείνεται ψηλά με μια γρήγορη κίνηση και το βλέμμα στο καλάθι) σε σχέση με τα ευρωπαϊκά πρωταθλήματα.

Οι Jiao et al. (2021) προτείναν μια Bayesian marked spatial point process για να μοντελοποιήσουν τις περιοχές του σουτ και τα αποτελέσματα τους. Για την επίτευξη του στόχου τους χρησιμοποιούν μια μη ομογενής Poisson point process , για την μοντελοποίηση των χωρικών μοτίβων και λογιστική παλινδρόμηση. Για την ερευνά του χρησιμοποιήθηκε και εδώ η NMF των Miller et al. (2014).Το αποτέλεσμα της έρευνας τους έδειξε ότι τα σημεία που σουτάρουν πιο συχνά οι παίκτες έχουν θετική σχέση με την αποτελεσματικότητα τους.

3.3.4 Ανάλυση στρατηγικών μέσω tracking data

Η χωρική ανάλυση είναι ένας τομέας που έχει ερευνηθεί λιγότερο, όμως με καινοτομίες όπως η εισαγωγή ειδικών καμερών που εστιάζουν στην κίνηση κάθε παίκτη, μπορούν να προκύψουν πιο διαισθητικά συμπεράσματα. Μια τέτοια ερευνά έκαναν οι Lucey et al. (2014) θέλοντας να ερευνήσουν τους αμυντικούς και επιθετικούς σχηματισμούς των ομάδων, με την βοήθεια tracking data. Χρησιμοποίησαν ακόμη t-test, για να μπορέσουν να δουν ποιοι είναι σημαντικοί παράγοντες για τις συνθέσεις. Επιθετικά συμπέραναν ότι οι ομάδες πρέπει να πασάρουν περισσότερο και να ντριμπλάρουν λιγότερο. Αμυντικά η μέση και η μέγιστη απόσταση από τον επιτιθέμενο, η απόσταση που καλύπτουν σε 3 δευτερόλεπτα και οι αλλαγές αμυντικών ρόλων είναι στατιστικά σημαντικές.

Οι Miller & Bornn (2017) χρησιμοποιώντας player-tracking data, κατασκεύασαν ένα μοντέλο που αναγνωρίζει τις επιθετικές στρατηγικές, με γνώμονα τις κινήσεις των παικτών στον χώρο. Για την κατασκευή του μοντέλου καταγράφηκαν οι στιγμές που οι παίκτες χαμηλώνουν ταχύτητα στιγμιαία και στην συνέχεια χρησιμοποίησαν έναν probabilistic clustering αλγόριθμο, για να ταξινομήσουν ποια δράση αντιστοιχεί στις στιγμές που καταγράφηκαν. Έπειτα χρησιμοποιούν το μοντέλο Latent Dirichlet Allocation(LDA), για να εντοπιστεί ποια στρατηγική (ένα σύνολο από δράσεις) χρησιμοποιείται σε κάθε κατοχή.

Οι Tian et al. (2019), είχαν ως σκοπό να ταξινομήσουν αμυντικές στρατηγικές, με την βοήθεια χωρικών δεδομένων. Για την ταξινόμηση χρησιμοποίησαν διάφορους classifiers και μετά από έρευνα κατέληξαν ότι καλύτερος classifier είναι η τεχνική Support Vector Machine(SVM), με ακρίβεια 68.9%.

4ο ΚΕΦΑΛΑΙΟ

Παρουσίαση Τεχνικών Μηχανικής Μάθησης

4.1 Εισαγωγή

Σε αυτό το κεφάλαιο γίνεται μια συνοπτική παρουσίαση των τεχνικών που θα χρησιμοποιηθούν στην συνέχεια.

4.2 Κατηγορίες Μηχανικής Μάθησης

Η εκθετική αύξηση των διαθέσιμων δεδομένων έχει υπάρξει κινητήρια δύναμη για την ανάπτυξη της μηχανικής μάθησης. Ως αποτέλεσμα η μηχανική μάθηση, έχει ένα μεγάλο εύρος κατηγοριών, για την διαχείριση κάθε προβλήματος. Οι βασικές κατηγορίες είναι οι εξής:

- Εποπτευομένη μάθηση (Supervised learning)
- Μη εποπτευομένη μάθηση (Unsupervised learning)
- Ενισχυτική μάθηση (Reinforcement learning)

Οι δυο πρώτες κατηγορίες θεωρούνται οι θεμελιώδεις τύποι μηχανικής μάθησης, ενώ και η ενισχυτική μάθηση θεωρείται ιδιαίτερα σημαντική. Αξίζει να αναφερθεί επίσης ότι υπάρχουν πολλές ακόμα κατηγορίες που είναι είτε συνδυασμοί ή παραλλαγές των βασικών τύπων (Semi-supervised, Ensemble learning), είτε ξεχωριστοί τύποι μικρότερης σημασίας (Batch, Online Learning).

4.2.1 Εποπτευομένη μάθηση (Supervised learning)

Η εποπτευομένη μάθηση είναι ένας τύπος μηχανικής μάθησης, με κύριο χαρακτηριστικό, την χρήση επισημασμένων δεδομένων εκπαίδευσης (labeled training data), για να βρεθεί η σχέση ανάμεσα στα δεδομένα εισόδου- εξόδου (input- output data). Μέσω αυτής της σχέσης, δημιουργείται ένα τεχνητό σύστημα, που μπορεί να προβλέψει τα δεδομένα εξόδου, όταν εισάγουμε καινούργια δεδομένα εισόδου (Liu, 2011).

Η επιβλεπομένη μάθηση, βρίσκει εφαρμογή, κατά κύριο λόγο, σε προβλήματα ταξινόμησης (classification) και παλινδρόμησης (regression). Το βασικό πλεονέκτημα της είναι η υψηλή ακρίβεια, που μπορούν να παρουσιάσουν τέτοια μοντέλα, καθώς και η ερμηνεία των αποτελεσμάτων. Στα μειονεκτήματα είναι η δυσκολία συλλογής τέτοιων δεδομένων, διότι η συλλογή τους είναι μια χρονοβόρα διαδικασία, ενώ τίθεται και το ζήτημα της ποιότητας των δεδομένων.

Για το θέμα που εξετάζουμε η επιβλεπομένη μάθηση, μπορεί να βοηθήσει στην πρόβλεψη αποτελεσμάτων.

4.2.2 Μη εποπτευομένη μάθηση (Unsupervised learning)

Στην μη εποπτευομένη μάθηση, ο σκοπός είναι η εύρεση μοτίβων ή σχέσεων στα δεδομένα, χωρίς την ύπαρξη κάποιας εκ των προτέρων γνώσης, όπως στην εποπτευομένη μάθηση με τα επισημασμένα δεδομένα. Τα δυο γνωστότερα παραδείγματα είναι της συσταδοποίησης (clustering) και της μείωσης διαστάσεων.

Στο πλαίσιο της διπλωματικής, η μη εποπτευομένη μάθηση, μπορεί να χρησιμοποιηθεί για τον τρόπο εξέλιξης του παιχνιδιού, μέσω συσταδοποίησης, όπως είχαν κάνει οι Bianchi et al.(2017).

4.2.3 Ενισχυτική μάθηση (Reinforcement learning)

Η ενισχυτική μάθηση είναι άλλος ένας βασικός τομέας της μηχανικής μάθησης. Η κύρια ιδέα βασίζεται σε ένα σύστημα επιβράβευσης και τιμωρίας. Πιο συγκεκριμένα ένας λήπτης αποφάσεων (agent), αλληλοεπιδρά με το περιβάλλον του προβλήματος και λαμβάνει αποφάσεις που επηρεάζουν την επιβράβευση και την τιμωρία του λήπτη. Στόχος είναι ο λήπτης αποφάσεων να αναπτύξει έναν τρόπο συμπεριφοράς (policy) που θα μεγιστοποιήσει όχι την άμεση, αλλά την αθροιστική επιβράβευση. Η διαδικασία μάθησης του λήπτη περιλαμβάνει μια ισορροπία αναμεσά στην αναζήτηση καινούργιων δράσεων και στην επιλογή δράσεων που φέρνουν μεγάλη επιβράβευση, με κριτήριο τις υπάρχουσες γνώσεις.

4.3 Μέθοδοι ταξινόμησης (Classification methods)

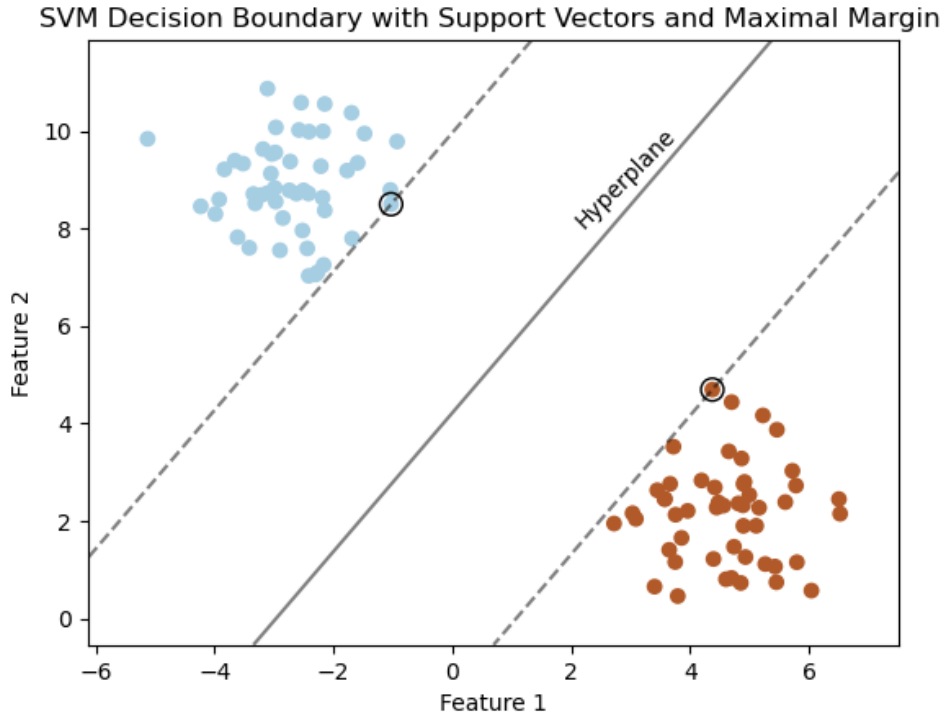
Η ταξινόμηση είναι μια θεμελιώδης έννοια στον χώρο της μηχανικής μάθησης. Στόχος των αλγορίθμων ταξινόμησης είναι η χρήση επισημασμένων δεδομένων (labeled data), για να μπορέσουν να ταξινομηθούν μη επισημασμένα δεδομένα (unlabeled data) σε τάξεις που γνωρίζουμε εκ των προτέρων.

Οι αλγόριθμοι που θα αναλυθούν παρακάτω:

- Μηχανές Διανυσματικής Υποστήριξης (Support Vector Machines)
- Κ Κοντινότεροι Γείτονες (K Nearest Neighbours)
- Λογιστική παλινδρόμηση (Logistic regression)
- Γραμμική διαχωριστική αναλυση (Linear discriminant analysis)
- Ο ταξινομητής Naïve Bayes (Naïve Bayes classifier)
- Δέντρα απόφασης (Decision trees)
- Τυχαία δάση (Random forests)

4.3.1 Μηχανές Διανυσματικής Υποστήριξης (Support Vector Machines)

Ο αλγόριθμος Support Vector Machines είναι ένας αλγόριθμος μηχανικής μάθησης που αναπτύχθηκε από τον Vladimir Vapnik, μαζί με συνεργάτες. Η βασική ιδέα του αλγορίθμου, είναι η εύρεση του βέλτιστου υπερεπίπεδου που θα μεγιστοποιεί την απόσταση αναμεσά στις τάξεις. Ενδεικτικά η μέθοδος παρουσιάζεται γραφικά στο σχήμα 4.1:



Σχήμα 4.1. Χρήση του αλγορίθμου SVM σε γραμμικά διαχωρίσιμα δεδομένα.

Στο παραπάνω σχήμα τα όρια απόφασης (διακεκομμένες ευθείες) καθορίζουν την μέγιστη απόσταση. Οι διακεκομμένες ευθείες καλούνται διανύσματα υποστήριξης και παράγονται από τα κυκλωμένα σημεία του σχήματος 4.1. Αυτά καθορίζουν αποκλειστικά την μέγιστη απόσταση. Για καλύτερη ταξινόμηση χρειάζεται να βρεθεί η μέγιστη απόσταση ανάμεσα στις τάξεις. Για την εύρεση της μέγιστης απόστασης, χρειάζεται να γνωρίζουμε αν τα δεδομένα μπορούν να διαχωριστούν τέλεια.

Συμβολίζουμε με M την απόσταση του ορίου απόφασης και του υπερεπίπεδου και θέτουμε αυθαίρετα $M = {}^{1}/_{\|\beta\|}$. Η εύρεση του βέλτιστου υπερεπίπεδου προκύπτει από την επίλυση του παρακάτω προβλήματος:

$$ελαχιστοποίηση\ του\ \|\beta\|$$

Υπό τον περιορισμό:

$$y_i(b_0 + b_1x_{i1} + b_2x_{i2} \ldots + b_px_{ip}) \geq 1, \iota = 1, \ldots, N$$

Η ελαχιστοποίηση του $\|\beta\|$ εξασφαλίζει την μεγιστοποίηση της απόστασης M.

Στις περισσότερες περιπτώσεις όμως δεν έχουμε τελειά διαχωρίσιμα δεδομένα. Μια τακτική που ακολουθείται είναι να εισάγουμε τις χαλαρές μεταβλητές (slack variables) ξ_i, που επιτρέπουν σε έναν αριθμό παρατηρήσεων να βρίσκονται εντός των ορίων που καθορίζουν οι μηχανές υποστήριξης ή ακόμα και στην λάθος πλευρά του υπερεπίπεδου. Έτσι στα μη διαχωρίσιμα δεδομένα χρειάζεται να μεγιστοποιηθεί η ίδια ποσότητα M, λαμβάνοντας υπόψιν αυτή την φορά τις ανισότητες:

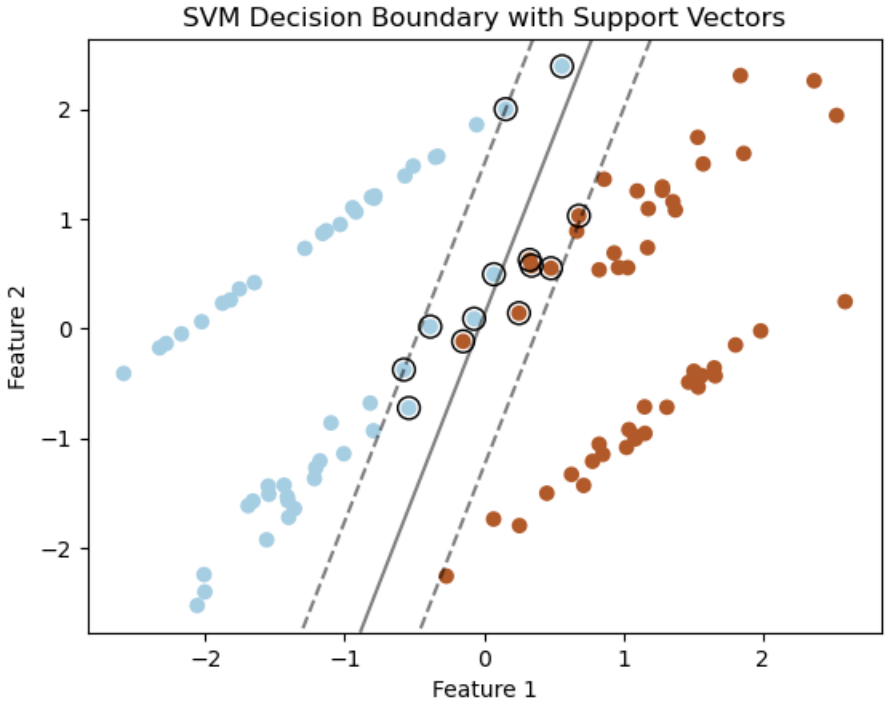
$$ελαχιστοποίηση\ του\ \|\beta\|$$

Υπό τους περιορισμούς:

$$y_i(b' \times x_i + b_0) \geq 1 - \xi_i$$
$$\xi_i \geq 0$$

$$\sum_{i=1}^{n} \xi_i \leq C$$

όπου ξ_i είναι μια επιπρόσθετη μεταβλητή που δίνει περιθώριο για λάθος ταξινόμηση ενός σημείου και C μια μη αρνητική υπερ-παράμετρος (James et al., 2013).



Σχήμα 4.2. Χρήση του αλγορίθμου SVM σε γραμμικά μη διαχωρίσιμα δεδομένα.

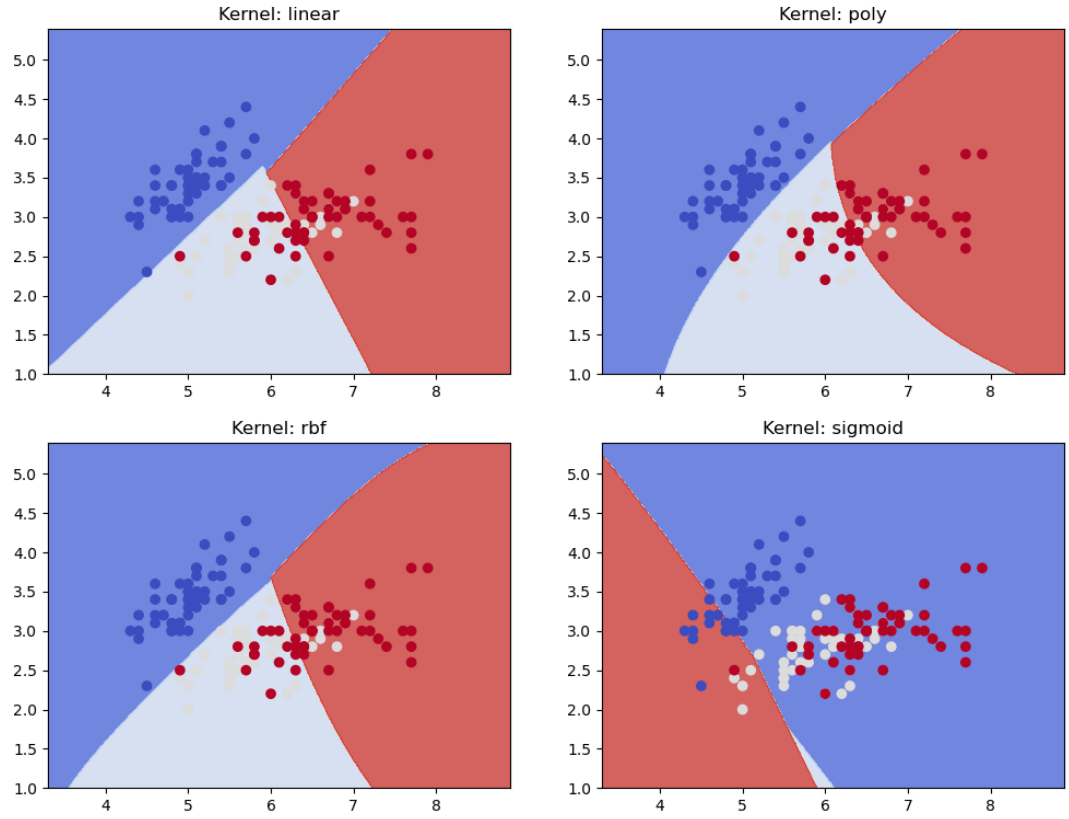
Για την διαχείριση μη γραμμικά διαχωρίσιμων δεδομένων, χρησιμοποιούνται οι συναρτήσεις kernel. Οι συναρτήσεις kernel απεικονίζουν μη γραμμικά διαχωρίσιμα δεδομένα σε έναν χώρο χαρακτηριστικών μεγαλύτερων διαστάσεων, όπου τα δεδομένα είναι γραμμικά διαχωρίσιμα. Για την επίτευξη αυτού του στόχου χρησιμοποιείται το τέχνασμα kernel (kernel trick).Το τέχνασμα kernel είναι μια τεχνική που επιτρέπει την προσαρμογή του ταξινομητή σε έναν χώρο υψηλότερων διαστάσεων. Αυτό προκύπτει μέσω υπολογισμού της ομοιότητας μεταξύ ζευγαριών των παρατηρήσεων. Η ομοιότητα υπολογίζεται μέσω μιας συνάρτησης kernel. Η επιλογή της συνάρτησης εξαρτάται από την φύση του προβλήματος. Οι πιο γνωστές συναρτήσεις kernel είναι οι εξής:

- Πολυώνυμη συνάρτηση kernel (Polynomial kernel function)
- Συνάρτηση ακτινωτής βάσης (Radial basis function (RBF))
- Σιγμοειδής συνάρτηση (Sigmoid function)
- Γραμμική συνάρτηση kernel (Linear kernel function)

Παρακάτω παρουσιάζονται οι τύποι της εκάστοτε συνάρτησης και στην συνέχεια στο σχήμα 4.3 γραφήματα που δείχνουν τις διαφορές που προκύπτουν από την επιλογή συνάρτησης (τα γραφήματα έγιναν στην Python).

Συνάρτηση	**Τύπος**
Linear Kernel Function	$K(\mathrm{x_i}, \mathrm{x_j}) = \mathrm{x_i}^{\mathrm{T}}\mathrm{x_j}$
Radial Basis Function	$K(\mathrm{x_i}, \mathrm{x_j}) = \mathrm{e}^{-\left(\frac{\lVert \mathrm{x_i} - \mathrm{x_j} \rVert^2}{2\sigma^2}\right)}$
Sigmoidal Kernel	$K(\mathrm{x_i}, \mathrm{x_j}) = \tanh(ax_i \times x_j - b)$
Polynomial Kernel Function	$K(\mathrm{x_i}, \mathrm{x_j}) = (\mathrm{x_i} \times \mathrm{x_j} + \mathrm{a})^{\mathrm{b}}$

Πίνακας 4.1 Συναρτήσεις kernel και οι τύποι τους



Σχήμα 4.3. Διαγραμματική απεικόνιση της διαφοράς που προκύπτει με την επιλογή διαφορετικής συνάρτησης

Οι συναρτήσεις kernel είναι ιδιαίτερα εύχρηστες υπολογιστικά, καθώς χρειάζεται να υπολογίσουν μόνο τα $K(\mathrm{x_i}, \mathrm{x_j})$ για όλα τα ζευγάρια i,j, χωρίς να βρίσκονται στον χώρο χαρακτηριστικών μεγαλύτερων διαστάσεων. Αυτό αποτελεί σημαντικό πλεονέκτημα, διότι συνήθως οι χώροι μεγαλύτερων διαστάσεων έχουν μεγάλο υπολογιστικό φόρτο.

4.3.2 Κ Κοντινότεροι γείτονες (K Nearest neighbours)

Ο αλγόριθμος Κ- Κοντινότεροι γείτονες (KNN), είναι ένας από τους γνωστότερους αλγορίθμους μηχανικής μάθησης. Η βασική ιδέα πίσω από τον αλγόριθμο, βρίσκεται στον υπολογισμό της απόστασης του σημείου που εισάγεται στα δεδομένα από τα υπόλοιπα σημεία και στην επιλογή του αριθμού 'γειτόνων', που βρίσκονται πιο κοντά στο σημείο που εξετάζουμε. Το πρόβλημα έγκειται στην επιλογή του μέτρου απόστασης και στην επιλογή του αριθμού των 'γειτόνων'.

Για την ταξινόμηση μιας άγνωστης παρατήρησης ακολουθούνται τα παρακάτω βήματα (James et al. 2013):

- Επιλέγεται ένας αριθμός k γειτόνων, δηλαδή ο αριθμός των κοντινότερων σημείων στην άγνωστη παρατήρηση.
- Επιλέγεται ένα μέτρο απόστασης.
- Έπειτα υπολογίζουμε την πιθανότητα το σημείο να ανήκει σε μια τάξη j:

$$P(Y = j) = \frac{N_j}{K}$$

- Η τάξη με την μεγαλύτερη πιθανότητα είναι αυτή που ταξινομείται το σημείο.

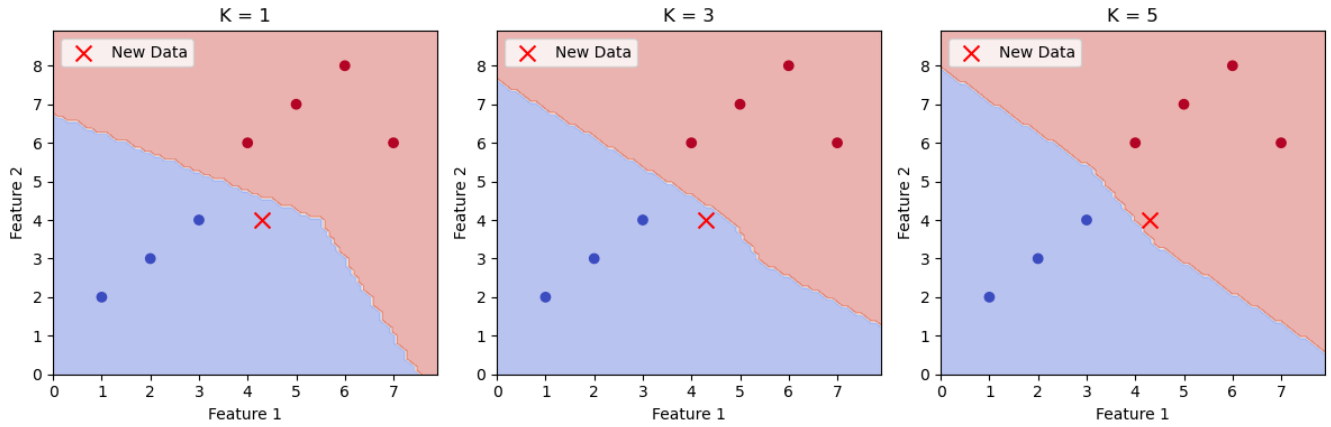
Για την επιλογή του μέτρου απόστασης υπάρχουν αρκετές επιλογές, με τις πιο συχνές να είναι η ευκλείδεια απόσταση και η απόσταση Manhattan.

Μέτρα απόστασης	**Τύπος**
Ευκλείδεια	$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
Manhattan	$\sum_{i=1}^{n}\lvert x_i - y_i\rvert$
Minkowski	$\left(\sum_{i=1}^{n}\lvert x_i - y_i\rvert^p\right)^{1/p}$
Chebyshev	$\max_i\lvert x_i - y_i\rvert$

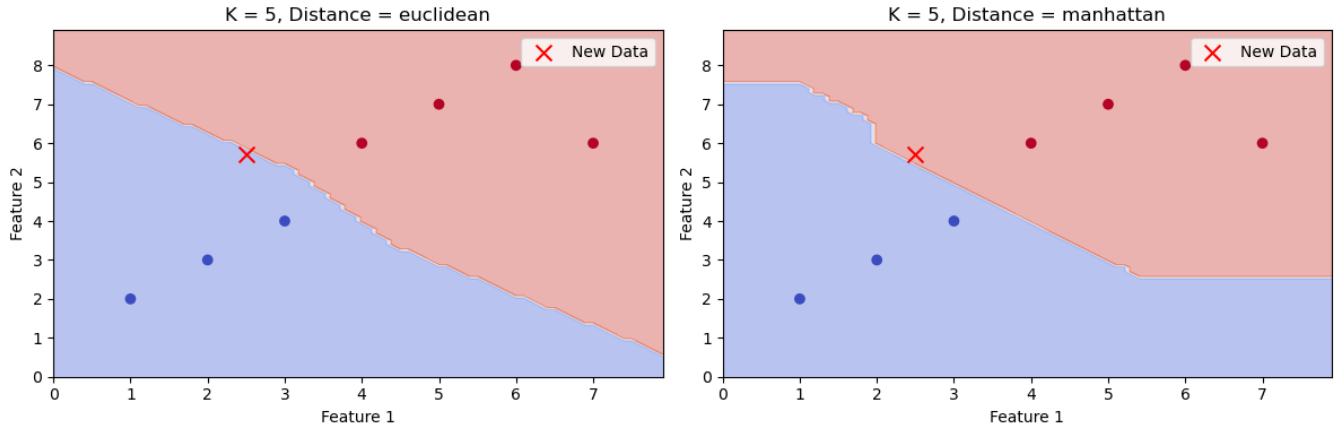
Πίνακας 4.2 Μέτρα απόστασης μαζί με τους τύπους τους.

Η κατάλληλη επιλογή αριθμού γειτόνων k είναι ιδιαίτερα σημαντική, διότι επηρεάζει την αποτελεσματικότητα του ταξινομητή (classifier). Η επιλογή ενός πολύ μικρού αριθμού k οδηγεί σε πολύ μικρή μεροληψία. Παρόλα αυτά αποφεύγεται, επειδή οδηγεί σε υπερπροσαρμογή (overfitting) και μεγάλη διακύμανση. Η επιλογή αριθμού γειτόνων και μέτρου απόστασης, διαφέρει ανάλογα με το πρόβλημα και απαιτεί εξέταση.

Παρακάτω παρατίθενται δυο σχήματα (σχήμα 4.4, σχήμα 4.5), όπου φαίνεται πως η επιλογή αριθμού 'γειτόνων' και μέτρων απόστασης επηρεάζει την τάξη που ταξινομείται το σημείο.



Σχήμα 4.4. Χρήση του αλγορίθμου KNN, με διαφορετικό αριθμό K. Παρατηρείται αλλαγή τάξης που τοποθετείται το σημείο για K=5



Σχήμα 4.5. Χρήση του αλγορίθμου KNN με διαφορετικά μέτρα απόστασης. Τα μέτρα απόστασης είναι η Ευκλείδεια απόσταση και η απόσταση Manhattan. Από το σχήμα φαίνεται πως επηρεάζονται η τάξη με βάση την απόσταση.

4.3.3 Λογιστική παλινδρόμηση (Logistic regression)

Η λογιστική παλινδρόμηση είναι ένα μοντέλο εποπτευομένης μηχανικής μάθησης, που χρησιμοποιείται σε προβλήματα ταξινόμησης. Η ιδέα πίσω από τον αλγόριθμο είναι η εύρεση των χαρακτηριστικών που προβλέπουν αποτελεσματικά μια δίτιμη εξαρτημένη μεταβλητή. Το μοντέλο της λογιστικής παλινδρόμησης δίνεται από την σιγμοειδή συνάρτηση:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

Όπου με β_0 συμβολίζεται ο σταθερός όρος και $\beta_1,\ldots,\beta_p$ συμβολίζονται οι συντελεστές των ανεξάρτητων μεταβλητών.

Το μοντέλο είναι κατασκευασμένο έτσι ώστε η εκτίμηση να βρίσκεται αναμεσά στο διάστημα [0,1].

Η λογιστική συνάρτηση μετά από πράξεις μπορεί να γραφτεί:

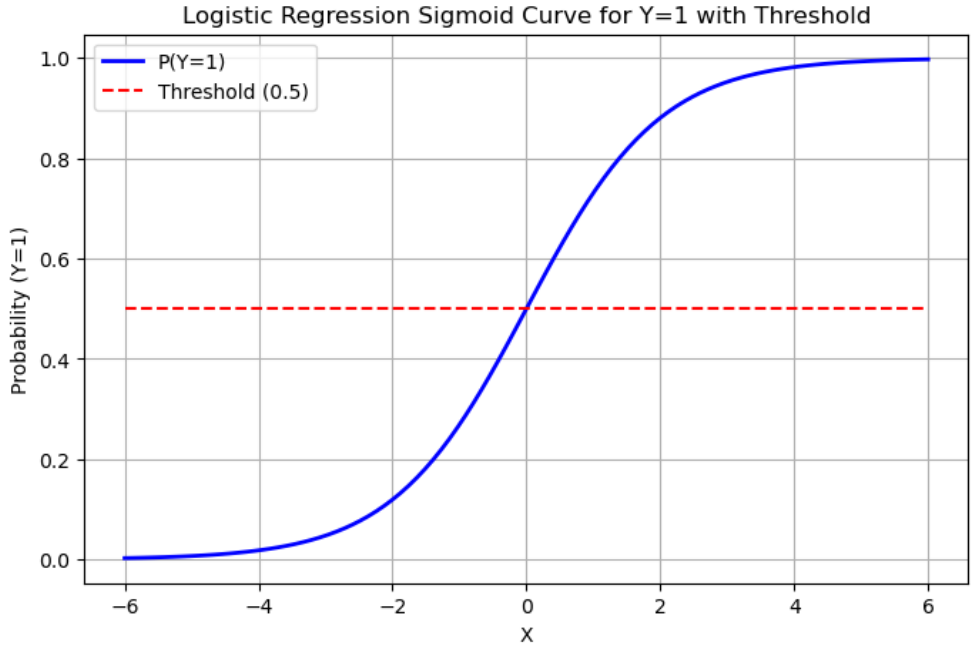
$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Όπου $\log\left(\frac{p(X)}{1-p(X)}\right)$ είναι ο λογάριθμος του λόγου των συμπληρωματικών πιθανοτήτων (log odds ratio).

Έτσι η λογιστική συνάρτηση μετατρέπεται σε ένα ισοδύναμο μοντέλο γραμμικής παλινδρόμησης με εξαρτημένη μεταβλητή το log odds ratio. Ένα σημαντικό πλεονέκτημα αυτής της μορφής είναι η ευκολία στην ερμηνεία των συντελεστών.

Η εύρεση των συντελεστών γίνεται με την μέθοδο της μέγιστης πιθανοφάνειας (James et al. 2013). Η βασική προϋπόθεση για τον καθορισμό των συντελεστών, είναι να οδηγούν όσο πιο κοντά γίνεται στο 0 ή 1, δηλαδή στην πραγματική τιμή μιας παρατήρησης. Οι συντελεστές μεγιστοποιούν την συνάρτηση πιθανοφάνειας.

Για την ταξινόμηση των τιμών που προβλέπει το μοντέλο. Χρειάζεται ένα σημείο αποκοπής (cutoff point). Οι τιμές που βρίσκονται πάνω από το κατώφλι ταξινομούνται ως επιτυχίες και αντίστοιχα οι τιμές κάτω από το κατώφλι ως αποτυχίες. Η συνήθης επιλογή του σημείου αποκοπής είναι το 0.5. Ενδεικτικά παρατίθεται στο σχήμα 4.6 ένα παράδειγμα λογιστικής παλινδρόμησης, με την διακεκομμένη γραμμή να είναι το σημείο αποκοπής.



Σχήμα 4.6. Απεικόνιση της καμπύλης της λογιστικής παλινδρόμησης. Η διακεκομμένη γραμμή απεικονίζει το κατώφλι.

Η λογιστική παλινδρόμηση είναι ένα ιδιαίτερα χρήσιμο εργαλείο στα basketball analytics, και συγκεκριμένα στην εύρεση των χαρακτηριστικών που μπορούν να προβλέψουν το αποτέλεσμα ενός αγώνα.

4.3.4 Γραμμική διαχωριστική ανάλυση (Linear discriminant analysis)

Η γραμμική διαχωριστική ανάλυση είναι μια μέθοδος εποπτευόμενης μάθησης, που χρησιμοποιείται για ταξινόμηση και μείωση διαστάσεων.

Βασική ιδέα του αλγορίθμου είναι η μείωση της διακύμανσης των σημείων εντός των τάξεων και η μεγιστοποίηση της διακύμανσης ανάμεσα στις τάξεις, για να διασφαλιστεί καλύτερη ταξινόμηση. Ο αλγόριθμος υποθέτει ότι τα δεδομένα ακολουθούν κανονική κατανομή και ότι οι διακυμάνσεις όλων των τάξεων είναι ίδιες (Balakrishnama, S. et al., 1998).

Αρχικά χρειάζεται ο υπολογισμός της διακύμανσης εντός και μεταξύ των τάξεων, όπως φαίνεται παρακάτω:

Η διακύμανση εντός των τάξεων:

$$S_w = \sum_{i=1}^{c} S_i$$

Όπου $S_i = \sum_{x \in D_i}^{n} (x - \mu_i)(x - \mu_i)^T$

Η διακύμανση μεταξύ των τάξεων:

$$S_b = \sum_j N_i(\mu_j - \mu) * (\mu_j - \mu)^T$$

Στην συνέχεια βρίσκουμε τις ιδιοτιμές του πίνακα:

$$S_w^{-1} S_b$$

Από την λύση της εξίσωσης εξετάζουμε τις ιδιοτιμές που προκύπτουν και κρατάμε τις μη μηδενικές. Από τα ιδιοδιανύσματα των μη μηδενικών τιμών σχηματίζεται ένας πίνακας W.

Πολλαπλασιάζοντας τον πίνακα W με τα αρχικά δεδομένα προκύπτει ο μετασχηματισμένος πίνακας με μειωμένες διαστάσεις. Για να προβλέψουμε την κλάση εκπαιδεύουμε έναν ταξινομητή στα μετασχηματισμένα δεδομένα.

Η γραμμική διαχωριστική ανάλυση έχει πολλά πλεονεκτήματα, όπως η μείωση διαστάσεων με διαχωρισμό τάξεων, η εύκολη ερμηνεία των αποτελεσμάτων και ανθεκτικότητα στις ακραίες τιμές. Όμως χρειάζεται σωστά ορισμένες τάξεις και είναι επιρρεπής στην υπερπροσαρμογή (overfitting)

4.3.5 Ο ταξινομητής Naïve Bayes (Naïve Bayes classifier)

Ο ταξινομητής Naïve Bayes είναι μια μέθοδος εποπτευόμενης μάθησης που εφαρμόζεται σε προβλήματα ταξινόμησης. Η μέθοδος βασίζεται στο θεμελιώδες θεώρημα του Bayes για την δεσμευμένη πιθανότητα. Κύριες υποθέσεις είναι η ανεξαρτησία των παραμέτρων (Hastie et al., 2009), που αν και δεν ανταποκρίνεται στα πραγματικά δεδομένα, κάνει τα προβλήματα ταξινόμησης πιο διαχειρίσιμα υπολογιστικά. Παρά το γεγονός ότι οι υποθέσεις παραβιάζονται, η μέθοδος θεωρείται ιδιαίτερα αποτελεσματική. Ο λόγος είναι ότι αν και η μέθοδος περιέχει μεροληψία δεν επηρεάζει σημαντικά τις εκ των υστέρων πιθανότητες που χρησιμοποιούνται για το συμπέρασμα.

Για την πρόβλεψη του ταξινομητή, θα πρέπει να γνωρίζουμε την εκ των προτέρων και τις συναρτήσεις πυκνότητας. Η εκ των υστέρων πιθανότητα έχει την μορφή:

$$P(Y = k|X = x) = \frac{\pi_\kappa \times f_{k1}(x_1) \times f_{k2}(x_2) \times \ldots \times f_{kp}(x_p)}{\sum_{l=1}^{K} \pi_l \times f_{l1}(x_1) \times f_{l2}(x_2) \times \ldots \times f_{lp}(x_p)}$$

Όπου με πκ συμβολίζεται η εκ των προτέρων πιθανότητα μια τυχαία παρατήρηση να ανήκει στην κ τάξη και με $f_{k\iota}(x_\iota)$ η συνάρτηση πυκνότητας του xi όταν ανήκει στην τάξη κi.

Η συνάρτηση πυκνότητας εξαρτάται από την φύση της μεταβλητής:

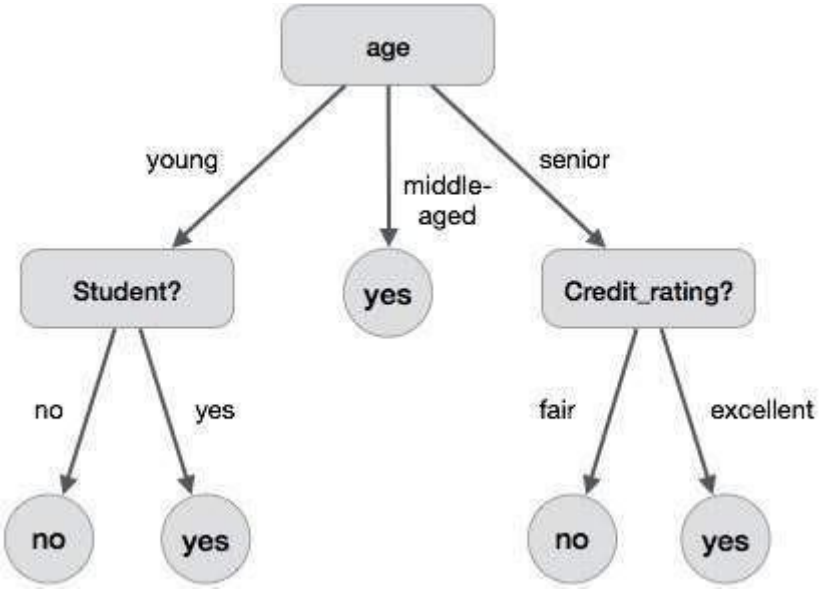
- Εάν το Xi είναι ποσοτική μεταβλητή υποθέτουμε ότι ακολουθεί κανονική κατανομή με την μέση τιμή και την διακύμανση να αλλάζουν ανάλογα την τάξη κ. Δηλαδή: $X_j|Y = k \sim N(\mu_{jk}, \sigma_{jk}^2)$
- Εάν το Xi είναι ποιοτική μεταβλητή τότε υποθέτουμε ότι ο ταξινομητής ακολουθεί πολυωνυμική κατανομή.

Η τυχαία παρατήρηση ταξινομείται όπου προκύπτει η μεγαλύτερη εκ των υστέρων πιθανότητα.

Ο ταξινομητής Naïve Bayes είναι μια μέθοδος ιδιαίτερα απλή, που δουλεύει αποτελεσματικά και προτιμάται όταν ο αριθμός των χαρακτηριστικών είναι μεγάλος.

4.3.6 Δέντρα απόφασης (Decision trees)

Τα δέντρα απόφασης είναι ένας αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται για ταξινόμηση. Η δομή του αλγορίθμου θυμίζει αυτή ενός δέντρου. Η δομή του μοντέλου αποτελείται από κόμβους (nodes), κλαδιά (branches) και φύλλα (leaves). Η ιδέα πίσω από τον αλγόριθμο έγκειται στον βέλτιστο διαχωρισμό του συνόλου δεδομένων. Αρχικά υπάρχει ο κόμβος ρίζα (root node) και μέσω επαναληπτικού διαχωρισμού (recursive splitting) δημιουργούνται διακλαδώσεις από κόμβους. Με κριτήριο τις αποφάσεις που λαμβάνονται από τις διακλαδώσεις, ο αλγόριθμος καταλήγει στην ταξινόμηση του σημείου στην τάξη που αντιστοιχεί στον τερματικό κόμβο (terminal node). Ένα βασικό υπόδειγμα παρατίθεται στο σχήμα 4.7:



Σχήμα 4.7 Ενδεικτική απεικόνιση δέντρου απόφασης (Πηγή: (Chapter 3 : Decision Tree Classifier — Theory, 2017))

Το πρόβλημα που δημιουργείται σε αυτή την διαδικασία είναι ο διαχωρισμός των δεδομένων. Σκοπός είναι οι κόμβοι να είναι όσο περισσότερο 'αγνοί', δηλαδή να περιέχουν όσο περισσότερες παρατηρήσεις από μια τάξη. Επιπλέον το δέντρο που δημιουργείται θα πρέπει να μην είναι υπερβολικά λεπτομερές, καθώς τότε ελλοχεύει ο κίνδυνος της υπερπροσαρμογής. Για αυτό τον λόγο συχνά χρειάζεται να χρησιμοποιηθεί κάποιο κριτήριο 'κλαδέματος' (pruning). Για την επίλυση τέτοιων προβλημάτων τα δημοφιλέστερα μέτρα είναι:

- **Σφάλμα λανθασμένης ταξινόμησης (Misclassification rate)**: Το σφάλμα λανθασμένης ταξινόμησης μετράει το ποσοστό των παρατηρήσεων που δεν έχουν ταξινομηθεί σωστά. Δίνεται από τον τύπο:

$$1 - \max_{k} \hat{p}_{mk}$$

με το $\widehat{p_{mk}}$ να είναι το ποσοστό των παρατηρήσεων της k τάξης στην m περιοχή.

- **Δείκτης Gini (Gini index)**: Ο δείκτης Gini αποτελεί ένα μέτρο διακύμανσης ανάμεσα στις τάξεις (James et al. (2013)). Σκοπός είναι η εύρεση της περιοχής και της τάξης που ελαχιστοποιούν τον δείκτη. Δίνεται από τον τύπο:

$$1 - \sum_{\mathrm{k}=1}^{\mathrm{K}} \hat{p}_{mk}\,(1 - \hat{p}_{mk})$$

- **<u>Εντροπία (Entropy)</u>**: Η εντροπία είναι ένα ακόμη εναλλακτικό μέτρο μέτρησης της αγνότητας των κόμβων. Όσο πιο αγνός είναι ένας κόμβος τόσο μικρότερη είναι η εντροπία. Δίνεται από τον τύπο:

$$\mathrm{H} = -\sum_{\mathrm{k}=1}^{\mathrm{K}} \hat{p}_{mk} \log \hat{p}_{mk}$$

Ο δείκτης Gini και η εντροπία είναι οι συνήθεις επιλογές για τα κριτήρια διαχωρισμού, ενώ για το κλάδεμα είναι το σφάλμα λανθασμένης ταξινόμησης (Hastie et al., 2009).

Τα δέντρα απόφασης είναι μια τεχνική που λόγω της εύκολης της ερμηνείας προτιμάται συχνά. Ένα σημαντικό μειονέκτημα τους όμως είναι ότι έχουν μεγάλη διακύμανση και συνεπώς η προβλεπτική τους ικανότητα δεν είναι τόσο καλή όσο άλλων ταξινομητών.

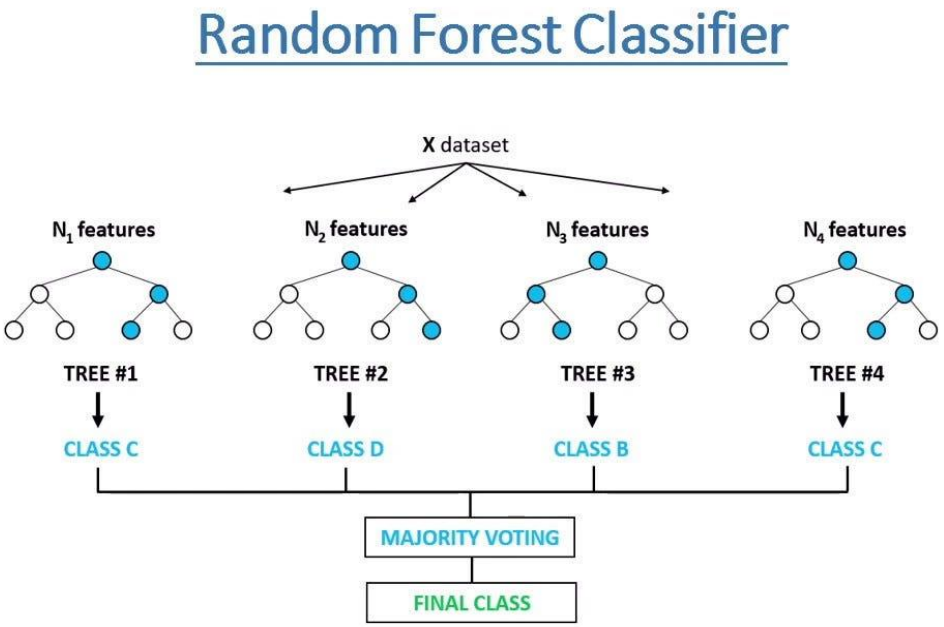
4.3.7 Τυχαία δάση (Random forests)

Τα τυχαία δάση είναι ένας αλγόριθμος εποπτευομένης μάθησης που εφαρμόζεται σε προβλήματα ταξινόμησης. Η μέθοδος είναι μια επέκταση του bootstrapped aggregating (bagging), για αυτό και θα γίνει αρχικά μια εισαγωγή σε αυτή την έννοια.

Η μέθοδος bagging χρησιμοποιεί έναν αριθμό B bootstrapped δειγμάτων, δηλαδή δειγμάτων που προκύπτουν από δειγματοληψία του αρχικού συνόλου με επανάθεση, για τον υπολογισμό μιας στατιστικής συνάρτησης. Χρησιμοποιώντας τον μέσο όρο των B δειγμάτων, μπορούμε να εκτιμήσουμε κάποιο μοντέλο. Η μέθοδος χρησιμοποιείται, διότι μειώνει την διακύμανση και βελτιώνει την απόδοση σε μοντέλα όπως αυτά των δέντρων απόφασης. Στην περίπτωση των δέντρων απόφασης επιπλέον η μεροληψία των bagged δέντρων είναι ίδια με την μεροληψία ενός μοναδικού δέντρου. Όμως, τα δέντρα που προκύπτουν μπορεί να έχουν μεγάλη συσχέτιση, εάν υπάρχουν χαρακτηριστικά που επηρεάζουν πολύ την πρόβλεψη.

Τα τυχαία δάση επεκτείνουν την παραπάνω τεχνική, αντιμετωπίζοντας το πρόβλημα της συσχέτισης. Στα τυχαία δάση επιλέγονται ξανά bootstrapped δείγματα, αλλά για κάθε δείγμα επιλέγονται τυχαία m από τις p μεταβλητές του συνόλου δεδομένων. Με αυτό τον τρόπο μειώνεται η συσχέτιση και οι προβλέψεις που προκύπτουν από τα δέντρα είναι πιο αξιόπιστες. Ο αριθμός m των τυχαίων χαρακτηριστικών που επιλέγονται είναι συνήθως $\sqrt{p}$ (James et al., 2013).

Η τελική ταξινόμηση δίνεται μέσα από τα αποτελέσματα των δέντρων απόφασης. Πιο συγκεκριμένα κάθε δέντρο απόφασης λαμβάνει μια ‘’ψήφο’’ και με κριτήριο την πλειοψηφία των προβλέψεων προκύπτει η τελική ταξινόμηση.



Σχήμα 4.8. Ενδεικτική απεικόνιση του ταξινομητή τυχαίων δασών.

Πηγή: Random Forest Classifier and its Hyperparameters | by Ankit Chauhan | Analytics Vidhya | Medium

4.4 Μέθοδοι παλινδρόμησης (Regression methods)

Η παλινδρόμηση είναι άλλο ένα θεμελιώδες κομμάτι της εποπτευομένης μάθησης. Οι μέθοδοι παλινδρόμησης στοχεύουν στην πρόβλεψη μιας συνεχούς μεταβλητής.

Πιο αναλυτικά, οι τεχνικές παλινδρόμησης χρησιμοποιούν τα επισημασμένα δεδομένα για την εύρεση σχέσης μεταξύ της μεταβλητής απόκρισης (response variable) και των ανεξάρτητων μεταβλητών και μέσω αυτής γίνονται προβλέψεις για μη επισημασμένα δεδομένα.

Οι τεχνικές που θα αναλυθούν παρακάτω είναι:

- Γραμμική παλινδρόμηση (Linear regression)
- Παλινδρόμηση ridge (Ridge regression)
- Παλινδρόμηση lasso (Lasso regression)
- Δέντρα παλινδρόμησης (Regression trees)

4.4.1 Γραμμική παλινδρόμηση (Linear Regression)

Η γραμμική παλινδρόμηση είναι η πιο γνωστή μέθοδος παλινδρόμησης. Σκοπός είναι η εύρεση του επιπέδου που εκφράζει καταλληλότερα την γραμμική σχέση ανάμεσα στην μεταβλητή στόχο (target variable) και τις ανεξάρτητές μεταβλητές. Όταν η μεταβλητή στόχος εξαρτάται από μια ανεξάρτητη μεταβλητή τότε έχουμε απλή γραμμική παλινδρόμηση. Στις περισσότερες περιπτώσεις και συγκεκριμένα στα πλαίσια αυτής της εργασίας οι μεταβλητές είναι περισσότερες. Σε αυτές τις περιπτώσεις εφαρμόζεται πολλαπλή γραμμική παλινδρόμηση.

Η γραμμική σχέση εκφράζεται από τον τύπο:

$$Y = \hat{b}_0 + \hat{b}_1 X + \varepsilon \ \ (Απλή)$$

$$Y = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \ldots + \hat{b}_p x_p + \varepsilon \ \ (Πολλαπλή)$$

Με Y συμβολίζεται η εξαρτημένη μεταβλητή, με xi (για i=1,...p) όλες οι ανεξάρτητες μεταβλητές. Το b0 είναι η σταθερά, ενώ τα bi παριστάνουν την μεταβολή της Y όταν η μεταβλητή xi αυξηθεί κατά μια μονάδα. Επιπλέον το ε αντιπροσωπεύει το σφάλμα ανάμεσα στην πραγματική και την εκτιμώμενη τιμή.

Για την εύρεση του κατάλληλου μοντέλου υπάρχουν διάφορες μέθοδοι, αλλά αυτή που χρησιμοποιείται κατά κύριο λόγο είναι η μέθοδος ελαχίστων τετράγωνων, όπου ελαχιστοποιούμε το άθροισμα των καταλοίπων ε.

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [Y_i - (\hat{b}_0 + \hat{b}_1 x_1)]^2 \ \ (Απλή)$$

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [Y_i - (\hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \ldots + \hat{b}_p x_p)]^2 \ \ (Πολλαπλή)$$

Από το παραπάνω προκύπτει ότι οι εκτιμητές ελαχίστων τετραγώνων δίνονται από την σχέση:

- Απλή γραμμική παλινδρόμηση:

$$\hat{b}_1 = \frac{n \sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$\hat{b}_0 = \frac{1}{n} \sum_{i=1}^{n} y_i - \hat{b}_1 \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Πολλαπλή γραμμική παλινδρόμηση:

$$\hat{b} = (X^T X)^{-1} X^T Y$$

Οι προσαρμοσμένες τιμές:

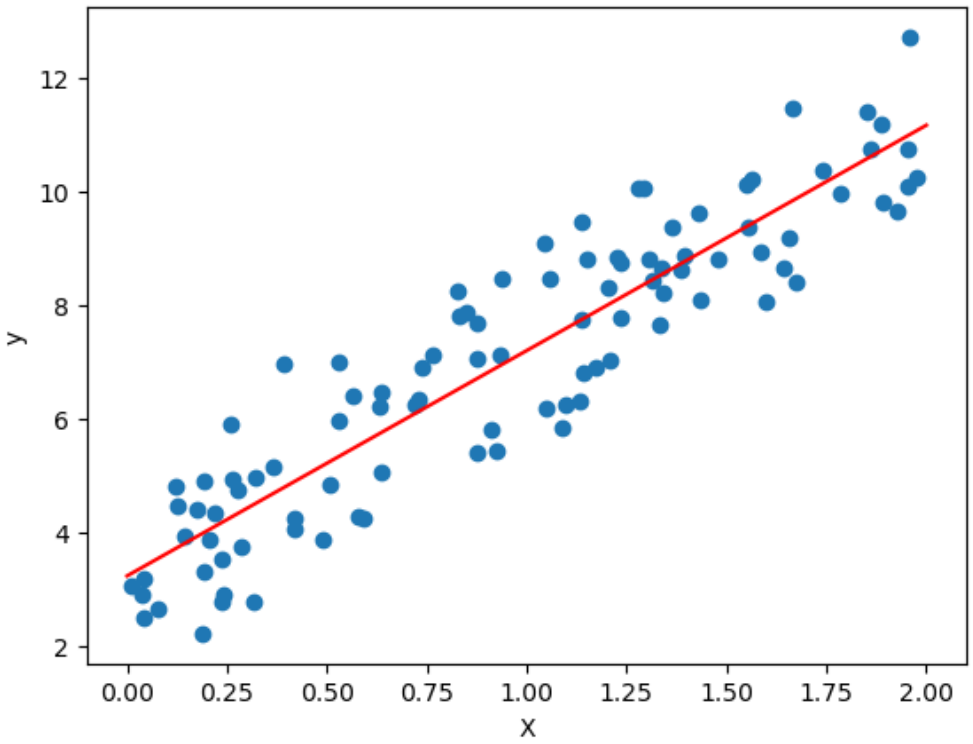
$$\hat{Y}_\iota = \hat{b}_0 + \hat{b}_1 X \ \ (Απλή)$$

$$\hat{Y} = X\hat{b} \ (Πολλαπλή)$$

Και τα εκτιμημένα κατάλοιπα:

$$\hat{\varepsilon} = Y - \hat{Y}$$

Ένα παράδειγμα γραμμικής παλινδρόμησης φαίνεται στο παρακάτω σχήμα:



Σχήμα 4.9. Ενδεικτικό παράδειγμα απλής γραμμικής παλινδρόμησης. Η κόκκινη ευθεία προκύπτει από τη μέθοδο ελαχίστων τετραγώνων.

Η μέθοδος γραμμικής παλινδρόμησης διακρίνεται για την απλότητα και την ευκολία στην ερμηνεία της, λόγω της υπόθεσης των γραμμικών σχέσεων. Παρόλα αυτά η υπόθεση της γραμμικής σχέσης την κάνει αναποτελεσματική σε μη γραμμικά δεδομένα, ενώ σε περιπτώσεις πολυσυγγραμικότητας, υπάρχει ο κίνδυνος οι συντελεστές να μην αποτυπώνουν την πραγματική σημασία μιας μεταβλητής.

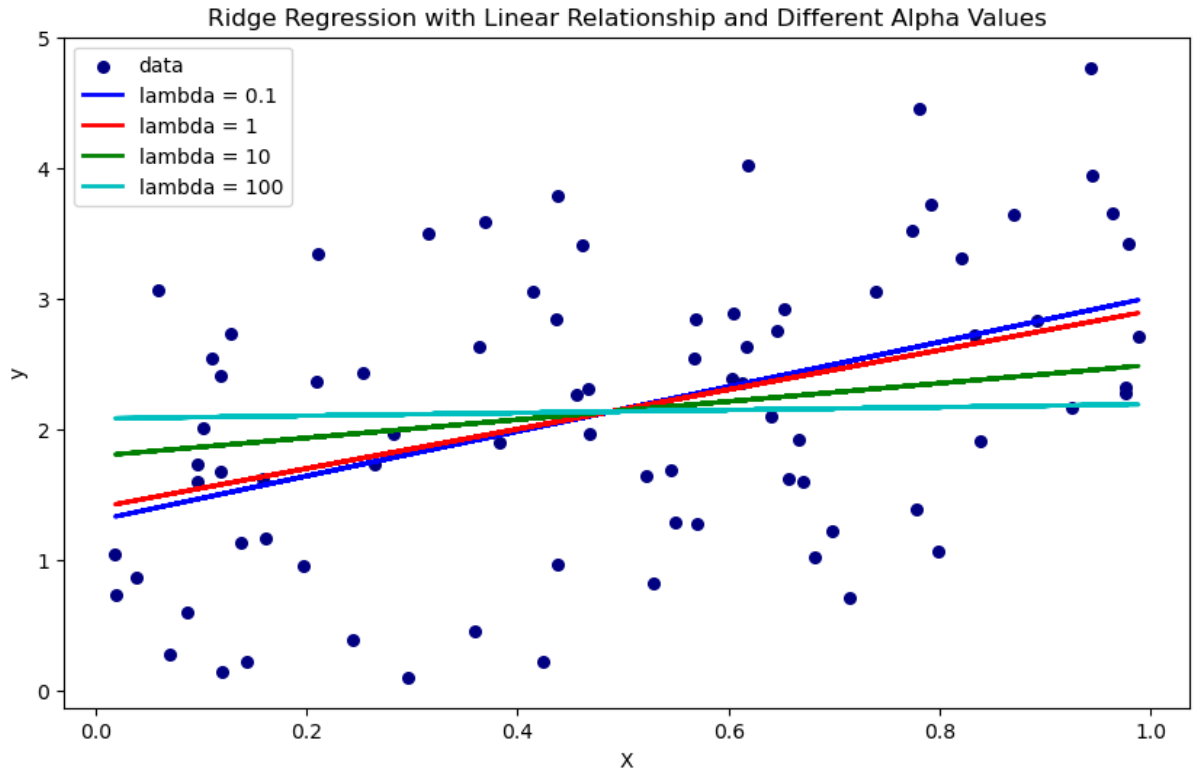
4.4.2 Παλινδρόμηση Ridge (Ridge regression)

Η παλινδρόμηση Ridge είναι μια μέθοδος συρρίκνωσης. Ένα πρόβλημα που μπορεί να προκύψει στην γραμμική παλινδρόμηση, είναι η μεγάλη συσχέτιση χαρακτηριστικών και συνεπώς η ύπαρξη πολυσυγγραμικότητας. Η ύπαρξη πολυσυγγραμικότητας οδηγεί σε μεγάλη διακύμανση και συνεπώς αναξιόπιστες προβλέψεις. Οι μέθοδοι συρρίκνωσης και σε αυτή την περίπτωση η παλινδρόμηση Ridge καταπολεμούν το πρόβλημα της πολυσυγγραμικότητας, μειώνοντας την διακύμανση (James et al., 2013).

Πιο αναλυτικά ο αλγόριθμος μειώνει την διακύμανση χρησιμοποιώντας έναν όρο ποινής, που μεταφέρει τις τιμές των συντελεστών προς το 0. Ο όρος ποινής στην παλινδρόμηση Ridge ονομάζεται κανονικοποίηση L2 και προστίθεται στο κλασικό άθροισμα ελαχίστων τετράγωνων. Οι συντελεστές που επιλέγονται είναι αυτοί που ελαχιστοποιούν την συνάρτηση:

$$\sum_{\mathrm{i}=1}^{\mathrm{n}}[\mathrm{Y_i} - b_o - \sum_{j=1}^{p} b_j\, x_{ij}]^2 + \lambda \sum_{j=1}^{p} b_j^2$$

Με το λ να είναι μια παράμετρος που καθορίζεται μετά από διασταυρουμένη επικύρωση (cross validation) και καθορίζει το μέγεθος της συρρίκνωσης των συντελεστών. Όσο μεγαλύτερο είναι το λ τόσο μικραίνουν οι συντελεστές και έτσι οι προβλέψεις είναι λιγότερο ευαίσθητες. Στο σχήμα 4.10 επιβεβαιώνεται ο παραπάνω ισχυρισμός. Φαίνεται ότι η γραμμή για λ=100 είναι αρκετά πιο επίπεδη από την ευθεία για λ=0.1. Ο καθορισμός του λ επίσης μεταβάλλει την μεροληψία και την διακύμανση ενός μοντέλου. Όταν το λ αυξάνεται, τότε αυξάνεται και η μεροληψία του μοντέλου, αλλά ταυτόχρονα μειώνεται η διακύμανση του.



Σχήμα 4.10. Ενδεικτική απεικόνιση της παλινδρόμησης Ridge για διαφορετικές τιμές του λ. Παρατηρούμε ότι όσο μεγαλώνει το λ τόσο η ευθεία γίνεται πιο επίπεδη.

Βασικό προτέρημα του αλγόριθμου είναι ότι αποφεύγει προβλήματα υπερπροσαρμογής, ενώ ακόμα η ύπαρξη του όρου ποινής επιτρέπει τον χειρισμό της πολυσυγγραμικότητας (σε αντίθεση με την γραμμική παλινδρόμηση). Ωστόσο ο αλγόριθμος, όσο και αν μειώνει τους συντελεστές η αύξηση του λ, το τελικό μοντέλο θα περιέχει όλα τα αρχικά χαρακτηριστικά. Οι συντελεστές τείνουν, αλλά δεν μηδενίζονται ποτέ. Έτσι γίνεται δυσκολότερη την ερμηνεία του μοντέλου, ειδικά όταν ο αριθμός των χαρακτηριστικών είναι μεγάλος.

4.4.3 Παλινδρόμηση lasso (Lasso regression)

Η παλινδρόμηση lasso είναι ένας αλγόριθμος εποπτευομένης μάθησης που χρησιμοποιείται για παλινδρόμηση. Ο αλγόριθμος ,όπως και η παλινδρόμηση ridge, αποτελεί άλλη μια μέθοδο συρρίκνωσης. Πιο αναλυτικά η παλινδρόμηση lasso χρησιμοποιεί επίσης έναν όρο ποινής, την κανονικοποίηση L1. Η κανονικοποίηση L1, όπως και η κανονικοποίηση L2, προστίθεται στο άθροισμα ελαχίστων τετραγώνων και αυτή την φορά αντί για τον όρο $\sum_{j=1}^{p} b_j^2$ χρησιμοποιειται

το άθροισμα της απόλυτης τιμής των συντελεστών b_i, δηλαδή $\sum_{j=1}^{p}|b_j|$. Επομένως η συνάρτηση που χρειάζεται να ελαχιστοποιηθεί είναι:

$$\sum_{\mathrm{i}=1}^{\mathrm{n}}[\mathrm{Y_i} - b_o - \sum_{j=1}^{p} b_j\, x_{ij}]^2 + \lambda \sum_{j=1}^{p}|b_j|$$

Όπου η παράμετρος λ λειτουργεί όπως και στην παλινδρόμηση ridge.
Ισοδύναμα το πρόβλημα ελαχιστοποίησης γράφεται ως:

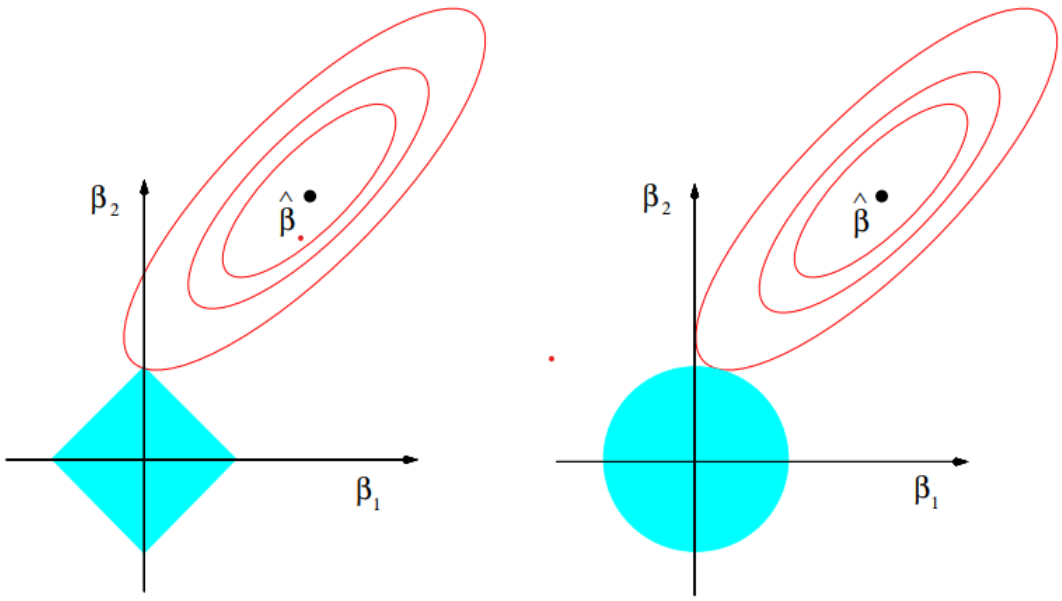
- Ελαχιστοποίηση του αθροίσματος:

$$\sum_{i=1}^{p}[\mathrm{Y_i} - b_o - \sum_{j=1}^{p} b_j\, x_{ij}]^2$$

- Υπό τον περιορισμό:

$$\sum_{j=1}^{p}|b_j| \leq t$$

Ένα από τα μειονεκτήματα της παλινδρόμησης Ridge ήταν ότι η αύξηση του λ, οδηγούσε τους συντελεστές προς το 0, αλλά δεν μηδενίζοταν, με αποτέλεσμα το τελικό μοντέλο να περιέχει όλα τα χαρακτηριστικά. Με την χρήση της κανονικοποίησης L1, η παλινδρόμηση Lasso δεν αντιμετωπίζει τέτοιο πρόβλημα, καθώς από τον περιορισμό $\sum_{j=1}^{p}|b_j| \leq t$ συμπεραίνουμε ότι η εφικτή περιοχή είναι ένα κυρτό σύνολο. Εξαιτίας του κυρτού συνόλου σχηματίζονται γωνίες και όποτε αυτές τέμνονται κάποιος συντελεστής εξαλείφεται. Αυτή η ιδιότητα της lasso κάνει την ερμηνεία του τελικού μοντέλου πιο κατανοητή από την παλινδρόμηση ridge. Στο σχήμα 4.11 φαίνεται η εφικτή περιοχή αριστερά για την παλινδρόμηση Lasso και δεξιά την Ridge.



Σχήμα 4.11. Γραφική απεικόνιση της παλινδρόμησης Lasso (αριστερά) και Ridge (δεξιά).Πηγή:Hastie et al.(2009)

4.4.4 Δέντρα παλινδρόμησης (Regression trees)

Τα δέντρα παλινδρόμησης είναι μια δημοφιλής μέθοδος παλινδρόμησης. Τα δέντρα παλινδρόμησης είναι πανομοιότυπα με τα δέντρα ταξινόμησης. Παρουσιάζουν την ίδια δομή και βασίζονται στην ίδια λογική, δηλαδή την εύρεση του διαχωρισμού και την κατασκευή του βέλτιστου δέντρου.

Για την κατασκευή του δέντρου θα πρέπει να δημιουργηθούν κόμβοι, που θα οδηγήσουν στην τελική πρόβλεψη. Για τα κριτήρια διαχωρισμού αρχικά χωρίζουμε τα δεδομένα με μια μεταβλητή j και ένα σημείο διαχωρισμού s, τέτοια ώστε να δημιουργούνται τα ημι- επίπεδα:

$$R_1(j,s) = \{X | X_j < s\} \; και \; R_2(j,s) = \{X | X_j \geq s\}$$

Στόχος είναι η εύρεση του ζευγαριού (j, s) που ελαχιστοποιεί την εξίσωση:

$$\sum_{i:\, x_i \in R_1(j,s)} (y_i - \bar{y}_{R_1})^2 + \sum_{i:\, x_i \in R_2(j,s)} (y_i - \bar{y}_{R_2})^2$$

Όπου τα $\bar{y}_{R_1}$ και $\bar{y}_{R_2}$ είναι οι μέσες τιμές των μεταβλητών απόκρισης για τις παρατηρήσεις που ανήκουν στα $R_1(j,s)$ και $R_2(j,s)$ αντίστοιχα. Αυτή η διαδικασία επαναλαμβάνεται διαδοχικά μέχρι να καταλήξει σε κάποιο κριτήριο διακοπής (stopping criterion). Ένα γνωστό κριτήριο διακοπής είναι η επιλογή ενός προκαθορισμένου μεγέθους των τερματικών κόμβων.

Το δέντρο, που προκύπτει συνήθως είναι αρκετά μεγάλο και ενδεχομένως να οδηγεί σε υπερπροσαρμογή. Έτσι δημιουργείται η ανάγκη κλαδέματος του. Μια λύση είναι το κλάδεμα κόστους πολυπλοκότητας (cost-complexity pruning), μια τεχνική κλαδέματος που μειώνει την πολυπλοκότητα ενός δέντρου απόφασης. Ο τύπος είναι:

$$\sum_{m=1}^{|T|} \sum_{x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$$

Όπου το πρώτο μέρος είναι το άθροισμα τετραγώνων υπολοίπων (RSS) για όλες τις παρατηρήσεις που ανήκουν στο διάστημα Ri και το γινόμενο α|T| αποτελείται από την παράμετρο α που καθορίζεται από τον χρήστη και το |T| που συμβολίζει τον αριθμό των τερματικών κόμβων ή φύλλων.

Η παράμετρος α καθορίζει το μέγεθος ενός δέντρου. Η αύξηση του οδηγεί σε μεγαλύτερο αριθμό κλαδέματος και επομένως ένα απλουστευμένο δέντρο. Όμως το υπερβολικό κλάδεμα μπορεί να έχει κόστος στην ακρίβεια του μοντέλου, για αυτό απαιτείται προσοχή στην επιλογή του α. Σκοπός είναι η εύρεση του α που ελαχιστοποιεί τον παραπάνω τύπο. Μια τακτική επιλογής της παραμέτρου είναι μέσω διασταυρούμενης επικύρωσης (cross-validation), που θα μελετήσουμε παρακάτω.

4.5 Μέθοδοι συσταδοποίησης (Clustering methods)

Η συσταδοποίηση είναι ένας τύπος μη εποπτευομένης μάθησης, που χρησιμοποιείται για την εύρεση αγνώστων υποομάδων στα δεδομένα. Συγκεκριμένα η συσταδοποίηση, στοχεύει στην κατασκευή υποομάδων, με τα εξής γνωρίσματα:

- Οι παρατηρήσεις της κάθε υποομάδας να είναι παρόμοιες μεταξύ τους.
- Οι ομάδες μεταξύ τους να παρουσιάζουν μεγάλες διαφορές.

Ο καθορισμός ομοιοτήτων και διάφορων εξαρτάται από τα δεδομένα.

Οι τεχνικές που θα μελετηθούν παρακάτω είναι οι K-means και η ιεραρχική συσταδοποίηση.

4.5.1 Συσταδοποίηση K-means (K-means clustering)

Η συσταδοποίηση με τον αλγόριθμο K- means είναι μια πολύ δημοφιλής μη ιεραρχική μέθοδος. Βασική ιδέα της μεθόδου είναι ο διαχωρισμός των δεδομένων σε K ομάδες, με την διακύμανση εντός των συστάδων να είναι όσο το δυνατόν μικρότερη. Για την επίτευξη μιας καλής συσταδοποίησης, ο αλγόριθμος ελαχιστοποιεί την διακύμανση στα δεδομένα εντός των υποομάδων, για να υπάρχει ομοιογένεια. Δηλαδή η ποσότητα που χρειάζεται να ελαχιστοποιηθεί είναι:

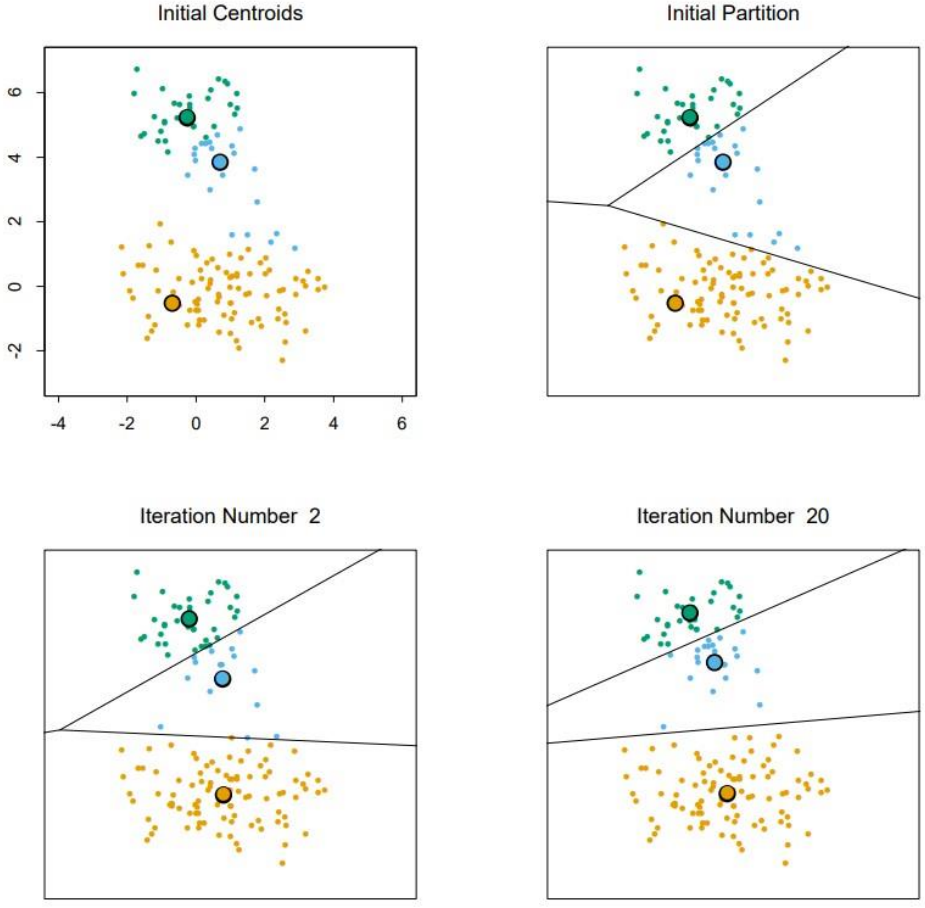
$$\frac{1}{|C_k|}\sum_{i,i' \in C_k}\sum_{j=1}^{p}\left(x_{ij} - x_{i'j}\right)^2$$

Όπου |Ck| είναι ο αριθμός των παρατηρήσεων σε μια συστάδα (James et al., 2013).

Επειδή η ελαχιστοποίηση της παραπάνω ποσότητας είναι ένα πολύ απαιτητικό πρόβλημα, χρησιμοποιείται η παρακάτω διαδικασία:

- Τοποθετούμε κάθε παρατήρηση τυχαία στις K συστάδες.
- Υπολογίζουμε το κέντρο βάρους κάθε συστάδας και στην συνέχεια τα χρησιμοποιούμε για να κατατάξουμε ξανά τις παρατηρήσεις σε K υποομάδες. (Για την εύρεση των κοντινότερων σημείων χρησιμοποιείται η ευκλείδεια απόσταση)
- Η διαδικασία αυτή επαναλαμβάνεται μέχρι να μην παρατηρηθεί καμία αλλαγή στις υποομάδες.

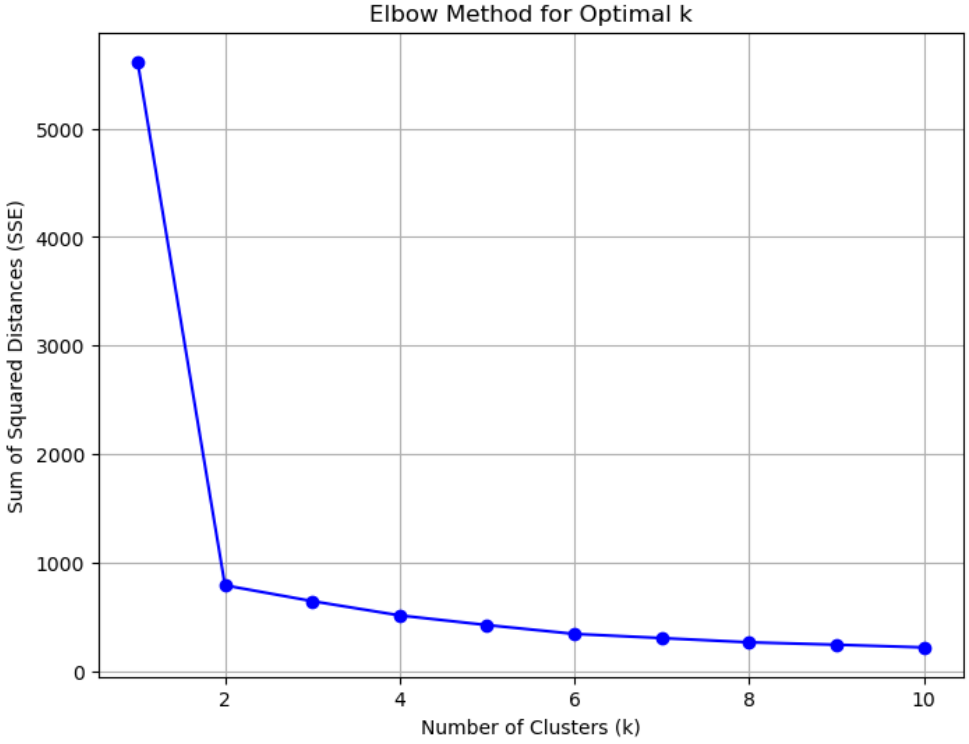
Η παραπάνω διαδικασία απεικονίζεται και γραφικά στο σχήμα 4.12.



Σχήμα 4.12. Διαγραμματική απεικόνιση του αλγορίθμου K means (Πηγή: Hastie et al.(2009))

Ο αλγόριθμος είναι υπολογιστικά οικονομικός, καθώς δεν χρειάζεται πολλές επαναλήψεις, για να πάρει την τελική του μορφή. Όμως ο σχηματισμός των υποομάδων, επηρεάζεται σε μεγάλο βαθμό από τον αρχικό καταμερισμό των μητρικών σημείων. Για αυτό τον λόγο χρειάζεται να γίνουν δοκιμές με διαφορετικά μητρικά σημεία.

Ένα ακόμη σημαντικό πρόβλημα είναι η επιλογή του αριθμού K. Η επιλογή καθορίζει εν πολλοίς την ποιότητα της συσταδοποίησης, για αυτό χρειάζεται γνώση του εκάστοτε προβλήματος. Υπάρχουν διάφορες τεχνικές επιλογής, με την πιο γνωστή να είναι η μέθοδος του αγκώνα (elbow method). Η μέθοδος του αγκώνα απεικονίζει γραφικά το άθροισμα τετραγώνων των υπολοίπων για διάφορες του K και θεωρεί ως κριτήριο επιλογής, το σημείο που σχηματίζεται ένας 'αγκώνας', δηλαδή η μείωση του αθροίσματος γίνεται πιο αργά. Στο παρακάτω σχήμα (4.13) ο αγκώνας εμφανίζεται όταν ο αριθμός των συστάδων είναι k=2.



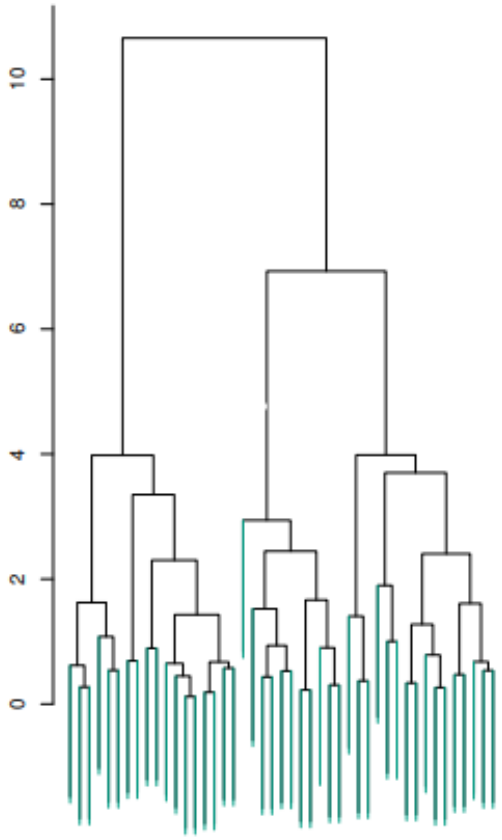
Σχήμα 4.13. Ενδεικτικό διάγραμμά για την μέθοδο αγκώνα. Από το διάγραμμα γίνεται σαφές ότι το K είναι γύρω στο 2.

4.5.2 Ιεραρχική συσταδοποίηση (Hierarchical clustering)

Η ιεραρχική συσταδοποίηση είναι μια μέθοδος συσταδοποίησης, για την οποία δεν χρειάζεται να καθοριστεί από την αρχή ένας αριθμός συστάδων. Η τεχνική χωρίζεται σε δυο κατηγορίες: την συσσωρευτική (agglomerative) και την διαιρετική (divisive). Λόγω της μικρής πρακτικής χρησιμότητας των διαιρετικών μεθόδων, θα εστιάσουμε στις συσσωρευτικές που βρίσκουν εφαρμογή στα πλαίσια της εργασίας.

Χαρακτηριστικό του αλγόριθμου είναι η επιλογή του αριθμού των συστάδων μέσω ενός δενδρογράμματος. Το δενδρόγραμμα ξεκινάει με όλες τις παρατηρήσεις να είναι ανεξάρτητες συστάδες και ενώνει διαδοχικά όλα τα σημεία. Όσο νωρίτερα ενωθούν κάποια σημεία ή κάποιες ομάδες με σημεία, τόσο πιο παρόμοια είναι μεταξύ τους. Για την επιλογή του αριθμού των υποομάδων, παρατηρούμε το σημείο στο όποιο παρατηρείται η μεγαλύτερη μεταβολή στον άξονα απόστασης.

Προχωρώντας στην λειτουργία του αλγόριθμου, αρχικά θεωρούμε, (όπως απεικονίζεται και στο παρακάτω δενδρόγραμμα) την κάθε παρατήρηση ανεξάρτητη και μέσω ενός μέτρου απόστασης (συνήθως ευκλείδεια απόσταση) ενώνουμε τα κοντινότερα σημεία.



Σχήμα 4.14. Ενδεικτική απεικόνιση ενός δενδροδιαγράμματος (Πηγή: James et al.(2013))

Από αυτή την διαδικασία δημιουργείται το πρόβλημα ένωσης ομάδων. Για τον υπολογισμό των αποστάσεων των ομάδων, υπάρχουν διάφορες τεχνικές. Οι πιο διαδεδομένες παρουσιάζονται παρακάτω:

- Μέθοδος της απλής συνένωσης (Single linkage method): Η μέθοδος της απλής συνένωσης είναι η πιο κλασική μέθοδος. Ανάμεσα σε δυο ομάδες διαλέγουμε την μικρότερη απόσταση που προκύπτει από ένα ζευγάρι σημείων των ομάδων.
- Μέθοδος της πλήρους συνένωσης (Complete linkage method): Αυτή την φορά επιλέγεται το ζευγάρι σημείων με την μεγαλύτερη απόσταση.
- Μέθοδος των μέσων (Weighted linkage method): Για την μέθοδο των μέσων υπολογίζεται η απόσταση όλων των ζευγαριών των συστάδων και υπολογίζουμε την μέση τιμή τους.
- Μέθοδος των κέντρων βάρους (Centroid method): Στην μέθοδο των κέντρων βάρους υπολογίζεται το κέντρο βάρους κάθε ομάδας.

Αν και έχει το βασικό πλεονέκτημα, ότι δεν καθορίζεται ο αριθμός των υποομάδων εξαρχής, η ιεραρχική συσταδοποίηση απαιτεί πολύ χρόνο υπολογιστικά. Επιπλέον η επιλογή του τρόπου συνένωσης επηρεάζει σε μεγάλο βαθμό την τελική ομαδοποίηση, για αυτό και χρειάζεται μεγάλη προσοχή στην επιλογή της.

4.6 Ανάλυση κυρίων συνιστωσών (Principal component analysis)

Πολύ συχνό φαινόμενο στα πραγματικά δεδομένα των basketball analytics είναι η ύπαρξη μεγάλου αριθμού χαρακτηριστικών. Πολλά από αυτά μπορεί να μην προσφέρουν ιδιαίτερες πληροφορίες, ενώ μπορεί να προκύπτουν και άλλα ζητήματα όπως της πολυσυγγραμικότητας. Η αναλυση κυρίων συνιστωσών (PCA) είναι ένα πολύ ισχυρό εργαλείο που μπορεί να βοηθήσει στην αντιμετώπιση αυτών των προβλημάτων, μετατρέποντάς τα χαρακτηριστικά σε ασυσχέτιστες μεταβλητές με μικρότερες διαστάσεις.

Στα πλαίσια της εργασίας ο Yin. (2014) χρησιμοποίησαν PCA σε δεδομένα του NBA, προτού κάνει ανάλυση παραγόντων για την ανάλυση της ικανότητας των αθλητών . Από τις 10 μεταβλητές που είχε αρχικά, κατέληξε σε 3 κύριες συνιστώσες.

	Component		
	1	2	3
Playing time (minute)	.979	-.058	-.083
Score	.947	-.161	.143
Steal	.927	-.153	-.012
Number of faults	.925	-.241	.197
Games played	.867	-.009	-.105
Rebound	.817	.252	-.497
Block shot	.765	.312	-.505
Assist	.732	-.456	.408
Field-goal percentage	.298	.795	.309
Free throw percentage	.466	.501	.572
Extraction method: Principal component analysis.			
a. Already extracted three components.			

Πίνακας 4.3. Απεικόνιση των 3 κυρίων συνιστωσών. Πηγή: Yin (2014)

Η ανάλυση κυρίων συνιστωσών όπως φαίνεται και παραπάνω απεικονίζει τα δεδομένα σε χαμηλότερες διαστάσεις. Αυτό γίνεται μέσω των κυρίων συνιστωσών, οι οποίες είναι γραμμικοί συνδυασμοί των αρχικών μεταβλητών. Κάθε κύρια συνιστώσα αποτελείται από n (αριθμός παρατηρήσεων) z σκορ, οπού για την πρώτη συνιστώσα το σκορ είναι:

$$z_{i1} = \varphi_{11}x_{i1} + \varphi_{21}x_{i2} + .. + \varphi_{p1}x_{ip}$$

Για τους συντελεστές στις κύριες συνιστώσες (φ_{ij}) ισχύει:

$$\sum_{j=1}^{p} {\varphi_{ij}}^2 = 1$$

Επιπλέον όλες οι μεταβλητές μετασχηματίζονται, έτσι ώστε να έχουν μέση τιμή 0.

Με δεδομένο τους παραπάνω περιορισμούς ο αλγόριθμος αναζητά σε κάθε κυριά συνιστώσα του συντελεστές που μεγιστοποιούν το άθροισμα:

$$\frac{1}{n}\sum_{\iota=1}^{\nu}\left(\sum_{j=1}^{p}\varphi_{j1}x_{ij}\right)^2$$

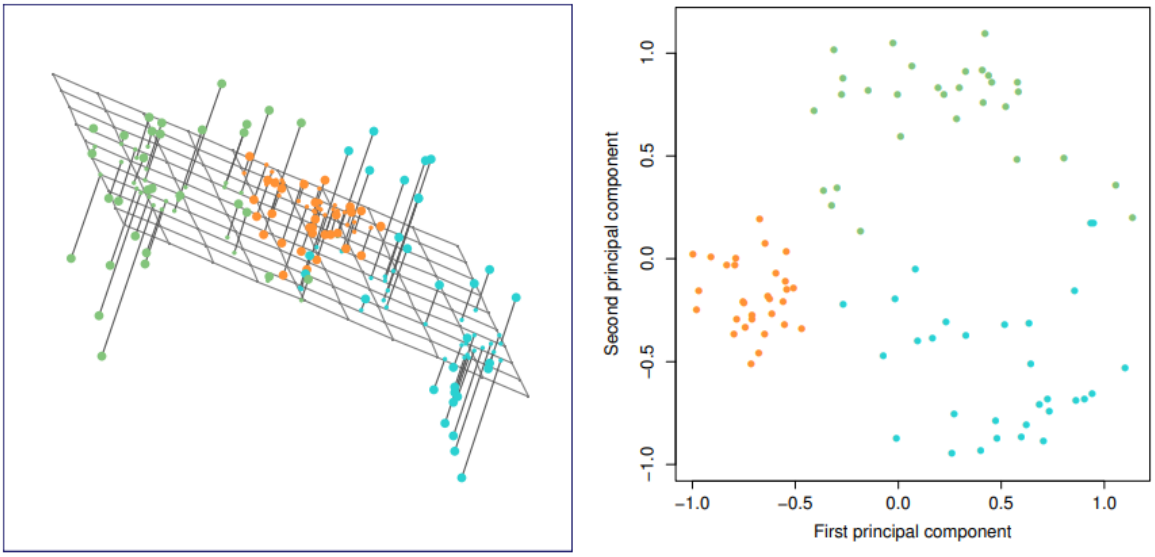
Ο στόχος είναι η μεγιστοποίηση της συνολικής διακύμανσης της πρώτης κύριας συνιστώσας. Για την λύση του προβλήματος βρίσκουμε τις ιδιοτιμές και τα ιδιοδιανύσματα του αθροίσματος. Το ιδιοδιάνυσμα της μεγαλύτερης ιδιοτιμής είναι αυτό που οδηγεί στην μεγιστοποίηση.

Για την εύρεση της δεύτερης κύριας συνιστώσας η διαδικασία που ακολουθείται είναι ίδια, με επιπλέον περιορισμό ότι το ιδιοδιάνυσμα της πρέπει να είναι κάθετο στο ιδιοδιάνυσμα της πρώτης. Αυτός ο περιορισμός εξασφαλίζει ότι οι δυο συνιστώσες θα είναι ασυσχέτιστες.

Η ανάλυση κυρίων συνιστωσών, όπως ειπώθηκε ήδη, είναι ο μετασχηματισμός των χαρακτηριστικών σε νέα χαρακτηριστικά που κρατάνε όσο περισσότερη πληροφορία γίνεται σε ένα χώρο μειωμένων διαστάσεων. Για αυτόν τον λόγο ελέγχουμε το ποσοστό της διακύμανσης που εξηγεί η κάθε συνιστώσα. Για τον υπολογισμό του ποσοστού παίρνουμε τον λόγο της διακύμανσης της συνιστώσας που ελέγχουμε (το άθροισμα που μεγιστοποιήθηκε παραπάνω) και της συνολικής διακύμανσης των δεδομένων. Δηλαδή:

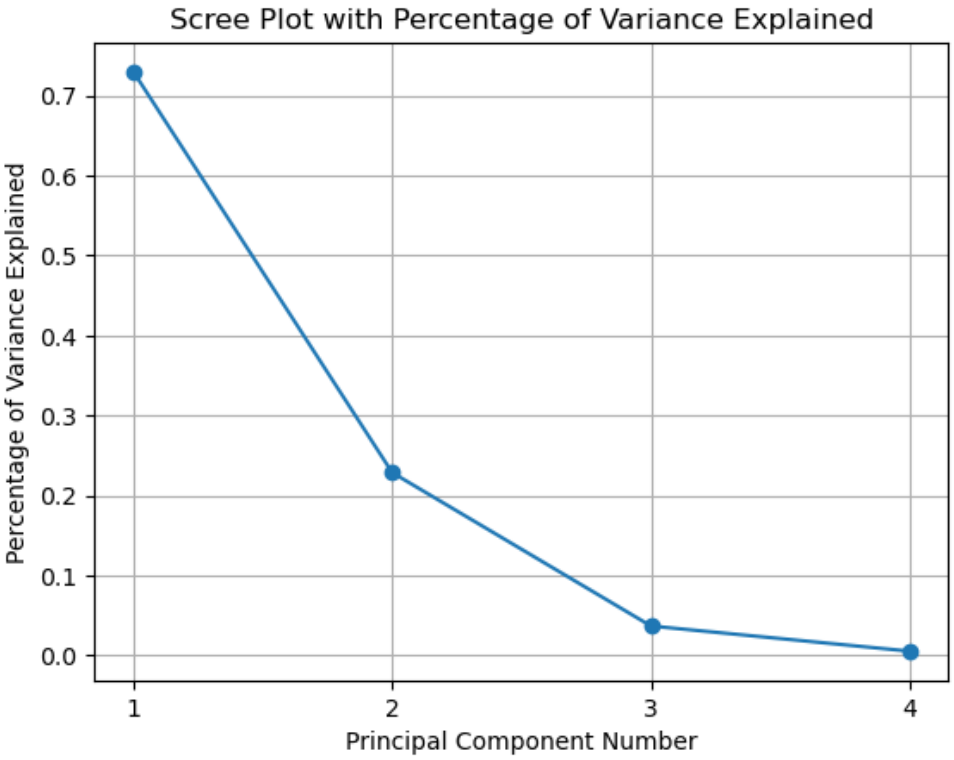
$$\frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\varphi_{j1}x_{ij}\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{p}x_{ij}^{\ 2}}$$

Στα παρακάτω δύο σχήματα βλέπουμε το αποτέλεσμα της PCA. Ένα αρχικά τρισδιάστατο σύνολο δεδομένων καταλήγει σε δύο διαστάσεις.



Σχήμα 4.15. Αριστερά: Το αρχικό σύνολο δεδομένων. Δεξιά: Γραφική απεικόνιση των δεδομένων μετά την PCA Πηγή: James et al. (2013)

Ένα πρόβλημα που δημιουργείται από την PCA είναι η επιλογή του αριθμού των συνιστωσών. Μια ευρέως χρησιμοποιούμενη τεχνική είναι το scree plot. Για την κατασκευή του τοποθετούμε σε φθίνουσα σειρά μεγέθους τις συνιστώσες στον οριζόντιο άξονα και στον κάθετο το ποσοστό της διακύμανσης που εξηγούν. Από το γράφημα ελέγχουμε σε ποιο σημείο η γραμμή αρχίζει να γίνεται επίπεδη, δηλαδή σε ποιο σημείο το ποσοστό της διακύμανσης που εξηγείται είναι πολύ μικρό. Στο επόμενο σχήμα φαίνεται ότι το ποσοστό της διακύμανσης που εξηγείται από την πρώτη κύρια συνιστώσα είναι φυσιολογικά πολύ μεγάλο (72.9%). Η δεύτερη κύρια συνιστώσα εξηγεί πολύ μικρότερο ποσοστό (22.8%) και στην 3η κύρια συνιστώσα φαίνεται η γραμμή να γίνεται επίπεδη, διότι το ποσοστό που εξηγείται είναι μόλις 3.6%.



Σχήμα 4.16. Ενδεικτική απεικόνιση ενός scree plot

4.7 Μέθοδοι επιλογής και αξιολόγησης μοντέλων (Model evaluation methods)

Όπως είδαμε παραπάνω, υπάρχουν αρκετά διαφορετικά μοντέλα που μπορούν να χρησιμοποιηθούν στην περίπτωση της παλινδρόμησης και της ταξινόμησης. Για αυτό τον λόγο χρειάζονται τεχνικές που βοηθάνε στην επιλογή μοντέλου και θα μετράνε την απόδοση των αλγορίθμων. Αφού επιλέξουμε κάποιο μοντέλο, ελέγχουμε πόσο καλά γενικεύεται σε δεδομένα ελέγχου. Οι μέθοδοι επιλογής και αξιολόγησης μοντέλων είναι ένα πολύ χρήσιμο εργαλείο για την μηχανική μάθηση, που βοηθάει στην επιλογή του βέλτιστου μοντέλου με βάση το εκάστοτε πρόβλημα.

4.7.1 Μέθοδοι επιλογής και αξιολόγησης μοντέλων ταξινόμησης

Πίνακας σύγχυσης (Confusion matrix)

Ο πίνακας σύγχυσης παρέχει μια σύνοψη όλων των ενδεχομένων που προκύπτουν από ένα μοντέλο ταξινόμησης. Τα ενδεχόμενα αυτά είναι:

- Αληθώς θετικό (True positive- TP): Είναι ο αριθμός των περιπτώσεων που έχουν προβλεφθεί σωστά από το μοντέλο.
- Ψευδώς θετικό (False positive- FP): Είναι ο αριθμός των περιπτώσεων που έχουν προβλεφθεί ως θετικές, ενώ στην πραγματικότητα δεν είναι.

- Αληθώς αρνητικό (True negative- TN): Είναι ο αριθμός των περιπτώσεων που έχουν προβλεφθεί σωστά ως αρνητικές.
- Ψευδώς αρνητικό (False negative- FN): Είναι ο αριθμός των περιπτώσεων που έχουν προβλεφθεί λανθασμένα ως αρνητικές.

		Actual values	
		+	-
Predicted values	+	True positive(TP)	False positive(FP)
	-	False negative(FN)	True negative (TN)

Πίνακας 4.4 Πίνακας σύγχυσης (Πηγή: Valero- Carreras et al.(2023))

Με βάση τα αποτελέσματα του πίνακα μπορούμε να χρησιμοποιήσουμε διάφορες μετρικές (metrics), για να αξιολογήσουμε τα μοντέλα.

- Ορθότητα (Accuracy): Είναι το ποσοστό των ορθά ταξινομημένων περιπτώσεων του μοντέλου. Υπολογίζεται ως: $\frac{TP+TN}{TP+TN+FP+FN}$
- Ακρίβεια (Precision): Η ακρίβεια είναι το ποσοστό των περιπτώσεων που προβλέφθηκαν σωστά ως θετικές. Υπολογίζεται ως: $\frac{TP}{TP+FP}$
- Ανάκληση (Recall): Η ανάκληση είναι το ποσοστό των πραγματικά θετικών περιπτώσεων που προβλέφθηκαν σωστά. Υπολογίζεται ως: $\frac{TP}{TP+FN}$
- F1 σκορ (F1 score): Είναι ο αρμονικός μέσος της ακρίβειας και της ανάκλησης. Υπολογίζεται ως: $\frac{2\times Ακριβεια \times Ανάκληση}{Ακριβεια+Ανάκληση}$

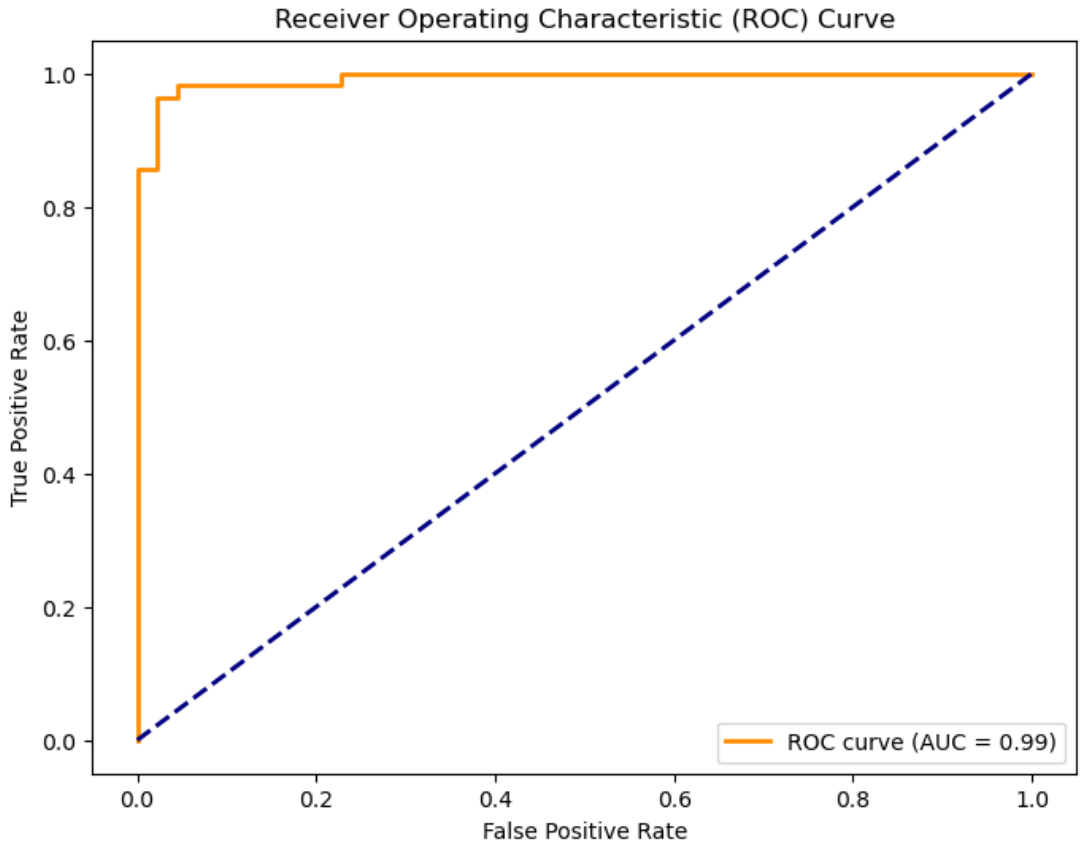
Η επιλογή των μετρικών εξαρτάται κάθε φορά από την φύση των δεδομένων και του προβλήματος. Για παράδειγμα η ορθότητα δεν είναι καλή επιλογή όταν οι τάξεις δεν είναι ισορροπημένες.

Καμπύλη ROC (ROC curve)

Η καμπύλη ROC είναι μια γραφική αναπαράσταση που χρησιμοποιείται για την αξιολόγηση μοντέλων ταξινόμησης. Ο οριζόντιος άξονας είναι το ποσοστό των ψευδώς θετικών περιπτώσεων (1-ειδικότητα), ενώ ο κάθετος το ποσοστό των αληθώς θετικών περιπτώσεων (ευαισθησία),

Η ιδανική καμπύλη προκύπτει από μεγιστοποίηση της ευαισθησίας και της ειδικότητας (δηλαδή τιμές της καμπύλης που βρίσκονται στην πάνω αριστερά πλευρά του διαγράμματος). Όμως δεν είναι πάντα η ιδανική λύση. Παραδείγματα ιατρικής φύσης μπορεί ορισμένες φορές να θεωρούν την μεγιστοποίηση ενός ποσοστού από τα παραπάνω πιο σημαντική, με αποτέλεσμα οι τιμές να κινούνται η στην κορυφή της δεξιάς πλευράς ή στην βάση της αριστερής.

Για την σύγκριση μοντέλων και τιμών αποκοπής χρησιμοποιείται το εμβαδόν κάτω από την καμπύλη (area under curve - AUC). Όλοι οι ταξινομητές σε ανεξάρτητα δεδομένα γενίκευσης έχουν τουλάχιστον AUC ίσο με 0.5. Όσο μεγαλύτερο είναι το εμβαδόν κάτω από την καμπύλη, τόσο πιο αξιόπιστος είναι ο ταξινόμητής. Ένα παράδειγμα καμπύλης ROC φαίνεται στο σχήμα 4.17.



Σχήμα 4.17. Ενδεικτική απεικόνιση μιας καμπύλης ROC, με AUC=0.99

4.7.2 Μέθοδοι επιλογής και αξιολόγησης μοντέλων παλινδρόμησης

4.7.2.1 Συντελεστής προσδιορισμού R^2 (Coefficient of determination R^2)

Ο συντελεστής προσδιορισμού R^2 εκφράζει το ποσοστό της διακύμανσης της εξαρτημένης μεταβλητής που εξηγείται από τις ανεξάρτητες μεταβλητές. Λαμβάνει τιμές ανάμεσα στο διάστημα [0,1]. Για τον υπολογισμό του παίρνουμε τον λόγο του αθροίσματος τετράγωνων παλινδρόμησης προς το συνολικό άθροισμα τετράγωνων.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Μεγάλη τιμή του R^2 ισοδυναμεί με καλή προσαρμογή του μοντέλου στα δεδομένα.

Παρόλα αυτά πρέπει να λαμβάνουμε υπόψιν, ότι όσο προσθέτουμε μεταβλητές ο συντελεστής προσδιορισμού θα μεγαλώνει διαρκώς. Ένας τρόπος για την εύρεση του βέλτιστου μοντέλου είναι ο έλεγχος της μεταβολής του συντελεστή μέχρι το σημείο, που η αύξηση στην διακύμανση που επιφέρει ένα πιο περίπλοκο μοντέλο είναι σχεδόν αμελητέα.

Μια άλλη τακτική είναι η χρήση του τροποποιημένου συντελεστή προσδιορισμού R^2_{adj}. Ο τροποποιημένος συντελεστής προσδιορισμού που επηρεάζεται από το πλήθος των μεταβλητών.

$$R^2_{adj} = 1 - \frac{n-1}{n-k-1}(1 - R^2)$$

Όπου k είναι ο αριθμός των ανεξαρτήτων μεταβλητών στο μοντέλο.

4.7.2.2 Κριτήριο αθροίσματος τετράγωνων υπολοίπων και μέσο τετραγωνικό υπόλοιπο

Το άθροισμα τετράγωνων υπολοίπων (SSE) είναι μια αρκετά δημοφιλής εξίσωση που χρησιμοποιείται για την εκτίμηση μοντέλων. Η ποσότητα δίνει την τετραγωνική διαφορά της πραγματικής και της προβλεπόμενης τιμής.

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Ένα επιθυμητό αποτέλεσμα είναι η όσο γίνεται μικρότερη διαφορά των δυο τιμών και ως αποτέλεσμα προτιμάται το μοντέλο με την μικρότερη τιμή.

Ένα μειονέκτημα του SSE είναι ότι σε περιπτώσεις που ο αριθμός των παρατηρήσεων είναι πολύ μεγάλος, μπορεί να δίνει παραπλανητικά αποτελέσματα. Μια λύση για αυτό είναι να χρησιμοποιήσουμε το μέσο τετραγωνικό υπόλοιπο (MSE), που χρησιμοποιεί και τον αριθμό παρατηρήσεων.

$$\mathrm{M}SE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$$

Το κριτήριο επιλογής παραμένει ίδιο, δηλαδή προτιμούμε ένα μοντέλο με μικρότερο MSE.

4.7.2.3 Διασταυρούμενη επικύρωση (Cross- validation)

Η διασταυρούμενη επικύρωση είναι μια πολύ γνωστή μέθοδος επαναδειγματοληψίας και είναι ένα βασικό εργαλείο για την εκτίμηση ενός μοντέλου. Η βασική ιδέα της μεθόδου είναι να χωρίσουμε τα δεδομένα, αποκλείοντας ένα υποσύνολο των δεδομένων και εκπαιδεύοντας τα υπόλοιπα στο μοντέλο και στην συνέχεια ελέγχουμε την απόδοση του με το υποσύνολο που δεν χρησιμοποιήθηκε στην εκπαίδευση.

Διασταυρούμενη επικύρωση χωρίς διαχωρισμούς (Leave one out cross validation-LOOCV)

Για να κάνουμε την LOOCV, θα πρέπει να εκπαιδεύσουμε το μοντέλο με όλες τις παρατηρήσεις του συνόλου δεδομένων, εκτός από μια. Η μια παρατήρηση θα είναι το σύνολο ελέγχου (test set). Από την εκπαίδευση προκύπτει μια εκτίμηση του μοντέλου, την οποία συγκρίνουμε με την παρατήρηση που έχει μείνει εκτός (Με χρήση του MSE για ένα δείγμα). Όμως μια δοκιμή, αν και έχει μικρή μεροληψία, έχει μεγάλη διακύμανση. Επομένως επαναλαμβάνουμε την διαδικασία n φορές και υπολογίζουμε την μέση τιμή των μέσων τετραγωνικών υπολοίπων:

$$CV = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

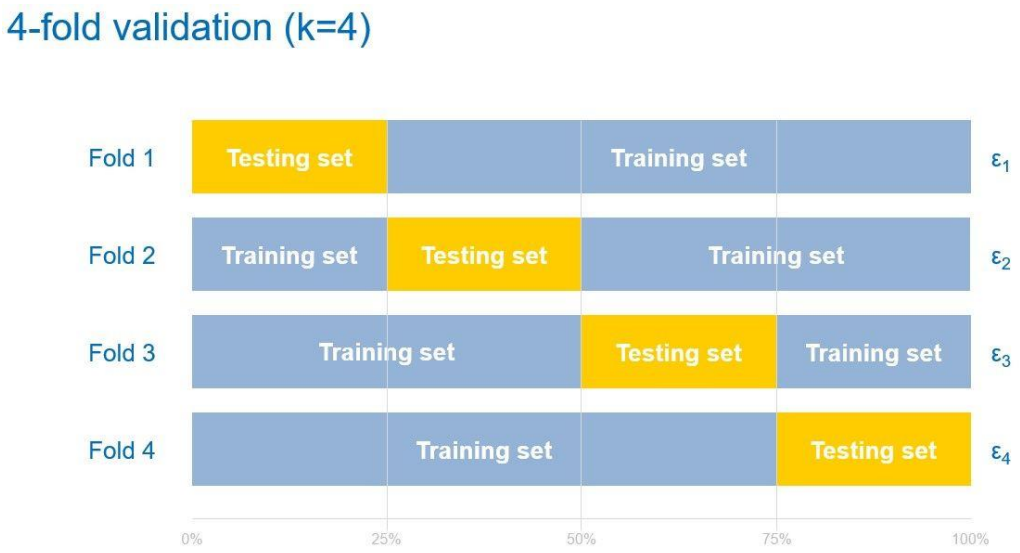
Όπου $MSE_i = (y_i - \hat{y}_i)^2$

Διασταυρούμενη επικύρωση k-πτυχών (k-fold cross validation)

Η διασταυρουμένη επικύρωση k-πτυχών αποτελεί μια γενίκευση της παραπάνω διαδικασίας. Συγκεκριμένα χωρίζουμε τα δεδομένα σε k πτυχές ίσου μεγέθους. Από τις k πτυχές κρατάμε μια εκτός για επικύρωση. Έπειτα εκπαιδεύουμε το μοντέλο χρησιμοποιώντας τις k-1 πτυχές και χρησιμοποιώντας το MSE συγκρίνουμε τις εκτιμώμενες τιμές με τις πραγματικές. Αυτό επαναλαμβάνεται k φορές (για όλες τι πτυχές) και η τελική εκτίμηση προκύπτει από τον μέσο όρο:

$$CV = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

Όπου $MSE_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$



Εικόνα 4.1. Ενδεικτική απεικόνιση ενός συνόλου δεδομένων που χωρίζεται σε 4 πτυχές

Η χρήση της k-fold, αν και έχει κόστος στην μεροληψία, έχει μικρότερη διακύμανση και δίνει πιο αξιόπιστα αποτελέσματα σε μεγάλα σύνολα δεδομένων. Επιπλέον είναι και υπολογιστικά πιο οικονομική. Η συνήθης επιλογή του k που συνήθως έχει μια ισορροπία ανάμεσα σε μεροληψία και διακύμανση είναι 5 ή 10.

Συμπερασματικά η διασταυρουμένη επικύρωση κάνοντας λίγες υποθέσεις για το μοντέλο, παρέχει αξιόπιστα αποτελέσματα και για αυτό είναι ιδιαίτερα δημοφιλής.

4.8 Προεπεξεργασία δεδομένων (Data Pre-processing)

Η προεπεξεργασία δεδομένων αποτελεί ένα βασικό βήμα που εφαρμόζεται πριν ξεκινήσει η χρήση των μοντέλων που έχουν αναφερθεί μέχρι στιγμής.

Τα πραγματικά δεδομένα σε πολλές περιπτώσεις μπορεί να περιέχουν ‘θόρυβο’, δηλαδή ελαττωματικά δεδομένα που οδηγούν σε μια προβληματική ανάλυση. Τέτοιου είδους προβλήματα είναι χαρακτηριστικά η ύπαρξη ακραίων ή και η πλήρης έλλειψη τιμών. Για αυτό τον λόγο υπάρχουν τεχνικές μετασχηματισμού των δεδομένων, για να είναι λειτουργικά.

4.8.1 Διαχείριση ελλειπουσών τιμών(Handling missing values)

Η έλλειψη τιμών είναι από τα συχνότερα προβλήματα που έχουν τα ακατέργαστα δεδομένα. Είναι σαφές ότι δεν μπορούν να διενεργηθούν αναλύσεις όταν υπάρχουν κενά στα δεδομένα.

Μια λύση σε αυτό το πρόβλημα είναι η εξάλειψη των σειρών που περιέχουν ελλείπουσες τιμές ή η εξάλειψη μιας στήλης αν περιέχει μεγάλο αριθμό ελλειπουσών τιμών. Με την εξάλειψη τους το σύνολο δεδομένων έχει λίγο μικρότερο αριθμό διαστάσεων, αλλά είναι λειτουργικό.

Μια άλλη τεχνική είναι η συμπλήρωση των ελλειπουσών τιμών, με μια σταθερή τιμή που ορίζεται από τον χρήστη. Συχνές επιλογές είναι η χρήση της μέσης τιμής, της διαμέσου ή και της επικρατούσας τιμής της στήλης.

Μια επιπλέον τεχνική είναι η συμπλήρωση των ελλειπουσών τιμών με τυχαίες παρατηρήσεις από το υπάρχον δείγμα. Αρχικά χωρίζουμε το δείγμα σε 2 υποσύνολα, ένα που περιέχει όλες τις ελλείπουσες τιμές και ένα που δεν περιέχει καμία. Έπειτα παίρνουμε τυχαία παρατηρήσεις από τα δυο υποσύνολα και σε κάθε περίπτωση που υπάρχουν ελλείπουσες τιμές, συμπληρώνουμε με μια τυχαία τιμή του πλήρες υποσυνόλου.

4.8.2 Κωδικοποίηση κατηγορικών μεταβλητών (Label Encoding)

Ένα συχνό πρόβλημα που έχουν τα ακατέργαστα δεδομένα είναι κατηγορικές μεταβλητές που έχουν ετικέτες αντί για τιμές. Για να γίνουν οι αναλύσεις, θα πρέπει να γίνει μετατροπή των ετικετών σε αριθμητική μορφή.

Η τεχνική αυτή λέγεται κωδικοποίηση κατηγορικών μεταβλητών. Στην κωδικοποίηση αντιμετωπίζουμε κάθε μεταβλητή με ετικέτες ξεχωριστά και εντοπίζουμε το πλήθος των ξεχωριστών ετικετών. Για κάθε ετικέτα αναθέτουμε έναν αριθμό ξεκινώντας συνήθως από το 0 ή 1. Η διαδικασία αυτή επαναλαμβάνεται για όλες τις κατηγορικές μεταβλητές.

Παρακάτω ένας ενδεικτικός πίνακας (4.5), με μια στήλη που περιέχει ετικέτες (‘City’) και χρησιμοποιώντας κωδικοποίηση, αναθέτουμε τιμές σε κάθε ετικέτα.

City	CityEncoded
New York	2
Paris	1
Tokyo	3
Paris	1
New York	2

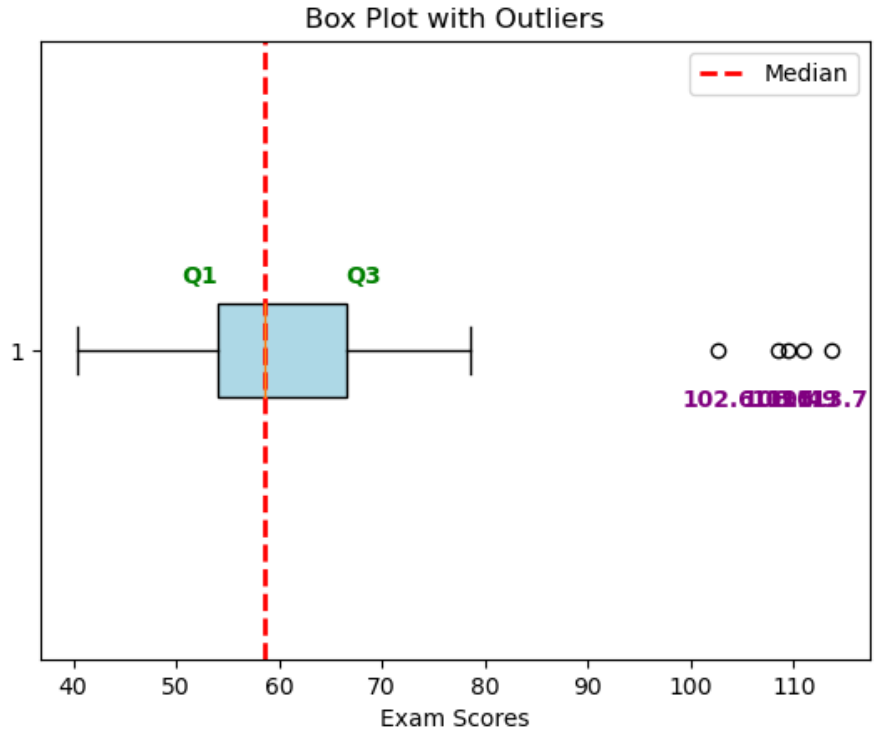
Πίνακας 4.5 Ενδεικτικό παράδειγμα κωδικοποίησης κατηγορικών μεταβλητών

4.8.3 Εντοπισμός ακραίων τιμών (Outlier detection)

Ως ακραίες τιμές σε ένα σύνολο δεδομένων θεωρούμε τις παρατηρήσεις που αποκλίνουν σημαντικά από την πλειοψηφία των δεδομένων. Οι ακραίες τιμές πρέπει να περιορίζονται, καθώς επηρεάζουν διάφορα στατιστικά μέτρα, όπως η μέση τιμή, και συνεπώς οδηγούν σε παραπλανητικά συμπεράσματα.

Θηκόγραμμα

Μια μέθοδος εντοπισμού των ακραίων τιμών είναι μέσω θηκογράμματος. Ένα θηκόγραμμα αποτελεί ένα χρήσιμο εργαλείο εντοπισμού ακραίων τιμών, διότι απεικονίζει ξεκάθαρα τις ακραίες τιμές. Ένα ενδεικτικό θηκόγραμμα παρουσιάζεται γραφικά στο σχήμα 4.19.



Σχήμα 4.19. Ενδεικτικό θηκόγραμμα

Το παραπάνω σχήμα αναπαριστά ένα απλό παράδειγμα θηκογράμματος. Το θηκόγραμμα αναπαριστά όλα τα βασικά χαρακτηριστικά μιας μεταβλητής. Το χρωματισμένο διάστημα καθορίζεται από το ενδοτεταρτημοριακό εύρος, δηλαδή την απόσταση του 1ου και 3ου τεταρτημόριου, ενώ η διακεκομμένη κόκκινη γραμμή είναι η διάμεσος. Ακόμη τα άκρα του σχήματος αντιπροσωπεύουν την ελάχιστη και μέγιστη τιμή. Υποψήφια ακραία τιμή σε ένα ιστόγραμμα θεωρούμε μια παρατήρηση x_i όταν ισχύει ένα από τα παρακάτω:

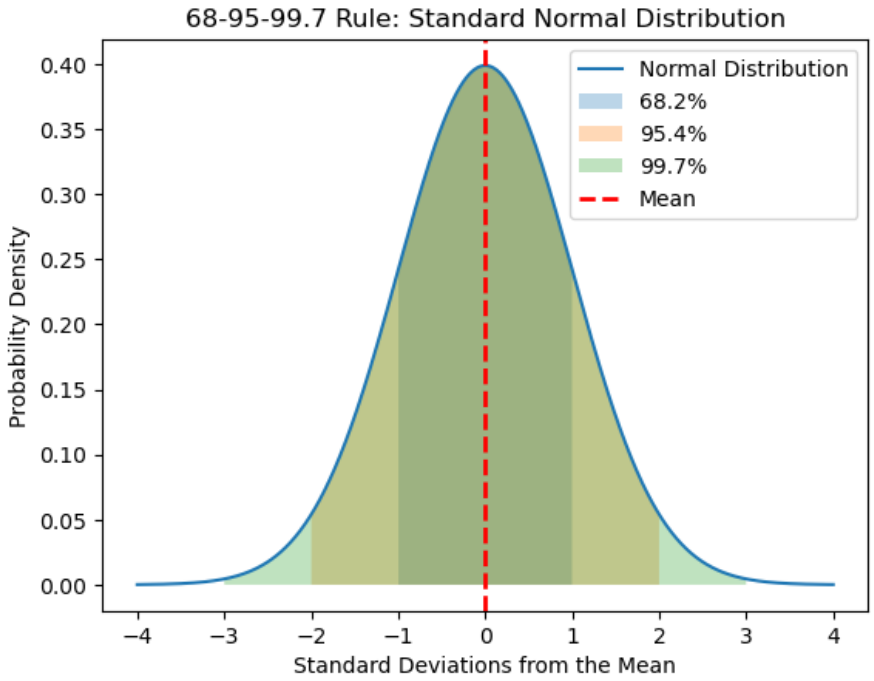
- $x_i < Q_1 - 1.5 \times IQR$
- $x_i > Q_3 + 1.5 \times IQR$

Χρήση της τυπικής απόκλισης

Υποθέτοντας ότι τα δεδομένα ακολουθούν κανονική κατανομή, ισχύει ότι:

- Το 68.2% των παρατηρήσεων βρίσκεται στο εύρος $\mu \pm \sigma$
- Το 95.4% των παρατηρήσεων βρίσκεται στο εύρος $\mu \pm 2\sigma$
- Το 99.7% των παρατηρήσεων βρίσκεται στο εύρος $\mu \pm 3\sigma$

Ο εμπειρικός κανόνας είναι ένα χρήσιμο εργαλείο, διότι οι παρατηρήσεις που βρίσκονται πέρα από το εύρος $\mu \pm 3\sigma$ μπορουν να θεωρηθουν ακραιες τιμες.



Σχήμα 4.20. Διαγραμματική απεικόνιση του κανόνα 68-95-99.7

Η χρήση της τυπικής απόκλισης είναι αρκετά δημοφιλής μέθοδος, που για να λειτουργήσει όμως χρειάζεται τα δεδομένα να ακολουθούν κανονική κατανομή.

Δάση απομόνωσης (Isolation forests)

Έκτος ων κλασικών μεθόδων, υπάρχουν και αλγοριθμικές τεχνικές για τον εντοπισμό των ακραίων τιμών.

Τα δάση απομόνωσης είναι μια διαδικασία παρόμοια με τα τυχαία δάση που μελετήθηκαν παραπάνω και έχουν ως σκοπό την απομόνωση των ακραίων τιμών.

Ο αλγόριθμος αρχικά διαλέγει τυχαία μια μεταβλητή από τα δεδομένα και στην συνέχεια διαλέγει ένα τυχαίο κατώφλι, για να ταξινομήσει τις παρατηρήσεις της μεταβλητής. Ανάλογα με το αν βρίσκεται πάνω ή κάτω από το κατώφλι η παρατήρηση ταξινομείται και έτσι σχηματίζεται ένας κόμβος. Ο διαχωρισμός συνεχίζεται μέχρι κάποια παρατήρηση να απομονωθεί ή μέχρι το δέντρο να φτάσει σε έναν συγκεκριμένο αριθμό κόμβων που έχει ορίσει ο χρήστης. Στην συνέχεια σχηματίζουμε ένα άλλο δέντρο μέχρι να απομονωθεί η επόμενη παρατήρηση και αυτό επαναλαμβάνεται για όλες τις παρατηρήσεις και όλες τις μεταβλητές.

Ο αλγόριθμος αναθέτει κάποια σκορ με βάση τον αριθμό των κόμβων που χρειάστηκε μια παρατήρηση για να απομονωθεί. Με βάση το σκορ μπορούμε να υποθέσουμε αν υπάρχουν ακραίες τιμές. Για παράδειγμα ένα χαμηλό σκορ είναι μια ισχυρή ένδειξη ότι η παρατήρηση είναι ακραία τιμή.

Εφόσον παρατηρηθούν ακραίες τιμές, χρειάζονται προσεκτική παρατήρηση, προτού ληφθούν δράσεις. Η συνήθης διαδικασία είναι η διαγραφή τους από το σύνολο δεδομένων. Όμως αυτό πρέπει να γίνεται μόνο αν η ύπαρξη της ακραίας τιμής οφείλεται σε λάθος εισαγωγή δεδομένων ή κάποιο άλλο σφάλμα.

Για την εξάλειψη ή μη μιας ακραίας τιμής χρειάζεται γνώση του τομέα που γίνεται η μελέτη. Ένα παράδειγμα που δίνει ευρύτερη κατανόηση στην παραπάνω πρόταση: Ο Wilt Chamberlain έχει το ρεκόρ πόντων στο NBA, σκοράροντας 100 πόντους το 1962. Όπως είναι προφανές αυτή η τιμή είναι ακραία αν αναλογιστούμε ότι οι καλύτεροι παίκτες έχουν μέσους ορούς πόντων που κυμαίνονται από 25-35 ανά παιχνίδι. Όμως είναι ένα ιστορικό ρεκόρ και δεν υπάρχει κάποιο σφάλμα που να δικαιολογεί την διαγραφή αυτής της παρατήρησης. Είναι όμως επίσης γνωστό ότι ο ανταγωνισμός 60 χρονιά μετά έχει ανεβεί δραστικά, επομένως το να επαναλάβει κάποιος αυτό το ρεκόρ είναι μια αρκετά σπάνια περίπτωση και ενδεχομένως η παρατήρηση να μην προσφέρει κάτι, αν γίνεται μελέτη του σύγχρονου τρόπου παιχνιδιού.

Έτσι συμπεραίνουμε ότι η διαχείριση των ακραίων τιμών διαφέρει ανάλογα με το πρόβλημα και το στόχο κάποιας μελέτης.

Εντοπισμός ακραίων τιμών για μεγαλύτερες διαστάσεις

Στα σύνολα δεδομένων με μεγάλες διαστάσεις, ο εντοπισμός ακραίων τιμών είναι μια πιο περίπλοκη διαδικασία. Οι ακραίες τιμές σε αυτή την περίπτωση χωρίζονται σε δυο είδη (Talagala et al., 2021):

- Στις global ακραίες τιμές, που επηρεάζουν την κατανομή του συνόλου δεδομένων.
- Στις τοπικές ακραίες τιμές που διαφέρουν αρκετά από ένα υποσύνολο ή μια συστάδα δεδομένων.

Επιπλέον η ενδεχόμενη πολυσυγγραμικότητα δυσχεραίνει τον εντοπισμό τους.

Μια γνωστή προσέγγιση είναι ο υπολογισμός της απόστασης Mahalanobis για κάθε παρατήρηση:

$$D(X_i) = \sqrt{(X_i - \mu)^T S^{-1} (X_i - \mu)}$$

Όπου το S είναι ο πίνακας διακύμανσης συνδιακύμανσης.

Στην συνέχεια συγκρίνουμε το $D(X_i)$ που προκύπτει με το σημείο αποκοπής $\sqrt{x^2_{p,0.975}}$,όπου $x^2_{p,0.975}$ είναι το ποσοστιαίο σημείο της κατανομής X^2 με p βαθμούς ελευθερίας. Τα $D(X_i)$ που βρίσκονται πάνω από το σημείο αποκοπής θεωρούνται ακραίες τιμές

4.8.4 Κλιμάκωση δεδομένων (Data scaling)

Η κλιμάκωση δεδομένων είναι άλλη μια απαραίτητη διαδικασία προεπεξεργασίας δεδομένων. Πολλές μεταβλητές σε ένα σύνολο δεδομένων συχνά έχουν πολύ μεγαλύτερη

κλίμακα, δηλαδή κάποιες λαμβάνουν πολύ μεγαλύτερες τιμές από άλλες. Συνέπεια αυτού είναι μια αναξιόπιστη ανάλυση, διότι κάποιες μεταβλητές επηρεάζουν περισσότερο ένα μοντέλο.

Η κλιμάκωση δεδομένων μετασχηματίζει τα δεδομένα, για να εξασφαλιστεί η ίση μεταχείριση των μεταβλητών.

Παρακάτω παρουσιάζονται οι πιο γνωστές μέθοδοι κλιμάκωσης:

- Κλιμάκωση min-max: Η κλιμάκωση min-max είναι μια γνωστή μέθοδος κλιμάκωσης, που μετασχηματίζει τα δεδομένα σε ένα συγκεκριμένο εύρος (συνήθως στο [0,1]). Για την εφαρμογή του χρειάζεται η εύρεση του ελάχιστου και μέγιστου σημείου κάθε μεταβλητής.

$$x_{trans} = \frac{x - X_{min}}{X_{max} - X_{min}}$$

- Κανονικοποίηση Z-σκορ (Z-score normalization): Στην κανονικοποίηση Z-σκορ τα δεδομένα μετασχηματίζονται, έτσι ώστε να έχουν μέση τιμή 0 και τυπική απόκλιση 1. Είναι μια ιδιαίτερα χρήσιμη μέθοδος, όταν υπάρχουνε πολλά διαφορετικά εύρη τιμών.

$$x_{trans} = \frac{x - mean(X)}{sd(X)}$$

- Ανθεκτική κλιμάκωση (Robust scaling): Η ανθεκτική κλιμάκωση είναι μια μέθοδος κλιμάκωσης δεδομένων που είναι ανθεκτική στις ακραίες τιμές. Αυτό συμβαίνει διότι, μετασχηματίζει τις τιμές με την χρήση της διαμέσου και του ενδοτεταρτημοριακού εύρους. Επιπλέον μπορεί να χρησιμοποιηθεί σε δεδομένα που δεν υπάρχει απαραίτητα η υπόθεση της κανονικότητας.

$$x_{trans} = \frac{x - median(X)}{IQR(X)}$$

4.9 Νευρωνικά δίκτυα (Neural networks)

Στο τελευταίο κομμάτι του κεφαλαίου θα γίνει μια ξεχωριστή αναφορά στα νευρωνικά δίκτυα. Τα νευρωνικά δίκτυα είναι μοντέλα μηχανικής μάθησης, που με την αύξηση της υπολογιστικής δύναμης, έχουν γίνει αρκετά δημοφιλή. Χρησιμοποιούνται σε προβλήματα παλινδρόμησης αλλά και ταξινόμησης. Τα μοντέλα των νευρωνικών δικτύων είναι πιο περίπλοκα από τους αλγόριθμους που μελετήθηκαν παραπάνω, αλλά συνήθως έχουν και μεγαλύτερη προβλεπτική ικανότητα. Ο σχεδιασμός των νευρωνικών δικτύων είναι εμπνευσμένος από την λειτουργία του ανθρώπινου εγκέφαλου. Τα μοντέλα λαμβάνουν μεγάλο όγκο πληροφοριών και μέσω νευρώνων που συνδέονται μεταξύ τους καταλήγουν σε πρόβλεψη. Οι τύποι που θα μελετηθούν παραπάνω είναι τα:

- Νευρωνικά δίκτυα πρόσθιας τροφοδότησης
- Συνελικτικά νευρωνικά δίκτυα

- Επαναλαμβανόμενα νευρωνικά δίκτυα
- Πιθανοτικά νευρωνικά δίκτυα

4.9.1 Νευρωνικά Δίκτυα πρόσθιας τροφοδότησης (Feedforward Neural Networks)

Πρόκειται για τον πιο βασικό τύπο νευρωνικών δικτυών. Τα νευρωνικά δίκτυα πρόσθιας τροφοδότησης χαρακτηρίζονται για την μια κατεύθυνση που ακολουθούν τα δεδομένα. Συγκεκριμένα οι μεταβλητές εισάγονται αρχικά στο στρώμα εισόδου (input layer). Στην συνέχεια μεταφέρονται διαδοχικά στα κρυφά στρωματά (hidden layers) μέχρι να καταλήξουν στο στρώμα εξόδου (output layer). Κάθε κρυφό στρώμα χρησιμοποιεί τις πληροφορίες από το προηγούμενο, μετασχηματίζοντας έτσι κάθε φορά τα βάρη.

Προχωρώντας στο κομμάτι των τύπων, θα αναφερθούμε στην περίπτωση που υπάρχουν δυο κρυφά στρώματα.

Ο τύπος του πρώτου κρυφού στρώματος:

$$h_k^{(1)}(X) = g(b^{(1)} + \sum_{j=1}^{p} w_{kj}^{(1)} X_j)$$

Για k=1,..,K1. Το $b^{(1)}$ είναι η μεροληψία του πρώτου κρυφού στρώματος. Με το g() να είναι μια μη γραμμική συνάρτηση ενεργοποίησης. Οι συναρτήσεις ενεργοποίησης είναι σημαντικές, διότι εισάγουν μη γραμμικότητα και δίνουν την δυνατότητα στα νευρωνικά μοντέλα να εντοπίζουν περίπλοκα μοτίβα. Παλαιότερα χρησιμοποιούταν κυρίως η σιγμοειδής συνάρτηση, αλλά πλέον η πιο δημοφιλής επιλογή είναι η συνάρτηση ενεργοποίησης ReLU, με τύπο:

$$g(z) = \max(0, z)$$

Τα $w_{k_j}^{(1)}$είναι τα βάρη των δεδομένων με τον δείκτη (1) να δηλώνει το πρώτο κρυφό στρώμα.

Για το δεύτερο κρυφό στρώμα υπολογίζουμε τον εξής τύπο:

$$h_l^{(2)}(X) = g(b^{(2)} + \sum_{j=1}^{K_1} w_{lj}^{(2)} h_k^{(1)}(X))$$

Με το l=1,…,K2. Από τον τύπο βλέπουμε πως συνδέονται τα δυο στρώματα μεταξύ τους, αφού το αποτέλεσμα του πρώτου χρησιμοποιείται για την εύρεση του επόμενου.

Το στρώμα εξόδου βρίσκεται ως εξής:

$$Z_m = \beta_{m0} + \sum_{l=1}^{K_2} \beta_{ml} h_l^{(2)}(X)$$

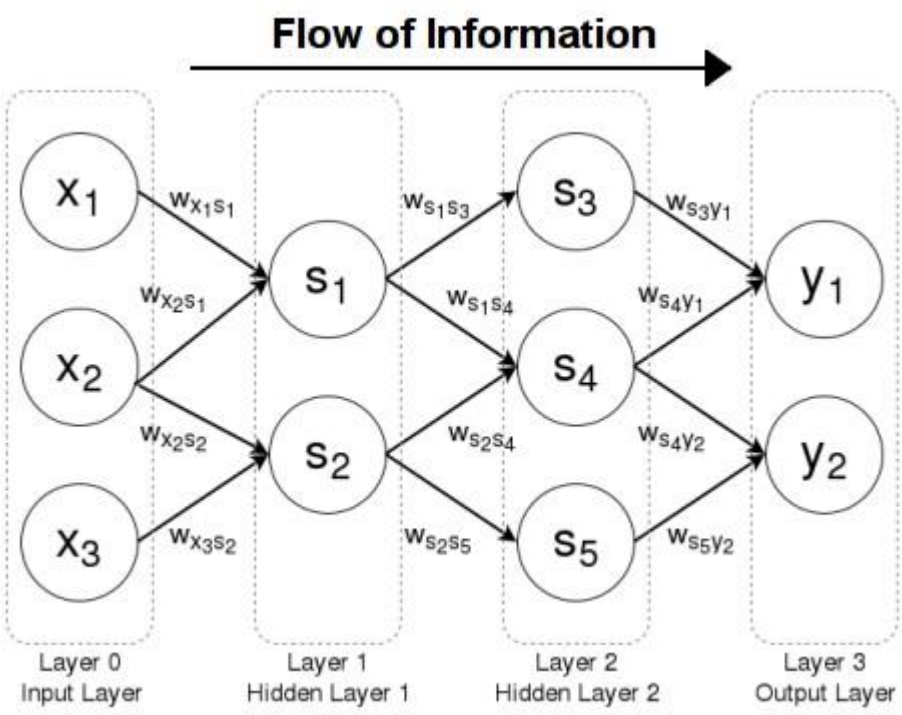
Όπου β_{ml} είναι οι συντελεστές του κρυφού στρώματος.

Σε προβλήματα ταξινόμησης, χρειάζεται επιπλέον να χρησιμοποιήσουμε την συνάρτηση ενεργοποίησης softmax με τύπο:

$$P(Y = m|X) = \frac{e^{Z_m}}{\sum_{l=0}^{k} e^{Z_l}}$$

Η συνάρτηση μετατρέπει τα αποτελέσματα του στρώματος εξόδου σε πιθανότητες που αθροίζουν στο 1. Η ταξινόμηση γίνεται στην τάξη με την μεγαλύτερη πιθανότητα.

Ένα νευρωνικό δίκτυο σαν αυτό που περιγραφικέ παραπάνω απεικονίζεται στο σχήμα 4.21.



Σχήμα 4.21 Ενδεικτικό νευρωνικό δίκτυο με δύο κρυφά στρώματα (Πηγή: Feedforward Neural Networks | Brilliant Math & Science Wiki)

4.9.2 Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks)

Τα συνελικτικά νευρωνικά δίκτυα (CNN) είναι ένα είδος νευρωνικών δικτύων, που χρησιμοποιείται κατά κύριο λόγο για την ανάλυση εικόνων. Η βασική ιδέα πίσω από τον αλγόριθμο είναι η ταξινόμηση εικόνων παρόμοια με τον τρόπο που τις ταξινομεί ένας ανθρώπινος εγκέφαλος. Πιο συγκεκριμένα τα CNN, μέσω της διαδικασίας που θα περιγράφει παρακάτω, αναγνωρίζουν μικρά χαρακτηριστικά σε μια εικόνα και συνδυάζοντας τα καταλήγουν στην αντίστοιχη πρόβλεψη.

Για την ανίχνευση των χαρακτηριστικών μιας εικόνας κατασκευάζονται πίνακες μικρών διαστάσεων που τονίζουν την περιοχή που πρέπει να εστιάσει ο αλγόριθμος. Αυτοί οι

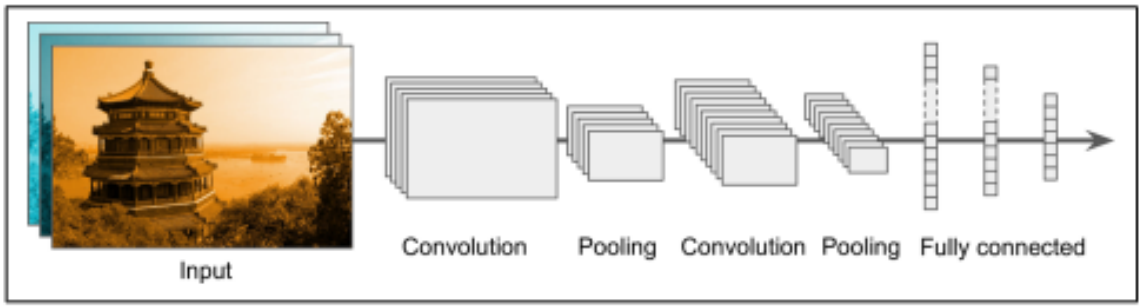
πίνακες λέγονται φίλτρα συμβολής (convolutional filters) και πολλαπλασιάζονται με υποσύνολα του πίνακα της αρχικής εικόνας. Με αυτό τον τρόπο τα φίλτρα συμβολής σαρώνουν όλη την εικόνα για την εύρεση χαρακτηριστικών ή μοτίβων.

Από την σάρωση των φίλτρων συμβολής στον αρχικό πίνακα, προκύπτουν πίνακες μικρότερων διαστάσεων που καλούνται συνεπτυγμένες εικόνες (convolved images). Μια συνεπτυγμένη εικόνα περιέχει μεγάλες τιμές στις περιοχές που μοιάζουν περισσότερο με τα φίλτρα συμβολής. Επιπλέον στην συνεπτυγμένη εικόνα εφαρμόζουμε μια συνάρτηση ενεργοποίησης (συνήθως την ReLU).

Ένα ακόμη βήμα που χρησιμοποιείται είναι η τεχνική pooling. Η τεχνική pooling μειώνει τις διαστάσεις του πίνακα και τον υπολογιστικό φόρτο (Géron, A. (2022)). Η πιο γνωστή τεχνική pooling είναι η max pooling, που κρατάει την μέγιστη τιμή από το υποσύνολο ενός πίνακα. Έτσι εκτός της μείωσης διαστάσεων, ο πίνακας διατηρεί την πληροφορία που έχει προκύψει από την συνεπτυγμένη εικόνα.

Η διαδικασία που αναλύθηκε παραπάνω, επαναλαμβάνεται μέχρι οι διαστάσεις να γίνουν αρκετά μικρές. Στην συνέχεια τοποθετούμε όλα τα στοιχεία των εικόνων σε ένα διάνυσμα μιας διάστασης και εφαρμόζουμε την συνάρτηση ενεργοποίησης softmax,για να δώσουμε πιθανότητες σε κάθε τάξη και να προχωρήσουμε στην ταξινόμηση.

Η διαδικασία απεικονίζεται ενδεικτικά στο σχήμα 4.22.



Σχήμα 4.22 Ενδεικτική απεικόνιση της διαδικασίας ενός συνελικτικού νευρωνικού δικτύου (Πηγή: Géron, A. (2022))

Για την καλύτερη γενίκευση του μοντέλου μια τεχνική είναι η αύξηση των δεδομένων (data augmentation). Η αύξηση των δεδομένων γίνεται στις ήδη υπάρχουσες εικόνες που εκπαιδεύουν το μοντέλο και ουσιαστικά αναπαράγουν τις ίδιες εικόνες με μικρές παραμορφώσεις όπως η περιστροφή και η μεγέθυνση, για την καλύτερη γενίκευση του μοντέλου σε άγνωστα δεδομένα.

4.9.3 Επαναλαμβανόμενα νευρωνικά δίκτυα (Recurrent Neural Networks)

Τα επαναλαμβανόμενα νευρωνικά δίκτυα (RNN) είναι άλλη μια ειδική κατηγορία νευρωνικών δικτύων, με χαρακτηριστικό την διαχείριση διαδοχικών δεδομένων. Τα RNN διαφέρουν από τα νευρωνικά δίκτυα που έχουν περιγράφει μέχρι τώρα, καθώς τα δεδομένα εισαγωγής είναι μια ακολουθία δεδομένων. Επίσης διαφέρουν στην δομή τους. Τα

νευρωνικά δίκτυα πρόσθιας τροφοδότησης ρέουν προς μια κατεύθυνση, δηλαδή το μοντέλο δέχεται τα δεδομένα εισαγωγής και καταλήγει στο στρώμα εξόδου. Για την δομή των επαναλαμβανόμενων νευρωνικών δικτύων παρατηρούμε ότι υπάρχουν συνδέσεις που γυρνάνε προς τα πίσω. Αυτό συμβαίνει γιατί το μοντέλο σε κάθε βήμα χρησιμοποιεί πέραν των δεδομένων εισαγωγής και τα αποτελέσματα του κρυφού στρώματος που προκύπτουν από το προηγούμενο βήμα. Με αυτό τον τρόπο οι νευρώνες αποκτούν μνήμη, κάτι που είναι απαραίτητο για τις προβλέψεις.

Πιο αναλυτικά για την δομή των RNN, το δίκτυο αποτελείται από τα στρώμα εισόδου, τα κρυφά στρωματά και το στρώμα εξόδου. Το στρώμα εισόδου περιέχει το διάνυσμα με τα διαδοχικά δεδομένα. Το κρυφό στρώμα είναι ένα είδος μνήμης που περιέχει το δίκτυο και αναβαθμίζεται σε κάθε νέα είσοδο δεδομένων. Η αναβάθμιση γίνεται μέσω του σταθμισμένου αθροίσματος του προηγούμενου κρυφού στρώματος και των δεδομένων του κάθε βήματος. Ακόμη στο κρυφό στρώμα χρησιμοποιείται και μια συνάρτηση ενεργοποίησης (συνήθως είναι η ReLU ή η συνάρτηση υπερβολικής εφαπτομένης). Το στρώμα εξόδου περιέχει τα αποτελέσματα του δικτύου.

Με τύπους τα παραπάνω εκφράζονται ως εξής:

Το κρυφό στρώμα στο βήμα t υπολογίζεται με τον παρακάτω τύπο:

$$h_t = g(b + W_X^T x_t + W_h^T h_{t-1})$$

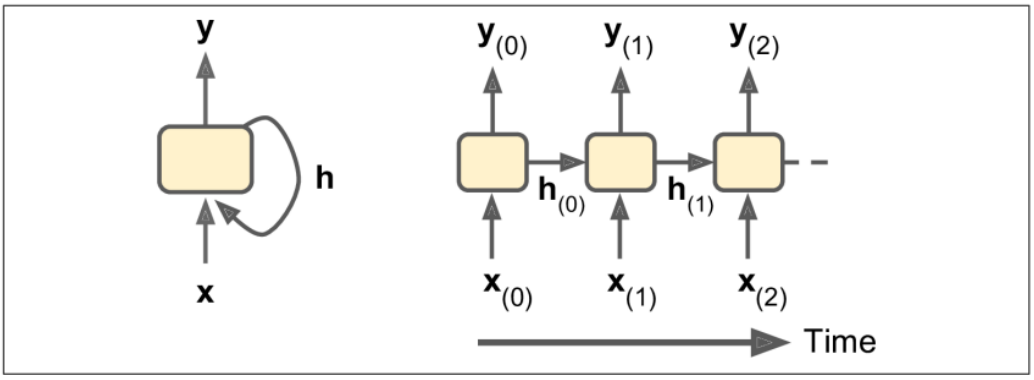
Όπου με b συμβολίζεται το διάνυσμα μεροληψίας. Ο όρος $W_X^T x_t$ αποτελείται από το τα δεδομένα εισαγωγής (x_t) και τα αντίστοιχα βάρη (W_X^T). Ο όρος $W_h^T h_{t-1}$ περιέχει την κρυφή κατάσταση στο προηγούμενο βήμα (h_{t-1}) και τα βάρη των αναδρομικών σχέσεων (W_h^T). Η συνάρτηση g() είναι η συνάρτηση ενεργοποίησης.

Για τον υπολογισμό του στρώματος εξόδου υπολογίζεται ο τύπος:

$$y_t = b_0 + Vh_t$$

Όπου b_0 είναι η μεροληψία, V ο πίνακας των συντελεστών του κρυφού στρώματος.

Επιπλέον όταν θέλουμε να κάνουμε ταξινόμηση χρησιμοποιούμε την συνάρτηση softmax, που αναφέρθηκε παραπάνω, για να μετατρέψουμε τις τιμές κάθε τάξης σε πιθανότητες. Ενδεικτικά παρατίθεται ένα επαναλαμβανόμενο νευρωνικό δίκτυο στο σχήμα 4.23.



Σχήμα 4.23 Ενδεικτική απεικόνιση ενός επαναλαμβανόμενου νευρωνικού δικτύου. (Πηγή: Géron, A. (2022))

Τα επαναλαμβανόμενα νευρωνικά δίκτυα είναι ένα πολύ χρήσιμο εργαλείο για την επεξεργασία διαδοχικών δεδομένων. Παρόλα αυτά μετά από έναν συγκεκριμένο αριθμό βημάτων δεν είναι τόσο αποτελεσματικά, διότι χάνουν μεγάλο κομμάτι της αρχικής τους μνήμης. Ένας τρόπος καταπολέμησης του προβλήματος είναι η χρήση των νευρωνικών δικτυών μακράς βραχυπρόθεσμης μνήμης (Long Short-Term Memory Networks). Τα δίκτυα μακράς βραχυπρόθεσμης μνήμης είναι σχεδιασμένα για να κρατάνε τα πιο σημαντικά δεδομένα εισαγωγής μακροπρόθεσμα και να τα χρησιμοποιούν οπότε χρειάζονται.

Τα RNN βρίσκουν πολλές εφαρμογές σε διάφορα πεδία όπως η μετάφραση κείμενων και η ανάλυση χρονοσειρών. Για την ανάλυση στον χώρο της καλαθοσφαίρισης μπορούν να χρησιμοποιηθούν για την πρόβλεψη μιας εύστοχης προσπάθειας.

4.9.4 Πιθανοτικά Νευρωνικά Δίκτυα (Probabilistic Neural Networks)

Τα πιθανοτικά νευρωνικά δίκτυα (PNN) είναι ένα είδος νευρωνικού δικτύου, που χρησιμοποιείται κυρίως για προβλήματα ταξινόμησης. Τα πιθανοτικά νευρωνικά δίκτυα μοντελοποιούν την σχέση των δεδομένων εισόδου και εξόδου μέσω πιθανοτήτων για την λήψη της τελικής απόφασης. Χαρακτηριστικό των PNN είναι η χρήση της μεθόδου του Parzen και στην συνέχεια ο κανόνας του Bayes.

Η μέθοδος Parzen Windows χρησιμοποιείται στα PNN για τον υπολογισμό της συνάρτησής πυκνότητας πιθανότητας. Η μέθοδος είναι μη παραμετρική και χρειάζεται τον ορισμό κάποιας συνάρτησης kernel. Η συνάρτηση kernel που χρησιμοποιείται κατά κύριο λόγο είναι η πολυμεταβλητή Γκαουσιανή συνάρτηση. Στην συνέχεια για την ταξινόμηση ενός καινούργιου σημείου, υπολογίζουμε το αποτέλεσμα της Γκαουσιανης συνάρτησης kernel για όλα τα δεδομένα εκπαίδευσης που ανήκουν στην εκάστοτε τάξη και η συνάρτηση πυκνότητας πιθανότητας προκύπτει από τον μέσο όρο. Η ταξινόμηση γίνεται στην τάξη με την μεγαλύτερη πιθανότητα.

Εκφράζοντας τα παραπάνω σε τύπους:

- Υπολογισμός της Γκαουσιανής συνάρτησης kernel:

$$f_{i,j} = \frac{1}{(2\pi)^{d/2}\sigma^d} exp^{(-\frac{(X-Xi,j)'(X-Xi,j)}{2\sigma^2})}$$

Με το d να είναι ο αριθμός των χαρακτηριστικών στα δεδομένα.

- Υπολογισμός του μέσου όρου.

$$P_k = \frac{1}{|C_k|}\sum_{j=1}^{|C_k|} f_{k,j}$$

όπου $|C_k|$είναι το μέγεθος ενός υποσυνόλου

- Ταξινόμηση του σημείου στην μεγαλύτερη εκ των υστέρων πιθανότητα.

$$argmax(P_k)$$

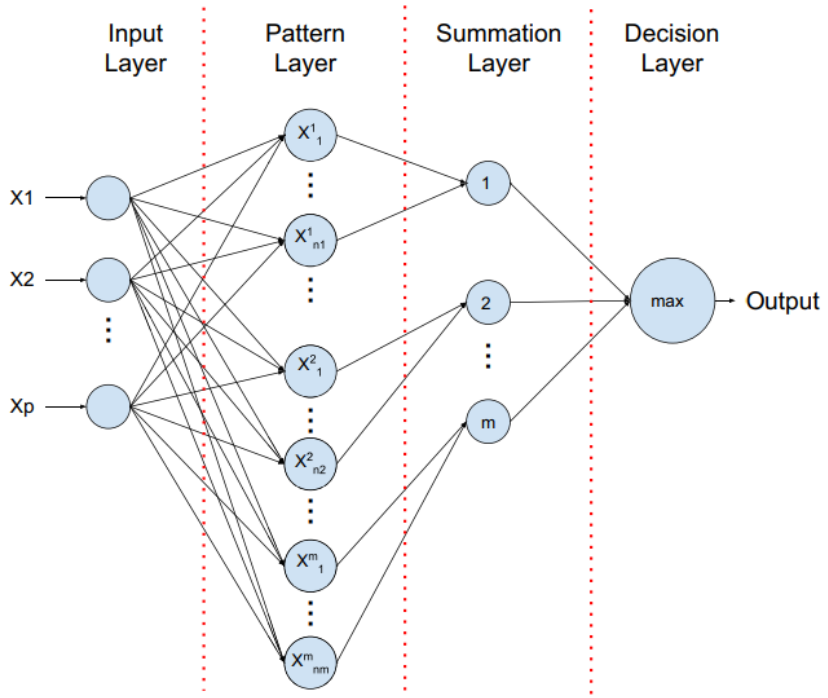
- Σε περίπτωση που υπάρχει γνώση των εκ των προτέρων πιθανοτήτων (π_k) και έχει οριστεί αυθαίρετα κάποιο κόστος για την ταξινόμηση σε μια τάξη k (l_k), τότε μετά τον υπολογισμό του μέσου όρου, υπολογίζεται η ποσότητα d_j.

$$d_k = P_k \pi_k l_k$$

- Για την ταξινόμηση υπολογίζεται η μέγιστη τιμή του dk.

$$argmax(d_k)$$

Αναφορικά με την δομή, τα πιθανοτικά νευρωνικά δίκτυα αποτελούνται από 4 βασικά στρωματά. Το στρώμα εισόδου (input layer) είναι το στρώμα που εισάγονται τα δεδομένα και κάθε κόμβος αντιπροσωπεύει ένα χαρακτηριστικό. Το επόμενο στρώμα είναι το στρώμα μοτίβου (pattern layer). Στο στρώμα μοτίβου υπολογίζονται όλες τις συναρτήσεις kernel. Στην συνέχεια οι τιμές που προκύπτουν μεταφέρονται στο στρώμα άθροισης (summation layer). Ο αριθμός των κόμβων στο στρώμα άθροισης είναι ίδιος με τον αριθμό των τάξεων (Specht D.F., 1990). Κάθε κόμβος περιέχει τις τιμές των δεδομένων εκπαίδευσης που ανήκουν στην εκάστοτε τάξη και υπολογίζει τον μέσο όρο τους. Το στρώμα εξόδου περιέχει την ταξινόμηση του αγνώστου σημείου με βάση την μέγιστη πιθανότητα των τάξεων. Ένα τυπικό πιθανοτικό νευρωνικό δίκτυο φαίνεται στο σχήμα 4.24.



Σχήμα 4.24 Ενδεικτική απεικόνιση ενός πιθανοτικού νευρωνικού δικτύου. (Πηγή: Mohebali et al. (2020))

Τα πιθανοτικά νευρωνικά δίκτυα αποτελούν μια χρήσιμη εναλλακτική των νευρωνικών δικτυών πρόσθιας τροφοδότησης, καθώς εκπαιδεύονται γρήγορα και είναι ανθεκτικά στα θορυβώδη δεδομένα. Για τις ανάγκες της εργασίας τα PNN είναι χρήσιμα για την πρόβλεψη του νικητή σε έναν αγώνα.

5ο ΚΕΦΑΛΑΙΟ

Εφαρμογές

5.1 Σκοπός της ανάλυσης

Η ανάλυση που θα ακολουθήσει αποτελείται από δυο βασικά μέρη:

- Η εξερεύνηση προβλεπτικών μοντέλων με διάφορους αλγόριθμους μηχανικής μάθησης, για την ανίχνευση της νίκης ή ήττας μιας ομάδας.
- Η χρήση συσταδοποίησης για την ταξινόμηση των παικτών του NBA με σκοπό τον προσδιορισμό του συγχρόνου τρόπου παιχνιδιού. Η συγκεκριμένη μέθοδος αντλεί έμπνευση από την έρευνα των Zhang et al. (2016). Στην συνέχεια επεκτείνεται αυτή η διαδικασία χωρίζοντας τις ομάδες σε 3 γκρουπ με βάση το ρεκόρ τους την σεζόν 2022-23, για να ελέγξουμε αν η ταξινόμηση των παικτών διαφοροποιείται ανάλογα με την τελική απόδοση των ομάδων.

Η ανάλυση πραγματοποιήθηκε στην γλώσσα Python

Τα δεδομένα που χρησιμοποιήθηκαν έχουν ληφθεί από το API πακέτο nba_api. Μέσω του συγκεκριμένου πακέτου παρέχεται άμεση πρόσβαση στα πραγματικά δεδομένα που υπάρχουν στην επίσημη ιστοσελίδα του NBA(nba.com). Για την εύρεση των δεδομένων που χρειάζονται, αρχικά πρέπει να εισαχθούν αρχικά όλοι οι υπό μελέτη αγώνες, καθώς και οι παίκτες που έπαιξαν τις συγκεκριμένες σεζόν. Τα παραπάνω βρίσκονται κωδικοποιημένα σε υποπακέτα του nba_api (nba_api.stats). Στην συνέχεια για την εύρεση των μεταβλητών που θα χρειαστούν για τις εφαρμογές, έγινε συλλογή μέσα από διαφορετικές ενότητες του πακέτου (boxscorefourfactorsv3, leaguegamefinder, boxscoreadvancedv2, teamgamelog).

5.2 Παρουσίαση των δεδομένων

Για την συσταδοποίηση έγιναν δοκιμές με διάφορους συνδυασμούς μεταβλητών και η τελική διάταξη που δίνει τα πιο ακριβή αποτελέσματα παρουσιάζεται στον πίνακα 5.1. Στις μεταβλητές έχουν συμπεριληφθεί και αναλυτικοί δείκτες όπως το PIE και το PER (NBA.com, Basketball-Reference.com)

ACTUAL_MINUTES	Λεπτά που αγωνίστηκε ο παίκτης
FG_PCT	Ποσοστό ευστοχίας προσπαθειών
FG3_PCT	Ποσοστό ευστοχίας τριπόντων

FT_PCT	Ποσοστό ευστοχίας στις ελεύθερες βολές
AVG_TOT_REB	Μέσος όρος συνολικών rebound
AVG_AST	Μέσος όρος assist
AVG_STL	Μέσος όρος κλεψιμάτων
AVG_TURNOVERS	Μέσος όρος λαθών
AVG_BLK	Μέσος ορός block
AVG_PTS	Μέσος όρος πόντων
percentagePointsMidrange2pt	Ποσοστό πόντων από το midrange
percentagePointsFastBreak	Ποσοστό πόντων από αντεπιθέσεις
percentagePointsFreeThrow	Ποσοστό πόντων από το ελεύθερες βολές
percentagePointsOffTurnovers	Ποσοστό πόντων από λάθη αντίπαλου
percentagePointsPaint	Ποσοστό πόντων από το ζωγραφιστό
percentageAssisted2pt	Ποσοστό πόντων που ήρθαν από assist και ήταν για 2 πόντους
percentageUnassisted2pt	Ποσοστό πόντων που δεν ήρθαν από assist και ήταν για 2 πόντους
percentageAssisted3pt	Ποσοστό πόντων που ήρθαν από assist και ήταν για 3 πόντους
percentageUnassisted3pt	Ποσοστό πόντων που δεν ήρθαν από assist και ήταν για 3 πόντους
playerPoints	Πόντοι που προήλθαν από το matchup του εκάστοτε παίκτη
matchupFieldGoalPercentage	Ποσοστό ευστοχίας από το matchup του εκάστοτε παίκτη

PIE	Δείκτης που μετράει την συνολική συνεισφορά ενός παίκτη με βάση τα στατιστικά του.
PER	Δείκτης που μετράει την επίδραση ενός παίκτη για κάθε λεπτό που αγωνίζεται

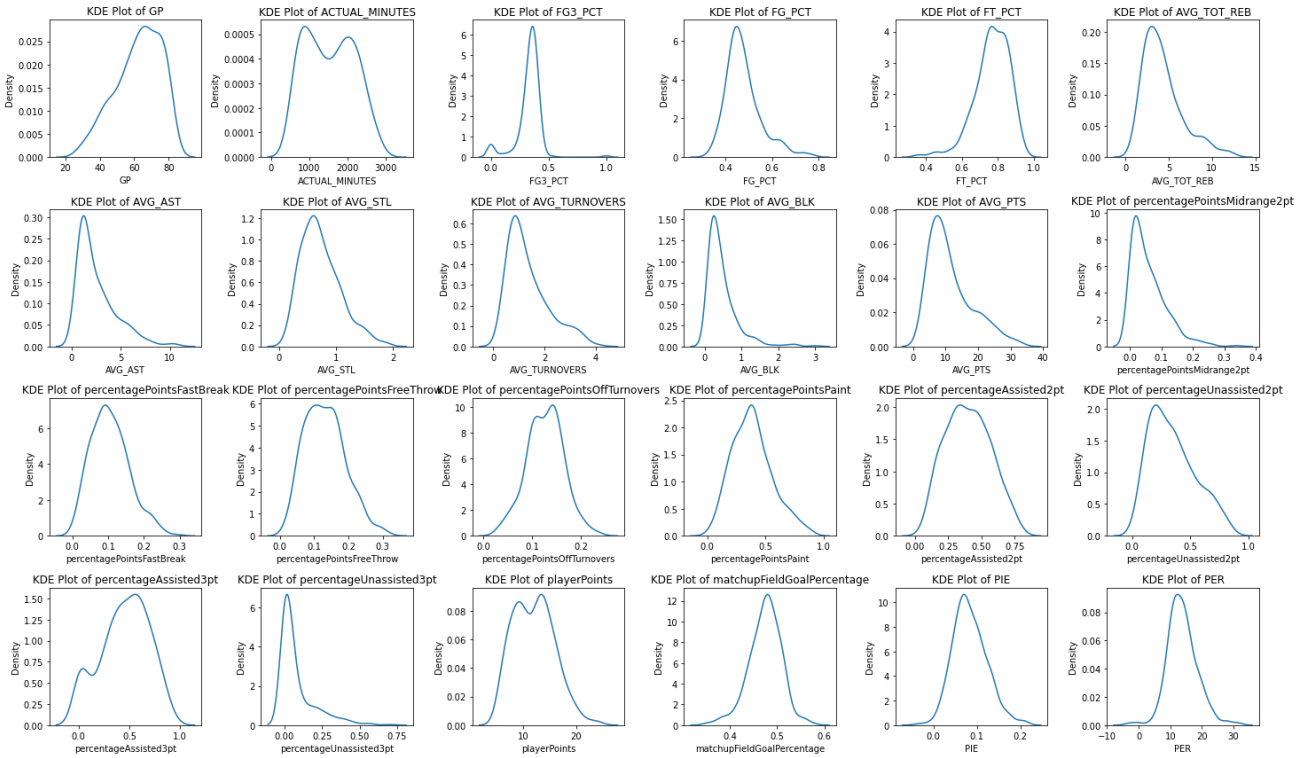
Πίνακας 5.1 Παρουσίαση των μεταβλητών

5.3 Προεπεξεργασία των δεδομένων

Για την ταξινόμηση των παικτών το σύνολο δεδομένων είναι για την σεζόν 2022-23 και αποτελείται αρχικά από 539 παρατηρήσεις.

Για την αποφυγή ακραίων ή ελαττωματικών τιμών χρησιμοποιήθηκαν οι παίκτες που αγωνίστηκαν σε τουλάχιστον 25 παιχνίδια και ο τελικός αριθμός παρατηρήσεων ήταν 366.

Παρακάτω παρατίθενται στο σχήμα 5.1 όλες οι κατανομές των μεταβλητών.



Σχήμα 5.1 Διαγράμματα πυκνότητας για όλες τις μεταβλητές που χρησιμοποιήθηκαν

Στο σύνολο δεδομένων δεν παρατηρήθηκαν ελλείπεις τιμές.

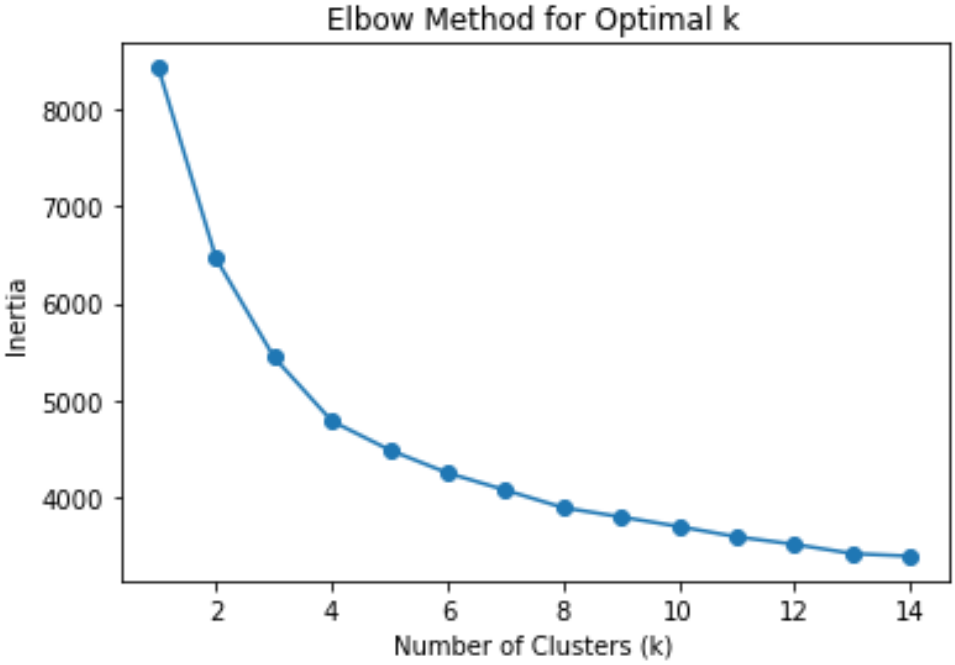
Λόγω ύπαρξης διαφορετικής κλίμακας στα δεδομένα πραγματοποιήθηκε τυποποίηση με χρήση της κανονικοποίησης Z-σκορ.

5.4 Εφαρμογή συσταδοποίησης

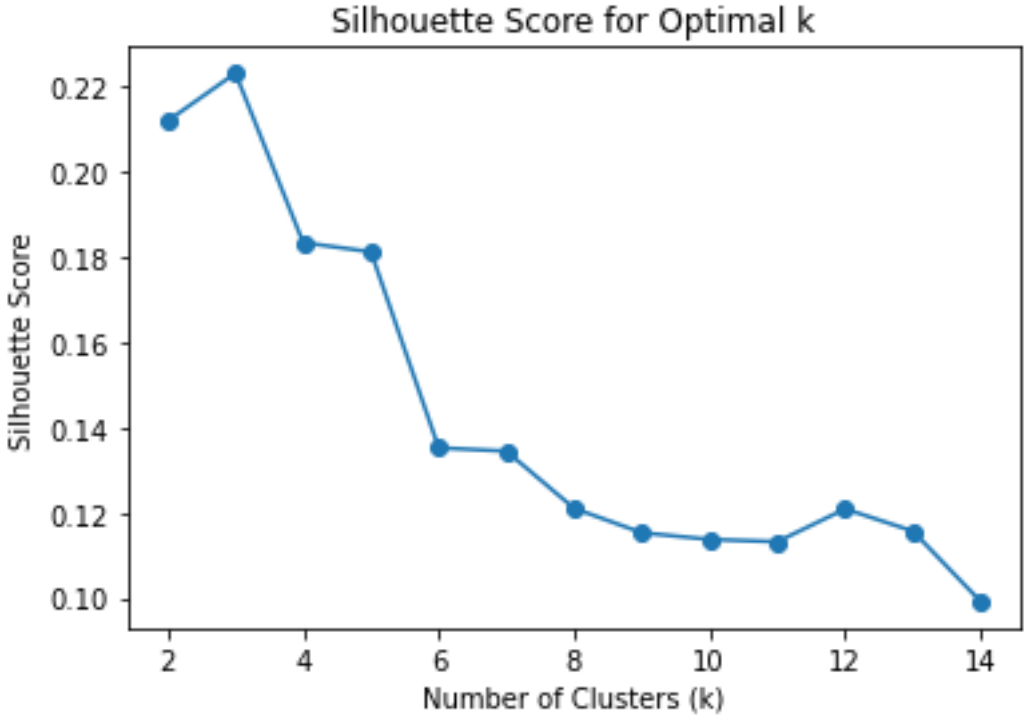
Για την τελική διάταξη χρησιμοποιήθηκαν για τις βασικές κατηγορίες (Πόντοι, rebound, assist, block, κλεψίματα) οι μεσοί όροι των παικτών. Εκτός αυτών έγιναν δοκιμές με τον συνολικό αριθμό και τον αριθμό ανά λεπτό των συγκεκριμένων κατηγοριών. Παρόλα αυτά οι συγκεκριμένες κατηγορίες δεν έδιναν ικανοποιητικά αποτελέσματα, για αυτό και καταλήξαμε στην χρήση των μέσων όρων.

Για την εφαρμογή αντλήθηκε έμπνευση από την έρευνα των Zhang et al. (2016), με την διαφορά ότι η συσταδοποίηση εφαρμόζεται για όλους τους παίκτες και δεν περιορίζεται μόνο στους guards. Όπως και στην προαναφερθείσα έρευνα έτσι και εδώ για την ταξινόμηση των παικτών χρησιμοποιήθηκε η μέθοδος συσταδοποίησης K-means. Το βασικότερο πρόβλημα της ανάλυσης είναι η επιλογή του κατάλληλου αριθμού K. Προτού προχωρήσουμε στις μεθόδους επιλογής, αξίζει να σημειωθεί ότι όπως αναφέρθηκε και στα προηγούμενα κεφάλαια, η καλαθοσφαίριση έχει εξελιχθεί ραγδαία τα τελευταία χρονιά με τις θέσεις πλέον να μην περιορίζονται στον βασική διάταξη (Point Guard-Shooting Guard, Small Forward, Power Forward, Center), αλλά με τους παίκτες να αναλαμβάνουν διαφορετικούς και πιο πληθωρικούς ρόλους. Για αυτό λοιπόν αναμένουμε ότι ο τελικός αριθμός K θα ξεπερνάει το 5, έτσι ώστε να πιάσει το πλήρες φάσμα του συγχρόνου τρόπου παιχνιδιού.

Για την επιλογή του αριθμού K θα χρησιμοποιηθούν οι μέθοδοι silhouette score και η μέθοδος αγκώνα. Χρησιμοποιώντας το εύρος 2 έως 15 προκύπτουν τα εξής σχήματα (σχήμα 5.2, σχήμα 5.3):



Σχήμα 5.2 Μέθοδος αγκώνα



Σχήμα 5.3 Silhouette score για τις διάφορες τιμές του K

Εξετάζοντας τα 2 σχήματα αρχικά βλέπουμε ότι φυσιολογικά όσο αυξάνεται ο αριθμός K τόσο μειώνεται. Από την μέθοδο αγκώνα δεν φαίνεται κάποιο χαρακτηριστικό σημείο, εκτός του K=4, το οποίο όμως είναι αρκετά μικρό και δεν μας δίνει κάποιο ουσιαστικό αποτέλεσμα.

Έτσι θα προχωρήσουμε στο silhouette score. Αγνοώντας τις συστάδες που προκύπτουν για K από 1 έως 5, βλέπουμε ότι το μεγαλύτερο score προκύπτει για K=6 και K=7. Παρόλα αυτά τα αποτελέσματα που δίνουν δεν είναι ικανοποιητικά, για αυτό και θα πρέπει να επιλεχθεί άλλος αριθμός συστάδων. Μετά από αναλυτική μελέτη στις συστάδες που προκύπτουν επιλέχθηκε το K=12, καθώς φαίνεται να δίνει τα πιο ακριβή αποτελέσματα, ενώ ταυτόχρονα δίνει και μια βαθύτερη εικόνα στα σύγχρονα είδη παικτών.

Εφαρμόζοντας τον αλγόριθμο K-means για K=12, προκύπτει ο παρακάτω πίνακας με τα κεντροειδή των συστάδων.

	0	1	2	3	4	5	6	7	8	9	10	11
ACTUAL_MINUTES	-1,22	0,45	0,22	0,90	0,04	1,12	0,83	0,97	-0,38	-1,12	-0,58	0,71
FG_PCT	-0,86	0,15	2,29	-0,24	0,48	-0,14	0,87	-0,24	-0,74	1,16	-0,46	-0,43
FG3_PCT	0,06	0,16	-2,45	0,30	0,30	0,36	-0,01	0,18	0,12	-0,65	0,31	0,45
FT_PCT	-0,05	0,02	-1,56	0,62	-0,45	0,89	0,35	0,49	0,35	-1,11	0,09	0,38
AVG_TOT_REB	-0,93	0,01	1,79	0,10	0,72	0,33	2,12	0,13	-0,78	-0,22	-0,64	-0,15
AVG_AST	-0,74	-0,20	-0,44	0,45	-0,34	2,30	1,06	1,55	0,25	-0,83	-0,63	-0,22
AVG_STL	-0,92	0,89	-0,09	0,17	-0,45	1,68	0,78	0,89	0,06	-0,80	-0,54	0,51
AVG_TURNOVERS	-0,87	-0,28	-0,10	0,63	-0,08	2,06	1,72	1,38	-0,11	-0,74	-0,77	-0,22
AVG_BLK	-0,65	0,00	1,69	-0,26	0,83	-0,05	1,44	-0,29	-0,55	0,31	-0,53	-0,20
AVG_PTS	-0,97	-0,29	-0,18	0,74	-0,10	2,02	1,87	1,30	-0,40	-0,90	-0,68	0,22
percentagePointsMidrange2pt	-0,56	-0,52	-0,64	0,15	-0,22	0,53	0,42	1,41	0,81	-0,77	-0,41	0,36
percentagePointsFastBreak	-0,96	1,89	-0,65	0,61	-0,39	0,16	-0,24	0,12	0,03	-0,87	0,07	0,79
percentagePointsFreeThrow	-0,93	0,16	0,37	0,98	0,17	1,16	1,44	0,67	-0,07	-0,43	-0,80	-0,43
percentagePointsOffTurnovers	-1,31	1,73	0,01	0,48	-0,16	0,53	0,11	0,06	0,26	-0,74	-0,19	0,55
percentagePointsPaint	-1,22	0,71	2,15	0,10	0,71	-0,14	0,91	0,02	-0,36	0,49	-0,71	-0,51
percentageAssisted2pt	-1,16	0,63	1,43	0,24	1,10	-1,08	0,80	-0,73	-0,81	0,04	-0,21	0,64
percentageUnassisted2pt	-1,06	0,03	-0,32	0,71	-0,37	1,99	0,48	1,67	0,68	-0,76	-0,84	-0,13
percentageAssisted3pt	-0,60	0,32	-1,85	0,80	0,13	-0,11	0,08	0,39	0,07	-1,42	0,28	1,27
percentageUnassisted3pt	-0,48	-0,45	-0,69	0,41	-0,55	3,17	0,09	1,28	0,21	-0,67	-0,39	-0,13
playerPoints	-1,23	0,05	0,98	0,44	0,63	0,81	1,40	0,73	-0,58	-0,77	-0,72	0,50
matchupFieldGoalPercentage	-0,93	0,19	0,05	0,00	0,32	0,18	-0,14	0,24	0,02	-0,29	0,16	0,46
PIE	-1,33	-0,09	0,87	0,45	0,21	1,57	2,07	0,81	-0,37	-0,37	-0,64	-0,22
PER	-0,49	-0,23	0,65	0,36	0,00	0,67	2,04	-0,03	-0,37	0,13	-0,44	-0,33

Πίνακας 5.2. Κεντροειδή των συστάδων

Παρακάτω επίσης παρατίθεται ένας επιπλέον πίνακας που δίνει μια σύντομη περιγραφή της εκάστοτε συστάδας μαζί με κάποια χαρακτηριστικά παραδείγματα παικτών που έχουν ταξινομηθεί.

Συστάδα	**Όνομα συστάδας**	**Χαρακτηριστικά παραδείγματα**
1	Role Players	Frank Ntilikina, George Hill
2	Perimeter defenders	Josh Hart, Herbert Jones
3	Classic Bigs	Jarrett Allen, Clint Capela
4	Combo Guards	Malik Monk, Immanuel Quickley
5	Modern Bigs	Brook Lopez, Miles Turner

6	Primary Ballhandlers	Steph Curry, Trae Young
7	Elite Bigs	Nikola Jokic, Joel Embiid, Lebron James
8	Primary scoring threats	Devin Booker, Jalen Brunson
9	Off the bench playmakers	TJ McConnell, Gabe Vincent
10	Off the bench bigs	Isaiah Hartenstein, Isaiah Jackson
11	Spot up shooters	Gary Harris, Troy Brown Jr
12	3 and D	Kentavious Caldwell-Pope, Klay Thompson

Πίνακας 5.3. Συνοπτική παρουσίαση των συστάδων

Από τα αποτελέσματα παρατηρείται ότι υπάρχει μια πλήρης εικόνα των αρχέτυπων που υπάρχουν στο NBA. Πιο συγκεκριμένα, η 1η συστάδα αποτελείται από παίκτες που συμπληρώνουν την 12αδα μιας ομάδας και έχουν ελάχιστο χρόνο συμμέτοχης.

Η 2η συστάδα αποτελείται από παίκτες που έχουν ως ειδικότητα την περιφερειακή άμυνα και οδηγούν σε λάθη τους αντίπαλους τους. Χαρακτηριστικό ότι έχουν πολύ μεγάλο κέντρο συστάδας για τα κλεψίματα. Επιπλέον οι συγκεκριμένοι παίκτες εκμεταλλεύονται τα λάθη για να σκοράρουν πόντους στην αντεπίθεση.

Η 3η συστάδα απεικονίζει το κλασικό πρότυπο ψηλών που υπάρχει από τις αρχές του αθλήματος. Παίκτες που έχουν μεγάλη έφεση στα rebound, στα block και αποτελούν αμυντικούς ογκόλιθους για τις ομάδες τους.

Η 4η συστάδα επιστρέφει στα guard και χαρακτηρίζει τους combo guards, δηλαδή παίκτες που έχουν την άνεση να αγωνίζονται με την μπάλα στα χέρια όταν λείπει ο βασικός χειριστης και ταυτόχρονα να έχουν την ευελιξία να λειτουργήσουν σε πιο συμπληρωματικούς ρόλους, κατευθύνοντας λιγότερο δηλαδή την επίθεση και αναλαμβάνοντας τον ρόλο του κλασικού shooting guard.

Η 5η συστάδα συνιστά το σύγχρονο μοντέλο ψηλών, δηλαδή παικτών που έχουν μεγαλύτερο επιθετικό ρεπερτόριο από τους κλασικούς ψηλούς που περιεγράφηκαν λίγο παραπάνω. Οι συγκεκριμένοι παίκτες μπορούν να απειλήσουν επιθετικά και στο ζωγραφιστό, αλλά και με το περιφερειακό σουτ. Επιπλέον είναι ικανοί αμυντικοί και έχουν αρκετά καλή ικανότητα στα rebound.

Η επόμενη συστάδα είναι για τους βασικούς χείριστες μιας ομάδας. Εδώ βρίσκονται κατά κύριο λόγο point guards που έχουν τον ρόλο του playmaker και είναι το σημείο αναφοράς της ομάδας τους. Οι συγκεκριμένοι παίκτες έχουν ένα πλήρες πακέτο ικανοτήτων, καθώς είναι ιδιαίτερα καλοί σκόρερ, με μεγάλη ικανότητα στο σουτ και στην επίθεση ένας εναντίον

ενός. Αυτοί οι παίκτες λαμβάνουν τον κύριο όγκο των αποφάσεων σε μια επίθεση, για αυτό αν και είναι εξίσου καλοί δημιουργοί, υποπίπτουν συνήθως σε μεγάλο αριθμό λαθών.

Η 7η συστάδα αποτελείται από τους καλυτέρους ψηλούς του πρωταθλήματος. Οι παίκτες αυτής της συστάδας συνδυάζουν το μέγεθος ενός ψηλού για να κυριαρχούν επιθετικά και να είναι κορυφαίοι αμυντικά, με την ικανότητα ενός παραδοσιακού guard να πασάρει και να σουτάρει. Δεν είναι τυχαίο ότι σε αυτή την συστάδα περιλαμβάνονται οι MVP των τελευταίων χρονών (Jokic, Embiid) και ένας, κατά γενική ομολογία, από τους καλυτέρους παίκτες που έχουν περάσει ποτέ από τα γήπεδα (Lebron James).

Προχωρώντας στην επόμενη συστάδα, παρατηρούμε ότι αποτελείται από τους παίκτες που κουβαλάνε το μεγαλύτερο επιθετικό φορτίο στην ομάδα τους. Οι συγκεκριμένοι παίκτες έχουν μεγάλη αποτελεσματικότητα στο σκοράρισμα, γεγονός που εξηγείται και από το μεγάλο ποσοστό προσπαθειών από το midrange, ένα σημείο του γηπέδου που πλέον οι περισσότεροι αποφεύγουν.

Οι συστάδες 9 και 10 απαρτίζονται από παίκτες που έρχονται από τον πάγκο και προσφέρουν διαφορετικά στοιχεία. Στην συστάδα 9 βρίσκονται κυρίως guards που μπορούν να προσφέρουν και στο σκοράρισμα και στην δημιουργία, ενώ στην συστάδα 10 βρίσκονται κυρίως ψηλοί που μπορούν να βοηθήσουν κυρίως αμυντικά με το μέγεθος τους.

Η 11η συστάδα είναι η συστάδα των παικτών με βασικό και σχεδόν αποκλειστικό ρολό την παραγωγή γρήγορου σκοραρίσματος μέσω του σουτ. Συνήθως οι συγκεκριμένοι παίκτες προσδίδουν στην επίθεση έναν απρόβλεπτο χαρακτήρα, καθώς προσπαθούν μέσω του σουτ από την περιφέρεια να ανοίξουν τις αντίπαλες άμυνες και να δώσουν χώρο στους υπόλοιπους συμπαίκτες.

Η τελευταία συστάδα αποτελεί μια διαρκώς αυξανομένη τάση στον χώρο της καλαθοσφαίρισης, και έχει λάβει την ονομασία 3 and D. Οι παίκτες αυτής της κατηγορίας είναι αμυντικοί υψηλού επιπέδου, που είναι επιπλέον πολύ χρήσιμοι επιθετικά, διότι είναι και πολύ επικίνδυνοι στα σουτ πίσω από την γραμμή των 3 πόντων.

Φαίνεται λοιπόν, ότι ο αλγόριθμος K means καλύπτει και δίνει μια ενδελεχή ματιά στις σύγχρονα είδη παικτών που επικρατούν.

Στην συνέχεια θα χρησιμοποιηθούν τα ίδια δεδομένα και η ίδια μέθοδος συσταδοποίησης, με την διαφορά να βρίσκεται στον διαχωρισμό των δεδομένων σε 3 υποκατηγορίες. Οι παίκτες χωρίζονται σε 3 κατηγορίες ανάλογα με την ομάδα που ανήκουν και το ρεκόρ που κατέγραψαν στην κανονική περίοδο. Οι ομάδες με τα 10 καλυτέρα ρεκόρ (Great teams) αποτελούν μια υποκατηγορία, οι επόμενες 10 (Mediocre teams) άλλη υποκατηγορία και οι τελευταίες 10 (Bad teams) άλλη μια συστάδα. Σκοπός του διαχωρισμού είναι να ερευνηθεί αν η κατανομή των συστάδων αλλάζει ανάλογα με τον αριθμό νικών μιας ομάδας, καθώς επίσης και αν τα δομικά χαρακτηριστικά των ομάδων παρουσιάζουν

διαφορές. Η συγκεκριμένη τεχνική μπορεί ενδεχομένως να παρέχει μια εσωτερική μάτια στα κομμάτια που διαφοροποιούν τις κορυφαίες ομάδες από τις υπόλοιπες.

Οι ομάδες που απαρτίζουν τις υποκατηγορίες είναι οι:

Great teams:

- Milwaukee Bucks
- Boston Celtics
- Philadelphia 76ers
- Denver Nuggets
- Cleveland Cavaliers
- Memphis Grizzles
- Sacramento Kings
- New York Knicks
- Brooklyn Nets
- Phoenix Suns

Mediocre teams:

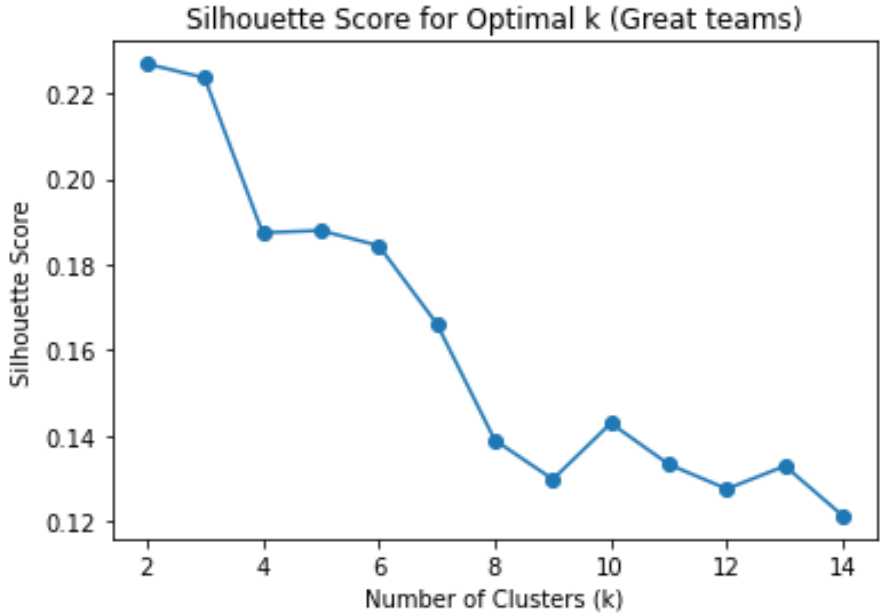
- Golden State Warriors
- Los Angeles Clippers
- Miami Heat
- Los Angeles Lakers
- Minnesota Timberwolves
- New Orleans Pelicans
- Atlanta Hawks
- Toronto Raptors
- Chicago Bulls
- Oklahoma City Thunder

Bad teams:

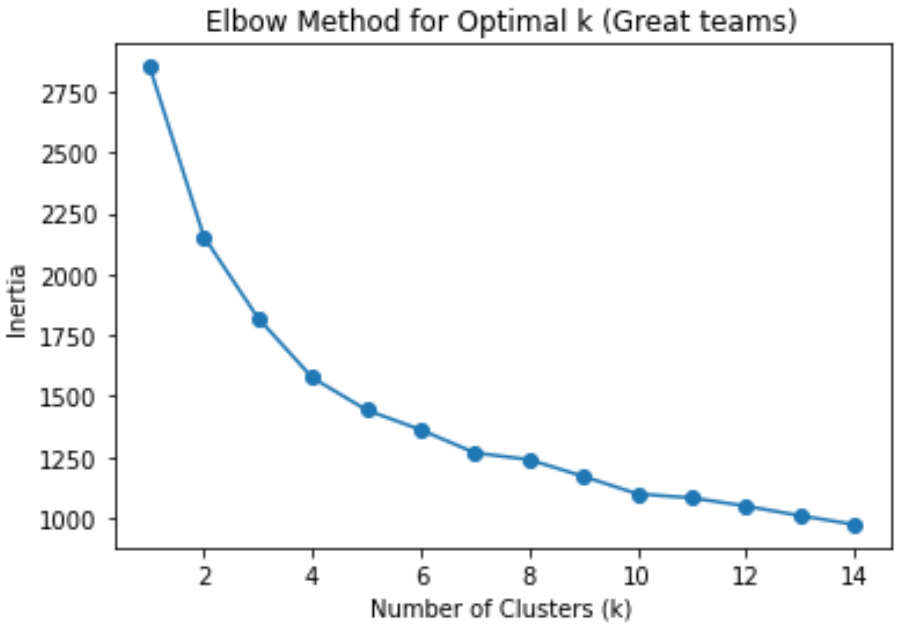
- Dallas Mavericks
- Utah Jazz
- Indiana Pacers
- Washington Wizards
- Orlando Magic
- Portland Trail Blazers
- Charlotte Hornets
- Houston Rockets
- San Antonio Spurs
- Detroit Pistons

Αρχικά θα πρέπει ξανά να βρούμε τον καλύτερο αριθμό συστάδων. Εφόσον πλέον τα δεδομένα έχουν μειωθεί, αναμένουμε ότι ο αριθμός των συστάδων θα είναι μικρότερος και ενδεχομένως να εντοπίσει ελαφρώς διαφορετικά μοτίβα παικτών από την εκδοχή με ολόκληρα τα δεδομένα. Ακόμη η λογική που χρησιμοποιήθηκε στην παραπάνω εφαρμογή, θα χρησιμοποιηθεί και εδώ, δηλαδή μας ενδιαφέρει ο αριθμός των συστάδων να ξεπερνάει το K=5, διότι τότε τα αποτελέσματα τείνουν να θυμίζουν την κλασική διάταξη μιας ομάδας.

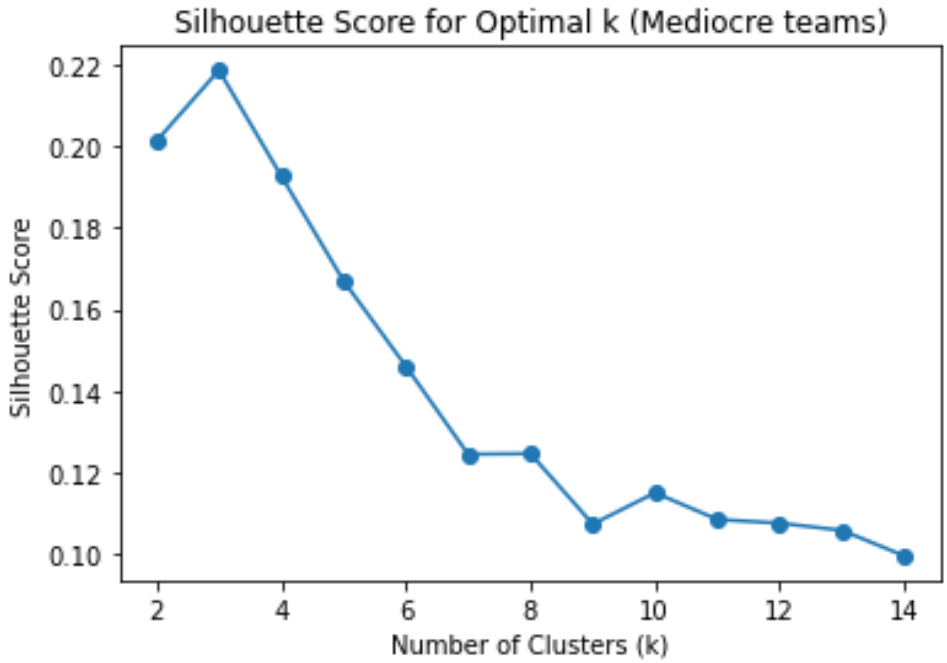
Εφαρμόζοντας τις μεθόδους silhouette score και την μέθοδο αγκώνα προκύπτουν τα παρακάτω σχήματα (σχήμα 5.4-5.9):



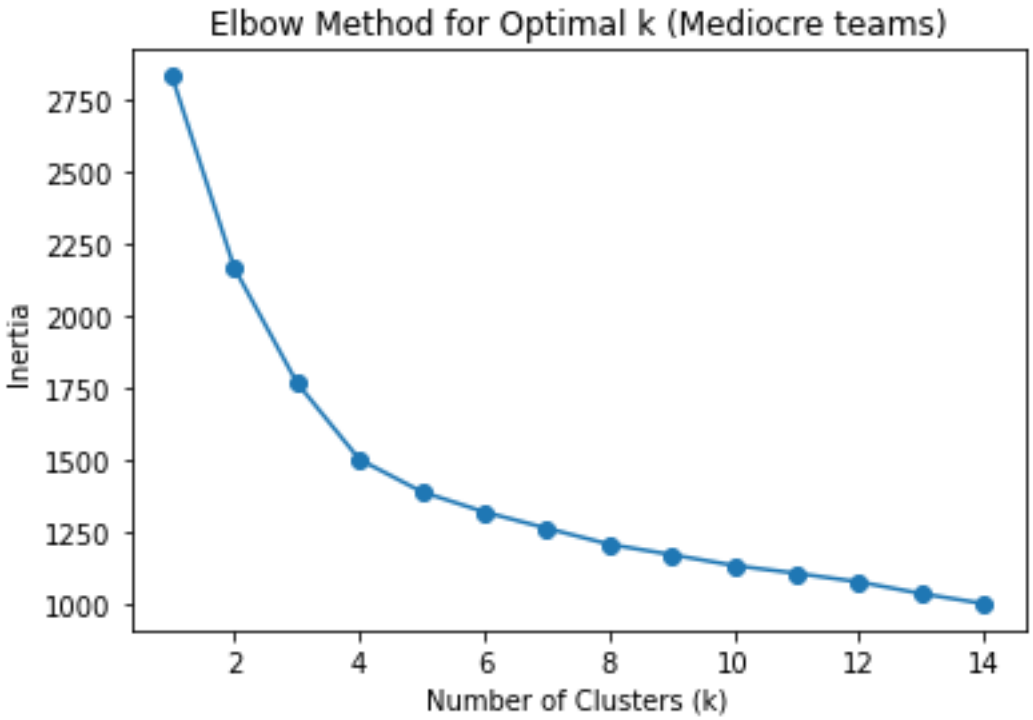
Σχήμα 5.4 Silhouette score για τις κορυφαίες ομάδες



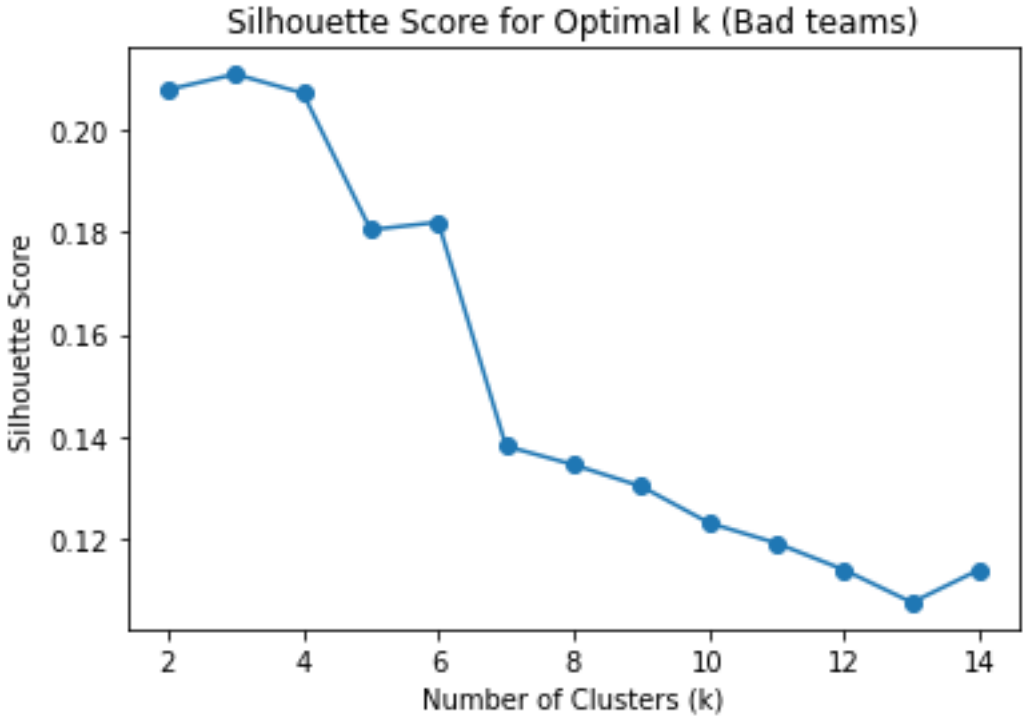
Σχήμα 5.5 Μέθοδος αγκώνα για τις κορυφαίες ομάδες



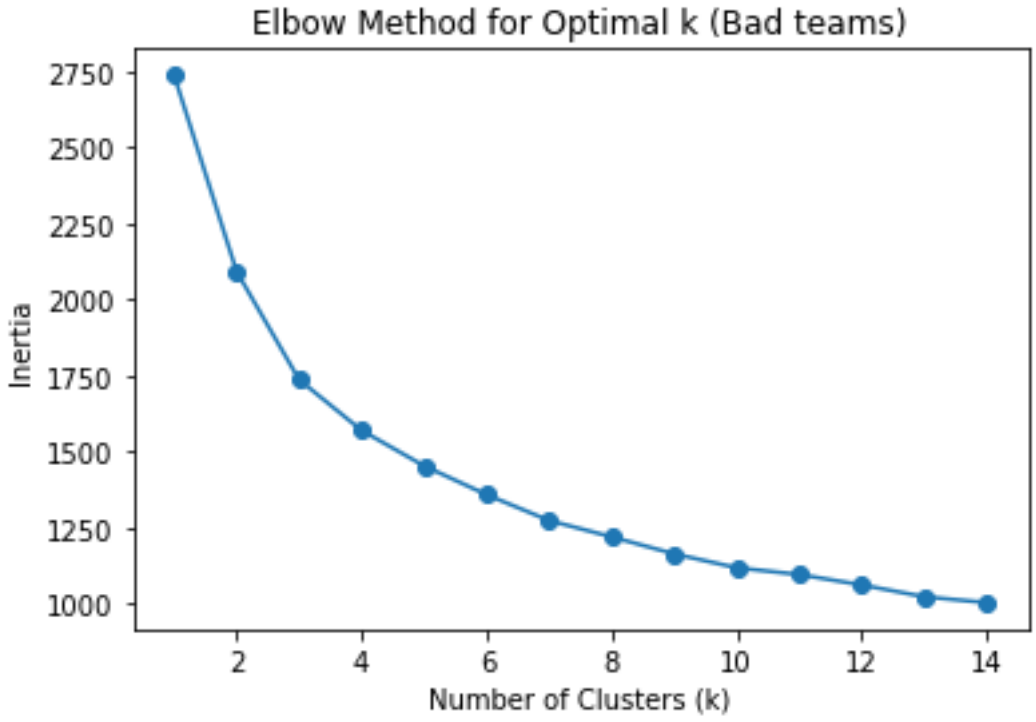
Σχήμα 5.6 Silhouette score για τις μέτριες ομάδες



Σχήμα 5.7 Μέθοδος αγκώνα για τις μέτριες ομάδες



Σχήμα 5.8 Silhouette score για τις κακές ομάδες



Σχήμα 5.9. Μέθοδος αγκώνα για τις κακές ομάδες

Από την μέθοδο του αγκώνα δεν φαίνεται να υπάρχει κάποιο σημείο που να δίνει ασφαλή συμπεράσματα. Παρατηρώντας τα σχήματα του silhouette score φαίνεται να υπάρχει και

στις 3 περιπτώσεις μεγάλη πτώση μετά το K=6. Εξετάζοντας αναλυτικά τις συστάδες που προκύπτουν ο τελικός αριθμός συστάδων παραμένει ο ίδιος.

Εφαρμόζοντας τον αλγόριθμο προκύπτουν οι ακόλουθοι πίνακες με τα κεντροειδή:

	BAD TEAMS					
	0	1	2	3	4	5
ACTUAL_MINUTES	-0,40025	0,1577	0,572494	-0,95614	0,732921	1,211081
FG_PCT	1,742016	-0,13557	-0,14289	-0,637	0,502671	-0,17422
FG3_PCT	-1,13697	0,121087	0,18808	-0,07578	0,767332	0,266788
FT_PCT	-1,2074	0,609401	0,118847	-0,17326	0,085829	0,757224
AVG_TOT_REB	0,753758	-0,37017	0,020239	-0,88016	1,701947	0,376396
AVG_AST	-0,67725	1,001834	-0,26698	-0,60325	-0,09534	1,593994
AVG_STL	-0,53064	1,146442	0,166661	-0,61133	-0,14712	0,813593
AVG_TURNOVERS	-0,36484	0,250613	-0,17769	-0,80039	0,582686	1,817813
AVG_BLK	1,16063	-0,34909	-0,37626	-0,46984	1,35876	-0,22575
AVG_PTS	-0,42664	0,023201	0,128208	-0,84961	0,628064	1,770577
percentagePointsMidrange2pt	-0,54225	1,323129	-0,31665	-0,36181	-0,12711	0,668739
percentagePointsFastBreak	-0,50982	0,785391	0,963846	-0,52114	-0,70044	0,315979
percentagePointsFreeThrow	0,147997	0,11146	0,119349	-0,7601	0,792144	0,809244
percentagePointsOffTurnovers	-0,23925	0,716297	0,937337	-0,65659	-0,3326	0,193178
percentagePointsPaint	1,361392	0,257588	0,125104	-0,84936	0,567453	-0,07274
percentageAssisted2pt	0,86816	-0,62199	0,753436	-0,71063	1,223215	-0,48136
percentageUnassisted2pt	-0,44894	1,215792	-0,23678	-0,7425	-0,13557	1,511865
percentageAssisted3pt	-1,37392	0,084881	0,832859	-0,38817	0,685901	0,50421
percentageUnassisted3pt	-0,59364	0,216685	-0,32594	-0,34745	-0,44351	1,988468
playerPoints	0,15582	-0,02575	0,199407	-1,01546	1,531476	0,847194
matchupFieldGoalPercentage	-0,00161	-0,05735	0,640445	-0,60964	0,464981	0,322891
PIE	0,38414	0,379049	-0,10364	-0,96306	0,818972	1,102584
PER	0,184289	0,372242	-0,14804	-0,35373	0,459818	0,16118

Πίνακας 5.4 Κεντροειδή για τις συστάδες των κακών ομάδων

BAD TEAMS		
Συστάδα	**Όνομα συστάδας**	**Χαρακτηριστικά παραδείγματα**
1	Backup Bigs	Walker Kessler, Isaiah Jackson
2	Combo Guards	Tyus Jones, Andrew Nembhard
3	3nD	Aaron Nesmith, Deni Avdija
4	Role Players	Usman Garuba, Jordan Nwora

5	Starting Bigs	Kristaps Porzingis, Lauri Markkanen
6	Primary Scoring Threats	Luka Doncic, Bradley Beal

Πίνακας 5.5 Συνοπτική παρουσίαση των συστάδων για τις κακές ομάδες

MEDIOCRE TEAMS						
	0	1	2	3	4	5
ACTUAL_MINUTES	0,169368	0,477412	-0,73261	0,792499	0,921027	-1,23179
FG_PCT	-0,72063	2,114841	0,157997	0,227308	0,046827	-0,64754
FG3_PCT	0,319912	-2,3905	0,141228	0,318781	0,342113	0,056315
FT_PCT	0,412357	-0,94355	-0,27671	0,103665	0,584984	-0,38857
AVG_TOT_REB	-0,54952	2,144781	-0,30669	0,327543	0,466941	-0,84793
AVG_AST	0,082735	-0,42499	-0,59489	-0,00925	1,589009	-0,82837
AVG_STL	0,188786	-0,23456	-0,61448	0,731719	0,904475	-0,95408
AVG_TURNOVERS	-0,08951	0,131238	-0,62446	0,038237	1,636767	-0,962
AVG_BLK	-0,61183	1,775458	-0,29723	0,683798	0,001118	-0,40593
AVG_PTS	-0,05205	0,154151	-0,63162	0,053238	1,659246	-1,04827
percentagePointsMidrange2pt	0,221927	-0,42512	-0,38434	0,021958	1,033653	-0,69624
percentagePointsFastBreak	0,1712	-0,91395	0,038426	0,864453	0,317677	-0,85274
percentagePointsFreeThrow	0,058637	0,858374	-0,60924	0,011144	0,982088	-0,83534
percentagePointsOffTurnovers	0,170158	-0,04701	-0,22593	0,995055	0,339239	-1,12538
percentagePointsPaint	-0,54979	1,878697	0,179824	0,513823	0,104955	-1,06783
percentageAssisted2pt	-0,38585	1,346076	0,371591	1,045919	-0,45083	-1,09395
percentageUnassisted2pt	0,464484	-0,24435	-0,58997	-0,17865	1,469465	-1,13104
percentageAssisted3pt	0,769843	-1,79032	-0,10907	0,542341	0,243805	-0,60578
percentageUnassisted3pt	0,250215	-0,71903	-0,47271	-0,45611	1,456956	-0,50578
playerPoints	-0,02408	1,293501	-0,55372	0,694664	0,650041	-1,26496
matchupFieldGoalPercentage	0,222596	0,384892	0,387284	0,122453	-0,23385	-0,79823
PIE	-0,3076	1,226466	-0,40063	0,02714	1,300163	-1,16999
PER	-0,22843	0,895091	-0,2007	-0,44965	0,860404	-0,46308

Πίνακας 5.6 Κεντροειδή για τις συστάδες των μέτριων ομάδων

MEDIOCRE TEAMS		
Συστάδα	**Όνομα συστάδας**	**Χαρακτηριστικά παραδείγματα**
1	Combo Guards	Austin Reaves, Coby White
2	Starting Bigs	Rudy Gobert, Bam Adebayo
3	Off the Bench Scorers	Kevin Love, Lonnie Walker IV
4	3nD	Alex Caruso, Klay Thompson
5	Star Cluster	Steph Curry, Lebron James
6	Role Players	Malachi Flynn, Max Christie

Πίνακας 5.7 Συνοπτική παρουσίαση των συστάδων για τις μέτριες ομάδες

	GREAT TEAMS					
	0	1	2	3	4	5
ACTUAL_MINUTES	0,772713	-0,17303	-1,09346	1,068806	0,872883	-0,73703
FG_PCT	-0,33315	2,042239	1,214685	0,955666	-0,16645	-0,64807
FG3_PCT	0,333675	-2,44531	-0,39241	-0,10276	0,26997	0,197222
FT_PCT	0,232524	-1,87084	-0,96826	-0,04624	0,646299	0,239646
AVG_TOT_REB	0,03272	1,662484	-0,14178	2,159967	0,333861	-0,79611
AVG_AST	0,051872	-0,17402	-0,81556	0,75929	1,633079	-0,50793
AVG_STL	0,558963	0,468988	-0,95799	0,420709	0,999026	-0,65021
AVG_TURNOVERS	-0,06093	0,104338	-0,71838	1,550322	1,581565	-0,62985
AVG_BLK	-0,1058	1,733381	0,207672	1,950579	-0,1851	-0,57898
AVG_PTS	0,152377	-0,2545	-0,85815	1,496456	1,581651	-0,67925
percentagePointsMidrange2pt	-0,15382	-0,71588	-0,72022	0,274694	1,119245	0,012854
percentagePointsFastBreak	0,781428	-0,22641	-0,94708	-0,57601	-0,03776	-0,10906
percentagePointsFreeThrow	0,15115	-0,0606	-0,35825	1,238142	1,068413	-0,64407
percentagePointsOffTurnovers	0,738865	0,043287	-0,98111	0,020006	0,065309	-0,26183
percentagePointsPaint	-0,02738	2,464995	0,472282	1,209431	-0,02714	-0,78797
percentageAssisted2pt	0,417773	1,408851	0,127629	1,270952	-0,77191	-0,55127
percentageUnassisted2pt	0,198973	-0,08881	-0,81551	0,149399	1,740192	-0,56842
percentageAssisted3pt	0,929277	-1,82136	-1,40179	-0,10168	0,186281	0,029272
percentageUnassisted3pt	-0,10383	-0,71291	-0,67815	-0,30082	2,117531	-0,33859
playerPoints	0,428	0,587946	-0,55538	2,105943	0,672192	-0,9061
matchupFieldGoalPercentage	0,121204	-0,54167	0,431403	0,327349	0,163204	-0,30059
PIE	-0,00149	0,742183	-0,36508	1,742647	1,239575	-0,81707
PER	-0,18194	0,613079	0,145453	2,014312	0,034166	-0,40782

Πίνακας 5.8 Κεντροειδή για τις συστάδες των κορυφαίων ομάδων

GREAT TEAMS		
Συστάδα	**Όνομα συστάδας**	**Χαρακτηριστικά παραδείγματα**
1	3nD	Derrick White, Dillon Brooks
2	Classic Bigs	Jarrett Allen, Mitchell Robinson
3	Bench Bigs	Isaiah Hartenstein, Bismack Biyombo
4	Modern Bigs	Joel Embiid, Nikola Jokic
5	Star Cluster	Donovan Mitchell, Kevin Durant
6	Off the bench scorers	Edmond Sumner, Jevon Carter

Πίνακας 5.9 Συνοπτική παρουσίαση των συστάδων για τις κορυφαίες ομάδες

Από τα αποτελέσματα μπορούμε να εξάγουμε μερικά χρήσιμα συμπεράσματα. Προηγουμένως έγινε εκτενής αναφορά στην εξέλιξη του παιχνιδιού και ότι πλέον θεωρείται αναγκαίο για τους 'ψηλούς' να εμπλουτίσουν το παιχνίδι τους. Παρόλα αυτά η 2η συστάδα (Classic Bigs) των κορυφαίων βάση ρεκόρ ομάδων απαρτίζεται αποκλειστικά από ψηλούς που έχουν βασικό ρόλο στην ομάδα τους και τα βασικά τους χαρακτηριστικά είναι αυτά που έκαναν τους ψηλούς του παρελθόντος κορυφαίους. Αυτοί οι παίκτες παρουσιάζουν μεγάλη έφεση στα rebound, block και είναι αμυντικές σταθερές στις ομάδες τους. Επομένως φαίνεται ότι οι ψηλοί που διαπρέπουν σε αυτούς τους τομείς μπορούν να σταθούν και μάλιστα να παίζουν βασικό ρόλο σε καλές ομάδες.

Μια επιπλέον παρατήρηση που φαίνεται είναι η έλλειψη καθαρής συστάδας με παίκτες που λογίζονται ως forward. Αντ' αυτού όλοι μοιράζονται σε συστάδες ανάλογα με τον ρόλο και το μέγεθος τους, δείχνοντας πόσο ευέλικτος είναι ο ρόλος του forward σε κάθε ομάδα. Το παραπάνω συμπέρασμα ενισχύει τον αρχικό ισχυρισμό ότι οι κλασικές θέσεις στην καλαθοσφαίριση δεν προσφέρουν παρά μια επιφανειακή εικόνα του παιχνιδιού.

Αξίζει ακόμη να αναφερθεί ότι ανεξαρτήτου ρεκόρ ομάδας ο ρόλος των παικτών που μπορούν να προσφέρουν στην άμυνα και να είναι διαρκείς απειλές πίσω από την γραμμή των 3 πόντων (συστάδα 3nD) είναι περιζήτητος, αφού παρουσιάζεται σε όλες τις κατηγορίες ομάδων.

Παρατηρώντας συνολικά τις συστάδες μπορούμε να διακρίνουμε μερικά χαρακτηριστικά που διαφοροποιούν τις ομάδες. Αρχικά φαίνεται ότι η παρουσία ενός καλού ψηλού παίκτη είναι τόσο επιδραστική όσο αυτή ενός πολύ καλού guard. Αυτό γίνεται σαφές κοιτώντας τις συστάδες των κορυφαίων ομάδων. Στις κορυφαίες ομάδες υπάρχουν 2 βασικές συστάδες που αποτελούνται αποκλειστικά από παίκτες από αγωνίζονται στις θέσεις 4-5 και η μια από αυτές αποτελείται από ψηλούς με κλασικά χαρακτηριστικά ενός center όπως το rebound, το παιχνίδι κοντά στο καλάθι και το αμυντικό παιχνίδι. Γίνεται κατανοητό λοιπόν ότι οι καλοί κλασικοί ψηλοί μπορούν να επηρεάσουν σε μεγάλο βαθμό το παιχνίδι των ομάδων τους. Αντίθετα στις άλλες 2 κατηγορίες υπάρχει μια βασική συστάδα ψηλών. Στις ομάδες με μέτρια και κακά ρεκόρ η απόδοση επηρεάζεται περισσότερο από combo guards, διότι μόνο αυτές έχουν συστάδα που τους περιέχει. Μια ένδειξη που δείχνει το κενό δυναμικότητας αναμεσά στις κακές με τις υπόλοιπες ομάδες είναι η έλλειψη κορυφαίων παικτών. Αυτό γίνεται σαφές, από το γεγονός, ότι οι ομάδες που βρίσκονται από πάνω έχουν και στις 2 κατηγορίες μια συστάδα που ταξινομεί τους κορυφαίους guard-forward (Star Cluster), εν αντιθέσει με την 3η κατηγορία.

Ρίχνοντας μια ματιά στα κεντροειδή, προκαλεί ενδιαφέρον το γεγονός ότι οι καλύτεροι σουτέρ 3 πόντων δεν φαίνεται να τοποθετούνται στις κορυφαίες, αλλά στις μέτριες ομάδες. Αυτό το εύρημα αποτελεί ένα χαρακτηριστικό παράδειγμα ότι αν και το τρίποντο είναι πλέον βασικό στοιχείο του παιχνιδιού, δεν είναι αποκλειστικός παράγοντας για την επιτυχία

μιας ομάδας. Επιπρόσθετα στρέφοντας την προσοχή σε αμυντικές μεταβλητές όπως τα κλεψίματα (AVG_STL) και τους πόντους που επιτρέπουν στους παίκτες που μαρκάρουν (playerPoints) φαίνεται ότι οι καλύτερες ομάδες δείχνουν μεγαλύτερη συνέπεια στην άμυνα.

5.5 Μοντέλα πρόβλεψης

Η πρόβλεψη αποτελεσμάτων είναι από τα πιο γνωστά και διαδεδομένα προβλήματα που υπάρχουν στον χώρο των basketball analytics. Με βάση τις έρευνες που αναφέρθηκαν στην βιβλιογραφική ανασκόπηση πάνω στο ζήτημα, θα χρησιμοποιηθούν αλγόριθμοι μηχανικής μάθησης για την πρόβλεψη της νίκης μιας ομάδας.

Τα δεδομένα που χρησιμοποιήθηκαν είναι για τις σεζόν 2020-21, 2021-22, 2022-23. Οι σεζόν 2020-21 και 2021-22 χρησιμοποιήθηκαν ως δεδομένα εκπαίδευσης και ως δεδομένα ελέγχου χρησιμοποιήθηκε η σεζόν 2022-23.

Πιο αναλυτικά για τα σύνολα δεδομένων η μεταβλητή στόχος είναι κατηγορική και αντιπροσωπεύει την νίκη ή την ήττα της ομάδας. Για τις ανεξάρτητες μεταβλητές χρησιμοποιήθηκαν 2 διαφορετικά σύνολα δεδομένων. Το ένα αποτελείται από τα κλασικά στατιστικά που περιέχονται στο τυπικό box score. Το 2ο αποτελείται από τις τέσσερις μεταβλητές που όρισε ο Oliver ως τους 4 πιο σημαντικούς παράγοντες για να κερδίσει μια ομάδα. Οι μεταβλητές παρουσιάζονται πιο αναλυτικά στον παρακάτω πίνακα.

BOX SCORE STATS	
PTS	Πόντοι που σημειωθήκαν
REB	Αριθμός rebound
AST	Αριθμός assist
STL	Αριθμός κλεψίματων
BLK	Αριθμός κοψίματων
TOV	Αριθμός λαθών

Πίνακας 5.8 Μεταβλητές που χρησιμοποιούνται για το box score και χρησιμοποιήθηκαν παρακάτω

FOUR FACTORS	
effectiveFieldGoalPercentage	Δείκτης που παρομοιάζει το ποσοστό ευστοχίας με μια τροποποίηση που δίνει μεγαλύτερο βάρος στο σουτ 3 πόντων
freeThrowAttemptRate	Ο λόγος των προσπαθειών από τις ελεύθερες βολές προς τις

	συνολικές προσπάθειες που παίρνει μια ομάδα
teamTurnoverPercentage	Το ποσοστό των συνολικών κάτοχων που καταλήγουν σε λάθος
offensiveReboundPercentage	Το ποσοστό των επιθετικών rebound

Πίνακας 5.9. Μεταβλητές που χρησιμοποιούνται για τους 4 παράγοντες και χρησιμοποιήθηκαν παρακάτω

Η πρόβλεψη του αποτελέσματος είναι ένα πρόβλημα ταξινόμησης και για αυτό θα χρησιμοποιηθούν οι παρακάτω αλγόριθμοι:

- Μηχανές διανυσματικής υποστήριξης (Support vector machines)
- Τυχαία δάση (Random forests)
- Δέντρα απόφασης (Decision trees)
- Ταξινομητής Naïve Bayes (Naïve Bayes classifier)
- Γραμμική διαχωριστική ανάλυση (Linear discriminant analysis)
- Λογιστική παλινδρόμηση (Logistic regression)
- Κ κοντινότεροι γείτονες (K nearest neighbors)
- Νευρωνικά δίκτυα (Neural networks)

Για την εύρεση των βέλτιστων παραμέτρων σε μερικούς αλγορίθμους χρησιμοποιήθηκαν οι τεχνικές βελτιστοποίησης παραμέτρων GridsearchCV και Optuna. Η τεχνική GridsearchCV διερευνάει όλα τα πιθανά αποτελέσματα μέσα από αναλυτική αναζήτηση. Η τεχνική Optuna είναι μια πιο αποτελεσματική σε θέμα χρόνου, διότι αντί να δοκιμάσει όλους τους πιθανούς συνδυασμούς, χρησιμοποιεί τον εκτιμητή Tree structured Parzen (TPE). Το εύρος σε κάθε περίπτωση καθορίζεται από τον χρήστη. Στα τελικά αποτελέσματα χρησιμοποιήθηκαν μόνο η μεγαλύτερη ακρίβεια που πρόκυπτε από τις 2 τεχνικές, με την GridsearchCV να παρουσιάζει ελάχιστα καλυτέρα αποτελέσματα αλλά με μεγαλύτερο κόστος χρονικά.

5.6 Εφαρμογή μοντέλων πρόβλεψης

Box score stats

Αρχικά θα γίνει εφαρμογή των τεχνικών μηχανικής μάθησης στα βασικά δεδομένα που παρέχονται στο box score. Τα αποτελέσματα που προκύπτουν φαίνονται στον κάτωθι πίνακα:

BOX SCORE STATS					
Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC**
SVM	75,61%	73,83%	79,35%	76,49%	82,54%
Random Forest	73,54%	70,17%	81,87%	75,57%	81,25%
Decision Tree	71,59%	67,66%	82,68%	74,42%	78,66%
Naive Bayes	74,63%	72,99%	78,21%	75,51%	81,55%
LDA	75,73%	74,71%	77,80%	76,22%	82,61%
Logistic Regression	75,77%	74,42%	78,54%	76,42%	82,62%
KNN	74,31%	73,04%	77,07%	75,00%	81,21%

Πίνακας 5.10 Τα αποτελέσματα που καταγράφηκαν για όλα τα διαφορετικά μοντέλα, με μεταβλητές τα box score στατιστικά

Παρατηρώντας τον πίνακα φαίνεται ότι η λογιστική παλινδρόμηση παρουσιάζει συνολικά την καλύτερη εικόνα.

Πιο αναλυτικά η λογιστική παλινδρόμηση πετυχαίνει το μεγαλύτερο ποσοστό ακρίβειας με 75.77%, με την γραμμική διαχωριστική ανάλυση να πετυχαίνει ελάχιστα μικρότερο ποσοστό (75.73%). Αρκετά χαμηλότερη ακρίβεια από τα υπόλοιπα μοντέλα έχουν επίσης τα δέντρα απόφασης (71.59%).

Προχωρώντας στην στήλη με την ορθότητα, τα περισσότερα μοντέλα κυμαίνονται σε ποσοστά κοντά στο 73-74%, εκτός των δέντρων απόφασης και των τυχαίων δασών. Τα δέντρα απόφασης μάλιστα καταγράφουν αρκετά χαμηλότερο ποσοστό με 67.66%. Τη μεγαλύτερη τιμή πετυχαίνει η γραμμική διαχωριστική ανάλυση.

Αν και τα δέντρα απόφασης παρουσίαζαν χαμηλότερα ποσοστά ορθότητας και ακρίβειας σε σχέση με τα υπόλοιπα μοντέλα, επιτυγχάνουν το μεγαλύτερο ποσοστό ανάκλησης με 82.68%. Έτσι μπορούμε να συμπεράνουμε ότι έχουν μεγάλη ικανότητα ανίχνευσης του αληθινού αριθμού νικών των ομάδων. Επιπλέον και τα δέντρα απόφασης παρουσιάζουν μεγάλο ποσοστό με 81.87%, ενώ όλα τα υπόλοιπα μοντέλα κυμαίνονται από 77-79%.

Για το f1-score όλα τα μοντέλα έχουν παρόμοια εικόνα με τις μηχανές διανυσματικής υποστήριξης να έχουν το μεγαλύτερο ποσοστό με 76.49% και τα δέντρα απόφασης το χαμηλότερο με 74.42%.

Στην συνέχεια ελέγχουμε την στήλη με την περιοχή κάτω από την καμπύλη, οπού η λογιστική παλινδρόμηση ξανά έχει οριακά το μεγαλύτερο ποσοστό με 82.62%, ενώ αξίζει να αναφερθεί ότι η γραμμική διαχωριστική ανάλυση έχει σχεδόν πανομοιότυπο αποτέλεσμα (82.61%). Τα δέντρα απόφασης για άλλη μια φορά έχουν το χαμηλότερο ποσοστό με 78.66%.

Four Factors

Εφαρμόζοντας τις τεχνικές οδηγούμαστε στα παρακάτω αποτελέσματα

FOUR FACTORS					
Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC**
SVM	71,06%	68,04%	79,43%	73,29%	79,67%
Random Forest	70,33%	67,48%	78,46%	72,56%	78,04%
Decision Tree	70,77%	68,53%	76,83%	72,44%	77,29%
Naive Bayes	70,24%	68,28%	75,61%	71,76%	78,35%
LDA	71,18%	68,18%	79,43%	73,38%	79,68%
Logistic Regression	71,02%	67,94%	79,59%	73,31%	79,68%
KNN	71,10%	68,52%	78,05%	72,98%	77,51%

Πίνακας 5.11 Τα αποτελέσματα που καταγράφηκαν για όλα τα διαφορετικά μοντέλα, με μεταβλητές τους 4 παράγοντες του Oliver

Από τον πίνακα παρατηρώντας όλες τις μετρικές βλέπουμε ότι το μοντέλο της γραμμικής διαχωριστικής ανάλυσης φαίνεται να αποδίδει καλύτερα.

Πιο συγκεκριμένα, η γραμμική διαχωριστική ανάλυση παρουσιάζει το μεγαλύτερο ποσοστό ακρίβειάς με 71.18%. Τα υπόλοιπα μοντέλα δεν φαίνεται να αποκλίνουν σημαντικά, καθώς όλα ξεπερνάνε το 70%.

Προχωρώντας στην στήλη της ακρίβειας παρατηρούμε ότι όλα τα μοντέλα φαίνεται βρίσκονται λίγο πιο πάνω ή λίγο πιο κάτω από το 68%.

Για την ανάκληση τα αποτελέσματα που προκύπτουν έχουν απόκλιση μεταξύ τους. Τα μοντέλα των δέντρων απόφασης και Naïve Bayes παρουσιάζουν μικρότερα ποσοστά σε σχέση με τα υπόλοιπα, με ποσοστά 76.83% και 75.61% αντίστοιχα. Αντίθετα τα υπόλοιπα μοντέλα ξεπερνάνε το 78% με τις SVM και LDA να φτάνουν το 79.43%.

Για το f1-score τα αποτελέσματα βρίσκονται αναμεσά στο 71-73% με την LDA να έχει το μεγαλύτερο ποσοστό με 73.38%.

Για την περιοχή κάτω από την καμπύλη βλέπουμε ότι τα μοντέλα LDA,SVM και λογιστικής παλινδρόμησης δίνουν τα υψηλότερα και σχεδόν ίδια ποσοστά με σχεδόν 79.7%. Τα δέντρα απόφασης ξανά παρουσιάζουν το χαμηλότερο ποσοστό με 77.29%.

5.7 Εφαρμογή με νευρωνικά δίκτυα

Για την πρόβλεψη αποτελεσμάτων χρησιμοποιήθηκαν τα εξής νευρωνικά δίκτυα:

- Νευρωνικά δίκτυα πρόσθιας τροφοδότησης(FFNN)
- Επαναλαμβανόμενα νευρωνικά δίκτυα (RNN)
- Συνελικτικά νευρωνικά Δίκτυα (CNN)

Με στόχο την μεγιστοποίηση της απόδοσης των μοντέλων, διαπιστώθηκε ότι για διαφορετικό αριθμό μεταβλητών μεταβάλλεται η δομή των νευρωνικών δικτύων. Πιο συγκεκριμένα, για τους 4 παράγοντες του Oliver τα αποτελέσματα ήταν βελτιωμένα όταν υπήρχαν 2 κρυμμένα στρώματα, ενώ για τα box score stats τα αποτελέσματα μεγιστοποιούνταν για 3 κρυφά στρώματα. Τα αποτελέσματα και για τις 2 βάσεις δεδομένων μετρά την εφαρμογή των μοντέλων παρουσιάζονται στον παρακάτω πίνακα:

BOX SCORE STATS					
Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**AUC Score**
FFNN	75,28%	72,15%	82,36%	76,91%	82,43%
RNN	74,15%	70,37%	83,41%	76,34%	82,22%
CNN	74,31%	71,51%	80,81%	75,88%	81,75%

Πίνακας 5.12 Τα αποτελέσματα που καταγράφηκαν για όλα τα διαφορετικά μοντέλα νευρωνικών δικτύων, με μεταβλητές τα box score στατιστικά

FOUR FACTORS					
Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**AUC Score**
FFNN	71,34%	69,40%	76,34%	72,70%	79,09%
RNN	71,54 %	69,17 %	77,72 %	73,20 %	79,48 %
CNN	71,99%	68,88%	80,24%	74,13%	79,66%

Πίνακας 5.13 Τα αποτελέσματα που καταγράφηκαν για όλα τα διαφορετικά μοντέλα, με μεταβλητές τους 4 παράγοντες του Oliver

Ξεκινώντας από τον πίνακα 5.12 παρατηρούμε ότι από τα 3 μοντέλα αυτό που φαίνεται να ξεχωρίζει είναι τα νευρωνικά δίκτυα πρόσθιας τροφοδότησης. Στο μοντέλο FFNN σημειώνονται τα μεγαλύτερα ποσοστά σε όλους τους δείκτες εκτός της ανάκλησης.

Συγκρίνοντας όμως το μοντέλο FFNN με το μοντέλο της λογιστικής παλινδρόμησης, βλέπουμε ότι αν και πλησιάζει αρκετά, υστερεί και στην ακρίβεια (75.28% έναντι75.77%) και στο AUC score (79.66% έναντι 82.62%).

Χρησιμοποιώντας για χαρακτηριστικά τους 4 παράγοντες παρατηρούμε από τον πίνακα 5.13 ότι και τα 3 μοντέλα δίνουν παρόμοια αποτελέσματα, με τα συνελικτικά νευρωνικά δίκτυα να παρέχουν ελαφρώς πιο ικανοποιητικά αποτελέσματα. Στο μοντέλο CNN παρατηρείται η μεγαλύτερη ακρίβεια (71.99%) και AUC score (79.66%) ανάμεσα στα νευρωνικά δίκτυα. Επιπλέον τα νευρωνικά δίκτυα πρόσθιας τροφοδότησης έχουν το μεγαλύτερο ποσοστό ορθότητας (69.40%). Συγκρίνοντας το μοντέλο συνελικτικών νευρωνικών δικτύων με το μοντέλο γραμμικής διαχωριστικής ανάλυσης, παρατηρείται ότι τα νευρωνικά δίκτυα παρουσιάζουν βελτίωση στην ακρίβεια (71.99% έναντι 71.18%) και έχουν σχεδόν πανομοιότυπα AUC score (79.66% έναντι 79.68%).

Συγκρίνοντας τα αποτελέσματα που προέκυψαν, βλέπουμε ότι τα box score στατιστικά ενός αγώνα έχουν μεγαλύτερη προβλεπτική ικανότητα τους παράγοντες του Oliver. Η επιλογή του αλγορίθμου δεν φαίνεται να επηρεάζει τόσο τα αποτελέσματα, όσο η επιλογή των χαρακτηριστικών, κάτι που συμπίπτει με το συμπέρασμα της έρευνας των Shi et al (2013).

6ο ΚΕΦΑΛΑΙΟ

Συμπεράσματα

Στην παρούσα εργασία καλύφθηκε αναλυτικά το αντικείμενο των basketball analytics. Έγινε μια εισαγωγή για την ιστορία και την εξέλιξη του αθλήματος από τα πρώτα χρονιά μέχρι σήμερα. Αναφέρθηκαν επίσης και άλλοι τομείς του παιχνιδιού, όπως οι παράγοντες που συμβάλλουν στην απόδοση μιας ομάδας και η ποικιλομορφία του παιχνιδιού που υπάρχει στα κορυφαία πρωταθλήματα.

Στην συνέχεια διενεργήθηκε μια ενδελεχής βιβλιογραφική ανασκόπηση, κατά την οποία δημιουργήθηκαν 3 κατηγορίες για την ανάλυση των βασικότερων πεδίων των basketball analytics.

Έπειτα έγινε μια συνοπτική παρουσίαση των μεθόδων μηχανικής μάθησης, των μεθόδων αξιολόγησης μοντέλων και των μεθόδων προεπεξεργασίας.

Στο επόμενο κεφάλαιο έγιναν εφαρμογές πάνω σε 2 διαφορετικούς τομείς των basketball analytics.

Αρχικά διερευνήθηκε η εξέλιξη του σύγχρονου τρόπου παιχνιδιού των παικτών του NBA μέσω της συσταδοποίησης K-means. Η έρευνα αυτή έγινε για όλο το σύνολο των ομάδων και ξεχωριστά για 3 κατηγορίες ομάδων με βάση τα ρεκόρ τους, με σκοπό την εύρεση διαφοροποιήσεων ανάλογα με τον αριθμό νικών. Από την έρευνα για όλες τις ομάδες προέκυψαν 12 συστάδες με διαφορετικά χαρακτηριστικά, ενώ η ξεχωριστή ανάλυση για τις 3 κατηγορίες οδήγησε σε 6 συστάδες. Από τις συστάδες που προέκυψαν φάνηκε μια σαφής στόχευση των ομάδων για παίκτες που έχουν αποτελεσματικότητα πίσω από την γραμμή του τριπόντου και ταυτόχρονα είναι συνεπείς στα αμυντικά τους καθήκοντα (Συστάδα 3nD).

Επιπρόσθετα παρατηρήθηκε ότι οι παίκτες που λογίζονται ως forwards δεν έχουν συστάδες που απευθύνονται σε αυτούς αποκλειστικά και κατανέμονται σε διαφορετικές συστάδες που μπορεί να περιέχουν κυρίως guard ή κυρίως centers. Από αυτό συμπεραίνουμε ότι ο ρόλος του forward στο σύγχρονο NBA είναι πολυδιάστατος και ανάλογα με τις ικανότητες του παίκτη και τις ανάγκες της ομάδας διαφοροποιείται.

Επιπλέον παρά την τάση των σύγχρονων ψηλών να έχουν χαρακτηριστικά που έχουν οι guards/forwards, φαίνεται ότι το κλασικό αρχέτυπο των center παραμένει βασικό στοιχείο των παικτών του NBA και μάλιστα αποτελούν κομμάτι των ομάδων με τα καλυτέρα ρεκόρ. Παρόλα αυτά αξίζει να σημειωθεί ότι ακόμα και οι ψηλοι με το πιο κλασικό σετ ικανοτήτων έχουν προσαρμοστεί στα σύγχρονα δεδομένα. Στο κεφάλαιο 2 έγινε αναφορά στον Roy

Hibbert που διαθέτει αρκετά παρόμοιο τρόπο παιχνιδιού με παίκτες όπως ο Mitchell Robinson και ο Jarrett Allen. Η διαφορά των 2 προαναφερθέντων με τονHibbert έγκειται στο κομμάτι της φυσικής κατάστασης. Οι παίκτες αυτοί για να ακολουθήσουν τον πολύ γρήγορο ρυθμό έγιναν πιο ελαφριοί και πιο ευκίνητοι για να συνάδουν με τους υπολοίπους συμπαίκτες τους.

Μελετώντας τις διάφορες των ομάδων των 3 κατηγοριών, γίνεται σαφές ότι η ποιότητα των centers επηρεάζει σε σημαντικό βαθμό την πορεία μιας ομάδας. Επίσης ένα χαρακτηριστικό των κορυφαίων ομάδων είναι ότι διαθέτουν παίκτες που εκτός από την επίθεση, έχουν μεγάλη έφεση στην άμυνα.

Στο δεύτερο κομμάτι των εφαρμογών, κατασκευάστηκαν μοντέλα πρόβλεψης με δεδομένα εκπαίδευσης τα αποτελέσματα των σεζόν 2020-21 και 2021-2022 και δεδομένα ελέγχου τα παιχνίδια της σεζόν 2022-23. Ως χαρακτηριστικά χρησιμοποιήθηκαν 2 παραλλαγές: Στην μια περίπτωση οι 4 παράγοντες του Oliver (2011) και στην δεύτερη τα στατιστικά που χρησιμοποιούνται στο box score.

Τα αποτελέσματα που προέκυψαν ήταν ικανοποιητικά και στις 2 περιπτώσεις. Χρησιμοποιώντας τους 4 παράγοντες του Oliver, όλα τα μοντέλα είχαν ακρίβεια που ξεπερνάει το 70%, με τα συνελικτικά νευρωνικά δίκτυα να παρουσιάζουν το μεγαλύτερο ποσοστό. Για τα στατιστικά του box score η ακρίβειά βελτιώθηκε με κάποια μοντέλα να ξεπερνάνε το 75% και την λογιστική παλινδρόμηση να έχει την μεγαλύτερη ακρίβεια.

ΠΑΡΑΡΤΗΜΑΤΑ

Π1 Κύρια μέτρα απόστασης

Μέτρα απόστασης	Τύπος
Ευκλείδεια	$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
Manhattan	$\sum_{i=1}^{n}\|x_i - y_i\|$
Minkowski	$\left(\sum_{i=1}^{n}\|x_i - y_i\|^p\right)^{1/p}$
Chebyshev	$\max_i\|x_i - y_i\|$

Π2 Αναλυτικά αποτελέσματα συσταδοποίησης

Συστάδα	Τύπος
1	A. Holiday|A. Coffey|B. Griffin|B. McGowens|C. Houstan|C. Thomas|C. Okeke|D. Nix|D. Saric|D. Robinson|E. Omoruyi|F. Ntilikina|G. Mathews|G. Hill|H. Highsmith|I. Wainright|J. Thor|J. Walker|J. Hardy|J. Bouknight|J. Toscano-Anderson|J. Hernangomez|J. Holiday|L. Waters III|M. Flynn|M. Beauchamp|M. Christie|M. Muscala|M. McBride|M. Moody|N.

	Alexander-Walker|O. Agbaji|P. Tucker|P. Mills|P. Pritchard|R. Hampton|R. Neto|R. Covington|R. McGruder| S. Mamukelashvili|S. Fontecchio|V. Cancar|W. Matthews
2	A. Caruso|A. Dosunmu|B. Brown|C. Boucher|D. Wright|H. Jones|I. Okoro|J. McDaniels|J. Johnson|J. McDaniels|J. Williams|J. Vanderbilt|J. Nowell|J. Konchar|J. Green|J. Hart|J. Winslow|K. Martin|O. Anunoby|T. Eason|T. Mann|T. Murphy III|
3	A. Drummond|B. Simmons|B. Clarke|C. Capela|D. Gafford|D. Ayton|D. Powell|E. Mobley|I. Zubac|J. Poeltl|J. Duren|J. Allen|K. Looney|M. Williams|M. Plumlee|M. Robinson|N. Claxton|O. Okongwu|R. Williams III|R. Gobert|S. Adams|. Kessler
4	A. Reaves|B. Mathurin|B. Bogdanovic|C. LeVert|C. Anthony|C. Sexton|D. White|D. Bane|F. Wagner|G. Hayward|H. Barnes|I. Quickley|J. Grant|J. Poole|J. Giddey|K. Johnson|K. Olynyk|K. Kuzma|K. Lowry|L. Markkanen|L. Dort|M. Brogdon|M. Monk|M. Bridges|M. Conley|N. Marshall|N. Powell|P. Banchero|R. Barrett|S. Bey|S. Barnes|T. Herro|T. Maxey
5	A. Gordon|A. Horford|A. Pokusevski|B. Portis|B. Bol|B. Lopez|C. Wood|D. Avdija|D. Green|D. Eubanks|I. Stewart|J. Smith|J. Green|J. Sochan|J. Collins|J. Valanciunas|J. Kuminga|J. Nurkic|K. Bates-Diop|K. Williams|K. Anderson|L. Nance Jr.|M. Bagley III|M. Kleber|M. Wagner|M. Turner|N.

	Reid|N. Richards|P. Washington|P. Achiuwa|R. Hachimura|S. Aldama|T. Bryant|T. Craig|W. Carter Jr.|Z. Collins
6	A. Edwards|D. Lillard|D. Garland|D. Mitchell|F. VanVleet|J. Morant|J. Harden|J. Holiday|K. Porter Jr.|K. Irving|L. Doncic|P. George|S. Gilgeous-Alexander|S. Curry|T. Young|T. Haliburton
7	A. Sengun|A. Davis|B. Adebayo|D. Sabonis|G. Antetokounmpo|J. Jackson Jr.|J. Tatum|J. Butler|J. Embiid|K. Towns|K. Durant|K. Porzingis|L. James|N. Jokic|N. Vucevic|Z. Williamson
8	A. Simons|B. Beal|B. Ingram|C. McCollum|C. Paul|D. Russell|D. Fox|D. DeRozan|D. Murray|D. Booker|J. Ivey|J. Brunson|J. Green|J. Murray|J. Brown|J. Clarkson|J. Randle|K. Leonard|K. Middleton|K. Hayes|L. Ball|M. Smart|M. Fultz|P. Siakam|R. Westbrook|S. Dinwiddie|T. Rozier|T. Jones|Z. LaVine|
9	A. Burks|B. Wesley|B. Hyland|C. Payne|C. Duarte|C. Joseph|D. Mitchell|D. Schroder|D. Smith Jr.|D. Graham|E. Gordon|G. Vincent|G. Dragic|J. Suggs|J. Carter|J. Wall|J. Goodwin|J. Alvarado|J. Okogie|J. Richardson|M. Branham|M. Morris Sr.|M. Morris|R. Jackson|R. Rubio|R. Gay|S. Curry|S. Milton|T. McConnell|T. Horton-Tucker|T. Maledon|T. Mann|T. Jerome|T. Jones|V. Oladipo
10	A. Gill|B. Biyombo|C. Bassey|C. Metu|C. Koloko|D. Bazley|D. Sharpe|D. Jordan|D. Jones Jr.|H. Diallo|I. Hartenstein|I. Jackson|J.

	Tate|J. Wiseman|J. Hayes|J. Sims|J. Landale|K. Jones|L. Kornet|M. Bamba|M. Harrell|P. Reed|R. Langford|T. Young|T. Watford|U. Garuba|W. Gabriel|X. Tillman|Z. Nnaji
11	A. Griffin|A. Wiggins|A. Lamb|A. Rivers|C. Reddish|C. Osman|C. Braun|D. Lee|D. House Jr.|D. Roddy|D. Wade|D. McDermott|D. Daniels|E. Sumner|G. Harris|G. Niang|G. Williams|I. Joe|I. Livers|J. Green|J. Williams|J. Robinson-Earl|J. Harris|J. Ingles|J. McLaughlin|J. Nwora|J. Christopher|K. Nunn|K. Knox II|K. Love|L. Stevens|L. Shamet|L. Walker IV|L. Kennard|M. Thybulle|N. Little|N. Batum|O. Toppin|O. Brissett|O. Dieng|P. Connaughton|P. Beverley|R. Bullock Jr.|S. Hauser|T. Warren|T. Davis|T. Ross|T. Lyles|T. Brown Jr.|W. Barton|Y. Watanabe|Z. Williams
12	A. Nesmith|A. Nembhard|A. Wiggins|B. Bogdanovic|B. Hield|C. Martin|C. Johnson|C. White|C. Kispert|D. Hunter|D. Melton|D. Vassell|D. Brooks|D. DiVincenzo|D. Finney-Smith|G. Trent Jr.|G. Allen|J. Smith Jr.|K. Murray|K. Oubre Jr.|K. Caldwell-Pope|K. Huerter|K. Thompson|M. Beasley|M. Strus|M. Porter Jr.|P. Williams|Q. Grimes|R. O'Neale|S. Sharpe|T. Prince|T. Hardaway Jr.|T. Harris

ΒΙΒΛΙΟΓΡΑΦΙΑ

Αναφορές

Alamar, B. (2018, June). Rockets, Spurs lead the way in NBA draft analytics. Ανάκτηση από ESPN.com: https://www.espn.com/nba/story/_/id/23762871/rockets-spurs-celtics-most-analytical-draft-teams-nba

Alonso, R. P., & Babac, M. B. (2022). Machine learning approach to predicting a basketball game outcome. International Journal of Data Science, 7(1), 60-77.

Apostolou, K. &. (2019). Sports Analytics algorithms for performance prediction. In 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA) (pp. 1-4). IEEE.

Arastey, G. M. (2019, November 27). HISTORY OF PERFORMANCE ANALYSIS: THE CONTROVERSIAL PIONEER CHARLES REEP. Ανάκτηση από www.sportperformanceanalysis.com: https://www.sportperformanceanalysis.com/article/history-of-performance-analysis-the-controversial-pioneer-charles-reep

Balakrishnama, S., & Ganapathiraju, A. (1998). Linear discriminant analysis-a brief tutorial. Institute for Signal and information Processing, 18(1998), 1-8.

Bianchi, F., Facchinetti, T., & Zuccolotto, P. (2017). Role revolution: towards a new meaning of positions in basketball. Electronic Journal of Applied Statistical Analysis, 10(3), 712-734.

Cervone, D., D'Amour, A., Bornn, L., & Goldsberry, K. (2014, February). Pointwise: Predicting points and valuing decisions in real time with nba optical tracking data. In Proceedings of the 8th MIT Sloan Sports Analytics Conference, Boston, MA, USA (Vol. 28, p. 3).

Chen, C. Y., Lai, W., Hsieh, H. Y., Zheng, W. H., Wang, Y. S., & Chuang, J. H. (2018, October). Generating defensive plays in basketball games. In Proceedings of the 26th ACM international conference on Multimedia (pp. 1580-1588).

Cheng, G., Zhang, Z., Kyebambe, M. N., & Kimbugwe, N. (2016). Predicting the outcome of NBA playoffs based on the maximum entropy principle. Entropy, 18(12), 450.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20, 273-297.

Csapo, P., & Raab, M. (2014). “Hand down, man down.” analysis of defensive adjustments in response to the hot hand in basketball using novel defense metrics. PloS one, 9(12), e114184.

Daly-Grafstein, D., & Bornn, L. (2020). Using in-game shot trajectories to better understand defensive impact in the NBA. Journal of Sports Analytics, 6(4), 235-242.

Erčulj, F., & Štrumbelj, E. (2015). Basketball shot types and shot success in different levels of competitive basketball. PloS one, 10(6), e0128885.

Fletcher, T. (2009). Support vector machines explained. Tutorial paper, 1-19.

Franks, A., Miller, A., Bornn, L., & Goldsberry, K. (2015). Characterizing the spatial structure of defensive skill in professional basketball.
Goldberg, M. (2017). Evaluating Lineups and Complementary Play Styles in the NBA (Doctoral dissertation).

Géron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. " O'Reilly Media, Inc.".

Goldsberry, K. (2019). Sprawlball: A visual tour of the new era of the NBA. Mariner Books.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.

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.

Horvat, T., Job, J., & Medved, V. (2018, September). Prediction of Euroleague games based on supervised classification algorithm k-nearest neighbours. In 6th International Congress on Support Sciences Research and Technology Support (Vol. 20, p. 21).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Jiao, J., Hu, G., & Yan, J. (2021). A Bayesian marked spatial point processes model for basketball shot chart. Journal of Quantitative Analysis in Sports, 17(2), 77-90.

Juravich, M., Salaga, S., & Babiak, K. (2017). Upper echelons in professional sport: The impact of NBA general managers on team performance. Journal of Sport Management, 31(5), 466-479.

Korn, F., Labrinidis, A., Kotidis, Y., Faloutsos, C., Kaplunovich, A., & Perkovic, D. (1998). Quantifiable data mining using principal component analysis.

Kubatko, J., Oliver, D., Pelton, K., & Rosenbaum, D. T. (2007). A starting point for analyzing basketball statistics. Journal of quantitative analysis in sports, 3(3).

Lewis, M. (2004). Moneyball: The art of winning an unfair game. WW Norton & Company.

Liu, B., & Liu, B. (2011). Supervised learning. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 63-132.

Loeffelholz, B., Bednar, E., & Bauer, K. W. (2009). Predicting NBA games using neural networks. Journal of Quantitative Analysis in Sports, 5(1).

López, F. A., Martínez, J. A., & Ruiz, M. SPATIAL PATTERN ANALYSIS OF SHOT ATTEMPTS IN BASKETBALL; THE CASE OF LA LAKERS ANÁLISIS ESPACIAL DE LANZAMIENTOS EN BALONCESTO; EL CASO DE LA LAKERS.

Lucey, P., Bialkowski, A., Carr, P., Yue, Y., & Matthews, I. (2014, February). How to get an open shot: Analyzing team movement in basketball using tracking data. In Proceedings of the 8th annual MIT SLOAN sports analytics conference.

Milanović, D., Selmanović, A., & Škegro, D. (2014). Characteristics and differences of basic types of offenses in European and American top-level basketball. In 7th International Scientific Conference on Kinesiology (p. 400).

Miller, A., Bornn, L., Adams, R., & Goldsberry, K. (2014, January). Factorized point process intensities: A spatial analysis of professional basketball. In International conference on machine learning (pp. 235-243). PMLR.

Miller, A. C., & Bornn, L. (2017, March). Possession sketches: Mapping NBA strategies. In Proceedings of the 2017 MIT Sloan Sports Analytics Conference (pp. 1-12).

Migliorati, M. (2020). Detecting drivers of basketball successful games: an exploratory study with machine learning algorithms. Electronic Journal of Applied Statistical Analysis, 13(2), 454-473.

Mohebali, B., Tahmassebi, A., Meyer-Baese, A., & Gandomi, A. H. (2020). Probabilistic neural networks: a brief overview of theory, implementation, and application. Handbook of probabilistic models, 347-367.

Oliver, D. (2011). Basketball on paper: rules and tools for performance analysis. U of Nebraska Press.

Osken, C., & Onay, C. (2022). Predicting the winning team in basketball: A novel approach. Heliyon, 8(12).

Papageorgiou, G. (2022). Data Mining in Sports: Daily NBA Player Performance Prediction.

Pelton, K. (2015). THE GREAT ANALYTICS RANKINGS. Ανάκτηση από ESPN.com: https://www.espn.com/espn/feature/story/_/id/12331388/the-great-analytics-rankings

Reich, B. J., Hodges, J. S., Carlin, B. P., & Reich, A. M. (2006). A spatial analysis of basketball shot chart data. The American Statistician, 60(1), 3-12.

Rosenbaum, Dan T. "Measuring How NBA Players Help their Teams Win." 2004. 82games.com, http://www.82games.com/comm30.htm.

Sarlis, V., & Tjortjis, C. (2020). Sports analytics—Evaluation of basketball players and team performance. Information Systems, 93, 101562.

Selmanović, A., Škegro, D., & Milanović, D. (2015). Basic characteristics of offensive modalities in the Euroleague and the NBA. Acta Kinesiologica, 9(2), 83-87.

Shah, R., & Romijnders, R. (2016). Applying deep learning to basketball trajectories. arXiv preprint arXiv:1608.03793.

Shi, Z., Moorthy, S., & Zimmermann, A. (2013, September). Predicting NCAAB match outcomes using ML techniques–some results and lessons learned. In ECML/PKDD 2013 Workshop on Machine Learning and Data Mining for Sports Analytics.

Sill, J. (2010, March). Improved NBA adjusted+/-using regularization and out-of-sample testing. In Proceedings of the 2010 MIT Sloan sports analytics conference.

Skinner, B. (2011). The price of anarchy in basketball. Journal of Quantitative Analysis in Sports, 6(1).

Skinner, B., & Guy, S. J. (2015). A method for using player tracking data in basketball to learn player skills and predict team performance. PloS one, 10(9), e0136393.

Specht, D. F. (1990). Probabilistic neural networks. Neural networks, 3(1), 109-118.

Štrumbelj, E., & Vračar, P. (2012). Simulating a basketball match with a homogeneous Markov model and forecasting the outcome. International Journal of Forecasting, 28(2), 532-542.

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.

Talagala, P. D., Hyndman, R. J., & Smith-Miles, K. (2021). Anomaly detection in high-dimensional data. Journal of Computational and Graphical Statistics, 30(2), 360-374.

Thakur, S., & Karthik, R. (2022). End of game shot selection for individual players in the NBA. Materials Today: Proceedings, 62, 4643-4650.

Tian, C., De Silva, V., Caine, M., & Swanson, S. (2019). Use of machine learning to automate the identification of basketball strategies using whole team player tracking data. Applied Sciences, 10(1), 24.

Valero-Carreras, D., Alcaraz, J., & Landete, M. (2023). Comparing two SVM models through different metrics based on the confusion matrix. Computers & Operations Research, 152, 106131.

Vaz de Melo, P. O., Almeida, V. A., Loureiro, A. A., & Faloutsos, C. (2012). Forecasting in the NBA and other team sports: Network effects in action. ACM Transactions on Knowledge Discovery from Data (TKDD), 6(3), 1-27.

Zhang, L., Lu, F., Liu, A., Guo, P., & Liu, C. (2016). Application of K-means clustering algorithm for classification of NBA guards. International Journal of Science and Engineering Applications, 5(1), 1-6.

Zhang, S., Lorenzo, A., Gómez, M. A., Liu, H., Gonçalves, B., & Sampaio, J. (2017). Players' technical and physical performance profiles and game-to-game variation in NBA. International Journal of Performance Analysis in Sport, 17(4), 466-483.

Zhang, S., Lorenzo, A., Gómez, M. A., Mateus, N., Gonçalves, B., & Sampaio, J. (2018). Clustering performances in the NBA according to players' anthropometric attributes and playing experience. Journal of sports sciences, 36(22), 2511-2520.

Zhang, S., Lorenzo, A., Woods, C. T., Leicht, A. S., & Gomez, M. A. (2019). Evolution of game-play characteristics within-season for the National Basketball Association. International Journal of Sports Science & Coaching, 14(3), 355-362.

Zhao, Y., Yang, R., Chevalier, G., Shah, R. C., & Romijnders, R. (2018). Applying deep bidirectional LSTM and mixture density network for basketball trajectory prediction. Optik, 158, 266-272.

ΔΙΑΔΙΚΤΥΟ

Ibm.com

ibm.com/topics/machine-learning

Nba.com

Basketball-reference.com

basketball-reference.com/about/per.html

nba.com/stats/help/glossary#pie

https://jr.nba.com/basketball-positions/

basketball-reference.com/nbl/seasons/1938.html

basketball-reference.com/leagues/NBA_2020_transactions.html

official.nba.com/rule-no-5-scoring-and-timing

jr.nba.com/3-point-shot

nba.com/news/history-mvp-award-winners