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, Ιούνιος 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 20η Ιουνίου 2025.

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

Αθήνα, Ιούνιος 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

Ευχαριστώ θερμά τον επιβλέποντα καθηγητή αυτής της διπλωματικής εργασίας, κ. Παναγιώτη Τσανάκα, για την συνεχή καθοδήγηση και εμπιστοσύνη του. Ευχαριστώ επίσης τον κ. Μάριο Κονιάρη για τις πολύτιμες συμβουλές του και τις ιδιαίτερα χρήσιμες συζητήσεις που είχαμε καθόλη την διάρκεια της εκπόνησης της διπλωματικής μου εργασίας. Θέλω να ευχαριστήσω ακόμη, τον συμφοιτητή μου Δημήτριο Χαραλάμπη για διάφορα πραγματάκια.

Θεόδωρος Άγγελος Μέξης,

Αθήνα, 20η Ιουνίου 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, Ιούνιος 2025.

URL:
FTP:

Περιεχόμενα

Περίληψη 5

Abstract 7

Ευχαριστίες 9

Περιεχόμενα 11

Κατάλογος σχημάτων 13

1. Εισαγωγή 15

2. Θεωρητικό υπόβαθρο 17
2.1 Τεχνικές Προβλέψεων 17
2.1.1 Ορισμός και Διαδικασία Πρόβλεψης 17
2.1.2 Κατηγορίες Μεθόδων Πρόβλεψης 17

3. Παρουσίαση Συνόλου Δεδομένων 19
3.1 Προεπεξεργασία Δεδομένων 19
3.2 Διαδικασία Επισημείωσης 20

4. Μέθοδοι Πρόβλεψης Δικαστικών Αποφάσεων 21
4.1 Προετοιμασία Κειμένων 21
4.2 Μέθοδοι Ταξινόμησης 21

5. Πειραματικά Αποτελέσματα - Ερμηνεία 23

6. Επίλογος 25

Παράρτημα 27

A. Ευρετήριο συμβολισμών 27

B. Ευρετήριο γλωσσών 29

C. Ευρετήριο αριθμών 31

Κατάλογος σχημάτων

Κεφάλαιο 1

Εισαγωγή

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. Πιο συγκεκριμένα, στην συγκεκριμένη 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 τα επιλεγμένα μοντέλα και προβλέψεις 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 διακρίνονται σε τρεις μεγάλες κατηγορίες σύμφωνα με τη διαδικασία παραγωγής τους:

Ποσοτικές Μέθοδοι. Οι ποσοτικές μέθοδοι αναφέρονται στην εφαρμογή στατιστικών μοντέλων χρονοσειρών ή αιτιοκρατικών μοντέλων επί μιας σειράς δεδομένων με σκοπό την αυτοματοποιημένη και συστηματική παραγωγή προβλέψεων. Οι στατιστικές προβλέψεις είναι άμεσα εφαρμόσιμες και αποδεκτά ακριβείς, ειδικότερα αν συνδυαστούν με κατάλληλα διαστήματα εμπιστοσύνης.

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, 2022 από τα συγκεκριμένα δικαστήρια και καλύπτουν διάφορους τομείς του δικαίου, παρέχοντας έτσι ένα πλούσιο και ποικιλόμορφο σύνολο δεδομένων για ανάλυση. 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.

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 φράσεων-κλειδιά (*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 PDF, χρησιμοποιήθηκε ειδικό εργαλείο εξαγωγής κειμένου, το οποίο επιτρέπει την ανάκτηση του πλήρους περιεχομένου των αποφάσεων. Αντίστοιχα, για τα αρχεία HTML, χρησιμοποιήθηκαν κατάλληλα εργαλεία κώδικα για την 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, καθορίσαμε μια λίστα από φράσεις-κλειδιά (*regular expressions*). Πιο συγκεκριμένα, δημιουργήθηκε μια λίστα με φράσεις-κλειδιά, τα οποία χρησιμοποιούνται συνήθως σε αποφάσεις που καταλήγουν σε αποδοχή. Οι φράσεις αυτές, παραδείγματος χάριν «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

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 TfidfVectorizer(), 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. Οι τιμές που χρησιμοποιήθηκαν σε αυτήν την εργασία επιλέχθηκαν εμπειρικά, έπειτα ωστόσο από πολλές δοκιμές. 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.

4.2 Μέθοδοι Ταξινόμησης

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*).

4.2.1 Τυχαία Δάση - Random Forests

Στην παρούσα προσέγγιση για τα Τυχαία Δάση (Ŕandom Forests) έγινε χρήση ενός πολυεπίπεδου συστήματος βελτιστοποίησης των υπερπαραμέτρων, προκειμένου να βρεθεί το βέλτιστο κούρδισμα τους. Αρχικά, χωρίσαμε τα δεδομένα μέσω τυχαίου καταμερισμού σε εκπαιδευτικό και δοκιμαστικό σύνολο, με σκοπό τη διατήρηση της ισορροπίας στις δύο κατηγορίες τους. Στη συνέχεια, αξιοποίησαμε τις τεχνικές GridSearchCV και RandomizedSearchCV, αναζητώντας έτσι τον βέλτιστο συνδυασμό υπερπαραμέτρων, όπως ο αριθμός των δέντρων και το βάθος τους. Πιο συγκεκριμένα, το RandomizedSearchCV προσφέρει τη δυνατότητα εξερεύνησης ενός ευρύτερου εύρους υπερπαραμέτρων πιο γρήγορα, καθώς αντί να εξετάζει κάθε πιθανό συνδυασμό, δοκιμάζει ένα υποσύνολο.

Κεφάλαιο 5

Πειραματικά Αποτελέσματα - Ερμηνεία

Κεφάλαιο 6

Επίλογος

Παράρτημα Α

Ευρετήριο συμβολισμών

$A \to B$: συνάρτηση από το πεδίο A στο πεδίο B.

Παράρτημα Β

Ευρετήριο γλωσσών

C++: πώς θα βγάλω λεφτά...
Haskell: η γλώσσα της ζωής μου αλλά πάνε οι μπύρες...
Javascript: χα, χα, χα...
Python: πώς θα τελειώνω για να πάω για μπύρες...

Παράρτημα C

Ευρετήριο αριθμών

17: ask Zachos.
42: life, the universe and everything — ask Douglas.