ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΗΛΕΚΤΡΟΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μέθοδοι Μηχανικής Μάθησης στον χώρο της Ιατρικής Φροντίδας (Healthcare)

Διπλωματικη Εργασια

του

ΔΗΜΗΤΡΙΟΥ ΚΑΡΑΜΑΝΗ

Επιβλέποντες: Χρήστος Μακρής

Πατρα, Ιούνιος 2023

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΗΛΕΚΤΡΟΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μέθοδοι Μηχανικής Μάθησης στον χώρο της Ιατρικής Φροντίδας (Healthcare)

Διπλωματικη Εργασια

του

ΔΗΜΗΤΡΙΟΥ ΚΑΡΑΜΑΝΗ

Επιβλέποντες: Χρήστος Μακρής

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την χχ Ιουνίου 2023.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Χρήστος Μακρής	Σπύρος Σιούτας	Δημήτριος Τσώλης
Αναπληρωτής Καθηγητής	Καθηγητής	Επίκουρος Καθηγητής

Πατρα, Ιούνιος 2023

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΗΛΕΚΤΡΟΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

Υπεύθυνη Δήλωση

Βεβαιώνω ότι είμαι συγγραφέας αυτής της πτυχιακής εργασίας, και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην πτυχιακή 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 πτυχιακή εργασία προετοιμάστηκε από εμένα προσωπικά ειδικά για τις απαιτήσεις του προγράμματος σπουδών του τμήματος Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής του Πανεπιστημίου Πατρών.

(Υπογραφή)

.....
Δημήτριος Καραμάνης

Περίληψη

Η εκπαίδευση ενός υπολογιστικού συστήματος στην νοηματική αναγνώριση κειμένου και την εξαγωγή συμπερασμάτων από αυτό, αποτελεί μια πρόκληση για την επιστημονική κοινότητα και την έρευνα.

Στόχος του συγκεκριμένου ερευνητικού πεδίου που δημιουργήθηκε από αυτήν την πρόκληση είναι η δημιουργία κατάλληλων εργαλείων για την υποστήριξη, υποβοήθηση και διευκόλυνση του ανθρώπινου δυναμικού.

Στην παρούσα διπλωματική εργασία δίνεται έμφαση στην χρήση ιατρικού κειμένου ως πηγή δεδομένων, στη σωστή προεπεξεργασία και στην κατηγοριοποίηση του σε κάποια ιατρική ειδικότητα με σχεδιασμό και χρήση του κατάλληλου μοντέλου. Οι μέθοδοι που χρησιμοποιούνται για την αναπαράσταση των λέξεων με διανύσματα είναι η μέθοδος tf-Idf και ο αλγόριθμος ενσωμάτωσης λέξεων Word2Vec. Οι αλγόριθμοι ταξινόμησης που χρησιμοποιήθηκαν για την εξόρυξη γνώσης στην υλοποίηση είναι η Λογιστική Παλινδρόμηση (Logistic Regression), ο αλγόριθμος Naïve-Bayes, ο αλγόριθμος Supported Vector Machine (SVM), ο αλγόριθμος k Nearest Neighbors (kNN) και τα νευρωνικά δίκτυα βαθιάς μάθησης (Neural Networks) με χρήση της βιβλιοθήκης Tensorflow/Keras.

Η εξαγωγή συμπερασμάτων από το υπολογιστικό σύστημα, με είσοδο ως δεδομένο μόνον την περιγραφή ενός ιατρικού περιστατικού, ίσως γίνει η απαρχή ώστε στο μέλλον να αποτελέσει σημαντικό εργαλείο στα χέρια των ιατρών, των ερευνητών και των επιστημόνων.

Λέξεις Κλειδιά

Εξόρυξη κειμένου, κατηγοριοποίηση κειμένου, νευρωνικά δίκτυα, λογιστική παλινδρόμηση text mining, multiclass classification, clinical text, tf-idf, word2vec, deep learning, logistic regression

Abstract

Training a system to understand natural language and extract knowledge out of natural text is a challenge for the scientific community. The purpose of this research field is to create tools to support and simplify the work of human resources.

This thesis focuses on medical text as a source, aiming to correctly preprocess it and train machine learning models to automatically identify the medical specialty to which it should be classified.

The techniques used to vectorize words are the inverse terms frequency (Tf-Idf) and the Word2Vec algorithm. Classification algorithms used are Logistic Regression, Naïve-Bayes, k Nearest Neighbors (kNN), Supported Vector Machines (SVM) and Neural Networks, using the Tensorflow/Keras library.

Creating a system able to extract knowledge out of medical text using only a medical transcription written in natural language may be a strong basis for future development of powerful tools for doctors, scientists and researchers.

Keywords

Machine Learning, Artificial Intelligence, tf-idf, Word2Vec, Logistic Regression, Naïve-Bayes, kNN, SVM, Neural Networks, text mining, text classification, medical text, NLP

στην οικογένεια μου, το θεμέλιο κάθε οικοδομήματος της ζωής μου

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω τον καθηγητή κ. Χρήστο Μακρή για την επίβλεψη αυτής της διπλωματικής εργασίας, για την υποστήριξη και καθοδήγηση του.

Επίσης ευχαριστώ ιδιαίτερα τον Δρ. Αριστείδη Βραχάτη για την έμπρακτη υποστήριξη του και την εξαιρετική συνεργασία που είχαμε μέσω συζητήσεων, οδηγιών και ανταλλαγής ιδεών.

Τέλος θα ήθελα να ευχαριστήσω τον παππού και τη γιαγιά, τους γονείς μου και τα έξι αδέρφια μου, τους φάρους που φωτίζουν το δρόμο της ζωής μου με διάκριση και αγάπη, σε κάθε βήμα, με κάθε κόστος, φροντίζοντας την ψυχή και την καρδιά μου.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Η Μηχανική Μάθηση (Machine Learning) αποτελεί τομέα της Τεχνητής Νοημοσύνης (Artificial Intelligence) και αναφέρεται στο σχεδιασμό και την υλοποίηση συστημάτων που μπορούν να εκπαιδευτούν και να παράγουν γνώση από ένα σύνολο δεδομένων. Για παράδειγμα θα μπορούσαμε από ένα σύνολο μηνυμάτων να εκπαιδεύσουμε ένα σύστημα ωστε να αναγνωρίζει ένα μήνυμα ως απειλητικό ή μη. Μετά το πέρας της διαδικασίας εκπαίδευσης, το σύστημα θα είναι σε θέση να αναγνωρίζει ένα οποιοδήποτε μήνυμα που δεν ανήκει στο σύνολο των ήδη γνωστών δεδομένων και να το χαρακτηρίζει ως προς το αν είναι απειλητικό για το χρήστη.

Αυτή η διαδικασία αποτελείται απο δύο βασικές έννοιες: την αναπαράσταση και τη γενίκευση. Ο όρος αναπαράσταση αναφέρεται στην προσαρμογή των δεδομένων (πχ το μήνυμα κειμένου) σε γλώσσα και μορφή κατανοητή για τον υπολογιστή. Αυτό επιτυγχάνεται με χρήση εξισώσεων και μεθόδων μετατροπής που συχνά μοιάζουν με τις λειτουργίες των νευρώνων. Ο όρος γενίκευση αναφέρεται στη δυνατότητα του εκπαιδευμένου συστήματος να λειτουργεί εξ ίσου καλά με δεδομένα πάνω στα οποία δεν έχει εκπαιδευτεί.

Η εξόρυξη δεδομένων (data mining) είναι βασικό κομμάτι της διαδικασίας εξαγωγής γνώσης από τα δεδομένα. Η άντληση πληροφορίας από ένα σύνολο δεδομένων συχνά άγνωστου περιεχομένου και η δημιουργία γνώσης απο αυτήν, αποτελεί μια πρόκληση που απαιτεί καλό σχεδιασμό και σε νοηματικό αλλά και σε προγραμματιστικό επίπεδο.

Προτού φτάσουμε σε αυτό το σημείο, καλή και απαραίτητη πρακτική αποτελεί το στάδιο της Προεπεξεργασίας (Preproccessing). Σε αυτό το στάδιο, εκτελούμε κάποιες διεργασίες στο σύνολο των δεδομένων ώστε να φτάσουν στην είσοδο του προγράμματος πιο ¨καθαρά¨, απομονώνοντας το θόρυβο με σκοπό να κρατήσουμε την πιο ουσιώδη πληροφορία. Στην ουσία, διατηρούνται μόνον οι πιο σημαντικές λέξεις στην πιο απλή μορφή τους και διαγράφονται άρθρα, συνηθισμένα ρήματα χωρίς σημαντική νοηματική συνεισφορά, σημεία στίξης και άλλες συχνά χρησιμοποιούμενες λέξεις, απαραίτητες για την επικοινωνία μεταξύ των ανθρώπων αλλά δίχως ιδιαίτερη επίδραση στη διαμόρφωση του νοήματος. Αυτό έχει ως αποτέλεσμα την εξοικονόμηση πόρων, υπολογιστικής ισχύος αλλά και την πιο στοχευμένη εκπαίδευση του συστήματος και φυσικά τη βελτίωση των αποτελεσμάτων.

Οι διεργασίες που επιτελούνται στο στάδιο της Προεπεξεργασίας είναι οι εξής:

- Καθαρισμός Δεδομένων (data cleansing)
- Ενοποίηση δεδομένων (Data integration)
- Μετασχηματισμός δεδομένων (Data transformation) και Διακριτοποίηση δεδομένων (Data discretization)
- Μείωση δεδομένων (Data reduction)

Κεφάλαιο 2

Related work

Η χρήση τεχνολογιών Machine Learning και NLP φαίνεται να απασχολεί ιδιαίτερα την ιατροδιαγνωστική κοινότητα, αφού αποτελεί έναν τομέα που θα επωφεληθεί ιδιαίτερα από τα αποτελέσματα τους. Υποστηρίζεται από πολλούς ότι εάν η απόδοση της αγγίξει επιθυμητά αποτελέσματα τότε οι επιτυχείς διαγνώσεις θα μπορούν να επιτυγχάνονται σε σημαντικά μικρότερο χρονικό διάστημα ίσως ξεπερνώντας σε ακρίβεια τις απόψεις των γιατρών οι οποίοι πολλές φορές επηρεάζονται από τα συναισθήματα τους. Αρκετές επιστημονικές ομάδες έχουν ασχοληθεί με τον τομέα αυτό λαμβάνοντας αρκετά υποσχόμενα αποτελέσματα, όμως θα πρέπει να σημειωθεί ότι οι περισσότεροι από αυτούς συμφωνούν στο ότι οι τεχνικές αυτές δεν στοχεύουν στην αντικατάσταση των ιατρών αλλά θα μπορούν να προσφέρουν πολύ σημαντική πληροφορία διευκολύνοντας ιδιαίτερα την λήψη αποφάσεων.

Οι Po-Hao Chen, Hanna Zafar, Maya Galperin-Aizenberg & Tessa Cook στην εργασία τους [3] ενσωματώνουν αλγόριθμους Επεξεργασίας Φυσικής Γλώσσας και Μηχανικής Μάθησης για την κατηγοριοποίηση της ογκολογικής απόκρισης στις αναφορές ακτινολογίας. Υποστηρίζουν ότι οι τεχνικές επεξεργασίας φυσικής γλώσσας (NLP) και μηχανικής μάθησης (ML) έχουν δείξει ότι εξάγουν με επιτυχία πληροφορίες από αναφορές μαγνητικής ή αξονικής τομογραφίας. Συνδυάζουν κάθε μία από τις τρεις τεχνικές NLP με πέντε αλγόριθμους ML για να προβλέψουν την ογκολογική διάγνωση του ασθενούς χρησιμοποιώντας το μη δομημένο κείμενο της αναφοράς της και συγκρίνουν την απόδοση κάθε συνδυασμού. Οι NLP αλγόριθμοι που χρησιμοποιήθηκαν είναι οι TF-IDF, TF, και το 16 bit hashing. Οι ML αλγόριθμοι από την άλλη είναι οι Logistic Regression, RDF, SVM, BPM και NN. Κατέληξαν στο ότι με όλες τις παραμέτρους βελτιστοποιημένες, το SVM είχε την καλύτερη απόδοση στο σύνολο δεδομένων δοκιμής, με μέση ακρίβεια 90,6 και βαθμολογία F 0,813. Η αλληλεπίδραση μεταξύ των αλγορίθμων ML και NLP και η επίδρασή τους στην ακρίβεια της ερμηνείας είναι πολύπλοκη. Η καλύτερη ακρίβεια επιτυγχάνεται όταν και οι δύο αλγόριθμοι βελτιστοποιούνται ταυτόχρονα.

Οι Shiva Kazempour Dehkordi & Hedieh Sajedi επιχειρούν πρόβλεψη ασθένειας με βάση τη πρώτιστη συνταγογράφηση με χρήση μεθόδων εξόρυξης δεδομένων. [4] Ο στόχος της έρευνάς τους είναι η χρήση μεθόδων εξόρυξης δεδομένων για την εξεύρεση γνώσης από ένα σύνολο δεδομένων που παρασχέθηκε από ένα ερευνητικό κέντρο. Αναλύοντας τα

φάρμακα που αγοράστηκαν από τον κάθε ασθενή, η προτεινόμενη μέθοδος τους στοχεύει στο να προβλέψει τον τύπο του γιατρού στον οποίο έχει ανατεθεί ο κάθε ασθενής και το είδος της νόσου από την οποία πάσχει. Χρησιμοποιούν τρεις αλγορίθμους εξόρυξης δεδομένων, οι οποίες ήταν το δέντρο απόφασης, ο Naïve Bayes και ο kNN. Βλέποντας όμως ότι κανένας από αυτούς δεν λειτούργησε σωστά, εφαρμόστηκε ένας ταξινομητής στοίβαξης, ο οποίος αποδείχθηκε ότι έχει μεγαλύτερη ακρίβεια από τους προηγούμενους. Στην πρώτη έκδοση του συνόλου δεδομένων, τρεις διαφορετικοί βασικοί αλγόριθμοι που περιλαμβάνουν kNN, Decision Tree και SVM εφαρμόστηκαν για ταξινόμηση, ενώ στη δεύτερη έκδοση, τέσσερις διαφορετικοί, όπως kNN, Decision Tree, Generalized Linear Model και Random Forest χρησιμοποιήθηκαν στον Stacking Operator. Επιχειρούν έτσι αρκετά πειράματα για να συγκριθεί η απόδοση διαφορετικών τεχνικών εξόρυξης δεδομένων για την πρόβλεψη των ασθενειών και τα αποτελέσματα δείχνουν ότι το προτεινόμενο Stacking Model έχει υψηλότερη ακρίβεια σε σύγκριση με άλλες τεχνικές εξόρυξης δεδομένων όπως το k-Narest Neighbor (kNN).

Στον γειτονικό τομέα της κτηνιατρικής οι Abdullah Awaysheh, Jeffrey Wilcke, François Elvinger, Loren Rees, Weiguo Fan συμφωνούν ότι οι μέθοδοι μηχανικής μάθησης μπορούν να βοηθήσουν στις διαδικασίες λήψης ιατρικών αποφάσεων τόσο σε κλινικό όσο και σε διαγνωστικό επίπεδο. [1] Στην έρευνά τους ρίχνουν μια μηχανιστική ματιά σε τρεις αρχετυπικούς αλγόριθμους μάθησης — naive Bayes, δέντρα αποφάσεων και νευρωνικά δίκτυα — που χρησιμοποιούνται συνήθως για την τροφοδοσία αυτών των εργαλείων υποστήριξης ιατρικών αποφάσεων. Επίσης εστιάζουν τη παρατήρηση τους στα σύνολα δεδομένων που χρησιμοποιούνται για την εκπαίδευση αυτών των αλγορίθμων και εξετάζουν μεθόδους για επικύρωση, αναπαράσταση δεδομένων, μετασχηματισμό και σωστή επιλογή features. Απέδειξαν ότι η ποιότητα των δεδομένων εισόδου έχει μεγάλο αντίκτυπο στη διαδικασία μηχανικής μάθησης και στην απόδοση αυτών των συστημάτων και παρουσιάζουν αποδεικτικά στοιχεία ότι αυτές οι εφαρμογές βελτιώνουν την ακρίβεια των ιατρικών διαγνώσεων και συμβάλλουν σε καλύτερα αποτελέσματα ασθενών

Κεφάλαιο 3

Θεωρητικό υπόβαθρο

Στη συνέχεια, θα αναφερθεί όλη η απαραίτητη θεωρία πίσω από τις τεχνολογίες, τις μεθόδους και τις τεχνικές που επιλέχθηκαν για την εκπόνηση της παρούσας διπλωματικής.

3.1 Ιστορική αναδρομή

Μπορεί μια μηχανή να σκεφτεί; Η συγκεκριμένη ερώτηση αποτυπώνει τη βάση για την ανάπτυξη της επιστήμης της μηχανικής μάθησης και διατυπώνεται για πρώτη φορά από τον Alan Turing στο βιβλίο του Computing Machinery and Intelligence τον Οκτώβριο του 1950. [17] Ήδη νωρίτερα όμως, πολλοί επιστήμονες προσπαθούν να εκφράσουν αντίστοιχες ανησυχίες. Όλα ξεκινούν το 1940, όταν σχεδιάζεται η πρώτη σκεπτόμενη μηχανή από τον J. Von Neumann, ενώ στη συνέχεια το 1943 οι McCulloch και Pitts κατασκευάζουν ένα ηλεκτρονικό σύστημα με σκοπό την προσομοίωση του εγκεφάλου και της λειτουργίας των νευρώνων, εισάγοντας έτσι για πρώτη φορά τους όρους «Electronic Brain» και «Tresholded Logic Unit». [6]

Το μοντέλο αλληλεπίδρασης των εγκεφαλικών κυττάρων, στο οποίο βασίζεται εν μέρει η μηχανική μάθηση, δημιουργείται το 1949 από τον Donald Hebb και περιγράφεται στο βιβλίο του The Organization of Behavior μαζί με θεωρίες για τον τρόπο που επικοινωνούν οι νευρώνες μεταξύ τους (βάρος νευρώνων) [5]. Στο τέλος της δεκαετίας, ο Turing θέτει διάφορα κριτήρια για τον χαρακτηρισμό μιας μηχανής ως «έξυπνης», τα οποία εκφράζονται με τη μορφή «ανάκρισης» της μηχανής από τρεις σταθμούς και η διαδικασία αυτή θα μένει γνωστή ως «δοκιμή Turing». [9]

Με την έννοια της σκεπτόμενης μηχανής να εμπεδώνεται στην επιστημονική κοινότητα, το 1952 ο Arthur Samuel έρχεται να την περιγράψει για πρώτη φορά με τον όρο «Μηχανική Μάθηση», ο οποίος παραμένει μέχρι και σήμερα. Στα πλαίσια ανάπτυξης του όρου από τη δική του σκοπιά, συνθέτει έναν υπολογιστικό κώδικα για το παιχνίδι της ντάμας, επιτρέποντας στο πρόγραμμα να επιλέγει την επόμενη κίνηση στο παιχνίδι και δίνοντάς του τη δυνατότητα μέσω διαφόρων μηχανισμών να βελτιώνεται. [5]

Παράλληλα, το 1957, ο Frank Rosenblatt συνδυάζοντας το μοντέλο αλληλεπίδρασης του Donald Hebb και τις θεωρίες του Arthur Samuel για την Μηχανική Μάθηση, αναπτύσσει

τον αλγόριθμο «Perceptron». Προτείνει, λοιπόν, ένα στοιχειώδες Τεχνητό Νευρωνικό Δίκτυο απλού αισθητήρα, το οποίο θα μπορεί να αναγνωρίσει γράμματα και αριθμούς. Το λογισμικό που κατασκευάζει με βάση τον συγκεκριμένο αλγόριθμο εγκαθίσταται σε μία μηχανή γνωστή ως «Mark 1 perceptron» και επιτρέπει τελικά μερικώς την αναγνώριση εικόνων, μη κατορθώνοντας την πολλαπλή αναγνώριση διαφόρων οπτικών προτύπων και σκορπίζοντας αμφιβολίες για τη λειτουργία της Μηχανικής Μάθησης με Νευρωνικά Δίκτυα. [5], [18], [8]

Αν και το 1967 γράφεται ο αλγόριθμος «πλησιέστερου γείτονα» (Nearest Neighbor Algorithm), ο οποίος αποτελεί τη βάση για την αναγνώριση προτύπων, το 1969 οι Minsky και Papert αποδεικνύουν μαθηματικά ότι τα Τεχνητά Νευρωνικά Δίκτυα ενός επιπέδου, όπως το Perceptron, δεν επιλύουν μη γραμμικά προβλήματα. [6], [5]

3.2 Ορισμοί

3.2.1 Τεχνητή Νοημοσύνη - Artificial Intelligence (AI)

Ο όρος «Τεχνητή Νοημοσύνη» ξεκίνησε ως η απλή θεωρία ότι η ανθρώπινη νοημοσύνη μπορεί να χρησιμοποιηθεί από μηχανές. Η σκέψη ότι μια μηχανή μπορεί να δημιουργήσει κάποιου είδους νοημοσύνη ενθουσίαζε τον άνθρωπο ήδη από την αρχή της ύπαρξης υπολογιστών. Παρ'όλα αυτά, ακόμη και σήμερα, ο ακριβής ορισμός της Τεχνητής Νοημοσύνης αποτελεί θέμα συζήτησης και διαφωνιών. Παρακάτω παρουσιάζονται μερικοί από τους ορισμούς που διατυπώνονται για την Τεχνητή Νοημοσύνη:

- Το πεδίο της μελέτης στον τομέα της επιστήμης των υπολογιστών, το οποίο ασχολείται με την ανάπτυξη υπολογιστικών μηχανών ικανών να υιοθετήσουν ανθρώπινες διαδικασίες, όπως η μάθηση, η προσαρμοστικότητα, η αυτοδιόρθωση, κλπ.

- Η ιδέα ότι οι μηχανές μπορούν να βελτιωθούν στην υιοθέτηση ορισμένων δυνατοτήτων, που κανονικά σχετίζονται με την ανθρώπινη νοημοσύνη, όπως η μάθηση, η προσαρμοστικότητα, η αυτοδιόρθωση, κλπ.

- Η εξέλιξη της ανθρώπινης νοημοσύνης ύστερα από τη χρήση υπολογιστών ως φυσική εξέλιξη του ανθρώπου, όπως στο παρελθόν η σωματική δύναμη επεκτάθηκε μετά στη χρήση μηχανικών εργαλείων.

- Η μελέτη διαφόρων βελτιωμένων τεχνικών προγραμματισμού για την αποτελεσματικότερη χρήση των υπολογιστών.

- Η επιστήμη που μελετά τη μίμηση της ανθρώπινης ευφυής συμπεριφοράς.

Οι ορισμοί αλλάζουν όσο η επιστήμη εξελίσσεται. Σήμερα, από μία απλή θεωρία, έχει μετατραπεί σε απτές εφαρμογες. Με σκοπό πάντα την ανάπτυξη υπολογιστικών συστημάτων, τα οποία θα πλησιάζουν την ανθρώπινη συμπεριφορά, μπορούμε να χωρίσουμε τους ορισμούς που έχουν δοθεί κατά καιρούς από την επιστημονική κοινότητα, ανάλογα τα συστήματα, στις παρακάτω κατηγορίες:

- Συστήματα που σκέφτονται ως άνθρωποι
- Συστήματα που δρουν ως άνθρωποι
- Συστήματα που σκέφτονται ορθολογικά
- Συστήματα που δρουν ορθολογικά [19], [7], [12]

3.2.2 Μηχανική Μάθηση – Machine Learning (ML)

Ένας από τους ευρύτερους ορισμούς της Μηχανικής Μάθησης δίνεται από τον Tom Mitchell:

«Το πεδίο της Μηχανικής Μάθησης ασχολείται με το ερώτημα πώς μπορούν να κατασκευαστούν υπολογιστικά προγράμματα τα οποία έχουν την ικανότητα να βελτιώνονται αυτόματα μέσω εμπειρίας.»[2]

Ξεκινώντας, πρέπει να τονιστεί ότι η Μηχανική Μάθηση αποτελεί ένα υποσύνολο της Τεχνητής Νοημοσύνης. Συχνά οι δύο επιστήμες συγχέονται λόγω των δυνατοτήτων της πρώτης ως προς τη «μάθηση» και την «λήψη αποφάσεων». Ωστόσο, όχι μόνο δεν συμπίπτουν, αλλά η Μηχανική Μάθηση αποτελεί ένα σημαντικό εργαλείο για την επίτευξη αξιοποίησης διαφόρων τεχνολογιών που σχετίζονται με την Τεχνητή Νοημοσύνη. Πιο συγκεκριμένα, αποτυπώνει τη βελτίωση μέσω της βιωματικής «μάθησης» που σχετίζεται με την ανθρώπινη νοημοσύνη, αξιοποιώντας υπολογιστικούς αλγόριθμους. Για την σταδιακή βελτίωση της απόδοσης υπολογιστικών συστημάτων χρησιμοποιούνται αλγόριθμοι, οι οποίοι αυτόματα «χτίζουν» ενα μαθηματικό μοντέλο, χρησιμοποιώντας μεγάλο αριθμό δεδομένων, γνωστών κι ως «τραινινγ δατα», για να προβλέψουν διάφορες εξόδους, χωρίς να είναι ειδικά προγραμματισμένοι για αυτές τις αποφάσεις. Στη συνέχεια, τα αποτελέσματα συγκρίνονται με ένα σύνολο γνωστών αποτελεσμάτων, υπολογίζεται η ακρίβεια των εξόδων του κάθε αλγόριθμου και επαναπροσαρμόζεται, ώστε να τελειοποιήσει την ικανότητα πρόβλεψής του. Μέσω πολλών επαναλήψεων και τροποποιήσεων, λοιπόν, το σύστημα καταφέρνει να «μάθει» να παίρνει αυτόνομες αποφάσεις. [5], [7]

Η διαδικασία αυτή εκμάθησης οριοθετείται από τον Tom Mitchell, ο οποίος καθορίζει μια συγκεκριμένη μορφή τυποποίησης για τους υπολογιστικούς όρους:

«Ένα υπολογιστικό πρόγραμμα λέγεται ότι μαθαίνει μέσω της εμπειρίας E, με γνώμονα μια τάξη διαφόρων εργασιών T και ένα μέτρο απόδοσης P, εάν η απόδοσή του σε εργασίες T, σύμφωνα με το μέτρο P, βελτιώνεται αναφορικά με την εμπειρία E.»

Συλλογιζόμενοι τον ορισμό αυτόν, μπορούμε να πούμε ότι αποτελεί ένα εργαλείο ταχτοποίησης των προβλημάτων Μηχανικής Μάθησης, διαχωρίζοντας τα συλλεχθέντα δεδομένα στο E, τις αποφάσεις που λαμβάνει το σύστημα στο T, και την επικύρωση των αποτελεσμάτων σύμφωνα με το P. Συνοψίζοντας, «Μηχανική Μάθηση είναι η εκπαίδευση ενός μοντέλου από δεδομένα, η οποία γενικεύει μια απόφαση έναντι ενός μέτρου απόδοσης.» Ερμηνεύοντας αυτόν τον ορισμό, θα μπορούσαμε να πούμε ότι η φράση «η εκπαίδευση ενός μοντέλου» παραπέμπει στη χρήση εκπαιδευτικών δεδομένων και παραδειγμάτων, η λέξη «μοντέλο» υπονοεί μία κατάσταση που αποκτάται μέσω εμπειρίας, το ρήμα «γενικεύει»

συμπεριλαμβάνει τη λήψη αποφάσεων με τη διαδικασία της πρόβλεψης, και τέλος, το «μέτρο απόδοσης» υποδηλώνει την ανάγκη του μοντέλου να προσαρμοστεί και να εκαπαιδευτεί, ώστε η τελική απόφαση να είναι άρτια. [5], [7], [2]

3.3 Διαδικασίες εκμάθησης Μηχανικής Μάθησης

Οι τεχνικές Μηχανικής Μάθησης διαφέρουν ανάλογα με τη φύση του προβλήματος και ταξινομούνται ανάλογα με τη φύση του συστήματος εκπαίδευσης σε τρεις μεγάλες κατηγορίες:

3.3.1 Μάθηση με Επίβλεψη / Επιτηρούμενη Μάθηση – Supervised Learning

Στην περίπτωση της επιτηρούμενης μάθησης, ή αλλιώς επιβλεπόμενης μάθησης, το υπολογιστικό σύστημα καλείται να «μάθει» μια γενική μέθοδο ώστε να αντιστοιχήσει τα δεδομένα εισόδου (βάση γνώσης) στα επιθυμητά αποτελέσματα, μετά από την επέμβαση ενός «επιβλέποντος», ο οποίος παρέχει κάποιες σωστές τιμές εξόδου για τη βάση γνώσης που αναλύεται. Είναι μια τεχνική, η οποία στοχεύει, δηλαδή, στον χαρακτηρισμό δεδομένων χρησιμοποιώντας ως γνώμονα κάποια δεδομένα εκπαίδευσης, που είναι ουσιαστικά ένα σύνολο παραδειγμάτων, το οποίο καθορίζει ο «επιβλέπων». [;] Λαμβάνοντας υπόψη ότι επαγωγική μάθηση ονομάζεται η επιχείρηση του ανθρώπινου εγκεφάλου να κατανοήσει το περιβάλλον του παρατηρώντας και δημιουργώντας απλοποιημένα μοντέλα, μπορούμε να πούμε ότι το σύστημα που εκπαιδεύεται με πλήρη επίβλεψη, μαθαίνει με επαγωγική μέθοδο, δημιουργώντας δηλαδή μοντέλα, να προβλέπει κάποιες άγνωστες ιδιότητες, χρησιμοποιώντας τις γνωστές, από τα παραδείγματα, ιδιότητες που εισάγει ο εκπαιδευτής. Πιο αναλυτικά, θα μπορούσαμε να θεωρήσουμε ότι ο «επιβλέπων» εκπαιδευτής εισάγει ένα σύνολο παραδειγμάτων εισόδου-εξόδου από το γνωστό σε εκείνον περιβάλλον (διανύσματα εκπαίδευσης). Στη συνέχεια, παρέχει στο μοντέλο την επιθυμητή ενέργεια που πρέπει να εκτελέσει, η οποία δίνεται από μια συνάρτηση περιγραφής δεδομένων (συνάρτηση στόχος - target function) , ώστε να προβλεφθεί σωστά η μεταβλητή εξόδου (εξαρτημένη) βάσει της μεταβλητής εισόδου (ανεξάρτητη, χαρακτηριστικό). Οι παράμετροι, τελικά, του συστήματος προσαρμόζονται συνδυαστικά από την επιρροή του επιβλέποντος και το σφάλμα απόδοσης (η διαφορά της επιθυμητής από την πραγματική απόκριση του συστήματος). Μετά από μια σειρά επαναληπτικών προσαρμογών επιτυγχάνεται η εκπαίδευση του συστήματος ως προς την επίγνωση του περιβάλλοντος. Επιτυγχάνεται, δηλαδή, η εκπαίδευση του συστήματος σύμφωνα με την αντίληψη και τη συμπεριφορά του εκπαιδευτή ως προς το περιβάλλον. Παράλληλα, αποθηκεύεται η επιτευχθής γνώση στη μακροπρόθεσμη μνήμη του συστήματος με σκοπό την προσπάθεια λειτουργίας του συστήματος στο συγκεκριμένο περιβάλλον χωρίς πια εκπαιδευτή. Αναμένεται, λοιπόν, από το σύστημα ότι για κάθε υπόθεση η που έχει βρεθεί να προσεγγίζει καλά τη συνάρτηση στόχο για ένα μεγάλο σύνολο διανυσμάτων εκπαίδευσης, θα προσεγγίζει το ίδιο καλά τη συνάρτηση στόχο και για περιπτώσεις που δεν είναι γνωστές. [20], [;]

Συμπερασματικά, θα μπορούσε να δοθεί ο ακόλουθος ορισμός:

Η μάθηση με επίβλεψη είναι η έρευνα εύρεσης αλγορίθμων που βασίζονται σε εξωτερικούς παράγοντες, με σκοπό να παράγουν γενικές υποθέσεις, που στη συνέχεια μπορούν να χρησιμοποιηθούν σε μελλοντικά άγνωστα παραδείγματα. Με άλλα λόγια ο στόχος της επιβλεπόμενης μάθησης είναι ο αλγόριθμος να ταξινομήσει τελικά σε ετικέτες τις κατηγορίες γνωστών παραδειγμάτων, ώστε μελλοντικά να μπορεί με «λογικό» τρόπο να προβλέψει, δηλαδή, να χαρακτηρίσει την ετικέτα ενός άγνωστου παραδείγματος. [20]

3.3.2 Μάθηση χωρίς Επίβλεψη / Μη Επιτηρούμενη Μάθηση – Unsupervised Learning

Στην περίπτωση της μάθησης χωρίς επίβλεψη, δεν έχουμε ετικέτες στα δεδομένα εισόδου και ο στόχος μας είναι να εξάγουμε δομικά χαρακτηριστικά (φεατυρες) και δομές (στρυςτυρες) από τα δεδομένα χωρίς καμία επιπλέον πληροφορία.

Ο σκοπός της μάθησης χωρίς επίβλεψη είναι η ανακάλυψη κρυφών δομικών χαρακτηριστικών των δεδομένων, η οποία μπορεί να οδηγήσει σε καλύτερη κατανόηση των δεδομένων. Οι μέθοδοι μάθησης χωρίς επίβλεψη χρησιμοποιούνται συνήθως σε προβλήματα αναγνώρισης μοτίβων (παττερν ρεςογνιτιον), συσταδοποίησης (ςλυστερινγ), μείωσης διαστάσεων (διμενσιοναλιτψ ρεδυςτιον) και αναδιάταξης δεδομένων (δατα ρεαρρανγεμεντ).

Η συσταδοποίηση είναι μια από τις πιο συνηθισμένες τεχνικές μάθησης χωρίς επίβλεψης και αποσκοπεί στον εντοπισμό ομάδων δεδομένων με κοινά χαρακτηριστικά. Οι αλγόριθμοι συσταδοποίησης αντιμετωπίζουν τα δεδομένα ως ένα σύνολο από δεδομένα που είναι άγνωστα και αποσκοπούν στον εντοπισμό ομάδων δεδομένων που είναι πιθανό να ανήκουν στην ίδια κατηγορία. Αυτή η διαδικασία είναι χρήσιμη σε πολλές εφαρμογές, όπως στην επεξεργασία εικόνας, στον εντοπισμό κοινοτήτων στα κοινωνικά δίκτυα, και στην ανίχνευση ανωμαλιών σε δεδομένα υγείας.

Μια άλλη κατηγορία αλγορίθμων μάθησης χωρίς επίβλεψη είναι η μείωση διαστάσεων. Στην περίπτωση αυτή, το στόχος είναι να μειωθεί η διάσταση των δεδομένων χωρίς να χαθεί η πληροφορία που περιέχουν. Οι αλγόριθμοι μείωσης διαστάσεων μπορούν να χρησιμοποιηθούν για την απομόνωση και απόρριψη περιττών χαρακτηριστικών (μείωση θορύβου), τη βελτίωση της επεξεργασίας των δεδομένων, και την ανίχνευση ενός πιθανού χαρακτηριστικού που επηρεάζει την απόδοση του συστήματος.

Τέλος, οι μέθοδοι μάθησης χωρίς επίβλεψης μπορούν να χρησιμοποιηθούν για την αναδιάταξη των δεδομένων. Αυτό είναι χρήσιμο όταν θέλουμε να οπτικοποιήσουμε τα δεδομένα ή να τα παρουσιάσουμε σε έναν ανθρώπινο χρήστη. Οι αλγόριθμοι αναδιάταξης δεδομένων αντιστρέφουν την αρχική μορφή των δεδομένων σε μια νέα μορφή που είναι πιο ευανάγνωστη από τον άνθρωπο.

Συνολικά, η μάθηση χωρίς επίβλεψη είναι μια πολύτιμη κατηγορία μηχανικής μάθησης, καθώς επιτρέπει στους υπολογιστές να ανακαλύψουν μοτίβα και δομές στα δεδομένα που δεν θα ήταν εύκολα εμφανή με το ανθρώπινο μάτι. Οι αλγόριθμοι μάθησης χωρίς επίβλεψη είναι επίσης χρήσιμοι για την αντιμετώπιση προβλημάτων όπου δεν υπάρχουν ετικέτες για τα

δεδομένα ή είναι δύσκολο να προσδιοριστούν. Επιπλέον, η μάθηση χωρίς επίβλεψη μπορεί να χρησιμοποιηθεί για τη βελτίωση των αποτελεσμάτων άλλων αλγορίθμων μάθησης, όπως η μάθηση με επίβλεψη.

Ωστόσο, η μάθηση χωρίς επίβλεψη έχει και ορισμένους περιορισμούς όσον αφορά στις δυνατότητες. Συγκεκριμένα, είναι δύσκολο να ελέγξουμε τα αποτελέσματα, καθώς δεν έχουμε κάποιο κριτήριο για να μετρήσουμε την επίδοση του αλγορίθμου. Επιπλέον, η μάθηση χωρίς επίβλεψη μπορεί να είναι επιρρεπής σε λάθη, καθώς οι αλγόριθμοι μπορεί να επιλέξουν λανθασμένες δομές ή μοτίβα.

Για την αντιμετώπιση αυτών των προκλήσεων, υπάρχουν διάφορες τεχνικές και αλγόριθμοι που χρησιμοποιούνται στη μάθηση χωρίς επίβλεψη. Ορισμένες από αυτές τις τεχνικές περιλαμβάνουν:

- Ανάλυση κύριων συνιστωσών (PCA): Η PCA είναι μια τεχνική που χρησιμοποιείται για τη μείωση της διάστασης των δεδομένων ενώ διατηρεί τις κύριες πληροφορίες. Αυτό επιτυγχάνεται με την εύρεση των κύριων συνιστωσών των δεδομένων και τη μετατροπή τους σε έναν χαμηλότερης διάστασης χώρο.

- Ομαδοποίηση (Clustering): Η ομαδοποίηση είναι μια τεχνική που χρησιμοποιείται για τον εντοπισμό ομάδων δεδομένων με κοινά χαρακτηριστικά. Οι αλγόριθμοι ομαδοποίησης χρησιμοποιούνται συνήθως για την εξερεύνηση των δεδομένων και την εντοπισμό δομών.

- Αυτοπροσδιοριστικά νευρωνικά δίκτυα (Autoencoders): Τα αυτοπροσδιοριστικά νευρωνικά δίκτυα είναι νευρωνικά δίκτυα που εκπαιδεύονται να αναγνωρίζουν τα δεδομένα εισόδου

3.3.3 Ενισχυτική Μάθηση – Reinforcement Learning

Η ενισχυτική μάθηση είναι η τρίτη κατηγορία μηχανικής μάθησης η οποία επικεντρώνεται στην εκπαίδευση ενός αλγορίθμου με τη χρήση ενός συστήματος ενισχύσεων, που του παρέχει πληροφορίες σχετικά με την επίδοσή του σε μια δεδομένη εργασία και προσαρμόζει τη συμπεριφορά του αλγορίθμου ανάλογα με τα αποτελέσματα.

Το σύστημα ενισχύσεων λειτουργεί με τον ακόλουθο τρόπο: ο αλγόριθμος επιχειρεί μια ενέργεια σε ένα περιβάλλον και παρατηρεί την αντίδραση που προκαλεί. Η παρατήρηση αυτή περνάει μέσω του συστήματος ενισχύσεων, που εκτιμά την ποιότητα της επίδοσης του αλγορίθμου και παράγει ένα σήμα ενίσχυσης που επηρεάζει την επόμενη επιλογή ενέργειας του αλγορίθμου. Κατά αυτόν τον τρόπο, ο αλγόριθμος μαθαίνει να προσαρμόζει τη συμπεριφορά του ώστε να μεγιστοποιεί την επίδοσή του σε μια συγκεκριμένη εργασία.

Μια από τις κύριες εφαρμογές της ενισχυτικής μάθησης είναι οι αυτόματοι πράκτορες (αυτονομους αγεντς), δηλαδή ρομπότ και άλλες συστήματα που μπορούν να λαμβάνουν αποφάσεις μέσα σε ένα δεδομένο περιβάλλον. Για παράδειγμα, ένα ρομπότ που μαθαίνει να περπατάει θα μπορούσε να χρησιμοποιήσει ενισχυτική μάθηση για να βελτιώσει την τεχνική του στο βάδισμα. Με τη χρήση ενός συστήματος ενισχύσεων, το ρομπότ θα μπορούσε να

εκτιμήσει την ποιότητα της κίνησής του και να προσαρμόζει τη συμπεριφορά του ώστε να βελτιώνεται σταδιακά.

Μια άλλη εφαρμογή της ενισχυτικής μάθησης είναι στα παιχνίδια, όπου οι αλγόριθμοι μπορούν να μάθουν πώς να επιλέγουν τις καλύτερες δυνατές κινήσεις σε παιχνίδια όπως το Γο και το σκάκι. Στην περίπτωση αυτή, το σύστημα ενισχύσεων λειτουργεί ως αντίπαλος του αλγορίθμου, παρέχοντας το σήμα ενίσχυσης κάθε φορά που ο αλγόριθμος κερδίζει ή χάνει ένα παιχνίδι. Κατά αυτόν τον τρόπο, ο αλγόριθμος μαθαίνει να προσαρμόζει τη συμπεριφορά του ώστε να επιτυγχάνει καλύτερα αποτελέσματα στο μέλλον.

Ένας ακόμη τομέας εφαρμογής της ενισχυτικής μάθησης είναι η ρομποτική. Τα ρομπότ μπορούν να μάθουν να αλληλεπιδρούν με το περιβάλλον τους χρησιμοποιώντας αλγόριθμους ενισχυτικής μάθησης. Για παράδειγμα, ένα ρομπότ που κινείται σε ένα άγνωστο περιβάλλον μπορεί να χρησιμοποιήσει ενισχυτική μάθηση για να μάθει να προσαρμόζει την κίνησή του στη βάση της ανταμοιβής που λαμβάνει από το περιβάλλον του, όπως την απόκτηση ενός στόχου ή την αποφυγή εμποδίων.

Τέλος, η ενισχυτική μάθηση μπορεί να χρησιμοποιηθεί σε εφαρμογές που απαιτούν λήψη αποφάσεων σε συνεχή χρονικό διάστημα, όπως η διαχείριση πόρων. Ένα σύστημα ενισχυτικής μάθησης μπορεί να μάθει πώς να διαχειρίζεται τους πόρους σε μια συστηματική βάση, ώστε να επιτυγχάνει τους στόχους του με τον καλύτερο δυνατό τρόπο.

Συνολικά, η ενισχυτική μάθηση αναδεικνύεται ως μια αναπτυσσόμενη κατηγορία μηχανικής μάθησης με αξιόλογες δυνατότητες και εφαρμογές σε πολλούς τομείς, όπως η ρομποτική, η αυτοματοποίηση και η επικοινωνία με τους ανθρώπους. Η δυνατότητα αυτόνομης μάθησης του συστήματος από την εμπειρία και η ανάπτυξη μεθόδων ανταμοιβής μπορούν να οδηγήσουν στη δημιουργία ευφυών συστημάτων, που μπορούν να επιλύσουν προβλήματα που θεωρούνταν πριν από λίγο καιρό απαγορευτικά. Παράλληλα, η ενισχυτική μάθηση προκαλεί ερωτήματα για την ασφάλεια και την ευθύνη των αποφάσεων που λαμβάνουν τα συστήματα, καθώς και για τις επιπτώσεις τους στο περιβάλλον και την κοινωνία. Επίσης αντιμετωπίζει προκλήσεις οι οποίες μεταξύ άλλων περιλαμβάνουν την ανάγκη για μεγάλα και πολύπλοκα σύνολα δεδομένων, την ανάγκη για επαρκή επικοινωνία μεταξύ του συστήματος και του περιβάλλοντος του, καθώς και την ανάγκη για εξασφάλιση της ασφάλειας και της ευθύνης των αποφάσεων που λαμβάνονται από το σύστημα.

Επιπλέον, τα ερωτήματα που προκύπτουν περιλαμβάνουν το πώς μπορεί να εξασφαλιστεί ότι το σύστημα θα μάθει μόνο επιθυμητές συμπεριφορές και δεν θα εμπλακεί σε επικίνδυνες δραστηριότητες, το πώς μπορεί να ελεγχθεί η απόδοση του συστήματος και να αναγνωριστούν και διορθωθούν πιθανά σφάλματα, και το πώς μπορεί να διασφαλιστεί η αποφασιστικότητα και η δικαιοσύνη των αποφάσεων που λαμβάνονται από το σύστημα.

3.4 Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing - NLP)

3.4.1 TF-IDF (Term Frequency - Inverse Document Frequency)

Το TF-IDF (Term Frequency - Inverse Document Frequency), είναι ένας τρόπος αξιολόγησης λέξεων που χρησιμοποιείται κυρίως στην ανάκτηση πληροφοριών και στην περίληψη. Δείχνει την σχετικότητα ενός όρου μέσα στο δωσμένο κείμενο.

Η λογική του είναι πως όσες περισσότερες φορές εμφανίζεται μια λέξη στο κείμενο, τόσο περισσότερο μεγαλώνει η αξία της (TF). Παράλληλα όμως, αν η λέξη αυτή εμφανίζεται συχνά και σε υπόλοιπα κείμενα του corpus, θα σημαίνει πως πρόκειται για κάποια κοινή λέξη, όχι τόσο σχετική ή ουσιώδης για το αρχικό κείμενο (IDF). Αυτό σημαίνει πως όσο μεγαλύτερη είναι η συλλογή κειμένων, τόσο πιο ακριβή θα είναι τα αποτελέσματα του TF-IDF.

Κάθε λέξη ή όρος, αποκτά το δικό του TF και IDF score. Το παράγωγό τους, ονομάζεται TF-IDF score του όρου. Όσο μεγαλύτερο το TF-IDF, τόσο πιο χρήσιμη είναι μία λέξη, και το αντίστροφο.

Με αυτόν λοιπόν τον τρόπο αξιολόγησης, παράγεται μέσα από το συνολικό κείμενο των πεδίων transcription όλων των εγγραφών του αρχείου ένας πίνακας λέξεων.

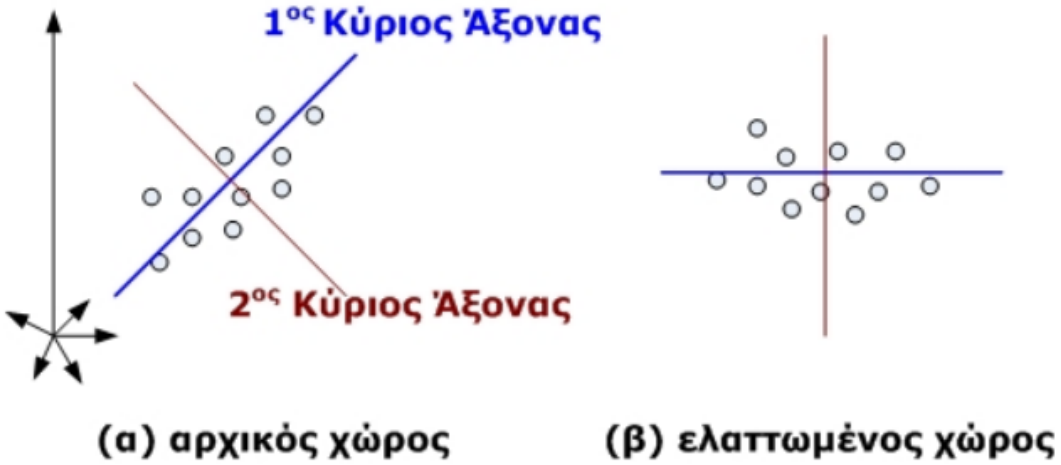
3.4.2 Η τεχνική t-SNE

Η τεχνική t-SNE αποτελεί μέθοδο ενσωμάτωσης (embedding technique), η οποία επιτρέπει την απεικόνιση δεδομένων πολλών διαστάσεων σε γραφική παράσταση δίδοντας σε κάθε σημείο δεδομένων μια θέση σε ένα χάρτη δύο ή τριών διαστάσεων. Η τεχνική είναι μια παραλλαγή της Στοχαστικής Ενσωμάτωσης Γειτόνων (Hinton & Roweis, 2002) που είναι πολύ πιο εύκολη στη βελτιστοποίηση και παράγει σημαντικά βελτιωμένες οπτικές αναπαραστάσεις μειώνοντας την τάση να συγκεντρώνονται σημεία στο κέντρο του χάρτη.

3.4.3 Μείωση διαστατικότητας: η μέθοδος PCA

Στη συνέχεια εφαρμόζουμε τη μέθοδο PCA στον πίνακα tf-idf. Η μέθοδος PCA (Principle Component Analysis - Ανάλυση Κύριων Συνιστωσών), η οποία αποτελεί μία γραμμική μέθοδο συμπίεσης δεδομένων η οποία συνίσταται από τον επαναπροσδιορισμό των συντεταγμένων ενός συνόλου δεδομένων σε ένα άλλο σύστημα συντεταγμένων το οποίο θα είναι καταλληλότερο στην επικείμενη ανάλυση δεδομένων.

Η μέθοδος αυτή προσπαθεί να υπολογίσει τους άξονες εκείνους στους οποίους παρατηρείται η μέγιστη διασπορά των δεδομένων. Μια γενικευμένη απεικόνιση ως παράδειγμα του πως λειτουργεί η συγκεκριμένη μέθοδος φαίνεται στο Σχήμα 3.1 :



Σχήμα 3.1: Παράδειγμα λειτουργίας της μεθόδου PCA

3.4.4 Η τεχνική SMOTE

Για την αντιμετώπιση του προβλήματος της ανισορροπίας των δεδομένων (imbalanced data), οι ερευνητές Μηχανικής Μάθησης χρησιμοποιούν:

1. Είτε τεχνικές oversampling, οι οποίες έχουν ως στόχο να δημιουργήσουν στιγμιότυπα σε κλάσεις που θεωρούνται μειονότητες σε σχέση με το μέσο όρο

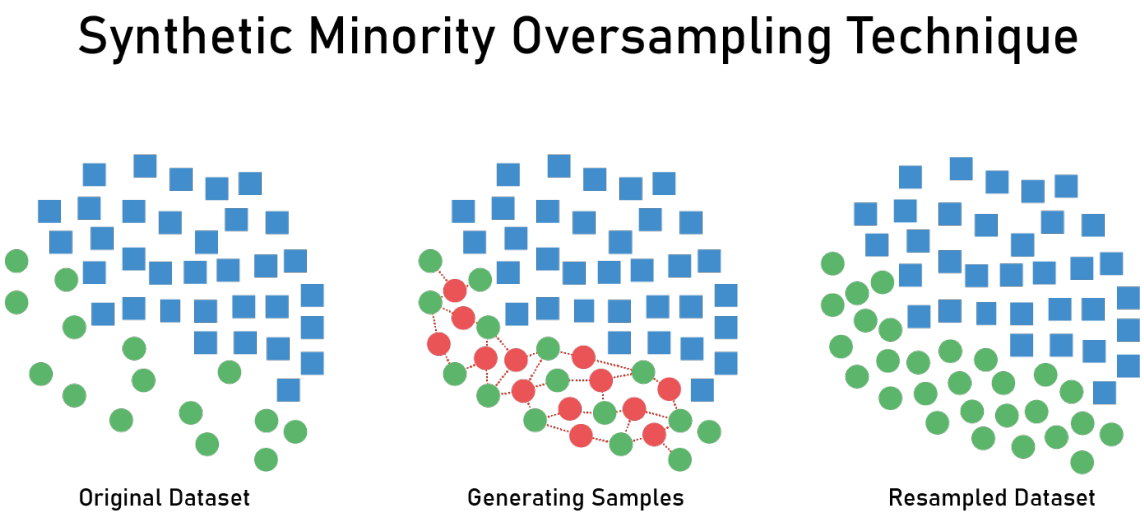
2. Είτε τεχνικές undersampling που ακολουθούν την αντίστροφη διαδικασία, αφαιρούν δηλαδή στιγμιότυπα απο τις κλάσεις με το μεγαλύτερο πληθυσμό.

Οι δύο αυτές διαδικασίες χρήζουν ιδιαίτερης προσοχής ώστε, στην πρώτη περίπτωση να μη δημιουργηθούν χαμηλής ποιότητας και αξίας δεδομένα, ή αλλιώς θόρυβος, και στη δεύτερη περίπτωση, να μη χαθεί χρήσιμη πληροφορία.

Ο αλγόριθμος του SMOTE (Synthetic Minority Oversampling Technique), όπως υποδηλώνει και το πλήρες όνομα του, εντοπίζει την κλάση ή τις κλάσεις με τα ελλιπή στιγμιότυπα και συνθέτει νέα στιγμιότυπα από τα ήδη υπάρχοντα, με σκοπό την εξισορρόπηση των στιγμιότυπων μεταξύ των κλάσεων του συνόλου δεδομένων. Τα στιγμιότυπα αυτά είναι συνθετικά και δημιουργούνται αξιοποιώντας τις τιμές των ήδη υπαρχόντων στιγμιοτύπων. Εφαρμόζει την προσέγγιση KNN (k-nearest neighbours), με βάση τους K πλησιέστερους γείτονες, δημιουργεί τα συνθετικά δείγματα στον χώρο της μειοψηφίας. Ο αλγόριθμος παίρνει τα διανύσματα χαρακτηριστικών από τους πλησιέστερους γείτονές του και υπολογίζει την απόσταση μεταξύ αυτών των διανυσμάτων. Η διαφορά πολλαπλασιάζεται με τυχαίο αριθμό μεταξύ (0, 1) και προστίθεται στο χαρακτηριστικό.

Στο σχήμα 3.2 γίνεται μια προσπάθεια επεξήγησης του τρόπου λειτουργίας του αλγορίθμου της τεχνικής SMOTE γραφικά. Τα στιγμιότυπα των κλάσεων με το μεγαλύτερο πληθυσμό απεικονίζονται με μπλε χρώμα, τα στιγμιότυπα των κλάσεων που αποτελούν

μειονότητες απεικονίζονται με πράσινο χρώμα και τα συνθετικά στιγμιότυπα που δημιουργεί η τεχνική SMOTE απεικονίζονται με κόκκινο χρώμα και ενσωματώνονται στην κλάση-μειονότητα:



Σχήμα 3.2: Γραφική επεξήγηση της λειτουργίας του αλγορίθμου SMOTE

3.5 Αλγόριθμοι Ταξινόμησης (Classification Algorithms)

3.5.1 Λογιστική Παλινδρόμηση (Logistic Regression)

Η Λογιστική Παλινδρόμηση (Logistic Regression), ή εναλλακτικά το λογιστικό υπόδειγμα πιθανότητας, αποτελεί στην ουσία ένα μη γραμμικό μοντέλο ταξινόμησης των τιμών μιας μεταβλητής απόκρισης Υ με βάση τη θεωρία πιθανοτήτων. Στο λογιστικό μοντέλο το σφάλμα δεν ακολουθεί την κανονική κατανομή. Επίσης, η εξαρτημένη μεταβλητή Υ είναι δυική (boolean), δηλαδή παίρνει τιμές 0 και 1 και αναφέρεται στην πραγματοποίηση ή όχι ενός γεγονότος. Η λογιστική παλινδρόμηση χρησιμοποιείται εκτενώς σε πολλές εφαρμογές όπου επιδιώκεται η πραγματοποίηση πρόβλεψης της παρουσίας ή απουσίας κάποιου χαρακτηριστικού ή γενονότος. Πρακτικά, αναπτύσσεται μια μη γραμμική συνάρτηση, βάσει της οποίας υπολογίζεται η πιθανότητα ένα στοιχείο να έχει ή όχι το χαρακτηριστικό το οποίο εξετάζεται την κάθε στιγμή.

Η συνάρτηση αυτή ονομάζεται λογιστική συνάρτηση και εκφράζεται από τον τύπο:

$$P_i = F(w_0 + wx_i) = \frac{1}{1+e^{-w_0 - wx_i}}$$

Τέλος, αξίζει να σημειωθεί ότι πρώτος ο James A. Ohlson, το 1980, χρησιμοποίησε το λογιστικό υπόδειγμα πιθανότητας σε έρευνα του σχετικά με την πρόβλεψη της πτώχευσης επιχειρήσεων. [14]

3.5.2 Naïve Bayes

Ο αλγόριθμος Naïve Bayes αποτελεί ένα σύνολο πιθανολογικών μοντέλων μηχανικής μάθησης, με σκοπό την επίλυση προβλημάτων ταξινόμησης, το οποίο βασίζεται στο θεώρημα πιθανοτήτων Bayes. Αξιοποιώντας το Θεώρημα Bayes ο αλγόριθμος, για κάθε δεδομένο εισόδου (πχ μήνυμα κειμένου, tweet, τίτλος άρθρου), προβλέπει, μέσα απο ένα σύνολο δεδομένων διακριτών κατηγοριών, την κατηγορία στην οποία ανήκει. Η πρόβλεψη αποτελεσμάτων από μη επισημασμένα δεδομένα έχει ως προϋπόθεση την ανεξαρτησία μεταξύ των χαρακτηριστικών (features), μια προϋπόθεση που εάν στην πραγματικότητα αποτελούσε εφικτό στόχο θα απλοποιούσε σημαντικά τη διαδικασία μάθησης.

Ο αλγόριθμος Naïve Bayes χαρακτηρίζεται ως πιθανολογικός λόγω του ότι η πρόβλεψη γίνεται αποδίδοντας μια πιθανότητα για κάθε δεδομένη κατηγορία και επιστρέφοντας εκείνη με τη μεγαλύτερη πιθανότητα. Οι πιθανότητες υπολογίζονται μεσω του θεωρήματος Bayes, το οποίο αξιολογεί πιθανολογικά το κάθε χαρακτηριστικό (feature) ξεχωριστά βάσει της ήδη υπάρχουσας γνώσης που σχετίζεται με αυτό.

Θεώρημα Bayes

$$P(A/B) = \frac{P(A) * P(B/A)}{P(B)}$$

όπου A και B γεγονότα.

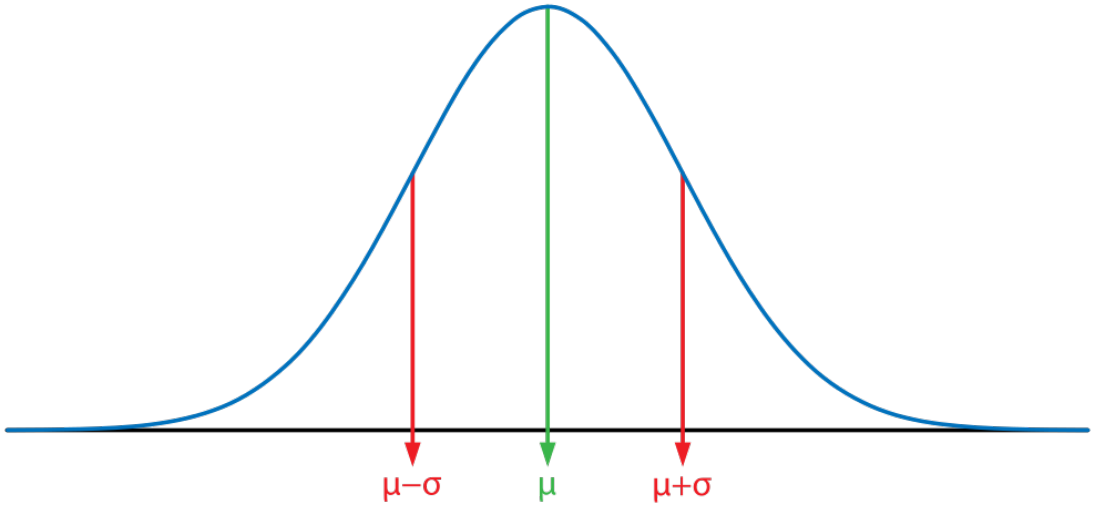
- P(A) και P(B) είναι οι πιθανότητες των A και B που είναι ανεξάρτητα μεταξύ τους.
- P(A/B), η υπό συνθήκη πιθανότητα, είναι η πιθανότητα του A δεδομένου του B να είναι αληθής.
- P(B/A), είναι η πιθανότητα του B δεδομένου του A να είναι αληθής.

[16]

Το Θεώρημα Bayes συσχετίζει την τρέχουσα πιθανότητα με την αρχική πιθανότητα. Μπορεί δηλαδή να υπολογιστεί η πιθανότητα να συμβεί ένα γεγονός A δεδομένου ενός γεγονότος B. Όταν πρόκειται για πρόβλημα ταξινόμησης το γεγονός B αναφέρεται στην απόδειξη και το γεγονός A στην υπόθεση.

$$P(class/features) = \frac{P(class) * P(features/class)}{P(features)}$$

- P(class/features) : εκ των υστέρων πιθανότητα



Σχήμα 3.3: Η κατανομή Gauss (Κανονική κατανομή)

- P(class) : εκ των προτέρων πιθανότητα κλάσης
- P(features/class) : Δεσμευμένη Πιθανότητα
- P(features) : εκ των υστέρων πιθανότητα ταξινομητή

Η P(class/features) καλείται συνάρτηση πιθανοφάνειας της κλάσης σε σχέση με το χαρακτηριστικό και χρησιμοποιείται για να δηλώσει ότι η κατηγορία για την οποία η συνάρτηση έχει μεγάλη τιμή έχει μεγαλύτερη πιθανότητα να είναι η σωστή κατηγορία αναφορικά με την πρόβλεψη. Να σημειωθεί ότι το γινόμενο της πιθανοφάνειας και της εκ των προτέρων πιθανότητας είναι αυτό που καθορίζει την τιμή της εκ των υστέρων πιθανότητας.

Η βιβλιοθήκη Scikit-learn της Python προσφέρει τρεις τύπους μοντέλων Naïve Bayes:

1. Multinomial Naïve Bayes: Το συγκεκριμένο μοντέλο χρησιμοποιείται κατα κόρον σε προβλήματα που αφορούν την ταξινόμηση κειμένου σε κατηγορίες που το χαρακτηρίζουν.

2. Bernoulli Naïve Bayes: Η διαφοροποίηση του με το Multinomial Naïve Bayes έγκειται στο οτι οι τιμές πρόβλεψης είναι αυστηρά δυαδικές(boolean). Αυτό σημαίνει οτι για κάθε δεδομένο η απάντηση του αλγορίθμου είναι της μορφής ΝΑΙ/ΟΧΙ, κάνοντας το μοντέλο χρήσιμο σε αντίστοιχες εφαρμογές όπως, για παράδειγμα, στην αξιολόγηση ενός κειμένου ως υβριστικού ή όχι.

3. Gaussian Naïve Bayes: Στα gaussian μοντέλα, όπως υποδηλώνει και το όνομα, οι προβλέψεις για κάθε κλάση ακολουθούν την κατανομή Gauss (Σχήμα 3.3). Επομένως οι προβλέψεις εκφράζονται με συνεχείς και όχι διακριτές τιμές.

. Η συνάρτηση πυκνότητας πιθανότητας της κανονικής κατανομής είναι:

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\,\pi\sigma^2}}\, ^{-\frac{(x-\mu)^2}{2\,\sigma^2}} \tag{3.1}$$

Όπου

- $'\mu'$ είναι η μέση τιμή της μεταβλητής
- $'\sigma'$ είναι η τυπική απόκλιση
- $'\sigma^2'$ είναι η διακύμανση.

3.5.3 Supported Vector Machine

Το πρωταρχικό μοντέλο του αλγορίθμου SVM (Supported Vector Machine) δημιουργήθηκε από τους Vladimir Vapnik και Alexey Chervonenkis το 1963. Το 1992, προτάθηκε από τους Bernhard Boser, Isabelle Guyon και Vladimir Vapnik ένα πιο ισχυρό μοντέλο το οποίο χρησιμοποιείται για την ταξινόμηση δεδομένων σε κατηγορίες και αποτελεί μια τεχνική που χρησιμοποιείται ευρέως στην κατηγοριοποίηση κειμένου λόγω της μεγάλης αποδοτικότητας της. Ο λόγος που θεωρείται ένας απ τους καλύτερους κατηγοριοποιητές στην ταξινόμηση κειμένου είναι η ικανότητα του μοντέλου να διαχειρίζεται μεγάλα σύνολα χαρακτήρων, όπως είναι ένα κείμενο φυσικής γλώσσας.

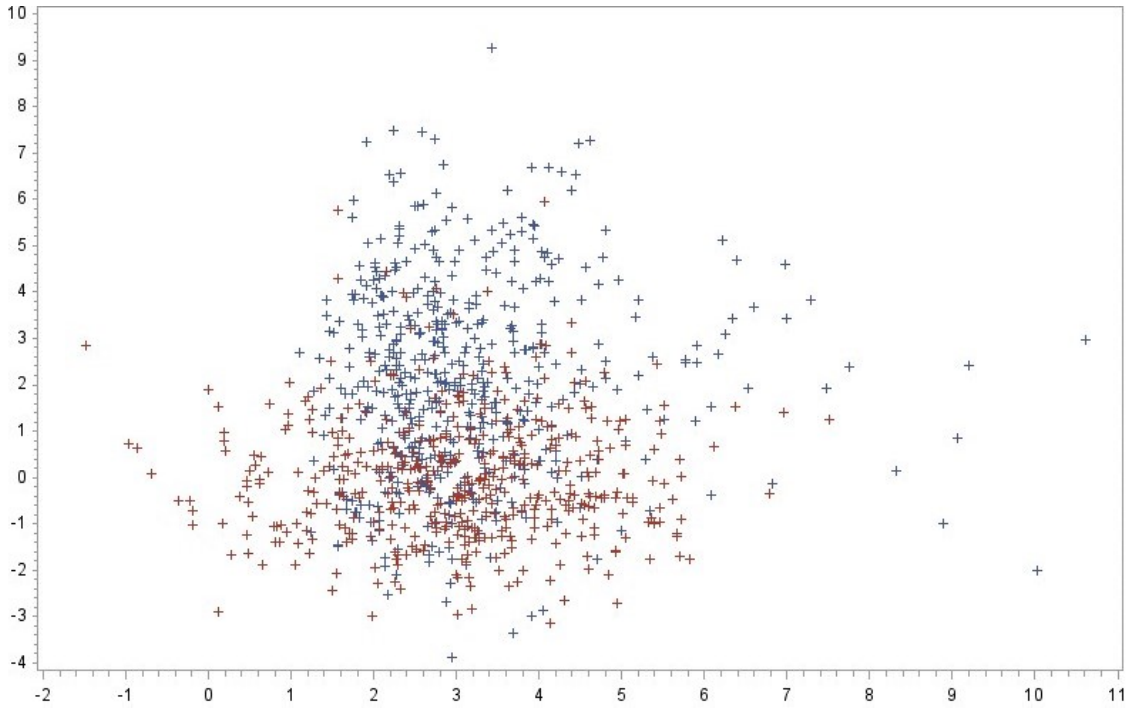
Ο αλγόριθμος λειτουργεί ως εξής:

Αρχικά παίρνει ως είσοδο το σύνολο δεδομένων εκπαίδευσης (training set) και δημιουργεί μία απεικόνιση του σε ένα πολυδιάστατο διανυσματικό χώρο. Σε αυτόν το χώρο προσπαθεί να εντοπίσει ένα πεδίο το οποίο να διαχωρίζει τα δύο υποσύνολα δεδομένων, ανάλογα με το αν ανήκουν ή όχι στην εκάστοτε κατηγορία, μεγιστοποιώντας την απόσταση ανάμεσα στο πεδίο και στα δεδομένα εκατέρωθεν του, όπως φαίνεται στο Σχήμα 3.4, όπου το διαχωριστικό πεδίο απεικονίζεται με πράσινο χρώμα και ο διαχωρισμός των δύο υποσυνόλων είναι ευδριάκριτος.

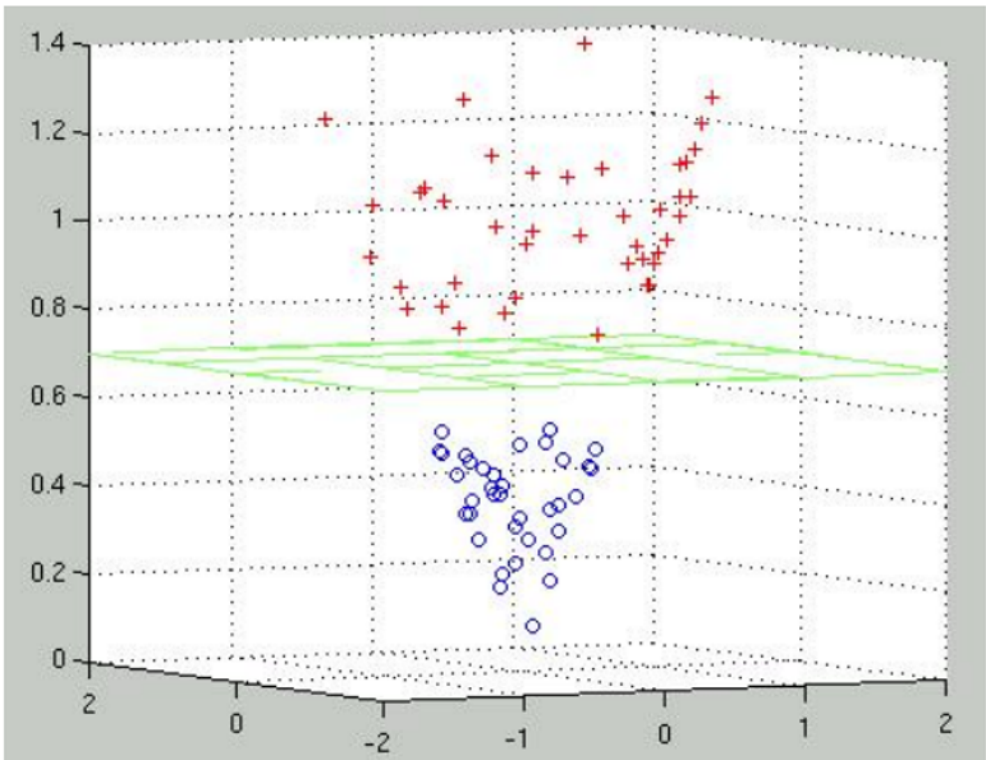
Χρησιμοποιώντας έπειτα αυτό το πεδίο, το οποίο ονομάζεται hyperplane., ο αλγόριθμος είναι ικανός να προβλέψει αποτελεσματικά την κατηγορία στην οποία ανήκει ένα άγνωστο μέχρι στιγμής δεδομένο εισόδου χαρτογραφώντας το στο χώρο και αποφασίζει σε ποια μεριά του διαχωριστικού πεδίου ανήκει.

Ο τρόπος του αλγορίθμου να επιλέξει ένα από τα πολλά διαχωριστικά πεδία που θα μπορούσαν να δημιουργηθούν στο διανυσματικό χώρο είναι να επιλέξει το πεδίο με τη μεγαλύτερη απόσταση από τα δεδομένα του εκπαιδευτικού συνόλου, μειώνοντας έτσι τις πιθανότητες σφάλματος στην πρόβλεψη.

3.5.4 k Nearest Neighbors - kNN



Σχήμα 3.4: Scatter Plot: Παράδειγμα χαρτογράφησης δεδομένων σε χώρο δύο διαστάσεων



Σχήμα 3.5: Αλγόριθμος SVM: Παράδειγμα χαρτογράφησης δεδομένων στον πολυδιάστατο διανυσματικό χώρο

3.6 Νευρωνικά Δίκτυα (Neural Networks - NN)

Ομοίως με τον άνθρωπο, ένα νευρωνικό δίκτυο απαιτεί εκπαίδευση για να λειτουργήσει σωστά. Η εκπαίδευση περιλαμβάνει τον προσδιορισμό των κατάλληλων συντελεστών βαρών μεταξύ των συνάψεων του, η οποία επιτυγχάνεται μέσω αλγορίθμων που βοηθούν στην εκμάθηση του περιβάλλοντος και τη βελτίωση της απόδοσης του. Η εκπαίδευση ενός νευρωνικού δικτύου είναι μια σταδιακή διαδικασία, η οποία ακολουθεί καθορισμένους κανόνες. Επαναλαμβανόμενες διαδικασίες ρύθμισης των βαρών στις συνάψεις μεταξύ των επιπέδων επιτρέπουν στο νευρωνικό δίκτυο να αποκτήσει περισσότερη "γνώση". Με αυτόν τον τρόπο, το νευρωνικό δίκτυο μπορεί να διπλασιάσει τον όγκο των γνώσεων του, καθιστώντας το πιο αποτελεσματικό στο να ανταποκρίνεται στις απαιτήσεις του περιβάλλοντος στο οποίο επιδρά.

3.6.1 Βαθιά Μάθηση (Deep Learning)

Η τεχνολογία του deep learning είναι μια προηγμένη μέθοδος μηχανικής μάθησης, η οποία βασίζεται στη χρήση νευρωνικών δικτύων, μια κατηγορία αλγορίθμων με μακρά ιστορία που ξεκινά από τον αλγόριθμο Perceptron του Rosenblatt το 1957 [15] και φτάνει στις σύγχρονες τεχνικές. Στον τομέα της οπτικής αναγνώρισης, οι Geoffrey Hinton και Yann LeCun έχουν κάνει σημαντικές προσπάθειες στην εξέλιξη αυτής της τεχνολογίας.

Αυτά τα δίκτυα είναι σχεδιασμένα να αναλύουν και να εξάγουν χρήσιμες πληροφορίες από μεγάλα σύνολα δεδομένων, σε αντίθεση με τις παραδοσιακές μεθόδους που βασίζονται σε κανόνες και χειροκίνητη επεξεργασία δεδομένων. Μαθηματικά, τα νευρωνικά δίκτυα είναι μη γραμμικές συναρτήσεις που εκτελούν πολλαπλά επίπεδα επεξεργασίας για την εξαγωγή συμπερασμάτων από τα δεδομένα εισόδου. Η μέθοδος αυτή επιτρέπει στα νευρωνικά δίκτυα να εξάγουν χαρακτηριστικά χωρίς τη χρήση ανθρώπινης παρέμβασης στην επεξεργασία των δεδομένων. Μέσω της επιβλεπόμενης μάθησης, τα νευρωνικά δίκτυα μπορούν να αναγνωρίσουν πρότυπα και σχέσεις στα δεδομένα εισόδου.

Για παράδειγμα, μπορούν να ταξινομήσουν εικόνες σε διαφορετικές κατηγορίες χρησιμοποιώντας ετικέτες. Σε αυτό το παράδειγμα, η μηχανική μάθηση μπορεί να χρησιμοποιηθεί για να εκπαιδεύσει ένα μοντέλο να αναγνωρίζει εάν μια εικόνα περιέχει ζώο ή άνθρωπο, βασιζόμενο σε ένα σύνολο εκπαίδευσης που περιλαμβάνει εικόνες ταξινομημένες ως απεικόνιση ανθρώπου ή απεικόνιση ζώου. Ενώ για τον ανθρώπινο εγκέφαλο φαντάζει εύκολο να διαχωριστεί η εικόνα ενός ανθρώπου από εκείνη ενός άλλου ζώου, είναι αρκετά δύσκολο να προσδιορίσει κανείς εν είδει μιας ακολουθίας βημάτων (αυτό που στην επιστήμη της πληροφορικής ονομάζεται αλγόριθμος) την διαδικασία που ακολουθεί η λογική σκέψη του για να κάνει αυτήν την ταξινόμηση.

Το μοντέλο μηχανικής μάθησης μπορεί να εξάγει αυτόματα χαρακτηριστικά από τις εικόνες, όπως γωνίες και συνήθως χρησιμοποιείται ένα νευρωνικό δίκτυο για να εκπαιδευτεί σε αυτά τα χαρακτηριστικά. Στη συνέχεια, το μοντέλο μπορεί να χρησιμοποιηθεί για να ταξινομήσει νέες εικόνες, βασιζόμενο στις παραμέτρους με τις οποίες έχει εκπαιδευτεί.

Η μηχανική μάθηση λοιπόν, χρησιμοποιείται για να δώσει σε έναν υπολογιστή την

ικανότητα να αναγνωρίζει πρότυπα και να προβλέπει αποτελέσματα με βάση τα δεδομένα που έχει "δει". Στην ουσία, ο υπολογιστής εκπαιδεύεται με ένα σύνολο δεδομένων εισόδου-εξόδου και μαθαίνει να αναγνωρίζει συσχετίσεις μεταξύ αυτών των δεδομένων, ώστε να μπορεί να κάνει προβλέψεις για νέα δεδομένα.

Το επιθυμητό αποτέλεσμα της επιβλεπόμενης μάθησης επιτυγχάνεται μέσω της βελτιστοποίησης και παραμετροποίησης μιας πολυπαραγοντικής ευέλικτης συνάρτησης, η οποία θα πρέπει να αφομοιώνει και να αναγνωρίζει τις ετικέτες μιας σειράς ετικετοποιημένων δεδομένων (labelled dataset), γνωστή ως σετ δεδομένων εκπαίδευσης (training dataset). Ο αλγόριθμος αναζητά μέσα στο χώρο όλων των πιθανών τιμών των παραμέτρων με στόχο να προσδιορίσει τη βέλτιστη συνάρτηση η οποία θα ταιριάζει με τα ετικετοποιημένα παραδείγματα με τη μεγαλύτερη δυνατή ακρίβεια. Στη συνέχεια θα χρησιμοποιήσει τη συνάρτηση αυτή για να ταξινομήσει νέες άγνωστες εικόνες που θα δοθούν προς ταξινόμηση.

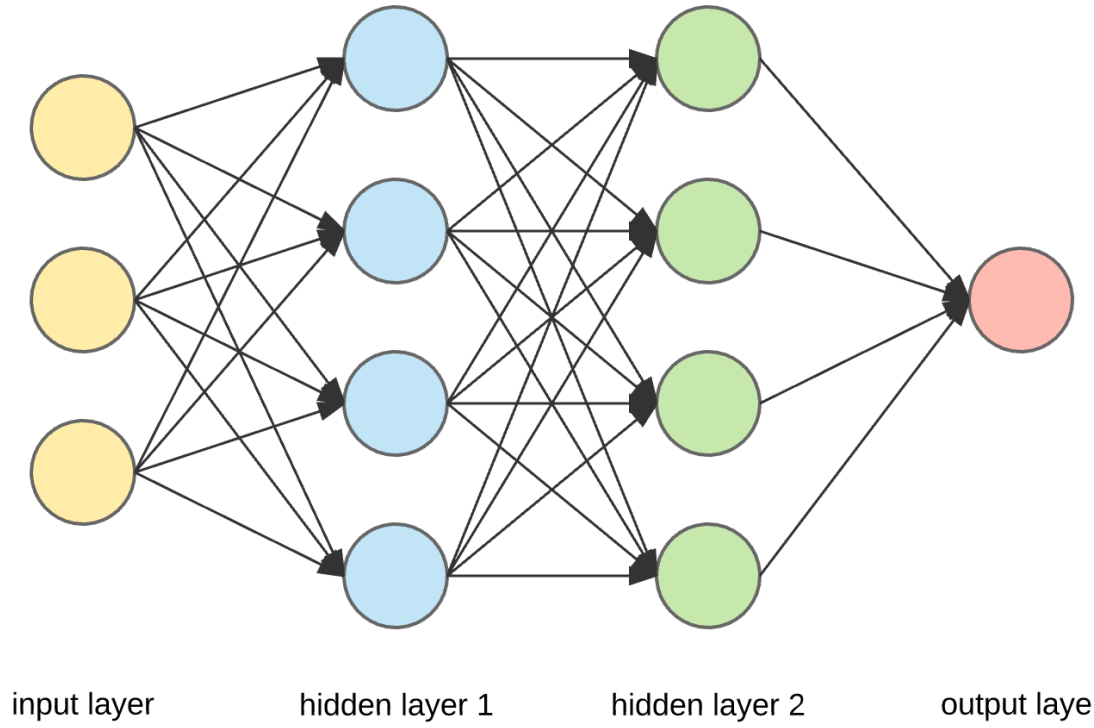
Ουσιαστικά, τα νευρωνικά δίκτυα είναι τόσο ευέλικτα που μπορούν να προσεγγίσουν μαθηματικά οποιαδήποτε συνάρτηση, χρησιμοποιώντας ένα μεγάλο αριθμό απλών συναρτήσεων που συνδέονται με διάφορους τρόπους. Μπορούν να συνδεθούν σειριακά ή παράλληλα, με τη μία μετασχηματιστική λειτουργία να αποτελεί την είσοδο για την επόμενη ή με πολλές μετασχηματιστικές λειτουργίες να εφαρμόζονται παράλληλα στα δεδομένα και το αποτέλεσμα να συνδυάζεται. Μια από τις βασικές τεχνικές της μηχανικής μάθησης είναι η υπερβολική προσαρμογή (overfitting), όπου το μοντέλο εκπαιδεύεται να ανταποκρίνεται πολύ στα δεδομένα εκπαίδευσης και δεν μπορεί να γενικεύσει σε νέα δεδομένα. Για να αποφευχθεί αυτό, χρησιμοποιούνται τεχνικές όπως η απόσυρση (dropout) και η κανονικοποίηση (regularization).

Αν και ιστορικά η Βαθιά Μάθηση ξεκινά να υπάρχει με την εμφάνιση του αλγόριθμου «perceptron», σήμερα χρησιμοποιείται ευρέως ως σύχγρονη τεχνική σε διάφορες εφαρμογές. Η Βαθιά Μάθηση είναι το πεδίο της Μηχανικής Μάθησης που χρησιμοποιεί πιο σύνθετες μορφές μοντέλων, διαχωρισμένων σε πολλά ιεραρχικά επίπεδα. Πιο συγκεκριμένα, αυτά τα μοντέλα ονομάζονται «Βαθιά νευρωνικά δίκτυα», ξεκινούν από ένα επίπεδο δεδομένων εισόδου, το οποίο στη συνέχεια περνά ξεχωριστά από αρκετά «κρυφά» επίπεδα αλλοιώντας κάθε φορά τα χαρακτηριστικά των δεδομένων εισόδου. Με αυτόν τον τρόπο, το μοντέλο αναπτύσσεται ως προς την κατανόηση των δεδομένων εισόδου, με αποτέλεσμα τη σταδιακή βελτίωση του αλγόριθμου ως προς νέα διαθέσιμα δεδομένα.

Ουσιαστικά, η υπολογιστική μηχανή αποκτά τη δυνατότητα να λειτουργεί με παρόμοιο τρόπο, όπως ο ανθρώπινος εγκέφαλος, να πράττει, δηλαδή, μαθαίντας φυσικά μέσα από παραδείγματα. Με τους αλγόριθμους της Βαθιάς Μάθησης, οι συνδέσεις των δεδομένων γίνονται με τη μορφή «δένδρου» σε πολλαπλά ιεραχικά επίπεδα, γεγονός που θυμίζει τις νευρωνικές συνδέσεις του ανθρώπινου εγκεφάλου, ο οποίος επίσης χρησιμοποιεί απλές υπολογιστικές μονάδες, γνωστές ως νευρώνες, για να πραγματοποιεί πολύπλοκους υπολογισμούς. Ωστόσο, είναι σημαντικό να σημειωθεί ότι τα νευρωνικά δίκτυα δεν αποτελούν ακριβή αντιγραφή του ανθρώπινου εγκεφάλου και οι σύγχρονες επιλογές αρχιτεκτονικών και συναρτήσεων έχουν σχεδιαστεί χωρίς να περιορίζονται στο να αντικατοπτρίζουν πλήρως τον τρόπο λειτουργίας του. [18], [7]

3.6.2 Μοντέλο Νευρωνικού Δικτύου (Neural Network Model)

Ένα νευρωνικό δίκτυο αποτελείται από πολλά μικρά υποσυστήματα που ονομάζονται νευρώνες. Κάθε νευρώνας λαμβάνει είσοδο από άλλους νευρώνες ή από το περιβάλλον και παράγει μια έξοδο, η οποία μπορεί να χρησιμοποιηθεί ως είσοδος σε άλλους νευρώνες.



Σχήμα 3.6: Απλό μοντέλο νευρωνικού δικτύου που αποτελείται επίπεδα εισόδου, επίπεδα εξόδου και κρυφά επίπεδα.

Κάθε νευρώνας έχει ένα σύνολο βαρών που αντιστοιχεί στις εισόδους του, καθώς και μια συνάρτηση ενεργοποίησης, η οποία καθορίζει το επίπεδο εξόδου που θα παραχθεί, δεδομένων των εισόδων και των σχετικών βαρών.

Τα νευρωνικά δίκτυα είναι οργανωμένα σε επίπεδα (layers), όπου κάθε στρώμα αποτελείται από πολλούς νευρώνες που επεξεργάζονται την ίδια είσοδο. Τα επίπεδα στα νευρωνικά δίκτυα συνήθως χωρίζονται σε τρία είδη: επίπεδο εισόδου (input layer), κρυφό επίπεδο (hidden layer) και επίπεδο εξόδου (output layer).

Το επίπεδο εξόδου αποτελείται από μια ή περισσότερες μονάδες που παράγουν την τελική έξοδο του δικτύου, δηλαδή την πρόβλεψη ή την ταξινόμηση του δεδομένου εισόδου.

Η διαδικασία εκπαίδευσης του νευρωνικού δικτύου συνήθως περιλαμβάνει τη ρύθμιση των βαρών των συνδέσμων μεταξύ των νευρώνων, ώστε να ελαχιστοποιηθεί η απόκλιση της έξοδου του δικτύου από την επιθυμητή έξοδο. Η διαδικασία αυτή γίνεται μέσω της χρήσης ενός αλγορίθμου βελτιστοποίησης, όπως η μέθοδος των ελαχίστων τετραγώνων ή η μέθοδος της ανάδρασης προς τα πίσω (backpropagation).

Κεφάλαιο 4

Πρακτικό Μέρος

4.1 Εργαλεία που χρησιμοποιήθηκαν

Η υλοποίηση αυτής της διπλωματικής έγινε χρησιμοποιώντας την γλώσσα προγραμματισμού Python. Παράλληλα, χρησιμοποιήθηκε ένα πολύ ισχυρό εργαλείο για ανάπτυξη κώδικα που ονομάζεται Jupyter Notebook.

Η Python είναι μία από τις πιο δημοφιλείς γλώσσες προγραμματισμού για τον χειρισμό δεδομένων και τον αναλυτικό προγραμματισμό (data analytics). Υπάρχουν διάφοροι λόγοι για αυτό, μερικοί εκ των οποίων είναι:

- Είναι εύκολη στην εκμάθηση: Η Python είναι μια ανθρώπινη γλώσσα προγραμματισμού, η οποία σημαίνει ότι είναι σχετικά εύκολο για τους ανθρώπους να την κατανοήσουν. Αυτό την καθιστά ιδανική για χρήση σε επιστημονικά εργαστήρια όπου οι ερευνητές δεν είναι απαραίτητα ειδικοί στον προγραμματισμό.

- Διαθέτει εκτεταμένη βιβλιοθήκη: Η Python διαθέτει μια εκτεταμένη βιβλιοθήκη που περιλαμβάνει πολλά εργαλεία για τον χειρισμό των δεδομένων. Αυτό σημαίνει ότι οι ερευνητές μπορούν να χρησιμοποιήσουν ένα ευρύ φάσμα εργαλείων και λειτουργιών για την ανάλυση των δεδομένων τους. Αυτά τα πακέτα και βιβλιοθήκες περιλαμβάνουν το NumPy για επιστημονικούς υπολογισμούς, το Pandas για ανάλυση δεδομένων και επεξεργασία, το Matplotlib για οπτικοποίηση δεδομένων, και πολλά άλλα.

- Η Python είναι μια γλώσσα ανοιχτού κώδικα, η οποία σημαίνει ότι ο κώδικας είναι διαθέσιμος για όλους να τον χρησιμοποιήσουν και να τον βελτιώσουν. Αυτό έχει ως αποτέλεσμα τη δημιουργία μιας πολύ ενεργής και ανοικτής κοινότητας που συνεργάζεται για να δημιουργήσει και να βελτιώσει τις βιβλιοθήκες και τα πακέτα που χρησιμοποιούνται στην ανάλυση δεδομένων.

Το Jupyter Notebook είναι μια διαδραστική πλατφόρμα ανάπτυξης λογισμικού, που χρησιμοποιείται για την εκτέλεση κώδικα σε πολλές γλώσσες προγραμματισμού, όπως η Python, η R, η Julia και άλλες. Επιτρέπει στους χρήστες να δημιουργούν και να κοινοποιούν ένα αρχείο σε μορφή notebook που περιέχει κώδικα, αποτελέσματα, εξηγήσεις, εικόνες, γραφήματα και πίνακες.

Τα notebooks αποτελούνται από σελίδες, οι οποίες μπορούν να περιέχουν κελιά κώδικα και κελιά κειμένου. Τα κελιά κώδικα περιέχουν τον κώδικα που εκτελείται σε πραγματικό χρόνο και παράγει τα αποτελέσματα, ενώ τα κελιά κειμένου χρησιμοποιούνται για να παρέχουν εξηγήσεις, οδηγίες ή σχόλια σχετικά με τον κώδικα. Το Jupyter Notebook επιτρέπει στους χρήστες να επεξεργάζονται τα κελιά κώδικα και κειμένου σε πραγματικό χρόνο, επιτρέποντας την αμφίδρομη επικοινωνία με τον κώδικα και τα αποτελέσματα που παράγει.

4.2 Το dataset

Στο κεφάλαιο αυτό αρχικά γίνεται μια περιγραφή το συνόλου δεδομένων (dataset) που χρησιμοποιήθηκε για την εκπόνηση της παρούσας εργασίας.

Πηγή dataset

Το θέμα του ιατρικού απορρήτου καθιστά τη συλλογή και διάθεση ιατρικών δεδομένων δύσκολη διαδικασία. Για το λόγο αυτό, είναι γνωστό στην επιστημονική κοινότητα ότι αποτελούν σπάνια και δύσκολα προσβάσιμη πληροφορία. Το συγκεκριμένο dataset αντλήθηκε από χρήστες της πλατφόρμας kaggle [10], με την κύρια πηγή άντλησης να είναι η ιστοσελίδα www.mtsamples.com. [11] Φυσικά, όλες οι εγγραφές είναι απολύτως ανώνυμες. Η πρόκληση του kaggle είναι η επιτυχής πρόβλεψη της ιατρικής ειδικότητας κάθε περιγραφής ιατρικού περιστατικού.

Δομή dataset

Η αρχική μορφή του dataset είναι ένα αρχείο csv με πέντε χιλιάδες (5000) εγγραφές και έξι (6) πεδία και η δομή του έχει ως εξής:

1. A/A - (Αύξων αριθμός εγγραφής)
2. description - (Περιγραφή)
3. medical specialty - (Ιατρική ειδικότητα)
4. sample name - (Όνομα δείγματος)
5. transcription - (Ιατρική συνταγογράφηση)
6. keywords - (Λέξεις - κλειδιά)

0	description	medical_specialty	sample_name	transcription	keywords
0	A 23-year-old white female presents with comp...	Allergy / Immunology	Allergic Rhinitis	SUBJECTIVE:, This 23-year-old white female pr...	allergy / immunology, allergic rhinitis, aller...
1	Consult for laparoscopic gastric bypass.	Bariatrics	Laparoscopic Gastric Bypass Consult - 2	PAST MEDICAL HISTORY:, He has difficulty climb...	bariatrics, laparoscopic gastric bypass, weigh...
2	Consult for laparoscopic gastric bypass.	Bariatrics	Laparoscopic Gastric Bypass Consult - 1	HISTORY OF PRESENT ILLNESS: , I have seen ABC ...	bariatrics, laparoscopic gastric bypass, heart...
3	2-D M-Mode. Doppler.	Cardiovascular / Pulmonary	2-D Echocardiogram - 1	2-D M-MODE: , ,1. Left atrial enlargement wit...	cardiovascular / pulmonary, 2-d m-mode, dopple...
4	2-D Echocardiogram	Cardiovascular / Pulmonary	2-D Echocardiogram - 2	1. The left ventricular cavity size and wall ...	cardiovascular / pulmonary, 2-d, doppler, echo...

Σχήμα 4.1: Οι πέντε (5) πρώτες στήλες του dataset

Το Σχήμα 4.1 απεικονίζει ενδεικτικά τις πέντε (5) πρώτες στήλες του dataset πριν υποστεί οποιαδήποτε επεξεργασία, όπως γίνεται η εισαγωγή του αρχείου csv στο πρόγραμμα.

Σε πρώτο στάδιο ανάλυσης του συνόλου δεδομένων και με χρήση κατάλληλων εντολών python εξάγονται πληροφορίες για τον αριθμό των κατηγοριών (Ιατρικών ειδικοτήτων), το όνομα κάθε κατηγορίας καθώς και το πλήθος εγγραφών που περιέχει.

Το Σχήμα 4.2 εμφανίζει τις σαράντα (40) κατηγορίες που περιέχει αρχικά το dataset καθώς και το πλήθος εγγραφών ανά κατηγορία.

Επίσης, δημιουργείται η συνάρτηση get_sentence_word_count(), η οποία μας επιστρέφει το πλήθος των διακριτών λέξεων και προτάσεων που υπάρχουν στο πεδίο transcription του εισαγόμενου αρχείου. Έτσι έχουμε μια πιο ολοκληρωμένη εικόνα για τη μορφή, τη δομή και την κατανομή των δεδομένων στο αρχείο.

Το επόμενο στάδιο που περιγράφεται στο κεφάλαιο 3 είναι το στάδιο της Προεπεξεργασίας. Η σωστή και καθαρή απεικόνιση των δεδομένων βοηθάει τον ερευνητή να εφαρμόσει πιο στοχευμένες και αποτελεσματικές τεχνικές προεπεξεργασίας κειμένου ώστε να αποφύγει την απώλεια χρήσιμης πληροφορίας.

```
Number of sentences in transcriptions column: 140214
Number of unique words in transcriptions column: 35805
===========Original Categories ======================
Cat:1  Allergy / Immunology : 7
Cat:2  Autopsy : 8
Cat:3  Bariatrics : 18
Cat:4  Cardiovascular / Pulmonary : 371
Cat:5  Chiropractic : 14
Cat:6  Consult - History and Phy. : 516
Cat:7  Cosmetic / Plastic Surgery : 27
Cat:8  Dentistry : 27
Cat:9  Dermatology : 29
Cat:10  Diets and Nutritions : 10
Cat:11  Discharge Summary : 108
Cat:12  ENT - Otolaryngology : 96
Cat:13  Emergency Room Reports : 75
Cat:14  Endocrinology : 19
Cat:15  Gastroenterology : 224
Cat:16  General Medicine : 259
Cat:17  Hematology - Oncology : 90
Cat:18  Hospice - Palliative Care : 6
Cat:19  IME-QME-Work Comp etc. : 16
Cat:20  Lab Medicine - Pathology : 8
Cat:21  Letters : 23
Cat:22  Nephrology : 81
Cat:23  Neurology : 223
Cat:24  Neurosurgery : 94
Cat:25  Obstetrics / Gynecology : 155
Cat:26  Office Notes : 50
Cat:27  Ophthalmology : 83
Cat:28  Orthopedic : 355
Cat:29  Pain Management : 61
Cat:30  Pediatrics - Neonatal : 70
Cat:31  Physical Medicine - Rehab : 21
Cat:32  Podiatry : 47
Cat:33  Psychiatry / Psychology : 53
Cat:34  Radiology : 273
Cat:35  Rheumatology : 10
Cat:36  SOAP / Chart / Progress Notes : 166
Cat:37  Sleep Medicine : 20
Cat:38  Speech - Language : 9
Cat:39  Surgery : 1088
Cat:40  Urology : 156
==================================
```

Σχήμα 4.2: Η αρχική κατανομή των δεδομένων στις κατηγορίες

4.3 Προεπεξεργασία (Preprocessing)

Στο κεφάλαιο αυτό παρουσιάζεται αναλυτικά το στάδιο της Προεπεξεργασίας (Preprocessing). Θα παρουσιαστούν εκτενώς όλες οι μέθοδοι και οι τεχνικές που χρησιμοποιήθηκαν, ο σκοπός τους και η τελική μορφή των δεδομένων.

Είναι σημαντικό να γίνει κατανοητή η σημαντικότητα αυτού του σταδίου για την ακρίβεια των τελικών αποτελεσμάτων. Το αρχικό μας σύνολο δεδομένων περιέχει δύο είδη πληροφορίας. Τη χρήσιμη για τον ερευνητικό σκοπό και όλη την υπόλοιπη η οποία, εάν δεν απομονωθεί σωστά, ενδέχεται να προκαλέσει αλλοίωση των αποτελεσμάτων, εκπαίδευση του συστήματος προς λανθασμένη κατεύθυνση και σίγουρα ανώφελη κατανάλωση υπολογιστικής ισχύος αλλά και χρόνου.

Σε αυτό το στάδιο ο ερευνητής πρέπει να παρατηρήσει σωστά την πληροφορία και να κρατήσει το ωφέλιμο κομμάτι με σκοπό να έχει πιο ποιοτικά δεδομένα, προσέχοντας στην προσπάθεια αυτήν να μην απωλέσει κομμάτι χρήσιμης πληροφορίας που θα εκπαίδευε σωστά το σύστημα.

4.3.1 Βήμα 1: Μείωση κατηγοριών και απομόνωση πεδίων

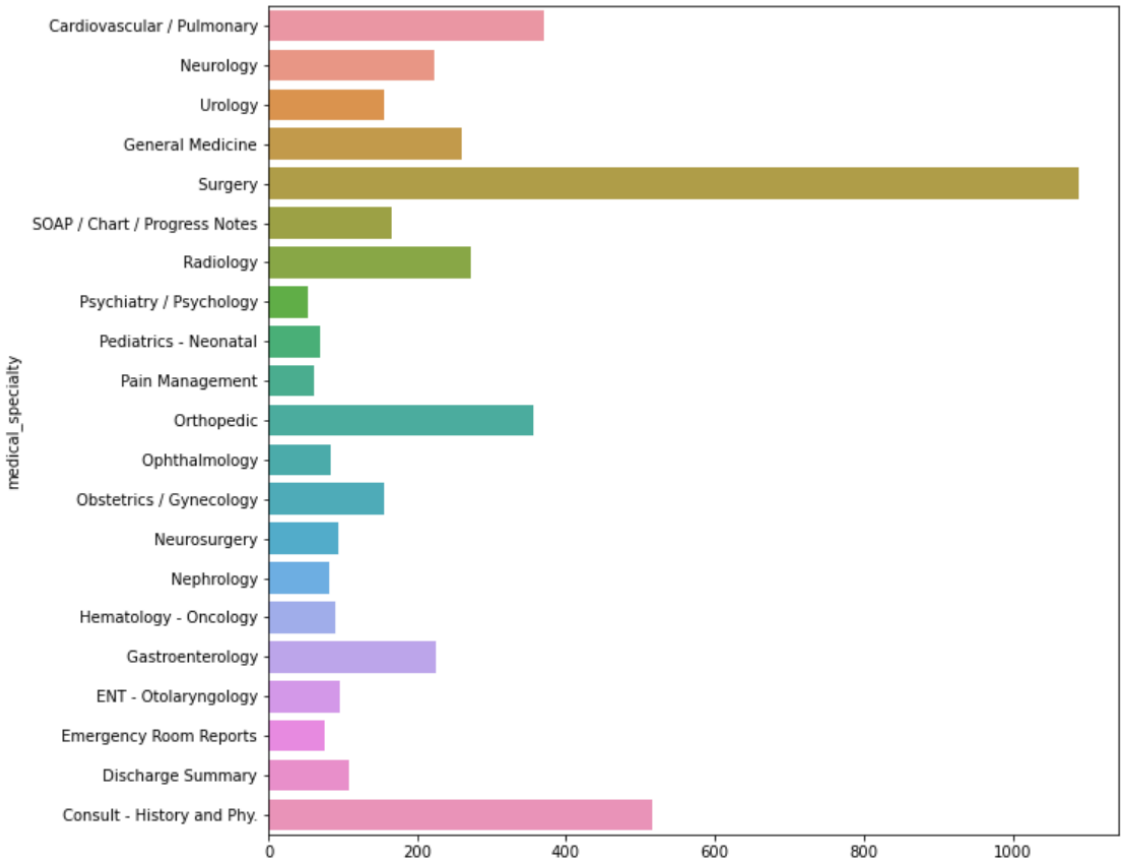
Είδαμε οτι στο συγκεκριμένο dataset υπάρχει μεγάλη ανομοιομορφία ως προς την κατανομή των δεδομένων στις διάφορες κατηγορίες. Αυτό αποτελεί ένα μείζον πρόβλημα στους αναλυτές δεδομένων, γνωστό ως Imbalanced Data, το οποίο οδήγησε την κοινότητα στην ανάπτυξη διάφορων τεχνικών αντιμετώπισης του. Το μεγαλύτερο πρόβλημα έγκειται στο ότι οι αλγόριθμοι ταξινόμησης υποθέτουν οτι όλες οι κλάσεις έχουν ίδο πλήθος δεδομένων και εκπαιδεύονται εξ΄ίσου σε όλες.

Επομένως, στο πρώτο βήμα, εξαλείφουμε τις μειονότητες με σκοπό να οδηγηθούμε σε ένα κάπως πιο ισορροπημένο σύνολο δεδομένων. Πρατικά, φιλτράρουμε τα δεδομένα έτσι ώστε να εξαλειφθούν οι κατηγορίες που έχουν λιγότερα απο 50 στοιχεία.

Εμφανίζουμε ξανά το πλήθος των κατηγοριών και των εγγραφών ανά κατηγορία, καθώς και τη γραφική απεικόνιση σε διάγραμμα μέσω της βιβλιοθήκης matplotlib. Τα αποτελέσματα φαίνονται στα Σχήματα 4.3 και 4.4

```
============Reduced Categories ======================
Cat:1  Cardiovascular / Pulmonary : 371
Cat:2  Consult - History and Phy. : 516
Cat:3  Discharge Summary : 108
Cat:4  ENT - Otolaryngology : 96
Cat:5  Emergency Room Reports : 75
Cat:6  Gastroenterology : 224
Cat:7  General Medicine : 259
Cat:8  Hematology - Oncology : 90
Cat:9  Nephrology : 81
Cat:10  Neurology : 223
Cat:11  Neurosurgery : 94
Cat:12  Obstetrics / Gynecology : 155
Cat:13  Ophthalmology : 83
Cat:14  Orthopedic : 355
Cat:15  Pain Management : 61
Cat:16  Pediatrics - Neonatal : 70
Cat:17  Psychiatry / Psychology : 53
Cat:18  Radiology : 273
Cat:19  SOAP / Chart / Progress Notes : 166
Cat:20  Surgery : 1088
Cat:21  Urology : 156
============ Reduced Categories ======================
```

Σχήμα 4.3: Το πλήθος εγγραφών ανά κατηγορία μετά την εξάλειψη των μειονοτήτων.



Σχήμα 4.4: Γραφικη απεικόνιση του σχήματος 4.1

Είδαμε ότι το σύνολο των δεδομένων αποτελείται απο έξι πεδία εκ των οποίων μας αφορούν μόνο δύο:

1) Το πεδίο medical_specialty το οποίο αναφέρεται στην κατηγορία/ειδικότητα στην οποία έγκειται το περιστατικό και

2) Το πεδίο transcription το οποίο είναι το κυρίως κείμενο και αναφέρεται στην περιγραφή του περιστατικού.

Επομένως τα δεδομένα φιλτράρονται ξανά με σκοπό να κρατηθεί μόνο η πληροφορία αυτών των δύο πεδίων και στη συνέχεια αποθηκεύονται σε έναν πίνακα, εμφανίζοντας στο τέλος το πλήθος στηλών και γραμμών.

Τα δεδομένα πλέον έχουν την εξής μορφή : [4597 rows x 2 columns]

4.3.2 Βήμα 2: Καθαρισμός κειμένου (text cleaning)

Στη συνέχεια εκτελείται η διαδικασία καθαρισμού κειμένου (text cleaning process.) Αρχικά ορίζονται δύο συναρτήσεις: η clean_text() και η lemmatize_text() οι οποίες επιτελούν τις παρακάτω διεργασίες:

1. Διαίρεση σε όρους (Tokenization): Με αυτόν τον όρο αναφερόμαστε τον κατακερματισμό μιας συμβολοσειράς σε διακριτές λέξεις. Ο κάθε όρος μπορεί να είναι ολόκληρη λέξη, μεμονομένος χαρακτήρας, αριθμός, σύμβολο ή σημείο στίξης. [13]. Σε αυτήν την εργασία, η κάθε εγγραφή του πεδίου transcription, που αποτελούν το κείμενο προς επεξεργασία, χωρίζεται σε λέξεις. Η μορφή των δεδομένων τώρα είναι ένας πίνακας όπου το κείμενο παρουσιάζεται σε μορφή λίστας με διακριτούς όρους.

2. Αφαίρεση λέξεων χωρίς νοηματική αξία stopwords removal: Ένα από τα πιο συνηθισμένα βήματα προεπεξεργασίας κειμένου φυσικής γλώσσας είναι η αφαίρεση λέξεων χωρίς ιδιαίτερη νοηματική αξία. Η βιβλιοθήκη NLTK της Python προσφέρει αυτή τη δυνατότητα, απομακρύνοντας λέξεις που δεν αποδίδουν χρήσιμη πληροφορία, αντιθέτως προσθέτουν στο κείμενο θόρυβο και όγκο. Αυτό γίνεται είτε μέσω έτοιμης λίστας stopwords, είτε εντοπίζοντας λέξεις που εμφανίζονται εξαιρετικά συχνά σε ένα κείμενο. Τέτοιες λέξεις μπορεί να είναι άρθρα, αντωνυμίες, ακόμα και πολύ συνηθισμένα ρήματα.

3. Μετατροπή κεφαλαίων χαρακτήρων σε πεζούς: Εδώ γίνεται χρήση της μεθόδου .lower() που προσφέρει η python για μετατροπή συμβολοσειρών, η οποία μετατρέπει οποιοδήποτε κεφαλαίο χαρακτήρα σε πεζό.

4. Απομάκρυνση αριθμών και συμβόλων: Οι αριθμοί, τα σύμβολα και τα σημεία στίξης, για το σκοπό της εργασίας, αποτελούν περιττή πληροφορία. Απαλάσσοντας το κείμενο από αυτά μειώνεται ο όγκος εργασίας, η απαιτούμενη υπολογιστική ισχύς και παράγονται πιο ποιοτικά δεδομένα. Έτσι με χρήση κατάλληλων εντολών της python, όλα αυτά αφαιρούνται.

5. Stemming Ο όρος Stemming αναφέρεται σε μια διαδικασία που παρέχεται απο τη βιβλιοθήκη NLTK της Python η οποία ανάγει όλες τις ομόρριζες λέξεις του κειμένου στην αρχική τους ρίζα, αδιαφορώντας για τις διαφορετικές καταλήξεις που ενδεχομένως να έχουν. Για παράδειγμα, η λέξη coding ανάγεται στη λέξη code.

6. Λημματοποίηση (Lemmatization) Ο όρος λημματοποίηση αναφέρεται στη διαδικασία επεξεργασίας κειμένου φυσικής γλώσσας με σκοπό να φέρει την κάθε λέξη σε μια μορφή τέτοια, ώστε να υπάρχει όσο το δυνατόν μικρότερη ποικιλία στο συνολικό λεξιλόγιο, επιδιώκοντας την επιστροφή των λέξεων στο αρχικό τους λήμμα. Η βασική διαφορά των δύο τελευταίων διαδικασιών ειναι οτι στη λημματοποίηση εξετάζεται η γλωσσολογική προέλευση κάθε όρου ώστε να βρεθεί το κατάλληλο λήμμα. Αποτελεί μια πιο εξελιγμένη διαδικασία καθώς το stemming δεν εξετάζει πληροφορίες που αφορούν τη γλωσσολογική προέλευση της λέξης, αλλά μόνο τη σχέση ρίζας - κατάληξης. Χρησιμοποιώντας και πάλι τη βιβλιοθήκη NLTK της Python, γίνεται ομαδοποίηση των λημμάτων μέσω της μεθόδου WordNetLemmatizer() με βάση το λεξικό WordNet Για παράδειγμα, στο lemmatization, η λέξη better ανάγεται στη λέξη good.

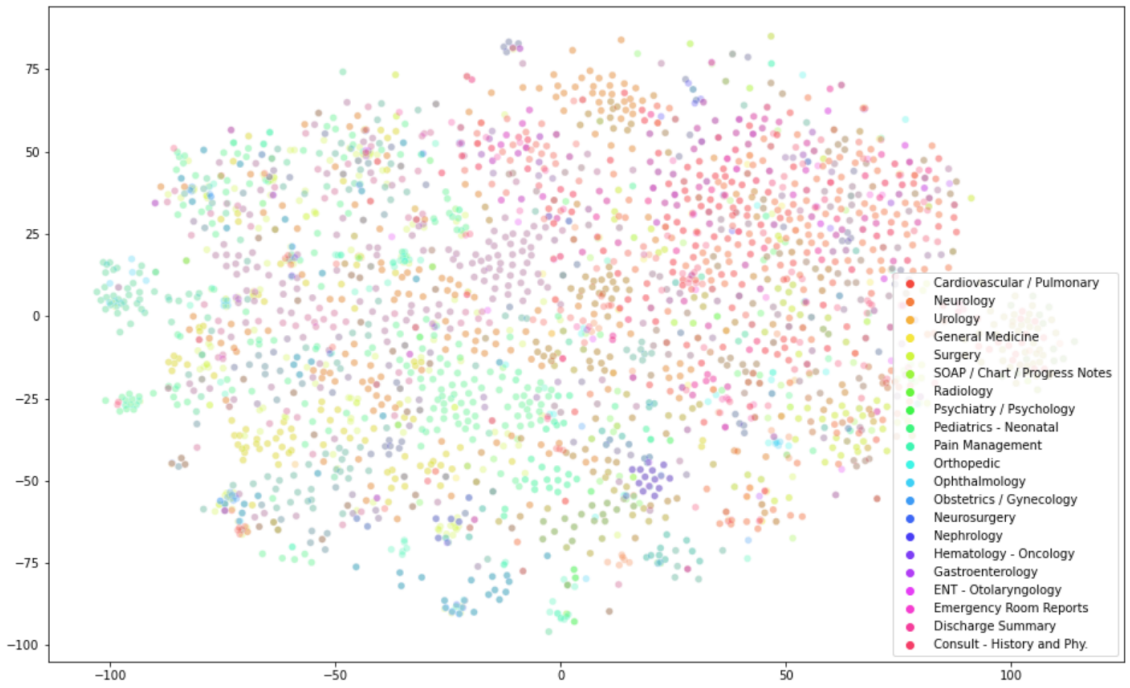
Σε αυτό το σημείο, η λίστα κειμένων είναι απαλλαγμένη από στοιχεία που προσδίδουν αχρείαστο όγκο, θόρυβο και περιττή πληροφορία. Στα επόμενα κεφάλαια θα χρησιμοποιηθεί το σύνολο δεδομένων με το ¨καθαρισμένο' κείμενο ως είσοδο στα προγράμματα με σκοπό την εκπαίδευση του συστήματος και την παραγωγή γνώσης.

4.4 Υλοποίηση Ταξινόμησης

Στο προηγούμενο κεφάλαιο φάνηκε αναλυτικά πως τα δεδομένα του dataset έφτασαν μέσω των διαφόρων διεργασιών της προεπεξεργασίας σε μια πιο ποιοτική και καθαρή μορφή. Για να φανεί χρήσιμη η πληροφορία αυτή, χρειάζεται το αδόμητο κείμενο να μετατραπεί σε κάποια κατανοητή μορφή για τους υπολογιστές, όπως πίνακες ή διανύσματα από χαρακτηριστικά (φεατυρες).

Αυτό γίνεται με χρήση της μεθόδου tf-idf, ο τρόπος λειτουργίας της οποίας αναλύθηκε σε προηγούμενο κεφάλαιο.

Έπειτα, το Σχήμα 4.5 εμφανίζει την απεικόνιση του πίνακα που προέκυψε από την tf-idf με χρήση της μεθόδου t-SNE όπου γίνεται εμφανές πως πολλές κατηγορίες-ειδικότητες αλληλοκαλύπτονται.



Σχήμα 4.5: Η γραφική απεικόνιση του αλγορίθμου t-SNE

Στη συνέχεια εφαρμόζεται η μέθοδος PCA στον πίνακα tf-idf με στόχο τη μείωση διαστατικότητας.

Η Sklearn (ή Scikit-learn) είναι μια βιβλιοθήκη της Python η οποία προσφέρει διάφορες δυνατότητες για επεξεργασία δεδομένων και χρησιμοποιείται συνήθως για ταξινόμηση, ομαδοποίηση και επιλογή μοντέλου.

Για να εκπαιδεύσουμε το μοντέλο χρησιμοποιώντας ένα συγκεκριμένο dataset θα πρέπει να το «τεστάρουμε» πάνω σε ένα δεύτερο dataset. Όταν έχουμε μόνο ένα, όπως στη δική μας περίπτωση, το χωρίζουμε στα δύο χρησιμοποιώντας τη μέθοδο train_test_split() της sklearn. Η μέθοδος αυτή χωρίζει το σύνολο των data arrays σε δύο υποσύνολα: training set (Σύνολο εκπαίδευσης) και test set (Σύνολο αξιολόγησης).

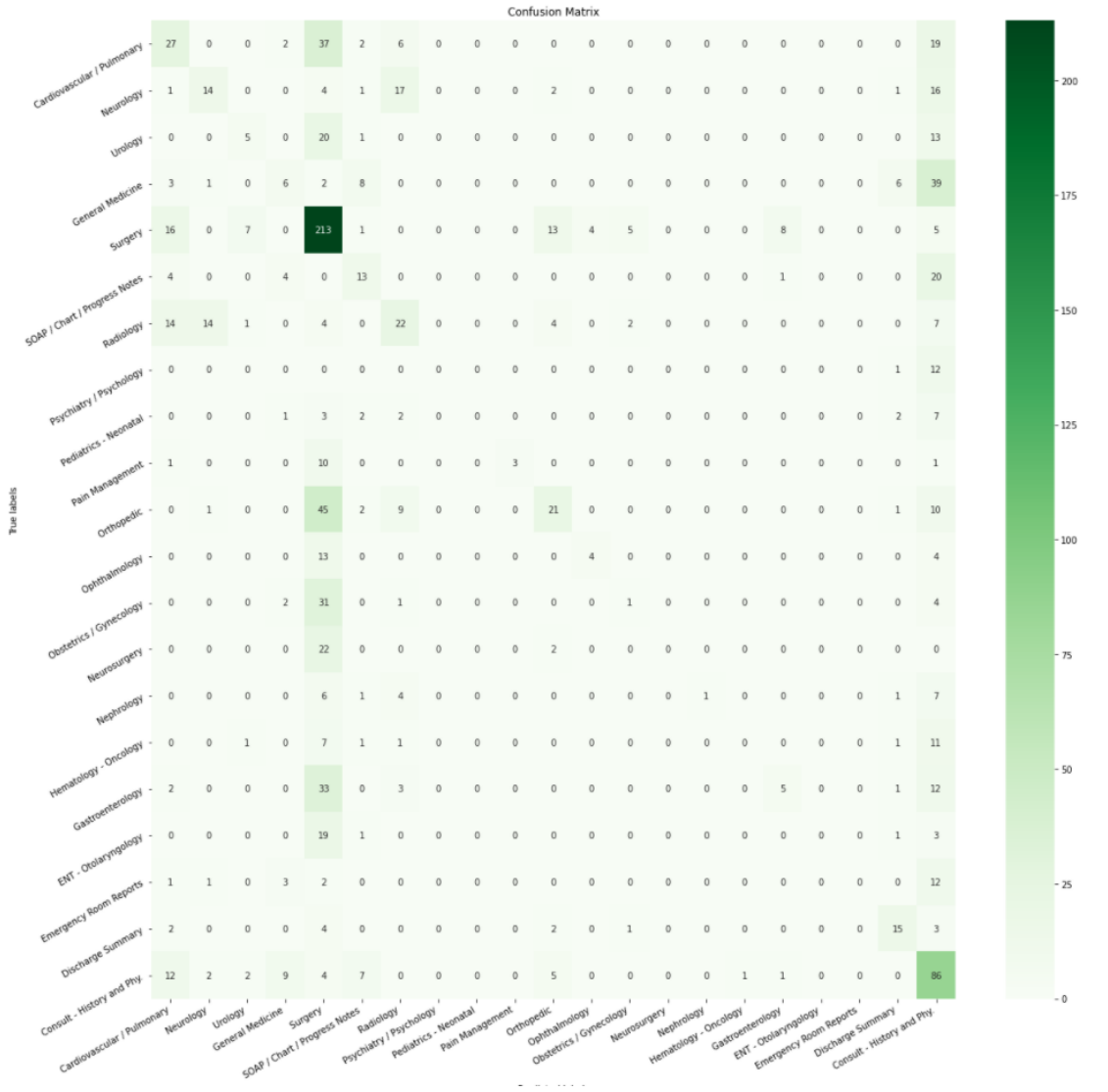
Έτσι, μετά την εφαρμογή του διαχωρισμού έχουμε:

- Train_Set_Size: (3447, 587)
- Test_Set_Size: (1150, 587)

Στη συνέχεια, εφαρμόζουμε στα δεδομένα Λογιστική Παλινδρόμηση (Logistic Regression) για να εκπαιδεύσουμε το μοντέλο στα training data και να κάνει την πρόβλεψη στα test data. Η εφαρμογή της λογιστικής παλινδρόμησης στη βιβλιοθήκη της Python scikit-learn μπορεί να προσεγγιστεί από την κλάση LogisticRegression.

Μετά απο αυτήν τη διαδικασία κατασκευάζουμε τον πίνακα σύγχυσης (confusion matrix). Πρόκειται για έναν MxM πίνακα, όπου το (i,j) στοιχείο του ισούται με το πλήθος των σημείων που, ενώ προέρχονται από την κλάση i, καταχωρούνται στην κλάση j. Δίνει πληροφορίες σχετικά με το αν κάποιες κλάσεις έχουν την τάση να συγχέονται με άλλες κλάσεις.

Από τη γραφική απεικόνιση του πίνακα σύγχυσης που φαίνεται στο Σχήμα 4.6, παρατηρούμε πως μεγαλύτερη σύγχυση υπάρχει σε συγκεκριμένες ειδικότητες όπως η χειρουργική, οι οποίες έχουν την ιδιότητα υπερκλάσης, αφού επικαλύπτονται με άλλες ειδικότητες εκ φύσεως.



Σχήμα 4.6: Η γραφική απεικόνιση του πίνακα σύγχυσης

Η βιβλιοθήκη sklearn προσφέρει τη μέθοδο classification_report() η οποία τυπώνει τις εκτιμήσεις διαφόρων μετρικών πάνω στην ποιότητα των αποτελεσμάτων πρόβλεψεις του αλγορίθμου.

Οι μετρικές αυτές είναι:

1. Ορθότητα (Accuracy)
2. Ανάκληση (Recall)
3. Ακρίβεια (Precision)
4. F-Measure

Τα πρώτα αποτελέσματα ακρίβειας ταξινόμησης ανά κατηγορία φαίνονται στον παρακάτω πίνακα (Σχήμα 4.7:

	precision	recall	f1-score	support
Cardiovascular / Pulmonary	0.33	0.29	0.31	93
Neurology	0.42	0.25	0.31	56
Urology	0.31	0.13	0.18	39
General Medicine	0.22	0.09	0.13	65
Surgery	0.44	0.78	0.57	272
SOAP / Chart / Progress Notes	0.33	0.31	0.32	42
Radiology	0.34	0.32	0.33	68
Psychiatry / Psychology	0.00	0.00	0.00	13
Pediatrics - Neonatal	0.00	0.00	0.00	17
Pain Management	1.00	0.20	0.33	15
Orthopedic	0.43	0.24	0.30	89
Ophthalmology	0.50	0.19	0.28	21
Obstetrics / Gynecology	0.11	0.03	0.04	39
Neurosurgery	0.00	0.00	0.00	24
Nephrology	1.00	0.05	0.10	20
Hematology - Oncology	0.00	0.00	0.00	22
Gastroenterology	0.33	0.09	0.14	56
ENT - Otolaryngology	0.00	0.00	0.00	24
Emergency Room Reports	0.00	0.00	0.00	19
Discharge Summary	0.50	0.56	0.53	27
Consult - History and Phy.	0.30	0.67	0.41	129
accuracy			0.38	1150
macro avg	0.31	0.20	0.20	1150
weighted avg	0.34	0.38	0.32	1150

Σχήμα 4.7: Πίνακας πρώτων αποτελεσμάτων μετρικών ταξινόμησης

Από το classification report παρατηρούμε ότι τα αποτελέσματα ακρίβειας είναι αρκετά χαμηλά. Αυτό οφείλεται στο οτι πολλές απο τις κατηγορίες αλληλοεπικαλύπτονται, όπως φαίνεται και στον πίνακα σύγχυσης. Επειδή αυτό συμβαίνει λόγω της φύσης των δεδομένων, καθώς σε κάποια περιστατικά δε γίνεται να αποφανθεί οτι εμπίπτουν μόνο σε μια συγκεκριμένη ιατρική ειδικότητα, στο επόμενο βήμα αφαιρούνται απο το σύνολο των δεδομένων οι ειδικότητες που δημιουργούν αυτόν το «θόρυβο».

Οι ειδικότητες που απαλείφονται είναι:

1. Surgery
2. SOAP / Chart / Progress Notes
3. Consult - History and Phy.
4. Emergency Room Reports
5. Discharge Summary
6. Pain Management
7. General Medicine

Επίσης, λόγω κοινού γνωστικού πεδίου, συγχωνεύονται οι εξής κατηγορίες:

1. Neurosurgery & Neurology
2. Nephrology & Urology

Ο πίνακας του πλήθους εγγραφών ανα κατηγορία μετά τη διαμόρφωση του έχει τη μορφή που παρουσιάζεται στο Σχήμα 4.8:

```
============Reduced Categories==========
Cat:1 Cardiovascular / Pulmonary : 371
Cat:2 ENT - Otolaryngology : 96
Cat:3 Gastroenterology : 224
Cat:4 Hematology - Oncology : 90
Cat:5 Neurology : 317
Cat:6 Obstetrics / Gynecology : 155
Cat:7 Ophthalmology : 83
Cat:8 Orthopedic : 355
Cat:9 Pediatrics - Neonatal : 70
Cat:10 Psychiatry / Psychology : 53
Cat:11 Radiology : 273
Cat:12 Urology : 237
============Reduced Categories==========
(2324, 2)
```

Σχήμα 4.8: Το πλήθος εγγραφών ανά κατηγορία μετά την απαλοιφή και τη συγχώνευση κάποιων κατηγοριών

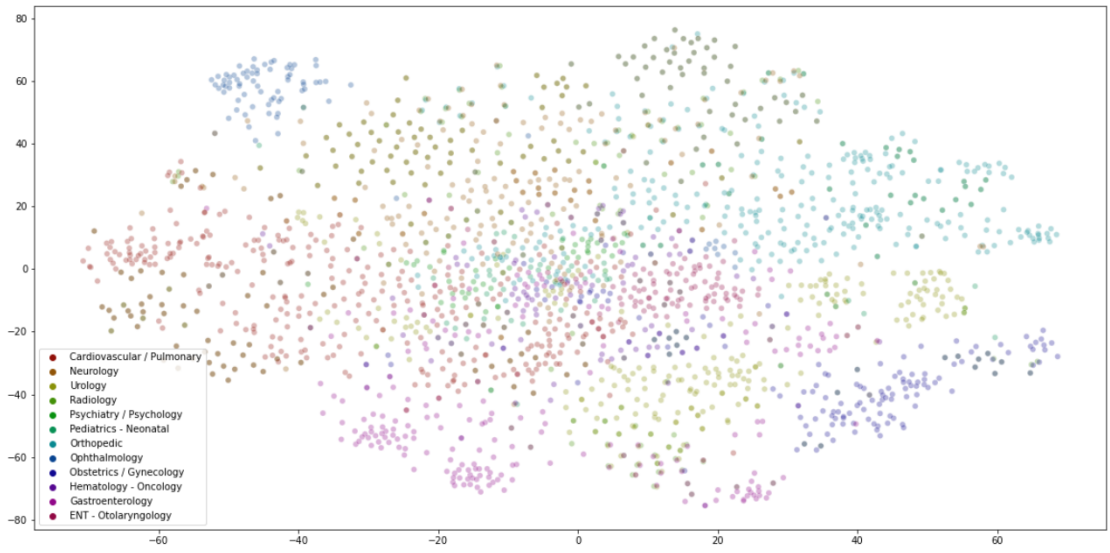
Το sciSpacy είναι ένα πακέτο της Python το οποίο περιέχει τα μοντέλα Spacy και είναι σχεδιασμένο για προεπεξεργασία σε κείμενα που περιέχουν οντότητες που αποτελούν ιατρική, επιστημονική ή κλινική ορολογία.

Αξιοποιώντας τη δυνατότητα αυτού του πακέτου, δημιουργείται η συνάρτηση process_text() η οποία περνάει κάθε εγγραφή του πεδίου transcription ως είσοδο στη μέθοδο nlp() του μοντέλου en_ner_bionlp13cg_md του πακέτου sciSpacy, η οποία επιστρέφει το σύνολο λέξεων που έχουν ιατρική, επιστημονική ή κλινική νοηματική αξία.

Έπειτα, εφαρμόζουμε αναδρομικά όλη τη διαδικασία:

1. Προεπεξεργασία (εφαρμογή sciSpacy)
2. Tf-Idf αξιολόγηση
3. Απεικόνιση του πίνακα Tf-Idf με χρήση της μεθόδου t-SNE
4. Εφαρμογή PCA στον πίνακα Tf-Idf
5. Εκ νέου διαχωρισμός δεδομένων σε training set και test set
6. Εφαρμογή λογιστικής παλινδρόμησης (logistic regression)
7. Κατασκευή Πίνακα Σύγχυσης (confusion matrix)
8. Εμφάνιση νέων αποτελεσμάτων μετρικών ποιότητας πρόβλεψης

Το Σχήμα 4.9 εμφανίζει την απεικόνιση του πίνακα που προέκυψε από την tf-idf με χρήση της μεθόδου t-SNE μετά την εκ νέου προεπεξεργασία των δεδομένων

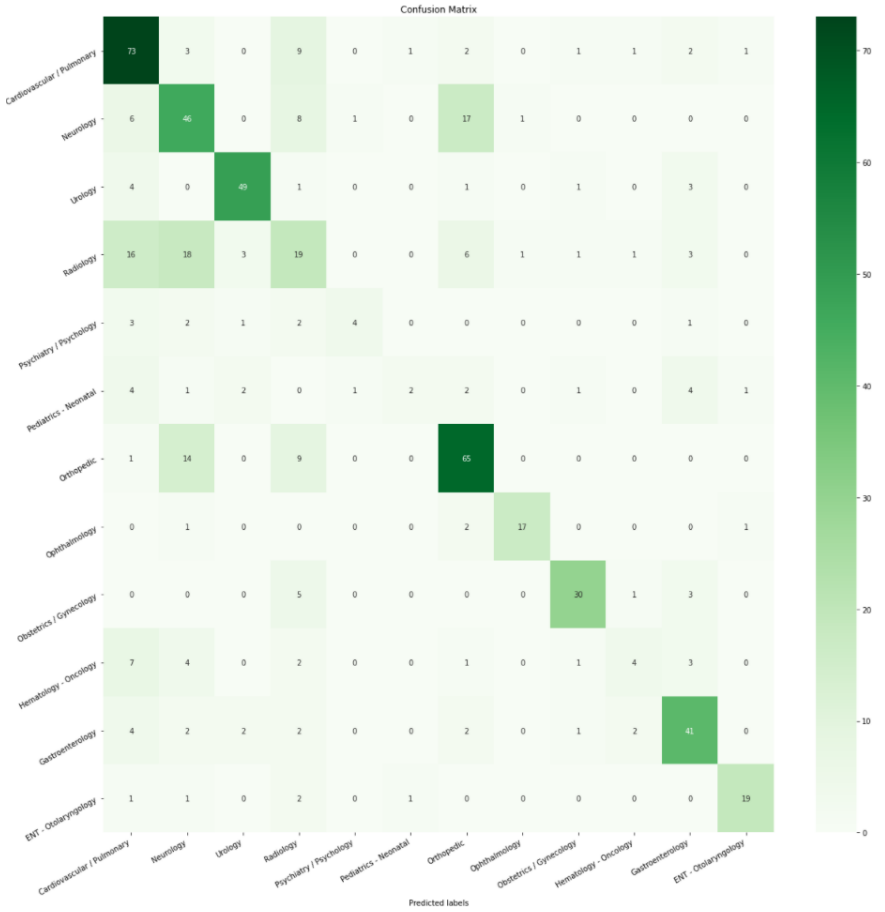


Σχήμα 4.9: Η νέα γραφική απεικόνιση του αλγορίθμου t-SNE

Ο νέος Πίνακας Σύγχυσης και ο νέος Πίνακας αποτελεσμάτων μετρικών ταξινόμησης φαίνονται στα σχήματα 4.11 και 4.10

	precision	recall	f1-score	support
Cardiovascular / Pulmonary	0.61	0.78	0.69	93
Neurology	0.50	0.58	0.54	79
Urology	0.86	0.83	0.84	59
Radiology	0.32	0.28	0.30	68
Psychiatry / Psychology	0.67	0.31	0.42	13
Pediatrics - Neonatal	0.50	0.11	0.18	18
Orthopedic	0.66	0.73	0.70	89
Ophthalmology	0.89	0.81	0.85	21
Obstetrics / Gynecology	0.83	0.77	0.80	39
Hematology - Oncology	0.44	0.18	0.26	22
Gastroenterology	0.68	0.73	0.71	56
ENT - Otolaryngology	0.86	0.79	0.83	24
accuracy			0.64	581
macro avg	0.65	0.58	0.59	581
weighted avg	0.63	0.64	0.62	581

Σχήμα 4.10: Πίνακας Αποτελεσμάτων μετρικών ταξινόμησης μετά την επεξεργασία με το πακέτο sciSpacy



Σχήμα 4.11: Πίνακας Σύγχυσης μετά την επεξεργασία με το πακέτο sciSpacy

Παρατηρούμε εμφανώς βελτιωμένα αποτελέσματα στα scores των διάφορων μετρικών. (προηγούμενα αποτελέσματα στο σχήμα 4.7) Παρ' όλα αυτά, εξακολουθεί να υπάρχει ανισορροπία στο πλήθος των υπαρχόντων δεδομένων μεταξύ των κατηγοριών.

Στα πλαίσια της συγκεκριμένης εργασίας ακολουθήθηκαν σε προηγούμενα βήματα και οι δύο προσεγγίσεις αντιμετώπισης της ανισορροπίας των δεδομένων (oversampling και undersampling: αφαιρέθηκαν οι κλάσεις που αποτελούσαν μειονότητες καθώς και η κλάση της χειρουρικής που περιείχε δεδομένα κατά πολύ μεγαλύτερα σε όγκο σε σχέση με τις υπόλοιπες κλάσεις τα οποία επίσης συγχέονταν με αυτές λόγω επικαλυπτόμενων γνωστικών πεδίων.

Μία ακόμα απλή προσέγγιση για την αντιμετώπιση του προβλήματος θα ήταν να χρησιμοποιηθεί η απλούστερη oversampling τεχνική η οποία υποδεικνύει τη δημιουργία διπλότυπων στιγμιοτύπων στις κατηγορίες που αποτελούν μειονότητες. Κάτι τέτοιο όμως, παρ΄ ότι φαινομενικά επιφέρει μεγαλύτερη ισορροπία στα δεδομένα, δεν προσθέτει καμία καινούρια πληροφορία στο μοντέλο εκπαίδευσης.

Γι΄ αυτό θα χρησιμοποιηθεί η τεχνική SMOTE (Synthetic Minority Oversampling Technique).

Πριν τη δημιουργία νέου dataset με την τεχνική SMOTE τα δεδομένα ήταν τα εξής:

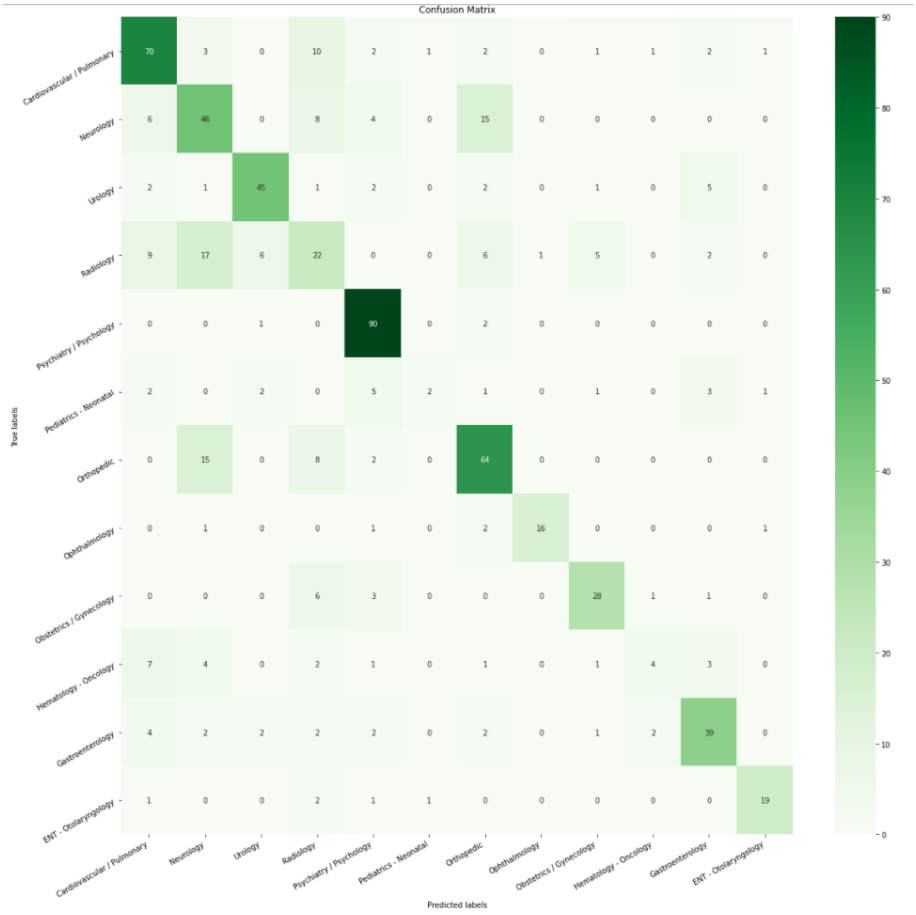
- Train_Set_Size: (1743, 696)
- Test_Set_Size: (581, 696)

Μετά τη δημιουργία νέου dataset με την τεχνική SMOTE τα δεδομένα είναι τα εξής:

- Train_Set_Size: (1981, 696)
- Test_Set_Size: (661, 696)

4.4.1 Λογιστική Παλινδρόμηση

Στη συνέχεια γίνεται εφαρμογή Λογιστικής Παλινδρόμησης και παράγεται ο νέος Πίνακας Σύγχυσης (Σχήμα 4.12):



Σχήμα 4.12: Πίνακας Σύγχυσης μετά την εφαρμογή της τεχνικής SMOTE

Τέλος, δημιουργείται ο πίνακας των τελικών αποτελεσμάτων μετρικών ταξινόμησης ο οποίος φαίνεται στο σχήμα 4.13.

	precision	recall	f1-score	support
Cardiovascular / Pulmonary	0.69	0.75	0.72	93
Neurology	0.52	0.58	0.55	79
Urology	0.80	0.76	0.78	59
Radiology	0.36	0.32	0.34	68
Psychiatry / Psychology	0.80	0.97	0.87	93
Pediatrics - Neonatal	0.50	0.12	0.19	17
Orthopedic	0.66	0.72	0.69	89
Ophthalmology	0.94	0.76	0.84	21
Obstetrics / Gynecology	0.74	0.72	0.73	39
Hematology - Oncology	0.50	0.17	0.26	23
Gastroenterology	0.71	0.70	0.70	56
ENT - Otolaryngology	0.86	0.79	0.83	24
accuracy			0.67	661
macro avg	0.67	0.61	0.63	661
weighted avg	0.66	0.67	0.66	661

Σχήμα 4.13: Logistic Regression: Πίνακας τελικών Αποτελεσμάτων μετρικών ταξινόμησης μετά την εφαρμογή της τεχνικής SMOTE

Τα τελικά αποτελέσματα είναι εμφανώς βελτιωμένα. Παρατηρούμε οτι σε συγκεκριμένες κατηγορίες όπως Neurology, Radiology και Hematology τα αποτελέσματα ταξινόμησης παραμένουν χαμηλά καθώς οι περιγραφές των περιστατικών εξακολουθούν να αφορούν παραπάνω απο μία ειδικότητες.

Στη συνέχεια, για την ταξινόμηση του κειμένου, δοκιμάζεται η επίδοση των αλγορίθμων:

1. Naïve Bayes
2. SVM
3. kNN

διατηρώντας ως είσοδο το dataset στην τελευταία μορφή που ταξινομήθηκε με τη μέθοδο της λογιστικής παλινδρόμησης, πριν την επεξεργασία για επίτευξη oversampling με την τεχνική SMOTE.

4.4.2 Αλγόριθμος Naïve Bayes

Ο πίνακας αποτελεσμάτων των μετρικών ταξινόμησης που παράγεται από τη μέθοδο classification_report() για τον Αλγόριθμο Naïve Bayes φαίνεται στο σχήμα 4.14.

	precision	recall	f1-score	support
Cardiovascular / Pulmonary	0.61	0.88	0.72	112
ENT - Otolaryngology	0.86	0.55	0.67	33
Gastroenterology	0.84	0.67	0.74	63
Hematology - Oncology	0.67	0.16	0.26	25
Neurology	0.58	0.45	0.51	96
Obstetrics / Gynecology	0.97	0.70	0.82	54
Ophthalmology	0.96	0.76	0.85	33
Orthopedic	0.62	0.84	0.71	104
Pediatrics - Neonatal	0.91	0.38	0.54	26
Psychiatry / Psychology	0.80	0.75	0.77	16
Radiology	0.33	0.45	0.38	65
Urology	0.71	0.69	0.70	71
accuracy			0.65	698
macro avg	0.74	0.61	0.64	698
weighted avg	0.69	0.65	0.65	698

Σχήμα 4.14: Naïve Bayes: Πίνακας Αποτελεσμάτων μετρικών ταξινόμησης

4.4.3 Αλγόριθμος SVM

Ο πίνακας αποτελεσμάτων των μετρικών ταξινόμησης που παράγεται από τη μέθοδο classification_report() για τον Αλγόριθμο SVM φαίνεται στο σχήμα 4.15.

	precision	recall	f1-score	support
Cardiovascular / Pulmonary	0.67	0.81	0.73	99
ENT - Otolaryngology	0.82	0.69	0.75	26
Gastroenterology	0.76	0.80	0.78	69
Hematology - Oncology	0.36	0.40	0.38	25
Neurology	0.69	0.60	0.64	100
Obstetrics / Gynecology	0.78	0.81	0.80	48
Ophthalmology	0.97	0.80	0.88	35
Orthopedic	0.75	0.80	0.77	110
Pediatrics - Neonatal	0.38	0.35	0.36	17
Psychiatry / Psychology	0.86	0.80	0.83	15
Radiology	0.30	0.32	0.31	75
Urology	0.87	0.68	0.77	79
accuracy			0.68	698
macro avg	0.68	0.66	0.67	698
weighted avg	0.69	0.68	0.68	698

Σχήμα 4.15: SVM: Πίνακας Αποτελεσμάτων μετρικών ταξινόμησης

4.4.4 Αλγόριθμος kNN

Ο πίνακας αποτελεσμάτων των μετρικών ταξινόμησης που παράγεται από τη μέθοδο classification_report() για τον Αλγόριθμο kNN φαίνεται στο σχήμα 4.16.

	precision	recall	f1-score	support
Cardiovascular / Pulmonary	0.59	0.86	0.70	111
ENT - Otolaryngology	0.74	0.52	0.61	27
Gastroenterology	0.79	0.79	0.79	71
Hematology - Oncology	0.70	0.25	0.37	28
Neurology	0.56	0.45	0.50	96
Obstetrics / Gynecology	0.75	0.68	0.71	40
Ophthalmology	1.00	0.65	0.78	31
Orthopedic	0.63	0.73	0.68	102
Pediatrics - Neonatal	0.50	0.52	0.51	21
Psychiatry / Psychology	0.83	0.62	0.71	16
Radiology	0.39	0.44	0.41	79
Urology	0.79	0.64	0.71	76
accuracy			0.63	698
macro avg	0.69	0.60	0.62	698
weighted avg	0.65	0.63	0.63	698

Σχήμα 4.16: kNN: Πίνακας Αποτελεσμάτων μετρικών ταξινόμησης

4.4.5 ΛΕΙΠΕΙ -> Word2Vec - Neural Network

Word Embedding

Ο όρος Word Embedding είναι ο συλλογικός όρος για τεχνικές μάθησης χαρακτηριστικών όπου οι λέξεις από το λεξιλόγιο αντιστοιχίζονται σε διανύσματα πραγματικών αριθμών. Αυτά τα διανύσματα υπολογίζονται από την πιθανοτική κατανομή για κάθε λέξη που εμφανίζεται πριν ή μετά από μια άλλη. Με άλλα λόγια, λέξεις που εκφράζουν παρόμοιες έννοιες συνήθως εμφανίζονται μαζί στο σώμα κειμένου, επομένως θα είναι κοντά και στον χώρο των διανυσμάτων.

Στα πλαίσια της παρούσας διπλωματικής, χρησιμοποιείται το πρώτο μοντέλο αυτής της οικογένειας: το Word2Vec της Google (2013). Άλλα δημοφιλή μοντέλα ενσωμάτωσης λέξης είναι το GloVe του Stanford (2014) και το FastText του Facebook (2016).

Το Word2Vec παράγει έναν χώρο διανυσμάτων, συνήθως με εκατοντάδες διαστάσεις, για κάθε διακριτή λέξη του κειμένου, έτσι ώστε οι λέξεις που εκφράζουν παρόμοιες έννοιες να βρίσκονται κοντά η μία στην άλλη στο διανυσματικό χώρο. Αυτό μπορεί να γίνει με δύο διαφορετικές προσεγγίσεις: ξεκινώντας από μια μεμονωμένη λέξη για να προβλέψουμε το συμφραζόμενο της (Skip-gram) ή ξεκινώντας από το συμφραζόμενο για να προβλέψουμε μια λέξη (Continuous Bag-of-Words).

Αρχικά μετατρέπουμε το σύνολο του dataset σε λίστα που περιέχει λίστες (list of lists), κάθε μία εκ των οποίων περιέχει το σύνολο των διακριτών λέξεων του κειμένου μιας εγγραφής.

Κατά την εφαρμογή του Word2Vec, θα πρέπει να καθοριστούν:

- Το επιθυμητό μέγεθος των διανυσματικών αναπαραστάσεων των λέξεων. Επιλέχθηκε η τιμή 300.
- Το παράθυρο (window), δηλαδή τη μέγιστη απόσταση μεταξύ της τρέχουσας και της προβλεπόμενης λέξης μέσα σε μια πρόταση. Στη συγκεκριμένη υλοποίηση επιλέχθηκε το μέσο μήκος του κειμένου στη συλλογή (TODO: ελεγχος)
- Τον αλγόριθμο εκπαίδευσης. Επιλέχθη η μέθοδος skip-grams (sg=1), καθώς γενικά παρουσιάζει καλύτερα αποτελέσματα.

Κεφάλαιο 5

Επίλογος

5.1 Αποτελέσματα

Συγκεντρωτικά, τα αποτελέσματα όλων των μεθόδων φαίνονται στον παρακάτω πίνακα:

Πίνακας Αποτελεσμάτων	
Αλγόριθμος	Ακρίβεια
Logistic Regression	0.64
Logistic Regression & SMOTE	0.67
Naive - Bayes	0.65
K Nearest Neighbors (KNN)	0.63
Support Vector Machines (SVM)	0.67
Word2Vec & Neural Network	0.68

Πίνακας 5.1: Συγκεντρωτικός Πίνακας Αποτελεσμάτων Ακρίβειας

5.2 Συμπεράσματα

Συμπερασματικά, το συγκεκριμένο dataset δεν είναι εύκολο να επεξεργαστεί με ακρίβεια καθώς δεν είναι «καθαρό». Με διάφορες τεχνικές φτάνουμε σε καλύτερα αποτελέσματα, ωστόσο αυτό δεν ειναι εφικτό να εφαρμοστεί σωστά σε οποιαδήποτε περιγραφή περιστατικού.

Μέσα από τις διάφορες προσεγγίσεις, με διαφορετική αναπαράσταση δεδομένων και διαφορετικές τεχνικές, καταφέραμε να φτάσουμε σε κάποια ικανοποιητικά ποσοστά ακρίβειας.

Το μεγαλύτερο ίσως εμπόδιο για την επίτευξη αυτού του στόχου ήταν η αντιμετώπιση της ανισορροπίας των δεδομένων, δηλαδή η άνιση κατανομή τους στις διάφορες κλάσεις, η

επικάλυψη των κλάσεων, η σωστή προεπεξεργασία των κειμένων ώστε να απομονωθεί η χρήσιμη πληροφορία και η σωστή παραμετροποίηση του νευρωνικού δικτύου για την αποτελεσματική εκπαίδευση του μοντέλου.

5.3 Μελλοντικές Επεκτάσεις

Ωστόσο, τα ποσοστά ακρίβειας που επιτεύχθησαν, αποτελούν αισιόδοξο σημάδι για μελλοντική δουλειά. Δημιουργώντας στο μέλλον ένα πιο ισορροπημένο και ακριβές dataset, θα μπορέσουμε να εκπαιδεύσουμε καλύτερα το σύστημα μας ωστε να αναπτύξουμε ένα ακόμα πιο ισχυρό και αποτελεσματικό εργαλείο. Πέρα από την πρόβλεψη ιατρικής ειδικότητας, μέσα απο την περιγραφή ιατρικού περιστατικού θα μπορούσε να δημιουργηθεί πρόγνωση για το μέτρο του επείγοντος, τη φαρμακευτική περίθαλψη που ενδεχομένως χρειστεί ή την πιθανότητα να χρειαστεί εισαγωγή και νοσηλεία, αντικείμενα και αποφάσεις που απασχολούν ιατρικό προσωπικό χωρίς απαραίτητα να χρήζουν ιατρικών γνώσεων.

Εν κατακλείδι, η δημιουργία χρήσιμων, ισχυρών και αποτελεσματικών εργαλείων στα χέρια ιατρών, ερευνητών και επιστημόνων αποτελεί ένα πεδίο με μεγάλο φάσμα προκλήσεων και προς επίλυση προβλημάτων, στοχεύοντας πάντα στην υποστήριξη και όχι στην αντικατάσταση επιστημονικά καταρτισμένου προσωπικού.

Βιβλιογραφία

[1] Abdullah Awaysheh, Jeffrey Wilcke, François Elvinger, Loren Rees, Weiguo Fan και Kurt L. Zimmerman. Review of medical decision support and machine-learning methods. *Veterinary Pathology*, 56(4):512–525, 2019.

[2] Jason Brownlee. What is machine learning?, 2013.

[3] Po Hao Chen, Hanna Zafar, Maya Galperin-Aizenberg και Tessa Cook. Integrating natural language processing and machine learning algorithms to categorize oncologic response in radiology reports. *Journal of digital imaging*, 31(2):178—184, 2018.

[4] Shiva Kazempour Dehkordi και Hedieh Sajedi. Prediction of disease based on prescription using data mining methods. *Health and Technology*, 9:37–44, 2018.

[5] Keith D. Foote. A brief history of machine learning, 2019.

[6] A. Georgouli. Μηχανική Μάθηση [chapter]. in georgouli, a. 2015. Τεχνητή νοημοσύνη [undergraduate textbook], 2015.

[7] Swiergosz A.M. Haeberle H.S. Helm, J.M. Machine learning and artificial intelligence: Definitions, applications, and future directions, 2020.

[8] https://blog.knoldus.com/introduction to perceptron neural network/. Introduction to perceptron: Neural network, 2017.

[9] https://researchdatapod.com/the history of machine learning/, χ.χ.

[10] https://www.kaggle.com/tboyle10/medicaltranscriptions, χ.χ.

[11] https://www.mtsamples.com/, χ.χ.

[12] Walter A. Kosters Peter van der Putten Joost N. Kok, Egbert J. W. Boers και Mannes Poel. Artificial intelligence: Definition, trends, techniques and cases, 2002.

[13] Christopher D. Manning, Prabhakar Raghavan και Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, USA, 2008.

[14] James A. Ohlson. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1):109–131, 1980.

[15] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review,*, 1958.

[16] K. Stuart, A.; Ord. Kendallś advanced theory of statistics: Volume i—distribution theory, 1994.

[17] A. M. Turing. I.—Computing Machinery and Intelligence. *Mind*, 1950.

[18] Αντωνιος Αθανασίου. *Πτυχιακή «Βαθιά Μάθηση με εφαρμογές στην Ιατρική»*. Σχολή Τεχνολογικών Εφαρμογών Τμήμα Αυτοματισμού, 2018.

[19] Ζαχόπουλος Γιώργος. *Master Thesis "Prediction of Aircraft trajectory using LSTM Neural Networks"*. Πανεπιστήμιο Πειραιώς – Τμήμα Πληροφορικής, 2018.

[20] Δημολίτσα Θεοδώρα Λελοβίτη Σοφία. *Νευρωνικά Δίκτυα και Μηχανική Μάθηση Εφαρμογές στη Βιοπληροφορική και στον Παγκόσμιο Ιστό*. Πανεπιστήμιο Θεσσαλίας, χ·χ·