



ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΔΗΜΙΟΥΡΓΙΑ ΝΕΥΡΩΝΙΚΩΝ ΔΙΚΤΥΩΝ ΓΙΑ ΤΗΝ
ΠΡΟΒΛΕΨΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΜΕΛΛΟΝΤΙΚΩΝ
ΠΟΔΟΣΦΑΙΡΙΚΩΝ ΑΓΩΝΩΝ ΚΑΙ ΤΗΝ ΑΠΟΔΟΣΗ ΠΑΙΚΤΩΝ

ΣΚΑΡΛΑΤΟΣ ΗΛΙΑΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΥΠΕΥΘΥΝΟΣ

.....ΚΟΛΟΜΒΑΤΣΟΣ ΚΩΝΣΤΑΝΤΙΝΟΣ.....

.....ΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣ.....

ΣΥΝΕΠΙΒΛΕΠΩΝ
(εφόσον υπάρχει)

.....Ονοματεπώνυμο Υπεύθυνου.....

.....Βαθμίδα.....

Λαμία13 ΣΕΠΤΕΜΒΡΙΟΥ..... έτος ...2026...



UNIVERSITY OF
THESSALY

SCHOOL OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE & TELECOMMUNICATIONS

DEVELOPMENT OF NEURAL NETWORKS FOR
PREDICTING THE OUTCOMES OF FUTURE FOOTBALL
MATCHES AND PLAYER PERFORMANCE

SKARLATOS ILIAS

FINAL THESIS

ADVISOR

.....KOLOMVATSOS KONSTANTINOS.....
.....ASSOCIATE PROFESSOR.....

CO ADVISOR

.....Full Name.....
.....Position.....

LamiaSEPTEMBER 13th..... year ...2026...

«Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις ⁽¹⁾, που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί **χωρίς να τα περικλείω σε εισαγωγικά** και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάστηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.

2. Δέχομαι ότι η αυτολεξεί **παράθεση χωρίς εισαγωγικά**, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσιάσή τους ως δική μου εργασία.

3. Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια

4. Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.

Ημερομηνία: ...13.../...09.../2026.....

Ο Δηλών Σκαρλάτος Ηλίας



(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.»

ΠΕΡΙΛΗΨΗ

Η παρούσα εργασία παρουσιάζει την ανάπτυξη ενός συστήματος πρόβλεψης ποδοσφαιρικών αγώνων και ατομικής απόδοσης παικτών με χρήση νευρωνικών δικτύων και στατιστικών μεθόδων. Για την ανάπτυξη του συστήματος χρησιμοποιήθηκαν ιστορικά δεδομένα ομάδων και παικτών από τέσσερα πρωταθλήματα, το Αγγλικό, το Σκωτσέζικο, το Ιταλικό και το Ελληνικό, καλύπτοντας πέντε αγωνιστικές περιόδους, από τη σεζόν 2021-2022 έως τη σεζόν 2025-2026. Σε επίπεδο ομάδας αναπτύχθηκε συνδυαστική αρχιτεκτονική LSTM-MLP, η οποία αξιοποιεί τόσο τη βραχυπρόθεσμη αγωνιστική εικόνα των ομάδων όσο και στατιστικά χαρακτηριστικά από προηγούμενους αγώνες. Το μοντέλο εκτιμά τον αναμενόμενο αριθμό γκολ κάθε ομάδας και, βάσει των εκτιμήσεων αυτών, δημιουργούνται κατανομές Poisson, ενώ εφαρμόζεται η διόρθωση Dixon-Coles για τη βελτίωση της μοντελοποίησης αποτελεσμάτων χαμηλού σκορ. Από τις τελικές κατανομές υπολογίζονται οι πιθανότητες νίκης της γηπεδούχου, ισοπαλίας και νίκης της φιλοξενούμενης ομάδας, καθώς και το πιθανότερο ακριβές σκορ.

Σε επίπεδο παίκτη αναπτύχθηκε ξεχωριστό μοντέλο πολλαπλών εργασιών (multi-task learning), το οποίο παράγει προβλέψεις για στοιχεία της ατομικής απόδοσης, όπως γκολ, ασίστ, σουτ, λεπτά συμμετοχής, πιθανότητα συμμετοχής στην αρχική ενδεκάδα και πιθανότητες λήψης κίτρινης ή κόκκινης κάρτας.

Το σύστημα υλοποιήθηκε σε περιβάλλον γραμμής εντολών, μέσω του οποίου παρέχονται λειτουργίες πρόβλεψης για την έκβαση αγώνων και την ατομική απόδοση των παικτών. Παράλληλα, υποστηρίζεται προσομοίωση της τελικής κατάταξης ενός πρωταθλήματος μέσω Monte Carlo, εκτίμηση της πιθανότητας κατάκτησης του τίτλου και υποβιβασμού, προβλέψεις για τους παίκτες με την υψηλότερη προβλεπόμενη επιθετική απόδοση, καθώς και προβλέψεις για τους παίκτες με τη μεγαλύτερη πιθανότητα να δεχτούν κίτρινη ή κόκκινη κάρτα.

Η αξιολόγηση του μοντέλου πρόβλεψης αγώνων πραγματοποιήθηκε με walk-forward evaluation σε τέσσερα χρονικά folds. Η ακρίβεια πρόβλεψης του τελικού αποτελέσματος κυμάνθηκε από 48,9% έως 54,0%, με μέση τιμή 50,8%, έναντι της μέσης ακρίβειας 43,4% του majority-class baseline. Η μέση ακρίβεια πρόβλεψης του ακριβούς σκορ ήταν 12,3%.

ABSTRACT

This thesis presents the development of a system for predicting football match outcomes and individual player performance using neural networks and statistical methods. Historical team and player data from four leagues (the English, Scottish, Italian and the Greek leagues) were used, covering five seasons, from 2021-2022 to 2025-2026. At team level, a hybrid LSTM-MLP architecture was developed, combining short-term team form with statistical features derived from previous matches. The model estimates the expected number of goals for each team and, based on these estimates, Poisson distributions are constructed, while the Dixon-Coles adjustment is applied to improve the modelling of low-scoring outcomes. From the resulting distributions, the probabilities of a home win, draw and away win are calculated, along with the most likely exact score.

At player level, a separate multi-task learning model was developed to generate predictions for individual performance indicators, including goals, assists, shots, expected minutes played, probability of being included in the starting eleven and the probabilities of receiving a yellow or red card.

The system was implemented in a command-line environment, providing functions for predicting match outcomes and individual player performance. In addition, it supports Monte Carlo simulation of final league standings, estimation of the title and relegation probabilities, as well as predictions for the players with the highest predicted attacking performance and the players with the highest probability of receiving yellow or red card.

The match prediction model was evaluated using a four-fold walk-forward evaluation procedure. Match outcome prediction accuracy ranged from 48.9% to 54.0%, with a mean accuracy of 50.8%, compared with a mean accuracy of 43.4% for the majority-class baseline. The mean exact-score prediction accuracy was 12.3%.

Table of Contents

ΠΕΡΙΛΗΨΗ	I
ABSTRACT	II
ΚΕΦΑΛΑΙΟ 1 - ΕΙΣΑΓΩΓΗ	3
1.1 ΓΕΝΙΚΗ ΕΙΣΑΓΩΓΗ	3
1.2 ΚΙΝΗΤΡΟ ΤΗΣ ΕΡΓΑΣΙΑΣ	3
1.3 ΠΡΟΒΛΗΜΑΤΑ-ΠΕΡΙΟΡΙΣΜΟΙ	4
1.4 ΣΤΟΧΟΣ ΤΗΣ ΕΡΓΑΣΙΑΣ	4
1.5 ΔΟΜΗ ΤΗΣ ΕΡΓΑΣΙΑΣ	4
ΚΕΦΑΛΑΙΟ 2 - ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΕΠΙΣΚΟΠΗΣΗ	5
2.1 ΣΤΑΤΙΣΤΙΚΗ ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ (POISSON ΜΟΝΤΕΛΑ)	5
(ΕΝΟΤΗΤΑ 2.1.Α) ΚΑΤΑΝΟΜΗ POISSON	5
(ΕΝΟΤΗΤΑ 2.1.Β) ΜΟΝΤΕΛΟΠΟΙΗΣΗ POISSON ΚΑΤΑ MAHER	5
(ΕΝΟΤΗΤΑ 2.1.Γ) ΔΙΪΘΡΩΣΗ DIXON-COLES	6
(ΕΝΟΤΗΤΑ 2.1.Δ) ΕΞΑΓΩΓΗ ΠΙΘΑΝΟΤΗΤΩΝ ΑΠΟΤΕΛΕΣΜΑΤΟΣ	6
(ΕΝΟΤΗΤΑ 2.1.Ε) ΣΥΣΤΗΜΑ ΑΞΙΟΛΟΓΗΣΗΣ ΕΛΟ	7
2.2 ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ ΓΙΑ ΔΕΔΟΜΕΝΑ ΑΚΟΛΟΥΘΩΝ (RNN/LSTM)	8
(ΕΝΟΤΗΤΑ 2.2.Α) ΕΙΣΑΓΩΓΗ	8
(ΕΝΟΤΗΤΑ 2.2.Β) ΕΠΑΝΑΛΑΜΒΑΝΟΜΕΝΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ ΚΑΙ ΧΡΟΝΙΚΗ ΕΞΑΡΤΗΣΗ	9
(ΕΝΟΤΗΤΑ 2.2.Γ) VANISHING ΚΑΙ EXPLODING GRADIENT	9
(ΕΝΟΤΗΤΑ 2.2.Δ) ΜΑΘΗΜΑΤΙΚΗ ΔΙΑΤΥΠΩΣΗ LSTM	9
2.3 MULTI-BRANCH ΑΡΧΙΤΕΚΤΟΝΙΚΕΣ ΚΑΙ EMBEDDINGS ΚΑΤΗΓΟΡΙΚΩΝ ΜΕΤΑΒΛΗΤΩΝ	10
(ΕΝΟΤΗΤΑ 2.3.Α) MULTILAYER PERCEPTRON (MLP)	10
(ΕΝΟΤΗΤΑ 2.3.Β) ΣΥΝΕΝΩΣΗ ΑΝΑΠΑΡΑΣΤΑΣΕΩΝ (FEATURE FUSION)	11
(ΕΝΟΤΗΤΑ 2.3.Γ) ENTITY EMBEDDINGS ΚΑΤΗΓΟΡΙΚΩΝ ΜΕΤΑΒΛΗΤΩΝ	11
2.4 MULTI-TASK LEARNING	11
(ΕΝΟΤΗΤΑ 2.4.Α) MULTI-TASK LEARNING ΚΑΙ ΚΟΙΝΗ ΑΝΑΠΑΡΑΣΤΑΣΗ	11
(ΕΝΟΤΗΤΑ 2.4.Β) ΕΤΕΡΟΓΕΝΕΙΣ ΣΥΝΑΡΤΗΣΕΙΣ ΑΠΩΛΕΙΑΣ (LOSS FUNCTIONS)	12
2.5 ΜΕΘΟΔΟΙ ΕΚΠΑΙΔΕΥΣΗΣ ΚΑΙ ΑΠΟΦΥΓΗ ΥΠΕΡΕΚΠΑΙΔΕΥΣΗΣ	12
(ΕΝΟΤΗΤΑ 2.5.Α) ΤΥΠΟΠΟΙΗΣΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΕΙΣΟΔΟΥ (FEATURE STANDARDIZATION)	12
(ΕΝΟΤΗΤΑ 2.5.Β) ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΜΕ ADAM OPTIMIZER	12
(ΕΝΟΤΗΤΑ 2.5.Γ) EARLY STOPPING & DROPOUT	13
(ΕΝΟΤΗΤΑ 2.5.Δ) ΑΠΟΦΥΓΗ DATA LEAKAGE ΚΑΙ ΧΡΟΝΙΚΟΣ ΔΙΑΧΩΡΙΣΜΟΣ ΔΕΔΟΜΕΝΩΝ	13
2.6 ΠΡΟΣΟΜΟΙΩΣΗ ΜΟΝΤΕ CARLO ΓΙΑ ΠΡΟΒΛΕΨΗ ΚΑΤΑΤΑΞΗΣ ΠΡΩΤΑΘΛΗΜΑΤΟΣ	14
ΚΕΦΑΛΑΙΟ 3 ΠΡΟΤΕΙΝΟΜΕΝΟ ΣΥΣΤΗΜΑ ΚΑΙ ΥΛΟΠΟΙΗΣΗ	15
3.1 ΣΥΝΟΛΙΚΗ ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ	15

3.2 ΠΗΓΗ, ΦΟΡΤΩΣΗ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ	16
(ΕΝΟΤΗΤΑ 3.2.Α) ΕΞΑΓΩΓΗ ΤΟΥ ΤΕΛΙΚΟΥ ΣΚΟΡ	17
(ΕΝΟΤΗΤΑ 3.2.Β) ΜΕΤΑΒΑΣΗ ΑΠΟ PLAYER-LEVEL ΣΕ TEAM-LEVEL ΔΕΔΟΜΕΝΑ	17
(ΕΝΟΤΗΤΑ 3.2.Γ) ΈΛΕΓΧΟΣ ΑΣΥΝΕΠΕΙΑΣ ΔΕΔΟΜΕΝΩΝ ΛΟΓΩ ΜΕΤΑΓΡΑΦΩΝ ΠΑΙΚΤΩΝ	18
3.3 ΜΗΧΑΝΙΚΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ	18
(ΕΝΟΤΗΤΑ 3.3.Α) ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΟΜΑΔΩΝ	18
(ΕΝΟΤΗΤΑ 3.3.Β) ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΠΑΙΚΤΩΝ	20
(ΕΝΟΤΗΤΑ 3.3.Γ) ΔΗΜΙΟΥΡΓΙΑ ΧΡΟΝΙΚΩΝ ΑΚΟΛΟΥΘΩΝ	21
(ΕΝΟΤΗΤΑ 3.3.Δ) ΤΥΠΟΠΟΙΗΣΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΚΑΙ ΔΙΑΧΕΙΡΙΣΗ ΕΛΛΙΠΩΝ ΤΙΜΩΝ	21
3.4 ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΤΟΥ ΜΟΝΤΕΛΟΥ ΟΜΑΔΩΝ	22
3.5 ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΤΟΥ ΜΟΝΤΕΛΟΥ ΠΑΙΚΤΩΝ	24
3.6 ΕΚΠΑΙΔΕΥΣΗ ΤΩΝ ΜΟΝΤΕΛΩΝ	26
(ΕΝΟΤΗΤΑ 3.6.Α) ΕΚΠΑΙΔΕΥΣΗ ΤΟΥ ΜΟΝΤΕΛΟΥ ΟΜΑΔΩΝ	26
(ΕΝΟΤΗΤΑ 3.6.Β) ΕΚΠΑΙΔΕΥΣΗ ΤΟΥ ΜΟΝΤΕΛΟΥ ΠΑΙΚΤΩΝ	26
3.7 ΔΙΑΔΙΚΑΣΙΑ ΠΑΡΑΓΩΓΗΣ ΠΡΟΒΛΕΨΕΩΝ	27
(ΕΝΟΤΗΤΑ 3.7.Α) ΠΡΟΒΛΕΨΗ ΑΓΩΝΑ	27
(ΕΝΟΤΗΤΑ 3.7.Β) ΠΡΟΒΛΕΨΗ ΠΑΙΚΤΩΝ	27
(ΕΝΟΤΗΤΑ 3.7.Γ) ΠΡΟΣΟΜΟΙΩΣΗ ΠΡΩΤΑΘΛΗΜΑΤΟΣ	28
(ΕΝΟΤΗΤΑ 3.7.Δ) ΠΡΟΣΘΕΤΕΣ ΠΡΟΒΛΕΨΕΙΣ ΠΑΙΚΤΩΝ ΣΕ ΕΠΙΠΕΔΟ ΠΡΩΤΑΘΛΗΜΑΤΟΣ	28
3.8 ΔΙΕΠΑΦΗ ΧΡΗΣΤΗ ΚΑΙ ΛΕΙΤΟΥΡΓΙΑ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ	29
 ΚΕΦΑΛΑΙΟ 4 - ΠΕΙΡΑΜΑΤΙΚΗ ΑΠΟΤΙΜΗΣΗ	 31
4.1 ΠΛΑΪΣΙΟ ΑΞΙΟΛΟΓΗΣΗΣ	31
4.2 ΣΥΝΟΛΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ΠΡΟΒΛΕΨΗΣ ΑΓΩΝΑ	31
4.3 ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΝΑ ΠΡΩΤΑΘΛΗΜΑ	32
4.4 ΑΝΑΛΥΣΗ ΠΡΟΒΛΕΨΕΩΝ ΤΕΛΙΚΟΥ ΑΠΟΤΕΛΕΣΜΑΤΟΣ	33
4.5 ΕΠΙΔΡΑΣΗ ΤΟΥ ΕΛΟ ΣΤΗΝ ΑΠΟΔΟΣΗ ΤΟΥ ΜΟΝΤΕΛΟΥ	34
4.6 ΕΞΕΛΙΞΗ ΤΗΣ ΕΚΠΑΙΔΕΥΣΗΣ ΤΩΝ ΜΟΝΤΕΛΩΝ	34
4.7 ΣΥΝΟΨΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΚΑΙ ΠΕΡΙΟΡΙΣΜΟΙ	36
 ΚΕΦΑΛΑΙΟ 5 - ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΠΡΟΕΚΤΑΣΕΙΣ	 36
5.1 ΓΕΝΙΚΑ ΣΥΜΠΕΡΑΣΜΑΤΑ	36
5.2 ΠΡΟΒΛΗΜΑΤΑ ΚΑΙ ΠΕΡΙΟΡΙΣΜΟΙ ΤΗΣ ΕΡΓΑΣΙΑΣ	37
5.3 ΜΕΛΛΟΝΤΙΚΕΣ ΠΡΟΕΚΤΑΣΕΙΣ	37
 ΒΙΒΛΙΟΓΡΑΦΙΑ	 39

ΚΕΦΑΛΑΙΟ 1 - Εισαγωγή

1.1 Γενική εισαγωγή

Το ποδόσφαιρο αποτελεί ένα από τα δημοφιλέστερα αθλήματα παγκοσμίως και συγκεντρώνει έντονο ενδιαφέρον τόσο σε αγωνιστικό όσο και σε επίπεδο ανάλυσης. Τα τελευταία χρόνια η αυξανόμενη αξιοποίηση στατιστικών στοιχείων έχει αλλάξει σημαντικά τον τρόπο με τον οποίο μπορεί να μελετηθεί η απόδοση ομάδων και παικτών. [1] Πληροφορίες που αφορούν τον αριθμό τερμάτων σε προηγούμενους αγώνες, την αμυντική και επιθετική απόδοση, τα αναμενόμενα γκολ, τα ατομικά στατιστικά παικτών και γενικά αποτελέσματα από προηγούμενους αγώνες μπορούν να χρησιμοποιηθούν για να κατανοηθεί η εξέλιξη της αγωνιστικής συμπεριφοράς μιας ομάδας.

Παρά το αυξανόμενο ενδιαφέρον για την ανάλυση τέτοιου είδους δεδομένων, η πρόβλεψη του αποτελέσματος ενός μελλοντικού αγώνα παραμένει ένα αρκετά σύνθετο πρόβλημα. Τα αποτελέσματα μιας αναμέτρησης δεν καθορίζονται από έναν μόνο παράγοντα, αλλά επηρεάζονται από το συνδυασμό πολλών διαφορετικών στοιχείων. Η πρόσφατη αγωνιστική κατάσταση των δύο ομάδων, η αμυντική και επιθετική τους αποτελεσματικότητα, η ισχύς του αντιπάλου, η απόδοση του κάθε παίκτη ξεχωριστά και ο παράγοντας της έδρας (με την έννοια ότι αν μια ομάδα παίζει στην έδρα της, πιθανότατα θα είναι και πιο αποτελεσματική στην αναμέτρηση), μπορούν να επηρεάσουν σε διαφορετικό βαθμό το τελικό αποτέλεσμα [2]. Επιπλέον, η εμφάνιση απρόβλεπτων γεγονότων κατά τη διάρκεια ενός αγώνα δυσκολεύει ακόμη περισσότερο την ακριβή πρόβλεψη της έκβασής του.

Η ανάπτυξη της μηχανικής μάθησης και των νευρωνικών δικτύων προσφέρει νέες δυνατότητες για την αντιμετώπιση τέτοιων προβλημάτων. Σε αντίθεση με απλούστερες στατιστικές προσεγγίσεις, τα νευρωνικά δίκτυα μπορούν να επεξεργαστούν ταυτόχρονα πολλαπλά χαρακτηριστικά και να προσαρμοστούν σε σύνθετες σχέσεις που ενδέχεται να υπάρχουν μεταξύ των δεδομένων [3].

Στο πλαίσιο της παρούσας πτυχιακής εργασίας εξετάζεται η αξιοποίηση τεχνικών μηχανικής μάθησης και νευρωνικών δικτύων για την πρόβλεψη μελλοντικών αγώνων ποδοσφαίρου, με βάση ιστορικά δεδομένα αγωνιστικής απόδοσης.

1.2 Κίνητρο της εργασίας

Η πρόβλεψη ποδοσφαιρικών αγώνων παρουσιάζει ιδιαίτερο ενδιαφέρον, καθώς αποτελεί ένα σύνθετο πρόβλημα στο οποίο συνδυάζονται ιστορικά δεδομένα, μεταβαλλόμενη αγωνιστική κατάσταση και υψηλός βαθμός αβεβαιότητας. Η απόδοση μιας ομάδας ή ενός μεμονωμένου παίκτη δεν διατηρείται σταθερή κατά τη διάρκεια της σεζόν, γεγονός που δημιουργεί την ανάγκη αξιοποίησης πρόσφατης αγωνιστικής συμπεριφοράς πέραν των συνολικών στατιστικών στοιχείων.

Βασικό κίνητρο της εργασίας είναι η διερεύνηση του κατά πόσο οι τεχνικές μηχανικής μάθησης μπορούν να χρησιμοποιήσουν τις υπάρχουσες πληροφορίες για να παράξουν

αξιόπιστες προβλέψεις. Αντί η πρόβλεψη να περιορίζεται αποκλειστικά στην επιλογή μεταξύ νίκης, ήττας και ισοπαλίας, ενδιαφέρον παρουσιάζει η δυνατότητα εκτίμησης πιθανοτήτων για τα διαφορετικά πιθανά αποτελέσματα ενός αγώνα, ώστε να προβάλλεται καλύτερα η αβεβαιότητα του αποτελέσματος.

1.3 Προβλήματα-Περιορισμοί

Μια από τις σημαντικότερες προκλήσεις κατά την ανάπτυξη της εργασίας ήταν η συλλογή επαρκούς όγκου κατάλληλων ποδοσφαιρικών δεδομένων. Για την εκπαίδευση ενός μοντέλου μηχανικής μάθησης, απαιτείται επαρκής αριθμός ιστορικών καταγραφών, οι οποίες να περιλαμβάνουν πληροφορίες τόσο για την απόδοση των ομάδων όσο και για την ατομική απόδοση των παικτών.

Κατά την αναζήτηση διαθέσιμων πηγών, διαπιστώθηκε πως δεν ήταν εύκολο να εντοπιστούν δεδομένα με την απαιτούμενη έκταση και λεπτομέρεια. Η πηγή που χρησιμοποιήθηκε εν τέλει ήταν η διαδικτυακή βάση δεδομένων FBref [4], από την οποία αντλήθηκαν στατιστικά αγώνων, ομάδων και παικτών. Παρότι η συγκεκριμένη πηγή παρείχε σημαντικό αριθμό δεδομένων, ο συνολικός όγκος και το εύρος τους παρέμειναν μικρότερα από επιθυμητά. Δεν εντοπίστηκε κάποια διαφορετική πηγή που να προσφέρει αντίστοιχα επαρκή και κατάλληλα οργανωμένα δεδομένα, ώστε να επεκταθεί περαιτέρω το σύνολο που χρησιμοποιήθηκε για την εκπαίδευση του μοντέλου.

1.4 Στόχος της Εργασίας

Στόχος της παρούσας εργασίας είναι η ανάπτυξη ενός ολοκληρωμένου συστήματος πρόβλεψης και ανάλυσης ποδοσφαιρικών δεδομένων, το οποίο δεν περιορίζεται μόνο στην εκτίμηση του αποτελέσματος ενός μεμονωμένου αγώνα, αλλά καλύπτει διαφορετικά επίπεδα ανάλυσης μιας ποδοσφαιρικής διοργάνωσης.

Συγκεκριμένα, το σύστημα έχει ως στόχο την πρόβλεψη του σκορ ενός αγώνα και της αναμενόμενης απόδοσης των παικτών που συμμετέχουν σε αυτόν, καθώς και την πρόβλεψη της τελικής κατάταξης ενός πρωταθλήματος. Επιπλέον, παρέχει εκτιμήσεις σχετικά με τον πιθανότερο πρωταθλητή και τις ομάδες που εμφανίζουν πιθανότητα υποβιβασμού. Σε επίπεδο παικτών, επιδιώκεται η εκτίμηση των παικτών με την υψηλότερη προβλεπόμενη επιθετική απόδοση καθώς και των παικτών που παρουσιάζουν τη μεγαλύτερη πιθανότητα να δεχθούν κίτρινη ή κόκκινη κάρτα. Τέλος, το σύστημα παρέχει τη δυνατότητα προβολής επιμέρους πληροφοριών και στατιστικών για κάθε παίκτη.

1.5 Δομή της εργασίας

Η εργασία οργανώνεται σε πέντε κεφάλαια. Μετά το εισαγωγικό κεφάλαιο, στο Κεφάλαιο 2 παρουσιάζεται η επισκόπηση της σχετικής βιβλιογραφίας και αναλύεται το θεωρητικό υπόβαθρο των τεχνικών που χρησιμοποιούνται στην εργασία. Ειδικότερα, παρουσιάζονται οι βασικές αρχές και η μαθηματική διατύπωση των μεθόδων που αξιοποιούνται για την ανάπτυξη των μοντέλων και την παραγωγή των προβλέψεων.

Στο Κεφάλαιο 3 παρουσιάζεται το προτεινόμενο σύστημα και περιγράφεται αναλυτικά η υλοποίησή του. Αναλύεται η συλλογή και η προεπεξεργασία των δεδομένων, η δημιουργία των απαραίτητων χαρακτηριστικών, η αρχιτεκτονική και η εκπαίδευση των μοντέλων, καθώς και η συνολική λειτουργία του συστήματος.

Στο Κεφάλαιο 4 παρουσιάζεται η πειραματική αποτίμηση του συστήματος, η διαδικασία αξιολόγησης και τα αποτελέσματα που προέκυψαν από τις προβλέψεις των μοντέλων.

Τέλος, στο Κεφάλαιο 5 παρουσιάζονται τα βασικά συμπεράσματα της εργασίας, οι περιορισμοί που προέκυψαν και πιθανές επεκτάσεις και βελτιώσεις του συστήματος στο μέλλον.

ΚΕΦΑΛΑΙΟ 2 - Βιβλιογραφική Επισκόπηση

2.1 Στατιστική μοντελοποίηση αποτελεσμάτων (Poisson μοντέλα)

(Ενότητα 2.1.α) Κατανομή Poisson

Η κατανομή Poisson αποτελεί μια διακριτή κατανομή πιθανότητας, η οποία χρησιμοποιείται για τη μοντελοποίηση του αριθμού εμφανίσεων ενός γεγονότος μέσα σε ένα συγκεκριμένο χρονικό ή χωρικό διάστημα. Μια τυχαία μεταβλητή X που ακολουθεί κατανομή Poisson χαρακτηρίζεται από την παράμετρο $\lambda > 0$, η οποία αποτελεί και την αναμενόμενη τιμή της μεταβλητής. [5]

Η συνάρτηση πιθανότητας της κατανομής Poisson δίνεται από τη σχέση:

$$P(i) = P\{X = i\} = \frac{e^{-\lambda} \lambda^i}{i!}, \quad i=0,1,\dots \quad [5] \quad (2.1)$$

Όπου i είναι ο αριθμός των εμφανίσεων του εξεταζόμενου γεγονότος και λ η παράμετρος της κατανομής. Η αναμενόμενη τιμή μιας τυχαίας μεταβλητής που ακολουθεί την κατανομή είναι ίση με λ [5]:

$$E[X] = \lambda \quad (2.2)$$

ενώ η διακύμανσή της είναι ίση με την ίδια παράμετρο [5]:

$$\text{Var}(X) = \lambda. \quad (2.3)$$

(Ενότητα 2.1.β) Μοντελοποίηση Poisson κατά Maher

Για την μοντελοποίηση του τελικού σκορ ενός ποδοσφαιρικού αγώνα μπορούν να χρησιμοποιηθούν δύο ανεξάρτητες κατανομές Poisson, μια για τον αριθμό τερμάτων της γηπεδούχου και μία για της φιλοξενούμενης. Η προσέγγιση αυτή έχει χρησιμοποιηθεί από τον Maher [6], ο οποίος έδειξε ότι το ανεξάρτητο μοντέλο Poisson μπορεί να παρέχει ικανοποιητική περιγραφή των αποτελεσμάτων ποδοσφαιρικών αγώνων.

Με βάση την παραπάνω προσέγγιση, στην παρούσα εργασία ορίζεται ως X_H ο αριθμός τερμάτων της γηπεδούχου και X_A ο αριθμός τερμάτων της φιλοξενούμενης. Οι δύο μεταβλητές μοντελοποιούνται μέσω κατανομών Poisson με παραμέτρους λ_H και λ_A αντίστοιχα:

$$X_H \sim \text{Poisson}(\lambda_H) \quad (2.4)$$

$$X_A \sim \text{Poisson}(\lambda_A) \quad (2.5)$$

όπου λ_H και λ_A εκφράζουν τον αναμενόμενο αριθμό τερμάτων της γηπεδούχου και τις φιλοξενούμενης ομάδας αντίστοιχα.

Υποθέτοντας ότι οι δύο μεταβλητές είναι ανεξάρτητες, από τον ορισμό της ανεξαρτησίας διακριτών τυχαίων μεταβλητών, προκύπτει [5]:

$$P(X_H = x, X_A = y) = P(X_H = x)P(X_A = y) \quad (2.6)$$

Και επομένως, εφαρμόζοντας τη συνάρτηση πιθανότητας Poisson της σχέσης (2.1), στις δύο μεταβλητές και χρησιμοποιώντας την υπόθεση ανεξαρτησίας, προκύπτει:

$$P(X_H = x, X_A = y) = \frac{e^{-\lambda_H} \lambda_H^x}{x!} \frac{e^{-\lambda_A} \lambda_A^y}{y!} \quad [5], [6] \quad (2.7)$$

Η σχέση (2.7) επιτρέπει τον υπολογισμό της πιθανότητας κάθε δυνατού τελικού σκορ ενός αγώνα, με βάση τις δύο αναμενόμενες τιμές λ_H και λ_A . Με αυτόν τον τρόπο σχηματίζεται μια κατανομή πιθανοτήτων για τα πιθανά σκορ, η οποία μπορεί στη συνέχεια να χρησιμοποιηθεί για την εξαγωγή των πιθανοτήτων των διαφορετικών εκβάσεων του αγώνα.

(Ενότητα 2.1.γ) Διόρθωση Dixon-Coles

Η υπόθεση ανεξαρτησίας μεταξύ των τερμάτων των δύο ομάδων δεν αποδίδει εξίσου καλά σε όλες τις πιθανές εκβάσεις ενός αγώνα. Οι Dixon και Coles [7] πρότειναν μία τροποποίηση του ανεξάρτητου μοντέλου Poisson, εισάγοντας έναν διορθωτικό παράγοντα για τα χαμηλά σκορ, όπως 0-0, 1-0, 0-1 και 1-1. Η κοινή πιθανότητα των τερμάτων των δύο ομάδων δίνεται στο μοντέλο των Dixon-Coles από τη σχέση [7]:

$$Pr(X_{i,j} = x, Y_{i,j} = y) = \tau_{\lambda,\mu}(x, y) \frac{\lambda^x \exp(-\lambda)}{x!} \frac{\mu^y \exp(-\mu)}{y!} \quad (2.8)$$

Όπου:

- $X_{i,j}$:αριθμός τερμάτων της γηπεδούχου ομάδας(i απέναντι στην ομάδα j),
- $Y_{i,j}$:αριθμός τερμάτων της φιλοξενούμενης ομάδας,
- x, y :οι αντίστοιχοι αριθμοί τερμάτων,
- λ, μ :οι αναμενόμενοι αριθμοί τερμάτων των δύο ομάδων,
- $\tau_{\lambda,\mu}$:ο διορθωτικός παράγοντας,
- ρ :παράμετρος εξάρτησης

Στη συνέχεια, ο διορθωτικός παράγοντας ορίζεται στο Dixon-Coles ως [7]:

$$\tau_{\lambda,\mu}(x, y) = \begin{cases} 1 - \lambda\mu\rho, & \text{αν } x = y = 0, \\ 1 + \lambda\rho, & \text{αν } x = 0, y = 1, \\ 1 + \mu\rho, & \text{αν } x = 1, y = 0, \\ 1 - \rho, & \text{αν } x = y = 1, \\ 1, & \text{διαφορετικά.} \end{cases} \quad (2.9)$$

(Ενότητα 2.1.δ) Εξαγωγή πιθανοτήτων αποτελέσματος

Η μοντελοποίηση του αριθμού τερμάτων αντί της απευθείας ταξινόμησης του αποτελέσματος σε νίκη, ισοπαλία ή ήττα επιτρέπει την παραγωγή ολόκληρης της κατανομής πιθανών σκορ. Από την κατανομή αυτή μπορούν στη συνέχεια να υπολογιστούν

τόσο οι πιθανότητες συγκεκριμένων σκορ όσο και οι πιθανότητες των τριών εκβάσεων του αγώνα.

Επιπλέον, η αντιμετώπιση κάθε δυνατού τελικού σκορ ως ξεχωριστή κατηγορία, θα οδηγούσε σε μεγάλο αριθμό κατηγοριών, πολλές από τις οποίες εμφανίζονται σπάνια στα διαθέσιμα δεδομένα.

Από την κοινή κατανομή πιθανότητας $Pr(X_{i,j} = x, Y_{i,j} = y)$ (2.8), είναι δυνατός ο υπολογισμός της πιθανότητας κάθε μιάς από τις τρεις εκβάσεις ενός αγώνα (νίκη γηπεδούχου, ισοπαλία, νίκη φιλοξενούμενης), αθροίζοντας την πιθανότητα σε όλα τα ζεύγη σκορ (x,y) που αντιστοιχούν στην εκάστοτε έκβαση [6],[7]:

$$\bullet \quad P(\text{Νίκη γηπεδούχου}) = \sum_{x>y} Pr(X_{i,j} = x, Y_{i,j} = y) \quad (2.10)$$

$$\bullet \quad P(\text{Ισοπαλία}) = \sum_{x=y} Pr(X_{i,j} = x, Y_{i,j} = y) \quad (2.11)$$

$$\bullet \quad P(\text{Νίκη φιλοξενούμενης}) = \sum_{x<y} Pr(X_{i,j} = x, Y_{i,j} = y) \quad (2.12)$$

Όπου το άθροισμα στην πράξη περιορίζεται σε ένα πεπερασμένο εύρος τιμών $x,y \in \{0,1,...,N\}$, με το N να επιλέγεται αρκετά μεγάλο ώστε η πιθανότητα των σκορ $>N$ που παραλείπεται να είναι αμελητέα, δεδομένων των ρεαλιστικών τιμών των παραμέτρων λ_H, λ_A στο ποδόσφαιρο.

(Ενότητα 2.1.ε) Σύστημα αξιολόγησης Elo

Το σύστημα Elo αποτελεί μέθοδο δυναμικής αξιολόγησης της σχετικής αγωνιστικής ισχύος μεταξύ δύο ομάδων. Κάθε ομάδα αντιστοιχίζεται σε μία αριθμητική τιμή αξιολόγησης, η οποία μεταβάλλεται διαδοχικά ανάλογα με τα αποτελέσματα των αγώνων της. Βασικό στοιχείο της μεθόδου είναι ότι το αναμενόμενο αποτέλεσμα μιας αναμέτρησης εξαρτάται από τη διαφορά μεταξύ των αξιολογήσεων των δύο αντιπάλων [31]. Οι Szczecinski και Djebbi ορίζουν τη διαφορά αξιολόγησης μεταξύ των δύο αντιπάλων μέσω της μεταβλητής v και χρησιμοποιούν τη λογιστική αθροιστική συνάρτηση κατανομής:

$$\Phi(v) = \frac{1}{1+10^{-v/\sigma}} \quad (2.13)$$

όπου $\sigma > 0$ είναι η παράμετρος κλίμακας της αξιολόγησης. Όσο μεγαλύτερη είναι η τιμή του v , δηλαδή όσο μεγαλύτερη είναι η διαφορά αξιολόγησης υπέρ της εξεταζόμενης ομάδας, τόσο μεγαλύτερη είναι και η αντίστοιχη αναμενόμενη επίδοσή της [31]. Η ενημέρωση της αξιολόγησης μετά την πραγματοποίηση του αγώνα δίνεται στο ίδιο μοντέλο από τη σχέση [31]:

$$\hat{\theta}_{n+1,i} = \hat{\theta}_{n,i} + K[s_i - \Phi(\Delta_i)] \quad (2.14)$$

όπου $\Delta_i = \hat{\theta}_{n,i} - \hat{\theta}_{n,j}$ είναι η διαφορά μεταξύ της τρέχουσας αξιολόγησης Elo της ομάδας i και της αξιολόγησης της αντίπαλης ομάδας j , ενώ K είναι η παράμετρος που ελέγχει το μέγεθος της μεταβολής της αξιολόγησης μετά από κάθε αποτέλεσμα.

Στην περίπτωση όπου λαμβάνονται υπόψη ισοπαλίες, η μεταβλητή s_i λαμβάνει την τιμή 1 σε περίπτωση νίκης, την τιμή $1/2$ σε περίπτωση ισοπαλίας και την τιμή 0 σε περίπτωση ήττας. Με αυτόν τον τρόπο, η μεταβολή της αξιολόγησης εξαρτάται από τη διαφορά μεταξύ της πραγματικής επίδοσης και της αναμενόμενης επίδοσης $\Phi(\Delta_i)$. Ένα αποτέλεσμα καλύτερο από το αναμενόμενο, αυξάνει την αξιολόγηση, ενώ ένα χειρότερο οδηγεί σε μείωσή της [31].

Στην παρούσα εργασία, οι αξιολογήσεις Elo των δύο ομάδων και η μεταξύ τους διαφορά αξιοποιούνται ως πρόσθετα χαρακτηριστικά εισόδου στο μοντέλο των ομάδων.

(Ενότητα 2.2.α) Εισαγωγή

Ένα συμβατικό feedforward νευρωνικό δίκτυο επεξεργάζεται κάθε είσοδο χωρίς να διατηρεί εσωτερική πληροφορία από προηγούμενα χρονικά βήματα. Για παράδειγμα, αν δοθούν ως είσοδος τα στατιστικά των N προηγούμενων αγώνων μιας ομάδας, το δίκτυο δε διαθέτει μηχανισμό μνήμης που να αποτυπώνει άμεσα τη χρονική εξάρτηση μεταξύ των διαδοχικών αγώνων. Όμως, στην αγωνιστική εξέλιξη μιας ομάδας η χρονική σειρά είναι σημαντική πληροφορία, γιατί μια ανοδική πορεία ή μια πτώση στην απόδοση έχουν διαφορετική προγνωστική αξία, ανάλογα με το πότε ακριβώς έγιναν σε σχέση με τον επόμενο αγώνα.

Γι' αυτό, τα δίκτυα Long Short-Term Memory (LSTM) αποτελούν ειδική μορφή επαναλαμβανόμενων νευρωνικών δικτύων (Recurrent Neural Networks-RNN), σχεδιασμένη για την επεξεργασία δεδομένων που παρουσιάζουν ακολουθιακή ή χρονική δομή [8]. Σε αντίθεση με ένα συμβατικό feedforward νευρωνικό δίκτυο, ένα RNN διαθέτει επαναλαμβανόμενες συνδέσεις και διατηρεί μια εσωτερική κατάσταση, μέσω της οποίας πληροφορίες από προηγούμενα χρονικά βήματα μπορούν να επηρεάζουν την επεξεργασία των επόμενων [9].

Η συγκεκριμένη ιδιότητα είναι σημαντική όταν οι παρατηρήσεις δεν είναι ανεξάρτητες μεταξύ τους αλλά σχηματίζουν μια χρονικά διατεταγμένη ακολουθία. Στην παρούσα εργασία, η πρόσφατη αγωνιστική συμπεριφορά μιας ομάδας μπορεί να αναπαρασταθεί ως ακολουθία προηγούμενων αγώνων, ώστε να διατηρείται η χρονική σειρά με την οποία εμφανίστηκαν οι αντίστοιχες παρατηρήσεις.

Παρότι τα απλά RNN έχουν τη δυνατότητα διατήρησης πληροφορίας από προηγούμενα βήματα, η εκπαίδευσή τους σε μεγαλύτερες ακολουθίες μπορεί να επηρεαστεί από τα προβλήματα vanishing gradient και exploding gradient. Κατά την οπισθοδιάδοση του σφάλματος (backpropagation) μέσα στον χρόνο, οι κλίσεις μπορούν να μειωθούν σε πολύ μικρές τιμές ή να αυξηθούν υπερβολικά, δυσκολεύοντας την εκμάθηση εξαρτήσεων μεταξύ χρονικά απομακρυσμένων παρατηρήσεων [9].

Τα LSTM χρησιμοποιούν μια εσωτερική κατάσταση μνήμης και μηχανισμούς πυλών (gates), οι οποίοι ελέγχουν τη ροή της πληροφορίας μέσα στο δίκτυο. Στη σύγχρονη μορφή της αρχιτεκτονικής χρησιμοποιούνται η πύλη εισόδου (input gate), η forget gate και η πύλη εξόδου (output gate), οι οποίες καθορίζουν αντίστοιχα ποια νέα πληροφορία θα εισαχθεί στη μνήμη, ποιο μέρος της προηγούμενης κατάστασης θα διατηρηθεί και ποια πληροφορία θα μεταφερθεί στην έξοδο [8], [10].

(Ενότητα 2.2.β) Επαναλαμβανόμενα Νευρωνικά Δίκτυα και χρονική εξάρτηση

Σε ένα επαναλαμβανόμενο νευρωνικό δίκτυο, η κατάσταση του δικτύου στο χρονικό βήμα t εξαρτάται τόσο από την τρέχουσα είσοδο όσο και από την κατάσταση του προηγούμενου χρονικού βήματος. Η γενική σχέση δίνεται Pascanu, Mikolov και Bengio ως [9]:

$$\bullet \quad \mathbf{h}_t = \mathbf{F}(\mathbf{h}_{t-1}, \mathbf{x}_t, \boldsymbol{\theta}) \quad (2.15)$$

Όπου $\mathbf{h}_t$ είναι η κρυφή κατάσταση του δικτύου στο χρονικό βήμα t , $\mathbf{h}_{t-1}$ η προηγούμενη κατάσταση, $\mathbf{x}_t$ η τρέχουσα είσοδος και $\boldsymbol{\theta}$ το σύνολο των παραμέτρων του μοντέλου.

(Ενότητα 2.2.γ) Vanishing και Exploding Gradient

Κατά την εκπαίδευση ενός RNN, το σφάλμα μεταδίδεται προς τα προηγούμενα χρονικά βήματα μέσω της διαδικασίας Backpropagation Through Time (BPTT). Στα RNN, η επίδραση ενός προηγούμενου χρονικού βήματος k σε ένα μεταγενέστερο χρονικό βήμα t εξαρτάται από το γινόμενο των διαδοχικών Jacobian πινάκων [9]:

$$\frac{\partial x_t}{\partial x_k} = \prod_{t \geq i > k} \frac{\partial x_i}{\partial x_{i-1}} = \prod_{t \geq i > k} W_{rec}^T \text{diag}(\sigma'(x_{i-1})) \quad (2.16)$$

Καθώς αυξάνεται η χρονική απόσταση $t-k$, το γινόμενο αυτό περιλαμβάνει όλο και περισσότερους πίνακες. Αν το μέτρο των αντίστοιχων όρων είναι μικρότερο από τη μονάδα, η κλίση μπορεί να μειώνεται εκθετικά καθώς μεταδίδεται στα προηγούμενα χρονικά βήματα. Το φαινόμενο αυτό ονομάζεται vanishing gradient και δυσκολεύει την εκμάθηση εξαρτήσεων μεταξύ χρονικά απομακρυσμένων παρατηρήσεων.

Αντίθετα, όταν το γινόμενο αυξάνεται σημαντικά κατά τη μετάδοση προς τα πίσω, το μέτρο της κλίσης μπορεί να αυξηθεί υπερβολικά, οδηγώντας στο exploding gradient. Στην περίπτωση αυτή, η διαδικασία εκπαίδευσης μπορεί να γίνει ασταθής λόγω των πολύ μεγάλων τιμών των gradients [9].

(Ενότητα 2.2.δ) Μαθηματική διατύπωση LSTM

Το LSTM αντιμετωπίζει το πρόβλημα του vanishing gradient εισάγοντας μια ξεχωριστή κατάσταση κυττάρου (cell state) $\mathbf{C}_t$, η οποία διατρέχει το δίκτυο με ελάχιστες γραμμικές παρεμβάσεις, επιτρέποντας στην πληροφορία να διατηρείται σε μεγαλύτερες χρονικές αποστάσεις χωρίς να εξασθενεί εκθετικά. Η ροή της πληροφορίας ρυθμίζεται από τις πύλες (gates).

Η αρχιτεκτονική LSTM, όπως προτάθηκε από τους Hochreiter και Schmidhuber [8], περιλάμβανε αποκλειστικά μηχανισμούς εισόδου και εξόδου. Η forget gate, που επιτρέπει στο δίκτυο να “ξεχνά” επιλεκτικά τμήματα της κατάστασης κυττάρου, προστέθηκε μεταγενέστερα από τους Gers, Schmidhuber και Cummins [10] και αποτελεί πλέον αναπόσπαστο τμήμα της σύγχρονης, ευρέως χρησιμοποιούμενης μορφής LSTM, όπως αυτή υλοποιείται και στο πλαίσιο εργασίας PyTorch που αξιοποιείται και στην παρούσα εργασία. Η σύγχρονη, ενοποιημένη μαθηματική διατύπωση που ενσωματώνει και τις δύο συνεισφορές (τον αρχικό μηχανισμό σταθερής διατήρησης του σφάλματος μέσω του cell state και τη forget gate) παρουσιάζεται συνήθως στη μορφή [11]:

$$\bullet \quad f_t = \sigma(W_f \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + b_f) \quad (2.17)$$

- $i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$ (2.18)

- $\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$ (2.19)

- $C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$ (2.20)

- $o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$ (2.21)

- $h_t = o_t \odot \tanh(C_t)$ (2.22)

Όπου f_t η forget gate, που καθορίζει ποιο μέρος της προηγούμενης κατάστασης κυττάρου C_{t-1} θα διατηρηθεί, i_t η input gate, που ελέγχει πόση νέα πληροφορία $\tilde{C}_t$ θα ενσωματωθεί στο κύτταρο, o_t η output gate που καθορίζει ποιο μέρος της ενημερωμένης κατάστασης C_t θα μεταφερθεί στο κρυφό διάνυσμα εξόδου h_t και $\odot$ που συμβολίζει τον στοιχείο κατά στοιχείο πολλαπλασιασμό. Η προσθετική, αντί για καθαρά πολλαπλασιαστική, ενημέρωση της σχέσης (2.20), είναι ο μηχανισμός που επιτρέπει στην κλίση να διαδίδεται προς τα πίσω στον χρόνο χωρίς την εκθετική συρρίκνωση που περιγράφεται στη σχέση (2.16), περιορίζοντας έτσι σημαντικά το πρόβλημα του vanishing gradient σε μεγάλο εύρος πρακτικών εφαρμογών [8], [10].

2.3 Multi-branch αρχιτεκτονικές και embeddings κατηγορικών μεταβλητών

(Ενότητα 2.3.α) Multilayer Perceptron (MLP)

Ένα Multilayer Perceptron (MLP) είναι ένα feedforward νευρωνικό δίκτυο που αποτελείται από διαδοχικά πλήρως συνδεδεμένα επίπεδα. Κάθε επίπεδο εφαρμόζει γραμμικό μετασχηματισμό στις εισόδους του και στη συνέχεια μία μη γραμμική συνάρτηση ενεργοποίησης, επιτρέποντας στο δίκτυο να μοντελοποιεί μη γραμμικές σχέσεις μεταξύ χαρακτηριστικών και της τελικής εξόδου [12]

Για μια μονάδα j , οι Rumelhart, Hinton και Williams [12] ορίζουν τη συνολική είσοδο ως:

$$x_j = \sum_i y_i w_{ji} \quad (2.23)$$

Όπου y_i είναι οι έξοδοι των μονάδων του προηγούμενου επιπέδου και w_{ji} τα αντίστοιχα βάρη των συνδέσεων. Σε κάθε μονάδα ενός πλήρως συνδεδεμένου επιπέδου μπορεί να προστεθεί ένας όρος πόλωσης (bias), ο οποίος ισοδυναμεί με μια επιπλέον είσοδο σταθερής τιμής 1 [12].

Στα κρυφά επίπεδα χρησιμοποιείται μη γραμμική συνάρτηση ενεργοποίησης. Για τη συνάρτηση Rectified Linear Unit (ReLU), οι Nair και Hinton [13], χρησιμοποιούν τη σχέση:

$$y = \max(0, x) \quad (2.24)$$

ώστε οι αρνητικές τιμές εισόδου να μηδενίζονται, ενώ οι θετικές να διατηρούνται αμετάβλητες.

(Ενότητα 2.3.β) Συνένωση αναπαραστάσεων (feature fusion)

Όταν διαφορετικοί κλάδοι ενός μοντέλου παράγουν ξεχωριστές αναπαραστάσεις δεδομένων, οι αναπαραστάσεις αυτές μπορούν να συνενωθούν (concatenation) σε μια κοινή αναπαράσταση πριν το τελικό στάδιο πρόβλεψης. Οι Baltrusaitis, Ahuja και Morency [14] εκφράζουν μαθηματικά μια κοινή αναπαράσταση ως: $\mathbf{x}_m = \mathbf{f}(\mathbf{x}_1, \dots, \mathbf{x}_n)$ (2.25) όπου, $\mathbf{x}_1, \dots, \mathbf{x}_n$ είναι οι επιμέρους αναπαραστάσεις και $\mathbf{x}_m$ η κοινή αναπαράσταση που προκύπτει από τη συνάρτηση $\mathbf{f}$.

(Ενότητα 2.3.γ) Entity Embeddings κατηγορικών μεταβλητών

Οι κατηγορικές μεταβλητές (όπως το πρωτάθλημα προέλευσης ενός αγώνα) μπορούν να αναπαρασταθούν μέσω διανυσμάτων πραγματικών αριθμών που μαθαίνονται μαζί με τις υπόλοιπες παραμέτρους του νευρωνικού δικτύου. Οι Guo και Berkahn [15] ορίζουν το entity embedding μιας κατηγορικής μεταβλητής x_i μέσω της αντιστοίχισης:

$$e_i: x_i \mapsto \mathbf{x}_i \quad (2.26)$$

όπου x_i είναι η τιμή της κατηγορικής μεταβλητής, e_i η συνάρτηση αντιστοίχισης και $\mathbf{x}_i$ το αντίστοιχο διάνυσμα πραγματικών τιμών. Το διάνυσμα αυτό δεν είναι προκαθορισμένο, αλλά οι τιμές του μαθαίνονται κατά την εκπαίδευση μαζί με τις υπόλοιπες παραμέτρους του νευρωνικού δικτύου [15].

2.4 Multi-Task Learning

(Ενότητα 2.4.α) Multi-Task Learning και κοινή αναπαράσταση

Η πολυεργασιακή μάθηση (Multi-Task Learning) αποτελεί προσέγγιση κατά την οποία περισσότερες από μία συσχετιζόμενες εργασίες μαθαίνονται ταυτόχρονα, αξιοποιώντας μια κοινή αναπαράσταση των δεδομένων. Με αυτόν τον τρόπο, η πληροφορία που αποκτάται κατά την εκμάθηση μιας εργασίας μπορεί να συμβάλει στην εκμάθηση των υπόλοιπων εργασιών [16].

Στα νευρωνικά δίκτυα, μια συνηθισμένη μορφή πολυεργασιακής μάθησης είναι το hard parameter sharing, κατά το οποίο τα κρυφά επίπεδα (hidden layers) είναι κοινά μεταξύ των επιμέρους εργασιών, ενώ κάθε εργασία διαθέτει ξεχωριστό επίπεδο εξόδου [17].

Η χρήση κοινής αναπαράστασης αντί για πλήρως ανεξάρτητα μοντέλα επιτρέπει στις επιμέρους εργασίες να αξιοποιούν κοινή πληροφορία, περιορίζοντας παράλληλα την ανάγκη για ξεχωριστό σύνολο παραμέτρων για κάθε εργασία [16], [17]. Όταν οι εργασίες σχετίζονται μεταξύ τους, η κοινή μάθηση μπορεί επίσης να λειτουργήσει ως μορφή κανονικοποίησης και να βελτιώσει τη γενίκευση του μοντέλου.

Στην παρούσα εργασία, η κοινή αναπαράσταση του παίκτη χρησιμοποιείται για την ταυτόχρονη πρόβλεψη διαφορετικών τύπων μεταβλητών μέσω ξεχωριστών εξόδων του μοντέλου.

(Ενότητα 2.4.β) Ετερογενείς συναρτήσεις απώλειας (Loss Functions)

Οι επιμέρους προβλέψεις ενός πολυεργασιακού μοντέλου μπορεί να αφορούν μεταβλητές διαφορετικής φύσης και συνεπώς να απαιτούν διαφορετικές συναρτήσεις απώλειας. Στην συγκεκριμένη εργασία χρησιμοποιούνται ξεχωριστές συναρτήσεις απώλειας για μεταβλητές καταμέτρησης, δυαδικές μεταβλητές και συνεχείς μεταβλητές.

Για τις μεταβλητές που εκφράζουν αριθμό γεγονότων χρησιμοποιείται η συνάρτηση απώλειας Poisson. Οι Suryadi, Chew και Ong [18] τη διατυπώνουν ως:

$$L_{Po}(x_j, \mu_j) = \mu_j - x_j \log \mu_j \quad (2.27)$$

όπου x_j είναι η παρατηρούμενη τιμή και μ_j ο μέσος ρυθμός της κατανομής Poisson που εκτιμάται από το νευρωνικό δίκτυο [18].

Για τις δυαδικές μεταβλητές χρησιμοποιείται η Binary Cross-Entropy (BCE). Για μια δυαδική παρατήρηση x_j και προβλεπόμενη πιθανότητα $p_j = P(x_j=1)$ [18], ο τύπος είναι:

$$L_B(x_j, p_j) = -x_j \log p_j - (1 - x_j) \log (1 - p_j) \quad (2.28)$$

όπου p_j είναι η εκτιμώμενη πιθανότητα η μεταβλητή να λάβει την τιμή 1.

Για τις συνεχείς μεταβλητές χρησιμοποιείται η Mean Squared Error (MSE), η οποία υπολογίζει το μέσο τετραγωνικό σφάλμα μεταξύ πραγματικών και προβλεπόμενων τιμών. Οι Mehta et al. [19], την ορίζουν ως:

$$E(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i(\mathbf{w}))^2 \quad (2.29)$$

όπου n είναι το πλήθος των παρατηρήσεων, y_i η πραγματική τιμή και $\hat{y}_i(\mathbf{w})$ η αντίστοιχη πρόβλεψη του νευρωνικού δικτύου με παραμέτρους $\mathbf{w}$ [19].

2.5 Μέθοδοι εκπαίδευσης και αποφυγή υπερεκπαίδευσης

(Ενότητα 2.5.α) Τυποποίηση χαρακτηριστικών εισόδου (Feature Standardization)

Τα χαρακτηριστικά εισόδου μπορεί να εκφράζονται σε διαφορετικές αριθμητικές κλίμακες. Γι' αυτό το λόγο εφαρμόζεται τυποποίηση τύπου Z-Score, ώστε κάθε χαρακτηριστικό να μετασχηματίζεται με βάση το μέσο όρο και την τυπική απόκλιση του. Οι Cao et al. [25], ορίζουν τον μετασχηματισμό ως:

$$v' = (v - \bar{x}) / (\sigma_x) \quad (2.30)$$

όπου v είναι η αρχική τιμή, $\bar{x}$ ο μέσος όρος και σ_x η τυπική απόκλιση της αντίστοιχης μεταβλητής. Μετά τον μετασχηματισμό, το χαρακτηριστικό έχει μέσο όρο μηδέν και τυπική απόκλιση ίση με ένα [25]. Οι παράμετροι $\bar{x}$ και σ_x υπολογίζονται από τα δεδομένα εκπαίδευσης και στη συνέχεια χρησιμοποιούνται για τον μετασχηματισμό των υπόλοιπων δεδομένων.

(Ενότητα 2.5.β) Βελτιστοποίηση με Adam Optimizer

Για τη βελτιστοποίηση των παραμέτρων ενός νευρωνικού δικτύου απαιτείται η επαναληπτική ενημέρωσή τους με βάση την κλίση των συναρτήσεων απώλειας. Στην εργασία, χρησιμοποιήθηκε ο αλγόριθμος Adam, ο οποίος βασίζεται σε προσαρμοστικές εκτιμήσεις του μέσου όρου και του μέσου τετραγώνου των κλίσεων (ροπές) [20]. Οι Kingma και Ba ορίζουν [20] :

$$g_t = \nabla_{\theta} f_t(\theta_{t-1}) \quad (2.31)$$

Και τις εκτιμήσεις των δύο ροπών:

$$\bullet \quad m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \quad (2.32)$$

$$\bullet \quad v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \quad (2.33)$$

Επειδή οι m_t και v_t ξεκινούν από μηδενικές αρχικές τιμές, οι πρώτες εκτιμήσεις τους επηρεάζονται από αυτή την αρχικοποίηση. Γι'αυτό το λόγο εφαρμόζεται διόρθωση (bias correction), ώστε να προκύψουν οι διορθωμένες εκτιμήσεις m_t και v_t [20]:

$$\bullet \quad \widehat{m}_t = (m_t) / (1 - \beta_1^t) \quad (2.34)$$

$$\bullet \quad \widehat{v}_t = (v_t) / (1 - \beta_2^t) \quad (2.35)$$

και οι παράμετροι ενημερώνονται σύμφωνα με:

$$\theta_t = \theta_{t-1} - \alpha (\widehat{m}_t) / (\sqrt{\widehat{v}_t} + \epsilon) \quad (2.36)$$

όπου g_t είναι η κλίση στο χρονικό βήμα t , m_t και v_t οι εκτιμήσεις της πρώτης και της δεύτερης ροπής αντίστοιχα, β_1, β_2 οι συντελεστές εκθετικής απόσβεσης και α ο ρυθμός μάθησης [20].

(Ενότητα 2.5.γ) Early stopping & Dropout

Κατά την εκπαίδευση ενός νευρωνικού δικτύου, η συνεχής μείωση του σφάλματος εκπαίδευσης δεν συνεπάγεται απαραίτητα αντίστοιχη βελτίωση της απόδοσης σε μη χρησιμοποιημένα δεδομένα. Το early stopping χρησιμοποιεί την απόδοση σε σύνολο επικύρωσης για να εντοπίσει το σημείο μετά το οποίο η περαιτέρω εκπαίδευση δεν βελτιώνει τη γενίκευση του μοντέλου [21]. Ο Prechelt [21] περιγράφει το validation-based on early stopping ως τεχνική για τον περιορισμό της υπερπροσαρμογής (overfitting), κατά την οποία η εκπαίδευση διακόπτεται όταν η επίδοση στο σύνολο επικύρωσης παύει να βελτιώνεται.

Στην παρούσα εργασία, μετά από κάθε εποχή παρακολουθείται η τιμή της συνάρτησης απώλειας στο σύνολο επικύρωσης. Εάν αυτή η τιμή δεν παρουσιάσει βελτίωση για προκαθορισμένο αριθμό διαδοχικών εποχών, η διαδικασία εκπαίδευσης διακόπτεται.

Επίσης, μια επιπλέον τεχνική που χρησιμοποιείται για τον περιορισμό της υπερπροσαρμογής είναι το Dropout, όπου κατά την εκπαίδευση ορισμένες μονάδες του δικτύου απενεργοποιούνται προσωρινά και τυχαία, μαζί με τις αντίστοιχες συνδέσεις τους. Με αυτόν τον τρόπο περιορίζεται η υπερβολική συνεξάρτηση μεταξύ συγκεκριμένων μονάδων του δικτύου [24]. Οι Srivastava et al. [24] περιγράφουν τη διαδικασία μέσω μιας δυαδικής τυχαίας μεταβλητής:

$$r_j^{(l)} \sim \text{Bernoulli}(p) \quad (2.37)$$

και η έξοδος του επιπέδου μετά την εφαρμογή Dropout δίνεται από [24]:

$$\tilde{y}^{(l)} = r^{(l)} * y^{(l)} \quad (2.38)$$

όπου $\tilde{y}^{(l)}$ είναι το διάνυσμα της εξόδου του επιπέδου l , $r^{(l)}$ το τυχαίο δυαδικό διάνυσμα Dropout και p η πιθανότητα διατήρησης μιας μονάδας [24].

(Ενότητα 2.5.δ) Αποφυγή Data Leakage και χρονικός διαχωρισμός δεδομένων

Data Leakage υπάρχει όταν το μοντέλο, κατά την εκπαίδευση ή την κατασκευή των χαρακτηριστικών του, χρησιμοποιεί πληροφορία που δεν θα ήταν διαθέσιμη τη στιγμή που θα έπρεπε να πραγματοποιήσει την πρόβλεψη [22]. Αυτό μπορεί να οδηγήσει σε υπερασπισιόδοξη εκτίμηση της πραγματικής απόδοσης του μοντέλου.

Στη συγκεκριμένη εργασία, για την αποφυγή αυτού του προβλήματος, τα χαρακτηριστικά κάθε αγώνα υπολογίζονται αποκλειστικά από προηγούμενες παρατηρήσεις. Επομένως, δεν

χρησιμοποιούνται στοιχεία από τον ίδιο αγώνα που πρόκειται να προβλεφθεί ούτε από μεταγενέστερους αγώνες.

Επιπλέον, επειδή τα δεδομένα έχουν χρονική διάταξη, ο διαχωρισμός σε σύνολα εκπαίδευσης, επικύρωσης και ελέγχου πραγματοποιείται με βάση τον χρόνο και όχι τυχαία [23]. Με αυτόν τον τρόπο, το μοντέλο εκπαιδεύεται σε παλαιότερα δεδομένα και αξιολογείται σε μεταγενέστερα, προσομοιώνοντας πιο ρεαλιστικά τη διαδικασία μελλοντικής πρόβλεψης.

2.6 Προσομοίωση Monte Carlo για πρόβλεψη κατάταξης πρωταθλήματος

Η μέθοδος Monte Carlo βασίζεται στην επαναλαμβανόμενη τυχαία δειγματοληψία από καθορισμένες κατανομές πιθανότητας και στη στατιστική ανάλυση των αποτελεσμάτων που προκύπτουν από τις προσομοιώσεις [26]. Για κάθε σύνολο τυχαίων τιμών, παράγεται μια δυνατή έξοδος του μοντέλου. Μετά από μεγάλο αριθμό επαναλήψεων, τα αποτελέσματα μπορούν να χρησιμοποιηθούν για την εκτίμηση των αντίστοιχων στατιστικών μεγεθών [26]. Ο Raychaudhuri [26] χρησιμοποιεί τον μέσο όρο των αποτελεσμάτων των επαναλαμβανόμενων προσομοιώσεων για την εκτίμηση της αναμενόμενης τιμής ενός μεγέθους που προκύπτει από τις προσομοιώσεις:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad [26] \quad (2.39)$$

όπου x_i είναι η τιμή της εξεταζόμενης ποσότητας στην i -οστή προσομοίωση και n ο συνολικός αριθμός των προσομοιώσεων. Επομένως, η προσέγγιση Monte Carlo επιτρέπει την εκτίμηση στατιστικών μεγεθών από την εμπειρική κατανομή των αποτελεσμάτων που παράγονται μέσω των επαναλήψεων [26].

Στην περίπτωση των ποδοσφαιρικών διοργανώσεων, η ίδια λογική μπορεί να χρησιμοποιηθεί για την προσομοίωση των αποτελεσμάτων ενός συνόλου αγώνων και την εκτίμηση της τελικής κατάταξης. Οι Koning et al. [27] εφαρμόζουν προσομοίωση σε ποδοσφαιρικά πρωταθλήματα, χρησιμοποιώντας κατανομές Poisson για την παραγωγή των αριθμών τερμάτων των ομάδων και επαναλαμβάνοντας τη διαδικασία ώστε να εκτιμηθούν πιθανότητες σχετικές με την τελική έκβαση της διοργάνωσης.

Στην παρούσα εργασία, για κάθε πιθανό ζεύγος ομάδων που εξετάζεται στην προσομοίωση, χρησιμοποιούνται οι παράμετροι λ_H και λ_A που παράγονται από το μοντέλο πρόβλεψης. Σε κάθε επανάληψη της προσομοίωσης παράγεται τυχαία ένα πιθανό σκορ για κάθε αγώνα από τις αντίστοιχες κατανομές Poisson και αποδίδονται οι βαθμοί στις ομάδες ανάλογα με το αποτέλεσμα. Η διαδικασία επαναλαμβάνεται πολλές φορές και από τις κατατάξεις που προκύπτουν υπολογίζονται οι μέσοι βαθμοί των ομάδων, καθώς και εμπειρικές πιθανότητες για γεγονότα όπως η κατάκτηση του πρωταθλήματος και ο υποβιβασμός.

ΚΕΦΑΛΑΙΟ 3 ΠΡΟΤΕΙΝΟΜΕΝΟ ΣΥΣΤΗΜΑ ΚΑΙ ΥΛΟΠΟΙΗΣΗ

3.1 Συνολική αρχιτεκτονική του συστήματος

Η υλοποίηση της παρούσας εργασίας βασίζεται σε ένα ολοκληρωμένο σύστημα επεξεργασίας ποδοσφαιρικών δεδομένων και παραγωγής προβλέψεων, το οποίο οργανώνεται σε διαδοχικά στάδια. Η συνολική διαδικασία ξεκινά από τη φόρτωση των αρχικών δεδομένων αγώνων και παικτών, συνεχίζει με την προεπεξεργασία και την κατασκευή των απαραίτητων χαρακτηριστικών και καταλήγει στην εκπαίδευση και χρήση των μοντέλων πρόβλεψης.

Τα αρχικά δεδομένα αποθηκεύονται σε αρχεία CSV και οργανώνονται ανά πρωτάθλημα και αγωνιστική περίοδο. Κατά τη φόρτωσή τους, τα επιμέρους αρχεία συνενώνονται σε ένα κοινό σύνολο δεδομένων, ενώ προστίθενται πληροφορίες που προσδιορίζουν τη σεζόν και το πρωτάθλημα στο οποίο ανήκει κάθε εγγραφή. Στη συνέχεια πραγματοποιείται ο απαραίτητος καθαρισμός των δεδομένων και δημιουργούνται τα χαρακτηριστικά που χρησιμοποιούνται στα μοντέλα.

Μετά το αρχικό στάδιο επεξεργασίας, η ροή του συστήματος διαχωρίζεται σε δύο βασικά υποσυστήματα. Το πρώτο λειτουργεί σε επίπεδο ομάδας και έχει ως στόχο την πρόβλεψη της αναμενόμενης επιθετικής απόδοσης της γηπεδούχου και της φιλοξενούμενης ομάδας. Το δεύτερο λειτουργεί σε επίπεδο παίκτη και χρησιμοποιείται για την πρόβλεψη διαφορετικών ατομικών στατιστικών και γεγονότων που σχετίζονται με την απόδοση ενός παίκτη.

Αυτός ο διαχωρισμός επιλέχθηκε επειδή τα δύο προβλήματα έχουν διαφορετική μορφή εισόδου και διαφορετικές μεταβλητές εξόδου. Στο μοντέλο ομάδων χρησιμοποιούνται χαρακτηριστικά που περιγράφουν τη συνολική αγωνιστική κατάσταση των δύο ομάδων και παράγονται δύο θετικές παράμετροι λ_H και λ_A , οι οποίες αντιστοιχούν στον αναμενόμενο αριθμό τερμάτων τους. Οι τιμές αυτές χρησιμοποιούνται στη συνέχεια για τη δημιουργία της κατανομής πιθανών σκορ και τον υπολογισμό των πιθανοτήτων νίκης γηπεδούχου, ισοπαλίας και νίκης φιλοξενούμενης.

Αντίθετα, το μοντέλο παικτών λαμβάνει πληροφορίες που αφορούν την πρόσφατη και συνολική αγωνιστική συμπεριφορά κάθε παίκτη και παράγει ταυτόχρονα διαφορετικού είδους προβλέψεις. Συγκεκριμένα, το μοντέλο διαθέτει ξεχωριστές εξόδους για μεταβλητές καταμέτρησης, για τον αναμενόμενο χρόνο συμμετοχής και για δυαδικά γεγονότα, όπως η συμμετοχή στην αρχική ενδεκάδα ή η λήψη καρτών. Η κοινή εσωτερική αναπαράσταση επιτρέπει την ταυτόχρονη εκπαίδευση των διαφορετικών αυτών εργασιών στο ίδιο μοντέλο. Πριν από την εκπαίδευση, τα αριθμητικά χαρακτηριστικά τυποποιούνται και τα δεδομένα οργανώνονται σε κατάλληλες χρονικές ακολουθίες και διανύσματα εισόδου. Επειδή οι παρατηρήσεις έχουν χρονική διάταξη, η διαδικασία εκπαίδευσης και αξιολόγησης πραγματοποιείται επίσης με χρονικό διαχωρισμό, ώστε οι πληροφορίες από μεταγενέστερους αγώνες να μην χρησιμοποιούνται για την εκπαίδευση πάνω σε

παλαιότερες παρατηρήσεις. Οι παράμετροι τυποποίησης υπολογίζονται αποκλειστικά από τα δεδομένα εκπαίδευσης.

Για την αξιολόγηση του μοντέλου ομάδων εφαρμόζεται η διαδικασία *walk-forward evaluation*, κατά την οποία δημιουργούνται διαδοχικά χρονικά σύνολα εκπαίδευσης, επικύρωσης και ελέγχου. Σε κάθε επανάληψη, το μοντέλο εκπαιδεύεται εκ νέου από την αρχή χρησιμοποιώντας μόνο τις παλαιότερες διαθέσιμες παρατηρήσεις, το αμέσως επόμενο χρονικό διάστημα χρησιμοποιείται ως σύνολο επικύρωσης (*validation*) και για την εφαρμογή του *early stopping*, ενώ η απόδοση αξιολογείται σε μεταγενέστερο χρονικό διάστημα ελέγχου. Με τον τρόπο αυτό εξετάζεται η συμπεριφορά του μοντέλου σε περισσότερες από μία χρονικές περιόδους. Για το μοντέλο παικτών, χρησιμοποιείται ξεχωριστός χρονικός διαχωρισμός σε σύνολα εκπαίδευσης και επικύρωσης, σύμφωνα με τη διαφορετική διαδικασία εκπαίδευσής του.

Μετά την ολοκλήρωση της εκπαίδευσης, τα τελικά εκπαιδευμένα μοντέλα χρησιμοποιούνται από τις επιμέρους λειτουργίες πρόβλεψης του συστήματος. Το μοντέλο ομάδων χρησιμοποιείται τόσο για την πρόβλεψη μεμονωμένων αγώνων όσο και ως βάση για την προσομοίωση ολόκληρου πρωταθλήματος με τη μέθοδο Monte Carlo. Το μοντέλο παικτών χρησιμοποιείται για την παραγωγή ατομικών προβλέψεων και για την εξαγωγή συγκεντρωτικών εκτιμήσεων σε επίπεδο πρωταθλήματος. Οι λειτουργίες αυτές ενοποιούνται σε μια κοινή διαδραστική διεπαφή γραμμής εντολών, μέσω της οποίας ο χρήστης μπορεί να επιλέξει τον τύπο πρόβλεψης που επιθυμεί.

3.2 Πηγή, φόρτωση και επεξεργασία δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν στην εργασία προέρχονται από τη διαδικτυακή πλατφόρμα FBref και περιλαμβάνουν πληροφορίες σχετικά με αγώνες, ομάδες και ατομικές επιδόσεις παικτών. Το σύνολο δεδομένων που χρησιμοποιήθηκε καλύπτει τέσσερα ποδοσφαιρικά πρωταθλήματα και πέντε αγωνιστικές περιόδους. Τα δεδομένα αποθηκεύονται τοπικά σε αρχεία CSV, οργανωμένα ιεραρχικά ανά πρωτάθλημα και στη συνέχεια ανά σεζόν.

Η συγκεκριμένη οργάνωση επιτρέπει στο σύστημα να επεξεργάζεται με ενιαίο τρόπο δεδομένα από διαφορετικές διοργανώσεις και χρονικές περιόδους. Κατά τη διαδικασία φόρτωσης, η συνάρτηση **CsvLoadMergeAndFeatureAdding()** διατρέπει αρχικά τους φακέλους των πρωταθλημάτων και στη συνέχεια τους αντίστοιχους φακέλους της κάθε σεζόν. Για κάθε αρχείο CSV δημιουργούνται επιπλέον οι μεταβλητές **season** και **league**, ώστε κάθε εγγραφή να διατηρεί την πληροφορία σχετικά με τη χρονική περίοδο και το πρωτάθλημα από το οποίο προέρχεται. Παράλληλα αποθηκεύεται και η ομάδα που αντιστοιχεί στο αρχείο προέλευσης μέσω της μεταβλητής `source_file_team`.

Τα επιμέρους σύνολα δεδομένων συνενώνονται στη συνέχεια σε κοινό DataFrame, το οποίο αποτελεί τη βάση για τα επόμενα στάδια επεξεργασίας. Η φόρτωση κάθε αρχείου πραγματοποιείται με κωδικοποίηση UTF-8-sig, ώστε να αντιμετωπίζεται σωστά η πιθανή παρουσία του UTF-8 Byte Order Mark [28] και να αποφεύγονται προβλήματα κατά την ανάγνωση των ονομάτων των στηλών ή άλλων χαρακτήρων. Επιπλέον, η διαδικασία φόρτωσης περιλαμβάνει κατάλληλο χειρισμό σφαλμάτων, ώστε αν υπάρξει τυχόν αποτυχία ανάγνωσης ενός μεμονωμένου αρχείου να μην επηρεάζει την επεξεργασία των υπόλοιπων αρχείων.

Μετά τη συνένωση πραγματοποιούνται αρχικοί έλεγχοι εγκυρότητας των εγγραφών. Συγκεκριμένα, χρησιμοποιείται λίστα των έγκυρων ομάδων κάθε πρωταθλήματος και

αφαιρούνται εγγραφές στις οποίες είτε η ομάδα είτε η αντίπαλος δεν ανήκουν στο αντίστοιχο σύνολο ομάδων. Η διαδικασία αυτή περιορίζει την παρουσία μη αναγνωρισμένων ή ασύμβατων εγγραφών στο dataset που θα χρησιμοποιηθεί στη συνέχεια για την κατασκευή των χαρακτηριστικών.

Η συνολική διαδικασία φόρτωσης και συνένωσης των επιμέρους αρχείων συνοψίζεται ως:

Είσοδος: Φάκελοι πρωταθλημάτων

Έξοδος: Ενοποιημένο σύνολο δεδομένων

Για κάθε πρωτάθλημα:

 Για κάθε σεζόν:

 Για κάθε αρχείο CSV:

 Φόρτωση του αρχείου

 Προσθήκη πρωταθλήματος

 Προσθήκη σεζόν

 Αποθήκευση επιμέρους συνόλου

Συνένωση όλων των επιμέρους συνόλων

Επιστροφή ενοποιημένου dataset

(Ενότητα 3.2.α) Εξαγωγή του τελικού σκορ

Στα αρχικά δεδομένα, το τελικό σκορ του αγώνα βρίσκεται αποθηκευμένο στη στήλη **Result** σε μορφή συμβολοσειράς. Για τις ανάγκες εκπαίδευσης του μοντέλου απαιτούνται δύο ξεχωριστές αριθμητικές μεταβλητές, μια για τα τέρματα που πέτυχε η ομάδα που εξετάζεται και μια για τα τέρματα που δέχθηκε. Γι' αυτό το λόγο υλοποιήθηκε η συνάρτηση **goalparsefromresult()**. Η συνάρτηση εφαρμόζει regular expression στη συμβολοσειρά του αποτελέσματος: $(\d+)\s*[-]\s*(\d+)$, με σκοπό την αναγνώριση των δύο αριθμητικών τμημάτων του σκορ. Η έκφραση υποστηρίζει τον χαρακτήρα hyphen minus (-) καθώς και en dashes (–), ώστε να αναγνωρίζονται διαφορετικές μορφές αναπαράστασης του τελικού σκορ. Οι δύο αριθμοί μετατρέπονται στη συνέχεια σε αριθμητικές τιμές και αποθηκεύονται στις νέες στήλες **GF** (Goals For column) και **GA** (Goals Against column).

Οι μεταβλητές αυτές χρησιμοποιούνται στη συνέχεια κατά τη δημιουργία του πίνακα αγώνων σε επίπεδο ομάδας. Η συνάρτηση **teamtable()** καλεί αρχικά τη διαδικασία εξαγωγής του σκορ και διατηρεί τις βασικές πληροφορίες κάθε αγώνα, όπως ημερομηνία, ομάδα, αντίπαλο, έδρα, αποτέλεσμα, τέρματα υπέρ και κατά, σεζόν και πρωτάθλημα.

Για την συμβολοσειρά: $(\d+)\s*[-]\s*(\d+)$ [29], το $\d$ αντιστοιχεί σε δεκαδικό ψηφίο, το + δηλώνει μια ή περισσότερες εμφανίσεις του προηγούμενου στοιχείου, το $\s$ σε χαρακτήρα κενού, ενώ η χρήση αγκυλών [] επιτρέπει την αποδοχή ενός από τους χαρακτήρες που περιέχονται σε αυτές.

(Ενότητα 3.2.β) Μετάβαση από player-level σε team-level δεδομένα

Τα αρχεία περιέχουν πολλαπλές εγγραφές που σχετίζονται με τον ίδιο αγώνα, επειδή οι πληροφορίες καταγράφονται και σε επίπεδο παικτών. Για την εκπαίδευση του μοντέλου των ομάδων απαιτείται ενιαία εγγραφή ανά ομάδα και αγώνα. Για τον λόγο αυτό, δημιουργείται ξεχωριστός πίνακας σε επίπεδο ομάδας, από τον οποίο αφαιρούνται οι επαναλαμβανόμενες εγγραφές του ίδιου αγώνα και διατηρούνται μόνο τα χαρακτηριστικά που αφορούν την ομάδα συνολικά. Η αφαίρεση των διπλότυπων εγγραφών

πραγματοποιείται μέσω της **drop_duplicates()**, χρησιμοποιώντας ως κριτήρια την ημερομηνία του αγώνα, την ομάδα και τον αντίπαλο.

(Ενότητα 3.2.γ) Έλεγχος ασυνέπειας δεδομένων λόγω μεταγραφών παικτών

Κατά τον έλεγχο των αρχικών δεδομένων εντοπίστηκε το πρόβλημα που σχετίζεται με τους παίκτες οι οποίοι μεταγράφηκαν σε διαφορετική ομάδα κατά τη διάρκεια της ίδιας αγωνιστικής περιόδου. Σε ορισμένες περιπτώσεις, το αρχείο μιας συγκεκριμένης ομάδας περιείχε εγγραφές παικτών με στήλη ομάδας (Squad) διαφορετική από την ομάδα στην οποία ανήκε το ίδιο το αρχείο, καθώς η διαδικασία συλλογής δεδομένων συμπεριλάμβανε τους αγώνες που ο παίκτης είχε αγωνιστεί με προηγούμενη ομάδα πριν τη μεταγραφή του. Αυτό δημιουργούσε κίνδυνο να συσχετιστούν στατιστικά που παράχθηκαν με μια προηγούμενη ομάδα με τη νέα ομάδα κατά τη κατασκευή των χαρακτηριστικών.

Για τον εντοπισμό τέτοιων περιπτώσεων διατηρείται κατά τη φόρτωση των δεδομένων η πληροφορία της ομάδας προέλευσης κάθε αρχείου μέσω της μεταβλητής `source_file_team`, η οποία εξάγεται από το όνομα του αρχείου. Στη συνέχεια, για κάθε συνδυασμό πρωταθλήματος, σεζόν και αρχείου προέλευσης, δηλαδή για κάθε ομάδα (league, season, source_file_team), προσδιορίζεται η επικρατέστερη τιμή (mode) της στήλης Squad μεταξύ όλων των εγγραφών του συγκεκριμένου αρχείου. Η τιμή αυτή θεωρείται η ομάδα αναφοράς του αρχείου και οι εγγραφές των οποίων η τιμή Squad διαφέρει από αυτή, αφαιρούνται από το σύνολο δεδομένων.

Η διαδικασία δεν εφαρμόζεται ξεχωριστά για κάθε παίκτη, αλλά στο σύνολο των εγγραφών κάθε αρχείου. Με αυτόν τον τρόπο η εκτίμηση της ομάδας αναφοράς βασίζεται στο συνολικό περιεχόμενο του αρχείου και δεν επηρεάζεται υπερβολικά από τους παίκτες με μικρό αριθμό συμμετοχών μετά από μια μεταγραφή. Παράλληλα, η χρήση της επικρατέστερης τιμής της Squad, αντί της άμεσης σύγκρισης με το όνομα του αρχείου, επιτρέπει την αντιμετώπιση πιθανών διαφορών ονοματοδοσίας μεταξύ των δύο.

3.3 Μηχανική χαρακτηριστικών

Μετά την ολοκλήρωση της προεπεξεργασίας, τα αρχικά δεδομένα μετασχηματίζονται σε σύνολο χαρακτηριστικών κατάλληλο για την εκπαίδευση των μοντέλων. Στο στάδιο αυτό υπολογίζονται χρονικά και συγκεντρωτικά στατιστικά που αποτυπώνουν την πρόσφατη αγωνιστική κατάσταση και τη συνολική πορεία μέσα στη σεζόν. Η διαδικασία εφαρμόζεται στα δεδομένα ομάδων και παικτών, με διαφορετικό σύνολο μεταβλητών σε κάθε περίπτωση.

(Ενότητα 3.3.α) Χαρακτηριστικά ομάδων

Η δημιουργία των βασικών στατιστικών χαρακτηριστικών σε επίπεδο ομάδας πραγματοποιείται από τη συνάρτηση **build_team_features()**, ενώ τα χαρακτηριστικά που σχετίζονται με την αξιολόγηση Elo υπολογίζονται ξεχωριστά πριν από τη διαμόρφωση του τελικού συνόλου δεδομένων. Αρχικά, οι εγγραφές ταξινομούνται ανά ομάδα, σεζόν και ημερομηνία, ώστε οι υπολογισμοί που ακολουθούν να πραγματοποιούνται με σωστή χρονική σειρά. Στη συνέχεια, τα δεδομένα ομαδοποιούνται ως προς την ομάδα και τη

σεζόν, ώστε τα στατιστικά μιας αγωνιστικής περιόδου να μην επηρεάζονται από τα δεδομένα προηγούμενης ή επόμενης σεζόν.

Για κάθε ομάδα δημιουργούνται κυλιόμενα χαρακτηριστικά που περιγράφουν την πρόσφατη αγωνιστική της κατάσταση. Συγκεκριμένα, για παράθυρο N προηγούμενων αγώνων υπολογίζεται ο μέσος αριθμός τερμάτων που πέτυχε η ομάδα, ο μέσος αριθμός τερμάτων που δέχθηκε και το ποσοστό νικών. Η τελευταία μεταβλητή υπολογίζεται μέσω μιας βοηθητικής δυαδικής μεταβλητής, η οποία λαμβάνει την τιμή 1 όταν η ομάδα έχει κερδίσει τον αγώνα και 0 για άλλο αποτέλεσμα.

Στην παρούσα υλοποίηση, το μέγεθος του κυλιόμενου παραθύρου ορίζεται σε $N=7$ αγώνες. Σημαντικό στοιχείο της υλοποίησης είναι ότι πριν από κάθε υπολογισμό κυλιόμενου μέσου εφαρμόζεται η μετατόπιση `shift(1)`. Με τον τρόπο αυτό, κατά την κατασκευή των χαρακτηριστικών ενός αγώνα δεν χρησιμοποιούνται τα στατιστικά του ίδιου αγώνα, αλλά αποκλειστικά αποτελέσματα που ήταν διαθέσιμα πριν από την πραγματοποίησή του. Η επιλογή αυτή αποτελεί πρακτική εφαρμογή της αποφυγής `data leakage`.

Για τα αρχικά στάδια μιας σεζόν, όταν είναι διαθέσιμες από μία έως $N-1$ προηγούμενες παρατηρήσεις, χρησιμοποιείται η παράμετρος `min_periods=1`. Έτσι, ο κυλιόμενος μέσος υπολογίζεται χρησιμοποιώντας όσους προηγούμενους αγώνες είναι διαθέσιμοι μέχρι τη συγκεκριμένη χρονική στιγμή, χωρίς να απαιτείται η συμπλήρωση ολόκληρου του παραθύρου.

Εκτός από τα κυλιόμενα χαρακτηριστικά, υπολογίζονται και χαρακτηριστικά `season-to-date`, τα οποία βασίζονται σε όλους τους προηγούμενους αγώνες της ομάδας μέσα στην τρέχουσα σεζόν. Για τον σκοπό αυτό χρησιμοποιείται `expanding mean`, επίσης μετά από `shift(1)`. Με αυτόν τον τρόπο παράγονται η μέση επιθετική απόδοση της σεζόν, η μέση αμυντική απόδοση και το ποσοστό νικών μέχρι το συγκεκριμένο χρονικό σημείο.

Επιπλέον δημιουργείται ένα χαρακτηριστικό που λαμβάνει υπόψη την έδρα του αγώνα. Για τον υπολογισμό του, οι εγγραφές ομαδοποιούνται όχι μόνο ανά ομάδα και σεζόν, αλλά και ανά τιμή της μεταβλητής `Venue`. Έτσι, ο κυλιόμενος μέσος αριθμός τερμάτων που πετυχαίνει μια ομάδα υπολογίζεται ξεχωριστά για τους αγώνες στους οποίους αγωνίζεται ως γηπεδούχος και για εκείνους στους οποίους αγωνίζεται ως φιλοξενούμενη. Δηλαδή, σε έναν αγώνα στον οποίο μια ομάδα αγωνίζεται ως γηπεδούχος, το αντίστοιχο χαρακτηριστικό αποτυπώνει τη μέση επιθετική της απόδοση στους προηγούμενους εντός έδρας αγώνες και όχι τη συνολική απόδοση ανεξαρτήτως έδρας. Με αυτόν τον τρόπο διατηρείται πληροφορία σχετικά με τη διαφοροποίηση της επιθετικής απόδοσης της ομάδας ανάλογα με την έδρα.

Εκτός από τα στατιστικά χαρακτηριστικά πρόσφατης φόρμας και αγωνιστικής περιόδου, στο μοντέλο ενσωματώνονται και χαρακτηριστικά που προκύπτουν από το σύστημα αξιολόγησης `Elo`. Στην παρούσα υλοποίηση, οι αξιολογήσεις υπολογίζονται ανεξάρτητα για κάθε πρωτάθλημα, ώστε να μην πραγματοποιείται άμεση σύγκριση μεταξύ ομάδων που αγωνίζονται σε διαφορετικά πρωταθλήματα.

Η αρχική αξιολόγηση κάθε ομάδας ορίζεται στις 1500 μονάδες, ενώ για την ενημέρωσή της μετά από κάθε αγώνα χρησιμοποιείται συντελεστής $K=20$ και παράμετρος κλίμακας $\sigma=400$. Για κάθε αναμέτρηση καταγράφεται η αξιολόγηση `Elo` και των δύο ομάδων πριν από την πραγματοποίηση του αγώνα. Η ενημέρωση των αξιολογήσεων πραγματοποιείται στη συνέχεια ταυτόχρονα για τις δύο ομάδες, με βάση τις τιμές που είχαν πριν από την αναμέτρηση. Με αυτόν τον τρόπο αποφεύγεται η χρήση πληροφορίας από το αποτέλεσμα του ίδιου αγώνα κατά τη δημιουργία των χαρακτηριστικών εισόδου.

Κατά την αλλαγή της αγωνιστικής περιόδου εφαρμόζεται μερική επαναφορά της αξιολόγησης προς μία βασική τιμή, πρακτική που έχει χρησιμοποιηθεί και σε εφαρμογές του Elo για δεδομένα πολλαπλών σεζόν [32]. Η γενική μορφή της διαδικασίας δίνεται από [32]:

$$E_{new} = E_{previous}(1 - r) + E_{base} * r$$

Στην παρούσα υλοποίηση επιλέγονται $r=0.25$ και $E_{base}=1500$. Συμβολίζοντας την αξιολόγηση Elo με R , η σχέση γίνεται:

$$R_{new} = 0.75R_{previous} + 0.25 * 1500$$

Επιπλέον, χρησιμοποιείται ως πρόσθετο χαρακτηριστικό η διαφορά μεταξύ των δύο αξιολογήσεων Elo της γηπεδούχου και της φιλοξενούμενης ομάδας πριν από την αναμέτρηση, ώστε το μοντέλο να διαθέτει έναν συνοπτικό δείκτη της σχετικής αγωνιστικής τους ισχύος. Η επιλογή αυτή ακολουθεί προηγούμενες προσεγγίσεις πρόβλεψης ποδοσφαιρικών αγώνων, στις οποίες η διαφορά των προ αγώνα αξιολογήσεων Elo χρησιμοποιείται ως μεταβλητή εισόδου [33].

(Ενότητα 3.3.β) Χαρακτηριστικά παικτών

Για τα δεδομένα παικτών, η κατασκευή των χαρακτηριστικών πραγματοποιείται επίσης με βάση αποκλειστικά προγενέστερες αγωνιστικές παρατηρήσεις, ώστε να αποφεύγεται η χρήση πληροφορίας από τον ίδιο τον αγώνα ή από μεταγενέστερες αναμετρήσεις. Ωστόσο, χρησιμοποιείται διαφορετικό σύνολο μεταβλητών, καθώς ο στόχος είναι η αποτύπωση της πρόσφατης ατομικής αγωνιστικής κατάστασης του παίκτη. Η δημιουργία των χαρακτηριστικών πραγματοποιείται από τη συνάρτηση **rollingavgfeatures()**. Αρχικά, οι αριθμητικές μεταβλητές μετατρέπονται σε κατάλληλη αριθμητική μορφή, ενώ τυχόν μη διαθέσιμες τιμές αντικαθίστανται με μηδενικές τιμές. Παράλληλα, η μεταβλητή *Start*, η οποία δηλώνει αν ένας παίκτης ξεκίνησε ως βασικός, μετατρέπεται σε δυαδική μορφή.

Οι βασικές μεταβλητές πάνω στις οποίες κατασκευάζονται τα χρονικά χαρακτηριστικά περιλαμβάνουν τα λεπτά συμμετοχής (*Min*), τα γκολ (*Gl*s), τις ασίστ (*Ast*), τα σουτ (*Sh*), τα σουτ εντός στόχου (*SoT*), τα φάουλ (*Fl*s), τα κερδισμένα φάουλ (*Fl*d), τις αναχαίτισεις (*Int*), τα κερδισμένα τάκλιν (*TklW*), τις κίτρινες και κόκκινες κάρτες (*CrdY*, *CrdR*), τα πέναλτι που παραχωρήθηκαν ή κερδήθηκαν (*PKcon*, *PKwon*), τις σέντρες (*Crs*) και τα οφσάιντ (*Off*). Η λίστα των μεταβλητών ορίζεται πριν από τη συνάρτηση **rollingavgfeatures()** μέσω της *player_columns_features*.

Οι εγγραφές ταξινομούνται ανά παίκτη, σεζόν και ημερομηνία και στη συνέχεια ομαδοποιούνται ανά παίκτη και σεζόν. Για κάθε διαθέσιμο χαρακτηριστικό υπολογίζονται δύο διαφορετικές χρονικές αναπαραστάσεις: ο κυλιόμενος μέσος όρος των τελευταίων $N=5$ αγώνων και ο μέσος όρος όλων των προηγούμενων αγώνων της τρέχουσας σεζόν. Και στις δύο περιπτώσεις εφαρμόζεται *shift(1)*, ώστε τα χαρακτηριστικά ενός αγώνα να βασίζονται αποκλειστικά σε προηγούμενες εμφανίσεις του παίκτη και να μην περιλαμβάνουν πληροφορία από τον ίδιο αγώνα.

Επιπλέον, για τη μεταβλητή *Start*, υπολογίζεται το *start_rate_last5*, το οποίο εκφράζει το ποσοστό των προηγούμενων πέντε αγώνων στους οποίους ο παίκτης ξεκίνησε ως βασικός. Με τον τρόπο αυτό παρέχεται στο μοντέλο πληροφορία σχετικά με το πόσο συστηματικά χρησιμοποιείται ο συγκεκριμένος παίκτης στην αρχική ενδεκάδα. Στον υπολογισμό του *start_rate_last5* εφαρμόζεται *shift(1)*, ώστε να μην χρησιμοποιείται η πληροφορία της βασικής ενδεκάδας του ίδιου αγώνα.

Επιπλέον, τα χαρακτηριστικά του παίκτη εμπλουτίζονται με πληροφορίες που αφορούν την αντίπαλη ομάδα, μέσω της συνάρτησης **build_player_match_dataset()**. Για κάθε εγγραφή

παίκτη πραγματοποιείται αντίστοιχη με τα χαρακτηριστικά της αντίπαλης ομάδας για την ίδια ημερομηνία αγώνα. Συγκεκριμένα, χρησιμοποιούνται χαρακτηριστικά που περιγράφουν την πρόσφατη και συνολική αμυντική της επίδοση, όπως ο κυλιόμενος μέσος όρος τερμάτων που έχει δεχθεί, ο αντίστοιχος μέσος όρος της τρέχουσας σεζόν, καθώς και το πρόσφατο και συνολικό ποσοστό νικών της. Με αυτόν τον τρόπο, κατά την εκπαίδευση του μοντέλου, η ατομική απόδοση του παίκτη συσχετίζεται όχι μόνο με την προηγούμενη αγωνιστική του συμπεριφορά, αλλά και με πληροφορίες που αφορούν την αντίπαλη ομάδα.

(Ενότητα 3.3.γ) Δημιουργία χρονικών ακολουθιών

Για την αξιοποίηση της χρονικής πληροφορίας από τα LSTM, τα δεδομένα μετατρέπονται σε ακολουθίες προηγούμενων αγώνων. Στο επίπεδο των ομάδων, η διαδικασία πραγματοποιείται από τη συνάρτηση **build_form_sequences()**, ενώ για τους παίκτες χρησιμοποιείται η **build_player_sequences()**. Και στις δύο περιπτώσεις, οι εγγραφές ταξινομούνται χρονικά και ομαδοποιούνται ανά ομάδα ή παίκτη και ανά σεζόν, ώστε οι ακολουθίες να μην περιλαμβάνουν δεδομένα από διαφορετικές αγωνιστικές περιόδους.

Οι χρονικές ακολουθίες αποτελούν διαφορετική μορφή εισόδου από τα συγκεντρωτικά χαρακτηριστικά των προηγούμενων υποενοτήτων. Αντί να συνοψίζουν το ιστορικό μέσω κυλιόμενων ή season-to-date μέσων όρων, διατηρούν ξεχωριστές παρατηρήσεις διαδοχικών προηγούμενων αγώνων, επιτρέποντας στο LSTM να επεξεργαστεί άμεσα την χρονική τους σειρά.

Για κάθε αγώνα ομάδας δημιουργείται μια ακολουθία που περιλαμβάνει τις παρατηρήσεις των τελευταίων $N=7$ αγώνων. Ο τρέχων αγώνας δεν συμπεριλαμβάνεται στην ακολουθία, καθώς η επιλογή του ιστορικού πραγματοποιείται μέχρι τη θέση που προηγείται της τρέχουσας εγγραφής. Με αυτόν τον τρόπο, η είσοδος του LSTM για έναν συγκεκριμένο αγώνα περιέχει αποκλειστικά πληροφορία που ήταν διαθέσιμη πριν από αυτόν.

Στην αρχή μιας σεζόν, δεν υπάρχουν πάντοτε επτά προηγούμενοι αγώνες. Για να διατηρείται σταθερό το μήκος της εισόδου του LSTM, οι θέσεις που δεν αντιστοιχούν σε διαθέσιμους προηγούμενους αγώνες συμπληρώνονται με μηδενικές τιμές (zero padding). Ειδικότερα, όταν δεν υπάρχει κανένας προηγούμενος αγώνας δημιουργείται μια πλήρως μηδενική ακολουθία, ενώ όταν υπάρχουν λιγότεροι από επτά, οι μηδενικές εγγραφές προστίθενται στην αρχή της ακολουθίας και ακολουθούν οι πραγματικές προηγούμενες παρατηρήσεις.

Η ίδια λογική εφαρμόζεται και στις ακολουθίες των παικτών, με διαφορετικό μήκος χρονικού παραθύρου. Για κάθε παίκτη και αγώνα, δημιουργείται ιστορικό των προηγούμενων $N=5$ εμφανίσεών του, ενώ σε περιπτώσεις ανεπαρκούς ιστορικού χρησιμοποιείται αντίστοιχο zero padding.

Κάθε ακολουθία αποθηκεύεται με κλειδί που αποτελείται από την ομάδα ή τον παίκτη και την ημερομηνία του αντίστοιχου αγώνα, ώστε να μπορεί στη συνέχεια να ανακτηθεί η κατάλληλη χρονική είσοδος κατά την εκπαίδευση και την πρόβλεψη. Στις ομάδες αποθηκεύονται ως (team, match_date) και στους παίκτες ως (player_name, match_date).

(Ενότητα 3.3.δ) Τυποποίηση χαρακτηριστικών και διαχείριση ελλειψών τιμών

Μετά τη δημιουργία των χαρακτηριστικών και των χρονικών ακολουθιών εφαρμόζεται τυποποίηση των αριθμητικών μεταβλητών, ώστε οι διαφορετικές κλίμακες τιμών να μην επηρεάζουν δυσανάλογα την εκπαίδευση των νευρωνικών μοντέλων. Για τον σκοπό αυτό

χρησιμοποιείται ο `StandardScaler`, ο οποίος μετασχηματίζει κάθε χαρακτηριστικό με βάση τον μέσο όρο και την τυπική απόκλιση που υπολογίζονται από τα δεδομένα εκπαίδευσης. Στο μοντέλο ομάδων χρησιμοποιούνται δύο διαφορετικοί μετασχηματιστές, ένας για τα στατικά χαρακτηριστικά και ένας για τις χρονικές ακολουθίες που εισάγονται στο LSTM. Πριν από την τυποποίηση των στατιστικών χαρακτηριστικών, τυχόν ελλειπείς τιμές αντικαθίστανται από τη διάμεσο του αντίστοιχου χαρακτηριστικού, η οποία υπολογίζεται αποκλειστικά από το σύνολο εκπαίδευσης. Στη συνέχεια, ο `static_scaler` προσαρμόζεται επίσης μόνο στις εγγραφές που ανήκουν στο `training set`.

Αντίστοιχα, για τις χρονικές ακολουθίες των ομάδων χρησιμοποιείται ξεχωριστός `sequence_scaler`. Κατά την προσαρμογή του `scaler` λαμβάνονται υπόψη μόνο οι πραγματικές παρατηρήσεις των ακολουθιών του συνόλου εκπαίδευσης, ενώ οι πλήρως μηδενικές γραμμές που έχουν προστεθεί λόγω `zero padding` εξαιρούνται. Μετά την προσαρμογή, ο ίδιος μετασχηματιστής εφαρμόζεται στις ακολουθίες της γηπεδούχου και της φιλοξενούμενης ομάδας.

Στο πλαίσιο της διαδικασίας `walk-forward evaluation`, η παραπάνω διαδικασία επαναλαμβάνεται ανεξάρτητα για κάθε χρονικό `fold`. Συγκεκριμένα, για κάθε `fold` υπολογίζονται εκ νέου οι διάμεσοι και οι παράμετροι τυποποίησης χρησιμοποιώντας αποκλειστικά τα δεδομένα εκπαίδευσης του συγκεκριμένου `fold`. Οι ίδιοι μετασχηματιστές εφαρμόζονται στη συνέχεια στα αντίστοιχα σύνολα επικύρωσης και ελέγχου, χωρίς να χρησιμοποιούνται δεδομένα από αυτά κατά την προσαρμογή των `scalers`. Με αυτό τον τρόπο αποφεύγεται η μεταφορά πληροφορίας από μεταγενέστερες χρονικές περιόδους προς τη διαδικασία εκπαίδευσης. Η συνάρτηση αξιολόγησης καλεί εκ νέου τη διαδικασία κανονικοποίησης για κάθε `fold` με το αντίστοιχο χρονικό όριο εκπαίδευσης.

Παρόμοια διαδικασία εφαρμόζεται και στα δεδομένα των παικτών. Για τα στατιστικά χαρακτηριστικά, οι ελλειπείς τιμές συμπληρώνονται με τη διάμεσο που προκύπτει από το `training set` και στη συνέχεια εφαρμόζεται `StandardScaler`, ο οποίος προσαρμόζεται αποκλειστικά στα δεδομένα εκπαίδευσης.

Για τις χρονικές ακολουθίες των παικτών δημιουργείται ξεχωριστός `sequence_scaler`, ο οποίος προσαρμόζεται μόνο σε προηγούμενες ακολουθίες που αντιστοιχούν σε ημερομηνίες πριν από το `train_cutoff`. Οι μηδενικές γραμμές που αντιστοιχούν σε `padding` εξαιρούνται από τον υπολογισμό των παραμέτρων του `scaler`. Οι τυποποιημένες ακολουθίες χρησιμοποιούνται κατά την εκπαίδευση του μοντέλου παικτών, ενώ οι αρχικές μη τυποποιημένες ακολουθίες διατηρούνται για την διαδικασία πρόβλεψης. Κατά το `inference` εφαρμόζεται εκ νέου ο ίδιος `sequence_scaler` που είχε προσαρμοστεί στο `training set`, ώστε η προεπεξεργασία να παραμένει συνεπής μεταξύ εκπαίδευσης και πρόβλεψης.

3.4 Αρχιτεκτονική του μοντέλου ομάδων

Το μοντέλο πρόβλεψης αγώνων υλοποιείται από την κλάση `SyndyasmenoMlPlstmModel` και ακολουθεί αρχιτεκτονική πολλαπλών εισόδων, συνδυάζοντας χρονικές ακολουθίες, στατικά χαρακτηριστικά και πληροφορία για το πρωτάθλημα. Η χρονική πληροφορία της γηπεδούχου και της φιλοξενούμενης ομάδας επεξεργάζεται μέσω LSTM, ενώ τα υπόλοιπα αριθμητικά χαρακτηριστικά του αγώνα επεξεργάζονται από ξεχωριστό MLP encoder.

Στα στατικά χαρακτηριστικά περιλαμβάνονται επίσης οι προ αγώνα αξιολογήσεις Elo των δύο ομάδων και η μεταξύ τους διαφορά.

Παράλληλα, το πρωτάθλημα αναπαρίσταται μέσω ενός εκπαιδευσιμου `embedding`.

Οι ακολουθίες των δύο ομάδων εισάγονται στον ίδιο LSTMEncoder, επομένως χρησιμοποιείται κοινό σύνολο βαρών για τη γηπεδούχο και τη φιλοξενούμενη ομάδα. Για κάθε ακολουθία, το LSTM επεξεργάζεται διαδοχικά τους προηγούμενους αγώνες και ως τελική χρονική αναπαράσταση χρησιμοποιείται η τελευταία κρυφή κατάσταση h_n . Με αυτόν τον τρόπο παράγεται ένα διάνυσμα που συνοψίζει την πρόσφατη αγωνιστική κατάσταση κάθε ομάδας.

Τα στατικά χαρακτηριστικά του αγώνα εισάγονται στον MLPDecoder. Ο συγκεκριμένος κλάδος αποτελείται από πλήρως συνδεδεμένα επίπεδα, συναρτήσεις ενεργοποίησης ReLU και Dropout και παράγει μία συμπυκνωμένη αναπαράσταση των μη χρονικών χαρακτηριστικών. Παράλληλα, κάθε πρωτάθλημα αντιστοιχίζεται σε έναν ακέραιο αριθμητικό δείκτη ($league_id$), ο οποίος χρησιμοποιείται ως είσοδος σε ένα επίπεδο Embedding. Η πληροφορία αυτή επιτρέπει στο μοντέλο να διατηρεί ξεχωριστή αναπαράσταση για κάθε πρωτάθλημα χωρίς να απαιτείται η χρήση ανεξάρτητου μοντέλου για κάθε ένα από αυτά.

Οι τέσσερις αναπαραστάσεις, δηλαδή η έξοδος LSTM της γηπεδούχου, η έξοδος LSTM της φιλοξενούμενης, η έξοδος MLP και το embedding του πρωταθλήματος, συνενώνονται σε ένα ενιαίο διάνυσμα. Το διάνυσμα αυτό εισάγεται σε ένα τελικό, πλήρως συνδεδεμένο τμήμα του δικτύου, το οποίο καταλήγει σε δύο αριθμητικές εξόδους, μία για κάθε ομάδα. Στην τελική παραμετροποίηση της υλοποίησης, το LSTM χρησιμοποιεί κρυφή διάσταση 64, ο MLP χρησιμοποιεί δύο κρυφά επίπεδα διαστάσεων 64 και 32 και παράγει έξοδο διάστασης 32, ενώ το embedding του πρωταθλήματος έχει διάσταση 8. Το τελικό κρυφό επίπεδο του prediction head έχει διάσταση 32. Συνεπώς, πριν από το τελικό prediction head σχηματίζεται διάνυσμα συνολικής διάστασης $64+64+32+8=168$.

Οι δύο τελικές εξόδοι του δικτύου αντιστοιχούν σε μη περιορισμένες τιμές z_H και z_A . Στην έξοδο της γηπεδούχου ομάδας προστίθεται μια εκπαιδεύσιμη παράμετρος $home_advantage$, ώστε η επίδραση της έδρας να εκτιμάται απευθείας από τα δεδομένα. Η παράμετρος αρχικοποιείται στην τιμή 0.1, αλλά μεταβάλλεται κατά την εκπαίδευση. Η μετατροπή των μη περιορισμένων εξόδων του δικτύου σε θετικές παραμέτρους έντασης βασίζεται στη λογαριθμική συνάρτηση σύνδεσης της Poisson παλινδρόμησης, της οποίας η αντίστροφη συνάρτηση είναι η εκθετική [30]. Οι τελικές παράμετροι έντασης τερμάτων υπολογίζονται ως:

$$\log \lambda_H = z_H + h$$

$$\log \lambda_A = z_A$$

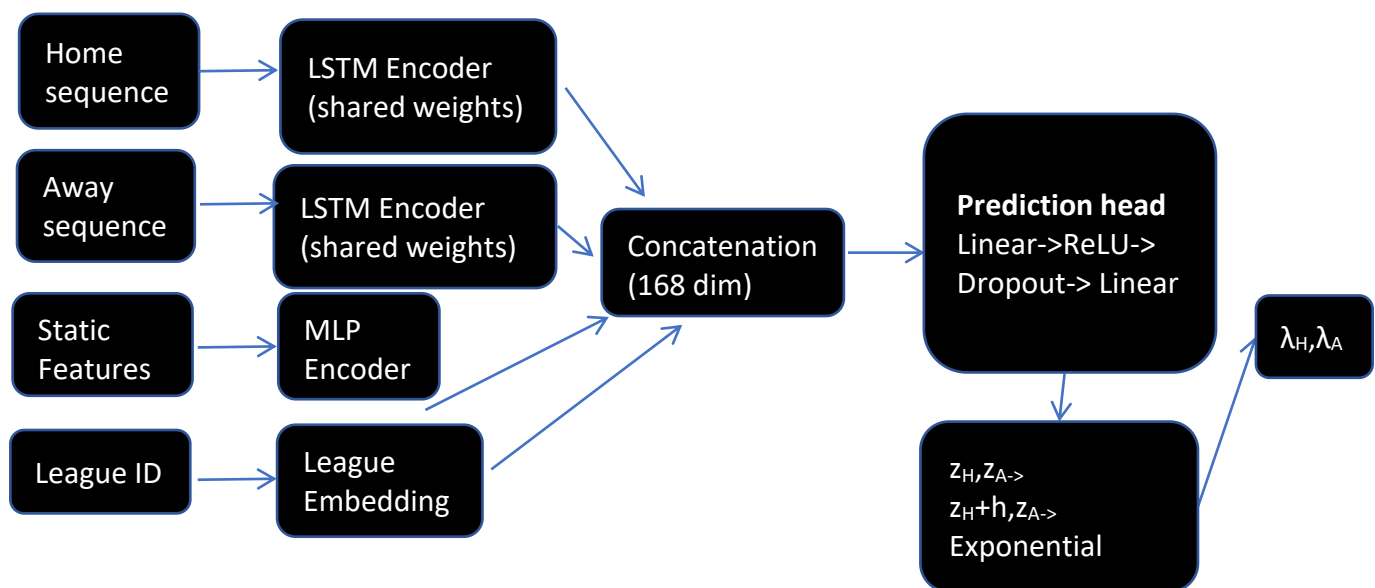
και επομένως:

$$\lambda_H = e^{z_H + h}$$

$$\lambda_A = e^{z_A}$$

όπου h είναι η εκπαιδεύσιμη παράμετρος έδρας. Η χρήση της εκθετικής συνάρτησης εξασφαλίζει ότι οι τιμές λ_H και λ_A είναι θετικές και μπορούν στη συνέχεια να χρησιμοποιηθούν ως παράμετροι στις κατανομές Poisson για την εκτίμηση του αριθμού τερμάτων των δύο ομάδων [6].

Η συνολική ροή της αρχιτεκτονικής του μοντέλου ομάδων συνοψίζεται ως:



Σχήμα 3.1 - Αρχιτεκτονική του μοντέλου πρόβλεψης αγώνων

3.5 Αρχιτεκτονική του μοντέλου παικτών

Το μοντέλο πρόβλεψης παικτών υλοποιείται από την κλάση **playerpredictionmodel** και ακολουθεί επίσης αρχιτεκτονική πολλαπλών εισόδων. Σε αυτή την περίπτωση συνδυάζονται η χρονική ακολουθία πρόσφατων εμφανίσεων του παίκτη, τα στατικά χαρακτηριστικά του εμπλουτισμένα με επιλεγμένα χαρακτηριστικά της αντίπαλης ομάδας, η θέση στην οποία αγωνίζεται και το πρωτάθλημα στο οποίο συμμετέχει. Η χρονική πληροφορία επεξεργάζεται από το LSTMEncoder, τα στατικά χαρακτηριστικά από το MLPEncoder, ενώ η θέση και το πρωτάθλημα μετατρέπονται αρχικά σε αριθμητικούς δείκτες και στη συνέχεια σε εκπαιδευσιμα διανύσματα μέσω των position_embedding και league_embedding. Οι τέσσερις αναπαραστάσεις συνενώνονται σε ένα κοινό διάνυσμα και εισάγονται σε ένα κοινό τμήμα του δικτύου (trunk), το οποίο αποτελείται από πλήρως συνδεδεμένο επίπεδο, συνάρτηση ενεργοποίησης ReLU και Dropout. Από την κοινή αυτή αναπαράσταση το δίκτυο διακλαδώνεται σε τρία ανεξάρτητα τμήματα εξόδου, καθένα από τα οποία είναι υπεύθυνο για διαφορετική κατηγορία προβλέψεων.

Το πρώτο τμήμα εξόδου αφορά την πρόβλεψη αριθμητικών μεταβλητών καταμέτρησης. Συγκεκριμένα, προβλέπονται τα γκολ (Gs), οι ασίστ (Ast), τα σουτ (Sh), τα σουτ εντός στόχου (SoT), τα φάουλ (Fs), οι αναχαιτίσεις (Int) και τα κερδισμένα τάκλιν (TklW). Στην έξοδο του συγκεκριμένου τμήματος εφαρμόζεται η εκθετική συνάρτηση, ώστε οι προβλεπόμενες τιμές να παραμένουν θετικές.

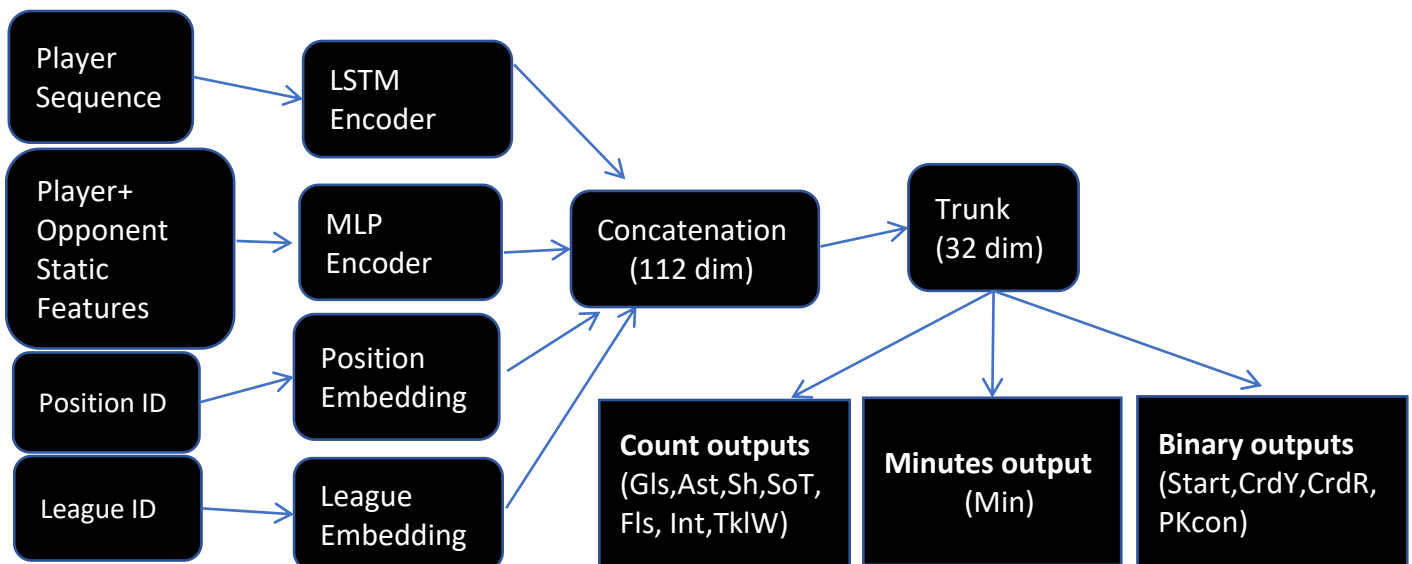
Το δεύτερο τμήμα εξόδου χρησιμοποιείται για την πρόβλεψη των αναμενόμενων λεπτών συμμετοχής του παίκτη. Η έξοδός του περνάει από τη συνάρτηση sigmoid και στη συνέχεια πολλαπλασιάζεται με τη μέγιστη τιμή των 120 λεπτών, με αποτέλεσμα η τελική πρόβλεψη να περιορίζεται στο διάστημα από 0 έως 120 λεπτά.

Το τρίτο τμήμα αφορά δυαδικά γεγονότα. Ειδικότερα, το μοντέλο εκτιμά αν ο παίκτης θα ξεκινήσει ως βασικός (Start), την πιθανότητα εμφάνισης κίτρινης ή κόκκινης κάρτας

(CrdY,CrdR), καθώς και την πιθανότητα παραχώρησης πέναλτι (PKcon). Οι έξοδοι παράγονται αρχικά σε μορφή logits και κατά τη διαδικασία πρόβλεψης μετατρέπονται σε πιθανότητες μέσω της συνάρτησης sigmoid.

Στην τελική παραμετροποίηση του μοντέλου παικτών, το LSTM χρησιμοποιεί κρυφή διάσταση 64, ενώ ο MLP Encoder παράγει διάνυσμα διάστασης 32. Τα embeddings της θέσης και του πρωταθλήματος έχουν διάσταση 8 το καθένα. Επομένως, μετά τη συνένωση των τεσσάρων αναπαραστάσεων, σχηματίζεται ένα κοινό διάνυσμα συνολικής διάστασης $64+32+8+8=112$. Το διάνυσμα αυτό εισάγεται στο κοινό τμήμα του δικτύου (trunk), το οποίο παράγει αναπαραστάση διάστασης 32 και τροφοδοτεί τα τρία ανεξάρτητα τμήματα εξόδου.

Η συνολική ροή της αρχιτεκτονικής του μοντέλου παικτών συνοψίζεται ως:



Σχήμα 3.2 - Αρχιτεκτονική του μοντέλου πρόβλεψης παικτών

3.6 Εκπαίδευση των μοντέλων

Για το μοντέλο ομάδων εφαρμόζεται διαδικασία walk-forward evaluation με τέσσερα διαδοχικά χρονικά folds. Σε κάθε fold, το χρονικά επόμενο σύνολο από το σύνολο εκπαίδευσης χρησιμοποιείται για επικύρωση και το μεταγενέστερο για backtesting.

Για το μοντέλο παικτών, χρησιμοποιείται ξεχωριστός χρονικός διαχωρισμός σε σύνολο εκπαίδευσης και σύνολο επικύρωσης.

(Ενότητα 3.6.α) Εκπαίδευση του μοντέλου ομάδων

Για το μοντέλο ομάδων, τα δεδομένα εκπαίδευσης οργανώνονται σε batches των 128 παρατηρήσεων και η ενημέρωση των βαρών πραγματοποιείται με τον βελτιστοποιητή Adam. Ως συνάρτηση απώλειας χρησιμοποιείται η PoissonNLLLoss, καθώς οι δύο έξοδοι του μοντέλου αντιστοιχούν στις παραμέτρους έντασης των γκολ της γηπεδούχου και της φιλοξενούμενης ομάδας. Η συνολική απώλεια κάθε batch υπολογίζεται ως το άθροισμα της Poisson απώλειας για τα γκολ των δύο ομάδων.

Για την αξιολόγηση, στο πρώτο fold χρησιμοποιούνται για εκπαίδευση οι αγώνες πριν από την 01/08/2022, για επικύρωση οι αγώνες από 01/08/2022 έως πριν από 01/12/2022 και για backtesting οι αγώνες από 01/12/2022 έως πριν από 01/08/2023. Στο δεύτερο fold, τα αντίστοιχα χρονικά όρια είναι 01/04/2023, 01/08/2023 και 01/04/2024, στο τρίτο 01/04/2024, 01/08/2024 και 01/04/2025, ενώ στο τέταρτο 01/04/2025, 01/08/2025 και 01/06/2026. Σε κάθε fold το μοντέλο εκπαιδεύεται εκ νέου από την αρχή, ενώ το σύνολο επικύρωσης χρησιμοποιείται για την παρακολούθηση του validation loss και την εφαρμογή early stopping.

Στην τελική παραμετροποίηση χρησιμοποιείται learning rate ίσο με 0.0003 και μέγιστος αριθμός 150 epochs. Παράλληλα εφαρμόζεται early stopping με patience ίσο με 20. Εάν η απώλεια επικύρωσης δεν παρουσιάσει βελτίωση για 20 συνεχόμενα epochs, η εκπαίδευση διακόπτεται. Κατά τη διαδικασία, διατηρείται η κατάσταση του μοντέλου με τη μικρότερη τιμή validation loss και αυτή επαναφέρεται στο τέλος της εκπαίδευσης.

(Ενότητα 3.6.β) Εκπαίδευση του μοντέλου παικτών

Για το μοντέλο παικτών χρησιμοποιείται ξεχωριστός χρονικός διαχωρισμός, όπου οι εγγραφές πριν από την 01/06/2025 χρησιμοποιούνται για εκπαίδευση, ενώ οι εγγραφές από την 01/06/2025 ως πριν από την 01/09/2025 χρησιμοποιούνται για επικύρωση.

Επίσης, χρησιμοποιείται ο βελτιστοποιητής Adam, αλλά η διαδικασία εκπαίδευσης περιλαμβάνει τρεις διαφορετικές συναρτήσεις απώλειας λόγω των διαφορετικών τύπων προβλέψεων. Για τις μεταβλητές καταμέτρησης χρησιμοποιείται PoissonNLLLoss, για τα δυαδικά γεγονότα BCEWithLogitsLoss, ενώ για τα λεπτά συμμετοχής χρησιμοποιείται MSELoss. Παρότι η έξοδος του μοντέλου εκφράζεται σε λεπτά στο διάστημα 0-120, κατά τον υπολογισμό της συγκεκριμένης απώλειας τόσο η προβλεπόμενη όσο και η πραγματική τιμή διαιρούνται με 120, ώστε η σύγκριση να πραγματοποιείται σε κανονικοποιημένη κλίμακα (0-1). Η συνολική απώλεια προκύπτει από το άθροισμα των τριών επιμέρους όρων. Σε αυτό το μοντέλο χρησιμοποιείται learning rate 0.0003, batch size 128, μέγιστος αριθμός 60 epochs και early stopping με patience ίσο με 10. Και σε αυτή τη περίπτωση διατηρείται η κατάσταση του μοντέλου που επιτυγχάνει τη μικρότερη validation loss και επαναφέρεται στο τέλος της εκπαίδευσης.

3.7 Διαδικασία παραγωγής προβλέψεων

Μετά την ολοκλήρωση της εκπαίδευσης, τα εκπαιδευμένα μοντέλα χρησιμοποιούνται σε κατάσταση αξιολόγησης (eval) για την παραγωγή νέων προβλέψεων. Κατά το inference εφαρμόζονται οι ίδιοι μετασχηματισμοί προεπεξεργασίας που είχαν υπολογιστεί από το σύνολο εκπαίδευσης, ώστε οι νέες είσοδοι να βρίσκονται στην ίδια κλίμακα με τα δεδομένα που χρησιμοποιήθηκαν κατά την εκπαίδευση. Η διαδικασία διαφέρει μεταξύ του μοντέλου ομάδων και του μοντέλου παικτών, ανάλογα με το είδος των παραγόμενων εξόδων.

(Ενότητα 3.7.α) Πρόβλεψη αγώνα

Για την πρόβλεψη ενός αγώνα χρησιμοποιείται η συνάρτηση **match_prediction()**. Αρχικά ανακτώνται τα πιο πρόσφατα διαθέσιμα στατικά χαρακτηριστικά και οι χρονικές ακολουθίες της γηπεδούχου και της φιλοξενούμενης ομάδας. Τα στατικά χαρακτηριστικά οργανώνονται σύμφωνα με τις μεταβλητές που χρησιμοποιήθηκαν κατά την εκπαίδευση, ενώ τυχόν ελλειπείς τιμές συμπληρώνονται με τις αντίστοιχες διαμέσους του training set. Στη συνέχεια εφαρμόζονται οι ίδιοι μετασχηματιστές `static_scaler` και `sequence_scaler` που είχαν εφαρμοστεί στα δεδομένα εκπαίδευσης.

Επιπλέον, για κάθε επιλεγμένο ζεύγος ομάδων υπολογίζεται εκ νέου η διαφορά των τελευταίων διαθέσιμων αξιολογήσεων `Elo` της γηπεδούχου και της φιλοξενούμενης, ώστε το χαρακτηριστικό `elo_diff` να αντιστοιχεί στη συγκεκριμένη αναμέτρηση.

Οι επεξεργασμένες είσοδοι μετατρέπονται σε `tensors` και εισάγονται στο εκπαιδευμένο μοντέλο, το οποίο παράγει τις δύο παραμέτρους έντασης λ_H και λ_A . Οι τιμές αυτές χρησιμοποιούνται στη συνέχεια από τη συνάρτηση **ypologismos_score()** για τον υπολογισμό των πιθανοτήτων των επιμέρους σκορ.

Για κάθε ομάδα υπολογίζονται οι πιθανότητες επίτευξης από 0 έως 6 τερμάτων και από τον συνδυασμό τους σχηματίζεται πίνακας πιθανοτήτων σκορ. Στις περιπτώσεις χαμηλού σκορ (0-0, 0-1, 1-0 και 1-1) εφαρμόζεται η διόρθωση Dixon-Coles με παράμετρο $\rho = -0.13$, ενώ στη συνέχεια ο πίνακας κανονικοποιείται ώστε το άθροισμα των πιθανοτήτων να ισούται με 1.

Από τον τελικό πίνακα υπολογίζονται η πιθανότητα νίκης γηπεδούχου, πιθανότητα ισοπαλίας και πιθανότητα νίκης φιλοξενούμενης. Ως πιθανότερο τελικό αποτέλεσμα επιλέγεται εκείνο που αντιστοιχεί στη μεγαλύτερη από τις τρεις πιθανότητες. Παράλληλα, ως πιθανότερο ακριβές σκορ επιλέγεται το κελί του πίνακα με τη μεγαλύτερη πιθανότητα. Τα αποτελέσματα παρουσιάζονται στον χρήστη μαζί με τις τρεις πιθανότητες τελικού αποτελέσματος.

(Ενότητα 3.7.β) Πρόβλεψη παικτών

Για την παραγωγή πρόβλεψης ενός παίκτη χρησιμοποιείται η συνάρτηση **playerprediction()**. Ως είσοδοι χρησιμοποιούνται τα πιο πρόσφατα διαθέσιμα στατικά χαρακτηριστικά του παίκτη, η χρονική ακολουθία των προηγούμενων εμφανίσεών του, η αγωνιστική του θέση και το πρωτάθλημα στο οποίο συμμετέχει. Για τα χαρακτηριστικά που αφορούν τον αντίπαλο χρησιμοποιούνται οι πιο πρόσφατες διαθέσιμες τιμές της ομάδας που έχει επιλεγεί ως αντίπαλος.

Πριν από την εισαγωγή τους στο μοντέλο, τα στατικά χαρακτηριστικά οργανώνονται σύμφωνα με τις μεταβλητές που χρησιμοποιήθηκαν κατά την εκπαίδευση και τυχόν

ελλιπείς τιμές συμπληρώνονται με τις αντίστοιχες διαμέσους του training set. Στη συνέχεια, τόσο τα στατικά χαρακτηριστικά όσο και η χρονική ακολουθία μετασχηματίζονται με τους αντίστοιχους scalers που είχαν προσαρμοστεί στα δεδομένα της εκπαίδευσης.

Η θέση και το πρωτάθλημα του παίκτη μετατρέπονται στους αντίστοιχους αριθμητικούς δείκτες που είχαν δημιουργηθεί κατά την εκπαίδευση. Οι επεξεργασμένες είσοδοι μετατρέπονται σε tensors και εισάγονται στο μοντέλο σε κατάσταση αξιολόγησης (eval), χωρίς υπολογισμό gradients. Το μοντέλο παράγει τις αριθμητικές προβλέψεις, την εκτίμηση των λεπτών συμμετοχής και τα logits των δυαδικών γεγονότων. Στα logits εφαρμόζεται η συνάρτηση sigmoid, ώστε να μετατραπούν σε πιθανότητες.

Το τελικό αποτέλεσμα περιλαμβάνει τις αναμενόμενες τιμές για τα επιμέρους αγωνιστικά στατιστικά, τα αναμενόμενα λεπτά συμμετοχής, καθώς και τις πιθανότητες συμμετοχής στην αρχική ενδεκάδα, κίτρινης/κόκκινης κάρτας και παραχώρησης πέναλτι. Για την πρόβλεψη συμμετοχής του παίκτη στην αρχική ενδεκάδα, τιμή πιθανότητας ίση ή μεγαλύτερη από 0.5 ερμηνεύεται ως θετική ένδειξη ότι ο παίκτης θα ξεκινήσει ως βασικός. Οι υπόλοιπες δυαδικές έξοδοι παρουσιάζονται απευθείας ως πιθανότητες.

(Ενότητα 3.7.γ) Προσομοίωση πρωταθλήματος

Εκτός από την πρόβλεψη μεμονωμένων αγώνων, το σύστημα παρέχει δυνατότητα προσομοίωσης ενός ολόκληρου πρωταθλήματος μέσω της συνάρτησης `simulate_league()`. Αρχικά, προσδιορίζονται οι ομάδες της πιο πρόσφατης διαθέσιμης σεζόν και ανακτώνται για καθεμία τα τελευταία διαθέσιμα στατικά χαρακτηριστικά και η πιο πρόσφατη χρονική ακολουθία. Στη συνέχεια, για κάθε δυνατό ζεύγος γηπεδούχου και φιλοξενούμενης ομάδας υπολογίζονται μέσω του μοντέλου οι αντίστοιχες παράμετροι λ_H και λ_A , οι οποίες αποθηκεύονται και επαναχρησιμοποιούνται στις προσομοιώσεις.

Η διαδικασία πραγματοποιείται με χίλιες επαναλήψεις. Σε κάθε προσομοίωση, ο αριθμός τερμάτων της γηπεδούχου και της φιλοξενούμενης παράγεται τυχαία από κατανομή Poisson με παραμέτρους τις αντίστοιχες τιμές λ_H και λ_A . Με βάση το σκορ που προκύπτει, αποδίδονται τρεις βαθμοί στη νικήτρια ομάδα ή ένας βαθμός σε κάθε ομάδα σε περίπτωση ισοπαλίας. Η διαδικασία επαναλαμβάνεται για όλα τα ζευγάρια του πρωταθλήματος και οι συνολικοί βαθμοί κάθε ομάδας καταγράφονται για κάθε προσομοίωση.

Μετά την ολοκλήρωση των προσομοιώσεων υπολογίζονται για κάθε ομάδα ο μέσος αριθμός βαθμών και η τυπική απόκλισή τους. Παράλληλα, η συχνότητα με την οποία μία ομάδα καταλαμβάνει την πρώτη θέση, χρησιμοποιείται ως εκτίμηση της πιθανότητας κατάκτησης του πρωταθλήματος, ενώ η συχνότητα εμφάνισής της στις τελευταίες θέσεις χρησιμοποιείται για την εκτίμηση της πιθανότητας υποβιβασμού. Με βάση τα αποτελέσματα αυτά, σχηματίζεται η τελική προβλεπόμενη κατάταξη του πρωταθλήματος.

Στην παρούσα υλοποίηση δεν χρησιμοποιούνται τα πραγματικά εναπομείναντα fixtures της σεζόν. Αντίθετα, προσομοιώνονται όλα τα δυνατά ζευγάρια ομάδων με εναλλαγή έδρας, βάσει της τελευταίας διαθέσιμης αγωνιστικής τους κατάστασης. Για κάθε ζεύγος ομάδων υπολογίζεται επίσης η αντίστοιχη διαφορά των τελευταίων διαθέσιμων αξιολογήσεων Elo πριν από την εισαγωγή των χαρακτηριστικών στο μοντέλο.

(Ενότητα 3.7.δ) Πρόσθετες προβλέψεις παικτών σε επίπεδο πρωταθλήματος

Το σύστημα παρέχει επίσης τη δυνατότητα παραγωγής συγκεντρωτικών προβλέψεων για τους παίκτες ενός επιλεγμένου πρωταθλήματος. Για τον σκοπό αυτό χρησιμοποιείται η

συνάρτηση **predict_all_league_players()**, η οποία εντοπίζει όλους τους διαθέσιμους παίκτες του πρωταθλήματος και για κάθε έναν ανακτά το πιο πρόσφατο διαθέσιμο σύνολο χαρακτηριστικών και την τελευταία χρονική ακολουθία των εμφανίσεών του. Στη συνέχεια, εφαρμόζεται η διαδικασία πρόβλεψης του μοντέλου παικτών και αποθηκεύονται οι αντίστοιχες εκτιμήσεις.

Με βάση τις προβλέψεις αυτές, οι παίκτες μπορούν να ταξινομηθούν σύμφωνα με διαφορετικά κριτήρια. Για την εκτίμηση των παικτών με τη μεγαλύτερη επιθετική συνεισφορά χρησιμοποιείται η προβλεπόμενη τιμή των γκολ, ενώ για τις πειθαρχικές προβλέψεις χρησιμοποιούνται οι πιθανότητες κίτρινης και κόκκινης κάρτας. Το σύστημα ταξινομεί τους παίκτες κατά φθίνουσα σειρά και εμφανίζει εκείνους με τις υψηλότερες αντίστοιχες τιμές.

Στην παρούσα υλοποίηση οι συγκεκριμένες κατατάξεις βασίζονται στο τελευταίο διαθέσιμο αγωνιστικό στιγμιότυπο κάθε παίκτη και όχι σε προσομοίωση όλων των μελλοντικών αγώνων της σεζόν. Επομένως, η κατάταξη βάσει αναμενόμενων γκολ αποτυπώνει κυρίως ποιοι παίκτες εμφανίζουν την υψηλότερη προβλεπόμενη επιθετική απόδοση με βάση τα τελευταία διαθέσιμα δεδομένα και δεν αποτελεί άμεση πρόβλεψη του συνολικού αριθμού τερμάτων που θα σημειώσουν μέχρι το τέλος της σεζόν.

3.8 Διεπαφή χρήστη και λειτουργία του συστήματος

Για την αλληλεπίδραση του χρήστη με το σύστημα, υλοποιήθηκε διεπαφή γραμμής εντολών (Command Line Interface), μέσω της οποίας παρέχεται πρόσβαση στις διαφορετικές λειτουργίες πρόβλεψης. Το κεντρικό μενού επιτρέπει την πρόβλεψη μεμονωμένου αγώνα, την εκτίμηση της τελικής κατάταξης ενός πρωταθλήματος, του πιθανού πρωταθλητή και των ομάδων με τη μεγαλύτερη πιθανότητα υποβιβασμού. Παράλληλα, παρέχονται λειτουργίες που αφορούν τους παίκτες, όπως η κατάταξη πιθανών κορυφαίων σκόρερ, οι παίκτες με τη μεγαλύτερη πιθανότητα κίτρινης ή κόκκινης κάρτας και η προβολή πληροφοριών για συγκεκριμένο παίκτη.

Στην περίπτωση πρόβλεψης αγώνα, ο χρήστης επιλέγει αρχικά το πρωτάθλημα που επιθυμεί και στη συνέχεια τη γηπεδούχο και τη φιλοξενούμενη ομάδα. Το σύστημα εμφανίζει το πιθανότερο αποτέλεσμα του αγώνα, το πιθανότερο ακριβές σκορ, καθώς και τις πιθανότητες νίκης της γηπεδούχου, ισοπαλίας και νίκης της φιλοξενούμενης ομάδας. Μετά την πρόβλεψη του αγώνα, δίνεται προαιρετικά η δυνατότητα παραγωγής προβλέψεων για τους παίκτες των δύο ομάδων, είτε για όλους τους διαθέσιμους παίκτες είτε για έναν συγκεκριμένο παίκτη.

Η διεπαφή λειτουργεί ως ενιαίο σημείο πρόσβασης στις επιμέρους δυνατότητες του συστήματος, συνδέοντας τις διαδικασίες προεπεξεργασίας και τα εκπαιδευμένα μοντέλα με τις τελικές λειτουργίες πρόβλεψης. Με αυτόν τον τρόπο, ο χρήστης μπορεί να αξιοποιήσει τις δυνατότητες του συστήματος χωρίς να απαιτείται άμεση αλληλεπίδραση με τον πηγαίο κώδικα.

1.Πρόβλεψη αγώνα (σκορ + παίκτες) 2.Πρόβλεψη κατάταξης πρωταθλήματος 3.Πρόβλεψη νικητή πρωταθλήματος 4.Πρόβλεψη υποβιβασμού 5.Πιθανός τοπ σκόρερ πρωταθλήματος 6.Παίκτες με τη μεγαλύτερη πιθανότητα κίτρινης 7.Παίκτες με τη μεγαλύτερη πιθανότητα κόκκινης 8.Πληροφορίες παίκτη 9.Έξοδος	Manchester City vs Tottenham (English) Πρόβλεψη: Πιο πιθανό αποτέλεσμα: Νίκη Manchester City (P=73.9%) Πιο πιθανό ΣΥΓΚΕΚΡΙΜΕΝΟ σκορ: Manchester City 2-0 Tottenham [11.1%] (Νίκη Manchester City: 73.9% Ισοπαλία: 17.2% Νίκη Tottenham: 8.9%)
--	---

Εικόνα 3.1 - Κεντρικό μενού του συστήματος

Εικόνα 3.2 - Παράδειγμα πρόβλεψης αγώνα

Παίκτης:Erling Haaland (FW-MF) Αντίπαλος: Tottenham ----- ----- Πιθανότητα να είναι στην αρχική 11άδα: YES (82.1%) Λεπτά που θα παίξει: 81.9 Γκολ που θα βάλει: 0.53 Πιθανές ασίστ: 0.16 Πιθανά σουτ: 3.06 Πιθανά σουτ στην εστία: 1.29 Πιθανά φάουλ: 0.87 Πιθανές επεμβάσεις: 0.2 Πιθανά τάκλιν: 0.35 Πιθανότητα κίτρινης κάρτας: 0.1 Πιθανότητα κόκκινης κάρτας: 0.002 Πιθανότητα να προκαλέσει πέναλτι: 0.0	Στοιχεία παίκτη: Erling Haaland position: FW-MF footed: Left height_cm: 191.0 weight_kg: 92.0 nationality: Norway Τρέχουσα ομάδα: Manchester City Πρωτάθλημα: English Πραγματικά στατιστικά: Gls: 0.76/αγώνα Ast: 0.24/αγώνα Sh: 3.65/αγώνα SoT: 1.71/αγώνα Fls: 0.68/αγώνα Int: 0.12/αγώνα TklW: 0.21/αγώνα Min: 84.21/αγώνα
---	---

Εικόνα 3.3 - Παράδειγμα πρόβλεψης παίκτη

Εικόνα 3.4 - Παράδειγμα πληροφοριών παίκτη

Πρόβλεψη κατάταξης Scotch (1000 simulations)						Πρόβλεψη υποβιβασμού English (1000 simulations)			
#	Ομάδα	Βαθμοί	±	Τίτλος	Υποβιβ.	#	Ομάδα	Βαθμοί	Υποβιβασμός
1	Celtic	44.9	5.8	52.9%	0.0%	1	Tottenham	32.3	62.2%
2	Hearts	42.9	6.0	33.3%	0.2%	2	Fulham	37.5	35.5%
3	Rangers	37.6	5.7	7.1%	0.7%	3	Crystal Palace	37.1	35.0%
4	Hibernian	34.9	5.8	4.4%	2.8%	4	Sunderland	38.4	28.1%
5	Motherwell	32.2	5.9	1.1%	6.5%	5	Leeds United	38.9	26.7%
6	Kilmarnock	27.5	6.0	0.4%	22.2%				
7	Falkirk	27.1	5.9	0.4%	27.2%				
8	Dundee United	26.5	6.0	0.3%	30.0%				
9	Dundee	25.1	5.6	0.1%	39.6%				
10	Aberdeen	24.6	5.7	0.0%	44.3%				
11	St Mirren	21.8	5.4	0.0%	58.0%				
12	Livingston	20.3	5.3	0.0%	68.5%				

Εικόνα 3.5 - Πρόβλεψη κατάταξης πρωταθλήματος

Εικόνα 3.6 - Πρόβλεψη ομάδων προς υποβίβαση

Παίκτες με μεγαλύτερη πιθανότητα κίτρινης κάρτας (English)				Παίκτες με μεγαλύτερη πιθανότητα κόκκινης κάρτας (English)			
#	Παίκτης	Ομάδα	P/αγώνα	#	Παίκτης	Ομάδα	P/αγώνα
1	Moisés Caicedo	Chelsea	25.1%	1	Emi Buendía	Aston Villa	1.8%
2	Saša Lukić	Fulham	24.7%	2	Anton Stach	Leeds United	1.6%
3	Emerson	Tottenham	23.6%	3	Luke Ayling	Leeds United	1.6%
4	Wilfred Ndidi	Leicester City	23.5%	4	Joelinton	Newcastle	1.5%
5	Ethan Ampadu	Leeds United	23.2%	5	Marcos Alonso	Chelsea	1.5%
6	James Garner	Everton	23.2%	6	Wesley Fofana	Chelsea	1.4%
7	Felipe	Nottingham	23.1%	7	Cristhian Mosquera	Arsenal	1.3%
8	Conor Gallagher	Tottenham	22.9%	8	Dejan Kulusevski	Tottenham	1.3%
9	João Gomes	Wolves	22.8%	9	Jonny Castro	Wolves	1.3%
10	Merlin Röhl	Everton	22.5%	10	Joshua Acheampong	Chelsea	1.3%
11	Rúben Neves	Wolves	22.5%	11	Luke Cundle	Wolves	1.3%
12	Casemiro	Manchester Utd	22.3%	12	Mason Holgate	Everton	1.3%
13	Cristhian Mosquera	Arsenal	22.1%	13	Milos Kerkez	Liverpool	1.3%
14	Sead Kolašinac	Arsenal	22.1%	14	Riccardo Calafiori	Arsenal	1.3%
15	Thomas Partey	Arsenal	22.1%	15	Felipe	Nottingham	1.2%

Εικόνες 3.7, 3.8 - Πρόβλεψη παικτών με τη μεγαλύτερη πιθανότητα λήψης κίτρινης/κόκκινης κάρτας

ΚΕΦΑΛΑΙΟ 4 - ΠΕΙΡΑΜΑΤΙΚΗ ΑΠΟΤΙΜΗΣΗ

Στο συγκεκριμένο κεφάλαιο παρουσιάζονται και αναλύονται τα αποτελέσματα της αξιολόγησης των μοντέλων που αναπτύχθηκαν. Για το μοντέλο πρόβλεψης αγώνων, εξετάζεται η ακρίβεια πρόβλεψης του τελικού αποτελέσματος, η ακρίβεια πρόβλεψης του ακριβούς σκορ και η απόδοση στα διαφορετικά πρωταθλήματα. Παράλληλα, παρουσιάζεται η συμπεριφορά του μοντέλου πρόβλεψης παικτών, οι διαφορετικές κατηγορίες εξόδων που παράγει και η εξέλιξη της διαδικασίας εκπαίδευσής του.

4.1 Πλαίσιο αξιολόγησης

Τα χρονικά διαστήματα που χρησιμοποιήθηκαν για την αξιολόγηση του μοντέλου αγώνων παρουσιάζονται στον Πίνακα 4.1.

Πίνακας 4.1 - Χρονικός διαχωρισμός των folds

Fold	Εκπαίδευση	Επικύρωση	Έλεγχος
1	Πριν από 01/08/2022	01/08/2022 - 30/11/2022	01/12/2022 - 31/07/2023
2	Πριν από 01/04/2023	01/04/2023 - 31/07/2023	01/08/2023 - 31/03/2024
3	Πριν από 01/04/2024	01/04/2024 - 31/07/2024	01/08/2024 - 31/03/2025
4	Πριν από 01/04/2025	01/04/2025 - 31/07/2025	01/08/2025 - 31/05/2026

4.2 Συνολικά αποτελέσματα πρόβλεψης αγώνα

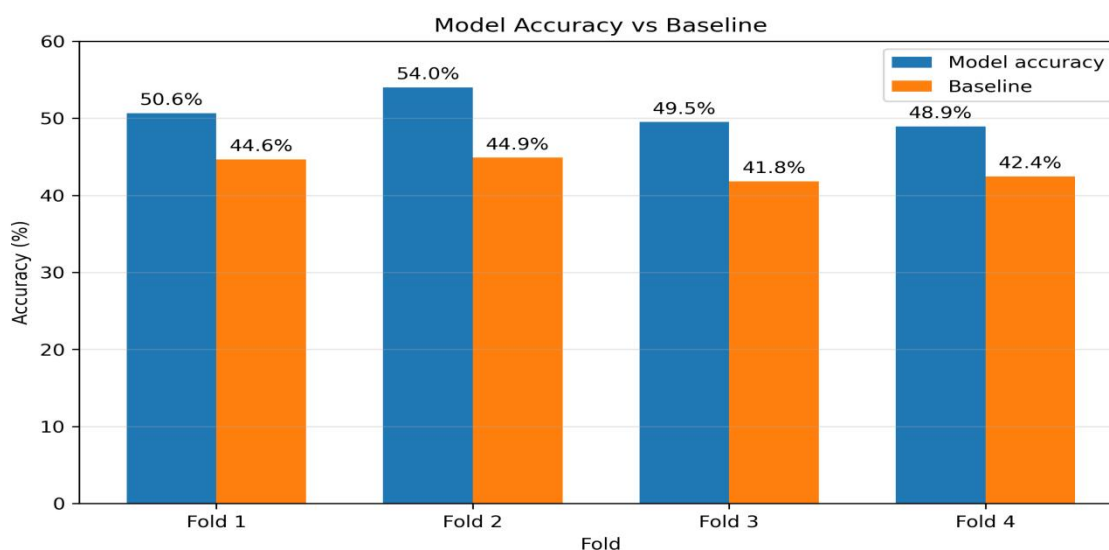
Τα αποτελέσματα της αξιολόγησης του μοντέλου πρόβλεψης αγώνων παρουσιάζονται στον πίνακα 4.2. Η ακρίβεια πρόβλεψης του τελικού αποτελέσματος κυμάνθηκε από 48,9% έως

και 54,0% στα τέσσερα folds, με μέση τιμή 50,8% και τυπική απόκλιση περίπου 2,0 ποσοστιαίων μονάδων.

Πίνακας 4.2 - Αποτελέσματα αξιολόγησης ανά fold

Fold	Αγώνες	Ακρίβεια αποτελέσματος	Ακρίβεια ακριβούς σκορ	Baseline
1	616	50,6%	14,1%	44,6%
2	800	54,0%	11,8%	44,9%
3	794	49,5%	12,3%	41,8%
4	740	48,9%	11,1%	42,4%
Μέσος όρος	-	50,8%	12,3%	43,4%

Σε όλα τα folds, η ακρίβεια του μοντέλου ήταν υψηλότερη από το αντίστοιχο baseline. Συνολικά, το μοντέλο παρουσίασε μέση βελτίωση περίπου 7,4 ποσοστιαίων μονάδων σε σχέση με το baseline. Η μεγαλύτερη ακρίβεια καταγράφηκε στο δεύτερο fold, όπου το μοντέλο έφτασε το 54,0%, ενώ η χαμηλότερη εμφανίστηκε στο τέταρτο fold με 48,9%. Παράλληλα, η ακρίβεια πρόβλεψης του ακριβούς σκορ παρέμεινε χαμηλότερη, με τιμές μεταξύ 11,1% και 14,1% και μέσο όρο περίπου 12,3%. Η σύγκριση της απόδοσης του μοντέλου με το baseline στα τέσσερα folds παρουσιάζεται στο Σχήμα 4.1.



Σχήμα 4.1: Σύγκριση της ακρίβειας του μοντέλου με το baseline στα τέσσερα folds

4.3 Αποτελέσματα ανά πρωτάθλημα

Η απόδοση του μοντέλου παρουσίασε διαφορές μεταξύ των πρωταθλημάτων και των χρονικών folds. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 4.3.

Πίνακας 4.3 - Ακρίβεια πρόβλεψης τελικού αποτελέσματος ανά πρωτάθλημα

Fold	Αγγλικό	Ιταλικό	Σκωτσέζικο	Ελληνικό
1	56,2%	43,5%	57,0%	48,3%
2	55,0%	53,3%	52,4%	55,1%
3	41,3%	53,5%	53,8%	49,7%
4	46,3%	47,1%	53,9%	-

Το σκωτσέζικο πρωτάθλημα παρουσίασε την πιο σταθερή συμπεριφορά, με την ακρίβεια να παραμένει πάνω από 52,0% σε όλα τα folds. Αντίθετα, στο αγγλικό πρωτάθλημα παρατηρήθηκε μεγαλύτερη μεταβολή, καθώς η ακρίβεια μειώθηκε από 56,2% στο πρώτο fold σε 41,3% στο τρίτο.

Στο Ιταλικό πρωτάθλημα η καλύτερη επίδοση παρατηρήθηκε στο τρίτο fold με 53,5%, ενώ στο ελληνικό η υψηλότερη ακρίβεια ήταν 55,1% στο δεύτερο fold. Για το ελληνικό πρωτάθλημα δεν υπήρχαν διαθέσιμοι αγώνες στο χρονικό διάστημα ελέγχου του τέταρτου fold και γι' αυτό το λόγο δεν υπολογίστηκε η αντίστοιχη ακρίβεια.

4.4 Ανάλυση προβλέψεων τελικού αποτελέσματος

Πέρα από τη συνολική ακρίβεια, εξετάστηκε και η κατανομή των προβλέψεων στις τρεις δυνατές κατηγορίες αποτελέσματος: νίκη γηπεδούχου, ισοπαλία και νίκη φιλοξενούμενης. Από την ανάλυση προέκυψε ότι το μοντέλο παρουσίασε δυσκολία στην επιλογή της ισοπαλίας ως πιθανότερου αποτελέσματος.

Συγκεκριμένα, στο πρώτο fold προβλέφθηκε μία μόνο ισοπαλία, ενώ στα υπόλοιπα τρία folds δεν προβλέφθηκε καμία. Επομένως, οι τελικές προβλέψεις του μοντέλου κατευθύνονταν σχεδόν αποκλειστικά στη νίκη μίας από τις δύο ομάδες.

Αυτή η συμπεριφορά αποτελεί έναν από τους βασικούς περιορισμούς του μοντέλου. Παρότι η συνολική ακρίβεια πρόβλεψης του μοντέλου ξεπέρασε το baseline σε όλα τα folds, η πολύ περιορισμένη πρόβλεψη ισοπαλιών δείχνει ότι η απόδοση δεν είναι ομοιόμορφη μεταξύ των τριών κατηγοριών αποτελέσματος. Στις εικόνες που ακολουθούν, αναγράφεται η κατανομή των πραγματικών αποτελεσμάτων και η κατανομή των προβλέψεων του μοντέλου.

=====	=====
=====	=====
BACKTEST -- 616 αγώνες (skipped: 0)	BACKTEST -- 800 αγώνες (skipped: 0)
=====	=====
Κατανομή πραγματικών αποτελεσμάτων: H=275	Κατανομή πραγματικών αποτελεσμάτων: H=359
D=154 A=187	D=207 A=234
Κατανομή ΠΡΟΒΛΕΨΕΩΝ μοντέλου: H=391	Κατανομή ΠΡΟΒΛΕΨΕΩΝ μοντέλου: H=595
D=1 A=224	D=0 A=205
Baseline (πάντα 'H'): 44.6%	Baseline (πάντα 'H'): 44.9%
Μοντέλο -- winner accuracy: 50.6%	Μοντέλο -- winner accuracy: 54.0%
Μοντέλο -- exact score accuracy: 14.1%	Μοντέλο -- exact score accuracy: 11.8%
=====	=====

Εικόνα 4.1 - Fold 1

Εικόνα 4.2 - Fold 2

```

=====
BACKTEST -- 794 αγώνες (skipped: 0)
=====
Κατανομή πραγματικών αποτελεσμάτων: H=332
D=199 A=263
Κατανομή ΠΡΟΒΛΕΨΕΩΝ μοντέλου: H=574
D=0 A=220
Baseline (πάντα 'H'): 41.8%
Μοντέλο -- winner accuracy: 49.5%
Μοντέλο -- exact score accuracy: 12.3%
=====

```

Εικόνα 4.3 - Fold 3

```

=====
BACKTEST -- 740 αγώνες (skipped: 0)
=====
Κατανομή πραγματικών αποτελεσμάτων: H=314
D=199 A=227
Κατανομή ΠΡΟΒΛΕΨΕΩΝ μοντέλου: H=552
D=0 A=188
Baseline (πάντα 'H'): 42.4%
Μοντέλο -- winner accuracy: 48.9%
Μοντέλο -- exact score accuracy: 11.1%
=====

```

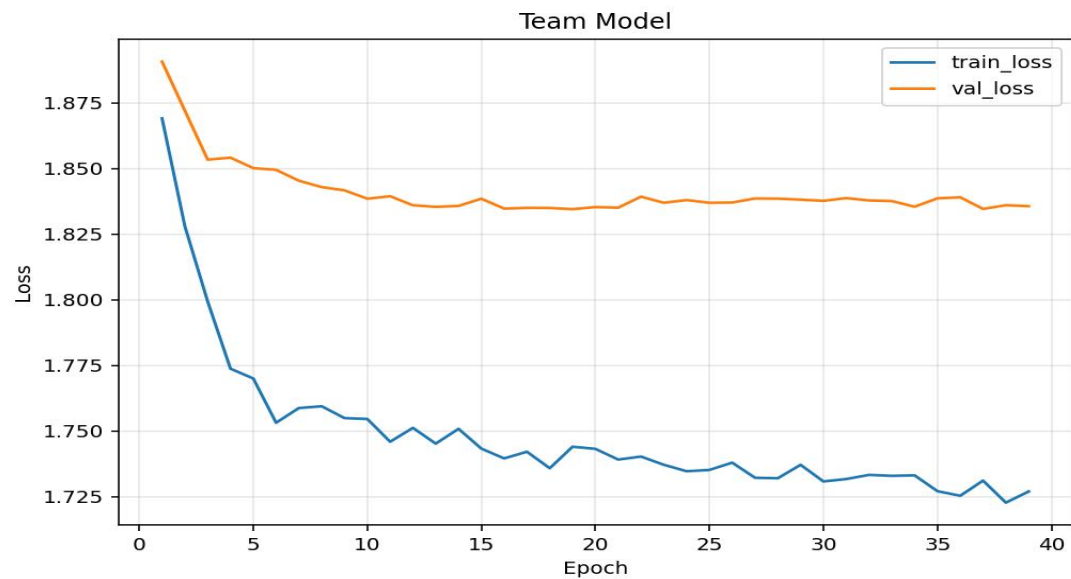
Εικόνα 4.4- Fold 4

4.5 Επίδραση του Elo στην απόδοση του μοντέλου

Για να εξεταστεί η συνεισφορά της μεταβλητής Elo, συγκρίθηκε η απόδοση του τελικού μοντέλου με αντίστοιχη εκδοχή χωρίς της συγκεκριμένη πληροφορία. Χωρίς τη χρήση του Elo, η μέση ακρίβεια πρόβλεψης του τελικού αποτελέσματος ήταν $49,9\% \pm 2,0\%$, ενώ με την ενσωμάτωση του αυξήθηκε σε $50,8\% \pm 2,0\%$. Η διαφορά αντιστοιχεί σε βελτίωση περίπου 0,9 ποσοστιαίων μονάδων. Παρότι η μεταβολή είναι μικρή, δείχνει ότι η πληροφορία σχετικής δυναμικότητας που παρέχει το Elo, συμβάλλει θετικά στη συνολική απόδοση του μοντέλου.

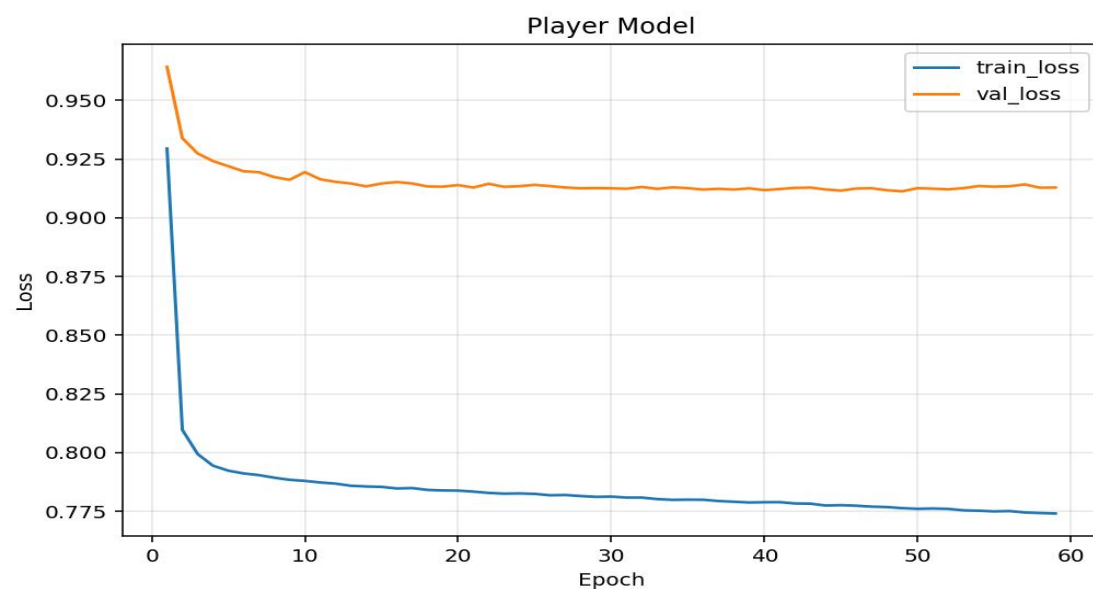
4.6 Εξέλιξη της εκπαίδευσης των μοντέλων

Κατά την εκπαίδευση των μοντέλων παρακολουθήθηκαν οι τιμές του training loss και validation loss, ενώ εφαρμόστηκε early stopping με διατήρηση της κατάστασης κάθε μοντέλου που αντιστοιχούσε στη χαμηλότερη τιμή validation loss. Η εξέλιξη αυτών των τιμών για το μοντέλο πρόβλεψης αγώνων και το μοντέλο πρόβλεψης παικτών παρουσιάζεται στα ακόλουθα σχήματα.



Σχήμα 4.2 - Εξέλιξη του training loss και validation loss κατά την εκπαίδευση του μοντέλου ομάδων

Στο Σχήμα 4.2, παρατηρείται σημαντική μείωση του training και validation loss στα πρώτα epochs. Στη συνέχεια, το validation loss παρουσιάζει μικρότερες μεταβολές και σταθεροποιείται, ενώ το training loss συνεχίζει να μειώνεται.



Σχήμα 4.3 - Εξέλιξη του training loss και validation loss κατά την εκπαίδευση του μοντέλου παικτών.

Όπως φαίνεται στο Σχήμα 4.3, το training loss μειώνεται σταθερά κατά τη διάρκεια της εκπαίδευσης, ενώ το validation loss παρουσιάζει έντονη αρχική μείωση και στη συνέχεια σταθεροποιείται. Η μεγαλύτερη μείωση του validation loss παρατηρείται στα αρχικά epochs.

4.7 Σύνοψη αποτελεσμάτων και περιορισμοί

Η συνολική εικόνα της αξιολόγησης δείχνει ότι το μοντέλο πρόβλεψης αγώνων παρουσίασε σταθερά καλύτερη επίδοση από το baseline στα τέσσερα χρονικά folds, αν και καταγράφηκαν διαφοροποιήσεις μεταξύ των επιμέρους πρωταθλημάτων. Παράλληλα, η ενσωμάτωση του Elo οδήγησε σε μικρή βελτίωση της συνολικής ακρίβειας.

Ο σημαντικότερος περιορισμός του μοντέλου αφορά την πρόβλεψη ισοπαλιών, καθώς αυτές επιλέχθηκαν αρκετά σπάνια ως πιθανότερο αποτέλεσμα. Αυτό δείχνει ότι παρά τη συνολική απόδοση του συστήματος, η συμπεριφορά του δεν είναι εξίσου ισορροπημένη μεταξύ των τριών κατηγοριών τελικού αποτελέσματος.

Όσον αφορά το μοντέλο πρόβλεψης παικτών, δεν πραγματοποιήθηκε ξεχωριστό backtesting με επιμέρους μετρικές για κάθε μεταβλητή εξόδου. Συνεπώς, η αξιολόγηση του συγκεκριμένου μοντέλου είναι πιο περιορισμένη σε σχέση με το μοντέλο πρόβλεψης αγώνων.

Τέλος, η απόδοση του συστήματος επηρεάζεται σε σημαντικό βαθμό από την ποιότητα και την επάρκεια των ιστορικών δεδομένων που χρησιμοποιούνται ως είσοδος.

ΚΕΦΑΛΑΙΟ 5 - Συμπεράσματα και μελλοντικές προεκτάσεις

5.1 Γενικά συμπεράσματα

Στην παρούσα εργασία αναπτύχθηκε ένα ολοκληρωμένο σύστημα πρόβλεψης ποδοσφαιρικών αγώνων και ατομικής απόδοσης παικτών, βασισμένο σε συνδυασμό νευρωνικών δικτύων και στατιστικών μεθόδων. Για την πρόβλεψη αγώνων αξιοποιήθηκαν χρονικές ακολουθίες πρόσφατης αγωνιστικής εικόνας και στατιστικά χαρακτηριστικά των ομάδων, ενώ για την παραγωγή πιθανοτήτων σκορ χρησιμοποιήθηκαν κατανομές Poisson σε συνδυασμό με τη διόρθωση Dixon - Coles.

Η αξιολόγηση του μοντέλου αγώνων έδειξε ότι η προτεινόμενη προσέγγιση μπορεί να ξεπεράσει ένα απλό majority-class baseline, επιτυγχάνοντας μέση ακρίβεια 50,8% έναντι 43,4%. Παράλληλα, η ενσωμάτωση του Elo παρείχε μικρή επιπλέον βελτίωση, ενώ η αξιολόγηση σε διαφορετικά χρονικά folds έδειξε σταθερή συμπεριφορά του μοντέλου.

Ιδιαίτερη σημασία έχει και η ανάπτυξη του μοντέλου παικτών, το οποίο επεκτείνει το σύστημα πέρα από την πρόβλεψη της τελικής έκβασης ενός αγώνα και επιτρέπει την παραγωγή εξατομικευμένων προβλέψεων για την αναμενόμενη απόδοση ενός παίκτη.

Συνολικά, το σύστημα που αναπτύχθηκε επιτυγχάνει τον βασικό στόχο της εργασίας, καθώς συνδυάζει διαφορετικές τεχνικές μοντελοποίησης σε μια ενιαία εφαρμογή ικανή να

παράγει προβλέψεις τόσο σε επίπεδο αγώνα όσο και σε επίπεδο παίκτη και πρωταθλήματος.

5.2 Προβλήματα και περιορισμοί της εργασίας

Κατά την ανάπτυξη της εργασίας, προέκυψαν ορισμένες δυσκολίες που σχετίζονται κυρίως με τη διαθεσιμότητα και τη διαχείριση των δεδομένων. Αρχικά, η εύρεση επαρκών και κατάλληλα δομημένων δεδομένων αποτέλεσε σημαντική πρόκληση, καθώς δεν ήταν εύκολο να εντοπιστεί μια ενιαία πηγή που να περιλαμβάνει τα απαιτούμενα στατιστικά για ομάδες και παίκτες σε όλα τα πρωταθλήματα και τις περιόδους που εξετάστηκαν. Γι' αυτό το λόγο αξιοποιήθηκαν δεδομένα από τον ιστότοπο FBref και αναπτύχθηκε διαδικασία αυτοματοποιημένης συλλογής τους μέσω scraping.

Παράλληλα, παρατηρήθηκαν διαφορές στην πληρότητα και στη διαθεσιμότητα των δεδομένων μεταξύ πρωταθλημάτων και αγωνιστικών περιόδων. Αυτό είχε ως αποτέλεσμα να μην είναι πάντα διαθέσιμος ο ίδιος όγκος πληροφορίας για όλες τις περιπτώσεις και να απαιτείται προσεκτική επεξεργασία και καθαρισμός των δεδομένων πριν από τη χρήση τους.

Ιδιαίτερη προσοχή απαιτήθηκε και στη χρονική οργάνωση των δεδομένων, ώστε να αποφεύγεται η χρήση πληροφορίας που δεν ήταν διαθέσιμη τη χρονική στιγμή μιας πρόβλεψης. Ο σωστός διαχωρισμός των δεδομένων και η δημιουργία των χρονικών ακολουθιών αποτέλεσαν κρίσιμο μέρος της διαδικασίας προεπεξεργασίας.

Στην περίπτωση των δεδομένων παικτών προέκυψαν επιπλέον δυσκολίες, κυρίως λόγω μεταγραφών, αλλαγών ομάδας μέσα στην ίδια αγωνιστική περίοδο και της ανάγκης αντιστοίχισης κάθε παίκτη με τη σωστή ομάδα και τον αντίπαλο της εκάστοτε πρόβλεψης. Η σωστή διαχείριση αυτών των περιπτώσεων ήταν απαραίτητη ώστε οι τελικές προβλέψεις να βασίζονται στα πιο πρόσφατα και κατάλληλα διαθέσιμα δεδομένα.

Ένας από τους βασικότερους περιορισμούς του μοντέλου πρόβλεψης αγώνων ήταν η πολύ περιορισμένη πρόβλεψη ισοπαλιών, καθώς το σύστημα επέλεγε σχεδόν πάντα μία από τις δύο ομάδες ως πιθανότερο νικητή.

5.3 Μελλοντικές προεκτάσεις

Μία από τις μελλοντικές κατευθύνσεις για τη βελτίωση του συστήματος είναι η καλύτερη αντιμετώπιση της πρόβλεψης ισοπαλιών. Η παρούσα υλοποίηση εμφάνισε περιορισμένη ικανότητα επιλογής ισοπαλίας ως πιθανότερου αποτελέσματος, επομένως θα μπορούσαν να εξεταστούν διαφορετικές τεχνικές εκπαίδευσης, στάθμιση των κατηγοριών ή μέθοδοι βαθμονόμησης των τελικών πιθανοτήτων.

Επιπλέον, το μοντέλο παικτών θα μπορούσε να αξιολογηθεί με ξεχωριστή διαδικασία backtesting και με κατάλληλες μετρικές για κάθε μεταβλητή εξόδου. Για τις αριθμητικές προβλέψεις θα μπορούσαν να χρησιμοποιηθούν μέτρα όπως MAE και RMSE, ενώ για τις δυαδικές προβλέψεις θα μπορούσαν να εξεταστούν μετρικές όπως accuracy, precision, recall και F1-score.

Μια ακόμη πιθανή επέκταση αφορά την ενσωμάτωση πρόσθετων δεδομένων, όπως τραυματισμοί, τιμωρίες, πιθανές ενδεκάδες, μεταγραφές και δεδομένα που σχετίζονται με την απόσταση που διανύει κάθε παίκτης κατά τη διάρκεια του αγώνα. Η χρήση περισσότερων πρωταθλημάτων και αγωνιστικών περιόδων θα μπορούσε να προσφέρει

μεγαλύτερο όγκο και εύρος δεδομένων και να ενισχύσει τη δυνατότητα γενίκευσης του συστήματος.

Ακόμη, το σύστημα θα μπορούσε να συνδεθεί με εξωτερική πηγή δεδομένων, ώστε να ενημερώνεται αυτόματα με νέα αποτελέσματα, στατιστικά ομάδων και παικτών πριν από κάθε αγωνιστική.

Σε επίπεδο μοντελοποίησης, θα μπορούσαν να εξεταστούν εναλλακτικές αρχιτεκτονικές νευρωνικών δικτύων, όπως GRU, Transformer ή μοντέλα με μηχανισμούς attention, καθώς και πιο συστηματικές διαδικασίες βελτιστοποίησης υπερπαραμέτρων. Παράλληλα, η προσομοίωση του πρωταθλήματος θα μπορούσε να επεκταθεί ώστε να βασίζεται στο πραγματικό πρόγραμμα υπολειπόμενων αγώνων κάθε διοργάνωσης.

Μια ακόμη μελλοντική βελτίωση θα μπορούσε να αφορά την καλύτερη ερμηνεία των προβλέψεων του μοντέλου. Συγκεκριμένα, θα μπορούσαν να χρησιμοποιηθούν τεχνικές που να δείχνουν ποια χαρακτηριστικά επηρέασαν περισσότερο κάθε πρόβλεψη. Με αυτόν τον τρόπο, ο χρήστης θα μπορεί να κατανοεί καλύτερα γιατί το σύστημα κατέληξε σε ένα συγκεκριμένο αποτέλεσμα.

Τέλος, μια γραφική ή διαδικτυακή διεπαφή θα μπορούσε να βελτιώσει σημαντικά την ευχρηστία του συστήματος και να επιτρέψει την παρουσίαση των προβλέψεων με πιο άμεσο και διαδραστικό τρόπο.

BIBΛΙΟΓΡΑΦΙΑ

- [1] Rein R, Memmert D. Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. Springerplus. 2016 Aug 24;5(1):1410. doi: 10.1186/s40064-016-3108-2. PMID: 27610328; PMCID: PMC4996805.
- [2] Pollard R. Home Advantage in Football: A Current Review of an Unsolved Puzzle . Open Sports Sci J, 2008; 1: . <http://dx.doi.org/10.2174/1875399X00801010012>
- [3] Bunker, R. P., & Thabtah, F. (2019). *A machine learning framework for sport result prediction*. Applied Computing and Informatics, 15(1), 27–33. DOI: 10.1016/j.aci.2017.09.005.
- [4] Sports Reference, LLC. FBref: Football Stats and History Statistics. Available:<https://fbref.com/en/>
- [5] Ross, S. M. (2014). *Introduction to probability models* (11th ed.). Academic Press.
- [6] Maher, M.J. (1982), Modelling association football scores. Statistica Neerlandica, 36: 109-118. <https://doi.org/10.1111/j.1467-9574.1982.tb00782.x>
- [7] Dixon, M.J. and Coles, S.G. (1997), Modelling Association Football Scores and Inefficiencies in the Football Betting Market. Journal of the Royal Statistical Society: Series C (Applied Statistics), 46: 265-280. <https://doi.org/10.1111/1467-9876.00065>
- [8] Sepp Hochreiter, Jürgen Schmidhuber; Long Short-Term Memory. *Neural Comput* 1997; 9 (8): 1735–1780. doi: <https://doi.org/10.1162/neco.1997.9.8.1735>
- [9] Razvan Pascanu, Tomas Mikolov, Yoshua Bengio. On the difficulty of training Recurrent Neural Networks. Proceedings of the 30th International Conference on Machine Learning, PMLR 28(3):1310-1318, 2013.
- [10] Gers FA, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM. *Neural Comput*. 2000 Oct;12(10):2451-71. doi: 10.1162/089976600300015015. PMID: 11032042.
- [11] Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2023). *Dive into Deep Learning*. Cambridge University Press. Available: <https://d2l.ai/>
- [12] Rumelhart, D., Hinton, G. & Williams, R. Learning representations by back-propagating errors. *Nature* 323, 533–536 (1986). <https://doi.org/10.1038/323533a0>
- [13] Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML'10). Omnipress, Madison, WI, USA, 807–814.
- [14] T. Baltrušaitis, C. Ahuja and L. -P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423-443, 1 Feb. 2019, doi: 10.1109/TPAMI.2018.2798607.
- [15] Guo, C., &Berkhahn, F.(2016).Entity embedding of categorical Variables. arXiv preprint arXiv:1604.06737.
- [16] Caruana, R. Multitask Learning. *Machine Learning* 28, 41–75 (1997). <https://doi.org/10.1023/A:1007379606734>
- [17] Ruder, S. (2017). An Overview of Multi-Task Learning in Deep Neural Networks. arXiv:1706.05098.

- [18] Suryadi, Chew, L.Y. & Ong, YS. Jacobian Granger causality for count and binary data with applications to causal network inference. *Sci Rep* 16, 3452 (2026). <https://doi.org/10.1038/s41598-025-33385-w>
- [19] Mehta P, Wang CH, Day AGR, Richardson C, Bukov M, Fisher CK, Schwab DJ. A high-bias, low-variance introduction to Machine Learning for physicists. *Phys Rep*. 2019 May 30;810:1-124. doi: 10.1016/j.physrep.2019.03.001.
- [20] Kingma, D.P., & Ba, J. (2017). Adam: A Method for Stochastic Optimization. 3rd International Conference on Learning Representations (ICLR). arXiv:1412.6980.
- [21] Prechelt, L. (1998). Early Stopping — But When? In G. B. Orr & K.-R. Müller (Eds.), *Neural Networks: Tricks of the Trade*, Lecture Notes in Computer Science, Vol. 1524, pp. 55–69. Springer. DOI: 10.1007/3-540-49430-8_3.
- [22] Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6(4), Article 15, 1-21. <https://doi.org/10.1145/2382577.2382579>
- [23] Cerqueira, V., Torgo, L., & Mozetič, I. (2020). Evaluating Time Series Forecasting Models: An Empirical Study on Performance Estimation Methods. *Machine Learning*, 109, 1997–2028. DOI: 10.1007/s10994-020-05910-7.
- [24] Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov; 15(56):1929–1958, 2014.
- [25] Cao, X.H., Stojkovic, I. & Obradovic, Z. A robust data scaling algorithm to improve classification accuracies in biomedical data. *BMC Bioinformatics* 17, 359 (2016). <https://doi.org/10.1186/s12859-016-1236-x>
- [26] S. Raychaudhuri, "Introduction to Monte Carlo simulation," *2008 Winter Simulation Conference*, Miami, FL, USA, 2008, pp. 91-100, doi: 10.1109/WSC.2008.4736059.
- [27] Koning, RH., Koolhaas, M., Renes, G., & Ridder, G. (2003). A simulation model for football championships. *European Journal of Operational Research*, 148(2), 268-276. [https://doi.org/10.1016/S0377-2217\(02\)00683-5](https://doi.org/10.1016/S0377-2217(02)00683-5)
- [28] Python Software Foundation.codecs — Codec registry and base classes. Python Documentation. Available: <https://docs.python.org/3/library/codecs.html>
- [29] Python Software Foundation.re — Regular expression operations.Python Documentation. Available: <https://docs.python.org/3/library/re.html>
- [30] Montesinos-Lopez, O. A., Montesinos-Lopez, J. C., Salazar, E., Barron, J. A., Montesinos-Lopez, A., Buenrostro-Mariscal, R., & Crossa, J. Application of a Poisson deep neural network model for the prediction of count data in genome-based prediction. *Plant Genome*. 2021; 14:e20118. <https://doi.org/10.1002/tpg2.20118>
- [31] Leszek Szczecinski, Aymen Djebbi (2019).Understanding and Pushing the Limits of the Elo Rating Algorithm. arXiv:1910.06081.
- [32] Zhong, Y. Enhancing game outcome prediction in the Chinese basketball league through a machine learning framework based on performance data. *Sci Rep* 15, 23788 (2025). <https://doi.org/10.1038/s41598-025-08882-7>
- [33] Arntzen H, Hvattum LM. Predicting match outcomes in association football using team ratings and player ratings. *Statistical Modelling*. 2021;21(5):449-470. doi:10.1177/1471082X20929881

