Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών
& Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Πρόβλεψη αποτελεσμάτων αγώνων ποδοσφαίρου με χρήση ευφυών συστημάτων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Κωνσταντίνου Π. Αλεξάκη

Επιβλέπων : Ανδρέας - Γεώργιος Σταφυλοπάτης Καθηγητής Ε.Μ.Π.

Συνεπιβλέπων: Γεώργιος Σιόλας ΕΔΙΠ Ε.Μ.Π.

Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών
& Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Πρόβλεψη αποτελεσμάτων αγώνων ποδοσφαίρου με χρήση ευφυών συστημάτων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Κωνσταντίνου Π. Αλεξάκη

Επιβλέπων : Ανδρέας - Γεώργιος Σταφυλοπάτης Καθηγητής Ε.Μ.Π.

Συνεπιβλέπων: Γεώργιος Σιόλας ΕΔΙΠ Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 8η Νοεμβρίου 2019.

.....	.....	.....
Σταφυλοπάτης Ανδρέας-Γεώργιος	**Στάμου Γεώργιος**	**Τσανάκας Παναγιώτης**
Καθηγητής ΕΜΠ	Αναπληρωτής Καθηγητής ΕΜΠ	Καθηγητής ΕΜΠ

Αθήνα, Νοέμβριος 2019

.....
Κωνσταντίνος Π. Αλεξάκης

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

Περίληψη

Η δημιουργία προβλέψεων με χρήση συστημάτων Μηχανικής Μάθησης έχει γνωρίσει τεράστια ανάπτυξη τα τελευταία χρόνια, ενώ πλέον χρησιμοποιείται και σε πολλούς και διαφορετικούς τομείς. Τα συστήματα Μηχανικής Μάθησης χρησιμοποιούν δεδομένα του παρελθόντος, που είναι γνωστό το αποτέλεσμα τους, και προσπαθούν να προβλέψουν τις νέες καταστάσεις, αναγνωρίζοντας τα πρότυπα και τα μοτίβα στα διαθέσιμα ιστορικά στοιχεία. Η παρούσα εργασία θα ασχοληθεί με την πρόβλεψη αποτελεσμάτων αγώνων ποδοσφαίρου του Αγγλικού πρωταθλήματος και σκοπό να εξετάσει αν κάποιο μοντέλο μπορεί να οδηγήσει σε κέρδος, χρησιμοποιώντας τις στοιχηματικές αποδόσεις μεγάλων ευρωπαϊκών εταιρειών. Συνολικά αναπτύχθηκαν επτά διαφορετικά μοντέλα, πετυχαίνοντας πολύ ικανοποιητικά και ελπιδοφόρα για το μέλλον αποτελέσματα.

Λέξεις Κλειδιά

Μηχανική Μάθηση, προβλέψεις, στοιχηματικές αποδόσεις, κέρδος, Αγγλικό πρωτάθλημα ποδοσφαίρου

Abstract

Making predictions using Machine Learning techniques has grown significantly the last decades and is widely used in different domains. The Machine learning systems exploit data from the past and try to recognize patterns in these, in order to predict the new class of the new data. In the herein presented thesis, Machine learning algorithms were applied to predict the outcome of matches from the English Premier League. The scope of this analysis is to examine and search for models able to lead to profit, using these predictions and the betting odds from large betting firms. Overall, seven different models have been developed, which resulted in a satisfactory and promising outcome.

Key Words

Machine Learning, predictions, betting odds, profit, English Premier League, football

Ευχαριστίες

Η παρούσα εργασία εκπονήθηκε στα πλαίσια του προπτυχιακού κύκλου σπουδών της Σχολής Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών και αποτελεί το τελευταίο μου βήμα πριν την απόκτηση του Διπλώματος. Η χαρά μου είναι μεγάλη και θα ήθελα να ευχαριστήσω όλους τους ανθρώπους που συνέβαλαν και βοήθησαν να φτάσω ως εδώ.

Αρχικά, θα ήθελα να ευχαριστήσω τον κ. Ανδρέα-Γεώργιο Σταφυλοπάτη, Καθηγητή Ε.ΜΠ., που ήταν επιβλέπων της εργασίας και μου επέτρεψε να ασχοληθώ με το συγκεκριμένο θέμα, που τόσο πολύ ήθελα, και να διευρύνω τους επιστημονικούς μου ορίζοντες. Ακολούθως, θα ήθελα να ευχαριστήσω τους κ. κ. Παναγιώτη Τσανάκα, Καθηγητή Ε.Μ.Π., και Γεώργιο Στάμου, Αναπληρωτή Καθηγητή Ε.Μ.Π., για την τιμή που μου έκαναν να αποτελέσουν την εξεταστική επιτροπή.

Θα ήθελα να ευχαριστήσω θερμά τον κ. Γεώργιο Σιόλα, ΕΔΙΠ Ε.Μ.Π., που ήταν ο συνεπιβλέπων της εργασίας και με βοήθησε σημαντικά στην σωστή προσέγγιση του προβλήματος. Η καθοδήγησή του και η στήριξη του, επιστημονική και ψυχολογική, βοήθησαν τα μέγιστα στην ολοκλήρωση της εργασίας.

Ένα μεγάλο ευχαριστώ οφείλω ακόμα στους φίλους και συμφοιτητές μου, που 5 χρόνια τώρα με στήριξαν και με βοήθησαν να φτάσω ως εδώ. Θα ήθελα να ευχαριστήσω ακόμα περισσότερο τους συνεργάτες που είχα όλα αυτά τα χρόνια, σε εργασίες και ασκήσεις, καθώς και τα άτομα που διαβάσαμε και μοχθήσαμε μαζί με τις ώρες.

Τέλος, ένα μεγάλο ευχαριστώ στους δικούς μου ανθρώπους, τους φίλους και την οικογένειά μου, που πάντα πίστευαν σε μένα και με στήριζαν. Ειδικότερα, στην αδερφή μου, που όλα αυτά τα χρόνια με γεμίζει αγάπη και ενέργεια να πετυχαίνω τους στόχους μου.

Αλεξάκης Κωνσταντίνος,
Αθήνα, 8η Νοεμβρίου 2019

Περιεχόμενα

Κεφάλαιο 1: Εισαγωγή

1.1 Εισαγωγή

Ο άνθρωπος από αρχαιοτάτων χρόνων είχε ενδιαφέρον για το μέλλον και την πρόβλεψη του, φοβούμενος και έχοντας αγωνία για το άγνωστο που θα ακολουθούσε. Ήθελε να γνωρίζει τι καιρό θα έχει το επόμενο διάστημα, για να οργανώσει τις δραστηριότητες του, ποια θα είναι η έκβαση ενός πολέμου, για να ξέρει αν θα βγει νικητής ή ηττημένος, τι φύλο θα έχει ένα παιδί που πρόκειται να έρθει σύντομα στη ζωή και πολλά ακόμα. Για να καταφέρει ο άνθρωπος να προβλέπει με "επιτυχία" το μέλλον, είχε εφεύρει αρκετούς τρόπους, όπως οι χρησμοί και τα Μαντεία, η παρατήρηση των οιωνών, τα σημάδια στα δέντρα και τον ουρανό και άλλους ευφάνταστους τρόπους.

Κατά το πέρασμα των αιώνων και την πρόοδο του, το πάθος του και η ανάγκη του να βρει τρόπους να προβλέπει το μέλλον παρέμεναν κυρίαρχοι στόχοι. Το μόνο που διαφοροποιούνταν ήταν το μέσο που θα χρησιμοποιούσε για να κάνει τις προβλέψεις του. Μπορεί αυτή τη φορά να μην περίμενε από την Πυθία κάποιον χρησμό, αλλά να "διάβαζε" το πέταγμα των πουλιών ή να κατέφευγε στη χαρτομαντεία. Σε κάθε περίπτωση πάντως, το ενδιαφέρον του ανθρώπου και η περιέργεια του για το μέλλον παρέμεναν πολύ σημαντικά.

Με την περαιτέρω πρόοδο του, ο άνθρωπος συνειδητοποίησε ότι το πιο ισχυρό όπλο του στη μάχη πρόβλεψης του μέλλοντος είναι το παρελθόν και οι εμπειρίες που έχει. Έτσι, ο άνθρωπος σχεδόν απαλλάχτηκε από τις προκαταλήψεις που είχε σχετικά με τα μέσα που χρησιμοποιούσε να προβλέψει το μέλλον. Λέμε σχεδόν, γιατί ακόμα υπάρχουν άτομα που πιστεύουν σε μη επιστημονικά μέσα πρόβλεψης του μέλλοντος, όπως η χαρτομαντεία και η παρατήρηση των οιωνών. Ίσως αυτό δικαιολογείται από την πολύπλοκη και σύνθετη φύση του ανθρώπου, που προσπαθεί να βρει εύκολους και εναλλακτικούς τρόπους για να λύσει τα προβλήματα του. Κάπως έτσι δικαιολογείται και η ύπαρξη "μάγων" στις φυλές και πολλών προφητών, που έχουν κάνει την εμφάνιση τους σε όλους αυτούς τους αιώνες.

Ο άνθρωπος, λοιπόν, συνειδητοποίησε ότι θα πρέπει να συγκεντρώσει πληροφορίες από το παρελθόν, να τις αναλύσει σωστά και μεθοδικά και ύστερα να παράξει τις προβλέψεις του. Εργαζόμενος πλέον με λογική και μεθοδικότητα και χρησιμοποιώντας εργαλεία της στατιστικής, ο άνθρωπος ξεκίνησε να παράγει πιο ορθολογικές προβλέψεις. Στηριζόμενος σε μαθηματικά εργαλεία, προσπάθησε να αναγνωρίσει μοτίβα και πρότυπα στα δεδομένα του παρελθόντος και βάσει αυτών να προβλέψει και το μέλλον.

Με την τεράστια ανάπτυξη της τεχνολογίας η συλλογή δεδομένων έγινε ακόμα πιο εύκολη. Πλέον, το άτομο δεν χρειαζόταν να καταγράφει με το χέρι γεγονότα του παρελθόντος, αλλά μπορούσε να έχει πρόσβαση σε αυτά μέσω του διαδικτύου. Αυτή η έκρηξη πληροφορίας, δημιούργησε ένα νέο πρόβλημα στον άνθρωπο, διέθετε τόση πληροφορία που δεν μπορούσε να την επεξεργαστεί και να τη μελετήσει μόνος του. Τα δεδομένα έπαψαν να είναι μερικές δεκάδες ή εκατοντάδες στοιχεία και άρχισαν να είναι χιλιάδες και εκατομμύρια. Τη λύση έδωσε η τεχνολογία, που δημιούργησε και το πρόβλημα. Χρησιμοποίησε λοιπόν, υπολογιστικά συστήματα και η επεξεργασία, η ανάλυση και αποθήκευση δεδομένων έγινε πιο εύκολη και γρήγορη.

Η Μηχανική Μάθηση ήρθε σε αυτό το σημείο και έδωσε ώθηση στις προβλέψεις του ανθρώπου, με αποτέλεσμα και η ίδια να εκτοξευτεί. Η Μηχανική Μάθηση χρησιμοποιώντας δεδομένα του παρελθόντος, αναλύοντας τα με στατιστικά εργαλεία και εφαρμόζοντας διαφορετικές μεθόδους, παράγει προβλέψεις για το μέλλον. Τα αποτελέσματα της είναι τόσο ικανοποιητικά, που ήδη η Μηχανική Μάθηση χρησιμοποιείται ευρέως στις περισσότερες εφαρμογές προβλέψεων.

Πλέον, ο άνθρωπος δεν ενδιαφέρεται μόνο για τη πρόγνωση του καιρού και άλλα ζητήματα, που τον ενδιέφεραν παλιότερα, αλλά ασχολείται και με δραστηριότητες που μπορούν να του αποφέρουν κέρδος. Εξάλλου, η μάχη για ευημερία και κέρδος είναι από τα κύρια χαρακτηριστικά της ανθρωπότητας, όχι μόνο τώρα αλλά πάντοτε. Στη σημερινή εποχή έχουν τεράστια ανάπτυξη τα τυχερά παιχνίδια, με το στοιχηματισμό σε αθλητικά γεγονότα ή σε παιχνίδια αριθμών να κυριαρχούν. Μέσα σε αυτά, μπορεί κανείς να συνυπολογίσει και το χρηματιστήριο, που αποτελεί μια εναλλακτική μορφή στοιχήματος, χωρίς ομάδες, αλλά με μετοχές. Από τα παραπάνω έχουμε ξεχωρίσει το στοίχημα σε ποδοσφαιρικά γεγονότα, μιας και το ποδόσφαιρο είναι το λαοφιλέστερο άθλημα και ένα άθλημα που περιτριγυρίζεται από τεράστια οικονομικά ποσά.

1.2 Περιγραφή και στόχος της εργασίας

Το ποδόσφαιρο αποτελείται από τρία διαφορετικά αποτελέσματα, τη νίκη γηπεδούχου, την ισοπαλία και την ήττα της φιλοξενούμενης ομάδας. Σε κάθε διαφορετικό αποτέλεσμα παρέχεται από τις στοιχηματικές εταιρείες κάποια απόδοση, που μπορεί το κάθε άτομο να ποντάρει σε αυτή και αν προβλέψει σωστά το αποτέλεσμα, να εισπράξει το κεφάλαιο του πολλαπλασιασμένο με αυτή την απόδοση. Ο άνθρωπος προσπαθεί να προβλέψει σωστά έναν αγώνα, προκειμένου να αποκτήσει κέρδος. Κάθε μέρα εκατομμύρια χρημάτων ποντάρονται σε στοιχηματικές αποδόσεις και τα άτομα ελπίζουν να αποκτήσουν κέρδος. Η παρούσα εργασία προσπαθεί να δημιουργήσει μοντέλα, που θα προβλέπουν το τελικό αποτέλεσμα αγώνων ποδοσφαίρου από το Αγγλικό Πρωτάθλημα, στοχεύοντας στη δημιουργία κέρδους μακροπρόθεσμα από το στοιχηματισμό σε αυτούς τους αγώνες. Να διευκρινιστεί εδώ, ότι στόχος της εργασίας είναι η ανάπτυξη κερδοφόρων μοντέλων και όχι μοντέλων που επιτυγχάνουν υψηλή ορθότητα (accuracy) στις προβλέψεις τους. Όπως θα φανεί και από τα συμπεράσματα της εργασίας, η υψηλή ορθότητα δεν είναι ανάλογη της κερδοφορίας.

Για τη δημιουργία αυτών των μοντέλων θα χρησιμοποιηθούν βασικοί αλγόριθμοι Μηχανικής Μάθησης. Δεδομένα για αυτό το πρόβλημα βρέθηκαν σχετικά εύκολα και μάλιστα για αρκετά χρόνια πίσω από το διαδίκτυο. Από τα δεδομένα αυτά, υπολογίστηκαν κι άλλα δεδομένα, όπως η βαθμολογία και άλλα στατιστικά που δεν ήταν διαθέσιμα. Στη συνέχεια, τα σημαντικότερα δεδομένα χρησιμοποιήθηκαν για την εκπαίδευση διάφορων μοντέλων Μηχανικής Μάθησης. Ιδιαίτερη αξία έχει η διαδικασία επιλογής των καλύτερων παραμέτρων για κάθε μοντέλο (grid search), όπου ως μέτρο δεν εφαρμόστηκε η ορθότητα, που είναι και η πιο συνηθισμένη συνθήκη, αλλά μία συνθήκη που αυξάνει την κερδοφορία των συστημάτων και η οποία θα αναλυθεί επαρκώς στη συνέχεια. Μετά τη δημιουργία των μοντέλων, εφαρμόστηκαν διαφορετικές προσεγγίσεις όσον αφορά τα παιχνίδια που θα επιλέγει κανείς να στοιχηματίσει. Για την επαρκή και ορθή σύγκριση και αξιολόγηση των συστημάτων χρησιμοποιήθηκε μια έννοια από τα οικονομικά, το Yield, το οποίο δείχνει σε μακροπρόθεσμη βάση αν ένα σύστημα είναι κερδοφόρο ή ζημιογόνο. Συνολικά θα δημιουργηθούν και θα συγκριθούν 7 διαφορετικά μοντέλα.

1.3 Οργάνωση εργασίας

Στο πρώτο κεφάλαιο δόθηκε μία σύντομη εισαγωγή και περιγραφή του προβλήματος, που πραγματεύεται η παρούσα εργασία, ενώ ορίστηκαν και οι στόχοι της.

Στο δεύτερο κεφάλαιο παρουσιάζεται το θεωρητικό υπόβαθρο των αλγορίθμων Μηχανικής Μάθησης, που χρησιμοποιούνται στην εργασία. Πιο συγκεκριμένα, κάθε αλγόριθμος αναλύεται συνοπτικά και περιγράφονται και οι παράμετροί που χρησιμοποιήθηκαν σε αυτή την εργασία.

Στο τρίτο κεφάλαιο παρουσιάζονται και επεξηγούνται ποδοσφαιρικές και στοιχηματικές έννοιες, που απαιτούνται για την καλύτερη κατανόηση του προβλήματος και της εργασίας. Ακόμα, ορίζεται και αναλύεται με παραδείγματα η αξία του Yield και του ROI.

Στο τέταρτο κεφάλαιο περιγράφονται και παρουσιάζονται τα χαρακτηριστικά, τα features, με τα οποία θα εκπαιδευτούν στη συνέχεια τα μοντέλα. Επιπλέον, παρουσιάζονται και οι συσχετίσεις μεταξύ των χαρακτηριστικών.

Στο πέμπτο κεφάλαιο παρουσιάζονται οι διαφορετικές υλοποιήσεις και τα αποτελέσματα τους, τα οποία σχολιάζονται συνοπτικά. Σε κάθε υλοποίηση ξεχωρίζονται και οι επικρατέστερες προσεγγίσεις.

Στο έκτο, και τελευταίο κεφάλαιο, παρουσιάζονται τα κυριότερα αποτελέσματα και οι παρατηρήσεις που προέκυψαν από την παρούσα εργασία, ενώ γίνεται και αναφορά σε πιθανές μελλοντικές επεκτάσεις.

Κεφάλαιο 2: Θεωρία

2.1 Μηχανικής Μάθηση – Θεωρητικό υπόβαθρο

Η Μηχανική Μάθηση είναι πεδίο της Επιστήμης Υπολογιστών, που αναπτύχθηκε από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην Τεχνητή νοημοσύνη. Τα τελευταία χρόνια αναπτύσσεται συνεχώς και έχει κάνει την εμφάνιση της σε αρκετούς τομείς, όπου η χρήση υπολογιστικών συστημάτων ήταν περιορισμένη ή ακόμα και ανύπαρκτη, όπως είναι ο έλεγχος παραγωγής εργοστασίων, η διαφήμιση, η αξιολόγηση των εργαζομένων και πολλά ακόμα. Δεν είναι τυχαίο που όλο και περισσότερες εταιρείες εντάσσουν τη Μηχανική Μάθηση στις λειτουργίες τους, ακολουθώντας τις τεχνολογικές εξελίξεις. Ο ορισμός που πρότεινε ο Tom M. Mitchell για τη Μηχανική Μάθηση είναι ο ακόλουθος: “ 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”, δηλαδή ένα υπολογιστικό σύστημα χρησιμοποιεί παλαιότερα δεδομένα και προσπαθεί να “μάθει” από αυτά τα δεδομένα, να αναγνωρίσει δηλαδή τα πρότυπα και τα μοτίβα που ακολουθούν αυτά, ώστε να ταξινομήσει νέα δεδομένα, να βρει ομοιότητες μεταξύ τους, να τα ομαδοποιήσει κλπ.

Υπάρχουν αρκετές κατηγορίες προβλημάτων Μηχανικής Μάθησης, ωστόσο η παρούσα εργασία θα ασχοληθεί αποκλειστικά με την επιβλεπόμενη μάθηση (supervised learning), όπου το υπολογιστικό σύστημα εκπαιδεύεται σε ένα σύνολο παραδειγμάτων, τα οποία συνοδεύονται με τις κατηγορίες-κλάσεις στις οποίες ανήκουν τα δεδομένα. Το σύστημα δηλαδή, χρησιμοποιεί την εμπειρία, παλιά δεδομένα τα οποία είναι διαθέσιμα, με σκοπό να προβλέψει μελλοντικές καταστάσεις, ταξινομώντας τα νέα δεδομένα στην κλάση που πιστεύει ότι ανήκουν. Να αναφερθεί εδώ, ότι γενικότερα στις προβλέψεις γίνεται η υπόθεση ότι το μοτίβο που επικρατεί στα μέχρι τώρα δεδομένα, ή έστω στα τελευταία, θα συνεχίσει να εμφανίζεται και στα επόμενα. Επομένως, είναι πραγματικά δύσκολο να είναι κανείς σίγουρος για τις προβλέψεις του, καθώς μπορεί το μοτίβο που μέχρι τώρα εμφανιζόταν να αλλάξει ή και να εξαφανιστεί.

Σε κάθε πρόβλημα της Μηχανικής Μάθησης υπάρχει και το σύνολο δεδομένων (dataset) με το οποίο θα εργαστεί κάποιος για την ανάπτυξη του συστήματος, που επιλύει το πρόβλημα. Για τη δημιουργία αυτού του συστήματος, το αρχικό dataset διαμερίζεται σε 2 υποσύνολα δεδομένων, το σύνολο δεδομένων εκπαίδευσης (training set), όπου το σύστημα θα προσπαθήσει να “μάθει” από τα δεδομένα και το σύνολο στο οποίο θα δοκιμαστεί η απόδοση του μοντέλου (test set). Με αυτό τον τρόπο σχεδιάζονται και εκπαιδεύονται οι περισσότεροι αλγόριθμοι Μηχανικής Μάθησης.

Η παρούσα εργασία χρησιμοποιεί δεδομένα από αγώνες του Αγγλικού Πρωταθλήματος Ποδοσφαίρου, με σκοπό να προβλέψει το τελικό αποτέλεσμα επόμενων αγώνων του ίδιου Πρωταθλήματος και διερευνά αν από τις παραπάνω προβλέψεις είναι δυνατή η δημιουργία κέρδους, χρησιμοποιώντας τις στοιχηματικές αποδόσεις. Για το σκοπό αυτό χρησιμοποιήθηκαν οι ακόλουθοι αλγόριθμοι: k-Nearest Neighbours, Gaussian Naive Bayes, Support Vector Machines και Multilayer Perceptron, οι οποίοι παρουσιάζονται συνοπτικά παρακάτω.

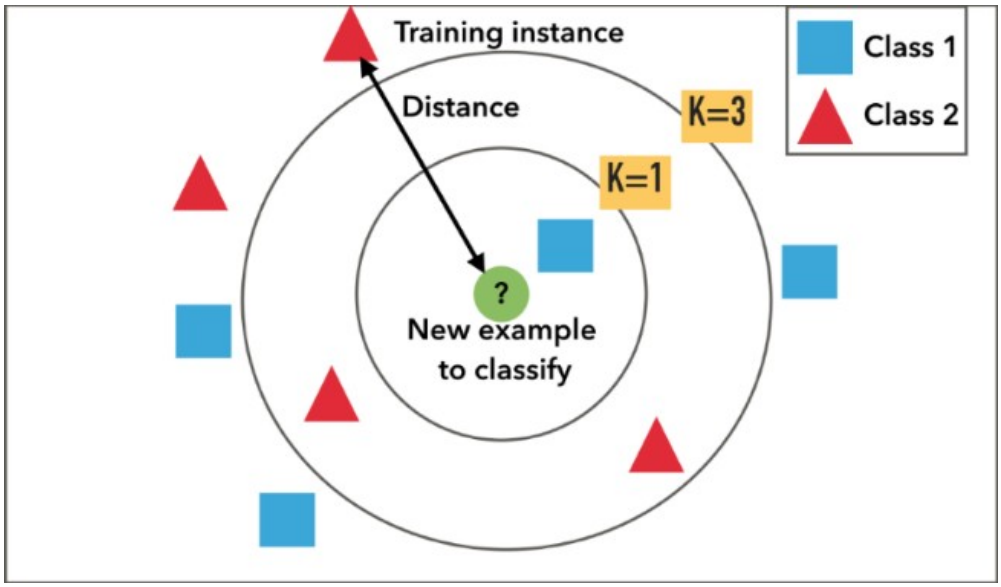
- k-Nearest Neighbours (k-NN)

Ένας από τους πιο απλούς και συνηθισμένους αλγορίθμους ταξινόμησης. Ο k-NN είναι μη παραμετρικός αλγόριθμος, δηλαδή δεν διαθέτει παραμέτρους που επηρεάζουν τη δομή του, ούτε κάνει υποθέσεις για τα δεδομένα, αλλά αντίθετα η δομή του καθορίζεται από τα δεδομένα της

εκπαίδευσης. Πιο συγκεκριμένα, όλα τα δεδομένα εκπαίδευσης αποθηκεύονται σε διανύσματα, απλοποιώντας πάρα πολύ τη διαδικασία εκπαίδευσης και το χρόνο που απαιτείται. Στη συνέχεια, για κάθε νέο στοιχείο-διάνυσμα που πρέπει να ταξινομηθεί, ακολουθείται η παρακάτω διαδικασία. Αρχικά, υπολογίζονται όλες οι αποστάσεις μεταξύ του νέου διανύσματος και όλων των διανυσμάτων εκπαίδευσης, ακολούθως ανιχνεύονται τα k κοντινότερα διανύσματα και ανάλογα με το που ανήκει η πλειοψηφία των k διανυσμάτων ταξινομείται αντίστοιχα και το διάνυσμα-είσοδος. Το k λοιπόν, είναι το πλέον σημαντικό στοιχείο του αλγορίθμου, γιατί ουσιαστικά είναι και αυτό που αποφασίζει πόσους γείτονες χρειάζεται να ληφθούν υπόψιν ώστε να ταξινομηθεί η είσοδος.

Παρακάτω, παρουσιάζεται ένα χαρακτηριστικό παράδειγμα. Έστω ότι υπάρχουν τα δεδομένα δύο κλάσεων, όπου η κλάση 1 παρουσιάζεται με τετράγωνο, η κλάση 2 με τρίγωνο και το διάνυσμα που είναι να ταξινομηθεί με κύκλο, στο κέντρο του σχήματος.

Αν επιλεγεί k=1, δηλαδή επιλεγεί να ταξινομηθεί η είσοδος στην ίδια κλάση με αυτή που βρίσκεται το πιο κοντινό στιγμιότυπο, τότε θα ταξινομηθεί το διάνυσμα στην κλάση 1. Αντίστοιχα, αν επιλεγεί k=3, δηλαδή την κλάση που πλειοψηφεί μεταξύ των 3 πιο κοντινών διανυσμάτων, τότε με ψήφους 2 προς 1 θα επιλεγόταν η κλάση 2 (τα 3 πιο κοντινά στιγμιότυπα στην είσοδο είναι 2 τρίγωνα και 1 τετράγωνο). Με ακριβώς όμοια λογική, αν επιλεγόταν οποιοδήποτε k, θα αντιστοιχούταν το διάνυσμα-είσοδος και στην αντίστοιχη κλάση που προκύπτει.



Ο αλγόριθμος κρίνεται αρκετά εύχρηστος και γρήγορος, πετυχαίνοντας πολύ καλά ποσοστά ορθότητας.

- Gaussian Naive Bayes (GNB)

Ένας αλγόριθμος αρκετά απλός, που στηρίζεται στο θεώρημα Bayes, το οποίο 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:

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)} \quad (2.1)$$

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:

$$X=(x_1,x_2,x_3, \dots , x_n)$$

τότε η πιθανότητα το στοιχείο να ανήκει στην κλάση y, δίνεται από τη σχέση:

$$P(y|x_1, x_{2,} \dots , x_n)=\frac{P(x_1|y)P(x_2|y)\dots P(x_n|y)}{P(x_1)P(x_2)\dots P(x_n)} \quad (2.2)$$

Οι πιθανότητες $P(x_1).P(x_2),\dots,P(x_n)$ και $P(x_1|y),P(x_2|y),\dots,P(x_n|y)$ είναι εύκολο να υπολογιστούν από τα δεδομένα εκπαίδευσης. Με αυτό 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 τη συγκεκριμένη είσοδο. Να αναφερθεί εδώ, πως προϋπόθεση για τη χρήση του Bayes είναι η στατιστική ανεξαρτησία των χαρακτηριστικών μεταξύ τους, γεγονός που είναι αρκετά δύσκολο μεταξύ των features σε ένα dataset.

Ο αλγόριθμος κρίνεται αρκετά εύχρηστος και γρήγορος, καθώς ούτε αυτός διαθέτει παραμέτρους και στηρίζεται αποκλειστικά στο Θεώρημα Bayes. Η ορθότητα του στην κατηγοριοποίηση των στοιχείων χαρακτηρίζεται ικανοποιητική.

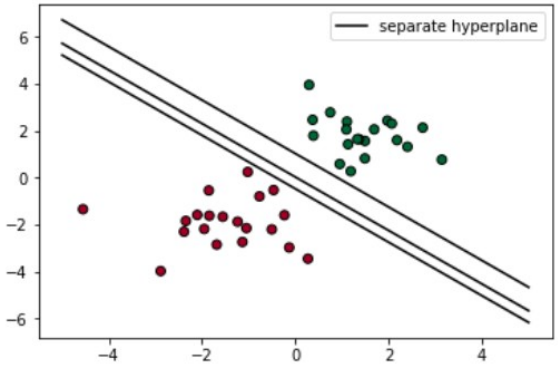
- Support Vector Machines (SVM)

Η βασική ιδέα του αλγορίθμου είναι η εύρεση μιας ευθείας, ενός υπερεπιπέδου γενικότερα, που σκοπό έχει να διαχωρίσει τις κλάσεις μεταξύ τους κατά το βέλτιστο τρόπο, δηλαδή να αυξήσει όσο γίνεται το περιθώριο (margin) ανάμεσα στα δεδομένα διαφορετικών κλάσεων. Αν τα δεδομένα είναι N διαστάσεων, τότε το υπερεπίπεδο θα είναι N-1 διαστάσεων με εξίσωση:

$$\beta_0+\beta_1*x_1+\beta_2*x_2+\dots+\beta_n*x_n=0 \quad (2.3)$$

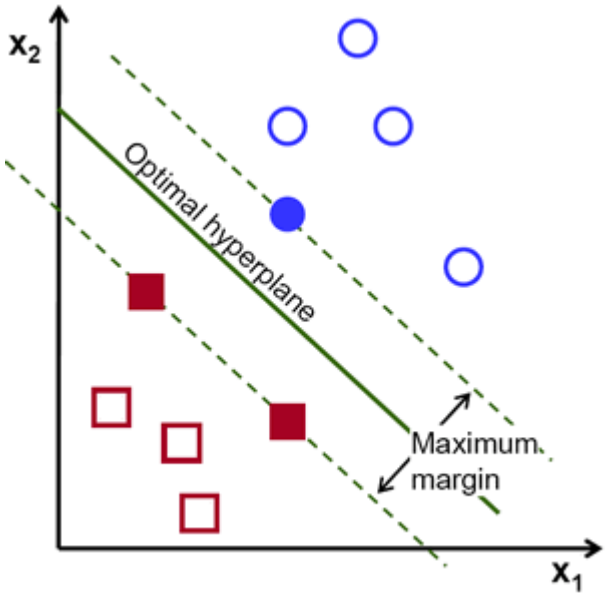
Τα δεδομένα εκπαίδευσης, αν είναι γραμμικώς διαχωρίσιμα, διαχωρίζονται με ένα υπερεπίπεδο και κάθε στοιχείο που είναι προς ταξινόμηση, ελέγχεται σε ποια πλευρά του υπερεπιπέδου βρίσκεται.

Έστω το ακόλουθο παράδειγμα, όπου η μια κλάση παρουσιάζεται με πράσινο χρώμα και η άλλη με κόκκινο. Είναι φανερό ότι, τα δεδομένα είναι γραμμικώς διαχωρίσιμα και μάλιστα υπάρχουν πολλές ευθείες που τα διαχωρίζουν επιτυχώς, όμως μια είναι η βέλτιστη.



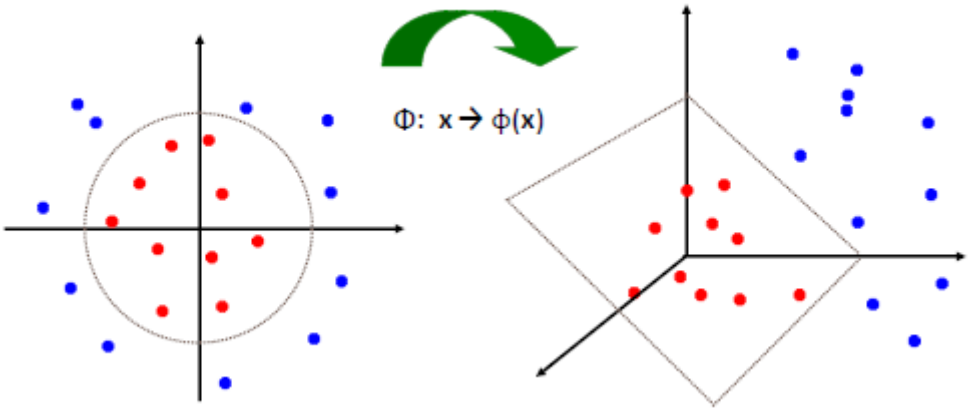
Τα στιγμιότυπα των κλάσεων που αποτελούν τα όρια της κλάσης με την άλλη κλάση αποτελούν τα Support Vectors και βάσει αυτών θα υπολογιστεί το βέλτιστο υπερεπίπεδο διαχωρισμού, με το οποίο θα επιτευχθεί το μεγαλύτερο δυνατό περιθώριο (απόσταση των support vectors με το υπερεπίπεδο), ώστε η ταξινόμηση μετά να έχει ανοχή σε θόρυβο και να ταξινομεί σωστά τα νέα δεδομένα

Παρακάτω παρουσιάζονται με χρωματισμένο το εσωτερικό τους τα support vectors και με συνεχόμενη γραμμή η βέλτιστη ευθεία διαχωρισμού, ενώ είναι φανερό και το περιθώριο (margin) που αναφέρθηκε προηγουμένως.



Στα προηγούμενα παραδείγματα, τα δεδομένα ήταν γραμμικώς διαχωρίσιμα, τι γίνεται όμως αν τα δεδομένα δεν είναι;

Μια λύση δίνει η kernel function (συνάρτηση πυρήνα) RBF (Radial Basis Function), όπου ουσιαστικά η συνάρτηση αυτή απεικονίζει τα δεδομένα σε έναν χώρο μεγαλύτερων διαστάσεων, όπου εκεί τα δεδομένα είναι γραμμικώς διαχωρίσιμα, όπως φαίνεται στην παρακάτω εικόνα.



Στο παράδειγμα αυτό, είναι φανερό πώς τα μη γραμμικώς διαχωρίσιμα δεδομένα, απεικονίστηκαν σε μεγαλύτερη διάσταση και βρέθηκε υπερεπίπεδο που διαχωρίζει τις δύο κλάσεις.

Υπάρχουν και άλλες συναρτήσεις πυρήνα όπως η Linear (γραμμική) και η Poly (πολυωνυμική), ωστόσο στην παρούσα εργασία χρησιμοποιήθηκε η kernel function RBF, καθώς τα δεδομένα του προβλήματος ήταν γραμμικώς μη διαχωρίσιμα.

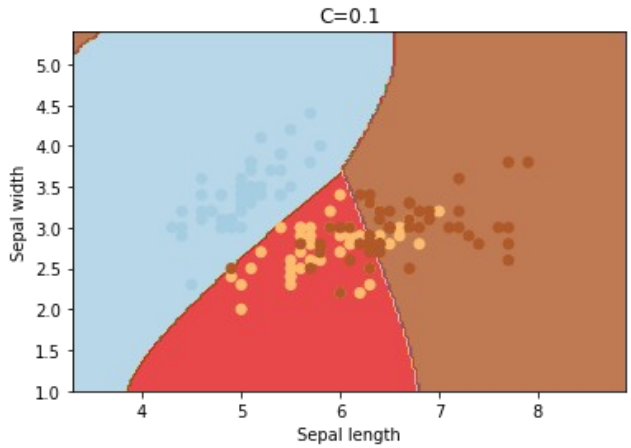
Ο αλγόριθμος SVM σε αντίθεση με τον k-NN και τον GNB, που παρουσιάστηκαν προηγουμένως, διαθέτει πολλές παραμέτρους, η επιλογή των οποίων επηρεάζει προφανώς και την απόδοση του αλγορίθμου. Ο καθορισμός των παραμέτρων και η τεχνική που χρησιμοποιήθηκε για την επιλογή τους,θα αναπτυχθεί εκτενέστερα στο κεφάλαιο με τις εφαρμογές και τα πειράματα.

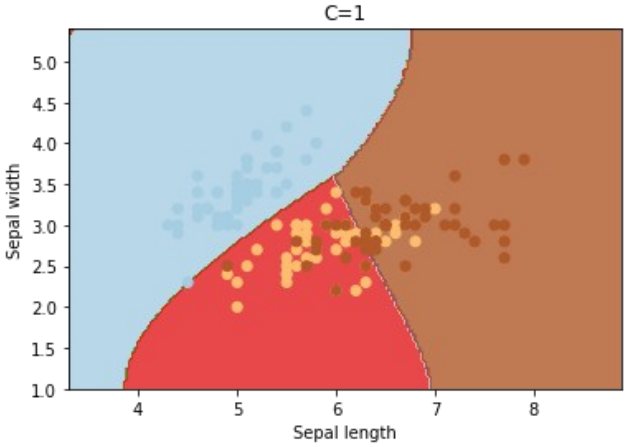
Από τις διαθέσιμες παραμέτρους του αλγορίθμου μόνο οι παρακάτω επιλέχθηκαν για τροποποίηση από την προκαθορισμένη τιμή τους:

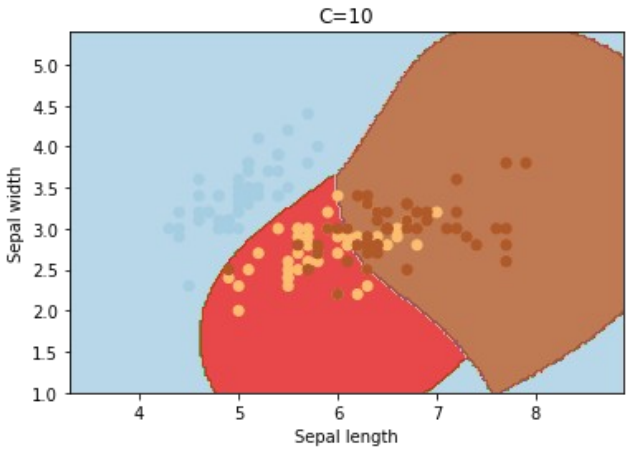
- C

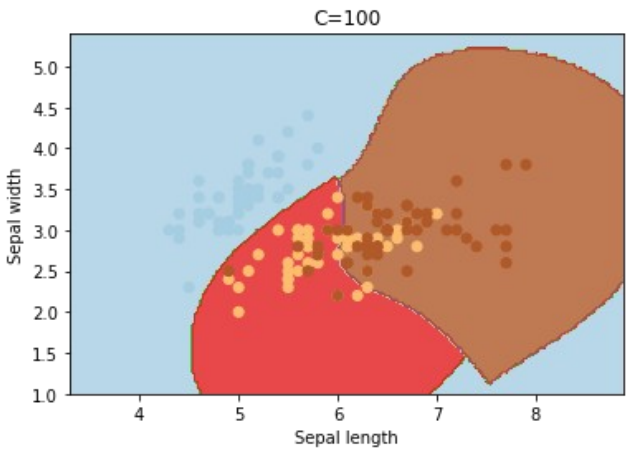
Είναι η ποινή που επιβάλλεται κατά τη λάθος ταξινόμηση ενός στοιχείου κατά την εκπαίδευση. Ελέγχει το όριο μεταξύ χαλαρού ορίου απόφασης και σωστής ταξινόμησης. Προφανώς, μεγάλη τιμή μπορεί να οδηγήσει σε overfiting (υπερεκπαίδευση).

Παρακάτω παρουσιάζονται παραδείγματα για τις διάφορες τιμές του C (0.1, 1, 10, 100) και το πώς ταξινομούνται οι 3 κλάσεις, οι οποίες εμφανίζονται στα σχήματα με γαλάζιο, κίτρινο και καφέ χρώμα. Γίνεται κατανοητό ότι με την αύξηση της τιμής της παραμέτρου C, ο αλγόριθμος γίνεται πιο αυστηρός και προσπαθεί να μειώσει τις λάθος ταξινομήσεις, γεγονός όμως που μπορεί να τον οδηγήσει σε υπερεκπαίδευση, δηλαδή να προσαρμοστεί τέλεια στα δεδομένα εκπαίδευσης και να ταξινομεί σωστά μόνο δεδομένα που είναι όμοια με αυτά της εκπαίδευσης.





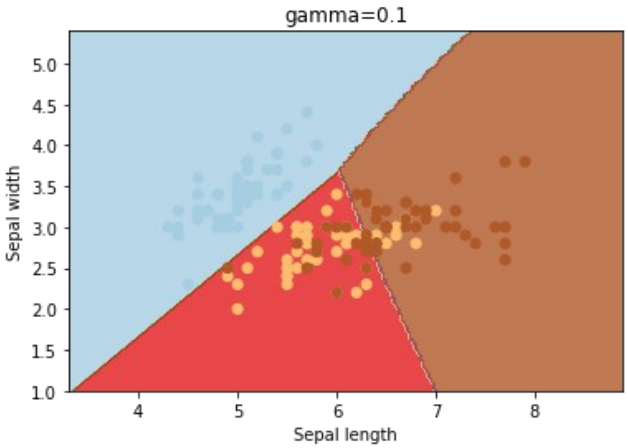


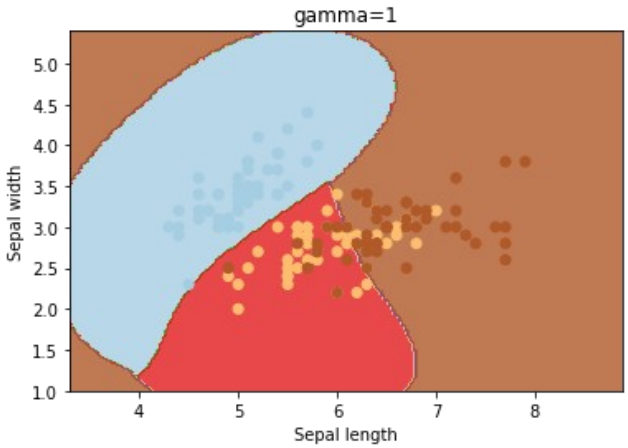


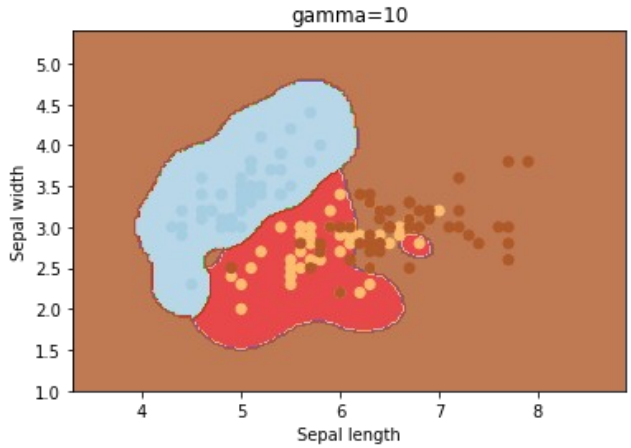
- gamma

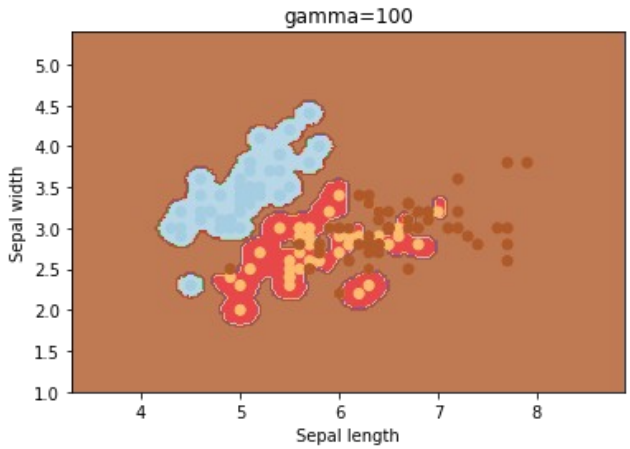
Είναι παράμετρος για μη γραμμικά υπερεπίπεδα και ορίζει την επίδραση που έχει κάθε στοιχείο της εκπαίδευσης. Μεγαλώνοντας την τιμή σε αυτή τη παράμετρο, το μοντέλο προσπαθεί να προσαρμοστεί όσο καλύτερα μπορεί στα δεδομένα εκπαίδευσης. Η παράμετρος μπορεί να οριστεί και ως το αντίστροφο της ακτίνας επιρροής των support vectors (μεγάλη τιμή οδηγεί σε μικρή ακτίνα επιρροής).

Παρακάτω παρουσιάζονται παραδείγματα για τις διάφορες τιμές του gamma (0.1, 1, 10, 100) και το πώς ο αλγόριθμος ταξινομεί τις 3 κλάσεις. Είναι φανερό και εδώ, ότι με αύξηση της τιμής ο αλγόριθμος προσαρμόζεται καλύτερα στα δεδομένα εκπαίδευσης, γεγονός που πάλι μπορεί να οδηγήσει σε overfittng. Ειναι χαρακτηριστικό αυτό στην εικόνα για gamma=100, όπου η ταξινόμηση έχει ταιριάξει τέλεια στα δεδομένα εκπαίδευσης.



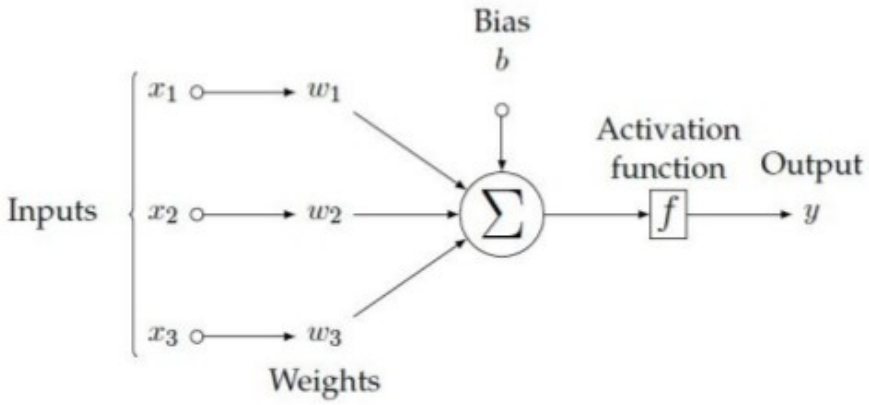






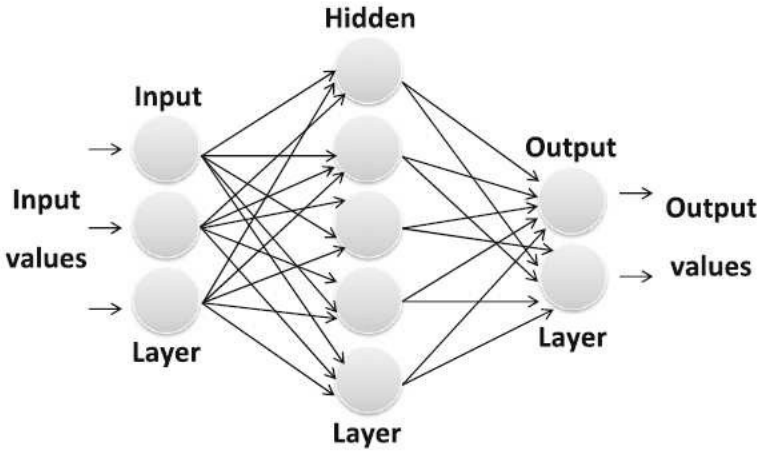
- Multilayer Perceptron (MLP)

Αποτελεί ένα δίκτυο νευρώνων, οι οποίοι δέχονται εισόδους και παράγουν μια έξοδο. Κάθε νευρώνας Perceptron δέχεται n εισόδους και στη συνέχεια κάθε είσοδος πολλαπλασιάζεται με το αντίστοιχο βάρος. Το άθροισμα αυτών των γινομένων δίνεται στη συνάρτηση ενεργοποίησης και παράγεται η έξοδος του νευρώνα, όπως δείχνει και η παρακάτω εικόνα:



Υπάρχουν αρκετές συναρτήσεις που χρησιμοποιούνται για συναρτήσεις ενεργοποίησης, οι πιο συνηθισμένες από αυτές είναι η βηματική, η σιγμοειδής, η υπερβολική εφαπτομένη και η relu.

Ένα δίκτυο τέτοιων νευρώνων ορίζεται σε επίπεδα, όπου κάθε επίπεδο προωθεί την έξοδο του στην είσοδο του επόμενου επιπέδου. Υπάρχει το επίπεδο εισόδου, το κρυφό επίπεδο (που μπορεί και να παραλείπεται) και το επίπεδο εξόδου. Στο επίπεδο εισόδου, οι νευρώνες δέχονται τις εισόδους του συστήματος και προωθούν τις εξόδους τους στο κρυφό επίπεδο (αν υπάρχει), ακολούθως οι νευρώνες του κρυφού επιπέδου παράγουν την έξοδο τους και την προωθούν στο επίπεδο εξόδου, όπου παράγεται και η τελική έξοδος. Να αναφερθεί εδώ, ότι το κρυφό επίπεδο μπορεί να αποτελείται από περισσότερα του ενός επιπέδου. Η λειτουργία αυτή παρουσιάζεται στο παρακάτω σχήμα:



Ο αλγόριθμος MLP δέχεται στο επίπεδο εισόδου τα χαρακτηριστικά ενός στοιχείου που είναι προς ταξινόμηση και στη συνέχεια παράγει ως έξοδο την κλάση στην οποία ταξινομεί το στοιχείο αυτό. Πολύ σημαντικό ρόλο σε αυτή τη διαδικασία παίζουν τα βάρη σε κάθε νευρώνα, τα οποία

διαφέρουν ανά νευρώνα και υπολογίζονται κατά τη διαδικασία της εκπαίδευσης. Μια πολύ συνηθισμένη διαδικασία εκπαίδευσης και ορισμού τω βαρών είναι η backpropagation, όπου τα βάρη ανανεώνονται ανάλογα με την απόσταση της επιθυμητής εξόδου και της πραγματικής μέχρι να συμπληρωθεί ένα πλήθος επαναλήψεων ή μέχρι να μην χρειάζονται άλλες ανανεώσεις βαρών.

Ο συγκεκριμένος αλγόριθμος είναι αρκετά απλός και θεωρείται από τους βασικούς αλγορίθμους ταξινόμησης απλών νευρωνικών δικτύων. Διαθέτει και αυτός αρκετές παραμέτρους, δίνοντας μεγάλη ευελιξία και ελευθερία σε αυτόν που σχεδιάζει το δίκτυο. Θα αναφερθούν και θα αναλυθούν παρακάτω μόνο οι παράμετροι που διαφοροποιήθηκε η τιμή τους από την προκαθορισμένη:

- solver

Η μέθοδος που θα χρησιμοποιηθεί για την βελτιστοποίηση των βαρών. Κάθε μέθοδος ακολουθεί διαφορετική προσέγγιση, μπορεί να είναι στοχαστική ή μπορεί να ανήκει στην οικογένεια μεθόδων Newton.

- alpha

Είναι παράμερος που χρησιμοποιείται για την κανονικοποίηση των βαρών, επιβάλλοντας ποινή σε βάρη με μεγάλη τιμή. Αυξάνοντας την τιμή του alpha μπορεί να οδηγηθεί το σύστημα σε overfitting, καθώς ενθαρρύνονται οι μικρές τιμές στα βάρη και τιμωρούνται οι μεγάλες τιμές.

- hidden layer sizes

Με την παράμετρο αυτή ορίζεται το πλήθος των επιπέδων στο κρυφό επίπεδο και το πλήθος των νευρώνων σε κάθε επίπεδο. Είναι δυνατόν έτσι, να καθοριστεί και η δομή του δικτύου, η οποία μπορεί να είναι αρκετά απλή ή και πολύ σύνθετη, τοποθετώντας πολλά επίπεδα εντός του κρυφού επιπέδου και πολλούς νευρώνες. Αν τα δεδομένα είναι γραμμικώς διαχωρίσιμα δεν χρειάζεται καν η ύπαρξη κρυφού επιπέδου. Τα δεδομένα είναι αυτά που θα καθορίσουν ποιο πρέπει να είναι το μέγεθος του κρυφού επιπέδου και το πλήθος των νευρώνων, ώστε ο αλγόριθμος να επιτυγχάνει τη μέγιστη δυνατή ορθότητα.

- max iter

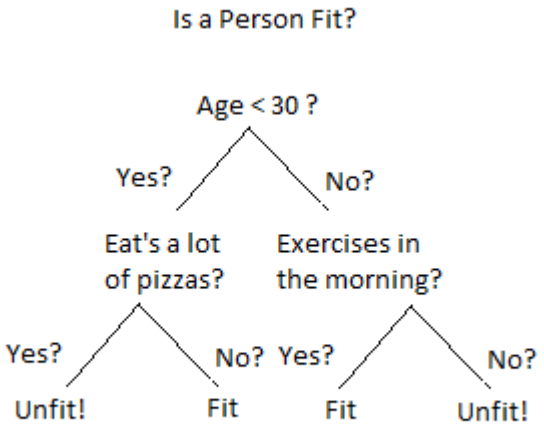
Είναι η παράμετρος που ορίζει τον μέγιστο αριθμό επαναλήψεων του solver (βελτιστοποίηση βαρών), αν δεν έχει τερματίσει αυτός λόγω μη βελτίωσης του score ή του loss παραπάνω από την προκαθορισμένη τιμή της παραμέτρου tol, η οποία είναι 1e-4.

◆ RANDOM FOREST

Ο ταξινομητής Random Forest, όπως αναφέρει και το όνομά του, αποτελεί ένα “δάσος” από Δέντρα αποφάσεων. Ένα Δέντρο απόφασης διαχωρίζει συνεχώς τα αρχικά δεδομένα, σε μικρότερα υποσύνολα μέχρι να φτάσει στους κόμβους φύλλα, όπου γίνεται εκεί η τελική ταξινόμηση σε κλάση ή η λήψη μιας απόφασης.

Στο ακόλουθο παράδειγμα παρουσιάζεται ένα Δέντρο Απόφασης, το οποίο επιθυμεί να ταξινομήσει τους ανθρώπους σε αθλητικούς ή όχι. Αρχικά, διαχωρίζονται τα δεδομένα ανάλογα με την ηλικία των ανθρώπων. Τα άτομα που είναι κάτω από 30 χρονών, διαχωρίζονται ξανά ανάλογα με το αν καταναλώνουν πολλή πίτσα. Αν το άτομο δεν καταναλώνει, τότε κατηγοριοποιείται στην

κλάση fit, ενώ αλλιώς στην κλάση unfit. Αντίστοιχα, τα άτομα τα οποία είναι πάνω από 30 χρονών, διαχωρίζοντα ανάλογα με το αν ασκούνται το πρωί. Τα άτομα που ασκούνται το πρωί, κατηγοριοποιούνται ως fit, ενώ αλλιώς ως unfit.



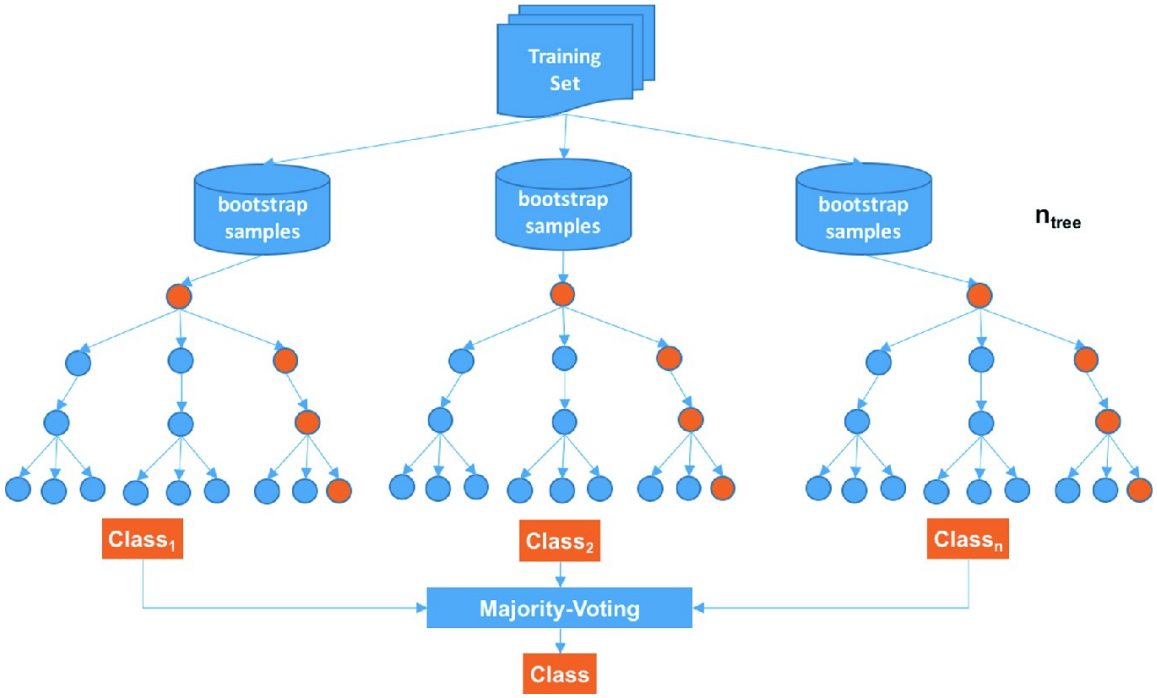
Γίνεται φανερό ότι όσο αναπτύσσεται το δέντρο, τα δεδομένα διαχωρίζονται σε κόμβους μέχρι να φτάσουν στο τέλος, στα φύλλα, όπου εκεί τα δεδομένα κατηγοριοποιούνται σε μία κλάση. Υπάρχουν αρκετά Δέντρα Απόφασης που μπορούν να δημιουργηθούν, με διαφορετικό πλήθος επιπέδων ή διαφορετικό τρόπο διαχωρισμού των δεδομένων, ωστόσο επιλέγεται αυτό που ταιριάζει καλύτερα στα δεδομένα εκπαίδευσης.

Κάθε Δέντρο Απόφασης, του Random Forest, ταξινομεί ένα ζητούμενο σε μία κλάση. Η κλάση που πλειοψηφεί μεταξύ των Δέντρων Αποφάσεων, είναι και η κλάση στην οποία ο αλγόριθμος Random Forest θα ταξινομήσει το ζητούμενο. Ωστόσο, κάθε Δέντρο Απόφασης διαφέρει σημαντικά με τα υπόλοιπα. Αυτό επιτυγχάνεται, καθώς ο αλγόριθμος Random Forest, εκμεταλλεύεται την ευαισθησία των Δέντρων Αποφάσεων, τόσο στα δεδομένα εκπαίδευσης, όσο και στην επιλογή των χαρακτηριστικών (features), με τα οποία γίνεται ο διαχωρισμός των δεδομένων σε κάθε επίπεδο. Κάθε δέντρο Απόφασης εκπαιδεύεται στο ίδιο πλήθος δεδομένων με το αρχικό dataset, επιλέγοντας τυχαία από από τα αρχικά δεδομένα με δικαίωμα επανατοποθέτησης. Δηλαδή, αν το αρχικό σύνολο δεδομένων εκπαίδευσης είναι [1, 2, 3, 4, 5, 6], τότε ένα πιθανό σύνολο εκπαίδευσης για ένα Δέντρο μπορεί να είναι το [1, 1, 2, 3, 5, 5], το οποίο περιέχει ίδιο πλήθος δεδομένων με το αρχικό σύνολο, με κάποια δεδομένα να βρίσκονται παραπάνω από μία φορά. Επιπλέον, κάθε Δέντρο διαλέγει τυχαία ένα υποσύνολο χαρακτηριστικών για τον διαχωρισμό των δεδομένων σε κάθε επίπεδο. Με αυτό τον τρόπο, δημιουργούνται διαφορετικά Δέντρα, που δεν είναι συσχετισμένα μεταξύ τους, παρά το γεγονός ότι τα Δέντρα επέλεξαν δεδομένα και σύνολο χαρακτηριστικών από τα ίδια σύνολα, ενώ κάθε Δέντρο εστιάζει σε διαφορετικά δεδομένα.

Με τη δημιουργία πολλών ανεξάρτητων Δέντρων, εξαλείφεται η προκατάληψη (bias), που μπορεί να υπάρχει στα δεδομένα, καθώς μερικά Δέντρα μπορεί να ταξινομήσουν ένα στοιχείο λάθος, όμως ρόλο δεν παίζει η επιλογή που κάνει ένα δέντρο, αλλά η επιλογή που θα κάνει η πλειοψηφία των Δέντρων. Αξίζει να αναφερθεί εδώ, πως η επιλογή χαρακτηριστικών γίνεται και αυτή στην τύχη.

Στο ακόλουθο παράδειγμα, παρουσιάζεται η λειτουργία του Random Forest με n Δέντρα Αποφάσεων, σχηματικά. Κάθε Δέντρο επιλέγει τυχαία δείγματα και χαρακτηριστικά από το training set με επανατοποθέτηση. Έτσι δημιουργούνται διαφορετικά Δέντρα, με καθένα να κάνει τη δική

του ταξινόμηση. Στο τέλος, τα Δέντρα ψηφίζουν την κλάση που έχουν ταξινομήσει το κάθε στοιχείο. Αυτή που υπερέχει έναντι των άλλων, είναι και η κλάση που επιλέγει ο αλγόριθμος.



Ο Random Forest είναι ένας πολύ ισχυρός ταξινομητής, ο οποίος τα πηγαίνει πολύ καλά σε σωρεία προβλημάτων και δεδομένων. Είναι λογικό, ότι λόγω της σύνθετης φύσης του, θα διαθέτει και πολλές παραμέτρους. Όπως και πριν, θα αναφερθούν παρακάτω μόνο οι παράμετροι που διαφοροποιήθηκαν από την προκαθορισμένη τιμή τους:

- n estimators

Είναι η παράμετρος που ορίζει το πλήθος των Δέντρων Αποφάσεων, που θα λάβει υπόψιν του ο αλγόριθμος. Μεγάλος αριθμός σε αυτή την παράμετρο, οδηγεί σε περισσότερα "ανεξάρτητα" Δέντρα και θεωρητικά πιο ακριβή ταξινόμηση, ωστόσο αυτό οδηγεί και σε μεγάλους χρόνους εκπαίδευσης και προσδιορισμό των παραμέτρων.

- max depth

Είναι η παράμετρος που ορίζει το βάθος που θα έχει κάθε Δέντρο Απόφασης. Προφανώς, όσο μεγαλύτερο και το βάθος του Δέντρου, τόσο περισσότεροι διαχωρισμοί των δεδομένων και περισσότερη πληροφορία αποθηκευμένη για τα δεδομένα εκπαίδευσης.

- min samples split

Είναι η παράμετρος που ορίζει το ελάχιστο πλήθος δειγμάτων που απαιτούνται για τον διαχωρισμό ενός εσωτερικού κόμβου στο Δέντρο Αποφάσεων. Η αύξηση της παραμέτρου αυτής, περιορίζει κάθε Δέντρο, καθώς διατηρεί περισσότερα δείγματα σε κάθε κόμβο και δεν τα

διαχωρίζει κατά τον βέλτιστο δυνατό τρόπο. Χαρακτηριστικό παράδειγμα είναι αν ορίσουμε την παράμετρο ίση με όλα τα δείγματα του συνόλου των δεδομένων εκπαίδευσης, τότε το μοντέλο δεν μαθαίνει από τα δεδομένα και έχουμε πρόβλημα υποεκπαίδευσης.

- min samples leaf

Όμοια με την προηγούμενη παράμετρο, ωστόσο αυτή ορίζει το ελάχιστο πλήθος δειγμάτων που απαιτούνται να βρίσκονται σε έναν κόμβο φύλλο. Όμοια η μεγάλη αύξηση της παραμέτρου αυτής, μπορεί να δημιουργήσει πρόβλημα υποεκπαίδευσης.

- max features

Η παράμετρος αυτή, ορίζει τη συνάρτηση από την οποία θα προκύψει το πλήθος των χαρακτηριστικών, που θα επιλέξει κάθε Δέντρο Απόφασης για τον διαχωρισμό των δειγμάτων σε κόμβους.

- Bootstrap

Αν η παράμετρος πάρει την τιμή False, τότε κάθε Δέντρο Απόφασης θα χρησιμοποιήσει όλο το σύνολο δεδομένων εκπαίδευσης και όχι κάποιο μέρος του. Αυτό επιτυγχάνεται με την κατάργηση της επανατοποθέτησης των δειγμάτων εκπαίδευσης, όπως περιγράφηκε προηγουμένως. Αν η τιμή είναι True, τότε κάθε Δέντρο επιλέγει τα δείγματα με επανατοποθέτηση, δημιουργώντας διαφοροποίηση μεταξύ των Δέντρων, ως προς το σύνολο εκπαίδευσης. Να αναφερθεί ξανά, ότι σε κάθε περίπτωση το πλήθος των δειγμάτων εκπαίδευσης για κάθε Δέντρο είναι ίσο με το αρχικό πλήθος, με τη διαφοροποίηση ότι αν είναι δυνατή η επανατοποθέτηση κατά την επιλογή, κάποια δείγματα υπάρχουν περισσότερες φορές και κάποια άλλα καθόλου.

Κεφάλαιο 3: Εισαγωγή στον ποδοσφαιρικό στοιχηματισμό

3.1 Ποδόσφαιρο και στοίχημα

Το ποδόσφαιρο αποτελεί το λαοφιλέστερο άθλημα στον κόσμο, με τους περισσότερους ανθρώπους να έχουν παρακολουθήσει ή και να έχουν παίξει ποδόσφαιρο τουλάχιστον μια φορά στη ζωή τους. Υπάρχουν αγώνες και διοργανώσεις όπου χιλιάδες φιλάθλων συρρέουν στο γήπεδο για να παρακολουθήσουν έναν αγώνα ποδοσφαίρου. Τι είναι όμως το ποδόσφαιρο και τι το κάνει τόσο αγαπητό;

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 γκολ. Η ομάδα με τα περισσότερα γκολ 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. Το ποδόσφαιρο γνωρίζει τεράστια αγάπη και αναγνώριση ανά τον κόσμο, κυρίως επειδή είναι φθηνό και προσιτό για κάποιον να δει ή να παίξει ποδόσφαιρο, καθώς μπορεί πολύ εύκολα να βρεθεί μια αυτοσχέδια μπάλα και να χωριστεί μια παρέα ατόμων και να παίξει. Ακόμα, καθοριστικό ρόλο παίζει και η άγρια ομορφιά του ποδοσφαίρου, με μεγάλες ανατροπές τα τελευταία λεπτά, αλλά και μεγάλες χαμένες ευκαιρίες, που θα μπορούσαν να αλλάξουν το αποτέλεσμα.

Ενας αγώνας έχει 3 δυνατές εκβάσεις, τη νίκη του γηπεδούχου, την ισοπαλία και τη νίκη του φιλοξενούμενου. Στη γλώσσα του στοιχήματος η νίκη γηπεδούχου λέγεται άσσος (“1”), η ισοπαλία Χ (“χι”) 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 επίπεδο , 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 ενός ανθρώπου να τοποθετήσει ένα ποσό στην πρόβλεψη της εξέλιξης κάποιων αγώνων, που προσφέρονται από εταιρείες στοιχηματισμού έναντι κάποιος 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.

Με το πλήθος των αγώνων όλη τη χρονιά να είναι τεράστιο και με τη μεγάλη αγάπη του κόσμου για το στοίχημα και το άθλημα, συμπεραίνεται ότι ο κόσμος του ποδοσφαίρου κατακλύζεται από εκατομμύρια χρημάτων. Αυτό δεν είναι αυθαίρετο συμπέρασμα, αλλά το δείχνουν και τα στοιχεία, κατατάσσοντας το ποδόσφαιρο ως την τρίτη μεγαλύτερη οικονομία στον κόσμο, με πλούσιους μισθούς παικτών, υπερσύγχρονες εγκαταστάσεις, μεγάλους στοιχηματικούς τζίρους, διαφημίσεις και πολλά άλλα.

Τι είναι όμως η απόδοση στο στοίχημα;

3.2 Αποδόσεις στο στοίχημα-Γκανιότα

Ο άνθρωπος πιστεύει ότι μπορεί να προβλέψει το τελικό σκορ ενός αγώνα, το πλήθος των γκολ, των τρίποντων, των κόρνερ, ακόμα και το ποια ομάδα θα ξεκινήσει το παιχνίδι στη σέντρα. Εκεί έρχονται οι στοιχηματικές εταιρείες και δίνουν μια απόδοση για το ενδεχόμενο που θέλει να προβλέψει ο παίχτης, με τον παίχτη να ποντάρει ένα ποσό στη συγκεκριμένη απόδοση. Αν ο παίχτης κερδίσει, τότε του επιστρέφεται το ποσό που πόνταρε πολλαπλασιασμένο με την απόδοση που έθεσε η στοιχηματική. Για παράδειγμα, έστω ότι ο παίχτης ποντάρει 100 € στη νίκη της ομάδας Α με απόδοση 2,1 , τότε αν η ομάδα Α νικήσει το παιχνίδι, δηλαδή σκοράρει περισσότερα γκολ ή πετύχει περισσότερους πόντους από την αντίπαλο της, τότε θα επιστραφούν στον παίχτη 2,1·100€ =210 €, καθαρό κέρδος δηλαδή για τον παίχτη 210€ - 100€ = 110€. Σε αντίθετη περίπτωση, αν ο παίχτης δεν προβλέψει σωστά δηλαδή, το ποσό που πόνταρε μένει στην εταιρεία που διοργάνωσε το στοίχημα.

Οι αποδόσεις των bookmakers, όπως λέγονται αυτοί που δημιουργούν τις αποδόσεις των εταιρειών, εξαρτώνται από πολλά γεγονότα, αγωνιστικά και μη, όπως για παράδειγμα από τη φόρμα κάθε ομάδας, από την ύπαρξη τραυματισμών ή απουσιών λόγω καρτών, από το αν θα υπάρχει κόσμος στο γήπεδο ή όχι, από το αν υπάρχουν φήμες για προβλήματα μεταξύ των παιχτών ή με τον προπονητή, από την προϊστορία των δύο ομάδων, από την απόσταση που έχει να διανύσει η φιλοξενούμενη ομάδα για το παιχνίδι, ακόμα και από τον διαιτητή, αν για παράδειγμα είναι αρκετά αυστηρός και μοιράζει εύκολα κάρτες και φάουλ, τότε η απόδοση να δοθούν κάρτες ή πολλά φάουλ θα είναι χαμηλή, καθώς αυτό το ενδεχόμενο είναι αρκετά πιθανό, οπότε και η εταιρεία θα το λάβει υπόψιν της. Οι αποδόσεις ενός γεγονότος αλλάζουν συνεχώς, ακόμα και μέρες πριν τον αγώνα μπορεί μια απόδοση να ανέβει ή να πέσει κατακόρυφα. Μπορεί για παράδειγμα, ο κορυφαίος παίχτης της ομάδας Α, να τραυματιστεί και να μην υπολογίζεται για το επόμενο ματς, γεγονός που μπορεί να οδηγήσει σε αλλαγές στο σετ των αποδόσεων, με αύξηση της απόδοσης νίκης της ομάδας Α και μείωση της απόδοσης ήττας της. Ή μπορεί να παρατηρηθούν υψηλά πονταρίσματα σε ένα ενδεχόμενο του αγώνα, πχ η ομάδα Α να σκοράρει παραπάνω από 2 τέρματα ή ο παίχτης Χ να αποβληθεί, τότε οι εταιρείες μειώνουν την αντίστοιχη απόδοση, ώστε να προστατευτούν, καθώς εικάζουν ότι ο κόσμος έχει κάποια πληροφορία που δεν έχουν οι ίδιες ή δεν τη λαμβάνουν υπόψιν τους. Αν συνεχιστούν πολύ υψηλά πονταρίσματα, μπορεί η απόδοση για το ενδεχόμενο αυτό να σταματήσει να είναι διαθέσιμη προς στοιχηματισμό.

Ο παίχτης, ακόμη, δύναται να στοιχηματίσει σε περισσότερα από ένα ενδεχόμενα, ακόμα και του ίδιου γεγονότος, με συνολική απόδοση το γινόμενο των επιμέρους αποδόσεων. Για παράδειγμα, έστω ότι ο παίχτης στοιχηματίζει ότι η ομάδα Α θα νικήσει με απόδοση 1,65, η ομάδα Β θα φέρει ισοπαλία με απόδοση 3,1 και η ομάδα Γ θα σκοράρει πάνω από 2 τέρματα στον αγώνα της με απόδοση 1,80. Έστω ότι ο παίχτης ποντάρει σε αυτό το στοίχημα 10€, τότε αν επαληθευτούν και τα 3 ενδεχόμενα, ο παίχτης θα εισπράξει 1,65·3,1·1,80·10 € = 92,07 €, καθώς οι αποδόσεις πολλαπλασιάζονται μεταξύ τους όπως είπαμε. Όμως, αν έστω και ένα ενδεχόμενο δεν επαληθευτεί, τότε το ποντάρισμα παραμένει στην εταιρεία και ο παίχτης χάνει. Το παραπάνω είναι και λογικό, καθώς ο παίχτης θέλει την τομή των 3 ενδεχομένων, δηλαδή να συμβούν και τα 3 ενδεχόμενα ταυτόχρονα, άρα οι πιθανότητες του μειώνονται και το ρίσκο του παίχτη αυξάνεται, επομένως πρέπει να αυξηθεί και η τελική απόδοση του συγκεκριμένου στοιχήματος. Ο λόγος που έχουμε πολλαπλασιασμό και όχι κάποια άλλη πράξη θα γίνει κατανοητός στη συνέχεια.

Η δημιουργία του σετ των αποδόσεων αποτελεί πολύ σημαντική δουλειά για τις εταιρείες και απαιτεί ιδιαίτερη προσοχή, καθώς το παραμικρό λάθος μπορεί να στοιχίσει χρήματα στην εταιρεία. Κάθε εταιρεία διαθέτει μια πολύ μεγάλη ομάδα ατόμων που ασχολείται ακριβώς με αυτή τη δουλειά, η οποία έχει αυτοματοποιηθεί τα τελευταία χρόνια με τη χρήση σύγχρονων

πληροφοριακών εργαλείων και των συνυπολογισμό πολλών μεταβλητών και πληροφοριών. Ωστόσο, οι bookmakers πρέπει να είναι ιδιαίτερα προσεχτικοί και σχολαστικοί στη δημιουργία αποδόσεων και να ελέγχουν τη διαδικασία. Όπως γίνεται εύκολα αντιληπτό, οι αποδόσεις περιέχουν μεγάλη δόση πληροφορίας, αποτυπώνοντας άμεσα τη γνώμη ενός ειδικού του στοιχήματος. Παρατηρώντας κανείς τις αποδόσεις, ακόμα και να μην γνωρίζει τις ομάδες ή ακόμα και το άθλημα στο οποίο θα αγωνιστούν οι δύο ομάδες, καταλαβαίνει ποια ομάδα λογίζεται ως φαβορί και συνεπώς έχει περισσότερες πιθανότητες για νίκη, αν στο παιχνίδι αναμένονται πολλά φάουλ ή κάρτες και πολλά άλλα. Να αναφέρουμε εδώ, ότι δημοσιεύσεις αναφέρουν ότι μπορούν να παραχθούν προβλέψεις μόνο με τις στοιχηματικές αποδόσεις των ομάδων[1, 9], με ορθότητα λίγο πάνω από το 50%. Αξίζει να αναφέρουμε ότι οι ίδιες δημοσιεύσεις επισημαίνουν τη σημαντικότητα του Asian handicap στην παραγωγή προβλέψεων αγώνων, ενός εναλλακτικού τρόπου στοιχήματος, διαφορετικό από το κλασικό 1X2 του ποδοσφαίρου.

Ποια η σχέση όμως μεταξύ των αποδόσεων των bookmakers και της πιθανότητας που οι ίδιοι πιστεύουν για ενδεχόμενο;

Η σχέση απόδοσης και πιθανότητας είναι αντιστρόφως ανάλογη, δηλαδή

$$\text{πιθανότητα} = \frac{1}{\textit{απόδοση}} \quad (3.1)$$

Για παράδειγμα, έστω ότι έχουμε το ακόλουθο σετ αποδόσεων:

	1	X	2
Απόδοση	2,65	4	2,2

Αυτό σημαίνει ότι η συγκεκριμένη στοιχηματική εταιρεία, χρησιμοποιώντας την παραπάνω σχέση δίνει τις ακόλουθες πιθανότητες:

	1	X	2	Σύνολο
πιθανότητα	37,7%	25%	45,4%	108,1%

Εδώ γίνεται και κατανοητό γιατί πολλαπλασιάζονται οι αποδόσεις μεταξύ τους, όταν ο παίχτης συνδυάζει πολλά ενδεχόμενα μαζί. Τα ενδεχόμενα είναι ανεξάρτητα μεταξύ τους, άρα η πιθανότητα να συμβούν ταυτόχρονα αποτελεί την τομή τους, δηλαδή το γινόμενο των επιμέρους πιθανοτήτων.

Παρατηρούμε στο παράδειγμα ότι οι πιθανότητες αθροίζουν πάνω από το 100%. Η διαφορά από το 100%, δηλαδή το 8,1% στο παραπάνω παράδειγμα, αποτελεί τη γκανιότα που θέτει η εταιρεία και είναι αυτή που εξασφαλίζει και το μακροπρόθεσμο κέρδος για την εταιρεία, δίνοντας σημαντικό πλεονέκτημα στην ίδια έναντι του παίχτη. Υψηλή γκανιότα, σημαίνει και μικρότερες αποδόσεις προς στοιχηματισμό, επομένως απαιτείται και καλύτερη ορθότητα στις προβλέψεις του παίχτη, αν θέλει να έχει μακροπρόθεσμο κέρδος. Η γκανιότα είναι αυτή που ξεχωρίζει κατά κύριο λόγο τις στοιχηματικές εταιρείες μεταξύ τους, μιας και οι υπηρεσίες που προσφέρουν είναι πάνω κάτω ανάλογες. Συνήθως στην ευρωπαϊκή αγορά η γκανιότα κυμαίνεται μεταξύ 4-10%, ενώ στις ασιατικές αγορές η γκανιότα κυμαίνεται στο 2-4%. Να αναφερθεί εδώ, ότι πολλές στοιχηματικές, για να προσελκύσουν περισσότερο κόσμο προσφέρουν μεγάλα αθλητικά γεγονότα με μηδενική γκανιότα.

Για να είναι όμως, ένα σύστημα κερδοφόρο πρέπει

$$\textit{ορθότητα συστήματος} > \frac{1}{\overline{\textit{απόδοση}}} \quad (3.2)$$

ή ισοδύναμα

$$\textit{ορθότητα συστήματος} \cdot \overline{\textit{απόδοση}} > 1$$

δηλαδή θα πρέπει το γινόμενο ορθότητας (accuracy) συστήματος (δηλαδή το κλάσμα του πλήθους των σωστών προβλέψεων προς το σύνολο των προβλέψεων) με τη μέση στοιχηματική απόδοση που προβλέπει επιτυχώς να είναι μεγαλύτερο της μονάδος. Το παραπάνω γινόμενο ισούται με τη μέση μοναδιαία αναμενόμενη επιστροφή χρημάτων, είναι δηλαδή τα χρήματα που αναμένεται να εισπράξει ένα άτομο αν ποντάρει τυχαία σε ένα ματς 1 €. Γίνεται εύκολα αντιληπτό, ότι η γκανιότα επηρεάζει άμεσα και την κερδοφορία του συστήματος, γιατί υψηλή γκανιότα σημαίνει χαμηλές αποδόσεις, απαιτώντας από το σύστημα να αυξήσει την ορθότητα του.

Ένα απλό παράδειγμα παρουσιάζεται παρακάτω, ώστε να γίνουν πιο κατανοητά τα παραπάνω. Έστω ότι υπάρχει ένα δίκαιο κέρμα, τότε οι πιθανότητες για τα δύο ενδεχόμενα ορίζονται ως εξής:

	Κορώνα	Γράμματα	Σύνολο
Πιθανότητα	50%	50%	100%

Οι αποδόσεις όμως που θα έθετε μια στοιχηματική ορίζονται ανάλογα με τη γκανιότα ως εξής:

Κορώνα	Γράμματα	Γκανιότα	Ελάχιστη ορθότητα
2	2	0%	50%
1,9	1,9	5,2%	52,6%
1,8	1,8	11,1%	55,5%

Με ελάχιστη ορθότητα ορίζουμε την ακρίβεια εκείνη του συστήματος που μακροπρόθεσμα δεν θα δημιουργήσει ούτε κέρδος , αλλά και ούτε ζημία. Πάνω από την ακρίβεια αυτή, το σύστημα αρχίζει και οδηγεί μακροπρόθεσμα σε κέρδος.

3.3 Αξιολόγηση συστημάτων

Ένα σύστημα το οποίο προβλέπει ένα ενδεχόμενο ενός αθλητικού γεγονότος και στη συνέχεια η πρόβλεψη του χρησιμοποιείται προς στοιχηματισμό, δεν είναι σωστό να αξιολογηθεί μόνο από την ορθότητα του στις προβλέψεις του, αλλά με το αν μακροπρόθεσμα οδηγεί σε κερδοφορία. Η κερδοφορία εξαρτάται, όπως αναφέρθηκε προηγουμένως, και από την ορθότητα του συστήματος, αλλά και από τη μέση απόδοση που έχουν τα σημεία που προβλέπει το σύστημα. Θα πρέπει δηλαδή να ικανοποιείται η συνθήκη (3.2). Ας δούμε το επόμενο παράδειγμα, έστω ότι έχουμε τους ακόλουθους αγώνες με τις αντίστοιχες αποδόσεις για κάθε σημείο και τις προβλέψεις δύο συστημάτων:

Κωδικός αγώνα	1	X	2	Αποτέλεσμα

101	**1,5**	4	7	1
102	2,1	3	**4**	2
103	4,1	**2,8**	1,9	X
104	**1,2**	5,5	8,5	1
105	**2,3**	3,5	2,8	1
106	6	**4,5**	1,4	X
107	**2,3**	3,2	2,9	1

Κωδικός αγώνα	Σύστημα Α	Σύστημα Β
101	**1**	X
102	1	**2**
103	2	1
104	**1**	X
105	**1**	2
106	2	**X**
107	**1**	X

Το σύστημα Α προβλέπει σωστά 4 στα 7 παιχνίδια, δηλαδή έχει ορθότητα (accuracy) 57,1% με μέση απόδοση των επιτυχημένων προβλέψεων του 1,825. Επομένως, το σύστημα Α αναμένουμε να βγάλει κέρδος σε αυτή τη σειρά αγώνων, γιατί ικανοποιείται η συνθήκη (3.2) . Αν σε κάθε παιχνίδι ποντάραμε 10€ στην πρόβλεψη που μας δίνει το σύστημα, τότε:

$K_A = 10 * 1,5 + 10 * 1,2 + 10 * 2,3 + 10 * 2,3 – 10 * 7 = 73 – 70 = 3$ €.
Πράγματι, το σύστημα Α είναι κερδοφόρο και αφήνει καθαρό κέρδος 3 €.

Το σύστημα Β προβλέπει εύστοχα 2 στα 7 παιχνίδια, δηλαδή έχει ορθότητα 28,5% με μέση απόδοση στις επιτυχημένες προβλέψεις του 4,25. Επομένως, και το σύστημα Β αναμένεται να βγάλει κέρδος σε αυτή τη σειρά αγώνων, γιατί ικανοποιείται και εδώ η συνθήκη (3.2). Αν σε κάθε παιχνίδι, ποντάραμε πάλι από 10€ στην πρόβλεψη που μας δίνει το σύστημα, τότε:

$K_B = 10 * 4 + 10 * 4,5 – 10 * 7 = 85 – 70 = 15$ €.
Πράγματι, το σύστημα Β είναι και αυτό κερδοφόρο και μάλιστα αφήνει καθαρό κέρδος 15 €.

Παρατηρούμε ότι το σύστημα Β, παρότι παρουσιάζει πολύ χαμηλότερη ορθότητα από το σύστημα Α (μάλιστα παρουσιάζει ακριβώς τη μισή ορθότητα από το σύστημα Α), καταφέρνει να επιτύχει σημαντικά μεγαλύτερο κέρδος. Όπως γίνεται κατανοητό, σημασία δεν έχει μόνο η ορθότητα του μοντέλου, αλλά και η δυσκολία των προβλέψεων που επιτυγχάνει. Στο παράδειγμα μας, το σύστημα Α προέβλεπε συστηματικά τη μικρότερη απόδοση, δηλαδή το ενδεχόμενο με την μεγαλύτερη πιθανότητα να επαληθευτεί σύμφωνα με τους bookmakers (3.1). Το σύστημα Β αντίθετα, προέβλεπε ενδεχόμενα με υψηλή απόδοση, δηλαδή χαμηλή πιθανότητα να εμφανιστούν, για το λόγο αυτό παρά τη χαμηλή ευστοχία του κατάφερε να επιτύχει υψηλότερο καθαρό κέρδος. Επομένως, μεταξύ των δύο συστημάτων, παρότι και τα δύο οδηγούν σε κέρδος σε αυτό το μικρό δείγμα αγώνων, το σύστημα Β κρίνεται καλύτερο, γιατί πετυχαίνει μεγαλύτερο καθαρό κέρδος.

Σε αυτό το παράδειγμα, ήταν εύκολο να απαντήσουμε στο ερώτημα ποιο σύστημα είναι καλύτερο, γιατί τα δύο συστήματα συγκρίθηκαν στο ίδιο πλήθος αγώνων και στο ίδιο συνολικό ποντάρισμα (70 € και στις δύο περιπτώσεις). Τι θα γινόταν αν τώρα εμφανιζόταν ένα σύστημα Γ με τις ακόλουθες προβλέψεις;

Αγώνας	101	102	103	104	105	106	107
Σύστημα Γ	No bet	No bet	No bet	**1**	**1**	No bet	X

Το σύστημα Γ παράγει προβλέψεις μόνο για 3 αγώνες και έχει ορθότητα 66,6% με μέση απόδοση στις στις επιτυχημένες προβλέψεις του 1,75. Ικανοποιείται πάλι η συνθήκη (3.2), επομένως και το σύστημα Γ, αναμένεται να είναι κερδοφόρο. Αν τώρα σε κάθε πρόβλεψη του συστήματος Γ ποντάραμε 20 €, μιας και δεν προβλέπει πολλά παιχνίδια, τότε:

$K_Γ$=20 * 1,2 + 20 * 2,3 – 20 * 3 = 70 – 60 = 10 €
Πράγματι, το σύστημα Γ είναι και αυτό κερδοφόρο, με καθαρό κέρδος 10 €.

Ποιο σύστημα είναι καλύτερο τώρα, το Α, το Β ή το Γ; Τα Α και τα Β, μπορούμε να τα συγκρίνουμε, γιατί συγκρίθηκαν στο ίδιο πλήθος αγώνων και με το ίδιο συνολικό ποντάρισμα, με το σύστημα Β, όπως είπαμε, να υπερέχει. Όμως, το σύστημα Γ δεν μπορεί να συγκριθεί άμεσα ούτε με το σύστημα Α, ούτε με το σύστημα Β, γιατί έχουμε μία ουσιαστική διαφοροποίηση, διαφορετικό ποντάρισμα (60 έναντι 70 €). Για το λόγο αυτό θα ορίσουμε εδώ δύο νέες έννοιες, το ROI (Return On Investment) και το Yield.

Οι δείκτες ROI και Yield χρησιμοποιούνται ευρέως στο στοίχημα και δείχνουν μακροπρόθεσμα την οικονομική πορεία του παίχτη και είναι δανεισμένοι από την επιστήμη των Οικονομικών. Οι δύο αυτοί όροι είναι αρκετά κοντινοί, όμως έχουν διαφορές και δεν θα πρέπει να συγχέονται. Πιο αναλυτικά:

- **ROI** είναι όπως δείχνει και το όνομα του δείκτη η επιστροφή της επένδυσης. Αποτελεί τη σχέση μεταξύ μεικτών εσόδων προς το συνολικό ποντάρισμα του παίχτη. Η σχέση με την οποία υπολογίζεται είναι:

$$ROI = \frac{Μεικτά\ έσοδα}{Συνολικό\ ποντάρισμα} \cdot 100\,\% \quad (3.3)$$

Προφανώς, αν η τιμή είναι μεγαλύτερη του 100%, τότε ο δείκτης δείχνει ότι ο παίχτης είχε κέρδος. Ενώ αν ο δείκτης είναι μικρότερος, ο παίχτης είχε απώλειες. Για παράδειγμα, αν ένα παίχτης έχει μεικτά έσοδα 150 € και το συνολικό του ποντάρισμα είναι 120 €, τότε το ROI= 125%, δηλαδή ο παίχτης αν πόνταρε το ίδιο ποσό σε όλα τα στοιχήματα του, ποντάρωντας συνολικά 100€, τότε θα είχε κέρδος 25 €. Αντίθετα, αν το ROI προέκυπτε 80%, τότε ο παίχτης αν πάλι έκανε ποντάρισμα 100 €, θα είχε ζημία 20 €

- **Yield** είναι ο δείκτης των καθαρών κερδών ή ζημίας προς το συνολικό ποντάρισμα. Αυτός ο δείκτης είναι πιο απλός και χρησιμοποιείται περισσότερο. Η σχέση με την οποία υπολογίζεται είναι:

$$Yield = \frac{Κέρδη\ ή\ Ζημία}{Συνολικό\ ποντάρισμα} \cdot 100\,\% \quad (3.4)$$

Προφανώς, θετικό Yield δείχνει κέρδος για τον παίχτη, ενώ σε αντίθετη περίπτωση δείχνει ζημία. Για παράδειγμα, αν ένα παίχτης ποντάρει συνολικά 120 €, με μεικτά έσοδα 150 €, όπως και στο προηγούμενο μας παράδειγμα, θα έχει καθαρό κέρδος 30 €, δηλαδή Yield=25%, δηλαδή αν ο παίχτης ποντάρει 100 €, τότε το καθαρό κέρδος θα είναι 25 €

Παρατηρούμε ότι οι δύο αυτοί δείκτες είναι αρκετά κοντινοί, απαλλαγμένοι από το ύψος των πονταρισμάτων, όμως παρουσιάζουν αρκετές διαφορές, για παράδειγμα το ROI=50% απέχει πολύ από το Yield=50 %. Στην πρώτη περίπτωση, έχει απομείνει το μισό κεφάλαιο από το συνολικό που διέθεσε ο παίχτης για τα πονταρίσματα του, ενώ στη δεύτερη περίπτωση ο παίχτης έχει κέρδος που ισούται με το μισό κεφάλαιο του συνολικού πονταρίσματος του. Τέλος, αξίζει να αναφερθεί ότι, ένα σύστημα ή άνθρωπος που προβλέπει αγώνες, θεωρείται αρκετά καλό αν έχει Yield της τάξεως 7-8% και περιζήτητο αν έχει Yield 10+%.

Η σχέση που συνδέει το Yield και το ROI είναι η ακόλουθη:

$$ROI = 100\% + Yield \quad (3.5)$$

Ας επιστρέψουμε στο παράδειγμα μας, ποιο σύστημα είναι καλύτερο, το Α,το Β ή το Γ;
Αφού ορίστηκαν οι παραπάνω δείκτες, μπορούμε πλέον τα συγκρίνουμε άμεσα τα συστήματα μας. Ακολούθως υπολογίζεται το Yield κάθε συστήματος:

- $Yield_A = \frac{3}{70} * 100\% = +4{,}28\%$
- $Yield_B = \frac{15}{70} * 100\% = +\mathbf{21{,}42\%}$
- $Yield_Γ = \frac{10}{60} * 100\% = +16{,}66\%$

Είναι φανερό ότι το σύστημα Β, έχει μεγαλύτερο Yield, επομένως κρίνεται καλύτερο από τα άλλα δύο συστήματα. Προφανώς, θα καταλήγαμε και στο ίδιο συμπέρασμα αν υπολογίζαμε και τον δείκτη ROI.

Να αποσαφηνίσουμε ότι το παραπάνω αποτελεί ένα σύντομο παράδειγμα σύγκρισης συστημάτων. Είναι αδύνατο να βγάλουμε ασφαλή συμπεράσματα όσον αφορά ποιο σύστημα είναι καλύτερο σε τόσο μικρό δείγμα αγώνων, κανονικά απαιτούνται τουλάχιστον μερικές εκατοντάδων αγώνων. Σε επόμενα κεφάλαια που θα χρειαστεί να αξιολογηθούν και να συγκριθούν συστήματα μεταξύ τους, θα χρησιμοποιηθούν οι παραπάνω δείκτες και πιο συγκεκριμένα ο δείκτης Yield.

Κεφάλαιο 4: Εξαγωγή χαρακτηριστικών (Feature Engineering)

4.1 Μελέτη χαρακτηριστικών

Ο κόσμος του ποδοσφαίρου περιβάλλεται από αυτόν του στοιχήματος, έτσι λοιπόν η πληροφορία είναι πλούσια και εύκολη προσβάσιμη σε όλους. Επομένως, ήταν εύκολο να βρεθούν πολλά και χρήσιμα στατιστικά και πληροφορίες για τους αγώνες του Αγγλικού Πρωταθλήματος που μελετά η παρούσα εργασία. Η ιστοσελίδα https://www.footballx-data.co.uk παρέχει πληροφορίες για όλους τους αγώνες των μεγάλων πρωταθλημάτων τα τελευταία χρόνια με ημερομηνίες, τελικό σκορ, σκορ ημιχρόνου, στατιστικά αγώνα, όπως φάουλ, κάρτες, τελικές προσπάθειες και άλλα, αποδόσεις για διάφορα σημεία των αγώνων από μεγάλες στοιχηματικές εταιρείες, το όνομα του Διαιτητή που διηύθυνε τον αγώνα κλπ, σε μορφή .csv. Τα δεδομένα βρίσκονται οργανωμένα σε γραμμές, όπου κάθε γραμμή αναφέρεται σε έναν αγώνα και περιέχει όλα τα στοιχεία του. Μάλιστα, δεν παρατηρούνται πολλά missing values, παρά μόνο μερικά σε κάποιες αποδόσεις κάποιων στοιχηματικών εταιρειών, τις οποίες δεν θα χρησιμοποιήσουμε σε αυτή την εργασία.

Για την καλύτερη οργάνωση των δεδομένων δημιουργήθηκε ένα jupyter notebook, το οποίο διατήρησε από το αρχικό dataset μόνο τα στοιχεία που κρίθηκαν σημαντικά, αλλά και δημιούργησε νέα χαρακτηριστικά για κάθε αγώνα, χρησιμοποιώντας δεδομένα από τη χρονιά 2007 έως και το 2018 (12 σαιζόν δηλαδή). Μερικά, από τα νέα χαρακτηριστικά που δημιουργήθηκαν είναι η βαθμολογία κάθε ομάδας πριν τον αγώνα, στατιστικά για κάθε ομάδα, όπως μέσο όρο γκολ που σκοράρει, διαφορά γκολ υπέρ και κατά, φόρμα εντός και εκτός έδρας στα τελευταία παιχνίδια. Επιπλέον, συνδυάζοντας τα στατιστικά κάθε ομάδας δημιουργήθηκαν και νέα features με τις διαφορές των στατιστικών κάθε ομάδας, όπως η διαφορά βαθμολογικής συγκομιδής στα τελευταία ματς, η διαφορά στο σκοράρισμα, η διαφορά στη δημιουργία προσπαθειών στον στόχο κλπ. Ακόμη, διατηρήθηκαν από το αρχικό dataset οι στοιχηματικές αποδόσεις για τα 3 πιθανά σημεία ενός αγώνα (νίκη γηπεδούχου, ισοπαλία και ήττα γηπεδούχου) από 5 μεγάλες εταιρείες και δημιουργήθηκαν νέα δεδομένα, όπως η μέγιστη απόδοση για κάθε σημείο, η διαφορά μέγιστης και ελάχιστης απόδοσης για το ίδιο σημείο από τις εταιρείες, ενώ διατηρήθηκαν και οι αποδόσεις για το asian handicap από την bet365. Τέλος, προστέθηκαν και οι οικονομικές αξίες των ρόστερ των ομάδων, καθώς και οι διαφορές τους, όπως αυτές δίνονται από την ανεξάρτητη ιστοσελίδα https://www.transxfermarkt.com/

Αξίζει να σημειωθεί πως στα δεδομένα μας έχουμε κρατήσει μόνο τους αγώνες από την $10^{η}$ αγωνιστική της Premier League μέχρι και την $32^{η}$. Οι λόγοι που μας ώθησαν στην παραπάνω ενέργεια είναι δύο. Ο πρώτος είναι πρακτικός, καθώς για να συλλέξουμε τη φόρμα κάθε ομάδας (5 τελευταίοι αγώνες ανεξαρτήτου έδρας και 3 τελευταίοι αγώνες εκτός έδρας και 3 τελευταίοι εκτός), απαιτείται να ολοκληρωθούν πρώτα τουλάχιστον 6-7 αγωνιστικές. Ο δεύτερος είναι εμπειρικός, καθώς η εμπειρία έχει δείξει ότι οι ομάδες παρουσιάζουν μη φυσιολογικά αποτελέσματα, στην μεν αρχή του πρωταθλήματος με τη δικαιολογία ότι δεν έχουν φορμαριστεί και βρει ρυθμό, ενώ τις τελευταίες αγωνιστικές γιατί είναι αδιάφορες, καθώς μόνο μερικές ομάδες έχουν απομείνει να κυνηγούν τις πρώτες θέσεις ή να μάχονται για την παραμονή στην κατηγορία.

Στον παρακάτω πίνακα παρουσιάζονται αναλυτικά τα χαρακτηριστικά που επιλέξαμε από το αρχικό dataset, αλλά και αυτά που δημιουργήσαμε:

ID	Σύντομη περιγραφή

0	Πλήθος αγώνων της ομάδας εντός έδρας
1	Σύνολο βαθμών του συνόλου των αγώνων της ομάδας εντός έδρας
2	Σύνολο βαθμών του συνόλου των εντός έδρας αγώνων της ομάδας εντός έδρας
3	Σύνολο βαθμών των τελευταίων 5 αγώνων ανεξαρτήτου έδρας της ομάδας εντός έδρας
4	Σύνολο βαθμών των τελευταίων 3 αγώνων εντός έδρας της ομάδας εντός έδρας
5	Διαφορά τερμάτων υπέρ και κατά της ομάδας εντός έδρας
6	Διαφορά τερμάτων υπέρ και κατά για τα τελευταία 5 ματς της εντός έδρας ομάδας
7	Διαφορά τερμάτων υπέρ και κατά των τελευταίων 3 εντός έδρας αγώνων της εντός έδρας ομάδας
8	Μέσος όρος γκολ υπέρ της εντός έδρας ομάδας ανά αγώνα
9	Μέσος όρος γκολ υπέρ των εντός έδρας ματς της εντός έδρας ομάδας ανά εντός έδρας αγώνα
10	Μέσος όρος γκολ κατά της εντός έδρας ομάδας ανά αγώνα
11	Μέσος όρος γκολ κατά των εντός έδρας ματς της εντός έδρας ομάδας ανά εντός έδρας αγώνα
12	Αποτέλεσμα τελευταίου αγώνα της εντός έδρας ομάδας
13	Αποτέλεσμα τελευταίου εντός έδρας αγώνα της εντός έδρας ομάδας
14	Μέσος όρος σουτ της εντός έδρας ομάδας ανά αγώνα
15	Σύνολο σουτ των τελευταίων 3 εντός έδρας αγώνων της εντός έδρας ομάδας
16	Μέσος όρος σουτ στον στόχο της εντός έδρας ομάδας ανά αγώνα
17	Σύνολο σουτ στον στόχο των τελευταίων 3 εντός έδρας αγώνων της εντός έδρας ομάδας
18	Μέσος όρος κόρνερ της εντός έδρας ομάδας ανά αγώνα
19	Σύνολο κόρνερ των τελευταίων 3 εντός έδρας αγώνων της εντός έδρας ομάδας
20	Πλήθος αγώνων της ομάδας εκτός έδρας
21	Σύνολο βαθμών του συνόλου των αγώνων της ομάδας εκτός έδρας
22	Σύνολο βαθμών του συνόλου των εκτός έδρας αγώνων της ομάδας εκτός έδρας
23	Σύνολο βαθμών των τελευταίων 5 αγώνων ανεξαρτήτου έδρας της ομάδας εκτός έδρας
24	Σύνολο βαθμών των τελευταίων 3 αγώνων εκτός έδρας της ομάδας εκτός έδρας
25	Διαφορά τερμάτων υπέρ και κατά της ομάδας εκτός έδρας
26	Διαφορά τερμάτων υπέρ και κατά για τα τελευταία 5 ματς της εκτός έδρας ομάδας
27	Διαφορά τερμάτων υπέρ και κατά των τελευταίων 3 εκτός έδρας αγώνων της εκτός έδρας ομάδας
28	Μέσος όρος γκολ υπέρ της εκτός έδρας ομάδας ανά αγώνα

29	Μέσος όρος γκολ υπέρ των εκτός έδρας ματς της εκτός έδρας ομάδας ανά εκτός έδρας αγώνα
30	Μέσος όρος γκολ κατά της εκτός έδρας ομάδας ανά αγώνα
31	Μέσος όρος γκολ κατά των εκτός έδρας ματς της εκτός έδρας ομάδας ανά εκτός έδρας αγώνα
32	Σύνολο βαθμών από τον τελευταίο αγώνα της εκτός έδρας ομάδας
33	Σύνολο βαθμών από τον τελευταίο εκτός έδρας αγώνα της εκτός έδρας ομάδας
34	Μέσος όρος σουτ της εκτός έδρας ομάδας ανά αγώνα
35	Σύνολο σουτ των τελευταίων 3 εκτός έδρας αγώνων της εκτός έδρας ομάδας
36	Μέσος όρος σουτ στον στόχο της εκτός έδρας ομάδας ανά αγώνα
37	Σύνολο σουτ στον στόχο των τελευταίων 3 εκτός έδρας αγώνων της εκτός έδρας ομάδας
38	Μέσος όρος κόρνερ της εκτός έδρας ομάδας ανά αγώνα
39	Διαφορά ποσοστού των βαθμών που έχει μαζέψει η εντός έδρας ομάδα και αυτών της εκτός έδρας ομάδας προς το σύνολο των βαθμών που θα μπορούσαν να είχαν μαζέψει οι ομάδες
40	Βαθμολογική διαφορά των ομάδων
41	Μέση διαφορά βαθμών που έχει μαζέψει η γηπεδούχος ομάδα εντός έδρας και αυτών που έχει μαζέψει η φιλοξενούμενη εκτός έδρας
42	Διαφορά βαθμών που έχουν μαζέψει οι δύο ομάδες στα τελευταία 5 ματς ανεξαρτήτου έδρας
43	Διαφορά βαθμών που έχει μαζέψει η γηπεδούχος ομάδα στα τελευταία 3 εντός έδρας ματς και των βαθμών που έχει μαζέψει η φιλοξενούμενη στα τελευταία 3 εκτός έδρας ματς
44	Διαφορά βαθμών που έχει μαζέψει η γηπεδούχος ομάδα από τον τελευταίο εντός έδρας αγώνα της και των βαθμών που έχει μαζέψει η φιλοξενούμενη ομάδα από τον τελευταίο εκτός έδρας αγώνα της
45	Διαφορά βαθμών που έχουν μαζέψει οι 2 ομάδες από τον τελευταίο αγώνα τους ανεξαρτήτου έδρας
46	(6)-(26)
47	(5)-(25)
48	Διαφορά γκολ υπέρ της εντός έδρας ομάδας των 3 τελευταίων εντός έδρας αγώνων της και της εκτός έδρας ομάδας των τελευταίων 3 εκτός έδρας αγώνων της
49	Διαφορά γκολ υπέρ της εντός έδρας ομάδας των εντός έδρας αγώνων της και της εκτός έδρας ομάδας των εκτός έδρας αγώνων της
50	Συνολική διαφορά γκολ υπέρ των 2 ομάδων
51	Συνολική διαφορά στα σουτ των 2 ομάδων
52	Διαφορά των σουτ της γηπεδούχου ομάδας εντός έδρας και της φιλοξενούμενης εκτός έδρας

53	Συνολική διαφορά στα σουτ στον στόχο των 2 ομάδων
54	Διαφορά των σουτ στον στόχο της γηπεδούχου ομάδας εντός έδρας και της φιλοξενούμενης εκτός έδρας
55	Συνολική διαφορά στα κόρνερ των 2 ομάδων
56	Διαφορά των κόρνερ της γηπεδούχου ομάδας εντός έδρας και της φιλοξενούμενης εκτός έδρας
57	Απόδοση νίκης γηπεδούχου από bet365
58	Απόδοση ισοπαλίας από bet365
59	Απόδοση νίκης φιλοξενούμενου από bet365
60	Μέγιστη απόδοση νίκης γηπεδούχου από 5 μεγάλες στοιχηματικές εταιρείες
61	Μέγιστη απόδοση ισοπαλίας από 5 μεγάλες στοιχηματικές εταιρείες
62	Μέγιστη απόδοση νίκης φιλοξενούμενου από 5 μεγάλες στοιχηματικές εταιρείες
63	Τύπος asian handicap
64	Μέση απόδοση νίκης γηπεδούχου με asian handicap
65	Μέση απόδοση νίκης φιλοξενούμενου με asian handicap
66	(60)-(57)
67	(61)-(58)
68	(62)-(59)
69	Πιθανότητα νίκης γηπεδούχου από bet365
70	Πιθανότητα ισοπαλίας από bet365
71	Πιθανότητα νίκης φιλοξενούμενου από bet365
72	Αξία γηπεδούχου ομάδας από το transfermarkt
73	Αξία φιλοξενούμενης ομάδας από το transfermarkt
74	(72)-(73)
75	Τελικό αποτέλεσμα αγώνα (1 για νίκη γηπεδούχου, 2 για ισοπαλία και 3 για νίκη φιλοξενούμενου)

Τα χαρακτηριστικά, όπως εύκολα γίνεται αντιληπτό, έχουν ομαδοποιηθεί σε 5 σύνολα:

- 0-19: Χαρακτηριστικά που αφορούν την γηπεδούχο ομάδα
- 20-38: Χαρακτηριστικά που αφορούν τη φιλοξενούμενη ομάδα
- 39-56: Χαρακτηριστικά με τις διαφορές των δύο ομάδων
- 57-71: Χαρακτηριστικά από τις στοιχηματικές εταιρείες για τον αγώνα
- 72-74: Χαρακτηριστικά που αφορούν τις αξίες των 2 ομάδων

Να σημειώσουμε πως όπου αναφέρονται διαφορές μεταξύ ενός χαρακτηριστικού της γηπεδούχου ομάδας και της φιλοξενούμενης, εννοείται χαρακτηριστικό γηπεδούχου μείον χαρακτηριστικό φιλοξενούμενης.

4.2 Συσχέτιση και συντελεστής συσχέτισης Pearson r

Με την ύπαρξη πολλών χαρακτηριστικών και μεταβλητών είναι σίγουρο ότι κάποια από αυτά θα συσχετίζονται, δηλαδή θα υπάρχει κάποια σχέση μεταξύ τους. Όπως για παράδειγμα, η υψηλή θερμοκρασία συσχετίζεται με τη συχνή χρήση aircondition σε μια πόλη ή την μεγάλη κατανάλωση σε παγωτά. Από την άλλη, είναι πολύ δύσκολο να μην υπάρχει κάποια συσχέτιση και οι δύο μεταβλητές να μην συνδέονται με κάποια σχέση.

Είναι σημαντικό όμως, να εντοπισθούν και να κατανοηθούν οι συσχετίσεις μεταξύ των δεδομένων για τη σωστή χρήση τους στα μοντέλα. Για το λόγο αυτό, θα χρησιμοποιηθεί βιβλιοθήκη της Python, που υπολογίζει τις γραμμικές συσχετίσεις όλων των μεταβλητών μεταξύ τους.

Έστω ότι έχουμε δύο μεταβλητές X,Y με n στοιχεία και θέλουμε να υπολογίσουμε κατά πόσο συσχετίζονται μεταξύ τους. Θα χρησιμοποιηθεί ο συντελεστής Pearson, όπου αποτελεί ένα μέτρο συσχέτισης ανεξάρτητο των μονάδων μέτρησης. Ο συντελεστής Pearson ή συντελεστής r προκύπτει από τη διαίρεση της συνδιακύμανσης των δύο μεταβλητών (πώς δηλαδή μεταβάλλονται οι δύο μεταβλητές μεταξύ τους) προς το γινόμενο των τυπικών αποκλίσεων σ_X και σ_Y.

Δηλαδή:

$$r = \frac{\sum_i^n (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_i^n (x_i - \bar{X})^2 \sum_i^n (y_i - \bar{Y})^2}} \quad (4.1)$$

,όπου xi, y_i είναι κάθε τιμή της μεταβλητής X και Y, ενώ $\bar{X}, \bar{Y}$ είναι οι μέσες τιμές των δύο μεταβλητών.

Ο συντελεστής Pearson δεν έχει μονάδες και παίρνει τιμές στο διάστημα [-1,+1]. Επομένως, για δύο μεταβλητές θα ισχύει $-1 \leq r \leq +1$. Για τις διάφορες τιμές του r έχουμε:

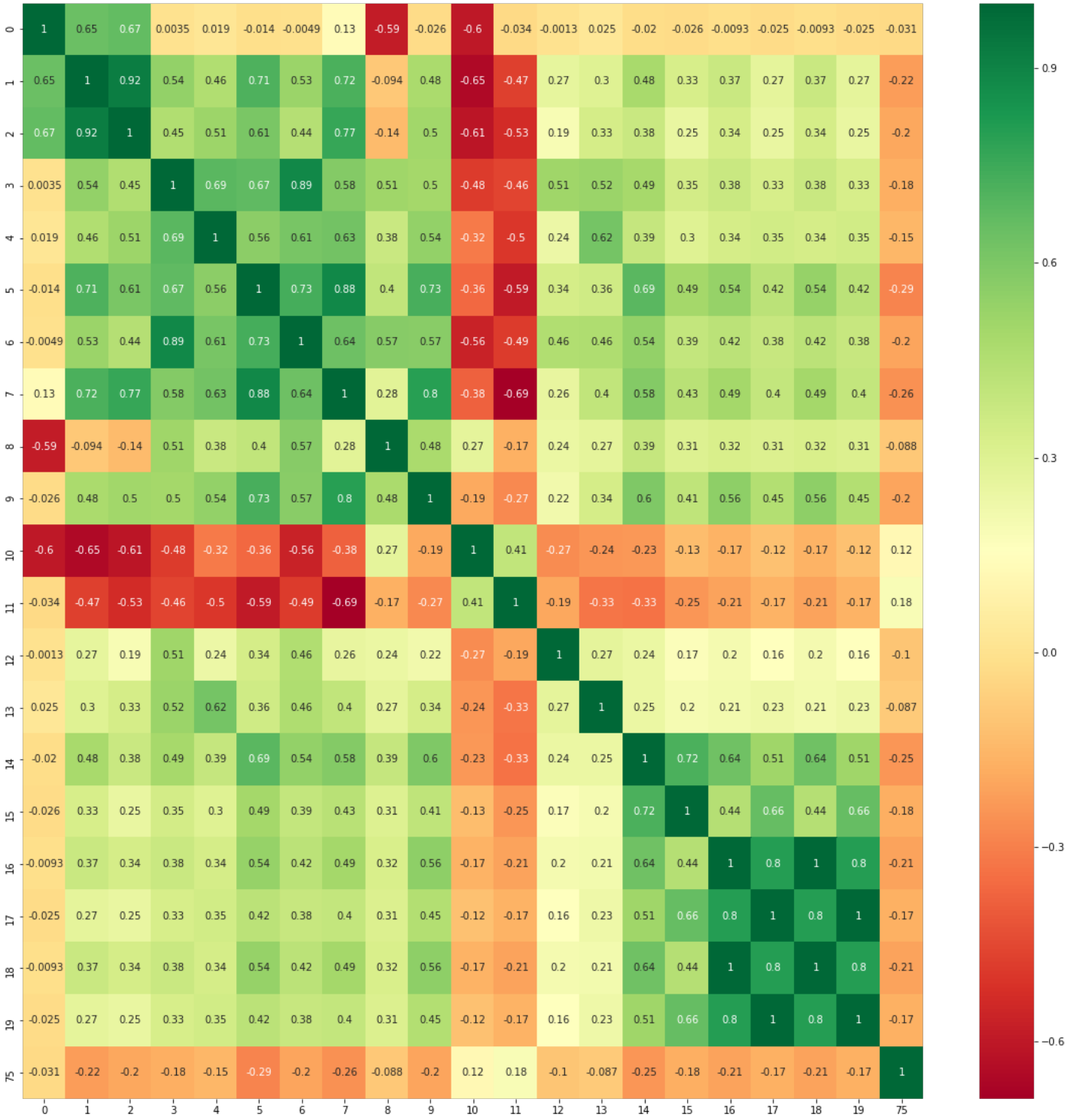
- αν $r = \pm 1$, τότε υπάρχει γραμμική συσχέτιση των δεδομένων.
- αν $r = 0$, τότε τα δεδομένα είναι πλήρως ανεξάρτητα.
- αν $r > 0$, τότε τα δεδομένα είναι θετικά συσχετισμένα, δηλαδή η αύξηση του ενός συνεπάγεται και αύξηση του άλλου.
- αν $r < 0$, τότε τα δεδομένα είναι αρνητικά συσχετισμένα, δηλαδή η αύξηση του ενός συνεπάγεται μείωση του άλλου.

Χρησιμοποιώντας βιβλιοθήκη της Python είναι πολύ εύκολο να αποτυπώσουμε σε δισδιάστατο χάρτη τις συσχετίσεις όλων των μεταβλητών μεταξύ τους. Μας ενδιαφέρει κυρίως η συσχέτιση κάθε χαρακτηριστικού με το τελικό αποτέλεσμα αγώνα, που είναι και η πρόβλεψη που θέλουμε να κάνουμε. Εννοείται ότι τα στοιχεία της διαγωνίου θα αποτελούνται από 1, καθώς κάθε στοιχείο της διαγωνίου αποτελεί την συσχέτιση κάθε μεταβλητής με τον εαυτό της. Αξίζει να παρατηρήσουμε και τις συσχετίσεις των χαρακτηριστικών μεταξύ τους, καθώς η εισαγωγή χαρακτηριστικών στην εκπαίδευση του συστήματος, που είναι αρκετά συσχετισμένα προσδίδουν θόρυβο στο σύστημα και μειώνουν την απόδοση του.

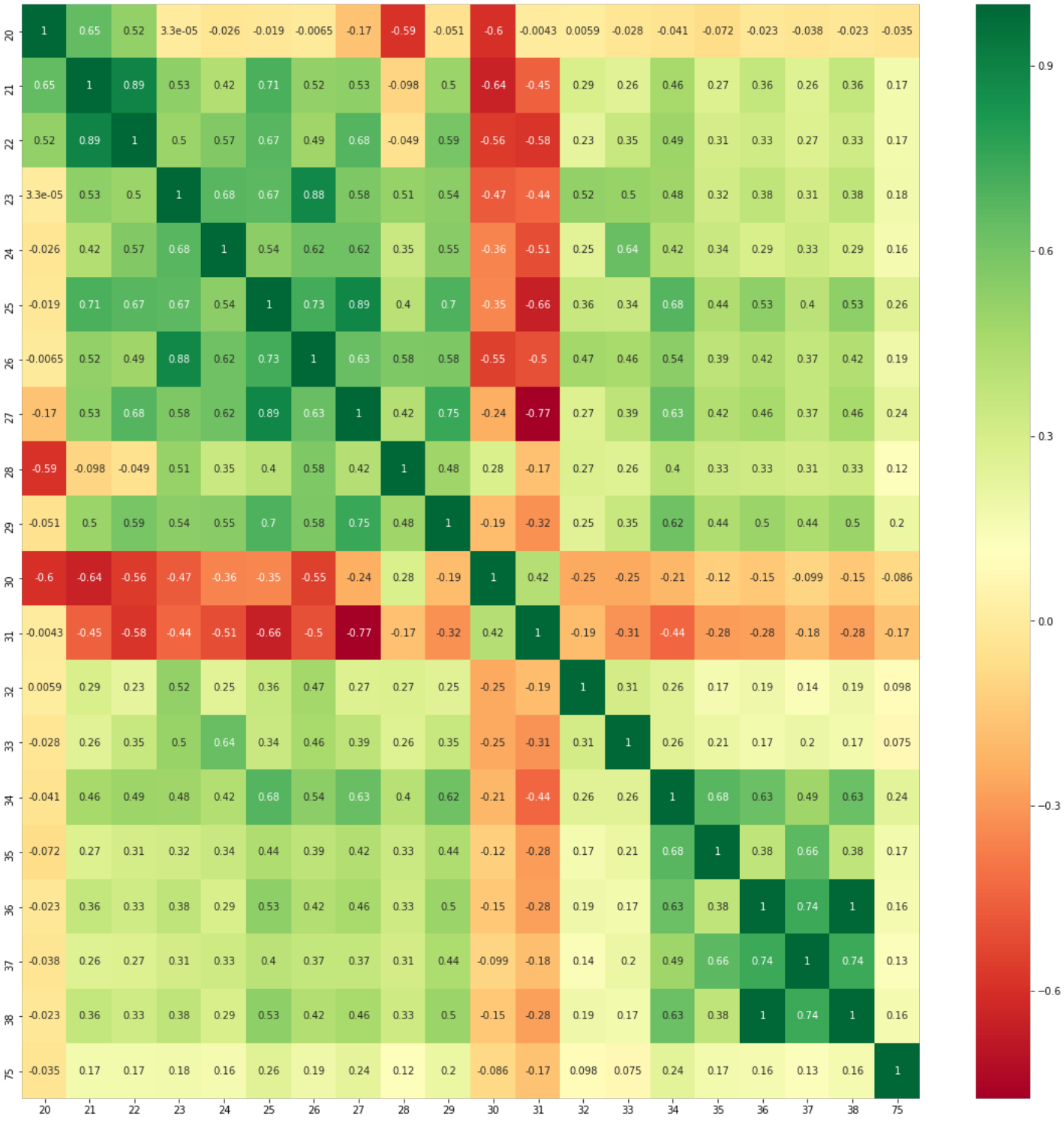
Για λόγους καλύτερης παρουσίασης του χάρτη συσχετίσεων, θα παρουσιαστούν οι επιμέρους χάρτες κάθε συνόλου που ορίσαμε προηγουμένως μαζί με το χαρακτηριστικό στόχο, το τελικό αποτέλεσμα. Με πράσινο χρώμα παρουσιάζονται οι θετικές αυτοσυσχετίσεις, ενώ με κόκκινο οι αρνητικές. Το σκούρο χρώμα υποδηλώνει ότι οι δύο μεταβλητές έχουν υψηλή αυτοσυσχέτιση και αντίθετα όσο πιο ανοιχτό χρώμα, τόσο χαμηλότερη αυτοσυσχέτιση. Στην τελευταία γραμμή παρουσιάζεται η συσχέτιση κάθε χαρακτηριστικού με το χαρακτηριστικό στόχο.

Παρακάτω, παρουσιάζονται οι πίνακες συσχέτισης Pearson κάθε συνόλου χαρακτηριστικών.

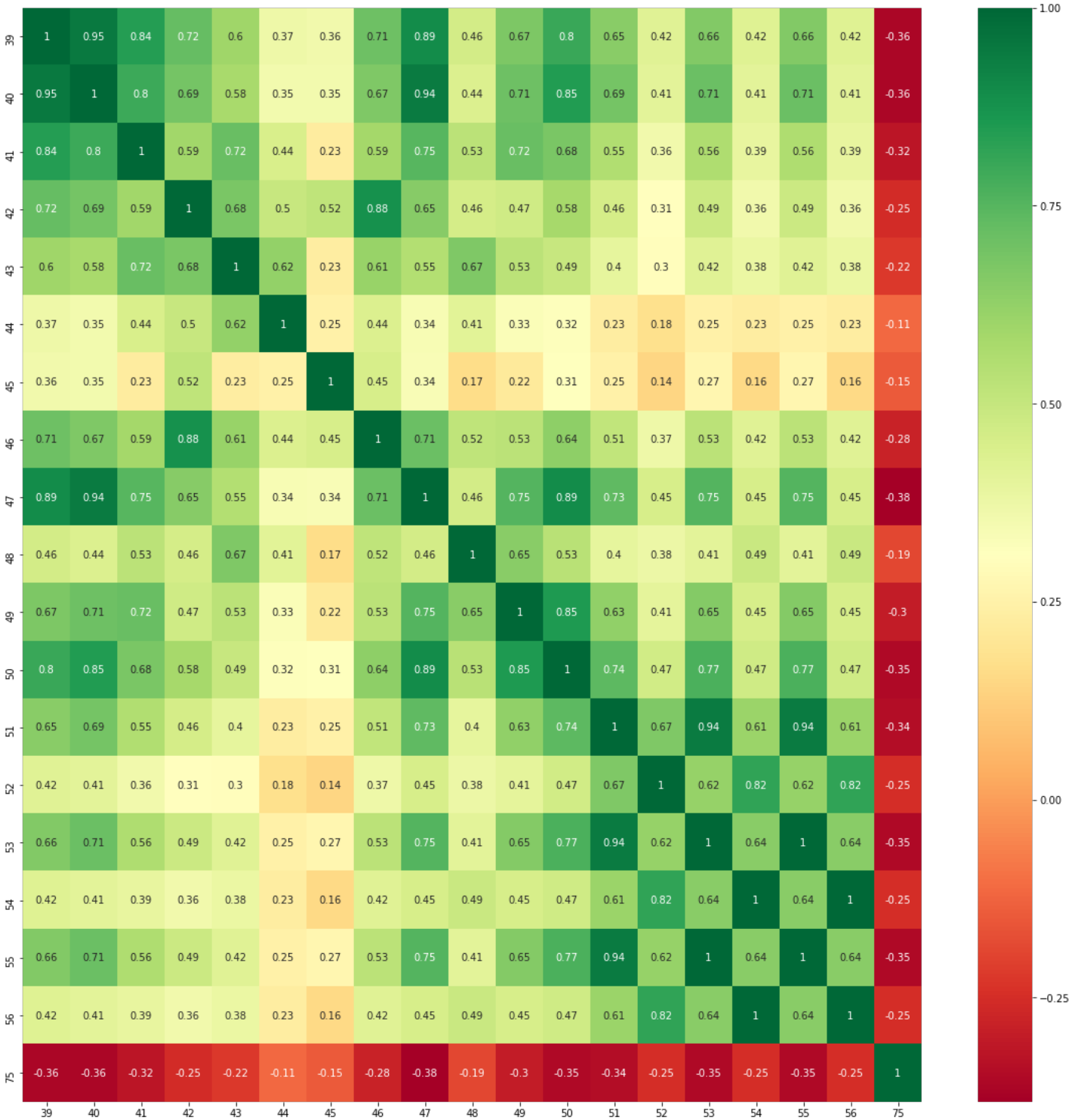
- Χαρακτηριστικά γηπεδούχου (0-19)



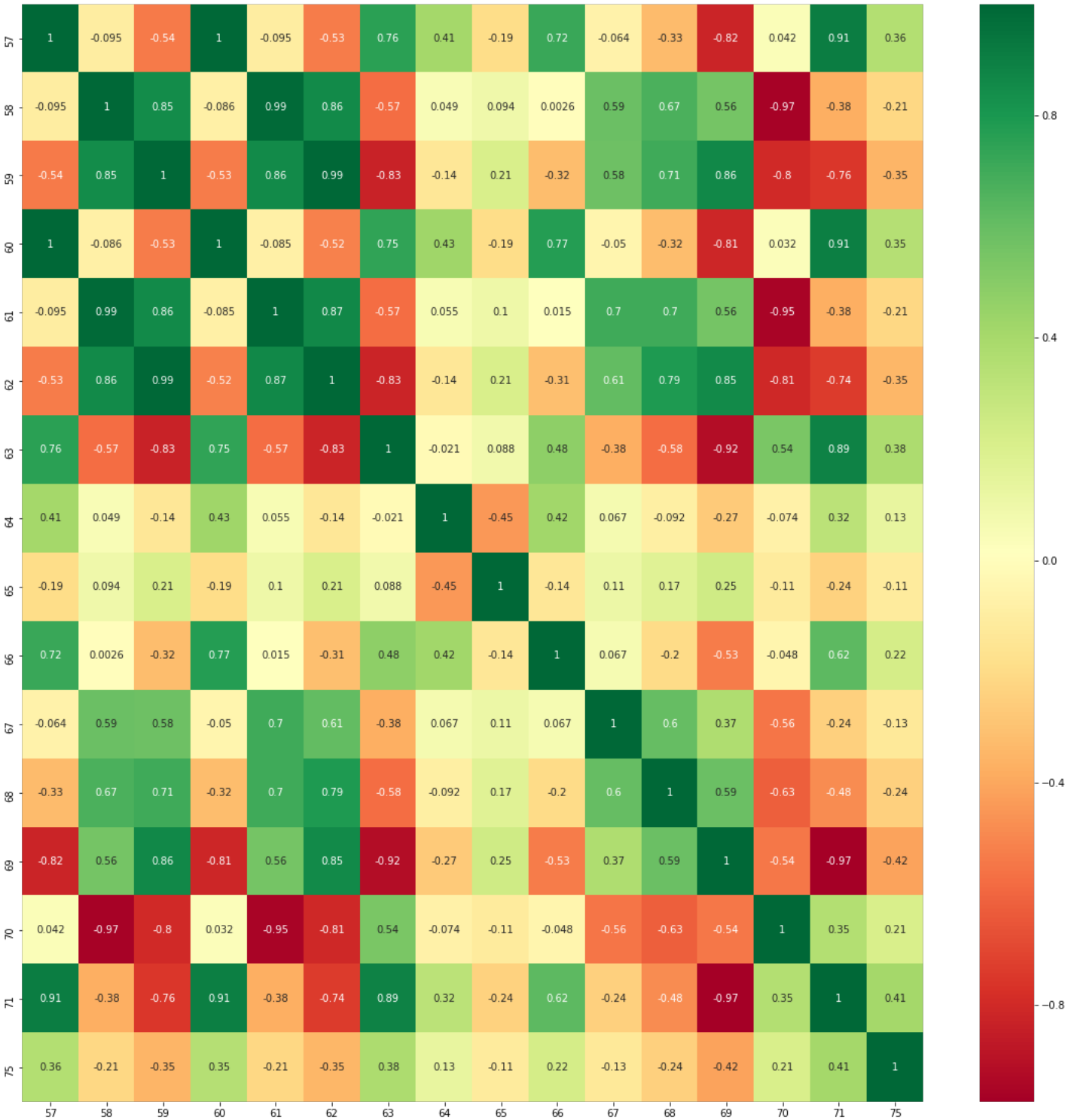
- Χαρακτηριστικά φιλοξενούμενης ομάδας (20-38)



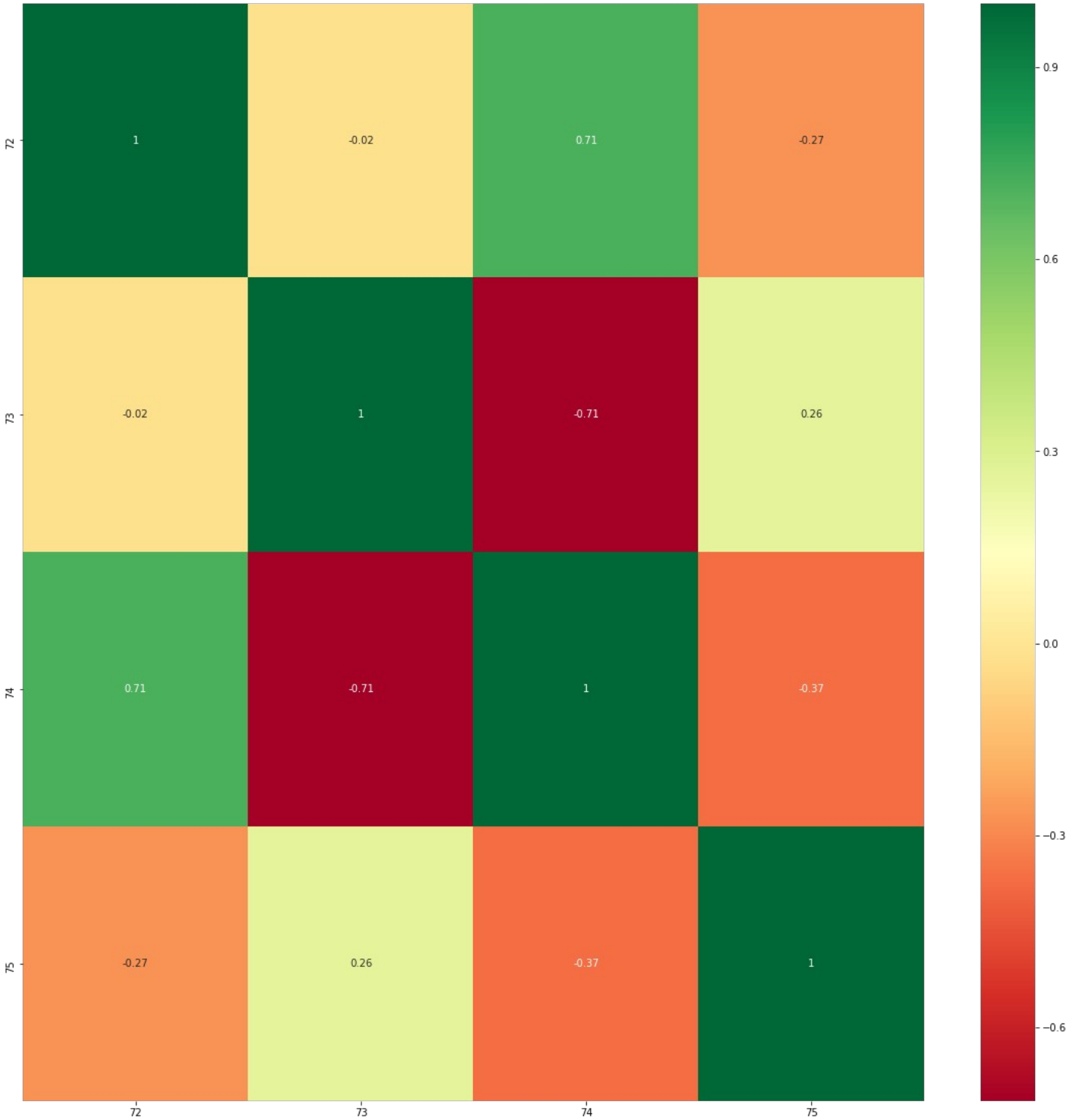
- Χαρακτηριστικά με τις διαφορές των δύο ομάδων (39-56)



- Χαρακτηριστικά που αφορούν τις στοιχηματικές αποδόσεις (57-71)



- Χαρακτηριστικά που αφορούν τις αξίες των ομάδων (72-75)



Με αυτό τον τρόπο, εποπτικά, προκύπτουν αρκετά συμπεράσματα για το κάθε χαρακτηριστικό ξεχωριστά, αλλά και για τις ομάδες χαρακτηριστικών, κοιτώντας σε κάθε χάρτη την τελευταία γραμμή ή στήλη. Όπως αναμενόταν, για τα χαρακτηριστικά 0-19, που περιγράφουν τα χαρακτηριστικά της γηπεδούχου ομάδος, όσο αυξάνουν, πχ όσα περισσότερα γκολ πετυχαίνει η ομάδα ή όσο περισσότερους βαθμούς μαζεύει στα παιχνίδια της, είναι λογικό το τελικό αποτέλεσμα να παίρνει και χαμηλότερη τιμή, δηλαδή να τείνει προς το 1, πράγμα που αποτυπώνεται με το κόκκινο χρώμα. Θυμίζουμε εδώ, πως η νίκη γηπεδούχου αντιστοιχεί στον αριθμό 1, η ισοπαλία στον αριθμό 2 και η νίκη του φιλοξενούμενου στο 3. Αντίστοιχα, στα χαρακτηριστικά που περιγράφουν τον φιλοξενούμενο, όπως αναμενόταν η αύξηση τους οδηγεί και σε αύξηση του χαρακτηριστικού του αποτελέσματος, καθώς αυτό δείχνει ότι η φιλοξενούμενη ομάδα είναι αρκετά δυνατή, έτσι είναι λογικό και το αποτέλεσμα να είναι υπέρ της, για το λόγο αυτό και στον χάρτη αποτυπώνονται με πράσινο χρώμα. Με όμοια λογική, και η κατηγορία με τις διαφορές των χαρακτηριστικών των δύο ομάδων, καθώς αύξηση της τιμής τους δείχνει υπεροχή του γηπεδούχου και άρα είναι λογικό το αποτέλεσμα να είναι 1 ή 2 (νίκη γηπεδούχου ή ισοπαλία) . Με όμοιο τρόπο σκέψης, προκύπτουν αντίστοιχα συμπεράσματα και στις υπόλοιπες κατηγορίες. Παρατηρεί κανείς εύκολα, ότι η κατηγορία χαρακτηριστικών με τις διαφορές των δύο ομάδων παρουσιάζεται με περισσότερο σκούρο χρώμα σε σχέση με τις υπόλοιπες, επομένως, όπως ήταν και αναμενόμενο, αποτελεί και την σημαντικότερη κατηγορία. Ενώ τέλος, χαρακτηριστικά που είναι πανομοιότυπα ή πολύ κοντινά παρουσιάζουν υψηλή συσχέτιση, όπως για παράδειγμα η απόδοση νίκης γηπεδούχου (57) και η πιθανότητα νίκης γηπεδούχου (69), που ουσιαστικά περιγράφουν το ίδιο ενδεχόμενο με άλλον τρόπο.

Τα κύρια χαρακτηριστικά που ξεχώρισαν από τους χάρτες, αλλά και από τις δοκιμές στα συστήματα που έγιναν είναι τα ακόλουθα *7, 25, 39, 41, 47, 53, 57, 59, 63, 71,* τα οποία και αποτέλεσαν τη βάση στα συστήματα μας. Ωστόσο, μετά από και άλλες δοκιμές που έγιναν, κάποια συστήματα απέδιδαν καλύτερα με άλλα χαρακτηριστικά και έτσι, υπήρξαν κάποιες προσθαφαιρέσεις.

Επιπλέον, για κάθε αλγόριθμο δοκιμάστηκε και η κανονικοποίηση του dataset, καθώς αρκετοί αλγόριθμοι παρουσιάζουν βελτιωμένη απόδοση σε κανονικοποιημένα δεδομένα, τα οποία είναι και πιο εύκολα να συγκριθούν μεταξύ τους. Σε κάθε περίπτωση, επιλέχθηκε η επιλογή που αλγόριθμος πετυχαίνει καλύτερη απόδοση στο validation set.

Κεφάλαιο 5: Πειραματικά αποτελέσματα

5.1 Εισαγωγή στα πειραματικά αποτελέσματα

Στο παρόν κεφάλαιο θα εξεταστούν τα μοντέλα και τα αποτελέσματα τους στο επιλεγμένο dataset. Τα δεδομένα από τις season 2007-08 μέχρι και τη 2018-19 του Αγγλικού Πρωταθλήματος Ποδοσφαίρου, όπως αναφέρθηκε και σε προηγούμενο κεφάλαιο, ανακτήθηκαν από την ιστοσελίδα https://www.football-data.co.uk σε μορφή .csv και στη συνέχεια, με τη βοήθεια jupyter notebook που δημιουργήθηκε για τις ανάγκες της εργασίας, παρήχθησαν και νέα χαρακτηριστικά όπως βαθμολογίες, μέσοι όροι σε τελικές προσπάθειες και γκολ κλπ.

Τα δεδομένα χωρίστηκαν ως εξής:

- season 2007-08 έως 2012-13: αποτελούν το σύνολο δεδομένων εκπαίδευσης κάθε αλγορίθμου που χρησιμοποιήθηκε.

- season 2013-14 και 2014-15: αποτελούν το validation set, όπου κάθε μοντέλο του ίδιου αλγορίθμου (η διαφορά είναι στις παραμέτρους και στα χαρακτηριστικά), αφού εκπαιδευθεί στο training set δοκιμάζει την απόδοση του σε αυτό το σύνολο δεδομένων, με τη χρήση ενός τροποποιημένου grid search που θα επεξηγηθεί αναλυτικά παρακάτω. Οι παράμετροι και τα χαρακτηριστικά του μοντέλου με την καλύτερη απόδοση επιλέγονται και ως αυτοί που θα χρησιμοποιηθούν στο τελικό μοντέλο.

- season 2015-16 έως 2018-19: αποτελούν το test set, όπου θα δοκιμαστεί η απόδοση του τελικού μοντέλου. Το τελικό μοντέλο προκύπτει από την επιλογή των επικρατέστερων παραμέτρων από το τροποποιημένο grid search με εκπαίδευση στο σύνολο δεδομένων της ένωσης του training set και του validation set.

Θα εφαρμοστούν συνολικά 7 διαφορετικές υλοποιήσεις για την επίλυση του προβλήματος που πραγματεύεται η παρούσα εργασία. Οι 5 υλοποιήσεις από τις 7, αποτελούν κλασικούς αλγορίθμους Μηχανικής Μάθησης, που μάλιστα παρουσιάστηκαν συνοπτικά στο κεφάλαιο 2 και είναι: οι k-Nearest Neighbours, Gaussian Naive Bayes, Support Vector Machines, Multilayer Perceptron και Random Forest. Μια επιπλέον υλοποίηση αποτελεί ο γεωμετρικός προσδιορισμός των κέντρων κάθε κλάσης και η κατηγοριοποίηση κάθε αγώνα στην κλάση στην οποία βρίσκεται πλησιέστερα. Ενώ τέλος, μια ακόμα υλοποίηση αποτελείται από το συνδυασμό δύο αλγορίθμων Μηχανικής Μάθησης, κατά την οποία ο ένας αλγόριθμος έχει εκπαιδευτεί και προβλέπει νίκη γηπεδούχου ή ισοπαλία και ο άλλος αντίστοιχα, μόνο ισοπαλία και νίκη φιλοξενούμενου. Κάθε υλοποίηση θα παρουσιαστεί και θα αναλυθεί παρακάτω με τα αποτελέσματά της.

Μετά την παραγωγή των τελικών προβλέψεων στο test set κάθε αλγορίθμου θα εφαρμοστούν 7 διαφορετικές επιλογές στοιχηματισμού πάνω στις προβλέψεις, ώστε να φανεί αν το σύστημα τελικά παράγει κέρδος και αν ναι, ποια προσέγγιση αποδίδει καλύτερα σε κάθε αλγόριθμο. Οι 7 προσεγγίσεις είναι οι ακόλουθες:

- 1-X-2: Σε αυτή την περίπτωση, επιλέγονται να παιχτούν όλα τα ματς του test set.

- 1: Σε αυτή την περίπτωση, επιλέγονται να παιχτούν μόνο τα παιχνίδια που ο αλγόριθμος έχει προβλέψει νίκη της γηπεδούχου ομάδος. Προφανώς, το ποντάρισμα τοποθετείται στον “άσσο”.

- X: Σε αυτή την περίπτωση, επιλέγονται να παιχτούν μόνο τα παιχνίδια που ο αλγόριθμος έχει προβλέψει ισοπαλία.

- 2: Σε αυτή την περίπτωση, επιλέγονται να παιχτούν μόνο τα παιχνίδια που ο αλγόριθμος έχει προβλέψει νίκη της φιλοξενούμενης ομάδος.

- 1-X: Σε αυτή την περίπτωση, επιλέγονται προς στοιχηματισμό μόνο οι αγώνες που ο αλγόριθμος προβλέπει νίκη γηπεδούχου ή ισοπαλία. Προφανώς, σε κάθε ματς τοποθετείται το ποντάρισμα και στην αντίστοιχη πρόβλεψη του αλγορίθμου.

- 1-2: Σε αυτή την περίπτωση, επιλέγονται προς στοιχηματισμό μόνο οι αγώνες που το σύστημα προβλέπει νίκη γηπεδούχου ή φιλοξενούμενου.

- X-2: Σε αυτή την περίπτωση, επιλέγονται προς στοιχηματισμό μόνο οι αγώνες που το σύστημα προβλέπει ισοπαλία ή νίκη του φιλοξενούμενου.

Όπως αναφέρθηκε και σε προηγούμενο κεφάλαιο, η σύγκριση μεταξύ όλων των μοντέλων για το ποια υλοποίηση και ποια προσέγγιση είναι πιο αποδοτική θα γίνει με τη χρήση του Yield, το οποίο αποτελεί έναν δείκτη κερδών ή ζημίας του παίχτη σε ένα μεγάλο σερί αγώνων.

5.2 Πειραματικό πρωτόκολλο

5.2.1 Πίνακας σύγχυσης και μετρικές αποδόσεις

Ο πίνακας σύγχυσης (confusion matrix) χρησιμοποιείται για την καλύτερη παρουσίαση και κατανόηση των προβλέψεων ενός συστήματος. Ο πίνακας σύγχυσης αποτελεί έναν τετραγωνικό πίνακα, μεγέθους όσο και το πλήθος των διαφορετικών κλάσεων. Κάθε γραμμή αντιπροσωπεύει το σύνολο των στοιχείων που ανήκουν πραγματικά στη συγκεκριμένη κλάση, ενώ κάθε στήλη αντιπροσωπεύει το σύνολο των στοιχείων που ταξινομήθηκαν σε αυτή την κλάση. Προφανώς, τα στοιχεία της διαγωνίου είναι αυτά που έχουν ταξινομηθεί στη σωστή κλάση. Παρακάτω, παρουσιάζεται ο confusion matrix που θα χρησιμοποιηθεί στη συνέχεια.

	Πρόβλεψη άσσου	Πρόβλεψη Ισοπαλία	Πρόβλεψη διπλού
Αποτέλεσμα άσσος			
Αποτέλεσμα Ισοπαλία			
Αποτέλεσμα διπλό			

Για την καλύτερη κατανόηση των προβλέψεων θα χρησιμοποιηθούν κάποιες μετρικές αποδόσεις, που μπορούν να υπολογιστούν εύκολα και μέσω του πίνακα σύγχυσης. Αυτές οι μετρικές είναι:

- το precision (ακρίβεια), το οποίο αποτελεί το κλάσμα των σωστά ταξινομημένων στοιχείων της κλάσης προς το σύνολο των στοιχείων που ταξινομήθηκαν στην κλάση αυτή. Από τον πίνακα σύγχυσης μπορεί να προκύψει από τη διαίρεση της τιμής του κελιού της κύριας διαγωνίου της γραμμής της συγκεκριμένης κλάσης, προς το άθροισμα της αντίστοιχης στήλης-κλάσης
- το recall (ανάκληση), που αποτελεί το κλάσμα των σωστά ταξινομημένων στοιχείων της κλάσης προς το σύνολο των στοιχείων της κλάσης αυτής. Από τον πίνακα σύγχυσης μπορεί να προκύψει από τη διαίρεση της τιμής της κύριας διαγωνίου της γραμμής της συγκεκριμένης κλάσης, προς το άθροισμα της γραμμής-κλάσης.
- και το f1-score, που αποτελεί τον σταθμισμένο μέσο όρο του precision και του recall. Υπολογίζεται από τον ακόλουθο τύπο:

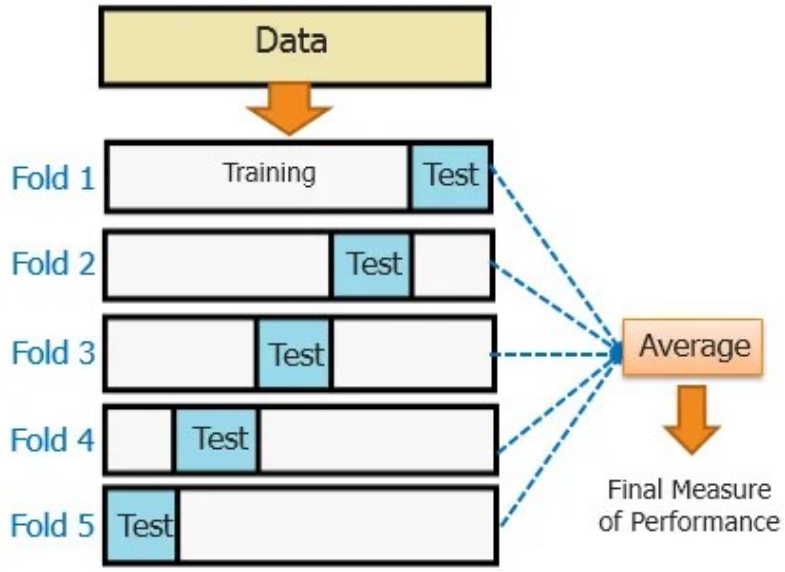
$$f1-score=\frac{2\cdot(recall\cdot precision)}{recall+precision} \quad (5.1)$$

5.2.2 Αναζήτηση πλέγματος (Grid search)

Οι περισσότεροι αλγόριθμοι Μηχανικής Μάθησης διαθέτουν παραμέτρους, οι οποίες παίζουν καθοριστικό ρόλο για τη δομή και τις αποφάσεις του συστήματος, γεγονός που επηρεάζει σημαντικά και την απόδοση τους .Το Grid Search είναι η διαδικασία αναζήτησης των βέλτιστων παραμέτρων κάθε αλγορίθμου. Στην ουσία, ορίζονται για κάθε παράμετρο οι τιμές που επιθυμούμε να εξετάσουμε και δοκιμάζονται όλοι οι δυνατοί συνδυασμοί μεταξύ τους. Για κάθε συνδυασμό παραμέτρων που επιλέγεται, εκπαιδεύεται το μοντέλο στο training set και δοκιμάζεται η απόδοση του σε ένα άλλο σετ. Υπάρχουν δύο τρόποι να γίνει αυτό, είτε με cross validation, είτε με την ύπαρξη ενός συνόλου, που ονομάζεται validation set. Σε κάθε περίπτωση, το σύστημα πρέπει να δοκιμαστεί σε δεδομένα, τα οποία δεν έχουν χρησιμοποιηθεί κατά την εκπαίδευση.

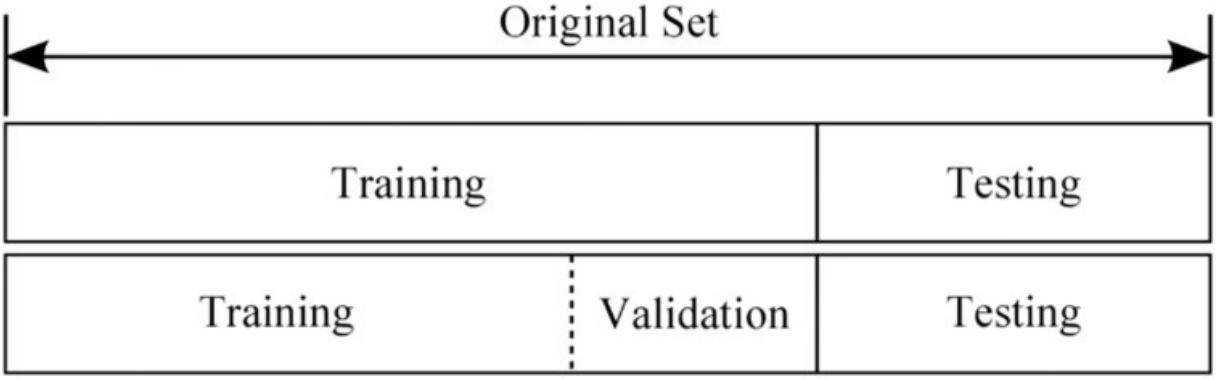
Το cross validation χωρίζει το training set σε k μέρη, όπου το k ορίζεται από το άτομο που αναπτύσσει το σύστημα. Στη συνέχεια, το σύστημα εκπαιδεύεται στα k-1 σύνολα και δοκιμάζεται η απόδοση του στο σύνολο που δεν χρησιμοποιήθηκε. Η διαδικασία είναι επαναληπτική, κάθε φορά επιλέγεται το σύνολο στο οποίο θα γίνει η αξιολόγηση και η εκπαίδευση γίνεται στα υπόλοιπα k-1. Η διαδικασία ολοκληρώνεται όταν όλα τα k σύνολα έχουν επιλεγεί. Προφανώς, για κάθε συνδυασμό προκύπτουν και k διαφορετικές αποδόσεις. Ο μέσος όρος των k αποδόσεων είναι και η απόδοση του συστήματος για τον συγκεκριμένο συνδυασμό παραμέτρων. Οι παράμετροι που βρίσκονται στον συνδυασμό με τον καλύτερο μέσο όρο, δηλαδή με την καλύτερη απόδοση επιλέγονται ως οι βέλτιστες παράμετροι του συστήματος. Είναι φανερό, ότι αυτή η διαδικασία είναι ακριβή σε χρόνο και υπολογιστικούς πόρους, καθώς έχουν να εκπαιδευτούν και να δοκιμαστούν πολλά μοντέλα. Για παράδειγμα, αν ένα μοντέλο έχει 4 παραμέτρους, με κάθε παράμετρο να μπορεί να πάρει 3 τιμές και επιλεγεί k=5, τότε θα πρέπει να πραγματοποιηθούν $3^4 \cdot 5 = 405$ δοκιμές. Προφανώς, μεγαλύτερη τιμή του k ή επιλογή προς εξέταση περισσότερων παραμέτρων ή διαφορετικών τιμών για κάθε παράμετρο, θα αύξανε ακόμα περισσότερο το πλήθος των δοκιμών. Αυτός, ο τρόπος εύρεσης παραμέτρων είναι αρκετά αποδοτικός, καθώς δεν υπάρχει απώλεια δεδομένων και εξετάζονται εξαντλητικά όλοι οι συνδυασμοί. Ακόμα, ένας λόγος που χρησιμοποιείται ευρέως είναι ότι διατίθεται έτοιμος προς υλοποίηση από τη βιβλιοθήκη sklearn της Python, αυξάνοντας περισσότερο την ευχρηστία του. Παρακάτω παρατίθονται σχηματικά τα όσα έχουν αναφερθεί για το cross validation, όπου το k=5, δηλαδή το training set διαμερίζεται σε 5

μέρη και λαμβάνουν χώρα 5 επαναλήψεις, όπου κάθε φορά επιλέγεται και διαφορετικό υποσύνολο για να δοκιμαστεί η απόδοση, ενώ στο τέλος προκύπτει ο μέσος όρος των 5 επαναλήψεων, αποτελώντας και την απόδοση του συνδυασμού:



Λόγω της υψηλής απαίτησης αυτού του τρόπου σε χρόνο και υπολογιστική ισχύ, επιλέχθηκε ο επόμενος τρόπος.

Η άλλη επιλογή έναντι του cross validation, είναι η ύπαρξη του validation set. Για να αποφευχθεί η παραπάνω επαναληπτική διαδικασία, το training set διαμερίζεται σε δύο υποσύνολα, το νέο training set και το validation set. Για κάθε δυνατό συνδυασμό παραμέτρων το μοντέλο εκπαιδεύεται στο νέο training set και δοκιμάζεται η απόδοση του στο validation set. Ο συνδυασμός παραμέτρων που επιτυγχάνει την υψηλότερη απόδοση στο validation set, είναι και ο τελικός συνδυασμός που θα χρησιμοποιηθεί και από το μοντέλο. Το τελικό μοντέλο θα εκπαιδευτεί στο αρχικό training set, δηλαδή στην ένωση του validation set και του νέου training set, και η απόδοση θα δοκιμαστεί στο test set. Η αναλογία training:validation:test συνηθίζεται στην πράξη να είναι 5:1:3 και για την παρούσα εργασία επιλέχθηκε 6:2:4. Αυτός ο τρόπος είναι αρκετά πιο γρήγορος από τον προηγούμενο, υστερεί βέβαια λίγο σε ακρίβεια, αλλά είναι αρκετά ικανοποιητικός. Παρακάτω δίνεται και σχηματικά ο διαχωρισμός του training σε υποσύνολα:



Ως μέτρο απόδοσης ενός συστήματος συνήθως χρησιμοποιείται η ορθότητα, δηλαδή πόσα αντικείμενα ταξινόμησε σωστά το σύστημα. Όμως, σκοπός της εργασίας δεν είναι η ανάπτυξη ενός

συστήματος που είναι έχει υψηλή ορθότητα στις προβλέψεις του, αλλά ενός συστήματος που οι προβλέψεις του, αν χρησιμοποιηθούν προς στοιχηματισμό, να οδηγούν σε όσο το δυνατόν μεγαλύτερο κέρδος. Όπως αναφέρθηκε και στο κεφάλαιο 2, για να είναι ένα σύστημα κερδοφόρο πρέπει να ικανοποιείται η συνθήκη (3.2) , δηλαδή το γινόμενο ορθότητας και μέσης στοιχηματικής απόδοσης που προβλέπει το σύστημα να είναι μεγαλύτερο της μονάδος. Το παραπάνω γινόμενο εκφράζει την μέση αναμενόμενη επιστροφή που θα έχει κάποιος αν ποντάρει 1 μονάδα τυχαία σε μία πρόβλεψη του συστήματος. Συνεπώς, για να είναι κερδοφόρο το σύστημα η μέση επιστροφή θα πρέπει να είναι μεγαλύτερο της μονάδος. Επομένως, ως μέτρο σύγκρισης στα μοντέλα χρησιμοποιήθηκε αυτό το γινόμενο.

Συνδυάζοντας όλα τα παραπάνω, το grid search υλοποιήθηκε με την ύπαρξη validation set και επιλογή των παραμέτρων από το μοντέλο που επιτυγχάνει το υψηλότερο γινόμενο ορθότητας και μέσης στοιχηματικής απόδοσης. Προφανώς, το γινόμενο αυτό απαιτείται να είναι και μεγαλύτερο της μονάδας, όπως υπαγορεύει η συνθήκη (3.2).

5.3 Πειραματικά αποτελέσματα

Παρακάτω θα παρουσιαστούν οι 7 διαφορετικές υλοποιήσεις και όλες οι προσεγγίσεις για κάθε υλοποίηση. Για κάθε υλοποίηση θα παρουσιαστούν τα features που επιλέχθηκαν τελικά και το αν έγινε κάποια επεξεργασία στο dataset, όπως κανονικοποίηση ή κλιμακοποίηση των δεδομένων, που βελτίωσε την απόδοσή της. Να σημειωθεί ότι για την επιλογή απόδοσης προς στοιχηματισμό, για κάθε πρόβλεψη των συστημάτων, επιλέχθηκε η μεγαλύτερη από τις διαθέσιμες αποδόσεις, δηλαδή οι αποδόσεις που βρίσκονται στα χαρακτηριστικά 60, 61, 62. Αυτό βέβαια, προϋποθέτει ότι ο παίχτης που ακολουθεί τις προβλέψεις του συστήματος έχει λογαριασμό σε πολλές στοιχηματικές εταιρείες και προφανώς σε κάθε λογαριασμό έχει διαθέσιμα χρήματα για στοιχηματισμό, το οποίο δεν είναι κάτι δύσκολο ή παράνομο. Ακολουθούν οι υλοποιήσεις:

1. k-Nearest Neighbours (kNN)

Ο αλγόριθμος αυτός έχει περιγραφεί θεωρητικά στο Κεφάλαιο 2. Τα features που επιλέχθηκαν στο τελικό μοντέλο είναι τα *39, 42, 43, 63, 69, 70, 71,* τα οποία μάλιστα και κανονικοποιήθηκαν, γιατί ο k-NN στηρίζεται στον υπολογισμό αποστάσεων, οπότε χαρακτηριστικά με τιμές που παίρνουν πολύ μικρές ή πολύ μεγάλες τιμές τιμές, δημιουργούν προβλήματα στον αλγόριθμο. Το παραπάνω ήταν κάτι που θεωρητικά ήταν αναμενόμενο, αλλά αποδείχθηκε και στις δοκιμές. Ως παράμετρος για το k, δηλαδή πόσοι γείτονες να ληφθούν υπόψιν για τον προσδιορισμό της κλάσης, επιλέχθηκε μέσω του grid search το 36.

Ακολουθεί το confusion matrix των προβλέψεων του συστήματος στις 4 επιλεγμένες χρονιές του Αγγλικού Πρωταθλήματος, που αποτελούν και το test, μαζί με την αναφορά και τα στατιστικά κατηγοριοποίησης του συστήματος:

	Πρόβλεψη άσσου	Πρόβλεψη Ισοπαλία	Πρόβλεψη διπλού
Αποτέλεσμα άσσος	357	17	52
Αποτέλεσμα Ισοπαλία	139	20	45
Αποτέλεσμα διπλό	124	12	114

	precision	recall	f1-score	support
Άσσος	0.58	0.84	0.68	426
Ισοπαλία	0.41	0.10	0.16	204
Διπλό	0.54	0.46	0.49	250

Ο συγκεκριμένος αλγόριθμος παρατηρεί κανείς ότι έχει πολύ καλή απόδοση στις νίκες γηπεδούχων (0.68 f1-score), καθώς ανιχνεύει και κατηγοριοποιεί σωστά το 84% των αγώνων που έληξαν με νίκη γηπεδούχου. Ακόμα, παρατηρεί κανείς ότι την πλειοψηφία των αγώνων τα κατηγοριοποιεί σε άσσο, με ακρίβεια (precision) στην πρόβλεψη νίκη γηπεδούχου 58%, ενώ ανιχνεύει και προβλέπει σχετικά πολύ λίγες ισοπαλίες. Η νίκη γηπεδούχου είναι το πιο συχνό αποτέλεσμα και αποτελεί περίπου το αποτέλεσμα περίπου των 50% των αγώνων, για αυτό και δικαιολογείται η τάση του αλγορίθμου σε πολλές νίκες γηπεδούχου. Σε αντίθεση, οι ισοπαλίες είναι το λιγότερο πιθανό αποτέλεσμα, με περίπου το 25% των αγώνων να ολοκληρώνονται με ισόπαλο σκορ. Ο αλγόριθμος σημείωσε αρκετά υψηλή απόδοση και στην πρόβλεψη ισοπαλιών και στην πρόβλεψη νίκη της φιλοξενούμενης ομάδας.

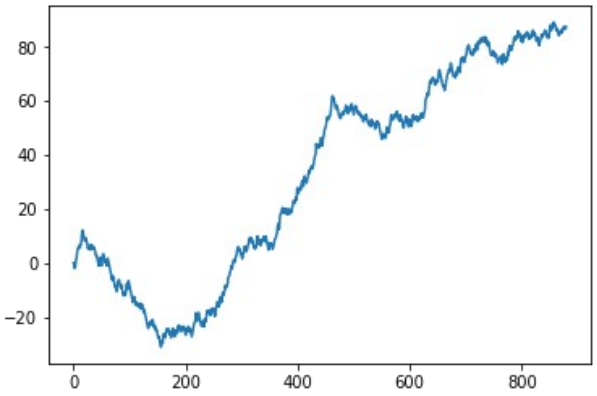
Ακολουθούν οι στοιχηματικές προσεγγίσεις, οι οποίες στηρίζονται στις προβλέψεις του k-NN:

Προσέγγιση	Πλήθος αγώνων που υπάρχει πρόβλεψη	Πλήθος επιτυχημένων προβλέψεων	Ακρίβεια	Καθαρό κέρδος σε μονάδες	Μέση απόδοση επιτυχημένων προβλέψεων	Yield
1X2	880	491	55.79%	**87.87**	1.97	**9.9%**
1	620	357	**57.58%**	64.62	1.91	**10.4%**
X	49	20	40.81%	21.74	3.57	**44%**
2	211	114	54.02%	1.50	1.86	0%
1X	669	377	56.35%	**86.36**	2.00	**12.90%**
12	831	471	56.67%	66.13	1.90	7.9%
X2	260	134	51.53%	23.25	2.11	8.9%

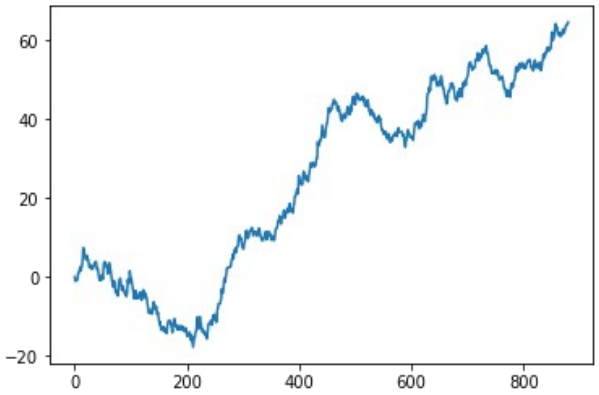
Να σημειωθεί ότι η στήλη με το καθαρό κέρδος, αναφέρεται σε μονάδες και όχι σε Ευρώ ή κάποιο άλλο νόμισμα. Για να γίνει πιο κατανοητό, αν κάποιος επέλεγε να στοιχιματίσει σε όλα τα παιχνίδια, δηλαδή επέλεγε την προσέγγιση 1X2 και σε κάθε αγώνα πόνταρε 10 €, τότε το καθαρό του κέρδος θα ήταν 87.87 € · 10 = 878.7 €. Δηλαδή, η μονάδα σε αυτό το παράδειγμα αντιστοιχεί σε 10 €, όσο και το ποντάρισμα σε κάθε αγώνα του παίχτη.

Με μια γρήγορη ανάγνωση, παρατηρεί κανείς ότι και οι 7 προσεγγίσεις δεν δημιουργούν ζημία στον παίχτη, ενώ υπάρχουν κάποιες που πετυχαίνουν πολύ υψηλό καθαρό κέρδος και πολύ υψηλό Yield. Το 44% Yield στην προσέγγιση που ο παίχτης ακολουθεί το σύστημα μόνο όταν προβλέπει το σύστημα ισοπαλία, κρίνεται πλασματικό λόγω του πολύ μικρού πλήθους αγώνων. Οι υπόλοιπες προσεγγίσεις με πλήθος αγώνων πάνω από 600 παιχνίδια κρίνονται με μεγαλύτερη ασφάλεια. Φυσικά, για να προκύψουν πιο ασφαλή συμπεράσματα θα μπορούσαν να προστεθούν κι άλλοι αγώνες, ωστόσο δεν ήταν δυνατή η εύρεση πλήρων στατιστικών για χρονιές πριν το 2007. Επιπλέον, να σημειωθεί ότι ο αλγόριθμος προβλέπει και υψηλές αποδόσεις, κοντά στο 2, και παρότι η ακρίβεια είναι λίγο πάνω από το 50%, ο αλγόριθμος είναι κερδοφόρος. Οι προσεγγίσεις

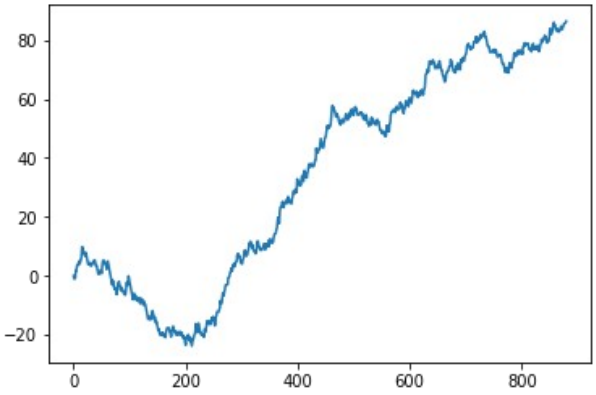
1X2, 1 και 1X είναι αυτές που ξεχώρισαν με πολύ υψηλό Yield και καθαρό κέρδος, σε ικανοποιητικό πλήθος αγώνων. Παρακάτω, παρουσιάζεται η πορεία των κερδών κατά την εξέλιξη των αγώνων στις 3 αυτές προσεγγίσεις.



Πορεία των κερδών σε μονάδες με 1X2



Πορεία των κερδών σε μονάδες με 1



Πορεία των κερδών σε μονάδες με 1X

Και στα 3 γραφήματα είναι φανερή η ανοδική τάση που επικρατεί στην εξέλιξη των καθαρών κερδών. Παρότι το δείγμα είναι μερικές εκατοντάδες αγώνων, μπορεί κανείς να υποθέσει λογικά ότι αυτή η αυξητική τάση θα συνεχίσει να υπάρχει και σε επόμενες προβλέψεις. Ωστόσο, και οι 3 προσεγγίσεις ξεκίνησαν με ζημία στην αρχή, κοντά στις 20 μονάδες, γεγονός που αποτρέπει μεγάλα πονταρίσματα, παρά το υψηλό κέρδος στο τέλος. Για παράδειγμα, αν ένα άτομο πόνταρε 20 € σε κάθε αγώνα, τότε θα είχε ζημία περίπου 400 € κοντά στην ολοκλήρωση 200 αγώνων, δηλαδή κοντά στο τέλος της πρώτης χρονιάς. Ωστόσο, αν συνέχιζε να εμπιστεύεται το σύστημα, θα τελείωνε την τετραετία με καθαρό κέρδος πάνω από 1200 €. Πράγμα που σημαίνει ότι ο παίχτης θα έπρεπε να διαθέσει 400 €, για να συνεχίσει να ακολουθεί τις προβλέψεις του συστήματος. Η αρχική πτώση ωστόσο, διαφέρει πολύ από την εξέλιξη των κερδών στη συνέχεια. Αυτό θα μπορούσε να δικαιολογηθεί από την ύπαρξη πολλών μη αναμενόμενων αποτελεσμάτων στην πρώτη season του test set. Στις υπόλοιπες 3 χρονιές το σύστημα και στις 3 προσεγγίσεις παρουσιάζει μεγάλη αυξητική τάση, καθώς από τις -20 μονάδες φτάνει ακόμα και στις +85 μονάδες, δηλαδή κάθε χρονιά σημειώνει κατά μέσο όρο κέρδος +35 μονάδες, ενώ γενικά στην τετραετία σημειώνει αύξηση +25 περίπου μονάδες.

2. Gaussian Naive Bayes (GNB)

Ο αλγόριθμος έχει περιγραφεί θεωρητικά στο Κεφάλαιο 2. Τα χαρακτηριστικά που επιλέχθηκαν για το συγκεκριμένο σύστημα ως επικρατέστερα είναι: *42, 44, 45, 47, 54, 63, 71,* στα οποία η κανονικοποίηση και η κλιμακοποίηση των δεδομένων δεν βοήθησε στην βελτίωση της απόδοσης του συστήματος. Αντίθετα, τη μείωσε, για το λόγο αυτό τα δεδομένα παρέμειναν ως είχαν,χωρίς καμία επεξεργασία. Ο αλγόριθμος δεν διαθέτει παραμέτρους.

Παρακάτω παρουσιάζεται το confusion matrix των προβλέψεων του συστήματος στις 4 επιλεγμένες χρονιές του Αγγλικού Πρωταθλήματος, που αποτελούν και το test, μαζί με την αναφορά και τα στατιστικά κατηγοριοποίησης του συστήματος:

	Πρόβλεψη άσσου	Πρόβλεψη Ισοπαλία	Πρόβλεψη διπλού
Αποτέλεσμα άσσος	288	61	77
Αποτέλεσμα Ισοπαλία	89	46	69
Αποτέλεσμα διπλό	60	44	146

	precision	recall	f1-score	support
Άσσος	0.66	0.68	0.67	426
Ισοπαλία	0.30	0.23	0.26	204
Διπλό	0.50	0.58	0.54	250

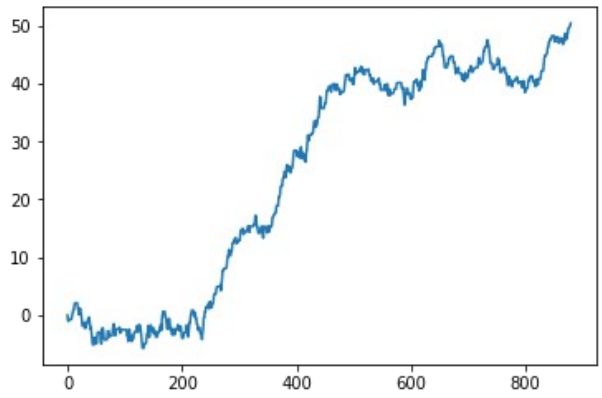
Παρατηρεί κανείς πως το συγκεκριμένο στοίχημα προβλέπει σε σωστές αναλογίες τις 3 κλάσεις, με το 50% περίπου των αγώνων να ταξινομούνται σε νίκη γηπεδούχου και τις άλλες δύο κλάσεις να μοιράζονται το υπόλοιπο ποσοστό. Ο αλγόριθμος επιτυγχάνει υψηλό precision στην κλάση του άσσου, δηλαδή όταν ταξινομεί έναν αγώνα σε Άσσο επαληθεύεται η πρόβλεψη του κατά 66%. Ακόμα, ο αλγόριθμος ανιχνεύει και ταξινομεί σωστά τους αγώνες που ανήκουν στις κλάσεις νίκη γηπεδούχου και νίκη φιλοξενούμενου, γεγονός που αποτυπώνεται και στο υψηλό f1-score. Το 23% recall στην κλάση των ισοπαλιών κρίνεται ως ικανοποιητικό, λόγω της δυσκολίας που υπάρχει στην πρόβλεψη της ισοπαλίας, μιας και είναι το πιο σπάνιο αποτέλεσμα και το πιο αμφίρροπο.

Ακολουθούν οι στοιχηματικές προσεγγίσεις, οι οποίες στηρίζονται στις προβλέψεις του GNB:

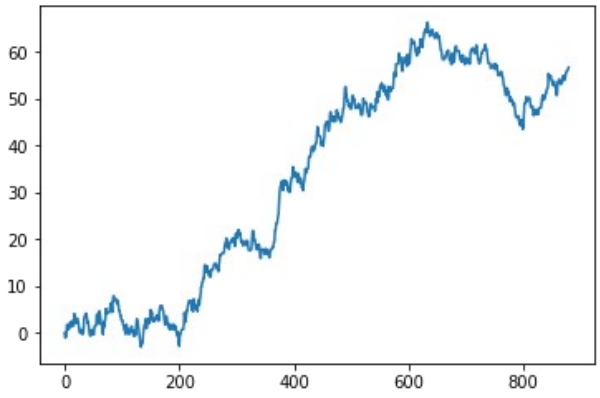
Προσέγγιση	Πλήθος αγώνων που υπάρχει πρόβλεψη	Πλήθος επιτυχημένων προβλέψεων	Ακρίβεια	Καθαρό κέρδος σε μονάδες	Μέση απόδοση επιτυχημένων προβλέψεων	Yield
1X2	880	480	54.54%	43.29	1.92	4.9%
1	437	288	**65.90%**	50.41	1.69	**11.5%**
X	151	46	30.46%	6.27	3.41	4.1%
2	292	146	50.00%	-13.40	1.90	-4.5%
1X	588	334	56.80%	**56.70**	1.93	**9.6%**

12	729	434	**59.53%**	37.00	1.76	5.0%
X2	443	192	43.34%	-7.13	2.27	-1.6%

Αρχικά, είναι εύκολο να διακρίνει κανείς ότι υπάρχουν προσεγγίσεις στο σύστημα αυτό που δημιουργούν ζημία στον παίχτη, έστω και μικρή. Είναι φανερό ότι η "προβληματική" κλάση είναι η νίκη του φιλοξενούμενου, το διπλό, καθώς είναι η μόνη κλάση που δημιουργεί ζημία όταν παίζεται μόνη της (-13.40 μονάδες). Επιπλέον, σε όποια προσέγγιση συμπεριλαμβάνεται στο στοιχηματισμό η συγκεκριμένη κλάση, μεταφέρει τη ζημία των 13.40 μονάδων. Για παράδειγμα, η προσέγγιση 1X2 έχει καθαρό κέρδος σε μονάδες 43.29, ενώ η προσέγγιση 1X έχει 56.70, διαφορά που ευθύνεται στη κλάση του διπλού. Η κλάση που αναφέρεται στη νίκη του φιλοξενούμενου δεν είναι κερδοφόρα, καθώς το γινόμενο 0,50 * 1.90 = 0.95 < 1 (συνθήκη 3.2). Το σύστημα φαίνεται να τα πηγαίνει εξαιρετικά στην πρόβλεψη νίκης γηπεδούχου με ποσοστό ακρίβειας 65.9% και μέση απόδοση στις επιτυχημένες προβλέψεις 1.69. Ως προς το Yield, είναι δύο οι προσεγγίσεις που ξεχωρίζουν, αυτή που επιλέγει να ασχοληθεί μόνο με τις προβλέψεις των άσσων (με Yield 11.5%) και αυτή που επιλέγει να συμπεριλάβει μόνο τις νίκες γηπεδούχων και τις ισοπαλίες για στοιχηματισμό (με Yield 9.6%). Οι δύο αυτές προσεγγίσεις κρίνονται ως οι επικρατέστερες αυτού του συστήματος, δοκιμασμένες μάλιστα και σε ικανοποιητικό πλήθος αγώνων. Παρακάτω, παρουσιάζεται η πορεία των κερδών κατά την εξέλιξη των αγώνων στις 2 αυτές προσεγγίσεις.



Πορεία των κερδών σε μονάδες με 1



Πορεία των κερδών σε μονάδες με 1X

Παρατηρεί κανείς εύκολα ότι και στα δύο γραφήματα τα καθαρά κέρδη σε μονάδες παρουσιάζουν αυξητική τάση σχεδόν σε όλη τη διάρκεια των 4 αγωνιστικών σαιζόν που αποτελούν το test του συστήματος. Ωστόσο, και οι δύο προσεγγίσεις ξεκίνησαν με στασιμότητα στα κέρδη και ίσως μια μικρή ζημία την πρώτη χρονιά, καθώς μετά τη συμπλήρωση περίπου 200 αγώνων, τα κέρδη βρίσκονται στο 0. Από κει και πέρα, η προσέγγιση με την επιλογή μόνο των αγώνων που το σύστημα προβλέπει νίκη γηπεδούχου ακολούθησε πολύ υψηλή αύξηση την $2^{η}$ και την $3^{η}$ χρονιά (με ετήσιο ρυθμό αύξησης κερδών σε μονάδες κοντά στο +20) και πιο ήπιο ρυθμό ανόδου την $4^{η}$ και τελευταία χρονιά (με ρυθμό αύξησης κοντά στο +10). Η προσέγγιση με την επιλογή τόσο των άσσων, όσο και των ισοπαλιών, προς στοιχηματισμό από τις προβλέψεις του συστήματος, κατά τη $2^{η}$ και $3^{η}$ χρονιά σημείωσε πολύ υψηλή αύξηση στα κέρδη (με ετήσιο ρυθμό αύξησης σε μονάδες κοντά στο +25), ωστόσο κατά την τελευταία χρονιά υπήρξε μια μικρή υποχώρηση στα κέρδη, περίπου -10 μονάδες. Και οι δύο προσεγγίσεις παρουσιάζουν αυξητική τάση κατά τη διάρκεια των τεσσάρων χρόνων, της τάξης περίπου των +13 μονάδων, και επιτυγχάνουν πολύ υψηλό Yield κοντά στο 10%, γεγονός που τις καθιστά ιδιαίτερα πετυχημένες.

3. Support Vector Machine (SVM)

Ο αλγόριθμος έχει περιγραφεί και αναλυθεί θεωρητικά στο Κεφάλαιο 2. Τα χαρακτηριστικά που επιλέχθηκαν και βελτιστοποιούσαν την απόδοση του συστήματος είναι: *39, 41, 44, 46, 57, 59, 63, 64, 65*. Στα χαρακτηριστικά αυτά, παρατηρήθηκε ότι η κλιμακοποίηση ή κανονικοποίηση των τιμών των δεδομένων, όχι μόνο δεν βελτίωνε την απόδοση του συστήματος, αλλά αντίθετα τη μείωνε, για το λόγο αυτό, δεν έγινε κάποια περαιτέρω επεξεργασία στις τιμές. Ως προς τις παραμέτρους του αλγορίθμου επιλέχθηκε συνάρτηση πυρήνα RBF, για το C η τιμή 1 και για το gamma η τιμή 1. Οι τιμές των παραμέτρων επιλέχθηκαν μετά από το grid search που περιγράφηκε στην προηγούμενη υποενότητα.

Παρακάτω παρουσιάζεται το confusion matrix των προβλέψεων του συστήματος στις 4 επιλεγμένες χρονιές του Αγγλικού Πρωταθλήματος μαζί με την αναφορά και τα στατιστικά κατηγοριοποίησης του συστήματος:

	Πρόβλεψη άσσου	Πρόβλεψη Ισοπαλία	Πρόβλεψη διπλού
Αποτέλεσμα άσσος	321	43	62
Αποτέλεσμα Ισοπαλία	126	36	42
Αποτέλεσμα διπλό	160	28	62

	precision	recall	f1-score	support
Άσσος	0.53	0.75	0.62	426
Ισοπαλία	0.34	0.18	0.23	204
Διπλό	0.37	0.25	0.30	250

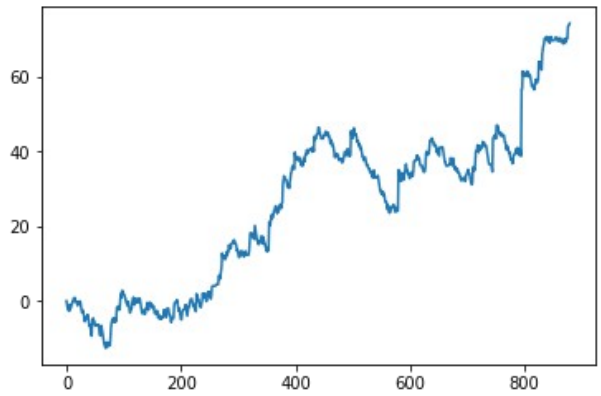
Παρατηρεί κανείς πως το σύστημα στην πλειοψηφία του προβλέπει νίκη γηπεδούχου και πολύ λιγότερο ισοπαλία ή νίκη φιλοξενούμενου. Η ακρίβεια στις προβλέψεις και στις 3 κλάσεις είναι αρκετά χαμηλή, ιδιαίτερα στην πρόβλεψη νίκης της φιλοξενούμενης ομάδας. Αυτό δεν αποτελεί πρόβλημα για το σύστημα, καθώς, όπως έχει ήδη αναφερθεί, στόχος της εργασίας δεν είναι η υψηλή ακρίβεια, αλλά η δημιουργία ενός κερδοφόρου συστήματος. Επομένως, όπως υπαγορεύει η συνθήκη (3.2), αν αυτή η χαμηλή ακρίβεια συνδυαστεί με υψηλή απόδοση, τότε το σύστημα μπορεί να είναι κερδοφόρο, πράγμα που αποτελεί και το ζητούμενο. Ο αλγόριθμος παρουσιάζει πολύ χαμηλό ποσοστό στην σωστή ανίχνευση ισοπαλιών και διπλών (χαμηλό recall) και μέτριο ως κακό ποσοστό ακρίβειας (precision) σε αυτές τις κλάσεις. Γεγονός που αποτυπώνεται και στο f1-score, που μόνο η κλάση νίκης γηπεδούχου παίρνει υψηλή τιμή.

Ακολουθούν οι στοιχηματικές προσεγγίσεις, οι οποίες στηρίζονται στις προβλέψεις του SVM:

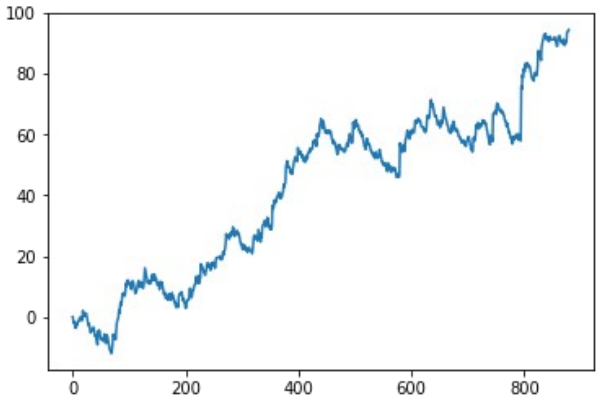
Προσέγγιση	Πλήθος αγώνων που υπάρχει πρόβλεψη	Πλήθος επιτυχημένων προβλέψεων	Ακρίβεια	Καθαρό κέρδος σε μονάδες	Μέση απόδοση επιτυχημένων προβλέψεων	Yield
1X2	880	419	47.61%	67.69	2.26	4.7%
1	607	321	52.88%	73.42	2.11	**12.0%**

X	107	36	33.64%	20.00	3.52	18.7%
2	166	62	37.34%	-25.73	2.26	-15.5%
1X	714	357	50.00%	**93.42**	2.26	**13.0%**
12	773	383	49.54%	47.69	2.14	6.1%
X2	273	98	35.89%	-5.73	2.72	-2.1%

Αρχικά, παρατηρεί κανείς ότι η προσέγγιση του ποντάρει κανείς μόνο στη νίκη του φιλοξενούμενου παρουσιάζει σημαντική ζημία, με Yield -15.5%. Γεγονός που φαίνεται από το πολύ χαμηλό ποσοστό ακρίβειας στις προβλέψεις του διπλού (37.34%) και την όχι και τόσο μεγάλη μέση απόδοση στις επιτυχείς προβλέψεις. Ωστόσο, και κέρδος να υπήρχε σε αυτή την προσέγγιση, είναι μικρό το δείγμα για να εξαχθούν συμπεράσματα. Αντίθετα, η προσέγγιση με μόνο νίκες γηπεδούχου φαίνεται πολύ καλή περίπτωση, με πολύ υψηλό Yield και σε ικανοποιητικό πλήθος αγώνων. Η προσέγγιση μόνο ισοπαλίες δείχνει να οδηγεί σε κέρδος, όμως το μικρό δείγμα αγώνων στο οποίο προβλέπει ισοπαλίες, καθιστούν τον οποιονδήποτε επιφυλακτικό. Όπως ήταν αναμενόμενο, οι προσεγγίσεις που περιλαμβάνουν στα πονταρίσματα τους τις νίκες φιλοξενούμενων, δεν παρουσιάζουν υψηλό Yield. Οι επικρατέστερες προσεγγίσεις είναι αυτές που κανείς ποντάρει μόνο στις νίκες γηπεδούχου (Yield 13%) και νίκες γηπεδούχου με ισοπαλίες (Yield 12% και καθαρό κέρδος σε μονάδες 93.42). Παρακάτω, παρουσιάζεται η πορεία των κερδών κατά την εξέλιξη των αγώνων στις 2 αυτές προσεγγίσεις.



Πορεία των κερδών σε προσέγγιση με 1



Πορεία των κερδών σε προσέγγιση με 1X

Παρατηρεί κανείς ότι και τα δύο γραφήματα παρουσιάζουν αυξητική τάση κατά τη διάρκεια και των 4 αγωνιστικών χρονιών. Στην αρχή, το σύστημα και στις δύο προσεγγίσεις παρουσιάζει μια πολύ μικρή ζημία, ωστόσο μέχρι το πέρας της πρώτης χρονιάς, έχουν καλύψει τη ζημία και μάλιστα έχουν δώσει ένα μικρό κέρδος. Η προσέγγιση μόνο με τη νίκη γηπεδούχου, τις επόμενες χρονιές παρουσιάζει καθαρά αυξητική πορεία, με αρκετά σκαμπανεβάσματα, ειδικότερα την $3^{η}$ χρονιά. Αντίθετα, η προσέγγιση με νίκη γηπεδούχου και ισοπαλία, παρουσιάζει πάλι καθαρά αυξητική πορεία, όμως με μικρότερα σκαμπανεβάσματα. Ακόμα, και στις 2 προσεγγίσεις παρατηρούνται απότομα άλματα, απόρροια των υψηλών αποδόσεων που προβλέπει επιτυχώς το σύστημα. Μάλιστα, το σύστημα προέβλεψε σωστά νίκη γηπεδούχου με απόδοση 19(!!), αλλά και άλλες πολύ υψηλές αποδόσεις. Το σύστημα γενικά παρουσιάζει χαμηλή ακρίβεια στις προβλέψεις του, ωστόσο επιδιώκει πολύ υψηλές στοιχηματικές αποδόσεις. Η ετήσια αύξηση του κέρδους σε μονάδες της προσέγγισης με 1 είναι +16, ενώ της άλλης +23. Πολύ υψηλές επιδόσεις και οι δύο.

4. Multilayer Perceptron (MLP)

Ο αλγόριθμος MLP έχει περιγραφεί και αναλυθεί στο Κεφάλαιο 2. Τα χαρακτηριστικά, που επιλέχθηκαν ως αυτά που ο αλγόριθμος επιτυγχάνει την υψηλότερη του απόδοση, είναι: *39, 41, 43, 44, 57, 58, 59, 63, 64, 65*. Τα χαρακτηριστικά δεν δέχτηκαν κάποια περαιτέρω επεξεργασία, καθώς ούτε η κανονικοποίηση, ούτε η κλιμακοποίηση των τιμών οδήγησε σε βελτίωση της απόδοσης του αλγορίθμου. Ως προς τις τιμές των παραμέτρων, επιλέχθηκε για solver το lbfgs, για alpha το 1e-05, για κρυφό επίπεδο επιλέχθηκαν 3 επίπεδα με μέγεθος διαδοχικά 20, 10, 5 και για max iter το 1300. Οι τιμές των παραμέτρων προέκυψαν από τη διαδικασία του grid search.

Παρακάτω παρουσιάζεται το confusion matrix των προβλέψεων του συστήματος στις 4 επιλεγμένες χρονιές του Αγγλικού Πρωταθλήματος μαζί με την αναφορά και τα στατιστικά κατηγοριοποίησης του συστήματος:

	Πρόβλεψη άσσου	Πρόβλεψη Ισοπαλία	Πρόβλεψη διπλού
Αποτέλεσμα άσσος	278	92	56
Αποτέλεσμα Ισοπαλία	108	57	19
Αποτέλεσμα διπλό	85	52	113

	precision	recall	f1-score	support
Άσσος	0.59	0.65	0.62	426
Ισοπαλία	0.28	0.28	0.28	204
Διπλό	0.54	0.45	0.49	250

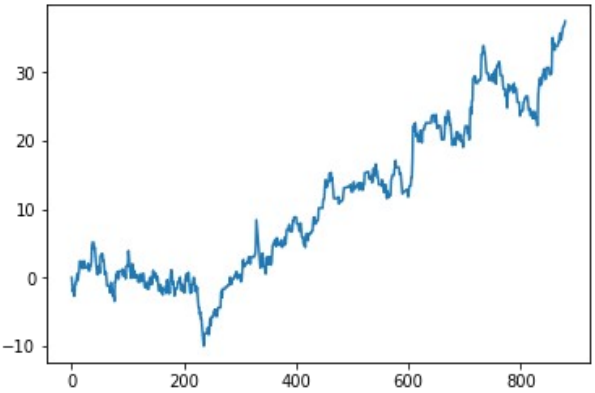
Ο αλγόριθμος MLP παρατηρεί κανείς ότι προβλέπει τις 3 κλάσεις σε σωστές αναλογίες, με τα περισσότερα ματς να ταξινομούνται ως νίκη γηπεδούχου και το υπόλοιπο περίπου 50% στις άλλες δύο κλάσεις. Ο αλγόριθμος παρουσιάζει ικανοποιητική ακρίβεια στις προβλέψεις του και στις 3 κλάσεις, με μικρότερη ακρίβεια στην πρόβλεψη ισοπαλιών, που είναι και η πιο δύσκολη κλάση, όπως έχει αναφερθεί. Ως προς την σωστή ανίχνευση και πρόβλεψη των κλάσεων, το σύστημα τα πηγαίνει ικανοποιητικά και στις 3 κλάσεις. Ειδικότερα, στην κλάση νίκης γηπεδούχου το recall είναι 65% και στην κλάση της νίκης φιλοξενούμενου 45%. Η γενικότερη απόδοση του αλγορίθμου κρίνεται ικανοποιητική με υψηλά f1-score σε όλες τις κλάσεις. Το f1-score στην κλάση της ισοπαλίας (28%) είναι μέχρι στιγμής το υψηλότερο που έχει σημειωθεί στους 4 πρώτους αλγορίθμους.

Ακολουθούν οι στοιχηματικές προσεγγίσεις, οι οποίες στηρίζονται στις προβλέψεις του MLP:

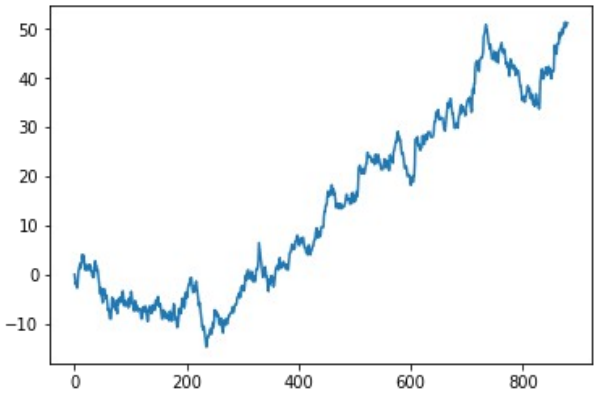
Προσέγγιση	Πλήθος αγώνων που υπάρχει πρόβλεψη	Πλήθος επιτυχημένων προβλέψεων	Ακρίβεια	Καθαρό κέρδος σε μονάδες	Μέση απόδοση επιτυχημένων προβλέψεων	Yield
1X2	880	448	50.90%	55.22	2.08	6.2%
1	471	278	**59.02%**	37.46	1.82	**7.9%**
X	201	57	28.35%	3.55	3.58	1.7%

2	208	113	54.32%	14.20	1.96	6.8%
1X	672	335	49.85%	41.02	2.12	6.1%
12	679	391	**57.58%**	**51.67**	1.86	**7.6%**
X2	409	170	41.54%	17.74	2.51	4.3%

Αρχικά, παρατηρεί κανείς ότι όλες οι προσεγγίσεις δημιουργούν έστω και μικρό κέρδος. Δεν υπάρχει δηλαδή κάποια προσέγγιση που να δημιουργεί ζημία, όπως ήταν στις προηγούμενες υλοποιήσεις η προσέγγιση μόνο νίκη φιλοξενούμενου. Ακόμα, παρατηρεί κανείς ότι οι περισσότερες προσεγγίσεις έχουν Yield περίπου 6-7% και μάλιστα σε ικανοποιητικό πλήθος αγώνων. Οι προσεγγίσεις που ξεχωρίζουν είναι αυτή που ακολουθεί το σύστημα μόνο στις νίκες γηπεδούχου και αυτή που το ακολουθεί στις νίκες γηπεδούχου και στις ισοπαλίες, με Yield 7.9% και 7.6% αντίστοιχα. Οι δυο προσεγγίσεις αυτές προσομοιώνονται σε ικανοποιητικό πλήθος αγώνων, οπότε χαρακτηρίζονται ως οι επικρατέστερες. Κανείς θα μπορούσε να πει ότι και η προσέγγιση, που ακολουθεί το σύστημα μόνο στις προβλέψεις νίκης του φιλοξενούμενου, είναι ικανοποιητική, ωστόσο το πλήθος των 208 αγώνων που δοκιμάστηκε κρίνεται ως επισφαλές. Παρακάτω, παρουσιάζεται η πορεία των κερδών κατά την εξέλιξη των αγώνων στις 2 επικρατέστερες προσεγγίσεις.



Πορεία των κερδών σε μονάδες με προσέγγιση 1



Πορεία των κερδών σε μονάδες με προσέγγιση 12

Παρατηρεί κανείς ότι και τα δύο γραφήματα, που παρουσιάζουν την πορεία των κερδών σε μονάδες των επικρατέστερων προσεγγίσεων στη διάρκεια των 4 season, ακολουθούν σταθερά αυξητική πορεία. Είναι ευδιάκριτο πάλι, ότι και οι δύο προσεγγίσεις ξεκινούν με στασιμότητα, καθώς δημιουργείται μια μικρή ζημία στην αρχή και ολοκληρώνουν την 1η χρονιά με οριακή ζημία. Ωστόσο, οι τρεις επόμενες χρονιές παρουσιάζουν καθαρά αυξητική πορεία, με ετήσια αύξηση περίπου +13 και +16 μονάδες αντίστοιχα. Η ζημία που παρατηρείται στην αρχή είναι της τάξης -10 και -15 μονάδες αντίστοιχα, όμως καλύπτεται σχετικά γρήγορα. Ακόμα, είναι φανερό από τις απότομες ανοδικές διακυμάνσεις των γραφικών παραστάσεων ότι και οι δύο προσεγγίσεις επιδιώκουν υψηλές στοιχηματικές αποδόσεις. Ενδιαφέρον αποτελεί, επίσης, ότι η προσέγγιση που δημιούργησε μεγαλύτερη ζημία στην αρχή, οδήγησε και σε μεγαλύτερο καθαρό κέρδος στο τέλος, ενώ η άλλη προσέγγιση με μικρότερη ζημία, οδήγησε και σε μικρότερο καθαρό κέρδος στο τέλος. Να σημειωθεί πως οι δύο προσεγγίσεις σε βάθος τετραετίας παρουσιάζουν ετήσια αυξητική τάση σε μονάδες +10 και +13 αντίστοιχα.

5. Random Forest

Ο αλγόριθμος Random Forest έχει περιγραφεί θεωρητικά στο κεφάλαιο 2. Τα χαρακτηριστικά που επιλέχθηκαν ως επικρατέστερα είναι: *39, 41, 43, 44, 57, 59, 63, 64, 65,* τα οποία δεν δέχτηκαν καμία παραπάνω επεξεργασία, καθώς καμία επεξεργασία δεν βελτίωσε την απόδοση του αλγορίθμου. Ως παράμετροι επιλέχθηκαν, για το n_estimators το 100, για το bootstrap η τιμή True, για το min_samples_leaf το 3, για το min_samples_split το 8, για το max_depth το 9 και για το max_features η τιμή log2. Οι παραπάνω τιμές προέκυψαν από τη διαδικασία του grid search.

Παρακάτω παρουσιάζεται το confusion matrix των προβλέψεων του συστήματος στις 4 επιλεγμένες χρονιές του Αγγλικού Πρωταθλήματος μαζί με την αναφορά και τα στατιστικά κατηγοριοποίησης του συστήματος:

	Πρόβλεψη άσσου	Πρόβλεψη Ισοπαλία	Πρόβλεψη διπλού
Αποτέλεσμα άσσος	327	25	74
Αποτέλεσμα Ισοπαλία	117	22	65
Αποτέλεσμα διπλό	98	14	138

	precision	recall	f1-score	support
Άσσος	0.60	0.77	0.68	426
Ισοπαλία	0.36	0.11	0.17	204
Διπλό	0.50	0.55	0.52	250

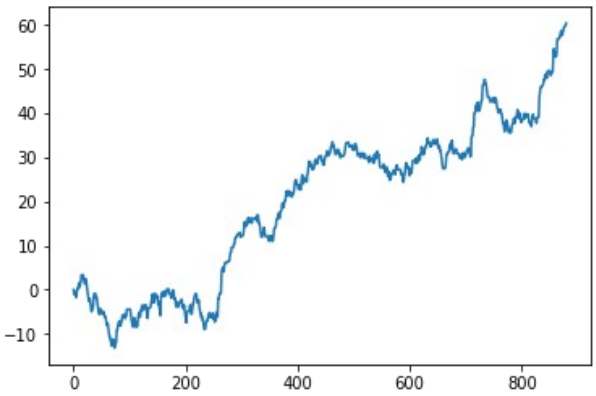
Αρχικά, παρατηρεί κανείς ότι ο αλγόριθμος προβλέπει κατά κύριο λόγο νίκη γηπεδούχου και νίκη φιλοξενούμενου. Η ισοπαλία κυμαίνεται σε χαμηλά επίπεδα σωστής ανίχνευσης (11%), ωστόσο η ακρίβεια θεωρείται καλή, για αυτό και το f1-score στην ισοπαλία δεν είναι πολύ χαμηλό. Στις άλλες δύο κλάσεις και η ακρίβεια, αλλά και η σωστή ανίχνευση τους είναι ιδιαίτερα υψηλά. Μάλιστα, στη νίκη γηπεδούχου το recall είναι 77% και το preciosion 60%, καταλήγοντας έτσι σε ένα πολύ υψηλό f1-score (68%). Όμοια και στη νίκη του φιλοξενούμενου, με τα ποσοστά να είναι λίγο πιο χαμηλά από τη νίκη γηπεδούχου. Ο αλγόριθμος δηλαδή έχει πολύ καλή συμπεριφορά στις νίκες, είτε γηπεδούχου, είτε φιλοξενούμενου, και σημαντικά πιο χαμηλή απόδοση στις ισοπαλίες.

Ακολουθούν οι στοιχηματικές προσεγγίσεις, οι οποίες στηρίζονται στις προβλέψεις του Random Forest:

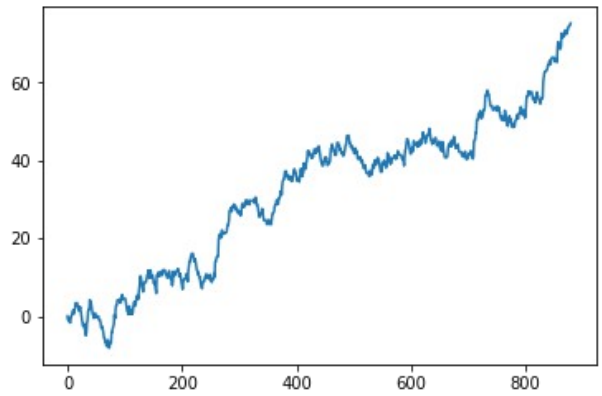
Προσέγγιση	Πλήθος αγώνων που υπάρχει πρόβλεψη	Πλήθος επιτυχημένων προβλέψεων	Ακρίβεια	Καθαρό κέρδος σε μονάδες	Μέση απόδοση επιτυχημένων προβλέψεων	Yield
1X2	880	487	55.34%	57.18	1.92	6.4%
1	542	327	**60.33%**	60.40	1.84	**11.1%**
X	61	22	36.06%	14.74	3.44	24.1%

2	277	138	49.81%	-17.95	1.87	-6.4%
1X	603	349	57.87%	**75.14**	1.94	**12.4%**
12	819	465	56.77%	42.45	1.85	5.1%
X2	338	160	47.33%	-3.21	2.09	-1.0%

Αρχικά, παρατηρεί κανείς ότι η προσέγγιση που ακολουθεί το σύστημα μόνο στις νίκες των φιλοξενούμενων ομάδων παρουσιάζει μικρή ζημία, η οποία μεταφέρεται και στις υπόλοιπες προσεγγίσεις που τις εμπεριέχουν (12, X2). Οι ισοπαλίες πετυχαίνουν πολύ υψηλό Yield (24.1%), που όμως μοιάζει να είναι πλασματικό λόγω του μικρού πλήθους αγώνων, που ταξινομήθηκαν σε ισοπαλία. Για το λόγο αυτό, η προσέγγιση μόνο με ισοπαλίες δεν θεωρείται ως μία από τις επικρατέστερες. Η προσέγγιση μόνο νίκες γηπεδούχων σημειώνει σημαντικό καθαρό κέρδος σε μονάδες (60.40), αλλά και υψηλό Yield (11.1%) σε ικανοποιητικό πλήθος αγώνων (542). Επιπλέον, ο συνδυασμός της νίκης γηπεδούχου με την ισοπαλία οδήγησε σε πολύ ενθαρρυντικά επίπεδα, με το Yield να φτάνει το 12.4%. Οι δύο αυτές προσεγγίσεις, που ξεχώρισαν με το υψηλό Yield και τα ποσοστά ακρίβειας, θεωρούνται ως οι επικρατέστερες για αυτή την υλοποίηση. Παρακάτω, παρουσιάζεται η πορεία των κερδών κατά την εξέλιξη των αγώνων στις δύο αυτές προσεγγίσεις.



Πορεία των κερδών σε μονάδες με προσέγγιση 1



Πορεία των κερδών σε μονάδες με προσέγγιση 1X

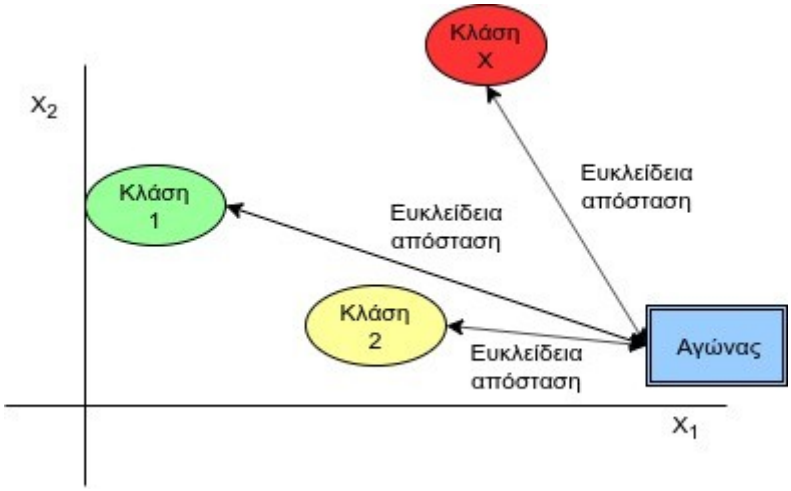
Αρχικά, είναι ευδιάκριτο ότι τα κέρδη σε μονάδες και των δύο προσεγγίσεων ακολουθούν αυξητική τάση σχεδόν σε όλη τη διάρκεια των 4 χρονιών. Παρατηρώντας κανείς τη γραφική αναπαράσταση των κερδών της προσέγγισης που ακολουθεί το σύστημα μόνο στις νίκες γηπεδούχου, διακρίνει κανείς μια μικρή ζημία κατά τα μισά της 1ης χρονιάς. Η ζημία αυτή παραμένει μέχρι και το τέλος της χρονιάς. Όμως, στη συνέχεια τα κέρδη αρχίζουν και ακολουθούν σταθερά ανοδική πορεία, με σποραδικές μεταπτώσεις. Η μέση ετήσια αύξηση σε κέρδη τα τελευταία τρία χρόνια είναι της τάξης +15 μονάδες. Από την άλλη, η προσέγγιση που ακολουθεί το σύστημα στις νίκες γηπεδούχων και στις ισοπαλίες, ξεκινάει με μια μικρή ζημία, ωστόσο ανακάμπτει γρήγορα και κλείνει την 1η χρονιά με κέρδος. Στη συνέχεια, η γραφική της παράσταση σημειώνει σταθερά ανοδική πορεία, με μικρότερες μεταπτώσεις από πριν. Η μέση ετήσια αυξητική πορεία της αυτά τα 4 χρόνια είναι της τάξης περίπου του +18 μονάδες. Είναι χαρακτηριστικό ότι και στα δύο γραφήματα υπάρχουν κατά διαστήματα μεγάλες αυξήσεις, δηλαδή επιτυχία του

συστήματος να προβλέπει σωστά μεγάλες αποδόσεις. Οι δύο αυτές προσεγγίσεις χαρακτηρίζονται ικανοποιητικές.

6. Γεωμετρική λύση

Η συγκεκριμένη υλοποίηση δεν είναι κάποιος ακόμα γνωστός αλγόριθμος Μηχανικής Μάθησης, όπως ήταν οι προηγούμενες υλοποιήσεις, αλλά αποτελεί μια διαφορετική προσέγγιση στο πρόβλημα που πραγματεύεται η παρούσα εργασία. Με μία απλή και πρωτότυπη παρατήρηση η τεχνική αυτή προσπαθεί να ανταγωνιστεί τις υπόλοιπες υλοποιήσεις. Αρχικά, η ιδέα ξεκίνησε από την απλή ερώτηση που μπορεί να κάνει κάποιος βλέποντας τα στοιχεία ενός αγώνα, με το αν το ματς αυτό, είναι πιο "κοντά" σε νίκη γηπεδούχου, σε ισοπαλία ή σε νίκη φιλοξενούμενου. Για το λόγο αυτό ήταν ανάγκη να προσδιοριστούν οι "εκπρόσωποι"κάθε αποτελέσματος (κάθε κλάσης). Για να γίνει αυτό, τα δεδομένα εκπαίδευσης χωρίστηκαν σε 3 σύνολα ανάλογα με το αποτέλεσμα τους, δηλαδή στο ένα σύνολο ενσωματώθηκαν μόνο οι "άσσοι", στο άλλο σύνολο μόνο οι ισοπαλίες και στο τελευταίο μόνο τα "διπλά¨. Αφού χωρίστηκαν έτσι τα δεδομένα, ως "εκπρόσωπος" κάθε συνόλου ορίστηκε το διάνυσμα που φέρει ως τιμές το μέσο όρο κάθε χαρακτηριστικού του συνόλου. Δηλαδή, ο "εκπρόσωπος" κάθε συνόλου είναι το γεωμετρικό κέντρο κάθε συνόλου, αν τα δεδομένα είχαν αναπαρασταθεί στο χώρο. Για κάθε διάνυσμα-αγώνα που το σύστημα χρειάζεται να ταξινομήσει σε μια κλάση, ή ισοδύναμα να προβλέψει το τελικό του αποτέλεσμα, υπολογίζεται η Ευκλείδεια του απόσταση από τους 3 εκπροσώπους. Τελικά, το διάνυσμα-αγώνας ταξινομείται στην κλάση, όπου απέχει τη μικρότερη απόσταση. Η Ευκλείδεια απόσταση εκφράζει πόσο όμοια είναι τα δεδομένα μεταξύ τους, με μικρότερη απόσταση να έχουν τα δεδομένα που είναι περισσότερο όμοια μεταξύ τους. Η συγκεκριμένη λύση ονομάστηκε γεωμετρική, γιατί ξεφεύγει από τους αλγορίθμους Μηχανικής Μάθησης και έρχεται πιο κοντά στη γεωμετρία, υπολογίζοντας κέντρα και ευκλείδειες αποστάσεις.

Ακολούθως, παρουσιάζεται σχηματικά ένα παράδειγμα ταξινόμησης ενός αγώνα, με την απλοποίηση ότι έχουν επιλεχτεί μόνο δύο χαρακτηριστικά (X_1 , X_2) ως τα επικρατέστερα. Οι 3 "εκπρόσωποι" των 3 κλάσεων παρουσιάζονται με πράσινο, κόκκινο και κίτρινο χρώμα, ενώ το διάνυσμα-αγώνας με μπλε. Στο ίδιο σχήμα φαίνονται, επιπλέον, και οι αποστάσεις του αγώνα από τις 3 κλάσεις. Όπως φαίνεται στο σχήμα, η κοντινότερη κλάση στον αγώνα είναι η κλάση της νίκης της φιλοξενούμενης ομάδας. Επομένως, η πρόβλεψη του συστήματος αυτού για τον συγκεκριμένο αγώνα θα είναι και νίκη της φιλοξενούμενης ομάδος, γιατί ο αγώνας "μοιάζει" περισσότερο με το μέσο αγώνα που τελειώνει "διπλό".



Τα χαρακτηριστικά που επιλέχθηκαν για το συγκεκριμένο σύστημα είναι: *2, 9, 13, 28, 40, 41, 48, 57, 58, 59, 63, 64, 65, 69*. Οι τιμές των features, όπως αναμενόταν (καθώς το συγκεκριμένο σύστημα βασίζεται στις αποστάσεις), κανονικοποιήθηκαν.

Παρακάτω παρουσιάζεται το confusion matrix των προβλέψεων του συστήματος στις 4 επιλεγμένες χρονιές του Αγγλικού Πρωταθλήματος μαζί με την αναφορά και τα στατιστικά κατηγοριοποίησης του συστήματος:

	Πρόβλεψη άσσου	Πρόβλεψη Ισοπαλία	Πρόβλεψη διπλού
Αποτέλεσμα άσσος	275	39	112
Αποτέλεσμα Ισοπαλία	92	30	82
Αποτέλεσμα διπλό	65	23	162

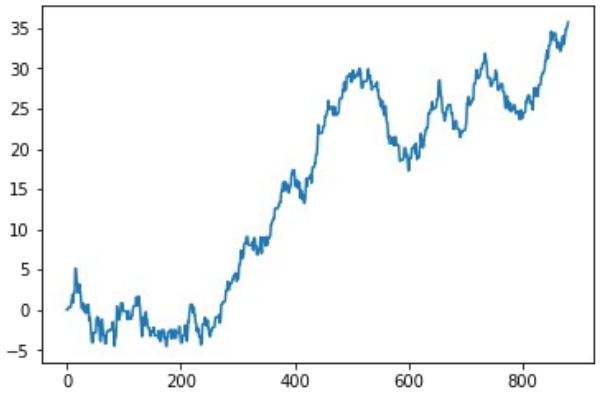
	precision	recall	f1-score	support
Άσσος	0.64	0.65	0.64	426
Ισοπαλία	0.33	0.15	0.20	204
Διπλό	0.46	0.65	0.53	250

Ο αλγόριθμος αυτός έχει μια τάση να κατηγοριοποιεί τους αγώνες σε νίκη γηπεδούχου και σε νίκη φιλοξενούμενου και πολύ λιγότερο σε ισοπαλία. Παρατηρεί κανείς πως τα πηγαίνει αρκετά καλά στις προβλέψεις του στις κλάσεις "άσσου" και "διπλού". Συγκεκριμένα, ανιχνεύει και ταξινομεί στη σωστή κλάση το 65% τόσο τις νίκης γηπεδούχου, όσο και της νίκης φιλοξενούμενου. Η ακρίβεια στην κλάση του "άσσου" είναι πολύ υψηλή, ενώ στην κλάση του διπλού είναι αρκετά πιο χαμηλή, ωστόσο το f1-score είναι πολύ ικανοποιητικό. Όσον αφορά τις ισοπαλίες, παρά την ικανοποιητική ακρίβεια του αλγορίθμου (33%), η ανίχνευση και σωστή ταξινόμηση της συγκεκριμένης κλάσης παραμένει πολύ χαμηλά.

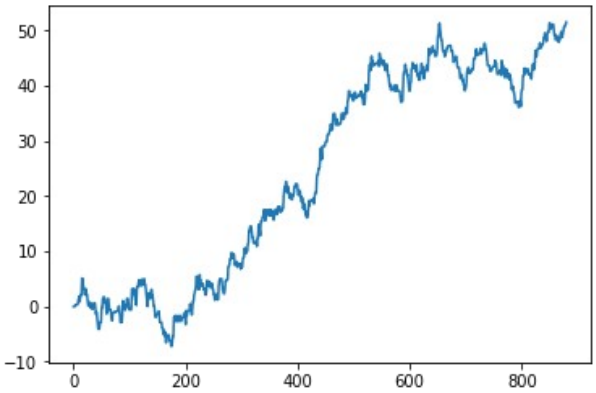
Ακολουθούν οι στοιχηματικές προσεγγίσεις, οι οποίες στηρίζονται στις προβλέψεις του συστήματος:

Προσέγγιση	Πλήθος αγώνων που υπάρχει πρόβλεψη	Πλήθος επιτυχημένων προβλέψεων	Ακρίβεια	Καθαρό κέρδος σε μονάδες	Μέση απόδοση επιτυχημένων προβλέψεων	Yield
1X2	880	467	53.06%	42.04	1.97	4.7%
1	432	275	**63.65%**	35.76	1.70	**8.2%**
X	92	30	32.60%	15.78	3.59	17.1%
2	356	162	45.50%	-9.50	2.13	-2.6%
1X	524	305	**58.20%**	**51.54**	1.88	**9.8%**
12	788	437	55.45%	26.26	1.86	3.3%
X2	448	192	42.85%	6.28	2.36	1.4%

Αρχικά, παρατηρώντας κανείς τον παραπάνω πίνακα συμπεραίνει πως μόνο μια προσέγγιση παρουσιάζει ζημία, αυτή που ακολουθεί το σύστημα μόνο στις νίκες των φιλοξενούμενων ομάδων. Όπως έχει αναφερθεί και στις προηγούμενες υλοποιήσεις, αυτή η ζημία μεταφέρεται και σε όλες τις άλλες προσεγγίσεις που περιλαμβάνουν "διπλά". Η προσέγγιση 1 φαίνεται να τα πηγαίνει πολύ καλά, πετυχαίνοντας υψηλή ακρίβεια και αρκετά υψηλό Yield, σε ικανοποιητικό πλήθος αγώνων (432), αποτελώντας έτσι μία από τις επικρατέστερες προσεγγίσεις. Η προσέγγιση X πετυχαίνει πολύ υψηλό Yield, όμως το πλήθος αγώνων που δοκιμάστηκε η συγκεκριμένη προσέγγιση είναι πολύ μικρό για να προκύψουν ασφαλή συμπεράσματα. Η προσέγγιση, όμως, 1X που συνδυάζει τις δύο τελευταίες προσεγγίσεις, μπορεί να χαρακτηριστεί ως μία από τις επικρατέστερες, πετυχαίνοντας πολύ υψηλό καθαρό κέρδος και πολύ υψηλό Yield. Οι υπόλοιπες προσεγγίσεις δεν παρουσιάζουν υψηλό Yield, γιατί περιέχουν "διπλά", οπότε εμπεριέχουν μέσα μια μικρή ζημία, για το λόγο αυτό, μόνο οι προσεγγίσεις 1 και 1X είναι οι επικρατέστερες. Παρακάτω, παρουσιάζεται η πορεία των κερδών κατά την εξέλιξη των αγώνων στις δύο αυτές προσεγγίσεις.



Πορεία των κερδών σε μονάδες με προσέγγιση 1



Πορεία των κερδών σε μονάδες με προσέγγιση 1X

Παρατηρώντας τα δύο γραφήματα είναι φανερή η ανοδική τάση που έχουν τα καθαρά κέρδη σε μονάδες και οι δύο προσεγγίσεις σχεδόν σε όλη τη διάρκεια των 4 αγωνιστικών χρονιών, που αποτελούν το test. Η πορεία των κερδών της προσέγγισης 1 την 1η χρονιά είναι πολύ κοντά στο 0, χωρίς να σημειώνονται καθαρά κέρδη ή ζημία. Τα 3 επόμενα χρόνια η προσέγγιση αυτή παρουσιάζει σταθερά ανοδική πορεία στα κέρδη της, με κάποιες μεταπτώσεις που όμως γρήγορα εξαλείφονται. Η μέση ετήσια αυξητική τάση αυτής της προσέγγισης είναι της τάξης περίπου +10 μονάδες. Από την άλλη, η προσέγγιση 1X κλείνει την 1η χρονιά με οριακό κέρδος και συνεχίζει με σταθερά αυξητική τάση τα επόμενα 3 χρόνια. Η μέση ετήσια αυξητική τάση αυτής της προσέγγισης είναι της τάξης +14 μονάδες, οδηγώντας προφανώς και σε μεγαλύτερο καθαρό κέρδος. Μια σημαντική διαφορά ανάμεσα στις δύο επικρατέστερες προσεγγίσεις είναι ότι η προσέγγιση 1X παρουσιάζει πολύ μικρότερες μεταπτώσεις συγκριτικά με την προσέγγιση 1. Τέλος, και οι δύο προσεγγίσεις σημειώνουν απότομες αυξήσεις κατά τη διάρκεια των τεσσάρων χρονιών, γεγονός που δείχνει ότι το σύστημα και εδώ προβλέπει σωστά υψηλές στοιχηματικές αποδόσεις.

7. 1X-X2

Η επόμενη υλοποίηση αποτελείται από έναν συνδυασμό αλγορίθμων Μηχανικής Μάθησης, που συνδυάζονται για να προκύψει η τελική πρόβλεψη για τον αγώνα. Η ιδέα πίσω από αυτή την υλοποίηση είναι ότι η ισοπαλία είναι το ενδιάμεσο αποτέλεσμα μεταξύ της νίκης γηπεδούχου και της νίκης του φιλοξενούμενου. Για το λόγο αυτό χρησιμοποιούνται δύο ταξινομητές, με τον έναν να εξειδικεύεται στη νίκη γηπεδούχου και στην ισοπαλία και τον άλλον να εξειδικεύεται στην ισοπαλία και στη νίκη του φιλοξενούμενου. Πιο αναλυτικά, οι 2 ταξινομητές εκπαιδεύονται σε διαφορετικά δεδομένα. Ο ταξινομητής Α εκπαιδεύεται σε δεδομένα που περιέχουν μόνο νίκες γηπεδούχων και ισοπαλίες, ενώ ο ταξινομητής Β μόνο σε δεδομένα που περιέχουν ισοπαλίες και νίκες φιλοξενούμενων. Αυτό έχει ως αποτέλεσμα ο ταξινομητής Α να προβλέπει μόνο "άσσους" και ισοπαλίες, γιατί δεν έχει εκπαιδευτεί σε "διπλά", και αντίστοιχα, ο ταξινομητής Β να προβλέπει μόνο ισοπαλίες και "διπλά". Ουσιαστικά, κάθε ταξινομητής προβλέπει πιο από τα δύο αυτά αποτελέσματα είναι πιο πιθανό. Στη συνέχεια, από την ταξινόμηση-πρόβλεψη των ταξινομητών προκύπτει και η τελική πρόβλεψη του συστήματος με τον ακόλουθο τρόπο:

- Αν οι δύο ταξινομητές συμφωνούν, δηλαδή προβλέπουν και οι δύο ισοπαλία, τότε η τελική πρόβλεψη του συστήματος είναι ισοπαλία.

- Αν ο ταξινομητής Α προβλέπει ισοπαλία και ο ταξινομητής Β νίκη της φιλοξενούμενης ομάδας, τότε το σύστημα προβλέπει "διπλό". Αυτό συμβαίνει, γιατί ο ταξινομητής Α θεωρεί πιο πιθανό το αποτέλεσμα να είναι ισόπαλο παρά νίκη γηπεδούχου και ο ταξινομητής Β, που εξειδικεύεται στην ισοπαλία και στη νίκη του φιλοξενούμενου, θεωρεί πιο πιθανό αποτέλεσμα το "διπλό". Επομένως, υπερισχύει η πρόβλεψη του ταξινομητή Β.

- Αν ο ταξινομητής Α προβλέπει νίκη γηπεδούχου και ο ταξινομητής Β ισοπαλία, τότε το σύστημα προβλέπει "άσσο". Ακριβώς με όμοια λογική με το παραπάνω, προκύπτει ότι υπερισχύει η πρόβλεψη του ταξινομητή Α.

- Αν ο ταξινομητής Α προβλέπει νίκη γηπεδούχου και ο ταξινομητής Β προβλέπει νίκη φιλοξενούμενου, τότε οι δύο ταξινομητές συμφωνούν ότι η ισοπαλία δεν είναι το πιο πιθανό αποτέλεσμα. Σε αυτή την περίπτωση, πρέπει να βρεθεί τρόπος να αποφασίζει το σύστημα πιο αποτέλεσμα να επιλέγει, "άσσο" ή "διπλό". Η λύση που δόθηκε ήταν και η πιο απλή, το σύστημα επιλέγει νίκη γηπεδούχου, γιατί αυτό είναι το πιο συχνό αποτέλεσμα με ποσοστό εμφάνισης κοντά στο 50%.

Κάθε ταξινομητής εκπαιδεύτηκε σε διαφορετικά χαρακτηριστικά και επιλέχτηκε ένας από τους αλγόριθμους Μηχανικής Μάθησης που αναφέρθηκαν προηγουμένως. Η διαδικασία επιλογής αλγορίθμου και χαρακτηριστικών έγινε κατά τα γνωστά με τη χρήση του grid search. Πιο συγκεκριμένα, ο ταξινομητής Α υλοποιήθηκε με χρήση του αλγορίθμου SVM με features: *39, 41, 57, 59, 63, 65, 71,* χωρίς να γίνει κάποια επεξεργασία στις τιμές των χαρακτηριστικών αυτών. Οι τιμές των παραμέτρων που επιλέχτηκαν ήταν για kernel το RBF, για C την τιμή 90 και για gamma το 1. Ο ταξινομητής Β υλοποιήθηκε με χρήση του αλγορίθμου k-NN με χαρακτηριστικά: *39, 41, 43, 44, 57, 59, 63, 65, 71,* τα οποία κανονικοποιήθηκαν. Ως παράμετρο για το πλήθος των γειτόνων που θα ληφθούν υπόψιν κατά την εκτέλεση του αλγορίθμου του ταξινομητή Β επιλέχτηκε το 3.

Παρακάτω παρουσιάζεται το confusion matrix των προβλέψεων του συστήματος στις 4 επιλεγμένες χρονιές του Αγγλικού Πρωταθλήματος μαζί με την αναφορά και τα στατιστικά κατηγοριοποίησης του συστήματος:

	Πρόβλεψη άσσου	Πρόβλεψη Ισοπαλία	Πρόβλεψη διπλού
Αποτέλεσμα άσσος	358	33	35
Αποτέλεσμα Ισοπαλία	161	26	17
Αποτέλεσμα διπλό	164	25	61

	precision	recall	f1-score	support
Άσσος	0.52	0.84	0.65	426
Ισοπαλία	0.31	0.13	0.18	204
Διπλό	0.54	0.24	0.34	250

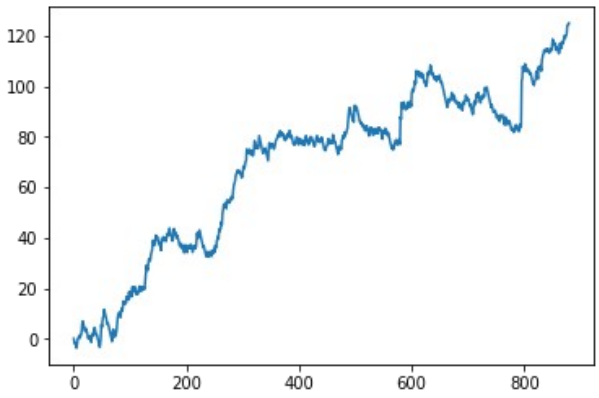
Το σύστημα παρατηρεί κανείς ότι προβλέπει κατά κύριο λόγο νίκη γηπεδούχου και πολύ λιγότερο τα άλλα δύο αποτελέσματα. Αυτό ήταν κάτι αναμενόμενο, καθώς η λύση που δόθηκε όταν και οι δύο ταξινομητές προβλέπουν νίκη, ήταν να υπερισχύει η πρόβλεψη του "άσσου", ως το πιο πιθανό αποτέλεσμα. Αυτό είχε ως αποτέλεσμα το παραπάνω, δηλαδή πολλές προβλέψεις "άσσων" και πολύ λιγότερες ισοπαλίες και "διπλά". Όπως ήταν αναμενόμενο, ο αλγόριθμος ανιχνεύει και προβλέπει σωστά την πλειονότητα των "άσσων" (85%), ωστόσο με ακρίβεια στο 52%, απόρροια της απλοποίησης που έγινε και ταξινομήθηκαν πολλά ματς στην κλάση νίκης γηπεδούχου. Στις άλλες δύο κλάσεις, η ακρίβεια είναι ικανοποιητική έως και καλή, όμως η ανίχνευση και σωστή πρόβλεψη (recall) κυμαίνεται σε πολύ χαμηλά επίπεδα, με αποτέλεσμα να προκύπτει στις κλάσεις αυτές και χαμηλό f1-score.

Ακολουθούν οι στοιχηματικές προσεγγίσεις, οι οποίες στηρίζονται στις προβλέψεις του συστήματος:

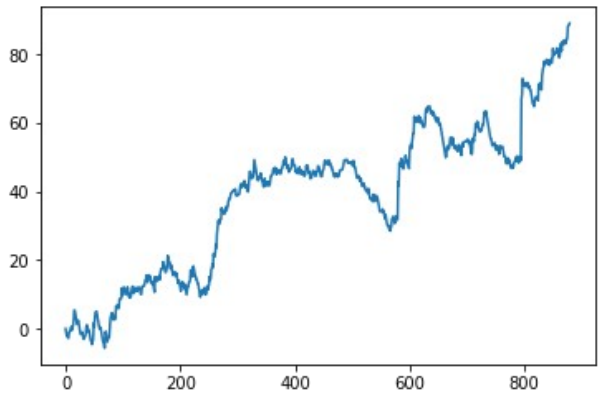
Προσέγγιση	Πλήθος αγώνων που υπάρχει πρόβλεψη	Πλήθος επιτυχημένων προβλέψεων	Ακρίβεια	Καθαρό κέρδος σε μονάδες	Μέση απόδοση επιτυχημένων προβλέψεων	Yield
1X2	880	445	50.56%	**125.42**	2.25	**14.2%**
1	683	358	52.41%	**89.10**	2.15	**13.0%**
X	84	26	30.95%	14.04	3.77	16.5%
2	113	61	53.98%	22.27	2.21	19.7%
1X	767	384	50.06%	**103.14**	2.26	**13.4%**
12	796	419	52.63%	**111.38**	2.16	**14.0%**
X2	197	87	44.16%	36.32	2.68	18.4%

Αρχικά, βλέποντας τον παραπάνω πίνακα είναι εμφανές ότι όλες οι προσεγγίσεις είναι αρκετά κερδοφόρες, σημειώνοντας μάλιστα παντού διψήφιο Yield. Ειδικά, η προσέγγιση 1X2 σημειώνει πολύ υψηλά καθαρά κέρδη σε μονάδες (125), με χαμηλή ακρίβεια στις προβλέψεις, που όμως ισοσταθμίζεται από την πολύ υψηλή μέση απόδοση επιτυχημένων προβλέψεων (2.25). Γίνεται εμφανές εδώ, ότι μπορεί ένα σύστημα να έχει χαμηλή ακρίβεια, ωστόσο αν επιτυγχάνει να προβλέψει σωστά υψηλές στοιχηματικές αποδόσεις, το σύστημα είναι κερδοφόρο. Όλες οι

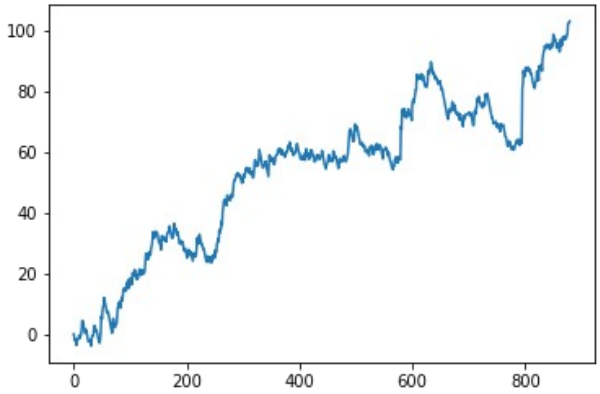
προσεγγίσεις είναι κερδοφόρες και έχουν υψηλό Yield, όμως αυτές που ξεχωρίζουν είναι οι 1X2, η 1, η 1X και η 12. Οι υπόλοιπες υλοποιούνται σε πολύ μικρό πλήθος αγώνων και δεν μπορούν να προκύψουν ασφαλή συμπεράσματα. Γενικά, να αναφερθεί ότι αυτή η υλοποίηση χαρακτηρίζεται από μέτρια και χαμηλή ακρίβεια, ωστόσο επιτυγχάνονται πολύ υψηλά καθαρά κέρδη. Παρακάτω, παρουσιάζεται η πορεία των κερδών κατά την εξέλιξη των αγώνων στις προσεγγίσεις που ξεχώρισαν.



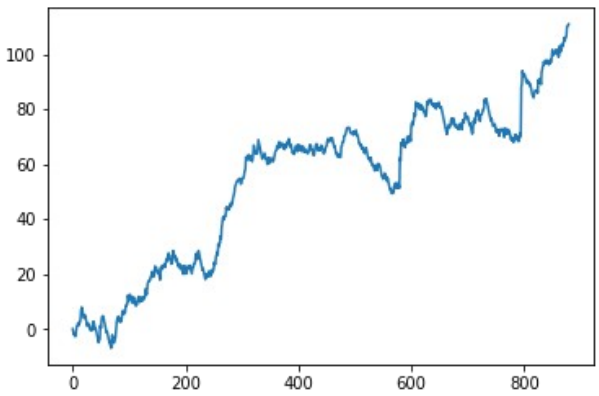
Πορεία των κερδών σε μονάδες με προσέγγιση 1X2



Πορεία των κερδών σε μονάδες με προσέγγιση 1



Πορεία των κερδών σε μονάδες με προσέγγιση 1X

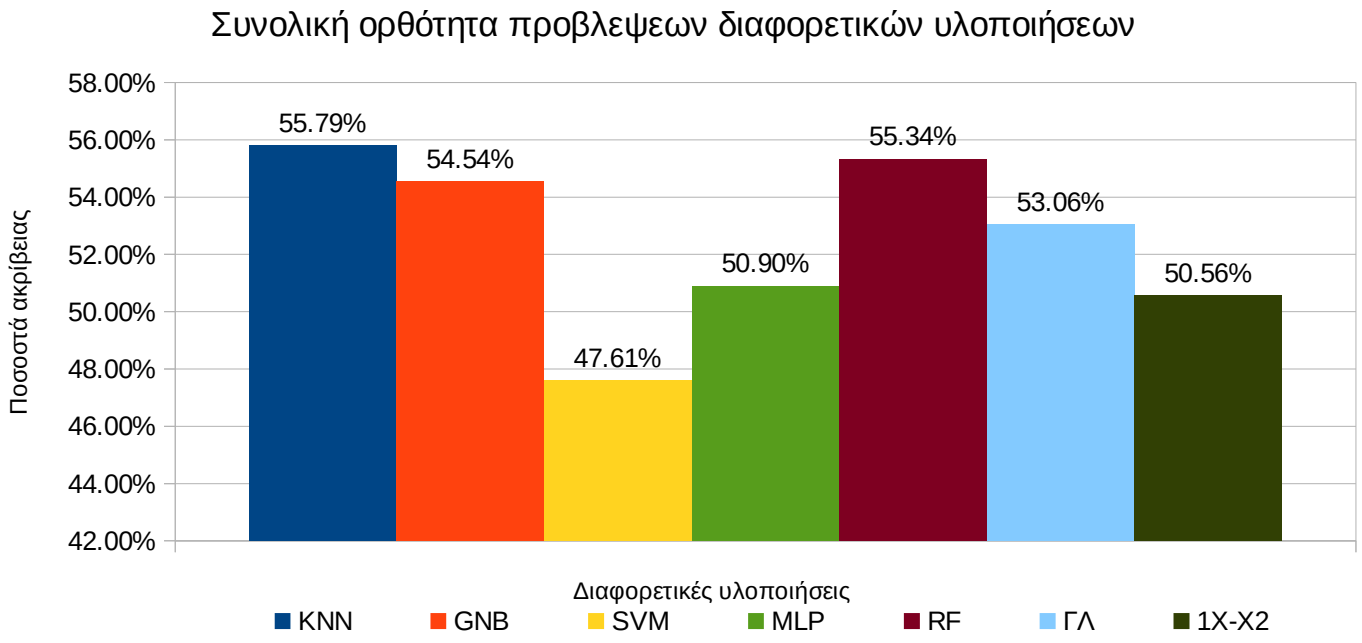


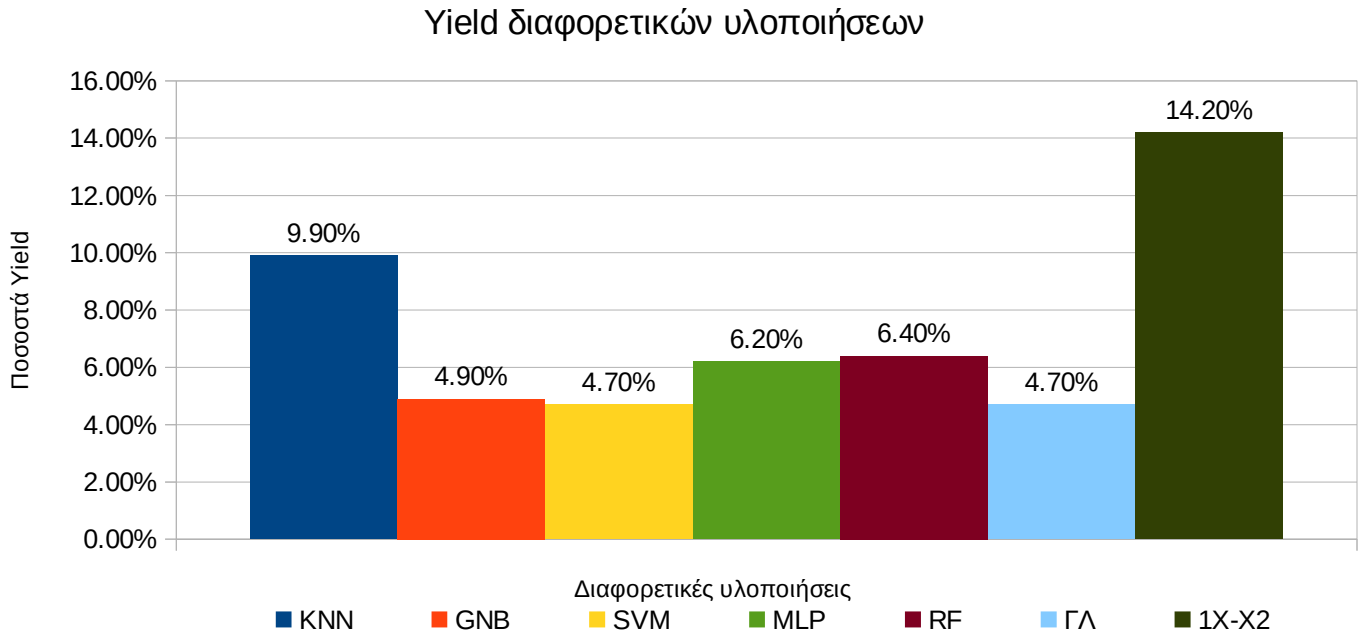
Πορεία των κερδών σε μονάδες με προσέγγιση 12

Παρατηρώντας κανείς τις γραφικές αναπαραστάσεις των κερδών σε μονάδες σε αυτές τις 4 προσεγγίσεις διακρίνει εύκολα την σταθερά ανοδική τάση που υπάρχει. Όλες οι προσεγγίσεις σημειώνουν σταθερή άνοδο από την $1^{η}$ κιόλας αγωνιστική χρονιά. Όπως φάνηκε στις προηγούμενες αναλύσεις την $1^{η}$ χρονιά υπήρχε οριακό κέρδος ή ζημία και τις επόμενες χρονιές υπήρξε σταθερή άνοδο. Αυτό παρατηρείται ότι εδώ δεν συμβαίνει. Επιπλέον, είναι φανερό σε αυτές τις προσεγγίσεις σπάνιες και απότομες μεταπτώσεις, όπως για παράδειγμα λίγο πριν τη συμπλήρωση 600 αγώνων και λίγο πριν τα 800. Αυτές οι μεταπτώσεις όμως, γρήγορα ξεπερνιούνται με μεγάλες ανόδους (δηλαδή εύστοχες υψηλές στοιχηματικές αποδόσεις) και το σύστημα συνεχίζει την ανοδική του πορεία. Η μέση ετήσια αυξητική τάση σε όλη τη διάρκεια αυτών των αγώνων είναι από +22 (προσέγγιση 1) μέχρι και +32 (προσέγγιση 1X2) σε μονάδες. Η υλοποίηση αυτή πέτυχε πολύ υψηλά νούμερα, και στο κέρδος και στο Yield.

5.4 Σύνοψη Μεθόδων

Παρατηρεί κανείς πως και οι 7 υλοποιήσεις πέτυχαν τον στόχο του να είναι κερδοφόρες. Μάλιστα, κάποιες υλοποιήσεις πέτυχαν και υψηλά καθαρά κέρδη και μεγάλο Yield. Στην υποενότητα αυτή, θα συγκριθούν οι υλοποιήσεις και θα απαντηθεί, ποια υλοποίηση τα πήγε καλύτερα από τις άλλες. Παρακάτω, παρουσιάζονται τα γραφήματα των υλοποιήσεων συνολικά (συμπεριλαμβάνονται όλοι οι αγώνες του test) ως προς την ακρίβεια τους και ως προς το Yield.





Όπως έχει ήδη αναφερθεί, θα αξιολογηθούν τα συστήματα ως προς το Yield που επιτυγχάνουν και όχι ως προς την ακρίβεια. Το γράφημα με τη συνολική ορθότητα των μεθόδων παρουσιάζεται εδώ για λόγους πληρότητας. Με μια πρώτη ματιά λοιπόν, η υλοποίηση 1X-X2 φαίνεται να ξεχωρίζει από τις υπόλοιπες, πετυχαίνοντας ένα πολύ υψηλό Yield (14% !!). Μάλιστα, όπως φάνηκε και στα αναλυτικά της αποτελέσματα προηγουμένως, σε όλες τις προσεγγίσεις το Yield ήταν αρκετά υψηλό. Η δεύτερη καλύτερη επίδοση είναι αυτή του k-NN, όπου πετυχαίνει Yield κοντά στο 10%. Οι υπόλοιπες υλοποιήσεις τα πήγαν αρκετά καλά και το Yield τους κυμάνθηκε στο 5-6%. Γενικά, τα αποτελέσματα είναι παραπάνω από ικανοποιητικά και ελπιδοφόρα για το μέλλον, μιας και ένας καλός Tipster πετυχαίνει Yield κοντά στο 8%. Φάνηκε λοιπόν, πως είναι δυνατή η ανάπτυξη συστημάτων Τεχνητής Νοημοσύνης, που μπορεί να προβλέπει αποτελέσματα αγώνων ποδοσφαίρου και χρησιμοποιώντας τις στοιχηματικές αποδόσεις να φτάσει σε κέρδος. Αυτός ήταν και ο κύριος στόχος της εργασίας, να δείξει το παραπάνω και να τονίσει τις εκπληκτικές δυνατότητες της Μηχανικής Μάθησης.

Όσον αφορά τώρα την ορθότητα των υλοποιήσεων, ο k-NN και ο Random Forest μαζί με τον Gaussian Naive Bayes, πέτυχαν ικανοποιητική ορθότητα, κοντά στο 55%, που κοιτώντας τα αποτελέσματα άλλων ερευνών της βιβλιογραφίας αποτελούν αρκετά καλά αποτελέσματα. Όπως έχει αναφερθεί και προηγουμένως, η υψηλή ορθότητα δεν ήταν και ο στόχος της παρούσας εργασίας.

Κεφάλαιο 6: Συμπεράσματα

6.1 Συμπεράσματα και παρατηρήσεις

Στο προηγούμενο κεφάλαιο παρουσιάστηκαν όλες οι υλοποιήσεις και τα αποτελέσματα τους, τα οποία ήταν κάτι παραπάνω από θετικά και ελπιδοφόρα. Οι υλοποιήσεις σημείωσαν κέρδος κατά τη διάρκεια αυτών των 4 χρόνων, διατηρώντας αυξητική τάση σε όλη σχεδόν την πορεία. Αυτός ήταν εξάλλου και ο στόχος της εργασίας, η ανάπτυξη μοντέλων που μπορούν να οδηγήσουν σε κέρδος αξιοποιώντας τις στοιχηματικές αποδόσεις. Σε όλη τη διάρκεια της εργασίας, από την αρχή μέχρι το τέλος, προέκυψαν ενδιαφέροντα συμπεράσματα και παρατηρήσεις, τα οποία θα αναφερθούν παρακάτω.

Τα κυριότερα συμπεράσματα και παρατηρήσεις που προέκυψαν και είναι άξια αναφοράς:

1. ***<u>Η υψηλή ορθότητα δεν συνεπάγεται απαραίτητα υψηλά κέρδη</u>***

Η υψηλή ορθότητα δεν καθιστά ένα σύστημα κερδοφόρο, ή αντίστοιχα, ένα σύστημα με χαμηλή ορθότητα δεν σημαίνει ότι είναι ζημιογόνο. Όπως αναφέρθηκε στα προηγούμενα κεφάλαια, η κερδοφορία εξαρτάται από το γινόμενο της μέσης στοιχηματικής απόδοσης που προβλέπει σωστά το σύστημα και της ορθότητας του συστήματος. Ως μέτρο κερδοφορίας, χρησιμοποιήθηκε το Yield. Όπως φάνηκε και από τα αποτελέσματα των υλοποιήσεων στο προηγούμενο κεφάλαιο, υψηλή ορθότητα δεν συνεπάγεται και υψηλό Yield. Στο προηγούμενο κεφάλαιο, παρουσιάστηκαν τα ποσοστά ακρίβειας και ορθότητας όλων των υλοποιήσεων, συμπεριλαμβάνοντας όλες τις προβλέψεις τους, και από κάτω γράφημα με το Yield κάθε υλοποίησης.

Όπως φάνηκε λοιπόν, η υλοποίηση 1X-X2 έχει την 6^{η} καλύτερη ορθότητα μεταξύ των συστημάτων, όμως έχει με διαφορά το υψηλότερο Yield. Ομοίως, παρατηρεί κανείς ότι η υλοποίηση SVM έχει τη χειρότερη ορθότητα, έχοντας όμως το ίδιο Yield με τη Γεωμετρική Λύση, η οποία έχει 6% καλύτερη ορθότητα. Επομένως, γίνεται φανερό ότι για την κατασκευή ενός κερδοφόρου συστήματος, η υψηλή ορθότητα δεν αποτελεί τη λύση, αλλά πρέπει να συνδυάζεται παράλληλα και με με ικανοποιητικές στοιχηματικές αποδόσεις.

2. ***<u>Η ορθότητα προβλέψεων απαλείφει τις ισοπαλίες</u>***

Στις πρώτες απόπειρες και στα πρώτα μοντέλα που σχεδιάστηκαν, το grid search επέλεγε τις παραμέτρους του μοντέλου που πετύχαινε την υψηλότερη ορθότητα στις προβλέψεις του. Με αυτό τον τρόπο, δημιουργήθηκαν μοντέλα με πολύ υψηλή ορθότητα (περίπου 57-60%), που όμως δεν προέβλεπαν καμία ισοπαλία. Αυτό συνέβαινε σε όλους τους αλγορίθμους Μηχανικής Μάθησης, που χρησιμοποιήθηκαν, εκτός από τον GNB, που δεν διαθέτει παραμέτρους. Σε παρόμοιο αποτέλεσμα κατέληξαν και αρκετές έρευνες που βρέθηκαν στο διαδίκτυο [6], οι οποίες πετύχαιναν υψηλή ορθότητα, όμως δεν προέβλεπαν ισοπαλίες. Ως προς το κέρδος, να αναφερθεί ότι τα μοντέλα αυτά πετύχαιναν σημαντικά μικρότερο κέρδος, ενώ σε αρκετές περιπτώσεις σημείωναν και ζημία, συγκριτικά με τα τελικά μοντέλα. Αυτό ήταν και το μεγαλύτερο πρόβλημα που συναντήθηκε

κατά την εκπόνηση της εργασίας, καθώς τα μοντέλα πετύχαιναν υψηλή ορθότητα, όμως δεν προέβλεπαν ισοπαλίες και δεν πετύχαιναν ικανοποιητικό κέρδος. Περνώντας ο χρόνος και καταλαβαίνοντας καλύτερα το πρόβλημα και τα ζητούμενα, προέκυψε η ιδέα για το grid search που περιγράφηκε σε προηγούμενο κεφάλαιο και τελικά χρησιμοποιήθηκε. Η ιδέα δηλαδή, ότι στο grid search πρέπει να επιλεγούν οι παράμετροι που οδηγούν σε πιο κερδοφόρο μοντέλο, με τη χρήση της συνθήκης (3.2), που υπαγορεύει ότι το γινόμενο ορθότητας και μέσης επιτυχημένης στοιχηματικής απόδοσης πρέπει να είναι μεγαλύτερο της μονάδος, οδήγησε στη λύση του προβλήματος. Το συμπέρασμα που προέκυψε είναι ότι η ισοπαλία είναι το πιο σπάνιο αποτέλεσμα, επομένως και το πιο δύσκολο να προβλεφθεί, αλλά παράλληλα και αυτό που έχει κατά μέσο όρο τη μεγαλύτερη απόδοση. Για το λόγο αυτό, όταν ζητούμενο ήταν η ορθότητα, τα μοντέλα επέλεγαν να προβλέπουν μόνο νίκη γηπεδούχου και νίκη φιλοξενούμενου, μη ρισκάροντας να προβλέψουν ισοπαλίες, οι οποίες είναι η πιο σπάνια κλάση. Όταν όμως, ζητούμενο έγινε η απόκτηση κέρδους, επιλέχτηκαν οι κατάλληλοι παράμετροι από τα μοντέλα, τα οποία άρχισαν να ρισκάρουν και να προβλέπουν ισοπαλίες. Η πρόβλεψη ισοπαλιών μείωσε την ορθότητα των προβλέψεων και αύξησε το κέρδος, καθώς προστέθηκαν στις επιτυχημένες αποδόσεις αυτές των ισοπαλιών που είναι αρκετά υψηλές. Επομένως, συμπεραίνεται ότι η υψηλή ορθότητα απορρίπτει, "σκοτώνει", τις ισοπαλίες στα διάφορα μοντέλα.

3. ***Οι στοιχηματικές αποδόσεις βοηθούν πολύ στις προβλέψεις***

Οι στοιχηματικές αποδόσεις, όπως έχει ήδη αναφερθεί, παράγονται από τους bookmakers, λαμβάνοντας υπόψιν τους πολλά δεδομένα και πληροφορίες που έχουν διαθέσιμα και είναι σημαντικά κατά την πρόβλεψη του αποτελέσματος ενός αγώνα. Ουσιαστικά, δημιουργούν προβλέψεις για τους αγώνες και ανάλογα διαμορφώνουν και τις αποδόσεις. Δηλαδή, αν ένα αποτέλεσμα είναι αρκετά πιθανό να πραγματοποιηθεί, τότε η απόδοση του θα είναι χαμηλή, ώστε η εταιρεία να επιστρέψει μικρό ποσό. Όπως γίνεται αντιληπτό, οι στοιχηματικές αποδόσεις περιέχουν τη γνώση ενός ειδικού και πολλή πληροφορία σχετικά με τον αγώνα, το οποίο τις καθιστά πολύ χρήσιμες για τα μοντέλα που αναπτύχθηκαν. Μάλιστα, όλα τα μοντέλα απέδωσαν καλύτερα με τη συμμετοχή των στοιχηματικών αποδόσεων στα χαρακτηριστά εκπαίδευσης τους. Μιλώντας για στοιχηματικές αποδόσεις, εννοούνται τόσο οι αποδόσεις για το τελικό αποτέλεσμα, όσο και οι αποδόσεις για το Asian Handicap, που βελτίωσαν σημαντικά την επίδοση των συστημάτων.

4. ***Ελπίδες για ανάπτυξη ισχυρού συστήματος προβλέψεων στο μέλλον***

Ίσως το πιο σημαντικό αποτέλεσμα αυτής της εργασίας είναι η ελπίδα, ότι μπορεί να αναπτυχθεί ένα ισχυρό μοντέλο, το οποίο σε βάθος χρόνου να οδηγεί σε κέρδος. Στην παρούσα εργασία, αναπτύχθηκαν 7 διαφορετικά μοντέλα, που σημείωσαν εξαιρετικές επιδόσεις στο test και θα έχει ενδιαφέρον να φανεί αν και μελλοντικά θα συνεχίσουν να δημιουργούν κέρδος. Η ελπίδα δικαιολογείται από το γεγονός, ότι το κέρδος σε κάθε μοντέλο ήρθε σταδιακά και μεθοδικά. Όπως φάνηκε και στις παρουσιάσεις των επικρατέστερων υλοποιήσεων στο προηγούμενο κεφάλαιο, η πορεία των κερδών σε μονάδες είχε σχεδόν σε όλη τη διάρκεια των 4 χρονιών σταθερή αυξητική τάση και μικρές μεταπτώσεις. Γεγονός, που δίνει ελπίδες ότι τα μοντέλα θα συνεχίσουν και στο μέλλον να παράγουν κέρδος, με τον ίδιο ή διαφορετικό ρυθμό. Να αναφερθεί εδώ, ότι τα Yield των επικρατέστερων υλοποιήσεων είναι αρκετά υψηλά και αξιοζήλευτα στον κόσμο του στοιχήματος (ένας πολύ καλός Tipster έχει Yield λίγο πάνω από το 10%). Είναι πολύ πιθανό, επομένως, στο μέλλον οι προβλέψεις να προκύπτουν από υπολογιστικά μοντέλα και να αλλάξουν τελείως τον

κόσμο του στοιχήματος. Κάτι τέτοιο, δεν μπορεί να θεωρείται απίθανο μιας και η ανάπτυξη της Μηχανικής Μάθησης στα πλαίσια των προβλέψεων είναι τεράστια τα τελευταία χρόνια και συνέχεια δημιουργούνται νέες προκλήσεις και δυνατότητες.

5. Η πρόσβαση σε πολλές και μεγάλες στοιχηματικές εταιρείες αυξάνουν το κέρδος

Η μεγάλη γκάμα διαθέσιμων στοιχηματικών εταιρειών είναι ένα πολύ σημαντικό όπλο για τους ανθρώπους που ασχολούνται με το στοίχημα. Ειδικά τα άτομα, που επιλέγουν να στοιχηματίζουν μόνο σε έναν αγώνα, έχουν τη δυνατότητα να αυξήσουν σημαντικά τα κέρδη τους. Ο λόγος είναι ότι κάθε εταιρεία προσφέρει διαφορετική απόδοση για κάθε αγώνα, με αυτό τον τρόπο ο παίχτης μπορεί να επιλέγει τη στοιχηματική εταιρεία που προσφέρει το σημείο που τον ενδιαφέρει με την υψηλότερη απόδοση. Αυτό εφαρμόστηκε και στην παρούσα εργασία, καθώς σε κάθε αγώνα υπήρχαν τα στοιχεία από 5 διαφορετικές στοιχηματικές εταιρείες και ανάλογα τη πρόβλεψη, το ποντάρισμα τοποθετούνταν στην υψηλότερη διαθέσιμη απόδοση, το οποίο αύξησε σημαντικά τα κέρδη σε μονάδες σε κάθε υλοποίηση. Κάτι το οποίο αποτελεί μειονέκτημα σε αυτή τη μέθοδο είναι ότι ο παίχτης θα πρέπει να διαθέτει λογαριασμό σε περισσότερες από μία εταιρείες, κάτι το οποίο συνεπάγεται περισσότερος χρόνος που απαιτείται για τη δημιουργία του στοιχήματος, καθώς το άτομο πρέπει να ψάξει και να συγκρίνει αποδόσεις, λιγότερη απόλαυση του παιχνιδιού, γιατί το άτομο προσθέτει μια ακόμα ασχολία γύρω από το στοίχημα, και το κυριότερο, το άτομο πρέπει να μοιράσει το κεφάλαιο που διαθέτει σε όλες τις εταιρείες, ώστε να είναι ανά πάσα στιγμή έτοιμος να στοιχηματίσει σε οποιαδήποτε εταιρεία. Σε κάθε περίπτωση, οι πολλές στοιχηματικές εταιρείες δίνουν σημαντικό πλεονέκτημα στον παίχτη και ένα άτομο, το οποίο ποντάρει κυρίως σε μονά παιχνίδια και στοχεύει σε μακροπρόθεσμο κέρδος, πρέπει να το λάβει σοβαρά υπόψιν του.

6. Διαχείριση δύσκολων στοιχηματικών χρονιών

Οι περισσότερες προσεγγίσεις έκλεισαν με οριακή ζημία ή πολύ χαμηλό κέρδος την 1^{η} χρονιά του test και στη συνέχεια ξεκίνησαν την ανοδική τους πορεία. Η συγκεκριμένη αγωνιστική season ήταν η χρονιά 2015-16, η οποία έμεινε στην ιστορία για τις πολλές εκπλήξεις της, με μεγαλύτερη φυσικά την κατάκτηση του πρωταθλήματος από τη Λέστερ. Μάλιστα, στην επίσημη ιστοσελίδα του Αγγλικού Πρωταθλήματος, η συγκεκριμένη χρονιά περιγράφεται ως η πιο απρόβλεπτη και εκπληκτική χρονιά στην ιστορία του Αγγλικού ποδοσφαίρου. Για να γίνει πιο κατανοητό αυτό, αρκεί κανείς να προσέξει την απόδοση που έδιναν οι αγγλικές εταιρείες για την κατάκτηση του πρωταθλήματος από τη Λεστερ, η οποία ήταν 5000, δηλαδή οι εταιρείες έδιναν 0.02% πιθανότητα στο συγκεκριμένο ενδεχόμενο. Αυτό δικαιολογείται από το γεγονός ότι η Λέστερ την προηγούμενη χρονιά τερμάτισε λίγο πάνω από τις τελευταίες θέσεις. Επιπλέον, κατά τη διάρκεια του πρωταθλήματος υπήρξαν και απρόβλεπτα αποτελέσματα από τις δυνατές ομάδες. Για παράδειγμα, η Τσέλσι, η προηγούμενη κάτοχος του τροπαίου, στα μισά του πρωταθλήματος βρισκόταν λίγο πάνω από τη ζώνη του υποβιβασμού και τερμάτισε 10^{η}. Η Λίβερπουλ τερμάτισε 8^{η} και έμεινε εκτός ευρωπαϊκών διοργανώσεων, ενώ η Μάντσεστερ Σίτυ, έχοντας δαπανήσει υπέρογκα ποσά για μεταγραφές, τερμάτισε 4^{η}. Γενικά, οι περισσότερες παραδοσιακά μεγάλες ομάδες έκαναν μέτρια χρονιά, δυσκολεύοντας πολύ τα προγνωστικά, καθώς κανένα ματς δεν ήταν εύκολο σε πρόβλεψη, γιατί μπορεί να είχε αγώνα η 6^{η} ομάδα με τη 15^{η}, όμως η 15^{η} μπορεί να ήταν η περσινή πρωταθλήτρια και αυτό άλλαζε πολύ τα δεδομένα. Αυτό έκανε τη χρονιά φοβερά απρόβλεπτη, γεγονός που αποτυπώνεται και στη δυσκολία των μοντέλων να δημιουργήσουν κέρδος. Ωστόσο, ιδιαίτερη σημασία έχει ότι τα συστήματα διαχειρίστηκαν καλά τη χρονιά αυτή και δεν

δημιούργησαν μεγάλη ζημία, κάποια μάλιστα σημείωσαν και οριακό κέρδος. Το παραπάνω είναι πολύ σημαντικό για οποιοδήποτε σύστημα επιδιώκει το κέρδος, γιατί μια δύσκολη χρονιά είναι πολύ πιθανό να εμφανιστεί. Επομένως, στόχος του είναι να διαχειριστεί σωστά αυτές τις χρονιές με τις λιγότερες δυνατές απώλειες, όπως ακριβώς έκαναν και τα μοντέλα της εργασίας.

7. ***Η προσέγγιση 1 ξεχώρισε στις περισσότερες υλοποιήσεις***

Παρατηρήθηκε ότι στις περισσότερες υλοποιήσεις, ως επικρατέστερη ήταν αυτή του 1, δηλαδή η προσέγγιση που ακολουθεί τις προβλέψεις του συστήματος μόνο όταν προβλέπει νίκη γηπεδούχου. Η νίκη γηπεδούχου, όπως έχει ήδη αναφερθεί πολλές φορές, είναι το πιο συχνό αποτέλεσμα και λογικά το πιο ασφαλές για πρόβλεψη, όπως έδειξαν τα αποτελέσματα όλων των υλοποιήσεων. Όλοι οι αλγόριθμοι πέτυχαν πολύ υψηλό f1-score και ικανοποιητική μέση στοιχηματική απόδοση επιτυχημένων προβλέψεων, καθιστώντας τον "άσσο" ασφαλή επιλογή. Αυτό μπορεί να δικαιολογηθεί από το γεγονός ότι τα συστήματα είχαν περισσότερα δεδομένα της κλάσης νίκης γηπεδούχου και εκπαιδεύτηκαν καλύτερα στη συγκεκριμένη κλάση. Ακόμη, η νίκη γηπεδούχου επαληθεύεται περίπου στο 50% των περιπτώσεων, γεγονός που αυξάνει ακόμα περισσότερο την καλή επίδοση στους "άσσους".

6.2 Μελλοντικές προεκτάσεις

Γενικότερα, η Μηχανική Μάθηση δεν έχει αναπτυχθεί πολύ σε προβλήματα προβλέψεων αθλητικών γεγονότων σε συνδυασμό με τις στοιχηματικές αποδόσεις και για αυτό υπάρχει πρόσφορο έδαφος να πραγματοποιηθεί τα επόμενα χρόνια. Η παρούσα εργασία έκανε μια πρώτη απόπειρα να δημιουργήσει ένα σύστημα, το οποίο θα παράγει προβλέψεις για το τελικά αποτελέσματα αγώνων του Αγγλικού Πρωταθλήματος Ποδοσφαίρου, οι οποίες θα χρησιμοποιούνται προς στοιχηματισμό, στοχεύοντας στη δημιουργία κέρδους μακροπρόθεσμα. Σε αυτό το πλαίσιο μπορεί κανείς να ασχοληθεί επιπλέον με:

- άλλα ενδεχόμενα ενός αγώνα ποδοσφαίρου πέρα από το τελικό αποτέλεσμα, όπως το σύνολο τερμάτων, το πλήθος των κόρνερ και των καρτών. Σε κάθε αγώνα, προσφέρονται πολλές επιλογές προς στοιχηματισμό πέρα από το κλασικό 1-X-2, δίνοντας τη δυνατότητα σε κάποιον να ασχοληθεί και με εναλλακτικά στοιχήματα.

- άλλα πρωταθλήματα πέρα από το Αγγλικό, όπως για παράδειγμα το Ισπανικό και το Γερμανικό ή και να ασχοληθεί σε αυτές τις χώρες, με τα πρωταθλήματα μικρότερων κατηγοριών. Στο διαδίκτυο υπάρχουν πολλά δεδομένα για τα περισσότερα πρωταθλήματα του κόσμου, επομένως είναι εύκολο και έχει ενδιαφέρον μια προσέγγιση και σε άλλα πρωταθλήματα και φυσικά η σύγκριση των αποτελεσμάτων με αυτά της παρούσας εργασίας.

- άλλα αθλήματα, όπως το μπάσκετ, το baseball και το τέννις. Το ποδόσφαιρο είναι το κυρίαρχο άθλημα, όμως δεδομένα συλλέγονται και για τα υπόλοιπα αθλήματα. Ειδικά, για το baseball, υπάρχουν συγκλονιστικά πολλά δεδομένα και για αρκετές χρονιές. Να αναφερθεί, ότι έχουν γίνει κάποιες προσπάθειες σε αρκετά αθλήματα πέρα από το ποδόσφαιρο, όμως αυτές στόχευαν στην υψηλή ορθότητα και όχι στη δημιουργία κέρδους, επωφελούμενος τις στοιχηματικές αποδόσεις.

Βιβλιογραφία

[1] N. Vlastakis, G. Dotsis, and R. N. Markellos, “Nonlinear modelling of European football scores using support vector machines,” Applied Economics, vol. 40, no. 1, pp. 111–118, 2008.

[2] Δ. Ιωάννου, “Μοντέλα Τεχνητής Νοημοσύνης για την πρόβλεψη αποτελεσμάτων σε αγώνες ποδοσφαίρου”, ΑΠΘ, 2013.

[3] A.C. Constantinou, N.E. Fenton, M. Neil, “Profiting from an inefficient association football gambling market: Prediction, risk and uncertainty using Bayesian networks”, Knowledge-Based Systems, Vol 50 (2013), pp. 60-86.

[4] R. Baboota, H. Kaur, “Predictive analysis and modeling football results using machine learning approach for English Premier League”, International Journal of Forecasting, (2018), Forthcoming.

[5] A. S. Timmaraju, A. Palnitkar, & V. Khanna, “Game ON! Predicting English Premier League Match Outcomes”, 2013.

[6] B. Ulmer & M. Fernandez, “Predicting Soccer Match Results in the English Premier League”, 2014.

[7] J. Hucaljuk and A. Rakipovic, “Predicting football scores using machine learning techniques,” in MIPRO, 2011 Proceedings of the 34th International Convention. IEEE, pp. 1623–1627, 2011.

[8] J. Goddard and I. Asimakopoulos, “Forecasting football results and the efficiency of fixed-odds betting,” Journal of Forecasting, vol. 23, no. 1, pp. 51–66, 2004.

[9] K. Odachowski and J. Grekow, “Using bookmaker odds to predict the final result of football matches,” in Knowledge Engineering, Machine Learning and Lattice Computing with Applications. Springer, pp. 196–205, 2013.

[10] M. Hall, “Correlation-based feature subset selection for machine learning,” Ph.D. dissertation, University of Waikato, Hamilton, New Zealand, 1999.

[11] A. Joseph, N. Fenton, and M. Neil, “Predicting football results using bayesian nets and other machine learning techniques,” Knowledge-Based Systems, vol. 19, no. 7, pp. 544–553, 2006.

[12] R. P. Bunker & F. Thabtah, “A machine learning framework for sport result prediction”, Applied Computing and informatics, vol. 15, issue 1, pp. 27-33, 2019

[13] N. Tax, Y. Joustra. Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach. Transactions on Knowledge and Data Engineering, 2015. pages 19

[14] A. McCabe and J. Trevathan, “Artificial intelligence in sports prediction,” in Proceedings of the 5th International Conference on Information Technology: New Generations, 2008, pp. 1194 1197.

[15] B. Hamadani. Predicting the outcome of NFL games using machine learning. Stanford University, 2006. pages 18

[16] Owramipur, F., Eskandarian, P., & Mozneb, F. S. (2013). “Football result prediction with Bayesian network in Spanish league-Barcelona team”. International Journal of Computer Theory and Engineering, 5, 812–815.

[17] Medium, Θεωρία, https://www.mexdium.com/

[18] Dataset, https://www.fooxtball-data.co.uk/englandm.php

[19] Αξία ομάδων, https://www.transxfermarkt.com/