Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Εντοπισμός διατακτικού σε μια δικαστική απόφαση με χρήση τεχνικών μηχανικής μάθησης

Διπλωματικη Εργασια

του

ΘΕΟΔΩΡΟΥ ΑΓΓΕΛΟΥ ΜΕΞΗ

Επιβλέπων: Παναγιώτης Τσανάκας
Καθηγητής Ε.Μ.Π.

Συνεπιβλέπων: Μάριος Κόνιαρης
Ε.ΔΙ.Π Ε.Μ.Π.

Αθήνα, Μάρτιος 2025

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Εντοπισμός διατακτικού σε μια δικαστική απόφαση με χρήση τεχνικών μηχανικής μάθησης

Διπλωματικη Εργασια

του

ΘΕΟΔΩΡΟΥ ΑΓΓΕΛΟΥ ΜΕΞΗ

Επιβλέπων: Παναγιώτης Τσανάκας
Καθηγητής Ε.Μ.Π.

Συνεπιβλέπων: Μάριος Κόνιαρης
Ε.ΔΙ.Π Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 14η Μαρτίου 2025.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Παναγιώτης Τσανάκας	Ανδρέας-Γεώργιος Σταφυλοπάτης	Ευγενία Τζαννίνη
Καθηγητής Ε.Μ.Π.	Ομότιμος Καθηγητής Ε.Μ.Π.	Επίκουρη Καθηγήτρια Ε.Μ.Π.

Αθήνα, Μάρτιος 2025

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα. Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

(Υπογραφή)

.....
Θεόδωρος Άγγελος Μέξης
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

14 Μαρτίου 2025

Περίληψη

Στην παρούσα διπλωματική εργασία διερευνάται το ζήτημα της πρόβλεψης δικαστικών αποφάσεων μέσω τεχνικών μηχανικής μάθησης και επεξεργασίας φυσικής γλώσσας (NLP). Αρχικά, πραγματοποιήθηκε εκτενής επισκόπηση της σχετικής βιβλιογραφίας, καθώς και συλλογή και ανάλυση ενός συνόλου δεδομένων που δημιουργήθηκε στο πλαίσιο της έρευνας, το οποίο περιλαμβάνει αποφάσεις του Αρείου Πάγου και του Εφετείου Πειραιώς. Στη συνέχεια, εφαρμόστηκαν διαδικασίες προεπεξεργασίας κειμένου, όπως καθαρισμός δεδομένων, μετατροπή τους σε αριθμητικές αναπαραστάσεις με την τεχνική TF-IDF καθώς και επισημείωση των αποφάσεων ως *αποδοχή* ή *απόρριψη*.

Ακολούθως, προσδιορίστηκε και περιγράφηκε αναλυτικά η μεθοδολογία εκπαίδευσης μοντέλων μηχανικής μάθησης για την πρόβλεψη δικαστικών αποφάσεων, ενώ παράλληλα εξετάστηκε η διαδικασία βελτιστοποίησης των υπερπαραμέτρων τους. Στο πειραματικό μέρος, πραγματοποιήθηκε εκτενής αξιολόγηση διαφόρων μοντέλων ταξινόμησης, αναδεικνύοντας τη σημασία της ποιοτικής προεπεξεργασίας των δεδομένων και της σωστής επιλογής υπερπαραμέτρων για τη βελτιστοποίηση της απόδοσης. Τα αποτελέσματα των πειραμάτων κατέδειξαν τον Random Forest και τον XGBoost ως τους πιο αποδοτικούς ταξινομητές, καθώς παρουσίασαν υψηλές τιμές Precision, Recall και F1-Score, διασφαλίζοντας σταθερότητα και αξιοπιστία στις προβλέψεις.

Λέξεις Κλειδιά

Δικαστικές αποφάσεις, Τεχνητή Νοημοσύνη, Μηχανική Μάθηση, NLP, Decision Trees, SVM, XGBoost, Logistic Regression, Random Forest

Abstract

This thesis explores the issue of judicial decision prediction using machine learning and natural language processing (NLP) techniques. Initially, an extensive review of the relevant literature was conducted, together with the collection and analysis of a data set created within the framework of this research, which includes decisions of the Supreme Court of Greece and the Court of Appeals of Piraeus. Text preprocessing procedures were applied, including data cleaning, transformation into numerical representations using the TF-IDF machine learning tool, as well as annotation of decisions.

Subsequently, the methodology for training machine learning models to predict court decisions was identified and analyzed in detail, while the process of optimizing their hyperparameters was also examined. In the experimental section, an extensive evaluation of five classification models was conducted, highlighting the importance of high-quality data preprocessing and the proper selection of hyperparameters for performance optimization. The experimental results demonstrated that Random Forest and XGBoost were the most effective classifiers, as they exhibited high Precision, Recall, and F1-Score values, ensuring stability and reliability in predictions.

Keywords

Judicial Decisions, Artificial Intelligence, Machine Learning, NLP, Decision Trees, SVM, XGBoost, Logistic Regression, Random Forest

στου αδελφό μου Κωνσταντίνο

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου Παναγιώτη Τσανάκα για την ευκαιρία που μου δόθηκε να εργαστώ στο εξαιρετικά ενδιαφέρον θέμα της διπλωματικής μου εργασίας. Οφείλω ένα τεράστιο ευχαριστώ στον Δρ. Μάριο Κόνιαρη, ο οποίος μου προσέφερε πραγματικά, κάθε δυνατή βοήθεια κατά την εκπόνηση της διπλωματικής μου εργασίας. Θα ήθελα επίσης να ευχαριστήσω την οικογένειά μου, τους φίλους και συμφοιτητές μου, που ήταν πλάι μου σε όλη την διάρκεια της φοιτητικής μου ζωής. Αφιερώνω την εργασία αυτή στον αδελφό μου Κωνσταντίνο.

Αθήνα, Μάρτιος 2025

Θεόδωρος Άγγελος Μέξης
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Η νομική επιστήμη βρίσκεται σε συνεχή διάλογο με τις κοινωνικές και τεχνολογικές εξελίξεις, επιδιώκοντας να ανταποκριθεί στις πολυδιάστατες ανάγκες της σύγχρονης κοινωνίας. Ωστόσο, η αυξανόμενη πολυπλοκότητα της νομοθεσίας και η διαρκής συσσώρευση δικαστικών αποφάσεων δημιουργούν σημαντικές προκλήσεις για τους νομικούς επαγγελματίες, όπως η διαχείριση μεγάλων όγκων δεδομένων, η εξασφάλιση συνέπειας και η αποφυγή σφαλμάτων. Σε αυτό το πλαίσιο, η εισαγωγή μεθόδων τεχνητής νοημοσύνης και ανάλυσης κειμένων στον νομικό τομέα προσφέρει τη δυνατότητα βελτίωσης της αποδοτικότητας και της ακρίβειας στις διαδικασίες λήψης αποφάσεων. Η παρούσα διπλωματική εργασία επικεντρώνεται στην πρόβλεψη δικαστικών αποφάσεων, αξιοποιώντας τεχνικές επεξεργασίας φυσικής γλώσσας, αναδεικνύοντας τις προκλήσεις και τις ευκαιρίες της εφαρμογής αυτών των μεθόδων στον τομέα της δικαιοσύνης.

1.1 Νομική Επιστήμη

1.1.1 Νομικά Κείμενα

Τα νομικά κείμενα είναι κείμενα τα οποία έχουν συνταχθεί για διάφορους σκοπούς, με βασικό τους χαρακτηριστικό ότι σχετίζονται με τον νόμο είτε λόγω της ιδιότητας του συντάκτη (π.χ δικαστής), είτε λόγω των αναφορών που περιέχουν σε άλλα νομικά κείμενα, ή λόγω της θεματολογίας τους που αφορά την ρύθμιση των δικαιωμάτων και των υποχρεώσεων ιδιωτών και θεσμών. Ορισμένοι από τους βασικούς τύπους νομικών κειμένων, ιεραρχημένοι βάσει της ισχύος τους ως πηγών δικαίου, είναι οι παρακάτω :

1. Συντάγματα : τα οποία αποτελούν τις θεμελιώδεις ραχές που αφορούν πως διοικείται ένα κράτος.

2. Νομοθεσία : η οποία περιλαμβάνει τους νόμους τους οποίους θεσπίζει ένα νομοθετικό σώμα και ρυθμίζουν τι είναι επιτρεπτό από τον νόμο. Συνήθως, οι νόμοι οργανώνονται θεματικά σε κώδικες παρόμοιας θεματολογίας (π.χ Ποινικός Κώδικας).

3. Δικαστικές Αποφάσεις : οι οποίες περιλαμβάνουν τα ενδιάμεσα ή τελικά αποτελέσματα μιας δίκης, την κρίση του δικαστηρίου για τα γεγονότα και τα επιχειρήματα της κάθε πλευράς καθώς και την απόφαση του δικαστηρίου σε γραπτή μορφή.

4. Συμβόλαια : τα οποία συνιστούν αμοιβαίες συμφωνίες μεταξύ συμβαλλόμενων μερών, με σκοπό να τηρηθούν αμοιβαίες υποχρεώσεις.

Τα νομικά κείμενα έχουν εξελιχθεί σημαντικά μέσα στις χιλιετίες που παρήλθαν από την πρώτη χρήση τους. Τα πρώτα κείμενα ιδιωτικού δικαίου ήταν συμβόλαια, διαθήκες και νομικές πράξεις εγγεγραμμένες σε πινακίδες από πηλό στην Σουμερία περίπου 5000 χρόνια πριν. Παρομοίως, τα πρώτα κείμενα δημοσίου δικαίου, όπως νόμοι, εμφανίστηκαν στην Μεσοποταμία με τους νόμους του βασιλιά Ουρ Ναμμού και αργότερα τον κώδικα του Χαμουραμπί να αποτελούν γνωστά παραδείγματα.

Στην Αρχαία Ελλάδα, οι νομικές ρυθμίσεις που αφορούσαν ιδιωτικές υποθέσεις, όπως θέματα κληρονομιάς, εμπορίου και συμβολαίων, διακρίνονταν από τις διατάξεις που ρύθμιζαν τη ζωή των πολιτών σε κάθε πόλη-κράτος. Η νομοθεσία και οι νόρμες διέφεραν ανά περιοχή, κυρίως λόγω της επιρροής της ρητορικής τέχνης, της διάδοσης της γνώσης των νόμων και της λειτουργίας των δικαστηρίων, καθώς και άλλων κοινωνικοπολιτικών παραμέτρων που καθόριζαν την ιδιότητα του πολίτη. Εμβληματικά παραδείγματα περιλαμβάνουν τους αυστηρούς νόμους του Δράκοντα (620 π.Χ.), οι οποίοι αργότερα μεταρρυθμίστηκαν από τον Σόλωνα (593 π.Χ.), προσδίδοντας μια πιο ισορροπημένη προσέγγιση στη νομοθεσία [1].

1.1.2 Δικαστήρια

Το **δικαστήριο** αποτελεί θεσμοθετημένο όργανο της δικαστικής εξουσίας, το οποίο απονέμει δικαιοσύνη στο πλαίσιο του κράτους δικαίου. Η λειτουργία του βασίζεται σε νομικά πρότυπα, τα οποία διασφαλίζουν την επίλυση διαφορών, την επιβολή ποινών για παραβάσεις του νόμου και τη διατήρηση της κοινωνικής συνοχής. Ως βασικός πυλώνας του δικαιοκρατικού συστήματος, το δικαστήριο ενσαρκώνει την ανεξαρτησία της δικαστικής εξουσίας και προστατεύει τα θεμελιώδη δικαιώματα και τις ελευθερίες των πολιτών [2]. Τα δικαστήρια κατηγοριοποιούνται με βάση τη δικαιοδοσία τους, τη φύση των υποθέσεων που εξετάζουν και τη θέση τους στην ιεραρχική δομή του δικαστικού συστήματος.

1. Ανάλογα με τη φύση των υποθέσεων

- Πολιτικά Δικαστήρια: Εξετάζουν διαφορές που προκύπτουν από τις σχέσεις μεταξύ ιδιωτών ή νομικών προσώπων, όπως διαφορές συμβατικού δικαίου, οικογενειακές υποθέσεις και ζητήματα περιουσιακού δικαίου.

- Ποινικά Δικαστήρια: Ασχολούνται με την εκδίκαση αδικημάτων που προβλέπονται στον ποινικό κώδικα, όπως εγκλήματα και πλημμελήματα.

- Διοικητικά Δικαστήρια: Διαχειρίζονται διαφορές μεταξύ πολιτών και της δημόσιας διοίκησης, όπως φορολογικές ενστάσεις ή προσφυγές κατά διοικητικών πράξεων.

- Συνταγματικά Δικαστήρια: Εξετάζουν ζητήματα συνταγματικότητας νόμων ή κρατικών ενεργειών.

2. Ανάλογα με την Ιεραρχία

- Κατώτερα Δικαστήρια: Αποτελούν το πρώτο επίπεδο απονομής της δικαιοσύνης, ασχολούμενα με υποθέσεις μικρότερης σημασίας, όπως τα Ειρηνοδικεία.
- Ανώτερα Δικαστήρια: Εκδικάζουν εφέσεις κατά αποφάσεων κατώτερων δικαστηρίων, όπως τα Εφετεία.
- Ανώτατα Δικαστήρια: Αποτελούν την κορυφή της ιεραρχίας και εξετάζουν σοβαρά νομικά ζητήματα, όπως ο Άρειος Πάγος και το Συμβούλιο της Επικρατείας.

3. Ειδικά Δικαστήρια

Ορισμένα δικαστήρια έχουν εξειδικευμένη δικαιοδοσία:

- Στρατιωτικά Δικαστήρια: Εξετάζουν υποθέσεις στρατιωτικού δικαίου.
- Εμπορικά Δικαστήρια: Αντιμετωπίζουν διαφορές που σχετίζονται με το εμπορικό δίκαιο, όπως οι πτωχεύσεις.
- Εργατικά Δικαστήρια: Διαχειρίζονται διαφορές εργοδότη και εργαζομένου.
- Διεθνή Δικαστήρια: Ρυθμίζουν διακρατικές διαφορές ή ζητήματα διεθνούς δικαίου.

Τα δικαστήρια αποτελούν τον θεματοφύλακα της νομιμότητας και της κοινωνικής δικαιοσύνης. Με την επίλυση νομικών διαφορών, διασφαλίζουν την ισονομία, προστατεύουν τα ανθρώπινα δικαιώματα και ενισχύουν την εμπιστοσύνη των πολιτών στο κράτος δικαίου. Παράλληλα, μέσα από τη νομολογία, διαμορφώνουν την ερμηνεία του δικαίου και τη λειτουργία της κοινωνίας συνολικά.

1.2 Νομική Πληροφορική

Η νομική πληροφορική είναι ένα διεπιστημονικό πεδίο που συνδυάζει τη νομική επιστήμη με τις τεχνολογίες πληροφορικής, με στόχο την ανάλυση, την οργάνωση, και την αυτοματοποίηση νομικών διαδικασιών. Πρόκειται για έναν ταχύτατα εξελισσόμενο τομέα, που επικεντρώνεται στη χρήση εργαλείων τεχνητής νοημοσύνης, ανάλυσης δεδομένων και μηχανικής μάθησης για τη διευκόλυνση της πρόσβασης στη δικαιοσύνη, τη βελτίωση της αποτελεσματικότητας της νομικής έρευνας και τη λήψη τεκμηριωμένων αποφάσεων.

Παρά την πρόοδο που έχει σημειωθεί στον τομέα της νομικής πληροφορικής, εξακολουθούν να υπάρχουν σημαντικές προκλήσεις που περιορίζουν την ευρεία υιοθέτηση και την αποτελεσματική εφαρμογή της. Μία από τις κυριότερες δυσκολίες είναι η μη δομημένη φύση των δεδομένων. Τα νομικά δεδομένα, όπως αποφάσεις δικαστηρίων, νομοθεσίες και συμβάσεις, είναι συχνά μη δομημένα και περιλαμβάνουν μεγάλες ποσότητες κειμένου. Αυτό καθιστά την ανάλυσή τους δύσκολη χωρίς τη χρήση εξειδικευμένων εργαλείων και τεχνολογιών. Επιπλέον, η απουσία τυποποιημένων φορμά δεδομένων περιπλέκει περαιτέρω τη διαδικασία επεξεργασίας και ενοποίησης.

Η γλωσσική ανάλυση στον χώρο της νομικής πληροφορικής αποτελεί επίσης μια περίπλοκη διαδικασία. Η νομική γλώσσα χαρακτηρίζεται από εξειδικευμένη ορολογία, περίπλοκη σύνταξη και συχνά αμφίσημη διατύπωση, η οποία εξαρτάται από το εκάστοτε πλαίσιο. Αυτά τα χαρακτηριστικά καθιστούν δύσκολη την αυτοματοποίηση διαδικασιών όπως η κατηγοριοποίηση δικαστικών αποφάσεων ή η πρόβλεψη εκβάσεων.

Επιπλέον, η προστασία προσωπικών δεδομένων αποτελεί κρίσιμη πρόκληση στον τομέα. Τα νομικά δεδομένα συχνά περιλαμβάνουν ευαίσθητες πληροφορίες, όπως ονόματα και διευθύνσεις. Η επεξεργασία τους πρέπει να συμμορφώνεται με αυστηρά κανονιστικά πλαίσια, όπως ο Γενικός Κανονισμός Προστασίας Δεδομένων (GDPR), γεγονός που περιορίζει τις διαθέσιμες πηγές δεδομένων ή απαιτεί διαδικασίες ανωνυμοποίησης.

Ένα άλλο πρόβλημα είναι η δυσκολία γενίκευσης των λύσεων που αναπτύσσονται. Ο τομέας της νομικής πληροφορικής πρέπει να λάβει υπόψη την ποικιλομορφία που υπάρχει μεταξύ διαφορετικών νομικών συστημάτων, δικαστηρίων και διαδικασιών. Αυτό καθιστά ιδιαίτερα δύσκολη τη δημιουργία γενικεύσιμων λύσεων που μπορούν να εφαρμοστούν σε πολλαπλά περιβάλλοντα.

Τέλος, η χρηστικότητα και η αποδοχή των εργαλείων πληροφορικής από τους νομικούς επαγγελματίες παραμένουν περιορισμένες. Πολλοί εμφανίζονται επιφυλακτικοί στη χρήση αυτοματοποιημένων συστημάτων, κυρίως λόγω έλλειψης εμπιστοσύνης στην ακρίβεια και την αξιοπιστία τους. Επιπλέον, υπάρχει ανησυχία ότι η χρήση τεχνητής νοημοσύνης μπορεί να υποκαταστήσει την ανθρώπινη κρίση, γεγονός που δημιουργεί αντιστάσεις στην υιοθέτηση τέτοιων εργαλείων.

1.3 Ορισμός προβλήματος της διπλωματικής εργασίας

Η αυτόματη ανάλυση νομικών εγγράφων αποτελεί μια κρίσιμη πρόκληση στη σύγχρονη νομική έρευνα και πρακτική, καθώς μπορεί να συμβάλει στην καλύτερη κατανόηση των νομικών αποφάσεων και στη βελτίωση της αποδοτικότητας των νομικών επαγγελματιών. Μία από τις πιο συχνές προσεγγίσεις στο πεδίο της επεξεργασίας φυσικής γλώσσας (NLP) είναι η χρήση τεχνικών μηχανικής μάθησης για την πρόβλεψη των δικαστικών αποφάσεων.

Ένα βασικό εμπόδιο στην ανάπτυξη τέτοιων συστημάτων είναι η πολυπλοκότητα της νομικής γλώσσας και η ανάγκη για ακριβή επεξεργασία των κειμένων. Επιπλέον, η υψηλή διακύμανση στα δεδομένα στις δικαστικές αποφάσεις αποτελεί σημαντικό παράγοντα που πρέπει να ληφθεί υπόψη κατά τη σχεδίαση των αλγορίθμων ταξινόμησης. Στόχος της εργασίας είναι να αναλύσει την απόδοση διαφορετικών τεχνικών NLP και μηχανικής μάθησης στην κατηγοριοποίηση των δικαστικών αποφάσεων, καθώς και να αξιολογήσει πώς τα αποτελέσματα μπορούν να αξιοποιηθούν στην πράξη από τους νομικούς επαγγελματίες.

Η παρούσα διπλωματική εργασία επικεντρώνεται σε ορισμένα από τα παραπάνω ζητήματα, επικετρώνοντας το ενδιαφέρον της στην πρόβλεψη δικαστικών αποφάσεων. Στόχος είναι να συμβάλει στη βελτίωση της διαχείρισης και αξιοποίησης νομικών κειμένων, προτείνοντας λύσεις που ενισχύουν την αποδοτικότητα και τη διαφάνεια στη δικαστική διαδικασία. Ως εκ τούτου, εστιάζει στην ταξινόμηση δικαστικών αποφάσεων με στόχο να προσδιορίσει αν μια απόφαση απορρίπτει ή αποδέχεται μια αίτηση. Η διαδικασία αυτή μπορεί να αποδειχθεί ιδιαίτερα χρήσιμη για τους νομικούς επαγγελματίες, καθώς μπορεί να τους επιτρέψει να

αναλύσουν πρότυπα στις αποφάσεις των δικαστηρίων, αλλά και να εντοπίσουν παράγοντες που συσχετίζονται με την έκβαση μιας υπόθεσης.

1.4 Συνεισφορά της διπλωματικής εργασίας

Η παρούσα εργασία αποσκοπεί στην ανάπτυξη ενός συστήματος πρόβλεψης δικαστικών αποφάσεων, το οποίο θα μπορούσε να αξιοποιηθεί σε ποικίλα πεδία.

Στο πλαίσιο της παρούσας διπλωματικής εργασίας, δημουργήθηκε ένα μοντέλο με στόχο την πρόβλεψη δικαστικών αποφάσεων, χρησιμοποιώντας ένα σύνολο δεδομένων που δημιουργήθηκε για το σκοπό αυτό. Αρχικά, συλλέχθηκαν και επισημειώθηκαν -όπως αναλύθηκε στο Κεφάλαιο 2- δικαστικές αποφάσεις, ταξινομώντας τις σε δύο κατηγορίες: **Αποδοχή** (0) και **Απόρριψη** (1), βάσει της τελικής έκβασης της κάθε υπόθεσης. Στη συνέχεια, πραγματοποιήθηκε η διαδικασία της προεπεξεργασίας των κειμένων, η οποία περιλάμβανε την αφαίρεση ειδικών χαρακτήρων κ.ο.κ, τη μετατροπή κειμένων σε αριθμητική αναπαράσταση (TF-IDF), ώστε τα δεδομένα να είναι κατάλληλα για την εφαρμογή αλγορίθμων μηχανικής μάθησης. Το πείραμα επικεντρώθηκε στη σύγκριση της απόδοσης πέντε διαφορετικών ταξινομητών: **Decision Trees (DT)**, **Random Forest (RF)**, **Support Vector Machines (SVM)**, **XGBoost (XGB)** και **Logistic Regression (LR)**. Οι ταξινομητές εκπαιδεύτηκαν και αξιολογήθηκαν με βάση την απόδοσή τους στο σύνολο δοκιμής (*test set*), με στόχο να διερευνηθεί η ικανότητά τους να προβλέπουν την έκβαση των δικαστικών αποφάσεων. Η απόδοσή τους μετρήθηκε με τους εξής δείκτες : Accuracy, Recall, Precision και F1-score. Όπως επισημαίνεται από τη μελέτη [3], η διαδικασία αυτή συχνά αναφέρεται εσφαλμένα ως *πρόβλεψη*, ενώ στην πραγματικότητα πρόκειται για κατηγοριοποίηση δικαστικών αποφάσεων βάσει των αποτελεσμάτων τους. Ωστόσο είναι ιδιαίτερα σημαντικό να αναφέρουμε ότι ένα τέτοιο σύστημα θα μπορούσε να αποτελέσει εργαλείο υποστήριξης για νομικούς επαγγελματίες, προσφέροντας μια προκαταρκτική εκτίμηση για την πιθανή έκβαση μιας υπόθεσης, βασισμένη σε προηγούμενες δικαστικές αποφάσεις. Ως εκ τούτου, η εν λόγω κατηγοριοποίηση δύναται να λειτουργήσει ως πρόβλεψη στα χέρια των νομικών επαγγελματιών. Επιπλέον, η εργασία αυτή μπορεί να συμβάλει στην καλύτερη κατανόηση των παραγόντων που επηρεάζουν την έκβαση μιας δικαστικής απόφασης, επιτρέποντας τη διεξαγωγή ποσοτικών αναλύσεων σε μεγάλης κλίμακας δεδομένα. Παράλληλα, μπορεί να λειτουργήσει ως σημείο εκκίνησης για τη δημιουργία εργαλείων που ενισχύουν την πρόσβαση στη δικαιοσύνη, παρέχοντας στους πολίτες πληροφορίες για τις πιθανότητες επιτυχίας μιας υπόθεσης πριν την υποβολή της.

Τέλος, η έρευνα αυτή έχει ακαδημαϊκό ενδιαφέρον, καθώς συμβάλλει στη διερεύνηση της εφαρμογής και της αποτελεσματικότητας διαφορετικών αλγορίθμων μηχανικής μάθησης στον τομέα της Νομικής.

1.5 Οργάνωση κειμένου

Στο Κεφάλαιο 2 αναλύεται το θεωρητικό υπόβαθρο των τεχνολογιών που χρησιμοποιούνται, καθώς και σχετικές επιστημονικές εργασίες. Το Κεφάλαιο 3 περιγράφει το σύνολο δεδομένων, τη διαδικασία προεπεξεργασίας των κειμένων και τη μεθοδολογία επισημείωσης.

Στο Κεφάλαιο 4 εστιάζω στη μεθοδολογία εκπαίδευσης των μοντέλων και στις τεχνικές αξιολόγησης. Στο Κεφάλαιο 5 παρουσιάζονται τα πειραματικά αποτελέσματα για την αξιολόγηση της απόδοσης. Τέλος, στο Κεφάλαιο 6 αναφέρονται τα συμπεράσματα της διπλωματικής εργασίας, αλλά και προτάσεις για μελλοντική δουλειά στον τομέα της νομικής πληροφορικής.

Κεφάλαιο 2

Θεωρητικό υπόβαθρο

Στο κεφάλαιο αυτό αρχικά παρουσιάζονται σχετικές εργασίες που μελετήθηκαν στα πλαίσια της παρούσας διπλωματικής εργασίας. Στη συνέχεια αναλύονται οι τεχνικές πρόβλεψης, περιγράφοντας τη διαδικασία παραγωγής τους, η οποία περιλαμβάνει τον καθορισμό προβλήματος, τη συλλογή και προεπεξεργασία δεδομένων, την επιλογή μεθόδων και την αξιολόγηση των μοντέλων. Ακολούθως, εξετάζεται η μηχανική μάθηση ως κεντρικό εργαλείο αυτών, αναλύοντας διαφορετικές προσεγγίσεις όπως η επιβλεπόμενη, μη επιβλεπόμενη και ενισχυτική μάθηση. Τέλος, περιγράφεται η λειτουργία και τα πλεονεκτήματα συγκεκριμένων μοντέλων μηχανικής μάθησης.

2.1 Σχετικές Εργασίες

Στην παρούσα ενότητα παρουσιάζονται σημαντικές επιστημονικές εργασίες που μελετήθηκαν στο πλαίσιο αυτής της διπλωματικής εργασίας. Η ανασκόπηση καλύπτει τόσο μελέτες που επικεντρώνονται άμεσα στην πρόβλεψη δικαστικών αποφάσεων, όσο και εργασίες που εντάσσονται στο ευρύτερο πεδίο της νομικής πληροφορικής. Στόχος είναι η ανάδειξη των κύριων μεθοδολογικών προσεγγίσεων και των τεχνολογικών εργαλείων που έχουν χρησιμοποιηθεί στη σχετική βιβλιογραφία. Οι εργασίες αυτές υπογραμμίζουν την ανάγκη για πολυδιάστατη ανάλυση των νομικών κειμένων, την ενσωμάτωση καινοτόμων τεχνολογιών και την αξιοποίηση δεδομένων για τη διαμόρφωση προβλέψεων που μπορούν να συμβάλουν στη διαφάνεια και την αποδοτικότητα της δικαιοσύνης. Όσον αφορά στις εργασίες πρόβλεψης δικαστικών αποφάσεων, στο [3] προτείνεται η κάτωθι κατηγοριοποίηση για την μελέτη των εργασιών πρόβλεψης δικαστικών αποφάσεων:

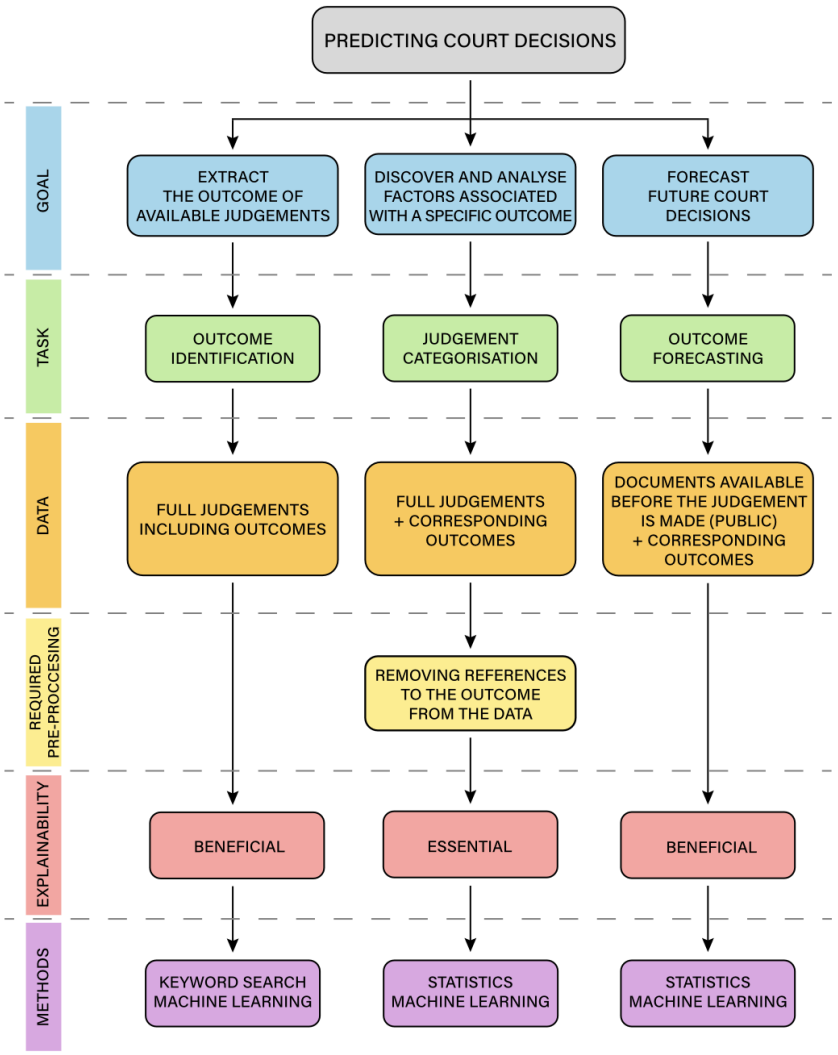
1. *Εντοπισμός ετυμηγορίας δικαστικής απόφασης* : οι εργασίες αναγνώρισης απόφασης δικαστικών αποφάσεων εστιάζουν στην εξαγωγή του αποτελέσματος από το κείμενο της δικαστικής απόφασης, όπου η ετυμυγορία υπάρχει στο κείμενο και το μοντέλο καλείται να την επισημάνει. Στην υποενότητα αυτή, θα παρουσιάσουμε εργασίες που ασχολήθηκαν με το έργο της αναγνώρισης του αποτελέσματος.

2. *Κατηγοριοποίηση αποφάσεων βάσει αποτελέσματος* : η κατηγοριοποίηση δικαστικών αποφάσεων βάσει της έκβασής τους ορίζεται ως η διαδικασία ταξινόμησης των δικαστικών αποφάσεων με βάση το αποτέλεσμά τους, χρησιμοποιώντας το κείμενο ή οποιαδήποτε άλλη πληροφορία δημοσιεύεται με την τελική απόφαση, αλλά αφαιρώντας την

ετυμηγορία της απόφασης.

3. *Πρόβλεψη δικαστικών αποφάσεων* : οι εργασίες πρόβλεψης δικαστικών αποφάσεων σχετίζονται με ένα ευρύ φάσμα ερευνών που συνδυάζουν την επεξεργασία φυσικής γλώσσας (NLP) και τη μηχανική μάθηση για την ανάλυση και πρόβλεψη δικαστικών αποφάσεων.

Η παραπάνω διάκριση των εργασιών αναπαρίσταται στο παρακάτω σχήμα 2.1, το οποίο προέρχεται από την εργασία [3], χωρίς τροποποιήσεις.



Σχήμα 2.1: *Διάγραμμα ροής που απεικονίζει τους στόχους και τις απαιτήσεις για τις τρεις εργασίες πρόβλεψης δικαστικών αποφάσεων*

2.1.1 Εντοπισμός ετυμηγορίας δικαστικής απόφασης

Οι εργασίες αναγνώρισης απόφασης δικαστικών αποφάσεων εστιάζουν στην εξαγωγή του αποτελέσματος από το κείμενο της δικαστικής απόφασης, όπου η ετυμυγορία υπάρχει στο κείμενο και το μοντέλο καλείται να την επισημάνει. Στην υποενότητα αυτή, θα παρουσιάσουμε εργασίες που ασχολήθηκαν με το έργο της αναγνώρισης του αποτελέσματος.

Μία από τις πρώτες μελέτες που προσπάθησε να προβλέψει δικαστικές αποφάσεις χρησιμοποιώντας το κείμενο της απόφασης είναι η [4]. Οι συγγραφείς χρησιμοποίησαν έναν δημοφιλή αλγόριθμο μηχανικής μάθησης, τονφ Support Vector Machine (SVM), για να προβλέψουν τις αποφάσεις του European Court of Human Rights (ECtHR). Το μοντέλο τους στόχευε στην πρόβλεψη της απόφασης του δικαστηρίου εξάγοντας τις διαθέσιμες πληροφορίες από σχετικές ενότητες των αποφάσεων του ECtHR και πέτυχε μέση ακρίβεια 79% για τρία ξεχωριστά άρθρα της ECHR. Αν και οι συγγραφείς απέκλεισαν την ίδια την ετυμηγορία (ή ολόκληρη την ενότητα που την περιείχε), εξακολουθούσαν να χρησιμοποιούν το υπόλοιπο κείμενο των αποφάσεων, το οποίο συχνά περιείχε συγκεκριμένες αναφορές στο τελικό αποτέλεσμα (π.χ. ¨Συνεπώς, υπάρχει παραβίαση του Άρθρου 3¨). Έτσι, παρόλο που η εργασία τους παρουσιάστηκε ως πρόβλεψη του αποτελέσματος δικαστικών υποθέσεων, η πραγματική της φύση περιοριζόταν στην αναγνώριση του αποτελέσματος.

Άλλες μελέτες που επικεντρώθηκαν στο ECtHR περιλαμβάνουν τις [5], [6] και [7]. Δεδομένου ότι στις [5] και [6] χρησιμοποίησαν το ίδιο σύνολο δεδομένων με τους [4], πραγματοποίησαν επίσης την ίδια εργασία αναγνώρισης αποτελέσματος. Στην εργασία [8] χρησιμοποίησαν παρόμοιες στατιστικές μεθόδους με την [4] και πέτυχαν ακρίβεια 88% με SVM, ενώ η [6] πέτυχε ακρίβεια 79% χρησιμοποιώντας ένα σύνολο μοντέλων SVM. Αντίθετα, η μελέτη [7] δημιούργησε ένα μεγαλύτερο σύνολο δεδομένων για το ίδιο δικαστήριο και με δυαδική ταξινόμηση (παραβίαση οποιουδήποτε άρθρου της ECHR έναντι μη παραβίασης) χρησιμοποίησε νευρωνικά μοντέλα, δίχως να αποκλείσει κάποιο τμήμα της απόφασης. Έτσι, η εργασία περιορίστηκε κι αυτή στην αναγνώριση του αποτελέσματος, με πολύ υψηλή ακρίβεια 96% χρησιμοποιώντας διάφορες στατιστικές μεθόδους. Αυτές οι μελέτες δείχνουν ότι η αυτόματη αναγνώριση του αποτελέσματος είναι σε μεγάλο βαθμό εφικτή για το ECtHR.

Οι μελέτες που βασίζονται στο ECtHR καταδεικνύουν δύο ευρείες κατηγορίες εργασιών που αποσκοπούν στην πρόβλεψη δικαστικών αποφάσεων, αλλά στην πραγματικότητα εκτελούν αναγνώριση αποτελέσματος.

Η πρώτη κατηγορία περιλαμβάνει μελέτες που ήταν μερικώς επιτυχημένες στην αφαίρεση των πληροφοριών που σχετίζονται με την ετυμηγορία. Εκτός από τις προαναφερθείσες μελέτες [4], [8] και [6], και η εργασία [9] αντιμετωπίζει το ίδιο πρόβλημα. Πιο συγκεκριμένα, στην εργασία αυτή, οι μελετητές επικεντρώθηκαν στο French Court of Cassation και πέτυχαν ακρίβεια έως 96%. Αν και αντικατέστησαν τις λέξεις που περιείχαν την ετυμηγορία, πολλές από τις λέξεις που αποδείχθηκαν σημαντικές για την πρόβλεψη του μοντέλου τους σχετίζονταν άμεσα με την περιγραφή του αποτελέσματος. Συνεπώς, δεν ήταν απόλυτα επιτυχημένοι στην απομάκρυνση των πληροφοριών που υποδήλωναν την ετυμηγορία.

Η δεύτερη κατηγορία περιλαμβάνει μελέτες που δεν φιλτράρουν καθόλου τις πληροφορίες της απόφασης, όπως οι [7]. Στη μελέτη [10], οι ερευνητές δηλώνουν ρητά ότι δεν αφαίρεσαν την πραγματική ετυμηγορία του Supreme Court of the Philippines (λόγω έλλειψης συνεκτικής δομής στις περιγραφές των αποφάσεων) κατά την πρόβλεψη των αποφάσεων του δικαστηρίου. Ωστόσο, η ακρίβειά τους ήταν σχετικά χαμηλή, μόλις 59%. Επιπλέον, υπάρχουν αρκετές μελέτες που δεν προσδιορίζουν καμία προεπεξεργασία για την αφαίρεση πληροφοριών που μπορεί να περιέχουν την ετυμηγορία. Παραδείγματα αποτελούν η εργασία [11], που εξετάζει υποθέσεις έφεσης από τα δικαστήρια της Βραζιλίας (με F1-score 79%), και η [12] που ασχολείται με υποθέσεις ανθρωποκτονίας δευτέρου βαθμού και διαφθοράς

που εκδικάστηκαν στο Court of São Paulo (με F1-score έως 98%).

2.1.2 Κατηγοριοποίηση αποφάσεων βάσει αποτελέσματος

Η κατηγοριοποίηση δικαστικών αποφάσεων βάσει της έκβασής τους ορίζεται ως η διαδικασία ταξινόμησης των δικαστικών αποφάσεων με βάση το αποτέλεσμά τους, χρησιμοποιώντας το κείμενο ή οποιαδήποτε άλλη πληροφορία δημοσιεύεται με την τελική απόφαση, αλλά αφαιρώντας την ετυμηγορία της απόφασης. Δεδομένου ότι τα αποτελέσματα τέτοιων υποθέσεων δημοσιεύονται και δεν χρειάζεται πλέον να ¨προβλεφθούν¨, αυτό το έργο είναι κυρίως χρήσιμο για την αναγνώριση παραγόντων (γεγονότων, επιχειρημάτων, δικαστών) που επηρεάζουν τις δικαστικές αποφάσεις μέσα στο κείμενο των αποφάσεων. Για να αποφευχθεί η περίπτωση το σύστημα να εντοπίσει το αποτέλεσμα μέσα στο ίδιο το κείμενο της απόφασης και προκειμένου να μάθει νέες πληροφορίες, πρέπει να αφαιρεθούν όλες οι αναφορές στην ετυμηγορία.

Αν και ένας αλγόριθμος μπορεί να επιτύχει πολύ καλά αποτελέσματα στην κατηγοριοποίηση, οι παραγόμενες κατηγορίες από μόνες τους δεν είναι ιδιαίτερα χρήσιμες. Δεδομένου ότι τα έγγραφα που χρησιμοποιεί το σύστημα είναι διαθέσιμα μόνο όταν οι αποφάσεις έχουν ήδη δημοσιευθεί, η κατηγοριοποίηση βάσει έκβασης δεν προσθέτει νέα πληροφορία (καθώς κάποιος μπορεί απλώς να εξάγει την ετυμηγορία από την απόφαση). Αυτή η άποψη υποστηρίζεται επίσης από τους [13], οι οποίοι τονίζουν ότι η ικανότητα κατηγοριοποίησης αποφάσεων χωρίς την εξήγηση του λόγου για αυτή την κατηγοριοποίηση δεν παρέχει χρήσιμες πληροφορίες και μπορεί ακόμη και να είναι παραπλανητική. Παρ' όλα αυτά, η απόδοση ενός μοντέλου μηχανικής μάθησης για την κατηγοριοποίηση δικαστικών αποφάσεων μπορεί να δώσει χρήσιμες πληροφορίες σχετικά με το πόσο πληροφοριακά είναι τα χαρακτηριστικά που χρησιμοποιούνται. Για να καταστεί δυνατή η εξαγωγή χαρακτηριστικών, είναι σημαντικό το σύστημα να μην είναι ένα ¨μαύρο κουτί¨ (όπως συμβαίνει με πολλά σύγχρονα νευρωνικά μοντέλα κατηγοριοποίησης). Επομένως, αντί για *¨πρόβλεψη δικαστικών αποφάσεων¨*, ο κύριος στόχος της κατηγοριοποίησης βάσει έκβασης θα πρέπει να είναι η *αναγνώριση των παραγόντων* που υποστηρίζουν τις κατηγοριοποιήσεις.

Καθώς εδώ συζητούνται μόνο δημοσιεύσεις που κατηγοριοποιούν δικαστικές αποφάσεις βάσει της έκβασης των υποθέσεων, θα αναφερόμαστε στη διαδικασία αυτή απλώς ως κατηγοριοποίηση δικαστικών αποφάσεων.

Μέσα σε αυτές τις μελέτες, διακρίνονται δύο ευρείες κατηγορίες ανάλογα με τον τύπο των δεδομένων που χρησιμοποιούν. Από τη μία πλευρά, οι περισσότερες μελέτες χρησιμοποιούν το ακατέργαστο κείμενο, επιλέγοντας ρητά τμήματα της απόφασης που δεν περιλαμβάνουν την ετυμηγορία. Από την άλλη πλευρά, υπάρχουν μελέτες που χρησιμοποιούν δεδομένα με χειροκίνητη επισημείωση ως βάση για την κατηγοριοποίηση. Στην εργασία [14], οι μελετητές χρησιμοποίησαν το ακατέργαστο κείμενο για να κατηγοριοποιήσουν (με accuracy 67%) τις αποφάσεις του Ανώτατου Δικαστηρίου της Ταϊλάνδης βάσει των πραγματικών περιστατικών και των νομικών διατάξεων που αφορούσαν εγκλήματα όπως δολοφονία, επίθεση, κλοπή, απάτη και δυσφήμιση, χρησιμοποιώντας στατιστικές και νευρωνικές μεθόδους. Ομοίως, στην [15] κατηγοριοποίησαν (με ακρίβεια έως 75%) αποφάσεις του ΕΔΔΑ χρησιμοποιώντας μόνο την ενότητα των πραγματικών περιστατικών κάθε απόφασης. Μάλιστα,εντόπισαν τους κύριους

παράγοντες (δηλαδή ακολουθίες λέξεων) για κάθε κατηγορία χρησιμοποιώντας μια μέθοδο υποστηρικτικών διανυσμάτων (SVM). Ακόμη, στην μελέτη [16] κατηγοριοποίησαν αποφάσεις του Ανώτατου Δικαστηρίου του Ηνωμένου Βασιλείου, συγκρίνοντας διάφορα συστήματα εκπαιδευμένα στο ακατέργαστο κείμενο (χωρίς την ετυμηγορία) και ανέφεραν ακρίβεια 69%, παρουσιάζοντας παράλληλα τους βασικούς προβλεπτικούς παράγοντες κάθε κατηγορίας. Περαιτέρω μελέτες επικεντρώθηκαν στο ΕΔΔΑ, χρησιμοποιώντας νευρωνικές μεθόδους για τη βελτίωση της απόδοσης της κατηγοριοποίησης, επιτυγχάνοντας ακρίβειες έως 82% [17]. Παρά τις υψηλές επιδόσεις ορισμένων συστημάτων, οι ερευνητές προειδοποιούν ενάντια στη χρήση τους για λήψη αποφάσεων. Πολλές μελέτες ισχυρίζονται ότι η κατηγοριοποίηση είναι χρήσιμη για τη νομική υποστήριξη, αλλά όπως αναφέρθηκε, από μόνη της δεν είναι χρήσιμη, καθώς η ετυμηγορία μπορεί να διαβαστεί απευθείας. Για να είναι ωφέλιμη, η κατηγοριοποίηση πρέπει να συνοδεύεται από τα πιο χαρακτηριστικά προβλεπτικά χαρακτηριστικά. Δυστυχώς, μόνο λίγες μελέτες παρέχουν αυτή την πληροφορία.

2.1.3 Πρόβλεψη δικαστικών αποφάσεων

Οι εργασίες πρόβλεψης δικαστικών αποφάσεων σχετίζονται με ένα ευρύ φάσμα ερευνών που συνδυάζουν την επεξεργασία φυσικής γλώσσας (NLP) και τη μηχανική μάθηση για την ανάλυση και πρόβλεψη δικαστικών αποφάσεων. Η μελέτη της εργασίας [4] αποτελεί μια σημαντική προσπάθεια πρόβλεψης αποφάσεων του Ευρωπαϊκού Δικαστηρίου Ανθρωπίνων Δικαιωμάτων βασισμένη αποκλειστικά σε κείμενα δικαστικών αποφάσεων. Χρησιμοποιώντας αλγόριθμους μηχανικής μάθησης, όπως οι Support Vector Machines (SVM) κατάφεραν να πετύχουν με υψηλή ακρίβεια (79%) στις προβλέψεις τους σχετικά με το αν υπήρξε ή όχι παραβίαση άρθρων της Σύμβασης των Ανθρωπίνων Δικαιωμάτων. Η ανάλυση υπογράμμισε τη σημασία των πραγματικών περιστατικών ως τον πλέον καθοριστικό παράγοντα για την πρόβλεψη αποφάσεων.

Η εργασία [8] εστιάζει στη σύγκριση της απόδοσης πέντε γνωστών αλγορίθμων μηχανικής μάθησης για την πρόβλεψη δικαστικών αποφάσεων, χρησιμοποιώντας αποκλειστικά κείμενα από υποθέσεις του Ευρωπαϊκού Δικαστηρίου Δικαιωμάτων του Ανθρώπου (ECtHR). Οι αλγόριθμοι που μελετήθηκαν είναι οι k-Nearest Neighbors (k-NN), Logistic Regression, Bagging, Random Forests και Support Vector Machines (SVM). Η μεθοδολογία βασίστηκε σε ένα σύνολο δεδομένων που περιλαμβάνει υποθέσεις σχετικές με τα άρθρα 3, 6 και 8 της Ευρωπαϊκής Σύμβασης Δικαιωμάτων του Ανθρώπου. Τα κείμενα μετατράπηκαν σε αριθμητικά χαρακτηριστικά μέσω N-gram αναπαραστάσεων (1-gram έως 4-gram) και θεματικής μοντελοποίησης (topic modeling) με χρήση spectral clustering. Οι επιλεγμένες περιοχές των εγγράφων περιλάμβαναν τμήματα όπως η διαδικασία, οι περιστάσεις, η σχετική νομοθεσία και η νομική ανάλυση. Για την αξιολόγηση της απόδοσης των αλγορίθμων, εφαρμόστηκε μέθοδος cross-validation 5 πτυχών, ενώ χρησιμοποιήθηκαν μετρικές όπως η ακρίβεια. Τα αποτελέσματα έδειξαν ότι ο αλγόριθμος SVM ξεπέρασε τους υπόλοιπους, ειδικά όταν χρησιμοποιήθηκαν χαρακτηριστικά που σχετίζονται με τα γεγονότα των υποθέσεων και τα θεματικά μοντέλα. Συγκεκριμένα, ο SVM πέτυχε τις καλύτερες επιδόσεις σε περιπτώσεις με περιορισμένο πλήθος δειγμάτων, λόγω της ικανότητάς του να διαχειρίζεται καλά μικρά σύνολα δεδομένων. Ενδιαφέρον παρουσιάζει το εύρημα ότι τα χαρακτηριστικά που προέρ-

χονται από την ενότητα «Περιστάσεις» είναι πιο σημαντικά για την πρόβλεψη των αποφάσεων από ό,τι η ίδια η νομική ανάλυση. Η εργασία καταλήγει ότι η επιλογή χαρακτηριστικών με βάση το περιεχόμενο και η χρήση αλγορίθμων με ισχυρές δυνατότητες κατηγοριοποίησης, όπως ο SVM, είναι κρίσιμες για τη βελτίωση της ακρίβειας στην πρόβλεψη δικαστικών αποφάσεων. Προτείνεται μελλοντική έρευνα στη χρήση τεχνικών εξαγωγής σημασιολογικών χαρακτηριστικών, όπως τα word embeddings, για περαιτέρω βελτίωση των αποτελεσμάτων.

Η εργασία [11] επικεντρώνεται στην πρόβλεψη δικαστικών αποφάσεων στη Βραζιλία, καθώς και στην εκτίμηση αν οι αποφάσεις αυτές θα είναι ομόφωνες. Για την επίτευξη αυτού του στόχου, οι συγγραφείς χρησιμοποίησαν ένα σύνολο δεδομένων με 4.043 δικαστικές υποθέσεις από το Ανώτατο Δικαστήριο της Πολιτείας Alagoas. Η μεθοδολογία βασίζεται σε τεχνικές Επεξεργασίας Φυσικής Γλώσσας (NLP) και αλγορίθμους μηχανικής μάθησης, συνδυάζοντας παραδοσιακούς ταξινομητές με μοντέρνα μοντέλα βαθιάς μάθησης. Οι παραδοσιακοί αλγόριθμοι που χρησιμοποιήθηκαν περιλαμβάνουν Gaussian Naive Bayes, Decision Trees, Support Vector Machines (SVM), Random Forests και XGBoost. Παράλληλα, αξιοποιήθηκαν σύγχρονα μοντέλα βαθιάς μάθησης όπως BERT (στην παραλλαγή BERT-Imbau, ειδικά εκπαιδευμένη για τα πορτογαλικά), LSTM, GRU, BiLSTM και Convolutional Neural Networks (CNN). Για την επεξεργασία των δεδομένων, οι συγγραφείς ανέπτυξαν έναν web scraper για την εξαγωγή νομικών αποφάσεων από δικαστικές ιστοσελίδες. Ακολούθησε προεπεξεργασία κειμένων, που περιλάμβανε κανονικοποίηση λέξεων (lowercasing), αφαίρεση στοπωορδς και stemming. Η αναπαράσταση των κειμένων έγινε με δύο προσεγγίσεις: χρήση tf-idf για τους παραδοσιακούς αλγορίθμους και word embeddings για τα βαθιά νευρωνικά δίκτυα. Τα πειράματα χωρίστηκαν σε πέντε σενάρια, τα οποία αφορούσαν είτε την πρόβλεψη της τελικής απόφασης (αποδοχή, απόρριψη, μερική αποδοχή) είτε την εκτίμηση αν η απόφαση ήταν ομόφωνη. Οι αλγόριθμοι αξιολογήθηκαν με χρήση 5-fold cross-validationκαι μετρικές όπως F1-score, ακρίβεια και ανάκληση. Τα αποτελέσματα έδειξαν ότι ο XGBoost πέτυχε την υψηλότερη απόδοση στην πρόβλεψη της τελικής απόφασης (F1-score 80,2%), ενώ το Random Forest και το BERT-Imbau παρουσίασαν εξαιρετική απόδοση στην πρόβλεψη της ομοφωνίας. Παρατηρήθηκε ότι τα κλασικά μοντέλα μηχανικής μάθησης είχαν καλύτερες επιδόσεις σε μικρότερα και μη ισορροπημένα σύνολα δεδομένων, ενώ τα βαθιά μοντέλα απέδωσαν καλύτερα όταν το σύνολο ήταν ισορροπημένο. Η εργασία καταλήγει ότι οι τεχνικές μηχανικής μάθησης μπορούν να προσφέρουν ουσιαστική υποστήριξη στους νομικούς επαγγελματίες, βελτιώνοντας την ακρίβεια και την αποδοτικότητα στην πρόβλεψη δικαστικών αποτελεσμάτων. Προτείνεται, τέλος, η περαιτέρω διερεύνηση πιο εξελιγμένων τεχνικών, όπως τα named entity recognition μοντέλα, για την ενίσχυση της πρόβλεψης.

Στην [18] εφαρμόστηκαν προηγμένα μοντέλα NLP και βαθιάς μάθησης (όπως GRUs, LSTMs με μηχανισμούς προσοχής) για την πρόβλεψη αποφάσεων ανώτερων δικαστηρίων της Τουρκίας, με έμφαση στις περιγραφές των πραγματικών περιστατικών. Τα αποτελέσματα απέδειξαν ότι οι μέθοδοι βαθιάς μάθησης μπορούν να επιτύχουν ακρίβεια έως και 93%, παρέχοντας ένα σταθερό σημείο αναφοράς για τη μελλοντική έρευνα.

Ακόμη, η [15] εξετάζει τον τρόπο με τον οποίο, τεχνικές μηχανικής μάθησης και φυσικής επεξεργασίας γλώσσας μπορούν να αξιοποιηθούν για την πρόβλεψη των αποφάσεων του Ευρωπαϊκού Δικαστηρίου Ανθρωπίνων Δικαιωμάτων. Οι συγγραφείς χρησιμοποίησαν δεδομένα από δημοσιευμένες αποφάσεις του ΕΔΑΔ, εφαρμόζοντας τον αλγόριθμο Support Vector

Machines (SVM) για την ταξινόμηση των υποθέσεων σε «παραβίαση» ή «μη παραβίαση» άρθρων της Ευρωπαϊκής Σύμβασης Ανθρωπίνων Δικαιωμάτων. Η μέθοδος που περιγράφεται στη συγκεκριμένη εργασία πέτυχε μέση ακρίβεια 75% για την πρόβλεψη παραβιάσεων ενώ παράλληλα, αναλύθηκαν τα χαρακτηριστικά που επηρεάζουν την πρόβλεψη, με στόχο την κατανόηση των παραγόντων που καθορίζουν τις αποφάσεις.

Η [19] προτείνει μια μέθοδο πρόβλεψης δικαστικών αποφάσεων σε υποθέσεις διαζυγίου, βασισμένη σε Markov Logic Networks (MLN). Η μέθοδος αυτή περιλαμβάνει την τυπική περιγραφή και εξαγωγή των νομικών παραγόντων, τη δημιουργία και εκπαίδευση ενός Markov Logic Network και την πρόβλεψη των δικαστικών αποφάσεων. Αρχικά, χρησιμοποιείται μια γραμματική δημιουργίας (generative grammar) για την τυπική περιγραφή των νομικών παραγόντων, οι οποίοι κατηγοριοποιούνται σε παράγοντες γεγονότων και αποτελεσμάτων. Στη συνέχεια, ένας μηχανισμός εξαγωγής γνώσης (knowledge extraction engine) ανιχνεύει αυτούς τους παράγοντες από δικαστικές αποφάσεις. Οι εξαγόμενες πληροφορίες χρησιμοποιούνται για την εκπαίδευση του MLN, το οποίο αποτελείται από συναρτήσεις και τύπους πρώτης τάξης (first-order logic formulas) με αντίστοιχα βάρη. Κατά την πρόβλεψη, τα δεδομένα εισόδου μετατρέπονται σε ένα αρχείο δοκιμών, το οποίο αξιοποιείται από το MLN για την εξαγωγή συμπερασμάτων. Τα πειράματα πραγματοποιήθηκαν με δεδομένα από το China Judgements Online και έδειξαν ότι η προσέγγιση είναι ανθεκτική σε διαφορετικά στυλ έκφρασης και παρέχει ερμηνεύσιμα αποτελέσματα, καθώς τα προβλεπόμενα αποτελέσματα συνοδεύονται από τους αντίστοιχους κανόνες επαγωγής. Επιπλέον, σε σύγκριση με άλλα μοντέλα μηχανικής μάθησης, η μέθοδος αυτή προσφέρει καλύτερη κατανόηση των αποφάσεων μέσω της χρήσης τυπικών νομικών περιγραφών και σαφών συσχετίσεων μεταξύ των παραγόντων.

Τέλος, η εργασία [20] επικεντρώνεται στην πρόβλεψη νομικών αποφάσεων από τα δικαστήρια του Ηνωμένου Βασιλείου, βασιζόμενη αποκλειστικά στο κείμενο των δικαστικών εγγράφων. Πρόκειται για την πρώτη μελέτη που στοχεύει στη δημιουργία ενός μοντέλου πρόβλεψης για το Ηνωμένο Βασίλειο, καθώς προηγούμενες έρευνες είχαν επικεντρωθεί σε άλλες δικαιοδοσίες όπως η ΕΕ, η Γαλλία και η Κίνα. Οι συγγραφείς ανέπτυξαν ένα νέο επισημασμένο σύνολο δεδομένων που περιλαμβάνει 4.959 δικαστικές αποφάσεις από το Ανώτατο Δικαστήριο του Ηνωμένου Βασιλείου. Τα δεδομένα αποκτήθηκαν μέσω web scraping και επισημάνθηκαν ως "allow" (αποδοχή) ή "dismiss" (απόρριψη) χρησιμοποιώντας αυτοματοποιημένες τεχνικές ανίχνευσης προτύπων. Η προεπεξεργασία των δεδομένων περιλάμβανε βήματα όπως μετατροπή σε πεζά, αφαίρεση αριθμών και stopwords, και λεμματοποίηση, μειώνοντας σημαντικά τον αριθμό μοναδικών λέξεων. Όσον αφορά στους αλγορίθμους μηχανικής μάθησης, εξετάστηκαν Support Vector Machines (SVM), Logistic Regression (LR), Random Forests (RF), k-Nearest Neighbors (k-NN), καθώς και νευρωνικά δίκτυα όπως Single Layer Perceptron (SLP) και Multi-Layer Perceptron (MLP). Η ρύθμιση των υπερπαραμέτρων έγινε μέσω cross-validated random search, ενώ η αξιολόγηση βασίστηκε σε μετρικές όπως ακρίβεια, F1-score, precision και recall. Τα αποτελέσματα έδειξαν ότι ο συνδυασμός TF-IDF με Logistic Regression ήταν ο πιο αποδοτικός, πετυχαίνοντας F1-score 69.02% και ακρίβεια 69.05%. Τα Random Forests και οι κλασικοί αλγόριθμοι παρήγαγαν επίσης ικανοποιητικά αποτελέσματα, ενώ οι μέθοδοι topic modeling και word embeddings είχαν χαμηλότερες επιδόσεις, πιθανώς λόγω του περιορισμένου όγκου δεδομένων. Παράλληλα,

η ανάλυση των πιο σημαντικών χαρακτηριστικών (n-grams) ανέδειξε λέξεις και φράσεις με ισχυρή προβλεπτική ικανότητα. Τέλος, η εργασία προτείνει ότι πιο πολύπλοκες αρχιτεκτονικές νευρωνικών δικτύων, όπως Hierarchical Attention Networks (HAN) και Recurrent Neural Networks (RNN), θα μπορούσαν να προσφέρουν βελτιωμένες επιδόσεις σε μελλοντικές μελέτες.

2.1.4 Εργασίες με δικαστικές αποφάσεις

Στην εργασία [21] οι συγγραφείς εστιάζουν στην αυτόματη περίληψη ελληνικών νομικών εγγράφων. Για την αντιμετώπιση της έλλειψης κατάλληλων συνόλων δεδομένων, δημιούργησαν ένα νέο, πλούσιο σε μεταδεδομένα σύνολο δεδομένων, το οποίο περιλαμβάνει επιλεγμένες αποφάσεις του Αρείου Πάγου με αντίστοιχες περιλήψεις και κατηγοριοποιήσεις. Στα πλαίσια της εργασίας εφαρμόζονται σύγχρονες μέθοδοι αφαιρετικής και εξαγωγικής περίληψης, ενώ τα αποτελέσματα αξιολογούνται τόσο με αυτόματους δείκτες όσο και μέσω ανθρώπινης αξιολόγησης. Τα αποτελέσματα έδειξαν ότι: (i) οι εξαγωγικές μέθοδοι αποδίδουν μέτρια, ενώ οι αφαιρετικές παράγουν πιο ευέλικτα και συνεκτικά κείμενα, αν και με χαμηλότερη συνάφεια και συνέπεια· (ii) υπάρχει ανάγκη για βελτιωμένους δείκτες που αποτυπώνουν καλύτερα τη συνεκτικότητα και τη συνάφεια νομικών περιλήψεων· (iii) η προσαρμογή μοντέλων τύπου BERT σε εξειδικευμένα καθήκοντα μπορεί να ενισχύσει σημαντικά την απόδοσή τους.

Η [22] αυτή προτείνει μια μέθοδο για την αυτόματη αναγνώριση των κατάλληλων νομικών διατάξεων που ισχύουν για μια υπόθεση, χρησιμοποιώντας machine learning (ML). Το πρόβλημα διατυπώνεται ως multi-label classification, καθώς μια υπόθεση μπορεί να σχετίζεται με πολλαπλές νομικές διατάξεις. Η μεθοδολογία περιλαμβάνει τέσσερα στάδια: απόκτηση δεδομένων μέσω web scraping, προεπεξεργασία κειμένου, δημιουργία αγωγού επεξεργασίας (pipelining) και ανάλυση της απόδοσης των μοντέλων. Τα δεδομένα προέρχονται από υποθέσεις του Indian Income Tax Act of 1963 και καταχωρούνται σε μορφή csv, αφού φιλτραριστούν για ανωμαλίες. Στην προεπεξεργασία εφαρμόζονται τεχνικές όπως stopword removal, POS tagging, stemming και TF-IDF vectorization. Το σύστημα περιλαμβάνει τέσσερα ML models: logistic regression, decision tree, Naïve Bayes και support vector machine (SVM), καθώς και ένα προσαρμοσμένο convolutional neural network (CNN). Τα πειράματα έδειξαν ότι το SVM με RBF kernel είχε την υψηλότερη ακρίβεια με F1-score 0.75, ενώ το CNN παρουσίασε χαμηλότερη απόδοση. Η εργασία καταλήγει στο ότι η μέθοδος μπορεί να επεκταθεί και σε άλλες νομικές διατάξεις και προτείνει τη χρήση recurrent neural networks (RNN) και long short-term memory (LSTM) για μελλοντικές βελτιώσεις.

Επιπλέον, στην [23] προτείνει μια στρατηγική βασισμένη στο BERT ([24]) για την ταξινόμηση νομικών υποθέσεων σε civil ή criminal. Το προτεινόμενο μοντέλο προσαρμόζει την τεχνική early stopping και βελτιστοποιεί το BERT-Base χρησιμοποιώντας υπερπαραμέτρους σε ένα σώμα επισημασμένων δικαστικών εγγράφων, εκμεταλλευόμενο την ικανότητα του BERT να συλλαμβάνει συμφραζόμενες πληροφορίες και ακριβείς σχέσεις μεταξύ λέξεων και προτάσεων. Ο μηχανισμός προσοχής (attention mechanism) στο BERT επιτρέπει στο μοντέλο να κατανοεί τις λεπτομέρειες των νομικών δεδομένων και να διακρίνει τις civil από τις criminal υποθέσεις βάσει του δεδομένου νομικού πλαισίου. Τα πειραματικά αποτελέσματα σε ένα μεγάλο σύνολο δεδομένων νομικών υποθέσεων αποδεικνύουν την αποτελεσματικότητα

της μεθόδου, επιτυγχάνοντας εξαιρετικά επίπεδα ακρίβειας, precision, recall και F1-score. Το μοντέλο βασισμένο στο BERT παρουσιάζει σημαντικές δυνατότητες για την αυτοματοποίηση της κατηγοριοποίησης υποθέσεων, διευκολύνοντας τη διαχείριση των υποθέσεων, υποστηρίζοντας τους νομικούς επαγγελματίες στη διαδικασία λήψης αποφάσεων και παρέχοντας στους πολίτες χρήσιμες αναλύσεις και προβλέψεις βάσει διαθέσιμων δεδομένων, με ακρίβεια 92.9% και F1-score 0.93.

Η εργασία [25] αυτή εξετάζει το ζήτημα της semantic bias στην απονομή δικαιοσύνης μέσω τεχνητής νοημοσύνης (AI), εστιάζοντας στον εντοπισμό και την ταξινόμηση μεροληπτικών στοιχείων σε νομικές αποφάσεις. Η ύπαρξη γνωστικών προκαταλήψεων (cognitive bias) σε μεγάλα νομικά δεδομένα μπορεί να οδηγήσει σε μη ηθικές ή ανακριβείς προβλέψεις, καθώς τα συστήματα AI εκπαιδεύονται σε αυτά. Η μελέτη χρησιμοποιεί το Chinese AI and Law (CAIL) dataset, το οποίο περιλαμβάνει εκατοντάδες υποθέσεις, για να αναλύσει πώς τα μοντέλα machine learning (ML) επηρεάζονται από τη σημασιολογική μεροληψία. Προτείνεται ένα ταξινομητικό μοντέλο που ανιχνεύει τη σημασιολογική προκατάληψη σε δικαστικές αποφάσεις, συμβάλλοντας έτσι στην αύξηση της νομιμότητας των δεδομένων εκπαίδευσης. Για την ταξινόμηση χρησιμοποιούνται διάφοροι ταξινομητές, όπως Support Vector Machine (SVM), Naïve Bayes (NB), Multi-Layer Perceptron (MLP) και K-Nearest Neighbour (KNN). Τα πειραματικά αποτελέσματα δείχνουν ότι το SVM πετυχαίνει την υψηλότερη ακρίβεια (96.90%), ακολουθούμενο από το NB (88.80%), το MLP (86.75%) και το KNN (85.66%). Τα μοντέλα επιτυγχάνουν υψηλή απόδοση στην πρόβλεψη αποτελεσμάτων με βάση την κατηγοριοποίηση της σημασιολογικής προκατάληψης. Η εργασία υπογραμμίζει την ανάγκη για δίκαια και αντικειμενικά συστήματα AI στο νομικό πεδίο, προτείνοντας την αυτοματοποίηση της ανάλυσης των δικαστικών αποφάσεων μέσω τεχνικών natural language processing (NLP) και ML. Η μελέτη προσφέρει μια νέα μεθοδολογία για τη μείωση της σημασιολογικής προκατάληψης, συμβάλλοντας στη βελτίωση της ακρίβειας και της διαφάνειας στη νομική απόδοση δικαιοσύνης.

Η [26] εξετάζει διάφορες μεθόδους ταξινόμησης κειμένου και διαφορετικούς συνδυασμούς embeddings, που εξάγονται από γλωσσικά μοντέλα στην πορτογαλική γλώσσα, καθώς και πληροφορίες για τη νομοθεσία που αναφέρεται στα αρχικά έγγραφα. Τα μοντέλα εκπαιδεύτηκαν σε ένα σύνολο δεδομένων (Golden Collection) που αποτελείται από 16.000 αρχικές αιτήσεις και κατηγορητήρια από το Court of Justice of the State of Ceará, στη Βραζιλία. Οι υποθέσεις ταξινομήθηκαν στις πέντε πιο αντιπροσωπευτικές κατηγορίες του CNJ: Common Civil Procedure, Execution of Extrajudicial Title, Criminal Action - Ordinary Procedure, Special Civil Court Procedure και Tax Enforcement. Το καλύτερο αποτέλεσμα επιτεύχθηκε με το μοντέλο BERT, το οποίο πέτυχε F1 score (macro) 0.88 στο πειραματικό σενάριο όπου η υπόθεση αναπαρίσταται μέσω ενός embedding που προκύπτει από τη συνένωση των κειμένων όλων των αιτήσεων που περιέχουν τουλάχιστον μία αναφορά στη νομοθεσία. Τα νομικά έγγραφα έχουν ιδιαίτερα χαρακτηριστικά, όπως μεγάλο μέγεθος, εξειδικευμένο λεξιλόγιο, επίσημη σύνταξη, σημασιολογία βασισμένη σε εκτεταμένη γνώση του νομικού τομέα και παραπομπές σε νόμους. Η ερμηνεία μας είναι ότι η αναπαράσταση του εγγράφου μέσω συμφραζομενικών embeddings που δημιουργεί το BERT, καθώς και η δισδιάστατη αρχιτεκτονική του μοντέλου με διπλή κατεύθυνση (bidirectional contexts), επιτρέπουν τη σύλληψη του εξειδικευμένου πλαισίου των νομικών εγγράφων με μεγαλύτερη ακρίβεια.

2.2 Τεχνικές Προβλέψεων

2.2.1 Ορισμός και Διαδικασία Πρόβλεψης

Ως πρόβλεψη μπορεί να οριστεί η εκτίμηση αβέβαιων μελλοντικών γεγονότων. Οι προβλέψεις μπορούν να γίνουν βασισμένες στην εμπειρία και την παρατήρηση, σε στατιστικές μεθόδους, καθώς και σε πολύπλοκα μαθηματικά μοντέλα. Χρησιμοποιούνται για τη βελτίωση της λήψης αποφάσεων και σχεδιασμού [27].

Η διαδικασία παραγωγής προβλέψεων είναι μια απαιτητική διαδικασία. Στην ακόλουθη παράγραφο θα περιγραφούν επιγραμματικά τα πέντε βασικά βήματα που είναι απαραίτητα για την παραγωγή και αξιολόγηση προβλέψεων:

1. *Καθορισμός του προβλήματος.* Συνιστά ένα από τα πιο σημαντικά και ταυτόχρονα δυσκολότερα μέρη της διαδικασίας παραγωγής προβλέψεων. Σε αυτό το βήμα γίνεται απόπειρα να καθοριστούν τα επιθυμητά μεγέθη που πρόκειται να προβλευθούν, καθώς και η μετέπειτα χρήση των προβλέψεων αυτών.

2. *Συλλογή των δεδομένων.* Η διαδικασία αυτή αποδεικνύεται συχνά χρονοβόρα, καθώς εκτός των μετρήσιμων αριθμητικών δεδομένων, σημαντική αποδεικνύεται και η χρήση διαθέσιμων εμπειρικών πληροφοριών για το αντικείμενο προς μελέτη.

3. *Προεπεξεργασία των δεδομένων.* Ένα καίριο βήμα για την παραγωγή προβλέψεων συνιστά η απόκτηση μιας ολοκληρωμένης αίσθησης των διαθέσιμων δεδομένων, έτσι ώστε να εντοπιστούν πιθανά λάθη, ασυνήθιστες τιμές, σημαντικές τάσεις ή εποχικότητα. Σκοπός της προεπεξεργασίας των δεδομένων είναι η δημιουργία ενός εξομαλυμένου συνόλου δεδομένων για την εφαρμογή των μοντέλων πρόβλεψης.

4. *Επιλογή μεθόδων πρόβλεψης.* Επιτυγχάνεται η ορθή επιλογή μοντέλων πρόβλεψης καθώς και η ιδιαίτερα σημαντική διαδικασία επιλογής των κατάλληλων παραμέτρων τους, ώστε να παραχθούν τα πλέον ακριβή αποτελέσματα.

5. *Χρήση και αξιολόγηση των μοντέλων πρόβλεψης.* Το τελικό στάδιο περιλαμβάνει την χρήση των επιλεγμένων μοντέλων ώστε να παραχθούν οι ζητούμενες προβλέψεις. Το κατά πόσο οι προβλέψεις των επιλεγμένων μοντέλων είναι ικανοποιητικές μπορεί να κριθεί μόνο με την πάροδο του χρόνου, και πιο συγκεκριμένα καθώς τα νέα δεδομένα γίνονται διαθέσιμα. Η αξιολόγηση και η μέτρηση της ακρίβειας των προβλέψεων επιτυγχάνεται με εξειδικευμένους στατιστικούς δείκτες.

2.2.2 Κατηγορίες Μεθόδων Πρόβλεψης

Οι μέθοδοι πρόβλεψης, σύμφωνα με την διαδικασία παραγωγής τους, διακρίνονται σε τρεις μεγάλες κατηγορίες :

1. **Ποσοτικές Μέθοδοι**. Οι ποσοτικές μέθοδοι αναφέρονται στην εφαρμογή στατιστικών μοντέλων χρονοσειρών ή αιτιοκρατικών μοντέλων επί μιας σειρά δεδομένων με σκοπό αυτοματοποιημένη και συστηματική παραγωγή προβλέψεων. Οι στατιστικές

προβλέψεις είναι αποδεκτά ακριβείς και εφαρμόσιμες, αν συνδυαστούν με κατάλληλα διαστήματα εμπιστοσύνης. Προϋποθέτουν ότι η συμπεριφορά της εκάστοτε χρονοσειράς θα συνεχιστεί στο μέλλον, κάτι το οποίο δεν συμβαίνει πάντα. Επιπροσθέτως, κύρια παραδοχή των μοντέλων αυτών συνιστά η σταθερή συσχέτιση μεταξύ του προς πρόβλεψη μεγέθους και άλλων παραγόντων, χωρίς ωστόσο να είναι απαραίτητη η ύπαρξη χρονική εξάρτησης. Η συλλογή των δεδομένων αποτελεί συχνά μια χρονοβόρα και ενίοτε δύσκολη διαδικασία, καθώς απαιτείται μεγάλο πλήθος ιστορικών δεδομένων προκειμένουν να παραχθούν οι ζητούμενες προβλέψεις. Τέτοια μοντέλα είναι οι μέθοδοι εκθετικής εξομάλυνσης, τα μοντέλα παλινδρόμησης, τα μοντέλα ARIMA και τα τεχνητά νευρωνικά δίκτυα.

2. **Κριτικές Μέθοδοι**. Οι κριτικές μέθοδοι πρόβλεψης δεν έχουν τις ίδιες απαιτήσεις σε δεδομένα όπως οι στατιστικές μέθοδοι. Τα δεδομένα των κριτικών μεθόδων αποτελούν προϊόν διαίσθησης, κρίσης και συσσωρευμένης γνώσης από πλευράς εμπειρογνωμόνων. Οι μέθοδοι αυτές μπορούν να λάβουν υπόψη ειδικά γεγονότα και ενέργειες, ενώ ταυτόχρονα έχουν τη δυνατότητα να αντισταθμίζουν ανεπάρκειες και ελλείψεις σε ιστορικά δεδομένα. Είναι κατάλληλες όταν θίγονται ηθικά ζητήματα που υπερισχύουν των οικονομικών και τεχνολογικών παραγόντων. Οι μέθοδοι αυτές πρέπει να λειτουργούν συμπληρωματικά με τις μεθόδους στατιστικής μελέτης. Ανάμεσα στις πιο διαδεδομένες μεθόδους συγκαταλέγονται η απλή κρίση, η μέθοδος Delphi και οι δομημένες αναλογίες.

3. **Τεχνολογικές Μέθοδοι**. Οι τεχνολογικές μέθοδοι πρόβλεψης αφορούν κυρίως μακροπρόθεσμα πλάνα τεχνολογικής, οικονομικής, κοινωνικής και πολιτικής φύσης. Διακρίνονται σε διερευνητικές και κανονιστικές. Οι πρώτες έχουν ως σημείο εκκίνησης το παρελθόν και το παρόν και στοχεύουν στη διερεύνηση όλων των πιθανών μελλοντικών περιπτώσεων. Οι κανονιστικές έχουν προκαθορισμένους στόχους και εξετάζουν τη δυνατότητα επίτευξής τους, σύμφωνα με τους υπάρχοντες περιορισμούς και διαθέσιμους πόρους.

2.2.3 Μηχανική Μάθηση

Η μηχανική μάθηση είναι υποπεδίο της επιστήμης των υπολογιστών που αναπτύχθηκε από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην τεχνητή νοημοσύνη. Η μηχανική μάθηση διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν να μαθαίνουν από τα δεδομένα και να κάνουν προβλέψεις σχετικά με αυτά. Οι αλγόριθμοι αυτοί βελτιώνουν τη συμπεριφορά τους σε κάποια εργασία χρησιμοποιώντας την εμπειρία τους. Τέτοιοι αλγόριθμοι λειτουργούν κατασκευάζοντας μοντέλα από πειραματικά δεδομένα, προκειμένου να κάνουν προβλέψεις βασιζόμενες στα δεδομένα ή να εξάγουν αποφάσεις που εκφράζονται ως το αποτέλεσμα. Ο Άρθουρ Σάμουελ ορίζει τη μηχανική μάθηση ως *¨Πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν, χωρίς να έχουν ρητά προγραμματιστεί¨*.

Ο τομέας της μηχανικής μάθησης αναπτύσσει τρεις τρόπους μάθησης, ανάλογους με τους τρόπους με τους οποίους μαθαίνει ο άνθρωπος:

1. **Επιβλεπόμενη μάθηση** *(Supervised Learning)*. Η επιβλεπόμενη μάθηση είναι η διαδικασία όπου ο αλγόριθμος κατασκευάζει μια συνάρτηση που απεικονίζει δεδομένες εισόδους (σύνολο εκπαίδευσης) σε γνωστές επιθυμητές εξόδους, με απώτερο στόχο τη γενίκευση της συνάρτησης αυτής και για εισόδους με άγνωστη έξοδο. Χρησιμοποιείται σε προβλήματα ταξινόμησης (classification), πρόγνωσης (prediction) και διερμηνείας (interpretation).

2. **Μη επιβλεπόμενη μάθηση** *(Unsupervised Learning)*. Στην μη επιβλεπόμενη μάθηση ο αλγόριθμος κατασκευάζει ένα μοντέλο για κάποιο σύνολο εισόδων υπό μορφή παρατηρήσεων χωρίς να γνωρίζει τις επιθυμητές εξόδους. Χρησιμοποιείται σε προβλήματα ανάλυσης συσχετισμών (association analysis) και ομαδοποίησης (clustering).

3. **Ενισχυτική μάθηση** *(Reinforcement Learning)*. Στην ενισχυτική μάθηση ο αλγόριθμος μαθαίνει μια στρατηγική ενεργειών μέσα από άμεση αλληλεπίδραση με το περιβάλλον. Χρησιμοποιείται κυρίως σε προβλήματα σχεδιασμού, όπως ο έλεγχος κίνησης ρομπότ και η βελτιστοποίηση εργασιών σε εργοστασιακούς χώρους.

2.3 Μοντέλα Πρόβλεψης - Ταξινόμησης

Στην παρούσα ενότητα παρουσιάζονται τα μοντέλα μηχανικής μάθησης που δοκιμάστηκαν στο πρόβλημα της πρόβλεψης δικαστικών αποφάσεων. Οι πιο γνωστές και χρήσιμες μέθοδοι για την πρόβλεψη έκβασης δικαστικών αποφάσεων -βάσει της βιβλιογραφίας- είναι τα Δέντρα Αποφάσεων (*Decision Trees*), τα Τυχαία Δάση (*Random Forest*), οι Μηχανές Υποστήριξης Διανυσμάτων «*SVMs*) καθώς και η Γραμμική Παλινδρόμηση (*Linear Regression*) [4].

2.3.1 Δέντρα Αποφάσεων - Decision Trees

Τα δέντρα αποφάσεων (decision trees) είναι ένας δημοφιλής και διαισθητικός αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται για προβλήματα ταξινόμησης και παλινδρόμησης [28]. Πρόκειται για μια δενδροειδή δομή όπου τα δεδομένα χωρίζονται διαδοχικά με βάση τα χαρακτηριστικά τους, με στόχο την κατηγοριοποίηση ή την πρόβλεψη μιας τιμής.

1. Ένα δέντρο αποφάσεων αποτελείται από κόμβους και κλάδους. Ο πρώτος κόμβος (ριζικός κόμβος) είναι ο κόμβος από τον οποίο ξεκινά η ανάλυση των δεδομένων. Στη συνέχεια, κάθε κόμβος διασπάται με βάση τις τιμές ενός χαρακτηριστικού, δημιουργώντας δύο ή περισσότερους υποκόμβους. Οι τελικοί κόμβοι, γνωστοί ως φύλλα, περιέχουν την τελική απόφαση (την κατηγορία ή την τιμή) που αντιστοιχεί στο εκάστοτε δείγμα.

2. Η επιλογή των χαρακτηριστικών που χρησιμοποιούνται για τη διάσπαση κάθε κόμβου γίνεται με βάση κριτήρια που βελτιστοποιούν την ποιότητα της διάσπασης. Στην ταξινόμηση, τα πιο κοινά κριτήρια είναι:

(α΄) Δείκτης Gini: Μετρά την ακαθαρσία (impurity) του κόμβου. Η τιμή του κυμαίνεται μεταξύ 0 (όλα τα δείγματα του κόμβου ανήκουν στην ίδια κατηγορία) και 0,5 (τα δείγματα κατανέμονται εξίσου μεταξύ δύο κατηγοριών).

(β΄) Εντροπία: Μετρά την αβεβαιότητα ή την ανομοιογένεια του κόμβου. Στόχος είναι η ελαχιστοποίηση της εντροπίας σε κάθε διάσπαση.

(γ΄) Παλινδρόμηση: στην παλινδρόμηση, η διάσπαση συνήθως βασίζεται στη μέση τετραγωνική απόκλιση ή στο σφάλμα παλινδρόμησης, όπου η μέθοδος προσπαθεί να ελαχιστοποιήσει την απόκλιση μεταξύ της προβλεπόμενης και της πραγματικής τιμής.

3. Δημιουργία του Δέντρου: Το δέντρο αποφάσεων δημιουργείται αναδρομικά, με κάθε διάσπαση να αυξάνει το βάθος του δέντρου και να διαιρεί τα δεδομένα σε μικρότερα και πιο ομοιογενή υποσύνολα. Η διαδικασία συνεχίζεται μέχρι να επιτευχθούν ορισμένα κριτήρια τερματισμού, όπως:

(α΄) Όλοι οι κόμβοι έχουν το ίδιο χαρακτηριστικό ή είναι αρκετά καθαροί.

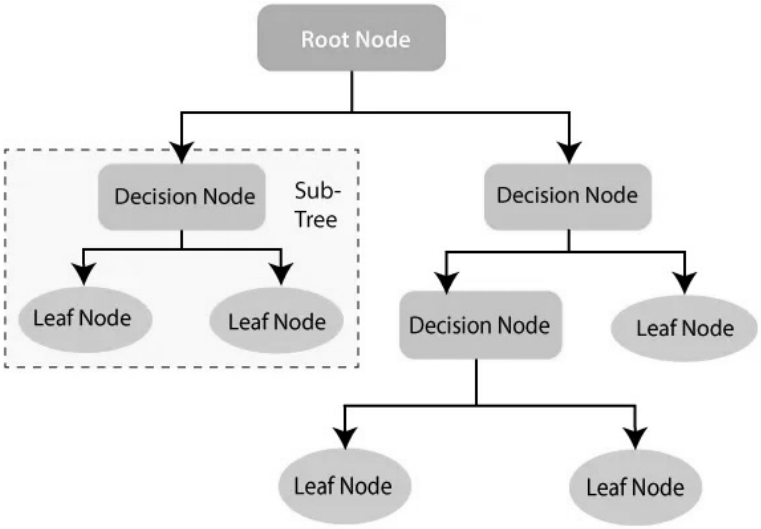
(β΄) Έχει φτάσει το μέγιστο προκαθορισμένο βάθος του δέντρου.

(γ΄) Κάθε κόμβος έχει λιγότερα από έναν ελάχιστο αριθμό δεδομένων.

4. Καθώς τα δέντρα αποφάσεων τείνουν να υπερεκπαιδεύονται, εφαρμόζεται μια τεχνική που ονομάζεται κλάδεμα (pruning). Το κλάδεμα απομακρύνει τους κόμβους που δεν προσφέρουν σημαντικές βελτιώσεις στην ακρίβεια, μειώνοντας το βάθος του δέντρου και βελτιώνοντας τη γενική του ικανότητα.

5. Τα δέντρα αποφάσεων είναι εύκολα στην κατανόηση και την ερμηνεία, καθώς η ιεραρχική διαδικασία διάσπασης είναι διαισθητική. Έχουν χαμηλό υπολογιστικό κόστος και είναι κατάλληλα για προβλήματα με πολλά χαρακτηριστικά, καθώς και για διαχείριση δεδομένων με ελλείψεις.

6. Τα δέντρα αποφάσεων τείνουν να υπερεκπαιδεύονται στα δεδομένα εκπαίδευσης, ειδικά αν το δέντρο δεν κλαδευτεί. Αυτό οδηγεί σε χαμηλή γενική ικανότητα και απόδοση σε νέα δεδομένα. Επιπλέον, είναι ευαίσθητα στις αλλαγές στα δεδομένα (π.χ., μια μικρή αλλαγή μπορεί να αλλάξει σημαντικά τη δομή του δέντρου).



Σχήμα 2.2: *Μοντέλο Δέντρων Απόφασης*

2.3.2 Τυχαία Δάση - Random Forests

Τα τυχαία δάση (Random Forests) είναι ένας από τους πιο δημοφιλείς αλγόριθμους μηχανικής μάθησης, ιδιαίτερα για προβλήματα ταξινόμησης και παλινδρόμησης. Αναπτύχθηκαν από τον Leo Breiman το 2001 [29] και βασίζονται σε μια προσέγγιση σύνολου (*ensemble learning*), δηλαδή συνδυάζουν πολλά ανεξάρτητα μοντέλα (*δέντρα αποφάσεων*) για να βελτιώσουν την ακρίβεια των προβλέψεων και να μειώσουν την πιθανότητα υπερεκπαίδευσης.

1. Κάθε τυχαίο δάσος αποτελείται από έναν αριθμό δέντρων αποφάσεων. Το δέντρο αποφάσεων (decision tree) είναι ένα μοντέλο που δημιουργεί μια σειρά από κανόνες, οι οποίοι βασίζονται στις τιμές των χαρακτηριστικών ενός δείγματος, για να καταλήξει σε μια απόφαση (π.χ., την κατηγορία στην οποία ανήκει το δείγμα ή την πρόβλεψη τιμής). Ωστόσο, τα δέντρα αποφάσεων έχουν την τάση να υπερεκπαιδεύονται, δηλαδή να προσαρμόζονται υπερβολικά στα δεδομένα εκπαίδευσης, μειώνοντας έτσι τη γενική ικανότητά τους σε νέα δεδομένα.

2. Τα τυχαία δάση καταπολεμούν το πρόβλημα της υπερεκπαίδευσης συνδυάζοντας πολλαπλά δέντρα, καθένα από τα οποία εκπαιδεύεται σε διαφορετικό υποσύνολο των δεδομένων εκπαίδευσης. Ο αλγόριθμος χρησιμοποιεί την τεχνική του bootstrap aggregation (ή bagging), δηλαδή επιλέγει τυχαία, με επανάληψη, ένα υποσύνολο των δεδομένων για την εκπαίδευση κάθε δέντρου. Αυτό βοηθά να δημιουργηθούν πιο ανεξάρτητα δέντρα και μειώνει την πιθανότητα υπερεκπαίδευσης.

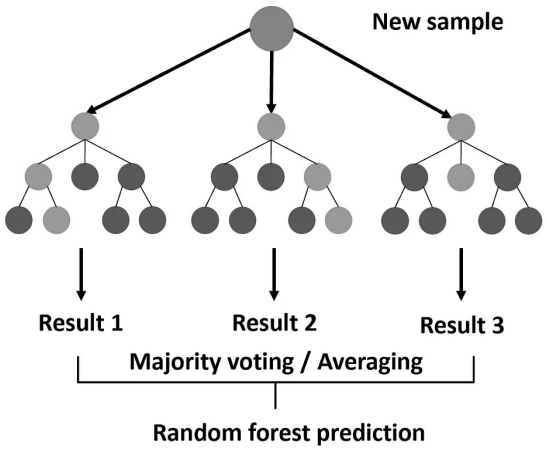
3. Σε κάθε κόμβο ενός δέντρου στο τυχαίο δάσος, ο αλγόριθμος εξετάζει ένα τυχαίο υποσύνολο χαρακτηριστικών, αντί για όλα τα διαθέσιμα χαρακτηριστικά, για να επιλέξει την καλύτερη διάσπαση. Αυτή η τυχαιότητα επιτρέπει τη δημιουργία δέντρων που είναι πιο διαφοροποιημένα μεταξύ τους και βελτιώνει τη γενική ικανότητα του μοντέλου.

4. Συνδυασμός Αποφάσεων: Στο τέλος, το τυχαίο δάσος συνδυάζει τις προβλέψεις από κάθε δέντρο για να παράγει την τελική απόφαση. Στην ταξινόμηση, γίνεται με ψηφο-

φορία πλειοψηφίας (majority voting), όπου η τελική πρόβλεψη είναι η κατηγορία που επιλέγεται πιο συχνά από τα δέντρα. Στην παλινδρόμηση, η τελική πρόβλεψη είναι ο μέσος όρος των προβλέψεων όλων των δέντρων.

5. Τα τυχαία δάση είναι ανθεκτικά στην υπερεκπαίδευση, ειδικά για μεγάλα σύνολα δεδομένων με υψηλή διαστατικότητα. Έχουν υψηλή ακρίβεια και μπορούν να χρησιμοποιηθούν για την εκτίμηση της σημασίας των χαρακτηριστικών, πράγμα που είναι πολύτιμο σε προβλήματα όπου χρειάζεται να κατανοήσουμε ποια χαρακτηριστικά έχουν μεγαλύτερη επιρροή.

6. Ένα από τα κύρια μειονεκτήματα των τυχαίων δασών είναι ότι απαιτούν περισσότερους υπολογιστικούς πόρους, ιδιαίτερα για μεγάλα σύνολα δεδομένων, και είναι πιο δύσκολο να ερμηνευτούν σε σχέση με απλούστερα μοντέλα.



Σχήμα 2.3: *Μοντέλο Τυχαίων Δασών*

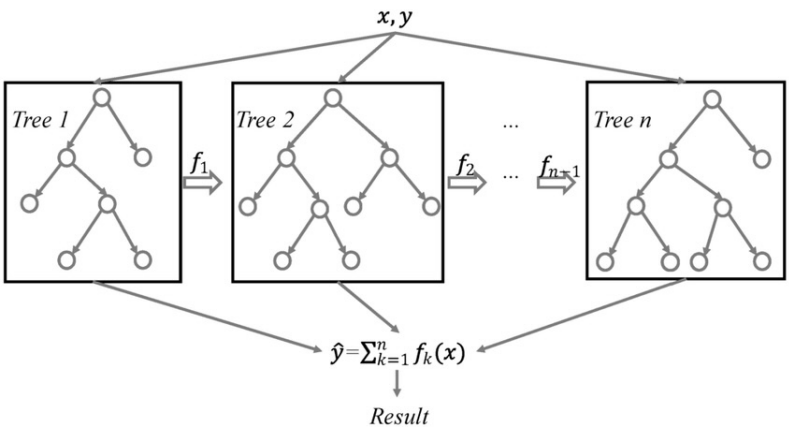
2.3.3 XGBoost Classifier

Η μέθοδος XGBoost είναι η συντομογραφία του Extreme Gradient Boosting και αποτελεί επίσης έναν αλγόριθμο ενίσχυσης [30]. Με τον όρο ενίσχυση (boosting), εννοείται μια οικογένεια αλγορίθμων συνδυασμού μοντέλων, οι οποίες αποσκοπούν στη μετατροπή αδύναμων μοντέλων με χαμηλή απόδοση, σε ισχυρά μοντέλα πρόβλεψης. Οι αλγόριθμοι αυτοί βρίσκουν εφαρμογή τόσο σε προβλήματα ταξινόμησης, όσο και παλινδρόμησης. Η μέθοδος Gradient Boosting, αποτελεί μια ισχυρή τεχνική συνδυασμού αδύναμων μοντέλων, με στόχο τη δημιουργία ενός ισχυρότερου. Αρχικά, εκπαιδεύεται ένα μοντέλο σε ένα υποσύνολο των δεδομένων εκπαίδευσης. Χρησιμοποιώντας αυτό το μοντέλο, γίνονται προβλέψεις σε ολόκληρο το σύνολο δεδομένων εκπαίδευσης και υπολογίζεται το σφάλμα του. Αυτές οι προβλέψεις είναι ανεπαρκείς και το αδύναμο μοντέλο θα πρέπει να ενισχυθεί σε μεταγενέστερες επαναλήψεις. Αυτός είναι ο λόγος για τον οποίον δημιουργείται ένα νέο μοντέλο, λαμβάνοντας υπόψη, τα σφάλματα που υπολογίστηκαν πριν, προκειμένου να διορθώσει τα λάθη του πρώτου. Τέλος, οι προβλέψεις του νέου μοντέλου συνδυάζονται με του προηγούμενου.

Το XGBoost υποστηρίζει πολλές τεχνικές για την ενίσχυση της απόδοσής του:

- Early stopping: Σταματά την εκπαίδευση όταν η απόδοση στο σύνολο επικύρωσης δεν βελτιώνεται μετά από έναν ορισμένο αριθμό επαναλήψεων.
- Feature importance: Προσφέρει εργαλεία για την ανάλυση της σημασίας των χαρακτηριστικών, ώστε να μπορούμε να εντοπίσουμε ποια χαρακτηριστικά συμβάλλουν περισσότερο στις προβλέψεις.
- Regularization: Περιλαμβάνει L1 (Lasso) και L2 (Ridge) κανονικοποίηση για την πρόληψη υπερεκπαίδευσης.
- Parallellization: Εκμεταλλεύεται την παραλληλία για να επιταχύνει τη διαδικασία εκπαίδευσης.

Η επιλογή παραμέτρων, όπως ο αριθμός των δέντρων, το μέγιστο βάθος κάθε δέντρου, η μάθηση ρυθμού (*learning rate*), και η ελάχιστη απώλεια μείωσης (*min child weight*), επηρεάζουν σημαντικά την απόδοση του XGBoost. Το XGBoost θεωρείται ένας από τους πιο ισχυρούς αλγόριθμους, ιδιαίτερα σε προβλήματα με μεγάλα και πολύπλοκα δεδομένα. Ωστόσο, λόγω της πολυπλοκότητάς του, μπορεί να είναι υπολογιστικά απαιτητικό και να χρειάζεται προσεκτική ρύθμιση των παραμέτρων για την επίτευξη της καλύτερης απόδοσης.



Σχήμα 2.4: *XGBoost Regression*

2.3.4 Μηχανές Υποστήριξης Διανυσμάτων - SVM

Οι Υποστηρικτικές Μηχανές Διανυσμάτων (Support Vector Machines - SVM) είναι ένας ισχυρός αλγόριθμος μηχανικής μάθησης για ταξινόμηση και παλινδρόμηση [31]. Η βασική ιδέα του SVM είναι η εύρεση μιας υπερ-επιφάνειας (ή υπερ-επίπεδου) που διαχωρίζει με τον καλύτερο δυνατό τρόπο τα δεδομένα σε κατηγορίες.

1. Ένα υπερ-επίπεδο είναι ένας γραμμικός διαχωριστής σε έναν χώρο δεδομένων πολλών διαστάσεων. Στην SVM, επιδιώκεται να βρεθεί το υπερ-επίπεδο που διαχωρίζει τα δεδομένα με το μέγιστο περιθώριο (maximum margin), δηλαδή το υπερ-επίπεδο που βρίσκεται όσο το δυνατόν πιο μακριά από τα σημεία των δύο κατηγοριών. Αυτό το περιθώριο επιτρέπει στο μοντέλο να έχει καλύτερη γενική ικανότητα και να αντέχει στις αλλαγές των δεδομένων.

2. Τα σημεία που βρίσκονται πλησιέστερα στο υπερ-επίπεδο διαχωρισμού ονομάζονται διανύσματα υποστήριξης (support vectors). Αυτά τα διανύσματα είναι τα πιο κρίσιμα δεδομένα για την εκπαίδευση του μοντέλου, καθώς η θέση τους καθορίζει το μέγιστο περιθώριο και την τελική θέση του υπερ-επιπέδου. Αν αλλάξουν αυτά τα σημεία, αλλάζει και το υπερ-επίπεδο διαχωρισμού.

3. Η βασική SVM είναι γραμμική, πράγμα που σημαίνει ότι χρησιμοποιεί έναν γραμμικό διαχωριστή για την ταξινόμηση των δεδομένων. Ωστόσο, σε πολλά προβλήματα τα δ εδομένα δεν είναι γραμμικά διαχωρίσιμα. Για να αντιμετωπιστεί αυτό, η SVM χρησιμοποιεί τον πυρήνα (kernel trick), δηλαδή έναν μετασχηματισμό που επιτρέπει τη μετατροπή των δεδομένων σε έναν χώρο υψηλότερων διαστάσεων, όπου τα δεδομένα γίνονται γραμμικά διαχωρίσιμα.

4. Η επιλογή του κατάλληλου πυρήνα εξαρτάται από τη φύση των δεδομένων. Οι πιο συνηθισμένοι πυρήνες είναι:

   (α') *Γραμμικός πυρήνας*: Χρησιμοποιείται όταν τα δεδομένα είναι γραμμικά διαχωρίσιμα.

   (β') *Πολυωνυμικός πυρήνας*: Κατάλληλος για μη γραμμικές σχέσεις.

   (γ') *Πυρήνας Radial Basis Function* (RBF): Είναι από τους πιο διαδεδομένους για μη γραμμικά δεδομένα και χρησιμοποιείται συχνά όταν τα δεδομένα έχουν πολύπλοκες σχέσεις.

   (δ') *Πυρήνας Sigmoid*: Χρησιμοποιείται σπανιότερα αλλά μπορεί να έχει εφαρμογές σε συγκεκριμένες περιπτώσεις.

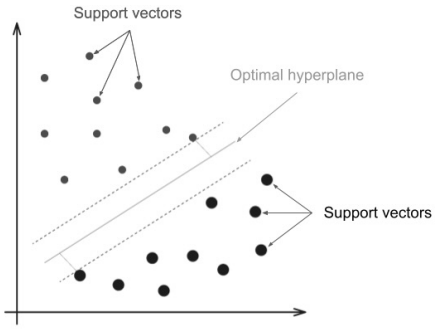
5. Παράμετροι C και γ (*Gamma*): Οι δύο κύριες παράμετροι που ρυθμίζουν την απόδοση της SVM είναι οι παράμετροι C και γ.

   (α') Παράμετρος C: Ρυθμίζει το μέγεθος του περιθωρίου του υπερ-επιπέδου διαχωρισμού. Μια υψηλή τιμή του C μειώνει το περιθώριο, προσπαθώντας να ταξινομήσει όλα τα δεδομένα σωστά. Αυτό μπορεί να οδηγήσει σε υπερεκπαίδευση. Μια χαμηλή τιμή του C αυξάνει το περιθώριο, αλλά μπορεί να επιτρέψει κάποια σφάλματα ταξινόμησης.

   (β') Παράμετρος γ: Στους πυρήνες RBF και πολυωνυμικού τύπου, η παράμετρος γ καθορίζει την επίδραση των μεμονωμένων δεδομένων στο διαχωριστικό υπερ-επίπεδο. Χαμηλές τιμές του γ κάνουν το μοντέλο να έχει πιο ευρύχωρες καμπύλες, ενώ υψηλές τιμές του γ κάνουν το μοντέλο να προσαρμόζεται στενά στα δεδομένα.

6. Η SVM είναι πολύ αποτελεσματική σε περιπτώσεις όπου τα δεδομένα έχουν σαφή διαχωριστικά όρια και είναι ιδανική για προβλήματα υψηλής διαστατικότητας. Είναι ανθεκτική στην υπερεκπαίδευση, ειδικά όταν χρησιμοποιείται με το κατάλληλο πυρήνα.

7. Η SVM μπορεί να είναι υπολογιστικά απαιτητική, ιδιαίτερα για μεγάλα σύνολα δεδομένων. Επίσης, η επιλογή του κατάλληλου πυρήνα και των παραμέτρων είναι δύσκολη και απαιτεί πειραματισμό. Σε προβλήματα όπου οι κατηγορίες δεν είναι καθαρά διαχωρίσιμες, η SVM μπορεί να παρουσιάσει υποδεέστερη απόδοση σε σχέση με άλλες μεθόδους.



Σχήμα 2.5: *Support Vector Machines*

2.3.5 Γραμμική Παλινδρόμηση - Linear Regression

Η γραμμική παλινδρόμηση είναι ένας απλός και ευρέως χρησιμοποιούμενος αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται για την πρόβλεψη μιας συνεχούς εξαρτημένης μεταβλητής (στόχου) από μία ή περισσότερες ανεξάρτητες μεταβλητές (χαρακτηριστικά). Ο αλγόριθμος επιδιώκει να βρει την καλύτερη δυνατή γραμμική σχέση ανάμεσα στην εξαρτημένη και τις ανεξάρτητες μεταβλητές [32].

1. Στη γραμμική παλινδρόμηση, το μοντέλο περιγράφεται από τη γραμμική εξίσωση:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

   όπου:

   - ψ είναι η εξαρτημένη μεταβλητή (η πρόβλεψη).
   - $x_1, x_2, \ldots, x_n$ είναι οι ανεξάρτητες μεταβλητές.
   - β_0 είναι το σταθερό όρο (intercept).
   - $\beta_1, \beta_2, \ldots, \beta_n$ είναι οι συντελεστές παλινδρόμησης (coefficients) για κάθε ανεξάρτητη μεταβλητή, οι οποίοι υποδεικνύουν το μέγεθος και την κατεύθυνση της επίδρασης κάθε ανεξάρτητης μεταβλητής στην εξαρτημένη μεταβλητή.
   - ε είναι το σφάλμα (error term), που αντιπροσωπεύει τη διαφορά μεταξύ των πραγματικών και των προβλεπόμενων τιμών.

2. Ανάλυση και Εκτίμηση των Συντελεστών: Οι συντελεστές εκτιμώνται με τη μέθοδο ελαχίστων τετραγώνων (*Ordinary Least Squares - OLS*). Η μέθοδος αυτή επιλέγει τις

τιμές των συντελεστών που ελαχιστοποιούν το άθροισμα των τετραγωνικών διαφορών (σφαλμάτων) μεταξύ των πραγματικών και των προβλεπόμενων τιμών:

Ελαχιστοποίηση του $\sum(y_{\text{πραγματικό}} - y_{\text{πρόβλεψη}})^2$

3. Υποθέσεις της γραμμικής παλινδρόμησης :

   (α΄) Γραμμικότητα: Υπάρχει γραμμική σχέση μεταξύ των ανεξάρτητων μεταβλητών και της εξαρτημένης μεταβλητής.

   (β΄) Κανονικότητα των Σφαλμάτων: Τα σφάλματα ακολουθούν κανονική κατανομή με μέσο όρο 0.

   (γ΄) Ομοσκεδαστικότητα: Η διασπορά των σφαλμάτων είναι σταθερή για όλες τις τιμές των ανεξάρτητων μεταβλητών (δηλαδή, δεν αλλάζει ανάλογα με την τιμή των μεταβλητών).

   (δ΄) Ανεξαρτησία των Σφαλμάτων: Τα σφάλματα δεν είναι συσχετισμένα μεταξύ τους (δεν υπάρχει αυτοσυσχέτιση).

   (ε΄) Μη πολυσυγγραμμικότητα: Οι ανεξάρτητες μεταβλητές δεν πρέπει να παρουσιάζουν ισχυρή συσχέτιση μεταξύ τους (αποφυγή πολυσυγγραμμικότητας), καθώς αυτό θα μπορούσε να οδηγήσει σε ασταθείς εκτιμήσεις συντελεστών.

4. Είδη Γραμμικής Παλινδρόμησης:

   (α΄) Απλή Γραμμική Παλινδρόμηση: Εξετάζει τη σχέση μεταξύ μιας εξαρτημένης μεταβλητής και μιας ανεξάρτητης μεταβλητής (π.χ., πρόβλεψη βάρους από το ύψος).

   (β΄) Πολλαπλή Γραμμική Παλινδρόμηση: Περιλαμβάνει πολλές ανεξάρτητες μεταβλητές και είναι κατάλληλη για πιο περίπλοκα προβλήματα όπου πολλοί παράγοντες επηρεάζουν το αποτέλεσμα.

5. Η γραμμική παλινδρόμηση είναι εύκολη στην ερμηνεία, γρήγορη στην εκπαίδευση και παρέχει ένα εύκολα κατανοητό μοντέλο. Επίσης, είναι πολύ χρήσιμη όταν επιδιώκεται η κατανόηση της σχέσης μεταξύ μεταβλητών.

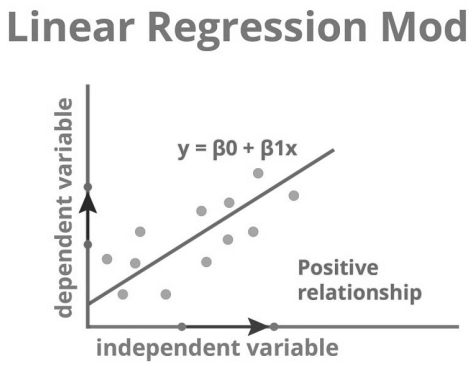
6. Η απόδοση της γραμμικής παλινδρόμησης μπορεί να είναι χαμηλή σε περιπτώσεις όπου οι σχέσεις είναι μη γραμμικές ή οι υποθέσεις της παραβιάζονται. Επίσης, είναι ευαίσθητη στις ακραίες τιμές (outliers) και μπορεί να επηρεαστεί από την πολυσυγγραμμικότητα.

Η γραμμική παλινδρόμηση (Linear Regression) αποτελεί ένα τρόπο μοντελοποίησης της σχέσης μεταξύ μιας μεταβλητής εξόδου και μιας ή περισσοτέρων εισόδων. Η μεταβλητή εξόδου ονομάζεται εξαρτημένη μεταβλητή, ενώ οι μεταβλητές εισόδου ονομάζονται ανεξάρτητες μεταβλητές. Στο μοντέλο της γραμμικής παλινδρόμησης γίνεται η υπόθεση ότι η σχέση αυτή είναι γραμμική, γεγονός που στην πραγματικότητα δε συμβαίνει συχνά. Σημαντική προϋπόθεση για την παραγωγή του μοντέλου αυτού είναι η απουσία συσχέτισης μεταξύ των ανεξάρτητων μεταβλητών. Συχνά, όταν οι ανεξάρτητες μεταβλητές είναι περισσότερες από μια, το μοντέλο ονομάζεται *πολλαπλή γραμμική παλινδρόμηση* (multiple linear regression). Το μοντέλο έχει την εξής μορφή :

$$Y = b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + b_i X_{ip} + e_i, \quad \text{για κάθε δείγμα } i = 1, 2, \ldots, n$$

Στην παραπάνω σχέση, Y είναι η εξαρτημένη μεταβλητή, X_{ij} είναι το i-οστό δείγμα της j ανεξάρτητης μεταβλητής X, όπου j=1,2,...,p. Ακόμη, b_i είναι οι παράμετροι του μοντέλου και e_i είναι η απόκλιση από της πραγματικές τιμές. Οι σταθερές παράμετροι b_i του μοντέλου, υπολογίζονται με τη μέθοδο των *κανονικών ελάχιστων τετραγώνων*, το πρόβλημα έγκειται στην ελαχιστοποίηση της συνάρτησης απωλειών :

$$L(X, y) = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{n} (b \cdot X_i - y_i)^2$$



Σχήμα 2.6: *Linear Regression Model*

2.3.6 Λογιστική Παλινδρόμηση - Logistic Regression

Η λογιστική παλινδρόμηση (*Logistic Regression*) είναι ένας αλγόριθμος ταξινόμησης που χρησιμοποιείται για τη μοντελοποίηση της πιθανότητας εμφάνισης μιας δυαδικής εξαρτημένης μεταβλητής. Αντί να προβλέπει απευθείας τιμές, προβλέπει πιθανότητες, οι οποίες μετατρέπονται σε κατηγορίες μέσω ενός κατωφλίου [33].

1. Το μοντέλο βασίζεται στη λογιστική συνάρτηση (*sigmoid function*), η οποία περιγράφεται ως:

   $$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}}$$

   όπου:

   $$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

   και:

   - $\hat{y}$: η πιθανότητα η παρατήρηση να ανήκει στην κατηγορία 1.
   - z: ο γραμμικός συνδυασμός των χαρακτηριστικών.
   - $x_1, x_2, \ldots, x_n$: οι ανεξάρτητες μεταβλητές3.
   - $\beta_0, \beta_1, \ldots, \beta_n$: οι συντελεστές του μοντέλου.

2. Η συνάρτηση κόστους που χρησιμοποιείται για την εκπαίδευση του μοντέλου είναι η αρνητική λογαριθμική πιθανοφάνεια (*log-loss*):

$$L(\beta) = -\frac{1}{m}\sum_{i=1}^{m}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

όπου:

- m: ο αριθμός των δειγμάτων.
- y_i: η πραγματική τιμή της εξαρτημένης μεταβλητής για το i-οστό δείγμα.
- $\hat{y}_i$: η πιθανότητα που προβλέπει το μοντέλο για το i-οστό δείγμα.

3. Οι εκτιμήσεις των συντελεστών β γίνονται μέσω της μέγιστης πιθανοφάνειας (*Maximum Likelihood Estimation - MLE*), ελαχιστοποιώντας τη συνάρτηση κόστους $L(\beta)$.

4. Υποθέσεις της λογιστικής παλινδρόμησης:

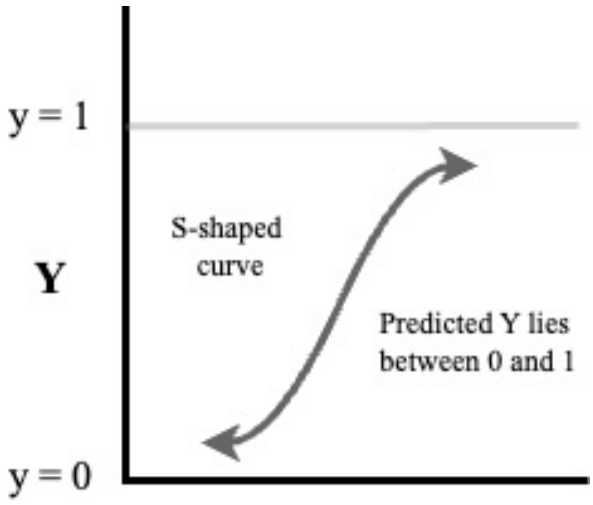
- Γραμμικότητα: Υπάρχει γραμμική σχέση μεταξύ των χαρακτηριστικών και του λογαρίθμου των πιθανοτήτων (*log-odds*):

$$\text{log-odds} = \log\left(\frac{P(y=1)}{P(y=0)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

- Ανεξαρτησία των παρατηρήσεων.
- Απουσία πολυσυγγραμμικότητας: Οι ανεξάρτητες μεταβλητές δεν πρέπει να είναι ισχυρά συσχετισμένες.

5. Το μοντέλο χρησιμοποιείται σε πλήθος εφαρμογών, όπως:

- Πρόβλεψη δυαδικών γεγονότων, π.χ., αν ένας πελάτης θα αγοράσει ένα προϊόν.
- Αναγνώριση απάτης, π.χ., αν μια συναλλαγή είναι ύποπτη.



Σχήμα 2.7: *Logistic Regression Model*

2.4 Μετρικές Απόδοσης

Οι μετρικές απόδοσης είναι εργαλεία που χρησιμοποιούνται για την αξιολόγηση της αποτελεσματικότητας και της ακρίβειας ενός αλγορίθμου ή μοντέλου μηχανικής μάθησης. Είναι ιδιαίτερα σημαντικές στην προσπάθειά μας να να ποσοτικοποιήσουμε το κατά πόσο καλά ένα μοντέλο προβλέπει ή ταξινομεί δεδομένα, επιτρέποντας έτσι τη σύγκριση διαφορετικών μοντέλων και φυσικά την επιλογή του καταλληλότερου.

2.4.1 Accuracy

Η ακρίβεια (accuracy) είναι ένας βασικός δείκτης απόδοσης των αλγορίθμων ταξινόμησης στη μηχανική μάθηση και εκφράζει το ποσοστό των σωστών προβλέψεων (θετικές και αρνητικές) σε σχέση με το συνολικό αριθμό των προβλέψεων [34]. Ο υπολογισμός της ακρίβειας δίνεται από τη σχέση :

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total Number of Samples}}$$

όπου:

- TP (True Positives) είναι οι περιπτώσεις στις οποίες το μοντέλο προβλέπει σωστά μια θετική κατηγορία.
- TN (True Negatives) είναι οι περιπτώσεις στις οποίες το μοντέλο προβλέπει σωστά μια αρνητική κατηγορία.

2.4.2 F1-Score

Το **F1-Score** είναι μια μετρική απόδοσης που χρησιμοποιείται για την αξιολόγηση μοντέλων ταξινόμησης, ειδικά όταν υπάρχει ανισορροπία μεταξύ των κατηγοριών ή όταν ενδιαφερόμαστε εξίσου για την *Precision* και την *Recall* [35]. Ορίζεται ως ο αρμονικός μέσος του *Precision* και του *Recall*:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

2.4.3 Precision και Recall

Η έννοια του *Precision* (Ευστοχία) και του *Recall* (Ανάκληση ή Ευαισθησία) εξηγείται ως εξής:

- **Precision:** Από όλες τις προβλέψεις που έγιναν ως θετικές, πόσες ήταν πράγματι σωστές:
$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$
- **Recall**: Από όλες τις θετικές περιπτώσεις στο σύνολο δεδομένων, πόσες αναγνωρίστηκαν σωστά:
$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Το **F1-Score** προσφέρει μία ενιαία τιμή που ισορροπεί το *Precision* και το *Recall*, κάτι που είναι κρίσιμο στις παρακάτω περιπτώσεις:

- Όταν υπάρχει ανισορροπία στις κλάσεις (π.χ. μία κλάση εμφανίζεται πολύ πιο συχνά από την άλλη).
- Όταν το *Accuracy* δίνει παραπλανητική εικόνα λόγω ασύμμετρων ψευδών προβλέψεων (*False Positives* και *False Negatives*).

Ένα υψηλό **F1 Score** υποδηλώνει ότι το μοντέλο έχει καλή ισορροπία μεταξύ *Precision* και *Recall*.

2.4.4 Mathews Correlation Coefficient - MCC

Η **Matthews Correlation Coefficient (MCC)** είναι μία μετρική που αξιολογεί την ποιότητα ενός ταξινομητή, ειδικά σε προβλήματα με ανισορροπία στις κλάσεις. Υπολογίζει τη συσχέτιση μεταξύ των πραγματικών τιμών και των προβλέψεων, λαμβάνοντας υπόψη όλους τους τύπους σφαλμάτων: *True Positives (TP)*, *True Negatives (TN)*, *False Positives (FP)*, και *False Negatives (FN)*. Ο μαθηματικός της ορισμός είναι:

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

- **Τιμές:** Η τιμή της MCC κυμαίνεται από -1 έως $+1$:
  - $+1$: Ιδανική ταξινόμηση (όλες οι προβλέψεις σωστές).
  - 0 : Τυχαία πρόβλεψη (καμία συσχέτιση).
  - -1 : Απόλυτα λανθασμένη ταξινόμηση (αντιστροφή όλων των προβλέψεων).
- Η MCC είναι κατάλληλη για ανισόρροπα σύνολα δεδομένων, καθώς ενσωματώνει όλες τις κατηγορίες λαθών στη σχέση του.
- Σε περιπτώσεις ανισορροπίας, η MCC θεωρείται πιο αξιόπιστη από το *F1 Score*, καθώς δίνει μια συνολική εικόνα για την απόδοση του ταξινομητή.

Η MCC είναι κατάλληλο:

- Για προβλήματα δυαδικής ή πολυκατηγορικής ταξινόμησης.
- Όταν υπάρχει σημαντική ανισορροπία στις κλάσεις.
- Όταν απαιτείται μια γενική μετρική που να ενσωματώνει όλες τις διαστάσεις των προβλέψεων (σωστά και λάθη).

Η **MCC** είναι μία από τις πιο αξιόπιστες μετρικές για την αξιολόγηση ταξινομητών, ιδιαίτερα σε περιπτώσεις με ανισορροπία στις κλάσεις. Παρέχει μια συνολική εικόνα της απόδοσης του μοντέλου, λαμβάνοντας υπόψη τόσο τις σωστές όσο και τις λανθασμένες προβλέψεις.

2.4.5 Τυπική Απόκλιση - Standard Deviation

Στη στατιστική, η Τυπική Απόκλιση είναι ένα μέτρο της ποσότητας παραλλαγής ή διασποράς ενός συνόλου τιμών. Μια χαμηλή τυπική απόκλιση υποδεικνύει ότι οι τιμές τείνουν να είναι κοντά στον μέσο όρο (ή αλλιώς την αναμενόμενη τιμή) του συνόλου, ενώ μια υψηλή τυπική απόκλιση υποδεικνύει ότι οι τιμές είναι κατανεμημένες σε ένα ευρύτερο εύρος. Η τυπική απόκλιση μπορεί να συντομευθεί σε SD και συνήθως αναπαρίσταται σε μαθηματικά κείμενα και εξισώσεις με το μικρό ελληνικό γράμμα σίγμα (σ) για την τυπική απόκλιση του πληθυσμού, ή το λατινικό γράμμα s για την τυπική απόκλιση του δείγματος. Η τυπική απόκλιση μιας τυχαίας μεταβλητής, δείγματος, στατιστικού πληθυσμού, συνόλου δεδομένων ή κατανομής πιθανοτήτων είναι η τετραγωνική ρίζα της διακύμανσης. Είναι αλγεβρικά απλούστερη, αν και στην πράξη λιγότερο ανθεκτική από την μέση απόλυτη απόκλιση [36].

Συγκεκριμένα, ο τύπος για την τυπική απόκλιση του πληθυσμού εκφράζεται από την εξίσωση :

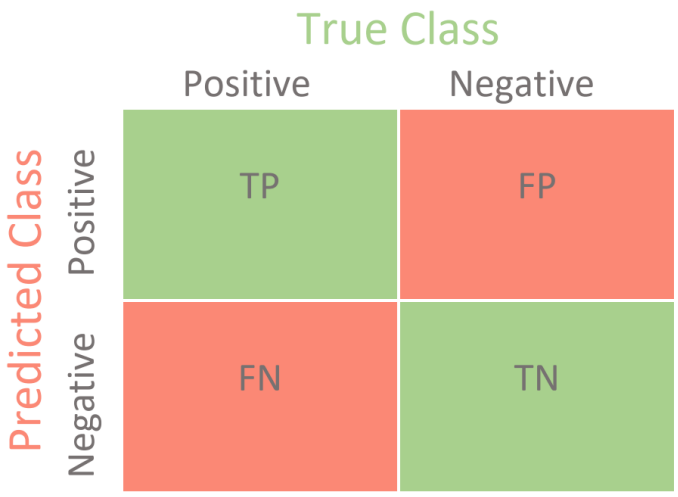
$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Αντίστοιχα, για το δείγμα :

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

2.4.6 Πίνακας Σύγχυσης - Confusion Matrix

Ο πίνακας σύγχυσης (*Confusion Matrix*) είναι ένας πίνακας που χρησιμοποιείται για την απεικόνιση της απόδοσης ενός ταξινομητή σε προβλήματα δυαδικής κατηγοριοποίησης. Απεικονίζει τον αριθμό των σωστών και λανθασμένων προβλέψεων που έκανε το μοντέλο, κατηγοριοποιημένες ανά κλάση [37]. Ο πίνακας έχει την εξής γενική μορφή για δυαδική ταξινόμηση:



Σχήμα 2.8: *Πίνακας απεικόνισης της απόδοσης ενός ταξινομητή σε προβλήματα δυαδικής κατηγοριοποίησης*

- **True Positives (TP)**: Οι περιπτώσεις όπου το μοντέλο προέβλεψε σωστά την θετική κλάση.
- **False Positives (FP)**: Οι περιπτώσεις όπου το μοντέλο προέβλεψε λανθασμένα τη θετική κλάση ενώ η πραγματική ήταν αρνητική.
- **False Negatives (FN)**: Οι περιπτώσεις όπου το μοντέλο προέβλεψε λανθασμένα την αρνητική κλάση ενώ η πραγματική ήταν θετική.
- **True Negatives (TN)**: Οι περιπτώσεις όπου το μοντέλο προέβλεψε σωστά την αρνητική κλάση.

Ο Confusion Matrix είναι χρήσιμος για τον υπολογισμό πολλών μετρικών απόδοσης, όπως:

- **Accuracy**:
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$
- **Precision**:
$$\text{Precision} = \frac{TP}{TP + FP}$$
- **Recall (ή Sensitivity)**:
$$\text{Recall} = \frac{TP}{TP + FN}$$
- **F1 Score**:
$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Ο Confusion Matrix προσφέρει μια ολοκληρωμένη εικόνα της απόδοσης ενός ταξινομητή, καθώς περιλαμβάνει όλες τις κατηγορίες λαθών (FP, FN) και σωστών προβλέψεων (TP, TN). Είναι ένα κρίσιμο εργαλείο για την ανάλυση της συμπεριφοράς ενός μοντέλου ταξινόμησης, ειδικά όταν χρησιμοποιείται μαζί με άλλες μετρικές απόδοσης.

Κεφάλαιο 3

Παρουσίαση Συνόλου Δεδομένων

Στο κεφάλαιο αυτό παρουσιάζονται οι πηγές των δικαστικών αποφάσεων που αποτέλεσαν το σύνολο δεδομένων της παρούσας διπλωματικής εργασίας και αναλύεται η διαδικασία συλλογής και επεξεργασίας των δικαστικών αποφάσεων που αποτέλεσαν το σύνολο δεδομένων της διπλωματικής εργασίας. Πιο συγκεκριμένα, περιγράφονται οι διαδικασίες προεπεξεργασίας των δεδομένων, καθώς και η διαδικασία επισημείωσης των αποφάσεων, με σκοπό την δυαδική κατηγοριοποίησή τους ως ¨αποδοχή¨ ή ¨απόρριψη¨. Επιπροσθέτως, παρουσιάζονται τα στατιστικά στοιχεία του συνόλου δεδομένων.

3.1 Πηγές Δικαστικών Αποφάσεων

Οι δικαστικές αποφάσεις που χρησιμοποιήθηκαν στην παρούσα εργασία συλλέχθηκαν από το Εφετείο Πειραιώς [1] και από τον Άρειο Πάγο [2]. Πρόκειται για αποφάσεις που εκδόθηκαν κατά τα έτη 2009, 2018, 2021 και 2022 από τα συγκεκριμένα δικαστήρια και καλύπτουν διάφορους τομείς του δικαίου.

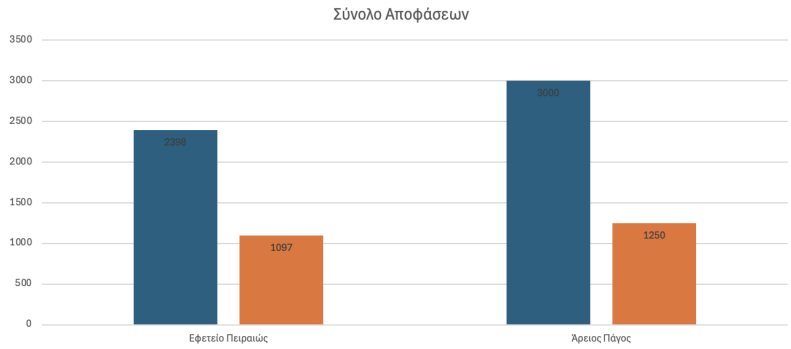
Η επιλογή του Εφετείου Πειραιώς και του Αρείου Πάγου για τη συλλογή των δικαστικών αποφάσεων δεν ήταν τυχαία. Αφενός, τα δύο αυτά δικαστήρια διαθέτουν τις αποφάσεις τους σε εύκολα προσβάσιμη και οργανωμένη μορφή μέσω των επίσημων ιστοσελίδων τους, διευκολύνοντας έτσι τη διαδικασία συλλογής. Αφετέρου, η συστηματική ταξινόμηση των αποφάσεων, σε συνδυασμό με τη θεματική ποικιλία που καλύπτουν, καθιστούν το παραγόμενο σύνολο δεδομένων αντιπροσωπευτικό για διάφορους τομείς του δικαίου.

Επιπλέον, η διαθεσιμότητα αποφάσεων από διαφορετικά έτη επιτρέπει τη μελέτη πιθανών χρονικών μεταβολών στις νομικές αποφάσεις, συμβάλλοντας στην πληρότητα και τη χρησιμότητα του συνόλου δεδομένων. Τόσο η προεπεξεργασία των δεδομένων αλλά και η επισημείωσή τους ήταν απαραίτητες διαδικασίες προκειμένου να δημιουργηθεί η τελική μορφή του συνόλου δεδομένων προς μελέτη. Οι λεπτομέρειες των διαδικασιών αυτών θα αναλυθούν παρακάτω.

Στο αρχικό dataset, ο αριθμός των αποφάσεων δεν ήταν ισομερής μεταξύ των κατηγοριών, γεγονός που θα μπορούσε να επηρεάσει αρνητικά την απόδοση των μοντέλων μηχανικής μάθησης. Συγκεκριμένα, η ύπαρξη ανισορροπίας στις κλάσεις μπορεί να οδηγήσει σε προκατειλημμένα μοντέλα που ευνοούν την επικρατέστερη κατηγορία, μειώνοντας έτσι

[1] (https://www.efeteio-peir.gr/?page_id=4017)
[2] (https://www.areiospagos.gr/nomologia/apofaseis.asp)

τη γενίκευση των προβλέψεων. Για την αντιμετώπιση αυτού του ζητήματος, εφαρμόστηκε μια διαδικασία εξισορρόπησης, ώστε να δημιουργηθεί ένα πιο αντιπροσωπευτικό dataset. Η παρακάτω γραφική απεικόνιση παρουσιάζει τη σύγκριση μεταξύ του αρχικού και του εξισορροπημένου συνόλου δεδομένων.



Σχήμα 3.1: *Κατανομή αποφάσεων πριν και μετά την εξισορρόπηση του συνόλου δεδομένων*

- Οι *μπλε* γραμμές αντιστοιχούν στον αρχικό αριθμό αποφάσεων που υπήρχαν στο dataset πριν τη διαδικασία εξισορρόπησης.

- Οι *πορτοκαλί* γραμμές δείχνουν τον αριθμό αποφάσεων μετά τη διαδικασία εξισορρόπησης, η οποία εφαρμόστηκε για τη δημιουργία ενός balanced dataset, γεγονός που βελτιώνει την αξιοπιστία των αποτελεσμάτων κατά την εφαρμογή μοντέλων μηχανικής μάθησης [10].

Ο παρακάτω πίνακας παρουσιάζει τα βασικά χαρακτηριστικά του συνόλου δεδομένων που χρησιμοποιήθηκε στη παρούσα διπλωματική εργασία, για τα δύο δικαστήρια που μελετήθηκαν: τον Άρειο Πάγο και το Εφετείο Πειραιώς.

- **Αριθμός Αποφάσεων:** Ο Άρειος Πάγος περιλαμβάνει 1250 αποφάσεις, ενώ το Εφετείο Πειραιώς 1197 αποφάσεις.

- **Μέσος Όρος Λέξεων:** Ο μέσος όρος λέξεων των αποφάσεων των δύο δικαστηρίων. Στις αποφάσεις του Αρείου Πάγου ο μέσος όρος λέξεων είναι 2470, ενώ στις αποφάσεις του Εφετείου Πειραιώς είναι 4361.

- **Αποδοχή/Απόρριψη:** Στον Άρειο Πάγο, 500 αποφάσεις επισημάνθηκαν -με τη διαδικασία της επισημείωσης όπως αναλύεται στην Ενότητα 3.4- ως «Αποδοχή» και 750 ως «Απόρριψη» ενώ για το Εφετείο Πειραιώς, 497 αποφάσεις επισημάνθηκαν ως «Αποδοχή» και 700 ως «Απόρριψη».

Ο πίνακας αναδεικνύει διαφοροποιήσεις στο μέγεθος και στη δομή των αποφάσεων ανάμεσα στα δύο δικαστήρια μετά την διαδικασία της προεπεξεργασίας και της επισημείωσης που αναλύονται παρακάτω.

Χαρακτηριστικό	**Άρειος Πάγος**	**Εφετείο Πειραιώς**
Αριθμός Αποφάσεων	1250	1197
Μ.Ο λέξεων	2470	4361
Αποδοχή	500	497
Απόρριψη	750	700

Πίνακας 3.1: *Συνοπτικός πίνακας χαρακτηριστικών του συνόλου δεδομένων*

3.2 Δομή δικαστικών αποφάσεων

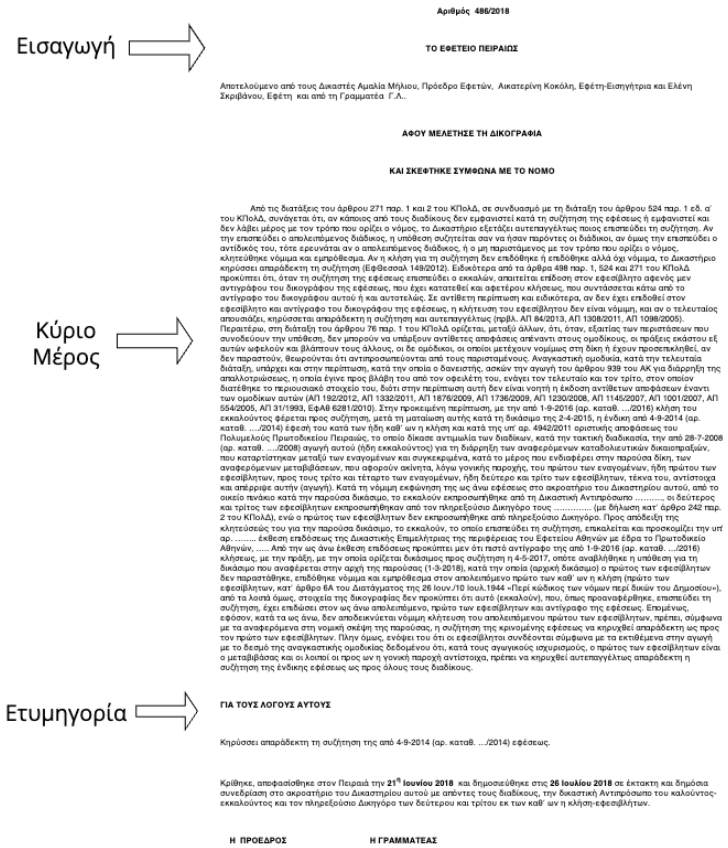
Για την καλύτερη κατανόηση των διαδικασιών που πραγματοποιήθηκαν στην διπλωματική εργασία είναι χρήσιμο να εξετάσουμε ποιά είναι η δομή μιας δικαστικής απόφασης, καθώς αποτελεί κρίσιμο στοιχείο για την ανάλυση και την πρόβλεψη της έκβασής τους [38]. Συνήθως, οι δικαστικές αποφάσεις περιλαμβάνουν τυποποιημένα τμήματα, όπως τον πρόλογο, όπου αναφέρονται οι διάδικοι, το νομικό πλαίσιο και η περιγραφή της υπόθεσης, το σκεπτικό, το οποίο παρουσιάζει την ανάλυση του δικαστηρίου, τους εφαρμοζόμενους νόμους και τη νομική επιχειρηματολογία και το διατακτικό, όπου καταγράφεται η τελική απόφαση και οι συνέπειές της.

Η συγκεκριμένη δομή καθιστά τις αποφάσεις ιδανικές για ανάλυση μέσω επεξεργασίας κειμένου, καθώς τα διαφορετικά μέρη μπορούν να απομονωθούν και να επισημανθούν, όπως αναλύεται στις ενότητες 3.3 και 3.4 διευκολύνοντας την κατηγοριοποίηση και την πρόβλεψη. Η τυποποίηση αυτή συμβάλλει στη δημιουργία συνόλων δεδομένων που ενισχύουν την εκπαίδευση αλγορίθμων μηχανικής μάθησης, υποστηρίζοντας την εξαγωγή χρήσιμων μοτίβων για τη δικαστική λογική και την πρόβλεψη της *θετικής* ή *αρνητικής* κατηγορίας, όπως ορίζεται στη μελέτη.

Μια δικαστική απόφαση, περιλαμβάνει την εισαγωγή, όπου αναφέρονται τα στοιχεία των διαδίκων, το δικαστήριο που εξέδωσε την απόφαση, οι δικαστές που συμμετείχαν στη σύνθεση του δικαστηρίου, καθώς και η ημερομηνία έκδοσης της απόφασης. Επίσης, περιγράφεται η φύση της υπόθεσης και αναφέρονται οι βασικές διατάξεις του νόμου που σχετίζονται με την υπόθεση. Στη συνέχεια, το κύριο μέρος περιλαμβάνει την περιγραφή των γεγονότων που οδήγησαν στη δικαστική διαμάχη, τις θέσεις των αντιδίκων και τα νομικά επιχειρήματα που προβάλλονται από κάθε πλευρά. Στο σημείο αυτό ενδέχεται να γίνεται και αναφορά σε προηγούμενες σχετικές αποφάσεις, που μπορεί να έχουν ρόλο στη νομική ερμηνεία της υπόθεσης. Το σκεπτικό αποτελεί την ανάλυση του δικαστηρίου και περιλαμβάνει τους νομικούς κανόνες που εφαρμόζονται στην υπόθεση, καθώς και τον τρόπο με τον οποίο αυτοί ερμηνεύονται. Εδώ, το δικαστήριο εξετάζει τα αποδεικτικά στοιχεία, αξιολογεί τη νομική βάση των ισχυρισμών των διαδίκων και εξηγεί τη λογική πίσω από την απόφασή του. Το σκεπτικό είναι ιδιαίτερα σημαντικό, καθώς αποτελεί τη νομική θεμελίωση της απόφασης και καθορίζει τη δυνατότητα άσκησης έφεσης ή αναίρεσης. Τέλος, μια δικαστική απόφαση περιλαμβάνει την ετυμηγορία, δηλαδή το τελικό και πλέον δεσμευτικό τμήμα της απόφασης, στο οποίο καταγράφεται η απόφαση του δικαστηρίου. Περιλαμβάνει το τελικό αποτέλεσμα, δηλαδή αν η αγωγή γίνεται δεκτή ή απορρίπτεται, καθώς και τυχόν επιβληθείσες κυρώσεις, χρηματικές αποζημιώσεις ή άλλα έννομα αποτελέσματα. Στην περίπτωση ανώτερων δικαστηρίων, μπορεί να περιλαμβάνει και οδηγίες προς τα κατώτερα δικαστήρια για την εκτέλεση

της απόφασης.

Στο Σχήμα 3.2, παρατίθενται η δικαστική απόφαση 486/2018 του Εφετείου Πειραιώς, προκειμένου να επισημανθούν τα βασικά μέρης που αποτελούν τη δομή μιας δικαστικής απόφασης, από το σύνολο δεδομένων που μελετάμε στην παρούσα διπλωματική εργασία :



Σχήμα 3.2: *Δικαστική απόφαση 486/2018 όπου διαφαίνονται τα βασικά μέρης μιας δικαστικής απόφασης του Εφ.Πειραιώς*

3.3 Προεπεξεργασία Δεδομένων

Ο πρωταρχικός στόχος της προεπεξεργασίας των δεδομένων είναι να προετοιμάσουμε το κείμενο, έτσι ώστε να διευκολύνουμε την διαδικασία ανάλυσης και την επεξεργασία τους. Οι συγκεκριμένες ενέργειες είναι απαραίτητες με σκοπό να φέρουμε τις δικαστικές αποφάσεις σε μορφή κατάλληλη για τα μοντέλα που θα εξετάσουμε στην συνέχεια. Η διαδικασία προεπεξεργασίας των κειμένων είναι ένα κρίσιμο βήμα στην προετοιμασία των δεδομένων για τη χρήση τους σε αλγορίθμους μηχανικής μάθησης. Για την προεπεξεργασία των δικαστικών αποφάσεων, ακολουθήσαμε μια σειρά από βήματα που στοχεύουν στην αργότερα αποτελεσματική ανάλυση των κειμένων από τα μοντέλα.

1. *Ανάκτηση κειμένου :* Τα δεδομένα που χρησιμοποιήθηκαν προήλθαν από δικαστικές αποφάσεις αποθηκευμένες σε αρχεία PDF και HTML. Για την εξαγωγή του κειμένου από τα αρχεία HTML, χρησιμοποιήθηκε το εργαλείο BeautifulSoup, το οποίο επιτρέπει

την ανάκτηση του πλήρους περιεχομένου των αποφάσεων. Αντίστοιχα, για τα αρχεία PDF και CSV, χρησιμοποιήθηκε η Python βιλβιοθήκη πανδας προκειμένουν να γίνει η εξαγωγή του κειμένου αγνοώντας tags κι άλλες δομικές πληροφορίες που περιέχονται στα αρχεία. Αυτή η διαδικασία εξασφάλισε ότι το κείμενο εξάγεται με συνέπεια και ακρίβεια, ανεξάρτητα από την πηγή του.

2. *Μετατροπή σε πεζά :* Όλα τα γράμματα μετατράπηκαν σε πεζά για να εξασφαλιστεί η συνέπεια και να αποφεύγεται η διάκριση μεταξύ κεφαλαίων και πεζών χαρακτήρων, που δεν θα πρόσθεταν κάποια αξία στην ανάλυση.

3. *Αφαίρεση τονισμού :* Οι τόνοι αφαιρέθηκαν από τις λέξεις, διευκολύνοντας έτσι την ταύτιση όρων με και χωρίς τόνο, όπως «δικαστής» και «δικαστης», τα οποία θα αντιμετωπίζονταν ως διαφορετικές λέξεις από τον αλγόριθμο.

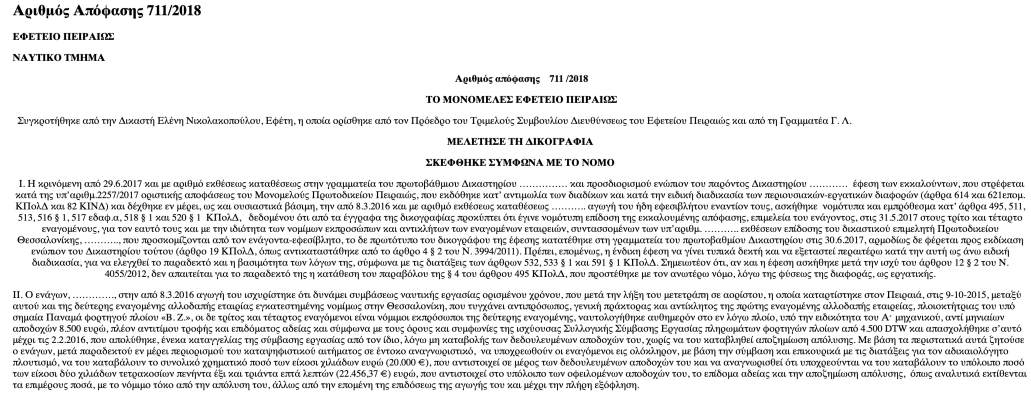
4. *Αφαίρεση σημείων στίξης :* Τα σημεία στίξης αφαιρέθηκαν, καθώς δεν προσφέρουν πληροφορίες χρήσιμες για την εκπαίδευση των μοντέλων πρόβλεψης. Αυτό περιλαμβάνει όλα τα σημεία στίξης, όπως κόμματα, τελείες, ερωτηματικά κ.λπ..

5. *Αφαίρεση ειδικών χαρακτήρων :* Αφαιρέθηκαν ειδικοί χαρακτήρες όπως η κάτω παύλα, που δεν προσθέτουν νόημα στο κείμενο και μπορεί να προκαλέσουν προβλήματα στη διαδικασία ανάλυσης.

6. *Αφαίρεση λέξεων-κλειδιά :* Αφαιρέθηκαν λέξεις-κλειδιά, δηλαδή συχνές λέξεις οι οποίες δεν φέρουν σημαντική σημασιολογική πληροφορία, με χρήση της λίστας που παρέχεται από το NLTK.[3]

Για να διευκολύνουμε την κατανόηση της διαδικασίας που ακολουθήσαμε στην προεπεξεργασία και την επισημείωση των δικαστικών αποφάσεων, παραθέτουμε μια σειρά από στιγμιότυπα οθόνης (screenshots) που απεικονίζουν τα βασικά στάδια της διαδικασίας. Αρχικά, παρουσιάζουμε ένα τμήμα μιας δικαστικής απόφασης στην αρχική της, ακατέργαστη (*raw*) μορφή, όπως συλλέχθηκε από τις διαθέσιμες πηγές. Στη συνέχεια, απεικονίζουμε το ίδιο κείμενο μετά την εφαρμογή των τεχνικών προεπεξεργασίας, όπως η αφαίρεση ειδικών χαρακτήρων, η μετατροπή του κειμένου σε πεζά και η αφαίρεση stopwords. Η οπτική αυτή αναπαράσταση συμβάλλει στην καλύτερη κατανόηση της μεθοδολογίας που εφαρμόσαμε και αναδεικνύει τη σημασία κάθε επιμέρους βήματος στη διαμόρφωση των δεδομένων για τη διαδικασία πρόβλεψης.

[3](https://github.com/hb20007/hands-on-nltk-tutorial/blob/main/7-1-NLTK-with-the-Greek-Script.ipynb)



Αριθμός Απόφασης 711/2018

ΕΦΕΤΕΙΟ ΠΕΙΡΑΙΩΣ

ΝΑΥΤΙΚΟ ΤΜΗΜΑ

Αριθμός απόφασης 711 /2018

ΤΟ ΜΟΝΟΜΕΛΕΣ ΕΦΕΤΕΙΟ ΠΕΙΡΑΙΩΣ

Συγκροτήθηκε από την Δικαστή Ελένη Νικολακοπούλου, Εφέτη, η οποία ορίσθηκε από τον Πρόεδρο του Τριμελούς Συμβουλίου Διευθύνσεως του Εφετείου Πειραιώς και από τη Γραμματέα Γ. Λ.

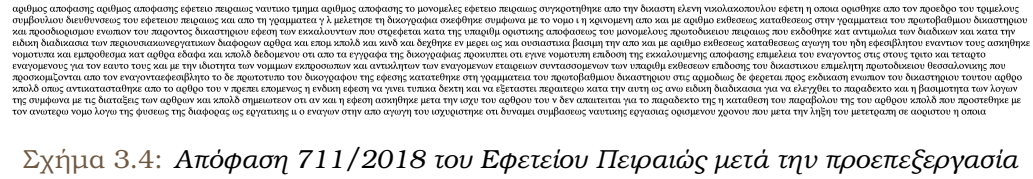
ΜΕΛΕΤΗΣΕ ΤΗ ΔΙΚΟΓΡΑΦΙΑ

ΣΚΕΦΘΗΚΕ ΣΥΜΦΩΝΑ ΜΕ ΤΟ ΝΟΜΟ

Ι. Η κρινόμενη από 29.6.2017 και με αριθμό εκθέσεως καταθέσεως στην γραμματεία του πρωτοβάθμιου Δικαστηρίου …… και προσδιορισμού ενώπιον του παρόντος Δικαστηρίου …… έφεση των εκκαλούντων, που στρέφεται κατά της υπ'αριθμ.2257/2017 οριστικής αποφάσεως του Μονομελούς Πρωτοδικείου Πειραιώς, που εκδόθηκε κατ' αντιμωλία των διαδίκων και κατά την ειδική διαδικασία των περιουσιακών-εργατικών διαφορών (άρθρα 614 και 621επομ. ΚΠολΔ και 82 ΚΙΝΔ) και δέχθηκε εν μέρει, ως και ουσιαστικά βάσιμη, την από 8.3.2016 και με αριθμό εκθέσεως καταθέσεως ……. αγωγή του ήδη εφεσιβλήτου εναντίον τους, ασκήθηκε νομότυπα και εμπρόθεσμα κατ' άρθρα 495, 511, 513, 516 § 1, 517 εδαφ.α, 518 § 1 και 520 § 1 ΚΠολΔ, δεδομένου ότι από τα έγγραφα της δικογραφίας προκύπτει ότι έγινε νομότυπη επίδοση της εκκαλουμένης απόφασης, επιμελεία του ενάγοντος, στις 31.5.2017 στους τρίτο και τέταρτο εναγομένους, για τον εαυτό τους και με την ιδιότητα των νομίμων εκπροσώπων και αντιπλήκτων των εναγομένων εταιρειών, συνταισσομένων των υπ'αριθμ. …… εκθέσεων επίδοσης του δικαστικού επιμελητή Πρωτοδικείου Θεσσαλονίκης, ……, που προσκομίζονται από τον ενάγοντα-εφεσίβλητο, το δε πρωτότυπο του δικογράφου της έφεσης κατατέθηκε στη γραμματεία του πρωτοβαθμίου Δικαστηρίου στις 30.6.2017, αρμοδίως δε φέρεται προς εκδίκαση ενώπιον του Δικαστηρίου τούτου (άρθρο 19 ΚΠολΔ, όπως αντικαταστάθηκε από το άρθρο 4 § 2 του Ν. 3994/2011). Πρέπει, επομένως, η ένδικη έφεση να γίνει τυπικά δεκτή και να εξεταστεί περαιτέρω κατά την αυτή ως άνω ειδική διαδικασία, για να ελεγχθεί το παραδεκτό και η βασιμότητα των λόγων της, σύμφωνα με τις διατάξεις των άρθρων 532, 533 § 1 και 591 § 1 ΚΠολΔ. Σημειωτέον ότι, αν και η έφεση ασκήθηκε μετά την ισχύ του άρθρου 12 § 2 του Ν. 4055/2012, δεν απαιτείται για το παραδεκτό της η κατάθεση του παραβόλου της § 4 του άρθρου 495 ΚΠολΔ, που προστέθηκε με τον ανωτέρω νόμο, λόγω της φύσεως της διαφοράς, ως εργατικής.

ΙΙ. Ο ενάγων, ……, στην από 8.3.2016 αγωγή του ισχυρίστηκε ότι δυνάμει συμβάσεως ναυτικής εργασίας ορισμένου χρόνου, που μετά την λήξη του μετετράπη σε αορίστου, η οποία καταρτίστηκε στον Πειραιά, στις 9-10-2015, μεταξύ αυτού και της δεύτερης εναγομένης αλλοδαπής εταιρίας εγκατεστημένης νομίμως στην Θεσσαλονίκη, που τυγχάνει αντιπρόσωπος, γενική πράκτορας και αντίκλητος της πρώτης εναγομένης αλλοδαπής εταιρείας, πλοιοκτήτριας του υπό σημαία Παναμά φορτηγού πλοίου «Β. Ζ.», οι δε τρίτος και τέταρτος εναγόμενοι είναι νόμιμοι εκπρόσωποι της δεύτερης εναγομένης, ναυτολογήθηκε αυθημερόν στο εν λόγω πλοίο, υπό την ειδικότητα του Α' μηχανικού, αντί μηνιαίων αποδοχών 8.500 ευρώ, πλέον αντιτίμου τροφής και επιδόματος αδείας και σύμφωνα με τους όρους και συμφωνίες της ισχύουσας Συλλογικής Σύμβασης Εργασίας πληρωμάτων φορτηγών πλοίων από 4.500 DTW και απασχολήθηκε σ' αυτό μέχρι τις 2.2.2016, που απολύθηκε, ένεκα καταγγελίας της σύμβασης εργασίας από τον ίδιο, λόγω μη καταβολής των δεδουλευμένων αποδοχών του, χωρίς να του καταβληθεί αποζημίωση απόλυσης. Με βάση τα περιστατικά αυτά ζητούσε ο ενάγων, μετά παραδεκτού εν μέρει περιορισμού του καταψηφιστικού αιτήματος σε έντοκο αναγνωριστικό, να υποχρεωθούν οι εναγόμενοι εις ολόκληρον, με βάση την σύμβαση και επικουρικά με τις διατάξεις για τον αδικαιολόγητο πλουτισμό, να του καταβάλουν το συνολικό χρηματικό ποσό των είκοσι χιλιάδων ευρώ (20.000 €), που αντιστοιχεί σε μέρος των δεδουλευμένων αποδοχών του και να αναγνωρισθεί ότι υποχρεούνται να του καταβάλουν το υπόλοιπο ποσό των είκοσι δύο χιλιάδων τετρακοσίων πενήντα έξι και τριάντα επτά λεπτών (22.456,37 €) ευρώ, που αντιστοιχεί στο υπόλοιπο των οφειλομένων αποδοχών του, το επίδομα αδείας και την αποζημίωση απόλυσης, όπως αναλυτικά εκτίθενται τα επιμέρους ποσά, με το νόμιμο τόκο από την απόλυση του, άλλως από την επομένη της επιδόσεως της αγωγής του και μέχρι την πλήρη εξόφληση.

Σχήμα 3.3: *Απόφαση 711/2018 του Εφετείου Πειραιώς σε ακατέργαστη μορφή*



αριθμος αποφασης αριθμος αποφασης εφετειο πειραιως ναυτικο τμημα αριθμος αποφασης το μονομελες εφετειο πειραιως συγκροτηθηκε απο την δικαστη ελενη νικολακοπουλου εφετη η οποια ορισθηκε απο τον προεδρο του τριμελους συμβουλιου διευθυνσεως του εφετειου πειραιως και απο τη γραμματεα γ λ μελετησε τη δικογραφια σκεφθηκε συμφωνα με το νομο ι η χρινομενη απο και με αριθμο εκθεσεως καταθεσεως στην γραμματεια του πρωτοβαθμιου δικαστηριου και προσδιορισμου ενωπιον του παροντος δικαστηριου εφεση των εκκαλουντων που στρεφεται κατα της υπαριθμ οριστικης αποφασεως του μονομελους πρωτοδικειου πειραιως που εκδοθηκε κατ αντιμωλια των διαδικων και κατα την ειδικη διαδικασια των περιουσιακωνεργατικων διαφορων αρθρα και επομ κπολδ και κινδ και δεχθηκε εν μερει ως και ουσιαστικα βασιμη την απο και με αριθμο εκθεσεως καταθεσεως αγωγη του ηδη εφεσιβλητου εναντιον τους ασκηθηκε νομοτυπα και εμπροθεσμα κατ αρθρα εδαφα και κπολδ δεδομενου οτι απο τα εγγραφα της δικογραφιας προκυπτει οτι εγινε νομοτυπη επιδοση της εκκαλουμενης αποφασης επιμελεια του εναγοντος στις στους τριτο και τεταρτο εναγομενους για τον εαυτο τους και με την ιδιοτητα των νομιμων εκπροσωπων και αντικλητων των εναγομενων εταιρειων συνταισσομενων των υπαριθμ εκθεσεων επιδοσης του δικαστικου επιμελητη πρωτοδικειου θεσσαλονικης που προσκομιζονται απο τον εναγονταεφεσιβλητο το δε πρωτοτυπο του δικογραφου της εφεσης κατατεθηκε στη γραμματεια του πρωτοβαθμιου δικαστηριου στις αρμοδιως δε φερεται προς εκδικαση ενωπιον του δικαστηριου τουτου αρθρο κπολδ οπως αντικατασταθηκε απο το αρθρο του ν πρεπει επομενως η ενδικη εφεση να γινει τυπικα δεκτη και να εξεταστει περαιτερω κατα την αυτη ως ανω ειδικη διαδικασια για να ελεγχθει το παραδεκτο και η βασιμοτητα των λογων της συμφωνα με τις διαταξεις των αρθρων και κπολδ σημειωτεον οτι αν και η εφεση ασκηθηκε μετα την ισχυ του αρθρου του ν δεν απαιτειται για το παραδεκτο της η καταθεση του παραβολου της του αρθρου κπολδ που προστεθηκε με τον ανωτερω νομο λογω της φυσεως της διαφορας ως εργατικης ι ο εναγων στην απο αγωγη του ισχυριστηκε οτι δυναμει συμβασεως ναυτικης εργασιας ορισμενου χρονου που μετα την ληξη του μετετραπη σε αοριστου η οποια

Σχήμα 3.4: *Απόφαση 711/2018 του Εφετείου Πειραιώς μετά την προεπεξεργασία*

3.4 Διαδικασία Επισημείωσης

Μετά την ολοκλήρωση της προεπεξεργασίας των κειμένων των δικαστικών αποφάσεων, προχωρήσαμε στην φάση της επισημείωσης των δεδομένων, προκειμένου να κατηγοριοποιήσουμε της αποφάσεις δυαδικά, δηλαδή ως αποδοχή ή απόρριψη. Η επισημείωση είναι ένα κρίσιμο βήμα στη διαδικασία ανάλυσης δεδομένων, ιδιαίτερα όταν χρησιμοποιούνται τεχνικές μηχανικής μάθησης. Η ακρίβεια της επισημείωσης ενός συνόλου δεδομένων, επηρεάζει άμεσα την απόδοση των μοντέλων πρόβλεψης που θα εκπαιδευτούν πάνω σε αυτά τα δεδομένα. Στην περίπτωση των δικαστικών αποφάσεων, το σωστό (*labeling*) των δεδομένων είναι καθοριστικό για την ανάπτυξη αξιόπιστων αλγορίθμων που μπορούν να βοηθήσουν στη βελτίωση της δικαστικής διαδικασίας. Η επισημείωση βασίστηκε στην ύπαρξη συγκεκριμένων (*regular expressions*) που χρησιμοποιούνται από τα δύο δικαστήρια, οι οποίες υποδηλώνουν την αποδοχή της αίτησης ή της έφεσης. Η διαδικασία που ακολουθήσαμε αναλύεται παρακάτω :

1. *Αναζήτηση στόχων - target words :* Προκειμένου να γίνει ορθή κατηγοριοποίηση των αποφάσεων που εξετάζουμε, σε προγραμματιστικό επίπεδο, καθορίσαμε μια λίστα από (*regular expressions*), η οποία χρησιμοπείται συνήθως σε αποφάσεις που καταλήγουν σε αποδοχή. Οι εκφράσεις αυτές, π.χ. «δέχεται τυπικά και κατ' ουσίαν» ή «δέχεται τυπικά και ουσιαστικά», επιλέχθηκαν με βάση την ανάλυση της γλώσσας που χρησιμοποιείται στα δικαστικά κείμενα που εξετάζουμε και αντιστοιχούν σε περιπτώσεις όπου το δικαστήριο κάνει αποδεκτή την αίτηση ή την έφεση. Οι αποφάσεις που περιείχαν αυτές τις φράσεις επισημάνθηκαν ως αποδοχή (με την ένδειξη 0), ενώ οι υπόλοιπες επισημάνθηκαν ως απορρίψη (με την ένδειξη 1).

2. *Δημιουργία συνόλου δεδομένων :* Μετά την αναζήτηση των λέξεων-στόχων, οι αποφάσεις επισημάνθηκαν και το αποτέλεσμα αποθηκεύτηκε σε ένα δομημένο σύνολο δεδομένων

(CSV αρχείο), το οποίο περιέχει για κάθε απόφαση το όνομα του αρχείου και την αντίστοιχη κατηγορία στην οποία ανήκει.

Με το πέρας της διαδικασίας της επισημείωσης, το σύνολο των δικαστικών αποφάσεων είναι πλέον έτοιμο για την επόμενη φάση της μελέτης μας, όπου θα εφαρμοστούν τεχνικές μηχανικής μάθησης για την εξαγωγή προβλέψεων σχετικά με την έκβαση μελλοντικών υποθέσεων.

Το διάγραμμα 3.5 αποτυπώνει τη μεθοδολογία που έχουμε ακολουθήσει μέχρι στιγμής στην εργασία μας. Ξεκινάμε από τις αρχικές δικαστικές αποφάσεις (Raw Decisions), στη συνέχεια προχωράμε στην προεπεξεργασία των δεδομένων (Preprocess), και τέλος στην επισημείωση (Annotation).

Σχήμα 3.5: *Διαδικασία από τις ακατέργαστες δικαστικές αποφάσεις έως την επισημείωση*

Κεφάλαιο 4

Προετοιμασία Δεδομένων και Ρύθμιση Υπερπαραμέτρων

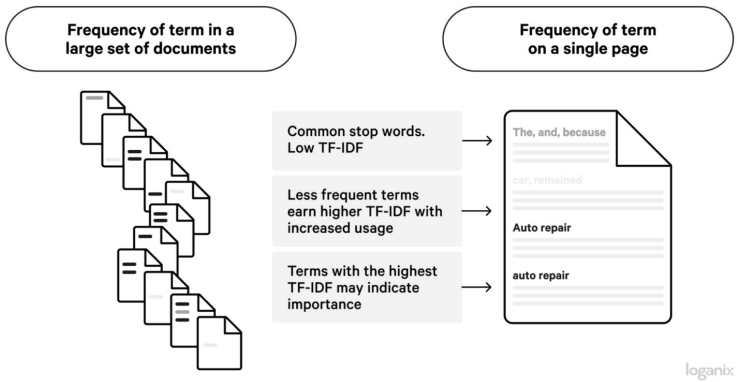
Στο παρόν κεφάλαιο περιγράφεται η απαραίτητη προετοιμασία των κειμένων των δικαστικών αποφάσεων, προκειμένου να αποκτήσουν μορφή κατάλληλη για να εξετάσουμε την απόδοση των διαφόρων μοντέλων μηχανικής μάθησης. Επιπλέον, παρουσιάζεται αναλυτικά η διαδικασία ρύθμισης των υπερπαραμέτρων των ταξινομητών που μελετήθηκαν, καθώς και τα αποτελέσματα αυτής.

4.1 Προετοιμασία Κειμένων

Προκειμένου να εξετάσουμε την αποτελεσματικότητα των διαφόρων ταξινομητών στο σύνολο των αποφάσεων που έχουμε στη διάθεσή μας ήταν απαραίτητο να αναπαραστίσουμε τα κείμενα των αποφάσεων σε αριθμητική μορφή. Η τεχνική (NLP - Natural Language Processing) που επιλέξαμε για την παρούσα εργασία είναι το TF-IDF (*Term Frequency-Inverse Document Frequency*). Το TF-IDF είναι μία από τις πιο διαδεδομένες και αποτελεσματικές τεχνικές για την αναπαράσταση αυτή, στην Επεξεργασία Φυσικής Γλώσσας (*NLP*) [39]. Το TF-IDF είναι ουσιαστικά μια αριθμητική στατιστική που προορίζεται να αντικατοπτρίζει τη σημασία μιας λέξης για ένα έγγραφο σε μια συλλογή ή ένα σώμα κειμένων. Πιο συγκεκριμένα, η τεχνική που εφαρμόσαμε στις αποφάσεις συνδυάζει δύο βασικές έννοιες: τη συχνότητα εμφάνισης μιας λέξης σε ένα έγγραφο (Term Frequency) και τη σπανιότητα αυτής της λέξης στο σύνολο των εγγράφων (Inverse Document Frequency). Αναλυτικότερα, η συνάρτηση TF μετρά πόσες φορές μια λέξη εμφανίζεται σε ένα έγγραφο σε σχέση με το συνολικό αριθμό λέξεων, ενώ η συνάρτηση IDF μειώνει τη βαρύτητα των όρων που εμφανίζονται σε πολλά έγγραφα, καθώς αυτοί δεν είναι τόσο διακριτικοί. Ο πολλαπλασιασμός των δύο αυτών μεγεθών οδηγεί σε μια μετρική που αναδεικνύει τους πιο ¨σημαντικούς¨ όρους για κάθε έγγραφο. Το TF-IDF χρησιμοποιείται ευρέως σε συστήματα ανάκτησης πληροφορίας και ταξινομήσεις κειμένων.

Στην παρούσα διπλωματική εργασία χρησιμοποιείται το αντικείμενο Tf-idfVectorizer(), το οποίο δέχεται ως είσοδο κείμενο και εξάγει διανύσματα (vectors) σε ένα μοντέλο διανυσματικού χώρου. Το αντικείμενο αυτό ανάλογα με τα ορίσματα που δέχεται διαχειρίζεται και διαφορετικά τα δεδομένα. Οι μεταβλητές maxdf και mindf, που αποτελούν επίσης παραμέτρους του Tf-idfVectorizer(), παίζουν εξίσου σημαντικό ρόλο και βοηθούν στη μείωση των

διαστάσεων κάθε μοντέλου, καθώς μπορούν ανάλογα με τις τιμές που θα λάβουν να περιορίσουν το εύρος του λεξιλογίου που δημιουργείται.. Όσον αφορά στο maxdf, με την τιμή 0,5 δηλώνεται ότι πρόκειται να αγνοηθούν όλοι οι όροι που εμφανίζονται σε πάνω από το 50 τοις εκατό των δεδομένων, ενώ σχετικά με την τιμή του mindf δηλώνεται ότι δεν θα ληφθούν υπόψη οι όροι που υπάρχουν σε λιγότερο από 10 έγγραφα.



Σχήμα 4.1: *Αναπαράσταση TF-IDF*

4.2 Ρύθμιση Υπερπαραμέτρων των Μοντέλων

Η σωστή λειτουργία πολλών από τους αλγορίθμους μηχανικής μάθησης, βασίζεται στη σωστή ρύθμιση των παραμέτρων τους. Οι υπερπαράμετροι (hyperparameters) συνιστούν τις μεταβλητές που χαρακτηρίζουν τη διαδικασία εκπαίδευσης του εκάστοτε μοντέλου. Οι τιμές τους πρέπει να ρυθμιστούν από τον προγραμματιστή προτού αρχίσει η διαδικασία εκπαίδευσης, εν αντιθέσει με τις απλές παραμέτρους του μοντέλου, η τιμή των οποίων υπολογίζεται αυτομάτως -κατά την εκπαίδευση- από το ίδιο το μοντέλο. Ως εκ τούτου, γίνεται κατανοητή η σημασία της επιλογής των κατάλληλων υπερπαραμέτρων για την αποδοτική λειτουργία κάθε μοντέλου. Ωστόσο, δεν υπάρχουν σαφείς και ακριβείς κανόνες που να προσδιορίζουν τον τρόπο με τον οποίο πρέπει να γίνει αυτή η επιλογή. Οι αλγόριθμοι που επιτελούν την ρύθμιση των υπερπαραμέτρων λειτουργούν σύμφωνα με τη μέθοδο δοκιμής-σφάλματος, προβαίνοντας σε συνεχόμενες επιλογές τιμών, εως ότου φτάσουν στις βέλτιστες τιμές. Επομένως, σημαντικό βήμα αποτελεί η επιλογή των τιμών εκείνων που θα υποβληθούν σε αυτή τη διαδικασία βελτιστοποίησης.

Υπάρχουν ελάχιστες καθολικές συμβουλές σχετικά με την επιλογή των τιμών αυτών, ενώ η τελική επιτυχία της διαδικασίας εξαρτάται σε μεγάλο βαθμό από την εμπειρία του προγραμματιστή. Η επιλογή τους πρέπει να γίνεται με σύνεση, καθώς κάθε παράμετρος που επιλέγεται να ρυθμιστεί μπορεί να αυξήσει εκθετικά τον απαιτούμενο αριθμό των δοκιμών. Αφού επιλεχθούν οι παράμετροι προς ρύθμιση, εφαρμόζονται αλγόριθμοι, οι οποίοι προελαύνουν το χώρο αναζήτησης που δημιουργείται, το μέγεθος του οποίου εξαρτάται από το πλήθος και το εύρος των υπερπαραμέτρων που έχουν επιλεχθεί να ρυθμιστούν. Οι δύο πλέον χρησιμοποιούμενοι από τους αλγορίθμους αυτούς είναι οι εξής :

1. **Αναζήτηση Πλέγματος** (*Grid Search*). Η αναζήτηση πλέγματος αποτελεί τον απλο-

ύστερο αλγόριθμο βελτιστοποίησης υπερπαραμέτρων. Ο αλγόριθμος αυτός εκτελεί μια εξαντλητική αναζήτηση στο προκαθορισμένο χώρο αναζήτησης που δημιουργείται. Ο χώρος αναζήτησης μπορεί να καταλήξει να αποτελεί ένα υπερεπίπεδο δεκάδων διαστάσεων, ανάλογα με το πλήθος των προς ρύθμιση παραμέτρων.

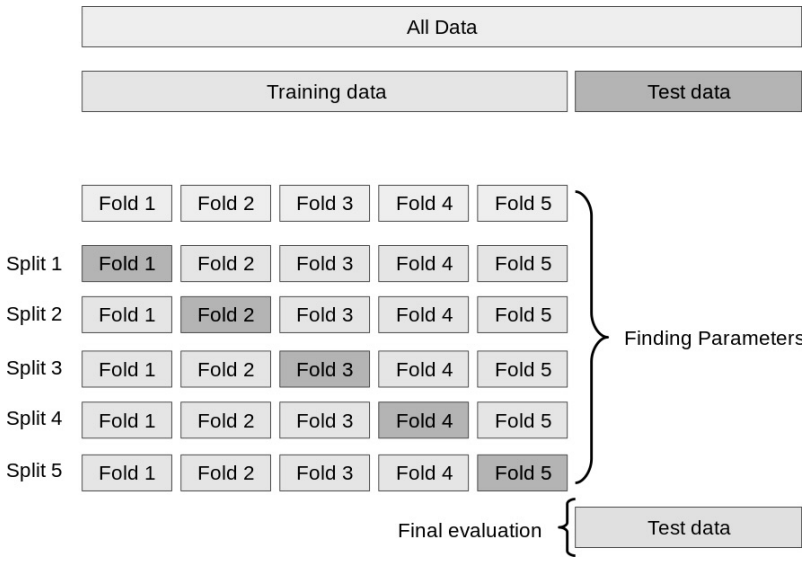
2. **Τυχαία Αναζήτηση** (*Randomized Search*). Στην τυχαία αναζήτηση, ο χώρος αναζήτησης διασχίζεται τυχαία έως ότου ικανοποιηθεί κάποιο κριτήριο τερματισμού, όπως ο αριθμός των επαναλήψεων. Ο αλγόριθμος αυτός δεν εγγυάται την εύρεση της βέλτιστης λύσης, αλλά λειτουργεί ικανοποιητικά σε προβλήματα που το πλήθος των υπερπαραμέτρων είναι μικρό, ενώ επιλέγεται επίσης, όταν δεν είναι διαθέσιμη μεγάλη υπολογιστική ισχύς.

Ένας κύριος λόγος μη αποδοτικότητας ενός μοντέλου μηχανικής μάθησης είναι η υπερπροσαρμογή (*overfitting*). Ο συγκεκριμένος όρος χρησιμοποιείται στην επιβλεπόμενη μάθηση για να δηλώσει την κατάσταση κατά την οποία ένα μοντέλο έχει εκπαιδευτεί και εξειδικευτεί στο σύνολο εκπαίδευσης του προβλήματος που εξετάζεται, με αποτέλεσμα να παρουσιάζει χαμηλή ακρίβεια στην πρόβλεψη του συνόλου δοκιμής.

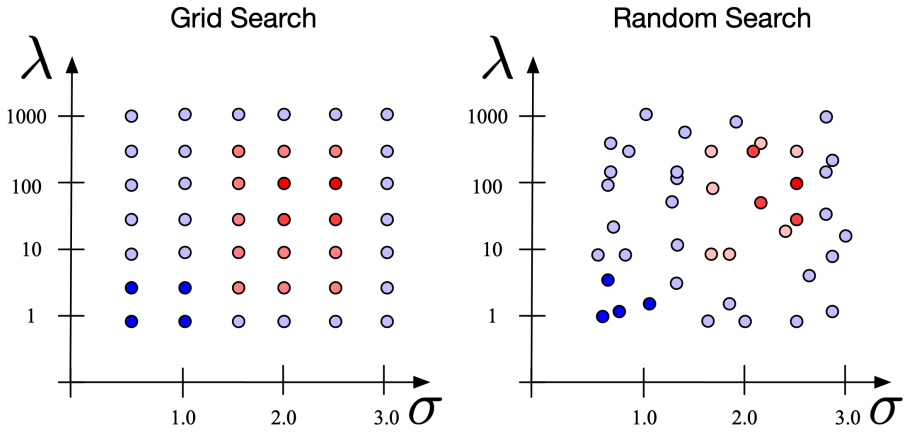
Προκειμένου να αντιμετωπιστεί το παραπάνω πρόβλημα, χρησιμοποιήθηκε η τεχνική *Cross-Validation*(CV), η οποία αποτελεί την πλέον ενδεικνυόμενη λύση. Με την τεχνική αυτή, δεν απαιτείται πλέον η δέσμευση ενός μέρους του συνόλου εκπαίδευσης σα σύνολο αξιολόγησης, με αποτέλεσμα τα μοντέλα να εκπαιδεύονταιμε το μέγιστο δυνατό αριθμό δειγμάτων. Ωστόσο, το σύνολο δοκιμής εξακολουθεί να υπάρχει για την τελική αξιολόγηση των μοντέλων. Η βασική πρακτική της εν λόγω τεχνικής ονομάζεται *k-fold Cross-Validation*. Πιο συγκεκριμένα, επιλέγεται ένας σταθερός αριθμός από *folds (πτυχές)*, δηλαδή συνεχόμενες διαιρέσεις των δεδομένων. Τα δεδομένα διαχωρίζονται σε *k* προσεγγιστικά ίσα folds και κάθε ένα στη συνέχεια θα χρησιμοποιηθεί επαναληπτικά για την αξιολόγηση, ενώ τα υπόλοιπα για την εκπαίδευση των μοντέλων. Τα k-1 folds χρησιμοποιούνται ως σύνολο εκπαίδευσης, ενώ το 1 fold λειτουργεί ως σύνολο αξιολόγησης. Η διαδικασία αυτή επαναλαμβάνεται συνολικά k φορές, διασφαλίζοντας ότι κάθε δείγμα του συνόλου δεδομένων αξιολογείται μία φορά και συμμετέχει k-1 φορές στο σύνολο εκπαίδευσης. Τυπικές τιμές του k είναι της τάξεως του 5 έως 10. Η συνολική αξιολόγηση του μοντέλου προκύπτει από τη μέση τιμή των επιμέρους αξιολογήσεων που προέκυψαν κατά τις k επαναλήψεις. Η διαδικασία αυτή, μπορεί να επαναληφθεί για κάθε τιμή των υπερπαραμέτρων που δοκιμάζονται, ώστε να επιλεχθούν τελικά οι βέλτιστες. Είναι σαφές ότι η πρακτική αυτή έχει μεγάλο υπολογιστικό κόστος, καθώς απαιτούνται πολλοί κύκλοι εκπαίδευσης του μοντέλου. Ωστόσο, η σημασία της έγκειται στο γεγονός ότι δε δεσμεύει μεγάλο μέρος των διαθέσιμων διεγμάτων προς αξιολόγηση, γεγονός θεμελιώδους σημασίας όταν ο αριθμός των δειγμάτων είναι περιορισμένος.

Στην παρούσα εργασία, τόσο η μέθοδος Grid Search όσο και η Randomized Search χρησιμοποιήθηκαν μέσω των σχετικών συναρτήσεων της βιβλιοθήκης scikit-learn, για την βέλτιστη ρύθμιση των υπερπαραμέτρων των μοντέλων. Η δεύτερη μέθοδος χρησιμοποιήθηκε, λόγω του απαγορευτικά υψηλού υπολογιστικού χρόνου που απαιτούταν, σε περιπτώσεις μοντέλων με μεγάλο αριθμό υπερπαραμέτρων και πολλές υπό δοκιμή τιμές. Οι δύο αυτοί αλγόριθμοι, δέχονται σαν ορίσματα το μοντέλο πρόβλεψης, τα σύνολα τιμών των υπερπαραμέτρων που θα δοκιμαστούν, την τεχνική Cross-Validation που θα εφαρμοστεί και τον

τρόπο με τον οποίο θα γίνει η αξιολόγηση (scoring) του μοντέλου. Στη συνέχεια, για κάθε δυνατό συνδυασμό των υπερπαραμέτρων του ορίσματος, εκτελείται η σχετική τεχνική Cross-Validation και προκύπτει η αξιολόγηση του εκάστοτε μοντέλου. Τέλος, ο αλγόριθμος επιλέγει το μοντέλο με εκείνες τις υπερπαραμέτρους που έδωσαν την υψηλότερη απόδοση στο σύνολο αξιολόγησης της τεχνικής Cross-Validation. Οι διαφορές στις τιμές των υπερπαραμέτρων μεταξύ των δύο δικαστηρίων οφείλονται στις διαφορετικές ιδιαιτερότητες των αντίστοιχων συνόλων δεδομένων, όπως το μέγεθος, η κατανομή των κλάσεων και η πολυπλοκότητα των κειμένων.



Σχήμα 4.2: *Σχηματική απεικόνιση της τεχνικής 5-fold Cross-Validations*



Σχήμα 4.3: *Grid vs Random Search Model*

4.3 Εκπαίδευση Μοντέλων

Το σύνολο δεδομένων χωρίστηκε σε δύο επιμέρους σύνολα, τα οποία χρησιμοποιήθηκαν για την εκπαίδευση και την αξιολόγηση των μοντέλων, αντίστοιχα. Το σύνολο εκπαίδευσης (*train set*) περιείχε το 67% του συνόλου δεδομένων, ενώ το σύνολο δοκιμής (*test set*) το

33%. Στα πλαίσια της εργασίας, πραγματοποιήθηκαν τρεις διαφορετικοί τρόποι διαχωρισμού (*split*) του συνόλου δεδομένων, σύμφωνα με τις παραπάνω αναλογίες. Σε κάθε έναν, οι υπερπαράμετροι των μοντέλων πρόβλεψης ρυθμίστηκαν εκ νέου, αξιολογήθηκαν οι παραχθείσες προβλέψεις και σχολιάστηκαν τα αντίστοιχα συμπεράσματα. Οι τιμές των υπερπαραμέτρων έγινε μέσω Grid Search και Randomized Search όπως αναφέραμε παραπάνω, με 5-fold Cross Validation. Τα σύνολα δεδομένων των δύο δικαστηρίων είναι παρόμοια σε μέγεθος και χαρακτηριστικά, ωστόσο μικρές διαφορές στη δομή των δεδομένων ή στη γραμμικότητα είναι αναμενόμενο να επηρεάζουν την απόδοση τιμών των υπερπαραμέτρων των μοντέλων, όπως π.χ. των solvers, οδηγώντας στην επιλογή διαφορετικών τιμών (π.χ. *liblinear* έναντι *saga*).

4.3.1 Random Forest

Στην υλοποίηση του μοντέλου 2.3.2 της παρούσας διπλωματικής εργασίας, τα δεδομένα χωρίστηκαν σε χαρακτηριστικά (*features*) και ετικέτες (*labels*). Η παράμετρος stratify = y εξασφαλίζει ότι η κατανομή των κατηγοριών στο εκπαιδευτικό και στο δοκιμαστικό σύνολο είναι ισόρροπη ως προς την αρχική κατανομή των κατηγοριών.

Το Random Forest αρχικοποιείται ως μοντέλο, και καθορίζεται ένα πλέγμα υπερπαραμέτρων (param_grid), που περιλαμβάνει παραμέτρους όπως ο αριθμός των δέντρων (n_estimators), το μέγιστο βάθος δέντρων (max_depth), ο ελάχιστος αριθμός παραδειγμάτων για διαχωρισμό ή φύλλα (min_samples_split και min_samples_leaf), και η μέθοδος επιλογής χαρακτηριστικών για διαχωρισμό (max_features).

Στους Πίνακες 4.1 και 4.2 παρουσιάζονται οι τιμές των υπερπαραμέτρων για τα δύο δικαστήρια, όπως προέκυψαν μετά τη διαδικασία τη ρύθμισής τους.

Υπερπαράμετρος	**Τιμή**
n_estimators	102
min_samples_split	10
min_samples_leaf	1
max_features	0.17
max_depth	30
bootstrap	False

Πίνακας 4.1: *Τιμές υπερπαραμέτρων του μοντέλου Random Forest για τις αποφάσεις του Αρείου Πάγου*

Υπερπαράμετρος	Τιμή
n_estimators	151
min_samples_split	2
min_samples_leaf	1
max_features	0.54
max_depth	23
bootstrap	True

Πίνακας 4.2: *Τιμές υπερπαραμέτρων του μοντέλου Random Forest για τις αποφάσεις του Εφετείου Πειραιώς*

4.3.2 SVM

Αντίστοιχα, για το μοντέλο 2.3.4, τα δεδομένα χωρίστηκαν σε χαρακτηριστικά και ετικέτες και στη συνέχεια κατανεμήθηκαν σε σύνολα εκπαίδευσης και δοκιμών, με τη μέθοδο train_test_split, διατηρώντας τις αναλογίες των κατηγοριών με το stratify, ενώ η επαναληψιμότητα εξασφαλίστηκε μέσω του random_state. Το SVM μοντέλο αρχικοποιείται και συνοδεύεται από ένα πλέγμα παραμέτρων (*param_grid*) που περιλαμβάνει διαφορετικές τιμές για τον παράγοντα κανονικοποίησης C, τον συντελεστή του πυρήνα gamma, και τους τύπους πυρήνα (linear και rbf), έτσι ώστε να εξερευνηθούν ποικίλες πιθανές ρυθμίσεις. Με τη βοήθεια του GridSearchCV και την πεντάπτυχη (*5-fold*) διασταυρούμενη επικύρωση, το μοντέλο βελτιστοποιείται, προκειμένου να μεγιστοποιήσει το F1 Score. Πιο συγκεκριμένα, στη διαδικασία επιλογής των καλύτερων υπερπαραμέτρων του SVM, ο αλγόριθμος GridSearchCV αξιολογεί τις διάφορες συνδυαστικές ρυθμίσεις παραμέτρων (C, gamma και kernel) με βάση το F1 Score. Παράλληλα, χρησιμοποιείται το n_jobs=-1 για την αξιοποίηση όλων των διαθέσιμων πόρων επεξεργασίας.

Υπερπαράμετρος	Τιμή
C	100
gamma	0.06
kernel	rbf

Πίνακας 4.3: *Τιμές υπερπαραμέτρων του μοντέλου SVM για τις αποφάσεις του Αρείου Πάγου*

Υπερπαράμετρος	Τιμή
C	0.5
gamma	0.12
kernel	rbf

Πίνακας 4.4: *Τιμές υπερπαραμέτρων του μοντέλου SVM για τις αποφάσεις του Εφετείου Πειραιώς*

4.3.3 Logistic Regression

Για την ανάπτυξη του μοντέλου 2.3.5, αρχικά τα δεδομένα διαχωρίστηκαν σε χαρακτηριστικά X και ετικέτες y, ενώ στη συνέχεια πραγματοποιήθηκε κατανομή τους σε σύνολα

εκπαίδευσης και δοκιμών, χρησιμοποιώντας τη μέθοδο *train_test_split*. Η διαδικασία αυτή διασφάλισε τη διατήρηση της αναλογίας των κατηγοριών μέσω της παραμέτρου *stratify*, ενώ για την επαναληψιμότητα χρησιμοποιήθηκε σταθερό *random_state*. Ακολούθως, εφαρμόστηκε κανονικοποίηση των δεδομένων χαρακτηριστικών με τη χρήση της κλάσης *StandardScaler*, ώστε να διασφαλιστεί ότι όλες οι μεταβλητές βρίσκονται στην ίδια κλίμακα. Στη συνέχεια, το μοντέλο λογιστικής παλινδρόμησης αρχικοποιήθηκε με μέγιστο αριθμό επαναλήψεων (*max_iter*) ορισμένο στις 2000 και εφαρμόστηκε πλέγμα παραμέτρων (*param_grid*), το οποίο περιελάμβανε διάφορες τιμές για την παράμετρο κανονικοποίησης C (αντίστροφο της ισχύος της ποινής), διαφορετικά είδη κανονικοποίησης (*penalty*) και επιλογές αλγορίθμων (*solver*).

Υπερπαράμετρος	**Τιμή**
C	1
penalty	l1
solver	saga

Πίνακας 4.5: *Τιμές υπερπαραμέτρων του μοντέλου Logistic Regression για τις αποφάσεις του Αρείου Πάγου*

Υπερπαράμετρος	**Τιμή**
C	10
penalty	l2
solver	liblinear

Πίνακας 4.6: *Τιμές υπερπαραμέτρων του μοντέλου Logistic Regression για τις αποφάσεις του Εφετείου Πειραιώς*

4.3.4 Decision Trees

Αντίστοιχα, για το μοντέλο 2.3.1, τα δεδομένα χωρίστηκαν σε χαρακτηριστικά και ετικέτες και στη συνέχεια κατανεμήθηκαν σε σύνολα εκπαίδευσης και δοκιμών, με τη μέθοδο *train_test_split*, διατηρώντας τις αναλογίες των κατηγοριών με τη χρήση του *stratify*, ενώ η επαναληψιμότητα εξασφαλίστηκε μέσω του *random_state*. Το μοντέλο αρχικοποιείται μέσω του *DecisionTreeClassifier* και εκπαιδεύεται στο σύνολο εκπαίδευσης χρησιμοποιώντας τη μέθοδο *fit*.

Υπερπαράμετρος	**Τιμή**
criterion	entropy
max_depth	5
min_samples_leaf	4
min_samples_split	2

Πίνακας 4.7: *Τιμές υπερπαραμέτρων του μοντέλου Decision Trees για τις αποφάσεις του Αρείου Πάγου*

Υπερπαράμετρος	Τιμή
criterion	gini
max_depth	7
min_samples_leaf	1
min_samples_split	2

Πίνακας 4.8: *Τιμές υπερπαραμέτρων του μοντέλου Decision Trees για τις αποφάσεις του Εφετείου Πειραιώς*

4.3.5 XGBoost Regression

Αυτή η υλοποίηση του 2.3.3 περιλαμβάνει τη χρήση του RandomizedSearchCV για τη βελτιστοποίηση των υπερπαραμέτρων του μοντέλου. Αρχικά, ορίζεται μια κατανομή παραμέτρων που περιλαμβάνει επιλογές όπως το πλήθος των δέντρων (n_estimators), το βάθος τους (max_depth), τον ρυθμό εκμάθησης (learning_rate), και άλλες παραμέτρους όπως το subsample και το gamma. Στη συνέχεια, το RandomizedSearchCV εκτελεί 25 τυχαίες δοκιμές (n_iter=25) με 5-πτυχή διασταυρούμενη επικύρωση (cv=5), βελτιστοποιώντας τη μέτρηση F1 score.

Υπερπαράμετρος	Τιμή
subsample	0.5
n_estimators	100
min_child_weight	5
max_depth	3
learning_rate	0.04
gamma	0.4
colsample_bytree	0.8

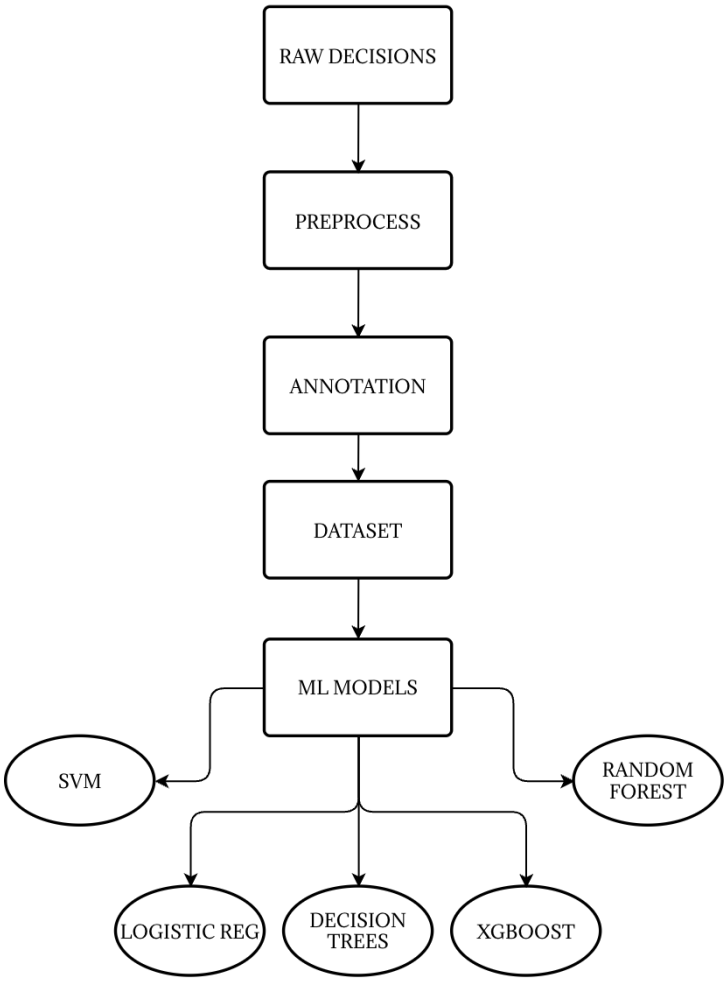
Πίνακας 4.9: *Τιμές υπερπαραμέτρων του μοντέλου XGB Regression για τις αποφάσεις του Αρείου Πάγου*

Υπερπαράμετρος	Τιμή
subsample	0.5
n_estimators	700
min_child_weight	2
max_depth	10
learning_rate	0.05
gamma	0.3
colsample_bytree	0.9

Πίνακας 4.10: *Τιμές υπερπαραμέτρων του μοντέλου XGB Regression για τις αποφάσεις του Εφετείου Πειραιώς*

4.3.6 Διαγραμματική Απεικόνιση

Το παρακάτω διάγραμμα απεικονίζει συνοπτικά το pipeline που ακολουθήθηκε στην παρούσα διπλωματική εργασία για την πρόβλεψη δικαστικών αποφάσεων μέσω μηχανικής μάθησης. Η διαδικασία ξεκινά από τις αρχικές (raw) δικαστικές αποφάσεις, οι οποίες στη συνέχεια υποβάλλονται σε προεπεξεργασία (preprocessing) για τη μετατροπή σε κατάλληλη μορφή για τα μοντέλα. Ακολούθως, προχωρήσαμε στο στάδιο της επισημείωσης (annotation), όπου οι αποφάσεις κατηγοριοποιούνται με βάση την έκβασή τους, δημιουργώντας έτσι ένα σύνολο δεδομένων (dataset).
Το τελικό dataset χρησιμοποιείται για την εκπαίδευση και αξιολόγηση αλγορίθμων μηχανικής μάθησης. Το διάγραμμα αυτό παρέχει μια σαφή εικόνα της ροής εργασιών, αναδεικνύοντας τα κρίσιμα στάδια επεξεργασίας των δεδομένων και την εφαρμογή των μοντέλων για την τελική πρόβλεψη των αποφάσεων.



Σχήμα 4.4: *Διάγραμμα Pipeline*

Κεφάλαιο 5

Πειραματικά Αποτελέσματα

Στο κεφάλαιο αυτό περιγράφεται η μεθοδολογία της παρούσα διπλωματικής εργασίας καθώς και η μετρικές απόδοσης των μοντέλων. Στη συνέχεια, παρουσιάζονται τα αποτελέσματα των μοντέλων ανά δικαστήριο, καθώς και τα συμπεράσματα που προέκυψαν από την αξιολόγησή τους. Για καθέναν από τους πέντε ταξινομητές που εξετάσαμε, τα μοντέλα που σχεδιάστηκαν, εκπαιδεύτηκαν στη συνέχεια με τα δεδομένα του συνόλου εκπαίδευσης. Η ρύθμιση των υπερπαραμέτρων τους έγινε με χρήση της τεχνικής cross-validation, όπως έχει ήδη αναλυθεί. Στη συνέχεια, εξετάστηκε η απόδοσή τους στο σύνολο αξιολόγησης (*validation set*) και φυσικά στο σύνολο δοκιμής (test set) [40]. Η τελική αξιολόγηση και σύγκριση των μοντέλων έγινε σύμφωνα με την απόδοσή τους στο σύνολο δοκιμής. Η αξιολόγηση των προβλέψεων έγινε χρησιμοποιώντας τις μετρικές απόδοσης που παρουσιάστηκαν στην Ενότητα 2.4.

5.1 Μεθοδολογία και Μετρικές

Για την αξιολόγηση της απόδοσης των μοντέλων ταξινόμησης δικαστικών αποφάσεων, ακολουθήθηκε μια μεθοδολογία που περιλαμβάνει την προεπεξεργασία των δεδομένων, την εκπαίδευση των μοντέλων και την ανάλυση των αποτελεσμάτων. Αρχικά, τα κείμενα των δικαστικών αποφάσεων μετατράπηκαν σε αριθμητική μορφή χρησιμοποιώντας τη μέθοδο TF-IDF (Term Frequency-Inverse Document Frequency), ενώ εφαρμόστηκαν τεχνικές όπως η αφαίρεση σημείων στίξης και stopwords. Στη συνέχεια, τα δεδομένα διαχωρίστηκαν σε σύνολο εκπαίδευσης (67%) και σύνολο δοκιμής (33%), ενώ για την αντιμετώπιση πιθανών προβλημάτων ανισορροπίας στις κατηγορίες χρησιμοποιήθηκαν στρατηγικές αναπροσαρμογής των δειγμάτων.

Για την επιλογή των βέλτιστων υπερπαραμέτρων εφαρμόστηκαν οι μέθοδοι Grid Search Cross-Validation και Randomized Search CV, εξασφαλίζοντας έτσι τη βελτιστοποίηση κάθε μοντέλου. Τα μοντέλα που αξιολογήθηκαν περιλαμβάνουν τον Random Forest, το Support Vector Machine (SVM), τη Logistic Regression, τα Decision Trees και τον XGBoost, με στόχο την επιλογή της πιο αποδοτικής προσέγγισης. Παράλληλα, για την αποφυγή overfitting και τη βελτίωση της γενίκευσης των μοντέλων, εφαρμόστηκε η μέθοδος k-fold Cross-Validation (k=5), επιτρέποντας μια πιο αξιόπιστη εκτίμηση της απόδοσης σε διαφορετικά υποσύνολα δεδομένων.

Η αξιολόγηση των μοντέλων πραγματοποιήθηκε με τη χρήση μετρικών απόδοσης, όπως

η Accuracy, το F1-score.

Για το Accuracy, ισχύει :

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total Number of Samples}}$$

Για το F1-score :

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

όπου:

- TP (True Positives) είναι οι περιπτώσεις στις οποίες το μοντέλο προβλέπει σωστά μια θετική κατηγορία.
- TN (True Negatives) είναι οι περιπτώσεις στις οποίες το μοντέλο προβλέπει σωστά μια αρνητική κατηγορία.

Επιπλέον, για τη βαθύτερη κατανόηση των επιδόσεων κάθε μοντέλου, αναλύθηκε ο πίνακας σύγχυσης (Confusion Matrix), ο οποίος παρέχει πληροφορίες σχετικά με τις σωστές και λανθασμένες ταξινομήσεις των αποφάσεων. Η συνδυαστική αξιολόγηση των παραπάνω μετρικών επιτρέπει μια σφαιρική αποτίμηση της απόδοσης των αλγορίθμων και συμβάλλει στην επιλογή της καταλληλότερης μεθόδου ταξινόμησης για το συγκεκριμένο πρόβλημα.

5.2 Αποτελέσματα για αποφάσεις Εφετείου Πειραιώς

Στην υποενότητα αυτή, γίνεται η συνοπτική παρουσίαση των αποτελεσμάτων των πειραμάτων που πραγματοποιήθηκαν για το σύνολο δεδομένων αποφάσεων του δικαστηρίου του Εφετείου Πειραιώς. Για κάθε μοντέλο που εξετάσαμε, καταγράφεται η απόδοσή του, σύμφωνα με την τιμή κάθε μετρικής απόδοσης, στο σύνολο δοκιμής (test set) και στο σύνολο αξιολόγησης (validation set), κατά τη διαδικασία του cross-validation.

Ο Πίνακας 5.2 αποτυπώνει τις επιδόσεις πέντε διαφορετικών αλγορίθμων (Random Forest, Decision Tree, Support Vector Machine, XGBoost και Logistic Regression) βάσει των μετρικών Precision, Recall, Accuracy και F1-Score στο σύνολο δοκιμής. Η ανάλυση των αποτελεσμάτων αποκαλύπτει σημαντικές διαφορές στις αποδόσεις των μοντέλων, οι οποίες προσφέρουν πολύτιμα συμπεράσματα για την καταλληλότητά τους, ανάλογα με τον στόχο της ταξινόμησης.

Ο XGBoost (XGB) ξεχωρίζει ως το πιο αποδοτικό μοντέλο, επιτυγχάνοντας τις υψηλότερες τιμές στις περισσότερες μετρικές απόδοσης. Με Precision 0.76, Recall 0.90 και F1-Score 0.83, καταφέρνει να διατηρήσει την ισορροπία μεταξύ ανίχνευσης θετικών περιπτώσεων - δηλαδή των αποφάσεων που επισημάνθηκαν ως απόρριψη- και ακρίβειας, καθιστώντας το εξαιρετική επιλογή για εφαρμογές που απαιτούν υψηλή ακρίβεια χωρίς να θυσιάζεται το recall. Αντίστοιχα, το Random Forest (RF) εμφανίζει παρόμοια απόδοση, με ιδιαίτερα υψηλό Recall (0.90), το οποίο υποδεικνύει ότι ανιχνεύει τη συντριπτική πλειονότητα των θετικών περιπτώσεων. Η Precision του (0.75) είναι ελαφρώς χαμηλότερη, όμως το F1-Score 0.82

δείχνει ότι αποτελεί εξίσου αξιόπιστη επιλογή, ιδίως όταν η ανίχνευση όλων των θετικών περιπτώσεων είναι κρίσιμη.

Το Support Vector Machine (SVM) παρουσιάζει τη μεγαλύτερη τιμή Recall(0.92) από όλα τα μοντέλα, γεγονός που σημαίνει ότι ανιχνεύει σχεδόν όλες τις θετικές περιπτώσεις. Ωστόσο, το Precision του (0.71) είναι χαμηλό, γεγονός που υποδηλώνει αυξημένο αριθμό ψευδώς θετικών προβλέψεων. Το χαρακτηριστικό αυτό καθιστά το SVM κατάλληλο για εφαρμογές όπου προτεραιότητα είναι η ανίχνευση όσο το δυνατόν περισσότερων θετικών περιπτώσεων, ακόμη κι αν αυτό συνεπάγεται αυξημένα ψευδή θετικά.

Αντίθετα, το Decision Tree (DT) και το Logistic Regression (LR) υπολείπονται σημαντικά σε απόδοση. Το DT εμφανίζει Precision 0.72 και Recall 0.69, ενώ το LR καταγράφει ακόμα χαμηλότερη Recall (0.67). Και τα δύο μοντέλα εμφανίζουν χαμηλά F1-Score (0.70), γεγονός που τα καθιστά λιγότερο αξιόπιστα συγκριτικά με τα υπόλοιπα. Η χαμηλή τους απόδοση δείχνει ότι πιθανώς δεν είναι κατάλληλα για σύνθετες ταξινομήσεις όπου απαιτείται ισορροπία μεταξύ ακρίβειας και ανάκλησης.

Συνολικά, ο XGBoost αναδεικνύεται ως ο πιο αποτελεσματικός ταξινομητής, ενώ το Random Forest μπορεί να προτιμηθεί σε περιπτώσεις όπου η ανίχνευση θετικών περιπτώσεων είναι υψίστης σημασίας. Η επιλογή του μοντέλου θα πρέπει να καθορίζεται από τη σχετική σημασία της ακρίβειας (Precision) έναντι της ανάκλησης (Recall), ανάλογα με τις ανάγκες της συγκεκριμένης εφαρμογής, όπως η απόρριψη ή η αποδοχή δικαστικών αποφάσεων.

	RF	**DT**	**SVM**	**XGB**	**LR**
Fold-1	0.76	0.70	0.78	0.70	0.70
Fold-2	0.74	0.69	0.80	0.69	0.70
Fold-3	0.73	0.72	0.80	0.71	0.72
Fold-4	0.71	0.72	0.79	0.72	0.72
Fold-5	0.71	0.69	0.77	0.69	0.69
Mean	0.78	0.72	0.81	0.72	0.73
SD	0.04	0.02	0.06	0.02	0.02

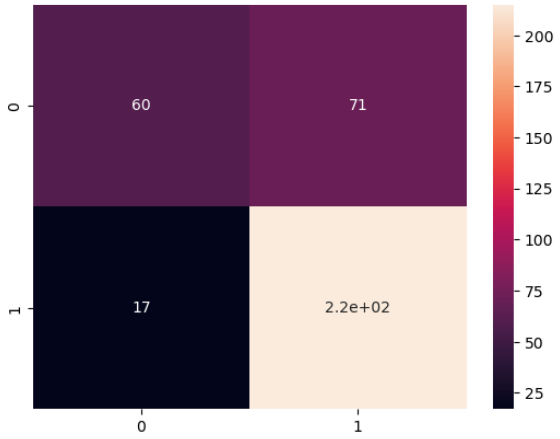
Πίνακας 5.1: *Μετρικές απόδοσης στο validation set για τις αποφάσεις Εφετείου Πειραιώς*

	RF	**DT**	**SVM**	**XGB**	**LR**
Precision	0.75	0.72	0.71	0.76	0.71
Recall	**0.90**	0.69	**0.92**	**0.90**	0.67
Accuracy	0.76	0.62	0.71	0.76	0.63
F1-Score	**0.82**	0.70	0.80	**0.83**	0.70

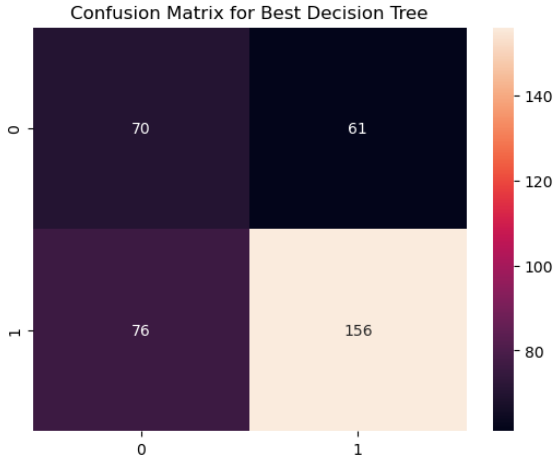
Πίνακας 5.2: *Μετρικές απόδοσης στο test set για τις αποφάσεις Εφετείου Πειραιώς*

5.2.1 Confusion Matrix μοντέλων

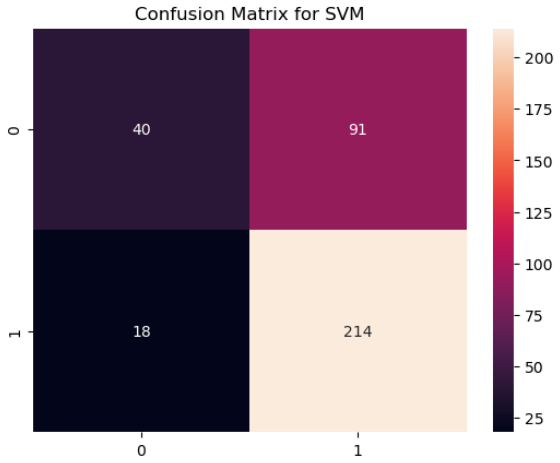
Στην υποενότητα αυτή, παρουσιάζονται γραφικά οι πίνακες σύγχυσης (*confusion matrix*) που προκύπτουν από τα πειράματα πρόβλεψης, παρέχοντας μια οπτική αναπαράσταση της απόδοσης των αλγορίθμων. Κάθε πίνακας δείχνει την κατανομή των σωστών και λανθασμένων προβλέψεων όπως έχει αναλυθεί στην υποενότητα Ενότητα 2.4.



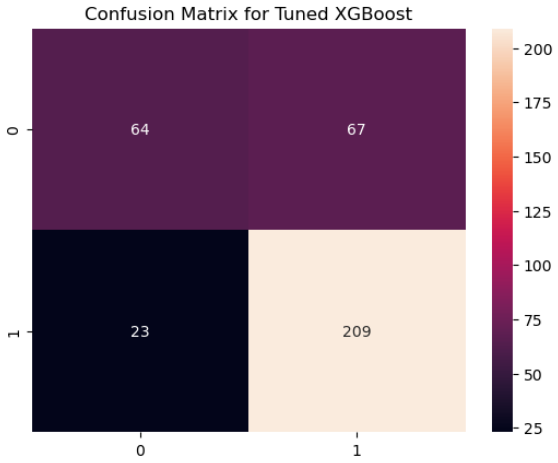
Σχήμα 5.1: *Random Forest CM για τις αποφάσεις Εφετείου Πειραιώς*



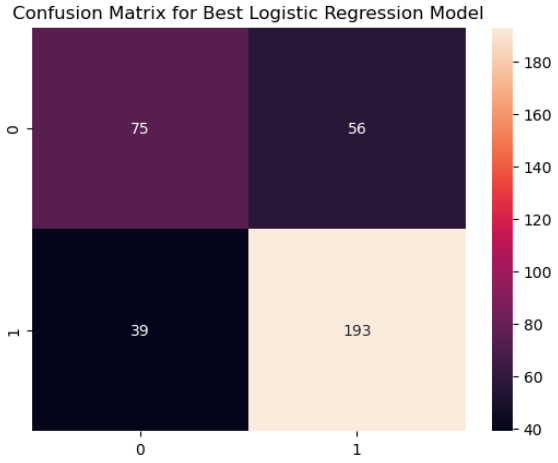
Σχήμα 5.2: *Decision Trees CM για τις απφάσεις Εφετείυ Πειραιώς*



Σχήμα 5.3: *SVM CM για τις αποφάσεις Εφετείου Πειραιώς*



Σχήμα 5.4: *XGBoost CM για τις αποφάσεις Εφετείου Πειραιώς*



Σχήμα 5.5: *Logistic Regression CM για τις αποφάσεις Εφετείου Πειραιώς*

5.3 Αποτελέσματα για αποφάσεις Αρείου Πάγου

Στην υποενότητα αυτή, γίνεται η συνοπτική παρουσίαση των αποτελεσμάτων των πειραμάτων που πραγματοποιήθηκαν για το σύνολο δεδομένων αποφάσεων του δικαστηρίου του Αρείου Πάγου. Για κάθε μοντέλο που εξετάσαμε, καταγράφεται η απόδοσή του, σύμφωνα με την τιμή κάθε μετρικής απόδοσης, στο σύνολο δοκιμής (*test set*) και στο σύνολο αξιολόγησης (*validation set*), κατά τη διαδικασία του cross-validation.

Ο Πίνακας 5.4 παρουσιάζει τις μετρικές απόδοσης πέντε αλγορίθμων : Random Forest - RF, Decision Tree - DT, Support Vector Machine - SVM, XGBoost - XGB και Logistic Regression - LR στο σύνολο δοκιμής.

Ο Random Forest (RF) εμφανίζει την υψηλότερη ισορροπία μεταξύ recall (0.86), precision (0.79) και F1-Score (0.82), γεγονός που τον καθιστά τον πιο αποδοτικό αλγόριθμο για την πρόβλεψη της θετικής κατηγορίας, δηλαδή των αποφάσεων που επισημάνθηκαν ως απόρριψη. Η υψηλή τιμή του recall υποδηλώνει ότι το μοντέλο καταφέρνει να εντοπίζει τις περισσότερες θετικές περιπτώσεις -δηλαδή των αποφάσεων που επισημάνθηκαν ως απόρριψη-

ενώ η τιμή του precision του υποδεικνύει ότι τα περισσότερα θετικά που προβλέπει είναι σωστά.

Ο XGBoost αποδίδει εξίσου καλά, με F1-Score 0.82, πολύ κοντά στον RF. Η τιμή του recall του (0.85) είναι σχεδόν ισοδύναμη με του RF, ενώ το precision του (0.78) δείχνει μικρότερη, αλλά αποδεκτή αποτελεσματικότητα. Ο XGBoost αποτελεί επίσης μια καλή επιλογή για μοντέλα όπου απαιτείται υψηλή ανάκληση.

Αντίθετα, ο Decision Tree (DT) έχει τη χαμηλότερη απόδοση μεταξύ των μοντέλων, με ακρίβεια 0.71 και χαμηλότερη τιμή recall (0.77). Αυτό το καθιστά λιγότερο αξιόπιστο, καθώς τα αποτελέσματα του έχουν μεγαλύτερη πιθανότητα λάθους τόσο στις θετικές όσο και στις αρνητικές προβλέψεις.

Ο Support Vector Machine (SVM) εμφανίζει υψηλή ακρίβεια (0.80), αλλά η σχετικά χαμηλότερη recall (0.75) δείχνει ότι μπορεί να αποτυγχάνει να εντοπίσει αρκετές θετικές περιπτώσεις. Το F1-Score του (0.78) υποδεικνύει ότι προσφέρει μια ισορροπημένη αλλά ελαφρώς υποδεέστερη απόδοση από τους RF και XGB.

Τέλος, η Logistic Regression (LR) έχει παρόμοιες επιδόσεις με τον SVM, με ευαισθησία (0.80), ακρίβεια (0.76) και F1-Score (0.78). Παρόλο που δεν ξεχωρίζει, παραμένει μια σταθερή επιλογή με αποδεκτή ακρίβεια και ανάκληση. Συνολικά, ο RF και ο XGB ξεχωρίζουν ως οι πιο αξιόπιστες επιλογές για την ανάλυση, ενώ οι DT και SVM αποδεικνύονται λιγότερο αποδοτικοί.

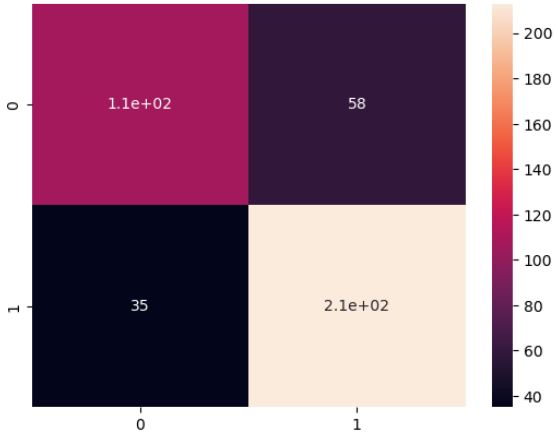
	RF	**DT**	**SVM**	**XGB**	**LR**
Fold-1	0.74	0.72	0.78	0.75	0.72
Fold-2	0.80	0.76	0.78	0.79	0.74
Fold-3	0.76	0.76	0.78	0.75	0.69
Fold-4	0.78	0.77	0.78	0.78	0.73
Fold-5	0.83	0.81	0.80	0.82	0.74
Mean	0.81	0.78	0.83	0.82	0.82
SD	0.03	0.03	0.01	0.03	0.01

Πίνακας 5.3: *Μετρικές απόδοσης στο validation set για τις αποφάσεις Αρείου Πάγου*

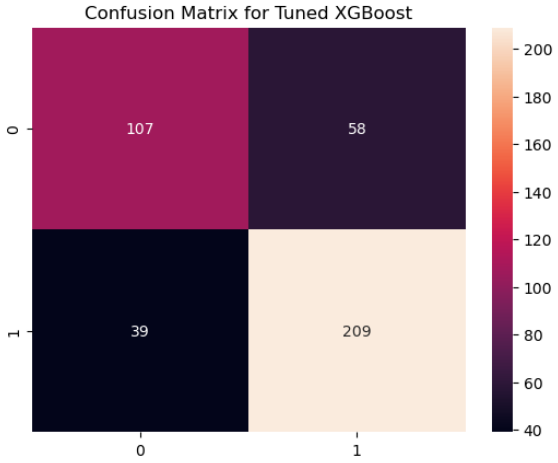
	RF	**DT**	**SVM**	**XGB**	**LR**
Precision	0.79	0.76	0.80	0.78	0.76
Recall	**0.86**	0.77	0.75	**0.85**	0.80
Accuracy	**0.77**	0.71	0.74	**0.77**	0.73
F1-Score	**0.82**	0.77	0.78	**0.82**	0.78

Πίνακας 5.4: *Μετρικές απόδοσης στο test set για τις αποφάσεις Αρείου Πάγου*

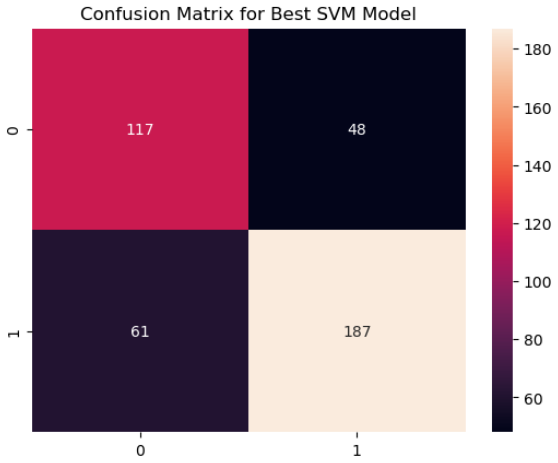
5.3.1 Confusion Matrix μοντέλων



Σχήμα 5.6: *Random Forest CM για τις αποφάσεις Αρείου Πάγου*

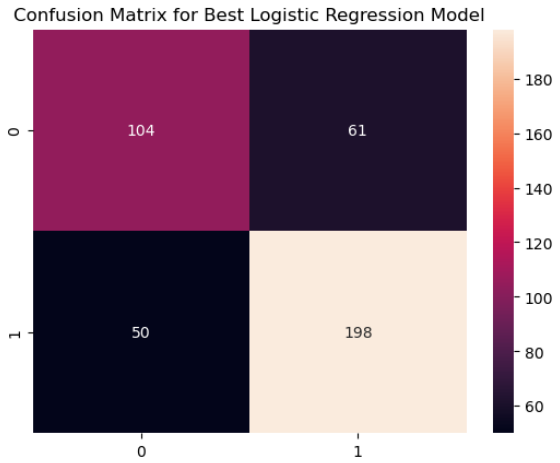


Σχήμα 5.7: *XGBoost CM για τις αποφάσεις Αρείου Πάγου*



Σχήμα 5.8: *SVM CM για τις αποφάσεις Αρείου Πάγου*

Σχήμα 5.9: *Decision Trees CM για τις αποφάσεις Αρείου Πάγου*



Σχήμα 5.10: *Logistic Regression CM για τις αποφάσεις Αρείου Πάγου*

5.4 Συμπεράσματα

Μετά το πέρας των πειραμάτων που πραγματοποιήθηκαν, είναι ασφαλές και χρήσιμο να εξάγουμε ορισμένα συμπεράσματα για την αποδοτικότητα των πέντε ταξινομητών που μελετήσαμε στις δικαστικές αποφάσεις των δύο δικαστηρίων. Πιο συγκεκριμένα, παρατηρούμε ότι τόσο ο Random Forest (RF) όσο και ο XGBoost (XGB) αναδεικνύονται σταθερά ως οι πιο αποδοτικοί αλγόριθμοι παρουσιάζοντας υψηλή ισορροπία μεταξύ Precision, Recall και F1-Score. Και στις δύο περιπτώσεις, ο RF υπερέχει ελαφρώς ως προς την ικανότητά του να ανιχνεύει τις θετικές περιπτώσεις, γεγονός που συμφωνεί και με την εργασία [18] (υψηλό Recall), ενώ ταυτόχρονα διατηρεί ικανοποιητική ακρίβεια στις προβλέψεις του. Ο XGBoost, αν και με ελαφρώς χαμηλότερο Precision, παραμένει μια ισχυρή εναλλακτική, ιδιαίτερα όταν απαιτείται υψηλή ανάκληση, δηλαδή σωστή ανίχνευση όσο το δυνατόν περισσότερων θετικών περιπτώσεων.

Αντίθετα, οι Decision Tree (DT) και Support Vector Machine (SVM) υπολείπονται σε αποδοτικότητα. Ο DT εμφανίζει συστηματικά τη χαμηλότερη απόδοση, γεγονός που τον καθιστά λιγότερο αξιόπιστο, ενώ ο SVM, παρά την υψηλή ακρίβεια, εμφανίζει χαμηλότερο

Recall, κάτι που περιορίζει την ικανότητά του να εντοπίζει όλες τις θετικές περιπτώσεις. Η Logistic Regression (LR), αν και δεν διακρίνεται, προσφέρει σταθερές επιδόσεις και μπορεί να είναι κατάλληλη για πιο απλές εφαρμογές. Συνολικά, τα αποτελέσματα δείχνουν ότι η επιλογή του καταλληλότερου μοντέλου εξαρτάται από το αν προτεραιότητα είναι το Precision ή η ανίχνευση όλων των θετικών περιπτώσεων, με τους RF και XGB να αποτελούν τις πιο αξιόπιστες επιλογές.

Κεφάλαιο 6

Επίλογος

6.1 Συμπεράσματα διπλωματικής εργασίας

Στην παρούσα διπλωματική εργασία μελετήθηκε το πρόβλημα της πρόβλεψης δικαστικών αποφάσεων με χρήση τεχνικών μηχανικής μάθησης και επεξεργασίας φυσικής γλώσσας (NLP). Αρχικά έγινε μια εκτεταμένη καταγραφή και μελέτη σχετικών εργασιών τεχνολογιών αιχμής, καθώς και συλλογή, και παρουσίαση του συνόλου δεδομένων που δημιουργήσαμε, το οποίο περιλαμβάνει αποφάσεις του Αρείου Πάγου και του Εφετείου Πειραιώς. Ακολούθως, πραγματοποίησαμε διαδικασίες προεπεξεργασίας κειμένου, όπως καθαρισμό δεδομένων, μετατροπή σε αριθμητικές αναπαραστάσεις TF-IDF και επισημείωση των αποφάσεων ως *αποδοχή* ή *απόρριψη*. Στη συνέχεια, καθορίστηκε και παρουσιάστηκε αναλυτικά η μεθοδολογία που απαιτείται για την εκπαίδευση μοντέλων μηχανικής μάθησης στο πρόβλημα της πρόβλεψης δικαστικών αποφάσεων. Ακολούθως, παρουσιάστηκε η μεθοδολογία της βέλτιστης ρύθμισης των υπερπαραμέτρων των μοντέλων και έγινε εκτεταμένη αξιολόγηση ενός μεγάλου αριθμού μοντέλων μηχανικής μάθησης στο πρόβλημα. Η εργασία υπογράμμισε τη σημασία της προεπεξεργασίας των δεδομένων και της σωστής επιλογής υπερπαραμέτρων για τη βελτιστοποίηση των αποτελεσμάτων. Μέσα από την ανάλυση και την εφαρμογή διαφόρων μοντέλων, τα πειραματικά αποτελέσματα ανέδειξαν τον Random Forest και τον XGBoost ως τους πιο αποδοτικούς ταξινομητές -γεγονός που συμφωνεί και με τις [18], [15] και [26]- με υψηλές τιμές σε Precision, Recall και F1-Score, προσφέροντας σταθερότητα και αξιοπιστία στις προβλέψεις.

6.2 Μελλοντική δουλειά στον τομέα

Η παρούσα εργασία μπορεί να επεκταθεί με την υιοθέτηση προηγμένων τεχνικών σημασιολογικής αναπαράστασης για την καλύτερη αποτύπωση των εννοιών και των σχέσεων μεταξύ των λέξεων στα κείμενα. Οι τεχνικές αυτές, όπως τα embeddings (Word2Vec [41], BERT [42]), επιτρέπουν στο μοντέλο να κατανοεί τη σημασία μιας λέξης ανάλογα με το συμφραζόμενο πλαίσιο, αποτυπώνοντας έτσι βαθύτερες νοηματικές συνδέσεις. Για παράδειγμα, λέξεις που χρησιμοποιούνται σε παρόμοια πλαίσια αποκτούν πιο κοντινές αριθμητικές αναπαραστάσεις, διευκολύνοντας την ανάλυση και τη λήψη αποφάσεων. Η εφαρμογή τέτοιων τεχνικών μπορεί να βελτιώσει σημαντικά την ακρίβεια των προβλέψεων και να προσφέρει πιο συνεκτική κατανόηση των δικαστικών αποφάσεων.

Η αξιολόγηση των αποτελεσμάτων από νομικούς επαγγελματίες αποτελεί μια σημαντική πτυχή για την ενίσχυση της αξιοπιστίας των προβλέψεων που προκύπτουν από τη μηχανική μάθηση και την επεξεργασία φυσικής γλώσσας. Η συμμετοχή των νομικών στην αξιολόγηση θα επιτρέπει να διασφαλιστεί ότι τα αποτελέσματα των μοντέλων είναι συνεπή με τις πραγματικές ανάγκες του νομικού κόσμου και ανταποκρίνονται στις σύνθετες και συχνά αμφισβητούμενες έννοιες της νομικής γλώσσας. Επιπλέον, η συνεργασία αυτή θα προσφέρει πολύτιμες ανατροφοδοτήσεις, οι οποίες θα μπορούσαν να οδηγήσουν σε βελτιώσεις στους αλγορίθμους και τις στρατηγικές ταξινόμησης, προσαρμόζοντας τις προσεγγίσεις στις απαιτήσεις της επαγγελματικής πράξης. Με αυτόν τον τρόπο, οι προβλέψεις θα γίνουν πιο χρήσιμες και εφαρμόσιμες σε καθημερινές νομικές καταστάσεις, ενισχύοντας την αξία της μεθόδου.

Από την άλλη πλευρά, η δυνατότητα γενίκευσης των αποτελεσμάτων σε διαφορετικά νομικά συστήματα και δεδομένα αποτελεί ένα κρίσιμο επόμενο βήμα για τη διεύρυνση της εφαρμογής αυτών των τεχνικών. Η ανάλυση της δυνατότητας προσαρμογής των μοντέλων σε διεθνή νομικά περιβάλλοντα και διαφορετικά σύνολα δεδομένων θα επιτρέψει την ανάπτυξη πιο ευέλικτων και επεκτάσιμων συστημάτων. Η ενσωμάτωση ποικίλων νομικών πλαισίων θα ενισχύσει την εφαρμοσιμότητα των τεχνικών ταξινόμησης σε παγκόσμιο επίπεδο και θα προσφέρει στους νομικούς τη δυνατότητα να επωφεληθούν από τα εργαλεία αυτά, ανεξάρτητα από το νομικό σύστημα με το οποίο ασχολούνται. Η επέκταση αυτής της προσέγγισης πέρα από τα τοπικά δεδομένα θα συμβάλει σημαντικά στην εξέλιξη της αυτόματης ανάλυσης νομικών εγγράφων σε παγκόσμια κλίμακα.

Επιπρόσθετα, θα μπορούσε να διερευνηθεί η ενσωμάτωση δεδομένων από διαφορετικές γλώσσες και δικαιοδοσίες, προκειμένου να μελετηθεί κατά πόσο η μεθοδολογία είναι μεταβιβάσιμη και αποτελεσματική σε διαφορετικά νομικά και πολιτισμικά πλαίσια. Μία άλλη ενδιαφέρουσα προοπτική είναι η ανάπτυξη ενός συστήματος ενισχυτικής μάθησης (reinforcement learning), το οποίο θα βελτιώνεται διαρκώς μέσα από την αλληλεπίδραση με πραγματικούς χρήστες, όπως νομικούς, δικαστές ή ερευνητές, ενσωματώνοντας εποικοδομητική ανατροφοδότηση.

Η δυνατότητα ενός νομικού επαγγελματία να έχει στη διάθεσή του μια κατηγοριοποίηση των δικαστικών αποφάσεων σε διάφορα νομικά είδη και υποθέσεις, συνδυασμένη με ένα μοντέλο που ταξινομεί και εντοπίζει συναφείς υποθέσεις, θα ήταν εξαιρετικά χρήσιμη. Μέσω αυτού του μοντέλου, ο νομικός θα μπορούσε να εντοπίζει παρόμοιες υποθέσεις με εκείνες που αναλαμβάνει, αποκτώντας έτσι πολύτιμες πληροφορίες για τις στρατηγικές που χρησιμοποιήθηκαν και τις πιθανές εκβάσεις. Αυτή η κατηγοριοποίηση και ανάλυση θα του επιτρέπει να προσαρμόσει την προσέγγισή του με βάση τα ιστορικά δεδομένα, εξοικονομώντας χρόνο και μειώνοντας τον κίνδυνο λαθών. Επιπλέον, η δυνατότητα εκτίμησης της πιθανότητας επιτυχίας ή αποτυχίας μιας υπόθεσης, βασισμένη σε παρόμοιες περιπτώσεις, θα αποτελούσε ένα σημαντικό εργαλείο για τη διαμόρφωση της στρατηγικής του. Αναγνωρίζοντας τα πρότυπα και τις τάσεις από παρόμοιες υποθέσεις, ο νομικός θα μπορούσε να ενισχύσει την αποτελεσματικότητα της νομικής του επιχειρηματολογίας και να παρέχει μια πιο τεκμηριωμένη εκτίμηση για την πιθανή έκβαση της υπόθεσης. Συνολικά, ένα τέτοιο σύστημα θα προσέφερε στους νομικούς επαγγελματίες τα εργαλεία που χρειάζονται για να είναι καλύτερα προετοιμασμένοι, να αναγνωρίζουν τις στρατηγικές που πρέπει να ακολουθήσουν και να εκτιμούν πιο αξιόπιστα τις πιθανές εκβάσεις των υποθέσεων που αναλαμβάνουν.

Τέλος, η παρούσα μελέτη μπορεί να αποτελέσει τη βάση για τη δημιουργία ολοκληρωμένων εργαλείων υποστήριξης απόφασης, τα οποία όχι μόνο θα προβλέπουν το πιθανό αποτέλεσμα, αλλά θα παρέχουν και επεξηγήσεις για τις προβλέψεις, προάγοντας τη διαφάνεια και την εμπιστοσύνη στη χρήση τέτοιων συστημάτων. Τέτοιες πρωτοβουλίες θα μπορούσαν να έχουν σημαντικό κοινωνικό αντίκτυπο, ενισχύοντας τη διαφάνεια και την αποδοτικότητα στον τομέα της δικαιοσύνης.

Βιβλιογραφία

[1] Dimitris P. Galanis. *Automatic Summarization of Court Judgements using Machine Learning*. Διπλωματική εργασία, National Technical University of Athens, 2022. DOI : http://dx.doi.org/10.26240/heal.ntua.24737.

[2] Βλάχου-Βλαχοπούλου, Μαγδαληνή-Χριστίνα. *Οι πηγές του Δικαίου*. ΠΜΣ Δημόσιο ίκαιο, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, 2022.

[3] Martijn Wieling Michel Vols Masha Medvedeva. *Rethinking the field of automatic prediction of court decisions*. *Artificial Intelligence and Law*, 2023. DOI: 10.1007/s10506-021-09306-3.

[4] Preotiuc Pietro D. Lampros V. Aletras N., Tsarapatsanis D. *Predicting judicial decisions of the European Court of Human Rights: A natural language, processing perspective*. *PeerJ Computer Science*, 2016. DOI : https://doi.org/10.7717/peerj-cs.93.

[5] Z. Liu και H. Chen. *A predictive performance comparison of machine learning models for judicial cases*. *Symposium Series on Computational Intelligence*, 2017. DOI : 10.1109/SSCI.2017.8285436.

[6] Andrea Visentin, Alessia Nardotto και Barry O'Sullivan. *Predicting Judicial Decisions: A Statistically Rigorous Approach and a New Ensemble Classifier*. *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, 2019. DOI : 10.1109/ICTAI.2019.00275.

[7] Alexandre Quemy και Robert Wrembel. *On Integrating and Classifying Legal Text Documents*. *Database and Expert Systems Applications*. Springer International Publishing, 2020. DOI : https://doi.org/10.1007/978-3-030-59003-1_25.

[8] Zhenyu Liu και Huanhuan Chen. *A predictive performance comparison of machine learning models for judicial cases*. *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2017. DOI : 10.1109/SSCI.2017.8285436.

[9] Octavia Maria Sulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P. Dinu και Josefvan Genabith. *Exploring the Use of Text Classification in the Legal Domain*. 2017. DOI : https://arxiv.org/abs/1710.09306.

[10] Michael Benedict L. Virtucio, Jeffrey A. Aborot, John Kevin C. Abonita, Roxanne S. Aviñante, Rother Jay B. Copino, Michelle P. Neverida, Vanesa O. Osiana, Elmer C. Peramo, Joanna G. Syjuco και Glenn Brian A. Tan. *Predicting Decisions of the Philippine Supreme Court Using Natural Language Processing and Machine Learning*.

2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), τόμος 02, 2018. DOI: 10.1109/COMPSAC.2018.10348.

[11] Santana O Oliveira Lage L. Lage-Freitas A, Allende-Cid H. *Predicting Brazilian Court Decisions*. *PeerJ Computer Science*, 2021. DOI : https://doi.org/10.7717/peerj-cs.904.

[12] Vithor Bertalan και Evandro Ruiz. *Predicting Judicial Outcomes in the Brazilian Legal System Using Textual Features*. 2022.

[13] Floris Bex και Henry Prakken. *On the relevance of algorithmic decision predictors for judicial decision making*. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. Association for Computing Machinery, 2021. DOI : https://doi.org/10.1145/3462757.3466069.

[14] Kankawin Kowsrihawat, Peerapon Vateekul και Prachya Boonkwan. *Predicting Judicial Decisions of Criminal Cases from Thai Supreme Court Using Bi-directional GRU with Attention Mechanism*. *2018 5th Asian Conference on Defense Technology (ACDT)*, 2018. DOI : 10.1109/ACDT.2018.8592948.

[15] Masha Medvedeva, Michel Vols και Martijn Wieling. *Using machine learning to predict decisions of the European Court of Human Rights*. *Artificial Intelligence and Law*, 2020. DOI: https://doi.org/10.1007/s10506-019-09255-y.

[16] Benjamin Strickson και Beatriz De La Iglesia. *Legal Judgement Prediction for UK Courts*. 2020. DOI : 10.1145/3388176.3388183.

[17] Ilias Chalkidis, Ion Androutsopoulos και Nikolaos Aletras. *Neural Legal Judgment Prediction in English*. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. DOI : 10.18653/v1/P19-1424.

[18] Emre Mumcuoğlu and Ceyhun E. Öztürk and Haldun M. Ozaktas and Aykut Koç. *Natural language processing in law: Prediction of outcomes in the higher courts of Turkey*. *Information Processing Management*, 2021. DOI : https://doi.org/10.1016/j.ipm.2021.102684.

[19] Jiajing Li, Guoying Zhang, Hongfei Yan, Longxue Yu και Tao Meng. *A Markov Logic Networks Based Method to Predict Judicial Decisions of Divorce Cases*. *2018 IEEE International Conference on Smart Cloud (SmartCloud)*, 2018. DOI : 10.1109/SmartCloud.2018.00029.

[20] Benjamin Strickson και Beatriz De La Iglesia. *Legal Judgement Prediction for UK Courts*. *Proceedings of the 3rd International Conference on Information Science and Systems*. Association for Computing Machinery, 2020. DOI : 10.1145/3388176.3388183.

[21] Marios Koniaris, Dimitris Galanis, Eugenia Giannini και Panayiotis Tsanakas. *Evaluation of Automatic Legal Text Summarization Techniques for Greek Case Law*. *Information*, 2023. DOI : https://doi.org/10.3390/info14040250.

[22] Souvik Sengupta και Vishwang Dave. *Predicting applicable law sections from judicial case reports using legislative text analysis with machine learning*. *Journal of Computational Social Science*, 2022. DOI : 10.1007/s42001-021-00135-7.

[23] Manjusha Singh Tomar και Vishal Gupta. *Legal Case Classification Using Machine Learning with NLP*. *2023 International Conference on Data Science, Agents Artificial Intelligence (ICDSAAI)*, 2023. DOI : 10.1109/ICDSAAI59313.2023.10452618.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee και Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. *CoRR*, αβ-σ/1810.04805, 2018. DOI: https://doi.org/10.48550/arXiv.1810.04805.

[25] Kashif Javed και Jianxin Li. *Artificial intelligence in judicial adjudication: Semantic biasness classification and identification in legal judgement (SBCILJ)*. *Heliyon*, 2024. DOI : 10.1016/j.heliyon.2024.e30184.

[26] André Aguiar, Raquel Silveira, Vládia Pinheiro, Vasco Furtado και João Araújo" Neto. *Text Classification in Legal Documents Extracted from Lawsuits in Brazilian Courts*. *Springer International Publishing*, 2021. DOI: 10.1007/978-3-030-91699-2_40.

[27] Πετρόπουλος Φ. και Ασημακόπουλος Β. *Επιχειρησιακές Προβλέψεις*. *Συμμετρία*, 2013. DOI: https://doi.org/10.1023/A:1010933404324.

[28] Lior Rokach και Oded Maimon. *The Data Mining and Knowledge Discovery Handbook*, τόμος 6. Decision Trees, 2005. URL: https://link.springer.com/book/10.1007/b107408.

[29] Breiman Leo. *The creation, structure, and interpretation of the legal text*. *Machine Learning*, 45, 2001. DOI: https://doi.org/10.1023/A:1010933404324.

[30] Tianqi Chen και Carlos Guestrin. *XGBoost: A Scalable Tree Boosting System*. ACM, 2016. DOI: http://dx.doi.org/10.1145/2939672.2939785.

[31] Corinna Cortes και Vladimir Vapnik. *Support-vector networks*. Machine learning, 1995. DOI: https://doi.org/10.1007/BF00994018.

[32] Jeffrey M. Stanton. *Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors*. Journal of Statistics Education, 2001. DOI: https://doi.org/10.1080/10691898.2001.11910537.

[33] Sandro Sperandei. *Understanding logistic regression analysis*. Biochem Med, 2014. DOI: 10.11613/BM.2014.003.

[34] Ashulekha Gupta, Rishabh Parmar, Pradeep Suri και Rajiv Kumar. *Determining Accuracy Rate of Artificial Intelligence Models using Python and R-Studio*. 2021. DOI : 10.1109/ICAC3N53548.2021.9725687.

[35] Marina Sokolova, Nathalie Japkowicz και Stan Szpakowicz. *Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation*. 2006.

[36] Edward H. Livingston. *The mean and standard deviation: what does it all mean? Journal of Surgical Research*, 119, 2004. DOI: https://doi.org/10.1016/j.jss.2004.02.008.

[37] James T Townsend. *Theoretical analysis of an alphabetic confusion matrix*. *Perception & Psychophysics*, 9:40–50, 1971.

[38] A. Joulin P. Bojanowski, E. Grave και T. Mikolov. *Enriching word vectors with subword information*. *Transactions of the Association for Computational Linguistics*, 5, 2017. DOI: https://doi.org/10.48550/arXiv.1607.04606.

[39] Claude Sammut και Geoffrey I. Webb. *Encyclopedia of Machine Learning*. Springer US, 2010. DOI : https://doi.org/10.1007/978-0-387-30164-8_832.

[40] Κωνσταντίνος Μέξης. *Πρόβλεψη κατανάλωσης φυσικού αερίου με τεχνολογίες επιστήμης δεδομένων*. Διπλωματική εργασία, Εθνικό Μετσόβιο Πολυτεχνείο, 2021. DOI : http://dx.doi.org/10.26240/heal.ntua.20648.

[41] Tomas Mikolov, Kai Chen, Greg Corrado και Jeffrey Dean. *Efficient Estimation of Word Representations in Vector Space*, 2013. DOI: https://doi.org/10.48550/arXiv.1301.3781.

[42] Jacob Devlin, Ming Wei Chang, Kenton Lee και Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. DOI: https://doi.org/10.48550/arXiv.1810.04805.