Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Ηλεκτρονικών Υπολογιστών

Τεχνικές και Εργαλεία Ελέγχου Γεγονότων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Του
Διονύσιου Πετρόπουλου,
Ιωάννη Κώνστα

Επιβλέπων Καθηγητής:
Ιωάννης Ζαχαράκης

ΠΑΤΡΑ 2023

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή

Πάτρα, / /2023

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ

1.

2.

3.

Υπεύθυνη Δήλωση Φοιτητή

Βεβαιώνω ότι είμαι συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης έχω αναφέρει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επίσης βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά ειδικά για τη συγκεκριμένη εργασία.

Η έγκριση της διπλωματικής εργασίας από το Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Πανεπιστημίου Πελοποννήσου δεν υποδηλώνει απαραιτήτως και αποδοχή των απόψεων του συγγραφέα εκ μέρους του Τμήματος.

Ευχαριστίες

Θα θέλαμε να ευχαριστήσουμε τον Καθ. κ. Ιωάννη Ζαχαράκη που με την καθοδήγηση και τις συμβουλές του συντέλεσε στην εκπόνηση της παρούσας διπλωματικής εργασίας. Επιπλέον θα θέλαμε να ευχαριστήσουμε τις οικογένειές μας για την στήριξη που μας προσέφεραν όλα αυτά τα χρόνια.

Περίληψη

Στην παρούσα διπλωματική εργασία θα μελετηθεί ο έλεγχος γεγονότων, ο οποίος αναφέρεται σε τεχνικές που επιδιώκουν την επαλήθευση πραγματικών πληροφοριών, προκειμένου να αναδειχθεί η ακρίβεια και η ορθότητα του περιεχομένου. Ως πρόβλημα, σήμερα, απασχολεί την ερευνητική κοινότητα και αποτελεί μία επιτακτική ανάγκη λόγω της πληθώρας της πληροφορίας που αναπαράγεται άκριτα και πολλές φορές με στόχους που εξυπηρετούν ιδιοτελείς σκοπούς. Στο πλαίσιο αυτό θα πραγματοποιηθεί η ανάπτυξη μίας εφαρμογής με αξιοποίηση των διαθέσιμων εργαλείων και θα επιδειχθεί σύγκριση της αποτελεσματικότητάς της μεταξύ υπαρχόντων αλγορίθμων.

Σκοπός της διπλωματικής είναι να εξεταστεί η δυνατότητα αξιοποίησης τεχνικών Μηχανικής Μάθησης με χρήση Επεξεργασίας Φυσικής Γλώσσας για την βιωσιμότητα μιας τέτοιας εφαρμογής.

Λέξεις Κλειδιά

Μηχανική Μάθηση, Επεξεργασία Φυσικής Γλώσσας, Εξόρυξη Δεδομένων, Λογιστική Παλινδρόμηση, Δέντρα Αποφάσεων, Τυχαίο Δάσος, Naïve Bayes, Μηχανή Διανυσμάτων Υποστήριξης, Στοχαστική Κάθοδος Κλίσης, Πολυωνυμικός Naïve Bayes, Ψευδοειδήσεις

Abstract

In this thesis we will study fact checking, which refers to techniques that seek to verify factual information in order to highlight the accuracy and correctness of the content. As a problem, it is currently a concern for the research community and an urgent need due to the abundance of information that is uncritically reproduced, often with goals that serve self-serving purposes. In this context, an application will be developed using the available tools and a comparison of its effectiveness among existing algorithms will be demonstrated.

The aim of the thesis is to explore the possibility of utilizing Machine Learning techniques with the use of Natural Language Processing for the viability of these kinds of applications.

Key Words

Machine Learning, Natural Language Processing, Data Mining, Logistic Regression, Decision Trees, Random Forest, Naïve Bayes, Support Vector Machine, Stochastic Gradient Descent, Multinomial Naïve Bayes, Fake News

Περιεχόμενα

Κεφάλαιο 1 Εισαγωγή

Αντικείμενο της παρούσας διπλωματικής εργασίας αποτελεί ο έλεγχος γεγονότων και συγκεκριμένα ο εντοπισμός των ψευδών ειδήσεων/πληροφοριών στα Μέσα Μαζικής Ενημέρωσης, καθώς και η μελέτη τεχνικών μηχανικής μάθησης για κατηγοριοποίηση κειμένου με σκοπό την καταπολέμηση του φαινομένου της παραπληροφόρησης.

Τα τελευταία χρόνια, η ραγδαία επέκταση του Διαδικτύου σε συνδυασμό με την εξοικείωση των ανθρώπων στα τεχνολογικά μέσα έχει ορίσει τα κοινωνικά δίκτυα (social media) και τις διαδικτυακές ιστοσελίδες ως τις κύριες πηγές πληροφόρησης για το μεγαλύτερο μέρος του πληθυσμού. Στην Ελλάδα, για το έτος 2021, περισσότερο από το 69% του πληθυσμού επιλέγει την άντληση ενημέρωσης μέσω των κοινωνικών δικτύων [1], ποσοστό σημαντικά υψηλό για το μέγεθος και τον πληθυσμό της χώρας. Η προσβασιμότητα, η τάχιστη διάδοση πληροφορίας, η διαδραστικότητα, το χαμηλό κόστος και η ευρεία αποδοχή τους έχει οδηγήσει κυβερνήσεις, οργανισμούς και δημόσια πρόσωπα να τα επιλέξουν ως κύριο μέσο ενημέρωσης των ατόμων. Ωστόσο, τα θετικά αυτά χαρακτηριστικά για την διάδοση της πληροφορίας συνοδεύονται με το μειονέκτημα ότι, χρόνο με το χρόνο, παρατηρείται αύξηση στη διασπορά των ψευδών ειδήσεων (fake news) στα social media. Το φαινόμενο αυτό ονομάζεται παραπληροφόρηση και έχει γίνει η μάστιγα της τεχνολογικής εποχής που ζούμε. Εξελίσσεται συνεχώς και αναπαράγεται με εκθετική άνοδο, ιδιαίτερα από οργανισμούς ή άτομα που αποσκοπούν στην ώθηση της ατζέντας τους.

Αν και το πρόβλημα των ψευδών ειδήσεων δεν είναι καινούργιο, η ανίχνευσή τους αποτελεί ένα πολύπλοκο έργο, δεδομένου της έλλειψης ελέγχου και εξάπλωσης του ψευδούς περιεχομένου αλλά και ότι οι άνθρωποι τείνουν να πιστεύουν παραπλανητικές πληροφορίες. Είναι δύσκολο για τους ανθρώπους να εντοπίσουν τις ψευδείς ειδήσεις. Μπορεί να υποστηριχθεί ότι ο μόνος τρόπος για ένα άτομο να εντοπίσει χειροκίνητα τις ψευδείς ειδήσεις είναι να έχει μια τεράστια ευρυμάθεια πάνω στο καλυπτόμενο θέμα. Ωστόσο, ακόμα και με αυτή τη γνώση, είναι εξαιρετικά δύσκολο να εντοπιστούν με επιτυχία και η διαδικασία είναι ιδιαίτερα χρονοβόρα, απασχολώντας συνάμα πολύ ανθρώπινο δυναμικό.

Στην παρούσα εργασία παρουσιάζουμε μια προσέγγιση βασισμένη σε αλγόριθμους μηχανικής μάθησης υπό επίβλεψη με χρήση επεξεργασίας φυσικής γλώσσας, με σκοπό την ταξινόμηση κειμένου ως μια προσπάθεια αντιμετώπισης του φαινομένου της παραπληροφόρησης. Συγκεκριμένα, θα συγκρίνουμε εφτά διαφορετικές επιβλεπόμενες μεθόδους ταξινόμησης, ονομαστικά, Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, Stochastic Gradient Descent και Support Vector Machines. Πιστεύουμε πως μελλοντικά, τέτοιες τεχνικές μπορούν να υιοθετηθούν μαζικά για την καταπολέμηση του εν λόγω φαινομένου, μειώνοντας σημαντικά τον χρόνο, το κόστος και το εργατικό δυναμικό σε σχέση με τον μακροχρόνια εδραιωμένο χειροκίνητο τρόπο ελέγχου γεγονότων.

Η εργασία αυτή αποτελείται από τρία βασικά μέρη. Αρχικά, περιγράφεται η γενική θεωρία που χρησιμοποιείται, περιλαμβάνοντας τους τομείς της μηχανικής μάθησης, της επεξεργασίας φυσικής γλώσσας και της εξόρυξης δεδομένων. Στη

συνέχεια, αναλύονται οι αλγόριθμοι που χρησιμοποιούνται για την εκπαίδευση και την ταξινόμηση των ψευδοειδήσεων, περιγράφοντας τις λειτουργίες και τις μεθόδους που περιλαμβάνουν. Τέλος, παρουσιάζεται η υλοποίηση του προγράμματος, περιγράφοντας την δομή, τις βιβλιοθήκες και τα εργαλεία που χρησιμοποιήθηκαν για την υλοποίηση του συστήματος καθώς και τις τεχνικές αξιολόγησής του.

Κεφάλαιο 2

2.1 Μηχανική Μάθηση

Ορισμός 1: Με βάση τον Tom Mitchell, ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από την εμπειρία E, ως προς κάποια κλάση εργασιών T και μέτρο απόδοσης P, αν η απόδοσή του σε εργασίες της κλάσης T, όπως μετριέται από το P, βελτιώνεται μέσω της εμπειρίας E [2].

Ορισμός 2: Ένα υπολογιστικό σύστημα ενσωματώνει Μηχανική Μάθηση, όταν κάτι αλλάζει στην συμπεριφορά του και μαθαίνει από αυτό, με τέτοιο τρόπο, έτσι ώστε, να αποδίδει καλύτερα στο μέλλον [3].

Η Μηχανική Μάθηση - ΜΜ (Machine Learning - ML) είναι ένας κλάδος της Τεχνητής Νοημοσύνης -ΤΝ (Artificial Intelligence - AI) που παρέχει στις μηχανές την ικανότητα να μαθαίνουν αυτόματα από δεδομένα και προηγούμενες εμπειρίες, ενώ παράλληλα εντοπίζουν μοτίβα για να κάνουν προβλέψεις με ελάχιστη ανθρώπινη παρέμβαση.

Οι μέθοδοι Μηχανικής Μάθησης επιτρέπουν στους υπολογιστές να λειτουργούν αυτόνομα χωρίς ρητό προγραμματισμό. Οι εφαρμογές ML τροφοδοτούνται με νέα δεδομένα και μπορούν να μαθαίνουν, να αναπτύσσονται, να εξελίσσονται και να προσαρμόζονται ανεξάρτητα.

Η Μηχανική Μάθηση αντλεί διορατικές πληροφορίες από μεγάλο όγκο δεδομένων αξιοποιώντας αλγορίθμους για τον εντοπισμό μοτίβων και μαθαίνει με μια επαναληπτική διαδικασία. Οι ML αλγόριθμοι χρησιμοποιούν μεθόδους υπολογισμού για να μαθαίνουν απευθείας από τα δεδομένα αντί να βασίζονται σε οποιαδήποτε προκαθορισμένη εξίσωση που μπορεί να χρησιμεύσει ως μοντέλο.

2.2 Τύποι Μηχανικής Μάθησης

Οι προσεγγίσεις Μηχανικής Μάθησης χωρίζονται παραδοσιακά σε τρεις μεγάλες κατηγορίες [4], οι οποίες αντιστοιχούν σε παραδείγματα μάθησης, ανάλογα με τη φύση του "σήματος" ή της "ανατροφοδότησης" που διαθέτει το σύστημα μάθησης, συγκεκριμένα:

2.2.1 Επιβλεπόμενη Μηχανική Μάθηση

Αυτό το είδος μάθησης θα χρησιμοποιηθεί στην παρούσα εφαρμογή. Οι αλγόριθμοι μάθησης με επίβλεψη δημιουργούν ένα μαθηματικό μοντέλο ενός συνόλου δεδομένων που περιέχει τόσο τις εισόδους όσο και τις επιθυμητές εξόδους. Τα δεδομένα είναι γνωστά ως δεδομένα εκπαίδευσης και αποτελούνται από ένα σύνολο παραδειγμάτων εκπαίδευσης. Κάθε παράδειγμα εκπαίδευσης έχει μία ή περισσότερες εισόδους και την επιθυμητή έξοδο, γνωστή και ως εποπτικό σήμα (supervisory signal). Στο μαθηματικό μοντέλο, κάθε παράδειγμα εκπαίδευσης αναπαρίσταται από έναν πίνακα ή διάνυσμα, που ονομάζεται διάνυσμα χαρακτηριστικών (feature vector) και τα δεδομένα εκπαίδευσης αναπαρίστανται από έναν πίνακα. Μέσω της επαναληπτικής βελτίωσης μιας αντικειμενικής συνάρτησης, οι αλγόριθμοι μάθησης με επίβλεψη μαθαίνουν μια συνάρτηση που μπορεί να χρησιμοποιηθεί για την πρόβλεψη της εξόδου που σχετίζεται με νέες εισόδους. Μια βέλτιστη συνάρτηση θα επιτρέψει στον αλγόριθμο να προσδιορίσει σωστά την έξοδο για εισόδους που δεν αποτελούσαν μέρος των δεδομένων εκπαίδευσης. Ένας αλγόριθμος που βελτιώνει την ακρίβεια των εξόδων ή των προβλέψεών του με την πάροδο του χρόνου θεωρείται ότι έχει μάθει να εκτελεί το συγκεκριμένο έργο.

Η Επιβλεπόμενη Μηχανική Μάθηση (Supervised Learning) ταξινομείται περαιτέρω σε δύο μεγάλες κατηγορίες:

Ταξινόμηση (Classification): Κατηγοριοποίηση διακριτών τιμών, για παράδειγμα, ναι ή όχι, αληθές ή ψευδές κ.λπ. Οι πραγματικές εφαρμογές αυτής της κατηγορίας είναι εμφανείς στην ανίχνευση ανεπιθύμητων μηνυμάτων και στο φιλτράρισμα ηλεκτρονικού ταχυδρομείου.

Παλινδρόμηση (Regression): Οι αλγόριθμοι παλινδρόμησης χειρίζονται προβλήματα κατηγοριοποίησης συνεχών τιμών όπου οι μεταβλητές εισόδου και εξόδου έχουν γραμμική σχέση. Είναι γνωστοί για την πρόβλεψη συνεχών μεταβλητών εξόδου. Παραδείγματα περιλαμβάνουν την πρόβλεψη του καιρού, την ανάλυση τάσεων της αγοράς κ.λπ.

2.2.2 Μη Επιβλεπόμενη Μάθηση

Η Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning) αναφέρεται σε μια τεχνική μάθησης χωρίς επίβλεψη. Εδώ, η μηχανή εκπαιδεύεται χρησιμοποιώντας ένα μη επισημασμένο σύνολο δεδομένων και έχει τη δυνατότητα να προβλέψει την έξοδο χωρίς καμία επίβλεψη. Ένας αλγόριθμος μάθησης χωρίς επίβλεψη στοχεύει στην ομαδοποίηση του μη ταξινομημένου συνόλου δεδομένων με βάση τις ομοιότητες, τις διαφορές και τα μοτίβα της εισόδου.

Η μη επιβλεπόμενη μηχανική μάθηση ταξινομείται περαιτέρω σε δύο τύπους:

Ομαδοποίηση (Clustering): Η τεχνική ομαδοποίησης αναφέρεται στην ομαδοποίηση αντικειμένων σε ομάδες με βάση παραμέτρους όπως ομοιότητες ή διαφορές μεταξύ των αντικειμένων.

Συσχέτιση (Association): Η μάθηση συσχετίσεων αναφέρεται στον εντοπισμό τυπικών σχέσεων μεταξύ των μεταβλητών ενός μεγάλου συνόλου δεδομένων. Προσδιορίζει την εξάρτηση διαφόρων στοιχείων δεδομένων και χαρτογραφεί τις συσχετιζόμενες μεταβλητές.

2.2.3 Ενισχυτική Μάθηση

Η Ενισχυτική Μάθηση (Reinforcement Learning) είναι ένας τομέας της Μηχανικής Μάθησης που ασχολείται με το πώς οι πράκτορες λογισμικού (software agents) θα έπρεπε να αναλαμβάνουν δράσεις σε ένα περιβάλλον ώστε να μεγιστοποιούν κάποια έννοια αθροιστικής ανταμοιβής. Εδώ, το στοιχείο TN καταγράφει αυτόματα το περιβάλλον του με τη μέθοδο hit & trial, αναλαμβάνει δράση, μαθαίνει από τις εμπειρίες του και βελτιώνει την απόδοση. Το στοιχείο ανταμείβεται για κάθε καλή ενέργεια και τιμωρείται για κάθε λανθασμένη κίνηση. Έτσι, το συστατικό ενισχυτικής μάθησης στοχεύει στη μεγιστοποίηση των ανταμοιβών εκτελώντας καλές ενέργειες. Λόγω της γενικότητάς του, ο τομέας μελετάται σε πολλούς άλλους κλάδους, όπως η θεωρία παιγνίων, η θεωρία ελέγχου, η επιχειρησιακή έρευνα, η θεωρία πληροφοριών, η βελτιστοποίηση βάσει προσομοίωσης, τα συστήματα πολλαπλών πρακτόρων, η νοημοσύνη σμήνους, η στατιστική και οι γενετικοί αλγόριθμοι.

Η ενισχυτική μάθηση χωρίζεται περαιτέρω σε δύο τύπους μεθόδων ή αλγορίθμων:

Θετική ενισχυτική μάθηση: Αυτό αναφέρεται στην προσθήκη ενός ενισχυτικού ερεθίσματος μετά από μια συγκεκριμένη συμπεριφορά του πράκτορα, η οποία καθιστά πιο πιθανό ότι η συμπεριφορά μπορεί να εμφανιστεί ξανά στο μέλλον, π.χ. προσθήκη μιας ανταμοιβής μετά από μια θεμιτή συμπεριφορά.

Μάθηση αρνητικής ενίσχυσης: Η μάθηση αρνητικής ενίσχυσης αναφέρεται στην ενίσχυση μιας συγκεκριμένης συμπεριφοράς που αποφεύγει ένα αρνητικό αποτέλεσμα, π.χ. η επικύρωση μετά από μια αθέμιτη συμπεριφορά.

2.3 Επιπλέον προσεγγίσεις Μηχανικής Μάθησης

2.3.1 Μάθηση με Ημιεπίβλεψη

Η μάθηση με Ημιεπίβλεψη (Semi-supervised learning) βρίσκεται μεταξύ της μάθησης χωρίς επίβλεψη (χωρίς καθόλου επισημειωμένα δεδομένα εκπαίδευσης) και της μάθησης με επίβλεψη (με πλήρως επισημειωμένα δεδομένα εκπαίδευσης). Ορισμένα από τα παραδείγματα εκπαίδευσης δεν έχουν ετικέτες εκπαίδευσης, ωστόσο πολλοί ερευνητές της μηχανικής μάθησης έχουν διαπιστώσει ότι τα μη επισημασμένα δεδομένα, όταν χρησιμοποιούνται σε συνδυασμό με ένα μικρό αριθμό επισημασμένων δεδομένων, μπορούν να επιφέρουν σημαντική βελτίωση στην ακρίβεια της μάθησης.

Στην μάθηση με ημιεπίβλεψη, οι ετικέτες εκπαίδευσης είναι θορυβώδεις, περιορισμένες ή ανακριβείς. Αυτό σημαίνει ότι ορισμένες ετικέτες μπορεί να είναι λανθασμένες, να μην είναι απολύτως ακριβείς ή ακόμη και να λείπουν. Αυτή η αβεβαιότητα μπορεί να προκαλέσει δυσκολίες κατά την εκπαίδευση μοντέλων Μηχανικής Μάθησης, καθώς το μοντέλο προσπαθεί να αντιστοιχίσει τα δεδομένα εισόδου στις ανακριβείς ετικέτες. Ωστόσο, οι ετικέτες αυτές είναι συχνά φθηνότερες στην απόκτησή τους, με αποτέλεσμα μεγαλύτερα αποτελεσματικά σύνολα εκπαίδευσης.

2.3.2. Βαθιά Μάθηση

Η Βαθιά Μάθηση (Deep Learning) είναι μια τεχνική μηχανικής μάθησης που διδάσκει στους υπολογιστές να κάνουν αυτό που είναι φυσικό για τους ανθρώπους: να μαθαίνουν από το παράδειγμα. Η βαθιά μάθηση είναι μια βασική τεχνολογία πίσω από τα αυτοκίνητα χωρίς οδηγό, που τους επιτρέπει να αναγνωρίζουν ένα σήμα στοπ ή να διακρίνουν έναν πεζό από μια κολόνα. Είναι το κλειδί για τον φωνητικό έλεγχο σε καταναλωτικές συσκευές, όπως τηλέφωνα, tablet, τηλεοράσεις και ηχεία hands-free. Η Βαθιά Μάθηση λαμβάνει μεγάλη προσοχή τελευταία και για καλό λόγο. Επιτυγχάνει αποτελέσματα που δεν ήταν εφικτά πριν.

Στη βαθιά μάθηση, ένα μοντέλο υπολογιστή μαθαίνει να εκτελεί εργασίες ταξινόμησης απευθείας από εικόνες, κείμενο ή ήχο. Τα μοντέλα βαθιάς μάθησης μπορούν να επιτύχουν κορυφαία ακρίβεια, που μερικές φορές ξεπερνά τις επιδόσεις σε ανθρώπινο επίπεδο. Τα μοντέλα εκπαιδεύονται χρησιμοποιώντας ένα μεγάλο σύνολο επισημασμένων δεδομένων και αρχιτεκτονικές νευρωνικών δικτύων που περιέχουν πολλά επίπεδα.

2.4 Επεξεργασία Φυσικής Γλώσσας

Η Επεξεργασία Φυσικής Γλώσσας (Natural Language processing ή NLP) [5] είναι ένα υποπεδίο της γλωσσολογίας, της επιστήμης των υπολογιστών και της Τεχνητής Νοημοσύνης που ασχολείται με τις αλληλεπιδράσεις μεταξύ των υπολογιστών και της ανθρώπινης γλώσσας, ιδίως με τον τρόπο προγραμματισμού των υπολογιστών για την επεξεργασία και την ανάλυση μεγάλων ποσοτήτων δεδομένων φυσικής γλώσσας. Στόχος είναι ένας υπολογιστής ικανός να "κατανοεί" το περιεχόμενο των εγγράφων, συμπεριλαμβανομένων των συμφραζόμενων αποχρώσεων της γλώσσας που περιέχεται σε αυτά. Αυτή η τεχνολογία μπορεί στη συνέχεια να εξάγει με ακρίβεια πληροφορίες και ιδέες που περιέχονται στα έγγραφα, καθώς και να κατηγοριοποιεί και να οργανώνει τα ίδια τα έγγραφα.

Οι προκλήσεις στην Επεξεργασία Φυσικής Γλώσσας περιλαμβάνουν συχνά την αναγνώριση ομιλίας, που αφορά τη μετατροπή του προφορικού λόγου σε γραπτή μορφή. Επίσης, υπάρχει η πρόκληση της κατανόησης φυσικής γλώσσας, που απαιτεί την αντίληψη της σημασίας και της δομής των γλωσσικών εκφράσεων. Τέλος, η δημιουργία φυσικής γλώσσας, η οποία αναφέρεται στη δημιουργία νέων γλωσσικών εκφράσεων, όπως αυτόματη παραγωγή κειμένου.

Με τη συνεχή ανάπτυξη και την εξέλιξη της Επεξεργασίας Φυσικής Γλώσσας, ανοίγονται νέες προοπτικές για την αποτελεσματική επεξεργασία και ανάλυση μεγάλων ποσοτήτων γλωσσικών δεδομένων, προσφέροντας μία βαθύτερη και πιο συνεπή κατανόηση της ανθρώπινης γλώσσας.

2.5 Εφαρμογές Επεξεργασίας Φυσικής Γλώσσας

Η Επεξεργασία Φυσικής Γλώσσας μπορεί να χρησιμοποιηθεί σε μια ποικιλία εφαρμογών, συμπεριλαμβανομένης της αυτόματης μετάφρασης, της ανίχνευσης ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου, της εξαγωγής πληροφοριών, της σύνοψης και της απάντησης ερωτήσεων. Οι εφαρμογές αυτές θα αναλυθούν περαιτέρω στις παρακάτω υποενότητες.

2.5.1 Μηχανική Μετάφραση

Το να γίνουν τα δεδομένα προσβάσιμα και διαθέσιμα σε όλους είναι ένα δύσκολο εγχείρημα, παρότι το μεγαλύτερο μέρος του πλανήτη είναι πλέον συνδεδεμένο στο διαδίκτυο. Ο γλωσσικός φραγμός αποτελεί σημαντικό εμπόδιο στην προσβασιμότητα των δεδομένων. Υπάρχουν πολλές διαφορετικές γλώσσες και κάθε μία έχει μοναδική δομή και σύνταξη προτάσεων. Με την Μηχανική Μετάφραση, οι φράσεις μεταφράζονται με τη χρήση μιας στατιστικής μηχανής όπως το Google Translate από τη μία γλώσσα στην άλλη. Το ζήτημα ωστόσο με την τεχνολογία αυτή δεν είναι η άμεση μετάφραση των λέξεων, αλλά η διατήρηση του νοήματος των προτάσεων, της γραμματικής και των χρόνων. Η στατιστική

μηχανική μάθηση συλλέγει όσο το δυνατόν περισσότερα δεδομένα, αναζητά παραλληλισμούς μεταξύ των δύο γλωσσών και στη συνέχεια αναλύει τα δεδομένα για να καθορίσει πόσο πιθανό είναι να αντιστοιχεί κάτι στη γλώσσα Α σε κάτι στη γλώσσα Β.

2.5.2 Κατηγοριοποίηση κειμένου

Ένα σύστημα κατηγοριοποίησης συλλέγει μεγάλες ροές δεδομένων, όπως επίσημα έγγραφα, αναφορές οδικών ατυχημάτων, δεδομένα της αγοράς και ειδησεογραφικά δελτία, και τα κατατάσσει σε προκαθορισμένες κατηγορίες ή ευρετήρια. Για παράδειγμα, το System44 [6] του Όμιλου Carnegie εξοικονομεί πολύ περισσότερο χρόνο εισάγοντας άρθρα του Reuters και κάνοντας τη δουλειά που συμβατικά κάνει εξειδικευμένο προσωπικό. Ορισμένες εταιρείες χρησιμοποιούν συστήματα ταξινόμησης για να ταξινομούν τα δελτία προβλημάτων ή τα αιτήματα παραπόνων ώστε να τα προωθούν στο κατάλληλο γραφείο. Μια άλλη εφαρμογή της ταξινόμησης κειμένου είναι τα φίλτρα ανεπιθύμητης αλληλογραφίας ηλεκτρονικού ταχυδρομείου. Τα φίλτρα ανεπιθύμητης αλληλογραφίας αποκτούν ολοένα και μεγαλύτερη σημασία ως πρώτη γραμμή άμυνας κατά των ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου. Το πρόβλημα των λανθασμένα αρνητικών και των λανθασμένα θετικών αποτελεσμάτων στα φίλτρα spam βρίσκεται στο επίκεντρο της Επεξεργασίας Φυσικής Γλώσσας, η οποία αντιμετωπίζει την πρόκληση εξαγωγής νοήματος από σειρές κειμένου. Οι λύσεις φιλτραρίσματος που εφαρμόζονται σε συστήματα ηλεκτρονικού ταχυδρομείου χρησιμοποιούν ένα σύνολο πρωτοκόλλων για να καθορίσουν ποια εισερχόμενα μηνύματα είναι ανεπιθύμητα και ποια όχι.

2.5.3 Εξαγωγή και Σύνοψη Πληροφοριών

Η εξαγωγή πληροφοριών εντοπίζει φράσεις ενδιαφέροντος από δεδομένα κειμένου. Σε πολλές εφαρμογές, η εξαγωγή οντοτήτων όπως ονόματα, τόποι, γεγονότα, ημερομηνίες, ώρες και τιμές είναι ένας ισχυρός τρόπος για τη σύνοψη πληροφοριών σχετικών με τις ανάγκες του χρήστη. Για τις μηχανές αναζήτησης, ο αυτόματος εντοπισμός σημαντικών πληροφοριών μπορεί να βελτιώσει την ακρίβεια και την αποτελεσματικότητα των στοχευμένων αναζητήσεων. Τα κρυφά μοντέλα Markov (Hidden Markov Models – HMM[1]) χρησιμοποιούνται για την εξαγωγή σχετικών περιοχών σε αρκετές ερευνητικές εργασίες. Αυτά τα εξαγόμενα τμήματα κειμένου χρησιμοποιούνται για να επιτρέπουν αναζητήσεις με βάση συγκεκριμένα πεδία, να εμφανίζουν αποδοτικά τα αποτελέσματα της αναζήτησης και να αντιστοιχίζουν αναφορές σε άρθρα.

* είναι ένα στατιστικό μοντέλο που μπορεί να χρησιμοποιηθεί για να περιγράψει την εξέλιξη παρατηρήσιμων γεγονότων που εξαρτώνται από εσωτερικούς παράγοντες, οι οποίοι δεν είναι άμεσα παρατηρήσιμοι. Ονομάζουμε το παρατηρούμενο γεγονός "σύμβολο" και τον αόρατο παράγοντα που διέπει την παρατήρηση "κατάσταση".

2.5.4 Συστήματα Διαλόγου

Σχετικά πρόσφατη τεχνολογία και μία από τις πιο επιθυμητές για την αυτοματοποίηση της ζωής του μέλλοντος. Τα συστήματα αυτά επιτρέπουν είτε γραπτή είτε φωνητική εντολή από τον χρήστη για την εκτέλεση κάποιας εργασίας ή απλής συνομιλίας όταν αυτά χρησιμοποιούν όλα τα επίπεδα της γλωσσικής επεξεργασίας. Οι λειτουργίες τους μπορεί να διαφέρουν ανάλογα το σύστημα, σε συγκεκριμένες, όπως οι ψηφιακοί βοηθοί Cortana των Windows και Siri της Apple, που απλά παρέχουν τις πληροφορίες που ζητάει ο χρήστης ψάχνοντας στο Διαδίκτυο. Αλλά και σε μία τεράστια πληθώρα δυνατοτήτων, όπως το Chat GPT της OpenAI, που όχι μόνο μπορεί να συνομιλήσει με τον χρήστη όπως ένας άνθρωπος, αλλά παράγει πρωτότυπες ιστορίες, ποιήματα και τραγούδια. Η πιο εντυπωσιακή ωστόσο λειτουργία του είναι ότι μπορεί να συγγράψει αποσπάσματα κώδικα, με αρκετή ακρίβεια, αρκεί ο χρήστης να τεκμηριώσει με σαφήνεια και ακρίβεια την εντολή που θα δώσει στο μοντέλο.

2.6 Ανακάλυψη Γνώσης και Εξόρυξη Δεδομένων

Οι όροι ανακάλυψη γνώσης σε βάσεις δεδομένων (KDD – Knowledge Discovery in Databases) και εξόρυξη δεδομένων (Data Mining) χρησιμοποιούνται συχνά συναλλακτικά. Στην πραγματικότητα, έχουν δοθεί πολλά άλλα ονόματα σε αυτή τη διαδικασία της ανακάλυψης χρήσιμων (κρυφών) προτύπων στα δεδομένα: εξόρυξη γνώσης, ανακάλυψη πληροφοριών, διερευνητική ανάλυση δεδομένων, συγκομιδή πληροφοριών και αναγνώριση προτύπων χωρίς επίβλεψη. Τα τελευταία χρόνια το KDD χρησιμοποιείται για να αναφερθεί σε μια διαδικασία που αποτελείται από πολλά βήματα, ενώ η εξόρυξη δεδομένων είναι μόνο ένα από αυτά τα βήματα. Ακολουθούν οι δύο ορισμοί με βάση την Margaret Dunham [7]:
Ορισμός 1: Η ανακάλυψη γνώσης σε βάσεις δεδομένων (KDD) είναι η διαδικασία εύρεσης χρήσιμων πληροφοριών και προτύπων στα δεδομένα.
Ορισμός 2: Η εξόρυξη δεδομένων είναι η χρήση αλγορίθμων για την εξαγωγή των πληροφοριών και μοτίβα που προκύπτουν από τη διαδικασία KDD.
Η KDD είναι μια διαδικασία που περιλαμβάνει πολλά διαφορετικά βήματα. Η είσοδος σε αυτή τη διαδικασία είναι τα δεδομένα και η έξοδος είναι οι χρήσιμες πληροφορίες που επιθυμούν οι χρήστες. Ωστόσο, ο στόχος μπορεί να είναι ασαφής ή ανακριβής. Η διεργασία είναι διαδραστική και μπορεί να απαιτεί πολύ χρόνο. Η διαδικασία KDD αποτελείται από τα ακόλουθα πέντε βήματα:

2.6.1 Επιλογή Δεδομένων

Τα δεδομένα που απαιτούνται για τη διαδικασία εξόρυξης δεδομένων μπορούν να ληφθούν από πολλές διαφορετικές και ετερογενείς πηγές δεδομένων. Αυτό το πρώτο βήμα αποκτά τα δεδομένα από διάφορες βάσεις δεδομένων, αρχεία και μη ηλεκτρονικές πηγές.

2.6.2 Προεπεξεργασία δεδομένων

Ένα πολύ συχνό φαινόμενο που παρατηρείται είναι η λανθασμένη μορφή των δεδομένων εισόδου στο σύστημα. Τα δεδομένα που πρόκειται να χρησιμοποιηθούν από τη διαδικασία μπορεί να είναι εσφαλμένα ή ελλιπή [8]. Μπορεί να υπάρχουν ανώμαλα δεδομένα (λέξεις ή μορφολογικοί τύποι που δεν ακολουθούν τα συνηθισμένα μοτίβα ή κανόνες συγκεκριμένης γλώσσας) από πολλαπλές πηγές που αφορούν διαφορετικούς τύπους δεδομένων και μετρικές. Πολλές διαφορετικές δραστηριότητες ενδέχεται να εκτελούνται σε αυτή τη στιγμή.

2.6.2.1Αφαίρεση σημείων στίξης

Από τις πιο βασικές διαδικασίες στον καθαρισμό λέξεων ενός κειμένου είναι η εύρεση και αφαίρεση των σημείων στίξης (removal of punctuation) [9]. Ωστόσο, καθώς μια τέτοια ενέργεια μπορεί να αλλοιώσει σημαντικά το νόημα του κειμένου πρέπει να εφαρμόζεται με προσοχή.

Παράδειγμα 2.6.2.1

Στο παρακάτω παράδειγμα, για την αφαίρεση των σημείων στίξης χρησιμοποιείται η μέθοδος «.maketrans» οπου δέχεται τρία ορίσματα, τα δύο πρώτα από τα οποία είναι κενές συμβολοσειρές και το τρίτο είναι η λίστα με τα σημεία στίξης που θέλουμε να αφαιρέσουμε. Αυτό λέει στη συνάρτηση να αντικαταστήσει όλα τα σημεία στίξης με «None».

```
a_string = '!hi. wh?at is the weat[h]er lik?e.'
new_string = a_string.translate(str.maketrans('', '',
string.punctuation))

print(new_string)

# Returns: hi what is the weather like
```

Εικόνα 2.1 [9]

Η μέθοδος maketrans() επιστρέφει έναν πίνακα αντιστοίχισης που μπορεί να χρησιμοποιηθεί με τη μέθοδο translate() για την αντικατάσταση ειδικών χαρακτήρων.

2.6.2.2 Κατάτμηση Λέξεων

Η Κατάτμηση Λέξεων (Word Tokenization) είναι το σπάσιμο του ακατέργαστου κειμένου σε μικρά κομμάτια. Η κατάτμηση σπάει το ακατέργαστο κείμενο σε λέξεις ή προτάσεις που ονομάζονται tokens. Αυτά τα tokens βοηθούν στην κατανόηση του πλαισίου ή στην ανάπτυξη του μοντέλου NLP. Επιπλέον βοηθά στην ερμηνεία του νοήματος του κειμένου αναλύοντας την αλληλουχία των λέξεων.

Παράδειγμα 2.6.2.2

Είσοδος: “Today was a sunny day”

Έξοδος: {‘Today’, ‘was’, ‘a’, ‘sunny’, ‘day’}

2.6.2.3 Αφαίρεση διακοπτούσων λέξεων

Διακοπτούσες (stopwords) λέγονται οι λέξεις της φυσικής γλώσσας που έχουν πολύ μικρή σημασία, όπως οι Αγγλικές "is", "an", "the", “has”, “and”, κ.λπ. [10]. Οι μηχανές αναζήτησης και άλλες πλατφόρμες ευρετηρίασης επιχειρήσεων συχνά φιλτράρουν αυτές τις λέξεις κατά την άντληση αποτελεσμάτων από τη βάση δεδομένων σε σχέση με τα ερωτήματα του χρήστη.

Οι διακοπτούσες λέξεις συχνά αφαιρούνται από το κείμενο πριν από την εκπαίδευση μοντέλων βαθιάς μάθησης και μηχανικής μάθησης, δεδομένου ότι εμφανίζονται σε αφθονία, συνεπώς παρέχουν ελάχιστες έως καθόλου μοναδικές πληροφορίες που μπορούν να χρησιμοποιηθούν για ταξινόμηση ή ομαδοποίηση.
Παράδειγμα 2.6.2.3
Είσοδος: “We’re going to remove the stopwords from this example”
Έξοδος: [‘going’, ‘remove’, ‘stopwords’, ‘example’]

2.6.2.4 Μετατροπή πεζών γραμμάτων

Ιδιαίτερα χρήσιμη διαδικασία στην προεπεξεργασία δεδομένων αποτελεί η μετατροπή των κεφαλαίων γραμμάτων σε πεζά (Lowercase conversion). Η τεχνική αυτή χρησιμοποιείται για την εξόρυξη γνώσης από πολλές μηχανές αναζήτησης και διευκολύνει την ταύτιση πανομοιότυπων λέξεων όπως π.χ. “Inflation” και “inflation”.

2.6.2.5 Στελεχοποίηση

Η Στελεχοποίηση (Stemming) είναι η διαδικασία απόκτησης του θέματος μιας λέξης. Θέμα είναι το μέρος στο οποίο προστίθενται τα κλιτικά επιθήματα (-ed, -ize, -de, -s, κ.λπ.). Το θέμα μιας λέξης δημιουργείται με την αφαίρεση του προθέματος ή του επιθέματος μιας λέξης. Έτσι, κάποιες φορές μπορεί να μην οδηγήσει σε πραγματικές λέξεις.

Παράδειγμα 2.6.2.5

Banks → Bank

Looked → Look

Looks → Look

2.6.2.6 Λημματοποίηση

Η διαδικασία της λημματοποίησης (lemmatization) λαμβάνει υπόψη διάφορους γλωσσικούς παράγοντες, συμπεριλαμβανομένων των συμφραζομένων των λέξεων, των ετικετών μέρους του λόγου (POS) και στους μορφολογικής ανάλυσης, για τον ακριβή προσδιορισμό της βασικής μορφής κάθε λέξης. Με την αναγωγή των λέξεων στα λήμματά τους, η λημματοποίηση αποσκοπεί στον χειρισμό των κλιτικών και παραγωγικών παραλλαγών, με αποτέλεσμα μια πιο συνοπτική και συνεπή αναπαράσταση των δεδομένων του κειμένου.

Όπως η στελεχοποίηση, έτσι και η λημματοποίηση μετατρέπει μια λέξη στη μορφή της ρίζας της. Η μόνη διαφορά είναι ότι η λημματοποίηση διασφαλίζει ότι η λέξη-ρίζα ανήκει στη γλώσσα, έτσι θα λάβουμε πάντα έγκυρες λέξεις. Σε εργασίες ταξινόμησης κειμένου, η λημματοποίηση βοηθά στη μείωση της διάστασης του λεξιλογίου, οποία με τη σειρά της βελτιώνει την απόδοση των αλγορίθμων μηχανικής μάθησης. Βοηθά επίσης στην ανάλυση συναισθήματος παρέχοντας μια συνεπή αναπαράσταση των λέξεων, επιτρέποντας την καλύτερη ανάλυση του συναισθηματικού τόνου και του συναισθήματος στο κείμενο.

2.6.3 Μετασχηματισμός Δεδομένων

Τα δεδομένα από διαφορετικές πηγές πρέπει να μετατραπούν σε μία κοινή μορφή για επεξεργασία. Ορισμένα δεδομένα μπορεί να κωδικοποιηθούν ή να μετασχηματιστούν σε πιο εύχρηστες μορφές. Η μείωση των δεδομένων μπορεί να χρησιμοποιηθεί για τη μείωση του αριθμού των πιθανών τιμών που εξετάζονται. Οι πιο συνήθεις τακτικές επιλογής χαρακτηριστικών των δεδομένων είναι η BOW (Bag of Words) ή Count Vectorizer και η στάθμιση όρων TF-IDF.

2.6.3.1. Bag Of Words

Η BOW [11] είναι μια αναπαράσταση κειμένου που περιγράφει την εμφάνιση λέξεων σε ένα έγγραφο. Απλώς παρακολουθούμε τον αριθμό των λέξεων και αγνοούμε τις γραμματικές λεπτομέρειες και τη σειρά των λέξεων. Ονομάζεται "σακούλα" λέξεων επειδή απορρίπτεται κάθε πληροφορία σχετικά με τη σειρά ή τη δομή των λέξεων στο έγγραφο. Το μοντέλο ασχολείται μόνο με το αν εμφανίζονται γνωστές λέξεις στο έγγραφο, όχι με το πού στο έγγραφο.

Παράδειγμα 2.6.3.1

Έστω ότι έχουμε δύο απλά αρχεία κειμένου:

(1) John likes to watch movies. Mary likes movies too.

(2) Mary also likes to watch football games.

Βήμα 1ο : δημιουργούνται οι εξής λίστες:

BoW1 = {"John","likes","to","watch","movies","Mary","likes","movies","too"}

BoW2 = {"Mary","also","likes","to","watch","football","games"}

Βήμα 2ο : Εντοπίζονται τα tf, δηλαδή η συχνότητα των όρων

BoW1 = {"John":1,"likes":2,"to":1,"watch":1,"movies":2,"Mary":1,"too":1};

BoW2 ={"Mary":1,"also":1,"likes":1,"to":1,"watch":1,"football":1,"games":1};

Βήμα 3ο : Μετασχηματίζουμε τις λίστες σε μαθηματικά διανύσματα με βάση την συχνότητα

BoW1 → [1,2,1,1,2,1,1]

BoW2 → [1,1,1,1,1,1,1]

2.6.3.2 TF-IDF

Στην ανάκτηση πληροφοριών, το TF-IDF [12] είναι ένα στατιστικό στοιχείο που αντικατοπτρίζει πόσο σημαντική είναι μια λέξη για ένα έγγραφο ή κείμενο. Η τιμή TF-IDF αυξάνεται αναλογικά με τον αριθμό των φορών που εμφανίζεται μια λέξη στο έγγραφο και αντισταθμίζεται από τον αριθμό των εγγράφων στο σώμα που περιέχουν τη λέξη, γεγονός που βοηθά στην προσαρμογή του γεγονότος ότι ορισμένες λέξεις εμφανίζονται γενικά πιο συχνά. Το TF-IDF είναι ένα από τα πιο δημοφιλή σχήματα στάθμισης όρων σήμερα.

TF (Term Frequency – Συχνότητα Όρων)

Το Term Frequency είναι ένα στοιχείο μέτρησης της συχνότητας εμφάνισης ενός όρου σε ένα έγγραφο. Υπολογίζεται διαιρώντας τον αριθμό των φορών που εμφανίζεται ο όρος στο έγγραφο με τον συνολικό αριθμό των όρων στο έγγραφο, το $tf(t,d)$, είναι η σχετική συχνότητα του όρου t στο έγγραφο d.

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f', d}$$

Όπου $f_{t,d}$ είναι η ακατέργαστη καταμέτρηση ενός όρου σε ένα έγγραφο, δηλαδή ο αριθμός των φορών που εμφανίζεται ο όρος t στο έγγραφο d. Να σημειωθεί πώς ο παρονομαστής είναι απλώς ο συνολικός αριθμός των όρων στο έγγραφο d (μετρώντας κάθε εμφάνιση του ίδιου όρου ξεχωριστά).

Αντίστροφη Συχνότητα Εγγράφων

Η αντίστροφη συχνότητα εγγράφων (Inverse Document Frequency – IDF) είναι ένα μέτρο αξιολόγησης του πόσο σημαντικός είναι ένας όρος σε μια συλλογή εγγράφων. Υπολογίζεται διαιρώντας τον συνολικό αριθμό των εγγράφων με τον αριθμό των εγγράφων που περιέχουν τον όρο και στη συνέχεια παίρνοντας τον λογάριθμο αυτής της τιμής.

$$idf(t,D) = log\frac{N}{|\{d \in D: t \in d\}|}$$

Όπου

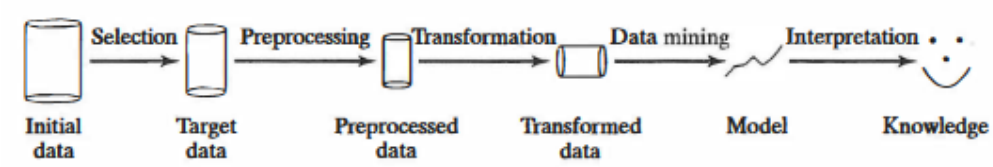
- N : ο συνολικός αριθμός των εγγράφων στα δεδομένα με $N = |D|$
- $|\{d \in D: t \in d\}|$: αριθμός εγγράφων στα οποία εμφανίζεται ο όρος t (δηλαδή, $tf(t,d) \neq 0$). Εάν ο όρος δεν υπάρχει στα δεδομένα, αυτό θα οδηγήσει σε διαίρεση με το μηδέν. Ως εκ τούτου, συνηθίζεται να προσαρμόζεται ο παρονομαστής σε $1 + |\{d \in D: t \in d\}|$

2.6.4 Εξόρυξη δεδομένων

Με βάση την εργασία εξόρυξης δεδομένων που εκτελείται, αυτό το βήμα εφαρμόζει αλγόριθμους στα μετασχηματισμένα δεδομένα για τη δημιουργία των επιθυμητών αποτελεσμάτων.

5. Ερμηνεία/αξιολόγηση

Ο τρόπος παρουσίασης των αποτελεσμάτων της εξόρυξης δεδομένων στους χρήστες είναι εξαιρετικά σημαντικός, διότι από αυτόν εξαρτάται η χρησιμότητα των αποτελεσμάτων. Σε αυτό το τελευταίο βήμα χρησιμοποιούνται διάφορες στρατηγικές οπτικοποίησης και GUI

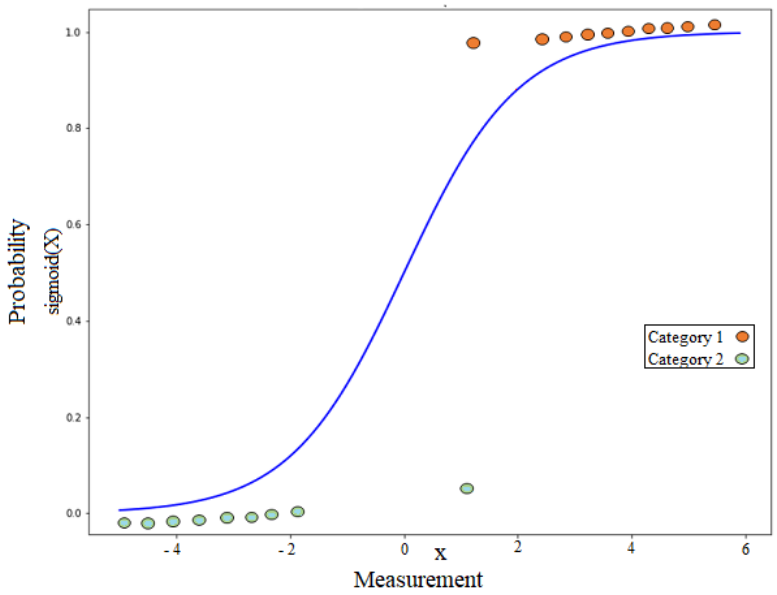


Εικόνα 2.2 Διαδικασία Εξόρυξης Γνώσης [7]

Κεφάλαιο 3
Αλγόριθμοι Μηχανικής Μάθησης

3.1 Λογιστική Παλινδρόμηση

Η λογιστική παλινδρόμηση (Logistic Regression) [13] [14]είναι μια εξαιρετικά ευέλικτη μέθοδος πρόβλεψης διχοτομικής ταξινόμησης, δηλαδή χρησιμοποιείται για την πρόβλεψη ενός δυαδικού αποτελέσματος ή κατάστασης, όπως ναι/όχι, επιτυχία/αποτυχία και αλήθεια/ψέμα. Επιπλέον προσπαθεί να ταιριάξει τα δεδομένα σε λογικές καμπύλες, με την πιο δημοφιλή να είναι η σιγμοειδής, όπως φαίνεται στην παρακάτω εικόνα 3.1. Είναι μία μονότονα αύξουσα συνάρτηση καθώς οι τιμές εξόδου της κυμαίνονται στο διάστημα [0,1], ώστε να μπορεί να εξαχθεί η επικρατέστερη πιθανότητα προς κάποια κατηγορία.



Εικόνα 3.1 Λογική Καμπύλη [13]

Όπως αναφέρθηκε προηγουμένως, η λογιστική παλινδρόμηση επιλύει προβλήματα δυαδικής φύσεως και η λογιστική συνάρτησή της είναι της μορφής:

$$p(x) = \frac{1}{1 + e^{-(x-\mu)/s}}$$

Μέσω της στρατηγικής «One vs Rest» μπορεί να προβλέψει παραπάνω από δύο κλάσεις, αν αυτό είναι επιθυμητό. Η πιθανότητα p να ανήκει το δεδομένο στην κατηγορία δίνεται από τον τύπο:

$$p = \frac{e^{(c_0 + c_1 x_1)}}{1 + e^{(c_0 + c_1 x_1)}}$$

Για προβλήματα παλινδρόμησης εφαρμόζουμε λογαριθμική συνάρτηση στην παραπάνω σχέση και προκύπτει:

$$log_e\left(\frac{p}{1-p}\right) = c_0 + c_1 x_1$$

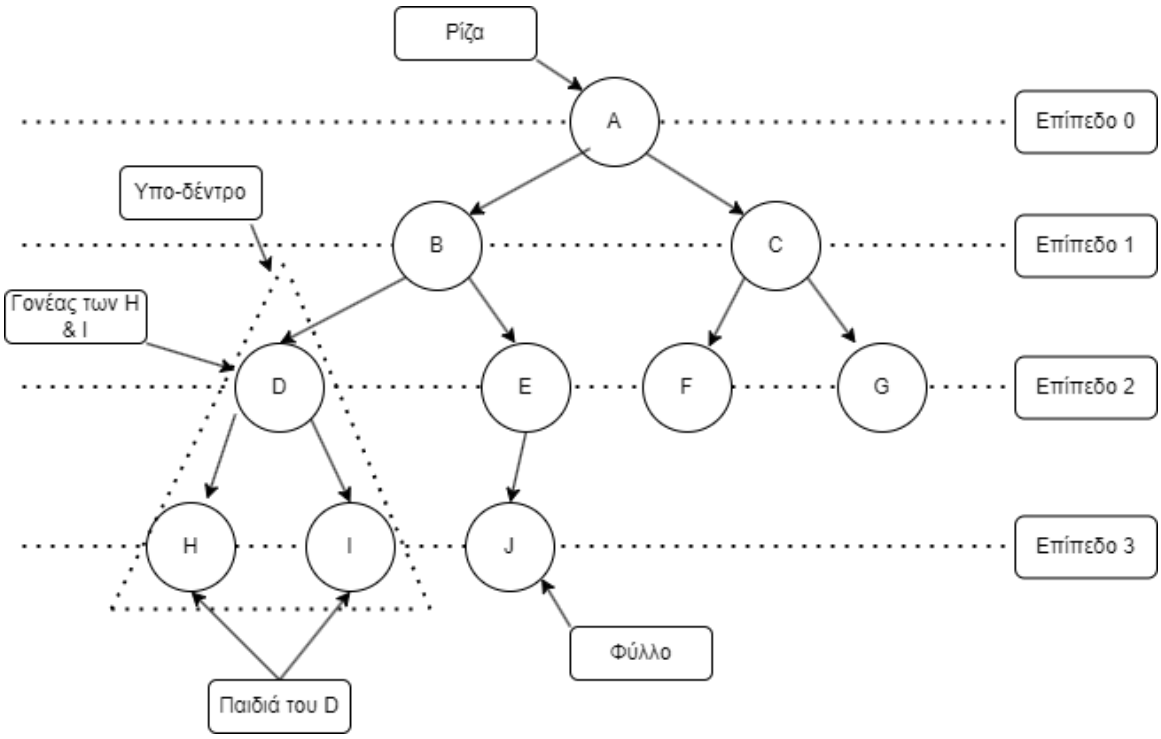
Η λογιστική παλινδρόμηση είναι μία ευρέως διαδεδομένη τεχνική κατηγοριοποίησης δεδομένων διότι δεν απαιτεί πολλούς υπολογιστικούς πόρους, είναι ιδιαίτερα προσαρμόσιμη και αποτελεσματική. Επίσης, εξάγει ακόμα καλύτερες προβλέψεις εάν αγνοηθούν χαρακτηριστικά που δεν σχετίζονται με την μεταβλητή εξόδου, καθώς και όσα σχετίζονται μεταξύ τους.

Το πρόβλημα με την λογιστική παλινδρόμηση είναι ότι δεν γίνεται να επιλύσουμε μη γραμμικά προβλήματα και δεν έχει καλή απόδοση με ανεξάρτητες μεταβλητές που δεν σχετίζονται με τη μεταβλητή εξόδου. Εν κατακλείδι, είναι ένας πολύ ισχυρός αλγόριθμος κατηγοριοποίησης, αλλά είναι αποδοτικός μόνο όταν βασίζεται σε καλά διακριβωμένες προβλεπόμενες πιθανότητες.

3.2 Δέντρα Αποφάσεων

Ο ταξινομητής δέντρων απόφασης (Decision Trees) [15] είναι ένας αλγόριθμος μάθησης με επίβλεψη που μπορεί να χρησιμοποιηθεί τόσο για εργασίες ταξινόμησης όσο και για εργασίες παλινδρόμησης. Τα δέντρα έχουν μία μη-γραμμική δομή δεδομένων και οργανώνονται σε ιεραρχική μορφή. Το δέντρο κατασκευάζεται χρησιμοποιώντας μια αναδρομική μέθοδο κατάτμησης, όπου τα δεδομένα χωρίζονται σε υποσύνολα με βάση τις τιμές των χαρακτηριστικών εισόδου. Κάθε εσωτερικός κόμβος του δέντρου αντιπροσωπεύει μια δοκιμή σε ένα χαρακτηριστικό εισόδου και κάθε κόμβος φύλλου αντιπροσωπεύει μια ετικέτα κλάσης ή μια προβλεπόμενη τιμή.

Εμβαθύνοντας, ο πρώτος κόμβος του δέντρου καλείται ρίζα (root node), υπάρχει σε κάθε δέντρο και είναι μοναδικός. Ο κόμβος που είναι προκάτοχος οποιουδήποτε άλλου ονομάζεται κόμβος πατέρας (parent node), ενώ ο διάδοχός του ονομάζεται κόμβος παιδί (child node). Η ρίζα είναι ο μόνος κόμβος που δεν έχει πατέρα. Επιπλέον οι κόμβοι χωρίς διαδόχους ονομάζονται φύλλα (leaf nodes). Τέλος, τα δέντρα με δύο ή παραπάνω επίπεδα συχνά σχηματίζουν άλλα δέντρα που λέγονται υποδέντρα (subtrees), τα οποία έχουν την ίδια μορφή και ιδιότητες. Ένα τυπικό παράδειγμα δομής ενός δέντρου απεικονίζεται στο παρακάτω σχήμα 3.2.



Εικόνα 3.2 Δομή Δέντρου Απόφασης

Πιο συγκεκριμένα το δέντρο κατασκευάζεται ξεκινώντας από τον κόμβο-ρίζα, ο οποίος αντιπροσωπεύει ολόκληρο το σύνολο των δεδομένων. Σε κάθε εσωτερικό κόμβο, ο αλγόριθμος επιλέγει το χαρακτηριστικό που χωρίζει καλύτερα τα δεδομένα σε υποσύνολα με την υψηλότερη καθαρότητα, με μια μετρική όπως το κέρδος πληροφορίας (information Gain) ή το Gini (βλ. παρακάτω). Μόλις επιλεχθεί ένα χαρακτηριστικό, τα δεδομένα μεταφέρονται στους αντίστοιχους κόμβους-παιδιά, οι οποίοι επαναλαμβάνουν τη διαδικασία επιλογής χαρακτηριστικών και διαχωρισμού των δεδομένων έως ότου ικανοποιηθεί ένα κριτήριο διακοπής. Το κριτήριο αυτό μπορεί να είναι ένα μέγιστο βάθος του δέντρου, ένας ελάχιστος αριθμός δειγμάτων σε έναν κόμβο φύλλου ή μια χαμηλή βελτίωση της καθαρότητας των δεδομένων. Κάθε κόμβος φύλλου αντιπροσωπεύει μια ετικέτα κλάσης ή μια προβλεπόμενη τιμή.

Όταν μια νέα είσοδος παρουσιάζεται στο δέντρο αποφάσεων, ακολουθεί τη διαδρομή από τον κόμβο ρίζας σε έναν κόμβο φύλλου δοκιμάζοντας την είσοδο σε σχέση με τις συνθήκες που καθορίζονται σε κάθε εσωτερικό κόμβο. Η ετικέτα κλάσης ή η προβλεπόμενη τιμή που σχετίζεται με τον κόμβο φύλλου είναι η έξοδος του δέντρου αποφάσεων.

Ο αλγόριθμος χρησιμοποιεί μια άπληστη προσέγγιση για την επιλογή του χαρακτηριστικού που χωρίζει τα δεδομένα σε υποσύνολα με τη μεγαλύτερη καθαρότητα σε κάθε εσωτερικό κόμβο, η οποία μπορεί να οδηγήσει σε υπερπροσαρμογή (overfitting) εάν το δέντρο γίνει πολύ βαθύ. Το overfitting συμβαίνει εάν το σύνολο των δεδομένων εκπαίδευσης είναι πολύ μεγάλο, οδηγώντας το δέντρο απόφασης να κατηγοριοποιεί με ιδιαίτερα μεγάλη ακρίβεια τα δεδομένα εκπαίδευσης (training set), αλλά αδυνατεί στα δεδομένα ελέγχου (test set). Για να ξεπεραστεί αυτός ο περιορισμός, έχουν προταθεί διάφορες τεχνικές, όπως το κλάδεμα (pruning), το bagging και το boosting.

Η εντροπία είναι ένα μέσο μέτρησης της τυχαιότητας των πληροφοριών που υποβάλλονται σε επεξεργασία. Όσο υψηλότερη είναι η εντροπία, τόσο πιο δύσκολο είναι να εξαχθούν συμπεράσματα από τις πληροφορίες αυτές. Μαθηματικά η εντροπία παρίσταται ως:

$$E(S) = \sum_{i}^{c} -p_i log_2 p_i$$

Όπου S είναι η τρέχουσα κατάσταση και p_i η πιθανότητα ενός γεγονότος I της κατάστασης S ή το ποσοστό της κλάσης i σε έναν κόμβο της κατάστασης S.

Το κέρδος πληροφορίας (information gain) ή IG είναι μια στατιστική ιδιότητα που μετρά πόσο καλά ένα δοσμένο χαρακτηριστικό διαχωρίζει τα παραδείγματα εκπαίδευσης σύμφωνα με την ταξινόμηση-στόχο τους. Η κατασκευή ενός δέντρου απόφασης έχει να κάνει με την εύρεση ενός χαρακτηριστικού που αποδίδει το μεγαλύτερο κέρδος πληροφορίας και τη μικρότερη εντροπία.

Το κέρδος πληροφορίας είναι μείωση της εντροπίας. Υπολογίζει τη διαφορά μεταξύ της εντροπίας πριν από τη διάσπαση της μέσης εντροπίας και μετά το διαχωρισμό του συνόλου δεδομένων με βάση τις δεδομένες τιμές των χαρακτηριστικών. Μαθηματικά, το IG αναπαρίσταται ως εξής:

$$IG = E(\pi\rho\iota\nu) - \sum_{j=1}^{K} E(j, \mu\varepsilon\tau\acute{\alpha})$$

Όπου «πριν» είναι το σύνολο των δεδομένων πριν το διαχωρισμό, K ο αριθμός των υποσυνόλων που δημιουργήθηκαν αφού έγινε η διάσπαση και (j, μετά) είναι το υποσύνολο j μετά το διαχωρισμό.

Υπάρχουν αρκετές αλγοριθμικές προσεγγίσεις αναφορικά με το μοντέλο ταξινόμησης του Δέντρου αποφάσεων, με τις πιο δημοφιλείς να είναι το ID3, C4.5 και το CART.

Το ID3 (Iterative Dichotomiser 3) χρησιμοποιεί την έννοια του κέρδους πληροφορίας για να καθορίσει το χαρακτηριστικό που χωρίζει καλύτερα τα δεδομένα σε κάθε εσωτερικό κόμβο. Το κέρδος πληροφορίας, IG, είναι ένα μέτρο της μείωσης της ασάφειας των δεδομένων μετά από μια διάσπαση και υπολογίζεται ως η διαφορά μεταξύ της εντροπίας, H, του γονικού κόμβου και του σταθμισμένου μέσου όρου των εντροπιών των παιδιών κόμβων, H(i). Το χαρακτηριστικό με το μεγαλύτερο κέρδος πληροφορίας (IG) επιλέγεται για τη διάσπαση των δεδομένων.

Το C4.5 είναι μια επέκταση του ID3 που χρησιμοποιεί την έννοια του λόγου κέρδους (Gain Ratio) αντί του κέρδους πληροφορίας (IG). Ο λόγος κέρδους είναι μια τροποποίηση του κέρδους πληροφορίας που λαμβάνει υπόψη τον αριθμό των αποτελεσμάτων ενός χαρακτηριστικού. Υπολογίζεται ως το κέρδος πληροφορίας διαιρούμενο με την εγγενή πληροφορία του χαρακτηριστικού, η οποία υπολογίζεται ως η εντροπία του χαρακτηριστικού. Το χαρακτηριστικό με τον υψηλότερο λόγο κέρδους επιλέγεται για το διαχωρισμό των δεδομένων. Μαθηματικά ο λόγος κέρδους αναπαρίσταται ως εξής:

$$Gain\ Ratio = \frac{IG}{SplitInfo} = \frac{E(\pi\rho\acute{\iota}\nu) - \sum_{j=1}^{K} E(j, \mu\varepsilon\tau\acute{\alpha})}{\sum_{j=1}^{K} w_j log_2 w_j}$$

Το CART (Δέντρο Ταξινόμησης και Παλινδρόμησης) χρησιμοποιεί την έννοια της μέτρησης Gini για να καθορίσει το χαρακτηριστικό που χωρίζει καλύτερα τα δεδομένα σε κάθε εσωτερικό κόμβο. Το Gini είναι ένα μέσο υπολογισμού της πιθανότητας λανθασμένης ταξινόμησης ενός τυχαία επιλεγμένου δείγματος και υπολογίζεται ως 1 μείον το άθροισμα των τετραγώνων των πιθανοτήτων των κλάσεων, p(i). Το χαρακτηριστικό με το χαμηλότερο Gini επιλέγεται για το διαχωρισμό των δεδομένων. Μαθηματικά, το Gini μπορεί να αναπαρασταθεί ως εξής:

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Ο Δείκτης Gini λειτουργεί με την κατηγορική μεταβλητή-στόχο "Επιτυχία" ή "Αποτυχία" και πραγματοποιεί μόνο δυαδικούς διαχωρισμούς. Η κατηγορική μεταβλητή-στόχος για τον υπολογισμό του δείκτη Gini αφορά την κατηγοριοποίηση των παρατηρήσεων σε ένα σύνολο δεδομένων.

Τα δέντρα απόφασης έχουν αρκετά πλεονεκτήματα σε σχέση με άλλες προσεγγίσεις κατηγοριοποίησης καθώς είναι εύκολα στην χρήση και αρκετά

αποτελεσματικά γεγονός που τα καθιστά από τους πιο δημοφιλείς κατηγοριοποιητές. Επίσης, λόγω της ανεξαρτησίας του δέντρου από το μέγεθος της βάσης δεδομένων, μπορούν να αποδώσουν πολύ καλά σε μεγάλες βάσεις δεδομένων (big data). Ωστόσο, τα μειονεκτήματά τους είναι ότι δεν μπορούν να χειριστούν με ευκολία συνεχή ή ελλιπή δεδομένα, εξαιτίας των πολλών και περίπλοκων διακλαδώσεων που δημιουργούνται με σκοπό την επίλυσή τους.

3.3 Gradient Boosting

Ο Gradient Boosting [16] είναι ένας αλγόριθμος μηχανικής μάθησης που ανήκει στην κατηγορία των μεθόδων συνόλου (ensemble methods). Μεταξύ άλλων, χρησιμοποιείται για την επίλυση προβλημάτων ταξινόμησης αλλά και παλινδρόμησης.

Ο αλγόριθμος αυτός βασίζεται στα δέντρα απόφασης για τα οποία μιλήσαμε παραπάνω, διότι διαιρεί τον χώρο των χαρακτηριστικών σε βαθμιαία μικρότερες περιοχές, μέχρι οι ετικέτες σε κάθε περιοχή να είναι όσο το δυνατό πιο όμοιες. Ωστόσο εκεί που διαφοροποιείται από αυτά είναι στην χρήση της τεχνικής “boosting”, καταπολεμώντας το βασικό πρόβλημα των δέντρων απόφασης, την υπερπροσαρμογή των δεδομένων.

Η ιδέα της ενίσχυσης (boosting) προέκυψε από την υπόθεση ότι ένας αδύναμος μαθητής μπορεί να τροποποιηθεί ώστε να γίνει καλύτερος. Η ενίσχυση δημιουργεί ένα μοντέλο συνόλου συνδυάζοντας διαδοχικά πολλά αδύναμα δέντρα αποφάσεων. Αναθέτει βάρη στην έξοδο των μεμονωμένων δέντρων. Στη συνέχεια, δίνει στις λανθασμένες ταξινομήσεις από το πρώτο δέντρο απόφασης υψηλότερο βάρος και είσοδο στο επόμενο δέντρο. Μετά από πολυάριθμους κύκλους, η μέθοδος boosting συνδυάζει αυτούς τους αδύναμους μαθητές διαδοχικά, θέτοντας ως είσοδο την έξοδο του προηγούμενου μαθητή, μέχρι να δημιουργηθεί ένας ενιαίος ισχυρός μαθητής πρόβλεψης όπου ελαχιστοποιεί όσο το περισσότερο δυνατό τα σφάλματα.

Ο Gradient Boosting χρησιμοποιεί τρία βασικά στοιχεία:

Α. Συνάρτηση Απώλειας

Η συνάρτηση απώλειας (Loss Function) που χρησιμοποιείται εξαρτάται από τον τύπο του προβλήματος που επιλύεται. Πρέπει να είναι διαφοροποιήσιμη, αλλά υποστηρίζονται πολλές τυπικές συναρτήσεις απώλειας και μπορούμε να ορίσουμε και τις δικές μας. Για παράδειγμα, η παλινδρόμηση μπορεί να χρησιμοποιεί τετραγωνικό σφάλμα και η ταξινόμηση μπορεί να χρησιμοποιεί λογαριθμική απώλεια.

Β. Αδύναμοι Μαθητές

Τα δέντρα αποφάσεων χρησιμοποιούνται ως ο αδύναμος μαθητής (weak learner) στον Gradient Boosting. Συγκεκριμένα χρησιμοποιούνται δέντρα παλινδρόμησης που δίνουν πραγματικές τιμές για τις διαχωριστικές γραμμές και των οποίων η

έξοδος μπορεί να προστεθεί μεταξύ τους, επιτρέποντας την προσθήκη των εξόδων των επόμενων μοντέλων και τη "διόρθωση" των υπολειμμάτων στις προβλέψεις.

Τα δέντρα κατασκευάζονται με άπληστο τρόπο, επιλέγοντας τα καλύτερα σημεία διάσπασης με βάση τις βαθμολογίες καθαρότητας (purity scores) όπως το Gini ή για την ελαχιστοποίηση της απώλειας. Συνήθως γίνεται περιορισμός στο πλήθος των αδύναμων μαθητών με συγκεκριμένους τρόπους, όπως ένας μέγιστος αριθμός επιπέδων, κόμβων, διαχωρισμών ή κόμβων φύλλων. Αυτό γίνεται για να εξασφαλιστεί ότι οι μαθητές παραμένουν αδύναμοι, αλλά μπορούν να κατασκευαστούν με άπληστο τρόπο.

Γ. Προσθετικό Μοντέλο

Τα δέντρα προστίθενται ένα-ένα και τα υπάρχοντα δέντρα στο μοντέλο δεν αλλάζουν (additive model). Χρησιμοποιείται μια διαδικασία βαθμιαίας κατάβασης για την ελαχιστοποίηση της απώλειας κατά την προσθήκη δέντρων.

Παραδοσιακά, η κάθοδος κλίσης (Gradient Descent) χρησιμοποιείται για την ελαχιστοποίηση ενός συνόλου παραμέτρων, όπως οι συντελεστές σε μια εξίσωση παλινδρόμησης ή τα βάρη σε ένα νευρωνικό δίκτυο. Μετά τον υπολογισμό του σφάλματος ή της απώλειας, τα βάρη ενημερώνονται για να ελαχιστοποιήσουν αυτό το σφάλμα.

Αντί για παραμέτρους, έχουμε αδύναμα υποδείγματα μαθητών ή πιο συγκεκριμένα δέντρα αποφάσεων. Μετά τον υπολογισμό της απώλειας, για να εκτελέσουμε τη διαδικασία βαθμιαίας καθόδου, πρέπει να προσθέσουμε ένα δέντρο στο μοντέλο που μειώνει την απώλεια (δηλαδή ακολουθεί την κλίση). Αυτό το κάνουμε με την παραμετροποίηση του δέντρου, στη συνέχεια τροποποιούμε τις παραμέτρους του δέντρου και κινούμαστε προς τη σωστή κατεύθυνση μειώνοντας την υπολειπόμενη απώλεια. Αυτή η προσέγγιση ονομάζεται λειτουργική κάθοδος κλίσης (functional gradient descent) ή κάθοδος κλίσης με συναρτήσεις.

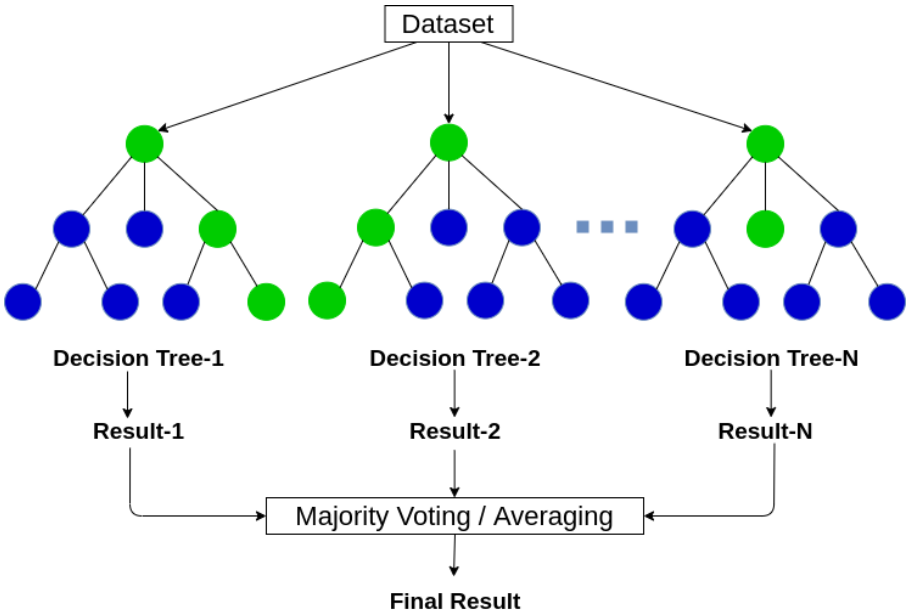
Η έξοδος για το νέο δέντρο προστίθεται στη συνέχεια στην έξοδο της υπάρχουσας ακολουθίας δέντρων σε μια προσπάθεια διόρθωσης ή βελτίωσης της τελικής εξόδου του μοντέλου. Έτσι, προστίθεται ένας σταθερός αριθμός δέντρων ή η εκπαίδευση σταματά όταν η απώλεια φτάσει σε ένα αποδεκτό επίπεδο ή δεν βελτιώνεται πλέον σε ένα εξωτερικό σύνολο δεδομένων επικύρωσης.

3.4 Τυχαία Δάση

Η τεχνική των τυχαίων δασών (random forest) [17] απέκτησε την ονομασία της από το γεγονός ότι εκμεταλλεύεται πολλαπλούς κατηγοριοποιητές δομής δέντρων απόφασης, ορίζοντας και αυτή μία ξεχωριστή δομή δέντρου. Συγκεκριμένα, η τεχνική των τυχαίων δασών δημιουργεί ένα σύνολο από δέντρα απόφασης (decision trees), όπου κάθε δέντρο παίρνει ένα τυχαία υποσύνολο των παρατηρήσεων και χαρακτηριστικών δεδομένων. Κατά τη διάρκεια της εκπαίδευσης, κάθε δέντρο αναπτύσσεται με βάση ένα διαφορετικό υποσύνολο των δεδομένων εκπαίδευσης και επιλέγει μία τυχαία τιμή των χαρακτηριστικών για την κατασκευή του. Έτσι, κάθε δέντρο του τυχαίου δάσους είναι μία ξεχωριστή δομή δέντρου απόφασης.

Χρησιμοποιεί τη μέθοδο συνόλων, η οποία συνδυάζει ένα σύνολο από n ίδια μοντέλα κατηγοριοποίησης $M_1, M_2, \dots, M_n$, έτσι ώστε να δημιουργήσει ένα πιο βελτιωμένο μοντέλο κατηγοριοποίησης M_ensemble. Έστω n υποσύνολα $D_1, D_2, \dots, D_n$ ενός συνόλου δεδομένων εκπαίδευσης D. Το κάθε υποσύνολο D_i χρησιμοποιείται για να τροφοδοτήσει και να δημιουργήσει το κάθε υπο-μοντέλο M_i. Αν στο τελικό μοντέλο δοθεί μια νέα πλειάδα προς κατηγοριοποίηση, τότε το κάθε υπο-μοντέλο θα «ψηφίσει» προβλέποντας την κλάση της πλειάδας. Στο τέλος, όλες οι ψήφοι των υπο-μοντέλων συγκεντρώνονται και μέσω πλειοψηφίας (ζυγισμένης ή όχι) ή άλλων τεχνικών, εξάγεται η τελική πρόβλεψη του μοντέλου για την πλειάδα.

Στην τεχνική κατηγοριοποίησης των Τυχαίων Δασών, όπως φαίνεται στην παρακάτω Εικόνα 3.3, το σύνολο εκπαίδευσης D τροφοδοτεί τις εισόδους n δέντρων απόφασης. Το κάθε δέντρο «χτίζεται» και κάνει μία πρόβλεψη, με βάση την διάσπαση χαρακτηριστικών που έχει εφαρμοστεί. Προσθέτοντας, τα δέντρα απόφασης αναπτύσσονται ως το μέγιστο βάθος τους και δεν γίνεται κλάδεμα των υπο-δέντρων. Επίσης, η επιλογή των χαρακτηριστικών για τον μετέπειτα διαχωρισμό τους σε κάθε κόμβο του δέντρου, δημιουργείται με τυχαίο τρόπο κάθε φορά. Στη συνέχεια, οι παραχθείσες προβλέψεις από τα n δέντρα απόφασης προστίθενται σε έναν αλγόριθμο που υπολογίζει την πλειοψηφία τους, επιλέγοντας την επικρατέστερη, η οποία θα δοθεί ως είσοδος για κάθε πλειάδα. Η τεχνική αυτή ονομάζεται Πλειοψηφική Ψήφος (Majority voting).



Εικόνα 3.3 Πλειοψηφική ψήφος [18]

Τα Τυχαία Δάση είναι από τους πιο αποδοτικούς και ανταγωνιστικούς κατηγοριοποιητές σε επίπεδο ακρίβειας. Η ακρίβειά τους, εξαρτάται από την ακρίβεια κάθε ξεχωριστού δέντρου απόφασης που δημιουργείται καθώς και της μεταξύ τους συσχέτισης. Είναι ανθεκτικά σε ακραίες τιμές και λάθη, διότι οι τιμές αυτές χάνονται μέσα στο πλήθος των Δασών απόφασης n. Επιπλέον, δεν πάσχουν από υπερπροσαρμογή, σε αντίθεση με τα Decision Trees, και αποδίδουν πολύ καλά σε μεγάλες βάσεις δεδομένων, αφού χρησιμοποιούν πολύ λιγότερα χαρακτηριστικά προς διαχωρισμό σε κάθε κόμβο – βήμα. Το πλήθος αυτών των χαρακτηριστικών δεν επηρεάζει αισθητά το αποτέλεσμα του μοντέλου και συνηθίζεται να είναι ίσο με $log_2 d + 1$.

Επιπρόσθετα, λόγω της μεγάλης πολυπλοκότητας του κατηγοριοποιητή, χρειάζεται πολλούς περισσότερους υπολογιστικούς πόρους. Αυτό το πρόβλημα μπορεί να μειωθεί παραλληλοποιώντας την διαδικασία δημιουργίας και πρόβλεψης δέντρων απόφασης, με χρήση πολλών πυρήνων επεξεργασίας. Τέλος, ο χρόνος που απαιτείται για την εκπαίδευση του μοντέλου είναι αρκετά περισσότερος σε σύγκριση με άλλα μοντέλα κατηγοριοποίησης, αλλά και αναλογικός με το πλήθος των δέντρων απόφασης και των δεδομένων εκπαίδευσης.

3.5 Πολυωνυμικός Naïve Bayes

Ο αλγόριθμος Naïve Bayes [19] είναι μια πιθανολογική προσέγγιση για την κατασκευή μοντέλων ταξινόμησης δεδομένων. Χρησιμοποιείται ως μία εναλλακτική λύση στην ομαδοποίηση με βάση την απόσταση K-Means και τα δάση δέντρων απόφασης και ασχολείται με την πιθανότητα ως την "εκτίμηση" ότι τα δεδομένα ανήκουν σε μια συγκεκριμένη κλάση. Υπάρχουν τα γκαουσιανά και τα πολυωνυμικά μοντέλα του Naïve Bayes.

Το πολυωνυμικό (multinomial) μοντέλο παρέχει τη δυνατότητα ταξινόμησης δεδομένων, τα οποία δεν μπορούν να αναπαρασταθούν αριθμητικά. Παρέχει τη δυνατότητα εκτέλεσης της κατηγοριοποίησης χρησιμοποιώντας μικρά σύνολα εκπαίδευσης, χωρίς να απαιτείται συνεχής επανεκπαίδευση και το κύριο πλεονέκτημά του είναι η σημαντικά μειωμένη πολυπλοκότητα.

Χρησιμοποιείται για την ταξινόμηση δεδομένων κειμένου σε διάφορες κατηγορίες, όπως spam/ham email ή ανάλυση συναισθήματος. Βασίζεται στο θεώρημα των πιθανοτήτων Bayes και ονομάζεται "αφελής" επειδή υποθέτει ανεξαρτησία μεταξύ των χαρακτηριστικών (λέξεων) στα δεδομένα, η οποία δεν είναι πάντα αληθής.

Για να καταλάβουμε πώς λειτουργεί ο Naïve Bayes, πρέπει πρώτα να κατανοήσουμε τον κανόνα του Bayes. Αυτό το μοντέλο πιθανοτήτων διατυπώθηκε από τον Thomas Bayes (1701-1761) και μπορεί να γραφτεί ως εξής:

Έστω ότι έχουμε την πιθανότητα P:

$$Μεταγενέστερη\ P = \frac{P\ Συνθήκης\ x\ Προηγούμενη\ P}{Πρόβλεψη\ Προηγούμενης\ P}$$

$$P(\frac{A}{B}) = \left(\frac{P(A \cap B)}{P(B)}\right) = \frac{P(A)xP(\frac{B}{A})}{P(B)}$$

Όπου,
PA η προηγούμενη πιθανότητα να συμβεί το Α
PBA η πιθανότητα συνθήκης του Β, δεδομένου ότι συνέβη το Α
PAB η πιθανότητα συνθήκης του Α, δεδομένου ότι συνέβη το Β
PB η πιθανότητα να συμβεί το Β
Η πιθανότητα να συμβεί ένα γεγονός Α όταν έχει ήδη συμβεί ένα άλλο γεγονός Β το οποίο συσχετίζεται με το Α ονομάζεται πιθανότητα συνθήκης.

$$P(\frac{B}{A}) = \left(\frac{P(A \cap B)}{P(A)}\right)$$

Η παραπάνω συνθήκη ισχύει μόνο όταν το P(A) είναι μεγαλύτερο του μηδενός.
Η προηγούμενη πιθανότητα ορίζεται ως την πιθανότητα που υπολογίσαμε πριν από την συλλογή νέων δεδομένων. Αυτή η πιθανότητα αναθεωρείται καθώς νέες πληροφορίες γίνονται διαθέσιμες για να παραχθούν ακριβέστερα αποτελέσματα.
Παράδειγμα 3.1: Αγώνας Ποδοσφαίρου

Weather	Sunny	Overcast	Rainy	Sunny	Sunny	Overcast	Rainy	Rainy	Sunny	Rainy	Sunny	Overcast	Overcast	Rainy
Play	No	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	No	Yes	Yes	No

Στον παραπάνω πίνακα 3.1, έχουμε ένα σύνολο δεδομένων για τις καιρικές συνθήκες που είναι ηλιόλουστες (Sunny), συννεφιασμένες (Overcast) και βροχερές (Rainy). Με βάση αυτό τον πίνακα θα προβλέψουμε αν θα πραγματοποιηθεί ένα υποθετικός αγώνας ποδοσφαίρου. Τώρα, πρέπει να προβλέψουμε την πιθανότητα για το αν οι παίκτες θα παίξουν με βάση τις καιρικές συνθήκες.
Αρχικά δημιουργούμε έναν πίνακα συχνοτήτων του συνόλου δεδομένων εκπαίδευσης που δίνεται στην παραπάνω περιγραφή του προβλήματος και καταγράφουμε την καταμέτρηση όλων των καιρικών συνθηκών έναντι της αντίστοιχης καιρικής συνθήκης.

Weather	Yes	No
Sunny	3	2
Overcast	4	0
Rainy	2	3
Total	9	5

Πίνακας 3.2
Βρίσκουμε τις πιθανότητες κάθε καιρικής συνθήκης και δημιουργούμε έναν πίνακα πιθανοτήτων.

Weather	Yes	No	
Sunny	3	2	=5/14(0.36)
Overcast	4	0	=4/14(0.29
Rainy	2	3	=5/14(0.36)
Total	9	5	
	=9/14(0.64)	=5/14(0.36)	

Πίνακας 3.3

Χρησιμοποιούμε την ακόλουθη εξίσωση για τον υπολογισμό της μεταγενέστερης πιθανότητας όλων των καιρικών συνθηκών:
P(A|B) = P(A) * P(B|A)/P(B)
*P(Yes|Sunny) = P(Yes) * P(Sunny|Yes) / P(Sunny)*
Παίρνουμε τις τιμές από τον παραπάνω πίνακα πιθανοτήτων και τις τοποθετούμε στον παραπάνω τύπο.
P(Sunny|Yes) = 3/9 = 0,33*, P(Yes)* = 0,64 και *P(Sunny)* = 0,36.
Επομένως, *P(Yes|Sunny)* = (0,64*0,33)/0,36 = 0,60
*P(No|Sunny) = P(No) * P(Sunny|No) / P(Sunny)*
Παίρνουμε τις τιμές από τον παραπάνω πίνακα πιθανοτήτων και τις τοποθετούμε στον παραπάνω τύπο.
P(Sunny|No) = 2/5 = 0,40*, P(No)* = 0,36 και *P(Sunny)* = 0,36
P(No|Sunny) = (0,36*0,40)/0,36 = 0,6 = 0,40

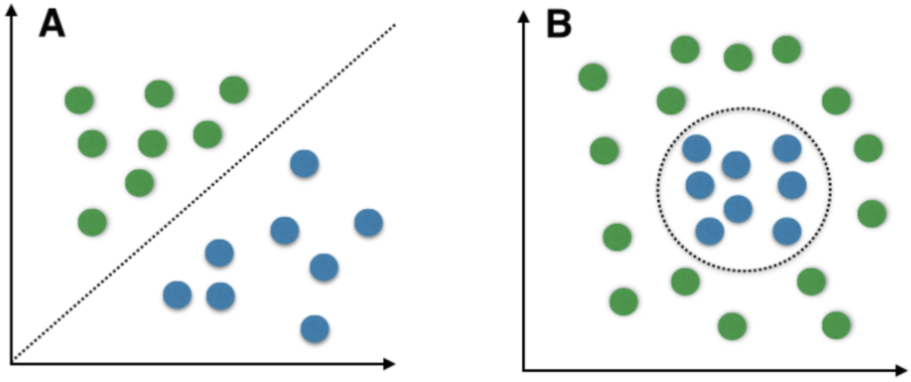
Η πιθανότητα να παίξει κάποιος σε συνθήκες ηλιοφάνειας είναι μεγαλύτερη. Ως εκ τούτου, ο παίκτης θα παίξει εάν ο καιρός είναι ηλιόλουστος.

3.6 Μηχανές Διανυσμάτων Υποστήριξης

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines – SVM) [21] είναι ένα σύνολο εποπτευόμενων μοντέλων μάθησης (supervised learning) με αλγορίθμους που χρησιμοποιούνται για την ταξινόμηση (classification) και την ανάλυση παλινδρόμησης (regression analysis). Οι θεμελιώδεις αρχές των SVM οι οποίες έχουν αναπτυχθεί από τον Vapnik, βασίζονται στη θεωρία της στατιστικής μάθησης (statistical learning). Έχουν χρησιμοποιηθεί σε ένα ευρύ φάσμα του πραγματικού κόσμου, όπως προβλήματα κατηγοριοποίησης κειμένου, αναγνώριση εικόνων, ήχου, ταξινόμηση και ανίχνευση δεδομένων. Επίσης έχουν μεγάλη εφαρμογή σε χρηματοοικονομικά θέματα, σημαντική συμβολή στον τομέα της βιοϊατρικής, όπως και σε εφαρμογές τεχνητής νοημοσύνης και ρομποτικής. Ο χρήστης πρέπει να κάνει εκτεταμένες προσπάθειες πολλές φορές μέχρι να βρει την βέλτιστη ρύθμιση των παραμέτρων που θα εισάγει στο SVM .Αυτή η διαδικασία είναι γνωστή ως επιλογή μοντέλου και σε πολλές περιπτώσεις μπορεί να αποδειχθεί αρκετά χρονοβόρα.
Η κεντρική ιδέα στην κατασκευή ενός αλγορίθμου SVM βασίζεται στο εσωτερικό γινόμενο μεταξύ του "support vector" $\boldsymbol{xi}$ και του διανύσματος $\boldsymbol{x}$ από την είσοδο. Τα support vectors αποτελούνται από ένα μικρό υποσύνολο των δεδομένων εκπαίδευσης. Σύμφωνα με το εσωτερικό αυτό γινόμενο, μπορούμε να κατασκευάσουμε διαφορετικές μηχανές εκμάθησης που χαρακτηρίζονται από δικές τους μη-γραμμικές επιφάνειες απόφασης. Στο πρόβλημα της ταξινόμησης , σκοπός μας είναι η εύρεση ενός βέλτιστου υπερεπιπέδου που χωρίζει τις δύο κλάσεις δεδομένων, μεγιστοποιώντας έτσι με αυτόν τον τρόπο το περιθώριο μεταξύ τους.
Τα SVM μπορούν να είναι δύο τύπων [22]:

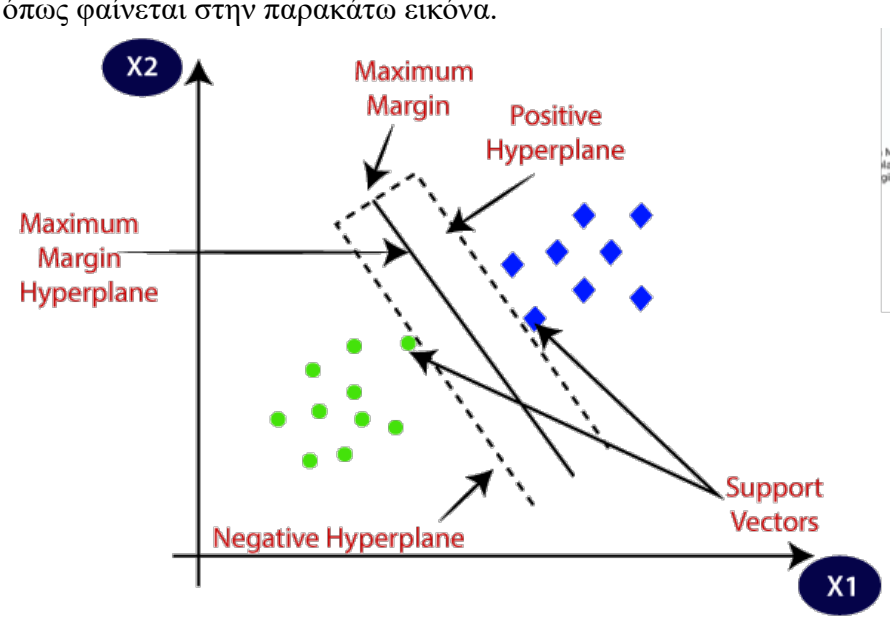
Γραμμικά SVM: Αυτό σημαίνει ότι αν ένα σύνολο δεδομένων μπορεί να ταξινομηθεί σε δύο κλάσεις με τη χρήση μιας μόνο ευθείας γραμμής, τότε τα δεδομένα αυτά ονομάζονται γραμμικά διαχωρίσιμα δεδομένα και ο ταξινομητής που χρησιμοποιείται ονομάζεται ταξινομητής γραμμικός (Linear) SVM.

Μη γραμμικά SVM: Το μη γραμμικό SVM χρησιμοποιείται για μη γραμμικά διαχωρισμένα δεδομένα, πράγμα που σημαίνει ότι εάν ένα σύνολο δεδομένων δεν μπορεί να ταξινομηθεί με τη χρήση μιας ευθείας γραμμής, τότε τα δεδομένα αυτά ονομάζονται μη γραμμικά δεδομένα και ο ταξινομητής που χρησιμοποιείται ονομάζεται μη γραμμικός (Non-linear) ταξινομητής SVM.



Εικόνα 3.7: Γραμμικό (A) και Μη Γραμμικό (B) SVM [23]

Στην προκείμενη περίπτωση αυτής της διπλωματικής, μας ενδιαφέρει μόνο η γραμμική περίπτωση των SVM, διότι έχουμε ένα δυαδικά διαχωρίσιμο πρόβλημα με δύο κλάσεις, όπως θα αναφερθεί λεπτομερώς στο επόμενο κεφάλαιο.
Τα δυαδικά SVM είναι βασισμένα πάνω στην αρχή της ελαχιστοποίησης του διορθωτικού κινδύνου από την υπολογιστική θεωρία (structural risk minimization principle from the computational learning theory). Τα δυαδικά SVM μοντέλα επιδιώκουν να διαχωρίζουν τα σημεία εκπαίδευσης σε δύο τάξεις και παίρνουν αποφάσεις για το που θα κατατάξουν τα δεδομένα που εξετάζονται βάσει των βοηθητικών διανυσμάτων (support vectors) που επιλέγονται ως τα μόνα αποτελεσματικά στοιχεία για τα ζεύγη εκπαίδευσης του μοντέλου. Με τον όρο support vectors εννοούμε τα διανύσματα των σημείων που δημιουργούνται από τα πιο κοντινά σημεία και των δύο κλάσεων ως προς την γραμμή διαχωρισμού τους, όπως φαίνεται στην παρακάτω εικόνα.



Εικόνα 3.8 Γραμμικός Διαχωρισμός SVM [24]

Δεδομένου ενός συνόλου N γραμμικών διαχωρισμένων σημείων $S = x(i) \in R \mid i = 1, 2, \ldots, N$, όπου κάθε xi σημείο ανήκει σε μια από τις δύο κλάσεις, που ορίζονται μαθηματικά ως $y(i)\ \varepsilon - 1, +1$. Το διάνυσμα που διαχωρίζει τα στοιχεία που ανήκουν στο S στις 2 κλάσεις $y(i)$ ονομάζεται hyper-plane. Το διάνυσμα hyper-plane μπορεί να οριστεί σαν ένα ζεύγος τιμών της μορφής (w,b) που ικανοποιεί τη συνθήκη $w * x + b = 0$, διαδοχικά δημιουργούνται τα παρακάτω διανύσματα, παράλληλα στο προαναφερθέν διάνυσμα:

$$w * x + b = 0\ x * xi \geq 1, if\ yi = +1\ x * xi \leq -1, if\ yi = -1$$

Έτσι βάσει των παραπάνω ο στόχος ενός binary SVM αλγορίθμου είναι να βρει το πιο σωστά διαχωρίσιμο hyper-plane (OSH – Optimal separating Hyper Plane) που έχει το μέγιστο περιθώριο των δύο πλευρών διαχωρισμού. Κατά την διάρκεια της κατηγοριοποίησης, ο αλγόριθμος SVM τις περισσότερες φορές παίρνει αποφάσεις που είναι βασισμένες στο OSH αντί ολόκληρης της βάσης δεδομένων των στοιχείων εκπαίδευσης. Στην ουσία βρίσκει από ποια πλευρά του OSH εντοπίζεται το σημείο δοκιμής. Αυτή η ιδιότητα καθιστά το SVM πολύ ανταγωνιστικό σε σύγκριση με άλλους παραδοσιακούς αλγορίθμους αναγνώρισης,

όσον αναφορά την αποτελεσματικότητα και ακρίβεια. Παρά τα πολλά πλεονεκτήματα των αλγορίθμων SVM, υπάρχουν κάποιες πολύ έντονες αδυναμίες μεταξύ των οποίων είναι: η επιλογή παραμέτρων, η αλγοριθμική πολυπλοκότητα που επηρεάζει το χρόνο εκπαίδευσης του ταξινομητή σε μεγάλα σύνολα δεδομένων και η υποβέλτιστη απόδοση των SVM σε μη ισορροπημένα δεδομένα.

3.7 Κατάβαση Πλαγιάς

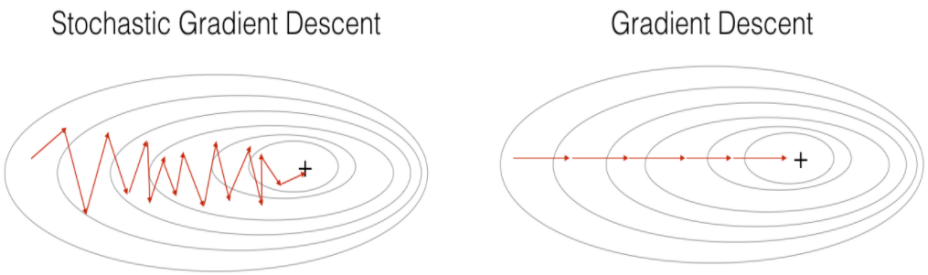
Ο αλγόριθμος Κατάβασης Πλαγιάς(Gradient Descent) [25] είναι ένας αλγόριθμος βελτιστοποίησης, με τη χρήση του οποίου επιχειρείται η ελαχιστοποίηση κάποιας συνάρτησης, με επαναληπτικά βήματα σταδιακής μεταβολής των εισόδων σε αυτή. Στόχος του είναι η ελαχιστοποίηση της συνάρτησης κόστους. Η συνάρτηση κόστους είναι συνάρτηση των εξόδων του Τεχνητού Νευρωνικού Δικτύου και εφόσον οι έξοδοι υπολογίζονται με βάση τις παραμέτρους του μοντέλου, μπορεί να θεωρηθεί συνάρτηση των παραμέτρων. Έτσι, με την Κατάβαση Πλαγιάς επιδιώκεται η επαναληπτική μεταβολή των παραμέτρων του μοντέλου, ώστε η συνάρτηση κόστους να προσεγγίσει κάποιο ελάχιστο. Όπως περιγράφει το όνομα του αλγορίθμου, η ελαχιστοποίηση επιτυγχάνεται με τον υπολογισμό του διανύσματος κλίσης της συνάρτησης κόστους ως προς τις παραμέτρους του δικτύου. Παρατηρούμε ότι το διάνυσμα αυτό έχει την κατεύθυνση της ταχύτερης αύξησης της συνάρτησης κόστους, ενώ το αντίθετό του την κατεύθυνση της ταχύτερης μείωσης. Επομένως, κατά τον αλγόριθμο Gradient Descent επιλέγονται διαδοχικά βήματα ανάλογα προς το αντίθετο τού διανύσματος κλίσης. Ο συντελεστής αναλογίας, ο οποίος καθορίζει πόσο μεγάλα θα είναι αυτά τα βήματα, ονομάζεται ρυθμός εκμάθησης (learning rate).
Για τη μαθηματική διατύπωση του βήματος του αλγορίθμου, θεωρούμε μία συνάρτηση *F(x)* καλά ορισμένη και παραγωγίσιμη στη γειτονιά ενός σημείου *α*. Τότε, η κατεύθυνση ταχύτερης μείωσης της συνάρτησης είναι αυτή της αντίθετης κλίσης (gradient) της *F* στο σημείο *α*, δηλαδή $-\nabla F$. Επομένως η επαναληπτική διαδικασία είναι αυτή που περιγράφεται στην εξίσωση 2.16.

$$\alpha(n) + 1 = \alpha(n) - \eta \nabla F\alpha(n) \quad (2.16)$$

Έτσι παράγεται μια ακολουθία διαδοχικών βημάτων, η οποία περιγράφεται από την ακολουθία σημείων *α0, α1, α2, . . ., με Fα(0) >= Fα(1) >= Fα(2) >= ...,* με σκοπό τελικά να οδηγηθούμε σε κάποιο a_m το οποίο θα αποτελεί ελάχιστο της συνάρτησης. Ο όρος *η* στην παραπάνω σχέση είναι ο ρυθμός εκμάθησης, ο οποίος πολλαπλασιάζεται με την υπολογισμένη κλίση και καθορίζει κατά πόσο θα μεταβληθεί η παράμετρος *α*.
Ένα από τα προβλήματα που μπορεί να συναντήσει ο Gradient Descent είναι τα τοπικά ελάχιστα. Στα τοπικά ελάχιστα το μέτρο της κλίσης της παραγώγου μηδενίζεται και έτσι ο αλγόριθμος «παγιδεύεται» στο τοπικό ελάχιστο, το οποίο δεν είναι βέλτιστη λύση για την ελαχιστοποίηση της συνάρτησης. Επίσης, άλλη μία πρόκληση είναι τα σαγματικά σημεία (saddle points), τα οποία δεν αποτελούν ακρότατο της συνάρτησης, παρόλα αυτά η κλίση μηδενίζεται και ο αλγόριθμος συγκλίνει και πάλι σε ένα υποβέλτιστο σημείο.

3.7.1 Στοχαστική Κατάβαση Πλαγιάς

Στην παρούσα Διπλωματική, θα χρησιμοποιήσουμε τον αλγόριθμο της Στοχαστικής Κατάβασης Πλαγιάς (Stochastic Gradient Descent – SGD) [26] έναντι του απλού GD για το μοντέλο ταξινόμησής μας.Ο αλγόριθμος της Στοχαστικής Κατάβασης Πλαγιάς (Stochastic Gradient Descent) βασίζεται στη γενική ιδέα της Κατάβασης Πλαγιάς που αναφέρθηκε παραπάνω, με τη διαφορά ότι η κλίση υπολογίζεται με βάση ένα υποσύνολο του σετ δεδομένων αντί να χρησιμοποιηθεί ολόκληρο. Πρακτικά χρησιμοποιείται μια στατιστική εκτίμηση, η οποία επιταχύνει σημαντικά τις επαναλήψεις και είναι σχεδόν απαραίτητη σε προβλήματα υψηλών διαστάσεων, όπως το δικό μας. Σημειώνεται ότι στην ακραία περίπτωση, η εκτίμηση της κλίσης μπορεί να γίνει από ένα μόνο δείγμα. Στην πράξη όμως, συνήθως χρησιμοποιούνται αρκετά δείγματα, τα οποία αποτελούν το λεγόμενο mini-batch. Έτσι συχνά η συγκεκριμένη παραλλαγή καλείται και mini-batch gradient descent. Ίσως φανεί λογικό εκ πρώτης όψεως να θεωρηθεί ότι η στοχαστική κατάβαση πλαγιάς, ως μία στατιστική προσέγγιση, είναι χαμηλότερης ποιότητας αλγόριθμος. Αντίθετα όμως, στην πραγματικότητα μπορεί να αντιμετωπίσει αρκετά αποτελεσματικά τα προβλήματα της παγίδευσης του αλγορίθμου σε τοπικά ελάχιστα και σαγματικά σημεία. Αυτό είναι απότοκο της τυχαιότητας που εισάγεται με τη χρήση λιγότερων δειγμάτων για την εκτίμηση, πράγμα που οδηγεί σε μία τυχαιότητα στην κατεύθυνση του διανύσματος κλίσης που υπολογίζεται. Δίνεται έτσι η δυνατότητα ο αλγόριθμος, μέσω της τυχαιότητας αυτής, να «αποδράσει» από τις περιοχές αυτές, όπως φαίνεται στην παρακάτω εικόνα.



Εικόνα 3.9 [27]

Κατά την απεικόνιση του αλγορίθμου στοχαστικής κατάβασης πλαγιάς, η διαδρομή που ακολουθούν οι ενημερώσεις των παραμέτρων μοιάζει με μια σειρά από ζιγκ-ζαγκ ή αιχμές. Αυτή η συμπεριφορά εμφανίζεται επειδή ο αλγόριθμος βασίζεται σε ολόκληρο το σύνολο δεδομένων για κάθε ενημέρωση παραμέτρου, με αποτέλεσμα να διατρέχει μπρος-πίσω κατά μήκος των περιγραμμάτων της συνάρτησης κόστους.

Από την άλλη πλευρά, η απλή κάθοδος πλαγιάς ενημερώνει τις παραμέτρους χρησιμοποιώντας ένα μόνο παράδειγμα εκπαίδευσης (ή μια μικρή δέσμη παραδειγμάτων) κάθε φορά. Αντί να υπολογίζει την κλίση σε ολόκληρο το σύνολο δεδομένων, η Κάθοδος Πλαγιάς υπολογίζει την κλίση της συνάρτησης κόστους με βάση ένα τυχαία επιλεγμένο δείγμα. Στη συνέχεια, οι ενημερώσεις των παραμέτρων πραγματοποιούνται με βάση αυτή την εκτιμώμενη κλίση. Γι’ αυτό το λόγο, η απεικόνισή της είναι μία ευθεία γραμμή.

Κεφάλαιο 4
Μοντέλο Ταξινόμησης Ψευδοειδήσεων

4.1Εισαγωγή

Σε αυτό το κεφάλαιο θα αναλύσουμε την διαδικασία ανάπτυξης και δημιουργίας του μοντέλου ταξινόμησης που υλοποιήσαμε. Συγκεκριμένα, το μοντέλο θα χρησιμοποιεί τους αλγορίθμους μηχανικής μάθησης που αναφέραμε στο κεφάλαιο 3 και θα πραγματοποιεί δυαδική ταξινόμηση ειδησεογραφικών άρθρων και σύντομων δηλώσεων, κατηγοριοποιώντας τα νέα είτε ως αληθή είτε ως ψευδή. Επιπλέον θα αναλυθούν τα εργαλεία και σύνολα δεδομένων που χρησιμοποιήθηκαν, οι μέθοδοι αξιολόγησης των αλγορίθμων καθώς και η παρουσίαση των ευρημάτων. Τέλος, θα αναλύσουμε τα αποτελέσματα και θα αντιπαραθέσουμε τα δυνατά και τα αδύναμα σημεία παράλληλα με τις πιθανές βελτιώσεις και επεκτάσεις του μοντέλου.

4.2 Εργαλεία Δημιουργίας Μοντέλου

Το πρακτικό κομμάτι της παρούσας διπλωματικής εργασίας εκτελέστηκε με την χρήση της γλώσσας Python. Η Python [28] είναι μια ευέλικτη γλώσσα προγραμματισμού υψηλού επιπέδου που διαθέτει μια πληθώρα βιβλιοθηκών, ενσωματωμένων εργαλείων και δομών δεδομένων που καθιστούν την επεξεργασία δεδομένων ευκολότερη. Επιπλέον, η Φιλοσοφία Προσανατολισμένη στα Αντικείμενα (Object-Oriented Philosophy) της Python καθιστά εύκολη τη δημιουργία μοντέλων ταξινόμησης και αναγνώρισης προτύπων με τη χρήση αλγορίθμων μηχανικής μάθησης. Στο πλαίσιο αυτό, η βιβλιοθήκες scikit-learn και pandas περιλαμβάνουν έτοιμες υλοποιήσεις αλγορίθμων ταξινόμησης με ενσωματωμένη χρήση επεξεργασίας φυσικής γλώσσας, συμπεριλαμβανομένων των προαναφερόμενων στο κεφάλαιο 2 αλγορίθμων που θα αξιοποιήσουμε.
Το περιβάλλον υλοποίησης του κώδικα είναι το GoogleColab [29], μία cloud-based (βασισμένη στο υπολογιστικό νέφος) πλατφόρμα ανοιχτού κώδικα που αναπτύχθηκε από την Google Research και επιτρέπει στους χρήστες να γράφουν και να εκτελούν κώδικα Python σε ένα πρόγραμμα περιήγησης στο διαδίκτυο. Το Colab είναι δωρεάν και προσφέρει πρόσβαση σε υπολογιστική υποδομή, όπως αποθήκευση, μνήμη, ικανότητα επεξεργασίας και μονάδες επεξεργασίας γραφικών και επεξεργασίας τανυστών. Έχει σχεδιαστεί για προγραμματιστές μηχανικής μάθησης, επιστήμονες δεδομένων, αναλυτές μεγάλων δεδομένων, ερευνητές ΤΝ και μαθητές Python. Το Colab παρέχει πολλά χαρακτηριστικά, όπως GPUs και TPUs, συνεργατική κωδικοποίηση, προεγκατεστημένες βιβλιοθήκες και αποθήκευση στο cloud, ενώ επιτρέπει τη σύνδεση με το GitHub για εύκολη εισαγωγή και εξαγωγή αρχείων κώδικα.

4.3 Σύνολα Δεδομένων

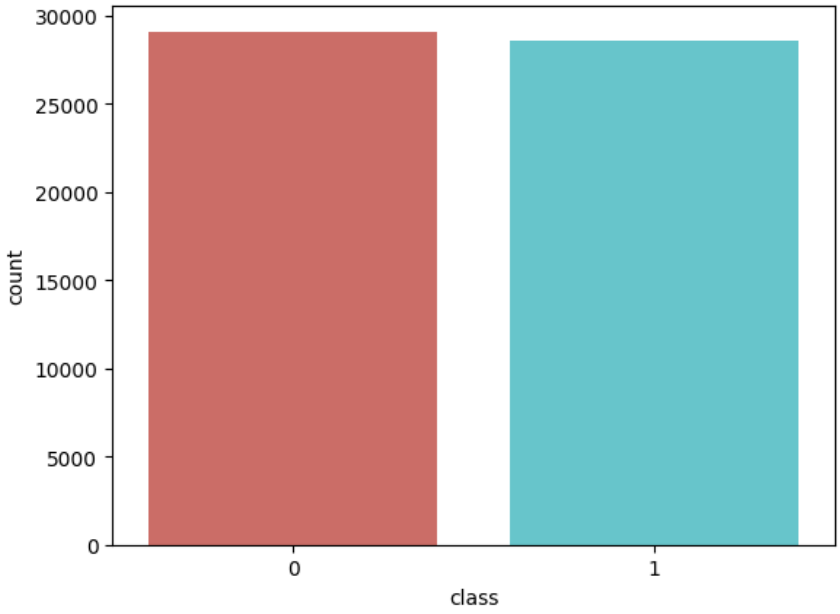
Τα σύνολα δεδομένων που χρησιμοποιήθηκαν έχουν κατανεμηθεί και αξιολογηθεί από το PolitiFact. Το PolitiFact είναι ένας δημοσιογραφικός οργανισμός και μέλος του IFCN (International Fact checking code of principles) που ελέγχει τα γεγονότα και επικεντρώνεται στον έλεγχο της ακρίβειας συγκεκριμένων δηλώσεων που γίνονται από πολιτικούς [30]. Το PolitiFact αξιολογεί τα σύνολα δεδομένων του για τις ψευδείς ειδήσεις συνεργαζόμενο με άλλους ειδησεογραφικούς οργανισμούς για τη δημιουργία ιστότοπων ελέγχου γεγονότων σε επίπεδο πολιτείας και συνεργαζόμενο με πλατφόρμες κοινωνικής δικτύωσης, όπως το Facebook και το TikTok, για την εξέταση δυνητικά παραπλανητικών αναρτήσεων [31]. Επιπλέον, το PolitiFact παράγει έσοδα μέσω συνεργασιών περιεχομένου, διαδικτυακής διαφήμισης και συνδρομών. Γνωστοποιεί ατομικές δωρεές ή επιχορηγήσεις άνω των 1.000 δολαρίων και οργανισμούς που συνεισφέρουν περισσότερο από το 5% των συνολικών εσόδων κατά το προηγούμενο ημερολογιακό έτος. Δεν δέχεται δωρεές από ανώνυμες πηγές, πολιτικά κόμματα, εκλεγμένους αξιωματούχους, υποψηφίους που διεκδικούν δημόσιο αξίωμα ή οποιαδήποτε άλλη πηγή που θα μπορούσε να θεωρηθεί σύγκρουση συμφερόντων.

Συγκεκριμένα, τα σύνολα δεδομένων που χρησιμοποιήθηκαν είναι:

1. **ISOT Fake News Dataset**
Το σύνολο δεδομένων αυτό περιέχει συνολικά 44.898 άρθρα, εκ των οποίων 21.417 είναι από το Reuters.com με αληθές περιεχόμενο, ενώ τα 23.481 με ψευδείς ειδήσεις είναι από διάφορες πηγές οι οποίες έχουν επισημανθεί από το PolitiFact ως αναξιόπιστες. Το σύνολο δεδομένων αυτό περιέχει διαφορετικούς τύπους άρθρων σε διάφορα
θέματα, ωστόσο, η πλειοψηφία των άρθρων εστιάζει σε θέματα πολιτικών και παγκόσμιων ειδήσεων [32].

2. **LIAR**
Το LIAR είναι ένα δημόσια διαθέσιμο σύνολο δεδομένων για την ανίχνευση ψευδών ειδήσεων. Το σύνολο δεδομένων αυτό περιλαμβάνει 12.800 σύντομες δηλώσεις με ανθρώπινη επισήμανση από το API του PolitiFact και κάθε δήλωση αξιολογείται από έναν συντάκτη του ως προς την αληθοφάνειά της . [33]

Επομένως, το συνολικό dataset θα αποτελείται από ένα συνδυασμό άρθρων και σύντομων δηλώσεων, εκ των οποίων θα είναι 29.092 ειδήσεις ψευδείς και 28.552 αληθείς, καθιστώντας το αρκετά ισορροπημένο, με μικρή διαφοροποίηση ανάμεσα στις δύο κλάσεις του (Εικόνα 4.1).



Εικόνα 4.1: Σύγκριση κλάσεων

4.4 Μετρικές Αξιολόγησης Αλγορίθμων

Οι μετρικές αξιολόγησης είναι ένα σημαντικό εργαλείο για την εκτίμηση της απόδοσης των αλγορίθμων μηχανικής μάθησης. Αυτές οι μετρικές καθορίζουν πόσο καλά λειτουργεί ένας αλγόριθμος σε σχέση με το σύνολο δεδομένων εκπαίδευσης και ελέγχου. Είναι σημαντικό να χρησιμοποιούμε μια ποικιλία μετρικών αξιολόγησης προκειμένου να έχουμε μία καλή εικόνα των επιδόσεων αλλά και να είμαστε σε θέση να βελτιώσουμε τη συνολική προβλεπτική ικανότητα του μοντέλου. Στην υλοποίησή μας, έχουμε ένα μοντέλο δυαδικής ταξινόμησης υπό επίβλεψη, άρα υπάρχουν μόνο δύο πιθανές κλάσεις εξόδου(δηλ. διχοτόμηση). Επομένως θα επικεντρωθούμε μόνο στην δυαδική ταξινόμηση και οι μετρικές που θα συγκρίνουμε είναι το Precision, Accuracy, recall, F1 score, Confusion Matrix και η καμπύλη ROC-AUC, οι οποίες περιγράφονται στη συνέχεια.

4.4.1 Precision:

Το Precision [34] εξηγεί πόσες από τις σωστά προβλεπόμενες περιπτώσεις αποδείχθηκαν πράγματι θετικές. Είναι χρήσιμο στις περιπτώσεις όπου τα False Positives (Λανθασμένως ταξινομημένα Θετικά) αποτελούν μεγαλύτερη προτεραιότητα από τα False Negatives (Λανθασμένως ταξινομημένα Αρνητικά). Υπολογίζεται από την παρακάτω εξίσωση:

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}}$$

Όπου,

- TP = True Positives (Σωστά ταξινομημένα Θετικά)
- FP = False Positives (Λανθασμένως ταξινομημένα Θετικά)

4.4.2 Accuracy:

Το Accuracy [34] είναι ένα από τα πιο βασικά μέτρα αξιολόγησης της απόδοσης των αλγορίθμων μηχανικής μάθησης. Υπολογίζει το ποσοστό των σωστών προβλέψεων του μοντέλου σε σχέση με το συνολικό αριθμό των προβλέψεων που έγιναν. Το μόνο αρνητικό του είναι ότι δεν λαμβάνει υπόψη τον αριθμό των λανθασμένων προβλέψεων, οπότε δεν είναι κατάλληλο για όλες τις περιπτώσεις. Για παράδειγμα, σε περιπτώσεις με ανισορροπία κλάσεων, κάτι που στην προκειμένη περίπτωση δεν ισχύει, διότι έχουμε ένα πολύ ισορροπημένο σύνολο δεδομένων. Επομένως, θα δώσουμε αρκετή βάση σε αυτή την μετρική. Η εξίσωση υπολογισμού της είναι η εξής:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP} + \boldsymbol{TN}}{(\boldsymbol{TP} + \boldsymbol{FP} + \boldsymbol{TN} + \boldsymbol{FN})}$$

Όπου,

- TP = True Positives (Σωστά ταξινομημένα Θετικά)
- FP = False Positives (Λανθασμένως ταξινομημένα Θετικά)
- TN = True Negatives (Σωστά ταξινομημένα Αρνητικά)
- FN = False Negatives (Λανθασμένως ταξινομημένα Αρνητικά)

4.4.3 Recall:

Είναι το μέτρο των σωστά αναγνωρισμένων θετικών περιπτώσεων (True Positives) από όλες τις πραγματικές θετικές περιπτώσεις [34]. Είναι σημαντικό όταν το ποσοστό των ψευδών αρνητικών (False Negatives) είναι υψηλό. Υπολογίζεται από τον τύπο:

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}}$$

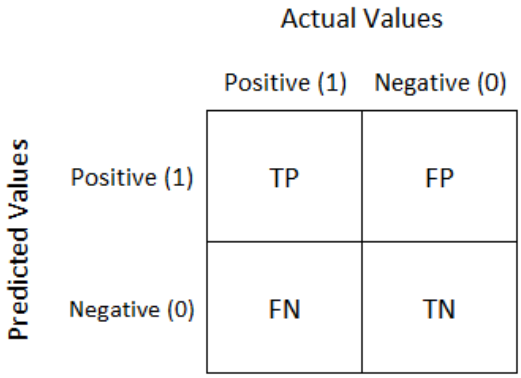
4.4.4 F1 score:

Πρόκειται για τον αρμονικό μέσο όρο του Accuracy και του Recall [34] και αποτελεί ένα καλύτερο μέτρο για τις εσφαλμένα ταξινομημένες περιπτώσεις από ότι το Accuracy. Παίρνει υπόψη τόσο την ακρίβεια (precision) όσο και την ανάκληση (recall) του συστήματος και παρέχει έναν αριθμητικό δείκτη που αντιπροσωπεύει τη συνολική απόδοση του αλγορίθμου. Ο δείκτης αυτός κυμαίνεται από 0 έως 1, με τιμές κοντά στο 1 να αντιπροσωπεύουν υψηλή απόδοση και τιμές κοντά στο 0 να αντιπροσωπεύουν χαμηλή απόδοση. Το F1 score είναι ιδιαίτερα χρήσιμο όταν υπάρχει ανισορροπία κλάσεων, καθώς λαμβάνει υπόψη και την ακρίβεια και την ανάκληση για κάθε κλάση. Ο τύπος υπολογισμού του είναι:

$$\boldsymbol{F1\ score} = \left(\frac{\boldsymbol{Recall^{-1}} + \boldsymbol{Precision^{-1}}}{\boldsymbol{2}}\right)^{\boldsymbol{-1}}$$

4.4.5 Confusion Matrix:

Ο πίνακας σύγχυσης ή Confusion Matrix [35] είναι πιο πολύ μια μέτρηση απόδοσης παρά αξιολόγησης για τα προβλήματα ταξινόμησης όπου η έξοδος μπορεί να είναι δύο ή περισσότερες κλάσεις. Μας βοηθά να οπτικοποιήσουμε τα ταξινομημένα δεδομένα καλύτερα ώστε να έχουμε μια κατά προσέγγιση εικόνα για την απόδοση του αλγορίθμου. Είναι ένας πίνακας με συνδυασμούς προβλεπόμενων και πραγματικών τιμών όπως φαίνεται στην παρακάτω εικόνα.

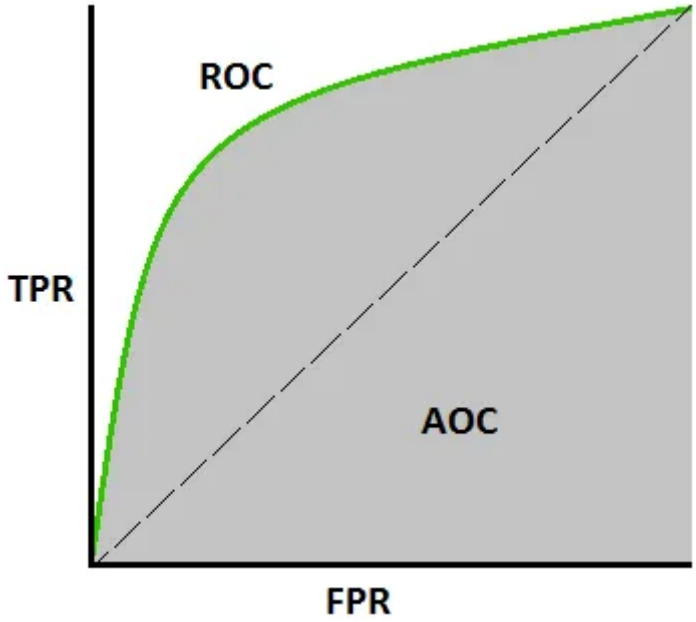


Εικόνα 4.2: Confusion Matrix [36]

Με βάση αυτόν τον πίνακα βλέπουμε κατά πόσο σωστά ταξινομήθηκαν οι τιμές των κλάσεων, συγκρίνοντας τις πραγματικές τιμές του άξονα X, με τις τιμές πρόβλεψης του άξονα Y. Εάν ο αλγόριθμος ταξινόμησε σωστά τις τιμές, τότε πρέπει η πλειοψηφία τους να βρίσκεται στο δεύτερο και στο τέταρτο τεταρτημόριο, δηλαδή στα True Positives και στα True Negatives αντίστοιχα. Πολλές τιμές στο πρώτο και τρίτο τεταρτημόριο (False Positives και False Negatives) σημαίνει ότι ο αλγόριθμος έκανε κακή κατηγοριοποίηση μεταξύ των κλάσεων.

4.4.6 ROC-AUC:

Η καμπύλη AUC – ROC [37] είναι μια μέτρηση απόδοσης για προβλήματα ταξινόμησης σε διάφορες τιμές κατωφλίου (threshold). Η ROC (receiver operating characteristic curve/καμπύλη λειτουργίας δείκτη) είναι μια καμπύλη πιθανοτήτων και η AUC (Area Under the ROC curve/Περιοχή κάτω από την καμπύλη ROC) αντιπροσωπεύει το βαθμό ή το μέτρο της διαχωρισιμότητας. Δείχνει κατά πόσο το μοντέλο είναι ικανό να διακρίνει μεταξύ των κλάσεων. Όσο υψηλότερη είναι η AUC, τόσο καλύτερο είναι το μοντέλο στο να προβλέπει τις κλάσεις 0 ως 0 και τις κλάσεις 1 ως 1. Κατ' αναλογία, όσο υψηλότερη είναι η AUC, τόσο καλύτερο είναι το μοντέλο στο να ταξινομεί τα δεδομένα.



Εικόνα 4.3: ROC-AUC [38]

Η καμπύλη ROC είναι ένα γράφημα που δείχνει την απόδοση ενός μοντέλου ταξινόμησης σε όλα τα όρια ταξινόμησης. Η καμπύλη αυτή απεικονίζει δύο παραμέτρους:

- TPR = True Positive Rate (Ποσοστό Σωστά ταξινομημένων Θετικών)
- FPR = False Positive Rate (Ποσοστό Λανθασμένως ταξινομημένων Θετικών)

Το TPR είναι συνώνυμο του Recall, επομένως ορίζεται ως:

$$\boldsymbol{TPR = \frac{TP}{TP + FN}}$$

Το FPR ορίζεται ως:

$$\boldsymbol{FPR = \frac{FP}{FP + TN}}$$

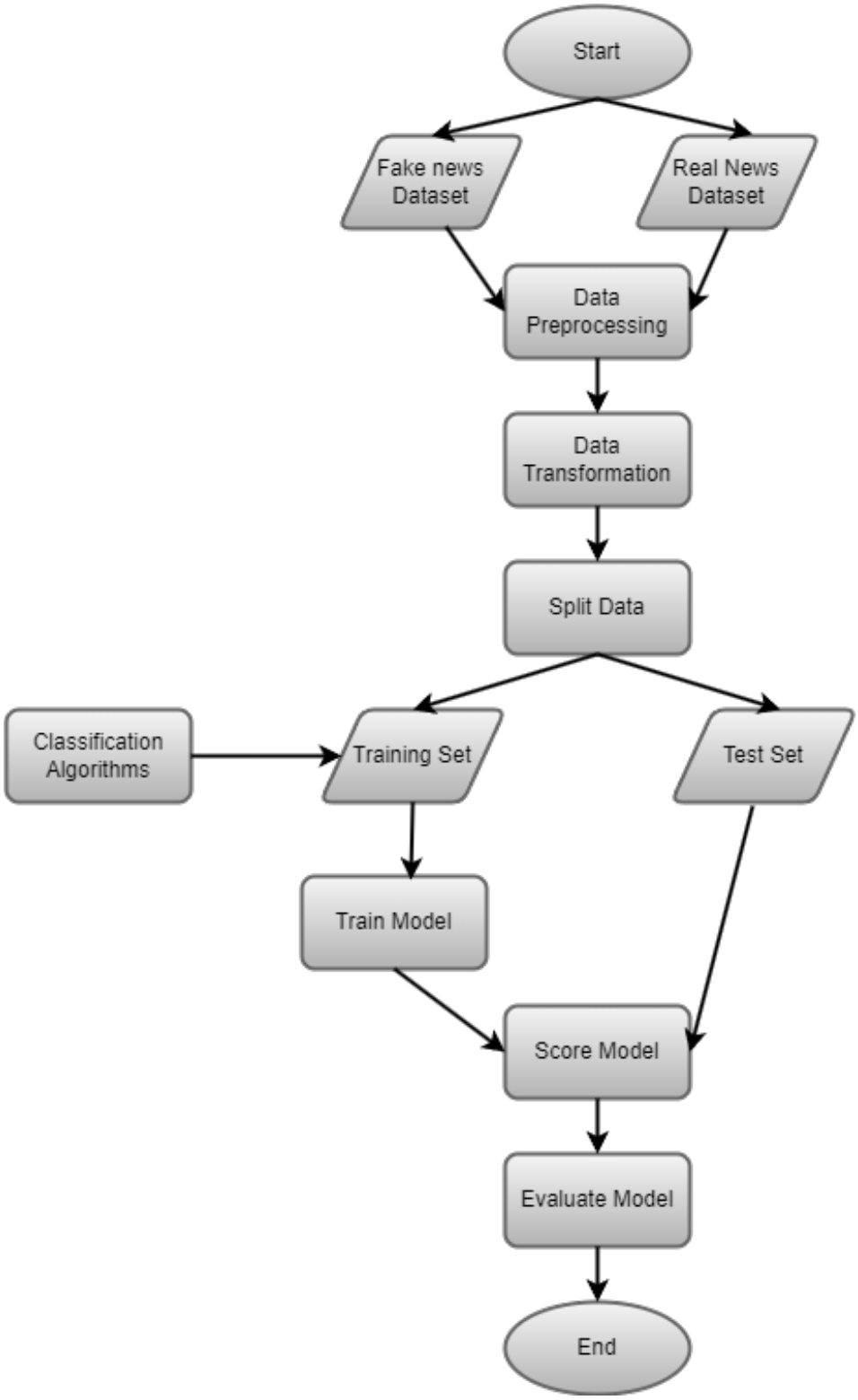
Μια καμπύλη ROC απεικονίζει το TPR σε σχέση με το FPR σε διαφορετικά κατώφλια ταξινόμησης. Η μείωση του κατωφλίου ταξινόμησης ταξινομεί

περισσότερα στοιχεία ως θετικά, αυξάνοντας έτσι τόσο τα False Positives όσο και τα True Positives.
Η τιμή της AUC κυμαίνεται από 0 έως 1. Ένα μοντέλο του οποίου οι προβλέψεις είναι 100% λανθασμένες έχει AUC 0, ενώ ένα άριστο μοντέλο του οποίου οι προβλέψεις είναι 100% σωστές έχει AUC 1. Στην περίπτωση που το AUC είναι 0,5, σημαίνει ότι το μοντέλο δεν έχει καμία απολύτως ικανότητα διαχωρισμού κλάσεων.
Η AUC είναι επιθυμητή για τους ακόλουθους δύο λόγους:

- Είναι αναλλοίωτη κλίμακας. Μετράει πόσο καλά κατατάσσονται οι προβλέψεις και όχι τις απόλυτες τιμές τους.
- Είναι αμετάβλητη ως προς το κατώφλι ταξινόμησης. Μετρά την ποιότητα των προβλέψεων του μοντέλου ανεξάρτητα από το κατώφλι ταξινόμησης που έχει επιλεγεί.

4.5 Δημιουργία Μοντέλου

Σκοπός αυτής της ενότητας είναι να παραθέσει μια λεπτομερή περιγραφή των βημάτων που ακολουθήθηκαν για τη δημιουργία του μοντέλου μηχανικής μάθησης και να εξηγήσει το σκεπτικό πίσω από κάθε απόφαση. Με τον τρόπο αυτό, επιδιώκουμε να δώσουμε στον αναγνώστη μια ολοκληρωμένη κατανόηση της διαδικασίας ανάπτυξης του μοντέλου και να παράσχουμε πληροφορίες που μπορεί να είναι χρήσιμες για μελλοντική έρευνα στον τομέα της μηχανικής μάθησης. Το παρακάτω διάγραμμα της εικόνας 4.4 σκιαγραφεί την δομή του μοντέλου και τα βήματα που ακολουθήθηκαν για την δημιουργία του.



Εικόνα 4.4: Μοντέλο Ανίχνευσης Ψευδοειδήσεων, Διάγραμμα Ροής

Βήμα 1

Συλλογή και Δημιουργία Συνόλων Δεδομένων

Στο αρχικό στάδιο της υλοποίησης του μοντέλου, απαιτείται η συλλογή και ενσωμάτωση των συνόλων δεδομένων που θα χρησιμοποιηθούν για την εκπαίδευσή του. Σύμφωνα με όσα αναφέρθηκαν στην ενότητα 4.3, τα δύο σύνολα δεδομένων που θα χρησιμοποιηθούν είναι το ISOT fake News και LIAR, τα οποία βρίσκονται σε μορφή .csv και είναι διαιρεμένα σε δύο αρχεία Excel το καθένα, ένα για τις αληθινές και ένα για τις ψευδείς ειδήσεις.

Για τη συγχώνευση των δύο συνόλων δεδομένων, χρησιμοποιήσαμε το Excel, συνδυάζοντας τα αρχεία με τις αληθινές και τις ψευδείς ειδήσεις αντίστοιχα. Με αυτόν τον τρόπο, καταλήξαμε στη δημιουργία δύο τελικών συνόλων δεδομένων, το ένα για τις αληθινές και το άλλο για τις ψευδείς ειδήσεις. Κάθε σύνολο αποτελείται από ένα αρχείο Excel, το οποίο περιλαμβάνει όλα τα στοιχεία των αντίστοιχων κατηγοριών.

Σημαντική προϋπόθεση κατά τη διαδικασία συγχώνευσης των δύο διαφορετικών συνόλων δεδομένων είναι η αντιστοίχιση των σωστών γραμμών και στηλών. Αυτό σημαίνει πως πρέπει να διασφαλίσουμε ότι οι στήλες που περιέχουν τα ίδια δεδομένα, όπως οι τίτλοι ή τα θέματα, έχουν την ίδια ονομασία και τοποθετούνται στις ίδιες θέσεις σε κάθε αρχείο. Επιπλέον, πρέπει να ελέγξουμε ότι οι γραμμές του ενός αρχείου αντιστοιχούν στις αντίστοιχες γραμμές του άλλου αρχείου. Αυτό εξασφαλίζει ότι τα δεδομένα είναι σωστά αντιστοιχισμένα και δεν θα προκαλέσουν σφάλματα στη διαδικασία εκπαίδευσης του μοντέλου.

Βήμα 2

Προεπεξεργασία δεδομένων

Έχοντας συλλέξει και ενσωματώσει τα σύνολα δεδομένων σε ένα τελικό, προχωράμε στην προεπεξεργασία των δεδομένων. Αρχικά εισάγουμε μία στήλη με όνομα “class” σε κάθε dataset. Για τις ψευδές ειδήσεις δίνουμε την τιμή «0» και για της αληθινές την τιμή «1».

```
dataset_fake["class"] = 0
dataset_true["class"] = 1
```

Εικόνα 4.5: Ορισμός κλάσης

Ο λόγος γι' αυτό είναι να μπορέσει το μοντέλο να μάθει να διακρίνει μεταξύ ψεύτικων και αληθινών ειδήσεων αναθέτοντάς τες στις αντίστοιχες κατηγορίες τους (0 ή 1, ψευδείς ή αληθινές) με βάση τα χαρακτηριστικά και τα μοτίβα που τις διαφοροποιούν.

Έπειτα, συγχωνεύουμε τα δύο σύνολα δεδομένων σε ένα ενιαίο που θα περιέχει τόσο τις αληθινές όσο και τις ψευδείς ειδήσεις, αντιστοιχίζοντας ταυτόχρονα τις στήλες μεταξύ τους.

```
dataset_merge = pd.concat([dataset_fake, dataset_true], axis =0 )
dataset_merge.head(47400)
```

	title	text	subject	date	class
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	31-Dec-17	0
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	31-Dec-17	0
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	30-Dec-17	0
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	29-Dec-17	0
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	25-Dec-17	0
...	...	...	...	...	...
18304	Mexico plans aid for Puerto Rico after Hurrica...	MEXICO CITY - Suffering itself after two maj...	worldnews	4-Oct-17	1
18305	Abandoned by tourists, Bali town counts cost o...	AMED, Indonesia - A Balinese town once bustl...	worldnews	2-Oct-17	1
18306	London's Angel station reopens after suspect p...	LONDON - British police said they carried ou...	worldnews	4-Oct-17	1
18307	London's Angel underground station closed due ...	- London s Angel underground station is clos...	worldnews	4-Oct-17	1
18308	UK police alerted to suspect package in London...	LONDON - British police were alerted to a su...	worldnews	4-Oct-17	1

47400 rows × 5 columns

```
dataset_merge.columns
```

```
Index(['title', 'text', 'subject', 'date', 'class'], dtype='object')
```

Εικόνα 4.6: Συγχώνευση των dataset

Η συγχώνευση των συνόλων βοηθά στην αποτροπή της μεροληψίας και βελτίωση της δυνατότητας γενίκευσης του μοντέλου. Επιπλέον παρέχει στο μοντέλο ένα πιο ισορροπημένο και αντιπροσωπευτικό δείγμα τόσο ψευδών όσο και αληθινών ειδήσεων, γεγονός που μπορεί να βελτιώσει την απόδοσή του.

Στη συνέχεια, αφαιρούμε τις στήλες «title», «subject» και «date», μένοντας μόνο με τις στήλες «class» και «text».

```
dataset = dataset_merge.drop(["title", "subject","date"], axis = 1)
```

Εικόνα 4.7: Αφαίρεση στηλών

Στην τρέχουσα περίπτωση, αυτές οι στήλες δεν είναι απαραίτητες για την κατηγοριοποίηση ψευδοειδήσεων, η διατήρησή τους μπορεί να προσθέσει περιττές πληροφορίες στο μοντέλο, αυξάνοντας το επίπεδο θορύβου (noise) και πιθανών οδηγώντας το σε υπερπροσαρμογή, κάτι που θα επηρεάσει άμεσα την ταξινόμηση των άρθρων.

Ακολούθως, κάνουμε ένα τυχαίο ανακάτεμα των άρθρων του συγχωνευμένου dataset.

```
dataset = dataset.sample(frac = 1)
```

```
dataset.head()
```

	text	class
6095	WASHINGTON - U.S. President Donald Trump's a...	1
364	MOSCOW - The Kremlin said on Monday Russian ...	1
14448	BEIRUT - Lebanon s foreign minister may not ...	1
18781	Is it simply pandering for votes, or is it pos...	0
14016	DUBLIN - The Irish government was on the ver...	1

Εικόνα 4.8: Τυχαίο ανακάτεμα άρθρων

Ανακατανέμοντας τυχαία τα δεδομένα αποφεύγουμε τυχόν μεροληψίες που μπορεί να προκύψουν κατά την διαδικασία της εκπαίδευσης. Εάν τα δεδομένα είναι διατεταγμένα με συγκεκριμένο τρόπο, όπως πριν το ανακάτεμα (βλ. Εικόνα 4.6, το πρώτο μισό του dataset είναι μόνο με τις ψευδείς, ενώ το δεύτερο μισό μόνο με τις αληθινές ειδήσεις), το μοντέλο μαθαίνει να κάνει προβλέψεις με βάση την σειρά αυτή, κάτι που θα αποτελέσει σε υπερπροσαρμογή. Με την τυχαία ανακατανομή των δεδομένων, εξασφαλίζουμε ότι το μοντέλο εκτίθεται σε όλα τα μέρη του συνόλου δεδομένων και μαθαίνει να κάνει ακριβείς προβλέψεις ανεξάρτητα από τη σειρά των παραδειγμάτων.

Βήμα 3

Μετασχηματισμός Δεδομένων

Τρίτο βήμα της δημιουργίας του μοντέλου αποτελεί ο μετασχηματισμός των δεδομένων. Αυτό επιτυγχάνεται με την δημιουργία μιας συνάρτησης που ονομάσαμε transform.

```
def transform(text):
    text = text.lower()
    text = re.sub('\[.*?\]', '', text)
    text = re.sub("\\W"," ",text)
    text = re.sub('https?://\S+|www\.\S+', '', text)
    text = re.sub('<.*?>+', '', text)
    text = re.sub('[%s]' % re.escape(string.punctuation), '', text)
    text = re.sub('\n', '', text)
    text = re.sub('\w*\d\w*', '', text)
    return text
```

```
dataset["text"] = dataset["text"].apply(transform)
```

Εικόνα 4.9: Μετασχηματισμός του κειμένου με τη συνάρτηση transform

Η συνάρτηση transform χρησιμοποιείται για τον μετασχηματισμό των δεδομένων του κειμένου αποτρέποντας το μοντέλο απ' το να μάθει περιττές συσχετίσεις μεταξύ των λέξεων με σκοπό την μείωση της υπερπροσαρμογής και την αύξηση της απόδοσής του. Συγκεκριμένα, ο κώδικας της παραπάνω Εικόνας 4.9 κάνει τα εξής:

- Μετατροπή όλων των κεφαλαίων γραμμάτων σε πεζά
- Αφαίρεση των ακγυλών και του περιεχομένου τους
- Αντικατάσταση όλων των μη λεκτικών χαρακτήρων με κενό
- Αφαίρεση των URL
- Αφαίρεση των ετικετών HTML
- Αφαίρεση των σημείων στίξης
- Αντικατάσταση όλων των "\n" (new line characters) με κενό
- Αφαίρεση όλων των αριθμών

Βήμα 4

Διαχωρισμός και παραμετροποίηση συνόλου δεδομένων

Έχοντας ολοκληρώσει την διαδικασία προεπεξεργασίας και μετασχηματισμού δεδομένων, πρέπει να χωρίσουμε το σύνολο δεδομένων σε δύο ξεχωριστά υποσύνολα, ένα για την εκπαίδευση του μοντέλου και ένα για την αξιολόγησή του. Πριν γίνει αυτό, πρέπει να αναθέσουμε μια εξαρτημένη και ανεξάρτητη μεταβλητή στις στήλες του συνόλου (βλ. Εικόνα 4.10).

```
x = dataset["text"]
y = dataset["class"]
```

Εικόνα 4.10: Ορισμός εξαρτημένης και ανεξάρτητης μεταβλητής

Ο παραπάνω κώδικας εξάγει τις στήλες text και class από το συγχωνευμένο σύνολο δεδομένων και τις αναθέτει στις μεταβλητές x και y, αντίστοιχα. Αυτό το βήμα είναι απαραίτητο πριν από το διαχωρισμό του συνόλου δεδομένων σε σύνολο εκπαίδευσης και δοκιμής, επειδή θέλουμε να χρησιμοποιήσουμε τη στήλη text ως χαρακτηριστικό εισόδου και τη στήλη class ως μεταβλητή-στόχο στο μοντέλο μηχανικής μάθησης.

Διαχωρίζοντας τα χαρακτηριστικά εισόδου (x) και τη μεταβλητή-στόχο (y) πριν από τη διάσπαση του συνόλου δεδομένων, εξασφαλίζουμε ότι τόσο το σύνολο εκπαίδευσης όσο και το σύνολο δοκιμής έχουν αντιπροσωπευτική κατανομή των διαφόρων κλάσεων στη μεταβλητή-στόχο. Αυτό είναι σημαντικό, διότι αν χωρίζαμε τυχαία το σύνολο δεδομένων χωρίς να διαχωρίσουμε τις μεταβλητές εισόδου και στόχου, μπορεί να καταλήξουμε σε ένα σύνολο εκπαίδευσης ή δοκιμής που είναι έντονα μεροληπτικό προς μία κλάση, γεγονός που μπορεί να επηρεάσει την ακρίβεια του μοντέλου μας.

Εν συνεχεία, διαχωρίζουμε το σύνολο δεδομένων σε δεδομένα εκπαίδευσης και δεδομένα δοκιμής (Εικόνα 4.11).

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.25)
```

Εικόνα 4.11: Διαχωρισμός συνόλου σε δεδομένα εκπαίδευσης και δοκιμής

Χωρίζουμε το σύνολο δεδομένων σε ένα σύνολο εκπαίδευσης και ένα σύνολο δοκιμής, προκειμένου να αξιολογήσουμε την απόδοση του μοντέλου μας. Το σύνολο train χρησιμοποιείται για την εκπαίδευση του μοντέλου και το σύνολο test χρησιμοποιείται για να αξιολογηθεί πόσο καλά το μοντέλο γενικεύεται σε νέα, αθέατα δεδομένα. Ο διαχωρισμός αυτός είναι ένα από τα πιο βασικά και σημαντικά βήματα στην δημιουργία ενός μοντέλου μηχανικής μάθησης. Αν δεν χωρίζαμε το σύνολο δεδομένων και αντί αυτού χρησιμοποιούσαμε όλα τα δεδομένα για την εκπαίδευση του μοντέλου, δεν θα μας έμεναν δεδομένα για να αξιολογήσουμε την απόδοσή του. Αυτό θα είχε ως αποτέλεσμα μια υπερβολικά αισιόδοξη αξιολόγηση της ακρίβειας του μοντέλου, καθώς αυτό θα είχε ήδη δει όλα τα δεδομένα. Στην συγκεκριμένη περίπτωση, τα δεδομένα χωρίστηκαν σε 75% για εκπαίδευση και 25% για αξιολόγηση.

Τέλος, μετατρέπουμε το κείμενο σε διανύσματα (Εικόνα 4.12).

```
[29] from sklearn.feature_extraction.text import TfidfVectorizer
0s

[30] vectorization = TfidfVectorizer()
14s  xv_train = vectorization.fit_transform(x_train)
     xv_test = vectorization.transform(x_test)
```

Εικόνα 4.12: Μετατροπή κειμένου σε διανύσματα

Η μετατροπή κειμένου σε αριθμητικά διανύσματα αποτελεί θεμελιώδες βήμα στα προβλήματα επεξεργασίας φυσικής γλώσσας (NLP). Οι αλγόριθμοι μηχανικής μάθησης δεν μπορούν να λειτουργήσουν απευθείας με δεδομένα κειμένου ως είσοδο, οπότε πρέπει να τα μετατρέψουμε σε αριθμητικά διανύσματα ώστε να μπορούν να εισαχθούν σε αυτούς. Για την μετατροπή αυτή, χρησιμοποιήσαμε την τεχνική TF-IDF, μέσω της συνάρτησης «TfidfVectorizer()».

Βήμα 5

Τελευταίο βήμα της υλοποίησης συνιστά την ενσωμάτωση των αλγορίθμων μηχανικής μάθησης για την εκπαίδευση του μοντέλου με τα μετασχηματισμένα δεδομένα, καθώς και την παρουσίαση των μετρικών αξιολόγησης με τα οπτικοποιημένα διαγράμματα Confusion Matrix και ROC-AUC curve που θα δούμε παρακάτω.

4.5 Ανάλυση Αποτελεσμάτων

Στην ενότητα αυτή θα παρουσιαστούν τα αποτελέσματα της εκπαίδευσης των διαφορετικών αλγορίθμων μηχανικής μάθησης. Η αξιολόγηση θα γίνει με βάση τις προαναφερόμενες μετρικές της ενότητας 4.4. Αφού γίνει η ανάλυση των αποτελεσμάτων, η αλγόριθμοι θα αξιολογηθούν με βάση της αποδοτικότητάς τους.

Στον παρακάτω πίνακα φαίνονται οι μετρικές Accuracy, Precision, Recall και F1 score για τον κάθε αλγόριθμο ξεχωριστά.

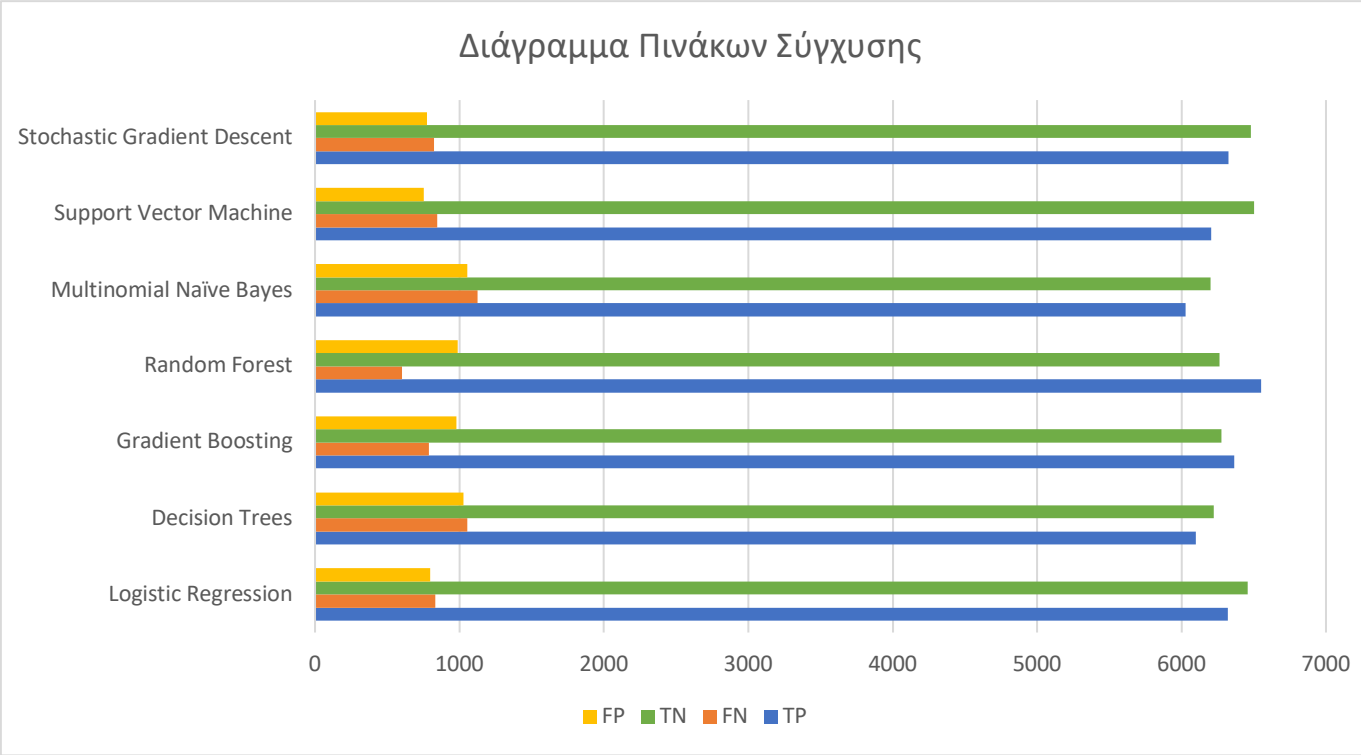
Αλγόριθμοι	**Accuracy**	**Precision**	**Recall**	**F1**
Logistic Regression	0.886	0.89	0.89	0.89
Decision Trees	0.856	0.86	0.86	0.86
Gradient Boosting	0.877	0.88	0.88	0.88
Random Forest	0.889	0.89	0.89	0.89
Multinomial Naïve Bayes	0.848	0.85	0.85	0.85
Support Vector Machine	0.889	0.89	0.89	0.89
Stochastic Gradient Descent	0.888	0.89	0.89	0.89

Πίνακας 4.1 Αποτελέσματα

Συνολικά, τα αποτελέσματα υποδεικνύουν ότι οι αλγόριθμοι έχουν καλή απόδοση στα σύνολα δεδομένων μας, με υψηλές τιμές για όλες τις μετρικές αξιολόγησης έχοντας συγκρίσιμη απόδοση μεταξύ τους.
Προκειμένου να γίνουν καλύτερα αντιληπτές οι διαφορές στην απόδοση των αλγορίθμων, συγκρίνουμε τα γραφήματα των Πινάκων Σύγχυσης και των καμπύλων ROC-AUC, όπως φαίνεται στον παρακάτω πίνακα.

Αλγόριθμοι	**Σωστά ταξινομημένα θετικά**	**Λανθασμένα ταξινομημένα αρνητικά**	**Σωστά ταξινομημένα αρνητικά**	**Λανθασμένα ταξινομημένα αρνητικά**	**AUC score**
Logistic Regression	6320	833	6457	798	0.89
Decision Trees	6100	1053	6225	1030	0.86
Gradient Boosting	6363	790	6276	979	0.88
Random Forest	6552	601	6265	990	0.89
Multinomial Naïve Bayes	6028	1125	6200	1055	0.85
Support Vector Machine	6206	847	6502	753	0.89
Stochastic Gradient Descent	6327	826	6478	777	0.89

Πίνακας 4.2 Αποτελέσματα Πίνακα Σύγχυσης



Πίνακας 4.3 Διάγραμμα Πίνακα Σύγχυσης

Συνολική Απόδοση: Όλοι οι αλγόριθμοι επιτυγχάνουν σχετικά υψηλές βαθμολογίες ακρίβειας (Accuracy), που κυμαίνονται από 0,848 έως 0,889. Αυτό δείχνει ότι η πλειονότητα των προβλέψεων που έγιναν είναι σωστές. Οι βαθμολογίες Precision, Recall και F1 παρουσιάζουν επίσης σταθερή απόδοση, με τιμές που κυμαίνονται από 0,85 έως 0,89, κάτι που είναι λογικό καθώς το σύνολο δεδομένων είναι ισορροπημένο, με ελάχιστη διαφοροποίηση μεταξύ των κλάσεών του. Οι βαθμολογίες AUC αποτελούν μέτρο της διαχωριστικής ικανότητας των αλγορίθμων. Η λογιστική παλινδρόμηση, το τυχαίο δάσος, ο Support Vector Machine και ο Gradient Boosting παρουσίασαν τις υψηλότερες βαθμολογίες AUC, που κυμαίνονται από 0,88 έως 0,89. Αυτό υποδηλώνει ότι οι συγκεκριμένοι αλγόριθμοι είναι πιο αποτελεσματικοί στη διάκριση μεταξύ ψεύτικων και πραγματικών ειδήσεων.

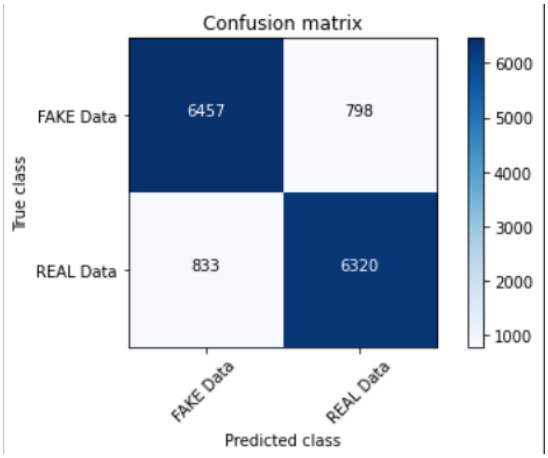
Λαμβάνοντας υπ' όψιν τις παραπάνω μετρικές, παρατηρούμε ότι εμφανές πλεονέκτημα παρουσιάζουν οι αλγόριθμοι Logistic Regression, Random Forest, Support Vector Machine και Stochastic Gradient Descent, όπου παρουσιάζουν μικρές διαφοροποιήσεις μεταξύ τους.

Logistic Regression: Η επιτυχία της λογιστικής παλινδρόμησης στην παρούσα μελέτη μπορεί να αποδοθεί στην ικανότητά της να μοντελοποιεί τη σχέση μεταξύ των χαρακτηριστικών (περιεχόμενο κειμένου) και της μεταβλητής-στόχου (ψεύτικες ή πραγματικές ειδήσεις) χρησιμοποιώντας μια λογιστική συνάρτηση. Πιο συγκεκριμένα:

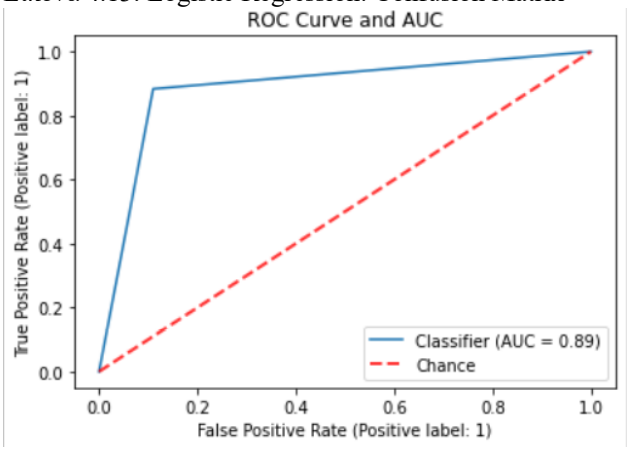
- Γραμμικότητα: Η λογιστική παλινδρόμηση υποθέτει γραμμική σχέση μεταξύ των χαρακτηριστικών και των λογαριθμικών αποδόσεων της μεταβλητής-στόχου. Παρόλο που αυτή η υπόθεση μπορεί να μην ισχύει ακριβώς σε δεδομένα κειμένου, μπορεί και πάλι να ανιχνεύσει σημαντικά μοτίβα στο χώρο

των χαρακτηριστικών, ειδικά όταν υπάρχουν ισχυρά διαχωριστικά χαρακτηριστικά που σχετίζονται με τις ψεύτικες και πραγματικές ειδήσεις.

- Αποτελεσματική εκπαίδευση: Η λογιστική παλινδρόμηση διαθέτει μια απλή και αποτελεσματική διαδικασία εκπαίδευσης, καθιστώντας την υπολογιστικά αποδοτική για μεγάλα σύνολα δεδομένων. Συγκλίνει γρήγορα και μπορεί να χειριστεί χώρους χαρακτηριστικών υψηλών διαστάσεων που συναντώνται συχνά σε εργασίες επεξεργασίας φυσικής γλώσσας.
- Πιθανολογική ερμηνεία: Η λογιστική παλινδρόμηση παρέχει πιθανολογικές εξόδους, δίνοντας την πιθανότητα ένα άρθρο ειδήσεων να ανήκει σε μια συγκεκριμένη κλάση. Η πιθανολογική ερμηνεία της λογιστικής παλινδρόμησης επιτρέπει μια πιο διαφοροποιημένη κατανόηση των προβλέψεων του μοντέλου, παρέχοντας εκτιμήσεις πιθανότητας μιας περίπτωσης να ανήκει σε μια συγκεκριμένη κλάση. Οι πληροφορίες αυτές μπορούν να καθοδηγήσουν τη λήψη αποφάσεων και να επιτρέψουν πιο τεκμηριωμένες ενέργειες στο πλαίσιο της ανίχνευσης ψευδών ειδήσεων.



Εικόνα 4.13: Logistic Regression: Confusion Matrix



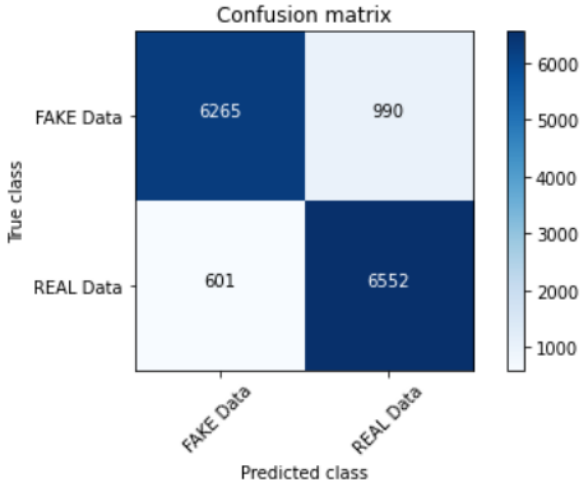
Εικόνα 4.14: Logistic Regression: ROC Curve και AUC

Random Forest: Το τυχαίο δάσος επέδειξε ισχυρές επιδόσεις στην ανίχνευση ψευδών ειδήσεων λόγω των εγγενών χαρακτηριστικών του:

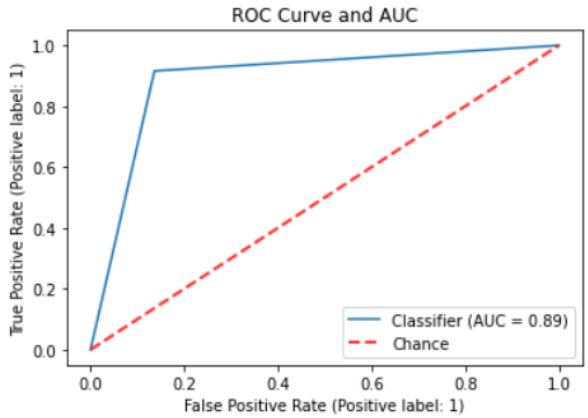
- Ensemble Learning: Το τυχαίο δάσος συνδυάζει πολλαπλά δέντρα απόφασης για να σχηματίσει ένα ισχυρό μοντέλο συνόλου. Κάθε δέντρο μαθαίνει από

ένα τυχαίο υποσύνολο χαρακτηριστικών και περιπτώσεων εκπαίδευσης, μειώνοντας την υπερπροσαρμογή και βελτιώνοντας τη γενίκευση.

- Σημασία χαρακτηριστικών: Το τυχαίο δάσος υπολογίζει τη σημασία των χαρακτηριστικών με βάση τη μείωση του δείκτη Gini κατά την κατασκευή των δέντρων. Αυτό επιτρέπει στον αλγόριθμο να εντοπίζει τα πιο κατατοπιστικά χαρακτηριστικά για τη διαφοροποίηση μεταξύ ψεύτικων και πραγματικών ειδήσεων, οδηγώντας σε ακριβή ταξινόμηση.
- Μη γραμμικές σχέσεις: Το τυχαίο δάσος μπορεί να συλλάβει μη γραμμικές σχέσεις μεταξύ των χαρακτηριστικών και της μεταβλητής-στόχου. Στο πλαίσιο της επεξεργασίας φυσικής γλώσσας, όπου οι σχέσεις μεταξύ των λέξεων και του νοήματός τους μπορεί να είναι πολύπλοκες, η ικανότητα μη γραμμικής μοντελοποίησης του τυχαίου δάσους μπορεί να τις συλλάβει αποτελεσματικά.



Εικόνα 4.15: Random Forest: Confusion Matrix

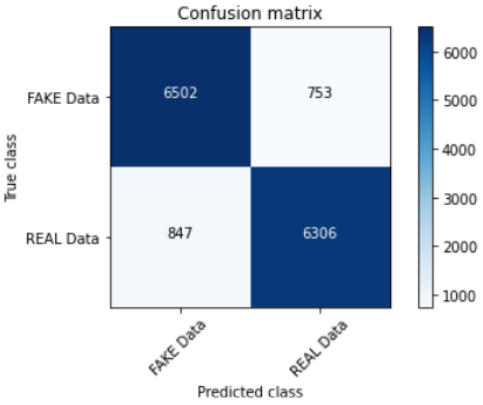


Εικόνα 4.16: Random Forest: ROC Curve και AUC

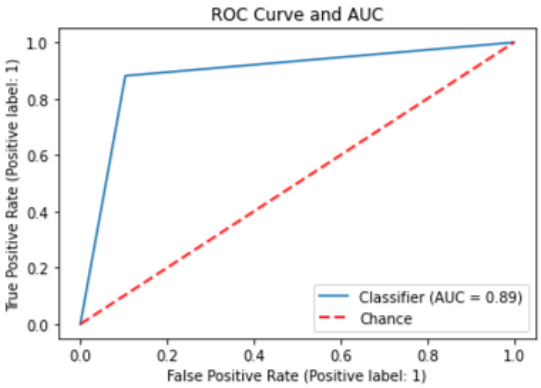
Support Vector Machine: Ο SVM είχε καλή απόδοση στην ικανότητα ανίχνευσης ψευδών ειδήσεων λόγω των μοναδικών χαρακτηριστικών του:

- Μεγιστοποίηση περιθωρίου: Ο SVM στοχεύει στην εύρεση ενός βέλτιστου ορίου απόφασης που μεγιστοποιεί το περιθώριο μεταξύ των διαφόρων κλάσεων. Με τη μεγιστοποίηση του περιθωρίου, ο SVM επιτυγχάνει καλύτερο διαχωρισμό μεταξύ των περιπτώσεων ψεύτικων και πραγματικών ειδήσεων, οδηγώντας σε βελτιωμένη απόδοση ταξινόμησης.

- Kernel συναρτήσεις: Ο SVM μπορεί να χρησιμοποιήσει συναρτήσεις Kernel για να μετατρέψει τα δεδομένα σε χώρους χαρακτηριστικών υψηλότερων διαστάσεων, όπου οι κλάσεις γίνονται πιο διαχωρίσιμες. Αυτό επιτρέπει στον SVM να συλλάβει πολύπλοκες σχέσεις και να βελτιώσει την ικανότητά του να διακρίνει μεταξύ διαφορετικών τύπων ειδησεογραφικών άρθρων.
- Κανονικοποίηση: Ο SVM ενσωματώνει τεχνικές κανονικοποίησης για τον έλεγχο της πολυπλοκότητας του μοντέλου και την αποφυγή της υπερπροσαρμογής. Αυτή η κανονικοποίηση βοηθά το μοντέλο να γενικεύει καλά σε αθέατα δεδομένα, οδηγώντας σε βελτιωμένη απόδοση.



Εικόνα 4.17: Support Vector Machine: Confusion Matrix

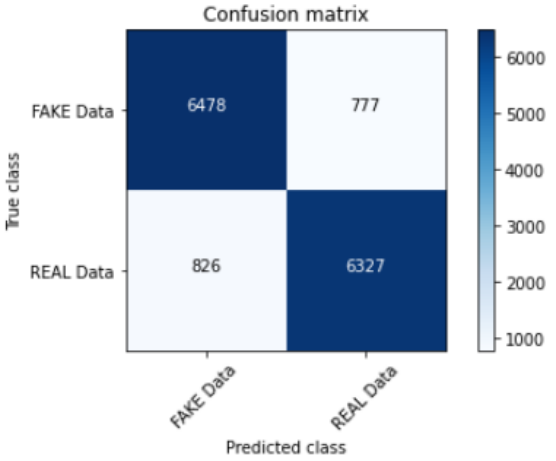


Εικόνα 4.18: Support Vector Machine: ROC Curve και AUC

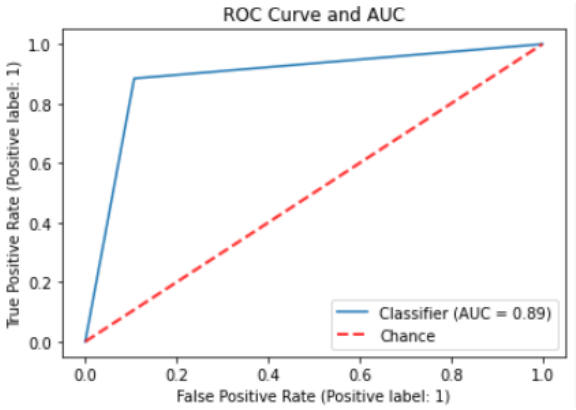
Stochastic Gradient Descent: Η αποτελεσματικότητά του στην ανίχνευση ψευδών ειδήσεων μπορεί να αποδοθεί στους ακόλουθους παράγοντες:

- Αποδοτικότητα: Ο SGD ενημερώνει τις παραμέτρους του μοντέλου με βάση μεμονωμένες περιπτώσεις εκπαίδευσης, καθιστώντας τον υπολογιστικά αποδοτικό, ιδίως για μεγάλα σύνολα δεδομένων.
- Ταχύτητα σύγκλισης: Η επαναληπτική φύση του SGD του επιτρέπει να συγκλίνει γρήγορα σε μια καλή λύση. Προσαρμόζεται καλά στις μεταβαλλόμενες κλίσεις, καθιστώντας τον κατάλληλο για σενάρια κατηγοριοποίησης ειδήσεων όπου νέα δεδομένα καταφθάνουν συνεχώς.

- Επιλογές κανονικοποίησης: Ο SGD παρέχει διάφορες τεχνικές κανονικοποίησης, όπως η κανονικοποίηση L1 ή L2, για τον έλεγχο της πολυπλοκότητας του μοντέλου και την αποφυγή της υπερπροσαρμογής. Αυτές οι τεχνικές κανονικοποίησης συμβάλλουν στη βελτίωση της ικανότητας του μοντέλου να γενικεύει και να κάνει ακριβείς προβλέψεις σε αθέατα δεδομένα. Στην περίπτωσή μας, χρησιμοποιήσαμε την L2. Η κανονικοποίηση L2, επίσης γνωστή ως κανονικοποίηση Ridge, προσθέτει έναν όρο ποινής στη συνάρτηση απώλειας του αλγορίθμου SGD που ενθαρρύνει μικρότερα βάρη στο μοντέλο. Αυτός ο όρος ποινής είναι ανάλογος του τετραγώνου του μεγέθους των βαρών. Στην περίπτωσή μας, η συγκεκριμένη κανονικοποίηση ήταν πιο αποδοτική από την L1 με βάση τις δοκιμές που πραγματοποιήθηκαν.



Εικόνα 4.19: Stochastic Gradient Descent: Confusion Matrix



Εικόνα 4.20: Stochastic Gradient Descent: ROC Curve και AUC

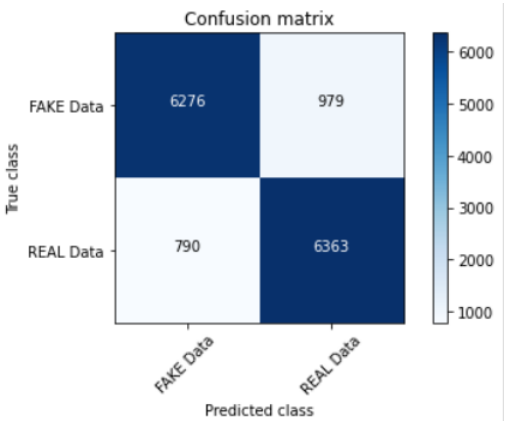
Gradient Boosting: Ο συγκεκριμένος αλγόριθμος παρουσίασε αρκετά καλά αποτελέσματα, αλλά λίγο χαμηλότερα από τους Logistic Regression, Random Forest, Support Vector Machine και Stochastic Gradient Descent. Οι λόγοι για αυτό δεν είναι ιδιαίτερα προφανείς.

Ο Gradient Boosting χρησιμοποιεί την μέθοδο Ensemble, συνδυάζοντας πολλαπλούς αδύναμους μαθητές, όπως ο Random Forest. Επιπλέον, χειρίζεται τόσο γραμμικές όσο και μη γραμμικές σχέσεις, παρόμοια με τους προαναφερόμενους αλγορίθμους. Επιπρόσθετα, χρησιμοποιεί προσαρμοστική μάθηση, που σημαίνει ότι αναθέτει διαφορετικά βάρη σε κάθε περίπτωση εκπαίδευσης με βάση τα σφάλματα που έγιναν

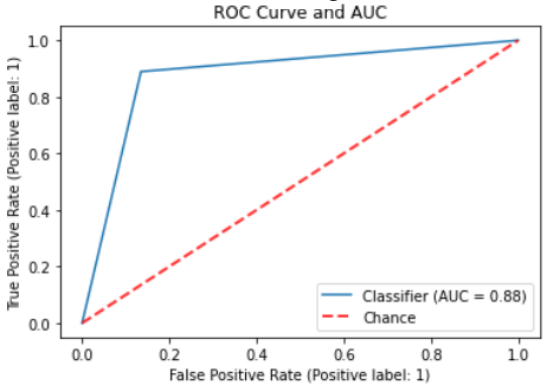
από προηγούμενα μοντέλα. Αυτό επιτρέπει στον αλγόριθμο να δίνει προτεραιότητα στα δείγματα που είναι πιο δύσκολο να ταξινομηθούν, βελτιώνοντας την ικανότητά του να αντιλαμβάνεται σύνθετα μοτίβα και να κάνει ακριβείς προβλέψεις.

Κάποιοι από τους λόγους που τον κάνουν να υστερεί ελαφρά ίσως είναι:

- Ευαισθησία σε υπερπαραμέτρους: Είναι πιθανό οι υπερπαράμετροι που επιλέχθηκαν για το Gradient Boosting να μην βελτιστοποιήθηκαν αποτελεσματικά, με αποτέλεσμα την ελαφρώς χαμηλότερη απόδοσή του σε σύγκριση με άλλους καλά συντονισμένους αλγορίθμους.
- Χαρακτηριστικά δεδομένων: Ο Gradient Boosting ίσως να είναι ευαίσθητος σε συγκεκριμένες πτυχές των δεδομένων που άλλοι αλγόριθμοι μπορούν να χειριστούν καλύτερα. Εάν το σύνολο δεδομένων περιέχει σημαντική ποσότητα θορύβου ή άσχετων χαρακτηριστικών, μπορεί να επηρεάσει την απόδοση του Gradient Boosting. Ο Gradient Boosting είναι ευαίσθητος στα θορυβώδη δεδομένα, κάτι που δεν τον βοηθάει στο συγκεκριμένο σύνολο, καθώς υπάρχουν πολλά αχρείαστα δεδομένα. Εάν υπάρχει θόρυβος στα δεδομένα εκπαίδευσης, μπορεί να διαδοθεί και να επηρεάσει αρνητικά την απόδοση του συνόλου.



Εικόνα 4.21: Gradient Boosting: Confusion Matrix



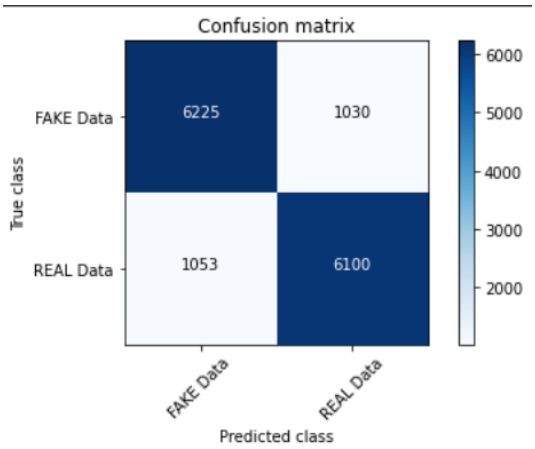
Εικόνα 4.22: Gradient Boosting: ROC Curve και AUC

Μειωμένη απόδοση παρατηρείται στους Αλγόριθμους Decision Trees και Multinomial Naïve Bayes, με μέσο Accuracy, Precision, Recall και F1 0.86 και 0.85, καθώς και AUC 0.86 και 0.85 αντίστοιχα. Τα χειρότερα αποτελέσματα των δύο αυτών αλγορίθμων ήταν αναμενόμενα για τους λόγους που θα εξηγήσουμε παρακάτω.

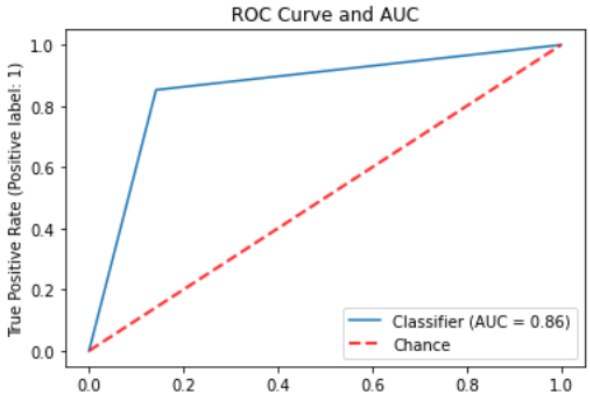
Δέντρα Απόφασης

Τα Δέντρα Απόφασης δεν είχαν τόσο καλή απόδοση όσο τους άλλους αλγορίθμους εξ' αιτίας των έμφυτων χαρακτηριστικών και περιορισμών της δομής τους, συγκεκριμένα:

- Χειρισμός πολυπλοκότητας: Τα Δέντρα Αποφάσεων είναι επιρρεπή στην υπερπροσαρμογή, ιδίως όταν πρόκειται για πολύπλοκα σύνολα δεδομένων. Εάν το σύνολο δεδομένων περιέχει περίπλοκες σχέσεις και εξαρτήσεις μεταξύ των χαρακτηριστικών, τα Δέντρα Αποφάσεων ενδέχεται να δυσκολεύονται να τις αποτυπώσουν με ακρίβεια. Αυτό μπορεί να οδηγήσει σε χαμηλότερες επιδόσεις γενίκευσης σε αθέατα δεδομένα.
- Σημασία χαρακτηριστικών: Τα Δέντρα Αποφάσεων βασίζονται σε μεγάλο βαθμό στη σημασία των μεμονωμένων χαρακτηριστικών κατά την κατασκευή του δέντρου. Εάν το σύνολο δεδομένων περιέχει άσχετα ή θορυβώδη χαρακτηριστικά (όπως το συγκεκριμένο) που δεν συμβάλλουν σημαντικά στη μεταβλητή-στόχο το Δέντρο Αποφάσεων μπορεί να γίνει μεροληπτικό ή λιγότερο αποτελεσματικό στην πραγματοποίηση ακριβών προβλέψεων.
- Περιορισμένη χωρητικότητα μοντέλου: Τα Δέντρα Αποφάσεων έχουν περιορισμένη ικανότητα να συλλαμβάνουν σύνθετα μοτίβα και σχέσεις σε σύγκριση με πιο εξελιγμένους αλγορίθμους, όπως τον Random Forest. Συνεπώς, ενδέχεται να δυσκολεύονται να χειριστούν δεδομένα υψηλών διαστάσεων ή περίπλοκα όρια αποφάσεων, με αποτέλεσμα χαμηλότερη απόδοση.
- Σύνολο Δεδομένων: Κάθε αλγόριθμος αποδίδει διαφορετικά ανάλογα με τα χαρακτηριστικά του συνόλου δεδομένων. Είναι πιθανό το συγκεκριμένο σύνολο δεδομένων να είχε ορισμένες ιδιότητες ή προκαταλήψεις που δεν ευθυγραμμίζονταν καλά με τα Δέντρα Αποφάσεων όπως ο θόρυβος, η ασάφεια των χαρακτηριστικών, η έλλειψη αριθμητικών στοιχείων και τα αραιά δεδομένα, κάνοντας δύσκολο τον εντοπισμό σημαντικών χαρακτηριστικών για αυτόν τον αλγόριθμο.
- Έλλειψη σφαιρικής βελτιστοποίησης: Τα δέντρα αποφάσεων χρησιμοποιούν έναν άπληστο αλγόριθμο που βελτιστοποιεί τη διάσπαση τοπικά σε κάθε κόμβο, χωρίς να λαμβάνει υπόψη τη συνολική δομή των δεδομένων. Τα σύνολα δεδομένων φυσικής γλώσσας παρουσιάζουν συχνά πολύπλοκες και λεπτές σχέσεις που απαιτούν σφαιρική κατανόηση. Τα δέντρα αποφάσεων ενδέχεται να δυσκολεύονται να συλλάβουν αυτές τις περίπλοκες εξαρτήσεις, με αποτέλεσμα να έχουν υποβέλτιστες επιδόσεις.



Εικόνα 4.23: Decision Trees: Confusion Matrix



Εικόνα 4.24: Decision Trees: ROC curve και AUC

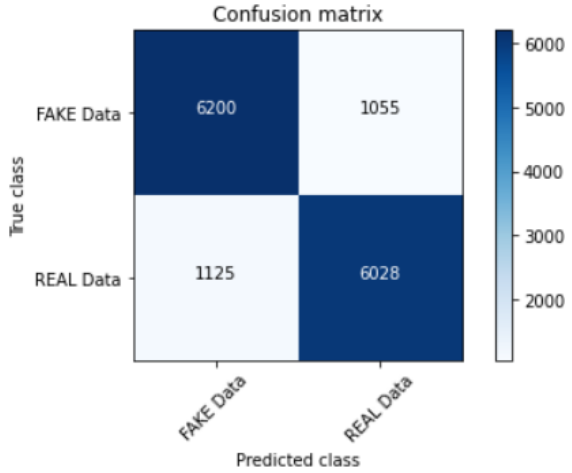
Multinomial Naïve Bayes

Ο Multinomial Naïve Bayes είχε επίσης συγκριτικά φτωχή απόδοση σε σχέση με τους υπόλοιπους αλγορίθμους. Οι πιθανοί λόγοι γι' αυτό είναι οι εξής:

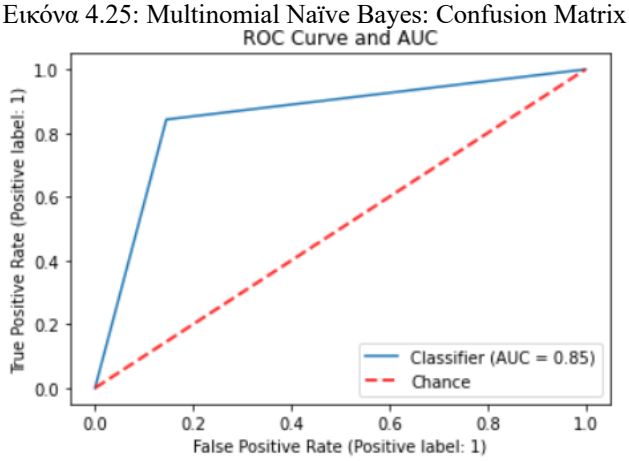
- Πολυωνυμική κατανομή: Ο MNB υποθέτει ότι οι πιθανότητες των χαρακτηριστικών ακολουθούν μια πολυωνυμική κατανομή, η οποία είναι πιο κατάλληλη για προβλήματα πολλαπλών κατηγοριών. Σε ένα σενάριο δυαδικής ταξινόμησης, όπου υπάρχουν μόνο δύο κλάσεις (ψεύτικες και πραγματικές ειδήσεις), η υποκείμενη κατανομή των χαρακτηριστικών δεν ευθυγραμμίζεται καλά με την πολυωνυμική υπόθεση. Αυτή η ασυμφωνία μπορεί να οδηγήσει σε υποβέλτιστες επιδόσεις.
- Αγνόηση των αλληλεπιδράσεων μεταξύ των χαρακτηριστικών: Ο MNB υποθέτει ότι τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους με βάση την ετικέτα της κλάσης. Αυτή η υπόθεση είναι γνωστή ως "bag of words", όπου η σειρά και οι αλληλεπιδράσεις μεταξύ των λέξεων δεν λαμβάνονται υπόψη. Ωστόσο, στην φυσική επεξεργασία γλώσσας, οι σχέσεις πλαισίου μεταξύ των λέξεων και οι αλληλεξαρτήσεις τους διαδραματίζουν κρίσιμο ρόλο στον

καθορισμό της κλάσης ενός εγγράφου. Αγνοώντας αυτές τις αλληλεπιδράσεις, ο MNB μπορεί να αποτύχει να συλλάβει τα διαφοροποιημένα μοτίβα και τις εξαρτήσεις που υπάρχουν στα δεδομένα κειμένου, οδηγώντας τον σε χαμηλότερες επιδόσεις.

- Έλλειψη συνεχών εκτιμήσεων πιθανοτήτων: Ο MNB παρέχει ξεχωριστές εκτιμήσεις πιθανότητας για κάθε κλάση με βάση την εμφάνιση των χαρακτηριστικών της. Στη δυαδική ταξινόμηση, η ύπαρξη συνεχών εκτιμήσεων πιθανότητας είναι ζωτική για τη λήψη αποφάσεων και την επιλογή κατωφλίου. Οι διακριτές εκτιμήσεις πιθανότητας του MNB ενδέχεται να περιορίζουν την ικανότητά του να παρέχει καλά βαθμονομημένες και λεπτομερείς προβλέψεις, επηρεάζοντας ενδεχομένως την απόδοσή του.



Εικόνα 4.25: Multinomial Naïve Bayes: Confusion Matrix



Εικόνα 4.26: Multinomial Naïve Bayes: ROC Curve και AUC

4.7 Συμπεράσματα

Μεταξύ των αλγορίθμων που αξιολογήθηκαν στη μελέτη μας, οι αλγόριθμοι Logistic Regression, Random Forest, Support Vector Machine και Stochastic Gradient Descent αναδείχθηκαν ως οι αλγόριθμοι με τις καλύτερες επιδόσεις για την ταξινόμηση ψευδών ειδήσεων. Αυτοί οι αλγόριθμοι επέδειξαν υψηλότερη ακρίβεια, βαθμολογίες AUC και πέτυχαν μεγαλύτερο αριθμό σωστά ταξινομημένων περιπτώσεων σε σύγκριση με τα άλλα μοντέλα.

Αξίζει να σημειωθεί, ότι παρόλο που οι αναφερόμενοι αλγόριθμοι έχουν επιτύχει σημαντικά αποτελέσματα, η ταξινόμηση ψευδοειδήσεων είναι ένα πρόβλημα που παρουσιάζει πολλές προκλήσεις και η απόδοση των αλγορίθμων μπορεί να διαφέρει ανάλογα με το επιλεγμένο σύνολο δεδομένων και τα χαρακτηριστικά που χρησιμοποιούνται.

Για το λόγο αυτό, προτείνεται η διερεύνηση προηγμένων μεθόδων συνόλου και αρχιτεκτονικών βαθιάς μάθησης για την περαιτέρω ενίσχυση της απόδοσης της ταξινόμησης ψευδών ειδήσεων. Η συνεχιζόμενη έρευνα σε αυτόν τον τομέα μπορεί να οδηγήσει στην ανάπτυξη αποτελεσματικότερων μοντέλων μηχανικής μάθησης που θα βοηθήσουν στην καταπολέμηση της εξάπλωσης της παραπληροφόρησης και της διάδοσης ψευδών ειδήσεων.

Συνοψίζοντας, η παρούσα μελέτη προσφέρει πολύτιμες πληροφορίες για τον τρόπο με τον οποίο οι αλγόριθμοι μηχανικής μάθησης μπορούν να χρησιμοποιηθούν για την αναγνώριση και ταξινόμηση των ψευδών ειδήσεων. Τα ευρήματα αυτής της μελέτης αποτελούν βάση για μελλοντικές έρευνες και αναπτύξεις σε αυτόν τον σημαντικό τομέα, με στόχο την ανάπτυξη αποτελεσματικών εργαλείων και τεχνικών για την αντιμετώπιση των ψευδών ειδήσεων και τη διατήρηση της αξιοπιστίας της πληροφορίας.

Κεφάλαιο 5
Προβληματισμοί και περιορισμοί

Η παρούσα διπλωματική εργασία αποσκοπεί στο να ρίξει φως σε ένα ευρύ και αρκετά καινούργιο πρόβλημα που πλήττει την ψηφιακή εποχή στην οποία ζούμε. Οι τρόποι προσέγγισής του από ερευνητές είναι πολυάριθμοι, καθώς υπάρχει πληθώρα επιλογών τόσο με τεχνικές μηχανικής μάθησης και επεξεργασίας φυσικής γλώσσας όσο και εξόρυξης δεδομένων και web scalping.

Ένα από τα προβλήματα που προέκυψαν κατά τη δημιουργία του μοντέλου ήταν η σωστή διαμόρφωση του συνόλου δεδομένων ώστε να μπορέσει να διασφαλιστεί η σωστή χρήση του από τους αλγορίθμους αλλά και να βελτιστοποιηθεί η απόδοσή τους. Επιπλέον, η υπερπροσαρμογή αποτέλεσε πρόβλημα από τις πρώτες κιόλας δοκιμές του μοντέλου, η οποία μειώθηκε σημαντικά με την περαιτέρω επέκταση του συνόλου δεδομένων, ενσωματώνοντας το LIAR dataset, καθώς και με την κατάλληλη προεπεξεργασία και μετασχηματισμό των δεδομένων.

Επιπρόσθετα, η βελτιστοποίηση των αλγοριθμικών υπερπαραμέτρων απαίτησε προσεκτικό χειρισμό και εξαντλητική δοκιμή διαφόρων συνδυασμών. Η κατάλληλη επιλογή των υπερπαραμέτρων είναι κρίσιμη για την απόδοση των αλγορίθμων, είναι μία χρονοβόρα διαδικασία και ιδιαίτερα επιρρεπής σε ανθρώπινα σφάλματα.

Κεφάλαιο 6
Βελτιώσεις και επεκτάσεις

Το μοντέλο που δημιουργήσαμε στην παρούσα ερευνητική είναι δυαδικής ταξινόμησης. Ωστόσο, όπως και ο κόσμος δεν είναι άσπρος και μαύρος, έτσι και οι πληροφορίες στο διαδίκτυο δεν είναι πάντα είτε αληθείς είτε ψευδείς. Ένα πιο σύνθετο μοντέλο μηχανικής μάθησης θα μπορούσε να κάνει πολυωνυμική ταξινόμηση, δίνοντας ως εξόδους παραπάνω από δύο αποτελέσματα. Κάποια είδηση μπορεί να περιέχει παραπλανητικό τίτλο αλλά οι πληροφορίες του άρθρου να είναι κατά βάση σωστές, ενώ κάποια άλλη είδηση μπορεί να περιέχει μερικές ανακρίβειες. Μία τέτοια πολυωνυμική ταξινόμηση θα μπορούσε να χρησιμοποιεί τελείως διαφορετικούς αλγορίθμους, καταλληλότερους για κατηγοριοποίηση τέτοιου είδους. Ο multinomial Naïve Bayes, που είναι δομημένος για αντίστοιχες πολυωνυμικές ταξινομήσεις, πιθανότατα να μας έδινε πιο υποσχόμενα αποτελέσματα σε ένα τέτοιο σενάριο.

Η περαιτέρω επέκταση του συνόλου δεδομένων, εάν και απαιτεί αρκετά παραπάνω πόρους και αυξάνει τον χρόνο εκπαίδευσης του μοντέλου, παραδίδει ακόμα πιο ακριβή αποτελέσματα και αυξάνει εκθετικά την απόδοση των αλγορίθμων, με βάση την διαθέσιμη βιβλιογραφία και μελέτες πάνω στο κομμάτι της κατηγοριοποίησης. Στη μηχανική μάθηση, η ύπαρξη περισσότερων δεδομένων μπορεί να βοηθήσει στην καταγραφή ενός ευρύτερου φάσματος μοτίβων και παραλλαγών που υπάρχουν στον πραγματικό κόσμο, οδηγώντας σε καλύτερη γενίκευση και βελτιωμένη απόδοση.

Ιστορικά, θεωρούνταν ότι για την επίτευξη μεγαλύτερης απόδοσης απαιτούνταν μεγάλη πρόοδος στις επιδόσεις των αλγορίθμων. Πρόσφατες ανακαλύψεις, όπως το GPT (Generative Pre-trained Transformer) του OpenAI, έχουν θέσει υπό αμφισβήτηση αυτή την υπόθεση. Τα μοντέλα σαν το GPT προ-εκπαιδεύονται σε τεράστιους όγκους δεδομένων κειμένου από το διαδίκτυο, χρησιμοποιώντας ένα ποικίλο σύνολο γλωσσικών προτύπων και σημασιολογικών συνδέσεων, επιτρέποντάς τους να παράγουν πιο συναφή, συνεκτικά και ακριβές αποτελέσματα.

Βιβλιογραφία

[1] D. A. Kalogeropoulos, "Digital News Report," 2021. [Online]. Available: https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021/greece.

[2] Tom Mitchell, McGraw Hill, "Machine Learning," 1997. [Online]. Available: http://www.cs.cmu.edu/~tom/mlbook.html.

[3] V. Kanade, "What is Machine Learning," 2022. [Online]. Available: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-ml/.

[4] M. Wolfe, "Machine Learning Methods," 2021. [Online]. Available: https://towardsdatascience.com/three-popular-machine-learning-methods-7cb2dcb40bd0.

[5] Diksha Khurana, Aditya Koli, Kiran Khatter, Sukhdev Singh, "Natural Language Processing: State of The Art, Current Trends and Challenges," Manav Rachna University, India, 2022. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/35855771/.

[6] D. M. J. Adams, "System44," 2008. [Online]. Available: https://www.hmhco.com/research/system-44-research-evidence-base.

[7] M. H. Duhman, "Data Mining, Introductory and advanced topics,," 2002. [Online]. Available: https://theswissbay.ch/pdf/Gentoomen%20Library/Data%20Mining/Dunham%20-%20Data%20Mining.pdf.

[8] J. Peralta, "Text preprocessing," 2022. [Online]. Available: https://www.geeksforgeeks.org/text-preprocessing-in-python-set-1/.

[9] "Removal of Punctuation," 2021. [Online]. Available: https://datagy.io/python-remove-punctuation-from-string/.

[10] U. Malik, "Removing Stop Words," 2020. [Online]. Available: https://stackabuse.com/removing-stop-words-from-strings-in-python/.

[11] G. L. Team, "Bag of Words Method," 2022. [Online]. Available: https://www.mygreatlearning.com/blog/bag-of-words/.

[12] J. D. Ullman, "Data Mining," 2011. [Online]. Available: http://i.stanford.edu/~ullman/mmds/ch1.pdf.

[13] "Logistic Regression," [Online]. Available: https://www.sciencedirect.com/topics/computer-science/logistic-regression.

[14] C. M. Bishop, "Pattern Recognition and Machine learning," 2006. [Online]. Available: http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf.

[15] N. S. Chauhan, "Decision Trees," 2022. [Online]. Available: https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html.

[16] J. Brownlee, "Gradient Boosting," 2016. [Online]. Available: https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/.

[17] L. Breiman, "Random Forest," 2001. [Online]. Available: https://link.springer.com/article/10.1023/a:1010933404324.

[18] A. Sharma, "Random Forest vs Decision Tree," 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/ .

[19] G. L. Team, "Naïve Bayes," 2022. [Online]. Available: https://www.mygreatlearning.com/blog/multinomial-naive-bayes-explained/.

[20] Sriram, "Multinomial Naïve Bayes," 2022. [Online]. Available: https://www.upgrad.com/blog/multinomial-naive-bayes-explained/.

[21] C. Cortes, V. Vapnik, and L. Saitta, "Support-Vector Networks Editor," Kluwer Academic Publishers, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050921012035.

[22] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," 2019. [Online]. Available: https://www.scribd.com/document/498832298/A-comprehensive-survey-on-support-vector-machine-classification-Applications-challenges-and-trends.

[23] "Introduction to Support Vector Machines," 2023. [Online]. Available: https://www.geeksforgeeks.org/introduction-to-support-vector-machines-svm/.

[24] "Support Vector Machine Algorithm," [Online]. Available: https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm.

[25] Κ. Κωνσταντίνος, "ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ," 2020. [Online]. Available: http://ikee.lib.auth.gr/record/323648/files/thesis.pdf.

[26] Y. Tian, "Recent Advances in Stochastic Gradient Descent in Deep Learning," 2023. [Online]. Available: https://www.mdpi.com/2227-7390/11/3/682.

[27] "Gradient Descent in Neural Network," [Online]. Available: https://studymachinelearning.com/optimization-algorithms-in-neural-network/.

[28] "Python," [Online]. Available: https://www.python.org/.

[29] "Google Collab," [Online]. Available: https://research.google.com/colaboratory/faq.html.

[30] IFCN, "International Fact-Checking Network," [Online]. Available: https://ifcncodeofprinciples.poynter.org.

[31] A. D. Holan, "PolitiFact," 2018. [Online]. Available: https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/.

[32] "ISOT FAKE NEWS dataset," [Online]. Available: https://onlineacademiccommunity.uvic.ca/isot/2022/11/27/fake-news-detection-datasets/.

[33] "LIAR dataset," [Online]. Available: https://paperswithcode.com/dataset/liar.

[34] P. Huilgol, "Accuracy vs. F1-Score," [Online]. Available: https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2.

[35] A. Ragan, "Confusion Matrix," 2018. [Online]. Available: https://towardsdatascience.com/taking-the-confusion-out-of-confusion-matrices-c1ce054b3d3e.

[36] S. K. Agrawal, "Metrics to Evaluate your Classification Model to take the right decisions," 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/.

[37] S. Narkhede, "Understanding AUC - ROC Curve," [Online]. Available: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.

[38] "ROC Curve and AUC," [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.