ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ

ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΔΕΙΚΤΕΣ ΓΙΑ ΤΗΝ ΑΞΙΟΛΟΓΗΣΗ ΤΗΣ ΑΠΟΔΟΣΗΣ ΠΑΙΚΤΩΝ ΚΑΙ ΟΜΑΔΩΝ ΣΕ ΑΓΩΝΕΣ ΜΠΑΣΚΕΤ ΚΑΙ ΠΑΡΑΓΟΝΤΕΣ ΠΟΥ ΤΟΥΣ ΕΠΗΡΕΑΖΟΥΝ

Αλέξανδρος Καμίτσης

Διπλωματική Εργασία

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

Πειραιάς

Ιούνιος 2023

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

ΔΕΙΚΤΕΣ ΓΙΑ ΤΗΝ ΑΞΙΟΛΟΓΗΣΗ ΤΗΣ ΑΠΟΔΟΣΗΣ ΠΑΙΚΤΩΝ ΚΑΙ ΟΜΑΔΩΝ ΣΕ ΑΓΩΝΕΣ ΜΠΑΣΚΕΤ ΚΑΙ ΠΑΡΑΓΟΝΤΕΣ ΠΟΥ ΤΟΥΣ ΕΠΗΡΕΑΖΟΥΝ

Αλέξανδρος Καμίτσης

Διπλωματική Εργασία

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

Πειραιάς

Ιούνιος 2023

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:

- Αναπληρωτής Καθηγητής Κωνσταντίνος Πολίτης (Επιβλέπων)
- Επίκουρος Καθηγητής Χαράλαμπος Ευαγγελάρας
- Επίκουρος Καθηγητής Ιωάννης Τριανταφύλλου

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

INDICES FOR EVALUATING PLAYER AND TEAM PERFORMANCE IN BASKETBALL GAMES, AND FACTORS AFFECTING THESE INDICES

by **Alexandros Kamitsis**

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

June 2023

Στους γονείς μου

Ιωάννη και Ευφροσύνη

Ευχαριστίες

Η συγκεκριμένη διπλωματική εργασία εκπονήθηκε στα πλαίσια του μεταπτυχιακού προγράμματος σπουδών «Εφαρμοσμένης Στατιστικής» του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης, του Πανεπιστημίου Πειραιώς. Δράττομαι της ευκαιρίας με την παρούσα παράγραφο, να ευχαριστήσω όλους τους ανθρώπους που συνέβαλαν στην πραγματοποίηση της παρούσας εργασίας. Αρχικά, θα ήθελα προφανώς να ευχαριστήσω θερμά τον Αναπληρωτή Καθηγητή κύριο Κωνσταντίνο Πολίτη, ο οποίος από τη πρώτη επικοινωνία μας ήταν πρόθυμος να με βοηθήσει και να με κατευθύνει σε κάτι εντελώς πρωτόγνωρο για εμένα, και παράλληλα να μου εμπιστευτεί ένα πάρα πολύ ενδιαφέρον θέμα, το οποίο συνδύαζε την στατιστική με την αγάπη μου για τον αθλητισμό. Θα ήθελα να τον ευχαριστήσω για τη συνεχή του καθοδήγηση και στήριξη σε όλη τη διάρκεια συγγραφής αυτής της μελέτης, καθώς και για όλες τις πολύτιμες υποδείξεις του, σχετικά με την κατάρτιση της διπλωματικής μου εργασίας, και όχι μόνο. Επιπροσθέτως, θα ήθελα να ευχαριστήσω από τα βάθη της καρδιάς μου τα αγαπημένα μου οικογενειακά πρόσωπα, για την ανιδιοτελή υλική και ηθική υποστήριξη που μου παρείχαν, καθ' όλη την διάρκεια των μεταπτυχιακών μου σπουδών και της ζωής μου γενικότερα. Ήταν και είναι πάντα εκεί για εμένα… Τέλος, θα ήθελα να ευχαριστήσω εξίσου θερμά το κύριο Αθανάσιο Φλωρόπουλο, μαθηματικό μου στο φροντιστήριο μέσης εκπαίδευσης, καθώς μου μετέφερε την αγάπη του για τα μαθηματικά, τον αναλυτικό τρόπο σκέψης για την επίλυση προβλημάτων, και γιατί ένα τεράστιο μερίδιο απ' ό,τι κατάφερα στις σπουδές μου και όπου έχω φτάσει ως σήμερα, οφείλεται σε εκείνον.

Περίληψη

Σε μια εποχή που είναι άρρηκτα συνδεδεμένη με τα δεδομένα, ο τομέας της ανάλυσης δεδομένων έχει αναδειχθεί σε ισχυρό εργαλείο σε διάφορους κλάδους, συμπεριλαμβανομένου και του αθλητισμού. Με την εκθετική ανάπτυξη της τεχνολογίας και τη διαθεσιμότητα τεράστιων ποσοτήτων δεδομένων, οι ομάδες και οι οργανισμοί είναι πλέον σε θέση να εξάγουν πολύτιμες πληροφορίες και να λαμβάνουν τεκμηριωμένες αποφάσεις για να αποκτήσουν ανταγωνιστικό πλεονέκτημα σε βάρος των αντιπάλων τους. Η ανάλυση δεδομένων στον αθλητισμό περιλαμβάνει τη συλλογή, την ερμηνεία και την οπτικοποίηση πολύπλοκων συνόλων δεδομένων, επιτρέποντας στις ομάδες να βελτιστοποιήσουν την απόδοση τους, να ενισχύσουν τις στρατηγικές ανάπτυξης των παικτών, να βελτιώσουν τις στρατηγικές του παιχνιδιού και τελικά να αποκτήσουν μεγαλύτερη επιτυχία εντός του αγωνιστικού χώρου. Από τα στατιστικά στοιχεία των παικτών έως τις μετρήσεις απόδοσης των αγώνων, η ανάλυση δεδομένων έχει φέρει επανάσταση στον τρόπο με τον οποίο παίζονται και αναλύονται τα αθλήματα, ανοίγοντας νέες ευκαιρίες για τις ομάδες και τους αθλητές να ξεκλειδώσουν τις πλήρεις δυνατότητες τους.

Στην παρούσα διπλωματική, χρησιμοποιώντας πραγματικά δεδομένα από την Ευρωλίγκα (EuroLeague), οποία θεωρείται η κορυφαία διασυλλογική διοργάνωση καλαθοσφαίρισης στην Ευρώπη, αναλύουμε με τεχνικές στατιστικής και εξετάζοντας όλες τις σεζόν από την δημιουργία της, ποιοι είναι οι σημαντικοί, προβλεπτικοί παράγοντες στη πρόκριση των ομάδων στα Playoffs και στο Final Four, καθώς και ποια είναι η συνεισφορά των καλύτερων παικτών με βάση το δείκτη αξιολόγησης της απόδοσης (Performance Index Rating - PIR) στη πορεία των ομάδων τους. Επιπροσθέτως, παρουσιάζουμε μια περιγραφική ανάλυση των μεταβλητών μας και απεικονίζουμε τα αποτελέσματα μας μέσω γραφικών παραστάσεων, γραφημάτων και πινάκων. Στη συνέχεια, εφαρμόζουμε μοντέλα λογιστικής παλινδρόμησης με σκοπό να βρούμε τις βασικές μεταβλητές που επηρεάζουν τη πρόκριση (ή μη) μιας ομάδας στις δύο φάσεις που εξετάζουμε. Τέλος, μέσω κατάλληλων τεχνικών μηχανικής μάθησης, διερευνούμε αν υπάρχουν ήδη υπάρχουσες κλάσεις στα δεδομένα μας με παρόμοια χαρακτηριστικά (clustering), και επιχειρούμε να φτιάξουμε αποτελεσματικά μοντέλα ταξινόμησης (classification), για να εξετάσουμε τη πρόκριση των ομάδων στις επιμέρους φάσης της διοργάνωσης.

Abstract

In a data-driven era, the field of data analytics has emerged as a powerful tool in various industries, including sports. With the exponential growth of technology and the availability of vast amounts of data, teams and organizations are now able to extract valuable insights and make informed decisions to gain a competitive advantage over their opponents. Data analytics in sports involves the collection, interpretation and visualization of complex data sets, allowing teams to optimize their performance, enhance player development strategies, improve game strategies and ultimately achieve greater success on the field. From player statistics to game performance metrics, data analytics has revolutionized the way sports are played and analyzed, opening up new possibilities for teams and athletes to unlock their full potential.

In this master thesis, using real data from the EuroLeague, which is considered the top club basketball competition in Europe, we use statistical techniques in order to analyze, by examining all the seasons since its creation, what are the important, predictive factors in the qualification of teams to the Playoffs and the Final Four, as well as what is the contribution of the five best players based on PIR to the progress of their teams. Additionally, we present a descriptive analysis of our variables and illustrate the results through graphs, charts, and tables. Then, we apply logistic regression models to find the key variables that affect a team's qualification (or not) in the two phases we examine. Finally, through appropriate machine learning techniques, we will investigate whether there are already existing clusters in our data with similar characteristics (clustering), and we try to build efficient classification models, to examine the qualification of teams in the individual phases of the competition.

Πίνακας περιεχομένων

Κατάλογος Σχημάτων και Πινάκων

ΚΕΦΑΛΑΙΟ 1ο - ΕΙΣΑΓΩΓΗ

Στην παρούσα διπλωματική εργασία, πραγματοποιείται μια ανάλυση της διοργάνωσης της Ευρωλίγκας (EuroLeague), η οποία θεωρείται η κορυφαία διασυλλογική διοργάνωση καλαθοσφαίρισης στην Ευρώπη, χρησιμοποιώντας έναν συνδυασμό στατιστικών τεχνικών και τεχνικών μηχανικής μάθησης. Η μελέτη πραγματοποιείται, έχοντας αντλήσει αριθμητικά δεδομένα από τη πρώτη σεζόν της διοργάνωσης (2000-2001), έως και τη περσινή σεζόν (2021-2022). Για να διασφαλιστεί μια ενδελεχής διερεύνηση, το σύνολο δεδομένων χωρίστηκε σε δύο διακριτές κατηγορίες, τις ομάδες που προκρίθηκαν στα Playoffs και στο Final Four, και αυτές που δε προκρίθηκαν. Αξιοποιώντας αυτές τις ξεχωριστές κλάσεις, αποσκοπούμε στο να παρέχουμε πληροφορίες σχετικά με τη επίδοση των συμμετεχουσών ομάδων, ρίχνοντας φως παράλληλα στους παράγοντες που συμβάλλουν στην επιτυχία και στις δύο αυτές φάσεις της Ευρωλίγκας.

Στο δεύτερο κεφάλαιο, γίνεται αρχικά μια εκτενής, ιστορική αναδρομή στην διοργάνωση. Στη συνέχεια, παρουσιάζονται οι ομάδες που συμμετείχαν διαχρονικά, οι μεταβλητές και τα δεδομένα που χρησιμοποιήσαμε στην ανάλυση μας, καθώς και ο τρόπος καταγραφής τους από το επίσημο site της EuroLeague. Τέλος, πραγματοποιείται και βιβλιογραφική επισκόπηση, με αναφορές σε υπάρχοντα επιστημονικά άρθρα, τα οποία είναι σχετικά με το θέμα της παρούσας εργασίας, καθώς και στα συμπεράσματα στα οποία αυτά καταλήγουν.

Στο τρίτο κεφάλαιο, καταγράφουμε αναλυτικά κάποια περιγραφικά μέτρα για τις επεξηγηματικές μεταβλητές μας σε σχέση με τις μεταβλητές απόκρισης Playoffs / Quarter-Finals (πρόκριση ή όχι στα playoffs) και Final 4 / Semi-Finals (πρόκριση στο Final Four). Παράλληλα, παρουσιάζουμε μια διαγραμματική απεικόνιση των δεδομένων των αναλύσεων μας, ώστε τα αποτελέσματα να γίνουν ακόμη πιο κατανοητά και συγκρίσιμα από τον αναγνώστη.

Στο τέταρτο κεφάλαιο, αρχικά γίνεται έλεγχος κανονικότητας για τις ποσοτικές μεταβλητές μας, για κάθε ανάλυση που πραγματοποιήθηκε ξεχωριστά στο τρίτο κεφάλαιο. Στη συνέχεια, υπολογίζονται οι συντελεστές συσχέτισης των μεταβλητών μας, για να δούμε τι σχέση υπάρχει μεταξύ τους (και αν υπάρχει), και τέλος, πραγματοποιούνται και έλεγχοι για την ισότητα των μέσων τιμών των πόντων και του δείκτη PIR, των δύο δειγμάτων (ομάδες που προκρίθηκαν και ομάδες που δεν προκρίθηκαν στις επιμέρους φάσεις), όπου αυτό κρίνεται απαραίτητο. Για όλους τους προαναφερθέντες ελέγχους, το επίπεδο σημαντικότητας με το οποίο εργαστήκαμε, ισούταν με 5%.

Στο πέμπτο κεφάλαιο, αποσκοπούμε στο να εξετάσουμε ποιες από τις μεταβλητές μας παίζουν καθοριστικό ρόλο στο αν μια ομάδα προκρίνεται ή όχι, στις δύο φάσεις της διοργάνωσης που εξετάζουμε. Για να γίνει αυτό, προσαρμόζουμε τα κατάλληλα γενικευμένα γραμμικά μοντέλα, με μεταβλητές απόκρισης τις κατηγορικές μεταβλητές Playoffs / Quarter-Finals και Final 4 / Semi-Finals , οι οποίες όπως έχουμε αναφέρει στα προηγούμενα κεφάλαια, αποτελούν ενδείξεις για το αν οι ομάδες προκρίθηκαν (ή όχι) στις αντίστοιχες φάσεις της διοργάνωσης, και στη συνέχεια ερμηνεύουμε τα αποτελέσματα μας.

Στο έκτο κεφάλαιο, μέσω της χρήσης τεχνικών μηχανικής μάθησης και της κατάλληλης επεξεργασίας των δεδομένων μας, αποσκοπούμε στο να εξάγουμε χρήσιμη γνώση/πληροφορία. Ειδικότερα, αφού πρώτα καταλήγουμε στις κατάλληλες μεταβλητές (feature selection) που είναι χρήσιμες για τους αλγορίθμους μας, χρησιμοποιούμε τεχνικές ομαδοποίησης (clustering) για να εντοπίσουμε μοτίβα και ομάδες παρατηρήσεων με παρόμοια χαρακτηριστικά μέσα στα δεδομένα μας, καθώς και τεχνικές κατηγοριοποίησης (classification), προκειμένου να προσαρμόσουμε κατάλληλα μοντέλα ταξινόμησης για το αν μια ομάδα θα προκρινόταν (ή όχι) στις δύο φάσεις της διοργάνωσης που εξετάζουμε.

Στο έβδομο και τελευταίο κεφάλαιο, γίνεται μια σύνοψη των ευρημάτων των προηγούμενων κεφαλαίων και παρουσιάζονται τα συμπεράσματα της παρούσας διπλωματικής εργασίας.

ΚΕΦΑΛΑΙΟ 2ο

2.1 Η διοργάνωση της EuroLeague

2.1.1 Ιστορική αναδρομή στη διοργάνωση της EuroLeague

Η Ευρωλίγκα (EuroLeague), γνωστή και ως «Turkish Airlines EuroLeague» για χορηγικούς λόγους, θεωρείται η κορυφαία, διασυλλογική διοργάνωση συλλόγων μπάσκετ/καλαθοσφαίρισης στην Ευρώπη. Το πρωτάθλημα αποτελείται από 18 ομάδες, εκ των οποίων οι 16 λαμβάνουν μακροχρόνιες άδειες, τα λεγόμενα «κλειστά συμβόλαια» και wild cards[1].

Το πρωτάθλημα διοργανώθηκε για πρώτη φορά το 1958, υπό την αιγίδα της FIBA[2], με τίτλο «Κύπελλο Πρωταθλητριών Ευρώπης της FIBA» (FIBA European Champions Cup), στη πορεία μετονομάστηκε σε «FIBA EuroLeague» ενώ από τη σεζόν 2000-2001, η νεοσύστατη, ιδιωτική εταιρία Euroleague Basketball, σε συνεργασία με ομάδες που ήταν μέλη μιας πανευρωπαϊκής κοινοπραξίας των κορυφαίων επαγγελματικών συλλόγων καλαθοσφαίρισης, που ονομάζεται «Ένωση Ευρωπαϊκών Επαγγελματικών Ομάδων Καλαθοσφαίρισης» (ULEB) ανέλαβαν τη διοργάνωση του, έως τη σεζόν 2004-2005. Στη συνέχεια, από την σεζόν 2004-2005 έως την σεζόν 2010-2011, το πρωτάθλημα διοργανώθηκε από την FIBA Europe. Τέλος, από την σεζόν 2010-2011 ονομάζεται πλέον «Turkish Airlines Euroleague» (TAE) για χορηγικούς λόγους.

Το Κύπελλο Πρωταθλητριών Ευρώπης της FIBA και η EuroLeague, θεωρούνται η ίδια διοργάνωση, με την αλλαγή του ονόματος να αποτελεί απλώς ένα re-branding.

Η ομάδα με τις περισσότερες κατακτήσεις της διοργάνωσης είναι η Real Madrid με έντεκα τίτλους, η οποία είναι και η πιο πρόσφατη πρωταθλήτρια, καθώς νίκησε τον Olympiacos στον τελικό του Final Four του 2023, που διεξήχθη στο Κάουνας της Λιθουανίας. Τον τίτλο έχουν κατακτήσει 22 σύλλογοι, 14 εκ των οποίων έχουν κερδίσει τον τίτλο περισσότερες από μία φορές.

[1] Wild card είναι μια θέση σε διοργάνωση, η οποία απονέμεται σε άτομο ή ομάδα που αποτυγχάνει να προκριθεί με τον κανονικό τρόπο, όπως για παράδειγμα, έχοντας υψηλή κατάταξη ή κερδίζοντας ένα προκριματικό στάδιο. Σε ορισμένες εκδηλώσεις/διοργανώσεις, επιλέγονται ελεύθερα από τους διοργανωτές.
(Πηγή : https://en.wikipedia.org/wiki/Wild_card_(sports))

[2] The International Basketball Federation, originally known as the Fédération internationale de basket-ball amateur.

Σχήμα 2.1 : Οι εναλλαγές στο λογότυπο της διοργάνωσης έως και σήμερα.

2.1.2 Δημιουργία της FIBA SuproLeague και σχίσμα στο Ευρωπαϊκό μπάσκετ

Η FIBA είχε χρησιμοποιήσει προηγουμένως το όνομα της EuroLeague για τη διοργάνωση από το 1996, αλλά δεν είχε ποτέ κατοχυρώσει το σήμα κατατεθέν. Καθώς η FIBA δεν είχε νομική προσφυγή για τη χρήση του ονόματος, ξεκίνησε ένα νέο πρωτάθλημα με το όνομα «FIBA SuproLeague». Η σεζόν 2000–2001 ξεκίνησε με δύο πλέον κορυφαίες ευρωπαϊκές επαγγελματικές συλλογικές διοργανώσεις μπάσκετ, την FIBA SuproLeague (που μετονομάστηκε από FIBA EuroLeague) και την Euroleague. Αποτέλεσμα αυτού, ήταν οι κορυφαίοι ευρωπαϊκοί σύλλογοι να χωριστούν στα δύο πρωταθλήματα. Ο Panathinaikos, η Maccabi Tel Aviv, η CSKA Moscow και η Efes Pilsen[3] παρέμειναν στη FIBA, ενώ ο Olympiacos, η Kinder Bologna, η Real Madrid Teka, η FC Barcelona, η Paf Wennington Bologna, η Zalgiris Kaunas, η Benetton Treviso, η ΑΕΚ και η Tau Ceramica εντάχθηκαν στην Euroleague.

Τον Μάιο του 2001, η Ευρώπη είχε δύο πρωταθλητές ηπείρου, τη Maccabi Tel Aviv της FIBA SuproLeague και την Kinder Bologna της Euroleague. Και οι δύο οργανισμοί συνειδητοποίησαν την ανάγκη να καταλήξουν σε μια ενιαία διοργάνωση. Για το λόγο αυτό, η Euroleague Basketball διαπραγματεύτηκε όρους και υπαγόρευσε τις διαδικασίες, με τις οποίες η FIBA συμφώνησε, θέτοντας και δικούς της όρους παράλληλα. Ως αποτέλεσμα, η ευρωπαϊκή διοργάνωση συλλόγων ενσωματώθηκε πλήρως κάτω από την ομπρέλα της Euroleague Basketball και οι ομάδες που αγωνίστηκαν στη FIBA SuproLeague κατά τη διάρκεια της σεζόν 2000–2001 εντάχθηκαν επίσης σε αυτήν.

Η εξουσία στο ευρωπαϊκό επαγγελματικό μπάσκετ ήταν διχασμένη μεταξύ συλλόγων και χωρών. Η FIBA παρέμεινε επικεφαλής των διοργανώσεων των εθνικών ομάδων (όπως το Ευρωμπάσκετ, το Παγκόσμιο Κύπελλο και οι Θερινοί Ολυμπιακοί Αγώνες), ενώ η Euroleague Basketball ανέλαβε τις ευρωπαϊκές επαγγελματικές διοργανώσεις συλλόγων. Από εκείνο το σημείο και μετά, οι διοργανώσεις «Korać Cup» και «Saporta Cup» της FIBA διήρκεσαν μία ακόμη σεζόν και στη συνέχεια η Euroleague Basketball ξεκίνησε το «ULEB Cup», τώρα γνωστό ως «EuroCup».

[3] Σήμερα Anadolu Efes.

2.1.3 Ιστορική αναδρομή στα διαφορετικά format της διοργάνωσης στο πέρασμα των χρόνων

Η διοργάνωση, όπως προαναφέρθηκε, ξεκίνησε την σεζόν 1957-58 (όλοι οι αγώνες διεξήχθησαν το ημερολογιακό έτος 1958) με την ονομασία «Κύπελλο Πρωταθλητριών», κατά τα πρότυπα της αντίστοιχης ποδοσφαιρικής διοργάνωσης της UEFA. Μέχρι τη σεζόν 1990-91, έπαιρναν μέρος μόνο οι πρωταθλήτριες ομάδες κάθε χώρας. Από την επόμενη σεζόν, άρχισαν να προστίθενται οι δευτεραθλήτριες ή/και τριταθλήτριες ομάδες από τις καλαθοσφαιρικά προηγμένες χώρες και έτσι η διοργάνωση άλλαξε μορφή. Τη σεζόν 1996-97 καταργήθηκαν οι προκριματικοί νοκ-άουτ αγώνες και οι ομάδες ξεκινούσαν την πορεία τους από την αρχική φάση των ομίλων, ταυτόχρονα όμως έχασαν το δικαίωμα συμμετοχής οι χώρες όπου το άθλημα δε γνώριζε ιδιαίτερη ανάπτυξη.

Την σεζόν 1965-66 δοκιμάστηκε για πρώτη φορά η διαδικασία Final Four[4] ανάμεσα στις τέσσερις ομάδες που θα έφταναν στη φάση των ημιτελικών, κατά τα πρότυπα του αμερικανικού κολεγιακού πρωταθλήματος. Πολύ σύντομα η ιδέα εγκαταλείφθηκε, για να επανέλθει πολλά χρόνια αργότερα (1987-88) και να καθιερωθεί έως και σήμερα. Μόνη εξαίρεση υπήρξε το 2001, όταν η διοργάνωση της FIBA περιλάμβανέ Final Four, ενώ η αντίστοιχη της Euroleague Basketball όχι.

Ένα σημαντικό χαρακτηριστικό των πρόσφατων διοργανώσεων, έχει να κάνει με τη δέσμευση κλειστών συμβολαίων ανάμεσα στη Euroleague Basketball και κάποιους συλλόγους που θεωρούνται ιδιαίτερα «εμπορικοί». Αυτό πρακτικά σημαίνει πως κάποιες ομάδες έχουν δικαίωμα συμμετοχής ασχέτως των επιδόσεών τους στα εθνικά πρωταθλήματα. Κάτι τέτοιο έχει συμβεί και με ελληνικούς συλλόγους, όπως π.χ. με τις ομάδες Olympiacos και ΑΕΚ που είχαν αγωνισθεί στο παρελθόν στην Ευρωλίγκα, ενώ δεν είχαν καταταγεί σε υψηλή θέση στο εθνικό πρωτάθλημα της προηγούμενης αγωνιστικής περιόδου.

- Σεζόν 2000-2001

  Συμμετείχαν συνολικά 24 ομάδες από 14 χώρες. Η πρώτη φάση ήταν η κανονική περίοδος (Regular Season), στην οποία οι διαγωνιζόμενες ομάδες κληρώθηκαν σε τέσσερις ομίλους, ο καθένας από τους οποίους περιείχε έξι ομάδες. Κάθε ομάδα έπαιξε με όλες τις άλλες ομάδες του ομίλου της εντός και εκτός έδρας, με αποτέλεσμα 10 παιχνίδια για κάθε ομάδα στην πρώτη φάση. Οι 4 πρώτες ομάδες σε κάθε όμιλο προκρίθηκαν στον επόμενο γύρο, το Top 16. Οι 16 ομάδες τέθηκαν

[4] Στον αθλητισμό, ο όρος «φάιναλ φορ» αναφέρεται στις τέσσερις τελευταίες ομάδες που απομένουν σε ένα τουρνουά πλέι οφ. Συνήθως οι ομάδες διαγωνίζονται στα δύο παιχνίδια ενός ημιτελικού (προτελευταίου) γύρου ενός τουρνουά ενός αποκλεισμού. Από αυτές τις ομάδες, οι δύο που κερδίζουν στον ημιτελικό γύρο παίζουν άλλο ένα παιχνίδι και ο νικητής είναι ο πρωταθλητής του τουρνουά. Σε ορισμένα τουρνουά, οι δύο ομάδες που χάνουν στον ημιτελικό γύρο, ανταγωνίζονται για την τρίτη θέση σε ένα παιχνίδι παρηγοριάς.
(Πηγή : https://en.wikipedia.org/wiki/Final_four)

μεταξύ τους αντιμέτωπες, σε μια σειρά best-of-three[5] αγώνων. Οι 8 κορυφαίες ομάδες προκρίθηκαν στα προημιτελικά και τέθηκαν μεταξύ τους αντιμέτωπες, πάλι σε μια σειρά best-of-three αγώνων. Οι 4 νικήτριες ομάδες προκρίθηκαν στα ημιτελικά και τέθηκαν αντιμέτωπες σε μια σειρά best-of-five[6] αγώνων. Οι 2 νικήτριες ομάδες των ημιτελικών, έπαιξαν μεταξύ τους σε μια σειρά τελικών best-of-five για το τίτλο του πρωταθλητή Ευρώπης.

- Σεζόν 2001-2002

Συμμετείχαν 41 ομάδες συνολικά με τους προκριματικούς γύρους και 32 ομάδες στην πρώτη φάση των ομίλων (οι πρωταθλητές εγχώριων πρωταθλημάτων από τα καλύτερα πρωταθλήματα και ένας μεταβλητός αριθμός άλλων συλλόγων από τα σημαντικότερα εθνικά πρωταθλήματα), αγωνίστηκαν σε σύστημα τουρνουά. Η πρώτη φάση ήταν η κανονική περίοδος, στην οποία οι διαγωνιζόμενες ομάδες κληρώθηκαν σε τέσσερις ομίλους, με τον καθένα να περιέχει οκτώ ομάδες. Κάθε ομάδα έπαιξε με όλες τις άλλες ομάδες του ομίλου της, εντός και εκτός έδρας, με αποτέλεσμα 14 παιχνίδια για κάθε ομάδα στην πρώτη φάση. Οι 4 πρώτες ομάδες από κάθε όμιλο προκρίθηκαν στον επόμενο γύρο, το Top 16. Οι 16 ομάδες χωρίστηκαν σε τέσσερις ομίλους, των τεσσάρων ομάδων ο καθένας. Κάθε ομάδα έπαιξε με κάθε άλλη ομάδα του ομίλου της δύο φορές, μία εντός και μία εκτός έδρας. Οι πρώτες ομάδες καθενός από τους τέσσερις ομίλους προκρίθηκαν στο Final Four.

- Σεζόν 2002-2003

Συμμετείχαν 24 ομάδες από 13 χώρες. Η πρώτη φάση ήταν η κανονική περίοδος, στην οποία οι διαγωνιζόμενες ομάδες κληρώθηκαν σε τρεις ομίλους, ο καθένας από τους οποίους περιείχε οκτώ ομάδες. Κάθε ομάδα έπαιξε με όλες τις άλλες ομάδες του ομίλου της εντός και εκτός έδρας, με αποτέλεσμα 14 παιχνίδια για κάθε ομάδα στην πρώτη φάση. Οι 5 πρώτες ομάδες σε κάθε όμιλο και η καλύτερη στην έκτη θέση προκρίθηκαν στον επόμενο γύρο. Οι ομάδες που προκρίθηκαν, χωρίστηκαν σε τέσσερις νέους ομίλους των τεσσάρων ομάδων ο καθένας, και πάλι υιοθετήθηκε ένα σύστημα round robin[7], με αποτέλεσμα 6 παιχνίδια για τη καθεμία ομάδα, με την πρώτη ομάδα κάθε ομίλου να προκρίνεται στο Final Four. Τα

[5] Η πρώτη ομάδα που θα φτάσει τις δύο νίκες παίρνει τη πρόκριση.

[6] Η πρώτη ομάδα που θα φτάσει τις τρεις νίκες παίρνει τη πρόκριση.

[7] Ένα τουρνουά στρογγυλής περιπέτειας (ή τουρνουά all-go-away) είναι ένας διαγωνισμός στον οποίο κάθε διαγωνιζόμενος συναντά κάθε άλλο συμμετέχοντα, συνήθως με τη σειρά. Ένα round robin τουρνουά έρχεται σε αντίθεση με ένα τουρνουά αποκλεισμού, στο οποίο οι συμμετέχοντες/ομάδες αποκλείονται μετά από έναν ορισμένο αριθμό ηττών.
(Πηγή : https://en.wikipedia.org/wiki/Round-robin_tournament)

tiebreakers[8] για την επίλυση πιθανών ισοβαθμιών ήταν πανομοιότυπα με αυτά που χρησιμοποιήθηκαν με την ολοκλήρωση της κανονικής περιόδου[9].

- Σεζόν 2003-2004

Την σεζόν 2003-2004 συμμετείχαν 24 ομάδες από 13 χώρες. Η πρώτη φάση και εδώ ήταν η κανονική περίοδος, στην οποία οι διαγωνιζόμενες ομάδες κληρώθηκαν σε τρεις ομίλους, ο καθένας από τους οποίους περιείχε οκτώ ομάδες. Κάθε ομάδα έπαιξε με όλες τις άλλες ομάδες του ομίλου της εντός και εκτός έδρας, με αποτέλεσμα 14 παιχνίδια για κάθε ομάδα στην πρώτη φάση. Οι 5 πρώτες ομάδες σε κάθε όμιλο και η καλύτερη στην έκτη θέση προκρίθηκαν στον επόμενο γύρο. Οι ομάδες που προκρίθηκαν, χωρίστηκαν σε τέσσερις ομίλους των τεσσάρων ομάδων, και πάλι υιοθετήθηκε ένα σύστημα round robin που είχε ως αποτέλεσμα 6 παιχνίδια για τη κάθε ομάδα, με την πρώτη κάθε ομίλου να προκρίνεται στο Final Four. Τα tiebreakers ήταν πανομοιότυπα με αυτά που χρησιμοποιήθηκαν στην κανονική σεζόν και εδώ.

Αυτή ήταν η τελευταία σεζόν στην οποία οι ομάδες προκρίθηκαν απευθείας από το Top 16 στο Final Four. Ένας προημιτελικός γύρος εισήχθη από τη σεζόν 2004-2005.

- Σεζόν 2004-2005

Επίσης 24 διαγωνιζόμενες ομάδες από 13 χώρες. Η πρώτη φάση ήταν η κανονική περίοδος, στην οποία οι διαγωνιζόμενες ομάδες κληρώθηκαν σε τρεις ομίλους, ο καθένας από τους οποίους περιείχε οκτώ ομάδες. Κάθε ομάδα έπαιξε με όλες τις άλλες ομάδες του ομίλου της εντός και εκτός έδρας, με αποτέλεσμα 14 παιχνίδια για κάθε ομάδα στην πρώτη φάση. Οι 5 πρώτες ομάδες σε κάθε όμιλο και η καλύτερη στην έκτη θέση προκρίθηκαν στον επόμενο γύρο. Οι ομάδες που πήραν τη πρόκριση, χωρίστηκαν σε τέσσερις ομίλους των τεσσάρων ομάδων ο καθένας, και πάλι υιοθετήθηκε ένα σύστημα round robin, με αποτέλεσμα 6 παιχνίδια για τη καθεμία ομάδα, με τις δύο κορυφαίες ομάδες κάθε ομίλου να προκρίνονται στα προημιτελικά. Τα tiebreakers ήταν πανομοιότυπα με αυτά που χρησιμοποιήθηκαν στην κανονική σεζόν. Κάθε προημιτελικός γύρος ήταν μια σειρά best-of-three

[8] Στα παιχνίδια και τα αθλήματα, ένα tiebreak χρησιμοποιείται για τον καθορισμό ενός νικητή μεταξύ των παικτών ή των ομάδων που ισοβαθμούν στο τέλος ενός διαγωνισμού ή ενός συνόλου διαγωνισμών.
(Πηγή : https://en.wikipedia.org/wiki/Tiebreaker)

[9] Εάν ένας ή περισσότεροι σύλλογοι ισοδυναμούσαν στα ρεκόρ νίκης-ήττας στο τέλος της κανονικής περιόδου, τα tiebreakers εφαρμόζονταν με την ακόλουθη σειρά:

1. Ρεκόρ σε αγώνες μεταξύ των ισόπαλων ομάδων.
2. Συνολική διαφορά πόντων στα παιχνίδια μεταξύ των ισόπαλων ομάδων.
3. Συνολική διαφορά πόντων σε όλους τους αγώνες των ομίλων (πρώτο tiebreak αν οι ισόπαλοι σύλλογοι δεν ήταν στον ίδιο όμιλο).
4. Βαθμοί σε όλους τους αγώνες των ομίλων.
5. Άθροισμα των πόντων που σημείωσαν και των πόντων που δέχθηκαν σε κάθε αγώνα του ομίλου.

αγώνων μεταξύ μιας ομάδας που είχε τερματίσει πρώτη σε έναν όμιλο στο Top 16, και μιας ομάδας που είχε τερματίσει δεύτερη σε έναν διαφορετικό όμιλο, με την πρώτη ομάδα να λαμβάνει το πλεονέκτημα έδρας. Οι 4 ομάδες που κέρδισαν τη προημιτελική σειρά, πήραν τη πρόκριση για τα Final Four.

- Σεζόν 2005-2006

Την σεζόν 2005-2006, το format της διοργάνωσης ήταν ακριβώς το ίδιο με τη προηγούμενη σεζόν.

- Σεζόν 2006-2007

Την σεζόν 2006-2007, το format της διοργάνωσης ήταν ακριβώς το ίδιο με τη προηγούμενη σεζόν.

- Σεζόν 2007-2008

Την σεζόν 2007-2008, το format της διοργάνωσης ήταν ακριβώς το ίδιο με τη προηγούμενη σεζόν.

- Σεζόν 2008-2009

Στη σεζόν συμμετείχαν 24 ομάδες από 13 χώρες και έγιναν αλλαγές στη μορφή για δύο από τις φάσεις της διοργάνωσης, την κανονική περίοδο και τους προημιτελικούς, σε σχέση με τα προηγούμενα χρόνια. Πρώτη φάση ήταν η κανονική περίοδος, στην οποία οι ομάδες κληρώθηκαν σε τέσσερις ομίλους, ο καθένας από τους οποίους περιείχε έξι ομάδες. Κάθε ομάδα έπαιξε με όλες τις άλλες ομάδες του ομίλου της, εντός και εκτός έδρας, με αποτέλεσμα 10 παιχνίδια για κάθε ομάδα στην πρώτη φάση. Οι 4 πρώτες ομάδες κάθε ομίλου προκρίθηκαν στον επόμενο γύρο. Αυτή ήταν η πρώτη χρονιά για τη συγκεκριμένη μορφή (τις προηγούμενες σεζόν, οι ομάδες χωρίζονταν σε τρεις ομίλους των οκτώ ομάδων, με τις 5 πρώτες ομάδες σε κάθε όμιλο συν την καλύτερη ομάδα στην έκτη θέση να προκρίνονται). Εάν δύο ή περισσότεροι σύλλογοι τερμάτιζαν ισόβαθμοι με ρεκόρ νικών-ηττών, τα tiebreakers εφαρμόζονταν με την σειρά που αναφέρθηκε παραπάνω (βλ. υποσημείωση 9). Οι ομάδες που προκρίθηκαν, χωρίστηκαν στη συνέχεια σε τέσσερις ομίλους των τεσσάρων ομάδων ο καθένας, και πάλι υιοθετήθηκε ένα σύστημα round-robin με αποτέλεσμα 6 παιχνίδια για τη καθεμία, με τις 2 πρώτες ομάδες κάθε ομίλου να προκρίνονται στους προημιτελικούς. Τα tiebreakers είναι πανομοιότυπα με αυτά που χρησιμοποιούνται στην κανονική σεζόν. Στους προημιτελικούς, οι κορυφαίες ομάδες από κάθε όμιλο των Top 16 έπαιξαν με τις δεύτερες ομάδες από διαφορετικό όμιλο, σε μια σειρά best-of-five, με τους νικητές αυτής της σειράς να προκρίνονται στο Final Four. Αυτή ήταν η πρώτη σεζόν στην οποία οι προημιτελικοί ήταν σειρές best-of-five, καθώς προηγουμένως ήταν σειρές best-of-three.

- Σεζόν 2009-2010

Στην κανονική περίοδο, συμμετείχαν 24 ομάδες από 13 χώρες. Αυτή η σεζόν σηματοδότησε την πρώτη φορά από τη σεζόν 2001-2002 που πραγματοποιήθηκαν προκριματικοί γύροι, για τον προσδιορισμό των δύο τελευταίων ομάδων που θα συμμετείχαν στην κανονική περίοδο. 8 ομάδες συνεπώς αγωνίστηκαν σε προκριματικούς γύρους, σε ένα τουρνουά νοκ-άουτ που αποτελούταν από αγώνες δύο σκελών. Οι 4 που προκρίνονταν από το πρώτο προκριματικό γύρο, αναμετρόντουσαν μεταξύ τους για τον δεύτερο προκριματικό γύρο, με τους 2 νικητές να παίρνουν τις δύο τελευταίες θέσεις για τη κανονική περίοδο της EuroLeague. Όλοι οι ηττημένοι σύλλογοι στους προκριματικούς γύρους υποβιβάστηκαν στη δεύτερη τη τάξει διοργάνωση της Euroleague Basketball, το EuroCup. Στη κανονική περίοδο, οι ομάδες κληρώθηκαν σε τέσσερις ομίλους, ο καθένας από τους οποίους περιείχε έξι ομάδες. Κάθε ομάδα έπαιξε με όλες τις άλλες ομάδες του ομίλου της, εντός και εκτός έδρας, με αποτέλεσμα 10 παιχνίδια για κάθε ομάδα στην πρώτη φάση. Οι 4 πρώτες ομάδες κάθε ομίλου προκρίθηκαν στον επόμενο γύρο, στο Top 16. Εκεί, κληρώθηκαν σε τέσσερις ομίλους των τεσσάρων ομάδων ο καθένας, όπου και πάλι υιοθετήθηκε ένα σύστημα round-robin με αποτέλεσμα 6 παιχνίδια για τη καθεμία, με τις 2 πρώτες ομάδες κάθε ομίλου να προκρίνονται στους προημιτελικούς, σε σειρές best-of-five. Κατά τα γνωστά, οι νικητές αυτής της σειράς προκρίθηκαν στο Final Four.

- Σεζόν 2010-2011

Συμμετείχαν 24 ομάδες από 14 χώρες. Οι προκριματικοί γύροι αποτελούνταν από τρεις γύρους, οι οποίοι έγιναν σε σειρές με αναμετρήσεις εντός και εκτός έδρας. Ο πρώτος προκριματικός γύρος είχε 16 ομάδες που έπαιζαν σε σειρές δύο αγώνων (εντός και εκτός έδρας), οι 8 νικητές των οποίων προκρίθηκαν στον δεύτερο προκριματικό γύρο. Οι ηττημένοι αυτής της σειράς των δύο αγώνων έπαιξαν στην διοργάνωση του Eurocup. Οι 4 νικητές του δεύτερου προκριματικού γύρου μπήκαν στον τελευταίο προκριματικό γύρο, από τον οποίο προέκυψαν οι 2 ομάδες που θα έπαιρναν μέρος στη κανονική περίοδο. Η δομή της διοργάνωσης από την κανονική περίοδο μέχρι και το Final Four, ήταν η ακριβώς ίδια με αυτή της προηγούμενης σεζόν.

- Σεζόν 2011-2012

Συμμετείχαν 24 ομάδες από 13 χώρες. Όπως και τη προηγούμενη σεζόν, 16 ομάδες συνολικά συμμετείχαν στους προκριματικούς γύρους. Οι προκριματικοί γύροι αποτελούνταν από δύο νοκ-άουτ τουρνουά σε μορφή Final Eight. Οι 2 νικήτριες ομάδες προκρίθηκαν στην κανονική περίοδο. Η δομή της διοργάνωσης από την κανονική περίοδο μέχρι και το Final Four ήταν και σε αυτή τη περίπτωση η ίδια με τη προηγούμενη σεζόν. Είναι γνωστό πως ομάδες που προέρχονταν από τον ίδιο όμιλο της κανονικής περιόδου, δεν συνέπιπταν στον ίδιο όμιλο στη φάση

Top 16. Άρχισε να γίνεται πλέον και μια προσπάθεια να μη συμπίπτουν ούτε ομάδες από την ίδια χώρα στον ίδιο όμιλο. Επιπροσθέτως, ομάδες από την ίδια πόλη όπως η Anadolu Efes, η Fenerbahçe Ülker και η Galatasaray Medical Park από την Κωνσταντινούπολη, ή ο Olympiacos και ο Panathinaikos από την Αθήνα, όπως και ομάδες που είχαν κοινή έδρα, δεν θα έπαιζαν εντός έδρας αγώνες την ίδια αγωνιστική.

- Σεζόν 2012-2013

Συνολικά 24 ομάδες από 13 χώρες συμμετείχαν στην διοργάνωση. Οι προκριματικοί γύροι διεξήχθησαν σαν ένα τουρνουά νοκ-άουτ με μονούς αγώνες, αποτελούμενο από οκτώ ομάδες. Ο νικητής προκρίθηκε στην κανονική περίοδο της EuroLeague. Η κανονική περίοδος έχει την ίδια μορφή με τις προηγούμενες σεζόν. Στον γύρο Top 16 πλέον, οι 16 ομάδες που είχαν προκριθεί από τη πρώτη φάση χωρίστηκαν σε δύο ομίλους των οκτώ ομάδων. Οι 4 πρώτοι κάθε ομίλου προκρίθηκαν στον προημιτελικό γύρο (όπου ακολουθήθηκε η ίδια δομή όπως παλαιότερα) και στη συνέχεια, οι νικητές του γύρου αυτού, πέρασαν στο Final Four.

- Σεζόν 2013-2014

Συμμετείχαν 24 ομάδες από 12 χώρες. Η δομή της διοργάνωσης από το προκριματικό γύρο μέχρι και το Final Four ήταν η ίδια ακριβώς με τη προηγούμενη σεζόν, με μόνη αλλαγή πως οι ομάδες που αποκλείστηκαν στην κανονική περίοδο, υποβιβάστηκαν στο EuroCup.

- Σεζόν 2014-2015

Στη σεζόν 2014-2015, συμμετείχαν 24 ομάδες από 12 χώρες. Η δομή της διοργάνωσης από το προκριματικό γύρο μέχρι και το Final Four ήταν η ίδια ακριβώς με τη προηγούμενη σεζόν.

- Σεζόν 2015-2016

Συνολικά 24 ομάδες από 12 χώρες συμμετείχαν στην σεζόν 2015-2016, με το ίδιο format στη διοργάνωση όπως τις προηγούμενες σεζόν.

- Σεζόν 2016-2017

Τον Νοέμβριο του 2015, η Euroleague Basketball και η πολυεθνική εταιρεία IMG, ειδικευμένη σε αθλητικά, μόδα και μάρκετινγκ υπέγραψαν δεκαετές συμβόλαιο συνεργασίας, αλλάζοντας τη δομή της διοργάνωσης προσφέροντας παράλληλα στις ομάδες και περισσότερα έσοδα. Αποφασίστηκε πως από τις 16 ομάδες που θα συμμετέχουν, οι 11 θα έχουν συμβόλαιο τύπου Α, δηλαδή σταθερή

παρουσία στο θεσμό για τα επόμενα δέκα χρόνια. Αυτές οι ομάδες είναι ο Panathinaikos, ο Olympiacos, η Real Madrid, η Barcelona, η Baskonia, η CSKA Moscow, η Fenerbahce, η Anadolu Efes, η Emporio Armani Milano, η Maccabi Tel Aviv και η Zalgiris Kaunas. Οι υπόλοιπες πέντε άδειες, οι οποίες θα εξασφαλίζονται μέσω της παρουσίας στις διοργανώσεις των χωρών, θα διανέμονται ως εξής : οι νικητές της Αδριατικής Λίγκας[10], του πρωταθλήματος Γερμανίας, της VTB League[11] (η φιναλίστ σε περίπτωση που το πρωτάθλημα κατακτήσει η CSKA Moscow), ο κάτοχος του Eurocup της προηγούμενης σεζόν, καθώς και μια ακόμα ομάδα από την Ισπανία. (Καλλιακμάνης, 2020)

Στη Euroleague λοιπόν από τη σεζόν 2016-2017, συμμετείχαν 16 ομάδες, όπου όλες τέθηκαν αντιμέτωπες μεταξύ τους σε μία κανονική περίοδο, σε τουρνουά round-robin, με 30 αγωνιστικές. Όταν περισσότερες από δύο ομάδες ήταν ισόβαθμες, η κατάταξη καθοριζόταν λαμβάνοντας υπόψιν τις νίκες που σημειώθηκαν στους αγώνες που έγιναν μόνο μεταξύ τους. Εάν η ισοβαθμία συνεχιζόταν μεταξύ ορισμένων, αλλά όχι όλων, των ομάδων, η κατάταξη των ομάδων που εξακολουθούσαν να είναι ισόβαθμες καθοριζόταν λαμβάνοντας και πάλι υπόψιν τις νίκες στα παιχνίδια που είχαν γίνει μόνο μεταξύ τους και επαναλαμβάνοντας την ίδια διαδικασία μέχρι να επιλυόταν η ισοβαθμία. Εάν η ισοβαθμία συνεχιζόταν, η κατάταξη θα καθοριζόταν από τη διαφορά πόντων υπέρ και κατά, στα παιχνίδια που έχουν γίνει μόνο μεταξύ των ομάδων που ήταν ισόβαθμες.

Οι 8 πρώτες ομάδες πέρασαν στον προημιτελικό γύρο σε σειρές αγώνων best-of-five. Οι δύο πρώτοι αγώνες διεξήχθησαν στην έδρα των τεσσάρων ομάδων που τερμάτισαν στην υψηλότερη θέση με το πέρας της κανονικής περιόδου (θέσεις 1-4), ο τρίτος αγώνας και (εάν χρειάστηκε) ο τέταρτος, διεξήχθησαν στην έδρα των επόμενων τεσσάρων ομάδων (θέσεις 5-8) και το πέμπτο παιχνίδι (εάν χρειάστηκε) διεξήχθη στην έδρα που έγιναν οι πρώτοι δύο αγώνες. Οι νικητές πήραν τη πρόκριση για το Final Four.

[10] Το ABA League (το οποίο μετονομάστηκε σε ABA League First Division το 2017) είναι το περιφερειακό επαγγελματικό πρωτάθλημα μπάσκετ ανδρών πρώτης κατηγορίας, που αρχικά περιλάμβανε συλλόγους από την πρώην Γιουγκοσλαβία (Βοσνία-Ερζεγοβίνη, Κροατία, Μαυροβούνιο, Βόρεια Μακεδονία, Σερβία και Σλοβενία). Το πρωτάθλημα συνυπάρχει παράλληλα με τα (μειωμένης δυναμικότητας) εθνικά πρωταθλήματα στη Βοσνία-Ερζεγοβίνη, την Κροατία, τη Βόρεια Μακεδονία, το Μαυροβούνιο, τη Σερβία και τη Σλοβενία. Όλοι οι σύλλογοι της Αδριατικής Λίγκας, εκτός από έναν, με την ολοκλήρωση της διοργάνωσης αυτής, συμμετέχουν στις διοργανώσεις της χώρας τους.
(Πηγή : https://en.wikipedia.org/wiki/ABA_League)

[11] Το VTB United League είναι ένα διεθνές επαγγελματικό πρωτάθλημα μπάσκετ ανδρών που ιδρύθηκε το 2008. Αποτελείται κυρίως από ρωσικούς συλλόγους, μαζί με έναν από τη Λευκορωσία και το Καζακστάν. Από το 2013, είναι η πρώτη βαθμίδα του ρωσικού επαγγελματικού μπάσκετ.
(Πηγή : https://en.wikipedia.org/wiki/VTB_United_League)

- Σεζόν 2017-2018

Συνολικά 16 ομάδες από 9 χώρες συμμετείχαν στο πρωτάθλημα, συμπεριλαμβανομένων 11 ομάδων με μακροχρόνια άδεια (συμβόλαιο τύπου Α) από τη σεζόν 2016-2017, 1 ομάδα προκρίθηκε από το EuroCup (κάτοχος) και οι 4 υψηλότερες ομάδες από την ABA League, τη γερμανική Bundesliga, τη VTB United League και τη ισπανική ACB. Η δομή της διοργάνωσης, σε όλες τις φάσεις της, ήταν η ίδια με τη περσινή σεζόν.

- Σεζόν 2018-2019

Η δομή της διοργάνωσης, σε όλες τις φάσεις της, ήταν η ίδια με τη περσινή σεζόν, όπως και ο τρόπος κατανομής των ομάδων που συμμετείχαν.

- Σεζόν 2019-2020

Στις 5 Ιουλίου 2018, η Euroleague Basketball συμφώνησε να επεκτείνει τη διοργάνωση, δίνοντας wild cards για δύο χρόνια στη γερμανική Bayern Munich και στη γαλλική LDLC ASVEL. Συνεπώς, συνολικά 18 ομάδες από 10 χώρες συμμετείχαν στην διοργάνωση τη σεζόν 2019-2020.

Στις 12 Μαρτίου 2020, η Euroleague Basketball ανέστειλε προσωρινά τις διοργανώσεις της λόγω της πανδημίας COVID-19[12], ενώ στις 25 Μαΐου, ακύρωσε τις διοργανώσεις της οριστικά. Ως συνέπεια της πανδημίας του COVID-19, δεν αναγνωρίστηκε καμία ομάδα ως πρωταθλήτρια για τη σεζόν εκείνη.

- Σεζόν 2020-2021

Καθώς η σεζόν αυτή ήταν η πρώτη που παίχτηκε μετά την εγκατάλειψη της προηγούμενης, λόγω ταξιδιωτικών περιορισμών που προκλήθηκαν από την πανδημία COVID-19, το διοικητικό συμβούλιο της EuroLeague αποφάσισε ότι οι 18 ομάδες από τις 10 χώρες της προηγούμενης σεζόν, θα παραμείνουν στο πρωτάθλημα. Το format της διοργάνωσης ήταν το ίδιο με τις προηγούμενες σεζόν.

- Σεζόν 2021-2022

Ξεκίνησαν τη σεζόν 18 ομάδες από 10 χώρες, ενώ χορηγήθηκε διετής wild card στην ALBA Berlin. Οι τρεις ρωσικές ομάδες ωστόσο (CSKA Moscow, Zenit Saint Petersburg και Unics Kazan) ανεστάλησαν λόγω της ρωσικής εισβολής στην

[12] Η νόσος του κορονοϊού 2019 (COVID-19) είναι μια μεταδοτική ασθένεια που προκαλείται από έναν ιό, το σοβαρό οξύ αναπνευστικό σύνδρομο κορονοϊός 2 (SARS-CoV-2). Το πρώτο γνωστό κρούσμα εντοπίστηκε στη Γουχάν της Κίνας τον Δεκέμβριο του 2019. Η ασθένεια εξαπλώθηκε γρήγορα σε όλο τον κόσμο, με αποτέλεσμα την πανδημία του COVID-19.
(Πηγή : https://en.wikipedia.org/wiki/COVID-19)

Ουκρανία[13] το Φεβρουάριο του 2022. Παράλληλα, καθώς η ρωσική εισβολή δεν σταμάτησε, τα ρεκόρ όλων των αγώνων της κανονικής περιόδου εναντίον ρωσικών ομάδων ακυρώθηκαν και τα ρεκόρ νικών-ηττών των ομάδων προσαρμόστηκαν ανάλογα, επηρεάζοντας δραματικά τη βαθμολογία του πρωταθλήματος στη κανονική περίοδο. Οι φάσεις των προημιτελικών και του Final Four είχαν την ίδια μορφή με τις προηγούμενες σεζόν.

2.2 Παρουσίαση των δεδομένων

Τα δεδομένα τα οποία χρησιμοποιήθηκαν στη παρούσα διπλωματική εργασία, αφορούν μέσους όρους παικτών και ομάδων, στην διοργάνωση της EuroLeague, ξεκινώντας από την αγωνιστική σεζόν 2000-2001 ως και τη σεζόν 2021-2022. Επίσης, έχουν καταγραφεί και οι δείκτες PIR, τόσο των παικτών, όσο και των ομάδων, για την ίδια περίοδο. (Ο δείκτης PIR περιγράφεται αναλυτικά παρακάτω)

Η καταγραφή τους έγινε με τη χρήση του προγράμματος επεξεργασίας υπολογιστικών φύλλων «Microsoft Excel». Πηγή άντλησής τους ήταν η επίσημη ιστοσελίδα της EuroLeague. Η ανάλυση έγινε με τη γλώσσα προγραμματισμού ανοιχτού κώδικα (open source) «R». Αναλυτικά, ο κώδικας παρουσιάζεται στο επισυναπτόμενο παράρτημα, στο τέλος της εργασίας.

Από τη σεζόν 2000-2001 που ιδρύθηκε η Ευρωλίγκα όπως είναι γνωστή σήμερα, μέχρι και τη προτελευταία ολοκληρωμένη σεζόν (2021-2022), έχουν συμμετάσχει, έστω και για μια σεζόν, συνολικά 87 ομάδες. Παρακάτω παρουσιάζονται αναλυτικά αυτές οι ομάδες, η χώρα προέλευσης τους και ο κωδικός της καθεμίας. Επειδή κατά τη πορεία των διαφορετικών σεζόν, ορισμένες ομάδες είχαν διάφορα ονόματα λόγω χορηγικών συμφωνιών, παρουσιάζονται οι ομάδες με τη σημερινή τους ονομασία.

Team	Country	Team id
Acqua S.Bernardo Cantù	Italy	64
A.S. Monaco Basket	France	87
AEK Athens	Greece	2
ALBA Berlin	Germany	25
Anadolu Efes S.K.	Turkey	28
Aris Thessaloniki	Greece	48

[13] Στις 24 Φεβρουαρίου 2022, η Ρωσία εισέβαλε στην Ουκρανία σε μια μεγάλη κλιμάκωση του Ρωσο-Ουκρανικού Πολέμου, ο οποίος ξεκίνησε το 2014. Η εισβολή πιθανότατα οδήγησε σε δεκάδες χιλιάδες θανάτους και από τις δύο πλευρές και προκάλεσε τη μεγαλύτερη προσφυγική κρίση στην Ευρώπη από τον Β' Παγκόσμιο Πόλεμο με περίπου 8 εκατομμύρια ανθρώπους να έχουν εκτοπιστεί εντός της χώρας μέχρι τα τέλη Μαΐου, καθώς και 7,8 εκατομμύρια Ουκρανούς να έχουν εγκαταλείψει τη χώρα ως και τις 8 Νοεμβρίου 2022. Μέσα σε πέντε εβδομάδες από την εισβολή, η Ρωσία γνώρισε τη μεγαλύτερη μετανάστευση από την Οκτωβριανή Επανάσταση του 1917. Η εισβολή αυτή, είχε και έχει προκαλέσει παράλληλα, παγκόσμιες ελλείψεις τροφίμων.
(Πηγή : https://en.wikipedia.org/wiki/2022_Russian_invasion_of_Ukraine)

Arka Gdynia	Poland	43
Basketball Club Neptūnas	Lithuania	80
Basketball Club Žalgiris	Lithuania	24
Baskets Oldenburg	Germany	60
BC Budivelnyk Kyiv	Ukraine	72
BC Khimki	Russia	61
BC Nizhny Novgorod	Russia	81
BC UNICS	Russia	67
BC Zenit Saint Petersburg	Russia	86
Beşiktaş Emlakjet	Turkey	69
Brose Bamberg	Germany	45
Budućnost VOLI	Montenegro	4
Cazoo Baskonia	Spain	21
Cholet Basket	France	63
Chorale Roanne Basket	France	55
Cibona Zagreb	Croatia	5
Club Baloncesto Estudiantes	Spain	1
Club Baloncesto Gran Canaria – Claret	Spain	85
Club Joventut Badalona	Spain	49
Darüşşafaka Basketbol	Turkey	83
Dinamo Banco di Sardegna Sassari	Italy	78
Élan Béarnais Pau-Lacq-Orthez	France	34
Élan Sportif Chalonnais	France	71
FC Barcelona Bàsquet	Spain	6
FC Bayern Munich Basketball	Germany	74
Fenerbahçe Beko	Turkey	52
Filou Oostende	Belgium	36
Fortitudo Kigili Bologna	Italy	15
Fraport Skyliners	Germany	13
Galatasaray S.K.	Turkey	65
GeVi Napoli	Italy	51
Hapoel Bank Yahav Jerusalem	Israel	7
KK Cedevita Junior	Croatia	70
KK Crvena zvezda mts	Serbia	73
KK Krka	Slovenia	30
KK Olimpija	Slovenia	22
KK Zadar	Croatia	23
KK Zagreb	Croatia	68
LDLC ASVEL	France	26
Le Mans Sarthe Basket	France	53
Limoges Cercle Saint-Pierre	France	79
London Towers	United Kingdom	8
Lugano Tigers	Switzerland	10
Maccabi Playtika Tel Aviv	Israel	31
Maroussi B.C.	Greece	62
MBC Dynamo Moscow	Russia	50

Mens Sana 1871 Basket (Montepaschi Siena)	Italy	40
Nanterre 92	France	75
Olympiacos Piraeus B.C.	Greece	12
Orléans Loiret Basket	France	59
Ovarense Basquetebol	Portugal	14
Pallacanestro Olimpia Milano	Italy	44
Pallacanestro Treviso (Benetton Basket)	Italy	3
Pallacanestro Virtus Roma	Italy	41
Panathinaikos Athens B.C.	Greece	32
Panionios Athens	Greece	57
PAOK B.C.	Greece	16
Partizan Mozzart Bet	Serbia	33
PBC CSKA Moscow	Russia	27
PBC Lokomotiv Kuban	Russia	76
PBC Ural Great Perm	Russia	39
Peristeri B.C.	Greece	17
PGE Turów Zgorzelec	Poland	82
Pınar Karşıyaka	Turkey	84
Real Madrid Baloncesto	Spain	18
RheinStars Köln	Germany	54
Rytas Vilnius	Lithuania	46
S.S. Felice Scandone (Avellino)	Italy	56
Saint Petersburg Lions	Russia	20
Scaligera Basket Verona	Italy	11
SIG Strasbourg	France	47
SLUC Nancy Basket	France	58
Spirou Charleroi	Belgium	19
Surne Bilbao Basket	Spain	66
Ulker	Turkey	37
Unicaja Málaga	Spain	38
Unione Sportiva Victoria Libertas Pallacanestro (Scavolini Pesaro)	Italy	35
Valencia Basket Club	Spain	42
Virtus Segafredo Bologna	Italy	9
WKS Śląsk Wrocław	Poland	29
Zastal Zielona Góra	Poland	77

Πίνακας 2.1 : Οι ομάδες που έχουν συμμετάσχει στη διοργάνωση της EuroLeague.

Στην ανάλυση που θα ακολουθήσει στα προσεχή κεφάλαια, χρησιμοποιήθηκαν δεδομένα με τις παρακάτω μεταβλητές :

Μεταβλητές	Επεξήγηση
Team	Όνομα της ομάδας
Team id	Κωδικός της ομάδας
Playoffs / Quarter-Finals	Ένδειξη για το αν η ομάδα προκρίθηκε στα playoffs (1:ναι , 0:όχι)
Final 4 / Semi-Finals	Ένδειξη για το αν η ομάδα προκρίθηκε στο Final Four (1:ναι , 0:όχι)
Final 4 position	Τελική κατάταξη στο Final Four
Points	Πόντοι
2PT %	Ποσοστό επιτυχημένων διπόντων
3PT %	Ποσοστό επιτυχημένων τριπόντων
FT %	Ποσοστό επιτυχημένων ελευθέρων βολών
OR	Επιθετικά “Ριμπάουντ”
DR	Αμυντικά “Ριμπάουντ”
TR	Συνολικά “Ριμπάουντ”
AST	“Ασίστ” (Τελικές πάσες που οδήγησαν σε καλάθι)
STL	Κλεψίματα
TO	Λάθη
BLK	Κοψίματα υπέρ της ομάδας
BLKA	Κοψίματα κατά της ομάδας
FC	Φάουλ στα οποία υπέπεσε η ομάδα
FD	Φάουλ τα οποία κέρδισε η ομάδα
PIR	Δείκτης αξιολόγησης της απόδοσης
Season	Κωδικός για τη κάθε σεζόν (1:2000-2001 , 2:2001-2002 κ.ο.κ)
Player	Όνομα του παίκτη

Πίνακας 2.2 : Οι μεταβλητές μας.

(Πηγή : https://www.euroleaguebasketball.net/euroleague/teams/)

Ο δείκτης **Performance Index Rating (PIR)** είναι ένας μαθηματικός στατιστικός τύπος μπάσκετ που χρησιμοποιείται από τις διοργανώσεις πρώτης και δεύτερης κατηγορίας της «Euroleague Basketball Company», την EuroLeague και το EuroCup, καθώς και από διάφορα ευρωπαϊκά εθνικά, εγχώρια και περιφερειακά πρωταθλήματα. Είναι μέρος του συστήματος αξιολόγησης μπάσκετ Tendex[14]. Αναφέρεται επίσης ποικιλοτρόπως ως Κατάταξη Δείκτη Απόδοσης, Βαθμολογία, Κατάταξη, Αξιολόγηση, Αποτίμηση και Αποδοτικότητα. Είναι παρόμοιο, αλλά όχι ακριβώς το ίδιο, με το στατιστικό του NBA, το δείκτη Efficiency (EFF).

[14] Το σύστημα Tendex είναι ένας μαθηματικός στατιστικός τύπος μπάσκετ που δημιουργήθηκε από τον αθλητικογράφο Dave Heeren, προκειμένου να προσδιοριστεί η απόδοση των παικτών μπάσκετ. Είναι γενικά αποδεκτό ως η αρχική φόρμουλα σταθμισμένης προηγμένης στατιστικής, που χρησιμοποιείται στο άθλημα του μπάσκετ.
(Πηγή : https://en.wikipedia.org/wiki/Tendex)

Ο ειδικός αυτός δείκτης αξιολόγησης υπολογίζεται ως εξής :

PIR = (Points + Rebounds + Assists + Steals + Blocks + Fouls Drawn) – (Missed Field Goals + Missed Free Throws + Turnovers + Shots Rejected/Blocked + Fouls Committed)

(Πηγή : https://en.wikipedia.org/wiki/Performance_Index_Rating)

2.3 Παρουσίαση προβλημάτων/στόχων της εργασίας

Στη παρούσα διπλωματική εργασία, έγινε προσπάθεια να απαντηθούν τα παρακάτω ερωτήματα :

1) Ποιοι είναι οι σημαντικοί, προβλεπτικοί παράγοντες στη πρόκριση των ομάδων στα playoffs και στο Final Four ;

2) Ποια είναι η συνεισφορά των καλύτερων παικτών με βάση το PIR στη πορεία των ομάδων τους ;

3) Είναι εύκολο να ανακαλύψουμε ομάδες μέσα στα δεδομένα μας, που να παρουσιάζουν παρόμοια χαρακτηριστικά ;

4) Μπορούμε να φτιάξουμε αποτελεσματικά μοντέλα ταξινόμησης, χρησιμοποιώντας διάφορες μεθόδους, προκειμένου να εξετάσουμε αν οι ομάδες προκρίθηκαν ή όχι στις επιμέρους φάσης της διοργάνωσης ;

2.4 Βιβλιογραφική επισκόπηση

Σε αυτή την ενότητα, γίνεται αναφορά σε υπάρχοντα επιστημονικά άρθρα τα οποία είναι σχετικά με το θέμα της παρούσας εργασίας, καθώς και στα συμπεράσματα στα οποία αυτά καταλήγουν.

Οι Marmarinos, C. et al. (2016) αποσκοπούσαν να ερευνήσουν ποιοι παράγοντες είναι στατιστικώς σημαντικοί για τη πρόκριση (ή μη) μιας ομάδας στα playoffs της EuroLeague. Συνέλλεξαν δεδομένα από 1514 αγώνες των σεζόν 2012-13, 2013-14 και 2014-15, και χρησιμοποιώντας τη μέθοδο της διαχωριστικής ανάλυσης για τη στατιστική τους ανάλυση, κατέληξαν στο συμπέρασμα πως τα αμυντικά ριμπάουντ, η αποτελεσματικότητα στην επίθεση και στην άμυνα, οι τελικές πάσες και ο περιορισμός των λαθών είναι οι ζητούμενοι διαχωριστικοί παράγοντες.

Οι Mikołajec, K. et al. (2021) αναζήτησαν τους παράγοντες που σχετίζονται με την απόδοση 10 ομάδων της EuroLeague, κατά τη διάρκεια 13 αγωνιστικών σεζόν. Χρησιμοποιώντας τη μέθοδο της ανάλυσης παλινδρόμησης, κατέληξαν πως οι μεταβλητές που έχουν τον μεγαλύτερο αντίκτυπο ήταν τα δίποντα που επιτευχθήκαν και επιχειρήθηκαν (2PT-made, 2PT-attempts), ο αριθμός των ελεύθερων βολών που επιτεύχθηκαν (1PT-made), ο αριθμός των τριπόντων που επιτευχθήκαν και επιχειρήθηκαν (3PT-made, 3PT-attempts), το πλήθος των ασίστ και ο αριθμός των φάουλ. Ειδικότερα κατέληξαν, πως επτά ομάδες είχαν ως στατιστικώς σημαντικότερο παράγοντα νίκης τον αριθμό των τριπόντων που επιτευχθήκαν, δύο ομάδες είχαν τον αριθμό των ελεύθερων βολών που επιτεύχθηκαν, και μια ομάδα τον αριθμό των φάουλ στα οποία υπέπεσε.

Οι Dogan, I. and Ersoz, Y. (2019) χρησιμοποιώντας τη μέθοδος της γραμμικής, διαχωριστικής ανάλυσης ερεύνησαν ποιοι παράγοντες παίζουν καθοριστικό ρόλο στη πρόκριση ομάδων από τους επιμέρους γύρους της EuroLeague. Συνέλεξαν στοιχεία για τις σεζόν 2010-2017 και κατέληξαν στα εξής συμπεράσματα : Το ποσοστό των δίποντών, το αμυντικό ριμπάουντ, τα φάουλ υπέρ και τα κοψίματα (μπλοκ) υπέρ είναι τα στοιχεία που έχουν τη μεγαλύτερη συμβολή στην πρόκριση των ομάδων στο γύρο Top 16. Επιπλέον, οι άλλες μεταβλητές που συμβάλλουν στην πρόκριση της ομάδας, είναι οι ασίστ και το ποσοστό των τριπόντων, απλώς σε μικρότερο βαθμό. Για τη πρόκριση των ομάδων στα playoffs, μεγαλύτερη συμβολή παίζουν το ποσοστό των δίποντών, τα κοψίματα υπέρ και τα λάθη. Σε μικρότερο βαθμό συμβάλλουν το ποσοστό των τριπόντων, το αμυντικό ριμπάουντ και τα φάουλ υπέρ. Όσον αφορά τη πρόκριση στο Final Four, το ποσοστό των τριπόντων είναι το σχετικό στατιστικό που έχει τη μεγαλύτερη συμβολή. Ακόμη, οι ασίστ, τα μπλοκ υπέρ και το αμυντικό ριμπάουντ συμβάλλουν, σε μικρότερο όμως βαθμό. Τέλος, το ποσοστό των τριπόντων είναι το στατιστικό με τη μεγαλύτερη συμβολή στην ανάδειξη μιας ομάδας ως πρωταθλήτρια, ενώ μικρότερο ρόλο παίζουν τα επιθετικά ριμπάουντ και τα ποσοστά των δίποντών.

Οι Mandić, R. et al. (2019) επιχείρησαν να συγκρίνουν στατιστικά στοιχεία ανάμεσα στη EuroLeague και στο NBA, από το 2000 ως το 2017. Συγκεντρώνοντας δεδομένα για τις δύο αυτές διοργανώσεις από τη σεζόν 2000-2001 μέχρι τη σεζόν 2016-2017 και χρησιμοποιώντας τόσο γραμμική παλινδρόμηση, όσο και κάποια εποπτικά μέσα-γραφήματα, κατέληξαν πως η EuroLeague μοιάζει αρκετά, ποσοτικά και ποιοτικά με το NBA. Δεν υπάρχουν διακριτές διαφορές σε ασίστ, κλεψίματα ή συνολικά ριμπάουντ και οι όποιες διαφορές υφίστανται στα μοτίβα εκτέλεσης δίποντών και τριπόντων, έχουν τη τάση να μειώνονται επίσης. Αντιθέτως, τα μπλοκ και ο ρυθμός αμυντικών ριμπάουντ είναι σημαντικά υψηλότερα στο NBA. Επίσης, υπάρχουν περισσότερα προσωπικά φάουλ ανά κατοχή στην EuroLeague, με μια πιθανή εξήγηση να είναι η αυστηρότερη διαιτησία ή/και το πιο επιθετικό παιχνίδι. Επιπροσθέτως, συμπέραναν πως υφίσταται σημαντική διαφορά στον ρυθμό του παιχνιδιού. Το ευρωπαϊκό μπάσκετ είναι πιο τακτικό, ενώ το NBA έχει μικρότερες κατοχές, λιγότερη έμφαση στο παιχνίδι άμυνας και τακτικής και, κατά συνέπεια, περισσότερες ανατροπές. Ωστόσο, θεωρούν πως το μπάσκετ του NBA μοιάζει περισσότερο με το ευρωπαϊκό μπάσκετ, στη φάση των playoffs, τα οποία διαθέτουν τις καλύτερες ομάδες και υψηλότερο επίπεδο

ανταγωνιστικότητας. Επιπλέον, η σταδιακή μείωση των διαφορών μεταξύ NBA και Ευρωλίγκας εξαρτάται εν πολλοίς και από την διαφορά των κανόνων. Οι πιο ουσιαστικές διαφορές στους κανόνες που απομένουν είναι οι 4 περίοδοι των 10 λεπτών, η διαφορετική απόσταση του τρίποντού και το ελαφρώς στενότερο γήπεδο. Η έρευνα εν τούτοις καταλήγει, πως η απόλυτη διαφορά ποιότητας ανάμεσα στις δύο διοργανώσεις δεν έχει μειωθεί πολύ, χωρίς ωστόσο να είναι εύκολο να μετρηθεί αυτή με άμεσο τρόπο. Έμμεσα στοιχεία από διεθνείς αγώνες όπου κυριαρχεί συνήθως η εθνική ομάδα των ΗΠΑ, υποδεικνύουν ότι η διαφορά μεταξύ των κορυφαίων παικτών του NBA και των κορυφαίων Ευρωπαίων παικτών είναι ακόμα ουσιαστική. Ένας άλλος έμμεσος δείκτης αυτού είναι καθαρά οικονομικός, καθώς οι αστέρες του NBA κερδίζουν 10 φορές περισσότερα χρήματα, (χωρίς τις χορηγίες) σε σχέση με τους καλύτερους παίκτες της EuroLeague.

Οι Csataljay, G. et al. (2009) επιδίωξαν να εντοπίσουν τους κρίσιμους δείκτες απόδοσης, οι οποίοι συμβάλλουν περισσότερο στη διάκριση των νικητριών και χαμένων ομάδων, σύμφωνα με τις διαφορές στα τελικά σκορ αγώνων του Ευρωπαϊκού πρωταθλήματος του 2007. Συνέλεξαν δεδομένα από την επίσημη ιστοσελίδα της διοργάνωσης και για τα 54 παιχνίδια της. Έγινε χρήση της ανάλυσης συστάδων προκειμένου να ταξινομηθούν οι αγώνες σε τρεις κατηγορίες : κλειστά παιχνίδια με διαφορές στο τελικό σκορ μεταξύ 1 και 9 πόντων, ισορροπημένα παιχνίδια (10-22 πόντοι) και ανισόρροπα παιχνίδια (22-34 πόντοι), καθώς και του ελέγχου "Wilcoxon signed ranks" για τη σύγκριση 18 δεικτών απόδοσης μεταξύ των νικητριών και των χαμένων ομάδων, για κάθε κατηγορία αγώνα. Αναλύοντας αρχικά όλους τους αγώνες του Ευρωπαϊκού πρωταθλήματος, κατέληξαν πως ήταν 13 οι σημαντικοί δείκτες απόδοσης. Πέρα από τους πόντους, σημαντικά επίσης ήταν το ποσοστό επιτυχημένων τριπόντων, το ποσοστό των επιτυχημένων ελεύθερων βολών και των αμυντικών ριμπάουντ. Στα κλειστά παιχνίδια, βρέθηκαν 6 σημαντικοί δείκτες απόδοσης, με το ποσοστό των επιτυχημένων ελεύθερων βολών να είναι ο πιο κρίσιμος από αυτούς. Στα ισορροπημένα παιχνίδια, η ανάλυση έδειξε 8 σημαντικούς δείκτες απόδοσης, με τους σημαντικότερους να είναι το ποσοστό δίποντών και το ποσοστό τριπόντων. Τέλος, η στατιστική ανάλυση των ανισόρροπων παιχνιδιών έδειξε 5 σημαντικούς δείκτες απόδοσης. Συμπερασματικά, τα αποτελέσματα από τα ισορροπημένα και τα ανισόρροπα παιχνίδια έδειξαν ότι κέρδιζαν οι ομάδες με την καλύτερη απόδοση στα περισσότερα στατιστικά στοιχεία του παιχνιδιού, ενώ στα κλειστά παιχνίδια, νικήτριες ήταν οι ομάδες οι οποίες είχαν καλύτερη απόδοση στα τρίποντα, στις ελεύθερες βολές και στο αμυντικό ριμπάουντ.

Στόχος των García, J. et al. (2013) ήταν να προσδιοριστούν οι δείκτες απόδοσης αγώνων μπάσκετ που διακρίνουν καλύτερα νικητές και ηττημένους, στην κανονική περίοδο και στα playoff. Συνέλεξαν δεδομένα από 323 παιχνίδια της ACB Spanish Basket League, από την κανονική περίοδο, και από τα playoffs. Με τη χρήση της μεθόδου της ανάλυσης συστάδων, χωρίστηκε αρχικά το δείγμα σε ισορροπημένα (ίση ή κάτω από 12 πόντους η τελική διαφορά), μη ισορροπημένα (μεταξύ 13 και 28 πόντων) και πολύ ανισόρροπα παιχνίδια (πάνω από 28 πόντους διαφορά), ενώ χρησιμοποιήθηκε και διαχωριστική ανάλυση για τον προσδιορισμό των δεικτών

απόδοσης σε αγώνες κανονικής περιόδου και playoffs, για κάθε κατηγορία αγώνων. Επίσης, μέσω της μεθόδου ANOVA εντοπίστηκαν διαφορές μεταξύ νικητριών και ηττημένων ομάδων, για τη κανονική περίοδο και για τα playoffs. Κατέληξαν στα εξής συμπεράσματα : Αρχικά, στη κανονική περίοδο, η απόδοση στις ασίστ, στα αμυντικά ριμπάουντ, στα επιτυχημένα δίποντα και τρίποντα έπαιξαν ρόλο στο διαχωρισμό των ομάδων σε νικήτριες και χαμένες. Σε αγώνες playoff, δεν εντοπίστηκαν μεταβλητές σημαντικές για το διαχωρισμό. Στη συνέχεια, για τα ισορροπημένα παιχνίδια, ξεχωρίζουν ως σημαντικοί παράγοντες οι ασίστ, τα αμυντικά ριμπάουντ και τα επιτυχημένα δίποντα. Στα παιχνίδια playoffs, η ANOVA αποκάλυψε πως η διαφορά μεταξύ νικητριών και ηττημένων ομάδων, σχετίζεται με τα αμυντικά ριμπάουντ και επιτυχημένα δίποντα. Τα αποτελέσματα για μη ισορροπημένα παιχνίδια έδειξαν ότι οι ασίστ παίζουν το πιο καθοριστικό ρόλο. Στα playoffs, η ANOVA εντόπισε πως τα επιτυχημένα τρίποντα και τα αμυντικά ριμπάουντ παίζουν σημαντικό ρόλο στο διαχωρισμό των ομάδων.

ΚΕΦΑΛΑΙΟ 3ο

Περιγραφική Ανάλυση

Σε αυτό το κεφάλαιο παρουσιάζονται αναλυτικά κάποια περιγραφικά μέτρα για τις επεξηγηματικές μεταβλητές μας σε σχέση με τις μεταβλητές απόκρισης **Playoffs / Quarter-Finals** (πρόκριση ή όχι στα playoffs) και **Final 4 / Semi-Finals** (πρόκριση ή όχι στο Final Four). Οι αναλύσεις που ακολουθούν βασίζονται στο διαχωρισμό των ομάδων σε αυτές που προκρίθηκαν στα playoffs και στο Final Four, και σε αυτές που δε προκρίθηκαν. Παράλληλα, παρουσιάζεται και διαγραμματική απεικόνιση των δεδομένων των αναλύσεων μας, ώστε τα αποτελέσματα να γίνουν ακόμη πιο κατανοητά και συγκρίσιμα από τον αναγνώστη.

3.1 Διερεύνηση για ελλιπή δεδομένα (missing values)

Αρχικά, γίνεται έλεγχος σε κάθε ανάλυση για την ύπαρξη ελλιπών τιμών στα δεδομένα μας. Τα αποτελέσματα δείχνουν ότι υπάρχουν missing values[15], κάτι το οποίο ήταν αναμενόμενο. Αφενός, από τη σεζόν 2001-2002 ως τη σεζόν 2003-2004 δεν υπήρξε φάση προημιτελικών/playoffs, αλλά κατευθείαν φάση ημιτελικών, και συνεπώς τα NA values εμφανίζονται για τη μεταβλητή Playoffs / Quarter-Finals. Αφετέρου, η σεζόν 2019-2020 διακόπηκε οριστικά νωρίτερα του αναμενόμενου, λόγω της πανδημίας του COVID-19 και συνεπώς, ούτε εκεί πραγματοποιήθηκαν οι φάσεις των playoffs και του Final Four. Για αυτό το λόγο, και επειδή έχουμε δεδομένα από πολλές σεζόν, αποφασίστηκε να μην ληφθούν υπόψιν οι σεζόν 2001-2004 , καθώς και η σεζόν 2019-2020.

Κανονικά, τα missing values πρέπει να συμπληρώνονται. Αυτό μπορεί να πραγματοποιηθεί είτε με τη χρήση κάποιας σταθεράς, είτε με την μέση τιμή του γνωρίσματος για όλα τα δείγματα που ανήκουν στην ίδια κατηγορία, είτε με τη διάμεσο όταν υπάρχει ασσυμετρία, ή ακόμη και με την χρήση της πιο πιθανής τιμής.

[15] Στα στατιστικά στοιχεία, τα ελλιπή δεδομένα (missing values) εμφανίζονται όταν δεν αποθηκεύεται καμία τιμή δεδομένων για τη μεταβλητή σε μια παρατήρηση. Τα ελλιπή δεδομένα είναι σύνηθες φαινόμενο και μπορεί να έχουν σημαντική επίδραση στα συμπεράσματα που μπορούν να εξαχθούν από τα δεδομένα. Υπάρχουν τρία κύρια προβλήματα που προκαλούν τα ελλείποντα δεδομένα : η έλλειψη δεδομένων μπορεί να προκαλέσει σημαντική μεροληψία, να κάνει τον χειρισμό και την ανάλυση των δεδομένων πιο «επίπονη» διαδικασία, και να δημιουργήσει μειώσεις στην αποτελεσματικότητα των αλγορίθμων.
(Πηγή : https://en.wikipedia.org/wiki/Missing_data)

(Πηγή : https://en.wikipedia.org/wiki/Imputation_(statistics))

3.2 Περιγραφικά μέτρα και διαγραμματική απεικόνιση των δεδομένων

Στην παράγραφο αυτή θα γίνει παρουσίαση των βασικών περιγραφικών μέτρων των ποσοτικών χαρακτηριστικών των δεδομένων μας, σε συνδυασμό με τις διαφορετικές αναλύσεις τις οποίες πραγματοποιήσαμε. Αυτό σημαίνει ότι από τα αποτελέσματα λείπουν τα χαρακτηριστικά εκείνα που αφορούν τις κατηγορικές μας μεταβλητές.

3.2.1 Γενικά χαρακτηριστικά των ομάδων, για όλες τις σεζόν, στο πέρασμα του χρόνου

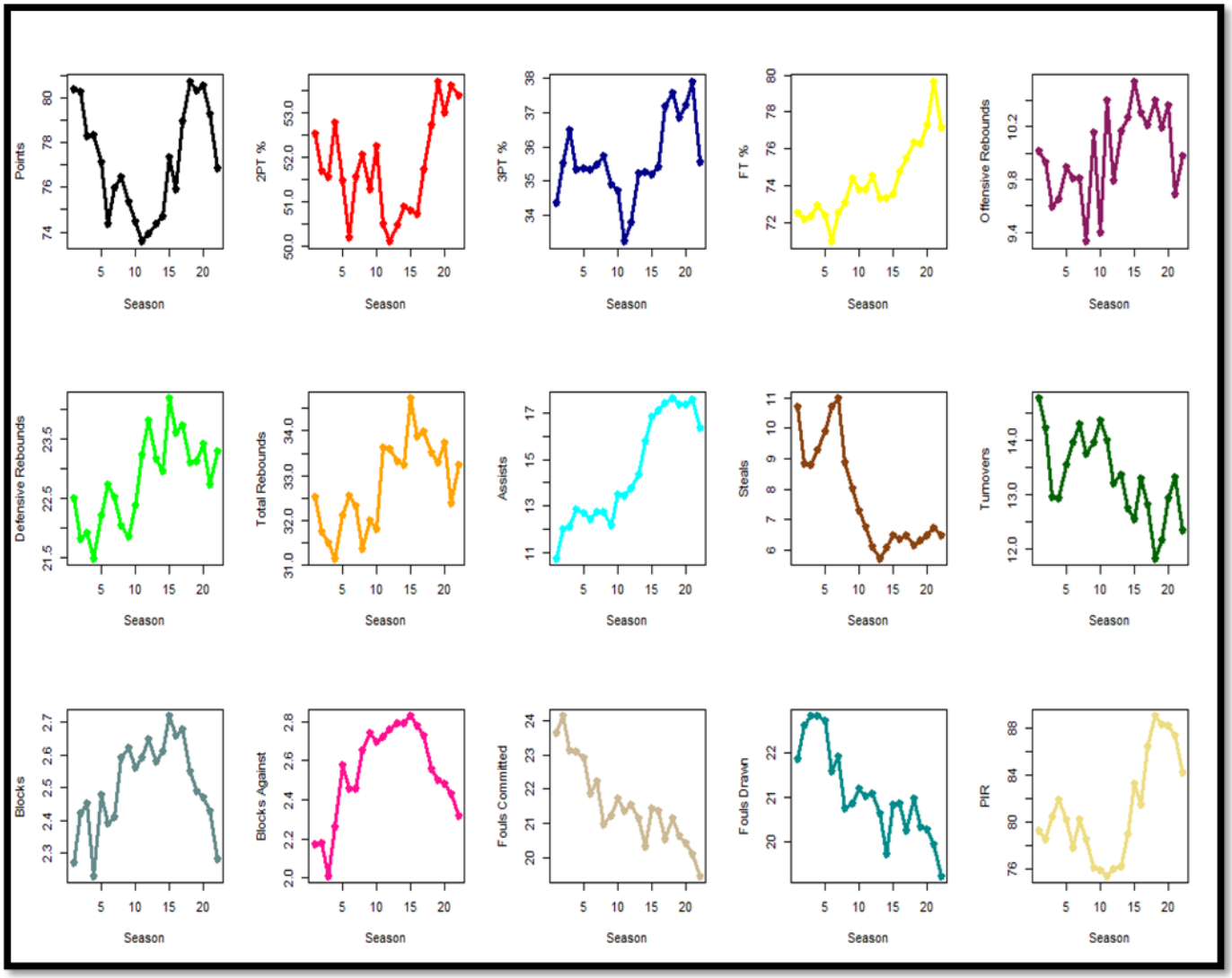
Σε αυτήν την ενότητα, καταγράψαμε, το μέσο όρο όλων των στατιστικών στοιχείων, για όλες τις ομάδες στο σύνολο των σεζόν. Σε αυτήν την ανάλυση, δεν έχουμε εξαιρέσει καμία σεζόν, άρα έχουμε δεδομένα για 22 αγωνιστικές σεζόν. Σκοπός της ανάλυσης μας ήταν να δούμε πώς μεταβάλλονται αυτά τα χαρακτηριστικά στο πέρασμα των χρόνων.

	Min.	1st. Qu.	Median	Mean	3rd. Qu.	Max.
Points	73,57	74,86	76,98	77,16	79,19	80,72
2PT %	50,1	50,83	51,62	51,77	52,66	53,7
3PT %	33,23	35,2	35,38	35,62	36,29	37,91
FT %	70,95	72,59	73,65	74,2	75,3	79,67
Offensive Rebounds	9,33	9,795	9,995	9,994	10,255	10,54
Defensive Rebounds	21,48	22,25	22,84	22,8	23,27	24,17
Total Rebounds	31,14	32,02	32,89	32,8	33,58	34,72
Assists	10,71	12,7	13,62	14,5	17,05	17,64
Steals	5,72	6,393	6,77	7,717	8,88	10,98
Turnovers	11,81	12,84	13,3	13,33	13,96	14,77
Blocks	2,23	2,422	2,52	2,506	2,605	2,72
Blocks Against	2,01	2,438	2,57	2,541	2,737	2,83
Fouls Commited	19,47	20,71	21,36	21,56	22,14	24,14
Fouls Drawn	19,22	20,41	20,91	21,11	21,8	22,85
PIR	75,35	77,91	80,16	81,04	83,94	89,04

Πίνακας 3.1 : Στατιστικά περιγραφικά μέτρα.

Στον παραπάνω πίνακα παρουσιάζονται η ελάχιστη τιμή, το πρώτο τεταρτημόριο, η διάμεσος, η μέση τιμή, το τρίτο τεταρτημόριο και η μέγιστη τιμή, για το μέσο όρο των μεταβλητών/στατιστικών των ομάδων.

Χρησιμοποιήσαμε αρχικά κάποια γραφήματα χρονοσειρών (time series plots), προκειμένου να αποκτήσουμε μια γενικότερη οπτική για το πώς μεταβάλλεται κάθε στατιστικό στοιχείο με τη πάροδο των σεζόν. Ένα γράφημα χρονοσειρών είναι ένα γράφημα που εμφανίζει δεδομένα που συλλέγονται σε μια χρονική ακολουθία, από οποιαδήποτε διαδικασία. Μπορεί να χρησιμοποιηθεί για να προσδιοριστεί η τάση των δεδομένων με την πάροδο του χρόνου, και εάν τα σημεία των δεδομένων είναι τυχαία ή παρουσιάζουν κάποιο οποιοδήποτε μοτίβο.



Σχήμα 3.1 : Time Series Plots των στατιστικών στοιχείων με τη πάροδο των σεζόν.

Στο παραπάνω γράφημα, στον κατακόρυφο άξονα υπάρχουν οι τιμές για το μέσο όρο κάθε στατιστικού στοιχείου, και στον οριζόντιο άξονα υπάρχουν οι τιμές των σεζόν. Υπενθυμίζεται πως η σεζόν (season) 1 αντιστοιχεί στη σεζόν 2000-2001, η σεζόν 2 στη σεζόν 2001-2002 κ.ο.κ.

Κάποια συμπεράσματα που μπορούμε να βγάλουμε είναι τα εξής : Από την σεζόν 2010-2011 και μετά, υπάρχει μια απότομη, ανοδική τάση για τους πόντους, για τα

ποσοστά εύστοχων δίποντων και τριπόντων, τα ποσοστά εύστοχων ελευθέρων βολών, για τα αμυντικά ριμπάουντ, τα συνολικά ριμπάουντ, για τις ασίστ και το συνολικό δείκτη PIR για τις ομάδες. Το παραπάνω συμπέρασμα ίσως οφείλεται στο γεγονός πως από τη σεζόν 2010-2011, υπήρξαν κάποιες αλλαγές στους κανονισμούς. Αρχικά, η γραμμή των τριών πόντων μετακινήθηκε πίσω, στα 6,75 μέτρα (ήταν στα 6,25 μ.) και το σχήμα της περιοχής «κλειδί»[16] , γνωστή και ως «ζωγραφιστό», άλλαξε από τραπεζοειδές σε ορθογώνιο. Επιπροσθέτως, εισήχθη ένα τόξο[17] εντός του «ζωγραφιστού», με οριακά ευρύτερη ακτίνα (1,25 μέτρα), μέσα στο οποίο δεν χρεώνονται οι παίκτες επιθετικά φάουλ, υπήρξαν τροποποιήσεις στο κανόνα των 24 δευτερολέπτων, ενώ υπήρξε αλλαγή και σχετικά με το σημείο του γηπέδου από το οποίο θα γινόταν επαναφορά της μπάλας, μετά από timeout, στα τελευταία δύο λεπτά του αγώνα. Τέλος, οι ποινές για τα φάουλ, και ειδικότερα για τα αντιαθλητικά φάουλ και συμπεριφορές, έγιναν πιο αυστηρές, ενώ αυτές για την παράνομη επιστροφή της μπάλας στο μετόπισθεν (backcourt violation), έγιναν ηπιότερες. Όλες οι παραπάνω αλλαγές οδήγησαν σε γρηγορότερες επιθέσεις από τις ομάδες, αύξηση στον αριθμό των κατοχών της μπάλας, καθώς και των ριμπάουντ, μιας και τώρα δε δινόντουσαν με την ίδια ευχέρεια επιθετικά φάουλ.

Αντιθέτως, υπάρχει πτωτική πορεία όσο περνάνε οι σεζόν, για τα κλεψίματα και τα λάθη, για τα φάουλ στα οποία υπέπεσαν οι ομάδες και για τα φάουλ που δέχθηκαν. Τα μπλοκ τα οποία έκαναν οι ομάδες και τα μπλοκ που δέχθηκαν φαίνεται να έφτασαν τη μέγιστη τιμή τους γύρω στην σεζόν 2014-2015 και μετά άρχισαν να ακολουθούν μια φθίνουσα πορεία. Τα επιθετικά ριμπάουντ έχουν πολύ μεγάλη διακύμανση, καθώς περνάνε οι σεζόν.

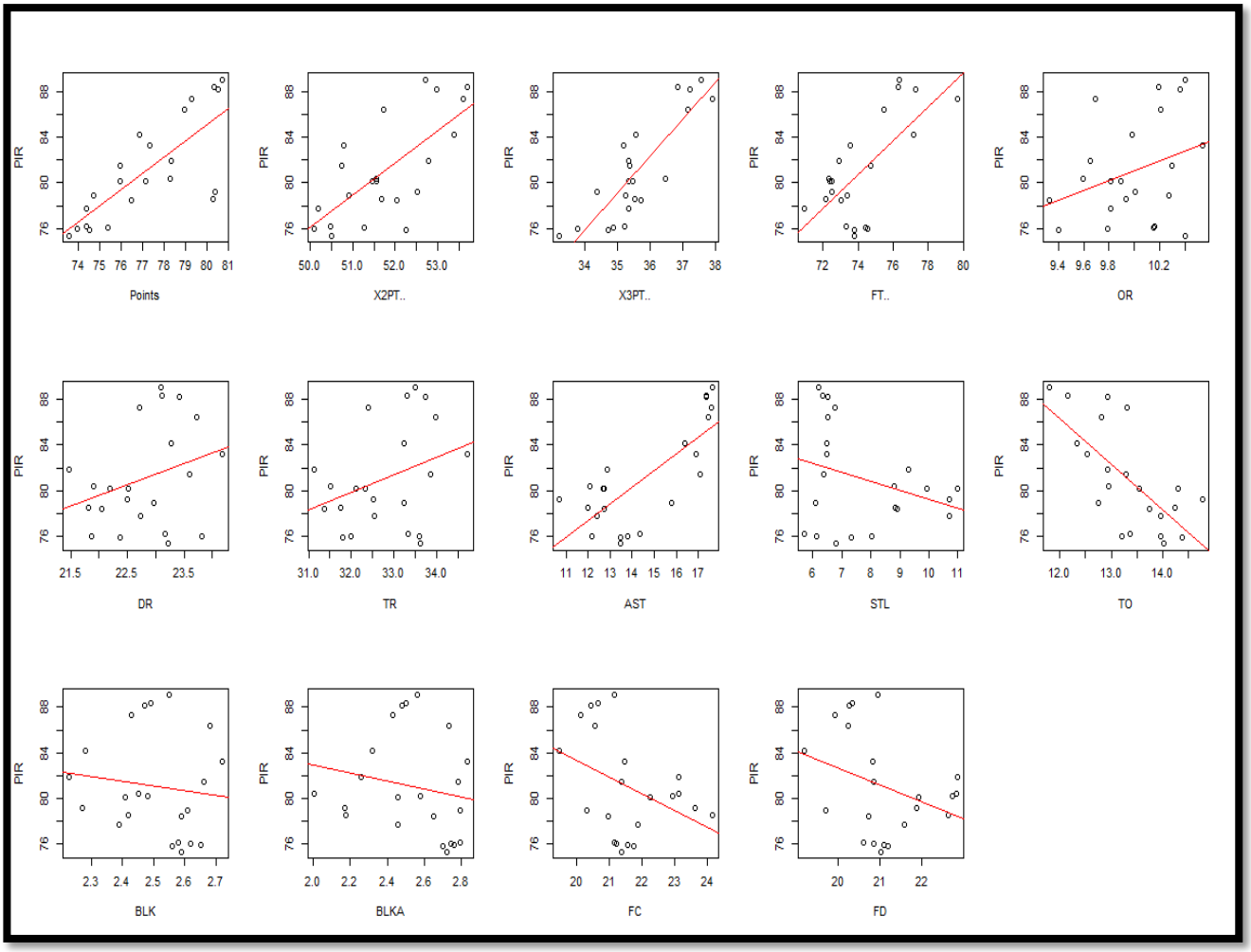
Μια έρευνα (Ibañez et al., 2018), κατέληξε σε παρόμοια συμπεράσματα, αναφορικά με τη σχέση των αλλαγών στους κανονισμούς και της αύξησης των ποσοστών ευστοχίας στα σουτ γενικότερα, την ελάττωση των λαθών, των κλεψιμάτων και των φάουλ στα οποία υπέπεσαν οι παίκτες, μεταξύ άλλων.

Στη συνέχεια, χρησιμοποιώντας διαγράμματα διασποράς (scatter plots), αναπαραστήσαμε τα ζεύγη που δημιουργούνται μεταξύ του δείκτη PIR και των υπόλοιπων μεταβλητών της ανάλυσης. Ένα διάγραμμα διασποράς είναι ένας τύπος γραφικής παράστασης που χρησιμοποιεί καρτεσιανές συντεταγμένες για την εμφάνιση τιμών για δύο μεταβλητές ενός συνόλου δεδομένων. Εάν τα σημεία είναι κωδικοποιημένα (χρώμα/σχήμα/μέγεθος), μπορεί να εμφανιστεί μία επιπλέον μεταβλητή. Τα δεδομένα εμφανίζονται ως μια συλλογή σημείων, καθένα από τα οποία

[16] Το κλειδί, που επίσημα αναφέρεται ως η λωρίδα ελευθέρων βολών, και στην καθομιλουμένη ως ζωγραφιστό, είναι μια περιοχή σε ένα γήπεδο μπάσκετ που περιβάλλει το καλάθι. Οροθετείται από την τελική γραμμή, τη γραμμή των ελεύθερων βολών και τις δύο πλάγιες γραμμές (γραμμές ελεύθερου σώματος), και συνήθως βάφεται σε ένα χαρακτηριστικό χρώμα. Είναι μια κρίσιμη περιοχή στο γήπεδο, όπου λαμβάνει χώρα μεγάλο μέρος της δράσης του παιχνιδιού.
(Πηγή : https://en.wikipedia.org/wiki/Key_(basketball))

[17] Πηγή : https://sportsfanfocus.com/restricted-area-basketball/

έχει τιμή για μια μεταβλητή που καθορίζει τη θέση του στον οριζόντιο άξονα και τιμή μιας άλλης μεταβλητής που καθορίζει τη θέση του στον κατακόρυφο άξονα.



Σχήμα 3.2 : Scatter Plots των στατιστικών στοιχείων με το δείκτη PIR.

Εξετάζοντας συνεπώς τη σχέση του PIR με τις υπόλοιπες μεταβλητές μας, και με βάση το παραπάνω σχήμα, βγάζουμε τα εξής συμπεράσματα :

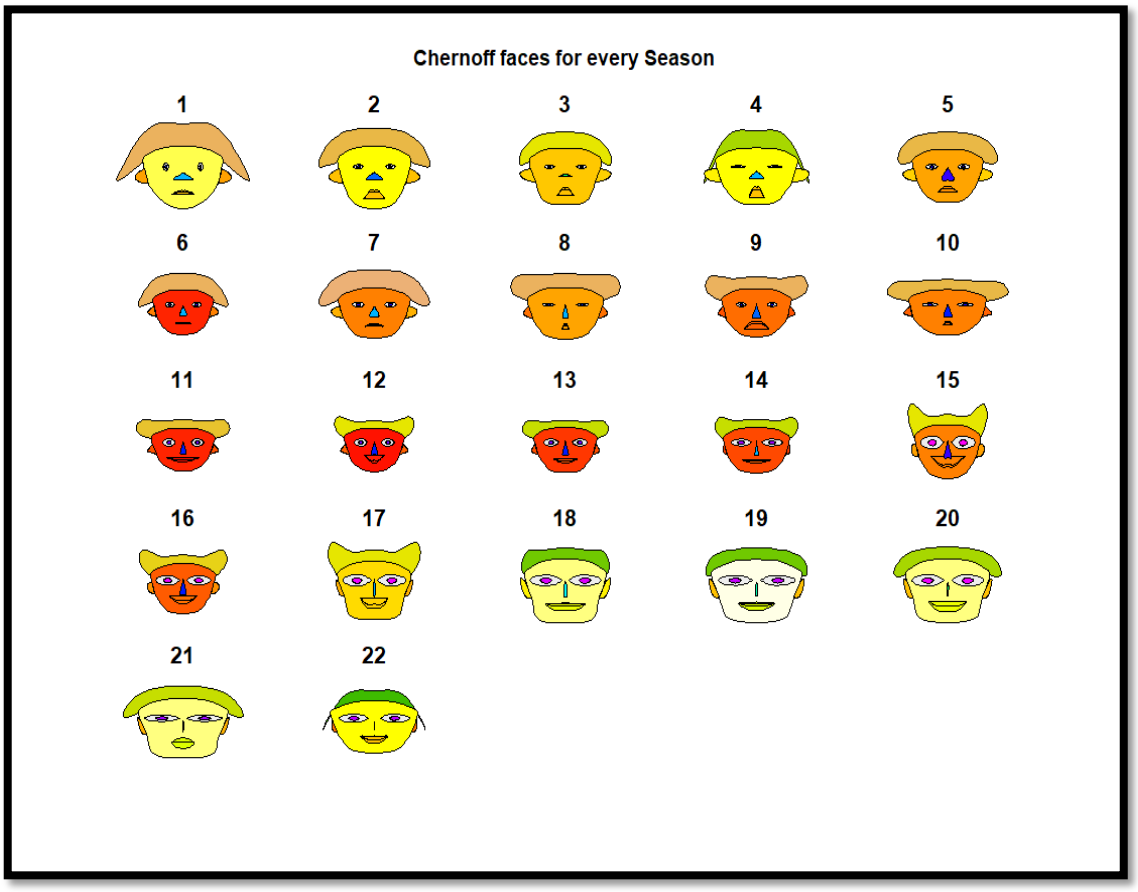
- Ο δείκτης PIR φαίνεται να έχει ισχυρή, θετική συσχέτιση με τους πόντους, το ποσοστό εύστοχων δίποντων, το ποσοστό εύστοχων τρίποντων, το ποσοστό εύστοχων βολών και τις ασίστ.

- Φαίνεται να υπάρχει χαμηλή, θετική συσχέτιση ανάμεσα στο PIR και στα επιθετικά ριμπάουντ, στα αμυντικά ριμπάουντ και στα συνολικά ριμπάουντ.

- Ο δείκτης PIR φαίνεται να έχει χαμηλή, αρνητική συσχέτιση με τα κλεψίματα, τα μπλοκ υπέρ της ομάδας, τα μπλοκ κατά της ομάδας, τα φάουλ στα οποία υπέπεσε μια ομάδα και τα φάουλ που κέρδισε μια ομάδα.
- Φαίνεται να υπάρχει υψηλή, αρνητική συσχέτιση ανάμεσα στο PIR και στα λάθη.

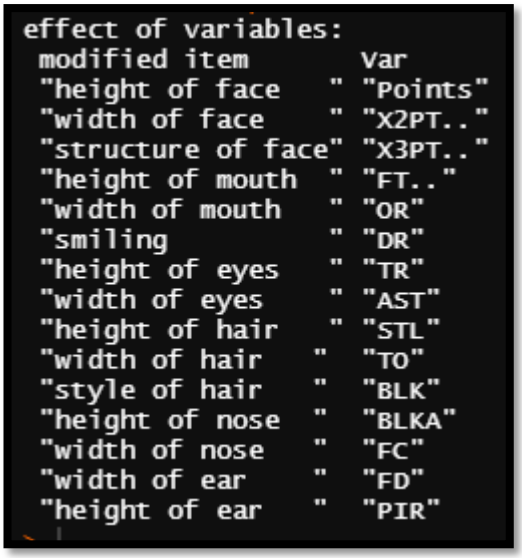
Στο τέταρτο κεφάλαιο, στην ενότητα 4.2.1, γίνεται ανάλυση των συσχετίσεων μεταξύ του δείκτη PIR και όλων των υπόλοιπων μεταβλητών, καθώς και των συμπερασμάτων τα οποία προκύπτουν.

Τέλος, χρησιμοποιήσαμε κάποιες αναπαραστάσεις που λέγονται Chernoff faces, προκειμένου να παρατηρήσουμε με ένα αρκετά ενδιαφέρον και κατανοητό τρόπο, το πως διαφέρουν οι κάθε σεζόν μεταξύ τους. Σύμφωνα με αυτή τη μέθοδο αναπαράστασης δεδομένων, σε κάθε χαρακτηριστικό του προσώπου αντιστοιχίζεται μια μεταβλητή, και η τελική μορφή συνεπώς κάθε προσώπου εξαρτάται από την αντιστοίχιση αυτή των μεταβλητών στα χαρακτηριστικά του. Η τεχνική στηρίζεται στη ευαισθησία του ανθρώπινου ματιού να εντοπίζει ακόμα και τις μικρότερες διαφορές στα πρόσωπα, καθώς κάποια χαρακτηριστικά έχουν μεγαλύτερη αναγνωρισιμότητα από άλλα. Εφαρμόζεται σε περιορισμένο πλήθος μεταβλητών, και εντοπίζει συσχετίσεις μεταξύ των μεταβλητών, καθώς και ακραίες παρατηρήσεις. (Κούτρας, 2021)



Σχήμα 3.3 : Chernoff faces για κάθε σεζόν.

Υπενθυμίζεται πως η σεζόν (season) 1 αντιστοιχεί στη σεζόν 2000-2001, η σεζόν 2 στη σεζόν 2001-2002 κ.ο.κ.



Σχήμα 3.4 : Chernoff faces, effect of variables.

Στο παραπάνω σχήμα παρουσιάζεται το πως το κάθε στατιστικό στοιχείο ανά σεζόν, επηρεάζει τη διαμόρφωση των προσώπων.

Συμπεραίνουμε πως οι σεζόν 11, 12, 13, 14 μοιάζουν αρκετά μεταξύ τους, όπως παραδείγματος χάρη στους πόντους, στα εύστοχα δίποντα, τρίποντα και ελεύθερες βολές, στα μπλοκ υπέρ των ομάδων. Ομοίως, φαίνεται να έχουν αρκετά κοινά στοιχεία οι σεζόν 1 με 4, οι σεζόν 5 με 10, οι σεζόν 15 με 17 και οι σεζόν 18 με 22. Γενικότερα, παρατηρείται ένα συγκεκριμένο μοτίβο, πως κοντινές σεζόν φαίνεται να έχουν και τη μεγαλύτερη ομοιότητα σχετικά με τα χαρακτηριστικά τους.

3.2.2 Σχέση ανάμεσα στην επίδοση των πέντε καλύτερων παικτών σε PIR, και στη πρόκριση των ομάδων τους

Για την επόμενη ανάλυση, υπολογίσαμε για κάθε ομάδα, τους μέσους όρους των πόντων και του PIR, των 5 καλύτερων παικτών τους, για όλες τις σεζόν. Θέλαμε να μελετήσουμε πώς αυτοί οι μ.ό. διαχωρίζουν τις ομάδες, αναλόγως με το αν προκρίθηκαν στα playoffs και στο Final Four. Όπως προαναφέρθηκε, στις αναλύσεις μας δε λήφθηκαν υπόψιν οι σεζόν 2001-2004 και η σεζόν 2019-2020.

Για την ανάλυση των ομάδων που προκρίθηκαν ή όχι στα προημιτελικά / Playoffs

Ομάδες που ΔΕΝ προκρίθηκαν	Min.	1st. Qu.	Median	Mean	3rd. Qu.	Max.
Points	7,46	9,86	10,48	10,69	11,48	15,16
PIR	8,20	10,52	11,36	11,50	12,42	16,12

Πίνακας 3.2 : Στατιστικά περιγραφικά μέτρα για τις ομάδες που δεν προκρίθηκαν στα playoffs.

Ομάδες που προκρίθηκαν	Min.	1st. Qu.	Median	Mean	3rd. Qu.	Max.
Points	8,16	10,09	10,92	10,96	11,82	14,42
PIR	9,20	11,54	12,48	12,53	13,40	17,08

Πίνακας 3.3 : Στατιστικά περιγραφικά μέτρα για τις ομάδες που προκρίθηκαν στα playoffs.

Για την ανάλυση των ομάδων που προκρίθηκαν ή όχι στα ημιτελικά / Final Four

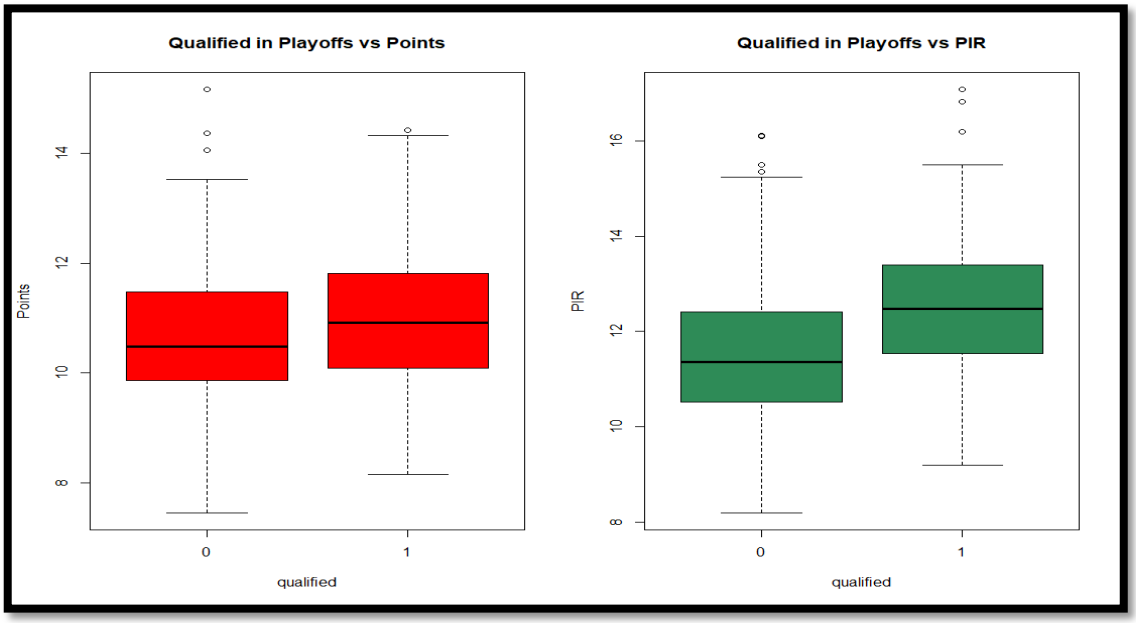
Ομάδες που ΔΕΝ προκρίθηκαν	Min.	1st. Qu.	Median	Mean	3rd. Qu.	Max.
Points	7,46	9,90	10,57	10,76	11,56	15,16
PIR	8,20	10,70	11,60	11,69	12,59	16,82

Πίνακας 3.4 : Στατιστικά περιγραφικά μέτρα για τις ομάδες που δεν προκρίθηκαν στο Final Four.

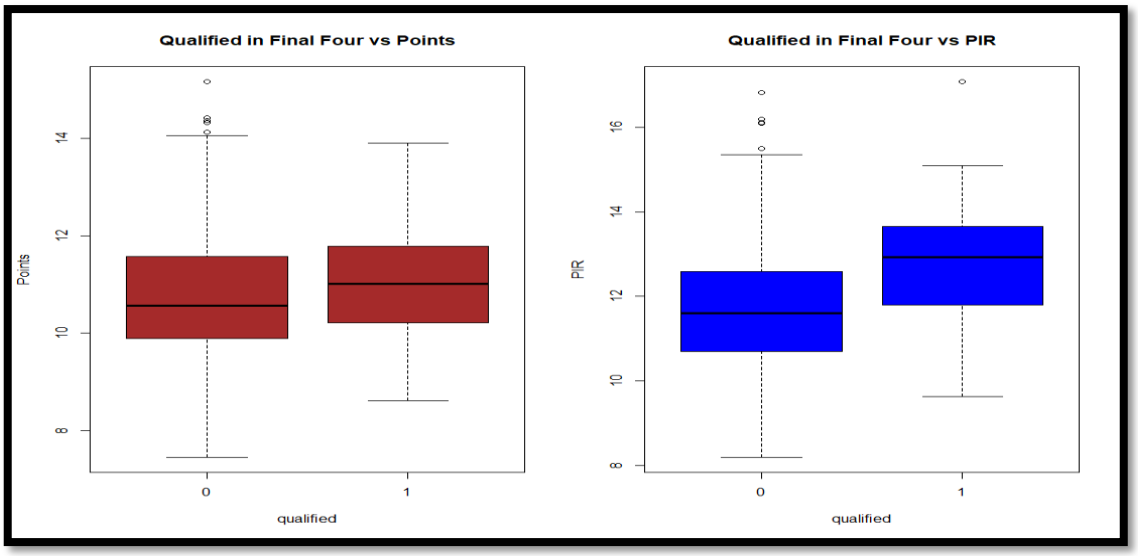
Ομάδες που προκρίθηκαν	Min.	1st. Qu.	Median	Mean	3rd. Qu.	Max.
Points	8,62	10,24	11,01	10,95	11,79	13,90
PIR	9,64	11,82	12,93	12,81	13,64	17,08

Πίνακας 3.5 : Στατιστικά περιγραφικά μέτρα για τις ομάδες που προκρίθηκαν στο Final Four.

Μια άλλη μέθοδος γραφικής ανάλυσης που χρησιμοποιήθηκε για τις συνεχείς μεταβλητές είναι το θηκόγραμμα (boxplot). Με αυτά τα γραφήματα, φαίνονται αναλυτικά σε ποιο ακριβώς σημείο απεικονίζονται οι έκτροπες τιμές (outliers), η ελάχιστη και μέγιστη τιμή, το ενδοτεταρτημοριακό εύρος, το πρώτο και τρίτο τεταρτημόριο, καθώς και η διάμεσος. Η κατασκευή θηκογραμμάτων είναι ιδιαίτερα χρήσιμη στην περίπτωση που θέλουμε να ελέγξουμε αν υπάρχουν έκτροπες τιμές στα δεδομένα που χρησιμοποιούμε, ή ακόμη και ακραίες τιμές (extreme values), οι οποίες συνήθως εξαιρούνται από την ανάλυση. (Καλλιακμάνης, 2020)



Σχήμα 3.5 : Boxplots για τη σχέση των πόντων και του PIR, αναφορικά με τη πρόκριση στα playoffs.

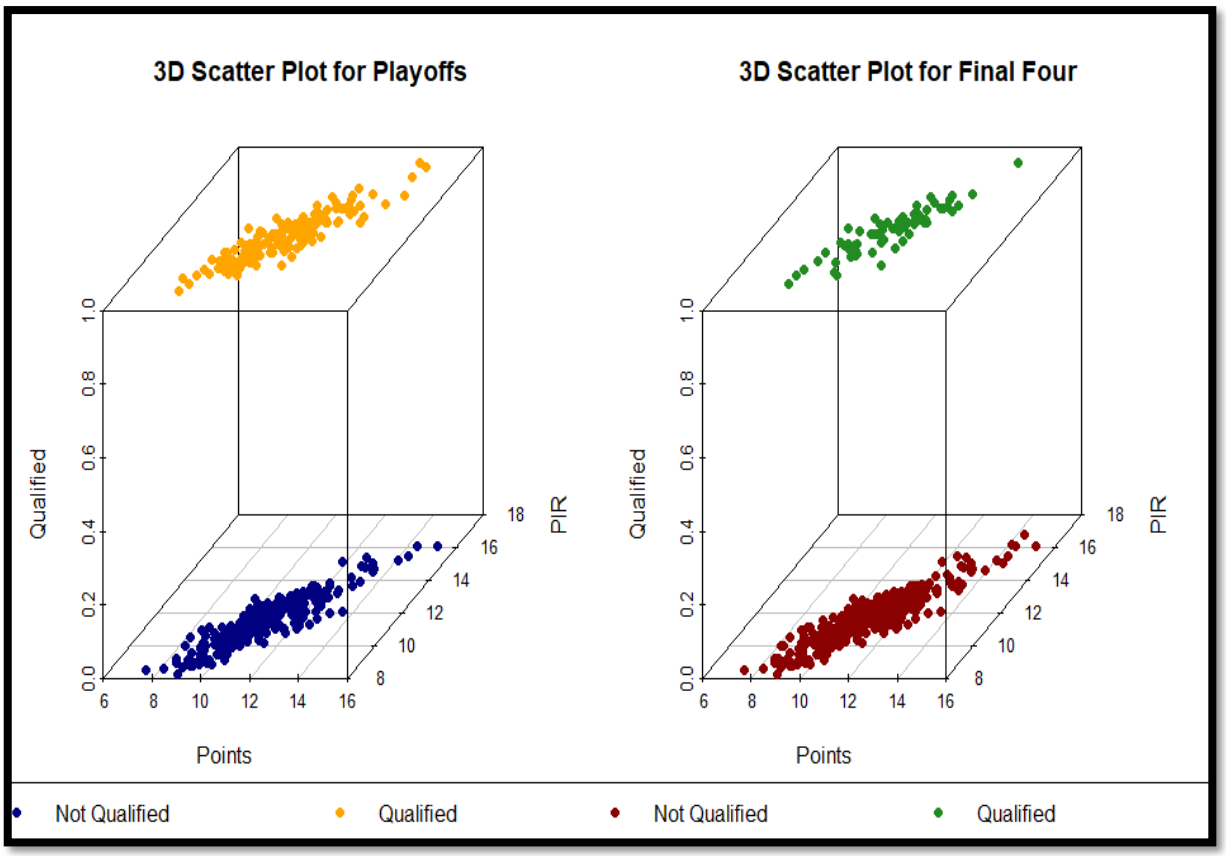


Σχήμα 3.6 : Boxplots για τη σχέση των πόντων και του PIR, αναφορικά με τη πρόκριση στο Final Four.

Παρατηρούμε πως η διάμεσος για τις ομάδες που προκρίθηκαν (qualified = 1) βρίσκεται υψηλότερα από ότι για τις ομάδες που δεν προκρίθηκαν, τόσο αναφορικά για τους πόντους (αριστερά boxplots στα δύο σχήματα), όσο και για το δείκτη PIR (δεξιά boxplots στα δύο σχήματα), κάτι το οποίο ήταν και αναμενόμενο. Παράλληλα, υπάρχουν τρεις έκτροπες τιμές όσον αφορά τους πόντους στις ομάδες που δεν προκρίθηκαν στα playoffs, τις ομάδες Hapoel Jerusalem (2000-2001), Nizhny Novgorod (2014-2015) και Armani Milano (2018-2019), και δύο για αυτές που προκρίθηκαν, τις Barcelona (2000-2001) και PAOK (2000-2001). Επιπροσθέτως, εμφανίστηκαν δύο έκτροπες τιμές όσον αφορά το PIR στις ομάδες που δεν προκρίθηκαν στα playoffs, τις Hapoel Jerusalem (2000-2001) και Nizhny Novgorod (2014-2015), καθώς και τρεις για αυτές που προκρίθηκαν, τις ομάδες Benetton Basket (2000-2001), PAOK (2000-2001) και Maccabi Tel-Aviv (2004-2005). Υπάρχουν πέντε outliers αναφορικά με τους πόντους στις ομάδες που δεν προκρίθηκαν στο Final Four, τις Hapoel Jerusalem (2000-2001), Nizhny Novgorod (2014-2015), Armani Milano (2018-2019), Barcelona (2000-2001) και PAOK (2000-2001), και κανένα για αυτές που προκρίθηκαν. Τέλος, εμφανίζονται τέσσερα outliers όσον αφορά το PIR στις ομάδες που δεν προκρίθηκαν στο Final Four, τις Benetton Basket (2000-2001), Hapoel Jerusalem (2000-2001), PAOK (2000-2001) και Nizhny Novgorod (2014-2015), και ένα για αυτές που προκρίθηκαν, τη Maccabi Tel-Aviv (2004-2005). Ολοκληρώνοντας, δε φαίνεται κάποια έντονη ασσυμετρία, είτε θετική, είτε αρνητική.

Ακολουθούν τώρα κάποια τρισδιάστατα διαγράμματα διασποράς (3D scatter plots). Τα τρισδιάστατα διαγράμματα διασποράς χρησιμοποιούνται για τη σχεδίαση των σημείων των δεδομένων σε τρεις άξονες, στην προσπάθεια να φανεί η σχέση μεταξύ των τριών μεταβλητών. Κάθε γραμμή στον πίνακα δεδομένων αντιπροσωπεύεται από έναν δείκτη, του οποίου η θέση εξαρτάται από τις τιμές του στις στήλες που ορίζονται στους άξονες X, Y και Z. Εδώ, ο άξονας X αντιστοιχεί στους πόντους, ο Y στο PIR και

ο Z στη κατηγορική μεταβλητή για τη πρόκριση της ομάδας στα playoffs και στο Final Four.



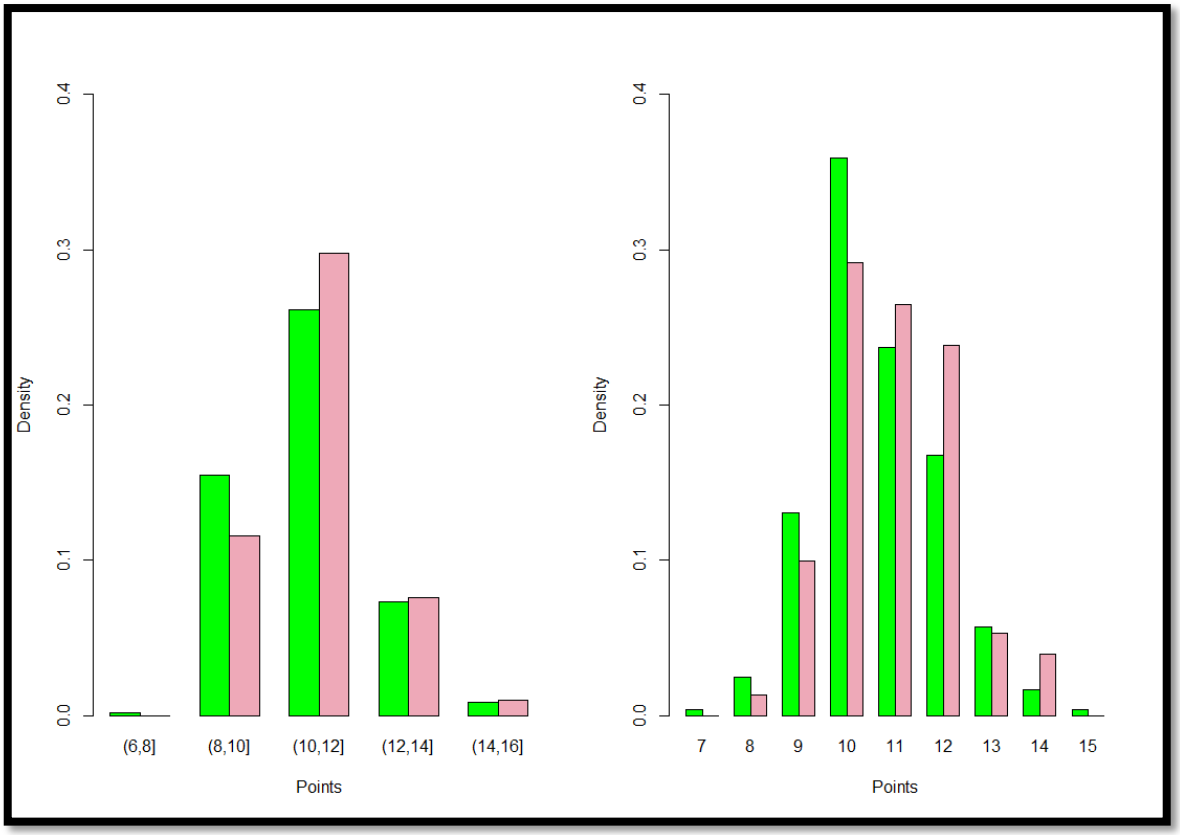
Σχήμα 3.7 : 3D scatter plots για τα playoffs και το Final Four.

Παρατηρώντας το παραπάνω σχήμα, μπορέσαμε να δούμε καθαρά πώς κατανέμονται οι ομάδες στο τρισδιάστατο χώρο, αναλόγως αν πήραν τη πρόκριση για τα playoffs (αριστερό σχήμα) και για το Final Four (δεξιό σχήμα).

Τέλος, παρουσιάζονται ιστογράμματα (histograms) για τους μέσους όρους των πόντων, και διαγράμματα πυκνότητας (density charts) για τους μέσους όρους του δείκτη PIR, προκειμένου να οπτικοποιήσουμε τα δεδομένα και να βγάλουμε κάποια επιπλέον συμπεράσματα.

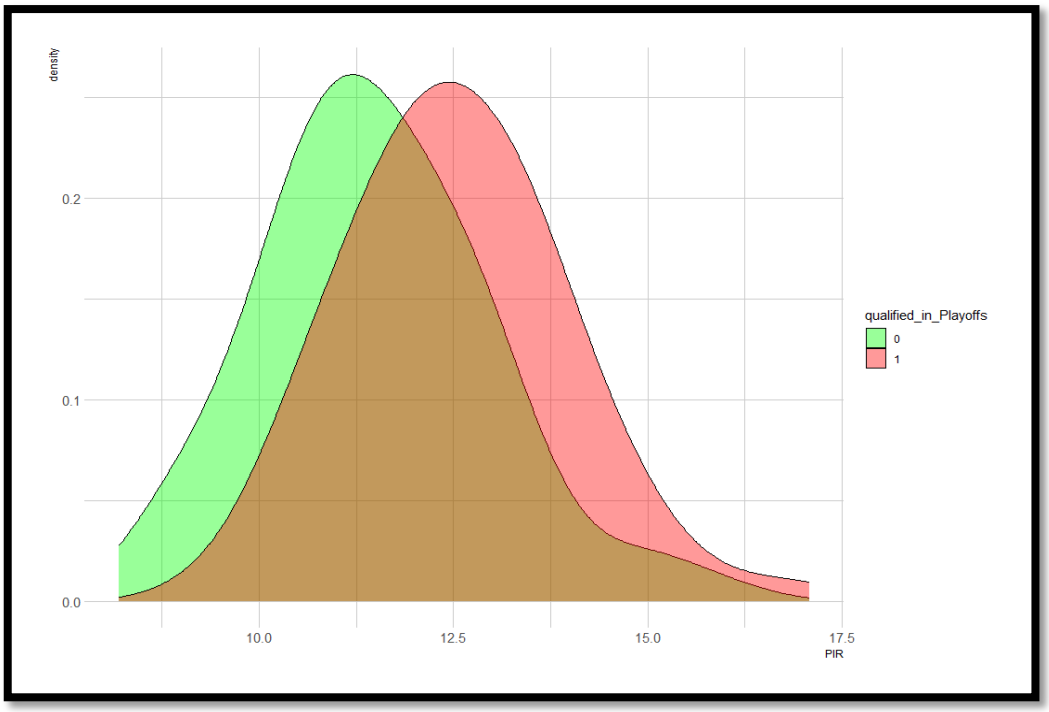
Το ιστόγραμμα είναι ένα δημοφιλές εργαλείο δημιουργίας γραφημάτων. Χρησιμοποιείται για τη σύνοψη διακριτών ή συνεχών δεδομένων που μετρούνται σε κλίμακα διαστήματος. Συχνά χρησιμοποιείται για να απεικονίσει τα κύρια χαρακτηριστικά της διανομής των δεδομένων σε μια βολική μορφή. Είναι επίσης χρήσιμο όταν ασχολούμαστε με μεγάλα σύνολα δεδομένων (περισσότερες από 100 παρατηρήσεις). Μπορεί να βοηθήσει στον εντοπισμό τυχόν ασυνήθιστων (έκτροπων) παρατηρήσεων (outliers) ή τυχόν κενών στα δεδομένα.

Τα διαγράμματα πυκότητας οπτικοποιούν την κατανομή των δεδομένων σε μια δεδομένη περίοδο, ενώ οι κορυφές δείχνουν το πού συγκεντρώνονται οι τιμές. Τα διαγράμματα πυκνότητας χρησιμοποιούνται επίσης για την παρατήρηση του σχήματος της κατανομής μιας μεταβλητής, σε ένα σύνολο δεδομένων. Στον κατακόρυφο άξονα είναι η συνάρτηση πυκνότητας πιθανότητας για την εκτίμηση της πυκνότητας του πυρήνα, ενώ στον οριζόντιο είναι οι τιμές της εκάστοτε μεταβλητής.



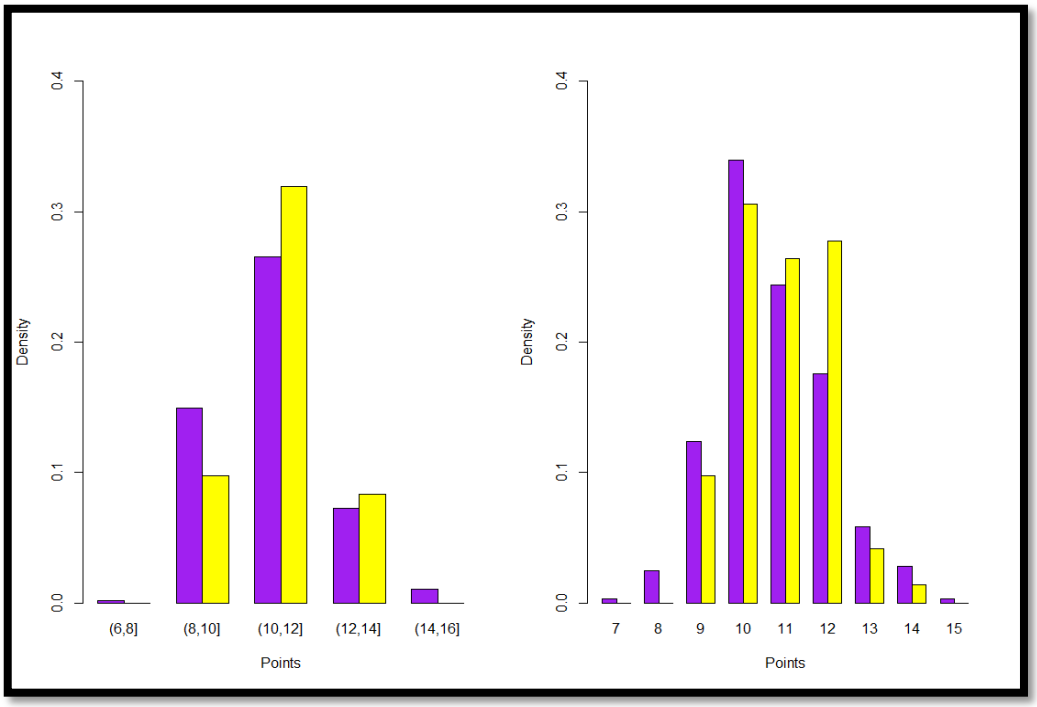
Σχήμα 3.8 : Ιστογράμματα για τους πόντους, σε σχέση με τη πρόκριση στα playoffs.

Στο παραπάνω σχήμα, παρουσιάζονται τα ιστογράμματα για τους μ.ό. των πόντων για τις ομάδες που δε προκρίθηκαν στα playoffs (πράσινο χρώμα) και αυτές που προκρίθηκαν (ροζ χρώμα). Αριστερά, στον οριζόντιο άξονα, οι πόντοι είναι χωρισμένοι σε κλάσεις των δύο, ενώ δεξιά οι τιμές είναι πιο «μεμονωμένες». Οι μέσοι όροι των πόντων φαίνεται να κατανέμονται κανονικά, τόσο για τις ομάδες που προκρίθηκαν, όσο και για αυτές που δεν προκρίθηκαν. Στον κατακόρυφο άξονα παρουσιάζεται η συνάρτηση πυκνότητας πιθανότητας, για την εκτίμηση της πυκνότητας του πυρήνα. Οι παίκτες των ομάδων που δε προκρίθηκαν στα playoffs φαίνεται να επιτυγχάνουν 10-12 πόντους κατά μέσο όρο, με μια συχνότητα μεγαλύτερη του 25%, και αυτοί των ομάδων που προκρίθηκαν, με συχνότητα περίπου 30%.



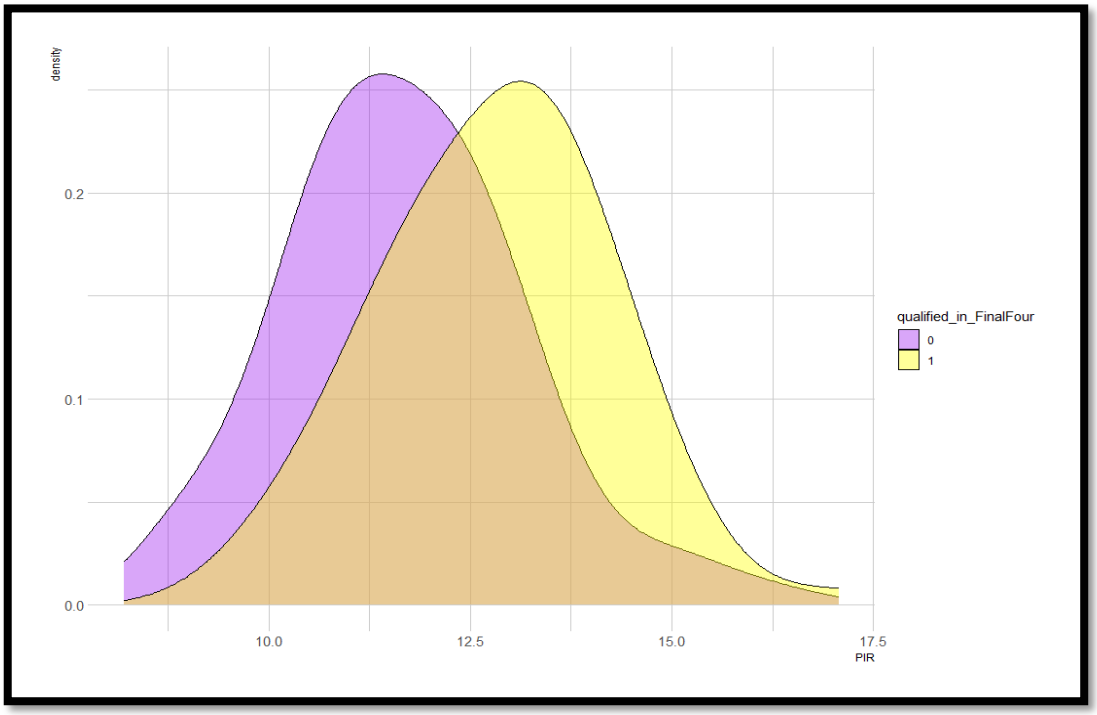
Σχήμα 3.9 : Density charts για το PIR, σε σχέση με τη πρόκριση στα playoffs.

Σύμφωνα με τα παραπάνω density charts, οι μέσοι όροι του PIR φαίνεται να κατανέμονται κανονικά, τόσο για τις ομάδες που δεν προκρίθηκαν στα playoffs, όσο και για αυτές που προκρίθηκαν. Η μέση τιμή των πόντων για τις ομάδες που δεν προκρίθηκαν, όπως είναι λογικό, είναι μικρότερη (πιο «αριστερά» στο σχήμα) σε σχέση με αυτή των ομάδων που προκρίθηκαν.



Σχήμα 3.10 : Ιστογράμματα για τους πόντους, σε σχέση με τη πρόκριση στο Final Four.

Στο παραπάνω σχήμα, παρουσιάζονται τα ιστογράμματα για τους μ.ό. των πόντων για τις ομάδες που δε προκρίθηκαν στο Final Four (μωβ χρώμα) και αυτές που προκρίθηκαν (κίτρινο χρώμα). Οι παίκτες των ομάδων που δε προκρίθηκαν στο Final Four φαίνεται να επιτυγχάνουν 10-12 πόντους κατά μέσο όρο, με μια συχνότητα μεγαλύτερη του 25%, και αυτοί των ομάδων που προκρίθηκαν, με συχνότητα κοντά στο 34%.



Σχήμα 3.11 : Density charts για το PIR, σε σχέση με τη πρόκριση στο Final Four.

Σύμφωνα με το παραπάνω σχήμα των density charts, οι μέσοι όροι του PIR φαίνεται να κατανέμονται κανονικά, τόσο για τις ομάδες που δε προκρίθηκαν στο Final Four, όσο και για αυτές που προκρίθηκαν. Ομοίως και εδώ, η μέση τιμή των πόντων για τις ομάδες που δεν προκρίθηκαν στο Final Four, όπως είναι λογικό, είναι μικρότερη (πιο «αριστερά» στο σχήμα) σε σχέση με αυτή των ομάδων που προκρίθηκαν.

Ο αναλυτικός έλεγχος κανονικότητας των μεταβλητών μας πραγματοποιείται στο τέταρτο κεφάλαιο, στην ενότητα 4.1.2 .

3.2.3 Σχέση ανάμεσα στην επίδοση των δέκα καλύτερων παικτών σε πόντους και των δέκα σε PIR, ανά σεζόν, και στη πρόκριση των ομάδων τους

Στη συνέχεια της ανάλυσής μας, πήραμε τους δέκα κορυφαίους παίκτες σε πόντους και PIR ανά σεζόν και είδαμε πόσοι από αυτούς αντιστοιχούν σε ομάδες που προκρίθηκαν στα playoffs, και σε ομάδες που προκρίθηκαν στο Final Four.

Για την ανάλυση των δέκα κορυφαίων παικτών σε πόντους, σε σχέση με τις ομάδες που προκρίθηκαν ή όχι στα προημιτελικά / Playoffs

Ομάδες που ΔΕΝ προκρίθηκαν	Min.	1st. Qu.	Median	Mean	3rd. Qu.	Max.
Points	13	14,97	15,80	16,56	17,30	26,00
PIR	11,60	14,30	16,45	17,12	19,07	30,50

Πίνακας 3.6 : Στατιστικά περιγραφικά μέτρα για τους 108 παίκτες των ομάδων που δεν προκρίθηκαν στα playoffs.

Ομάδες που προκρίθηκαν	Min.	1st. Qu.	Median	Mean	3rd. Qu.	Max.
Points	13,10	15,00	16,20	16,57	17,38	22,90
PIR	12,80	14,88	17,20	17,52	19,52	30,90

Πίνακας 3.7 : Στατιστικά περιγραφικά μέτρα για τους 72 παίκτες των ομάδων που προκρίθηκαν στα playoffs.

Για την ανάλυση των δέκα κορυφαίων παικτών σε πόντους, σε σχέση με τις ομάδες που προκρίθηκαν ή όχι στα ημιτελικά / Final Four

Ομάδες που ΔΕΝ προκρίθηκαν	Min.	1st. Qu.	Median	Mean	3rd. Qu.	Max.
Points	13	15	15,80	16,62	17,30	26
PIR	11,60	14,30	16,50	17,13	18,95	30,90

Πίνακας 3.8 : Στατιστικά περιγραφικά μέτρα για τους 147 παίκτες των ομάδων που δεν προκρίθηκαν στο Final Four.

Ομάδες που προκρίθηκαν	Min.	1st. Qu.	Median	Mean	3rd. Qu.	Max.
Points	14,50	15	16,10	16,33	17	19,70
PIR	13,40	15,90	17,50	17,96	19,70	25,20

Πίνακας 3.9 : Στατιστικά περιγραφικά μέτρα για τους 33 παίκτες των ομάδων που προκρίθηκαν στο Final Four.

Για την ανάλυση των δέκα κορυφαίων παικτών σε PIR, σε σχέση με τις ομάδες που προκρίθηκαν ή όχι στα προημιτελικά / Playoffs

Ομάδες που ΔΕΝ προκρίθηκαν	Min.	1st. Qu.	Median	Mean	3rd. Qu.	Max.
Points	10	13,10	14,90	15,51	17,27	26
PIR	15	17,05	18,15	18,92	19,95	30,50

Πίνακας 3.10 : Στατιστικά περιγραφικά μέτρα για τους 94 παίκτες των ομάδων που δεν προκρίθηκαν στα playoffs.

Ομάδες που προκρίθηκαν	Min.	1st. Qu.	Median	Mean	3rd. Qu.	Max.
Points	9,80	13,32	14,70	15,21	16,70	22,90
PIR	14,60	17,30	18,05	18,82	19,70	30,90

Πίνακας 3.11 : Στατιστικά περιγραφικά μέτρα για τους 86 παίκτες των ομάδων που προκρίθηκαν στα playoffs.

Για την ανάλυση των δέκα κορυφαίων παικτών σε PIR, σε σχέση με τις ομάδες που προκρίθηκαν ή όχι στα ημιτελικά / Final Four

Ομάδες που ΔΕΝ προκρίθηκαν	Min.	1st. Qu.	Median	Mean	3rd. Qu.	Max.
Points	10	13,25	14,90	15,58	17,30	26
PIR	15	17,20	18,10	18,96	19,90	30,90

Πίνακας 3.12 : Στατιστικά περιγραφικά μέτρα για τους 131 παίκτες των ομάδων που δεν προκρίθηκαν στο Final Four.

Ομάδες που προκρίθηκαν	Min.	1st. Qu.	Median	Mean	3rd. Qu.	Max.
Points	9,80	13,30	14,60	14,79	16,30	19,70
PIR	14,60	17	17,90	18,66	19,70	25,20

Πίνακας 3.13 : Στατιστικά περιγραφικά μέτρα για τους 49 παίκτες των ομάδων που προκρίθηκαν στο Final Four.

Συνοψίζοντας, και στις δύο αναλύσεις μας σε αυτήν την ενότητα, είχαμε συνολικά 180 παίκτες. Σχετικά με την ανάλυση των δέκα κορυφαίων παικτών σε πόντους, είχαμε 108 παίκτες που αντιστοιχούν σε ομάδες που δεν προκρίθηκαν στα playoffs, και 72 σε ομάδες που προκρίθηκαν. Αντίστοιχα, 147 παίκτες αντιστοιχούν σε ομάδες που δεν προκρίθηκαν στο Final Four, και 33 σε ομάδες που προκρίθηκαν.

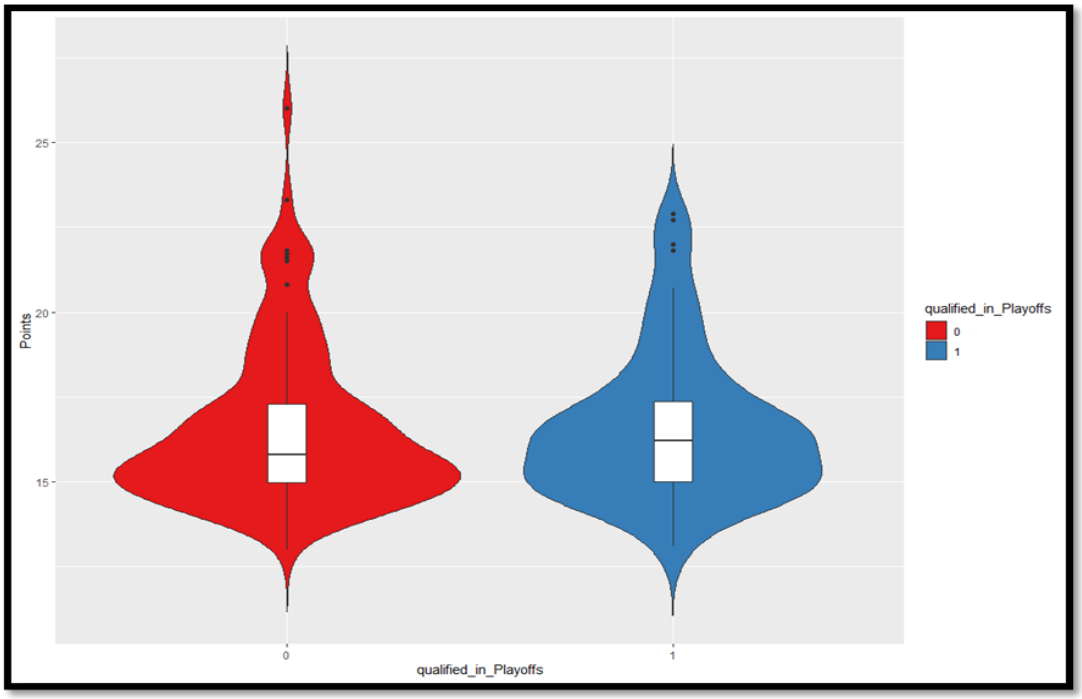
Αναφορικά τώρα με την ανάλυση των δέκα κορυφαίων παικτών σε PIR, είχαμε 94 παίκτες που αντιστοιχούν σε ομάδες που δεν προκρίθηκαν στα playoffs, και 86 σε

ομάδες που προκρίθηκαν. Αντίστοιχα, 131 παίκτες αντιστοιχούν σε ομάδες που δεν προκρίθηκαν στο Final Four, και 49 σε ομάδες που προκρίθηκαν.

Μια παραλλαγή των θηκογραμμάτων αποτελούν τα violin plots, τα οποία κερδίζουν ολοένα και περισσότερο έδαφος όσον αφορά τον τομέα του exploratory data analysis (EDA). Η ερμηνεία είναι παρόμοια με αυτή ενός απλού boxplot. Η μόνη διαφορά τους είναι ότι παρουσιάζουν και μια εκτίμηση της συνάρτησης πυκνότητας για τα δεδομένα. (Καλλιακμάνης, 2020)

Γενικότερα, τα violin plots δίνουν πιο πολλές πληροφορίες από ένα boxplot. Ενώ τα θηκογράμματα εμφανίζουν μόνο συνοπτικά στατιστικά στοιχεία, όπως τη διάμεσο και το διατεταρτημόριο, το διάγραμμα βιολιού δείχνει την πλήρη κατανομή των δεδομένων. Η διαφορά είναι ιδιαίτερα χρήσιμη, όταν η κατανομή των δεδομένων είναι πολυκόρυφη (περισσότερες από μία κορυφές). Σε αυτή την περίπτωση, ένα violin plot δείχνει την ύπαρξη διαφορετικών κορυφών, τη θέση τους και το σχετικό πλάτος τους. (Wikipedia, Violin plot 2022)

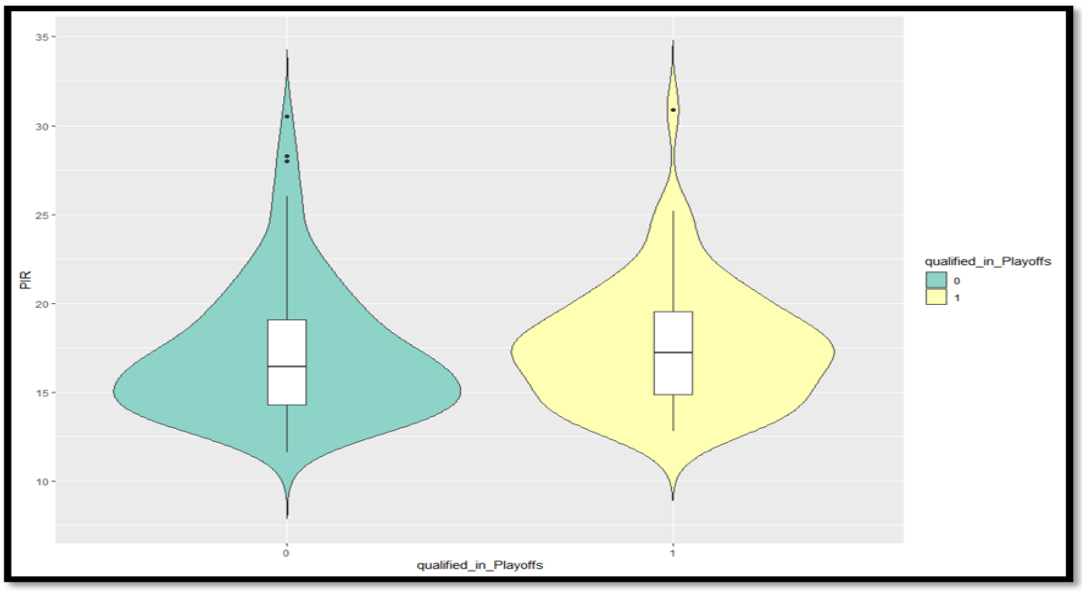
Για την ανάλυση των δέκα κορυφαίων παικτών σε πόντους



Σχήμα 3.12 : Violin plots για τους πόντους σε σχέση με τη πρόκριση στα playoffs.

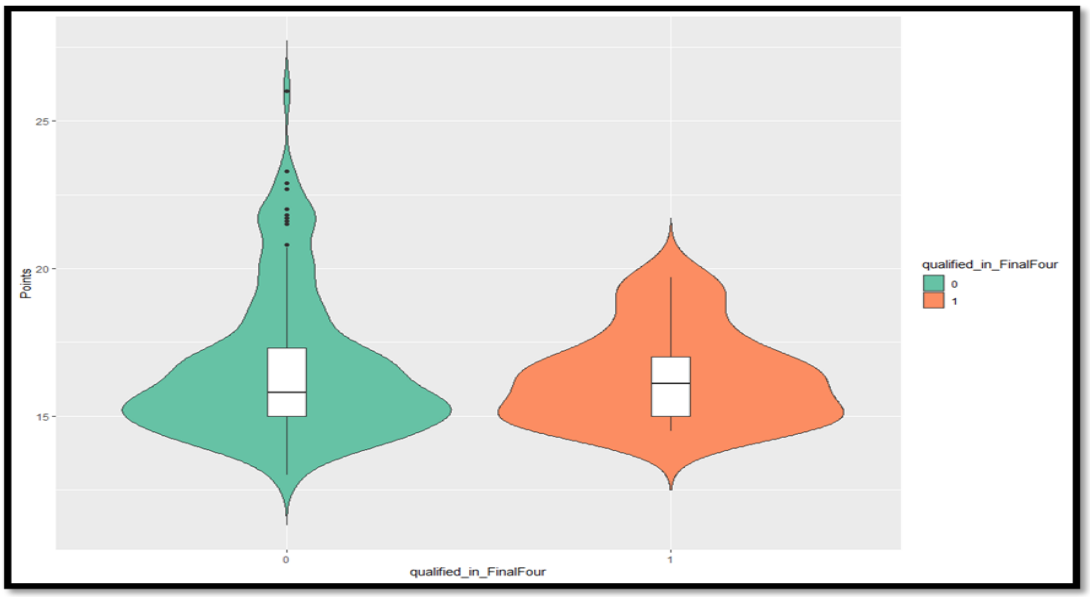
Παρατηρούμε πως η διάμεσος για τις ομάδες των παικτών που προκρίθηκαν στα playoffs (qualified = 1) βρίσκεται υψηλότερα απ' ότι για τις ομάδες που δεν προκρίθηκαν, κάτι το οποίο ήταν και αναμενόμενο. Παράλληλα, φαίνεται να υπάρχουν εννέα έκτροπες τιμές για τις ομάδες που δεν προκρίθηκαν στα playoffs, και τέσσερις

για αυτές που προκρίθηκαν, οι παίκτες Dejan Tomasevic (2000-2001), Panagiotis Liadelis (2000-2001), Louis Bullock (2000-2001) και Alexey Shved (2017-2018).



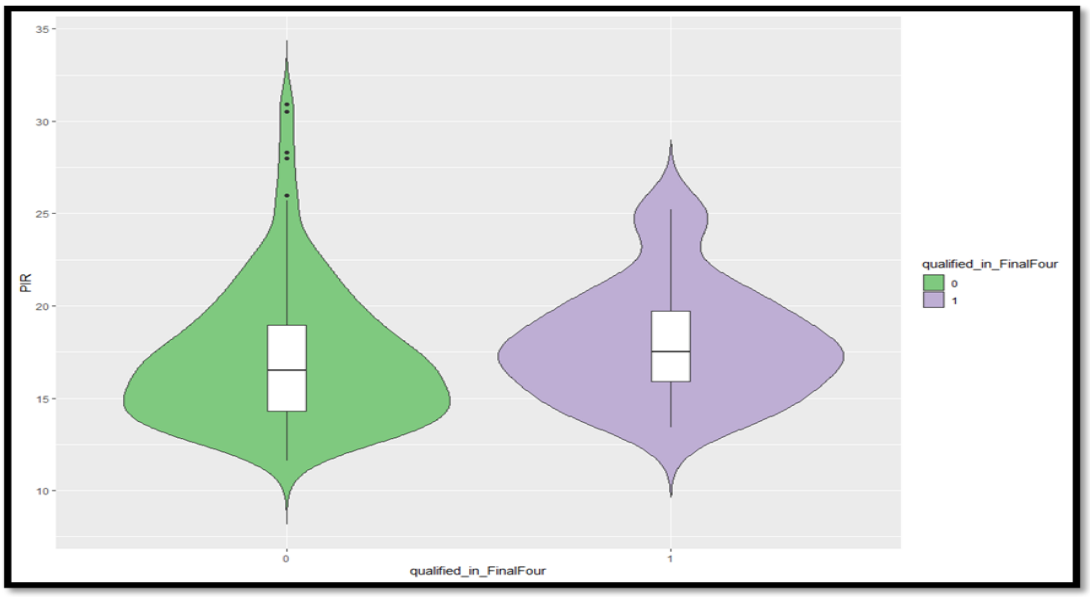
Σχήμα 3.13 : Violin plots για το PIR σε σχέση με τη πρόκριση στα playoffs.

Εξετάζοντας τους παίκτες με το υψηλότερο PIR, παρατηρείται πως η διάμεσος για τις ομάδες των παικτών που προκρίθηκαν στα playoffs βρίσκεται υψηλότερα απ' ότι για τις ομάδες που δεν προκρίθηκαν, κάτι το οποίο ήταν αναμενόμενο και εδώ. Παράλληλα, φαίνεται να υπάρχουν τρεις έκτροπες τιμές για τις ομάδες που δεν προκρίθηκαν, οι Dereck Hamilton (2000-2001), Dejan Milojevic (2004-2005) και Curtis Borchardt (2009-2010), και μια για αυτές που προκρίθηκαν, ο Dejan Tomasevic (2000-2001).



Σχήμα 3.14 : Violin plots για τους πόντους σε σχέση με τη πρόκριση στο Final Four.

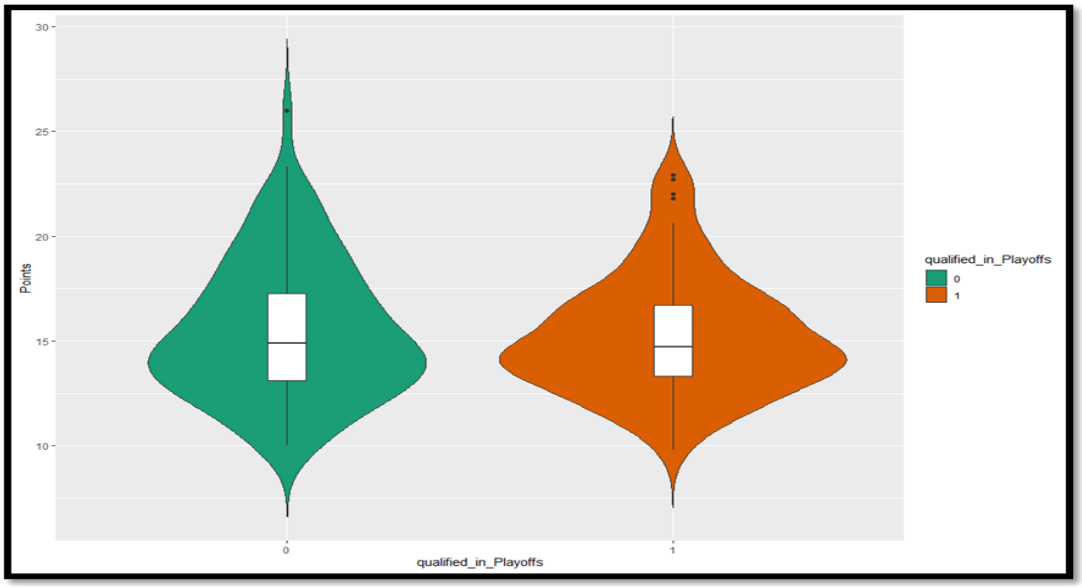
Παρατηρούμε πως η διάμεσος για τις ομάδες των παικτών που προκρίθηκαν στο Final Four βρίσκεται υψηλότερα απ' ότι για τις ομάδες που δεν προκρίθηκαν. Παράλληλα, υπάρχουν πολλές έκτροπες τιμές για τις ομάδες που δεν προκρίθηκαν στο Final Four, και καμία για αυτές που προκρίθηκαν.



Σχήμα 3.15 : Violin plots για το PIR σε σχέση με τη πρόκριση στο Final Four.

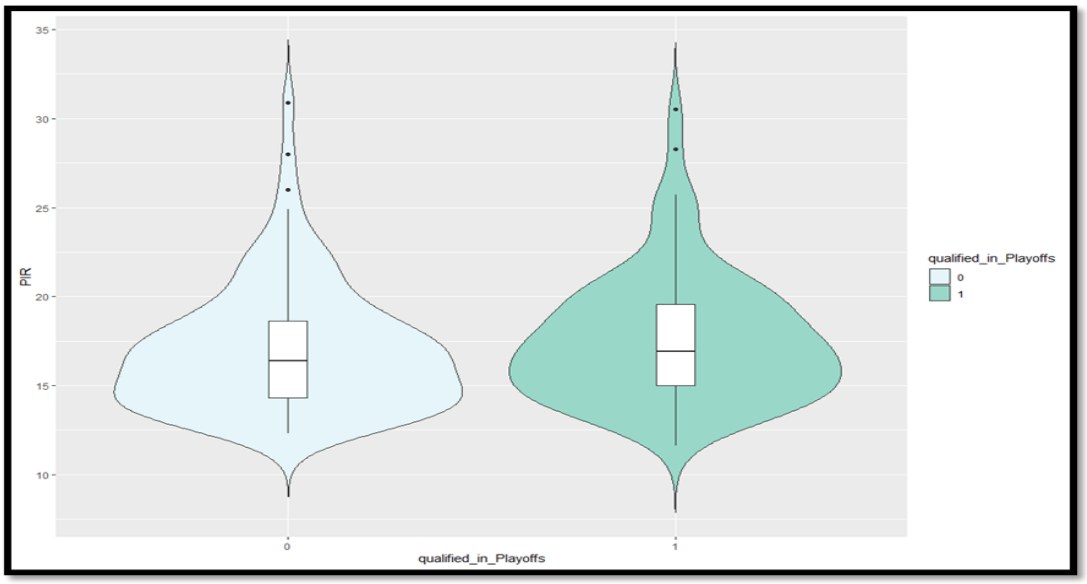
Και σε αυτή τη περίπτωση, η διάμεσος για τις ομάδες των παικτών που προκρίθηκαν στο Final Four βρίσκεται υψηλότερα απ' ότι για τις ομάδες που δεν προκρίθηκαν. Παράλληλα, υπάρχουν πέντε έκτροπες τιμές για τις ομάδες που δεν προκρίθηκαν στο Final Four, και καμία για αυτές που προκρίθηκαν.

Για την ανάλυση των δέκα κορυφαίων παικτών σε PIR.



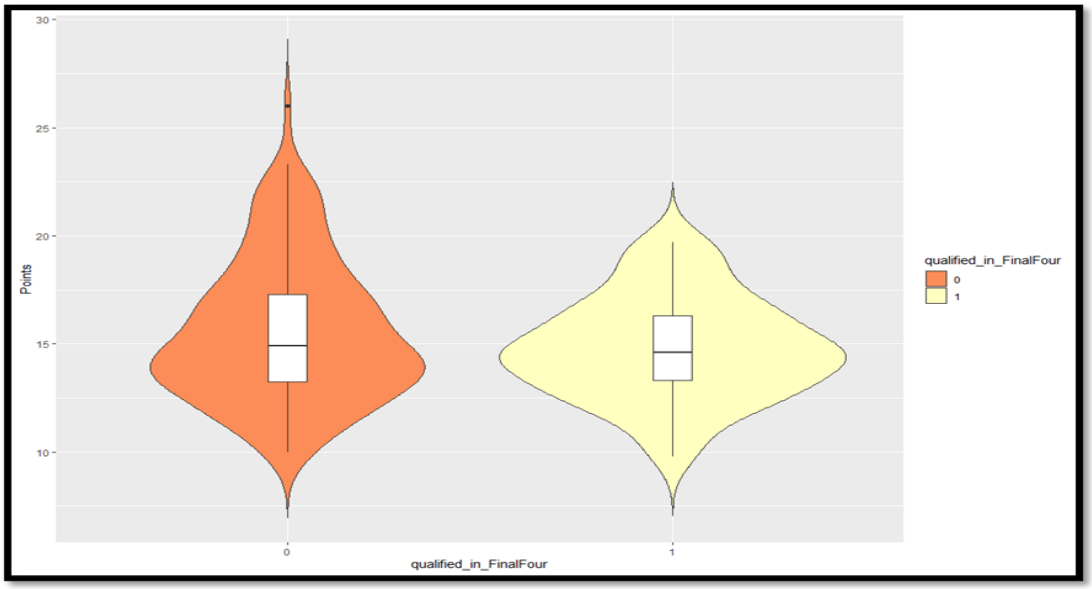
Σχήμα 3.16 : Violin plots για τους πόντους σε σχέση με τη πρόκριση στα playoffs.

Εργαζόμενοι τώρα με τους κορυφαίους παίκτες σε PIR, συμπεραίνουμε πως η διάμεσος για τις ομάδες των παικτών που δεν προκρίθηκαν στα playoffs βρίσκεται υψηλότερα απ' ότι για τις ομάδες που προκρίθηκαν, κάτι το οποίο είναι ενδιαφέρον. Παράλληλα, υπάρχει μια έκτροπη τιμή για τις ομάδες που δεν προκρίθηκαν στα playoffs, ο Alphonso Ford (2000-2001), και τέσσερις για αυτές που προκρίθηκαν.



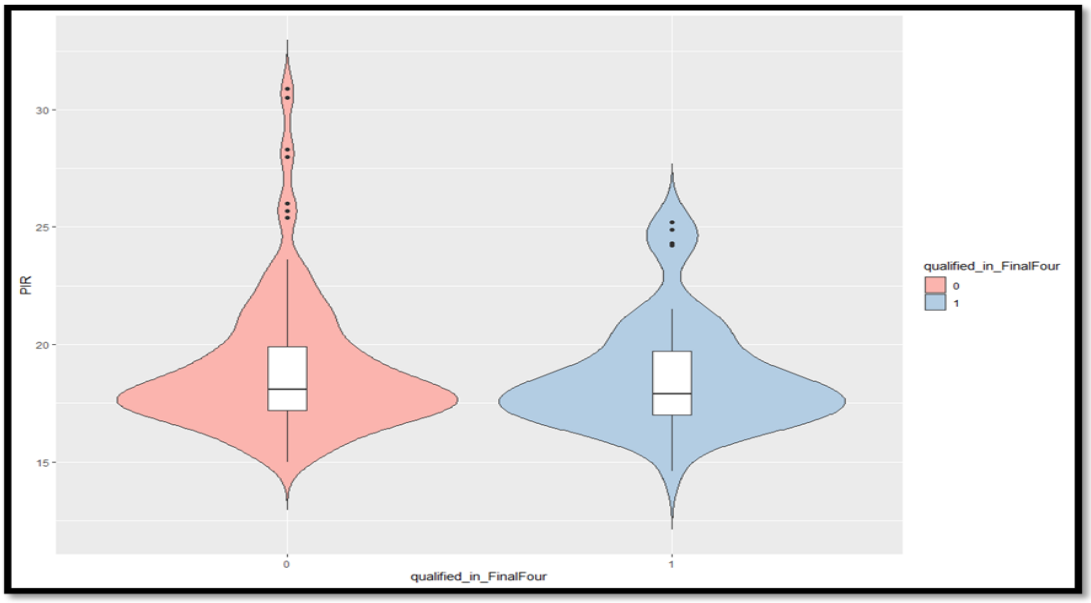
Σχήμα 3.17 : Violin plots για το PIR σε σχέση με τη πρόκριση στα playoffs.

Τώρα, παρατηρείται πως η διάμεσος για τις ομάδες των παικτών που προκρίθηκαν στα playoffs βρίσκεται υψηλότερα απ' ότι για τις ομάδες που δεν προκρίθηκαν. Παράλληλα, υπάρχουν τρία outliers για τις ομάδες που δεν προκρίθηκαν και δύο για αυτές που προκρίθηκαν.



Σχήμα 3.18 : Violin plots για τους πόντους σε σχέση με τη πρόκριση στο Final Four.

Παρατηρούμε πως η διάμεσος για τις ομάδες των παικτών που προκρίθηκαν στο Final Four βρίσκεται χαμηλότερα απ' ότι για τις ομάδες που δεν προκρίθηκαν. Εμφανίζεται μια έκτροπη τιμή για τις ομάδες που δεν προκρίθηκαν στο Final Four, ο Alphonso Ford (2000-2001), και καμία για αυτές που προκρίθηκαν.



Σχήμα 3.19 : Violin plots για το PIR σε σχέση με τη πρόκριση στο Final Four.

Τέλος, η διάμεσος για τις ομάδες των παικτών που προκρίθηκαν στο Final Four βρίσκεται ελαφρώς χαμηλότερα απ' ότι για τις ομάδες που δεν προκρίθηκαν. Υπάρχουν επτά έκτροπες τιμές για τις ομάδες που δεν προκρίθηκαν στο Final Four, και τέσσερις για αυτές που προκρίθηκαν, οι Gregor Fucka (2000-2001), Anthony Parker (2004-2005), Andrei Kirilenko (2011-2012) και Nando De Colo (2015-2016).

Ολοκληρώνοντας, δε φαίνεται γενικά κάποια έντονη ασσυμετρία, είτε θετική, είτε αρνητική σε **όλα** τα παραπάνω violin plots.

Ένα σημαντικό συμπέρασμα στο οποίο καταλήγουμε είναι πως παίκτες με αρκετά υψηλές επιδόσεις, τόσο σε πόντους, όσο και σε PIR, αντιστοιχούν σε ομάδες που δε προκρίθηκαν στις επιμέρους φάσεις που εξετάζουμε. Αυτό φαίνεται κοιτάζοντας τα αριστερά violin plots σε κάθε σχήμα, από αυτά που προηγήθηκαν. Συνεπώς, συμπεραίνουμε πως για μια ομάδα να προκριθεί στις επόμενες φάσεις, δεν είναι αρκετός ένας παίκτης με υψηλό δείκτη PIR, αλλά χρειάζονται περισσότεροι παίκτες με μια σχετικά καλή επίδοση.

3.2.4 Μελέτη των πενήντα καλύτερων παικτών σε PIR διαχρονικά

Τέλος, καταγράψαμε τους πενήντα κορυφαίους παίκτες σε PIR διαχρονικά, και μελετήσαμε την κατανομή τους σε ομάδες που προκρίθηκαν ή όχι στις δύο φάσεις.

Για τα προημιτελικά / Playoffs

Ομάδες που ΔΕΝ προκρίθηκαν	Min.	1st. Qu.	Median	Mean	3rd. Qu.	Max.
Points	11,7	16,4	18,5	18,3	20,8	26
PIR	20	21	21,50	22,79	23,60	30,50

Πίνακας 3.14 : Στατιστικά περιγραφικά μέτρα για τους 25 παίκτες των ομάδων που δεν προκρίθηκαν στα playoffs.

Ομάδες που προκρίθηκαν	Min.	1st. Qu.	Median	Mean	3rd. Qu.	Max.
Points	12,10	14,50	16,90	17,22	19,70	22,90
PIR	19,80	20,40	21,10	21,97	22,60	30,90

Πίνακας 3.15 : Στατιστικά περιγραφικά μέτρα για τους 25 παίκτες των ομάδων που προκρίθηκαν στα playoffs.

Για τα ημιτελικά / Final Four

Ομάδες που ΔΕΝ προκρίθηκαν	Min.	1st. Qu.	Median	Mean	3rd. Qu.	Max.
Points	11,70	15,12	18,50	18,28	21,48	26
PIR	19,80	20,48	21,45	22,50	23,12	30,90

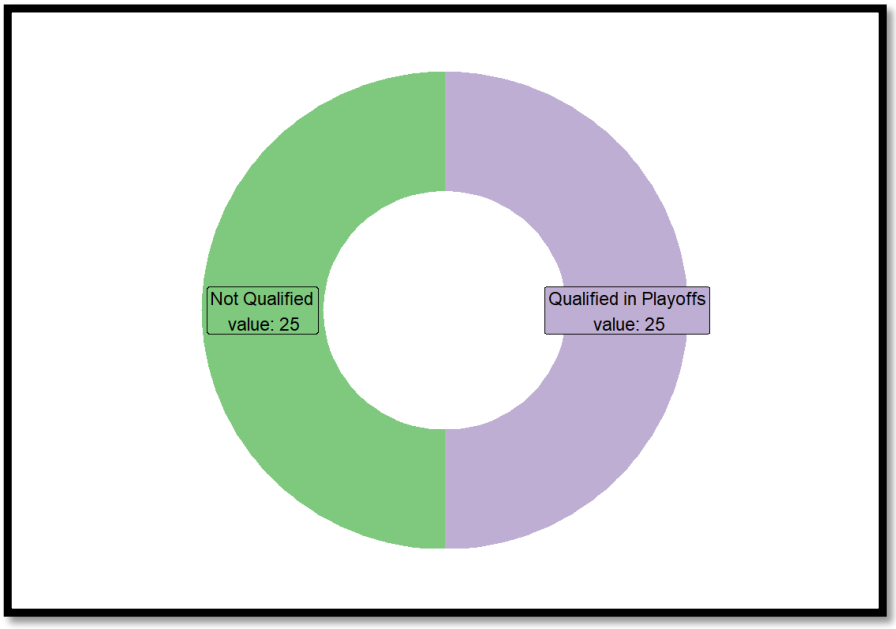
Πίνακας 3.16 : Στατιστικά περιγραφικά μέτρα για τους 38 παίκτες των ομάδων που δεν προκρίθηκαν στο Final Four.

Ομάδες που προκρίθηκαν	Min.	1st. Qu.	Median	Mean	3rd. Qu.	Max.
Points	12,10	14,40	15,40	16,09	18,27	19,70
PIR	19,80	20,65	21,10	22,02	24,23	25,20

Πίνακας 3.17 : Στατιστικά περιγραφικά μέτρα για τους 12 παίκτες των ομάδων που προκρίθηκαν στο Final Four.

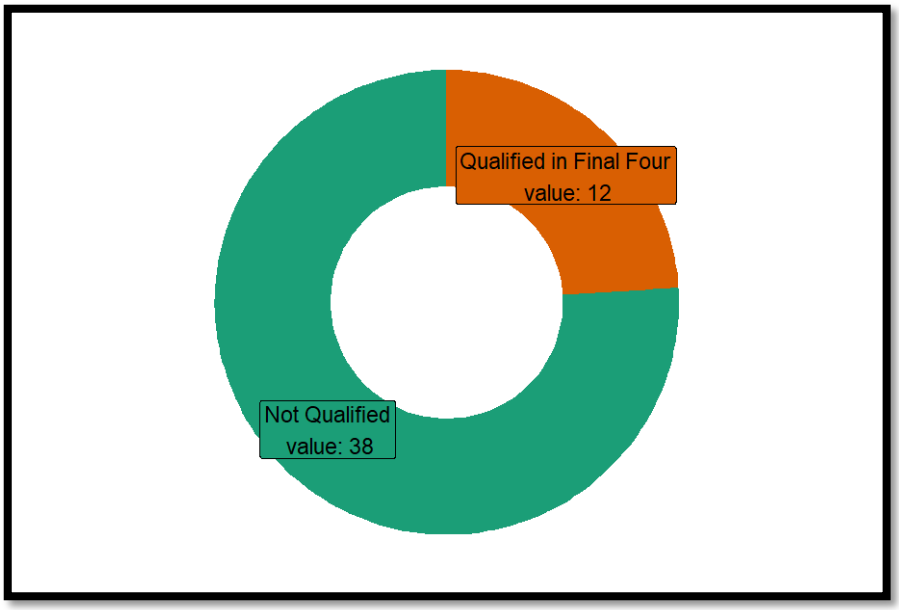
Ακολουθούν τα διαγράμματα donut/doughnut charts, τα οποία αποτελούν μια παραλλαγή των κλασσικών διαγραμμάτων πίτας (pie charts), και απλώς μας δίνουν μια

καθαρότερη οπτική για τις αναλογίες σχετικά με το πώς κατανέμονται οι παίκτες σε ομάδες που προκρίθηκαν (ή όχι) στις φάσεις που εξετάζουμε.



Σχήμα 3.20 : Donut chart για τα playoffs.

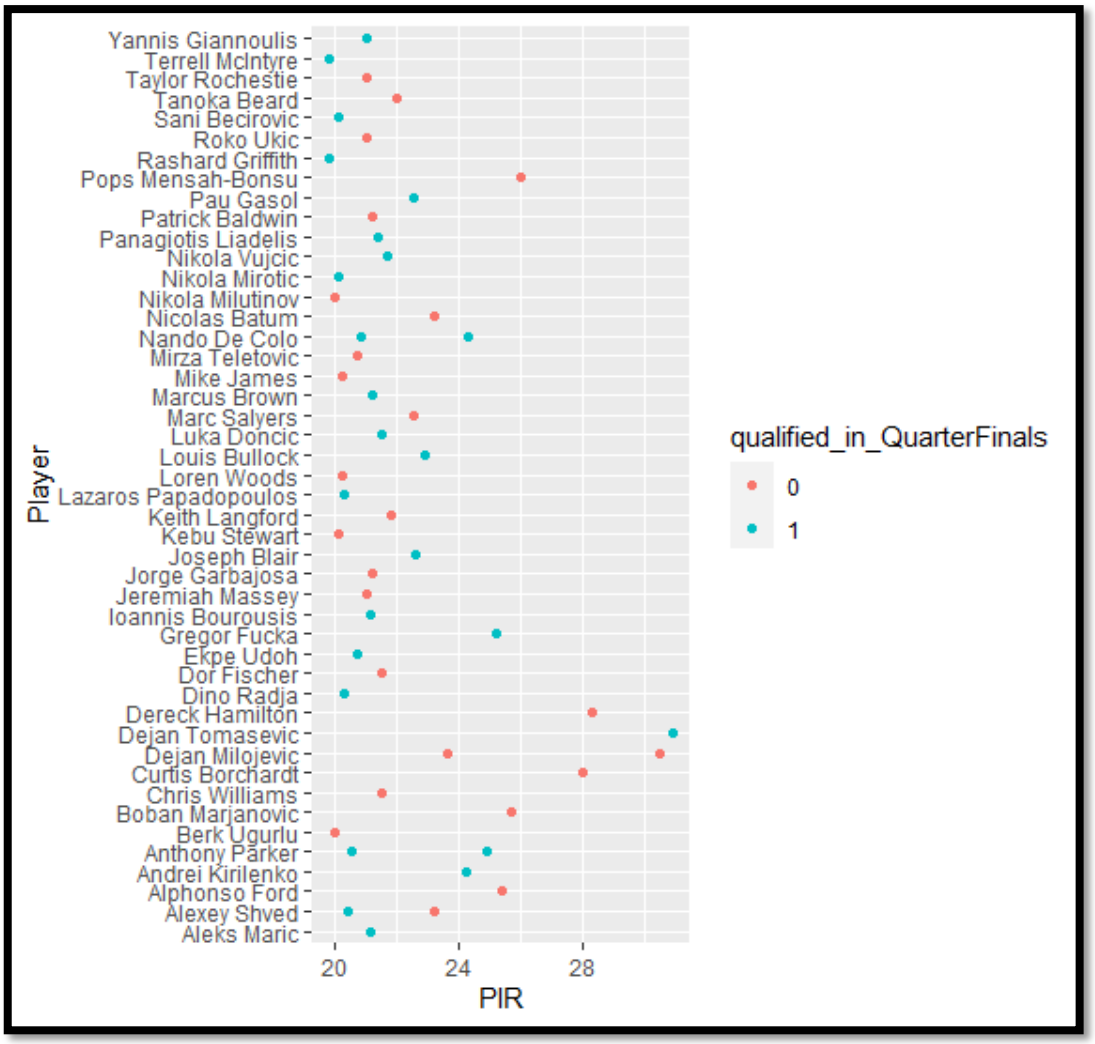
Παρατηρούμε πως στο σύνολο των πενήντα παικτών, οι μισοί (είκοσι πέντε) ανήκουν σε ομάδες που δεν προκρίθηκαν στα playoffs, και οι άλλοι μισοί σε ομάδες που προκρίθηκαν.



Σχήμα 3.21 : Donut chart για το Final Four.

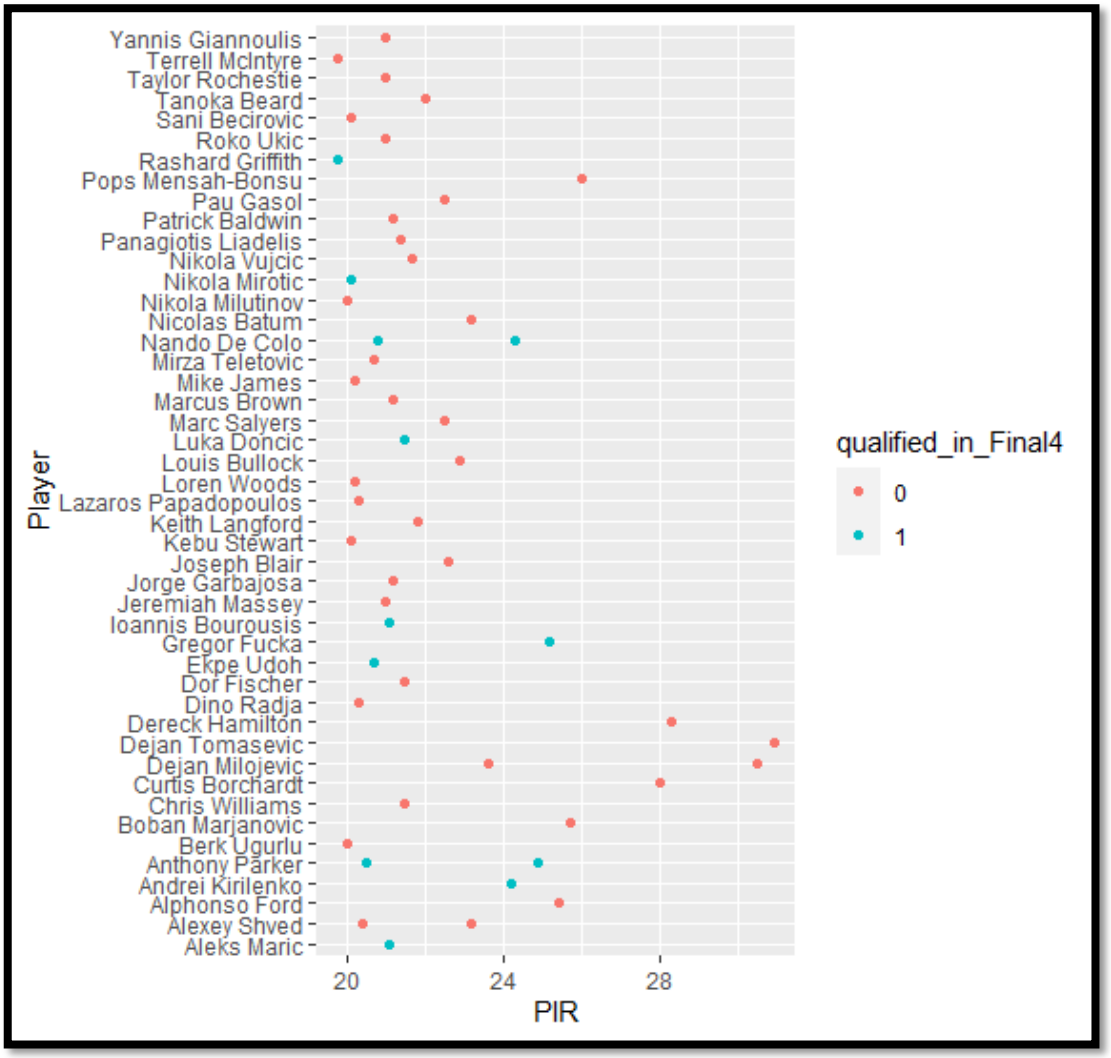
Αντίστοιχα, παρατηρούμε πως στο σύνολο των πενήντα παικτών, οι δώδεκα ανήκουν σε ομάδες που προκρίθηκαν στο Final Four, και οι υπόλοιποι τριάντα οχτώ σε ομάδες που δεν προκρίθηκαν.

Τέλος, τα παρακάτω διαγράμματα μας δίνουν μια γραφική απεικόνιση για τους πενήντα παίκτες, και την αντιστοιχία τους σε ομάδες που προκρίθηκαν σε playoffs και Final Four.



Σχήμα 3.22 : Δείκτης PIR και παίκτες, αναφορικά με τα playoffs.

Παρατηρούμε πως τέσσερις στους πέντε παίκτες διαχρονικά, με το υψηλότερο PIR, αντιστοιχούν σε ομάδες που δεν προκρίθηκαν στα playoffs. Αντίστοιχα, ένας μόνο παίκτης, ο Dejan Tomasevic αντιστοιχεί σε ομάδα που προκρίθηκε (Buducnost, σεζόν 2000-2001). Γενικά, όπως προαναφέρθηκε, 25 παίκτες αντιστοιχούν σε ομάδες που δεν προκρίθηκαν στα playoffs και 25 σε ομάδες που προκρίθηκαν. Υπήρχαν και παίκτες με παραπάνω από μια εμφανίσεις στο διάγραμμα, όπως π.χ. ο Dejan Milojevic.



Σχήμα 3.23 : Δείκτης PIR και παίκτες, αναφορικά με το Final Four.

Παρατηρούμε πως πέντε στους πέντε πρώτους παίκτες διαχρονικά, με το υψηλότερο PIR, αντιστοιχούν σε ομάδες που δεν προκρίθηκαν στο Final Four. Γενικά, 38 παίκτες αντιστοιχούν σε ομάδες που δεν προκρίθηκαν στο Final Four και 12 σε ομάδες που προκρίθηκαν. Υπήρχαν και εδώ παίκτες με παραπάνω από μια εμφανίσεις στο διάγραμμα, όπως π.χ. ο Alexey Shved.

ΚΕΦΑΛΑΙΟ 4ο

Σε αυτό το κεφάλαιο, αρχικά γίνεται έλεγχος κανονικότητας για τις ποσοτικές μεταβλητές μας, για κάθε ανάλυση που πραγματοποιήθηκε ξεχωριστά στο προηγούμενο κεφάλαιο. Στη συνέχεια, υπολογίζονται οι συντελεστές συσχέτισης των μεταβλητών μας, για να δούμε τι σχέση υπάρχει μεταξύ τους (και αν υπάρχει), και τέλος, πραγματοποιούνται και έλεγχοι για την ισότητα των μέσων τιμών των πόντων και του δείκτη PIR, των δύο δειγμάτων (ομάδες που προκρίθηκαν και ομάδες που δεν προκρίθηκαν στις επιμέρους φάσεις), όπου αυτό κρίνεται απαραίτητο.

ΣΧΟΛΙΟ : Για όλους τους ελέγχους, το επίπεδο σημαντικότητας με το οποίο εργαζόμαστε, ισούται με 5%.

4.1 Έλεγχοι κανονικότητας των δεδομένων

Σε αυτή τη παράγραφο, για κάθε ανάλυση που προηγήθηκε, γίνονται έλεγχοι κανονικότητας των ποσοτικών μεταβλητών μας. Στην στατιστική, ο Έλεγχος Κανονικότητας (Test of Normality), είναι σε πολλές περιπτώσεις ο σπουδαιότερος στατιστικός έλεγχος υποθέσεων, και χρησιμοποιείται για να εξεταστεί αν η κατανομή μιας μεταβλητής είναι συμβατή με την Κανονική Κατανομή. Σε επίπεδο συνόλου δεδομένων, ο εν λόγω έλεγχος δύναται να προσδιορίσει εάν το σύνολο δεδομένων είναι καλά μοντελοποιημένο (προέρχεται) από την κανονική κατανομή. Η μεγάλη σημασία αυτού του ελέγχου φαίνεται και από το γεγονός πως με βάση το αποτέλεσμα του ελέγχου, δύναται να αποφασίσουμε αν θα χρησιμοποιηθεί παραμετρικός ή μη παραμετρικός έλεγχος για την εξέταση της εκάστοτε μηδενικής υπόθεσης, ή για την περαιτέρω στατιστική ανάλυση των δεδομένων μας. Η μορφή του συγκεκριμένου στατιστικού ελέγχου υποθέσεων είναι η εξής :

Μηδενική υπόθεση H_0 *: Το δείγμα προέρχεται από κανονική κατανομή.*

VS

Εναλλακτική υπόθεση H_1 *: Το δείγμα δεν προέρχεται από κανονική κατανομή.*

Υπάρχουν διαφορετικοί τρόποι για να ελέγξουμε αν τα δεδομένα μας προέρχονται από κανονική κατανομή. Παρακάτω αναφέρουμε ενδεικτικά κάποιους, αλλά δεν θα τους χρησιμοποιήσουμε όλους.

Έλεγχος Kolmogorov – Smirnov (K – S)

Ο έλεγχος K – S είναι ένας μη παραμετρικός έλεγχος ο οποίος εξετάζει την καλή προσαρμογή ενός τυχαίου δείγματος σε μία δεδομένη κατανομή. Βασίζεται στη διαφορά της εμπειρικής συνάρτησης κατανομής που προέρχεται από το δείγμα, και της αναμενόμενης συνάρτησης κατανομής, υπό την μηδενική υπόθεση της κανονικότητας, ή γενικά της μηδενικής υπόθεσης για μια γνωστή κατανομή (π.χ. Ομοιόμορφη, Poisson, Εκθετική). Ένα σημαντικό μειονέκτημα του είναι πως χρειάζεται να είναι γνωστές οι παράμετροι της κατανομής της συνεχούς τυχαίας μεταβλητής, προκειμένου να είναι αποτελεσματικός. (Μανωλέσου, 2015)

Έλεγχος Shapiro – Wilk

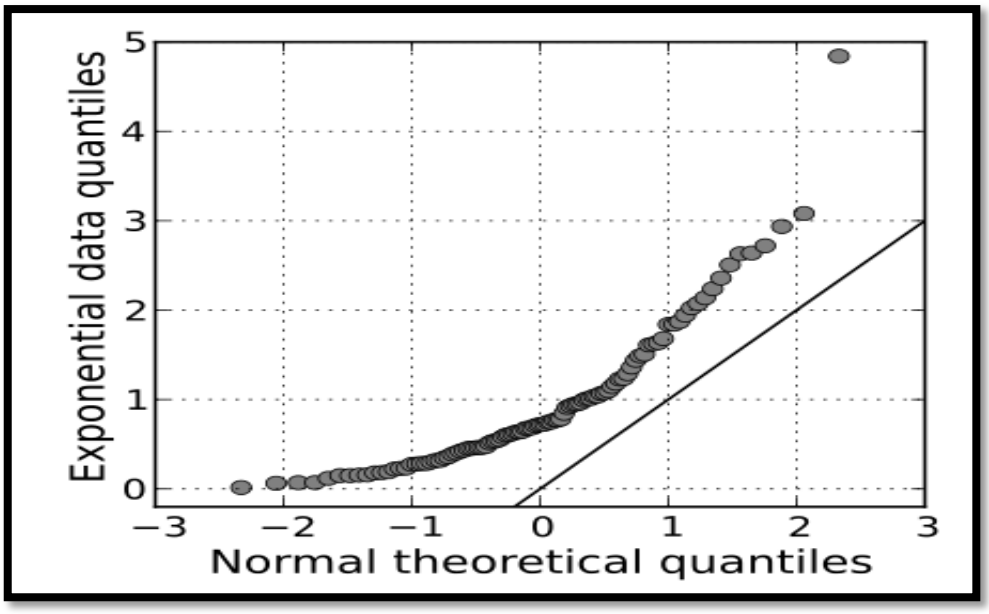
Ο έλεγχος Shapiro – Wilk προτιμάται έναντι του ελέγχου Komogorov – Smirnov για μικρά δείγματα (n<50), έχει γενικά μεγάλη ισχύ, και είναι και αυτός μη παραμετρικός. Η στατιστική συνάρτηση ελέγχου αποτιμά το πόσο κοντά είναι τα εμπειρικά ποσοστιαία σημεία του δείγματος, από τα αντίστοιχα θεωρητικά ποσοστιαία σημεία της αντίστοιχης κανονικής κατανομής. (Αντζουλάκος, 2021)

Έλεγχος Lilliefors

Όπως προαναφέρθηκε, για να εφαρμοστεί ο έλεγχος Kolmogorov – Smirnov, πρέπει να είναι εκ των προτέρων γνωστές οι παράμετροι της κατανομής του πληθυσμού. Σε περίπτωση που αυτές δεν είναι γνωστές, για τον έλεγχο της κανονικότητας του πληθυσμού είναι προτιμότερο να χρησιμοποιήσουμε μια παραλλαγή του ελέγχου K-S, τον έλεγχο Lilliefors. (Αντζουλάκος, 2021)

Q-Q plots

Ένα Q-Q plot είναι ένα γράφημα πιθανότητας για τη γραφική σύγκριση δύο κατανομών πιθανότητας, απεικονίζοντας τα ποσοστημόρια της μιας, σε σχέση με την άλλη. Το γράφημα Q-Q χρησιμοποιείται για να συγκρίνουμε τα σχήματα των συναρτήσεων κατανομής, παρέχοντας μια γραφική άποψη για τον τρόπο με τον οποίο ιδιότητες, όπως η θέση, η κλίμακα και η ασυμμετρία, είναι παρόμοιες ή διαφορετικές στις δύο κατανομές. Επίσης, χρησιμοποιείται για να συγκρίνει τις συλλογές δεδομένων ή θεωρητικές κατανομές. Η χρήση του γραφήματος για τη σύγκριση δύο δειγμάτων δεδομένων, μπορεί να θεωρηθεί ως μια μη-παραμετρική προσέγγιση για τη σύγκριση των κατανομών τους. (Μανωλέσου, 2015)



Σχήμα 4.1 : Παράδειγμα ενός Q-Q plot.

(Πηγή : https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot)

Εμείς κάνουμε χρήση του ελέγχου Lilliefors για τον έλεγχο κανονικότητας των δεδομένων μας, από τη βιβλιοθήκη "nortest" της R, σε επίπεδο σημαντικότητας όπως προαναφέρθηκε, ίσο με 5%.

4.1.1 Έλεγχος κανονικότητας για τους μέσους όρους των γενικών χαρακτηριστικών των ομάδων, για όλες τις σεζόν

Αναφορικά με τους μέσους όρους των μεταβλητών των πόντων, των ποσοστών εύστοχων δίποντων, ποσοστών των ελεύθερων βολών, των επιθετικών ριμπάουντ, των αμυντικών ριμπάουντ, των συνολικών ριμπάουντ, των ασίστ, των λαθών, των μπλοκ υπέρ της ομάδας, των μπλοκ κατά της ομάδας, των φάουλ στα οποία υπέπεσε η ομάδα, των φάουλ που δέχτηκε η ομάδα, του δείκτη PIR, καθώς και για τη μεταβλητή για τις σεζόν, συμπεραίνουμε πως δεν μπορούμε να απορρίψουμε την υπόθεση της κανονικότητας σε επίπεδο σημαντικότητας 5%. Αντίστοιχα, για τους μέσους όρους των ποσοστών των εύστοχων τριπόντων και των κλεψιμάτων, απορρίπτουμε την υπόθεση της κανονικότητας.

Ακολουθεί ο πίνακας με τις μεταβλητές και τις αντίστοιχες p-values του ελέγχου Lilliefors :

Μεταβλητές	p-values
Points	0,4579
2PT %	0,7104
3PT %	0,01749
FT %	0,09257
OR	0,3829
DR	0,7398
TR	0,09662
AST	0,09243
STL	0,0009687
TO	0,7691
BLK	0,478
BLKA	0,156
FC	0,3103
FD	0,2503
PIR	0,2301
Season	0,989

Πίνακας 4.1 : Αποτελέσματα ελέγχων κανονικότητας.

4.1.2 Έλεγχος κανονικότητας στην ανάλυση των πέντε καλύτερων παικτών σε PIR

Αρχικά, θα εξετάσουμε τη κανονικότητα των πόντων και του δείκτη PIR, αναφορικά με τη φάση των playoffs. Αναφορικά με τους παίκτες των ομάδων που δεν προκρίθηκαν στα playoffs, η υπόθεση της κανονικότητας για τους πόντους απορρίπτεται, ενώ για το δείκτη PIR, δε μπορούμε να απορρίψουμε την μηδενική μας υπόθεση. Αντίστοιχα, για τις ομάδες που προκρίθηκαν στα playoffs, δε μπορούμε να απορρίψουμε την υπόθεση τόσο για τους πόντους, όσο και για το δείκτη PIR.

Στη συνέχεια, εξετάζουμε τη κανονικότητα των πόντων και του δείκτη PIR, αναφορικά με το Final Four. Για τους παίκτες των ομάδων που δεν προκρίθηκαν στο Final Four, οι πόντοι τους δεν προέρχονται από κανονική κατανομή, ενώ ο δείκτης PIR τους προέρχεται. Αντίστοιχα, για τις ομάδες που προκρίθηκαν στο Final Four, τόσο για τους πόντους, όσο και για το δείκτη PIR, δε μπορούμε να απορρίψουμε την υπόθεση της κανονικότητας των μεταβλητών.

Ακολουθούν οι πίνακες με τις μεταβλητές και τις αντίστοιχες p-values του ελέγχου Lilliefors, για κάθε φάση :

Ομάδες που ΔΕΝ προκρίθηκαν στα playoffs	p - values
Points	0,0009344
PIR	0,2359

Πίνακας 4.2 : Αποτελέσματα ελέγχων κανονικότητας για ομάδες που δε προκρίθηκαν στα playoffs.

Ομάδες που προκρίθηκαν στα playoffs	p - values
Points	0,1979
PIR	0,7054

Πίνακας 4.3 : Αποτελέσματα ελέγχων κανονικότητας για ομάδες που προκρίθηκαν στα playoffs.

Ομάδες που ΔΕΝ προκρίθηκαν στο Final Four	p - values
Points	0,0003327
PIR	0,2004

Πίνακας 4.4 : Αποτελέσματα ελέγχων κανονικότητας για ομάδες που δε προκρίθηκαν στο Final Four.

Ομάδες που προκρίθηκαν στο Final Four	p - values
Points	0,425
PIR	0,548

Πίνακας 4.5 : Αποτελέσματα ελέγχων κανονικότητας για ομάδες που προκρίθηκαν στο Final Four.

4.1.3 Έλεγχος κανονικότητας στην ανάλυση των δέκα καλύτερων παικτών σε πόντους, και των δέκα καλύτερων σε PIR, ανά σεζόν

- Ο έλεγχος για τους δέκα καλύτερους παίκτες σε πόντους, ανά σεζόν.

Για τις ομάδες των παικτών, που δεν προκρίθηκαν στα playoffs, τόσο για τους πόντους, όσο και για το δείκτη PIR, απορρίπτεται η υπόθεση της κανονικότητας. Αντίστοιχα, αναφορικά με τις ομάδες που προκρίθηκαν στα playoffs, για τους πόντους απορρίπτεται η υπόθεση της κανονικότητας, ενώ δεν μπορούμε να την απορρίψουμε για το δείκτη PIR.

Για τις ομάδες που δεν προκρίθηκαν στο Final Four, τόσο για τους πόντους, όσο και για το δείκτη PIR, απορρίπτεται η υπόθεση της κανονικότητας. Αντίστοιχα, για τις

ομάδες που προκρίθηκαν στο Final Four, τόσο για τους πόντους, όσο και για το δείκτη PIR, δεν μπορούμε να απορρίψουμε την υπόθεση της κανονικότητας.

Ακολουθούν οι πίνακες με τις μεταβλητές και τις αντίστοιχες p-values του ελέγχου Lilliefors, για κάθε φάση :

Ομάδες που ΔΕΝ προκρίθηκαν στα playoffs	p - values
Points	3,42e-06
PIR	0,0002168

Πίνακας 4.6 : Αποτελέσματα ελέγχων κανονικότητας για ομάδες που δε προκρίθηκαν στα playoffs.

Ομάδες που προκρίθηκαν στα playoffs	p - values
Points	0,004382
PIR	0,1668

Πίνακας 4.7 : Αποτελέσματα ελέγχων κανονικότητας για ομάδες που προκρίθηκαν στα playoffs.

Ομάδες που ΔΕΝ προκρίθηκαν στο Final Four	p - values
Points	5,665e-08
PIR	0,000375

Πίνακας 4.8 : Αποτελέσματα ελέγχων κανονικότητας για ομάδες που δε προκρίθηκαν στο Final Four.

Ομάδες που προκρίθηκαν στο Final Four	p - values
Points	0,1731
PIR	0,34

Πίνακας 4.9 : Αποτελέσματα ελέγχων κανονικότητας για ομάδες που προκρίθηκαν στο Final Four.

- Ο έλεγχος για τους δέκα καλύτερους παίκτες σε PIR, ανά σεζόν.

Για τις ομάδες των παικτών, που δεν προκρίθηκαν στα playoffs, τόσο για τους πόντους, όσο και για το δείκτη PIR, απορρίπτεται η υπόθεση της κανονικότητας. Αντίστοιχα, αναφορικά με τις ομάδες που προκρίθηκαν στα playoffs, η υπόθεση της κανονικότητας απορρίπτεται τόσο για τους πόντους, όσο και για το δείκτη PIR.

Για τις ομάδες που δεν προκρίθηκαν στο Final Four, τόσο για τους πόντους, όσο και για το δείκτη PIR, απορρίπτεται η υπόθεση της κανονικότητας. Αντίστοιχα, για τις

ομάδες που προκρίθηκαν στο Final Four, η υπόθεση της κανονικότητας δεν μπορεί να απορριφθεί για τους πόντους, ενώ απορρίπτεται για το δείκτη PIR.

Ακολουθούν οι πίνακες με τις μεταβλητές και τις αντίστοιχες p-values του ελέγχου Lilliefors, για κάθε φάση :

Ομάδες που ΔΕΝ προκρίθηκαν στα playoffs	p - values
Points	0,02018
PIR	0,0001324

Πίνακας 4.10 : Αποτελέσματα ελέγχων κανονικότητας για ομάδες που δε προκρίθηκαν στα playoffs.

Ομάδες που προκρίθηκαν στα playoffs	p - values
Points	0,001506
PIR	2,655e-08

Πίνακας 4.11 : Αποτελέσματα ελέγχων κανονικότητας για ομάδες που προκρίθηκαν στα playoffs.

Ομάδες που ΔΕΝ προκρίθηκαν στο Final Four	p - values
Points	0,002388
PIR	1,537e-08

Πίνακας 4.12 : Αποτελέσματα ελέγχων κανονικότητας για ομάδες που δε προκρίθηκαν στο Final Four.

Ομάδες που προκρίθηκαν στο Final Four	p - values
Points	0,1355
PIR	0,0008219

Πίνακας 4.13 : Αποτελέσματα ελέγχων κανονικότητας για ομάδες που προκρίθηκαν στο Final Four.

4.1.4 Έλεγχος κανονικότητας στην ανάλυση των πενήντα καλύτερων παικτών σε PIR διαχρονικά

Για τις ομάδες των παικτών, που δεν προκρίθηκαν στα playoffs, η υπόθεση της κανονικότητας δεν μπορεί να απορριφθεί για τους πόντους, ενώ για το δείκτη PIR τους απορρίπτεται. Αντίστοιχα, για τις ομάδες που προκρίθηκαν στα playoffs, τόσο για τους πόντους, όσο και για το δείκτη PIR, απορρίπτεται η υπόθεση.

Για τις ομάδες που δεν προκρίθηκαν στο Final Four, η υπόθεση της κανονικότητας για τους πόντους δεν μπορεί να απορριφθεί, ενώ για το δείκτη PIR απορρίπτεται. Τέλος, αναφορικά με τις ομάδες που προκρίθηκαν στο Final Four, δε μπορούμε να απορρίψουμε την υπόθεση της κανονικής κατανομής για τους πόντους, ενώ απορρίπτεται για το δείκτη PIR.

Ακολουθούν οι πίνακες με τις μεταβλητές και τις αντίστοιχες p-values του ελέγχου Lilliefors, για κάθε φάση :

Ομάδες που ΔΕΝ προκρίθηκαν στα playoffs	p - values
Points	0,8983
PIR	0,007113

Πίνακας 4.14 : Αποτελέσματα ελέγχων κανονικότητας για ομάδες που δε προκρίθηκαν στα playoffs.

Ομάδες που προκρίθηκαν στα playoffs	p - values
Points	0,01174
PIR	0,002214

Πίνακας 4.15 : Αποτελέσματα ελέγχων κανονικότητας για ομάδες που προκρίθηκαν στα playoffs.

Ομάδες που ΔΕΝ προκρίθηκαν στο Final Four	p - values
Points	0,5656
PIR	0,0005096

Πίνακας 4.16 : Αποτελέσματα ελέγχων κανονικότητας για ομάδες που δε προκρίθηκαν στο Final Four.

Ομάδες που προκρίθηκαν στο Final Four	p - values
Points	0,2185
PIR	0,01723

Πίνακας 4.17 : Αποτελέσματα ελέγχων κανονικότητας για ομάδες που προκρίθηκαν στο Final Four.

4.2 Συντελεστές συσχέτισης των μεταβλητών

Σε αυτή τη παράγραφο, για κάθε ανάλυση που προηγήθηκε, θα εξεταστούν οι συσχετίσεις ανάμεσα στις μεταβλητές μας. Εξετάζοντας τις συσχετίσεις, συμπεραίνουμε για τον βαθμό της αλληλεξάρτησης ανάμεσα σε δύο, ή περισσότερες μεταβλητές. Στην στατιστική, η εξάρτηση είναι οποιαδήποτε στατιστική σχέση μεταξύ δύο τυχαίων μεταβλητών ή δύο συνόλων δεδομένων. Η συσχέτιση αναφέρεται σε μια ευρεία κατηγορία στατιστικών σχέσεων με τη συμμετοχή της εξάρτησης, αν και σε κοινή χρήση συχνότερα αναφέρεται στο βαθμό με τον οποίο δύο μεταβλητές έχουν μια γραμμική σχέση η μία με την άλλη. Οι συσχετίσεις είναι χρήσιμες, διότι μπορεί να υποδείξουν μια προγνωστική σχέση που μπορεί να αξιοποιηθεί στην πράξη.

Θα αναφέρουμε τρία μέτρα συσχέτισης, το συντελεστή γραμμικής συσχέτισης του Pearson, το συντελεστή γραμμικής συσχέτισης του Spearman, και το δείκτη Kendall.

Συντελεστής συσχέτισης του Pearson (r)

Το πιο γνωστό μέτρο της εξάρτησης μεταξύ δύο ποσοτήτων είναι ο συντελεστής συσχέτισης Pearson, που συνήθως ονομάζεται απλά «συντελεστής συσχέτισης». Είναι το πηλίκο της διαίρεσης της συνδιακύμανσης των δύο μεταβλητών, με το γινόμενο των τυπικών αποκλίσεων. Χρησιμοποιείται για τον έλεγχο της σχέσης μεταξύ δύο ποσοτικών μεταβλητών, όταν ακολουθούν και οι δύο κανονική κατανομή. (Καλλιακμάνης, 2020)

Ο γνωστός συντελεστής συσχέτισης $\rho_{X,Y}$ μεταξύ δύο τυχαίων μεταβλητών X και Y με τις αναμενόμενες τιμές μ_X και μ_Y και τυπική απόκλιση σ_X και σ_Y ορίζεται ως:

$$\rho_{X,Y} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{E[(X-\mu_X)\cdot(Y-\mu_Y)]}{\sigma_X \cdot \sigma_Y} \ ,$$

όπου E είναι η αναμενόμενη τιμή, cov είναι η συνδιακύμανση, και corr είναι μια ευρέως χρησιμοποιούμενη εναλλακτική συντομογραφία για το συντελεστή συσχέτισης.

Ο συντελεστής συσχέτισης Pearson ορίζεται μόνο αν και οι δύο τυπικές αποκλίσεις είναι πεπερασμένες και μη μηδενικές. Είναι απόρροια της ανισότητας Cauchy - Schwarz πως η συσχέτιση δεν μπορεί να υπερβαίνει το 1, κατά απόλυτη τιμή. Ο συντελεστής συσχέτισης είναι συμμετρικός, δηλαδή ισχύει corr(X,Y) = corr(Y,X).

Ο συντελεστής συσχέτισης Pearson παίρνει τιμές +1 σε περίπτωση μίας τέλειας (αύξουσας) γραμμικής σχέσης (θετική συσχέτιση), -1 σε περίπτωση μίας τέλειας φθίνουσας (αντίστροφης) γραμμικής σχέσης (αρνητική συσχέτιση), και κάποια τιμή μεταξύ -1 και +1 σε όλες τις άλλες περιπτώσεις, που δείχνει το βαθμό της γραμμικής

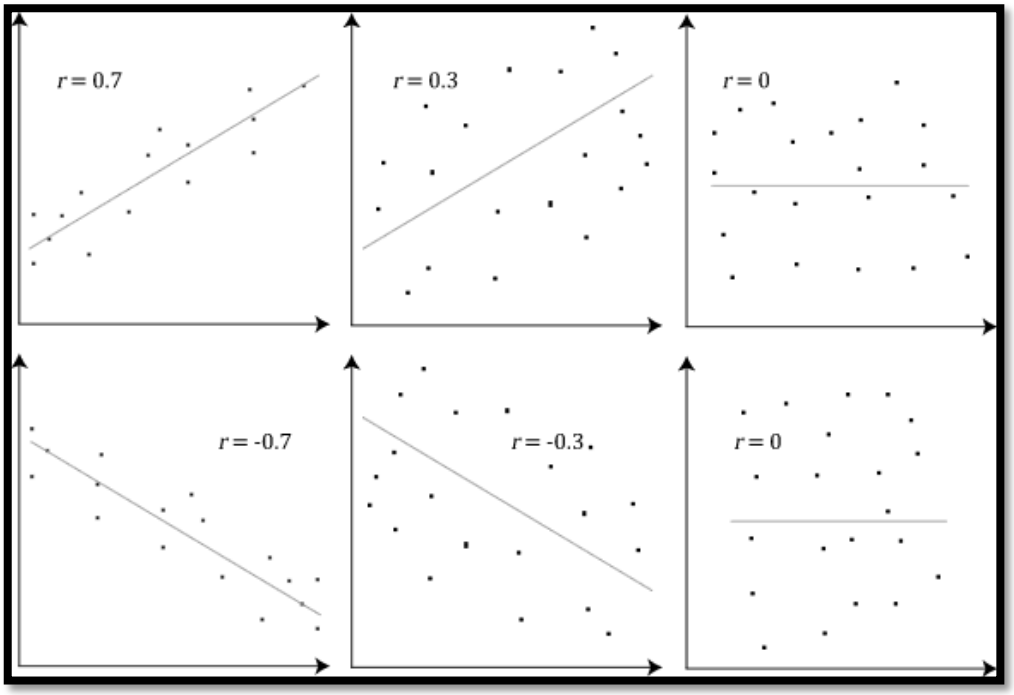
εξάρτησης μεταξύ των μεταβλητών. Καθώς πλησιάζει το μηδέν, υπάρχει λιγότερη σχέση (πιο κοντά σε ασυσχέτιστες μεταβλητές). Όσο πιο κοντά είναι ο συντελεστής είτε στο -1 ή στο +1, τόσο ισχυρότερη είναι η συσχέτιση μεταξύ των μεταβλητών. Αν οι μεταβλητές είναι ανεξάρτητες, ο συντελεστής συσχέτισης Pearson είναι 0, αλλά το αντίστροφο δεν είναι αληθές, διότι ο συντελεστής συσχέτισης ανιχνεύει μόνο γραμμική εξάρτηση μεταξύ των δύο μεταβλητών.

Αν έχουμε μια σειρά από n μετρήσεις των X και Y γραμμένες ως x_i και y_i για i = 1, 2, ..., n, τότε ο δειγματικός συντελεστής συσχέτισης μπορεί να χρησιμοποιηθεί για την εκτίμηση του πληθυσμιακού συντελεστή συσχέτισης Pearson r μεταξύ X και Y. Ο δειγματικός συντελεστής συσχέτισης γράφεται :

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Γενικότερα, τόσο για τον απλό συντελεστή συσχέτισης, όσο και για τον δειγματικό, έχουμε ότι οι τιμές τους είναι μεταξύ του -1 και του +1.

Παρακάτω παρουσιάζεται μια εικόνα για την διαγραμματική απεικόνιση του συντελεστή συσχέτισης r, όταν υπάρχει θετική, αρνητική ή και καμία συσχέτιση μεταξύ των δεδομένων :



Σχήμα 4.2 : Διαγράμματα διασποράς για παρατήρηση συσχετίσεων.

(Πηγή : https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php)

Συντελεστής συσχέτισης του Spearman (ρ)

Ο συγκεκριμένος συντελεστής, που συμβολίζεται επίσης με το γράμμα ρ ή r_s, χρησιμοποιείται σε ένα μη-παραμετρικό πλαίσιο που μας δίνει τη συσχέτιση ανάμεσα στην τάξη (rank) που παίρνουν οι τιμές των δύο μεταβλητών. Ουσιαστικά, δείχνει κατά πόσο μπορεί να περιγραφεί η σχέση που έχουν δύο μεταβλητές μέσω μίας μονότονης συνάρτησης. Χρησιμοποιείται για να ελέγξει τη σχέση που έχουν δύο μεταβλητές, όταν η μια είναι συνεχής, με κανονική κατανομή και η άλλη δεν ακολουθεί την κανονική κατανομή ή είναι κατηγορική, όπως και στη περίπτωση όπου και οι δύο δεν ακολουθούν κανονική κατανομή. (Σπυριδάκης, 2022)

Ο συγκεκριμένος συντελεστής είναι μια εξειδίκευση του συντελεστή συσχέτισης του Pearson στην κατάταξη των τιμών των παρατηρήσεων των δύο μεταβλητών, οπότε για το σύνολο του πληθυσμού για δύο μεταβλητές X και Y, υπολογίζεται ως εξής :

$$r_s = \rho_{rg_X, rg_Y} = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X}\, \sigma_{rg_Y}},$$

όπου rg_X, rg_Y είναι οι μεταβλητές που αφορούν την κατάταξη των παρατηρήσεων των μεταβλητών X και Y.

Αναφορικά με τον υπολογισμό του δειγματικού συντελεστή, προκύπτει ότι ξεχωρίζουν δύο περιπτώσεις, ανάλογα με τα αποτελέσματα που έχουν οι κατατάξεις των μεταβλητών. Στην περίπτωση που η κατάταξη δεν δίνει περιπτώσεις ισοβαθμίας, ο συντελεστής συσχέτισης υπολογίζεται ως εξής :

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)},$$

όπου d_i είναι η διαφορά μεταξύ των δύο βαθμίδων κάθε παρατήρησης και n είναι ο αριθμός των παρατηρήσεων.

Αντίστοιχα, στην περίπτωση που υπάρχει τουλάχιστον μία περίπτωση ισοβαθμίας στην κατάταξη των παρατηρήσεων μίας εκ των δύο μεταβλητών, ο δειγματικός συντελεστής συσχέτισης υπολογίζεται με τον ίδιο ακριβώς τρόπο που υπολογίστηκε ο δειγματικός συντελεστής συσχέτισης του Pearson :

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Για τον συγκεκριμένο συντελεστή ισχύει, όπως και για τον συντελεστή συσχέτισης του Pearson, ότι λαμβάνει τιμές μεταξύ του -1 και του +1. Για τιμή κοντά στο +1, ισχύει ότι οι μεταβλητές σχετίζονται άριστα μέσω μίας μονότονης συνάρτησης, και πως μια αύξηση της τιμής του Y σημαίνει αύξηση της τιμής του X. Αντίστοιχα, για τιμή κοντά στο -1, ισχύει ότι οι μεταβλητές σχετίζονται άριστα μέσω μίας μονότονης συνάρτησης, και πως μία αύξηση της τιμής του Y σημαίνει μείωση της τιμής του X.

Συντελεστής συσχέτισης του Kendall (τ)

Ο συντελεστής συσχέτισης του Kendall, γνωστός και ως συντελεστής συμφωνίας του Kendall, μοιάζει με τον συντελεστή ρ του Spearman, ως προς το ότι υπολογίζεται με βάση την τάξη (rank) μεγέθους των παρατηρήσεων και όχι με βάση τις παρατηρήσεις αυτές καθεαυτές. Χρησιμοποιείται για τον έλεγχο της σχέσης μεταξύ δύο μεταβλητών, όταν αυτές είναι κατηγορικές, διατάξιμες. Επιπλέον, η κατανομή του δεν εξαρτάται από την κατανομή των μεταβλητών X και Y, όταν αυτές είναι ανεξάρτητες και συνεχείς. Το κύριο πλεονέκτημα του μέτρου αυτού, σε σχέση με το μέτρο ρ του Spearman, είναι ότι τείνει πιο γρήγορα στην κανονική κατανομή. (Croux & Dehon, 2010)

Ένα άλλο πλεονέκτημα βρίσκεται στο γεγονός ότι μπορεί άμεσα να ερμηνευθεί μέσω των πιθανοτήτων με τις οποίες παρατηρούμε εναρμονισμένα (concordant) ζεύγη τιμών και μη εναρμονισμένα (discordant) ζεύγη τιμών. Δύο παρατηρήσεις ονομάζονται εναρμονισμένες, αν και τα δύο μέλη της μίας παρατήρησης είναι μεγαλύτερα (ή μικρότερα) από τα αντίστοιχα μέλη της άλλης παρατήρησης. Αν η διάταξη των πρώτων μελών τους είναι αντίθετη από την διάταξη των δεύτερων μελών τους, οι παρατηρήσεις ονομάζονται μη εναρμονισμένες.

Τα ζεύγη των παρατηρήσεων (X_i , Y_j) και (X_k , Y_k) για τα οποία ισχύει $X_j = X_k$ ή/και $Y_j = Y_k$, δεν είναι ούτε εναρμονισμένα ούτε μη εναρμονισμένα και ονομάζονται ισοβαθμούντα (tied). Συνεπώς, ο συντελεστής συσχέτισης του Kendall ανάμεσα σε δύο τυχαίες μεταβλητές με n παρατηρήσεις, ορίζεται ως εξής :

$$\tau = \frac{(\#\, concordant) - (\#\, discordant\ pairs)}{\frac{n(n-1)}{2}}$$

4.2.1 Έλεγχος συσχετίσεων για τους μέσους όρους των γενικών χαρακτηριστικών των ομάδων, για όλες τις σεζόν

Αρχικά, θα εξετάσουμε τη συσχέτιση ανάμεσα στις μεταβλητές που προέρχονται από κανονική κατανομή και είναι ποσοτικές, χρησιμοποιώντας το συντελεστή Pearson.

Παρακάτω, παρουσιάζεται ο πίνακας των συσχετίσεων για τις κανονικές, ποσοτικές μεταβλητές μας :

	Points	2PT %	FT %	OR	DR	TR	AST	TO	BLK	BLKA	FC	FD	PIR
Points	1	0,73	0,37	0,14	-0,11	-0,04	0,28	-0,32	-0,37	-0,57	0,16	0,11	0,77
2PT %	0,73	1	0,61	-0,17	-0,22	-0,22	0,33	-0,3	-0,53	-0,48	-0,21	-0,2	0,7
FT %	0,37	0,61	1	0,25	0,41	0,38	0,78	-0,52	0,07	0,11	-0,72	-0,7	0,7
OR	0,14	-0,17	0,25	1	0,64	0,82	0,53	-0,41	0,38	0,39	-0,3	-0,37	0,32
DR	-0,11	-0,22	0,41	0,64	1	0,96	0,7	-0,44	0,53	0,56	-0,58	-0,62	0,31
TR	-0,04	-0,22	0,38	0,82	0,96	1	0,7	-0,47	0,53	0,55	-0,53	-0,59	0,34
AST	0,28	0,33	0,78	0,53	0,7	0,7	1	-0,76	0,36	0,37	-0,75	-0,71	0,76
TO	-0,32	-0,3	-0,52	-0,41	-0,44	-0,47	-0,76	1	-0,17	-0,12	0,53	0,42	-0,7
BLK	-0,37	-0,53	0,07	0,38	0,53	0,53	0,36	-0,17	1	0,84	-0,36	-0,34	-0,13
BLKA	-0,57	-0,48	0,11	0,39	0,56	0,55	0,37	-0,12	0,84	1	-0,52	-0,5	-0,19
FC	0,16	-0,21	-0,72	-0,3	-0,58	-0,53	-0,75	0,53	-0,36	-0,52	1	0,95	-0,39
FD	0,11	-0,2	-0,7	-0,37	-0,62	-0,59	-0,71	0,42	-0,34	-0,5	0,95	1	-0,34
PIR	0,77	0,7	0,7	0,32	0,31	0,34	0,76	-0,7	-0,13	-0,19	-0,39	-0,34	1

Πίνακας 4.18 : Πίνακας συσχετίσεων για τις κανονικές, ποσοτικές μεταβλητές.

Προκειμένου να σχολιαστούν τα αποτελέσματα, πρέπει να οριστούν τα όρια των τιμών της συσχέτισης, σύμφωνα με τα οποία συμπεραίνουμε αν δύο μεταβλητές έχουν τέλεια, ισχυρή, μέτρια ή ασθενής συσχέτιση. Οι τιμές αυτές δεν αλλάζουν σε περίπτωση θετικής ή αρνητικής συσχέτισης, και βασίζονται στην τιμή του συντελεστή. (Evans, 1996)

- Τέλεια Συσχέτιση: Εάν η τιμή του συντελεστή είναι πολύ κοντά στο ± 1.
- Πολύ Υψηλή Συσχέτιση: Εάν η τιμή του συντελεστή κυμαίνεται μεταξύ ± 0,80 και ± 1 .
- Υψηλή Συσχέτιση: Εάν η τιμή κυμαίνεται μεταξύ ± 0,60 και ± 0,79 .
- Μέτρια Συσχέτιση: Εάν η τιμή κυμαίνεται μεταξύ ± 0,40 και ± 0,59 .
- Ασθενής Συσχέτιση: Όταν η τιμή κυμαίνεται μεταξύ ± 0,20 και ± 0,39 .

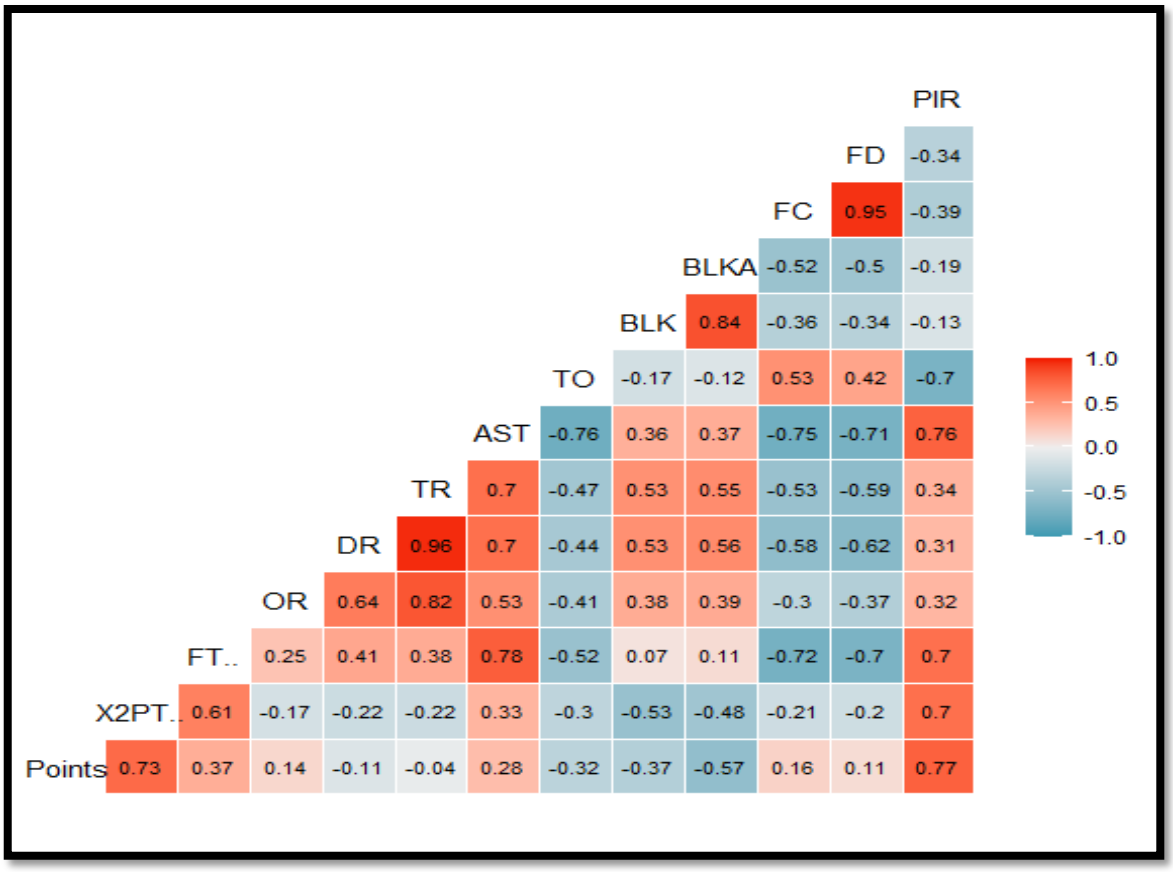
- Πολύ Ασθενής Συσχέτιση: Εάν η τιμή του συντελεστή είναι κάτω από ± 0,19 .

- Χωρίς Συσχέτιση: Όταν η τιμή είναι μηδέν.

Αναλυτικά, θα σχολιαστούν οι συσχετίσεις του δείκτη PIR με τα υπόλοιπα στατιστικά στοιχεία. Ο δείκτης PIR έχει υψηλή, θετική συσχέτιση με τους πόντους, το ποσοστό δίποντων, το ποσοστό των ελευθέρων βολών και με τις ασίστ. Έχει ασθενή, θετική συσχέτιση με τα επιθετικά ριμπάουντ, τα αμυντικά ριμπάουντ και τα συνολικά ριμπάουντ. Έχει υψηλή, αρνητική συσχέτιση με τα λάθη. Παράλληλα, ο δείκτης έχει πολύ ασθενή, αρνητική συσχέτιση με τα μπλοκ των ομάδων και τα μπλοκ κατά των ομάδων, ενώ έχει υψηλή, αρνητική συσχέτιση με τα λάθη στα οποία υπέπεσαν οι ομάδες. Τέλος, έχει ασθενή, αρνητική συσχέτιση τόσο με τα φάουλ που κέρδισαν οι ομάδες, όσο και με αυτά που έκαναν.

Παρακάτω, παρουσιάζεται το άνω τριγωνικό μέρος του κορελογράμματος και το heatmap για τις συσχετίσεις, ώστε να έχουμε μια καλύτερη οπτικοποίηση των σχέσεων μεταξύ των μεταβλητών, αναλόγως των χρωμάτων.

Στο σχήμα που ακολουθεί, τα μπλε τετράγωνα αντιστοιχούν σε αρνητικές συσχετίσεις και τα κόκκινα σε θετικές. Όσο πιο έντονο το χρώμα, τόσο μεγαλύτερη κατ' απόλυτη τιμή η συσχέτιση.

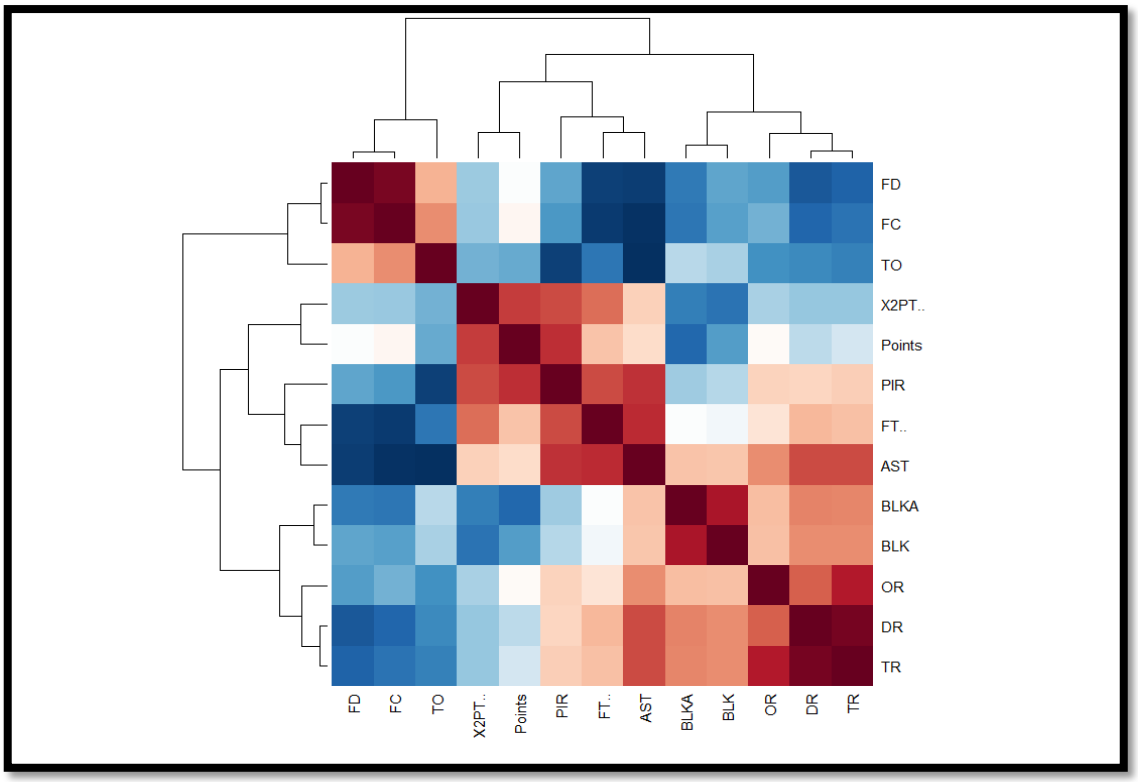


Σχήμα 4.3 : Κορελόγραμμα για τις κανονικές, ποσοτικές μεταβλητές.

Αξίζει επίσης να σημειωθεί η υψηλή, θετική συσχέτιση των πόντων με το ποσοστό των εύστοχων δίποντων, κάτι που είναι αναμενόμενο, όπως και η μέτρια σχέση των πρώτων με τα μπλοκ κατά των ομάδων, κάτι που επίσης είναι λογικό. Εξίσου λογικό είναι να υπάρχει πολύ υψηλή, θετική σχέση ανάμεσα στις δύο κατηγορίες των φάουλ, ανάμεσα στις δύο κατηγορίες των μπλοκ, ανάμεσα στα αμυντικά ριμπάουντ με τα συνολικά ριμπάουντ, καθώς και στα επιθετικά ριμπάουντ με τα συνολικά.

Οι ασίστ έχουν υψηλή, θετική συσχέτιση με τα αμυντικά ριμπάουντ, κάτι που ίσως να μεταφράζεται πως με το «κατέβασμα» ενός ριμπάουντ, οι ομάδες συχνά επιτίθενται άμεσα και σκοράρουν, μοιράζοντας ασίστ. Αντιθέτως, προκαλεί εντύπωση η αρνητική και υψηλή σχέση, ανάμεσα στις ασίστ και στα λάθη, όπως και ανάμεσα στις ασίστ με τις δύο κατηγορίες φάουλ.

Στο heatmap, με το έντονο κόκκινο χρώμα παρουσιάζονται οι θετικά, συσχετισμένες μεταβλητές και με το μπλε, οι αρνητικά, συσχετισμένες μεταβλητές. Ακόμη παρατηρούμε ότι οι ισχυρότερες, θετικά συσχετίσεις συγκεντρώνονται κυρίως στην κάτω δεξιά γωνία του heatmap. Όσο πιο έντονα τα χρώματα, τόσο πιο ισχυρή η συσχέτιση ανάμεσα στις μεταβλητές.



Σχήμα 4.4 : Heatmap για τις κανονικές, ποσοτικές μεταβλητές.

Στη συνέχεια, παρουσιάζεται η σχέση της κανονικής, κατηγορική μεταβλητής για τις σεζόν (Season) με τις υπόλοιπες κανονικές μεταβλητές, με βάση το συντελεστή του Spearman.

	Season
Points	0,14
2PT %	0,30
FT %	0,91
OR	0,48
DR	0,70
TR	0,65
AST	0,92
TO	-0,65
BLK	0,34
BLKA	0,36
FC	-0,89
FD	-0,86
PIR	0,54
Season	1

Πίνακας 4.19 : Πίνακας συσχετίσεων για τις κανονικές, ποσοτικές μεταβλητές με τη μεταβλητή των σεζόν.

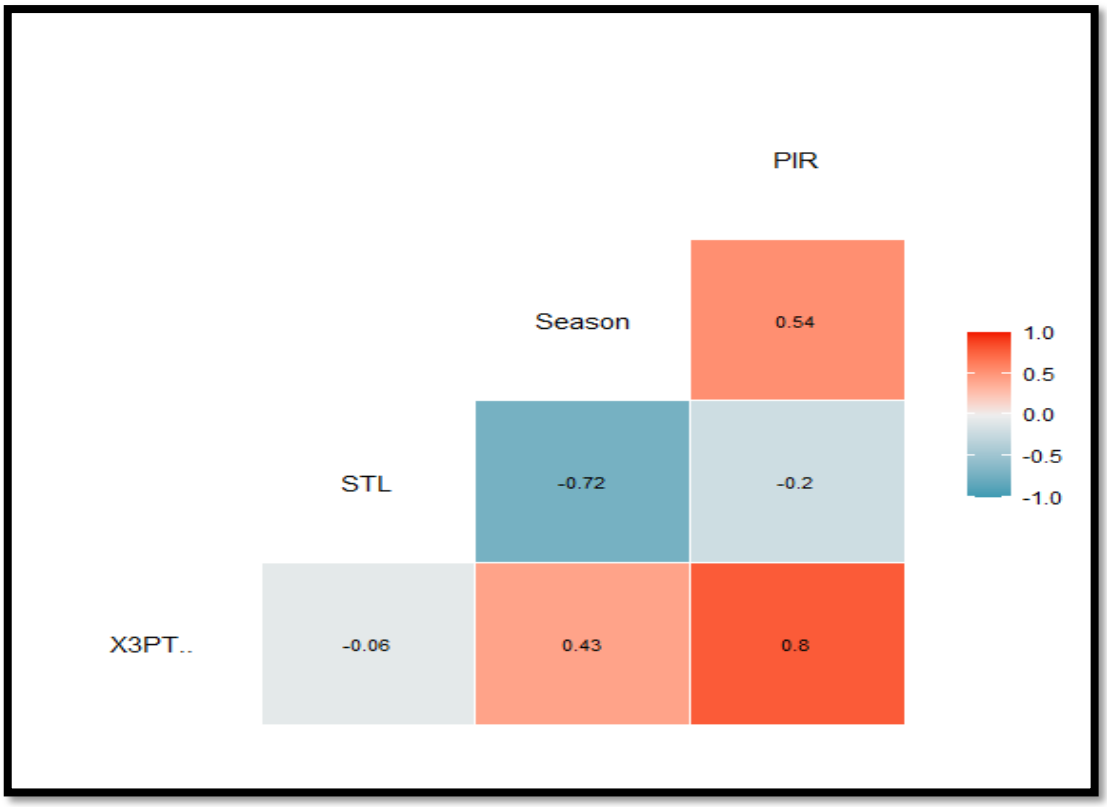
Από το παραπάνω πίνακα, φαίνεται ενδεικτικά πως ο δείκτης PIR έχει μια μέτρια, συσχέτιση με τη κατηγορική μεταβλητή των σεζόν. Η σχέση των σεζόν με το ποσοστό εύστοχων ελεύθερων βολών και με τις ασίστ είναι θετική και πολύ ισχυρή. Αντιθέτως, οι σεζόν είχαν πολύ υψηλή, αρνητική σχέση με τις δύο κατηγορίες των φάουλ.

Στη συνέχεια, εξετάζουμε τη σχέση των μη κανονικών μεταβλητών, με το δείκτη PIR και με τη μεταβλητή της σεζόν.

	3PT %	STL	Season	PIR
3PT %	1	-0,06	0,43	0,80
STL	-0,06	1	-0,72	-0,20
Season	0,43	-0,72	1	0,54
PIR	0,80	-0,20	0,54	1

Πίνακας 4.20 : Πίνακας συσχετίσεων για τις μη κανονικές, ποσοτικές μεταβλητές με αυτές των σεζόν και του PIR.

Ο δείκτης PIR έχει μια πολύ υψηλή, θετική συσχέτιση με το ποσοστό εύστοχων τριπόντων και μια ασθενή, αρνητική συσχέτιση με τα κλεψίματα.



Σχήμα 4.5 : Κορελόγραμμα για τις μη κανονικές, ποσοτικές μεταβλητές με αυτές των σεζόν και του PIR.

Τέλος, εξετάζουμε τη σχέση των μη κανονικών μεταβλητών, με τις κανονικές μεταβλητές.

	Points	2PT %	FT %	OR	DR	TR	AST
3PT %	0,68	0,64	0,36	-0,02	-0,02	-0,02	0,48
STL	0,09	0,09	-0,63	-0,52	-0,66	-0,63	-0,71
	TO	BLK	BLKA	FC	FD	PIR	
3PT %	-0,47	-0,21	-0,38	-0,43	-0,31	0,80	
STL	0,65	-0,58	-0,61	0,62	0,65	-0,20	

Πίνακας 4.21 : Πίνακας συσχετίσεων για τις μη κανονικές μεταβλητές με τις κανονικές.

Είναι απολύτως λογική η υψηλή, θετική σχέση ανάμεσα στο ποσοστό εύστοχων τριπόντων με τους πόντους, όπως και αυτή των λαθών με τα κλεψίματα. Ενδιαφέρουσα αντιθέτως, είναι η υψηλή, αρνητική σχέση των κλεψιμάτων με τις ασίστ, δηλαδή ομάδες που έκαναν πολλά κλεψίματα, φαίνεται να μην μοίραζαν ιδιαίτερα πολλές ασίστ, καθώς είναι πολύ πιθανό να εκδήλωναν άμεσους αιφνιδιασμούς.

4.2.2 Έλεγχος συσχετίσεων για την ανάλυση των πέντε καλύτερων παικτών σε PIR

Σε αυτήν την ενότητα, θα ελέγξουμε τη σχέση που έχουν μεταξύ τους οι ποσοτικές μεταβλητές των πόντων και του δείκτη PIR των πέντε καλύτερων παικτών, χωρισμένοι πάλι σε αυτούς των οποίων οι ομάδες προκρίθηκαν στις εκάστοτε φάσεις που εξετάζουμε, και σε αυτούς των οποίων δεν προκρίθηκαν οι ομάδες. Στην ενότητα 4.1.2, έγινε έλεγχος για τη κανονικότητα των μεταβλητών αυτών, για κάθε φάση.

Αρχικά, για τις ομάδες που δε προκρίθηκαν στα playoffs, οι πόντοι και ο δείκτης PIR παρουσιάζουν υψηλή, θετική συσχέτιση (0,73) με βάση το συντελεστή του Spearman. Για τις ομάδες που προκρίθηκαν στα playoffs, οι πόντοι και ο δείκτης PIR παρουσιάζουν μια υψηλή, θετική συσχέτιση (0,79) επίσης, με βάση το συντελεστή του Pearson αυτή τη φορά.

Για τις ομάδες που δεν προκρίθηκαν στο Final Four, οι πόντοι και ο δείκτης PIR παρουσιάζουν υψηλή, θετική συσχέτιση (0,74) με βάση το συντελεστή του Spearman. Τέλος, για τις ομάδες που προκρίθηκαν στο Final Four, οι πόντοι και ο δείκτης PIR παρουσιάζουν υψηλή, θετική συσχέτιση (0,77) επίσης, με βάση τώρα το συντελεστή του Pearson.

4.2.3 Έλεγχος συσχετίσεων για την ανάλυση των δέκα καλύτερων παικτών σε πόντους, και των δέκα καλύτερων σε PIR, ανά σεζόν

- Ο έλεγχος για τους δέκα καλύτερους παίκτες σε πόντους, ανά σεζόν.

Αρχικά, για τους παίκτες των ομάδων που δε προκρίθηκαν στα playoffs, οι πόντοι και ο δείκτης PIR παρουσιάζουν μια μέτρια, θετική συσχέτιση (0,52) με βάση το συντελεστή του Spearman. Για τις ομάδες που προκρίθηκαν στα playoffs, οι πόντοι και ο δείκτης PIR παρουσιάζουν υψηλή, θετική συσχέτιση (0,66) με βάση τον ίδιο συντελεστή.

Για τις ομάδες που δε προκρίθηκαν στο Final Four, οι πόντοι και ο δείκτης PIR παρουσιάζουν μέτρια, θετική συσχέτιση (0,59) με βάση το συντελεστή του Spearman. Τέλος, για τους παίκτες των ομάδων που προκρίθηκαν στο Final Four, οι πόντοι και ο δείκτης PIR παρουσιάζουν υψηλή, θετική συσχέτιση (0,63) επίσης, με βάση το συντελεστή του Pearson.

- Ο έλεγχος για τους δέκα καλύτερους παίκτες σε PIR, ανά σεζόν.

Για τους παίκτες των ομάδων που δε προκρίθηκαν στα playoffs, οι πόντοι και ο δείκτης PIR παρουσιάζουν μέτρια, θετική συσχέτιση (0,47) με βάση το συντελεστή του Spearman. Για τις ομάδες που προκρίθηκαν στα playoffs, οι πόντοι και ο δείκτης PIR παρουσιάζουν επίσης μέτρια, θετική συσχέτιση (0,44) με βάση τον ίδιο συντελεστή.

Για τις ομάδες που δε προκρίθηκαν στο Final Four, οι πόντοι και ο δείκτης PIR παρουσιάζουν μέτρια, θετική συσχέτιση (0,47) με βάση το συντελεστή του Spearman. Τέλος, για τους παίκτες των ομάδων που προκρίθηκαν στο Final Four, οι πόντοι και ο δείκτης PIR παρουσιάζουν επίσης μια μέτρια, θετική συσχέτιση (0,42) με βάση το συντελεστή του Spearman.

4.2.4 Έλεγχος συσχετίσεων για την ανάλυση των πενήντα καλύτερων παικτών σε PIR διαχρονικά

Για τους παίκτες των ομάδων που δε προκρίθηκαν στα playoffs, οι πόντοι και ο δείκτης PIR παρουσιάζουν μέτρια, θετική συσχέτιση (0,46) με βάση το συντελεστή του Spearman. Για τις ομάδες που προκρίθηκαν στα playoffs, οι πόντοι και ο δείκτης PIR παρουσιάζουν μια ασθενή, θετική συσχέτιση (0,29) με βάση τον ίδιο συντελεστή.

Για τις ομάδες που δε προκρίθηκαν στο Final Four, οι πόντοι και ο δείκτης PIR παρουσιάζουν μέτρια, θετική συσχέτιση (0,42) με βάση το συντελεστή του Spearman. Τέλος, για τους παίκτες των ομάδων που προκρίθηκαν στο Final Four, οι πόντοι και ο δείκτης PIR παρουσιάζουν μια μέτρια, θετική συσχέτιση (0,54) με βάση το συντελεστή του Spearman.

4.3 Έλεγχος για την ισότητα μέσων τιμών δύο δειγμάτων (t-tests)

Το t-test είναι μια στατιστική μέθοδος που χρησιμοποιείται για τη σύγκριση των μέσων δύο ομάδων. Συχνά χρησιμοποιείται στον έλεγχο υποθέσεων για να προσδιοριστεί, εάν μια διαδικασία ή θεραπεία έχει πράγματι επίδραση στον πληθυσμό που ενδιαφέρει, ή εάν δύο ομάδες είναι διαφορετικές μεταξύ τους. Μπορεί να χρησιμοποιηθεί, μόνο όταν συγκρίνονται οι μέσοι όροι δύο ομάδων (γνωστό και ως σύγκριση κατά ζεύγη).

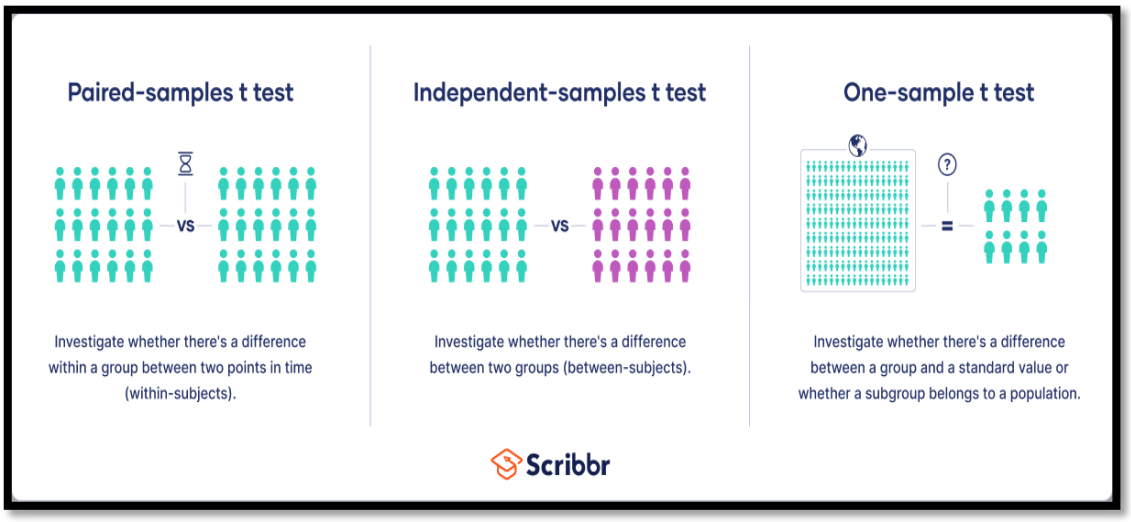
Το t-test είναι ένα παραμετρικό τεστ διαφοράς, που σημαίνει ότι κάνει τις ίδιες υποθέσεις για τα δεδομένα, με άλλα παραμετρικά τεστ. Προϋποθέτει τα δεδομένα να

- είναι ανεξάρτητα.

- κατανέμονται (περίπου) κανονικά.
- έχουν παρόμοια διακύμανση, σε κάθε ομάδα που συγκρίνεται (γνωστή και ως ομοιογένεια διακύμανσης).

Εάν τα δεδομένα δεν ταιριάζουν με αυτές τις παραδοχές, χρησιμοποιούνται μη παραμετρικές εναλλακτικές του t-test, όπως το τεστ Wilcoxon Signed-Rank, για δεδομένα με άνισες διακυμάνσεις.

Υπάρχουν διαφορετικά ήδη t-tests που χρησιμοποιούνται, ανάλογα με το εάν οι ομάδες που συγκρίνονται προέρχονται από έναν μόνο πληθυσμό ή δύο διαφορετικούς πληθυσμούς, και εάν γίνεται έλεγχος της διαφοράς σε μια συγκεκριμένη κατάσταση.



Σχήμα 4.6 : Διαφορετικές κατηγορίες των t-tests.

(Πηγή : https://www.scribbr.com/)

- Εάν οι ομάδες προέρχονται από έναν μόνο πληθυσμό (π.χ. μέτρηση πριν και μετά από μια πειραματική θεραπεία), χρησιμοποιείται ένα paired t-test.
- Εάν οι ομάδες προέρχονται από δύο διαφορετικούς πληθυσμούς (π.χ. δύο διαφορετικά είδη ή άτομα από δύο ξεχωριστές πόλεις), χρησιμοποιείται ένα two-sample t-test (γνωστό και ως ανεξάρτητο t-test).
- Εάν υπάρχει μια ομάδα που συγκρίνεται με μια τυπική τιμή (π.χ., σύγκριση της οξύτητας ενός υγρού με ουδέτερο pH=7), χρησιμοποιείται ένα one-sample t-test.

Αν μας ενδιαφέρει να εξετάσουμε το αν οι μέσες τιμές δύο πληθυσμών διαφέρουν μεταξύ τους, γίνεται χρήση ενός αμφίπλευρου t-test (two-sample t-test).

Αν εξετάζουμε το αν ο μέσος όρος ενός πληθυσμού είναι μεγαλύτερος ή μικρότερος από μια τιμή ελέγχου, γίνεται χρήση ενός μονόπλευρου t-test (one-sample t-test).

Το t-test εκτιμά την πραγματική διαφορά μεταξύ των μέσων τιμών δύο ομάδων, χρησιμοποιώντας το πηλίκο της διαφοράς των μέσων των ομάδων, προς το από κοινού τυπικό σφάλμα και των δύο ομάδων. Ο τύπος για το τεστ δύο δειγμάτων (γνωστό και ως το Student's t-test) είναι:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

όπου t είναι η τιμή t της κατανομής, $\bar{x}_1$ και $\bar{x}_2$ είναι οι μέσοι όροι των δύο ομάδων που συγκρίνονται, s^2 είναι το ομαδοποιημένο τυπικό σφάλμα των δύο ομάδων και n_1 και n_2 είναι ο αριθμός των παρατηρήσεων σε καθεμία από τις ομάδες.

Μπορεί να συγκριθεί η υπολογιζόμενη τιμή t με τις τιμές σε ένα γράφημα κρίσιμων τιμών (π.χ. πίνακας κατανομής t Student) για να προσδιοριστεί, εάν η τιμή t είναι μεγαλύτερη από ό,τι θα αναμενόταν τυχαία. Εάν ναι, μπορούμε να απορρίψουμε τη μηδενική υπόθεση και να συμπεράνουμε ότι οι δύο ομάδες προέρχονται από πληθυσμούς με διαφορετική μέση τιμή.

Κατά την αναφορά των αποτελεσμάτων, οι πιο σημαντικές τιμές που πρέπει να συμπεριληφθούν είναι η τιμή t, η p-value και οι βαθμοί ελευθερίας για το τεστ. Με αυτά βγαίνει συμπέρασμα, για το εάν η διαφορά μεταξύ των δύο ομάδων είναι στατιστικώς σημαντική (δηλαδή ότι είναι απίθανο να έχει συμβεί αυτό το αποτέλεσμα τυχαία). (Bevans, 2022)

Στην παρούσα εργασία, μπορεί τα δεδομένα μας να μην ακολουθούν όλα κανονική κατανομή, αλλά επειδή είναι πολλά στο πλήθος τους (n>50), από το Κεντρικό Οριακό Θεώρημα, μπορούμε να χρησιμοποιήσουμε χωρίς πρόβλημα τα κατάλληλα t-tests, σε επίπεδο σημαντικότητας 5% . Αν είχαμε μικρό πλήθος μη κανονικών δεδομένων, θα έπρεπε να χρησιμοποιήσουμε μη παραμετρικά τεστ, όπως το Wilcoxon signed rank test, το Mann-Whitney U Test, ή το Kruskal-Wallis test.

Εφόσον δεν έχουμε κάνει κάποια υπόθεση για την ισότητα των διακυμάνσεων, θα χρησιμοποιήσουμε το “Welch Two Sample t-test”, για τις αναλύσεις μας.

4.3.1 Έλεγχος για την ισότητα των μέσων τιμών δύο δειγμάτων για την ανάλυση των πέντε καλύτερων παικτών σε PIR

Σε αυτήν την ενότητα, πραγματοποιήθηκαν t-tests για τη σύγκριση των μέσων τιμών των πόντων και του δείκτη PIR, ανάμεσα στις ομάδες που προκρίθηκαν στις δύο φάσεις και σε αυτές που δεν προκρίθηκαν, με βάση τους πέντε καλύτερους παίκτες ανά σεζόν.

Αρχικά, αναφορικά με τη φάση των playoffs, οι μέσες τιμές των πόντων διαφέρουν σημαντικά, αφού η p-value του ελέγχου είναι μικρότερη από το επίπεδο σημαντικότητας 5%. Η τιμή t είναι -2,137 , οι βαθμοί ελευθερίας είναι 317,99 και η p-value του ελέγχου ισούται με 0,03336. Οι ομάδες που προκρίθηκαν στα playoffs είχαν μεγαλύτερο μέσο όρο πόντων, έναντι αυτών που δεν προκρίθηκαν. Το ίδιο συμπέρασμα σχετικά με τις μέσες τιμές, προκύπτει και για το δείκτη PIR (p = 2,006e-11), καθώς οι ομάδες που προκρίθηκαν είχαν μεγαλύτερο μέσο όρο PIR.

```
        Welch Two Sample t-test

data:  k1$Points and k2$Points
t = -2.137, df = 317.99, p-value = 0.03336
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.52990673 -0.02188351
sample estimates:
mean of x mean of y
 10.68596  10.96185
```

Πίνακας 4.22 : Output για τον έλεγχο των πόντων των ομάδων, στα playoffs.

```
        Welch Two Sample t-test

data:  k1$PIR and k2$PIR
t = -6.9467, df = 329.25, p-value = 2.006e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.3199249 -0.7373434
sample estimates:
mean of x mean of y
 11.49706  12.52570
```

Πίνακας 4.23 : Output για τον έλεγχο του PIR των ομάδων, στα playoffs.

Στη συνέχεια, εξετάζοντας τη φάση του Final Four, συμπεράναμε πως οι μέσες τιμές των πόντων δε διαφέρουν (για τις δύο κατηγορίες ομάδων), αφού η p-value είναι μεγαλύτερη τώρα από το επίπεδο σημαντικότητας 0,05 , ενώ οι μέσες τιμές του PIR διαφέρουν (p = 1,231e-08), καθώς οι ομάδες που προκρίθηκαν στο Final Four είχαν μεγαλύτερο μέσο όρο PIR, όπως θα ήταν αναμενόμενο.

```
        Welch Two Sample t-test

data:  k3$Points and k4$Points
t = -1.3005, df = 115.45, p-value = 0.196
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4940917  0.1024251
sample estimates:
mean of x mean of y
 10.75556  10.95139
```

Πίνακας 4.24 : Output για τον έλεγχο των πόντων των ομάδων, στο Final Four.

```
        Welch Two Sample t-test

data:  k3$PIR and k4$PIR
t = -6.1539, df = 110.93, p-value = 1.231e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.4827248 -0.7604233
sample estimates:
mean of x mean of y
 11.68537  12.80694
```

Πίνακας 4.25 : Output για τον έλεγχο του PIR των ομάδων, στο Final Four.

4.3.2 Έλεγχος για την ισότητα των μέσων τιμών δύο δειγμάτων για την ανάλυση των δέκα καλύτερων παικτών σε πόντους, και των δέκα καλύτερων σε PIR, ανά σεζόν

Σε αυτήν την ενότητα, πραγματοποιήθηκαν τώρα t-tests για τη σύγκριση των μέσων τιμών των πόντων και του δείκτη PIR, ανάμεσα στις ομάδες που προκρίθηκαν στις δύο φάσεις και σε αυτές που δεν προκρίθηκαν, με βάση τους δέκα καλύτερους παίκτες σε πόντους και τους δέκα καλύτερους σε PIR.

- Ο έλεγχος για τους δέκα καλύτερους παίκτες σε πόντους, ανά σεζόν.

Αρχικά, αναφορικά με τα playoffs, οι μέσες τιμές των πόντων δεν διαφέρουν (p = 0,9787), όπως και οι μέσες τιμές του δείκτη PIR (p = 0,4654).

```
        Welch Two Sample t-test

data:  k5$Points and k6$Points
t = -0.026799, df = 159.51, p-value = 0.9787
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6916274  0.6731089
sample estimates:
mean of x mean of y
 16.56019  16.56944
```

Πίνακας 4.26 : Output για τους πόντους των ομάδων στα playoffs, για τους δέκα κορυφαίους παίκτες σε πόντους.

```
        Welch Two Sample t-test

data:  k5$PIR and k6$PIR
t = -0.73176, df = 161.46, p-value = 0.4654
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.4520593  0.6668741
sample estimates:
mean of x mean of y
 17.12407  17.51667
```

Πίνακας 4.27 : Output για το PIR των ομάδων στα playoffs, για τους δέκα κορυφαίους παίκτες σε πόντους.

Εξετάζοντας στη συνέχεια τη φάση του Final Four, παρατηρούμε πως οι μέσες τιμές τόσο για τους πόντους (p = 0,3928), όσο και για το PIR δεν διαφέρουν σημαντικά (p = 0,1826).

```
        Welch Two Sample t-test

data:  k7$Points and k8$Points
t = 0.85965, df = 72.077, p-value = 0.3928
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3821263  0.9615944
sample estimates:
mean of x mean of y
 16.61701  16.32727
```

Πίνακας 4.28 : Output για τους πόντους των ομάδων στο Final Four, για τους δέκα κορυφαίους παίκτες σε πόντους.

```
        Welch Two Sample t-test

data:  k7$PIR and k8$PIR
t = -1.3498, df = 54.583, p-value = 0.1826
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.0675278  0.4034585
sample estimates:
mean of x mean of y
 17.12857  17.96061
```

Πίνακας 4.29 : Output για το PIR των ομάδων στο Final Four, για τους δέκα κορυφαίους παίκτες σε πόντους.

- Ο έλεγχος για τους δέκα καλύτερους παίκτες σε PIR, ανά σεζόν.

Αναφορικά τώρα με τη φάση των playoffs, οι μέσες τιμές των πόντων δε διαφέρουν σημαντικά (p = 0,5104), όπως και οι μέσες τιμές για τους δείκτες PIR (p = 0,8092).

```
        Welch Two Sample t-test

data:  k9$Points and k10$Points
t = 0.65961, df = 176.97, p-value = 0.5104
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6006104  1.2036781
sample estimates:
mean of x mean of y
 15.50851  15.20698
```

Πίνακας 4.30 : Output για τους πόντους των ομάδων στα playoffs, για τους δέκα κορυφαίους παίκτες σε PIR.

```
        Welch Two Sample t-test

data:  k9$PIR and k10$PIR
t = 0.24181, df = 176.76, p-value = 0.8092
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.7012589  0.8971025
sample estimates:
mean of x mean of y
 18.92234  18.82442
```

Πίνακας 4.31 : Output για το PIR των ομάδων στα playoffs, για τους δέκα κορυφαίους παίκτες σε PIR.

Στη συνέχεια, εξετάζοντας τη φάση του Final Four, οι μέσες τιμές των πόντων δεν διαφέρουν σε ε.σ. 5% (p = 0,07959). Το ίδιο συμπέρασμα ισχύει και για το δείκτη PIR (p = 0,4772).

```
        Welch Two Sample t-test

data:  k11$Points and k12$Points
t = 1.7681, df = 119.81, p-value = 0.07959
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.09395049  1.66192058
sample estimates:
mean of x mean of y
 15.57786  14.79388
```

Πίνακας 4.32 : Output για τους πόντους των ομάδων στο Final Four, για τους δέκα κορυφαίους παίκτες σε PIR.

```
        Welch Two Sample t-test

data:  k11$PIR and k12$PIR
t = 0.71337, df = 105.14, p-value = 0.4772
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5290371  1.1236468
sample estimates:
mean of x mean of y
 18.95649  18.65918
```

Πίνακας 4.33 : Output για το PIR των ομάδων στο Final Four, για τους δέκα κορυφαίους παίκτες σε PIR.

4.3.3 Έλεγχος για την ισότητα των μέσων τιμών δύο δειγμάτων για την ανάλυση των πενήντα καλύτερων παικτών σε PIR διαχρονικά

Τέλος, σε αυτήν την ενότητα, πραγματοποιήθηκαν t-tests για τη σύγκριση των μέσων τιμών των πόντων και του δείκτη PIR, ανάμεσα στις ομάδες που προκρίθηκαν στις δύο φάσεις και σε αυτές που δεν προκρίθηκαν, με βάση τους πενήντα καλύτερους παίκτες σε PIR διαχρονικά.

Εξετάζοντας τα playoffs, οι μέσες τιμές των πόντων δεν διαφέρουν σημαντικά (p = 0,2734), όπως και οι μέσες τιμές για το δείκτη PIR (p = 0,287).

```
        Welch Two Sample t-test

data:  k13$Points and k14$Points
t = 1.1081, df = 47.576, p-value = 0.2734
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.8832708  3.0512708
sample estimates:
mean of x mean of y
   18.300    17.216
```

Πίνακας 4.34 : Output για τους πόντους των ομάδων στα playoffs, για τους πενήντα κορυφαίους διαχρονικά.

```
        Welch Two Sample t-test

data:  k13$PIR and k14$PIR
t = 1.0771, df = 46.62, p-value = 0.287
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.7119163  2.3519163
sample estimates:
mean of x mean of y
   22.792    21.972
```

Πίνακας 4.35 : Output για το PIR των ομάδων στα playoffs, για τους πενήντα κορυφαίους διαχρονικά.

Ολοκληρώνοντας, εξετάζοντας τη φάση του Final Four, υπάρχει ισχυρή ένδειξη ότι οι μέσες τιμές των πόντων διαφέρουν σημαντικά (p = 0,02541). Κάτι που είναι ιδιαιτέρως ενδιαφέρον είναι πως οι ομάδες που δεν προκρίθηκαν στο Final Four φαίνεται να έχουν μεγαλύτερο μέσο όρο πόντων, σε σχέση με αυτές που προκρίθηκαν. Αντιθέτως, οι μέσοι όροι των δεικτών PIR δεν διαφέρουν σημαντικά, για τις δύο κατηγορίες ομάδων (p = 0,5251).

```
        Welch Two Sample t-test

data:  k15$Points and k16$Points
t = 2.3678, df = 26.689, p-value = 0.02541
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2915062 4.0935815
sample estimates:
mean of x mean of y
 18.28421  16.09167
```

Πίνακας 4.36 : Output για τους πόντους των ομάδων στο Final Four, για τους πενήντα κορυφαίους διαχρονικά.

```
        Welch Two Sample t-test

data:  k15$PIR and k16$PIR
t = 0.64401, df = 26.686, p-value = 0.5251
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.051659  2.013063
sample estimates:
mean of x mean of y
 22.49737  22.01667
```

Πίνακας 4.37 : Output για το PIR των ομάδων στο Final Four, για τους πενήντα κορυφαίους διαχρονικά.

ΚΕΦΑΛΑΙΟ 5ο

Σε αυτό το κεφάλαιο, αποσκοπούμε στο να εξετάσουμε ποιες από τις μεταβλητές μας παίζουν καθοριστικό ρόλο στο αν μια ομάδα προκρίνεται ή όχι, στις δύο φάσεις της διοργάνωσης που εξετάζουμε. Για να γίνει αυτό, προσαρμόζουμε τα κατάλληλα γενικευμένα γραμμικά μοντέλα, με μεταβλητές απόκρισης τις κατηγορικές μεταβλητές **Playoffs / Quarter-Finals** και **Final 4 / Semi-Finals** , οι οποίες όπως έχουμε αναφέρει στα προηγούμενα κεφάλαια, αποτελούν ενδείξεις για το αν οι ομάδες προκρίθηκαν (ή όχι) στις αντίστοιχες φάσεις της διοργάνωσης, και στη συνέχεια ερμηνεύουμε τα αποτελέσματα μας.

5.1 Εισαγωγή στην Ανάλυση Παλινδρόμησης

Ο κλάδος της Στατιστικής που εξετάζει τη σχέση μεταξύ δύο ή περισσότερων μεταβλητών, με απώτερο στόχο την πρόβλεψη μιας απ' αυτές μέσω των άλλων λέγεται ανάλυση παλινδρόμησης (regression analysis). Η απλούστερη περίπτωση παλινδρόμησης είναι η απλή γραμμική παλινδρόμηση (simple linear regression) κατά την οποία χρησιμοποιούμε μόνο μια μεταβλητή X, η οποία ονομάζεται ανεξάρτητη μεταβλητή, και μια δεύτερη μεταβλητή Y, η οποία ονομάζεται εξαρτημένη μεταβλητή ή μεταβλητή απόκρισης, και η οποία μπορεί να προσεγγιστεί ικανοποιητικά από μία γραμμική συνάρτηση της X. Αντιθέτως, η παλινδρόμηση στην οποία υπάρχουν περισσότερες από μία ανεξάρτητες μεταβλητές X_1, X_2,…, X_p λέγεται πολλαπλή παλινδρόμηση και είναι ουσιαστικά μια επέκταση της απλής γραμμικής παλινδρόμησης.

Το (στατιστικό) μοντέλο της πολλαπλής παλινδρόμησης έχει την εξής μορφή:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

όπου i=1,…,ν , Y_i είναι η i-οστή τιμή της εξαρτημένης μεταβλητής Y, β_0, β_1,…, β_p είναι οι τιμές των παραμέτρων του μοντέλου και x_{i1},…, x_{ip} είναι οι i-οστές τιμές των p ανεξάρτητων μεταβλητών και ε_i τα σφάλματα.

Για το παραπάνω μοντέλο ισχύουν οι παρακάτω υποθέσεις :

- Οι ποσότητες β_0, β_1,…, β_p είναι άγνωστες παράμετροι.
- Τα x_{i1},…, x_{ip} είναι γνωστοί αριθμοί. Πιο συγκεκριμένα, είναι οι τιμές των ανεξάρτητων (ή προβλεπουσών) μεταβλητών κατά την i-οστή επανάληψη του πειράματος. Οι τιμές αυτές καθορίζονται από τον ερευνητή που εκτελεί το πείραμα.
- Το Y_i είναι η τιμή της εξαρτημένης μεταβλητής (ή μεταβλητής απόκρισης) κατά την i-οστή επανάληψη του πειράματος. Το Y_i είναι τυχαία μεταβλητή.

- Τα ε_i , i=1,...,ν είναι τυχαία σφάλματα με μέση τιμή 0 και διασπορά σ^2 .
- Τα σφάλματα ε_i και ε_j που αντιστοιχούν σε διαφορετικές επαναλήψεις του πειράματος ($j \neq i$) θεωρούνται ασυσχέτιστα, δηλαδή ισχύει $Cov(\varepsilon_i,\varepsilon_j) = 0$ για $j \neq i$.

(Κούτρας, 2021)

Βασική προϋπόθεση για την εφαρμογή του συγκεκριμένου μοντέλου είναι η κανονικότητα των σφαλμάτων και κατ' επέκταση της εξαρτημένης μεταβλητής, πράγμα το οποίο παραβιάζεται και για τις δύο κατηγορικές μας μεταβλητές, σ' όλες τις αναλύσεις μας. Δεδομένου ότι πρόκειται για δίτιμες, κατηγορικές μεταβλητές, δεν έχει νόημα να εξετάσουμε τη κανονικότητα τους. Συνεπώς, μπορεί να εξεταστεί η προσαρμογή Γενικευμένων Γραμμικών Μοντέλων (ΓΓΜ) με εξαρτημένες μεταβλητές, τις παραπάνω δίτιμες.

5.2 Γενικευμένα Γραμμικά Μοντέλα

Σε ένα ΓΓΜ, η μεταβλητή απόκρισης εισέρχεται στο μοντέλο μέσω της μέσης της τιμής,

$$\eta_i = g(\mu_i) = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij}$$

Με άλλα λόγια, δε γίνεται **καμία υπόθεση για την κατανομή των σφαλμάτων** στο μοντέλο παρά μόνο για την κατανομή της Y. Η συνάρτηση $g(\mu_i)$ ονομάζεται συνάρτηση σύνδεσης και συνδέει τη μέση τιμή μi της εξαρτημένης μεταβλητής με το σταθερό κομμάτι του μοντέλου μας. Ο παραπάνω τύπος μοντέλων αφορά τις κατανομές, οι οποίες ανήκουν στην εκθετική οικογένεια κατανομών. Οι εκτιμήσεις των παραμέτρων του εκάστοτε μοντέλου βρίσκονται με τη βοήθεια της συνάρτησης πιθανοφάνειας, και συνεπώς έχουν μια σειρά από επιθυμητές ιδιότητες.

Τα πλεονεκτήματα των ΓΓΜ έναντι της συνήθους παλινδρόμησης είναι τα εξής :

- Πολύ μεγαλύτερο φάσμα εφαρμογών. Χρησιμοποιούνται και σε περιπτώσεις όπου δεν μπορεί να υποτεθεί ότι η κατανομή της Y είναι κανονική, ούτε καν προσεγγιστικά π.χ. όταν η Y παίρνει μόνο δύο τιμές (Bernoulli).
- Οι εκτιμητές των παραμέτρων προκύπτουν με τη μέθοδο μέγιστης πιθανοφάνειας, άρα έχουν μια σειρά από επιθυμητές ιδιότητες.
- Στις περισσότερες περιπτώσεις, δε χρειάζεται να υποθέσουμε σταθερή διακύμανση για τις τιμές της Y.

- Με την ενοποιημένη θεωρία των ΓΓΜ, δε χρειάζεται να χρησιμοποιήσουμε διαφορετικό μοντέλο ανάλογα με το αν οι ερμηνευτικές μεταβλητές είναι ποσοτικές ή ποιοτικές (ή μείξη των δύο).
- Στην περίπτωση που όλες οι μεταβλητές είναι κατηγορικές, τα ΓΓΜ αποτελούν ένα βασικό τρόπο ανάλυσης σε πίνακες συνάφειας.

(Πολίτης, 2021)

Στο πλαίσιο ενός ΓΓΜ, μία διάκριση που μας ενδιαφέρει είναι αν τα δεδομένα μας είναι ομαδοποιημένα, οπότε μιλάμε συνήθως για διωνυμικά δεδομένα (binomial data), ή αν δεν είναι ομαδοποιημένα, οπότε γνωρίζουμε για κάθε άτομο στο δείγμα την τιμή της απόκρισης (συνήθως 0 = αποτυχία, 1 = επιτυχία). Τότε μιλάμε για δίτιμα δεδομένα (binary data). Στη δικιά μας περίπτωση, οι μεταβλητές απόκρισης είναι δίτιμες όπως έχουμε αναφέρει, με τη τιμή 0 να συμβολίζει τη μη πρόκριση της ομάδας στις εξεταζόμενες φάσεις, και τη τιμή 1 τη πρόκριση αντίστοιχα.

Για δίτιμα δεδομένα, η κατανομή πιθανότητας των Y_i είναι

$$P[Y_i = y_i] = p_i^{y_i}(1 - p_i)^{1-y_i}$$

όπου $y_i = 0,1$.

Τρεις συναρτήσεις σύνδεσης που χρησιμοποιούμε για δίτιμα δεδομένα:

1. **Logit**

$$\eta_i = logit(p_i) = log\left[\frac{p_i}{1 - p_i}\right]$$

2. **Probit**

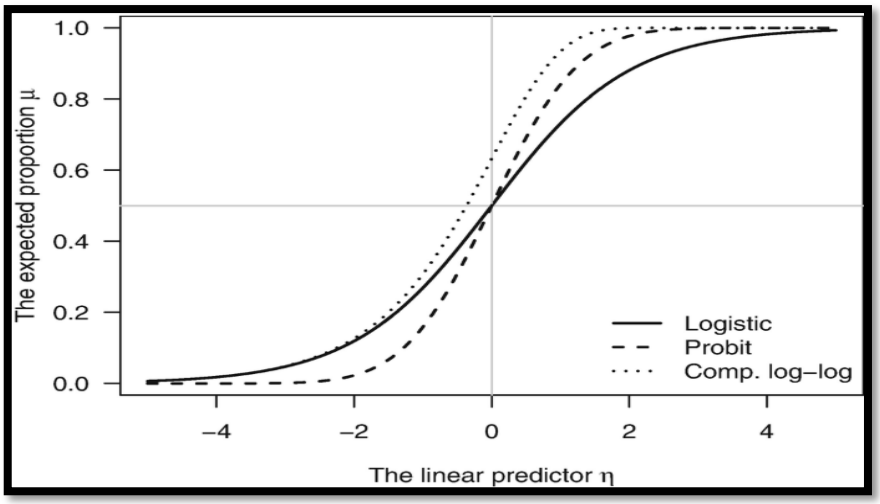
$$\eta_i = probit(p_i) = \phi^{-1}(p_i)$$

όπου Φ είναι η αθροιστική συνάρτηση κατανομής της τυποποιημένης κανονικής.

3. **Complementary log log**

$$\eta_i = log[-log(1 - p_i)]$$

Όταν χρησιμοποιούμε ως συνάρτηση σύνδεσης τη συνάρτηση logit, τότε μιλάμε για ένα μοντέλο λογιστικής παλινδρόμησης.



Σχήμα 5.1 : Γραφική αναπαράσταση των συναρτήσεων σύνδεσης.

5.3 Λογιστική Παλινδρόμηση

Η λογιστική παλινδρόμηση χρησιμοποιείται όταν επιθυμούμε να προβλέψουμε την απουσία ή την παρουσία ενός χαρακτηριστικού, ή την εμφάνιση ενός συμβάντος. Είναι μια γενίκευση της απλής γραμμικής παλινδρόμησης, για την περίπτωση όπου η εξαρτημένη μεταβλητή Y είναι δίτιμη (δηλαδή παίρνει την τιμή 0, όταν απουσιάζει το χαρακτηριστικό ή την τιμή 1, όταν υπάρχει το χαρακτηριστικό).

Το γραμμικό μοντέλο είναι αδύνατο να χρησιμοποιηθεί, όταν η μεταβλητή Y είναι δυαδική και έχουμε τα εξής τρία προβλήματα:

1. Τα σφάλματα δεν είναι κανονικά.
2. Τα σφάλματα έχουν άνισες διασπορές.
3. Περιορισμός στη συνάρτηση απόκρισης (η προβλεπόμενη πιθανότητα θα πρέπει να ανήκει στα διάστημα (0,1))

Παρόλο που στα δύο πρώτα προβλήματα είναι δυνατό σε κάποιες περιπτώσεις να τα παραλείψουμε και να χρησιμοποιήσουμε την γραμμική παλινδρόμηση, εφαρμόζοντας κάποιες άλλες τεχνικές, το τρίτο πρόβλημα μας το απαγορεύει ρητά, γιατί δεν μπορεί να αντιμετωπιστεί με διαφορετικό τρόπο.

Η σημαντικότερη διαφοροποίηση μεταξύ λογιστικής και γραμμικής παλινδρόμησης βασίζεται στη φύση της επιλεγμένης μεταβλητής απόκρισης, η οποία στην μεν πρώτη μπορεί να είναι κατηγορική, (τακτική ή ονομαστική), στη δε δεύτερη αποκλειστικά ποσοτική και συνεχής. Ενώ κατά την κλασική γραμμική παλινδρόμηση η εκτίμηση των παραμέτρων α και β_i γίνεται με τη μέθοδο των ελάχιστων τετραγώνων, κατά τη λογιστική παλινδρόμηση η εκτίμηση των παραμέτρων γίνεται με τη μέθοδο της μέγιστης πιθανοφάνειας, δηλαδή επιλέγονται οι πιο πιθανοφανείς εκτιμήσεις των

παραμέτρων, προκειμένου να οδηγήσουν στα παρατηρούμενα αποτελέσματα. Ως επακόλουθο, η πρώτη παραδέχεται την ύπαρξη ομοιογένειας (ομοσκεδαστικότητας) στα υπολείμματα των αποκρίσεων, ενώ στη δεύτερη αναπτύσσεται πάντα ετεροσκεδαστικότητα σε κάθε προβλεπόμενη τιμή εξαιτίας του μεταβαλλόμενου ποσοστού διακύμανσης που αναλογεί σε αυτήν. Διακρίνονται τρεις τύποι λογιστικής παλινδρόμησης, ανάλογα με την φύση της εξαρτημένης κατηγορικής μεταβλητής η οποία μπορεί να είναι:

1. Δίτιμη ή δυαδική ή διχοτομική (binary) ή διμερής εξαρτημένη μεταβλητή. Συνίσταται από δύο κατηγορίες, όπως π.χ. είναι οι εκβάσεις επιτυχία/αποτυχία.

2. Διατάξιμη (ordinal) μεταβλητή. Η εξαρτημένη μεταβλητή συνίσταται από τρεις ή περισσότερες κατηγορίες, μεταξύ των οποίων ισχύει η έννοια της διάταξης, όπως π.χ. σε μια ερώτηση της κλίμακας διαφωνώ : καθόλου, λίγο, μέτρια, αρκετά, πολύ.

3. Ονομαστική (Nominal) ή πολυωνυμική (polynomial) ή πολυτομική (polychotomus) ή κατηγορική αδιαβάθμητη (non-ordered categorical) ή πολυμερής μεταβλητή απόκρισης. Περιέχει τρεις ή περισσότερες κατηγορίες, χωρίς κάποια φυσική διαβάθμιση, όπως π.χ. το χρώμα αντικειμένων.

(Καλλιακμάνης, 2020)

Το μοντέλο logit

Όπως αναφέραμε νωρίτερα, η λογιστική παλινδρόμηση είναι το γενικευμένο γραμμικό μοντέλο για δίτιμες αποκρίσεις, όταν η συνάρτηση σύνδεσης είναι η logit. Αν Y είναι μία δίτιμη απόκριση με P(Y = 1) = π = E(Y), το μοντέλο της λογιστικής παλινδρόμησης εκφράζεται ως εξής :

$$logit(\pi) \equiv log\left(\frac{\pi}{1-\pi}\right) = b_0 + b_1 \cdot x_1 + \cdots + b_p \cdot x_p$$

Αντιστρέφοντας την logit, βλέπουμε ότι το μοντέλο της λογιστικής παλινδρόμησης εκφράζει την πιθανότητα :

$$\pi = \frac{e^{b_0 + b_1 \cdot x_1 + \cdots + b_p \cdot x_p}}{1 + e^{b_0 + b_1 \cdot x_1 + \cdots + b_p \cdot x_p}}$$

ή ισοδύναμα την σχετική πιθανότητα (odds) ως :

$$\frac{\pi}{1-\pi} = e^{b_0 + b_1 \cdot x_1 + \cdots + b_p \cdot x_p}$$

Συνεπώς, η logit συνάρτηση εκφράζει το λογάριθμο της σχετικής πιθανότητας για ένα γεγονός (odds), δηλαδή το λογάριθμο της πιθανότητας να συμβεί ένα γεγονός προς την πιθανότητα να μην συμβεί το γεγονός. Οι συντελεστές b_i δείχνουν πόσο αλλάζει το logit βασισμένο στις τιμές των επεξηγηματικών μεταβλητών. Η logit είναι μια γνησίως αύξουσα συνάρτηση.

Το μοντέλο probit

Η ιδέα της συνάρτησης probit δημοσιεύτηκε από τον Chester Ittner Bliss σε ένα άρθρο του (1934) στο Science[18] , σχετικά με τον τρόπο διαχείρισης δεδομένων, όπως το ποσοστό παρασίτων που σκοτώθηκαν από ένα φυτοφάρμακο. Στον ίδιο οφείλεται η ονομασία του, ως αποτέλεσμα της ένωσης μερών των λέξεων **prob**ability un**it**. Η προσαρμογή των μοντέλων probit δεν διαφέρει πολύ από την προσαρμογή των μοντέλων logit. Η σημαντική τους διαφορά είναι πως τα πρώτα χρησιμοποιούν την αθροιστική Gaussian κατανομή.

Ο σκοπός του μοντέλου είναι να εκτιμήσει την πιθανότητα μια παρατήρηση με συγκεκριμένα χαρακτηριστικά να εμπίπτει σε μια συγκεκριμένη από τις κατηγορίες. Επιπλέον, η ταξινόμηση των παρατηρήσεων, με βάση τις προβλεπόμενες πιθανότητες είναι ένας τύπος μοντέλου για δυαδική ταξινόμηση.

Έστω ότι Y είναι μια δίτιμη, εξαρτημένη μεταβλητή με μόνο δύο τιμές, 0 και 1, και έστω X μια επεξηγηματική μεταβλητή που επηρεάζει το αποτέλεσμα της Y. Η μορφή που έχει το μοντέλο αυτό είναι :

$$P(Y = 1|X) = \phi(b_0 + b_1 \cdot x_1 + \cdots + b_p \cdot x_p)$$

όπου Φ είναι η αθροιστική συνάρτηση κατανομής της τυπικής κανονικής κατανομής.

Το μοντέλο clog-log

Το complementary log log μοντέλο ταιριάζει σε περιπτώσεις που η P(Y =1) πλησιάζει γρήγορα τη μονάδα, αλλά αργά το 0. Το μοντέλο αυτό έχει τη μορφή :

$$cloglog(\pi) \equiv log(-log(1 - \pi_x)) = b_0 + b_1 \cdot x_1 + \cdots + b_p \cdot x_p$$

όπου $\pi_x = P(Y = 1|X = x)$.

[18] Το Science, επίσης ευρέως γνωστό ως Science Magazine, είναι το ακαδημαϊκό περιοδικό της Αμερικανικής Ένωσης για την Προώθηση της Επιστήμης και ένα από τα κορυφαία ακαδημαϊκά περιοδικά στον κόσμο. Εκδόθηκε για πρώτη φορά το 1880, κυκλοφορεί σήμερα εβδομαδιαία και έχει βάση συνδρομητών περίπου ίση με 130.000. Επειδή οι συνδρομές των ιδρυμάτων και η διαδικτυακή πρόσβαση εξυπηρετούν μεγαλύτερο κοινό, το εκτιμώμενο αναγνωστικό κοινό είναι πάνω από 400.000 άτομα.
(Πηγή : https://en.wikipedia.org/wiki/Science_(journal))

5.4 Προσαρμογή μοντέλων λογιστικής παλινδρόμησης

5.4.1 Μοντέλο για τη φάση των Quarter-Finals / Playoffs

Αρχικά, προσαρμόσαμε ένα μοντέλο λογιστικής παλινδρόμησης, έχοντας ως μεταβλητή απόκρισης τη κατηγορική μεταβλητή "Playoffs / Quarter-Finals", και ανεξάρτητες μεταβλητές τις :

Μεταβλητές	Επεξήγηση
Points	Πόντοι
2PT %	Ποσοστό επιτυχημένων διπόντων
3PT %	Ποσοστό επιτυχημένων τριπόντων
FT %	Ποσοστό επιτυχημένων ελευθέρων βολών
OR	Επιθετικά "Ριμπάουντ"
DR	Αμυντικά "Ριμπάουντ"
TR	Συνολικά "Ριμπάουντ"
AST	"Ασίστ" (Τελικές πάσες που οδήγησαν σε καλάθι)
STL	Κλεψίματα
TO	Λάθη
BLK	Κοψίματα υπέρ της ομάδας
BLKA	Κοψίματα κατά της ομάδας
FC	Φάουλ στα οποία υπέπεσε η ομάδα
FD	Φάουλ τα οποία κέρδισε η ομάδα
PIR	Δείκτης Performance Index Rating
PIR_INDIV	Δείκτης Performance Index Rating του καλύτερου παίκτη της κάθε ομάδας για την αντίστοιχη σεζόν
PIR_BEST_FIVE	Μέσος όρος του δείκτη Performance Index Rating για τους πέντε καλύτερους παίκτες της κάθε ομάδας για την αντίστοιχη σεζόν

Πίνακας 5.1 : Μεταβλητές που χρησιμοποιήθηκαν στα μοντέλα.

Είναι σημαντικό να αναφερθεί πως και σε αυτό το κεφάλαιο, δεν λήφθηκαν υπόψιν οι σεζόν 2001-2004, καθώς δεν υπήρχαν προημιτελικές φάσεις, όπως και η σεζόν 2019-2020, η οποία δεν ολοκληρώθηκε λόγω της πανδημίας του COVID-19. Ξεκινώντας την ανάλυση και λαμβάνοντας υπόψη τα αποτελέσματα του ελέγχου συσχετίσεων από το προηγούμενο κεφάλαιο, χρησιμοποιείται ένα από τα πλέον σημαντικά κριτήρια για την εύρεση του κατάλληλου μοντέλου που περιγράφει καλύτερα την μεταβλητή απόκρισης, το κριτήριο του AIC (Akaike's Information Criterion). Για ένα μοντέλο με k παραμέτρους, το AIC ορίζεται ως

$$AIC = 2k - ln(L)$$

όπου L είναι η μέγιστη πιθανοφάνεια για το συγκεκριμένο μοντέλο. Ανάμεσα σε διαφορετικά μοντέλα, το «βέλτιστο» μοντέλο θεωρείται αυτό που έχει τη μικρότερη τιμή για το δείκτη AIC.

Έγινε χρήση της συνάρτησης **step** του πακέτου 'MASS' στην R. Η προεπιλογή σχετικά με την κατεύθυνση που θα χρησιμοποιηθεί (προς τα πίσω αποκλεισμός, ή προς τα εμπρός επιλογή των παραμέτρων που θα είναι στο μοντέλο μας) είναι αυτή που χρησιμοποιεί και τις δύο αυτές περιπτώσεις (both). Καταλήξαμε συνεπώς πως το καταλληλότερο μοντέλο είναι αυτό που σαν ανεξάρτητες μεταβλητές έχει τις PIR, AST, STL, FT%, FC, Points, TO και 2PT% , με δείκτη AIC ίσο με 348,99 .

Χρησιμοποιώντας την εντολή summary(), πήραμε πληροφορίες για το μοντέλο που προσαρμόσαμε, αναφορικά με τις τιμές των παραμέτρων του μοντέλου, καθώς και βασικές μετρήσεις που αναδεικνύουν τη στατιστική σημαντικότητα ολόκληρου του μοντέλου και της κάθε ανεξάρτητης μεταβλητής. Μια μέθοδος για να ελέγξουμε τη σημαντικότητα μιας μεταβλητής σε ένα μοντέλο λογιστικής παλινδρόμησης είναι ο έλεγχος του Wald. Από την θεωρία εκτιμητικής, οι ε.μ.π. για τις παραμέτρους ενός γενικευμένου γραμμικού μοντέλου έχουν ασυμπτωτικά κανονική κατανομή. Συνεπώς, για μεγάλο αριθμό παρατηρήσεων, μπορούμε να χρησιμοποιήσουμε την κανονική κατανομή για να εξετάσουμε τον έλεγχο υποθέσεων με μορφή :

Μηδενική υπόθεση $H_0 : \beta = 0$
VS
Εναλλακτική υπόθεση $H_1 : \beta \neq 0$

Η στατιστική συνάρτηση που χρησιμοποιείται είναι η

$$Z = \frac{\hat{\beta}}{s.e.(\hat{\beta})}$$

όπου ασυμπτωτικά ακολουθεί την N(0,1). Με $\widehat{\beta}$ συμβολίζεται η εκτιμήτρια της παραμέτρου της εκάστοτε ανεξάρτητης μεταβλητής.

(Πολίτης, 2021)

```
> summary(model2)

Call:
glm(formula = Playoffs...Quarter.Finals ~ (PIR + AST + STL +
    FT.. + FC + Points + TO + X2PT..), family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3295  -0.6247  -0.2169   0.6670   2.2815

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.65909    4.71665  -1.200 0.230213
PIR          0.39397    0.05788   6.807 9.96e-12 ***
AST         -0.48137    0.09469  -5.084 3.70e-07 ***
STL         -0.30087    0.08533  -3.526 0.000422 ***
FT..        -0.08892    0.03752  -2.370 0.017803 *
FC           0.47508    0.10255   4.633 3.61e-06 ***
Points      -0.34818    0.08906  -3.910 9.24e-05 ***
TO          -0.31931    0.11440  -2.791 0.005253 **
X2PT..       0.18876    0.07101   2.658 0.007859 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 526.45  on 395  degrees of freedom
Residual deviance: 330.99  on 387  degrees of freedom
AIC: 348.99

Number of Fisher Scoring iterations: 5
```

Πίνακας 5.2 : Output για το προσαρμοσμένο μοντέλο.

Με βάση τα αποτελέσματα του ελέγχου Wald, αποδείχθηκε πως όλες οι παραπάνω μας μεταβλητές ήταν στατιστικώς σημαντικές, καθώς για όλες αυτές, οι αντίστοιχες p-values ήταν μικρότερες του επιπέδου σημαντικότητας 5%. Παράλληλα, ο αλγόριθμος μεγιστοποίησε το λογάριθμο της πιθανοφάνειας μετά από 5 επαναλήψεις, γεγονός που χαρακτηρίζεται θετικό, καθώς όσο λιγότερες είναι οι επαναλήψεις, τόσο πιο ευσταθές είναι και το μοντέλο.

Ωστόσο, στο παραπάνω πίνακα 5.2 παρατηρήσαμε πως οι συντελεστές των περισσότερων μεταβλητών του μοντέλου μας, είχαν αρνητικό πρόσημο. Το γεγονός αυτό έρχεται σε αντίθεση με τη κοινή λογική, καθώς φαίνεται παραδείγματος χάρη οι πόντοι και οι ασίστ να έχουν αρνητική επίδραση στη πρόκριση μιας ομάδας. Αυτό το αποτέλεσμα είναι πολύ πιθανό να οφείλεται στην ύπαρξη υψηλής γραμμικής σχέσης μεταξύ ορισμένων επεξηγηματικών μεταβλητών, δηλαδή στην ύπαρξη πολυσυγγραμμικότητας, την οποία εξετάσαμε παρακάτω.

Έλεγχος Πολυσυγγραμικότητας

Η πολυσυγγραμικότητα (multicollinearity) δεν επηρεάζει την πρόβλεψη της μεταβλητής απόκρισης, προκαλεί όμως σύγχυση στην εκτίμηση των συντελεστών του μοντέλου (προκύπτουν μεγάλα τυπικά σφάλματα), δηλαδή δε μπορούν να καθοριστούν οι επιδράσεις των επεξηγηματικών μεταβλητών στη μεταβλητή απόκρισης. Η

πολυσυγγραμμικότητα μπορεί να ανιχνευθεί κυρίως με τη βοήθεια της ανοχής (tolerance) και του συντελεστής πληθωρισμού διακύμανσης (VIF).

Η ανοχή (Senaviratna & A. Cooray, 2019) είναι το ποσοστό της διακύμανσης μιας επεξηγηματικής μεταβλητής, που δεν μπορεί να εξηγηθεί από τις άλλες ανεξάρτητες μεταβλητές. Εξ' ορισμού, η ανοχή οποιασδήποτε συγκεκριμένης επεξηγηματικής μεταβλητής ισούται με $1 - R_i^2$, όπου R_i^2 είναι ο συντελεστής προσδιορισμού που προκύπτει από την παλινδρόμηση των άλλων μεταβλητών πάνω στην i-οστή μεταβλητή. Τιμές ανοχής κοντά στο 1 δείχνουν ότι υπάρχει μικρή πολυσυγγραμμικότητα, ενώ μια τιμή κοντά στο μηδέν υποδηλώνει ότι η πολυσυγγραμμικότητα μπορεί να αποτελεί πρόβλημα.

Ο συντελεστής πληθωρισμού διακύμανσης (VIF) έχει τύπο

$$VIF = \frac{1}{1 - R_i^2}$$

και δείχνει πόσο διογκώνεται η διακύμανση της εκτίμησης του συντελεστή από την ύπαρξη πολυσυγγραμμικότητας. Η τετραγωνική ρίζα του VIF υποδηλώνει πόσο μεγαλύτερο είναι το τυπικό σφάλμα, σε σύγκριση με το πόσο θα ήταν, εάν αυτή η μεταβλητή δεν ήταν συσχετισμένη με τις άλλες επεξηγηματικές μεταβλητές.

Τιμές VIF που υπερβαίνουν το 10 συχνά θεωρείται ότι υποδεικνύουν πολυσυγγραμμικότητα, αλλά σε ασθενέστερα μοντέλα (κάτι που συμβαίνει συχνά στην λογιστική παλινδρόμηση) τιμές πάνω από το 5 μπορεί να είναι αιτία ανησυχίας.

Εμείς χρησιμοποιήσαμε τη συνάρτηση **vif** από το πακέτο 'car' της R προκειμένου να εξετάσουμε την ύπαρξη πολυσυγγραμικότητας στο μοντέλο που προσαρμόσαμε.

Μεταβλητές	Τιμές VIF
PIR	**10,324**
AST	**3,214**
STL	**2,250**
FT%	**1,317**
FC	**2,211**
Points	**6,572**
TO	**1,371**
2PT%	**1,664**

Πίνακας 5.3 : Τιμές VIF για τις μεταβλητές του μοντέλου.

Από το παραπάνω πίνακα, φαίνεται να υπάρχει πρόβλημα πολυσυγγραμμικότητας ανάμεσα στους πόντους και στο δείκτη PIR. Αυτό επιβεβαιώνει και το συμπέρασμα του σχήματος 3.2 του τρίτου κεφαλαίου, όπου είδαμε ότι ο δείκτης PIR φάνηκε να έχει μια ισχυρή, θετική συσχέτιση με τους πόντους.

Συνεπώς, μιας και η μεταβλητή του PIR είναι πιο «σημαντική» για την ανάλυση μας, αποφασίσαμε να αφαιρέσουμε τη μεταβλητή των πόντων, και ακολουθώντας την ίδια διαδικασία με παραπάνω, να προσαρμόσουμε ένα νέο μοντέλο. Αυτό περιλάμβανε τις μεταβλητές 2PT%, DR, TO, FD, 3PT% και AST.

```
> summary(model3)

Call:
glm(formula = Playoffs...Quarter.Finals ~ (X2PT.. + DR + TO + 
    FD + X3PT.. + AST), family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max  
-2.2842  -0.6774  -0.2408   0.6965   2.4378  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -32.34158    4.24465  -7.619 2.55e-14 ***
X2PT..        0.33536    0.05958   5.629 1.81e-08 ***
DR            0.46857    0.09806   4.778 1.77e-06 ***
TO           -0.54248    0.10823  -5.012 5.37e-07 ***
FD            0.37400    0.09780   3.824 0.000131 ***
X3PT..        0.14141    0.05136   2.753 0.005903 ** 
AST          -0.14664    0.07344  -1.997 0.045847 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 526.45  on 395  degrees of freedom
Residual deviance: 342.88  on 389  degrees of freedom
AIC: 356.88

Number of Fisher Scoring iterations: 5
```

Πίνακας 5.4 : Output για το νέο προσαρμοσμένο μοντέλο.

Με βάση τα αποτελέσματα του ελέγχου Wald, αποδείχθηκε πως όλες οι παραπάνω μεταβλητές ήταν στατιστικώς σημαντικές, καθώς για όλες αυτές, οι αντίστοιχες p-values ήταν μικρότερες του επιπέδου σημαντικότητας 5%. Παράλληλα, ο αλγόριθμος μεγιστοποίησε το λογάριθμο της πιθανοφάνειας μετά από 5 επαναλήψεις, γεγονός που χαρακτηρίζεται όπως αναφέραμε παραπάνω θετικό, καθώς όσο λιγότερες είναι οι επαναλήψεις, τόσο πιο ευσταθές είναι και το μοντέλο.

Ελέγχοντας τους δείκτες VIF, είδαμε πως δεν υπήρχε πλέον θέμα πολυσυγγραμικότητας των μεταβλητών μας.

Μεταβλητές	Τιμές VIF
2PT%	1,192
DR	1,244
TO	1,285
FD	1,331
3PT%	1,222
AST	2,019

Πίνακας 5.5 : Τιμές VIF για τις μεταβλητές του μοντέλου.

Ο έλεγχος του Wald αποτελεί μία καλή περίπτωση ελέγχου για τη στατιστική σημαντικότητα των ανεξάρτητων μεταβλητών σε ένα μοντέλο λογιστικής παλινδρόμησης, ωστόσο δεν αποτελεί την μοναδική περίπτωση. Η βασική έννοια, η οποία χρησιμοποιείται για να ελέγξουμε την καλή προσαρμογή ενός μοντέλου είναι αυτή της απόκλισης (Πολίτης, 2021). Η έννοια αυτή είναι παρόμοια με αυτή του αθροίσματος τετραγώνων των καταλοίπων που υπάρχει στα κανονικά γραμμικά μοντέλα, καθώς αποτελεί μέτρο της ανερμήνευτης μεταβλητότητας του μοντέλου. Ως απόκλιση θεωρούμε τον λογάριθμο της πιθανοφάνειας του προσαρμοσμένου μοντέλου. Διαισθητικά, όσο πιο μικρή είναι η απόκλιση ενός μοντέλου, τόσο πιο κοντά είναι στο κορεσμένο μοντέλο, και αυτό παρέχει ένδειξη καλής προσαρμογής.

Στη γενική περίπτωση, η κατανομή της απόκλισης δεν είναι γνωστή. Λόγω αυτού, αξιοποιείται ο έλεγχος του λόγου πιθανοφανειών, όπου στην ουσία αποτελεί την διαφορά της απόκλισης του κορεσμένου μοντέλου (αυτού που αποτελείται από τόσες μεταβλητές, όσες και οι παρατηρήσεις μας) και του προσαρμοσμένου μοντέλου, η οποία εκφράζεται από τον παρακάτω τύπο :

$$D_1 - D_2 = -2\left[\frac{log\, L(reduced\ model)}{log\, L\,(saturated\ model)}\right]$$

Η συγκεκριμένη ποσότητα γνωρίζουμε ότι ακολουθεί την κατανομή X_p^2, όπου ισχύει $p = df_1 - df_2$, ενώ $L()$ είναι η συνάρτηση πιθανοφάνειας.

Αξιοποιώντας τώρα την εντολή anova(), πήραμε τις μετρήσεις των αποκλίσεων για το μοντέλο, όταν προστίθεται κάθε φορά μια μεταβλητή (από αυτές που έχουν επιλεχθεί) και για τη συνολική απόκλιση.

```
> anova(model3,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Playoffs...Quarter.Finals

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     395     526.45
X2PT..  1   86.810       394     439.64 < 2.2e-16 ***
DR      1   30.632       393     409.00 3.120e-08 ***
TO      1   23.008       392     386.00 1.613e-06 ***
FD      1   33.704       391     352.29 6.415e-09 ***
X3PT..  1    5.330       390     346.96   0.02096 *
AST     1    4.077       389     342.88   0.04347 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Πίνακας 5.6 : Output για τον έλεγχο απόκλισης του μοντέλου.

Τέλος, για δίτιμα δεδομένα, προτείνεται ο έλεγχος των Hosmer - Lemeshow, για τον έλεγχο της ολικής επάρκειας ενός μοντέλου. Διατάσσουμε τις παρατηρήσεις ανάλογα με την προβλεπόμενη πιθανότητα επιτυχίας. Στη συνέχεια, χωρίζουμε τις διατεταγμένες παρατηρήσεις σε g ομάδες, με ίσο περίπου αριθμό παρατηρήσεων, και για καθεμία από αυτές καταγράφουμε τον αριθμό επιτυχιών και αποτυχιών, σχηματίζοντας έτσι έναν πίνακα διαστάσεων g×2 . Η στατιστική συνάρτηση του ελέγχου, είναι η

$$X_{HL} = \sum_{g=1}^{G} \frac{(O_g - E_g)^2}{E_g\left(1 - \frac{E_g}{n_g}\right)}$$

όπου O_g είναι οι παρατηρηθείσες τιμές, E_g είναι οι εκτιμώμενες τιμές και n_g είναι ο αριθμός των παρατηρήσεων για την g-οστή ομάδα, και G είναι ο αριθμός των ομάδων. Η στατιστική συνάρτηση ακολουθεί την κατανομή X^2_{g-2}.

Ο έλεγχος των Hosmer – Lemeshow έχει την μορφή :

Μηδενική υπόθεση H_0 *: Οι παρατηρηθείσες τιμές της Y δε διαφέρουν σημαντικά, από τις εκτιμώμενες τιμές.*

VS

Εναλλακτική υπόθεση H_1 *: όχι H_0*

Γενικά, δεν θέλουμε να απορρίπτεται η μηδενική υπόθεση, ώστε το μοντέλο που προσαρμόζουμε να είναι επαρκές.

```
> y <- Playoffs...Quarter.Finals
> hl.test <- hoslem.test(model3$y, fitted(model3))
> hl.test

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  model3$y, fitted(model3)
X-squared = 8.8293, df = 8, p-value = 0.3569
```

Πίνακας 5.7 : Output για τον έλεγχο Hosmer – Lemeshow.

Ο έλεγχος των Hosmer-Lemeshow έδειξε ότι η μηδενική υπόθεση δεν απορρίπτεται, καθώς το p-value ήταν ίσο με 0,3569, μεγαλύτερο από το επίπεδο σημαντικότητας μας 5%. Συνεπώς, έχουμε και σε αυτή την περίπτωση μία πολύ ισχυρή ένδειξη καλής προσαρμογής του μοντέλου μας.

Στη συνέχεια, αναπαραστήσαμε τα κατάλοιπα απόκλισης (deviance residuals), που είναι αυτά που χρησιμοποιούνται ευρέως στη λογιστική παλινδρόμηση. Τα κατάλοιπα απόκλισης είναι οι τετραγωνικές ρίζες των προσθετέων στο άθροισμα

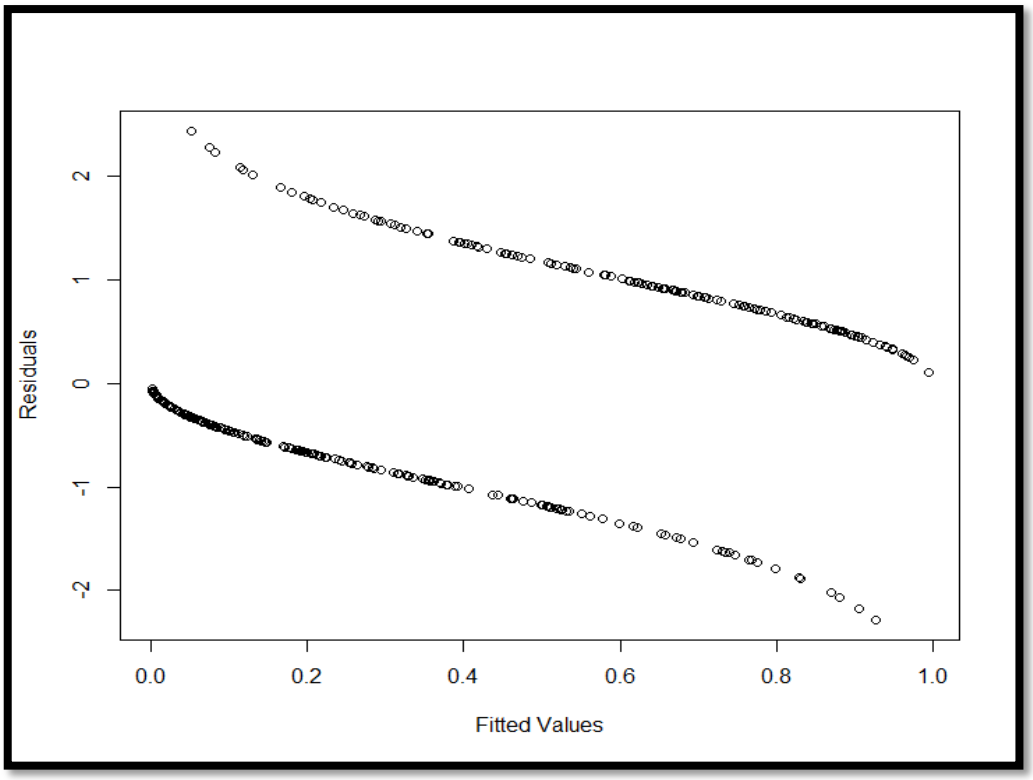
$$-2\sum\left[y_i \log\left(\frac{\hat{p}_i}{y_i}\right) + (1-y_i)\log\left(\frac{1-\hat{p}_i}{1-y_i}\right)\right]$$

όπου σε κάθε τετραγωνική ρίζα βάζουμε το ίδιο πρόσημο με αυτό που υπάρχει στο αντίστοιχο response residual. Τα κατάλοιπα αυτά ονομάζονται κατάλοιπα απόκρισης και ισούνται, σε ένα μοντέλο με δίτιμα δεδομένα, με

$$e_i = y_i - \hat{y}_i = y_i - \hat{p}_i$$

αφού η προβλεπόμενη τιμή $(\hat{y}_i)$ για κάθε παρατήρηση (y_i), είναι η εκτιμώμενη πιθανότητα επιτυχίας $(\hat{p}_i)$.

Το διάγραμμα των καταλοίπων απόκλισης έχει μορφή δύο καμπυλών, με τη μία να αντιπροσωπεύει τις θετικές τιμές καταλοίπων, που είναι και οι προβλέψεις για την τιμή 1 της μεταβλητής απόκρισης και την δεύτερη να έχει αρνητικές τιμές, που αντιπροσωπεύει τις προβλέψεις για την τιμή 0. Το παρακάτω διάγραμμα έχει αυτή τη μορφή, και συμπεραίνουμε συνεπώς ότι υπάρχει καλή προσαρμογή του μοντέλου.



Σχήμα 5.2 : Κατάλοιπα απόκλισης του μοντέλου λογιστικής παλινδρόμησης που προσαρμόστηκε.

Ερμηνεία των συντελεστών των παραμέτρων

Το μοντέλο που προσαρμόσαμε, με βάση το πίνακα 5.4, έχει τη μορφή :

$$log\left(\frac{\pi}{1-\pi}\right) = -32{,}342 + 0{,}335 \cdot 2PT\% + 0{,}469 \cdot DR - 0{,}542 \cdot TO + 0{,}374 \cdot FD + 0{,}141 \cdot 3PT\% - 0{,}147 \cdot AST$$

Απαλείφοντας τον λογάριθμο, έχουμε ότι η εξίσωση παίρνει την μορφή :

$$\frac{\pi}{1-\pi} = exp(-32{,}342 + 0{,}335 \cdot 2PT\% + 0{,}469 \cdot DR - 0{,}542 \cdot TO + 0{,}374 \cdot FD + 0{,}141 \cdot 3PT\% - 0{,}147 \cdot AST)$$

Συνεπώς για το σχετικό λόγο πιθανοτήτων προκύπτουν τα εξής ενδιαφέροντα συμπεράσματα :

- Ο σχετικός λόγος πιθανοτήτων για τη μεταβλητή 2PT% ισούται με $e^{0{,}335} \approx 1{,}398$. Αυτό σημαίνει πως για κάθε αύξηση του ποσοστού των εύστοχων

δίποντων κατά μια μονάδα, ο σχετικός λόγος πιθανοτήτων για τη πρόκριση αυξάνεται κατά περίπου 39,8% (1,398 – 1 = 0,398).

- Ο σχετικός λόγος πιθανοτήτων για τη μεταβλητή DR ισούται με $e^{0,469} \approx 1,598$. Αυτό σημαίνει πως για κάθε αύξηση των αμυντικών ριμπάουντ κατά ένα, ο σχετικός λόγος πιθανοτήτων για τη πρόκριση αυξάνεται κατά περίπου 59,8% (1,598 – 1 = 0,598).

- Ο σχετικός λόγος πιθανοτήτων για τη μεταβλητή TO ισούται με $e^{-0,542} \approx 0,581$. Αυτό σημαίνει πως για κάθε αύξηση του πλήθους των λαθών κατά ένα, ο σχετικός λόγος πιθανοτήτων για τη πρόκριση μειώνεται κατά περίπου 41,9% (1 – 0,581 = 0,419).

- Ο σχετικός λόγος πιθανοτήτων για τη μεταβλητή FD ισούται με $e^{0,374} \approx 1,454$. Αυτό σημαίνει πως για κάθε αύξηση των φάουλ τα οποία κερδίζει μια ομάδα κατά ένα, ο σχετικός λόγος πιθανοτήτων για τη πρόκριση αυξάνεται κατά περίπου 45,4% (1,454 – 1 = 0,454).

- Ο σχετικός λόγος πιθανοτήτων για τη μεταβλητή 3PT% ισούται με $e^{0,141} \approx 1,152$. Αυτό σημαίνει πως για κάθε αύξηση του ποσοστού των εύστοχων τρίποντων κατά μια μονάδα, ο σχετικός λόγος πιθανοτήτων για τη πρόκριση αυξάνεται κατά περίπου 15,2% (1,152 – 1 = 0,152).

- Ο σχετικός λόγος πιθανοτήτων για τη μεταβλητή AST ισούται με $e^{-0,147} \approx 0,864$. Αυτό σημαίνει πως για κάθε αύξηση των ασίστ κατά μια, ο σχετικός λόγος πιθανοτήτων για τη πρόκριση μειώνεται κατά περίπου 13,6% (1 – 0,864 = 0,136).

Εναλλακτικά :

- Η αύξηση κατά 1% του ποσοστού εύστοχων δίποντων, αυξάνει την σχετική πιθανότητα πρόκρισης πολλαπλασιαστικά κατά $e^{0,335} \approx 1,398$ φορές.

- Η αύξηση των λαθών κατά μια μονάδα, μειώνει την σχετική πιθανότητα πρόκρισης πολλαπλασιαστικά κατά $e^{-0,542} \approx 0,581$ φορές.

Με την ίδια λογική ερμηνεύονται και οι υπόλοιπες μεταβλητές.

Μέτρα Προσαρμογής

Στην κλασσική παλινδρόμηση με κανονικές αποκρίσεις, ένα αριθμητικό μέτρο που χρησιμοποιείται για την αξιολόγηση της προσαρμογής ενός μοντέλου είναι ο συντελεστής προσαρμογής :

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

ο οποίος ουσιαστικά είναι το ποσοστό της μεταβλητότητας των Y_i που «ερμηνεύεται» από το μοντέλο.

Πολλές φορές χρησιμοποιείται και ο προσαρμοσμένος συντελεστής :

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

που προκύπτει διαιρώντας καθένα από τα SSE, SST με τους βαθμούς ελευθερίας τους.

Για την αξιολόγηση ενός λογιστικού μοντέλου (ή, γενικότερα, ενός ΓΓΜ) δεν υπάρχει ένα γενικά αποδεκτό μέτρο, αντίστοιχο του R^2. Υπάρχουν ωστόσο διάφορα μέτρα που έχουν προταθεί, τα οποία συνήθως είναι γνωστά ως pseudo R^2. Έστω L_M είναι η μέγιστη πιθανοφάνεια υπό το μοντέλο M και L_0 είναι η μέγιστη πιθανοφάνεια του μοντέλου μόνο με το σταθερό όρο. Τα κυριότερα μέτρα προσαρμογής είναι τα ακόλουθα.

1. McFadden

Είναι το πιο δημοφιλές pseudo R^2. Τιμές μεταξύ 0,2 και 0,4 υποδηλώνουν καλή προσαρμογή του μοντέλου μας. Ο τύπος είναι

$$1 - \frac{log\, L_M}{log\, L_0}$$

2. McFadden adjusted

Μία παραλλαγή του παραπάνω, η οποία θέτει ποινή για το πλήθος k των παραμέτρων, και έχει με τύπο

$$1 - \frac{log\, L_M - k}{log\, L_0}$$

3. Cox & Snell

Ο λόγος των πιθανοφανειών δείχνει τη βελτίωση του μοντέλου, σε σχέση με αυτό που έχει μόνο το σταθερό όρο. Όσο μικρότερος είναι ο λόγος, τόσο μεγαλύτερη βελτίωση θα υπάρχει. Ο τύπος είναι

$$1 - (L_0 \,/\, L_M)^{2/n}$$

4. Nagelkerke Cragg & Uhler

Τροποποίηση του παραπάνω, όταν διαιρεθεί με τη μέγιστη τιμή του. Προσαρμόζεται έτσι ώστε να παίρνει τη τιμή 1, όταν έχουμε τέλεια προσαρμογή. Ο τύπος του είναι

$$1 - \frac{(L_0 \,/\, L_M)^{2/n}}{1 - L_0^{2/n}}$$

5. AIC (Akaike Information Criterion)

Το κριτήριο πληροφορίας του Akaike είναι αρκετά δημοφιλές κριτήριο επιλογής μοντέλου, όχι μόνο για δίτιμες αποκρίσεις. Επιβάλλει ποινή για το πλήθος k των παραμέτρων. Όσο μικρότερο τόσο καλύτερο το προσαρμοσμένο μοντέλο. Δίνεται από τον τύπο

$$AIC = -2\,log\,L_M + 2k$$

6. BIC (Bayesian Information Criterion)

Είναι γνωστό και ως κριτήριο πληροφορίας του Schawrz. Όπως και στο AIC, επιλέγουμε την ελάχιστη τιμή του. Δίνεται από τον τύπο

$$BIC = -2\,log\,LM + k(log\,n)$$

όπου n είναι το πλήθος των παρατηρήσεων.

Το AIC αντανακλά τον κίνδυνο ένα μοντέλο να υπερπροσαρμοστεί (overfitting), ενώ το BIC να υποπροσαρμοστεί (underfitting). (Πολίτης, 2021)

Παρουσιάζονται στο παρακάτω πίνακα τα μέτρα προσαρμογής για το μοντέλο που προσαρμόσαμε.

Μέτρα Προσαρμογής	Τιμές
McFadden	**0,349**
McFadden adjusted	**0,322**
Cox & Snell	**0,371**

Nagelkerke Cragg & Uhler	0,504
AIC	356,883
BIC	384,753

Πίνακας 5.8 : Μέτρα προσαρμογής για το προσαρμοσμένο μοντέλο.

Παρατηρούμε ότι το pseudo R^2 του McFadden είναι εντός του διαστήματος (0,2 , 0,4) κάτι το οποίο δείχνει καλή προσαρμογή για το μοντέλο μας. Διαισθητικά, αυτό το αντιλαμβανόμαστε καλύτερα με το τέταρτο pseudo R^2, όπου είναι προσαρμοσμένο για να παίρνει την τιμή 1 (όπως αναφέραμε) για την τέλεια προσαρμογή.

Καμπύλη ROC (Receiver Operating Characteristic curve)

Αρχικά, πρέπει να ορίσουμε δύο νέους όρους, αναφορικά με την ταξινόμηση τιμών, προκειμένου να σχολιάσουμε τη καμπύλη ROC. Με τον όρο ευαισθησία (sensitivity) του μοντέλου εννοούμε το ποσοστό επιτυχιών που «ταξινομούνται» σωστά, και με τον όρο ειδικότητα (specificity) εννοούμε το ποσοστό των αποτυχιών που «ταξινομούνται» σωστά.

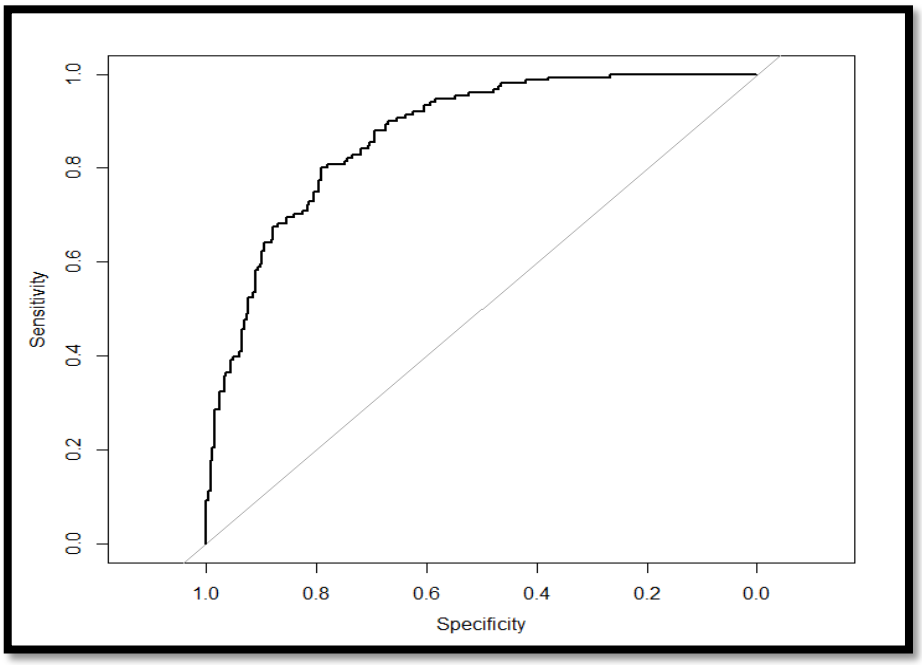
Η καμπύλη ROC είναι η γραφική παράσταση της ευαισθησίας έναντι της ποσότητας (1-ειδικότητα), των οποίων οι τιμές μεταβάλλονται για κάθε τιμή της διαχωριστικής τιμής c. Οι καμπύλες ROC εμφανίζονται στα διαγνωστικά τεστ, όταν χρησιμοποιείται κάποια μέτρηση για να κάνουμε τη διάγνωση κάποιας νόσου και τίθεται θέμα της «καλύτερης» τιμής της μέτρησης cut off για να γίνει η διάγνωση. Η επάνω αριστερή γωνία θεωρείται η βέλτιστη θέση σε ένα γράφημα ROC, δείχνοντας ένα υψηλό ποσοστό αληθώς θετικών και ένα χαμηλό ποσοστό ψευδώς θετικών.

Με άλλα λόγια, η καμπύλη ROC είναι η γραφική παράσταση της συνάρτησης ευαισθησία(c) = 1-ειδικότητα(c) , για διάφορες τιμές του threshold c .

Η περιοχή κάτω από την καμπύλη ROC (**A**rea **U**nder the **C**urve – AUC) αποτελεί ένα γενικό μέτρο της ακρίβειας της λανθασμένης ταξινόμησης, αντιπροσωπεύοντας το γενικό ποσοστό των ορθώς ταξινομημένων τιμών. Εκτιμάται με παραμετρικές και μη μεθόδους, και εκτιμά την πιθανότητα να είναι ίσες οι προβλέψεις με τις πραγματικές παρατηρήσεις. Όσο μεγαλύτερος είναι ο δείκτης αυτός, τόσο μεγαλύτερη θεωρείται η προβλεπτική ικανότητα του μοντέλου (AUC $\geq$ 0,8 για ένα καλό μοντέλο). Αν ισούται με 0,50 τότε το μοντέλο προβλέπει τελείως τυχαία. Το μέτρο AUC είναι ιδιαίτερα χρήσιμο για τα σύνολα δεδομένων με μη ισορροπημένη κατανομή της μεταβλητής απόκρισης (η μία κλάση κυριαρχεί έναντι της άλλης). Η τιμή 1 αντιπροσωπεύει την πλήρη ακρίβεια.

Για την εύρεση του βέλτιστου σημείου αποκοπής, χρησιμοποιείται το κριτήριο μεγιστοποίησης της συνάρτσησης h(c) = ευαισθησία + ειδικότητα - 1, το οποίο είναι γνωστό ως Δείκτης του Youden. (Μπερσίμης, 2022)

Είναι φανερό πως οι καμπύλες ROC χρησιμοποιούνται κυρίως για τα προβλήματα ταξινόμησης. Εμείς θα τις χρησιμοποιήσουμε ως ένα καλό και εποπτικό μέσο αξιολόγησης των μοντέλων που προσαρμόσαμε.



Σχήμα 5.3 : Καμπύλη ROC για το προσαρμοσμένο μοντέλο.

Το μοντέλο κρίνεται ιδιαίτερα επιτυχημένο, καθώς η καμπύλη ROC απέχει πολύ από την διαγώνιο και το εμβαδόν κάτω από την καμπύλη ισούται με 0,87. Η τιμή πλησιάζει στο 1, που σημαίνει ότι η συνολική ακρίβεια του μοντέλου είναι υψηλή και κατά συνέπεια, υψηλές θα είναι και οι τιμές της ευαισθησίας και της ειδικότητας. Συνεπώς, είναι καλύτερο από ένα τυχαίο μοντέλο (με εμβαδόν ίσο με 0,5) και επομένως οι προβλέψεις του μοντέλου με τις πραγματικές παρατηρήσεις είναι πολύ κοντά. Τέλος, το cutoff σημείο που είναι πιο κοντά στην επάνω αριστερή γωνία έχει τιμή 0,393469 και θεωρείται το βέλτιστο όσον αφορά τη διάκριση μεταξύ των ομάδων που προκρίθηκαν στα playoffs και αυτών που δεν προκρίθηκαν.

5.4.2 Προσαρμογή μοντέλου με αλληλεπιδράσεις 2ης τάξης

Μέχρι εδώ, προσαρμόσαμε ένα μοντέλο, με γνώμονα τη βασική θεωρία που διέπει τα ΓΓΜ, το οποίο θα μας βοηθήσει να κατανοήσουμε ποιοι είναι οι βασικοί παράγοντες που επηρεάζουν το αν μια ομάδα προκρίνεται ή όχι στη φάση των Playoffs. Ωστόσο, οι παράγοντες που απαρτίζουν το μοντέλο δεν αφορούν απλά και μόνο τις ανεξάρτητες μεταβλητές που επιλέχθηκαν με βάση τη στατιστική τους σημαντικότητα. Βασικό πεδίο για την ακόμα καλύτερη προσαρμογή του μοντέλου παίζουν και οι παράγοντες αλληλεπίδρασης.

Γενικά, λέμε ότι δύο μεταβλητές Α και Β αλληλοεπιδρούν ως προς μία τρίτη μεταβλητή Υ όταν η επίδραση της μιας μεταβλητής (Α) στην Υ εξαρτάται από το επίπεδο της άλλης μεταβλητής (Β). Μία βασική αρχή είναι ότι αν συμπεριλάβουμε τον όρο αλληλεπίδρασης Α:Β στο μοντέλο, τότε θα πρέπει να συμπεριληφθούν και οι κύριες επιδράσεις των Α και Β, ακόμα και αν αυτές δεν είναι στατιστικά σημαντικές. (Πολίτης, 2021)

Προσπαθήσαμε να προσαρμόσουμε ένα μοντέλο με αλληλεπιδράσεις 2ης τάξης, ώστε να είναι πιο εύκολη η ερμηνεία των συντελεστών. Παρ' όλα αυτά, μέσω της ίδιας διαδικασίας με πριν (χρήση της συνάρτησης step και δοκιμές πολλών μοντέλων), δε μπορέσαμε να καταλήξουμε σε ένα τέτοιο ικανοποιητικό μοντέλο, καθώς όλες οι πιθανές μεταβλητές αλληλεπιδράσεων 2ης τάξης έβγαιναν στατιστικώς μη σημαντικές.

5.4.3 Μοντέλο για τη φάση των Semi-Finals / Final Four

Σε αυτό το σημείο, ακολουθήσαμε ακριβώς τις ίδιες διαδικασίες με τις παραπάνω παραγράφους, με τη βασική διαφορά να είναι στο γεγονός πως ως μεταβλητή απόκρισης πήραμε τώρα τη κατηγορική μεταβλητή " Final 4 / Semi-Finals ", δηλαδή εξετάσαμε τώρα τη πρόκριση (ή όχι) στο Final Four.

Μέσω της συνάρτησης **step**, πήραμε πως το κατάλληλο μοντέλο ήταν αυτό με ανεξάρτητες μεταβλητές τις PIR, Points, FC, AST, 3PT%, BLKA, 2PT%, TO και OR. Όλες οι επεξηγηματικές μεταβλητές ήταν στατιστικώς σημαντικές, αφού τα p-values τους για τον έλεγχο του Wald ήταν όλα μικρότερα του επιπέδου σημαντικότητας. Επιπροσθέτως, ο αλγόριθμος μεγιστοποίησε το λογάριθμο της πιθανοφάνειας μετά από 7 επαναλήψεις.

```
> summary(model6)

Call:
glm(formula = Final.4...Semi.Finals ~ (PIR + Points + FC + AST +
    X3PT.. + BLKA + X2PT.. + TO + OR), family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9278  -0.4246  -0.1378  -0.0195   3.5369

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -26.71392    6.78974  -3.934 8.34e-05 ***
PIR           0.53534    0.08763   6.109 1.00e-09 ***
Points       -0.75362    0.14848  -5.076 3.86e-07 ***
FC            0.56924    0.14629   3.891 9.98e-05 ***
AST          -0.38145    0.10249  -3.722 0.000198 ***
X3PT..        0.29635    0.10472   2.830 0.004655 **
BLKA          0.84818    0.42096   2.015 0.043917 *
X2PT..        0.39209    0.12428   3.155 0.001605 **
TO           -0.49052    0.17902  -2.740 0.006143 **
OR            0.50389    0.20813   2.421 0.015476 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 375.52  on 395  degrees of freedom
Residual deviance: 202.66  on 386  degrees of freedom
AIC: 222.66

Number of Fisher Scoring iterations: 7
```

Πίνακας 5.9 : Output για το προσαρμοσμένο μοντέλο.

Ωστόσο, παρατηρήσαμε πως οι συντελεστές για τους πόντους και τις ασίστ ήταν αρνητικοί ξανά, και αυτό ερχόταν σε αντίθεση με τη «κοινή λογική». Ελέγχοντας για την ύπαρξη πολυσυγγραμικότητας, παρατηρήσαμε πάλι υψηλή σχέση ανάμεσα στους πόντους και στο δείκτη PIR, κάτι που ήταν και αναμενόμενο. Πραγματοποιώντας την ίδια διαδικασία με την ανάλυση για τη φάση των Playoffs, αφαιρέσαμε τη μεταβλητή Points και ξανά προσαρμόσαμε ένα μοντέλο. Το νέο μοντέλο περιλάμβανε τις μεταβλητές FD, TO, 2PT%, DR και 3PT%.

```
> summary(model8)

Call:
glm(formula = Final.4...Semi.Finals ~ (FD + TO + X2PT.. + DR + 
    X3PT..), family = binomial(link = logit))

Deviance Residuals:
     Min        1Q    Median        3Q       Max  
-1.74577  -0.45674  -0.15983  -0.03898   2.76840  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -44.58929    6.40699  -6.959 3.42e-12 ***
FD            0.53807    0.11766   4.573 4.80e-06 ***
TO           -0.58661    0.14751  -3.977 6.99e-05 ***
X2PT..        0.34126    0.07245   4.710 2.47e-06 ***
DR            0.55105    0.12894   4.274 1.92e-05 ***
X3PT..        0.22667    0.06919   3.276  0.00105 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 375.52  on 395  degrees of freedom
Residual deviance: 224.19  on 390  degrees of freedom
AIC: 236.19

Number of Fisher Scoring iterations: 7
```

Πίνακας 5.10 : Output για το προσαρμοσμένο μοντέλο.

Παρατηρήσαμε πως όλες οι επεξηγηματικές μας μεταβλητές ήταν στατιστικώς σημαντικές, καθώς για όλες αυτές, οι αντίστοιχες p-values ήταν μικρότερες του επιπέδου σημαντικότητας 5%. Παράλληλα, ο αλγόριθμος μεγιστοποίησε το λογάριθμο της πιθανοφάνειας μετά από 7 επαναλήψεις.

```
> anova(model8,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Final.4...Semi.Finals

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     395     375.52
FD      1   24.994       394     350.52 5.752e-07 ***
TO      1   41.797       393     308.73 1.012e-10 ***
X2PT..  1   51.115       392     257.61 8.712e-13 ***
DR      1   21.500       391     236.11 3.538e-06 ***
X3PT..  1   11.924       390     224.19 0.0005541 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Πίνακας 5.11 : Output για τον έλεγχο απόκλισης του μοντέλου.

Όλες οι μεταβλητές που προστίθενται στο μοντέλο, προσφέρουν κάτι σημαντικό, αναφορικά με την απόκλιση στο νέο μας προσαρμοσμένο μοντέλο.

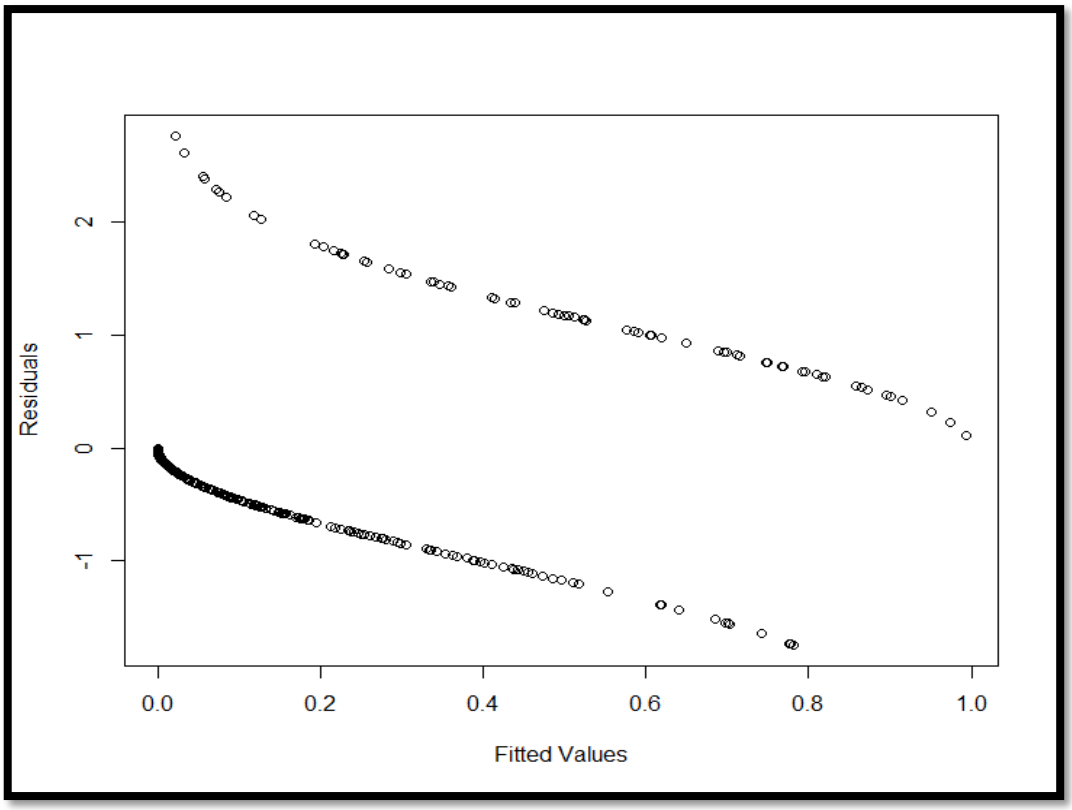
```
> y <- Final.4...Semi.Finals
> hl.test <- hoslem.test(model8$y, fitted(model8))
> hl.test

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  model8$y, fitted(model8)
X-squared = 4.8793, df = 8, p-value = 0.7704
```

Πίνακας 5.12 : Output για τον έλεγχο Hosmer – Lemeshow.

Ο έλεγχος των Hosmer-Lemenshow έδειξε ότι η μηδενική υπόθεση δεν απορρίπτεται, καθώς το p-value ήταν ίσο με 0,7704 και άρα μεγαλύτερο από το επίπεδο σημαντικότητας μας 5%. Συνεπώς, έχουμε και σε αυτή την περίπτωση μία πολύ ισχυρή ένδειξη καλής προσαρμογής του μοντέλου μας.



Σχήμα 5.4 : Κατάλοιπα απόκλισης του μοντέλου λογιστικής παλινδρόμησης που προσαρμόστηκε.

Το διάγραμμα των καταλοίπων απόκλισης έχει και σε αυτή τη περίπτωση τη μορφή που επιθυμούμε, ώστε να συμπεράνουμε υπέρ της καλής προσαρμογή του μοντέλου.

Ερμηνεία των συντελεστών των παραμέτρων

Το μοντέλο που προσαρμόσαμε, με βάση το πίνακα 5.10, έχει τη μορφή :

$$log\left(\frac{\pi}{1-\pi}\right) = -44{,}589 + 0{,}538 \cdot FD - 0{,}587 \cdot TO + 0{,}341 \cdot 2PT\% + 0{,}551 \cdot DR + 0{,}227 \cdot 3PT\%$$

Απαλείφοντας τον λογάριθμο, έχουμε ότι η εξίσωση παίρνει την μορφή :

$$\frac{\pi}{1-\pi} = e^{-44{,}589 + 0{,}538 \cdot FD - 0{,}587 \cdot TO + 0{,}341 \cdot 2PT\% + 0{,}551 \cdot DR + 0{,}227 \cdot 3PT\%}$$

Συνεπώς για την σχετικό λόγο πιθανοτήτων προκύπτουν τα εξής ενδιαφέροντα συμπεράσματα :

- Ο σχετικός λόγος πιθανοτήτων για τη μεταβλητή FD ισούται με $e^{0,538} \approx 1,713$. Αυτό σημαίνει πως για κάθε αύξηση των φάουλ τα οποία κερδίζει μια ομάδα κατά ένα, ο σχετικός λόγος πιθανοτήτων για τη πρόκριση αυξάνεται κατά περίπου 71,3% (1,713 – 1 = 0,713).

- Ο σχετικός λόγος πιθανοτήτων για τη μεταβλητή TO ισούται με $e^{-0,587} \approx 0,556$. Αυτό σημαίνει πως για κάθε αύξηση του πλήθους των λαθών κατά ένα, ο σχετικός λόγος πιθανοτήτων για τη πρόκριση μειώνεται κατά περίπου 44,4% (1 – 0,556 = 0,444).

- Ο σχετικός λόγος πιθανοτήτων για τη μεταβλητή 2PT% ισούται με $e^{0,341} \approx 1,407$. Αυτό σημαίνει πως για κάθε αύξηση του ποσοστού των εύστοχων δίποντων κατά μια μονάδα, ο σχετικός λόγος πιθανοτήτων για τη πρόκριση αυξάνεται κατά περίπου 40,7% (1,407 – 1 = 0,407).

- Ο σχετικός λόγος πιθανοτήτων για τη μεταβλητή DR ισούται με $e^{0,551} \approx 1,735$. Αυτό σημαίνει πως για κάθε αύξηση των αμυντικών ριμπάουντ κατά ένα, ο σχετικός λόγος πιθανοτήτων για τη πρόκριση αυξάνεται κατά περίπου 73,5% (1,735 – 1 = 0,735).

- Ο σχετικός λόγος πιθανοτήτων για τη μεταβλητή 3PT% ισούται με $e^{0,227} \approx 1,254$. Αυτό σημαίνει πως για κάθε αύξηση του ποσοστού των εύστοχων δίποντων κατά μια μονάδα, ο σχετικός λόγος πιθανοτήτων για τη πρόκριση αυξάνεται κατά περίπου 25,4% (1,254 – 1 = 0,254).

Εναλλακτικά :

- Η αύξηση των φάουλ που κέρδισε μια ομάδα κατά μια μονάδα, αυξάνει την σχετική πιθανότητα πρόκρισης πολλαπλασιαστικά κατά $e^{0,538} \approx 1,713$ φορές.

- Η αύξηση των λαθών κατά μια μονάδα, μειώνει την σχετική πιθανότητα πρόκρισης πολλαπλασιαστικά κατά $e^{-0,587} \approx 0,556$ φορές.

Με την ίδια λογική πάλι ερμηνεύονται οι υπόλοιπες μεταβλητές.

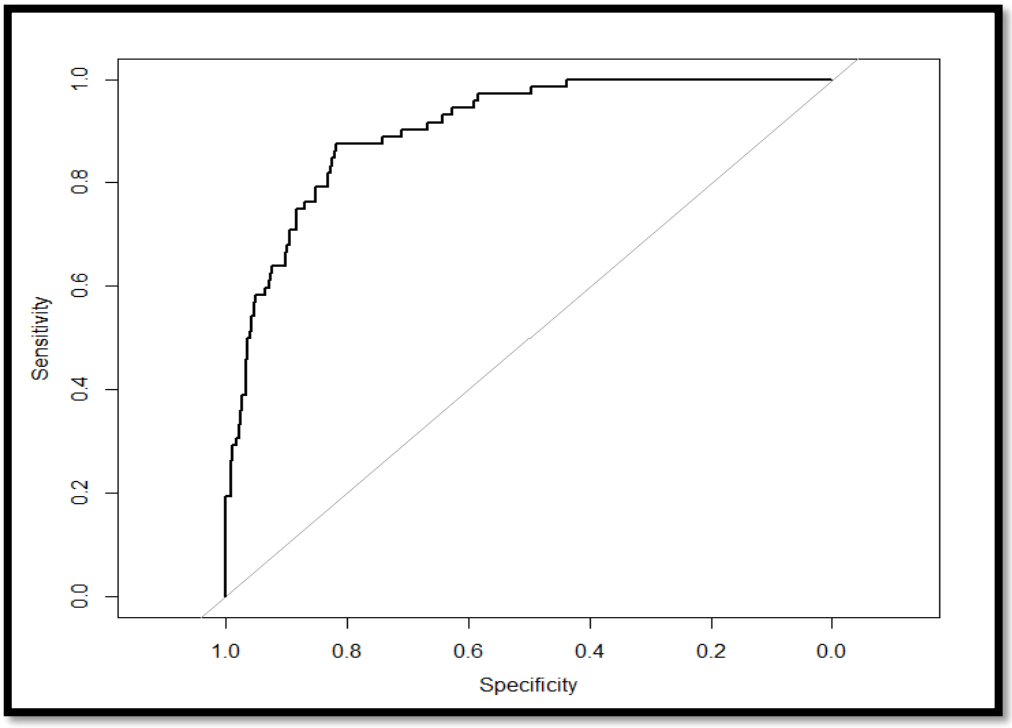
Μέτρα Προσαρμογής

Μέτρα Προσαρμογής	Τιμές
McFadden	0,403
McFadden adjusted	0,371
Cox & Snell	0,318
Nagelkerke Cragg & Uhler	0,518
AIC	236,188
BIC	260,077

Πίνακας 5.13 : Μέτρα προσαρμογής για το προσαρμοσμένο μοντέλο.

Παρατηρήσαμε πως είχαμε ένα καλά προσαρμοσμένο μοντέλο, με βάση τα παραπάνω μέτρα προσαρμογής. Το pseudo R^2 του McFadden είναι λίγο μεγαλύτερο του 0,4 , κάτι το οποίο όντως δείχνει καλή προσαρμογή για το μοντέλο μας.

Καμπύλη ROC



Σχήμα 5.5 : Καμπύλη ROC για το προσαρμοσμένο μοντέλο.

Η καμπύλη ROC απέχει αρκετά από την διαγώνιο και το εμβαδόν κάτω από την καμπύλη ισούται με 0,9039. Η τιμή είναι αρκετά κοντά στο 1, που σημαίνει ότι η συνολική ακρίβεια του μοντέλου είναι πολύ υψηλή και κατά συνέπεια, υψηλές θα είναι και οι τιμές της ευαισθησίας και της ειδικότητας. Συνεπώς, είναι πολύ καλύτερο από

ένα τυχαίο μοντέλο (με εμβαδόν ίσο με 0,5) και επομένως οι προβλέψεις του μοντέλου με τις πραγματικές παρατηρήσεις είναι πολύ κοντά. Τέλος, το cutoff σημείο που είναι πιο κοντά στην επάνω αριστερή γωνία έχει τιμή 0,1900682 και θεωρείται το βέλτιστο όσον αφορά τη διάκριση μεταξύ των ομάδων που προκρίθηκαν στα Final Four και αυτών που δεν προκρίθηκαν.

5.4.4 Προσαρμογή μοντέλου με αλληλεπιδράσεις 2ης τάξης

Προσαρμόσαμε και πάλι ένα μοντέλο με αλληλεπιδράσεις 2ης τάξης. Εργαζόμενοι όπως και πριν (χρήση της συνάρτησης **step**), πήραμε σαν κατάλληλο μοντέλο αυτό με επεξηγηματικές μεταβλητές τις 2PT%, 3PT%, DR, FD, TO καθώς και τις αλληλεπιδράσεις 3PT% × FD και 2PT% × 3PT% .

```
> summary(model_inter2)

Call:
glm(formula = Final.4...Semi.Finals ~ X2PT.. + X3PT.. + DR +
    FD + TO + X3PT..:FD + X2PT..:X3PT.., family = binomial(link = logit))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.07690  -0.44994  -0.16384  -0.03831   2.79429

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   5.08116   67.38776   0.075   0.9399
X2PT..       -1.80469    1.01611  -1.776   0.0757 .
X3PT..       -1.15865    1.83326  -0.632   0.5274
DR            0.56922    0.13280   4.286 1.82e-05 ***
FD            3.40985    1.35628   2.514   0.0119 *
TO           -0.62484    0.15151  -4.124 3.72e-05 ***
X3PT..:FD    -0.07753    0.03636  -2.132   0.0330 *
X2PT..:X3PT..  0.05873    0.02782   2.111   0.0348 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 375.52  on 395  degrees of freedom
Residual deviance: 213.18  on 388  degrees of freedom
AIC: 229.18

Number of Fisher Scoring iterations: 7
```

Πίνακας 5.14 : Output για το προσαρμοσμένο μοντέλο με αλληλεπιδράσεις.

Σύμφωνα και εδώ με τον έλεγχο του Wald, οι μεταβλητές 2PT% και 3PT% δεν ήταν στατιστικώς σημαντικές (αντίστοιχα p-values μικρότερα του ε.σ. 5%), και συνεπώς δεν θα έπρεπε να μπουν στο μοντέλο μας. Όμως, η αντίστοιχη μεταβλητή αλληλεπίδρασης τους ήταν στατιστικώς σημαντική, και συνεπώς, όπως αναφέραμε νωρίτερα, έπρεπε να συμπεριληφθούν στο μοντέλο μας.

Παρατηρήσαμε ωστόσο για άλλη μια φορά, αρνητικούς συντελεστές για τις επεξηγηματικές μεταβλητές 2PT% και 3PT% , το οποίο μας προβλημάτισε για την ύπαρξη πολυσυγγραμικότητας. Κοιτώντας πάλι τις τιμές VIF, πήραμε το παρακάτω πίνακα.

Μεταβλητές	Τιμές VIF
2PT%	207,301
3PT%	923,130
DR	1,103
FD	166,784
TO	1,289
3PT% × FD	288,834
2PT% × 3PT%	1005,667

Πίνακας 5.15 : Τιμές VIF για τις μεταβλητές του μοντέλου.

Από το παραπάνω πίνακα, φαίνεται να υπάρχει μεγάλο πρόβλημα πολυσυγρραμμικότητας ανάμεσα στις μεταβλητές για τα ποσοστά των εύστοχων δίποντων και τρίποντων, όπως και με την μεταβλητή για την αλληλεπίδραση τους. Το παραπάνω οφείλεται ειδικότερα στην ύπαρξη δομικής πολυσυγγραμικότητας (structural multicollinearity). Η δομική πολυσυγγραμμικότητα (Hair Jr. et al., 2019, p.96) αναφέρεται στο φαινόμενο όπου δύο ή περισσότερες ανεξάρτητες μεταβλητές συσχετίζονται σε μεγάλο βαθμό, επειδή «μετρούν» την ίδια υποκείμενη έννοια. Μπορεί να οδηγήσει σε μεροληπτικές εκτιμήσεις, ανακριβείς προβλέψεις, ενώ μπορεί να είναι πρόβλημα στην μοντελοποίηση μας, όταν ο στόχος είναι να εκτιμηθούν τα μοναδικά αποτελέσματα κάθε ανεξάρτητης μεταβλητής στην εξαρτημένη μεταβλητή. Οι αιτίες της δομικής πολυσυγγραμμικότητας περιλαμβάνουν τη χρήση πολλαπλών μέτρων της ίδιας δομής/έννοιας, συμπεριλαμβανομένων μεταβλητών που είναι εννοιολογικά περιττές, και μεταβλητών που μετρούν διαφορετικές πτυχές της ίδιας έννοιας. Εδώ, αυτό είναι πολύ πιθανό να συμβαίνει, καθώς οι μεταβλητές 2PT% και 3PT% θα μπορούσε να πει κανείς πως όντως μετράνε διαφορετικές πτυχές της ίδιας έννοιας, δηλαδή των εύστοχων καλαθιών/σουτ μιας ομάδας.

Η δομική πολυσυγγραμμικότητα στα μοντέλα παλινδρόμησης αντιμετωπίζεται καλά με «κεντρικοποίηση» (centering) των δεδομένων στη διάμεσο τους (Kraemer & Blasey, 2004), (Glantz & Slinker, 2001). Πρόκειται για μια διαδικασία μετασχηματισμού των μεταβλητών, κατά την οποία από τη κάθε τιμή της κάθε μεταβλητής, αφαιρείται η διάμεσος τιμή της. Αυτή τη διαδικασία εφαρμόσαμε και εμείς, και προσαρμόσαμε ένα νέο μοντέλο με αλληλεπιδράσεις $2^{ου}$ βαθμού, χρησιμοποιώντας τα νέα κεντρικοποιημένα μας δεδομένα.

Καταλήξαμε λοιπόν στο ίδιο μοντέλο με πριν, με τη διαφορά πως τώρα όλες οι μεταβλητές μας ήταν στατιστικώς σημαντικές, καθώς όλες είχαν p-value για τον έλεγχο

Wald μικρότερη του επιπέδου σημαντικότητας 5%. Παράλληλα, οι συντελεστές φάνηκε να είναι πιο «λογικοί» πλέον, με σκοπό την ερμηνεία του μοντέλου μας.

```
> summary(model_inter3)

Call:
glm(formula = Final.4...Semi.Finals ~ X2PT.. + X3PT.. + DR +
    FD + TO + X3PT..:FD + X2PT..:X3PT.., family = binomial(link = logit),
    data = data_centered)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.07690  -0.44994  -0.16384  -0.03831   2.79429

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.16933    0.34705  -9.132  < 2e-16 ***
X2PT..         0.27733    0.08010   3.463 0.000535 ***
X3PT..         0.25343    0.09361   2.707 0.006782 **
DR             0.56922    0.13280   4.286 1.82e-05 ***
FD             0.66150    0.13491   4.903 9.43e-07 ***
TO            -0.62484    0.15151  -4.124 3.72e-05 ***
X3PT..:FD     -0.07753    0.03636  -2.132 0.032973 *
X2PT..:X3PT..  0.05873    0.02782   2.111 0.034759 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 375.52  on 395  degrees of freedom
Residual deviance: 213.18  on 388  degrees of freedom
AIC: 229.18

Number of Fisher Scoring iterations: 7
```

Πίνακας 5.16 : Output για το τελικό, προσαρμοσμένο μοντέλο με αλληλεπιδράσεις.

Επιπροσθέτως, ο αλγόριθμος μεγιστοποίησε το λογάριθμο της πιθανοφάνειας μετά από 7 επαναλήψεις, ενώ ελέγχοντας τους δείκτες VIF, είδαμε πως δεν υπήρχε τώρα θέμα πολυσυγγραμικότητας των μεταβλητών μας.

Μεταβλητές	Τιμές VIF
2PT%	1,288
3PT%	2,407
DR	1,103
FD	1,650
TO	1,289
3PT% × FD	2,171
2PT% × 3PT%	1,458

Πίνακας 5.17 : Τιμές VIF για τις μεταβλητές του μοντέλου.

```
> anova(model_inter3,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Final.4...Semi.Finals

Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                          395     375.52
X2PT..        1   69.612       394     305.91 < 2.2e-16 ***
X3PT..        1   26.447       393     279.46 2.709e-07 ***
DR            1   22.515       392     256.94 2.085e-06 ***
FD            1   13.978       391     242.97  0.000185 ***
TO            1   18.778       390     224.19 1.468e-05 ***
X3PT..:FD     1    6.811       389     217.38  0.009057 **
X2PT..:X3PT.. 1    4.197       388     213.18  0.040507 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Πίνακας 5.18 : Output για τον έλεγχο απόκλισης του μοντέλου με αλληλεπιδράσεις.

Πήραμε τις μετρήσεις των αποκλίσεων του μοντέλου ξανά, για όταν προστίθεται κάθε φορά μια νέα μεταβλητή (από αυτές που έχουν επιλεχθεί), καθώς και για τη συνολική απόκλιση.

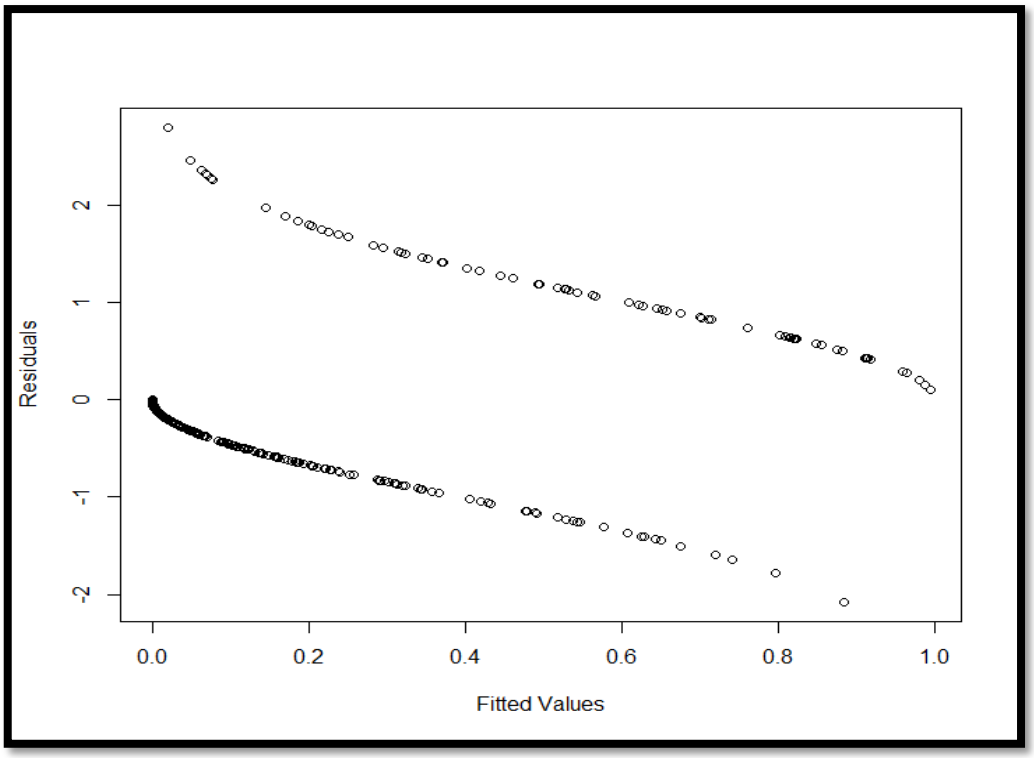
```
> y <- Playoffs...Quarter.Finals
> hl.test <- hoslem.test(model3$y, fitted(model3))
> hl.test

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  model3$y, fitted(model3)
X-squared = 8.8293, df = 8, p-value = 0.3569
```

Πίνακας 5.19 : Output για τον έλεγχο Hosmer – Lemeshow.

Ο έλεγχος των Hosmer-Lemenshow έδειξε ότι η μηδενική υπόθεση δεν απορρίπτεται ούτε εδώ, καθώς το p-value ήταν ίσο με 0,3569, μεγαλύτερο από το επίπεδο σημαντικότητας μας 5%. Συνεπώς, έχουμε και σε αυτή την περίπτωση μία πολύ ισχυρή ένδειξη καλής προσαρμογής του μοντέλου μας.



Σχήμα 5.6 : Κατάλοιπα απόκλισης του μοντέλου λογιστικής παλινδρόμησης που προσαρμόστηκε.

Τέλος, και εδώ καταλήξαμε ότι υπάρχει καλή προσαρμογή του μοντέλου μας, αφού το διάγραμμα των καταλοίπων απόκλισης έχει την αναμενόμενη μορφή.

Ερμηνεία των συντελεστών των παραμέτρων του μοντέλου με αλληλεπιδράσεις

Το μοντέλο που προσαρμόσαμε, με βάση το πίνακα 5.16, έχει τη μορφή :

$$log\left(\frac{\pi}{1-\pi}\right) = -3{,}169 + 0{,}277 \cdot 2PT\% + 0{,}253 \cdot 3PT\% + 0{,}569 \cdot DR + 0{,}661 \cdot FD - 0{,}625 \cdot TO - 0{,}078 \cdot 3PT\% \times FD + 0{,}059 \cdot 2PT\% \times 3PT\%$$

Απαλείφοντας τον λογάριθμο, έχουμε ότι η εξίσωση παίρνει την μορφή :

$$\frac{\pi}{1-\pi} = exp(-3{,}169 + 0{,}277 \cdot 2PT\% + 0{,}253 \cdot 3PT\% + 0{,}569 \cdot DR + 0{,}661 \cdot FD - 0{,}625 \cdot TO - 0{,}078 \cdot 3PT\% \times FD + 0{,}059 \cdot 2PT\% \times 3PT\%)$$

- Ο σχετικός λόγος πιθανοτήτων για τη μεταβλητή 2PT% ισούται με $e^{0,277} \approx 1,320$. Αυτό σημαίνει πως για κάθε αύξηση του ποσοστού των εύστοχων δίποντων κατά μια μονάδα, ο σχετικός λόγος πιθανοτήτων για τη πρόκριση αυξάνεται κατά περίπου 32% (1,320 – 1 = 0,320).

- Ο σχετικός λόγος πιθανοτήτων για τη μεταβλητή 3PT% ισούται με $e^{0,253} \approx 1,288$. Αυτό σημαίνει πως για κάθε αύξηση του ποσοστού των εύστοχων τρίποντων κατά μια μονάδα, ο σχετικός λόγος πιθανοτήτων για τη πρόκριση αυξάνεται κατά περίπου 28,8% (1,288 – 1 = 0,288).

- Ο σχετικός λόγος πιθανοτήτων για τη μεταβλητή DR ισούται με $e^{0,569} \approx 1,767$. Αυτό σημαίνει πως για κάθε αύξηση των αμυντικών ριμπάουντ κατά ένα, ο σχετικός λόγος πιθανοτήτων για τη πρόκριση αυξάνεται κατά περίπου 76,7% (1,767 – 1 = 0,767).

- Ο σχετικός λόγος πιθανοτήτων για τη μεταβλητή FD ισούται με $e^{0,661} \approx 1,938$. Αυτό σημαίνει πως για κάθε αύξηση των φάουλ τα οποία κερδίζει μια ομάδα κατά ένα, ο σχετικός λόγος πιθανοτήτων για τη πρόκριση αυξάνεται κατά περίπου 93,8% (1,938 – 1 = 0,938).

- Ο σχετικός λόγος πιθανοτήτων για τη μεταβλητή TO ισούται με $e^{-0,625} \approx 0,535$. Αυτό σημαίνει πως για κάθε αύξηση του πλήθους των λαθών κατά ένα, ο σχετικός λόγος πιθανοτήτων για τη πρόκριση μειώνεται κατά περίπου 46,5% (1 – 0,535 = 0,465).

- Η διαφορά μεταξύ των λογαρίθμων του λόγου σχετικών πιθανοτήτων, που αντιστοιχεί στην αύξηση του ποσοστού εύστοχων τρίποντων κατά μια μονάδα για δύο ομοιογενείς ομάδες που διαφέρουν κατά ένα φάουλ που κέρδισαν, ισούται με -0,078. Φαίνεται παράλληλα πως οι πιο εύστοχες ομάδες στα τρίποντα κέρδιζαν και τα λιγότερα φάουλ.

- Η διαφορά μεταξύ των λογαρίθμων του λόγου σχετικών πιθανοτήτων, που αντιστοιχεί στην αύξηση του ποσοστού εύστοχων δίποντων κατά μια μονάδα για δύο ομοιογενείς ομάδες που διαφέρουν κατά ένα 1% ποσοστό εύστοχων τρίποντων, ισούται με 0,059. Παράλληλα, μπορούμε να πούμε πως οι πιο εύστοχες ομάδες στα δίποντα, είχαν καλά ποσοστά και στα τρίποντα επίσης.

Εναλλακτικά :

- Η αύξηση κατά 1% του ποσοστού εύστοχων δίποντων, αυξάνει την σχετική πιθανότητα πρόκρισης πολλαπλασιαστικά κατά $e^{0,277} \approx 1,320$ φορές.

- Η αύξηση των λαθών κατά μια μονάδα, μειώνει την σχετική πιθανότητα πρόκρισης πολλαπλασιαστικά κατά $e^{-0,625} \approx 0,535$ φορές.

Με την ίδια λογική ερμηνεύονται και οι υπόλοιπες μεταβλητές.

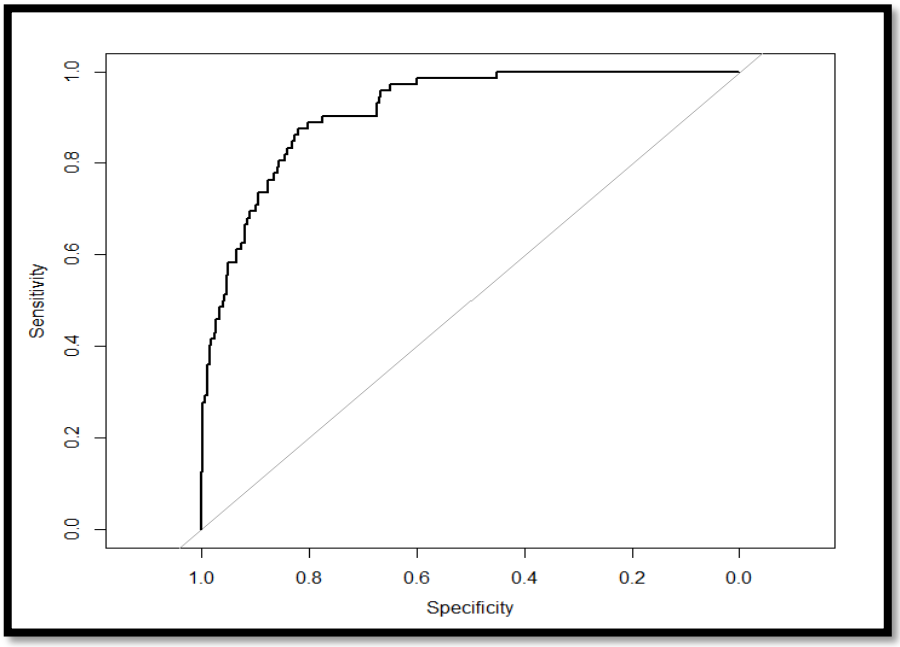
Μέτρα Προσαρμογής

Μέτρα Προσαρμογής	Τιμές
McFadden	0,432
McFadden adjusted	0,390
Cox & Snell	0,336
Nagelkerke Cragg & Uhler	0,549
AIC	229,180
BIC	261,031

Πίνακας 5.20 : Μέτρα προσαρμογής για το προσαρμοσμένο μοντέλο με αλληλεπιδράσεις.

Από το παραπάνω πίνακα, φαίνεται επίσης πως είχαμε ένα καλά προσαρμοσμένο μοντέλο, έχοντας όπως είπαμε και αλληλεπιδράσεις $2^{ης}$ τάξης. Παρατηρούμε ότι το pseudo R^2 του McFadden είναι λίγο μεγαλύτερο του 0,4 , κάτι το οποίο δείχνει και σε αυτή τη περίπτωση καλή προσαρμογή για το μοντέλο μας.

Καμπύλη ROC



Σχήμα 5.7 : Καμπύλη ROC για το προσαρμοσμένο μοντέλο με αλληλεπιδράσεις.

Το μοντέλο κρίνεται ιδιαίτερα επιτυχημένο και τώρα, καθώς η καμπύλη ROC απέχει πολύ από την διαγώνιο και το εμβαδόν κάτω από την καμπύλη ισούται με 0,915. Η τιμή πλησιάζει πολύ στο 1, που σημαίνει ότι η συνολική ακρίβεια του μοντέλου είναι αρκετά υψηλή και κατά συνέπεια, υψηλές θα είναι και οι τιμές της ευαισθησίας και της ειδικότητας. Συνεπώς, είναι αρκετά καλύτερο από ένα τυχαίο μοντέλο (με εμβαδόν ίσο με 0,5) και επομένως οι προβλέψεις του μοντέλου με τις πραγματικές παρατηρήσεις είναι πολύ κοντά. Τέλος, το cutoff σημείο που είναι πιο κοντά στην επάνω αριστερή γωνία έχει τιμή 0,1860383 και θεωρείται το βέλτιστο όσον αφορά τη διάκριση μεταξύ των ομάδων που προκρίθηκαν στο Final Four και αυτών που δεν προκρίθηκαν.

ΚΕΦΑΛΑΙΟ 6ο

Σε αυτό το κεφάλαιο, μέσω της χρήσης τεχνικών μηχανικής μάθησης και της κατάλληλης επεξεργασίας των δεδομένων μας, αποσκοπούμε στο να εξάγουμε χρήσιμη γνώση/πληροφορία. Ειδικότερα, αφού πρώτα καταλήγουμε στις κατάλληλες μεταβλητές (feature selection) που ήταν χρήσιμες για τους αλγορίθμους μας, χρησιμοποιούμε τεχνικές ομαδοποίησης (clustering) για να εντοπίσουμε μοτίβα και ομάδες παρατηρήσεων με παρόμοια χαρακτηριστικά μέσα στα δεδομένα μας, καθώς και τεχνικές κατηγοριοποίησης (classification), προκειμένου να προσαρμόσουμε κατάλληλα μοντέλα ταξινόμησης για το αν μια ομάδα θα προκρινόταν (ή όχι) στις δύο φάσεις της διοργάνωσης που εξετάζουμε. Για άλλη μια φορά, αφαιρούμε από την ανάλυση μας τις σεζόν που δεν είχαν φάση των playoffs, όπως και τη σεζόν που διακόπηκε λόγω της πανδημίας, ενώ είναι σημαντικό να αναφέρουμε επίσης πως σαν μεταβλητές απόκρισης (μεταβλητές για τις κλάσεις των δεδομένων) για τις μεθόδους που ακολουθούν, χρησιμοποιούμε τις κατηγορικές μεταβλητές **Playoffs / Quarter-Finals** και **Final 4 / Semi-Finals** , οι οποίες όπως έχουμε αναφέρει στα προηγούμενα κεφάλαια, αποτελούν ενδείξεις για το αν οι ομάδες προκρίθηκαν (ή όχι) στις αντίστοιχες φάσεις της διοργάνωσης.

6.1 Εισαγωγή στην Εξόρυξη Δεδομένων

Με τον όρο εξόρυξη δεδομένων (Data Mining) εννοούμε την εξεύρεση μιας (ενδιαφέρουσας, αυτονόητης, μη προφανούς και πιθανόν χρήσιμης) πληροφορίας από μεγάλες βάσεις δεδομένων, με χρήση αλγορίθμων ομαδοποίησης ή κατηγοριοποίησης και των αρχών της στατιστικής, της τεχνητής νοημοσύνης, της μηχανικής μάθησης και των συστημάτων βάσεων δεδομένων. Στόχος της εξόρυξης δεδομένων είναι η πληροφορία που θα εξαχθεί και τα πρότυπα που θα προκύψουν, να έχουν δομή κατανοητή προς τον άνθρωπο, έτσι ώστε να τον βοηθήσουν να πάρει τις κατάλληλες αποφάσεις.

Η εξόρυξη δεδομένων, ως ένα πραγματικά διεπιστημονικό αντικείμενο, μπορεί να οριστεί με πολλούς διαφορετικούς τρόπους. Ακόμη και ο όρος «εξόρυξη δεδομένων» δεν παρουσιάζει πραγματικά όλα τα σημαντικά συστατικά στοιχεία της εικόνας. Παραδείγματος χάρη, για να αναφερθούμε στην εξόρυξη χρυσού από πετρώματα ή άμμο, λέμε εξόρυξη χρυσού, αντί για εξόρυξη πετρωμάτων ή άμμου. Αντίστοιχα, η εξόρυξη δεδομένων θα έπρεπε να είχε ονομαστεί πιο κατάλληλα «εξόρυξη γνώσης από δεδομένα». Ωστόσο, ο συντομότερος όρος, «εξόρυξη γνώσης» μπορεί να μην αντικατοπτρίζει την έμφαση που δίνεται σε εξόρυξη από μεγάλες ποσότητες δεδομένων. Έτσι, ένας τέτοιος λανθασμένος όρος που φέρει τόσο τον όρο «δεδομένα», όσο και τον όρο «εξόρυξη» έγινε δημοφιλής επιλογή. Επιπλέον, πολλοί άλλοι όροι

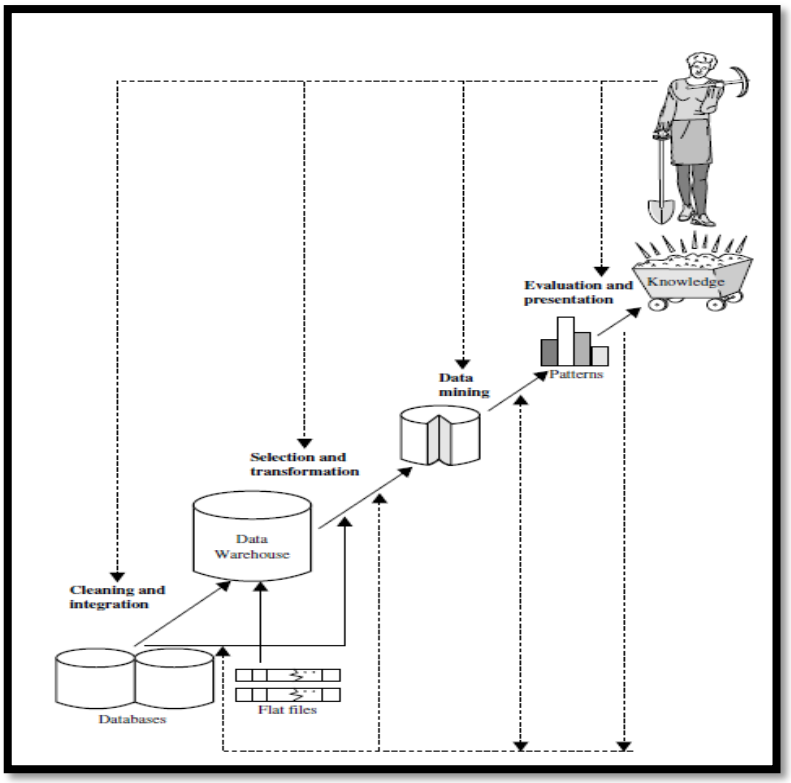
έχουν παρόμοια σημασία με την εξόρυξη δεδομένων. Για παράδειγμα, εξόρυξη γνώσης από δεδομένα, εξόρυξη γνώσης, ανάλυση δεδομένων/μοτίβων, εκσκαφή δεδομένων.

Πολλοί άνθρωποι αντιμετωπίζουν την εξόρυξη δεδομένων ως συνώνυμο ενός άλλου ευρέως χρησιμοποιούμενου όρου, «ανακάλυψη γνώσης από δεδομένα» (Knowledge Discovery from Data - KDD), ενώ άλλοι θεωρούν την εξόρυξη δεδομένων ως ένα απλό ουσιαστικό βήμα στη διαδικασία ανακάλυψης γνώσης. Η διαδικασία ανακάλυψης γνώσης παρουσιάζεται στο σχήμα 6.1 ως μια επαναληπτική ακολουθία των ακόλουθων βημάτων:

1. Καθαρισμός δεδομένων (για την απομάκρυνση του θορύβου και των ασυνεπών δεδομένων).
2. Ολοκλήρωση (integration) δεδομένων (όπου μπορούν να συνδυαστούν πολλαπλές πηγές δεδομένων).
3. Επιλογή των δεδομένων (όπου τα δεδομένα που σχετίζονται με την ανάλυση ανακτώνται από τη βάση δεδομένων).
4. Μετασχηματισμός δεδομένων (όπου τα δεδομένα μετασχηματίζονται και ενοποιούνται σε μορφές κατάλληλες για εξόρυξη με την εκτέλεση κατάλληλων πράξεων).
5. Εξόρυξη δεδομένων (μια ουσιαστική διαδικασία όπου εφαρμόζονται μέθοδοι για την εξαγωγή δεδομένων).
6. Αξιολόγηση των μοτίβων (για τον εντοπισμό των πραγματικά ενδιαφερόντων μοτίβων που αντιπροσωπεύουν τη γνώση, με βάση τα μέτρα ενδιαφέροντος).
7. Παρουσίαση της γνώσης (όπου οι τεχνικές οπτικοποίησης και αναπαράστασης χρησιμοποιούνται για την παρουσίαση της «εξορυγμένης» γνώσης στους χρήστες).

(Han et al., 2012)

Τα βήματα 1 έως 4 είναι διαφορετικές μορφές προ-επεξεργασίας δεδομένων, όπου τα δεδομένα προετοιμάζονται για την εξόρυξη. Το πέμπτο βήμα εξόρυξης δεδομένων μπορεί να αλληλεπιδρά με τον χρήστη ή με μια βάση γνώσεων. Το ενδιαφέροντα μοτίβα παρουσιάζονται στο χρήστη και μπορούν να αποθηκευτούν ως νέα γνώση στη βάση γνώσης. Η προηγούμενη άποψη δείχνει την εξόρυξη δεδομένων ως ένα βήμα στη διαδικασία ανακάλυψης γνώσης, αν και ουσιαστικό, επειδή αποκαλύπτει κρυμμένα πρότυπα για αξιολόγηση. Ωστόσο, στη βιομηχανία, στα μέσα μαζικής ενημέρωσης και στο ερευνητικό περιβάλλον, ο όρος εξόρυξη δεδομένων χρησιμοποιείται συχνά για να αναφέρεται σε ολόκληρη τη διαδικασία ανακάλυψης γνώσης (ίσως επειδή ο όρος είναι συντομότερος από την ανακάλυψη γνώσης από δεδομένα). Ως εκ τούτου, υιοθετούμε μια ευρεία θεώρηση της λειτουργικότητας της εξόρυξης δεδομένων : **Η εξόρυξη δεδομένων είναι η διαδικασία ανακάλυψης ενδιαφερόντων μοτίβων και γνώσης, από μεγάλες ποσότητες δεδομένων**. Οι πηγές δεδομένων μπορεί να περιλαμβάνουν βάσεις δεδομένων, αποθήκες δεδομένων, το διαδίκτυο, άλλα αποθετήρια πληροφοριών ή δεδομένα που μεταφέρονται με ροή στο σύστημα με δυναμικό τρόπο.



Σχήμα 6.1 : Διαδικασία εξόρυξης γνώσης.

(Πηγή : Han et al., 2012)

6.2 Εισαγωγή στην Μηχανική Μάθηση

Η μηχανική μάθηση και η εξόρυξη δεδομένων είναι στενά συνδεδεμένα πεδία που και τα δύο περιλαμβάνουν την ανάλυση μεγάλων συνόλων δεδομένων. Η εξόρυξη δεδομένων, όπως προαναφέρθηκε, είναι η διαδικασία εξαγωγής χρήσιμων πληροφοριών και γνώσεων από τα δεδομένα, ενώ η μηχανική μάθηση είναι ένα υπό-πεδίο της τεχνητής νοημοσύνης που περιλαμβάνει τη δημιουργία συστημάτων που μπορούν να μαθαίνουν από τα δεδομένα και να βελτιώνουν την απόδοσή τους με την πάροδο του χρόνου.

Ένας τρόπος να σκεφτεί κανείς τη σχέση μεταξύ μηχανικής μάθησης και εξόρυξης δεδομένων είναι ότι η μηχανική μάθηση είναι ένα εργαλείο που μπορεί να χρησιμοποιηθεί για την εκτέλεση εργασιών εξόρυξης δεδομένων. Οι αλγόριθμοι μηχανικής μάθησης μπορούν να μαθαίνουν αυτόματα μοτίβα και σχέσεις στα δεδομένα, και αυτό μπορεί να χρησιμοποιηθεί για τον εντοπισμό χρήσιμων πληροφοριών. Για παράδειγμα, οι αλγόριθμοι μηχανικής μάθησης μπορούν να χρησιμοποιηθούν για τον εντοπισμό συστάδων παρόμοιων δεδομένων, την ταξινόμηση

δεδομένων σε διαφορετικές κατηγορίες, ή την πρόβλεψη μελλοντικών αποτελεσμάτων με βάση ιστορικά δεδομένα.

Ένας άλλος τρόπος να σκεφτεί κανείς τη σχέση μεταξύ μηχανικής μάθησης και εξόρυξης δεδομένων είναι ότι η εξόρυξη δεδομένων μπορεί να χρησιμοποιηθεί για την προετοιμασία δεδομένων για αλγορίθμους μηχανικής μάθησης. Οι τεχνικές εξόρυξης δεδομένων μπορούν να χρησιμοποιηθούν για την προ-επεξεργασία και τον καθαρισμό των δεδομένων, τον εντοπισμό σχετικών χαρακτηριστικών και τη μείωση της διαστατικότητας των δεδομένων, καθιστώντας ευκολότερη τη μάθηση των αλγορίθμων μηχανικής μάθησης από τα δεδομένα.

Στην πράξη, τα όρια μεταξύ της μηχανικής μάθησης και της εξόρυξης δεδομένων είναι συχνά δυσδιάκριτα και τα δύο πεδία χρησιμοποιούνται συχνά εναλλακτικά. Πολλές τεχνικές εξόρυξης δεδομένων βασίζονται σε αλγορίθμους μηχανικής μάθησης και πολλοί αλγόριθμοι μηχανικής μάθησης χρησιμοποιούνται για εργασίες εξόρυξης δεδομένων. Καθώς ο τομέας της επιστήμης των δεδομένων συνεχίζει να εξελίσσεται, μπορούμε να περιμένουμε ότι η μηχανική μάθηση και η εξόρυξη δεδομένων θα συνεχίσουν να αποτελούν σημαντικά εργαλεία για την ανάλυση και την εξαγωγή συμπερασμάτων από μεγάλα σύνολα δεδομένων. (Hastie et al., 2009)

Υπάρχουν πολλές διαφορετικές τεχνικές μηχανικής μάθησης που μπορούν να χρησιμοποιηθούν ανάλογα με τη συγκεκριμένη εργασία που θέλουμε να πραγματοποιήσουμε, καθώς και από τα διαθέσιμα δεδομένα και πόρους. Ακολουθούν οι συνήθεις τύποι τεχνικών μηχανικής μάθησης, σύμφωνα με τον Bishop (2006).

- Επιβλεπόμενη μάθηση (Supervised Learning) : Στην επιβλεπόμενη μάθηση, ο αλγόριθμος μηχανικής μάθησης εκπαιδεύεται σε επισημασμένα δεδομένα, όπου η σωστή έξοδος/κατηγορία είναι γνωστή. Ο αλγόριθμος μαθαίνει να αντιστοιχίζει τις εισόδους στις εξόδους και μπορεί να χρησιμοποιηθεί για να κάνει προβλέψεις σε νέα δεδομένα. Ορισμένοι συνήθεις αλγόριθμοι μάθησης με επίβλεψη περιλαμβάνουν τη γραμμική παλινδρόμηση, τη λογιστική παλινδρόμηση, τα δέντρα αποφάσεων και τα νευρωνικά δίκτυα.

- Μάθηση χωρίς επίβλεψη (Unsupervised Learning) : Στη μάθηση χωρίς επίβλεψη, ο αλγόριθμος μηχανικής μάθησης εκπαιδεύεται σε μη επισημασμένα δεδομένα και ο στόχος είναι να εντοπιστούν μοτίβα και δομές μέσα στα ίδια δεδομένα. Αυτό μπορεί να περιλαμβάνει εργασίες όπως η ομαδοποίηση, η ανίχνευση «ανωμαλιών» και η μείωση διαστάσεων. Ορισμένοι συνήθεις αλγόριθμοι μάθησης χωρίς επίβλεψη περιλαμβάνουν την ομαδοποίηση K-means (την οποία θα χρησιμοποιήσουμε εμείς προσεχώς) και την ανάλυση κύριων συνιστωσών (Principle Component Analysis - PCA).

- Μάθηση με ημιεπίβλεψη (Semi-Supervised Learning) : Στην ημι-επιβλεπόμενη μάθηση, ο αλγόριθμος μηχανικής μάθησης εκπαιδεύεται σε έναν συνδυασμό επισημασμένων και μη επισημασμένων δεδομένων. Αυτό μπορεί να είναι

χρήσιμο σε περιπτώσεις όπου η επισήμανση δεδομένων είναι δαπανηρή ή χρονοβόρα.

- Ενισχυτική μάθηση (Reinforcement Learning) : Στην ενισχυτική μάθηση, ο αλγόριθμος μηχανικής μάθησης μαθαίνει να λαμβάνει αποφάσεις με βάση την ανατροφοδότηση από το περιβάλλον. Ο αλγόριθμος λαμβάνει «ανταμοιβές» ή τιμωρίες με βάση τις ενέργειές του, και μαθαίνει να μεγιστοποιεί την ανταμοιβή με την πάροδο του χρόνου. Η ενισχυτική μάθηση έχει χρησιμοποιηθεί σε εφαρμογές όπως τα ηλεκτρονικά παιχνίδια, η ρομποτική και η αυτόνομη οδήγηση.

- Βαθιά μάθηση (Deep Learning) : Η βαθιά μάθηση είναι ένας τύπος μηχανικής μάθησης που χρησιμοποιεί νευρωνικά δίκτυα με πολλαπλά επίπεδα. Αυτά τα δίκτυα μπορούν να μάθουν πολύπλοκες, ιεραρχικές αναπαραστάσεις δεδομένων και έχουν χρησιμοποιηθεί σε εφαρμογές, όπως η αναγνώριση εικόνας και ομιλίας, η επεξεργασία φυσικής γλώσσας, καθώς και η ανακάλυψη φαρμάκων.

6.3 Επιλογή χαρακτηριστικών (Feature Selection)

Πριν ξεκινήσουμε με τις τεχνικές μηχανική μάθησης για να εξάγουμε γνώση, είναι σημαντικό να επιλέξουμε τις κατάλληλες μεταβλητές που θα είναι χρήσιμες για τις αναλύσεις μας, να πραγματοποιήσουμε δηλαδή επιλογή χαρακτηριστικών. Η επιλογή χαρακτηριστικών είναι η διαδικασία επιλογής ενός υποσυνόλου σχετικών χαρακτηριστικών/μεταβλητών από ένα μεγαλύτερο σύνολο χαρακτηριστικών για τη δημιουργία ενός προγνωστικού μοντέλου, ή για τη μείωση της διαστατικότητας ενός συνόλου δεδομένων. Αυτό γίνεται συχνά για να βελτιωθεί η ακρίβεια του μοντέλου, να μειωθεί η υπερ-προσαρμογή (over-fitting) και να επιταχυνθεί η διαδικασία εκπαίδευσης του μοντέλου. Υπάρχουν δύο βασικές κατηγορίες των μεθόδων επιλογής χαρακτηριστικών :

- Μέθοδοι χωρίς επίβλεψη (Unsupervised) : Η επιλογή χαρακτηριστικών χωρίς επίβλεψη αναφέρεται στη μεθόδους που δεν χρειάζονται την κατηγορία εξόδου για την επιλογή χαρακτηριστικών. Χρησιμοποιούνται συνεπώς για δεδομένα χωρίς να είναι γνωστές οι κλάσεις στις οποίες ανήκουν.

- Μέθοδοι με επίβλεψη (Supervised) : Επιβλεπόμενη επιλογή χαρακτηριστικών αναφέρεται στις μεθόδους που χρειάζονται την κατηγορία εξόδου για την επιλογή χαρακτηριστικών. Χρησιμοποιούν τις μεταβλητές-στόχους για να προσδιορίσουν τις μεταβλητές που μπορούν να αυξήσουν την αποτελεσματικότητα του μοντέλου.

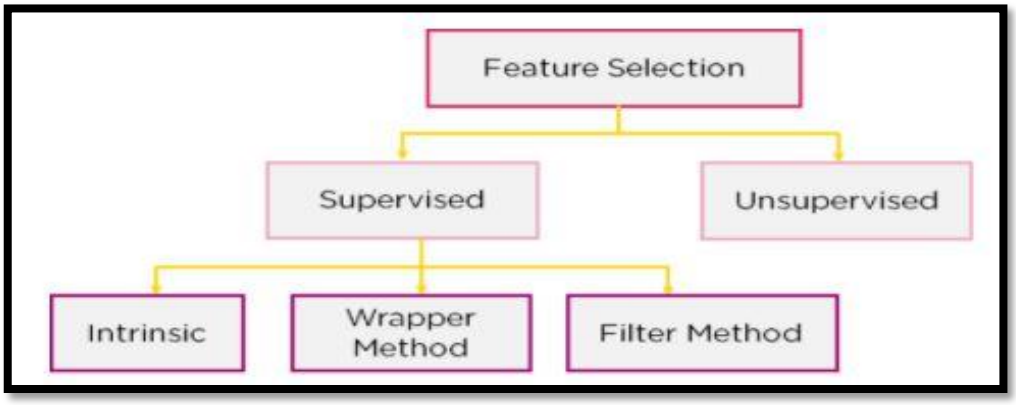
Με τη σειρά τους, οι supervised μέθοδοι μπορούν να χωριστούν σε τρεις υπό-κατηγορίες, τις μεθόδους «φίλτρου» (Filter), τις μεθόδους «περιτύλιξης» (Wrapper) και τις ενσωματωμένες μεθόδους (Intrinsic or Embedded).

- Μέθοδος φίλτρου : Σε αυτή τη μέθοδο, τα χαρακτηριστικά επιλέγονται με βάση τη σχέση τους με την μεταβλητή εξόδου, ή τον τρόπο με τον οποίο συσχετίζονται με την μεταβλητή αυτή. Χρησιμοποιούμε τη συσχέτιση για να ελέγξουμε αν τα χαρακτηριστικά συσχετίζονται θετικά ή αρνητικά με τις μεταβλητές απόκρισης και απορρίπτουμε τα χαρακτηριστικά αναλόγως. Χρησιμοποιούνται για τον έλεγχο συσχέτισης π.χ. το Chi-Square Test, το Fisher's Score κ.ά.

- Μέθοδος «περιτύλιξης» : Σύμφωνα με αυτή τη μέθοδο, διαχωρίζονται τα δεδομένα σε υποσύνολα και εκπαιδεύεται ένα μοντέλο, χρησιμοποιώντας αυτά. Με βάση την έξοδο του μοντέλου, προθέτονται και αφαιρούνται χαρακτηριστικά, εκπαιδεύοντας ξανά το μοντέλο. Διαμορφώνει τα υποσύνολα χρησιμοποιώντας μια «άπληστη» προσέγγιση και αξιολογεί την ακρίβεια όλων των πιθανών συνδυασμών των χαρακτηριστικών. Παραδείγματα τέτοιων μεθόδων είναι η επιλογή προς τα εμπρός (Forward Selection), εξάλειψη προς τα πίσω (Backwards Elimination) κ.λ.π.

- Ενσωματωμένη μέθοδος : Αυτή η μέθοδος συνδυάζει τις ιδιότητες τόσο της μεθόδου φίλτρου, όσο και της μεθόδου περιτύλιξης για τη δημιουργία του καλύτερου υποσυνόλου. Αυτή η μέθοδος φροντίζει για την επαναληπτική διαδικασία εκπαίδευσης του μοντέλου, διατηρώντας παράλληλα το κόστος υπολογισμού στο ελάχιστο δυνατό. Τέτοιες μέθοδοι είναι οι Lasso και Ridge παλινδρομήσεις.

(Menon, 2023)



Σχήμα 6.2 : Κατηγοριοποίηση των μεθόδων επιλογής χαρακτηριστικών.

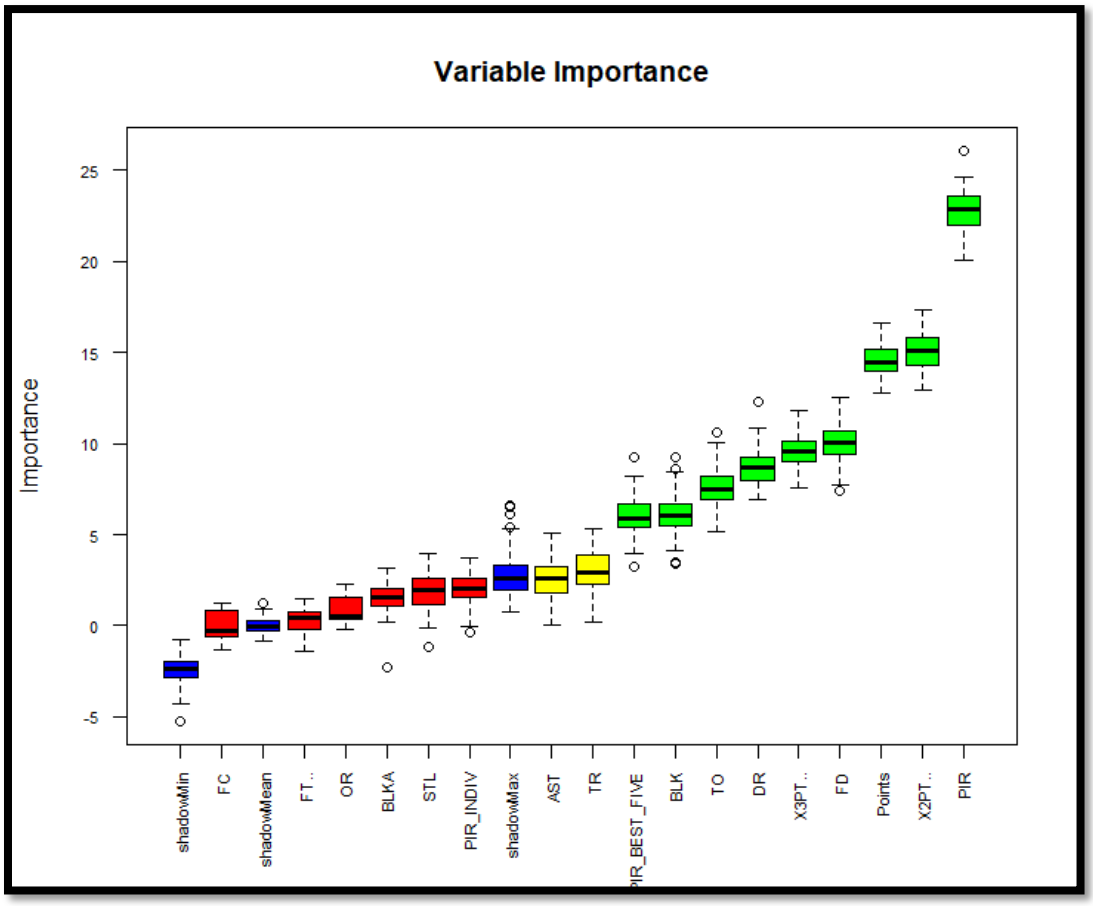
(Πηγή : https://www.simplilearn.com/tutorials/machine-learning-tutorial/feature-selection-in-machine-learning)

Εμείς στη παρούσα εργασία, αποφασίσαμε να χρησιμοποιήσουμε τη μέθοδο **Boruta** (wrapper method), προκειμένου να επιλέξουμε τις κατάλληλες μεταβλητές για τις αναλύσεις μας. Ένας λόγος που διαλέξαμε αυτή τη μέθοδο για την επιλογή των χαρακτηριστικών μας είναι επειδή δε λαμβάνει υπόψιν τις συσχετίσεις ανάμεσα στις επεξηγηματικές μεταβλητές με την μεταβλητή απόκρισης. Ο συγκεκριμένος αλγόριθμος παρουσιάστηκε από τους Πολωνούς ερευνητές Kursa και Rudnicki το 2010. Το όνομα του προέρχεται από μια σλαβική, μυθολογική φιγούρα που ενσαρκώνει το πνεύμα του δάσους. Βασίζεται στον αλγόριθμο ταξινόμησης τυχαίου δάσους (Random Forest Classifier), τον οποίο αναλύουμε στην ενότητα 6.5.3. Προσπαθεί να συλλάβει όλα τα σημαντικά χαρακτηριστικά/μεταβλητές που μπορεί να έχει το σύνολο δεδομένων, αναφορικά με τη μεταβλητή απόκρισης.

Αρχικά, προσθέτει τυχαιότητα στο σύνολο δεδομένων που έχουμε διαθέσιμο, δημιουργώντας ανακατεμένα αντίγραφα όλων των χαρακτηριστικών που ονομάζονται σκιώδη χαρακτηριστικά (shadow features). Στη συνέχεια, εκπαιδεύει έναν ταξινομητή τυχαίου δάσους στο σύνολο δεδομένων και εφαρμόζει ένα μέτρο σημαντικότητας χαρακτηριστικών, όπως το Mean Decrease Accuracy, ώστε να αξιολογήσει τη σημασία κάθε χαρακτηριστικού. Σε κάθε επανάληψη, ο αλγόριθμος συγκρίνει τα Z-scores των σκιωδών χαρακτηριστικών και των αρχικών χαρακτηριστικών, για να διαπιστώσει αν τα δεύτερα είχαν καλύτερη απόδοση από τα πρώτα. Εάν ναι, ο αλγόριθμος θα χαρακτηρίσει το χαρακτηριστικό ως σημαντικό. Στην ουσία, ο αλγόριθμος προσπαθεί να επικυρώσει τη σημασία του χαρακτηριστικού συγκρίνοντάς το με τα τυχαία ανακατεμένα αντίγραφα, γεγονός που αυξάνει την ανθεκτικότητα. Αυτό γίνεται απλά συγκρίνοντας τον αριθμό των περιπτώσεων που ένα χαρακτηριστικό τα πήγε καλύτερα σε σχέση με τα σκιώδη χαρακτηριστικά, χρησιμοποιώντας μια διωνυμική κατανομή. Ο αλγόριθμος σταματάει μετά από έναν προκαθορισμένο αριθμό επαναλήψεων, ή αν όλα τα χαρακτηριστικά έχουν είτε επιβεβαιωθεί, είτε απορριφθεί. (Kursa & Rudnicki, 2010)

6.3.1 Επιλογή χαρακτηριστικών για τη φάση των Playoffs

Χρησιμοποιώντας τη μέθοδο Boruta που περιγράψαμε παραπάνω, πήραμε το παρακάτω σχήμα.



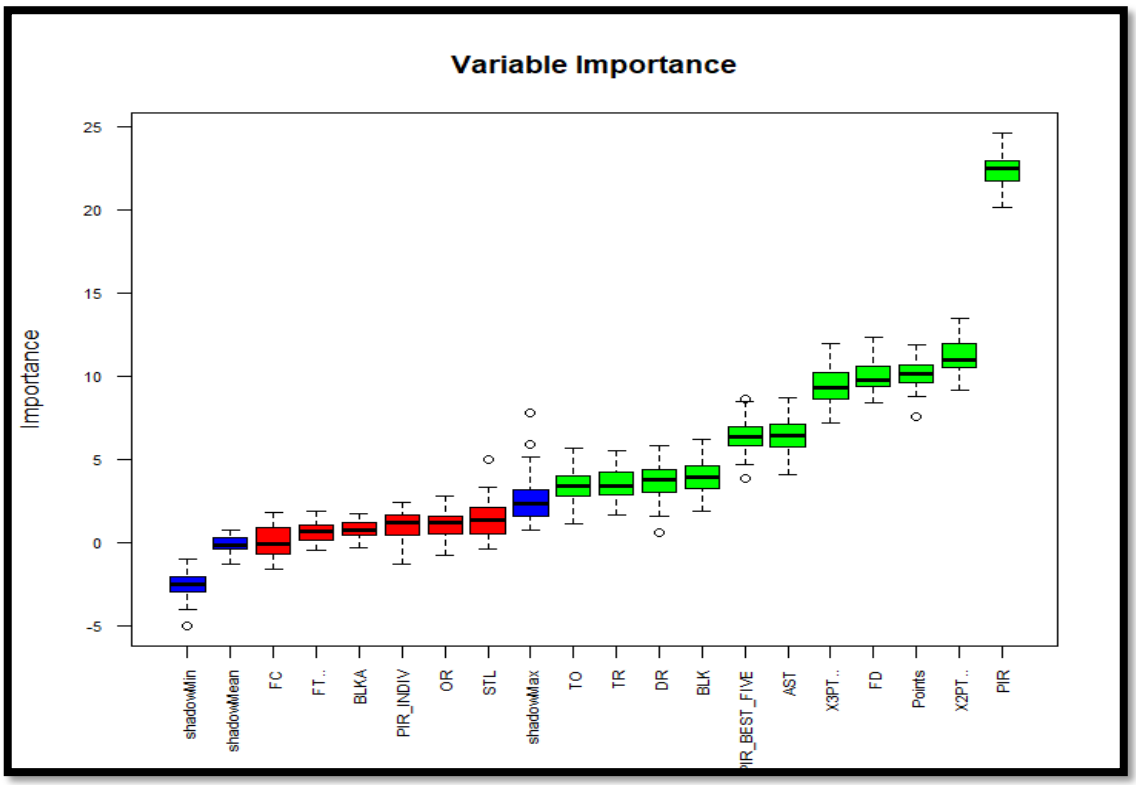
Σχήμα 6.3 : Γραφική αναπαράσταση της σημαντικότητας των μεταβλητών με τη μέθοδο Boruta για τα playoffs.

Παρατηρούμε ότι έγινε μια ταξινόμηση των μεταβλητών μας σε τρία επίπεδα, τα οποία στο σχήμα δηλώνονται με αντίστοιχα χρώματα. Με πράσινο χρώμα είναι οι μεταβλητές που έχουν επιβεβαιωθεί (confirmed) από τον αλγόριθμο του random forest. Συνεπώς, καταλήξαμε πως οι πιο σημαντικές μεταβλητές είναι ο δείκτης PIR, το ποσοστό εύστοχων δίποντων, οι πόντοι, τα φάουλ που κέρδισαν οι ομάδες, το ποσοστό εύστοχων τρίποντων, τα αμυντικά ριμπάουντ, τα λάθη, τα κοψίματα/μπλοκ και ο μέσος όρος του δείκτη PIR των πέντε καλύτερων παικτών κάθε ομάδας. Με κόκκινο χρώμα είναι οι μεταβλητές που δεν έχουν εγκριθεί από τον ταξινομητή. Με κίτρινο χρώμα απεικονίζονται οι tentative μεταβλητές, δηλαδή οι μεταβλητές που είναι στην κρίση του εκάστοτε αναλυτή αν θα συμπεριληφθούν στην μελέτη, και η συνεισφορά τους στο μοντέλο είναι υπό διερεύνηση. Από τη μία πλευρά, η συμπερίληψη αυτών των μεταβλητών μπορεί ενδεχομένως να βελτιώσει την ακρίβεια του μοντέλου, «αιχμαλωτίζοντας» πρόσθετη πληροφορία που δεν αποδόθηκε από τα επιλεγμένα χαρακτηριστικά. Από την άλλη πλευρά, η συμπερίληψη υπερβολικά πολλών χαρακτηριστικών στο μοντέλο μπορεί να αυξήσει τον κίνδυνο over-fitting. Τέλος, με μπλε χρώμα δηλώνονται τα σκιώδη χαρακτηριστικά (shadow features) που όπως αναφέρθηκε στη περιγραφή του αλγορίθμου, δεν είναι πραγματικές μεταβλητές, απλά είναι χρήσιμα στη διαδικασία προκειμένου να αποφασίσουμε αν μια μεταβλητή είναι

σημαντική ή όχι. Οι τιμές τους αντιστοιχούν στην ελάχιστη τιμή (min), την μέση τιμή (average) και τη μέγιστη τιμή (max) του Z-score.

6.3.2 Επιλογή χαρακτηριστικών για τη φάση του Final Four

Εργαζόμενοι ομοίως με παραπάνω, καταλήξαμε πως οι πιο σημαντικές μεταβλητές για τη φάση του Final Four είναι ο δείκτης PIR, το ποσοστό εύστοχων δίποντων, οι πόντοι, τα φάουλ που κέρδισαν οι ομάδες, το ποσοστό εύστοχων τρίποντων, οι ασίστ, ο μέσος όρος του δείκτη PIR των πέντε καλύτερων παικτών κάθε ομάδας, τα κοψίματα/μπλοκ, τα αμυντικά ριμπάουντ, τα συνολικά ριμπάουντ και τα λάθη. Αυτό φαίνεται στο παρακάτω σχήμα.



Σχήμα 6.4 : Γραφική αναπαράσταση της σημαντικότητας των μεταβλητών με τη μέθοδο Boruta για το Final Four.

Μεταβλητές για τη φάση των Playoffs	Μεταβλητές για τη φάση του Final Four
PIR	PIR
2PT%	2PT%
Points	Points
FD	FD
3PT%	3PT%
DR	AST
TO	PIR_BEST_FIVE

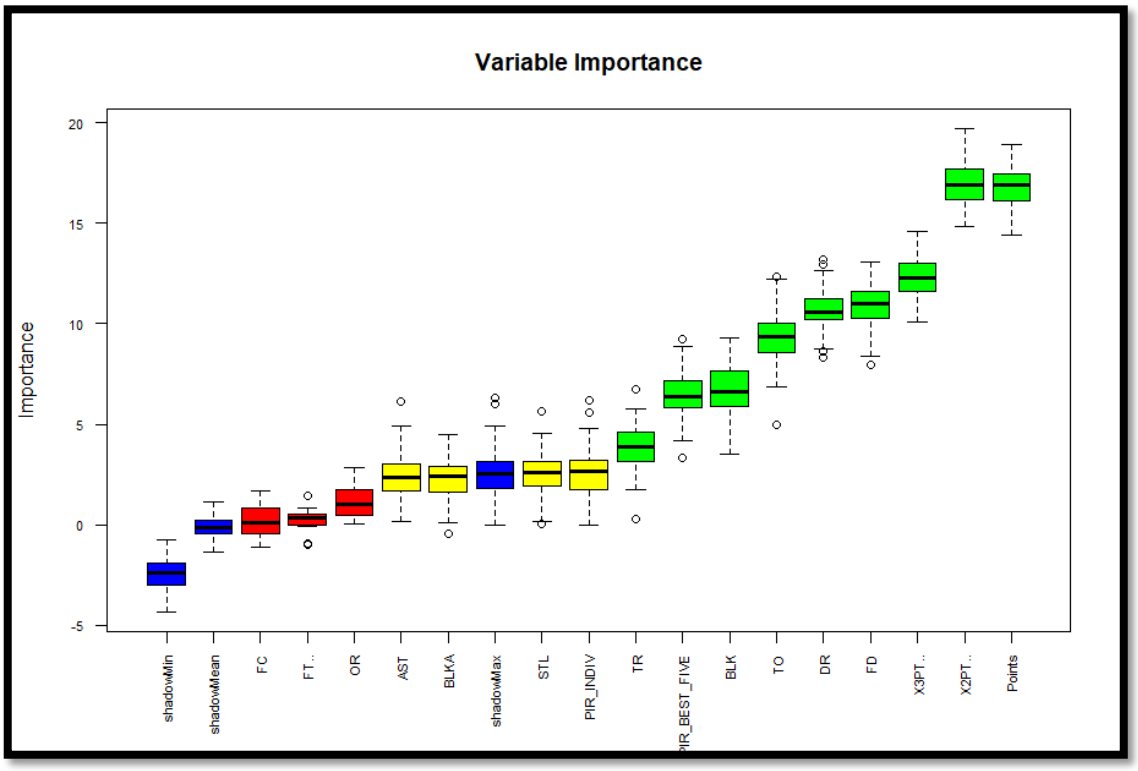
BLK	BLK
PIR_BEST_FIVE	DR
TR *	TR
AST *	TO

Πίνακας 6.1 : Χαρακτηριστικά που χρησιμοποιήθηκαν στις δύο φάσεις.

(* Οι TR και AST δε συμπεριλήφθηκαν στην ανάλυση, για αυτή τη φάση)

Από το παραπάνω πίνακα, παρατηρούμε ότι και για τις δυο φάσεις της διοργάνωσης που εξετάζουμε, τα πέντε πρώτα χαρακτηριστικά τα οποία είχαν και τη μεγαλύτερη σημασία ώστε να χρησιμοποιηθούν στις αναλύσεις μας, ήταν ακριβώς τα ίδια.

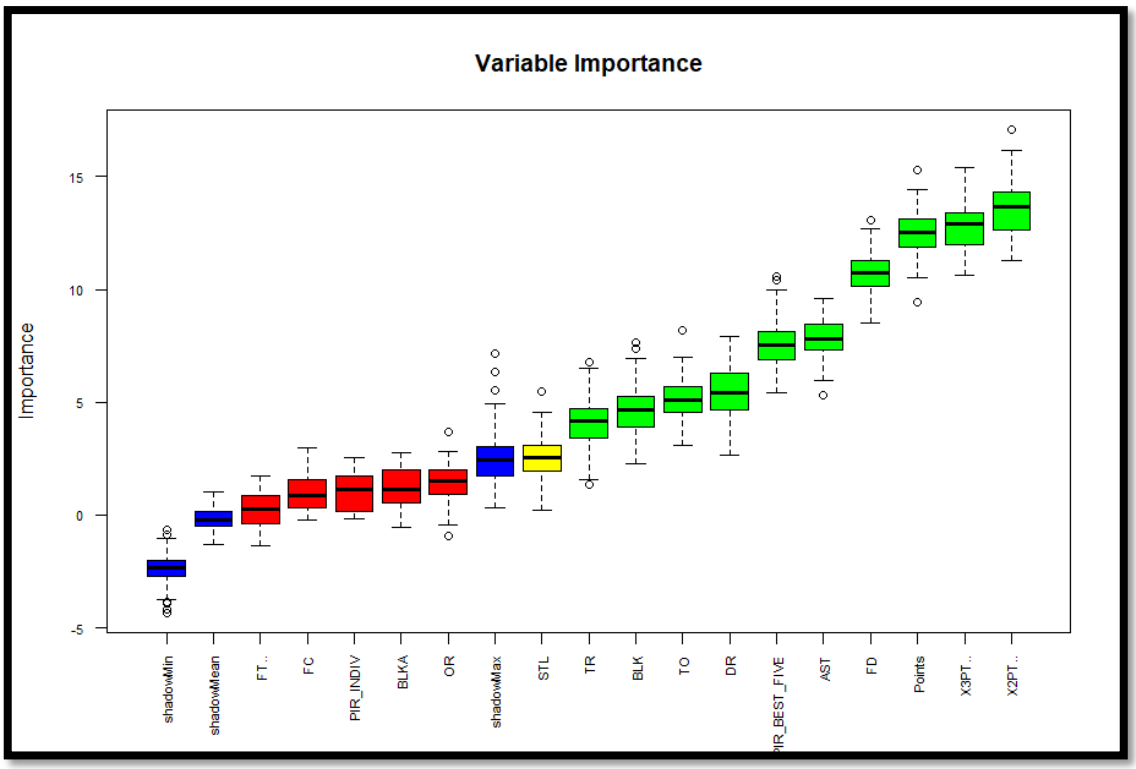
Όπως είδαμε και στο δεύτερο κεφάλαιο, ο δείκτης PIR είναι άρρηκτα συνδεδεμένος με τις υπόλοιπες μεταβλητές των δεδομένων μας, καθώς προκύπτει από συνδυασμό αυτών. Για αυτό το λόγο, αποφασίσαμε να τρέξουμε πάλι τους αλγορίθμους για τη μέθοδο Boruta, έχοντας αφαιρέσει τώρα τη μεταβλητή του δείκτη, θέλοντας απλώς να δούμε τι αποτελέσματα θα είχαμε αυτή τη φορά.



Σχήμα 6.5 : Γραφική αναπαράσταση της σημαντικότητας των μεταβλητών με τη μέθοδο Boruta για τα playoffs, χωρίς το PIR.

Παρατηρήσαμε πως για τη φάση των playoffs, πιο σημαντικές μεταβλητές παρέμειναν οι πόντοι, το ποσοστό εύστοχων δίποντων, το ποσοστό εύστοχων τρίποντων, τα φάουλ που κέρδισαν οι ομάδες, τα αμυντικά ριμπάουντ, τα λάθη, τα κοψίματα/μπλοκ και ο

μέσος όρος του δείκτη PIR των πέντε καλύτερων παικτών κάθε ομάδας, ενώ πλέον θεωρήθηκε σημαντικό και το πλήθος των συνολικών ριμπάουντ.



Σχήμα 6.6 : Γραφική αναπαράσταση της σημαντικότητας των μεταβλητών με τη μέθοδο Boruta για το Final Four, χωρίς το PIR.

Αντίστοιχα, για τη φάση του Final Four, πιο σημαντικές μεταβλητές παρέμειναν το ποσοστό εύστοχων δίποντων, το ποσοστό εύστοχων τρίποντων, οι πόντοι, τα φάουλ που κέρδισαν οι ομάδες, οι ασίστ, ο μέσος όρος του δείκτη PIR των πέντε καλύτερων παικτών κάθε ομάδας, τα αμυντικά ριμπάουντ, τα λάθη, τα κοψίματα/μπλοκ και τα συνολικά ριμπάουντ. Στην ουσία, δεν άλλαξε κάτι ιδιαίτερα, παρά μόνο η σημασία της κάθε μεταβλητής.

Τελευταίο βήμα που πρέπει να κάνουμε πριν αρχίσουμε την ανάλυσή μας με τις τεχνικές μηχανικής μάθησης για την ομαδοποίηση (clustering), είναι να κανονικοποιήσουμε τα δεδομένα μας. Τα μοντέλα ομαδοποίησης είναι αλγόριθμοι, οι οποίοι βασίζονται στην απόσταση. Για να μετρήσουν συνεπώς τις ομοιότητες μεταξύ των παρατηρήσεων και να σχηματίσουν συστάδες, χρησιμοποιούν ένα μέτρο απόστασης. Έτσι, χαρακτηριστικά με μεγάλες αποστάσεις θα έχουν και μεγαλύτερη επίδραση στην ομαδοποίηση. Επομένως, απαιτείται κανονικοποίηση των μεταβλητών πριν από τη δημιουργία ενός μοντέλου συσταδοποίησης. Εμείς χρησιμοποιήσαμε την εντολή **scale** στην R, η οποία αφαιρεί από κάθε τιμή της μεταβλητής τον μέσο όρο κάθε στήλης, και στη συνέχεια διαιρεί κάθε στοιχείο με τη τυπική του απόκλιση.

6.4 Ανάλυση κατά συστάδες (Cluster Analysis / Clustering)

Η ανάλυση κατά συστάδες (ή αλλιώς ομαδοποίηση/συσταδοποίηση) είναι μια θεμελιώδης τεχνική στην εξόρυξη δεδομένων και τη μηχανική μάθηση, η οποία περιλαμβάνει την κατανομή ενός συνόλου αντικειμένων σε ομάδες ή συστάδες, έτσι ώστε τα αντικείμενα εντός της μιας ομάδας να μοιάζουν περισσότερο μεταξύ τους, να είναι δηλαδή πιο ομοιογενή, σε σχέση με εκείνα των άλλων ομάδων. Η συσταδοποίηση χρησιμοποιείται για διάφορους σκοπούς, όπως ο εντοπισμός φυσικών ομαδοποιήσεων παρόμοιων αντικειμένων, η ανακάλυψη ακραίων τιμών ή ανωμαλιών σε δεδομένα (outlier detection), και η μείωση της διαστατικότητας σε μεγάλα σύνολα δεδομένων. Η επιλογή του αλγορίθμου εξαρτάται από τη φύση των δεδομένων και τους στόχους της ανάλυσης. Η συσταδοποίηση έχει πολλές εφαρμογές σε ποικίλους τομείς, όπως το μάρκετινγκ, οι βιο-επιστήμες, τα χρηματοοικονομικά κ.ά. (Tan et al., 2018)

Ένα από τα βασικά στοιχεία της συσταδοποίησης είναι ο ορισμός ενός μέτρου απόστασης που μπορεί να μετρήσει την ομοιότητα ή τη διαφορετικότητα μεταξύ των αντικειμένων. Ένα μέτρο απόστασης καθορίζει τον τρόπο υπολογισμού της απόστασης μεταξύ δύο οποιωνδήποτε αντικειμένων στο σύνολο δεδομένων. Η επιλογή τους εξαρτάται από τη φύση των δεδομένων και τον τύπο του αλγορίθμου ομαδοποίησης που χρησιμοποιείται. Οι αλγόριθμοι συσταδοποίησης διαφέρουν ως προς την ικανότητά τους να χειρίζονται τις διαφορετικές αποστάσεις και τύπους δεδομένων.

Υπάρχουν διάφορα μέτρα απόστασης που χρησιμοποιούνται, μερικά από τα οποία παρατίθενται παρακάτω :

Ευκλείδεια απόσταση

Περιορίζεται σε διανύσματα πραγματικών τιμών. Υπολογίζει το μήκος της ευθείας γραμμής μεταξύ δύο σημείων και έχει τύπο

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Απόσταση Manhattan

Αυτή είναι μια άλλη επίσης δημοφιλής μετρική απόστασης, η οποία μετρά την απόλυτη τιμή μεταξύ δύο σημείων. Αναφέρεται επίσης ως «απόσταση ταξί» ή «απόσταση οικοδομικού τετραγώνου», καθώς συνήθως απεικονίζεται με ένα πλέγμα, απεικονίζοντας τον τρόπο με τον οποίο μπορεί κανείς να πλοηγηθεί από μια διεύθυνση σε μια άλλη μέσω των δρόμων της πόλης. Δίνει περίπου ίδια αποτελέσματα με την ευκλείδεια απόσταση, εκτός από την περίπτωση που υπάρχουν έκτροπες παρατηρήσεις

(outliers) οπότε, επειδή τους δίνει μικρότερο βάρος (η διαφορά δεν υψώνεται στο τετράγωνο), μπορεί να οδηγήσει σε πιο ανθεκτικά αποτελέσματα. Έχει τύπο

$$d(x,y) = (\sum_{i=1}^{m} |x_i - y_i|)$$

Απόσταση Minkowski

Αυτό το μέτρο απόστασης είναι η γενικευμένη μορφή των μετρικών της Ευκλείδειας απόστασης και της απόστασης Manhattan. Η παράμετρος λ (λ≥1) επιτρέπει τη δημιουργία άλλων μετρικών απόστασης. Για λ=2, προκύπτει η ευκλείδεια απόσταση, ενώ για λ=1 έχουμε την απόσταση Manhattan. Ο τύπος της είναι

$$d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^{\lambda}\right)^{1/\lambda}$$

Κάποιοι ερευνητές (Tan et al., 2018) κατηγοριοποιούν τις μεθόδους συσταδοποίησης σε τέσσερις κατηγορίες, τις μεθόδους που βασίζονται στα κέντρα (ή αλλιώς «κατάτμησης»), τις ιεραρχικές μεθόδους, τις μεθόδους που βασίζονται στη πυκνότητα και σε αυτές που βασίζονται στα μοντέλα.

Μέθοδοι «κατάτμησης» (Partitioning methods)

Οι μέθοδοι αυτές, όπως ο K-means (που θα χρησιμοποιήσουμε εμείς προσεχώς), χωρίζουν τα δεδομένα σε μη επικαλυπτόμενες συστάδες, ελαχιστοποιώντας μια συνάρτηση κριτηρίου. Αυτές οι μέθοδοι είναι υπολογιστικά αποδοτικές και μπορούν να λειτουργήσουν καλά σε μεγάλα σύνολα δεδομένων, αλλά είναι ευαίσθητες στην αρχική επιλογή του πλήθους των συστάδων.

Ιεραρχικές μέθοδοι (Hierarchical methods)

Οι ιεραρχικές μέθοδοι, όπως η συσσωρευτική συσταδοποίηση (επίσης θα τη χρησιμοποιήσουμε προσεχώς), δημιουργούν μια ένθετη ιεραρχία συστάδων με τη διαδοχική συγχώνευση των πιο όμοιων συστάδων ή σημείων, με βάση ένα μέτρο απόστασης. Αυτές οι μέθοδοι μπορούν να συλλάβουν πολύπλοκες δομές στα δεδομένα, αλλά μπορεί να είναι υπολογιστικά δαπανηρές και ευαίσθητες στην επιλογή της απόστασης.

Μέθοδοι με βάση την πυκνότητα (Density-based methods)

Οι μέθοδοι με βάση την πυκνότητα, όπως ο DBSCAN, ομαδοποιούν τα σημεία δεδομένων που βρίσκονται εντός ενός καθορισμένου ορίου απόστασης ή πυκνότητας μεταξύ τους. Αυτές οι μέθοδοι μπορούν να χειριστούν συστάδες αυθαίρετου σχήματος και θορυβώδη δεδομένα, αλλά είναι ευαίσθητες στην επιλογή του μέτρου απόστασης και στις ρυθμίσεις των παραμέτρων τους.

Μέθοδοι βασισμένες σε μοντέλα (Model-based methods)

Οι μέθοδοι αυτές, βασίζονται σε μοντέλα, όπως τα μοντέλα μίξης Gauss, υποθέτουν ότι τα δεδομένα παράγονται από ένα πιθανολογικό μοντέλο και εκτιμούν τις παραμέτρους του μοντέλου για τον εντοπισμό συστάδων. Αυτές οι μέθοδοι μπορούν να χειριστούν πολύπλοκες κατανομές δεδομένων και να παρέχουν πιθανολογικές αναθέσεις συστάδων, αλλά μπορεί να είναι υπολογιστικά δαπανηρές και ευαίσθητες στην επιλογή του μοντέλου και των ρυθμίσεων των παραμέτρων.

6.4.1 Εφαρμογή αλγορίθμου K-Means για τη φάση των playoffs

Ο αλγόριθμος K-means είναι ένας δημοφιλής αλγόριθμος ομαδοποίησης που χρησιμοποιείται στην εξόρυξη δεδομένων και τη μηχανική μάθηση. Η βασική ιδέα πίσω από τον αλγόριθμο K-means είναι να χωρίσει ένα σύνολο n σημείων δεδομένων σε k συστάδες, όπου k είναι μια παράμετρος που ορίζεται από τον χρήστη. Τα βήματα του αλγορίθμου είναι τα εξής :

1. Αρχικοποίηση : Ο χρήστης επιλέγει k αρχικά κέντρα, τυχαία από το σύνολο δεδομένων.
2. Ανάθεση : Κάθε σημείο δεδομένων ανατίθεται στη συστάδα της οποίας το κέντρο είναι πλησιέστερα σε αυτό, χρησιμοποιώντας ένα μέτρο απόστασης, όπως η ευκλείδεια απόσταση.
3. Ενημέρωση: Τα κέντρα των k συστάδων ενημερώνονται με τον υπολογισμό του μέσου όρου όλων των σημείων που έχουν ανατεθεί στη συγκεκριμένη συστάδα.
4. Επανάληψη: Τα βήματα 2 και 3 επαναλαμβάνονται μέχρι σύγκλισης, δηλαδή μέχρι η ανάθεση των σημείων στις συστάδες να μην αλλάζει πλέον.

Ο αλγόριθμος K-means είναι ευαίσθητος στην αρχική επιλογή των κέντρων και μπορεί να κολλήσει σε τοπικά ελάχιστα. Ως εκ τούτου, συχνά εκτελείται πολλές φορές με διαφορετικές αρχικοποιήσεις για να αυξηθούν οι πιθανότητες εύρεσης μιας καλής λύσης.

Ο αλγόριθμος παράλληλα έχει και ορισμένους περιορισμούς. Ένας περιορισμός είναι ότι υποθέτει ότι οι συστάδες είναι σφαιρικές και ότι έχουν ίσες αποκλίσεις. Αυτό

μπορεί να οδηγήσει σε κακή απόδοση, όταν τα δεδομένα δεν είναι καλά διαχωρισμένα ή έχουν πολύπλοκα γεωμετρικά σχήματα. Ένας άλλος περιορισμός είναι ότι απαιτεί από τον χρήστη να καθορίσει τον αριθμό των συστάδων k (όπως αναφέρθηκε παραπάνω), ο οποίος μπορεί να μην είναι γνωστός εκ των προτέρων. Για την επιλογή της κατάλληλης τιμής του k, μπορούν να χρησιμοποιηθούν διάφορες τεχνικές, όπως η μέθοδος του αγκώνα και η ανάλυση «σιλουέτας», τις οποίες θα χρησιμοποιήσουμε και εμείς.

Μέθοδος αγκώνα (Elbow Method)

Η μέθοδος του αγκώνα είναι μια τεχνική για την επιλογή του αριθμού των συστάδων στην ομαδοποίηση K-means. Η βασική ιδέα πίσω από τη μέθοδο του αγκώνα είναι να σχεδιαστεί το άθροισμα των τετραγώνων εντός συστάδας (within-cluster sum of squares - WCSS) ως συνάρτηση του αριθμού συστάδων k και να βρεθεί ένας «αγκώνας» ή μια καμπύλη στο διάγραμμα. Το άθροισμα τετραγώνων εντός συστάδας μετρά το άθροισμα των τετραγωνικών αποστάσεων μεταξύ κάθε σημείου δεδομένων και του κέντρου της συστάδας στην οποία έχει ανατεθεί, και δίνεται από τον ακόλουθο τύπο :

$$WCSS = \sum_{i} \sum_{j} \|x_i - c_j\|^2$$

όπου x_i είναι η i-οστή παρατήρηση, c_j είναι το κέντρο της j-οστής συστάδας και ||.|| δηλώνει την ευκλείδεια απόσταση.

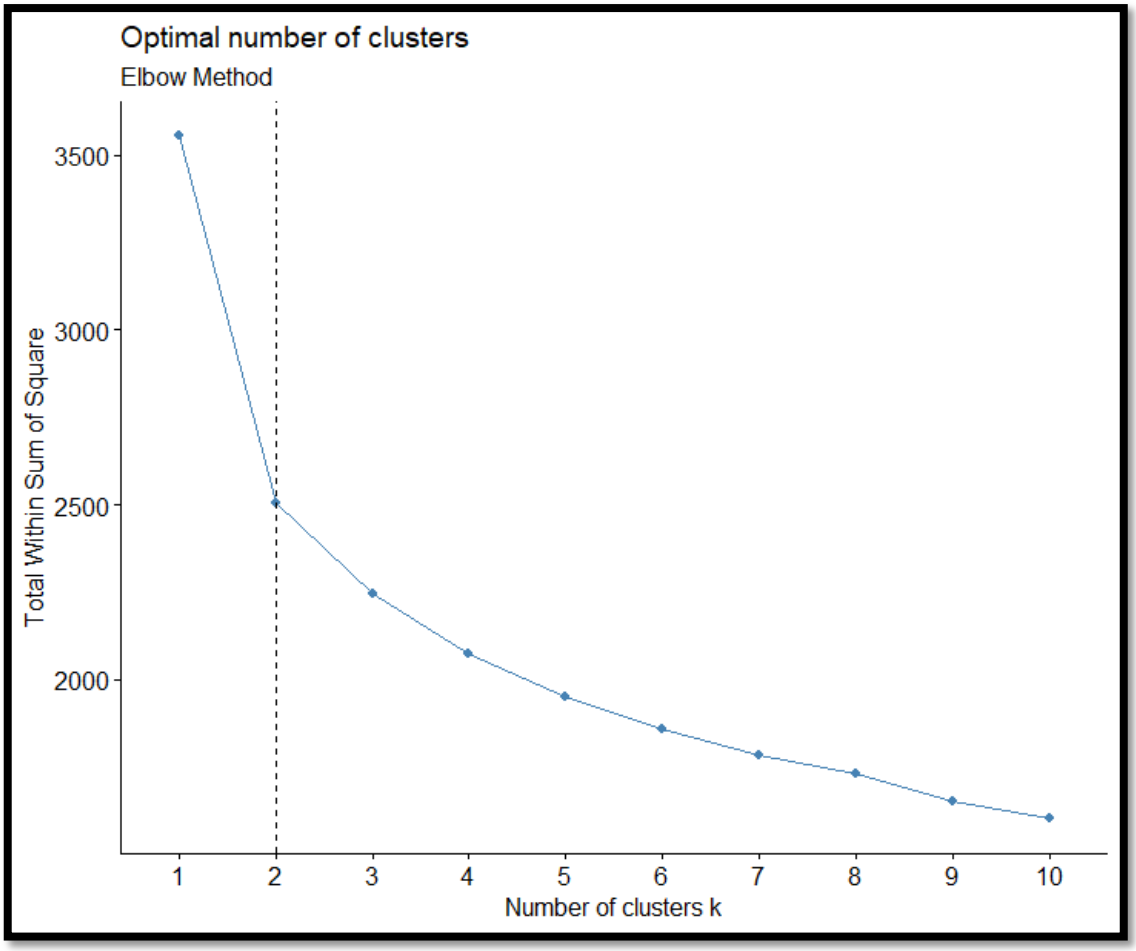
Για την εφαρμογή της μεθόδου του αγκώνα, εκτελούνται τα ακόλουθα βήματα :

1. Εκτελείται ο αλγόριθμος K-means για ένα εύρος τιμών του k, για παράδειγμα από 1 έως 10.
2. Για κάθε τιμή του k, υπολογίζεται το WCSS.
3. Σχεδιάζεται το WCSS ως συνάρτηση του k.
4. Αναζητείται ένας αγκώνας στο διάγραμμα, δηλαδή ένα σημείο όπου το WCSS αρχίζει να εξισώνεται ή να μειώνεται με μικρότερο ρυθμό.

(Tan et al., 2018)

Η διαίσθηση πίσω από τη μέθοδο του αγκώνα είναι ότι καθώς αυξάνεται ο αριθμός των συστάδων k, το WCSS θα πρέπει να μειώνεται, καθώς οι μικρότερες συστάδες είναι πιο «συμπαγείς». Ωστόσο, πέρα από έναν ορισμένο αριθμό συστάδων, η βελτίωση του WCSS γίνεται οριακή, καθώς οι συστάδες γίνονται πολύ μικρές για να έχουν νόημα. Η μέθοδος του αγκώνα επιδιώκει να προσδιορίσει αυτόν τον βέλτιστο αριθμό συστάδων. Είναι σημαντικό να σημειωθεί ότι η μέθοδος του αγκώνα δεν είναι

αλάνθαστη και μπορεί να μην δίνει πάντα σαφή ένδειξη του βέλτιστου αριθμού συστάδων. Στην πράξη, χρησιμοποιείται συχνά ως σημείο εκκίνησης, ενώ άλλες μέθοδοι, όπως η ανάλυση «σιλουέτας», μπορούν να χρησιμοποιηθούν για επικύρωση.



Σχήμα 6.7 : Γράφημα για τη μέθοδο του αγκώνα για τα playoffs.

Παρατηρώντας το παραπάνω γράφημα, συμπεραίνουμε πως ένα καλό πλήθος για τις αρχικές μας συστάδες ισούται με δυο, καθώς βλέπουμε ότι μετά από αυτή τη τιμή, η γραμμή που αντιστοιχεί στο WCSS έχει όλο και μικρότερη κλίση, καθώς μεγαλώνουν οι τιμές του k.

Μέθοδος «σιλουέτας» (Silhouette Method)

Η μέθοδος της «σιλουέτας» είναι μια άλλη τεχνική για την επιλογή του αριθμού των συστάδων στην συσταδοποίηση μέσω του αλγορίθμου K-means. Η μέθοδος αξιολογεί την ποιότητα της συσταδοποίησης, υπολογίζοντας έναν συντελεστή «σιλουέτας» (silhouette coefficient) για κάθε σημείο δεδομένων. Ο συντελεστής αυτός μετρά πόσο όμοιο είναι ένα σημείο δεδομένων με σημεία τη δικής του συστάδας, σε σύγκριση με σημεία άλλων συστάδων. Κυμαίνεται από -1 έως 1, με τις τιμές πιο κοντά στο 1 να υποδηλώνουν καλύτερη ομαδοποίηση.

Για τη μέθοδο σιλουέτας, εκτελούνται τα ακόλουθα βήματα :

1. Εκτελείται ο αλγόριθμος K-means για ένα εύρος τιμών του k, για παράδειγμα από 1 έως 10 συστάδες.
2. Για κάθε τιμή του k, υπολογίζεται ο μέσος συντελεστής σιλουέτας σε όλα τα σημεία δεδομένων.
3. Σχεδιάζεται ο μέσος συντελεστής σιλουέτας ως συνάρτηση του k.
4. Αναζητείται η μέγιστη τιμή του μέσου συντελεστή σιλουέτας, η οποία υποδεικνύει τον βέλτιστο αριθμό συστάδων.

(Tan et al., 2018)

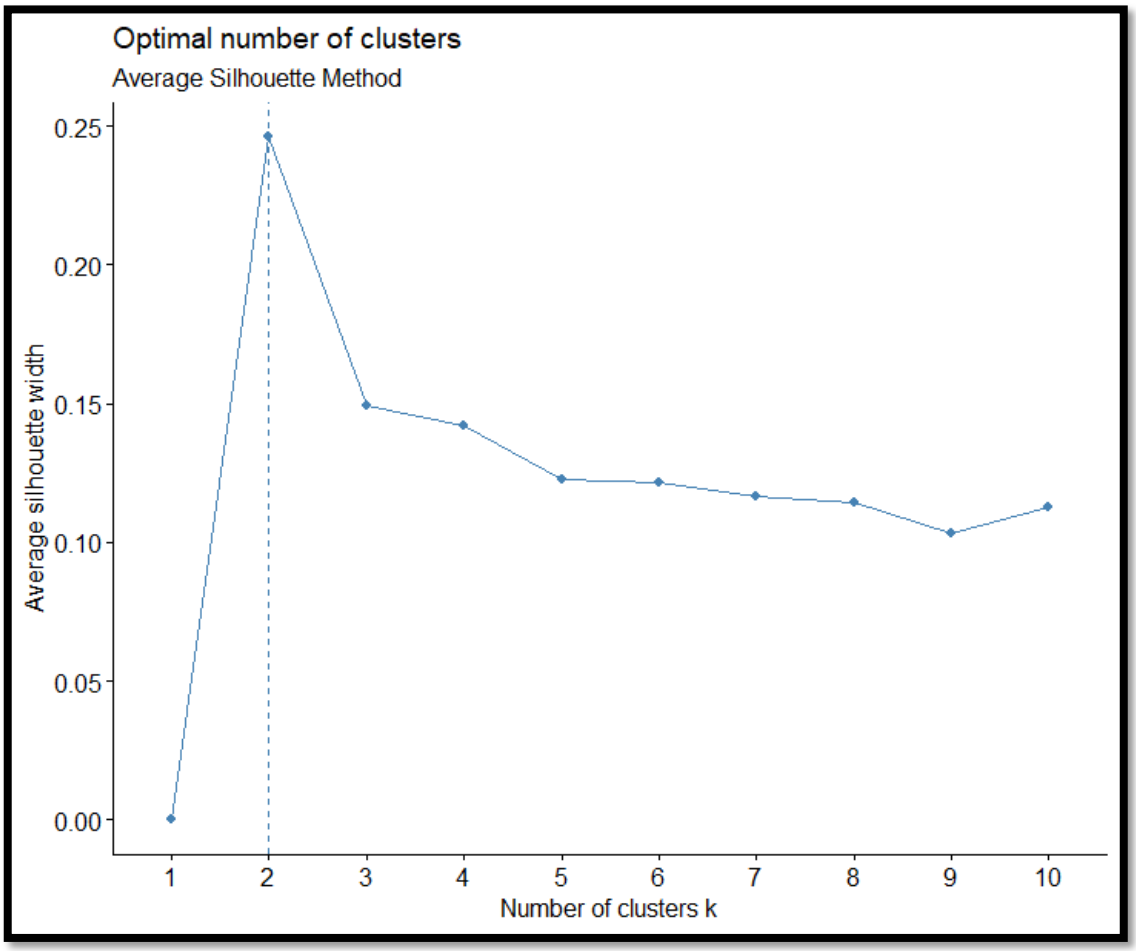
Αναφορικά τώρα με τον συντελεστή σιλουέτας, αυτός υπολογίζεται ως εξής. Υπολογίζουμε τη μέση απόσταση μεταξύ του σημείου i και όλων των άλλων σημείων που βρίσκονται στην ίδια συστάδα. Συμβολίζουμε την ποσότητα αυτή ως $a(i)$. Στη συνέχεια, υπολογίζουμε τη μέση απόσταση μεταξύ του σημείου και όλων των σημείων, που βρίσκονται στην πλησιέστερη γειτονική συστάδα. Η ποσότητα αυτή συμβολίζεται ως $b(i)$. Υπολογίζουμε τον συντελεστή σιλουέτας για το σημείο i ως

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Ο μέσος συντελεστής σιλουέτας σε όλα τα σημεία δεδομένων υπολογίζεται στη συνέχεια, ως ο μέσος όρος των συντελεστών σιλουέτας για κάθε σημείο δεδομένων.

Η διαίσθηση πίσω από τη μέθοδο της σιλουέτας είναι ότι για να υπάρχει μια καλή ομαδοποίηση, η μέση απόσταση μεταξύ των σημείων εντός μιας συστάδας πρέπει να είναι μικρή και η μέση απόσταση μεταξύ των σημείων σε διαφορετικές συστάδες πρέπει να είναι μεγάλη. Ο συντελεστής σιλουέτας μετρά πόσο καλά ικανοποιείται αυτή η ιδιότητα και μπορεί να χρησιμοποιηθεί για τη σύγκριση της ποιότητας της

συσταδοποίησης, για διάφορες τιμές του k. Όπως προαναφέρθηκε, μπορούμε να χρησιμοποιήσουμε τη μέθοδο της σιλουέτας προκειμένου να επιβεβαιώσουμε τα αποτελέσματα από τη μέθοδο του αγκώνα.



Σχήμα 6.8 : Γράφημα για τη μέθοδο της σιλουέτας για τα playoffs.

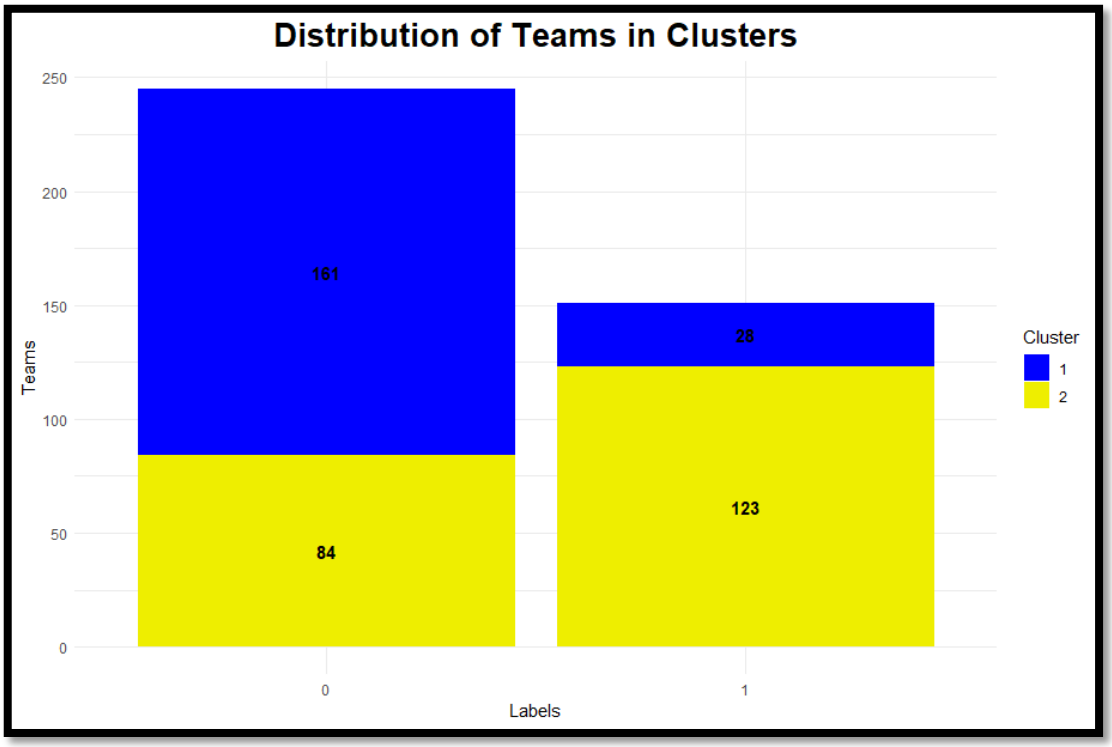
Παρατηρώντας το παραπάνω γράφημα, επιβεβαιώνουμε το συμπέρασμα στο οποίο καταλήξαμε παραπάνω με τη μέθοδο του αγκώνα. Συγκεκριμένα, βλέπουμε πως ένα καλό πλήθος για τις αρχικές μας συστάδες θα ήταν για k=2 αφού για αυτή τη τιμή, έχουμε τη μέγιστη τιμή του συντελεστή σιλουέτας.

Εφαρμόζοντας τον αλγόριθμο K-means, καταλήξαμε στο εξής συμπέρασμα σχετικά με τη κατανομή των δεδομένων/ομάδων μας στις δύο συστάδες :

- Από τις 245 ομάδες που δε προκρίθηκαν στα playoffs, οι 161 ομάδες βρέθηκαν στην $1^{η}$ συστάδα (cluster 1) και οι 84 ομάδες βρέθηκαν στην $2^{η}$ συστάδα (cluster 2).

- Από τις 151 ομάδες που προκρίθηκαν στα playoffs, οι 28 ομάδες βρέθηκαν στην $1^{η}$ συστάδα (cluster 1) και οι 123 ομάδες βρέθηκαν στην $2^{η}$ συστάδα (cluster 2).

Το παραπάνω συμπέρασμα φαίνεται και στο παρακάτω stacked bar plot.



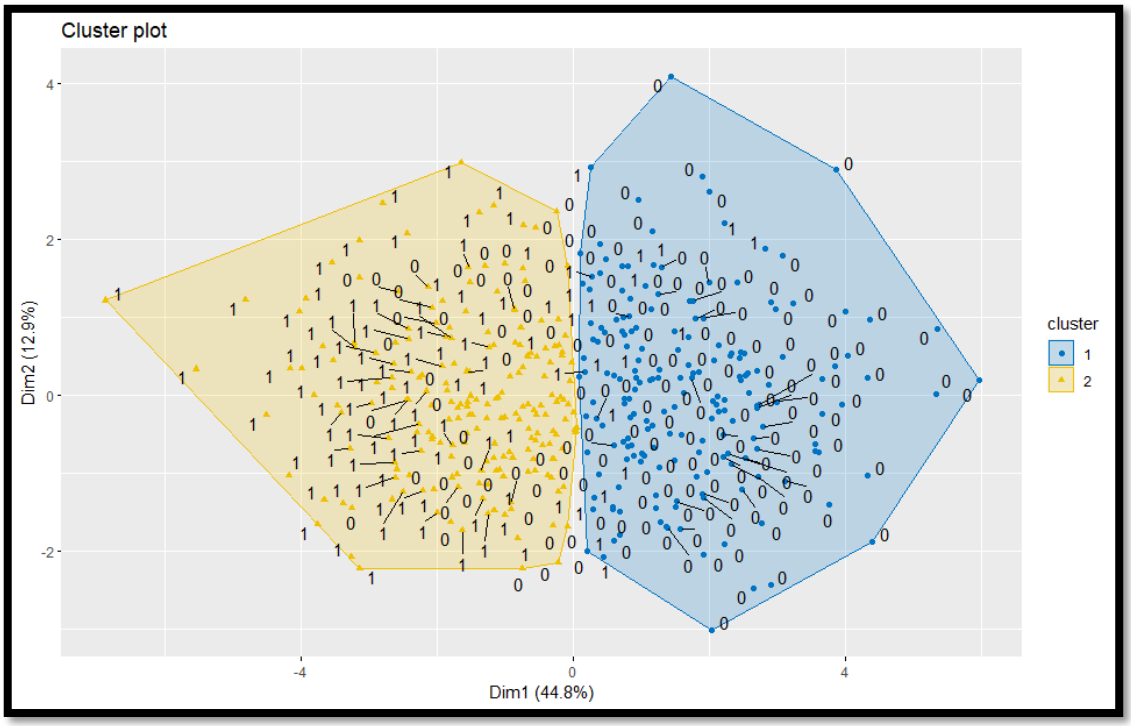
Σχήμα 6.9 : Stacked bar plot για τα playoffs.

Τώρα, θέλουμε να απεικονίσουμε τα δεδομένα μας, με τέτοιο τρόπο ώστε να φαίνονται οι δύο συστάδες που δημιούργησε ο αλγόριθμος μας. Το πρόβλημα είναι ότι το σύνολο των δεδομένων μας περιέχει περισσότερες από δύο μεταβλητές, και έτσι δε μπορούμε να χρησιμοποιήσουμε ένα κλασικό διάγραμμα διασποράς, ώστε να έχουμε ένα ικανοποιητικό αποτέλεσμα. Μια λύση για αυτό το πρόβλημα συνεπώς, είναι η μείωση του αριθμού των διαστάσεων, με την εφαρμογή ενός αλγορίθμου μείωσης διαστασιμότητας, όπως η ανάλυση κύριων συνιστωσών (PCA), που δημιουργεί νέες μεταβλητές, οι οποίες είναι γραμμικοί συνδυασμοί των αρχικών μας μεταβλητών και προσπαθούν να ερμηνεύσουν (μεταφέρουν) όσο το δυνατόν περισσότερη πληροφορία (μεταβλητότητα) από το αρχικό σύνολο δεδομένων. Η πρώτη κύρια συνιστώσα ερμηνεύει το μεγαλύτερο ποσοστό της μεταβλητότητας, η δεύτερη το αμέσως επόμενο κ.ο.κ.

Το scatter plot από το πακέτο ‘factoextra’ , το οποίο εμείς θα κάνουμε χρήση, χρησιμοποιεί αυτομάτως τις δύο πρώτες κύριες συνιστώσες, προκειμένου να αναπαραστήσει την εκάστοτε συσταδοποίηση.

Ο αλγόριθμος K-Means για δύο αρχικά κέντρα μας έδωσε τη συσταδοποίηση, όπως φαίνεται στο παρακάτω γράφημα. Παρουσιάζονται τα δεδομένα σε κάθε συστάδα, μαζί με ενδεικτικά (λόγω μεγάλου πλήθους) σε ορισμένα, τη κατηγορία στην οποία ανήκαν εξ αρχής, με βάση το αν προκρίθηκαν ή όχι στα playoffs. Η μια συστάδα που δημιουργήθηκε περιλαμβάνει 189 παρατηρήσεις, ενώ η άλλη περιλαμβάνει 207.

Παράλληλα, βλέπουμε πως η πρώτη κύρια συνιστώσα ερμηνεύει το 44,8% της μεταβλητότητας των αρχικών μας δεδομένων, ενώ η δεύτερη το 12,9%.



Σχήμα 6.10 : Cluster plot για τη συσταδοποίηση μέσω του K-means, για τα playoffs.

Μέτρα Αξιολόγησης της συσταδοποίησης

Υπάρχουν διάφοροι μέθοδοι για την αξιολόγηση της ομαδοποίησης, συμπεριλαμβανομένων των εσωτερικών, των εξωτερικών και των σχετικών μέτρων. Η επιλογή του μέτρου αξιολόγησης εξαρτάται από τη εκάστοτε εφαρμογή και τους στόχους της συσταδοποίησης.

- Τα **εσωτερικά μέτρα** αξιολογούν την ποιότητα της ομαδοποίησης, με βάση αποκλειστικά τα δεδομένα και τον αλγόριθμο ομαδοποίησης, χωρίς να χρησιμοποιούν εξωτερικές πληροφορίες ή τις ετικέτες/κλάσεις που συνοδεύουν κάθε σημείο/παρατήρηση. Παραδείγματα εσωτερικών μέτρων περιλαμβάνουν το άθροισμα τετραγώνων εντός της συστάδας (WCSS) και τον συντελεστή σιλουέτας, τα οποίο είδαμε παραπάνω.

- Τα **εξωτερικά μέτρα** αξιολογούν την ποιότητα της συσταδοποίησης συγκρίνοντας τα αποτελέσματα της συσταδοποίησης με κάποιο εξωτερικό κριτήριο, όπως οι ετικέτες των κλάσεων στις οποίες ανήκουν τα σημεία. Παράδειγμα εξωτερικού μέτρου είναι ο προσαρμοσμένος δείκτη Rand.

- Τα **<u>σχετικά μέτρα</u>** αξιολογούν την ποιότητα της συσταδοποίησης συγκρίνοντας τα αποτελέσματα διαφορετικών αλγορίθμων συσταδοποίησης, ή διαφορετικών ρυθμίσεων για τις παραμέτρους του ίδιου αλγορίθμου. Παραδείγματα τέτοιων σχετικών μέτρων είναι ο δείκτης Dunn και ο δείκτης Davies-Bouldin.

Εμείς χρησιμοποιήσαμε για την αξιολόγηση των αποτελεσμάτων μας το συντελεστή σιλουέτας, το προσαρμοσμένο δείκτη Rand και το δείκτη Dunn.

<u>Προσαρμοσμένος δείκτης Rand</u>

Ο προσαρμοσμένος δείκτης Rand (Adjusted Rand Index - ARI) είναι ένα μέτρο συμφωνίας μεταξύ δύο ομαδοποιήσεων ενός συνόλου δεδομένων, παραλλαγή του δείκτη Rand, ο οποίος μετρά το ποσοστό των ζευγαριών των σημείων δεδομένων που αντιστοιχούν στις ίδιες ή διαφορετικές συστάδες, και στις δύο ομαδοποιήσεις. Ο ARI λαμβάνει υπόψιν την τύχη, και ειδικότερα την πιθανότητα συμφωνίας λόγω τύχης. Αυτό είναι σημαντικό, επειδή ο δείκτης Rand μπορεί να παράγει υψηλές τιμές ακόμη και όταν η συμφωνία οφείλεται στην τύχη, ιδίως όταν ο αριθμός των συστάδων είναι μεγάλος. Ο δείκτης κυμαίνεται από -1 έως 1, με τις τιμές κοντά στο 1 να υποδηλώνουν υψηλή συμφωνία μεταξύ των δύο ομαδοποιήσεων, τιμές κοντά στο 0 να υποδηλώνουν τυχαία συμφωνία και τιμές κοντά στο -1 να υποδηλώνουν διαφωνία μεταξύ των δύο ομαδοποιήσεων. Ορίζεται ως εξής :

$$ARI = \frac{RI - Expected\_RI}{\max\{RI\} - Expected\_RI}$$

όπου RI είναι ο δείκτης Rand, $Expected\_RI$ είναι η αναμενόμενη τιμή του δείκτη Rand υπό την υπόθεση της ανεξαρτησίας μεταξύ των δύο ομαδοποιήσεων και $\max\{RI\}$ είναι η μέγιστη δυνατή τιμή του δείκτη Rand.

Ο ARI έχει πολλά πλεονεκτήματα έναντι άλλων μέτρων αξιολόγησης της συσταδοποίησης. Πρώτον, δεν επηρεάζεται από τον αριθμό των συστάδων ή το μέγεθος του συνόλου δεδομένων, σε αντίθεση με ορισμένα άλλα μέτρα. Δεύτερον, είναι συμμετρικός, που σημαίνει ότι δεν ευνοεί τη μία ομαδοποίηση έναντι της άλλης. Τρίτον, λαμβάνει υπόψη την πιθανότητα συμφωνίας κατά τύχη, καθιστώντας το πιο ανθεκτικό στο θόρυβο και στις τυχαίες διακυμάνσεις των δεδομένων. (Tan et al., 2018)

Δείκτης Dunn

Ο δείκτης Dunn είναι ένα μέτρο εγκυρότητας των συστάδων που χρησιμοποιείται για την αξιολόγηση της ποιότητας των αποτελεσμάτων της συσταδοποίησης. Μετρά τη συμπαγή δομή των συστάδων και τον διαχωρισμό μεταξύ των συστάδων και ορίζεται ως ο λόγος της ελάχιστης απόστασης μεταξύ των συστάδων, προς τη μέγιστη απόσταση εντός των συστάδων. Όσο υψηλότερος είναι ο δείκτης Dunn, τόσο καλύτερα είναι τα αποτελέσματα της συσταδοποίησης, καθώς δείχνει ότι οι συστάδες είναι συμπαγείς και καλά διαχωρισμένες. Συγκεκριμένα, ορίζεται ως :

$$DI = \frac{\min\{d(i,j)\}}{\max\{D(k)\}}$$

όπου $d(i,j)$ είναι η απόσταση μεταξύ του i-οστού και του j-οστού σημείου δεδομένων, $D(k)$ είναι η διάμετρος της k-οστής συστάδας και n είναι ο συνολικός αριθμός των σημείων δεδομένων. Ισχύει $1 \leq i < j \leq n$.

Ακολουθεί ο πίνακας με της τιμές των μέτρων αξιολόγησης για την ομαδοποίηση της για τα Playoffs.

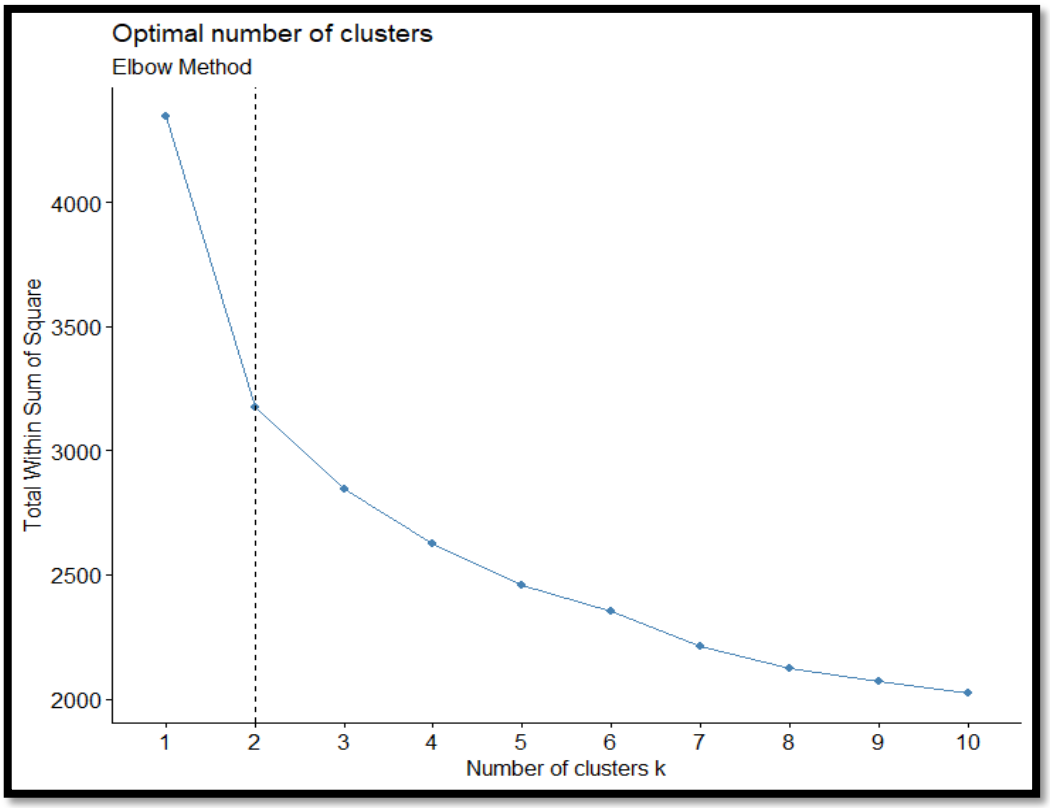
Μέτρα Αξιολόγησης	Τιμές
Silhouette Coefficient	**0,246**
Adjusted Rand Index	**0,796**
Dunn Index	**0,103**

Πίνακας 6.2 : Μέτρα αξιολόγησης για την ομαδοποίηση των playoffs.

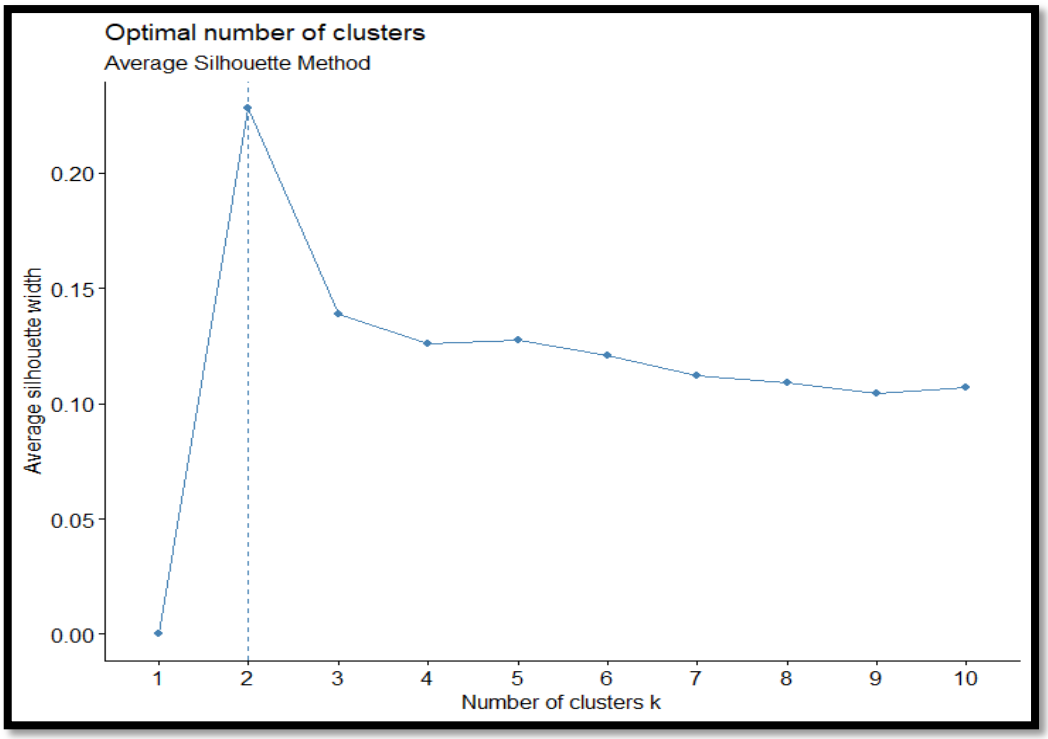
Σύμφωνα με το παραπάνω πίνακα, η ομαδοποίηση της φαίνεται να είναι αρκετά καλή με βάση τον προσαρμοσμένο δείκτη Rand, αλλά όχι αποτελεσματική με βάση τα άλλα δύο μέτρα.

6.4.2 Εφαρμογή αλγορίθμου K-Means για τη φάση του Final Four

Στη συνέχεια, ακολουθήσαμε ακριβώς την ίδια διαδικασία με την προηγούμενη ενότητα, απλώς τώρα για τη φάση του Final Four. Αρχικά, τόσο μέσω της μεθόδου του αγκώνα, όσο και μέσω της μεθόδου της σιλουέτας, καταλήξαμε πως πάλι το βέλτιστο πλήθος αρχικών συστάδων ισούταν με δύο.



Σχήμα 6.11 : Γράφημα για τη μέθοδο του αγκώνα για το Final Four.



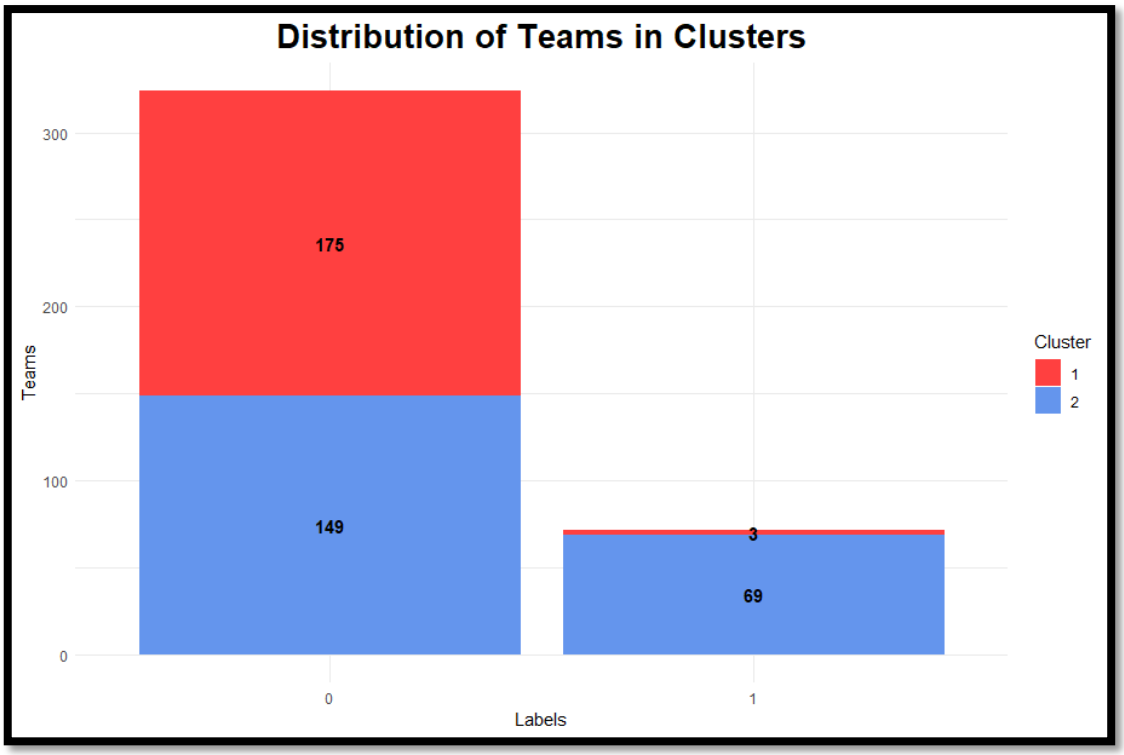
Σχήμα 6.12 : Γράφημα για τη μέθοδο της σιλουέτας για το Final Four.

Εφαρμόζοντας τον αλγόριθμο, καταλήξαμε στο εξής συμπέρασμα σχετικά με τη κατανομή των δεδομένων/ομάδων μας στις δύο κλάσεις :

- Από τις 324 ομάδες που δε προκρίθηκαν στο Final Four, οι 175 ομάδες βρέθηκαν στην $1^{η}$ συστάδα (cluster 1) και οι 149 ομάδες βρέθηκαν στην $2^{η}$ συστάδα (cluster 2).

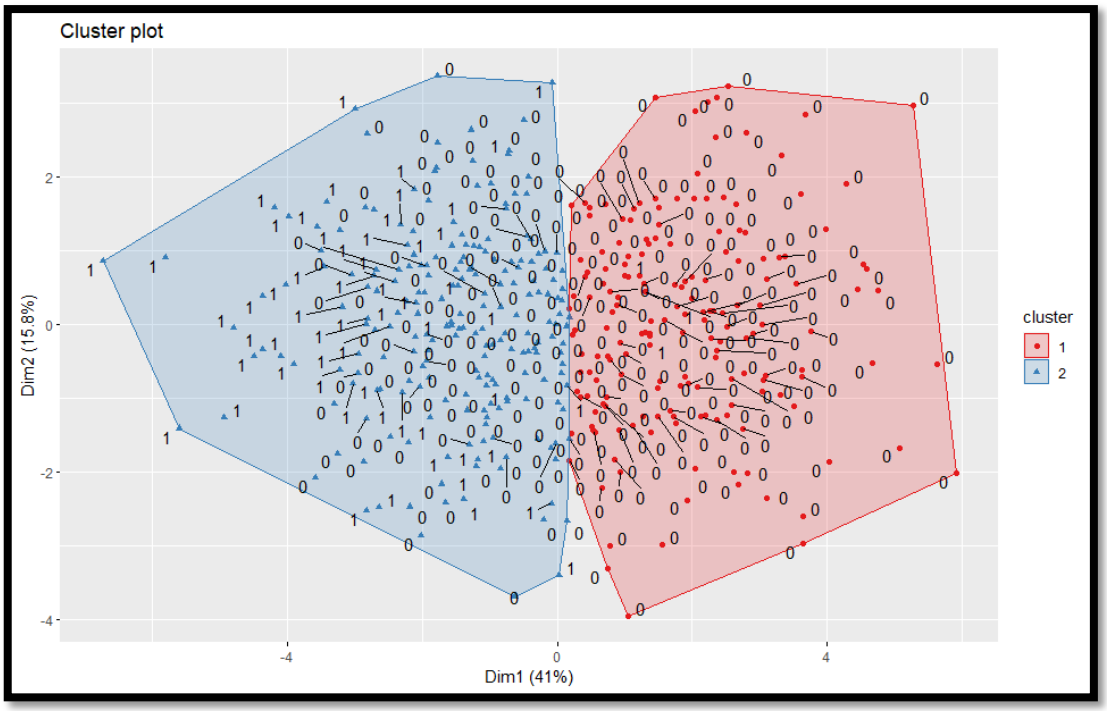
- Από τις 72 ομάδες που προκρίθηκαν στο Final Four, μόλις 3 ομάδες βρέθηκαν στην $1^{η}$ συστάδα (cluster 1) και οι 69 υπόλοιπες ομάδες βρέθηκαν στην $2^{η}$ συστάδα (cluster 2).

Το παραπάνω συμπέρασμα φαίνεται και στο παρακάτω stacked bar plot.



Σχήμα 6.13 : Stacked bar plot για το Final Four.

Συνεπώς, εφαρμόσαμε τον αλγόριθμο K-Means πάλι για δύο αρχικά κέντρα για τα δεδομένα μας και πήραμε τη παρακάτω ομαδοποίηση. Η μια συστάδα που δημιουργήθηκε περιλαμβάνει 178 παρατηρήσεις, ενώ η άλλη περιλαμβάνει 218. Παράλληλα, βλέπουμε πως η πρώτη κύρια συνιστώσα ερμηνεύει το 41% της μεταβλητότητας των αρχικών μας δεδομένων, ενώ η δεύτερη το 15,8%.



Σχήμα 6.14 : Cluster plot για τη συσταδοποίηση μέσω του K-means, για το Final Four.

Τέλος, ακολουθεί ο πίνακας με τις τιμές των μέτρων αξιολόγησης για την ομαδοποίηση, για τη φάση του Final Four αυτή τη φορά.

Μέτρα Αξιολόγησης	Τιμές
Silhouette Coefficient	**0,228**
Adjusted Rand Index	**0,956**
Dunn Index	**0,148**

Πίνακας 6.3 : Μέτρα αξιολόγησης για την ομαδοποίηση του Final Four.

Σύμφωνα με τον παραπάνω πίνακα, η ομαδοποίηση μας φαίνεται να είναι πάρα πολύ καλή και πάλι με βάση τον προσαρμοσμένο δείκτη Rand, αλλά καθόλου αποτελεσματική με βάση τα άλλα δύο μέτρα.

6.4.3 Εφαρμογή αλγορίθμου ιεραρχικής, συσσωρευτικής συσταδοποίησης για τη φάση των playoffs

Η ιεραρχική, συσσωρευτική ομαδοποίηση (Hierarchical Agglomerative Clustering ή Agglomerative Nesting - agnes) είναι μια τεχνική ομαδοποίησης η οποία ξεκινά με κάθε σημείο δεδομένων να αποτελεί τη δική του συστάδα, και συγχωνεύει επαναληπτικά το πλησιέστερο ζεύγος συστάδων μέχρι όλα τα σημεία να ανήκουν σε μια ενιαία συστάδα. Η ιεραρχική συσσωρευτική συσταδοποίηση έχει πολλά πλεονεκτήματα, όπως η απλότητα και η ικανότητά της να χειρίζεται διαφορετικούς

τύπους δεδομένων. Ωστόσο, μπορεί να είναι υπολογιστικά δαπανηρή για μεγάλα σύνολα δεδομένων και τα αποτελέσματά της μπορεί να είναι ευαίσθητα στην επιλογή του κατάλληλου μέτρου απόστασης και της μεθόδου σύνδεσης, στις οποίες αναφερόμαστε παρακάτω.

Αυτός ο αλγόριθμος μπορεί να αναπαρασταθεί με τη χρήση ενός δενδρογράμματος, το οποίο είναι ένα δενδροειδές διάγραμμα που δείχνει τη σειρά και τις αποστάσεις των συστάδων, κατά τη διάρκεια της διαδικασίας.

Τα βασικά βήματα για την ιεραρχική συσσωρευτική συσταδοποίηση είναι τα εξής :

1. Αρχικοποίηση κάθε σημείου δεδομένων ως μια συστάδα.
2. Υπολογισμός του πίνακα απόστασης μεταξύ όλων των ζευγών συστάδων.
3. Προσδιορισμός του πλησιέστερου ζεύγους συστάδων, με βάση κάποιο μέτρο απόστασης, για τα οποία κάναμε αναφορά παραπάνω, όπως π.χ. την Ευκλείδεια απόσταση.
4. Συγχώνευση των δύο πλησιέστερων συστάδων σε μια ενιαία συστάδα.
5. Ενημέρωση του πίνακα αποστάσεων, ώστε να αντικατοπτρίζει τις νέες αποστάσεις μεταξύ της συγχωνευμένης συστάδας και των υπόλοιπων συστάδων.
6. Επανάληψη των βημάτων 3-5 έως ότου όλα τα σημεία δεδομένων ανήκουν σε μια ενιαία συστάδα, ή έως ότου ικανοποιηθεί ένα κριτήριο διακοπής (π.χ. να προκύψει ένας επιθυμητός αριθμός συστάδων).

Η απόσταση μεταξύ των συστάδων μπορεί να υπολογιστεί με διάφορες μεθόδους σύνδεσης. Οι πιο συνηθισμένες μέθοδοι σύνδεσης είναι οι εξής :

- Ελάχιστη ή απλή σύνδεση (minimum or single linkage) : Στην απλή σύνδεση, η απόσταση μεταξύ δύο συστάδων ορίζεται ως η ελάχιστη απόσταση μεταξύ δύο οποιωνδήποτε σημείων εντός των δύο συστάδων. Τείνει να παράγει μεγάλες, «χαλαρές» συστάδες.

- Μέγιστη ή πλήρης σύνδεση (maximum or complete linkage) : Στην πλήρη σύνδεση, η απόσταση ορίζεται ως η μέγιστη απόσταση μεταξύ δύο οποιωνδήποτε σημείων μέσα στις συστάδες. Τείνει να παράγει πιο συμπαγείς συστάδες.

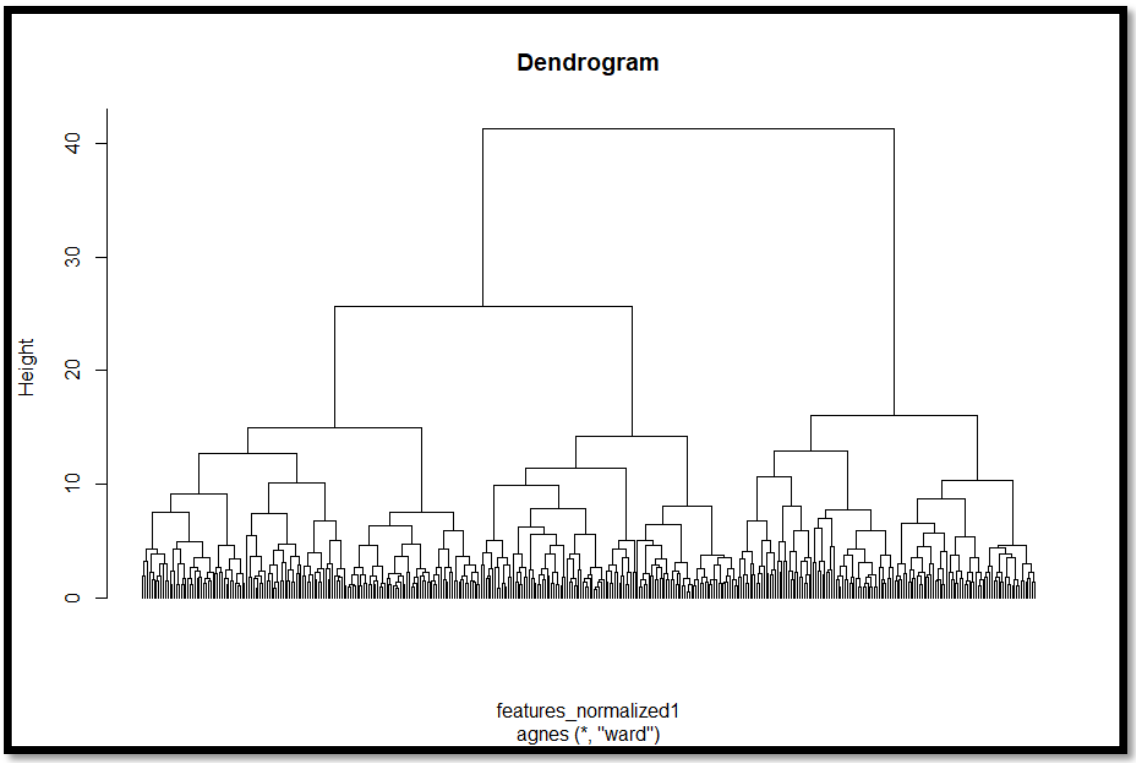
- Μέση σύνδεση (mean or average linkage) : Στη μέση σύνδεση, η απόσταση ορίζεται ως η μέση απόσταση μεταξύ όλων των ζευγών σημείων στις συστάδες.

- Μέθοδος των κέντρων βάρους (centroid linkage) : Αυτή η μέθοδος υπολογίζει τα κέντρα βάρους των δύο συστάδων, και ορίζει ως απόσταση τους, την απόσταση των κέντρων αυτών.

- Μέθοδος του Ward : Η μέθοδος ελαχιστοποιεί τη συνολική διακύμανση εντός της συστάδας. Σε κάθε βήμα, οι δύο συστάδες με την ελάχιστη απόσταση, συγχωνεύονται. Τείνει να παράγει συμπαγείς και σφαιρικές συστάδες. Ωστόσο, μπορεί να είναι ευαίσθητη στις ακραίες τιμές και μπορεί να μην λειτουργεί καλά με μη σφαιρικές συστάδες, ή με μικτούς τύπους δεδομένων. Εμείς και για τις δύο φάσεις της διοργάνωσης που εξετάζουμε, χρησιμοποιήσαμε τη μέθοδο αυτή.

(Tan et al., 2018)

Εκτελώντας λοιπόν το αλγόριθμο της ιεραρχικής συσσωρευτικής συσταδοποίησης με συνάρτηση σύνδεσης του Ward και για τη φάση των playoffs, πήραμε το παρακάτω δενδρόγραμμα.

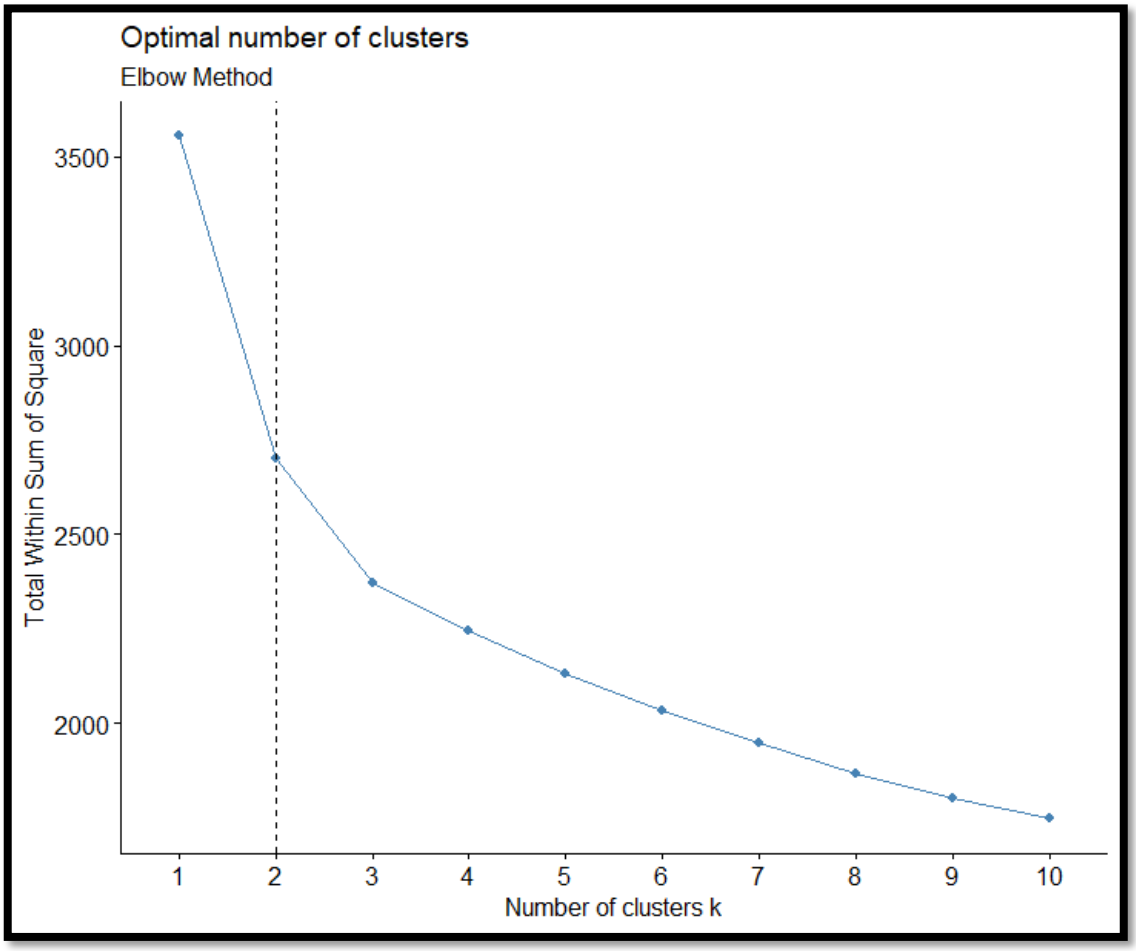


Σχήμα 6.15 : Δενδρόγραμμα για τα playoffs.

Τα δενδρογράμματα είναι ιεραρχικά διαγράμματα που χρησιμοποιούνται για την αναπαράσταση των σχέσεων μεταξύ αντικειμένων ή μεταβλητών. Κάθε παρατήρηση στο κάτω μέρος του διαγράμματος αποτελεί και μια συστάδα μόνη της, και όσο ανεβαίνουμε, καταλήγουμε σε μια ενιαία συστάδα. Προκειμένου να σχηματίσουμε

εμείς συστάδες με στοιχεία που έχουν παρόμοια χαρακτηριστικά είναι αναγκαία η διαίρεση των κλάδων του διαγράμματος σε συγκεκριμένο ύψος. Η απόφαση για την κοπή ενός δενδρογράμματος εξαρτάται από το ερευνητικό ερώτημα και τον επιθυμητό αριθμό συστάδων. Η κοπή σε υψηλότερο επίπεδο οδηγεί σε λιγότερες το πλήθος, αλλά μεγαλύτερες συστάδες, ενώ η κοπή σε χαμηλότερο επίπεδο δημιουργεί περισσότερες το πλήθος, μικρότερες συστάδες. Μια κατάλληλη μέθοδος για την κοπή των δενδρογραμμάτων είναι η μέθοδος του αγκώνα, την οποία είδαμε αναλυτικότερα στις προηγούμενες ενότητες. Η κοπή των δενδρογραμμάτων είναι ένα ουσιαστικό βήμα στην ανάλυση συσταδοποίησης, καθώς παρέχει πληροφορίες για την υποκείμενη δομή των δεδομένων και βοηθά στον εντοπισμό σημαντικών μοτίβων.

Εφαρμόζοντας τη μέθοδο του αγκώνα, παρατηρήσαμε πως πρέπει να κόψουμε το δενδρόγραμμα με τέτοιο τρόπο, ώστε να δημιουργηθούν δύο συστάδες.



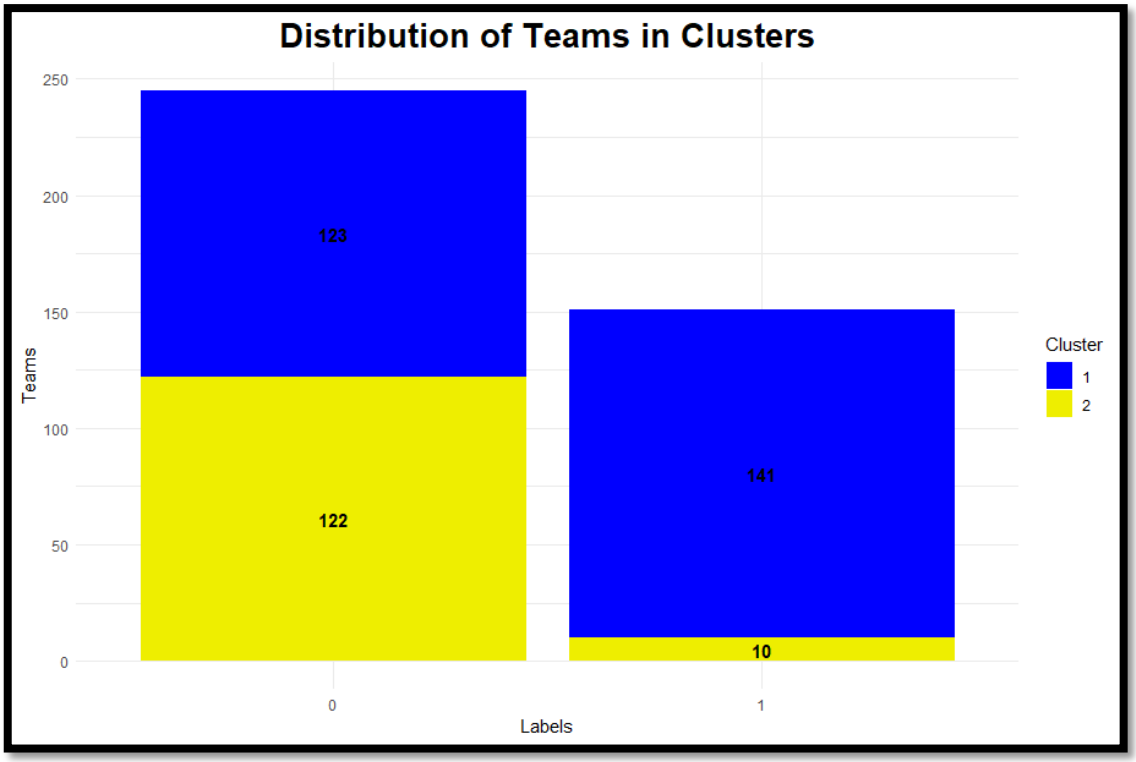
Σχήμα 6.16 : Γράφημα για τη μέθοδο του αγκώνα για τα playoffs.

Εφαρμόζοντας τον αλγόριθμο, καταλήξαμε στο εξής συμπέρασμα σχετικά με τη κατανομή των δεδομένων/ομάδων μας στις δύο κλάσεις :

- Από τις 245 ομάδες που δε προκρίθηκαν στα playoffs, οι 123 ομάδες βρέθηκαν στην 1η συστάδα (cluster 1) και οι 122 ομάδες βρέθηκαν στην 2η συστάδα (cluster 2).

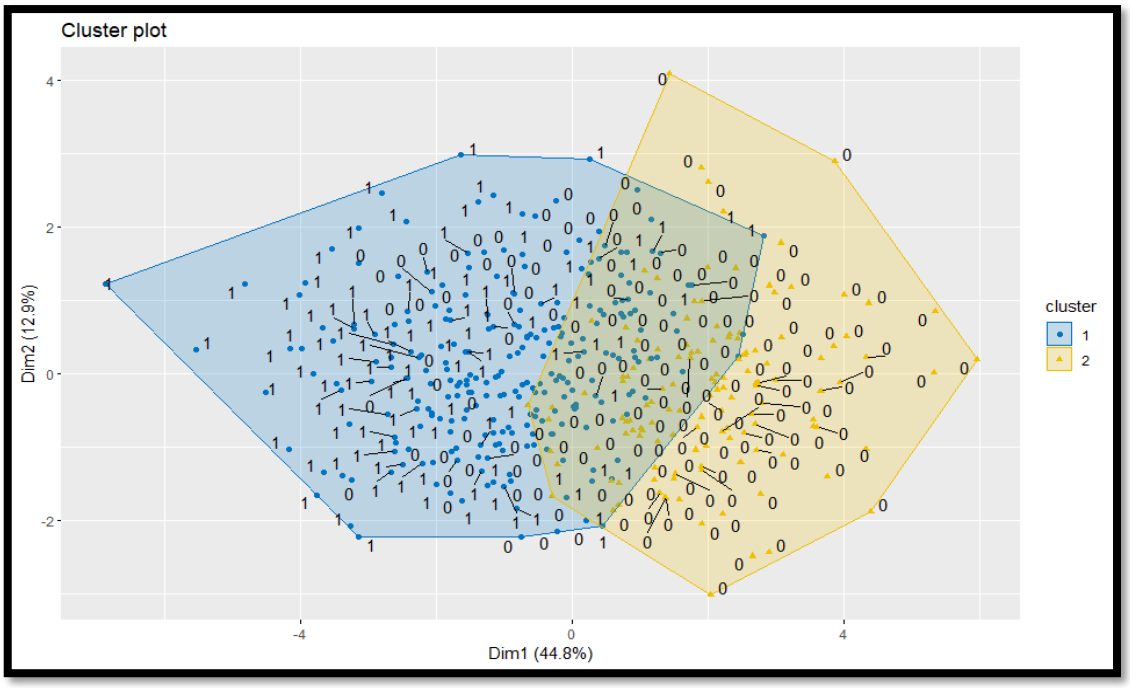
- Από τις 151 ομάδες που προκρίθηκαν στα playoffs, οι 141 ομάδες βρέθηκαν στην 1η συστάδα (cluster 1) και οι υπόλοιπες 10 βρέθηκαν στην 2η συστάδα (cluster 2).

Το παραπάνω συμπέρασμα φαίνεται και στο παρακάτω stacked bar plot.



Σχήμα 6.17 : Stacked bar plot για τα playoffs.

Κόβοντας λοιπόν κατάλληλα το δενδρόγραμμα, έχουμε δύο συστάδες, οι οποίες αναπαρίστανται στο παρακάτω cluster plot. Φαίνεται πως η μια συστάδα «επικαλύπτει» αρκετά την άλλη. Η μια περιλαμβάνει 264 παρατηρήσεις, ενώ η άλλη περιλαμβάνει 132. Παράλληλα, βλέπουμε πως η πρώτη κύρια συνιστώσα ερμηνεύει το 44,8% της μεταβλητότητας των αρχικών μας δεδομένων, ενώ η δεύτερη το 12,9%.



Σχήμα 6.18 : Cluster plot για τη συσταδοποίηση μέσω της agnes, για τα playoffs.

Ακολουθεί τώρα ο πίνακας με τις τιμές των μέτρων αξιολόγησης για την ομαδοποίηση μας, για τα Playoffs.

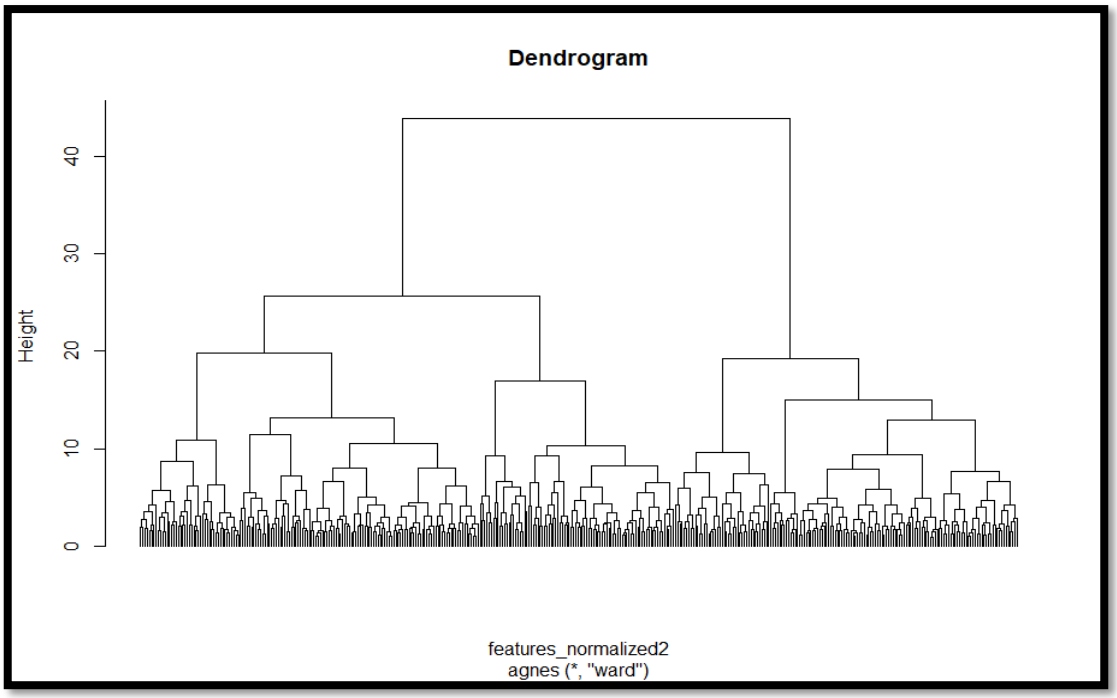
Μέτρα Αξιολόγησης	Τιμές
Silhouette Coefficient	**0,218**
Adjusted Rand Index	**0,923**
Dunn Index	**0,118**

Πίνακας 6.4 : Μέτρα αξιολόγησης για την ομαδοποίηση των playoffs.

Σύμφωνα με τον παραπάνω πίνακα, η ομαδοποίηση μας φαίνεται να είναι εξαιρετικά καλή με βάση τον προσαρμοσμένο δείκτη Rand, αλλά όχι αποτελεσματική με βάση τα άλλα δύο μέτρα.

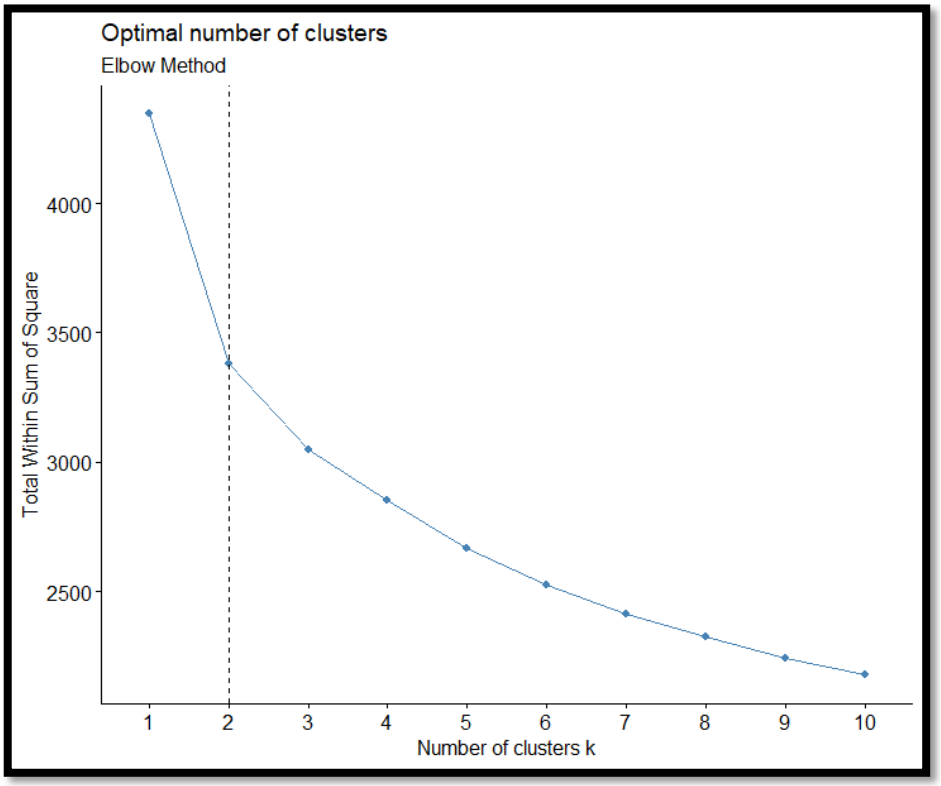
6.4.4 Εφαρμογή αλγορίθμου ιεραρχικής, συσσωρευτικής συσταδοποίησης για τη φάση του Final Four

Σε αυτό το σημείο, ακολουθήσαμε ακριβώς την ίδια διαδικασία με την προηγούμενη ενότητα, απλώς τώρα για τη φάση του Final Four. Εφαρμόσαμε τη μέθοδο της ιεραρχικής συσσωρευτικής συσταδοποίησης, με συνάρτηση σύνδεσης του Ward και πήραμε το παρακάτω δενδρόγραμμα.



Σχήμα 6.19 : Δενδρόγραμμα για το Final Four.

Χρησιμοποιώντας τώρα τη μέθοδο του αγκώνα, καταλήξαμε πως πάλι το βέλτιστο πλήθος συστάδων ισούταν με δύο.



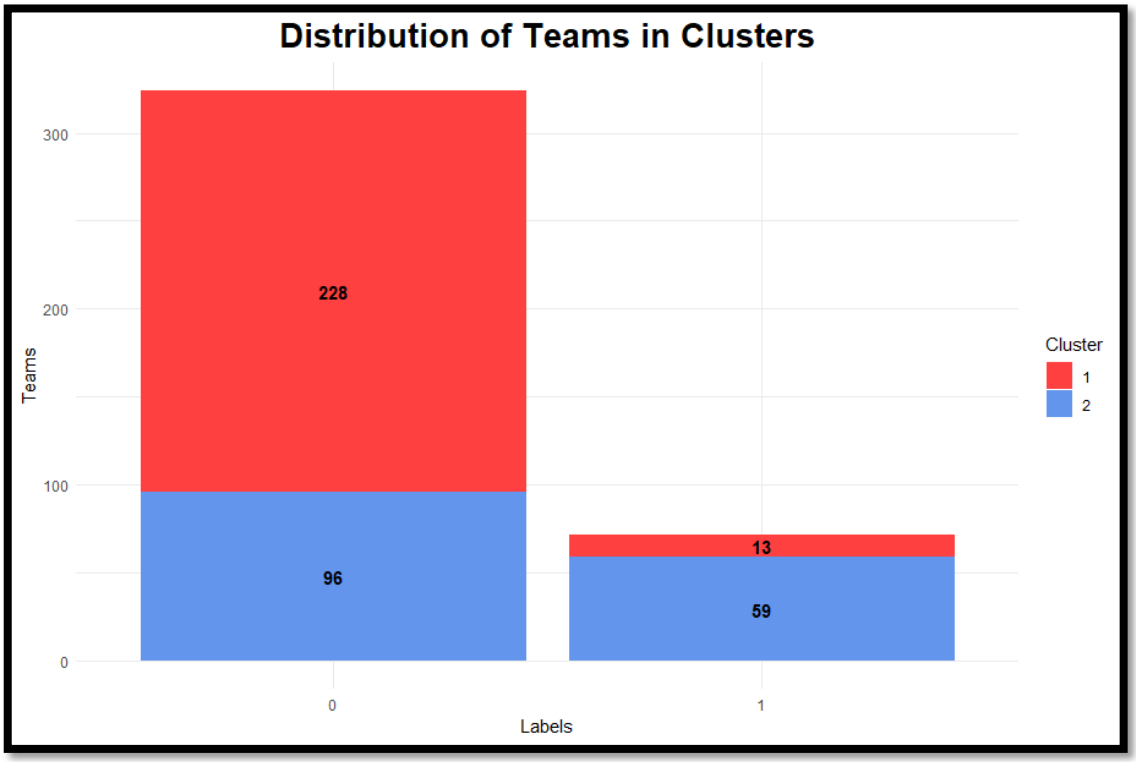
Σχήμα 6.20 : Γράφημα για τη μέθοδο του αγκώνα για το Final Four.

Εφαρμόζοντας τον αλγόριθμο, καταλήξαμε στο εξής συμπέρασμα σχετικά με τη κατανομή των δεδομένων/ομάδων μας στις δύο κλάσεις :

- Από τις 324 ομάδες που δε προκρίθηκαν στο Final Four, οι 228 ομάδες βρέθηκαν στην 1η συστάδα (cluster 1) και οι 96 ομάδες βρέθηκαν στην 2η συστάδα (cluster 2).

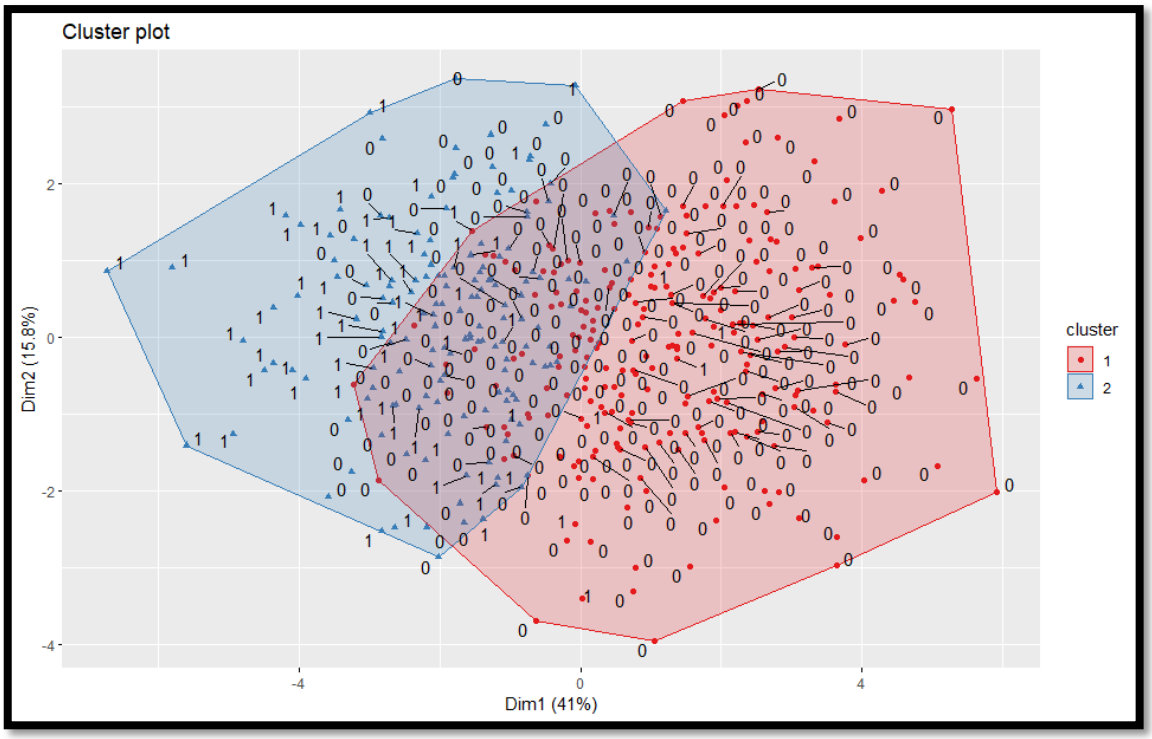
- Από τις 72 ομάδες που προκρίθηκαν στο Final Four, οι 13 ομάδες βρέθηκαν στην 1η συστάδα (cluster 1) και οι 59 υπόλοιπες ομάδες βρέθηκαν στην 2η συστάδα (cluster 2).

Το παραπάνω συμπέρασμα φαίνεται και στο παρακάτω stacked bar plot.



Σχήμα 6.21 : Stacked bar plot για το Final Four.

Οι δύο συστάδες που προέκυψαν, αναπαρίστανται στο παρακάτω cluster plot. Φαίνεται η μια συστάδα να «επικαλύπτει» την άλλη και σε αυτή τη περίπτωση. Η μια περιλαμβάνει 241 παρατηρήσεις, ενώ η άλλη περιλαμβάνει 155. Παράλληλα, βλέπουμε πως η πρώτη κύρια συνιστώσα ερμηνεύει το 41% της μεταβλητότητας των αρχικών μας δεδομένων, ενώ η δεύτερη το 15,8%.



Σχήμα 6.22 : Cluster plot για τη συσταδοποίηση μέσω της agnes, για το Final Four.

Τέλος, ακολουθεί ο πίνακας με τις τιμές των μέτρων αξιολόγησης για την ομαδοποίηση, για τη φάση του Final Four τώρα.

Μέτρα Αξιολόγησης	Τιμές
Silhouette Coefficient	**0,182**
Adjusted Rand Index	**0,819**
Dunn Index	**0,127**

Πίνακας 6.5 : Μέτρα αξιολόγησης για την ομαδοποίηση του Final Four.

Σύμφωνα με τον παραπάνω πίνακα, η ομαδοποίηση μας φαίνεται να είναι καλή και πάλι με βάση τον προσαρμοσμένο δείκτη Rand, αλλά με βάση τα άλλα δύο μέτρα, δε κρίνεται επιτυχημένη.

Συνοπτικά, παρατηρήσαμε πως και στις τέσσερις περιπτώσεις ομαδοποίησης που πραγματοποιήσαμε, τα δεδομένα μας διαχωριζόντουσαν καλύτερα σε δύο ξεχωριστές ομάδες. Αυτό το γεγονός κρίνεται και αναμενόμενο, δεδομένου ότι οι δύο βασικές κλάσεις των παρατηρήσεων μας ήταν οι ομάδες που προκρίθηκαν στις επιμέρους φάσεις της διοργάνωσης, και αυτές που δεν προκρίθηκαν. Συνεπώς, αυτό θα μπορούσε να αποτελεί το κοινό χαρακτηριστικό ανάμεσα στα στοιχεία/παρατηρήσεις των εκάστοτε δύο συστάδων.

6.5 Κατηγοριοποίηση / Ταξινόμηση (Classification)

Η κατηγοριοποίηση/ταξινόμηση είναι μια θεμελιώδης εργασία στη μηχανική μάθηση, όπου ο στόχος είναι να αντιστοιχηθεί μια παρατήρηση από το σύνολο δεδομένων σε μια κατηγορία, με βάση τα χαρακτηριστικά της. Στόχος είναι η εκμάθηση ενός μοντέλου που μπορεί να προβλέψει με ακρίβεια την κατηγορία των νέων παρατηρήσεων. Με την ακριβή κατηγοριοποίηση των δεδομένων, η ταξινόμηση μας επιτρέπει να λαμβάνουμε καλύτερες αποφάσεις, να εντοπίζουμε μοτίβα και να αποκτούμε βαθύτερες γνώσεις για τον κόσμο γύρω μας. Έχει εφαρμοστεί σε ένα ευρύ φάσμα τομέων, όπως π.χ. η βιολογία, τα χρηματοοικονομικά και οι κοινωνικές επιστήμες. Στη βιολογία, η ταξινόμηση χρησιμοποιείται για την κατηγοριοποίηση διαφορετικών ειδών με βάση τη γενετική τους σύσταση, τα φυσικά χαρακτηριστικά και τη συμπεριφορά τους. Στα χρηματοοικονομικά, χρησιμοποιείται για την ανίχνευση «δόλιων» συναλλαγών με τον εντοπισμό μοτίβων στα δεδομένα που υποδηλώνουν δόλια συμπεριφορά. Στις κοινωνικές επιστήμες, χρησιμοποιείται για την ομαδοποίηση των ανθρώπων με βάση τα δημογραφικά χαρακτηριστικά, τα ενδιαφέροντα και τη συμπεριφορά τους.

Η επιτυχία των αλγορίθμων κατηγοριοποίησης εξαρτάται σε μεγάλο βαθμό από την ποιότητα των δεδομένων που χρησιμοποιούνται για την εκπαίδευση. Ένα καλά επισημασμένο (well-labeled) σύνολο δεδομένων είναι ζωτικής σημασίας για την ακριβή εκμάθηση των υποκείμενων προτύπων και τη γενίκευση του αλγορίθμου σε νέα δεδομένα. Η επιλογή χαρακτηριστικών (όπως είδαμε στην ενότητα 6.3) είναι μια άλλη σημαντική πτυχή της ταξινόμησης, όπου επιλέγονται τα πιο σχετικά χαρακτηριστικά για τη μείωση του θορύβου και τη βελτίωση της ακρίβειας του μοντέλου.

Μία από τις προκλήσεις της ταξινόμησης είναι η αντιμετώπιση ανισοβαρών συνόλων δεδομένων, όπου μία κλάση (ή κατηγορία) έχει σημαντικά λιγότερες παρατηρήσεις από τις άλλες. Αυτό μπορεί να οδηγήσει σε μεροληψία στο μοντέλο, όπου αυτό τείνει να προβλέπει συχνότερα την πλειοψηφούσα κλάση, με αποτέλεσμα κακές επιδόσεις για τη μειοψηφούσα κλάση. Για να αντιμετωπιστεί αυτό, έχουν αναπτυχθεί τεχνικές όπως η επαναδειγματοληψία, η μάθηση με ευαισθησία στο κόστος και οι μέθοδοι συνόλου, με σκοπό την εξισορρόπηση της κατανομής των κλάσεων και τη βελτίωση της απόδοσης στην κλάση της μειονότητας. (Hastie et al., 2009 ; Bishop, 2006)

Υπάρχουν διάφοροι αλγόριθμοι ταξινόμησης, ο καθένας με τα δυνατά και τα αδύνατα σημεία του. Η επιλογή του αλγορίθμου εξαρτάται και πάλι από το εκάστοτε πρόβλημα που αντιμετωπίζουμε, καθώς και τη φύση των δεδομένων μας.

Ακολουθούν συνοπτικά, ορισμένα παραδείγματα δημοφιλών αλγορίθμων ταξινόμησης, κάποιους από τους οποίους αναφέρουν και οι Hastie et al. (2009).

- Λογιστική παλινδρόμηση : Ένας απλός και ευρέως χρησιμοποιούμενος αλγόριθμος που μοντελοποιεί την πιθανότητα ένα σημείο δεδομένων να ανήκει σε μια συγκεκριμένη κλάση με βάση έναν γραμμικό συνδυασμό των

χαρακτηριστικών του. Την είδαμε αναλυτικά στο προηγούμενο κεφάλαιό μας, χωρίς να τη χρησιμοποιήσουμε για εργασία κατηγοριοποίησης.

- Γραμμική Διαχωριστική Ανάλυση (Linear Discriminant Analysis - LDA) : Ο αλγόριθμος αναλύει τα χαρακτηριστικά ενός συνόλου δεδομένων και προβάλλει τα δείγματα σε έναν χώρο χαμηλότερης διάστασης, ο οποίος μπορεί να χρησιμοποιηθεί για να κατηγοριοποιήσει νέα δείγματα. Στόχος είναι να βρεθεί ένας γραμμικός συνδυασμός των χαρακτηριστικών που μπορεί να χωρίσει τα δείγματα σε διαφορετικές κατηγορίες.

- Δέντρα αποφάσεων (Decision Trees) : Ένας αλγόριθμος βασισμένος σε δέντρα, που χωρίζει τα δεδομένα αναδρομικά με βάση τις τιμές των χαρακτηριστικών τους για να σχηματίσει μια δομή που μοιάζει με δέντρο.

- Τυχαίο δάσος (Random Forest) : Ένας ευρέως χρησιμοποιούμενος αλγόριθμος συνόλου που συνδυάζει πολλαπλά δέντρα απόφασης για να βελτιώσει την απόδοση της ταξινόμησης και να ελαττώσει το over-fitting. Θα χρησιμοποιήσουμε αυτόν τον αλγόριθμο προσεχώς στην ανάλυση μας.

- Μηχανές διανυσμάτων υποστήριξης (Support Vector Machines - SVM) : Ένας ισχυρός αλγόριθμος που διαχωρίζει τα δεδομένα σε κλάσεις, χρησιμοποιώντας ένα υπερ-επίπεδο σε έναν χώρο χαρακτηριστικών υψηλής διάστασης. Θα τον χρησιμοποιήσουμε επίσης στην ανάλυση μας.

- Naive Bayes : Ένας πιθανοτικός αλγόριθμος που μοντελοποιεί την πιθανότητα ένα σημείο των δεδομένων να ανήκει σε μια συγκεκριμένη κλάση, με βάση το θεώρημα του Bayes, υποθέτοντας ανεξαρτησία μεταξύ των χαρακτηριστικών του.

- K-Nearest Neighbors (KNN) : Ένας απλός αλγόριθμος που ταξινομεί ένα σημείο δεδομένων με βάση τις ετικέτες κλάσης (labels) των K - πλησιέστερων γειτόνων του στο χώρο των χαρακτηριστικών.

- Gradient Boosting : Ένας δημοφιλής αλγόριθμος συνόλου που συνδυάζει πολλαπλούς αδύναμους ταξινόμητές με σκοπό να σχηματίσει έναν ισχυρό, χρησιμοποιώντας την «κάθοδο κλίσης» (gradient descent) .

- Νευρωνικά δίκτυα (Neural Networks) : Ένας ευέλικτος αλγόριθμος που χρησιμοποιεί ένα σύνολο διασυνδεδεμένων κόμβων για να μάθει πολύπλοκες μη γραμμικές σχέσεις μεταξύ των χαρακτηριστικών εισόδου και των ετικετών της κλάσης εξόδου.

Πριν συνεχίσουμε με τις τεχνικές κατηγοριοποίησής μας, είναι σημαντικό να αναφέρουμε δύο πράγματα.

- Έχουμε ήδη πραγματοποιήσει επιλογή χαρακτηριστικών για τις δύο φάσεις που εξετάζουμε (ενότητα 6.3) και στη συνέχεια έχουμε κανονικοποιήσει και τα δεδομένα μας. Με άλλα λόγια, θα χρησιμοποιήσουμε τις ίδιες μεταβλητές τις οποίες χρησιμοποιήσαμε και στο clustering, για τις δύο φάσεις.

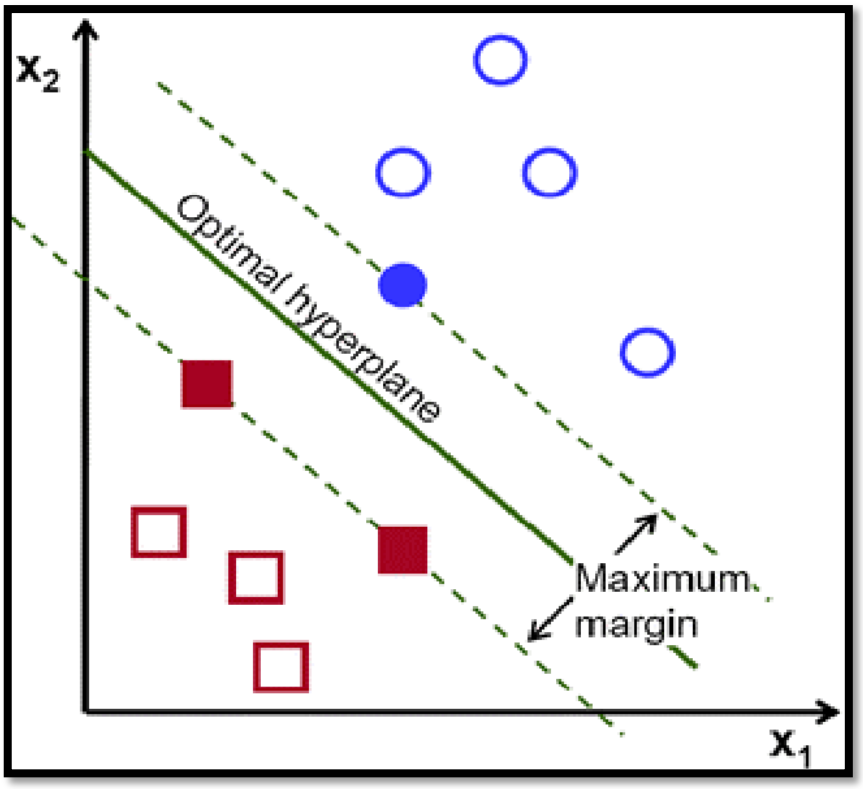
- Στη συνέχεια, θα πραγματοποιήσουμε διαχωρισμό των δεδομένων μας. Στη μηχανική μάθηση, ο διαχωρισμός των δεδομένων σε σύνολα εκπαίδευσης (training sets) και δοκιμών (testing sets) είναι απαραίτητος για την αξιολόγηση της απόδοσης του μοντέλου. Η ιδέα είναι πως ένα μέρος των δεδομένων χρησιμοποιείται για να εκπαιδευτεί το μοντέλο και στη συνέχεια, να δοκιμαστεί σε ένα άλλο μέρος των δεδομένων, και συγκεκριμένα, σε αθέατα δεδομένα. Εάν εκπαιδευτεί το μοντέλο σε ολόκληρο το σύνολο δεδομένων μόνο, μπορεί να προσαρμοστεί υπερβολικά στα δεδομένα εκπαίδευσης και να έχει κακές επιδόσεις στα νέα δεδομένα. Εκτός από το διαχωρισμό των δεδομένων σε σύνολα εκπαίδευσης και δοκιμής, είναι επίσης κοινή πρακτική να χωρίζεται περαιτέρω το σύνολο εκπαίδευσής σε ένα μικρότερο σύνολο εκπαίδευσης και ένα σύνολο επικύρωσης (validation set). Αυτό επιτρέπει την αξιολόγηση της απόδοσης του μοντέλου κατά τη διάρκεια της εκπαίδευσης.

Εμείς χωρίσαμε τα δεδομένα μας σε αναλογία 70% - 30% ανάμεσα σε δεδομένα εκπαίδευσης και δεδομένα δοκιμής. Συχνά χρησιμοποιείται και 80% - 20% αναλογία.

6.5.1 Εφαρμογή αλγορίθμου SVM για τη φάση των playoffs

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVM) είναι ένας τύπος αλγορίθμου μάθησης με επίβλεψη, που μπορεί να χρησιμοποιηθεί τόσο για εργασίες ταξινόμησης, όσο και παλινδρόμησης. Η βασική ιδέα πίσω από τα SVMs είναι η εύρεση ενός υπερ-επιπέδου (hyperplane) που διαχωρίζει στοιχεία δύο κλάσεων σε έναν χώρο υψηλών διαστάσεων. Τα SVMs είναι ιδιαίτερα χρήσιμα, όταν τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα, καθώς χρησιμοποιούν ένα «τέχνασμα» πυρήνα για να μετατρέψουν τα δεδομένα σε έναν χώρο υψηλότερων διαστάσεων, όπου γίνονται τελικά διαχωρίσιμα.

Ο στόχος ενός SVM είναι να βρει το υπερ-επίπεδο που μεγιστοποιεί το περιθώριο μεταξύ των δύο κλάσεων. Περιθώριο ορίζεται ως η απόσταση μεταξύ του υπερ-επιπέδου και των πλησιέστερων σημείων από κάθε κλάση. Με τη μεγιστοποίηση του περιθωρίου, τα SVMs είναι σε θέση να βρουν ένα όριο απόφασης που είναι πιο ανθεκτικό στο θόρυβο και τις ακραίες τιμές. Τα σημεία των δεδομένων που βρίσκονται πλησιέστερα στο υπερ-επίπεδο ονομάζονται διανύσματα υποστήριξης και ορίζουν το υπερ-επίπεδο. (Hastie et al., 2009)



Σχήμα 6.23 : Γραφική απεικόνιση της επιλογής του μέγιστου περιθωρίου μεταξύ των σημείων των κατηγοριών.

(Πηγή : https://www.researchgate.net/figure/Optimal-Hyperplane-and-Margin-of-SVM_fig3_338698374)

Ο αλγόριθμος λειτουργεί μετασχηματίζοντας αρχικά τα δεδομένα σε έναν χώρο χαρακτηριστικών υψηλότερων διαστάσεων, όπου είναι πιο πιθανό να είναι γραμμικά διαχωρίσιμα. Αυτός ο μετασχηματισμός γίνεται με τη χρήση μιας συνάρτησης πυρήνα, η οποία απεικονίζει τα δεδομένα στον χώρο υψηλότερων διαστάσεων χωρίς στην πραγματικότητα να υπολογίζει τις συντεταγμένες των δεδομένων στον εν λόγω χώρο. Η επιλογή της συνάρτησης πυρήνα είναι κρίσιμη, καθώς καθορίζει το σχήμα του ορίου απόφασης. Ακολουθούν μερικές από τις πιο συνηθισμένες.

Πολυωνυμική συνάρτηση πυρήνα (Polynomial Kernel Function)

Η πολυωνυμική συνάρτηση πυρήνα χρησιμοποιείται όταν τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα στον αρχικό χώρο εισόδου. Αυτή η συνάρτηση πυρήνα απεικονίζει τα διανύσματα εισόδου σε έναν χώρο χαρακτηριστικών υψηλότερης διάστασης, χρησιμοποιώντας μια πολυωνυμική συνάρτηση. Ο βαθμός της πολυωνυμικής συνάρτησης συμβολίζεται με d και είναι μια υπερπαράμετρος που πρέπει να ρυθμιστεί. Οι παράμετροι α και c ελέγχουν την επιρροή των όρων υψηλότερης τάξης στο πολυώνυμο. Έχει τύπο

$$k(x, y) = (\alpha x^T y + c)^d$$

Συνάρτηση ακτινωτής βάσης (Radial Basis Function - RBF)

Η συνάρτηση πυρήνα RBF είναι επίσης γνωστή ως συνάρτηση πυρήνα Gauss. Αυτή η συνάρτηση πυρήνα απεικονίζει τα διανύσματα εισόδου σε έναν άπειρης διάστασης χώρο χαρακτηριστικών, χρησιμοποιώντας μια συνάρτηση Gauss (μπορεί και Laplace). Η παράμετρος γ ελέγχει το πλάτος της «Γκαουσιανής» συνάρτησης και πρέπει να ρυθμιστεί. Η συνάρτηση πυρήνα RBF είναι κατάλληλη όταν το όριο απόφασης είναι εξαιρετικά μη γραμμικό. Αυτός είναι ο πιο συχνά χρησιμοποιούμενος πυρήνας, και αυτός που θα χρησιμοποιήσουμε και εμείς στην εφαρμογή του αλγορίθμου. Έχει τύπο

$$k(x, y) = exp(\gamma \, \|x - y\|^2)$$

Σιγμοειδής συνάρτηση πυρήνα (Sigmoid Kernel Function)

Η σιγμοειδής συνάρτηση πυρήνα χρησιμοποιείται επίσης όταν τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα στον αρχικό χώρο εισόδου. Αυτή η συνάρτηση πυρήνα απεικονίζει τα διανύσματα εισόδου σε έναν χώρο χαρακτηριστικών υψηλότερων διαστάσεων χρησιμοποιώντας μια σιγμοειδή συνάρτηση. Βρίσκει περισσότερη εφαρμογή στα νευρωνικά δίκτυα, ως συνάρτηση ενεργοποίησης. Οι παράμετροι α και c ελέγχουν το σχήμα της σιγμοειδούς συνάρτησης και πρέπει να ρυθμιστούν. Έχει τύπο

$$k(x, y) = tanh(\alpha x^T y + c)$$

Γραμμική συνάρτηση πυρήνα (Linear Kernel Function)

Η γραμμική συνάρτηση πυρήνα είναι η απλούστερη και πιο συχνά χρησιμοποιούμενη συνάρτηση πυρήνα στο SVM. Υπολογίζει το τετραγωνικό γινόμενο μεταξύ δύο διανυσμάτων εισόδου στον αρχικό χώρο εισόδου. Αυτή η συνάρτηση πυρήνα είναι κατάλληλη όταν τα δεδομένα είναι γραμμικά διαχωρίσιμα, δηλαδή όταν οι κλάσεις μπορούν να διαχωριστούν από μια ευθεία γραμμή ή ένα υπερ-επίπεδο. Έχει τύπο

$$k(x, y) = x^T y$$

(Schölkopf & Smola, 2002)

Αφού μετασχηματιστούν τα δεδομένα, το SVM βρίσκει το υπερ-επίπεδο που μεγιστοποιεί το περιθώριο μεταξύ των κλάσεων. Αυτό γίνεται με την επίλυση ενός προβλήματος βελτιστοποίησης με περιορισμούς, όπου ο στόχος είναι η μεγιστοποίηση του περιθωρίου με τον περιορισμό ότι τα σημεία των δεδομένων ταξινομούνται σωστά.

Ο αλγόριθμος έχει πολλά πλεονεκτήματα για τα προβλήματα ταξινόμησης. Μπορεί να χειριστεί τόσο γραμμικά, όσο και μη γραμμικά διαχωρίσιμα δεδομένα, χάρη στη χρήση των συναρτήσεων πυρήνα. Επιπροσθέτως, είναι λιγότερο επιρρεπής σε υπερ-προσαρμογή από άλλους αλγορίθμους ταξινόμησης, καθώς επιδιώκει να

μεγιστοποιήσει το περιθώριο μεταξύ των κλάσεων, αντί να προσαρμόζει στενά τα δεδομένα. Ωστόσο, έχει και ορισμένους περιορισμούς, όπως η ευαισθησία του στην επιλογή της συνάρτησης πυρήνα και στις υπερ-παραμέτρους του αλγορίθμου, σαν τη παράμετρο κανονικοποίησης και τις παραμέτρους του πυρήνα. Επίσης, μπορεί να είναι υπολογιστικά δαπανηρός, ιδίως όταν πρόκειται για μεγάλα σύνολα δεδομένων ή πολύπλοκες συναρτήσεις πυρήνα. Τέλος, μπορεί να είναι δύσκολο να ερμηνευτεί, καθώς το υπερ-επίπεδο στον χώρο χαρακτηριστικών υψηλής διάστασης μπορεί να μην αντιστοιχεί άμεσα σε ένα απλό όριο απόφασης στον αρχικό χώρο χαρακτηριστικών. Παρ' όλα αυτά, είναι ιδιαίτερα χρήσιμος όταν οι υποκείμενες σχέσεις μεταξύ των χαρακτηριστικών και της μεταβλητής-στόχου είναι πολύπλοκες ή δεν είναι καλά καθορισμένες.

Η R προσφέρει διάφορα πακέτα για τα SVMS, με ορισμένα από τα πιο δημοφιλή να είναι τα 'e1071' , ' kernlab' , 'caret' και 'libsvm' . Κάθε πακέτο προσφέρει μοναδικά χαρακτηριστικά και δυνατότητες για τις εκάστοτε εργασίες που μας ενδιαφέρουν. Εμείς χρησιμοποιήσαμε το πρώτο πακέτο, το οποίο προσφέρει υποστήριξη τόσο για εργασίες ταξινόμησης, όσο και παλινδρόμησης, ενώ περιλαμβάνει και τις παραπάνω συναρτήσεις πυρήνα (γραμμική, πολυωνυμική, RBF, σιγμοειδής). Το πακέτο περιλαμβάνει επίσης πρόσθετα βοηθητικά προγράμματα για την προεπεξεργασία δεδομένων και την οπτικοποίηση αποτελεσμάτων, την ρύθμιση των παραμέτρων και διάφορες μετρικές αξιολόγησης μοντέλων. Χρησιμοποιήσαμε επίσης το πακέτο 'caret' για την οπτικοποίηση των αποτελεσμάτων μας.

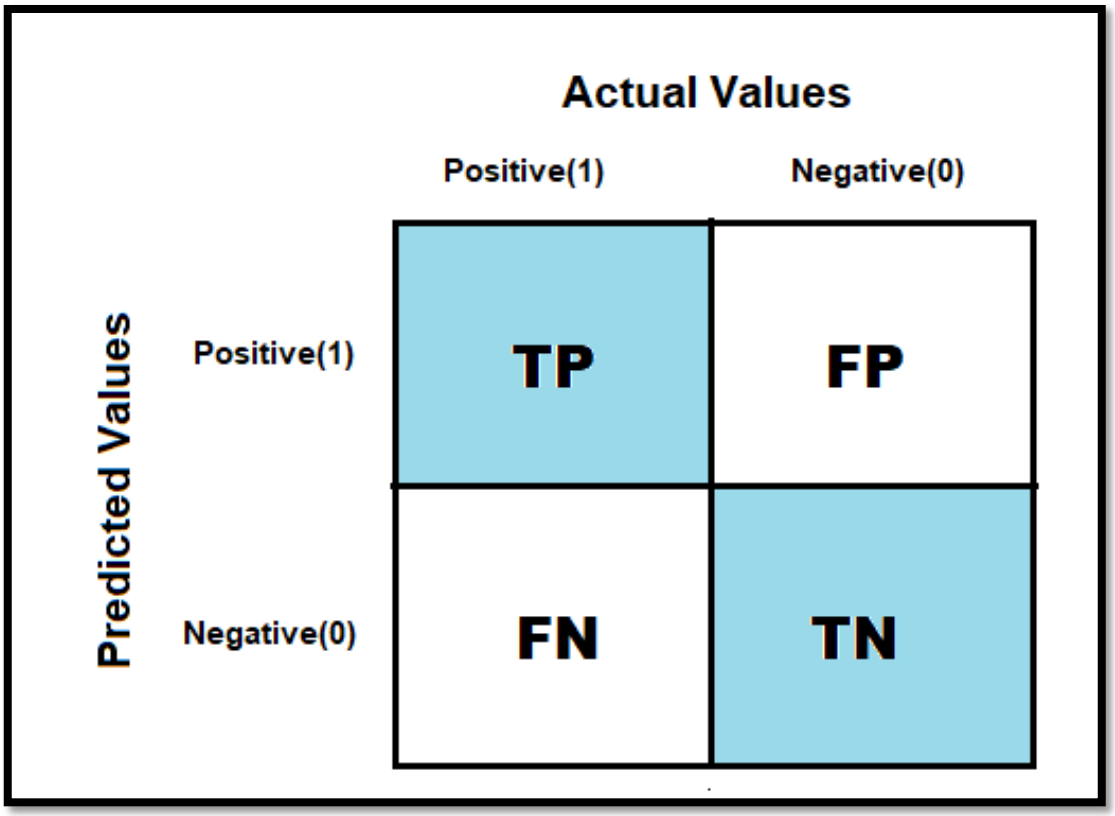
Μέτρα Αξιολόγησης της κατηγοριοποίησης

Μια άλλη σημαντική πτυχή της ταξινόμησης γενικότερα, είναι η αξιολόγηση της απόδοσης του μοντέλου. Οι συνήθεις μετρικές περιλαμβάνουν την ορθότητα (accuracy), την ακρίβεια (precision), την ανάκληση (recall), το F1 score, την καμπύλη ROC (την οποία χρησιμοποιήσαμε στο προηγούμενο κεφάλαιο, για να αξιολογήσουμε τα προσαρμοσμένα μοντέλα λογιστικής παλινδρόμησης), καθώς και το πίνακα σύγχυσης (confusion matrix). Αυτές οι μετρικές παρέχουν πληροφορίες για τα δυνατά και αδύνατα σημεία του μοντέλου και βοηθούν στον εντοπισμό περιοχών για βελτίωση.

Ο πίνακας σύγχυσης είναι ένας πίνακας που χρησιμοποιείται συνήθως για την αξιολόγηση της απόδοσης ενός μοντέλου ταξινόμησης. Βοηθά στην οπτικοποίηση των ποσοστών των αληθώς θετικών, αληθώς αρνητικών, ψευδώς θετικών και ψευδώς αρνητικών τιμών ενός μοντέλου. Ο όρος θετικές αναφέρεται στις τιμές της κατηγορικής μεταβλητής που ισούνται με 1, και ο όρος αρνητικές όταν έχουν τιμή ίση με 0.

- **Αληθώς θετικές (True Positives - TP) :** Αναφέρεται στον αριθμό των περιπτώσεων που προβλέφθηκαν σωστά ως θετικές από το μοντέλο.
- **Αληθώς αρνητικές (True Negatives - TN) :** Αναφέρεται στον αριθμό των περιπτώσεων που προβλέφθηκαν σωστά ως αρνητικές από το μοντέλο.

- **Ψευδώς θετικές (False Positives - FP) :** Αναφέρεται στον αριθμό των περιπτώσεων που προβλέφθηκαν ως θετικές από το μοντέλο, αλλά στην πραγματικότητα ήταν αρνητικές.
- **Ψευδώς αρνητικές (False Negatives - FN) :** Αναφέρεται στον αριθμό των περιπτώσεων που προβλέφθηκαν ως αρνητικές από το μοντέλο, αλλά ήταν στην πραγματικότητα θετικές.



Σχήμα 6.24 : Παράδειγμα ενός πίνακα συσχέτισης.

(Πηγή : https://www.linkedin.com/pulse/understanding-confusion-matrix-kartik-lokare)

Τα διαγώνια στοιχεία του πίνακα αντιπροσωπεύουν τις σωστές προβλέψεις που έγιναν από το μοντέλο, ενώ τα στοιχεία εκτός διαγωνίου αντιπροσωπεύουν τις λανθασμένες προβλέψεις. Με βάση τις τιμές στον πίνακα σύγχυσης, μπορούν να υπολογιστούν τα μέτρα που αναφέραμε παραπάνω.

Ορθότητα (Accuracy)

Πρόκειται για μια από τις πιο συχνά χρησιμοποιούμενες μετρικές για εργασίες ταξινόμησης. Μετρά το ποσοστό των σωστών προβλέψεων που πραγματοποιεί το μοντέλο επί του συνολικού αριθμού των προβλέψεων. Υπολογίζεται ως

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Ακρίβεια (Precision)

Είναι ο λόγος των αληθώς θετικών αποτελεσμάτων προς τον συνολικό αριθμό των προβλεπόμενων θετικών αποτελεσμάτων. Μετρά την ικανότητα του μοντέλου να εντοπίζει σωστά θετικές περιπτώσεις. Υπολογίζεται ως

$$\frac{TP}{TP + FP}$$

Ανάκληση (Recall)

Η ανάκληση είναι ο λόγος των αληθώς θετικών προς τον συνολικό αριθμό των πραγματικών θετικών. Μετρά την ικανότητα του μοντέλου να αναγνωρίζει σωστά όλες τις θετικές περιπτώσεις. Υπολογίζεται ως

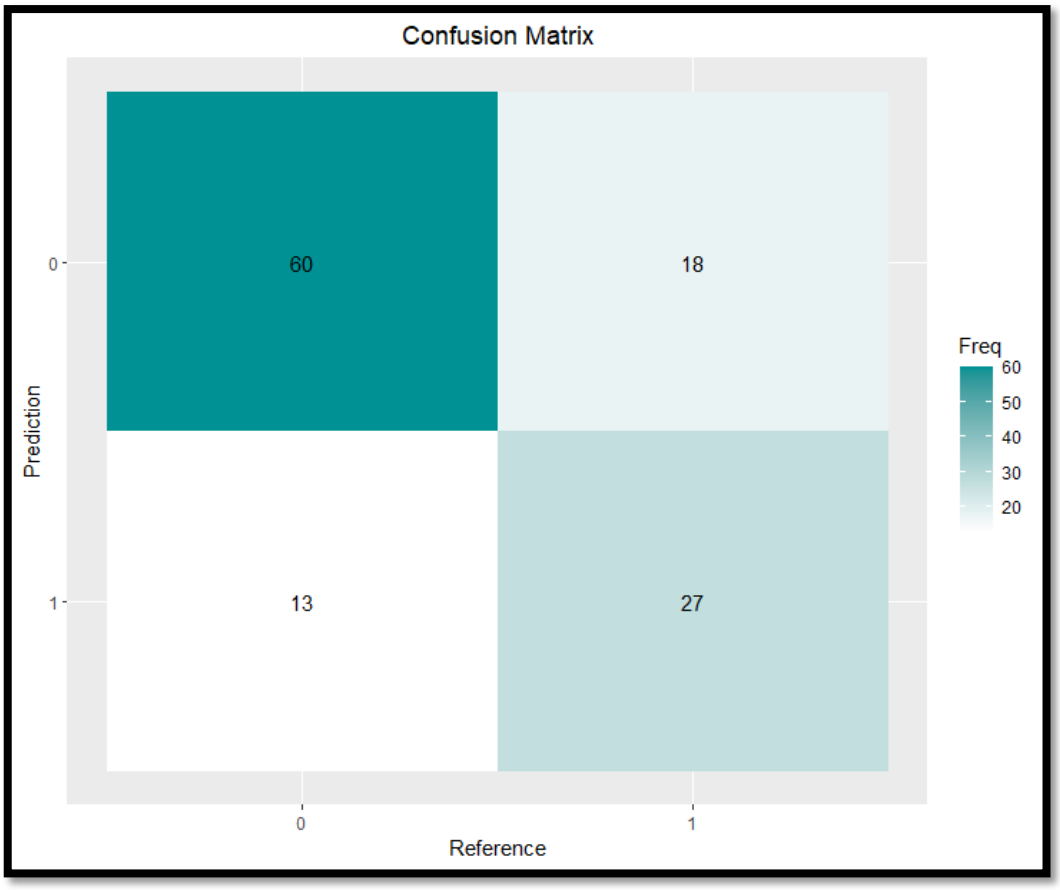
$$\frac{TP}{TP + FN}$$

F1 Score

Το F1 score είναι ο αρμονικός μέσος της ακρίβειας και της ανάκλησης. Αποτελεί μια ισορροπία μεταξύ ακρίβειας και ανάκλησης και δίνει ένα συνολικό μέτρο της απόδοσης του μοντέλου. Υπολογίζεται ως

$$2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Εφαρμόσαμε συνεπώς τον αλγόριθμο των SVM, έχοντας ως συνάρτηση πυρήνα την συνάρτηση ακτινωτής βάσης (RBF) με τη τιμή της παραμέτρου γ να ισούται με 1. Παρακάτω δίνεται ο πίνακας σύγχυσης της ταξινόμησης που προέκυψε.



Σχήμα 6.25 : Πίνακας σύγχυσης για τα playoffs.

Παρατηρούμε ότι ταξινομήθηκαν σωστά 60 παρατηρήσεις στις ομάδες που δεν προκρίθηκαν στη φάση των playoffs (διαχρονικά), και 27 στις ομάδες που προκρίθηκαν. Παράλληλα, το μοντέλο μας ταξινόμησε λανθασμένα 13 ομάδες που δεν προκρίθηκαν, σε ομάδες που πήραν την πρόκριση, και 18 ομάδες που προκρίθηκαν, στην κατηγορία των αποκλεισμένων ομάδων. Συνολικά, ταξινόμησε ορθά 87 παρατηρήσεις και 31 λανθασμένα, με βάση το σύνολο δοκιμών.

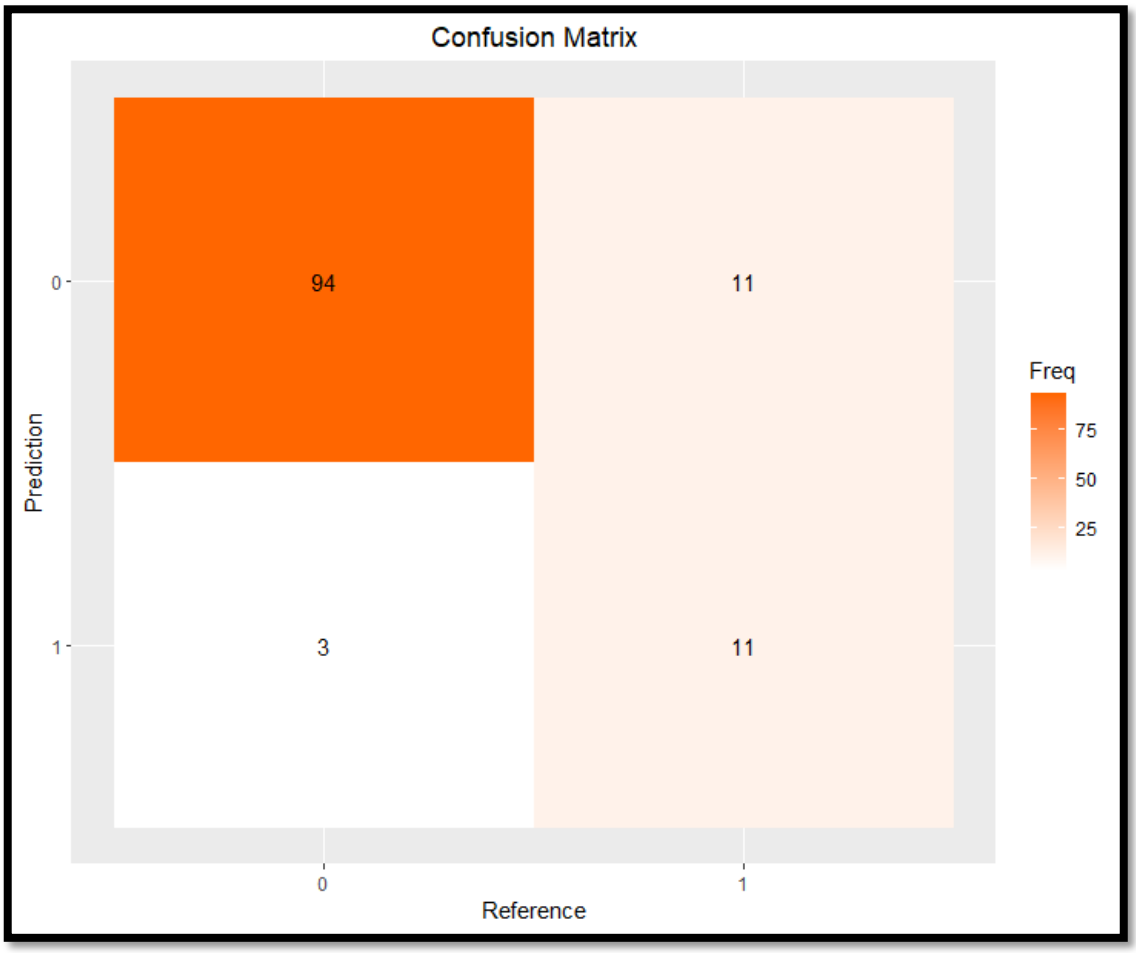
Ακολουθεί τώρα ο πίνακας με τις τιμές των μέτρων αξιολόγησης για την κατηγοριοποίηση, για τα Playoffs. Γενικά, φαίνεται να έχει πραγματοποιηθεί μια μέτρια προς καλή ταξινόμηση από το μοντέλο μας.

Μέτρα Αξιολόγησης	Τιμές
Accuracy	**0,737**
Precision	**0,675**
Recall	**0,600**
F1 score	**0,635**

Πίνακας 6.6 : Μέτρα αξιολόγησης για την κατηγοριοποίηση των playoffs.

6.5.2 Εφαρμογή αλγορίθμου SVM για τη φάση του Final Four

Σε αυτό το σημείο, ακολουθήσαμε ακριβώς την ίδια διαδικασία με την προηγούμενη ενότητα, απλώς τώρα για τη φάση του Final Four. Εφαρμόσαμε τον αλγόριθμο των SVM με συνάρτηση πυρήνα την RBF, τιμή για τη παράμετρο γ ίση με 1 και πήραμε το παρακάτω πίνακα σύγχυσης.



Σχήμα 6.26 : Πίνακας σύγχυσης για το Final Four.

Παρατηρούμε ότι ταξινομήθηκαν σωστά 94 παρατηρήσεις στις ομάδες που δεν προκρίθηκαν στη φάση του Final Four (διαχρονικά), και 11 στις ομάδες που προκρίθηκαν. Παράλληλα, το μοντέλο μας ταξινόμησε λανθασμένα 3 ομάδες που δεν προκρίθηκαν, σε ομάδες που πήραν την πρόκριση, και 11 ομάδες που προκρίθηκαν, στην κατηγορία των αποκλεισμένων ομάδων. Οι 3 αυτές ομάδες ήταν η Benetton (2005-2006), η οποία δεν πέρασε καν στα προημιτελικά/playoffs εκείνης της σεζόν, η Barcelona (2006-007), η οποία αποκλείστηκε στα προημιτελικά εκείνης της σεζόν από τη Unicaja, χάνοντας με 2-1 στη σειρά των αγώνων, και ο Olympiacos (2013-2014), ο οποίος αποκλείστηκε στα playoffs, χάνοντας από τη Real Madrid με 3-2 στη σειρά των

αγώνων. Συνολικά, το μοντέλο ταξινόμησε ορθά 105 παρατηρήσεις και μόλις 14 λανθασμένα, με βάση το σύνολο δοκιμών.

Ακολουθεί ο πίνακας με τις τιμές των μέτρων αξιολόγησης για την κατηγοριοποίηση, για το Final Four. Γενικά, φαίνεται να έχει πραγματοποιηθεί μια καλύτερη ταξινόμηση από το μοντέλο μας, σε αυτή τη φάση της διοργάνωσης.

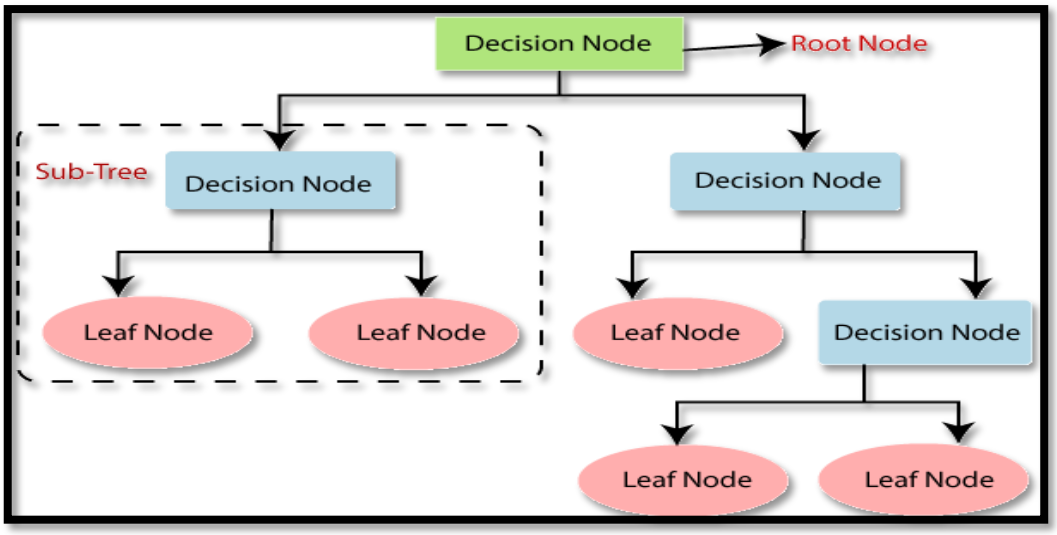
Μέτρα Αξιολόγησης	Τιμές
Accuracy	**0,882**
Precision	**0,786**
Recall	**0,500**
F1 score	**0,611**

Πίνακας 6.7 : Μέτρα αξιολόγησης για την κατηγοριοποίηση του Final Four.

6.5.3 Εφαρμογή αλγορίθμου Random Forest για τη φάση των playoffs

Προκειμένου να αναλύσουμε τον αλγόριθμο του τυχαίου δάσους (Random Forest), πρέπει πρώτα να κάνουμε μια σύντομη αναφορά στα δέντρα αποφάσεων (Decision Trees).

Τα δέντρα απόφασης είναι ένας δημοφιλής αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται ευρέως σε διάφορους τομείς, λόγω της ερμηνευσιμότητας και της απλότητάς του. Λειτουργεί με την κατασκευή ενός δενδροειδούς μοντέλου αποφάσεων και της μελέτης των πιθανών συνεπειών τους, όπου κάθε εσωτερικός κόμβος (decision node) αναπαριστά μια απόφαση βάσει ενός χαρακτηριστικού, κάθε κλάδος αναπαριστά ένα αποτέλεσμα της απόφασης και κάθε κόμβος φύλλου (leaf node) αναπαριστά μια ετικέτα ταξινόμησης. Τα δέντρα αποφάσεων μπορούν να χειριστούν τόσο κατηγορικά, όσο και αριθμητικά δεδομένα και μπορούν να χρησιμοποιηθούν τόσο για εργασίες παλινδρόμησης, όσο και για εργασίες ταξινόμησης. Ωστόσο, μπορεί να υποφέρουν από προβλήματα υπερ-προσαρμογής, όταν το εκάστοτε δέντρο γίνεται πολύ πολύπλοκο και ευαίσθητο στο θόρυβο των δεδομένων.

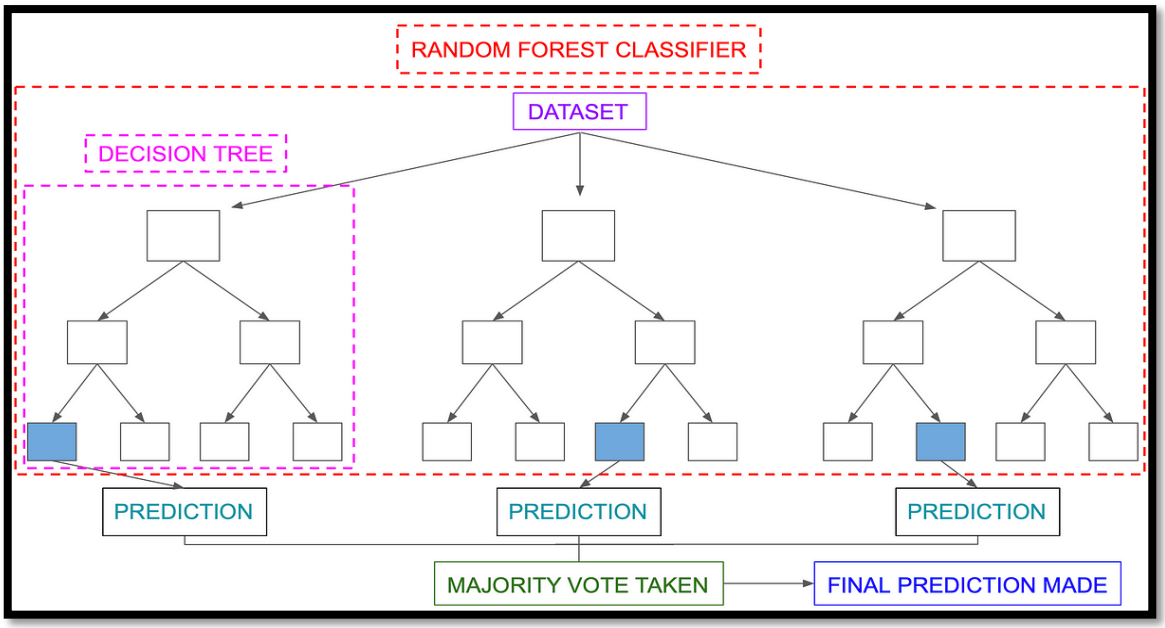


Σχήμα 6.27 : Παράδειγμα ενός δέντρου αποφάσεων.

(Πηγή : https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm)

Ο αλγόριθμος του τυχαίου δάσους (Random Forest) έχει κερδίσει σημαντική προσοχή τα τελευταία χρόνια, λόγω της ικανότητάς του να χειρίζεται πολύπλοκα προβλήματα ταξινόμησης. Πρόκειται για μια μέθοδο μάθησης συνόλου που συνδυάζει πολλαπλά δέντρα απόφασης για τη βελτίωση της ακρίβειας και τη μείωση της υπερ-προσαρμογής του μοντέλου. Ο αλγόριθμος έχει χρησιμοποιηθεί ευρέως σε διάφορους τομείς, όπως η χρηματοδότηση, η υγειονομική περίθαλψη, το μάρκετινγκ κ.ά., όπου η υψηλή ακρίβεια και η αξιοπιστία είναι ζωτικής σημασίας. Ο κατηγοριοποιητής τυχαίου δάσους είναι γνωστός για την επεκτασιμότητα, την αποτελεσματικότητα και την ικανότητά του να χειρίζεται θορυβώδη δεδομένα, καθώς και ελλείπουσες τιμές.

Ο αλγόριθμος λειτουργεί με την κατασκευή ενός συνόλου δέντρων απόφασης, όπου κάθε δέντρο εκπαιδεύεται σε ένα τυχαία επιλεγμένο υποσύνολο δεδομένων και ένα τυχαία επιλεγμένο υποσύνολο χαρακτηριστικών. Κατά την ταξινόμηση, το τυχαίο δάσος συνδυάζει τις προβλέψεις όλων των δέντρων για να καταλήξει σε μια τελική πρόβλεψη, η οποία έχει αποδειχθεί ότι είναι πιο ακριβής και αξιόπιστη από τις προβλέψεις των μεμονωμένων δέντρων. Το τυχαίο δάσος όπως αναφέραμε παραπάνω, έχει αποδειχθεί ότι υπερτερεί έναντι πολλών άλλων αλγορίθμων ταξινόμησης σε διάφορες εφαρμογές, ιδίως όταν πρόκειται για δεδομένα υψηλής διάστασης και με θόρυβο, καθιστώντας έτσι μια δημοφιλής επιλογή για τους επιστήμονες δεδομένων και τους επαγγελματίες της μηχανικής μάθησης. (Breiman L., 2001)



Σχήμα 6.28 : Παράδειγμα ενός τυχαίου δάσους.

(Πηγή : https://medium.com/analytics-vidhya/machine-learning-decision-trees-and-random-forest-classifiers-81422887a544)

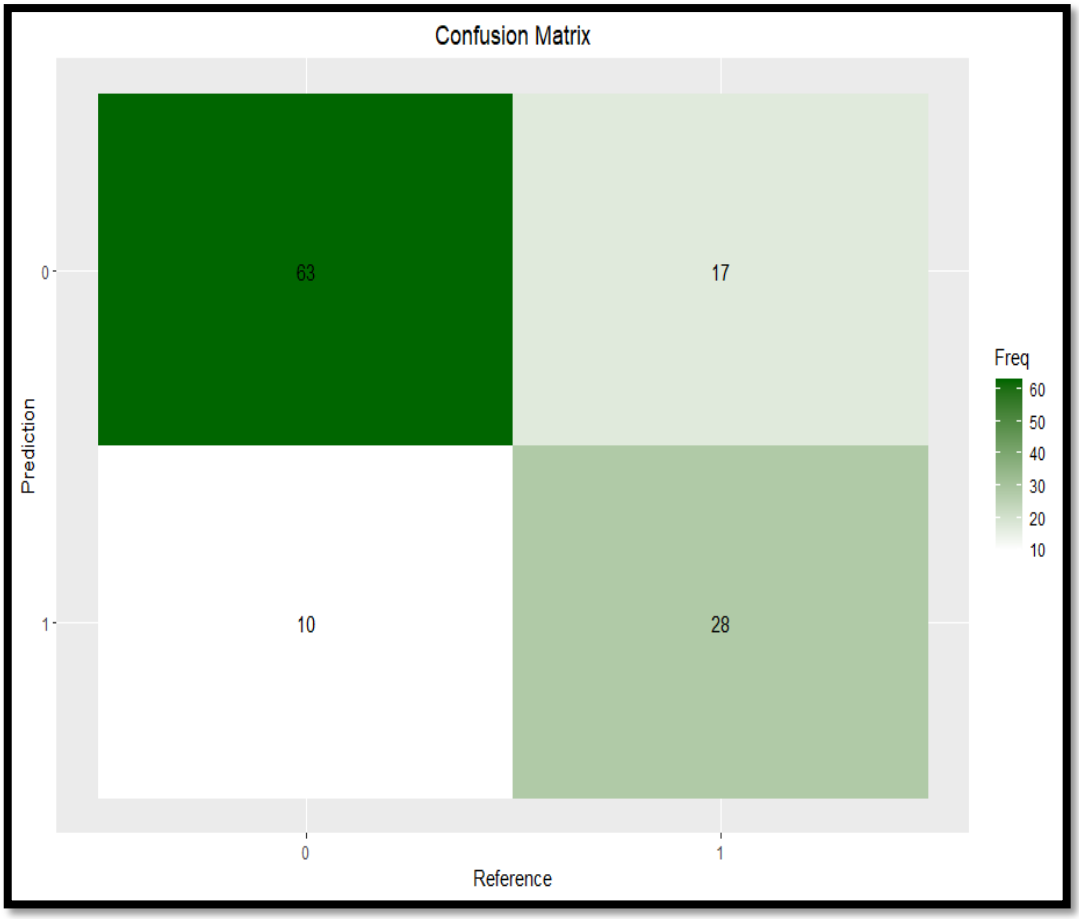
Για τη αλγόριθμο του τυχαίου δάσους, εκτελούνται τα ακόλουθα βήματα σύμφωνα με τον Καλλιακμάνη (2020) :

1. Δημιουργείται ένα bootstrapped σύνολο δεδομένων. Ένα τέτοιο σύνολο δεδομένων είναι ένα δείγμα δεδομένων που δημιουργείται με τυχαία επιλογή παρατηρήσεων από ένα αρχικό σύνολο δεδομένων με επανάθεση. Η διαδικασία επιλογής δεδομένων με επανάθεση σημαίνει ότι ορισμένες από τις αρχικές παρατηρήσεις, μπορεί να εμφανίζονται περισσότερες από μία φορές στο bootstrapped σύνολο, ενώ άλλες μπορεί να μην εμφανίζονται και καθόλου.
2. Δημιουργούνται δέντρα αποφάσεων.
3. Γυρίζουμε ξανά στο 1ο βήμα και επαναλαμβάνουμε την διαδικασία.
4. Προβλέπεται το αποτέλεσμα για μια καινούργια παρατήρηση, χρησιμοποιώντας το άθροισμα από όλα τα δέντρα μαζί. Η συγκεκριμένη διαδικασία είναι ευρέως γνωστή ως bagging.
5. Τέλος, γίνεται η αξιολόγηση του μοντέλου. Στην πραγματικότητα, το 36,8% (περίπου το 1/3) του αρχικού συνόλου δεν περιλαμβάνεται στο bootstrapped dataset. Αυτό το σύνολο λέγεται Out-of-Bag (OOB) και χρησιμοποιείται για να ελέγξει την ακρίβεια του μοντέλου, δηλαδή εάν είναι αποτελεσματικό ή όχι.

Στην R, όταν προσαρμόζουμε ένα τέτοιο μοντέλο, “by default” ο αριθμός των δέντρων απόφασης στο δάσος είναι 500 και ο αριθμός των πιθανών χαρακτηριστικών που χρησιμοποιούνται για κάθε διαχωρισμό του συνόλου είναι 3. Μέσα από δοκιμές, είδαμε πως με 6 πιθανά χαρακτηριστικά, είχαμε λίγο καλύτερες τιμές για τα μέτρα αξιολόγησης, χωρίς να είναι ουσιαστική η διαφορά. Δεδομένου ότι θέλουμε συνήθως

όσο το δυνατόν λιγότερες μεταβλητές, αποφασίσαμε να συνεχίσουμε με την ανάλυση μας, με την επιλογή των 3 χαρακτηριστικών.

Μετά την εφαρμογή του αλγορίθμου για τη φάση των playoffs, προέκυψε ο παρακάτω πίνακα σύγχυσης.



Σχήμα 6.29 : Πίνακας σύγχυσης για τα playoffs.

Παρατηρούμε ότι ταξινομήθηκαν σωστά 63 παρατηρήσεις στις ομάδες που δεν προκρίθηκαν στη φάση των playoffs (διαχρονικά), και 28 στις ομάδες που προκρίθηκαν. Παράλληλα, το μοντέλο μας ταξινόμησε λανθασμένα 10 ομάδες που δεν προκρίθηκαν, σε ομάδες που πήραν την πρόκριση, και 17 ομάδες που προκρίθηκαν, στην κατηγορία των αποκλεισμένων ομάδων. Συνολικά, ταξινόμησε ορθά 91 παρατηρήσεις και 27 λανθασμένα, με βάση το σύνολο δοκιμών.

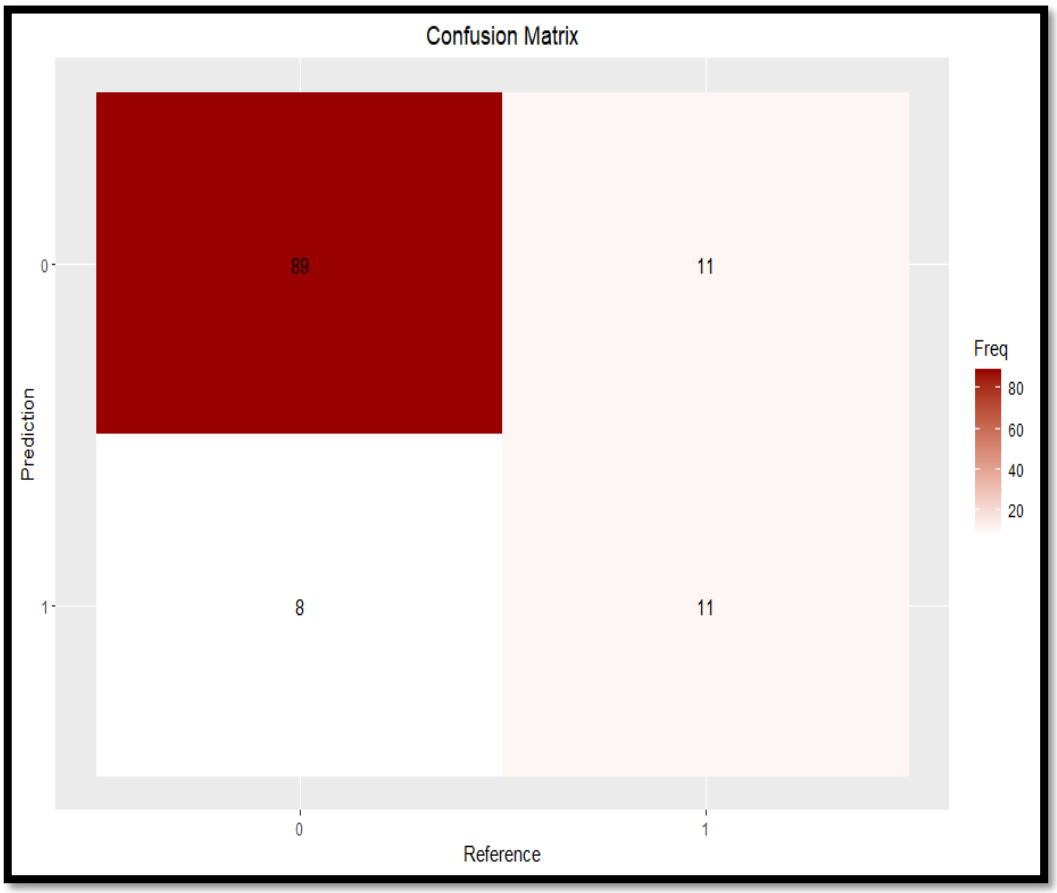
Ακολουθεί τώρα ο πίνακας με τις τιμές των μέτρων αξιολόγησης για την κατηγοριοποίηση, για τα Playoffs. Γενικά, φαίνεται να έχει πραγματοποιηθεί μια καλή ταξινόμηση από το μοντέλο μας.

Μέτρα Αξιολόγησης	Τιμές
Accuracy	**0,771**
Precision	**0,737**
Recall	**0,622**
F1 score	**0,675**

Πίνακας 6.8 : Μέτρα αξιολόγησης για την κατηγοριοποίηση των playoffs.

6.5.4 Εφαρμογή αλγορίθμου Random Forest για τη φάση του Final Four

Σε αυτό το σημείο της ανάλυσης, ακολουθήσαμε ακριβώς την ίδια διαδικασία με την προηγούμενη ενότητα, απλώς τώρα για τη φάση του Final Four. Εφαρμόσαμε τον αλγόριθμο τυχαίου δάσους, για 500 δέντρα και 3 πιθανά χαρακτηριστικά για διαχωρισμό, και πήραμε το παρακάτω πίνακα σύγχυσης.



Σχήμα 6.30 : Πίνακας σύγχυσης για το Final Four.

Παρατηρούμε ότι ταξινομήθηκαν σωστά 89 παρατηρήσεις στις ομάδες που δεν προκρίθηκαν στη φάση του Final Four (διαχρονικά), και 11 στις ομάδες που προκρίθηκαν. Παράλληλα, το μοντέλο μας ταξινόμησε λανθασμένα 8 ομάδες που δεν προκρίθηκαν, σε ομάδες που πήραν την πρόκριση, και 11 ομάδες που προκρίθηκαν, στην κατηγορία των αποκλεισμένων ομάδων. Συνολικά, ταξινόμησε ορθά 100 παρατηρήσεις και 19 λανθασμένα, με βάση το σύνολο δοκιμών.

Ακολουθεί τώρα ο πίνακας με τις τιμές των μέτρων αξιολόγησης για την κατηγοριοποίηση, για το Final Four. Φαίνεται εδώ να μην έχει πραγματοποιηθεί μια αποτελεσματική κατηγοριοποίηση. Το μοντέλο αποδίδει αρκετά καλά όσον αφορά τη συνολική ακρίβειά του, αλλά δυσκολεύεται να πραγματοποιήσει ακριβείς θετικές προβλέψεις.

Μέτρα Αξιολόγησης	Τιμές
Accuracy	**0,840**
Precision	**0,579**
Recall	**0,500**
F1 score	**0,537**

Πίνακας 6.9 : Μέτρα αξιολόγησης για την κατηγοριοποίηση του Final Four.

ΚΕΦΑΛΑΙΟ 7ο - ΣΥΜΠΕΡΑΣΜΑΤΑ

Για τις ανάγκες αυτής της ανάλυσης, συλλέξαμε δεδομένα από το επίσημο site της EuroLeague. Τα δεδομένα μας αφορούσαν όλες τις σεζόν από την αρχή της διοργάνωσης, τη σεζόν 2000-2001 έως και τη περσινή σεζόν 2021-2022. Η μελέτη επικεντρώθηκε κυρίως στην εύρεση των σημαντικών, προβλεπτικών παραγόντων για τη πρόκριση των ομάδων στα Playoffs και στο Final Four. Εξετάσαμε παράγοντες όπως, οι συνολικοί πόντοι μιας ομάδας, τα ποσοστά εύστοχων δίποντων και τρίποντων, τα ποσοστά ευστοχίας στις ελεύθερες βολές, οι ασίστ, τα λάθη, τα κλεψίματα, τα μπλοκ, τα αμυντικά, επιθετικά και συνολικά ριμπάουντ μιας ομάδας, τα φάουλ στα οποία υπέπεσε μια ομάδα, αλλά και μεταβλητές που παρείχαν πιο συνδυαστική πληροφορία, όπως ο ειδικός δείκτης αξιολόγησης PIR. Στην πρώτη φάση της ανάλυσης μας, η μεταβλητή απόκρισης που χρησιμοποιήσαμε ήταν η πρόκριση ή όχι της ομάδας στις επιμέρους φάσεις της διοργάνωσης, ενώ στην συνέχεια εφαρμόσαμε τεχνικές μηχανικής μάθησης προκειμένου να εξάγουμε χρήσιμη γνώση/πληροφορία αναφορικά με τα δεδομένα μας.

Στο δεύτερο κεφάλαιο παρουσιάσαμε αναλυτικά μια ιστορική αναδρομή στην EuroLeague και στο format της διοργάνωσης στο πέρασμα των χρόνων. Στη συνέχεια, παρουσιάσαμε τις ομάδες που έχουν συμμετάσχει διαχρονικά στη διοργάνωση, καθώς και τις μεταβλητές που χρησιμοποιήσαμε. Τέλος, θεωρήσαμε σημαντικό να γίνει αναφορά σε υπάρχοντα επιστημονικά άρθρα και μελέτες, τα οποία είναι σχετικά με το θέμα της παρούσας εργασίας, καθώς και στα συμπεράσματα στα οποία αυτά κατέληγαν.

Στο τρίτο κεφάλαιο, παρουσιάστηκαν αναλυτικά κάποια περιγραφικά μέτρα για τις επεξηγηματικές μεταβλητές μας, σε σχέση με τις μεταβλητές απόκρισης Playoffs / Quarter-Finals (πρόκριση ή όχι στα playoffs) και Final 4 / Semi-Finals (πρόκριση στο Final Four). Οι αναλύσεις σε αυτό το κεφάλαιο βασίστηκαν στο διαχωρισμό των ομάδων σε αυτές που προκρίθηκαν στα playoffs και στο Final Four, και σε αυτές που δεν προκρίθηκαν. Παράλληλα, δόθηκε και μια διαγραμματική απεικόνιση των δεδομένων των αναλύσεων μας, ώστε τα αποτελέσματα να γίνουν ακόμη πιο κατανοητά και συγκρίσιμα από τον αναγνώστη. Αρχικά, πραγματοποιήσαμε έλεγχο για την ύπαρξη ελλιπών τιμών στα δεδομένα μας, για κάθε ανάλυση μας. Τα αποτελέσματα έδειξαν ότι υπήρχαν missing values, καθώς από τη σεζόν 2001-2002 ως τη σεζόν 2003-2004 δεν υπήρξε φάση προημιτελικών/playoffs, αλλά κατευθείαν φάση ημιτελικών, και συνεπώς οι ελλιπείς τιμές εμφανίστηκαν για τη μεταβλητή Playoffs / Quarter-Finals. Αφετέρου, η σεζόν 2019-2020 διακόπηκε οριστικά νωρίτερα του αναμενόμενου, λόγω της πανδημίας του COVID-19 και συνεπώς, ούτε εκεί πραγματοποιήθηκαν οι φάσεις των playoffs και του Final Four. Για αυτό το λόγο, και επειδή είχαμε δεδομένα από πολλές σεζόν, αποφασίστηκε να μην ληφθούν υπόψιν οι σεζόν 2001-2004 , καθώς και η σεζόν 2019-2020.

Στη συνέχεια, εξετάσαμε τα χαρακτηριστικά των ομάδων, για όλες τις σεζόν, με σκοπό να δούμε πώς μεταβάλλονταν αυτά στο πέρασμα των χρόνων. Συγκεκριμένα, καταγράψαμε το μέσο όρο όλων των στατιστικών στοιχείων, για όλες τις ομάδες στο σύνολο των σεζόν, ενώ δεν εξαιρέσαμε καμία σεζόν και συνεπώς είχαμε δεδομένα για 22 αγωνιστικές σεζόν. Χρησιμοποιήσαμε αρχικά κάποια γραφήματα χρονοσειρών, προκειμένου να αποκτήσουμε μια γενικότερη οπτική για το πώς μεταβάλλεται κάθε στατιστικό στοιχείο με τη πάροδο των χρόνων. Κάποια συμπεράσματα στα οποία καταλήξαμε ήταν τα εξής : Από την σεζόν 2010-2011 και μετά, υπήρχε μια απότομη, ανοδική τάση για τους πόντους, για τα ποσοστά εύστοχων δίποντων και τριπόντων, τα ποσοστά εύστοχων ελευθέρων βολών, για τα αμυντικά ριμπάουντ, τα συνολικά ριμπάουντ, για τις ασίστ και το συνολικό δείκτη PIR των ομάδων. Σχολιάσαμε αναλυτικά το παραπάνω συμπέρασμα στην ενότητα 3.2.1, και προσθέσαμε πως ίσως να οφειλόταν στο γεγονός πως από τη σεζόν 2010-2011, υπήρξαν κάποιες αλλαγές στους κανονισμούς, οι οποίες είναι πιθανό να οδήγησαν σε γρηγορότερες επιθέσεις από τις ομάδες, καθώς και αύξηση στον αριθμό των κατοχών της μπάλας, όπως και των ριμπάουντ. Αντιθέτως, υπήρχε πτωτική πορεία για τα κλεψίματα, τα λάθη, για τα φάουλ στα οποία υπέπεσαν οι ομάδες και για τα φάουλ που δέχθηκαν, καθώς περνούσαν οι σεζόν. Τα μπλοκ τα οποία έκαναν οι ομάδες και τα μπλοκ που δέχθηκαν φαίνεται να έφτασαν τη μέγιστη τιμή τους γύρω στην σεζόν 2014-2015 και μετά άρχισαν να ακολουθούν μια φθίνουσα πορεία. Τα επιθετικά ριμπάουντ παρουσίασαν πολύ μεγάλη διακύμανση.

Στη συνέχεια αυτής της ανάλυσης, μέσω scatter plots, εξετάσαμε τη σχέση του PIR με τις υπόλοιπες μεταβλητές μας και βγάλαμε τα εξής συμπεράσματα : Ο δείκτης αξιολόγησης φάνηκε να έχει ισχυρή, θετική συσχέτιση με τους πόντους, το ποσοστό εύστοχων δίποντων, το ποσοστό εύστοχων τρίποντων, το ποσοστό εύστοχων βολών και τις ασίστ. Φάνηκε να υπάρχει χαμηλή, θετική συσχέτιση ανάμεσα στο PIR και στα επιθετικά ριμπάουντ, στα αμυντικά ριμπάουντ, όπως και στα συνολικά ριμπάουντ. Ο δείκτης PIR φαίνεται να έχει χαμηλή, αρνητική συσχέτιση με τα κλεψίματα, τα μπλοκ υπέρ της ομάδας, τα μπλοκ κατά της ομάδας, τα φάουλ στα οποία υπέπεσε μια ομάδα και τα φάουλ που κέρδισε μια ομάδα. Τέλος, φάνηκε να υπάρχει υψηλή, αρνητική συσχέτιση ανάμεσα στο δείκτη και στα λάθη. Επιπροσθέτως, μέσω της χρήσης των Chernoff faces, συμπεράναμε πως οι σεζόν 2011-2014 έμοιαζαν αρκετά μεταξύ τους, όπως παραδείγματος χάρη στους πόντους, στα εύστοχα δίποντα, τρίποντα και ελεύθερες βολές, στα μπλοκ υπέρ των ομάδων. Ομοίως, φάνηκε να έχουν αρκετά κοινά στοιχεία οι σεζόν 2000-2001 με 2003-2004, οι σεζόν 2004-2005 με 2009-2010, οι σεζόν 2014-2015 με 2016-2017 και οι σεζόν 2017-2018 με 2021-2022.

Για την επόμενη ανάλυση του κεφαλαίου, υπολογίσαμε για κάθε ομάδα, τους μέσους όρους των πόντων και του PIR, των πέντε καλύτερων παικτών τους, για όλες τις σεζόν. Θέλαμε να μελετήσουμε πώς αυτοί οι μ.ό. διαχώριζαν τις ομάδες, αναλόγως με το αν προκρίθηκαν στα playoffs και στο Final Four. Όπως προαναφέρθηκε, στις αναλύσεις μας δε λήφθηκαν υπόψιν οι σεζόν 2001-2004 και η σεζόν 2019-2020. Αρχικά, χρησιμοποιώντας θηκογράμματα, είδαμε πως η διάμεσος για τις ομάδες που προκρίθηκαν, βρισκόταν υψηλότερα από ότι για τις ομάδες που δεν προκρίθηκαν, τόσο

αναφορικά για τους πόντους, όσο και το δείκτη PIR, κάτι το οποίο ήταν και αναμενόμενο. Παράλληλα, εμφανίστηκαν και έκτροπες τιμές στις περιπτώσεις αυτές, οι οποίες παρουσιάστηκαν αναλυτικά στην ενότητα 3.2.2.

Φτιάχνοντας τα κατάλληλα ιστογράμματα και διαγράμματα πυκνότητας στην ίδια ανάλυση, συμπεράναμε πως οι παίκτες των ομάδων που δε προκρίθηκαν στα playoffs αρχικά, σκόραραν 10-12 πόντους κατά μέσο όρο, με μια συχνότητα μεγαλύτερη του 25%, και αυτοί των ομάδων που προκρίθηκαν, με συχνότητα περίπου 30%. Παράλληλα, οι μέσοι όροι του PIR φάνηκε να κατανέμονται κανονικά, τόσο για τις ομάδες που δεν προκρίθηκαν στα playoffs, όσο και για αυτές που προκρίθηκαν. Αναφορικά με το Final Four, οι παίκτες των ομάδων που δε προκρίθηκαν σκόραραν 10-12 πόντους κατά μέσο όρο, με μια συχνότητα μεγαλύτερη του 25%, και αυτοί των ομάδων που προκρίθηκαν, με συχνότητα κοντά στο 34%, ενώ οι μέσοι όροι του PIR κατανεμόντουσαν κανονικά, τόσο για τις ομάδες που δε προκρίθηκαν στο Final Four, όσο και για αυτές που προκρίθηκαν.

Η επόμενη ανάλυση του τρίτου κεφαλαίου αφορούσε τους δέκα κορυφαίους παίκτες σε πόντους και PIR ανά σεζόν. Χρησιμοποιώντας violin plots, καταλήξαμε σε αναμενόμενα αποτελέσματα. Η διάμεσος που αφορούσε ομάδες που προκρίθηκαν στις φάσεις της διοργάνωσης βρισκόταν υψηλότερα (σε μια περίπτωση ήταν ελαφρώς χαμηλότερα), τόσο αναφορικά με τους καλύτερους παίκτες σε πόντους, όσο και αυτούς σε PIR. Το σημαντικότερο συμπέρασμα στο οποίο καταλήξαμε ήταν πως παίκτες με αρκετά υψηλές επιδόσεις, τόσο σε πόντους, όσο και σε PIR, αντιστοιχούσαν σε ομάδες που δε προκρίθηκαν στις επιμέρους φάσεις της διοργάνωσης. Συνεπώς, συμπεράναμε πως για μια ομάδα να προκριθεί στις επόμενες φάσεις, δεν ήταν αρκετός ένας παίκτης με υψηλό δείκτη PIR, αλλά περισσότεροι παίκτες με μια σχετικά καλή επίδοση.

Στη τελευταία ανάλυση αυτού του κεφαλαίου, εξετάσαμε τους πενήντα καλύτερους παίκτες με βάση το PIR διαχρονικά και χρησιμοποιήσαμε donut charts ως εποπτικά μέσα. Παρατηρήσαμε πως στο σύνολο των πενήντα παικτών, οι μισοί (είκοσι πέντε) ανήκαν σε ομάδες που δεν προκρίθηκαν στα playoffs, και οι άλλοι μισοί σε ομάδες που προκρίθηκαν. Αντίστοιχα, οι δώδεκα ανήκαν σε ομάδες που προκρίθηκαν στο Final Four, και οι υπόλοιποι τριάντα οχτώ σε ομάδες που δεν προκρίθηκαν. Παρατηρήσαμε ακόμα πως τέσσερις στους πέντε παίκτες διαχρονικά, με το υψηλότερο PIR, ανήκαν σε ομάδες που δεν προκρίθηκαν στα playoffs και ένας μόνο παίκτης, ο Dejan Tomasevic, ανήκε σε ομάδα που προκρίθηκε (Buducnost, σεζόν 2000-2001). Επιπλέον, και οι πρώτοι πέντε παίκτες διαχρονικά με το υψηλότερο PIR, ανήκαν σε ομάδες που δεν προκρίθηκαν στο Final Four, κάτι που επιβεβαιώνει το προηγούμενο συμπέρασμα μας πως δεν αρκεί ένας μόνο παίκτης με υψηλό PIR για να προκριθεί μια ομάδα.

Στο τέταρτο κεφάλαιο, αρχικά πραγματοποιήθηκε έλεγχος κανονικότητας για τις ποσοτικές μεταβλητές μας, για κάθε ανάλυση που πραγματοποιήθηκε ξεχωριστά στο προηγούμενο κεφάλαιο. Στη συνέχεια, υπολογίσαμε τους συντελεστές συσχέτισης των μεταβλητών και τέλος, κάναμε ελέγχους για την ισότητα των μέσων τιμών των πόντων και του δείκτη PIR, των δύο δειγμάτων (ομάδες που προκρίθηκαν και ομάδες που δεν προκρίθηκαν στις επιμέρους φάσεις), οπουδήποτε αυτό κρίθηκε απαραίτητο.

Αναφορικά με τους μέσους όρους των μεταβλητών των πόντων, των ποσοστών εύστοχων δίποντων, ποσοστών των ελεύθερων βολών, των επιθετικών ριμπάουντ, των αμυντικών ριμπάουντ, των συνολικών ριμπάουντ, των ασίστ, των λαθών, των μπλοκ υπέρ της ομάδας, των μπλοκ κατά της ομάδας, των φάουλ στα οποία υπέπεσε η ομάδα, των φάουλ που δέχτηκε η ομάδα, του δείκτη PIR, καθώς και για τη μεταβλητή για τις σεζόν, είδαμε πως δε μπορούσαμε να απορρίψουμε την υπόθεση της κανονικότητας σε επίπεδο σημαντικότητας 5%. Αντιθέτως, για τους μέσους όρους των ποσοστών των εύστοχων τριπόντων και των κλεψιμάτων, απορρίψαμε την υπόθεση της κανονικότητας. Ο δείκτης PIR είχε υψηλή, θετική συσχέτιση με τους πόντους, το ποσοστό δίποντων, το ποσοστό των ελευθέρων βολών και με τις ασίστ. Παρουσίασε ασθενή, θετική συσχέτιση με τα επιθετικά ριμπάουντ, τα αμυντικά ριμπάουντ και τα συνολικά ριμπάουντ. Παράλληλα, ο δείκτης είχε πολύ ασθενή, αρνητική συσχέτιση με τα μπλοκ των ομάδων και τα μπλοκ κατά των ομάδων, ενώ είχε υψηλή, αρνητική συσχέτιση με τα λάθη στα οποία υπέπεσαν οι ομάδες. Τέλος, παρουσίασε μια ασθενή, αρνητική συσχέτιση τόσο με τα φάουλ που κέρδισαν οι ομάδες, όσο και με αυτά που έκαναν. Επιπλέον, ο δείκτης PIR παρουσίασε μια μέτρια συσχέτιση με τη κατηγορική μεταβλητή των σεζόν, μια πολύ υψηλή, θετική συσχέτιση με το ποσοστό εύστοχων τριπόντων και μια ασθενή, αρνητική συσχέτιση με τα κλεψίματα. Αναφορικά με τις σεζόν ειδικότερα, όπως είδαμε και από το σχήμα 3.1, ο δείκτης PIR είχε μια πτωτική τάση μέχρι και την σεζόν 2010-2011, ενώ αμέσως μετά, παρουσίασε μια πολύ απότομη ανοδική πορεία, μέχρι τη σεζόν 2016-2017. Εμφανίστηκαν και ενδιαφέρουσες συσχετίσεις μεταξύ των υπόλοιπων μεταβλητών, οι οποίες παρουσιάζονται αναλυτικά στην ενότητα 4.2.1.

Η επόμενη ανάλυση αφορούσε τους πέντε καλύτερους παίκτες κάθε ομάδας, με βάση το PIR. Αναφορικά με τους παίκτες των ομάδων που δεν προκρίθηκαν στα playoffs, η υπόθεση της κανονικότητας για τους πόντους απορρίφθηκε, σε αντίθεση με το δείκτη PIR, όπου δεν μπορέσαμε να απορρίψουμε την μηδενική υπόθεση. Οι δύο αυτές μεταβλητές παρουσίασαν υψηλή, θετική συσχέτιση (0,73) με βάση το συντελεστή του Spearman. Αντίστοιχα, για τις ομάδες που προκρίθηκαν στα playoffs, είδαμε ότι οι πόντοι και ο δείκτης PIR προσαρμόζονται ικανοποιητικά στην κανονική κατανομή. Ακόμα, παρουσίασαν μια υψηλή, θετική συσχέτιση (0,79) επίσης, με βάση το συντελεστή του Pearson αυτή τη φορά. Οι μέσες τιμές των πόντων και του δείκτη αξιολόγησης διέφεραν σημαντικά, αφού οι p-value του ελέγχου "Welch Two Sample t-test" ήταν μικρότερες από το επίπεδο σημαντικότητας 5%. Παράλληλα, οι ομάδες που προκρίθηκαν είχαν μεγαλύτερο μέσο όρο πόντων και PIR.

Για τους παίκτες των ομάδων που δεν προκρίθηκαν στο Final Four, οι πόντοι τους δεν προέρχονταν από κανονική κατανομή, σε αντίθεση με το δείκτη PIR. Οι πόντοι και ο δείκτης PIR παρουσίασαν υψηλή, θετική συσχέτιση (0,74) με βάση το συντελεστή του Spearman. Αντίστοιχα, για τις ομάδες που προκρίθηκαν στο Final Four, τόσο για τους πόντους, όσο και για το δείκτη PIR, δε μπορέσαμε να απορρίψουμε την υπόθεση της κανονικότητας των μεταβλητών. Παρουσίασαν και εδώ υψηλή, θετική συσχέτιση (0,77) με βάση τώρα το συντελεστή του Pearson. Συμπεράναμε πως οι μέσες τιμές των

πόντων δε διέφεραν σημαντικά, σε αντίθεση με τις μέσες τιμές του δείκτη, καθώς οι ομάδες που προκρίθηκαν στο Final Four είχαν μεγαλύτερο μέσο όρο PIR, όπως ήταν και αναμενόμενο.

Ακολούθησε η ανάλυση η οποία αφορούσε αρχικά τους δέκα καλύτερους παίκτες σε πόντους, ανά σεζόν. Για τις ομάδες των παικτών, που δεν προκρίθηκαν στα playoffs, τόσο για τους πόντους, όσο και για το δείκτη PIR, απορρίφθηκε η υπόθεση της κανονικότητας. Παρουσίασαν μια μέτρια, θετική συσχέτιση (0,52) με βάση το συντελεστή του Spearman. Αντίστοιχα, αναφορικά με τις ομάδες που προκρίθηκαν στα playoffs, για τους πόντους απορρίφθηκε η υπόθεση της κανονικότητας, κάτι που δεν μπορέσαμε να κάνουμε για το δείκτη PIR. Αυτές οι μεταβλητές παρουσίασαν υψηλή, θετική συσχέτιση (0,66) με βάση τον ίδιο συντελεστή. Οι μέσες τιμές των πόντων δεν διέφεραν σημαντικά, όπως και οι μέσες τιμές του δείκτη PIR.

Για τις ομάδες που δεν προκρίθηκαν στο Final Four, τόσο για τους πόντους, όσο και για το δείκτη PIR, απορρίφθηκε η υπόθεση της κανονικότητας. Παρουσίασαν επίσης μια μέτρια, θετική συσχέτιση (0,59) με βάση το συντελεστή του Spearman. Αντίστοιχα, για τις ομάδες που προκρίθηκαν στο Final Four, τόσο για τους πόντους, όσο και για το δείκτη PIR, δεν μπορέσαμε να απορρίψουμε την υπόθεση της κανονικότητας, ενώ εμφάνισαν μια υψηλή, θετική συσχέτιση (0,63) επίσης, με βάση το συντελεστή του Pearson. Είδαμε επίσης πως οι μέσες τιμές τόσο για τους πόντους, όσο και για το PIR δεν διέφεραν σημαντικά.

Ο έλεγχος για τους δέκα καλύτερους παίκτες σε PIR ανά σεζόν, έδειξε πως για τις ομάδες των παικτών, που δεν προκρίθηκαν στα playoffs, τόσο για τους πόντους, όσο και για το δείκτη PIR, απορρίφθηκε η υπόθεση της κανονικότητας. Οι δύο αυτές μεταβλητές παρουσίασαν μέτρια, θετική συσχέτιση (0,47) με βάση το συντελεστή του Spearman. Αντίστοιχα, αναφορικά με τις ομάδες που προκρίθηκαν στα playoffs, η υπόθεση της κανονικότητας απορρίφθηκε επίσης τόσο για τους πόντους, όσο και για το δείκτη PIR, ενώ η συσχέτιση μεταξύ τους ήταν μέτρια και θετική (0,44) με βάση τον ίδιο συντελεστή. Τόσο οι μέσες τιμές των πόντων, όσο και του δείκτη αξιολόγησης, δε διέφεραν σημαντικά.

Για τις ομάδες που δεν προκρίθηκαν στο Final Four, τόσο για τους πόντους, όσο και για το δείκτη PIR, απορρίφθηκε και εδώ η υπόθεση της κανονικότητας. Παρουσίασαν μια μέτρια, θετική συσχέτιση (0,47) με βάση το συντελεστή του Spearman. Αντίστοιχα, για τις ομάδες που προκρίθηκαν στο Final Four, δε μπορέσαμε να απορρίψουμε την υπόθεση της κανονικότητας για τους πόντους, ενώ το πετύχαμε για το δείκτη PIR. Παράλληλα, εμφάνισαν μια μέτρια, θετική συσχέτιση (0,42) με βάση το συντελεστή του Spearman. Τόσο οι μέσες τιμές των πόντων, όσο και του δείκτη αξιολόγησης, δε διέφεραν σημαντικά και σε αυτή τη περίπτωση.

Η τελευταία ανάλυση αφορούσε τους πενήντα καλύτερους παίκτες σύμφωνα με το PIR. Για τις ομάδες των παικτών, που δεν προκρίθηκαν στα playoffs, η υπόθεση της κανονικότητας δεν μπορούσε να απορριφθεί για τους πόντους, ενώ απορρίφθηκε για

το δείκτη PIR. Οι δύο μεταβλητές παρουσίασαν μέτρια, θετική συσχέτιση (0,46) με βάση το συντελεστή του Spearman. Αντίστοιχα, για τις ομάδες που προκρίθηκαν στα playoffs, τόσο για τους πόντους, όσο και για το δείκτη PIR, απορρίφθηκε η υπόθεση της κανονικότητας, ενώ εμφάνισαν και μια ασθενή, θετική συσχέτιση (0,29) με βάση τον ίδιο συντελεστή. Τόσο οι μέσες τιμές των πόντων, όσο και του δείκτη αξιολόγησης, δε διέφεραν σημαντικά.

Για τις ομάδες που δεν προκρίθηκαν στο Final Four, η υπόθεση της κανονικότητας για τους πόντους δεν μπορούσε να απορριφθεί, ενώ για το δείκτη PIR απορρίφθηκε. Παράλληλα, παρουσίασαν μια μέτρια, θετική συσχέτιση (0,42) με βάση το συντελεστή του Spearman. Τέλος, αναφορικά με τις ομάδες που προκρίθηκαν στο Final Four, δε μπορούσαμε να απορρίψουμε την υπόθεση της κανονικής κατανομής για τους πόντους, ενώ το πετύχαμε για το δείκτη PIR. Η συσχέτιση μεταξύ τους ήταν μέτρια και θετική (0,54) με βάση το συντελεστή του Spearman. Εμφανίστηκε ισχυρή ένδειξη ότι οι μέσες τιμές των πόντων διέφεραν σημαντικά, σε αντίθεση με τους μέσους όρους των δεικτών PIR. Ένα αρκετά ενδιαφέρον αποτέλεσμα ήταν πως οι ομάδες που δεν προκρίθηκαν στο Final Four, φάνηκε να έχουν μεγαλύτερο μέσο όρο πόντων, σε σχέση με αυτές που προκρίθηκαν.

Στο πέμπτο κεφάλαιο θέλαμε να εξετάσουμε ποιες από τις μεταβλητές μας έπαιζαν καθοριστικό ρόλο στο αν μια ομάδα προκρίνεται ή όχι, στις δύο φάσεις της διοργάνωσης. Για αυτό το λόγο, προσαρμόσαμε τα κατάλληλα γενικευμένα γραμμικά μοντέλα με μεταβλητές απόκρισης τις κατηγορικές μεταβλητές Playoffs / Quarter-Finals και Final 4 / Semi-Finals , οι οποίες αποτελούσαν ενδείξεις για το αν οι ομάδες προκρίθηκαν (ή όχι) στις αντίστοιχες φάσεις της διοργάνωσης. Η σκέψη μας ήταν να προσαρμόσουμε συνολικά τέσσερα μοντέλα, δύο για τη φάση των playoffs (ένα χωρίς, και ένα με αλληλεπιδράσεις δευτέρου βαθμού) και δύο για τη φάση του Final Four, ίδιας λογικής.

Για τα playoffs, το τελικό μοντέλο που προσαρμόσαμε είχε ως στατιστικά σημαντικές μεταβλητές το ποσοστό εύστοχων δίποντων, τα αμυντικά ριμπάουντ, τα λάθη, τα φάουλ που κέρδισε μια ομάδα, το ποσοστό εύστοχων τριπόντων και τις ασίστ. Το μοντέλο είχε καλή προσαρμογή, το οποίο φάνηκε και από τον έλεγχο Hosmer-Lemeshow (p-value = 0,3569), και από τα μέτρα προσαρμογής, αλλά και από τη καμπύλη ROC (AUC = 0,87). Αντιθέτως, δε μπορέσαμε να καταλήξουμε σε ένα ικανοποιητικό μοντέλο με αλληλεπιδράσεις δευτέρου βαθμού, καθώς όλες οι πιθανές μεταβλητές αλληλεπιδράσεων έβγαιναν στατιστικώς μη σημαντικές.

Για το Final Four, το τελικό μοντέλο που προσαρμόσαμε είχε ως στατιστικά σημαντικές μεταβλητές το ποσοστό εύστοχων δίποντων, τα αμυντικά ριμπάουντ, τα λάθη, τα φάουλ που κέρδισε μια ομάδα και το ποσοστό εύστοχων τριπόντων. Το μοντέλο είχε καλή προσαρμογή και εδώ, το οποίο φάνηκε από τον έλεγχο Hosmer-Lemeshow (p-value = 0,7704), από τα μέτρα προσαρμογής, αλλά και από τη καμπύλη ROC (AUC = 0,9039).

Το αντίστοιχο μοντέλο με αλληλεπιδράσεις δευτέρας τάξης είχε ως στατιστικά σημαντικές μεταβλητές το ποσοστό εύστοχων δίποντων, τα αμυντικά ριμπάουντ, τα λάθη, τα φάουλ που κέρδισε μια ομάδα, το ποσοστό εύστοχων τριπόντων, καθώς και τις αλληλεπιδράσεις ανάμεσα στα τρίποντα και τα φάουλ που κέρδισε η ομάδα (3PT% × FD), και ανάμεσα στα δίποντα και στα τρίποντα (2PT% × 3PT%). Το μοντέλο είχε καλή προσαρμογή και σε αυτή τη περίπτωση, το οποίο προέκυψε και από τον έλεγχο Hosmer-Lemeshow (p-value = 0,3569), και από τα αντίστοιχα μέτρα προσαρμογής, αλλά και από τη καμπύλη ROC (AUC = 0,915).

Στο έκτο κεφάλαιο μέσω της χρήσης τεχνικών μηχανικής μάθησης και της κατάλληλης επεξεργασίας των δεδομένων μας, αποσκοπούσαμε να εξάγουμε χρήσιμη γνώση/πληροφορία. Ειδικότερα, αφού πρώτα καταλήξαμε στις κατάλληλες μεταβλητές (feature selection) που ήταν χρήσιμες για τους αλγορίθμους μας, χρησιμοποιήσαμε τεχνικές ομαδοποίησης (K-means και Hierarchical Agglomerative clustering) για να εντοπίσουμε μοτίβα και ομάδες παρατηρήσεων με παρόμοια χαρακτηριστικά μέσα στα δεδομένα μας, καθώς και τεχνικές κατηγοριοποίησης (Support Vector Machines και Random Forest), προκειμένου να προσαρμόσουμε κατάλληλα μοντέλα ταξινόμησης για το αν μια ομάδα θα προκρινόταν (ή όχι) στις δύο φάσεις της διοργάνωσης που εξετάζουμε.

Αρχικά, με τη μέθοδο Boruta, καταλήξαμε πως για τη φάση των playoffs, οι πιο σημαντικές μεταβλητές για τα μοντέλα μας ήταν ο δείκτης PIR, το ποσοστό εύστοχων δίποντων, οι πόντοι, τα φάουλ που κέρδισαν οι ομάδες, το ποσοστό εύστοχων τριπόντων, τα αμυντικά ριμπάουντ, τα λάθη, τα μπλοκ και ο μέσος όρος του δείκτη PIR των πέντε καλύτερων παικτών κάθε ομάδας, ενώ για την φάση του Final Four, οι πιο σημαντικές μεταβλητές ήταν ο δείκτης PIR, το ποσοστό εύστοχων δίποντων, οι πόντοι, τα φάουλ που κέρδισαν οι ομάδες, το ποσοστό εύστοχων τριπόντων, οι ασίστ, ο μέσος όρος του δείκτη PIR των πέντε καλύτερων παικτών κάθε ομάδας, τα μπλοκ, τα αμυντικά ριμπάουντ, τα συνολικά ριμπάουντ και τα λάθη.

Αναφορικά με τα playoffs, καταλήξαμε και μέσω της μεθόδου των K-μέσων και μέσω της ιεραρχικής, συσσωρευτικής συσταδοποίησης πως το βέλτιστο πλήθος συστάδων για τα δεδομένα μας ήταν δύο. Οι δύο ομαδοποιήσεις μας έδωσαν καλά αποτελέσματα σύμφωνα με τον προσαρμοσμένο δείκτη Rand (0,796 και 0,923 αντίστοιχα), αλλά φάνηκαν λιγότερο αποτελεσματικές με βάση τα άλλα δυο μέτρα που χρησιμοποιήσαμε, το συντελεστή σιλουέτας και το δείκτη Dunn. Παρόμοια αποτελέσματα πήραμε και για τη φάση του Final Four. Οι δύο μέθοδοι μας έδειξαν ως βέλτιστο πλήθος συστάδων τις δυο, με πολύ καλά αποτελέσματα σύμφωνα με το προσαρμοσμένο δείκτη Rand (0,956 και 0,819 αντίστοιχα), αλλά όχι τόσο καλές σύμφωνα με τα άλλα μέτρα αξιολόγησης. Συμπερασματικά, σύμφωνα με τον προσαρμοσμένο δείκτη Rand, φαίνεται να πετύχαμε τη καλύτερη ομαδοποίηση των δεδομένων μας, για τη φάση του Final Four, με τη μέθοδο K-Means. Το να έχουμε ως βέλτιστο πλήθος συστάδων τις δυο σε κάθε περίπτωση, ήταν κάτι που φάνταζε εξ αρχής λογικό, αφού τα δεδομένα μας χωριζόντουσαν εκ των προτέρων σε δύο μεγάλες

ομάδες/κατηγορίες, τις ομάδες που προκρίθηκαν και αυτές που δε προκρίθηκαν στις επιμέρους φάσεις.

Χρησιμοποιήσαμε στη συνέχεια τις μεθόδους των Support Vector Machines και Random Forest, προκειμένου να φτιάξουμε αποτελεσματικά μοντέλα ταξινόμησης των δεδομένων μας στις δύο κλάσεις που είχαμε. Αναφορικά με τα playoffs, και οι δύο μέθοδοι μας έδωσαν σχετικά καλές ταξινομήσεις, κρίνοντας την ακρίβειά (precision) τους (0,675 και 0,737 αντίστοιχα). Υπενθυμίζουμε ότι η ακρίβεια αξιολογεί την ικανότητα του μοντέλου να εντοπίζει ορθά τις ομάδες που προκρίθηκαν στις φάσεις της διοργάνωσης που εξετάζουμε. Ομοίως για το Final Four, είχαμε σχετικά αποτελεσματικά μοντέλα κατηγοριοποίησης (0,786 και 0,579 αντίστοιχα οι τιμές της ακρίβειας) για τις δύο μεθόδους. Συμπερασματικά, συγκρίνοντας την ακρίβεια του κάθε μοντέλου, που είναι ιδιαίτερα χρήσιμή όταν τα δεδομένα είναι μη ισορροπημένα (imbalanced) όπως εδώ, φαίνεται ότι το καλύτερο μοντέλο ταξινόμησης ήταν αυτό για τη φάση του Final Four, με τη μέθοδο των Support Vector Machines.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική

Αντζουλάκος, Δ. (2021-2022). Διαφάνειες μαθήματος «Ανάλυση δεδομένων με τη χρήση στατιστικών πακέτων: Εισαγωγή στην R», ΠΜΣ «Εφαρμοσμένη Στατιστική», Πανεπιστήμιο Πειραιώς.

Καλλιακμάνης, Δ. (2020). «Στατιστικά μοντέλα για την απόδοση μιας ομάδας στο μπάσκετ: Ποια στατιστικά στοιχεία είναι καθοριστικά για τη απόδοση της ομάδας, σε ετήσια βάση», Διπλωματική εργασία, ΠΜΣ «Εφαρμοσμένη Στατιστική», Πανεπιστήμιο Πειραιώς.

Κούτρας, Μ. (2022). Σημειώσεις μαθήματος «Εφαρμοσμένη Πολυμεταβλητή Ανάλυση», ΠΜΣ «Εφαρμοσμένη Στατιστική», Πανεπιστήμιο Πειραιώς.

Κούτρας, Μ. (2021-2022). Σημειώσεις μαθήματος «Ανάλυση παλινδρόμησης και ανάλυση διακύμανσης», ΠΜΣ «Εφαρμοσμένη Στατιστική», Πανεπιστήμιο Πειραιώς.

Μανωλέσου, Α. (2015). «Στατιστικοί έλεγχοι κανονικότητας», Αποθετήριο Κάλλιπος, Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, Αθήνα.

Μπερσίμης, Σ. (2022-2023). Διαφάνειες μαθήματος «Βιοστατιστική και Στατιστικές Μέθοδοι στην Επιδημιολογία», ΠΜΣ «Εφαρμοσμένη Στατιστική», Πανεπιστήμιο Πειραιώς.

Πολίτης, Κ. (2022). Διαφάνειες μαθήματος «Γενικευμένα Γραμμικά Μοντέλα», ΠΜΣ «Εφαρμοσμένη Στατιστική», Πανεπιστήμιο Πειραιώς.

Πολίτης, Κ. (2022). Σημειώσεις μαθήματος «Γενικευμένα Γραμμικά Μοντέλα: Η έννοια της αλληλεπίδρασης», ΠΜΣ «Εφαρμοσμένη Στατιστική», Πανεπιστήμιο Πειραιώς.

Σπυριδάκης, Α. (2022). «Στατιστική ανάλυση για τους παράγοντες που επηρεάζουν την απόφαση των ομάδων ποδοσφαίρου στις Ευρωπαϊκές διοργανώσεις», Διπλωματική εργασία, ΠΜΣ «Εφαρμοσμένη Στατιστική», Πανεπιστήμιο Πειραιώς.

Ξένη

Bishop, C.M. (2006) "Pattern recognition and machine learning.", New York: Springer.

Breiman, L. (2001) "Random Forests", *Machine Learning*, **45**, pp. 5–32 https://doi.org/10.1023/A:1010933404324

Croux, C., and Dehon, C. (2010). "Influence Functions of the Spearman and Kendall Correlation Measures" , *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1585216

Csataljay, G. et al. (2009). "Performance indicators that distinguish winning and losing teams in basketball", *International Journal of Performance Analysis in Sport*, 9(1), pp. 60–66. https://doi.org/10.1080/24748668.2009.11868464

Dogan, I. and Ersoz, Y. (2019). "The important game-related statistics for qualifying next rounds in Euroleague", *Montenegrin Journal of Sports Science and Medicine*, 8(1), pp. 43–50. https://doi.org/10.26773/mjssm.190307

Evans, R. H. (1996). "An Analysis of Criterion Variable Reliability in Conjoint Analysis.", Perceptual and Motor Skills, 82(3), pp. 988–990. https://doi.org/10.2466/pms.1996.82.3.988

García, J. et al. (2013) "Identifying basketball performance indicators in regular season and playoff games", *Journal of Human Kinetics*, 36(1), pp. 161–168. https://doi.org/10.2478/hukin-2013-0016

Glantz, S.A., Slinker, B.K. and Neilands, T.B. (2016) "Primer of Applied Regression & Analysis of variance.", 3rd edn. New York: McGraw-Hill Education.

Hair, J.F. et al. (2019) "Multivariate Data Analysis". Andover, Hampshire, United Kingdom: Cengage.

Han, J., Kamber, M. and Pei, J. (2012) "Data Mining: Concepts and Techniques.", 3rd edn. Cambridge, MA: Morgan Kaufmann.

Hastie, T., Friedman, J. and Tisbshirani, R. (2009) "The elements of Statistical Learning: Data Mining, Inference, and prediction.", New York: Springer.

Ibañez, S.J. et al. (2018). "The impact of rule modifications on elite basketball teams' performance", Journal of Human Kinetics, 64(1), pp. 181–193. https://doi.org/10.1515/hukin-2017-0193

Kraemer, H.C. and Blasey, C.M. (2004) "Centring in regression analyses: A strategy to prevent errors in statistical inference", *International Journal of Methods in Psychiatric Research*, 13(3), pp. 141–151. https://doi.org/10.1002/mpr.170

Kursa, M.B. and Rudnicki, W.R. (2010) "Feature selection with the Boruta Package" *Journal of Statistical Software*, 36(11). https://doi.org/10.18637/jss.v036.i11

Mandić, R. et al. (2019). "Trends in NBA and Euroleague Basketball: Analysis and comparison of statistical data from 2000 to 2017", PLOS ONE, 14(10). https://doi.org/10.1371/journal.pone.0223524

Marmarinos, C. et al. (2016). "Game-related statistics that discriminate playoffs teams from the rest of the competition in Euroleague Basketball", *Journal of Athletic Enhancement*, 05(06). https://doi.org/10.4172/2324-9080.1000245

Mikołajec, K. et al. (2021)."How to win the Basketball Euroleague? Game Performance Determining Sports Results during 2003–2016 matches", *Journal of Human Kinetics*, 77(1), pp. 287–296. https://doi.org/10.2478/hukin-2021-0050

Schölkopf, B. and Smola, A.J. (2002) "Learning with kernels: Support vector machines, regularization, optimization, and beyond.", Cambridge, MA: MIT Press.

Senaviratna, N.A. and A. Cooray, T.M. (2019) "Diagnosing multicollinearity of logistic regression model", Asian Journal of Probability and Statistics, pp. 1–9. https://doi.org/10.9734/ajpas/2019/v5i230132

Štrumbelj, E. et al. (2013). "A decade of euroleague basketball: An analysis of trends and recent rule change effects", *Journal of Human Kinetics*, 38, pp. 183–189. https://doi.org/10.2478/hukin-2013-0058

Tan, P.-N. et al. (2018) "Introduction to data mining.", 2nd edn. Boston, MA: Pearson Education.

Σύνδεσμοι

https://www.euroleaguebasketball.net/euroleague/
https://www.euroleaguebasketball.net/euroleague/teams/
https://en.wikipedia.org/wiki/EuroLeague
https://en.wikipedia.org/wiki/Wild_card_(sports)
https://en.wikipedia.org/wiki/Final_four
https://en.wikipedia.org/wiki/Round-robin_tournament
https://en.wikipedia.org/wiki/Tiebreaker
https://en.wikipedia.org/wiki/ABA_League
https://en.wikipedia.org/wiki/VTB_United_League

https://en.wikipedia.org/wiki/COVID-19
https://en.wikipedia.org/wiki/2022_Russian_invasion_of_Ukraine
https://en.wikipedia.org/wiki/Performance_Index_Rating
https://en.wikipedia.org/wiki/Tendex
https://en.wikipedia.org/wiki/Missing_data
https://en.wikipedia.org/wiki/Imputation_(statistics)
https://en.wikipedia.org/wiki/Key_(basketball)
https://sportsfanfocus.com/restricted-area-basketball/
https://en.wikipedia.org/wiki/Violin_plot

https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot

https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php
https://www.scribbr.com/
https://en.wikipedia.org/wiki/Probit_model

https://en.wikipedia.org/wiki/Science_(journal)
https://en.wikipedia.org/wiki/Training,_validation,_and_test_data_sets

https://www.researchgate.net/figure/Optimal-Hyperplane-and-Margin-of-SVM_fig3_338698374
https://www.linkedin.com/pulse/understanding-confusion-matrix-kartik-lokare
https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm
https://medium.com/analytics-vidhya/machine-learning-decision-trees-and-random-forest-classifiers-81422887a544
https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch9/histo/5214822-eng.htm

https://www.simplilearn.com/tutorials/machine-learning-tutorial/feature-selection-in-machine-learning

https://www.andreaperlato.com/mlpost/feature-selection-using-boruta-algorithm/

ΠΑΡΑΡΤΗΜΑ

```
# Εισαγωγή βιβλιοθηκών/πακέτων
library(readxl)
library(ggplot2)
library(dbplyr)
library(dplyr)
library(tidyr)
library(aplpack)
library(scatterplot3d)
library(purrr)
library(hrbrthemes)
library(plotrix)
library(nortest)
library(GGally)
library(MASS)
library(car)
library(DescTools)
library(ResourceSelection)
library(pROC)
library(Boruta)
library(cluster)
library(factoextra)
library(ggrepel)
library(clValid)
library(fossil)
library(clusterSim)
library(caTools)
library(e1071)
library(caret)
```

```
library(randomForest)


# ΚΕΦΑΛΑΙΟ 3ο

# Εισαγωγή δεδομένων

excel_data3 <- read_excel("C:\\Users\\alex_\\Desktop\\ΔΙΠΛΩΜΑΤΙΚΗ\\means.xlsx")

# μέσοι όροι των στατιστικών στοιχείων ανά σεζόν για ΟΛΕΣ τις ομάδες.

dfdata3 <- data.frame(excel_data3)


sum(is.na(dfdata3))

summary(dfdata3)


#Time Series Plot

attach(dfdata3)

par(mfrow=c(3,5))


plot(Season, Points, type="o", lwd=3,col="black", xlab="Season", ylab="Points")

plot(Season, X2PT.., type="o", lwd=3, col="red", xlab="Season", ylab="2PT %")

plot(Season, X3PT.., type="o", lwd=3, col="blue4", xlab="Season", ylab="3PT %")

plot(Season, FT.., type="o", lwd=3, col="yellow", xlab="Season", ylab="FT %")

plot(Season, OR, type="o", lwd=3, col="maroon4", xlab="Season", ylab="Offensive Rebounds")

plot(Season, DR, type="o", lwd=3, col="green", xlab="Season", ylab="Defensive Rebounds")

plot(Season, TR, type="o", lwd=3, col="orange1", xlab="Season", ylab="Total Rebounds")

plot(Season, AST, type="o", lwd=3, col="cyan", xlab="Season", ylab="Assists")

plot(Season, STL, type="o", lwd=3, col="chocolate4", xlab="Season", ylab="Steals")

plot(Season, TO, type="o", lwd=3, col="darkgreen", xlab="Season", ylab="Turnovers")

plot(Season, BLK, type="o", lwd=3, col="paleturquoise4", xlab="Season", ylab="Blocks")

plot(Season, BLKA, type="o", lwd=3, col="deeppink", xlab="Season", ylab="Blocks Against")

plot(Season, FC, type="o", lwd=3, col="wheat3", xlab="Season", ylab="Fouls Committed")

plot(Season, FD, type="o", lwd=3, col="darkcyan", xlab="Season", ylab="Fouls Drawn")
```

```
plot(Season, PIR, type="o", lwd=3 ,col="lightgoldenrod", xlab="Season", ylab="PIR")
```

 # Scatter Plots των στατιστικών στοιχείων με το δείκτη PIR.

```
par(mfrow=c(3,5))

plot(Points,PIR)
abline(lm(PIR~Points),col="red")
plot(X2PT..,PIR)
abline(lm(PIR~X2PT..),col="red")
plot(X3PT..,PIR)
abline(lm(PIR~X3PT..),col="red")
plot(FT..,PIR)
abline(lm(PIR~FT..),col="red")
plot(OR,PIR)
abline(lm(PIR~OR),col="red")
plot(DR,PIR)
abline(lm(PIR~DR),col="red")
plot(TR,PIR)
abline(lm(PIR~TR),col="red")
plot(AST,PIR)
abline(lm(PIR~AST),col="red")
plot(STL,PIR)
abline(lm(PIR~STL),col="red")
plot(TO,PIR)
abline(lm(PIR~TO),col="red")
plot(BLK,PIR)
abline(lm(PIR~BLK),col="red")
plot(BLKA,PIR)
abline(lm(PIR~BLKA),col="red")
plot(FC,PIR)
```

```
abline(lm(PIR~FC),col="red")

plot(FD,PIR)

abline(lm(PIR~FD),col="red")


# Chernoff faces

data <- subset(dfdata3, select = -c(Season)) ; data

faces(data, plot.faces = T, main = 'Chernoff faces for every Season')


# Πέντε καλύτεροι παίκτες σε PIR.

excel_data8 <- read_excel("C:\\Users\\alex_\\Desktop\\ΔΙΠΛΩΜΑΤΙΚΗ\\top average 5 players (PIR).xlsx")

dfdata8 <- data.frame(excel_data8)


clean2 <- subset(dfdata8,(Season != 2 & Season != 3 & Season != 4 & Season != 20))

# dataset χωρίς τις σεζόν με τα nan values στα προημιτελικά και στο covid.

attach(clean2)

sum(is.na(clean2))


# Περιγραφικά μέτρα

clean2%>%split(clean2$Playoffs...Quarter.Finals)%>%map(summary)

clean2%>%split(clean2$Final.4...Semi.Finals)%>%map(summary)


# Boxplots

par(mfrow=c(1,2))


b1 <- boxplot(Points~Playoffs...Quarter.Finals,data=clean2,main="Qualified in Playoffs vs Points",  xlab="qualified",ylab="Points",col="red")

b2 <- boxplot(PIR~Playoffs...Quarter.Finals,data=clean2,main="Qualified in Playoffs vs PIR", xlab="qualified",ylab="PIR",col="seagreen")

b3 <- boxplot(Points~Final.4...Semi.Finals,data=clean2,main="Qualified in Final Four vs Points", xlab="qualified",ylab="Points",col="brown")
```

```
b4 <- boxplot(PIR~Final.4...Semi.Finals,data=clean2,main="Qualified in Final Four vs PIR",
xlab="qualified",ylab="PIR",col="blue")


# Outlier Detection

out <- boxplot.stats(clean2$Points)$out

out_ind <- which(clean2$Points %in% c(out))

out_ind

clean2[out_ind, ]


out <- boxplot.stats(clean2$PIR)$out

out_ind <- which(clean2$PIR %in% c(out))

out_ind

clean2[out_ind, ]


# 3D Scatter Plots

attach(clean2)

par(mfrow=c(1,2))


cc1 <- subset(clean2,select=c('Points','PIR','Playoffs...Quarter.Finals'))

colors <- c("navy", "orange")

colors <- colors[as.factor(Playoffs...Quarter.Finals)]

scatterplot3d(cc1, pch = 16, color=colors, box=T, main="3D Scatter Plot for Playoffs", xlab =
"Points", ylab = "PIR", zlab = "Qualified")

legend("bottom", legend = c('Not Qualified','Qualified'), col =  c("navy", "orange"), pch = 16,
inset = -0.25, xpd = TRUE, horiz = TRUE)


cc2 <- subset(clean2,select=c('Points','PIR','Final.4...Semi.Finals'))

colors <- c("darkred", "forestgreen")

colors <- colors[as.factor(Final.4...Semi.Finals)]

scatterplot3d(cc2, pch = 16, color=colors, box=T, main="3D Scatter Plot for Final Four", xlab
= "Points", ylab = "PIR", zlab = "Qualified")
```

```
legend("bottom", legend = c('Not Qualified','Qualified'), col =  c("darkred", "forestgreen"), pch = 16,  inset = -0.25, xpd = TRUE, horiz = TRUE)

# Histograms

k1 <- subset(clean2, Playoffs...Quarter.Finals == 0 )

k2 <- subset(clean2, Playoffs...Quarter.Finals == 1 )


k3 <- subset(clean2, Final.4...Semi.Finals == 0 )

k4 <- subset(clean2, Final.4...Semi.Finals == 1 )


par(mfrow=c(1,2))

l1 <- list(k1$Points,k2$Points)

breaks1 <- pretty(unlist(l1))

levs1 <- levels(cut(unlist(l1), breaks=breaks1))

multhist(l1, ylim=c(0,0.40), xlab = "Points", ylab="Density", col = c("green", "pink2"), freq=F, breaks=breaks1, names.arg = levs1)

multhist(l1, ylim=c(0,0.40), xlab = "Points", ylab="Density", col = c("green", "pink2"), freq=F, breaks=seq(6.5,16,by=1))


l2 <- list(k3$Points,k4$Points)

breaks2 <- pretty(unlist(l2))

levs2 <- levels(cut(unlist(l2), breaks=breaks2))

multhist(l2, ylim=c(0,0.40), xlab = "Points", ylab="Density", col = c("purple", "yellow"), freq=F, breaks=breaks2, names.arg = levs2)

multhist(l2, ylim=c(0,0.40), xlab = "Points", ylab="Density", col = c("purple", "yellow"), freq=F, breaks=seq(6.5,16,by=1))


# Density charts

qualified_in_Playoffs <- factor(Playoffs...Quarter.Finals)

qualified_in_FinalFour <- factor(Final.4...Semi.Finals)


p1 <- ggplot(data=clean2, aes(x=Points, group=Playoffs...Quarter.Finals, fill=qualified_in_Playoffs)) + geom_density(adjust=1.5, alpha=.4) + theme_ipsum()

p1 + scale_fill_brewer(palette = 'Set1')
```

```
p2 <- ggplot(data=clean2, aes(x=Points, group=Final.4...Semi.Finals, fill=qualified_in_FinalFour)) + geom_density(adjust=1.5, alpha=.4) + theme_ipsum()

p2 + scale_fill_manual(values = c('orange', 'blue'))


p3 <- ggplot(data=clean2, aes(x=PIR, group=Playoffs...Quarter.Finals, fill=qualified_in_Playoffs)) + geom_density(adjust=1.5, alpha=.4) + theme_ipsum()

p3 + scale_fill_manual(values = c('green', 'red'))


p4 <- ggplot(data=clean2, aes(x=PIR, group=Final.4...Semi.Finals, fill=qualified_in_FinalFour)) + geom_density(adjust=1.5, alpha=.4) + theme_ipsum()

p4 + scale_fill_manual(values = c('purple','yellow'))
```

 # Δέκα κορυφαίοι παίκτες σε πόντους και PIR ανά σεζόν.

```
excel_data6 <- read_excel("C:\\Users\\alex_\\Desktop\\ΔΙΠΛΩΜΑΤΙΚΗ\\top 10 in Points for each season.xlsx")

dfdata6 <- data.frame(excel_data6)


excel_data7 <- read_excel("C:\\Users\\alex_\\Desktop\\ΔΙΠΛΩΜΑΤΙΚΗ\\top 10 in PIR for each season.xlsx")

dfdata7 <- data.frame(excel_data7)


clean4 <- subset(dfdata6,(Season != 2 & Season != 3 & Season != 4 & Season != 20))

# Points dataset χωρίς τις σεζόν με τα nan values στα προημιτελικά και στο covid


clean5 <- subset(dfdata7,(Season != 2 & Season != 3 & Season != 4 & Season != 20))

# PIR dataset χωρίς τις σεζόν με τα nan values στα προημιτελικά και στο covid


sum(is.na(clean4)) ; sum(is.na(clean5))


clean4%>%split(clean4$Playoffs...Quarter.Finals)%>%map(summary)

clean4%>%split(clean4$Final.4...Semi.Finals)%>%map(summary)
```

```
clean5%>%split(clean5$Playoffs...Quarter.Finals)%>%map(summary)

clean5%>%split(clean5$Final.4...Semi.Finals)%>%map(summary)


# Violin Plots

qualified_in_Playoffs <- factor(clean4$Playoffs...Quarter.Finals)


k1 <- ggplot(clean4, aes(x=qualified_in_Playoffs, y=Points, fill=qualified_in_Playoffs)) + geom_violin(trim=FALSE) + geom_boxplot(width=0.1, fill = 'white')

k1 + scale_fill_brewer(palette="Set1")


k2 <- ggplot(clean4, aes(x=qualified_in_Playoffs, y=PIR, fill=qualified_in_Playoffs)) + geom_violin(trim=FALSE) + geom_boxplot(width=0.1, fill = 'white')

k2  + scale_fill_brewer(palette="Set3")


qualified_in_FinalFour <- factor(clean4$Final.4...Semi.Finals)


k3 <- ggplot(clean4, aes(x=qualified_in_FinalFour, y=Points, fill=qualified_in_FinalFour)) + geom_violin(trim=FALSE) + geom_boxplot(width=0.1, fill = 'white')

k3 + scale_fill_brewer(palette="Set2")


k4 <- ggplot(clean4, aes(x=qualified_in_FinalFour, y=PIR, fill=qualified_in_FinalFour)) + geom_violin(trim=FALSE) + geom_boxplot(width=0.1, fill = 'white')

k4 + scale_fill_brewer(palette="Accent")


qualified_in_Playoffs <- factor(clean5$Playoffs...Quarter.Finals)


k5 <- ggplot(clean5, aes(x=qualified_in_Playoffs, y=Points, fill=qualified_in_Playoffs)) + geom_violin(trim=FALSE) + geom_boxplot(width=0.1, fill = 'white')

k5 + scale_fill_brewer(palette="Dark2")


k6 <- ggplot(clean4, aes(x=qualified_in_Playoffs, y=PIR, fill=qualified_in_Playoffs)) + geom_violin(trim=FALSE) + geom_boxplot(width=0.1, fill = 'white')

k6  + scale_fill_brewer(palette="BuGn")
```

```
qualified_in_FinalFour <- factor(clean5$Final.4...Semi.Finals)


k7 <- ggplot(clean5, aes(x=qualified_in_FinalFour, y=Points, fill=qualified_in_FinalFour)) +
geom_violin(trim=FALSE) + geom_boxplot(width=0.1, fill = 'white')

k7 + scale_fill_brewer(palette="Spectral")


k8 <- ggplot(clean5, aes(x=qualified_in_FinalFour, y=PIR, fill=qualified_in_FinalFour)) +
geom_violin(trim=FALSE) + geom_boxplot(width=0.1, fill = 'white')

k8 + scale_fill_brewer(palette="Pastel1")
```

 # Outlier Detection

```
out <- boxplot.stats(clean4$Points)$out

out_ind <- which(clean4$Points %in% c(out))

out_ind

clean4[out_ind, ]


out <- boxplot.stats(clean4$PIR)$out

out_ind <- which(clean4$PIR %in% c(out))

out_ind

clean4[out_ind, ]


out <- boxplot.stats(clean5$Points)$out

out_ind <- which(clean5$Points %in% c(out))

out_ind

clean5[out_ind, ]


out <- boxplot.stats(clean5$PIR)$out

out_ind <- which(clean5$PIR %in% c(out))

out_ind
```

```
clean5[out_ind, ]


# Πενήντα καλύτεροι σε PIR διαχρονικά.

excel_data9 <- read_excel("C:\\Users\\alex_\\Desktop\\ΔΙΠΛΩΜΑΤΙΚΗ\\Δεδομένα Επίσημα\\each team's top 5 players (2000-2022) - PIR.xlsx")

dfdata9 <- data.frame(excel_data9)


clean6 <- subset(dfdata9,(Season != 2 & Season != 3 & Season != 4 & Season != 20))

# dataset χωρίς τις σεζόν με τα nan values στα προημιτελικά και στο covid

clean6 <- clean6[1:50,] ; clean6

sum(is.na(clean6))


clean6%>%split(clean6$Playoffs...Quarter.Finals)%>%map(summary)

clean6%>%split(clean6$Final.4...Semi.Finals)%>%map(summary)


# Donut Charts

data1 <- data.frame(category=c("Qualified in Playoffs", "Not Qualified"), count=c(25, 25))

data1$fraction <- data1$count / sum(data1$count)

data1$ymax <- cumsum(data1$fraction)

data1$ymin <- c(0, head(data1$ymax, n=-1))

data1$labelPosition <- (data1$ymax + data1$ymin) / 2

data1$label <- paste0(data1$category, "\n value: ", data1$count)


ggplot(data1, aes(ymax=ymax, ymin=ymin, xmax=4, xmin=3, fill=category)) + geom_rect() + geom_label( x=3.5, aes(y=labelPosition, label=label), size=6) + scale_fill_brewer(palette='Accent') + coord_polar(theta="y") + xlim(c(2, 4)) + theme_void() + theme(legend.position = "none")


data2 <- data.frame(category=c("Qualified in Final Four", "Not Qualified"), count=c(38, 12))

data2$fraction <- data2$count / sum(data2$count)

data2$ymax <- cumsum(data2$fraction)

data2$ymin <- c(0, head(data2$ymax, n=-1))
```

```
data2$labelPosition <- (data2$ymax + data2$ymin) / 2

data2$label <- paste0(data2$category, "\n value: ", data2$count)

ggplot(data2, aes(ymax=ymax, ymin=ymin, xmax=4, xmin=3, fill=category)) + geom_rect() + geom_label( x=3.5, aes(y=labelPosition, label=label), size=6) + scale_fill_brewer(palette='Dark2') + coord_polar(theta="y") + xlim(c(2, 4)) + theme_void() + theme(legend.position = "none")
```

 # Γράφημα για παίκτες και ομάδες που προκρίθηκαν (ή όχι) στις φάσεις.

```
attach(clean6)

qualified_in_Final4 <- as.factor(clean6$Final.4...Semi.Finals)

qualified_in_QuarterFinals <- as.factor(clean6$Playoffs...Quarter.Finals)


q1 <- qplot(PIR,Player,data=clean6,color=qualified_in_QuarterFinals)

q1

q2 <- qplot(PIR,Player,data=clean6,color=qualified_in_Final4)

q2
```

ΚΕΦΑΛΑΙΟ 4ο

 # Έλεγχος Κανονικότητας (Lilliefors test)

```
excel_data3 <- read_excel("C:\\Users\\alex_\\Desktop\\ΔΙΠΛΩΜΑΤΙΚΗ\\means.xlsx")

# μέσοι όροι των στατιστικών στοιχείων ανά σεζόν για ΟΛΕΣ τις ομάδες.

dfdata3 <- data.frame(excel_data3)

attach(dfdata3)


lillie.test(Points)

lillie.test(X2PT..)

lillie.test(X3PT..)

lillie.test(FT..)

lillie.test(OR)

lillie.test(DR)

lillie.test(TR)
```

```
lillie.test(AST)

lillie.test(STL)

lillie.test(TO)

lillie.test(BLK)

lillie.test(BLKA)

lillie.test(FC)

lillie.test(FD)

lillie.test(PIR)

lillie.test(Season)
```

 # Πέντε καλύτεροι σε PIR.

```
excel_data8 <- read_excel("C:\\Users\\alex_\\Desktop\\ΔΙΠΛΩΜΑΤΙΚΗ\\top average 5 players (PIR).xlsx")

dfdata8 <- data.frame(excel_data8)

clean2 <- subset(dfdata8,(Season != 2 & Season != 3 & Season != 4 & Season != 20))

# dataset χωρίς τις σεζόν με τα nan values στα προημιτελικά και στο covid

attach(clean2)


k1 <- subset(clean2, Playoffs...Quarter.Finals == 0 )

k2 <- subset(clean2, Playoffs...Quarter.Finals == 1 )

k3 <- subset(clean2, Final.4...Semi.Finals == 0 )

k4 <- subset(clean2, Final.4...Semi.Finals == 1 )


lillie.test(k1$Points)

lillie.test(k1$PIR)

lillie.test(k2$Points)

lillie.test(k2$PIR)

lillie.test(k3$Points)

lillie.test(k3$PIR)

lillie.test(k4$Points)

lillie.test(k4$PIR)
```

 # Δέκα κορυφαίοι παίκτες σε πόντους και PIR ανά σεζόν.

```
excel_data6 <- read_excel("C:\\Users\\alex_\\Desktop\\ΔΙΠΛΩΜΑΤΙΚΗ\\top 10 in Points for each season.xlsx")

dfdata6 <- data.frame(excel_data6)


excel_data7 <- read_excel("C:\\Users\\alex_\\Desktop\\ΔΙΠΛΩΜΑΤΙΚΗ\\top 10 in PIR for each season.xlsx")

dfdata7 <- data.frame(excel_data7)


clean4 <- subset(dfdata6,(Season != 2 & Season != 3 & Season != 4 & Season != 20))

# Points dataset χωρίς τις σεζόν με τα nan values στα προημιτελικά και στο covid


clean5 <- subset(dfdata7,(Season != 2 & Season != 3 & Season != 4 & Season != 20))

# PIR dataset χωρίς τις σεζόν με τα nan values στα προημιτελικά και στο covid


attach(clean4) ; attach(clean5)


k5 <- subset(clean4, Playoffs...Quarter.Finals == 0 )

k6 <- subset(clean4, Playoffs...Quarter.Finals == 1 )

k7 <- subset(clean4, Final.4...Semi.Finals == 0 )

k8 <- subset(clean4, Final.4...Semi.Finals == 1 )

k9 <- subset(clean5, Playoffs...Quarter.Finals == 0 )

k10 <- subset(clean5, Playoffs...Quarter.Finals == 1 )

k11 <- subset(clean5, Final.4...Semi.Finals == 0 )

k12 <- subset(clean5, Final.4...Semi.Finals == 1 )


lillie.test(k5$Points)

lillie.test(k5$PIR)

lillie.test(k6$Points)
```

```
lillie.test(k6$PIR)

lillie.test(k7$Points)

lillie.test(k7$PIR)

lillie.test(k8$Points)

lillie.test(k8$PIR)

lillie.test(k9$Points)

lillie.test(k9$PIR)

lillie.test(k10$Points)

lillie.test(k10$PIR)

lillie.test(k11$Points)

lillie.test(k11$PIR)

lillie.test(k12$Points)

lillie.test(k12$PIR)
```

 # Πενήντα καλύτεροι σε PIR διαχρονικά

```
excel_data9 <- read_excel("C:\\Users\\alex_\\Desktop\\ΔΙΠΛΩΜΑΤΙΚΗ\\Δεδομένα Επίσημα\\each team's top 5 players (2000-2022) - PIR.xlsx")

dfdata9 <- data.frame(excel_data9)

clean6 <- subset(dfdata9,(Season != 2 & Season != 3 & Season != 4 & Season != 20))

# dataset χωρίς τις σεζόν με τα nan values στα προημιτελικά και στο covid

clean6 <- clean6[1:50,]

attach(clean6)


k13 <- subset(clean6, Playoffs...Quarter.Finals == 0 )

k14 <- subset(clean6, Playoffs...Quarter.Finals == 1 )

k15 <- subset(clean6, Final.4...Semi.Finals == 0 )

k16 <- subset(clean6, Final.4...Semi.Finals == 1 )

lillie.test(k13$Points)

lillie.test(k13$PIR)

lillie.test(k14$Points)

lillie.test(k14$PIR)
```

```
lillie.test(k15$Points)

lillie.test(k15$PIR)

lillie.test(k16$Points)

lillie.test(k16$PIR)


# Συντελεστές Συσχέτισης

# Heatmaps και Κορελογράμματα


# δεδομένα με τις κανονικές μεταβλητές

data_norm <- subset(dfdata3, select = -c(Season,X3PT..,STL))

round(cor(data_norm, method='pearson'), 2)


source("http://www.sthda.com/upload/rquery_cormat.r")

require("corrplot")

rquery.cormat(data_norm)

cormat<-rquery.cormat(data_norm, graphType="heatmap")

rquery.cormat(data_norm, type="flatten", graph=TRUE)


ggcorr(data_norm, method = c("everything" , "pearson"), label=T, label_size = 3, label_round
= 2)

round(cor(data_norm,Season, method='spearman'), 2)


# δεδομένα με τις μη κανονικές μεταβλητές

data_not_norm <- subset(dfdata3, select = c(X3PT..,STL)) ; data_not_norm

round(cor(data_not_norm,method = 'spearman'), 2)


mm <- data.frame(data_not_norm,Season,PIR)

round(cor(mm,method = 'spearman'), 2)

ggcorr(mm, method = c("everything" , "spearman"), label=T, label_size = 3, label_round = 2)


round(cor(data_not_norm,data_norm,method = 'spearman'), 2)
```

 # Πέντε καλύτεροι σε PIR.

```
k1 <- subset(clean2, Playoffs...Quarter.Finals == 0 )
k2 <- subset(clean2, Playoffs...Quarter.Finals == 1 )
k3 <- subset(clean2, Final.4...Semi.Finals == 0 )
k4 <- subset(clean2, Final.4...Semi.Finals == 1 )

round(cor(k1$Points, k1$PIR, method = 'spearman'), 2)
round(cor(k2$Points, k2$PIR, method = 'pearson'), 2)
round(cor(k3$Points, k3$PIR, method = 'spearman'), 2)
round(cor(k4$Points, k4$PIR, method = 'pearson'), 2)
```

 # Δέκα κορυφαίοι παίκτες σε πόντους και PIR ανά σεζόν.

```
k5 <- subset(clean4, Playoffs...Quarter.Finals == 0 )
k6 <- subset(clean4, Playoffs...Quarter.Finals == 1 )
k7 <- subset(clean4, Final.4...Semi.Finals == 0 )
k8 <- subset(clean4, Final.4...Semi.Finals == 1 )
k9 <- subset(clean5, Playoffs...Quarter.Finals == 0 )
k10 <- subset(clean5, Playoffs...Quarter.Finals == 1 )
k11 <- subset(clean5, Final.4...Semi.Finals == 0 )
k12 <- subset(clean5, Final.4...Semi.Finals == 1 )

round(cor(k5$Points, k5$PIR, method = 'spearman'), 2)
round(cor(k6$Points, k6$PIR, method = 'spearman'), 2)
round(cor(k7$Points, k7$PIR, method = 'spearman'), 2)
round(cor(k8$Points, k8$PIR, method = 'pearson'), 2)
round(cor(k9$Points, k9$PIR, method = 'spearman'), 2)
round(cor(k10$Points, k10$PIR, method = 'spearman'), 2)
round(cor(k11$Points, k11$PIR, method = 'spearman'), 2)
round(cor(k12$Points, k12$PIR, method = 'spearman'), 2)
```

```
# Πενήντα καλύτεροι σε PIR διαχρονικά.
k13 <- subset(clean6, Playoffs...Quarter.Finals == 0 )
k14 <- subset(clean6, Playoffs...Quarter.Finals == 1 )
k15 <- subset(clean6, Final.4...Semi.Finals == 0 )
k16 <- subset(clean6, Final.4...Semi.Finals == 1 )

round(cor(k13$Points, k13$PIR, method = 'spearman'), 2)
round(cor(k14$Points, k14$PIR, method = 'spearman'), 2)
round(cor(k15$Points, k15$PIR, method = 'spearman'), 2)
round(cor(k16$Points, k16$PIR, method = 'spearman'), 2)

# Έλεγχος για ισότητα μέσων τιμών (t-tests)
# Πέντε καλύτεροι σε PIR.
k1 <- subset(clean2, Playoffs...Quarter.Finals == 0 )
k2 <- subset(clean2, Playoffs...Quarter.Finals == 1 )
k3 <- subset(clean2, Final.4...Semi.Finals == 0 )
k4 <- subset(clean2, Final.4...Semi.Finals == 1 )

t.test(k1$Points, k2$Points, paired = F)
t.test(k1$PIR, k2$PIR, paired = F)
t.test(k3$Points, k4$Points, paired = F)
t.test(k3$PIR, k4$PIR, paired = F)

# Δέκα κορυφαίοι παίκτες σε πόντους και PIR ανά σεζόν.
k5 <- subset(clean4, Playoffs...Quarter.Finals == 0 )
k6 <- subset(clean4, Playoffs...Quarter.Finals == 1 )
k7 <- subset(clean4, Final.4...Semi.Finals == 0 )
k8 <- subset(clean4, Final.4...Semi.Finals == 1 )
k9 <- subset(clean5, Playoffs...Quarter.Finals == 0 )
k10 <- subset(clean5, Playoffs...Quarter.Finals == 1 )
k11 <- subset(clean5, Final.4...Semi.Finals == 0 )
```

```
k12 <- subset(clean5, Final.4...Semi.Finals == 1 )


t.test(k5$Points, k6$Points, paired = F)

t.test(k5$PIR, k6$PIR, paired = F)

t.test(k7$Points, k8$Points, paired = F)

t.test(k7$PIR, k8$PIR, paired = F)

t.test(k9$Points, k10$Points, paired = F)

t.test(k9$PIR, k10$PIR, paired = F)

t.test(k11$Points, k12$Points, paired = F)

t.test(k11$PIR, k12$PIR, paired = F)
```

 # Πενήντα καλύτεροι σε PIR διαχρονικά.

```
k13 <- subset(clean6, Playoffs...Quarter.Finals == 0 )

k14 <- subset(clean6, Playoffs...Quarter.Finals == 1 )

k15 <- subset(clean6, Final.4...Semi.Finals == 0 )

k16 <- subset(clean6, Final.4...Semi.Finals == 1 )


t.test(k13$Points, k14$Points, paired = F)

t.test(k13$PIR, k14$PIR, paired = F)

t.test(k15$Points, k16$Points, paired = F)

t.test(k15$PIR, k16$PIR, paired = F)
```

ΚΕΦΑΛΑΙΟ 5ο

```
excel_data <- read_excel("C:\\Users\\alex_\\Desktop\\ΔΙΠΛΩΜΑΤΙΚΗ\\TEAMS EVERY YEAR.xlsx")

dfdata <- data.frame(excel_data)

sum(is.na(dfdata))


clean1 <- subset(dfdata,(Season != 2 & Season != 3 & Season != 4 & Season != 20))

# dataset χωρίς τις σεζόν με τα nan values στα προημιτελικά και στο covid

attach(clean1)
```

```
sum(is.na(clean1))


data <- clean1[c(6:20,22:23)]

##################### Ανάλυση για Playoffs ############################

Playoffs...Quarter.Finals <- as.factor(Playoffs...Quarter.Finals)

model_start1 <- glm(Playoffs...Quarter.Finals ~ 1, family=binomial(link=logit))


model1<-glm(Playoffs...Quarter.Finals ~ Points+X2PT..+X3PT..+FT..+OR+DR+TR+AST+STL+TO+BLK+BLKA+FC+FD+PIR+PIR_INDIV+PIR_BEST_FIVE, family=binomial(link=logit))

summary(model1)

anova(model1,test="Chisq")


# STEP μέθοδος χωρίς αλληλεπιδράσεις 2ου βαθμού

stepMod1 <- step(model_start1, scope = list(lower = model_start1, upper = model1),direction = "both")


model2<-glm(Playoffs...Quarter.Finals ~ (PIR+AST+STL+FT..+FC+Points+TO+X2PT..), family=binomial(link=logit))

summary(model2)

anova(model2,test="Chisq")



# Έλεγχος για πολυσυγγραμικότητα

round(vif(model2), 3)


## αφαιρώ την μεταβλητή Points ##

model1_new <- glm(Playoffs...Quarter.Finals ~ X2PT..+X3PT..+FT..+OR+DR+TR+AST+STL+TO+BLK+BLKA+FC+FD+PIR+PIR_INDIV+PIR_BEST_FIVE, family=binomial(link=logit))

summary(model1_new)

anova(model1_new,test="Chisq")
```

```
stepMod1_new <- step(model_start1, scope = list(lower = model_start1, upper = model1_new),direction = "both")


model2_new <- glm(Playoffs...Quarter.Finals ~ (AST+TO+DR+X2PT..+X3PT..+FD+BLK+OR), family=binomial(link=logit))

summary(model2_new)

anova(model2_new,test="Chisq")

round(vif(model2_new), 3)


stepMod2_new <- step(model_start1, scope = list(lower = model_start1, upper = model2_new),direction = "both")


model3 <- glm(Playoffs...Quarter.Finals ~ (X2PT.. + DR + TO + FD + X3PT.. + AST), family=binomial(link=logit))

summary(model3)

anova(model3,test="Chisq")

round(vif(model3), 3)


# Μέτρα Προσαρμογής

round(PseudoR2(model3, which = 'all'), 3)


# Έλεγχος Hosmer-Lemeshow

y <- Playoffs...Quarter.Finals

hl.test <- hoslem.test(model3$y, fitted(model3))

hl.test


# Deviance Residuals

res1 <- resid(model3, type = 'deviance')

fit1 <- fitted(model3)

plot(fit1, res1, xlab='Fitted Values', ylab='Residuals')


# ROC καμπύλη
```

```
pred1=predict(model3,type="response")

plot.roc(clean1$Playoffs...Quarter.Finals,pred1)

auc(clean1$Playoffs...Quarter.Finals,pred1)

coords(roc(clean1$Playoffs...Quarter.Finals,pred1),"best",best.method="youden")
```

 # STEP μέθοδος με αλληλεπιδράσεις 2ου βαθμού

```
model4<-glm(Playoffs...Quarter.Finals ~ (X2PT.. + DR + TO + FD + X3PT.. + AST)^2, family=binomial(link=logit))

summary(model4)

anova(model4,test="Chisq")


stepMod2 <- step(model_start1, scope = list(lower = model_start1, upper = model4),direction = "both")


model_inter1 <- glm(Playoffs...Quarter.Finals ~ X2PT.. + DR + TO + FD + X3PT.. + AST, family=binomial(link=logit))

summary(model_inter1)

anova(model_inter1,test="Chisq")

round(vif(model_inter1), 3)


round(PseudoR2(model_inter1, which = 'all'), 3)


y <- Playoffs...Quarter.Finals

hl.test <- hoslem.test(model_inter1$y, fitted(model_inter1))

hl.test

res2 <- resid(model_inter1, type = 'deviance')

fit2 <- fitted(model_inter1)

plot(fit2, res2, xlab='Fitted Values', ylab='Residuals')


pred2=predict(model_inter1,type="response")

plot.roc(clean1$Playoffs...Quarter.Finals,pred2)

auc(clean1$Playoffs...Quarter.Finals,pred2)
```

```
coords(roc(clean1$Playoffs...Quarter.Finals,pred2),"best",best.method="youden")



##################### Ανάλυση για Final Four ############################

Final.4...Semi.Finals<- as.factor(Final.4...Semi.Finals)


model_start2 <- glm(Final.4...Semi.Finals ~ 1, family=binomial(link=logit))


model5<-glm(Final.4...Semi.Finals ~
Points+X2PT..+X3PT..+FT..+OR+DR+TR+AST+STL+TO+BLK+BLKA+FC+FD+PIR+PIR
_INDIV+PIR_BEST_FIVE, family=binomial(link=logit))

summary(model5)

anova(model5,test="Chisq")


stepMod3 <- step(model_start2, scope = list(lower = model_start2, upper = model5),direction
= "both")


# STEP μέθοδος χωρίς αλληλεπιδράσεις

model6 <-glm(Final.4...Semi.Finals ~
(PIR+Points+FC+AST+X3PT..+BLKA+X2PT..+TO+OR), family=binomial(link=logit))

summary(model6)

anova(model6,test="Chisq")

round(vif(model6), 3)


## αφαιρώ τη μεταβλητή Points ##

model_start2 <- glm(Final.4...Semi.Finals ~ 1, family=binomial(link=logit))


model4_new<-glm(Final.4...Semi.Finals ~
X2PT..+X3PT..+FT..+OR+DR+TR+AST+STL+TO+BLK+BLKA+FC+FD+PIR+PIR_INDI
V+PIR_BEST_FIVE, family=binomial(link=logit))

summary(model4_new)

anova(model4_new,test="Chisq")
```

```
stepMod3 <- step(model_start2, scope = list(lower = model_start2, upper = model4_new),direction = "both")


model7 <-glm(Final.4...Semi.Finals ~ (PIR + FD + PIR_BEST_FIVE + TO + X2PT.. + DR + X3PT..), family=binomial(link=logit))

summary(model7)

anova(model7,test="Chisq")


model8 <-glm(Final.4...Semi.Finals ~ (FD + TO + X2PT.. + DR + X3PT..), family=binomial(link=logit))

summary(model8)

anova(model8,test="Chisq")

round(vif(model8), 3)


round(PseudoR2(model8, which = 'all'), 3)


y <- Final.4...Semi.Finals

hl.test <- hoslem.test(model8$y, fitted(model8))

hl.test


res3 <- resid(model8, type = 'deviance')

fit3 <- fitted(model8)

plot(fit3, res3, xlab='Fitted Values', ylab='Residuals')


pred3=predict(model8,type="response")

plot.roc(clean1$Final.4...Semi.Finals,pred3)

auc(clean1$Final.4...Semi.Finals,pred3)

coords(roc(clean1$Final.4...Semi.Finals,pred3),"best",best.method="youden")


# STEP μέθοδος με αλληλεπιδράσεις 2ου βαθμού

model4_new<-glm(Final.4...Semi.Finals ~ (FD + TO + X2PT.. +  DR + X3PT..)^2, family=binomial(link=logit))
```

```
summary(model4_new)

anova(model4_new,test="Chisq")


stepMod4 <- step(model_start2, scope = list(lower = model_start2, upper = model4_new),direction = "both")


model_inter2 <- glm(Final.4...Semi.Finals ~  X2PT.. + X3PT.. + DR + FD + TO + X3PT..:FD + X2PT..:X3PT..  , family=binomial(link=logit))

summary(model_inter2)

anova(model_inter2,test="Chisq")

round(vif(model_inter2), 3)


# «Κεντροποίηση» των δεδομένων

num_vars <- sapply(data, is.numeric)

data_centered <- lapply(data[, num_vars], function(x) x - median(x))

data_centered


model_start2 <- glm(Final.4...Semi.Finals ~ 1, family=binomial(link=logit),data=data_centered)


model4_new<-glm(Final.4...Semi.Finals ~ (FD + TO + X2PT.. + DR + X3PT..)^2, family=binomial(link=logit), data=data_centered)

summary(model4_new)

anova(model4_new,test="Chisq")


stepMod4 <- step(model_start2, scope = list(lower = model_start2, upper = model4_new),direction = "both")

model_inter3 <- glm(Final.4...Semi.Finals ~ X2PT.. + X3PT.. + DR + FD + TO + X3PT..:FD + X2PT..:X3PT..  , family=binomial(link=logit), data=data_centered)


summary(model_inter3)

anova(model_inter3,test="Chisq")

round(vif(model_inter3), 3)
```

```
round(PseudoR2(model_inter3, which = 'all'), 3)


y <- Playoffs...Quarter.Finals

hl.test <- hoslem.test(model_inter3$y, fitted(model_inter3))

hl.test


res4 <- resid(model_inter3, type = 'deviance')

fit4 <- fitted(model_inter3)

plot(fit4, res4, xlab='Fitted Values', ylab='Residuals')


pred4=predict(model_inter3,type="response")

plot.roc(clean1$Final.4...Semi.Finals,pred4)

auc(clean1$Final.4...Semi.Finals,pred4)

coords(roc(clean1$Final.4...Semi.Finals,pred4),"best",best.method="youden")
```

ΚΕΦΑΛΑΙΟ 6ο

Σε αυτό το κεφάλαιο, οι εντολές που χρησιμοποιήσαμε είναι οι ίδιες για τις δύο φάσεις της διοργάνωσης, αλλάζοντας κάθε φορά την αντίστοιχη μεταβλητή απόκρισης. Συνεπώς, θα παραθέσουμε **μόνο** τους κώδικες σχετικά με τη φάση των playoffs, για τις μεθόδους clustering και classification που εφαρμόσαμε.

```
excel_data <- read_excel("C:\\Users\\alex_\\Desktop\\ΔΙΠΛΩΜΑΤΙΚΗ\\TEAMS EVERY YEAR.xlsx")

dfdata <- data.frame(excel_data)

sum(is.na(dfdata))


clean1 <- subset(dfdata,(Season != 2 & Season != 3 & Season != 4 & Season != 20))

# dataset χωρίς τις σεζόν με τα nan values στα προημιτελικά και στο covid

attach(clean1)

sum(is.na(clean1))

data <- clean1[c(6:20,22:23)]
```

```
# Feature Selection (Boruta Method)

features <- data

labels1 <- Playoffs...Quarter.Finals

set.seed(18)

boruta_res1 <- Boruta(labels1~.,data=na.omit(features),doTrace=2)

selected_features1 <- getSelectedAttributes(boruta_res1, withTentative = F)

features_selected1 <- features[,selected_features1]

features_normalized1 <- scale(features_selected1)

plot(boruta_res1, cex.axis=.7, las=2, xlab="", main="Variable Importance")


# Elbow Plot

fviz_nbclust(features_normalized1, kmeans, method = "wss") + geom_vline(xintercept = 2, linetype = 2) + labs(subtitle = "Elbow Method")


# Silhouette Plot

fviz_nbclust(features_normalized1, kmeans, method = "silhouette") + labs(subtitle = "Average Silhouette Method")


# K-Means Clustering

set.seed(18)

kmeans_res1 <- kmeans(features_normalized1, centers = 2)

cluster_assignments1 <- kmeans_res1$cluster


# Cluster plot

cluster_plot <- fviz_cluster(list(data = features_normalized1, cluster = cluster_assignments1), geom = 'point', palette = "jco", repel = TRUE, show.clust.cent = FALSE)

cluster_plot <- cluster_plot + geom_text_repel(aes(label = labels1), color = "black", size = 4)


# Stacked bar plot

df <- data.frame(labels=rep(c('0', '1'), each=2), clusters=rep(c('1', '2'), times=2), teams=c(161, 84, 28, 123))
```

```
ggplot(df, aes(fill = clusters, y = teams, x = labels)) + geom_bar(position = 'stack', stat = 'identity') + geom_text(aes(label = teams), position = position_stack(vjust = 0.5), color = "black", fontface = "bold", size = 4) + theme_minimal() +  labs(x = 'Labels', y = 'Teams', title = 'Distribution of Teams in Clusters') + theme(plot.title = element_text(hjust = 0.5, size = 20, face = 'bold') ) + scale_fill_manual(name = 'Cluster', values = c('blue1', 'yellow2'))

# Μέτρα Αξιολόγησης

silhouette_avg1 <- silhouette(kmeans_res1$cluster, dist(features_normalized1))

dunn_index1 <- dunn(dist(features_normalized1), kmeans_res1$cluster)

adj_rand_index1 <- adj.rand.index(kmeans_res1$cluster, labels1)


# Hierarchical Agglomerative Clustering

hc1 <- agnes(features_normalized1, method = "ward")

pltree(hc1, cex = 0.6, hang = -1, main = "Dendrogram", labels = F)

pred_labels1 <- cutree(hc1, k = 2)


# Cluster plot

cluster_plot <- fviz_cluster(list(data = features_normalized1, cluster = pred_labels1), geom = 'point', palette = "jco", repel = TRUE, show.clust.cent = FALSE)

cluster_plot <- cluster_plot + geom_text_repel(aes(label = labels1), color = "black", size = 4)


# Stacked bar plot

df <- data.frame(labels=rep(c('0', '1'), each=2), clusters=rep(c('1', '2'), times=2), teams=c(123, 122, 141, 10))


ggplot(df, aes(fill = clusters, y = teams, x = labels)) + geom_bar(position = 'stack', stat = 'identity') + geom_text(aes(label = teams), position = position_stack(vjust = 0.5), color = "black", fontface = "bold", size = 4) + theme_minimal() + labs(x = 'Labels', y = 'Teams', title = 'Distribution of Teams in Clusters') + theme(plot.title = element_text(hjust = 0.5, size = 20, face = 'bold')) + scale_fill_manual(name = 'Cluster', values = c('blue1', 'yellow2'))


# Μέτρα Αξιολόγησης

silhouette_avg1 <- silhouette(pred_labels1, dist(features_normalized1))

dunn_index1 <- dunn(dist(features_normalized1), pred_labels1)

adj_rand_index1 <- adj.rand.index(pred_labels1, labels1)
```

 # Splitting the Dataset

```
my_data <- data.frame(features_normalized1,labels1)

data <- data.frame(features_normalized1)

labels1 = factor(labels1)


set.seed(18)

split = sample.split(my_data$labels1, SplitRatio = 0.70)

training_set = subset(my_data, split == TRUE)

test_set = subset(my_data, split == FALSE)
```

 # Support Vector Machines

```
classifier = svm(formula = labels1 ~ ., data = training_set, type = 'C-classification', kernel = 'radial')

y_pred = predict(classifier, newdata = test_set)
```

 # Confusion Matrix

```
cm <- confusionMatrix(data=y_pred, reference=factor(test_set$labels1), dnn = c("Prediction", "Reference"))


plt <- as.data.frame(cm$table)

plt$Prediction <- factor(plt$Prediction, levels=rev(levels(plt$Prediction)))


ggplot(plt, aes(Reference,Prediction, fill= Freq)) + geom_tile() + geom_text(aes(label=Freq)) + scale_fill_gradient(low="white", high="#009194") + labs(x = "Reference",y = "Prediction",title = "Confusion Matrix") + scale_x_discrete(labels=c("0","1")) + scale_y_discrete(labels=c("1","0")) + theme(plot.title = element_text(hjust = 0.5))
```

 # Μέτρα Αξιολόγησης

```
TP <- confusion_matrix[2, 2]

FP <- confusion_matrix[2, 1]
```

```
TN <- confusion_matrix[1, 1]

FN <- confusion_matrix[1, 2]



accuracy <- (TP + TN) / sum(confusion_matrix)

precision <- TP / (TP + FP)

recall <- TP / (TP + FN)

f1_score <- 2 * precision * recall / (precision + recall)


# Random Forest Classifier

set.seed(18)

classifier_RF = randomForest(formula = labels1 ~ . , data = training_set)

y_pred = predict(classifier_RF, newdata = test_set)


# Confusion Matrix

cm <- confusionMatrix(data=y_pred, reference=factor(test_set$labels1), dnn = c("Prediction",
"Reference"))


plt <- as.data.frame(cm$table)

plt$Prediction <- factor(plt$Prediction, levels=rev(levels(plt$Prediction)))


ggplot(plt, aes(Reference,Prediction, fill= Freq)) + geom_tile() + geom_text(aes(label=Freq))
+ scale_fill_gradient(low="white", high="#006600") + labs(x = "Reference",y =
"Prediction",title='Confusion Matrix') + scale_x_discrete(labels=c("0","1")) +
scale_y_discrete(labels=c("1","0")) + theme(plot.title = element_text(hjust = 0.5))


# Μέτρα Αξιολόγησης

TP <- confusion_matrix[2, 2]

FP <- confusion_matrix[2, 1]

TN <- confusion_matrix[1, 1]

FN <- confusion_matrix[1, 2]
```

```
accuracy <- (TP + TN) / sum(confusion_matrix)

precision <- TP / (TP + FP)

recall <- TP / (TP + FN)

f1_score <- 2 * precision * recall / (precision + recall)
```