ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

# Οικονομική Βελτιστοποίηση Εκμετάλλευσης Προβλέψεων Ποδοσφαίρου από Μηχανική Μάθηση

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

**Αριστείδη Βασιλάκη**

**Επιβλέπων :** Ασκούνης Δημήτριος
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2022

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

# Οικονομική Βελτιστοποίηση Εκμετάλλευσης Προβλέψεων Ποδοσφαίρου από Μηχανική Μάθηση

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

**Αριστείδη Βασιλάκη**

**Επιβλέπων :** Ασκούνης Δημήτριος
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 12η Οκτωβρίου 2022.

| | | |
|---|---|---|
| ..... | ..... | ..... |
| Δημήτριος Ασκούνης | Ιωάννης Ψαρράς | Χρυσόστομος Δούκας |
| Καθηγητής Ε.Μ.Π. | Καθηγητής Ε.Μ.Π. | Καθηγητής Ε.Μ.Π. |

Αθήνα, Οκτώβριος 2022

.....

**Αριστείδης Βασιλάκης**

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

© 2022 – All rights reserved

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

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

# Περίληψη

Η εξαγωγή προβλέψεων φαβορί ποδοσφαιρικών αναμετρήσεων με τη χρήση βελτιστοποιημένων μοντέλων Μηχανικής Μάθησης αποτελεί αντικείμενο μεγάλου αριθμού ερευνητικών δημοσιεύσεων. Όμως, δεν έχει παρουσιαστεί επαρκής έρευνα που να καταδεικνύει τα μοντέλα μηχανικής μάθησης που είναι καταλληλότερα στην εξαγωγή πιθανοτήτων που αντιστοιχούν στην εκπλήρωση των πιθανών αποτελεσμάτων για κάθε αγώνα. Η εργασία θα παρουσιάσει μια μεθοδολογία εξαγωγής βέλτιστων ποσών τοποθέτησης στο παίγνιο του στοιχήματος που βασίζεται στο κριτήριο Kelly. Η ποιότητα των προγνώσεων πιθανοτήτων είναι απόλυτα συνδεδεμένη με τη κερδοφορία που επιτυγχάνεται με τη χρήση αυτής της μεθόδου. Με αυτόν τον τρόπο προκύπτουν συμπεράσματα για τους ταξινομητές. Ταυτόχρονα, αποδεικνύεται πως η προσέγγιση του Kelly Criterion για το παίγνιο αποτελεί τη βέλτιστη οπότε η εργασία παρουσιάζει και μια ρεαλιστική μελέτη που αφορά τη βέλτιστη χρήση αυτών των προβλέψεων στο παίγνιο. Λαμβάνονται αποτελέσματα από προσομοιώσεις παιξίματος με τη χρήση ιστορικών δεδομένων για 3 διαφορετικούς συνδυασμούς μοντέλων Μηχανικής Μάθησης και της τεχνικής εξαγωγής ποσών τοποθέτησης στο παίγνιο που εισάγει η εργασία για τα κύρια πρωταθλήματα 4 μεγάλων ευρωπαϊκών κρατών. Τα αποτελέσματα που προκύπτουν είναι αρκετά ικανοποιητικά.



**Λέξεις Κλειδιά:**

Μηχανική Μάθηση, προβλέψεις, πιθανότητες, στοίχημα , κέρδος, ευρωπαϊκά πρωταθλήματα, κριτήριο Kelly, ποδόσφαιρο

# Abstract

The objective of determining the winner of upcoming European football matches with the aid of Machine Learning algorithms is a topic which many researchers have focused on. However, little amount of research has focused on determining the probabilities of occurrence for each end result of a match. In this study, a comprehensive methodology of determining the optimal wagers when betting is presented which is based on the Kelly Criterion. The performance of this methodology is absolutely connected with the quality of the probability predictions. Thus, conclusions can be made regarding the best performing classifiers in that sense. At the same time, it is mathematically proven that the Kelly Criterion's approach is optimal regarding the betting game so this study also presents comprehensive results as to the optimal use of the predictions to maximize profitability. In the end, very satisfactory results are presented from simulations that correspond to 4 major European leagues using 4 different combinations of Machine Learning models and the aforementioned technique for determining the optimal wager.



**Keywords:**

Machine Learning, predictions, probabilities, betting, profit, European Championships, Kelly Criterion, European football.

# Ευχαριστίες

Η εργασία αυτή αφιερώνεται σε όλους όσοι στάθηκαν δίπλα μου κατά τη διάρκεια των τελευταίων 6 ετών εντός και εκτός του πανεπιστημίου. Φίλοι και οικογένεια βοήθησαν καθοριστικά με ποικίλους τρόπους έτσι ώστε αυτή η πορεία να αποδειχθεί πολύ πιο πλούσια σε εμπειρίες και γνώση από ό,τι διαφαινόταν στην αρχή της.

Θέλω επίσης να ευχαριστήσω τους κ.κ. Κωνσταντίνο Αλεξάκη και Δημήτριο Ασκούνη, Καθηγητή Ε.Μ.Π., για την ευκαιρία να ασχοληθώ με ένα τόσο ενδιαφέρον θέμα εντός των πλαισίων της σχολής και την πολύ καλή συνεργασία μας καθ' όλη τη διάρκεια της εκπόνησης αυτής της εργασίας κατά τους τελευταίους 7 μήνες. Επίσης, θα ήθελα να ευχαριστήσω και τους κ.κ. Ιωάννη Ψαρρά και Χρυσόστομο Δούκα, Καθηγητές Ε.Μ.Π., για τη συμμετοχή τους στην τριμελή επιτροπή.

# Περιεχόμενα

**Οικονομική Βελτιστοποίηση Εκμετάλλευσης Προβλέψεων Ποδοσφαίρου από Μηχανική Μάθηση** ..... **1**

**Οικονομική Βελτιστοποίηση Εκμετάλλευσης Προβλέψεων Ποδοσφαίρου από Μηχανική Μάθηση** ..... **3**

**1 Εισαγωγή** ..... **1**

1.1 Θεμελίωση και Περιγραφή της Εργασίας ..... 1

1.2 Οργάνωση της Εργασίας ..... 4

**2 Θεωρητικό Υπόβαθρο** ..... **5**

2.1 Μηχανική Μάθηση (Machine Learning) ..... 5

*2.1.1 Gaussian Naive Bayes (GNB)* ..... *7*

*2.1.2 Linear Discriminant Analysis (LDA)* ..... *8*

*2.1.3 Quadratic Discriminant Analysis (QDA)* ..... *9*

*2.1.4 Multinomial Logistic Regression (M-LR)* ..... *10*

*2.1.5 eXtreme Gradient Boosting (XGBoost)* ..... *12*

*2.1.6 Cross Validation* ..... *15*

2.2 Ποδόσφαιρο και Στοιχηματισμός ..... 16

*2.2.1 Γενικά* ..... *16*

*2.2.2 Τεχνική Arbitrage* ..... *18*

2.3 Κριτήριο Kelly ..... 19

**3 Δομή της Κατασκευής και Υλοποίηση της** ..... **25**

3.1 Εποπτική Περιγραφή ..... 25

3.2 Στάδιο 1: Ταξινομητής ..... 26

3.3 Στάδιο 2: Κριτήριο Kelly ..... 28

3.4 Σύνοψη ..... 31

**4 Πειραματική Διαδικασία** ..... **34**

4.1 Εισαγωγή ..... 34

4.2 Διαμόρφωση τελικών δεδομένων εισόδου και συνδυασμών παραμέτρων ..... 34

4.3 Επίδοση των μοντέλων στα δεδομένα δοκιμής. ..... 40

*4.3.1 Premier League ..... 42*

*4.3.2 Bundesliga ..... 47*

*4.3.3 La Liga ..... 52*

*4.3.4 Serie A ..... 57*

**5 Συμπεράσματα ..... 62**

5.1 Συμπεράσματα και Παρατηρήσεις..... 62

5.2 Μελλοντικές Προεκτάσεις..... 64

**6 Βιβλιογραφία ..... 65**

# 1 *Εισαγωγή*

## *1.1 Θεμελίωση και Περιγραφή της Εργασίας*

Το αντικείμενο της πρόβλεψης αγώνων ποδοσφαίρου αποτελεί ένα θέμα το οποίο έχει απασχολήσει πολλούς ερευνητές. Αρχικά, προτού υπάρξει η ραγδαία ανάπτυξη της Επιστήμης Υπολογιστών, οι έρευνες παρουσίαζαν αμιγώς στατιστικές μεθόδους που χρησιμοποιούντο για να εξαχθούν τα αποτελέσματα των αγώνων [1], [2]. Στη συνέχεια, ειδικά κατά την τελευταία δεκαετία, η σχετική έρευνα προσανατολίστηκε στη χρήση μεθόδων προβλέψεων που βασίζονται σε μοντέλα Μηχανικής Μάθησης [3], [4], [5] με αρκετά ικανοποιητικά αποτελέσματα που αναφέρονται και αναλύονται στις εργασίες τους.

Κατά κανόνα, τόσο στις εργασίες που αφορούν τη πρόβλεψη ποδοσφαίρου όσο και σε άλλες εργασίες που χρησιμοποιείται η Μηχανική Μάθηση για να προβλεφθούν μελλοντικά αποτελέσματα κάθε είδους (π.χ. αν κάποιος ασθενής πρόκειται να νοσήσει από κάποια ασθένεια ή αν πρόκειται να εμφανιστεί κάποιο καιρικό φαινόμενο σε μια περιοχή του πλανήτη) αξιολογούν τα αποτελέσματα τους κυρίως με βάση την ορθότητα των προβλέψεων. Δηλαδή, πόσες φορές επαληθεύεται η πρόγνωση που δίνει το εργαλείο στο μέλλον. Ένας άλλος τρόπος αξιολόγησης των αποτελεσμάτων θα μπορούσε να προκύψει από την «ικανότητα» ενός μοντέλου όχι μόνο να δώσει κάποια τελική μελλοντική πρόγνωση αλλά να παρέχει κάποια μετρική που να αντιπροσωπεύει πόση βεβαιότητα υπάρχει για το αποτέλεσμα. Αυτό είναι ιδιαιτέρως χρήσιμο διότι πολλές φορές τα δεδομένα εισόδου σε κάποιο μοντέλο ταξινόμησης δεν είναι δυνατό να αποτυπώσουν πλήρως την εικόνα που υπάρχει ώστε να έχουν τη δυνατότητα να καταλήξουν σε συμπέρασμα.

Για να γίνουν πιο σαφή τα προηγούμενα, θα αναφερθεί το τετριμμένο παράδειγμα της πρόβλεψης της ρίψης κάποιου νομίσματος. Τα τελικά ενδεχόμενα όταν φθάσει στο έδαφος

είναι είτε η επάνω επιφάνεια να είναι «Κεφαλή» ή «Γράμματα». Εάν χρησιμοποιηθεί κάποιος ταξινομητής για να προβλεφθεί το αποτέλεσμα αλλά δεν υπάρχουν πληροφορίες όπως η τροχιά, η γωνία ή άλλες σχετικά με τη φύση της ρίψης, είναι αναμενόμενο η ορθότητα των προβλέψεων γύρω στο 50%. Ας γίνει η υπόθεση όμως ότι επιπρόσθετα από την τελική πρόγνωση, είχε ζητηθεί και η πιθανότητα που εκτιμά το μοντέλο να επαληθευθεί αυτή. Έτσι, κάποιος ταξινομητής επιστρέφει κάθε φορά πως αυτή είναι ίση με 50% αλλά μόνο 45% των προβλέψεων που δίνει επαληθεύονται. Οι προβλέψεις όμως ενός 2$^{ου}$ ταξινομητή επαληθεύονται 55 στις 100 φορές αλλά η πιθανότητα που εκτιμά ότι θα επαληθευτούν είναι 95%. Μια πρόχειρη ανάγνωση των αποτελεσμάτων τους είναι ότι ο πρώτος είναι πιο αποτελεσματικός από τον 2$^{ο}$. Όμως, μάλλον τούτο δεν θα ήταν αλήθεια. Γιατί εάν πραγματοποιούνταν 100 ρίψεις για να εξαχθούν τα προηγούμενα αποτελέσματα, είναι αρκετά πιθανό να παρατηρηθεί αυτή η διαφορά στην ορθότητα και να μην οφείλεται σε κάποια διαφορά αποτελεσματικότητας μεταξύ τους. Όμως, ο πρώτος δίνοντας κάθε φορά πιθανότητα για την πρόβλεψη του ίση με 50% κατέδειξε πως δεν υπάρχει έντονη σύνδεση των αποτελεσμάτων και των δεδομένων εισόδου ενώ ο 2$^{ος}$ δεν κατάφερε να εξάγει αυτό το ορθό συμπέρασμα. Έτσι, προκύπτει τελικά πως μάλλον ο πρώτος ταξινομητής είναι αποτελεσματικότερος και θα ήταν γενικά προτιμότερη η χρήση του σε κάποιο πρόβλημα ταξινόμησης όπου θα δίνονταν χρήσιμα δεδομένα για την πρόβλεψη του αποτελέσματος.

Με την ανάπτυξη της χρήσης τεχνικών που βασίζονται σε μοντέλα Μηχανικής Μάθησης στην καθημερινότητα μας, το προηγούμενο ζήτημα για την επίδοση των μοντέλων ταξινόμησης γίνεται εντελώς επίκαιρο και αποκτά κοινωνικές διαστάσεις. Διότι, για παράδειγμα, εάν κάποιο ερευνητικό κέντρο συνέκρινε τη χρήση διαφορετικών ταξινομητών για να βοηθήσει κάποιο νοσοκομείο στην ταξινόμηση απεικονίσεων όγκων ασθενών σε καλοήθεις και κακοήθεις, μπορεί κάποιος ερευνητής να κατέληγε να προτείνει τη χρήση του ταξινομητή πως παρατήρησε ότι δίνει συνολικά την υψηλότερη ορθότητα προβλέψεων κατά την πειραματική διαδικασία αλλά δεν δίνει αντιπροσωπευτικές πιθανότητες (όχι απαραίτητα γιατί δεν είναι «ικανός» ταξινομητής αλλά γιατί το υποκείμενο μαθηματικό μοντέλο που τον περιγράφει δεν είναι πολύ αποτελεσματικό στο να προσφέρει πιθανότητες του δείγματος να ανήκει σε διαφορετικές ομάδες). Όμως, δεδομένου ότι το ιατρικό προσωπικό καλείται να λάβει αποφάσεις για τη χορήγηση κάποιων αγωγών με ενδεχομένως ισχυρές παρενέργειες, θα ήταν ομολογουμένως πιο χρήσιμη για τους ιατρούς να έχουν καλή εικόνα των πιθανοτήτων νόσησης και όχι τόσο των τελικών προγνώσεων.

Τα προηγούμενα διαφορετικά παραδείγματα ήταν απολύτως απαραίτητα για να τεκμηριωθεί η αξία της παρούσας εργασίας διότι ασχολείται και με αυτό το ζήτημα. Συγκεκριμένα, επιχειρεί να αναδείξει ταξινομητές που βασίζονται σε μοντέλα Μηχανικής Μάθησης οι οποίοι δίνουν αντιπροσωπευτικές πιθανότητες πραγμάτωσης των προβλέψεων

τους. Αυτό θα γίνει θεσπίζοντας ένα στιβαρό πλαίσιο αξιολόγησης τους αποδεικνύοντας πως η επίδοση σε μια μετρική που θα επινοηθεί συνδέεται απευθείας με την αποτελεσματικότητα ενός ταξινομητή στην εξαγωγή πιθανοτήτων. Όλο αυτό θα γίνει υπό το πρίσμα της μελέτης ποδοσφαιρικών αναμετρήσεων. Πιο συγκεκριμένα το θέμα επελέγη διότι:

- Το ποδόσφαιρο δίνει ένα πολύ απλό και αξιόπιστο πλαίσιο αξιόλογης των πιθανοτήτων που εξάγονται για τα τελικά αποτελέσματα. Αυτό είναι το παίγνιο του στοιχήματος διότι μπορεί να αποδειχθεί πως η επίδοση σε αυτό, ακολουθώντας τη μεθοδολογία που παρουσιάζεται, είναι απόλυτα συνδεδεμένη με το πόσο αντιπροσωπευτικές είναι η πιθανότητες πραγμάτωσης των τελικών αποτελεσμάτων (βλ. κεφ. 2.3). Αντίθετα, παρόμοιο πλαίσιο που θα αφορούσε το πεδίο διάγνωσης όγκων και εξαγωγής έπειτα βέλτιστων συνδυασμών επικίνδυνων φαρμακευτικών αγωγών που αναφέρθηκε πριν θα ήταν γιγαντιαίο εγχείρημα συγκριτικά.
- Όπως αναφέρθηκε προηγουμένως, υπάρχει μεγάλη έρευνα επάνω στην εξαγωγή τελικών προγνώσεων. Η έρευνα έχει γενικά καταλήξει στα μοντέλα που είναι αποτελεσματικότερα σε αυτό το έργο (και συγκεκριμένα είναι ταξινομητές που βασίζονται σε δένδρα αποφάσεων όπως οι XGBoost και Random Forest [4], [3]) αλλά όχι στο ποιοι δίνουν πιο αντιπροσωπευτικές πιθανότητες για τα τελικά αποτελέσματα του κάθε αγώνα. Αυτό ενδεχομένως να ήταν ουσιαστικά πιο χρήσιμο διότι τα δεδομένα που υπάρχουν πριν από αθλητικούς αγώνες δεν είναι δυνατό να δώσουν πλήρη εικόνα της πραγματικότητας κατά παρόμοιο τρόπο με το παράδειγμα της ρίψης νομίσματος. Για παράδειγμα, δεν υπάρχουν δεδομένα που να αφορούν την ψυχολογία των ομάδων ή τραυματισμούς πριν τον αγώνα. Έτσι, είναι μάλλον αδύνατο να προκύψουν πολύ υψηλά ποσοστά ορθότητας προβλέψεων (και η μέχρι στιγμής έρευνα το έχει επαληθεύσει [3], [6]). Υπό αυτές τις συνθήκες, τα συμπεράσματα λοιπόν αυτής εργασίας δύνανται να αναδείξουν τους καταλληλότερους ταξινομητές από τη σκοπιά μια διαφορετικής μετρικής που ενδεχομένως να είναι και καταλληλότερη για πολλές χρήσεις. Έτσι, η εργασία καλύπτει ένα σημαντικό κενό στη σχετική έρευνα των προβλέψεων.

Το πεδίο της βελτιστοποίησης της οικονομικής επίδοσης στο στοιχηματικό παίγνιο που αφορά ποδοσφαιρικούς αγώνες έχει απασχολήσει στο παρελθόν και άλλους ακαδημαϊκούς ερευνητές [7], [8], [9]. Η παρούσα εργασία όμως διαφοροποιείται σημαντικά και έτσι συμπληρώνει τη σχετική έρευνα. Διότι παρουσιάζει ένα συστηματικό. λεπτομερές και μαθηματικά αξιόπιστο πλαίσιο το οποίο δεν έχει ξαναπαρουσιαστεί στο παρελθόν για τη συγκεκριμένη εφαρμογή. Αυτό αποτελεί το έτερο ζήτημα που ασχολείται η εργασία από το

οποίο προκύπτει και ο τίτλος της και βασίζεται στη χρήση του κριτηρίου Kelly με κάποιες διαφοροποιήσεις (βλ. κεφ. 2.3). Τέλος, γίνεται μεγάλος αριθμός πειραμάτων και έτσι προκύπτει μια σαφέστερη εικόνα για την επίδοση των μεθόδων.

Από την προσπάθεια συνολικά που παρουσιάζεται στην εργασία εξάγονται πολύ ικανοποιητικά αποτελέσματα σε τέτοιο βαθμό που γίνεται εύκολα αντιληπτή η επιτυχία της εργασίας σε ό,τι αφορά τους στόχους που έθεσε. Συγκεκριμένα, οι πειραματικές δοκιμές έδειξαν ξεκάθαρα ποιοι ταξινομητές δίνουν πιο αντιπροσωπευτικές πιθανότητες πραγμάτωσης αποτελεσμάτων ποδοσφαιρικών αναμετρήσεων ενώ η μέθοδος που παρουσιάστηκε οδηγεί πράγματι σε πολύ καλύτερη επίδοση στο παίγνιο σε σχέση με αυτή που παρουσιάζεται σε άλλα συγγράμματα. Όλα τα συμπεράσματα παρουσιάζονται στο κεφάλαιο 5.

## *1.2 Οργάνωση της Εργασίας*

Στο πρώτο κεφάλαιο δίνεται μια εισαγωγή στο πρόβλημα που πραγματεύεται η εργασία και περιγράφεται συνοπτικά η πρόταση που δίνεται για την αντιμετώπιση του.

Στο δεύτερο κεφάλαιο παρουσιάζονται κυρίως το θεωρητικό υπόβαθρο των μοντέλων μηχανικής μάθησης που χρησιμοποιούνται (αλγόριθμοι και παράμετροι) και του κριτηρίου Kelly (μεθοδολογία και τροποποιήσεις). Τέλος, παρουσιάζονται έννοιες που αφορούν τον ποδοσφαιρικό στοιχηματισμό.

Στο τρίτο κεφάλαιο παρουσιάζεται εποπτικά η συνολική δομή της πειραματικής διάταξης. Συγκεκριμένα, γίνεται αναφορά στη ροή των δεδομένων από την εισαγωγή των δεδομένων για την εκπαίδευση των ταξινομητών έως τα αποτελέσματα που προκύπτουν για το βέλτιστο δυνατό ποντάρισμα σε μελλοντικούς αγώνες και περιγράφονται με περισσότερη λεπτομέρεια οι υλοποιήσεις των ταξινομητών και της δομής που βασίζεται στο κριτήριο Kelly.

Στο τέταρτο κεφάλαιο παρουσιάζονται τα πειραματικά αποτελέσματα και ορισμένες πιο ειδικές λεπτομέρειες σχετικά με την υλοποίηση της δομής του κεφαλαίου 3. Τέλος, τα αποτελέσματα αξιολογούνται και συγκρίνονται για να μπορεί να γίνει αποτελεσματική εξαγωγή συμπερασμάτων στη συνέχεια.

Τέλος, στο πέμπτο κεφάλαιο αναφέρονται τα κύρια συμπεράσματα που προέκυψαν από την ενασχόληση με την εργασία και μελλοντικά βήματα που μπορεί να γίνουν επάνω στα ευρήματα της.

# 2 *Θεωρητικό Υπόβαθρο*

## *2.1 Μηχανική Μάθηση (Machine Learning)*

Η Μηχανική Μάθηση αποτελεί σήμερα έναν ανερχόμενο κλάδο του πεδίου της Τεχνητής Νοημοσύνης η οποία με τη σειρά της υπάγεται στην Επιστήμη Υπολογιστών. Αφορά την επινόηση υπολογιστικών μεθόδων που «μαθαίνουν», δηλαδή αξιοποιούν δεδομένα με σκοπό τη βελτίωση της επίδοσης τους σε ένα σύνολο διεργασιών. Τα τελευταία χρόνια αναπτύσσεται συνεχώς και έχει διεισδύσει σε πολλούς τομείς (με κάποιους από αυτούς να μην περιλάμβαναν προηγουμένως την εκτεταμένη χρήση υπολογιστικών συστημάτων) βοηθώντας στο να γίνονται πιο αποτελεσματικά διεργασίες ομαδοποίησης, λήψης αποφάσεων, διάγνωσης κ.α. Αυτοί οι τομείς απαιτούν την αναγνώριση μοτίβων σε δεδομένα με τον τρόπο που ο ανθρώπινος εγκέφαλος δύναται να ενεργήσει, αλλά σε πολλαπλάσιο όγκο δεδομένων που μπορεί μόνο ένας υπολογιστής να διαχειριστεί. Είναι αλήθεια άλλωστε πως κάποια μοντέλα μηχανικής μάθησης έχουν επηρεαστεί από τη δομή του ανθρώπινου εγκεφάλου. Κατά τον Tom M. Mitchell, η μηχανική μάθηση ορίζεται ως εξής: «Ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από μια εμπειρία E ως προς μια κλάση εργασιών T και ένα μέτρο επίδοσης P, αν η επίδοση του σε εργασίες της κλάσης T, όπως αποτιμάται από το μέτρο P, βελτιώνεται με την εμπειρία E». Πιο απλά, μιλάμε για μηχανική μάθηση όταν ένα υπολογιστικό σύστημα χρησιμοποιεί παλιά δεδομένα για να «μάθει» από αυτά ώστε να ταξινομήσει πιο αποτελεσματικά νέα δεδομένα.

Υπάρχουν αρκετοί τύποι μοντέλων Μηχανικής Μάθησης αλλά η παρούσα εργασία θα ασχοληθεί αποκλειστικά με την επιβλεπόμενη μάθηση (supervised learning), όπου ένα μοντέλο εκπαιδεύεται σε πλήθος παραδειγμάτων με το κάθε ένα από αυτά να «δηλώνεται» πως ανήκει σε μια ομάδα με σκοπό στη συνέχεια το μοντέλο να ταξινομεί αχαρακτήριστα νέα παραδείγματα στις προηγούμενες ομάδες σωστά. Η υπόθεση που γίνεται πίσω από αυτή τη προσπάθεια είναι ότι στα δεδομένα μας υπάρχουν μοτίβα που θα συνδέουν κάθε ομάδα με ένα πλήθος παραδειγμάτων και ότι αυτά παραμένουν πολύ παρόμοια ή ίδια κατά την πάροδο του χρόνου μελέτης των παραδειγμάτων. Αξίζει να σημειωθεί ότι υπάρχουν προβλήματα για τα οποία η προηγούμενη υπόθεση ισχύει στο ακέραιο (π.χ. ταξινόμηση όγκων σε καλοήθεις και κακοήθεις από ιατρικές απεικονίσεις) και άλλα που αυτά τα μοτίβα μπορεί να μεταβάλλονται

ραγδαία (πρόβλεψη αποτελεσμάτων αθλητικών αγώνων διότι αρκετές φορές στοιχεία της φύσης του αθλήματος αλλάζουν). Είναι προφανές πως η δυνατότητα επιτυχίας στην επίλυση του προβλήματος μειώνεται εάν ανήκουν στη 2η κατηγορία. Αυτή η πραγματικότητα σε συνδυασμό με τη χρήση δεδομένων που ενδεχομένως να μην δίνουν αρκετά πλήρη εικόνα για τα παραδείγματα που αφορούν (διότι είναι εύλογο να διαπιστώσει κανείς πως η εικόνα ακριβώς πριν έναν αγώνα εξαρτάται από πολύ περισσότερους παράγοντες από προηγούμενα αριθμητικά στατιστικά των ομάδων) έχει οδηγήσει σε σχετικά χαμηλά ποσοστά ορθότητας προβλέψεων στη μέχρι σήμερα δημοσιευμένη έρευνα.

Η παρούσα εργασία εστιάζει σε μεγάλα πρωταθλήματα ποδοσφαίρου της Ευρώπης. Συγκεκριμένα, χρησιμοποιούνται δημοσίως διαθέσιμα δεδομένα που αφορούν την τελευταία 15ετία για τα κύρια πρωταθλήματα της Αγγλίας, της Ισπανίας, της Γερμανίας και της Ιταλίας. Η επίδοση σε αυτά τα πρωταθλήματα δίνει μια καλή εικόνα της αποτελεσματικότητας των μεθόδων που θα χρησιμοποιηθούν καθώς διακρίνονται για τον υψηλό επαγγελματισμό τους αλλά το κάθε ένα από αυτά περιλαμβάνει το ειδικό ταπεραμέντο του εκάστοτε πληθυσμού που αφορά. Τα στατιστικά στοιχεία, όπως θα αναλυθεί και πιο μετά, περιλαμβάνουν παλαιά scores, shoots, corners, βαθμολογίες και αποδόσεις αμέσως πριν αρχίσει ο αγώνας.

Διαφορετικά μοντέλα Μηχανικής Μάθησης διακρίνονται για την επίδοση τους σε διαφορετικές μετρικές. Για παράδειγμα, στο πρόβλημα της πρόβλεψης αποτελεσμάτων ποδοσφαιρικών αγώνων μπορούμε να χρησιμοποιήσουμε μοντέλα τα οποία βασίζονται σε σύνολο από δένδρα αποφάσεων (όπως Random Forest, XGBoost κ.α.). Αυτά, δίνουν συνήθως το μέγιστο δυνατό ποσοστό ορθών προβλέψεων [3] [4] αλλά υστερούν στην ορθότητα της τριπλέτας πιθανοτήτων που δίνουν για τα τρία ενδεχόμενα τέλους του αγώνα. Αντίθετα, μπορούμε να χρησιμοποιήσουμε μοντέλα που βασίζονται σε πιο παραδοσιακές στατιστικές μεθόδους και δίνουν παραδοσιακά πιο ορθές πιθανότητες θυσιάζοντας όμως σε κάποιο βαθμό την ορθότητα των τελικών προβλέψεων (όπως ο GNB και η λογιστική παλινδρόμηση). Στη παρούσα εργασία θα αναδειχθούν οι διαφορετικές επιδόσεις των μοντέλων και προφανώς θα προκύψουν συμπεράσματα για τις καλύτερες πρακτικές.

Στη συνέχεια, παρουσιάζονται και περιγράφονται συνοπτικά το σύνολο των διαφόρων μοντέλων Μηχανικής Μάθησης που χρησιμοποιούνται σε αυτήν την εργασία. Επειδή υπάρχει εκτενής βιβλιογραφία για το μαθηματικό υπόβαθρο των ακολούθων δημόσια διαθέσιμη (εδώ και δεκαετίες για τα περισσότερα), γίνεται προσπάθεια ο αναγνώστης να αποκτήσει ποιοτική και εποπτική κατανόηση των θεμάτων και όχι να μπορεί να τα αναπαραγάγει από την ακόλουθη θεωρία. Είναι προφανές ότι κατά την εκπόνηση της εργασίας δεν χρειάστηκε να δημιουργηθούν τα εργαλεία που αντιστοιχούν σε αυτή τη θεωρία αλλά ήταν απαραίτητο να υπάρχει ισχυρή κατανόηση των δυνατοτήτων τους και των περιορισμών τους που οδήγησε στην τελική επιλογή τους.

### 2.1.1 Gaussian Naive Bayes (GNB)

Είναι ένας αρκετά απλός αλγόριθμος που στηρίζεται στο θεώρημα Bayes και υποθέτει πως η κατανομή που ακολουθούν τα χαρακτηριστικά είναι Gauss. Ο αλγόριθμος υπολογίζει την πιθανότητα ενός ενδεχομένου να υλοποιηθεί με την προϋπόθεση ότι ισχύει μια συνθήκη. Πιο συγκεκριμένα, το θεώρημα αναφέρει ότι η πιθανότητα να συμβεί το ενδεχόμενο A με την υπόθεση ότι ισχύει ένα ενδεχόμενο B, δίνεται από τον παρακάτω τύπο:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Όπου $P(B|A)$ είναι η πιθανότητα να πραγματοποιηθεί το ενδεχόμενο B με την υπόθεση ότι ισχύει το $A$.

Εντελώς ανάλογα, ο αλγόριθμος χρησιμοποιεί το θεώρημα Bayes για να ταξινομήσει ένα διάνυσμα εισόδου. Έστω ότι κάθε στοιχείο εισόδου αποτελείται από n χαρακτηριστικά, δηλαδή έστω:

$$X = (x_1, x_2, \dots, x_n)$$

Τότε η πιθανότητα το στοιχείο να ανήκει κλάση $y$, δίνεται από τη σχέση:

$$P(y|x_1, x_2, \dots, x_n) = \frac{P(x_1|y)P(x_2|y) \dots P(x_n|y)}{P(x_1)P(x_2) \dots P(x_n)}$$

Και δεδομένου ότι τα χαρακτηριστικά της κάθε κλάσης ακολουθούν Gauss, θα ισχύει:

$$P(x_n|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(x_n-\mu_y)^2}{2\sigma_y^2}}$$

Όπου $\mu_y$ είναι η μέση τιμή της κατανομής και $\sigma_y$ η τυπική απόκλιση της. Αυτές υπολογίζονται από τα δεδομένα εκπαίδευση με τέτοιο τρόπο έτσι ώστε κάθε μια κατανομή Gauss που αντιστοιχεί σε feature να περιγράφει όσο το δυνατόν καλύτερα την πραγματική κατανομή των αριθμών.

Με αυτόν τον τρόπο προκύπτουν διαφορετικές πιθανότητες για κάθε διαφορετική κλάση. Προφανώς, η κλάση με τη μεγαλύτερη πιθανότητα αποτελεί και την κλάση που θα κατηγοριοποιήσει ο αλγόριθμος για τη συγκεκριμένη είσοδο. Να αναφερθεί εδώ, πως προϋπόθεση για τη χρήση του Bayes είναι η στατιστική ανεξαρτησία των χαρακτηριστικών μεταξύ τους. Στη συνήθη περίπτωση που αυτό δεν είναι δυνατό χρησιμοποιούνται τεχνικές μείωσης των διαστάσεων των features έτσι ώστε το dataset που προκύπτει από αυτή τη μετατροπή να περιέχει (όσο γίνεται) στατιστικά ανεξάρτητα features.

Ο αλγόριθμος κρίνεται αρκετά εύχρηστος και γρήγορος. Αυτός ο αλγόριθμος συνήθως δεν δίνει τελικές προβλέψεις με πολύ αποτελεσματικό τρόπο (σε σχέση με άλλους αλγόριθμους πιο σύγχρονους) αλλά σε ορισμένα προβλήματα επιστρέφει αρκετά αντιπροσωπευτικές

πιθανότητες για ένα δείγμα να ανήκει σε κάποια κλάση (το οποίο είναι ιδιαίτερα χρήσιμο σε προβλήματα παρόμοιας φύσης με αυτό της παρούσας εργασίας).

### *2.1.2 Linear Discriminant Analysis (LDA)*

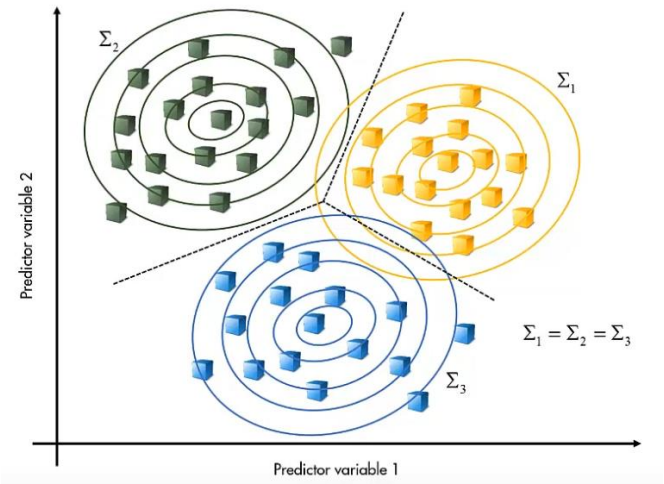
Είναι ένας επίσης απλός αλγόριθμος ο οποίος επιχειρεί να ταξινομήσει δείγματα σε κλάσεις διαχωρίζοντας τον χώρο $R^n$ (που προβάλλονται τα δείγματα) σε περιοχές με γραμμικά σύνορα. Αυτός ο αλγόριθμος χρησιμοποιείται για να αντιμετωπισθούν οι συνέπειες της βασικής υπόθεσης της στατιστικής ανεξαρτησίας των features του GNB διότι υποθέτει ότι τα features (δηλαδή εδώ οι συνιστώσες του $R^n$) δεν είναι στατιστικά ανεξάρτητα για τα δείγματα των κλάσεων. Βέβαια, ο LDA υποθέτει πως η στατιστική εξάρτηση μεταξύ των features είναι η ίδια για όλες τις κατανομές που περιγράφουν τις κλάσεις καθώς και ότι η τυπική απόκλιση για τις επιμέρους κατανομές των κλάσεων σε κάθε συνιστώσα του $R^n$ είναι η ίδια.

Πιο συγκεκριμένα, για κάθε μία από τις κλάσεις δημιουργείται μια κατανομή Gauss (στον χώρο $R^n$) της οποίας η μέση τιμή $\mu$ και ο πίνακας συνδιακύμανσης $\Sigma$ υπολογίζονται με τέτοιο τρόπο που να περιγράφουν με τον καλύτερο τρόπο τη πραγματική κατανομή των δειγμάτων στον χώρο $R^n$. Ο πίνακας $\Sigma$ εκφράζει την τυπική απόκλιση των επιμέρους μονοδιάστατων κατανομών Gauss και τις συνδιακυμάνσεις ανά ζεύγη μεταξύ τους. Βέβαια, συνέπεια της υπόθεσης του LDA που αναφέρθηκε στη προηγούμενη παράγραφο είναι ότι ο πίνακας $\Sigma$ είναι ίδιος για κάθε μία από τις κατανομές για κάθε κλάση. Αυτό απεικονιστικά στην πράξη «τοποθετεί» κατανομές Gauss ίδιου «μεγέθους και σχήματος» στο $R^n$ (όπως φαίνεται στην εικόνα 1).

Έστω ένα νέο δείγμα της μορφής (μετά τη διαδικασία εκπαίδευσης):

$$X = (x_1, x_2, \dots, x_n)$$

και έχουμε m κλάσεις $y_1, y_2, \dots, y_m$. Το ταξινομούμε το $X$ στην κλάση που είναι πιθανότερο να ανήκει υπολογίζοντας τα $P(y_1|X), P(y_2|X), \dots, P(y_n|X)$ σύμφωνα με τον κανόνα Bayes. Εάν αποφασίσουμε να χωρίσουμε σε περιοχές τον $R^n$ που αντιστοιχούν σε κάθε κλάση κατά αυτό τον τρόπο, προκύπτει πως αυτές θα διαχωρίζονται από γραμμικά σύνορα (όπως φαίνεται στην εικόνα 1 με τις διακεκομμένες γραμμές) ως άμεση συνέπεια της υπόθεσης του LDA για κατανομές με ίδιους πίνακες συνδιακύμανσης.




**Εικόνα 1 Δείγματα 3 κλάσεων σε δισδιάστατο χώρο που διαχωρίζονται αποτελεσματικά με γραμμικά σύνορα που προκύπτουν από LDA. Οι κυκλικές γραμμές απεικονίζουν τη κατανομή Gauss που προσαρμόστηκε για να περιγράφει τα δείγματα της κάθε κλάσης. Προσοχή ότι ως αποτέλεσμα της υπόθεσης του LDA, έχουν το ίδιο σχήμα. Πηγή: [10]**

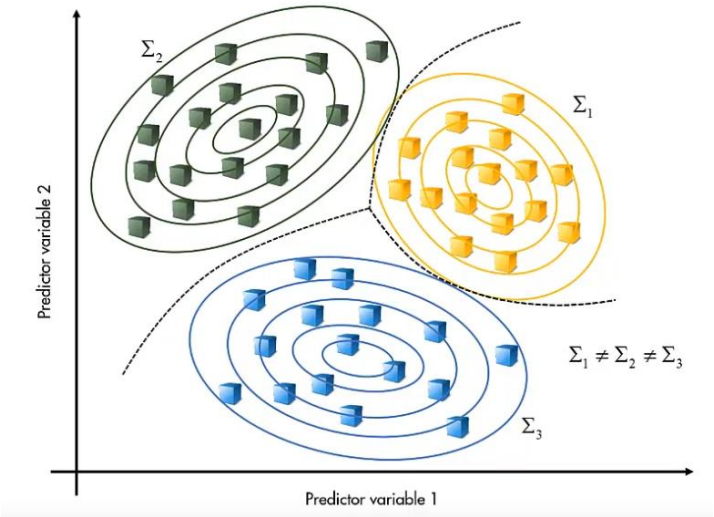
Τέλος, αξίζει να σημειωθεί το ακόλουθο. Οι αλγόριθμοι GNB και QDA είναι πολύ παρόμοιοι διότι βασίζονται στον κανόνα Bayes και κάνουν χρήση της κατανομής Gauss. Ο GNB μπορεί να περιγραφθεί με τον ίδιο τρόπο που έγινε η περιγραφή του LDA προηγουμένως με τη διαφορά ότι θα ήταν μηδενικές οι συνδιακυμάνσεις μεταξύ των επιμέρους κατανομών των κλάσεων στις συνιστώσες του $R^n$ αλλά η τυπική απόκλιση (η κύρια διαγώνιος του πίνακα συνδιακύμανσης) δεν χρειάζεται να παραμένει η ίδια ανάμεσα σε καμία από αυτές.

### *2.1.3 Quadratic Discriminant Analysis (QDA)*

Ο αλγόριθμος QDA είναι παρόμοιος με τον LDA. Η μοναδική διαφορά τους είναι πως δεν γίνεται καμία υπόθεση στατιστικής εξάρτησης μεταξύ των features των κλάσεων σε αντίθεση με τους LDA και GNB (βλ. προηγούμενες σελίδες). Ως εκ τούτου, προκύπτουν διαφορετικοί πίνακες συνδιακύμανσης $\Sigma$ για κάθε κατανομή που αντιστοιχεί σε κλάση. Έτσι, ο QDA τείνει να εφαρμόζεται πιο αποτελεσματικά σε δεδομένα κατά την εκπαίδευση και να επιτυγχάνει ορθότερη ταξινόμηση νέων δειγμάτων (εάν τα δεδομένα εκπαίδευσης είναι αρκετά για να δίνουν αντιπροσωπευτική εικόνα της πραγματικής κατανομής). Όμως, απαιτεί περισσότερους υπολογισμούς για τις περισσότερες παραμέτρους που διαθέτει ενώ οι αλγόριθμοι GNB και LDA πρακτικά αποτελούν υποπεριπτώσεις του με επιπλέον περιορισμούς.

Τέλος, αξίζει να σημειωθεί ότι απεικονιστικά τα σύνορα των περιοχών που ανήκουν σε κάθε κλάση (όπως αυτά ορίστηκαν κατά την περιγραφή του LDA) δεν θα είναι απαραίτητα γραμμικά αλλά θα μοιάζουν με τις διακεκομμένες γραμμές στο γράφημα της εικόνας 2. Υπάρχει επίσης περίπτωση να δημιουργούνται περιοχές που ανήκουν στην ίδια κλάση και να

μην είναι ενωμένες στον χώρο πράγμα το οποίο προκύπτει πως είναι αδύνατο με τον αλγόριθμο LDA.




**Εικόνα 2 Δείγματα 3 κλάσεων σε δισδιάστατο χώρο που διαχωρίζονται αποτελεσματικά με γραμμικά σύνορα που προκύπτουν από QDA. Τα ελλειψοειδή σχήματα απεικονίζουν τη κατανομή Gauss που προσαρμόστηκε για να περιγράφει τα δείγματα της κάθε κλάσης. Προσοχή στα μη (απαραίτητα) γραμμικά σύνορα. Πηγή: [10]**

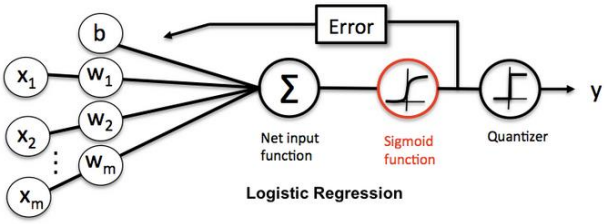
### *2.1.4 Multinomial Logistic Regression (M-LR)*

Είναι γνωστός στη βιβλιογραφία ο αλγόριθμος Logistic Regression (λογιστική παλινδρόμηση). Αυτός χρησιμοποιείται για προβλήματα ταξινόμησης στα οποία τα δείγματα ανήκουν σε 2 κλάσεις αποκλειστικά. Ο αλγόριθμος αυτός λειτουργεί ως ακολούθως.

Έστω $k$ δείγματα στον χώρο $R^n$ της μορφής $x_i = (x_{i1}, x_{i2}, \dots, x_{in})$, $1 \leq i \leq k$ που αποτελούν το σύνολο εκπαίδευσης με κάθε ένα από αυτά να ανήκει σε μια εκ των δύο κλάσεων $y_1$,$y_2$. Ορίζεται η εξής συνάρτηση:

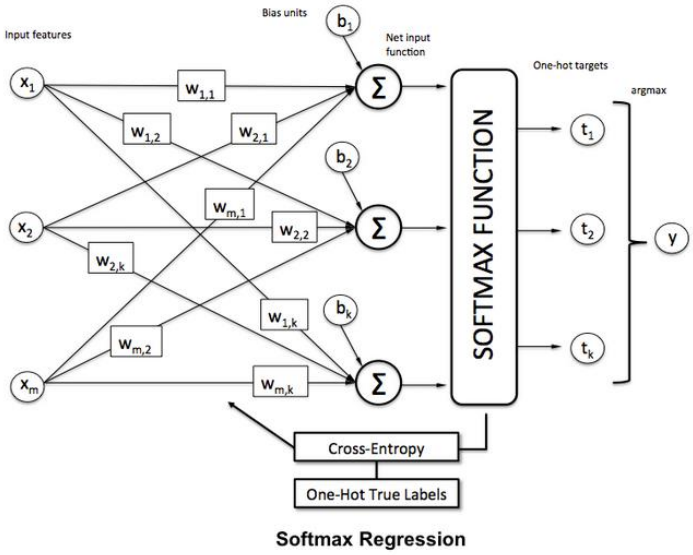
$$y = \frac{1}{1 + e^{x_i * w + b}}$$

Με $w = (w_1, w_2, \dots, w_n)$ και $b$ τα βάρη που προσδιορίζονται κατά τη διάρκεια εκπαίδευσης. Συχνά για αυτό τον σκοπό χρησιμοποιείται η μέθοδος της μέγιστης πιθανοφάνειας.




**Εικόνα 3 Σχηματική αναπαράσταση ταξινομητή Λογιστικής Παλινδρόμησης. Πηγή: [11]**

Εάν εμείς θέλουμε να χρησιμοποιήσουμε αυτήν την τεχνική για προβλήματα στα οποία τα δείγματα μπορεί να ανήκουν σε 3 ή περισσότερες κλάσεις, δημιουργούμε πολλαπλούς δυαδικούς ταξινομητές των οποίων οι έξοδοι συνδυάζονται σε μία έξοδο (με τη χρήση της συνάρτησης Softmax η οποία θα αναφερθεί στη συνέχεια) όπως φαίνεται και στο παρακάτω σχήμα.




**Εικόνα 4 Σχηματική αναπαράσταση ταξινομητή Softmax Regression. Πηγή: [11]**

Η δομή που προκύπτει ονομάζει Softmax Regression ή Multinomial Logistic Regression. Ο προσδιορισμός των βαρών μπορεί να γίνει (μεταξύ άλλων τρόπων) με τη χρήση κάποιας μετρικής σφάλματος στην έξοδο της συνάρτησης Softmax.

Τέλος, ορίζεται ως συνάρτηση Softmax η ακόλουθη συνάρτηση για κάποια:

$$y = (y_1, y_2, \dots, y_n), \;\; y_i = \frac{e^{-x_i}}{\sum_{j=1}^{n} e^{-x_j}}$$

Όπου $x = (x_1, x_2, \dots, x_n)$ το διάνυσμα εισόδου και $y = (y_1, y_2, \dots, y_n)$ το διάνυσμα εξόδου. Έχει τις εξής ιδιότητες:

- $y_1, y_2, \dots, y_n \in [0,1]$
- $\sum_{i=1}^{n} y_i = 1$

Οι δύο παραπάνω ιδιότητες δίνουν την παραπλανητική εντύπωση πως το διάνυσμα εξόδου της αντιστοιχεί σε πιθανότητες κάποιο δείγμα να αντιστοιχεί σε κάποια κλάση αλλά αυτό εξαρτάται από τον αλγόριθμο που τροφοδοτεί την Softmax η οποία απλώς κάνει έναν μετασχηματισμό. Όντως, υπάρχουν αλγόριθμοι των οποίων ο τρόπος που γίνονται οι υπολογισμοί εντός τους, τους καθιστά κατάλληλους για περιπτώσεις όπου απαιτείται να λαμβάνεται η πιθανότητα του δείγματος να ανήκει σε κάποια κλάση αλλά υπάρχουν και άλλοι οι οποίοι δεν είναι κατάλληλοι. Ο αλγόριθμος Multinomial Logistic Regression μπορεί να χρησιμοποιηθεί για τον προηγούμενο σκοπό εάν η κατανομή των δεδομένων είναι τέτοια που να εφαρμόζεται αποτελεσματικά η σιγμοειδής συνάρτηση σε αυτά.

### *2.1.5 eXtreme Gradient Boosting (XGBoost)*

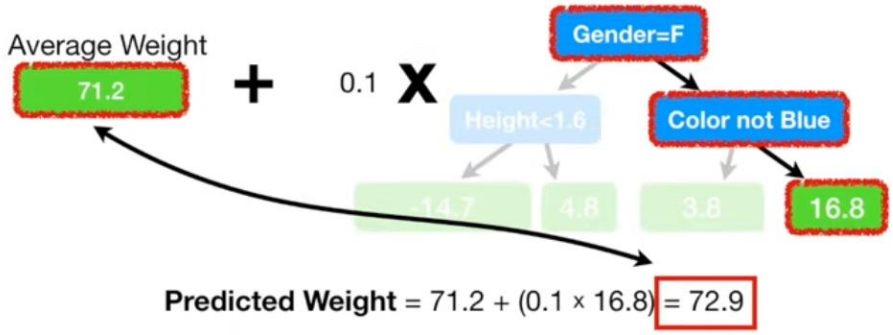
Αποτελεί έναν σύγχρονο αλγόριθμο ο οποίος βασίζεται στη δημιουργία ενός συνόλου από δένδρα αποφάσεων και στην ταξινόμηση δειγμάτων με τον τρόπο που θα αναλυθεί στη συνέχεια. Παρόλο που στηρίζεται σε πολλαπλά δένδρα αποφάσεων, διαφέρει πολύ στη φιλοσοφία του σε σχέση με τον Random Forest όπως θα γίνει αντιληπτό στη συνέχεια.

Όταν χρησιμοποιείται ο XGBoost για classification είναι εγγενώς κατάλληλος για προβλήματα ταξινόμησης των οποίων τα δείγματα μπορεί να ανήκουν σε μια από ακριβώς 2 κλάσεις. Στην περίπτωση όπου χρησιμοποιείται ο XGBoost για προβλήματα ταξινόμησης multiclass (όπου τα δείγματα μπορεί να ανήκουν σε 3 ή και παραπάνω κλάσεις) δημιουργείται ένα σχήμα όμοιο με αυτό του Softmax Regression (βλ. περιγραφή του Softmax Regression προηγουμένως). Δηλαδή δημιουργούνται πολλαπλά μοντέλα XGBoost που εκπαιδεύονται σε δυαδικά υποπροβλήματα και οι έξοδος συνδυάζεται με τη χρήση της συνάρτησης Softmax.

Δεδομένων των παραπάνω, για να γίνει αντιληπτός ο τρόπους λειτουργίας του XGBoost αρκεί να αναλυθούν δύο παραδείγματα χρήσης του αλγόριθμου. Στο ένα θα γίνεται (δυαδική) ταξινόμηση και στο άλλο παλινδρόμηση.

#### *2.1.5.1 XGboost Regression*

Έστω $k$ δείγματα στον χώρο $R^n$ της μορφής $x_i = (x_{i1}, x_{i2}, \dots, x_{in}),\ 1 \leq i \leq k$ που αποτελούν το σύνολο εκπαίδευσης με κάθε ένα από αυτά να αντιστοιχεί σε μια τιμή στο $R$. Κατά τη διάρκεια της εκπαίδευσης, ο αλγόριθμος ουσιαστικά δημιουργεί διαδοχικά δένδρα αποφάσεων. Για κάθε νέο δείγμα προς ταξινόμηση η έξοδος του αλγόριθμου είναι το αριθμητικό άθροισμα που προκύπτει από τη τιμή του φύλλου που προκύπτει, διασχίζοντας κάθε δένδρο απόφασης όπως στην παρακάτω εικόνα.




**Εικόνα 5 Πρόβλεψη τιμής βάρους σε δείγματα από το δάσος δένδρων που δημιουργήθηκε στην εκπαίδευση. Να σημειωθεί ότι το 0.1 αποτελεί το learning rate. Η τρόπος πρόβλεψης διαφέρει ουσιαστικά από τον αλγόριθμο Random Forest. Πηγή: [12]**

Τα δένδρα αποφάσεων δημιουργούνται ως εξής:

1) Δημιουργείται ένα φύλλο-δένδρο με μια τιμή που προστίθεται έπειτα σε όλες τις προβλέψεις (βλ. "average weight" στην εικόνα 8). Η αριθμητική τιμή δεν έχει ουσιαστική σημασία διότι δημιουργούνται επόμενα δένδρα που λαμβάνουν υπόψιν το σφάλμα.
2) Έπειτα, για τη δομή που έχει προκύψει (μπορεί να είναι απλά ένα φύλλο ή αργότερα δάσος από δένδρα) υπολογίζουμε το "similarity score" όπως ορίζεται από την ακόλουθη παράσταση:

$$similarity = \frac{(\sum(actualvalue - predictedvalue))^2}{numberofincorrectpredictions + \lambda}$$

Αυτό το σκορ δεν αναδεικνύει ακριβώς πόσο εσφαλμένες ήταν οι προβλέψεις αλλά δίνει μια εικόνα για το πόσο «εσφαλμένος» είναι ο «μέσος όρος» αυτών. Αυτό γιατί εάν υπάρχουν σφάλματα αντίθετα και ίσα μεταξύ τους θα αλληλοακυρωθούν στην παράσταση. Ο συντελεστής $\lambda$ αποτελεί έναν «παράγοντα εξομάλυνσης» διότι οδηγεί τον αλγόριθμο να περιορίζει τυχόντα ψηλά scores σε περιπτώσεις όπου υπάρχουν ελάχιστες εσφαλμένες προβλέψεις. Αυτό μπορεί να αντιμετωπίσει την υπερεκπαίδευση και είναι πιο χρήσιμο στο βήμα 3. Εάν η δομή περιλαμβάνει επιπλέον δένδρα και όχι μόνο το αρχικό δένδρο-φύλλο τότε ορίζεται η παράμετρος "learning rate" που καθορίζει το βάρος του αποτελέσματος του κάθε δένδρου στο άθροισμα όπως φαίνεται στην εικόνα 8.
3) Στη συνέχεια, επιχειρούμε να κάνουμε έναν περαιτέρω διαχωρισμό δεδομένων χρησιμοποιώντας κάποιο κατώφλι. Έτσι προκύπτουν δύο νέα υποσύνολα

δειγμάτων. Δημιουργούμε ένα ακόμη δένδρο γι' αυτόν σκοπό. Αναζητούμε τον βέλτιστο διαχωρισμό για να δημιουργήσουμε τη διακλάδωση του δένδρου. Γι' αυτό το σκοπό, ορίζουμε το score "gain" ως ακολούθως:

$$gain = similarity_{leftnode} + similarity_{rightnode} - similarity_{node}$$

Είναι ευκολονόητο πως υψηλότερες τιμές του gain συνεπάγονται και αποτελεσματικότερο διαχωρισμό. Αξίζει να σημειωθεί επίσης πως εδώ φαίνεται η αξία της χρήσης μη τετραγωνισμένων διαφορών στο similarity. Με τον τρόπο που συμβαίνει αναδεικνύεται πιο αποτελεσματικά χρησιμοποιώντας το gain όταν υπάρχει αποτελεσματικότερος διαχωρισμός. Στο τέλος επιλέγεται ο διαχωρισμός με το υψηλότερο gain. Επαναλαμβάνουμε τη διαδικασία του βήματος 3 μέχρι να φθάσουμε στο μέγιστο επιτρεπόμενο βάθος.

4) Κλαδεύουμε το δένδρο του βήματος 3 ως εξής. Ορίζουμε τη παράμετρο $\gamma$. Για κάθε φύλλο του δένδρου εάν η μετρική similarity είναι μικρότερη από την τιμή που ορίσαμε για τη $\gamma$ το αφαιρούμε μαζί με το άλλο φύλλο που προκύπτει από τον ίδιο «γονέα» στο δένδρο. Ο «γονέας» πλέον γίνεται φύλλο. Επαναλαμβάνουμε τη διαδικασία μέχρι η $\gamma$ να είναι μικρότερη από κάθε similarity score.
5) Επαναλαμβάνουμε από το βήμα 2 μέχρι να τελειώσει είτε ο μέγιστος χρόνος εκπαίδευσης που ορίσαμε ή το σφάλμα μειωθεί κάτω από ένα προκαθορισμένο όριο ή να εξαντλήσουμε το πλήθος των διαθέσιμων δένδρων-εκτιμητών

### *2.1.5.2 XGBoost Classification*

Έστω $k$ δείγματα στον χώρο $R^n$ της μορφής $x_i = (x_{i1}, x_{i2}, \dots, x_{in}),\ 1 \le i \le k$ που αποτελούν το σύνολο εκπαίδευσης με κάθε ένα από αυτά να ανήκει ή όχι σε μια κλάση. Έξοδος ίση με 1 δηλώνει ότι το δείγμα αντιστοιχεί στην κλάση και 0 ότι δεν αντιστοιχεί. Κατά τη διάρκεια της εκπαίδευσης στο προηγούμενο σύνολο δειγμάτων, ο αλγόριθμος ουσιαστικά δημιουργεί διαδοχικά δένδρα αποφάσεων όμοια με την ταξινόμηση προηγουμένως. Για κάθε νέο δείγμα προς ταξινόμηση η έξοδος του αλγόριθμου είναι το αριθμητικό άθροισμα που προκύπτει από τη τιμή του φύλλου που προκύπτει διασχίζοντας κάθε δένδρο απόφασης όμοια με την εικόνα 8.

Η μοναδική ουσιαστική διαφορά μεταξύ της παλινδρόμησης και της ταξινόμησης είναι η διαφοροποίηση του προαναφερθέν similarity score. Αυτό πλέον ορίζεται ως:

$$similarity = \frac{\left(\sum(actualvalue - predictedvalue)\right)^2}{\sum(previousprobability(1 - previousprobability)) + \lambda}$$

Με το $previousprobability$ να είναι η τιμή μεταξύ 0 και 1 που επιστρέφει ο αλγόριθμος για κάθε δείγμα εισόδου. Με αυτό τον τρόπο ουσιαστικά «ενθαρρύνεται» η επιστροφή αποτελεσμάτων πιο κοντά στις ακραίες τιμές 0 και 1.

Έχοντας αναφέρει όλα τα παραπάνω για τον XGBoost αναφέρονται συνοπτικά παρακάτω οι παράμετροι εκπαίδευσης που θα ασχοληθούμε στη συνέχεια.

- gamma ($\gamma$). Οδηγεί στη δημιουργία «κοντύτερων» (και έμμεσα περισσοτέρων) δένδρων. Προλαμβάνει την υπερεκπαίδευση.
- Αριθμός εκτιμητών. Καθορίζει τον μέγιστο αριθμό των δένδρων που θα δημιουργηθούν.
- lambda ($\lambda$). Οδηγεί σε εντονότερη αγνόηση διαχωρισμών σε σύνολα αποτελούμενα από ελάχιστα δείγματα. Προλαμβάνει την υπερεκπαίδευση.
- Μέγιστο βάθος. Καθορίζει το μέγιστο βάθος του κάθε δένδρου.
- Learning rate. Περιορίζει το βάρος απόφασης του κάθε δένδρου. Προλαμβάνει την υπερεκπαίδευση.
- Subsample. Σε κάθε δένδρο χρησιμοποιείται ένα υποσύνολο το δειγμάτων εκπαίδευσης. Προλαμβάνει την υπερεκπαίδευση.
- colsample_bytree. Σε κάθε δένδρο χρησιμοποιείται ένα υποσύνολο των features το δειγμάτων εκπαίδευσης. Προλαμβάνει την υπερεκπαίδευση.

### *2.1.6 Cross Validation*

Κλείνοντας τη θεωρία για το Machine Learning θα γίνει αναφορά σε μια τεχνική βελτιστοποίησης παραμέτρων που χρησιμοποιείται ευρέως και σε αυτήν την εργασία. Η εκπαίδευση αποτελεί τη διαδικασία κατά την οποία ένα μοντέλο αυτό-διαμορφώνει κάποιες εσωτερικές παραμέτρους που το οδηγούν φαινομενικά να «εξοικειωθεί» με κάποια παλαιά δεδομένα, αναγνωρίζοντας πρότυπα δίνοντας του (μοντέλου) τη δυνατότητα ταξινόμησης νέων αταξινόμητων δεδομένων. Η εκπαίδευση όμως έχει και κάποιες (περιοριστικές) παραμέτρους και ο προσδιορισμός αυτών με συστηματικό τρόπο μπορεί να οδηγήσει σε διαφορετικές εσωτερικές παραμέτρους του μοντέλου που οδηγούν ενίοτε σε καλύτερες προβλέψεις.

Το ζήτημα είναι όμως ακριβώς ο συστηματικός τρόπος προσδιορισμού των παραμέτρων της εκπαίδευσης. Δηλαδή, εφόσον δεν χρησιμοποιούνται τα δεδομένα δοκιμής κατά την εκπαιδευτική διαδικασία, πως θα μπορούσαν να αξιολογηθούν δυο διαφορετικά μοντέλα όμοιου τύπου που έχουν εκπαιδευτεί σύμφωνα με διαφορετικές παραμέτρους εκπαίδευσης? Ένα σύνηθες σχήμα είναι η χρήση ενός τρίτου συνόλου εκτός αυτού των δειγμάτων εκπαίδευσης και δοκιμής, το οποίο είναι διαθέσιμο κατά την διαδικασία

εκπαίδευσης και επάνω σε αυτό αξιολογείται η επίδοση των όμοιων τύπου μοντέλων. Αυτό το τρίτο σύνολο ονομάζεται σύνολο επαλήθευσης (validation set). Υπάρχει όμως το ενδεχόμενο αυτό το μεμονωμένο σύνολο (το οποίο είναι συνήθως μικρό σε σχέση με τα άλλα δύο) να μην δίνει αντιπροσωπευτική εικόνα της συνολικής επίδοσης του μοντέλου. Η τακτική που ακολουθείται για να λυθεί αυτό το πρόβλημα είναι ο διαμελισμός του συνόλου εκπαίδευσης σε $N$ ίσα μέρη εκ των οποίων τα $N-1$ αποτελούν το σύνολο εκπαίδευσης και αυτό που περισσεύει το σύνολο επαλήθευσης. Προκύπτουν $N$ τελικοί συνδυασμοί που είναι δυνατό να γίνει αυτό και χρησιμοποιούνται όλοι εξ αυτών για να αξιολογηθεί το μοντέλο. Συγκεκριμένα λαμβάνουμε τον μέσο όρο των επιδόσεων στους συνδυασμούς καθώς και τη τυπική απόκλιση. Ο συνδυασμός των παραμέτρων με τα καλύτερα scores στις προηγούμενες δύο μετρικές είναι και αυτός που τελικά επιλέγεται για την εκπαίδευση του τελικού μοντέλου επάνω σε όλο το σύνολο εκπαίδευσης. Αυτή η τεχνική ονομάζεται cross-validation.

## *2.2 Ποδόσφαιρο και Στοιχηματισμός*

### *2.2.1 Γενικά*

Το ποδόσφαιρο αποτελεί το λαοφιλέστερο άθλημα στον κόσμο με πληθώρα επαγγελματικών πρωταθλημάτων. Δεν θα γίνει αναφορά στον τρόπο παιξίματος πέρα από τα βασικά που χρειάζονται για τα πλαίσια αυτής της εργασίας, καθώς ο αναγνώστης μάλλον ήδη γνωρίζει με λεπτομέρεια το θέμα αλλά ακόμα και αν όχι, μπορεί πολύ εύκολα να βρει σχετικές πληροφορίες. Αξίζει να αναφερθεί πως το ποδόσφαιρο είναι υπεύθυνο για ένα κομμάτι της παγκόσμιας οικονομίας που αντιστοιχεί σε πολλά δισεκατομμύρια ευρώ. Αυτό έχει ως συνέπεια να υπάρχει μεγάλο ερευνητικό ενδιαφέρον γύρω από την ερμηνεία και πρόβλεψη αποτελεσμάτων.

Αρχικά, ένας αγώνας ποδοσφαίρου αποτελείται από 2 ομάδες των 11 παιχτών που αγωνίζονται για 90 λεπτά και σκοπό έχουν να τοποθετήσουν όσο το δυνατόν περισσότερες φορές την μπάλα στην αντίπαλη εστία και να πετύχουν goal(s). Η ομάδα με τα περισσότερα goals στο τέλος του 90λεπτου ανακηρύσσεται νικήτρια, ενώ σε περίπτωση ισοπαλίας, ανάλογα με τη φύση της διοργάνωσης, ενδέχεται να υπάρξει και επιπλέον χρόνος παιχνιδιού (παράταση). Τον αγώνα επιβλέπει ένας διαιτητής, με την βοήθεια 3 ή 5 ακόμα ατόμων, ο οποίος είναι υπεύθυνος για την τήρηση των κανονισμών. Ενας αγώνας έχει 3 δυνατές εκβάσεις. Αυτές είναι η νίκη της γηπεδούχου, η ισοπαλία και η νίκη της φιλοξενούμενης ομάδας. Στη γλώσσα του στοιχήματος η νίκη γηπεδούχου λέγεται άσσος («1»), η ισοπαλία «Χ» και η νίκη φιλοξενούμενου διπλό «2». Το συνηθέστερο αποτέλεσμα σε έναν αγώνα είναι η νίκη γηπεδούχου με ποσοστό κοντά στο 50%, η νίκη φιλοξενούμενου μετά και η ισοπαλία τελευταία.

Αγώνες ποδοσφαίρου υπάρχουν καθ' όλη τη διάρκεια της χρονιάς, είτε σε συλλογικό είτε σε επίπεδο εθνικών ομάδων. Σε συλλογικό επίπεδο υπάρχει το εθνικό πρωτάθλημα, όπου μια ομάδα αγωνίζεται με όλες τις ομάδες της ίδιας κατηγορίας εντός και εκτός έδρας, ενώ αρκετές ομάδες έχουν και ευρωπαϊκές υποχρεώσεις αγωνιζόμενες με τις υπόλοιπες καλύτερες ομάδες της Ευρώπης. Σε επίπεδο εθνικών ομάδων, υπάρχουν οι μεγάλες διοργανώσεις που λαμβάνουν μέρος το καλοκαίρι και κάποιοι φιλικοί ή προκριματικοί αγώνες την υπόλοιπη χρονιά, οπότε και διεξάγονται οι αγώνες πρωταθλήματος.

Ως στοίχημα, κανείς αναφέρεται στην επιλογή να τοποθετήσει κάποιος ένα ποσό στην πρόβλεψη της εξέλιξης κάποιων αγώνων, που προσφέρονται από εταιρείες στοιχηματισμού έναντι κάποιας απόδοσης. Το στοίχημα είναι διαδεδομένο σε όλο τον κόσμο, χωρίς εξαιρέσεις, με πολύ κόσμο να επιλέγει και άλλα αθλήματα, πέρα του ποδοσφαίρου για να στοιχηματίσει. Ο στοιχηματισμός και γενικότερα η προσπάθεια επίτευξης κέρδους είναι χαρακτηριστικά της ανθρώπινης φύσης.

Η απόδοση στο παίγνιο του στοιχήματος αποτελεί έναν αριθμό που καθορίζει το μέγεθος της επιστροφής χρημάτων σε περίπτωση στοιχηματικής νίκης. Ορίζεται με αρκετούς διαφορετικούς τρόπους. Στην Ευρώπη έχει επικρατήσει κυρίως ότι εκφράζει τον αριθμό που πολλαπλασιάζεται το ποντάρισμα και προκύπτουν τα χρήματα που επιστρέφονται σε περίπτωση στοιχηματικής νίκης. Έτσι απόδοση ίση με 2.25 σημαίνει πως θα επιστραφούν στον παίκτη 112.5 ευρώ σε περίπτωση νίκης ενώ απόδοση μικρότερη του 1 δεν έχει νόημα να υπάρχει με αυτόν τον ορισμό. Σε αυτήν την εργασία θα ασχοληθούμε με τον πλέον παραδοσιακό τρόπο στοιχηματισμού ποδοσφαίρου όπου ένας αγώνας σε ένα πρωτάθλημα μπορεί να τελειώσει αποκλειστικά με νίκη της γηπεδούχου, ήττα ή ισοπαλία (1-2-X). Έτσι, μια εταιρεία στοιχημάτων θα καθορίζει μια τριπλέτα αριθμών που θα αντιστοιχούν στις αποδόσεις που προσφέρει για τα στοιχήματα. Μια τέτοια εταιρεία στοχεύει οι αποδόσεις να αντιστοιχούν στις πραγματικές πιθανότητες εκπλήρωσης των ενδεχομένων, αλλά για διάφορους συγκυριακούς και μη λόγους ενδέχεται να τη συμφέρει να αποκλίνουν σε ένα βαθμό [13]. Μεμονωμένα και για μια μόνο ευκαιρία τοποθέτησης στοιχήματος (περισσότερα σχετικά με αυτό στην υποενότητα «κριτήριο Kelly») ένα στοίχημα συμφέρει να τοποθετηθεί εάν:

$$\text{απόδοση} * \text{πιθανότητα εκπλήρωσης} > 1$$

Η 'απόδοση' ορίζεται με τον προαναφερθέντα τρόπο για να είναι σωστή η προηγούμενη σχέση. Έτσι μπορεί κάποιος να θεωρήσει ότι η απόδοση υποδηλώνει μια εκτιμώμενη πιθανότητα εκπλήρωσης από τη στοιχηματική εταιρεία. Αυτό δεν ισχύει απόλυτα στη πραγματικότητα [13] οπότε δεν θα πρέπει να ταυτίζονται. Η μόνη χρήσιμη ερμηνεία που μπορεί να δοθεί είναι ότι εάν η πιθανότητα εκπλήρωσης ενός ενδεχομένου είναι υψηλότερη από το αντίστροφο της απόδοσης, τότε «μεμονωμένα» ένα στοίχημα «αξίζει» να τοποθετηθεί. Αυτές οι έννοιες θα αναλυθούν σε βάθος στη συνέχεια.

Κλείνοντας, θα γίνει αναφορά στη γκανιότα. Αυτή αποτελεί έναν εκ των πολλών τρόπων που προσπαθεί να εξασφαλίσει κέρδος μια στοιχηματική εταιρεία. Στην ουσία εκφράζει το μέγεθος της μείωσης των προσφερόμενων αποδόσεων (της τριπλέτας) συνολικά και όχι με τρόπο που η μείωση της μιας να οδηγεί στην αύξηση της άλλης. Με τη χρήση της γκανιότας μια στοιχηματική εταιρεία εισάγει έμμεσα μια χρέωση σε κάθε στοίχημα που τοποθετείται η οποία της δίνει ένα μακροπρόθεσμο κέρδος ανεξάρτητα από ένα «προβάδισμα» γνώσης που ενδεχομένως έχει ή άλλα πιο σύνθετα στατιστικά φαινόμενα τα οποία τη καθιστούν επίσης κερδοφόρα σε σχέση με τους παίκτες. Η γκανιότα μετριέται σε ποσοστιαίες μονάδες και υπολογίζεται ως εξής:

$$Γκανιότα = (\sum \frac{1}{αποδόσεις} - 1) * 100\%$$

Υψηλότερη γκανιότα σημαίνει πως απαιτούνται ορθότερες προβλέψεις για να γίνει ένας παίκτης μακροπρόθεσμα κερδοφόρος γιατί αυξάνεται η «χρέωση» ανά τοποθέτηση στοιχήματος. Η γκανιότα ανά τριπλέτα (σε στοιχήματα ίδιας εταιρείας) είναι προφανώς μη αρνητικός αριθμός αλλά είναι δυνατόν να συνδυαστούν αποδόσεις από διαφορετικές εταιρείες για να δώσουν συνολικά αρνητική γκανιότα. Αυτή είναι δυνατό να εξασφαλίσει στον παίκτη μακροπρόθεσμο κέρδος με απλές υπολογιστικές μεθόδους χωρίς τη δημιουργία πρόβλεψης (τεχνική arbitrage). Η εταιρείες μεταξύ τους προσπαθούν να περιορίζουν τέτοιους συνδυασμούς-ευκαιρίες και πέρα από σπάνιες πλέον εξαιρέσεις το επιτυγχάνουν.

### *2.2.2 Τεχνική Arbitrage*

Στη σύγχρονη εποχή, ανά πάσα στιγμή δίνεται η δυνατότητα στοιχηματισμού σε πολλές εταιρείες. Έτσι, ανά πάσα στιγμή ένας παίκτης έχει πάρα πολλές διαθέσιμες επιλογές για να τοποθετήσει στοιχήματα και αναπόφευκτα προκύπτουν διαφορές τις οποίες μπορεί να εκμεταλλευτεί έτσι ώστε να έχει όφελος ακόμα και χωρίς τη δύναμη των προγνωστικών. Αυτό μπορεί να συμβεί εάν ποντάρίσματα που αφορούν τον ίδιο αγώνα αλλά διαφορετικά τελικά ενδεχόμενα από διαφορετικές εταιρίες συνδυάζονται και προκύπτει αρνητική γκανιότα. Για παράδειγμα έστω οι ακόλουθες αποδόσεις του πίνακα:

| ΕΤΑΙΡΙΑ Α: ΑΠΟΔΟΣΗ ΝΙΚΗΣ ΓΗΠΕΔΟΥΧΟΥ | ΕΤΑΙΡΙΑ Β: ΑΠΟΔΟΣΗ ΙΣΟΠΑΛΙΑΣ | ΕΤΑΙΡΙΑ Γ: ΑΠΟΔΟΣΗ ΙΣΟΠΑΛΙΑΣ |
|---|---|---|
| **4** | 4 | 4 |

Και ο παίκτης ποντάρει 1 χρηματική μονάδα σε κάθε ενδεχόμενο. Τότε θα έχει ποντάρει συνολικά 3 χρηματικές μονάδες αλλά ανεξάρτητα από την έκβαση του αγώνα θα καταλήξει με 4 χρηματικές μονάδες και εν τέλει το κέρδος του θα είναι ίσο με 1.

Το προηγούμενο παράδειγμα αποτελεί ακραία περίπτωση διότι προκύπτει γκανιότα ίση με -25%. Στην πραγματικότητα μπορεί να δημιουργηθεί οριακά αρνητική γκανιότα μερικών ποσοστιαίων μονάδων. Οι εταιρείες καταβάλλουν σημαντικές προσπάθειες για να μην συμβαίνει αυτό αλλά δεν είναι πάντα επιτυχείς. Το βέλτιστο ποντάρισμα ακόμα και σε περιπτώσεις που είναι δυνατό το arbitrage μπορεί να προσδιορισθεί φυσιολογικά και με τη χρήση του Kelly criterion.

## *2.3 Κριτήριο Kelly*

Έστω το ακόλουθο παράδειγμα. Η εθνική ομάδα της Ελλάδας αντιμετωπίζει αυτήν της Σλοβενίας σε έναν knock out αγώνα basketball. Τα πιθανά αποτελέσματα είναι μόνο δύο (νίκη Ελλάδας ή Σλοβενίας) και έμπειροι αναλυτές εκτιμούν σωστά εκ των προτέρων ότι αυτά είναι πραγματικά ισοπίθανα. Κάποια εταιρεία δίνει απόδοση 2.25 στη νίκη της Σλοβενίας και 1.8 στη νίκη της Ελλάδας. Ένας συχνός παίκτης (ας πούμε ο Γιώργος) έχοντας μάθει όλα τα παραπάνω αισθάνεται ότι εντόπισε μια νέα ευκαιρία για να αποκτήσει περισσότερα κέρδη. Αρχικά, εκτιμά (ορθά) ότι μια εταιρεία δίνει αυτές τις αποδόσεις - παρά τις (σωστές για την υπόθεση του παραδείγματος) εκτιμήσεις των αναλυτών – διότι εκμεταλλεύεται τον συναισθητισμό του όλου θέματος (που θα οδηγήσει πολλούς να ποντάρουν στην Ελλάδα) για να μεγιστοποιήσει τα κέρδη του. Αντιλαμβάνεται ότι συνέπεια αυτής της τακτικής είναι να παρασυρθεί ελαφρώς προς τα πάνω η απόδοση της Σλοβενίας.

Ο Γιώργος αποφασίζει να ποντάρει στη νίκη της Σλοβενίας που εκτιμά ότι αποτελεί ευκαιρία. Έστω ότι διαθέτει 100 ευρώ διαθέσιμά για στοιχηματισμό για τις επόμενες 30 μέρες το οποίο μάλιστα είναι ένα όριο που του έχει επιβάλλει η σύζυγος του (Ελένη). Επιλέγει να ποντάρει 50 από αυτά στην Σλοβενία αν και ο ίδιος θα ήθελε να στοιχηματίσει ελαφρώς περισσότερα αλλά γνωρίζει πως αν τα χάσει η «ζημιά» για τον ίδιο της αποχής από το παιχνίδι θα είναι μεγάλη. Προσπαθεί να πιέσει την Ελένη έτσι ώστε να ποντάρουν περισσότερα χρήματα (που δεν θα προσμετρώνται στα προηγούμενα 100) αλλά η πρόταση του απορρίπτεται. Η Ελένη ομολογουμένως δεν αντιλαμβάνεται (και μάλλον ούτε θέλει) τη συλλογιστική του πορεία. Εστιάζει στο γεγονός ότι ο Γιώργος μονίμως χάνει λεφτά σε βάθος χρόνου όσο στοιχηματίζει. Θα εξηγηθεί στη συνέχεια γιατί και τα δύο μέρη στην προηγούμενη αντιπαράθεση «έχουν δίκιο» και η σύντομη ιστορία των πρωταγωνιστών θα βοηθήσει να εξοικειωθεί κάποιος αναγνώστης με τις έννοιες που θεμελιώνουν τη μεθοδολογία που προτάσσεται.

Πράγματι, το ποντάρισμα στην Σλοβενία αποτελεί «ευκαιρία». Διότι εάν υπολογίσουμε απλά την αναμενόμενη τιμή του τελικού κεφαλαίου θα είναι $100 - 50 + 0.5 * 2.25 * 50 = 106.25 > 100$. Εδώ βέβαια αξίζει να αναφερθεί πως ο προηγούμενος υπολογισμός που δείχνει την αξία της ευκαιρίας γίνεται με την προϋπόθεση ότι για μέχρι

μερικές δεκάδες ευρώ το όφελος για τον Γιώργο να κερδίζει $X$ χρήματα και η ζημιά να χάσει $X$ θα έχουν ίσο και αντίθετο αντίκτυπο στην καθημερινότητά του. Οπότε όντως ο Γιώργος εντόπισε μια στοιχηματική ευκαιρία.

Ο Γιώργος έχοντας υπολογίσει τη μέση τιμή του κέρδους του από το στοίχημα απογοητεύεται από το μικρό μέσο κέρδος του (γι' αυτό και ήθελε να ποντάρει παραπάνω). Σκέφτεται όμως πως εν τέλει δεν υπάρχει πρόβλημα που δεν μπορεί να ποντάρει όσα θέλει διότι πιστεύει πως αν εντοπίζει συνεχώς παρόμοιες «ευκαιρίες» και ποντάρει θα καταφέρει να συσσωρεύσει πολλά κέρδη επαναποντάροντας συνεχώς με τον ίδιο τρόπο. Μάλιστα, συγκρίνει τον εαυτό του με κάποιους φίλους του που σταθερά ποντάρουν σε αυτό που αντιλαμβάνονται ως φαβορί (λίγο ή πολύ ανεξάρτητα από την απόδοση του αποτελέσματος) και νιώθει πιο έξυπνος με το σύστημα που μόλις επινόησε.

Η πραγματικότητα όμως θα έρθει και θα διαψεύσει τις ελπίδες του. Ας γίνει η υπόθεση ότι ο Γιώργος όντως εντοπίζει ορθά στοιχηματικές ευκαιρίες κατά τον προηγούμενο τρόπο. Δίχως βλάβη της γενικότητας ας υποθέσουμε επίσης ότι οι επόμενες ευκαιρίες που θα εντοπίσει ο Γιώργος θα είναι της μορφής του προηγούμενου αγώνα μπάσκετ και ότι οι αποδόσεις θα είναι ίσες με αυτές που αναφέρθηκαν πριν. Αν αρχικά η Σλοβενία κερδίσει τότε ο Γιώργος σύμφωνα με τα προηγούμενα θα έχει πλέον κεφάλαιο ίσο με 162.5 ευρώ. Στον επόμενο αγώνα επιλέγει φυσιολογικά την ίδια τακτική και ποντάρει το μισό κεφάλαιο αλλά τυχαίνει να χάσει, οπότε μένει με κεφαλαίο ίσο με 81.25 ευρώ. Εύκολα μπορεί να καταλάβει κανείς ότι έπειτα από N διαδοχικές στοιχηματικές νίκες αρκούν N διαδοχικές ήττες για να καταλήξει ζημιογόνος ο Γιώργος. Επίσης, εάν νικήσει η Ελλάδα τότε ο Γιώργος θα μείνει με 50 ευρώ τα οποία θα γίνουν πάλι 81.25 αν νικήσει το επόμενο στοίχημα. Δηλαδή, δεν αρκούν N διαδοχικές νίκες έπειτα από ήττες ούτε καν για να ισοφαρίσει. Είναι ευκολονόητο πως αυτό συμβαίνει διότι σε κάθε ήττα το κεφάλαιο υποδιπλασιάζεται αλλά σε κάθε νίκη δεν διπλασιάζεται. Έτσι, δεδομένου του ότι σε βάθος χρόνου θα χάσει όσες φορές κερδίσει τα στοιχήματα που θα βάλει σε αυτούς τους αγώνες, η πιθανότητα που έχει ο Γιώργος να ξεμείνει από χρήματα τείνει να γίνει βεβαιότητα όσο επιμένει να στοιχηματίζει κατά αυτόν τον τρόπο που μεμονωμένα είναι "ευκαιρία".

Αυτό συμβαίνει διότι παρόλο που για τον ίδιο η απώλεια μερικών δεκάδων ευρώ είναι «αντίθετη» και «ίση» με το κέρδος τους στη καθημερινότητά του, δεν ισχύει το ίδιο εάν ο στόχος είναι η ανάπτυξη του κεφαλαίου του. Η μείωση 50% του αρχικού κεφαλαίου είναι πολύ πιο ζημιογόνα από την αύξηση κατά 62.5%. Εδώ αξίζει να αναφερθεί ότι η ιστορία του Γιώργου ενώ μοιάζει αρκετά συνηθισμένη στο χώρο των καθημερινών παικτών, δεν διαφέρει σχεδόν καθόλου από τις τακτικές που υλοποιούνται στη δημοσιευμένη έρευνα που χρησιμοποιεί Μηχανική Μάθηση για να στοιχηματίσει σε αθλήματα [3], [7]. Ο Γιώργος κατάφερε να παραμείνει ζημιογόνος ακόμα και όταν στην υπόθεση του παραδείγματος είχε ορθές προβλέψεις σύμφωνα με όλα τα δεδομένα που υπήρχαν πριν την έναρξη του αγώνα. Στην

πραγματικότητα, αυτή η υπόθεση δεν ισχύει σχεδόν ποτέ. Επομένως, φαίνεται πως είναι απόλυτα καθοριστικός ο τρόπος πονταρίσματος στην πορεία που θα ακολουθήσουν τα κέρδη.

Εάν είναι επιθυμητή η κερδοφορία σε βάθος χρόνου χρειάζεται με κάποιο τρόπο να ληφθεί υπόψη ότι η μείωση ενός μεριδίου $X$ του κεφαλαίου προκαλεί μεγαλύτερη ζημία από το όφελος της ισόποσης αύξησης. Αρχικά, ας επιστρέψουμε στην προηγούμενη υπόθεση των όμοιων στοιχημάτων που ποντάρει ο Γιώργος. Είναι εύλογο να διαπιστώσουμε ότι εφόσον δεν υπάρχει καμία μεταβολή των παραμέτρων του προβλήματος -δεδομένου ενός κεφαλαίου $K$ σε κάθε ευκαιρία μεμονωμένου στοιχηματισμού-, το κεφάλαιο που θα αποφασίσουμε να ποντάρουμε (εάν τελικά αυτό συμφέρει να γίνει εντελώς) θα είναι ένα σταθερό κεφάλαιο $l * K$ με $0 \leq l \leq 1$. Εδώ χρειάζεται να τονιστεί πως το $l$ είναι ανεξάρτητο του $K$. Αυτό διότι σε διαφορετική περίπτωση θα εκφραζόταν κάποια υποκειμενική αξία για τον παίκτη που έχουν διαφορετικά ύψη του $K$ η οποία δεν γίνεται να συνάδει με το έργο της αμιγούς μεγιστοποίησης της κερδοφορίας. Έτσι, θεμελιώνουμε ότι στη προσπάθεια για να μεγιστοποιηθεί το κεφάλαιο σε βάθος χρόνου χρειάζεται να διατίθεται ένα σταθερό μέρος $l$ του κεφαλαίου. Εφαρμόζοντας αυτή τη στρατηγική, σε κάθε γύρο στοιχηματισμού ο λόγος του αναμενόμενου τελικού κεφαλαίου προς αυτού στο τέλος του προηγούμενου θα παραμείνει ίδιος. Αυτό έχει ως συνέπεια ότι αναμένεται το κεφάλαιο να ακολουθήσει μια (περίπου) αυξητική ή μειωτική εκθετική πορεία ανάλογα φυσικά με το $l$.

Το πρόβλημα της βέλτιστης διαμόρφωσης του στοιχηματισμού από τη πλευρά του παίκτη ανάγεται πλέον στη μεγιστοποίηση του εκθετικού ρυθμού ανάπτυξης του κεφαλαίου. Αυτό το μέγεθος $G$ εκφράζεται από την ακόλουθη παράσταση:

$$G = \lim_{N \to \infty} \frac{1}{N} \log_2 \frac{V_N}{V_0}$$

Όπου $N$ οι γύροι στοιχηματισμού, $V_0$ το αρχικό κεφάλαιο και $V_N$ το κεφάλαιο του τέλος του $N$-οστού γύρου. Για παράδειγμα, $G = 1$ ισούται με διπλασιασμό (κατά μέσο όρο) του κεφαλαίου σε κάθε γύρο στοχηματισμού. Επίσης έχουμε:

$$V_N = (1 + (o - 1) * l)^W (1 - l)^L V_0$$

Όπου $W$ το πλήθος των στοιχηματικών επιτυχιών και $L$ των αποτυχιών και $o$ η απόδοση του στοιχήματος. Επομένως ισχύει:

$$G = \lim_{N \to \infty} \left( \frac{W}{N} \log(1 + (o - 1) * l) + \frac{L}{N} \log(1 - l) \right)$$
$$= plog(1 + (o - 1)l) + (1 - p)log(1 - l)$$

Όπου $p$ η πιθανότητα στοιχηματικής νίκης. Εξετάζοντας απλά τη μονοτονία της παραπάνω συνάρτησης ως προς τη ποσότητα $l$ εύκολα ευρίσκεται ότι η παραπάνω παράσταση μεγιστοποιείται για:

$$l = p - \frac{1-p}{o-1}$$

Το οποίο για το παράδειγμα του Γιώργου ισούται με 0.1.

Η παραπάνω διαδικασία προσδιορισμού του βέλτιστου ποσού πονταρίσματος μεγιστοποιώντας τον εκθετικό ρυθμό ανάπτυξης του κεφαλαίου ονομάζεται κριτήριο Kelly (Kelly criterion). Ο John Larry Kelly Jr. υπήρξε ερευνητής στα Bell Labs. Διεξάγοντας έρευνα μαζί με τον Claude Shannon δημιοσίευσε το σύγγραμα "A New Interpretation of Information Rate" στο οποίο πολύ ευρηματικά συνέδεσε την αποτελέσματα του Shannon επάνω στη μέγιστη ροή πληροφορίας μέσα από ένα θορυβώδες κανάλι μετάδοσης δεδομένων. Συγκεκριμένα, στο εν λόγω σύγγραμα, ο Kelly αρχικά φέρνει το παράδειγμα ενός παίκτη που παίζει ένα στοίχημα όμοιο με του Γιώργου με τη διαφορά ότι ο παίκτης μαθαίνει την πληροφορία για τον νικητή εκ των προτέρων με πιθανότητα $p$. Αυτό υλοποιείται στο σύγγραμα χρησιμοποιώντας ένα κανάλι θορυβώδες επικοινωνίας μέσα από το οποίο η πιθανότητα να μεταδοθεί σωστά η πληροφορία για τον νικητή πριν γίνει ο αγώνας είναι $p$. Ο Kelly δείχνει πως το μέγιστο $G$ που επιτυγχάνεται είναι όμοιο με τον μέγιστο ρυθμό πληροφορίας που μπορεί να μεταδοθεί από το κανάλι!

Ο Kelly απεβίωσε σε μικρή ηλικία πιο πριν από όταν έγιναν δημοφιλή τα ευρήματα του τα οποία πήραν δημοσιότητα όταν πιο σύγχρονοι μας ερευνητές τα αξιοποίησαν. Είναι ευρέως γνωστό ότι ο μετέπειτα ερευνητής και συνεργάτης του Shannon ονόματι Ed Thorp αξιοποίησε τα ευρήματα του Kelly (πριν γίνουν ευρέως γνωστά) με αποτέλεσμα να αποκομίσει πολλά κέρδη σε επενδυτικές δραστηριότητες [14]. Στη συνέχεια προς τα τέλη του 20$^{ου}$ αιώνα, η τεχνική που παρουσίασε ο Kelly με πολλές παραλλαγές έγινε πια κοινή πρακτική στην επιτυχημένη επενδυτική κοινότητα .

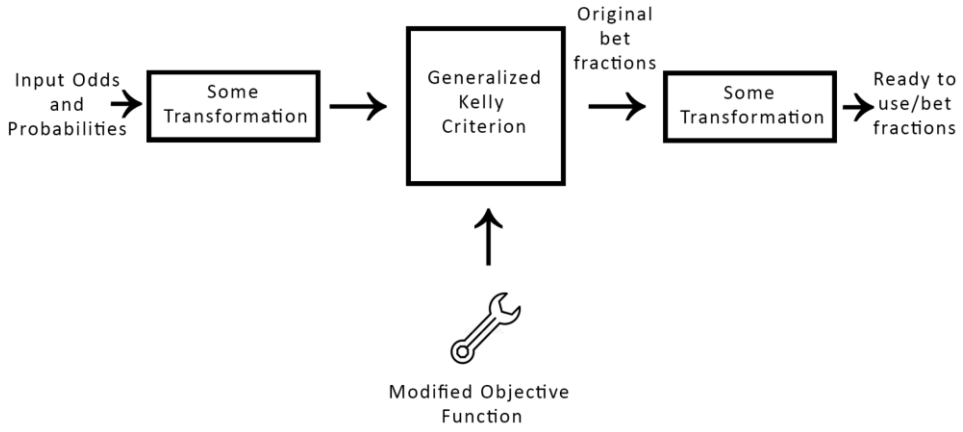
Είναι προφανές ότι σε περίπτωση περισσότερων από 2 τελικών ενδεχομένων (όπως σε έναν αγώνα 1-2-X) η μεθοδολογία παραμένει η ίδια και μεταβάλλεται η παράσταση $V_N$. Πλέον η εύρεση των βέλτιστων τοποθετήσεων στο παίγνιο γίνεται με υπολογιστικές μεθόδους και δεν μπορεί να εξαχθεί ένας συνεπτυγμένος τύπος όπως έγινε προηγουμένως στο απλό παράδειγμα του Γιώργου. Σε αυτό το σημείο, είναι σημαντικό να αναφερθούν τα ακόλουθα.

- Στις προηγούμενες σελίδες, έγινε η ουσιαστική διαπίστωση ότι η μείωση ενός μεριδίου $X$ του κεφαλαίου προκαλεί μεγαλύτερη ζημία από το όφελος της ισόποσης αύξησης. Προκύπτει η όμορφη παρατήρηση από τις προηγούμενες σχέσεις ότι η λογαριθμική καμπυλότητα είναι η πλέον κατάλληλη για να «εκφράσει» αυτή την ανισορροπία. Σε αντίθεση με το μέσο τελικό κεφάλαιο έπειτα από N γύρους που προκύπτει από το στάθμισμα των τελικών πιθανών κεφαλαίων με βάρη της αντίστοιχες πιθανότητες εκπλήρωσης (και μεγιστοποιείται με

ποντάρισμα όλου του κεφαλαίου), το βέλτιστο ποντάρισμα κατά Kelly προκύπτει από το ίδιο στάθμισμα των λογάριθμων των πιθανών τελικών κεφαλαίων.

- Κάποιος ενδεχομένως να εξακολουθεί να υποστηρίζει ότι είναι προτιμότερο το ποντάρισμα "all in" διότι έτσι όντως μεγιστοποιείται το αναμενόμενο τελικό κεφάλαιο. Σίγουρα, εντός ορισμένων $N$ γύρων αυτό εξαρτάται από την αξία που έχει για τον επενδυτή η πιθανότητα να ξεμείνει από κεφάλαιο. Οπότε αν δεν τον απασχολεί ιδιαίτερα η ζημιά, όντως το βέλτιστο όφελος για τον ίδιο δεν εκφράζεται από το πιθανοτικό στάθμισμα των λογάριθμων. Είναι αλήθεια όμως ότι όσο γίνονται περισσότεροι οι στοιχηματικοί γύροι, η πιθανότητα η τεχνική Kelly να δώσει το υψηλότερο κεφάλαιο τείνει να γίνεται βεβαιότητα και ο παίκτης "all in" θα καταλήξει σίγουρα χρεοκοπημένος.
- Εξετάζοντας τις προηγούμενες μαθηματικές σχέσεις του Kelly criterion είναι αντιληπτό ότι θεωρητικά εξαλείφεται η πιθανότητα ολοκληρωτικής χρεοκοπίας. Αυτό γιατί καθώς ένα ενδεχόμενο τελικό κεφάλαιο θα προσέγγιζε το μηδέν, θα οδηγούσε τη παράσταση προς μεγιστοποίηση $G$ να τείνει στο $-\infty$. Στην πράξη όμως είναι δυνατή η χρεοκοπία διότι το χρήμα διαθέτει πάντα κάποια ελάχιστη υποδιαίρεση.

Τέλος, είναι κρίσιμο να αναφερθεί το εξής. Το κριτήριο Kelly σχεδόν ποτέ δεν είναι δυνατό να εφαρμοστεί στην πραγματική ζωή "out of the box". Μικρά σφάλματα στις ακριβείς προβλέψεις που είναι σίγουρα αναπόφευκτα αλλάζουν καθοριστικά το βέλτιστο ποσό που προκύπτει με αποτέλεσμα η πορεία του κεφαλαίου να γίνεται «εκρηκτική» προς όλες τις κατευθύνσεις και συνήθως να καταλήγει να εξανεμίζεται. Στην πράξη αυτό αντιμετωπίζεται ποντάροντας πιο συντηρητικά (κατ' ουσίαν παρεμβαίνοντας στην τεχνική του Kelly Criterion) από ότι υποδεικνύεται άμεσα από το κριτήριο. Έτσι περιορίζονται τα κέρδη αλλά γίνεται λιγότερο πιθανή η καταστροφή. Αυτό μπορεί να επιτευχθεί προσαρμόζοντας τις παραμέτρους εισόδου τού υλοποιημένου αλγόριθμού, τη συνάρτηση βελτιστοποίησης ή τα αποτελέσματα εξόδου. Αντικείμενο αυτής της εργασίας θα είναι και να βρεθεί μια καλή πρακτική για να γίνει αυτή η παρέμβαση στο Kelly Criterion. Έχοντας κατά νου όλα τα παραπάνω, εάν ο Γιώργος δεν ήταν σίγουρος ότι οι εκτιμητές δίνουν απολύτως ορθές εκτιμήσεις θα ήταν μάλλον καλύτερο να ποντάρει μέχρι 5 ευρώ από το σύνολο των 100 που διαθέτει.



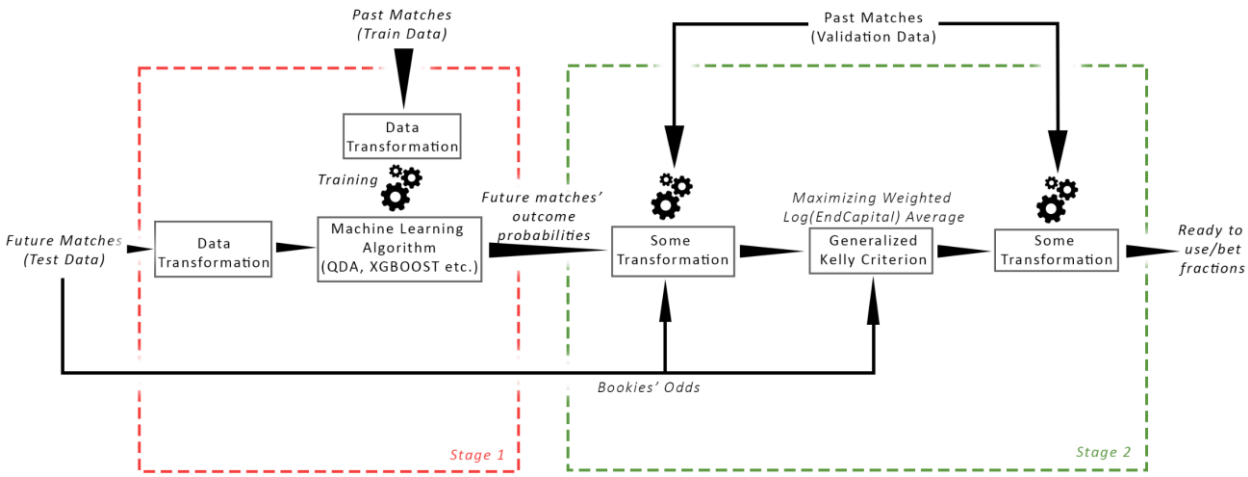

**Εικόνα 6 Σχηματικό διάγραμμα που απεικονίζονται όλοι οι πιθανοί τρόποι παρέμβασης στο "καθαρό" κριτήριο Kelly. Η χρήση του για πραγματικά προβλήματα συνήθως απαιτεί τροποποιήσεις**

# 3

# *Δομή της Κατασκευής και Υλοποίηση της*

## *3.1 Εποπτική Περιγραφή*

Συνεχίζοντας από τη θεωρία του προηγούμενου κεφαλαίου θα προχωρήσουμε εδώ σταδιακά στη πράξη περιγράφοντας τα κύρια σημεία της δομής που αφορά η εργασία. Ο στόχος (όπως έχει ήδη αναφερθεί) είναι να παρουσιαστεί μια δομή προσδιορισμού στοιχηματικών επιλογών η οποία θα είναι όσο το δυνατό περισσότερο έτοιμη για να χρησιμοποιηθεί στον πραγματικό κόσμο εντός βέβαια των χρονικών πλαισίων μιας εξάμηνης διπλωματικής εργασίας. Η δομή αυτή θα έχει ως εισόδους τα δημοσίως διαθέσιμα παλαιά ποδοσφαιρικά στατιστικά και στο τέλος θα επιστρέφει ποσά που προτείνει για στοιχηματισμό ως ποσοστά του συνολικού κεφαλαίου για μελλοντικούς αγώνες. Ενδιάμεσα θα χωρίζεται σε δύο στάδια. Το πρώτο υλοποιεί μοντέλα Μηχανική Μάθησης τα οποία εκπαιδεύονται επάνω στα στατιστικά και έπειτα επιστρέφουν πιθανότητες πραγματοποίησης των διαφόρων τελικών ενδεχομένων των μελλοντικών αγώνων (νίκη, ήττα ή ισοπαλία της γηπεδούχου κάθε φορά). Το δεύτερο στάδιο αποτελεί μια ειδική υλοποίηση του κριτηρίου Kelly η οποία θα επιστρέφει ποσά χρημάτων χρησιμοποιώντας ως εισόδους τις πιθανότητες του 1$^{ου}$ σταδίου και τις αποδόσεις των εταιριών. Αυτό η διαδικασία γίνεται για κάθε αγώνα ξεχωριστά θεωρώντας πως όλοι αγώνες γίνονται σειριακά και μπορούμε να ποντάρουμε ξεχωριστά σε κάθε έναν. Δεδομένου ότι (όπως εξηγήθηκε στο προηγούμενο κεφάλαιο) το 2$^{ο}$ στάδιο επιστρέφει τα βέλτιστα ποσά εάν οι πιθανότητες που λαμβάνει είναι αντιπροσωπευτικές της πραγματικότητας, το γενικότερο πρόβλημα του προσδιορισμού των βέλτιστων στοιχηματικών ποσών ανάγεται στην εύρεση των πλέον αντιπροσωπευτικών πιθανοτήτων για τα τελικά αποτελέσματα του κάθε αγώνα. Η όποια παραμετροποίηση του 2$^{ου}$ σταδίου γίνεται διότι οι πιθανότητες που προκύπτουν από το 1$^{ο}$ δεν είναι απόλυτα αντιπροσωπευτικές.




Εικόνα 7 Διαγραμματική αναπαράσταση της δομής της εργασίας.

## 3.2 Στάδιο 1: Ταξινομητής

Για τις προβλέψεις αγώνων γενικά έχουν παρουσιαστεί υλοποιήσεις που βασίζονται είτε σε πιο παραδοσιακές στατιστικές μεθόδους είτε και σε τεχνικές Μηχανική Μάθησης (περισσότερο τα τελευταία χρόνια). Ανάμεσα βέβαια στις διάφορες τεχνικές (Μηχανικής Μάθησης) υπάρχουν μοντέλα που βασίζονται σε πιο παραδοσιακές στατιστικές μεθόδους όπως ο Gaussian Naive Bayes (βλ. κεφάλαιο θεωρίας) οι οποίες επιχειρούν να υπολογίσουν με κατάλληλο τρόπο παραμέτρους γνωστών στατιστικών κατανομών ή καμπυλών που περιγράφουν τα δείγματα αλλά και άλλες πιο σύγχρονες που προσπαθούν με εντελώς διαφορετικές μεθόδους να δημιουργήσουν έναν ταξινομητή όπως τα νευρωνικά δίκτυα ή μοντέλα που υλοποιούν δένδρα αποφάσεων. Έτσι, η χρήση της Μηχανική Μάθησης σήμερα ουσιαστικά καλύπτει μια ποικιλόμορφη γκάμα πιθανών λύσεων για το πρόβλημα της ταξινόμησης και θα χρησιμοποιηθεί σε αυτή την εργασία.

Παρόλο το έντονο ενδιαφέρον όμως, το πρόβλημα στην έρευνα της εύρεσης του νικητή πριν έναν αγώνα ποδοσφαίρου αποτελεί ένα πρόβλημα για το οποίο δεν έχουν υπάρξει αρκετά ικανοποιητικές λύσεις. Αυτό διότι δεν είναι ιδιαίτερα ικανοποιητικό ένα ποσοστό ορθών προβλέψεων που κυμαίνεται συνήθως μεταξύ 50-75% [3], [4], [5], [15] ενώ αν απλά «ποντάραμε» για νικητή την γηπεδούχο ομάδα το ποσοστό αυτό θα ήταν περίπου 50%. Για αυτό το στάδιο στην κατασκευή της εργασίας όμως το πρόβλημα είναι γενικότερο από αυτό και έχει να κάνει με τον υπολογισμό των πιθανοτήτων των τριών αποτελεσμάτων. Εδώ δεν υπάρχει ακριβώς σχετική δημοσιευμένη έρευνα που να επισημαίνει τις καλές πρακτικές. Ως εκ τούτου, η επιλογή των διάφορων υλοποιήσεων που θα δοκιμαστούν θα γίνει έχοντας κατά νου ποια μοντέλα δίνουν τις καλύτερα ποσοστά επιτυχίας στο πρόβλημα της εύρεσης του νικητή και ποια μοντέλα είναι καταλληλότερα για την εξαγωγή πιθανοτήτων. Γενικά νεότερες δομές

που δεν βασίζονται σε στατιστικές μεθόδους τείνουν να μην δίνουν αντιπροσωπευτικές πιθανότητες στα προβλήματα που εφαρμόζονται διότι πολλές φορές εξαιτίας της δομής του είναι εύκολο να δοθεί το φαβορί ενός αγώνα αλλά δύσκολο να αποτυπωθούν σε πιθανότητες πόσο «εκτιμούν» ως φαβορί την κάθε ομάδα. Οι αλγόριθμοι/μοντέλα Μηχανικής Μάθησης που θα χρησιμοποιηθούν στη παρούσα εργασία είναι κάποιοι από όσους αναφέρθηκαν στο κεφάλαιο της θεωρίας.

Ένα σημαντικό ζήτημα που προκύπτει σε αυτό το στάδιο είναι επίσης η επιλογή των δεδομένων εισόδου. Τα δεδομένα εισόδου στα μοντέλα Μηχανική Μάθησης είναι κατά γενική ομολογία εξίσου ή και περισσότερα σημαντικά από τη φύση των ίδιων των μοντέλων. Στα πλαίσια αυτής εργασίας πρόκειται να χρησιμοποιηθούν δημοσίως διαθέσιμα δεδομένα. Έτσι, ο καθένας μπορεί να αναπαραγάγει τα αποτελέσματα και ταυτόχρονα φαίνεται η δύναμη των ελεύθερα διαθέσιμων δεδομένων του διαδικτύου. Τα ελεύθερα δεδομένα που επιλέχθηκαν είναι αυτά που είναι συγκεντρωμένα στη σελίδα football-data.co.uk και fifaindex.com. Από των σύνολο των πρωταθλημάτων για τα οποία υπάρχουν συγκεντρωμένα στατιστικά εμείς θα ασχοληθούμε με τα κύρια πρωταθλήματα της Αγγλίας (Premier League), της Γερμανίας (Bundesliga), της Ιταλίας (Serie A) και της Ισπανίας (La Liga). Τα στατιστικά για κάθε αγώνα σε κάθε σεζόν περιλαμβάνουν χαρακτηριστικά που αφορούν τα γκολ που επετεύχθησαν, τα shoots, corners και πλήθος κίτρινων και κόκκινων καρτών. Επίσης, επειδή αυτά είναι καταγεγραμμένα με σειρά χρονολογική μπορεί κανείς και να εξάγει ανά πάσα ημερομηνία τη βαθμολογική κατάταξη. Τα στατιστικά των αγώνων στη συνέχεια μετασχηματίζονται και εξάγονται νέα χαρακτηριστικά που αξιοποιούνται από τους ταξινομητές. Εδώ αξίζει να σημειωθεί πως τα δεδομένα που αναφέρθηκαν επ' ουδενί δεν δίνουν μια εντελώς αντιπροσωπευτική εικόνα της πραγματικότητας πριν τον κάθε αγώνα και ενδεχομένως αυτό να αποτελεί το κύριο αίτιο πίσω από τα «φτωχά» αποτελέσματα της χρήσης των αλγορίθμων μηχανικής μάθησης στη πρόβλεψη ποδοσφαιρικών αγώνων. Τραυματισμοί, προπονητικές επιλογές για τη προσπάθεια που επιλέγεται να «δοθεί» σε κάποιον αγώνα, ψυχολογικοί παράγοντες και πιο λεπτομερή στατιστικά αγώνων σίγουρα θα δίνουν μια καλύτερη εικόνα αλλά δεν ευρέθησαν.

Ουσιαστικά, το πρόβλημα δημιουργίας του βέλτιστου ταξινομητή έχει τις ακόλουθες «μεταβλητές» στην πορεία που θα ακολουθηθεί. Πρώτη «μεταβλητή» αποτελεί ακριβώς το ίδιο το μοντέλο (π.χ. k-NN, XGBoost κ.α.). Μετά έχουμε για κάθε μοντέλο τις λειτουργικές παραμέτρους του εάν αυτές υπάρχουν (π.χ. αριθμός γειτόνων για k-NN). Τέλος, έχουμε τα features που χρησιμοποιούνται για τα μοντέλα που δεν έχουν τη δυνατότητα να δίνουν μεγαλύτερη «σημασία» στα σημαντικότερα εξ αυτών και λιγότερη στα πιο «ασήμαντα».

## *3.3 Στάδιο 2: Κριτήριο Kelly*

Το αντικείμενο εδώ είναι η εξαγωγή ποσοστών του κεφαλαίου που προτείνεται να τοποθετηθούν σε στοιχηματικές επιλογές ανά αγώνα. Όπως εξηγήθηκε και στη θεωρία, το κριτήριο Kelly στον πραγματικό κόσμο, όταν αυτό χρησιμοποιείται "out of the box", οδηγεί στον εξανεμισμό του κεφαλαίου με τη λύση για αυτό να είναι να θυσιαστεί το δυναμικό των κερδών υπέρ ενός πιο «συντηρητικού» παιξίματος. Αυτό το παίξιμο, ναι μεν θα δίνει αυξημένη κερδοφορία σε σχέση με άλλες μεθόδους αλλά θα περιορίζει πολύ «τολμηρές» κινήσεις που θα οδηγούσαν σε απώλεια κεφαλαίου εάν προκύπτουν από εσφαλμένες εκτιμήσεις. Γι' αυτό και τοποθετούνται μετασχηματισμοί στην είσοδο και στην έξοδο του «Kelly» των οποίων ο προσδιορισμός θα γίνει κατά τη διαδικασία του validation που αναφέρθηκε προηγουμένως. Εδώ πρέπει να αναφερθεί ότι θεωρητικά θα μπορούσε να γίνει και τροποποίηση στην αντικειμενική συνάρτηση μεγιστοποίησης του Kelly (βλ. κεφάλαιο θεωρίας) αλλά δεν θα γίνει λόγω ανεπαρκούς χρόνου για να εξεταστεί και αυτή η λύση κατά την εκπόνηση της παρούσας εργασίας. Στην ευρύτερη βιβλιογραφία παρόλα αυτά, συνήθως προσδιορίζονται μονάχα (γραμμικοί) μετασχηματισμοί στην έξοδο του αλγορίθμου (στα τελικά ποσά) [8] [16]

«Πατώντας» επάνω σε όσα αναφέρθηκαν στη θεωρία, το 2ο στάδιο έχει ως στόχο την εύρεση των ποσών στοιχηματισμού $b_h, b_d, b_a$ για νίκη, ισοπαλία ή ήττα της γηπεδούχου αντίστοιχα τα οποία μεγιστοποιούν την ακόλουθη συνάρτηση (που αποτελεί τον αναμενόμενο εκθετικό ρυθμό αύξησης του κεφαλαίου).

$$G = p_h \log(1 + (o_h - 1)b_h - b_d - b_a) + p_d \log(1 - b_h + (o_d - 1)b_d - b_a) + p_a \log(1 - b_h - b_d + (o_a - 1)b_a)$$

Όπου $o_h, o_d, o_a$ οι αποδόσεις των στοιχηματικών εταιριών για νίκη, ισοπαλία και ήττα της γηπεδούχου και $p_h, p_d, p_a$ οι εκτιμήσεις που προκύπτουν από το 1ο στάδιο για τα τρία ενδεχόμενα πάλι.

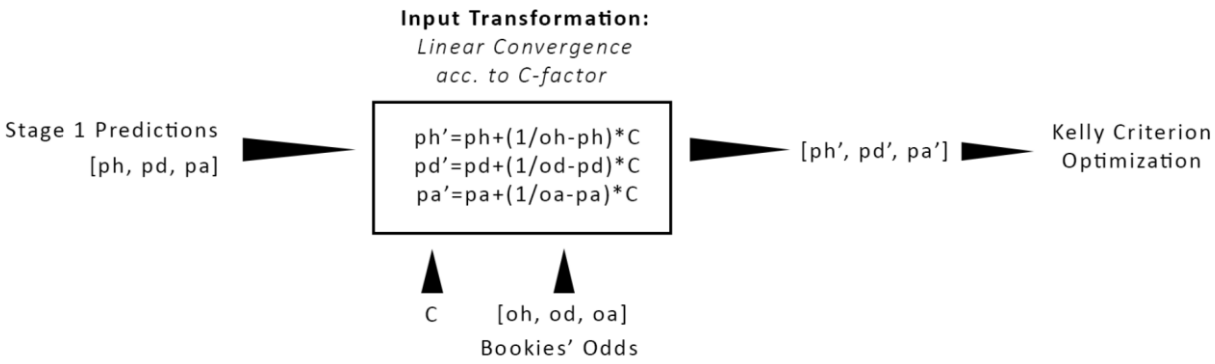
Όσο αφορά τις υλοποιήσεις των μετασχηματισμών στην έξοδο, έγιναν τα εξής. Αρχικά, ακολουθώντας κοινές ερευνητικές πρακτικές [8] [16], δοκιμάστηκε ο απλός γραμμικός μετασχηματισμός που δίνει ως βέλτιστο ποσοστό ίσο με την έξοδο του Kelly διαιρεμένη με μια σταθερά μεγαλύτερη ή ίση με 1 (Fractional Kelly). Στη συνέχεια και σε συνδυασμό με το προηγούμενο, δοκιμάστηκαν μέθοδοι δημιουργίας low ή high pass κατωφλίου στα στοιχήματα έτσι ώστε να ποντάρονται ή όχι ποσοστά που είναι ίσα ή μεγαλύτερα από έναν αριθμό. Αυτό έδωσε την έμπνευση για τον ορισμό μιας παραμετρικής συνάρτησης «ράμπας» η οποία θα επιστρέφει έναν διαφορετικό συντελεστή διαίρεσης αναλόγως με το ποσοστό που επιλέγεται να ποντάριστεί. Οι παραπάνω μετασχηματισμοί στα τελικά ποσά που δίνει το κριτήριο Kelly έδωσαν καλύτερα αποτελέσματα γενικά από την έρευνα που υπάρχει διαθέσιμη αλλά υπήρχαν σημαντικές αστάθειες σε κάποια πρωταθλήματα.

Κατά τις δοκιμές του συστήματος, κάποια στιγμή έγινε η ακόλουθη παρατήρηση. Δίχως βλάβη της γενικότητας ας υποθέσουμε ότι σε έναν αγώνα υπήρχε μόνο ένα ενδεχόμενο, αυτό της νίκης γηπεδούχου, που θα μπορούσε να είναι κερδοφόρο δεδομένων των αποδόσεων και των εκτιμήσεων. Εάν η πιθανότητα νίκης από το 1ο στάδιο ήταν απλά λίγο μεγαλύτερη από αυτήν που «θέτει» ως κατώφλι η απόδοση για να είναι κερδοφόρο μεμονωμένα το στοίχημα (βλ. θεωρία) τότε το κριτήριο Kelly επέστρεφε φυσιολογικά ένα μικρό ποσοστό του κεφαλαίου ως προτεινόμενο ποντάρισμα (και μόνο για αυτό το στοίχημα). Εάν όμως η διαφορά μεταξύ των προηγούμενων δύο μεγεθών ήταν πολύ μεγάλη, το κριτήριο έδινε ένα σαφώς μεγαλύτερο προτεινόμενο ποντάρισμα (για αυτό το στοίχημα) αλλά επιπλέον πρότεινε ποσοστά για ποντάρισμα και στα ενδεχόμενα ισοπαλίας ή ήττας τα οποία μεμονωμένα δεν ήταν κερδοφόρα. Αυτό όντως προκύπτει από την διαδικασία μεγιστοποίησης του λογάριθμού και «εκλαϊκευμένα» μπορεί να εξηγηθεί από το γεγονός ότι ναι μεν ένα στοίχημα είναι συμφέρον να πονταριστεί με ένα μεγάλο ποσοστό αλλά από ένα σημείο και μετά το όφελος από ένα σεβαστό ποντάρισμα στην ισοπαλία ή στη ήττα (σε περίπτωση που δεν νικηθεί το στοίχημα στη νίκη) συμφέρει σε ένα βαθμό. Αυτό, διότι τα αναμενόμενα κέρδη εξ αιτίας της διαφοράς της πρόβλεψης και της απόδοσης για (την περίπτωση της νίκης της γηπεδούχου) είναι τόσα που θα «απορροφήσουν» εύκολα αυτή τη «δικλείδα ασφαλείας» που τοποθετείται στα άλλα δυο ενδεχόμενα. Έτσι, ένας μετασχηματισμός στην έξοδο, όπως αυτός περιγράφεται στην προηγούμενη παράγραφο, θα μειώσει εξίσου στοιχήματα σε πολλαπλά ενδεχόμενα σαν και αυτά που περιγράφησαν. Έτσι, δημιουργείται ένα «διαστρεβλωμένο» ποντάρισμα που δεν είναι αποτελεσματικό. Αυτό γέννησε την ανάγκη να γίνει διαφορετική παρέμβαση στο στάδιο 2 η οποία αντί να επεμβαίνει στα ποσοστά στο τέλος, επεμβαίνει στη σχέση μεταξύ πιθανοτήτων από το 1ο στάδιο και στις αποδόσεις διότι έτσι θα αλλάζει αναφανδόν η κατανομή των ποσοστών στα ενδεχόμενα ανάλογα με την ένταση της παρέμβασης στο καθαρό κριτήριο Kelly του σταδίου 2. Πιο συγκεκριμένα, δοκιμάζεται η εισαγωγή ενός παράγοντα σύγκλισης ο οποίος οδηγεί σε μετασχηματισμό των πιθανοτήτων έτσι ώστε να συγκλίνουν με τις πιθανότητες που «εκφράζονται» από τις αποδόσεις. Τα αποτελέσματα από αυτή τη παρέμβαση ήταν καθοριστικά και αυτή αποτελεί μια «μικρή» καινοτομία που εισάγεται σε αυτήν την εργασία σε σχέση με άλλες.

Παρατηρήθηκε λοιπόν πως η παρέμβαση στην είσοδο του Kelly δίνει πολύ καλύτερα αποτελέσματα σε σχέση με αυτά που προκύπτουν με τη παρέμβαση στην έξοδο. Μάλιστα, φάνηκε πως η καλύτερη λύση περιλάμβανε παρέμβαση αποκλειστικά στην είσοδο και καθόλου στην έξοδο. Έχοντας αυτό κατά νου, οι επιλογές παραμένουν βέβαια πολλές. Μια απλή λύση αρχικά είναι η αναλογική μείωση της απόστασης των προβλέψεων από τις πιθανότητες που εκφράζουν οι αποδόσεις. Αυτό μπορεί να εξελιχθεί έπειτα σε μια παραμετρική συνάρτηση «ράμπας» όπως αναφέρθηκε προηγουμένως κλπ. Εν τέλει, εξαιτίας κυρίως των περιορισμών

που θέτει η μελέτη μιας εξάμηνης εργασίας και του γεγονότος ότι οι προσομοιώσεις που γίνονται είναι πολύ χρονοβόρες υπολογιστικά, η παρούσα εργασία θα παρουσιάσει μονάχα τη χρήση ενός συντελεστή αναλογικής σύγκλισης των προβλέψεων και των αποδόσεων. Εξάλλου, οποιαδήποτε άλλη λύση ουσιαστικά θα διαφοροποιούταν στον τρόπο που διαστρεβλώνει το βέλτιστο σημείο που προκύπτει από την αντικειμενική συνάρτηση. Η άποψη του συγγραφέα μέχρι στιγμής είναι ότι αυτό θα είχε πολύ μεγαλύτερο ενδιαφέρον να γίνει τροποποιώντας την αντικειμενική συνάρτηση και τούτο αποτελεί από μόνο του ένα θέμα ικανό να θεμελιώσει μια ολόκληρη εργασία (και δεν θα μελετηθεί εδώ). Η λύση του γραμμικού συντελεστή σύγκλισης τελικά δίνει πολύ ικανοποιητικά αποτελέσματα.

Έτσι, για την υλοποίηση που ακολουθείται στην εργασία μετασχηματίζονται οι προβλέψεις στην είσοδο του Kelly χωρίς να μεταβάλλονται έπειτα τα ποσά. Ορίζεται ένα παράγοντας $C \in [0,1]$ τέτοιος έτσι ώστε να αφαιρείται από τις προβλέψεις ένα μερίδιο της διαφοράς τους με τα ανάστροφο των αποδόσεων ανάλογα με την τιμή του. Έτσι, σχηματικά προκύπτει:




Φαίνεται πως $C = 0$ ισοδυναμεί με μη χρήση των προβλέψεων του 1ου σταδίου και $C = 1$ ισοδυναμεί με χρήση αυτούσιων των προβλέψεων. Όσο πιο κοντά το $C$ προσεγγίζει το 0, το ποντάρισμα γίνεται πιο «συντηρητικό».

Αξίζει να σημειωθεί το ακόλουθο που θα γίνει πολύ σαφές στο επόμενο κεφάλαιο. Ο αναγνώστης θα μπορούσε να σκεφτεί πως $C = 0$ ισοδυναμεί με μηδενικό ποντάρισμα διότι δεν αποτελούν κανένα από τα τρία στοιχήματα μεμονωμένα «συμφέροντα» προς τοποθέτηση. Αυτό δεν ισχύει όμως στην περίπτωση που προκύπτει αρνητική γκανιότα (βλ. θεωρία) η οποία μπορεί να προκύψει εάν συνδυάζονται αποδόσεις από διαφορετικές εταιρίες με τον κατάλληλο τρόπο. Σε αυτή τη περίπτωση, υπάρχει τρόπος να τοποθετηθούν στοιχήματα που οδηγούν σε υψηλή κερδοφορία και αυτός μπορεί να προσδιοριστεί από το κριτήριο Kelly. Στον χώρο του καθημερινού στοιχήματος, η τεχνική του ποντάρίσματος σε αγώνες που προκύπτει αρνητική γκανιότα αποκαλείται «arbitrage» και οι εταιρίες καταβάλλουν μεγάλες προσπάθειες να μην προκύπτουν ευκαιρίες για να συμβαίνει αυτό συνδυάζοντας τις αποδόσεις που προσφέρουν.

## *3.4 Σύνοψη*

Έχοντας διασαφηνίσει τα ειδικά σημεία της δομής που παρουσιάζεται στην εργασία θα αναφερθεί τώρα πως αυτά σημεία θα «αλληλοεπιδρούν» οδηγώντας στη τελική υλοποίηση. Κατά την υλοποίηση της δομής γίνεται η βελτιστοποίηση του ταξινομητή στο 1ο στάδιο και ο προσδιορισμός του $C$ στο 2ο. Αυτά δεν είναι δυνατό να θεωρηθεί πως είναι ανεξάρτητα. Πιο αποτελεσματικά μοντέλα μηχανικής μάθησης θα δίνουν πιο αντιπροσωπευτικές προβλέψεις και θα επιτρέπουν σε μεγαλύτερο $C$ οδηγώντας συνήθως σε υψηλότερη κερδοφορία και το αντίθετο. Επίσης, οι μεταβλητές των μοντέλων του 1ου σταδίου (παράμετροι μοντέλων και δεδομένα εισόδου) δεν είναι δυνατό να αξιολογηθούν σωστά εάν δεν εξεταστεί η κερδοφορία συνολικά η οποία είναι συνάρτηση και του παράγοντα $C$. Έτσι, χρειάζεται να αντιμετωπιστεί «συνολικά» το ζήτημα της εύρεσης των βέλτιστων μεταβλητών και για τα δύο στάδια.

Η πληρέστερη τακτική για τον προσδιορισμό των μεταβλητών της κατασκευής συνολικά θα ήτο μια αναζήτηση πλέγματος (grid search) λαμβάνοντας όλους τους πιθανούς συνδυασμούς των δεδομένων εκπαίδευσης, της τιμής $C$, και των παραμέτρων των μοντέλων Μηχανικής Μάθησης και αξιολογώντας τους με τη τεχνική του cross-validation (βλ. θεωρία) επάνω στα δεδομένα εκπαίδευσης. Αυτό χρονικά είναι ανεπίτρεπτο για τα δεδομένα τις παρούσας εργασίας οπότε θα χρειαστεί να γίνει ένας συμβιβασμός που θα οδηγήσει σε μια πολύ καλή λύση αλλά δεν είναι πολύ χρονοβόρα η διαδικασία.

Εδώ αξίζει να σημειωθεί πως θα γίνουν δοκιμές με 4 μοντέλα ταξινομητών σε 4 διαφορετικά πρωταθλήματα. Κάθε πρωτάθλημα θα προσδιορίζει και τη «δική του» δομή που είναι βέλτιστή. Οι ταξινομητές που θα χρησιμοποιηθούν είναι οι k-NN, XGBoost, Multinomial Logistic Regression και Quadratic Discriminant Analysis. Καθένας από αυτούς παρουσιάζει ιδιαιτερότητες και στα πλαίσια της εργασίας διαμορφώνουμε τον ακόλουθο πίνακα με αυτές που μας αφορούν.

| | **«Αυτόματη» διαλογή σημαντικών features.** | **Έχει παραμέτρους εκπαίδευσης μοντέλου.** | **Χρονοβόρα Εκπαίδευση** |
|---|---|---|---|
| **XGBoost** | ΝΑΙ | ΝΑΙ | ΠΟΛΥ |
| **LR** | ΟΧΙ | ΟΧΙ | ΛΙΓΟ |
| **QDA** | ΟΧΙ | ΟΧΙ | ΛΙΓΟ |

Ο πίνακας βοηθάει έτσι ώστε να «σπάσει» η τακτική που θα ακολουθηθεί σε υποπεριπτώσεις. Αρχικά θα γίνει αναφορά στην υποπερίπτωση που αφορά τα μοντέλα LR και QDA όπου θα γίνει και η πιο εκτεταμένη αναζήτηση διότι το επιτρέπει ο μικρός χρόνος εκπαίδευσης

μεμονωμένα. Εδώ, οι παράμετροι της δομής είναι ο παράγοντας $C$ και τα δεδομένα εισόδου ελλείψει παραμέτρων εκπαίδευσης των μοντέλων. Θα ακολουθηθούν τα εξής βήματα

1) Ξεκινάμε την αναζήτηση θέτοντας μια συντηρητική τιμή για το $C$ ίση με 0.1. Κατά τη πειραματική διαδικασία απεδείχθη πως συνήθως οδηγεί σε πονταρίσματα που δεν δίνουν πολύ υψηλά κέρδη αλλά ταυτόχρονα δεν υπάρχει ιδιαίτερα ασταθής πορεία του κεφαλαίου. Αναδεικνύει σταθερά μια κατεύθυνση προς τη βελτιστοποίηση των άλλων παραμέτρων. Έτσι, ξεκινάμε με το σύνολο των features και εκτελούμε την ακόλουθη διαδικασία για να αφαιρέσουμε όσα επιδεινώνουν την απόδοση.
    a. Αφαιρούμε ένα feature (κάθε φορά) από το σύνολο και εξετάζουμε την επίδοση του μοντέλου στα δεδομένα εκπαίδευσης με τη τεχνική του cross-validation.
    b. Αποφασίζουμε να αφαιρέσουμε οριστικά από το σύνολο το feature του οποίου η απουσία βελτιώνει τα μέγιστα την κερδοφορία. Εάν με καμία αφαίρεση δεν βελτιώνεται η επίδοση τότε σταματάμε. Διαφορετικά, επαναλαμβάνουμε από το βήμα 'a'.
2) Χρησιμοποιώντας τα feature που προέκυψαν από το 1ο βήμα, επιχειρούμε να προσδιορίσουμε το βέλτιστο $C$ αξιολογώντας και πάλι την κερδοφορία του μοντέλου με τη χρήση της τεχνικής του cross-validation επάνω στα δεδομένα εκπαίδευσης. Επαναλαμβάνουμε από το βήμα 1 χρησιμοποιώντας το $C$ που προσδιορίστηκε αλλά χρησιμοποιώντας το αρχικό σύνολο των features. Σταματάμε σε κάθε βήμα εάν δεν υπάρχει αύξηση στην κερδοφορία έπειτα από το πέρας του.

Πρέπει να σημειωθεί ότι ο τρόπος που αξιολογείται η κερδοφορία είναι ο εξής. Σε κάθε δοκιμή του cross-validation με τους διάφορους συνδυασμούς των δεδομένων λαμβάνεται ως score ο λογάριθμος του τελικού κεφαλαίου (σε κάθε δοκιμή το αρχικό κεφάλαιο είναι σταθερό). Ο λογάριθμος είναι αντιπροσωπευτικός της επίδοσης ενός μεγέθους που αυξάνεται εκθετικά ενώ το απόλυτο ποσό δεν είναι. Τα scores από τις δοκιμές του cross-validation δίνουν μια μέση τιμή $\mu$ και μια τυπική απόκλιση $\sigma$. Η παράσταση $2\mu - \sigma$ δίνει το τελικό σκορ (το οποίο χρησιμοποιείται στην αξιολόγηση των βημάτων παραπάνω). Έτσι, ναι μεν κύριο λόγο έχει η μέση επίδοση αλλά συνυπολογίζεται σε σημαντικό βαθμό και η σταθερότητα των αποτελεσμάτων.

Σε ό,τι αφορά τη χρήση του μοντέλου XGBoost όπου οι χρόνοι εκπαίδευσης είναι πολλαπλάσιοι εξαιτίας εν μέρει της πληθώρας παραμέτρων εκπαίδευσης θα ακολουθηθεί η επόμενη λιγότερο εκτεταμένη στρατηγική που εκμεταλλεύεται το γεγονός ότι δεν είναι απαραίτητη η διαλογή δεδομένων.

- Ξεκινάμε την αναζήτηση εκπαιδεύοντας μοντέλα για να προσδιοριστεί η τιμή του $C$ μεταξύ 0.0 και 0.8 και των παραμέτρων εκπαίδευσης του ταξινομητή. Σταματάμε εκεί που μεγιστοποιείται η τιμή της παράστασης $2\mu - \sigma$ όπως προηγουμένως.

# 4 Πειραματική Διαδικασία

## 4.1 Εισαγωγή

Αυτό το κεφάλαιο αποτελεί το 3ο και τελευταίο βήμα στην πορεία από τη θεωρία στη πράξη. Η αρχή έγινε αναλύοντας την αναγκαία θεωρία, στη συνέχεια «πατώντας» επάνω της χτίστηκε μια γενικευμένη δομή με παραμετρικά χαρακτηριστικά στην οποία έγιναν κάποιες παραδοχές για εξοικονόμηση σημαντικού χρόνοι και τέλος τώρα θα παρουσιαστούν οι λεπτομέρειες της και τα αποτελέσματα που λαμβάνουμε.

## 4.2 Διαμόρφωση τελικών δεδομένων εισόδου και συνδυασμών παραμέτρων

Όπως αναφέρθηκε και στο προηγούμενο κεφάλαιο χρησιμοποιούνται τα δημοσίως διαθέσιμα δεδομένα από τη σελίδα football-data.co.uk και fifaindex.com για την Premier League (PL), La Liga (LL), Bundesliga (BL) και Serie A (SA). Ο γενικός τρόπος σκέψης κατά τη δημιουργία των δεδομένων είναι ότι για κάθε αγώνα-δείγμα δημιουργούνται χαρακτηριστικά που αφορούν την επίδοση της κάθε ομάδας στο παρελθόν. Δεν αντιμετωπίζουμε τους αγώνες ως αναμετρήσεις μεταξύ επώνυμων ομάδων αλλά ως events αναμέτρησης μιας γηπεδούχου και μιας φιλοξενούμενης δεδομένων κάποιων επιδόσεων. Τα χαρακτηριστικά που αναφέρθηκαν προκύπτουν από δημοσίως διαθέσιμα στατιστικά δεδομένα για παλαιούς αγώνες (της σελίδας football-data.co.uk) και από ιστορικές κατατάξεις των ομάδων (της σελίδα fifaindex.com). Έχοντας ανατρέξει σε αρκετή βιβλιογραφία [4] [3] [5] [15], έγινε η επιλογή να ακολουθηθεί γενικά η μεθοδολογία της εργασίας [4] διότι κρίθηκε ως η πληρέστερη.

Πιο συγκεκριμένα, τα δεδομένα αγώνων από το football-data.co.uk περιέχουν γραμμές δείγματα της μορφής:

| Date | HomeTeam | AwayTeam | FTHG | FTAG | HC | AC | HST | AST | HC | AC | MaxH |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **9/8/19** | Liverpool | Norwich | 4 | 1 | 11 | 2 | 7 | 5 | 11 | 2 | 1.16 |

| MaxD | MaxA | AvgH | AvgD | AvgA | FTR | … |
|---|---|---|---|---|---|---|
| **10** | 23 | 1.14 | 8.75 | 19.83 | H | … |

Όπου FTHG: Full Time Home Goals, FTAG: Full Time Away Goals, HC: Home Corners, AC: Away Corners, HST: Home Shots on Target, AST: Away Shots on Target, MaxH/D/A: Maximum Home/Draw/Away Odds, AvgH/D/A: Average Home/Draw/Away Odds.

Επίσης, τα δεδομένα από το fifaindex είναι της ακόλουθης μορφής για κάθε σεζόν:

| Teams | Attack | Middle | Defense | Overall (Average) |
|---|---|---|---|---|
| **Man United** | 86 | 81 | 82 | 82 |
| **Man City** | 85 | 83 | 80 | 82 |
| **Chelsea** | 82 | 80 | 81 | 81 |
| **…** | … | … | … | … |

Με τα scores “Attack”, “Middle”, “Defence” να είναι ανάλογα των επιδόσεων μιας ομάδας στην επίθεση (όπως αυτές εκτιμώνται για το ετήσιο βιντεοπαιχνίδι «FIFA») , στο κέντρο και στην άμυνα ενώ το “Overall” αποτελεί τον μέσο όρο των προηγουμένων.

Η εργασία [4] χρησιμοποιεί αυτά τα δεδομένα για να εξάγει νέα χαρακτηριστικά που τα χωρίζει σε δύο μεγάλες κατηγορίες. Η πρώτη περιέχει χαρακτηριστικά που αφορούν μονάχα τη γηπεδούχο ή τη φιλοξενούμενη (π.χ. τα συνολικά γκολ στη σεζόν που έχουν επιτευχθεί για κάθε μια ομάδα) και η δεύτερη αποτελείται από τις διαφορές μεταξύ των ίδιων χαρακτηριστικών μεταξύ γηπεδούχου και φιλοξενούμενης. Αρχικά, επιλέγουμε να χρησιμοποιήσουμε αποκλειστικά τα δεδομένα της 2ης ομάδας στα μοντέλα διότι αυτά της 1ης ομολογουμένως από μόνα τους δεν είναι σε μορφή που μπορεί να αποτυπωθεί έντονα η εικόνα της διαφοράς μεταξύ των ομάδων. Επίσης, η χρήση ταυτοχρόνως και των δύο μάλλον δεν θα προσέφερε ουσιαστική νέα πληροφορία που θα ήτο καθοριστική ενώ παράλληλα θα αύξανε δραματικά τους χρόνους εκπαίδευσης και θα επιδείνωνε την επίδοση σε ορισμένα μοντέλα. Αυτά που προκύπτουν από αυτή τη πρώτη διαλογή είναι αυτά:

| ΚΩΔΙΚΗ ΟΝΟΜΑΣΙΑ | ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ |
|---|---|

| | |
|---|---|
| **FormDifferential** | Διαφορά «φόρμας» μεταξύ ομάδων. Η «φόρμα» (Form) ορίζεται στην εργασία και αποτελεί μια μετρική που εκφράζει τη δυναμική της ομάδας επιβραβεύοντας τις προηγούμενες νίκες απέναντι σε ικανές ομάδες αυξάνοντας (ή μειώνοντας σε ήττες) τη φόρμα. (βλ. [4]) |
| **StreakDifferential** | Εκφράζει της διαφορά μεταξύ συνεχόμενων επιτυχιών της ομάδας ανεξαρτήτως δυσκολίας. (βλ. [4]) |
| **STKPP** | Εκφράζει τη διαφορά των περασμένων σουτ στον στόχο κατά τη διάρκεια των 'k' περασμένων αγώνων. |
| **GKPP** | Εκφράζει τη διαφορά των περασμένων γκολ κατά τη διάρκεια των 'k' περασμένων αγώνων. |
| **CKPP** | Εκφράζει τη διαφορά των περασμένων κόρνερς κατά τη διάρκεια των 'k' περασμένων αγώνων. |
| **RelAttack** | Διαφορά μεταξύ του score των ομάδων στην επίθεση από τη κατάταξη του fifaindex.com. |
| **RelMidField** | Διαφορά μεταξύ του score των ομάδων στο κέντρο από τη κατάταξη του fifaindex.com. |

| | |
|---|---|
| **RelDefense** | Διαφορά μεταξύ του score των ομάδων στην άμυνα από τη κατάταξη του fifaindex.com. |
| **RelOverall** | Διαφορά μεταξύ του score των ομάδων στην άμυνα από τη κατάταξη του fifaindex.com. |
| **GDDifferential** | Διαφορά μεταξύ των αθροισμάτων του των διαφορών των γκολ των προηγούμενων αγώνων κάθε ομάδας |
| **StWeightedDifferential** | Όπως το «StreakDifferential» αλλά μετρούν περισσότερο οι τελευταίοι αγώνες (βλ. [4]) |
| **ODDSH12X** | Μέση απόδοση μεταξύ των στοιχηματικών εταιρειών για τον ενδεχόμενο νίκης της γηπεδούχου. |
| **ODDSD12X** | Μέση απόδοση μεταξύ των στοιχηματικών εταιρειών για τον ενδεχόμενο ισοπαλίας. |
| **ODDSA12X** | Μέση απόδοση μεταξύ των στοιχηματικών εταιρειών για τον ενδεχόμενο νίκης της φιλοξενούμενης. |

Εδώ πρέπει να αναφερθεί πως τα χαρακτηριστικά ODDSH12X, ODDSD12X και ODDSA12X δεν χρησιμοποιήθηκαν στην εργασία [4] αλλά τοποθετήθηκαν έχοντας ως έμπνευση άλλες εργασίες [3].

Έχοντας δημιουργήσει αυτά τα features για κάθε δείγμα του συνόλου δεδομένων, περαιτέρω διαλογή θα εξαρτάται πλέον από το κάθε μοντέλο που θα δοκιμασθεί και έγινε όπως αναλύθηκε στο προηγούμενο κεφάλαιο μαζί με τον παράγοντα σύγκλισης $C$.

Έτσι καταλήγουμε στους επόμενους πίνακες που παρουσιάζουν τους τελικούς βέλτιστους συνδυασμούς πρωταθλημάτων, μοντέλων, υποσυνόλων δεδομένων εισόδου και

παραγόντων $C$ που προέκυψαν κατά τη διαδικασία της εκπαίδευσης. Η εκπαίδευση σε όλες τις περιπτώσεις έγινε με τη χρήση δεδομένων που αφορούν τα έτη μεταξύ 2007 και 2017

**Αγγλικό Πρωτάθλημα: Premier League**

| | FEATURES | C-factor |
|---|---|---|
| M-LR | **STDIFF, RELMID, GDDIFF, WSTDIFF, ODDSH12X, ODDSD12X** | **0.47** |
| QDA | **STKPP, STDIFF, CKPP, RELDEF, RELOALL, FDIFF** | **0.19** |

| | PARAMETERS | C-factor |
|---|---|---|
| XGBoost | **colsample_bytree = 0.1**<br>**gamma = 0. 27296351513034095**<br>**learning_rate = 0.46117292680572036**<br>**max_depth = 200**<br>**n_estimators = 460**<br>**reg_alpha = 10.0**<br>**reg_lambda = 2.0816956991712323**<br>**subsample = 0.1** | **0.10** |

**Ισπανικό Πρωτάθλημα: La Liga**

| | FEATURES | C-factor |
|---|---|---|
| M-LR | **CKPP, RELATT, RELMID, RELDEF, RELOALL, GDDIFF, WSTDIFF, ODDSD12X** | **0.19** |
| QDA | **STDIFF, CKPP, RELATT, RELDEF, WSTDIFF** | **0.17** |

| | PARAMETERS | C-factor |
|---|---|---|
| XGBoost | **colsample_bytree = 0.9428840500922321**<br>**gamma = 0.08947222410809427**<br>**learning_rate = 0.34317689475042074**<br>**max_depth = 188**<br>**n_estimators = 500**<br>**reg_alpha = 10.0** | **0.13** |

| | | |
|---|---|---|
| | **reg_lambda = 10**<br>**subsample = 1.0** | |

**<u>Γερμανικό Πρωτάθλημα: Bundesliga</u>**

| | FEATURES | C-factor |
|---|---|---|
| M-LR | **STKPP, STDIFF, CKPP, RELDEF, GDDIFF, FDIFF, ODDSH12X, ODDSA12X** | **0.35** |
| QDA | **GKPP, RELATT, RELMID, RELDEF, RELOALL, WSTDIFF, ODDSD12X** | **0.15** |

| | PARAMETERS | C-factor |
|---|---|---|
| XGBoost | **colsample_bytree = 1.0**<br>**gamma = 0.335210274410828**<br>**learning_rate = 0.5**<br>**max_depth = 200**<br>**n_estimators = 277**<br>**reg_alpha = 6.717021481922724**<br>**reg_lambda = 5.389337239233785**<br>**subsample = 0.1** | **0.13** |

**<u>Ιταλικό Πρωτάθλημα: Serie A</u>**

| | FEATURES | C-factor |
|---|---|---|
| M-LR | **STDIFF, RELMID, GDDIFF, ODDSA12X** | **0.33** |
| QDA | **STKPP', GKPP, RELMID, RELOALL, GDDIFF, FDIFF, ODDSH12X, ODDSD12X, ODDSA12X** | **0.12** |

| | PARAMETERS | C-factor |
|---|---|---|
| XGBoost | **colsample_bytree = 0.19973887868917634**<br>**gamma = 0.09448721514307101**<br>**learning_rate = 0.24842991795274494**<br>**max_depth = 129** | **0.07** |

| | | |
|---|---|---|
| | **n_estimators = 146**<br>**reg_alpha = 1e-09**<br>**reg_lambda = 1e-09**<br>**subsample = 0.7496990428599569** | |

## *4.3 Επίδοση των μοντέλων στα δεδομένα δοκιμής.*

Εδώ θα παρουσιασθούν τα τελικά αποτελέσματα των δοκιμών των συνδυασμών μοντέλων και πρωταθλημάτων ποδοσφαίρου. Θα παρουσιασθεί η πορεία του κεφαλαίου γραφικά με την πάροδο των παιχνιδιών που αφορούν τις σεζόν μεταξύ 2017 και 2020.

Κατά τη διάρκεια των δοκιμών έγινε η διαπίστωση πως για μεγάλο μερίδιο των παιχνιδιών, η δομή του σταδίου 2 (βλ. κεφάλαιο 3) πρότεινε ποσά προς στοιχηματισμό που πετύχαιναν πολύ μεγάλη ανάπτυξη του κεφαλαίου κάνοντας arbitrage (βλ. θεωρία) αν επιλεγόταν $C = 0$ (δηλαδή εάν δεν χρησιμοποιούντο οι προβλέψεις στα δεδομένα εισόδου του 2ου σταδίου). Μαζί με τα αποτελέσματα λοιπόν για το βέλτιστο $C$, θα παρουσιαστούν και αυτά για $C = 0$ που οφείλονται σε arbitrage. Η σύγκριση των δύο αυτών ειδών αποτελεσμάτων δίνει μια ακόμα αντιπροσωπευτικότερη εικόνα της επίδοσης των μοντέλων πρόβλεψης.

Τέλος, μαζί με την παρουσίαση της πορείας του κεφαλαίου γραφικά θα υπολογιστεί και ο μέσος ρυθμός εκθετικής ανάπτυξης του κεφαλαίου ο οποίος δίνεται από τον τύπο:

$$G = \frac{1}{total_events}\log_{10}\left(\frac{endcapital}{startcapital}\right)$$

Χρήση άλλων μετρικών όπως το ROI (ή το παρεμφερές με αυτό Yield) δεν θα γίνει καθώς δεν ήταν αντιπροσωπευτικοί της επίδοσης του συστήματος για τους εξής λόγους:

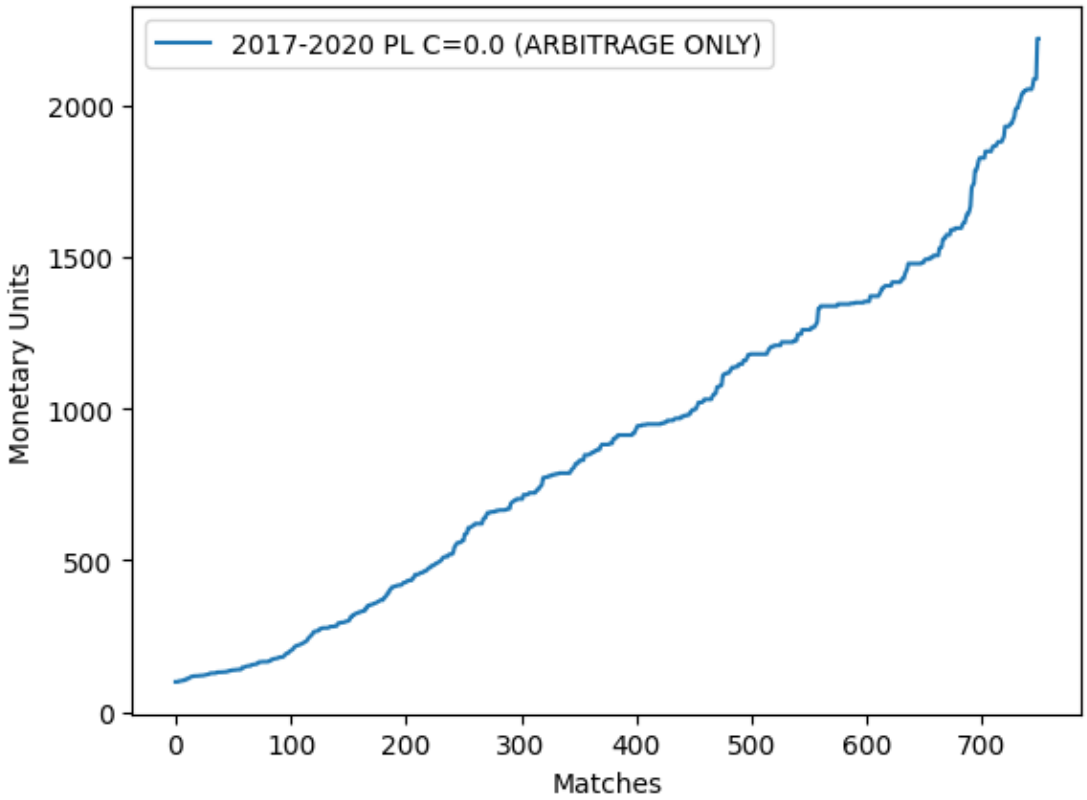
- Αν θεωρήσουμε ολόκληρη την στοιχηματική πορεία σαν μια ενιαία επένδυση (το οποίο μάλλον είναι και το πιο σωστό όπως θα εξηγηθεί μετά), τότε το ROI που φυσιολογικά δίνεται από την παράσταση $\frac{\mu\epsilon\iota\kappa\tau\acute{\alpha}\ \acute{\epsilon}\sigma o\delta\alpha}{\sigma\upsilon\nu o\lambda\iota\kappa\acute{o}\ \pi o\nu\tau\acute{\alpha}\rho\iota\sigma\mu\alpha} * 100\%$ θα ήταν ίσο με τον λόγο του τελικού κεφαλαίου προς το αρχικό κεφάλαιο. Όπως θα φανεί στην πορεία, υπάρχουν περιπτώσεις π.χ. όπου το τελικό κεφάλαιο είναι 15 φορές μεγαλύτερο του αρχικού ενώ σε άλλες μπορεί να είναι (μόνο) 10πλάσιο. Μπορεί σε απόλυτα νούμερα να είναι τεράστια η διαφορά με το ROI να αποτυπώνει ακριβώς αυτό και κάποιος αξιολογητής να εξάγει αυτό ως συμπέρασμα, αλλά σε βάθος χρόνου π.χ. 700 αγώνων (ο οποίος είναι περίπου ο αριθμός των αγώνων στην Premier League) δεκαπλασιασμός του κεφαλαίου προκύπτει από $G \cong 0.003289$

και 15πλασιασιμός από $G \cong 0.003869$. Εδώ η διαφορά δεν φαίνεται τόσο μεγάλη και δεδομένου πως το σύστημα επιχειρεί να μεγιστοποιήσει ακριβώς αυτό το μέγεθος, όντως δεν είναι τόσο διαφορετική η επίδοση. Και είναι πράγματι εύκολο να καταλάβει και διαισθητικά κάποιος πως ένα μέγεθος που αυξάνεται εκθετικά, αναπτύσσεται με τη μορφή χιονοστιβάδας και έπειτα από την πάροδο μεγάλου χρονικού διαστήματος, αλλαγές μεγάλες σε απόλυτα νούμερα είναι μικρές από τη σκοπιά της εκθετικής ανάπτυξης. Μια μικρή αλλαγή στο μέγεθος $G$ που μπορεί να οφείλεται και σε συστηματικά τυχαία γεγονότα είναι δυνατό μετά από πολύ χρόνο να αλλάξει εντελώς τα απόλυτα νούμερα.

- Αν θεωρήσουμε ο παρονομαστής της παράστασης του ROI που αναφέρεται παραπάνω είναι το άθροισμα των συνολικών πονταρισμάτων καθ' όλη τη πορεία, τότε αυτή πραγματικά θα ήταν μια λάθος ερμηνεία. Αφενός γιατί ο παίκτης εξ' αρχής «δέσμευσε» μονάχα το αρχικό του κεφάλαιο με τα ενδιάμεσα πονταρίσματα να χρησιμοποιούν αυτό μαζί με τα κέρδη του (εάν υπάρχουν), αφετέρου γιατί το ROI θα μεγιστοποιούταν χρησιμοποιώντας τη στρατηγική «all in» σε πολλά «μεμονωμένα» κερδοφόρα στοιχήματα όπως ακριβώς και η αναμενόμενη τιμή του τελικού κεφαλαίου (βλ. θεωρία). Αυτή ακριβώς η στρατηγική θα οδηγούσε με βεβαιότητα όμως σε πλήρη απώλεια κεφαλαίου με την πάροδο του χρόνου.

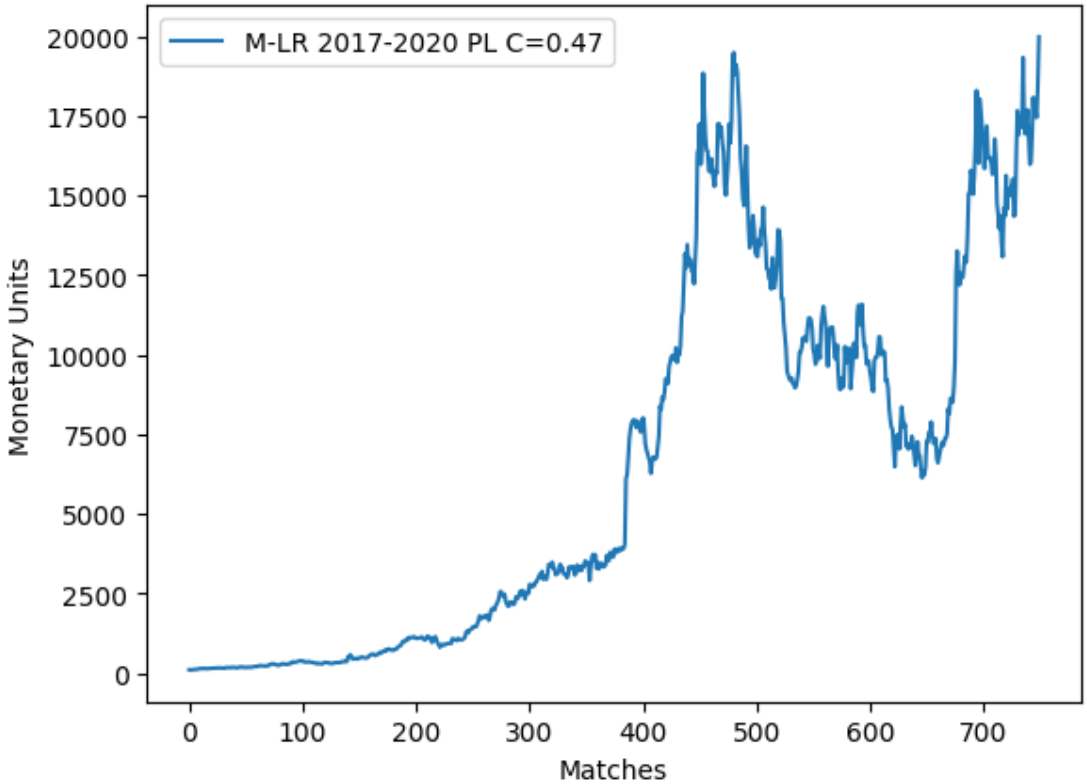
### *4.3.1 Premier League*

#### *4.3.1.1 Χωρίς προβλέψεις ταξινομητή (C=0)*



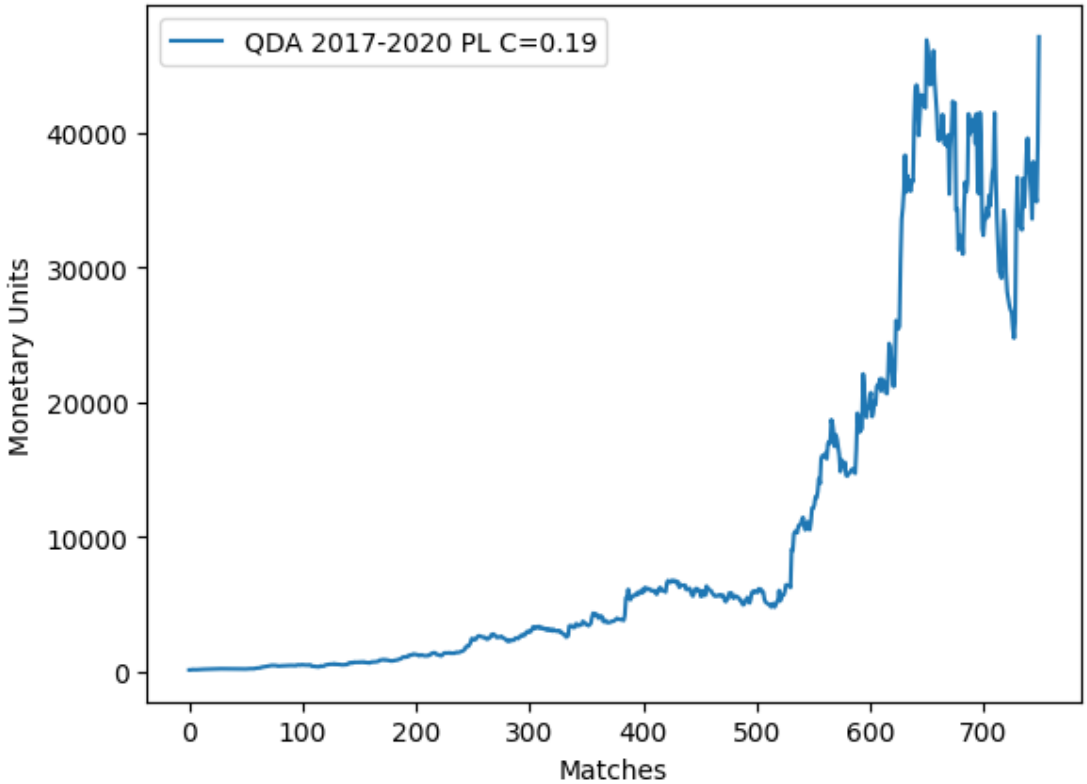

| ΑΡΧΙΚΟ ΚΕΦΑΛΑΙΟ | ΤΕΛΙΚΟ ΚΕΦΑΛΑΙΟ | G | ΟΡΘΟΤΗΤΑ ΠΡΟΒΛΕΨΕΩΝ |
|---|---|---|---|
| 100 | 2218 | $41.32 * 10^{-4}$ | 56.2% |

#### 4.3.1.2 *Multinomial Logistic Regression*



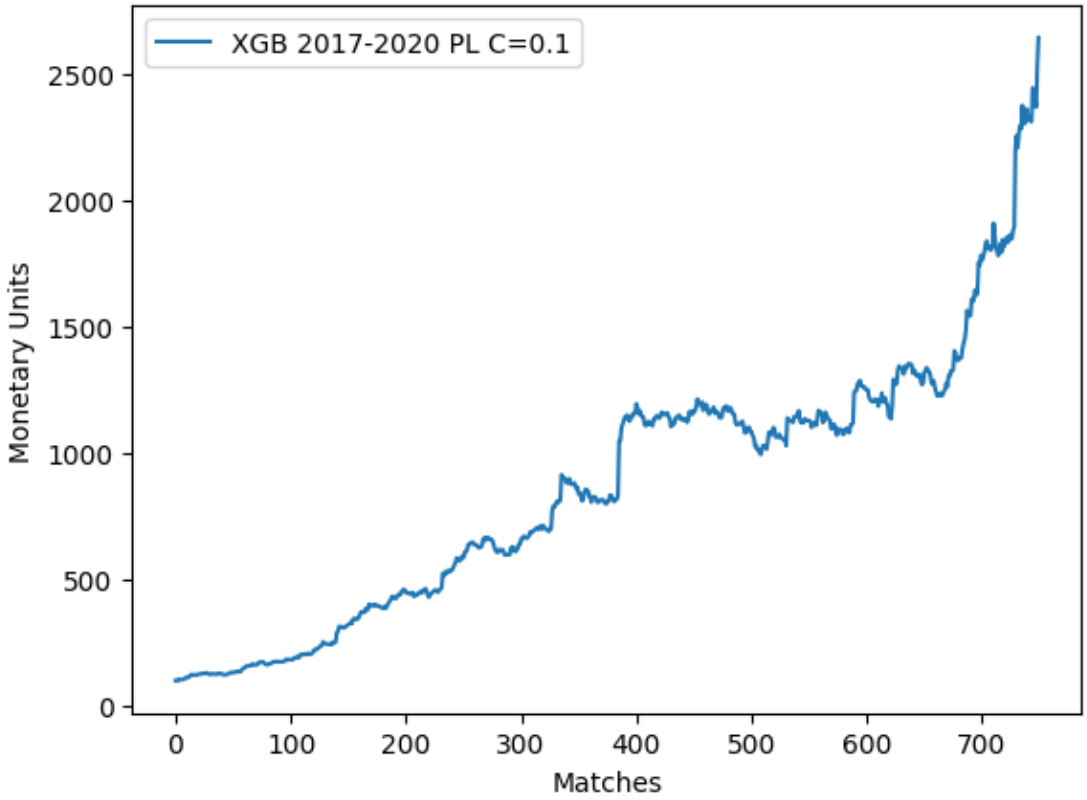

| ΑΡΧΙΚΟ ΚΕΦΑΛΑΙΟ | ΤΕΛΙΚΟ ΚΕΦΑΛΑΙΟ | G | ΟΡΘΟΤΗΤΑ ΠΡΟΒΛΕΨΕΩΝ |
|---|---|---|---|
| 100 | 19981 | $70.63 * 10^{-4}$ | 56.2% |

#### 4.3.1.3 *Quadratic Discriminant Analysis*



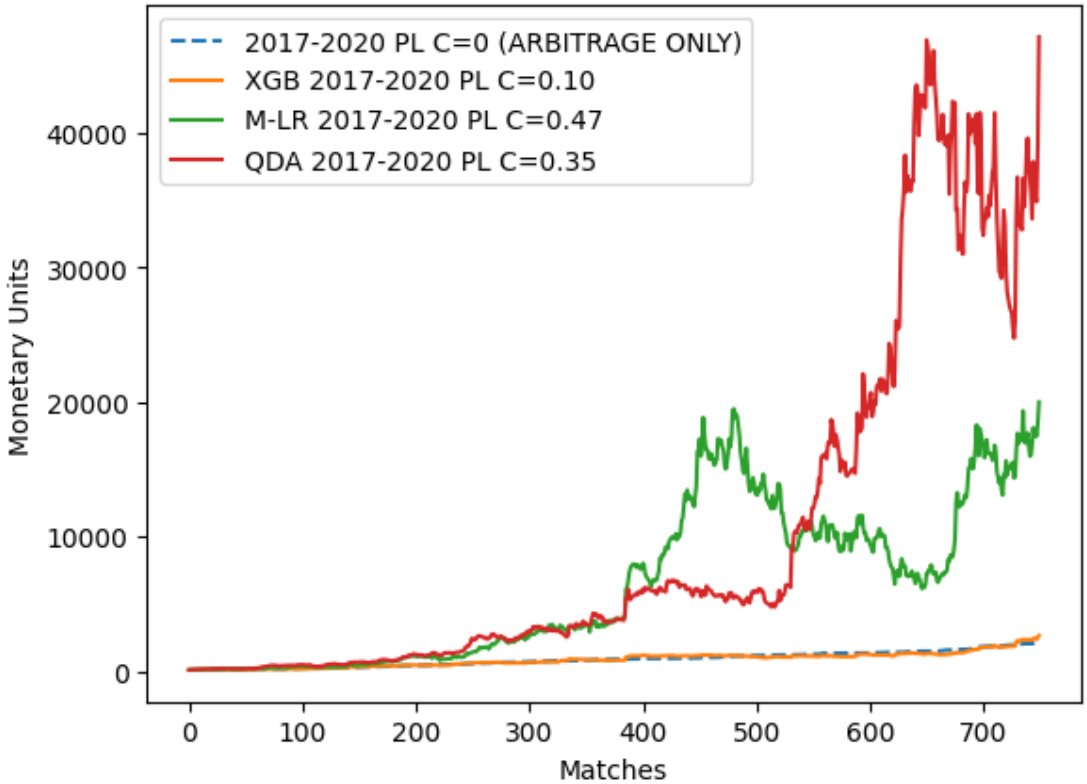

| ΑΡΧΙΚΟ ΚΕΦΑΛΑΙΟ | ΤΕΛΙΚΟ ΚΕΦΑΛΑΙΟ | G | ΟΡΘΟΤΗΤΑ ΠΡΟΒΛΕΨΕΩΝ |
|---|---|---|---|
| 100 | 47098 | $82.06 * 10^{-4}$ | 55.9% |

#### 4.3.1.4 XGBoost




| ΑΡΧΙΚΟ ΚΕΦΑΛΑΙΟ | ΤΕΛΙΚΟ ΚΕΦΑΛΑΙΟ | G | ΟΡΘΟΤΗΤΑ ΠΡΟΒΛΕΨΕΩΝ |
|---|---|---|---|
| 100 | 2643 | $43.66 * 10^{-4}$ | 55.8% |

#### 4.3.1.5 *Συγκεντρωτικά*

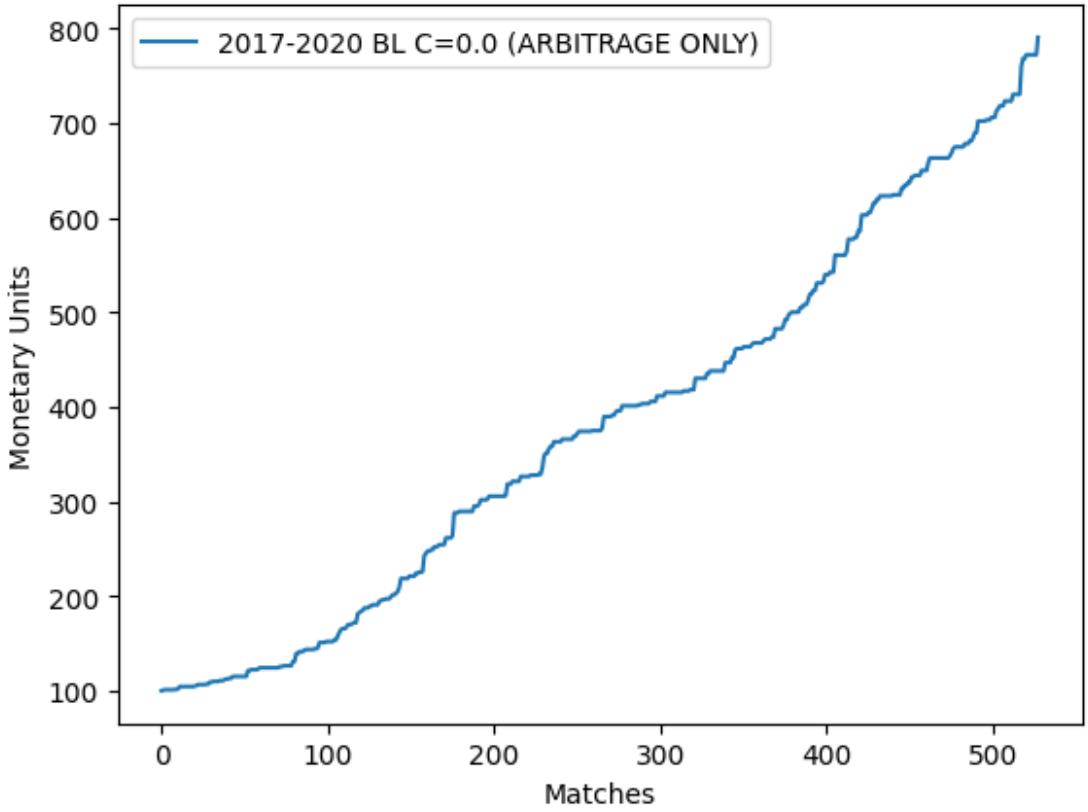



| ΤΑΞΙΝΟΜΗΤΗΣ | ΑΡΧ. ΚΕΦΑΛΑΙΟ | ΤΕΛ. ΚΕΦΑΛΑΙΟ | G | ΟΡΘΟΤΗΤΑ ΠΡΟΒΛΕΨΕΩΝ |
|---|---|---|---|---|
| ΧΩΡΙΣ | **100** | 2218 | $41.32 * 10^{-4}$ | 56.2% |
| XGBoost | | 2647 | $43.66 * 10^{-4}$ | 55.8% |
| M-LR | | 19981 | $70.63 * 10^{-4}$ | 56.2% |
| QDA | | 47098 | $82.06 * 10^{-4}$ | 55.9% |

- Στο γράφημα φαίνεται τεράστια η διαφορά της επίδοσης μεταξύ του καλύτερου ταξινομητή (QDA) και της πορείας αναφοράς χωρίς ταξινομητή. Από τη σκοπιά όμως της εκθετικής ανάπτυξης είναι μόνο διπλάσια.
- Ο XGBoost δεν προσφέρει ουσιαστικό πλεονέκτημα σε σχέση με την εφαρμογή μόνο του arbitrage.
- Η Premier League είναι το πρωτάθλημα στο οποίο η χρήση των ταξινομητών δίνει το μεγαλύτερο πλεονέκτημα. Αυτό θα φανεί στη συνέχεια σε σύγκριση με τα υπόλοιπα πρωταθλήματα.
- Η ορθότητα των προβλέψεων θα έδινε εντελώς λάθος εικόνα για την κερδοφορία.

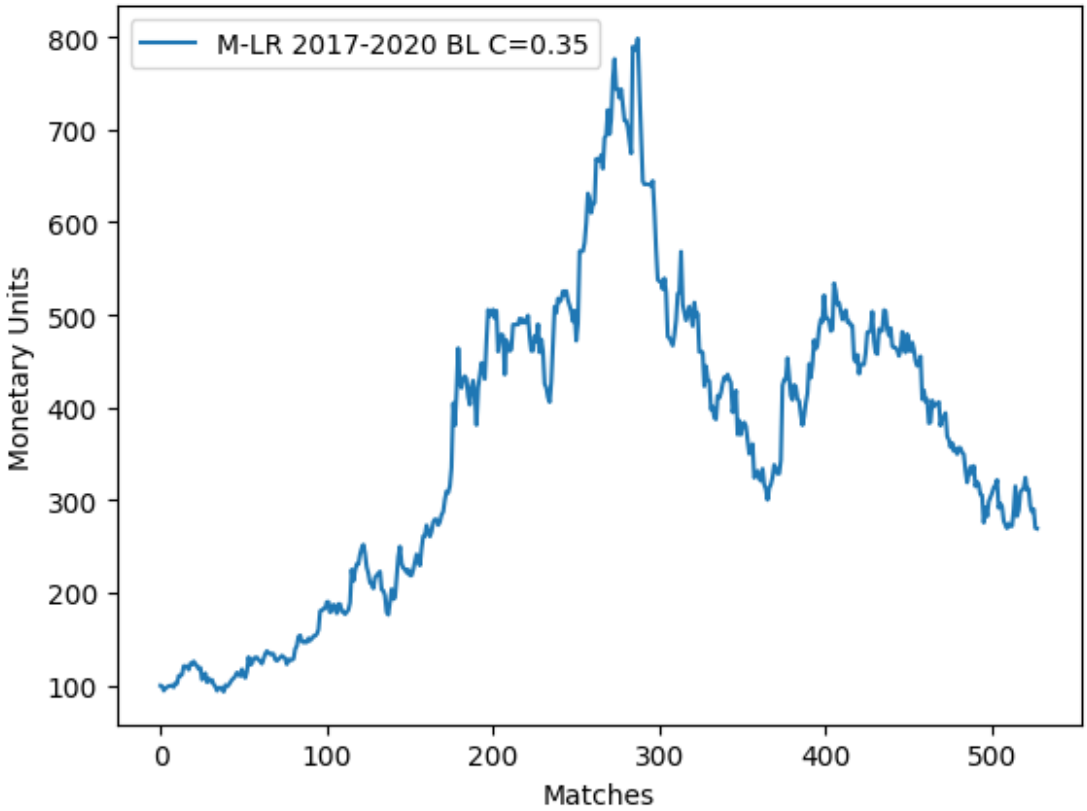
### *4.3.2 Bundesliga*

#### *4.3.2.1 Χωρίς προβλέψεις ταξινομητή*



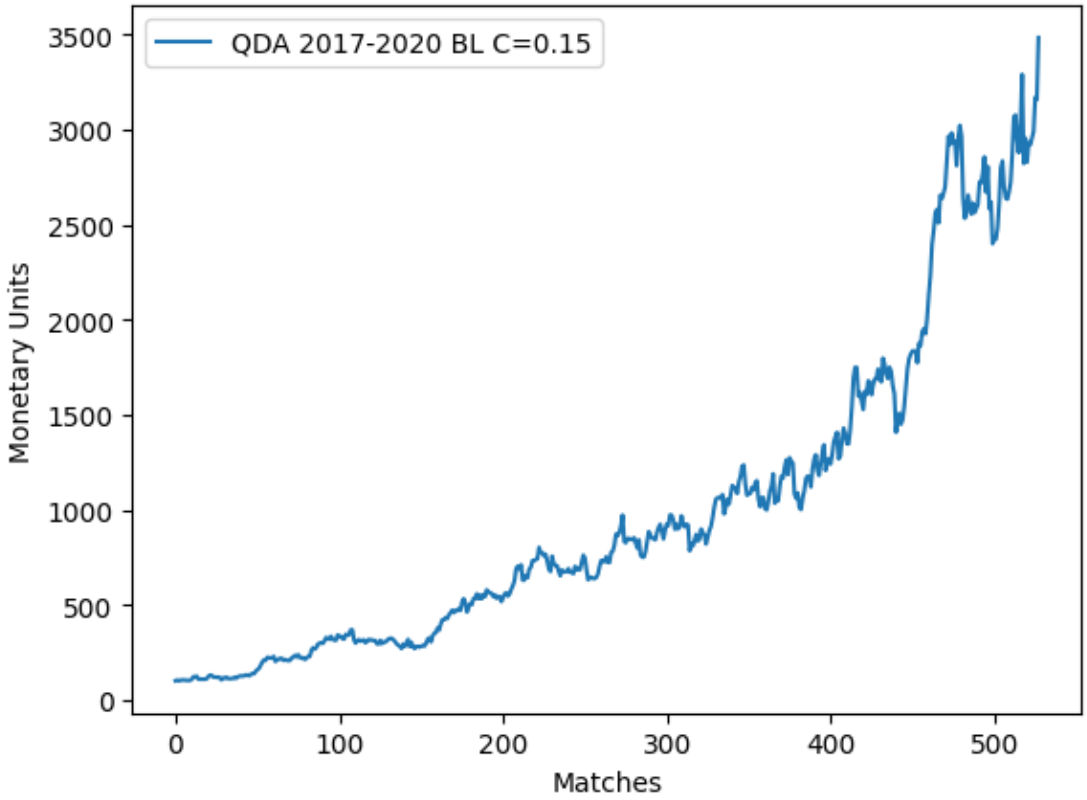

| ΑΡΧΙΚΟ ΚΕΦΑΛΑΙΟ | ΤΕΛΙΚΟ ΚΕΦΑΛΑΙΟ | G | ΟΡΘΟΤΗΤΑ ΠΡΟΒΛΕΨΕΩΝ |
|---|---|---|---|
| 100 | 790 | $39.15 * 10^{-4}$ | 53.5% |

#### 4.3.2.2 *Multinomial Logistic Regression*



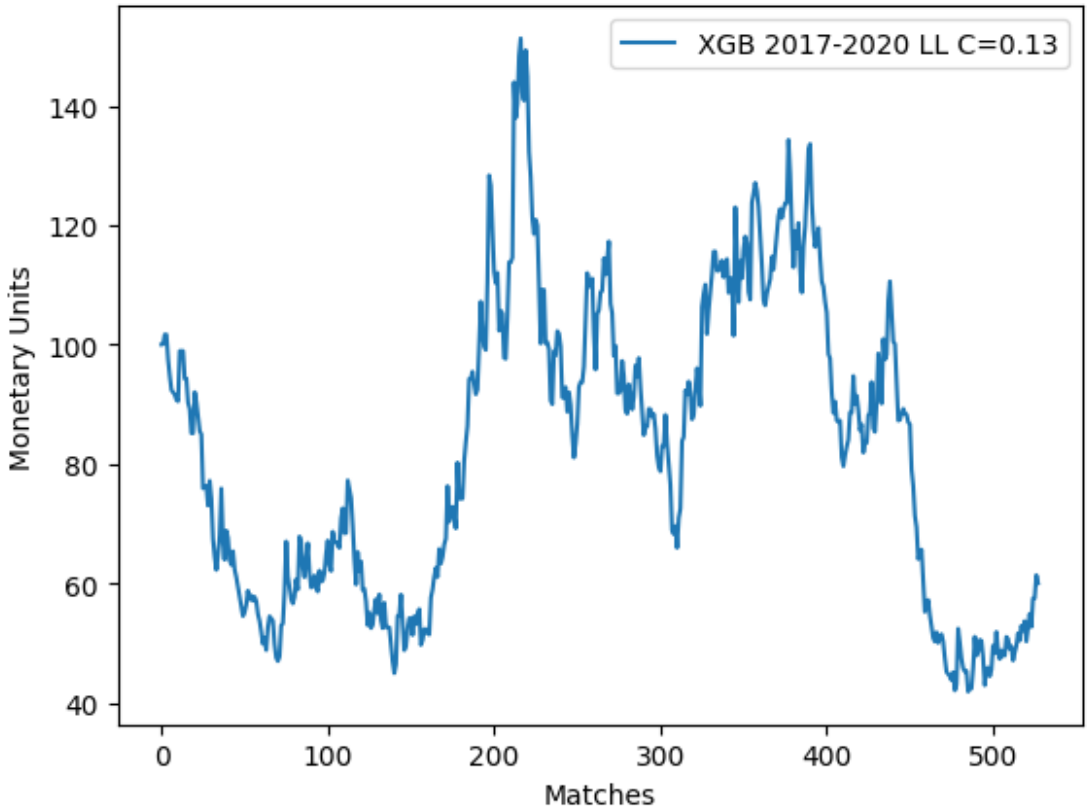

| ΑΡΧΙΚΟ ΚΕΦΑΛΑΙΟ | ΤΕΛΙΚΟ ΚΕΦΑΛΑΙΟ | G | ΟΡΘΟΤΗΤΑ ΠΡΟΒΛΕΨΕΩΝ |
|---|---|---|---|
| 100 | 269 | $18.77 * 10^{-4}$ | 52.8% |

#### 4.3.2.3 *Quadratic Discriminant Regression*



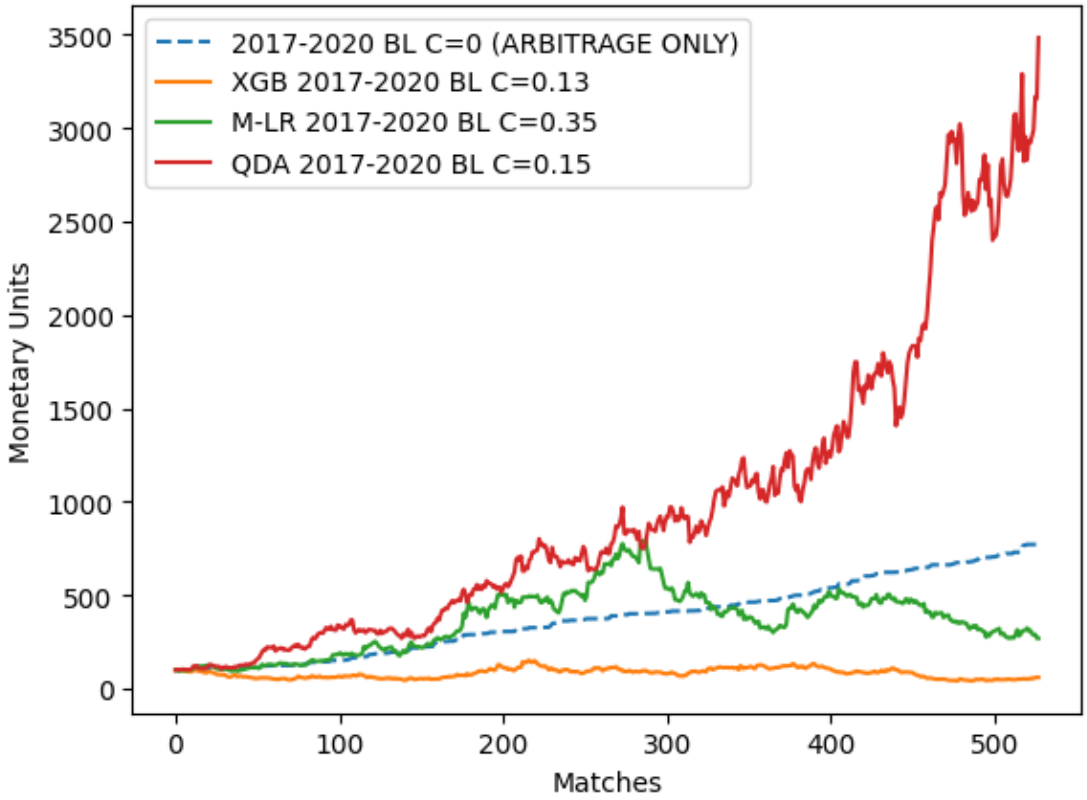

| ΑΡΧΙΚΟ ΚΕΦΑΛΑΙΟ | ΤΕΛΙΚΟ ΚΕΦΑΛΑΙΟ | G | ΟΡΘΟΤΗΤΑ ΠΡΟΒΛΕΨΕΩΝ |
|---|---|---|---|
| 100 | 3485 | $67.26 * 10^{-4}$ | 52.6% |

#### 4.3.2.4 XGBoost




| ΑΡΧΙΚΟ ΚΕΦΑΛΑΙΟ | ΤΕΛΙΚΟ ΚΕΦΑΛΑΙΟ | G | ΟΡΘΟΤΗΤΑ ΠΡΟΒΛΕΨΕΩΝ |
|---|---|---|---|
| 100 | 60 | $-9.67 * 10^{-4}$ | 45.9% |

#### 4.3.2.5 *Συγκεντρωτικά*

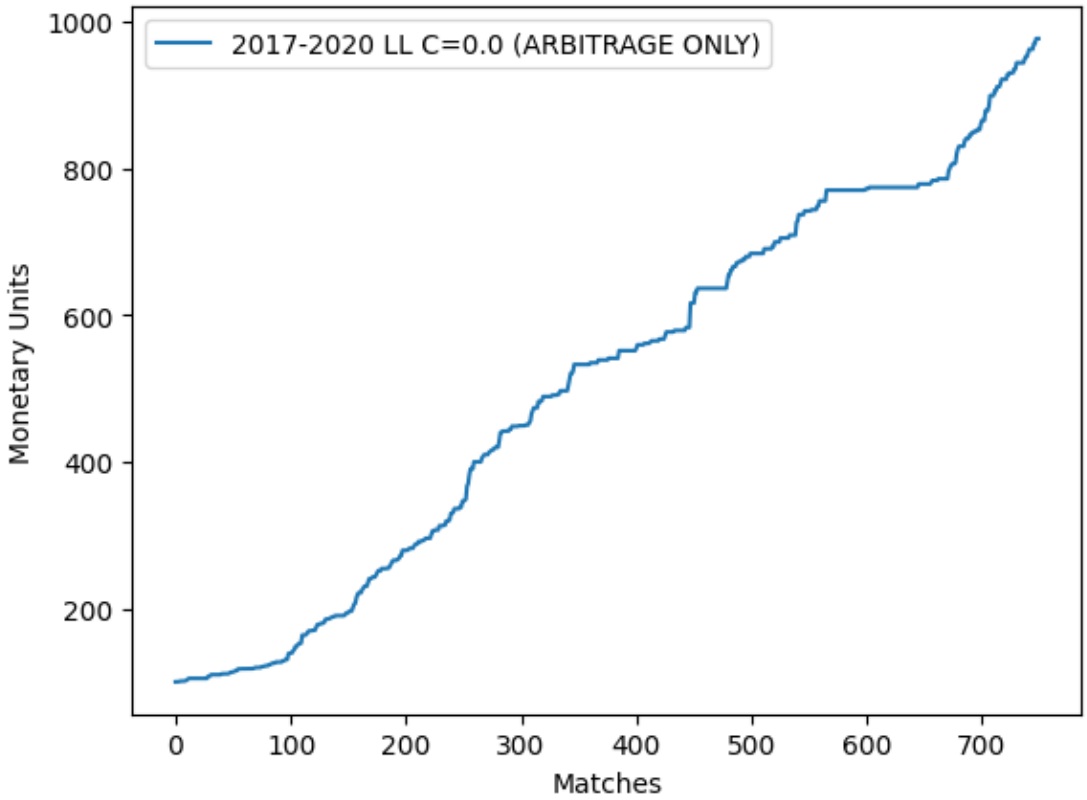



| ΤΑΞΙΝΟΜΗΤΗΣ | ΑΡΧ. ΚΕΦΑΛΑΙΟ | ΤΕΛ. ΚΕΦΑΛΑΙΟ | G | ΟΡΘΟΤΗΤΑ ΠΡΟΒΛΕΨΕΩΝ |
|---|---|---|---|---|
| ΧΩΡΙΣ | **100** | 790 | $39.15 * 10^{-4}$ | 53.5% |
| XGBoost | | 60 | $-9.67 * 10^{-4}$ | 45.9% |
| M-LR | | 269 | $18.77 * 10^{-4}$ | 52.8% |
| QDA | | 3485 | $67.26 * 10^{-4}$ | 52.6% |

- Η επίδοση των ταξινομητών στην Bundesliga αποτελεί σταθερά τη χειρότερη σε σχέση με τα άλλα πρωταθλήματα.
- Ο XGBoost οδηγεί σε απώλεια χρημάτων. Αποτελεί τον μόνο συνδυασμό από όλους που δοκιμάσθηκαν που οδηγεί σε μερική απώλεια κεφαλαίου.
- Για άλλη μια φορά, η καλύτερη πορεία είναι αυτή που χρησιμοποιείται ο ταξινομητής QDA.
- Η ορθότητα των προβλέψεων θα έδινε εντελώς λάθος εικόνα για την κερδοφορία

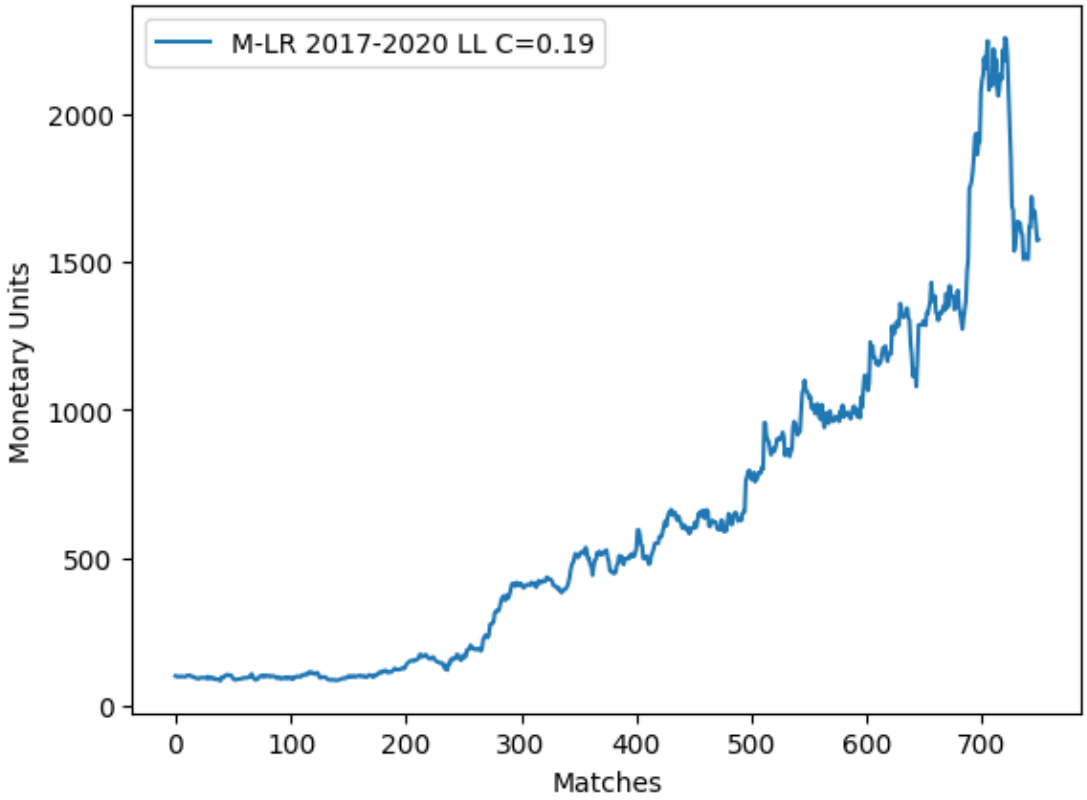
### 4.3.3 La Liga

#### 4.3.3.1 Χωρίς προβλέψεις ταξινομητή



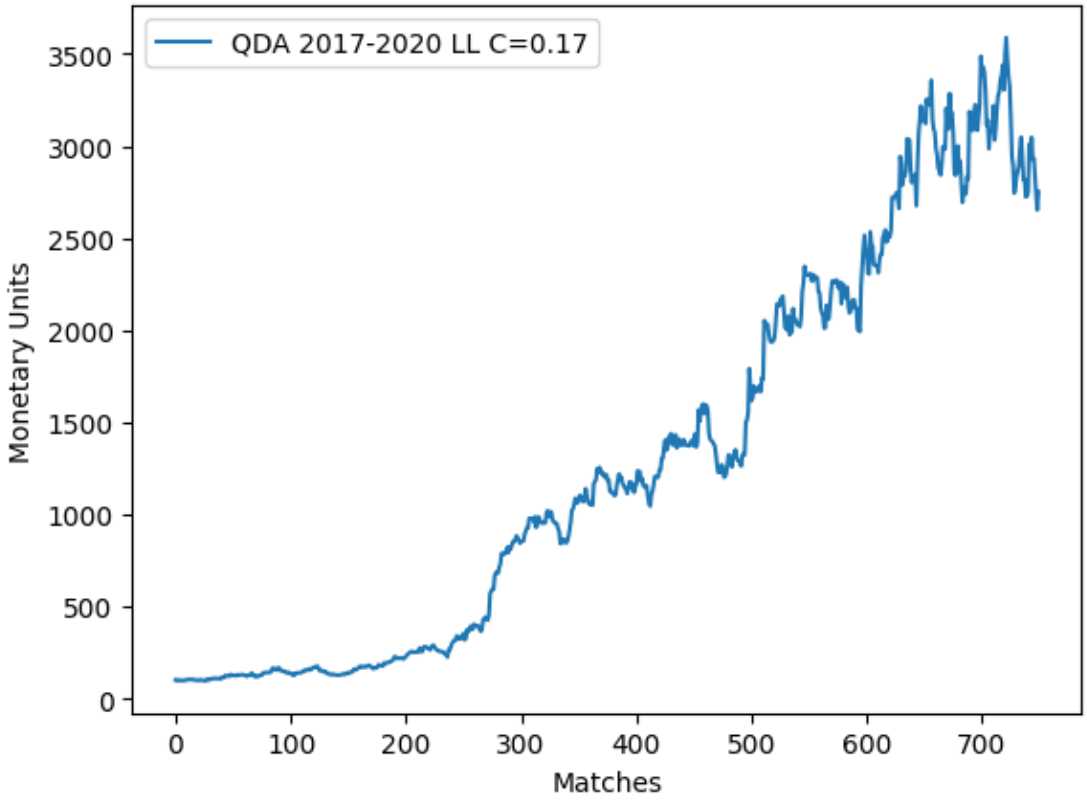

| ΑΡΧΙΚΟ ΚΕΦΑΛΑΙΟ | ΤΕΛΙΚΟ ΚΕΦΑΛΑΙΟ | G | ΟΡΘΟΤΗΤΑ ΠΡΟΒΛΕΨΕΩΝ |
|---|---|---|---|
| 100 | 976 | $30.38 * 10^{-4}$ | 52.5% |

#### 4.3.3.2 *Multinomial Logistic Regression*



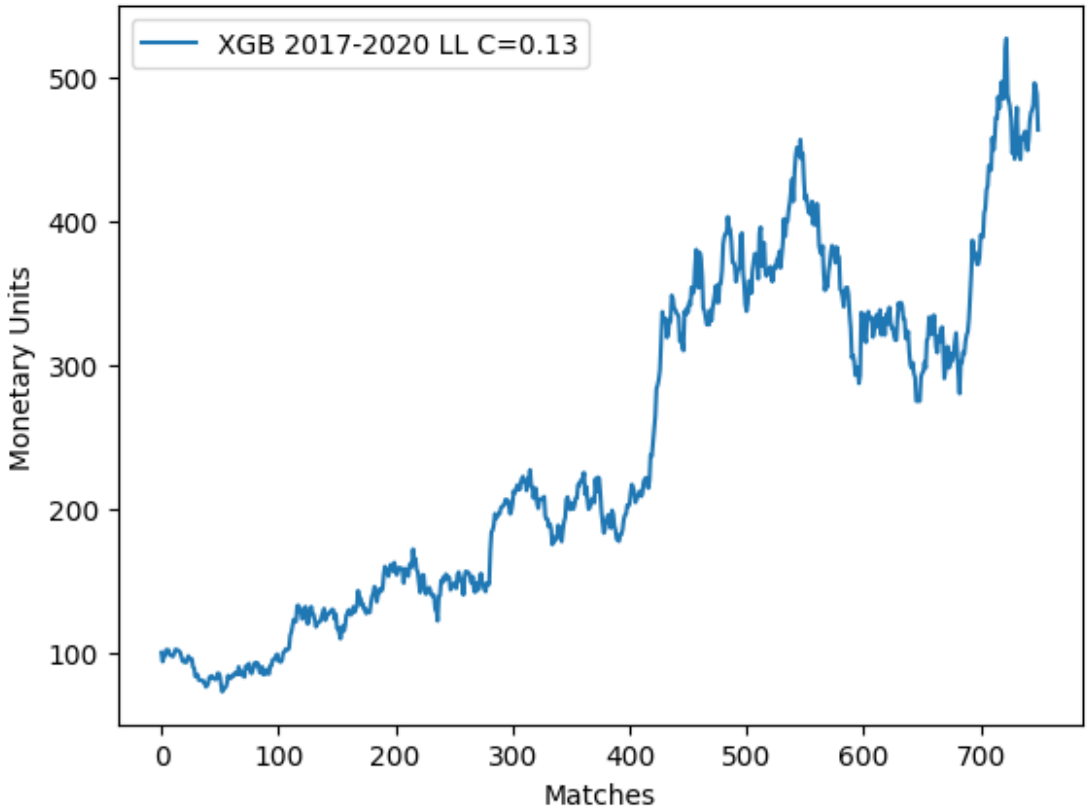

| ΑΡΧΙΚΟ ΚΕΦΑΛΑΙΟ | ΤΕΛΙΚΟ ΚΕΦΑΛΑΙΟ | G | ΟΡΘΟΤΗΤΑ ΠΡΟΒΛΕΨΕΩΝ |
|---|---|---|---|
| 100 | 1575 | $36.75 * 10^{-4}$ | 52.0% |

#### 4.3.3.3 *Quadratic Discriminant Analysis*



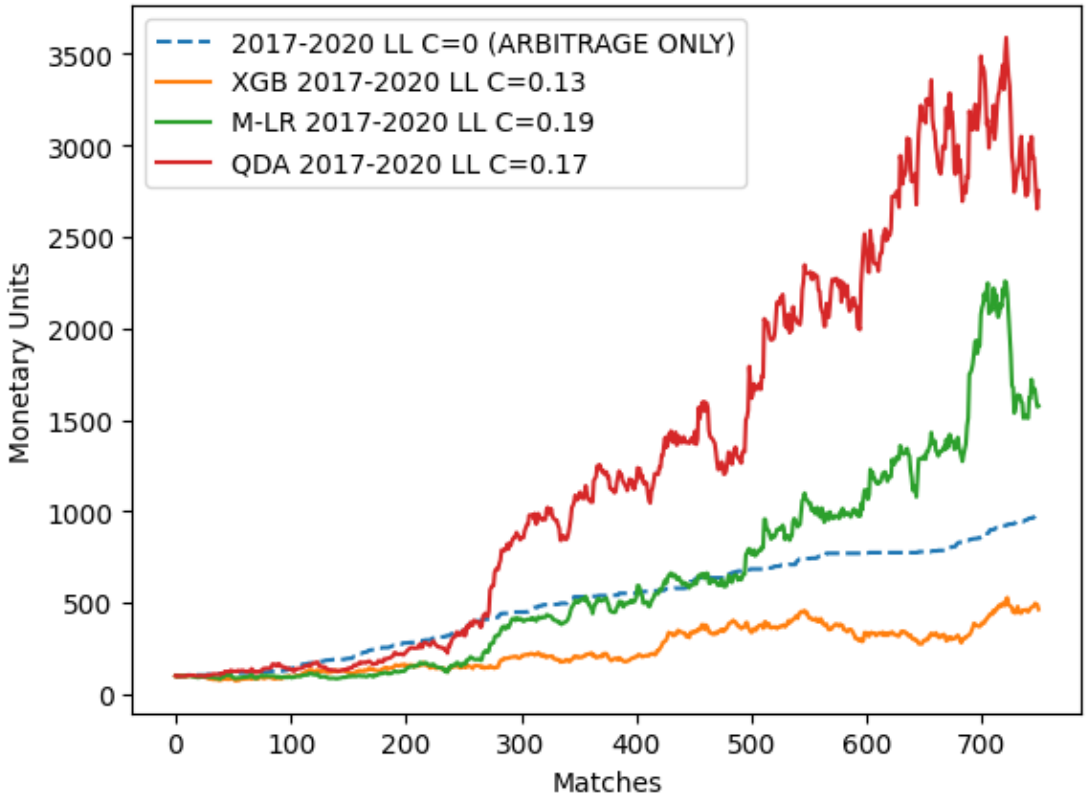

| ΑΡΧΙΚΟ ΚΕΦΑΛΑΙΟ | ΤΕΛΙΚΟ ΚΕΦΑΛΑΙΟ | G | ΟΡΘΟΤΗΤΑ ΠΡΟΒΛΕΨΕΩΝ |
|---|---|---|---|
| 100 | 2752 | $44.20 * 10^{-4}$ | 51.8% |

#### 4.3.3.4 XGBoost




| ΑΡΧΙΚΟ ΚΕΦΑΛΑΙΟ | ΤΕΛΙΚΟ ΚΕΦΑΛΑΙΟ | G | ΟΡΘΟΤΗΤΑ ΠΡΟΒΛΕΨΕΩΝ |
|---|---|---|---|
| 100 | 464 | $20.45 * 10^{-4}$ | 51.7% |

#### 4.3.3.5 *Συγκεντρωτικά*

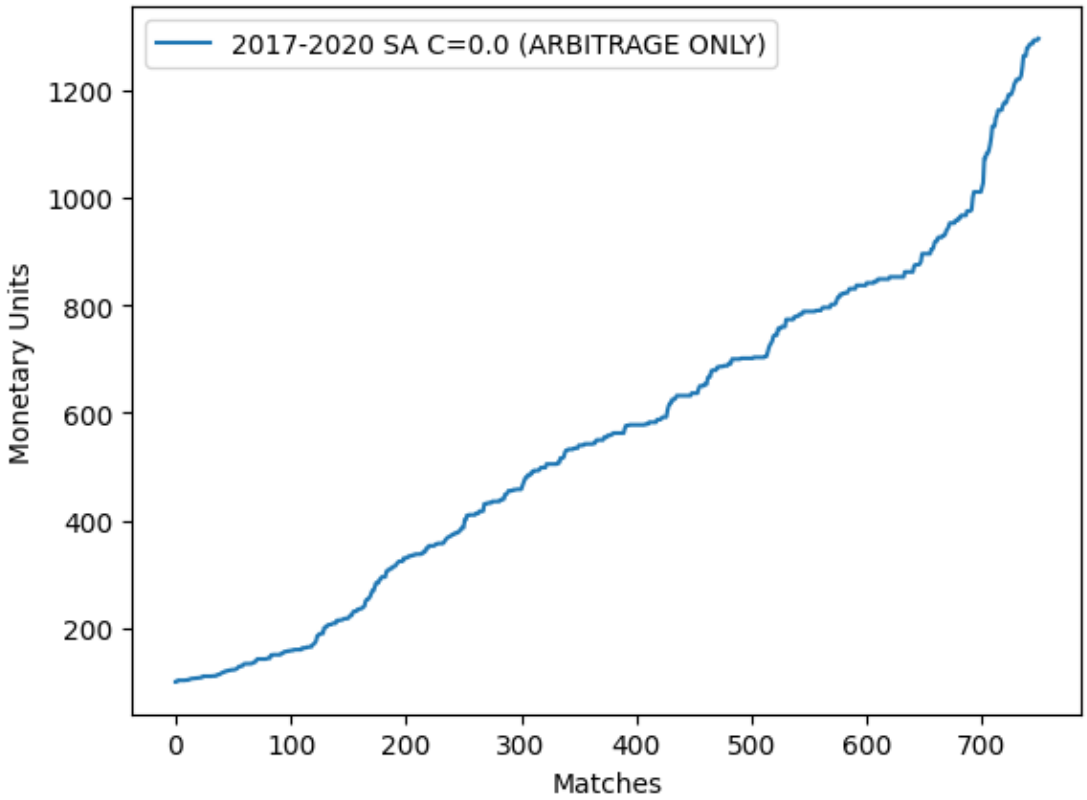



| ΤΑΞΙΝΟΜΗΤΗΣ | ΑΡΧ. ΚΕΦΑΛΑΙΟ | ΤΕΛ. ΚΕΦΑΛΑΙΟ | G | ΟΡΘΟΤΗΤΑ ΠΡΟΒΛΕΨΕΩΝ |
|---|---|---|---|---|
| ΧΩΡΙΣ | **100** | 976 | $30.38 * 10^{-4}$ | 52.5% |
| XGBoost | | 464 | $20.45 * 10^{-4}$ | 51.7% |
| M-LR | | 1575 | $36.75 * 10^{-4}$ | 52.0% |
| QDA | | 2752 | $44.20 * 10^{-4}$ | 51.8% |

- Για άλλη μια φορά, ο XGBoost δίνει τη χειρότερη επίδοση και ο QDA την καλύτερη.
- Η διαφορά της επίδοσης των ταξινομητών σε βάθος χρόνο ακολουθεί σε αυτό το πρωτάθλημα την πιο ξεκάθαρη πορεία.
- Η ορθότητα των προβλέψεων θα έδινε εντελώς λάθος εικόνα για την κερδοφορία

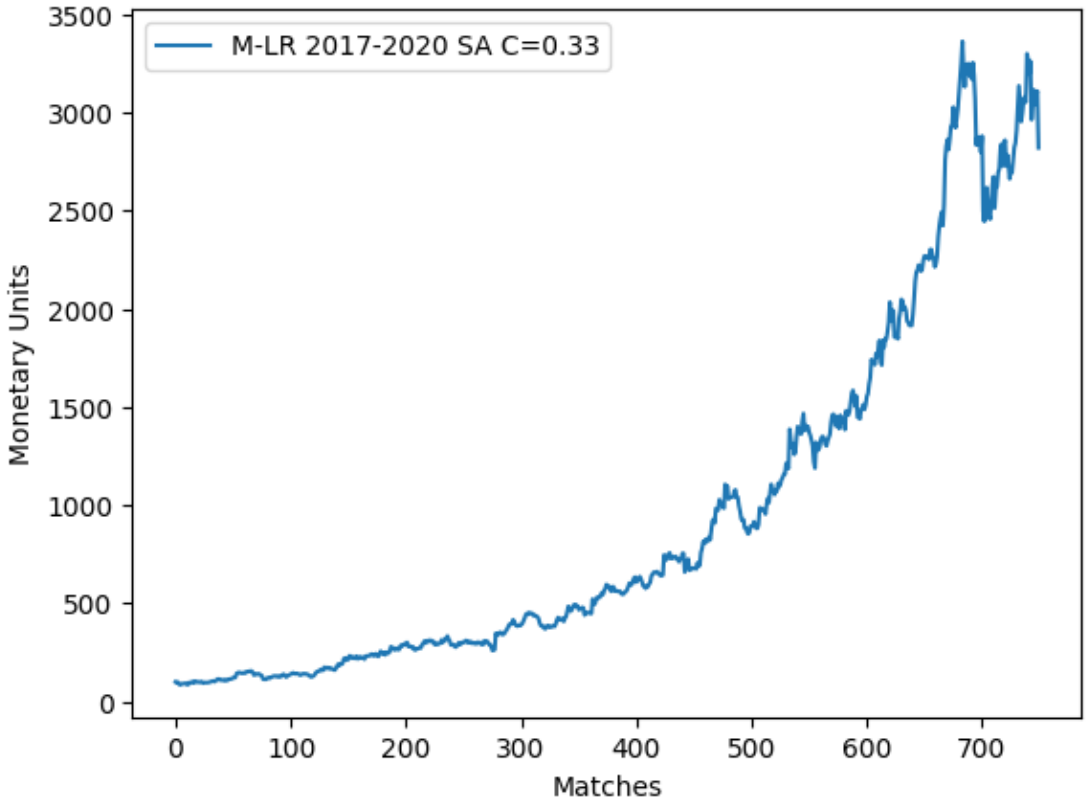
### 4.3.4 Serie A

#### 4.3.4.1 *Χωρίς προβλέψεις ταξινομητή*



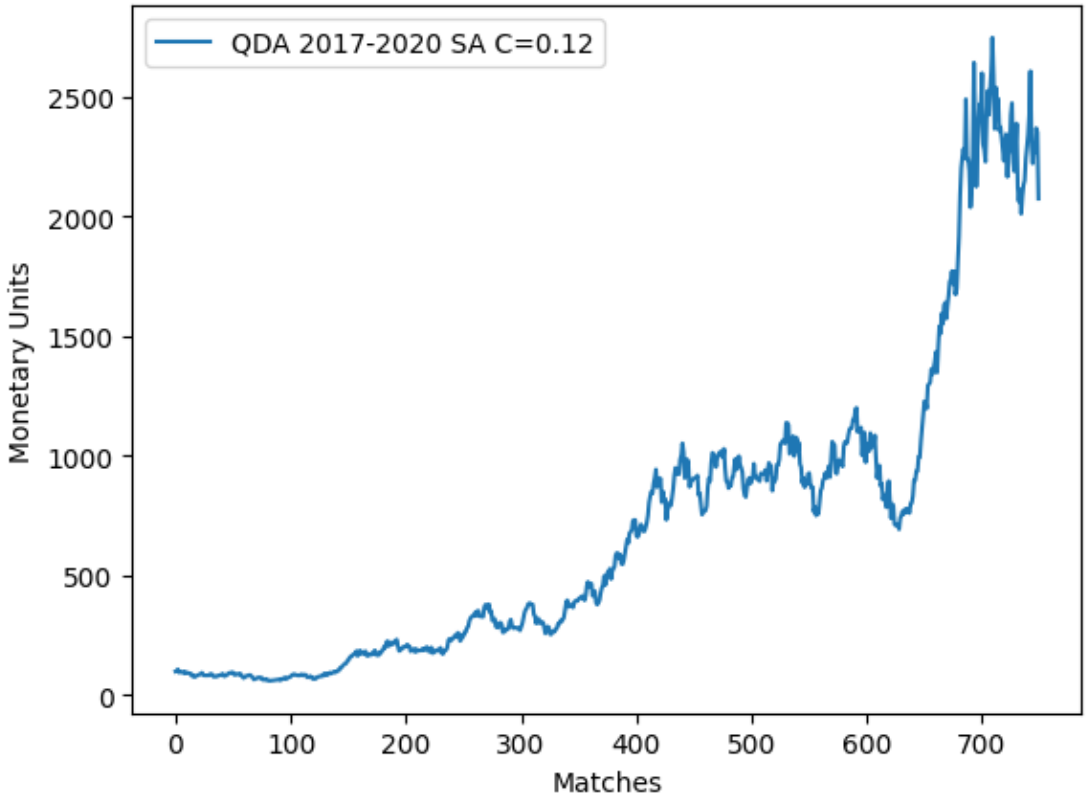

| ΑΡΧΙΚΟ ΚΕΦΑΛΑΙΟ | ΤΕΛΙΚΟ ΚΕΦΑΛΑΙΟ | G | ΟΡΘΟΤΗΤΑ ΠΡΟΒΛΕΨΕΩΝ |
|---|---|---|---|
| 100 | 1295 | $34.14 * 10^{-4}$ | 54.6% |

#### 4.3.4.2 *Multinomial Logistic Regression*



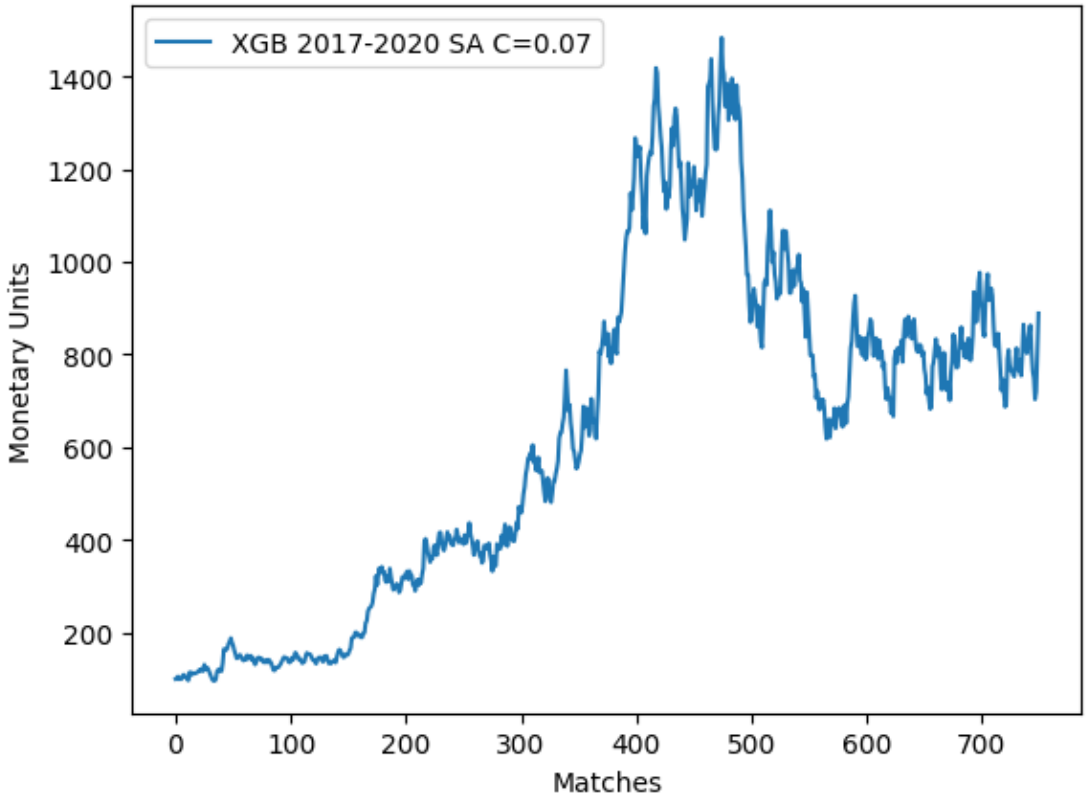

| ΑΡΧΙΚΟ ΚΕΦΑΛΑΙΟ | ΤΕΛΙΚΟ ΚΕΦΑΛΑΙΟ | G | ΟΡΘΟΤΗΤΑ ΠΡΟΒΛΕΨΕΩΝ |
|---|---|---|---|
| 100 | 2821 | $44.53 * 10^{-4}$ | 54.3% |

#### 4.3.4.3 *Quadratic Discriminant Analysis*



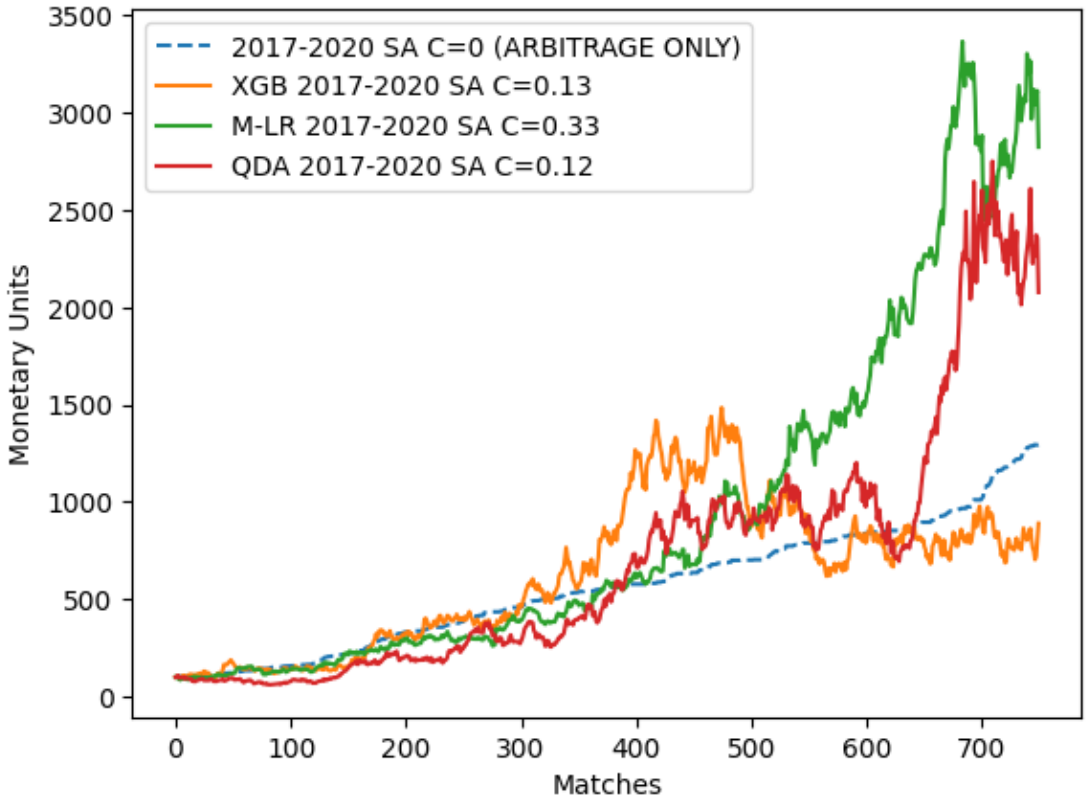

| ΑΡΧΙΚΟ ΚΕΦΑΛΑΙΟ | ΤΕΛΙΚΟ ΚΕΦΑΛΑΙΟ | G | ΟΡΘΟΤΗΤΑ ΠΡΟΒΛΕΨΕΩΝ |
|---|---|---|---|
| 100 | 2075 | $40.44 * 10^{-4}$ | 53.7% |

#### 4.3.4.4 XGBoost




| ΑΡΧΙΚΟ ΚΕΦΑΛΑΙΟ | ΤΕΛΙΚΟ ΚΕΦΑΛΑΙΟ | G | ΟΡΘΟΤΗΤΑ ΠΡΟΒΛΕΨΕΩΝ |
|---|---|---|---|
| 100 | 889 | $29.12 * 10^{-4}$ | 50.7% |

#### 4.3.4.5 *Συγκεντρωτικά*




| ΤΑΞΙΝΟΜΗΤΗΣ | ΑΡΧ. ΚΕΦΑΛΑΙΟ | ΤΕΛ. ΚΕΦΑΛΑΙΟ | G | ΟΡΘΟΤΗΤΑ ΠΡΟΒΛΕΨΕΩΝ |
|---|---|---|---|---|
| ΧΩΡΙΣ | **100** | 1295 | $34.14 * 10^{-4}$ | 54.6% |
| XGBoost | | 889 | $29.12 * 10^{-4}$ | 50.7% |
| M-LR | | 2821 | $44.53 * 10^{-4}$ | 54.3% |
| QDA | | 2075 | $40.44 * 10^{-4}$ | 53.7% |

- Εδώ έχουμε το μοναδικό πρωτάθλημα στο οποίο η επίδοση του ταξινομητή M-LR είναι (οριακά) καλύτερη από τον QDA
- Επίσης οι επιδόσεις είναι πιο συγκρίσιμες μεταξύ τους σε σχέση με αυτές των ταξινομητών σε άλλα πρωταθλήματα.
- Η ορθότητα των προβλέψεων θα έδινε εντελώς λάθος εικόνα για την κερδοφορία

# 5 *Συμπεράσματα*

## *5.1 Συμπεράσματα και Παρατηρήσεις*

Έχοντας κατά νου τους στόχους που είχαν τεθεί για την εργασία οι οποίοι αναλύθηκαν στο εισαγωγικό κεφάλαιο και επιγραμματικά ήταν ο προσδιορισμός της βέλτιστης χρήσης προβλέψεων ποδοσφαίρου σε ό,τι αφορά την επίδοση στο παίγνιο του στοιχήματος και η ανάδειξη αποτελεσματικών ταξινομητών από τη σκοπιά των πιθανοτήτων, μπορούμε να βγάλουμε τα ακόλουθα συνολικά συμπεράσματα από με βάση τα όσα έχουν γραφθεί:

- Σε συνέχεια όσων αναφέρθηκαν στη θεωρία για το κριτήριο Kelly Criterion, χρειάζεται να τονισθεί το εξής. Σε πιθανή ερώτηση ενός αναγνώστη για το αν θα ήταν δυνατό να υπήρχε άλλη καλύτερη στρατηγική από το Kelly Criterion, η απάντηση θα ήταν κατηγορηματικά όχι. Αυτό γιατί μπορεί εύκολα να δειχθεί ότι εφόσον το βέλτιστο ποσό που τοποθετείται σε κάποια επιλογή στο παίγνιο του στοιχήματος είναι ένα σταθερό μερίδιο του κεφαλαίου και ανεξάρτητο από αυτό (το οποίο φυσιολογικά θα εξαρτάται από τις αποδόσεις και τις προβλέψεις), η πορεία του κεφαλαίου θα είναι γεωμετρική (αυξητική ή μειούμενη). Επομένως, μέγιστη κερδοφορία ισοδυναμεί με ταχύτερη γεωμετρική (ή εναλλακτικά εκθετική) ανάπτυξη του κεφαλαίου. Το Kelly Criterion επιχειρεί να μεγιστοποιήσει ακριβώς αυτόν τον ρυθμό ανάπτυξης οπότε ισοδύναμα επιχειρεί να μεγιστοποιήσει την κερδοφορία. Οποιαδήποτε παρέμβαση έχει νόημα να γίνει αφορά μονάχα υπό το πρίσμα του Kelly Criterion μεταβάλλοντας τα δεδομένα εισόδου, εξόδου ή την αντικειμενική συνάρτηση όπως αναλύθηκε στην θεωρία, και γίνεται μονάχα για να αντιμετωπισθεί το γεγονός ότι υπάρχει σφάλμα στις προβλέψεις. Σε κάθε περίπτωση όμως, το καλύτερο πλαίσιο μελέτης του θέματος είναι αυτό του κριτηρίου Kelly.
- Η διαφορά της επίδοσης στο παίγνιο που επιτυγχάνεται με τη χρήση των μεθόδων της παρούσας εργασίας σε σχέση με άλλες ερευνητικές προσπάθειες που χρησιμοποίησαν Μηχανική Μάθηση [3], [5], [7] είναι πολύ μεγάλη. Αυτή δεν

οφείλεται στην ανωτερότητα των τεχνικών πρόβλεψης που δοκιμάσθηκαν αυτή τη φορά αλλά στη μεταβολή της μεθοδολογίας χρήσης των προβλέψεων. Φαίνεται πως ο τρόπος που αξιοποιούνται οι προβλέψεις είναι πιο καθοριστικός στην επίδοση στο παίγνιο από ότι η επιλογή ανάμεσα σε κάποια αποτελεσματικά (κατά γενική ομολογία στην σχετική έρευνα) μοντέλα εξαγωγής τους.

- Είναι ξεκάθαρο ότι είναι διαφορετικό ένας ταξινομητής να παράγει μονάχα απόλυτες προβλέψεις από το να επιστρέφει πιθανότητες που έχει ένα δείγμα να ανήκει σε κάποιες κατηγορίες. Είναι διαφορετικό δηλαδή να ζητηθούν οι πιθανότητες νίκης, ισοπαλίας ή ήττας της γηπεδούχου από το καταλήξει το μοντέλο απλά σε ένα φαβορί. Έτσι, υπάρχουν μοντέλα που αποδίδουν καλύτερα στο ένα πρόβλημα και άλλα που αποδίδουν στο άλλο. Για παράδειγμα, η υλοποίηση που βασίζεται στον ταξινομητή QDA δίνει σχεδόν πάντα την μέγιστη κερδοφορία στα πρωταθλήματα στα αποτελέσματα του 4$^{ου}$ κεφαλαίου αλλά δεν δίνει το ψηλότερο accuracy στις προβλέψεις. Ακριβώς το αντίθετο μπορεί να διατυπωθεί γενικά για τον αλγόριθμο XGBoost. Εδώ να σημειωθεί ότι υψηλότερη κερδοφορία με τη χρήση του κριτηρίου Kelly είναι ισοδύναμη με περισσότερο αντιπροσωπευτικές πιθανότητες στις προβλέψεις συνολικά. Φαίνεται πως ταξινομητές που επιχειρούν να εφαρμόσουν γνωστές παραμετρικές στατιστικές δομές επάνω στα δείγματα είναι πιο αποτελεσματικές στο να δίνουν πιθανότητες από πιο σύγχρονα μοντέλα Μηχανικής Μάθησης. Οι αλγόριθμοι M-LR και QDA έδωσαν πάντα υψηλότερη κερδοφορία στα αποτελέσματα.
- Επίσης πρέπει να επισημανθεί ότι γενικά ο αλγόριθμος XGBoost δίνει γενικά το υψηλότερο accuracy στη βιβλιογραφία. Αυτή η πραγματικότητα φαίνεται να μην επαληθεύεται στα ευρήματα αυτής της εργασίας αλλά αυτό μπορεί να δικαιολογηθεί εξαιτίας του γεγονότος ότι οι τελικές υπερπαράμετροι εκπαίδευσης που προκύπτουν με την τεχνική cross validation είναι αυτές που δίνουν μέγιστη κερδοφορία στα δεδομένα επαλήθευσης. Δεν είναι αυτές που θα έδιναν την υψηλότερη ορθότητα προβλέψεων.
- Λαμβάνουμε τα καλύτερα αποτελέσματα στην Premier League. Προφανώς, δεν μπορεί να προσδιορισθεί ακριβώς γιατί συμβαίνει αυτό αλλά φαίνεται πως οφείλεται στον αυξημένο επαγγελματισμό αυτού του αγγλικού πρωταθλήματος που καθιστά τα αποτελέσματα των αγώνων λιγότερα ευάλωτα απέναντι σε ψυχολογικούς και άλλους εξωαθλητικούς συγκυριακούς παράγοντες. Ακόμα και από το πλαίσιο μελέτης μονάχα της ορθότητας των προβλέψεων, η βιβλιογραφία έχει γενικά αναδείξει αυτό το πρωτάθλημα ως πιο «προσιτό» για τους ταξινομητές [17].

- Η εντυπωσιακή ανάπτυξη του αρχικού κεφαλαίου στο τέλος των προσομοιώσεων του κεφαλαίου 4 οφείλεται στην ύπαρξη μιας σχετικά αδύναμης γεωμετρικής ανάπτυξης σε κάθε γύρο του παιγνίου η οποία γιγαντώνει το κεφάλαιο εξαιτίας των πολλών γύρων παιχνιδιού (δηλαδή των πολλών αγώνων σε βάθος τριών ετών). Αυτό δεν αναφέρεται για να υποβαθμιστεί η αξία των αποτελεσμάτων αλλά γιατί υπάρχει μια πολύ ουσιαστική συνέπεια. Ελαφρά χειρότερες αποδόσεις, δηλαδή αποδόσεις εξίσου «ορθές» πιθανοτικά αλλά με υψηλότερη γκανιότα, θα μπορούσαν να ανατρέψουν αυτήν την ισχνή γεωμετρική αύξηση και να οδηγήσουν γεωμετρική μείωση του κεφαλαίου.

## *5.2 Μελλοντικές Προεκτάσεις*

Οι μελλοντικές προεκτάσεις που θα μπορούσαν να γίνουν αφορούν κυρίως τα επιμέρους κομμάτια της δομής που παρουσιάστηκε με σκοπό να αντικατασταθούν από περισσότερο κατάλληλα για τη εκάστοτε χρήση τους. Πιο συγκεκριμένα:

- Χρήση πιο αποτελεσματικών ταξινομητών. Η επίδοση του συστήματος μπορεί να βελτιωθεί έτη περισσότερο εάν χρησιμοποιηθούν μηχανισμού πρόβλεψης που θα δίνουν πιο αντιπροσωπευτικές προβλέψεις.
- Χρήση πιο αντιπροσωπευτικών δεδομένων. Είναι αλήθεια πως δεν δίνεται όλη η εικόνα για τη δυνατότητα νίκης μιας ομάδας από μονάχα τα δεδομένα που χρησιμοποιήθηκαν. Έτσι η χρήση πιο αντιπροσωπευτικών συνόλων δεδομένων ή εφαρμογή της δομής σε άλλα αθλήματα στα οποία αντικειμενικά στατιστικά δίνουν καλύτερη εικόνα για τα επόμενα αποτελέσματα μπορεί την καταστήσει περισσότερα κερδοφόρα.
- Πειραματισμός με άλλους τρόπους παρέμβασης. Σε αυτή την εργασία παρουσιάστηκε ένας τρόπος παρέμβασης στο γενικευμένο κριτήριο Kelly ο οποίος δίνει καλά αποτελέσματα. Ενδεχομένως να υπάρχουν πιο αποτελεσματικοί τρόποι για να γίνει αυτό.

# 6 *Βιβλιογραφία*


[1] D. Karlis και I. Ntzoufras, «Robust fitting of football prediction models,» *IMA Journal of Management Mathematics,* τόμ. 22, αρ. 2, pp. 171-182, 2010.

[2] L. M. Hvattum και H. Arntzen, «Using ELO ratings for match result prediction in association football,» *International Journal of Forecasting,* τόμ. 26, αρ. 3, pp. 460-470, 2010.

[3] Κ. Αλεξάκης, «Πρόβλεψη αποτελεσμάτων αγώνων ποδοσφαίρου με χρήση ευφυών συστημάτων,» Ε.Μ.Π., Αθήνα, 2019.

[4] H. K. Rahul Baboota, «Predictive analysis and modelling football results using machine learning approach for English Premier League,» *International Journal of Forecasting,* τόμ. 35, αρ. 2, pp. 741-755, 2019.

[5] Δ. Ιωάννου, «Μοντέλα Τεχνητής Νοημοσύνης για την Πρόβλεψη Αποστελεσμάτων σε Αγώνες Ποδοσφαίρου,» Α.Π.Θ., Θεσσαλονίκη, 2013.

[6] J. Kozaka και S. Głowaniaa, «Heterogeneous ensembles of classifiers in predicting Bundesliga football results,» *Procedia Computer Science,* τόμ. 192, pp. 1573-1582, 2021.

[7] J. Stuebinger και J. Knoll, «Beat the Bookmaker – Winning Football Bets with Machine Learning,» https://link.springer.com/chapter/10.1007/978-3-030-04191-5_21, 2018.

[8] M. L. A.R. Hassan, «Profiting from the English Premier League: Predictive elicitation, the Kelly Criterion, and black swans,» *International Journal of Sports Finance,* τόμ. 12, αρ. 4, pp. 386-395, 2017.

[9] Constantinou, Fenton και Neil, «Profiting from an inefficient association football gambling market: Prediction, risk and uncertainty using Bayesian networks,» *Knowledge-Based Systems,* τόμ. 50, pp. 60-86, 2013.

[10] «Matlab (Youtube Channel),» [Ηλεκτρονικό]. Available: https://www.youtube.com/c/matlab. [Πρόσβαση October 2022].

[11] [Ηλεκτρονικό]. Available: https://sebastianraschka.com/faq/docs/softmax_regression.html. [Πρόσβαση October 2022].

[12] «StatQuest with Josh Starmer,» [Ηλεκτρονικό]. Available: https://www.youtube.com/c/joshstarmer. [Πρόσβαση October 2022].

[13] H. S. Shin, «Optimal Betting Odds Against Insider Traders,» *The Economic Journal,* τόμ. 101, αρ. 408, pp. 1179-1185, 1991.

[14] E. O. Thorp, A Man for All Markets: From Las Vegas to Wall Street, How I Beat the Dealer and the Market, 2017.

[15] S. G. Jan Kozak, «Heterogeneous ensembles of classifiers in predicting Bundesliga football results,» *Procedia Computer Science,* τόμ. 192, pp. 1573-1582, 2021.

[16] E. Thorp, «The Kelly Criterion in Blackjack Sports Betting, and the Stock Market,» σε *Handbook of Asset and Liability Management, Volume I,* Elsevier B.V, 2006.

[17] E. Štrumbelj και M. R. Šikonja, «Online bookmakers' odds as forecasts: The case of European soccer leagues,» *International Journal of Forecasting,* τόμ. 26, αρ. 3, pp. 482-488, 2010.

[18] Sibony, John, Tlemsani, Abdelfatah, Hamchi, Youssef, Gjergji, Monika, Shurmanov, Evgueny, Frunza και Marius, «Neural Networks Predictive Modeling for Football Betting,» Available at SSRN: https://ssrn.com/abstract=3655700, 2020.

[19] J. L. Kelly, «A new interpretation of information rate,» *The Bell System Technical Journal,* τόμ. 35, αρ. 4, pp. 917-926, 4 July 1956.

[20] S. Guan και X. Wang, «Optimization analysis of football match prediction model based on neural network,» *Neural Computing and Applications,* τόμ. 34, pp. 2525-2541, 2022.

[21] S. Jain και E. Tiwari, «Soccer Result Prediction Using Deep Learning and Neural Networks,» σε *Intelligent Data Communication Technologies and Internet of Things,* 2021, pp. 697-707.

[22] N. Zaveri, S. Tiwari, P. Shinde, U. Shah και L. K. Teli, «Prediction of Football Match Score and Decision Making Process,» *International Journal on Recent and Innovation Trends on Computing and Communication,* τόμ. 6, αρ. 2, 2018.

[23] W. Gourh, K. Poojary, M. Vengarai και N. Parkar, «Football Prognosis using Machine Learning Algorithm XGBoost: A Literature Review,» 2020.

[24] «FIFA Index,» [Ηλεκτρονικό]. Available: https://www.fifaindex.com/. [Πρόσβαση October 2022].

[25] «Football-Data,» [Ηλεκτρονικό]. Available: https://www.football-data.co.uk/. [Πρόσβαση October 2022].