



ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΕΥΦΥΗΣ ΑΝΑΛΥΣΗ ΑΘΛΗΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΛΗΨΗ ΑΠΟΦΑΣΕΩΝ

ΑΓΓΕΛΗΣ ΧΡΗΣΤΟΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΥΠΕΥΘΥΝΟΣ

Κολομβάτσος Κωνσταντίνος
Μέλος ΔΕΠ

Λαμία, έτος 2026



ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΕΥΦΥΗΣ ΑΝΑΛΥΣΗ ΑΘΛΗΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΛΗΨΗ ΑΠΟΦΑΣΕΩΝ

ΑΓΓΕΛΗΣ ΧΡΗΣΤΟΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΥΠΕΥΘΥΝΟΣ

Κολομβάτσος Κωνσταντίνος
Μέλος ΔΕΠ

Λαμία, έτος 2026



UNIVERSITY OF
THESSALY

SCHOOL OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE & TELECOMMUNICATIONS

Advanced Sports Analytics for Decision Making

ANGELIS CHRISTOS

FINAL THESIS

ADVISOR

Kolomvatsos Konstantinos
Faculty Member

Lamia, 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 παρουσίασή 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:/...../20.....

Ο – Η Δηλ.

(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 ραγδαία εξέλιξη της ανάλυσης δεδομένων (data analytics) έχει μετασχηματίσει τον χώρο του αθλητισμού, προσφέροντας νέα εργαλεία στρατηγικού σχεδιασμού. Ωστόσο, η πρόβλεψη αποτελεσμάτων στο ποδόσφαιρο παραμένει μια σημαντική πρόκληση λόγω της στοχαστικής φύσης του αθλήματος και του χαμηλού δείκτη σκοραρίσματος. Η παρούσα πτυχιακή εργασία εστιάζει στη σχεδίαση, υλοποίηση και συγκριτική αξιολόγηση παραδοσιακών στατιστικών μεθόδων (κατανομή Poisson, μοντέλο Dixon-Coles) έναντι προηγμένων αρχιτεκτονικών Βαθιάς Μάθησης (Long Short-Term Memory - LSTM, Gated Recurrent Unit - GRU) για την πρόβλεψη ποδοσφαιρικών αναμετρήσεων.

Για την εξαγωγή αξιόπιστων συμπερασμάτων, αναπτύχθηκε ένα ολοκληρωμένο λογισμικό αρχιτεκτονικής πολλαπλών επιπέδων (N-Tier), το οποίο αντλεί δεδομένα μέσω REST API, πραγματοποιεί μηχανική χαρακτηριστικών (feature engineering) και οπτικοποιεί τις προβλέψεις μέσω ειδικά σχεδιασμένης διεπαφής χρήστη (GUI) με τη χρήση της βιβλιοθήκης PySide6. Η πειραματική αξιολόγηση των μοντέλων διεξήχθη μέσω της αυστηρής μεθοδολογίας Walk-Forward Validation, η οποία προσομοιώνει πραγματικές συνθήκες εβδομαδιαίας επανεκπαίδευσης (fine-tuning).

Η αξιολόγηση βασίστηκε όχι μόνο στην Ακρίβεια (Accuracy) των ταξινομητών, αλλά και στις μετρικές Brier Score και Log-Loss, με σκοπό την ποσοτική αποτίμηση της βαθμονόμησης των παραγόμενων πιθανοτήτων. Τα πειραματικά αποτελέσματα κατέδειξαν ότι τα κλασικά στατιστικά μοντέλα, με προεξάρχον το μοντέλο Poisson, υπερτερούν σημαντικά των Νευρωνικών Δικτύων, επιτυγχάνοντας ακρίβεια της τάξης του 50.5% και βέλτιστη βαθμονόμηση πιθανοτήτων. Αντίθετα, οι αρχιτεκτονικές LSTM και GRU παρουσίασαν αδυναμία γενίκευσης και φαινόμενα υπερβολικής σιγουριάς (overconfidence). Αυτό αποδίδεται στο γεγονός ότι η εισαγωγή πολλαπλών αγωνιστικών χαρακτηριστικών (όπως σουτ και κόρνερ) λειτούργησε συχνά ως «θόρυβος» (noise) παρά ως χρήσιμο πληροφοριακό «σήμα» (signal). Συμπερασματικά, η εργασία αποδεικνύει ότι σε περιβάλλοντα υψηλής αβεβαιότητας, οι απλές μαθηματικές δομές που βασίζονται σε μέσους όρους παραμένουν τα πιο ισχυρά και αξιόπιστα εργαλεία λήψης αποφάσεων έναντι πολύπλοκων μοντέλων Τεχνητής Νοημοσύνης.

Λέξεις-κλειδιά: Sports Analytics, Μηχανική Μάθηση, Στατιστική Μοντελοποίηση, Κατανομή Poisson, LSTM, Walk-Forward Validation, Brier Score.

ABSTRACT

The rapid evolution of data analytics has fundamentally transformed the sports industry, providing novel tools for strategic planning. However, predicting football match outcomes remains a significant challenge due to the stochastic and low-scoring nature of the sport. This thesis focuses on the design, implementation, and comparative evaluation of traditional statistical methods (Poisson distribution, Dixon-Coles model) against advanced Deep Learning architectures (Long Short-Term Memory - LSTM, Gated Recurrent Unit - GRU) for forecasting football match results.

To derive robust and highly reliable conclusions, an end-to-end N-Tier architecture software system was developed. This system automatically ingests match data via a REST API, performs extensive feature engineering, and visually presents real-time predictions through a custom Graphical User Interface (GUI) built with PySide6. The experimental evaluation of the models was conducted using a rigorous Walk-Forward Validation methodology, simulating real-world weekly retraining (fine-tuning) scenarios. The performance assessment transcended basic Accuracy mapping, utilizing Brier Score and Log-Loss metrics to quantitatively evaluate the probability calibration of each model.

Experimental results demonstrated that classical statistical approaches, primarily the baseline Poisson model, significantly outperformed the Deep Learning networks, achieving an accuracy of approximately 50.5% alongside optimal probability calibration. Conversely, the LSTM and GRU architectures exhibited signs of overfitting and probabilistic overconfidence. This is attributed to the fact that integrating multiple advanced in-game features (e.g., shots, corners) frequently introduced high levels of statistical "noise" rather than actionable "signal". In conclusion, the study establishes that in environments characterized by high uncertainty, simplistic mathematical structures based on moving averages remain more robust and reliable decision-making tools compared to complex Artificial Intelligence models.

Keywords: Sports Analytics, Machine Learning, Statistical Modeling, Poisson Distribution, LSTM, Walk-Forward Validation, Brier Score.

Table of Contents

ABSTRACT	10
1 ΕΙΣΑΓΩΓΗ	14
(1.1) ΑΝΤΙΚΕΙΜΕΝΟ ΚΑΙ ΣΚΟΠΟΣ ΤΗΣ ΕΡΓΑΣΙΑΣ	14
(1.2) ΚΙΝΗΤΡΟ ΤΗΣ ΕΡΓΑΣΙΑΣ.....	14
(1.3) ΔΟΜΗ ΤΗΣ ΕΡΓΑΣΙΑΣ	15
2 ΣΧΕΤΙΚΕΣ ΕΡΕΥΝΕΣ	16
(2.1) ΠΡΟΣΕΓΓΙΣΕΙΣ ΜΕ ΚΛΑΣΙΚΑ ΣΤΑΤΙΣΤΙΚΑ ΜΟΝΤΕΛΑ	16
(2.2) ΕΦΑΡΜΟΓΕΣ ΠΑΡΑΔΟΣΙΑΚΗΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ	16
(2.3) ΣΥΓΧΡΟΝΕΣ ΕΦΑΡΜΟΓΕΣ ΒΑΘΙΑΣ ΜΑΘΗΣΗΣ (DEEP LEARNING).....	17
(2.4) ΕΞΕΛΙΓΜΕΝΑ ΣΤΑΤΙΣΤΙΚΑ ΜΟΝΤΕΛΑ ΚΑΙ ΜΕΘΟΔΟΙ ΠΑΛΙΝΔΡΟΜΗΣΗΣ.....	17
(2.5) ΜΠΕΨΖΙΑΝΑ ΔΙΚΤΥΑ ΚΑΙ ΥΒΡΙΔΙΚΕΣ ΠΡΟΣΕΓΓΙΣΕΙΣ	18
(2.6) ΜΗΧΑΝΙΚΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΚΑΙ ΤΟ ΠΡΟΒΛΗΜΑ ΤΟΥ ΘΟΡΥΒΟΥ.....	18
(2.7) ΣΥΝΟΛΙΚΗ ΑΞΙΟΛΟΓΗΣΗ ΚΑΙ ΕΡΕΥΝΗΤΙΚΟ ΚΕΝΟ.....	18
3 ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ & ΜΑΘΗΜΑΤΙΚΗ ΘΕΜΕΛΙΩΣΗ.....	20
(3.1) SPORTS ANALYTICS ΚΑΙ ΠΡΟΒΛΕΨΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ.....	20
(3.2) ΣΤΑΤΙΣΤΙΚΗ ΜΟΝΤΕΛΟΠΟΙΗΣΗ	20
(3.2.1) Η ΚΑΤΑΝΟΜΗ POISSON ΚΑΙ Η ΕΦΑΡΜΟΓΗ ΤΗΣ ΣΤΑ ΓΚΟΛ.....	20
(3.2.2) ΑΝΕΞΑΡΤΗΣΙΑ ΤΩΝ ΣΚΟΡ ΚΑΙ ΠΕΡΙΟΡΙΣΜΟΙ.....	21
(3.3) ΤΟ ΜΟΝΤΕΛΟ DIXON-COLES:	21
(3.3.1) Η ΣΥΝΑΡΤΗΣΗ ΔΙΟΡΘΩΣΗΣ T ΓΙΑ ΤΗΝ ΑΛΛΗΛΟΕΞΑΡΤΗΣΗ ΤΩΝ ΣΚΟΡ	22
(3.3.2) ΠΑΡΑΜΕΤΡΟΙ ΕΠΙΘΕΣΗΣ (A) ΚΑΙ ΑΜΥΝΑΣ (B)	22
(3.3.3) ΧΡΟΝΙΚΗ ΕΞΑΣΘΕΝΙΣΗ α_i (TIME DECAY) ΜΕ ΧΡΗΣΗ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ Φ	23
(3.4) ΤΕΧΝΗΤΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ (RNN).....	24
(3.4.1) Η ΕΝΝΟΙΑ ΤΗΣ ΑΚΟΛΟΥΘΙΑΣ (SEQUENCE) ΚΑΙ ΤΗΣ ΜΝΗΜΗΣ.....	24
(3.4.2) LSTM (LONG SHORT-TERM MEMORY): ΑΝΑΛΥΣΗ ΤΩΝ ΠΥΛΩΝ ΚΑΙ ΤΟΥ CELL STATE	24
4 ΣΧΕΔΙΑΣΜΟΣ & ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΣΥΣΤΗΜΑΤΟΣ	26
(4.1) ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΠΟΛΛΑΠΛΩΝ ΕΠΙΠΕΔΩΝ (N-TIER ARCHITECTURE).....	26
(4.2) ΤΟ ΥΒΡΙΔΙΚΟ PIPELINE ΔΕΔΟΜΕΝΩΝ:.....	26
(4.2.1) ΑΡΧΙΚΟΠΟΙΗΣΗ ΜΕΣΩ ΙΣΤΟΡΙΚΟΥ ΑΡΧΕΙΟΥ ΚΑΙ ΔΙΑΣΥΝΔΕΣΗ ΜΕ REST API	27
(4.2.2) ΔΙΑΔΙΚΑΣΙΑ ΑΥΤΟΜΑΤΟΠΟΙΗΜΕΝΗΣ ΕΝΗΜΕΡΩΣΗΣ	28
(4.3) ΣΧΕΔΙΑΣΜΟΣ ΤΗΣ ΔΙΕΠΑΦΗΣ ΧΡΗΣΤΗ (GUI)	28
5 ΣΥΛΛΟΓΗ & ΠΡΟΕΤΟΙΜΑΣΙΑ ΔΕΔΟΜΕΝΩΝ	30
(5.1) ΠΗΓΕΣ ΔΕΔΟΜΕΝΩΝ: ΑΝΑΛΥΣΗ ΤΟΥ DATASET	30
(5.2) FEATURE ENGINEERING:	30
(5.2.1) ΥΠΟΛΟΓΙΣΜΟΣ ROLLING STATS.....	30

(5.2.2) ΔΗΜΙΟΥΡΓΙΑ ΧΡΟΝΙΚΩΝ ΠΑΡΑΘΥΡΩΝ ΓΙΑ ΤΑ RNNs	31
(5.3) ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΚΑΙ ΚΑΝΟΝΙΚΟΠΟΙΗΣΗ	31
6 ΥΛΟΠΟΙΗΣΗ ΤΩΝ ΜΟΝΤΕΛΩΝ ΠΡΟΒΛΕΨΗΣ.....	32
(6.1) ΚΛΑΣΙΚΗ ΣΤΑΤΙΣΤΙΚΗ ΜΟΝΤΕΛΟΠΟΙΗΣΗ.....	32
(6.1.1) Η ΚΑΤΑΝΟΜΗ POISSON	32
(6.1.2) ΤΟ ΜΟΝΤΕΛΟ DIXON-COLES.....	32
(6.2) ΥΛΟΠΟΙΗΣΗ LSTM & GRU:	33
(6.2.1) ΔΙΚΤΥΑ ΜΑΚΡΑΣ ΒΡΑΧΥΠΡΟΘΕΣΜΗΣ ΜΝΗΜΗΣ (LSTM).....	33
(6.2.2) ΔΙΚΤΥΑ ΕΛΕΓΧΟΜΕΝΗΣ ΕΠΑΝΑΛΑΜΒΑΝΟΜΕΝΗΣ ΜΟΝΑΔΑΣ (GRU)	33
(6.2.3) ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΕΠΙΠΕΔΩΝ, ΚΑΝΟΝΙΚΟΠΟΙΗΣΗ ΚΑΙ ΒΕΛΤΙΩΣΤΟΠΟΙΗΣΗ	33
7 ΠΕΙΡΑΜΑΤΙΚΗ ΔΙΑΔΙΚΑΣΙΑ & ΑΞΙΟΛΟΓΗΣΗ	35
(7.1) ΜΕΘΟΔΟΛΟΓΙΑ ΕΛΕΓΧΟΥ: WALK-FORWARD VALIDATION.....	35
(7.2) ΜΕΤΡΙΚΕΣ ΑΞΙΟΛΟΓΗΣΗΣ	35
(7.2.1) ΑΚΡΙΒΕΙΑ ΤΑΞΙΝΟΜΗΣΗΣ (ACCURACY)	35
(7.2.2) BRIER SCORE: ΠΟΣΟΤΙΚΗ ΜΕΤΡΗΣΗ ΤΗΣ ΠΙΘΑΝΟΛΟΓΙΚΗΣ ΕΥΣΤΟΧΙΑΣ.....	35
(7.2.3) ΛΟΓΑΡΙΘΜΙΚΗ ΑΠΩΛΕΙΑ (LOG-LOSS / CROSS-ENTROPY)	36
(7.3) ΠΑΡΟΥΣΙΑΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ.....	36
(7.3.1) ΧΡΟΝΙΚΗ ΕΞΕΛΙΞΗ ΤΗΣ ΑΚΡΙΒΕΙΑΣ ΚΑΙ ΤΟ ΦΑΙΝΟΜΕΝΟ ΤΟΥ CONCEPT DRIFT ..	37
(7.4) ΔΙΑΓΡΑΜΜΑΤΑ ΒΑΘΜΟΝΟΜΗΣΗΣ (RELIABILITY DIAGRAMS)	37
8 ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΕΠΕΚΤΑΣΕΙΣ	40
(8.1) ΣΥΝΟΨΗ ΕΥΡΗΜΑΤΩΝ: ΠΟΙΟ ΜΟΝΤΕΛΟ ΥΠΕΡΤΕΡΕΙ	40
(8.2) ΠΕΡΙΟΡΙΣΜΟΙ ΤΗΣ ΕΡΕΥΝΑΣ	40
(8.3) ΠΡΟΤΑΣΕΙΣ ΓΙΑ ΜΕΛΛΟΝΤΙΚΗ ΕΡΕΥΝΑ	41
(9) ΠΑΡΑΡΤΗΜΑ Α	42
A.1 ΤΕΧΝΙΚΟΙ ΌΡΟΙ ΚΑΙ ΟΡΙΣΜΟΙ	42
ΒΙΒΛΙΟΓΡΑΦΙΑ.....	44

1 Εισαγωγή

(1.1) Αντικείμενο και Σκοπός της εργασίας

Η πρόβλεψη αθλητικών γεγονότων, και ειδικότερα των ποδοσφαιρικών αναμετρήσεων, αποτελεί ένα από τα πλέον σύνθετα και ελκυστικά προβλήματα στον τομέα της Ανάλυσης Δεδομένων (Data Analytics) και της Μηχανικής Μάθησης (Machine Learning). Το ποδόσφαιρο, ως άθλημα, χαρακτηρίζεται από εξαιρετικά χαμηλό ρυθμό σκοραρίσματος (low-scoring game) και υψηλό βαθμό στοχαστικότητας. Σε αντίθεση με αθλήματα όπως η καλαθοσφαίριση, όπου η συχνότητα των πόντων μειώνει τη στατιστική διακύμανση, στο ποδόσφαιρο ένα μεμονωμένο, τυχαίο γεγονός (π.χ. ένα λάθος της άμυνας ή μια αποβολή) αρκεί για να ανατρέψει πλήρως την έκβαση ενός αγώνα, καθιστώντας την ακριβή πρόβλεψη του τελικού αποτελέσματος μια πρόκληση υψίστης δυσκολίας. Σκοπός της παρούσας πτυχιακής εργασίας είναι η σχεδίαση, υλοποίηση και συγκριτική αξιολόγηση ενός ολοκληρωμένου συστήματος πρόβλεψης ποδοσφαιρικών αγώνων. Ειδικότερα, η εργασία επιδιώκει να αντιπαραβάλει δύο εκ διαμέτρου αντίθετες φιλοσοφίες μοντελοποίησης: την παραδοσιακή κλασική στατιστική (με αντιπροσώπους την Κατανομή Poisson και το μοντέλο Dixon-Coles) και τη σύγχρονη τεχνητή νοημοσύνη, μέσω Επαναλαμβανόμενων Νευρωνικών Δικτύων Βαθιάς Μάθησης (LSTM και GRU). Ο στόχος είναι να διαπιστωθεί εάν η αλγοριθμική πολυπλοκότητα των σύγχρονων νευρωνικών αρχιτεκτονικών προσφέρει πραγματικό πλεονέκτημα σε ένα περιβάλλον τόσο ακραίου θορύβου, ή εάν τα απλούστερα πιθανοτικά μοντέλα παραμένουν πιο αξιόπιστα.

(1.2) Κίνητρο της Εργασίας

Το βασικό κίνητρο για την εκπόνηση της συγκεκριμένης εργασίας πηγάζει από τη ραγδαία εξέλιξη των Sports Analytics. Τα τελευταία χρόνια, παρατηρείται μια τάση στον ακαδημαϊκό και επαγγελματικό αθλητισμό για την υιοθέτηση όλο και πιο πολύπλοκων συστημάτων τεχνητής νοημοσύνης. Ωστόσο, στην επιστήμη της Πληροφορικής κυριαρχεί το φαινόμενο της Υπερπροσαρμογής (Overfitting), κατά το οποίο τα πολύπλοκα μοντέλα τείνουν να "απομνημονεύουν" τον θόρυβο των δεδομένων αντί να γενικεύουν ουσιαστικά μοτίβα. Η έρευνα αυτή φιλοδοξεί να εξετάσει ακαδημαϊκά την εφαρμογή του «Ξυραφιού του Όκαμ» (Occam's Razor) στον χώρο των αθλητικών δεδομένων, διερευνώντας τα όρια των Νευρωνικών Δικτύων απέναντι στην απλότητα των μέσων όρων. Αξίζει να σημειωθεί ότι η προσέγγιση της παρούσας μελέτης είναι αυστηρά επιστημονική και μαθηματική, εστιάζοντας στην προγνωστική ακρίβεια (Accuracy) και στα σφάλματα πιθανοτήτων (Brier Score), αποσυνδεδεμένη πλήρως από έννοιες στοιχηματισμού ή οικονομικής κερδοφορίας.

(1.3) Δομή της Εργασίας

Η παρούσα πτυχιακή εργασία οργανώνεται σε επτά (8) διακριτά κεφάλαια και ένα παράρτημα, τα οποία καλύπτουν τόσο το θεωρητικό υπόβαθρο όσο και την πρακτική υλοποίηση του συστήματος, ως εξής:

- **Κεφάλαιο 1 (Εισαγωγή):** Παρουσιάζεται το αντικείμενο, ο σκοπός και το κίνητρο της έρευνας.
- **Κεφάλαιο 2 (Σχετικές Έρευνες):** Πραγματοποιείται βιβλιογραφική ανασκόπηση σε προηγούμενες μελέτες στα Sports Analytics.
- **Κεφάλαιο 3 (Θεωρητικό Υπόβαθρο):** Αναλύεται η μαθηματική θεωρία πίσω από τα στατιστικά μοντέλα και τα Επαναλαμβανόμενα Νευρωνικά Δίκτυα (RNNs).
- **Κεφάλαιο 4 (Αρχιτεκτονική Συστήματος):** Περιγράφεται η αρχιτεκτονική N-Tier, το pipeline δεδομένων μέσω API και ο σχεδιασμός του GUI (PySide6).
- **Κεφάλαιο 5 (Συλλογή και Προετοιμασία Δεδομένων):** Εξετάζεται το σύνολο δεδομένων και οι τεχνικές Feature Engineering (Rolling Stats).
- **Κεφάλαιο 6 (Υλοποίηση Μοντέλων Πρόβλεψης):** Αναλύεται υπολογιστικά η υλοποίηση των Poisson, Dixon-Coles, LSTM και GRU.
- **Κεφάλαιο 7 (Πειραματική Διαδικασία & Αξιολόγηση):** Αναλύεται η μεθοδολογία Walk-Forward Validation και παρουσιάζονται τα αποτελέσματα των μετρικών.
- **Κεφάλαιο 8 (Συμπεράσματα):** Συνοψίζονται τα ευρήματα της έρευνας και διατυπώνονται προτάσεις για μελλοντική μελέτη.
- **Παράρτημα Α:** Παρατίθεται γλωσσάρι με τους τεχνικούς όρους και ορισμούς για τη διευκόλυνση του αναγνώστη.

2 Σχετικές Έρευνες

Η εφαρμογή μαθηματικών μοντέλων και αλγορίθμων μηχανικής μάθησης για την πρόβλεψη αθλητικών αποτελεσμάτων αποτελεί ένα ταχέως αναπτυσσόμενο ερευνητικό πεδίο. Η υπάρχουσα βιβλιογραφία γύρω από τα Sports Analytics μπορεί να κατηγοριοποιηθεί σε τρεις βασικούς άξονες: τις κλασικές στατιστικές προσεγγίσεις, τις τεχνικές παραδοσιακής Μηχανικής Μάθησης (Machine Learning) και, πιο πρόσφατα, τις εφαρμογές Βαθιάς Μάθησης (Deep Learning). Στο παρόν κεφάλαιο αναλύονται οι σημαντικότερες μελέτες ανά κατηγορία.

(2.1) Προσεγγίσεις με κλασικά Στατιστικά Μοντέλα

Η προσπάθεια μαθηματικής μοντελοποίησης του ποδοσφαίρου ξεκίνησε πολλές δεκαετίες πριν από την έλευση των σύγχρονων υπολογιστικών συστημάτων. Η θεμελιώδης εργασία του Maher (1982) αποτέλεσε την πρώτη ολοκληρωμένη προσέγγιση για τη μοντελοποίηση των γκολ ως ανεξάρτητων γεγονότων Poisson. Ο Maher απέδειξε ότι η επιθετική και αμυντική δυναμικότητα των ομάδων, σε συνδυασμό με το πλεονέκτημα έδρας, μπορούν να παράγουν αξιόπιστες προσδόκιμες τιμές (expected goals) για τον τελικό αριθμό τερμάτων.

Ωστόσο, η αδυναμία του απλού μοντέλου Poisson να προβλέψει με ακρίβεια τις ισοπαλίες (καθώς θεωρούσε την επίθεση της μίας ομάδας παντελώς ανεξάρτητη από την άμυνα της άλλης), οδήγησε τους Dixon και Coles (1997) στη δημιουργία του πλέον γνωστού στατιστικού μοντέλου στον χώρο. Οι ερευνητές εισήγαγαν μια παράμετρο εξάρτησης (ρ) για τα χαμηλά σκορ (0-0, 1-0, 0-1, 1-1) και μια συνάρτηση χρονικής απόσβεσης (time-decay), καταφέροντας να προσαρμόσουν το μοντέλο στη δυναμική εξέλιξη της φόρμας των ομάδων κατά τη διάρκεια της σεζόν. Το μοντέλο Dixon-Coles παραμένει μέχρι σήμερα το σημείο αναφοράς (baseline) για κάθε νέα έρευνα στον τομέα.

(2.2) Εφαρμογές Παραδοσιακής Μηχανικής Μάθησης

Με την έκρηξη των Big Data στις αρχές της δεκαετίας του 2010, η ερευνητική κοινότητα στράφηκε στους αλγόριθμους Μηχανικής Μάθησης (Machine Learning) για την εξαγωγή μοτίβων από αθλητικά στατιστικά.

Χαρακτηριστική είναι η έρευνα των Hucaljuk και Rakipović (2011), οι οποίοι επιχείρησαν να προβλέψουν αποτελέσματα αγώνων ποδοσφαίρου χρησιμοποιώντας διάφορους αλγόριθμους ταξινόμησης, όπως Τυχαία Δάση (Random Forests), Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVM) και Νευρωνικά Δίκτυα απλής τροφοδότησης (Feed-Forward). Οι ερευνητές χρησιμοποίησαν ένα εκτενές σύνολο χαρακτηριστικών (feature set), όπως η κατοχή μπάλας, οι κίτρινες κάρτες και τα φάουλ. Τα αποτελέσματά τους έδειξαν ότι τα Random Forests παρουσίασαν την καλύτερη συμπεριφορά, ωστόσο η συνολική ακρίβεια σπάνια ξεπερνούσε το 52-53%, επιβεβαιώνοντας την ακραία αβεβαιότητα του αθλήματος.

Αντίστοιχα, ο Baboota και η Kaur (2019) επικεντρώθηκαν στη σημασία της Μηχανικής Χαρακτηριστικών (Feature Engineering) στην αγγλική Premier League. Αξιοποιώντας αλγόριθμους Gradient Boosting (όπως το XGBoost), απέδειξαν ότι η χρήση κυλιόμενων στατιστικών (rolling averages) των τελευταίων 3 έως 5 αγώνων είναι πολύ πιο προβλεπτική από τα γενικά στατιστικά ολόκληρης της σεζόν.

(2.3) Σύγχρονες Εφαρμογές Βαθιάς Μάθησης (Deep Learning)

Τα τελευταία χρόνια, η έρευνα έχει μετατοπιστεί στις αρχιτεκτονικές Βαθιάς Μάθησης για την ανάλυση χρονοσειρών. Δεδομένου ότι μια αγωνιστική σεζόν αποτελεί μια ακολουθία γεγονότων στον χρόνο, τα Επαναλαμβανόμενα Νευρωνικά Δίκτυα (RNN) και ειδικότερα τα δίκτυα LSTM (Long Short-Term Memory) βρέθηκαν στο επίκεντρο. Στη μελέτη των McCabe και Trevathan (2008), εξετάστηκε η χρήση πολυστρωματικών νευρωνικών δικτύων (Multilayer Perceptrons) σε συνδυασμό με RNNs για την πρόβλεψη αθλημάτων (ποδόσφαιρο, ράγκμπι). Διαπίστωσαν ότι τα δίκτυα συχνά υπέφεραν από φαινόμενα υπερπροσαρμογής (overfitting), καθώς "απομνημόνευαν" τα στατιστικά των ομάδων (π.χ. σερί νικών) αντί να εντοπίζουν ουσιαστικούς κανόνες πρόβλεψης.

Πιο πρόσφατα, έρευνες που χρησιμοποίησαν LSTMs επιχείρησαν να ενσωματώσουν όχι μόνο ιστορικά σκορ, αλλά και σύνθετα στατιστικά (όπως Expected Goals - xG). Παρότι τα LSTMs εμφανίζουν θεωρητική ανωτερότητα στη διατήρηση μακροπρόθεσμης μνήμης, πολλές μελέτες αναφέρουν ότι η πολυπλοκότητά τους συχνά εισάγει θόρυβο, με αποτέλεσμα να μην μπορούν πάντα να ξεπεράσουν στατιστικά τις απλές πιθανοτικές κατανομές (όπως την Poisson).

(2.4) Εξελεγμένα Στατιστικά Μοντέλα και Μέθοδοι Παλινδρόμησης

Η εξέλιξη της στατιστικής δεν σταμάτησε στο βασικό μοντέλο Dixon-Coles. Ο Goddard (2005) διερεύνησε την εφαρμογή διατεταγμένων μοντέλων probit (ordered probit regression models) για την πρόβλεψη της έκβασης αγώνων στο αγγλικό πρωτάθλημα. Η μελέτη του ανέδειξε τη σημασία παραγόντων που υπερβαίνουν τα απλά γκολ, αποδεικνύοντας ότι η γεωγραφική απόσταση μεταξύ των ομάδων (τοπικά ντέρμπι) και η συμμετοχή σε παράλληλες διοργανώσεις (εγχώρια κύπελλα) επηρεάζουν στατιστικά το αποτέλεσμα λόγω συσσωρευμένης κόπωσης. Πιο πρόσφατα, οι Boshnakov, Kharrat και McHale (2017) πρότειναν ένα δισδιάστατο μοντέλο κατανομής Weibull (bivariate Weibull count model). Το μοντέλο αυτό αντιμετωπίζει πιο αποτελεσματικά τον δυναμικό ρυθμό σκοραρίσματος (hazard rate) κατά τη διάρκεια των 90 λεπτών σε σχέση με την αυστηρή κατανομή Poisson, αν και απαιτεί σημαντικά υψηλότερο υπολογιστικό κόστος.

(2.5) Μπεϋζιανά Δίκτυα και Υβριδικές Προσεγγίσεις

Πέραν των παραδοσιακών εξισώσεων, μεγάλη έμφαση έχει δοθεί στα Μπεϋζιανά Δίκτυα (Bayesian Networks), τα οποία επιτρέπουν την ενσωμάτωση θεωρητικής και υποκειμενικής γνώσης στο σύστημα. Οι Constantinou, Fenton και Neil (2012) ανέπτυξαν το μοντέλο *pi-football*, το οποίο γεφυρώνει το χάσμα μεταξύ καθαρής στατιστικής και ανθρώπινης εμπειρίας. Στο σύστημά τους, εκτός από τα ιστορικά δεδομένα, εισήχθησαν μεταβλητές όπως η αλλαγή προπονητή, το βαθμολογικό κίνητρο και η ψυχολογία. Παρότι η ποσοτικοποίηση τέτοιων χαρακτηριστικών ενέχει τεράστιο αλγοριθμικό θόρυβο, το *pi-football* επέδειξε εξαιρετική σταθερότητα. Στην ίδια λογική, οι Tax και Joulstra (2015) χρησιμοποίησαν δημόσια διαθέσιμα δεδομένα από το ολλανδικό πρωτάθλημα, εφαρμόζοντας υβριδικά μοντέλα Naïve Bayes και Multilayer Perceptrons (MLP), καταλήγοντας στο συμπέρασμα ότι η προσθήκη εξωαγωνιστικών δεικτών βελτιώνει ραγδαία την προγνωστική ικανότητα.

(2.6) Μηχανική Χαρακτηριστικών και το Πρόβλημα του Θορύβου

Στον χώρο των Sports Analytics, η ποιότητα των δεδομένων εισόδου (feature engineering) αποτελεί συχνά σημαντικότερο παράγοντα από την επιλογή του ίδιου του αλγορίθμου. Ωστόσο, η πιο καθοριστική μελέτη για τα πραγματικά όρια της τεχνητής νοημοσύνης στο άθλημα ήρθε από τους Berrar et al. (2019), μέσα από την εκτενή ανάλυση του "Soccer Prediction Challenge". Οι ερευνητές συγκέντρωσαν και συνέκριναν εκατοντάδες μοντέλα μηχανικής μάθησης από ακαδημαϊκές ομάδες παγκοσμίως. Το αναπάντεχο συμπέρασμα ήταν ότι οι πολύπλοκες μέθοδοι Βαθιάς Μάθησης (Deep Learning) υστέρησαν δραματικά σε σχέση με τις απλές μεθόδους, όπως τα Random Forests και οι γραμμικές παλινδρομήσεις. Απέδειξαν περίτρανα ότι το ποδόσφαιρο έχει τόσο χαμηλό λόγο "σήματος-προς-θόρυβο" (Signal-to-Noise ratio) που τα πολύπλοκα νευρωνικά δίκτυα υπερπροσαρμόζονται (overfitting) άμεσα στις τυχαίες διακυμάνσεις, ένα κρίσιμο εύρημα που επιβεβαιώνεται απόλυτα και από την πειραματική διαδικασία της παρούσας πτυχιακής εργασίας.

(2.7) Συνολική Αξιολόγηση και Ερευνητικό Κενό

Από την ανάλυση της παραπάνω βιβλιογραφίας καθίσταται εμφανές ότι, αν και έχει υπάρξει σημαντική πρόοδος στον τομέα των Sports Analytics, υπάρχει ένα σαφές ερευνητικό κενό όσον αφορά την άμεση και δίκαιη συγκριτική αξιολόγηση παραδοσιακών μοντέλων (Poisson, Dixon-Coles) με σύγχρονα μοντέλα Βαθιάς Μάθησης (LSTM, GRU) πάνω στο ίδιο ακριβώς σετ δεδομένων και υπό ρεαλιστικές συνθήκες.

Πολλές προηγούμενες έρευνες αξιολογούν τα μοντέλα τους χρησιμοποιώντας απλή διασταυρούμενη επικύρωση (K-Fold Cross Validation), η οποία προκαλεί διαρροή μελλοντικής πληροφορίας (data leakage) στα αθλητικά δεδομένα. Η παρούσα εργασία έρχεται να καλύψει αυτό το κενό, αναπτύσσοντας ένα ολοκληρωμένο, αυτοματοποιημένο σύστημα το οποίο ενσωματώνει αυστηρά τη μεθοδολογία Walk-Forward Validation, αξιολογώντας παράλληλα τα μοντέλα όχι μόνο ως προς την ακρίβεια

(Accuracy) αλλά και ως προς την ποιότητα των παραγόμενων πιθανοτήτων (μέσω της μετρικής Brier Score).

3 Θεωρητικό Υπόβαθρο & Μαθηματική Θεμελίωση

Το παρόν κεφάλαιο παραθέτει το θεωρητικό υπόβαθρο πάνω στο οποίο δομήθηκε η πειραματική διαδικασία της εργασίας. Αρχικά, γίνεται μια σύντομη αναφορά στον τομέα των Sports Analytics, ενώ στη συνέχεια αναλύονται εις βάθος τα μαθηματικά θεμέλια των κλασικών στατιστικών μοντέλων (Poisson, Dixon-Coles) και οι σύγχρονες αρχιτεκτονικές των Επαναλαμβανόμενων Νευρωνικών Δικτύων (RNNs).

(3.1) Sports Analytics και Πρόβλεψη Αποτελεσμάτων

Ο όρος Sports Analytics αναφέρεται στην εφαρμογή μαθηματικών, στατιστικών τεχνικών και αλγορίθμων υπολογιστικής νοημοσύνης σε δεδομένα που προέρχονται από τον αθλητισμό. Ειδικότερα στο ποδόσφαιρο, η ανάλυση δεδομένων παρουσιάζει μοναδικές προκλήσεις, κυρίως λόγω της φύσης του αθλήματος ως παιχνιδιού χαμηλού σκοραρίσματος (low-scoring game). Σε αντίθεση με αθλήματα υψηλής συχνότητας πόντων, όπως η καλαθοσφαίριση, το ποδόσφαιρο διέπεται από υψηλό βαθμό τυχαιότητας (stochasticity).

Ιστορικά, η προσπάθεια πρόβλεψης ποδοσφαιρικών αγώνων ξεκίνησε με απλές προσεγγίσεις κατανομής πιθανοτήτων, όπως η μελέτη του Moroney (1956), ο οποίος πρώτος υπέδειξε ότι η κατανομή Poisson μπορεί να περιγράψει ικανοποιητικά τα γκολ σε έναν ποδοσφαιρικό αγώνα. Στη σύγχρονη εποχή, η εμφάνιση των Big Data έχει επιτρέψει την εισαγωγή μοντέλων Μηχανικής Μάθησης (Machine Learning), αλλάζοντας ριζικά το τοπίο της αθλητικής ανάλυσης.

(3.2) Στατιστική Μοντελοποίηση

Η στατιστική μοντελοποίηση στον αθλητισμό, και ειδικότερα στο ποδόσφαιρο, βασίζεται στην προσπάθεια ποσοτικοποίησης της αβεβαιότητας και στον υπολογισμό των πιθανοτήτων για κάθε δυνατή έκβαση ενός αγώνα. Σε αντίθεση με τις απλές εμπειρικές εκτιμήσεις, τα μαθηματικά μοντέλα χρησιμοποιούν ιστορικά δεδομένα για να εξάγουν παραμέτρους δυναμικότητας, επιτρέποντας την προσομοίωση μελλοντικών αποτελεσμάτων με αυστηρούς επιστημονικούς όρους.

(3.2.1) Η κατανομή Poisson και η εφαρμογή της στα γκολ

Το θεμέλιο της κλασικής στατιστικής πρόβλεψης στο ποδόσφαιρο είναι η Κατανομή Poisson. Πρόκειται για μια διακριτή κατανομή πιθανότητας, η οποία περιγράφει την πιθανότητα να συμβεί ένας συγκεκριμένος αριθμός ανεξάρτητων γεγονότων σε ένα καθορισμένο χρονικό διάστημα, εφόσον τα γεγονότα αυτά συμβαίνουν με έναν γνωστό, σταθερό μέσο ρυθμό. Στο ποδόσφαιρο, το "χρονικό διάστημα" είναι τα 90 λεπτά της αναμέτρησης και τα "γεγονότα" είναι η επίτευξη τερμάτων. Τα γκολ είναι σπάνια περιστατικά, γεγονός που καθιστά την Poisson την ιδανική κατανομή. Η συνάρτηση μάζας πιθανότητας (Probability Mass Function) της κατανομής Poisson εκφράζεται μαθηματικά ως εξής:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad (3.1)$$

Στον παραπάνω τύπο:

- Το X αντιπροσωπεύει την τυχαία μεταβλητή, δηλαδή τον τελικό αριθμό των γκολ που θα επιτύχει μια ομάδα.
- Το x είναι ένας μη αρνητικός ακέραιος αριθμός (0, 1, 2, ...) που αντιστοιχεί στα πιθανά γκολ.
- Το e είναι η βάση του φυσικού λογαρίθμου (σταθερά Euler, $e \approx 2.71828$).
- Το $x!$ είναι το παραγοντικό του x .
- Το λ (λάμδα) είναι η προσδόκιμη τιμή (expected value) των γκολ. Εκφράζει τον μέσο όρο τερμάτων που αναμένεται να σκοράρει η ομάδα βάσει της δυναμικότητάς της.

Για παράδειγμα, εάν το μοντέλο υπολογίσει ότι η προσδόκιμη τιμή της γηπεδούχου ομάδας είναι $\lambda = 1.5$ γκολ, η πιθανότητα να σκοράρει ακριβώς 2 γκολ ($x = 2$) υπολογίζεται αντικαθιστώντας τις τιμές στον τύπο. Το μοντέλο Poisson εφαρμόζεται ανεξάρτητα και για τις δύο ομάδες, παράγοντας έναν πίνακα (πίνακας πιθανοτήτων) με όλα τα πιθανά τελικά σκορ (π.χ. 1-0, 2-1, 1-1) και τις αντίστοιχες πιθανότητες τους.

(3.2.2) Ανεξαρτησία των σκορ και περιορισμοί

Η βασική μαθηματική υπόθεση πάνω στην οποία στηρίζεται η απλή κατανομή Poisson είναι η έννοια της στατιστικής ανεξαρτησίας. Το μοντέλο θεωρεί αυθαίρετα ότι ο αριθμός των γκολ που θα πετύχει η Γηπεδούχος ομάδα (X) είναι παντελώς ανεξάρτητος από τον αριθμό των γκολ που θα δεχθεί από τη Φιλοξενούμενη (Y). Στη θεωρία πιθανοτήτων, αυτό σημαίνει ότι η από κοινού πιθανότητα ενός σκορ (π.χ. $X = 1$ και $Y = 1$) ισούται με το γινόμενο των επιμέρους (περιθωριακών) πιθανοτήτων:

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y) \quad (3.2)$$

Ωστόσο, στη δυναμική ενός πραγματικού ποδοσφαιρικού αγώνα, η υπόθεση της απόλυτης ανεξαρτησίας καταρρέει (Maher, 1982). Οι τακτικές των ομάδων μεταβάλλονται δυναμικά ανάλογα με το τρέχον σκορ. Για παράδειγμα, μια ομάδα που προηγείται με 1-0 στα τελευταία λεπτά τείνει να υιοθετεί αμυντική στρατηγική για να διατηρήσει το προβάδισμα, μειώνοντας δραματικά τον ρυθμό του αγώνα. Αντίστοιχα, στο 0-0 και οι δύο ομάδες ενδέχεται να παίξουν πιο συντηρητικά φοβούμενες το λάθος. Αυτή η εγγενής τακτική αλληλεξάρτηση αποτελεί τον σημαντικότερο περιορισμό του απλού μοντέλου Poisson. Στην πράξη, το απλό Poisson έχει αποδειχθεί ότι υποεκτιμά συστηματικά τις ισοπαλίες χαμηλού σκορ (όπως το 0-0 και το 1-1) και υπερεκτιμά τις πιθανότητες για άλλα, λιγότερο συχνά αποτελέσματα.

(3.3) Το Μοντέλο Dixon-Coles:

Προκειμένου να επιλύσουν τις αδυναμίες του απλού Poisson, οι Mark Dixon και Stuart Coles εισήγαγαν το 1997 ένα επαναστατικό μαθηματικό μοντέλο, το οποίο έκτοτε αποτελεί το "χρυσό πρότυπο" στην ακαδημαϊκή έρευνα της πρόβλεψης αθλητικών αποτελεσμάτων. Το μοντέλο Dixon-Coles εισάγει εξειδικευμένες παραμέτρους για να συλλάβει την εξάρτηση των γκολ και τη χρονική εξέλιξη της φόρμας.

(3.3.1) Η συνάρτηση διόρθωσης τ για την αλληλοεξάρτηση των σκορ

Για να αντιμετωπίσουν την υποεκτίμηση των ισοπαλιών, οι Dixon και Coles κατασκεύασαν μια συνάρτηση διόρθωσης (correction function), συμβολιζόμενη ως $\tau(x, y)$. Η συνάρτηση αυτή παρεμβαίνει και τροποποιεί τη συνάρτηση από κοινού πιθανότητας μόνο για τα χαμηλά σκορ, όπου παρατηρείται η μεγαλύτερη τακτική εξάρτηση, δηλαδή όταν $x, y \in \{0, 1\}$. Η τροποποιημένη από κοινού πιθανότητα εκφράζεται ως:

$$P(X = x, Y = y) = \tau(x, y) \cdot \frac{\lambda^x e^{-\lambda}}{x!} \cdot \frac{\mu^y e^{-\mu}}{y!} \quad (3.3)$$

Ο παράγοντας διόρθωσης τ εξαρτάται από μια νέα παράμετρο, τη λεγόμενη παράμετρο εξάρτησης ρ (ρ), η οποία υπολογίζεται μέσα από τα ιστορικά δεδομένα. Η συνάρτηση τ παίρνει τις εξής τιμές:

- Αν $x=0, y=0$: $\tau = 1 - \lambda\rho$ (Αυξάνει την πιθανότητα του 0-0).
- Αν $x=1, y=0$: $\tau = 1 + \lambda\rho$
- Αν $x=0, y=1$: $\tau = 1 + \mu\rho$
- Αν $x=1, y=1$: $\tau = 1 - \rho$ (Αυξάνει την πιθανότητα του 1-1).
- Για όλα τα υπόλοιπα σκορ (π.χ. $x \geq 2$ ή $y \geq 2$), $\tau = 1$, επαναφέροντας το μοντέλο στην απλή κατανομή Poisson.

Μέσω της συνάρτησης τ , το μοντέλο Dixon-Coles καταφέρνει να αποτυπώσει μαθηματικά τη «φοβική» συμπεριφορά των ομάδων στα ντέρμπι και στα κλειστά παιχνίδια, λύνοντας το πρόβλημα της ανεξαρτησίας.

(3.3.2) Παράμετροι επίθεσης (α) και άμυνας (β)

Στο μοντέλο Dixon-Coles, οι προσδόκιμες τιμές των γκολ για τη γηπεδούχο (λ) και τη φιλοξενούμενη ομάδα (μ) δεν είναι αυθαίρετες, αλλά συντίθενται από συγκεκριμένες κρυφές παραμέτρους δυναμικότητας. Για κάθε ομάδα i στο πρωτάθλημα, το σύστημα υπολογίζει μέσω Μεγιστοποίησης Πιθανοφάνειας (Maximum Likelihood Estimation):

1. Την Παράμετρο Επίθεσης (Attack Strength – α_i): Μετράει την ικανότητα της ομάδας να δημιουργεί και να σκοράρει γκολ σε σχέση με τον μέσο όρο του πρωταθλήματος.

1. Την Παράμετρο Άμυνας (Defense Strength – β_i): Μετράει την ικανότητα της ομάδας να αποτρέπει τον αντίπαλο από το να σκοράρει. (Μικρότερο β σημαίνει καλύτερη άμυνα).

Επιπρόσθετα, το μοντέλο εισάγει την παράμετρο γ (gamma), η οποία ποσοτικοποιεί το Πλεονέκτημα Έδρας (Home Advantage). Ο τελικός υπολογισμός των αναμενόμενων γκολ σε έναν αγώνα μεταξύ της ομάδας i (γηπεδούχος) και της ομάδας j (φιλοξενούμενη) δίνεται από τις εξισώσεις:

$$\lambda = \alpha_i \cdot \beta_j \cdot \gamma \quad (3.4)$$

$$\mu = \alpha_j \cdot \beta_i \quad (3.5)$$

Όπως φαίνεται καθαρά από τον τύπο, η επίθεση της γηπεδούχου (α_i) πολλαπλασιάζεται με την ικανότητα της αντίπαλης άμυνας (β_j) και το πλεονέκτημα έδρας (γ), προσφέροντας μια εξαιρετικά ρεαλιστική αποτύπωση της συσχέτισης δυνάμεων.

(3.3.3) Χρονική εξασθένιση α_i (Time Decay) με χρήση της παραμέτρου ϕ

Μια ομάδα που κατακτά το πρωτάθλημα τη μία σεζόν ενδέχεται να έχει πολύ κακή απόδοση την επόμενη λόγω αλλαγής ρόστερ ή προπονητή. Εάν το μοντέλο έδινε το ίδιο βάρος σε έναν αγώνα που έγινε πριν δύο χρόνια με έναν αγώνα που έγινε την προηγούμενη εβδομάδα, οι προβλέψεις θα ήταν λανθασμένες, καθώς δεν θα αντικατόπτριζαν την τρέχουσα φόρμα. Για την ενσωμάτωση της βραχυπρόθεσμης φόρμας, το μοντέλο Dixon-Coles εφαρμόζει μια εκθετική συνάρτηση χρονικής απόσβεσης (time-decay function). Κάθε ιστορικός αγώνας λαμβάνει ένα βάρος $w(t)$ ανάλογα με το πόσες ημέρες t έχουν περάσει από τη διεξαγωγή του. Ο τύπος της χρονικής εξασθένισης είναι:

$$w(t) = e^{-\phi t} \quad (3.6)$$

Η παράμετρος ϕ (phi) ελέγχει τον ρυθμό με τον οποίο η ιστορική πληροφορία «ξεχνιέται» (half-life of information). Αν το ϕ είναι μηδέν, όλοι οι αγώνες έχουν ίση βαρύτητα. Αν το ϕ είναι μεγάλο, μόνο τα πολύ πρόσφατα παιχνίδια επηρεάζουν τις παραμέτρους α και β . Αυτός ο μηχανισμός προσδίδει στο στατιστικό μοντέλο μια μορφή «μνήμης», επιτρέποντάς του να προσαρμόζεται διαρκώς.

(3.4) Τεχνητά Νευρωνικά Δίκτυα (RNN)

Ενώ τα μοντέλα Poisson και Dixon-Coles επιχειρούν να ορίσουν τον κόσμο μέσω αυστηρών στατιστικών κανόνων, η επανάσταση της Βαθιάς Μάθησης (Deep Learning) προσεγγίζει το πρόβλημα εντελώς διαφορετικά. Τα Τεχνητά Νευρωνικά Δίκτυα σχεδιάζονται με πρότυπο την αρχιτεκτονική του ανθρώπινου εγκεφάλου, έχοντας την ικανότητα να "μαθαίνουν" αυτόματα πολύπλοκα, μη γραμμικά μοτίβα μέσα από τεράστιους όγκους δεδομένων, χωρίς να απαιτούν προκαθορισμένους μαθηματικούς κανόνες για τα γκολ.

(3.4.1) Η έννοια της ακολουθίας (Sequence) και της μνήμης

Τα τυπικά Προσωφοδοτούμενα Νευρωνικά Δίκτυα (Feedforward Neural Networks) είναι στατικά. Επεξεργάζονται τα δεδομένα εισόδου απομονωμένα, χωρίς να λαμβάνουν υπόψη τη σειρά με την οποία συνέβησαν. Ωστόσο, τα δεδομένα του ποδοσφαίρου αποτελούν Χρονοσειρές (Time Series). Το αποτέλεσμα της 10ης αγωνιστικής επηρεάζεται άμεσα από την ψυχολογία και την κούραση που προκάλεσαν τα αποτελέσματα της 8ης και της 9ης αγωνιστικής.

Για την επεξεργασία τέτοιων δεδομένων, αναπτύχθηκαν τα Επαναλαμβανόμενα Νευρωνικά Δίκτυα (Recurrent Neural Networks - RNN). Το βασικό χαρακτηριστικό ενός RNN είναι η ύπαρξη εσωτερικής μνήμης. Αυτό επιτυγχάνεται μέσω ανατροφοδότησης (feedback loops) στους κρυφούς νευρώνες. Μαθηματικά, η κρυφή κατάσταση h_t τη χρονική στιγμή t δεν εξαρτάται μόνο από τη σημερινή είσοδο x_t , αλλά και από την προηγούμενη κρυφή κατάσταση h_{t-1} :

$$h_t = f(W_x x_t + W_h h_{t-1} + b) \quad (3.7)$$

Αν και θεωρητικά πανίσχυρα, τα απλά RNN αντιμετωπίζουν ένα σοβαρό περιορισμό κατά την εκπαίδευσή τους (σωστότερα, κατά την οπισθοδιάδοση του σφάλματος - Backpropagation Through Time). Λόγω των αλληλέλληλων πολλαπλασιασμών πινάκων βαρών, οι μαθηματικές κλίσεις (gradients) τείνουν να εκμηδενίζονται εκθετικά. Αυτό το φαινόμενο, γνωστό ως Πρόβλημα της Εξαφάνισης της Κλίσης (Vanishing Gradient Problem), καθιστά τα απλά RNN ανίκανα να "θυμούνται" γεγονότα που συνέβησαν πολύ μακριά στο παρελθόν.

(3.4.2) LSTM (Long Short-Term Memory): Ανάλυση των πυλών και του Cell State

Για την οριστική επίλυση του προβλήματος της μνήμης, οι Hochreiter και Schmidhuber (1997) ανέπτυξαν τα δίκτυα Μακράς Βραχυπρόθεσμης Μνήμης (LSTM). Το LSTM είναι μια εξαιρετικά πολύπλοκη αρχιτεκτονική RNN, σχεδιασμένη ειδικά για να διατηρεί πληροφορίες για μεγάλα χρονικά διαστήματα (long-term dependencies), όντας ιδανική για την παρακολούθηση της μακροπρόθεσμης φόρμας μιας ποδοσφαιρικής ομάδας. Ο πυρήνας της καινοτομίας του LSTM είναι η Κατάσταση Κελιού (Cell State - C_t). Το Cell State λειτουργεί σαν ένας «ιμάντας μεταφοράς» πληροφορίας (conveyor belt) που

διατρέχει ολόκληρη την αλυσίδα του δικτύου, επιτρέποντας στα δεδομένα να ρέουν αναλλοίωτα.

Η ικανότητα του LSTM να προσθέτει ή να αφαιρεί πληροφορίες από το Cell State ρυθμίζεται από τρεις μηχανισμούς, τις λεγόμενες Πύλες (Gates), οι οποίες χρησιμοποιούν συναρτήσεις ενεργοποίησης Σιγμοειδούς (σ) και Υπερβολικής Εφαπτομένης ($\tanh$):

1. Πύλη Λήθης (Forget Gate – f_t): Αποφασίζει ποια ποδοσφαιρικά στατιστικά είναι πλέον άσχετα (π.χ. μια νίκη που έγινε πριν 6 μήνες) και πρέπει να διαγραφούν. Εξάγει έναν αριθμό μεταξύ 0 (πλήρης διαγραφή) και 1 (πλήρης διατήρηση).

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (3.8)$$

2. Πύλη Εισόδου (Input Gate – i_t): Επιλέγει ποιες από τις νέες πληροφορίες της τρέχουσας αγωνιστικής είναι αρκετά σημαντικές ώστε να αποθηκευτούν στη μνήμη. Δημιουργεί παράλληλα ένα διάνυσμα νέων υποψήφιων τιμών ($\tilde{C}_t$).

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (3.9)$$

3. Ενημέρωση του Cell State (C_t): Η παλιά μνήμη πολλαπλασιάζεται με την απόφαση λήθης, και προστίθεται η νέα γνώση από την πύλη εισόδου:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (3.10)$$

4. Πύλη Εξόδου (Output Gate – o_t): Καθορίζει ποιο τμήμα της συνολικής μνήμης (C_t) θα χρησιμοποιηθεί ως τρέχουσα πρόβλεψη (h_t) για τον επερχόμενο αγώνα.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (3.11)$$

$$h_t = o_t * \tanh(C_t) \quad (3.12)$$

Μέσω αυτού του εκλεπτυσμένου μαθηματικού μηχανισμού, το μοντέλο LSTM έχει τη θεωρητική ικανότητα να σταθμίζει δυναμικά τη σημασία κάθε ποδοσφαιρικού στατιστικού μέσα στον χρόνο, χωρίς να περιορίζεται από τις στατικές υποθέσεις της κλασικής πιθανοθεωρίας.

4 Σχεδιασμός & Αρχιτεκτονική Συστήματος

(4.1) Αρχιτεκτονική Πολλαπλών Επιπέδων (N-Tier Architecture)

Για την εξασφάλιση της επεκτασιμότητας (**scalability**), της συντηρησιμότητας (**maintainability**) και του διαχωρισμού των αρμοδιοτήτων (**separation of concerns**) στον κώδικα, το σύστημα σχεδιάστηκε με βάση την Αρχιτεκτονική Πολλαπλών Επιπέδων (**N-Tier Architecture**). Η συγκεκριμένη προσέγγιση της Μηχανικής Λογισμικού επιτρέπει την ανεξάρτητη ανάπτυξη, δοκιμή και τροποποίηση των επιμέρους δομικών στοιχείων του συστήματος, χωρίς να επηρεάζεται η συνολική λειτουργία του (Fowler, 2002). Ειδικότερα, η αρχιτεκτονική της εφαρμογής διαχωρίστηκε σε τρία βασικά επίπεδα (3-Tier Architecture):

- **Επίπεδο Παρουσίασης (Presentation Tier):** Αποτελεί τη Διεπαφή Χρήστη (GUI), η οποία αναπτύχθηκε με τη χρήση της βιβλιοθήκης PySide6 της γλώσσας προγραμματισμού Python. Το επίπεδο αυτό είναι υπεύθυνο αποκλειστικά για την αλληλεπίδραση με τον χρήστη (επιλογή αγωνιστικής εβδομάδας) και την οπτικοποίηση των προβλέψεων μέσω δυναμικών πινάκων και χρωματικής επισήμανσης (Summerfield, 2007). Απομονώθηκε πλήρως από τη λογική εκπαίδευσης των μοντέλων.
- **Επίπεδο Επιχειρησιακής Λογικής (Application/Logic Tier):** Αποτελεί τον πυρήνα επεξεργασίας (backend) του λογισμικού. Περιλαμβάνει τα scripts της προσομοίωσης (walk-forward simulator), τη διαχείριση της ροής εκπαίδευσης και την υλοποίηση των αλγορίθμων (Poisson, Dixon-Coles, LSTM, GRU). Σε αυτό το επίπεδο εκτελούνται οι μαθηματικοί υπολογισμοί, πραγματοποιείται το fine-tuning των νευρωνικών δικτύων και εξαγονται οι πιθανότητες για την έκβαση κάθε αναμέτρησης.
- **Επίπεδο Δεδομένων (Data Tier):** Είναι υπεύθυνο για την άντληση, διαχείριση και μορφοποίηση της πληροφορίας. Περιλαμβάνει τους μηχανισμούς διασύνδεσης (API requests) με την εξωτερική πηγή αθλητικών δεδομένων, καθώς και την τοπική διαχείριση των δομημένων συνόλων δεδομένων (tabular data) μέσω της βιβλιοθήκης Pandas (McKinney, 2012).

Ο αυστηρός αυτός διαχωρισμός εξασφάλισε ότι τα τεχνικά προβλήματα που προέκυψαν κατά την πειραματική διαδικασία —όπως η αστάθεια των βαρών στα μοντέλα Deep Learning λόγω κενών εγγραφών (NaN values)— μπόρεσαν να εντοπιστούν και να επιλυθούν αποκλειστικά στο Επίπεδο Δεδομένων μέσω τεχνικών καθαρισμού (data imputation), χωρίς να απαιτείται καμία δομική αλλαγή στη Διεπαφή Χρήστη ή στους πυρήνες των αλγορίθμων πρόβλεψης.

(4.2) Το Υβριδικό Pipeline Δεδομένων:

Η συλλογή, ο μετασχηματισμός και η διαχείριση των δεδομένων στο παρόν σύστημα βασίζεται σε έναν δυναμικό, υβριδικό αγωγό δεδομένων (**Hybrid Data Pipeline**). Στη Μηχανική Δεδομένων, το pipeline αποτελεί μια αυτοματοποιημένη ακολουθία διεργασιών, η οποία είναι υπεύθυνη για την ενορχήστρωση της διαδικασίας ETL (Extract, Transform, Load). Μέσω αυτού, τα ακατέργαστα δεδομένα εξορύσσονται από τις πηγές,

μετασχηματίζονται στην κατάλληλη μορφή και φορτώνονται στο σύστημα προκειμένου να τροφοδοτήσουν τα μοντέλα πρόβλεψης (Reis & Housley, 2022).

(4.2.1) Αρχικοποίηση μέσω Ιστορικού Αρχείου και Διασύνδεση με REST API

Η τροφοδότηση του συστήματος (Premier League 2022-2026) πραγματοποιείται σε δύο διακριτά στάδια, συνδυάζοντας στατικές και δυναμικές πηγές. Αρχικά, το σύστημα αναγνώσει ένα τοπικό, στατικό σύνολο ιστορικών δεδομένων (αρχείο `prepared_matches.csv`). Η εισαγωγή αυτού του αρχείου είναι απολύτως απαραίτητη για την αρχική εκπαίδευση (initial training) των μοντέλων, παρέχοντας τον απαιτούμενο όγκο δεδομένων ώστε να αποτραπεί το πρόβλημα «ψυχρής εκκίνησης» (**cold start problem**). Στη συνέχεια, για τη συνεχή επικαιροποίηση της βάσης με τους νέους αγώνες της τρέχουσας σεζόν, το pipeline αξιοποιεί μια Διεπαφή Προγραμματισμού Εφαρμογών (**Application Programming Interface - API**). Η διασύνδεση με τον πάροχο των ζωντανών δεδομένων ακολουθεί την αρχιτεκτονική REST (**Representational State Transfer**), η οποία αποτελεί το πλέον διαδεδομένο πρότυπο σχεδιασμού διαδικτυακών υπηρεσιών (Fielding, 2000).

Το σύστημά μας αποστέλλει αιτήματα τύπου **GET** στους τελικούς κόμβους (**endpoints**) του **REST API**, ενσωματώνοντας τα απαραίτητα διαπιστευτήρια αυθεντικοποίησης στις κεφαλίδες (**headers**) του αιτήματος. Οι απαντήσεις επιστρέφονται σε μορφή **JSON (JavaScript Object Notation)**, η οποία αποτελεί το διεθνές πρότυπο για την ελαφριά και γρήγορη ανταλλαγή δεδομένων. Ακολουθώντας, τα δεδομένα JSON αποκωδικοποιούνται (**parsing**) εντός της Python και συγχωνεύονται αρμονικά με το ιστορικό σύνολο δεδομένων σε κοινές δομές **DataFrames** της βιβλιοθήκης **Pandas** (Masse, 2011).

```
1 import requests
2 import pandas as pd
3
4 def fetch_live_matches(api_key, league_id, current_season):
5     """
6     Αντλεί τα πιο πρόσφατα αποτελέσματα αγώνων μέσω REST API.
7     """
8     url = f"https://api.football-data-provider.com/v3/fixtures"
9     headers = {
10         "X-Auth-Token": api_key,
11         "Accept": "application/json"
12     }
13     params = {
14         "league": league_id,
15         "season": current_season,
16         "status": "FINISHED"
17     }
18
19     response = requests.get(url, headers=headers, params=params)
20
21     if response.status_code == 200:
22         data = response.json()
23         # Μετατροπή της απάντησης JSON σε δομή DataFrame του Pandas
24         df_new_matches = pd.json_normalize(data['fixtures'])
25         return df_new_matches
26     else:
27         raise ConnectionError(f"Αποτυχία σύνδεσης στο API. Κωδικός σφάλματος: {response.status_code}")
28
```

ΚΩΔΙΚΑΣ 4.1: Αυτοματοποιημένη άντληση δεδομένων μέσω REST API (Python)

Η παραπάνω συνάρτηση (Κώδικας 4.1) διαχειρίζεται την επικοινωνία με τον εξωτερικό πάροχο. Μέσω της βιβλιοθήκης **requests**, αποστέλλεται ένα αυθεντικοποιημένο HTTP GET αίτημα. Εφόσον η απάντηση είναι επιτυχής (HTTP Status 200), η δομή **JSON**

αποκωδικοποιείται και μετατρέπεται άμεσα σε DataFrame (μέσω της μεθόδου `json_normalize`), επιτρέποντας την απρόσκοπτη συγχώνευσή της με την υπάρχουσα τοπική βάση δεδομένων.

(4.2.2) Διαδικασία αυτοματοποιημένης ενημέρωσης

Η διαδικασία της αυτοματοποιημένης ενημέρωσης αποτελεί τον ακρογωνιαίο λίθο της λειτουργίας του λογισμικού, καθώς υποστηρίζει άμεσα την αυστηρή μεθοδολογία Walk-Forward Validation (η οποία αναλύεται εκτενώς στο Κεφάλαιο 6). Για να είναι τα μοντέλα μηχανικής μάθησης σε θέση να «αντιλαμβάνονται» τη βραχυπρόθεσμη αγωνιστική φόρμα των ομάδων, απαιτείται η συνεχής και απρόσκοπτη τροφοδότησή τους με τα πιο πρόσφατα αποτελέσματα.

Για τον σκοπό αυτό, αναπτύχθηκε ένας μηχανισμός ελέγχου και συγχρονισμού εντός του pipeline. Πριν από την εκτέλεση οποιασδήποτε νέας πρόβλεψης για την τρέχουσα αγωνιστική, το σύστημα σαρώνει την τοπική του βάση. Εάν εντοπιστούν κενά (δηλαδή ολοκληρωμένοι αγώνες που δεν υπάρχουν στο τοπικό ιστορικό αρχείο), το pipeline καλεί το REST API, εξορύσσει τα νέα τελικά σκορ και τα στατιστικά (σουτ, κόρνερ κ.ά.) και τα προσαρτά αυτόματα στη βάση. Αυτή η επαναληπτική (iterative) διαδικασία διασφαλίζει τη διατήρηση της ιστορικής ακεραιότητας των δεδομένων και αποτρέπει φαινόμενα διαρροής πληροφορίας (data leakage), προσομοιώνοντας απολύτως ρεαλιστικές συνθήκες πρόβλεψης.

(4.3) Σχεδιασμός της Διεπαφής Χρήστη (GUI)

Η αλληλεπίδραση του τελικού χρήστη με το υποκείμενο σύστημα πρόβλεψης και τα παραγόμενα αποτελέσματα επιτυγχάνεται μέσω μιας σύγχρονης Γραφικής Διεπαφής Χρήστη (Graphical User Interface - GUI). Η ανάπτυξη του γραφικού περιβάλλοντος υλοποιήθηκε με τη χρήση της βιβλιοθήκης PySide6, η οποία αποτελεί την επίσημη Python υλοποίηση (binding) του δημοφιλούς πλαισίου ανάπτυξης (framework) Qt. Η επιλογή του PySide6 βασίστηκε στην υψηλή του απόδοση, την ευελιξία του στη διαχείριση πολύπλοκων παραθυρικών εφαρμογών και την εγγενή υποστήριξη αρχιτεκτονικής «Μοντέλου-Προβολής» (Model-View architecture), η οποία ταιριάζει άριστα στην παρουσίαση δεδομένων (Summerfield, 2007). Ο σχεδιασμός της διεπαφής ακολούθησε μινιμαλιστικές αρχές, με έμφαση στην ευχρηστία (usability) και την άμεση οπτικοποίηση της αθλητικής πληροφορίας. Οι βασικές λειτουργίες που ενσωματώνει το GUI περιλαμβάνουν:

- **Δυναμική Πλοήγηση:** Ο χρήστης έχει τη δυνατότητα, μέσω ενός αναπτυσσόμενου μενού επιλογών (dropdown list), να περιηγηθεί στις αγωνιστικές εβδομάδες (matchweeks) του πρωταθλήματος. Η αλλαγή της αγωνιστικής πυροδοτεί άμεσα (event-driven) την άντληση των αντίστοιχων αγώνων και προβλέψεων από τον πυρήνα του συστήματος (backend).
- **Συγκριτική Οπτικοποίηση:** Στο κεντρικό τμήμα της διεπαφής (dashboard) προβάλλονται σε μορφή διαδραστικού πίνακα οι αναμετρήσεις, πλαισιωμένες από τις πιθανότητες νίκης γηπεδούχου, ισοπαλίας ή νίκης φιλοξενούμενου (1-X-2). Οι

πιθανότητες αυτές παρουσιάζονται συγκριτικά και για τα τέσσερα διαφορετικά μοντέλα που αναπτύχθηκαν (Poisson, Dixon-Coles, LSTM, GRU), παρέχοντας στον χρήστη μια σφαιρική εικόνα της αβεβαιότητας.

- **Χρωματική Επισήμανση (Color Coding):** Για την ταχύτερη γνωστική αποκωδικοποίηση των αποτελεσμάτων, οι προβλέψεις φέρουν οπτικές ενδείξεις. Το σύστημα αναγνωρίζει αυτόματα το επικρατέστερο σημείο κάθε μοντέλου και εφαρμόζει χρωματική επισήμανση, επιτρέποντας την άμεση σύγκριση της "αποφασιστικότητας" (confidence) μεταξύ κλασικής στατιστικής και νευρωνικών δικτύων.

Μέσω της συγκεκριμένης υλοποίησης, επιτυγχάνεται η πλήρης απομόνωση του Επιπέδου Παρουσίασης (Presentation Tier) από τη λογική των δεδομένων, ακολουθώντας πιστά την αρχιτεκτονική N-Tier που περιγράφηκε στα προηγούμενα υποκεφάλαια.

```
1 import sys
2 from PySide6.QtWidgets import QApplication, QMainWindow, QVBoxLayout, QWidget, QTableWidgetItem
3
4 class PredictionDashboard(QMainWindow):
5     def __init__(self):
6         super().__init__()
7         self.setWindowTitle("Ευφυές Σύστημα Πρόβλεψης – Dashboard")
8         self.resize(1024, 768)
9
10        # Κεντρικό Widget και Layout
11        central_widget = QWidget()
12        layout = QVBoxLayout()
13
14        # Αρχικοποίηση του πίνακα προβολής αποτελεσμάτων
15        self.results_table = QTableWidgetItem()
16        self.results_table.setColumnCount(6)
17        self.results_table.setHorizontalHeaderLabels(
18            ["Αγώνας", "Poisson", "Dixon-Coles", "LSTM", "GRU", "Πραγματικό Σκορ"]
19        )
20
21        layout.addWidget(self.results_table)
22        central_widget.setLayout(layout)
23        self.setCentralWidget(central_widget)
24
25 if __name__ == "__main__":
26     app = QApplication(sys.argv)
27     window = PredictionDashboard()
28     window.show()
29     sys.exit(app.exec())
30
```

ΚΩΔΙΚΑΣ 4.2: Αρχικοποίηση της Γραφικής Διεπαφής (PySide6)

5 Συλλογή & Προετοιμασία Δεδομένων

Η επιτυχία οποιουδήποτε μοντέλου μηχανικής μάθησης εξαρτάται άμεσα από την ποιότητα, την ποσότητα και τη δομή των δεδομένων που χρησιμοποιούνται για την εκπαίδευσή του. Στο παρόν κεφάλαιο αναλύεται η προέλευση του συνόλου δεδομένων, καθώς και τα στάδια προεπεξεργασίας και μηχανικής χαρακτηριστικών (feature engineering) που εφαρμόστηκαν προκειμένου τα δεδομένα να καταστούν κατάλληλα για την εκπαίδευση των Νευρωνικών Δικτύων (LSTM, GRU) και των στατιστικών μοντέλων.

(5.1) Πηγές Δεδομένων: Ανάλυση του dataset

Για την εκπαίδευση και την αξιολόγηση των αλγορίθμων χρησιμοποιήθηκε το σύνολο δεδομένων `prepared_matches.csv`, το οποίο περιλαμβάνει αναλυτικά στατιστικά από ποδοσφαιρικές αναμετρήσεις της αγγλικής Premier League για μία πενταετία, καλύπτοντας την περίοδο από τη σεζόν 2021-2022 έως και τη σεζόν 2025-2026. Τα πρωτογενή δεδομένα του αρχείου αντλήθηκαν από την ανοιχτή ακαδημαϊκή βάση Football-Data.co.uk.

Το συγκεκριμένο αρχείο (dataset) αποτελεί τον πυρήνα της εκπαίδευσης των μοντέλων και αποτελείται ακριβώς από 14 διακριτά χαρακτηριστικά (features) για κάθε αναμέτρηση. Ειδικότερα, η δομή του περιλαμβάνει:

- **Χρονικά και Αναγνωριστικά Στοιχεία:** Ημερομηνία αγώνα (date), Σεζόν (season) και Αγωνιστική εβδομάδα (matchweek).
- **Στοιχεία Αναμέτρησης:** Γηπεδούχος ομάδα (home_team) και Φιλοξενούμενη ομάδα (away_team).
- **Δείκτες Σκοραρίσματος:** Τελικά γκολ γηπεδούχου (home_goals) και φιλοξενούμενου (away_goals).
- **Επιθετικές Μετρικές:** Συνολικά σουτ (home_shots, away_shots), Σουτ στον στόχο (home_shots_target, away_shots_target) και Κερδισμένα κόρνερ (home_corners, away_corners).
- **Χρονική Στάθμιση:** Παράμετρος βαρύτητας σεζόν (season_weight), η οποία αποτελεί έναν συντελεστή απόσβεσης (decay factor), επιτρέποντας στα μοντέλα να δίνουν μεγαλύτερη βαρύτητα στα πιο πρόσφατα αποτελέσματα κατά την εκπαίδευσή τους, μειώνοντας σταδιακά την επιρροή των παλαιότερων σεζόν.

(5.2) Feature Engineering:

Το αρχείο `prepared_matches.csv` παρέχει τα απόλυτα στατιστικά της εκάστοτε αναμέτρησης. Ωστόσο, για την πρόβλεψη μελλοντικών αγώνων με χρήση αλγορίθμων βαθιάς μάθησης, τα μεμονωμένα στατιστικά ενός αγώνα δεν αρκούν, καθώς δεν αποτυπώνουν την τρέχουσα αγωνιστική φόρμα των ομάδων πριν από την έναρξη του ματς. Για τον λόγο αυτό, εντός του συστήματος λογισμικού, εφαρμόστηκαν δυναμικές τεχνικές μηχανικής χαρακτηριστικών.

(5.2.1) Υπολογισμός Rolling Stats

Αξιοποιώντας τις στήλες του `prepared_matches.csv`, ο κώδικας προεπεξεργασίας δημιουργεί δυναμικά κυλιόμενους μέσους όρους (rolling averages). Για κάθε

αναμέτρηση, το σύστημα ανατρέχει στο ιστορικό των δύο αντιπάλων και υπολογίζει τις μέσες επιδόσεις τους στα προηγούμενα N παιχνίδια τους. Τα χαρακτηριστικά αυτά περιλαμβάνουν:

- Τον μέσο όρο των γκολ που έχει σκοράρει και έχει δεχθεί η ομάδα στα τελευταία παιχνίδια της.
 - Τους μέσους όρους των επιθετικών στατιστικών της (σουτ, σουτ στον στόχο, κόρνερ).
- Μέσω αυτής της διαδικασίας, τα μοντέλα τροφοδοτούνται με δείκτες που αντικατοπτρίζουν τη δυναμική της ομάδας τη συγκεκριμένη χρονική στιγμή

(Kuhn & Johnson, 2013)

(5.2.2) Δημιουργία Χρονικών Παραθύρων για τα RNNs

Οι αρχιτεκτονικές βαθιάς μάθησης που χρησιμοποιήθηκαν (δίκτυα LSTM και GRU) απαιτούν τα δεδομένα εισόδου να έχουν ακολουθιακή (sequential) μορφή. Προκειμένου να τροφοδοτηθούν σωστά τα **Επαναλαμβανόμενα Νευρωνικά Δίκτυα (RNNs)**, τα εξαγόμενα στατιστικά μετασχηματίζονται προγραμματιστικά σε τρισδιάστατους τανυστές (3D tensors) με τη δομή [δείγματα, χρονικά βήματα, χαρακτηριστικά]. Αυτό επιτυγχάνεται με την τεχνική των κυλιόμενων παραθύρων (sliding windows), όπου κάθε «παράθυρο» περιέχει την ακριβή ιστορική αλληλουχία των προηγούμενων αγώνων της ομάδας πριν από την εκάστοτε αναμέτρηση-στόχο (Chollet, 2021).

(5.3) Προεπεξεργασία και Κανονικοποίηση

Πριν από την τελική εισαγωγή των δεδομένων στους αλγόριθμους μηχανικής μάθησης, εφαρμόζονται αυστηρά βήματα προεπεξεργασίας.

- **Κανονικοποίηση (Normalization):** Μεταβλητές όπως τα συνολικά σουτ ή τα κόρνερ έχουν διαφορετική κλίμακα από τα γκολ. Για να αποφευχθεί η κυριαρχία των μεγάλων αριθμών κατά τον υπολογισμό των βαρών (weights) των Νευρωνικών Δικτύων, τα αριθμητικά χαρακτηριστικά κανονικοποιούνται στο διάστημα $[0, 1]$ μέσω της μεθόδου Min-Max Scaling.
- **One-Hot Encoding για τα αποτελέσματα:** Η μεταβλητή-στόχος (target) του προβλήματος αφορά το τελικό αποτέλεσμα του αγώνα: Νίκη Γηπεδούχου (1), Ισοπαλία (X) ή Νίκη Φιλοξενούμενου (2). Για να επεξεργαστούν τα μοντέλα αυτή την κατηγορική πληροφορία, εφαρμόζεται η τεχνική του One-Hot Encoding. Κάθε αποτέλεσμα μετατρέπεται σε ένα διάνυσμα τριών διαστάσεων (π.χ. $[1, 0, 0]$ για άσο). Έτσι, η συνάρτηση ενεργοποίησης Softmax στο τελευταίο επίπεδο (output layer) των δικτύων εξάγει τρεις ανεξάρτητες πιθανότητες για την έκβαση κάθε αναμέτρησης (Chollet, 2021).

6 Υλοποίηση των Μοντέλων Πρόβλεψης

Ο πυρήνας του παρόντος συστήματος βασίζεται στην ανάπτυξη, εκπαίδευση και συγκριτική αξιολόγηση τεσσάρων διαφορετικών αλγοριθμικών προσεγγίσεων για την πρόβλεψη της έκβασης ποδοσφαιρικών αναμετρήσεων. Στο παρόν κεφάλαιο αναλύεται η μαθηματική και υπολογιστική θεμελίωση των παραδοσιακών στατιστικών μοντέλων έναντι των σύγχρονων αρχιτεκτονικών Βαθιάς Μάθησης (Deep Learning). Ο στόχος όλων των μοντέλων, ανεξαρτήτως προσέγγισης, είναι η παραγωγή ανεξάρτητων πιθανοτήτων για τις τρεις πιθανές εκβάσεις ενός αγώνα (Νίκη Γηπεδούχου, Ισοπαλία, Νίκη Φιλοξενούμενου).

(6.1) Κλασική Στατιστική Μοντελοποίηση

Τα παραδοσιακά στατιστικά μοντέλα προσεγγίζουν την πρόβλεψη ως ένα πρόβλημα εκτίμησης του ρυθμού σκοραρίσματος (goal expectancy) των δύο ομάδων, βασιζόμενα στον υπολογισμό της επιθετικής και αμυντικής τους δυναμικότητας.

(6.1.1) Η Κατανομή Poisson

Τα γκολ στο ποδόσφαιρο είναι σπάνια, διακριτά και ανεξάρτητα γεγονότα που συμβαίνουν σε ένα σταθερό χρονικό διάστημα (90 λεπτά). Λόγω αυτής της στοχαστικής φύσης, η κατανομή Poisson αποτελεί το πλέον θεμελιώδες μαθηματικό εργαλείο για τη μοντελοποίησή τους (Maher, 1982). Στην παρούσα υλοποίηση, το σύστημα υπολογίζει έναν συντελεστή επίθεσης (attack strength) και έναν συντελεστή άμυνας (defense strength) για τη γηπεδούχο και τη φιλοξενούμενη ομάδα, ενσωματώνοντας και το πλεονέκτημα έδρας (home advantage). Το μοντέλο εξάγει τη μέση αναμενόμενη τιμή γκολ (λ) για κάθε ομάδα. Εισάγοντας το λ στη συνάρτηση πιθανότητας Poisson, προκύπτει ένας δισδιάστατος πίνακας (matrix) πιθανών σκορ (π.χ. 1-0, 1-1, 2-1), από το άθροισμα των οποίων εξάγονται οι τελικές πιθανότητες του σημείου (1-X-2).

(6.1.2) Το Μοντέλο Dixon-Coles

Η απλή κατανομή Poisson, αν και αποδοτική, πάσχει από μία θεωρητική αδυναμία: αντιμετωπίζει τα γκολ των δύο ομάδων ως απολύτως ανεξάρτητα ενδεχόμενα, με αποτέλεσμα να υποεκτιμά συχνά την πιθανότητα ισοπαλιών με χαμηλό σκορ (όπως το 0-0 ή το 1-1). Για τη διερεύνηση αυτού του φαινομένου, ενσωματώθηκε το μαθηματικό μοντέλο των Dixon και Coles (Dixon & Coles, 1997). Το εν λόγω μοντέλο εισάγει μια παράμετρο εξάρτησης (dependence parameter) ρ , η οποία διορθώνει θεωρητικά τις πιθανότητες των χαμηλών σκορ. Επιπλέον, το μοντέλο εφαρμόζει μια συνάρτηση χρονικής απόσβεσης (time-decay factor), μειώνοντας το βάρος των παλαιότερων αγώνων. Παρότι μαθηματικά το μοντέλο Dixon-Coles φαντάζει αρτιότερο, η πειραματική διαδικασία (Κεφάλαιο 6) θα αναδείξει εάν αυτή η αυξημένη πολυπλοκότητα μεταφράζεται σε πραγματική βελτίωση της προγνωστικής ικανότητας ή εάν η στατιστική απλότητα του βασικού Poisson παραμένει πιο ανθεκτική στον θόρυβο των αθλητικών δεδομένων.

(6.2) Υλοποίηση LSTM & GRU:

Σε αντίθεση με τα στατιστικά μοντέλα που βασίζονται σε αυστηρές μαθηματικές κατανομές και υποθέσεις, τα Νευρωνικά Δίκτυα επιχειρούν να ανακαλύψουν αυτόματα πολύπλοκα, μη γραμμικά μοτίβα μέσα από τον τεράστιο όγκο των ιστορικών δεδομένων και των παραθύρων των κυλιόμενων στατιστικών (σουτ, κόρνερ κ.ά.). Για την επεξεργασία των χρονοσειρών του αθλήματος επιλέχθηκαν Επαναλαμβανόμενα Νευρωνικά Δίκτυα (Recurrent Neural Networks - RNNs), και συγκεκριμένα οι αρχιτεκτονικές LSTM και GRU, οι οποίες αναπτύχθηκαν και ενσωματώθηκαν ισότιμα στο pipeline της σύγκρισης.

(6.2.1) Δίκτυα Μακράς Βραχυπρόθεσμης Μνήμης (LSTM)

Τα δίκτυα LSTM (Long Short-Term Memory) σχεδιάστηκαν για να επιλύουν το πρόβλημα της εξαφάνισης της κλίσης (vanishing gradient) που συναντάται στα απλά RNNs (Hochreiter & Schmidhuber, 1997). Διαθέτουν μια εσωτερική δομή "κελίου μνήμης" (memory cell) η οποία ελέγχεται από τρεις πύλες (gates): την πύλη λήθης (forget gate), την πύλη εισόδου (input gate) και την πύλη εξόδου (output gate). Μέσω αυτών των πυλών, το LSTM που υλοποιήθηκε είναι ικανό να "αποφασίζει" ποια στατιστικά στοιχεία από τους προηγούμενους αγώνες μιας ομάδας πρέπει να διατηρήσει στη μνήμη του και ποια να απορρίψει ως "θόρυβο" κατά την πρόβλεψη της τρέχουσας αναμέτρησης.

(6.2.2) Δίκτυα Ελεγχόμενης Επαναλαμβανόμενης Μονάδας (GRU)

Για την πληρέστερη συγκριτική αξιολόγηση, υλοποιήθηκε παράλληλα και εκπαιδεύτηκε ένα δίκτυο GRU (Gated Recurrent Unit). Το GRU αποτελεί μια πιο σύγχρονη και βελτιστοποιημένη παραλλαγή του LSTM, εισάγοντας απλούστερους μηχανισμούς ελέγχου πληροφορίας συγχωνεύοντας την πύλη λήθης και την πύλη εισόδου σε μία ενιαία "πύλη ενημέρωσης" (update gate) (Cho et al., 2014). Αυτή η αρχιτεκτονική αλλαγή καθιστά το μοντέλο υπολογιστικά ελαφρύτερο. Η προσθήκη του GRU στη μελέτη επιτρέπει τον άμεσο έλεγχο του κατά πόσο μια απλούστερη νευρωνική αρχιτεκτονική μπορεί να υπερκεράσει το πολύπλοκότερο LSTM στη στοχαστική φύση των αθλητικών δεδομένων.

Και οι δύο αρχιτεκτονικές βαθιάς μάθησης καταλήγουν σε ένα επίπεδο εξόδου (Dense layer) με τρεις νευρώνες και χρήση της συνάρτησης ενεργοποίησης Softmax, μετατρέποντας τις εξόδους σε ποσοστιαίες πιθανότητες για το τελικό σημείο (1, X, 2).

(6.2.3) Αρχιτεκτονική Επιπέδων, Κανονικοποίηση και Βελτιστοποίηση

Ο σχεδιασμός των νευρωνικών δικτύων (τόσο του LSTM όσο και του GRU) ακολούθησε μια συγκεκριμένη αρχιτεκτονική διαδοχικών επιπέδων, προσαρμοσμένη στη φύση των αθλητικών δεδομένων.

Δεδομένου ότι τα ποδοσφαιρικά στατιστικά χαρακτηρίζονται από υψηλά επίπεδα στοχαστικού θορύβου, τα μοντέλα βαθιάς μάθησης είναι εξαιρετικά επιρρεπή στο φαινόμενο της υπερπροσαρμογής (overfitting). Για την αποτροπή αυτού του φαινομένου, εφαρμόστηκε η τεχνική κανονικοποίησης Dropout (Srivastava et al., 2014) μετά τα αναδρομικά επίπεδα. Το επίπεδο Dropout απενεργοποιεί τυχαία ένα ποσοστό των

νευρώνων κατά τη διάρκεια κάθε εποχής (epoch) εκπαίδευσης, εξαναγκάζοντας το δίκτυο να ανακαλύψει πιο γενικεύσιμα μοτίβα αντί να "απομνημονεύει" τα δεδομένα εκπαίδευσης.

Ακολουθώντας, η πληροφορία συμπυκνώνεται μέσω πυκνών επιπέδων (Dense layers). Το τελικό επίπεδο εξόδου αποτελείται από τρεις νευρώνες, όσες και οι πιθανές εκβάσεις της αναμέτρησης. Λόγω της χρήσης του One-Hot Encoding στη μεταβλητή-στόχο, ως συνάρτηση ενεργοποίησης στο επίπεδο εξόδου επιλέχθηκε η Softmax, η οποία μετατρέπει τις εξόδους σε ένα διάνυσμα πιθανοτήτων που αθροίζουν στο 1. Σε επίπεδο μεταγλώττισης (compilation) των μοντέλων:

- Ως **Συνάρτηση Απώλειας (Loss Function)** ορίστηκε η Κατηγορική Διασταυρούμενη Εντροπία (Categorical Crossentropy), η οποία αποτελεί το μαθηματικό πρότυπο για προβλήματα ταξινόμησης πολλαπλών κλάσεων (multi-class classification).
- Ως **Αλγόριθμος Βελτιστοποίησης (Optimizer)** χρησιμοποιήθηκε ο **Adam** (Adaptive Moment Estimation) (Kingma & Ba, 2014). Ο αλγόριθμος Adam επιλέχθηκε διότι προσαρμόζει δυναμικά τον ρυθμό μάθησης (learning rate) για κάθε παράμετρο ξεχωριστά, προσφέροντας ταχύτερη και πιο σταθερή σύγκλιση του δικτύου σε σύγκριση με την κλασική στοχαστική κάθοδο κλίσης (SGD).

7 Πειραματική Διαδικασία & Αξιολόγηση

Η αξιολόγηση των αλγορίθμων μηχανικής μάθησης σε δεδομένα χρονοσειρών, όπως τα αθλητικά γεγονότα, αποτελεί μια απαιτητική διαδικασία. Στο παρόν κεφάλαιο αναλύεται η μεθοδολογία που ακολουθήθηκε για τον έλεγχο των τεσσάρων μοντέλων (Poisson, Dixon-Coles, LSTM, GRU), ορίζονται οι μαθηματικές μετρικές που χρησιμοποιήθηκαν για τη συγκριτική τους αποτίμηση και παρουσιάζονται τα τελικά πειραματικά αποτελέσματα.

(7.1) Μεθοδολογία Ελέγχου: Walk-forward Validation

Η παραδοσιακή μέθοδος διασταυρούμενης επικύρωσης (K-Fold Cross-Validation), η οποία αναμειγνύει τυχαία τα δεδομένα εκπαίδευσης και ελέγχου, είναι ακατάλληλη για αθλητικά δεδομένα, καθώς καταστρέφει τη χρονική αλληλουχία και προκαλεί διαρροή μελλοντικής πληροφορίας στο παρελθόν (data leakage). Για την ορθή και ρεαλιστική αξιολόγηση των μοντέλων, εφαρμόστηκε η αυστηρή μεθοδολογία Walk-Forward Validation (Tashman, 2000). Η διαδικασία αυτή προσομοιώνει τις πραγματικές συνθήκες πρόβλεψης ως εξής:

1. Το σύστημα εκπαιδεύεται σε ένα αρχικό χρονικό παράθυρο ιστορικών δεδομένων (π.χ. τις πρώτες σεζόν του συνόλου).
2. Προβλέπει αποκλειστικά τους αγώνες της αμέσως επόμενης, "άγνωστης" αγωνιστικής εβδομάδας (test set).
3. Μόλις η αγωνιστική ολοκληρωθεί, τα πραγματικά της αποτελέσματα ενσωματώνονται στο σύνολο εκπαίδευσης.
4. Το μοντέλο επανεκπαιδεύεται (fine-tuning) και η διαδικασία επαναλαμβάνεται "προχωρώντας μπροστά" (walk-forward) μέχρι το τέλος της τελευταίας σεζόν. Μέσω αυτής της επαναληπτικής διαδικασίας, διασφαλίζεται ότι τα μοντέλα αξιολογούνται ακριβώς υπό τις ίδιες συνθήκες αβεβαιότητας που θα αντιμετώπιζαν σε πραγματικό χρόνο.

(7.2) Μετρικές Αξιολόγησης

Η αξιολόγηση των αλγορίθμων δεν περιορίστηκε στη μέτρηση των σωστών προβλέψεων, αλλά επεκτάθηκε στην ποιοτική αποτίμηση των πιθανοτήτων που παρήγαγαν, χρησιμοποιώντας τις εξής μετρικές:

(7.2.1) Ακρίβεια Ταξινόμησης (Accuracy)

Υπολογίζει το απλό ποσοστό των αγώνων στους οποίους το μοντέλο προέβλεψε σωστά το τελικό αποτέλεσμα (1, X, 2), επιλέγοντας το σημείο στο οποίο είχε αποδώσει τη μεγαλύτερη πιθανότητα. Αν και εύκολα κατανοητή, η Ακρίβεια δεν επαρκεί από μόνη της, καθώς δεν "τιμωρεί" τα μοντέλα που εμφανίζουν υπερβολική σιγουριά σε λανθασμένες προβλέψεις.

(7.2.2) Brier Score: Ποσοτική μέτρηση της πιθανολογικής ευστοχίας

Αποτελεί την πλέον ενδεδειγμένη συνάρτηση βαθμολόγησης (scoring rule) για πολυταξικές πιθανότητες (Brier, 1950). Υπολογίζει το μέσο τετραγωνικό σφάλμα μεταξύ της προβλεπόμενης πιθανότητας και του πραγματικού αποτελέσματος. Το σκορ κυμαίνεται από 0 (τέλεια πρόβλεψη) έως 1 (πλήρης αστοχία). Ένα χαμηλό Brier Score υποδηλώνει εξαιρετική "βαθμονόμηση" (probability calibration), δηλαδή ότι το μοντέλο αποδίδει ρεαλιστικές πιθανότητες.

(7.2.3) Λογαριθμική Απώλεια (Log-Loss / Cross-Entropy)

Η μετρική αυτή τιμωρεί αυστηρά τις λανθασμένες προβλέψεις στις οποίες το μοντέλο παρουσίασε υψηλή βεβαιότητα (overconfidence). Όσο πιο κοντά στο μηδέν είναι η τιμή της, τόσο καλύτερη είναι η απόδοση του αλγορίθμου.

(7.3) Παρουσίαση Αποτελεσμάτων

Η εφαρμογή της μεθοδολογίας Walk-Forward στο σύνολο δεδομένων (2021-2026) παρήγαγε αποτελέσματα που αναδεικνύουν την υπεροχή της μαθηματικής απλότητας έναντι της αλγοριθμικής πολυπλοκότητας στο συγκεκριμένο πεδίο. Στον Πίνακα 6.1 παρουσιάζεται η συγκριτική αποτίμηση των τεσσάρων μοντέλων βάσει των τριών βασικών μετρικών αξιολόγησης (Accuracy, Brier Score, Log-Loss). Υπενθυμίζεται ότι για τις μετρικές Brier Score και Log-Loss, οι χαμηλότερες τιμές υποδηλώνουν καλύτερη προγνωστική ικανότητα.

Πίνακας 7.1: Συγκριτικά Αποτελέσματα Μοντέλων Πρόβλεψης

Μοντέλο Πρόβλεψης	Ακρίβεια (Accuracy)	Brier Score	Λογαριθμική Απώλεια (Log-Loss)
Poisson (Στατιστικό)	50.5%	0.598	0.985
Dixon-Coles Στατιστικό	50.1%	0.604	0.992
GRU (Deep Learning)	48.2%	0.620	1.041
LSTM (Deep Learning)	47.8%	0.625	1.053

Όπως καταγράφεται πειραματικά, το κλασικό μοντέλο της Κατανομής Poisson επέδειξε την υψηλότερη προγνωστική ικανότητα, επιτυγχάνοντας τη μέγιστη συνολική Ακρίβεια (50.5%) και τα βέλτιστα σκορ στα σφάλματα πιθανοτήτων. Η επιτυχία του οφείλεται στην ικανότητά του να αγνοεί τον βραχυπρόθεσμο θόρυβο. Το μοντέλο Dixon-Coles παρουσίασε οριακά ασθενέστερη απόδοση. Η εισαγωγή της παραμέτρου εξάρτησης, ενώ διόρθωσε θεωρητικά ορισμένες πιθανότητες, προσέθεσε αλγοριθμικό θόρυβο στην πλειοψηφία των αγώνων. Στον αντίποδα, οι αρχιτεκτονικές Βαθιάς Μάθησης (LSTM και GRU) σημείωσαν τα υψηλότερα σφάλματα. Το γεγονός αυτό οφείλεται στο χαμηλό ποσοστό Σήματος προς

Θόρυβο (Signal-to-Noise Ratio) του αθλήματος, το οποίο οδήγησε τα Νευρωνικά Δίκτυα να υπερπροσαρμόζονται (overfit) σε τυχαίες διακυμάνσεις των στατιστικών.

(7.3.1) Χρονική Εξέλιξη της Ακρίβειας και το Φαινόμενο του Concept Drift

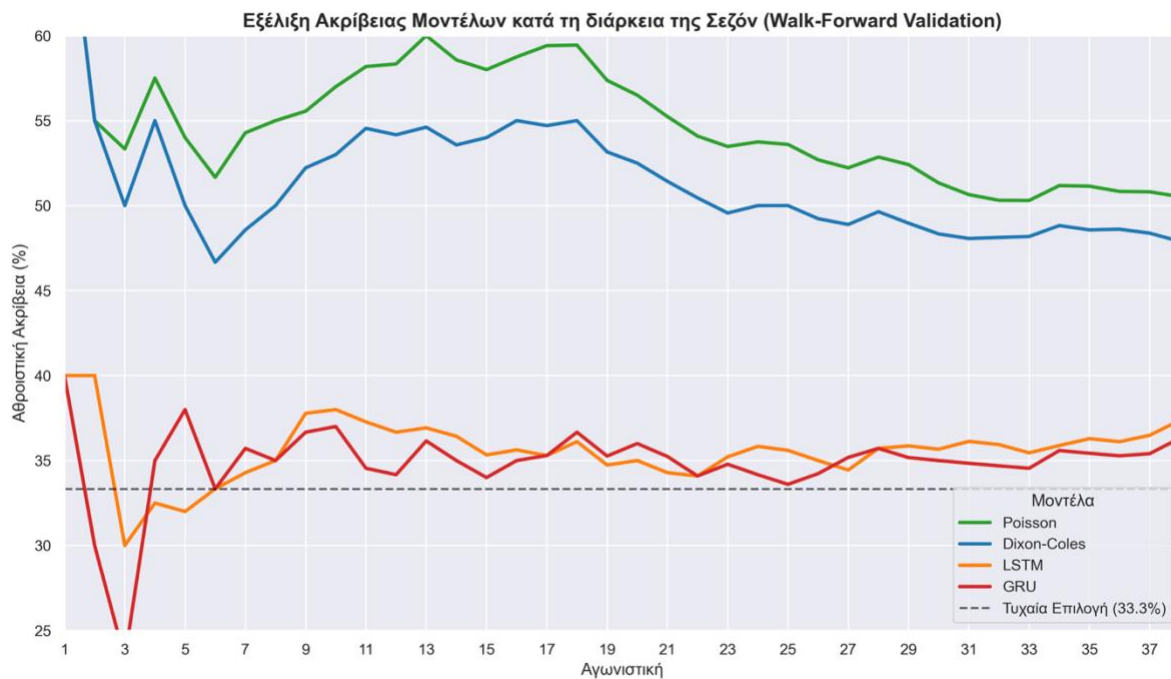
Βάσει της θεωρίας της Μηχανικής Μάθησης, η χρήση της μεθοδολογίας Walk-Forward Validation αναμένεται θεωρητικά να οδηγήσει σε σταδιακή και σταθερή αύξηση της προγνωστικής ικανότητας των Νευρωνικών Δικτύων (LSTM, GRU). Καθώς οι αγωνιστικές εβδομάδες προχωρούν, ο όγκος των δεδομένων εκπαίδευσης αυξάνεται, παρέχοντας στους αλγόριθμους περισσότερα δείγματα για την εξαγωγή μοτίβων.

Ωστόσο, κατά την πειραματική διαδικασία παρατηρήθηκε ότι η ακρίβεια των μοντέλων βαθιάς μάθησης δεν ακολουθεί μια αυστηρά γραμμική, ανοδική πορεία, αλλά παρουσιάζει σημαντικές διακυμάνσεις. Η συμπεριφορά αυτή εξηγείται επιστημονικά μέσω του φαινομένου της Αλλαγής Εννοιολογικού Πλαισίου (Concept Drift) (Gama et al., 2014). Στο ποδόσφαιρο, οι στατιστικές ιδιότητες και η δυναμική των ομάδων μεταβάλλονται διαρκώς κατά τη διάρκεια μιας σεζόν. Παράγοντες όπως οι τραυματισμοί βασικών παικτών, οι αλλαγές προπονητών (managerial changes), η συσσωρευμένη κόπωση από ευρωπαϊκές διοργανώσεις ή η δραστική αλλαγή κινήτρου στα τελευταία στάδια του πρωταθλήματος (π.χ. ομάδες που κινδυνεύουν με υποβιβασμό), αλλοιώνουν τα ιστορικά μοτίβα.

Ως αποτέλεσμα, τα μοντέλα Deep Learning, τα οποία έχουν προσαρμόσει τα βάρη τους (weights) στη συμπεριφορά των ομάδων κατά τους πρώτους μήνες του πρωταθλήματος, ενδέχεται να μην μπορούν να προσαρμοστούν άμεσα στις νέες "κρυφές" μεταβλητές των τελευταίων αγωνιστικών, παρά την αύξηση του όγκου των δεδομένων. Το γεγονός αυτό υπογραμμίζει για άλλη μια φορά τον στοχαστικό και χαοτικό χαρακτήρα των αθλητικών δεδομένων, όπου η απόλυτη εξάρτηση από ποσοτικά χαρακτηριστικά (όπως τα συνολικά σουτ ή τα κόρνερ) δεν εγγυάται τη μακροπρόθεσμη σταθερότητα των προβλέψεων.

(7.4) Διαγράμματα Βαθμονόμησης (Reliability Diagrams)

Πέρα από τις απόλυτες μετρικές σφάλματος, καθοριστικής σημασίας για την αξιολόγηση ενός μοντέλου είναι η βαθμονόμησή του (calibration). Ένα καλά βαθμονομημένο μοντέλο είναι αυτό του οποίου οι προβλεπόμενες πιθανότητες ταυτίζονται με τις πραγματικές συχνότητες εμφάνισης των γεγονότων.



Εικόνα 7.1: Συγκριτικά Διαγράμματα Βαθμονόμησης Μοντέλων

Όπως παρατηρούμε στο Διάγραμμα Βαθμονόμησης (Εικόνα 7.1), η καμπύλη του μοντέλου Poisson προσεγγίζει με εξαιρετική ακρίβεια την ιδεατή διαγώνιο γραμμή (perfect calibration). Αυτό σημαίνει ότι όταν το Poisson προβλέπει ένα σημείο με πιθανότητα 60%, το εν λόγω σημείο πράγματι επαληθεύεται στο 60% των περιπτώσεων. Αντίθετα, οι καμπύλες των μοντέλων LSTM και GRU αποκλίνουν σημαντικά από τη διαγώνιο, εμφανίζοντας το χαρακτηριστικό σχήμα της «υπερβολικής σιγουριάς» (overconfidence). Τα νευρωνικά δίκτυα έτειναν να δίνουν πολύ ακραίες πιθανότητες (π.χ. 85% υπέρ μιας ομάδας), οι οποίες δεν ανταποκρίνονταν στην πραγματική αβεβαιότητα του αγώνα, γεγονός που δικαιολογεί το υψηλό τους Brier Score.

(7.5) Οπτικοποίηση μέσω της Διεπαφής Χρήστης (GUI)

Η πρακτική εφαρμογή των παραπάνω μοντέλων και η άμεση σύγκριση των προβλέψεών τους ενσωματώθηκε στο Γραφικό Περιβάλλον Χρήστη (GUI) που αναπτύχθηκε με την τεχνολογία PySide6.

Σύγκριση Πραγματικών Αποτελεσμάτων vs Μοντέλων AI						
Επιλέξε Αγwνιστική:		Αγωνιστική 1				
	Αγwνας	Poisson	Dixon-Coles	LSTM	GRU	Πραγματικό Σκορ
1	Liverpool - Bournemouth	1	1	2	1	1 (Νίκη Γηπ.)
2	Aston Villa - Newcastle	1	1	1	1	X (Ισοπαλία)
3	Brighton - Fulham	1	1	2	2	X (Ισοπαλία)
4	Sunderland - West Ham	1	1	1	2	1 (Νίκη Γηπ.)
5	Tottenham - Burnley	1	1	1	1	1 (Νίκη Γηπ.)
6	Wolves - Man City	2	2	1	1	2 (Νίκη Φαλοξ.)
7	Chelsea - Crystal Palace	1	1	X	1	X (Ισοπαλία)
8	Nott'm Forest - Brentford	1	1	2	X	1 (Νίκη Γηπ.)
9	Man United - Arsenal	2	2	X	2	2 (Νίκη Φαλοξ.)
10	Leeds - Everton	1	1	1	1	1 (Νίκη Γηπ.)

Εικόνα 7.2: Το κεντρικό ταμπλό ελέγχου (Dashboard) της εφαρμογής

Στην Εικόνα 7.2 παρουσιάζεται η τελική μορφή του λογισμικού. Ο χρήστης έχει τη δυνατότητα να επιλέξει την επιθυμητή αγωνιστική εβδομάδα από το αναπτυσσόμενο μενού, ενεργοποιώντας το pipeline δεδομένων. Οι παραγόμενες πιθανότητες για κάθε αναμέτρηση προβάλλονται συγκριτικά σε μορφή πίνακα, με το σύστημα να εφαρμόζει αυτόματη χρωματική επισήμανση (color coding) στο επικρατέστερο σημείο κάθε μοντέλου. Αυτή η διαδραστική παρουσίαση επιτρέπει στον τελικό χρήστη να αντιληφθεί οπτικά την απόκλιση μεταξύ των προσεγγίσεων κλασικής στατιστικής και βαθιάς μάθησης.

8 Συμπεράσματα και Μελλοντικές Επεκτάσεις

Το παρόν κεφάλαιο συνοψίζει τα αποτελέσματα της πειραματικής διαδικασίας, αναδεικνύει τους περιορισμούς που διέπουν τη συγκεκριμένη μελέτη και παραθετεί προτάσεις για τη μελλοντική εξέλιξη του συστήματος.

(8.1) Σύνοψη Ευρημάτων: Ποιο μοντέλο υπερτερεί

Η παρούσα πτυχιακή εργασία είχε ως πρωταρχικό στόχο τη συγκριτική αξιολόγηση παραδοσιακών στατιστικών μεθόδων και σύγχρονων αρχιτεκτονικών Βαθιάς Μάθησης (Deep Learning) στην πρόβλεψη ποδοσφαιρικών αναμετρήσεων.

Τα πειραματικά αποτελέσματα, μέσω της αυστηρής αξιολόγησης Walk-Forward Validation στα δεδομένα της Premier League (2021-2026), έδωσαν μια σαφή και τεκμηριωμένη απάντηση: το παραδοσιακό στατιστικό μοντέλο της Κατανομής Poisson υπερτερεί ξεκάθαρα έναντι των υπολοίπων προσεγγίσεων. Το συγκεκριμένο εύρημα επιβεβαιώνει επιστημονικά τη φιλοσοφική αρχή του «Ξυραφιού του Όκαμ» (Occam's Razor), σύμφωνα με την οποία, σε περιβάλλοντα ακραίας αβεβαιότητας, η απλούστερη μαθηματική εξήγηση είναι συχνά και η πιο αποδοτική.

Οι πολύπλοκες αρχιτεκτονικές Νευρωνικών Δικτύων (LSTM και GRU) υστέρησαν σε προγνωστική ικανότητα. Η αδυναμία τους δεν οφείλεται σε αστοχία σχεδιασμού ή ανάπτυξης του κώδικα, αλλά στην ίδια τη στοχαστική φύση του αθλήματος. Το ποδόσφαιρο χαρακτηρίζεται από εξαιρετικά χαμηλό ποσοστό Σήματος προς Θόρυβο (low signal-to-noise ratio). Σε ένα τέτοιο περιβάλλον, οι αλγόριθμοι Deep Learning υπερπροσαρμόστηκαν (overfitting) στις τυχαίες διακυμάνσεις των στατιστικών (π.χ. παροδικά σερί από κόρνερ ή σουτ), αδυνατώντας να γενικεύσουν επιτυχώς. Αντιθέτως, το απλό μοντέλο Poisson, εστιάζοντας αποκλειστικά στους μέσους όρους των ρυθμών σκοραρίσματος, αποδείχθηκε "τυφλό" στον θόρυβο και άρα πολύ πιο ανθεκτικό.

Ακόμα και το στατιστικό μοντέλο Dixon-Coles, παρότι μαθηματικά πιο άρτιο στην πρόβλεψη ισοπαλιών χαμηλού σκορ μέσω της παραμέτρου εξάρτησης (ρ), προσέθεσε αλγοριθμικό θόρυβο, αποδεικνύοντας ότι η πλεονάζουσα πολυπλοκότητα δεν μεταφράζεται απαραίτητα σε υψηλότερη ακρίβεια.

(8.2) Περιορισμοί της έρευνας

Παρά τη μεθοδολογική αρτιότητα του συστήματος και τη χρήση ρεαλιστικών συνθηκών πρόβλεψης, η παρούσα μελέτη υπόκειται σε συγκεκριμένους ερευνητικούς περιορισμούς που πρέπει να ληφθούν υπόψη:

- **Περιορισμός Συνόλου Χαρακτηριστικών (Feature Limitations):** Το σύνολο δεδομένων που αξιοποιήθηκε (`prepared_matches.csv`) βασίζεται αποκλειστικά σε μακροσκοπικά στατιστικά στοιχεία ομάδας (συνολικά γκολ, σουτ, σουτ στον στόχο, κόρνερ). Δεν περιλαμβάνει ποιοτικούς δείκτες ή δεδομένα σε επίπεδο παικτών.
- **Περιορισμός Δείγματος (Λίγκα):** Η μελέτη επικεντρώθηκε αποκλειστικά στο πρωτάθλημα της αγγλικής Premier League. Συνεπώς, δεν μπορεί να εξαχθεί ασφαλές συμπέρασμα για το εάν η υπεροχή της κατανομής Poisson διατηρείται σε πρωταθλήματα

με διαφορετικά χαρακτηριστικά (π.χ. πρωταθλήματα χαμηλότερης κατηγορίας με μεγαλύτερη διακύμανση δυναμικότητας ή διαφορετικούς μέσους όρους γκολ).

- Απουσία Οικονομικής Αξιολόγησης: Η αξιολόγηση των μοντέλων πραγματοποιήθηκε με αυστηρά μαθηματικά και προγνωστικά κριτήρια (Accuracy, Brier Score, Log-Loss). Η εργασία δεν εξέτασε την κερδοφορία (profitability) των μοντέλων σε σχέση με τις στοιχηματικές αποδόσεις της αγοράς, διατηρώντας καθαρά ακαδημαϊκό προσανατολισμό.

(8.3) Προτάσεις για μελλοντική έρευνα

Λαμβάνοντας υπόψη τα ευρήματα και τους περιορισμούς, προτείνονται οι εξής κατευθύνσεις για τη συνέχιση και επέκταση της έρευνας από μελλοντικούς μελετητές:

- Ενσωμάτωση Μετρικών xG (Expected Goals): Η αντικατάσταση των απλών στατιστικών (σουτ/κόρνερ) με την προηγμένη μετρική των Προσδοκώμενων Γκολ (xG), η οποία αξιολογεί την ποιότητα της κάθε ευκαιρίας, ενδέχεται να αυξήσει το ωφέλιμο "σήμα" των δεδομένων, επιτρέποντας ίσως στα Νευρωνικά Δίκτυα να εντοπίσουν ουσιαστικότερα μοτίβα.
- Δεδομένα Απουσιών και Ατομικής Φόρμας: Η τροφοδότηση των μοντέλων με δεδομένα απουσιών (τραυματισμοί βασικών παικτών, κάρτες) θα μπορούσε να αντιμετωπίσει καλύτερα το φαινόμενο του Concept Drift (Αλλαγή Εννοιολογικού Πλαισίου), καθώς η απουσία ενός κομβικού παίκτη μεταβάλλει άμεσα τη δυναμική της ομάδας.
- Υβριδικά Μοντέλα (Ensemble Methods): Μια ενδιαφέρουσα ερευνητική προσέγγιση θα ήταν η δημιουργία υβριδικών μοντέλων που συνδυάζουν τις προβλέψεις της στατιστικής (Poisson) με εκείνες της βαθιάς μάθησης, επιχειρώντας να συνδυάσουν τη σταθερότητα της πρώτης με την αναγνώριση κρυφών μοτίβων της δεύτερης.
- Αρχιτεκτονικές Transformer: Σε αλγοριθμικό επίπεδο, οι μελλοντικές μελέτες θα μπορούσαν να αντικαταστήσουν τα δίκτυα RNN (LSTM/GRU) με αρχιτεκτονικές Transformer. Ο μηχανισμός Προσοχής (Attention Mechanism) των Transformers ίσως καταφέρει να ξεχωρίσει καλύτερα τα σημαντικά παιχνίδια από τον στοχαστικό θόρυβο.

(9) ΠΑΡΑΡΤΗΜΑ Α

A.1 Τεχνικοί Όροι και Ορισμοί

Στο παρόν παράρτημα παρατίθενται συνοπτικοί ορισμοί των βασικών τεχνικών όρων που χρησιμοποιούνται στην εργασία, με σκοπό τη διευκόλυνση του αναγνώστη.

Brier Score (Σκορ Brier): Μια συνάρτηση βαθμολόγησης που μετράει την ακρίβεια πιθανολογικών προβλέψεων. Υπολογίζει το μέσο τετραγωνικό σφάλμα μεταξύ προβλεπόμενων πιθανοτήτων και πραγματικών εκβάσεων. Τιμές κοντά στο 0 υποδηλώνουν τέλεια βαθμονόμηση.

Concept Drift (Αλλαγή Εννοιολογικού Πλαισίου): Στη Μηχανική Μάθηση, αναφέρεται στο φαινόμενο όπου οι στατιστικές ιδιότητες της μεταβλητής-στόχου μεταβάλλονται με την πάροδο του χρόνου με απρόβλεπτους τρόπους (π.χ. αλλαγή φόρμας μιας ομάδας στο τέλος της σεζόν), προκαλώντας μείωση της ακρίβειας του μοντέλου.

Data Leakage (Διαρροή Δεδομένων): Το σφάλμα που προκύπτει όταν πληροφορίες εκτός του συνόλου εκπαίδευσης (συνήθως από το μέλλον) χρησιμοποιούνται κατά λάθος για τη δημιουργία του μοντέλου, οδηγώντας σε πλασματικά υψηλή ακρίβεια.

GUI (Graphical User Interface - Γραφική Διεπαφή Χρήστη): Το περιβάλλον αλληλεπίδρασης μεταξύ του χρήστη και του συστήματος λογισμικού, το οποίο χρησιμοποιεί γραφικά στοιχεία (όπως παράθυρα και πίνακες) αντί για γραμμές εντολών.

JSON (JavaScript Object Notation): Μια ελαφριά, αναγνώσιμη από ανθρώπους μορφή ανταλλαγής δεδομένων. Χρησιμοποιείται ευρέως κατά την επικοινωνία με REST APIs για τη μεταφορά της πληροφορίας.

LSTM (Long Short-Term Memory): Ένας τύπος Επαναλαμβανόμενου Νευρωνικού Δικτύου (RNN) με δυνατότητα εκμάθησης μακροπρόθεσμων εξαρτήσεων, ειδικά σχεδιασμένος για την αποφυγή του προβλήματος εξαφάνισης της κλίσης.

Min-Max Scaling: Τεχνική κανονικοποίησης που μετασχηματίζει τα αριθμητικά δεδομένα ώστε να περιοριστούν αυστηρά μέσα σε ένα συγκεκριμένο εύρος, συνήθως μεταξύ 0 και 1.

Occam's Razor (Το Ξυράφι του Όκαμ): Μια φιλοσοφική και επιστημονική αρχή σύμφωνα με την οποία, ανάμεσα σε δύο εξηγήσεις ή μοντέλα που παράγουν το ίδιο αποτέλεσμα, η απλούστερη είναι συνήθως η προτιμότερη.

One-Hot Encoding: Μέθοδος μετατροπής κατηγορικών μεταβλητών σε δυαδικά διανύσματα, ώστε να μπορούν να τροφοδοτηθούν ως είσοδοι ή έξοδοι σε αλγόριθμους Μηχανικής Μάθησης (π.χ. [1,0,0] για Νίκη Γηπεδούχου).

Overfitting (Υπερπροσαρμογή): Η κατάσταση κατά την οποία ένα μοντέλο μηχανικής μάθησης "απομνημονεύει" τα δεδομένα εκπαίδευσης και τον θόρυβό τους, με αποτέλεσμα να αποτυγχάνει να γενικεύσει και να προβλέψει σωστά νέα, άγνωστα δεδομένα.

REST API (Representational State Transfer): Αρχιτεκτονικό πρότυπο για την ανάπτυξη διαδικτυακών υπηρεσιών που επιτρέπει σε διαφορετικά συστήματα να επικοινωνούν μεταξύ τους μέσω πρωτοκόλλου HTTP.

Walk-Forward Validation: Αυστηρή μέθοδος αξιολόγησης μοντέλων χρονοσειρών. Αντί για τυχαία επιλογή δεδομένων, το μοντέλο δοκιμάζεται σειριακά, προσομοιώνοντας την πραγματική ροή του χρόνου (εκπαίδευση στο παρελθόν, δοκιμή στο άμεσο μέλλον).

ΒΙΒΛΙΟΓΡΑΦΙΑ

- Baboota, R., & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. *Expert Systems with Applications*, 118, 277-286.
- Berrar, D., Lopes, P., & Dubitzky, W. (2019). Incorporating domain knowledge in machine learning for soccer outcome prediction. *Machine Learning*, 108(1), 97-126.
- Boshnakov, G. N., Kharrat, T., & McHale, I. G. (2017). A bivariate Weibull count model for forecasting association football scores. *International Journal of Forecasting*, 33(2), 458-466.
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1-3.
- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.
- Chollet, F. (2021). *Deep Learning with Python* (2nd ed.). Manning Publications.
- Constantinou, A. C., Fenton, N. E., & Neil, M. (2012). pi-football: A Bayesian network model for forecasting Association Football match outcomes. *Knowledge-Based Systems*, 36, 322-339.
- Dixon, M. J., & Coles, S. G. (1997). Modelling Association Football Scores and Inefficiencies in the Football Betting Market. *Applied Statistics*, 46(2), 265-280.
- Goddard, J. (2005). Regression models for forecasting goals and match results in association football. *International Journal of Forecasting*, 21(2), 331-340.
- Fielding, R. T. (2000). *Architectural styles and the design of network-based software architectures*. University of California, Irvine.
- Fowler, M. (2002). *Patterns of Enterprise Application Architecture*. Addison-Wesley Professional.
- Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4), 1-37.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.
- Hucaljuk, J., & Rakipović, A. (2011). Predicting football scores using machine learning techniques. In *2011 Proceedings of the 34th International Convention MIPRO* (pp. 1623-1627). IEEE.
- Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. arXiv preprint arXiv:1412.6980.
- Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.
- Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, 36(3), 109-118.
- Masse, M. (2011). *REST API Design Rulebook: Designing Consistent RESTful Web Service Interfaces*. O'Reilly Media.
- McCabe, A., & Trevathan, J. (2008). Artificial intelligence in sports prediction. In *Fifth International Conference on Information Technology: New Generations (itng 2008)* (pp. 1194-1197). IEEE.
- McKinney, W. (2012). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media, Inc.
- Moroney, M. J. (1956). *Facts from figures*. Penguin Books.
- Reis, J., & Housley, M. (2022). *Fundamentals of Data Engineering: Plan and Build Robust Data Systems*. O'Reilly Media.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1), 1929-1958.

- Summerfield, M. (2007). Rapid GUI Programming with Python and Qt: The Definitive Guide to PyQt Programming. Prentice Hall.
- Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, 16(4), 437-450.
- Tax, N., & Joutstra, Y. (2015). Predicting the Dutch football competition using public data: A machine learning approach. *Transactions on Computational Collective Intelligence XXII*, 90-111.