ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικό και Καποδιστριακό
Πανεπιστήμιο Αθηνών

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ»

Δ Ι Π Λ Ω Μ Α Τ Ι Κ Η Ε Ρ Γ Α Σ Ι Α

«Αξιοποίηση μηχανικής μάθησης για τη μελέτη ιατροβιολογικών δεδομένων σχετικών με τον καρκίνο»

του

ΝΙΚΟΛΑΟΥ ΣΠΥΡΟΥ

Πτυχιούχου Ιατρικής Σχολής Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης

Αθήνα, Σεπτέμβριος 2019

HELLENIC REPUBLIC
National and Kapodistrian University of Athens

SCHOOL OF SCIENCE
FACULTY OF BIOLOGY

MASTER IN "BIOINFORMATICS"

Master Diploma Thesis

«Utilization of Machine Learning methods in data analysis related to Cancer»

NIKOLAOS SPYROU

MD, Aristotle University of Thessaloniki

ATHENS (2019)

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικό και Καποδιστριακό
Πανεπιστήμιο Αθηνών

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ»

Δ Ι Π Λ Ω Μ Α Τ Ι Κ Η Ε Ρ Γ Α Σ Ι Α

«Αξιοποίηση μηχανικής μάθησης για τη μελέτη ιατροβιολογικών δεδομένων σχετικών με τον καρκίνο»

Τριμελής εξεταστική επιτροπή

Επίκουρος Καθηγητής Ηρακλής Βαρλάμης (Επιβλέπων)
Τμήμα Πληροφορικής και Τηλεματικής,
Χαροκόπειο Πανεπιστήμιο

Καθηγητής Ιωάννης Τρουγκάκος
Τμήμα Βιολογίας,
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Δρ. Ιωάννης Αλμυράντης
Ερευνητής Α' Ιωάννης Αλμυράντης,
ΕΚΕΦΕ «Δημόκριτος»

Ευχαριστίες

Η παρούσα μελέτη δεν θα μπορούσε να φτάσει στην τελική της μορφή χωρίς την πολύτιμη συνεισφορά αρκετών ανθρώπων.

Καταρχήν, θέλω να ευχαριστήσω από καρδιάς τον Δρ. Γιώργο Γιαννακόπουλο, ερευνητή στο Εργαστήριο Μηχανικής Γνώσης και Λογισμικού (SKEL) του ΕΚΕΦΕ "Δημόκριτος", που δέχθηκε να αναλάβει την άμεση επίβλεψη αυτής της διπλωματικής εργασίας δεδομένης της χρονικής πίεσης που υπήρχε για την περάτωσή της. Ήταν πάντα ενθαρρυντικός, αισιόδοξος και ανεκτικός. Μέσω της τακτικής και συστηματικής ανάδρασης που δεχόμουν επάνω στα βήματα αυτής της εργασίας, διδάχθηκα να ξεπερνώ εμπόδια με δημιουργικό τρόπο και να είμαι ευέλικτος, χωρίς όμως να ξεφεύγω από τα όρια της επιστημονικής μεθόδου. Ταυτόχρονα, με βοήθησε στο να αναπτύξω αυτοπειθαρχία και να μάθω μέσω του παραδείγματός του, με ποιόν τρόπο λειτουργεί βέλτιστα μια επιστημονική ομάδα. Η καθοδήγησή του μέσα στον κόσμο της Μηχανικής Μάθησης αποτέλεσε ένα πολύ ευχάριστο ταξίδι.

Επίσης, θα ήθελα να ευχαριστήσω ιδιαιτέρως τον Επίκουρο Καθηγητή Δρ. Ηρακλή Βαρλάμη, του Τμήματος Πληροφορικής και Τηλεματικής του Χαροκόπειου Πανεπιστημίου, ο οποίος μου έκανε την τιμή να είναι επιβλέπων της παρούσας εργασίας.

Στη συνέχεια, θα ήθελα να ευχαριστήσω τον Καθηγητή Δρ. Ιωάννη Τρουγκάκο και Πρόεδρο του MSc "Βιοπληροφορική" καθώς και τον Δρ. Ιωάννη Αλμυράντη, Ερευνητή Α' του ΕΚΕΦΕ "Δημόκριτος", για τη συμμετοχή τους στην τριμελή εξεταστική επιτροπή της παρούσας εργασίας.

Η επιλογή και ταξινόμηση των ερωτήσεων που απαρτίζουν το σύνολο δεδομένων της παρούσας εργασίας έγινε με την καθοριστική συνδρομή του Δημήτρη Ράπτη τον οποίο και ευχαριστώ θερμά. Ευχαριστώ επίσης τον Αναστάσιο Νεντίδη, ερευνητή του ΕΚΕΦΕ "Δημόκριτος" ο οποίος με καθοδήγησε γύρω από λεπτομέρειες του BioASQ Challenge επάνω στο οποίο βασίστηκε η παρούσα εργασία. Θα ήθελα τέλος να ευχαριστήσω τη συμφοιτήτριά μου στο μεταπτυχιακό, Χριστίνα Βασιλοπούλου για τη στήριξη και καθοδήγησή της.

Στη Γιόνα,τη Δωροθέα και τη μικρούλα

"The greatest opportunity offered by AI is not reducing errors or workloads, or even curing cancer: it is the opportunity to restore the precious and time-honored connection and trust—the human touch—between patients and doctors. Not only would we have more time to come together, enabling far deeper communication and compassion, but also we would be able to revamp how we select and train doctors."

— Eric Topol, Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again

Περίληψη

Σε μια εποχή που η επιστημονική πληροφορία παράγεται με ραγδαίους ρυθμούς, η ανάγκη της σταχυολόγησης της από ερευνητές του βιοϊατρικού χώρου και κλινικούς ιατρούς γίνεται άκρως σημαντική. Για αυτόν το λόγο, μια από τις μεγαλύτερες σύγχρονες προκλήσεις της Ιατρικής Πληροφορικής και της Βιοπληροφορικής είναι η δημιουργία συστημάτων που να παρέχουν με αυτοματοποιημένο τρόπο, αξιόπιστες απαντήσεις σε ερωτήσεις βιοϊατρικού περιεχομένου, σε πραγματικό χρόνο.

Κύριος σκοπός της παρούσας εργασίας είναι η ανάλυση και ταξινόμηση ερωτήσεων ιατρικού-κλινικού περιεχομένου με έμφαση στις ερωτήσεις που αφορούν την Ογκολογία. Η επεξεργασία των ερωτήσεων επιχειρήθηκε με μεθόδους της Μηχανικής Μάθησης και της Επεξεργασίας Φυσικής Γλώσσας με απώτερο σκοπό τη εύρεση των βέλτιστων απαντήσεων σε αυτές σύμφωνα με τους κανόνες της Ιατρικής Βασισμένη σε Στοιχεία.

Το σύνόλο δεδομένων της παρούσας εργασίας, αντλήθηκε από τη βάση δεδομένων του BioASQ Challenge. Το BioASQ Challenge συνίσταται σε ένα παγκόσμιο διαγωνισμό επεξεργασίας επιστημονικών κειμένων και δημιουργίας συστημάτων αυτοματοποιημένης απάντησης στο βιοϊατρικό χώρο. Οι ερωτήσεις χαρακτηρίστηκαν και κατηγοριοποιήθηκαν από δύο ανεξάρτητους ειδικούς, σύμφωνα με κανόνες της Ιατρικής Βασισμένη σε Στοιχεία.

Η αλληλουχία της επεξεργασίας των ερωτήσεων διαχωρίστηκε σε τρία στάδια, καθένα από τα οποία εστίασε σε διαφορετικές ιδιότητες των ερωτήσεων. Στο πρώτο, επιχειρήθηκε η αυτόματη ταξινόμηση των ερωτήσεων σε ερωτήσεις τύπου background, δηλαδή σε ερωτήσεις των οποίων η απάντηση είναι γενικής φύσης, ξεκάθαρη και παγιωμένη και σε ερωτήσεις foreground των οποίων η απάντηση αφορά συγκεκριμένες κλινικές περιπτώσεις και/ή αφορά γνώση που δεν έχει ακόμη ξεκάθαρη απάντηση και αποτελεί αντικείμενο αντιπαράθεσης. Το δεύτερο στάδιο αφορούσε την αυτόματη ανίχνευση της ύπαρξης των στοιχείων PICO (patient, intervention, comparison, outcome) στις ερωτήσεις foreground. Το τρίτο και τελευταίο στάδιο περιείχε την κατηγοριοποίηση των ερωτήσεων σε ερωτήσεις που αφορούν τη θεραπεία, τη διάγνωση, την πρόγνωση ασθενειών καθώς και τη συσχέτιση διαφόρων παραγόντων με ασθένειες.

Σε αυτό το πλαίσιο, καταφέραμε να χαρακτηρίσουμε και να ταξινομήσουμε με επιτυχία, με συνδυασμό διάφορων μοντέλων Μηχανικής Μάθησης κλινικές ερωτήσεις του BioASQ challenge. Συγκεκριμένα, σε καθένα από τα στάδια κατασκευάσαμε μοντέλα που ταξινόμησαν τις ερωτήσεις σε στατιστικά σημαντικό βαθμό σε σχέση με την τυχαιότητα και αρκετές φορές σε επίπεδα συγκρίσιμα με αυτά της ιδανικής ταξινόμησης. Η απόδοση των μοντέλων θεωρήθηκε σημαντική όταν συγκρίθηκε και με τη συμφωνία ανθρώπων - κριτών στην ταξινόμηση των ερωτήσεων.

Η κατηγοριοποίηση των ερωτήσεων με αυτόν τον τρόπο αποσκοπεί στη διευκόλυνση της αναζήτησης των βέλτιστων απαντήσεων στους τύπους κλινικών μελετών που αντιστοιχούν σε κάθε κατηγορία. Φιλοδοξία της παρούσας εργασίας είναι να θέ-

σει τα θεμέλια για τη δημιουργία ενός ολοκληρωμένου συστήματος αυτοματοποιημένης απάντησης βιοϊατρικών ερωτήσεων που να παρέχει απαντήσεις σεβόμενο την ιεράρχηση των στοιχείων της σύγχρονης Ιατρικής Βασισμένης σε Στοιχεία.

Περιεχόμενα

1 Εισαγωγή

1.1 Επισκόπηση ενότητας (Overview)

Σε αυτή την ενότητα θα γίνει εισαγωγή στις βασικές έννοιες στις οποίες θα στηριχθεί η παρούσα εργασία. Αρχικά, θα γίνει αναφορά στην ανάγκη του σύγχρονου χώρου της Υγείας και ειδικά της Ογκολογίας για ποιοτικά συστήματα αυτοματοποιημένης απάντησης κλινικών ερωτήσεων. Στη συνέχεια θα γίνει μια περιεκτική εισαγωγή στην Ιατρική Βασισμένη σε Στοιχεία και τους κανόνες της, πάνω στους οποίους θα στηριχτεί η λογική των κύριων πειραμάτων-ταξινομήσεων του πειραματικού σκέλους της παρούσας εργασίας. Έπειτα, Θα επεξηγηθεί η φιλοσοφία και η δομή του BioASQ challenge, της πηγής δηλαδή του συνόλου δεδομένων.

Στο δεύτερο μεγάλο σκέλος της εισαγωγής θα επιχειρηθεί η εισαγωγή και επεξήγηση εννοιών της Μηχανικής Μάθησης καθώς και της Επεξεργασίας Φυσικής Γλώσσας. Πάνω σε αυτές, θα βασιστεί η δόμηση των υπολογιστικών εργαλείων-ταξινομητών, του πυρήνα δηλαδή του το πειραματικού σκέλους της παρούσας εργασίας. Τέλος, θα γίνει αναφορά σε προηγούμενες ερευνητικές προσπάθειες στο ίδιο αντικείμενο, και θα τεθούν οι στόχοι τους οποίους θα προσπαθήσει να εκπληρώσει το παρόν πόνημα.

1.2 Εποχή της βιοϊατρικής Πληροφορίας

1.2.1 Ιατρική βασισμένη σε στοιχεία(Evidence Based Medicine)

Η Ιατρική βασισμένη σε Στοιχεία (Evidence Based Medicine, EBM) είναι η κριτική, εξορθολογισμένη και μεθοδική χρήση των βέλτιστων και πιό σύγχρονων στοιχείων που προκύπτουν από ερευνητικές μελέτες, για τη λήψη αποφάσεων σχετικά με τη φροντίδα των ασθενών[1]. Η EBM συγχωνεύει την κλινική εμπειρία με τις αξίες του ασθενούς και τις καλύτερες διαθέσιμες ερευνητικές πληροφορίες[1]. Από πλευράς του ιατρού, η άσκηση της EBM προυποθέτει την ικανότητα αποτελεσματικής αναζήτησης στη βιβλιογραφία και στη συνέχεια, την εφαρμογή επίσημων κανόνων για την αξιολόγηση και ιεράρχηση των αποτελεσμάτων που η αναζήτηση παρέχει. Η κύρια διαφορά της EBM σε σχέση με την κλασσική Ιατρική, δεν έγκειται στο ότι η πρώτη λαμβάνει υπόψιν της στοιχεία ενώ η δεύτερη όχι, αλλά στο ότι η πρώτη ιεραρχεί τα στοιχεία και απαιτεί τα βέλτιστα για την καθοδήγηση της διάγνωσης, πρόγνωσης και θεραπείας των ασθενών.

1.2.1.1 Ογκολογία, εκεί που τα νέα δεδομένα προκύπτουν με ραγδαίο ρυθμό

Στη σύγχρονη εποχή, οι ιατροί που ασκούν την ειδικότητα της Παθολογικής Ογκολογίας έρχονται αντιμέτωποι με αρκετές προκλήσεις σχετικά με τη διαχείριση της νέας πληροφορίας. Ως και τον Σεπτέμβριο του 2019 η βάση

δεδομένων PubMed περιείχε περισσότερα από 4.5 εκατομμύρια άρθρα υπό τον MeSH (Medical Subject Heading) όρο 'neoplasms'. Επιπροσθέτως, περισσότερα από 100.000 νέα άρθρα προστίθενται στη βάση ετησίως από το 2011. Για να συγκρίνουμε αυτές τις ποσότητες με αυτές άλλων ιατρικών ειδικοτήτων, υπό τον MeSH όρο "Heart Diseases" υπάρχουν 1.3 εκατομμύρια άρθρα, ενώ υπό τον όρο "Diabetes Mellitus" 475.000 άρθρα. Για να μπορέσει κάποιος να είναι ενήμερος (να διαβάσει μία φορά κάθε νέο άρθρο που βγαίνει) μόνο στην προκύπτουσα βιβλιογραφία της Ογκολογίας, θα έπρεπε να διαβάζει 17 περίπου άρθρα ανά ώρα, 16 ώρες την ημέρα.

Αυτή η παραγωγή τεράστιων ποσοτήτων πληροφορίας και η ανικανότητα ποιοτικής διαχείρισής της από αυτούς που καλούνται να την κρίνουν και να λάβουν ενημερωμένες αποφάσεις μέσω αυτής, συνιστά μια από τις μεγαλύτερες προκλήσεις της εποχής για την Ογκολογία που απαιτεί στρατηγικές λύσεις. Η Εθνική Ακαδημία Μηχανικής (National Academy of Engineering) των Η.Π.Α. χαρακτηριστικά, έχει προτείνει για το παραπάνω πρόβλημα, ότι οι ιατροί χρειάζονται επικουρικά, συστηματικά ηλεκτρονικά συστήματα που να βρίσκουν πληροφορίες στοχευμένα για τη θεραπεία συγκεκριμένων ασθενών και συστήματα υποστήριξης κλινικών αποφάσεων που να προσφέρουν εξατομικευμένες προτάσεις τη στιγμή που ο κλινικός γιατρός τις έχει ανάγκη[2]. Μέχρι τώρα υπάρχουν κάποια τέτοια

συστήματα υποστήριξης των οποίων η ευχρηστία βελτιώνεται συνεχώς, παρόλα αυτά παραμένουν αρκετά μακριά ακόμη από το ιδεατό[3].

1.2.1.2 Κλινικές ερωτήσεις

Η άσκηση της Ιατρικής βασισμένης σε στοιχεία κατά τη διάρκεια της φροντίδας των ασθενών, πολλές φορές έχει σαν εφαλτήριο μια ερώτηση που κάνει ο κλινικός ιατρός, της οποίας η απάντηση θα καθοδηγήσει τις επιλογές του κατά τη διαχείριση των ασθενών. Οι κλινικές ερωτήσεις, για να αποφέρουν τις βέλτιστες απαντήσεις πρέπει να πληρούν ορισμένα κριτήρια[1]. Ως κλινικές ερωτήσεις ορίζονται οι ερωτήσεις εκείνες των οποίων αντικείμενο είναι πληροφορίες γύρω από ασθένειες ή/και ασθενείς και αποσκοπούν στην επιλογή της βέλτιστης πρακτικής από πλευράς του θεράποντα ιατρού. Στις επόμενες ενότητες θα γίνει περιγραφή των τύπων των κλινικών ερωτήσεων με βάση αυτά. Η περιγραφή αυτή θα αποτελέσει στη συνέχεια βάση των ταξινομήσεων της παρούσας εργασίας.

1.2.1.3 Κλινικές ερωτήσεις τύπου background και foreground

Οι κλινικές ερωτήσεις μπορούν να διαχωριστούν σε background και foreground ερωτήσεις [4, 5, 6, 7]. Οι ερωτήσεις background ρωτάνε για γενικές γνώσεις επάνω σε μια πάθηση, κατάσταση, διαδικάσία ή οντότητα. Αυτές οι ερωτήσεις ρωτάνε το ποιός, τι, που, πότε, πως και γιατί σχετικά με παθήσεις, διαγνωστικά τεστ, θερ-

απείες κλπ. Για παράδειγμα:

- Ποιά είνα τα κύρια συμπτώματα της πνευμονίας;
- Ποιά είναι η ειδικότητα και η ευαισθησία του strep-test[1];
- Ποιά είναι η γενετική αιτία του συνδρόμου Down;

Από την άλλη πλευρά, οι ερωτήσεις foreground ρωτάνε για συγκεκριμένες πληροφορίες έτσι ώστε να ενημερωθούν οι κλινικές αποφάσεις των ιατρών. Τυπικά, αυτές οι ερωτήσεις αφορούν συγκεκριμένους ασθενείς ή ομάδα ασθενών με συγκεκριμένα χαρακτηριστικά. Οι ερωτήσεις foreground τείνουν να είναι πιό συγκεκριμένες και πολύπλοκες από τις ερωτήσεις background. Συχνά, οι ερωτήσεις foreground διερευνούν συγκρίσεις π.χ. σύκριση δύο φαρμάκων, δύο θεραπειών, δύο διαγνωστικών tests κλπ. Για παράδειγμα:

- Σε ασθενείς με καρκίνο του πνεύμονα η σισπλατίνη είναι πιό αποτελεσματική από την καρβοπλατίνη[2];
- Είναι το ηλεκτρονικό τσιγάρο αποτελεσματικό στη διακοπή του καπνίσματος;
- Μπορεί η μετφορμίνη[3] να μειώσει την επίπτωση του καρκίνου του εντέρου;

[1] διαγνωστικό εργαλείο για τη διάγνωση της στρεπτοκοκκικής φαρυγγίτιδας
[2] χημειοθεραπευτικό φάρμακο
[3] αντιδιαβητικό φάρμακο

Η διάκριση των ερωτήσεων στις δύο αυτές κατηγορίες, την κατηγορία foreground και την κατηγορία background είναι μεγάλης σημασίας για τη μεθοδολογία που πρέπει να ακολουθηθεί στην πορεία αναζήτησης της βέλτιστης απάντησης.

Οι ερωτήσεις τύπου background είναι ερωτήσεις των οποίων το αντικείμενο δεν είναι αντικείμενο μελέτης της σύγχρονης ερευνητικής δραστηριότητας[5]. Οι απαραίτητες πληροφορίες που χρειάζονται για την πλήρη απάντησή τους μπορεί να ανευρεθεί σε συγγράμματα αναφοράς, τα οποία δεν χρειάζεται κατά ανάγκη να είναι τελείως εκσυγχρονισμένα.

Αντίθετα ερωτήσεις οι ερωτήσεις τύπου foreground είναι λεπτομερείς ερωτήσεις, συνήθως προσαρμοσμένες σε συγκεκριμένη κατηγορία ασθενών, που ερωτούν για την αποτελεσματικότητα κάποιας συγκεκριμένης θεραπείας, την απόδοση κάποιου διαγνωστικού εργαλείου ή τη συσχέτιση με κάποιο συγκεκριμένο προγνωστικό, επιβαρυντικό ή ελαφρυντικό παράγοντα. Για τη βέλτιστη απάντηση σε τέτοιου είδους ερωτήσεις οι απαραίτητες πληροφορίες κατά πάσα πιθανότητα χρειάζεται να ανευρεθούν σε βάσεις δεδομένων ιατρικών μελετών υπό τη μορφή των αρτιέστερα σχεδιασμένων, πιό σύγχρονων και με χαρακτηριστικά πληθυσμού της μελέτης εγγύτερα σε αυτά του υπό μελέτη ασθενούς, κλινικών μελετών. Όπως θα παρατηρηθεί και στη συνέχεια, ανάλογα με την επιμέρους κατηγοριοποίηση των ερωτήσεων foreground σε ερωτήσεις διάγνωσης,συσχέτισης, πρόγνωσης και

θεραπείας κάθε μια από αυτές τις κατηγορίες μπορεί να απαντηθεί βέλτιστα με πληροφορίες από διαφορετική κατηγορία κλινικών μελετών η καθεμιά[5].

1.2.1.4 Το μοντέλο PICO Οι ερωτήσεις background όπως αναφέρθηκε είναι ερωτήσεις το αντικείμενο τον οποίων είναι γνώση καλά θεμελιωμένη, δεν έχουν απαραίτητα συγκεκριμένη δομή[5]. Οι ερωτήσεις foreground ωστόσο έχουν δομή η οποία είναι σχετικά ελαστική, ωστόσο πρέπει να περιέχει συγκεκριμένα στοιχεία για να είναι σαφής[5]. Μια καλά δομημένη ερώτηση foreground πρέπει να διαθέτει τέσσερα στοιχεία που συνοψίζονται με το ακρωνύμιο PICO [8]. Αναλυτικά:

- P: patient, problem, population

  Ποιά είναι τα χαρακτηριστικά του πληθυσμού που μας ενδιαφέρει;

- I: intervention, Prognostic factor, Exposure

  Ποιά είναι η παρέμβαση της οποίας το αποτέλεσμα θέλουμε να μελετήσουμε;

- C: Comparison

  Τι θα συγκρίνουμε με την παρέμβασή μας; Μια άλλη θεραπεία, φάρμακο, placebo[4], ένα άλλο διαγνωστικό εργαλείο;

- O: Outcome

[4]Η έννοια του placebo περιγράφει μια ουσία χωρίς φαρμακευτική δράση που χρησιμοποιείται στις ομάδες ελέγχου των κλινικών μελετών. Αν ένα φάρμακο έχει στατιστικά σημαντικότερη επίδραση από το placebo, τότε αποδεικνύεται η δραστικότητά του.

Τι προσπαθούμε να πετύχουμε, να μετρήσουμε, να επηρεάσουμε; Πως θα αξιολογήσουμε την αποτελεσματικότητα της παρέμβασής μας;

Για παράδειγμα, αν η ερώτηση:

"Είναι το ηλεκτρονικό τσιγάρο αποτελεσματικό στη διακοπή του καπνίσματος;"

αναλυθεί σε στοιχεία PICO τότε προκύπτει:

- P: καπνιστές
- I: ηλεκτρονικό τσιγάρο
- C: placebo (όπου δεν λέγεται ρητά και εξετάζεται μια παρέμβαση και όχι διαγνωστικό εργαλείο ή παράγοντας κινδύνου τότε υπονοείται ότι το C είναι placebo)
- O: διακοπή του καπνίσματος

1.2.1.5 Ερωτήσεις foreground που αφορούν θεραπεία, διάγνωση, πρόγνωση και επιβάρυνση

Οι ερωτήσεις foreground μπορούν να διαχωριστούν περαιτέρω σε τέσσερις κατηγορίες: τις σχετιζόμενες με θεραπεία, διάγνωση, πρόγνωση και κίνδυνο-επιβαρυντικό παράγοντα (π.χ. έκθεση σε βλαπτικό παράγοντα)[6].

Οι ερωτήσεις που αφορούν **θεραπεία** περιλαμβάνουν το σύνολο εκείνο των

ερωτήσεων που έχουν σαν αντικείμενό τους την εξέταση της επίδρασης μιας παρέμβασης σε κάποια άλλη παράμετρο (π.χ. ίαση από μια ασθένεια ή αύξηση της επιβίωσης) σχετιζόμενη με κάποια ασθένεια ή κατάσταση.

Παράδειγμα:

"Is palcociclib [5]effective in the treatment of lung cancer?"

Οι ερωτήσεις foreground σχετιζόμενες με **διάγνωση** αφορούν την εξέταση ποσοτικών παραμέτρων και ποιοτικών (ευαισθησία, ειδικότητα, Area under the curveκ.α.) παραμέτρων διαγνωστικών δοκιμασιών για τη διάγνωση κάποιας ασθένειας. Οι ερωτήσεις μπορεί να περιέχουν και σύγκριση διαγνωστικών εργαλείων.

Παράδειγμα:

"Serum amylase[6] or abdominal CT[7] is more sensitive in the diagnosis of pancreatitis?"

Οι ερωτήσεις **πρόγνωσης** συνήθως ρωτούν για το αν υπάρχει και ποιά είναι η συσχέτιση κάποιας παραμέτρου με την έκβαση κάποιας ασθένειας. Αυτή η παράμετρος μπορεί να είναι κάποιος βιοδείκτης ή κάποια άλλη συνιστώσα που με κάποιο τρόπο να σχετίζεται με τη βαρύτητα μιας ασθένειας, την ανταπόκριση σε κάποια θεραπεία κ.α.

Παράδειγμα:

[5]Νέο φάρμακο της ομάδας των αναστολέων κυκλινών εναντίον του καρκίνου του μαστού
[6]παγκρεατικό ένζυμο
[7]αξονική τομογραφία

"Does PD-L1[8] positivity affect bladder cancer prognosis?"

Τέλος, οι ερωτήσεις **επιβάρυνσης -συσχέτισης** εξετάζουν τη συσχέτιση κάποιας παραμέτρου με κάποια ασθένεια. Αυτή η συσχέτιση μπορεί να είναι θετική, δηλαδή η παράμετρος αυτή να αυξάνει της πιθανότητα για κάποια ασθένεια, οπότε και ονομάζεται παράγοντας κινδύνου (π.χ. κάπνισμα για τον καρκίνο του πνεύμονα), μπορεί όμως να είναι και αρνητική, δηλαδή η παράμετρος να μειώνει την πιθανότητα για την εμφάνιση μιας πάθησης (π.χ. άθληση για το έμφραγμα του μυοκαρδίου).

Παράδειγμα:

"Is Vitamin D deficiency associated with multiple sclerosis?"

1.2.1.6 Τύποι κλινικών μελετών και αντιστοίχισή τους στους τύπους κλινικών ερωτήσεων

Γνωρίζοντας τον τύπο της ερώτησης foreground, μπορούμε να διαλέξουμε με ακρίβεια, τον τύπο μελέτης που θα απαντήσει καλύτερα στην ερώτησή μας σύμφωνα με την ιεράρχηση των στοιχείων της Ιατρικής Βασισμένης σε Στοιχεία (Σχήμα 1).

Για να εξεταστεί στη συνέχεια η αντιστοίχιση του τύπου της κλινικής ερώτησης foreground με το βέλτιστο τύπο κλινικής μελέτης που μπορεί να δώσει την απάντηση,

[8] υποδοχέας κυττάρων του ανοσοποιητικού. Η ανακάλυψή του έφερε την επανάσταση στην ογκολογία (βλ. ανοσοθεραπεία)

θα οριστούν εδώ περιεκτικά οι έννοιες της μελέτης κοόρτης, της μελέτης ασθενών μαρτύρων, της τυχαιοποιημένης μελέτης καθώς και της συστηματικής ανασκόπησης και μεταανάλυσης[1]. Η επεξήγηση που ακολουθεί δεν είναι συστηματική ή εξαντλητική αλλά αποσκοπεί στην αποσαφήνιση κάποιων ενδεικτικών εννοιών. Συγκεκριμένα:

-Μελέτη **κοόρτης** είναι ένα είδος μελέτης κατά την οποία μια ομάδα ατόμων (κοόρτη) η οποία εκτίθεται σε έναν συγκεκριμένο παράγοντα κινδύνου (π.χ. κάπνισμα) και προοπτικά στο χρόνο υπολογίζεται ο αριθμός των ατόμων που εμφανίζουν κάποιο συγκεκριμένο νόσημα (π.χ. καρκίνος πνεύμονα) με σκοπό τη διερεύνηση αιτιολογικών σχέσεων

-Μελέτη **ασθενών μαρτύρων** είναι ένα είδος μελέτης στην οποία δύο ομάδες ατόμων από τις οποίες η μια περιέχει ασθενείς με κάποια συγκεκριμένη ασθένεια (π.χ. καρκίνος πνεύμονα) ενώ η άλλη δεν έχει αυτή την ασθένεια (μάρτυρες) εξετάζονται αναδρομικά (κοιτώντας στο παρελθόν)για το αν είχαν εκτεθεί σε κάποιο συγκεκριμένο παράγοντα (π.χ. κάπνισμα) με σκοπό τη διερεύνηση συσχετίσεων

-**Τυχαιοποιημένη ελεγχόμενη** μελέτη (Randomized Controlled Trial-RCT)είναι ένα είδος μελέτης κατά την οποία δύο ομάδες ατόμων που έχουν μια συγκεκριμένη πάθηση (π.χ. καρκίνος πνεύμονα) λαμβάνουν η μεν πρώτη μια συγκεκριμένη θεραπεία (π.χ. Nivolumab[9]) η δε δεύτερη μια θεραπεία ελέγχου (θεραπεία placebo

[9]νέο αντικαρκινικό φάρμακο της κατηγορίας της ανοσοθεραπείας

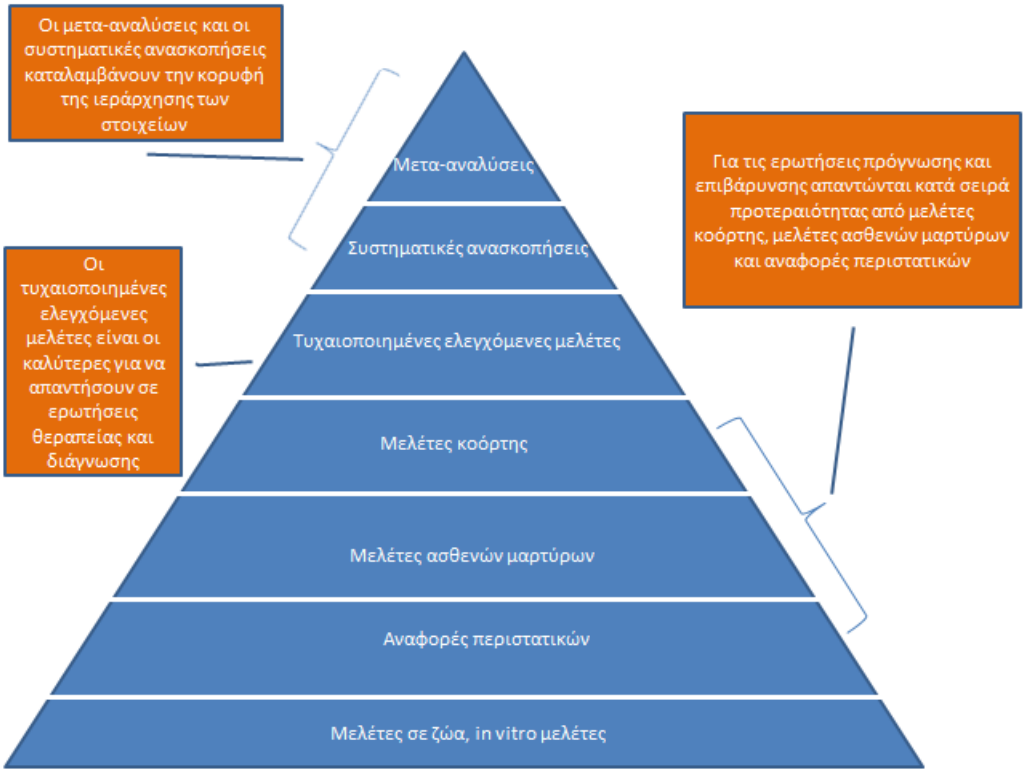
ή θεραπεία που ως τη στιγμή εκείνη θεωρείτε η βέλτιστη πρακτική) με σκοπό τη εύρεση νέων θεραπειών ή σύγκριση θεραπειών μεταξύ τους

-Συστηματικη ανασκοπήση-Μεταανάλυση είναι ένα είδος μελέτης το οποίο συγκεντρώνει τα ευρήματα άλλων μελετών ίδιου είδους (π.χ. συγκεντρώνει δύο τυχαιοποιημένες μελέτες που ελέγξαν την αποτελεσματικότητα του Nivolumab στον καρκίνο του πνεύμονα) και τα συνδυάζει σε ένα συμπέρασμα είτε ποιοτικά (ανασκόπηση) είτε/και ποσοτικά (μεταανάλυση). Αυτό το είδος μελέτης θεωρείται οτι κατέχει την κορυφή της πυραμίδας ιεράρχησης των κλινικών μελετών.

Αντιστοιχίζοντας τον τύπο της ερώτησης foreground με τους τύπους των κλινικών μελετών, εάν για παράδειγμα θέλουμε να απαντήσουμε σε μια foreground ερώτηση πρόγνωσης ή κινδύνου-επιβαρυντικού παράγοντα, πρέπει να ανατρέξουμε κατά σειρά προτεραιότητας σε μελέτες κοόρτης, μελέτες ασθενών μαρτύρων και τέλος σε σειρές περιστατικών ή μεμονωμένες αναφορές περιστατικού.

Εάν από την άλλη πλευρά θέλουμε να δώσουμε απάντηση σε μια foreground ερώτηση θεραπείας ή διάγνωσης, πρέπει να ανατρέξουμε σε τυχαιοποιημένες κοντρολαρισμένες μελέτες (Randomised Controlled trials-RCTs).

Ανεξάρτητα από το είδος της foreground ερώτησης που θα θέσουμε, το κορυφαίο είδος μελέτης που μπορούμε να χρησιμοποιήσουμε για να την απαντήσουμε είναι οι συστηματικές ανασκοπήσεις και μεταναλύσεις. Ιδανικά, μια σύγχρονη



Σχήμα 1: Η πυραμίδα της ιεραρχίας των ιατρικών μελετών σύμφωνα με την Ιατρική Βασισμένη σε Στοιχεία. Κατά σειρά αξίας των στοιχείων: μετα-αναλύσεις, συστηματικές ανασκοπήσεις, τυχαιοποιηένες ελεγχόενες μελέτες, μελέτες κοόρτης, μελέτες ασθενών μαρτύρων, αναφορές περιστατικών, μελέτες σε ζώα και in vitro μελέτες, βλ. 1.2.1.4

συστηματική ανασκόπηση με μεταανάλυση που εξέτασε επακριβώς το ερώτημα που θέσαμε μπορεί να παρέχει ακριβώς την πληροφορία που χρειάζεται ο κλινικός ιατρός[1].

1.2.2 To BioASQ Challenge

Το BioASQ είναι ένας διαγωνισμός κατασκευής λογισμικού επάνω στην σημασιολογική καταλογογράφηση ιατροβιολογικών εννοιών (semantic indexing) καθώς και στην αυτοματοποιημένη σύνθεση απαντήσεων σε ερωτήσεις ιατροβιολογικού περιεχομένου [9]. Οι διαγωνιζόμενοι πρέπει να χαρακτηρίσουν σημασιολογικά, αρχεία από μεγάλες βάσεις ιατροβιολογικών δεδομένων (π.χ. Medline) και να δημιουργήσουν απαντήσεις σε πραγματικές βιοϊατρικές ερωτήσεις, στα αγγλικά. Το BioASQ λειτουργεί σε ετήσια βάση από το 2013 λαμβάνοντας επιχορηγήσεις από την Ευρωπαική Ένωση, το Εθνικό Ινστιτούτο Υγείας των Η.Π.Α. καθώς και από την εταιρεία Google[10].

Το BioASQ Challenge μεταξύ άλλων, προσπαθεί να αντιμετωπίσει ένα πρόβλημα που καθημερινά αντιμετωπίζουν οι εργαζόμενοι στον τομέα της Υγείας: την ανεύρεση και τη σύνθεση πληροφορίας πάνω σε ένα θέμα, από πολλές, μεγάλες και καθημερινά μεταβαλλόμενες, συν τω χρόνω όλο και μεγαλύτερες βάσεις δεδομένων. Οι υπάρχουσες μηχανές αναζήτησης βάσεων δεδομένων, όπως το Pubmed και το GoPubmed, αντιμετωπίζουν μόνο μερικώς το πρόβλημα αυτό. Η αναζήτηση απαντήσεων σε ερωτήσεις απαιτεί τη προσεκτική δόμηση αναζητήσεων για την εύρεση των βέλτιστων απαντήσεων που πολλές φορές ακόμη και αν γίνει σωστά, δεν επιφέρει τα επιθυμητά αποτελέσματα[11]. Επιπρόσθετα, ο χρήστης των μηχανών

αναζήτησης πρέπει να φιλτράρει, να διαβάσει και να συνθέσει μόνος του τις πληροφορίες μέσα από διαφορετικές ενδεχομένως πηγές έτσι ώστε να παράγει ορθές και κατανοητές απαντήσεις επάνω στο υπό ερώτηση ζήτημα.

Στα προβλήματα αυτά έρχονται να συνεπικουρήσουν τα συστήματα αυτοματοποιημένων απαντήσεων (QAS -Question answering systems) που προκύπτουν μέσα από τη συμμετοχή διαφορετικών ομάδων στο BioASQ challenge[12]. Η σημασιολογική καταλογογράφηση, δηλαδή η επισήμανση κειμένων με έννοιες από ήδη καθιερωμένες εννοιολογικές ταξονομίες ή οντολογίες, παρέχει ένα μέσο για να συνδυαστούν διαφορετικές πηγές και διευκολύνει το ταίριασμα ερωτήσεων-απαντήσεων. Πρόσφατα έχουν αναπτυχθεί διάφορες μέθοδοι που χρησιμοποιούν υπάρχουσες ταξονομίες για να καταλογογραφήσουν κείμενα και να πραγματοποιήσουν σημασιολογική αναζήτηση. Το BioASQ πάνω σε αυτή την ανάγκη έχει δημιουργήσει λεπτομερώς ορισμένες δοκιμασίες που εντάσσονται περαιτέρω σε ένα σύστημα αξιολόγησης, αποσκοπώντας στην βελτιστοποίηση των μεθόδων αυτών με σκοπό την βελτίωση των συστημάτων QA. Λόγω της κατά σειρά ετών λειτουργίας του διαγωνισμού, το BioASQ κάθε χρόνο δημιουργεί ρεαλιστικά, υψηλής ποιότητας σύνολα δεδομένων και καθορίζει μέτρα αξιολόγησης των διαγωνιζόμενων συστημάτων[12].

1.2.2.1 Δομή του BioASQ

Το BioASQ challenge διαχωρίζεται σε δύο μεγάλες υποομάδες δοκιμασιών, το Task A και το Task B. Το Task A στοχεύει στη δημιουργία συστημάτων για την αυτοματοποιημένη εξαγωγή οντολογιών από επιστημονικές μελέτες που δημοσιεύονται στο PubMed. Το Task B, το οποίο θα αποτελέσει επίκεντρο της παρούσας μελέτης στοχεύει στη δημιουργία συστημάτων αυτοματοποιημένης παραγωγής απαντήσεων σε βιοϊατρικές ερωτήσεις.

1.2.2.2 Οι ερωτήσεις του Bioasq Challenge

Η παρούσα εργασία άντλησε το σύνολο δεδομένων της από τις ερωτήσεις του Task B. Οι ερωτήσεις που έχουν δημοσιευθεί ως τώρα στο Task B του BioASQ είναι 3200 στον αριθμό (σε αυτό τον αριθμό συμπεριλαμβάνονται και οι ερωτήσεις εκπαίδευσης που περιέχουν και ιδανικές απαντήσεις, όπως αυτές ορίζονται στον διαγωνισμό αλλά και ερωτήσεις δοκιμής). Το σύνολο των ερωτήσεων έχει κατασκευαστεί από ειδικούς του βιοϊατρικού τομέα στην Ευρώπη και ως προς το περιεχόμενο διατρέχει ένα ευρύ φάσμα θεματολογιών στο χώρο της Βιολογίας, της Ιατρικής, της Βιοπληροφορικής και συναφών πεδίων [10]. Οι ερωτήσεις μπορούν να διαχωριστούν και με βάση το περιεχόμενό τους αλλά και το είδος της βιβλιογραφίας βάσει του οποίου μπορεί να γίνει βέλτιστη σύνθεση μιας απάντησης. Στα πλαίσια της παρούσας εργασίας, ένας πρώτος διαχωρισμός των ερωτήσεων θα μπορούσε να ξεχωρίζει τις ερωτήσεις σε:

- Ιατρικές
- Βιολογικές
- Άλλες

Οι Ιατρικές ερωτήσεις, περιλαμβάνουν ερώτηση για περιεχόμενο σχετικό με ασθένειες, επιδημιολογία, συμπτώματα, παθογένεια ασθενειών, φάρμακα, θεραπείες, προγνωστικούς παράγοντες και συσχετίσεις ασθενειών μεταξύ τους.

Οι Βιολογικές, περιλαμβάνουν ερωτήσεις για περιεχόμενο σχετικό Βιολογία και τους κλάδους της (κυτταρική και μοριακή, συστημική βιολογία κ.α.) αλληλεπιδράσεις μορίων, εργαστηριακές μεθόδους κ.α.

Πέρα των δύο αυτών κατηγοριών το BioASQ challenge περιέχει πληθώρα άλλων ερωτήσεων, η συστηματική ταξινόμηση των οποίων ξεφεύγει από το σκοπό της παρούσας εργασίας(Φαρμακολογία, Βιοπληροφορική, Ειδήσεις και Νέα του βιοϊατρικού χώρου κ.α.).

Οι δημοσιευμένες ερωτήσεις του BioASQ ανακοινώνονται σε αρχεία της μορφής .json[10] και έχουν την εξής δομή:

- **body**: Το σώμα της ερώτησης, η ερώτηση δηλαδή αυτή καθαυτή

[10] JavaScript Object Notation

- **type**: Ο τύπος της ερώτησης που καθορίζει το ποιά πρέπει να είναι η δομή της απάντησης που αντιστοιχεί σε αυτή. Οι τύποι των ερωτήσεων είναι οι εξής: "yes/no" όπου η απάντηση έχει τη μορφή Ναι/Όχι, "factoid" όπου η απάντηση έχει τη μορφή λίγων λέξεων(ονόματα, φράσεις) ,"list" όπου η απάντηση έχει τη μορφή μιας λίστας λέξεων και "summary" όπου η απάντηση είναι περιγραφικού τύπου με μορφή περίληψης

- **id**: Ο μοναδικός κωδικός της ερώτησης ο οποίος χρησιμοποιείται για indexing

1.3 Προηγούμενες μελέτες επάνω στην ανάλυση βιοϊατρικών ερωτήσεων

Αρκετές ερευνητικές ομάδες έχουν προσπαθήσει στο παρελθόν να κατηγοριοποιήσουν κλινικές ερωτήσεις, άλλες σαν αυτόνομο εγχείρημα ενώ άλλες ως ένα από τα αρχικά βήματα ενός συστήματος αυτοματοποιημένων απαντήσεων σε κλινικές ερωτήσεις[13].

Μια μελέτη επιχείρησε να διαχωρήσει κλινικές ερωτήσεις σε κατηγορίες με βάση ανθρώπους-κριτές, φτάνοντας σε επίπεδα συμφωνίας μεταξύ των κριτών (δείκτη Cohen) 0.53[14]. Οι κριτές ανέδειξαν τρεις τύπους κλινικών ερωτήσεων, αυτές που αφορούν θεραπεία, αυτές που αφορούν συμπτώματα και αυτές που αφορούν διαγνωστικές δοκιμασίες[14]. Οι Seol και συνεργάτες αναλύοντας κλινικές

ερωτήσεις όρισαν τέσσερις κατηγορίες, που συμπίπτουν με τις κατηγορίες foreground ερωτήσεων που υιοθετήθηκαν στην παρούσα μελέτη(θεραπεία, διάγνωση, πρόγνωση, συσχέτιση)[15]. Άλλοι ερευνητές όρισαν εκ των προτέρων διάφορους τύπους κλινικών ερωτήσεων και στη συνέχεια προσάρμοσαν την εκάστοτε κλινική ερώτηση στη δομή μιας εκ των κατηγοριών αυτών[16].

Οι κλινικές ερωτήσεις σπάνια περιέχουν όλα τα στοιχεία PICO, όπως διαπιστώθηκε από μια μελέτη ανίχνευσης τους από ανθρώπους/κριτές[17]. Μια άλλη ερευνητική ομάδα επιχείρησε να ανιχνεύσει τα στοιχεία αυτά σε περιλήψεις ιατρικών μελετών χρησιμοποιώντας διάφορους ταξινομητές με αρκετή επιτυχία[18]. Πιο συγκεκριμένα πετύχαν f-scores 86.3% για το P, 67% για το I και 56.6% για το O. Μια άλλη ομάδα κατάφερε να ανιχνεύσει κάποια από τα στοιχεία PICO (το P) από την πρώτη πρόταση μιας ιατρικής μελέτης με μεγαλύτερη ακρίβεια σε σχέση με το να εξέταζε την ιατρική μελέτη στο σύνολό της[19]. Πρόσφατα ερευνητές δημιούργησαν ένα μοντέλο ανίχνευσης των στοιχείων PICO μεγάλης ακρίβειας, βασισμένο στα νευρωνικά δίκτυα, που κατηγοριοποιεί κάθε πρόταση σε μια περίληψη μελέτης λαμβάνοντας υπόψιν και το πλαίσιο της και όχι αξιολογώντας την ως μεμονωμένη[20]. Τα στοιχεία PICO έχουν επίσης χρησιμοποιηθεί ως στοιχεία για την αναζήτηση και την άντληση σχετικών μελετών[21]. Όλες οι προαναφερθείσες μελέτες χρησιμοποίησαν για τη δημιουργία του συνόλου δεδομένων τους

περιλήψεις ιατρικών μελετών.

Για να αποφύγουν το θόρυβο πληροφορίας που προκύπτει από το ελεύθερο κείμενο στις κλινικές ερωτήσεις, η ομάδα που δημιούργησε το σύστημα αυτοματοποιημένων απαντήσεων AskHERMES δημιούργησαν ένα 'καλούπι' της μορφής PICO όπου ο ιατρός καλείται να συμπληρώσει τα στοιχεία της ερώτησης στα αντίστοιχα πεδία[22]. Αυτή η μέθοδος βέβαια περιορίζει αρκετά τους βαθμούς ελευθερίας που μπορεί να χρειάζεται να υπάρχουν σε μια σύνθετη κλινική ερώτηση[22].

Οι Niu και συνεργάτες εξέτασαν την εξαγωγή σημασιολογικών οντοτήτων από ιατρικές ερωτήσεις και εξέτασαν τη σχέση αυτών σε μια κλινική ερώτηση/πρόταση πετυχαίνοντας ποσοστό 83% σε ακρίβεια και ανάκληση στη δοκιμασία εύρεσης της ύπαρξης του στοιχείου O [23]. Σύμφωνα με τη γνώση του γράφοντος αυτή είναι η μοναδική μελέτη που επιχείρησε να εντοπίσει στοιχεία PICO σε κλινικές ερωτήσεις με αυτοματοποιημένο τρόπο.

Επιπροσθέτως, η κατηγοριοποίηση κλινικών ερωτήσεων με τη βοήθεια της μηχανικής μάθησης σε ερωτήσεις τύπου foreground και background καθώς και η κατηγοριοποίηση των ερωτήσεων foreground με τη σειρά τους σε ερωτήσεις T, D, P, H (treatment, diagnosis, prognosis, harm) επιχειρήθηκε για πρώτη φορά στην παρούσα μελέτη. Τέλος, η κατηγοριοποίηση του υποσυνόλου των ιατρικών ερωτήσεων που δημοσιεύονται στο BioASQ challenge με σκοπό την ανεύρεση

στοχευμένων απαντήσεων σε αυτές σύμφωνα με τους κανόνες της Ιατρικής Βασισμένης σε Στοιχεία (Evidence Based Medicine) επιχειρήθηκε επίσης, για πρώτη φορά στην παρούσα μελέτη.

Συνοπτικά, στην παρούσα εργασία επιχειρήθηκε για πρώτη φορά η ανάλυση και κατηγοριοποίηση ερωτήσεων κλινικού περιεχομένου του BioASQ challenge. Καταρχήν, επιχειρήθηκε επίσης για πρώτη φορά η κατηγοριοποίηση κλινικών ερωτήσεων με τη βοήθεια της Μηχανικής Μάθησης σε ερωτήσεις τύπου foreground και background, όπως επίσης και σε ερωτήσεις θεραπείας, διάγνωσης, πρόγνωσης και επιβάρυνσης. Σινάμα, επιχειρήθηκε και η κατηγοριοποίηση των ερωτήσεων με βάση τα στοιχεία PICO, ένα εγχείρημα που είχαν αναλάβει και άλλες ερευνητικές ομάδες στο παρελθόν, κυρίως στα πλαίσια περιλήψεων κλινικών μελετών

1.4 Μηχανική Μάθηση(Machine learning)

1.4.1 Εισαγωγή

Ο τομέας της Μηχανικής Μάθησης (Machine learning-ML) αποτελεί έναν από τους κλάδους του τομέα της Τεχνητής Νοημοσύνης (Artificial Intelligence-AI)[24]. Το ευρύτερο πεδίο της AI περιγράφει το σύνολο πεδίων της Επιστήμης των Υπολογιστών στο οποίο οι υπολογιστές "μιμούνται" μέσω ειδικών αλγορίθμων ανθρώπινες "γνωστικές" λειτουργίες όπως η "μάθηση" και η "επίλυση προβλη-

μάτων". Η Μηχανική Μάθηση άρχισε να αναδεικνύεται ως ξεχωριστό πεδίο στους κόλπους της Τεχνητής Νοημοσύνης, καθώς οι αλγόριθμοι που κατασκευάζονταν αρχικά σε ακαδημαικά πλαίσια ανακατευθύνθηκαν από διαφορετικές ερευνητικές ομάδες στην επίλυση προβλημάτων του πραγματικού κόσμου (real world problems) μέσω της εκπαίδευσης υπολογιστών από πραγματικά σύνολα δεδομένων.

Ο αρχικός ορισμός της Μηχανικής Μάθησης (Machine Learning) τέθηκε το 1959 από τον Arthur Samuel ως το πεδίο της επιστήμης των υπολογιστών που μελετά και βασίζεται στην δυνατότητα των υπολογιστών να μαθαίνουν να εκτελούν μία εργασία χωρίς να έχουν προγραμματιστεί εντελώς (explicitly) για αυτήν [25]. Πιο απλά, αυτός ο ορισμός διαχωρίζει τη Μηχανική Μάθηση από την "παραδοσιακή" επιστήμη των υπολογιστών που κάνει χρήση αλγορίθμων που ορίζουν επακριβώς το τι θα κάνει ο υπολογιστής για να πραγματοποίσει κάποιον υπολογισμό ή να λύσει κάποιο πρόβλημα. Η Μηχανική Μάθηση αντίθετα, κάνει χρήσει αλγορίθμων που επιτρέπουν στους υπολογιστές να εκπαιδεύονται επάνω σε δεδομένα εισόδου και μέσω μαθηματικής-στατιστικής ανάλυσης να παράγουν έξοδο, ανάλογη του προβλήματος[26].

Ο Tom Mitchell το 1997, απέδωσε στην Μηχανική Μάθηση έναν πιο σύγχρονο ορισμό, ο οποίος είναι: "Ένα πρόγραμμα θα λέγεται ότι μαθαίνει από μία εμπειρία E ως προς ένα σύνολο εργασιών T και ένα μέτρο απόδοσης P, αν η απόδοσή του σε

εργασίες του συνόλου εργασιών T, όπως προσδιορίζεται από το μέτρο απόδοσης P, βελτιώνεται διαμέσου της εμπειρίας E"[26].

Το πρώτο πρόγραμμα που μπορούσε να "μάθει" κατά τη διάρκεια της εκτέλεσής του ήταν ένα πρόγραμμα που μπορούσε να παίξει ντάμα και δημιουργήθηκε από τον πρωτοπόρο Arthur Samuel το 1952. Το 1958 ο Frank Rosenblatt δημιούργησε το πρώτο τεχνητό νευρωνικό δίκτυο (βλ. 1.4.3.5). Σταδιακά, στατιστικές μέθοδοι άρχισαν να τροποποιούνται και να ενσωματώνονται σε νέα εργαλεία. Από τη δεκαετία του 1990 έως και σήμερα, το πεδίο της Μηχανικής Μάθησης είναι ένα συνεχώς και ταχέως εξελισσόμενο πεδίο έρευνας με τεράστιο πλήθος αλγορίθμων αλλά και ευρύτατο φάσμα εφαρμογής (Επιστήμη, Οικονομία, Υγεία κ.α.) καθώς έχει καταστεί σαφές ότι με την αρωγή των μοντέλων της Μηχανικής Μάθησης μπορεί να επιτευχθεί βελτιστοποίηση πολλαπλών παραμέτρων προς όφελος του ανθρώπου και της κοινωνίας.

Οι τύποι αλγορίθμων μάθησης που χρησιμοποιούνται στα πλαίσια της Μηχανικής Μάθησης διαφέρουν ως προς τη μέθοδο που χρησιμοποιούν, τον τύπο δεδομένων που δέχονται ως είσοδο και παράγουν ως έξοδο αλλά και τον τύπο προβλημάτων που καλούνται να λύσουν[26]. Αδρά χωρίζονται στις εξής ομάδες:

-Εποπτευόμενη μάθηση(Supervised learning)

Οι αλγόριθμοι εποπτευόμενης μάθησης δημιουργούν ένα μαθηματικό μοντέλο από

ένα σύνολο δεδομένων που περιέχει και τα δεδομένα εισόδου αλλά και τα δεδομένα εξόδου. Ιδανικά ο αλγόριθμος πρέπει μετά τη διαδικασία εκπαίδευσης να προβλέπει την έξοδο για δεδομένα εισόδου που δεν αποτελούσαν κομμάτι του συνόλου δεδομένων εκπαίδευσης.

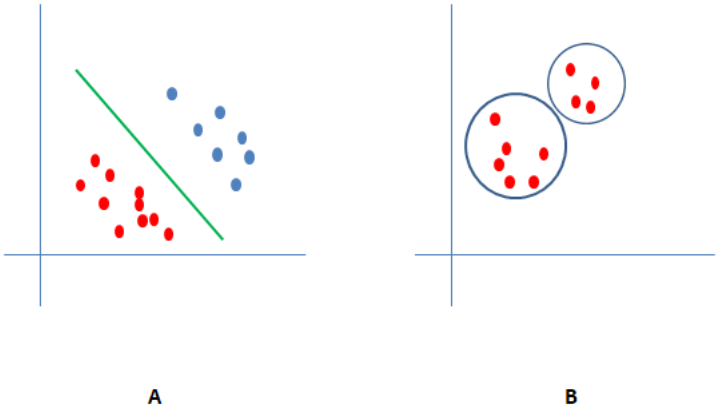
Οι αλγόριθμοι εποπτευόμενης μάθησης περιέχουν τους αλγόριθμους ταξινόμησης (οι αλγόριθμοι που θα χρησιμοποιηθούν στην παρούσα εργασία ανήκουν σε αυτήν την κατηγορία) και τους αλγόριθμους παλινδρόμησης. Στην περίπτωση της ταξινόμησης, οι δυνατές τιμές εξόδου περιορίζονται σε ένα πεπερασμένο σύνολο τιμών (κατηγορίες, classes), ενώ στην περίπτωση της παλινδρόμησης η τιμή εξόδου μπορεί να είναι οποιαδήποτε αριθμητική τιμή από ένα σύνολο τιμών.

-Μη-εποπτευόμενη μάθηση(Unsupervised learning)

Οι αλγόριθμοι μη εποπτευόμενης μάθησης δέχονται σύνολα δεδομένων που περιέχουν μόνο τις εισόδους και βρίσκουν οι ίδιοι τη δομή των δεδομένων δημιουργόντας ομάδες (clusters) στιγμιοτύπων που "μοιάζουν" μεταξύ τους.

-Άλλοι τύποι μάθησης

Υπάρχει πλήθος άλλων ομάδων αλγορίθμων Μηχανικής Μάθησης. Ενδεικτικά, υπάρχει η ημιεπιβλεπόμενη μάθηση όπου τμήμα του συνόλου δεδομένων περειέχει και την είσοδο αλλά και την έξοδο, ενώ το υπόλοιπο περιέχει μόνο την είσοδο. Η μάθηση θετικής ανατροφοδότησης περιλαμβάνει τη βελτιστοποίηση συμπεριφοράς



Σχήμα 2: Εποπτευόμενη (A) και μη-εποπτευόμενη (B) μάθηση. Στην A περίπτωση, ο αλγόριθμος βρίσκει τη βέλτιστη διαχωριστική γραμμή μεταξύ δύο προκαθορισμένων κλάσεων στιγμιοτύπων. Στην B περίπτωση, ο αλγόριθμος κατηγοριοποιεί τα στιγμιότυπα σύμφωνα με την "ομοιότητά" τους σε κλάσεις που ο ίδιος δημιουργεί

αλγορίθμων σε διαφορετικά περιβάλλοντα με σκοπό την μέγιστη αθροιστική συγκέντρωση επιβράβευσης (cumulative reward). Τέλος ενδεικτικά να αναφέρουμε και τη μάθηση χαρακτηριστικών η οποία στοχεύει στην βελτιστοποίηση της αναπαράστασης των δεδομένων εισόδου κατά τη διαδικασία εκπαίδευσης.

Στην παρούσα εργασία το σύνολο των αλγορίθμων Μηχανικής Μάθησης που θα χρησιμοποιηθούν ανήκει στο χώρο της ταξινόμησης της επιβλεπόμενης Μηχανικής Μάθησης.

1.4.2 Σύνολο δεδομένων, σύνολο εκπαίδευσης, σύνολο αξιολόγησης)

Κατά τη διαδικασία κατασκευής ενός μοντέλου Μηχανικής Μάθησης χρησιμοποιούνται δεδομένα που προέρχονται από διάφορα σύνολα δεδομένων[26]. Συνήθως, τα σύνολα δεδομένων διαχωρίζονται στις εξής τρεις κατηγορίες:

- Σύνολο εκπαίδευσης(training set): Αυτό το σύνολο δεδομένων χρησιμοποιείται στην εκπαίδευση και οριστικοποίηση των παραμέτρων μοντέλου
- Σύνολο εγκυρότητας(validation set): Αυτό το σύνολο δεδομένων χρησιμοποιείται για την αποφυγή υπερπροσαρμογής των δεδομένων με ταυτοχρονη απώλεια της γενικευσιμότητας (ρύθμιση των υπερπαραμέτρων του μοντέλου όπως π.χ. αριθμώς κρυφών νευρώνων σε ένα νευρωνικό δίκτυο)
- Σύνολο αξιολόγησης(test set): Αφότου εκπαιδευθεί το υπό μελέτη μοντέλο, στη συνέχεια αξιολογείται η απόδοσή του με βάση την επιτυχία του στο να προβλέπει τις σωστές εξόδους για τις εισόδους που δέχεται από το σύνολο αξιολόγησης. Το σύνολο αξιολόγησης μπορεί να αποτελεί τμήμα του συνόλου εκπαίδευσης, μπορεί όμως να αποτελεί και ξεχωριστό σύνολο δεδομένων.

Υπάρχουν διάφορες μέθοδοι για την υλοποίηση του διαχωρισμού του συνόλου δεδομένων στα προαναφερθέντα υποσύνολα. Στην παρούσα εργασία ο διαχωρισμός

έγινε με βάση τη μέθοδο stratified k-fold cross validation (βλ.2.4)

1.4.3 Ταξινόμηση και ταξινομητές(Classification-Classifiers)

Μια δοκιμασία η οποία έχει φανεί οτι μπορεί να εκτελεστεί με μεγάλη επιτυχία από εργαλεία της μηχανικής μάθησης είναι η ταξινόμηση (classification) [27]. Γενικά, η ταξινόμηση αποτελεί τη διαδικασία κατηγοριοποίησης ενός προτύπου (pattern/instance) σε διάφορες γνωστές κατηγορίες (categories,≥2 στον αριθμό). Παραδείγματος χάριν στην παρούσα εργασία μία από τις δοκιμασίες ταξινόμησης αποτελεί η ταξινόμηση ιατρικών ερωτήσεων (instances) στις κατηγορίες background και foreground (categories). Αντίστοιχες δοκιμασίες κατηγοριοποίησης μπορούν να υλοποιηθούν σε κάθε τομέα της ανθρώπινης δραστηριότητας.

Ένα σύστημα που μπορεί να φέρει εις πέρας μια διαδικασία ταξινόμησης, λαμβάνει σαν είσοδο ένα σύνολο αριθμητικών τιμών (feature values)το οποίο προκύπτει από ιδιότητες του προς ταξινόμηση instance, εν συνεχεία υλοποιεί τη διαδικασία ταξινόμησης και σαν έξοδο παράγει την κατηγορία στην οποία κατέταξε το στιγμιότυπο (output).

Συχνά, επιλέγεται η αναπαράσταση ενός διανύσματος, το οποίο ονομάζεται διάνυσμα χαρακτηριστικών (feature vector). Κάθε στιγμιότυπο χαρακτηρίζεται από F ζευγάρια τιμών (χ,y), όπου F είναι ο αριθμός των χαρακτηριστικών, έτσι

ώστε κάθε στιγμιότυπο να έχει μία συγκεκριμένη θέση σε έναν F-διάστατο χώρο που ονομάζεται χώρος των χαρακτηριστικών (feature space). Σε ένα στιγμιότυπο, το x αντιστοιχεί στην αριθμητική τιμή ενός χαρακτηριστικού και ονομάζεται γνώρισμα (attribute) και το y αντιστοιχεί στο όνομα της κατηγορίας στην οποία ανήκει το στιγμιότυπο (class label).

Πολλές φορές κάποια χαρακτηριστικά δε συμβάλλουν στην ταξινόμηση των στιγμιοτύπων σε κατηγορίες, και όταν το πλήθος των χαρακτηριστικών είναι μεγάλο, πραγματοποιείται ένα επιπλέον βήμα, αυτό της επιλογής των πιο σημαντικών χαρακτηριστικών (feature selection). Στο παράδειγμα που προαναφέρθηκε για την ταξινόμηση ερωτήσεων στις δύο κατηγορίες, ως χαρακτηριστικά μιας πρότασης (features) θα μπορούσε κάποιος να καταχωρήσει τον αριθμό, τη σειρά και τη συχνότητα των λέξεων, γραμματικές ή/και συντακτικές ιδιότητες της πρότασης κ.ο.κ. Το γεγονός οτι κάθε ερώτηση τελειώνει με ερωτηματικό, σαν χαρακτηριστικό μιας ερώτησης-στιγμιοτύπου δεν προσφέρει πληροφορία αξιοποιήσιμη για την βελτίσωση της διαδικασίας ταξινόμησης, καθώς όλες οι ερωτήσεις instances φέρουν ερωτηματικό στο πέρας τους, ανεξάρτητα από το αν ανήκουν εν τέλει στην κατηγορία background ή foreground.

Στις επόμενες τέσσερις ενότητες (1.4.3.1 έως και 1.4.3.5) θα παρουσιαστούν τέσσερα είδη κεντρικής σημασίας ταξινομητών που χρησιμοποιήθηκαν στις ταξι-

νομήσεις της παρούσας εργασίας: ο ταξινομητής Dummy, ο ταξινομητής Naive Bayes, τα Τυχαία Δάση(Random Forests), οι Μηχανές Διανυσματικής Υποστήριξης(Support Vector Machines - SVM) και τα Νευρωνικά Δίκτυα (Neural Networks).

-

1.4.3.1 Ταξινομητής Dummy Ο ταξινομητής dummy, είναι ένας ταξινομητής ο οποίος χρησιμοποιείται ως ταξινομητής αναφοράς (baseline classifier) για τη σύγκριση της απόδοσης άλλων ταξινομητών. Ο ταξινομητής dummy αντί να χρησιμοποιήσει χαρακτηριστικά (features) των στιγμιοτύπων για να εκπαιδευθεί στην ταξινόμησή τους και να την υλοποιήσει σε ένα σύνολο δοκιμής, ακολουθεί απλούς κανόνες για να κατανήμει τα στιγμιότυπα σε κατηγορίες. Οι απλοί αυτοί κανόνες για παράδειγμα μπορεί να είναι:

- Κατηγοριοποίηση των παρατηρήσεων στην πιο συχνή κατηγορία (πάντα, η ταξινόμηση κάθε νέας παρατήρησης θα κατηγοριοποιείται ως αυτή που απαντάται συχνότερα στο σύνολο εκπαίδευσης)
- Ομοιόμορφη κατανομή στις κατηγορίες
- Διαστρωμάτωση (stratification) δηλαδή κατανομή των παρατηρήσεων σε κατηγορίες με διατήρηση της πιθανότητας κατανομής που παρατηρείται στο σύνολο εκπαίδευσης

Η υλοποίηση του ταξινομητή Dummy στην παρούσα εργασία έγινε μέσω του module του DummyClassifier του Scikit learn (βλ. 2.3).

1.4.3.2 Ταξινομητής Naive Bayes Οι ταξινομητές Naive Bayes είναι πιθανοκρατικοί ταξινομητές που βασίζονται στο θεώρημα του Bayes για την ταξινόμηση στιγμιοτύπων σε κλάσεις[25]. Ονομάζονται αθώοι (naive) καθώς λειτουργούν με την παραδοχή ότι δεδομένης της κλάσης ενός στιγμιότυπου, κάθε χαρακτηριστικό του είναι ανεξάρτητο από όλα τα υπόλοιπα χαρακτηριστικά δηλαδή αριθμητική τιμή ενός χαρακτηριστικού σε ένα στιγμιότυπο δεν επηρεάζει την αριθμητική τιμή οποιουδήποτε άλλου. Πιο συγκεκριμένα, όπως προκύπτει από το θεώρημα δεσμευμένης πιθανότητας του Bayes, εάν Ck είναι το ενδεχόμενο του να ανήκει μια παρατήρηση στην κλάση k και χ ένα χαρακτηριστικό των στιγμιοτύπων x έχουμε:

$$P(\mathbf{Ck}|\mathbf{x}) = P(\mathbf{Ck})\frac{P(\mathbf{x}|\mathbf{Ck})}{P(\mathbf{x})} \tag{1}$$

Αν επεκτείνουμε τον κανόνα αυτό σε πολλά χαρακτηριστικά $(x_1,\ldots,x_n)$ προκύπτει ότι:

$$\mathrm{p}(\mathrm{C}_k \mid x_1,\ldots,x_n)p(C_k,x_1,\ldots,x_n) = p(C_k)\ p(x_1 \mid C_k)\ p(x_2 \mid C_k)\ p(x_3 \mid$$

$C_k)$ $\cdots$

Στην περίπτωση που έχουμε δεδομένα κειμένου και χρησιμοποιούμε τη συχνότητα λέξεων ως χαρακτηριστικά κάθε στιγμιοτύπου/ερώτησης όπως στην παρούσα εργασία, η παραπάνω εξίσωση μεθερμηνεύεται ως εξής. Η πιθανότητα μιας πρότασης να ανήκει σε μία κλάση προτάσεων, ισούται με την πιθανότητα του να ανήκει σε μια κλάση προτάσεων χωρίς να ξέρουμε κανένα χαρακτηριστικό (αρχική κατανομή των προτάσεων σε κλάσεις) επί την πιθανότητα του να περιέχει μια συγκεκριμένη λέξη δεδομένου ότι ανήκει σε αυτή την κλάση κ.ο.κ.

Τελικό βήμα στη διαδικασία ταξινόμησης αποτελεί η αντιστοίχηση των πιθανοτήτων σε ταξινομήσεις των στιγμιοτύπων σε κλάσεις. Αυτό επιτυγχάνεται με κάποιους κανόνες απόφασης που συνήθως συγκρίνουν τις υπολογισμένες πιθανότητες για το ανήκειν ενός στιγμιοτύπου σε κάθε κλάση και στη συνέχεια ταξινομούν το στιγμιότυπο στην κλάση με τη μεγαλύτερη πιθανότητα.

Η απόδοση των ταξινομητών Bayes έχει φανεί οτι πλησιάζει την απόδοση των μηχανών διανυσματικής υποστήριξης σε δοκιμασίες κατηγοριοποίησης κειμένου που χρησιμοποιούν τη μέθοδο TF-IDF.

Η υλοποίηση του ταξινομητή Naive Bayes στην παρούσα εργασία έγινε μέσω του module του MultinomialNB του Scikit learn (βλ. 2.3).

1.4.3.3 Δέντρα Αποφάσεων και Τυχαία δάση (Decision trees, Random Forests)

Στην παρούσα εργασία χρησιμοποιήθηκαν τα τυχαία δάση με σκοπό την ταξινόμηση των ερωτήσεων. Για την κατανόηση της λειτουργίας του τυχαίου δάσους, πρέπει να προηγηθεί η επεξήγηση της δομικής του μονάδας, του δέντρου αποφάσεων[28]. Το δέντρο αποφάσεων αποτελεί θεμελιώδες εργαλείο της Μηχανικής Μάθησης για την ταξινόμηση στιγμιοτύπων σε κλάσεις. Διαισθητικά ένα δέντρο αποφάσεων, προσομοιάζει σε ένα φυσικό δέντρο. Κάθε σημείο διακλάδωσης ονομάζεται κόμβος του δένδρου. Το αρχικό σημείο του δέντρου ονομάζεται αρχικός κόμβος ή ρίζα ενώ κάθε σημείο διακλάδωσης περιφερικότερα της ρίζας ονομάζεται εσωτερικός κόμβος. Οι εξωτερικοί ή αλλιώς τελικοί κόμβοι ή φύλα είναι καταληκτικά σημεία των κλαδιών του δέντρου και αντιστοιχούν ο καθένας σε μια από τις κλάσεις στις οποίες πρέπει να ταξινομηθούν τα στιγμιότυπα του συνόλου δεδομένων. Σε κάθε κόμβο εκτός από τη ρίζα εισέρχεται μια κατευθυνόμενη ακμή από άλλον κόμβο.

Κατά τη διαδικασία ταξινόμησης, γίνεται διάβαση του δέντρου αποφάσεων από την ρίζα έως κάποιο από τα φύλλα. Το ποιά από τις εξερχόμενες ακμές κάθε κόμβου θα επιλεγεί κατά τη διάβαση του δέντρου καθορίζεται απο τη λεγόμενη συνθήκη ελέγχου, η οποία συνήθως καθορίζεται με βάση κάποιο από τα χαρακτηριστικά των στιγμιοτύπων.

Σχήμα 3: Παράδειγμα ταξινόμησης στιγμιότυπου με δέντρο αποφάσεων στο πρόβλημα ταξινόμησης της διατήρησης ή απόρριψης προσωπικών αντικειμένων. Η ταξινόμηση γίνεται στις κλάσεις "το κρατάω" και "το πετάω" οι οποίες αντιστοιχούν στα φύλλα του δέντρου, μέσω τριών κόμβων(δύο εσωτερικών)

Η διαδικασία εκπαίδευσης ενός δέντρου αποφάσεων περιλαμβάνει τη δημιουργία ενός μεγάλου αριθμου διαφορετικών δέντρων με διαφορετική συνδεσμολογία κόμβων κάθε φορά. Για κάθε τέτοιο δέντρο υπολογίζεται στη συνέχεια το κόστος που προκύπτει από τις λάθος κατηγοριοποιήσεις στιγμιοτύπων σε κλάσεις και επιλέγεται εν τέλει το δέντρο που ελαχιστοποιεί το κόστος αυτό.

Πολλές φορές ο αριθμός των χαρακτηριστικών είναι πολύ μεγάλος, με αποτέλεσμα ο διαφορετικός αριθμός των πιθανών δέντρων αποφάσεων να είναι τεράστιος. Υπάρχουν δύο ευρέως χρησιμοποιούμενοι τρόποι για την επίλυση αυτού του προβλήματος. Ο πρώτος στοχεύει στον περιορισμό του μέγιστου βάθους που μπορεί να φτάσει ένα δέντρο, δηλαδή στη μέγιστη απόσταση μεταξύ της ρίζας και των φύλων

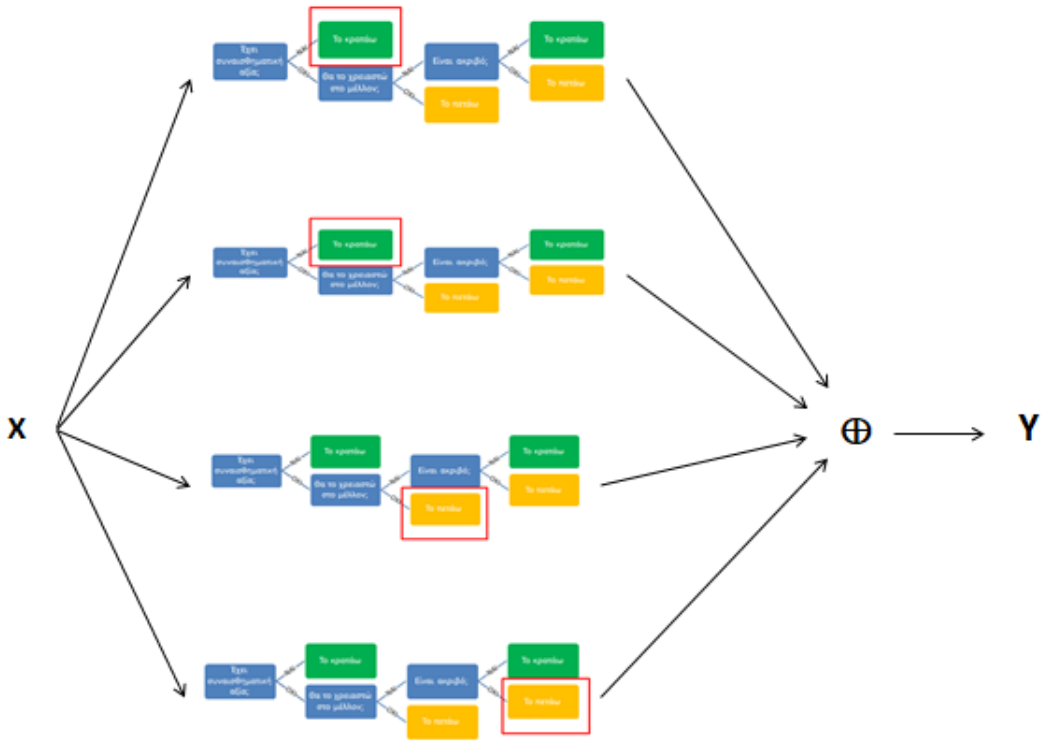
που μπορεί να έχει ένα δέντρο. Ο δεύτερος αποκαλείται pruning και αποκόπτει τα κομμάτια εκείνα του δέντρου που εμπεριέχουν κόμβους με χαρακτηριστικά χαμηλής σημαντικότητας(που δεν έχουν δηλαδή μεγάλη ικανότητα διαχωρισμού των στιγμιοτύπων σε κλάσεις). Στην παρούσα εργασία έγινε χρήση του πρώτου τρόπου, με ρύθμιση της παραμέτρου βάθους στο 40, δηλαδή στην μη κατασκευή δέντρων με απόσταση ρίζας-φύλλου μεγαλύτερη του 40.

Τα τυχαία δάση αποτελούν μια συνδυαστική μέθοδο ταξινόμησης (ensemble method) η οποία χρησιμοποιεί για ταξινομητές διαφορετικά δέντρα απόφασης, και συνδυάζει τις αποφάσεις/ταξινομήσεις αυτών με διάφορους τρόπους. Ο αριθμός των δέντρων που θα αποτελούν το δάσος καθορίζεται εξαρχής στο μοντέλο. Στην παρούσα εργασία κατά σύμβαση δημιουργήθηκαν σε όλες τις ταξινομήσεις δάση των 100 δέντρων. Ο πιό κλασσικός τρόπος καθορισμού της τελικής ταξινόμησης ενός στιγμιοτύπου από το δάσος είναι η επιλογή της κλάσης που επέλεξε η πλειοψηφία των δέντρων απόφασης (majority voting)[29].

Η υλοποίηση του ταξινομητή Random Forest στην παρούσα εργασία έγινε μέσω του module του RandomForestClassifier του Scikit learn (βλ. 2.3).

1.4.3.4 Μηχανές διανυσματικής υποστήριξης (Support Vector Machines - SVM)

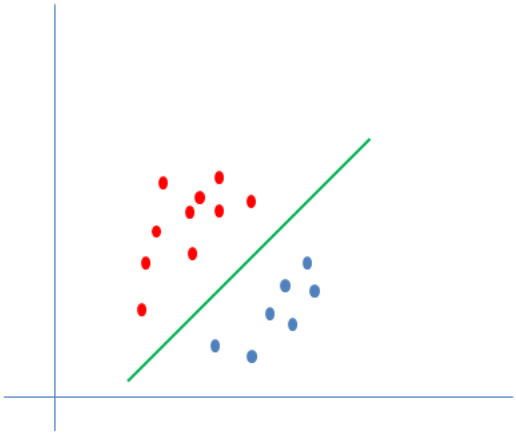
Μια Μηχανή Διανυσματικής Υποστήριξης (Support Vector



Σχήμα 4: Παράδειγμα ταξινόμησης στιγμιότυπου από ένα τυχαίο δάσος με επιλογή της κλάσης που επέλεξε η πλειοψηφία των δέντρων απόφασης. Τα μεμονωμένα δέντρα αποφάσεων είναι αυτά του σχήματος 3

Machines - SVM) είναι ένα είδος ταξινομητή ο οποίος έχει την ιδιότητα να διαχωρίζει στιγμιότυπα σε κλάσεις μέσω διαχωριστικών υπερεπιπέδων. Συγκεκριμένα, δεδομένου ενός κατονομασμένου (labeled) συνόλου στιγμιοτύπων, τα SVM's παράγουν το βέλτιστο υπερεπίπεδο το οποίο κατηγοριοποιεί νέα στιγμιότυπα. Εάν τα στιγμιότυπα είναι γραμμικά διαχωρίσιμα, τότε το υπερεπίπεδο είναι απλώς μια ευθεία που διαχωρίζει τα στιγμιότυπα σε κλάσεις με το κάθε εκατέρωθεν ημιεπίπεδο

που αυτή ορίζει να αποτελεί το χώρο της κάθε κλάσης.

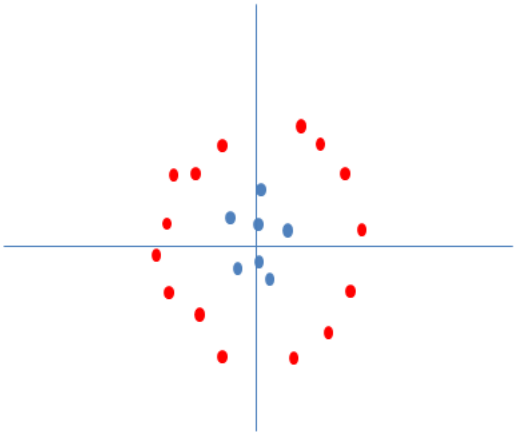


Σχήμα 5: Γραμμικός διαχωρισμός στιγμιοτύπων στο δισδιάστατο χώρο.

Υπάρχει όμως περίπτωση τα στιγμιότυπα δύο κλάσεων να μην μπορούν να διαχωριστούν από μια ευθεία να μην είναι δηλαδή γραμμικά διαχωρίσιμα στον δισδιάστατο χώρο.

Σε αυτή την περίπτωση, η μέθοδος που ακολουθούν τα SVM's είναι να προσθέσουν περισσότερες διαστάσεις στον χώρο ύπαρξης των στιγμιοτύπων.

Προσθέτωντας διαστάσεις τα στιγμιότυπα μπορεί να γίνουν γραμμικά διαχωρίσιμα από μια οντότητα που ονομάζεται υπερεπίπεδο. Τα υπερεπίπεδα μπορούν να εκ-



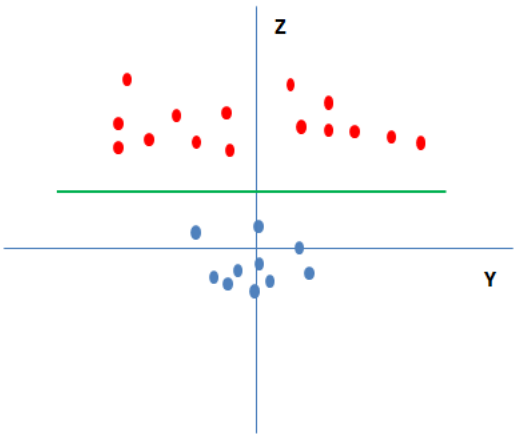
Σχήμα 6: Εδώ τα στιγμιότυπα διαφορετικών κλάσεων δεν είναι γραμμικά διαχωρίσιμα στον δισδιάστατο χώρο. Τα SVM's θα προσπαθήσουν να τα διαχωρήσουν γραμμικά σε πιό υψηλές διαστάσεις.

φραστούν ως το σύνολο των σημείων $\vec{x}$ που ικανοποιούν την παρακάτω σχέση:

$$\vec{w} \cdot \vec{x} - b = 0 \tag{2}$$

όπου $\vec{w}$ το ορθοκανονικό στο υπερεπίπεδο διάνυσμα και b παράμετρος τηςπου προκύπτει από τη διαδικασία εκπαίδευσης.

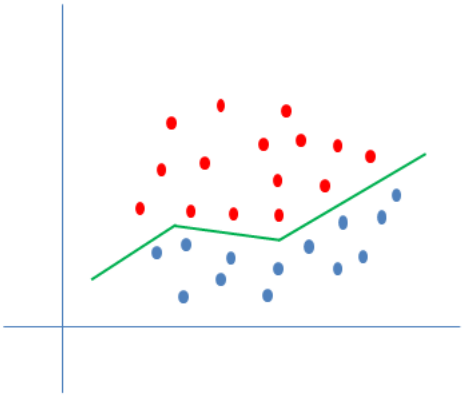
Εάν τα δεδομένα είναι διαχωρίσιμα με μη γραμμικό τρόπο τότε η δημιουργία του υπερεπιπέδου ενσωματώνει κάποιο κόστος λ για κάθε στιγμιότυπο που κατηγοριοποιεί με λανθασμένο τρόπο και ο στόχος είναι το υπερεπίπεδο να είναι τέτοιο που το κόστος λ να ελαχιστοποιείται.



Σχήμα 7: Τα μη γραμμικά διαχωρίσιμα στιγμιότυπα του προηγούμενου σχήματος είναι τώρα γραμμικά διαχωρίσιμα με τη θέασή τους από τις 3 διαστάσεις. Είναι δυνατή η κατασκευή γραμμικού υπερεπιπέδου.

Για να μετασχηματίσουμε το υπερεπίπεδο από τις περισσότερες διαστάσεις που αυτό εμφανίστηκε, στον δισδιάστατο χώρο από τον οποίο ξεκίνησε η όλη διαδικασία, χρησιμοποιείται η λεγόμενη μέθοδος kernel. Ενώ το υπερεπίπεδο στο χώρο πολλών διαστάσεων μπορεί να ήταν γραμμικό, όταν γίνεται μετασχηματισμός στον δισδιάστατο χώρο μέσω του kernel η διαχωριστική καμπύλη που προκύπτει είναι συνήθως μη-γραμμική. Υπάρχουν διάφορες μέθοδοι kernels με τις πιο συχνά χρησιμοποιούμενες να είναι η πολυωνυμική, η εκθετική και radial basis function η οποία και χρησιμοποιήθηκε στα μοντέλα ταξινόμησης της παρούσας εργασίας.

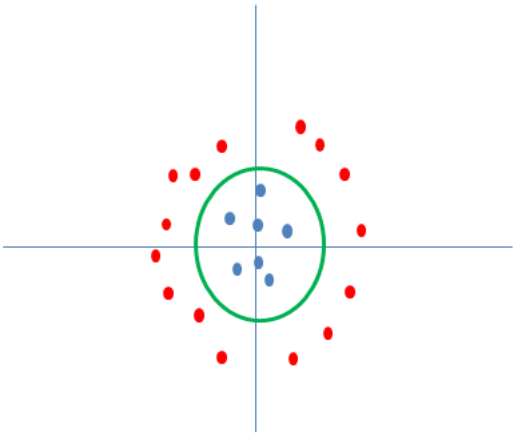
Μια σημαντική παράμετρος των SVM's είναι η παράμετρος gamma. Αυτή



Σχήμα 8: Σε περίπτωση που δεν είναι εφικτός ο γραμμικός διαχωρισμός από υπερεπίπεδο μπορεί να γίνει μη γραμμικός διαχωρισμός στιγμιοτύπων με ελαχιστοποίηση του κόστους λανθασμένης ταξινόμησης

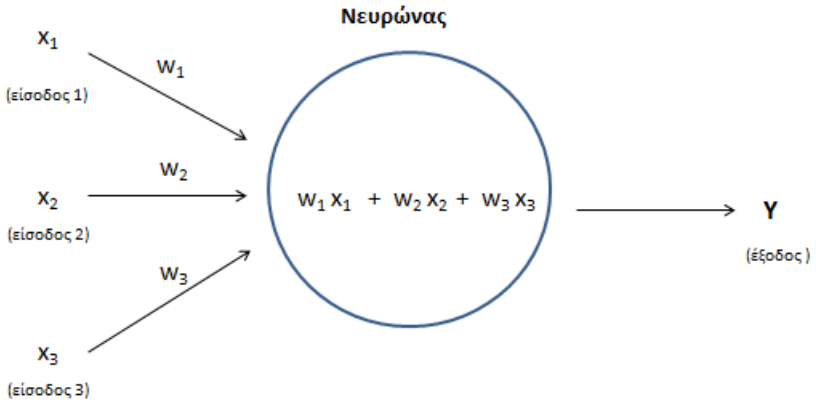
η παράμετρος καθορίζει τον βαθμό επιρροής διαφορετικών (σε απόσταση από το υπερεπίπεδο) στιγμιοτύπων στη διαμόρφωση του διαχωριστικού υπερεπιπέδου. Πιο συγκεκριμένα εάν ένα μοντέλο SVM έχει μεγάλο gamma τότε μόνο τα κοντινά στο υπερεπίπεδο στιγμιότυπα επηρεάζουν τη διαμόρφωση του υπερεπιπέδου, ενώ σε μικρότερες τιμές του gamma απομακρυσμένες από το υπερεπίπεδο παρατηρήσεις επηρεάζουν τη διαμόρφωσή του.

Η υλοποίηση του ταξινομητή SVM στην παρούσα εργασία έγινε μέσω του module του SVC του Scikit learn (βλ. 2.3).



Σχήμα 9: Μέσω της χρήσης kernel το γραμμικό υπερεπίπεδο του προηγούμενου σχήματος 7 μετασχηματίστηκε σε μη γραμμική δισδιάστατη καμπύλη

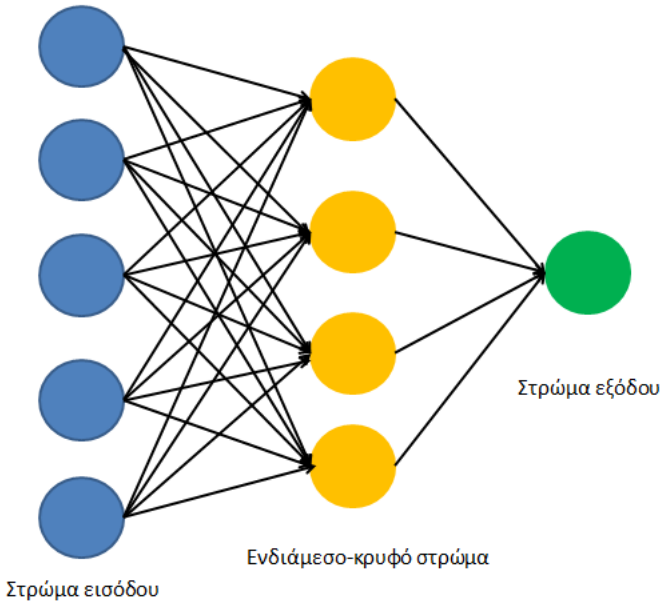
1.4.3.5 Νευρωνικά δίκτυα (Neural Networks) Η μέθοδος των Τεχνητών Νευρωνικών δικτύων αποτελεί μία από τις πιό δημοφιλείς και αποτελεσματικές μεθόδους της μηχανικής μάθησης[30]. Τα τεχνητά νευρωνικά δίκτυα μιμούνται τα βιολογικά νευρωνικά δίκτυα προσομοιάζοντας τη λειτουργίες των βιολογικών νευρικών κυττάρων. Ένα τεχνητό νευρωνικό δίκτυο απαρτίζεται από ένα διαστρωματωμένο σύνολο από απλό στοιχείο (νευρώνας) το οποίο είναι ικανό μόνο να αθροίζει την είσοδό του (input) και να μετασχηματίζει αυτή σε έξοδο μέσω μιας συνάρτησης απόφασης. Για ταξινόμηση συνήθως χρησιμοποιούνται νευρωνικά δίκτυα εμπρόσθιας τροφοδότησης πολλών επιπέδων όπου ένας νευρώνας επικοινωνεί μέσω κατευθυνόμενων ακμών με τους νευρώνες του επόμενου επιπέδου.



Σχήμα 10: Νευρώνας, η δομική μονάδα του νευρωνικού δικτύου. Ο νευρώνας σταθμίζει μέσω των βαρών τις εισερχόμενες σε αυτόν τιμές, τις αθροίζει και τέλος, προωθεί το αποτέλεσμα στους νευρώνες του επόμενου επιπέδου

Οι νευρώνες οργανώνονται σε επίπεδα: ένα επίπεδο εισόδου, το οποίο αντιστοιχεί στα χαρακτηριστικά(features) των στιγμιοτύπων του συνόλου δεδομένων, ένα ή περισσότερα κρυμμένα επίπεδα (hidden layers) και ένα επίπεδο εξόδου, το οποίο αντιστοιχεί στις κατηγορίες ταξινόμησης. Κάθε νευρώνας αθροίζει τις εισερχόμενες σε αυτόν τιμές(από το προηγούμενο επίπεδο) και μέσω μιας συνάρτησης απόφασης(η συνάρτηση αυτή μπορεί να πάρει πολλές μορφές ανάλογα με το πρόβλημα ταξινόμησης) παράγει μια τιμή output την οποία με τη σειρά του θα μεταβιβάσει στους νευρώνες του επόμενου επιπέδου. Στόχος της εκπαίδευσης του νευρωνικού

δικτύου είναι ο υπολογισμός των βαρών (της ισχύος της σύνδεσης δύο νευρώνων ή αλλιώς το πόσο η πληροφορία που μεταβιβάζεται από κάποιον νευρώνα στον επόμενο, μπορεί να τον επηρεάσει) των συνδέσεων μεταξύ των νευρώνων με στόχο να μειώσει το ποσοστό σφάλματος(την 'απόσταση' που είχαν οι προβλέψεις από την 'πραγματικότητα') ταξινόμησης.Τα βάρη που προκύπτουν από τη διαδικασία ενός κύκλου εκπαίδευσης χρησιμοποιούνται για να υπολογίσουμε οποία τα σταθμισμένα αθροίσματα, δηλαδή την είσοδο σε κάθε νευρώνα.



Σχήμα 11: Αναπαράσταση ενός νευρωνικού δικτύου. Το πρώτο στρώμα νευρώνων αποτελείται από τόσους νευρώνες όσα και τα χαρακτηριστικά των υπό ταξινόμηση στιγμιοτύπων. Ο αριθμός των νευρώνων στα κρυμμένα επίπεδα ποικίλει. Σε δοκιμασίες δυαδικής ταξινόμησης, όπως αυτές της παρούσας εργασίας ο νευρώνας εξόδου είναι ένας και δίνει σαν έξοδο την κλάση στην οποία κατηγοριοποίησε το στιγμιότυπο που του δόθηκε.

Για την ταξινόμηση ενός νέου στιγμιοτύπου οι τιμές των χαρακτηριστικών εφαρμόζονται στους νευρώνες εισόδου του ANN. Αυτές οι τιμές σταθμίζονται σύμφωνα με συνδέσεις μεταξύ των νευρώνων και τα σταθμισμένα αθροίσματά τους υπολογίζονται σε κάθε νευρώνα του επόμενου επιπέδου νευρώνων. Τα κανονικοποιημένα αποτελέσματα στους νευρώνες εξόδου καθορίζουν το αποτέλεσμα της ταξινόμησης (κατηγορία). Η κανονικοποίηση μπορεί να γίνει με διάφορους τρόπους. Ένας από τους βασικότερους είναι η κανονικοποίηση με συνάρτηση λογιστικής παλινδρόμησης, η οποία έχει τη μορφή:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

, όπου x η τιμή που προέκυψε μετά από το σταθμισμένο άθροισμα εισόδου του νευρώνα εξόδου. Η τιμή εξόδου συγκρίνεται με κάποιο δεδομένο κατώφλι (συνήθως το 0.5) και ανάλογα με το αν είναι μεγαλύτερη ή μικρότερη το στιγμιότυπο κατηγοριοποιείται στη μία ή την άλλη κατηγορία της διαδικής ταξινόμησης.

Στην παρούσα εργασία τα χαρακτηριστικά των στιγμιοτύπων ήταν οι τιμές TF-IDF των ζευγών όρων-ερωτήσεων. Ο αριθμός δηλαδή των νευρώνων εισόδου των νευρωνικών δικτύων της παρούσας μελέτης ήταν ίσως με τον αριθμό των ζευγών όρων-ερωτήσεων.

Η υλοποίηση του ταξινομητή NN στην παρούσα εργασία έγινε μέσω του module του MLPClassifier του Scikit learn (βλ. 2.3).

1.5 Επεξεργασία Φυσικής Γλώσσας(Natural Language Processing-NLP)

Η Επεξεργασία Φυσικής Γλώσσας(Natural Language Processing-NLP) αποτελεί έναν υποκλάδο της Γλωσσολογίας, της Επιστήμης των Υπολογιστών και της Τεχνητής Νοημοσύνης αντικείμενο του οποίου αποτελεί η αλληλεπίδραση μεταξύ των υπολογιστών και των ανθρωπίνων (φυσικών) γλωσσών και συγκεκριμένα, ο προγραμματισμός των υπολογιστών για να επεξεργάζονται και να αναλύουν μεγάλες ποσότητες δεδομένων φυσικής γλώσσας[31]. Στην παρούσα ενότητα γίνεται αναφορά σε δύο έννοιες που αφορούν την παρούσα εργασία, την κατηγοριοποίηση κειμένων καθώς και της μεθόδου συχνότητας όρων-αντίστροφης συχνότητας κειμένων (Term Frequency-Inverse Document Frequency, TF-IDF).

1.5.1 Σχέσεις-κατηγοριοποίηση κειμένων

Η κατηγοριοποίηση κειμένου, οργανώνει την πληροφορία, συσχετίζοντας κάθε κείμενο με τις σχετικότερες έννοιες, από ένα σύνολο εννοιών [32]. Η κατηγοριοποίηση απαιτεί ένα σύνολο εννοιών που έχουν οριστεί εκ των προτέρων (κατηγορίες) καθώς και πληροφορίες για το τι είδος κειμένων ανήκουν σε κάθε

κατηγορία[33]. Η κατηγοριοποίηση γενικά λαμβάνει χώρα σε δύο φάσεις: τη φάση εκπαίδευσης, κατά την οποία ο ταξινομητής μαθαίνει ποια χαρακτηριστικά αντιπροσωπεύουν βέλτιστα κάθε κατηγορία και τη φάση της κατηγοριοποίησης κατά την οποία, νέα αταξινόμητα κείμενα ταξινομούνται στη βέλτιστη κατηγορία. Οι διάφορες μεθοδολογίες κατηγοριοποίησης διαφέρουν ως προς το πως οι αναπαριστώνται οι έννοιες και τα κείμενα, το πως εξάγονται τα χαρακτηριστικά (και πως απονέμεται σε αυτά σχετική βαρύτητα - weighting) και στο πως υπολογίζεται η ομοιότητα των εννοιών αλλά και των κειμένων. Μια απλή και αποτελεσματική μέθοδος είναι η μέθοδος Rocchio κατά την οποία τα κείμενα εκπαίδευσης χρησιμοποιούνται για τη δημιουργία ενός αντιπροσωπευτικού διανύσματος για κάθε κατηγορία [34]. Κατά τη διάρκεια της ταξινόμηση, υπολογίζεται το διάνυσμα της ομοιότητας μεταξύ του κειμένου που πρέπει να κατηγοριοποιηθεί και των διανυσμάτων των εννοιών, και το αρχείο κατηγοριοποιείται στην πιό σχετική έννοια. Στη μέθοδο του k-πλησιέστερου γείτονα, ένα διάνυσμα δημιουργείται για κάθε κείμενο εκπαίδευσης. Κατά την ταξινόμηση, συγκρίνεται το διάνυσμα του κειμένου που πρέπει να ταξινομηθεί, με αυτά των κειμένων εκπαίδευσης. Τα k πιό όμοια κείμενα, συνεισφέρουν από μία ψήφο για την αντίστοιχη έννοια και το κείμενο ταξινομείται στην κατηγορία που συγκεντρώνει τις περισσότερες ψήφους. Άλλοι ερευνητές, χρησιμοποιήσαν τα Support Vector Machines(SVM) στην ταξινόμηση κειμένου

[35]. Αυτοί οι ταξινομητές, λειτουργούν όμοια με την προηγούμενη μέθοδο αρχικά, δηλαδή αντιστοιχούν διανύσματα στα κείμενα εκπαίδευσης και επιλέγουν τα k πιό αντιπροσωπευτικά. Στη συνέχεια προβάλλουν τα διανύσματα σε περισσότερες διαστάσεις, με αποτέλεσμα τα νέα γνωρίσματα (features) που προκύπτουν εκεί, χρησιμοποιούνται για να ταξινομηθούν τα κείμενα με γραμμικό τρόπο. Αν και τα SVM φαίνεται ότι έχουν την καλύτερη απόδοση στην ταξινόμηση, το υβριδικό μοντέλο k-πλησιέστερου γείτονα/μεθόδου Rocchio βρέθηκε να υπερέχει ελαφρώς της απλής μεθόδου k-πλησιέστερου γείτονα[32].Άλλοι ερευνητές έχουν χρησιμοποιήσει και άλλες μεθόδους για ταξινόμηση κειμένου, όπως τα νευρωνικά δίκτυα, την αφελή μέθοδο Bayes, τη μέθοδο ελαχίστων τετραγώνων, αλλά και συνδυασμό μεθόδων. Αν και τα SVM φαίνεται ότι έχουν την καλύτερη απόδοση στην ταξινόμηση, το υβριδικό μοντέλο k-πλησιέστερου γείτονα/μεθόδου Rocchio βρέθηκε να υπερέχει ελαφρώς της απλής μεθόδου k-πλησιέστερου γείτονα[32].

1.5.2 Συχνότητα όρων-αντίστροφη συχνότητα κειμένων (Term Frequency-Inverse Document Frequency, TF-IDF)

Ένας καθολικά αποδεκτός τρόπος για να βρει κανείς πόσο σημαντική είναι μια λέξη σε μια ομάδα/κλάση κειμένων είναι η Συχνότητα Όρων-Αντίστροφη συχνότητα Κειμένων (Term Frequency-Inverse Document Frequency, TF-IDF), μια μέθοδος κατασκευασμένη από την Spärck Jones το 1972. Η μέθοδος αυτή κρατάει ακόμη

και σήμερα κεντρική θέση στον τρόπο λειτουργίας των περισσότερων μηχανών αναζήτησης[36].

Για να γίνει κατανοητή η έννοια του TF-IDF θα οριστούν καταρχήν οι έννοιες του TF και του IDF.

-TF: Ας υποθέσουμε πως μεταξύ ενός συνόλου κειμένων επιθυμεί κανείς να βρει τα πιό σχετικά με το ερώτημα "καρκίνος μαστού". Ένας απλός τρόπος να ξεκινήσει, θα ήταν να αφαιρέσει τελείως κείμενα που δεν περιέχουν καμία από τις λέξεις "καρκίνος" και "μαστός". Τα εναπομείναντα κείμενα θα μπορούσαν να ιεραρχηθούν ανάλογα με το πόσες φορές περιέχουν καθεμιά από τις λέξεις. Παρόλα αυτά το μέγεθος των κειμένων ποικίλλει και έτσι π.χ. ένα μεγάλο κείμενο μπορεί να περιέχει μια λέξη περισσότερες φορές από ένα μικρότερο κείμενο χωρίς να λαμβάνεται υπόψιν η 'πυκνότητα' της λέξης. Για αυτό το λόγο αρκετές φορές η συχνότητα προσαρμόζεται στο μέγεθος του κειμένου με διάφορες μεθόδους (λογαρίθμηση, adjusted frequency κ.α.). Ο πιο συχνά χρησιμοποιούμενος τρόπος υπολογισμού της TF είναι ο λόγος της συχνότητας μιας λέξης μέσα στο κείμενο διά τον αριθμό των λέξεων του κειμένου.

-IDF: Ένας όρος μπορεί να είναι συχνός σε ένα κείμενο, αυτό όμως δε συνεπάγεται απαραίτητα ότι είναι ειδικός και μας προσφέρει πληροφορίες για αυτό. Μπορεί για παράδειγμα ένας όρος να είναι γενικά συχνός και να παρουσιάζει αυξημένη

συχνότητα σε ένα κείμενο για λόγους μη σχετιζόμενους με την κατηγορία στην οποία ανήκει το κείμενο. Αυτό το πρόβλημα μπορεί να αντιμετωπιστεί με την έννοια του της αντίστροφης συχνότητας του όρου στα κομμάτια κειμένου, η οποία αντιστοιχεί στον λογάριθμο του αντίστροφου κλάσματος του αριθμού των κειμένων που περιέχουν τον εν λόγω όρο διά το σύνολο των διαθέσιμων κειμένων.

-TF-IDF: Έχοντας ορίσει τους όρους TF και IDF μπορεί να οριστεί τώρα ο όρος TF-IDF ως το γινόμενό τους. Η TF-IDF μπορεί να λάβει τιμές μεγαλύτερες ή ίσες του μηδενός. Η τιμή αυτή υπολογίζεται για κάθε ζεύγος όρου-κειμένου και όσο πιο μεγάλη είναι τόσο περισσότερη 'πληροφορία' θεωρείται ότι φέρει ο συγκεκριμένος όρος για το συγκεκριμένο κείμενο.

Η υλοποίηση του αλγορίθμου TF-IDF στην παρούσα εργασία έγινε μέσω του module του TfidfVectorizer του Scikit learn (βλ. 2.3).

1.6 Στόχος της παρούσας εργασίας

Εως αυτό το σημείο έγινε μια εισαγωγική αναφορά στις έννοιες που θα χρησιμοποιηθούν στην παρούσα εργασία δηλαδή: στην Ιατρική Βασισμένη σε Στοιχεία και τη θέση της Ογκολογίας σε αυτή, στο BioASQ challenge και τις βιοϊατρικές ερωτήσεις καθώς και στη Μηχανική Μάθηση και τις μεθόδους της. Φάνηκε επίσης η σημασία της δημιουργίας συστημάτων αυτόματης απάντησης σε κλινικές ερωτή-

σεις με βάση συγκεκριμένους κανόνες ιεράρχησης της πληροφορίας.

Κύριο έργο της παρούσας εργασίας είναι η επεξεργασία και ανάλυση των κλινικών ερωτήσεων του BioASQ challenge με μεθόδους της Μηχανικής Μάθησης και της Επεξεργασίας Φυσικής Γλώσσας με απώτερο σκοπό τη εύρεση των βέλτιστων απαντήσεων σε αυτές σύμφωνα με τους κανόνες της Ιατρική Βασισμένη σε Στοιχεία. Οι ερωτήσεις καλύπτουν το ευρύτερο φάσμα της Ιατρικής, με έμφαση στον τομέα της Ογκολογίας.

Η επεξεργασία αυτή θα επιχειρηθεί σε **τρία** διαφορετικά στάδια, καθένα από τα οποία θα εστιάζει σε διαφορετικές ιδιότητες των ερωτήσεων. Οι στόχοι παρούσας εργασίας περιλαμβάνουν τη δημιουργία του συνόλου δεδομένων καθώς και αυτά τα στάδια επεξεργασίας και συνοψίζονται σε:

- Δημιουργία ενός συνόλου δεδομένων η βάση του οποίου θα αποτελείται από κλινικές ερωτήσεις του BioASQ challenge οι οποίες στη συνέχεια θα χαρακτηριστούν από ειδικούς, σύμφωνα με κριτήρια της Ιατρικής Βασισμένης σε Στοιχεία.

- Το **πρώτο** στάδιο θα επιχειρήσει την αυτόματη ταξινόμηση των ερωτήσεων σε ερωτήσεις τύπου foreground και background δηλαδή σε ερωτήσεις των οποίων η απάντηση είναι γενικής φύσης, ξεκάθαρη και παγιωμένη και σε

ερωτήσεις των οποίων η απάντηση αφορά συγκεκριμένες κλινικές περιπτώσεις και/ή αφορά γνώση που δεν έχει ακόμη ξεκάθαρη απάντηση και αποτελεί αντικείμενο αντιπαράθεσης.

- Το **δεύτερο** στάδιο θα περιλαμβάνει την αυτόματη ανίχνευση της ύπαρξης των στοιχείων PICO (patient, intervention, comparison, outcome) στις ερωτήσεις foreground.

- Το **τρίτο** και τελευταίο στάδιο θα επιχειρήσει την αυτόματη κατηγοριοποίηση των ερωτήσεων σε ερωτήσεις που αφορούν θεραπεία, διάγνωση, πρόγνωση και επιβάρυνση -συσχέτιση.

Μέσω των τριών αυτών κατηγοριοποιήσεων, φιλοδοξία της παρούσας εργασίας είναι να θέσει τα θεμέλια για τη δημιουργία ενός συστήματος αυτοματοποιημένης απάντησης βιοϊατρικών ερωτήσεων που να παρέχει απαντήσεις σεβόμενο την ιεράρχηση των στοιχείων της σύγχρονης Ιατρική Βασισμένη σε Στοιχεία.

2 Μέθοδοι

2.1 Επισκόπηση ενότητας (Overview)

Η ενότητα αυτή θα παρουσιάσει κατά σειρά, τις μεθόδους που χρησιμοποιήθηκαν κατά την εκπόνηση της παρούσας εργασίας. Αρχικά θα αναλυθεί ο τρόπος που δημιουργήθηκαν τα σύνολα δεδομένων για τις διαφορετικές ταξινομήσεις που επιχειρήθηκαν, δηλαδή το πως έγινε η επιλογή και η κατονομασία (labeling) των ερωτήσεων. Στη συνέχεια θα περιγραφεί ο τρόπος με τον οποίο χωρίστηκαν τα δεδομένα σε σύνολα κατάλληλα για την εκπαίδευση των ταξινομητών. Θα περιγραφεί η προεπεξεργασία των δεδομένων, δηλαδή ο τρόπος με τον οποίο οι ερωτήσεις μετασχηματίστηκαν από σύνολα λέξεων, σε αριθμούς, έτσι ώστε να δοθούν ως είσοδος στους ταξινομητές. Τέλος, θα αναλυθεί το πως αξιολογήθηκε η απόδοση των ταξινομητών, πως οι αποδόσεις συγκρίθηκαν μεταξύ τους αλλά και πως συγκρίθηκαν με την απόδοση ανθρώπων-κριτών που εκτέλεσαν τις αντίστοιχες ταξινομήσεις και την ιδανική ταξινόμηση.

2.2 Δημιουργία του συνόλου δεδομένων(dataset creation)

2.2.1 Κατηγοριοποίηση από κριτές

Στην παρούσα εργασία πραγματοποιήθηκαν διάφορες ταξινομήσεις ερωτήσεων από μοντέλα της Μηχανικής Μάθησης. Οι ταξινομήσεις ήταν επιβλεπόμενες (supervised) δηλαδή στα δεδομένα εισαγωγής σε κάθε ταξινομητή, δίδονταν το κείμενο

της ερώτησης καθώς και η κατηγορία στην οποία αυτή άνηκε. Η απονομή της κατηγορίας σε κάθε ερώτηση έγινε από δύο κριτές-ανθρώπους, για κάθε μια κατηγοριοποίση ξεχωριστά[37]. Τα **κριτήρια** που έπρεπε να πληρούν οι κριτές ήταν δύο:

- Έπρεπε να είναι ειδικοί στο υπό μελέτη αντικείμενο. Στην παρούσα μελέτη η οποία ανέλυσε ερωτήσεις ιατρικού-κλινικού περιεχομένου, αυτό θεωρήθηκε ισοδύναμο με την κατοχή πτυχίου Ιατρικής Σχολής.
- Έπρεπε να γνωρίζουν ρητώς τους κανόνες που πρέπει να πληροί μία ερώτηση για να ανήκει σε καθεμιά κατηγορία. Αυτό επιτεύχθηκε μετά από ανάγνωση και από τους δύο κριτές, συγκεκριμένης βιβλιογραφίας αναφοράς[1, 38, 8].

Ο κριτής Α ήταν ο γράφων και η κατηγοριοποιήσεις του θεωρήθηκαν ως κατηγοριοποιήσεις αναφοράς (baseline) για τις αξιολογήσεις των μοντέλων. Οι δοκιμασίες ταξινόμησης πραγματοποιήθηκαν ξεχωριστά για τα δεδομένα ταξινόμησης και των δύο κριτών. Επίσης όπως αναφέρεται στην πορεία, ο βαθμός συμφωνίας των δύο κριτών ποσοτικοποιήθηκε με τον δείκτη k(βλ. 2.9).

2.2.2 Διαδικασία δημιουργίας του συνόλου δεδομένων

Τα πρωτογενή δεδομένα που χρησιμοποιήθηκαν στο σύνολο των αναλύσεων της παρούσας μελέτης είχαν τη μορφή ερωτήσεων. Οι ερωτήσεις αντλήθηκαν από

το σύνολο των ερωτήσεων που είχαν δημοσιευθεί στο BioAsq Challenge. Πιο συγκεκριμένα, δημιουργήθηκε ένα αρχικό αρχείο που περιείχε το σύνολο των ερωτήσεων που δημοσιεύθηκαν ως σύνολα εκπαίδευσης σε όλες της φάσης του Task B του BioAsq Challenge (η δομή του BioAsq challenge περιγράφεται στην ενότητα 1.2.2.1).

Στόχος της δημιουργίας του αρχικού συνόλου δεδομένων ήταν η συγκέντρωση ικανού αριθμού ερωτήσεων που να πληροί το κριτήριο της κλινικής ερώτησης(βλ. 1.2.1.2). Κατά σύμβαση ο αρχικός αριθμός των κλινικών ερωτήσεων ορίστηκε το 320. Για να γίνει η διαδικασία δημιουργίας του συνόλου δεδομένων πληρέστερα κατανοητή, κάτω από την περιγραφή κάθε βήματος θα δίνεται ένα παράδειγμα με κάποιες ερωτήσεις από το σύνολο δεδομένων. Το αρχικό σύνολο δεδομένων περιείχε ερωτήσεις από όλο το φάσμα των κλινικών ερωτήσεων του BioASQ, ωστόσο, η αντιπροσώπευση των ερωτήσεων που είχαν αντικείμενο την **Ογκολογία**Ογκολογία ήταν μεγαλύτερη σε σχέση με αυτή άλλων ιατρικών ειδικοτήτων(30%)
Τα βήματα της δημιουργίας του συνόλου δεδομένων που ακολουθήθηκαν συνοψίζονται ως εξής:

- Δημιουργία προγράμματος για την προσπέλαση κατά τυχαίο τρόπο του συνόλου των έως τον Ιούλιο του 2019 δημοσιευθέντων στο BioASQ ερωτήσεων μία προς μία.

παράδειγμα:

Ερ. 1 "What are the symptoms of Meigs'[11] Syndrome?"

Ερ. 2 "Is the monoclonal antibody Trastuzumab[12] (Herceptin) of potential use in the treatment of prostate cancer?"

Ερ. 3 "Is it possible to purify pseudopodia[13] to be used for proteomic analysis?"

- Κατηγοριοποίηση των ερωτήσεων σε δύο κατηγορίες, τις κλινικές και τις μη κλινικές εως τη συγκέντρωση 320 κλινικών ερωτήσεων (αρχικό σύνολο δεδομένων).

  παράδειγμα:

  Ερ. 1 "What are the symptoms of Meigs' Syndrome?" η ερώτηση κατηγοριοποιήθηκε ως κλινική ερώτηση και εντάχθηκε στο αρχικό σύνολο δεδομένων

  Ερ. 2 "Is the monoclonal antibody Trastuzumab (Herceptin) of potential use in the treatment of prostate cancer?" η ερώτηση κατηγοριοποιήθηκε ως κλινική ερώτηση και εντάχθηκε στο αρχικό σύνολο δεδομένων

  Ερ. 3 "Is it possible to purify pseudopodia to be used for proteomic analysis?" η ερώτηση κατηγοριοποιήθηκε ως μη κλινική ερώτηση και αποκλείστηκε από το αρχικό σύνολο δεδομένων

[11] σύνδρομο που αναπτύσσεται στα πλαίσια νεοπλασίας των ωοθηκών
[12] μονοκλωνικό αντίσωμα που έφερε την επανάσταση στη θεραπεία του καρκίνου του μαστού
[13] παροδική προσεκβολή κυτταρικών σχηματισμών

- Κατηγοριοποίηση των ερωτήσεων σε ερωτήσεις foreground και background από δύο ανεξάρτητους κριτές (κριτής Α και κριτής Β), με σκοπό την εκπαίδευση μοντέλων ταξινόμησης στη διάκριση ερωτήσεων στις δύο αυτές κατηγορίες.

  παράδειγμα:

  Ερ. 1 "What are the symptoms of Meigs' Syndrome?" η ερώτηση κατηγοριοποιήθηκε ως ερώτηση background και από τους δύο κριτές

  Ερ. 2 "Is the monoclonal antibody Trastuzumab (Herceptin) of potential use in the treatment of prostate cancer?" η ερώτηση κατηγοριοποιήθηκε ως ερώτηση foreground και από τους δύο κριτές

- Δημιουργία υποσυνόλου δεδομένων που αποτελείται μόνο από τις ερωτήσεις που κατηγοριοποιήθηκαν ως foreground

  παράδειγμα:

  Μόνο η Ερ. 2 του βήματος 3 θα χρησιμοποιηθεί στις περαιτέρω αναλύσεις.

  Ερ. 2 "Is the monoclonal antibody Trastuzumab (Herceptin) of potential use in the treatment of prostate cancer?" (μόνο οι ερωτήσεις foreground θα χρησιμοποιηθούν στη συνέχεια της ανάλυσης)

- Κατηγοριοποίηση των foreground ερωτήσεων από δύο ανεξάρτητους κριτές

(κριτής Α και κριτής Β) με δυαδικό τρόπο (δηλαδή κάθε κριτής ανέθετε την τιμή '1' εάν μια ερώτηση περιείχε στοιχείο P, το ίδιο και για τις κατηγορίες I, C, O και αντίστοιχα για τις κατηγορίες T, D, P, H ενώ ανέθετε την τιμή '0' για μια ερώτηση εάν αυτή δεν περιείχε στοιχείο της κατηγορίας P και αντίστοιχα για τις άλλες κατηγορίες) στις κατηγορίες P, I, C, O και T, D, P, H όπως αυτές περιγράφονται στην Εισαγωγή.

παράδειγμα:

Ερ. 2 "Is the monoclonal antibody Trastuzumab (Herceptin) of potential use in the treatment of prostate cancer?" η ερώτηση κατηγοριοποιήθηκε από τον κριτή Α (αντίστοιχα και από τον κριτή Β) με τα εξής δυαδικά στοιχεία:

P: "1" (εμμέσως αναφέρεται στην ερώτηση ότι γίνεται αναφορά σε ασθενείς με καρκίνο του προστάτη)

I: "1" (γίνεται σαφής αναφορά στην παρέμβαση δηλαδή στο μονοκλωνικό αντίσωμα trastuzumab)

C: "0" (δε γίνεται αναφορά σε ομάδα σύγκρισης δηλαδή έναντι ποιάς παρέμβασης θα αντιπαραβληθεί η αποτελεσματικότητα της παρέμβασης)

O: "0" (δε γίνεται σαφής αναφορά για το επί ποιάς βάσης θα γίνει η σύγκριση αποτελεσματικότητας)

T: "1" (η ερώτηση αφορά θεραπευτική παρέμβαση)

D: "0" (η ερώτηση δεν αφορά διαγνωστικό εργαλείο)

P: "0" (η ερώτηση δεν αφορά προγνωστικό παράγοντα)

H: "0" (η ερώτηση δεν αφορά συσχέτιση, επιβαρυντικό ή προστατευτικό παράγοντα με/για κάποια πάθηση)

Εν κατακλείδι,τα σύνολα δεδομένων που δημιουργήθηκαν αποτελούνται από **ζεύγη ερωτήσεων με δυαδικές τιμές** ("0" ή "1") δηλαδή από σύνολα ερωτήσεων με την αντίστοιχη κατηγορία στην οποία ταξινομήθηκαν σε καθεμιά από τις προαναφερθείσες ταξινομήσεις.

2.3 Υλοποίηση των ταξινομήσεων, το πακέτο Scikit-learn

Για την υλοποίηση σε μορφή κώδικα των ταξινομήσεων της παρούσας εργασίας χρησιμοποιήθηκε το πακέτο Scikit-learn[39]. Το Scikit-learn αποτελείται από ένα τεράστιο σύνολο σύγχρονων αλγορίθμων Μηχανικής Μάθησης, γραμμένο στη γλώσσα προγραμματισμού Python και δημιουργημένο έτσι ώστε να μπορεί να χρησιμοποιηθεί από μη ειδικούς της Επιστήμης των Υπολογιστών (high level coding- κώδικας περισσότερο προσβάσιμος στην κατανόηση). Στόχος του Scikit-learn, όπως και το ονομά του περιγράφει, είναι η δημιουργία προσβάσιμου κώδικα

Μηχανικής Μάθησης σε επιστήμονες όλων των κλάδων. Σε αρκετά σημεία της παρούσας εργασίας γίνεται μνεία των επιμέρους στοιχείων του Scikit-learn που χρησιμοποιήθηκαν στις διαφορετικές διεργασίες.

2.4 Διαχωρισμός σε σύνολο εκπαίδευσης και σύνολο αξιολόγησης(training set, test set), η μέθοδος k-Fold Cross-Validation

Όπως ειπώθηκε στην 1.4.2, απαραίτητο στάδιο κάθε ταξινόμησης είναι ο διαχωρισμός του συνόλου δεδομένων σε σύνολο εκπαίδευσης (training set) και σύνολο αξιολόγησης(test set).

Το Cross-Validation είναι μια διαδικασία δειγματοληψίας που χρησιμοποιείται στην αξιολόγηση μοντέλων μηχανικής μάθησης σε ένα περιορισμένο δείγμα δεδομένων [40, 41]. Η διαδικασία χαρακτηρίζεται από μια παράμετρο κ η οποία αντιστοιχεί στον αριθμό των υποσυνόλων στις οποίες θα χωριστεί το σύνολο των δεδομένων. Αν για παράδειγμα το δείγμα χωριστεί σε 10 υποομάδες τότε μπορούμε να αναφερθούμε στη διαδικασία σαν 10-Fold Cross-Validation. Η μέθοδος χρησιμοποιείται ευρέως γιατί είναι λιγότερο μεροληπτική από άλλες μεθόδους διαχωρισμού των δεδομένων σε ομάδες εκπαίδευσης-αξιολόγησης (σε σχέση π.χ. μέσω της απλής μεθόδου train-test split[14]). Η διαδικασία γίνεται ως εξής:

[14] απλός διαχωρισμός του συνόλου σε δύο υποσύνολα, το σύνολο εκπαίδευσης και το σύνολο αξιολόγησης

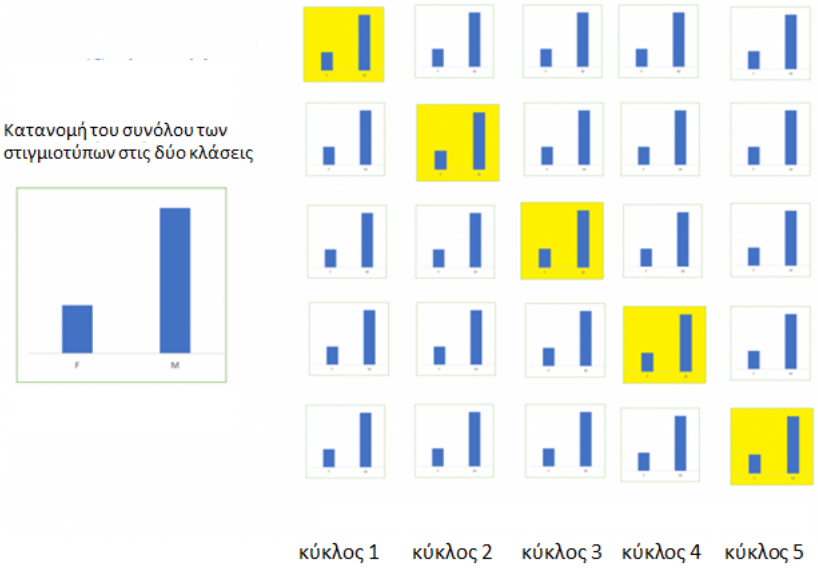
- Το σύνολο δεδομένων ανακατεύεται (shuffling)
- Το σύνολο δεδομένων διαχωρίζεται σε k ομάδες
- Κάθε ομάδα επιλέγεται ως ομάδα αξιολόγησης, ενώ χρησιμοποιούμε τα δεδομένα όλων των υπολοίπων ομάδων ως δεδομένα εκπαίδευσης. Στη συνέχεια το μοντέλο εκπαιδεύεται και εν τέλει αξιολογείται. Η διαδικασία επαναλαμβάνεται k φορές έτσι ώστε όλες οι ομάδες να χρησιμοποιηθούν ως ομάδες αξιολόγησης.
- Η απόδοση του μοντέλου καθορίζεται από το σύνολο των αποδόσεών του σε κάθε κύκλο εκπαίδευσης-αξιολόγησης (πχ με παράθεση του μέσου όρου και της τυπικής απόκλισης των αποδόσεων).

Παρατίθεται μια ιστοσελίδα με ένα ολοκληρωμένο παράδειγμα επάνω στο k-Fold Cross-Validation[40].

2.5 Διαστρωμάτωση (Stratification) -stratified k-Fold cross validation

Πολλές φορές, το σύνολο δεδομένων το οποίο χρησιμοποιείται στην εκπαίδευση και στη συνέχεια αξιολόγηση ενός μοντέλου δεν είναι ισορροπημένο, δηλαδή τα στιγμιότυπα μιας κατηγορίας μπορεί να υποεκπροσωπούνται ή να υπερεκπροσωπούνται όταν χωρίσουμε το σύνολο δεδομένων σε ομάδα εκαίδευσης και ομάδα

αξιολόγησης. Αυτό μπορεί επηρεάσει αρνητικά την ανάλυσή μας. Η διαδικασία της διαστρωμάτωσης εξασφαλίζει, ότι η αναλογία με την οποία εμφανίζονται τα δεδομένα πριν γίνει ο διαχωρισμός σε ομάδα εκπαίδευσης και αξιολόγησης, θα συνεχίσει να υπάρχει (όσο αυτό είναι εφικτό) και στις ομάδες εκπαίδευσης και αξιολόγησης ξεχωριστά [42].



Σχήμα 12: Αναπαράσταση της διαδικασίας του stratified 5-fold cross-validation. Το σύνολο δεδομένων χωρίζεται σε πέντε υποσύνολα (ίσα με τον αριθμό k) και η εκπαίδευση και αξιολόγηση πραγματοποιούνται αντίστοιχες φορές. Κάθε φορά ένα από τα πέντε υποσύνολα(κίτρινο) χρησιμοποιείται ως σύνολο αξιολόγησης ενώ τα υπόλοιπα μαζί ως σύνολο εκπαίδευσης. Σημαντικό είναι πως διατηρείται η αναλογία του αρχικού συνόλου σχετικά με την κατανομή σε κλάσεις στα προκύπτοντα υποσύνολα

Συνδυάζοντας τη μέθοδο k-Fold Cross-Validation, με τη μέθοδο του stratification, έχουμε έναν αξιόπιστο τρόπο διαχωρισμού των δεδομένων (όπως φαίνεται

σχηματικά στο 12). Για την υλοποίηση του stratified k-fold cross validation της παρούσας εργασίας χρησιμοποιήθηκε το module StratifiedKFold του Scikit learn.

2.6 Προεπεξεργασία δεδομένων(Data Preprocessing)

Οι ταξινομήσεις της παρούσας εργασίας αφορούν καταχώρηση βιοϊατρικών ερωτήσεων σε κατηγορίες. Οι οντότητες που πρέπει να ταξινομηθούν είναι κείμενα που αποτελούνται από γράμματα/λέξεις και σημεία στίξης. Η είσοδος(input) που μπορεί να καταχωρηθεί στον ταξινομητή ωστόσο, πρέπει να έχει τη μορφή του διανύσματος αριθμητικών τιμών(feature vector). Η διαδικασία μετατροπής του συνόλου των αρχικών προς ταξινόμηση δεδομένων (σύνολο των ερωτήσεων) σε διανύσματα χαρακτηριστικών αποτελεί τη διαδικασία της προεπεξεργασίας[43]. Η προεπεξεργασία στην παρούσα εργασία έγινε σε μια σειρά διαδοχικών σταδίων. Για την πιο κατανοητή παρουσίασή τους θα χρησιμοποιηθεί το παράδειγμα μιας ερώτησης("Is pembrolizumab effective in testicular cancer?"). Τα στάδια της προεπεξεργασίας είναι τα εξής:

- Αρχική μορφή της ερώτησης:

  "Is pembrolizumab[15] effective in testicular cancer?"

- Μετατροπή των κεφαλαίων γραμμάτων σε μικρά (έτσι ώστε π.χ. το Is και

[15] νέο αντικαρκινικό φάρμακο της κατηγορίας της ανοσοθεραπείας

το is να θεωρούνται η ίδια λέξη):

"is pembrolizumab effective in testicular cancer?"

- Αφαίρεση των σημείων στίξης:

  "is pembrolizumab effective in testicular cancer"

- Μετατροπή του κειμένου σε λίστα λέξεων:

  ("is", "pembrolizumab", "effective", "in", "testicular", "cancer")

- Stemming-διατήρηση του κορμού-ρίζας των λέξεων με περικοπή των καταλήξεων. Η διαδικασία του stemming είναι μια πολύπλοκη διαδικασία με λεπτομερείς γλωσσολογικούς κανόνες. Στην παρούσα εργασία για την υλοποίηση του stemming χρησιμοποιήθηκε ο αλγόριθμος Porter[16] [44]:

  ("is", "pembrolizumab", "effect", "in", "testic", "cancer") 1.5.2)[45].

Μετά την αλληλουχία αυτών των βημάτων προεπεξεργασίας τα δεδομένα έχουν πλέον την μορφή του διανύσματος χαρακτηριστικών και είναι έτοιμα για να δοθούν ως είσοδος στους διάφορους ταξινομητές.

[16] Αλγόριθμος Porter Stemmer. Περιγραφή της λειτουργίας του αλγορίθμου μπορείτε να βρείτε εδώ

2.7 Αξιολόγηση της απόδοσης των μοντέλων μηχανικής μάθησης

Αφότου κατασκευαστεί ένα μοντέλο μηχανικής μάθησης που να πραγματοποιεί ταξινόμηση, σημαντικό είναι να μπορεί να ποσοτικοποιηθεί το πόσο καλά έκανε την ταξινόμηση και εν συνεχεία να το συγκριθεί με άλλα μοντέλα που έκαναν την ίδια δοκιμασία[46]. Η αξιολόγηση βασίζεται στον αριθμό των σωστών και των λάθος προγνώσεων που έγιναν από το μοντέλο στο σύνολο ελέγχου. Έστω πως η ταξινόμησή απαιτούσε τον διαχωρισμό κομματιών κειμένου σε δύο κατηγορίες την κατηγορία '1' και την κατηγορία '0'. Σε μια τέτοια δυαδική ταξινόμηση (ταξινόμηση των δεδομένων σε δύο κατηγορίες)μπορούν να χρησιμοποιήσουμε τις έννοιες των:

- αληθώς θετικών, δηλαδή των τιμών ήταν όντως '1' και το μοντέλο προέβλεψε σωστά σαν '1'(True Positives-TP)
- αληθώς αρνητικών, δηλαδή αυτών των παρατηρήσεων που το μοντέλο τις ταξινόμησε σαν '0' και ήταν όντως '0'(True Negatives-TN)
- αυτών που ήταν '0' και το μοντέλο τις ταξινόμησε σαν '1'(False Positives-FP)
- αυτών που ήταν '1' και το μοντέλο τις ταξινόμησε σαν '0' (False Negatives-FN)

Ένα από τα βασικά μέτρα αξιολόγησης μοντέλων ταξινόμησης είναι η ακρίβεια (accuracy). Η ακρίβεια υπολογίζεται από τον τύπο και είναι ο λόγος του αριθμού των σωστών προγνώσεων προς τον άθροισμα όλων των προγνώσεων που πραγματοποιήθηκαν.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

Ένα άλλο μέγεθος για την εκτίμηση της απόδοσης ενός μοντέλου ταξινόμησης είναι η πιστότητα (precision), εκφράζει τη συνέπεια στις εκτιμήσεις του μοντέλου στο να κάνει σωστές προβλέψεις και δίνεται από τον παρακάτω τύπο.

$$PREC = \frac{TP}{TP + FP}$$

Επιπροσθέτως η ανάκληση(recall) είναι άλλο ένα σημαντικό μέτρο αξιολόγησης μοντέλων και εκφράζει το ποσοστό των πραγματικά θετικών τιμών που προβλέφθηκαν από το μοντέλο επιτυχώς σαν θετικές και δίνεται από τον παρακάτω τύπο.

$$REC = \frac{TP}{TP + FN}$$

Τα μέτρα της πιστότητας και της ανάκλησης αν και δίνουν σημαντικές πληροφορίες για την απόδοση του μοντέλου, δεν είναι ικανά για να το περιγράψουν σφαιρικά. Ένα μοντέλο λόγου χάρι, μπορεί να έχει υψηλή πιστότητα και χαμηλή ανάκληση ή το αντίθετο, χαμηλή πιστότητα και υψηλή ανάκληση. Την ανεπάρκεια αυτή

των μέτρων αυτών έρχεται να καλύψει ένα άλλο μέτρο, το μέτρο F1, το οποίο ενσωματώνει την πιστότητα και την ανάκληση στον τύπο του[46]. Το **μέτρο F1**, κοινώς αποδεκτό ως ένα από τα πιό αξιόπιστα για την αξιολόγηση μοντέλων ταξινόμησης, θα είναι και το μέτρο αξιολόγησης που θα **χρησιμοποιηθεί στην παρούσα εργασία** για την αξιολόγηση και σύγκριση των μοντέλων.

$$F1 = \frac{2 * PREC * REC}{PREC + REC}$$

Το μέτρο F1 είναι ο λεγόμενος αρμονικός μέσος της πιστότητας και της ανάκλησης και μπορεί να λάβει ένα φάσμα τιμών από 0 έως 1 και να θεωρείται καλύτερος, δηλαδή να σημαίνει μεγαλύτερη αξιοπιστία για το μοντέλο που περιγράφει όσο περισσότερο πλησιάζει τη μονάδα.

Τα υπολογιστικά εργαλεία για την υλοποίηση της αξιολόγησης των μοντέλων πάρθηκαν από το module f1score του Scikit learn.

2.8 Στατιστική ανάλυση

Το μέτρο F1 μπορεί από μόνο του να καταδείξει την αξία της απόδοσης ενός μοντέλου ταξινόμησης, γίνεται όμως ακόμη πιο χρήσιμο, όταν το αντιπαραβληθεί με τα μέτρα F1 της απόδοσης άλλων μοντέλων, βλέποντάς το δηλαδή ως συγκριτική συνιστώσα[46]. Μια δοκιμασία ταξινόμησης μπορεί λόγου χάρι να είναι εκ φύσεως εύκολη και για αυτό το μέτρο F1 που προκύπτει για ένα μοντέλο να είναι υψηλό.

Αν συγκριθεί όμως με την απόδοση ενός άλλου, απλοικού μοντέλου και αυτές δεν φανούν να διαφέρουν σημαντικά, τότε η αξία της αρχικής μας παρατήρησης χάνεται.

Η σύγκριση των F1 πρέπει να ποσοτικοποιηθεί και αυτό μπορεί να γίνει με αντικειμενικό τρόπο χρησιμοποιώντας κατάλληλα στατιστικά εργαλεία. Στην παρούσα εργασία, κάθε μοντέλο αξιολογήθηκε επάνω σε ένα test set k-φορές, διότι χρησιμοποιήθηκε η μέθοδος του k-fold cross validation. Ως εκ τούτου προέκυψαν k τιμές F1. Ο σκοπός της στατιστικής ανάλυσης ήταν να συγκριθεί η απόδοση όλων των μοντέλων (dummy predictor, naive bayes model, random forest, support vector machinem, neural network) ανά ζεύγη και να δειχθεί η στατιστική σημαντικότητα των συγκρίσεων. Η στατιστική ανάλυση που πραγματοποιήθηκε συνοψίζεται στα εξής στάδια:

- Περιγραφική στατιστική-εύρεση του μέσου όρου και της τυπικής απόκλισης των F1 κάθε μοντέλου
- Εκτίμηση κανονικότητας κατανομής-στην εκτίμηση αυτή χρησιμοποιήθηκε η δοκιμασία Shapiro-Wilk [47].
- Σύγκριση της απόδοσης των μοντέλων κατά ζεύγη-οι τιμές F1 κάθε μοντέλου συγκρίθηκαν με αυτές κάθε άλλου. Ανάλογα με το αν οι κατανομές

τιμών F1 των συγκρινόμενων μοντέλων ήταν κανονικές (αυτό προέκυψε από το στάδιο β) χρησιμοποιήθηκε και διαφορετική στατιστική δοκιμασία για την εν λόγω σύγκριση. Πιο συγκεκριμένα, εάν η κατανομή όλων των δεδομένων είναι κανονική, τότε χρησιμοποιείται η δοκιμασία independent t-test(η οποιά έχει ως προυπόθεση για την χρησιμοποίησή της την κανονικότητα των δεδομένων), ενώ εάν η κατανομή των δεδομένων δεν είναι κανονική, τότε χρησιμοποιείται η δοκιμασία Wilcoxon-Mann-Whitney[47]

Ως όριο στατιστικής σημαντικότητας τέθηκε κατά συνθήκη το a = 0.05. Για την διεκπεραίωση της στατιστικής ανάλυσης χρησιμοποιήθηκε το module Scipy.

2.9 Συμφωνία μεταξύ κριτών-σχολιαστών (Interannotator/Interjudge agreement)

Αρκετές φορές, όταν γίνεται προσπάθεια να οριστούν τα στοιχεία που πρέπει να πληροί ένα κείμενο για να ανήκει σε μια κατηγορία, αυτό δεν μπορεί να γίνει με επαρκώς σαφή ή/και αντικειμενικό τρόπο. Αυτό μπορεί να οφείλεται σε πολλούς παράγοντες, ο κυριότερος από τους οποίους είναι η διαφορετική αντίληψη δύο ξεχωριστών ανθρώπων/κριτών για την ίδια έννοια, ακόμη και εάν αυτή οι άνθρωποι διαθέτουν παρόμοιου βεληνεκούς εκπαίδευση στο υπό μελέτη αντικείμενο. Το πόσο καλά μπορούν δύο διακριτοί κριτές να κατηγοριοποιούν κομμάτια κειμένου στην ίδια κατηγορία, ορίζει την έννοια της συμφωνίας κριτών (interjudge

agreement)[48]. Μέσω αυτής της έννοιας μπορούμε να κατανοήσουμε δύο ποιοτικά στοιχεία για την κατηγοριοποίηση κειμένων που έχουμε παράξει.

- Πόσο *εύκολο* είναι να ορίσεις μια κατηγορία; Εάν π.χ. οι κριτές κάνουν τις ίδιες επιλογές κατηγορίες για τα περισσότερα κομμάτια κειμένου, τότε η κατηγορία είναι μάλλον επαρκώς ορισμένη.

- Πόσο *αξιόπιστη* είναι η κατηγοριοποίηση;

Για την ποσοτικοποίηση του interjudge agreement μπορούν να χρησιμοποιηθούν διάφορες μέθοδοι. Η πιο απλή μέθοδος είναι η εκτίμηση του ποσοστού συμφωνίας μεταξύ των κριτών. Αυτή η προσέγγιση είναι μεροληπτική. Για παράδειγμα εάν ο αριθμός των κατηγοριών είναι μικρός, π.χ. 2 κατηγορίες, τότε εκ των προτέρων, η πιθανότητα κατηγοριοποίησης στην ίδια κατηγορία από τους δύο κριτές είναι υψηλή (1/2). Παίρνοντας υπόψιν τα παραπάνω, στην παρούσα μελέτη χρησιμοποιήθηκε ο δείκτης Kappa του Cohen (Cohen's Kappa - CK) για την ποσοτικοποίηση του interjudge agreement[49]. Ο δείκτης CK πλεονεκτεί του απλού ποσοστού συμφωνίας επειδή συνεκτιμά την πιθανότητα οι δύο κριτές να συμφωνήσουν απλά και μόνο κατά τύχη. Ο δείκτης CK ισούται με:

$$\frac{P(a) - P(e)}{1 - P(e)}$$

όπου: P(a) η πραγματική συμφωνία των κριτών και P(e) η συμφωνία των κριτών που θα αναμενόταν εάν οι κριτές ταξινομούσαν με τυχαίο τρόπο τα δεδομένα.

Η ερμηνεία της αριθμητικής τιμής του δείκτη CK μπορεί να γίνει με τον παρακάτω τρόπο, αν και έχει ασκηθεί κριτική ως προς την σχετικά αυθαίρετη οριοθέτηση[50].

- < 0: καμία συμφωνία
- 0–0.20: ολίγη συμφωνία
- 0.21–0.40: κάποια συμφωνία
- 0.41–0.60: μέτρια συμφωνία
- 0.61–0.80: σημαντική συμφωνία
- 0.81–1: σχεδόν πλήρης συμφωνία

Η υλοποίηση της μέτρησης του CK έγινε μέσω του αντίστοιχου πακέτου του cohenkappascore του Scikit learn.

2.10 Ιδανική απόδοση (Topline Performance)

Ως baseline απόδοση των ταξινομητών, όπως έχει ειπωθεί, θα χρησιμοποιηθεί η απόδοση του τυχαίου ταξινομητή (dummy). Από την άλλη πλευρά, η ιδανική

απόδοση/topline performance θα προκύψει ως εξής[46]:

- Λαμβάνεται η κατηγοριοποίηση του κριτή A και επί αυτής, η κατηγοριοποίηση του κριτή B θεωρείται ως ταξινόμηση. Από αυτή την ταξινόμηση εξάγεται τιμή F1
- Γίνεται το αντίστοιχο για την κατηγοριοποίηση του κριτή B με την κατηγοριοποίηση του κριτή A να θεωρείται ως ταξινόμηση. Από αυτή την ταξινόμηση εξάγεται τιμή F1
- Υπολογίζεται ο μέσος όρος των τιμών

Η τιμή που προκύπτει θεωρείται ιδανική απόδοση, δηλαδή απόδοση που πέτυχαν οι πιο ειδικοί ταξινομητές, δηλαδή άνθρωποι, ειδικοί του αντικειμένου. Οι ταξινομητές Μηχανικής Μάθησης θα συγκριθούν εν τέλει και με αυτήν την παράμετρο με σκοπό τον καθορισμό της σχετικής ποιότητας της απόδοσής τους.

3 Αποτελέσματα

3.1 Επισκόπηση ενότητας (Overview)

Τα αποτελέσματα της παρούσας μελέτης παρουσιάζονται σε τρείς ενότητες οι οποίες αντιστοιχούν στις τρεις ομάδες ταξινομήσεων των κλινικών ερωτήσεων. Η παρουσίαση γίνεται σε μορφή πινάκων. Κάθε πίνακας αφορά την απόδοση των 5 ταξινομητών που χρησιμοποιήθηκαν επάνω σε κατηγοριοποίηση που είχε γίνει από έναν από τους δύο κριτές. Κατά σύμβαση παρουσιάζεται πρώτα η απόδοση των ταξινομητών σε κατηγοριοποίηση του κριτή Α και αμέσως μετά η απόδοση σε αυτή του κριτή Β.

Κάθε πίνακας χωρίζεται σε δύο μέρη. Το πρώτο αφορά στην απόδοση των μοντέλων ταξινόμησης. Το μέτρο σύγκρισης των μοντέλων είναι η $\overline{F1}$ δηλαδή η μέση τιμή του F1 score η οποία υπολογίστηκε από τα επιμέρους F1 scores που πέτυχαν οι ταξινομητές στους 3 κύκλους cross validation που πραγματοποιήθηκαν σε κάθε ταξινόμηση. Δίπλα στην μέση τιμή καταγράφεται η τυπική απόκλιση των F1 scores.

Το δεύτερο μέρος κάθε πίνακα περιγράφει την κατά ζεύγη σύγκριση των τιμών του πρώτου μέρους του πίνακα για τα διάφορα μοντέλα. Η στατιστική δοκιμασία που χρησιμοποιήθηκε σε όλες τις συγκρίσεις ήταν η Wilcoxon-Mann-Whitney καθώς όλες οι κατανομές ήταν μη κανονικές. Η αριθμητικές τιμές που καταγράφονται

στον παρόντα πίνακα είναι οι p-values της δοκιμασίας αυτής.

Παραλείφθηκαν από τα αποτελέσματα οι ταξινομήσεις C της ομάδας PICO καθώς και οι D, P της ομάδας TDPH καθώς οι ερωτήσεις του συνόλου δεδομένων ανήκαν σχεδόν εξολοκλήρου σε μια από τις δύο κλάσεις που η καθεμιά από αυτές τις ταξινομήσεις ορίζει.

Χάριν εξοικονόμησης χώρου οι ονομασίες των μοντέλων στους πίνακες γίνεται με βάση συντομογραφίες οι οποίες ορίζονται ως εξής:

- DU: ταξινομητής dummy
- BA: ταξινομητής Bayes
- RF: ταξινομητής Random Forest
- SVM: ταξινομητής Support Vector Machine
- NN: ταξινομητήςNeural Network

3.2 Ταξινόμηση Foreground/Background

Η ταξινόμηση του αρχικού συνόλου δεδομένων των κλινικών ερωτήσεων σε ερωτήσεις foreground και background (όπως αυτές ορίζονται στο υποκεφάλαιο 1.2.1.3 της Εισαγωγής), αποτελεί την κυριότερη ταξινόμηση της παρούσας εργασίας. Το αρχικό σύνολο δεδομένων περιείχε 320 ερωτήσεις. Ο κριτής Α (ταξινομητής αναφοράς) κατηγοριοποίησε 73 ερωτήσεις, ως ερωτήσεις foreground και 247 ερωτήσεις ως ερωτήσεις background. Με βάση αυτή την κατηγοριοποίηση (labeling) δημιουργήθηκε ένα αρχείο με ζεύγη ερωτήσεων - κατηγοριών το οποίο δόθηκε ως είσοδος στους πέντε προαναφερθέντες ταξινομητές. Οι ταξινομητές εκπαιδεύθηκαν και αξιολογήθηκαν με την μέθοδο που αναφέρεται στην επισκόπηση. Στον πίνακα 1 φαίνονται τα αποτελέσματα αυτής της ταξινόμησης. Ο ταξινομητής των νευρωνικών δικτύων (NN) απέδωσε σημαντικά καλύτερα από την τυχαία ταξινόμηση καθώς και από όλους τους άλλους ταξινομητές. Η διαφορά του μάλιστα από την τυχαία ταξινόμηση ήταν της τάξης του 0.2. Οι ταξινομητές BA, RF και SVM απέδωσαν σημαντικά χειρότερα από τον ταξινομητή DU, δηλαδή από την τυχαία ταξινόμηση.

Στην ίδια ταξινόμηση, ο κριτής Β κατηγοριοποίησε 106 ερωτήσεις, ως ερωτήσεις foreground και 214 ερωτήσεις ως ερωτήσεις background. Στον πίνακα 2 φαίνονται τα αποτελέσματα της ταξινόμησης από τους 5 ταξινομητές με βάση την

F1 score :$\overline{F1} \pm SD$				
DU 0.44±0.00	BA 0.03±0.04	RF 0.36±0.11	SVM 0.13±0.03	NN 0.62±0.05
Wilcoxon-Mann-Whitney p-values				
DU vs BA 0.000	RF vs DU 0.374	RF vs BA 0.016	RF vs SVM 0.047	RF vs NN 0.043
SVM vs DU 0.000	SVM vs BA 0.048	SVM vs NN 0.000	NN vs DU 0.010	NN vs BA 0.000

Πίνακας 1: Ταξινόμηση Foreground/Background-Κριτής Α

κατηγοριοποίηση του κριτή Β. Καλύτερα από όλους τους ταξινομητές, απέδωσε ο RF, αν και η υπεροχή του δεν έφτασε τη στατιστική σημαντικότητα όταν συγκρίθηκε με την τυχαία ταξινόμηση.

F1 score :$\overline{F1} \pm SD$				
DU 0.28±0.03	BA 0.20±0.04	RF 0.43±0.12	SVM 0.15±0.04	NN 0.19±0.27
Wilcoxon-Mann-Whitney p-values				
DU vs BA 0.073	RF vs DU 0.143	RF vs BA 0.057	RF vs SVM 0.035	RF vs NN 0.305
SVM vs DU 0.024	SVM vs BA 0.316	SVM vs NN 0.868	NN vs DU 0.669	NN vs BA 0.957

Πίνακας 2: Ταξινόμηση Foreground/Background-Κριτής Β

3.3 Ταξινόμηση PICO

Αυτή η ομάδα ταξινομήσεων αφορά μόνο τις ερωτήσεις foreground και περιλαμβάνει τις ταξινομήσεις των ερωτήσεων ανάλογα με το αν αυτές περιέχουν τα στοιχεία PICO, όπως αυτά ορίζονται στο υποκεφάλαιο 1.2.1.4 της εισαγωγής. Ως ερωτήσεις foreground θεωρήθηκαν όσες ερωτήσεις του αρχικού συνόλου δεδομένων κατηγοριοποιήθηκαν ως ερωτήσεις foreground από τον κριτή αναφοράς (κριτή A). Ως αρχικό σύνολο δεδομένων για τις ταξινομήσεις PICO θεωρήθηκαν λοιπόν, οι 73 ερωτήσεις foreground του κριτή αναφοράς. Παραλείφθηκε από τα αποτελέσματα η ταξινόμηση C της ομάδας PICO καθώς οι ερωτήσεις του συνόλου δεδομένων ανήκαν σχεδόν εξολοκλήρου σε μια από τις δύο κλάσεις της ταξινόμησης σύμφωνα με την ταξινόμηση και των δύο κριτών.

3.3.1 Ταξινόμηση P

Ο κριτής A κατηγοριοποίησε 53 ερωτήσεις, ως ερωτήσεις που περιείχαν στοιχείο P και 20 ερωτήσεις ως ερωτήσεις που το στοιχείο P απουσίαζε. Με βάση αυτή την κατηγοριοποίηση δημιουργήθηκε ένα αρχείο με ζεύγη ερωτήσεων - κατηγοριών το οποίο δόθηκε ως είσοδος στους πέντε ταξινομητές. Στον πίνακα 3 φαίνονται τα αποτελέσματα της ταξινόμησης από τους 5 ταξινομητές με βάση την κατηγοριοποίηση του κριτή A. Όλοι οι ταξινομητές ήταν σημαντικά καλύτεροι από την

τυχαία ταξινόμηση. Ο ταξινομητής RF πέτυχε το καλύτερο σκορ, χωρίς αυτό να υπερέχει σημαντικά των υπολοίπων.

F1 score :$\overline{F1} \pm SD$				
DU 0.63±0.03	BA 0.83±0.00	RF 0.86±0.01	SVM 0.83±0.00	NN 0.86±0.03
Wilcoxon-Mann-Whitney p-values				
DU vs BA 0.001	RF vs DU 0.000	RF vs BA 0.049	RF vs SVM 0.337	RF vs NN 0.910
SVM vs DU 0.001	SVM vs BA 1.000	SVM vs NN 0.259	NN vs DU 0.001	NN vs BA 0.259

Πίνακας 3: Ταξινόμηση P-Κριτής A

Ο κριτής B κατηγοριοποίησε 52 ερωτήσεις, ως ερωτήσεις που περιείχαν στοιχείο P και 21 ερωτήσεις ως ερωτήσεις που το στοιχείο P απουσίαζε. Στον πίνακα 4 φαίνονται τα αποτελέσματα της ταξινόμησης από τους 5 ταξινομητές με βάση αυτή την κατηγοριοποίηση. Όλοι οι ταξινομητές ήταν σημαντικά καλύτεροι από την τυχαία ταξινόμηση χωρίς ωστόσο κάποιος να ξεχωρίζει σημαντικά.

F1 score :$\overline{F1} \pm SD$				
DU 0.54±0.01	BA 0.83±0.00	RF 0.83±0.01	SVM 0.83±0.00	NN 0.83±0.00
Wilcoxon-Mann-Whitney p-values				
DU vs BA 0.000	RF vs DU 0.000	RF vs BA 0.498	RF vs SVM 0.498	RF vs NN 0.540
SVM vs DU 0.000	SVM vs BA 1.000	SVM vs NN 0.895	NN vs DU 0.000	NN vs BA 0.895

Πίνακας 4: Ταξινόμηση P-Κριτής B

Η συμφωνία των κριτών στην ταξινόμηση P ήταν 0.62 δηλαδή ήταν σημαντική

(βλ. πίνακα 13 και υποκεφάλαιο 2.9 για επεξήγηση).

3.3.2 Ταξινόμηση I

Ο κριτής Α κατηγοριοποίησε 68 ερωτήσεις, ως ερωτήσεις που περιείχαν στοιχείο I και 5 ερωτήσεις ως ερωτήσεις που το στοιχείο I απουσίαζε. Στον πίνακα 5 φαίνονται τα αποτελέσματα της ταξινόμησης από τους 5 ταξινομητές με βάση την κατηγοριοποίηση του αυτή. Όλοι οι ταξινομητές ήταν σημαντικά καλύτεροι από την τυχαία ταξινόμηση, χωρίς κάποιος να υπερέχει σημαντικά των υπολοίπων.

F1 score :$\overline{F1} \pm SD$				
DU 0.89±0.00	BA 0.96±0.00	RF 0.96±0.00	SVM 0.96±0.00	NN 0.96±0.00
Wilcoxon-Mann-Whitney p-values				
DU vs BA 0.000	RF vs DU 0.000	RF vs BA 1.000	RF vs SVM 1.000	RF vs NN 1.000
SVM vs DU 0.000	SVM vs BA 1.000	SVM vs NN 1.000	NN vs DU 0.000	NN vs BA 1.000

Πίνακας 5: Ταξινόμηση I-Κριτής Α

Ο κριτής Β κατηγοριοποίησε 63 ερωτήσεις, ως ερωτήσεις που περιείχαν στοιχείο I και 10 ερωτήσεις ως ερωτήσεις που το στοιχείο I απουσίαζε. Στον πίνακα 6 φαίνονται τα αποτελέσματα της ταξινόμησης από τους 5 ταξινομητές με βάση την κατηγοριοποίηση του αυτή. Όλοι οι ταξινομητές ήταν σημαντικά καλύτεροι από την τυχαία ταξινόμηση, χωρίς κάποιος να υπερέχει σημαντικά των υπολοίπων.

F1 score :$\overline{F1} \pm SD$				
DU 0.91±0.00	BA 0.96±0.01	RF 0.96±0.01	SVM 0.96±0.01	NN 0.96±0.01
Wilcoxon-Mann-Whitney p-values				
DU vs BA 0.002	RF vs DU 0.002	RF vs BA 1.000	RF vs SVM 1.000	RF vs NN 1.000
SVM vs DU 0.002	SVM vs BA 1.000	SVM vs NN 1.000	NN vs DU 0.002	NN vs BA 1.000

Πίνακας 6: Ταξινόμηση I-Κριτής B

Η συμφωνία των κριτών στην ταξινόμηση I ήταν 0 δηλαδή μικρού βαθμού (βλ. πίνακα 13 και υποκεφάλαιο 2.9 για επεξήγηση).

3.3.3 Ταξινόμηση O

Ο κριτής A κατηγοριοποίησε 59 ερωτήσεις, ως ερωτήσεις που περιείχαν στοιχείο O και 14 ερωτήσεις ως ερωτήσεις που το στοιχείο O απουσίαζε. Στον πίνακα 7 φαίνονται τα αποτελέσματα της ταξινόμησης από τους 5 ταξινομητές με βάση την κατηγοριοποίηση του αυτή. Όλοι οι ταξινομητές ήταν σημαντικά καλύτεροι από την τυχαία ταξινόμηση, χωρίς κάποιος να υπερέχει σημαντικά των υπολοίπων.

Ο κριτής B κατηγοριοποίησε 52 ερωτήσεις, ως ερωτήσεις που περιείχαν στοιχείο O και 14 ερωτήσεις ως ερωτήσεις που το στοιχείο O απουσίαζε. Στον πίνακα 8 φαίνονται τα αποτελέσματα της ταξινόμησης από τους 5 ταξινομητές με βάση την κατηγοριοποίηση του αυτή. Όλοι οι ταξινομητές ήταν σημαντικά καλύτεροι από

F1 score :$\overline{F1} \pm SD$				
DU 0.75±0.03	BA 0.90±0.01	RF 0.90±0.01	SVM 0.89±0.01	NN 0.89±0.02
Wilcoxon-Mann-Whitney p-values				
DU vs BA 0.002	RF vs DU 0.002	RF vs BA 0.453	RF vs SVM 0.453	RF vs NN 0.583
SVM vs DU 0.002	SVM vs BA 1.000	SVM vs NN 0.836	NN vs DU 0.006	NN vs BA 0.836

Πίνακας 7: Ταξινόμηση O-Κριτής A

την τυχαία ταξινόμηση, χωρίς κάποιος να υπερέχει σημαντικά των υπολοίπων.

F1 score :$\overline{F1} \pm SD$				
DU 0.65±0.10	BA 0.83±0.00	RF 0.81±0.03	SVM 0.83±0.00	NN 0.82±0.02
Wilcoxon-Mann-Whitney p-values				
DU vs BA 0.074	RF vs DU 0.109	RF vs BA 0.407	RF vs SVM 0.407	RF vs NN 0.808
SVM vs DU 0.074	SVM vs BA 1.000	SVM vs NN 0.427	NN vs DU 0.094	NN vs BA 0.427

Πίνακας 8: Ταξινόμηση O-Κριτής B

Η συμφωνία των κριτών στην ταξινόμηση O ήταν 0 δηλαδή μικρού βαθμού (βλ. πίνακα 13 και υποκεφάλαιο 2.9 για επεξήγηση).

3.4 Ταξινόμηση σε ερωτήσεις θεραπείας, διάγνωσης, πρόγνωσης και συσχέτισης(ταξινόμηση TDPH)

Σε αυτή την ομάδα ταξινομήσεων, οι ερωτήσεις ταξινομήθηκαν ανάλογα με το αν περιέχουν στοιχεία που αφορούν θεραπεία, διάγνωση, πρόγνωση και επιβάρυνση-συσχέτιση, όπως αυτά ορίζονται στο υποκεφάλαιο 1.2.1.5 της Εισαγωγής. Ως αρχικό σύνολο δεδομένων για τις ταξινομήσεις TDPH θεωρήθηκαν και εδώ, οι 73 ερωτήσεις foreground του κριτή αναφοράς. Παραλείφθηκαν από τα αποτελέσματα οι ταξινομήσεις D και P, καθώς οι ερωτήσεις του συνόλου δεδομένων ανήκαν σχεδόν εξολοκλήρου σε μια από τις δύο κλάσεις καθεμιάς από τις δύο ταξινομήσεις σύμφωνα με την κατηγοριοποίηση και των δύο κριτών.

3.4.1 Ταξινόμηση T

Ο κριτής A κατηγοριοποίησε 31 ερωτήσεις, ως ερωτήσεις T (που αφορούσαν θεραπεία) και 42 ερωτήσεις, ως ερωτήσεις που δεν περιείχαν στοιχεία T. Στον πίνακα 9 φαίνονται τα αποτελέσματα της ταξινόμησης από τους 5 ταξινομητές με βάση την κατηγοριοποίηση του αυτή. Οι ταξινομητές BA και RF ήταν σημαντικά καλύτεροι από την τυχαία ταξινόμηση, χωρίς κάποιος από τους δύο να υπερέχει σημαντικά του άλλου. Ο ταξινομητής NN είχε μηδενική απόδοση στην ταξινόμηση.

Ο κριτής B κατηγοριοποίησε 18 ερωτήσεις, ως ερωτήσεις T και 55 ερωτήσεις,

F1 score :$\overline{F1} \pm SD$				
DU 0.51±0.06	BA 0.83±0.07	RF 0.83±0.09	SVM 0.63±0.19	NN 0.00±0.00
Wilcoxon-Mann-Whitney p-values				
DU vs BA 0.008	RF vs DU 0.014	RF vs BA 0.940	RF vs SVM 0.261	RF vs NN 0.000
SVM vs DU 0.461	SVM vs BA 0.235	SVM vs NN 0.010	NN vs DU 0.000	NN vs BA 0.000

Πίνακας 9: Ταξινόμηση T-Κριτής A

ως ερωτήσεις που δεν περιείχαν στοιχεία T. Στον πίνακα 10 φαίνονται τα αποτελέσματα της ταξινόμησης από τους 5 ταξινομητές με βάση την κατηγοριοποίηση του αυτή. Μόνο ο ταξινομητής RF ήταν σημαντικά καλύτερος από την τυχαία ταξινόμηση. Οι ταξινομητές NN, BA και SVM είχαν σχεδόν μηδενική απόδοση στην ταξινόμηση.

F1 score :$\overline{F1} \pm SD$				
DU 0.33±0.06	BA 0.00±0.00	RF 0.53±0.21	SVM 0.10±0.13	NN 0.00±0.00
Wilcoxon-Mann-Whitney p-values				
DU vs BA 0.001	RF vs DU 0.276	RF vs BA 0.024	RF vs SVM 0.071	RF vs NN 0.024
SVM vs DU 0.084	SVM vs BA 0.374	SVM vs NN 0.374	NN vs DU 0.001	NN vs BA nan

Πίνακας 10: Ταξινόμηση T-Κριτής B

Η συμφωνία των κριτών στην ταξινόμηση T ήταν 0.61 δηλαδή σημαντικού βαθμού (βλ. πίνακα 13 και υποκεφάλαιο 2.9 για επεξήγηση).

3.4.2 Ταξινόμηση Η

Ο κριτής Α κατηγοριοποίησε 34 ερωτήσεις, ως ερωτήσεις Η (που αφορούσαν επιβαρυντικό παράγοντα-συσχέτιση) και 39 ερωτήσεις, ως ερωτήσεις που δεν περιείχαν στοιχεία Η. Στον πίνακα 11 φαίνονται τα αποτελέσματα της ταξινόμησης από τους 5 ταξινομητές με βάση την κατηγοριοποίηση του αυτή. Μόνο ο ταξινομητής RF ήταν σημαντικά καλύτερος από την τυχαία ταξινόμηση. Οι ταξινομητές BA και NN είχαν μηδενική απόδοση.

F1 score :$\overline{F1} \pm SD$				
DU 0.33±0.06	BA 0.00±0.00	RF 0.53±0.21	SVM 0.10±0.13	NN 0.00±0.00
Wilcoxon-Mann-Whitney p-values				
DU vs BA 0.001	RF vs DU 0.276	RF vs BA 0.024	RF vs SVM 0.071	RF vs NN 0.024
SVM vs DU 0.084	SVM vs BA 0.374	SVM vs NN 0.374	NN vs DU 0.001	NN vs BA nan

Πίνακας 11: Ταξινόμηση Η-Κριτής Α

Ο κριτής Β κατηγοριοποίησε 45 ερωτήσεις, ως ερωτήσεις Η και 28 ερωτήσεις, ως ερωτήσεις που δεν περιείχαν στοιχεία Η. Στον πίνακα 12 φαίνονται τα αποτελέσματα της ταξινόμησης από τους 5 ταξινομητές με βάση την κατηγοριοποίηση του αυτή. Οι ταξινομητές NN και SVM ήταν σημαντικά καλύτεροι από την τυχαία ταξινόμηση.

Η συμφωνία των κριτών στην ταξινόμηση Η ήταν 0.59 δηλαδή μέτριου βαθμού

F1 score :$\overline{F1} \pm SD$				
DU 0.58±0.11	BA 0.80±0.04	RF 0.80±0.02	SVM 0.82±0.03	NN 0.82±0.03
Wilcoxon-Mann-Whitney p-values				
DU vs BA 0.060	RF vs DU 0.054	RF vs BA 0.980	RF vs SVM 0.500	RF vs NN 0.503
SVM vs DU 0.043	SVM vs BA 0.631	SVM vs NN 0.959	NN vs DU 0.044	NN vs BA 0.619

Πίνακας 12: Ταξινόμηση H-Κριτής B

(βλ. πίνακα 13 και υποκεφάλαιο 2.9 για επεξήγηση).

3.5 Δείκτες Cohen's Kappa

Στον πίνακα 2.9 παρουσιάζονται συγκεντρωτικά οι τιμές του δείκτη συμφωνίας k στις διάφορες ταξινομήσεις. Η επεξήγηση του δείκτη γίνεται στο υποκεφάλαιο 2.9 της Εισαγωγής.

Cohen's Kappa					
FB	P	I	O	T	H
0,48	0,62	0	0	0,61	0,59

Πίνακας 13: Ο δείκτης Cohen's Kappa ανά ταξινόμηση

3.6 Topline επίδοση

Topline performance					
FB	P	I	O	T	H
0.89	0.93	0.75	0.73	0.81	0.62

Πίνακας 14: Topline performance ανά ταξινόμηση

4 Συζήτηση

Για τη διεκπεραίωση των ταξινομήσεων της παρούσας εργασίας δημιουργήθηκε, χαρακτηρίστηκε από ειδικούς και χρησιμοποιήθηκε ένα σύνολο δεδομένων αποτελούμενο από κλινικές ερωτήσεις και την κατηγοριοποίησή τους σε διάφορες κατηγορίες από ανθρώπους-κριτές, ειδικούς του του χώρου. Το σύνολο δεδομένων (βλέπε6), χρησιμοποίησε τμήμα των ερωτήσεων του BioASQ Challenge και με τη συμβολή ειδικών χαρακτηρίστηκε με τους τρόπους που περιγράφονται στο 2.2.2 και μπορεί να χρησιμοποιηθεί μελλοντικά από άλλες ερευνητικές ομάδες για τη δοκιμή διαφορετικών αλγορίθμων ή/και την κατασκευή συστημάτων αυτοματοποιημένης απάντησης ερωτήσεων. Η διαφορά του συνόλου δεδομένων της παρούσας εργασίας σε σχέση με τις αυθεντικές ερωτήσεις του BioASQ είναι η ενσωμάτωση πληροφορίας της Ιατρικής Βασισμένης σε Στοιχεία σε μέρος των δημοσιευμένων κλινικών ερωτήσεων.

Η κατηγοριοποίηση κλινικών ερωτήσεων σε τύπου foreground και τύπου background μέσω μεθόδων Μηχανικής Μάθησης επιχειρήθηκε για πρώτη φορά στην παρούσα εργασία. Τα νευρωνικά δίκτυα φάνηκε να επιτυγχάνουν σημαντικά μεγαλύτερο F1 score σε σχέση με όλους τους υπόλοιπους ταξινομητές στην ανάλυση επί της κατηγοριοποίησης του κριτή A (κατηγοριοποίηση αναφοράς) ενώ δεν ξεχώρησε η απόδοση κάποιου ταξινομητή στην κατηγοριοποίηση του κριτή B (βλ. Πίνακες 1, 2). Ενδεικτικά η μήτρα σύγχυσης (confusion matrix) των νευρωνικών δικτύων

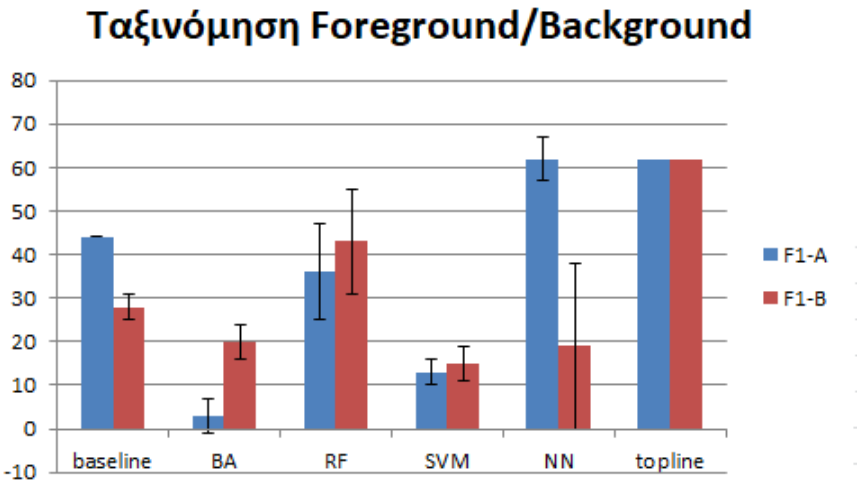
στην κατηγοριοποίηση foreground είχε ως εξής:

		Κατηγοριοποίηση κριτή Α	
		Foreground	Background
Πρόβλεψη μοντέλου	Foreground	17	7
	Background	8	74

Πίνακας 15: Μήτρα σύγχυσης για την κατηγοριοποίηση foreground/background από το μοντέλο νευρωνικού δικτύου σε ένα κύκλο cross validation

Η συμφωνία των κριτών στην ταξινόμηση foreground background έφτασε σε επίπεδα δείκτη Cohen's kappa 0.48, μια τιμή που χαρακτηρίζεται ως μέτρια συμφωνία. Όπως παρατηρούμε στο σχήμα 13 η απόδοση των νευρωνικών δικτύων ήταν ισάξια της ιδανικής επίδοσης (βλ. 2.10). Δεδομένης αυτής της άριστης απόδοσης αλλά και της μέτριας συμφωνίας για αυτή την ταξινόμηση, θεωρούμε οτι ο διαχωρισμός των ερωτήσεων σε ερωτήσεις foreground/background, η πρώτη και σημαντικότερη ταξινόμηση της παρούσας εργασίας, ήταν επιτυχής. Αυτό το είδος ταξινόμησης επιχειρείται για πρώτη φορά, γεγονός που καθιστά αυτό το εύρημα ακόμη πιο σημαντικό.

Στην ταξινόμηση των προτάσεων ανάλογα με το αν περιέχουν ή όχι τα στοιχεία PICO, η απόδοση των ταξινομητών κυμάνθηκε σε πιο χαμηλά επίπεδα, επιτυγχάνοντας ωστόσο τις περισσότερες φορές στατιστική σημαντικότητα σε σχέση με την τυχαία ταξινόμηση. Πιο συγκεκριμένα, στην ταξινόμηση P και οι 5 ταξινομητές



Σχήμα 13: Γραφική σύγκριση της απόδοσης των ταξινομητών σε σχέση με τις baseline και topline αποδόσεις στην ταξινόμηση foreground/background. Με μπλε χρώμα απεικονίζονται οι αποδόσεις των ταξινομητών στην κατηγοριοποίηση αναφοράς ενώ με κόκκινο στην κατηγοριοποίηση του κριτή Β.

ήταν σημαντικά καλύτεροι σε σχέση με την τυχαία ταξινόμηση για την κατηγοριοποίηση αναφοράς. Χωρίς να διαφέρουν σημαντικά μεταξύ τους (βλ. Πίνακες 3, 4) ήταν ισάξιοι με την κορυφαία απόδοση(topline performance, βλ. σχήμα 14). Στην κατηγοριοποίηση του κριτή Β μόνο ο ταξινομητής RF (random forest) απέδωσε σημαντικά καλύτερα από την τυχαιότητα αλλά υποδεέστερα από την κορυφαία επίδοση(βλ. σχήμα 14). Ενδεικτικά η μήτρα σύγχυσης (confusion matrix) των νευρωνικών δικτύων στην κατηγοριοποίηση P είχε ως εξής:

Στην πρόβλεψη της ύπαρξης των στοιχείων I και O όλοι οι ταξινομητές απέδωσαν σημαντικά καλύτερα από την τυχαία ταξινόμηση για τις κατηγοριοποιήσεις

		Κατηγοριοποίηση κριτή Β	
		Περιέχει P	Δεν περιέχει P
Πρόβλεψη μοντέλου	Περιέχει P	12	5
	Δεν περιέχει P	7	0

Πίνακας 16: Μήτρα σύγχυσης για την κατηγοριοποίηση P από το μοντέλο νευρωνικού δικτύου σε ένα κύκλο cross validation

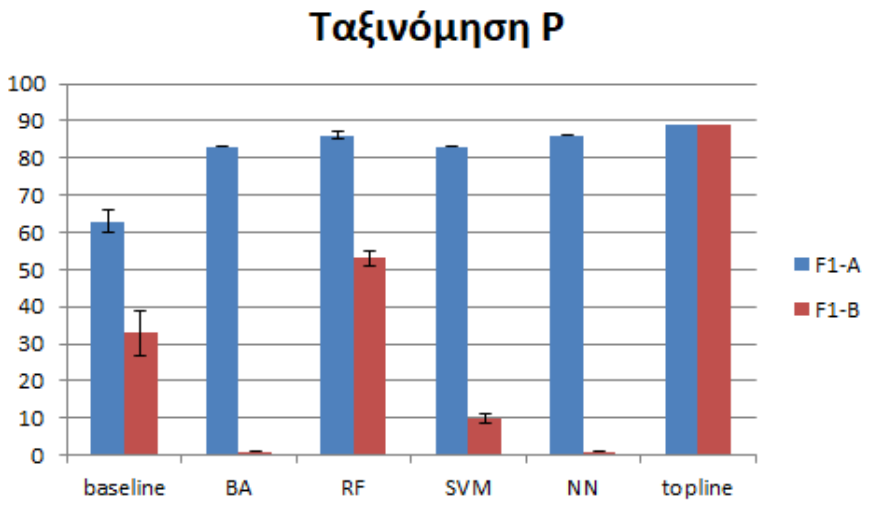
και από τους δύο κριτές (βλ. σχήματα 16, 15) . Παρόλα αυτά, όπως φαίνεται και στην παρακάτω μήτρα σύγχυσης (17) το γεγονός πως οι ερωτήσεις που δεν περιείχαν στοιχεία I ή O ήταν σημαντικά λιγότερες από αυτές που είχαν έκανε πιθανώς τα μοντέλα να αποδίδουν φαινομενικά καλύτερα διότι επέλεγαν πάντα την ταξινόμηση "1", ότι δηλαδή η ερώτηση περιέχει το στοιχείο σε σχέση με τον τυχαίο ταξινομητή που ισομοίραζε τις προβλέψεις του (stratification).

		Κατηγοριοποίηση κριτή Β	
		Περιέχει O	Δεν περιέχει O
Πρόβλεψη μοντέλου	Περιέχει O	20	0
	Δεν περιέχει O	5	0

Πίνακας 17: Μήτρα σύγχυσης για την κατηγοριοποίηση O από το μοντέλο SVM σε ένα κύκλο cross validation

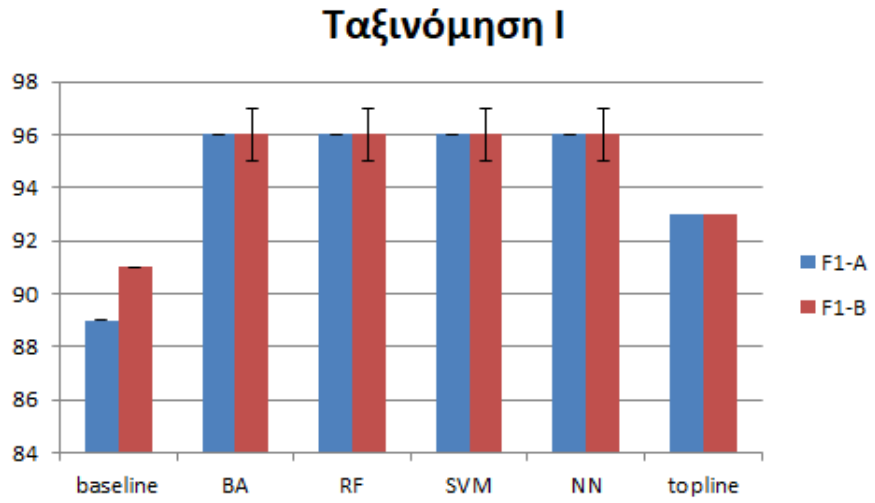
Οι δείκτες συμφωνίας των κριτών για τις κατηγοριοποιήσεις PICO ποίκιλλαν αρκετά. Ενώ στο στοιχείο P οι κριτές συμφώνησαν σε μέτριο βαθμό (k = 0.62) στα στοιχεία I και O οι κριτές η συμφωνία ήταν πτωχή. Το γεγονός αυτό μπορεί να οφείλεται στην εν γένει οντολογική ευρήτητα των δύο αυτών στοιχείων αλλά

και στην άνιση κατανομή των ερωτήσεων σε ομάδες, καθώς οι περισσότερες από τις ερωτήσεις περιείχαν τα στοιχεία I, O με αποτέλεσμα μεγάλο κομμάτι της συμφωνίας των κριτών να αποδίδεται στην τύχη επηρεάζοντας κατά συνέπεια άρδειν την τιμή του δείκτη k. Οι Niu και συνεργάτες περιγράφουν F1 δείκτη της τάξης του 0.83 στην κατηγοριοποίηση ερωτήσεων με βάση το στοιχείο O[23]. Αναφέρουν επίσης πως 53% των αρχικών τους προτάσεων περιείχαν στοιχείο O. Η παρούσα μελέτη αν και πέτυχε υψηλότερο F1 score στην εύρεση των στοιχείων O, είχε και υψηλότερο ποσοστό στοιχείων O (68%).

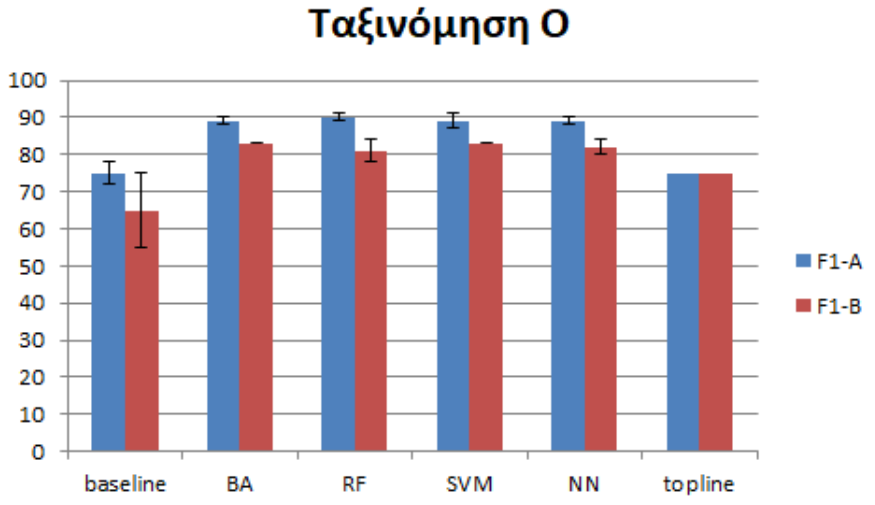


Σχήμα 14: Γραφική σύγκριση της απόδοσης των ταξινομητών σε σχέση με τις baseline και topline αποδόσεις στην ταξινόμηση P

Η ταξινόμηση των ερωτήσεων σε ερωτήσεις θεραπείας, διάγνωσης, πρόγνωσης και κινδύνου, συσχέτισης επιτεύχθηκε με επιτυχία από αρκετά μοντέλα μηχανικής μάθησης. Πιο συγκεκριμένα, στην ταξινόμηση T οι ταξινομητής Bayes, Ran-

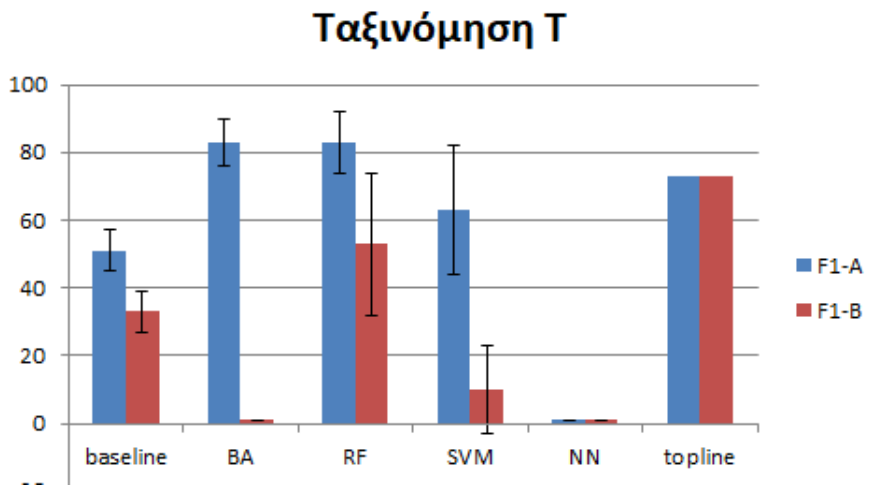


Σχήμα 15: Γραφική σύγκριση της απόδοσης των ταξινομητών σε σχέση με τις baseline και topline αποδόσεις στην ταξινόμηση I. Με μπλε χρώμα απεικονίζονται οι αποδόσεις των ταξινομητών στην κατηγοριοποίηση αναφοράς ενώ με κόκκινο στην κατηγοριοποίηση του κριτή B.

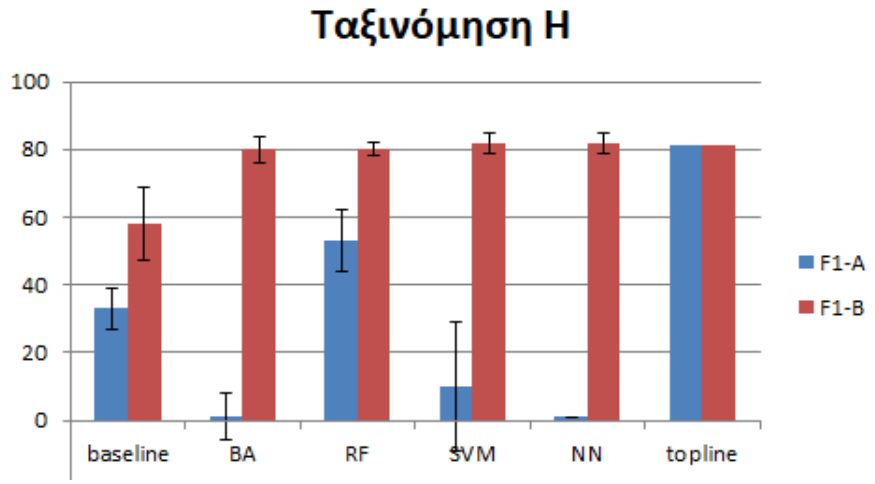


Σχήμα 16: Γραφική σύγκριση της απόδοσης των ταξινομητών σε σχέση με τις baseline και topline αποδόσεις στην ταξινόμηση O. Με μπλε χρώμα απεικονίζονται οι αποδόσεις των ταξινομητών στην κατηγοριοποίηση αναφοράς ενώ με κόκκινο στην κατηγοριοποίηση του κριτή B.

dom Forest απέδωσαν στατιστικά σημαντικότερα από τον ταξινομητή dummy για την κατηγοριοποίηση αναφοράς και ήταν ισάξια της topline performance (βλ. σχήμα 17). Στην ταξινόμηση Η ο ταξινομητής Random Forest απέδωσε σημαντικά καλύτερα από την τυχαία ταξινόμηση στην κατηγοριοποίηση αναφοράς ενώ όλοι οι ταξινομητές απέδωσαν καλύτερα από την τυχαιότητα σε αυτή του κριτή Β και αγγίξαν όλοι την topline performance χωρίς κάποιος από αυτούς να ξεχωρίζει σημαντικά. Η οριακά σημαντική συμφωνία των κριτών στις κατηγοριοποιήσεις Τ και Η (k=0.61 και 0.59 αντίχτοιχα) υπογραμμίζει τη σημασία της στατιστικά σημαντικής επιτυχίας στην ταξινόμηση των ερωτήσεων σε αυτές τις κατηγορίες.



Σχήμα 17: Γραφική σύγκριση της απόδοσης των ταξινομητών σε σχέση με τις baseline και topline αποδόσεις στην ταξινόμηση Τ. Με μπλε χρώμα απεικονίζονται οι αποδόσεις των ταξινομητών στην κατηγοριοποίηση αναφοράς ενώ με κόκκινο στην κατηγοριοποίηση του κριτή Β.



Σχήμα 18: Γραφική σύγκριση της απόδοσης των ταξινομητών σε σχέση με τις baseline και topline αποδόσεις στην ταξινόμηση Η. Με μπλε χρώμα απεικονίζονται οι αποδόσεις των ταξινομητών στην κατηγοριοποίηση αναφοράς ενώ με κόκκινο στην κατηγοριοποίηση του κριτή Β.

Η παρούσα εργασία είχε περιορισμούς. Ο αριθμός των κριτών που κατηγοριοποίησαν τις ερωτήσεις θα μπορούσε να είναι μεγαλύτερος. Ακόμη, ο αριθμός των ερωτήσεων foreground θα μπορούσε να γίνει ενδεχομένως μεγαλύτερος και να προκύψουν αναλύσεις μεγαλύτερης στατιστικής ισχύος. Επίσης η πηγή του συνόλου δεδομένων κλινικών ερωτήσεων θα μπορούσε να είναι πιο εξειδικευμένη στις κλινικές ερωτήσεις αντί του BioAsq, του οποίου οι ερωτήσεις είναι ευρέος φάσματος και οι κλινικές ερωτήσεις ένα υποσύνολο μόνο αυτού. Ωστόσο, αυτό καθιστά την παρούσα μελέτη ένα εργαλείο διαχωρισμού των ερωτήσεων του BioAsq και η δημιουργία στοχευμένων απαντήσεων για τις ερωτήσεις που κατηγοριοποιούνται ως κλινικές, σύμφωνα με τους κανόνες της Ιατρικής Βασισμένης

σε στοιχεία.

Στην παρούσα εργασία χρησιμοποιήθηκε η μέθοδος TF-IDF για την μετατροπή των ερωτήσεων σε διανύσματα για τη μετέπειτα επεξεργασία τους από τους ταξινομητές. Στο μέλλον οι ερωτήσεις θα μπορούσαν να υποστούν διαφορετική επεξεργασία με ενδεχόμενη ενσωμάτωση σημασιολογικών, συντακτικών και γραμματικών ιδιοτήτων των ερωτήσεων ως features με διάφορες τεχνικές της Επεξεργασίας Φυσικής Γλώσσας. Για την ανίχνευση των στοιχείων PICO συγκεκριμένα, στην παρούσα εργασία επιχειρήθηκε η ταξινόμηση των προτάσεων ανάλογα με το αν κάθε πρόταση περιείχε ή όχι κάθε μια από τις συνιστώσες PICO. Θα ήταν δόκιμο, να επιχειρηθεί η ακριβής εντόπιση μέσα σε κάθε πρόταση των όρων που αντιστοιχούν σε κάθε συνιστώσα με τεχνικές της Μηχανικής Μάθησης για την εξαγωγή οντοτήτων από κείμενο.

Τα δεδομένα της παρούσας εργασίας θα μπορούσαν να αξιοποιηθούν με διάφορους τρόπους. Η ανάλυση ερωτήσεων από μόνη της δεν μπορεί να αποτελεί ένα ολοκληρωμένο εγχείρημα. Ωστόσο μπορεί να αποτελέσει εφαλτήριο για τη βελτιστοποίηση της εύρεσης/δημιουργίας των απαντήσεων που λειτουργούν σε αυτές. Αν μια ερώτηση είναι τύπου background, για να βρούμε απάντηση δε χρειάζεται απαραίτητα να ανατρέξουμε στα πιο πρόσφατα στοιχεία καθώς συγγράμματα αναφοράς (reference textbooks) συνήθως επαρκούν. Αν από την άλλη πλευρά

η ερώτηση είναι foreground δηλαδή αφορά συγκεκριμένο ασθενή με συγκεκριμένο ερώτημα της οποίας την απάντηση η επιστημονική κοινότητα δεν έχει ακόμη οριστικοποιήσει, η βέλτιστη απάντηση καλό είναι να αναζητηθεί σε σύγχρονες(π.χ. Pubmed) πηγές, που ιδανικά θα περιέχουν συμπεράσματα από συστηματικές ανασκοπήσεις/ μετα-αναλύσεις. Ανάλογα με το αν μια ερώτηση foreground κατηγοριοποιείται ως ερώτηση θεραπείας ή ερώτηση κινδύνου/συσχέτισης, το είδος επιστημονικής μελέτης που βέλτιστα θα απαντά σε αυτή είναι διαφορετικό. Στην πρώτη περίπτωση η βέλτιστη απάντηση καλό είναι να προέρχεται από τυχαιοποιημένη ελεγχόμενη μελέτη (ή μετα-ανάλυση πολλών εξ αυτών) ενώ στη δεύτερη από μελέτη κοόρτης (ή μετα-ανάλυση πολλών εξ αυτών).

Πάνω στα ευρήματα της παρούσας μελέτης ευελπιστούμε μελλοντικά να στηριχτεί η δημιουργία ενός συστήματος υποστήριξης κλινικών αποφάσεων που θα συνεπικουρεί στο έργο των ιατρών δίνοντας την ευκαιρία για εξόρυξη της βέλτιστης πληροφορίας σε πραγματικό χρόνο.

5 Ανακεφαλαίωση και Μελλοντικοί στόχοι

Ανακεφαλαιώνοντας, στην παρούσα εργασία επιχειρήσαμε και καταφέραμε να χαρακτηρίσουμε και να ταξινομήσουμε, με συνδυασμό διάφορων μοντέλων Μηχανικής Μάθησης ένα υποσύνολο κλινικών ερωτήσεων του BioASQ challenge. Σε καθένα από τα τρία στάδια επεξεργασίας (βλ. υποκεφάλαιο 1.6 της Εισαγωγής) εκπαιδεύσαμε και αξιολογήσαμε 5 μοντέλα Μηχανικής Μάθησης στην ταξινόμηση των ερωτήσεων αυτών (βλ. Σχήμα 19). Συγκεκριμένα:

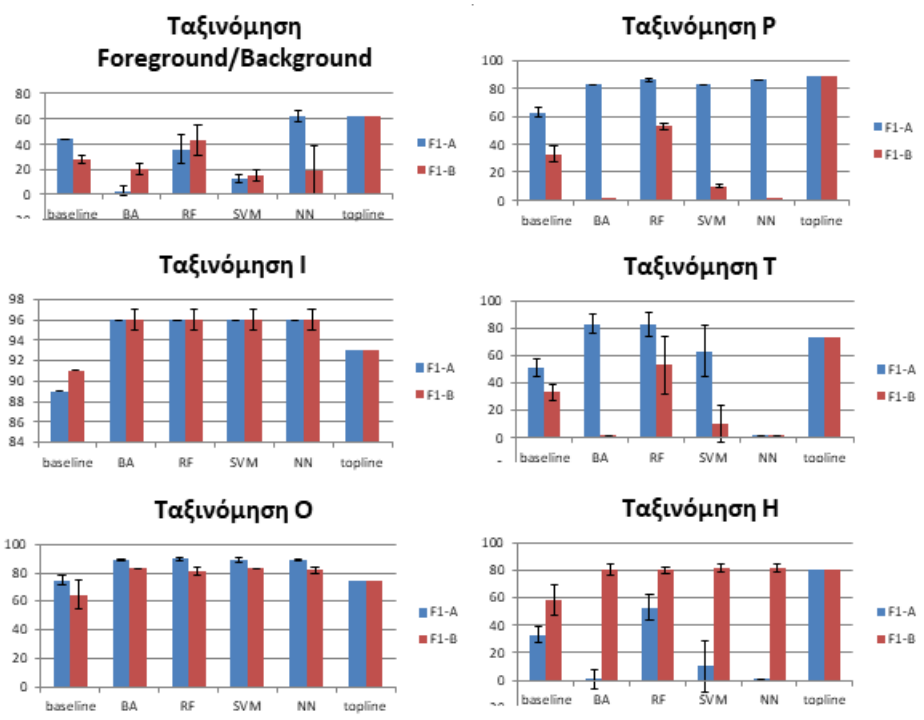
- Στο **πρώτο στάδιο** οι ταξινομητές NN(F1: 0.62 vs 0.44 της τυχαίας ταξινόμησης, για την κατηγοριοποίηση αναφοράς) και RF (F1: 0.43 vs 0.28 της τυχαίας ταξινόμησης, για την κατηγοριοποίηση κριτή B) ταξινόμησαν τις ερωτήσεις σημαντικά καλύτερα από την τυχαία ταξινόμηση. Η συμφωνία κριτών σε αυτή την ταξινόμηση ήταν μέτρια (k=0.48). Η απόδοση των NN ήταν ισάξια της topline performance, δηλαδή της ιδανικής απόδοσης ταξινόμησης με ανθρώπινα κριτήρια.

- Στο **δεύτερο στάδιο**, δηλαδή τον χαρακτηρισμό για τα στοιχεία PICO, οι ταξινομήσεις είχαν ως εξής: Στην κατηγοριοποίηση P όλοι οι ταξινομητές πέτυχαν F1 scores σημαντικά καλύτερα από την τυχαιότητα στην κατηγοριοποίηση αναφοράς και ισάξια με την ιδανική απόδοση. Στις κατηγοριοποιήσεις I και O ωστόσο, από τις μήτρες σύγχυσης φάνηκε οτι οι ταξινομητές συστηματικά ταξινομούσαν τις ερωτήσεις στην επικρατέστερη στο σύνολο

εκπαίδευσης κατηγορία. Μόνο στην κατηγορία P η συμφωνία κριτών ήταν σημαντική (k=0.62) ενώ στις κατηγορίες I και O ήταν μικρού βαθμού (k=0), γεγονός που αναδεικνύει την εγγενή δυσκολία των ταξινομήσεων αυτών.

- Στο **τρίτο στάδιο** δηλαδή στην ταξινόμηση των ερωτήσεων ανάλογα με το αντικείμενό τους (TDPH, θεραπεία, διάγνωση, πρόγνωση, συσχέτιση, επιβάρυνση), όσον αφορά στην ταξινόμηση T οι ταξινομητές BA και RF (0.83 vs 0.51 και οι δύο, κατηγοριοποίηση αναφοράς) είχαν σημαντικά καλή απόδοση, ισάξια της ιδανικής απόδοσης. Στην ταξινόμηση H, οι ταξινομητές RF (0.53 vs 0.33 της τυχαίας ταξινόμησης, για την κατηγοριοποίηση αναφοράς), SVM και NN (0.82 vs 0.58 και οι δυο έναντι της τυχαίας ταξινόμησης, για την κατηγοριοποίηση κριτή B) ταξινόμησαν τις ερωτήσεις σημαντικά καλύτερα από την τυχαία ταξινόμηση και απέδωσαν παρόμοια με την ιδανική απόδοση. Η συμφωνία κριτών ήταν οριακά σημαντική και μέτρια για τις ταξινομήσεις T και H αντίστοιχα (T: k=0.61, H: k=0.59).

Εν κατακλείδι δηλαδή, σε όλες τις ταξινομήσεις τουλάχιστον μία ταξινόμηση ήταν σημαντικά καλύτερη από την τυχαιότητα, με τις περισσότερες από αυτές που ήταν σημαντικές να πλησιάζουν την ιδανική απόδοση(με εξαίρεση τις εγγενείς δυσκολίες στις κατηγοριοποιήσεις I και O).



Σχήμα 19: Γραφική σύγκριση της απόδοσης των ταξινομητών σε σχέση με τις baseline και topline αποδόσεις σε όλες τις ταξινομήσεις συγκεντρωτικά. Με μπλε χρώμα απεικονίζονται οι αποδόσεις των ταξινομητών στην κατηγοριοποίηση αναφοράς ενώ με κόκκινο στην κατηγοριοποίηση του κριτή Β.

Οι ταξινομήσεις αφορούσαν κλινικές ερωτήσεις με ένα μεγάλο ποσοστό αυτών να αφορούν τον τομέα της Ογκολογίας. Τα αποτελέσματά μας αφορούν λοιπόν και την Ογκολογία, μπορούν ωστόσο να γενικευθούν και για λοιπές ιατρικές ειδικότητες.

Μελλοντικοί στόχοι μας, με εφαλτήριο την παρούσα εργασία αποτελούν:

- Η διεύρυνση του συνόλου δεδομένων των ερωτήσεων, συμπεριλαμβάνοντας

το σύνολο των δημοσιευθέντων κλινικών ερωτήσεων στο BioASQ challenge με σκοπό την αύξηση της στατιστικής ισχύος των ταξινομήσεων

- Η δημιουργία ενός συνόλου δεδομένων κλινικών ερωτήσεων αποκλειστικά για το πεδίο της Ογκολογίας με τη συμβολή ειδικών του τομέα με σκοπό την κατασκευή συστήματος επεξεργασίας ερωτήσεων που να αφορά αυτόν μόνο τον κλάδο και τις τελευταίες εξελίξεις στον τομέα

- Η χρήση άλλων μεθόδων Επεξεργασίας Φυσικής Γλώσσας που να εξορύσσουν με στοχευμένο τρόπο την πληροφορία εκεί που η μέθοδος TF-IDF φάνηκε να μη λειτουργεί καλά (ταξινομήσεις I, O)

- Η δημιουργία ενός ολοκληρωμένου συστήματος αυτόματης απάντησης κλινικών ερωτήσεων που να σέβεται την ιεράρχηση των στοιχείων της σύγχρονης Ιατρικής Βασισμένης σε Στοιχεία. Η παρούσα εργασία μπορεί να αποτελέσει τη βάση ενός τέτοιου συστήματος. Για να πραγματοποιηθεί κάτι τέτοιο πρέπει να υλοποιηθούν και τα υπόλοιπα στάδια σύνθεσης απάντησης (αναζήτηση και απόκτηση κειμένων, εύρεση σχετικών σημείων στα κείμενα, σύνθεση απάντησης κ.α.)

Ευελπιστούμε η παρούσα εργασία να αποτέλεσε ένα ευχάριστο ανάγνωσμα αλλά και ερέθισμα για περαιτέρω μελέτη πάνω στα αντικείμενα που πραγματεύθηκε,

όπως αυτά της Ιατρικής Βασισμένης σε Στοιχεία αλλά και της Μηχανικής Μάθησης. Και οι δύο τομείς αποτελούν και θα αποτελούν δύο πολύ σημαντικά εργαλεία προόδου της ανθρωπότητας.

6 Συμπληρωματικό υλικό(Supplemental Material)

Ο κώδικας που γράφτηκε στα πλαίσια της υλοποίησης της παρούσας εργασίας, όπως και τα σύνολα δεδομένων που χρησιμοποιήθηκαν για την παραγωγή των αποτελεσμάτων είναι διαθέσιμα σε αυτή την διεύθυνση. Πληροφορίες για τα περιεχόμενα του κάθε αρχείου βρίσκονται στο αρχείο README.md του καταλόγου αυτής της διεύθυνσης.

Παρακάτω παρατίθενται κάποιες online πηγές που κατά την άποψη του γράφοντα αποτελούν ομαλή και εκπαιδευτική εισαγωγή στα αντίστοιχα πεδία.

-Introduction in coding in Python: Codecademy

-Machine Learning: Machine Learning by Andrew Ng

-Introduction to Scikit-Learn: Introduction to Scikit-Learn

7 Βιβλιογραφία

References

[1] Sharon E Straus, Paul Glasziou, W Scott Richardson, and R Brian Haynes. *Evidence-Based Medicine E-Book: How to Practice and Teach EBM*. Elsevier Health Sciences, 2018.

[2] Matthew J Rioth, Travis J Osterman, and Jeremy L Warner. Advances in website information resources to aid in clinical practice. In *American Society of Clinical Oncology educational book/ASCO American Society of Clinical Oncology Meeting*, volume 35, pages e608–e615, 2015.

[3] Pamala A Pawloski, Gabriel A Brooks, Matthew E Nielsen, and Barbara A Olson-Bullis. A systematic review of clinical decision support systems for clinical oncology practice. *Journal of the National Comprehensive Cancer Network*, 17(4):331–338, 2019.

[4] Guides: Evidence-based medicine resource guide: Clinical questions, pico, study designs. http://guides.dml.georgetown.edu/ebm/ebmclinicalquestions.

[5] Jeff Weinfeld and Kate Finkelstein. How to answer your clinical questions more efficiently. *Family practice management*, 12(7):37, 2005.

[6] Karen Davies. Comparative analysis of questions posed by hospital-based physicians and physicians based in the primary care sector. *Journal of Hospital Librarianship*, 19(1):33–48, 2019.

[7] James W Drisko and Melissa D Grady. The steps of evidence-based practice in clinical practice: An overview. In *Evidence-Based Practice in Clinical Social Work*, pages 39–68. Springer, 2019.

[8] Cristina Mamédio da Costa Santos, Cibele Andrucioli de Mattos Pimenta, and Moacyr Roberto Cuce Nobre. The pico strategy for the research question construction and evidence search. *Revista latino-americana de enfermagem*, 15(3):508–511, 2007.

[9] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artiéres, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):138, Apr 2015.

[10] http://www.bioasq.org/.

[11] Rezarta Islamaj Dogan, G Craig Murray, Aurélie Névéol, and Zhiyong Lu. Understanding pubmed® user search behavior through log analysis. *Database*, 2009, 2009.

[12] Ioannis A Kakadiaris, George Paliouras, and Anastasia Krithara. Proceedings of the 6th bioasq workshop a challenge on large-scale biomedical semantic indexing and question answering. In *Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*, 2018.

[13] D Douglas Miller and Eric W Brown. Artificial intelligence in medical practice: the question to the answer? *The American journal of medicine*, 131(2):129–133, 2018.

[14] John W Ely, Jerome A Osheroff, Paul N Gorman, Mark H Ebell, M Lee Chambliss, Eric A Pifer, and P Zoe Stavri. A taxonomy of generic clinical questions: classification study. *Bmj*, 321(7258):429–432, 2000.

[15] Yoon-Ho Seol, David R Kaufman, Eneida A Mendonça, James J Cimino, and Stephen B Johnson. Scenario-based assessment of physicians' information needs. In *Medinfo*, pages 306–310, 2004.

[16] James J Cimino, Anthony Aguirre, Stephen B Johnson, and Ping Peng. Generic queries for meeting clinical information needs. *Bulletin of the Medical Library Association*, 81(2):195, 1993.

[17] Xiaoli Huang, Jimmy Lin, and Dina Demner-Fushman. Evaluation of pico as a knowledge representation for clinical questions. In *AMIA annual symposium proceedings*, volume 2006, page 359. American Medical Informatics Association, 2006.

[18] Florian Boudin, Jian-Yun Nie, Joan C Bartlett, Roland Grad, Pierre Pluye, and Martin Dawes. Combining classifiers for robust pico element detection. *BMC medical informatics and decision making*, 10(1):29, 2010.

[19] Ke-Chun Huang, I-Jen Chiang, Furen Xiao, Chun-Chih Liao, Charles Chih-Ho Liu, and Jau-Min Wong. Pico element detection in medical text without metadata: Are first sentences enough? *Journal of biomedical informatics*, 46(5):940–946, 2013.

[20] Di Jin and Peter Szolovits. PICO element detection in medical text via long short-term memory neural networks. In *Proceedings of the BioNLP 2018 workshop*, pages 67–75, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[21] Eya Znaidi, Lynda Tamine, and Chiraz Latiri. Answering pico clinical questions: A semantic graph-based approach. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 232–237. Springer, 2015.

[22] YongGang Cao, Feifan Liu, Pippa Simpson, Lamont Antieau, Andrew Bennett, James J Cimino, John Ely, and Hong Yu. Askhermes: An online question answering system for complex clinical questions. *Journal of biomedical informatics*, 44(2):277–288, 2011.

[23] Yun Niu and Graeme Hirst. Analysis of semantic classes in medical text for question answering. In *Proceedings of the Conference on Question Answering in Restricted Domains*, pages 54–61, 2004.

[24] Philip C Jackson. *Introduction to artificial intelligence*. Courier Dover Publications, 2019.

[25] Jaime G Carbonell, Ryszard S Michalski, and Tom M Mitchell. An overview of machine learning. In *Machine learning*, pages 3–23. Elsevier, 1983.

[26] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2009.

[27] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160:3–24, 2007.

[28] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[29] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[30] P Liang and NK Bose. Neural network fundamentals with graphs, algorithms and applications. *Mac Graw-Hill*, 1996.

[31] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551, 2011.

[32] Susan Gauch, Aravind Chandramouli, and Shankar Ranganathan. Training a hierarchical classifier using inter document relationships. *Journal of the American Society for Information Science and Technology*, 60(1):47–58, 2009.

[33] Charu C Aggarwal and ChengXiang Zhai. A survey of text classification algorithms. In *Mining text data*, pages 163–222. Springer, 2012.

[34] Joseph John Rocchio. The smart retrieval system: Experiments in automatic document processing. *Relevance feedback in information retrieval*, pages 313–323, 1971.

[35] Susan Dumais et al. Using svms for text categorization. *IEEE Intelligent Systems*, 13(4):21–23, 1998.

[36] Prafulla Bafna, Dhanya Pramod, and Anagha Vaidya. Document clustering: Tf-idf approach. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pages 61–66. IEEE, 2016.

[37] Todd Kulesza, Saleema Amershi, Rich Caruana, Danyel Fisher, and Denis Charles. Structured labeling for facilitating concept evolution in machine learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3075–3084. ACM, 2014.

[38] Grace YT Cheng. A study of clinical questions posed by hospital clinicians. *Journal of the Medical Library Association*, 92(4):445, 2004.

[39] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[40] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145.

Montreal, Canada, 1995.

[41] Stuart Russell and Peter Norvig. Ai a modern approach. *Learning*, 2(3):4, 2005.

[42] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. On the stratification of multi-label data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 145–158. Springer, 2011.

[43] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537, 2011.

[44] Peter Willett. The porter stemming algorithm: then and now. *Program*, 40(3):219–223, 2006.

[45] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[46] Donald Michie, David J Spiegelhalter, CC Taylor, et al. Machine learning. *Neural and Statistical Classification*, 13, 1994.

[47] Dennis Cooke and GM Clarke. *A basic course in statistics*. Arnold, 1989.

[48] Joseph L Fleiss, Bruce Levin, Myunghee Cho Paik, et al. The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2(212-236):22–23, 1981.

[49] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282, 2012.

[50] Nicole J-M Blackman and John J Koval. Interval estimation for cohen's kappa as a measure of agreement. *Statistics in medicine*, 19(5):723–741, 2000.