ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΓΙΑ ΤΟΥΣ ΠΑΡΑΓΟΝΤΕΣ ΠΟΥ ΕΠΗΡΕΑΖΟΥΝ ΤΗΝ ΑΠΟΔΟΣΗ ΤΩΝ ΟΜΑΔΩΝ ΠΟΔΟΣΦΑΙΡΟΥ ΣΤΙΣ ΕΥΡΩΠΑΪΚΕΣ ΔΙΟΡΓΑΝΩΣΕΙΣ

Ανδρέας Ε. Σπυριδάκης

Διπλωματική Εργασία

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

Πειραιάς
Φεβρουάριος 2022

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**

ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΓΙΑ ΤΟΥΣ ΠΑΡΑΓΟΝΤΕΣ ΠΟΥ ΕΠΗΡΕΑΖΟΥΝ ΤΗΝ ΑΠΟΔΟΣΗ ΤΩΝ ΟΜΑΔΩΝ ΠΟΔΟΣΦΑΙΡΟΥ ΣΤΙΣ ΕΥΡΩΠΑΪΚΕΣ ΔΙΟΡΓΑΝΩΣΕΙΣ

Ανδρέας Ε. Σπυριδάκης

Διπλωματική Εργασία

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

Πειραιάς
Φεβρουάριος 2022

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)
- Αναπληρωτής Καθηγητής Γεώργιος Τζαβελάς
- Επίκουρος Καθηγητής Χαράλαμπος Ευαγγελάρας

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

STATISTICAL ANALYSIS FOR THE FACTORS AFFECTING THE PERFORMANCE OF TEAMS IN EUROPEAN FOOTBALL

By

Andreas E. Spyridakis

MSc Dissertation

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

Piraeus, Greece
February 2022

Ευχαριστίες

Με την παρούσα παράγραφο θα ήθελα να ευχαριστήσω όλους τους ανθρώπους που συνέβαλλαν στην εκπόνηση της παρούσας διπλωματικής εργασίας. Πρωτίστως, θα ήθελα να ευχαριστήσω τον Αναπληρωτή Καθηγητή κ. Κωνσταντίνο Πολίτη, ο οποίος ως επιβλέπων της διπλωματικής και παρόλη την καινούργια κατάσταση που διαμόρφωσε η πανδημία, ήταν πάντα πολύ βοηθητικός με τις συμβουλές του, με τις γνώσεις του και με την ώθηση που έδωσε να ασχοληθώ με αυτό που θα ήθελα ιδανικά να συνδυάσω την επιστήμη μου. Η αμεσότητα και η μέριμνα που υπήρξε αποτέλεσε καθοριστική για την ολοκλήρωση της διαδικασίας εκπόνησης της διπλωματικής. Θα ήθελα, επίσης, να ευχαριστήσω την οικογένειά μου, όπως και το στενό μου περιβάλλον για την υλική και ηθική υποστήριξη που έλαβα κατά τη διαδικασία αυτή.

ΠΕΡΙΛΗΨΗ

Τα τελευταία χρόνια, η επιστήμη της στατιστικής έχει γνωρίσει ραγδαία ανάπτυξη. Καθοριστικό ρόλο σε αυτό έχει παίξει η αντίστοιχη ανάπτυξη της επιστήμης των υπολογιστών, η οποία έδωσε τη δυνατότητα ανάλυσης ενός μεγάλου όγκου δεδομένων, αλλά και την διεκπεραίωση εκατομμυρίων υπολογισμών σε ελάχιστο χρόνο. Ο συνδυασμός των δύο παραπάνω επιστημών έχει φέρει στο προσκήνιο τη μηχανική μάθηση, με την ανάπτυξη αλγορίθμων, οι οποίοι μπορούν να διαχειριστούν κάθε τύπο δεδομένων. Με αυτόν τον τρόπο, η ανάλυση των δεδομένων έχει μπει σε πολλούς νέους κλάδους. Ένας από αυτούς είναι και ο αθλητισμός. Παρόλο που τα δεδομένα παλαιότερα αναλύονταν σε μεγαλύτερο βαθμό στην καλαθοσφαίριση, τα τελευταία 5-10 χρόνια έχουν μπει δυναμικά και στο ποδόσφαιρο.

Στην παρούσα μελέτη γίνεται μια προσπάθεια να βρεθούν οι μεταβλητές εκείνες που επηρεάζουν σημαντικότερα το παιχνίδι μιας ποδοσφαιρικής ομάδας, με βάση τα βασικά δεδομένα που μπορεί να δει ο κάθε φίλαθλος. Τα δεδομένα προέρχονται από την κορυφαία ευρωπαϊκή διασυλλογική διοργάνωση. Χρησιμοποιήθηκαν στην αρχή μέθοδοι περιγραφικής στατιστικής ανάλυσης, για την εξαγωγή των πρώτων συμπερασμάτων. Επιπλέον, έγινε υπολογισμός των συσχετίσεων ανάμεσα στις μεταβλητές. Στη συνέχεια, με τη χρήση τριών αλγορίθμων μηχανικής μάθησης, έναν για την επιλογή χαρακτηριστικών και δύο για κατηγοριοποίηση με βάση και τα αποτελέσματα της επιλογής χαρακτηριστικών, βγαίνουν τα πρώτα πολύ βασικά αποτελέσματα. Τέλος, γίνεται χρήση γενικευμένων γραμμικών μοντέλων, έτσι ώστε να ποσοτικοποιηθεί η επιρροή των παραγόντων που κρίθηκαν σημαντικοί για την έκβαση ενός ποδοσφαιρικού αγώνα.

ABSTRACT

In recent years, the science of statistics has grown rapidly. The corresponding development of computer science has played a decisive role in this, which enabled the analysis of a large amount of data, but also the processing of millions of calculations in a minimum amount of time. The combination of the above two disciplines has brought machine learning to the forefront, with the development of algorithms that can handle any type of data. In this way, data analysis has entered many new fields. One of them is sports. Although the data used to be analyzed to a greater extent in basketball, in the last 5-10 years they have entered dynamically in football as well.

In the present study, an attempt is made to find those variables that most significantly affect the performance of a football team, based on the basic data that can be seen by each fan. The data comes from the leading European inter-club organization. Descriptive statistical analysis methods are initially used to draw the first conclusions. In addition, the correlations between the variables were calculated. Then, using three machine learning algorithms, one for feature selection and two for classification based on feature selection results, so that the first very basic results come out. Finally, generalized linear models are used, in order to quantify the influence of the factors that were considered important for the outcome of a football game.

Περιεχόμενα

ΕΙΣΑΓΩΓΗ

Η επιστήμη της Στατιστικής και η πρόοδός της, σε συνδυασμό με την γιγάντια ανάπτυξη που έχει ο κλάδος της Πληροφορικής τα τελευταία χρόνια, έχει φέρει στο προσκήνιο νέες μεθόδους και τεχνικές οι οποίες έχουν δώσει στον άνθρωπο τη δυνατότητα να μπορεί να διαχειριστεί έναν μεγάλο όγκο δεδομένων. Αν συνδυαστεί αυτή η ανάπτυξη με το γεγονός ότι η Στατιστική βρίσκει έδαφος σε όλους τους τομείς της ανθρώπινης δραστηριότητας, γίνεται εύκολα αντιληπτό ότι δημιουργούνται νέα μονοπάτια στην παραπέρα βελτίωσή της. Ένα από αυτά τα πεδία εφαρμογής είναι και ο χώρος του Αθλητισμού, ξεκινώντας από άλλα αθλήματα και κατά κύριο λόγο το μπάσκετ (Καλλιακμάνης, 2020). Πριν από μία δεκαετία κυκλοφόρησε η ταινία με τίτλο Moneyball, βασισμένη στο ομώνυμο βιβλίο του Michael Lewis (Moneyball: The art of winning an unfair game), το οποίο αναδείκνυε την προσπάθεια που έκανε η ομάδα baseball των Oakland Athletics με βάση την αγωνιστική εικόνα παικτών που αφορούσε τα στατιστικά τους, να δημιουργήσει μία ομάδα η οποία θα ανταγωνιζόταν τις μεγάλες και πλούσιες ομάδες του πρωταθλήματος. Παρόλα αυτά, τη μεγαλύτερη εξέλιξη μπορούμε να δούμε στο ποδόσφαιρο.

Το ποδόσφαιρο είναι ένα ομαδικό άθλημα το οποίο έχει προσφέρει αμέτρητες συγκινήσεις σε εκατομμύρια φιλάθλους και αθλητές σε όλο τον κόσμο, κάνοντάς το έτσι μακράν το πιο διαδεδομένο από τα υπόλοιπα, ακόμη και ατομικά αθλήματα. Η εμπορευματοποίησή του, η οποία έχει αποφέρει τεράστια κέρδη για τις επιχειρήσεις οι οποίες δραστηριοποιούνται στο κομμάτι αυτό, συμπεριλαμβάνοντας και τους ίδιους τους συλλόγους, έχει κάνει πολύ κόσμο δίκαια να σκέφτεται ότι καλύτερη ομάδα είναι εκείνη η οποία πληρώνει τα περισσότερα για πανάκριβες μεταγραφές και συμβόλαια ποδοσφαιριστών. Το ίδιο φυσικά συμβαίνει και όταν αναφερόμαστε σε άλλα ομαδικά αθλήματα. Παρόλα αυτά, μπορούμε να δούμε ότι στο ποδόσφαιρο σε μεγαλύτερο βαθμό από τα άλλα αθλήματα συνεχίζονται να γίνονται μεγάλες εκπλήξεις και ομάδες οι οποίες θεωρούνται ασθενέστερες να καταφέρνουν να αποδώσουν παραπάνω από το αναμενόμενο.

Η παραπάνω διαπίστωση προκύπτει λόγω του γεγονότος ότι το ποδόσφαιρο είναι ένα άθλημα το οποίο έχει μεγαλύτερο βαθμό τυχαιότητας από το μπάσκετ ή άλλα αθλήματα, επειδή έχει χαμηλά σκορ. Αυτό σημαίνει ότι το αποτέλεσμα μπορεί να επηρεαστεί από πολλά τυχαία γεγονότα, όπως η τροχιά που παίρνει η μπάλα ή μια λάθος απόφαση του διαιτητή. Αυτό μας δίνει σαν αποτέλεσμα ότι η καλύτερη ομάδα στο ποδόσφαιρο νικάει λιγότερο συχνά. Γι αυτό κοιτάμε την ομάδα με βάση συγκεκριμένους δείκτες και όχι μόνο το αποτέλεσμα (Rasmus Ankersen, TEDx Manchester, 2018).

Η εφαρμογή μεθόδων στατιστικής ανάλυσης στον τομέα του ποδοσφαίρου ουσιαστικά ξεκίνησε από τα τέλη της δεκαετίας του 1990. Η αρχή έγινε με την εφαρμογή μεθόδων παλινδρόμησης Poisson, όπου στην ουσία φάνηκε ότι μπορεί να προβλέψει ικανοποιητικά τον

αριθμό των γκολ που επιτυγχάνει μια ομάδα, όπου εν τέλει χρησιμοποιήθηκε για την πρόβλεψη της τελικής βαθμολογίας στα εξεταζόμενα πρωταθλήματα (Karlis and Ntzoufras, 2000). Ωστόσο τα συγκεκριμένα αποτελέσματα αφορούσαν δεδομένα τα οποία ήταν επικεντρωμένα στην απόδοση της κάθε ομάδας ξεχωριστά. Παρόλα αυτά, τα αποτελέσματα είναι μια αρκετά σημαντική βάση, καθώς αφορά και μια διοργάνωση πρωταθλήματος που δεν περιέχει knock-out φάσεις.

Ωστόσο, η έρευνα δεν αφορούσε μόνο τα μεγάλα πρωταθλήματα και με την ανάπτυξη νέων μοντέλων, δόθηκε η δυνατότητα στους άμεσα ενδιαφερόμενους να μελετήσουν και κομμάτια τα οποία δεν αφορούν απλά τον ρυθμό επίτευξης τερμάτων ή μελέτες τις οποίες μπορούσαν να προβλέψουν τον τελικό νικητή της διοργάνωσης, αλλά και τους τρόπους με τους οποίους επιτυγχάνονται αυτά (Armatas et al., 2009), πράγμα το οποίο μας κάνει να πούμε ότι οι στατιστικές μέθοδοι ξεφεύγουν από το συνηθισμένο και δεν αφορούν απλά και μόνο τον στοιχηματισμό, αλλά μπορούν να εφαρμοστούν και για την βελτίωση της αγωνιστικής εμφάνισης των ομάδων.

Σημαντικά αποτελέσματα είναι αυτά που αφορούσαν τη διοργάνωση του UEFA Champions League και των δύο μεγάλων διοργανώσεων των εθνικών ομάδων, το UEFA EURO και το FIFA World Cup. Ένα σημαντικό σκέλος αυτών των ερευνών έκανε μοντελοποίηση μη λαμβάνοντας υπόψη την ομάδα που αφορούσαν αυτά τα δεδομένα (Panaretos, 2003, Szwarc, 2007). Αυτό από τη στιγμή που δε μας αφορά να προβλέψουμε το τελικό αποτέλεσμα, αποτελεί μια απλούστερη μέθοδο. Όπως σε όλες τις έρευνες, οι πίνακες συνάφειας αποτελούν βασικό εργαλείο (Yiannakos, Armatas, 2006).

Σημαντικό παράγοντα στις παραπάνω προσπάθειες αφορά και η εισαγωγή ως παράγοντα της δυναμικής της οποίας έχουν οι ομάδες των οποίων εξετάζουμε. Συγκεκριμένα, η βάση αποτελεί το κάθε φορά Ranking που μπορεί να έχει η ομάδα (Liu et al., 2015) ή το πρωτάθλημα της χώρας που ανήκει (Leontijevic et al., 2018), προσπαθώντας έτσι να δημιουργηθούν κάποιου είδους προφίλ για τις ομάδες που αφορούσαν τα συγκεκριμένα Rankings. Αυτό καθώς και η παραπέρα προσπάθεια ανάλυσης παραγόντων που αφορούν γενικότερα την βελτίωση της αγωνιστικής εικόνας των ομάδων, χωρίς αυτό απαραίτητα να αφορά τους άμεσους παράγοντες που οδηγούν στην επίτευξη τερμάτων.

Στο συγκεκριμένο σημείο είναι απαραίτητο να αναφερθεί ότι σημαντικό ρόλο και πεδίο ανάπτυξης δραστηριοτήτων, δίνοντας μεγάλη αξιοπιστία στα αποτελέσματα, έχει και η ακριβής καταγραφή των δεδομένων, συγκεντρώνοντας δεδομένα και για κατηγορίες τις οποίες μέχρι πρότινος δεν είχε δοθεί ιδιαίτερη βαρύτητα. Συγκεκριμένα, έχουμε ότι τη μεγαλύτερη συμβολή την έχει το σύστημα της OPTA, το οποίο, όπως και τα περισσότερα, καταγράφει τα δεδομένα με ανάλυση μέσω βίντεο, έχοντας την απόλυτη εμπιστοσύνη των συλλόγων αλλά και των αναλυτών απόδοσης (Liu et al., 2013). Μάλιστα, τα τελευταία χρόνια έχει διαδοθεί ευρέως ο υπολογισμός των αναμενόμενων γκολ (Expected Goals – xG) και αναμενόμενων τελικών

πασών (Expected Assists – xA), το οποίο επί της ουσίας υπολογίζει την ποιότητα ενός σουτ ή μίας τελικής πάσας, βασισμένο σε παράγοντες όπως η απόσταση από το τέρμα, η τροχιά της μπάλας κ.ά. Ο υπολογισμός γίνεται με τη βοήθεια πολλών καταγραφών που έχουν γίνει από την παραπάνω εταιρεία. Αυτά τα δύο και ιδιαίτερα τα αναμενόμενα γκολ χρησιμοποιούνται κυρίως ως δείκτης της απόδοσης μιας ομάδας.

Στη συγκεκριμένη διατριβή θα αναλύσουμε τα βασικά στατιστικά δεδομένα που αφορούν τη διοργάνωση του UEFA Champions League για τις τελευταίες 3 αγωνιστικές περιόδους που ολοκληρώθηκαν πριν την έναρξη της πανδημίας του SARS-COV-2. Ο λόγος που επιλέχθηκαν τα δεδομένα που περιέχονται στην ιστοσελίδα της UEFA (www.uefa.com) είναι αφενός η έλλειψη της τεχνολογίας για την συγκέντρωση των δεδομένων μέσω της ανάλυσης βίντεο και αφετέρου για να εξετάσουμε την επάρκεια των συγκεκριμένων δεδομένων, τα οποία αφορούν το σύνολο του αγώνα και όχι μόνο το κομμάτι της επίτευξης γκολ. Σκοπός μας να αναδείξουμε πλευρές του παιχνιδιού που παίζουν καθοριστικό ρόλο. Επίσης, σε συνέχεια προηγούμενων μελετών (Barreira et al., 2014), θα γίνει απόπειρα να αναδειχθεί η σημασία του πεδίου που αφορά την ανακατάληψη του παιχνιδιού (Ball Recovery).

Πιο αναλυτικά, στο Κεφάλαιο 1 της παρούσας διατριβής γίνεται μια εισαγωγή στα δεδομένα που θα αξιοποιηθούν και η περιγραφική τους ανάλυση, έτσι ώστε να προκύψουν κάποια πρώτα συμπεράσματα αναφορικά με τους παράγοντες εκείνους που είναι καθοριστικοί για την απόδοση μιας ποδοσφαιρικής ομάδας. Δίνεται επιπλέον η δυνατότητα να γίνει διασταύρωση των αποτελεσμάτων με αυτά που προκύπτουν στα επόμενα κεφάλαια, έτσι ώστε να αποδειχθεί η αξιοπιστία τους.

Στο Κεφάλαιο 2 γίνεται έλεγχος για την εξεύρεση συσχετίσεων ανάμεσα στα χαρακτηριστικά του συνόλου των δεδομένων. Λόγω της ύπαρξης διαφορετικού τύπου μεταβλητών στο σύνολο των δεδομένων, υπήρξε η ανάγκη για τη χρήση παραπάνω από δύο μεθόδων. Συγκεκριμένα, χρησιμοποιήθηκαν ο συντελεστής συσχέτισης του Pearson για να μελετηθούν οι συσχετίσεις ανάμεσα σε συνεχείς μεταβλητές, ο συντελεστής συσχέτισης του Spearman για συσχετίσεις που αφορούν συνεχείς και διακριτές μεταβλητές και οι πίνακες συνάφειας αποκλειστικά ανάμεσα σε διακριτές μεταβλητές.

Στο Κεφάλαιο 3 είχαμε την δοκιμή αλγορίθμων εξόρυξης δεδομένων για την εξαγωγή βασικών συμπερασμάτων αναφορικά με τις μεταβλητές που αποτυπώνουν τα αποτελέσματα ενός αγώνα. Αρχικά χρησιμοποιήθηκε ο αλγόριθμος random forest έτσι ώστε να γίνει feature selection στα δεδομένα. Να φανεί, δηλαδή, με βάση τον συγκεκριμένο αλγόριθμο ποιες είναι οι μεταβλητές αυτές που καθορίζουν το αποτέλεσμα ενός αγώνα. Εν συνεχεία, έγινε δοκιμή 2 αλγόριθμων κατηγοριοποίησης (classification), που σαν στόχο είχε την επιβεβαίωση των αποτελεσμάτων που είχε ο random forrest και την ανάδειξη του καταλληλότερου αλγορίθμου για τα συγκεκριμένα δεδομένα.

Τέλος, στο Κεφάλαιο 4 έγινε η προσαρμογή δύο γενικευμένων γραμμικών μοντέλων με βάση τις δύο μεταβλητές που χρησιμοποιήθηκαν ως μεταβλητές απόκρισης. Ο συγκεκριμένος τρόπος ανάλυσης των δεδομένων έδωσε τη δυνατότητα ποσοτικοποίησης της σχέσης που υπάρχει ανάμεσα στις μεταβλητές απόκρισης και στις μεταβλητές που κρίθηκε ότι τις επηρεάζουν.

ΚΕΦΑΛΑΙΟ 1

ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ

Για να γίνει μια επαρκής στατιστική ανάλυση πρέπει να έχουμε λάβει υπόψη μας όλες τις παραμέτρους που αφορούν τα δεδομένα μας και την επεξεργασία τους. Έτσι θα αναλύσουμε όλα τα στάδια της διαδικασίας που ακολουθήθηκε, ξεκινώντας από την περιγραφή της διοργάνωσης.

1.1 Για το UEFA Champions League

Η διοργάνωση UEFA Champions League είναι η κορυφαία διοργάνωση σε επίπεδο ποδοσφαιρικών συλλόγων στην Ευρώπη. Διεξάγεται κάθε χρόνο και προκρίνονται οι κορυφαίες ομάδες από τα εγχώρια ευρωπαϊκά πρωταθλήματα, ανάλογα με τη βαθμολογία που έχει η χώρα στην ειδική βαθμολογία της UEFA. Η διοργάνωση αποτελείται από 4 προκριματικούς γύρους, οι οποίοι καταλήγουν στην φάση των ομίλων. Στη φάση αυτή προκρίνονται συνολικά 32 ομάδες οι οποίες χωρίζονται σε 8 ομίλους των 4 ομάδων. Για να διεξαχθεί η κλήρωση των ομίλων, οι ομάδες χωρίζονται σε 4 γκρουπ δυναμικότητας. Βασικό κριτήριο αυτού του διαχωρισμού είναι η βαθμολογία της UEFA σε επίπεδο συλλόγων. Οι κορυφαίες δύο ομάδες από κάθε όμιλο προκρίνονται στον δεύτερο γύρο της διοργάνωσης, όπου ξεκινάνε και οι knock-out φάσεις. Συγκεκριμένα, οι ομάδες από τους ομίλους πηγαίνουν στη «Φάση των 16», όπου κληρώνονται οι ομάδες που τερμάτισαν πρώτες στον όμιλο ενάντια σε αυτές που τερμάτισαν δεύτερες. Οι αντίπαλοι από εκείνη τη φάση μέχρι τον τελικό της διοργάνωσης αναμετρώνται σε διπλά παιχνίδια, δηλαδή μία φορά στην έδρα της ομάδας που θα κληρωθεί πρώτη από το ζευγάρι και μία στην έδρα της δεύτερης. Ο τελικός της διοργάνωσης διεξάγεται σε ένα παιχνίδι σε έδρα η οποία είναι γνωστή από την αρχή της αγωνιστικής περιόδου, με την παρουσία φιλάθλων και των δύο ομάδων. Αν λάβουμε υπόψη μας όλα τα παραπάνω, αυτό μας δίνει συνολικά 125 αγώνες σε μία αγωνιστική περίοδο, από τους οποίους έχουμε συνολικά 96 στη φάση των ομίλων και 29 στη Knock-Out φάση.

1.2 Για τα δεδομένα

Τα δεδομένα μας αφορούν την διοργάνωση του UEFA Champions League για τις αγωνιστικές περιόδους 2016-2017, 2017-2018 και 2018-2019. Από το δείγμα μας έχουν αφαιρεθεί οι τελικοί της διοργάνωσης. Αυτό συμβαίνει λόγω του γεγονότος ότι διεξάγονται σε ουδέτερη έδρα και με την παρουσία οπαδών και των δυο ομάδων, πράγμα που δεν αντιπροσωπεύει την μελέτη που θέλουμε να πραγματοποιήσουμε. Επιπλέον, δεν έχουν ληφθεί υπόψη τα παιχνίδια που πήγαν στην παράταση, λόγω του ότι είχαν διάρκεια μεγαλύτερη των

90 λεπτών. Με βάση τα παραπάνω προκύπτει ότι οι συνολικοί αγώνες που συμπεριλήφθηκαν στη μελέτη είναι 371. Η καταγραφή τους έγινε με τη βοήθεια του προγράμματος επεξεργασίας υπολογιστικών φύλλων Microsoft Excel. Πηγή άντλησής τους ήταν η επίσημη ιστοσελίδα της Ένωσης Ευρωπαϊκών Ποδοσφαιρικών Ομοσπονδιών (UEFA). Η ανάλυση έγινε μέσω του στατιστικού πακέτου της R.

Οι εγγραφές του συνόλου των δεδομένων μας αποτελούνται από τις αποδόσεις που είχαν οι ομάδες στους αγώνες που έλαβαν μέρος ανά χαρσακτηριστικό. Ξεκινώντας από τα ονόματα (χαρακτηριστικό Team) των παρατηρήσεων, αυτά έχουν χωριστεί με βάση, εκτός από το όνομα της ομάδας, τον γύρο στον οποίο αγωνίζονται. Αυτό ήταν κάτι το οποίο μας έδωσε σημαντική βοήθεια στην καταγραφή των δεδομένων.

1.3 Για τα χαρακτηριστικά

Παρακάτω θέτουμε τα βασικά χαρακτηριστικά που αναλύουμε τα δεδομένα μας:

Goals	Γκολ που πέτυχε η ομάδα στον αγώνα
Tattempts	Συνολικές προσπάθειες για γκολ
OnTarget	Προσπάθειες στο στόχο
Blocked	Προσπάθειες που αποκρούστηκαν
Corners	Κόρνερ
Offsides	Οφσάιντ
Possesion	Κατοχή
PassAcc	Ακρίβεια πασαρίσματος
PassT	Συνολικές πάσες
PassC	Επιτυχημένες πάσες
Distance	Η απόσταση που κάλυψε η ομάδα (το άθροισμα του συνόλου των παικτών που αγωνίστηκαν)
BallsRec	Ανακατάληψη μπάλας
Tackles	Τάκλιν
Blocks	Μπλοκαρίσματα

Clearances	Αποκρούσεις
Yellow	Κίτρινες Κάρτες
Red	Κόκκινες Κάρτες
FoulsCom	Κερδισμένα Φάουλ
Home.Away	Έδρα (0: Εκτός, 1:Εντός)
Round	Γύρος (0: Φάση ομίλων,1: Φάση Knock-out)
Ranking	Κατάταξη με βάση τη βαθμολογία της UEFA

Πίνακας 1.1: Μεταβλητές που συμπεριλήφθηκαν στα δεδομένα.

Για τις ανάγκες της ανάλυσής μας δημιουργήσαμε 3 επιπλέον μεταβλητές

GoalsCon	Τα γκολ που δέχθηκε η ομάδα στον αγώνα
Group	Γκρουπ Δυναμικότητας
Result	Αποτέλεσμα αγώνα για την ομάδα (0: Ήττα, 1:Ισοπαλία, 2: Νίκη)

Πίνακας 1.2: Μεταβλητές που δημιουργήθηκαν από τα δεδομένα.

Ο τρόπος με τον οποίο δημιουργήθηκαν οι συγκεκριμένες μεταβλητές, οι οποίες δεν αναφέρονται στα δεδομένα που αντλήσαμε από την ιστοσελίδα της UEFA, φαίνεται στο Παράρτημα Ε με τον κώδικα της R. Οι λόγοι για τους οποίους προχωρήσαμε στην συγκεκριμένη προσθήκη είναι οι εξής:

1. Γενικώς αυτό που παρατηρούμε για τα χαρακτηριστικά που έχουμε συμπεριλάβει στη μελέτη μας, είναι ότι ένας βασικός τρόπος για να εξετάσουμε την αποτελεσματικότητα του τρόπου με τον οποίο επιτίθεται μια ομάδα είναι να έχουμε σαν μεταβλητή-στόχο (Target Variable) τα γκολ που σημείωσε. Ωστόσο, παρατηρούμε ότι αν θέλαμε να κάνουμε κάτι αντίστοιχο για τον τρόπο με τον οποίο αμύνεται μια ομάδα, τότε θα είχαμε ένα αρχικό έλλειμα στο ζήτημα της μεταβλητής-στόχου. Αναμφίβολα, η μελέτη μας δε θα σταθεί μόνο στις συγκεκριμένες μεταβλητές, ωστόσο η προσθήκη της μεταβλητής GoalsCon είναι κάτι το οποίο δεν επηρεάζει την επάρκεια των δεδομένων μας.

2. Για την μεταβλητή Group έχουμε ότι στην ουσία δημιουργήσαμε τα 4 γκρουπ δυναμικότητας, αλλά όχι με βάση τον τρόπο με τον οποίο γίνεται αυτός από την UEFA. Ο τρόπος που πραγματοποιήθηκε αυτός ο διαχωρισμός είναι με βάση τα τεταρτημόρια της μεταβλητής Ranking. Άρα, έχουμε ότι το πρώτο Group αντιστοιχεί στο πρώτο τεταρτημόριο κλπ. Αυτό συμβαίνει λόγω του γεγονότος ότι το dataset περιέχει

παρατηρήσεις από 3 σεζόν με αποτέλεσμα η κατάταξη των ομάδων να αλλάζει από τη μία σεζόν στην άλλη.

3. Παρόλη την προσθήκη της μεταβλητής GoalsCon για τους λόγους που προαναφέρθηκαν, υπάρχει ένα κενό αναφορικά με το κατά πόσο θα μπορούμε να βγάλουμε ασφαλή συμπεράσματα εν γένει για το αποτέλεσμα ενός αγώνα. Αυτό χρειάζεται διότι μπορεί π.χ. ένας αγώνας να κριθεί υπέρ μιας ομάδας, αλλά να υπάρχει αθροιστικά ένας μεγάλος αριθμός τερμάτων, πράγμα το οποίο μπορεί να μας έδινε μια πολύ κακή αμυντική λειτουργία. Το να προσθέσουμε την μεταβλητή Result θα μας δώσει τη δυνατότητα να μπορέσουμε να λειτουργήσουμε συμπληρωματικά με έναν επιπλέον παράγοντα στην μελέτη μας, μαζί με τους προηγούμενους.

1.4 Έρευνα για missing values

Οι ελλιπείς τιμές (missing values) είναι στοιχεία του συνόλου των δεδομένων μας τα οποία για διάφορους λόγους που έχουν να κάνουν με την καταγραφή των δεδομένων δεν καταγράφηκαν, με αποτέλεσμα στα στατιστικά πακέτα που επεξεργαζόμαστε τη βάση δεδομένων να εμφανίζονται με τη μορφή NA. Η ύπαρξη πολλών τέτοιων τιμών μπορεί να προκαλέσει προβλήματα στα αποτελέσματα της ανάλυσής μας, παρόλο που οι μέθοδοι οι οποίες θα χρησιμοποιηθούν θα είναι κατάλληλες.

Λόγω της έλλειψης ανάλογης μελέτης για τις συγκεκριμένες αγωνιστικές περιόδους, δεν χρησιμοποιήθηκε κάποια έτοιμη βάση δεδομένων. Αυτό είχε σαν αποτέλεσμα να δημιουργηθεί από την αρχή. Μέσα από αυτή τη διαδικασία μας δόθηκε η δυνατότητα να ανακαλύψουμε την ύπαρξη ή μη των ελλιπών τιμών χωρίς να χρησιμοποιήσουμε κάποιους από τους γνωστούς αλγόριθμους της εξόρυξης δεδομένων.

Έτσι παρατηρήθηκε ότι υπάρχουν 4 στοιχεία της βάσης τα οποία περιέχουν ελλιπείς τιμές. Τα συγκεκριμένα αφορούσαν 2 αγώνες και ο λόγος ύπαρξής τους ήταν το γεγονός ότι δεν είχαν καταγραφή από την πηγή άντλησης των δεδομένων. Τα 4 αυτά στοιχεία αφορούσαν μόνο ένα συγκεκριμένο χαρακτηριστικό, το Blocks.

Ο τρόπος με τον οποίο αντιμετωπίστηκε το ζήτημα είναι να αντικαταστήσουμε με βάση την τιμή που βρήκαμε για τη διάμεσο, που για το χαρακτηριστικό Blocks είναι ίση με 3. Αυτός ο τρόπος, δεδομένου του μικρού αριθμού των παρατηρήσεων που ήταν ελλιπείς δεν επηρεάζει τα αποτελέσματα που έχουμε στα βασικά περιγραφικά στοιχεία, όπως και στα μοντέλα που θα

χρησιμοποιηθούν αργότερα. Αυτό θα φανεί και στον υπολογισμό των βασικών περιγραφικών μέτρων που θα έχουμε παρακάτω.

1.5 Περιγραφική Ανάλυση

Στην ενότητα αυτή θα προχωρήσουμε στην παρουσίαση των βασικών περιγραφικών μέτρων των χαρακτηριστικών των οποίων έχει νόημα να προβούμε σε μια τέτοια διαδικασία. Αυτό σημαίνει ότι από τα αποτελέσματα λείπουν τα χαρακτηριστικά εκείνα που αφορούν κατηγορικές μεταβλητές. Ωστόσο, θα δοθούν κάποιες επιπλέον πληροφορίες οι οποίες είναι χρήσιμες για την περιγραφή του συνόλου των δεδομένων και ενισχύουν κάποια βασικά συμπεράσματα της μελέτης. Όπως είναι ευρύτερα γνωστό, αυτές οι μέθοδοι δε μπορούν να μας δώσουν κάποιο ασφαλές στατιστικό συμπέρασμα. Μπορούν, παρόλα αυτά να δώσουν μια βασική κατεύθυνση στην μελλοντική μας αναζήτηση.

Τα βασικά περιγραφικά μέτρα τα οποία θα αξιοποιήσουμε στην έρευνά μας για κάθε χαρακτηριστικό ξεχωριστά είναι το πρώτο και το τρίτο τεταρτημόριο, δηλαδή οι τιμές εκείνες που αποτελούν το άνω φράγμα για το 25% και το 75% των παρατηρήσεων, αντίστοιχα, η διάμεσος που είναι το άνω φράγμα για το 50% των παρατηρήσεων, ο δειγματικός μέσος και η μέγιστη και η ελάχιστη τιμή. Επιπροσθέτως θα δούμε και την δειγματική τυπική απόκλιση κάθε χαρακτηριστικού.

Τα αποτελέσματα για το σύνολο των δεδομένων φαίνονται παρακάτω:

Χαρακτηριστικό	Min.	1st Qu.	Median	Mean	3rd Qu.	Max	St. Dev.
Goals	0	0	1	1.523	2	8	1.444946
Tattempts	1	9	12	12.89	16	34	5.7996
OnTarget	0	3	4	4.682	6	17	2.841798
Blocked	0	2	3	3.191	4	15	2.274314
Corners	0	3	5	4.982	7	19	3.116447
Offsides	0	1	2	2.411	3.750	11	1.956211
Possesion	28	42	50	50	57.75	72	9.705862
PassAcc	64	80	85	83.82	88	96	5.685699
PassT	196	400	497	508.9	599.8	1096	147.0548
PassC	125	321.5	416	433	528	1014	147.6985
Distance	92.3	107	110.3	110.2	113	143	4.785989
BallsRec	25	42	48	48.13	53	85	8.484652
Tackles	0	2	4	3.889	5	13	2.406396
Blocks	0	2	3	3.179	4	15	2.266687
Clearances	1	10	15	15.92	21	50	7.739662
Yellow	0	1	2	1.964	3	6	1.314313
Red	0	0	0	0.0876	0	2	0.2876353
FoulsCom	3	9	12	12.1	15	30	4.027271

Ranking	1	7	17	25.52	37	108	24.57633

Πίνακας 1.3: Περιγραφικά μέτρα για το σύνολο των συνεχών μεταβλητών.

Από τα παραπάνω αποτελέσματα, εκτός από τις κατηγορικές μεταβλητές, έχει εξαιρεθεί και η μεταβλητή GoalsCon. Ο λόγος είναι ότι τα αποτελέσματα είναι ακριβώς τα ίδια με την μεταβλητή Goals και αυτό είναι φυσιολογικό από την ίδια τη δομή του dataset.

Ενδεικτικά μπορούμε να πούμε ότι οι ομάδες σκόραραν κατά μέσο όρο 1.523 γκολ ανά αγώνα, δημιουργούσαν 12.89 ευκαιρίες εκ των οποίων οι 4.682 κατέληγαν προς την εστία και οι 3.191 είχαν την επιτυχή επέμβαση του τερματοφύλακα.

Επιπλέον, παρατηρούμε ότι στα παραπάνω αποτελέσματα έχουμε προσθέσει και αυτά της μεταβλητής Ranking. Εδώ τονίζεται ότι, όπως αναφέρθηκε στην Ενότητα 1.3, η μεταβλητή Group που αφορά τα Γκρουπ Δυναμικότητας έγινε με βάση τα τεταρτημόρια της Ranking. Άρα, με βάση τα αποτελέσματα του προηγούμενου Πίνακα, έχουμε τον ακριβή διαχωρισμό των 4 Γκρουπ:

ΓΚΡΟΥΠ 1: Οι ομάδες που είχαν κατάταξη από 1 μέχρι 6

ΓΚΡΟΥΠ 2: Οι ομάδες που είχαν κατάταξη από 7 μέχρι 16

ΓΚΡΟΥΠ 3: Οι ομάδες που είχαν κατάταξη από 17 μέχρι 36

ΓΚΡΟΥΠ 4: Οι ομάδες που είχαν κατάταξη από 37 και κάτω.

Με βάση λοιπόν αυτά τα αποτελέσματα, γίνεται εμφανές ότι ένας επιπλέον λόγος που καταλήξαμε σε αυτόν τον διαχωρισμό είναι και το γεγονός ότι αυτός μας δίνει τον ίδιο αριθμό παρατηρήσεων ανάμεσα στα γκρουπ.

Οι παραπάνω μετρήσεις εκτός από μία πρώτη εικόνα για το που κυμαίνονται τα δεδομένα μας είναι ιδιαίτερα χρήσιμη και στην περίπτωση όπου θέλουμε να εξετάσουμε τι γίνεται στις διάφορες κατηγορίες των κατηγορικών μας μεταβλητών. Οι τυχόν διαφορές που μπορεί να υπάρξουν στα βασικά αυτά στοιχεία ανάμεσα στις κατηγορίες αποτελεί μια πρώτη ένδειξη ότι ο εκάστοτε παράγοντας ίσως να επηρεάζει τα αποτελέσματα του χαρακτηριστικού.

Πρώτος και σημαντικότερος παράγοντας που θα ξεκινήσουμε την ανάλυσή μας είναι ο παράγοντας Έδρα, άρα ξεκινάμε με την παράθεση των βασικών περιγραφικών μέτρων ανάλογα με τις κατηγορίες της μεταβλητής Home.Away.

Σημαντικότερος είναι καθώς αν παρατηρήσουμε την συντριπτική πλειοψηφία των δημοσιεύσεων οι οποίες έχουν γίνει πάνω στην Ανάλυση Δεδομένων που αφορούν το ποδόσφαιρο, ακόμα και αυτές που πήραν δεδομένα αποκλειστικά από το UEFA Champions League, θα δούμε ότι βγαίνει σαν βασικό συμπέρασμα ότι ο παράγοντας έδρα επηρεάζει σημαντικά τα αποτελέσματα στα βασικά στατιστικά αποτελέσματα των ομάδων, αν όχι σε όλα.

Στους Πίνακες Α.1 και Α.2 του παραρτήματος Α παρουσιάζουμε τα αποτελέσματα που είχαμε ανά κατηγορία της Home.Away. Υπενθυμίζουμε ότι όταν παίρνει την τιμή 0

αναφερόμαστε στα εκτός έδρας αποτελέσματα και όταν παίρνει την τιμή 1 αναφερόμαστε στα εντός.
Αναφορικά με τα αποτελέσματα που παρατηρούνται ανάμεσα στις 2 αυτές κατηγορίες έχουμε να αναφέρουμε τα εξής στοιχεία:

1. Παρατηρείται ότι οι ομάδες που αγωνίζονται εντός έδρας έχουν καλύτερη επίδοση στο σκοράρισμα, αφού ο δειγματικός μέσος τους είναι 1.768 έναντι 1.278 για τις εκτός έδρας. Επίσης παρατηρείται μια καλύτερη έφεση στη δημιουργία ευκαιριών, αφού υπερτερούν και στις συνολικές (14.47 έναντι 11.32) και στις εύστοχες (5.291 έναντι 4.073). Γενικά, μπορούμε να πούμε ότι οι ομάδες που αγωνίζονται στην έδρα τους έχουν καλύτερα αποτελέσματα αναφορικά με τα χαρακτηριστικά που αφορούν το επιθετικό κομμάτι του παιχνιδιού.

2. Τα πράγματα είναι κάπως διαφορετικά αναφορικά με τα χαρακτηριστικά τα οποία αφορούν το αμυντικό κομμάτι. Ενώ παρατηρούμε μια καλύτερη επίδοση για τις ομάδες που αγωνίζονται εντός έδρας για τα χαρακτηριστικά BallsRec (47.79 έναντι 48.47) και Tackles (3.642 έναντι 4.137), δεν ισχύει το ίδιο και για τα χαρακτηριστικά Blocks και Clearances.

Στους Πίνακες Α.3 και Α.4 μπορούμε να δούμε τα αποτελέσματα ανάλογα με τις κατηγορίες της μεταβλητής Round.

Εδώ μπορούμε να παρατηρήσουμε αν ελέγξουμε κυρίως τον δειγματικό μέσο και το διάμεσο ότι υπάρχουν κάποιες διαφορές ανάμεσα στις δύο κατηγορίες, χωρίς ωστόσο να είναι στην ίδια κλίμακα με αυτές που είχαμε στην προηγούμενη περίπτωση της μεταβλητής Home.Away. Μάλιστα, αυτές παρουσιάζουν και μεγαλύτερη ανομοιομορφία σε σχέση με πριν. Για παράδειγμα, ενώ βλέπουμε ότι, στην περίπτωση όπου μιλάμε για την κατηγορία όπου η Round παίρνει την τιμή 1, υπάρχει μεγαλύτερη τιμή για τα γκολ που σημειώνονται, δεν ισχύει το ίδιο και για τις φάσεις που δημιουργούνται. Αυτό από μόνο του αποτελεί μια αντίφαση. Η συνέχεια (Πίνακες Α.5, Α.6, Α.7 και Α.8) αφορά τις κατηγορίες της μεταβλητής Group.

Το βασικότερο συμπέρασμα που προκύπτει εξετάζοντας, πάλι ενδεικτικά, μόνο τους δειγματικούς μέσους κάθε χαρακτηριστικό, χωρίς στην ουσία να υπάρχει βλάβη της γενικότητας είναι ότι υπάρχουν σημαντικές διαφορές σχεδόν σε όλα τα χαρακτηριστικά που εξετάζουμε ανάμεσα στις κατηγορίες της μεταβλητής Group. Συγκεκριμένα, παρατηρείται μια τάση για καλύτερη απόδοση στο επιθετικό κομμάτι ανάλογα με την θέση που έχει λάβει η ομάδα στην βαθμολογία της UEFA. Ανάμεικτα είναι τα αποτελέσματα αναφορικά με το αμυντικό σκέλος. Ενώ παρατηρείται μια πτωτική τάση όσο «κοιτάμε ψηλότερα» την

βαθμολογία των ομάδων αναφορικά με τις μεταβλητές Blocks και Clearances, δεν ισχύει το ίδιο και με τις BallsRec και Tackles, καθώς υπάρχουν αυξομειώσεις.
Τελευταίο χαρακτηριστικό που θα εξετάσουμε (Πίνακες Α.9, Α.10 και Α.11) τις κατηγορίες του είναι το Result.

Όπως και στις προηγούμενες κατηγορικές που εξετάσαμε, παρατηρούμε ότι υπάρχει διαφορά ανάμεσα στις κατηγορίες αναφορικά με το επιθετικό κομμάτι όσο βελτιώνεται το αποτέλεσμα του αγώνα. Αυτό που φαίνεται, σίγουρα, είναι μια ραγδαία αύξηση στο αριθμό των τερμάτων, το οποίο θεωρείται κάπως προφανές. Ωστόσο, δεν ισχύει το ίδιο για τα υπόλοιπα χαρακτηριστικά του επιθετικού κομματιού. Αν και υπάρχει αύξηση, δεν είναι το ίδιο σημαντική. Σε αυτά που αφορούν το αμυντικό σκέλος, βλέπουμε ότι δεν μπορεί να προκύψει κάποιο προφανές συμπέρασμα, καθώς όλα αυξάνονται αν πάρουμε τις διαφορές που υπάρχουν στην κατηγορία 0 και στην 1, ενώ βλέπουμε μια μείωση αν πάμε από την κατηγορία 1 στην κατηγορία 2.

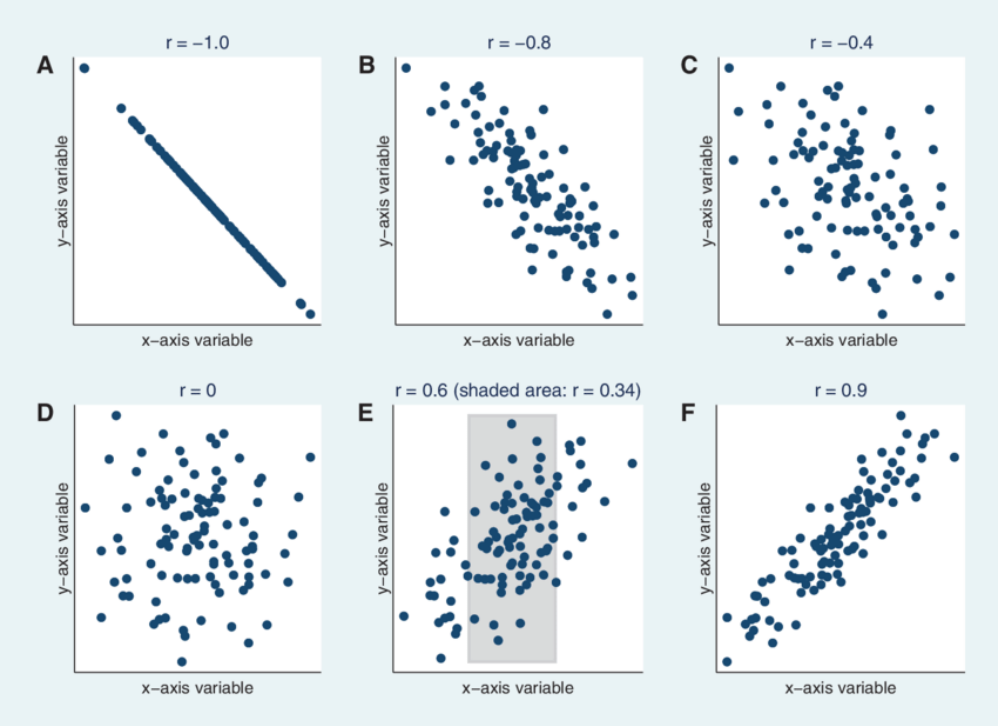
1.6 Γραφική Ανάλυση

Η γραφική μέθοδος για την ανάλυση των αποτελεσμάτων θα μπορέσει να μας δώσει επιπλέον στοιχεία για την εξαγωγή συμπερασμάτων. Ωστόσο, όπως και για την ανάλυση με τη χρήση περιγραφικών μέτρων που αξιοποιήσαμε προηγουμένως, δεν προκύπτουν ασφαλή στατιστικά συμπεράσματα μόνο με τη χρήση αυτών των μεθόδων.

Στη μελέτη αυτή θα χρησιμοποιηθούν δυο βασικοί τύποι διαγραμμάτων. Το διάγραμμα διασποράς (Scatter Plot) και το θηκόγραμμα (Boxplot).

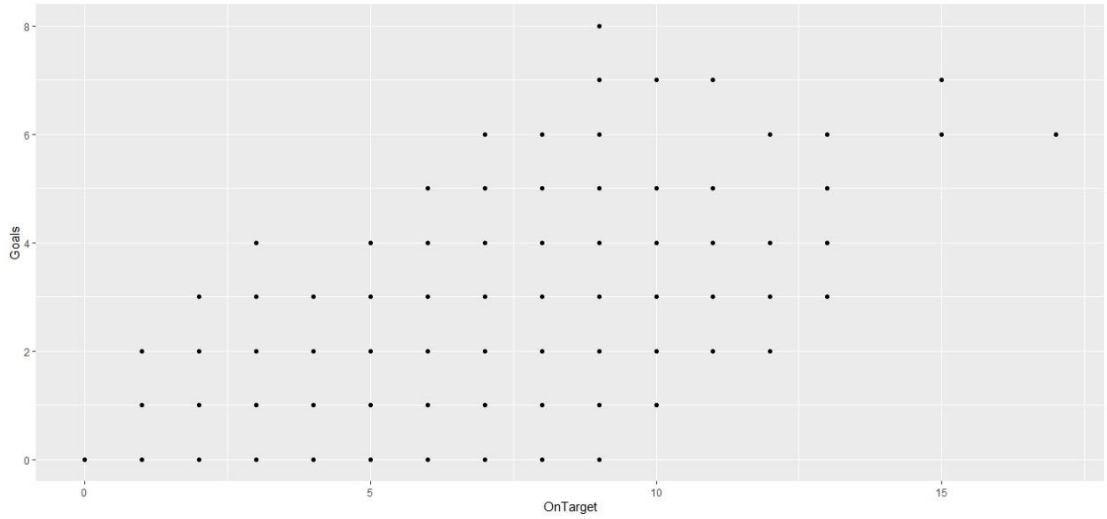
Τα διαγράμματα διασποράς χρησιμοποιούνται για να μελετηθεί η περίπτωση δύο χαρακτηριστικά να έχουν κάποια συσχέτιση, με ποιο χαρακτηριστική τη γραμμική συσχέτιση των μεταβλητών. Απεικονίζουν τις παρατηρούμενες τιμές του συνόλου των δεδομένων μας για τα χαρακτηριστικά που μας ενδιαφέρει να μελετήσουμε σαν σημεία στο επίπεδο με βάση το Καρτεσιανό σύστημα συντεταγμένων, με τη μία να βρίσκεται στον άξονα x και την δεύτερη στον άξονα y. Για να επιλέξουμε ποια από τις μεταβλητές θα μπει στον κάθε άξονα, αρκεί να σκεφτούμε ποια μεταβλητή μας ενδιαφέρει να μελετήσουμε αναφορικά με τη συμπεριφορά της. Στην περίπτωση που η συσχέτιση που έχουμε είναι θετική, δηλαδή η αύξηση της τιμής του χαρακτηριστικού που έχουμε στον άξονα των x (ανεξάρτητη μεταβλητή) σηματοδοτεί αυτόματα και την αύξηση της τιμής του χαρακτηριστικού που έχουμε στον άξονα των y (εξαρτημένη μεταβλητή), αυτό θα φαίνεται στο διάγραμμα διασποράς με μια συνεχή άνοδο των σημείων. Αντίστοιχα, αν η συσχέτιση είναι αρνητική, δηλαδή αν με την αύξηση της τιμής της ανεξάρτητης μεταβλητής παίρνουμε συνεχώς μια μείωση των τιμών της εξαρτημένης μεταβλητής, αυτό θα εκφραστεί με μια συνεχή κάθοδο των σημείων στο διάγραμμα διασποράς. Τέλος, στην περίπτωση που τα δύο χαρακτηριστικά μας είναι ασυσχέτιστα, δηλαδή η αύξηση της τιμής της ανεξάρτητης μεταβλητής δεν συνεπάγεται αυστηρά αύξηση ή μείωση της

εξαρτημένης μεταβλητής, αυτό θα μας δώσει μια τυχαία διασπορά στο γράφημα. Στην παρακάτω εικόνα μπορούμε να δούμε παραδείγματα διαγραμμάτων διασποράς για διάφορες τιμές του συντελεστή συσχέτισης των δύο χαρακτηριστικών.



Διάγραμμα 1.1: Χαρακτηριστικά διαγράμματα διασποράς (Πηγή: https://www.researchgate.net/figure/A-F-Scatter-plots-with-data-sampled-from-simulated-bivariate-normal-distributions-with_fig1_323388613)

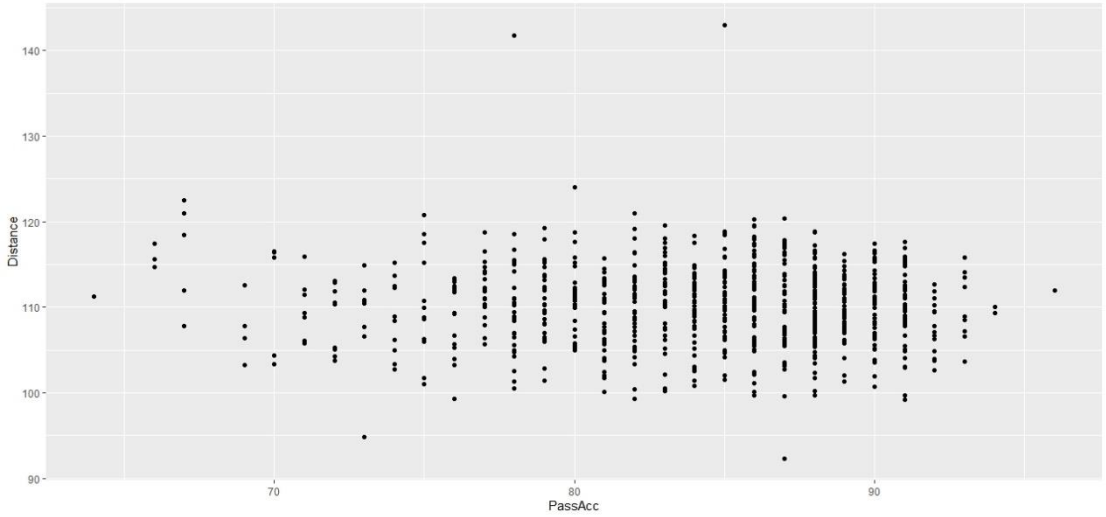
Στη μελέτη των χαρακτηριστικών που έχουμε στο παρόν σύνολο δεδομένων, η προσπάθεια να μελετήσουμε τη συσχέτιση που μπορεί να έχουν τα χαρακτηριστικά μεταξύ τους παρουσίασε αρκετές δυσκολίες. Ο βασικότερος λόγος που φαίνεται ότι συμβαίνει αυτό αφορά κυρίως το είδος των συγκεκριμένων χαρακτηριστικών και το πεδίο τιμών τους. Ενώ γενικότερα τα διαγράμματα διασποράς έχουν εφαρμογή σε όλους τους τύπους χαρακτηριστικών, βλέπουμε ότι στην περίπτωση μας που έχουμε διακριτά χαρακτηριστικά μικρό εύρος τιμών δημιουργούνται διαγράμματα τα οποία δεν βοηθούν στο να έχουμε μια πρώτη, αλλά όχι ασφαλή, εκτίμηση. Καθώς το μεγαλύτερο πλήθος των χαρακτηριστικών μας έχουν την παραπάνω μορφή, οι εκτιμήσεις που έχουμε από τον συγκεκριμένο τρόπο οπτικοποίησης των αποτελεσμάτων είναι πολύ λίγες.



Διάγραμμα 1.2: Διάγραμμα διασποράς της OnTarget με την Goals.

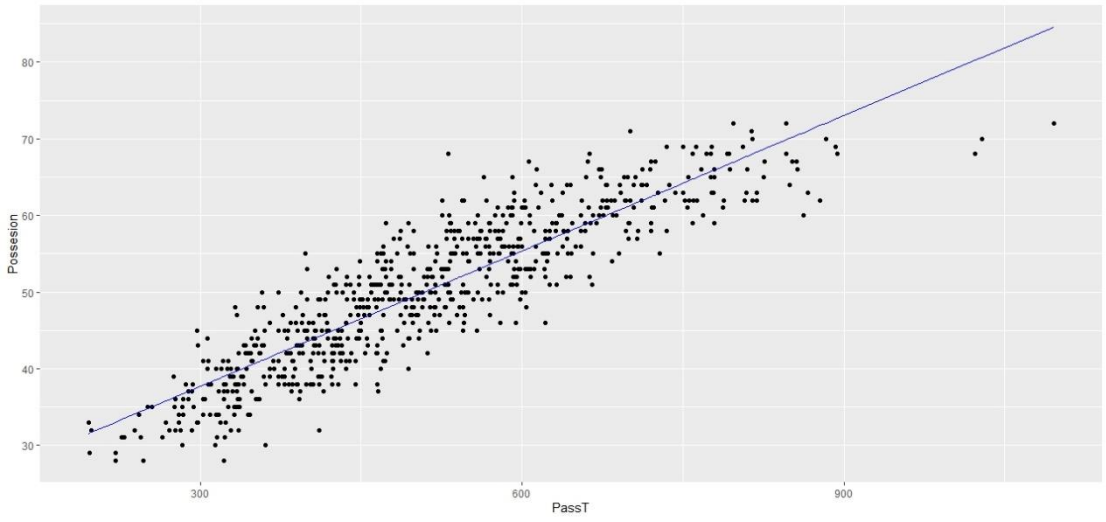
Στο διάγραμμα που προηγείται μπορούμε να δούμε καθαρά το πρόβλημα. Το συγκεκριμένο διάγραμμα είναι το Scatter Plot που έγινε για τις μεταβλητές OnTarget και Goals. Ο λόγος προφανής. Ο ισχυρισμός που θέλουμε να επιβεβαιώσουμε με το συγκεκριμένο διάγραμμα είναι ότι όσες περισσότερες εύστοχες ευκαιρίες κάνει μία ομάδα, τόσο περισσότερα γκολ θα πετύχει. Κάτι που μπορεί να επιβεβαιωθεί με ευκολία. Ωστόσο, όπως βλέπουμε στο προηγούμενο διάγραμμα, παρόλο που φαίνεται μια θετική συσχέτιση, η διασπορά που υπάρχει στο διάγραμμα λόγω του πεδίου τιμών των συγκεκριμένων χαρακτηριστικών δεν μας την εξασφαλίζει.

Υπάρχουν όμως και περιπτώσεις χαρακτηριστικών στις οποίες το διάγραμμα διασποράς μας έδωσε μια πρώτη εικόνα. Ένα ερώτημα το οποίο υπήρξε ήταν το κατά πόσο η ακρίβεια στις πάσες επηρεάζει τη συνολική απόσταση που διανύει μια ομάδα. Υπενθυμίζουμε εδώ ότι η απόσταση που περιγράφεται είναι το άθροισμα των αποστάσεων που έκαναν όλοι οι παίκτες της ομάδας που συμμετείχαν στον αγώνα. Η υπόθεση είναι ότι αν η απόσταση και η ακρίβεια στις πάσες είναι αρνητικά συσχετισμένες, αυτό θα ήταν ένα κίνητρο για τις ομάδες να επικεντρώνονται στην βελτίωση της ακρίβειας πάσας, οπωσδήποτε σε περίπτωση που η ομάδα λόγω συνεχόμενων αγώνων και έντασης θα έχει έλλειμμα φυσικής κατάστασης. Όπως θα δούμε και στο παρακάτω διάγραμμα διασποράς, αυτό δεν επιβεβαιώνεται από τα δεδομένα, τουλάχιστον σαν πρώτη εικόνα. Η διασπορά δείχνει ότι τα συγκεκριμένα χαρακτηριστικά είναι ασυσχέτιστα.

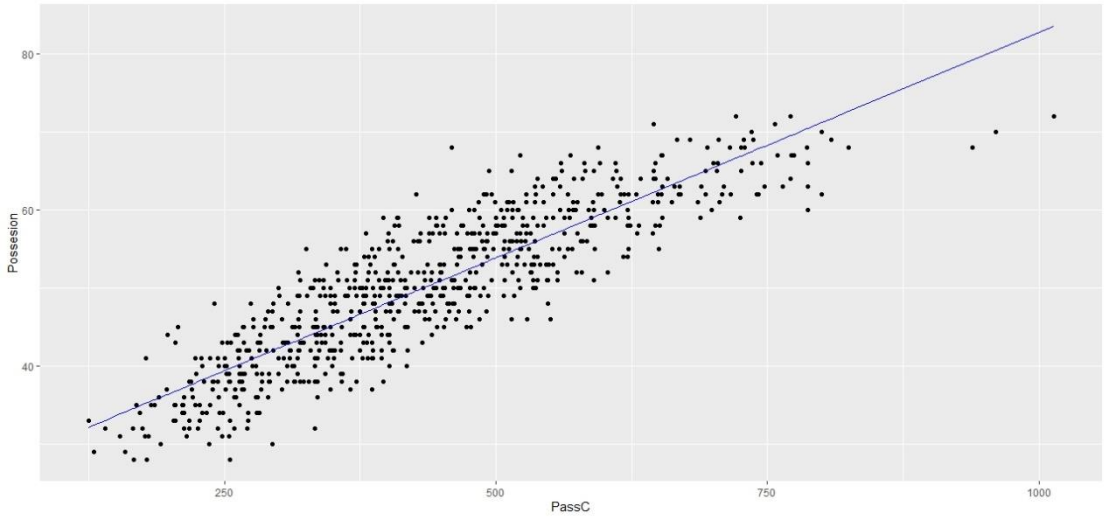


Διάγραμμα 1.3: Διάγραμμα διασποράς της PassAcc με την Distance.

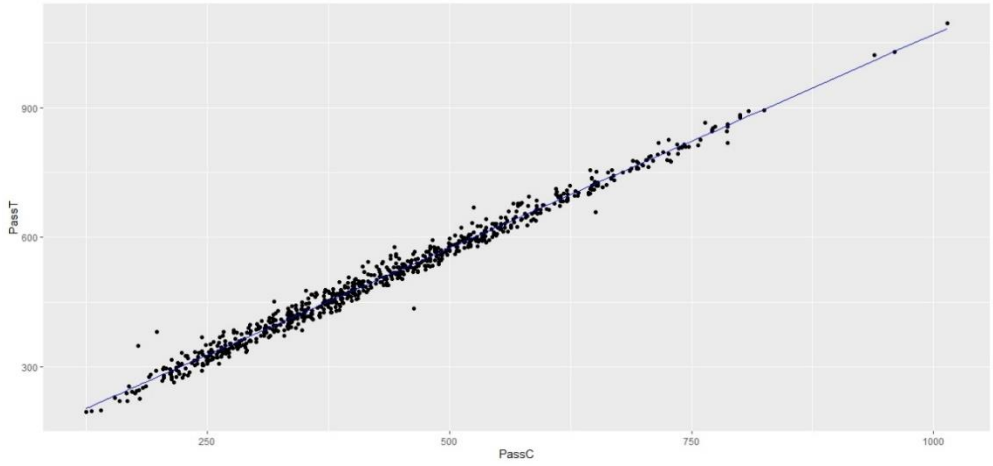
Ωστόσο ένα βασικό συμπέρασμα που προέκυψε από τη μελέτη μέσω των διαγραμμάτων διασποράς είναι η σχέση που έχουν οι συνολικές και οι επιτυχημένες πάσες που πραγματοποιεί η ομάδα σε σχέση με την κατοχή. Από τα διαγράμματα φαίνεται ότι και τα δυο αυτά χαρακτηριστικά έχουν θετική συσχέτιση και μάλιστα γραμμική με το ποσοστό κατοχής, όμως οι συνολικές πάσες είναι αυτές που έχουν μικρότερη διασπορά γύρω από την ευθεία παλινδρόμησης σε σχέση με τις επιτυχημένες πάσες. Λαμβάνοντας υπόψη και το γεγονός ότι ένα μοντέλο παλινδρόμησης μεταξύ αυτών των δύο χαρακτηριστικών φαίνεται να είναι μια αρκετά ικανοποιητική προσέγγιση, μπορούν να βγουν χρήσιμα συμπεράσματα αναφορικά με την επίδραση που έχει το σωστό passing game στο παιχνίδι κατοχής, δηλαδή στην προσέγγιση που έχει μια ομάδα να κρατάει στην κατοχή της την μπάλα και να ελέγχει τον ρυθμό του παιχνιδιού, έτσι ώστε να μπορέσει να είναι αποτελεσματική και στα δυο σκέλη του. Δηλαδή, να πετυχαίνει τη νίκη με ανέπαφη εστία.



Διάγραμμα 1.4: Διάγραμμα διασποράς της PassT με την Possession.



Διάγραμμα 1.5: Διάγραμμα διασποράς της PassC με την Possession.

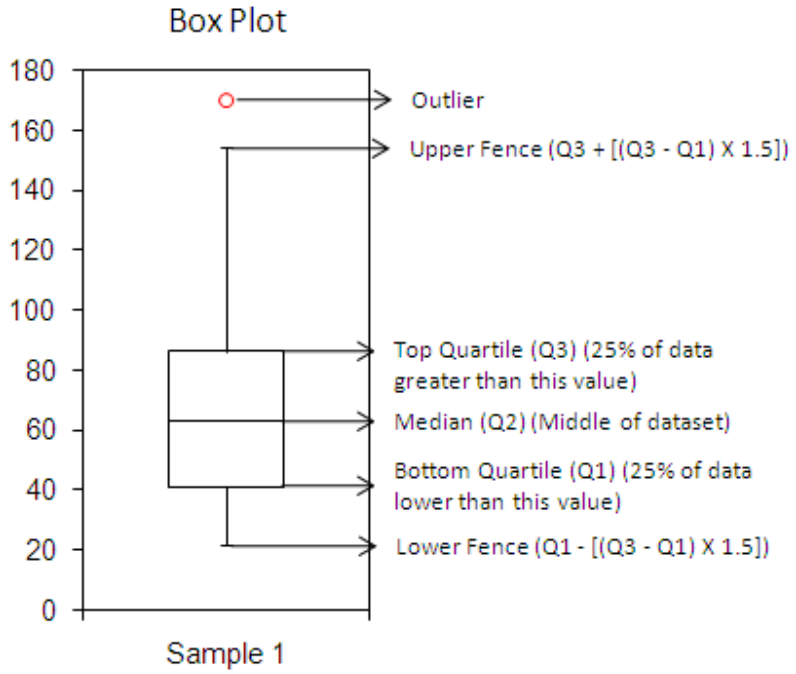


Διάγραμμα 1.6: Διάγραμμα διασποράς της PassC με την PassT.

Το θηκόγραμμα, αντιθέτως, θα φανεί πολύ χρήσιμο στην οπτικοποίηση των αποτελεσμάτων που είχαμε στο προηγούμενο σκέλος της περιγραφικής ανάλυσης που αφορούσε τα περιγραφικά μέτρα που υπολογίστηκαν. Γενικότερα, σαν διάγραμμα αποτελεί μια καλή λύση στην παρουσίαση της διασποράς που έχουν τα δεδομένα. Από αυτό προκύπτει και η βοήθεια που μπορεί να δώσει στην εξέταση της κανονικότητας των δεδομένων. Ένα πλήρες συμμετρικό θηκόγραμμα είναι αρκετά πιθανό να αντιστοιχεί σε κανονικά δεδομένα, χωρίς να αναιρούμε την εφαρμογή των γνωστών ελέγχων και γραφικών απεικονίσεων της κανονικότητας, πράγμα το οποίο θα γίνει και στη συνέχεια. Επιπλέον, το θηκόγραμμα βοηθάει και στην εξέταση ύπαρξης ακραίων παρατηρήσεων (outliers).

Ο λόγος για τον οποίο αναφερθήκαμε μόνο στον δειγματικό μέσο σαν μέτρο σύγκρισης των αποτελεσμάτων έναντι π.χ. της διαμέσου ήταν διπλός. Αφενός, διότι είναι μια αμερόληπτη εκτιμήτρια της μέσης τιμής σε ένα αρκετά ευρύ πλήθος κατανομών και επίσης η διάμεσος, όπως και τα υπόλοιπα περιγραφικά μέτρα απεικονίζονται επαρκώς μέσω του θηκογράμματος. Σε αρκετές περιπτώσεις και ειδικά σε αυτές που έχουν αρκετές ακραίες παρατηρήσεις, η χρήση της διαμέσου είναι εκείνη που ενδείκνυται σε σχέση με τον δειγματικό μέσο.

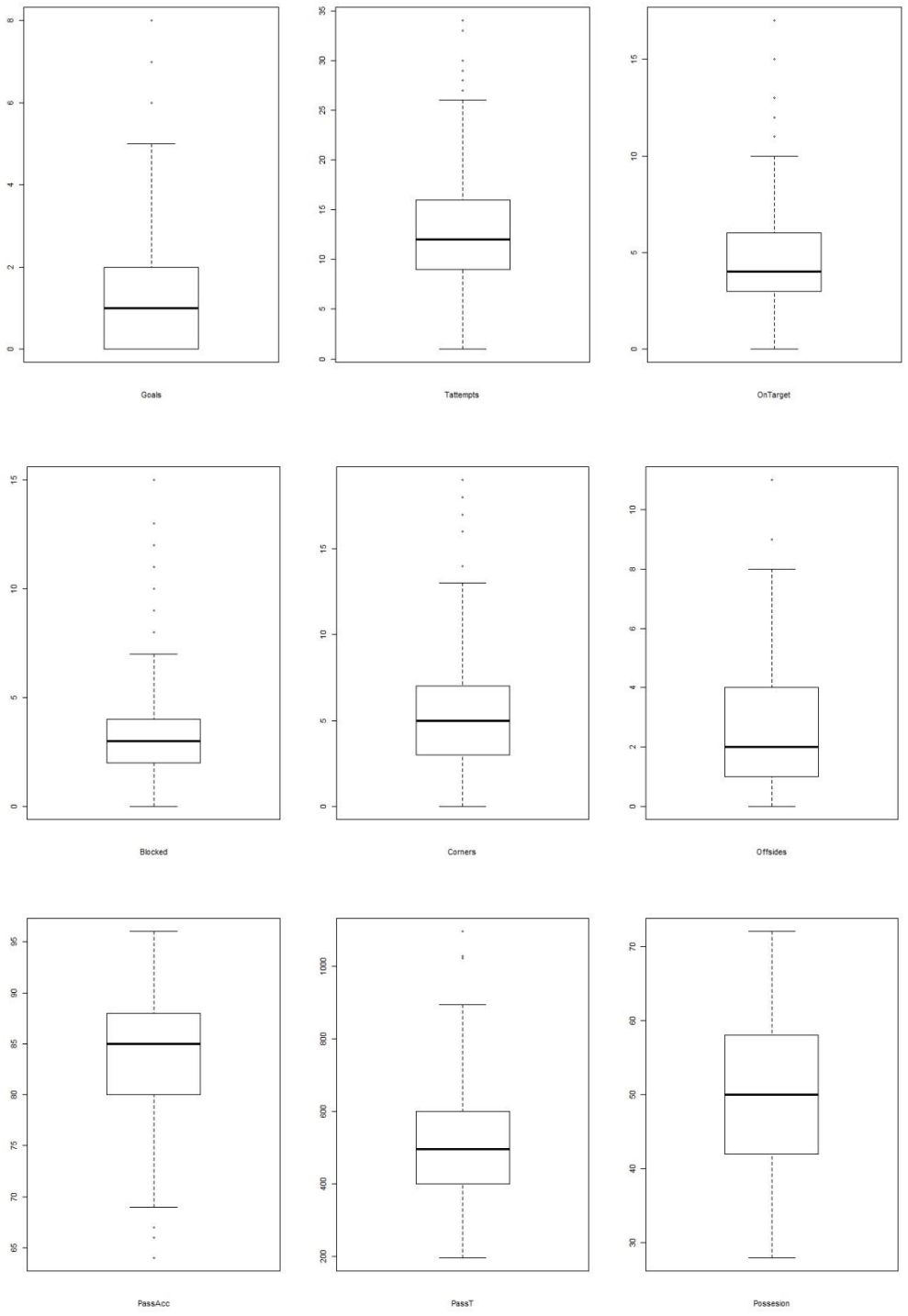
Αναλυτικά, το θηκόγραμμα αποτελείται από ένα παραλληλόγραμμο το οποίο παριστάνει την περιοχή ανάμεσα στο πρώτο και το τρίτο τεταρτημόριο. αυτά αποτελούν τις δύο από τις τέσσερεις πλευρές του. Οι άλλες δυο ουσιαστικά αναπαριστούν το ενδοτεταρτημοριακό εύρος (Interquartile Range – IQR), το οποίο ισοδυναμεί με την απόσταση ανάμεσα στα δύο. Ανάμεσά τους εκτείνεται μια τρίτη ευθεία που απεικονίζει την διάμεσο. Επίσης, εκτείνονται οι τιμές που αφαιρείται 1.5 φορά το IQR από το πρώτο τεταρτημόριο και προστίθεται στο τρίτο τεταρτημόριο. Όποιες εκτείνονται έξω από αυτά τα δυο όρια θεωρούνται outliers.

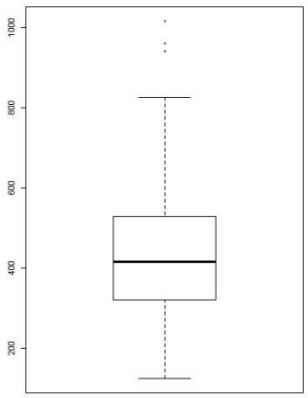


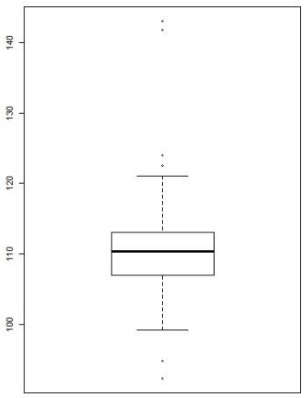
Διάγραμμα 1.6: Θηκόγραμμα με τα βασικά σημεία μελέτης (Πηγή: https://www.listendata.com/2014/08/how-to-read-box-plot.html)

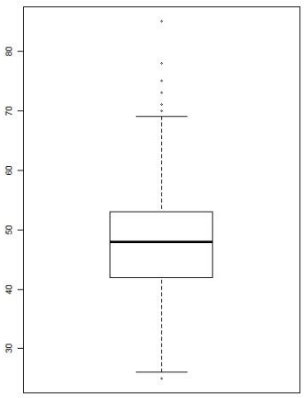
Ο επιπλέον λόγος για τον οποίο θα χρησιμοποιηθούν τα θηκογράμματα στην περιγραφή των δεδομένων και στην οπτικοποίηση των αποτελεσμάτων, αφορά κυρίως στο να ελέγξουμε τις διαφορές που υπάρχουν στις κατηγορίες των μεταβλητών που ελέγξαμε και με τη χρήση των περιγραφικών στατιστικών μέτρων. Αυτό αποτελεί μεν έναν πιο εύκολο και σύντομο τρόπο και θα ενισχύσει την προσπάθεια που θα γίνει για την κατάρτιση του

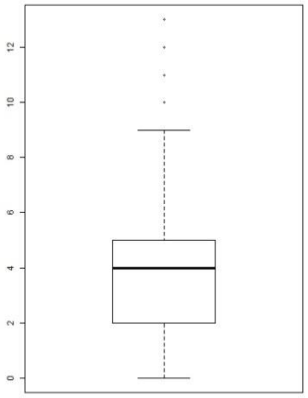
κατάλληλου μοντέλου. Η αρχική παρουσίαση είναι για το σύνολο των χαρακτηριστικών, για να μπορέσουμε να έχουμε ένα μέτρο σύγκρισης για τυχόν διαφορές ανάμεσα στις επιμέρους κατηγορίες. Το μέτρο σύγκρισης αφορά σε μια αρχική υπόθεση περί σημαντικότητας ή μη των διαφορών.

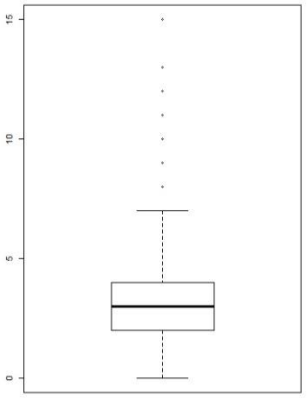


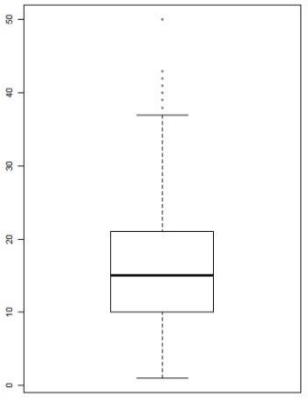

PassC

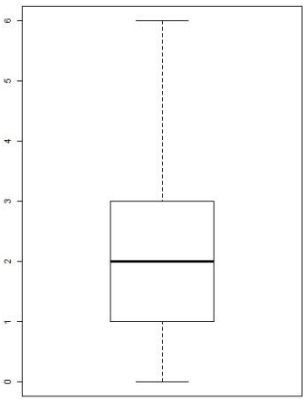

Distance

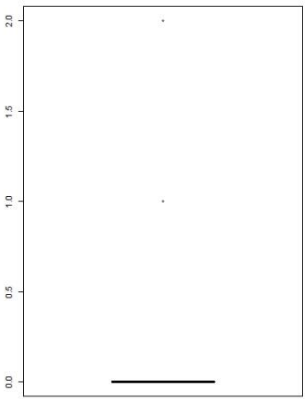

BallsRec

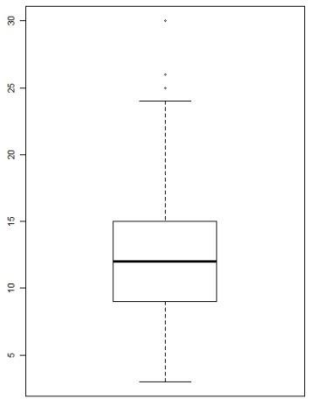

Tackles


Blocks

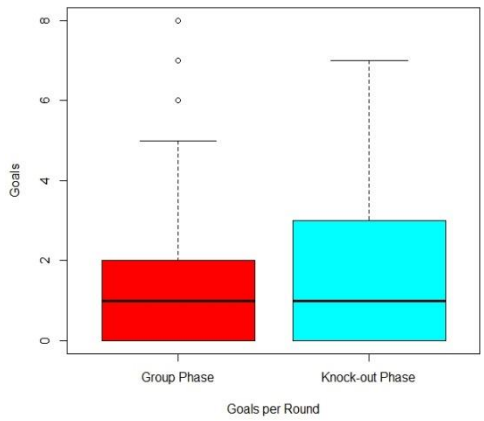

Clearances

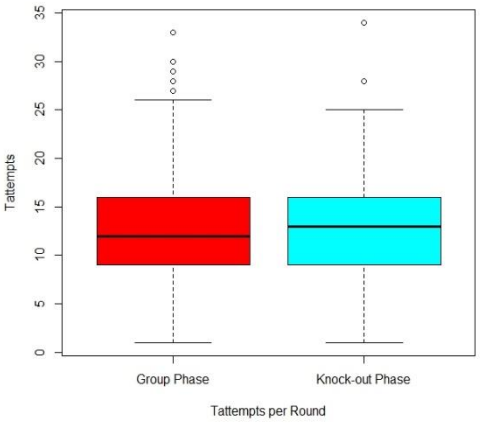

Yellow

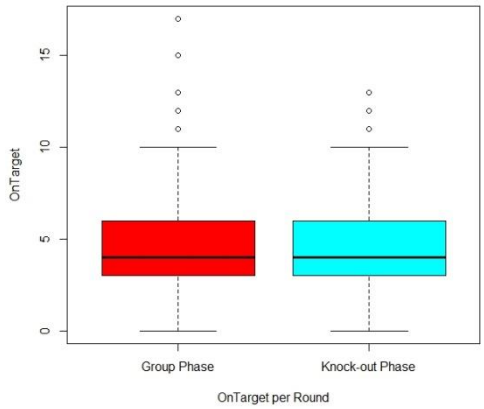

Red

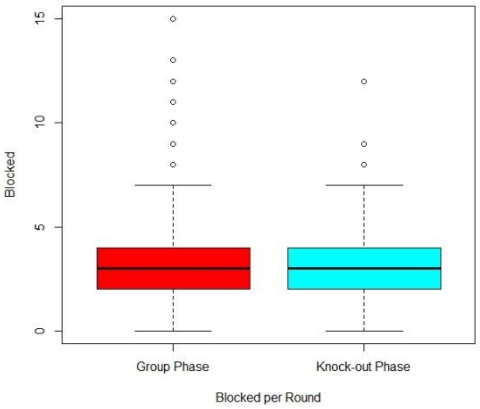

FoulsCom

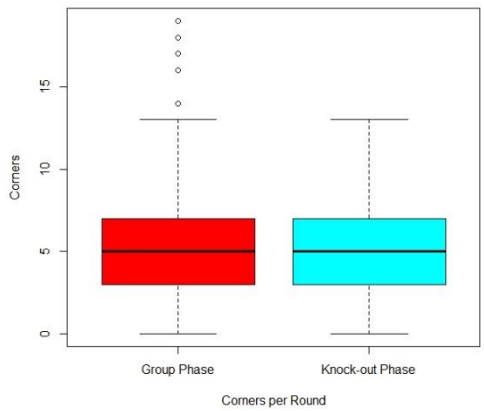
Σε πρώτη φάση έγινε σύγκριση ανάμεσα στις δύο κατηγορίες της μεταβλητής Round. Φανερώνεται και με την σχεδίαση των θηκογραμμάτων ότι οι δύο κατηγορίες δείχνουν να μην έχουν διαφορές στα βασικά μέτρα που υπολογίστηκαν. Οι περισσότερες παρατηρούνται στο κομμάτι της διασποράς των παρατηρήσεων, ωστόσο αναφορικά με την διάμεσο διαφορές φαίνονται στα χαρακτηριστικά Tattempts, PassAcc, Distance, Clearances. Οι διαφορές αυτές είναι ελάχιστες και δεν μπορεί σε καμία περίπτωση να βγει συμπέρασμα αναφορικά με τη σημαντικότητά τους.

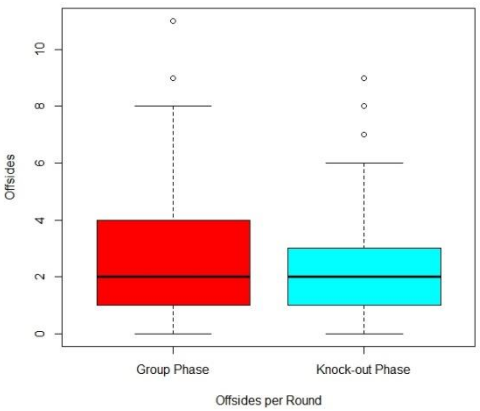


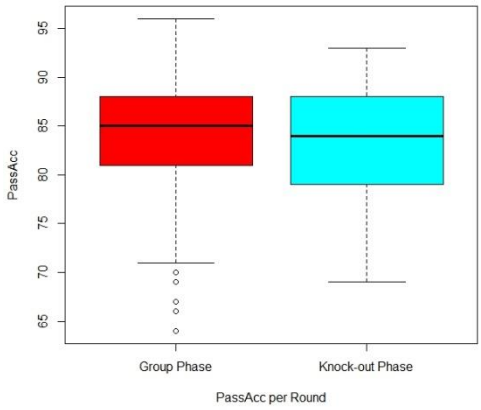


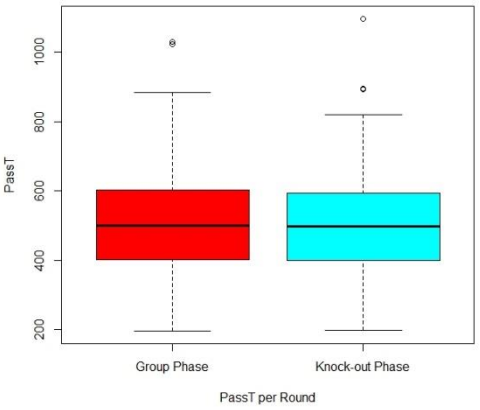


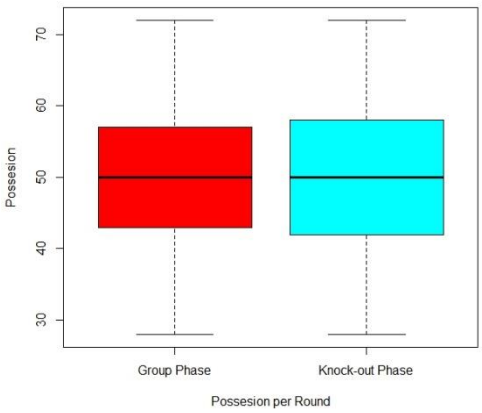


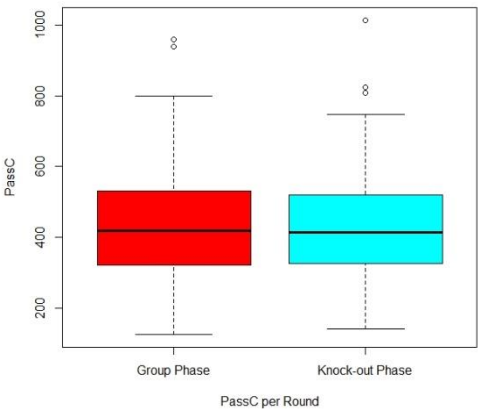


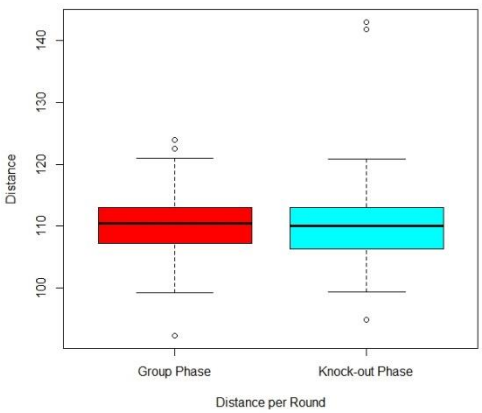


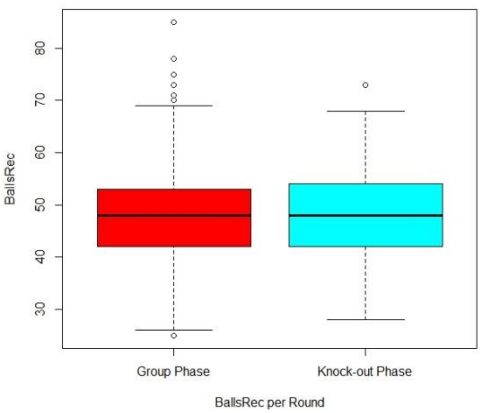

PassAcc
65
70
75
80
85
90
95
Group Phase
Knock-out Phase
PassAcc per Round

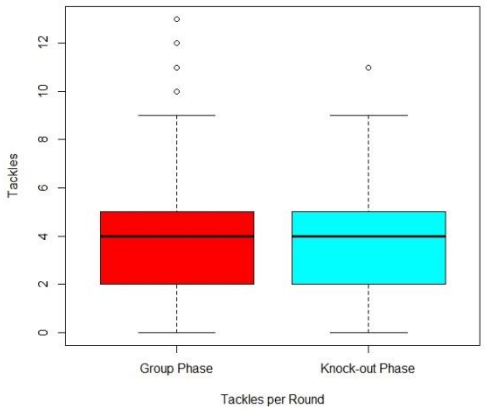

PassT
200
400
600
800
1000
Group Phase
Knock-out Phase
PassT per Round

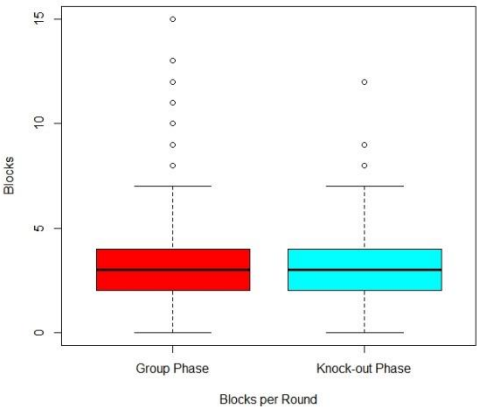

Possesion
30
40
50
60
70
Group Phase
Knock-out Phase
Possesion per Round

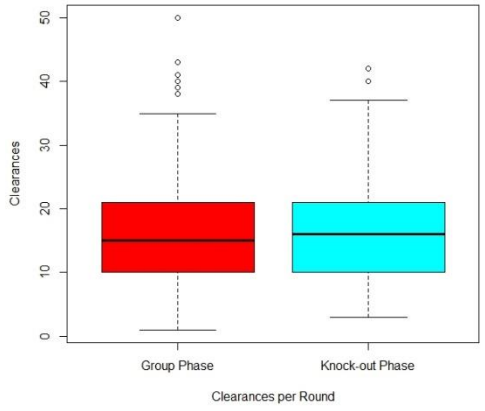

PassC
200
400
600
800
1000
Group Phase
Knock-out Phase
PassC per Round

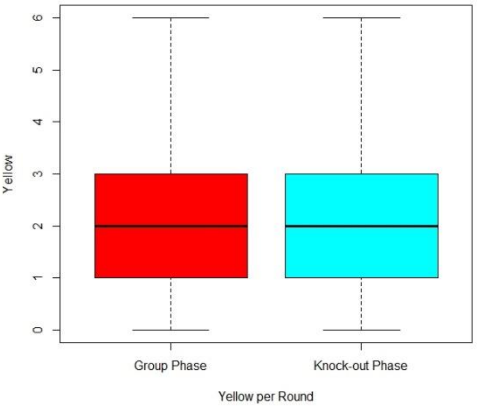

Distance
100
110
120
130
140
Group Phase
Knock-out Phase
Distance per Round

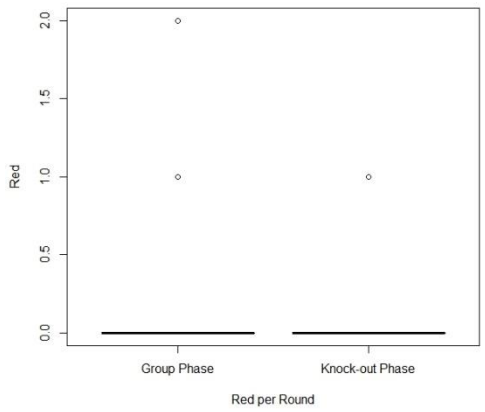

BallsRec
30
40
50
60
70
80
Group Phase
Knock-out Phase
BallsRec per Round

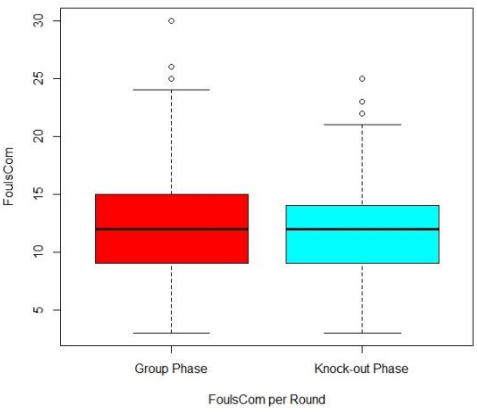

Tackles
0
2
4
6
8
10
12
Group Phase
Knock-out Phase
Tackles per Round


Blocks
0
5
10
15
Group Phase
Knock-out Phase
Blocks per Round

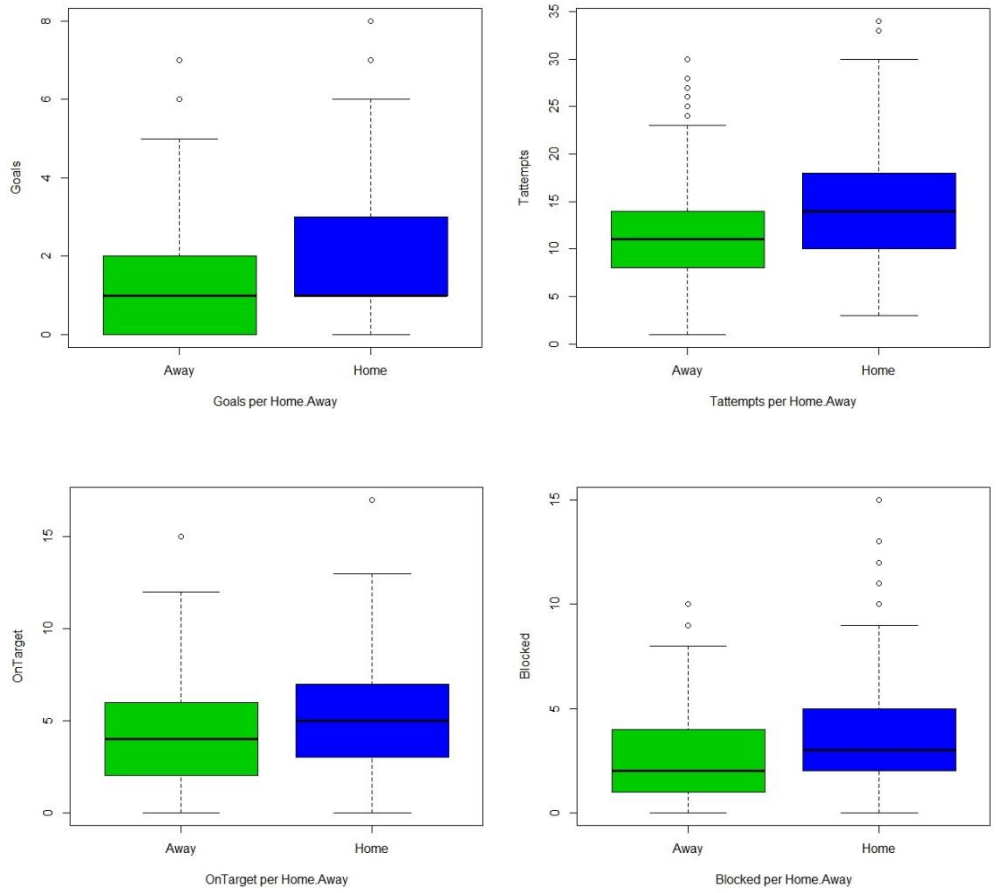

Clearances
0
10
20
30
40
50
Group Phase
Knock-out Phase
Clearances per Round

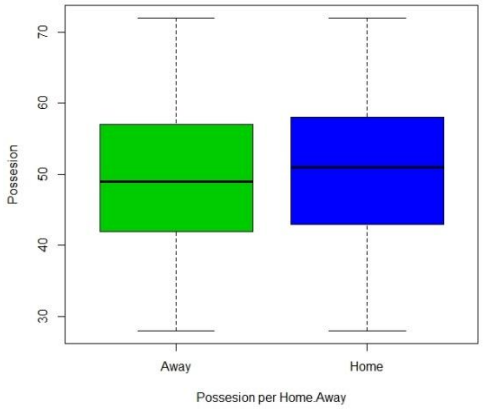

Yellow
0
1
2
3
4
5
6
Group Phase
Knock-out Phase
Yellow per Round

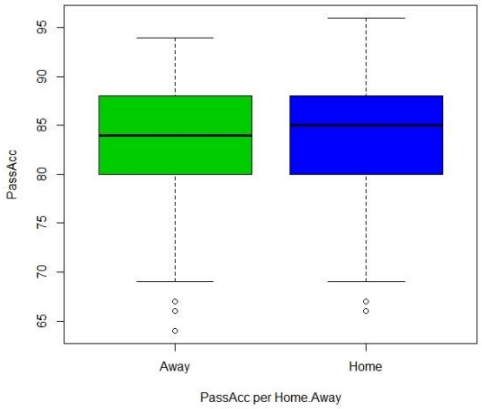

Red
0.0
0.5
1.0
1.5
2.0
Group Phase
Knock-out Phase
Red per Round

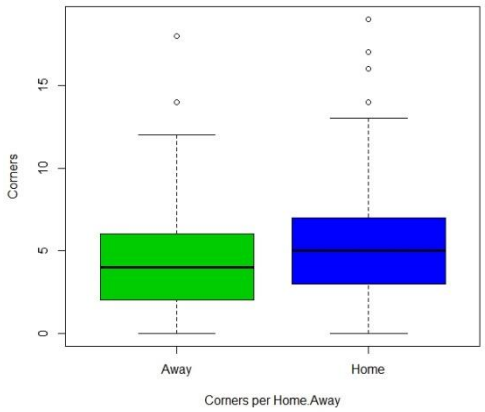

FoulsCom
5
10
15
20
25
30
Group Phase
Knock-out Phase
FoulsCom per Round

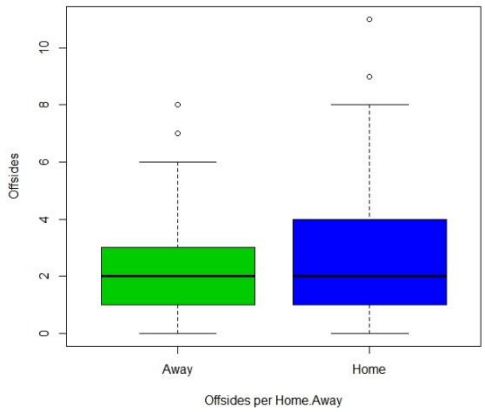
Δεν ισχύει το ίδιο, όπως διαπιστώθηκε και σε προηγούμενη ενότητα, με την επίδραση που πιθανά να έχει στα διάφορα χαρακτηριστικά η μεταβλητή Home.Away. Φαίνεται ότι παίρνουμε διαφορετικά αποτελέσματα στις δύο κατηγορίες μας, με την περίπτωση των εντός έδρας αποτελεσμάτων να δείχνουν υπεροχή στις περισσότερες κατηγορίες. Οι μόνες κατηγορίες που υπερέχουν οι ομάδες οι οποίες είναι εκτός έδρας είναι οι Blocks και Clearances όπως είχε διαπιστωθεί. Οι αποστάσεις μεταξύ των διαμέσων, που μπορούμε να το εκλάβουμε σαν ένδειξη σημαντικής διαφοράς των αποτελεσμάτων στα δεδομένα των δύο κατηγοριών μας και πιθανής επίδρασης του παράγοντα, δείχνουν σε αρκετές περιπτώσεις να είναι σημαντικές. Εξαιρώντας τις μεταβλητές Goals, Offsides, Yellow, Red και FoulsCom, οι υπόλοιπες έχουν διαφορές στις διαμέσους ανάλογα με το αν μία ομάδα αγωνίζεται στην έδρα της ή όχι.

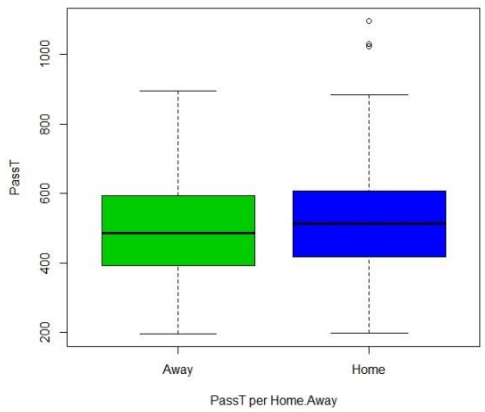


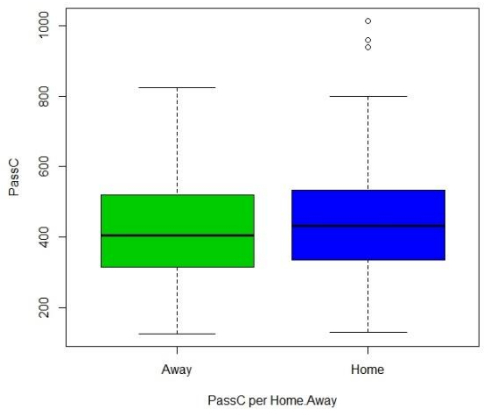

Possesion
30
40
50
60
70
Away
Home
Possesion per Home.Away

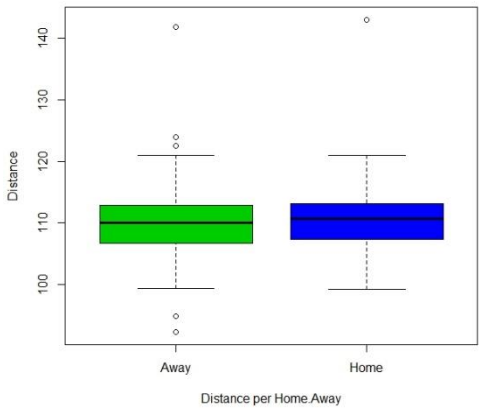

PassAcc
65
70
75
80
85
90
95
Away
Home
PassAcc per Home.Away

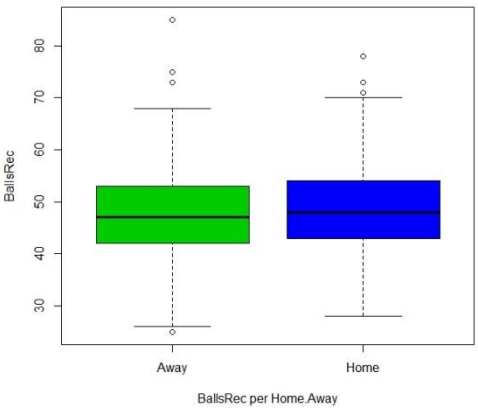

Corners
0
5
10
15
Away
Home
Corners per Home.Away

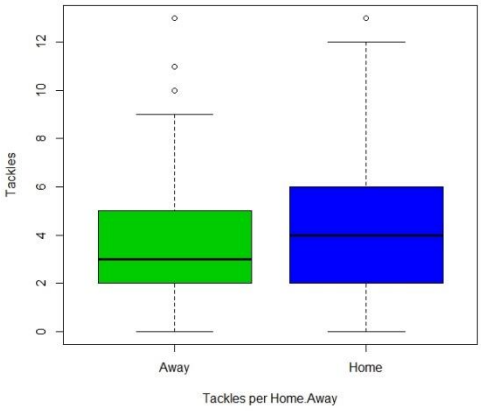

Offsides
0
2
4
6
8
10
Away
Home
Offsides per Home.Away

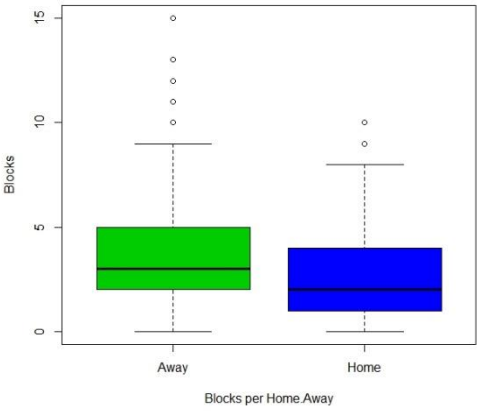

PassT
200
400
600
800
1000
Away
Home
PassT per Home.Away

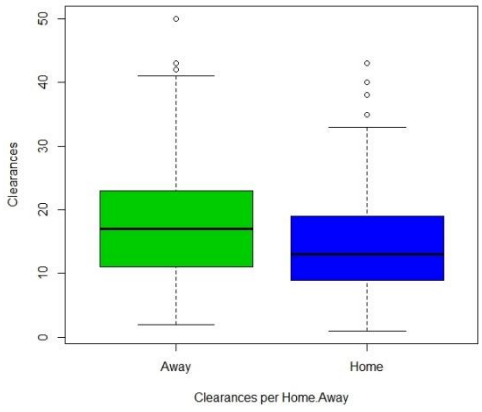

PassC
200
400
600
800
1000
Away
Home
PassC per Home.Away

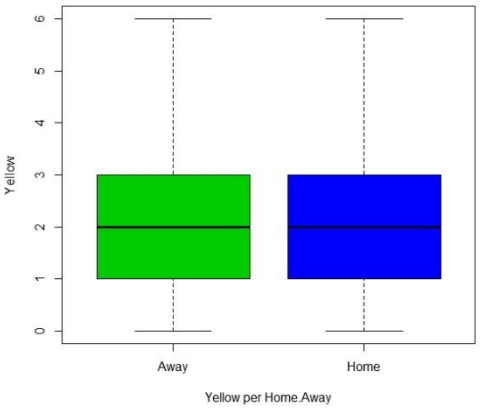

Distance
100
110
120
130
140
Away
Home
Distance per Home.Away

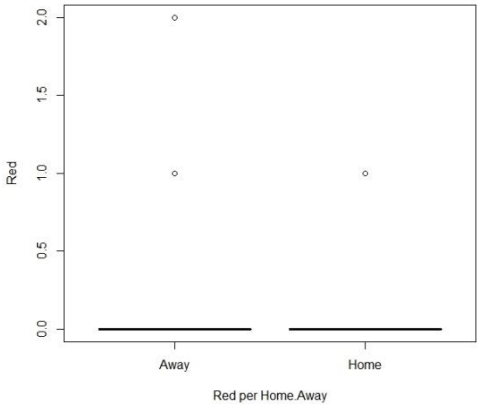

BallsRec
30
40
50
60
70
80
Away
Home
BallsRec per Home.Away

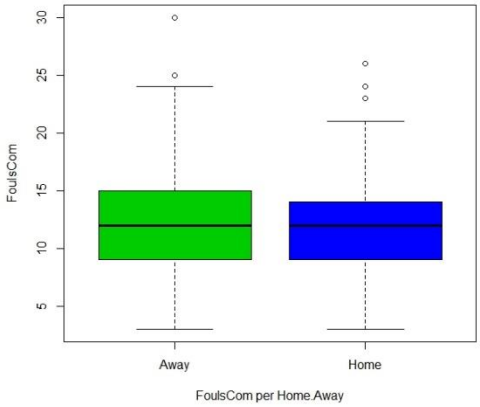

Tackles
0
2
4
6
8
10
12
Away
Home
Tackles per Home.Away


Blocks
0
5
10
15
Away
Home
Blocks per Home.Away


Clearances
0
10
20
30
40
50
Away
Home
Clearances per Home.Away

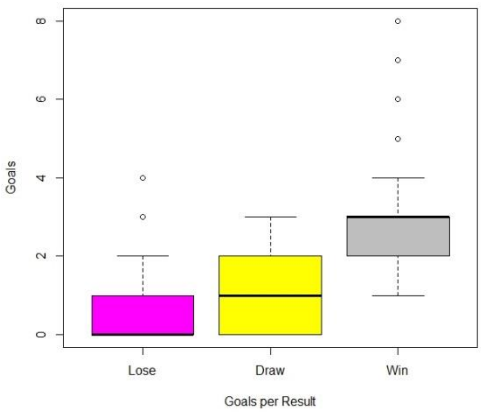

Yellow
0
1
2
3
4
5
6
Away
Home
Yellow per Home.Away

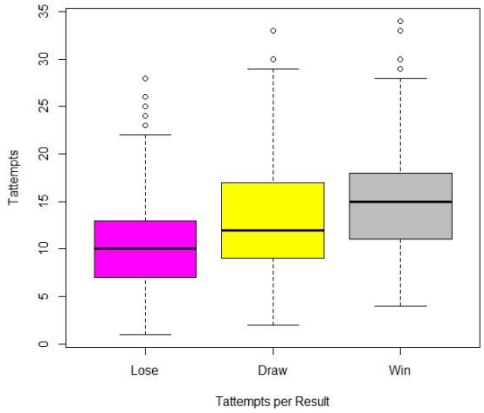


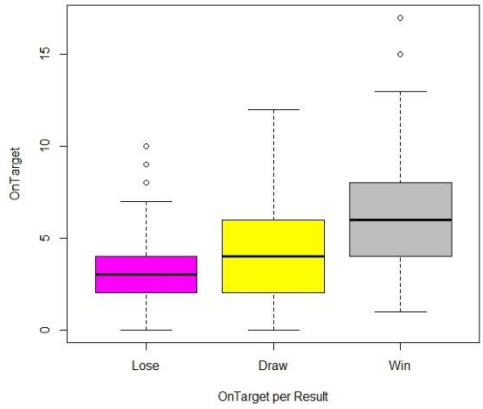


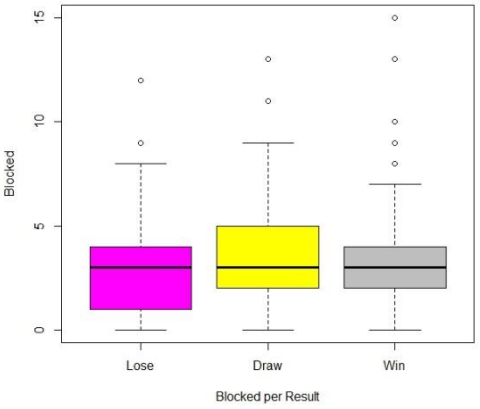
Αρκετά ξεκάθαρα είναι τα πράγματα για τη μεταβλητή που αφορά το αποτέλεσμα του αγώνα για την κάθε ομάδα. Η επιρροή του αποτελέσματος φαίνεται ξεκάθαρα στα χαρακτηριστικά που αφορούν την επιθετική απόδοση της ομάδας, με την κατηγορία Νίκη να μας δίνει τα πιο ανεβασμένα αποτελέσματα, η ισοπαλία ακολουθεί σε σημαντικά χαμηλότερες τιμές και την ήττα να έχει τις χειρότερες. Οι διαφορές που υπάρχουν μας δίνουν την δυνατότητα να πούμε ότι υπάρχει σοβαρή ένδειξη σημαντικής επιρροής του αποτελέσματος.

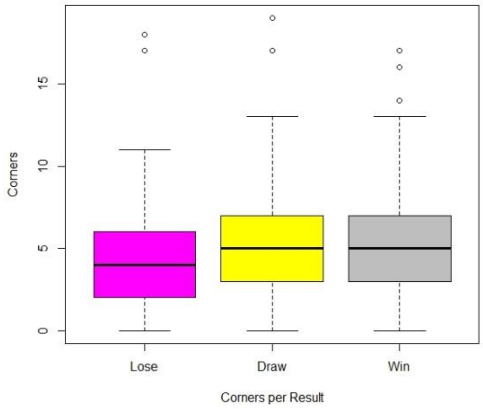
Για τα αποτελέσματα που αντιστοιχούν στο αμυντικό κομμάτι τα πράγματα είναι αβέβαια. Ενώ φαίνεται ότι υπάρχουν κάποιες διαφορές, αυτές είναι μικρότερες και όχι στην ίδια αντιστοιχία με τις προηγούμενες που αναφέρθηκαν. Οπότε δυσκολεύει και η οποιαδήποτε αρχική υπόθεση.

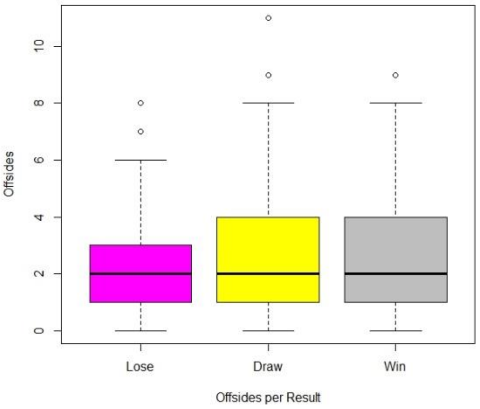


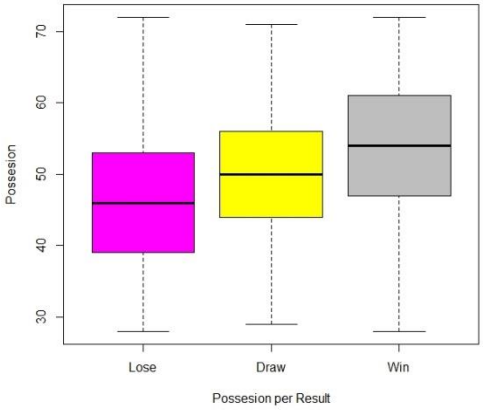


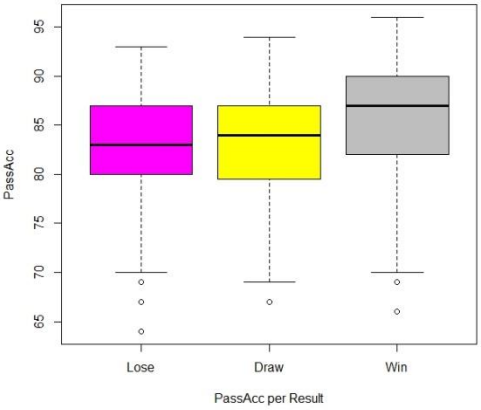

OnTarget
0
5
10
15
Lose
Draw
Win
OnTarget per Result

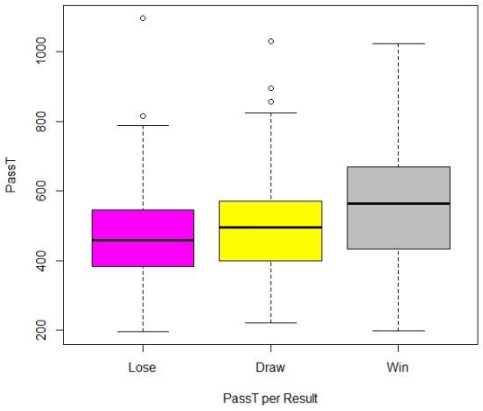

Blocked
0
5
10
15
Lose
Draw
Win
Blocked per Result

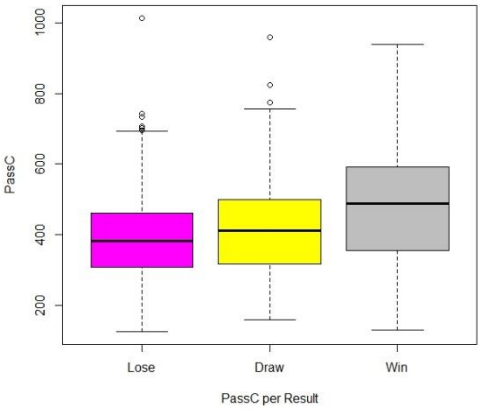

Corners
0
5
10
15
Lose
Draw
Win
Corners per Result

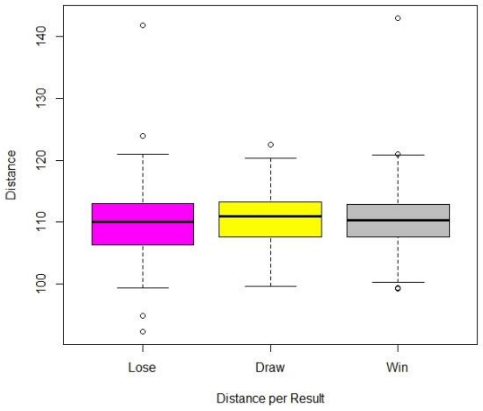

Offsides
0
2
4
6
8
10
Lose
Draw
Win
Offsides per Result

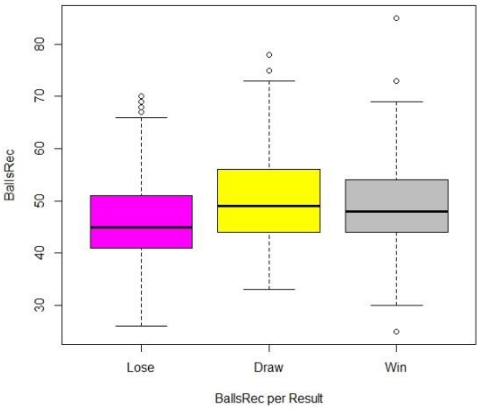

Possesion
30
40
50
60
70
Lose
Draw
Win
Possesion per Result

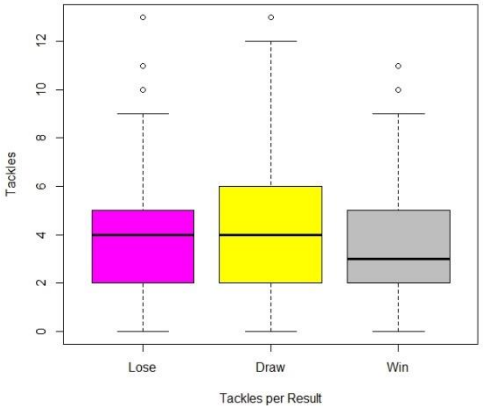

PassAcc
65
70
75
80
85
90
95
Lose
Draw
Win
PassAcc per Result

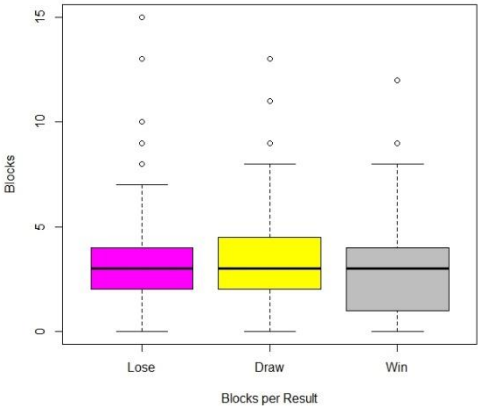

PassT
200
400
600
800
1000
Lose
Draw
Win
PassT per Result

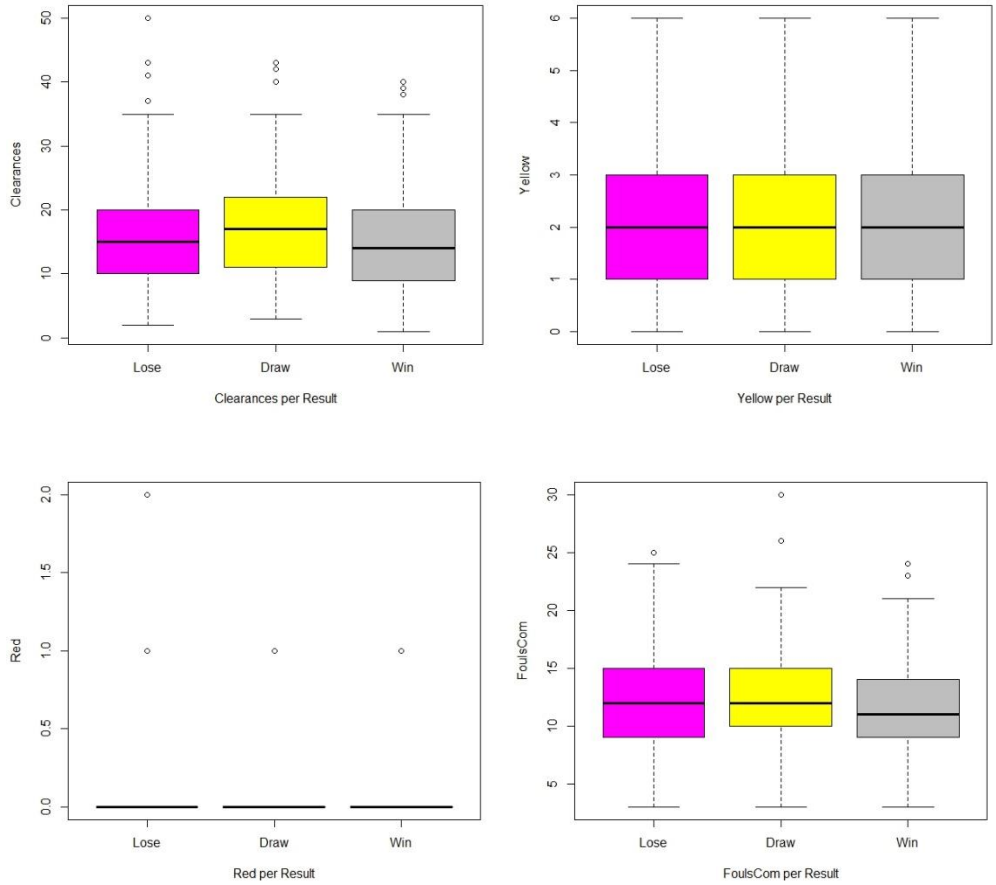

PassC
200
400
600
800
1000
Lose
Draw
Win
PassC per Result


Distance
100
110
120
130
140
Lose
Draw
Win
Distance per Result


BallsRec
30
40
50
60
70
80
Lose
Draw
Win
BallsRec per Result

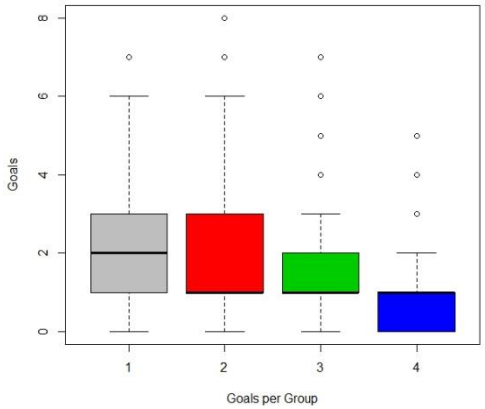

Tackles
0
2
4
6
8
10
12
Lose
Draw
Win
Tackles per Result

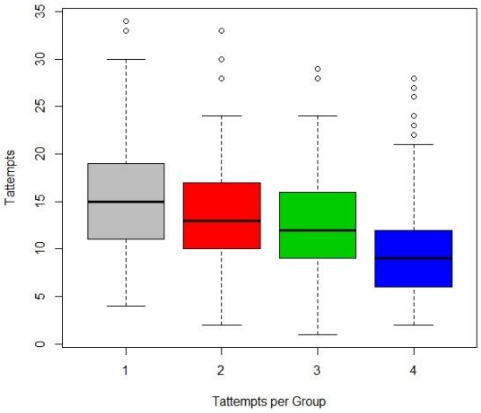

Blocks
0
5
10
15
Lose
Draw
Win
Blocks per Result

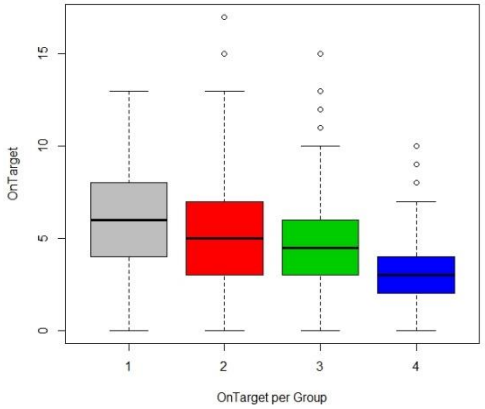


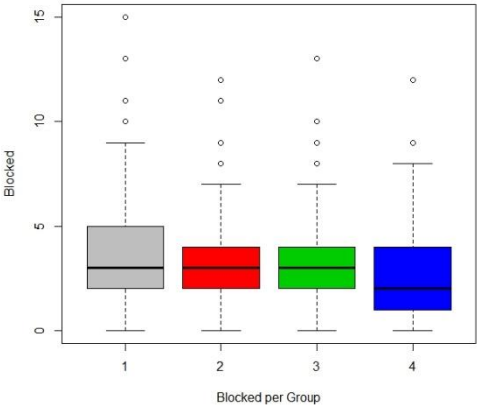
Τα ίδια αποτελέσματα αναφορικά με το επιθετικό κομμάτι των χαρακτηριστικών έχουμε αν δούμε τα θηκογράμματα αναφορικά με την επιρροή του γκρουπ δυναμικότητας της ομάδας. Βλέπουμε μια διαβάθμιση των διαμέσων ανάλογα με το πόσο ψηλά βρίσκεται η ομάδα στην ειδική βαθμολογία της UEFA. Οι διαφορές που φαίνονται μας δίνουν το δικαίωμα να πούμε ότι υπάρχει ένδειξη επιρροής.

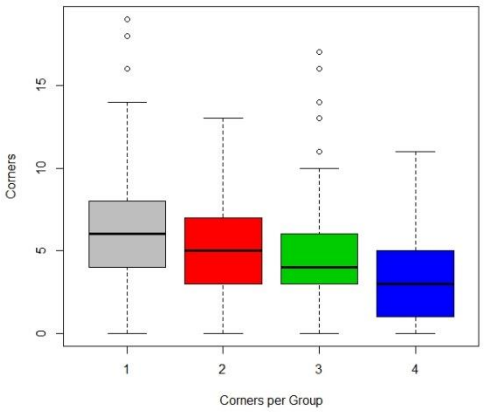
Αναφορικά με το αμυντικό κομμάτι, εδώ τα πράγματα είναι διαφορετικά σε σχέση με την μεταβλητή Result. Εδώ ναι μεν οι διαφορές που υπάρχουν δεν είναι οι ίδιες με το επιθετικό κομμάτι, αλλά η ανομοιομορφία στις διαφορές παρουσιάζεται μόνο στις BallsRec και Blocks. Στην πρώτη παρουσιάζεται διαφορά μόνο στο τρίτο γκρουπ δυναμικότητας σε σχέση με τα υπόλοιπα που είναι ισοδύναμα και στην Blocks διαφορά παρατηρείται μόνο στο πρώτο γκρουπ. Στην Tackles το πρώτο και το δεύτερο γκρουπ και αντίστοιχα το τρίτο και το τέταρτο είναι ισοδύναμα, με το πρώτο και το δεύτερο να έχουν μικρότερη διάμεσο. Στην Clearances φαίνεται ότι υπάρχει διαβάθμιση ανάλογα με το γκρουπ, με το πρώτο να έχει την μικρότερη διάμεσο και το τέταρτο τη μεγαλύτερη. Εν ολίγοις, στα δύο συγκεκριμένα χαρακτηριστικά υπάρχει ένδειξη επιρροής του χαρακτηριστικού Group.

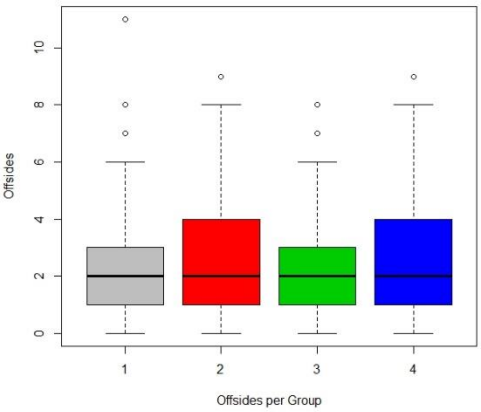

Goals
0
2
4
6
8
1
2
3
4
Goals per Group

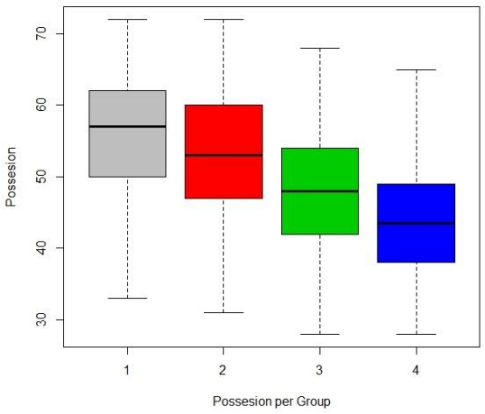

Tattempts
0
5
10
15
20
25
30
35
1
2
3
4
Tattempts per Group

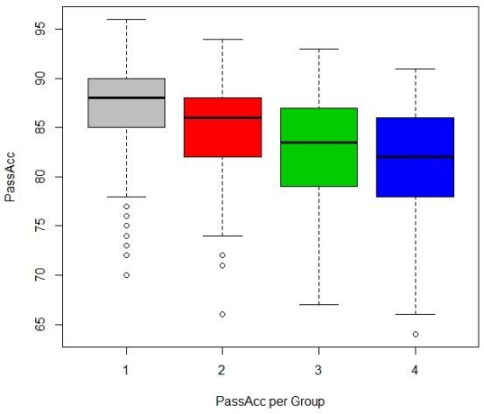

OnTarget
0
5
10
15
1
2
3
4
OnTarget per Group

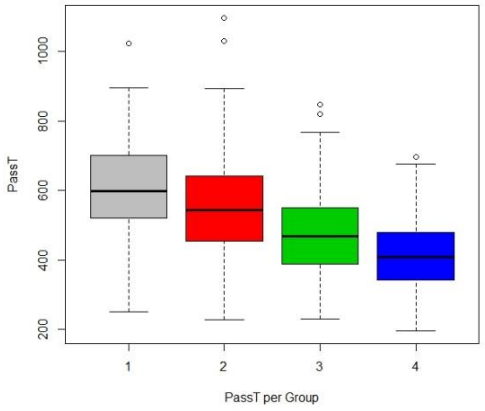

Blocked
0
5
10
15
1
2
3
4
Blocked per Group

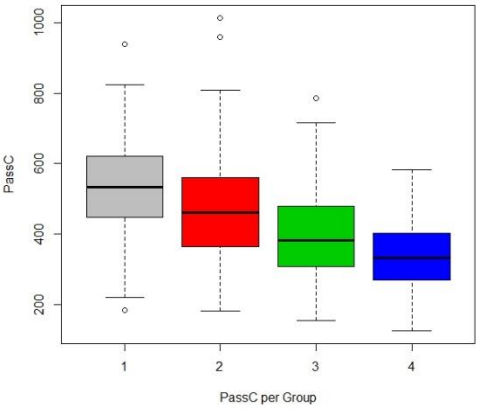

Corners
0
5
10
15
1
2
3
4
Corners per Group

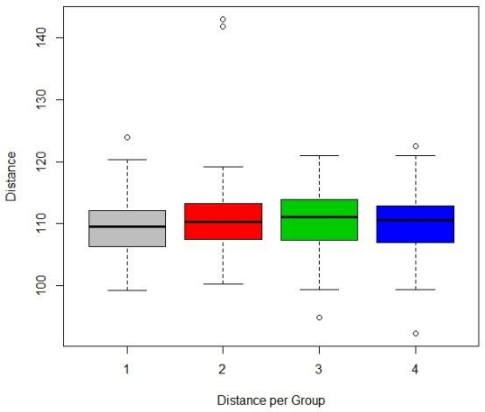

Offsides
0
2
4
6
8
10
1
2
3
4
Offsides per Group

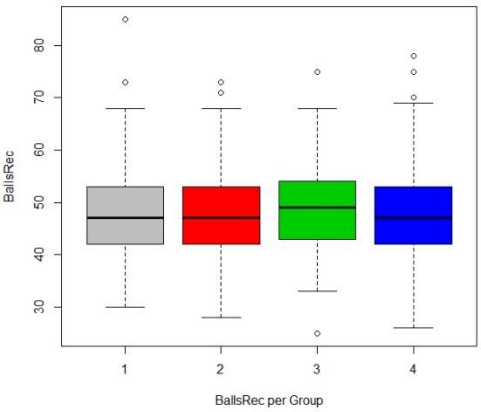

Possesion
30
40
50
60
70
1
2
3
4
Possesion per Group

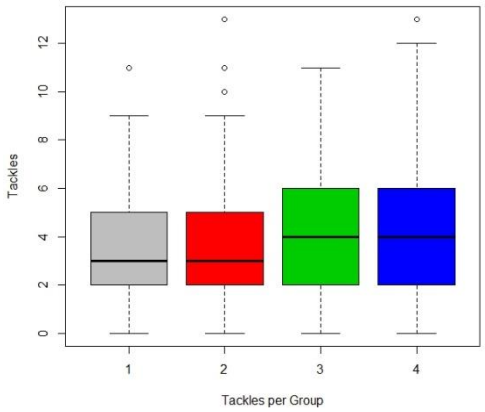

PassAcc
65
70
75
80
85
90
95
1
2
3
4
PassAcc per Group

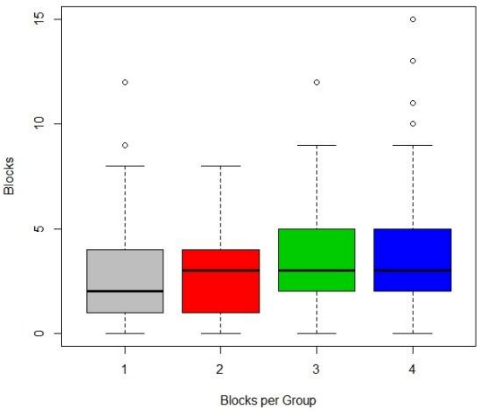

PassT
200
400
600
800
1000
1
2
3
4
PassT per Group

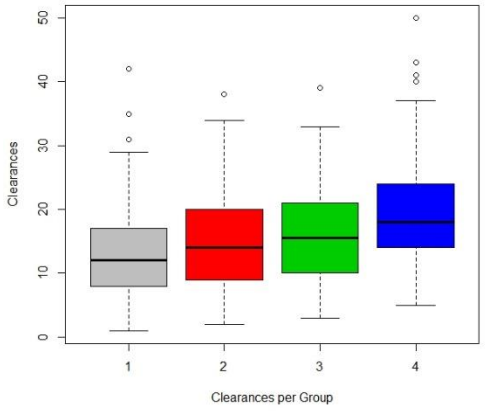

PassC
200
400
600
800
1000
1
2
3
4
PassC per Group

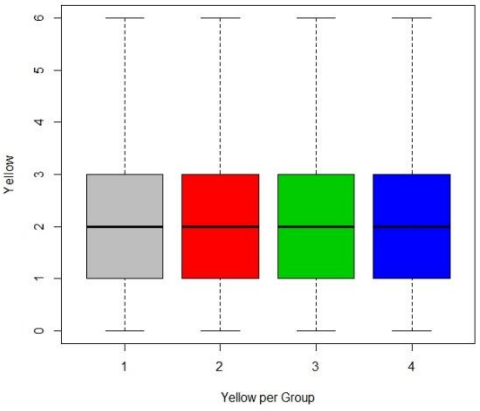

Distance
100
110
120
130
140
1
2
3
4
Distance per Group

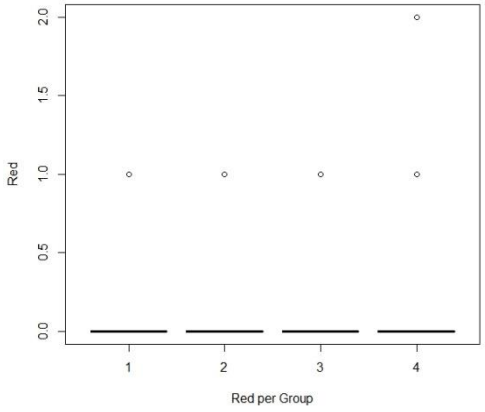

BallsRec
30
40
50
60
70
80
1
2
3
4
BallsRec per Group

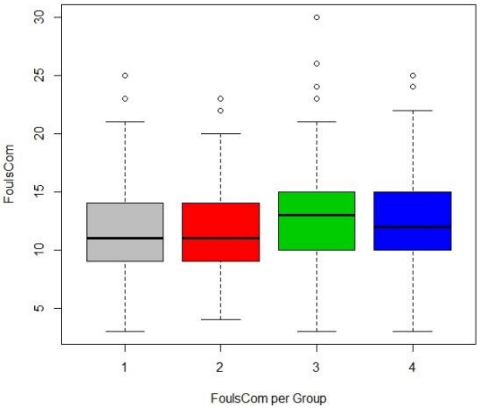

Tackles
0
2
4
6
8
10
12
1
2
3
4
Tackles per Group


Blocks
0
5
10
15
1
2
3
4
Blocks per Group


Clearances
0
10
20
30
40
50
1
2
3
4
Clearances per Group


Yellow
0
1
2
3
4
5
6
1
2
3
4
Yellow per Group


Red
0.0
0.5
1.0
1.5
2.0
1
2
3
4
Red per Group


FoulsCom
5
10
15
20
25
30
1
2
3
4
FoulsCom per Group

Αναφορικά με την υπόθεση που εξετάζει το κατά πόσο τα δεδομένα μας να ακολουθούν την κανονική κατανομή, αν λάβουμε υπόψη τα αρχικά θηκογράμματα η πρώτη ένδειξη είναι ότι στην συντριπτική πλειοψηφία των αποτελεσμάτων παραβιάζεται η συμμετρία. Τα μόνα χαρακτηριστικά που δείχνουν να έχουν συμμετρικά θηκογράμματα είναι αυτά που αφορούν τις BallsRec και FoulsCom. Οπότε, μια αρχική ένδειξη κανονικότητας μπορεί να υπάρξει μόνο για αυτά μέσω αυτής της μεθόδου. Για περαιτέρω διερεύνηση θα χρειαστεί να στραφούμε σε άλλες μεθόδους οπτικοποίησης για την εξέταση της υπόθεσης της κανονικότητας.

Λαμβάνοντας υπόψη όλες τις μεθόδους που πραγματοποιήθηκε αυτή η πρώτη και πολύ βασική ανάλυση των δεδομένων προκύπτουν κάποια βασικά συμπεράσματα:

- Για την μεταβλητή Home.Away, υπάρχει ένδειξη ότι επηρεάζει τις τιμές της πλειοψηφίας των μεταβλητών. Ωστόσο, δεν υπάρχει ομοιομορφία αναφορικά με την επιρροή. Φαίνεται ότι στις μεταβλητές που αφορούν την απόδοση της ομάδας στο επιθετικό σκέλος, τα πράγματα είναι καλύτερα για τις ομάδες που αγωνίζονται εντός έδρας, ενώ σε αυτές που αφορούν το αμυντικό σκέλος υπάρχουν αντιφάσεις.
- Η μεταβλητή Round δείχνει ότι δεν έχει πολύ σημαντική επιρροή στα δεδομένα. Οι μόνες περιπτώσεις που φαίνεται στα διαγράμματα κάποια διαφορά ανάμεσα στις δύο φάσεις της διοργάνωσης είναι στην Tattempts και στην PassAcc. Ωστόσο, επειδή η συγκεκριμένη ανάλυση είναι περιγραφική και δεν μπορεί να δώσει την ακρίβεια που δίνει ένα στατιστικό μοντέλο, θα εξεταστεί περαιτέρω στα επόμενα κεφάλαια.
- Αναμενόμενα αποτελέσματα υπήρξαν για τις μεταβλητές Group και Result. Στην πλειοψηφία των περιπτώσεων φαίνεται ότι στην Group έχουμε καλύτερες τιμές στο Group 1. Αντίστοιχα, ακολουθεί το δεύτερο Group, μετά το τρίτο και το τέταρτο. Στην περίπτωση της Result, προκύπτει ότι οι ηττημένοι έχουν τα χαμηλότερα στατιστικά και μετά οι ισόπαλοι, με τους νικητές να έχουν τα υψηλότερα. Αυτό σημαίνει ότι και στις δύο περιπτώσεις υπάρχουν ενδείξεις για διαφορές ανάμεσα στα επίπεδα, συνεπώς και για επιρροή των μεταβλητών αυτών στα δεδομένα.

ΚΕΦΑΛΑΙΟ 2

ΕΛΕΓΧΟΣ ΣΥΣΧΕΤΙΣΕΩΝ

Στην προσπάθεια να γίνει ξεκάθαρο ποιοι είναι οι παράγοντες που επηρεάζουν σε μεγάλο βαθμό την απόδοση, άρα κατά συνέπεια και την επιτυχία, των ομάδων σε μία τόσο μεγάλου βεληνεκούς ποδοσφαιρική διοργάνωση όπως είναι το UEFA Champions League και να καταρτιστεί το κατάλληλο στατιστικό μοντέλο, το οποίο θα δίνει ξεκάθαρες απαντήσεις και εν τέλει θα συμβάλλει στην προσπάθεια που κάνουν οι σύλλογοι και το τεχνικό τους επιτελείο να βελτιώσουν την αποδοτικότητα του συνόλου, γεννιέται η ανάγκη να διαχειριστούν έναν τεράστιο όγκο πληροφορίας.

Όπως προκύπτει από το προηγούμενο Κεφάλαιο, όπου παρουσιάστηκαν τα χαρακτηριστικά που περιέχονται στο αρχείο δεδομένων που απασχολεί την παρούσα μελέτη, εκτός από τις παρατηρήσεις που είναι πολλές σε πλήθος λόγω και της μορφής που έχουν τα δεδομένα μας, έχουμε ότι τα χαρακτηριστικά αυτά είναι εξίσου πολλά σε πλήθος. Αυτό είναι αρκετά σημαντικό να υπάρχει στο σύνολο των δεδομένων, έτσι ώστε να αξιοποιείται το σύνολο της πληροφορίας που έχει στα χέρια του ο ερευνητής, ωστόσο η χρήση του και εν τέλει η αξιοποίησή του στο τελικό μοντέλο δεν μπορεί να συμβάλλει αποτελεσματικά στην απάντηση των κύριων ερωτημάτων. Επίσης, λόγω της προσαρμογής του στο σύνολο των δεδομένων το μοντέλο ενδέχεται να γίνει αρκετά αναξιόπιστο στο να προβλέψει μελλοντικά αποτελέσματα σε ένα νέο σύνολο δεδομένων.

Εκτός από την εύρεση του κατάλληλου μοντέλου, το οποίο είναι κάτι που θα ερευνηθεί σε μελλοντικά κεφάλαια, υπάρχει και ένα πρότερο στάδιο στην ανάλυση των δεδομένων που βοηθάει στο να υπάρξει ένα πρώτο «ξεκαθάρισμα» μεταξύ των χαρακτηριστικών που εν τέλει θα αξιοποιηθούν για την προσαρμογή των μοντέλων. Αυτό γίνεται με τον έλεγχο συσχετίσεων. Η συσχέτιση αφορά μία ευρεία κατηγορία στατιστικών σχέσεων, ωστόσο στην πιο συνηθισμένη της μορφή αφορά την εξέταση ύπαρξης γραμμικής σχέσης ανάμεσα σε δύο μεταβλητές.

Στην παρούσα εργασία θα εξεταστούν οι συσχετίσεις μεταξύ των χαρακτηριστικών, έτσι ώστε να υπάρξει μια πρώτη εικόνα των χαρακτηριστικών εκείνων τα οποία είναι καθοριστικά για την απόδοση μιας ομάδας. Αυτή, προφανώς, δεν είναι μια προσπάθεια η οποία θεωρείται καθοριστική, καθώς δεν έχει την ακρίβεια που έχει η κατάρτιση ενός μοντέλου. Θα αξιοποιηθούν οι γνωστοί συντελεστές συσχέτισης, όπως και οι πίνακες συνάφειας, για τη μελέτη της σχέσης ανάμεσα στις κατηγορικές μεταβλητές.

2.1 Μεθοδολογία

Σε αυτή την ενότητα θα παρουσιάσουμε τους συντελεστές συσχέτισης που θα χρησιμοποιηθούν για αυτήν την αρχική μελέτη και του ελέγχου με βάση τους πίνακες συνάφειας.

2.1.1 Συντελεστής συσχέτισης του Pearson

Αποτελεί την περίπτωση συντελεστή συσχέτισης, η οποία είναι η πιο ευρεία διαδεδομένη. Στην πλειοψηφία των μελετών που συμπεριλαμβάνουν την έννοια της συσχέτισης και γενικότερα την αξιοποιούν σαν εργαλείο, αναφέρεται απλά ως «συντελεστής συσχέτισης». Εξετάζει την ύπαρξη γραμμικής σχέσης μεταξύ δύο μεταβλητών και συμβολίζεται με το ελληνικό γράμμα ρ.

Στην περίπτωση ενός πληθυσμού, ορίζεται ως το πηλίκο της συνδιακύμανσης και του γινομένου των διακυμάνσεων των δύο εξεταζόμενων μεταβλητών. Άρα, για δύο μεταβλητές X και Y με μέσες τιμές μ_X και μ_Y και διακυμάνσεις σ_X και σ_Y, αντίστοιχα, ο συντελεστής συσχέτισης του Pearson υπολογίζεται ως εξής:

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

Αντίστοιχα, ο δειγματικός συντελεστής συσχέτισης, ο οποίος συμβολίζεται με το γράμμα r, υπολογίζεται με βάση τις εκτιμήτριες των παραπάνω ποσοτήτων, με συνέπεια να έχουμε ότι

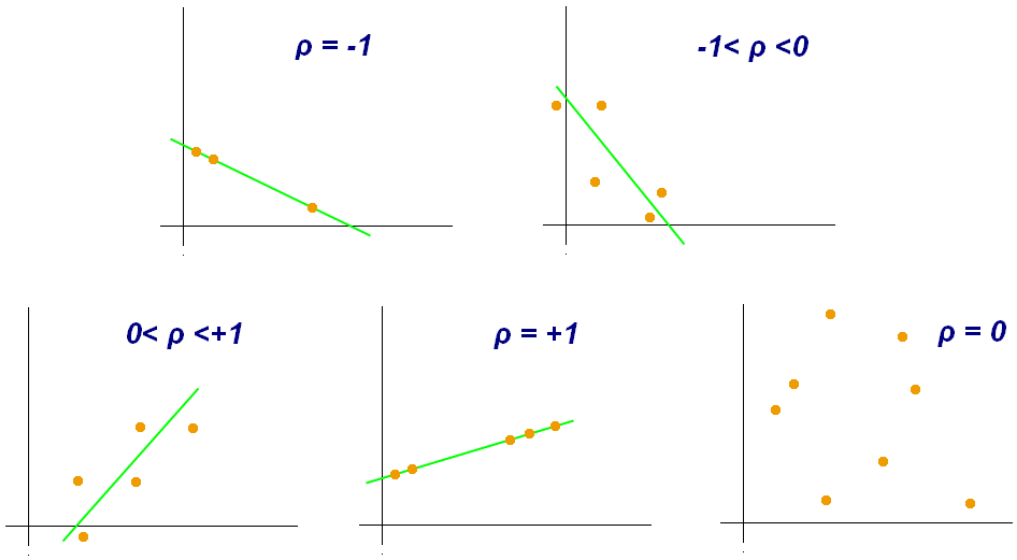
$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

όπου n είναι το μέγεθος του δείγματος, x_i, y_i είναι η i-οστή παρατήρηση του δείγματος και $\bar{x}, \bar{y}$ οι δειγματικοί μέσοι των δύο μεταβλητών.

Για τον δειγματικό συντελεστή συσχέτισης αυτό που γνωρίζουμε είναι ότι αποτελεί μία εκτίμηση του συντελεστή συσχέτισης. Ωστόσο, για να αποτελέσει μία πολύ καλή προσέγγισή του, θα πρέπει να ισχύει η βασική προϋπόθεση της κανονικότητας των δειγμάτων. Αυτό διότι στην περίπτωση που μιλάμε για κανονικά και μεγάλα σε μέγεθος δείγματα, ο δειγματικός συντελεστής συσχέτισης είναι ασυμπτωτικά αμερόληπτη και αποτελεσματική εκτιμήτρια του συντελεστή συσχέτισης, πράγμα το οποίο σημαίνει ότι είναι δύσκολο να βρεθεί μία ακριβέστερη εκτιμήτρια από αυτήν.

Γενικότερα, για τον συγκεκριμένο συντελεστή συσχέτισης, όπως και για τον δειγματικό, έχουμε ότι οι τιμές του είναι μεταξύ του -1 και του 1. Όσο η τιμή του πλησιάζει την τιμή 1, τόσο καλύτερη θετική συσχέτιση έχουν οι μεταβλητές X και Y. Αυτό σημαίνει ότι

μια αύξηση της Y σημαίνει αύξηση της X και το ανάποδο. Αντίστοιχα, όσο η τιμή του πλησιάζει την τιμή -1, τόσο καλύτερη αρνητική συσχέτιση έχουν οι μεταβλητές X και Y. Αυτό σημαίνει ότι μια αύξηση της Y σημαίνει μείωση της X και το ανάποδο. Στην περίπτωση της μηδενικής συσχέτισης έχουμε ότι οι δύο μεταβλητές δεν έχουν γραμμική σχέση. Παρακάτω φαίνονται συνοπτικά παραδείγματα της αποτύπωσης των παραπάνω περιπτώσεων στα διαγράμματα διασποράς.



(Πηγή: Wikipedia)

Συνεπώς, από τα παραπάνω προκύπτει ότι θα αξιοποιηθεί ο παραπάνω συντελεστής για την μελέτη της συσχέτισης μεταξύ των μεταβλητών των οποίων οι τιμές βρίσκονται σε ένα συνεχές διάστημα, αφού πρώτα γίνει έλεγχος κανονικότητας, χωρίς αυτό να αποτελεί την απαραίτητη προϋπόθεση.

2.1.2 Συντελεστής συσχέτισης του Spearman

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 βαθμολογία που παίρνουν οι τιμές των δύο μεταβλητών. Ουσιαστικά, δείχνει κατά πόσο μπορεί να περιγραφεί η σχέση που έχουν δύο μεταβλητές μέσω μίας μονότονης συνάρτησης.

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 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:

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

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:

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

Για τον συγκεκριμένο συντελεστή ισχύει, όπως και για τον συντελεστή συσχέτισης του Pearson, ότι λαμβάνει τιμές μεταξύ του -1 και του 1. Από την ερμηνεία που δόθηκε στην αρχή της παραγράφου, ισχύει ότι μία τιμή κοντά στο 1 σημαίνει ότι σχετίζονται άριστα μέσω μίας μονότονης συνάρτησης και μία αύξηση της τιμής του Y σημαίνει αύξηση της τιμής του X. Ανάλογα, μια τιμή κοντά στο -1 σημαίνει ότι σχετίζονται άριστα μέσω μίας μονότονης συνάρτησης και μία αύξηση της τιμής του Y σημαίνει μείωση της τιμής του X.

Τέλος, ο συντελεστής συσχέτισης του Spearman χρησιμοποιείται για να ελέγξει τη σχέση που έχουν δύο μεταβλητές και στην περίπτωση που μία εξ αυτών είναι διακριτή, ακόμα και στην περίπτωση των κατηγορικών μεταβλητών. Για το δείγμα που αφορά τη συγκεκριμένη μελέτη, θα χρησιμοποιηθεί για να ελεγχθεί η συσχέτιση ανάμεσα στις κατηγορικές και τις συνεχείς μεταβλητές.

2.1.3 Πίνακες συνάφειας

Οι περιπτώσεις των συντελεστών συσχέτισης που αναφέρθηκαν στις προηγούμενες δύο ενότητες αφορούσαν τις περιπτώσεις που εξετάζεται η σχέση μεταξύ δύο μεταβλητών όταν τουλάχιστον μία από τις δύο μεταβλητές είναι μία ποσοτική μεταβλητή. Στην περίπτωση που θα εξετάσουμε τώρα θα δούμε πως εξετάζεται η σχέση που έχουν δύο ποιοτικές (κατηγορικές) μεταβλητές, που ονομάζεται συνάφεια.

Βασικό συστατικό της συγκεκριμένης μεθόδου είναι ο πίνακας συνάφειας των δύο μεταβλητών. Ως πίνακα συνάφειας ορίζουμε τον πίνακα που συγκεντρώνει τις συχνότητες

ανάλογα με την τιμή που παίρνει η κάθε μεταβλητή. Οι γραμμές του αφορούν τις συχνότητες που εντοπίζονται οι τιμές (κατηγορίες) που παίρνει η μία από τις δύο ποιοτικές μεταβλητές, ενώ οι στήλες αφορούν αντίστοιχα τις τιμές για τη δεύτερη μεταβλητή. Η πιο απλή περίπτωση που εμφανίζεται είναι αυτή του 2X2 πίνακα, ωστόσο μπορεί να χρησιμοποιηθεί και σε μεγαλύτερες διαστάσεις, όπως φαίνεται παρακάτω:

		Columns (v_2)				
		1	2	…	c	Total
Rows (v_1)	1	n_{11}	n_{12}	…	n_{1c}	$n_{1.}$
	2	n_{21}	n_{22}	…	n_{2c}	$n_{2.}$
	⋮	⋮	⋮	⋮	⋮	⋮
	r	n_{r1}	n_{r2}	…	n_{rc}	$n_{r.}$
Total		$n_{.1}$	$n_{.2}$	…	$n_{.c}$	$n_{..} = N$

Πίνακας 2.1: Πίνακας συνάφειας (Πηγή: https://www.semanticscholar.org/)

Ο πίνακας συνάφειας αποτελεί την βάση πάνω στην οποία διεξάγεται ο έλεγχος υποθέσεων της ύπαρξης σχέσης ανάμεσα στις δύο μεταβλητές, με την διεξαγωγή του χ^2 ελέγχου ανεξαρτησίας. Ο συγκεκριμένος έλεγχος έχει τις εξής υποθέσεις:

$$H_0: \textit{οι μεταβλητές είναι ανεξάρτητες} - H_1: \textit{Οι μεταβλητές δεν είναι ανεξάρτητες}$$

Η στατιστική συνάρτηση του ελέγχου είναι η χ^2, η οποία είναι μια συνάρτηση που αφορά της συχνότητες του πίνακα συνάφειας υπολογίζεται ως εξής:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

όπου

$$\chi^2 = \eta\ \textit{στατιστική συνάρτηση ελέγχου}$$
$$O = \eta\ \textit{παρατηρούμενη συχνότητα κάθε κελιού}$$
$$E = \eta\ \textit{αναμενόμενη συχνότητα κάθε κελιού} = \frac{\textit{άθροισμα γραμμής} \cdot \textit{άθροισμα στήλης}}{\textit{αριθμός παρατηρήσεων}}$$

Η τιμή που παίρνουμε από την παραπάνω στατιστική συνάρτηση, η οποία έχει βαθμούς ελευθερίας ίσους με (r-1)(c-1), όπου r το πλήθος των τιμών της μεταβλητής που βρίσκεται στις γραμμές του πίνακα συνάφειας και c το πλήθος των τιμών της μεταβλητής που βρίσκεται στις στήλες του πίνακα συνάφειας, συγκρίνεται με την κρίσιμη τιμή $\chi^2_{(r-1)(c-1);a}$, όπου α το δεδομένο επίπεδο στατιστικής σημαντικότητας. Στην περίπτωση που $\chi^2 \geq \chi^2_{(r-1)(c-1);a}$, έχουμε στατιστικά σημαντικό αποτέλεσμα, δηλαδή υπάρχει ισχυρή ένδειξη υπέρ της H_1. Στην

συγκεκριμένη μελέτη, η παραπάνω διαδικασία θα χρησιμοποιηθεί για την μελέτη της ύπαρξης σχέσης ανάμεσα στις ποιοτικές μας μεταβλητές.

2.2 Έλεγχος κανονικότητας

Όπως φάνηκε και στην περιγραφή του συντελεστή συσχέτισης του Pearson, η κανονικότητα των δεδομένων αποτελεί έναν σημαντικό παράγοντα για να αποδειχθεί το κατά πόσο η εκτίμηση του πληθυσμιακού συντελεστή συσχέτισης μέσω του δειγματικού είναι αποτελεσματική. Για αυτόν τον λόγο, όπως και για μελλοντική χρήση στην περίπτωση εξέτασης του κατάλληλου στατιστικού μοντέλου για την περιγραφή των δεδομένων, σε αυτήν την ενότητα θα εξεταστεί η κανονικότητα των ποσοτικών μεταβλητών του δείγματος.

Με τον όρο κανονικότητα των δεδομένων εννοούμε την περίπτωση το δείγμα μας να προέρχεται από ένα σύνολο δεδομένων που ακολουθεί την κανονική κατανομή. Το παραπάνω εξακριβώνεται με τη διεξαγωγή του ελέγχου υποθέσεων που έχει σαν υποθέσεις τις παρακάτω:

$$H_0: Το\ δείγμα\ προέρχεται\ από\ την\ κανονική\ κατανομή$$
$$- H_1: Το\ δείγμα\ δεν\ προέρχεται\ από\ την\ κανονική\ κατανομή$$

Για τη διεξαγωγή του ελέγχου υπάρχουν αρκετές μέθοδοι. Ωστόσο, η καθεμία έχει και περιπτώσεις όπου παρουσιάζει αδυναμίες έναντι των άλλων. Για παράδειγμα, ένας από τους γνωστούς ελέγχους κανονικότητας είναι αυτός των Shapiro-Wilk, ο οποίος είναι αρκετά αποτελεσματικός καθώς επικεντρώνεται μόνο στην κανονική κατανομή, ωστόσο δεν συμπεριφέρεται το ίδιο καλά για δείγματα μεγαλύτερα των 50 παρατηρήσεων.

Σε τέτοιες περιπτώσεις προτείνεται ο έλεγχος των Kolmogorov-Smirnov, ο οποίος χρησιμοποιείται και για άλλες συνεχείς κατανομές. Παρόλα αυτά, για τον απλό έλεγχο χρειάζεται να είναι γνωστές οι παράμετροι της κατανομής για να προχωρήσουμε στη διαδικασία. Σε αντίθετη περίπτωση, χρησιμοποιείται ο έλεγχος του Lilliefors, όπου αποτελεί μία παραλλαγή του Kolmogorov-Smirnov. Αυτός ο έλεγχος θα χρησιμοποιηθεί και στο συγκεκριμένο δείγμα.

Διεξάγοντας τον συγκεκριμένο έλεγχο με χρήση της R, φάνηκε ότι σε επίπεδο στατιστικής σημαντικότητας 5% ότι υπάρχει στατιστικά σημαντικό αποτέλεσμα, αφού για όλες τις ποσοτικές μεταβλητές το p-value ήταν κατά πολύ μικρότερο του 0.05. Αυτό αποτελεί μια ισχυρή ένδειξη ότι τα δεδομένα μας δεν προέρχονται από την κανονική κατανομή. Για κάποιες από τις δεδομένες μεταβλητές ήταν αναμενόμενο, καθώς έχουν διακριτές τιμές, οπότε δεν κρίνεται αναγκαίο να διατυπωθεί κάποιο αποτέλεσμα. Αυτές είναι οι Goals, GoalsCon, Yellow, Red. Στη συνέχεια, έχουμε ότι για τις μεταβλητές OnTarget, Blocked, Corners, Offsides, PassAcc, Tackles, Blocks το p-value είναι μικρότερο του $2.2X10^{-16}$. Τέλος, ο παρακάτω πίνακας δείχνει τα αποτελέσματα του ελέγχου στις υπόλοιπες μεταβλητές:

Μεταβλητή	p-value
Tattempts	$1.362X10^{-13}$
Possession	0.0002656
PassT	0.0009907
PassC	$9.003X10^{-5}$
Distance	0.02418
BallsRec	$7.435X10^{-7}$
Clearances	$8.584X10^{-10}$
FoulsCom	$3.423X10^{-10}$

Πίνακας 2.2: Αποτελέσματα ελέγχων κανονικότητας

2.3 Αποτελέσματα συσχετίσεων

Στην ενότητα αυτή θα παρουσιαστούν τα ευρήματα που μας έδωσε η χρήση των 3 περιπτώσεων ελέγχων που παρουσιάστηκαν στην ενότητα 2.1. Συνεπώς, διακρίνουμε τρεις διαφορετικές περιπτώσεις, αντίστοιχα.

2.3.1 Συσχετίσεις μεταξύ ποσοτικών μεταβλητών

Εδώ έχουμε τα αποτελέσματα που μας έδωσε ο έλεγχος συσχέτισης με τον συντελεστή του Pearson. Στην συγκεκριμένη μελέτη δεν συμπεριλήφθηκαν οι μεταβλητές Yellow και Red, παρόλο που συγκαταλέγονται στις ποσοτικές μεταβλητές καθώς αυτές είχαν αρκετά μικρό εύρος τιμών, οπότε κρίθηκε σκόπιμο να εξεταστούν με τον συντελεστή συσχέτισης του Spearman. Το ίδιο θα έπρεπε να ισχύει και για την μεταβλητή Goals, ωστόσο επειδή υπάρχει τέτοια δυνατότητα μέσω των δύο συντελεστών, θα εξεταστεί η συσχέτισή της και στις δύο περιπτώσεις. Ομοίως και για την GoalsCon. Ο λόγος είναι ότι αποτελεί μία από τις βασικές μεταβλητές που μπορεί να βοηθήσει στην εξέταση της σημασίας των υπόλοιπων χαρακτηριστικών, οπότε είναι χρήσιμο να αντληθεί οποιαδήποτε πληροφορία για τη συμπεριφορά της. Αντίστοιχα, η GoalsCon μπορεί να δώσει συμπεράσματα, κυρίως στο αμυντικό σκέλος. Παρακάτω φαίνεται ο πίνακας συσχετίσεων για τις ποσοτικές μεταβλητές:

	Goals	Tattempts	OnTarget	Blocked	Corners	Offsides	Possesion	PassAcc	PassT
Goals	1	0.38774	0.65509	0.07052	0.12610	0.14061	0.28276	0.26472	0.29330
Tattempts	0.38774	1	0.74272	0.69163	0.59797	0.08422	0.54351	0.42766	0.48261
OnTarget	0.65509	0.74272	1	0.27563	0.41869	0.10341	0.42872	0.38042	0.41321
Blocked	0.07052	0.69163	0.27563	1	0.50732	0.04265	0.36789	0.25609	0.30235
Corners	0.12610	0.59797	0.41869	0.50732	1	0.08375	0.49835	0.33569	0.41350
Offsides	0.14061	0.08422	0.10341	0.04265	0.08375	1	0.12404	0.12770	0.09925
Possesion	0.28276	0.54351	0.42872	0.36789	0.49835	0.12404	1	0.69611	0.89302
PassAcc	0.26472	0.42766	0.38042	0.25609	0.33569	0.12770	0.69611	1	0.77922
PassT	0.29330	0.48261	0.41321	0.30235	0.41350	0.09925	0.89302	0.77922	1
PassC	0.30659	0.48412	0.42334	0.29720	0.41083	0.10083	0.88097	0.83038	0.99321
Distance	0.03051	0.07146	0.04105	0.02319	-0.00995	0.06008	-0.11358	-0.05285	0.01175
BallsRec	0.04940	0.06285	0.05652	0.03280	0.00518	0.12670	0.05846	-0.20326	0.05457
Tackles	-0.06136	-0.09279	-0.06592	-0.07059	-0.07529	0.11544	-0.17919	-0.12663	-0.14545
Blocks	-0.11023	-0.28436	-0.21131	-0.22891	-0.22937	-0.10855	-0.37010	-0.29014	-0.37201
Clearances	-0.15133	-0.42745	-0.32818	-0.28519	-0.37973	-0.13968	-0.53848	-0.51117	-0.57119
FoulsCom	-0.09826	-0.09175	-0.11824	-0.02792	-0.07179	-0.03138	-0.16933	-0.22791	-0.28107
GoalsCon	-0.26299	-0.26429	-0.26112	-0.10687	-0.17987	-0.03747	-0.28304	-0.12983	-0.24309

	PassC	Distance	BallsRec	Tackles	Blocks	Clearances	FoulsCom	GoalsCon
Goals	0.30659	0.03051	0.04940	-0.06136	-0.11023	-0.15133	-0.09826	-0.26299
Tattempts	0.48412	0.07146	0.06285	-0.09279	-0.28436	-0.42745	-0.09175	-0.26429
OnTarget	0.42334	0.04105	0.05652	-0.06592	-0.21131	-0.32818	-0.11824	-0.26112
Blocked	0.29720	0.02319	0.03280	-0.07059	-0.22891	-0.28519	-0.02792	-0.10687
Corners	0.41083	-0.00995	0.00518	-0.07529	-0.22937	-0.37973	-0.07179	-0.17987
Offsides	0.10083	0.06008	0.12670	0.11544	-0.10855	-0.13968	-0.03138	-0.03747
Possesion	0.88097	-0.11358	0.05846	-0.17919	-0.37010	-0.53848	-0.16933	-0.28304
PassAcc	0.83038	-0.05285	-0.20326	-0.12663	-0.29014	-0.51117	-0.22791	-0.12983
PassT	0.99321	0.01175	0.05457	-0.14545	-0.37201	-0.57119	-0.28107	-0.24309
PassC	1	-0.00314	0.00099	-0.14902	-0.36350	-0.57268	-0.28374	-0.23069
Distance	-0.00314	1	0.21298	0.02010	0.06606	0.13109	-0.07444	-0.07558
BallsRec	0.00099	0.21298	1	0.04860	-0.05795	0.05275	-0.01684	-0.20244
Tackles	-0.14902	0.02010	0.04860	1	0.04965	0.09615	-0.04255	0.06942
Blocks	-0.36350	0.06606	-0.05795	0.04965	1	0.41734	-0.01324	0.07682
Clearances	-0.57268	0.13109	0.05275	0.09615	0.41734	1	0.03547	-0.01159
FoulsCom	-0.28374	-0.07444	-0.01684	-0.04255	-0.01324	0.03547	1	-0.04817
GoalsCon	-0.23069	-0.07558	-0.20244	0.06942	0.07682	-0.01159	-0.04817	1

Πίνακας 2.3: Πίνακας συσχέτισης με βάση τον συντελεστή του Pearson.

Αρχικά, πριν γίνει η ανάλυση των αποτελεσμάτων, πρέπει να διασαφηνιστεί η ερμηνεία του συντελεστή συσχέτισης. Δηλαδή, να οριστούν τα όρια εκείνα των τιμών της συσχέτισης, τα οποία θεωρείται ότι δύο μεταβλητές είναι ισχυρά, μέτρια ή ασθενώς συσχετισμένες. Οι τιμές αυτές δεν αλλάζουν σε περίπτωση θετικής ή αρνητικής συσχέτισης, βασίζονται με λίγα λόγια στην απόλυτη τιμή του συντελεστή. Έτσι, ορίζουμε ότι:

Τιμή συντελεστή	Ερμηνεία συσχέτισης
$\|r\|\leq 0.3$	Ασθενής
$0.3<\|r\|\leq 0.5$	Μέτρια
$\|r\|\geq 0.5$	Ισχυρή

Πίνακας 2.4: Ερμηνεία συντελεστή συσχέτισης (Πηγή: statistics.laerd.com)

Έτσι αν ο συντελεστής συσχέτισης είναι π.χ. r=-0.20, τότε έχουμε ότι οι δύο μεταβλητές έχουν ασθενή αρνητική συσχέτιση.

Με βάση την παραπάνω ερμηνεία προέκυψαν τα εξής αποτελέσματα:

1. Η μεταβλητή Goals, που όπως είναι φυσιολογικό αποτελεί κρίσιμο παράγοντα για την επιτυχία μιας ομάδας, αφού αφορά το σκοράρισμα, έχει ισχυρή θετική συσχέτιση με την OnTarget (0.655), ενώ ταυτόχρονα έχει μέτρια συσχέτιση με την Tattempts. Δηλαδή, μια αύξηση στην τιμή της OnTarget αυξάνει γραμμικά την τιμή της Goals.

2. Ωστόσο, η Tattempts σχετίζεται ισχυρά με τις μεταβλητές OnTarget, Blocked, Corners και Possession. Σε όλες τις περιπτώσεις η συσχέτιση είναι θετική. Επίσης, ισχυρή θετική συσχέτιση παρατηρείται μεταξύ της Blocked και της Corners. Για την τελευταία παρατηρήθηκε μια θετική συσχέτιση με την Possession πολύ κοντά στο όριο που έχει τεθεί για την ισχυρή, ωστόσο συνεχίζει να θεωρείται μέτρια (0.498).

3. Αξιοσημείωτα αποτελέσματα προέκυψαν αναφορικά με τις μεταβλητές που αφορούν το κομμάτι της δημιουργίας μιας ομάδας (Possession, PassAcc, PassT, PassC). Τα αποτελέσματα έδειξαν ισχυρές θετικές συσχετίσεις ανάμεσα σε αυτές τις 4 στις μεταξύ τους συγκρίσεις, σε σημείο σχεδόν απόλυτης ταύτισης, όπως στην περίπτωση των PassT και PassC.

4. Με τις παραπάνω μεταβλητές παρουσίασε ισχυρές αρνητικές συσχετίσεις η μεταβλητή Clearances. Αυτό είναι ουσιαστικά και το μόνο αξιόλογο εύρημα που εντοπίστηκε για τις μεταβλητές που αφορούν το αμυντικό σκέλος (Balls Rec, Tackles, Clearances, Blocks, Fouls Com).

5. Τέλος, να συμπληρωθεί ότι μέτριες και κυρίως ασθενείς συσχετίσεις είχαν οι μεταβλητές GoalsCon, Distance και Offsides. Σε γενικές γραμμές, κάποιες από τις μεταβλητές που δεν έδωσαν κάποιο σημαντικό αποτέλεσμα με χρήση του συντελεστή συσχέτισης του Pearson ή του Spearman, θα εξεταστούν και με τη χρήση του πίνακα συνάφειας, καθώς αποδεδειγμένα είναι παράγοντες που έχουν δείξει σημαντική συνεισφορά, όπως η BallsRec.

2.3.2 Συσχετίσεις μεταξύ ποσοτικών και ποιοτικών μεταβλητών

Στην ενότητα αυτή θα παρουσιαστούν τα αποτελέσματα που προέκυψαν από τη χρήση του συντελεστή συσχέτισης του Spearman. Οι κανόνες κάτω από τους οποίους ερμηνεύονται τα κάθε φορά αποτελέσματα δεν διαφέρουν ανά συντελεστή, συνεπώς θα ερμηνευθούν με τον ίδιο τρόπο και σε αυτήν την περίπτωση. Παρακάτω παρουσιάζονται τα αποτελέσματα των συσχετίσεων.

	Yellow	Red	Home.Away	Round
Goals	-0.14275	-0.061748	0.16401	0.03686
Tattempts	-0.20100	-0.074974	0.27275	0.00014
OnTarget	-0.19503	-0.083412	0.20305	-0.01662
Blocked	-0.09170	-0.002720	0.19949	0.00141
Corners	-0.09320	-0.012047	0.18421	0.02062
Offsides	-0.01307	-0.102627	0.06035	-0.04734
Possesion	-0.13721	-0.055162	0.08956	0.00018
PassAcc	-0.18348	-0.018563	0.04620	-0.04975
PassT	-0.20806	-0.061236	0.06769	-0.01943
PassC	-0.20709	-0.054626	0.06123	-0.02624
Distance	-0.16595	-0.149675	0.04919	-0.03204
BallsRec	-0.03385	-0.092670	0.04564	-0.01172
Tackles	0.07100	-0.043972	0.09933	-0.03940
Blocks	0.11429	0.040473	-0.20379	0.00390
Clearances	0.12178	-0.030417	-0.21337	0.03103
FoulsCom	0.37333	0.059544	-0.06512	-0.02903

	Ranking	Group	GoalsCon	Result
Goals	-0.30158	-0.27848	-0.26118	0.71924
Tattempts	-0.36622	-0.35450	-0.26182	0.37797
OnTarget	-0.36066	-0.34115	-0.25918	0.51241
Blocked	-0.17970	-0.16910	-0.09203	0.08210
Corners	-0.33869	-0.32987	-0.15420	0.16761
Offsides	-0.02867	-0.00330	-0.03480	0.10925
Possesion	-0.48149	-0.46298	-0.25839	0.28569
PassAcc	-0.46061	-0.42836	-0.16456	0.23579
PassT	-0.52090	-0.49376	-0.22173	0.26407

PassC	-0.52305	-0.49330	-0.21030	0.26294
Distance	0.06593	0.06608	-0.03709	0.04794
BallsRec	0.01857	0.01580	-0.19058	0.15074
Tackles	0.08320	0.06885	0.05155	-0.04334
Blocks	0.18498	0.18627	0.07664	-0.08419
Clearances	0.31236	0.28580	0.00197	-0.01853
FoulsCom	0.13135	0.13224	-0.00598	-0.08250

Πίνακας 2.5: Πίνακας συσχέτισης με βάση τον συντελεστή του Spearman.

Αν και φαίνεται ότι σε γενικές γραμμές δεν προέκυψαν αρκετές ισχυρές συσχετίσεις ανάμεσα στις ποσοτικές και τις ποιοτικές μεταβλητές του δείγματος, προκύπτουν τρία πολύ σημαντικά αποτελέσματα:

1. Φαίνεται ότι υπάρχει ισχυρή αρνητική συσχέτιση ανάμεσα στις 4 μεταβλητές που αφορούν το δημιουργικό σκέλος του παιχνιδιού και την δυναμικότητα της ομάδας. Αυτό δεν εκφράζεται εξ ολοκλήρου με ισχυρές συσχετίσεις ανάμεσα στις μεταβλητές αυτές και την μεταβλητή Group ή την μεταβλητή Ranking, μέσω της οποίας προέκυψε και η μεταβλητή Group. Ωστόσο, βλέπουμε ότι υπάρχει ισχυρή αρνητική συσχέτιση ανάμεσα στην Ranking και τις PassT (-0.520) και PassC (-0.523), η οποία οριακά είναι κάτω από το όριο του -0.50 όταν πηγαίνουμε στην Group (-0.493 και -0.493, αντίστοιχα), κάτι που ενδεχομένως να ευθύνεται το γεγονός ότι η Group είναι μια ξεκάθαρα κατηγορική μεταβλητή. Επιπλέον, μέτρια συσχέτιση υπάρχει μεταξύ των δύο μεταβλητών και των μεταβλητών Possession και PassAcc. Για να προκύψουν ακριβέστερα συμπεράσματα και να μπορέσει να αποτελέσει αυτό ένα στοιχείο στα ζητούμενα που θέτει η μελέτη, τα παραπάνω εξετάζονται και με τους πίνακες συνάφειας.

2. Η μεταβλητή Result έχει ισχυρή θετική συσχέτιση με την Goals (0.719) και με την OnTarget (0.512). Αυτό δείχνει ότι η αύξηση των γκολ και των ευκαιριών στο στόχο βελτιώνει μέσω μονότονης συνάρτησης και το αποτέλεσμα του αγώνα που έχει μια ομάδα.

3. Για τις μεταβλητές Yellow, Red, Round και GoalsCon δεν προέκυψε κάποια ισχυρή συσχέτιση. Αν αυτό συνδυαστεί και με τα αποτελέσματα του Κεφαλαίου 1 για τις πρώτες τρεις και με του συντελεστή συσχέτισης του Pearson για την τελευταία, ίσως εδώ να έχουμε μια ένδειξη ότι σε γενικές γραμμές οι 4 αυτές μεταβλητές δεν αποτελούν σημαντικούς παράγοντες για την βελτίωση της απόδοσης των συλλόγων.

2.3.3 Συσχετίσεις μεταξύ ποσοτικών μεταβλητών

Σε αυτή την ενότητα θα εξεταστούν οι συσχετίσεις που αφορούν αποκλειστικά τις ποσοτικές μεταβλητές, πράγμα το οποίο ελέγχθηκε με τη μέθοδο του πίνακα συνάφειας. Να τονιστεί ότι, όπως αναφέρθηκε παραπάνω, θα εξεταστούν και συσχετίσεις ανάμεσα σε ποιοτικές και κάποιες ποσοτικές μεταβλητές, όπου η συγκεκριμένη μελέτη, αλλά και προηγούμενες (Barreira et al., 2014) έχουν αναδείξει τη σημασία τους, ωστόσο τα αποτελέσματα με τους δύο προηγούμενους συντελεστές δεν την επιβεβαίωσαν. Αυτό θα πραγματοποιηθεί με το να γίνει διακριτοποίηση αυτών των μεταβλητών, δηλαδή να την μετατρέψουμε σε μία κατηγορική μεταβλητή. Για να γίνει πράξη θα χρειαστεί να χωρίσουμε το δείγμα σε κατηγορίες, οι οποίες θα βαζίζονται στα τεταρτημόρια που υπολογίστηκαν στο Κεφάλαιο 1. Οι έλεγχοι διεξάγονται σε επίπεδο στατιστικής σημαντικότητας 5%.

Αρχικά, παρατηρείται μια σημαντική συσχέτιση μεταξύ της μεταβλητής Group και της μεταβλητής Round. Υπενθυμίζουμε ότι η μεταβλητή Round περιγράφει τη φάση στην οποία βρισκόταν η κάθε ομάδα-παρατήρηση, δηλαδή αν ήταν στη φάση των ομίλων ή στις νοκ-αουτ φάσεις. Φάνηκε ότι οι δύο μεταβλητές έχουν σχέση μεταξύ τους και τα δεδομένα του πίνακα συνάφειας δείχνουν το προφανές, ότι δηλαδή οι ομάδες που βρίσκονται στις υψηλότερες θέσεις στην βαθμολογία της UEFA είναι και αυτές που στην πλειονότητά τους προκρίνονται στις επόμενες φάσεις. Αντίστοιχα, οι ομάδες που βρίσκονται στις τελευταίες θέσεις έχουν στην συντριπτική τους πλειοψηφία προδιαγεγραμμένη μοίρα εκτός των επόμενων φάσεων των ομίλων, κάτι το οποίο, ωστόσο, δεν είναι σε απόλυτο βαθμό. Το p-value του αντίστοιχου χ^2 ελέγχου είναι κατά πολύ μικρότερο του 5%, συνεπώς έχουμε στατιστικά σημαντικό αποτέλεσμα.

Άμεση συνέπεια του προηγούμενου αποτελέσματος ήταν ο έλεγχος της ύπαρξης σχέσης μεταξύ της μεταβλητής Group και της Result. Ο λόγος είναι ότι εφόσον η τάση να προκρίνονται στις επόμενες φάσεις της διοργάνωσης οι ομάδες που βρίσκονται στα υψηλά κλιμάκια του Ranking, θα πρέπει να υπάρχει και ένα αντίκτυπο στα αποτελέσματα που έχουν οι ομάδες ανά Group.

Τα αποτελέσματα του ελέγχου χ^2 έδειξαν ότι υπάρχει στατιστικά σημαντική σχέση μεταξύ του Group και της μεταβλητής Result. Συγκεκριμένα, οι ομάδες που βρίσκονται υψηλότερα στο Ranking της UEFA δείχνουν ότι έχουν τη μερίδα του λέοντος στις νίκες, ενώ αντίθετα οι ομάδες που βρίσκονται στο τελευταίο Group δυναμικότητας είναι αυτές που δυσκολεύονται περισσότερο στο αποτέλεσμα. Το p-value του ελέγχου είναι κατά πολύ μικρότερο του 0.05.

Επόμενο πολύ σημαντικό αποτέλεσμα που προέκυψε ήταν αυτό κατά τον έλεγχο ύπαρξης σχέσης μεταξύ της Result και της Home.Away που δηλώνει αν η ομάδα αγωνίστηκε

εκτός ή εντός έδρας. Τα αποτελέσματα του ελέγχου χ^2 μας έδωσαν ένα p-value ίσο με 0.0002194, πολύ μικρότερο του 0.05. Από αυτό , όπως και από τον πίνακα συνάφειας φαίνεται περίτρανα ότι ο παράγοντας έδρα είναι στατιστικά σημαντικός στην έκβαση του αποτελέσματος.

Στη συνέχεια η μελέτη περιλαμβάνει τα αποτελέσματα που προέκυψαν από την εξέταση των μεταβλητών μέσω της διαδικασίας της διακριτοποίησης. Η αρχή γίνεται με την μεταβλητή BallsRec. Σε προηγούμενα σημεία της μελέτης έχει γίνει αιτιολόγηση αυτής της επιλογής, ωστόσο είναι αρκετά αξιοσημείωτη η μη εμφάνιση συσχέτισης με βάση τα αποτελέσματα των δύο συντελεστών συσχέτισης. Ωστόσο, αυτό διαψεύδεται με θεαματικό τρόπο χρησιμοποιώντας τον πίνακα συνάφειας. Ο έλεγχος χ^2 που διεξήχθη μεταξύ της Result και της διακριτοποιημένης BallsRec (BallsRec2), η οποία έχει 3 κατηγορίες με την πρώτη να αφορά το πρώτο τεταρτημόριο (με BallsRec μικρότερο ή ίσο του 42), τη δεύτερη το δεύτερο και τρίτο τεταρτημόριο (μεταξύ 43 και 53) και την τρίτη να αφορά τις τιμές που είναι μεγαλύτερες του 53, έδειξε ότι υπάρχει στατιστικά σημαντική σχέση μεταξύ των δύο αυτών μεταβλητών, με το p-value να είναι κατά πολύ μικρότερο του 0.05.

Εν συνεχεία, εξετάζονται οι περιπτώσεις που προέκυψαν από τις συσχετίσεις με βάση τον συντελεστή συσχέτισης του Spearman. Υπενθυμίζεται ότι θα γίνει έλεγχος ανάμεσα στις διακριτοποιημένες μεταβλητές που αφορούν το δημιουργικό σκέλος του παιχνιδιού και τις μεταβλητές Group και Result, καθώς τα αποτελέσματα έδειξαν μέτριες κατά κύριο λόγο συσχετίσεις, όπως και μια ανομοιομορφία στα αποτελέσματα μεταξύ αυτών των μεταβλητών όταν συγκρίθηκαν με την Ranking, κάτι το οποίο δεν θα έπρεπε να υφίσταται.

Με βάση λοιπόν αυτά οι μεταβλητές Possesion, PassAcc, PassT και PassC χωρίστηκαν σε 4 κατηγορίες, ανάλογα με τις τιμές των τεταρτημόριων τους. Από αυτά προέκυψαν στατιστικά σημαντικά αποτελέσματα σε όλες τις περιπτώσεις, με τις τιμές των p-values να είναι κατά πολύ μικρότερες 0.05, πράγμα που σημαίνει ότι έχουμε σημαντική σχέση ανάμεσα στις 4 αυτές μεταβλητές και τις Group και Result.

Κλείνοντας, παρατίθεται ο πίνακας με τις τιμές των p-values των παραπάνω ελέγχων. Οι πίνακες συνάφειας, όπως και τα ακριβή αποτελέσματα των ελέγχων με χρήση της R περιέχονται στο Παράρτημα Β.

Ελεγχόμενες μεταβλητές	**p-value**
Group, Round	$2.481 \cdot 10^{-16}$
Group, Result	$< 2.2 \cdot 10^{-16}$
Home.Away, Result	0.0002194
BallsRec2, Result	$6.795 \cdot 10^{-6}$
Possesion2, Group	$< 2.2 \cdot 10^{-16}$
Possesion2, Result	$9.012 \cdot 10^{-13}$
PassAcc2, Group	$< 2.2 \cdot 10^{-16}$

PassAcc2, Result	$6.665 \cdot 10^{-14}$
PassT2, Group	$<2.2 \cdot 10^{-16}$
PassT2, Result	$5.564 \cdot 10^{-13}$
PassAcc2, Group	$<2.2 \cdot 10^{-16}$
PassAcc2, Result	$1.161 \cdot 10^{-12}$

Πίνακας 2.6: Αποτελέσματα ελέγχων χ^2 για τους πίνακες συνάφειας.

2.4 ΣΥΜΠΕΡΑΣΜΑΤΑ

Συνδυάζοντας τα αποτελέσματα που προέκυψαν από την ανάλυση του συγκεκριμένου Κεφαλαίου, όπως ειπώθηκε εισαγωγικά, δεν είναι εφικτό ο αναλυτής των δεδομένων αυτών να βγάλει ένα τελικό συμπέρασμα αναφορικά με τα βασικά ερωτήματα της μελέτης, ωστόσο τα αποτελέσματα αυτά δίνουν κάποιες πρώτες απαντήσεις και βοηθούν καταλυτικά στην καλύτερη κατάρτιση του μοντέλου που θα δώσει την πλήρη εικόνα και το οποίο αξιοποιώντας οποιονδήποτε τρόπο θα φανεί πιο αποτελεσματικό.

Αρχικά, ένα πρώτο συμπέρασμα που προκύπτει από τις συσχετίσεις μιας από τις βασικές μεταβλητές που αποτελεί ένδειξη της αποτελεσματικότητας και της επιτυχίας μίας ομάδας, όπως τα γκολ που επιτυγχάνει, είναι ότι η σχέση η οποία ενδεχομένως μπορεί να έχει με το ποσοστό κατοχής είναι έμμεση και όχι άμεση, όπως πιθανά να πιστεύει ένα συντριπτικά μεγάλο κομμάτι του κόσμου που ασχολείται με το άθλημα. Αυτό εκφράστηκε με ισχυρή συσχέτιση της κατοχής με τις συνολικές προσπάθειες για γκολ, αλλά όχι με την ίδια την επίτευξη τερμάτων. Συνεπώς, το να διατηρεί μία ομάδα την κατοχή δεν της δίνει άμεσα το αποτέλεσμα. Χρειάζεται και η δημιουργία ουσιαστικών ευκαιριών για γκολ, ευκαιριών οι οποίες θα έχουν πολλές προοπτικές για να καταλήξουν στο τέρμα. Αυτό είναι και το συμπέρασμα από το γεγονός ότι η μεταβλητή Goals είχε ισχυρή συσχέτιση μόνο με την μεταβλητή OnTarget. Η αδύναμη συσχέτιση της τελευταίας με την μεταβλητή Blocked, η οποία συνδυάζεται με μέτρια με την μεταβλητή Corners, μας δείχνει περίτρανα ότι από τις τρεις επιλογές που έχει ένα σουτ το οποίο έχει κατεύθυνση εντός στόχου, το γκολ είναι η πιο καλά συσχετισμένη, πράγμα που σημαίνει ότι έχει και τις περισσότερες πιθανότητες να προκύψει.

Επίσης, φάνηκε ότι οι μεταβλητές που αφορούν το δημιουργικό σκέλος έχουν μέτρια και ισχυρή αρνητική συσχέτιση με τις μεταβλητές που απεικονίζουν τη δυναμικότητα που έχει ο σύλλογος στην διοργάνωση του UEFA Champions League, πράγμα που δείχνει άλλη μια σχέση μεταξύ της κατοχής και της ικανότητας της ομάδας. Γενικότερα, φάνηκε ότι αυτές οι μεταβλητές έχουν ισχυρή θετική συσχέτιση μεταξύ τους, άρα πιθανά να δοθεί η δυνατότητα κάποιες από αυτές να μείνουν εκτός μοντέλου, για να αποφευχθεί η περίπτωση του να προκύψει ένα μοντέλο με overfitting, με όποια προβλήματα έχουν αυτού του είδους τα μοντέλα.

Εκτός δείγματος στη διαδικασία προσαρμογής του κατάλληλου μοντέλου θα παραμείνουν και οι μεταβλητές Yellow, Red, GoalsCon και Round, καθώς μόνο σε μία

περίπτωση κάποια από αυτές εμφάνισε σημαντική σχέση και αυτή ήταν μεταξύ της Round και της Group, σχέση η οποία μελετήθηκε αποκλειστικά για να βγουν κάποια συμπεράσματα για την κατανομή των συλλόγων αναφορικά με τις knock-out φάσεις και δεν θα βοηθήσει στην προσαρμογή μοντέλου.

Τέλος, αξίζει να αναφερθεί ότι ο παράγοντας έδρα δείχνει να αποτελεί έναν στατιστικά σημαντικό παράγοντα, κρίνοντας συνδυαστικά από τα αποτελέσματα αυτού και του προηγούμενου κεφαλαίου, ο οποίος σίγουρα έχει θέση στην κατάρτιση του κατάλληλου στατιστικού μοντέλου, όπως και η μεταβλητή που αφορά τα γκρουπ δυναμικότητας.

ΚΕΦΑΛΑΙΟ 3

ΑΛΓΟΡΙΘΜΟΙ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

ΕΙΣΑΓΩΓΗ

Στο προηγούμενο κεφάλαιο, πραγματοποιήθηκε μία προσπάθεια, μέσω της χρήσης ενός βασικού εργαλείου, όπως είναι η συσχέτιση, να προκύψουν κάποια πρώτα συμπεράσματα όσων αφορούν την σημαντικότητα των διαφόρων χαρακτηριστικών που περιέχονται στο δείγμα, με σκοπό να υπάρξει μία πρώτη απάντηση στα βασικά ερωτήματα που απασχολούν τη συγκεκριμένη μελέτη, με βασικότερο τη μελέτη των παραγόντων που επηρεάζουν την απόδοση μιας ποδοσφαιρικής ομάδας.

Η μελέτη των συσχετίσεων έδωσε μία πρώτη απάντηση, ωστόσο σε πολλές περιπτώσεις παρατηρήθηκε ότι υπάρχει μία σχετική αβεβαιότητα αναφορικά με τη σημασία κάποιων συσχετίσεων. Σε αυτές τις περιπτώσεις αναφέρονται κυρίως αυτές των μέτριων συσχετίσεων που παρατηρήθηκαν, όπως και οι περιπτώσεις των κατηγορικών μεταβλητών που μπορούν να παίξουν τον ρόλο της εξαρτημένης μεταβλητής στο μοντέλο μας, με σημαντικότερη την Result. Συνεπώς, σε αυτό το κεφάλαιο, θα γίνει μια βαθύτερη προσέγγιση στα παραπάνω ζητήματα.

Η έλλειψη κανονικότητας του δείγματος που ελέγχθηκε, όπως και ο μεγάλος όγκος δεδομένων που περιέχει, δίνει το έναυσμα για να αξιοποιηθούν μέθοδοι εξόρυξης δεδομένων. Μέθοδοι, οι οποίες χρησιμοποιούνται ευρύτατα καθώς αναπτύσσονται μεγάλου όγκου βάσεις δεδομένων. Ο βασικότερος λόγος είναι ότι λόγω της αλγοριθμικής τους προσέγγισης είναι πιο ανθεκτικές σε αυτούς τους μεγάλους όγκους δεδομένων.

Όπως έχει αναφερθεί (Fayyad et al., 1996) «η εξόρυξη δεδομένων είναι η εφαρμογή ειδικών αλγορίθμων για εξαγωγή προτύπων από δεδομένα». Δεδομένου, λοιπόν, της πλειάδας χαρακτηριστικών και εγγραφών που περιέχονται στο εξεταζόμενο δείγμα, θα δώσει μία αρκετά ικανοποιητική προσέγγιση. Στόχος είναι μέσω της χρήσης, αρχικά, αλγορίθμων επιλογής χαρακτηριστικών (feature selection), να γίνει διασταύρωση των αποτελεσμάτων που υπήρξαν με τη χρήση κλασσικών στατιστικών μεθόδων, όπως του ελέγχου συσχετίσεων. Η βάση αυτής της επιλογής θα είναι οι μεταβλητές που έχουν θεωρηθεί, συμπεριλαμβανομένης και της μεταβλητής Result, ότι μπορούν να παίξουν τον ρόλο της εξαρτημένης μεταβλητής του δείγματος.

Η συνέχεια του κεφαλαίου περιλαμβάνει, όπως προδίδει και ο τίτλος του, τη χρήση μεθόδων κατηγοριοποίησης (classification), με στόχο να καταρτιστεί ένα μοντέλο το οποίο θα μπορεί να προβλέπει το αποτέλεσμα ενός αγώνα ή όποιας άλλης μεταβλητής παίζει το ρόλο της εξαρτημένης, έχοντας ως βάση τα χαρακτηριστικά τα οποία επιλέχθηκαν με την προηγούμενη διαδικασία. Η συγκεκριμένη διαδικασία παίζει το ρόλο του προτύπου που θα εξάγουμε από τα δεδομένα και θα αξιολογηθεί ως προς την απόδοσή της με συγκεκριμένες μετρήσεις, που θα αναφερθούν στην ανάλογη ενότητα. Συνδυαστικά, η έρευνα αυτή δείχνει σε

ικανοποιητικό βαθμό και το κατά πόσο συγκεκριμένοι αλγόριθμοι που υπάρχουν στη βιβλιογραφία, μπορούν να χρησιμοποιηθούν για την ανάλυση δεδομένων που αφορούν το ποδόσφαιρο, πράγμα το οποίο θα φανεί ιδιαίτερα χρήσιμο, ώστε να αποκλείει την προσπάθεια χρήσης αναξιόπιστων μεθόδων από μελλοντικούς ερευνητές, γλιτώνοντας έτσι και χρόνο στο να πραγματοποιηθεί μία έγκαιρη και έγκυρη ανάλυση.

3.1 Επιλογή Χαρακτηριστικών (Feature Selection)

Η εποχή που διανύουμε έχει δώσει τη δυνατότητα στον άνθρωπο, μέσω και της εξέλιξης της πληροφορικής, να διαχειριστεί και να προχωρήσει σε ανάλυση από έναν τεράστιο όγκο από δεδομένα. Υπάρχουν βάσεις δεδομένων, οι οποίες μπορεί να περιέχουν και εκατομμύρια εγγραφές. Το ζήτημα δεν εξαντλείται μόνο στο πεδίο των παρατηρήσεων (εγγραφών), καθώς μπορεί να περιέχει και μια πλειάδα χαρακτηριστικών, από την οποία ο στατιστικός ή ο αναλυτής θα πρέπει να βγάλει αποτελέσματα για το ποιες από αυτές είναι σημαντικές και βοηθούν στην εξαγωγή συμπερασμάτων για τα εκάστοτε ερευνητικά ερωτήματα.

Σε αυτές τις περιπτώσεις χρησιμοποιούνται τεχνικές και μέθοδοι που αφορούν το κομμάτι του Feature Selection. Η σκέψη είναι απλή. Αφορά την επιλογή των χαρακτηριστικών, από έναν μεγάλο όγκο χαρακτηριστικών, τα οποία είναι σημαντικά σε σχέση με το δείγμα συνολικά ή την εξαρτημένη μεταβλητή που τίθεται στο επίκεντρο. Η χρήση τους στην ανάλυση δεδομένων γίνεται όλο και πιο καθοριστική στην τελική επιλογή του κατάλληλου μοντέλου. Επίσης, έχουν μεγάλη συνεισφορά και σε πεδία που δεν απασχολούν τους Αναλυτές.

Πιο συγκεκριμένα, τα οφέλη τους συνοπτικά είναι τα εξής: διευκόλυνση της οπτικοποίησης (visualization) και της κατανόησης των δεδομένων, μείωση των απαιτήσεων που αφορούν στις μετρήσεις και την αποθήκευση των δεδομένων, μείωση χρόνων που αφορούν την εκπαίδευση (training) του μοντέλου και την εν γένει χρήση του (Guyon et al., 2003).

Συνεπώς, σε αυτή τη φάση, με βάση και τις μεταβλητές που έχουν επιλεχθεί να εξεταστούν ως εξαρτημένες από το σύνολο των δεδομένων, θα χρησιμοποιηθούν τέτοιοι αλγόριθμοι, οι οποίοι θα καθορίσουν και το σύνολο των ανεξάρτητων μεταβλητών που θα κληθούν να περιγράψουν την εξαρτημένη. Η συνέχεια της ενότητας αυτής αφορά τη χρήση και τα αποτελέσματα αυτών των αλγορίθμων.

3.1.1 Ο *αλγόριθμος* Random Forest

Ο συγκεκριμένος αλγόριθμος αποτελεί έναν από τους πιο γνωστούς αλγόριθμους στο αντικείμενο της Μηχανικής Μάθησης (Machine Learning). Έχει δείξει (Chen et al., 2020) ότι έχει μεγάλη ακρίβεια τόσο στην ανάδειξη των χαρακτηριστικών με τη μεγαλύτερη επιρροή στο δείγμα, όσο και στην περίπτωση που αξιοποιείται για το ζήτημα της κατηγοριοποίησης των

δεδομένων. Επιπλέον, ένα θετικό χαρακτηριστικό του είναι η ταχύτητα υπολογισμού των αποτελεσμάτων. Επιπλέον, μπορεί να διαχειριστεί αποτελεσματικά μεγάλα σύνολα δεδομένων, με έναν μεγάλο αριθμό χαρακτηριστικών (Καλλιακμάνης, 2020)

Ο συγκεκριμένος αλγόριθμος αποτελεί στην ουσία μία γενίκευση των δέντρων απόφασης (decision trees), με πολλές επαναλήψεις της συγκεκριμένης μεθόδου. Τα δέντρα απόφασης αποτελούν μία μέθοδο κατηγοριοποίησης, στην οποία ουσιαστικά γίνεται διαχωρισμός των παρατηρήσεων που ανήκουν σε κάθε κατηγορία με βάση τις τιμές που έχουν σε αυτά. Βασιζόμενοι σε μία συνάρτηση που αναδεικνύει τον βαθμό ακαθαρσίας (degree of impurity), επιλέγουν τον διαχωρισμό που έχει με τον μικρότερο βαθμό.

Με βάση τα παραπάνω, τα βήματα που συνθέτουν τον αλγόριθμο Random Forest έχουν ως εξής:

1. Επιλέγεται ένα τυχαίο δείγμα από το αρχικό σύνολο δεδομένων

2. Χρησιμοποιείται η μέθοδος των δέντρων απόφασης για το επιλεγμένο τυχαίο δείγμα.

3. Επιλέγεται από τον χρήστη ο αριθμός των δέντρων (επαναλήψεων) που επιθυμεί να πραγματοποιηθούν και επαναλαμβάνονται ανάλογες φορές τα Βήματα 1 και 2.

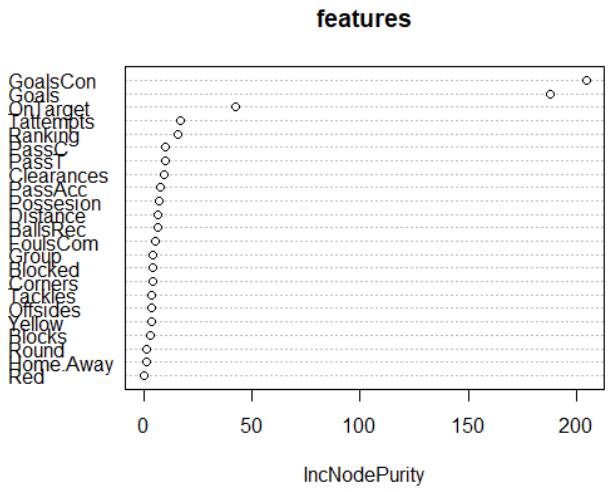
4. Για ένα νέο σημείο, κάθε ένα από τα παραπάνω δέντρα προβλέπει την κατηγορία που ανήκει το συγκεκριμένο σημείο και καταχωρείται στο νέο σημείο η κατηγορία με τις περισσότερες εμφανίσεις.

Τα παραπάνω, ωστόσο, είναι τα βήματα που αφορούν τον συγκεκριμένο αλγόριθμο και αφορούν το κομμάτι της κατηγοριοποίησης, τα αποτελέσματα της οποίας δεν θα αξιοποιηθούν. Όμως, εφαρμόζοντας τον συγκεκριμένο κώδικα μέσω του στατιστικού πακέτου της R δίνεται η δυνατότητα στη συνέχεια να χρησιμοποιηθούν οι ανάλογες εντολές για τον υπολογισμό της σημαντικότητας της κάθε μεταβλητής ξεχωριστά. Για τον υπολογισμό της σημαντικότητας της κάθε μεταβλητής αξιοποιούνται τα out-of-bag (OOB) δεδομένα, δηλαδή τα δεδομένα τα οποία δεν χρησιμοποιήθηκαν καθόλου στη διαδικασία του αλγορίθμου. Στο παράρτημα Δ του κώδικα που χρησιμοποιήθηκε για την παρούσα μελέτη φαίνονται οι εντολές που χρησιμοποιήθηκαν σε όλα τα πεδία της ανάλυσης. Συγκεκριμένα, χρησιμοποιήθηκε η βιβλιοθήκη randomForest, που τη βρίσκουμε εγκαθιστώντας το ομώνυμο πακέτο της R. Για καλύτερη κατανόηση των αποτελεσμάτων, ακόμα και από μη στατιστικούς, θα παρουσιαστούν και ανάλογα διαγράμματα (plots), για βαθύτερη κατανόηση του ζητήματος.

Η χρήση του αλγορίθμου έγινε, λαμβάνοντας ως εξαρτημένες μεταβλητές εκείνες που για τους περισσότερους ανθρώπους που εργάζονται και ζουν από το ποδόσφαιρο, συμπεριλαμβανομένων και των οπαδών, είναι εκείνες που έχουν εν τέλει τη μεγαλύτερη σημασία και έχουν την δυνατότητα να κρίνουν τα πάντα: Τα γκολ και το αποτέλεσμα. Αντικειμενικά, όσο ποιοτικό και σύγχρονο ποδόσφαιρο να παίζει μία ποδοσφαιρική ομάδα, όταν αυτό δεν καταλήγει στην επίτευξη των στόχων, όπως το να σκοράρει και το να παίρνει τις νίκες, που φέρνουν και τους τίτλους, κανείς δεν μένει ευχαριστημένος μόνο από τον τρόπο παιχνιδιού. Ίσα ίσα που αυτός ο τρόπος είναι που πρέπει να εξυπηρετεί αυτούς τους δύο σκοπούς.

Ο λόγος για τον οποίο δεν χρησιμοποιείται μόνο μία από τις δύο είναι διότι μέσω της χρήσης και των δύο θα μας δοθεί με μεγαλύτερη ακρίβεια το αποτέλεσμα, που αφορά τις μεταβλητές, άρα τους παράγοντες, που επηρεάζουν την πραγματική απόδοση μιας ομάδας. Αυτό ισχύει διότι με τη χρήση π.χ. της μεταβλητής Result, που αφορά το αποτέλεσμα του αγώνα, δεν λαμβάνονται υπόψη αγώνες οι οποίοι μπορεί να έληξαν ισόπαλοι, αλλά με μεγάλο αριθμό γκολ. Αυτό μπορεί να σημαίνει, από τη μία, ότι μία ομάδα είχε μία κακή αμυντική απόδοση, αλλά επίσης, καθώς το δείγμα μας αποτελείται από ομάδες που αγωνίστηκαν στη μεγαλύτερη διασυλλογική διοργάνωση του πλανήτη, μπορεί να σημαίνει και ότι μιλάμε για δύο υψηλού επιπέδου ομάδες. Άρα πρέπει να συμπεριληφθεί και η μεταβλητή Goals.

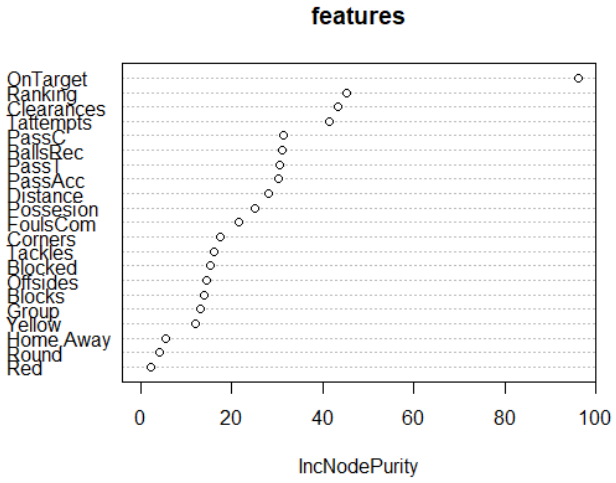
Ξεκινώντας από την μεταβλητή Result η εφαρμογή της Random Forest στο σύνολο των μεταβλητών έδειξε σημαντικές μόνο τις μεταβλητές Goals και GoalsCon, που αφορούν τα γκολ που πέτυχε μία ομάδα και αυτά που δέχτηκε και αρκετά πιο πίσω την OnTarget. Αυτό είναι απολύτως αναμενόμενο, καθώς ακόμα και ο κώδικας που χρησιμοποιήθηκε στην R χρησιμοποίησε τις πρώτες δύο μεταβλητές για να υπολογίσει την μεταβλητή Result. Συνεπώς, ως πρώτο αποτέλεσμα δείχνει να μην βοηθάει στην εξαγωγή συμπερασμάτων.



Διάγραμμα 3.1: Σημαντικότητα μεταβλητών της Random Forrest με μεταβλητή απόκρισης την Result.

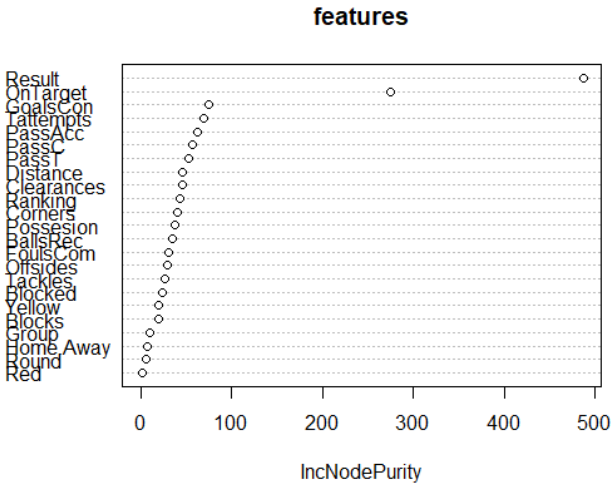
Ωστόσο, λόγω της τεράστιας αυτής διαφοράς, αλλά και του γεγονότος ότι αυτή η μέθοδος προϋποθέτει την προσαρμογή ενός μοντέλου Random Forest, θα μπορούσε να υπάρξει και κάποιος διαφορετικός τρόπος για να βγει ένα ασφαλές αποτέλεσμα για τις υπόλοιπες μεταβλητές. Αυτός ο τρόπος έχει να κάνει με την προσαρμογή ενός μοντέλου, το οποίο δεν θα περιέχει τις μεταβλητές Goals και GoalsCon, αλλά θα περιέχει όλες τις υπόλοιπες.

Με την προσαρμογή του παραπάνω μοντέλου εμφανίστηκαν αρκετά σημαντικές διαφορές, αναφορικά με τη σημαντικότητα των υπόλοιπων μεταβλητών. Όπως θα δούμε παρακάτω, εκτός από την σημαντική διαφορά που έχει η OnTarget με τις υπόλοιπες, σημαντικό ρόλο φαίνεται να παίζουν και οι μεταβλητές Ranking, Clearances και Tattempts.



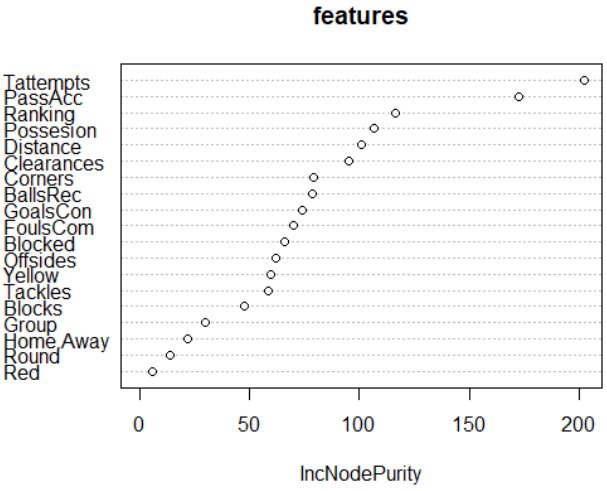
Διάγραμμα 3.2: Σημαντικότητα μεταβλητών της Random Forrest με μεταβλητή απόκρισης την Result χωρίς τις Goals και GoalsCon.

Για την μεταβλητή Goals είναι φανερό ότι, ακολουθώντας την ίδια διαδικασία, μπορούμε να δούμε αρκετά κοινά χαρακτηριστικά με την μεταβλητή Result, ωστόσο υπάρχουν και διαφορές. Ξεκινώντας από την προσαρμογή του μοντέλου του αλγόριθμου Random Forest, παρατηρείται ότι οι μοναδικές μεταβλητές που φαίνονται να είναι σημαντικές είναι η Result και η OnTarget. Αυτό το αποτέλεσμα διασταυρώνεται απόλυτα με το πλήρες μοντέλο που είχε σαν εξαρτημένη τη μεταβλητή Result. Συνεπώς, για μεγαλύτερη διερεύνηση και των υπόλοιπων χαρακτηριστικών, έγινε προσαρμογή και του μοντέλου που δεν περιέχει αυτές τις δύο μεταβλητές.



Διάγραμμα 3.3: Σημαντικότητα μεταβλητών της Random Forrest με μεταβλητή απόκρισης την Goals.

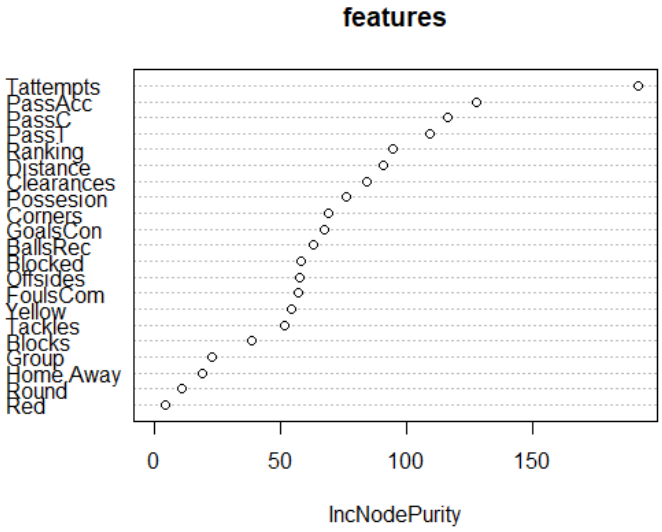
Η συγκεκριμένη ανάλυση δείχνει εμφανώς να έχει διαφορές με την προηγούμενη, ωστόσο εκτός από την περίπτωση της Tattempts, δεν φαίνεται να υπάρχει κάποια μεταβλητή που να επηρεάζει με σημαντική διαφορά από τις υπόλοιπες τα αποτελέσματα του μοντέλου. Αυτό όμως ενδέχεται να έχει και εξήγηση το γεγονός ότι τα επόμενα τρία χαρακτηριστικά, τα οποία δείχνουν να έχουν μεγάλη σημασία την προσαρμογή του μοντέλου, είναι οι PassAcc, PassT και PassC, οι οποίες όπως παρατηρήθηκε και στο προηγούμενο Κεφάλαιο είναι υψηλά συσχετισμένες, συνεπώς ενδέχεται στην περίπτωση αυτή να υπάρχει overfitting.



Διάγραμμα 3.4: Σημαντικότητα μεταβλητών της Random Forrest με μεταβλητή απόκρισης την Goals, εξαιρώντας την Result και την OnTarget.

Άρα το επόμενο βήμα είναι η προσαρμογή ενός μοντέλου, το οποίο θα περιέχει μόνο την PassAcc από αυτές τις τρεις, δεδομένου ότι εκφράζει το ποσοστό επιτυχημένων πασών, άρα είναι το πηλίκο της PassC προς την PassT, οπότε κρίνεται πιο επιτακτικό να παραμείνει αυτή στο δείγμα.

Αυτή τη φορά τα αποτελέσματα είναι πιο ξεκάθαρα. Εκτός από την ξεκάθαρη υπεροχή της Tattempts και την παραμονή στη δεύτερη θέση της PassAcc, υπάρχει άλλη μία ομάδα χαρακτηριστικών, η οποία δείχνει να είναι πιο μπροστά σε σημασία από το υπόλοιπο σώμα των χαρακτηριστικών. Αυτή περιέχει τις Ranking, Possession, Distance και Clearances. Οπότε, αναφορικά με το μοντέλο που θα έχει σαν εξαρτημένη μεταβλητή την Goals, οι μεταβλητές που θα αξιοποιηθούν θα είναι η OnTarget, Tattempts, PassAcc, Ranking, Possession, Distance και Clearances.



Διάγραμμα 3.5: Σημαντικότητα μεταβλητών της Random Forrest με μεταβλητή απόκρισης την Goals, εξαιρώντας την Result, OnTarget, PassT και PassC.

3.1.2 Συμπεράσματα από τη χρήση του αλγορίθμου

Από τη χρήση του αλγόριθμου της Random Forest μπορούμε με βεβαιότητα να πούμε ότι στο κομμάτι της αξιολόγησης των χαρακτηριστικών και της επιλογής των καλύτερων εξ αυτών, ο αλγόριθμος λειτούργησε, έχοντας δεδομένα τα οποία μέχρι στιγμής δεν έχουν επεξεργαστεί από την πρόσφατη βιβλιογραφία.

Τα αποτελέσματα αυτά αν συγκριθούν με εκείνα του Κεφαλαίου 2, που χρησιμοποιήθηκαν οι μέθοδοι που αφορούν τη χρήση συντελεστή συσχέτισης, θα δούμε ότι συμπίπτουν σε πολύ μεγάλο βαθμό. Επιβεβαιώθηκε και εδώ η μεγάλη επιρροή που έχουν η μεταβλητή OnTarget και η μεταβλητή Tattempts, σε μικρότερο βαθμό από την πρώτη, όπως και οι μεταβλητές που αφορούν την κατοχή. Η μεταξύ τους ισχυρή συσχέτιση, που αποτυπώθηκε και στο προηγούμενο Κεφάλαιο, έδωσε και τη δυνατότητα να εντοπιστεί και εν τέλει να διορθωθεί το πρόβλημα του overfitting που παρουσιάστηκε. Συνεχίζοντας, καταγράφεται ως σημαντικό η μεγάλη επιρροή που έχουν οι μεταβλητές Clearances και Ranking. Αυτό μας δίνει το δικαίωμα να διατυπώσουμε ότι η μεταβλητή Group, που είναι αποτέλεσμα στο πεδίο του data preprocessing, φάνηκε χρήσιμη αποκλειστικά για το πεδίο της περιγραφικής ανάλυσης.

Παρόλο που αναζητήθηκαν τα χαρακτηριστικά που έχουν τη μεγαλύτερη σημασία για την κατάρτιση ενός μοντέλου εξόρυξης δεδομένων, η δουλειά αναφορικά με τα ερευνητικά ερωτήματα της παρούσας μελέτης δεν έχει σε καμία περίπτωση ολοκληρωθεί. Αντιθέτως, θα πρέπει να γίνει χρήση τεχνικών κατηγοριοποίησης. Οι λόγοι είναι ότι, από τη μία, θα έχουμε άλλον έναν τρόπο για να επιβεβαιώσουμε στην πράξη ότι επαρκούν οι συγκεκριμένες μεταβλητές για την εξαγωγή συμπερασμάτων και, από την άλλη, θα μπορέσουμε να πραγματοποιήσουμε μία πρώτη μοντελοποίηση, ούτος ώστε να εξηγηθεί και ο τρόπος με τον οποίο επιδρούν τα χαρακτηριστικά που επιλέχθηκαν.

Τέλος, αξίζει να τονιστεί ότι τα συγκεκριμένα χαρακτηριστικά είναι οι βασικές στατιστικές μετρήσεις, που καταγράφονται σε έναν ποδοσφαιρικό αγώνα και είναι αναγνώσιμες από το φίλαθλο κοινό. Συνεπώς, είναι βέβαιο ότι δεν αποτελούν την αρχή και το τέλος των χαρακτηριστικών που μπορούν να ληφθούν υπόψη σε μία έρευνα. Αποτελούν, ωστόσο, μία πρωταρχική έρευνα με σύγχρονα δεδομένα, τα οποία μπορεί ένας μελλοντικός ερευνητής να βασιστεί πάνω τους και να αναπτύξει μία μελλοντική θεωρία, αλλά και ένας φίλαθλος ή επαγγελματίας του αθλήματος να μπορεί να εστιάσει στα σημαντικά.

3.2 Χρήση αλγόριθμων κατηγοριοποίησης για την μεταβλητή Result

Σε αυτήν την ενότητα θα εξεταστούν βασικοί αλγόριθμοι κατηγοριοποίησης, βασιζόμενοι και στα αποτελέσματα που έδωσε το πεδίο της επιλογής χαρακτηριστικών.

Σαν ένας γενικότερος ορισμός «η κατηγοριοποίηση είναι η διαδικασία της μάθησης μιας συνάρτησης – στόχου f που κατευθύνει κάθε σετ χαρακτηριστικών x σε μία από τις προκαθορισμένες κατηγορίες y. Η συνάρτηση – στόχος είναι, επίσης, γνωστή ως ένα μοντέλο κατηγοριοποίησης» (Tan et al., 2005).

Μεταφράζοντας τον παραπάνω ορισμό στο δείγμα που εξετάζει η παρούσα μελέτη, έχοντας το σετ χαρακτηριστικών, όπως αυτό επιλέχθηκε για την μεταβλητή Result μέσω του

αλγόριθμου της Random Forest, η προσπάθεια θα επικεντρωθεί στην εύρεση το κατάλληλου μοντέλου κατηγοριοποίησης, το οποίο θα μπορέσει να «διαβάσει» σε ικανοποιητικό βαθμό την συγκεκριμένη μεταβλητή, στηριζόμενοι μόνο στα επιλεγμένα χαρακτηριστικά, αντί για το συνολικό δείγμα.

Ο λόγος για τον οποίο θα γίνει αυτή η προσπάθεια μόνο για την μεταβλητή Result και όχι και για την Goals είναι ξεκάθαρος: «Οι τεχνικές κατηγοριοποίησης είναι πιο κατάλληλες για να περιγράψουν ή να προβλέψουν datasets με δίτιμες ή κατηγορικές μεταβλητές. Είναι λιγότερο αποτελεσματικά στις διατακτικές μεταβλητές επειδή δεν περιλαμβάνουν τη σειρά ανάμεσα στις κατηγορίες» (Tan et al., 2005). Η Goals είναι μία μεταβλητή, η οποία παρόλο που έχει ένα εύρος τιμών που αφορά τις ακέραιες τιμές από 0 έως 8 στο εξεταζόμενο δείγμα, δεν αποτελεί μια κατηγορική μεταβλητή. Συνεπώς, δεν θα βοηθήσει στην συγκεκριμένη περίπτωση η χρήση τέτοιων μεθόδων.

Στις επόμενες δύο υποενότητες θα γίνουν ξεκάθαρες δύο πλευρές της πρακτικής που ακολουθείται, όταν χρησιμοποιούνται τεχνικές κατηγοριοποίησης, οι οποίες είναι απαραίτητο να ληφθούν υπόψη κατά τη διαδικασία της έρευνας για το κατάλληλο μοντέλο.

3.2.1 Overfitting

Ορισμένες φορές παρατηρείται το φαινόμενο ένα μοντέλο να έχει τέλεια προσαρμογή στα δεδομένα. Να μπορεί, δηλαδή, αυτή η συνάρτηση – στόχος να κατευθύνει κάθε παρατήρηση στην κατηγορία, στην οποία πραγματικά ανήκει. Ένα τέτοιο φαινόμενο δεν σημαίνει απαραίτητα ότι έχει γίνει η τέλεια δουλειά, πόσο μάλλον ότι η έρευνά μας έχει τελειώσει. Μπορεί πολύ πιθανά να μιλάμε για μια περίπτωση υπερπροσαρμογής (overfitting) του μοντέλου. Τέτοιες περιπτώσεις συμβαίνουν όταν η προσαρμογή του μοντέλου επηρεάζεται από τη μορφή που έχουν τα δεδομένα που λαμβάνει ως αρχικά δεδομένα (training set) με σκοπό τον ακριβή υπολογισμό της συνάρτησης – στόχου. Αυτό έχει σαν αποτέλεσμα, όταν προσπαθούμε να δοκιμάσουμε το μοντέλο μας σε νέα δεδομένα (test set), αυτό να μην δίνει τα αναμενόμενα αποτελέσματα, με αποτέλεσμα να έχουμε μία λανθασμένη προσέγγιση.

Στην συγκεκριμένη περίπτωση και για την αποφυγή οποιασδήποτε τέτοιας περίπτωσης, θα μοιράσουμε τα δεδομένα μας σε δύο sets (training set και test set), με το μικρότερο ποσοστό (30%) των δεδομένων μας να πηγαίνει στο test set, έτσι ώστε να πάρουμε μία προσέγγιση που δεν επηρεάζεται από τέτοια φαινόμενα.

3.2.2 Μέτρηση της απόδοσης ενός αλγορίθμου κατηγοριοποίησης

Η απόδοση ενός αλγόριθμου κατηγοριοποίησης συνίσταται στο κατά πόσο ο αλγόριθμος κατηγοριοποιεί σωστά τις εγγραφές. Δηλαδή σε τι ποσοστό τα αποτελέσματα που

προκύπτουν από τη χρήση του αλγορίθμου συμπίπτουν με τα πραγματικά. Συνεπώς, η μέτρηση της απόδοσης θα χρειαστεί μεθόδους μέτρησης, οι οποίες δηλώνουν τα παραπάνω στοιχεία.

Η πρώτη από αυτές είναι η μήτρα σύγχυσης. Η μήτρα σύγχυσης είναι ένας πίνακας διαστάσεων 2 X 2, ο οποίος αποτυπώνει το πλήθος των ορθά και των λανθασμένα κατανεμημένων εγγραφών. Οι γραμμές του δείχνουν τις πραγματικές κατηγορίες που έχουν καταγραφεί και οι στήλες τις εκτιμημένες. Άρα, γίνεται αντιληπτό ότι το πλήθος των εγγραφών που έχουν διαφορετική πραγματική από εκτιμημένη κατηγορία είναι οι εσφαλμένες περιπτώσεις του δείγματος. Πιο συγκεκριμένα, στον πίνακα οι σωστές κατανομές αποτυπώνονται στην κύρια διαγώνιό του (στην παρακάτω αποτύπωση TP και TN) και οι λανθασμένες στα υπόλοιπα.

Confusion Matrix. Table 2 The outcomes of classification into positive and negative classes

		Assigned Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

Πίνακας 3.1: Μήτρα σύγχυσης (Πηγή: Sammut, C., & Webb, G. I. (Eds.). (2010). *Encyclopedia of Machine Learning.* doi:10.1007/978-0-387-30164-8)

Ο παραπάνω τρόπος διαμορφώνει και τη δεύτερη μέθοδο μέτρησης, που έχει να κάνει με το ποσοστό ορθής κατηγοριοποίησης. Ουσιαστικά, μιλάμε για το πηλίκο των ορθώς ταξινομημένων εγγραφών προς τις συνολικές (δηλαδή ορθές και λανθασμένες).

3.3 Ο αλγόριθμος Naïve Bayes

Ο αλγόριθμος Naïve Bayes είναι ένας αλγόριθμος, ο οποίος χρησιμοποιεί τον κανόνα του Bayes, με την υπόθεση ότι, δεδομένης της κατηγορίας που ανήκουν, τα χαρακτηριστικά είναι ανεξάρτητα. Ο συγκεκριμένος αλγόριθμος έχει βοηθήσει πολύ σε δεδομένα που αφορούν έγγραφα και γενικότερα αρχεία κειμένου (Ting et al., 2011), πράγμα που έχει βοηθήσει πολύ στον εντοπισμό fake news (Granik et al.,2017).

Υπενθυμίζεται ότι, θεωρώντας μία κατηγορία y και μία εγγραφή x, ο κανόνας του Bayes μας δίνει ότι

$$P(y|x) = \frac{P(y)P(x|y)}{P(x)}$$

Επειδή ως x έχει οριστεί μία οποιαδήποτε εγγραφή του συνόλου δεδομένων και οι εγγραφές αποτελούνται από τιμές στα χαρακτηριστικά του συνόλου δεδομένων, η υπόθεση της ανεξαρτησίας μας δίνει ότι:

$$P(x|y) = \prod_{i=1}^{n} P(x_i|y)$$

όπου x_i είναι η τιμή του i-οστού χαρακτηριστικού για την εγγραφή x και n είναι το πλήθος των χαρακτηριστικών.

Με βάση, λοιπόν, το δείγμα που υπάρχει, υπολογίζονται οι δεσμευμένες πιθανότητες $P(x_i|y)$. Στη συνέχεια, για τον υπολογισμό μιας νέας μεταβλητής, με βάση και τις εξισώσεις που αναφέρθηκαν παραπάνω, υπολογίζεται η πιθανότητα η νέα παρατήρηση να ανήκει σε κάθε μία από τις κατηγορίες της εξαρτημένης μεταβλητής. Ο αλγόριθμος θα κατατάξει τη νέα μεταβλητή εκεί που συγκεντρώνεται η μεγαλύτερη πιθανότητα (Πελέκης, 2019).

Ο λόγος για τον οποίο επιλέχθηκε να χρησιμοποιηθεί ο συγκεκριμένος αλγόριθμος είναι διότι:

1. Βασίζεται σε ένα βασικό κανόνα της θεωρίας πιθανοτήτων, συνεπώς μπορεί να γίνει εύκολα κατανοητός και από ανθρώπους που δεν ασχολούνται με το αντικείμενο της Μηχανικής Μάθησης.

2. Στην απόδοσή του παίζει καθοριστικό ρόλο η έννοια της ανεξαρτησίας, η οποία δεν έχει αποδοθεί πλήρως στα χαρακτηριστικά που εξετάζουμε στο παρόν σύνολο δεδομένων. Παρόλο που βρέθηκαν σημαντικές συσχετίσεις στο κεφάλαιο 2, η χρήση του είναι μία ακόμα ένδειξη της σημασίας τους.

3. Είναι ένας αλγόριθμος μηχανικής μάθησης, ο οποίος χρησιμοποιείται ευρέως και σε πολλούς τύπους δεδομένων.

3.3.1 Αποτελέσματα της εφαρμογής του Naïve Bayes στα δεδομένα

Προτού αναλυθούν τα αποτελέσματα της χρήσης του αλγορίθμου Naïve Bayes στα δεδομένα, αξίζει να αναφερθεί ότι στις συνεχείς και τις διακριτές μεταβλητές εφαρμόστηκε η τεχνική της κανονικοποίησης, με σκοπό οι μεταβλητές αυτές να έχουν ένα κοινό διάστημα που λαμβάνουν τις τιμές (αυτό είναι το [0,1]). Ο λόγος που πάρθηκε αυτή η απόφαση είναι το γεγονός ότι υπάρχουν μεταβλητές με 3ψήφιες τιμές, όπως η Distance ή η PassC και μεταβλητές

με μονοψήφιες τιμές, όπως η Goals. Αυτό ενδέχεται να έδινε αποτελέσματα τα οποία θα επηρεάζονταν σε σημαντικό βαθμό από τις μεταβλητές με τις μεγάλες τιμές, επικαλύπτοντας τη σημαντικότητα των μεταβλητών με τις μικρότερες. Υπενθυμίζεται ότι, τα δεδομένα μας έχουν χωριστεί σε Training και Test Set, επομένως τα αποτελέσματα που θα παρουσιαστούν αφορούν το Test Set.

Αναφορικά με τα τελικά αποτελέσματα, η αρχή γίνεται εισάγοντας το σύνολο των χαρακτηριστικών στο μοντέλο, με εξαρτημένη μεταβλητή την Result. Φαίνεται ότι το μοντέλο, δεδομένου του μεγέθους δείγματος, έχει μία καλή προσαρμογή στα δεδομένα. Ωστόσο, το πλήθος των λανθασμένων κατηγοριοποιήσεων είναι αρκετά υψηλό. Συγκεκριμένα, το ποσοστό των σωστά κατανεμημένων παρατηρήσεων είναι 73.87% και η μεγαλύτερη ασάφεια παρατηρείται στις ισοπαλίες, όπου έχουμε 28 λανθασμένες κατηγοριοποιήσεις από τις συνολικά 52 ισοπαλίες. Αυτό έχει την ουσία του, καθώς γενικότερα η ισοπαλία είναι ένα αποτέλεσμα που δεν μπορεί εύκολα να προβλεφθεί από τη σκοπιά της μέτρησης των βασικών δεικτών απόδοσης της κάθε ομάδας ξεχωριστά, αλλά σε συνδυασμό.

	Ήττα	**Ισοπαλία**	**Νίκη**
Ήττα	72	5	8
Ισοπαλία	14	24	14
Νίκη	3	14	68

Πίνακας 3.2: Αποτελέσματα Naïve Bayes για το σύνολο των χαρακτηριστικών

Βελτιωμένα αποτελέσματα δόθηκαν όταν χρησιμοποιήθηκαν τα χαρακτηριστικά που με βάση τον αλγόριθμο Random Forest κρίθηκε ότι παίζουν σημαντικότερο ρόλο για τη μεταβλητή Result. Συγκεκριμένα, το μοντέλο περιείχε τις μεταβλητές Goals, GoalsCon, Tattempts, OnTarget, Clearances και Ranking. Ο αλγόριθμος Naïve Bayes έδωσε ένα ποσοστό 79.72% στις σωστές κατανομές, με το αποτέλεσμα της ισοπαλίας να είναι κάπως βελτιωμένο, με 28 από τις 52 παρατηρήσεις να είναι σωστά κατανεμημένες. Τα συγκεκριμένα αποτελέσματα επιβεβαιώνουν σε ένα βαθμό την χρήση του αλγόριθμου Feature Selection, καθώς φαίνεται ότι η χρήση του συνόλου των δεδομένων δημιουργεί κάποιο θόρυβο, που σαν αποτέλεσμα έχει τη μειωμένη απόδοση του αλγορίθμου.

	Ήττα	**Ισοπαλία**	**Νίκη**
Ήττα	76	7	2
Ισοπαλία	8	32	12
Νίκη	0	16	69

Πίνακας 3.3: Αποτελέσματα Naïve Bayes για τα χαρακτηριστικά της μεθόδου Feature Selection

Παρόλα αυτά, στα αποτελέσματα και ειδικότερα σε αυτά που αφορούν το κομμάτι της ισοπαλίας, αναδεικνύεται για μια ακόμα φορά η μεγάλη συμβολή που έχουν οι μεταβλητές Goals και GoalsCon. Στην προσπάθεια να προσαρμοστεί ένα μοντέλο το οποίο οι δύο αυτές μεταβλητές ήταν εντελώς απούσες είχαμε αντιστρόφως ανάλογα αποτελέσματα από τα προηγούμενα. Συγκεκριμένα, το 59.45% με μόλις 3 παρατηρήσεις που αφορούσαν την ισοπαλία να κατανέμονται ορθά και πολλές λανθασμένες στις άλλες δύο κατηγορίες. Ωστόσο, φαίνεται ότι ένα μοντέλο με μόνο τις άλλες 4 μεταβλητές στέκεται αρκετά καλά αναφορικά με την διαμόρφωση του αποτελέσματος, καθώς οι μεταβλητές Goals και GoalsCon προσθέτουν λίγο πάνω από 20% απόδοσης.

	Ήττα	**Ισοπαλία**	**Νίκη**
Ήττα	67	2	16
Ισοπαλία	29	3	20
Νίκη	23	0	62

Πίνακας 3.4: Αποτελέσματα Naïve Bayes για τα χαρακτηριστικά της μεθόδου Feature Selection, εξαιρουμένων των μεταβλητών Goals και GoalsCon

Τέλος, αξίζει να αναφερθεί ότι τα αποτελέσματα που δίνει η αποκλειστική χρήση των μεταβλητών Goals και GoalsCon είναι και αυτά που, όπως αναμενόταν, δίνουν και την μεγαλύτερη απόδοση. Το συγκεκριμένο μοντέλο έδωσε μία απόδοση 89.18%, ωστόσο δεν εξυπηρετεί κανένα επιστημονικό ερώτημα της παρούσας μελέτης.

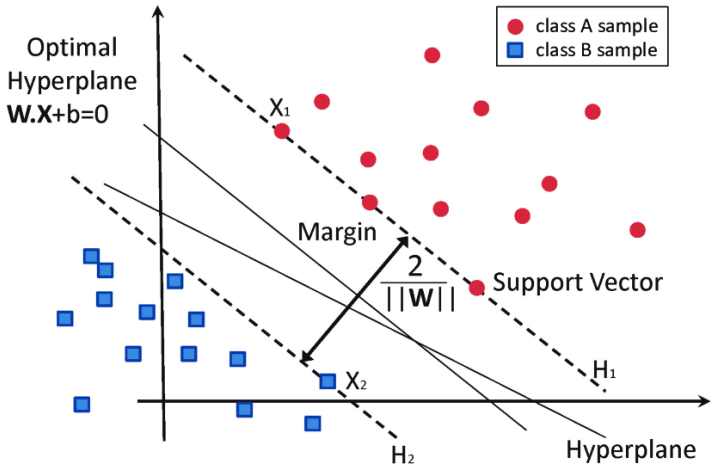
	Ήττα	**Ισοπαλία**	**Νίκη**
Ήττα	73	12	0
Ισοπαλία	0	52	0
Νίκη	0	12	73

Πίνακας 3.5: Αποτελέσματα Naïve Bayes για το μοντέλο που περιέχει μόνο τις μεταβλητές Goals και GoalsCon

3.4 Ο αλγόριθμος Support Vector Machine

Οι Support Vector Machines (SVMs) είναι μία κατηγορία αλγορίθμων που χρησιμοποιούνται για κατηγοριοποίηση και παλινδρόμηση. Στην απλούστερη μορφή τους, αυτή της κατηγοριοποίησης μεταξύ δύο κατηγοριών, βρίσκει ένα υπερπλάνο (hyperplane) που διαχωρίζει τις δύο κατηγορίες με το μεγαλύτερο δυνατό περιθώριο. Βασίζεται μόνο στα δεδομένα που βρίσκονται στα σύνορα αυτού του περιθωρίου, που ονομάζονται support vectors.

Αυτό οδηγεί σε μία μεγάλη ακρίβεια στην κατηγοριοποίηση και υποστηρίζει μεθόδους βελτιστοποίησης που επιτρέπουν στον αλγόριθμο να είναι αποτελεσματικός σε μεγάλη ποσότητα δεδομένων (Sammut C., & Webb, G. I. (Eds.). (2010). *Encyclopedia of Machine Learning*. doi:10.1007/978-0-387-30164-8).



Διάγραμμα 3.6: Λειτουργία αλγορίθμου SVM (Πηγή: García-Gonzalo, E., Fernández-Muñiz, Z., Nieto, P. J. G., Sánchez, A. B., Fernández, M. M, *Hard-Rock Stability Analysis for Span Design in Entry-Type Excavations with Learning Classifiers*, Materials 2016)

Το υπερπλάνο δημιουργείται χρησιμοποιώντας τις συναρτήσεις Kernel. Για τα δεδομένα μας θα χρησιμοποιηθεί η συνάρτηση RBF, μία από τις 4 πιο συνηθισμένες Kernel που υποστηρίζει η R.

Kernel	Equation
Linear	$K(x, y) = x \cdot y$
Sigmoid	$K(x, y) = \tanh(ax.y + b)$
Polynomial	$K(x, y) = (1 + x \cdot y)^d$
KMOD	$K(x, y) = a\left[\exp\left(\frac{\gamma}{\|\|x-y\|\|^2+\sigma^2}\right) - 1\right]$
RBF	$K(x, y) = \exp(-a\|\|x - y\|\|^2)$
Exponential RBF	$K(x, y) = \exp(-a\|\|x - y\|\|)$

Πίνακας 3.6: Βασικές συναρτήσεις Kernel (Πηγή: Chou, J. S., Chiu, C. K., Farfoura, M., Al-Taharwa, I. (2011) *Optimizing the Prediction Accuracy of Concrete Compressive Strength Based on a Comparison of Data-Mining Techniques*, Journal of Computing in Civil Engineering)

3.4.1 Αποτελέσματα της εφαρμογής του SVM στα δεδομένα

Η χρήση του συγκεκριμένου αλγορίθμου απορρέει και από το γεγονός ότι χρησιμοποιείται και είναι πολύ αποδοτικός σε πολλούς τύπους δεδομένων. Συνεπώς, ένας επιπλέον στόχος είναι ο έλεγχος της απόδοσης του συγκεκριμένου αλγορίθμου στα ποδοσφαιρικά δεδομένα. Όπως και στον αλγόριθμο Naïve Bayes, έτσι και εδώ, θα υπάρξει κανονικοποίηση στα αριθμητικά δεδομένα.

Ξεκινώντας από τη χρήση του στο μοντέλο που χρησιμοποιεί όλα τα χαρακτηριστικά του δείγματος, είναι φανερό ότι πρόκειται για μία πολύ καλή προσαρμογή. Συγκεκριμένα, κατηγοριοποιήθηκε σωστά το 81.98% των παρατηρήσεων, που από μόνο του είναι ένα πολύ καλό ποσοστό. Φαίνεται ότι στην περίπτωση αυτή υπάρχει μία τάση να κατηγοριοποιεί κάποιες ήττες ως ισοπαλίες και κάποιες ισοπαλίες ως νίκες, ωστόσο αυτό δεν επηρεάζει σε μεγάλο βαθμό την αποτελεσματικότητά του.

	Ήττα	**Ισοπαλία**	**Νίκη**
Ήττα	76	8	1
Ισοπαλία	12	30	10
Νίκη	0	9	76

Πίνακας 3.7: Αποτελέσματα SVM για το σύνολο των χαρακτηριστικών

Τα αποτελέσματα είναι εξαιρετικά αν αξιοποιηθεί το μοντέλο που περιέχει τις μεταβλητές που προέκυψαν μέσω του αλγορίθμου Feature Selection. Ο αλγόριθμος έδωσε μια ακρίβεια 99.54%, η οποία είναι ιδιαίτερα υψηλή.

	Ήττα	**Ισοπαλία**	**Νίκη**
Ήττα	84	1	0
Ισοπαλία	0	52	0
Νίκη	0	0	85

Πίνακας 3.8: Αποτελέσματα SVM για τα χαρακτηριστικά της μεθόδου Feature Selection

Όπως και στην περίπτωση του Naïve Bayes, έτσι και σε αυτήν την περίπτωση εξετάστηκε, λόγω της μεγάλης επιρροής που έχουν οι μεταβλητές Goals και GoalsCon στα αποτελέσματα, αφού η εξαρτημένη μεταβλητή είναι η Result, η απουσία αυτών των δύο μεταβλητών. Τα αποτελέσματα έδειξαν περίπου την ίδια εικόνα, αφού το 61.71% των παρατηρήσεων κατηγοριοποιήθηκε ορθά, δείχνοντας έτσι ότι η εισαγωγή στο μοντέλο των υπόλοιπων μεταβλητών είναι ουσιαστική.

	Ήττα	Ισοπαλία	Νίκη
Ήττα	68	1	16
Ισοπαλία	22	1	29
Νίκη	17	0	68

Πίνακας 3.9: Αποτελέσματα SVM για τα χαρακτηριστικά της μεθόδου Feature Selection, εξαιρουμένων των μεταβλητών Goals και GoalsCon

Τέλος, ένα επιπλέον καλό δείγμα γραφής στη χρήση του SVM στα συγκεκριμένα δεδομένα είναι και το γεγονός ότι στην περίπτωση όπου το μοντέλο που περιέχει μόνο τις μεταβλητές Goals και GoalsCon, το ποσοστό ορθών κατηγοριοποιήσεων είναι ίσο με 99.54%. Προέκυψε, δηλαδή, μία λάθος κατανομή. Ωστόσο, αυτό το μοντέλο δε θα χρησιμοποιηθεί, καθώς δεν προσφέρει καμία ουσιαστική ερμηνεία στα δεδομένα.

	Ήττα	Ισοπαλία	Νίκη
Ήττα	84	0	1
Ισοπαλία	0	52	0
Νίκη	0	0	85

Πίνακας 3.10: Αποτελέσματα SVM για το μοντέλο που περιέχει μόνο τις μεταβλητές Goals και GoalsCon

Κλείνοντας αυτή την ενότητα, μπορεί με βεβαιότητα να τονισθεί ότι ο συγκεκριμένος αλγόριθμος είναι αρκετά πιο αποδοτικός από τον αλγόριθμο Naïve Bayes και αυτό φαίνεται συγκρίνοντας όλες τις προσαρμογές των μοντέλων που έγιναν. Η ακρίβεια στην κατηγοριοποίηση που έγινε σε κάθε περίπτωση ήταν μεγαλύτερη για τον SVM. Συνεπώς, το συμπέρασμα που προκύπτει είναι ότι ο συγκεκριμένος αλγόριθμος είναι κατάλληλος για την εξόρυξη γνώσης από δεδομένα που αφορούν το ποδόσφαιρο. Επιπλέον, η αποτελεσματική του χρήση επιβεβαίωσε ως αποτελεσματική και την επιλογή χαρακτηριστικών που έγινε στην περίπτωση του αλγορίθμου Random Forest. Ωστόσο, αυτή η επιβεβαίωση έγινε μόνο για την περίπτωση που εξαρτημένη μεταβλητή ήταν η Result. Συνεπώς, η συνέχεια πρέπει να περιλαμβάνει και κάτι ανάλογο για την Goals, που είναι η δεύτερη μεταβλητή που μπορεί να χρησιμοποιηθεί ως εξαρτημένη.

Με βάση το σύνολο των αποτελεσμάτων τόσο από τη χρήση του αλγορίθμου Feature Selection, όσο και από την κατηγοριοποίηση για την μεταβλητή Result, προκύπτει ότι οι μεταβλητές που δείχνουν να επηρεάζουν τις τιμές των δύο εξαρτημένων μεταβλητών είναι αυτές που φαίνονται στον παρακάτω πίνακα. Για την Result, από την τελική επιλογή εξαιρέθηκαν οι Goals και GoalsCon, λόγω του γεγονότος ότι φαίνεται να είναι τετριμμένη η επιλογή τους. Για την Goals, εξαιρέθηκε για τους ίδιους λόγους, η Result, όπως και λόγω ότι αποτελεί την δεύτερη εξαρτημένη μεταβλητή.

Για την μεταβλητή Result	**Για την μεταβλητή Goals**
OnTarget	OnTarget
Tattempts	Tattempts
Ranking	PassAcc
Clearances	Ranking
	Possession
	Distance
	Clearances

Πίνακας 3.11: Επιλογή μεταβλητών με βάση τις μεθόδους του κεφαλαίου 3.

Η αποτελεσματική χρήση των παραπάνω μεθόδων μπορεί να οδηγήσει τους μελλοντικούς αναλυτές, αλλά και τους ποδοσφαιρικούς συλλόγους σε βαθύτερα συμπεράσματα, αν αυτά συνδυαστούν με τα κατάλληλα δεδομένα.

ΚΕΦΑΛΑΙΟ 4

ΠΡΟΣΑΡΜΟΓΗ ΓΕΝΙΚΕΥΜΕΝΩΝ ΓΡΑΜΜΙΚΩΝ ΜΟΝΤΕΛΩΝ

ΕΙΣΑΓΩΓΗ

Στο προηγούμενο κεφάλαιο χρησιμοποιήθηκαν μέθοδοι που αφορούσαν το πεδίο της μηχανικής μάθησης, με απώτερο σκοπό να γίνει ένα βασικό βήμα στην απάντηση των βασικών ερωτημάτων της παρούσας μελέτης. Τα αποτελέσματα που προέκυψαν έδωσαν εκτός από τους παράγοντες εκείνους που επηρεάζουν την απόδοση των ομάδων και έναν «οδηγό» για μελλοντικούς αναλυτές, αναφορικά με τις κατάλληλες μεθόδους ανάλυσης δεδομένων που αφορούν το ποδόσφαιρο. Ωστόσο, όπως είναι γενικότερα γνωστό για τις μεθόδους αυτές, δεν μπορεί να προκύψει κάποιο ποσοτικό μέτρο αναφορικά με τη σημαντικότητά τους και κατά συνέπεια χάνουν σε ακρίβεια, με αποτέλεσμα να υπάρχει έλλειψη στο κομμάτι της στατιστικής συμπερασματολογίας, παρόλη την πολύτιμη βοήθειά τους στο κομμάτι της πρόβλεψης.

Όπως είναι ευρύτερα γνωστό, 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) (Κούτρας, 2019). Συνήθως, όταν αναφερόμαστε στην ανάλυση παλινδρόμησης εννοούμε την γραμμική παλινδρόμηση, η οποία προβλέπει τη συμπεριφορά μίας μεταβλητής (εξαρτημένη μεταβλητή ή μεταβλητή απόκρισης) μέσω των μετρήσεων που έχουμε πάρει για μια (απλή γραμμική παλινδρόμηση) ή περισσότερες (πολλαπλή γραμμική παλινδρόμηση) μεταβλητές (ανεξάρτητες μεταβλητές). Το μοντέλο της πολλαπλής παλινδρόμησης συνοψίζεται στην παρακάτω σχέση:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, i = 1, \dots, \nu$$

όπου 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. Ωστόσο, βασική προϋπόθεση για την εφαρμογή του συγκεκριμένου μοντέλου είναι η κανονικότητα των σφαλμάτων και κατ' επέκταση της εξαρτημένης μεταβλητής, πράγμα το οποίο όπως έχει φανεί στο προηγούμενο κεφάλαιο παραβιάζεται και στις δύο περιπτώσεις που έχουν τεθεί ως εξαρτημένες.

Συγκεκριμένα, βλέπουμε ότι η μεταβλητή Result είναι μία κατηγορική μεταβλητή 3 τιμών και η Goals παίρνει διακριτές τιμές. Συνεπώς, μπορεί να εξεταστεί η προσαρμογή Γενικευμένων Γραμμικών Μοντέλων (ΓΓΜ) με εξαρτημένες τις παραπάνω μεταβλητές. Στα συγκεκριμένα μοντέλα η μεταβλητή απόκρισης εισέρχεται στο μοντέλο μέσω της μέσης της τιμής

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

Με άλλα λόγια, δεν γίνεται καμία υπόθεση για την κατανομή των σφαλμάτων του μοντέλου παρά μόνο για την κατανομή της Y (Πολίτης, 2020). Η συνάρτηση $g(\mu_i)$ ονομάζεται συνάρτηση σύνδεσης και συνδέει τη μέση τιμή της εξαρτημένης μεταβλητής μ_i με το σταθερό κομμάτι του μοντέλου.

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, συνεπώς έχουν μια σειρά από επιθυμητές ιδιότητες.

Στη συνέχεια του κεφαλαίου θα παρουσιαστεί μία ανάλυση με βάση τα ΓΓΜ, η οποία θα βασιστεί πάνω στην εικόνα που υπάρχει για τις δύο μεταβλητές απόκρισης (Goals και Result), όπως επίσης και μία λεπτομερής συμπερασματολογία των αποτελεσμάτων της. Αυτό σε συνδυασμό με τα αποτελέσματα του προηγούμενου κεφαλαίου. Για τις ανάγκες της συγκεκριμένης μελέτης, χρειάστηκε να μετατρέψουμε την μεταβλητή Result σε μία νέα μεταβλητή, την Result2, η οποία, όπως και πριν, παίρνει την τιμή 0 στην περίπτωση της ήττας, αλλά παίρνει την τιμή 1 στην περίπτωση ισοπαλίας ή νίκης.

Προτού γίνει η ανάλυση και η εξήγηση των δύο μοντέλων της οικογένειας των ΓΓΜ που θα αξιοποιηθούν, θα προηγηθεί ένα πρώτο βήμα που κρίνεται απαραίτητο να εφαρμοστεί με βάση τα προηγούμενα και για λόγους αξιοπιστίας των αποτελεσμάτων.

4.1 Έλεγχος Πολυσυγγραμμικότητας

Πολλές φορές στα δεδομένα που χρησιμοποιούνται για διάφορες εφαρμογές, παρατηρείται το φαινόμενο οι ανεξάρτητες μεταβλητές που χρησιμοποιούνται σε μοντέλα παλινδρόμησης να είναι μεταξύ τους υψηλά συσχετισμένες. Αυτό είναι ένα φαινόμενο το οποίο πρέπει να κάνει τους αναλυτές ιδιαίτερα δύσπιστους ως προς τα αποτελέσματα και την εξαγωγή συμπερασμάτων που προκύπτουν από την προσαρμογή ενός μοντέλου, καθώς πιθανές υψηλές συσχετίσεις αλλάζουν σε μεγάλο βαθμό την ερμηνεία που δίνει ένα μοντέλο, οδηγώντας σε λανθασμένα συμπεράσματα.

Στην περίπτωση που κάποιες ανεξάρτητες μεταβλητές ενός μοντέλου πολλαπλής παλινδρόμησης έχουν πολύ μεγάλη συσχέτιση μεταξύ τους, τότε λέμε ότι υπάρχει πολυσυγγραμμικότητα (Κούτρας, 2019). Η συνηθέστερη περίπτωση αναφοράς περιπτώσεων

πολυσυγγραμμικότητας αφορά τα μοντέλα πολλαπλής γραμμικής παλινδρόμησης, ωστόσο η μελέτη της μπορεί να βοηθήσει και στα Γενικευμένα Γραμμικά Μοντέλα. Συνεπώς, στα πλαίσια της παρούσας μελέτης θα γίνει πρωτίστως ένας έλεγχος πολυσυγγραμμικότητας ανάμεσα για το σύνολο των ανεξάρτητων μεταβλητών που αφορούν τα ΓΓΜ που θα χρησιμοποιηθούν, έτσι ώστε να γίνει μία πρώτη διαλογή για την κατάρτιση των μοντέλων.

Ο δείκτης που χρησιμοποιείται ευρέως ως κριτήριο για την ύπαρξη πολυσυγγραμμικότητας είναι ο παράγοντας διόγκωσης διακύμανσης (Variance Inflation Factor – VIF). Θεωρώντας ότι το μοντέλο μας έχει p ανεξάρτητες μεταβλητές, ο παράγοντας αυτός για κάθε μεταβλητή ορίζεται ως εξής:

$$VIF_k = \frac{1}{1 - R_k^2}, k = 1,2, \dots, p - 1$$

όπου R_k^2 είναι ο συντελεστής προσδιορισμού του μοντέλου που χρησιμοποιεί σαν εξαρτημένη μεταβλητή την k και ως ανεξάρτητες τις υπόλοιπες. Για το σύνολο του δείγματος χρησιμοποιείται ο μέσος όρος του παραπάνω αποτελέσματος

$$\overline{VIF} = \frac{1}{p - 1}\sum_{k=1}^{p-1} VIF_k$$

Από τα παραπάνω γίνεται αντιληπτό ότι, στην περίπτωση ύπαρξης συσχέτισης ανάμεσα στην εξεταζόμενη μεταβλητή και τις υπόλοιπες ανεξάρτητες μεταβλητές, ο συντελεστής προσδιορισμού πλησιάζει το 1, που σημαίνει ότι ο $\overline{VIF}$ γίνεται πολύ μεγάλος. Το ανάποδο ισχύει σε περίπτωση μη ύπαρξης συσχέτισης. Συνεπώς, στην περίπτωση μη ύπαρξης πολυσυγγραμμικότητας ο $\overline{VIF}$ πλησιάζει το 1. Ο γενικότερος κανόνας που ισχύει για τον VIF_k είναι ότι στην περίπτωση που $VIF_k > 10$, τότε υπάρχει πρόβλημα πολυσυγγραμμικότητας. Ωστόσο, σε πιο αδύναμα μοντέλα όπως τα ΓΓΜ, το πρόβλημα παρατηρείται όταν $VIF_k > 2.5$ (Senaviratna et al., 2019).

Στη φάση της παρούσας μελέτης η μεθοδολογία που ακολουθείται αφορά την προσαρμογή των πλήρων ΓΓΜ που θα χρησιμοποιηθούν με μεταβλητές απόκρισης τις Goals και Result2 και στη συνέχεια της χρήσης μέσω της R της συνάρτησης vif του πακέτου DAAG.

Για το έλεγχο πολυσυγγραμμικότητας στο μοντέλο με εξαρτημένη την μεταβλητή Result2 προκύπτουν τα παρακάτω αποτελέσματα:

```
> vif(model11)
  OnTarget  Tattempts    Blocked    Corners  Possesion    PassAcc      PassT
    2.2579     4.6601     2.5860     1.7815     6.5681     6.5221   177.7200
     PassC   Distance   BallsRec    Tackles     Blocks     Yellow        Red
  198.6100     1.2860     1.3638     1.1035     1.3052     1.3551     1.1317
  FoulsCom Home.Away1     Round1      Group    Ranking Clearances
    1.3821     1.2316     1.2341     3.8459     3.7265     2.1126
```

Πίνακας 4.1: Αποτελέσματα R για τον παράγοντα VIF στο πλήρες μοντέλο με εξαρτημένη την μεταβλητή Result2.

Όπως είναι φανερό, τα παραπάνω αποτελέσματα έδωσαν πολύ μεγάλες τιμές στον συντελεστή VIF_k για τις μεταβλητές PassT και PassC. Αυτό συνδυαστικά με τα αποτελέσματα του κεφαλαίου 2 είναι απόλυτα φυσιολογικό, καθώς παρατηρήθηκαν υψηλές συσχετίσεις ανάμεσα σε αυτές τις δύο μεταβλητές και στις PassAcc και Possession. Συνεπώς, η παρουσία των μεταβλητών PassT και PassC δημιουργεί πρόβλημα πολυσυγγραμμικότητας στο μοντέλο, με τα ανάλογα προβλήματα στην ερμηνεία του. Οπότε θα χρειαστεί να αφαιρεθούν πριν γίνει η ανάλογη μελέτη των πιο κατάλληλων μεταβλητών. Για τις μεταβλητές PassAcc και Possession, όπως και για τις υπόλοιπες, όπου είχαμε $VIF_k > 2.5$, παρόλο που εγείρουν έναν προβληματισμό ως προς την ύπαρξη πολυσυγγραμμικότητας, το γεγονός ότι οι τιμές είναι πολύ κοντά στο όριο του 2.5 μας κάνει να αναμένουμε μέχρι την τελική επιλογή μεταβλητών για την τελική απόφαση. Οπότε, η κατάρτιση του τελικού μοντέλου θα επιφέρει και έναν τελικό έλεγχο.

Ανάλογα αποτελέσματα παίρνουμε και για το πλήρες μοντέλο με μεταβλητή απόκρισης την Goals. Αν και παρατηρούνται υψηλότερες τιμές στα αποτελέσματα ανά μεταβλητή, το γενικότερο συμπέρασμα δεν αλλάζει. Έτσι και σε αυτή την περίπτωση, προτού περάσουμε στην επιλογή των κατάλληλων μεταβλητών, θα χρειαστεί να αφαιρέσουμε τις μεταβλητές PassT και PassC, καθώς παρατηρείται το φαινόμενο της πολυσυγγραμμικότητας.

```
> vif(model21)
  OnTarget  Tattempts    Blocked    Corners  Possesion    PassAcc      PassT
    3.7344     7.0404     2.6186     1.7396     6.8211     7.2281   211.0300
     PassC   Distance   BallsRec    Tackles     Blocks     Yellow        Red
  236.5500     1.2428     1.2950     1.0642     1.3983     1.4260     1.1297
  FoulsCom Home.Away1     Round1      Group    Ranking Clearances
    1.3966     1.1935     1.2002     3.9559     3.8623     2.1320
```

Πίνακας 4.2: Αποτελέσματα R για τον παράγοντα VIF στο πλήρες μοντέλο με εξαρτημένη την μεταβλητή Goals.

4.2 Εφαρμογή Λογιστικής Παλινδρόμησης για το αποτέλεσμα μιας ομάδας

Η λογιστική παλινδρόμηση είναι η πιο βασική και συνηθισμένη μορφή ΓΓΜ που αξιοποιείται σε διάφορες εφαρμογές, όπως η υγεία, η οικονομία κλπ. Αφορά δεδομένα που η εξαρτημένη τους μεταβλητή είναι μία μεταβλητή που λαμβάνει 2 τιμές, όπου η μία θεωρείται επιτυχία και η άλλη αποτυχία. Συνεπώς, η υπόθεση που αφορά τη λογιστική παλινδρόμηση ξεκινάει από το γεγονός ότι η μεταβλητή μας ακολουθεί την κατανομή Bernoulli. Οπότε, η μέση της τιμή αφορά την πιθανότητα επιτυχίας. Η συνάρτηση σύνδεσης που θα χρησιμοποιηθεί σε αυτή τη μελέτη είναι η logit, η οποία εκφράζεται με τον παρακάτω τύπο:

$$\log\left[\frac{p_i}{1-p_i}\right] = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

Ο λόγος $\frac{p_i}{1-p_i}$ ονομάζεται σχετική πιθανότητα (odds). Εκτός από την logit υπάρχουν και άλλες δύο συναρτήσεις σύνδεσης, η probit, η οποία έχει παρόμοια συμπεριφορά με την logit και η clog – log.

Όπως έχει αναφερθεί σε προηγούμενες ενότητες, η μεταβλητή που αφορά το αποτέλεσμα που είχε μία ομάδα – παρατήρηση σε έναν αγώνα εκφράζεται με τη μορφή της μεταβλητής Result, μιας κατηγορικής μεταβλητής, η οποία λαμβάνει 3 τιμές, ανάλογα με το αποτέλεσμα του αγώνα. Ωστόσο, όπως ειπώθηκε και προηγουμένως, η υπόθεση με την οποία μελετάται ένα μοντέλο λογιστικής παλινδρόμησης είναι ότι η εξαρτημένη μεταβλητή είναι μία μεταβλητή Bernoulli. Συνεπώς, για να πληροί η μεταβλητή Result αυτές τις προδιαγραφές, έπρεπε να γίνει η μετατροπή της σε Result2.

Ξεκινώντας την ανάλυση και λαμβάνοντας υπόψη τα αποτελέσματα του ελέγχου πολυσυγγραμμικότητας που προέκυψαν στο προηγούμενο κεφάλαιο, χρησιμοποιείται ένα από τα πλέον σημαντικά κριτήρια για την εύρεση του κατάλληλου μοντέλου που περιγράφει καλύτερα την μεταβλητή απόκρισης, αυτό 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. Για να επιλέξουμε ανάμεσα σε διαφορετικά μοντέλα με βάση αυτό το κριτήριο, «βέλτιστο» μοντέλο θεωρείται αυτό που έχει τη μικρότερη τιμή AIC.

Ο απλούστερος τρόπος για να γίνει αυτό μέσω της R είναι με τη χρήση της συνάρτησης step, η οποία έχει σαν βασικό όρισμα το μοντέλο που εξετάζεται. Η προεπιλογή αναφορικά με την κατεύθυνση που θα χρησιμοποιηθεί (προς τα πίσω αποκλεισμός ή προς τα εμπρός επιλογή των παραμέτρων) είναι αυτή που χρησιμοποιεί και τις δύο περιπτώσεις (both). Με βάση αυτά τα αποτελέσματα (βλ. Παράρτημα Γ), προέκυψε ότι καταλληλότερο μοντέλο είναι αυτό που σαν ανεξάρτητες μεταβλητές έχει τις OnTarget, Corners, Possession, BallsRec, Yellow, FoulsCom, Home.Away, Round, Group, Ranking, Clearances.

Μπορεί το κριτήριο AIC να έχει μια ευρεία χρήση και να είναι ένας εύκολος και άμεσος τρόπος για την κατάρτιση αξιόπιστων μοντέλων, ωστόσο σε καμία περίπτωση δεν πρέπει να θεωρείται ότι ολοκληρώνει τη μελέτη του συνόλου δεδομένων. Με βάση το μοντέλο που καταρτίστηκε από το κριτήριο, χρησιμοποιήθηκε η συνάρτηση summary(), η οποία στην περίπτωση των ΓΓΜ έχει ακριβώς την ίδια λειτουργία με την περίπτωση της γραμμικής παλινδρόμησης. Δηλαδή, την απεικόνιση των τιμών των παραμέτρων του μοντέλου, όπως και των βασικών μετρήσεων που αναδεικνύουν τη στατιστική σημαντικότητα ολόκληρου του μοντέλου και της κάθε ανεξάρτητης μεταβλητής.

Αρχικά, το βασικό σημείο που επικεντρώνεται η ανάλυση είναι η στατιστική σημαντικότητα των μεταβλητών. Η συνάρτηση summary() εμφανίζει τα αποτελέσματα που προκύπτουν από τον έλεγχο του Wald. Ο έλεγχος του Wald προκύπτει από το γεγονός ότι οι εκτιμητές μέγιστης πιθανοφάνειας, όπως είναι και οι παράμετροι του μοντέλου, ακολουθούν ασυμπτωτικά την κανονική κατανομή (Agresti, 2002). Άρα, μπορεί να χρησιμοποιηθεί η κανονική κατανομή για τον έλεγχο υποθέσεων

$$H_0: \beta = 0 \; \acute{\epsilon}\nu\alpha\nu\tau\iota \; \tau\eta\varsigma \; H_1: \beta \neq 0$$

Η στατιστική συνάρτηση που χρησιμοποιείται είναι η

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

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

Κατά την εξέταση του μοντέλου που προέκυψε από το κριτήριο AIC, παρατηρήθηκε ότι η μεταβλητή Yellow είναι στατιστικά μη σημαντική, με το p-value του ελέγχου του Wald να είναι μεγαλύτερο του 0.05 (Πίνακας Γ.2). Αυτό το αποτέλεσμα συνάδει με τα αποτελέσματα

της περιγραφικής στατιστικής ανάλυσης του κεφαλαίου 2, που έδειξε ότι δεν υπήρχαν σημαντικές διαφορές για τις διάφορες τιμές της, οπότε η μεταβλητή αυτή θα εξαιρεθεί από το μοντέλο.

Η αλλαγή αυτή στις ανεξάρτητες μεταβλητές και ο εκ νέου υπολογισμός των παραμέτρων του μέσω της μεθόδου μέγιστης πιθανοφάνειας διαφοροποιεί την κατάσταση αναφορικά με τις τιμές των παραμέτρων, όπως και σε σχέση με τη στατιστική τους σημαντικότητα, καθώς η αλλαγή στην εκτίμηση της παραμέτρου επηρεάζει αντικειμενικά την τιμή της στατιστικής συνάρτησης του ελέγχου του Wald.

Στην περίπτωση αυτή, η προσαρμογή του μοντέλου που δεν περιέχει την μεταβλητή Yellow έδωσε σημαντικές αλλαγές στη σημαντικότητα των υπόλοιπων. Συγκεκριμένα, οι μεταβλητές FoulsCom και Group φαίνεται ότι είναι στατιστικά μη σημαντικές, με την πρώτη να έχει p-value ίσο με 0.085716 και τη δεύτερη ίσο με 0.053809 (Πίνακας Γ.3). Οι τιμές αυτές θεωρούνται οριακές, ωστόσο αν αυτές συνδυαστούν με τα αποτελέσματα των προηγούμενων μεθόδων, επιβεβαιώνεται το γεγονός ότι πρέπει να θεωρηθούν στατιστικά μη σημαντικές.

Η τελική μορφή του μοντέλου που επιλέγεται είναι αυτή που εξαιρεί τις δύο αυτές μεταβλητές. Είναι φανερό ότι στον έλεγχο του Wald, όλες οι υπόλοιπες μεταβλητές είναι στατιστικά σημαντικές.

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, 2020). 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.

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) και του προσαρμοσμένου μοντέλου, η οποία εκφράζεται από τον παρακάτω τύπο:

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

Για τη συγκεκριμένη ποσότητα γνωρίζουμε ότι ακολουθεί την κατανομή x_p^2, με $p = df_1 - df_2$. Συνεπώς, έχουμε ότι για ένα ΓΓΜ του οποίου η τιμή που προκύπτει από τον παραπάνω τύπο είναι στατιστικά σημαντική, έχουμε ισχυρή ένδειξη καλής προσαρμογής του μοντέλου, οπότε το μοντέλο μπορεί να αξιοποιηθεί για την ανάλυση των δεδομένων.

Παρόλα αυτά, για δίτιμα δεδομένα προτείνεται ο έλεγχος των Hosmer-Lemeshow, ο οποίος έχει σαν μηδενική υπόθεση την

$$H_0: \textit{Οι παρατηρηθείσες τιμές της } Y \textit{ δεν διαφέρουν από τις εκτιμώμενες τιμές.}$$

Η μέθοδος που χρησιμοποιεί ο έλεγχος έχει ως εξής: Οι παρατηρήσεις διατάσσονται με αύξουσα σειρά με βάση την προβλεπόμενη πιθανότητα επιτυχίας (την εκτίμηση που προκύπτει από το μοντέλο). Στη συνέχεια, η συγκεκριμένη διάταξη χωρίζεται σε g ομάδες, πράγμα το οποίο δίνει ίδιο περίπου αριθμό παρατηρήσεων. Η στατιστική συνάρτηση του ελέγχου είναι η

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n'_k \bar{\pi}_k)^2}{n'_k \bar{\pi}_k (1 - \bar{\pi}_k)},$$

όπου o_k οι παρατηρηθείσες τιμές της μεταβλητής απόκρισης, n'_k ο αριθμός των παρατηρήσεων κάθε ομάδας και $\bar{\pi}_k$ η μέση εκτιμημένη πιθανότητα. Η παραπάνω στατιστική συνάρτηση ακολουθεί την κατανομή x^2 με g-2 βαθμούς ελευθερίας (Hosmer D., Lemeshow S., 2000).

Αναφορικά με τον έλεγχο καλής προσαρμογής χρησιμοποιήθηκαν και οι δύο μέθοδοι, αφού καταλήξαμε στην επιλογή μεταβλητών μέσω του ελέγχου του Wald. Ακολουθώντας την ανάλογη διαδικασία στην R, αξιοποιήθηκε η συνάρτηση anova(), έτσι ώστε να δώσει τις μετρήσεις των αποκλίσεων για το μοντέλο όταν προστίθεται από μία μεταβλητή (από αυτές που έχουν επιλεχθεί) και για τη συνολική. Οι πρώτες αποτελούν και μία αντίστοιχη μέθοδο του ελέγχου του Wald, αφού στην ουσία εξετάζεται η διαφορά στην απόκλιση που φέρνει η μεταβλητή, άρα και η σημαντικότητά της. Έπειτα υπολογίστηκε και το αντίστοιχο p-value για τη διαφορά, το οποίο έδωσε μία τιμή ίση περίπου με 0.7393, πράγμα που σημαίνει ότι έχουμε καλή προσαρμογή του μοντέλου, αφού δεν διαφέρει σημαντικά από το κορεσμένο.

```
> anova(model15)
Analysis of Deviance Table

Model: binomial, link: logit

Response: Result2

Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev
NULL                         741     988.39
OnTarget    1  155.204       740     833.19
Possesion   1    5.288       739     827.90
BallsRec    1   20.325       738     807.58
Ranking     1   14.998       737     792.58
Clearances  1   59.870       736     732.71
Home.Away   1    9.313       735     723.40
Corners     1    3.551       734     719.85
Round       1   11.772       733     708.07
> 1-pchisq(708.07,733)
[1] 0.7392972
```

Πίνακας 4.3: Αποτελέσματα ελέγχου απόκλισης του επιλεγμένου μοντέλου.

Αντίστοιχα, ο έλεγχος των Hosmer-Lemenshow έδειξε ότι η μηδενική υπόθεση δεν απορρίπτεται, καθώς το p-value είναι ίσο με 0.9127. Συνεπώς, έχουμε και σε αυτή την περίπτωση μία πολύ ισχυρή ένδειξη καλής προσαρμογής του μοντέλου.

```
> hoslem.test(model152$y, fitted(model152))

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  model152$y, fitted(model152)
X-squared = 3.3201, df = 8, p-value = 0.9127
```

Πίνακας 4.4: Αποτελέσματα ελέγχου Hosmer-Lemenshow για το επιλεγμένο μοντέλο.

Ωστόσο, σε σχέση με την μεταβλητή Corners, τίθενται ορισμένα ζητήματα που πρέπει να απαντηθούν για να υπάρξει κατάληξη στο μοντέλο που αφορά την λογιστική παλινδρόμηση. Αρχικά, το γεγονός ότι έχουμε αρνητικό συντελεστή, που στην ερμηνεία του μοντέλου αυτό σημαίνει μείωση της εκτιμώμενης πιθανότητας, αυτό εξηγείται αν συνδυαστεί με τα αποτελέσματα του Κεφαλαίου 2, όπου βλέπουμε μέτρια αρνητική συσχέτιση με την μεταβλητή Clearances, πράγμα που αν τον συνδυάσουμε με την ασθενή συσχέτιση με την μεταβλητή Result, δείχνει ότι το αποτέλεσμα του μοντέλου είναι λογικό και μπορεί να εξηγηθεί. Επιπλέον ζήτημα, αν δούμε τα αποτελέσματα του μοντέλου αναφορικά με τον έλεγχο του Wald και με την απόκλιση που μας δίνει η μεταβλητή Corners, τίθενται στο ότι στην περίπτωση του ελέγχου του Wald τα αποτελέσματα δίνουν ένα p-value οριακά μικρότερο του 0.05 (περίπου 0.048), ενώ στην περίπτωση του ελέγχου της απόκλισης οριακά μεγαλύτερο (περίπου 0.06). Συνεπώς, δημιουργείται ζήτημα αναφορικά με την σημαντικότητα της μεταβλητής Corners.

Μία πιθανή αιτία ενδέχεται να αποτελεί η μετατροπή της Result σε Result2. Ο λόγος είναι ότι μπλέκονται μαζί με τις νίκες και οι ισοπαλίες, που όπως είδαμε και στο Κεφάλαιο 3 που χρησιμοποιήθηκαν οι αλγόριθμοι μηχανικής μάθησης, οι ισοπαλίες είχαν μεγάλο βαθμό τυχαιότητας, με αποτέλεσμα να μην υπάρχει σωστή κατηγοριοποίηση για τις περισσότερες εγγραφές που είχαν σαν αποτέλεσμα την ισοπαλία, με αποτέλεσμα να χάνεται μεγάλο ποσοστό ορθής κατηγοριοποίησης. Άρα, για να διαπιστώσουμε αν και κατά πόσο η μεταβλητή Corners είναι πράγματι στατιστικά σημαντική, πραγματοποιήθηκε η μετατροπή της Result σε Result3, η οποία παίρνει την τιμή 1 όταν το αποτέλεσμα είναι νίκη και την τιμή 0 σε αντίθετη περίπτωση. Στη συνέχεια, προσαρμόστηκε ένα μοντέλο με τις ίδιες μεταβλητές που επιλέχθηκαν στην περίπτωση της Result2 (Πίνακας Γ.5 - Παράρτημα Γ).

Από τα παραπάνω φαίνεται ότι το αποτέλεσμα που αφορά την μεταβλητή Corners είναι στατιστικά σημαντικό και στις δύο περιπτώσεις ελέγχων. Άρα, συνδυάζοντας τα παραπάνω, καταλήγουμε στο συμπέρασμα ότι η μεταβλητή Corners είναι στατιστικά σημαντική. Αναφορικά με κάποιες από τις παραπάνω που εμφανίζονται ως μη στατιστικά σημαντικές, αναφέρουμε ότι θα στηριχτούμε στο μοντέλο που είχε σαν μεταβλητή απόκρισης την Result2. Ο λόγος για την επιλογή αυτή είναι ότι με βάση τους κανονισμούς που ισχύουν στις διοργανώσεις της UEFA, αλλά και γενικότερα τη δομή των διοργανώσεων, πρέπει να θεωρηθεί σαν θετικό αποτέλεσμα η ισοπαλία. Σε αυτές τις διοργανώσεις τα ισόπαλα αποτελέσματα δίνουν εκτός από τον 1 βαθμό στη φάση των ομίλων και χρηματικό έπαθλο. Επιπλέον, με βάση τον κανονισμό του εκτός έδρας γκολ, οι ομάδες που αγωνίζονται εκτός έδρας έχουν συμφέρον από την επίτευξη ισοπαλίας σε σχέση με αυτές που αγωνίζονται εντός έδρας.

Ωστόσο, ακόμα και με αυτή την παραδοχή, χρειάζεται να εξεταστεί περαιτέρω η μεταβλητή Round. Από τα αποτελέσματα του model16 φαίνεται περίτρανα ότι η συγκεκριμένη μεταβλητή είναι μη σημαντική, σε σχέση με το model15, όπου ήταν σημαντική. Με βάση αυτό, η υπόθεση είναι ότι επηρεάζει η αλλαγή που έγινε από τη Result2 σε Result3. Δηλαδή, το γεγονός ότι στην πρώτη περίπτωση η ισοπαλία βρίσκεται μαζί με τις νίκες και στη δεύτερη μαζί με τις ήττες.

Όμως, στην περίπτωση της Result3, η οποία έχει αποκλειστικά τις νίκες σαν το επιτυχές αποτέλεσμα, φαίνεται ότι η μεταβλητή αυτή δεν αποτελεί σε καμία περίπτωση σημαντικό παράγοντα για το μοντέλο. Σε αυτό έρχονται να συμφωνήσουν και τα αποτελέσματα που έχουν προκύψει από τα προηγούμενα κεφάλαια. Συγκεκριμένα, είχαμε στο κεφάλαιο 1 τα θηκογράμματα που προέκυψαν δείχνουν ότι δεν υπάρχουν σημαντικές διαφορές ανά επίπεδο της Round. Συνδυάζοντάς τα με τους πίνακες Α.3 και Α.4 του παραρτήματος Α, φαίνεται η ταύτιση των αποτελεσμάτων. Στο κεφάλαιο 2 δεν βρέθηκε σημαντική συσχέτιση, παρά μόνο με την Group. Τέλος, στο κεφάλαιο 3 για την μεταβλητή Result φαίνεται ότι δεν είναι μέσα στις μεταβλητές που θεωρούνται σημαντικές για να επιτευχθεί ένα καλό ποσοστό κατηγοριοποίησης των δεδομένων. Συνεπώς, η μεταβλητή Round είναι στατιστικά μη σημαντική και το τελικό μοντέλο χωρίς αλληλεπιδράσεις είναι το model17.

```
> summary(model17)

Call:
glm(formula = Result2 ~ OnTarget + Possesion + BallsRec + Ranking +
    Clearances + Home.Away + Corners, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6821  -0.7971   0.3583   0.7800   1.9895

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.63515    0.96732  -7.893 2.95e-15 ***
OnTarget     0.43276    0.04773   9.067  < 2e-16 ***
Possesion    0.05770    0.01304   4.424 9.68e-06 ***
BallsRec     0.04392    0.01126   3.901 9.57e-05 ***
Ranking     -0.02127    0.00427  -4.982 6.30e-07 ***
Clearances   0.11882    0.01613   7.368 1.73e-13 ***
Home.Away1   0.62520    0.19318   3.236  0.00121 **
Corners     -0.06961    0.03675  -1.894  0.05825 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 988.39  on 741  degrees of freedom
Residual deviance: 719.85  on 734  degrees of freedom
AIC: 735.85

Number of Fisher Scoring iterations: 5
> anova(model17)
Analysis of Deviance Table

Model: binomial, link: logit

Response: Result2

Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev
NULL                         741     988.39
OnTarget    1  155.204       740     833.19
Possesion   1    5.288       739     827.90
BallsRec    1   20.325       738     807.58
Ranking     1   14.998       737     792.58
Clearances  1   59.870       736     732.71
Home.Away   1    9.313       735     723.40
Corners     1    3.551       734     719.85
```

Πίνακας 4.5: Αποτελέσματα model17

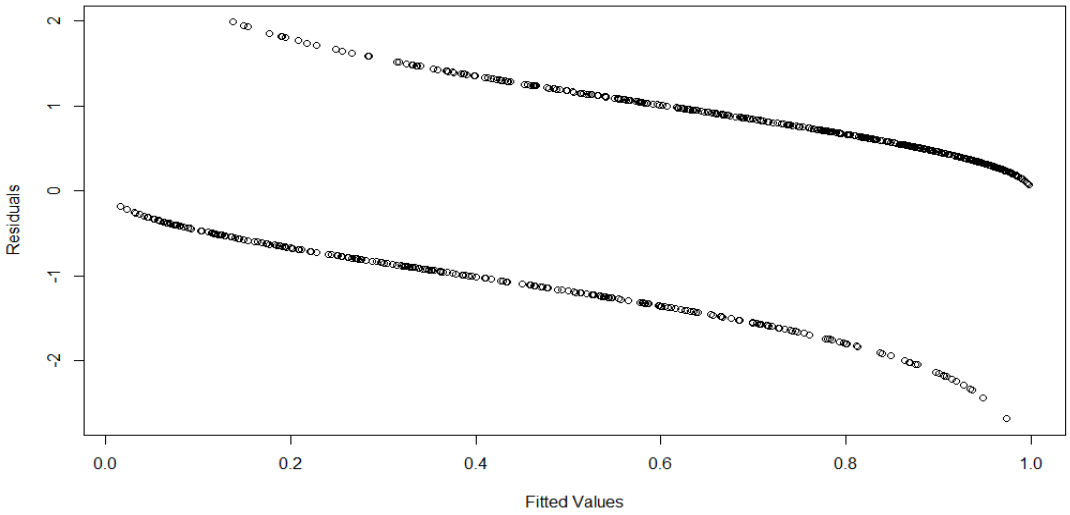
```
> hoslem.test(model172$y, fitted(model172))

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  model172$y, fitted(model172)
X-squared = 12.618, df = 8, p-value = 0.1257
```

Πίνακας 4.6: Αποτελέσματα ελέγχου Hosmer-Lemenshow για το επιλεγμένο μοντέλο.

Σε αντίθεση με τα μοντέλα γραμμικής παλινδρόμησης, στα ΓΓΜ υπάρχουν 4 είδη καταλοίπων. Στα πλαίσια της παρούσας μελέτης θα γίνει ανάλυση μέσω των καταλοίπων απόκλισης, που είναι αυτά που χρησιμοποιούνται ευρέως.



Διάγραμμα 4.1: Κατάλοιπα του μοντέλου λογιστικής παλινδρόμησης που επιλέχθηκε.

Σε αντίθεση με τα μοντέλα γραμμικής παλινδρόμησης, στη λογιστική παλινδρόμηση το διάγραμμα καταλοίπων 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 την τιμή 0. Συνεπώς, αφού το παραπάνω διάγραμμα έχει αυτή τη μορφή, συμπεραίνουμε ότι υπάρχει καλή προσαρμογή του μοντέλου.

4.3 Εφαρμογή μοντέλου παλινδρόμησης Poisson για τη μεταβλητή Goals

Η ανάλυση που έγινε στην προηγούμενη ενότητα με την εφαρμογή και εν τέλει την προσαρμογή ενός μοντέλου λογιστικής παλινδρόμησης με ανεξάρτητη μεταβλητή την Result2, μία μεταβλητή που αφορά διωνυμικά δεδομένα, δεν είναι η μοναδική περίπτωση ΓΓΜ. Η επόμενη σημαντική περίπτωση που αφορά την εκθετική οικογένεια κατανομών είναι η κατανομή Poisson. Όταν η κατανομή της εξαρτημένης μεταβλητής είναι η κατανομή Poisson, τότε μιλάμε για παλινδρόμηση Poisson.

Η παλινδρόμηση Poisson έχει σαν συνάρτηση σύνδεσης την

$$g(\mu_i) = \log(\mu_i)$$

συνεπώς το μοντέλο της παλινδρόμησης Poisson περιγράφεται από τον τύπο

$$log(\mu_i) = \beta_0 + \sum_{j=1}^{k} \beta_j \chi_{ij}$$

όπου μ_i η μέση τιμή της εξαρτημένης μεταβλητής.

Στην παρούσα μελέτη πραγματοποιήθηκε προσπάθεια για την προσαρμογή ενός μοντέλου Poisson βασιζόμενοι στα ίδια δεδομένα με την περίπτωση της λογιστικής παλινδρόμησης, έχοντας όμως σαν μεταβλητή απόκρισης την μεταβλητή Goals. Παλαιότερες μελέτες (Karlis D., Ntzoufras I., 2000) έχουν δείξει ότι η μεταβλητή που αφορά τα γκολ που επιτυγχάνονται ακολουθεί την κατανομή Poisson. Παρόλο που τα δεδομένα αυτά αφορούν κυρίως την τελική βαθμολογία του πρωταθλήματος, η παρούσα μελέτη θα λάβει υπόψη την συγκεκριμένη υπόθεση για δεδομένα που αφορούν μία διοργάνωση που απαρτίζεται και από τα δύο συστήματα διεξαγωγής (όμιλοι και νοκ-αουτ), όπως επίσης ότι τα δεδομένα αφορούν τις μετρήσεις ανά αγώνα.

Μετά τη διαδικασία για τον έλεγχο πολυσυγγραμμικότητας που πραγματοποιήθηκε σε προηγούμενη ενότητα, έγινε χρήση της εντολής step για την εύρεση του καταλληλότερου μοντέλου με βάση το κριτήριο AIC (βλ. Παράρτημα Δ). με βάση αυτό, το καταλληλότερο μοντέλο είναι αυτό με ανεξάρτητες μεταβλητές τις OnTarget, Tattempts, Ranking, Corners, Home.Away, Possession και Clearances. Προσαρμόζοντας αυτό το μοντέλο φαίνεται ότι υπάρχει στατιστικά σημαντική προσαρμογή στα δεδομένα.

Όπως φαίνεται και από τον παραπάνω πίνακα, όλες οι μεταβλητές είναι στατιστικά σημαντικές. Η κρίσιμη τιμή για την κατανομή χ^2 είναι $\chi^2_{1,0.95} = 3.841459$ και οι διαφορές που έχουν οι αποκλίσεις εισάγοντας κάθε μία μεταβλητή είναι μεγαλύτερες από αυτή. Συνεπώς, δεν χρειάζεται να προχωρήσουμε σε κάποια αλλαγή. Επιπλέον, προχωρώντας σε έναν έλεγχο καλής προσαρμογής του μοντέλου, υπολογίστηκε το p-value της συνολικής διαφοράς της απόκλισης του μοντέλου από το κορεσμένο, όπου ακολουθεί την κατανομή χ^2 και αυτό είναι περίπου ίσο με 0.985, οπότε μιλάμε για ένα επαρκές μοντέλο. Στο παράρτημα Γ (Πίνακας Γ.8) φαίνεται και ο πίνακας που προκύπτει από την συνάρτηση summary() για τις τιμές των παραμέτρων του μοντέλου.

Ωστόσο, το μοντέλο αυτό πέφτει σε μία αντίφαση. Η παράμετρος που αντιστοιχεί στην μεταβλητή Tattempts έχει αρνητική τιμή, με αποτέλεσμα κατά την ερμηνεία του μοντέλου ή κατά τον υπολογισμό μιας πρόβλεψης να καταλήγουμε στο συμπέρασμα ότι αν αυξηθεί κατά 1 η τιμή της Tattempts, τότε θα υπάρξει μείωση της προβλεπόμενης τιμής της μεταβλητής Goals. Αυτό έρχεται σε πλήρη αντίφαση με τα αποτελέσματα που είχαμε στο Κεφάλαιο 2 με τις συσχετίσεις, όπου η μεταβλητή Tattempts είχε (μέτρια) θετική συσχέτιση με την Goals

(Κεφάλαιο 2). Από τον ίδιο πίνακα παρατηρούμε ότι η ίδια μεταβλητή έχει ισχυρή συσχέτιση με την Tattempts. Αυτό είναι μία ισχυρή ένδειξη ότι παρουσιάζεται πολυσυγγραμμικότητα στο model23. Συνεπώς, έχουμε ότι πρέπει να αφαιρεθεί η συγκεκριμένη μεταβλητή από το μοντέλο.

Έτσι, προκύπτει το model24, που είναι και το τελικό μοντέλο. Εδώ παρατηρείται το εξής: Ενώ στον πίνακα summary με τις παραμέτρους προκύπτει ότι είναι όλες σημαντικές, δεν ισχύει το ίδιο και με τον πίνακα με τη διαφορά των αποκλίσεων. Επειδή όμως, στα ΓΓΜ παίζει ρόλο και η σειρά με την οποία εισάγουμε τις μεταβλητές, αλλάζοντας τη σειρά προκύπτει και η συμφωνία των δύο μεθόδων.

```
> summary(model24)

Call:
glm(formula = Goals ~ Corners + Home.Away + Possesion + OnTarget +
    Ranking + Clearances, family = poisson)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.34014  -1.06597  -0.09284   0.48827   2.70558

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.012070   0.274402  -3.688 0.000226 ***
Corners     -0.056619   0.011202  -5.054 4.32e-07 ***
Home.Away1   0.159563   0.063877   2.498 0.012491 *
Possesion    0.010151   0.004339   2.340 0.019300 *
OnTarget     0.186240   0.010588  17.590  < 2e-16 ***
Ranking     -0.005613   0.001632  -3.439 0.000584 ***
Clearances   0.014075   0.004974   2.830 0.004657 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1095.25  on 741  degrees of freedom
Residual deviance:  663.98  on 735  degrees of freedom
AIC: 2037.9

Number of Fisher Scoring iterations: 5
> anova(model24)
Analysis of Deviance Table

Model: poisson, link: log

Response: Goals

Terms added sequentially (first to last)
           Df Deviance Resid. Df Resid. Dev
NULL                         741    1095.25
Corners     1   15.574       740    1079.68
Home.Away   1   22.498       739    1057.18
Possesion   1   67.057       738     990.13
OnTarget    1  306.845       737     683.28
Ranking     1   11.399       736     671.88
Clearances  1    7.900       735     663.98

> 1-pchisq(663.98,735) #Ελεγχος επάρκειας
[1] 0.9711572
```

Πίνακας 4.7: Πίνακας παραμέτρων και διαφοράς αποκλίσεων για το τελικό μοντέλο Poisson χωρίς αλληλεπιδράσεις.

Όπως φαίνεται και από τον πίνακα, επιβεβαιώνεται και η επάρκεια του μοντέλου αφού το p-value του ζητούμενου ελέγχου είναι περίπου 0.971.

4.4 Έλεγχος αλληλεπιδράσεων

Στις προηγούμενες ενότητες έγινε μια προσπάθεια για να προσαρμοστούν δύο μοντέλα, με γνώμονα τη βασική θεωρία που διέπει τα ΓΓΜ, τα οποία θα αποτελέσουν τον καταλύτη για την εξαγωγή συμπερασμάτων αναφορικά με τους βασικούς παράγοντες που επηρεάζουν την απόδοση μιας ποδοσφαιρικής ομάδας. 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.

Σε γενικές γραμμές, όταν λέμε ότι δύο μεταβλητές Α και Β αλληλεπιδρούν ως προς μία τρίτη μεταβλητή Υ, εννοούμε ότι η επίδραση της μεταβλητής Α ως προς την Υ εξαρτάται από το επίπεδο της μεταβλητής Β (Πολίτης, 2020). Αντίστοιχα, αυτός ο ορισμός γενικεύεται και σε περιπτώσεις όπου μιλάμε για παραπάνω από δύο μεταβλητές.

Στην παρούσα μελέτη εξετάστηκαν όλοι οι συνδυασμοί αλληλεπιδράσεων μεταξύ των μεταβλητών που κρίθηκαν ότι επηρεάζουν την απόδοση ενός ποδοσφαιρικού αγώνα, αναφορικά και με τα δύο μοντέλα που κρίθηκαν κατάλληλα ως προς τις κύριες επιδράσεις τους. Οι μέθοδοι που χρησιμοποιήθηκαν ήταν οι ίδιοι με αυτές που χρησιμοποιήθηκαν για την επιλογή των κατάλληλων μεταβλητών – κύριων επιδράσεων. Ωστόσο, η τελική επιλογή των αλληλεπιδράσεων και των μεταβλητών κρίθηκε από δύο σημαντικούς παράγοντες.

Ο πρώτος είναι ότι με βάση έναν κανόνα (Mundry R., 2014), που σε πολλές δημοσιεύσεις ονομάζεται και αρχή της συνέπειας, όταν βρίσκουμε μία στατιστικά σημαντική αλληλεπίδραση, τότε στο μοντέλο πρέπει να περιέχονται οι μεταβλητές που απαρτίζουν την συγκεκριμένη αλληλεπίδραση, όπως επίσης και οι αλληλεπιδράσεις μικρότερης τάξης από τις ίδιες μεταβλητές. Έτσι, αν π.χ. έχουμε μία στατιστικά σημαντική αλληλεπίδραση ανάμεσα σε 3 μεταβλητές, τότε το μοντέλο πρέπει οπωσδήποτε να περιέχει τις 3 κύριες επιδράσεις, όπως και τις 3 αλληλεπιδράσεις ανάμεσα σε 2 μεταβλητές κάθε φορά.

Ο δεύτερος παράγοντας είναι το γεγονός ότι ένα στατιστικό μοντέλο πρέπει σε μεγάλο βαθμό να μπορεί να δώσει στην εκάστοτε ανάλυση μία ερμηνεία για το πως διαμορφώνονται τα δεδομένα που μελετώνται, με απώτερο σκοπό να χρησιμοποιηθεί για την επίτευξη των στόχων του εκάστοτε εγχειρήματος. Με βάση αυτό, για να συμπεριλάβουμε μία αλληλεπίδραση στο μοντέλο πρέπει σε μεγάλο βαθμό να μπορούμε να την ερμηνεύσουμε και να βγάλουμε τα αντίστοιχα συμπεράσματα.

Οι παραπάνω δύο παράγοντες μας οδηγούν στο συμπέρασμα ότι για την επιλογή του κατάλληλου μοντέλου με αλληλεπιδράσεις, η μελέτη μόνο των αλληλεπιδράσεων που αφορούν δύο μεταβλητές είναι εκείνη που μπορεί να δώσει την καλύτερη ερμηνεία στα δεδομένα μας.

Ξεκινώντας από την μελέτη του πλήρους μοντέλου με μεταβλητή απόκρισης την Result2, αυτού δηλαδή που περιέχει τις τις δυνατές αλληλεπιδράσεις μεταξύ των 8 κύριων επιδράσεων που προέκυψαν από την λογιστική παλινδρόμηση, παρατηρήθηκε μία σημαντική αντίφαση στα αποτελέσματα. Με βάση τα αποτελέσματα που προέκυψαν από την χρήση του ελέγχου του Wald, έχουμε ότι δεν προκύπτει κάποια στατιστικά σημαντική αλληλεπίδραση ανά 2 κύριες επιδράσεις. Ωστόσο, με τη χρήση τις συνάρτησης anova() στην R, όπου τις δείχνει τα αποτελέσματα και τις διαφορές στην απόκλιση από την ένταξη στο μοντέλο μιας αλληλεπίδρασης, τα πορίσματα είναι τελείως διαφορετικά. Σε τέτοιες περιπτώσεις, σε μικρά ή μέτρια δείγματα, ο έλεγχος με τον λόγο πιθανοφανειών, δηλαδή τις διαφορές στην απόκλιση, είναι πιο αξιόπιστος (Agresti, 2002).

Η χρήση του πλήρους μοντέλου τις έδειξε ότι οι ανά 2 αλληλεπιδράσεις που δίνουν διαφορά στην απόκλιση μεγαλύτερη του 3.84159, που αυτό σημαίνει ότι στον έλεγχο με βάση την κατανομή χ^2 και με 1 βαθμό ελευθερίας θα δίνεται ένα p-value μικρότερο του 0.05, είχαν οι αλληλεπιδράσεις τις Home.Away με τις BallsRec και Clearances και η αλληλεπίδραση τις Ranking με την Clearances. Παρόλα αυτά, κατά την προσαρμογή του μοντέλου που περιέχει τις αλληλεπιδράσεις αυτές και τις κύριες επιδράσεις, η αλληλεπίδραση τις Ranking με την Clearances έδειξε ότι δεν προσφέρει κάτι ιδιαίτερο στο μοντέλο (Πίνακας Γ.9). Συνεπώς, το μοντέλο που προτιμήθηκε με βάση την λογιστική παλινδρόμηση για τη μελέτη τις μεταβλητής Result2 είναι αυτό που περιέχει τις 2 αλληλεπιδράσεις (Πίνακας Γ.10).

Όπως φαίνεται, ωστόσο, από τον παραπάνω πίνακα, το μοντέλο δίνει κάποια αποτελέσματα που προκαλούν προβληματισμό. Ενώ στο επιλεγμένο μοντέλο χωρίς αλληλεπιδράσεις model17 η μεταβλητή Home.Away έχει τυπική απόκλιση περίπου 0.19 και μια παράμετρο ίση περίπου με 0.62, στην προκειμένη περίπτωση βλέπουμε την τιμή της παραμέτρου να πηγαίνει στο 2.58 με μια τυπική απόκλιση στο 1.15, κάνοντας την εκτίμηση αυτή αρκετά αβέβαιη. Δεδομένου ότι σε περίπτωση που επιλεχθούν οι συγκεκριμένες αλληλεπιδράσεις θα χρειαζόταν να παραμείνει η κύρια επίδραση, σημαίνει ότι το μοντέλο αυτό δε θα ευσταθούσε. Το ίδιο ισχύει και στην περίπτωση που κρατήσουμε την μία από τις δύο αλληλεπιδράσεις. Οπότε, καταλήγουμε στο συμπέρασμα να μην αξιοποιήσουμε τις δύο αυτές αλληλεπιδράσεις, παραμένοντας στην επιλογή του model17.

Για το μοντέλο Poisson με μεταβλητή απόκρισης την Goals, τα αποτελέσματα έδειξαν ότι η ανά δύο αλληλεπίδραση που είναι στατιστικά σημαντική με βάση την διαφορά στην απόκλιση είναι η αλληλεπίδραση της OnTarget με τη μεταβλητή Ranking. Ωστόσο, εδώ παρατηρείται το εξής πρόβλημα. Οι μεταβλητές αυτές είναι συνεχείς, με αποτέλεσμα να μην μπορούμε να προχωρήσουμε σε στατιστική συμπερασματολογία για την αλληλεπίδραση, αφού

αυτό προϋποθέτει την ύπαρξη διαφορετικών επιπέδων για μία μεταβλητή. Επιπλέον, όταν μιλάμε για μία αλληλεπίδραση μεταξύ δύο συνεχών μεταβλητών, μπορεί η αλληλεπίδραση αυτών των δύο να μη μεταφράζεται ως γραμμική συνάρτηση της μίας με την άλλη, όπως είναι συνήθως (Jaccard J., Turrisi R., 2003).

Με βάση την παραπάνω εργασία, ένας τρόπος για να λυθεί αυτό το πρόβλημα είναι η δημιουργία 5-10 γκρουπ, με σχεδόν ίδιο αριθμό παρατηρήσεων, για τουλάχιστον μία από τις δύο μεταβλητές.

Λόγω του γεγονότος ότι η αλληλεπίδραση που είναι στατιστικά σημαντική αφορά την μεταβλητή OnTarget, πραγματοποιήθηκε διακριτοποίησή της με τρεις ξεχωριστούς τρόπους. Σκοπός είναι να φτάσουμε όσο πιο κοντά γίνεται στο αρχικό μοντέλο, με τις 2 αλληλεπιδράσεις που εντοπίστηκε ότι δίνουν στατιστικά σημαντική διαφορά στην απόκλιση του μοντέλου.

Ο πρώτος αφορούσε τον διαχωρισμό σε 5 γκρουπ, αφού έχουμε προχωρήσει σε διαδικασία κεντρικοποίησης της μεταβλητής OnTarget. Με βάση αυτό, υπολογίστηκε ο μέσος και η τυπική απόκλιση και δημιουργήθηκε η μεταβλητή OnTarget2 με βάση τον παρακάτω τύπο:

$$OnTarget2 = \frac{OnTarget - Mean}{Standard\ Deviation}$$

Η νέα κατηγορική μεταβλητή OnTarget_Cat λαμβάνει την τιμή 0, όταν η OnTarget2 παίρνει τιμές μικρότερες ή ίσες του -0.9437089, την τιμή 1 όταν λαμβάνει τιμές μεγαλύτερες του -0.9437089 και μικρότερες ή ίσες του -0.239974, την τιμή 2 όταν λαμβάνει τιμές μεγαλύτερες του -0.239974 και μικρότερες ή ίσες του 0.1118939, την τιμή 3 όταν λαμβάνει τιμές μεγαλύτερες του 0.1118939και μικρότερες ή ίσες του 0.8156291 και την τιμή 4 όταν λαμβάνει τιμές μεγαλύτερες του 0.8156291.

Στη συνέχεια, υπολογίστηκε το μοντέλο παλινδρόμησης Poisson που έχει ως μεταβλητή απόκρισης την Goals, την OnTarget_Cat (αντί της OnTarget) και τις υπόλοιπες ανεξάρτητες μεταβλητές. Τα αποτελέσματα που παρουσιάζονται στον πίνακα Γ.11 (Παράρτημα Γ) δείχνουν ότι το μοντέλο δεν βρίσκεται πολύ κοντά στα αποτελέσματα του μοντέλου με τις συνεχείς μεταβλητές, αφού βγαίνει στατιστικά μη σημαντικό αποτέλεσμα τόσο για την κύρια επίδραση όσο και για την αλληλεπίδραση με την Ranking, αφού για τους δύο βαθμούς ελευθερίας η κρίσιμη τιμή της χ^2 είναι περίπου 9.5, ενώ η διαφορά στις αποκλίσεις για την αλληλεπίδραση που συζητάμε είναι 3.02. Με βάση και τον έλεγχο του Wald προκύπτει ότι η αλληλεπίδραση δεν είναι στατιστικά σημαντική.

Οι επόμενοι δύο τρόποι δεν ξεφεύγουν πολύ από τον παραπάνω τρόπο σκέψης. Αφορούν την κατηγοριοποίηση της OnTarget σε 4 και 5 επιμέρους γκρουπ, αυτή τη φορά χωρίς μετασχηματισμό. Η πρώτη αφορά την OnTarget_Cat1, η οποία λαμβάνει την τιμή 0, όταν η

OnTarget παίρνει τιμές μικρότερες ή ίσες του 3, την τιμή 1 όταν λαμβάνει τιμές μεγαλύτερες του 3 και μικρότερες ή ίσες του 4, την τιμή 2 όταν λαμβάνει τιμές μεγαλύτερες του 4 και μικρότερες ή ίσες του 6 και την τιμή 3 όταν λαμβάνει τιμές μεγαλύτερες του 6. Η OnTarget_Cat2 λαμβάνει την τιμή 0, όταν η OnTarget2 παίρνει τιμές μικρότερες ή ίσες του 2, την τιμή 1 όταν λαμβάνει τιμές μεγαλύτερες του 2 και μικρότερες ή ίσες του 4, την τιμή 2 όταν λαμβάνει τιμές μεγαλύτερες του 4 και μικρότερες ή ίσες του 5, την τιμή 3 όταν λαμβάνει τιμές μεγαλύτερες του 5 και μικρότερες ή ίσες του 7 και την τιμή 4 όταν λαμβάνει τιμές μεγαλύτερες του 7.

Ωστόσο και σε αυτές τις περιπτώσεις δεν είχαμε στατιστικά σημαντικά αποτελέσματα για την αλληλεπίδραση (Πίνακας Γ.12). Συνεπώς, το επόμενο που θα χρειαστεί να γίνει είναι να γίνει κατηγοριοποίηση της μεταβλητής Ranking.

Στην περίπτωση της Ranking έχουμε έτοιμη την κατηγοριοποίησή της. Αυτή είναι η μεταβλητή Group, η οποία αποτελείται από 4 γκρουπ. Αντικαθιστώντας στο μοντέλο την Ranking με την Group, παρατηρούνται καλύτερα αποτελέσματα αναφορικά με τη σημαντικότητα του μοντέλου. Προέκυψε και σε αυτή την περίπτωση η ανάγκη να γίνει αλλαγή στη σειρά των μεταβλητών, καθώς φάνηκε ότι αρχικά αυτό επηρέαζε τις αποκλίσεις κάποιων μεταβλητών. Οπότε, το μοντέλο το οποίο κρίνεται καταλληλότερο στην περίπτωση της παλινδρόμησης Poisson, συμπεριλαμβάνοντας και όρο αλληλεπίδρασης είναι το model29.

```
> anova(model29)
Analysis of Deviance Table

Model: poisson, link: log

Response: Goals

Terms added sequentially (first to last)


               Df Deviance Resid. Df Resid. Dev
NULL                           741    1095.25
Corners         1   15.574       740    1079.68
Group           3   68.759       737    1010.92
Home.Away       1   27.803       736     983.12
Possesion       1   30.616       735     952.50
OnTarget        1  281.008       734     671.49
Clearances      1    7.644       733     663.85
Group:OnTarget  3   11.822       730     652.03
> summary(model29)

Call:
glm(formula = Goals ~ Corners + Group + Home.Away + Possesion +
    OnTarget + Clearances + OnTarget:Group, family = poisson)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3337  -1.0091  -0.1180   0.4872   2.7598
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.904166   0.309284  -2.923  0.00346 **
Corners         -0.058833   0.011288  -5.212 1.87e-07 ***
Group2          -0.114275   0.186262  -0.614  0.53953
Group3          -0.194772   0.193072  -1.009  0.31307
Group4          -0.880888   0.213443  -4.127 3.67e-05 ***
```

```
Home.Away1       0.156979   0.064045   2.451  0.01424 *
Possesion        0.010192   0.004405   2.314  0.02069 *
OnTarget         0.173275   0.019271   8.992  < 2e-16 ***
Clearances       0.014384   0.005015   2.868  0.00413 **
Group2:OnTarget  0.004781   0.024056   0.199  0.84248
Group3:OnTarget  0.004708   0.026080   0.181  0.85675
Group4:OnTarget  0.110463   0.034334   3.217  0.00129 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1095.25  on 741  degrees of freedom
Residual deviance:  652.03  on 730  degrees of freedom
AIC: 2036

Number of Fisher Scoring iterations: 5
```

Πίνακας 4.8: Αποτελέσματα ελέγχου αποκλίσεων και παραμέτρων για το μοντέλο παλινδρόμησης Poisson με αλληλεπίδραση.

4.5 Ερμηνεία των αποτελεσμάτων

Από τους πίνακες 4.5 και 4.8 που υπολογίστηκαν μέσω της χρήσης της συνάρτησης summary(), προκύπτουν οι παράμετροι των κυρίων επιδράσεων και των αλληλεπιδράσεων που κρίθηκαν στατιστικά σημαντικές για τις μεταβλητές Goals και Result2. Συνδυάζοντας αυτές τις τιμές των παραμέτρων και την περιγραφή των δύο μοντέλων (λογιστική παλινδρόμηση και Poisson), θα βγουν και τα βασικά συμπεράσματα για την ακριβέστερη επιρροή των συγκεκριμένων παραγόντων στις δύο αυτές μεταβλητές και κατά συνέπεια στην βάση που πρέπει να έχει μία ποδοσφαιρική ομάδα, ώστε να οδηγηθεί με ασφαλέστερο τρόπο στην επιτυχία.

Για το μοντέλο λογιστικής παλινδρόμησης, υπενθυμίζεται ότι η εξίσωση που περιγράφει το μοντέλο είναι η

$$\log\left[\frac{p_i}{1-p_i}\right] = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

οπότε, με βάση και τα αποτελέσματα του πίνακα 4.5, η ακριβής εξίσωση του μοντέλου που περιγράφει την πιθανότητα επιτυχίας p_i, δηλαδή την πιθανότητα νίκης ή και ισοπαλίας, είναι η εξής

$$\begin{aligned}\log\left[\frac{\hat{p}}{1-\hat{p}}\right] = \; & -7.6351 + 0.4327 \times OnTarget + 0.0577 \times Possession \\ & + 0.0439 \times BallsRec - 0.0212 \times Ranking + 0.1188 \times Clearances \\ & + 0.6252 \times Home.Away - 0.0696 \times Corners\end{aligned}$$

Απαλείφοντας τον λογάριθμο έχουμε ότι η εξίσωση παίρνει την μορφή

$$\frac{\hat{p}}{1-\hat{p}} = e^{-7.6351} e^{0.4327 \times OnTarget} e^{0.0577 \times Possession} e^{0.0439 \times BallsRec} e^{-0.0212 \times Ranking} e^{0.1188 \times Clearances} e^{0.6252 \times Home.Away} e^{-0.0696 \times Corners}$$

Συνεπώς για την εκτιμώμενη σχετική πιθανότητα προκύπτουν τα εξής συμπεράσματα:

- Η αύξηση κατά 1 των ευκαιριών στο στόχο (OnTarget) αυξάνει την σχετική πιθανότητα επιτυχούς αποτελέσματος (νίκης ή ισοπαλίας) πολλαπλασιαστικά κατά $e^{0.4327} = 1.5414$ φορές.
- Η αύξηση κατά 1 του ποσοστού κατοχής αυξάνει την σχετική πιθανότητα επιτυχούς αποτελέσματος πολλαπλασιαστικά κατά $e^{0.0527} = 1.0593$ φορές.
- Η αύξηση κατά 1 των ανακαταλήψεων της κατοχής αυξάνει την σχετική πιθανότητα επιτυχούς αποτελέσματος πολλαπλασιαστικά κατά $e^{0.0439} = 1.0448$.
- Η αύξηση κατά 1 της θέσης στην κατάταξη της UEFA μειώνει την σχετική πιθανότητα επιτυχούς αποτελέσματος, αφού αυτή επιδρά πολλαπλασιαστικά κατά $e^{-0.0212} = 0.8089$ φορές. Εδώ να σημειωθεί ότι με τον όρο «αύξηση» εννοούμε την αριθμητική αύξηση, δηλαδή την πτώση στην κατάταξη.
- Η αύξηση κατά 1 των αποκρούσεων αυξάνει την σχετική πιθανότητα επιτυχούς αποτελέσματος (νίκης ή ισοπαλίας) πολλαπλασιαστικά κατά $e^{0.1188} = 1.1261$.
- Οι ομάδες που αγωνίζονται στην έδρα τους έχουν αυξημένη σχετική πιθανότητα πολλαπλασιαστικά κατά $e^{0.6252} = 1.8686$ φορές για επιτυχές αποτέλεσμα σε σχέση με τις φιλοξενούμενες.
- Η αύξηση κατά 1 των κόρνερ μειώνει την σχετική πιθανότητα επιτυχούς αποτελέσματος, εφόσον υπάρχει επίδραση πολλαπλασιαστική κατά $e^{-0.0696} = 0.9327$.

Αντίστοιχα, για το μοντέλο Poisson ο γενικός τύπος είχε οριστεί ως

$$log(\mu_i) = \beta_0 + \sum_{j=1}^{k} \beta_j \chi_{ij}$$

οπότε με βάση και τα αποτελέσματα του πίνακα 4.8

$$\begin{aligned} \log[\mu_i] = {} & -0.9041 - 0.0588 \times Corners - 0.1142 \times Group2 - 0.1947 \times Group3 \\ & - 0.8808 \times Group4 + 0.1569 \times Home.Away1 + 0.0101 \times Possession \\ & + 0.1732 \times OnTarget + 0.0143 \times Clearances \\ & + 0.0047 \times OnTarget{:}\, Group2 + 0.0047 \times OnTarget{:}\, Group3 \\ & + 0.1104 \times OnTarget{:}\, Group4 \end{aligned}$$

Με την απαλοιφή του λογαρίθμου έχουμε ότι

$$\mu_i = e^{-0.9041} e^{-0.0588 \times Corners} e^{-0.1142 \times Group2} e^{-0.1947 \times Group3} e^{-0.8808 \times Group4} e^{0.1569 \times Home.Away1} e^{0.0101 \times Possession} e^{0.1732 \times OnTarget} e^{0.0143 \times Clearances} e^{0.0047 \times OnTarget:Group2} e^{0.0047 \times OnTarget:Group3} e^{0.1104 \times OnTarget:Group4}$$

Συνεπώς, τα συμπεράσματα που προκύπτουν είναι τα παρακάτω

- Η αύξηση κατά 1 των κόρνερ μειώνει την μέση τιμή Goals κατά 0.9428 φορές.
- Όταν οι θέση της ομάδας είναι μικρότερη του 17 και μεγαλύτερη ή ίση του 7 (δηλαδή όταν είμαστε στο δεύτερο τεταρτημόριο της Group), η μέση τιμή των Goals μειώνεται κατά 0.8920 φορές σε σχέση με την περίπτωση που η θέση της ομάδας είναι στο πρώτο τεταρτημόριο (μικρότερη από 7). Εδώ υπενθυμίζεται ότι μιλάμε για κατάταξη, συνεπώς όσο μικρότερη είναι κατά απόλυτη τιμή η κατάταξη, τόσο το καλύτερο.
- Όταν οι θέση της ομάδας είναι μικρότερη του 37 και μεγαλύτερη ή ίση του 17 (δηλαδή όταν είμαστε στο τρίτο τεταρτημόριο της Group), η μέση τιμή των Goals μειώνεται κατά 0.8230 φορές σε σχέση με την περίπτωση που η θέση της ομάδας είναι στο πρώτο.
- Όταν οι θέση της ομάδας είναι μεγαλύτερη ή ίση του 37 (δηλαδή όταν είμαστε στο τέταρτο τεταρτημόριο της Group), η μέση τιμή των Goals μειώνεται κατά 0.4144 φορές σε σχέση με την περίπτωση που η θέση της ομάδας είναι στο πρώτο.
- Οι ομάδες που αγωνίζονται στην έδρα τους έχουν αυξημένη μέση τιμή Goals κατά 1.1698 φορές σε σχέση με τις φιλοξενούμενες.
- Η αύξηση κατά 1 του ποσοστού κατοχής της μπάλας αυξάνει τη μέση τιμή των Goals κατά 1.0101.
- η αύξηση των ευκαιριών στον στόχο αυξάνει την μέση τιμή κατά 1.1891 φορές.
- Η αύξηση κατά 1 των αποκρούσεων αυξάνει την μέση τιμή των Goals κατά 1.0144 φορές.

Για τις αλληλεπιδράσεις ισχύουν τα εξής:

- Η αύξηση κατά 1 των ευκαιριών στο στόχο για τις ομάδες που βρίσκονται στο τέταρτο τεταρτημόριο της Group, σε σχέση με την αύξηση κατά 1 των ευκαιριών στο στόχο για αυτές που βρίσκονται στο πρώτο τεταρτημόριο της OnTarget, αυξάνει πολλαπλασιαστικά την μέση τιμή των Goals κατά $e^{0.1104} = 1.1167$ φορές.
- Για τα υπόλοιπα 2 γκρουπ σε σχέση με το πρώτο δεν παρατηρούνται στατιστικά σημαντικές αλληλεπιδράσεις.

Τα αποτελέσματα που προέκυψαν από τα δύο μοντέλα αποτελούν λογική συνέχεια των αποτελεσμάτων που είχαν παρατεθεί στα προηγούμενα κεφάλαια. Ωστόσο, σε μεγάλο βαθμό επιτεύχθηκε ο βασικός στόχος της αριθμητικής αποτύπωσης της επιρροής που έχουν οι παράγοντες που επιλέχθηκαν ως οι σημαντικότεροι. Φαίνεται για ακόμη μία φορά ότι ο βασικότερος παράγοντας επιτυχίας αποτελεί το σύνολο των εύστοχων προσπαθειών. Οι αποκρούσεις αποτελούν τον σημαντικότερο παράγοντα στο αμυντικό σκέλος, αφού και στις δύο περιπτώσεις κρίθηκε σημαντικός. Περαιτέρω διερεύνηση χρειάζεται ο παράγοντας της ανακατάληψης της κατοχής (BallsRec). Υπάρχουν έρευνες (Barreira et al., 2014), οι οποίες μελετούν τη συμπεριφορά αυτής της μεταβλητής, ως μιας η οποία έχει σημαντική επίδραση στο παιχνίδι. Φάνηκε η σημαντικότητά της στο αποτέλεσμα ενός αγώνα, ωστόσο χρειάζεται να ληφθεί υπόψη αν συμπεριλάβουμε ακριβέστερα στατιστικά δεδομένα.

Τα συγκεκριμένα δεδομένα είναι τα βασικά δεδομένα που μπορεί να αντλήσει ένας οποιοσδήποτε φίλος του αθλήματος. Αποτελούν την πρώτη εικόνα που μπορεί να φανεί για την απόδοση των ομάδων.

Τα αποτελέσματα της μελέτης είναι ένα πρώτο βήμα μελέτης της απόδοσης των ομάδων με σύγχρονα δεδομένα, προσπαθώντας να μείνουμε ανεπηρέαστοι από της επιπτώσεις την πανδημίας, ωστόσο δεν αποτελούν μετρήσεις τις οποίες μπορούν να μπουν αμέσως στο πεδίο της βελτίωσης της αγωνιστικής εικόνας. Η βάση αυτή που δίνουν μπορεί, παρόλα αυτά, να γίνει η βάση για μελλοντικές αναλύσεις πάνω σε δεδομένα που δίνουν μεγαλύτερη ακρίβεια για την εικόνα ενός ποδοσφαιρικού αγώνα.

ΚΕΦΑΛΑΙΟ 5

ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα μελέτη έγινε προσπάθεια να μελετηθούν ποιοι είναι τελικά οι βασικοί παράγοντες που επηρεάζουν την απόδοση των ομάδων ποδοσφαίρου, παίρνοντας σαν δείγμα αυτές που βρίσκονται στο υψηλότερο επίπεδο. Οι μεταβλητές που εξετάστηκαν αφορούσαν τα βασικά στατιστικά που μπορούν να δουν και οι μη στατιστικοί σε κάθε ποδοσφαιρικό αγώνα της κορυφαίας διοργάνωσης σε επίπεδο συλλόγων. Ο σκοπός ήταν από τη μία να μπορούν οι φίλαθλοι να κρίνουν παρακολουθώντας έναν αγώνα τα σημεία υπερτερούσε ή υστερούσε η ομάδα τους, χωρίς να χρειάζεται να μπουν στη διαδικασία να κάνουν στατιστική ανάλυση. Από την άλλη, έγινε μια ανάλυση λαμβάνοντας υπόψη σύγχρονα δεδομένα, για τις 3 τελευταίες σεζόν πριν από την πανδημία.

Αναφορικά με τη δομή των δεδομένων που αναλύθηκαν, αυτά έρχονται σε συμφωνία με αρκετές μελέτες ως προς το γεγονός ότι αποτελούνταν από τα στατιστικά κάθε αγώνα (Lago-Peñas & Lago-Ballesteros, 2011, Szwarc, 2007 κ.ά.) , έναντι άλλων που χρησιμοποιούσαν τα στατιστικά της κάθε ομάδας στο τέλος της χρονιάς (Karlis & Ntzoufras, 2000). Αυτός ο τρόπος δίνει τη δυνατότητα να συγκεντρωθεί ένα αρκετά μεγάλο δείγμα, πράγμα που δίνει μεγαλύτερη ακρίβεια στο αποτέλεσμα. Ζήτημα το οποίο τίθεται για παραπέρα μελέτη είναι η ανεξαρτησία του δείγματος.

Ως προς τα αποτελέσματα της μελέτης, η περιγραφική ανάλυση των δεδομένων έδειξε ότι τα βασικά δεδομένα που αφορούν το ποδόσφαιρο έχουν στην πλειοψηφία τους ένα μικρό εύρος τιμών, κάνοντας δύσκολη την ανάλυσή τους με τη χρήση διαγράμματος διασποράς (scatter plot). Ωστόσο, η χρήση θηκογράμματος φαίνεται ότι αποτελεί μια ασφαλή επιλογή. Οι διαφορές στα επίπεδα των βασικών κατηγορικών μεταβλητών (Home.Away, Group, Result που αναδείχθηκαν αποτέλεσαν και τα πρώτα βασικά συμπεράσματα αναφορικά με τους παράγοντες που επηρεάζουν την απόδοση των ομάδων, αλλά και αναφορικά με το ποιες συνεχείς μεταβλητές είναι αυτές που δείχνουν καθοριστικές. Το τελευταίο έγινε εφικτό να το υποθέσουμε από το γεγονός ότι ενώ σε κάποιες κατηγορικές μεταβλητές, οι οποίες είναι αποδεδειγμένη και από παλαιότερες μελέτες (Lago-Peñas & Lago-Ballesteros, 2011) η σημασία τους, υπήρχαν αρκετές μεταβλητές οι οποίες δεν έδειχναν διαφορές.

Ο έλεγχος των συσχετίσεων ενώ δεν έδωσε σημαντικά αποτελέσματα αναφορικά με το ποιες μεταβλητές είναι σημαντικές εκτός από την περίπτωση της OnTarget, αποτέλεσε ένα πολύτιμο εργαλείο για την μελέτη της πολυσυγγραμμικότητας στην προσαρμογή των ΓΓΜ.

Αναφορικά με τις μεθόδους εξόρυξης δεδομένων που αξιοποιήθηκαν, ο αλγόριθμος Random Forest φαίνεται ότι λειτουργεί πολύ καλά και μπορεί να αξιοποιηθεί σε δεδομένα ποδοσφαίρου. Επιπλέον, τα αποτελέσματα της κατηγοριοποίησης με βάση τους αλγόριθμους

Naïve Bayes και Support Vector Machine, έδειξαν σημαντική διαφορά στο ποσοστό τον ορθά κατηγοριοποιημένων παρατηρήσεων υπέρ του δεύτερου. Ωστόσο, δεδομένου ότι τα ποσοστά αυτά στον Naïve Bayes είναι κοντά στο 80%, αποτελεί και αυτός μια αξιόπιστη λύση.

Τα δύο μοντέλα που προέκυψαν από την χρήση των ΓΓΜ δείχνουν ότι ο πιο σημαντικός παράγοντας, με μεγάλη διαφορά από τον επόμενο, είναι τα σουτ στο στόχο. Ο λόγος είναι ότι μια επιπλέον ευκαιρία στο στόχο μπορεί να δώσει στη σχετική πιθανότητα επιτυχούς αποτελέσματος μία αύξηση περίπου 54% και στην μέση τιμή των γκολ που επιτυγχάνονται περίπου στο 19%. Αυτό μπορεί να αποδειχθεί ακόμα πιο σημαντικό αν αναλογιστούμε ότι δεν υπήρξε κάποιο στατιστικά σημαντικό αποτέλεσμα για τα συνολικά σουτ. Από την συγκεκριμένη παρατήρηση φαίνεται το κομμάτι εκείνο που συχνά κάνει το ποδόσφαιρο αρκετά ενδιαφέρον σε σχέση με τα άλλα παιχνίδια, ότι δηλαδή δεν είναι αναγκαίο να έχει μια ομάδα την πρωτοβουλία του παιχνιδιού και να δημιουργεί πολλές ευκαιρίες, αλλά αυτές να είναι χτισμένες αποτελεσματικά. Αξιοποιώντας με αυτό τον σκοπό όλο το έμψυχο δυναμικό που μπορεί να έχει ένας σύλλογος, μπορούν και οι θεωρητικά πιο αδύναμες ομάδες να κάνουν έστω και μικρές υπερβάσεις.

Παρόλο που στις περισσότερες αναλύσεις ποδοσφαιρικών αγώνων που παρακολουθούμε λαμβάνονται κυρίως υπόψη τα ποσοστά κατοχής της μπάλας ή σε άλλες περιπτώσεις τα λεπτά κατοχής ως μέτρο για την απόδοση μιας ομάδας, αν και φάνηκε ότι και στις δύο περιπτώσεις είναι ένας στατιστικά σημαντικός παράγοντας δεν αποτελεί εκείνον ο οποίος καθορίζει σε μεγαλύτερο βαθμό το αποτέλεσμα. Συνεπώς, αυτό το συμπέρασμα μας βοηθάει στο να αντιληφθούμε ότι ένα αυξημένο ποσοστό της κατοχής της μπάλας αποτελεί μια πολύ καλή απόδειξη για την καλή κυκλοφορία της, η οποία έχει σημαντικό αντίκτυπο στο αποτέλεσμα.

Αρκετά σημαντική και στις δυο περιπτώσεις δείχνει να είναι η αύξηση που δίνει ο παράγοντας της έδρας. Αυτό είναι ένα αποτέλεσμα το οποίο φάνηκε και από την περιγραφική ανάλυση, ωστόσο μέσα από την μελέτη των ΓΓΜ επιτεύχθηκε και μία ποσοτικοποίηση αυτής της σχέσης.

Όπως, επίσης, φάνηκε και η σημασία που έχει η δυναμική μιας ομάδας. Παρόλα αυτά, στην περίπτωση του μοντέλου λογιστικής παλινδρόμησης ήταν προτιμότερο να ελεγχθεί με την μορφή του Ranking, σε αντίθεση με το μοντέλο Poisson που έδειξε να ανταποκρίνεται η μεταβλητή Group. Βασιζόμενοι σε αυτό, μπορούμε να πούμε ότι είναι χρήσιμο να εξετάζονται και οι δύο περιπτώσεις σε ανάλογες μελλοντικές έρευνες.

Ένα αποτέλεσμα το οποίο είναι επίσης άξιο αναφοράς στους σημαντικούς παράγοντες και στα δύο μοντέλα, είναι το γεγονός ότι κρίθηκε στατιστικά σημαντική η μεταβλητή Corners και μάλιστα με αρνητικό πρόσημο, πράγμα που σημαίνει ότι υπάρχει μείωση στην σχετική πιθανότητα επιτυχούς αποτελέσματος και στην μέση τιμή των γκολ. Μια λογική εξήγηση είναι ότι, αν αυτό το αποτέλεσμα συνδυαστεί με την ισχυρή συσχέτιση ανάμεσα στη συγκεκριμένη

μεταβλητή και στις μεταβλητές Tattemps και Blocked, σε αντιπαράθεση με την μέτρια συσχέτιση που είχε με την OnTarget, τα Corners είναι ένας παράγοντας ο οποίος ερμηνεύεται ως μια αποτυχημένη επιθετική προσπάθεια. Ανεξαρτήτως, λοιπόν, του γεγονότος ότι ένα κόρνερ μπορεί να φέρει την επιτιθέμενη ομάδα σε πλεονεκτική θέση, καθώς μπορεί να συγκεντρώσει περισσότερους παίκτες στην περιοχή της αμυνόμενης, η αύξηση των κόρνερ σημαίνει πολλές αποτυχημένες προσπάθειες για γκολ, το οποίο εκφράστηκε και στα δύο μοντέλα.

Αξίζει επίσης να ειπωθεί ότι στο μοντέλο λογιστικής παλινδρόμησης, εν αντιθέσει με το μοντέλο, κρίθηκε σημαντική η BallsRec που αφορά την ανακατάληψη της κατοχής, παράγοντας ο οποίος έχει αρχίσει να μπαίνει προς μελέτη από τις περισσότερες ομάδες, ως η αρχή μιας αποτελεσματικής επιθετικής πρωτοβουλίας. Ο συγκεκριμένος παράγοντας, αν και έχει αποδειχθεί και σε άλλες μελέτες η χρησιμότητά του (Barreira et al., 2014) χρειάζεται ακόμα περισσότερη προσπάθεια στην αποσαφήνισή του.

Σημαντική επίσης θετική συνεισφορά είχε και η μεταβλητή που αφορά τις αποκρούσεις από την άμυνα. Εδώ αξίζει να αναφερθεί και η σημασία που έχει η προσθήκη στα εξεταζόμενα μοντέλα μεταβλητών που αφορούν τόσο την επίθεση, όσο και την άμυνα. Αυτό διότι, όπως φαίνεται ξεκάθαρα από τις επιδόσεις των ομάδων τα τελευταία χρόνια στα μεγάλα πρωταθλήματα, η υπεροχή και η τελική νίκη σε έναν ποδοσφαιρικό αγώνα επηρεάζεται σε μεγαλύτερο βαθμό από την αμυντική απόδοση. Άρα μπορούμε να πούμε ότι ένα επιπλέον θετικό της συγκεκριμένης μελέτης ήταν αυτό, σε αντίθεση με μελέτες που έχουν λάβει υπόψη τους μόνο τα επιθετικά αποτελέσματα (Leontijević et al., 2018).

Τέλος, συγκρίνοντας τα αποτελέσματα που έδωσαν τα δύο ΓΓΜ σε σχέση με αυτά που προέκυψαν από τη χρήση των αλγόριθμων εξόρυξης δεδομένων, μπορούμε να πούμε ότι, ενώ δίνονται απαντήσεις στα ανάλογα ερωτήματα της παρούσας μελέτης από τη χρήση των αλγορίθμων, η ποιότητα των αποτελεσμάτων σε καμία περίπτωση δεν συγκρίνεται με αυτή των ΓΓΜ. Ενώ στην πλειοψηφία των περιπτώσεων αναδείχθηκαν οι σημαντικοί παράγοντες που επιδρούν, δεν ήταν με την ίδια ακρίβεια, καθώς όπως φαίνεται ειδικά στην περίπτωση που είχαμε την μεταβλητή Result σαν μεταβλητή απόκρισης, οι παράγοντες που κρίθηκαν σημαντικοί ήταν αρκετά λιγότεροι. Όπως και το ότι αυτοί οι αλγόριθμοι δεν δίνουν ποτέ τη δυνατότητα της ποσοτικοποίησης της σχέσης ανάμεσα στις ανεξάρτητες μεταβλητές και την μεταβλητή απόκρισης.

ΠΑΡΑΡΤΗΜΑ Α

ΠΙΝΑΚΕΣ ΚΕΦΑΛΑΙΟΥ 1

Home.Away=0							
Χαρακτηριστικό	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**	**St. Dev.**
Goals	0	0	1	1.278	2	7	1.286525
Tattempts	1	8	11	11.32	14	30	5.16245
OnTarget	0	2	4	4.073	6	15	2.552707
Blocked	0	1	2	2.763	4	10	2.042208
Corners	0	2	4	4.375	6	18	2.754319
Offsides	0	1	2	2.261	3	8	1.824487
Possesion	28	42	49	49.15	57	72	9.673346
PassAcc	64	80	84	83.52	88	94	5.817624
PassT	196	392.5	486	498.6	594	894	143.6666
PassC	125	314.5	405	423.4	520.5	825	144.1882
Distance	92.3	106.7	110.1	110	112.8	141.8	4.937621
BallsRec	25	42	47	47.79	53	85	8.612715
Tackles	0	2	3	3.642	5	13	2.310564
Blocks	0	2	3	3.612	5	15	2.403886
Clearances	1	11	17	17.48	23	50	7.836608
Yellow	0	1	2	2.089	3	6	1.347736
Red	0	0	0	0.0835	0	2	0.2866852
FoulsCom	3	9	12	12.4	15	30	4.271515

Πίνακας Α.1: Περιγραφικά στατιστικά των μεταβλητών για τα εκτός έδρας αποτελέσματα (Home.Away = 0).

Home.Away=1							
Χαρακτηριστικό	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**	**St. Dev.**
Goals	0	1	1	1.768	3	8	1.550993
Tattempts	3	10	14	14.47	18	34	5.978428
OnTarget	0	3	5	5.291	7	17	2.985805
Blocked	0	2	3	3.62	5	15	2.412475
Corners	0	3	5	5.59	7	19	3.335203
Offsides	0	1	2	2.561	4	11	2.071287
Possesion	28	43	51	50.84	58	72	9.678102
PassAcc	66	80	85	84.11	88	96	5.542769
PassT	197	416.5	514	519.3	606	1096	149.8509
PassC	130	334	432	442.6	532.5	1014	150.7134
Distance	99.2	107.3	110.7	110.4	113.2	143	4.627169

BallsRec	28	43	48	48.47	54	78	8.352428
Tackles	0	2	4	4.137	6	13	2.476915
Blocks	0	1	2	2.747	4	10	2.033625
Clearances	1	9	13	14.35	19	43	7.324237
Yellow	0	1	2	1.838	3	6	1.269493
Red	0	0	0	0.0916 4	0	1	0.2889127
FoulsCom	3	9	12	11.81	14	26	3.749604

Πίνακας Α.2: Περιγραφικά στατιστικά των μεταβλητών για τα εντός έδρας αποτελέσματα (Home.Away = 1).

Round=0							
Χαρακτηριστικό	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**	**St. Dev.**
Goals	0	0	1	1.491	2	8	1.42765
Tattempts	1	9	12	12.93	16	33	5.894614
OnTarget	0	3	4	4.727	6	17	2.888721
Blocked	0	2	3	3.2	4	15	2.294053
Corners	0	3	5	4.99	7	19	3.222189
Offsides	0	1	2	2.457	4	11	1.964429
Possesion	28	43	50	49.99	57	72	9.579596
PassAcc	64	81	85	83.94	88	96	5.690606
PassT	196	402.2	499	510.7	603	1029	146.1804
PassC	125	320	417	435.1	530.5	960	147.1897
Distance	92.3	107.2	110.4	110.2	112.9	124	4.435068
BallsRec	25	42	48	48.2	53	85	8.60629
Tackles	0	2	4	3.943	5	13	2.437072
Blocks	0	2	3	3.184	4	15	2.284349
Clearances	1	10	15	15.8	21	50	7.761312
Yellow	0	1	2	1.908	3	6	1.296231
Red	0	0	0	0.0885 4	0	2	0.2903802
FoulsCom	3	9	12	12.15	15	30	4.076754

Πίνακας Α.3: Περιγραφικά στατιστικά των μεταβλητών για τα αποτελέσματα της φάσης ομίλων (Round = 0).

Round=1							
Χαρακτηριστικό	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**	**St. Dev.**
Goals	0	0	1	1.633	3	7	1.502693
Tattempts	1	9	13	12.77	16	34	5.472411
OnTarget	0	3	4	4.524	6	13	2.675822
Blocked	0	2	3	3.163	4	12	2.211003
Corners	0	3	5	4.958	7	13	2.726642
Offsides	0	1	2	2.253	3	9	1.924914

Possesion	28	43	50	50	58	72	10.16172
PassAcc	69	79	84	83.38	88	93	5.663862
PassT	199	400	496.5	503	593	1096	150.3476
PassC	140	326.8	413	425.8	519	1014	149.6737
Distance	94.8	106.3	110	110.1	113	143	5.857703
BallsRec	25	42	48	47.86	53.75	73	8.067824
Tackles	0	2	4	3.705	5	11	2.294309
Blocks	0	2	3	3.163	4	12	2.211003
Clearances	1	10	16	16.3	21	42	7.674807
Yellow	0	1	2	2.157	3	6	1.361656
Red	0	0	0	0.0843 4	0	1	0.2787339
FoulsCom	3	9	12	11.95	14	25	3.858491

Πίνακας Α.4: Περιγραφικά στατιστικά των μεταβλητών για τα αποτελέσματα της φάσης knock-out (Round = 1).

Group=1							
Χαρακτηριστικό	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**	**St. Dev.**
Goals	0	1	2	1.978	3	7	1.514357
Tattempts	4	11	15	15.57	19	34	6.083947
OnTarget	0	4	6	5.848	8	13	2.762649
Blocked	0	2	3	3.736	5	15	2.484517
Corners	0	4	6	6.36	8	19	3.29611
Offsides	0	1	2	2.393	3	11	1.940458
Possesion	33	50	57	55.69	62	72	8.591279
PassAcc	70	85	88	87.13	90	96	4.511607
PassT	251	519	598.5	602.2	699.8	1022	140.0612
PassC	183	448	533.5	530.1	622	939	140.9741
Distance	99.2	106.3	109.5	109.3	112.1	124	4.211283
BallsRec	30	42	47	47.96	53	85	8.405554
Tackles	0	2	3	3.775	5	11	2.288526
Blocks	0	1	2	2.674	4	12	1.990232
Clearances	1	8.25	12	13.21	17	42	6.786206
Yellow	0	1	2	1.904	3	6	1.287437
Red	0	0	0	0.0842 7	0	1	0.2785754
FoulsCom	3	9	11	11.45	14	25	3.499834

Πίνακας Α.5: Περιγραφικά στατιστικά των μεταβλητών για τα αποτελέσματα του πρώτου γκρουπ δυναμικότητας (Group = 1).

Group=2							
Χαρακτηριστικό	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**	**St. Dev.**
Goals	0	1	1	1.781	3	8	1.640682

Tattempts	2	10	13	13.55	17	33	5.603694
OnTarget	0	3	5	5.084	7	17	3.032532
Blocked	0	2	3	3.27	4	12	2.233611
Corners	0	3	5	5.208	7	13	2.652479
Offsides	0	1	2	2.36	4	9	1.917786
Possesion	31	47	53	53.07	60	72	9.089651
PassAcc	66	82	86	84.92	88	94	5.063941
PassT	227	454.2	543	556	640.8	1096	148.7245
PassC	180	364.2	461	476.7	560.5	1014	150.1453
Distance	100.2	107.5	110.3	110.7	113.3	147	5.288424
BallsRec	28	42	47	47.77	53	73	8.441743
Tackles	0	2	3	3.556	5	13	2.428571
Blocks	0	1	3	2.742	4	8	1.772534
Clearances	2	9.25	14	14.66	20	38	7.096627
Yellow	0	1	2	1.837	3	6	1.276195
Red	0	0	0	0.0842 7	0	1	0.2785754
FoulsCom	4	9	11	11.57	14	23	3.996956

Πίνακας Α.6: Περιγραφικά στατιστικά των μεταβλητών για τα αποτελέσματα του δεύτερου γκρουπ δυναμικότητας (Group = 2).

Group=3							
Χαρακτηριστικό	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**	**St. Dev.**
Goals	0	1	1	1.459	2	7	1.313975
Tattempts	1	9	12	12.7	16	29	5.149477
OnTarget	0	3	4.5	4.679	6	15	2.751922
Blocked	0	2	3	3.097	4	13	2.152111
Corners	0	3	4	949	6	17	3.264015
Offsides	0	1	2	2.495	3	8	1.854993
Possesion	28	42	48	47.85	54	68	8.603937
PassAcc	67	79	83.5	82.56	87	93	5.297499
PassT	229	389	468	472	550	846	119.6511
PassC	154	307.8	382.5	394.6	478.2	786	119.2556
Distance	94.8	107.5	111.1	110.6	113.9	121	4.99696
BallsRec	25	43	49	48.78	54	75	8.48938
Tackles	0	2	4	4.031	6	11	2.275642
Blocks	0	2	3	3.367	5	12	2.225206
Clearances	3	10	15.5	16.12	21	39	7.353458
Yellow	0	1	2	1.98	3	6	1.308588
Red	0	0	0	0.0816 3	0	1	0.2745054
FoulsCom	3	10	13	12.8	15	30	4.328344

Πίνακας Α.7: Περιγραφικά στατιστικά των μεταβλητών για τα αποτελέσματα του τρίτου γκρουπ δυναμικότητας (Group = 3).

Group=4							
Χαρακτηριστικό	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**	**St. Dev.**
Goals	0	0	1	0.9211	1	5	1.05363
Tattempts	2	6	9	9.968	12	28	4.963257
OnTarget	0	2	3	3.216	4	10	2.120887
Blocked	0	1	2	2.705	4	12	2.125052
Corners	0	1.25	3	3.516	5	11	2.514984
Offsides	0	1	2	2.389	3.75	9	2.114674
Possesion	28	38	43.5	44	49	65	8.13185
PassAcc	64	78	82	80.98	86	91	5.783536
PassT	196	342.8	407	415.6	479.5	695	102.4162
PassC	125	268	333	340.8	402	582	101.895
Distance	92.3	107	110.6	110	112.9	122.5	4.484118
BallsRec	26	42	47	47.95	53	78	8.623051
Tackles	0	2	4	4.163	6	13	2.592156
Blocks	0	2	3	3.868	5	15	2.72914
Clearances	5	14	18	19.42	24	50	8.245029
Yellow	0	1	2	2.121	3	6	1.372858
Red	0	0	0	0.1	0	2	0.3178965
FoulsCom	3	10	12	12.49	15	25	4.061364

Πίνακας Α.8: Περιγραφικά στατιστικά των μεταβλητών για τα αποτελέσματα του τέταρτου γκρουπ δυναμικότητας (Group = 4).

Result=0							
Χαρακτηριστικό	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**	**St. Dev.**
Goals	0	0	0	0.5544	1	4	0.687942
Tattempts	1	7	10	10.5	13	28	4.865533
OnTarget	0	2	3	3.165	4	10	2.078774
Blocked	0	1	3	2.919	4	12	2.130549
Corners	0	2	4	4.379	6	18	2.905237
Offsides	0	1	2	2.168	3	8	1.789952
Possesion	28	39	46	46.86	53	72	9.510629
PassAcc	64	80	83	82.69	87	93	5.308259
PassT	196	383	458	467	545	1096	128.6163
PassC	125	308	382	391.2	462	1014	126.1363
Distance	92.3	106.3	110	109.8	113	141.8	5.184278
BallsRec	26	41	45	46.18	51	70	8.181314
Tackles	0	2	4	3.937	5	13	2.421197
Blocks	0	2	3	3.326	4	15	2.278692
Clearances	2	10	15	15.55	20	50	7.30199
Yellow	0	1	2	2.189	3	6	1.360624

Red	0	0	0	0.1228	0	2	0.3393332
FoulsCom	3	9	12	12.28	15	25	4.164549

Πίνακας Α.9: Περιγραφικά στατιστικά των μεταβλητών για τα αποτελέσματα που αφορούσαν τις ήττες (Result=0).

Result=1							
Χαρακτηριστικό	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**	**St. Dev.**
Goals	0	0	1	1.07	2	3	0.8890589
Tattempts	2	9	12	13.02	17	33	6.100018
OnTarget	0	2	4	4.32	6	12	2.533302
Blocked	0	2	3	3.419	5	13	2.46835
Corners	0	3	5	5.047	7	19	3.185886
Offsides	0	1	2	2.39	4	11	2.103883
Possesion	29	44	50	50	56	71	8.649105
PassAcc	67	79.75	84	83.32	87	94	5.528984
PassT	221	400	495.5	499.4	570.8	1029	134.7976
PassC	159	317.8	410.5	421.4	498.2	960	136.1472
Distance	99.6	107.6	110.9	110.5	113.2	122.5	4.210543
BallsRec	33	44	49	50.02	56	78	8.733851
Tackles	0	2	4	4.157	6	13	2.534858
Blocks	0	2	3	3.384	4.25	13	2.435933
Clearances	3	11	17	17.43	22	43	8.101752
Yellow	0	1	2	1.983	3	6	1.24475
Red	0	0	0	0.08721	0	1	0.2829653
FoulsCom	3	10	12	12.74	15	30	4.116169

Πίνακας Α.10: Περιγραφικά στατιστικά των μεταβλητών για τα αποτελέσματα που αφορούσαν τις ισοπαλίες (Result=1).

Result=2							
Χαρακτηριστικό	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**	**St. Dev.**
Goals	1	2	3	2.765	3	8	1.372836
Tattempts	4	11	15	15.21	18	34	5.525212
OnTarget	1	4	6	6.418	8	17	2.73431
Blocked	0	2	3	3.326	4	15	2.272503
Corners	0	3	5	5.547	7	17	3.178166
Offsides	0	1	2	2.667	4	9	1.997651
Possesion	28	47	54	53.13	61	72	9.516886
PassAcc	66	82	87	85.24	90	96	5.85755
PassT	197	433	564	556.7	669	1022	157.5719
PassC	130	354	487	481.9	591	939	160.0385

Distance	99.2	107.6	110.3	110.4	112.9	143	4.682531
BallsRec	25	44	48	48.93	54	85	8.265347
Tackles	0	2	3	3.681	5	11	2.299681
Blocks	0	1	3	2.909	4	12	2.125986
Clearances	1	9	14	15.37	20	40	7.850342
Yellow	0	1	2	1.726	3	6	1.270816
Red	0	0	0	0.0526 3	0	1	0.2236897
FoulsCom	3	9	11	11.54	14	24	3.765478

Πίνακας Α.10: Περιγραφικά στατιστικά των μεταβλητών για τα αποτελέσματα που αφορούσαν τις νίκες (Result=2).

ΠΑΡΑΡΤΗΜΑ Β

ΠΙΝΑΚΕΣ ΚΕΦΑΛΑΙΟΥ 2

Round		Φάση ομίλων	Νοκ-αουτ	
Group	1	108	70	178
	2	132	46	178
	3	150	46	196
	4	186	4	190
		576	166	742

```
> chisq.test(Group, Round)

        Pearson's Chi-squared test

data:  Group and Round
X-squared = 75.767, df = 3, p-value = 2.481e-16
```

Πίνακας Β.1: Πίνακας συνάφειας και αποτελέσματα ελέγχου χ^2 για τις μεταβλητές Group και Round.

Result		Ήττα	Ισοπαλία	Νίκη	
Group	1	33	40	105	178
	2	62	37	79	178
	3	80	44	72	196
	4	110	51	29	190
		285	172	285	742

```
> chisq.test(Group, Result)

        Pearson's Chi-squared test

data:  Group and Result
X-squared = 87.515, df = 6, p-value < 2.2e-16
```

Πίνακας Β.2: Πίνακας συνάφειας και αποτελέσματα ελέγχου χ^2 για τις μεταβλητές Group και Result.

Result		Ήττα	Ισοπαλία	Νίκη	
Home.Away	Εκτός	167	86	118	371
	Εντός	118	86	167	371
		285	172	285	742

```
> chisq.test(Home.Away, Result)

        Pearson's Chi-squared test

data:  Home.Away and Result
X-squared = 16.849, df = 2, p-value = 0.0002194
```

Πίνακας Β.3: Πίνακας συνάφειας και αποτελέσματα ελέγχου χ^2 για τις μεταβλητές Home.Away και Result.

Result		Ήττα	Ισοπαλία	Νίκη	
BallsRec2	≤42	106	34	60	200
	43-53	127	82	148	357
	≥54	52	56	77	185
		285	172	285	742

```
> chisq.test(BallsRec2,Result)

        Pearson's Chi-squared test

data:  BallsRec2 and Rsult
X-squared = 29.3, df = 4, p-value = 6.795e-06
```

Πίνακας Β.4: Πίνακας συνάφειας και αποτελέσματα ελέγχου χ^2 για τις μεταβλητές BallsRec2 και Result.

Group		1	2	3	4	
Possesion2	≤42	19	24	57	87	187
	43-50	30	51	58	62	201
	54-57.75	43	42	52	31	168
	>57.75	86	61	29	10	186
		178	178	196	190	742

```
> chisq.test(Possesion2,Group)

        Pearson's Chi-squared test

data:  Possesion2 and Group
X-squared = 154.67, df = 9, p-value < 2.2e-16
```

Πίνακας Β.5: Πίνακας συνάφειας και αποτελέσματα ελέγχου χ^2 για τις μεταβλητές Group και Possesion2.

Result		Ήττα	Ισοπαλία	Νίκη	
Possesion2	≤42	105	31	51	187
	43-50	84	58	59	201
	54-57.75	46	52	70	168
	>57.75	50	31	105	186

	285	172	285	742

```
> chisq.test(Possesion2,Result)

        Pearson's Chi-squared test

data:  Possesion2 and Result
X-squared = 68.325, df = 6, p-value = 9.012e-13
```

Πίνακας Β.6: Πίνακας συνάφειας και αποτελέσματα ελέγχου χ^2 για τις μεταβλητές Possesion2 και Result.

Group		**1**	**2**	**3**	**4**	
PassAcc2	**≤80**	17	37	65	73	192
	81-85	35	50	59	67	211
	86-88	39	48	53	37	177
	>88	87	43	19	13	162
		178	178	196	190	742

```
> chisq.test(PassAcc2,Group)

        Pearson's Chi-squared test

data:  PassAcc2 and Group
X-squared = 140.04, df = 9, p-value < 2.2e-16
```

Πίνακας Β.7: Πίνακας συνάφειας και αποτελέσματα ελέγχου χ^2 για τις μεταβλητές PassAcc2 και Group.

Result		**Ήττα**	**Ισοπαλία**	**Νίκη**	
PassAcc2	**≤80**	78	50	64	192
	81-85	106	47	58	211
	86-88	71	50	56	177
	>88	30	25	107	162
		285	172	285	742

```
> chisq.test(PassAcc2,Result)

        Pearson's Chi-squared test

data:  PassAcc2 and Result
X-squared = 73.835, df = 6, p-value = 6.665e-14
```

Πίνακας Β.8: Πίνακας συνάφειας και αποτελέσματα ελέγχου χ^2 για τις μεταβλητές PassAcc2 και Result.

Group		**1**	**2**	**3**	**4**	
PassT2	**≤400**	16	26	55	90	187

	401-497	22	40	63	60	185
	498-599.8	52	53	50	29	184
	>599.8	88	59	28	11	186
		178	178	196	190	742

Πίνακας Β.9: Πίνακας συνάφειας και αποτελέσματα ελέγχου χ^2 για τις μεταβλητές PassT2 και Result.

Result		**Ήττα**	**Ισοπαλία**	**Νίκη**	
PassT2	**≤400**	89	44	54	187
	401-497	91	43	51	185
	498-599.8	68	52	64	184
	>599.8	37	33	116	186
		285	172	285	742

```
> chisq.test(PassC2,Group)

        Pearson's Chi-squared test

data:  PassC2 and Group
X-squared = 187.18, df = 9, p-value < 2.2e-16
```

Πίνακας Β.10: Πίνακας συνάφειας και αποτελέσματα ελέγχου χ^2 για τις μεταβλητές Group και PassC2.

Group		**1**	**2**	**3**	**4**	
PassC2	**≤321.5**	13	25	58	90	186
	322-416	25	43	59	59	186
	417-528	48	54	54	30	186
	>529	92	56	25	11	184
		178	178	196	190	742

```
> chisq.test(PassT2,Group)

        Pearson's Chi-squared test

data:  PassT2 and Group
X-squared = 177.6, df = 9, p-value < 2.2e-16
```

Πίνακας Β.11: Πίνακας συνάφειας και αποτελέσματα ελέγχου χ^2 για τις μεταβλητές Group και PassT2.

Result		**Ήττα**	**Ισοπαλία**	**Νίκη**	
PassC2	**≤321.5**	86	45	55	186

	322-416	92	44	50	186
	417-528	69	52	65	186
	>529	38	31	115	184
		285	172	285	742

```
> chisq.test(PassC2,Result)

        Pearson's Chi-squared test

data:  PassC2 and Result
X-squared = 67.787, df = 6, p-value = 1.161e-12
```

Πίνακας Β.12: Πίνακας συνάφειας και αποτελέσματα ελέγχου χ^2 για τις μεταβλητές Result και PassC2.

ΠΑΡΑΡΤΗΜΑ Γ

ΠΙΝΑΚΕΣ ΚΕΦΑΛΑΙΟΥ 4

```
> model12<-glm(Result2~OnTarget+Tattempts+Blocked+Corners+Possesion
+              +PassAcc+Distance+BallsRec+Tackles+Blocks+Yellow+Red
+              +FoulsCom+Home.Away+Round+Group+Ranking+Clearances, family
=binomial)
> step(model12)
Start:  AIC=732.05
Result2 ~ OnTarget + Tattempts + Blocked + Corners + Possesion +
    PassAcc + Distance + BallsRec + Tackles + Blocks + Yellow +
    Red + FoulsCom + Home.Away + Round + Group + Ranking + Clearances

             Df Deviance    AIC
- Distance    1   694.05 730.05
- Red         1   694.14 730.14
- PassAcc     1   694.33 730.33
- Tackles     1   694.58 730.58
- Blocks      1   695.30 731.30
- Blocked     1   695.51 731.51
- Tattempts   1   695.77 731.77
<none>            694.05 732.05
- Yellow      1   697.26 733.26
- Corners     1   697.51 733.51
- Group       1   697.82 733.82
- Ranking     1   698.89 734.89
- FoulsCom    1   700.00 736.00
- Possesion   1   704.84 740.84
- Round       1   705.24 741.24
- Home.Away   1   706.52 742.52
- BallsRec    1   707.84 743.84
- OnTarget    1   729.05 765.05
- Clearances  1   754.57 790.57

Step:  AIC=730.05
Result2 ~ OnTarget + Tattempts + Blocked + Corners + Possesion +
    PassAcc + BallsRec + Tackles + Blocks + Yellow + Red + FoulsCom +
    Home.Away + Round + Group + Ranking + Clearances

             Df Deviance    AIC
- Red         1   694.14 728.14
- PassAcc     1   694.33 728.33
- Tackles     1   694.58 728.58
- Blocks      1   695.30 729.30
- Blocked     1   695.51 729.51
- Tattempts   1   695.78 729.78
<none>            694.05 730.05
- Yellow      1   697.28 731.28
- Corners     1   697.51 731.51
- Group       1   697.82 731.82
- Ranking     1   698.89 732.89
- FoulsCom    1   700.04 734.04
- Round       1   705.24 739.24
- Possesion   1   705.38 739.38
- Home.Away   1   706.52 740.52
- BallsRec    1   708.78 742.78
- OnTarget    1   729.05 763.05
- Clearances  1   754.80 788.80

Step:  AIC=728.14
Result2 ~ OnTarget + Tattempts + Blocked + Corners + Possesion +
    PassAcc + BallsRec + Tackles + Blocks + Yellow + FoulsCom +
```

```
    Home.Away + Round + Group + Ranking + Clearances

             Df Deviance    AIC
- PassAcc     1   694.41 726.41
- Tackles     1   694.69 726.69
- Blocks      1   695.38 727.38
- Blocked     1   695.61 727.61
- Tattempts   1   695.89 727.89
<none>            694.14 728.14
- Corners     1   697.66 729.66
- Group       1   697.84 729.84
- Yellow      1   697.99 729.99
- Ranking     1   699.18 731.18
- FoulsCom    1   700.25 732.25
- Round       1   705.31 737.31
- Possesion   1   705.71 737.71
- Home.Away   1   706.56 738.56
- BallsRec    1   708.98 740.98
- OnTarget    1   729.18 761.18
- Clearances  1   755.57 787.57

Step:  AIC=726.41
Result2 ~ OnTarget + Tattempts + Blocked + Corners + Possesion +
    BallsRec + Tackles + Blocks + Yellow + FoulsCom + Home.Away +
    Round + Group + Ranking + Clearances
             Df Deviance    AIC
- Tackles     1   694.96 724.96
- Blocks      1   695.63 725.63
- Blocked     1   695.88 725.88
- Tattempts   1   696.17 726.17
<none>            694.41 726.41
- Corners     1   698.10 728.10
- Group       1   698.17 728.17
- Yellow      1   698.23 728.23
- Ranking     1   699.48 729.48
- FoulsCom    1   700.27 730.27
- Round       1   706.09 736.09
- Home.Away   1   706.62 736.62
- BallsRec    1   709.74 739.74
- Possesion   1   712.31 742.31
- OnTarget    1   729.89 759.89
- Clearances  1   757.11 787.11
Step:  AIC=724.96
Result2 ~ OnTarget + Tattempts + Blocked + Corners + Possesion +
    BallsRec + Blocks + Yellow + FoulsCom + Home.Away + Round +
    Group + Ranking + Clearances

             Df Deviance    AIC
- Blocks      1   696.13 724.13
- Blocked     1   696.44 724.44
- Tattempts   1   696.68 724.68
<none>            694.96 724.96
- Yellow      1   698.50 726.50
- Corners     1   698.56 726.56
- Group       1   698.86 726.86
- Ranking     1   699.90 727.90
- FoulsCom    1   700.52 728.52
- Round       1   706.92 734.92
- Home.Away   1   708.13 736.13
- BallsRec    1   710.61 738.61
- Possesion   1   712.32 740.32
- OnTarget    1   730.28 758.28
- Clearances  1   757.61 785.61

Step:  AIC=724.13
Result2 ~ OnTarget + Tattempts + Blocked + Corners + Possesion +
    BallsRec + Yellow + FoulsCom + Home.Away + Round + Group +
    Ranking + Clearances

             Df Deviance    AIC
- Blocked     1   697.87 723.87
- Tattempts   1   698.03 724.03
<none>            696.13 724.13
- Yellow      1   699.30 725.30
```

```
- Corners     1   699.60 725.60
- Group       1   699.96 725.96
- Ranking     1   701.03 727.03
- FoulsCom    1   701.15 727.15
- Round       1   708.09 734.09
- Home.Away   1   708.47 734.47
- BallsRec    1   711.41 737.41
- Possesion   1   712.51 738.51
- OnTarget    1   731.17 757.17
- Clearances  1   766.78 792.78

Step:  AIC=723.87
Result2 ~ OnTarget + Tattempts + Corners + Possesion + BallsRec +
    Yellow + FoulsCom + Home.Away + Round + Group + Ranking +
    Clearances

             Df Deviance    AIC
- Tattempts   1   698.30 722.30
<none>            697.87 723.87
- Yellow      1   701.05 725.05
- Group       1   702.04 726.04
- Corners     1   702.43 726.43
- Ranking     1   702.63 726.63
- FoulsCom    1   702.88 726.88
- Round       1   710.03 734.03
- Home.Away   1   710.06 734.06
- BallsRec    1   713.25 737.25
- Possesion   1   714.11 738.11
- OnTarget    1   753.08 777.08
- Clearances  1   767.57 791.57

Step:  AIC=722.3
Result2 ~ OnTarget + Corners + Possesion + BallsRec + Yellow +
    FoulsCom + Home.Away + Round + Group + Ranking + Clearances

             Df Deviance    AIC
<none>            698.30 722.30
- Yellow      1   701.72 723.72
- Corners     1   702.43 724.43
- Group       1   702.52 724.52
- Ranking     1   703.00 725.00
- FoulsCom    1   703.54 725.54
- Round       1   710.35 732.35
- Home.Away   1   711.58 733.58
- BallsRec    1   713.91 735.91
- Possesion   1   716.07 738.07
- Clearances  1   767.58 789.58
- OnTarget    1   793.07 815.07

Call:  glm(formula = Result2 ~ OnTarget + Corners + Possesion + BallsRec
+
    Yellow + FoulsCom + Home.Away + Round + Group + Ranking +
    Clearances, family = binomial)

Coefficients:
(Intercept)     OnTarget      Corners    Possesion     BallsRec       Yel
low
   -7.20120      0.42078     -0.07762      0.05653      0.04443     -0.14
389
   FoulsCom   Home.Away1       Round1        Group      Ranking   Clearan
ces
    0.05856      0.71902     -0.82668     -0.33654     -0.01543      0.12
752

Degrees of Freedom: 741 Total (i.e. Null);  730 Residual
Null Deviance:       988.4
Residual Deviance: 698.3  AIC: 722.3
```

Πίνακας Γ.1: Επιλογή μεταβλητών με τον αλγόριθμο step.

```
> model13<-glm(Result2~OnTarget+Corners+Possesion
+              +BallsRec+Yellow+FoulsCom+Home.Away+Round
+              +Group+Ranking+Clearances, family=binomial)
> summary(model13)

Call:
glm(formula = Result2 ~ OnTarget + Corners + Possesion + BallsRec +
    Yellow + FoulsCom + Home.Away + Round + Group + Ranking +
    Clearances, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7817  -0.7617   0.3251   0.7641   2.0454

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.201204   1.153233  -6.244 4.26e-10 ***
OnTarget     0.420782   0.048718   8.637  < 2e-16 ***
Corners     -0.077619   0.037945  -2.046 0.040801 *
Possesion    0.056531   0.013648   4.142 3.44e-05 ***
BallsRec     0.044433   0.011527   3.855 0.000116 ***
Yellow      -0.143887   0.078210  -1.840 0.065805 .
FoulsCom     0.058559   0.025861   2.264 0.023548 *
Home.Away1   0.719015   0.199724   3.600 0.000318 ***
Round1      -0.826676   0.239122  -3.457 0.000546 ***
Group       -0.336545   0.164131  -2.050 0.040319 *
Ranking     -0.015433   0.007224  -2.136 0.032647 *
Clearances   0.127524   0.016758   7.610 2.74e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 988.39  on 741  degrees of freedom
Residual deviance: 698.30  on 730  degrees of freedom
AIC: 722.3

Number of Fisher Scoring iterations: 5
```

Πίνακας Γ.2: Αποτελέσματα για το μοντέλο λογιστικής παλινδρόμησης με ανεξάρτητες τις μεταβλητές που προέκυψαν από το κριτήριο AIC.

```
> model14<-glm(Result2~OnTarget+Corners+Possesion
+              +BallsRec+FoulsCom+Home.Away+Round
+              +Group+Ranking+Clearances, family=binomial)
> summary(model14)

Call:
glm(formula = Result2 ~ OnTarget + Corners + Possesion + BallsRec +
    FoulsCom + Home.Away + Round + Group + Ranking + Clearances,
    family = binomial)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7702  -0.7795   0.3278   0.7678   2.1033


Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.327111   1.151144  -6.365 1.95e-10 ***
OnTarget     0.428063   0.048384   8.847  < 2e-16 ***
Corners     -0.078432   0.037731  -2.079 0.037644 *
```

```
Possesion     0.056452   0.013618   4.145 3.39e-05 ***
BallsRec      0.044829   0.011490   3.902 9.55e-05 ***
FoulsCom      0.041080   0.023906   1.718 0.085716 .
Home.Away1    0.725543   0.199174   3.643 0.000270 ***
Round1       -0.859872   0.238522  -3.605 0.000312 ***
Group        -0.314637   0.163162  -1.928 0.053809 .
Ranking      -0.016281   0.007184  -2.266 0.023446 *
Clearances    0.126104   0.016685   7.558 4.10e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 988.39  on 741  degrees of freedom
Residual deviance: 701.72  on 731  degrees of freedom
AIC: 723.72

Number of Fisher Scoring iterations: 5
```

Πίνακας Γ.3: Αποτελέσματα για το μοντέλο λογιστικής παλινδρόμησης με ανεξάρτητες τις μεταβλητές που προέκυψαν από το κριτήριο AIC εκτός της μεταβλητής Yellow.

```
> model15<-glm(Result2~OnTarget+Possesion+BallsRec+Ranking
+              +Clearances+Home.Away+Corners+Round, family=binomial)
> summary(model15)

Call:
glm(formula = Result2 ~ OnTarget + Possesion + BallsRec + Ranking +
    Clearances + Home.Away + Corners + Round, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7612  -0.7836   0.3415   0.7893   2.0535

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.331613   0.981479  -7.470 8.02e-14 ***
OnTarget     0.428132   0.048173   8.887  < 2e-16 ***
Possesion    0.056894   0.013162   4.323 1.54e-05 ***
BallsRec     0.044277   0.011388   3.888 0.000101 ***
Ranking     -0.026961   0.004687  -5.753 8.78e-09 ***
Clearances   0.123881   0.016454   7.529 5.12e-14 ***
Home.Away1   0.662935   0.195720   3.387 0.000706 ***
Corners     -0.074100   0.037509  -1.976 0.048209 *
Round1      -0.801230   0.233945  -3.425 0.000615 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 988.39  on 741  degrees of freedom
Residual deviance: 708.07  on 733  degrees of freedom
AIC: 726.07

Number of Fisher Scoring iterations: 5
```

Πίνακας Γ.4: Αποτελέσματα για το μοντέλο λογιστικής παλινδρόμησης με ανεξάρτητες τις μεταβλητές που προέκυψαν από το κριτήριο AIC εκτός των μεταβλητών Yellow, FoulsCom και Group.

```
> model16<-glm(Result3~OnTarget+Possesion+BallsRec+Ranking
+              +Clearances+Home.Away+Corners+Round, family=binomial)
> summary(model16)

Call:
glm(formula = Result3 ~ OnTarget + Possesion + BallsRec + Ranking +
    Clearances + Home.Away + Corners + Round, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5419  -0.7636  -0.3867   0.8326   2.3239
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.653704   0.988734  -5.718 1.08e-08 ***
OnTarget     0.446731   0.045237   9.875  < 2e-16 ***
Possesion    0.037596   0.013233   2.841  0.00450 **
BallsRec     0.014691   0.010723   1.370  0.17066
Ranking     -0.029797   0.005518  -5.400 6.67e-08 ***
Clearances   0.080387   0.015597   5.154 2.55e-07 ***
Home.Away1   0.593421   0.195718   3.032  0.00243 **
Corners     -0.101564   0.034966  -2.905  0.00368 **
Round1      -0.079133   0.224544  -0.352  0.72452
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 988.39  on 741  degrees of freedom
Residual deviance: 722.02  on 733  degrees of freedom
AIC: 740.02

Number of Fisher Scoring iterations: 5

> anova(model16)
Analysis of Deviance Table

Model: binomial, link: logit

Response: Result3

Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev
NULL                         741     988.39
OnTarget    1  190.238       740     798.16
Possesion   1    2.574       739     795.58
BallsRec    1    2.360       738     793.22
Ranking     1   29.076       737     764.15
Clearances  1   26.158       736     737.99
Home.Away   1    7.194       735     730.79
Corners     1    8.651       734     722.14
Round       1    0.124       733     722.02
```

Πίνακας Γ.5: Αποτελέσματα για το μοντέλο λογιστικής παλινδρόμησης με μεταβλητή απόκρισης την Result3.

```
> step(model22)
Start:  AIC=2044.04
Goals ~ OnTarget + Tattempts + Blocked + Corners + Possesion +
    PassAcc + Distance + BallsRec + Tackles + Blocks + Yellow +
    Red + FoulsCom + Home.Away + Round + Group + Ranking + Clearances

            Df Deviance    AIC
- Red        1   646.11 2042.0
- Blocked    1   646.32 2042.2
- PassAcc    1   646.50 2042.4
- Yellow     1   646.51 2042.4
- Blocks     1   646.62 2042.5
- FoulsCom   1   646.74 2042.7
```

```
- BallsRec    1   646.77 2042.7
- Group       1   646.95 2042.9
- Round       1   646.96 2042.9
- Ranking     1   647.28 2043.2
- Distance    1   647.35 2043.3
- Tackles     1   647.70 2043.6
<none>            646.11 2044.0
- Possesion   1   650.80 2046.7
- Clearances  1   651.26 2047.2
- Tattempts   1   654.57 2050.5
- Home.Away   1   654.84 2050.8
- Corners     1   656.92 2052.8
- OnTarget    1   823.55 2219.5

Step:  AIC=2042.04
Goals ~ OnTarget + Tattempts + Blocked + Corners + Possesion +
    PassAcc + Distance + BallsRec + Tackles + Blocks + Yellow +
    FoulsCom + Home.Away + Round + Group + Ranking + Clearances
            Df Deviance    AIC
- Blocked     1   646.32 2040.2
- PassAcc     1   646.50 2040.4
- Yellow      1   646.55 2040.5
- Blocks      1   646.62 2040.5
- FoulsCom    1   646.75 2040.7
- BallsRec    1   646.78 2040.7
- Group       1   646.95 2040.9
- Round       1   646.98 2040.9
- Ranking     1   647.30 2041.2
- Distance    1   647.39 2041.3
- Tackles     1   647.70 2041.6
<none>            646.11 2042.0
- Possesion   1   650.82 2044.8
- Clearances  1   651.31 2045.2

- Tattempts   1   654.57 2048.5
- Home.Away   1   654.84 2048.8
- Corners     1   656.93 2050.8
- OnTarget    1   823.60 2217.5

Step:  AIC=2040.25
Goals ~ OnTarget + Tattempts + Corners + Possesion + PassAcc +
    Distance + BallsRec + Tackles + Blocks + Yellow + FoulsCom +
    Home.Away + Round + Group + Ranking + Clearances

            Df Deviance    AIC
- PassAcc     1   646.73 2038.7
- Yellow      1   646.74 2038.7
- Blocks      1   646.79 2038.7
- FoulsCom    1   646.98 2038.9
- BallsRec    1   647.00 2038.9
- Group       1   647.14 2039.1
- Round       1   647.18 2039.1
- Ranking     1   647.50 2039.4
- Distance    1   647.55 2039.5
- Tackles     1   647.97 2039.9
<none>            646.32 2040.2
- Possesion   1   650.95 2042.9
- Clearances  1   651.61 2043.5
- Home.Away   1   654.98 2046.9
- Corners     1   656.93 2048.9
- Tattempts   1   658.81 2050.7
- OnTarget    1   867.79 2259.7
Step:  AIC=2038.66
Goals ~ OnTarget + Tattempts + Corners + Possesion + Distance +
    BallsRec + Tackles + Blocks + Yellow + FoulsCom + Home.Away +
    Round + Group + Ranking + Clearances

            Df Deviance    AIC
- BallsRec    1   647.13 2037.1
- Yellow      1   647.14 2037.1
- Blocks      1   647.18 2037.1
```

```
- FoulsCom    1   647.23 2037.2
- Round       1   647.46 2037.4
- Group       1   647.69 2037.6
- Ranking     1   647.92 2037.8
- Distance    1   648.11 2038.0
- Tackles     1   648.34 2038.3
<none>            646.73 2038.7
- Clearances  1   651.63 2041.6
- Possesion   1   654.88 2044.8
- Home.Away   1   655.13 2045.1
- Corners     1   657.88 2047.8
- Tattempts   1   659.18 2049.1
- OnTarget    1   873.83 2263.8

Step:  AIC=2037.06
Goals ~ OnTarget + Tattempts + Corners + Possesion + Distance +
    Tackles + Blocks + Yellow + FoulsCom + Home.Away + Round +
    Group + Ranking + Clearances

             Df Deviance    AIC
- Blocks      1   647.51 2035.4
- Yellow      1   647.54 2035.5
- FoulsCom    1   647.62 2035.5
- Round       1   647.86 2035.8
- Group       1   648.09 2036.0
- Ranking     1   648.28 2036.2
- Tackles     1   648.65 2036.6
- Distance    1   648.86 2036.8
<none>            647.13 2037.1
- Clearances  1   652.23 2040.2
- Home.Away   1   655.47 2043.4
- Possesion   1   655.51 2043.4
- Corners     1   658.35 2046.3
- Tattempts   1   659.57 2047.5
- OnTarget    1   874.66 2262.6
Step:  AIC=2035.44
Goals ~ OnTarget + Tattempts + Corners + Possesion + Distance +
    Tackles + Yellow + FoulsCom + Home.Away + Round + Group +
    Ranking + Clearances

             Df Deviance    AIC
- Yellow      1   647.85 2033.8
- FoulsCom    1   647.91 2033.8
- Round       1   648.21 2034.1
- Group       1   648.46 2034.4
- Ranking     1   648.63 2034.6
- Tackles     1   649.09 2035.0
- Distance    1   649.30 2035.2
<none>            647.51 2035.4
- Clearances  1   653.55 2039.5
- Home.Away   1   655.51 2041.4
- Possesion   1   655.53 2041.5
- Corners     1   658.67 2044.6
- Tattempts   1   660.07 2046.0
- OnTarget    1   875.56 2261.5
Step:  AIC=2033.77
Goals ~ OnTarget + Tattempts + Corners + Possesion + Distance +
    Tackles + FoulsCom + Home.Away + Round + Group + Ranking +
    Clearances

             Df Deviance    AIC
- FoulsCom    1   648.05 2032.0
- Round       1   648.46 2032.4
- Group       1   648.74 2032.7
- Ranking     1   649.04 2033.0
- Tackles     1   649.52 2033.5
- Distance    1   649.69 2033.6
<none>            647.85 2033.8
- Clearances  1   653.68 2037.6
- Possesion   1   655.81 2039.7
- Home.Away   1   656.03 2040.0
- Corners     1   659.15 2043.1
- Tattempts   1   660.21 2044.1
- OnTarget    1   879.32 2263.2
```

```
Step:  AIC=2031.98
Goals ~ OnTarget + Tattempts + Corners + Possesion + Distance +
    Tackles + Home.Away + Round + Group + Ranking + Clearances

             Df Deviance    AIC
- Round       1   648.66 2030.6
- Group       1   648.88 2030.8
- Ranking     1   649.31 2031.2
- Tackles     1   649.81 2031.7
- Distance    1   649.84 2031.8
<none>            648.05 2032.0
- Clearances  1   653.78 2035.7
- Possesion   1   655.81 2037.7
- Home.Away   1   656.18 2038.1
- Corners     1   659.40 2041.3
- Tattempts   1   660.26 2042.2
- OnTarget    1   880.08 2262.0

Step:  AIC=2030.59
Goals ~ OnTarget + Tattempts + Corners + Possesion + Distance +
    Tackles + Home.Away + Group + Ranking + Clearances

             Df Deviance    AIC
- Group       1   649.70 2029.6
- Ranking     1   650.11 2030.0
- Distance    1   650.44 2030.4
- Tackles     1   650.47 2030.4
<none>            648.66 2030.6
- Clearances  1   654.80 2034.7
- Possesion   1   656.09 2036.0
- Home.Away   1   657.12 2037.0
- Corners     1   660.21 2040.1
- Tattempts   1   660.79 2040.7
- OnTarget    1   880.26 2260.2

Step:  AIC=2029.62
Goals ~ OnTarget + Tattempts + Corners + Possesion + Distance +
    Tackles + Home.Away + Ranking + Clearances

             Df Deviance    AIC
- Distance    1   651.34 2029.3
- Tackles     1   651.39 2029.3
<none>            649.70 2029.6
- Clearances  1   655.98 2033.9
- Possesion   1   657.89 2035.8
- Home.Away   1   658.21 2036.1
- Corners     1   661.10 2039.0
- Tattempts   1   661.69 2039.6
- Ranking     1   662.03 2040.0
- OnTarget    1   880.78 2258.7
Step:  AIC=2029.26
Goals ~ OnTarget + Tattempts + Corners + Possesion + Tackles +
    Home.Away + Ranking + Clearances

             Df Deviance    AIC
- Tackles     1   653.13 2029.0
<none>            651.34 2029.3
- Clearances  1   658.36 2034.3
- Possesion   1   658.84 2034.8
- Home.Away   1   660.18 2036.1
- Tattempts   1   662.30 2038.2
- Corners     1   663.06 2039.0
- Ranking     1   663.66 2039.6
- OnTarget    1   880.86 2256.8

Step:  AIC=2029.05
Goals ~ OnTarget + Tattempts + Corners + Possesion + Home.Away +
    Ranking + Clearances

             Df Deviance    AIC
<none>            653.13 2029.0
- Clearances  1   660.02 2033.9
- Home.Away   1   661.27 2035.2
- Possesion   1   661.52 2035.5
```

```
- Tattempts   1   663.98 2037.9
- Corners     1   664.81 2038.7
- Ranking     1   665.44 2039.4
- OnTarget    1   881.38 2255.3

Call:  glm(formula = Goals ~ OnTarget + Tattempts + Corners + Possesion +
    Home.Away + Ranking + Clearances, family = poisson)

Coefficients:
(Intercept)     OnTarget    Tattempts      Corners    Possesion   Home.Aw
ay1
  -0.997398     0.218576    -0.029231    -0.040833     0.012713     0.182
829
    Ranking   Clearances
  -0.005578     0.013228

Degrees of Freedom: 741 Total (i.e. Null);  734 Residual
Null Deviance:       1095
Residual Deviance: 653.1   AIC: 2029
```

Πίνακας Γ.6: Επιλογή μεταβλητών με τον αλγόριθμο step για το μοντέλο παλινδρόμησης Poisson.

```
> anova(model23)
Analysis of Deviance Table

Model: poisson, link: log

Response: Goals

Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev
NULL                         741    1095.25
OnTarget    1   383.36       740     711.89
Tattempts   1    20.28       739     691.61
Ranking     1    12.67       738     678.94
Corners     1     9.50       737     669.44
Home.Away   1     5.45       736     663.99
Possesion   1     3.98       735     660.02
Clearances  1     6.89       734     653.13

> 1-pchisq(653.13,734)
[1] 0.9852238
```

Πίνακας Γ.7: Πίνακας απόκλισης για το μοντέλο με ανεξάρτητες μεταβλητές αυτές που προέκυψαν από το κριτήριο AIC για το μοντέλο παλινδρόμησης Poisson.

```
> summary(model23)

Call:
glm(formula = Goals ~ OnTarget + Tattempts + Ranking + Corners +
    Home.Away + Possesion + Clearances, family = poisson)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.27318  -1.04734  -0.08473   0.50659   2.66121

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.997398   0.275387  -3.622 0.000293 ***
OnTarget     0.218576   0.014520  15.053  < 2e-16 ***
Tattempts   -0.029231   0.008966  -3.260 0.001113 **
Ranking     -0.005578   0.001636  -3.410 0.000650 ***
Corners     -0.040833   0.012089  -3.378 0.000731 ***
Home.Away1   0.182829   0.064277   2.844 0.004449 **
Possesion    0.012713   0.004393   2.894 0.003806 **
Clearances   0.013228   0.005008   2.641 0.008256 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1095.25  on 741  degrees of freedom
Residual deviance:  653.13  on 734  degrees of freedom
AIC: 2029.1

Number of Fisher Scoring iterations: 5
```

Πίνακας Γ.8: Πίνακας τιμών των παραμέτρων του μοντέλου Poisson, όπως προέκυψε από το κριτήριο AIC.

```
> anova(model19)
Analysis of Deviance Table

Model: binomial, link: logit

Response: Result2

Terms added sequentially (first to last)


                     Df Deviance Resid. Df Resid. Dev
NULL                                   741     988.39
OnTarget              1  155.204       740     833.19
Possesion             1    5.288       739     827.90
BallsRec              1   20.325       738     807.58
Ranking               1   14.998       737     792.58
Clearances            1   59.870       736     732.71
Home.Away             1    9.313       735     723.40
Corners               1    3.551       734     719.85
Clearances:Home.Away  1    5.382       733     714.46
BallsRec:Home.Away    1    7.370       732     707.09
Ranking:Clearances    1    1.039       731     706.06
```

Πίνακας Γ.9: Αποτελέσματα διαφοράς αποκλίσεων για το μοντέλο λογιστικής παλινδρόμησης με τις 3 στατιστικά σημαντικές αλληλεπιδράσεις.

```
> anova(model110)
Analysis of Deviance Table

Model: binomial, link: logit

Response: Result2

Terms added sequentially (first to last)


                     Df Deviance Resid. Df Resid. Dev
NULL                                   741     988.39
OnTarget              1  155.204       740     833.19
Possesion             1    5.288       739     827.90
BallsRec              1   20.325       738     807.58
Ranking               1   14.998       737     792.58
Clearances            1   59.870       736     732.71
Home.Away             1    9.313       735     723.40
Corners               1    3.551       734     719.85
Clearances:Home.Away  1    5.382       733     714.46
BallsRec:Home.Away    1    7.370       732     707.09

> summary(model110)

Call:
glm(formula = Result2 ~ OnTarget + Possesion + BallsRec + Ranking +
    Clearances + Home.Away + Corners + Clearances:Home.Away +
    BallsRec:Home.Away, family = binomial)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8329  -0.7547   0.3391   0.7854   2.0481

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -8.784400   1.153669  -7.614 2.65e-14 ***
OnTarget              0.441729   0.048427   9.122  < 2e-16 ***
Possesion             0.057819   0.013145   4.399 1.09e-05 ***
BallsRec              0.076475   0.016877   4.531 5.86e-06 ***
```

Πίνακας Γ.10: Τελικό μοντέλο λογιστικής παλινδρόμησης με αλληλεπίδραση.

```
> anova(model26)
Analysis of Deviance Table

Model: poisson, link: log

Response: Goals

Terms added sequentially (first to last)


                     Df Deviance Resid. Df Resid. Dev
NULL                                   741    1095.25
OnTarget_Cat          4   387.99       737     707.27
Corners               1    15.78       736     691.49
Ranking               1    11.21       735     680.28
Home.Away             1     6.38       734     673.90
Possesion             1     4.03       733     669.87
Clearances            1     4.88       732     664.99
OnTarget_Cat:Ranking  4     3.02       728     661.97

> qchisq(0.95,4)
[1] 9.487729
```

Πίνακας Γ.11: Αποτελέσματα ελέγχου αποκλίσεων για το μοντέλο με βάση την OnTarget_Cat.

```
> anova(model27)
Analysis of Deviance Table

Model: poisson, link: log

Response: Goals

Terms added sequentially (first to last)


                      Df Deviance Resid. Df Resid. Dev
NULL                                    741    1095.25
OnTarget_Cat1          3   365.80       738     729.45
Ranking                1     8.06       737     721.39
Corners                1    16.49       736     704.90
Home.Away              1     5.87       735     699.03
Possesion              1     3.77       734     695.26
Clearances             1     5.19       733     690.07
OnTarget_Cat1:Ranking  3     3.54       730     686.53

> anova(model28)
Analysis of Deviance Table

Model: poisson, link: log

Response: Goals

Terms added sequentially (first to last)


                      Df Deviance Resid. Df Resid. Dev
NULL                                    741    1095.25
OnTarget_Cat2          2  232.309       739     862.94
Ranking                1   29.376       738     833.57
Corners                1    2.257       737     831.31
Home.Away              1   16.338       736     814.97
Possesion              1   18.193       735     796.78
Clearances             1    1.434       734     795.35
OnTarget_Cat2:Ranking  1    0.367       733     794.98
```

Πίνακας Γ.12: Αποτελέσματα ελέγχου αποκλίσεων για το μοντέλο με κατηγοριοποίηση με βάση τις πραγματικές τιμές της OnTarget.

ΠΑΡΑΡΤΗΜΑ Δ

ΚΩΔΙΚΑΣ ΣΤΗΝ R

```
data1<-read.table('2016-17 2.txt',header=TRUE)
data2<-read.table('2017-18 2.txt',header=TRUE)
data3<-read.table('2018-19 2.txt',header=TRUE)
data<-rbind(data1,data2,data3)
attach(data)
data<-data[, -1]

#Δημιουργία Μεταβλητής Group
summary(Ranking)
length(Ranking)
Group=NULL
for (i in 1:742){
  if (Ranking[i]<7){
    Group[i]=1
  } else if (Ranking[i]>=7 & Ranking[i]<17) {
    Group[i]=2
  } else if (Ranking[i]>=17 & Ranking[i]<37) {
    Group[i]=3
  } else {
    Group[i]=4
  }
}
data<-cbind(data,Group)


#Δημιουργία Μεταβλητής GoalsCon
GoalsCon=NULL
for (i in 1:742){
  if (i%%2==1){
    GoalsCon[i]=Goals[i+1]
  } else {
    GoalsCon[i]=Goals[i-1]
  }
}
data<-cbind(data,GoalsCon)

#Δημιουργία Μεταβλητής Result
Result=NULL
for (i in 1:742){
  if (Goals[i]-GoalsCon[i]>0){
    Result[i]=2
  } else if (Goals[i]-GoalsCon[i]==0) {
    Result[i]=1
  } else {
    Result[i]=0
  }
}
data<-cbind(data,Result)
```

```
#Κεφάλαιο 1
#Περιγραφικά μέτρα
summary(Goals)
summary(Tattempts)
summary(OnTarget)
summary(Blocked)
summary(Corners)
summary(Offsides)
summary(Possesion)
summary(PassAcc)
summary(PassT)
summary(PassC)
summary(Distance)
summary(BallsRec)
summary(Tackles)
summary(Blocks)
summary(Clearances)
summary(Yellow)
summary(Red)
summary(FoulsCom)
sd(Goals)
sd(Tattempts)
sd(OnTarget)
sd(Blocked)
sd(Corners)
sd(Offsides)
sd(Possesion)
sd(PassAcc)
sd(PassT)
sd(PassC)
sd(Distance)
sd(BallsRec)
sd(Tackles)
sd(Blocks)
sd(Clearances)
sd(Yellow)
sd(Red)
sd(FoulsCom)

#Περιγραφικά μέτρα ανά κατηγορία της Home.Away
summary(Goals[Home.Away==0])
summary(Tattempts[Home.Away==0])
summary(OnTarget[Home.Away==0])
summary(Blocked[Home.Away==0])
summary(Corners[Home.Away==0])
summary(Offsides[Home.Away==0])
summary(Possesion[Home.Away==0])
summary(PassAcc[Home.Away==0])
summary(PassT[Home.Away==0])
summary(PassC[Home.Away==0])
summary(Distance[Home.Away==0])
summary(BallsRec[Home.Away==0])
summary(Tackles[Home.Away==0])
summary(Blocks[Home.Away==0])
summary(Clearances[Home.Away==0])
```

```
summary(Yellow[Home.Away==0])
summary(Red[Home.Away==0])
summary(FoulsCom[Home.Away==0])
sd(Goals[Home.Away==0])
sd(Tattempts[Home.Away==0])
sd(OnTarget[Home.Away==0])
sd(Blocked[Home.Away==0])
sd(Corners[Home.Away==0])
sd(Offsides[Home.Away==0])
sd(Possesion[Home.Away==0])
sd(PassAcc[Home.Away==0])
sd(PassT[Home.Away==0])
sd(PassC[Home.Away==0])
sd(Distance[Home.Away==0])
sd(BallsRec[Home.Away==0])
sd(Tackles[Home.Away==0])
sd(Blocks[Home.Away==0])
sd(Clearances[Home.Away==0])
sd(Yellow[Home.Away==0])
sd(Red[Home.Away==0])
sd(FoulsCom[Home.Away==0])

summary(Goals[Home.Away==1])
summary(Tattempts[Home.Away==1])
summary(OnTarget[Home.Away==1])
summary(Blocked[Home.Away==1])
summary(Corners[Home.Away==1])
summary(Offsides[Home.Away==1])
summary(Possesion[Home.Away==1])
summary(PassAcc[Home.Away==1])
summary(PassT[Home.Away==1])
summary(PassC[Home.Away==1])
summary(Distance[Home.Away==1])
summary(BallsRec[Home.Away==1])
summary(Tackles[Home.Away==1])
summary(Blocks[Home.Away==1])
summary(Clearances[Home.Away==1])
summary(Yellow[Home.Away==1])
summary(Red[Home.Away==1])
summary(FoulsCom[Home.Away==1])
sd(Goals[Home.Away==1])
sd(Tattempts[Home.Away==1])
sd(OnTarget[Home.Away==1])
sd(Blocked[Home.Away==1])
sd(Corners[Home.Away==1])
sd(Offsides[Home.Away==1])
sd(Possesion[Home.Away==1])
sd(PassAcc[Home.Away==1])
sd(PassT[Home.Away==1])
sd(PassC[Home.Away==1])
sd(Distance[Home.Away==1])
sd(BallsRec[Home.Away==1])
sd(Tackles[Home.Away==1])
sd(Blocks[Home.Away==1])
sd(Clearances[Home.Away==1])
```

```
sd(Yellow[Home.Away==1])
sd(Red[Home.Away==1])
sd(FoulsCom[Home.Away==1])

#Περιγραφικά μέτρα ανά κατηγορία της Round
summary(Goals[Round==0])
summary(Tattempts[Round==0])
summary(OnTarget[Round==0])
summary(Blocked[Round==0])
summary(Corners[Round==0])
summary(Offsides[Round==0])
summary(Possesion[Round==0])
summary(PassAcc[Round==0])
summary(PassT[Round==0])
summary(PassC[Round==0])
summary(Distance[Round==0])
summary(BallsRec[Round==0])
summary(Tackles[Round==0])
summary(Blocks[Round==0])
summary(Clearances[Round==0])
summary(Yellow[Round==0])
summary(Red[Round==0])
summary(FoulsCom[Round==0])
sd(Goals[Round==0])
sd(Tattempts[Round==0])
sd(OnTarget[Round==0])
sd(Blocked[Round==0])
sd(Corners[Round==0])
sd(Offsides[Round==0])
sd(Possesion[Round==0])
sd(PassAcc[Round==0])
sd(PassT[Round==0])
sd(PassC[Round==0])
sd(Distance[Round==0])
sd(BallsRec[Round==0])
sd(Tackles[Round==0])
sd(Blocks[Round==0])
sd(Clearances[Round==0])
sd(Yellow[Round==0])
sd(Red[Round==0])
sd(FoulsCom[Round==0])

summary(Goals[Round==1])
summary(Tattempts[Round==1])
summary(OnTarget[Round==1])
summary(Blocked[Round==1])
summary(Corners[Round==1])
summary(Offsides[Round==1])
summary(Possesion[Round==1])
summary(PassAcc[Round==1])
summary(PassT[Round==1])
summary(PassC[Round==1])
summary(Distance[Round==1])
summary(BallsRec[Round==1])
summary(Tackles[Round==1])
summary(Blocks[Round==1])
```

```
summary(Clearances[Round==1])
summary(Yellow[Round==1])
summary(Red[Round==1])
summary(FoulsCom[Round==1])
sd(Goals[Round==1])
sd(Tattempts[Round==1])
sd(OnTarget[Round==1])
sd(Blocked[Round==1])
sd(Corners[Round==1])
sd(Offsides[Round==1])
sd(Possesion[Round==1])
sd(PassAcc[Round==1])
sd(PassT[Round==1])
sd(PassC[Round==1])
sd(Distance[Round==1])
sd(BallsRec[Round==1])
sd(Tackles[Round==1])
sd(Blocks[Round==1])
sd(Clearances[Round==1])
sd(Yellow[Round==1])
sd(Red[Round==1])
sd(FoulsCom[Round==1])

#Περιγραφικά μέτρα ανά κατηγορία της Group
summary(Goals[Group==1])
summary(Tattempts[Group==1])
summary(OnTarget[Group==1])
summary(Blocked[Group==1])
summary(Corners[Group==1])
summary(Offsides[Group==1])
summary(Possesion[Group==1])
summary(PassAcc[Group==1])
summary(PassT[Group==1])
summary(PassC[Group==1])
summary(Distance[Group==1])
summary(BallsRec[Group==1])
summary(Tackles[Group==1])
summary(Blocks[Group==1])
summary(Clearances[Group==1])
summary(Yellow[Group==1])
summary(Red[Group==1])
summary(FoulsCom[Group==1])
sd(Goals[Group==1])
sd(Tattempts[Group==1])
sd(OnTarget[Group==1])
sd(Blocked[Group==1])
sd(Corners[Group==1])
sd(Offsides[Group==1])
sd(Possesion[Group==1])
sd(PassAcc[Group==1])
sd(PassT[Group==1])
sd(PassC[Group==1])
sd(Distance[Group==1])
sd(BallsRec[Group==1])
sd(Tackles[Group==1])
sd(Blocks[Group==1])
```

```
summary(Goals[Group==2])
summary(Tattempts[Group==2])
summary(OnTarget[Group==2])
summary(Blocked[Group==2])
summary(Corners[Group==2])
summary(Offsides[Group==2])
summary(Possesion[Group==2])
summary(PassAcc[Group==2])
summary(PassT[Group==2])
summary(PassC[Group==2])
summary(Distance[Group==2])
summary(BallsRec[Group==2])
summary(Tackles[Group==2])
summary(Blocks[Group==2])
summary(Clearances[Group==2])
summary(Yellow[Group==2])
summary(Red[Group==2])
summary(FoulsCom[Group==2])
sd(Goals[Group==2])
sd(Tattempts[Group==2])
sd(OnTarget[Group==2])
sd(Blocked[Group==2])
sd(Corners[Group==2])
sd(Offsides[Group==2])
sd(Possesion[Group==2])
sd(PassAcc[Group==2])
sd(PassT[Group==2])
sd(PassC[Group==2])
sd(Distance[Group==2])
sd(BallsRec[Group==2])
sd(Tackles[Group==2])
sd(Blocks[Group==2])
sd(Clearances[Group==2])
sd(Yellow[Group==2])
sd(Red[Group==2])
sd(FoulsCom[Group==2])

summary(Goals[Group==3])
summary(Tattempts[Group==3])
summary(OnTarget[Group==3])
summary(Blocked[Group==3])
summary(Corners[Group==3])
summary(Offsides[Group==3])
summary(Possesion[Group==3])
summary(PassAcc[Group==3])
summary(PassT[Group==3])
summary(PassC[Group==3])
summary(Distance[Group==3])
summary(BallsRec[Group==3])
summary(Tackles[Group==3])
summary(Blocks[Group==3])
summary(Clearances[Group==3])
summary(Yellow[Group==3])
summary(Red[Group==3])
summary(FoulsCom[Group==3])
```

```
sd(Goals[Group==3])
sd(Tattempts[Group==3])
sd(OnTarget[Group==3])
sd(Blocked[Group==3])
sd(Corners[Group==3])
sd(Offsides[Group==3])
sd(Possesion[Group==3])
sd(PassAcc[Group==3])
sd(PassT[Group==3])
sd(PassC[Group==3])
sd(Distance[Group==3])
sd(BallsRec[Group==3])
sd(Tackles[Group==3])
sd(Blocks[Group==3])
sd(Clearances[Group==3])
sd(Yellow[Group==3])
sd(Red[Group==3])
sd(FoulsCom[Group==3])

summary(Goals[Group==4])
summary(Tattempts[Group==4])
summary(OnTarget[Group==4])
summary(Blocked[Group==4])
summary(Corners[Group==4])
summary(Offsides[Group==4])
summary(Possesion[Group==4])
summary(PassAcc[Group==4])
summary(PassT[Group==4])
summary(PassC[Group==4])
summary(Distance[Group==4])
summary(BallsRec[Group==4])
summary(Tackles[Group==4])
summary(Blocks[Group==4])
summary(Clearances[Group==4])
summary(Yellow[Group==4])
summary(Red[Group==4])
summary(FoulsCom[Group==4])
sd(Goals[Group==4])
sd(Tattempts[Group==4])
sd(OnTarget[Group==4])
sd(Blocked[Group==4])
sd(Corners[Group==4])
sd(Offsides[Group==4])
sd(Possesion[Group==4])
sd(PassAcc[Group==4])
sd(PassT[Group==4])
sd(PassC[Group==4])
sd(Distance[Group==4])
sd(BallsRec[Group==4])
sd(Tackles[Group==4])
sd(Blocks[Group==4])
sd(Clearances[Group==4])
sd(Yellow[Group==4])
sd(Red[Group==4])
sd(FoulsCom[Group==4])
```

```
#Περιγραφικά μέτρα ανά κατηγορία της Result
summary(Goals[Result==0])
summary(Tattempts[Result==0])
summary(OnTarget[Result==0])
summary(Blocked[Result==0])
summary(Corners[Result==0])
summary(Offsides[Result==0])
summary(Possesion[Result==0])
summary(PassAcc[Result==0])
summary(PassT[Result==0])
summary(PassC[Result==0])
summary(Distance[Result==0])
summary(BallsRec[Result==0])
summary(Tackles[Result==0])
summary(Blocks[Result==0])
summary(Clearances[Result==0])
summary(Yellow[Result==0])
summary(Red[Result==0])
summary(FoulsCom[Result==0])
sd(Goals[Result==0])
sd(Tattempts[Result==0])
sd(OnTarget[Result==0])
sd(Blocked[Result==0])
sd(Corners[Result==0])
sd(Offsides[Result==0])
sd(Possesion[Result==0])
sd(PassAcc[Result==0])
sd(PassT[Result==0])
sd(PassC[Result==0])
sd(Distance[Result==0])
sd(BallsRec[Result==0])
sd(Tackles[Result==0])
sd(Blocks[Result==0])
sd(Clearances[Result==0])
sd(Yellow[Result==0])
sd(Red[Result==0])
sd(FoulsCom[Result==0])

summary(Goals[Result==1])
summary(Tattempts[Result==1])
summary(OnTarget[Result==1])
summary(Blocked[Result==1])
summary(Corners[Result==1])
summary(Offsides[Result==1])
summary(Possesion[Result==1])
summary(PassAcc[Result==1])
summary(PassT[Result==1])
summary(PassC[Result==1])
summary(Distance[Result==1])
summary(BallsRec[Result==1])
summary(Tackles[Result==1])
summary(Blocks[Result==1])
summary(Clearances[Result==1])
summary(Yellow[Result==1])
summary(Red[Result==1])
summary(FoulsCom[Result==1])
```

```
sd(Goals[Result==1])
sd(Tattempts[Result==1])
sd(OnTarget[Result==1])
sd(Blocked[Result==1])
sd(Corners[Result==1])
sd(Offsides[Result==1])
sd(Possesion[Result==1])
sd(PassAcc[Result==1])
sd(PassT[Result==1])
sd(PassC[Result==1])
sd(Distance[Result==1])
sd(BallsRec[Result==1])
sd(Tackles[Result==1])
sd(Blocks[Result==1])
sd(Clearances[Result==1])
sd(Yellow[Result==1])
sd(Red[Result==1])
sd(FoulsCom[Result==1])

summary(Goals[Result==2])
summary(Tattempts[Result==2])
summary(OnTarget[Result==2])
summary(Blocked[Result==2])
summary(Corners[Result==2])
summary(Offsides[Result==2])
summary(Possesion[Result==2])
summary(PassAcc[Result==2])
summary(PassT[Result==2])
summary(PassC[Result==2])
summary(Distance[Result==2])
summary(BallsRec[Result==2])
summary(Tackles[Result==2])
summary(Blocks[Result==2])
summary(Clearances[Result==2])
summary(Yellow[Result==2])
summary(Red[Result==2])
summary(FoulsCom[Result==2])
sd(Goals[Result==2])
sd(Tattempts[Result==2])
sd(OnTarget[Result==2])
sd(Blocked[Result==2])
sd(Corners[Result==2])
sd(Offsides[Result==2])
sd(Possesion[Result==2])
sd(PassAcc[Result==2])
sd(PassT[Result==2])
sd(PassC[Result==2])
sd(Distance[Result==2])
sd(BallsRec[Result==2])
sd(Tackles[Result==2])
sd(Blocks[Result==2])
sd(Clearances[Result==2])
sd(Yellow[Result==2])
sd(Red[Result==2])
sd(FoulsCom[Result==2])
```

```
#θηκογράμματα ανά μεταβλητή

par(mfcol=c(1,3))
boxplot(Goals,xlab='Goals')
boxplot(Tattempts,xlab='Tattempts')
boxplot(OnTarget,xlab='OnTarget')
par(mfcol=c(1,1))

par(mfcol=c(1,3))
boxplot(Blocked,xlab='Blocked')
boxplot(Corners,xlab='Corners')
boxplot(Offsides,xlab='Offsides')
par(mfcol=c(1,1))

par(mfcol=c(1,3))
boxplot(PassAcc,xlab='PassAcc')
boxplot(PassT,xlab='PassT')
boxplot(Possesion,xlab='Possesion')
par(mfcol=c(1,1))

par(mfcol=c(1,3))
boxplot(PassC,xlab='PassC')
boxplot(Distance,xlab='Distance')
boxplot(BallsRec,xlab='BallsRec')
par(mfcol=c(1,1))

par(mfcol=c(1,3))
boxplot(Tackles,xlab='Tackles')
boxplot(Blocks,xlab='Blocks')
boxplot(Clearances,xlab='Clearances')
par(mfcol=c(1,1))

par(mfcol=c(1,3))
boxplot(Yellow,xlab='Yellow')
boxplot(Red,xlab='Red')
boxplot(FoulsCom,xlab='FoulsCom')
par(mfcol=c(1,1))

#θηκογράμματα ανά μεταβλητή σε σχέση με την Round

par(mfcol=c(1,2))
boxplot(Goals~Round, xlab='Goals per Round', col=rainbow(2), names=c("Group
Phase", "Knock-out Phase"))
boxplot(Tattempts~Round, xlab='Tattempts per Round', col=rainbow(2),
names=c("Group Phase", "Knock-out Phase"))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(OnTarget~Round, xlab='OnTarget per Round', col=rainbow(2),
names=c("Group Phase", "Knock-out Phase"))
boxplot(Blocked~Round, xlab='Blocked per Round', col=rainbow(2),
names=c("Group Phase", "Knock-out Phase"))
par(mfcol=c(1,1))
```

```
par(mfcol=c(1,2))
boxplot(Corners~Round, xlab='Corners per Round', col=rainbow(2),
names=c("Group Phase", "Knock-out Phase"))
boxplot(Offsides~Round, xlab='Offsides per Round', col=rainbow(2),
names=c("Group Phase", "Knock-out Phase"))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(PassAcc~Round, xlab='PassAcc per Round', col=rainbow(2),
names=c("Group Phase", "Knock-out Phase"))
boxplot(PassT~Round, xlab='PassT per Round', col=rainbow(2), names=c("Group
Phase", "Knock-out Phase"))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(Possesion~Round, xlab='Possesion per Round', col=rainbow(2),
names=c("Group Phase", "Knock-out Phase"))
boxplot(PassC~Round, xlab='PassC per Round', col=rainbow(2), names=c("Group
Phase", "Knock-out Phase"))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(Distance~Round, xlab='Distance per Round', col=rainbow(2),
names=c("Group Phase", "Knock-out Phase"))
boxplot(BallsRec~Round, xlab='BallsRec per Round', col=rainbow(2),
names=c("Group Phase", "Knock-out Phase"))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(Tackles~Round, xlab='Tackles per Round', col=rainbow(2),
names=c("Group Phase", "Knock-out Phase"))
boxplot(Blocks~Round, xlab='Blocks per Round', col=rainbow(2),
names=c("Group Phase", "Knock-out Phase"))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(Clearances~Round, xlab='Clearances per Round', col=rainbow(2),
names=c("Group Phase", "Knock-out Phase"))
boxplot(Yellow~Round, xlab='Yellow per Round', col=rainbow(2),
names=c("Group Phase", "Knock-out Phase"))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(Red~Round, xlab='Red per Round', col=rainbow(2), names=c("Group
Phase", "Knock-out Phase"))
boxplot(FoulsCom~Round, xlab='FoulsCom per Round', col=rainbow(2),
names=c("Group Phase", "Knock-out Phase"))
par(mfcol=c(1,1))

#θηκογράμματα ανά μεταβλητή σε σχέση με την Home.Away

par(mfcol=c(1,2))
boxplot(Goals~Home.Away, xlab='Goals per Home.Away', col=c(3,4),
names=c("Away", "Home"))
boxplot(Tattempts~Home.Away, xlab='Tattempts per Home.Away', col=c(3,4),
```

```
names=c("Away", "Home"))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(OnTarget~Home.Away, xlab='OnTarget per Home.Away', col=c(3,4),
names=c("Away", "Home"))
boxplot(Blocked~Home.Away, xlab='Blocked per Home.Away', col=c(3,4),
names=c("Away", "Home"))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(Corners~Home.Away, xlab='Corners per Home.Away', col=c(3,4),
names=c("Away", "Home"))
boxplot(Offsides~Home.Away, xlab='Offsides per Home.Away', col=c(3,4),
names=c("Away", "Home"))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(Possesion~Home.Away, xlab='Possesion per Home.Away', col=c(3,4),
names=c("Away", "Home"))
boxplot(PassAcc~Home.Away, xlab='PassAcc per Home.Away', col=c(3,4),
names=c("Away", "Home"))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(PassT~Home.Away, xlab='PassT per Home.Away', col=c(3,4),
names=c("Away", "Home"))
boxplot(PassC~Home.Away, xlab='PassC per Home.Away', col=c(3,4),
names=c("Away", "Home"))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(Distance~Home.Away, xlab='Distance per Home.Away', col=c(3,4),
names=c("Away", "Home"))
boxplot(BallsRec~Home.Away, xlab='BallsRec per Home.Away', col=c(3,4),
names=c("Away", "Home"))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(Tackles~Home.Away, xlab='Tackles per Home.Away', col=c(3,4),
names=c("Away", "Home"))
boxplot(Blocks~Home.Away, xlab='Blocks per Home.Away', col=c(3,4),
names=c("Away", "Home"))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(Clearances~Home.Away, xlab='Clearances per Home.Away', col=c(3,4),
names=c("Away", "Home"))
boxplot(Yellow~Home.Away, xlab='Yellow per Home.Away', col=c(3,4),
names=c("Away", "Home"))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(Red~Home.Away, xlab='Red per Home.Away', col=c(3,4), names=c("Away",
"Home"))
```

```
boxplot(FoulsCom~Home.Away, xlab='FoulsCom per Home.Away', col=c(3,4),
names=c("Away", "Home"))
par(mfcol=c(1,1))

#θηκογράμματα ανά μεταβλητή σε σχέση με την Result

par(mfcol=c(1,2))
boxplot(Goals~Result, xlab='Goals per Result', col=c(6,7,8), names=c("Lose",
"Draw", "Win"))
boxplot(Tattempts~Result, xlab='Tattempts per Result', col=c(6,7,8),
names=c("Lose", "Draw", "Win"))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(OnTarget~Result, xlab='OnTarget per Result', col=c(6,7,8),
names=c("Lose", "Draw", "Win"))
boxplot(Blocked~Result, xlab='Blocked per Result', col=c(6,7,8),
names=c("Lose", "Draw", "Win"))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(Corners~Result, xlab='Corners per Result', col=c(6,7,8),
names=c("Lose", "Draw", "Win"))
boxplot(Offsides~Result, xlab='Offsides per Result', col=c(6,7,8),
names=c("Lose", "Draw", "Win"))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(Possesion~Result, xlab='Possesion per Result', col=c(6,7,8),
names=c("Lose", "Draw", "Win"))
boxplot(PassAcc~Result, xlab='PassAcc per Result', col=c(6,7,8),
names=c("Lose", "Draw", "Win"))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(PassT~Result, xlab='PassT per Result', col=c(6,7,8), names=c("Lose",
"Draw", "Win"))
boxplot(PassC~Result, xlab='PassC per Result', col=c(6,7,8), names=c("Lose",
"Draw", "Win"))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(Distance~Result, xlab='Distance per Result', col=c(6,7,8),
names=c("Lose", "Draw", "Win"))
boxplot(BallsRec~Result, xlab='BallsRec per Result', col=c(6,7,8),
names=c("Lose", "Draw", "Win"))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(Tackles~Result, xlab='Tackles per Result', col=c(6,7,8),
names=c("Lose", "Draw", "Win"))
boxplot(Blocks~Result, xlab='Blocks per Result', col=c(6,7,8),
names=c("Lose", "Draw", "Win"))
par(mfcol=c(1,1))
```

```
par(mfcol=c(1,2))
boxplot(Clearances~Result, xlab='Clearances per Result', col=c(6,7,8),
names=c("Lose", "Draw", "Win"))
boxplot(Yellow~Result, xlab='Yellow per Result', col=c(6,7,8),
names=c("Lose", "Draw", "Win"))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(Red~Result, xlab='Red per Result', col=c(6,7,8), names=c("Lose",
"Draw", "Win"))
boxplot(FoulsCom~Result, xlab='FoulsCom per Result', col=c(6,7,8),
names=c("Lose", "Draw", "Win"))
par(mfcol=c(1,1))

#θηκογράμματα ανά μεταβλητή σε σχέση με την Group

par(mfcol=c(1,2))
boxplot(Goals~Group, xlab='Goals per Group', col=c(8,10,11,12))
boxplot(Tattempts~Group, xlab='Tattempts per Group', col=c(8,10,11,12))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(OnTarget~Group, xlab='OnTarget per Group', col=c(8,10,11,12))
boxplot(Blocked~Group, xlab='Blocked per Group', col=c(8,10,11,12))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(Corners~Group, xlab='Corners per Group', col=c(8,10,11,12))
boxplot(Offsides~Group, xlab='Offsides per Group', col=c(8,10,11,12))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(Possesion~Group, xlab='Possesion per Group', col=c(8,10,11,12))
boxplot(PassAcc~Group, xlab='PassAcc per Group', col=c(8,10,11,12))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(PassT~Group, xlab='PassT per Group', col=c(8,10,11,12))
boxplot(PassC~Group, xlab='PassC per Group', col=c(8,10,11,12))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(Distance~Group, xlab='Distance per Group', col=c(8,10,11,12))
boxplot(BallsRec~Group, xlab='BallsRec per Group', col=c(8,10,11,12))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(Tackles~Group, xlab='Tackles per Group', col=c(8,10,11,12))
boxplot(Blocks~Group, xlab='Blocks per Group', col=c(8,10,11,12))
par(mfcol=c(1,1))

par(mfcol=c(1,2))
boxplot(Clearances~Group, xlab='Clearances per Group', col=c(8,10,11,12))
boxplot(Yellow~Group, xlab='Yellow per Group', col=c(8,10,11,12))
par(mfcol=c(1,1))
```

```
par(mfcol=c(1,2))
boxplot(Red~Group, xlab='Red per Group', col=c(8,10,11,12))
boxplot(FoulsCom~Group, xlab='FoulsCom per Group', col=c(8,10,11,12))
par(mfcol=c(1,1))

#Διαγράμματα διασποράς

library(ggplot2)

ggplot(data,aes(x=PassAcc,y=Distance))+
  geom_point()

y1<-lm(Possesion~PassT, data=data)
ggplot(data,aes(x=PassT,y=Possesion))+
  geom_point()+
  geom_line(aes(x=PassT,y=predict(y1)), colour='blue')

y2<-lm(Possesion~PassC, data=data)
ggplot(data,aes(x=PassC,y=Possesion))+
  geom_point()+
  geom_line(aes(x=PassC,y=predict(y2)), colour='blue')

y3<-lm(PassT~PassC, data=data)
ggplot(data,aes(x=PassC,y=PassT))+
  geom_point()+
  geom_line(aes(x=PassC,y=predict(y3)), colour='blue')

ggplot(data,aes(x=Tattempts,y=Goals))+
  geom_point()

ggplot(data,aes(x=OnTarget,y=Goals))+
  geom_point()

ggplot(data,aes(x=FoulsCom,y=GoalsCon))+
  geom_point()

#Κεφάλαιο 2
# Έλεγχοι Κανονικότητας

library(nortest)
lillie.test(Goals)
lillie.test(Tattempts)
lillie.test(OnTarget)
lillie.test(Blocked)
lillie.test(Corners)
lillie.test(Offsides)
lillie.test(Possesion)
lillie.test(PassAcc)
lillie.test(PassT)
lillie.test(PassC)
lillie.test(Distance)
lillie.test(BallsRec)
lillie.test(Tackles)
lillie.test(Blocks)
lillie.test(Clearances)
```

```
lillie.test(Yellow)
lillie.test(Red)
lillie.test(FoulsCom)

# Συσχετίσεις
cor(data[,c(2:16,19,24)], method="pearson")
cor(data[,c(2:16,19)], data[,c(17,18,20:25)], method="spearman")
table(Group,Round)
chisq.test(Group, Round)
table(Group, Result)
chisq.test(Group, Result)
table(Home.Away, Result)
chisq.test(Home.Away, Result)

#Δημιουργία μεταβλητής BallsRec2
BallsRec2=NULL
for (i in 1:742){
  if (BallsRec[i]<=42){
    BallsRec2[i]=0
  } else if (BallsRec[i]<=53) {
    BallsRec2[i]=1
  } else {
    BallsRec2[i]=2
  }
}
data<-cbind(data,BallsRec2)
table(BallsRec2,Result)
chisq.test(BallsRec2,Result)

#Δημιουργία μεταβλητής Possesion2
Possesion2=NULL
for (i in 1:742){
  if (Possesion[i]<=42){
    Possesion2[i]=1
  } else if (Possesion[i]<=50) {
    Possesion2[i]=2
  } else if (Possesion[i]<=57.75) {
    Possesion2[i]=3
  } else {
    Possesion2[i]=4
  }
}
data<-cbind(data,Possesion2)
table(Possesion2,Group)
chisq.test(Possesion2,Group)
table(Possesion2,Result)
chisq.test(Possesion2,Result)

#Δημιουργία μεταβλητής PassAcc2
PassAcc2=NULL
for (i in 1:742){
  if (PassAcc[i]<=80){
    PassAcc2[i]=1
  } else if (PassAcc[i]<=85) {
    PassAcc2[i]=2
```

```
} else if (PassAcc[i]<=88) {
    PassAcc2[i]=3
  } else {
    PassAcc2[i]=4
  }
}
data<-cbind(data,PassAcc2)
table(PassAcc2,Group)
chisq.test(PassAcc2,Group)
table(PassAcc2,Result)
chisq.test(PassAcc2,Result)

#Δημιουργία μεταβλητής PassT2
PassT2=NULL
for (i in 1:742){
  if (PassT[i]<=400){
    PassT2[i]=1
  } else if (PassT[i]<=497) {
    PassT2[i]=2
  } else if (PassT[i]<=599.8) {
    PassT2[i]=3
  } else {
    PassT2[i]=4
  }
}
data<-cbind(data,PassT2)
table(PassT2,Group)
chisq.test(PassT2,Group)
table(PassT2,Result)
chisq.test(PassT2,Result)

#Δημιουργία μεταβλητής PassC2
PassC2=NULL
for (i in 1:742){
  if (PassC[i]<=321.5){
    PassC2[i]=1
  } else if (PassC[i]<=416) {
    PassC2[i]=2
  } else if (PassC[i]<=528) {
    PassC2[i]=3
  } else {
    PassC2[i]=4
  }
}
data<-cbind(data,PassC2)
table(PassC2,Group)
chisq.test(PassC2,Group)
table(PassC2,Result)
chisq.test(PassC2,Result)

#Κεφάλαιο 3
#Feature Selection
library(randomForest)
```

```
features<-randomForest(Result~., data=data)
importance(features)
varImpPlot(features)

features<-randomForest(Result~Tattempts+ OnTarget+ Blocked+ Corners
                       + Offsides+ Possesion+ PassAcc+ PassT+ PassC
                       + Distance+ BallsRec+ Clearances
                       + Home.Away+ Group+ FoulsCom+ Red+ Yellow
                       + Blocks+ Ranking+ Round+ Tackles, data=data)
importance(features)
varImpPlot(features)

features<-randomForest(Goals~., data=data)
importance(features)
varImpPlot(features)

features<-randomForest(Goals~Tattempts+ Blocked+ Corners
                       + Offsides+ Possesion+ PassAcc+ PassT+ PassC
                       + Distance+ BallsRec+ Clearances+ GoalsCon
                       + Home.Away+ Group+ FoulsCom+ Red+ Yellow
                       + Blocks+ Ranking+ Round+ Tackles, data=data)
importance(features)
varImpPlot(features)

features<-randomForest(Goals~Tattempts+ Blocked+ Corners
                       + Offsides+ Possesion+ PassAcc
                       + Distance+ BallsRec+ Clearances+ GoalsCon
                       + Home.Away+ Group+ FoulsCom+ Red+ Yellow
                       + Blocks+ Ranking+ Round+ Tackles, data=data)
importance(features)
varImpPlot(features)

#Classification για την Result
set.seed(123)
data$Result<-factor(data$Result, levels = c(0,1,2))
library(caTools)
split = sample.split(data$Result, SplitRatio = 0.70)
training_set = subset(data, split == TRUE)
test_set = subset(data, split == FALSE)
training_set[1:18] = scale(training_set[1:18])
test_set[1:18] = scale(test_set[1:18])

#Naive Bayes
library(e1071)
classifier<- naiveBayes(x = training_set[c(1:3, 15, 21, 23)],
                        y = training_set$Result)
y_pred <- predict(classifier, newdata = test_set[c(1:3, 15, 21, 23)])
cm <- table(test_set[, 24], y_pred); cm

classifier<- naiveBayes(x = training_set[c(2:3, 15, 21)],
                        y = training_set$Result)
y_pred <- predict(classifier, newdata = test_set[c(2:3, 15, 21)])
cm <- table(test_set[, 24], y_pred); cm
```

```
classifier <- naiveBayes(x = training_set[c(1:23)],
                         y = training_set$Result)
y_pred <- predict(classifier, newdata = test_set[c(1:23)])
cm <- table(test_set[, 24], y_pred); cm

#Support Vector Machine
classifier <- svm(formula = Result ~ .,
                  data = training_set,
                  type = 'C-classification',
                  kernel = 'radial')
y_pred <- predict(classifier, newdata = test_set[-24])
cm<- table(test_set[, 24], y_pred); cm

classifier <- svm(formula = Result ~
Goals+GoalsCon+Tattempts+OnTarget+Ranking+Clearances,
                  data = training_set,
                  type = 'C-classification',
                  kernel = 'radial')
cm <- table(test_set[, 24], y_pred); cm
y_pred <- predict(classifier, newdata = test_set)

classifier <- svm(formula = Result ~ Goals+GoalsCon,
                  data = training_set,
                  type = 'C-classification',
                  kernel = 'radial')
y_pred <- predict(classifier, newdata = test_set)
cm <- table(test_set[, 24], y_pred); cm

classifier <- svm(formula = Result ~ Tattempts+OnTarget+Ranking+Clearances,
                  data = training_set,
                  type = 'C-classification',
                  kernel = 'radial')
y_pred <- predict(classifier, newdata = test_set)
cm <- table(test_set[, 24], y_pred); cm

#Κεφάλαιο 4
#Δημιουργία Μεταβλητής Result2
Result2=NULL
for (i in 1:742){
  if (Result[i]==2 | Result[i]==1){
    Result2[i]=1
  } else {
    Result2[i]=0
  }
}
data<-cbind(data,Result2)


#Δημιουργία Μεταβλητής Result3
Result3=NULL
for (i in 1:742){
  if (Result[i]==2){
    Result3[i]=1
  } else {
    Result3[i]=0
Round<-factor(Round)
```

```
  }
}
data<-cbind(data,Result3)
Home.Away<-factor(Home.Away)

#Λογιστική Παλινδρόμηση για την μεταβλητή Result2
library(DAAG)
library(ResourceSelection)
model11<-glm(Result2~OnTarget+Tattempts+Blocked+Corners+Possesion
             +PassAcc+PassT+PassC+Distance+BallsRec+Tackles+Blocks+Yellow+Red
             +FoulsCom+Home.Away+Round+Group+Ranking+Clearances,
family=binomial)
vif(model11)

model12<-glm(Result2~OnTarget+Tattempts+Blocked+Corners+Possesion
              +PassAcc+Distance+BallsRec+Tackles+Blocks+Yellow+Red
              +FoulsCom+Home.Away+Round+Group+Ranking+Clearances,
family=binomial)
step(model12)

model13<-glm(Result2~OnTarget+Corners+Possesion
             +BallsRec+Yellow+FoulsCom+Home.Away+Round
             +Group+Ranking+Clearances, family=binomial)
summary(model13)

model14<-glm(Result2~OnTarget+Corners+Possesion
             +BallsRec+FoulsCom+Home.Away+Round
             +Group+Ranking+Clearances, family=binomial)
summary(model14)

model15<-glm(Result2~OnTarget+Possesion+BallsRec+Ranking
              +Clearances+Home.Away+Corners+Round, family=binomial)
summary(model15)
anova(model15)
1-pchisq(708.07,733)

model16<-glm(Result3~OnTarget+Ranking+Clearances+Possesion
              +Home.Away+Corners+BallsRec+Round, family=binomial)
summary(model16)
anova(model16)

model17<-glm(Result2~OnTarget+Possesion+BallsRec+Ranking
              +Clearances+Home.Away+Corners, family=binomial)
summary(model17)
anova(model17)
table(Result, Round)

res<-resid(model17, type = 'deviance')
fit<-fitted(model17)
plot(fit, res, xlab='Fitted Values', ylab='Residuals')

y<-Result2
model172<-glm(y~OnTarget+Possesion+BallsRec+Ranking
              +Clearances+Home.Away+Corners, family=binomial)
```

```
hoslem.test(model172$y, fitted(model172))
#Έλεγχος αλληλεπιδράσεων

model18<-glm(Result2~OnTarget*Possesion*BallsRec*Ranking
             *Clearances*Home.Away*Corners, family=binomial)
summary(model18)
anova(model18)

model19<-glm(Result2~OnTarget+Possesion+BallsRec+Ranking
             +Clearances+Home.Away+Corners+Clearances:Home.Away
             +BallsRec:Home.Away+Ranking:Clearances, family=binomial)
anova(model19)
model110<-glm(Result2~OnTarget+Possesion+BallsRec+Ranking
             +Clearances+Home.Away+Corners
             +BallsRec:Home.Away, family=binomial)
anova(model110)
summary(model110)

#Poisson για την μεταβλητή Goals
library(DAAG)
library(ResourceSelection)
model21<-glm(Goals~OnTarget+Tattempts+Blocked+Corners+Possesion

+PassAcc+PassT+PassC+Distance+BallsRec+Tackles+Blocks+Yellow+Red
             +FoulsCom+Home.Away+Round+Group+Ranking+Clearances,
family=poisson)
vif(model21)

model22<-glm(Goals~OnTarget+Tattempts+Blocked+Corners+Possesion
             +PassAcc+Distance+BallsRec+Tackles+Blocks+Yellow+Red
             +FoulsCom+Home.Away+Round+Group+Ranking+Clearances,
family=poisson)
step(model22)

model23<-glm(Goals~OnTarget+Tattempts+Ranking
             +Corners+Home.Away+Possesion+Clearances, family=poisson)
summary(model23)

model24<-glm(Goals~Corners+Home.Away+Possesion+OnTarget+Ranking
             +Clearances, family=poisson)
summary(model24)
anova(model24)
1-pchisq(663.98,735) #Έλεγχος επάρκειας

#Έλεγχος αλληλεπιδράσεων

model25<-glm(Goals~OnTarget*Group*Corners*Home.Away
             *Possesion*Clearances, family=poisson)
step(model25)
anova(model25)

#Διακριτοποίηση Μεταβλητής OnTarget
```

```
quantile(OnTarget2, probs = seq(0, 1, 0.2))
OnTarget2<-(OnTarget-4.682)/2.841978
OnTarget_Cat=NULL
for (i in 1:742){
  if (OnTarget2[i]<=-0.9437089){
    OnTarget_Cat[i]=0
  } else if (OnTarget2[i]>-0.9437089 & OnTarget2[i]<=-0.239974) {
    OnTarget_Cat[i]=1
} else if (OnTarget2[i]>-0.239974 & OnTarget2[i]<=0.1118939) {
    OnTarget_Cat[i]=2
  } else if (OnTarget2[i]>0.1118939 & OnTarget2[i]<=0.8156291) {
    OnTarget_Cat[i]=3
  } else {
    OnTarget_Cat[i]=4
  }
}
data<-cbind(data,OnTarget_Cat)
OnTarget_Cat<-factor(OnTarget_Cat)

OnTarget_Cat1=NULL
for (i in 1:742){
  if (OnTarget[i]<=3){
    OnTarget_Cat1[i]=0
  } else if (OnTarget[i]>3 & OnTarget[i]<=4) {
    OnTarget_Cat1[i]=1
  } else if (OnTarget[i]>4 & OnTarget[i]<=6) {
    OnTarget_Cat1[i]=2
  } else {
    OnTarget_Cat1[i]=3
  }
}

data<-cbind(data,OnTarget_Cat1)
OnTarget_Cat1<-factor(OnTarget_Cat1)

OnTarget_Cat2=NULL
for (i in 1:742){
  if (OnTarget[i]<=2){
    OnTarget_Cat2[i]=0
  } else if (OnTarget[i]>2 & OnTarget2[i]<=4) {
    OnTarget_Cat2[i]=1
  } else if (OnTarget2[i]>4 & OnTarget2[i]<=5) {
    OnTarget_Cat2[i]=2
  } else if (OnTarget2[i]>5 & OnTarget2[i]<=7) {
    OnTarget_Cat2[i]=3
  } else {
    OnTarget_Cat2[i]=4
  }
}
data<-cbind(data,OnTarget_Cat2)
OnTarget_Cat2<-factor(OnTarget_Cat2)

#Μοντέλα με αλληλεπίδραση για τη διακριτοποιημένη OnTarget
model26<-glm(Goals~OnTarget_Cat+Corners+Ranking+Home.Away+Possesion
             +Clearances+OnTarget_Cat:Ranking , family=poisson)
```

```
anova(model26)

model27<-glm(Goals~OnTarget_Cat1+Ranking+Corners+Home.Away
             +Possesion+Clearances+OnTarget_Cat1:Ranking, family=poisson)
anova(model27)

model28<-glm(Goals~OnTarget_Cat2+Ranking+Corners+Home.Away+Possesion
             +Clearances+OnTarget_Cat2:Ranking, family=poisson)
anova(model28)

model29<-glm(Goals~Corners+Group+Home.Away+Possesion+OnTarget
             +Clearances+OnTarget:Group , family=poisson)
anova(model29)
```

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική

Ανζουλάκος Δ. (2017) Ανάλυση Δεδομένων με την Χρήση Στατιστικών Πακέτων : Εισαγωγή στην R, Σημειώσεις για το μάθημα «Ανάλυση Δεδομένων με χρήση στατιστικών πακέτων» του ΠΜΣ στην Εφαρμοσμένη Στατιστική του Πανεπιστημίου Πειραιώς.

Δαμιανού Χ., Κούτρας Μ. (1991) Εισαγωγή στη Στατιστική, Μέρος Ι, Εκδόσεις Συμμετρία

Δαμιανού Χ., Κούτρας Μ. (1996) Εισαγωγή στη Στατιστική, Μέρος ΙΙ, Εκδόσεις Συμμετρία

Καλλιακμάνης Δ. (2020) Στατιστικά μοντέλα για την απόδοση μιας ομάδας μπάσκετ: ποια στατιστικά στοιχεία είναι καθοριστικά για την απόδοση της ομάδας, σε ετήσια βάση.

Κούτρας Μ. (2012) Εισαγωγή στη θεωρία πιθανοτήτων και εφαρμογές, Εκδόσεις Σταμούλη

Κούτρας Μ. (2019) Ανάλυση Παλινδρόμησης και ανάλυση διακύμανσης, Σημειώσεις για το μάθημα «Ανάλυση Παλινδρόμησης και ανάλυση διακύμανσης» του ΠΜΣ στην Εφαρμοσμένη Στατιστική του Πανεπιστημίου Πειραιώς.

Πελέκης Ν. (2019) Στατιστικές Μέθοδοι Εξόρυξης Δεδομένων, Σημειώσεις για το μάθημα «Στατιστικές Μέθοδοι Εξόρυξης Δεδομένων» του ΠΜΣ στην Εφαρμοσμένη Στατιστική του Πανεπιστημίου Πειραιώς.

Πολίτης Κ. (2020) Γενικευμένα Γραμμικά Μοντέλα, Σημειώσεις του μαθήματος «Γενικευμένα Γραμμικά Μοντέλα» του ΠΜΣ στην Εφαρμοσμένη Στατιστική του Πανεπιστημίου Πειραιώς.

Ξένη

Agresti, A. (2002) Categorical Data Analysis 2nd Edition, Wiley-Interscience, Florida.

Akaike, H. (1974) A New Look at the Statistical Model Identification, IEEE Transactions on automatic control, Vol. AC-19, December 1974.

Armatas, V., Yiannakos, A., Zaggelidis,G., Papadopoulou S., Fragkos, N. (2009) GOAL SCORING PATTERNS IN GREEK TOP LEVELED SOCCER MATCHES, Journal of Physical Education an Sport, Vol 23, No. 2

Barreira, D., Garganta, J., Guimaraẽs, P., Machado, J., Anguera, M. T. (2014) Ball recovery patterns as a performance indicator in elite soccer, J Sports Engineering and Technology, Vol. 228(1) 61–72.

Bekris, E., Mylonis, E., Sarakinos, A., Gissis, I., Gioldasis, A. , Sotiropoulos, A. (2013) Offense and defense statistical indicators that determine the Greek Superleague teams placement on the table 2011 – 12, Journal of Physical Education and Sport, 13(3), Art 55, pp 338 – 347.

Chen, R. C. , Dewi, C., Huang, S. W., Caraka, R. E. (2020) Selecting critical features for data classification based on machine learning methods, Journal of Big Data (2020) 7:52

Dobson, A. J., (2001) An Introduction to Generalized Linear Models, 2nd Edition, Chapman & Hall, Florida.

Dunham, M. (2003) Data Mining – Introductory and Advanced Topics, Prentice Hall

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996) From Data Mining to Knowledge Discovery in Databases, AI Magazine, Vol. 17, No. 3.

García-Gonzalo, E., Fernández-Muñiz, Z., Nieto, P. J. G., Sánchez, A. B., Fernández, M. M (2016) Hard-Rock Stability Analysis for Span Design in Entry-Type Excavations with Learning Classifiers, Materials

Granik, M. & Mesyura, V. (2017) Fake News Detection Using Naive Bayes Classifier, IEEE First Ukraine Conference on Electrical and Computer Engineering.

Guyon, I. & Elisseeff, A. (2003) An Introduction to Variable and Feature Selection, Journal of Machine Learning Research 3, 1157-1182.

Hassani, H., Saporta, G., Sirimal, Silva, E. (2014) Data Mining and Official Statistics: The Past, the Present and the Future, Big Data, March 2014, Vol. 2, No. 1.

Hosmer, D. & Lemeshow, S. (2000) Applied Logistic Regression, John Wiley & Sons, USA.

Hosmer, D., Hosmer, T., Le Cessie, S., Lemeshow, S. (1997) A comparison of goodness-of-fit tests for the logistic regression model, Statistics in medicine, Vol. 16, 965-980.

Jaccard, J. & Turrisi, R. (2003) Interaction effects in multiple regression, 2nd edition, SAGE Publications, Iowa City.

Karlis, D. & Ntzoufras, I. (2000) On Modelling Soccer Data, Student, Vol. 3, No. 4, 229-244.

Lago-Peñas, C., Lago-Ballesteros, J. (2011) Game location and team quality effects on performance profiles in professional soccer, Journal of Sports Science and Medicine, 10, 465-471.

Leontijević, B., Janković, A., Tomić, L. (2018) Attacking performance profile of football teams in different national leagues according to UEFA rankings for club competitions, Physical Education and Sport, Vol. 16, No 3, 697 – 708.

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

Liu, H., Hopkins, W., Gómez, M. A., Molinuevo, J. S. (2013) Inter-operator reliability of live football match statistics from OPTA Sportsdata, International Journal of Performance Analysis in Sport, 13, 803-821.

Liu, H., Yi, Q., Giménez, J. V., Gómez, M. A., Lago-Peñas, C. (2015) Performance profiles of football teams in the UEFA Champions League considering situational efficiency, International Journal of Performance Analysis in Sport, 15, 371-390.

Mackenzie, R. & Cushion, C. (2013) Performance analysis in football: A critical review and implications for future research, Journal of Sports Sciences, Vol. 31, No. 6.

Mitchell, T. (1997) Machine Learning, McGraw-Hill, New York.

Mundry, R. (2014) Statistical Issues and Assumptions of Phylogenetic Generalized Least Squares, Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology, Springer-Verlag, Berlin.

Panaretos, V. (2002) A Statistical Analysis of the European Soccer Champions League, Joint Statistical Meetings - Section on Statistics in Sports.

Reif, D., Motsinger, A., McKinney, B., Crowe, Jr. J., Moore, J. (2014) Feature Selection using a Random Forests Classifier for the Integrated Analysis of Multiple Data Types

Sammut, C. & Webb, G. (2011) Encyclopedia of Machine Learning, Springer, New York.

Senaviratna, N. A. M. R., Cooray, T. M. J. A. (2019) Diagnosing Multicollinearity of Logistic Regression Model, Asian Journal of Probability and Statistics, Vol. 5, No. 2, 1-9.

Szwarc, A. (2007) Efficacy of Successful and Unsuccessful Soccer Teams Taking Part in Finals of Champions League, Medsportpress, Volume 13, No. 2, 221-225.

Tan P. N., Steinbach M., Kumar V. (2005) Introduction to Data Mining, Pearson, Minnesota.

Ting, S.L., Ip, W.H., Tsang, A. (2011) Is Naïve Bayes a Good Classifier for Document Classification?, International Journal of Software Engineering and Its Applications Vol. 5, No. 3.

Yiannakos, A. & Armatas, V. (2006) Evaluation of the goal scoring patterns in European Championship in Portugal 2004, International Journal of Performance Analysis in Sport, Vol. 6, No. 1, 178-188.

Σύνδεσμοι

https://www.youtube.com/watch?v=Sy2vc9lW5r0

https://www.uefa.com

https://www.researchgate.net/figure/A-F-Scatter-plots-with-data-sampled-from-simulated-bivariate-normal-distributions-with_fig1_323388613

https://www.listendata.com/2014/08/how-to-read-box-plot.html

https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

https://www.semanticscholar.org/