ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Πρόβλεψη Αποτελεσμάτων Αγώνων Αντισφαίρισης με Χρήση Τεχνικών Μηχανικής Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Ροϊδούλη Βασιλικής Ηλιάννας

Επιβλέπων : Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2023

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Πρόβλεψη Αποτελεσμάτων Αγώνων Αντισφαίρισης με Χρήση Τεχνικών Μηχανικής Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Ροϊδούλη Βασιλικής Ηλιάννας

Επιβλέπων : Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 13η Ιουλίου 2023.

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

Αθήνα, Ιούλιος 2023

.....

ΡΟΪΔΟΥΛΗ ΒΑΣΙΛΙΚΗ ΗΛΙΑΝΝΑ

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

Περίληψη

Η αντισφαίριση έχει γίνει ένα παγκοσμίως αναγνωρισμένο και εξαιρετικά δημοφιλές άθλημα, που χαρακτηρίζεται από την αξιοσημείωτη πρόοδο των επαγγελματικών τουρνουά όπως τα ATP Masters, ATP Cup και Grand Slam. Ταυτόχρονα, ο τομέας του Αθλητικής Ανάλυσης γνώρισε σημαντική ανάπτυξη τον 21ο αιώνα, φέρνοντας επανάσταση στην ανάλυση και την κατανόηση διαφόρων αθλημάτων. Το Sports Analytics περιλαμβάνει ένα ευρύ φάσμα εφαρμογών, αξιοποιώντας εκτεταμένα σύνολα δεδομένων για την πρόβλεψη των αποτελεσμάτων, τη βελτιστοποίηση της επιλογής παικτών και την αξιολόγηση των στρατηγικών παιχνιδιού. Οι ακριβείς προβλέψεις των αθλητικών αποτελεσμάτων είναι κρίσιμης σημασίας σε αυτό το πλαίσιο, ιδιαίτερα με την άνοδο των διαδικτυακών αθλητικών προγνωστικών. Αυτή η διπλωματική εργασία στοχεύει να χρησιμοποιήσει αλγόριθμους και τεχνικές μηχανικής μάθησης για την πρόβλεψη των αποτελεσμάτων των αγώνων αντισφαίρισης. Χρησιμοποιώντας Προηγμένες Τεχνικές Ανάλυσης Δεδομένων, η έρευνα στοχεύει στην ανάπτυξη ενός ισχυρού Μοντέλου Πρόβλεψης ικανού να προβλέπει με ακρίβεια τα αποτελέσματα του αγώνα. Η επιτυχής εφαρμογή ενός τέτοιου μοντέλου θα είχε εκτεταμένες επιπτώσεις, ιδιαίτερα στον κλάδο των αθλητικών προβλέψεων

Λέξεις Κλειδιά:

Μηχανική Μάθηση, Αντισφαίριση, Πρόβλεψη Αποτελέσματος Αγώνα, Αθλητική Ανάλυση

Abstract

Tennis has become a globally recognized and extremely popular sport, characterized by the remarkable progress of professional tournaments such as the ATP Masters, ATP Cup and Grand Slam. At the same time, the field of Sports Analytics has seen significant growth in the 21st century, revolutionizing the analysis and understanding of various sports. Sports Analytics encompasses a wide range of applications, leveraging extensive datasets to predict outcomes, optimize player selection and evaluate game strategies. Accurate predictions of sports results are critical in this context, especially with the rise of online sports predictions. This thesis aims to use machine learning algorithms and techniques to predict the results of tennis matches. Using Advanced Data Analysis Techniques, the research aims to develop a powerful Prediction Model capable of accurately predicting match results. The successful implementation of such a model would have far-reaching implications, particularly in the sports prediction industry.

Keywords:

Machine Learning, Tennis, Predicting Results of a Match, Sports Analytics

Ευχαριστίες

Η ολοκλήρωση των προπτυχιακών μου σπουδών και η συγγραφή της παρούσας διπλωματικής εργασίας σηματοδοτούν το κλείσιμο ενός σημαντικού κεφαλαίου στην ζωή μου. Σκεπτόμενη όλα τα φοιτητικά μου χρόνια, είμαι γεμάτη ευγνωμοσύνη προς τα άτομα που μου στάθηκαν και συνέλαβαν στο ταξίδι αυτό. Πρώτα και κύρια, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ.Δημήτριο Ασκούνη, για την ευκαιρία που μου έδωσε με την ανάθεση ενός τόσο ενδιαφέροντος θέματος για την εκπόνηση της διπλωματικής μου. Επίσης, θα ήθελα να ευχαριστήσω και τον διδακτορικό φοιτητή Κωνσταντίνο Αλεξάκη, καθώς η συνεργασία, η υποστήριξη και η καθοδήγησή του σε όλη την διάρκεια εκπόνησης της διπλωματικής μου εργασίας ήταν πολύτιμες. Ακόμη, οφείλω ευγνωμοσύνη στους αγαπημένους μου φίλους και συμφοιτητές, που υπήρξαν πυλώνες στο φοιτητικό μου ταξίδι. Μαζί, έχουμε μοιραστεί αμέτρητες στιγμές που θα έχουν για πάντα μια ξεχωριστή θέση στην καρδιά μου. Η παρουσία τους μου έφερε στιγμές ξέγνοιστης χαράς, γέλιου και αγάπης και η υποστήριξή τους μου έδωσε την αυτοπεποίθηση να κυνηγήσω τα όνειρα μου και να ξεπεράσω τις προκλήσεις που εμφανίζονται στο δρόμο μου. Επιπλέον, δεν θα ήμουν ο άνθρωπος που είμαι σήμερα, ούτε θα είχα πετύχει τα όνειρά μου, χωρίς την αγάπη και την φροντίδα της μητέρας μου, Μαρίας, του πατέρα μου, Δημήτρη και της αδερφής μου, Χριστίνας. Η συνέχης υποστήριξή τους, η καθοδήγησή και η ενθάρρυνσή τους ήταν πολύτιμες για εμένα. Ήταν εκεί για μένα σε κάθε βήμα της διαδρομής, προσφέροντας μου συμβουλές και αγάπη. Τους είμαι αιωνίως ευγνώμων και τους αφιερώνω αυτή την εργασία ως σύμβολο της αγάπης και της εκτίμησής μου.

Πίνακας περιεχομένων

1 *Εισαγωγή*

Εδώ και πολλά χρόνια, η αντισφαίριση έχει κερδίσει σημαντική παγκόσμια αναγνώριση και πλέον θεωρείται ένα εξαιρετικά δημοφιλές άθλημα παγκοσμίως. Η εξέλιξη των επαγγελματικών τουρνουά Αντισφαίρισης, όπως τα ATP Masters, ATP Cup και Grand Slams, συμπεριλαμβανομένων των US Open, French Open, Wimbledon και Australian Open δείχνει την αξιοσημείωτη πρόοδο και βελτίωση του αθλήματος με την πάροδο των χρόνων. Παράλληλα με αυτή την ανάπτυξη, ο 21ος αιώνας γνώρισε μια σημαντική άνοδο στον τομέα των Αθλητικών Αναλυτικών Στοιχείων, φέρνοντας επανάσταση στον τρόπο με τον οποίο αναλύονται και κατανοούνται διάφορα αθλήματα. Τα Αθλητικά Αναλυτικά στοιχεία περιλαμβάνουν ένα ευρύ φάσμα εφαρμογών, από την πρόβλεψη πασών στο ποδόσφαιρο και την βελτιστοποίηση της επιλογής παικτών στο μπέιζμπολ και το χόκεϊ έως την αξιολόγηση αγώνων ποδοσφαίρου. Αυτές οι εξελίξεις αξιοποιούν εκτεταμένα σύνολο δεδομένων για να καθοδηγήσουν τα καθεστώτα εκπαίδευσης και την λήψη στρατηγικών αποφάσεων.

Τα τελευταία χρόνια, η Αθλητική Βιομηχανία έχει γίνει μάρτυρας της ακμάζουσας εμφάνισης των αθλητικών προγνωστικών ως μια κερδοφόρα παγκόσμια επιχείρηση, που δημιουργεί έσοδα δισεκατομμυρίων δολαρίων. Οι αθλητικές προβλέψεις βασίζονται σε προγνώσεις που αποδίδουν κέρδη σε διαφορετικά αθλητικά αποτελέσματα. Η ακρίβεια αυτών των αποδόσεων γίνεται όλο και πιο κρίσιμη με την άνοδο των διαδικτυακών αθλητικών προγνωστικών παγκοσμίως [1]. Κατά συνέπεια, υπάρχει μια αυξανόμενη ζήτηση για αξιόπιστα μοντέλα πρόβλεψης που μπορούν να προβλέψουν με ακρίβεια τα αποτελέσματα αθλητικών γεγονότων, όπως οι αγώνες αντισφαίρισης.

Ο πρωταρχικός στόχος αυτής της διπλωματικής εργασίας είναι να χρησιμοποιήσει Αλγόριθμους και Τεχνικές Μηχανικής Μάθησης για την πρόβλεψη των αποτελεσμάτων των

αγώνων αντισφαίρισης. Αξιοποιώντας τη δύναμη των Προηγμένων Τεχνικών Ανάλυσης Δεδομένων, αυτή η έρευνα στοχεύει να αναπτύξει ένα ισχυρό Μοντέλο Πρόβλεψης ικανό να προβλέπει με ακρίβεια τα αποτελέσματα των αγώνων. Η επιτυχής εφαρμογή ενός τέτοιου μοντέλου θα μπορούσε να έχει εκτεταμένες επιπτώσεις, ιδιαίτερα για τη βιομηχανία αθλητικών προγνώσεων, όπου η ακρίβεια των αποδόσεων παίζει καθοριστικό ρόλο.

Η εφαρμογή των αλγορίθμων Μηχανικής Μάθησης στην πρόβλεψη των αποτελεσμάτων των αγώνων αντισφαίρισης έχει σημαντικές δυνατότητες για πολλούς ενδιαφερόμενους. Πρώτον, οι ακριβείς προβλέψεις αγώνων μπορούν να βελτιώσουν τη συνολική εμπειρία αθλητικών προγνωστικών τόσο για τους λάτρεις του αθλήματος όσο και για τους παίκτες, δίνοντάς τους τη δυνατότητα να λαμβάνουν πιο ενημερωμένες αποφάσεις. Δεύτερον, οι αθλητικοί αναλυτές, οι προπονητές και οι παίκτες μπορούν να αξιοποιήσουν αυτές τις προβλέψεις για να αποκτήσουν πολύτιμες πληροφορίες για τις στρατηγικές του αντιπάλου, την απόδοση των παικτών και τη συνολική δυναμική του αγώνα. Επιπλέον, τα αποτελέσματα της έρευνας μπορούν να συμβάλουν στο υπάρχον σύνολο γνώσεων στους τομείς της Μηχανικής Μάθησης, της Αθλητικής Ανάλυσης και της Προγνωστικής Μοντελοποίησης.

Η παρούσα διπλωματική εργασία είναι οργανωμένη σε επτά κεφάλαια, καθένα από τα οποία εξετάζει συγκεκριμένες πτυχές των στόχων και της μεθοδολογίας της έρευνας. Τα κεφάλαια αυτά οργανώνονται με τον παρακάτω τρόπο.

Στο Κεφάλαιο 2 γίνεται μια ολοκληρωμένη ανάλυση, εστιάζοντας σε διάφορες πτυχές που σχετίζονται με την έρευνα. Εξετάζεται το παιχνίδι της αντισφαίρισης, μαζί με τις περιπλοκές των προγνωστικών στη σφαίρα της αντισφαίρισης. Επιπλέον, τα μοντέλα Μηχανικής Μάθησης που χρησιμοποιούνται σε αυτή τη μελέτη, συμπεριλαμβανομένων των Λογιστικής Παλινδρόμησης (Logistic Regression), Τυχαίων Δασών (Random Forest), Μηχανών Διανυσμάτων Υποστήριξης (Support Vector Machines), Τεχνητών Νευρωνικών Δικτύων (Artificial Neural Networks) και ενός αλγόριθμου Ψηφοφορίας (Voting Algorithm), διερευνώνται διεξοδικά.

Στο Κεφάλαιο 3 πραγματοποιείται μια εκτενής ανασκόπηση της υπάρχουσας βιβλιογραφίας και ερευνητικών μελετών που σχετίζονται με το αντικείμενο της παρούσας διπλωματικής. Αξιοσημείωτες εργασίες και μεθοδολογίες στον τομέα της πρόβλεψης αγώνων αντισφαίρισης, των αλγορίθμων Μηχανικής Μάθησης και των εφαρμογών τους αναλύονται και συντίθενται κριτικά.

Το Κεφάλαιο 4 εμβαθύνει στο σύνολο δεδομένων που χρησιμοποιήθηκε σε αυτή την έρευνα, περιγράφοντας τα χαρακτηριστικά και τη διαδικασία απόκτησής του. Επιπλέον, περιγράφονται λεπτομερώς οι Τεχνικές Προεπεξεργασίας (preprocessing) που εφαρμόζονται για τη διασφάλιση της ποιότητας και της καταλληλότητας των δεδομένων για ανάλυση. Τα βήματα που έγιναν για τον Καθαρισμό και τον Μετασχηματισμό του συνόλου δεδομένων διευκρινίζονται, διευκολύνοντας έτσι την ακριβή και ουσιαστική ανάλυση.

Το Κεφάλαιο 5 εστιάζει στην Εξαγωγή Χαρακτηριστικών (Feature Engineering), ένα κρίσιμο στάδιο στη γραμμή Ανάλυσης Δεδομένων. Εδώ, συγκεκριμένα χαρακτηριστικά εξάγονται ή δημιουργούνται από το σύνολο δεδομένων για την ενίσχυση των προγνωστικών δυνατοτήτων των μοντέλων Μηχανικής Μάθησης που χρησιμοποιούνται. Το σκεπτικό πίσω από τις τεχνικές Επιλογής Χαρακτηριστικών και Μηχανικής επεξηγείται, παρέχοντας πληροφορίες για τον Μετασχηματισμό και την αύξηση των δεδομένων.

Στο Κεφάλαιο 6, διευκρινίζεται η διαδικασία Επιλογής Χαρακτηριστικών (Feature Selection), με την οποία τα πιο σχετικά και επιδραστικά χαρακτηριστικά προσδιορίζονται και διατηρούνται για μεταγενέστερη ανάλυση. Διάφορες μέθοδοι Επιλογής Χαρακτηριστικών, όπως η Ανάλυση Συσχέτισης και η αναδρομική Εξάλειψη Χαρακτηριστικών, χρησιμοποιούνται για να διασφαλιστεί η συμπερίληψη σημαντικών μεταβλητών, ενώ μετριάζονται οι επιπτώσεις του πλεονασμού ή του θορύβου.

Η πειραματική φάση περιγράφεται λεπτομερώς στο Κεφάλαιο 7, περιγράφοντας τις μεθοδολογίες και τις διαδικασίες που αναλήφθηκαν για την εκπαίδευση και την αξιολόγηση των επιλεγμένων μοντέλων Μηχανικής Μάθησης. Η απόδοση και η προγνωστική ακρίβεια κάθε μοντέλου αξιολογούνται χρησιμοποιώντας κατάλληλες μετρήσεις αξιολόγησης. Τα αποτελέσματα που προέκυψαν από τα πειράματα αναλύονται, συζητούνται και συγκρίνονται για να εξακριβωθεί η αποτελεσματικότητα της προτεινόμενης προσέγγισης.

Το Κεφάλαιο 8 χρησιμεύει ως επιστέγασμα της έρευνας, παρέχοντας μια περιεκτική περίληψη των κύριων ευρημάτων, γνώσεων και συνεισφορών που έγιναν από αυτή τη διπλωματική. Επιπρόσθετα, οι περιορισμοί που προέκυψαν κατά τη διάρκεια της μελέτης αναγνωρίζονται και συζητούνται. Επιπλέον, σκιαγραφούνται πιθανοί δρόμοι για μελλοντική έρευνα και βελτιώσεις στα μοντέλα πρόβλεψης, προσφέροντας ευκαιρίες για περαιτέρω εξερεύνηση και ανάπτυξη στον τομέα της πρόβλεψης αγώνων Αντισφαίρισης χρησιμοποιώντας αλγόριθμους Μηχανικής Μάθησης.

2 *Θεωρητικό Υπόβαθρο*

2.1 Το Παιχνίδι της Αντισφαίρισης

Η αντισφαίριση είναι ένα παγκοσμίως αναγνωρισμένο άθλημα που παρουσιάζει ευρεία απήχηση και προσελκύει αθλητές από διαφορετικά έθνη σε όλο τον κόσμο, με συμμετοχή σε περισσότερες από εκατό χώρες. Η επικράτηση του είναι ιδιαίτερα αξιοσημείωτη τόσο σε ελίτ επίπεδο, όσο και σε επαγγελματικό, καθώς παίζεται ενεργά σε σημαντικό αριθμό χωρών, ξεπερνώντας τις τριάντα στο ελίτ επίπεδο και σε πάνω από εξήντα έθνη στο επαγγελματικό. Οι αγώνες αντισφαίρισης περιλαμβάνουν τη χρήση ρακέτας, δίνοντας στο στρατηγικό παιχνίδι και την αθλητική ικανότητα. Μπορούν να παιχτούν είτε από δύο άτομα σε αγώνες μονού είτε από δύο ζευγάρια σε αγώνες διπλού. Ωστόσο, για να διατηρηθεί η σαφήνεια και η εστίαση, αυτή η μελέτη επικεντρώνεται στον τομέα της πρόβλεψης αγώνων αντισφαίρισης ατομικού ανδρών.

Σε έναν αγώνα αντισφαίρισης, οι παίκτες αναλαμβάνουν διαφορετικούς διακριτούς ρόλους: ο ένας αναλαμβάνει το ρόλο αυτού που σερβίρει, υπεύθυνου για την έναρξη κάθε πόντου, ενώ ο άλλος τον ρόλο αυτού που δέχεται, τοποθετημένος στην απέναντι πλευρά του ορθογωνίου γηπέδου, το οποίο χωρίζεται από ένα δίχτυ. Η επιλογή της επιφάνειας του γηπέδου, συμπεριλαμβανομένου του χώματος, του σκληρού ή του χόρτου, ποικίλλει ανάλογα με τα τουρνουά, με κάθε επιφάνεια να επιβάλλει μοναδικά χαρακτηριστικά και στρατηγικές παιχνιδιού. Ο αγώνας εξελίσσεται καθώς οι παίκτες συμμετέχουν σε μία συνεχή ανταλλαγή χτυπημάτων της μπάλας, με στόχο να κερδίσουν πόντους μέσω επιδέξιας τοποθέτησης σουτ και στρατηγικού παιχνιδιού. Ένα παιχνίδι περιλαμβάνει μία σειρά πόντων, κατά την οποία ο παίκτης που σερβίρει προσπαθεί να συγκεντρώσει τέσσερις πόντους με προβάδισμα

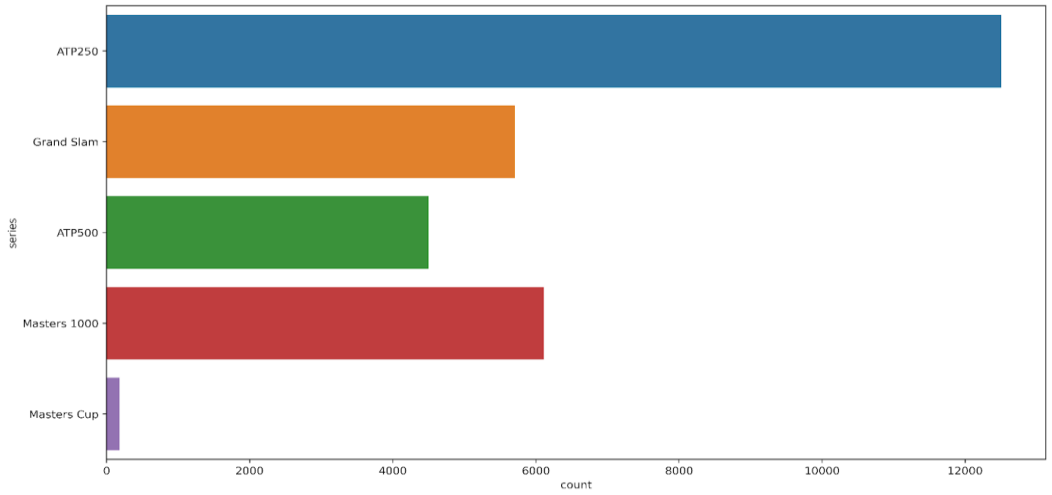
τουλάχιστον δύο πόντων για να εξασφαλίσει την νίκη του παιχνιδιού. Το σύστημα βαθμολόγησης που χρησιμοποιείται στην αντισφαίριση ακολουθεί μία αντισυμβατική ακολουθία 0, 15, 30 και 40, με το σκορ 40-40 να αναφέρεται ως ισοπαλία ή deuce. Στη συνέχεια, ο παίκτης που κατακτά τον επόμενο πόντο κερδίζει το πλεονέκτημα και η εξασφάλιση ενός ακόμη πόντου οδηγεί στη νίκη του παιχνιδιού. Οι παίκτες εναλλάσσουν τα καθήκοντά τους μετά από κάθε αγώνα, συμβάλλοντας στη δίκαιη κατανομή των ευκαιριών και ενισχύοντας τον αγωνιστικό χαρακτήρα του αθλήματος.

Στο πλαίσιο των σετ, ένας παίκτης πρέπει να κερδίσει τουλάχιστον έξι παιχνίδια, διατηρώντας παράλληλα προβάδισμα τουλάχιστον δύο αγώνων, για να εξασφαλίσει το σετ. Ωστόσο, εάν το παιχνίδι φτάσει σε ισοπαλία 6-6 στα παιχνίδια, παίζεται ένα παιχνίδι τάι μπρέικ (tie break) για να καθοριστεί ο νικητής. Αυτό το μοναδικό παιχνίδι απαιτεί από τον πρώτο παίκτη να συγκεντρώσει τουλάχιστον επτά πόντους με προβάδισμα δύο πόντων. Τελικά, ένας αγώνας καθορίζεται από τον παίκτη που βγαίνει νικητής κερδίζοντας την πλειοψηφία του προκαθορισμένου αριθμού σετ, που μπορεί να κυμαίνεται από τρία έως πέντε ανάλογα με τους συγκεκριμένους κανονισμούς του τουρνουά.

Η σφαίρα του αθλήματος της Αντισφαίρισης σε επαγγελματικό επίπεδο περιλαμβάνει μια ευρεία γκάμα τουρνουά, που εκτείνονται σε περίπου έντεκα μήνες του ημερολογιακού έτους. Αυτά τα τουρνουά διέπονται σχολαστικά από την Ένωση Επαγγελματιών Αντισφαίρισης (Association of Tennis Professionals, ATP) [2] διασφαλίζοντας την τήρηση των τυποποιημένων κανόνων και κανονισμών, όταν πρόκειται για τουρνουά ανδρών. Μεταξύ αυτών των τουρνουά [Σχ.2.1], υπάρχουν τέσσερα διακεκριμένα γεγονότα γνωστά ως τουρνουά Grand Slam, που περιλαμβάνουν το Wimbledon, το Roland Garros, το US Open και το Australian Open. Κάθε ένα από αυτά έχει τα δικά του μοναδικά χαρακτηριστικά, όπως τα γήπεδα με χόρτο του Wimbledon, τα χωμάτινα γήπεδα του Roland Garros, και τα σκληρά γήπεδα των US Open και Australian Open.

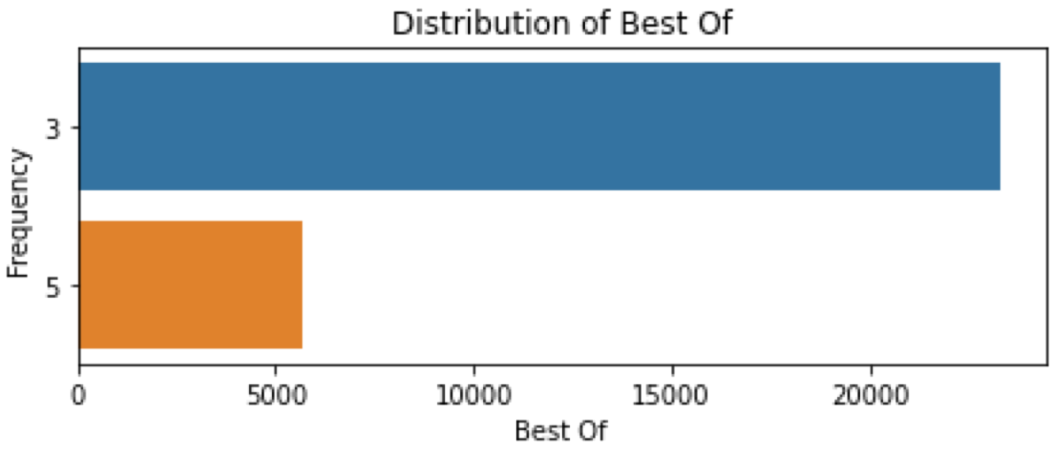
Είναι σημαντικό να σημειωθεί ότι τα συστήματα βαθμολόγησης που χρησιμοποιούνται σε αυτά τα τουρνουά παρουσιάζουν αξιοσημείωτες παραλλαγές. Ενώ τα περισσότερα τουρνουά χρησιμοποιούν τη μορφή best-of-3 σετ για αγώνες, οι αγώνες Grand Slam για άνδρες ακολουθούν μια πιο απαιτητική μορφή best-of-5 σετ [Σχ.2.2]. Επιπλέον, παρατηρούνται διακριτές προσεγγίσεις για την επίλυση ισόπαλων σετ, με το US Open, για παράδειγμα, να χρησιμοποιεί μηχανισμό τάι μπρέικ σετ όταν το σκορ φτάσει στο 6-6, ενώ τα άλλα τουρνουά Grand Slam υιοθετούν μια προσέγγιση πλεονεκτήματος σετ. Σε αυτό το πλαίσιο, ο νικητής

καθορίζεται από τον παίκτη που θα φτάσει πρώτος σε τουλάχιστον έξι σετ, ενώ διατηρεί προβάδισμα τουλάχιστον δύο σετ.



Σχήμα 2.1: Κατανομή Αγώνων ανά επίπεδο τουρνουά

Αυτές οι διαφοροποιήσεις στον προγραμματισμό των τουρνουά, στις επιφάνειες των γηπέδων και στα συστήματα βαθμολογίας συμβάλλουν στο ποικίλο και μαγευτικό τοπίο της αντισφαίρισης σε επαγγελματικό επίπεδο, προσθέτοντας πολυπλοκότητα και στρατηγικά στοιχεία που διαφοροποιούν κάθε αγώνα στην κατηγορία Grand Slam.



Σχήμα 2.2: Κατανομή Αγώνων ανά best - of

2.2 Προγνωστική στην Αντισφαίριση

Στη σφαίρα της πρόβλεψης των αγώνων αντισφαίρισης, προκύπτουν δύο βασικές κατηγορίες: οι προβλέψεις πριν από τον αγώνα και οι προβλέψεις εντός του αγώνα. Οι προβλέψεις πριν από τον αγώνα αφορούν αποκλειστικά προγνωστικά που τοποθετούνται πριν από την έναρξη ενός αγώνα, ενώ τα προγνωστικά εντός του αγώνα περιλαμβάνουν προβλέψεις που έγιναν κατά τη διάρκεια του αγώνα, ο οποίος βρίσκεται σε εξέλιξη. Μέσω αυτών των τομέων των προγνωστικών, μπορούν να τοποθετηθούν προβλέψεις σε ένα ευρύ φάσμα χαρακτηριστικών αγώνα, συμπεριλαμβανομένου του αποτελέσματος του αγώνα, των βαθμολογιών που ορίζονται και του συνολικού αριθμού παιχνιδιών που θα παιχτούν. Ωστόσο, για τους σκοπούς αυτής της μελέτης, η εστίαση θα περιοριστεί στις προβλέψεις πριν από τον αγώνα που στοχεύουν συγκεκριμένα τον νικητή του αγώνα. Αυτή η απόφαση βασίζεται στη μεγαλύτερη διαθεσιμότητα δεδομένων ιστορικών αποδόσεων που σχετίζονται με αυτόν τον συγκεκριμένο τύπο προβλέψεων.

Είναι επιτακτική ανάγκη να αναγνωρίσουμε ότι ο ανθρώπινος παράγοντας μπορεί να επηρεάσει σημαντικά τα αποτελέσματα των αθλητικών γεγονότων. Διάφορα στοιχεία, όπως λάθη που διαπράχθηκαν από διαιτητές ή παίκτες, περιστατικά τραυματισμών ή ατυχημάτων, διακυμάνσεις στην απόδοση των φαβορί και το αυξημένο κίνητρο των αουτσάιντερ που ανταγωνίζονται ισχυρότερους αντιπάλους, έχουν τη δυνατότητα να επηρεάσουν τα αποτελέσματα. Όταν συμμετέχουν σε έναν αγώνα, και οι δύο παίκτες εισέρχονται με κοινό στόχο να εξασφαλίσουν τη νίκη, καθώς ο ανταγωνισμός θα έπαυε να υπάρχει χωρίς αυτή τη στρατηγική. Επιπλέον, η τύχη μπορεί να ασκήσει σημαντική επιρροή στον καθορισμό του τελικού αποτελέσματος ενός αγώνα. Αυτοί οι πολύπλευροι παράγοντες υπογραμμίζουν τη δυναμική και απρόβλεπτη φύση των αθλημάτων, ενισχύοντας την ιδέα ότι πολλά στοιχεία πέρα από την απλή ικανότητα και στρατηγική μπορούν να διαμορφώσουν τα τελικά αποτελέσματα.

Ωστόσο, είναι σημαντικό να αναγνωρίσουμε ότι η Μηχανική Μάθηση, παρά τις δυνατότητές της ως πολύτιμο εργαλείο για την πρόβλεψη αθλητικών αποτελεσμάτων, δεν μπορεί να προσφέρει αλάνθαστη ακρίβεια ή να εξασφαλίσει εγγυημένα κέρδη. Ενώ η ικανότητα πρόβλεψης ανατροπών μπορεί να υπάρχει σε κάποιο βαθμό, η ίδια η ουσία των ανατροπών εξαρτάται από την έννοια της απρόβλεπτης κατάστασης. Εάν οι προβλέψεις ήταν σταθερά ακριβείς, θα έπαυαν να ταξινομούνται ως ανατροπές. Η Μηχανική Μάθηση, παράλληλα με άλλες στατιστικές μεθοδολογίες, χρησιμεύει ως πρακτικό μέσο για την αντιμετώπιση προβλημάτων του πραγματικού κόσμου. Ακόμη και τα πιο επιτυχημένα άτομα είναι βέβαιο

ότι θα αντιμετωπίσουν ήττες σε προσπάθειες πρόγνωσης του αποτελέσματος, όπως οι πιο κυρίαρχοι ανταγωνιστές μπορεί να αντιμετωπίσουν πισωγυρίσματα κατά τη διάρκεια διαφορετικών σταδίων ενός αγώνα. Ομολογουμένως, η εφαρμογή της Μηχανικής Μάθησης προσφέρει γνώσεις και βοηθά στη λήψη τεκμηριωμένων αποφάσεων, αλλά δεν μπορεί να εξαλείψει τις εγγενείς αβεβαιότητες και διακυμάνσεις που είναι εγγενείς στα αθλητικά αποτελέσματα και τις προγνωστικές επιχειρήσεις.

Το σύνολο δεδομένων που χρησιμοποιήθηκε σε αυτή τη μελέτη χρησιμοποιεί προγνωστικά χαρακτηριστικά που προέρχονται από εξέχουσες εταιρείες του Ηνωμένου Βασιλείου, που περιλαμβάνουν και τους δύο αγωνιζόμενους σε έναν συγκεκριμένο αγώνα αντισφαίρισης και αποκτήθηκαν μέσω του ιστότοπου tennis-data.co.uk [9]. Αυτές οι πρόσφατες αποδόσεις στοιχημάτων, που συνήθως συγκεντρώνονται από περισσότερες από 20 εταιρείες προβλέψεων, χρησιμεύουν για να αντιπροσωπεύουν τις πιο ενημερωμένες πληροφορίες πριν από τον αγώνα. Οι μέσες αποδόσεις αξιοποιούνται, καθώς προσφέρουν τις πιο ευνοϊκές συνθήκες για έναν παίκτη. Επίσης, η μελέτη ενσωματώνει τη διαφορά «spread» μεταξύ του μέσου όρου και των καλύτερων τιμών που διατίθενται στην αγορά τόσο για το φαβορί όσο και για το αουτσάιντερ. Είναι σημαντικό να αναγνωρίσουμε ότι οι εταιρείες προγνωστικών δεν παρέχουν «δίκαιες» πιθανότητες που μπορούν να μεταφραστούν άμεσα σε πιθανότητες γεγονότων.

Η απουσία δίκαιων αποδόσεων από τις προγνωστικές εταιρίες μπορεί να αποδοθεί σε δύο βασικούς λόγους. Πρώτον, οι εταιρείες ενσωματώνουν ένα περιθώριο, γνωστό ως «overround», το οποίο αντιπροσωπεύει το περιθώριο κέρδους τους. Για παράδειγμα, εάν η απόδοση είναι 1,25 για το φαβορί και 3,30 για το αουτσάιντερ, και υπάρχουν άτομα που ενδιαφέρονται να τοποθετήσουν τις προβλέψεις τους και στους δύο παίκτες, με ίδια ζήτηση και για τα δύο αποτελέσματα, τότε ο παίκτης μπορεί να προβλέψει κέρδος περίπου 10% με βάση τον υπολογισμό 1/1,25 + 1/3,30, που δίνει σύνολο 1,10. Χωρίς την εφαρμογή περιθωρίων, οι εταιρείες δεν θα μπορούσαν να παράγουν κέρδη. Συνήθως, οι εταιρείες παραμερίζουν ένα μέρος του μεριδίου τους από τις προγνώσεις που τοποθετούνται, πριν από τη διανομή των κερδών. Κατά συνέπεια, ακόμη και όταν οι πιθανότητες νίκης και ήττας είναι ίσες, οι εμφανιζόμενες πιθανότητες θα είναι ελαφρώς μικρότερες από 2,00, ανάλογα με το μέγεθος του περιθωρίου κέρδους. Η συνολική πιθανότητα όλων των πιθανών αποτελεσμάτων θα αθροίζεται πάντα σε περισσότερο από 100%, φτάνοντας ενδεχομένως στο 105,26%. Αυτό το πρόσθετο ποσοστό αντιπροσωπεύει το περιθώριο κέρδους του παίκτη, το οποίο είναι το κέρδος που μπορεί να περιμένει ο παίκτης όταν οι προβλέψεις είναι ομοιόμορφα ισορροπημένες.

Δεύτερον, οι εταιρείες προγνωστικών διαθέτουν την ικανότητα να προσαρμόζουν τις αποδόσεις που προσφέρουν προκειμένου να εξισορροπήσουν τη ζήτηση, να μεγιστοποιήσουν τον όγκο προβλέψεων και να αξιοποιήσουν τις προκαταλήψεις των παικτών. Αυτές οι προσαρμογές μπορεί να αποκλίνουν από τις «αληθινές» πιθανότητες. Όπως και άλλοι, οι εταιρείες δεν μπορούν να προβλέψουν με ακρίβεια το αποτέλεσμα κάθε αγώνα. Η επίτευξη μίας ισορροπημένης πρόβλεψης είναι ένα δύσκολο έργο, το οποίο σημαίνει ότι κατά καιρούς, οι παίκτες μπορεί να πληρώσουν περισσότερο από το συνολικό ποσό που τοποθέτησαν ή να κερδίσουν περισσότερο από το αναμενόμενο. Για να αντιμετωπίσουν αυτό το ζήτημα, οι εταιρείες προσαρμόζουν τις αποδόσεις τους με βάση το ποσό των χρημάτων που τοποθετούνται σε κάθε πιθανό αποτέλεσμα, συμβάλλοντας έτσι στη δημιουργία μίας πιο αξιόπιστης πρόγνωσης. Οι παίκτες που τοποθετούν τα προγνωστικά τους μπορεί να διαθέτουν εσωτερικές πληροφορίες ή προσωπική εμπειρία που μπορούν έμμεσα να αξιοποιηθούν για να βελτιώσουν τις πιθανότητες τους και να ενισχύσουν την ακρίβειά τους.

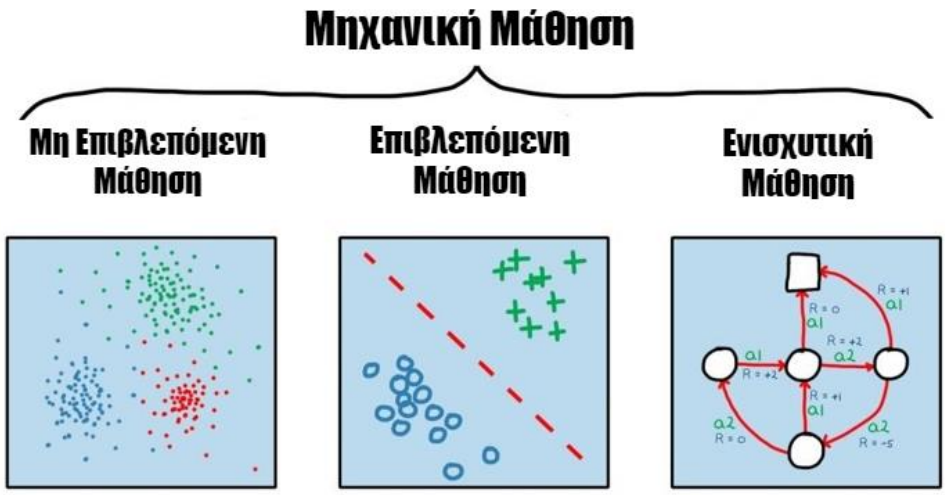
Ως εκ τούτου, αυτή η μελέτη ενσωματώνει τις αποδόσεις προγνωστικών ως είσοδο στα υπό εξέταση μοντέλα, αποφεύγοντας να κάνει ρητές προσαρμογές. Ο στόχος είναι να χρησιμοποιηθούν ευέλικτα μοντέλα ικανά να προσαρμόσουν τυχόν μη γραμμικές σχέσεις που μπορεί να υπάρχουν μεταξύ των αποδόσεων προγνωστικών και των πιθανοτήτων των αποτελεσμάτων του αγώνα. Επιτρέποντας στα μοντέλα να καταγράφουν περίπλοκες αλληλεπιδράσεις και πιθανές μη γραμμικότητες, η ανάλυση μπορεί να αξιοποιήσει αποτελεσματικά τις πληροφορίες που περιέχονται στις αποδόσεις προγνωστικών για να βελτιώσει την ακρίβεια των προβλέψεων αποτελεσμάτων. Η χρήση τέτοιων ευέλικτων μοντέλων διασφαλίζει ότι αντιμετωπίζονται επαρκώς οι εγγενείς πολυπλοκότητες της σχέσης μεταξύ των αποδόσεων και των αποτελεσμάτων των αγώνων, ενισχύοντας τη συνολική ευρωστία και αποτελεσματικότητα των μοντέλων πρόβλεψης που χρησιμοποιούνται σε αυτή τη μελέτη.

2.3 Μηχανική Μάθηση

Η Μηχανική Μάθηση είναι ένα υποσύνολο της Τεχνητής Νοημοσύνης και περιλαμβάνει την ανάπτυξη αλγορίθμων που διαθέτουν την ικανότητα να μαθαίνουν από δεδομένα. Χρησιμοποιεί στατιστικές τεχνικές και υπολογιστική ισχύ για την αυτόματη εξαγωγή μοτίβων και γνώσεων από τεράστιες ποσότητες πληροφοριών. Με την επαναληπτική βελτίωση των μοντέλων της μέσω διαδικασιών εκπαίδευσης, η Μηχανική Μάθηση επιτρέπει στους υπολογιστές να εντοπίζουν πολύπλοκες σχέσεις, να κάνουν ακριβείς προβλέψεις και να

προσαρμόζονται σε νέα δεδομένα. Αυτή η τεχνολογία έχει βρει εφαρμογές σε διάφορους τομείς, όπως η αναγώριση εικόνας και ομιλίας καθώς και η επεξεργασία φυσικής γλώσσας. Οι συνεχείς πρόοδοι και οι δυνατότητες για βελτίωση των διαδικασιών λήψης αποφάσεων καθιστούν την Μηχανική Μάθηση μία πολλά υποσχόμενη περιοχή στον ακαδημαϊκό χώρο.

2.3.1 Βασικά Είδη Μηχανικής Μάθησης



Σχήμα 2.3: Σχηματική Αναπαράσταση των κατηγοριών της Μηχανικής Μάθησης

2.3.1.1 Επιβλεπόμενη Μάθηση

Η Επιβλεπόμενη Μάθηση (Supervised Learning) [Σχ.2.3] είναι μια θεμελιώδης προσέγγιση στην Μηχανική Μάθηση που περιλαμβάνει την εκπαίδευση ενός μοντέλου σε δεδομένα με ετικέτα. Στοχεύει στην δημιουργία μιας αντιστοίχισης μεταξύ των χαρακτηριστικών εισόδου και των ανίστοιχων ετικετών εξόδου, μαθαίνοντας από ζεύγη παραδειγμάτων που παρέχονται κατά την φάση της εκπαίδευσης. Αυτός ο τύπος μάθησης βασίζεται στην διαθεσιμότητα ενός υψηλής ποιότητας συνόλου δεδομένων, όπου κάθε σημείο δεδομένων συσχετίζεται με μία γνωστή τιμή – στόχο. Οι επιβλεπόμενοι αλγόριθμοι μάθησης αναλύουν τα δεδομένα εκπαίδευσης και παράγουν ένα μοντέλο το οποίο μπορεί να χρησιμοποιηθεί για προβλέψεις νέων, αόρατων δεδομένων. Η ακρίβεια και η απόδοση των εποπτευόμενων μοντέλων μάθησης εξαρτώνται σε μεγάλο βαθμό από την ποιότητα και την αντιπροσωπευτικότητα των δεδομένων εκπαίδευσης, καθώς και από την επιλογή του αλγορίθμου και τις ρυθμίσεις παραμέτρων του.

2.3.1.2 Μη Επιβλεπόμενη Μάθηση

Η Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning) [Σχ.2.3] ασχολείται με την ανάλυση και την εξαγωγή σημαντικών προτύπων από δεδομένα χωρίς ετικέτα. Σε αντίθεση με την Επιβλεπόμενη Μάθηση, λειτουργεί χωρίς σαφείς ετικέτες στόχων ή προκαθορισμένα αποτελέσματα, καθιστώντας την κατάλληλη για διερευνητική ανάλυση και ανακάλυψη κρυψών δομών σε πολύπλοκα σύνολα δεδομένων. Οι αλγόριθμοι Μη Επιβλεπόμενης Μάθησης χρησιμοποιούν τεχνικές όπως η ομαδοποίηση, η μείωση διαστάσεων και η ανίχνευση ανωμαλιών για να αποκαλύψουν εγγενείς σχέσεις και ομαδοποιήσεις μέσα στα δεδομένα. Αξιοποιώντας στατιστικές μεθόδους και αλγορίθμους βελτιστοποίησης, αυτά τα μοντέλα αποκαλύπτουν πολύτιμες γνώσεις, εντοπίζουν ακραίες τιμές και διευκολύνουν την λήψη αποφάσεων βάσει δεδομένων. Η ευελιξία και η ικανότητα χειρισμού μη δομημένων και διαφορετικών δεδομένων καθιστούν την Μη Επιβλεπόμενη Μάθηση απαραίτητο εργαλείο σε τομείς όπως η εξόρυξη δεδομένων, η αναγνώριση προτύπων και η ανάλυση δεδομένων.

2.3.1.3 Ενισχυτική Μάθηση

Η Ενισχυτική Μάθηση (Reinforcement Learning) [Σχ.2.3] εστιάζει στην δυναμική αλληλεπίδραση μεταξύ ενός πράκτορα και του περιβάλλοντος του. Επικεντρώνεται γύρω από την έννοια της μάθησης μέσω δοκιμής και λάθους για την μεγιστοποίηση των μακροπρόθεσμων ανταμοιβών. Στην Ενισχυτική Μάθηση, ένας πράκτορας αναλαμβάνει ενέργειες σε ένα περιβάλλον, λαμβάνει ανατροφοδότηση με την μορφή επιβράβευσης ή τιμωρίας και ενημερώνει τις στρατηγικές του με βάση αυτά τα αποτελέσματα. Μέσα από μια διαδικασία εξερεύνησης, ο πράκτορας μαθαίνει τις βέλτιστες πολιτικές για την μεγιστοποίηση των ανταμοιβών με την πάροδο του χρόνου. Αυτό η κατηγορία μάθησης φαίνεται πολλά υποσχόμενη στην επίλυση πολύπλοκων διαδοχικών προβλημάτων λήψης αποφάσεων, όπως ο έλεγχος κίνησης των ρομποτ, η μάθηση επιτραπέζιων παιχνιδιών και η κατανομή πόρων. Οι αλγόριθμοι Ενισχυτικής Μάθησης, όπως το Q – Learning, προσφέρουν ένα ισχυρό πλαίσιο για αυτόνομη μάθηση, επιτρέποντας στους πράκτορες να προσαρμόσουν και να βελτιώσουν την συμπεριφορά τους μέσω της συνεχούς αλληλεπίδρασης με το περιβάλλον.

2.3.2 Αλγόριθμοι Μηχανικής Μάθησης

Στην ενότητα αυτή αναλύονται οι αλγόριθμοι που χρησιμοποιούνται στην παρούσα διπλωματική, οι οποίοι ανήκουν στην κατηγορία της Επιβλεπόμενης Μάθησης. Όπως

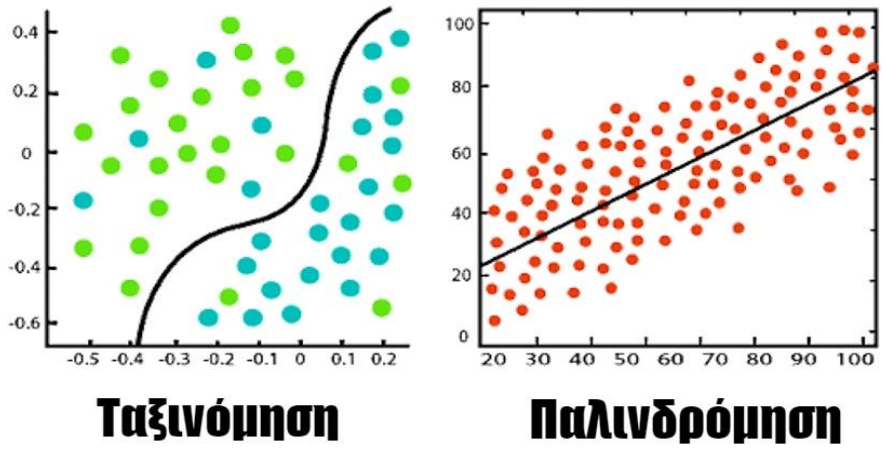
αναφέραμε και παραπάνω τα συστήματα Επιβλεπόμενης Μάθησης έχουν σχεδιαστεί για να αποκτούν γνώση εξάγοντας μια συνάρτηση από ένα παρεχόμενο σύνολο παραδειγμάτων εκπαίδευσης. Κάθε ένα από αυτά τα παραδείγματα περιλαμβάνει ένα διάνυσμα εισόδου, που αντιπροσωπεύει τα διάφορα χαρακτηριστικά του συνόλου δεδομένων, καθώς και μια αντίστοιχη τιμή εξόδου. Στο πλαίσιο της πρόβλεψης των αποτελεσμάτων των αγώνων αντισφαίρισης, τα ιστορικά δεδομένα αντισφαίρισης χρησιμεύουν ως βάση για τη δημιουργία παραδειγμάτων εκπαίδευσης. Αυτό συνεπάγεται την ενσωμάτωση των σχετικών χαρακτηριστικών του αγώνα και των παικτών στο διάνυσμα εισόδου, ενώ η τιμή εξόδου αντιπροσωπεύει το πραγματικό αποτέλεσμα του αγώνα.

Μια σημαντική πρόκληση για την κατασκευή αποτελεσματικών αλγορίθμων Μηχανικής Μάθησης για την πρόβλεψη αγώνων αντισφαίρισης έγκειται στη συνετή επιλογή των σχετικών χαρακτηριστικών. Ο εντοπισμός των πιο κατατοπιστικών και σημαντικών παραγόντων που συμβάλλουν στα αποτελέσματα των αγώνων είναι ζωτικής σημασίας για τη δημιουργία ακριβών μοντέλων πρόβλεψης. Η διαδικασία περιλαμβάνει προσεκτική εξέταση και ανάλυση των διαφόρων χαρακτηριστικών του αγώνα, των μετρήσεων απόδοσης των παικτών, των ιστορικών τάσεων και άλλων σχετικών παραγόντων που ενδέχεται να επηρεάσουν το τελικό αποτέλεσμα. Η σωστή επιλογή χαρακτηριστικών διασφαλίζει ότι τα μοντέλα πρόβλεψης καταγράφουν τους βασικούς παράγοντες που διέπουν το αποτέλεσμα των αγώνων αντισφαίρισης, ενισχύοντας έτσι την ακρίβεια και την αξιοπιστία των προβλέψεων.

Η επιλογή των σχετικών χαρακτηριστικών απαιτεί μια ολοκληρωμένη κατανόηση του παιχνιδιού της αντισφαίρισης, μαζί με τις στατιστικές τεχνικές που χρησιμοποιούνται στη Μηχανική Χαρακτηριστικών. Απαιτείται εξειδίκευση στον τομέα για να διακρίνουμε ποια χαρακτηριστικά είναι πιο πιθανό να έχουν σημαντικό αντίκτυπο στα αποτελέσματα των αγώνων και για να αποφευχθεί η συμπερίληψη περιττών ή άσχετων πληροφοριών. Η ποιότητα της επιλογής χαρακτηριστικών επηρεάζει άμεσα την απόδοση και την ερμηνευτικότητα των μοντέλων Μηχανικής Μάθησης που χρησιμοποιούνται στην πρόβλεψη των αγώνων αντισφαίρισης. Επομένως, μια σχολαστική και ενημερωμένη προσέγγιση για την επιλογή χαρακτηριστικών είναι ζωτικής σημασίας για τη δημιουργία ισχυρών και αποτελεσματικών μοντέλων πρόβλεψης στη σφαίρα της αντισφαίρισης.

Κατά την εφαρμογή αλγορίθμων Μηχανικής Μάθησης για την πρόβλεψης του αποτελέσματος ενός αγώνα αντισφαίρισης, υπάρχουν δύο κύριες προσεγγίσεις που μπορούν να υιοθετηθούν [Σχ.2.4]:

- Προσέγγιση Παλινδρόμησης (Regression): Σε αυτήν την προσέγγιση, η έξοδος είναι ένας πραγματικός αριθμός που αντιπροσωπεύει μια συνεχή τιμή. Στο πλαίσιο της πρόβλεψης αγώνων αντισφαίρισης, η έξοδος μπορεί να αντιπροσωπεύει άμεσα την πιθανότητα νίκης του αγώνα ή τις πιθανότητες των παικτών να κερδίσουν έναν πόντο στο σερβίς τους. Με την εκτίμηση αυτών των πιθανοτήτων, καθίσταται δυνατός ο προσδιορισμός της πιθανότητας νίκης ενός αγώνα.
- Προσέγγιση Δυαδικής Ταξινόμησης (Binary Classification): Στην περίπτωση αυτή, ο στόχος είναι να ταξινομηθούν οι αγώνες σε δύο κατηγορίες: «νίκη» ή «ήττα». Οι αλγόριθμοι ταξινόμησης μπορούν να χρησιμοποιηθούν για την ταξινόμηση των αγώνων με βάση τα χαρακτηριστικά εισόδου. Αυτοί οι αλγόριθμοι όχι μόνο παρέχουν τις προβλεπόμενες ετικέτες κλάσεων, αλλά αποδίδουν επίσης ένα μέτρο βεβαιότητας ή εμπιστοσύνης στην ταξινόμηση. Αυτή η τιμή εμπιστοσύνης μπορεί να ερμηνευτεί ως η πιθανότητα νίκης του αγώνα, χρησιμεύοντας ως χρήσιμος δείκτης για την πρόβλεψη του αποτελέσματος.



Σχήμα 2.4: Σχηματική Αναπαράσταση των προσεγγίσεων της Επιβλεπόμενης Μάθησης

Και οι δύο προσεγγίσεις παλινδρόμησης και δυαδικής ταξινόμησης έχουν τα πλεονεκτήματά τους, ανάλογα με τις συγκεκριμένες απαιτήσεις και προτιμήσεις του προβλήματος πρόβλεψης. Τα μοντέλα παλινδρόμησης παρέχουν συνεχείς εκτιμήσεις πιθανοτήτων, επιτρέποντας περισσότερες διαφοροποιημένες προβλέψεις. Από την άλλη πλευρά, τα μοντέλα δυαδικής ταξινόμησης προσφέρουν ένα πιο απλό αποτέλεσμα ταξινόμησης, με το πρόσθετο πλεονέκτημα της παροχής τιμών εμπιστοσύνης που μπορούν να ερμηνευθούν ως πιθανότητες. Η επιλογή μεταξύ αυτών των προσεγγίσεων εξαρτάται από τη φύση του προβλήματος, τη διαθεσιμότητα δεδομένων και το επιθυμητό επίπεδο ευκρίνειας στις προβλέψεις.

Οι ερευνητές και οι επαγγελματίες μπορούν να πειραματιστούν και με τις δύο προσεγγίσεις για να προσδιορίσουν ποια αποφέρει τα πιο ακριβή και αξιόπιστα αποτελέσματα στον τομέα της πρόβλεψης αγώνων αντισφαίρισης. Εμείς στην παρούσα φάση εστιάζουμε στην υιοθέτηση της προσέγγισης της Δυαδικής Ταξινόμησης.

2.3.2.1 Λογιστική Παλινδρόμηση

Ο αλγόριθμος Λογιστικής Παλινδρόμησης (Logistic Regression) [3] είναι ένας ευρέως χρησιμοποιούμενος αλγόριθμος για προβλήματα Ταξινόμησης. Όταν οι επιστήμονες αντιμετωπίζουν ένα νέο πρόβλημα ταξινόμησης, η Λογιστική Παλινδρόμηση είναι μεταξύ των πρώτων αλγορίθμων που εξετάζουν. Ανήκει στην κατηγορία των Γραμμικών Ταξινομητών και χρησιμοποιεί τη λογιστική συνάρτηση (logit function), γνωστή και ως σιγμοειδής συνάρτηση (sigmoid function), για να μοντελοποιήσει την σχέση μεταξύ των χαρακτηριστικών εισόδου και την πιθανότητα να ανήκει σε μία συγκεκριμένη κλάση. Ο αλγόριθμος έχει σχεδιαστεί ειδικά για να προβλέπει παρατηρήσεις σε διακριτές κλάσεις ή κατηγορίες, καθιστώντας τον κατάλληλο για προβλήματα Ταξινόμησης.

Το μοντέλο Λογιστικής Παλινδρόμησης αποτελείται από ένα σύνολο βαρών ή παραμέτρων, οι οποίοι συμβολίζονται ως «β» και αντιστοιχούν στα χαρακτηριστικά εισόδου «Χ» που χρησιμοποιούνται για την πρόβλεψη. Αυτά τα βάρη προσαρμόζονται μέσω της διαδικασίας ελαχιστοποίησης της λογιστικής απώλειας (logistic loss), η οποία είναι ένα μέτρο της απόκλισης μεταξύ των προβλεπόμενων πιθανοτήτων και των πραγματικών αποτελεσμάτων. Για να μετατραπεί ο γραμμικός συνδυασμός χαρακτηριστικών σε μία τιμή πιθανότητας που κυμαίνεται από 0 έως 1, χρησιμοποιείται η λογιστική συνάρτηση φ(z). Η εξίσωση υπολογισμού της πιθανολογικής πρόβλεψης περιλαμβάνει την άθροιση των γινομένων κάθε χαρακτηριστικού «Xi» και του αντίστοιχου βάρους του «βi», και στην συνέχεια την εφαρμογή της λογιστικής συνάρτησης φ(z) για να ληφθεί η τιμή της πιθανότητας.

$$P\ (y = 1|X;\ B) = \varphi(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)$$

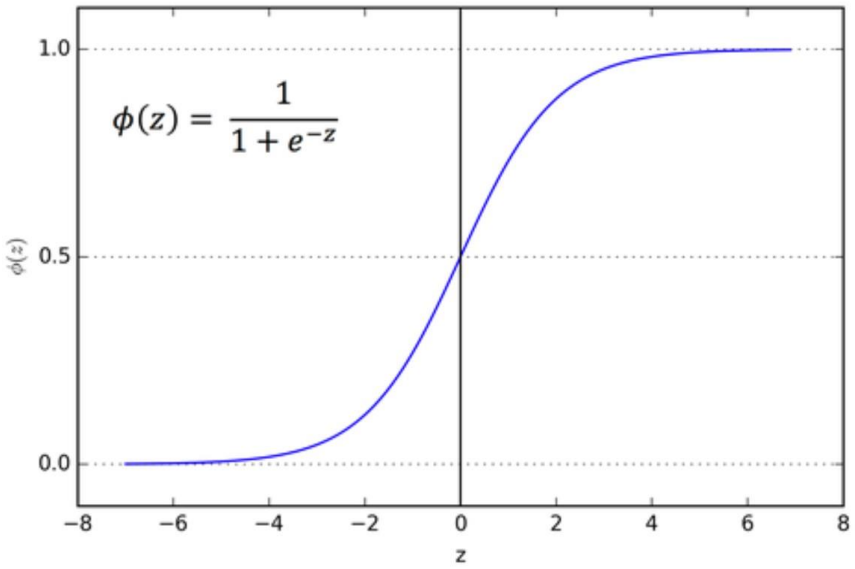
Όπου,

P (y = 1|x; β): η πιθανότητα θετικής κλάσης δεδομένων των χαρακτηριστικών εισόδου «Χ» και των βαρών «β».

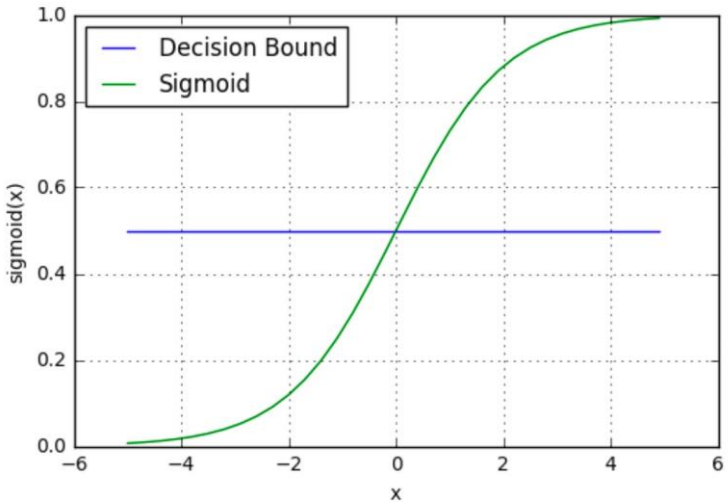
φ(z): η λογιστική (σιγμοειδής) συνάρτηση

Η σιγμοειδής συνάρτηση χαρακτηρίζεται από καμπύλη σχήματος «S», η οποία συνήθως αναφέρεται ως σιγμοειδής καμπύλη [Σχ.2.5]. Η σιγμοειδής συνάρτηση είναι μια μορφή της λογιστικής συνάρτησης και ορίζεται μαθηματικά ως εξής:

$$\varphi(z) = 1 / (1 + e^{-z})$$



Σχήμα 2.5: Σιγμοειδής Συνάρτηση



Σχήμα 2.6: Όριο Απόφασης

Η σιγμοειδής συνάρτηση μετατρέπει τις τιμές πιθανότητας, που κυμαίνονται από το 0 έως το 1, σε διακριτές ετικέτες κλάσεων. Για να προσδιοριστεί η ετικέτα κλάσης, επιλέγεται ένα

όριο απόφασης ή μια τιμή κατωφλίου. Πάνω από αυτή την τιμή κατωφλίου, οι τιμές πιθανότητας αντιστοιχίζονται στην κλάση 1, ενώ κάτω από το όριο, αντιστοιχίζονται στην κλάση 0. Συνήθως, το όριο απόφασης ορίζεται στο 0,5. Για παράδειγμα, εάν η τιμή πιθανότητας είναι 0,8 (που είναι μεγαλύτερη από 0,5), η παρατήρηση θα αντιστοιχιστεί στην κλάση 1. Αντίθετα, εάν η τιμή πιθανότητας είναι 0,2 (η οποία είναι μικρότερη από 0,5), η παρατήρηση θα αντιστοιχιστεί στην κλάση 0. [Σχ.2.6]

Η Λογιστική Παλινδρόμηση προσφέρει πολλά πλεονεκτήματα, καθιστώντας την δημοφιλή επιλογή σε διάφορους τομείς. Πρώτον, ξεχωρίζει για την απλότητα και την ερμηνευτικότητά της. Υποθέτοντας μια γραμμική σχέση μεταξύ των χαρακτηριστικών εισόδου και του λογάριθμου των πιθανοτήτων κλάσης στόχου, διασφαλίζει την διαφάνεια στην κατανόηση της συμπεριφοράς του μοντέλου. Επιπλέον, η λογιστική συνάρτηση μετατρέπει αυτές τις πιθανότητες σε ένα σημαντικό εύρος από 0 έως 1, διευκολύνοντας την ερμηνεία των προβλέψεων.

Δεύτερον, η Λογιστική Παλινδρόμηση παρουσιάζει σχετικά αποδοτικούς χρόνους εκπαίδευσης σε σύγκριση με πιο περίπλοκα μοντέλα. Ως γραμμικό μοντέλο, επιτρέπει την ταχεία εκτέλεση της διαδικασίας βελτιστοποίησης για την εκτίμηση των παραμέτρων του μοντέλου. Αυτό το χαρακτηριστικό είναι ιδιαίτερα πολύτιμο όταν πρόκειται για μεγάλης κλίμακας σύνολα δεδομένων ή εφαρμογές σε πραγματικό χρόνο όπου οι γρήγορες προβλέψεις είναι απαραίτητες.

Τέλος, η Λογιστική Παλινδρόμηση μειώνει τον κίνδενο της υπερπροσαρμογής (overfitting), σε αντίθεση με πιο περίπλοκα μοντέλα όπως τα Βαθιά Νευρωνικά Δίκτυα. Ο μειωμένος αρθμός υπερπαραμέτρων του βοηθάει την αποφυγή της εισαγωγής θορύβου στα δεδομένα, με αποτέλεσμα βελτιωμένη απόδοση γενίκευσης σε αόρατα δεδομένα.

Ο αλγόριθμος Λογιστικής Παλινδρόμησης έχει μερικές σημαντικές υπερπαραμέτρους που πρέπει να ρυθμιστούν για να βελτιστοποιήσουν την απόδοση του μοντέλου. Αυτές οι υπερπαράμετροι περιλαμβάνουν την ποινή (penalty), την μεταβλητή C, τον επιλυτή (solver) και τον μέγιστο αριθμό επαναλήψεων (max_iter). Ειδικότερα:

- penalty: Η υπερπαράμετρος ποινής καθορίζει τον τύπο της κανονικοποίησης που εφαρμόζεται στο μοντέλο. Η κανονικοποίηση βοηθά στην αποφυγή της υπερπροσαρμογής προσθέτοντας έναν όρο ποινής στη συνάρτηση απώλειας. Οι διαθέσιμες επιλογές για την ποινή είναι «L1», «L2» ή καμία. Η κανονικοποίηση

«L1» προάγει την αραιότητα ενθαρρύνοντας ορισμένα βάρη να είναι ακριβώς μηδενικά, ενώ η τακτοποίηση «L2» ενθαρρύνει μικρά βάρη, αλλά δεν τα αναγκάζει να μηδενιστούν.

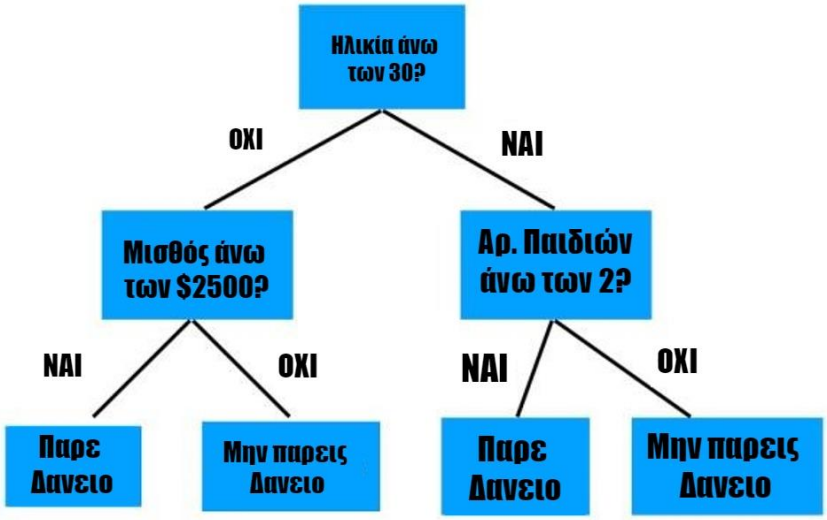
- C: Η υπερπαράμετρος C ελέγχει το αντίστροφο της ισχύος κανονικοποίησης. Καθορίζει το μέγεθος της κανονικοποίησης που εφαρμόζεται στο μοντέλο. Μικρότερες τιμές του C έχουν ως αποτέλεσμα ισχυρότερη κανονικοποίηση, ενώ μεγαλύτερες τιμές του C μειώνουν την ποσότητα της κανονικοποίησης. Με άλλα λόγια, μια μικρότερη τιμή C δίνει έμφαση στην ισχυρότερη κανονικοποίηση, οδηγώντας σε απλούστερα μοντέλα, ενώ μια μεγαλύτερη τιμή C επιτρέπει στο μοντέλο να ταιριάζει καλύτερα στα δεδομένα εκπαίδευσης.
- solver: Η υπερπαράμετρος επιλυτή καθορίζει τον αλγόριθμο που χρησιμοποιείται για τη βελτιστοποίηση της αντικειμενικής συνάρτησης Λογιστικής Παλινδρόμησης. Διαφορετικοί λύτες έχουν διαφορετικά υπολογιστικά χαρακτηριστικά, συμπεριφορά σύγκλισης και απαιτήσεις μνήμης. Η επιλογή του επιλυτή μπορεί να επηρεάσει τον χρόνο εκπαίδευσης και την ικανότητα του μοντέλου να χειρίζεται μεγάλα σύνολα δεδομένων. Ορισμένες κοινές επιλογές επίλυσης περιλαμβάνουν τα «liblinear», «newton – cg», «lbfgs», «sag» και «saga».
- max_iter: Η υπερπαράμετρος max iter ορίζει τον μέγιστο αριθμό επαναλήψεων που επιτρέπεται να συγκλίνει ο επιλυτής. Εάν ο επιλυτής αποτύχει να συγκλίνει εντός του καθορισμένου αριθμού επαναλήψεων, μπορεί να υποδεικνύει ότι το μοντέλο δεν είναι κατάλληλο για το δεδομένο σύνολο δεδομένων ή ότι ο ρυθμός εκμάθησης πρέπει να προσαρμοστεί. Η αύξηση του max_iter μπορεί να επιτρέψει στον επιλυτή να συγκλίνουν περισσότερες επαναλήψεις, αλλά μπορεί επίσης να αυξήσει τον χρόνο εκπαίδευσης.

Η κατάλληλη επιλογή των τιμών των υπερπαραμέτρων είναι κρίσιμη για την επίτευξη της βέλτιστης απόδοσης του μοντέλου. Αυτό συχνά απαιτεί μια διαδικασία συντονισμού υπερπαραμέτρων, όπως η χρήση Διασταυρούμενης Επικύρωσης (cross-validation) ή Αναζήτησης Πλέγματος (Grid Search), για τον εντοπισμό του βέλτιστου συνδυασμού υπερπαραμέτρων για ένα συγκεκριμένο σύνολο δεδομένων με σκοπό την βελτίωση της προγνωστική του απόδοσης και της ικανότητας γενίκευσης.

2.3.2.2 Τυχαία Δάση

Για να μπορέσουμε να δώσουμε μια περιγραφή του Τυχαίου Δάσους (Random Forest) [4], πρέπει πρώτα να συζητήσουμε τα Δέντρα Απόφασης (Decision Trees). Τα Δέντρα Απόφασης

είναι μία μέθοδος που μπορεί να χρησιμοποιηθεί τόσο για ταξινόμηση όσο και για παλινδρόμηση. Αποτελούν μια ιεραρχική δομή κόμβων, όπου κάθε κόμβος αντιπροσωπεύει ένα χαρακτηριστικό και έναν αντίστοιχο κανόνα απόφασης. Με την αναδρομική κατάτμηση των δεδομένων με βάση αυτούς τους κανόνες, τα δέντρα αποφάσεων δημιουργούν ένα μοντέλο που μοιάζει με διάγραμμα ροής που καθοδηγεί την διαδικασία πρόβλεψης [Σχ.2.7].



Σχήμα 2.7: Απλό Δέντρο Αποφάσεων

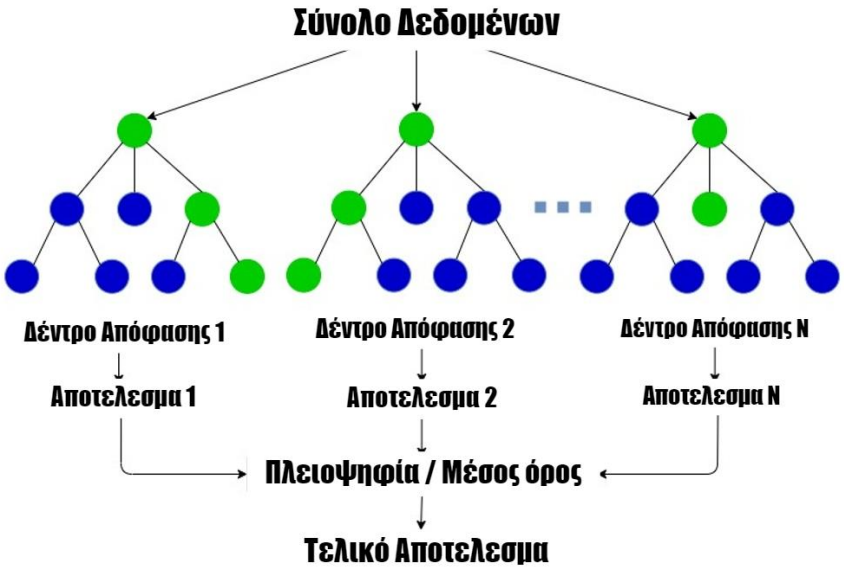
Σε κάθε κόμβο, το δέντρο επιλέγει το πιο διακριτικό χαρακτηριστικό για να διαχωρίσει τα δεδομένα, με στόχο την μεγιστοποίηση του κέρδους πληροφοριών ή την ελαχιστοποίηση της ακαθαρσίας. Τα τελικά φύλλα του δέντρου περιέχουν το προβλεπόμενο αποτέλεσμα. Ενώ τα δέντρα αποφάσεων προσφέρουν ερμηνευσιμότητα και ικανότητα χειρισμού τόσο κατηγορικών όσο και αριθμητικών δεδομένων, ενδέχεται να υποφέρουν από υπερβολική προσαρμογή και περιορισμένη απόδοση γενίκευσης. Για να ξεπεραστούν αυτοί οι περιορισμοί, ο αλγόριθμος Τυχαίων Δασών συνδυάζει πολλαπλά δέντρα απόφασης για να παράγει πιο ακριβείς και ισχυρές προβλέψεις.

Ο αλγόριθμος Τυχαίων Δασών είναι διάσημος για την ευελιξία και τη φιλική προς τον χρήστη φύση του και είναι ευρέως αναγνωρισμένος ως ένας από τους πιο προσιτούς αλγόριθμους. Η μεθοδολογία του περιλαμβάνει την κατασκευή πολλαπλών δέντρων αποφάσεων με βάση το παρεχόμενο σύνολο δεδομένων, την εξαγωγή προβλέψεων από κάθε μεμονωμένο δέντρο και, τελικά, την επιλογή της βέλτιστης λύσης μέσω ενός μηχανισμού ψηφοφορίας [Σχ.2.8].

Ο αλγόριθμος Τυχαίων Δασών προσφέρει αρκετά αξιοσημείωτα πλεονεκτήματα. Πρώτον, επιδεικνύει ευελιξία στην αποτελεσματική αντιμετώπιση προβλημάτων τόσο ταξινόμησης

όσο και παλινδρόμησης, καθιστώντας το εφαρμόσιμο σε διάφορους τομείς. Επιπλέον, ο αλγόριθμος είναι γνωστός για την υψηλή ακρίβεια και στιβαρότητά του. Αυτό προκύπτει από τη χρήση ενός σημαντικού αριθμού δέντρων αποφάσεων για τη δημιουργία προβλέψεων, μειώνοντας έτσι τις προκαταλήψεις και μετριάζοντας τον κίνδυνο υπερπροσαρμογής. Ένα άλλο πλεονέκτημα έγκειται στην ικανότητα του αλγορίθμου να χειρίζεται τιμές που λείπουν, μέσω δύο διακριτών προσεγγίσεων: τον υπολογισμό διαμέσου τιμών για συνεχείς μεταβλητές και τον υπολογισμό σταθμισμένων μέσων εγγύτητας για δεδομένα που λείπουν. Αυτή η ευελιξία συμβάλλει στην προσαρμοστικότητα του αλγορίθμου σε σενάρια πραγματικού κόσμου. Επιπλέον, ο αλγόριθμος Τυχαίων Δασών διευκολύνει την επιλογή χαρακτηριστικών, επιτρέποντας τον εντοπισμό των πιο σημαντικών χαρακτηριστικών από το σύνολο δεδομένων εκπαίδευσης. Αυτή η δυνατότητα επιλογής χαρακτηριστικών ενισχύει την ερμηνευτικότητα του μοντέλου και βοηθά στην εστιασμένη ανάλυση των πιο ενημερωτικών χαρακτηριστικών.

Ωστόσο, παρουσιάζει επίσης ορισμένους περιορισμούς. Το κυριότερο μεταξύ αυτών είναι η υπολογιστική του πολυπλοκότητα. Η χρήση μεγάλου αριθμού δέντρων αποφάσεων για την πραγματοποίηση προβλέψεων έχει ως αποτέλεσμα αυξημένες υπολογιστικές απαιτήσεις. Κάθε δέντρο στο δάσος πρέπει να παράγει προβλέψεις για την ίδια είσοδο, ακολουθούμενη από μια διαδικασία ψηφοφορίας, που οδηγεί σε χρονοβόρες λειτουργίες. Τα μοντέλα Τυχαίων Δασών είναι γενικά πιο δύσκολα στην ερμηνεία σε σύγκριση με μεμονωμένα δέντρα απόφασης. Ενώ τα δέντρα απόφασης προσφέρουν απλές προβλέψεις, η συλλογική φύση των προβλέψεων των Τυχαίων Δασών περιπλέκει την ερμηνευσιμότητα του μοντέλου.



Σχήμα 2.8: Αλγόριθμος Τυχαίων Δασεών

Στο πλαίσιο της Επιλογής Χαρακτηριστικών, ο αλγόριθμος Τυχαίων Δασών προσφέρει μια πολύτιμη προσέγγιση για την κατάταξη της σημασίας των μεταβλητών. Η διαδικασία περιλαμβάνει την προσαρμογή του αλγόριθμου του τυχαίου δάσους στο σύνολο δεδομένων, επιτρέποντας τη μέτρηση μεταβλητής σημασίας. Κατά τη διάρκεια αυτής της διαδικασίας προσαρμογής, ο αλγόριθμος καταγράφει και υπολογίζει τον μέσο όρο του σφάλματος out-of-bag για κάθε σημείο δεδομένων σε ολόκληρο το δάσος. Μετά την εκπαίδευση του τυχαίου δάσους, αξιολογείται η σημασία κάθε χαρακτηριστικού, που αντιπροσωπεύεται από το j-ο χαρακτηριστικό. Για να επιτευχθεί αυτό, οι τιμές του j-ου χαρακτηριστικού μετατίθενται στα δεδομένα εκπαίδευσης και το σφάλμα out-of-bag υπολογίζεται εκ νέου σε αυτό το διαταραγμένο σύνολο δεδομένων. Στη συνέχεια, η βαθμολογία σπουδαιότητας για το χαρακτηριστικό j υπολογίζεται με τον μέσο όρο της διαφοράς στο σφάλμα out-of-bag πριν και μετά τη μετάθεση σε όλα τα δέντρα στο δάσος. Για να εξασφαλιστεί η συνέπεια, αυτή η βαθμολογία στη συνέχεια κανονικοποιείται με την τυπική απόκλιση αυτών των διαφορών.

Τα χαρακτηριστικά που αποδίδουν υψηλότερες βαθμολογίες θεωρούνται πιο σημαντικά, ενώ εκείνα με χαμηλότερες βαθμολογίες θεωρούνται λιγότερο επιδραστικά. Με βάση αυτή την κατάταξη, επιλέγονται τα πιο σημαντικά χαρακτηριστικά για περαιτέρω ανάλυση και κατασκευή μοντέλων, ενώ τα λιγότερο σημαντικά μπορούν να αγνοηθούν. Χρησιμοποιώντας αυτή τη μεθοδολογία επιλογής χαρακτηριστικών στον αλγόριθμο Τυχαίων Δασών, μπορεί να εντοπιστεί και να χρησιμοποιηθεί ένα εκλεπτυσμένο σύνολο μεταβλητών, ενισχύοντας ενδεχομένως την απόδοση και την ερμηνευτικότητα του προκύπτοντος μοντέλου.

Ο αλγόριθμος Τυχαίων Δασών προσφέρει μια σειρά από υπερπαραμέτρους που μπορούν να ρυθμιστούν προσεκτικά για την επίτευξη βέλτιστης απόδοσης. Μερικές από αυτές τις υπερπαράμετρους περιλαμβάνουν:

- n_estimators: ο αριθμός των δέντρων που θα συμπεριληφθούν στο δάσος. Γενικά, ένας μεγαλύτερος αριθμός δέντρων οδηγεί σε βελτιωμένη απόδοση. Ωστόσο, αυξάνει επίσης τον υπολογιστικό χρόνο. Θα πρέπει να δοθεί ιδιαίτερη προσοχή στην επιλογή μιας κατάλληλης τιμής για αυτήν την υπερπαράμετρο ώστε να εξισορροπηθεί μεταξύ της απόδοσης και της υπολογιστικής απόδοσης.
- max_depth: το μέγιστο βάθος κάθε δέντρου απόφασης μέσα στο δάσος. Τα βαθύτερα δέντρα έχουν τη δυνατότητα να καταγράφουν πιο περίπλοκες σχέσεις στα δεδομένα, αλλά μπορεί επίσης να είναι επιρρεπή σε υπερβολική προσαρμογή. Είναι σημαντικό να γίνει επιλογή μιας βέλτιστης τιμής για το max_depth για την εξασφάλιση μιας ισορροπίας μεταξύ της πολυπλοκότητας του μοντέλου και της γενίκευσης.

- min_samples split: ο ελάχιστος αριθμός δειγμάτων που απαιτούνται για τον διαχωρισμό ενός εσωτερικού κόμβου κατά την κατασκευή ενός δέντρου αποφάσεων. Η αύξηση αυτής της τιμής μπορεί να βοηθήσει στην αποφυγή της υπερπροσαρμογής επιβάλλοντας έναν ορισμένο όγκο δεδομένων πριν δημιουργηθεί ένας διαχωρισμός. Ο σωστός συντονισμός του min_samples_split συμβάλλει στην εύρεση της κατάλληλης αντιστάθμισης μεταξύ της καταγραφής λεπτών μοτίβων και της αποτροπής της υπερβολικής προσαρμογής.
- min_samples_leaf: ο ελάχιστος αριθμός δειγμάτων που απαιτούνται για να σχηματιστεί ένας κόμβος φύλλου σε ένα δέντρο αποφάσεων. Η αύξηση αυτής της τιμής, όπως το min_samples_split, μπορεί να βοηθήσει στον μετριασμό της υπερπροσαρμογής. Ορίζοντας μια κατάλληλη τιμή για το min_samples_leaf, ο αλγόριθμος διασφαλίζει ότι κάθε κόμβος φύλλου περιέχει επαρκή αριθμό δειγμάτων για καλή γενίκευση.
- bootstrap: ελέγχει εάν χρησιμοποιείται τυχαία δειγματοληψία με αντικατάσταση κατά την κατασκευή μεμονωμένων δέντρων. Εισάγοντας την τυχαιότητα μέσω της δειγματοληψίας bootstrap, ο αλγόριθμος Τυχαίων Δασών μπορεί να βελτιώσει την απόδοση. Ωστόσο, είναι σημαντικό να σημειωθεί ότι αυτή η τυχαιότητα αυξάνει επίσης τη διακύμανση του μοντέλου. Θα πρέπει να δοθεί ιδιαίτερη προσοχή στην επιλογή της κατάλληλης ρύθμισης για την παράμετρο bootstrap, λαμβάνοντας υπόψη τα ειδικά χαρακτηριστικά του συνόλου δεδομένων και την επιθυμητή αντιστάθμιση μεταξύ απόδοσης και σταθερότητας.

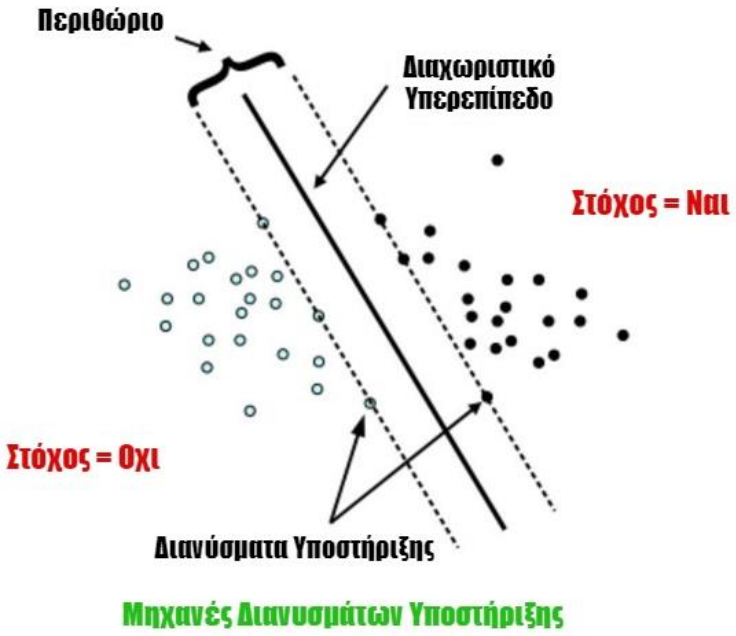
Η βελτιστοποίηση αυτών των υπερπαραμέτρων είναι κρίσιμη για την επίτευξη της καλύτερης δυνατής απόδοσης με τον αλγόριθμο Τυχαίων Δασών. Αυτή η διαδικασία συχνά περιλαμβάνει τεχνικές όπως διασταυρούμενη επικύρωση ή αναζήτηση πλέγματος για τη συστηματική διερεύνηση διαφορετικών συνδυασμών τιμών υπερπαραμέτρων και επιλογή της διαμόρφωσης που αποφέρει τη βέλτιστη αντιστάθμιση μεταξύ μεροληψίας και διακύμανσης.

2.3.2.3 Μηχανές Διανυσμάτων Υποστήριξης

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVMs) [5] είναι μια κατηγορία ισχυρών αλγορίθμων Μηχανικής Μάθησης που χρησιμοποιούνται για προβλήματα ταξινόμησης, παλινδρόμησης και ανίχνευσης ακραίων στοιχείων. Οι μηχανές αυτές υπερέχουν ως δυαδικοί Γραμμικοί Ταξινομητές, εκχωρώντας νέα σημεία δεδομένων σε προκαθορισμένες κατηγορίες. Δεν περιορίζονται σε προβλήματα Γραμμικής Ταξινόμησης, αλλά διαθέτουν την ικανότητα να εκτελούν αποτελεσματικά Μη Γραμμική Ταξινόμηση

αξιοποιώντας το «Κόλπο του Πυρήνα» (kernel trick). Το κόλπο του πυρήνα επιτρέπει στα SVMs να αντιστοιχίσουν έμμεσα τα δεδομένα εισόδου σε χώρους χαρακτηριστικών υψηλών διαστάσεων, όπου ο γραμμικός διαχωρισμός γίνεται εφικτός. Εφαρμόζοντας μια κατάλληλη συνάρτηση πυρήνα, μετατρέπουν τα αρχικά δεδομένα σε έναν χώρο υψηλότερων διαστάσεων, όπου μπορούν να αποτυπωθούν αποτελεσματικά πολύπλοκες σχέσεις και μοτίβα. Αυτό τις επιτρέπει να χειρίζονται δεδομένα που δεν διαχωρίζονται γραμμικά στον αρχικό χώρο χαρακτηριστικών, παρέχοντας μια πιο ευέλικτη και ισχυρή ικανότητα ταξινόμησης.

Ένα «Υπερεπίπεδο» (Hyperplane) στο πλαίσιο των Μηχανών Διανυσμάτων Υποστήριξης αναφέρεται σε ένα «Όριο Απόφασης» (Threshold) που διαχωρίζει ένα δεδομένο σύνολο σημείων δεδομένων με διαφορετικές ετικέτες κλάσης. Οι ταξινομητές στοχεύουν να βρουν ένα υπερεπίπεδο με το «μέγιστο Περιθώριο» (maximum Margin), το οποίο είναι ο μεγαλύτερος διαχωρισμός μεταξύ του υπερεπίπεδου και των πλησιέστερων σημείων δεδομένων [Σχ.2.9]. Αυτό το υπερεπίπεδο, γνωστό ως «Υπερεπίπεδο Μέγιστου Περιθωρίου», ορίζει έναν Γραμμικό Ταξινομητή που ονομάζεται «Ταξινομητής Μέγιστου Περιθωρίου» [Σχ.2.10]. Τα «Διανύσματα Υποστήριξης» (Support Vectors) είναι τα δείγματα σημείων δεδομένων που βρίσκονται πιο κοντά στο υπερεπίπεδο. Αυτά τα σημεία δεδομένων παίζουν κρίσιμο ρόλο στον καθορισμό της Διαχωριστικής Γραμμής ή του Υπερεπίπεδου υπολογίζοντας τα περιθώρια.

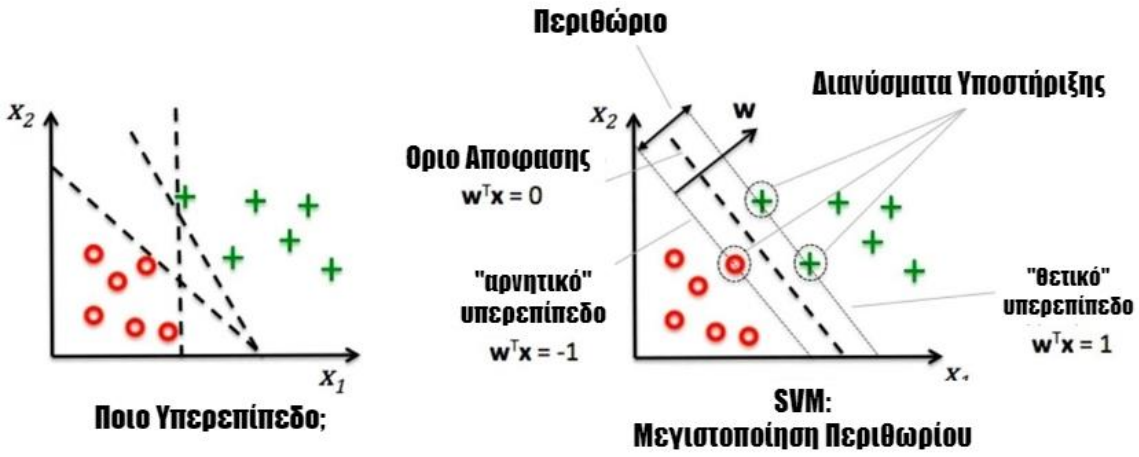


Σχήμα 2.9: Μηχανή Διανυσμάτων Υποστήριξης

Τα περιθώρια αντιπροσωπεύουν το χάσμα διαχωρισμού μεταξύ του υπερεπίπεδου και των πλησιέστερων σημείων δεδομένων. Υπολογίζονται ως η κάθετη απόσταση από το υπερεπίπεδο προς τα διανύσματα υποστήριξης ή τα πλησιέστερα σημεία δεδομένων. Η μεγιστοποίηση του χάσματος ή του περιθωρίου διαχωρισμού είναι ένας βασικός στόχος των Μηχανών Διανυσμάτων Υποστήριξης, καθώς επιτρέπει μεγαλύτερο περιθώριο εμπιστοσύνης στην απόφαση ταξινόμησης.

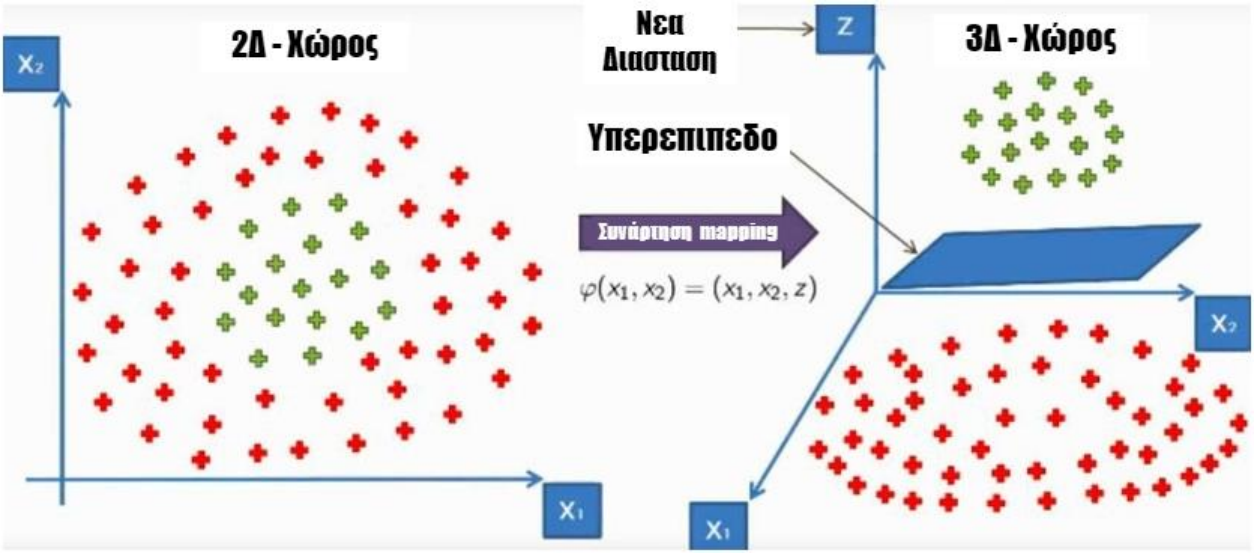
Το υπερεπίπεδο διαχωρίζει τα σημεία δεδομένων σε δύο κατηγορίες και τα διανύσματα υποστήριξης είναι τα σημεία δεδομένων που βρίσκονται στο ή πιο κοντά στο περιθώριο. Μεγιστοποιώντας το περιθώριο, οι Μηχανές Διανυσμάτων Υποστήριξης στοχεύουν να βρουν το βέλτιστο υπερεπίπεδο που παρέχει τον καλύτερο διαχωρισμό μεταξύ των κλάσεων.

Στο Σχήμα 2.10, τα σημεία δεδομένων δύο διαφορετικών κατηγοριών διαχωρίζονται από το Υπερεπίπεδο Μέγιστου Περιθωρίου. Τα Διανύσματα Υποστήριξης, που αντιπροσωπεύονται από τα επισημασμένα σημεία δεδομένων, παίζουν σημαντικό ρόλο στον προσδιορισμό της θέσης και του προσανατολισμού του υπερεπίπεδου. Μεγιστοποιώντας το περιθώριο, οι Μηχανές Διανυσμάτων Υποστήριξης προσπαθούν να βρουν ένα υπερεπίπεδο, που όχι μόνο διαχωρίζει αποτελεσματικά τις κλάσεις, αλλά παρέχει επίσης ένα ισχυρό και σίγουρο Όριο Απόφασης.



Σχήμα 2.10: Οπτική Αναπαράστση της έννοιας του Μέγιστου Περιθωρίου και του Υπερεπιπέδου Μέγιστου Περιθωρίου

Σε σενάρια όπου τα δείγματα σημείων δεδομένων είναι ευρέως διασκορπισμένα και δεν μπορούν να διαχωριστούν από ένα Γραμμικό Υπερεπίπεδο, οι Μηχανές Διανυσμάτων Υποστήριξης χρησιμοποιούν μια τεχνική που ονομάζεται όπως αναφέρθηκε και παραπάνω «Κόλπο του Πυρήνα». Αυτή η προσέγγιση περιλαμβάνει τον μετασχηματισμό του χώρου εισόδου σε χώρο υψηλότερης διάστασης, επιτρέποντας τον καλύτερο διαχωρισμό των σημείων δεδομένων [Σχ.2.11].



Σχήμα 2.11: Κόλπο του Πυρήνα – Μετασχηματισμός χώρου εισόδου σε χώρο υψηλότερης διάστασης για καλύτερο διαχωρισμό των σημείων δεδομένων

Στο Σχήμα 2.11, το αρχικό σύνολο δεδομένων απεικονίζεται σε ένα δισδιάστατο χώρο όπου δεν είναι δυνατός ο γραμμικός διαχωρισμός. Ωστόσο, εφαρμόζοντας το Κόλπο του Πυρήνα, χρησιμοποιείται μια συνάρτηση αντιστοίχισης (mapping function) για να μετατραπεί ο χώρος εισόδου σε έναν χώρο υψηλότερης διάστασης, στο σχήμα σε έναν τρισδιάστατο χώρο. Αυτός ο μετασχηματισμός επιτρέπει στα σημεία δεδομένων να διαχωριστούν αποτελεσματικά χρησιμοποιώντας ένα γραμμικό όριο.

Ουσιαστικά, ένας Πυρήνας μπορεί να οριστεί ως μια συνάρτηση που μετατρέπει τα δεδομένα εισόδου από χώρο χαμηλής διάστασης σε χώρο υψηλότερης διάστασης. Αυτός ο μετασχηματισμός επιτρέπει τη μετατροπή Μη Γραμμικά Διαχωρίσιμων Σημείων σε Γραμμικά Διαχωρίσιμα, εισάγοντας πρόσθετες διαστάσεις. Το Κόλπο του Πυρήνα παίζει σημαντικό ρόλο στην ενίσχυση της ακρίβειας του Ταξινομητή, ιδιαίτερα όταν αντιμετωπίζουμε προκλήσεις Μη Γραμμικού Διαχωρισμού. Μαθηματικά, μια συνάρτηση πυρήνα μπορεί να αναπαρασταθεί ως εξής:

$$K(x, y) = \Phi(x) \cdot \Phi(y)$$

Όπου,

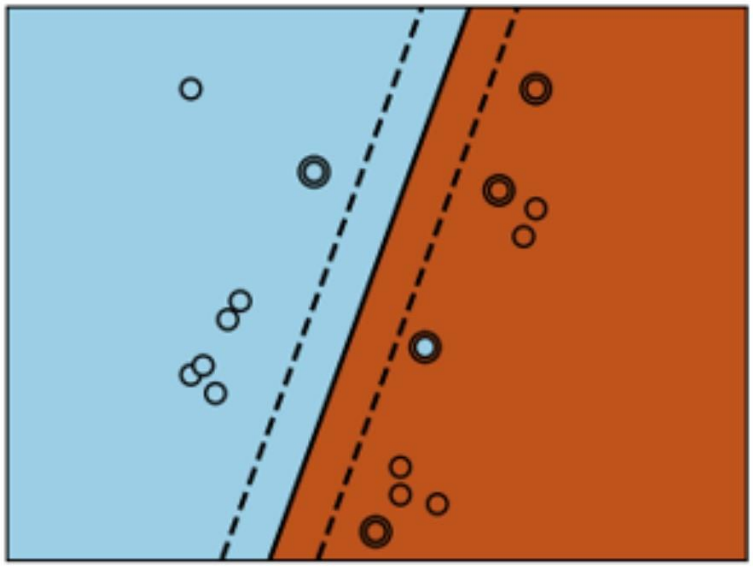
K : η συνάρτηση του πυρήνα,

x, y: τα σημεία δεδομένων εισόδου

Φ: η συνάρτηση μετασχηματισμού που αντιστοιχίζει τα σημεία δεδομένων στον χώρο υψηλότερης διάστασης.

Υπάρχουν αρκετοί δημοφιλείς πυρήνες, όπως ο Γραμμικός Πυρήνας (Linear Kernel), ο Πολυωνυμικός Πυρήνας (Polynomial Kernel), ο Πυρήνας της Ακτινικής Συνάρτησης Βάσης (Radial Basis Function Kernel) και ο Σιγμοειδής Πυρήνας (Sigmoid Kernel). Κάθε πυρήνας εξυπηρετεί έναν συγκεκριμένο σκοπό και παρουσιάζει ξεχωριστά χαρακτηριστικά.

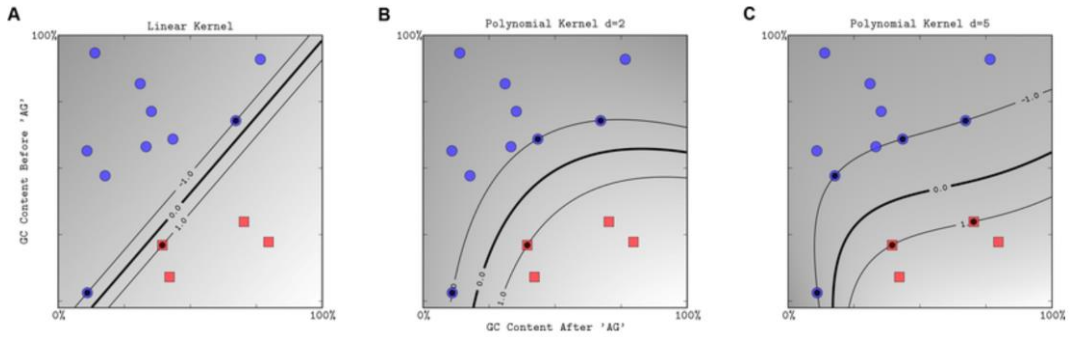
Ο Γραμμικός Πυρήνας [Σχ.2.12] αντιπροσωπεύει μια γραμμική συνάρτηση και χρησιμοποιείται όταν τα δεδομένα μπορούν να διαχωριστούν από μία μόνο γραμμή. Χρησιμοποιείται συχνά, ιδιαίτερα σε σενάρια που περιλαμβάνουν μεγάλο αριθμό χαρακτηριστικών ταξινόμησης κειμένου. Η εκπαίδευση με έναν γραμμικό πυρήνα τείνει να είναι ταχύτερη καθώς απαιτεί μόνο βελτιστοποίηση της παραμέτρου κανονικοποίησης C.



Σχήμα 2.12: Γραμμικός Πυρήνας

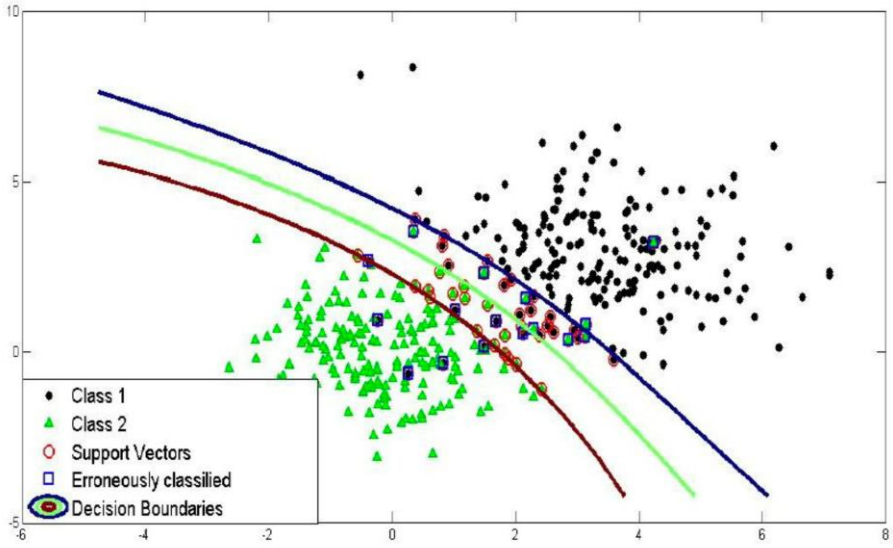
Ο Πολυωνυμικός Πυρήνας [Σχ.2.13] καταγράφει την ομοιότητα μεταξύ των διανυσμάτων σε έναν χώρο χαρακτηριστικών χρησιμοποιώντας πολυωνυμικούς συνδυασμούς των αρχικών μεταβλητών. Αυτός ο πυρήνας είναι δημοφιλής σε προβλήματα Επεξεργασίας Φυσικής

Γλώσσας, με τον τετραγωνικό βαθμό (degree = 2) να είναι η πιο κοινή επιλογή για την αποφυγή της υπερπροσαρμογής.



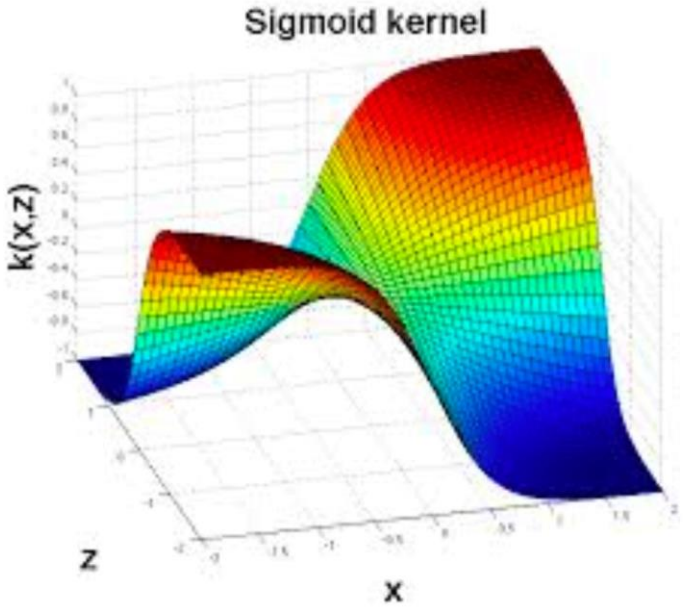
Σχήμα 2.13: Πολυωνυμικός Πυρήνας

Ο Πυρήνας Ακτινικής Συνάρτησης Βάσης [Σχ.2.14] είναι ένας ευέλικτος πυρήνας που χρησιμοποιείται όταν δεν υπάρχει προηγούμενη γνώση σχετικά με τα δεδομένα. Προσφέρει μια λύση γενικής χρήσης και ορίζεται ως μέτρο ομοιότητας μεταξύ δύο δειγμάτων, X και y, χρησιμοποιώντας την εξίσωση RBF. Αυτός ο πυρήνας χρησιμοποιείται ευρέως λόγω της ικανότητάς του να καταγράφει σύνθετες σχέσεις στα δεδομένα.



Σχήμα 2.14: Πυρήνας Ακτινικής Βάσης Συνάρτησης

Ο Σιγμοειδής Πυρήνας [Σχ.2.15] βρίσκει τις ρίζες του στα Νευρωνικά Δίκτυα και μπορεί να θεωρηθεί ως πληρεξούσιος για τέτοια δίκτυα. Ο Σιγμοειδής Πυρήνας απεικονίζεται ως μια ομαλή συνάρτηση σε σχήμα σιγμοειδούς.



Σχήμα 2.15: Σιγμοειδής Πυρήνας

Οι Μηχανές Διανυσμάτων Υποστήριξης προσφέρουν πολλά πλεονεκτήμα για τα προβλήματα ταξινόμησης. Επηρεάζονται λιγότερο από δεδομένα υψηλών διαστάσεων, χάρη στην εξάρτησή τους από τα διανύσματα υποστήριξης κοντά στα όρια απόφασης. Η στέρεη θεωρητική τους βάση που βασίζεται στην ελαχιστοποίηση του δομικού κινδύνου και στη μεγιστοποίηση των περιθωρίων παρέχει ισχυρές εγγυήσεις γενίκευσης. Οι Μηχανές Διανυσμάτων Υποστήριξης είναι λιγότερο επιρρεπείς σε υπερπροσαρμογή λόγω του βέλτιστου διαχωρισμού των κατηγοριών, διατηρώντας παράλληλα καλή ισορροπία. Ωστόσο, έχουν περιορισμούς όπως υπολογιστικό κόστος για μεγάλα σύνολα δεδομένων και πολύπλοκους πυρήνες. Είναι επίσης λιγότερο ερμηνεύσιμα σε σύγκριση με άλλους αλγόριθμους. Παρόλα αυτά, εξακολουθούν να χρησιμοποιούνται ευρέως λόγω της αποτελεσματικότητάς τους στη Γραμμική και Μη Γραμμική Ταξινόμηση, την ευρωστία και την αξιοπιστία τους σε διάφορους τομείς.

Ο αλγόριθμος των Μηχανών Διανυσμάτων Υποστήριξης προσφέρει μια σειρά από υπερπαραμέτρους που μπορούν να ρυθμιστούν προσεκτικά για την επίτευξη βέλτιστης απόδοσης. Μερικές από αυτές τις υπερπαράμετρους περιλαμβάνουν τον πυρήνα (kernel), την παράμετρο κανονικοποίησης C, gamma (γ), βαθμός (degree):

- kernel: ο πυρήνας καθορίζει τον τύπο του ορίου απόφασης που θα χρησιμοποιήσει η Μηχανή Διανυσμάτων Υποστήριξης. Οι κοινές επιλογές, όπως αναλύθηκε και παραπάνω, περιλαμβάνουν τον γραμμικό, τον πολυωνυμικό, τον πυρήνα με ακτινική συνάρτηση βάσης και τον σιγμοειδή. Η επιλογή του πυρήνα εξαρτάται από τα χαρακτηριστικά του συνόλου δεδομένων και την επιθυμητή πολυπλοκότητα του ορίου απόφασης.
- C: η παράμετρος κανονικοποίησης ελέγχει την αντιστάθμιση μεταξύ της μεγιστοποίησης του περιθωρίου και της ελαχιστοποίησης του σφάλματος εκπαίδευσης. Μια μικρότερη τιμή C οδηγεί σε μεγαλύτερο περιθώριο, αλλά μπορεί να επιτρέψει περισσότερα λάθη εκπαίδευσης, ενώ μια μεγαλύτερη τιμή C μειώνει το περιθώριο αλλά επιβάλλει αυστηρότερη τήρηση των δεδομένων εκπαίδευσης.
- gamma: η παράμετρος αυτή είναι αποκλειστικά για τον πυρήνα ακτινικής συνάρτησης βάσης. Καθορίζει την επιρροή μεμονωμένων παραδειγμάτων εκπαίδευσης. Μια μικρότερη τιμή gamma υποδεικνύει ένα ευρύτερο εύρος επιρροής, με αποτέλεσμα ένα πιο ομαλό όριο αποφάσης, ενώ μια μεγαλύτερη τιμή gamma εστιάζει σε πιο κοντικά παραδείγματα εκπαίδευσης, με αποτέλεσμα ένα πιο περίπλοκο και στενά προσαρμοσμένο όριο.
- degree: η παράμετρος αυτή είναι αποκλειστικά για τον πολυωνυμικό πυρήνα. Καθορίζει τον βαθμό της πολυωνυμικής συνάρτησης που χρησιμοποιείται για τον μετασχηματισμό των χαρακτηριστικών εισόδου. Οι τιμές υψηλότερου βαθμού επιτρέπουν πιο σύνθετα όρια αποφάσεων, αλλά μπορεί επίσης να αυξήσουν τον κίνδυνο υπερπροσαρμογής.

Η βελτιστοποίηση αυτών των υπερπαραμέτρων είναι κρίσιμη για την επίτευξη της καλύτερης δυνατής απόδοσης με τον αλγόριθμο Μηχανών Διανυσμάτων Υποστήριξης. Αυτή η διαδικασία συχνά περιλαμβάνει τεχνικές όπως διασταυρούμενη επικύρωση ή αναζήτηση πλέγματος για τη συστηματική διερεύνηση διαφορετικών συνδυασμών τιμών υπερπαραμέτρων και επιλογή της διαμόρφωσης που αποφέρει τη βέλτιστη αντιστάθμιση μεταξύ μεροληψίας και διακύμανσης.

2.3.2.4 Νευρωνικά Δίκτυα

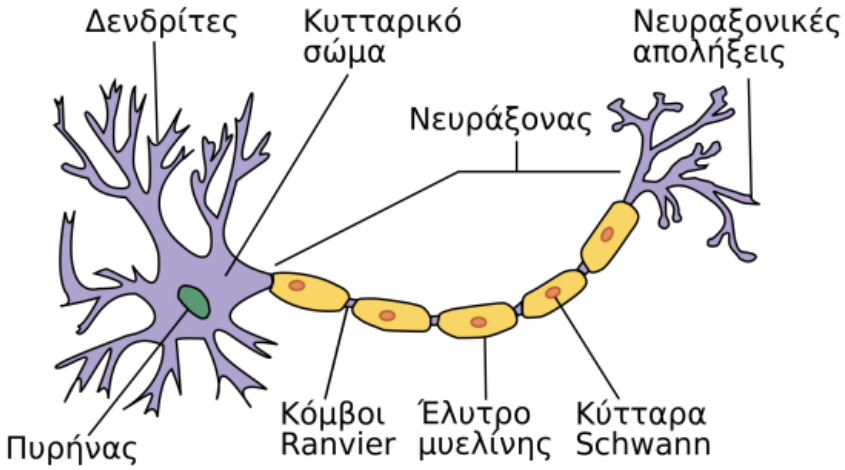
Τα Νευρωνικά Δίκτυα (Neural Networks) είναι μια ισχυρή και δημοφιλής κατηγορία μοντέλων στον τομέα της Τεχνητής Νοημοσύνης και της Μηχανικής Μάθησης. Εμπνευσμένα

από την δομή και την λειτουργία του ανθρώπινου εγκεφάλου, τα Νευρωνικά Δίκτυα έχουν σχεδιαστεί για να προσομοιώνουν την συμπεριφορά των διασυνδεδεμένων νευρώνων και την ικανότητά τους να επεξεργάζονται και να μαθαίνουν από δεδομένα.

Βιολογικά Νευρωνικά Δίκτυα

Νευρωνικό Δίκτυο ονομάζεται ένα κύκλωμα διασυνδεδεμένων νευρώνων [Σχ.2.16], οι οποίοι μεταδίδουν και επεξεργάζονται ηλεκτρικά και χημικά σήματα για να επιτρέψουν πολύπλοκες γνωστικές λειτουργίες και συμπεριφορές. Η μελέτη των βιολογικών νευρωνικών δικτύων αποκαλύπτει ιδέες για τους μηχανισμούς που διέπουν την αντίληψη, την μάθηση, την μνήμη και την λήψη αποφάσεων. Οι ερευνητές χρησιμοποιούν διάφορες τεχνικές, συμπεριλαμβανομένων ανατομικών μελετών, ηλεκτροφυσιολογίας και υπολογιστικής μοντελοποίησης, για να αποκρυπτογραφήσουν τα περίπλοκα μοτίβα συνδεσιμότητας, την συναπτική πλαστικότητα και τις αρχές επεξεργασίας πληροφοριών αυτών των δικτύων. Η κατανόηση των δομικών και λειτουργικών ιδιοτήτων των βιολογίκων νευρωνικών δικτύων παρέχει την βάση για την ανάπτυξη τεχνητών νευρωνικών δικτύων.

να καταγράφει σύνθετες σχέσεις στα δεδομένα.



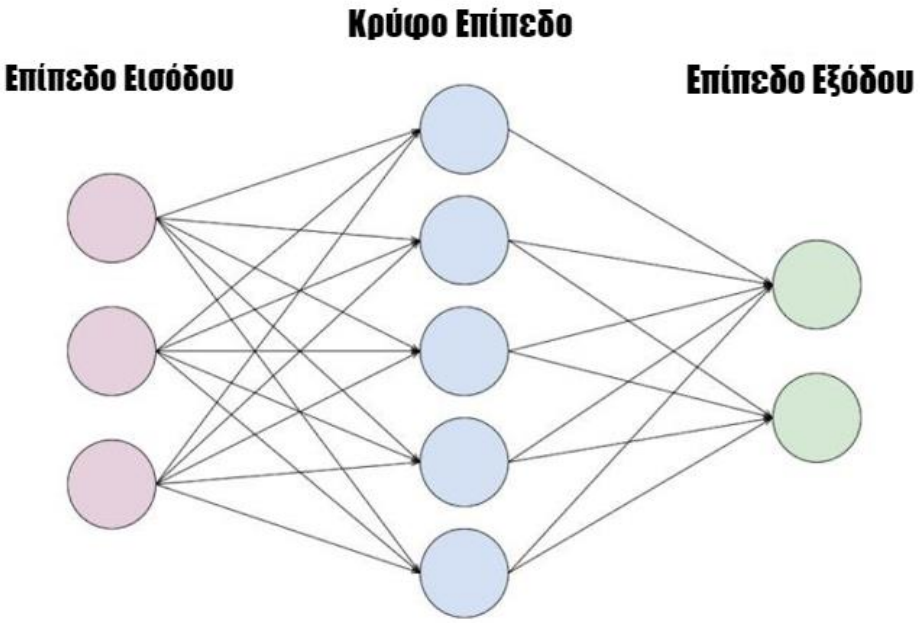
Σχήμα 2.16: Σχηματικό διάγραμμα ενός βιολογικού νευρώνα

Τεχνητά Νευρωνικά Δίκτυα

Ένα Τεχνητό Νευρωνικό Δίκτυο (Artificial Neural Networks) είναι ένα υπολογιστικό μοντέλο εμπνευσμένο από την δομή και την λειτουργία των βιολογικών νευρωνικών δικτύων. Πρόκειται για ένα μαθηματικό πλαίσιο που αποτελείται από διασυνδεδεμένους τεχνητούς

νευρώνες, γνωστούς και ως κόμβους ή μονάδες, οργανωμένους σε επίπεδα. Τα Τεχνητά Νευρωνικά Δίκτυα έχουν σχεδιαστεί για να επεξεργάζονται και να μαθαίνουν από δεδομένα, επιτρέποντας τους να κάνουν προβλέψεις, να αναγνωρίζουν μοτίβα και να εκτελούν διάφορες εργασίες.

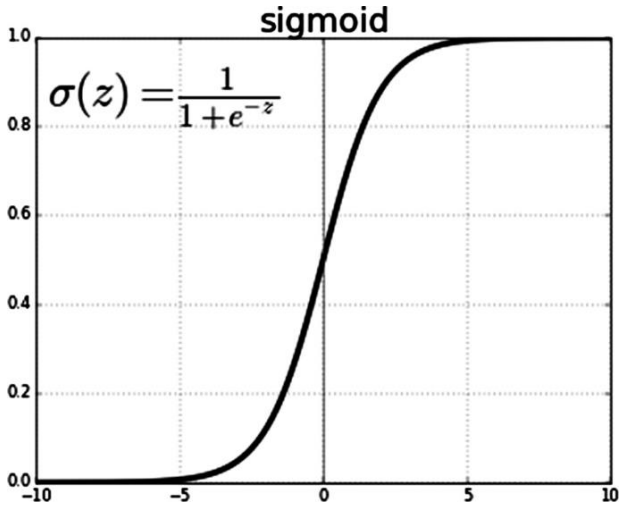
Οι Τεχνητοί Νευρώνες σε ένα Τεχνητό Νευρωνικό Δίκτυο λαμβάνουν σήματα εισόδου, εκτελούν μαθηματικές πράξεις σε αυτά και παράγουν ένα σήμα εξόδου. Κάθε σύνδεση μεταξύ νευρώνων συνδέεται με ένα βάρος που αντιπροσωπεύει τη δύναμη ή τη σημασία αυτής της σύνδεσης. Αυτά τα βάρη προσαρμόζονται κατά τη διαδικασία εκμάθησης για τη βελτιστοποίηση της απόδοσης του δικτύου. Η δομή ενός Τεχνητού Νευρωνικού Δικτύου [Σχ.2.17] συνήθως περιλαμβάνει ένα στρώμα εισόδου, ένα ή περισσότερα κρυφά επίπεδα και ένα στρώμα εξόδου. Το επίπεδο εισόδου λαμβάνει τα ακατέργαστα δεδομένα, τα οποία στη συνέχεια υποβάλλονται σε επεξεργασία μέσω των κρυφών επιπέδων. Τα κρυφά επίπεδα μετασχηματίζουν τα δεδομένα εισόδου εκτελώντας υπολογισμούς με βάση τις σταθμισμένες συνδέσεις μεταξύ των νευρώνων και σύμφωνα με μία συνάρτηση η οποία ονομάζεται συνάρτηση ενεργοποίησης. Τέλος, το επίπεδο εξόδου παράγει τις επιθυμητές προβλέψεις ή ταξινομήσεις με βάση τα μαθημένα μοτίβα και τα χαρακτηριστικά.



Σχήμα 2.17: Σχηματικό διάγραμμα ενός τεχνητού νευρωνικού δικτύου

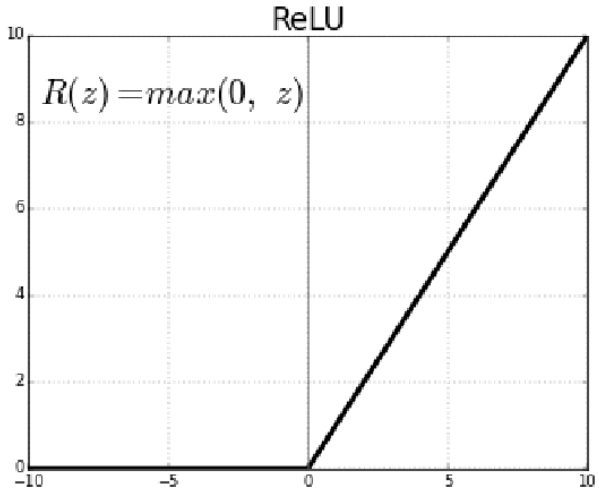
Οι συναρτήσεις ενεργοποίησης παίζουν κεντρικό ρόλο στον προσδιορισμό της εξόδου ενός νευρώνα και στην εισαγωγή μη γραμμικών μετασχηματισμών στο δίκτυο. Προσθέτοντας μη γραμμικότητα, τα τεχνητά νευρωνικά δίκτυα καθίστανται ικανά να μοντελοποιούν πολύπλοκες σχέσεις στα δεδομένα. Οι κοινές συναρτήσεις ενεργοποίησης που χρησιμοποιούνται περιλαμβάνουν την σιγμοειδή (sigmoid), την ReLU (Rectified Linear Unit) και υπερβολική εφαπτομένη (tanh).

- Σιγμοειδής συνάρτηση: πρόκειται για μια δημοφιλή επιλογή, λόγω της ομαλής καμπύλης σε σχήμα «S» [Σχ.2.18]. Αντιστοιχίζει το σταθμισμένο άθροισμα των εισόδων σε ένα εύρος μεταξύ 0 και 1, καθιστώντας το κατάλληλο για προβλήματα δυαδικής ταξινόμησης. Οι σιγμοειδείς συναρτήσεις εισάγουν τη μη γραμμικότητα, επιτρέποντας στο δίκτυο να μάθει περίπλοκα όρια αποφάσεων. Ωστόσο, υποφέρουν από το πρόβλημα της εξαφάνισης της κλίσης (gradient vanishing problem), περιορίζοντας την αποτελεσματικότητά τους σε Βαθιά Δίκτυα.



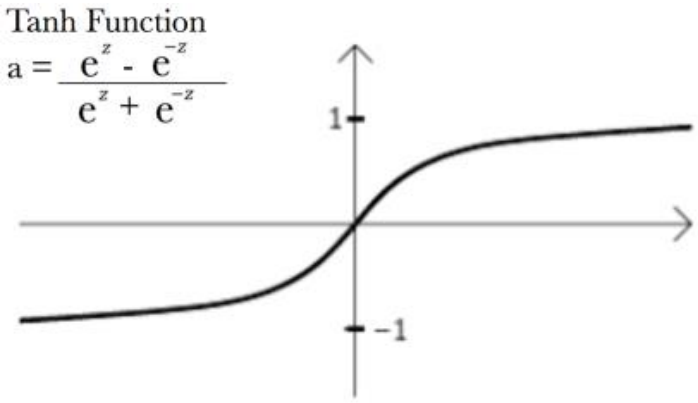
Σχήμα 2.18: Σχηματικό διάγραμμα της σιγμοειδούς συνάρτησης

- ReLU: πρόκειται για μια ευρέως χρησιμοποιούμενη συνάρτηση ενεργοποίησης που αντικαθιστά τις αρνητικές εισόδους με μηδέν [Σχ.2.19] και διατηρεί αμετάβλητες τις θετικές εισόδους. Προσφέρει υπολογιστική αποτελεσματικότητα και μετριάζει το πρόβλημα της κλίσης που εξαφανίζεται, επιτρέποντας την αποτελεσματική εκπαίδευση σε Βαθιά Δίκτυα. Οι λειτουργίες ενεργοποίησης ReLU είναι ευεργετικές για τη μοντελοποίηση αραιών και ιεραρχικών αναπαραστάσεων, συμβάλλοντας στη βελτιωμένη απόδοση σε διάφορες εργασίες.



Σχήμα 2.19: Σχηματικό διάγραμμα της συνάρτησης RELU

- Υπερβολική Εφαπτομένη: πρόκειται για μια επέκταση της σιγμοειδούς συνάρτησης, αντιστοιχίζοντας το σταθμισμένο άθροισμα των εισόδων σε μια περιοχή μεταξύ -1 και 1 [Σχ.2.20]. Οι συναρτήσεις αυτές εισάγουν συμμετρία γύρω από το μηδέν και παρέχουν ισχυρότερες διαβαθμίσεις σε σύγκριση με τις σιγμοειδείς συναρτήσεις. Είναι κατάλληλες για την αποτύπωση πιο περίπλοκων σχέσεων και παρουσιάζουν καλύτερες ιδιότητες σύγκλισης κατά τη διάρκεια της προπόνησης.



Σχήμα 2.20: Σχηματικό διάγραμμα της συνάρτησης της υπερβολικής εφαπτομένης

Οι συναρτήσεις ενεργοποίησης θα πρέπει να διαθέτουν βασικές ιδιότητες, όπως η διαφοροποίηση, η συνέχεια και ο μη κορεσμός, για να καταστεί δυνατή η αποτελεσματική εκπαίδευση και βελτιστοποίηση των Τεχνητών Νευρωνικών Δικτύων. Η

επιλογή της λειτουργίας ενεργοποίησης εξαρτάται από τις απαιτήσεις του προβλήματος, την αρχιτεκτονική του δικτύου και τις επιθυμητές ιδιότητες, όπως η αραιότητα, η ομαλότητα ή η υπολογιστική απόδοση.

Για να εκφράσουμε την έξοδο ενός νευρώνα n [Σχ.2.21] ενός κρυφού επιπέδου, μπορούμε να χρησιμοποιήσουμε την παρακάτω συνάρτηση:

$$y_n = f\,(\,\Sigma w_i x_i + \theta_n\,)$$

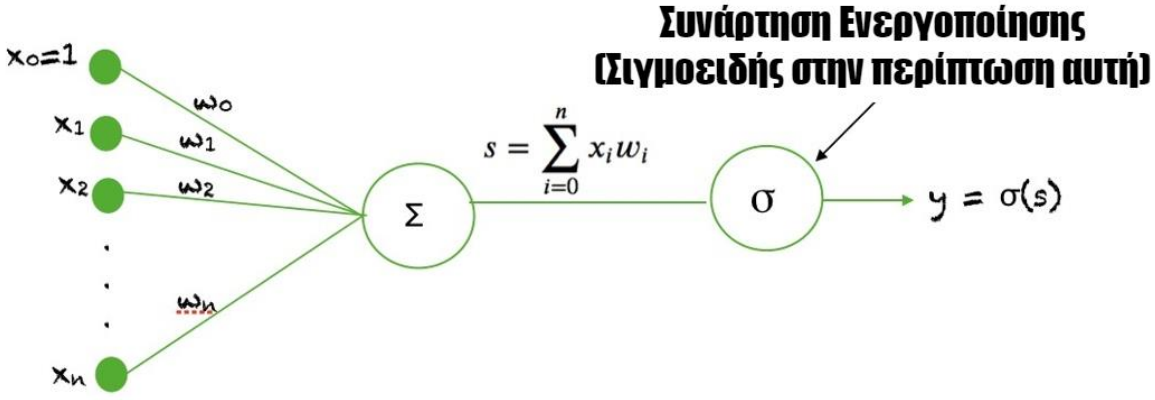
όπου

xi : i-οστή είσοδος από τις k που έχει το πρόβλημα

wi : το βάρος της διασύνδεσης με την i-οστή είσοδο

θk : η πόλωση για τον νευρώνα n (ανεξάρτητη του προβλήματος)

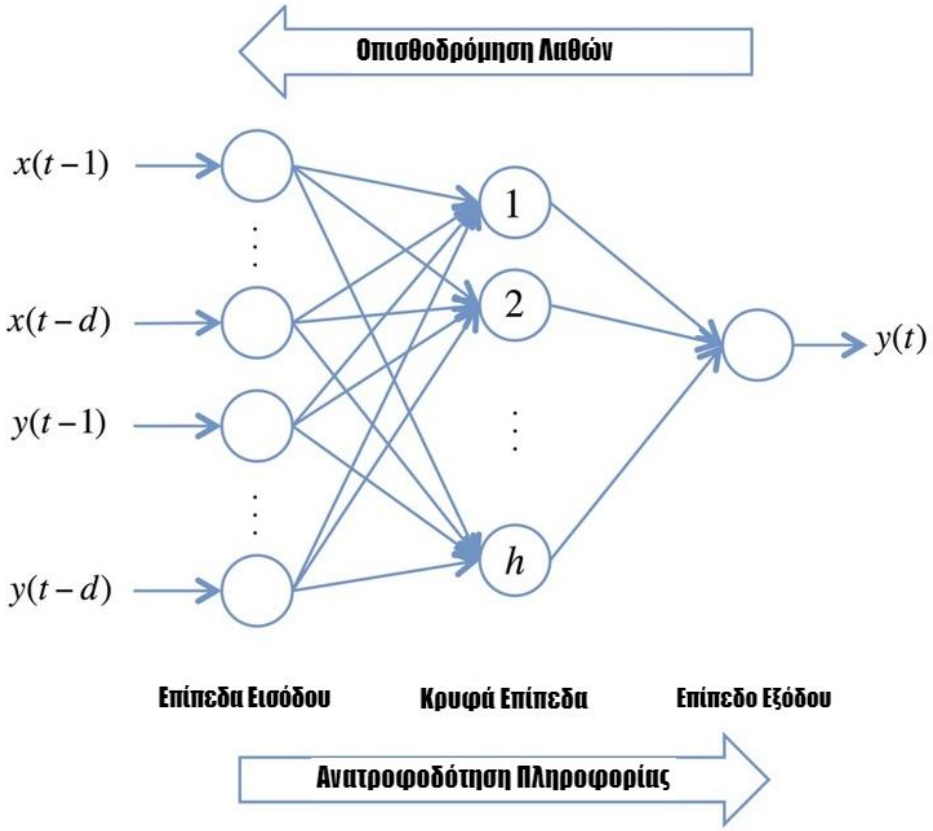
f(x) : συνάρτηση ενεργοποίησης



Σχήμα 2.21: Σχηματικό διάγραμμα υπολογισμού της εξόδου ενός νευρώνα

Η διαδικασία εκμάθησης περιλαμβάνει την εκπαίδευση του δικτύου σε ένα επισημασμένο σύνολο δεδομένων. Κατά τη διάρκεια της προπόνησης, αξιολογείται η απόδοση του δικτύου και τα βάρη προσαρμόζονται επαναληπτικά χρησιμοποιώντας αλγόριθμους, όπως ο «backpropagation» και η ανατροφοδότηση («feedforward») [Σχ.2.22]. Κατά την ανατροφοδότηση, τα δεδομένα ρέουν μέσω του δικτύου από το επίπεδο εισόδου στο επίπεδο εξόδου. Κάθε νευρώνας σε ένα επίπεδο λαμβάνει εισόδους από το προηγούμενο επίπεδο, εκτελεί ένα σταθμισμένο άθροισμα των εισόδων και εφαρμόζει μία συνάρτηση ενεργοποίησης για να παράγει μία έξοδο. Αυτή η διαδικασία επαναλαμβάνεται στρώμα προς στρώμα μέχρι να δημιουργηθεί η τελική έξοδος. Από την άλλη, ο backpropagation υπολογίζει το σφάλμα μεταξύ της εξόδου του δικτύου και της επιθυμητής εξόδου και στη συνέχεια

διαδίδει αυτό το σφάλμα προς τα πίσω μέσω του δικτύου για να ενημερώσει τα βάρη, επιτρέποντας στο δίκτυο να μάθει και να βελτιώσει τις προβλέψεις του.



Σχήμα 2.22: Σχηματικό διάγραμμα της διαδικασίας εκμάθησης σε ένα νευρωνικό δίκτυο, κάνοντας χρήση των feedforward και backpropagation

Επιπλεόν, ο αλγόριθμος «Κλίσης Κατάβασης» (Gradient Descent) [Σχ.2.23] είναι ένας ευρέως χρησιμοποιούμενος αλγόριθμος βελτιστοποίησης κατά την διαδικασία του backpropagation, που στόχο έχει την ελαχιστοποίηση του σφάλματος και την βελτιστοποίηση της απόδοσης του δικτύου. Προσαρμόζει επαναληπτικά τα βάρη προς την κατεύθυνση της αρνητικής κλίσης, προχωρώντας σταδιακά προς το ελάχιστο της επιφάνειας σφάλματος. Οι αλγόριθμοι βελτιστοποίησης, όπως ο Adam, ενισχύουν την αποτελεσματικότητα και την ταχύτητα σύγκλισης του gradient descent.

Τα Τεχνητά Νευρωνικά Δίκτυα παρέχουν πολλές υπερπαραμέτρους για συντονισμό, οι οποίες μπορούν να επηρεάσουν σημαντικά την απόδοση του μοντέλου. Αυτές οι υπερπαράμετροι καθορίζουν την αρχιτεκτονική, την διαδικασία εκπαίδευσης και τις τεχνικές κανονικοποίησης που χρησιμοποιούνται στα Τεχνητά Νευρωνικά Δίκτυα. Μερικές από αυτές τις υπερπαράμετρους περιλαμβάνουν τον αριθμό στρωμάτων και νευρώνων (Number of Layers

and Neurons), την συνάρτηση ενεργοποίησης (Activation Function), τον ρυθμό εκμάθησης (Learning Rate), το μέγεθος παρτίδας (Batch Size), την τεχνική κανονοποιήσης (Regularization), τον ρυθμό εγκατάλειψης (Dropout Rate), τον βελτιστοποιητή (Optimizer) και τον αριθμό εποχών εκπαίδευσης (Number of Training Epochs).

- Number of Layers and Neurons: Η αρχιτεκτονική ενός Τεχνητού Νευρωνικού Δικτύου περιλαμβάνει τον προσδιορισμό του αριθμού των στρωμάτων και του αριθμού των νευρώνων σε κάθε στρώμα. Αυτό περιλαμβάνει την επιλογή των επιπέδων εισόδου, κρυφών και εξόδου. Ο πειραματισμός με διαφορετικές διαμορφώσεις επιπέδων μπορεί να βοηθήσει στην εύρεση της βέλτιστης ισορροπίας μεταξύ της πολυπλοκότητας του μοντέλου και της ικανότητας εκμάθησης.
- Activation Function: Οι συναρτήσεις ενεργοποίησης εισάγουν μη γραμμικότητα στο δίκτυο και επηρεάζουν τον τρόπο με τον οποίο οι νευρώνες σε κάθε στρώμα διαδίδουν σήματα. Οι δημοφιλείς επιλογές περιλαμβάνουν τις sigmoid, tanh και ReLU. Η επιλογή των κατάλληλων συναρτήσεων ενεργοποίησης για διαφορετικά επίπεδα μπορεί να επηρεάσει την ικανότητα του δικτύου να μαθαίνει πολύπλοκα μοτίβα και να αποφεύγει ζητήματα όπως η εξαφάνιση ή η έκρηξη κλίσεων (vanishing or exploding gradients).
- Learning Rate: Ο ρυθμός εκμάθησης ελέγχει το μέγεθος του βήματος κατά τη διάρκεια της κλίσης κατάβασης (gradient descent), επηρεάζοντας την ταχύτητα και τη σύγκλιση της διαδικασίας βελτιστοποίησης. Η ρύθμιση του ρυθμού εκμάθησης σε υψηλές τιμές μπορεί να οδηγήσει σε υπέρβαση της βέλτιστης λύσης, ενώ ένας πολύ χαμηλός ρυθμός εκμάθησης μπορεί να οδηγήσει σε αργή σύγκλιση. Ο πειραματισμός με διαφορετικούς ρυθμούς μάθησης είναι ζωτικής σημασίας για την εύρεση της βέλτιστης τιμής.
- Batch Size: Το μέγεθος παρτίδας καθορίζει τον αριθμό των παραδειγμάτων εκπαίδευσης που υποβλήθηκαν σε επεξεργασία πριν από την ενημέρωση των βαρών του δικτύου κατά τη διάρκεια κάθε επανάληψης εκπαίδευσης. Τα μεγαλύτερα μεγέθη παρτίδων μπορούν να οδηγήσουν σε ταχύτερη προπόνηση λόγω παραλληλισμού, ενώ τα μικρότερα μεγέθη παρτίδων μπορούν να παρέχουν πιο συχνές ενημερώσεις βάρους και δυνητικά καλύτερη γενίκευση. Η εύρεση ενός κατάλληλου μεγέθους παρτίδας συχνά περιλαμβάνει μια αντιστάθμιση μεταξύ της υπολογιστικής απόδοσης και της απόδοσης του μοντέλου.
- Regularization: Η κανονικοποίηση βοηθά στην αποφυγή της υπερπροσαρμογής προσθέτοντας πρόσθετους περιορισμούς ή ποινές στη λειτουργία απώλειας κατά τη διάρκεια της προπόνησης. Οι κοινές τεχνικές κανονικοποίησης περιλαμβάνουν τις κανονικοποιήσεις L1 και L2 (απώλεια βάρους), εγκατάλειψη (dropout) και πρόωρη

διακοπή (early stopping). Ο συντονισμός των υπερπαραμέτρων κανονικοποίησης μπορεί να εξισορροπήσει την πολυπλοκότητα του μοντέλου και να αποτρέψει την υπερπροσαρμογή.

  - Early Stopping: Η πρόωρη διακοπή στοχεύει στην αποφυγή της υπερπροσαρμογής και στην βελτίωση της γενίκευσης παρακολουθώντας την απόδοση του μοντέλου σε ένα σύνολο επικύρωσης και σταματώντας την διαδικασία εκπαίδευσης όταν η απόδοση αρχίζει να επιδεινώνεται. Περιλαμβάνει την διαίρεση των διαθέσιμων δεδομένων σε σύνολα εκπαίδευσης και επικύρωσης. Κατά την διάρκεια της εκπαίδευσης, η απόδοση του μοντέλου αξιολογείται περιοδικά χρησιμοποιώντας το σύνολο επικύρωσης. Η απόδοση του μοντέλου συνήθως αξιολογείται χρησιμοποιώντας μια μέτρηση απόδοσης, όπως η ακρίβεια.

- Dropout Rate: Το Dropout είναι μια τεχνική κανονικοποίησης που απενεργοποιεί τυχαία ένα κλάσμα νευρώνων κατά τη διάρκεια της εκπαίδευσης για να μειώσει τις αλληλεξαρτήσεις και να ενισχύσει τη γενίκευση του μοντέλου. Ο ρυθμός εγκατάλειψης ελέγχει την πιθανότητα απενεργοποίησης ενός νευρώνα. Υψηλότερος ρυθμός εγκατάλειψης προωθεί μεγαλύτερη ευρωστία έναντι της υπερπροσαρμογής, αλλά ο υπερβολικά υψηλός ρυθμός μπορεί να εμποδίσει την ικανότητα εκμάθησης του δικτύου.
- Optimizer: Η επιλογή του βελτιστοποιητή καθορίζει τον αλγόριθμο που χρησιμοποιείται για την ενημέρωση των βαρών του δικτύου κατά τη διάρκεια της εκπαίδευσης. Οι δημοφιλείς βελτιστοποιητές περιλαμβάνουν το Stochastic Gradient Descent (SGD), το Adam, το RMSprop και το AdaGrad. Διαφορετικοί βελτιστοποιητές έχουν διαφορετικά χαρακτηριστικά όσον αφορά την ταχύτητα σύγκλισης, την αντίσταση στα τοπικά ελάχιστα και την προσαρμοστικότητα σε διαφορετικούς τύπους δεδομένων.
- Number of Training Epochs: Ο αριθμός των εποχών εκπαίδευσης καθορίζει πόσες φορές ολόκληρο το σύνολο δεδομένων εκπαίδευσης περνά μέσα από το δίκτυο κατά τη διάρκεια της εκπαίδευσης. Οι ανεπαρκείς εποχές μπορεί να οδηγήσουν σε μη προσαρμογή, ενώ οι υπερβολικές εποχές μπορεί να οδηγήσουν σε υπερπροσαρμογή. Είναι σημαντικό να παρακολουθείται η απόδοση εκπαίδευσης και επικύρωσης για να προσδιοριστεί ο βέλτιστος αριθμός εποχών.

Αυτές οι υπερπαράμετροι αλληλεπιδρούν μεταξύ τους και μπορούν να έχουν σημαντικό αντίκτυπο στην απόδοση του Τεχνητού Νευρωνικού Δικτύου. Ο σωστός συντονισμός μέσω του πειραματισμού, όπως η χρήση τεχνικών όπως η αναζήτηση πλέγματος ή η τυχαία

αναζήτηση, παίζουν κρίσιμο ρόλο στην εύρεση του βέλτιστου συνδυασμού υπερπαραμέτρων που μεγιστοποιούν την απόδοση και τις δυνατότητες γενίκευσης του μοντέλου.

2.3.3 Μετρικές Απόδοσης και Αξιολόγησης των Μοντέλων

Οι μετρικές απόδοσης σε ένα πρόβλημα Ταξινόμησης διαδραματίζουν κρίσιμο ρόλο στην αξιολόγηση της αποτελεσματικότητας και της ακρίβειας των μοντέλων πρόβλεψης. Αυτές οι μετρικές παρέχουν ποσοτικά μέτρα για την αξιολόγηση της απόδοσης των αλγορίθμων ταξινόμησης συγκρίνοντας τις προβλεπόμενες ετικέτες κλάσης τους με τις αληθινές ετικέτες κλάσεων από το σύνολο δεδομένων. Η πλήρης κατανόηση των μετρήσεων απόδοσης είναι απαραίτητη για τη μέτρηση των προγνωστικών δυνατοτήτων του μοντέλου και τη λήψη τεκμηριωμένων αποφάσεων. Οι μετρικές που χρησιμοποιήθηκαν στην παρούσα μελέτη είναι οι ακόλουθες:

2.3.3.1 Πίνακας Σύγχυσης

Πρόκειται για μια αναπαράσταση πίνακα της απόδοσης ενός μοντέλου ταξινόμησης που συγκρίνει τις προβλεπόμενες ετικέτες κλάσεων με τις πραγματικές ετικέτες κλάσεων από το σύνολο δεδομένων. Παρέχει μια λεπτομερή ανάλυση των προβλέψεων του μοντέλου, επιτρέποντας μια πιο λεπτομερή ανάλυση της απόδοσής του.

Ο πίνακας σύγχυσης είναι συνήθως οργανωμένος ως ένας πίνακας 2x2 [Σχ.2.25] για προβλήματα δυαδικής ταξινόμησης, αλλά μπορεί να χρησιμοποιηθεί και για προβλήματα ταξινόμησης πολλαπλών κλάσεων. Ο πίνακας περιλαμβάνει τέσσερα βασικά στοιχεία:

- True Positives (TP): ο αριθμός των περιπτώσεων που προβλέφθηκαν σωστά ως θετικές.
- True Negatives (TN): ο αριθμός των περιπτώσεων που προβλέφθηκαν σωστά ως αρνητικές.
- False Positives (FP): ο αριθμός των περιπτώσεων που έχουν προβλεφθεί λανθασμένα ως θετικές.
- False Negatives (FN): ο αριθμός των περιπτώσεων που προβλέφθηκαν λανθασμένα ως αρνητικές.

2.3.3.2 Ακρίβεια

Πρόκειται για μία θεμελιώδη μετρική που υπολογίζει την αναλογία των σωστά προβλεπόμενων δειγμάτων επί του συνολικού αριθμού δειγμάτων στο σύνολο δεδομένων. Παρέχει μια γενική επισκόπηση της προγνωστικής ακρίβειας του μοντέλου. Μαθηματικά, η ακρίβεια ορίζεται ως:

$$accuracy = (TP + TN) / (TP + TN + FP + FN)$$

όπου:

TP: True Positives

TN: True Negatives

FP: False Positives

FN: False Negatives

Η ακρίβεια ερμηνεύεται εύκολα και συχνά εκφράζεται ως ποσοστό, που κυμαίνεται από το 0% έως το 100%. Επίσης, λειτουργεί καλά όταν το σύνολο δεδομένων έχει ισορροπημένες κλάσεις, που σημαίνει ότι ο αριθμός παρουσιών σε κάθε κλάση είναι περίπου ίσος. Διαφορετικά, μπορεί να είναι παραπλανητική και καλό είναι να ληφθούν υπόψη κι άλλες μετρικές.

2.3.3.3 Λογιστική Απώλεια

Πρόκειται για μία ευρέως χρησιμοποιούμενη μετρική απόδοσης για την αξιολόγηση της ακρίβειας των μοντέλων ταξινόμησης. Σε αντίθεση με άλλες μετρικές, λαμβάνει υπόψη τις προβλεπόμενες πιθανότητες των ετικετών κλάσεων και όχι μόνο τις ίδιες τις προβλεπόμενες ετικέτες κλάσεων. Ουσιαστικά, υπολογίζει την ανομοιότητα μεταξύ των προβλεπόμενων πιθανοτήτων και των αληθινών ετικετών κλάσης. Όσο χαμηλότερη είναι η τιμή της λογιστικής απώλειας, τόσο καλύτερη είναι η απόδοση του μοντέλου. Μαθηματικά, ορίζεται ως ο αρνητικός λογάριθμος της προβλεπόμενης πιθανότητας που αποδίδεται στην ετικέτα αληθινής κλάσης.

$$log\ loss = - (1/n) * \sum[y * log(p) + (1-y) * log(1-p)]$$

όπου:

n: ο συνολικός αριθμός δειγμάτων στο σύνολο δεδομένων

y: η πραγματική ετικέτα κλάσης (0 ή 1)

p: η προβλεπόμενη πιθανότητα της θετικής κλάσης.

Μια λογιστική απώλεια 0 υποδεικνύει ένα τέλειο μοντέλο. Οσο η τιμή ανεβαίνει [Σχ.2.26] (φτάνοντας έως το 1), τόσο χαμηλότερη απόδοση θα έχει το μοντέλο. Η λογιστική απώλεια ενθαρρύνει τα μοντέλα να παράγουν καλά βαθμονομημένες εκτιμήσεις πιθανοτήτων και τιμωρεί τις υπερβολικά σίγουρες ή κακώς βαθμονομημένες προβλέψεις. Επίσης, πρόκειται για μια ιδιαίτερα ευαίσθητη μετρική σε μικρές αλλαγές στις πιθανότητες, ειδικότερα όταν αυτές είναι κοντά στο 0 ή στο 1. Αυτή η ευαισθησία το καθιστά αποτελεσματικό για δίκαιη σύγκριση και λεπτομερή αξιολόγηση των διάφορων πιθανοτικών μοντέλων.

3 *Βιβλιογραφική Ανασκόπηση*

Στον τομέα της πρόβλεψης του αποτελέσματος σε έναν αγώνα αντισφαίρισης, υπάρχει αυξανόμενο ενδιαφέρον για την χρήση τεχνικών Μηχανικής Μάθησης και Ανάλυσης Δεδομένων. Στόχος αυτού του κεφαλαίου είναι να παρέχει μια ολοκληρωμένη ανασκόπηση της τρέχουσας έρευνας και της βιβλιογραφίας σχετικά με τα μοντέλα Μηχανικής Μάθησης για την πρόβλεψη αγώνων αντισφαίρισης. Διευρευνά τα δυνατά σημεία, τους περιορισμούς και τις επιπτώσεις αυτών των μοντέλων. Κατανοώντας τις προηγμένες («state-of-the-art») προσεγγίσεις και μεθοδολογίες, οι ερευνητές και οι επαγγελματίες μπορούν να συγκρίνουν και να αξιολογήσουν τα δικά τους μοντέλα, αποκτώντας γνώσεις για την ακρίβεια, την ευρωστία και την γενίκευση τους. Μέχρι στιγμής, έχει χρησιμοποιηθεί μια ποικιλία μεθοδολογιών συμπεριλαμβανομένων στατιστικών μοντέλων, ανάλυσης χρονοσειρών, προσεγγίσεων ταξινόμησης και ανάλυσης παλινδρόμησης.

Θα πρέπει να σημειωθεί ότι οι συγκρίσεις ή οι γενικεύσεις των αποτελεσμάτων πρέπει να γίνονται προσεκτικά, λαμβάνοντας υπόψη τις διαφορετικές μεθοδολογίες και τις πηγές δεδομένων που χρησιμοποιούνται. Η ποικιλομορφία στις προσεγγίσεις υπογραμμίζει την αναγκαιότητα της προσεχτικής ερμηνείας και αξιολόγησης των ευρημάτων. Επιπλέον, υποδεικνύει την ανάγκη για πρόσθετη έρευνα για την διερεύνηση της αξιοπιστίας και της εφαρμοσιμότητας των μοντέλων πρόβλεψης σε διαφορετικά πλαίσια και χρονικά πεδία. Αναγνωρίζοντας αυτούς τους παράγονες, μπορεί να επιτευχθεί μια πιο ολοκληρωμένη κατανόηση των δυνατοτήτων και των περιορισμών των μοντέλων πρόβλεψης αγώνων αντισφαίρισης σε ένα ακαδημαϊκό περιβάλλον.

Μερικές από τις αξιοσημείωτες μελέτες που πραγματοποιήθηκαν σε αυτόν το πεδίο έρευνας αναφέρονται εν συντομία.

Οι Barnett και Clarke χρησιμοποιούν δεδομένα ιστορικού αγώνα για να προβλέψουν μεμονωμένα σημεία και να υπολογίσουν την πιθανότητα του αποτελέσματος ολόκληρου του αγώνα με βάση μια αλυσίδα Markov [7]. Ομοίως, ο Knottenbelt και οι συναργάτες αναλύουν ένα μοντέλο Markov που αποφέρει απόδοση στοιχήματος περίπου 4% [8].

Οι Scheibehenne και Broeder παρουσιάζουν στοιχεία που υποδεικνύουν ότι η απλή αναγνώριση των ονομάτων των παικτών από ερασιτέχνες παίκτες και λαϊκούς μπορεί να ξεπεράσει τις προβλέψεις που βασίζονται σε κατάταξη και σπορά των ειδικών [9]. Ωστόσο, οι αποδόσεις των διαδικτυακών προγνώσεων έχουν δείξει ακόμα καλύτερη απόδοση στην πρόβλεψη αγώνων. Αυτό δείχνει την αξία της ενσωμάτωσης εξωτερικών πηγών πληροφοριών, όπως οι αποδόσεις προγνώσεων.

Οι Somboonphokkaphan, Phimoltares και Lursinsap χρησιμοποίησαν ανάλυση χρονοσειρών και μια μέθοδο βασισμένη στο Multilayer Perceptron (MLP) για την πρόβλεψη των αποτελεσμάτων των αγώνων αντισφαίρισης [10]. Έλαβαν χαρακτηριστικά από το ιστορικό χρονοσειρών των δεδομένων και χρησιμοποίησαν την αποθήκευση δεδομένων για ανάλυση ταξινόμησης. Τόσο τα περιβαλλοντικά όσο και τα στατιστικά δεδομένα ελήφθησαν υπόψη για την εκτίμηση των αποτελεσμάτων του αγώνα.

Οι Del Corral και Prieto-Rodriguez διεξάγουν μια ολοκληρωμένη μελέτη χρησιμοποιώντας μοντέλα probit βαθμονομημένα με την προηγούμενη απόδοση, τα φυσικά χαρακτηριστικά και τα χαρακτηριστικά αγώνα [11]. Η ανάλυση αποκαλύπτει ότι οι πληροφορίες κατάταξης διαδραματίζουν κρίσιμο ρόλο στην ακρίβεια της πρόβλεψης. Οι διαφορές ηλικίας εμφάνισαν επίσης σημαντική επίδραση στα αποτελέσματα των αγώνων τόσο για άνδρες όσο και για γυναίκες, αν και με διαφορετικά πρότυπα.

Οι Panjan, Sarabon και Filipcic επέδειξαν τη δυνατότητα πρόβλεψης των αποτελεσμάτων του αγώνα αποκλειστικά με βάση τα χαρακτηριστικά των παικτών ως παραμέτρους εισαγωγής [12]. Χρησιμοποίησαν διάφορα φυσικά χαρακτηριστικά του παίκτη, όπως το σωματικό βάρος, και χρησιμοποίησαν διαδικασίες ταξινόμησης και μηχανικής μάθησης για να παράγουν τις προβλέψεις.

Οι McHale και Morton χρησιμοποιούν ένα μοντέλο πιθανοτήτων για ζευγαρωμένες συγκρίσεις, το οποίο ξεπερνάει τα μοντέλα που βασίζονται σε λογιστική παλινδρόμηση όσον αφορά τόσο την πρόβλεψη αγώνα όσο και τις εφικτές αποδόσεις στοιχημάτων [13]. Το μοντέλο τους ενσωματώνει τις προηγούμενες επιδόσεις των παικτών και την επιφάνεια του διαγωνισμού. Οι Lyocsa και Vyrost διερευνούν επίσης μοντέλα σύγκρισης ζευγαρώματος και ερευνούν διάφορους κανόνες στοιχηματισμού με βάση τις πιθανότητες και τις κατατάξεις [14].

Ο Ma και οι συνεργάτες του χρησιμοποιούν λογιστική παλινδρόμηση για να προβλέψει τα αποτελέσματα του αγώνα, ενσωματώνοντας μεταβλητές που σχετίζονται με τα χαρακτηριστικά του παίκτη και του αγώνα [15]. Η προσέγγισή τους πέτυχε ψευδο-R2 περίπου 80% και εντόπισε με ακρίβεια τον νικητή σε πάνω από το 90% των περιπτώσεων.

Ο Kovalchik κατηγοριοποιεί τα μοντέλα πρόβλεψης σε τρεις κύριους τύπους: μοντέλα σύγκρισης που βασίζονται σε παλινδρόμηση, μοντέλα που βασίζονται σε σημεία και σε μοντέλα σύγκρισης ζευγαρώματος [16]. Αυτά τα μοντέλα στοχεύουν να αναλύσουν διαφορετικές πτυχές του παιχνιδιού και να παρέχουν πληροφορίες για το πιθανό αποτέλεσμα. Επιπλέον, αρκετές μελέτες έχουν χρησιμοποιήσει τις αποδόσεις των προγνωστικών ως σημείο αναφοράς για σύγκριση.

Οι Lisi και Zanella χρησιμοποιούν ένα μοντέλο λογιστικής παλινδρόμησης που ενσωμάτωσε χαρακτηριστικά όπως η κατάταξη, η ηλικία των παικτών, ο παράγοντας πλεονέκτημα έδρας και πληροφορίες που προέρχονται από τις αποδόσεις των στοιχημάτων [17]. Αναφέρουν απόδοση περίπου 16% μέσω μιας στρατηγικής στοιχήματος με βάση τις προβλέψεις τους.

Οι Gu και Saaty αναπτύσσουν ένα μοντέλο πρόβλεψης συνδυάζοντας δεδομένα και κρίσεις ειδικών χρησιμοποιώντας ένα αναλυτικό μοντέλο διεργασίας δικτύου (Network Process Model) [18]. Αν και βασίζεται σε ένα μικρό μέγεθος δείγματος λιγότερων από 100 αγώνων, το μοντέλο τους πετυχαίνει ακρίβεια πρόβλεψης περίπου 85%.

Ο Ingram προτείνει ένα μοντέλο που βασίζεται σε σημεία, χρησιμοποιώντας μια ιεραρχική προσέγγιση Bayes για την πρόβλεψη αγώνων [19]. Αυτό το μοντέλο ενσωματώνει πληροφορίες επιφάνειας, τουρνουά και ημερομηνίας αγώνα, οδηγώντας σε αποτελέσματα συγκρίσιμα με άλλες κατηγορίες μοντέλων.

Ο Gorgi και οι συνεργάτες του προτείνει ένα δυναμικό στατιστικό μοντέλο που αντιπροσωπεύει τις χρονικά μεταβαλλόμενες ικανότητες των παικτών σε διαφορετικούς τύπους επιφάνειας γηπέδου [20]. Αυτό το μοντέλο αποδίδει ανώτερη απόδοση σε σύγκριση με μοντέλα που βαθμονομήθηκαν αποκλειστικά με βάση τις πληροφορίες κατάταξης.

Οι Candila και Palazzo διερευνούν την χρήση των Τεχνητών Νευρωνικών Δικτύων με διάφορες μεταβλητές εισόδου για την πρόβλεψη αποτελεσμάτων σε αγώνες αντισφαίρισης [21]. Το Τεχνητό Νευρωνικό Δίκτυο σχεδιάστηκε κάνοντας χρήση της «ανατροφοδότησης» και του «backpropagation» και βελτιστοποιήθηκε χρησιμοποιώντας τον αλγόριθμο Adam. Οι κόμβοι εισόδου περιελάμβαναν μεταβλητές από υπάρχουσες προσεγγίσεις για πρόβλεψη των αγώνων, καθώς και μεταβλητές που δημιουργήθηκαν σκόπιμα για την καταγραφή πρόσθετων πτυχών. Το προτεινόμενο μοντέλο ξεπέρασε τέσσερις από τις πέντε ανταγωνιστικές προσεγγίσεις και ήταν τουλάχιστον όσο καλή όσο το μοντέλο παλινδρόμησης LZR των Lisi και Zanella. Οι συγγραφείς προτείνουν επίσης τέσσερις στρατηγικές στοιχηματισμού που βασίζονται σε εκτιμώμενα ποσοστά δειγμάτων πιθανοτήτων για να βοηθήσουν στην επιλογή αγώνα για στοιχήματα. Συγκρίνοντας την απόδοση των επενδύσεων του προτεινόμενου μοντέλου με το LZR, το πρώτο απέδωσε σταθερά υψηλότερες καθαρές αποδόσεις.

Ο Serano χρησιμοποιεί ένα μοντέλο λογιστικής παλινδρόμησης το οποίο εκπαίδευσε σε τέσσερις διαφορετικές σεζόν, το οποίο κατόρθωσε να ξεπεράσει σε απόδοση ένα βασικό μοντέλο το οποίο στηριζόταν μόνο στις βαθμολογίες ATP [22]. Για την βελτίωση της απόδοσης, εκπαιδεύτηκε ένα κρυφό νευρωνικό δίκτυο και συγκρίθηκε με το μοντέλο λογιστικής παλινδρόμησης. Το νευρωνικό δίκτυο μπόρεσε και ξεπέρασε το μοντέλο λογιστικής παλινδρόμησης τόσο στην λογιστική απώλεια όσο και στην ακρίβεια. Επιπλέον, οι προβλέψεις του νευρωνικού δικτύου μεταφράστηκαν σε κερδοφορία στην αγορά διαδικτυακών στοιχημάτων, βελτιώνοντας τις στρατηγικές στοιχηματισμού.

Η Wilkens παρουσιάζει μία ανάλυση περίπου 39.000 αγώνων αντισφαίρισης για τις χρονιές από 2010 έως 2019, χρησιμοποιώντας ένα ολοκληρωμένο σύνολο δεδομένων και συμπέρανε ότι η κατάταξη των παικτών και οι αποδόσεις στοιχημάτων παρέχουν τις πιο πολύτιμες πληροφορίες για την πρόβλεψη των αγώνων [23]. Τα δεδομένα ιστορικού αγώνων και παικτών έχουν ελάχιστη επίδραση στην ακρίβεια της πρόβλεψης. Οι τεχνικές Μηχανικής Μάθησης επιτυγχάνουν περίπου 70% ακρίβεια, παρόμοια με τις βασικές προσεγγίσεις.

Τα χωροχρονικά δεδομένα έχουν αποδειχθεί πολύτιμα για την πρόβλεψη βολών στο τένις και την πρόβλεψη αποτελεσμάτων αγώνα. Οι Wei, Lucey, Morgan και Sridharan προσπάθησαν να προβλέψουν τα αποτελέσματα διαφόρων αθλημάτων, συμπεριλαμβανομένης της αντισφαίρισης, χρησιμοποιώντας χωροχρονική ανάλυση δεδομένων [24]. Ομοίως, οι Timmaraju, Palnitkar και Khanna ανέπτυξαν έναν αλγόριθμο πρόβλεψης ποδοσφαίρου με βάση την προηγούμενη απόδοση (KPP) για την εκτίμηση των αποτελεσμάτων των ποδοσφαιρικών αγώνων [25].

Στον τομέα της πρόβλεψης αθλημάτων, η εφαρμογή τεχνικών Μηχανικής Μάθησης αντιπροσωπεύει έναν σχετικά νέο και αναδυόμενο τομέα. Στον τομέα της αντισφαίρισης, μέχρι στιγμής έχει διεξαχθεί μόνο ένας περιορισμένος αριθμός μελετών. Παρά το ευρύ φάσμα προσεγγίσεων, πηγών δεδομένων, μεθόδων βαθμονόμησης και μετρήσεων αξιολόγησης που χρησιμοποιούνται, οι αναφερόμενες ακρίβειες πρόβλεψης κυμαίνονται γενικά γύρω από το εύρος 70-80%, με ορισμένους ισχυρισμούς να φτάνουν έως και το 99%. Αξίζει να σημειωθεί ότι οι περισσότερες μελέτες συμφωνούν ότι αυτά τα μοντέλα συνήθως δεν μπορούν να ξεπεράσουν τις προβλέψεις που συνάγονται από τις αποδόσεις των στοιχημάτων. Ωστόσο, είναι σημαντικό να ληφθεί υπόψη ότι αυτές οι αναλύσεις βασίζονται συχνά σε σχετικά σύντομες περιόδους συνήθως ενός έτους ή λιγότερο.

4 *Σύνολο Δεδομένων*

Οι τεχνικές Μηχανικής Μάθησης έχουν κερδίσει μεγάλη προσοχή στον τομέα της Αθλητικής Ανάλυσης (Sports Analytics), ιδιαίτερα στην πρόβλεψη των αποτελεσμάτων των αγώνων αντισφαίρισης. Η αντισφαίριση, ως ένα δημοφιλές άθλημα με πληθώρα δεδομένων, παρέχει μια εξαιρετική ευκαιρία διερεύνησης της αποτελεσματικότητας των αλγορίθμων Μηχανικής Μάθησης. Ωστόσο, η ακρίβεια και η αξιοπιστία αυτών των προβλέψεων εξαρτώνται σε μεγάλο βαθμό από την ποιότητα του συνόλου δεδομένων που χρησιμοποιείται.

Η σημασία της χρήσης ενός συνόλου δεδομένων υψηλής ποιότητας στην έρευνα της Μηχανικής Μάθησης δεν μπορεί να υπερεκτιμηθεί. Ένα σχολαστικά επιμελημένο και αντιπροσωπευτικό σύνολο δεδομένων χρησιμεύει ως βάση για την ανάπτυξη ακριβών μοντέλων. Στο πλαίσιο της πρόβλεψης των αποτελεσμάτων των αγώνων αντισφαίρισης, αρκετές κρίσιμες πτυχές της προετοιμασίας δεδομένων αναλαμβάνουν κεντρικό ρόλο στη διασφάλιση της ακεραιότητας και της αποτελεσματικότητας των αλγορίθμων Μηχανικής Μάθησης. Αυτές οι πτυχές περιλαμβάνουν παράγοντες όπως η αξιοπιστία της πηγής δεδομένων, η πληρότητα των διαδικασιών καθαρισμού δεδομένων, η προσπάθεια επίτευξης ισορροπίας δεδομένων και η επίτευξη συμμετρικής αναπαράστασης χαρακτηριστικών. Κάθε μία από αυτές τις πτυχές συμβάλλει σημαντικά στη συνολική ευρωστία και αξιοπιστία των μοντέλων πρόβλεψης που χρησιμοποιούνται στη μελέτη.

4.1 Προέλευση Συνόλου Δεδομένων

Η προέλευση του συνόλου δεδομένων διαδραματίζει κρίσιμο ρόλο στη διασφάλιση της αξιοπιστίας των δεδομένων που χρησιμοποιούνται στην έρευνα για την πρόβλεψη των αποτελεσμάτων των αγώνων αντισφαίρισης. Αξιοποιώντας δεδομένα από αξιόπιστες πηγές, οι ερευνητές μπορούν να εξασφαλίσουν την ακρίβεια, την πληρότητα και την ακεραιότητα των πληροφοριών. Αυτή η προσεκτική επιλογή συνόλων δεδομένων υψηλής ποιότητας αποτελεί μια σταθερή βάση για ανάλυση, δίνοντας τη δυνατότητα στους ερευνητές να αναπτύξουν ισχυρά μοντέλα πρόβλεψης και να δημιουργήσουν αξιόπιστες πληροφορίες στον τομέα της Αθλητικής Ανάλυσης. Η χρήση τέτοιων δεδομένων υψηλής ποιότητας ενισχύει τη συνολική ακεραιότητα και αξιοπιστία των ευρημάτων και συμβάλλει στην προώθηση της κατανόησης των προβλέψεων αγώνων αντισφαίρισης μέσω τεχνικών Μηχανικής Μάθησης.

Το Διαδίκτυο έχει φέρει επανάσταση στην προσβασιμότητα και τη διαθεσιμότητα των δεδομένων αγώνων αντισφαίρισης, καθιστώντας το πολύτιμο πόρο για αναλύσεις αθλημάτων. Ιστότοποι όπως το atpworldtour.com προσφέρουν εκτενείς πληροφορίες για τα προφίλ παικτών, τα αποτελέσματα των αγώνων και τα στατιστικά στοιχεία, δίνοντας τη δυνατότητα στους ερευνητές να αναλύσουν τις επιδόσεις μεμονωμένων παικτών σε συγκεκριμένους αγώνες. Επιπλέον, πλατφόρμες όπως το tennis-data.co.uk παρέχουν δομημένα ιστορικά δεδομένα σε μορφές όπως αρχεία CSV ή Excel. Αυτά τα σύνολα δεδομένων προσφέρουν πλήθος πληροφοριών και μπορούν εύκολα να ληφθούν για ανάλυση.

Για πιο ολοκληρωμένα και σύνθετα σύνολα δεδομένων, οι ερευνητές μπορεί να εξετάσουν το ενδεχόμενο αγοράς δεδομένων μέσω διαδικτύου, τα οποία μπορούν να παρέχουν ένα ευρύτερο χρονικό διάστημα και ένα ευρύ φάσμα χαρακτηριστικών που περιλαμβάνει αγώνες και παίκτες. Η πληθώρα δεδομένων αγώνων αντισφαίρισης στο Διαδίκτυο δίνει τη δυνατότητα στους ερευνητές να εξερευνήσουν και να αναλύσουν διάφορες πτυχές του αθλήματος χρησιμοποιώντας προηγμένες αναλυτικές τεχνικές.

Τα δεδομένα αγώνων αντισφαίρισης που χρησιμοποιήθηκαν στη μελέτη προέρχονται από ένα σύνολο δεδομένων ανοιχτού κώδικα που είναι διαθέσιμο στο GitHub, συγκεκριμένα το αποθετήριο αντισφαίρισης αγώνων ATP του Jeff Sackmann [26]. Αυτό το σύνολο δεδομένων περιλαμβάνει αποτελέσματα αγώνων από την Open Era, ξεκινώντας το 1968, έως τον πιο πρόσφατο μήνα με τακτικές ενημερώσεις. Παρέχει μια πλούσια και εκτενή συλλογή δεδομένων αντιστοίχισης για ανάλυση. Συγκεκριμένα, το σύνολο δεδομένων περιλαμβάνει συμπληρωματικά στατιστικά στοιχεία αγώνων για τους πιο πρόσφατους αγώνες, ξεκινώντας

από το 2000. Αυτά τα στατιστικά προσφέρουν πολύτιμες πληροφορίες για διάφορες μετρήσεις απόδοσης, όπως ο αριθμός των άσων, τα διπλά σφάλματα και άλλοι σχετικοί δείκτες. Επιπλέον, το σύνολο δεδομένων ενσωματώνει δεδομένα προγνωστικών και πιθανοτήτων νίκης-ήττας από εξειδικευμένες υπηρεσίες στον τομέα, που λαμβάνονται από τον ιστότοπο Tennis-Data [27]. Αυτές οι πρόσθετες πληροφορίες σχετικά με τις αποδόσεις και τα δεδομένα πρόγνωσης εκτείνονται από το 2010 και μετά, προσθέτοντας μια μοναδική διάσταση στην ανάλυση.

Η περιεκτική φύση αυτού του συνόλου δεδομένων επιτρέπει τη διεξοδική εξέταση των χαρακτηριστικών των αγώνων και των παικτών, καθώς και των αποτελεσμάτων μεμονωμένων αγώνων. Οι ερευνητές μπορούν να αξιοποιήσουν αυτό το σύνολο δεδομένων για να πραγματοποιήσουν εις βάθος αναλύσεις σε διάφορους παράγοντες που επηρεάζουν τα αποτελέσματα των αγώνων και να διερευνήσουν τις τάσεις απόδοσης στο άθλημα της αντισφαίρισης. Η διαθεσιμότητα αυτού του συνόλου δεδομένων, μαζί με τις τακτικές ενημερώσεις του, παρέχει στους ερευνητές έναν πολύτιμο πόρο για τη διεξαγωγή έρευνας, την απόκτηση γνώσεων και τη συμβολή στη γνώση και την κατανόηση των αθλητικών πτυχών των αγώνων αντισφαίρισης.

Σε αυτή τη μελέτη, η ερευνητική προσέγγιση βασίζεται στη χρήση δεδομένων τα οποία έχουν ληφθεί από την Ένωση Επαγγελματιών Αντισφαίρισης (Association of Tennis Professionals, ATP) για αγώνες απλού ανδρών. Η Ένωση Επαγγελματιών Αντισφαίρισης είναι ένας διάσημος οργανισμός που διέπει το άθλημα της επαγγελματικής αντισφαίρισης ανδρών παγκοσμίως. Το σύνολο δεδομένων που ελήφθη από την ATP χρησιμεύει ως πολύτιμη πηγή πληροφοριών για την ανάλυση και τη διερεύνηση διαφόρων πτυχών των αγώνων ατομικού ανδρών, συμπεριλαμβανομένων των επιδόσεων των παικτών, των αποτελεσμάτων των αγώνων και άλλων σχετικών παραγόντων. Ακόμη, το σύνολο δεδομένων συμπληρώνεται με δεδομένα από τα τουρνουά Grand Slam που διοργανώνει η Διεθνής Ομοσπονδία Αντισφαίρισης (International Tennis Federation, ITF) [28]. Τα τουρνουά Grand Slam, όπως το Australian Open, το French Open, το Wimbledon και το US Open, είναι γεγονότα κύρους που προσελκύουν κορυφαίους παίκτες και δημιουργούν πληθώρα δεδομένων.

Για να ενσωματωθούν χαρακτηριστικά τόσο του παίκτη όσο και του αγώνα από διαφορετικές πηγές, δύο σύνολα δεδομένων συγχωνεύτηκαν στην συγκεκριμένη μελέτη. Το πρώτο σύνολο δεδομένων, που ελήφθη από το αποθετήριο του JeffSackmann, παρέχει πολύτιμες πληροφορίες σχετικά με τα χαρακτηριστικά του παίκτη και του αγώνα, τα οποία όπως αναφέρθηκε παραπάνω λαμβάνονται από τους οργανισμούς ATP και ITF. Το δεύτερο σύνολο

δεδομένων, που ελήφθη από το tennis-data.co.uk, περιλαμβάνει δεδομένα απόδοσης που εμπλουτίζουν περαιτέρω την ανάλυση.

Με τη συγχώνευση αυτών των συνόλων δεδομένων, δημιουργείται ένα ολοκληρωμένο σύνολο δεδομένων, αποτελούμενο από 31.539 αγώνες αντισφαίρισης. Αυτό το σύνολο δεδομένων χρησιμεύει ως βάση για την εκπαίδευση και την αξιολόγηση των μοντέλων πρόβλεψης. Για να εξασφαλιστεί η ευρωστία και η ακρίβεια, το σύνολο δεδομένων χωρίζεται σε τρία υποσύνολα: ένα σύνολο εκπαίδευσης (training set), ένα σύνολο επικύρωσης (validation set) και ένα σύνολο δοκιμών (test set). Το σύνολο εκπαίδευσης περιέχει 22.898 αγώνες, παρέχοντας ένα μεγάλο και ποικίλο σύνολο δεδομένων για την εκπαίδευση των μοντέλων. Το σύνολο επικύρωσης περιλαμβάνει 3.686 αγώνες, επιτρέποντας τη λεπτομέρεια και την αξιολόγηση της απόδοσης των μοντέλων κατά τη διαδικασία ανάπτυξης. Τέλος, το σύνολο δοκιμών αποτελείται από 4.955 αγώνες και χρησιμεύει ως ανεξάρτητο σύνολο για την αξιολόγηση των προγνωστικών ικανοτήτων των μοντέλων.

Ο διαχωρισμός του συνόλου δεδομένων πραγματοποιήθηκε με βάση συγκεκριμένες χρονικές περιόδους. Το σύνολο εκπαίδευσης περιλαμβάνει αγώνες από το 2010 έως το 2018, το σύνολο επικύρωσης περιλαμβάνει αγώνες από το 2019 έως το 2020 και το σύνολο δοκιμών περιλαμβάνει αγώνες από το 2021 έως το 2022. Αυτός ο χωρισμός βάσει χρόνου επέτρεψε την αξιολόγηση της απόδοσης των μοντέλων σε διαφορετικές περιόδους, παρέχοντας πληροφορίες στη δυνατότητα γενίκευσής τους και στις προγνωστικές τους ικανότητες με την πάροδο του χρόνου.

Χρησιμοποιώντας αυτό το προσεκτικά διαχωρισμένο σύνολο δεδομένων, τα μοντέλα πρόβλεψης μπορούν να εκπαιδευτούν, να επικυρωθούν και να δοκιμαστούν σε διαφορετικά χρονικά πλαίσια, διασφαλίζοντας μια ολοκληρωμένη αξιολόγηση της απόδοσής τους στην πρόβλεψη των αποτελεσμάτων των αγώνων αντισφαίρισης.

4.2 Καθαρισμός Δεδομένων

Ο καθαρισμός δεδομένων είναι ένα κρίσιμο βήμα για την προετοιμασία ενός συνόλου δεδομένων για ανάλυση, ειδικά όταν χρησιμοποιούνται αλγόριθμοι Μηχανικής Μάθησης. Τα ακατέργαστα σύνολα δεδομένων συχνά περιέχουν ασυνέπειες, τιμές που λείπουν, ακραίες τιμές και άλλες ατέλειες που μπορούν να επηρεάσουν αρνητικά την απόδοση και την

ακρίβεια των μοντέλων. Με την εφαρμογή κατάλληλων τεχνικών προεπεξεργασίας, όπως ο χειρισμός τιμών που λείπουν, η ανίχνευση ακραίων τιμών και η κανονικοποίηση των δεδομένων, το σύνολο δεδομένων μπορεί να καθαριστεί και να τυποποιηθεί, οδηγώντας σε βελτιωμένη ακρίβεια και γενίκευση των μοντέλων.

Στο πλαίσιο αυτής της μελέτης, είναι σημαντικό να αντιμετωπιστεί το ζήτημα των τιμών που λείπουν στο σύνολο δεδομένων. Ανάλογα με τη φύση των μεταβλητών (αριθμητικές ή κατηγορικές), χρησιμοποιούνται διαφορετικές στρατηγικές για τον χειρισμό τιμών που λείπουν. Για τις αριθμητικές μεταβλητές, οι τιμές που λείπουν καταλογίζονται χρησιμοποιώντας τον μέσο όρο που υπολογίζεται από τα διαθέσιμα δεδομένα, ενώ για τις κατηγορικές μεταβλητές, οι τιμές που λείπουν αντικαθίστανται με την τιμή που εμφανίζεται πιο συχνά στο σύνολο δεδομένων. Χρησιμοποιώντας αυτές τις τεχνικές καταλογισμού, καταβάλλονται προσπάθειες για τον μετριασμό των επιπτώσεων των ελλιπών δεδομένων και τη διασφάλιση της ακεραιότητας και της πληρότητας της ανάλυσης. Η λογιστική για τις πληροφορίες που λείπουν είναι ζωτικής σημασίας για την αποφυγή μεροληπτικών ερμηνειών και τη διατήρηση της συνολικής ποιότητας και αξιοπιστίας των ευρημάτων της μελέτης.

Για τη δημιουργία αξιόπιστων προφίλ παικτών για κάθε αγώνα, χρησιμοποιούνται οι βαθμολογίες ATP τη στιγμή του αγώνα. Αυτές οι βαθμολογίες χρησιμεύουν ως δείκτης της σχετικής θέσης των παικτών και βοηθούν στη διάκριση μεταξύ του ευνοούμενου παίκτη και του αουτσάιντερ. Αν και η κατάταξη ATP δεν είναι ο μόνος καθοριστικός παράγοντας για την ανάλυση των αποτελεσμάτων του αγώνα, η ενσωμάτωσή της επιτρέπει τη συμπερίληψη μιας κρίσιμης μεταβλητής που επηρεάζει την πιθανότητα νίκης.

Με την τήρηση αυτής της σύμβασης, η μελέτη μπορεί να αναπτύξει ένα Μοντέλο Βασικής γραμμής, όπου ο ευνοούμενος παίκτης αναδεικνύεται σταθερά ως ο νικητής. Αυτή η προσέγγιση επιτρέπει τον υπολογισμό της πιθανότητας ο ευνοούμενος παίκτης να κερδίσει τον αγώνα και διευκολύνει την εξέταση της κατανομής πιθανοτήτων των αποτελεσμάτων του αγώνα, ιδιαίτερα για τον ευνοούμενο παίκτη. Μια τέτοια ανάλυση συμβάλλει σε μια πιο ολοκληρωμένη κατανόηση της δυναμικής και των παραγόντων που επηρεάζουν τα αποτελέσματα των αγώνων.

Για να εξασφαλιστεί η καταλληλότητα και η αξιοπιστία του συνόλου δεδομένων για ανάλυση, εκτελούνται διάφορα βήματα προεπεξεργασίας. Πρώτον, επιλέγεται ένα συγκεκριμένο εύρος ημερομηνιών για την εστίαση της ανάλυσης εντός μιας καθορισμένης

περιόδου, προωθώντας τη συνέπεια και τη συνοχή στο σύνολο δεδομένων. Στη συνέχεια, τα διπλά αρχεία εξαλείφονται για να αποφευχθεί ο πλεονασμός και να διατηρηθεί η ακεραιότητα των δεδομένων. Επιπλέον, οι αγώνες που σχετίζονται με το Davis Cup και τους Ολυμπιακούς Αγώνες εξαιρούνται από το σύνολο δεδομένων. Αυτοί οι αγώνες διαφέρουν σημαντικά από τα μεμονωμένα τουρνουά λόγω της ομαδικής πτυχής και των στρατηγικών που εμπλέκονται, τα οποία θα μπορούσαν να εισαγάγουν μεροληψία και να επηρεάσουν την αποτελεσματικότητα των μοντέλων πρόβλεψης. Οι ημιτελείς αγώνες, που προκύπτουν από αναστολές ή τραυματισμούς, αφαιρούνται επίσης καθώς ενδέχεται να μην αντικατοπτρίζουν με ακρίβεια την απόδοση του επηρεαζόμενου παίκτη, καθιστώντας τους ακατάλληλους για προπονητικά μοντέλα πρόβλεψης. Ομοίως, οι αγώνες που κατηγοριοποιούνται ως ”Walkovers”, όπου ένας παίκτης κερδίζει από προεπιλογή λόγω αποχώρησης ή απουσίας του αντιπάλου, αποκλείονται καθώς δεν παρέχουν πολύτιμα δεδομένα προπόνησης. Αυτά τα βήματα προεπεξεργασίας βελτιώνουν το σύνολο δεδομένων εξαλείφοντας άσχετες και δυνητικά παραπλανητικές αντιστοιχίσεις, βελτιώνοντας την ποιότητα και την αξιοπιστία των επόμενων αναλύσεων και προβλέψεων.

4.3 Αναπαράσταση Χαρακτηριστικών

Η αναπαράσταση χαρακτηριστικών είναι μια κρίσιμη πτυχή για την ακριβή πρόβλεψη των αποτελεσμάτων του αγώνα αντισφαίρισης. Επιλέγοντας προσεκτικά και εξάγοντας σχετικές λειτουργίες, όπως η κατάταξη παικτών, οι μετρήσεις απόδοσης και ο τύπος επιφάνειας, μπορούμε να καταγράψουμε σημαντικούς παράγοντες που επηρεάζουν τα αποτελέσματα του αγώνα. Τεχνικές προεπεξεργασίας όπως η κανονικοποίηση ή η κλιμάκωση χαρακτηριστικών μπορούν να εφαρμοστούν για να διασφαλιστεί η συμβατότητα και η βέλτιστη χρήση αυτών των χαρακτηριστικών στα μοντέλα Μηχανικής Μάθησης.

Για την εκπαίδευση ενός εποπτευόμενου αλγόριθμου Μηχανικής Μάθησης, απαιτείται ένα σύνολο δεδομένων με ετικέτα. Στο πλαίσιο της πρόβλεψης αγώνα αντισφαίρισης, κάθε παράδειγμα προπόνησης αποτελείται από ένα διάνυσμα χαρακτηριστικών εισαγωγής (X) που περιέχει διάφορα χαρακτηριστικά που σχετίζονται με τον παίκτη και τον αγώνα και μια αντίστοιχη τιμή στόχου (y) που υποδεικνύει το αποτέλεσμα του αγώνα.

Το διάνυσμα χαρακτηριστικών εισόδου περιλαμβάνει ένα ευρύ φάσμα πληροφοριών, συμπεριλαμβανομένων των χαρακτηριστικών παικτών, συγκεκριμένων λεπτομερειών αγώνα και άλλων σχετικών παραγόντων. Αυτά τα χαρακτηριστικά χρησιμεύουν ως βάση για τον

αλγόριθμο για να αποκαλύψει μοτίβα και σχέσεις μέσα στα δεδομένα, επιτρέποντας πιο ακριβείς προβλέψεις. Η τιμή στόχος αντιπροσωπεύει το αποτέλεσμα του αγώνα, με τις ήττες να σημειώνονται ως 0 και τις νίκες ως 1. Είναι σημαντικό να σημειωθεί ότι οι ημιτελείς αγώνες εξαιρούνται από τα δεδομένα εκπαίδευσης, με αποτέλεσμα μια δυαδική αναπαράσταση του αποτελέσματος.

Μόλις ο εποπτευόμενος αλγόριθμος μάθησης εκπαιδευτεί αποτελεσματικά χρησιμοποιώντας τα επισημασμένα παραδείγματα, δημιουργείται ένα μοντέλο πρόβλεψης. Αυτό το μοντέλο μπορεί στη συνέχεια να χρησιμοποιηθεί για την πρόβλεψη της έκβασης μελλοντικών αγώνων αντισφαίρισης με βάση τα χαρακτηριστικά εισόδου που παρέχονται. Αξιοποιώντας το εκπαιδευμένο μοντέλο, οι ενδιαφερόμενοι και οι λάτρεις του αθλήματος μπορούν να αποκτήσουν πολύτιμες γνώσεις και εκτιμήσεις πιθανοτήτων για επερχόμενους αγώνες. Μέσω της ενσωμάτωσης επισημασμένων παραδειγμάτων εκπαίδευσης και του προπονημένου μοντέλου πρόβλεψης, στόχος μας είναι να βελτιώσουμε την κατανόησή μας για τη δυναμική των αγώνων αντισφαίρισης και να επιτρέψουμε ακριβείς προβλέψεις σε αυτόν τον τομέα.

4.4 Εξισορρόπηση Συνόλου Δεδομένων

Η εξισορρόπηση του συνόλου δεδομένων είναι ένα κρίσιμο βήμα για την αποφυγή μεροληψίας και τη διασφάλιση δίκαιης εκπροσώπησης διαφορετικών κατηγοριών, όπως το αποτέλεσμα αγώνων (π.χ. νίκη ή ήττα). Τα μη ισορροπημένα σύνολα δεδομένων, όπου μια κατηγορία είναι σημαντικά πιο διαδεδομένη από άλλες, μπορεί να οδηγήσει σε ασύμμετρες προβλέψεις και ανακριβή αποτελέσματα. Για την αντιμετώπιση αυτού του ζητήματος, μπορούν να χρησιμοποιηθούν διάφορες τεχνικές, συμπεριλαμβανομένης της υπερδειγματοληψίας, της υποδειγματοληψίας ή της δημιουργίας σύνθετων δεδομένων, οι οποίες στοχεύουν στην επίτευξη ισορροπημένης κατανομής κλάσεων εντός του συνόλου δεδομένων.

Στο πλαίσιο της πρόβλεψης αγώνα αντισφαίρισης, δεδομένου ότι η πιθανότητα κάθε παίκτης να κερδίσει ή να χάσει αναμένεται να είναι περίπου 0,5 σε έναν τυχαίο αγώνα, η επίτευξη ισορροπίας περιλαμβάνει την τυχαία επιλογή παραδειγμάτων και την ανάθεσή τους είτε ως νίκες (1) είτε ως ήττες (0). Αυτή η τυχαία επιλογή εξασφαλίζει ίση αντιπροσώπευση θετικών και αρνητικών αποτελεσμάτων κατά τη διάρκεια της εκπαιδευτικής διαδικασίας, επιτρέποντας στο μοντέλο να μάθει από μια ποικιλία παραδειγμάτων [Πιν.4.1].

Επιπλέον, για να διατηρηθεί η συνέπεια, τα αντίστοιχα διανύσματα χαρακτηριστικών μεταξύ των παικτών μπορούν να εναλλάσσονται εάν είναι απαραίτητο. Αυτό διασφαλίζει ότι τα χαρακτηριστικά που σχετίζονται με τη νίκη ή την ήττα ενός παίκτη είναι σωστά ευθυγραμμισμένα, ενισχύοντας την ακεραιότητα και τη συνέπεια του συνόλου δεδομένων. Εξισορροπώντας το σύνολο δεδομένων, οι ερευνητές μπορούν να μετριάσουν τον αντίκτυπο της ανισορροπίας τάξης και να βελτιώσουν την ακρίβεια των προβλέψεων, οδηγώντας σε πιο αξιόπιστες και αμερόληπτες γνώσεις στον τομέα της πρόβλεψης αγώνων αντισφαίρισης.

4.5 Συμμετρική Αναπαράσταση Χαρακτηριστικών

Για να αναπτυχθεί ένα ισχυρό και ακριβές μοντέλο πρόβλεψης αγώνων αντισφαίρισης, είναι απαραίτητο να ληφθούν υπόψη τα μοναδικά χαρακτηριστικά των παικτών που συμμετέχουν σε έναν αγώνα. Αυτό απαιτεί τη συμπερίληψη δύο τιμών για κάθε μεταβλητή που σχετίζεται με τους παίκτες. Μια ευρέως χρησιμοποιούμενη προσέγγιση είναι ο υπολογισμός της διαφοράς μεταξύ αυτών των τιμών, δημιουργώντας έτσι ένα χαρακτηριστικό που ενσωματώνει τη σχετική ανισότητα μεταξύ των παικτών.

Για παράδειγμα, εάν η κατάταξη ATP επιλεγεί ως χαρακτηριστικό κλειδί, η διαφορά κατάταξης μπορεί να υπολογιστεί ως RANK = RANK1 − RANK2. Αυτή η μεθοδολογία, όπως καταδεικνύεται από τους Clarke και Dyte (2000) [29] στο μοντέλο Λογιστικής Παλινδρόμησης, θεωρεί τη διαφορά στην κατάταξη ως κρίσιμο προγνωστικό παράγοντα. Εναλλακτικά, θα μπορούσε κανείς να ενσωματώσει και τις δύο τιμές μιας μεταβλητής ως διακριτά χαρακτηριστικά. Ωστόσο, έρευνα του O'Malley [30] έδειξε ότι η χρήση της διαφοράς στις μεταβλητές μεταξύ των παικτών συχνά αρκεί για την επίτευξη υψηλής ακρίβειας πρόβλεψης. Στην πραγματικότητα, σε ένα ιεραρχικό μοντέλο, ο O'Malley διαπίστωσε ότι η διαφορά στις πιθανότητες νίκης μεταξύ των παικτών είχε τη μέγιστη επιρροή, καθιστώντας έτσι τις μεμονωμένες πιθανότητες ασήμαντες.

Εναλλακτικά, μια άλλη ευρέως χρησιμοποιούμενη προσέγγιση είναι ο υπολογισμός της αναλογίας μεταξύ των τιμών, όπως για παράδειγμα RANK = RANK1 / RANK2, η οποία καταγράφει τη σχετική αναλογία ή δύναμη μεταξύ των παικτών. Αυτό το χαρακτηριστικό που βασίζεται στην αναλογία επιτρέπει μια διαφορετική οπτική γωνία για τα χαρακτηριστικά των παικτών και τον πιθανό αντίκτυπό τους στα αποτελέσματα των αγώνων. Λαμβάνοντας υπόψη

την αναλογία των τιμών των χαρακτηριστικών, το μοντέλο μπορεί να αναγνωρίσει καταστάσεις όπου το χαρακτηριστικό ενός παίκτη είναι σημαντικά υψηλότερο ή χαμηλότερο από τον άλλο, παρέχοντας πολύτιμες πληροφορίες σχετικά με το σχετικό πλεονέκτημα ή μειονέκτημα κάθε παίκτη. Αυτή η αναπαράσταση που βασίζεται σε αναλογίες προσφέρει μια ξεχωριστή προοπτική για τα χαρακτηριστικά των παικτών και μπορεί να διερευνηθεί παράλληλα με άλλα χαρακτηριστικά για να ενισχύσει την προγνωστική ακρίβεια και την ερμηνευτικότητα του μοντέλου πρόβλεψης αγώνων αντισφαίρισης.

Η επιβολή της συμμετρίας στο μοντέλο είναι επωφελής για τη διασφάλιση συνεπών και αμερόληπτων προβλέψεων, ανεξάρτητα από την επισήμανση των παικτών. Χρησιμοποιώντας τις διαφορές στις μεταβλητές ως χαρακτηριστικά εισόδου, επιτυγχάνεται ένα συμμετρικό μοντέλο. Αντίθετα, ένα μη συμμετρικό μοντέλο μπορεί να αποδίδει διάφορους βαθμούς σημασίας σε χαρακτηριστικά για διαφορετικούς παίκτες, οδηγώντας σε ανόμοιες προβλέψεις με βάση την επισήμανση των παικτών. Για παράδειγμα, ένα μοντέλο Λογιστικής Παλινδρόμησης μπορεί να αποδώσει μεγαλύτερη βαρύτητα στην κατάταξη ενός παίκτη σε σύγκριση με εκείνη ενός άλλου. Με την υιοθέτηση ενός μόνου χαρακτηριστικού που καταγράφει τη διαφορά στην κατάταξη των παικτών, οι πιθανές προκαταλήψεις μπορούν να μετριαστούν και ο χώρος των χαρακτηριστικών μπορεί να μειωθεί αποτελεσματικά στο μισό. Αυτή η μείωση στη διάσταση των χαρακτηριστικών χρησιμεύει για να μετριάσει τη διακύμανση του μοντέλου, να μειώσει τον κίνδυνο υπερβολικής προσαρμογής και να αυξήσει τη γενίκευση του προγνωστικού μοντέλου.

Αγκαλιάζοντας αυτήν τη μεθοδολογία, η οποία ενσωματώνει διαφορές στα χαρακτηριστικά των παικτών ως στοιχεία εισαγωγής, είναι δυνατό να προωθηθεί ένα συμμετρικό και αμερόληπτο μοντέλο πρόβλεψης αγώνων αντισφαίρισης που παράγει αξιόπιστα και συνεπή αποτελέσματα αγώνα ανεξάρτητα από την επισήμανση των παικτών. Επιπλέον, αυτή η προσέγγιση συμβάλλει στη λιτότητα του μοντέλου, ενισχύοντας την ερμηνευτικότητα και μειώνοντας την πιθανότητα υπερβολικής προσαρμογής, ενισχύοντας έτσι την ακρίβεια και την αξιοπιστία του μοντέλου πρόβλεψης.

5 *Εξαγωγή Χαρακτηριστικών*

Τα τελευταία χρόνια, η εφαρμογή τεχνικών Μηχανικής Μάθησης σε διάφορους τομείς έχει επιδείξει αξιοσημείωτη επιτυχία, συμπεριλαμβανομένου του τομέα της αθλητικής ανάλυσης. Στο τένις, η ακριβής πρόβλεψη των αποτελεσμάτων των αγώνων έχει γίνει ένας ολοένα και πιο επιθυμητός στόχος. Ένας κρίσιμος παράγοντας που επηρεάζει σε μεγάλο βαθμό την ακρίβεια και την αξιοπιστία των μοντέλων πρόβλεψης σε αυτόν τον τομέα είναι η Εξαγωγή Χαρακτηριστικών. Η Εξαγωγή Χαρακτηριστικών παίζει καθοριστικό ρόλο στη μετατροπή των ακατέργαστων δεδομένων σε μια ουσιαστική αναπαράσταση που συλλαμβάνει τις σχετικές πληροφορίες που απαιτούνται για ακριβείς προβλέψεις. Στο πλαίσιο της πρόβλεψης αποτελεσμάτων αγώνων αντισφαίρισης, δίνει τη δυνατότητα στους ερευνητές να αποκαλύψουν τα υποκείμενα μοτίβα και τα χαρακτηριστικά που επηρεάζουν τα αποτελέσματα των αγώνων.

Επιλέγοντας προσεκτικά και εξάγοντας τα πιο σχετικά χαρακτηριστικά από τα διαθέσιμα δεδομένα, οι ερευνητές μπορούν να συλλάβουν αποτελεσματικά τους βασικού παράγοντες που συμβάλλουν στο τελικό αποτέλεσμα ενός αγώνα αντισφαίρισης. Αυτά τα χαρακτηριστικά μπορεί να περιλαμβάνουν διάφορες πτυχές, όπως κατάταξη παικτών, ιστορικές επιδόσεις, στυλ παιχνιδιού και επιφάνεια γηπέδου. Κάθε ένα από αυτά τα χαρακτηριστικά περιέχει πολύτιμες πληροφορίες και προγνωστικές δυνατότητες, επιτρέποντας την ανάπτυξη ισχυρών μοντέλων, ικανών να κάνουν ακριβείς προβλέψεις για τα αποτελέσματα του αγώνα.

Η Εξαγωγή Χαρακτηριστικών χρησιμεύει επίσης ως μέσο μείωσης διαστάσεων. Αντιμετωπίζει την πρόκληση των δεδομένων υψηλών διαστάσεων, κοινώς γνωστή ως κατάρα της διάστασης (curse of dimensionality), εντοπίζοντας τα πιο ενημερωτικά χαρακτηριστικά. Εξάγοντας πιο ουσιαστικά χαρακτηριστικά, οι ερευνητές μπορούν να μετριάσουν την πολυπλοκότητα και τις υπολογιστικές απαιτήσεις που σχετίζονται με τη μοντελοποίηση, οδηγώντας σε πιο αποτελεσματικά και κλιμακούμενα μοντέλα πρόβλεψης. Συνεπώς, η διαδικασία αυτή ενισχύει την ερμηνευσιμότητα των μοντέλων πρόβλεψης, παρέχοντας πολύτιμες πληροφορίες σχετικά με τους παράγοντες που επηρεάζουν τα αποτελέσματα των αγώνων και υποστηρίζοντας τις διαδικασίες λήψης στρατηγικών αποφάσεων στον τομέα των αναλυτικών στοιχείων της αντισφαίρισης.

Για να μπορέσουν οι αλγόριθμοι Μηχανικής Μάθησης να μάθουν τη σχέση μεταξύ εισόδων και επιθυμητών εξόδων, είναι απαραίτητη η ύπαρξη ενός συνόλου εκπαίδευσης. Αυτό σημαίνει ότι το αρχικό σύνολο ακατέργαστων δεδομένων πρέπει να μετατραπεί σε ένα σύνολο ενημερωτικών χαρακτηριστικών, όπου κάθε εγγραφή δεδομένων αντιπροσωπεύει έναν αγώνα αντισφαίρισης με την πάροδο του χρόνου, μαζί με την επιθυμητή τιμή στόχο, δηλαδή μία νίκη ή μία ήττα. Πριν από την εξαγωγή των χαρακτηριστικών, είναι σημαντικό το ακατέργαστο σύνολο δεδομένων να ταξινομηθεί χρονολογικά κατά ημερομηνία. Αυτό διασφαλίζει ότι κάθε εξαγωγέας χαρακτηριστικών χρησιμοποιεί δεδομένα μόνο από το παρελθόν, αποφεύγοντας τη διαρροή δεδομένων που θα μπορούσε να προκύψει εάν συμπεριληφθούν πληροφορίες από μελλοντικούς αγώνες.

Είναι σημαντικό να σημειωθεί ότι κατά την εξαγωγή χαρακτηριστικών, σε αυτή τη μελέτη χρησιμοποιείται μια προσέγγιση μέσου όρου που λαμβάνει υπόψη όλα τα προηγούμενα παιχνίδια εντός περιόδου ενός έτους. Αυτή η προσέγγιση παρέχει εκτιμήσεις που αντικατοπτρίζουν την απόδοση ενός παίκτη με μεγαλύτερη ακρίβεια από ό,τι αν λαμβάνονταν υπόψη μόνο τα πιο πρόσφατα παιχνίδια ή το σύνολο της καριέρας του.

Τα εξαγόμενα χαρακτηριστικά μπορούν να κατηγοριοποιηθούν σε δύο κύριες ομάδες: Χαρακτηριστικά περιβάλλοντος και Χαρακτηριστικά παίκτη. Τα χαρακτηριστικά περιβάλλοντος είναι πανομοιότυπα και για τους δύο παίκτες σε ένα παιχνίδι και περιλαμβάνουν χαρακτηριστικά, όπως το επίπεδο του τουρνουά (π.χ. ATP 250, ATP 500, Grand Slam), την τοποθεσία του τουρνουά, την επιφάνεια του γηπέδου (π.χ. γρασίδι, χώμα, σκληρό γήπεδο) και τον συγκεκριμένο γύρο του αγώνα (π.χ. πρώτος γύρος, προημιτελικοί). Κατά συνέπεια, αυτά τα χαρακτηριστικά μπορούν να αντιπροσωπεύονται από μία μόνο τιμή ανά παιχνίδι.

Από την άλλη πλευρά, τα χαρακτηριστικά του παίκτη, όπως η ηλικία και η κατάταξη, διαφέρουν μεταξύ των παικτών και απαιτούν ξεχωριστές αναπαραστάσεις για κάθε παίκτη. Επομένως, δημιουργούνται δύο διακριτά σύνολα αυτών των χαρακτηριστικών, ένα για τον παίκτη Α και ένα άλλο για τον παίκτη Β, για να μοντελοποιήσουν με ακρίβεια τα αντίστοιχα χαρακτηριστικά τους. Τέτοια χαρακτηριστικά μπορεί επίσης να είναι το προτιμώμενο χέρι του παίκτη (αριστερόχειρας ή δεξιόχειρας), η ημερομηνία γέννησης και η χώρα προέλευσης. Αυτά τα ειδικά χαρακτηριστικά του παίκτη συμβάλλουν στην κατανόηση του στυλ, της εμπειρίας και του υπόβαθρου του κάθε παίκτη, τα οποία μπορούν να επηρεάσουν τα αποτελέσματα του αγώνα.

5.1 Νίκες και Ήττες

Ενα από τα θεμελιώδη χαρακτηριστικά που χρησιμοποιούνται για την αναπαράσταση ενός παίκτη είναι το ποσοστό νίκης-ήττας (win-loss record). Ωστόσο, η αναπαράσταση αυτού ως απλού ποσοστού μπορεί να μην αποτυπώσει αποτελεσματικά την πραγματική δύναμη ενός παίκτη, ειδικά όταν ο παίκτης που έχει κερδίσει τον μοναδικό του αγώνα θα είχε ποσοστό νίκης 100%, παρόλο που μπορεί να μην αντικατοπτρίζει με ακρίβεια τη συνολική του ικανότητα σε σύγκριση με κάποιον με ποσοστό νίκης 80% σε μεγαλύτερο μέγεθος δείγματος. Παρομοίως, ένας παίκτης που έχασε τον πρώτο του αγώνα μπορεί να μην είναι απαραίτητα τόσο αδύναμος όσο υπονοεί το ποσοστό νίκης-ήττας.

Για να αντιμετωπιστεί αυτό, ο απόλυτος αριθμός νικών και ήττων χρησιμοποιούνται ως χαρακτηριστικά, παρέχοντας μια πιο ενημερωτική αναπαράσταση. Ωστόσο, δεν έχουν όλες οι νίκες και οι ήττες την ίδια σημασία, καθώς παράγοντες όπως το επίπεδο του τουρνουά και η δύναμη του αντιπάλου παίζουν επίσης ρόλο. Επιπλέον, οι ικανότητες των παικτών μπορούν να εξελιχθούν με την πάροδο του χρόνου, απαιτώντας την εξέταση διαφορετικών χρονικών περιόδων. Για την αντιμετώπιση αυτών των προκλήσεων, οι αγώνες κατηγοριοποιούνται με βάση το τελευταίο εξάμηνο, έτος και καριέρα του παίκτη (ξεκινώντας από το 2010). Αυτό έχει ως αποτέλεσμα τα ακόλουθα χαρακτηριστικά [Πίν. 5.1]: Νίκες, Ήττες και Αγώνες Καριέρας, Νίκες, Ήττες και Αγώνες για την περίοδο του τελευταίου χρόνου, Νίκες, Ήττες για την περίοδο του τελευταίου εξαμήνου.

Επιπλέον, δημιουργείται ένα πρόσθετο χαρακτηριστικό, το οποίο αντιπροσωπεύει το ποσοστό νικών των αγώνων του τελευταίου έτους. Αυτό προσφέρει ένα πιο πρόσφατο μέτρο της απόδοσης ενός παίκτη. Επιπλέον, ενσωματώνονται χαρακτηριστικά όπως οι Τίτλοι Καριέρας

(ο αριθμός των πρωταθλημάτων που κέρδισε ο παίκτης) και οι Τελικοί Καριέρας (ο αριθμός των τελικών που συμμετείχε ο παίκτης), παρέχοντας πληροφορίες για την επιτυχία ενός παίκτη σε αγώνες υψηλών προγνωστικών.

Χρησιμοποιώντας αυτά τα χαρακτηριστικά, μπορεί να αποτυπωθεί μια πιο ολοκληρωμένη αναπαράσταση της απόδοσης και των επιτευγμάτων ενός παίκτη, επιτρέποντας πιο ακριβείς προβλέψεις σε μοντέλα αποτελεσμάτων αγώνων αντισφαίρισης.

Χαρακτηριστικά
Wins Career – Συνολικές Νίκες Καριέρας
Ratio Wins Career – Συνολικές Νίκες Καριέρας (σε κλασματική αναπαράσταση)
Losses Career – Συνολικές Ήττες Καριέρας
Matches Career – Συνολικοί Αγώνες Καριέρας
Wins Year – Συνολικές Νίκες Του Τελευταίου Χρόνου
Losses Year – Συνολικές Ήττες του Τελευταίου Χρόνου
Matches Year – Συνολικοί Αγώνες του Τελευταίου Χρόνου
Wins Semester – Συνολικές Νίκες του Τελευταίου Εξαμήνου
Losses Semester – Συνολικές Ήττες του Τελευταίου Εξαμήνου
Titles Career – Συνολικοί Τίτλοι Καριέρας
Ratio Titles Career – Συνολικοί Τίτλοι Καριέρας (σε κλασματική αναπαράσταση)
Finals Career – Συνολικοί Τελικοί Καριέρας
Ratio Finals Career – Συνολικοί Τελικοί Καριέας (σε κλασματική αναπαράσταση)
Matches Percentage – Ποσοστό Νίκης σε αγώνες του Τελευταίου Χρόνου

Πίνακας 5.1 : Χαρακτηριστικά σχετικά με τις Νίκες και Ήττες των αντίπαλων παικτών σε χρονικές περιόδους όπως καριέρα, τελευταίος χρόνος και τελευταίο εξάμηνο

5.2 Βαθμολογία Κατάταξης

Για να ενσωματωθούν μη γραμμικές διαφορές στην ποιότητα του παίκτη, χρησιμοποιείται μια λογαριθμική κλίμακα για να εκφράσει τη διαφορά κατάταξης μεταξύ του φαβορί και του

αουτσάιντερ. Αυτός ο λογαριθμικός μετασχηματισμός είναι ιδιαίτερα σημαντικός για τους κορυφαίους παίκτες, καθώς παρέχει μια πιο ακριβή απεικόνιση της κρισιμότητας της διαφοράς κατάταξης [Πιν. 5.2]. Ομοίως, η διαφορά στους βαθμούς κατάταξης μεταξύ του φαβορί και του αουτσάιντερ αντιπροσωπεύεται επίσης σε λογαριθμική κλίμακα, προσφέροντας μια πιο λεπτή κατανόηση της διαφοράς δύναμης μεταξύ των παικτών σε σύγκριση με την εξέταση μόνο της κατάταξης.

Επιπλέον, εισάγεται ένα άλλο χαρακτηριστικό για να αντιπροσωπεύσει τη συνολική ποιότητα του αγώνα. Υπολογίζεται ως το άθροισμα των λογαριθμικών μετασχηματισμών της κατάταξης του παίκτη Α και του παίκτη Β. Λαμβάνοντας υπόψη τους λογαριθμικούς μετασχηματισμούς της κατάταξης και των δύο παικτών, η ποιότητα του αγώνα παρέχει ένα ολοκληρωμένο μέτρο της ανταγωνιστικής δύναμης και ισορροπίας μεταξύ των δύο αντιπάλων. Αυτό το χαρακτηριστικό συμβάλλει στην προγνωστική ακρίβεια του μοντέλου ενσωματώνοντας μια αξιολόγηση της ανταγωνιστικότητας του αγώνα ως σημαντικό παράγοντα για τον καθορισμό του αποτελέσματος.

Χαρακτηριστικά
Rank (log) – ATP Κατάταξη Παικτών σε λογαριθμική κλίμακα
Rank Points (log) – ATP Πόντοι Κατάταξης Παικτών σε λογαριθμική κλίμακα
Match Quality – Χαρακτηριστικό Ποιότητας Αγώνα

Πίνακας 5.2 : Χαρακτηριστικά σχετικά με την βαθμολογία κατάταξης

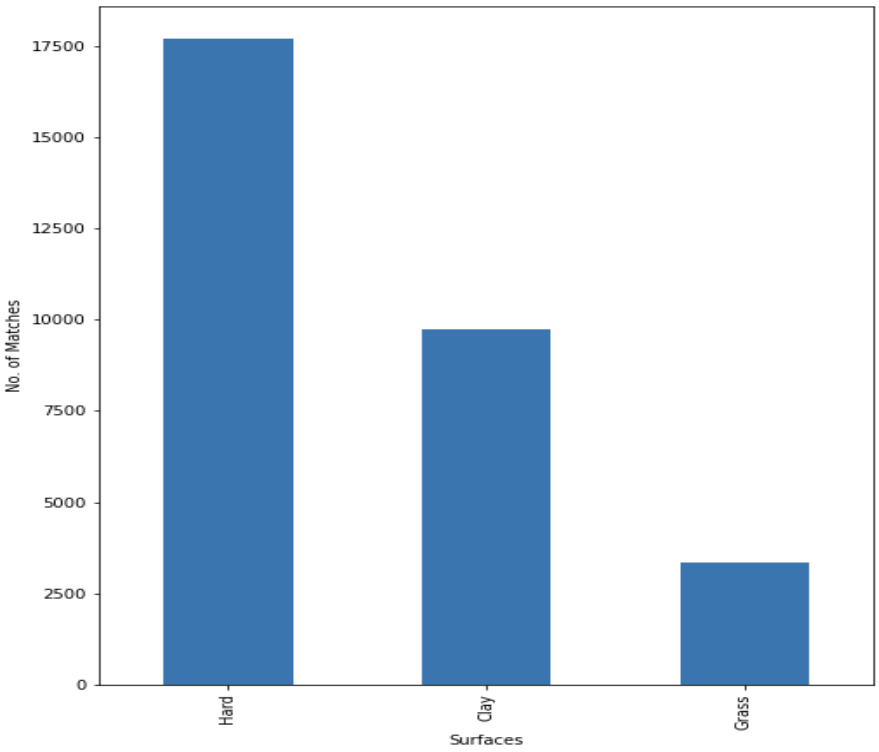
Με την ενσωμάτωση αυτών των χαρακτηριστικών, το μοντέλο πρόβλεψης μπορεί να καταγράψει τις μη γραμμικές διαφορές στην ποιότητα των παικτών και τη συνολική ποιότητα του αγώνα, επιτρέποντας πιο ακριβείς και διορατικές προβλέψεις των αποτελεσμάτων του αγώνα αντισφαίρισης.

5.3 Επιφάνεια Γηπέδου

Στο άθλημα της αντισφαίρισης, η επιφάνεια του γηπέδου έχει σημαντικό αντίκτυπο στην αναμενόμενη έκβαση ενός αγώνα, ανεξάρτητα από τα επίπεδα δεξιοτήτων των παικτών. Διαφορετικοί παίκτες τείνουν να υπερέχουν ή να αγωνίζονται σε συγκεκριμένες επιφάνειες

γηπέδων [Σχ. 5.1]. Για την αποτύπωση της διακύμανσης στην απόδοση του παίκτη σε όλες τις επιφάνειες, λαμβάνουμε υπόψη τον αριθμό των νικών και των ήττων για κάθε παίκτη σε μια συγκεκριμένη επιφάνεια.

Παρόμοια με την προηγούμενη αναπαράσταση, αντιπροσωπεύουμε αυτό το χαρακτηριστικό χρησιμοποιώντας τον απόλυτο αριθμό νικών και ήττων [Πίν. 5.3]. Ωστόσο, σε αντίθεση με άλλα χαρακτηριστικά, δεν είναι απαραίτητο να διαχωριστεί περαιτέρω αυτό το χαρακτηριστικό βάσει χρονικών περιόδων. Συνήθως, η προτίμηση και η απόδοση ενός παίκτη σε μια συγκεκριμένη επιφάνεια παραμένουν σχετικά σταθερές σε όλη την καριέρα του.



Σχήμα 5.1: Πλήθος αγώνων στις διαφορετικές επιφάνειες

Χαρακτηριστικά
Wins Surface – Συνολικές Νίκες σε επιφάνεια γηπέδου ίδια με τον αγώνα που εξετάζεται
Losses Surface – Συνολικές Ήττες σε επιφάνεια γηπέδου ίδια με τον αγώνα που εξετάζεται
Surface Percentage – Ποσοστό Νίκης σε αγώνες σε επιφάνεια γηπέδου ίδια με τον αγώνα που εξετάζεται

Πίνακας 5.3 : Χαρακτηριστικά σχετικά με την επιφάνεια γηπέδου

Αυτά τα χαρακτηριστικά παρέχουν πολύτιμες πληροφορίες σχετικά με το ιστορικό και το ποσοστό επιτυχίας ενός παίκτη σε διαφορετικές επιφάνειες, επιτρέποντας στο μοντέλο να λαμβάνει υπόψη τα πλεονεκτήματα και τις αδυναμίες της συγκεκριμένης επιφάνειας κάθε παίκτη. Ενσωματώνοντας τα χαρακτηριστικά που σχετίζονται με την επιφάνεια στο μοντέλο πρόβλεψης, παρέχεται στο μοντέλο η δυνατότητα να καταγράφει και να λαμβάνει υπόψη τον αντίκτυπο της αγωνιστικής επιφάνειας στα αποτελέσματα των αγώνων. Αυτή η συμπερίληψη ενισχύει την ακρίβεια και την αξιοπιστία των προβλέψεων του μοντέλου, λαμβάνοντας υπόψη τα πλεονεκτήματα και τις αδυναμίες κάθε παίκτη για συγκεκριμένη επιφάνεια.

5.4 Πλεονέκτημα Έδρας

Στο άθλημα της αντισφαίρισης, η έννοια του πλεονεκτήματος έδρας (home advantage) έχει παρατηρηθεί ότι έχει σημαντικό αντίκτυπο στα αποτελέσματα των αγώνων, ιδιαίτερα στο γύρο των ανδρών. Η έρευνα του Koning [31] έχει ρίξει φως σε αυτό το φαινόμενο, υποδεικνύοντας ότι οι παίκτες που αγωνίζονται στη χώρα τους έχουν ένα αξιοσημείωτο πλεονέκτημα έναντι των αντιπάλων τους. Για την ενσωμάτωση αυτού του χαρακτηριστικού στο σύνολο δεδομένων, εισάγεται μια δυαδική μεταβλητή, η οποία λαμβάνει την τιμή 1 εάν ο αγώνας παίζεται στην πατρίδα ενός παίκτη, υποδεικνύοντας την ύπαρξη πλεονεκτήματος έδρας, και 0 εάν ο αγώνας παίζεται αλλού. Συμπεριλαμβάνοντας αυτό το χαρακτηριστικό στο σύνολο δεδομένων [Πίν 5.4], στοχεύουμε να συλλάβουμε και να λάβουμε υπόψη την επιρροή του πλεονεκτήματος έδρας στις προβλέψεις αγώνων.

Χαρακτηριστικά
Home advantage – Πλεονέκτημα έδρας

Πίνακας 5.4 : Χαρακτηριστικό σχετικά με το πλεονέκτημα έδρας

5.5 Ιστορικότητα μεταξύ δύο παικτών

Το ιστορικό μεταξύ δύο παικτών είναι μια σημαντική πτυχή που πρέπει να λαμβάνεται υπόψη κατά την πρόβλεψη του αποτελέσματος των μελλοντικών αγώνων τους. Είναι προφανές ότι ένα συγκεκριμένο στυλ παιχνιδιού μπορεί να είναι πιο πλεονεκτικό σε σχέση με ένα άλλο στυλ παιχνιδιού. Μερικές φορές, ένας παίκτης μπορεί να παλέψει ενάντια σε έναν αντίπαλο που βρίσκεται κάτω από αυτόν, όπως το ρεκόρ 11-8 του Federer εναντίον του Nalbandian. Αυτό το χαρακτηριστικό θα μείωνε την προβλεπόμενη πιθανότητα να κερδίσει ο Federer αν αγωνιζόταν σήμερα. ΄Ενα άλλο παράδειγμα είναι το ρεκόρ 5-6 του Nadal απέναντι στον Davydenko. Το χαρακτηριστικό της ιστορικότητας [Πιν. 5.6] χρησιμοποιείται για να αναπαραστήσει τη σχέση μεταξύ των παικτών, η οποία υπολογίζεται ως η διαφορά μεταξύ του αριθμού των αγώνων που κέρδισε κάθε παίκτης σε κοινούς αγώνες.

Χαρακτηριστικά
Head to Head – Ιστορικό αγώνων μεταξύ των δύο παικτών

Πίνακας 5.5 : Χαρακτηριστικό σχετικά με το ιστορικό των αντιπαλόμενων παικτών

5.6 Τουρνουά

Στο πλαίσιο των τουρνουά αντισφαίρισης, τα προηγούμενα επιτεύγματα ενός παίκτη μπορεί να έχουν ψυχολογικό αντίκτυπο στην επακόλουθη απόδοσή του στο ίδιο τουρνουά. Για να ληφθεί υπόψη αυτή η επιρροή, γίνεται ενσωμάτωση των χαρακτηριστικών που αντικατοπτρίζουν το ιστορικό απόδοσης σε τουρνουά ενός παίκτη [Πιν. 5.6]. Αυτά τα χαρακτηριστικά περιλαμβάνουν τον συνολικό αριθμό νικών που έχει πετύχει ένας παίκτης στο τουρνουά, τον συνολικό αριθμό των ηττών και το ποσοστό νικών στο τουρνουά. Με την ενσωμάτωση αυτών των χαρακτηριστικών στο σύνολο δεδομένων, γίνεται εφικτή η καταγραφή των κινήτρων και των αποτελεσμάτων που σχετίζονται με την εμπιστοσύνη που

μπορεί να έχουν οι επιτυχίες ή οι αποτυχίες σε προηγούμενο τουρνουά στις επόμενες επιδόσεις ενός παίκτη εντός του ίδιου τουρνουά.

Χαρακτηριστικά
Wins Tour – Συνολικές Νίκες σε αγώνες τουρνουά ίδιου με τον αγώνα που εξετάζεται
Losses Tour – Συνολικές Ήττες σε αγώνες τουρνουά ίδιου με τον αγώνα που εξετάζεται
Tour Percentage – Ποσοστό Νίκης σε αγώνες τουρνουά ίδιου με τον αγώνα που εξετάζεται

Πίνακας 5.6 : Χαρακτηριστικά σχετικά με το τουρνουά

5.7 Πρόσφατοι Αγώνες

Στο άθλημα της αντισφαίρισης, όπως και σε άλλα αθλήματα, η συσσώρευση κούρασης μπορεί να έχει αρνητικές επιπτώσεις στην απόδοση ενός παίκτη. Επιστημονικές μελέτες [32] έχουν δείξει ότι το παρατεταμένο παιχνίδι αγώνων οδηγεί σε μείωση της λειτουργίας των σκελετικών μυών και αυτές οι αρνητικές επιπτώσεις είναι ιδιαίτερα έντονες όταν οι αγώνες παίζονται σε διαδοχικές ημέρες. Για την αντιμετώπιση αυτού του παράγοντα, ένα σημαντικό χαρακτηριστικό ενσωματώνεται στο μοντέλο πρόβλεψης, δηλαδή ο αριθμός των παιχνιδιών που έπαιξε κάθε παίκτης τις τελευταίες δύο εβδομάδες [Πιν. 5.7]. Αυτό το χαρακτηριστικό χρησιμεύει ως ένδειξη του πρόσφατου ψυχολογικού φορτίου και της κόπωσης του αγώνα.

Ωστόσο, είναι επίσης σημαντικό να ληφθούν υπόψη οι επιπτώσεις της αδράνειας στην απόδοση του παίκτη στην αντισφαίριση. Οι επιστημονικές μελέτες έχουν δείξει ότι μια μέτρια περίοδος ανάπαυσης και αποκατάστασης μπορεί να είναι επωφελής, ενισχύοντας δυνητικά τις πιθανότητες ενός παίκτη να κερδίσει στον επόμενο αγώνα του. Συγκεκριμένα, έως και τρεις εβδομάδες αδράνειας έχουν συσχετιστεί με βελτιωμένα αποτελέσματα απόδοσης. Πέρα από αυτό το όριο, ωστόσο, η παρατεταμένη αδράνεια μπορεί να οδηγήσει σε μειωμένη απόδοση. Για να αποτυπωθεί αυτή η πτυχή, ο αριθμός των ημερών από τον τελευταίο αγώνα που έπαιξε ένας παίκτης συμπεριλαμβάνεται ως χαρακτηριστικό. Το μοντέλο πρόβλεψης αποκτά έτσι την δυνατότητα να αξιολογεί την πιθανή επίδραση των περιόδων ανάπαυσης στην απόδοση του παίκτη και να προσαρμόζει τις προβλέψεις ανάλογα.

Επιπλέον, οι τάσεις απόδοσης των παικτών μπορούν να παίξουν ρόλο στην πρόβλεψη των αποτελεσμάτων του αγώνα. Οι παίκτες με νικηφόρα σερί είναι πιο πιθανό να συνεχίσουν να κερδίζουν, ενώ οι παίκτες σε σερί ήττες είναι πιο πιθανό να συνεχίσουν να βιώνουν ήττες. Για

να ληφθούν υπόψη αυτά τα σερί, προστίθενται στο μοντέλο, χαρακτηριστικά που αποτυπώνουν τις συνεχόμενες νίκες και τις συνεχόμενες ήττες, επιτρέποντας του να καταγράφει την ορμή και τους ψυχολογικούς παράγοντες που σχετίζονται με την πρόσφατη απόδοση ενός παίκτη.

Με την ενσωμάτωση αυτών των χαρακτηριστικών που σχετίζονται με το πρόσφατο ψυχολογικό φορτίο του αγώνα, την αδράνεια και τα σερί απόδοσης, το μοντέλο πρόβλεψης μπορεί να εξηγήσει τον αντίκτυπο της συσσωρευμένης κόπωσης, των περιόδων ανάπαυσης και των ψυχολογικών παραγόντων στην απόδοση ενός παίκτη. Αυτό επιτρέπει μια πιο ολοκληρωμένη και ακριβή αξιολόγηση των αποτελεσμάτων των αγώνων στην αντισφαίριση.

Χαρακτηριστικά
Matches Recent – Πλήθος αγώνων τις τελευταίες 15 ημέρες
Ratio Matches Recent – Πλήθος αγώνων τις τελευταίες 15 ημέρες (σε κλασματική αναπαράσταση)
Winning Streak – Σερί συνεχόμενων νικών
Losing Streak – Σερί συνεχόμενων ηττών
Days Inactive – Πλήθος ημερών αδράνειας

Πίνακας 5.7 : Χαρακτηριστικά σχετικά με τους πρόσφατους αγώνες και την ψυχολογική και σωματική κόπωση των παικτών

5.8 Σύστημα Κατάταξης

Κατά την αξιολόγηση της δύναμης ενός παίκτη, χρησιμοποιείται συνήθως το σύστημα κατάταξης ATP, με βάση τους βαθμούς κατάταξης που κέρδισαν σε τουρνουά κατά τη διάρκεια του προηγούμενου έτους. Ωστόσο, αυτό το σύστημα κατάταξης έχει ορισμένους περιορισμούς όταν πρόκειται για την αξιολόγηση της τρέχουσας κατάστασης ενός παίκτη και την εξέταση της ποιότητας των αντιπάλων που αντιμετωπίζει. Επιπλέον, το σύστημα επιτρέπει στους παίκτες να κερδίζουν μόνο πόντους χωρίς την πιθανότητα να τους χάσουν, κάτι που μπορεί να οδηγήσει σε διογκωμένες βαθμολογίες για παίκτες με μέτρια αποτελέσματα που αγωνίζονται συχνά.

Δεδομένων αυτών των μειονεκτημάτων, εναλλακτικά συστήματα αξιολόγησης όπως το Elo Rating System έχουν διερευνηθεί ως πιθανές βελτιώσεις. Το σύστημα αξιολόγησης Elo, που αναπτύχθηκε αρχικά για το σκάκι, έχει προσαρμοστεί για διάφορα αθλήματα, συμπεριλαμβανομένης της αντισφαίρισης. Σε αντίθεση με το σύστημα κατάταξης ATP, οι αξιολογήσεις Elo λαμβάνουν υπόψη την ποιότητα των αντιπάλων και επιτρέπουν προσαρμογές βαθμολογίας με βάση τα αποτελέσματα του αγώνα. Το αναμενόμενο αποτέλεσμα ενός αγώνα μπορεί να προσδιοριστεί λαμβάνοντας υπόψη τη διαφορά στις βαθμολογίες μεταξύ των δύο παικτών:

$$E_A = 1 / 1 + 10 (R_B - R_A / 400)$$

όπου

E_A *: το αναμενόμενο ποσοστό νίκης του Παίκτη Α*

R_i *: η βαθμολογία Elo του Παίκτη i.*

Λαμβάνοντας υπόψη τις αξιολογήσεις Elo ως χαρακτηριστικό στο μοντέλο πρόβλεψης, καθίσταται δυνατό να αποτυπωθεί η σχετική δύναμη και η προσαρμοστικότητα ενός παίκτη με βάση την απόδοσή του απέναντι σε διαφορετικούς αντιπάλους. Οι αξιολογήσεις Elo παρέχουν μια δυναμική και διαφοροποιημένη αναπαράσταση της ικανότητας του παίκτη, λαμβάνοντας υπόψη τόσο τις νίκες όσο και τις ήττες και λαμβάνοντας υπόψη τη δύναμη του αντίπαλου παίκτη. Αυτή η εναλλακτική προσέγγιση μπορεί ενδεχομένως να βελτιώσει την ακρίβεια και την αξιοπιστία του προγνωστικού μοντέλου, προσφέροντας μια πιο ολοκληρωμένη αξιολόγηση της τρέχουσας κατάστασης και της ανταγωνιστικής ικανότητας ενός παίκτη.

Διερευνώντας τη χρήση του Elo ως εναλλακτικού συστήματος αξιολόγησης, το μοντέλο πρόβλεψης μπορεί να ξεπεράσει ορισμένους από τους περιορισμούς που σχετίζονται με το σύστημα κατάταξης ATP, οδηγώντας τελικά σε βελτιωμένες αξιολογήσεις της δύναμης των παικτών και πιο ακριβείς προβλέψεις των αποτελεσμάτων του αγώνα.

Το σύστημα Elo υποθέτει ότι η απόδοση ενός παίκτη ακολουθεί μια κανονική κατανομή και το αναμενόμενο αποτέλεσμα ενός αγώνα μπορεί να εκτιμηθεί λαμβάνοντας υπόψη τη διαφορά στις αξιολογήσεις Elo μεταξύ των παικτών. Στο πλαίσιο της αντισφαίρισης, το σύστημα αξιολόγησης Elo αντιμετωπίζει τους περιορισμούς της κατάταξης ATP που συζητήθηκαν προηγουμένως. Μετά από κάθε αγώνα, ο παίκτης που κερδίζει κερδίζει πόντους Elo από τον παίκτη που χάνει, με τον αριθμό των πόντων που ανταλλάσσονται με βάση τη

διαφορά Elo πριν από τον αγώνα. Αυτός ο μηχανισμός προσαρμογής επιτρέπει μια πιο δυναμική αναπαράσταση της ικανότητας του παίκτη και υπολογίζει τη σχετική δύναμη των αντιπάλων.

Έρευνες έχουν δείξει ότι το σύστημα αξιολόγησης Elo [33] αποδίδει καλύτερα από την κατάταξη ATP στην πρόβλεψη των αποτελεσμάτων των αγώνων αντισφαίρισης. Παρέχει μια πιο ακριβή αξιολόγηση της απόδοσης του παίκτη και μπορεί να χρησιμοποιηθεί ως πολύτιμο χαρακτηριστικό στο μοντέλο πρόβλεψης. Το χαρακτηριστικό αξιολόγησης Elo [Πιν. 5.8] αντιπροσωπεύει τη βαθμολογία ενός παίκτη ως ενιαίο αριθμό, υποδεικνύοντας το σχετικό επίπεδο δεξιοτήτων του σε σύγκριση με άλλους παίκτες.

Χαρακτηριστικά
Elo Rating – Βαθμολογία Κατάταξης Elo
Ratio Elo Rating – Βαθμολογία Κατάταξης Elo (σε κλασματική αναπαράσταση)

Πίνακας 5.8 : Χαρακτηριστικά σχετικά με την βαθμολογία κατάταξης Elo

Ωστόσο, αξίζει να σημειωθεί ότι το σύστημα αξιολόγησης Elo έχει τους δικούς του περιορισμούς. Υποθέτει ότι η απόδοση του παίκτη ακολουθεί μια κανονική κατανομή και δεν λαμβάνει υπόψη τις διακυμάνσεις στη συνέπεια της απόδοσης ή τις τυπικές αποκλίσεις μεταξύ των παικτών. Η ενσωμάτωση της βαθμολογίας Elo ως χαρακτηριστικό μπορεί να βελτιώσει την προγνωστική ακρίβεια του μοντέλου και να παρέχει πολύτιμες πληροφορίες για τα σχετικά δυνατά σημεία των παικτών σε αγώνες αντισφαίρισης.

5.9 Στατιστικά Χαρακτηριστικά Αγώνα

Η ενσωμάτωση συγκεκριμένων στατιστικών χαρακτηριστικών των παικτών στο μοντέλο πρόβλεψης μπορεί να βελτιώσει σημαντικά την παραγωγή του. Η συμπερίληψη χαρακτηριστικών [Πιν. 2.9] παρέχει πολύτιμες πληροφορίες για το στυλ και την απόδοση ενός παίκτη.

Ειδικότερα, το ποσοστό των επιτυχημένων πρώτων σερβίς που έγιναν από έναν παίκτη, υποδεικνύει την ικανότητά του να εκκινήσει τον πόντο αποτελεσματικά και να κερδίσει

δυνητικά πλεονέκτημα όπως και να παρέχει ισχυρά και ακριβή σερβίς. Το ποσοστό επιτυχίας του παίκτη να κερδίσει πόντους όταν εκτελεί το 1ο και το 2ο σερβίς του, παρέχουν επίσης πολύτιμες πληροφορίες σχετικά με την ικανότητα ενός παίκτη να εκμεταλλεύεται τα σερβίς του και να κυριαρχεί στα παιχνίδια που αυτός σερβίρει. Το ποσοστό επιτυχίας του παίκτη στην επιστροφή του 1ου και του 2ου σερβίς το αντιπάλου του, δείχνουν την ικανότητα του παίκτη να εξουδετερώνει τα σερβίς του αντιπάλου και να ασκεί πίεση στα παιχνίδια. Ο αριθμός των σημείων διακοπής (breakpoints) που μετατρέπονται από έναν παίκτη, υπογραμμίζει την ικανότητά του να εκμεταλλεύεται κρίσιμες ευκαιρίες και να διαταράσσει το σερβίς του αντιπάλου. Η ταχύτητα σερβίς, που υπολογίζεται με βάση την αναλογία των άσων προς τα διπλά σφάλματα, παρέχει μια εκτίμηση της ταχύτητας σερβίς ενός παίκτη. Χρησιμεύει ως δείκτης της ισχύος τους και μπορεί να επηρεάσει σημαντικά την ικανότητά τους να κυριαρχούν στα παιχνίδια που σερβίρουν.

Χαρακτηριστικά
1st Serve (Year) – ποσοστό επιτυχημένων πρώτων σερβίς στην περίοδο του τελευταίου χρόνου
2nd Serve (Year) – ποσοστό επιτυχημένων δεύτερων σερβίς στην περίοδο του τελευταίου χρόνου
1st Serve Points Won (Year) – ποσοστό επιτυχημένων πόντων κατά την πραγματοποίηση του πρώτου σερβίς στην περίοδο του τελευταίου χρόνου
2nd Serve Points Won (Year) – ποσοστό επιτυχημένων πόντων κατά την πραγματοποίηση του δεύτερου σερβίς στην περίοδο του τελευταίου χρόνου
1st Serve Return Points Won (Year) – ποσοστό επιτυχημένων πόντων κατά την επιστροφή του πρώτου σερβίς στην περίοδο του τελευταίου χρόνου
2nd Serve Return Points Won (Year) – ποσοστό επιτυχημένων πόντων κατά την επιστροφή του δεύτερου σερβίς στην περίοδο του τελευταίου χρόνου
Break Points Converted (Year) – ποσοστό break πόντων που μετατράπηκαν στην περίοδο του τελευταίου χρόνου
Serving Speed (Year) – η ταχύτητα του σερβίς στην περίοδο του τελευταίου χρόνου
Service Games (Year) – ποσοστό νικηφόρων παιχνιδιών (games) στην περίοδο του τελευταίου χρόνου
Break Points Saved (Year) – ποσοστό break πόντων που «σώθηκαν» στην περίοδο του τελευταίου χρόνου

Πίνακας 5.9 : Χαρακτηριστικά σχετικά με την βαθμολογία κατάταξης Elo

Ενσωματώνοντας τέτοια στατιστικά χαρακτηριστικά στο σύνολο δεδομένων, το μοντέλο πρόβλεψης μπορεί να αξιοποιήσει αυτές τις πληροφορίες για να αποκτήσει βαθύτερες γνώσεις σχετικά με την απόδοση των παικτών. Αυτά τα χαρακτηριστικά επιτρέπουν στο μοντέλο να καταγράφει τις ικανότητες σερβίς και επιστροφής ενός παίκτη, την επιτυχία του σε κρίσιμες στιγμές όπως τα σημεία διακοπής και τη συνολική του ικανότητα σερβιρίσματος. Τελικά, η συμπερίληψη αυτών των χαρακτηριστικών ενισχύει την ικανότητα του μοντέλου να κάνει πιο ακριβείς προβλέψεις των αποτελεσμάτων των αγώνων αντισφαίρισης.

5.10 Προγνωστικά Χαρακτηριστικά

Το σύνολο δεδομένων περιλαμβάνει σημαντικά χαρακτηριστικά προγνωστικών που μπορούν να παρέχουν πολύτιμες πληροφορίες για την πρόβλεψη των αποτελεσμάτων του αγώνα [Πιν. 5.10]. Δύο από αυτά τα χαρακτηριστικά είναι το B365 και το Spread. Το B365 αντιπροσωπεύει τις πιθανότητες που προσφέρει η εταιρεία προγνωστικών Bet365 για τη νίκη ενός παίκτη σε έναν αγώνα. Αυτό το χαρακτηριστικό αντικατοπτρίζει την αντιληπτή πιθανότητα να κερδίσει ένας παίκτης με βάση την αξιολόγηση αυτών που θέτουν προγνωστικών και την αγορά. Μπορεί να είναι ένας χρήσιμος δείκτης του κοινού συναισθήματος και των προσδοκιών γύρω από την απόδοση ενός παίκτη. Αποτυπώνεται επίσης η διαφορά (spread) μεταξύ του μέσου όρου και των καλύτερων τιμών που διατίθενται στην αγορά τόσο για το φαβορί όσο και για το αουτσάιντερ. Αυτά τα χαρακτηριστικά πρόγνωσης αντικατοπτρίζουν την αξιολόγηση της αγοράς για τις πιθανότητες νίκης των παικτών και μπορούν να υποδείξουν την αντιληπτή δύναμη κάθε παίκτη. Η διαφορά μεταξύ του μέσου όρου και των καλύτερων τιμών προσφέρει πληροφορίες για τον βαθμό εμπιστοσύνης που έχουν οι εταιρείες προγνωστικών αποτελεσμάτων στις αποδόσεις τους, καθώς και για πιθανές αποκλίσεις στην αγορά. Με την ενσωμάτωση αυτών των χαρακτηριστικών στο σύνολο δεδομένων, το μοντέλο πρόβλεψης μπορεί να αξιοποιήσει αυτές τις πληροφορίες για να αποκτήσει πρόσθετες προοπτικές σχετικά με τη δυναμική του αγώνα και ενδεχομένως να βελτιώσει την ακρίβεια των προβλέψεων αποτελεσμάτων.

Χαρακτηριστικά
B365 – Αποδόσεις νίκης αγώνα, σύμφωνα με την εταιρεία Bet365
Ratio_B365 – Αποδόσεις νίκης αγώνα, σύμφωνα με την εταιρεία Bet365 (σε κλασματική αναπαράσταση)
Spread – Διαφορά μεταξύ του μέσου όρου και των καλύτερων τιμών αποδόσεων που διατίθενται στην αγορά

Πίνακας 5.10 : Χαρακτηριστικά σχετικά με τα προγνωστικά και τις αποδόσεις

5.11 Χαρακτηριστικά ήδη υπάρχοντα στο Σύνολο Δεδομένων

Το σύνολο δεδομένων περιλαμβάνει ακόμη χαρακτηριστικά [Σχ.5.11] εκ των προτέρων, τα οποία δεν δημιουργήθηκαν, αλλά χρησιμοποιήθηκαν όπως παρέχονται με μόνη τροποποίηση την εφαρμογή των μετασχηματισμών διαφοράς ή κλάσματος, όπως αναφέρθηκε παραπάνω.

Χαρακτηριστικά
Tourney ID – Χαρακτηριστικό αναγνωριστικό του τουρνουά
Location – Τοποθεσία διεξαγωγής τουρνουά
Surface – Επιφάνεια Γηπέδου που διεξάγεται το τουρνουά
Draw Size – Αριθμός συμμετεχόντων στο τουρνουά
Series – Επίπεδο τουρνουά
Date – Ημερομηνία Διεξαγωγής Αγώνα στο Τουρνουά
Best Of – Τρόπος Ανάδειξης Νικητή στο Τουρνουά
Round – Γύρος του Αγώνα στο Τουρνουά
Age – Ηλικία Παίκτη κατά την διεξαγωγή του Τουρνουά
Ratio Age – Ηλικία Παίκτη κατά την διεξαγωγή του Τουρνουά (σε κλασματική αναπαράσταση)
Height – Ύψος Παίκτη κατά την διεξαγωγή του Τουρνουά

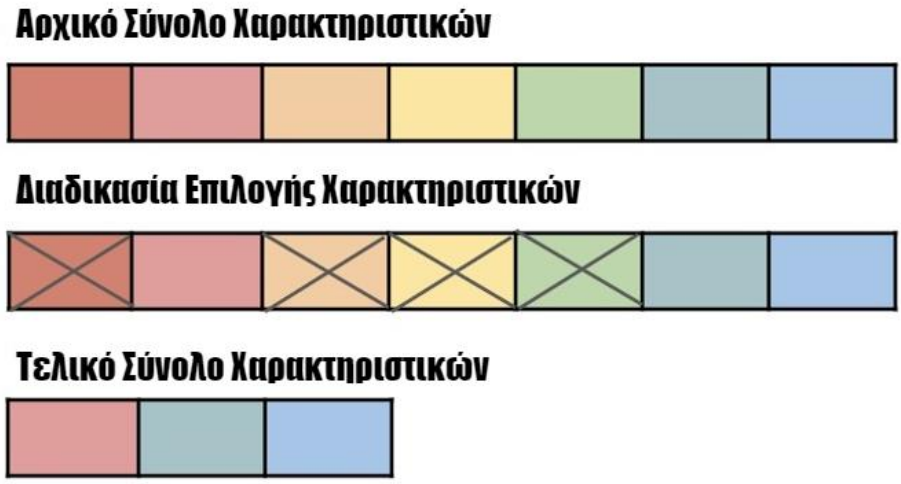
Ratio Height - Ύψος Παίκτη κατά την διεξαγωγή του Τουρνουά (σε κλασματική αναπαράσταση)
Preferred Hand – Προτιμώμενο Χέρι Παίκτη

Πίνακας 5.10 : Χαρακτηριστικά που υπάρχουν ήδη στο Σύνολο Δεδομένων

6

Εξαγωγή Χαρακτηριστικών

Η Επιλογή Χαρακτηριστικών (Feature Selection) [Σχ. 6.1] είναι μια κρίσιμη διαδικασία στη σφαίρα της Μηχανικής Μάθησης και της Ανάλυσης Δεδομένων, όπου ένα υποσύνολο σχετικών χαρακτηριστικών ή μεταβλητών επιλέγεται από μια μεγαλύτερη ομάδα διαθέσιμων επιλογών σε ένα σύνολο δεδομένων. Ο πρωταρχικός της στόχος είναι να εντοπίσει τα πιο ενημερωτικά και διακριτικά χαρακτηριστικά που συμβάλλουν σημαντικά στο τρέχον πρόβλημα πρόβλεψης ή ανάλυσης, ενώ απορρίπτει τα άσχετα ή περιττά χαρακτηριστικά.

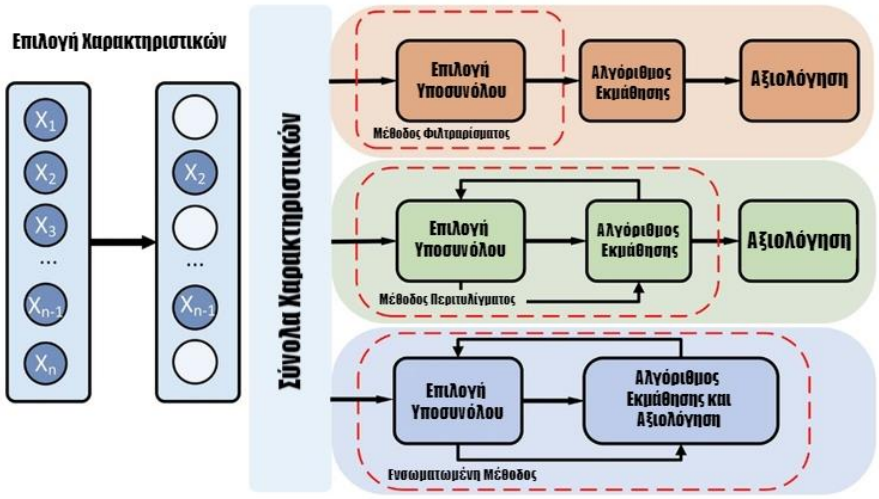


Σχήμα 6.1 : Μία απλή σχηματική αναπαράσταση της Επιλογής Χαρακτηριστικών

Η σημασία της Επιλογής Χαρακτηριστικών δεν μπορεί να υπερεκτιμηθεί, καθώς παίζει σημαντικό ρόλο στη βελτίωση της απόδοσης του μοντέλου, της ερμηνευσιμότητας και της

υπολογιστικής του αποτελεσματικότητας. Η διαδικασία Επιλογής Χαρακτηριστικών καθοδηγείται από την πρόθεση διατήρησης βασικών πληροφοριών, μετριάζοντας ταυτόχρονα τις αρνητικές επιπτώσεις του θορύβου, της υπερπροσαρμογής και των προκλήσεων που δημιουργούνται από σύνολα δεδομένων υψηλών διαστάσεων. Επιλέγοντας ένα μειωμένο σύνολο χαρακτηριστικών, το μοντέλο αποκτά βελτιωμένη ερμηνευτικότητα και ενισχύει την ικανότητά του να γενικεύει καλά σε νέες, αόρατες περιπτώσεις δεδομένων. Επιπλέον, η επιλογή χαρακτηριστικών μπορεί να οδηγήσει σε αξιοσημείωτα υπολογιστικά οφέλη, επιτρέποντας ταχύτερη εκπαίδευση και εξαγωγή συμπερασμάτων, ιδιαίτερα όταν αντιμετωπίζουμε σύνολα δεδομένων που περιέχουν σημαντικό αριθμό χαρακτηριστικών.

Οι μέθοδοι Επιλογής Χαρακτηριστικών μπορούν να ταξινομηθούν ευρέως σε τρεις κατηγορίες: Μεθόδους Φιλτραρίσματος (Filter Method), Μεθόδους Περιτυλίγματος (Wrapper Method), και Ενσωματωμένες Μεθόδους (Embedded Methods) [Σχ.6.2]. Οι Μέθοδοι Φιλτραρίσματος αξιολογούν τα χαρακτηριστικά ανεξάρτητα από τον επιλεγμένο αλγόριθμο εκμάθησης, χρησιμοποιώντας στατιστικές ιδιότητες όπως συσχέτιση ή το κέρδος πληροφοριών για την επιλογή χαρακτηριστικών. Οι Μέθοδοι περιτυλίγματος, από την άλλη πλευρά, αξιολογούν τα υποσύνολα χαρακτηριστικών χρησιμοποιώντας έναν συγκεκριμένο αλγόριθμο εκμάθησης και αξιολογούν την απόδοσή τους σε ένα σύνολο επικύρωσης, αναζητώντας ενεργά ένα βέλτιστο υποσύνολο χαρακτηριστικώ εξερευνώντας διάφορους συνδυασμούς. Οι Ενσωματωμένες Μέθοδοι περιλαμβάνουν την ενσωμάτωση της Επιλογής Χαρακτηριστικών μέσα στον ίδιο τον αλγόριθμο εκμάθησης, βελτιστοποιώντας ταυτόχρονα τη διαδικασία επιλογής χαρακτηριστικών και την κατασκευή του μοντέλου.



Σχήμα 6.2 : Μία απλή σχηματική αναπαράσταση των διαφορετικών μεθόδων Επιλογής Χαρακτηριστικών

6.1 Μέθοδοι Φιλτραρίσματος

Οι Μέθοδοι Φιλτραρίσματος στην Επιλογή Χαρακτηριστικών χρησιμοποιούν την κατάταξη χαρακτηριστικών ως μέτρηση αξιολόγησης. Αυτές οι μέθοδοι αξιολογούν τη συνάφεια των χαρακτηριστικών αποδίδοντάς τους βαθμολογίες με βάση στατιστικά τεστ, τα οποία μετρούν τη συσχέτισή τους με τη μεταβλητή στόχο. Τα χαρακτηριστικά που πέφτουν κάτω από ένα καθορισμένο όριο απορρίπτονται, ενώ επιλέγονται εκείνα που υπερβαίνουν το όριο. Το προκύπτον υποσύνολο επιλεγμένων χαρακτηριστικών, στη συνέχεια χρησιμοποιείται ως είσοδος για τον επιλεγμένο αλγόριθμο ταξινόμησης. Οι Μέθοδοι Φιλτραρίσματος είναι ανεξάρτητες από τον αλγόριθμο ταξινόμησης, ο οποίος βοηθά στον μετριασμό της υπερπροσαρμογής και της μεροληψίας ταξινομητή. Αυτή η ανεξαρτησία ωστόσο σημαίνει ότι οι Μέθοδοι Φιλτραρίσματος δεν λαμβάνουν υπόψη την αλληλεπίδραση μεταξύ των χαρακτηριστικών και του ταξινομητή κατά την Επιλογή Χαρακτηριστικών. Κατά συνέπεια, το επιλεγμένο σύνολο χαρακτηριστικών είναι πιο γενικό και δεν έχει βελτιστοποιηθεί για κάποιον συγκεκριμένο ταξινομητή, γεγονός που μπορεί να έχει ως αποτέλεσμα μειωμένη απόδοση πρόβλεψης σε σύγκριση με τις Μεθόδους Περιτυλίγματος ή τις Ενσωματωμένες Μεθόδους. Από τις Μεθόδους Φιλτραρίσματος, αυτή που χρησιμοποιήθηκε στην παρούσα μελέτη είναι η Συσχέτιση.

6.1.1 Συσχέτιση

Η Συσχέτιση (Correlation) είναι μια ευρέως χρησιμοποιούμενη τεχνική στην Επιλογή Χαρακτηριστικών που αξιολογεί τη σχέση μεταξύ των χαρακτηριστικών και της μεταβλητής στόχου. Η θεωρητική βάση της συσχέτισης βρίσκεται στην έννοια της μέτρησης της ισχύος και της κατεύθυνσης της γραμμικής σχέσης μεταξύ των μεταβλητών. Προσδιορίζει ποσοτικά τον βαθμό στον οποίο δύο μεταβλητές ποικίλλουν μαζί, παρέχοντας πληροφορίες για τη συσχέτισή τους.

Ένα από τα βασικά πλεονεκτήματα της χρήσης της συσχέτισης ως μεθόδου Επιλογής Χαρακτηριστικών είναι η απλότητα και η ευκολία ερμηνείας της. Με τον υπολογισμό του συντελεστή συσχέτισης, όπως ο συντελεστής συσχέτισης Pearson, οι ερευνητές μπορούν γρήγορα να εντοπίσουν χαρακτηριστικά που έχουν ισχυρή γραμμική σχέση με τη μεταβλητή στόχο. Αυτό επιτρέπει τον εντοπισμό δυνητικά σημαντικών χαρακτηριστικών που μπορούν να επηρεάσουν το αποτέλεσμα του μοντέλου.

Ο συντελεστής συσχέτισης Pearson υπολογίζεται ως εξής:

$$\rho_{X,Y} = cov(X, Y) / \sigma_X \sigma_Y$$

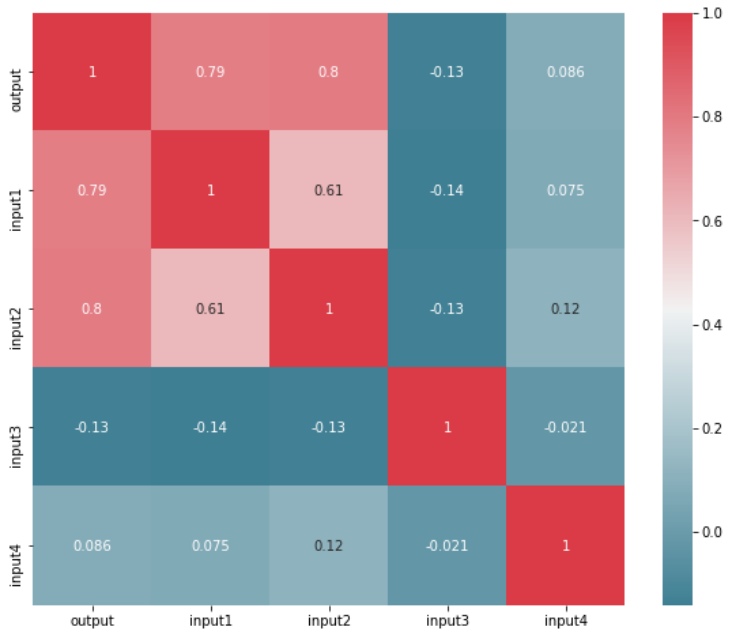
όπου

cov : η διακύμανση

σ_X*: η τυπική απόκλιση της X*

σ_Y*: η τυπική απόκλιση της Y*

Οι συντελεστές συσχέτισης εμφανίζονται σε μία δομή που ονομάζεται πίνακας συσχέτισης (correlation matrix) [Σχ.6.3]. Ο πίνακας απεικονίζει την συσχέτιση μεταξύ όλων των πιθανών ζευγών τιμών σε έναν πίνακα. Είναι ένα ισχυρό εργαλείο για την σύνοψη ενός μεγάλου συνόλου δεδομένων και τον εντοπισμό και την οπτικοποίηση μοτίβων στα δεδομένα. Ένας πίνακας συσχέτισης αποτελείται από γραμμές και στήλες που δείχνουν τις μεταβλητές. Κάθε κελί σε έναν πίνακα περιέχει τον συντελεστή συσχέτισης.



Σχήμα 6.3 : Παράδειγμα ενός Πίνακα Συσχέτισης

Είναι σημαντικό να σημειωθεί ότι η συσχέτιση έχει και τους περιορισμούς της. Πρώτον, η συσχέτιση καταγράφει μόνο γραμμικές σχέσεις και μπορεί να μην ανιχνεύει μη γραμμικούς συσχετισμούς μεταξύ των χαρακτηριστικών και της μεταβλητής στόχου. Αυτό σημαίνει ότι σημαντικές μη γραμμικές σχέσεις μπορεί να παραβλεφθούν χρησιμοποιώντας μόνο τη συσχέτιση. Επιπλέον, η συσχέτιση δεν παρέχει πληροφορίες σχετικά με την αιτιότητα μεταξύ των μεταβλητών. Υποδεικνύει απλώς τον βαθμό συσχέτισης.

Οσον αφορά τη συνεισφορά του στον γενικό στόχο της επιλογής χαρακτηριστικών, η συσχέτιση μπορεί να βοηθήσει στη μείωση της διάστασης εντοπίζοντας χαρακτηριστικά υψηλής συσχέτισης που μπορεί να είναι περιττά ή να παρέχουν περιττές πληροφορίες. Με την κατάργηση τέτοιων χαρακτηριστικών, η διάσταση του συνόλου δεδομένων μπορεί να μειωθεί, οδηγώντας σε πιο αποτελεσματικά και απλοποιημένα μοντέλα. Επιπλέον, η επιλογή χαρακτηριστικών με βάση τη συσχέτισή τους με τη μεταβλητή στόχο μπορεί ενδεχομένως να βελτιώσει την απόδοση του μοντέλου συμπεριλαμβάνοντας μόνο τα πιο σχετικά χαρακτηριστικά που σχετίζονται στενά με το αποτέλεσμα. Τέλος, η ανάλυση συσχέτισης ενισχύει την ερμηνευτικότητα επισημαίνοντας τις σχέσεις μεταξύ των χαρακτηριστικών και της μεταβλητής στόχου, επιτρέποντας στους ερευνητές να αποκτήσουν γνώσεις σχετικά με τους υποκείμενους παράγοντες που οδηγούν τα παρατηρούμενα μοτίβα.

Συνολικά, ενώ η συσχέτιση παρέχει μια πολύτιμη αρχική αξιολόγηση της σχέσης μεταξύ των χαρακτηριστικών και της μεταβλητής στόχου, θα πρέπει να συμπληρωθεί με άλλες τεχνικές επιλογής χαρακτηριστικών για την καταγραφή μη γραμμικών σχέσεων και την εξέταση πρόσθετων παραγόντων όπως η αιτιότητα και η γνώση του τομέα.

6.2 Μέθοδοι Περιτυλίγματος

Σε αντίθεση με τις μεθόδους Φιλτραρίσματος, οι μέθοδοι Περιτυλίγματος χρησιμοποιούν την απόδοση ενός επιλεγμένου αλγόριθμου ταξινόμησης για να επιλέξουν το καλύτερο υποσύνολο χαρακτηριστικών. Αυτό το πλεονεκτικό γνώρισμα των μεθόδων Περιτυλίγματος έχει αποδειχθεί ότι αποδίδει υψηλότερη προγνωστική απόδοση σε σύγκριση με τις μεθόδους Φιλτραρίσματος. Ωστόσο, η εξαντλητική αναζήτηση όλων των πιθανών συνδυασμών χαρακτηριστικών είναι υπολογιστικά μη πρακτική. Για να αντιμετωπιστεί αυτό, χρησιμοποιούνται ευρετικές στρατηγικές αναζήτησης όπως η Διαδοχική Αναζήτηση, οι Γενετικοί Αλγόριθμοι ή η Βελτιστοποίηση Αποικιών Μυρμηγκιών για τη δημιουργία υποσυνόλων χαρακτηριστικών. Αυτά τα υποσύνολα χρησιμοποιούνται στη συνέχεια για την εκπαίδευση και την αξιολόγηση του αλγόριθμου ταξινόμησης, με την απόδοση να αξιολογείται συνήθως χρησιμοποιώντας μετρήσεις όπως το AUC (περιοχή κάτω από τη χαρακτηριστική καμπύλη λειτουργίας του δέκτη). Το υποσύνολο που επιτυγχάνει την καλύτερη απόδοση θεωρείται το βέλτιστο υποσύνολο.

Οι μέθοδοι Περιτυλίγματος εξετάζουν σιωπηρά τις εξαρτήσεις χαρακτηριστικών, συμπεριλαμβανομένων των αλληλεπιδράσεων και των πλεονασμών, κατά τη διαδικασία επιλογής. Ωστόσο, λόγω της υπολογιστικής έντασης που εμπλέκεται στη δημιουργία και την αξιολόγηση υποσυνόλων χαρακτηριστικών, οι μέθοδοι Περιτυλίγματος είναι πιο απαιτητικές υπολογιστικά σε σύγκριση με τις μεθόδους Φιλτραρίσματος και τις Ενσωματωμένες.

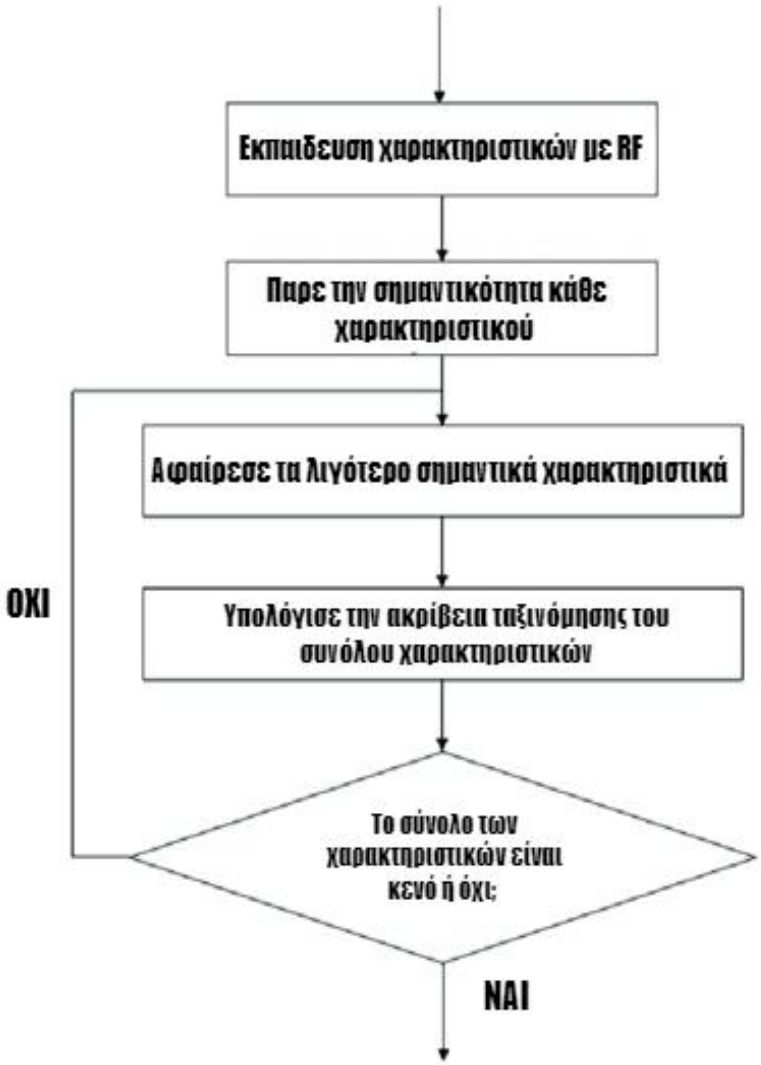
Είναι σημαντικό να σημειωθεί ότι οι μέθοδοι Περιτυλίγματος εξαρτώνται από τον συγκεκριμένο ταξινομητή που χρησιμοποιείται. Επομένως, δεν υπάρχει καμία εγγύηση ότι τα επιλεγμένα χαρακτηριστικά θα παραμείνουν βέλτιστα όταν χρησιμοποιείται διαφορετικός Ταξινομητής. Σε ορισμένες περιπτώσεις, η χρήση της απόδοσης ταξινομητή ως οδηγού για την επιλογή χαρακτηριστικών μπορεί να οδηγήσει σε ένα υποσύνολο χαρακτηριστικών με καλή ακρίβεια στο σύνολο δεδομένων εκπαίδευσης, αλλά κακή γενίκευση σε εξωτερικά σύνολα δεδομένων, υποδηλώνοντας υψηλότερο κίνδυνο υπερπροσαρμογής.

Σε αντίθεση με τις μεθόδους Φιλτραρίσματος, οι οποίες παράγουν μια ταξινομημένη λίστα χαρακτηριστικών, οι μέθοδοι Περιτυλίγματος αποδίδουν ένα μόνο υποσύνολο χαρακτηριστικών «καλύτερων» ως έξοδο. Αν και αυτό εξαλείφει την ανάγκη καθορισμού ενός βέλτιστου ορίου ή αριθμού χαρακτηριστικών, καθιστά λιγότερο σαφές ποια χαρακτηριστικά είναι σχετικά πιο σημαντικά στο επιλεγμένο υποσύνολο. Συνολικά, οι μέθοδοι Περιτυλίγματος μπορούν να επιτύχουν καλύτερη απόδοση ταξινόμησης, αλλά μπορεί να είναι λιγότερο αποτελεσματικές στην αποκάλυψη της σχέσης μεταξύ των χαρακτηριστικών και της κλάσης-στόχου.

Από τις μεθόδους Περιτυλίγματος, αυτή που χρησιμοποιήθηκε στην παρούσα μελέτη είναι η Αναδρομική Εξάλειψη Χαρακτηριστικών.

6.2.1 Αναδρομική Εξάλειψη Χαρακτηριστικών

Η Αναδρομική Εξάλειψη Χαρακτηριστικών (Recursive Feature Elimination, RFE) είναι μια τεχνική Επιλογής Χαρακτηριστικών που στοχεύει στην επαναληπτική εξάλειψη των λιγότερο ενημερωτικών χαρακτηριστικών από ένα σύνολο δεδομένων με βάση τη σημασία τους για την προγνωστική απόδοση του μοντέλου. Τα θεωρητικά θεμέλια της RFE βρίσκονται στην έννοια της οπισθοδρομικής εξάλειψης, όπου τα χαρακτηριστικά αφαιρούνται σταδιακά για να ενισχυθεί η ικανότητα του μοντέλου να γενικεύει και να βελτιώνει την απόδοσή του [Σχ 6.4].



Σχήμα 6.4 : Σχηματική Αναπαράσταση του αλγορίθμου RFE

Ένα από τα βασικά πλεονεκτήματα της RFE είναι η ικανότητά της να εξετάζει τις αλληλεξαρτήσεις μεταξύ των χαρακτηριστικών κατά τη διαδικασία εξάλειψης. Αξιολογώντας τον αντίκτυπο της κατάργησης κάθε χαρακτηριστικού στην απόδοση του μοντέλου, λαμβάνει υπόψη τις πιθανές αλληλεπιδράσεις και πλεονασμούς μεταξύ των χαρακτηριστικών, οδηγώντας σε ένα πιο ισχυρό υποσύνολο χαρακτηριστικών. Αυτή η εξέταση των αλληλεπιδράσεων χαρακτηριστικών μπορεί να αποκαλύψει κρυφές σχέσεις και να βελτιώσει την κατανόηση των υποκείμενων δεδομένων από το μοντέλο.

Ένα άλλο πλεονέκτημα είναι η ευελιξία του να εργάζεται με διαφορετικούς αλγόριθμους Μηχανικής Μάθησης. Μπορεί να χρησιμοποιηθεί σε συνδυασμό με ένα ευρύ φάσμα μοντέλων, συμπεριλαμβανομένης της γραμμικής παλινδρόμησης, των μηχανών διανυσμάτων

υποστήριξης και των δέντρων αποφάσεων. Αυτή η ευελιξία επιτρέπει την εφαρμογή της RFE σε διάφορους τύπους συνόλων δεδομένων και εργασίες πρόβλεψης μοντελοποίησης.

Ωστόσο, η RFE έχει επίσης πιθανούς περιορισμούς. Η υπολογιστική πολυπλοκότητα της RFE αυξάνεται με τον αριθμό των χαρακτηριστικών, καθιστώντας το πιο χρονοβόρο για μεγάλα σύνολα δεδομένων με χώρους χαρακτηριστικών υψηλών διαστάσεων. Επιπλέον, η απόδοση εξαρτάται σε μεγάλο βαθμό από την επιλεγμένη μέτρηση απόδοσης και τα κριτήρια διακοπής για την εξάλειψη χαρακτηριστικών. Η σωστή επικύρωση και η επιλογή αυτών των παραμέτρων είναι ζωτικής σημασίας για την αποφυγή κακής προσαρμογής ή υπερβολικής προσαρμογής του μοντέλου.

Όσον αφορά τον γενικό στόχο της επιλογής χαρακτηριστικών, η RFE συμβάλλει στη μείωση της διάστασης εξαλείφοντας συστηματικά λιγότερο πληροφοριακά χαρακτηριστικά. Με την επαναληπτική κατάργηση χαρακτηριστικών χαμηλής σημασίας, βοηθά στη δημιουργία ενός συμπαγούς υποσυνόλου χαρακτηριστικών που συλλαμβάνει τις πιο σχετικές πληροφορίες για πρόβλεψη. Αυτή η μείωση της διάστασης όχι μόνο βελτιώνει την υπολογιστική απόδοση αλλά μειώνει επίσης τον κίνδυνο υπερβολικής προσαρμογής μειώνοντας την πολυπλοκότητα του μοντέλου.

Επιπλέον, η RFE μπορεί να βελτιώσει την απόδοση του μοντέλου εστιάζοντας στα πιο ενημερωτικά χαρακτηριστικά. Καταργώντας άσχετες ή περιττές λειτουργίες, βοηθά στην ανάδειξη των βασικών χαρακτηριστικών των δεδομένων που συμβάλλουν σε ακριβείς προβλέψεις. Αυτή η διαδικασία εξάλειψης χαρακτηριστικών διασφαλίζει ότι το μοντέλο βασίζεται στα πιο διακριτικά και ενημερωτικά χαρακτηριστικά, οδηγώντας σε βελτιωμένη απόδοση πρόβλεψης.

Επιπλέον, η RFE μπορεί να βελτιώσει την ερμηνευτικότητα απλοποιώντας το μοντέλο και τονίζοντας τη σημασία συγκεκριμένων χαρακτηριστικών. Επιλέγοντας ένα υποσύνολο χαρακτηριστικών που είναι πιο σχετικά με την πρόβλεψη, βελτιώνει την ερμηνευτικότητα του μοντέλου παρέχοντας πληροφορίες για τους βασικούς παράγοντες που οδηγούν τις προβλέψεις. Αυτό δίνει τη δυνατότητα στους ερευνητές και στα ενδιαφερόμενα μέρη να αποκτήσουν καλύτερη κατανόηση των υποκείμενων δεδομένων και να λάβουν τεκμηριωμένες αποφάσεις με βάση τα καθορισμένα σημαντικά χαρακτηριστικά.

6.3 Ενσωματωμένες Μέθοδοι

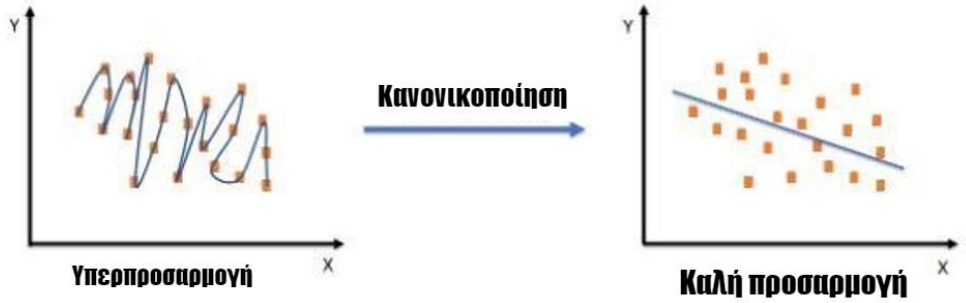
Οι Ενσωματωμένες Μέθοδοι ενσωματώνουν την επιλογή χαρακτηριστικών στον αλγόριθμο ταξινόμησης. Σε αυτήν την προσέγγιση, κατά τη διάρκεια της εκπαίδευσης, ο ταξινομητής προσαρμόζει τις εσωτερικές του παραμέτρους και αποδίδει βάρη ή σημασία σε κάθε χαρακτηριστικό με βάση τη συμβολή τους στην επίτευξη της βέλτιστης ακρίβειας ταξινόμησης. Κατά συνέπεια, η αναζήτηση για το καλύτερο υποσύνολο χαρακτηριστικών και η κατασκευή μοντέλου γίνονται ταυτόχρονα σε ένα μόνο βήμα. Παραδείγματα ενσωματωμένων μεθόδων περιλαμβάνουν αλγόριθμους που βασίζονται σε δέντρα αποφάσεων (όπως Decision Trees, Random Forests και Gradient Boosting) και επιλογή χαρακτηριστικών χρησιμοποιώντας Regularization μοντέλα (όπως LASSO ή Elastic Net). Οι μέθοδοι Regularization λειτουργούν συνήθως με γραμμικούς ταξινομητές (όπως Support Vector Machines ή Logistic Regression) τιμωρώντας ή συρρικνώνοντας τους συντελεστές των χαρακτηριστικών που δεν συμβάλλουν σημαντικά στο μοντέλο. Παρόμοια με τις μεθόδους φιλτραρίσματος, οι αλγόριθμοι που βασίζονται σε Decision Trees και οι μέθοδοι Regularization παρέχουν μια ταξινομημένη λίστα χαρακτηριστικών. Οι αλγόριθμοι που βασίζονται σε Decision Trees κατατάσσουν τη σημασία των χαρακτηριστικών χρησιμοποιώντας μετρικές, ενώ οι οι μέθοδοι Regularization κατατάσσουν τα χαρακτηριστικά με βάση το μέγεθος των συντελεστών (coefficients) τους.

Οι Ενσωματωμένες Μέθοδοι χρησιμεύουν ως μια ενδιάμεση λύση μεταξύ των μεθόδων φιλτραρίσματος και περιτυλίγματος, συνδυάζοντας τα πλεονεκτήματα και των δύο προσεγγίσεων. ΄Οπως οι μέθοδοι φιλτραρίσματος, οι ενσωματωμένες μέθοδοι είναι υπολογιστικά πιο ¨ελαφριές' από τις μεθόδους περιτυλίγματος (αν και εξακολουθούν να είναι πιο απαιτητικές από τις μεθόδους φιλτραρίσματος). Επιτυγχάνουν αυτό το μειωμένο υπολογιστικό φορτίο, ενώ ενσωματώνουν τη μεροληψία του ταξινομητή στην επιλογή χαρακτηριστικών, με αποτέλεσμα βελτιωμένη απόδοση του ταξινομητή, παρόμοια με αυτή που στοχεύουν να επιτύχουν οι μέθοδοι περιτυλίγματος. Ορισμένες ενσωματωμένες μέθοδοι, όπως τα Random Forests, επιτρέπουν αλληλεπιδράσεις χαρακτηριστικών, συμπεριλαμβανομένων αλληλεπιδράσεων υψηλότερης τάξης (πέρα από ζεύγη). Σε αντίθεση με μεθόδους φιλτραρίσματος, το Random Forest δεν εξαλείφει αυτόματα τα περιττά χαρακτηριστικά και η παρουσία περιττών χαρακτηριστικών μπορεί στην πραγματικότητα να μειώσει την απόδοση του αλγορίθμου. Για να αντιμετωπιστεί αυτό, μια προσέγγιση είναι το φιλτράρισμα των περιττών χαρακτηριστικών πριν από την εφαρμογή του Random Forest. Μια άλλη πιθανή λύση είναι η συγκέντρωση των πληροφοριών που μεταφέρονται από περιττές χαρακτηριστικά. Από την άλλη πλευρά, οι μέθοδοι ¨τιμώρησης' όπως το LASSO

μπορούν να απορρίψουν περιττά χαρακτηριστικά, αλλά δεν έχουν ενσωματωμένη ικανότητα ανίχνευσης αλληλεπιδράσεων μεταξύ των χαρακτηριστικών. Οι όροι αλληλεπίδρασης πρέπει να περιλαμβάνονται ρητά στην ανάλυση, συνήθως λαμβάνοντας υπόψη εξαντλητικά όλους τους όρους αλληλεπίδρασης ανά ζεύγη για τα χαρακτηριστικά. Ωστόσο, αυτή η προσέγγιση μπορεί να είναι ανακριβής και υπολογιστικά απαγορευτική σε σύνολα δεδομένων υψηλών διαστάσεων. Οι στρατηγικές δύο σταδίων ή οι υβριδικές στρατηγικές που μειώνουν τον χώρο αναζήτησης μπορούν να θεωρηθούν ως εναλλακτικές λύσεις σε αυτήν την πρόκληση. Από τις ενσωματωμένες μεθόδους, αυτές που χρησιμοποιήθηκαν στην παρούσα μελέτη είναι η κανονικοποίηση και η σημαντικότητα των χαρακτηριστικών.

6.3.1 Κανονικοποίηση

Η κανονικοποίηση (Regularization) χρησιμεύει ως πολύτιμη τεχνική για τον μετριασμό της υπερπροσαρμογής και την ενίσχυση της ικανότητας γενίκευσης των μοντέλων Μηχανικής Μάθησης [Σχ.6.5]. Περιλαμβάνει την εισαγωγή ενός όρου ποινής στην αντικειμενική συνάρτηση του μοντέλου, που δίνει κίνητρο για την επιλογή απλούστερων μοντέλων με μικρότερους συντελεστές ή παραμέτρους. Αυτή η διαδικασία τακτοποίησης διευκολύνει τη μείωση της εξάρτησης του μοντέλου σε συγκεκριμένα χαρακτηριστικά και ελαχιστοποιεί τον κίνδυνο υπερπροσαρμογής, με αποτέλεσμα το μοντέλο να προσαρμόζεται υπερβολικά στα δεδομένα εκπαίδευσης και να παρουσιάζει χαμηλότερη απόδοση όταν έρχεται αντιμέτωπο με νέα, αόρατα δεδομένα.



Σχήμα 6.5: Σχηματική Αναπαράσταση της επίδρασης της κανονικοποίησης στα σημεία δεδομένων

Η σημασία της εφαρμογής κανονικοποίησης στα δεδομένα πριν από την εκπαίδευση του μοντέλου πηγάζει από διάφορους παράγοντες. Πρώτον, βοηθά στην πρόληψη της κυριαρχίας των χαρακτηριστικών. Οι περιπτώσεις όπου τα χαρακτηριστικά διαθέτουν διαφορετικές κλίμακες μπορεί να προκαλέσουν προκλήσεις κατά τη διάρκεια της μαθησιακής διαδικασίας,

ιδιαίτερα με μοντέλα που είναι ευαίσθητα στην κανονικοποίηση των χαρακτηριστικών, όπως οι Μηχανές Διανυσμάτων Υποστήριξης. Τα χαρακτηριστικά με μεγαλύτερες κλίμακες μπορούν να επηρεάσουν δυσανάλογα τη διαδικασία βελτιστοποίησης, οδηγώντας σε μη βέλτιστα αποτελέσματα. Με την κανονικοποίηση των δεδομένων, όλα τα χαρακτηριστικά φέρονται σε συγκρίσιμη κλίμακα, διασφαλίζοντας ότι κανένα χαρακτηριστικό δεν υπερκαλύπτει τη διαδικασία εκμάθησης. Συνεπώς, ο όρος «κανονικοποίηση» προϋποθέτει μια πιο ισορροπημένη επίδραση σε όλα τα χαρακτηριστικά.

Δεύτερον, η κλιμάκωση των δεδομένων διευκολύνει τη δίκαιη κανονικοποίηση. Ο στόχος της κανονικοποίησης είναι να επιτευχθεί μια ισορροπία μεταξύ της αποτελεσματικής προσαρμογής των δεδομένων εκπαίδευσης και της αποφυγής της υπερπροσαρμογής. Ελλείψει κανονονικοποίησης χαρακτηριστικών, τα χαρακτηριστικά με μεγαλύτερες κλίμακες μπορεί να αποδώσουν μεγαλύτερους συντελεστές σε σύγκριση με χαρακτηριστικά με μικρότερες κλίμακες, όχι απαραίτητα λόγω της εγγενούς σημασίας τους, αλλά απλώς λόγω της κλίμακας τους. Αυτή η ασυμφωνία μπορεί να οδηγήσει σε άδικη κατανομή των ποινών της κανονικοποίησης, όπου ορισμένα χαρακτηριστικά τιμωρούνται σε μεγάλο βαθμό ενώ σε άλλα υπάρχει μεγαλύτερη επιείκεια. Μέσω της κανονικοποίησης των δεδομένων, η ποινή κανονικοποίησης κατανέμεται δίκαια σε όλα τα χαρακτηριστικά, επιτρέποντας στο μοντέλο να λαμβάνει αμερόληπτες αποφάσεις κατά τη διάρκεια της φάσης εκπαίδευσης.

Συνοπτικά, η εφαρμογή της κανονικοποίησης στα δεδομένα μέσω κλιμάκωσης διαδραματίζει κεντρικό ρόλο στη διασφάλιση ότι η διαδικασία κανονικοποίησης παραμένει δίκαιη, αμερόληπτη και λαμβάνει υπόψη τη σχετική σημασία όλων των χαρακτηριστικών. Αυτή η πρακτική συμβάλλει στη βελτίωση της ικανότητας του μοντέλου να γενικεύει αποτελεσματικά σε αόρατα δεδομένα και μετριάζει τις προκλήσεις που προκύπτουν από διακυμάνσεις στις διάφορες κλίμακες των χαρακτηριστικών.

6.4 Χειροκίνητη Επιλογή Χαρακτηριστικών

Αυτή η προσέγγιση περιλαμβάνει τη μη αυτόματη επιλογή χαρακτηριστικών με βάση τη γνώση ή τη διαίσθηση στον συγκεκριμένο τομέα. Απαιτεί την κρίση των ειδικών και περιλαμβάνει επαναληπτική δοκιμή διαφορετικών συνδυασμών χαρακτηριστικών για τον προσδιορισμό του υποσυνόλου που αποδίδει την καλύτερη απόδοση.

7 Πειραματικό Μέρος

7.1 Λογιστική Παλινδρόμηση

Η πρόβλεψη του αποτελέσματος των αγώνων αντισφαίρισης έχει σημαντικές επιπτώσεις για διάφορους ενδιαφερόμενους, συμπεριλαμβανομένων των παικτών, των προπονητών και τους λάτρεις των αθλημάτων. Η Λογιστική Παλινδρόμηση έχει χρησιμοποιηθεί ευρέως ως μοντέλο ταξινόμησης σε προβλήματα πρόβλεψης αθλημάτων. Η πειραματική μεθοδολογία που χρησιμοποιήθηκε σε αυτή την μελέτη είχε ως στόχο να αναπτύξει ένα, όσο το δυνατόν, ισχυρό και ακριβές μοντέλο Λογιστικής Παλινδρόμησης για την πρόβλεψη του αποτελέσματος ενός αγώνα αντισφαίρισης.

Η ερευνητική διαδικασία αποτελείται από πολλά κρίσιμα βήματα, συμπεριλαμβανομένης της ανάπτυξης ενός μοντέλου, που χρησιμοποιείται ως βάση για την περαιτέρω ανάλυση. Βήματα επίσης αποτελούν η επιλογή χαρακτηριστικών, μέσω της ανάλυσης της συσχέτισης, της αναδρομικής εξάλειψης χαρακτηριστικών και της ανάλυσης σημασίας των χαρακτηριστικών, καθώς και ο συντονισμός υπερπαραμέτρων και τέλος η αξιολόγηση σε ένα σύνολο δοκιμής. Η απόδοση του μοντέλου αξιολογήθηκε χρησιμοποιώντας μετρικές, όπως η ακρίβεια, η λογιστική απώλεια και ο πίνακας σύγχυσης (confusion matrix).

Για να δημιουργηθεί μια βάση σύγκρισης, δημιουργήθηκε το βασικό μοντέλο χρησιμοποιώντας την κατάταξη παίκτη ως μοναδικό χαρακτηριστικό [Πιν.7.1]. Αυτό το αρχικό μοντέλο πέτυχε στο σύνολο εκπαίδευσης ακρίβεια 66,9% και στο σύνολο επικύρωσης ακρίβεια 62,1%, παρέχοντας έτσι ένα σημείο αναφοράς βάσει του οποίου θα μπορούσαν να μετρηθούν οι επόμενες βελτιώσεις.

# Χαρακτηριστικά	Ακρίβεια Train	Λογ.Απώλεια Train	Ακρίβεια Val	Λογ. Απώλεια Val
1	66.9 %	0.635	62.1 %	0.669

Πίνακας 7.1: Αποτελέσματα του Baseline Μοντέλου, με βάση την κατάταξη ATP των παικτών

Για να ενισχυθεί η προγνωστική ισχύς του μοντέλου, όλα τα διαθέσιμα χαρακτηριστικά χρησιμοποιήθηκαν στην επόμενη δοκιμή. Αυτό το μοντέλο, που χρησιμοποιεί 63 χαρακτηριστικά, επέδειξε βελτιωμένη απόδοση με ακρίβεια 73,7% στο σύνολο εκπαίδευσης και ακρίβεια 71,6% στο σύνολο επικύρωσης [Πιν. 7.2].

# Χαρακτηριστικά	Ακρίβεια Train	Λογ.Απώλεια Train	Ακρίβεια Val	Λογ. Απώλεια Val
63	73.7 %	0.524	71.6 %	0.557

Πίνακας 7.2: Αποτελέσματα του Μοντέλου, με βάση όλα τα χαρακτηριστικά που υπάρχουν στο Σύνολο Δεδομένων

Στην συνέχεια, προκειμένου να εντοπιστεί το σύνολο δεδομένων που δίνει το βέλτιστο αποτέλεσμα στο μοντέλο μας, πραγματοποιήθηκε μια παράλληλη διαδικασία επιλογής χαρακτηριστικών και ρύθμισης υπερπαραμέτρων, η οποία έχεις ως εξής:

1. Εκτέλεση αρχικής βελτιστοποίησης υπερπαραμέτρων, χρησιμοποιώντας όλα τα χαρακτηριστικά που βρίσκονται στο σύνολο δεδομένων.
2. Εκτέλεση επιλογής χαρακτηριστικών με βάση την Αναδρομική Εξάλειψη Δεδομένων (Recursive Feature Elimination) με σκοπό την απόκτηση ενός νέου βέλτιστου συνόλου δεδομένων
3. Επαναβελτιστοποίηση των υπερπαραμέτρων στο νέο σύνολο δεδομεένων.

Τα αποτελέσματα της διαδικασίας αυτής παρουσιάζονται παρακάτω [Πιν. 7.3]. Οι βέλτιστες υπερπαράμετροι επιλέχθηκαν με βάση την Επιλογή Πλέγματος (Grid Search) και επιλέχθηκαν μέσω ενός εύρους των παρακάτω τιμών:

- Παράμετρος κανονικοποίησης C: {0.01, 0.1, 0.2, 0.4, 0.6, 1.0}

- Παράμετρος επιλυτή (solver) : {newton-cg, lbfgs, liblinear}
- Παράμετρος (penalty) : {l2, None}

# Χαρακτηριστικά	Ακρίβεια Train	Λογ.Απώλεια Train	Ακρίβεια Val	Λογ. Απώλεια Val	C	solver	penalty
63	73.6 %	0.524	71.4 %	0.557	0.1	newton-cg	L2
60	73.6%	0.524	71.3 %	0.557	0.2	liblinear	L2
55	73.6%	0.524	71.3 %	0.557	0.1	newton-cg	L2
50	73.6%	0.524	71.1 %	0.557	0.2	newton-cg	L2

Πίνακας 7.3: Αποτελέσματα ύστερα από επιλογή χαρακτηριστικών και ρύθμιση υπερπαραμέτρων

Όπως βλέπουμε και παραπάνω, η καλύτερη διαμόρφωση προκύπτει για # χαρακτηριστικά και με τις υπερπαραμέτρους:

# Χαρακτηριστικά	Ακρίβεια Train	Λογ.Απώλεια Train	Ακρίβεια Val	Λογ. Απώλεια Val	C	solver	penalty
63	73.6 %	0.524	71.4 %	0.557	0.1	newton-cg	L2

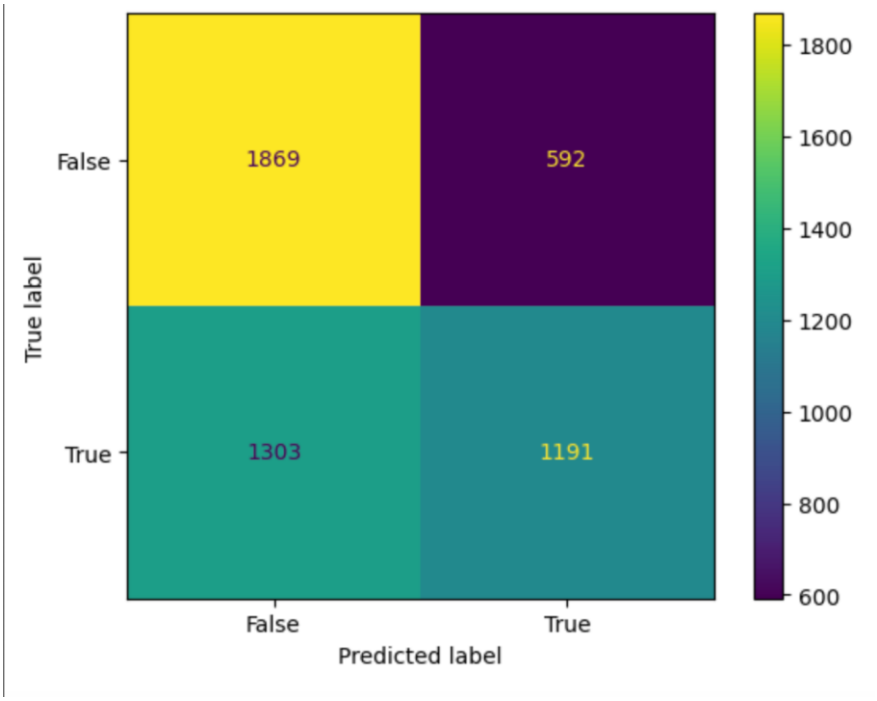
Πίνακας 7.4: Αποτελέσματα ύστερα από επιλογή χαρακτηριστικών και ρύθμιση υπερπαραμέτρων

Η απόδοση του τελικού μοντέλου αξιολογήθηκε στο ανεξάρτητο σύνολο δοκιμής για την αξιολόγηση των δυνατοτήτων γενίκευσής του. Τα αποτελέσματα αποκάλυψαν μια ακρίβεια δοκιμής 71,6%, παρέχοντας στοιχεία για την αποτελεσματικότητα του προτεινόμενου μοντέλου λογιστικής παλινδρόμησης στην πρόβλεψη του αποτελέσματος των αγώνων αντισφαίρισης.

# Χαρακτηριστικά	Ακρίβεια Train	Λογ.Απώλεια Train	Ακρίβεια Test	Λογ. Απώλεια Test	C	solver	penalty
63	73.2 %	0.528	71.6 %	0.557	0.1	newton-cg	L2

Πίνακας 7.5: Τελικό μοντέλο

Για το τελικό μοντέλο προκύπτει και ο παρακάτω πίνακας σύγχυσης (confusion matrix) [Σχ.7.1] :



Σχήμα 7.1: Πίνακας Σύγχυσης για το τελικό μοντέλο της Λογιστικής Παλινδρόμησης

7.2 Τυχαία Δάση

Παρακάτω παρουσιάζεται η πειραματική ανάλυση του Μοντέλου Τυχαίων Δασών (Random Forest) για την πρόβλεψη του αποτελέσματος των αγώνων αντισφαίρισης. Η μελέτη επικεντρώνεται στη διερεύνηση διαφόρων πτυχών του μοντέλου, συμπεριλαμβανομένου του συντονισμού υπερπαραμέτρων, της κλίμακας δεδομένων, της επιλογής χαρακτηριστικών και της αξιολόγησης απόδοσης. Τα πειραματικά αποτελέσματα πα ρέχουν πληροφορίες για τη βελτιστοποίηση της ακρίβειας και της ικανότητας γενίκευσης του μοντέλου.

Το μοντέλο αρχικά εκπαιδεύεται και επικυρώνεται χρησιμοποιώντας και τα 63 χαρακτηριστικά του συνόλου δεδομένων [Πιν.7.6]. Παραδόξως, το μοντέλο επιτυγχάνει τέλεια ακρίβεια εκπαίδευσης 100% και ακρίβεια επικύρωσης 92,2%. Αυτά τα αποτελέσματα εγείρουν ανησυχίες σχετικά με πιθανή υπερπροσαρμογή και την ανάγκη για περαιτέρω ρύθμιση του μοντέλου.

# Χαρακτηριστικά	Ακρίβεια Train	Λογ.Απώλεια Train	Ακρίβεια Val	Λογ. Απώλεια Val
63	100 %	0.098	92.2 %	0.317

Πίνακας 7.6: Αποτελέσματα του Μοντέλου, με βάση όλα τα χαρακτηριστικά που υπάρχουν στο Σύνολο Δεδομένων

Στην συνέχεια, προκειμένου να εντοπιστεί το σύνολο δεδομένων που δίνει το βέλτιστο αποτέλεσμα στο μοντέλο μας, πραγματοποιήθηκε μια παράλληλη διαδικασία επιλογής χαρακτηριστικών και ρύθμισης υπερπαραμέτρων, η οποία έχεις ως εξής:

1. Εκτέλεση αρχικής βελτιστοποίησης υπερπαραμέτρων, χρησιμοποιώντας όλα τα χαρακτηριστικά που βρίσκονται στο σύνολο δεδομένων.
2. Εκτέλεση επιλογής χαρακτηριστικών με βάση την Αναδρομική Εξάλειψη Δεδομένων (Recursive Feature Elimination) με σκοπό την απόκτηση ενός νέου βέλτιστου συνόλου δεδομένων
3. Επαναβελτιστοποίηση των υπερπαραμέτρων στο νέο σύνολο δεδομεένων.

Τα αποτελέσματα της διαδικασίας αυτής παρουσιάζονται παρακάτω [Πιν. 7.7]. Οι βέλτιστες υπερπαράμετροι επιλέχθηκαν με βάση την Επιλογή Πλέγματος (Grid Search) και επιλέχθηκαν μέσω ενός εύρους των παρακάτω τιμών:

- max_depth = {3, 5}
- n_estimators = {50, 70, 100, 130}
- min_samples_split = {5, 10, 20, 30}
- min_samples_leaf = {1, 5, 10, 15, 20}
- bootstrap = {True, False}

Ο καλύτερος συνδυασμός υπερπαραμέτρων βρέθηκε να είναι max_depth=5, n_estimators=70, min_samples_split=10, min_samples_leaf=20 και bootstrap=True. Αυτή η διαμόρφωση αποδίδει ακρίβεια εκπαίδευσης 79.8% και ακρίβεια επικύρωσης 79.7% [Πίν 7.8]. Ο αντίκτυπος της κλίμακας δεδομένων, ειδικά χρησιμοποιώντας το StandardScaler, αξιολογείται στο μοντέλο. Ωστόσο, παρατηρείται ότι η κλιμάκωση δεδομένων δεν οδηγεί σε καμία βελτίωση στην απόδοση του μοντέλου. Τα αποτελέσματα παραμένουν ίδια με το μη

συντονισμένο μοντέλο, υποδεικνύοντας ότι η κλίμακα δεδομένων έχει περιορισμένη επιρροή σε αυτό το πλαίσιο.

# Χαρακτηριστικά	Ακρίβεια Train	Λογ.Απώλεια Train	Ακρίβεια Val	Λογ. Απώλεια Val	n_estimators	min samples leaf	min samples split	max depth
63	79.2 %	0.476	78.7 %	0.497	70	20	5	5
60	78.5%	0.486	77.5 %	0.509	70	20	10	5
50	79.4%	0.47	78.1 %	0.493	50	5	10	5
40	79.4%	0.466	79.3 %	0.484	70	20	10	5
30	82.3%	0.454	81.6%	0.454	100	30	10	5

Πίνακας 7.7: Αποτελέσματα ύστερα από επιλογή χαρακτηριστικών και ρύθμιση υπερπαραμέτρων

Η απόδοση του τελικού μοντέλου αξιολογήθηκε στο ανεξάρτητο σύνολο δοκιμής για την αξιολόγηση των δυνατοτήτων γενίκευσής του. Τα αποτελέσματα αποκάλυψαν μια ακρίβεια δοκιμής 79.7%, παρέχοντας στοιχεία για την αποτελεσματικότητα του προτεινόμενου μοντέλου τυχαίων δασών στην πρόβλεψη του αποτελέσματος των αγώνων αντισφαίρισης [Πιν 7.9].

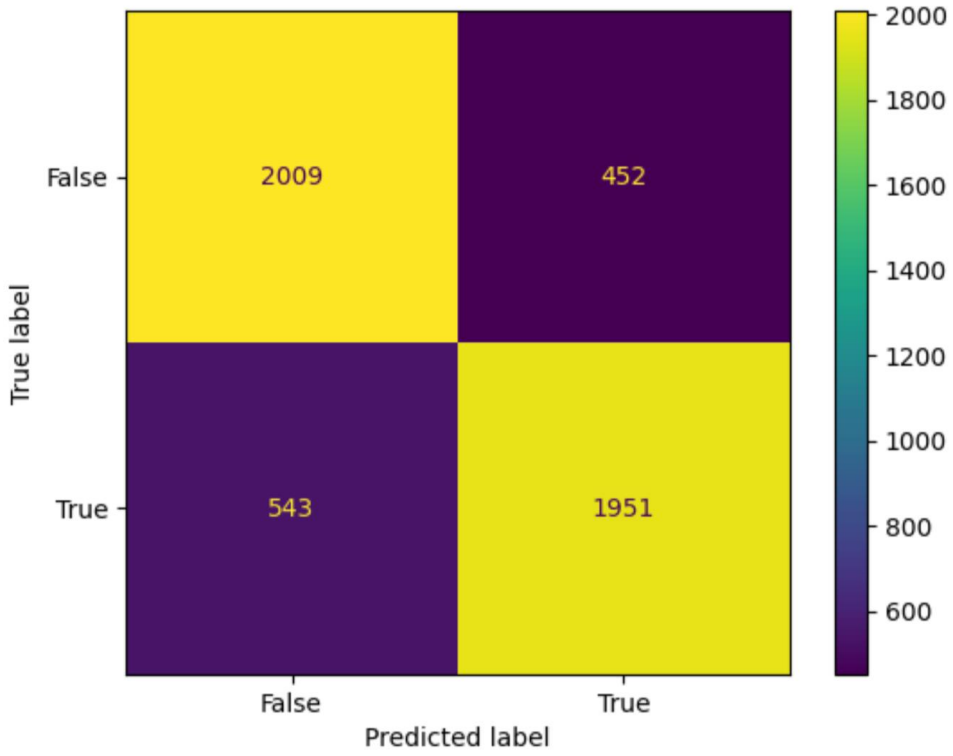
# Χαρακτηριστικά	Ακρίβεια Train	Λογ.Απώλεια Train	Ακρίβεια Val	Λογ. Απώλεια Val	n_estimators	min samples leaf	min samples split	max depth
63	79.2 %	0.476	78.7 %	0.497	70	20	5	5

Πίνακας 7.8: Αποτελέσματα του Μοντέλου, με βάση όλα τα χαρακτηριστικά που υπάρχουν στο Σύνολο Δεδομένων

# Χαρακτηριστικά	Ακρίβεια Train	Λογ.Απώλεια Train	Ακρίβεια Test	Λογ. Απώλεια Test	n_estimators	min samples leaf	min samples split	max depth
40	79.8%	0.466	79.7 %	0.472	70	20	10	5

Πίνακας 7.9: Τελικό μοντέλο Τυχαίου Δάσους

Για το τελικό μοντέλο προκύπτει και ο παρακάτω πίνακας σύγχυσης (confusion matrix) [Σχ.7.2] :



Σχήμα 7.2: Πίνακας Σύγχυσης για το τελικό μοντέλο του αλγορίθμου Τυχαίου Δάσους

7.3 Μηχανές Υποστήριξης Διανυσμάτων

Παρακάτω παρουσιάζεται η πειραματική ανάλυση του Μοντέλου Μηχανής Υποστήριξης Διανυσμάτων (Support Vector Machine) για την πρόβλεψη του αποτελέσματος των αγώνων αντισφαίρισης. Η μελέτη επικεντρώνεται στη διερεύνηση διαφόρων πτυχών του μοντέλου, συμπεριλαμβανομένης της κανονικοποίησης, της κλίμακας δεδομένων, της επιλογής χαρακτηριστικών και του συντονισμού υπερπαραμέτρων. Τα πειραματικά αποτελέσματα υπογραμμίζουν την επίδραση αυτών των παραγόντων στην ακρίβεια και την απόδοση του μοντέλου.

Το μοντέλο αρχικά εκπαιδεύεται και επικυρώνεται χρησιμοποιώντας και τα 63 χαρακτηριστικά του συνόλου δεδομένων [Πιν.7.10]. Τα ληφθέντα αποτελέσματα δείχνουν ακρίβεια εκπαίδευσης 66,6% και ακρίβεια επικύρωσης 59,5%. Αυτά τα αποτελέσματα δείχνουν ότι η απόδοση του μοντέλου δεν είναι η βέλτιστη.

# Χαρακτηριστικά	Ακρίβεια Train	Λογ.Απώλεια Train	Ακρίβεια Val	Λογ. Απώλεια Val
63	73.7 %	0.524	71.6 %	0.557

Πίνακας 7.10: Αποτελέσματα του Μοντέλου, με βάση όλα τα χαρακτηριστικά που υπάρχουν στο Σύνολο Δεδομένων

Για να βελτιωθεί η απόδοση του μοντέλου, εφαρμόζεται κανονικοποίηση προσαρμόζοντας την υπερπαράμετρο C. Δοκιμάζονται διάφορες τιμές του C [Πιν.7.11], που βρίσκονται στο σύνολο [0.1, 1, 10, 100, 150, 200, 250, 300, 350].

C	Ακρίβεια Train	Ακρίβεια Val
0.1	65.7 %	58.7 %
1	66.6 %	59.5 %
10	68.4 %	61.9 %
50	69.6 %	65.4 %
100	70.4 %	67.3 %
150	70.9 %	68.4 %
200	71.3 %	69.3 %
250	71.8 %	70.4 %
300	72.2 %	71.4 %
350	72.5 %	72.2 %

Πίνακας 7.11: Αποτελέσματα Δοκιμής υπερπαραμέτρου C

Τα αποτελέσματα καταδεικνύουν μια σταδιακή βελτίωση τόσο στην ακρίβεια εκπαίδευσης όσο και στην ακρίβεια επικύρωσης καθώς το C αυξάνεται. Η παράμετρος τακτοποίησης με την καλύτερη απόδοση βρέθηκε ότι είναι C = 350, με ακρίβεια εκπαίδευσης 72,5% και ακρίβεια επικύρωσης 72,2%. [Πιν. 7.12]

C	Ακρίβεια Train	Ακρίβεια Val
350	72.5 %	72.2 %

Πίνακας 7.12: Αποτελέσματα Δοκιμής υπερπαραμέτρου C

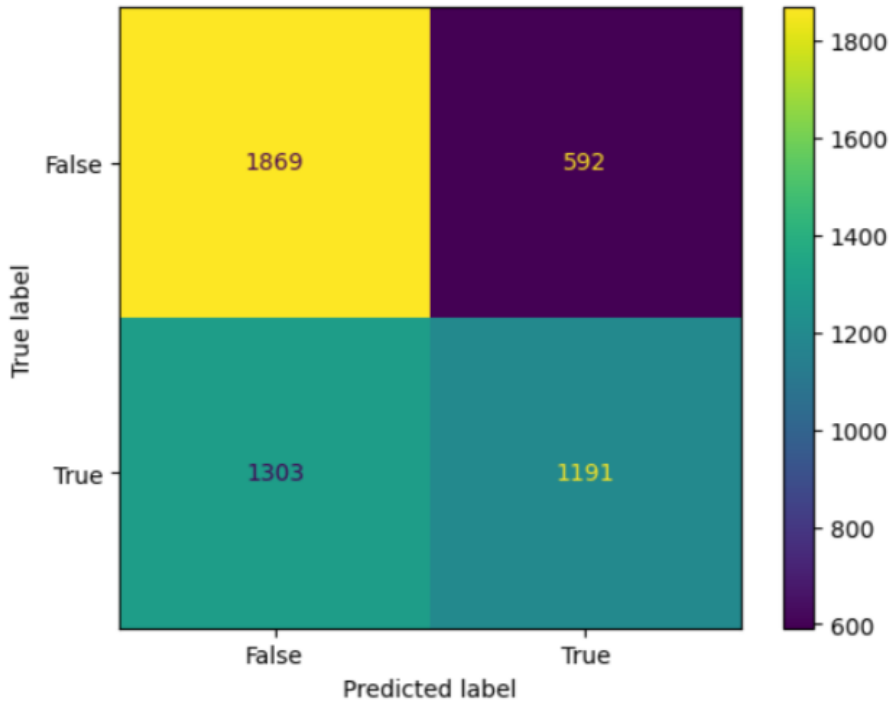
Η διαδικασία επιλογής χαρακτηριστικών απαιτεί πολλούς υπολογιστικούς πόρους, καθώς και πολύ υψηλούς χρόνους εκπαίδευσης, γεγονός που την καθιστά πολύ περιοριστική για την συγκεκριμένη διπλωματική εργασία.

Η απόδοση του τελικού μοντέλου αξιολογήθηκε στο ανεξάρτητο σύνολο δοκιμής για την αξιολόγηση των δυνατοτήτων γενίκευσής του. Τα αποτελέσματα αποκάλυψαν μια ακρίβεια δοκιμής 79.7%, παρέχοντας στοιχεία για την αποτελεσματικότητα του προτεινόμενου μοντέλου τυχαίων δασών στην πρόβλεψη του αποτελέσματος των αγώνων αντισφαίρισης [Πιν 7.13].

Ακρίβεια Train	Ακρίβεια Test
74.5 %	73.4 %

Πίνακας 7.12: Τελικό Μοντέλο

Για το τελικό μοντέλο προκύπτει και ο παρακάτω πίνακας σύγχυσης (confusion matrix) [Σχ.7.3] :



Σχήμα 7.3: Πίνακας Σύγχυσης για το τελικό μοντέλο του αλγορίθμου Τυχαίου Δάσους

7.4 Τεχνητό Νευρωνικό Δίκτυο

Παρακάτω παρουσιάζεται η πειραματική ανάλυση των Τεχνητών Νευρωνικών Δικτύων (Artificial Neural Networks) για την πρόβλεψη του αποτελέσματος των αγώνων αντισφαίρισης. Η μελέτη επικεντρώνεται στη διερεύνηση διαφόρων πτυχών του μοντέλου, συμπεριλαμβανομένης της κλίμακας δεδομένων, του συντονισμού παραμέτρων και της αξιολόγησης απόδοσης. Τα πειραματικά αποτελέσματα παρέχουν πληροφορίες για τη βελτιστοποίηση της ακρίβειας και της ικανότητας γενίκευσης του μοντέλου.

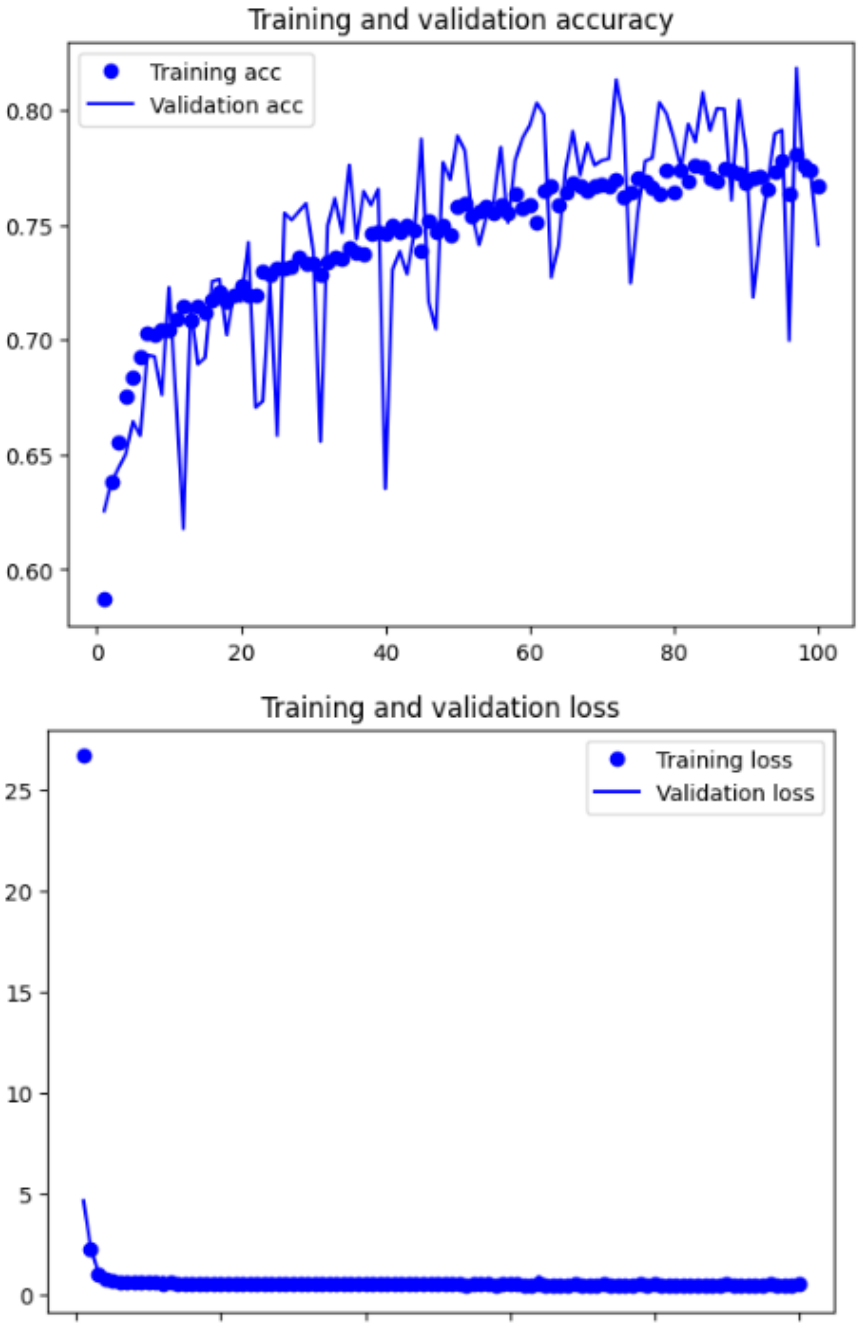
Το σύνολο δεδομένων που χρησιμοποιείται για την εκπαίδευση του μοντέλου αποτελείται από 63 χαρακτηριστικά που σχετίζονται με τα χαρακτηριστικά του αγώνα αντισφαίρισης. Το μοντέλο έχει διαμορφωθεί με ένα κρυφό επίπεδο και έναν καθορισμένο αριθμό εποχών, μέγεθος παρτίδας, βήματα ανά εποχή, βήματα επικύρωσης και βήματα δοκιμής.Αρχικά, το μοντέλο εκπαιδεύεται και αξιολογείται χωρίς καμία κλιμάκωση δεδομένων [Σχ.7.4].

Στη συνέχεια, το σύνολο δεδομένων κλιμακώνεται χρησιμοποιώντας το StandardScaler και το μοντέλο επανεκπαιδεύεται και αξιολογείται [Σχ. 7.5]. Τα αποτελέσματα δείχνουν βελτίωση. Τα αποτελέσματα φαίνονται στον Πίνακα 7.13.

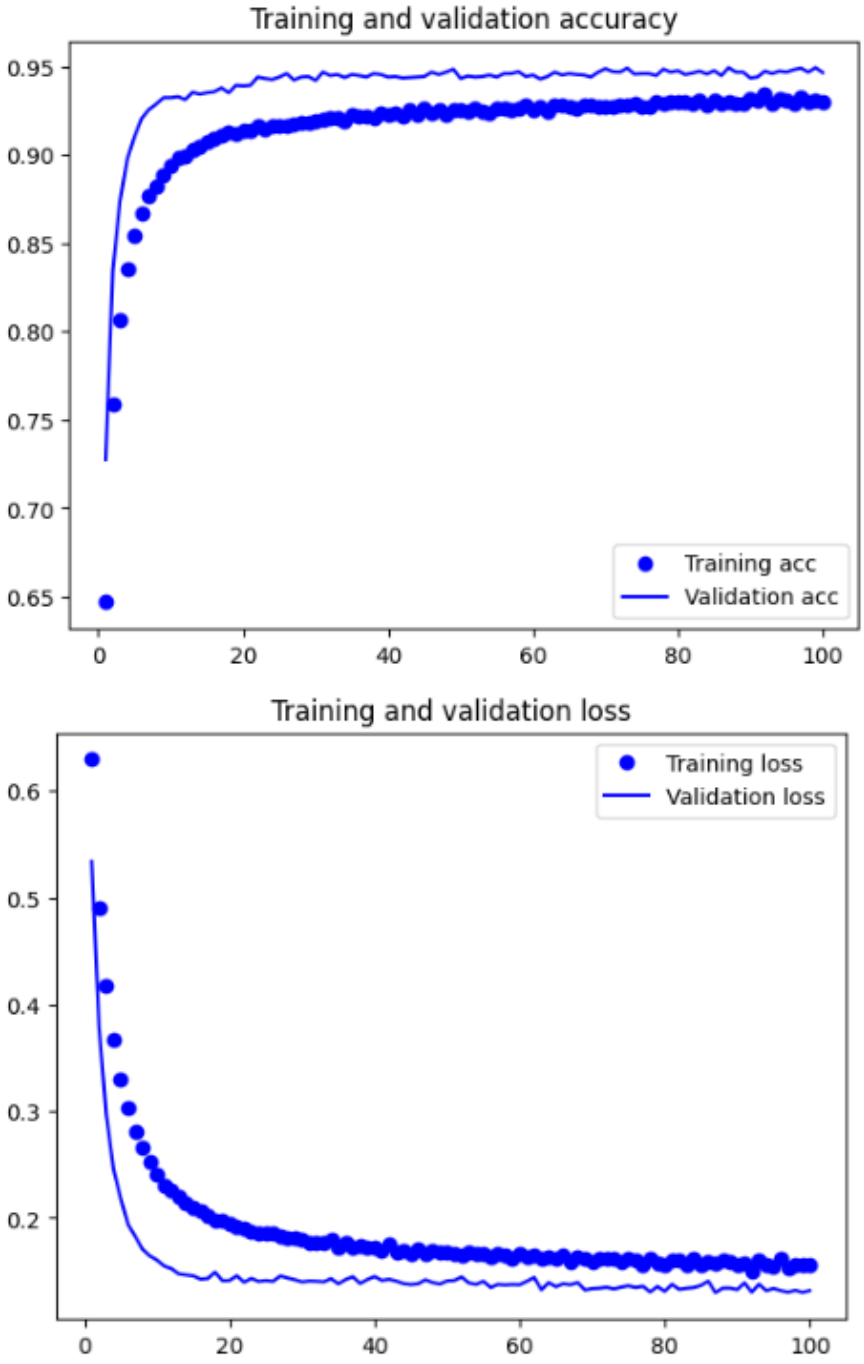
Κλιμάκωση Δεδομένων	**Ακρίβεια Train**	**Λογ.Απώλεια Train**	**Ακρίβεια Val**	**Λογ. Απώλεια Val**
Οχι	78 %	0.484	81.8 %	0.436
Ναι	92 %	0.157	94.9 %	0.130

Πίνακας 7.13: Αποτελέσματα του Μοντέλου, με και χωρίς χρήση κλιμάκωσης δεδομένων και 10 νευρώνες στο κρυφό επίπεδο

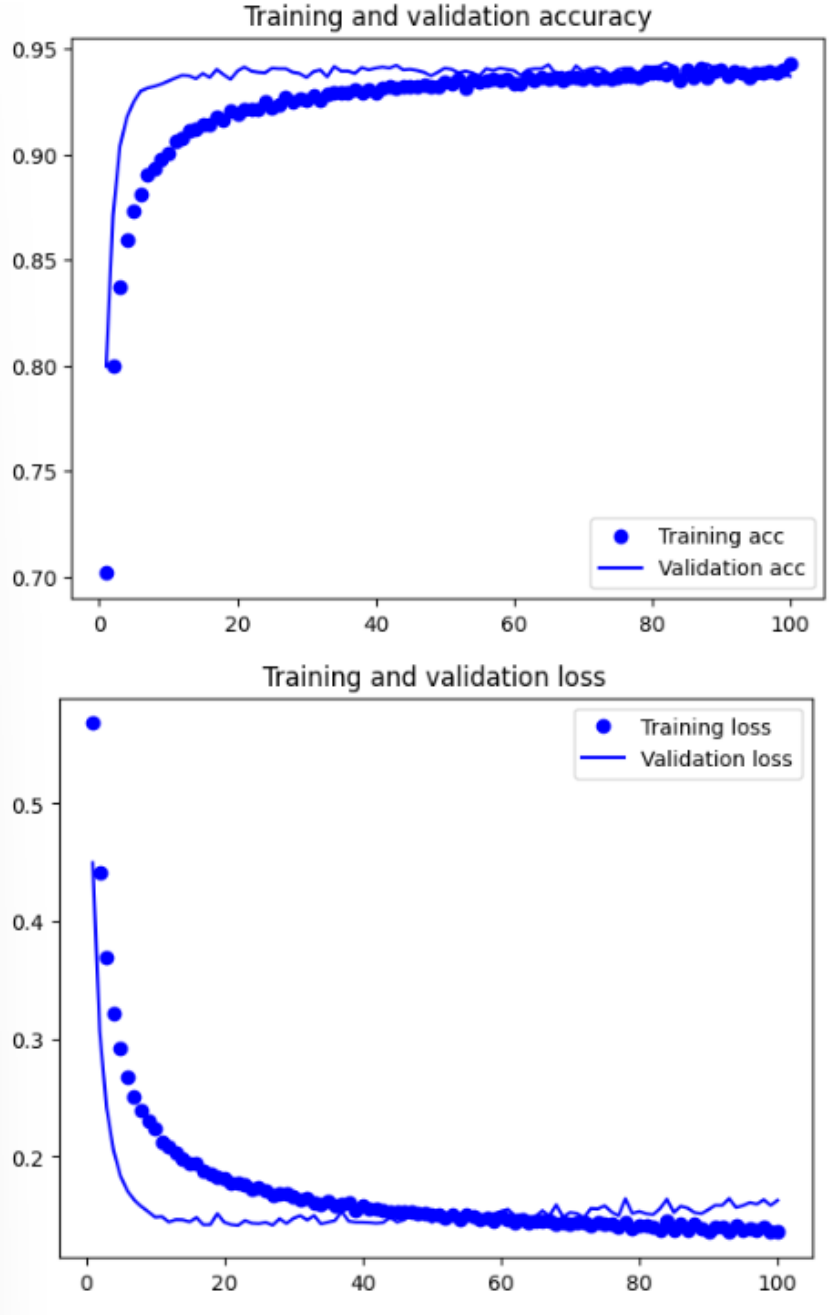
Στο στάδιο της ρύθμισης των υπερπαραμέτρων του μοντέλου πραγματοποιείται έλεγχος για τις τιμές του αριθμού των κόμβων στο κρυφό επίπεδο, του ποσοστού εγκατάλειψης, του ρυθμιστή κανονικοποίησης και της πρόωρης διακοπής. Πιο συγκεκριμένα έχουμε ότι ο αντίκτυπος του αριθμού των κόμβων στο κρυφό επίπεδο διερευνάται με τη δοκιμή τιμών των 10, 20, 30, 40, 50, 60, 70, 80, 90 και 100 κόμβων. Τα αποτελέσματα φαίνονται τόσο συγκεντρωτικά στον Πίνακα 7.14, όσο και πιο αναλυτικά για κάθε περίπτωση στα Σχήματα 7.6, 7.7, 7.8, 7.9, 7.10, 7.11.



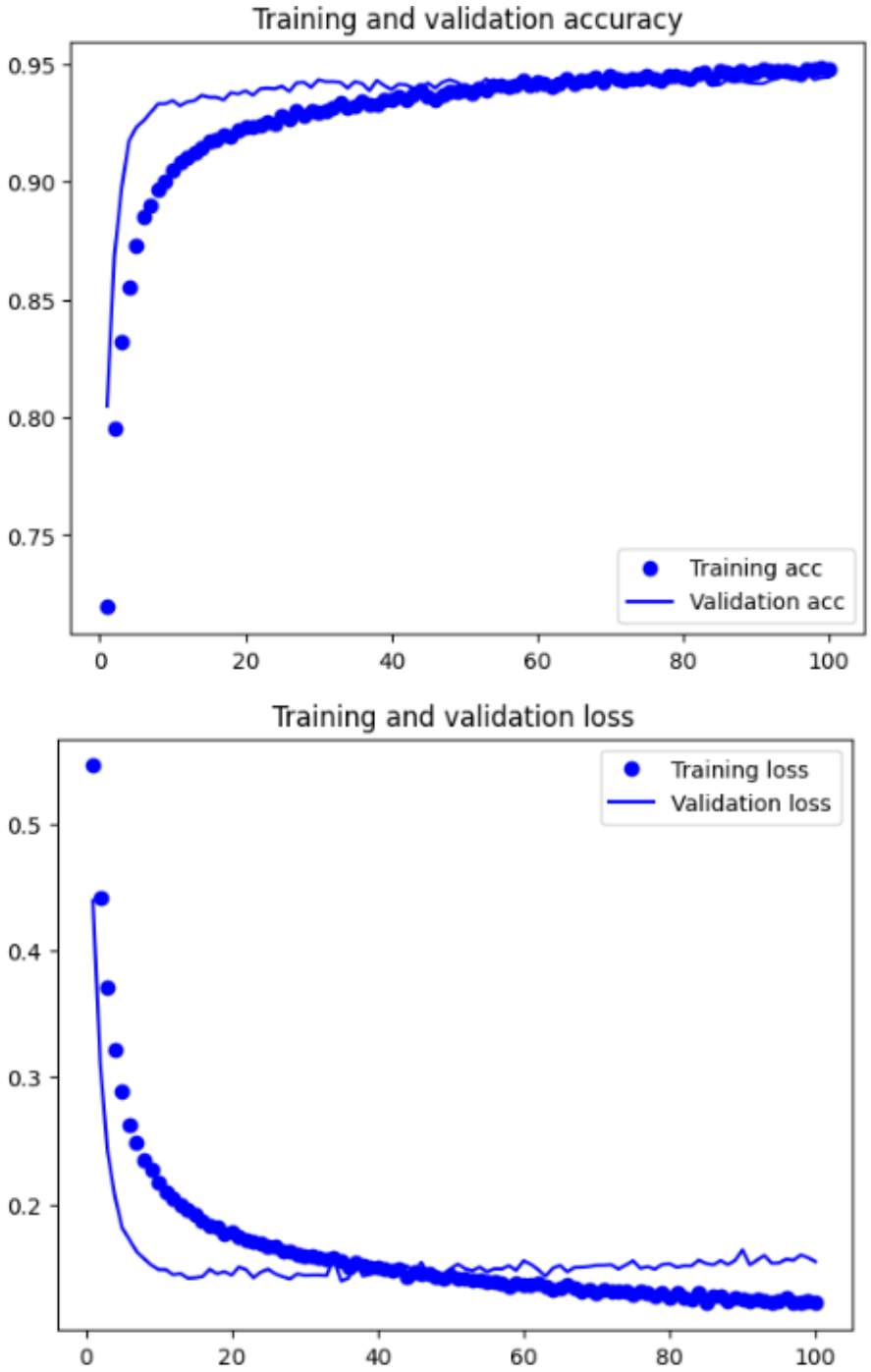
Σχήμα 7.4: Ακρίβεια και Λογιστική Απώλεια του Τεχνητού Νευρωνικού Δικτύου χωρίς χρήση κλιμάκωσης και 10 νευρώνες στο κρυφό επίπεδο.



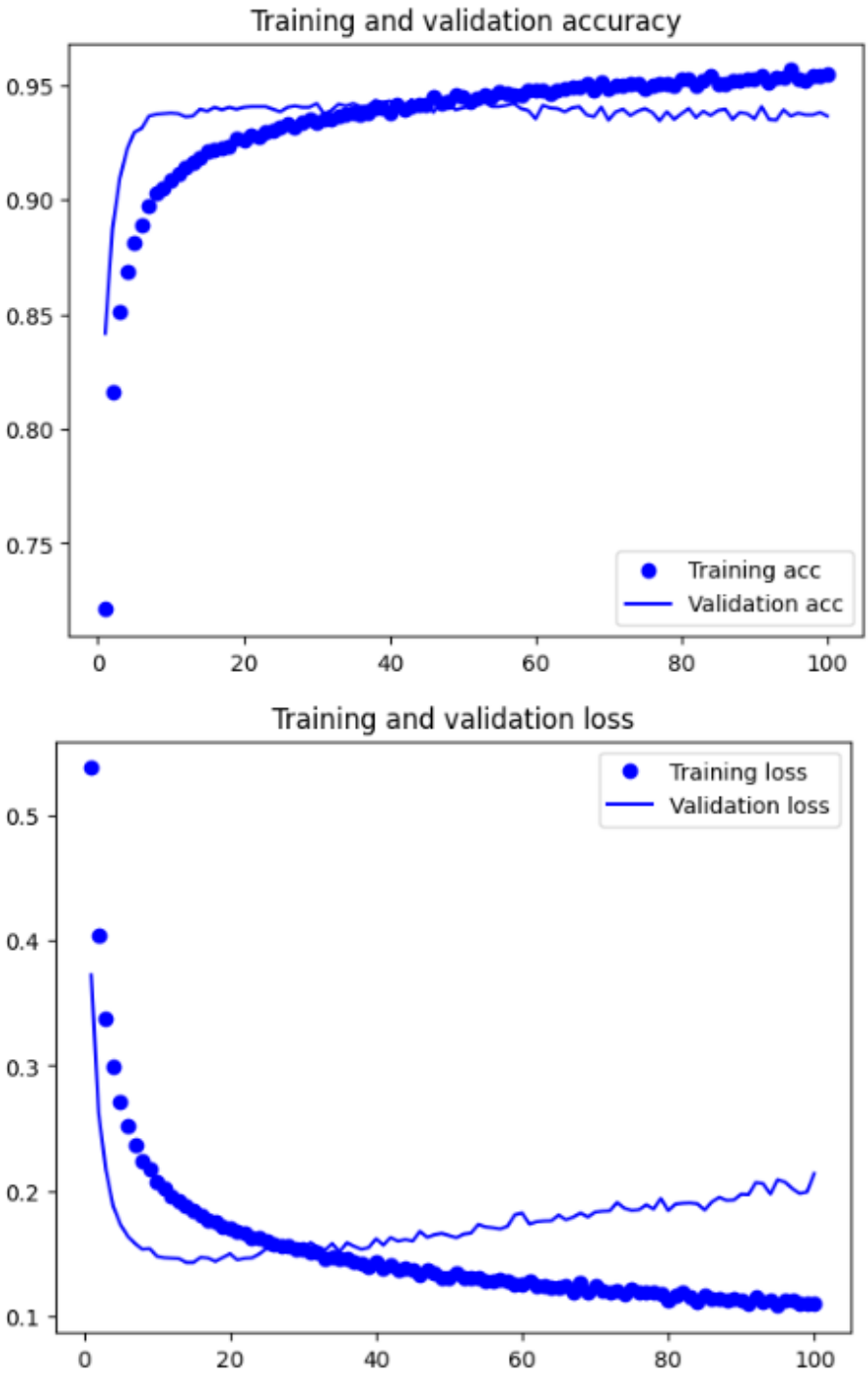
Σχήμα 7.5: Ακρίβεια και Λογιστική Απώλεια του Τεχνητού Νευρωνικού Δικτύου με χρήση κλιμάκωσης και 10 νευρώνες στο κρυφό επίπεδο.



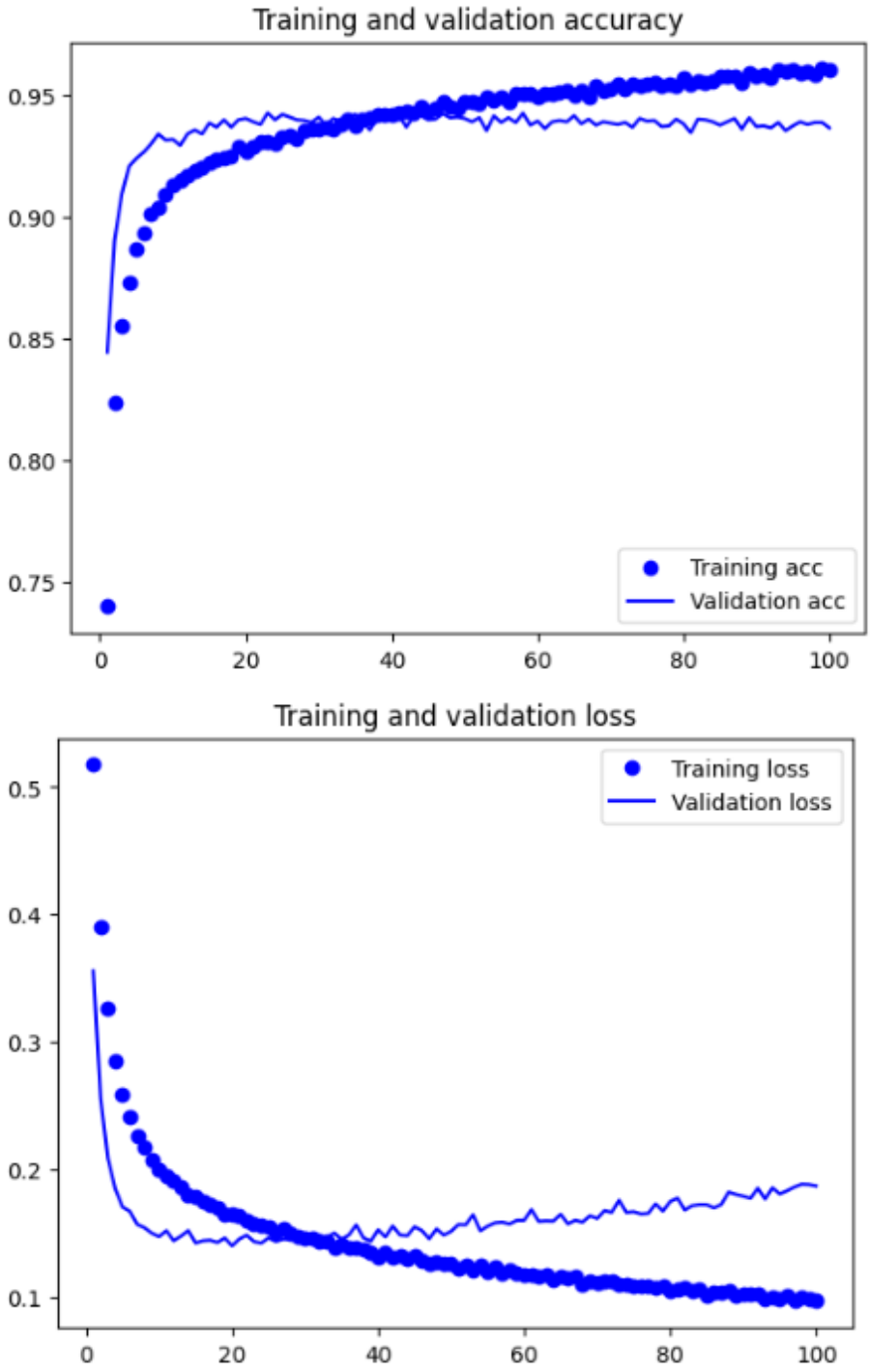
Σχήμα 7.6: Ακρίβεια και Λογιστική Απώλεια του Τεχνητού Νευρωνικού Δικτύου με χρήση κλιμάκωσης και 20 νευρώνες στο κρυφό επίπεδο.



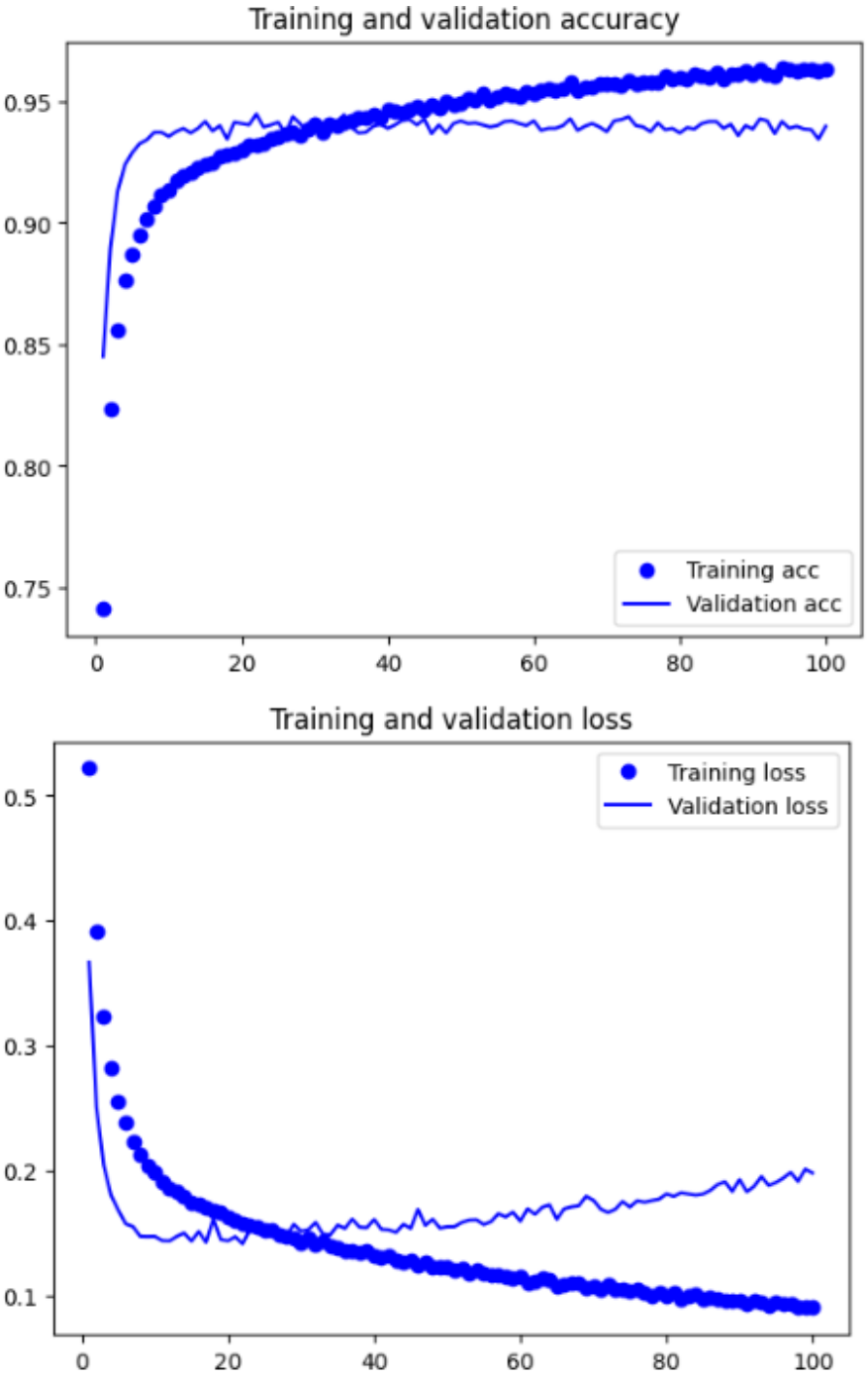
Σχήμα 7.7: Ακρίβεια και Λογιστική Απώλεια του Τεχνητού Νευρωνικού Δικτύου με χρήση κλιμάκωσης και 30 νευρώνες στο κρυφό επίπεδο.



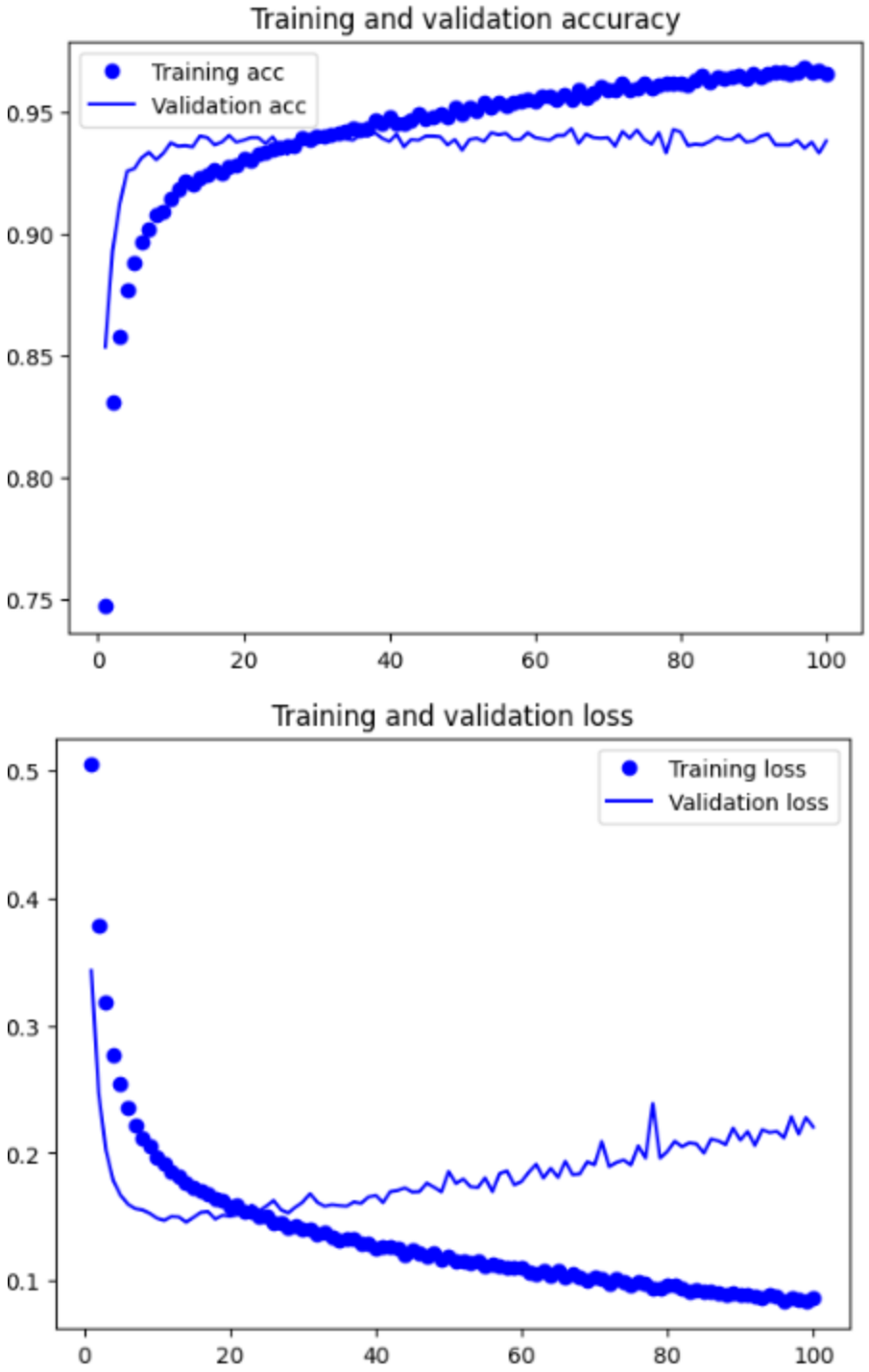
Σχήμα 7.8: Ακρίβεια και Λογιστική Απώλεια του Τεχνητού Νευρωνικού Δικτύου με χρήση κλιμάκωσης και 40 νευρώνες στο κρυφό επίπεδο.



Σχήμα 7.9: Ακρίβεια και Λογιστική Απώλεια του Τεχνητού Νευρωνικού Δικτύου με χρήση κλιμάκωσης και 50 νευρώνες στο κρυφό επίπεδο.



Σχήμα 7.10: Ακρίβεια και Λογιστική Απώλεια του Τεχνητού Νευρωνικού Δικτύου με χρήση κλιμάκωσης και 60 νευρώνες στο κρυφό επίπεδο.



Σχήμα 7.11: Ακρίβεια και Λογιστική Απώλεια του Τεχνητού Νευρωνικού Δικτύου με χρήση κλιμάκωσης και 70 νευρώνες στο κρυφό επίπεδο.

Τα αποτελέσματα δείχνουν ότι η χρήση 10, 20 και 30 κόμβων [Σχ.7.3] επιτυγχάνουν καλύτερη απόδοση. Ωστόσο, χρησιμοποιώντας πάνω απο 30 κόμβους βλέπουμε ότι η εκπαίδευση καταλήγει σε υπερπροσαρμογή και όσο προχωράνε οι εποχές η απόδοση και η λογιστική απώλεια του συνόλου δοκιμής και συνόλου αξιολόγησης δεν συμβαδίζουν.

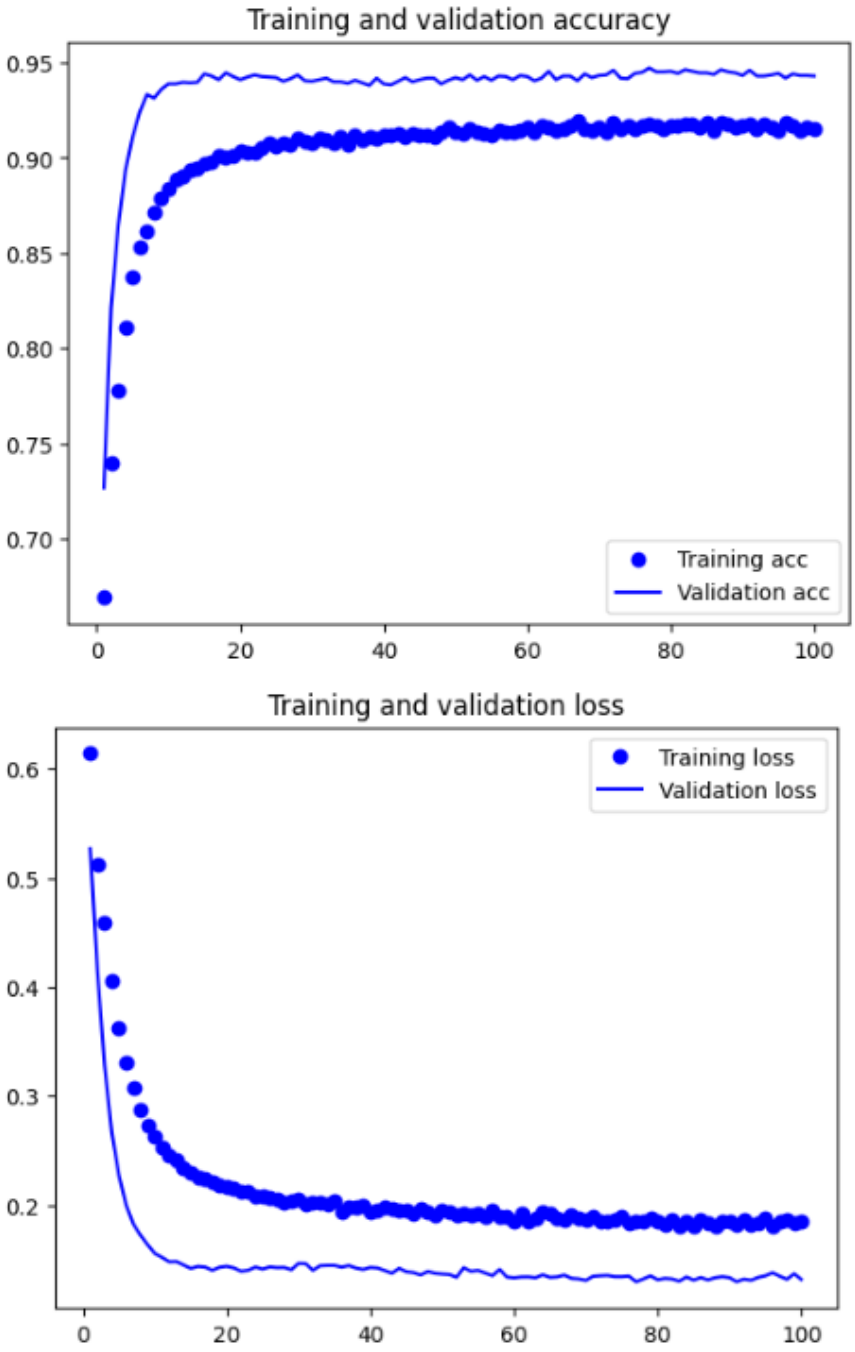
Αριθμός Νευρώνων	Ακρίβεια Train	Λογ.Απώλεια Train	Ακρίβεια Val	Λογ. Απώλεια Val
10	92.7 %	0.157	94.9 %	0.130
20	92.1 %	0.177	93.9 %	0.141
30	93.2 %	0.154	94.2 %	0.139
40	91.8 %	0.188	93.9 %	0.142
50	92.7 %	0.165	94.0 %	0.140
60	93.2 %	0.157	94.4 %	0.141
70	92.3 %	0.176	94.0 %	0.145

Πίνακας 7.14: Αποτελέσματα του Μοντέλου, με χρήση κλιμάκωσης δεδομένων και ρύθμιση του αριθμού των νευρώνων στο κρυφό επίπεδο

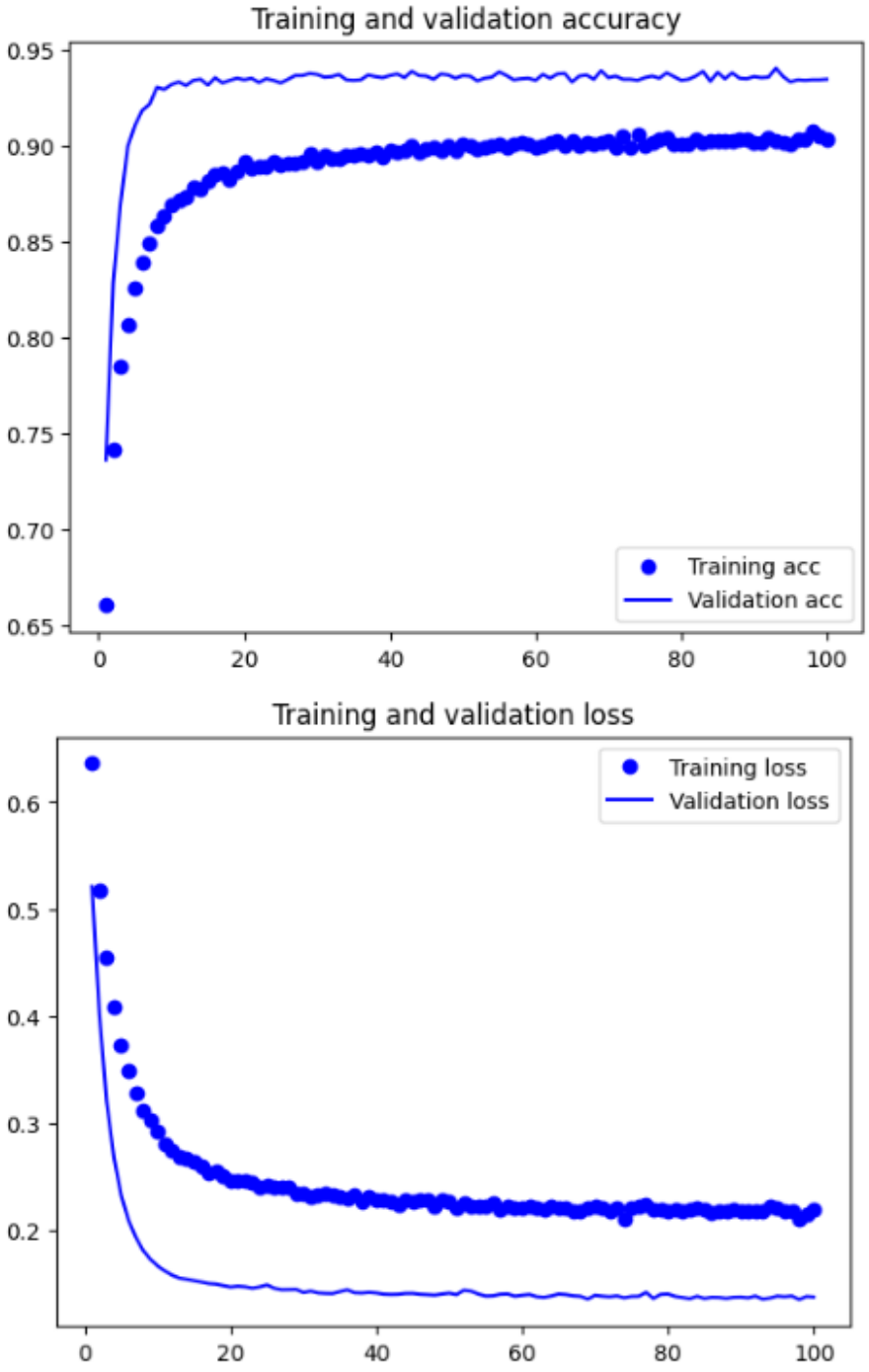
Διαφορετικά ποσοστά εγκατάλειψης (0,1, 0,2, 0,3, 0,4 και 0,5) ελέγχονται για το κρυφό στρώμα με 10, 20 και 30 κόμβους. Τα αποτελέσματα φαίνονται στους Πίνακες 7.15, 7.16 και 7.17. Παρατίθενται επίσης και τα διαγράμματα από την ακρίβεια και την λογιστική απώλεια των συνδυασμών για 10 νευρώνες στο κρυφό επίπεδο. Ανάλογα είναι τα αποτελέσματα και για τους 20 και 30 νευρώνες. Φαίνεται ότι όσο αυξάνεται το ποσοστό εγκατάλειψης, τόσο χειροτερεύει η απόδοση του μοντέλου.

Αριθμός Νευρώνων	Ποσοστό Εγκατάλειψης	Ακρίβεια Train	Λογ.Απώλεια Train	Ακρίβεια Val	Λογ. Απώλεια Val
10	0.1	91.8 %	0.185	94.7 %	0.129
10	0.2	90.7 %	0.210	93.4 %	0.135
10	0.3	88.4 %	0.236	93.1 %	0.143
10	0.4	86.0 %	0.296	92.1 %	0.160
10	0.5	83.2 %	0.325	91.6 %	0.169

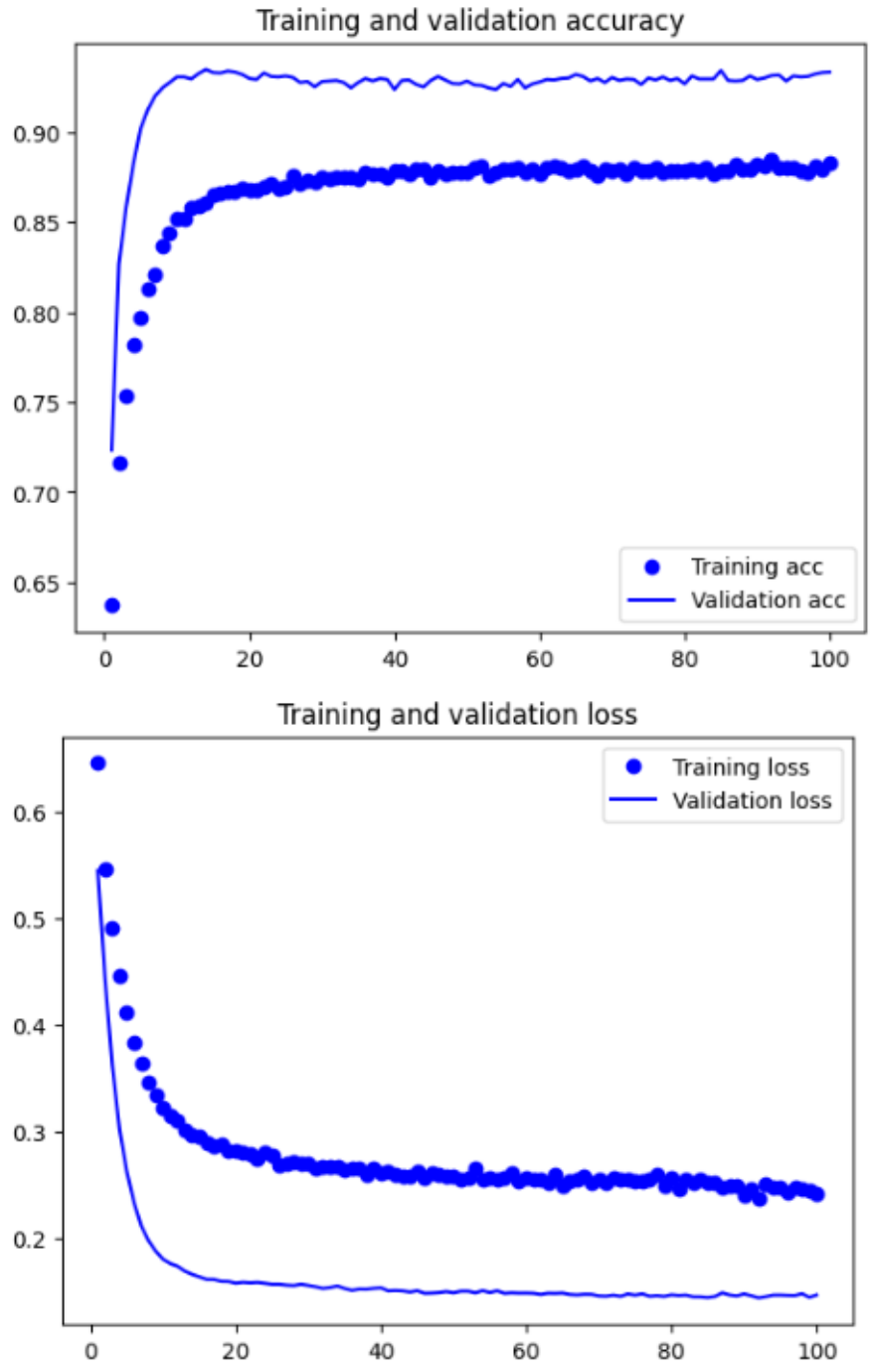
Πίνακας 7.15: Αποτελέσματα του Μοντέλου, με χρήση κλιμάκωσης δεδομένων, 10 νευρώνες στο κρυφό επίπεδο και ρύθμιση διαφορετικών τιμών ποσοστού εγκατάλειψης



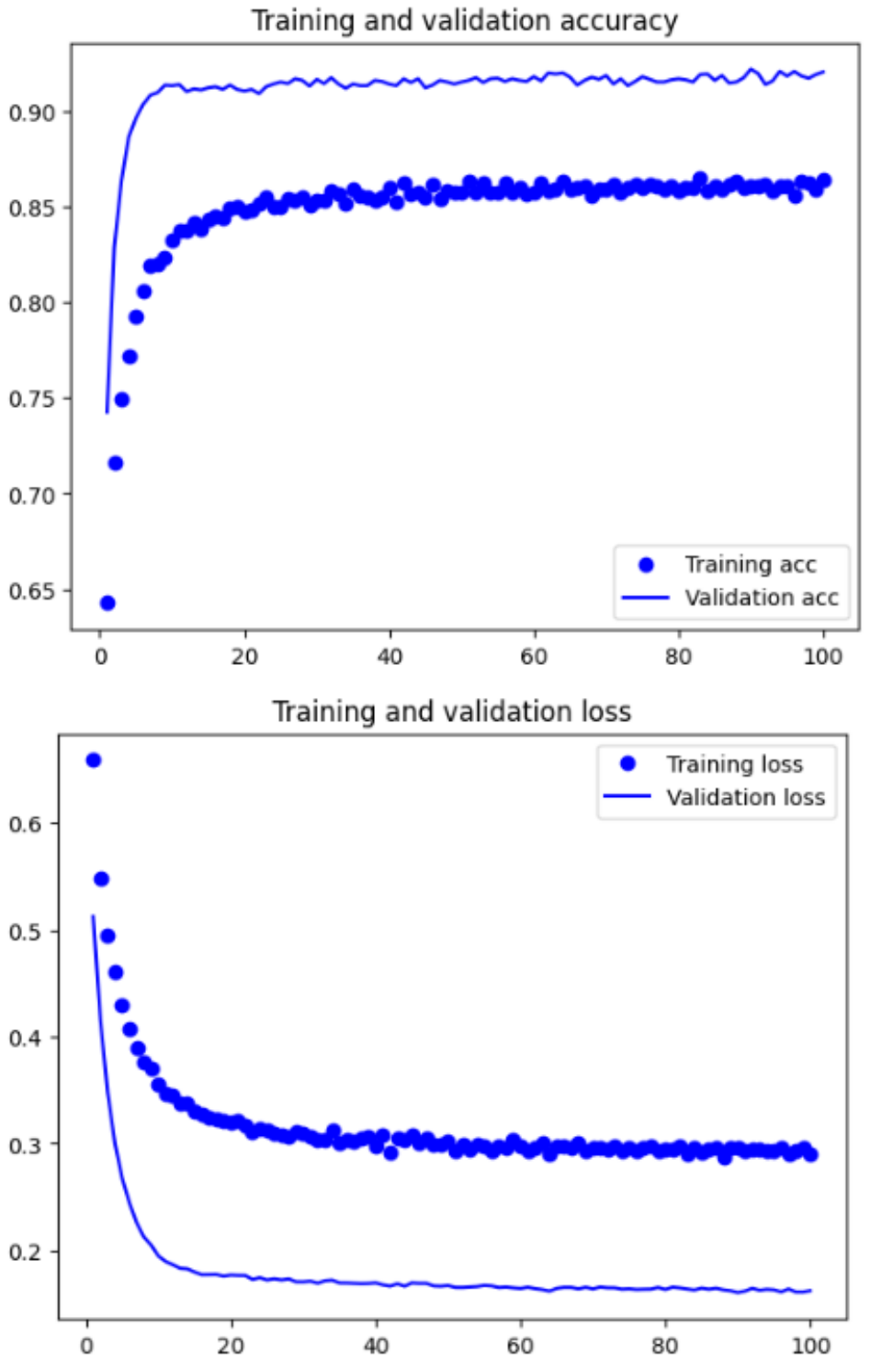
Σχήμα 7.12: Ακρίβεια και Λογιστική Απώλεια του Τεχνητού Νευρωνικού Δικτύου με χρήση κλιμάκωσης, 10 νευρώνες στο κρυφό επίπεδο και ποσοστό εγκατάλειψης 0.1.



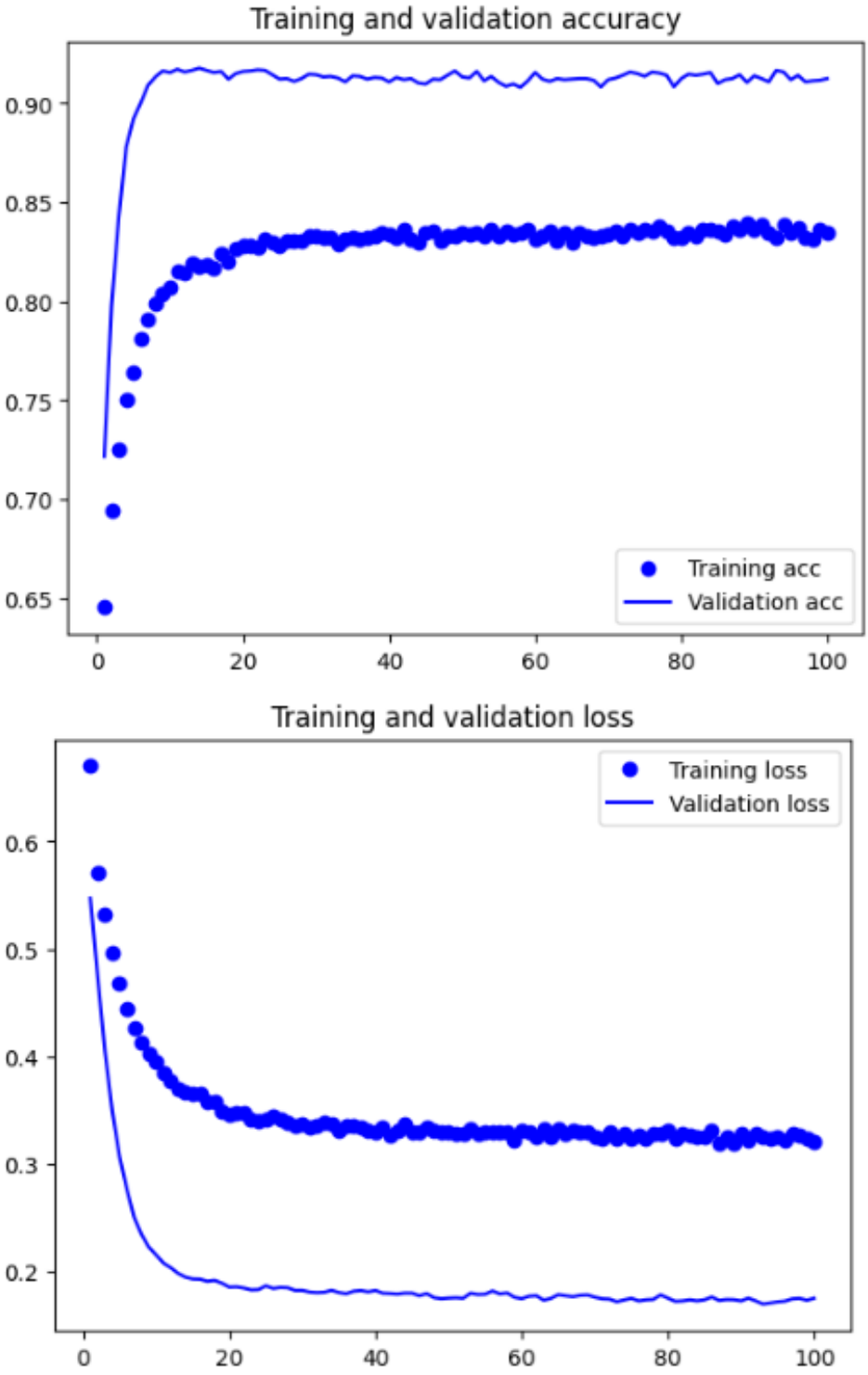
Σχήμα 7.13: Ακρίβεια και Λογιστική Απώλεια του Τεχνητού Νευρωνικού Δικτύου με χρήση κλιμάκωσης, 10 νευρώνες στο κρυφό επίπεδο και ποσοστό εγκατάλειψης 0.2.



Σχήμα 7.14: Ακρίβεια και Λογιστική Απώλεια του Τεχνητού Νευρωνικού Δικτύου με χρήση κλιμάκωσης, 10 νευρώνες στο κρυφό επίπεδο και ποσοστό εγκατάλειψης 0.3



Σχήμα 7.15: Ακρίβεια και Λογιστική Απώλεια του Τεχνητού Νευρωνικού Δικτύου με χρήση κλιμάκωσης, 10 νευρώνες στο κρυφό επίπεδο και ποσοστό εγκατάλειψης 0.4



Σχήμα 7.16: Ακρίβεια και Λογιστική Απώλεια του Τεχνητού Νευρωνικού Δικτύου με χρήση κλιμάκωσης, 10 νευρώνες στο κρυφό επίπεδο και ποσοστό εγκατάλειψης 0.5.

Αριθμός Νευρώνων	Ποσοστό Εγκατάλειψης	Ακρίβεια Train	Λογ.Απώλεια Train	Ακρίβεια Val	Λογ. Απώλεια Val
20	0.1	93.3 %	0.153	94.8 %	0.128
20	0.2	92.3 %	0.172	95.1 %	0.124
20	0.3	91.5 %	0.187	94.4 %	0.131
20	0.4	90.4 %	0.213	94.5 %	0.129
20	0.5	89.3 %	0.229	94.0%	0.139

Πίνακας 7.16: Αποτελέσματα του Μοντέλου, με χρήση κλιμάκωσης δεδομένων, 20 νευρώνες στο κρυφό επίπεδο και ρύθμιση διαφορετικών τιμών ποσοστού εγκατάλειψης

Αριθμός Νευρώνων	Ποσοστό Εγκατάλειψης	Ακρίβεια Train	Λογ.Απώλεια Train	Ακρίβεια Val	Λογ. Απώλεια Val
30	0.1	93.5 %	0.151	94.5 %	0.132
30	0.2	92.9 %	0.162	94.8 %	0.128
30	0.3	92.1 %	0.178	94.8 %	0.127
30	0.4	91.6 %	0.187	94.6 %	0.124
30	0.5	91.1 %	0.196	94.5%	0.131

Πίνακας 7.17: Αποτελέσματα του Μοντέλου, με χρήση κλιμάκωσης δεδομένων, 30 νευρώνες στο κρυφό επίπεδο και ρύθμιση διαφορετικών τιμών ποσοστού εγκατάλειψης

Η τελική αξιολόγηση πραγματοποιείται συνδυάζοντας το σύνολο εκπαίδευσης με το σύνολο επικύρωσης και το μοντέλο δοκιμάζεται στο ανεξάρτητο σύνολο δοκιμών. Τα ληφθέντα αποτελέσματα δείχνουν απώλεια 0,137 και ακρίβεια 93,9% στο σύνολο δοκιμών, υποδεικνύοντας την ικανότητα του μοντέλου να γενικεύει και να κάνει ακριβείς προβλέψεις σε αόρατα δεδομένα [Πιν.7.18].

Αριθμός Νευρώνων	Ποσοστό Εγκατάλειψης	Ακρίβεια Train	Λογ.Απώλεια Train	Ακρίβεια Val	Λογ. Απώλεια Val
20	0.1	92.8 %	0.160	94.0 %	0.130

Πίνακας 7.18: Τελικό Μοντέλο

Η πειραματική ανάλυση του μοντέλου Τεχνητού Νευρωνικού Δικτύου για την πρόβλεψη των αποτελεσμάτων του αγώνα αντισφαίρισης καταδεικνύει τη σημασία της κλίμακας δεδομένων και του συντονισμού των παραμέτρων.

8 *Συζήτηση*

8.1 Συμπεράσματα

Αυτό το κεφάλαιο παρουσιάζει τα αποτελέσματα που προέκυψαν από την εφαρμογή απλών αλγορίθμων Μηχανικής Μάθησης για την πρόβλεψη των αποτελεσμάτων των αγώνων αντισφαίρισης. Η έρευνα χρησιμοποίησε ένα μεγάλο σύνολο δεδομένων που περιλαμβάνει τόσο προϋπάρχοντα χαρακτηριστικά όσο και δημιουργημένα. Τέσσερις διαφορετικοί αλγόριθμοι χρησιμοποιήθηκαν για σκοπούς πρόβλεψης, δηλαδή Λογιστικής Παλινδρόμησης, Τυχαίων Δασών, Μηχανές Διανυσμάτων Υποστήριξης και ένα απλό Τεχνητό Νευρωνικό Δίκτυο.

Για την αξιολόγηση της απόδοσης αυτών των αλγορίθμων, χρησιμοποιήθηκαν διάφορες μετρικές, όπως η ακρίβεια, η λογιστική απώλεια και ο πίνακας σύγχυσης. Το σύνολο δεδομένων χωρίστηκε σε σύνολα εκπαίδευσης, επικύρωσης και δοκιμής για την αξιολόγηση των προγνωστικών δυνατοτήτων των μοντέλων. Τεχνικές διασταυρούμενης επικύρωσης, όπως η διασταυρούμενη επικύρωση k-fold, χρησιμοποιήθηκαν επίσης για να εξασφαλιστεί η ευρωστία και να μετριαστεί η υπερπροσαρμογή.

Τα αποτελέσματα έδειξαν διαφορετικά επίπεδα επιτυχίας στους διάφορους αλγορίθμους. Η Λογιστική Παλινδρόμηση πέτυχε ακρίβεια περίπου 71.6%, ξεπερνώντας το 70% προηγούμενων ερευνών. Το Τυχαίο Δάσος περίπου 80% και η Μηχανή Διανυσμάτων Υποστήριξης περίπου 75%. Από την άλλη, το Τεχνητό Νευρωνικό Δίκτυο πέτυχε ακρίβεια επίσης κοντά στο 92%. [Πιν. 8.1]

Μοντέλο	Ακρίβεια Train	Ακρίβεια Test
Λογιστική Παλινδρόμηση	73.2 %	71.6 %
Τυχαίο Δάσος	79.8%	79.7 %
Μηχανή Υποστήριξης Διανυσμάτων	74.5 %	73.4 %
Τεχνητό Νευρωνικό Δίκτυο	92.8 %	92.0 %

Πίνακας 8.1: Τελικά Αποτελέσματα

Τα παραπάνω αποτελέσματα υποδεικνύουν πολλά υποσχόμενες προγνωστικές δυνατότητες των μοντέλων που εφαρμόστηκαν. Ωστόσο, κατά τη διάρκεια της ερευνητικής διαδικασίας παρουσιάστηκαν αρκετοί περιορισμοί, οι οποίοι πρέπει να ληφθούν υπόψη.

8.2 Περιορισμοί

8.2.1 Σύνολο Δεδομένων

Η χρήση ενός συνόλου δεδομένων ανοιχτού κώδικα εισήγαγε πιθανούς περιορισμούς όσον αφορά την ποιότητα και την πληρότητα των δεδομένων. Παρόλο που έγιναν προσπάθειες για την προεπεξεργασία και τον καθαρισμό του συνόλου δεδομένων, ενδέχεται να εξακολουθούν να υπάρχουν εγγενείς προκαταλήψεις ή πληροφορίες που λείπουν. Η απόκτηση ενός πιο ολοκληρωμένου και αξιόπιστου δεδομένων μέσω αγοράς ή συνεργασίας με οργανισμούς αντισφαίρισης θα μπορούσε να βελτιώσει την ακρίβεια των προβλέψεων.

8.2.2 Υπολογιστικοί Πόροι

Οι υπολογιστικοί πόροι αποτέλεσαν περιορισμό κατά την εφαρμογή των αλγορίθμων και ιδιαίτερα των Μηχανών Διανυσμάτων Υποστήριξης (SVM). Ο χρόνος εκπαίδευσης που απαιτείται για την εκπαίδευση των μοντέλων ήταν πολύ μεγαλύτερος σε σύγκριση με τους άλλους αλγόριθμους, καθιστώντας τόσο λιγότερο πρακτικό για εφαρμογές πρόβλεψης σε πραγματικό χρόνο, όσο και ακριβό στην εκπαίδευση και στην αναζήτηση των βέλτιστων υπερπαραμέτρων. Η διερεύνηση εναλλακτικών προσεγγίσεων ή πιο αποτελεσματικών υλοποιήσεων των αλγορίθμων αυτών θα μπορούσε να βοηθήσει στην υπέρβαση αυτού του

περιορισμού. Επίσης, οι υψηλοί χρόνοι εκπαίδευσης, καθιστούν επίπονη και υπολογιστά ακριβή την εις βάθος εξερεύνηση των δυνατοτήτων που έχουν τα μοντέλα αυτά.

8.3 Μελλοντική Έρευνα

8.3.1 Αλγόριθμοι Μηχανικής Μάθησης

Η επιλογή των αλγορίθμων περιορίστηκε σε Λογιστική Παλινδρόμηση, Τυχαία Δάση, Μηχανές Διανυσμάτων Υποστήριξης και ένα απλό Τεχνητό Νευρωνικό Δίκτυο. Ενώ αυτά τα μοντέλα έδειξαν πολλά υποσχόμενα αποτελέσματα, άλλες προηγμένες τεχνικές Μηχανικής Μάθησης, όπως οι αρχιτεκτονικές Gradient Boosting ή Βαθιά Μάθηση, θα μπορούσαν να διερευνηθούν σε μελλοντική έρευνα για την περαιτέρω βελτίωση της ακρίβειας πρόβλεψης.

8.3.2 Εξαγωγή Χαρακτηριστικών

Η πλειονότητα των χαρακτηριστικών που εξάγαμε, όπως αναφέρονται στο κεφάλαιο 5, αποδείχθηκε ότι έχουν επιρροή στις προβλέψεις των μοντέλων μας. Η συνεχής εξερεύνηση και βελτίωση του συνόλου χαρακτηριστικών που χρησιμοποιείται για την πρόβλεψη θα μπορούσε να οδηγήσει σε βελτιωμένη ακρίβεια. Ο εντοπισμός πιο ενημερωτικών χαρακτηριστικών ή η ενσωμάτωση γνώσεων για συγκεκριμένο τομέα θα μπορούσε να ενισχύσει την ικανότητα των μοντέλων να καταγράφουν σχετικά μοτίβα και τάσεις. Για παράδειγμα, οι καιρικές συνθήκες (θερμοκρασία, άνεμος) μπορεί να ευνοήσουν ένα συγκεκριμένο στυλ παιχνιδιού.

8.3.3 Αγώνες Αντισφαίρισης Γυναικών

Η έρευνα μας περιορίστηκε σε αγώνες αντρικού επαγγελματικού επιπέδου (ATP), λόγω της μεγαλύτερης διαθεσιμότητας αποδόσεων για αυτούς τους αγώνες στο σύνολο δεδομένων μας. Ο κώδικας που παράχθηκε είναι αρκετά γενικός και θα μπορούσε να εφαρμοστεί και για προβλέψεις σε αγώνες αντισφαίρισης γυναικείου επαγγελματικού επιπέδου (Woman Tennis Association). Ωστόσο, καθώς διαφορετικά χαρακτηριστικά μπορεί να είναι σχετικά με την πορεία ενός αγώνα γυναικών, η εφαρμογή των μοντέλων μας σε αυτούς τους αγώνες απαιτεί εκ νέου ανάλυση χαρακτηριστικών, επιλογή χαρακτηριστικών και ρύθμιση υπερπαραμέτρων.

Αντιμετωπίζοντας αυτές τις πτυχές σε μελλοντική έρευνα, το πεδίο της πρόβλεψης αποτελεσμάτων αγώνων αντισφαίρισης με χρήση της Μηχανικής Μάθησης μπορεί να προχωρήσει, παρέχοντας πολύτιμες πληροφορίες για τους λάτρεις των σπορ, τους αναλυτές και τη βιομηχανία αθλητικών στοιχημάτων.

8.3.4 Αλγόριθμοι Συνόλου

Η έρευνα θα μπορούσε να εμβαθύνει και σε αλγορίθμου που υπολογίζουν το τελικό αποτέλεσμα πρόβλεψης χρησιμοποιώντας είτε τον μέσο όρο των προβλέψεων των μοντέλων που χρησιμοποιήθηκαν, είτε το αποτέλεσμα που έχει η πλειοψηφία των μοντέλων. Κάπως έτσι θα μπορούσε να βελτιωθεί η απόδοση του τελικού μας μοντέλου, παρέχοντας πιο αξιόπιστες προβλέψεις, καθώς διαφορετικά μοντέλα αναγνωρίζουν διαφορετικά μοτίβα μεταξύ των δεδομένων και των χαρακτηριστικών.

9 *Βιβλιογραφία*

[1] Humphrey T., “Tennis Trading on Betfair”, 2014, Trading Publications.

[2] https://www.atptour.com/en

[3] Scikit-learn, “sklearn.linear_model.LogisticRegression–scikit-learn 0.23.0 documentation” [Online]

[4] Scikit-learn, “sklearn.ensemble.RandomForestClassifier–scikit-learn 0.23.0 documentation” [Online]

[5] Scikit-learn, “sklearn.svm.SVC–scikit-learn 0.23.0 documentation” [Online]

[6] Scikit-learn, “sklearn.ensemble.VotingClassifier–scikit-learn0.23.0 documentation” [Online]

[7] Barnett T., Brown A., and Clarke S.R., “Developing a tennis model that reflects outcomes of tennis matches”, 2006 in proceedings of the 8th Australasian Conference on Mathematics and Computers in Sport, Coolangatta, Queensland, pp.178-188.

[8] Knottenbelt W.J., Spanias D., and Madurska A.M., “A Common-opponent Stochastic Model for Predicting the Outcome of Professional Tennis Matches”, 2012, Computers and Mathematics with Applications, 64, pp. 3820-3827.

[9] Scheibehenne B., and Broeder A., “Predicting Wimbledon 2005 Tennis Results by Mere Player Name Recognition”, 2007, International Journal of Forecasting, 23, pp.415-426.

[10] Somboonphokkaphan A., Phimoltares S. and Lursinsap C., “Tennis Winner Prediction Based on Time-Series History with Neural Modeling”, 2009, Proceedings of the International Multi-Conference of Engineers and Computer Scientists, Hong Kong.

[11] Del Corral J., and Prieto-Rodriguez J., “Are Differences in Ranks Good Predictors for Grand Slam Tennis Matches?”, 2010, International Journal of Forecasting, 26, pp.551-563

[12] Panjan A., Šarabon N., and Filipčič A., "Prediction of the successfulness of tennis players with machine learning methods.", 2010, Kinesiology 42.1., pp.98-106.

[13] McHale I., and Morton A., “A Bradley-Terry Type Model for Forecasting Tennis Match Results”, 2010, International Journal of Forecasting, 27, pp.619-630

[14] Lyócsa Š., Výrost T., “To bet or not to bet: a reality check for tennis betting market efficiency”, 2018, Applied Economics, 50(20), pp.2251-2272.

[15] Ma S.M., Liu C.C., Tan Y., and Ma S.C., “Winning Matches in Grand Slam Men’s Singles: An Analysis of Player Performance-related Variables from 1991 to 2008”, 2013, Journal of Sports Sciences, 31, pp. 1147-1155.

[16] Kovalchik S.A., “Searching for the GOAT of Tennis Win Prediction”, 2016, Journal of Quantitative Analysis in Sports, 12, 127-138.

[17] Lisi F., and Zanella G., “Tennis Betting: Can Statistics Beat Bookmakers?”, 2017, Electronic Journal of Applied Statistical Analysis, 10, pp. 790-808.

[18] Gu W., and Saaty T.L., “Predicting the Outcome of a Tennis Tournament: Based on Both Data and Judgments”, 2019, Journal of Systems Science and Systems Engineering, 28, pp.317-343.

[19] Ingram M., “A Point-based Bayesian Hierarchical Model to Predict the Outcome of Tennis Matches”, 2019, Journal of Quantitative Analysis in Sports, 15, pp.313-325.

[20] Gorgi P., Koopman S. J., and Lit R., “The analysis and forecasting of tennis matches by using a high dimensional dynamic model”, 2019, Journal of the Royal Statistical Society Series A: Statistics in Society, 182(4), pp.1393-1409.

[21] Candila V., and Scognamillo A., “On the Longshot Bias in Tennis Betting Markets: The Casco Normalization”, 2017, Universita Degli Studi di Salerno.

[22] De Seranno A., “Predicting Tennis Matches Using Machine Learning”, 2020.

[23] Wilkens S. “Sports prediction and betting models in the machine learning age: The case of tennis”, 2021, Journal of Sports Analytics, 7(2), pp.99-117.

[24] Wei X., Lucey P., Morgan S., and Sridharan S., “Predicting shot locations in tennis using spatiotemporal data”, 2013, In 2013 International Conference on Digital Image Computing: Techniques and Applications (DICTA) (pp. 1-8). IEEE.

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

[26] J. Sackmann, “GitHub – JeffSackmann/tennis_atp: ATP Tennis Rankings, Results, and Stats”, 2015, [Online]

[27] Tennis – Data, “Tennis Betting – Tennis Results – Tennis Odds”, 2007. [Online]

[28] https://www.itftennis.com/en/

[29] Clarke S.R., Dyte D., “Using Official Ratings to Simulate Major Tennis Tournaments”, 2000, International transaction in operational research, 7(6), pp.585-594.

[30] O’Malley A.J. “Probability formulas and statistical analysis in tennis”, 2008, Journal of Quantitative Analysis in Sports, 4(2).

[31] Koning R.H., “Home advantage in professional tennis”, (2010), Journal of Sports Sciences, 26(3), pp.551-563.

[32] Reid M., and Duffield R., “The development of fatigue during match-play tennis”, 2014, pp.7-11

[33] Williams L.V., Liu C., Dixon L., Gerard H., ”How well do Elo-based ratings predict professional tennis matches?”, 2021, Journal of Quantitative Analysis in Sports, 17(2), pp.91-105.