Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Πρόβλεψη Δικαστικών Αποφάσεων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΘΕΟΔΩΡΟΣ ΑΓΓΕΛΟΣ ΜΕΞΗΣ

Επιβλέπων : Παναγιώτης Τσανάκας
Καθηγητής Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2025

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Πρόβλεψη Δικαστικών Αποφάσεων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΘΕΟΔΩΡΟΣ ΑΓΓΕΛΟΣ ΜΕΞΗΣ

Επιβλέπων : Παναγιώτης Τσανάκας
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 15η Φεβρουαρίου 2025.

Νικόλαος Σ. Παπασπύρου Καθηγητής Ε.Μ.Π.	Πέτρος Παπαδόπουλος Επίκ. Καθηγητής Ε.Μ.Π.	Γεώργιος Νικολάου Αν. Καθηγητής Ε.Κ.Π.Α.

Αθήνα, Φεβρουάριος 2025

Θεόδωρος Άγγελος Μέξης

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

Περίληψη

Λέξεις κλειδιά

Μηχανική Μάθηση, Τεχνητή Νοημοσύνη, Δικαστικές Αποφάσεις, Πρόβλεψη

Abstract

Key words

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.

Θεόδωρος Άγγελος Μέξης,

Αθήνα, 15η Φεβρουαρίου 2025

Η εργασία αυτή είναι επίσης διαθέσιμη ως Τεχνική Αναφορά CSD-SW-TR-42-17, Εθνικό Μετσόβιο Πολυτεχνείο, Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών, Εργαστήριο Τεχνολογίας Λογισμικού, Φεβρουάριος 2025.

URL:
FTP:

Περιεχόμενα

Κατάλογος σχημάτων

Κεφάλαιο 1

Εισαγωγή

Η νομοθεσία αναπτύσσεται και εξελίσσεται συνεχώς για να ανταποκριθεί στις νέες και μεταβαλλόμενες ανάγκες των κοινωνιών, αντιδρώντας στις κοινωνικές, πολιτικές, οικονομικές και τεχνολογικές αλλαγές. Οι συνεχείς μεταβολές στη νομοθεσία και η ραγδαία αύξηση του αριθμού των δεδικασμένων υποθέσεων προσθέτουν ένα ολοένα αυξανόμενο βάρος στους νομικούς επαγγελματίες. Αυτό οδηγεί φυσικά στο ερώτημα αν μπορεί να παρασχεθεί κάποια μηχανική υποστήριξη στον τομέα αυτό. Ένα τέτοιο σύστημα θα διευκόλυνε τη δουλειά δικηγόρων, εισαγγελέων και δικαστών, καθώς και άλλων συναφών επαγγελματιών, και θα μπορούσε να έχει θετική συμβολή στο δημόσιο συμφέρον εξοικονομώντας χρόνο, μειώνοντας τα λάθη και βελτιώνοντας τη συνέπεια των δικαστικών αποφάσεων. Οι υπολογιστές μπορούν να αναλύσουν τεράστια σύνολα δεδομένα νομικών κειμένων.

1.0.1 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 κειμέναν, 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.

1.0.2 Δικαστήρια

Κεφάλαιο 2

Θεωρητικό υπόβαθρο

2.1 Τεχνικές Προβλέψεων

2.1.1 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.

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.

2.1.2 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 μοντέλως 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 απαραίτητ η ύπαρξη χρονική εξάρτησης. Η συλλογή των δεδομένων αποτελεί συχνά μια χρονοβόρα και ενίοτε δύσκολη διαδικασία, καθώς απαιτείται μεγ άλο 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 [Φ13].

2.1.3 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.

2.2 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*).[1]

2.2.1 Δέντρα Αποφάσεων - Decision Trees

Τα δέντρα αποφάσεων (decision trees) είναι ένας δημοφιλής και διαισθητικός αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται για προβλήματα ταξινόμησης και παλινδρόμησης. 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:

   (a) 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).

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

   (c) 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:

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

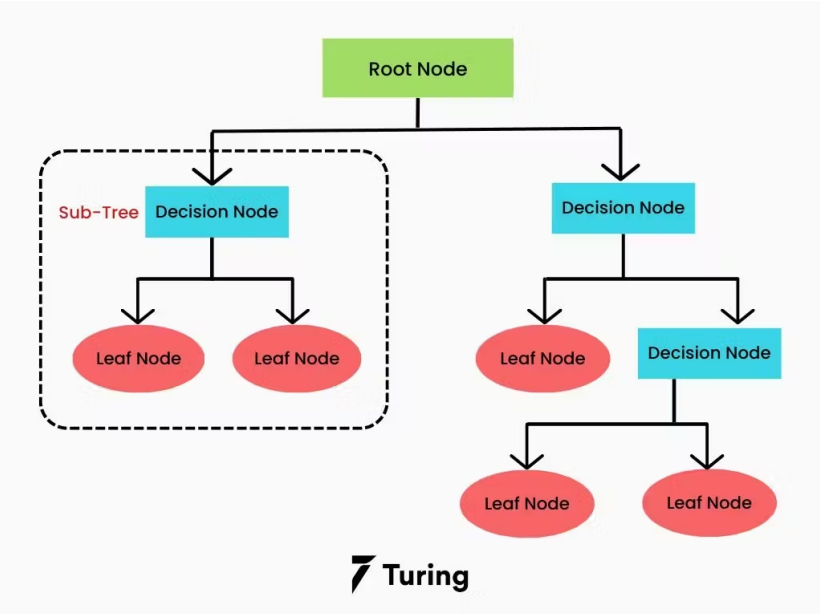
   (b) Έχει φτάσει το μέγιστο προκαθορισμένο βάθος του δέντρου.

   (c) 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 2.1: Μοντέλο Δέντρων Απόφασης

2.2.2 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 [2] 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.

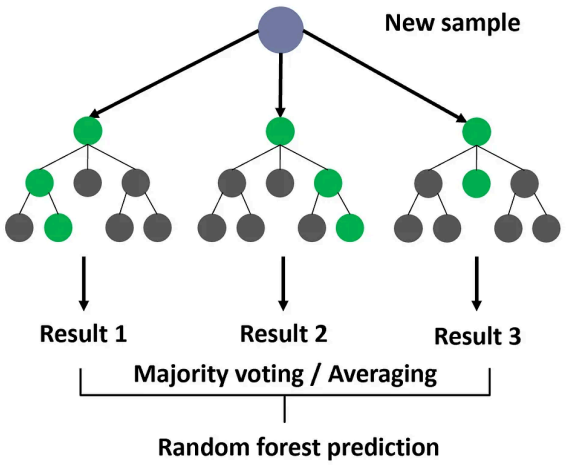
2. Τα τυχαία δάση καταπολεμούν το πρόβλημα της υπερεκπαίδευσης συνδυάζοντας πολλαπλά δέντρα, καθένα από τα οποία εκπαιδεύεται σε διαφορετικό υποσύνολο των δεδομένων εκπαίδευσης. Ο αλγόριθμος χρησιμοποιεί την τεχνική του bootstrap aggregation (ή bagging), δηλαδή επιλέγει τυχαία, με επανάληψη, ένα υποσύνολο 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 ψηφοφορία 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 2.2: Μοντέλο Τυχαίων Δασών

2.2.3 XGBoost Regression

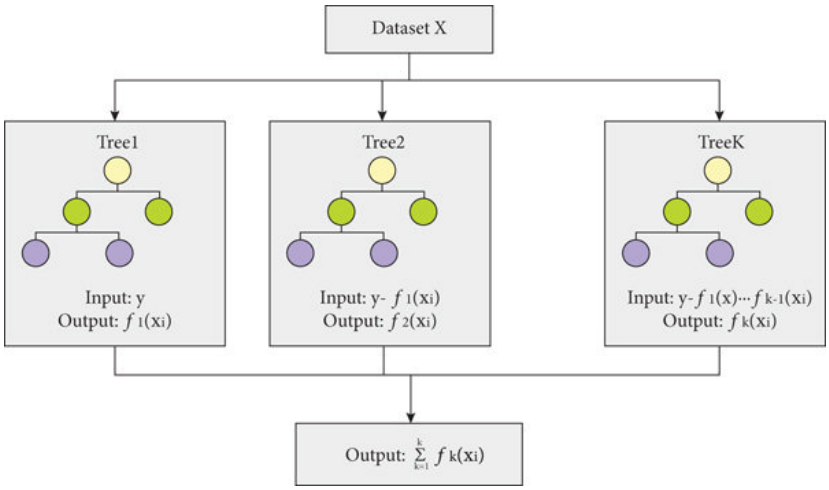
Η μέθοδος XGBoost είναι η συντομογραφία του Extreme Gradient Boosting και αποτελεί επίσης έναν αλγόριθμο ενίσχυσης. Με τον όρο ενίσχυση (boosting), εννοείται μια οικογένεια αλγορίθμων συνδυασμού μοντέλων, οι οποίες αποσκοπούν στη μετατροπή αδύναμων μοντέλων με χαμηλή απόδοση, σε ισχυρά μοντέλα πρόβλεψης. 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 2.3: XGBoost Regression

2.2.4 Μηχανές Υποστήριξης Διανυσμάτων - SVM

Οι Υποστηρικτικές Μηχανές Διανυσμάτων (Support Vector Machines - SVM) είναι ένας ισχυρός αλγόριθμος μηχανικής μάθησης για ταξινόμηση και παλινδρόμηση. 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 δεδομένα 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:

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

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

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

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

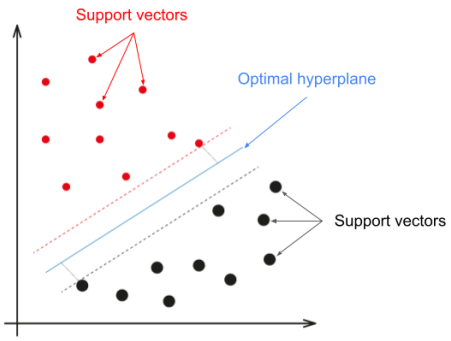
5. Παράμετροι C και γ (*Gamma*): Οι δύο κύριες παράμετροι που ρυθμίζουν την απόδοση της SVM είναι οι παράμετροι C και γ.

   (a) 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.

   (b) 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 2.4: Support Vector Machines

2.2.5 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.

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:

   - y είναι η εξαρτημένη μεταβλητή (η πρόβλεψη).
   - $x_1, x_2, \ldots, x_n$ είναι οι ανεξάρτητες μεταβλητές.
   - $\beta_0(intercept).\beta_1, \beta_2, \ldots, \beta_n$ 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 :

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

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

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

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

   (e) 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.

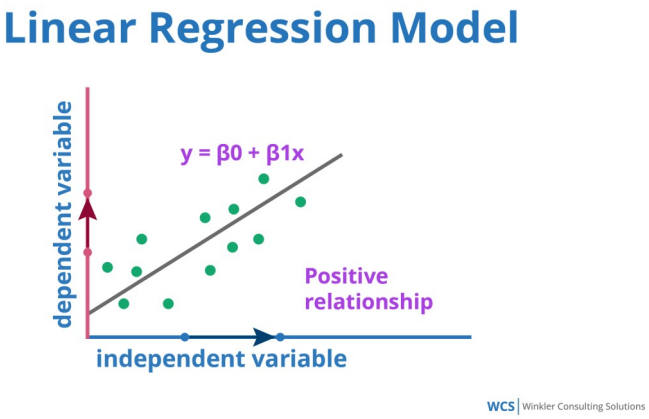
4. Είδη Γραμμικής Παλινδρόμησης:

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

   (b) 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.



Σχήμα 2.5: Linear Regression Model

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 :

$$\text{Y} = \text{b}_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + b_i X_{ip} + e_i, \quad \text{για κάθε δείγμα } i = 1, 2, \ldots, n$$

Στην παραπάνω σχέση, Y είναι η εξαρτημένη μεταβλητή, $\text{X}_{ij} i-jX, j = 1, 2, ..., p., b_i e_i.b_i$, *κανονικών ελάχιστων τ*

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

2.2.6 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.

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 μεταβλητές (χαρακτηριστικά).
- $\beta_0, \beta_1, \ldots, \beta_n$: 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*):

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

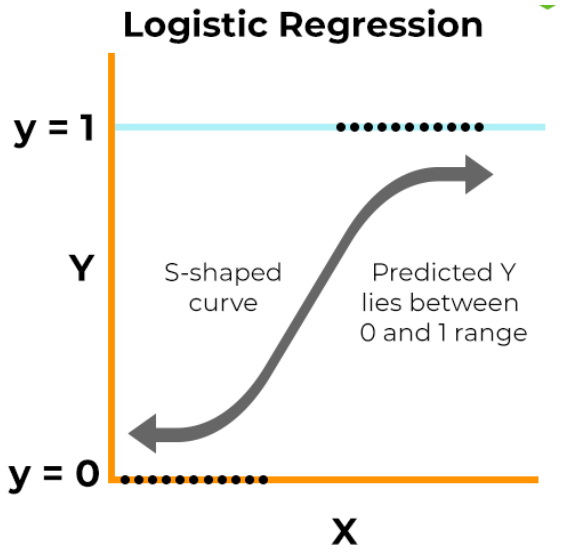
   - Ανεξαρτησία των παρατηρήσεων.
   - Απουσία πολυσυγγραμμικότητας: Οι ανεξάρτητες μεταβλητές δεν πρέπει να είναι ισχυρά συσχετισμένες.

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.

2.3 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.



Σχήμα 2.6: Logistic Regression Model

2.3.1 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. 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.

2.3.2 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*. Ορίζεται ως ο αρμονικός μέσος του *Precision* και του *Recall*:

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

Precision και Recall

Η έννοια του *Precision* (Ευστοχία) και του *Recall* (Ανάκληση) 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*.

2.3.3 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.

2.3.4 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.

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

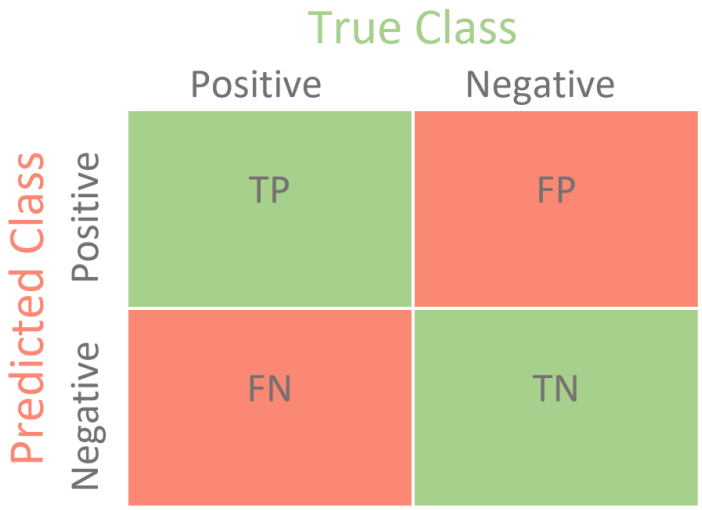
$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Αντίστοιχα, για το δείγμα :

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

2.3.5 Πίνακας Σύγχυσης - Confusion Matrix

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



Σχήμα 2.7: Confusion Matrix

- **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:

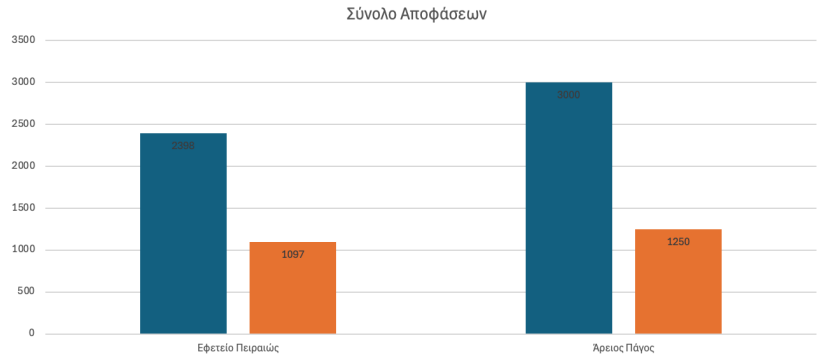
- **Accuracy:**
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$
- **Precision:**
$$\text{Precision} = \frac{TP}{TP + FP}$$
- **Recall (ή Sensitivity):**
$$\text{Recall} = \frac{TP}{TP + FN}$$
- **F1 Score:**
$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

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

Οι δικαστικές αποφάσεις που χρησιμοποιήθηκαν στην παρούσα εργασία συλλέχθηκαν από το Εφετείο Πειραιώς (https://www.efeteio-peir.gr/?page_id=4017) και από τον Άρειο Πάγο (https://www.areiospagos.gr/nomologia/apofaseis.asp). Πρόκειται για αποφάσεις που ελήφθησαν 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 (2009, 2018, 2021, 2022) 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.



Σχήμα 3.1: Αποφάσεις Δικαστηρίων

Η παραπάνω γραφική απεικόνιση παρουσιάζει τον αριθμό αποφάσεων στα δεδομένα που χρησιμοποιούνται για ανάλυση, τόσο στο αρχικό όσο και στο τελικό σύνολο. Συγκεκριμένα:

- 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.

Δικαστήριο	Αριθμός Αποφάσεων	Μ.Ο λέξεων	Αποδοχή	Απόρριψη	Μέγεθος
Άρειος Πάγος	1250	2470	500	750	0.97 MB
Εφετείο Πειραιώς	1197	4361	497	700	65.60 MB

Πίνακας 3.1: Συνοπτικός πίνακας χαρακτηριστικών του συνόλου δεδομένων

3.1 Προεπεξεργασία Δεδομένων

Ο πρωταρχικός στόχος της προεπεξεργασίας των δεδομένων είναι να προετοιμάσουμε το κείμενο, αφαιρώντας περιττούς χαρακτήρες, αριθμούς, και αγγλικούς χαρακτήρες, 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. *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.

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

3. *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.

4. *Αφαίρεση αριθμών :* Οι αριθμοί αφαιρέθηκαν από τα κείμενα, καθώς σε πολλές περιπτώσεις δεν παρέχουν ουσιαστικές πληροφορίες για την ανάλυση, ιδιαίτερα όταν δεν συνδέονται με κρίσιμες πληροφορίες για το νόημα των αποφάσεων.

5. *Αφαίρεση αγγλικών χαρακτήρων :* Επειδή οι δικαστικές αποφάσεις είναι στα ελληνικά, οποιοσδήποτε αγγλικός χαρακτήρας αφαιρέθηκε από τα δεδομένα.

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

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

3.2 Διαδικασία Επισημείωσης

Μετά την ολοκλήρωση της προεπεξεργασίας των κειμένων των δικαστικών αποφάσεων, προχωρήσαμε στην φάση της επισημείωσης των δεδομένων, προκειμένου να κατηγοριοποιήσουμεε 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, η σωστή ετικετοποίησης (*labeling*) των δεδομένων είναι καθοριστική 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 :

(a) *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 pandas 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.

(b) *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).

(c) *Δημιουργία συνόλου δεδομένων :* Μετά την αναζήτηση των λέξεων-στόχων, οι αποφάσεις επισημάνθηκαν κατάλληλα, 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.

Κεφάλαιο 4

Μέθοδοι Πρόβλεψης Δικαστικών Αποφάσεων

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

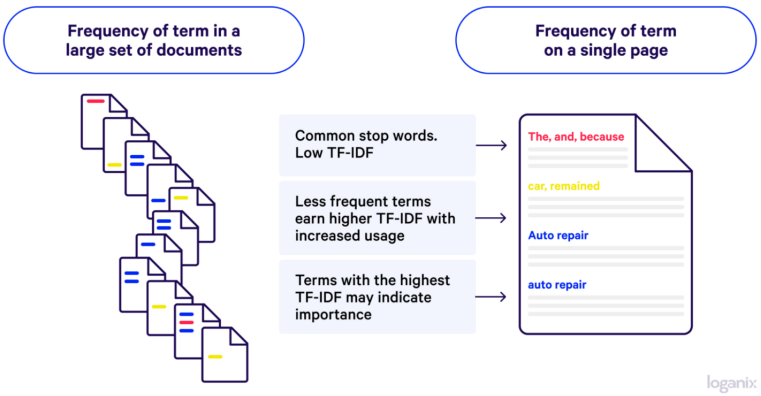
Προκειμένου να εξετάσουμε την αποτελεσματικότητα των διαφόρων ταξινομητών στο σύνολο των αποφάσεων που έχουμε στη διάθεσή μας ήταν απαραίτητο να αναπαραστίσουμε τα κείμενα των αποφάσεων σε αριθμητική μορφή.

Το TF-IDF (*Term Frequency-Inverse Document Frequency*) είναι μία από τις πιο διαδεδομένες και αποτελεσματικές τεχνικές για την αναπαράσταση αυτή, στην Επεξεργασία Φυσικής Γλώσσας (*NLP*). 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.

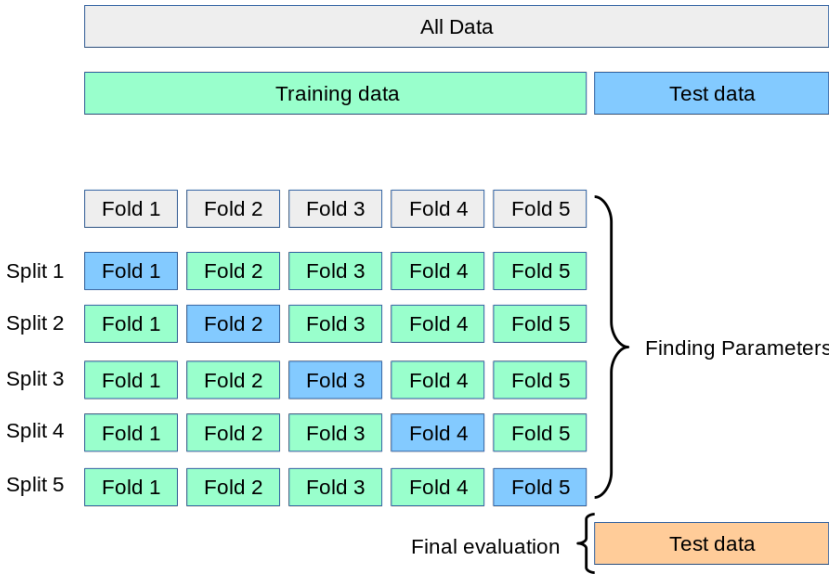
Υπάρχουν ελάχιστες καθολικές συμβουλές σχετικά με την επιλογή των τιμών αυτών, ενώ η τελική επιτυχία της διαδικασίας εξαρτάται σε μεγάλο βαθμό από την εμπειρία του προγραμματιστή. Η επιλογή τους πρέπει να γίνεται με σύνεση, καθώς κάθε παράμετρος που επιλέγεται να ρυθμιστεί μπορεί να αυξήσει εκθετικά τον απαιτούμενο αριθμό των δοκιμών. Αφού επιλεχθούν οι παράμετροι προς ρύθμιση, εφαρμόζονται αλγόριθμοι, οι οποίοι προελαύνουν το χώρο αναζήτησης που 85 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 :

i. **Αναζήτηση Πλέγματος** (*Grid Search*). Η αναζήτηση πλέγματος αποτελεί τον απλούστερο 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.
ii. **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.

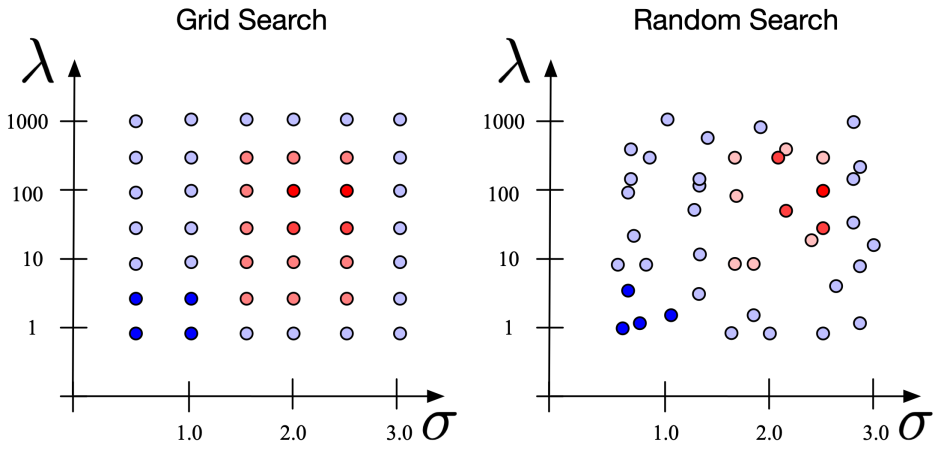
Ένας κύριος λόγος μη αποδοτικότητας ενός μοντέλου μηχανικής μάθησης είναι η υπερπροσαρμογή (*overfitting*). Ο συγκεκριμένος όρος χρησιμοποιείται στην επιβλεπόμενη μάθηση για να δηλώσει την κατάσταση κατά την οποία ένα μοντέλο έχει εκπαιδευτεί και εξειδικευτεί στο σύνολο εκπαίδευσης του προβλήματος που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



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 (textittest 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.

4.3.1 Random Forest

Στην υλοποίηση του μοντέλου Random Forest 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 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 για τις αποφάσεις του Εφετείου Πειραιώς

4.3.2 SVM

Αντίστοιχα, για το μοντέλο SVM (*Support Vector Machine*), 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 για τις αποφάσεις του Αρείου Πάγου

Υπερπαράμετρος	**Τιμή**
C	0.5
gamma	0.12
kernel	rbf

Πίνακας 4.4: Τιμές υπερπαραμέτρων για τις αποφάσεις του Εφετείου Πειραιώς

4.3.3 Logistic Regression

Για την ανάπτυξη του μοντέλου λογιστικής παλινδρόμησης (*Logistic Regression*), 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	11
solver	saga

Πίνακας 4.5: Τιμές υπερπαραμέτρων για τις αποφάσεις του Αρείου Πάγου

Υπερπαράμετρος	**Τιμή**
C	10
penalty	12
solver	liblinear

Πίνακας 4.6: Τιμές υπερπαραμέτρων για τις αποφάσεις του Εφετείου Πειραιώς

4.3.4 Decision Trees

Αντίστοιχα, για το μοντέλο Δέντρων Απόφασης (*Decision Trees*), 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 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 για τις αποφάσεις του Εφετείου Πειραιώς

4.3.5 XGBoost Regression

Αυτή η υλοποίηση του XGBoost 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 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 για τις αποφάσεις του Εφετείου Πειραιώς

Κεφάλαιο 5

Πειραματικά Αποτελέσματα - Ερμηνεία

Στο κεφάλαιο αυτό περιγράφεται η υλοποίηση των μοντέλων πρόβλεψης, παρουσιάζονται τα αποτελέσματά εκτέλεσής τους όσον αφορά την ποιότητα της πρόβλεψης, 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 εκπαίδευσης (*train set*), καθώς και κατά τη διαδικασία του cross-validation, στο σύνολο αξιολόγησης (*validation set*). Η τελική αξιολόγηση και σύγκριση των μοντέλων έγινε σύμφωνα με την απόδοσή τους στο σύνολο δοκιμής (test set). Η αξιολόγηση των προβλέψεων έγινε χρησιμοποιώντας τις μετρικές απόδοσης που παρουσιάστηκαν στην Ενότητα 2.3.

5.1 Αξιολόγηση αποτελεσμάτων για αποφάσεις Εφετείου Πειραιώς

Στην υποενότητα αυτή γίνεται η συνοπτική παρουσίαση των αποτελεσμάτων των πειραμάτων που πραγματοποιήθηκαν για το σύνολο δεδομένων 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*), στο σύνολο αξιολόγησης (*validation set*), κατά τη διαδικασία του cross-validation, καθώς και στο σύνολο εκπαίδευσης (*train set*).

5.1.1 Απόδοση μοντέλων στο σύνολο εκπαίδευσης

	RF	**DT**	**SVM**	**XGB**	**LR**
Precision	0.75	0.72	0.71	0.76	0.72
Recall	0.90	0.69	0.92	0.90	0.83
Accuracy	0.76	0.62	0.71	0.76	0.74
F1-Score	0.82	0.70	0.80	0.83	0.80

Πίνακας 5.1: Μετρικές απόδοσης στο train set

5.1.2 Απόδοση μοντέλων στο σύνολο αξιολόγησης

	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 5.2: Μετρικές απόδοσης στο validation set

5.1.3 Απόδοση μοντέλων στο σύνολο δοκιμής

	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 5.3: Μετρικές απόδοσης στο test set

Ο Πίνακας 5.3 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.

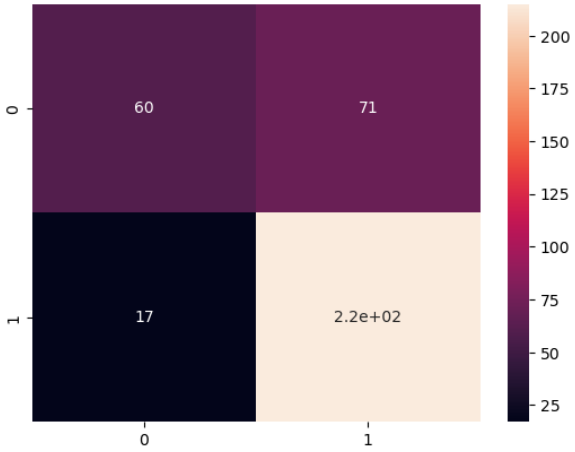
Το Support Vector Machine (SVM) παρουσιάζει τη μεγαλύτερη τιμή Recall (0.92) από όλα τα μοντέλα, γεγονός που σημαίνει ότι ανιχνεύει σχεδόν όλες τις θετικές περιπτώσεις. Ωστόσο, το Precision του (0.71) είναι χαμηλό, γεγονός που υποδηλώνει αυξημένο αριθμό ψευδώς θετικών προβλέψεων. Το χαρακτηριστικό αυτό 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.

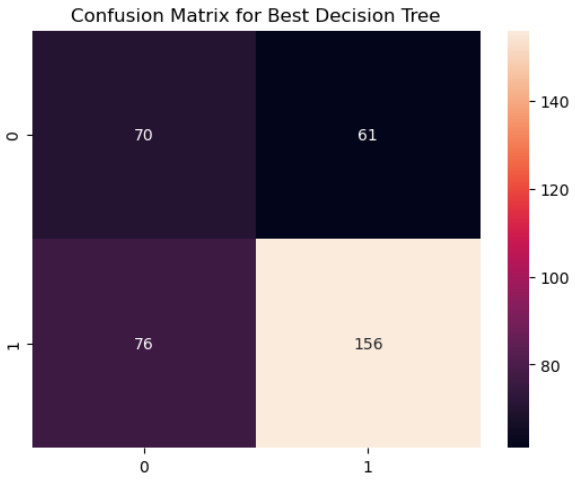
Συνολικά, ο XGBoost αναδεικνύεται ως ο πιο αποτελεσματικός ταξινομητής, ενώ το Random Forest μπορεί να προτιμηθεί σε περιπτώσεις όπου η ανίχνευση θετικών περιπτώσεων είναι υψίστης σημασίας. Η επιλογή του μοντέλου θα πρέπει να καθορίζεται από τη σχετική σημασία της ακρίβειας (Precision) έναντι της ανάκλησης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

5.1.4 Confusion Matrix μοντέλων

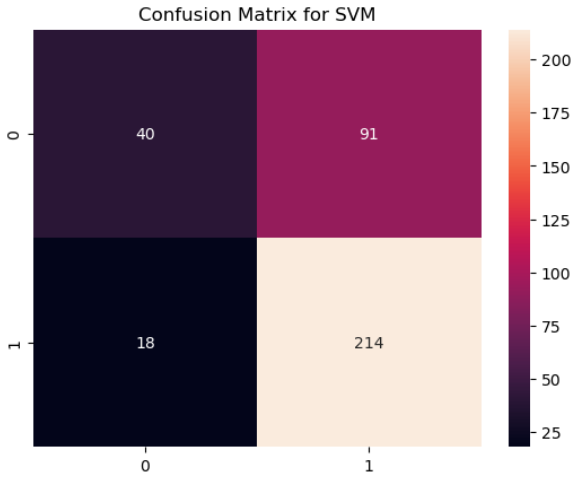
Στην υποενότητα αυτή, παρουσιάζονται γραφικά οι πίνακες σύγχυσης (*confusion matrix*) που προκύπτουν από τα πειράματα πρόβλεψης, παρέχοντας μια οπτική αναπαράσταση της απόδοσης των αλγορίθμων. Κάθε πίνακας δείχνει την κατανομή των σωστών και λανθασμένων προβλέψεων όπως έχει αναλυθεί στην υποενότητα Ενότητα 2.3.



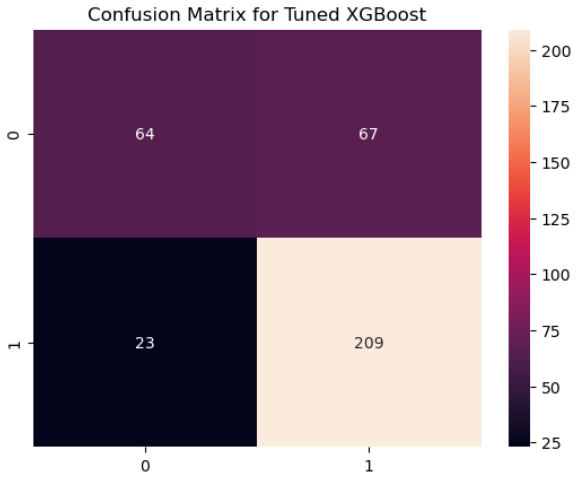
Σχήμα 5.1: Randrom Forest CM



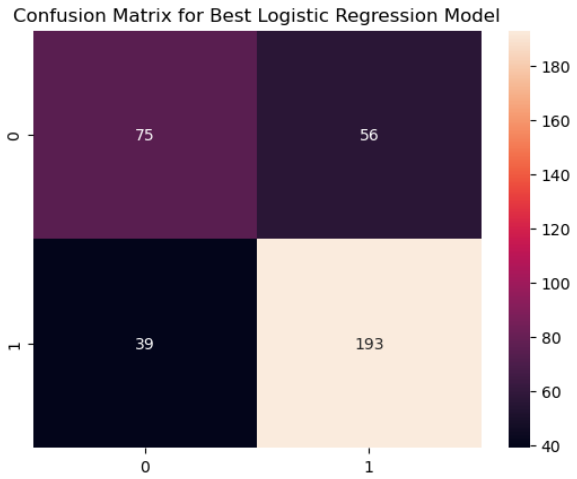
Σχήμα 5.2: Decision Trees CM



Σχήμα 5.3: SVM CM



Σχήμα 5.4: XGBoost CM



Σχήμα 5.5: Logistic Regression CM

5.2 Αξιολόγηση αποτελεσμάτων για αποφάσεις Αρείου Πάγου

Ομοίως, 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*), στο σύνολο αξιολόγησης (*validation set*), κατά τη διαδικασία του cross-validation, καθώς και στο σύνολο εκπαίδευσης (*train set*).

5.2.1 Απόδοση μοντέλων στο σύνολο εκπαίδευσης

	RF	**DT**	**SVM**	**XGB**	**LR**
Precision	0.75	0.72	0.71	0.88	0.77
Recall	0.90	0.69	0.92	0.86	0.83
Accuracy	0.76	0.62	0.71	0.85	0.74
F1-Score	0.82	0.70	0.80	0.88	0.80

Πίνακας 5.4: Μετρικές απόδοσης στο train set

5.2.2 Απόδοση μοντέλων στο σύνολο αξιολόγησης

	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 5.5: Μετρικές απόδοσης στο validation set

5.2.3 Απόδοση μοντέλων στο σύνολο δοκιμής

	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 5.6: Μετρικές απόδοσης στο test set

Ο *Πίνακας 5.6* 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.

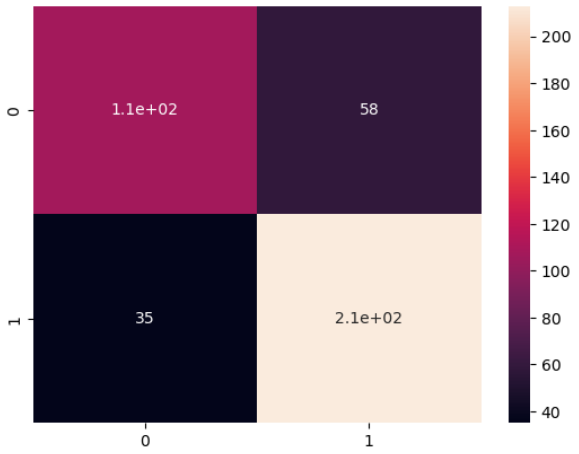
Αντίθετα, ο Decision Tree (DT) έχει τη χαμηλότερη απόδοση μεταξύ των μοντέλων, με ακρίβεια 0.71 και χαμηλότερη τιμή recall (0.77). Αυτό το καθιστά λιγότερο αξιόπιστο, καθώς τα αποτελέσματα του έχουν μεγαλύτερη πιθανότητα λάθους τόσο στις θετικές όσο και στις αρνητικές προβλέψεις.

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.

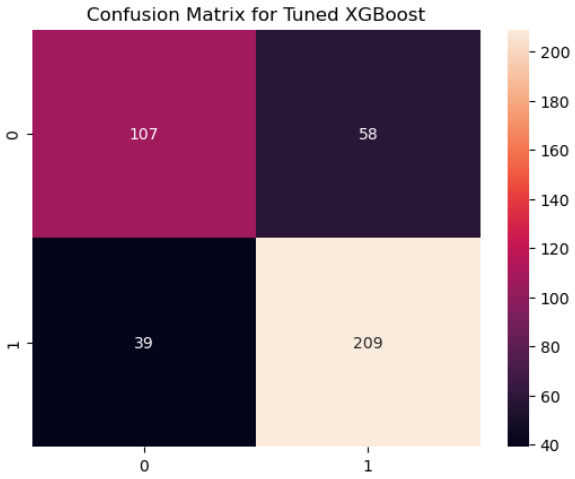
Τέλος, η Logistic Regression (LR) έχει παρόμοιες επιδόσεις με τον SVM, με ευαισθησία (0.80), ακρίβεια (0.76) και F1-Score (0.78). Παρόλο που δεν ξεχωρίζει, παραμένει μια σταθερή επιλογή με αποδεκτή ακρίβεια και ανάκληση, ειδικά για πιο απλές εφαρμογές.

Συνολικά, ο RF και ο XGB ξεχωρίζουν ως οι πιο αξιόπιστες επιλογές για την ανάλυση, ενώ οι DT και SVM αποδεικνύονται λιγότερο αποδοτικοί.

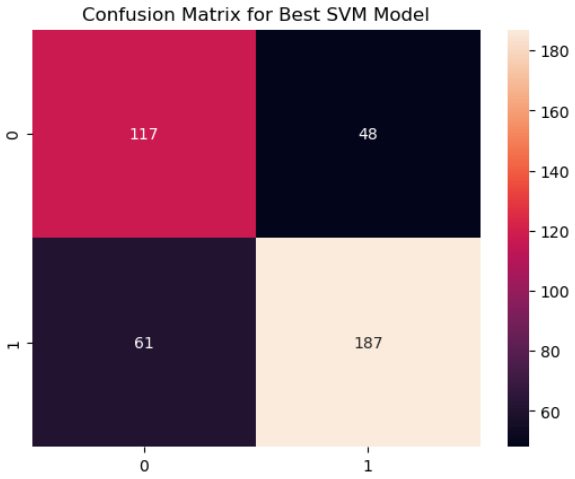
5.2.4 Confusion Matrix μοντέλων



Σχήμα 5.6: Random Forest CM

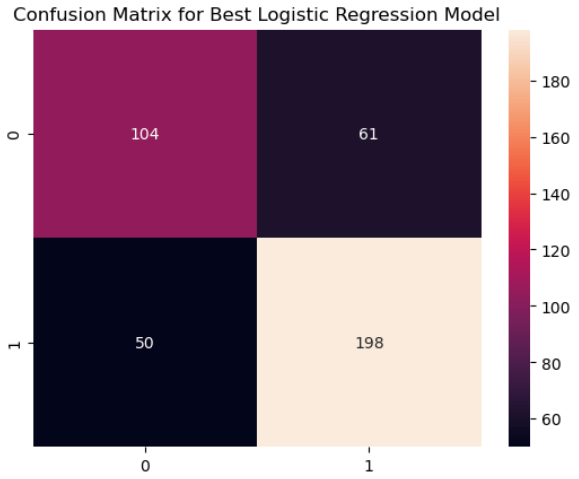


Σχήμα 5.7: XGBoost CM



Σχήμα 5.8: SVM CM

Σχήμα 5.9: Decision Trees CM



Σχήμα 5.10: Logistic Regression CM

5.3 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 (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.

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

Στην παρούσα διπλωματική εργασία μελετήθηκε το πρόβλημα της πρόβλεψης δικαστικών αποφάσεων με χρήση τεχνικών επιστήμης δεδομένων και μηχανικής μάθησης. Αρχικά έγινε μια εκτεταμένη καταγραφή και μελέτη σχετικών εργασιών τεχνολογιών αιχμής, καθώς και συλλογή, και παρουσίαση διαθέσιμων συνόλων δεδομένων.

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.

Βιβλιογραφία

[1] 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.

[2] Leo Breiman. The creation, structure, and interpretation of the legal text. *Machine Learning*, 45, 2001.

[3] Lior Rokach and Oded Maimon. Decision trees. *The Data Mining and Knowledge Discovery Handbook*, 6, 2005.

[4] P. Tiersma. The creation, structure, and interpretation of the legal text. 2022.

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

[6] Πετρόπουλος Φ. and Ασημακόπουλος Β. Επιχειρησιακές Προβλέψεις. *Συμμετρία*, 2013.