



ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εδνικόν και Καποδιστριακόν
Πανεπιστήμιον Αδηνών
ΙΔΡΥΘΕΝ ΤΟ 1837

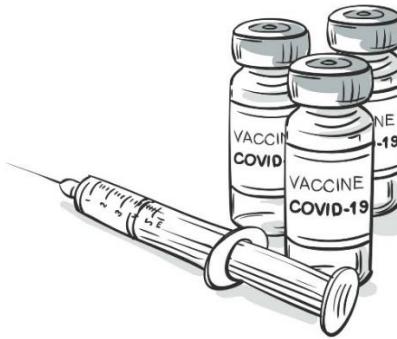
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εδνικό και Καποδιστριακό
Πανεπιστήμιο Αδηνών

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ
ΣΠΟΥΔΩΝ «ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ-
ΥΠΟΛΟΓΙΣΤΙΚΗ ΒΙΟΛΟΓΙΑ»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Πρόβλεψη της σοβαρότητας των παρενεργειών των εμβολίων κατά της COVID-19 με χρήση αλγορίθμων μηχανικής μάθησης»



Ελλη Ραμμένου

Πτυχιούχος Τμήματος Βιολογίας, Πανεπιστήμιο Πατρών

ΑΘΗΝΑ 2024



HELLENIC REPUBLIC

National and Kapodistrian
University of Athens

EST. 1837

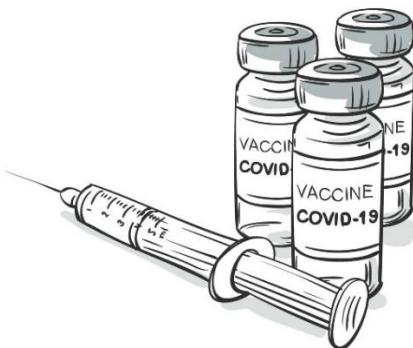
HELLENIC REPUBLIC
National and Kapodistrian
University of Athens

SCHOOL OF SCIENCE
DEPARTMENT OF BIOLOGY

MASTER IN «BIOINFORMATICS-
COMPUTATIONAL BIOLOGY»

Master Diploma Thesis

**«Prediction of severity of COVID-19 vaccines using
machine learning algorithms»**



ELLI RAMMENOU

BSc Biology, University of Patras

ATHENS 2024



ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εδνικόν και Καποδιστριακόν
Πανεπιστήμιον Αδηνών
ΙΔΡΥΘΕΝ ΤΟ 1837

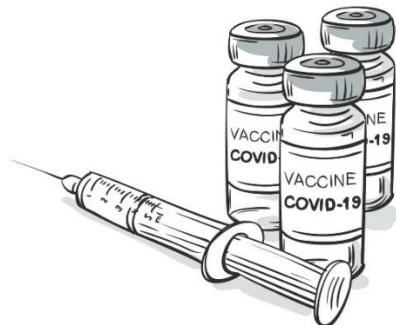
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εδνικό και Καποδιστριακό
Πανεπιστήμιο Αδηνών

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ
ΣΠΟΥΔΩΝ «ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ-
ΥΠΟΛΟΓΙΣΤΙΚΗ ΒΙΟΛΟΓΙΑ»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Πρόβλεψη της σοβαρότητας των παρενεργειών των εμβολίων
κατά της COVID-19 με χρήση αλγορίθμων μηχανικής μάθησης»



Τριμελής εξεταστική επιτροπή

Αναπληρώτρια Καδηγήτρια Βασιλική Οικονομίδου (Επιβλέπουσα)
Τομέας Βιολογίας Κυττάρου & Βιοφυσικής
Τμήμα Βιολογίας, ΕΚΠΑ
Καδηγητής Ιωάννης Τρουγκάκος
Τομέας Βιολογίας Κυττάρου & Βιοφυσικής
Τμήμα Βιολογίας, ΕΚΠΑ
Ιωάννης Μιχαλόπουλος, Ειδικός Λειτουργικός Επιστήμονας Β'
Ίδρυμα Ιατροβιολογικών Ερευνών Ακαδημίας Αδηνών

Ευχαριστίες

Θα ήθελα να απευθύνω ευχαριστίες στους ανθρώπους που με βοήθησαν και με στήριξαν κατά τη διάρκεια της εκπόνησης της συγκεκριμένης διπλωματικής εργασίας.

Αρχικά, θα ήθελα να ευχαριστήσω θερμά τον Ειδικό Λειτουργικό Επιστήμονα Β' του Ιδρύματος Ιατροβιολογικών Ερευνών της Ακαδημίας Αθηνών, Δρ. Ιωάννη Μιχαλόπουλο, μέλος της εξεταστικής μου επιτροπής που μου προσέφερε μία θέση στο εργαστήριό του και μου παρείχε τα απαραίτητα εφόδια για την πορεία μου ως βιοπληροφορικός, καθώς και χρήσιμες συμβουλές σχετικά με τη γενικότερη επιστημονική προσέγγιση. Θα ήθελα επίσης να απευθύνω τις ευχαριστίες μου στην Αναπληρώτρια Καθηγήτρια Βασιλική Οικονομίδου του Τμήματος Βιολογίας του Καποδιστριακού Πανεπιστημίου Αθηνών, στον Καθηγητή Ιωάννη Τρουγκάκο του Τμήματος Βιολογίας του Καποδιστριακού Πανεπιστημίου Αθηνών, επίσης μέλη της τριμελούς μου εξεταστικής επιτροπής, οι οποίοι υπήρξαν καθηγητές μου κατά τη διάρκεια του Προγράμματος Μεταπτυχιακών Σπουδών μεταφέροντας μου χρήσιμες πληροφορίες και γνώσεις για ποικίλα αντικείμενα του κλάδου της Βιοπληροφορικής.

Θα ήθελα επίσης να ευχαριστήσω το φίλο και συνεργάτη μου Δημήτριο Γεωργίου για την πολύτιμη βιόθειά του στο κομμάτι της μηχανικής μάθησης και της κατασκευής του μοντέλου, τον Βασίλειο Ζωγόπουλο για την πολύτιμη βιόθειά του στο τεχνικό κομμάτι της εργασίας, καθώς και τους υπόλοιπους συμφοιτητές μου στο εργαστήριο του Δρ. Ιωάννη Μιχαλόπουλου για το ευχάριστο κλίμα συνεργασίας.

Είμαι ιδιαίτερα ευγνώμων στους αγαπημένους μου φίλους και τους γονείς μου που ήταν δίπλα μου όλο αυτό το διάστημα και με στήριξαν στην απόφασή μου να ακολουθήσω αυτόν τον δρόμο.

Περιεχόμενα

1	Εισαγωγή.....	3
1.1	Iοί.....	3
1.1.1	Γενικές πληροφορίες.....	3
1.1.2	Ιστορική αναδρομή	3
1.1.3	Ορισμός.....	4
1.1.4	Δομή	4
1.1.5	Ταξινόμηση	4
1.2	Κορωνοϊοί.....	9
1.2.1	Ιστορική αναδρομή	9
1.2.2	Ταξινομική και γενικά χαρακτηριστικά κορονοϊών	10
1.2.3	Αλληλεπίδραση ιού – ξενιστή	11
1.2.4	SARS - CoV	12
1.2.5	SARS – CoV – 2	13
1.3	Ασθένεια CoViD-19.....	17
1.4	Εμβόλια κατά της CoViD-19	18
1.4.1	Εμβόλιο των Pfizer–BioNTech mRNA BNT162b2	18
1.4.2	Εμβόλιο της Moderna (mRNA-1273)	23
1.4.3	Εμβόλιο της AstraZeneca	24
1.4.4	Εμβόλιο της Johnson & Johnson (Janssen)	25
1.4.5	Σύνοψη χαρακτηριστικών εμβολίων	28
1.5	Βάση Δεδομένων EudraVigilance	29
1.5.1	Εισαγωγή	29
1.5.2	Δεδομένα της EudraVigilance	29
1.5.3	Αναζήτηση στη βάση.....	33
1.6	Μηχανική μάθηση.....	38
1.6.1	Ορισμός.....	38
1.6.2	Διαχωρισμός συνόλου δεδομένων	39
1.6.3	Τύποι Μηχανικής Μάθησης.....	40
1.6.4	SHAP (SHapley Additive exPlanations).....	42
2	Σκοπός.....	44
3	Μέθοδοι.....	45
3.1	Σύνολο δεδομένων – Γενικά Χαρακτηριστικά.....	45
3.2	Parsing του συνόλου δεδομένων	46
3.2.1	Γλώσσα PHP	46

3.2.2	Parser	47
3.3	Δημιουργία βάσης δεδομένων	49
3.3.1	Γλώσσα SQL.....	50
3.3.2	MySQL DBMS.....	51
3.3.3	MySQL Workbench.....	53
3.3.4	Διάγραμμα οντοτήτων - συσχετίσεων (ERD)	54
3.4	Μηχανικη μαθηση.....	59
3.4.1	Δημιουργία αρχείου input για την εκπαίδευση του μοντέλου	59
3.4.2	Γλώσσα Python.....	60
3.4.3	Data pre-processing (προ-επεξεργασία δεδομένων)	61
3.4.4	Διαχωρισμός συνόλου δεδομένων	65
3.4.5	Αλγόριθμοι κατηγοριοποίησης	65
3.4.6	SHAP (SHapley Additive exPlanations).....	76
3.5	Μετρικές αξιολόγησης	77
3.5.1	Accuracy	79
3.5.2	Recall	80
3.5.3	Precision	80
3.5.4	F1 score	81
3.6	Γενικές τεχνικές πληροφορίες	82
4	Αποτελέσματα	83
4.1	Περιγραφή συνόλου δεδομένων	83
4.1.1	Συνολικός αριθμός εγγραφών	83
4.1.2	Περιγραφικά στοιχεία	83
4.2	Διάγραμμα οντοτήτων – συσχετίσεων.....	88
4.3	Προ-επεξεργασία δεδομένων	91
4.4	Κλάση του μοντέλου ταξινόμησης	91
4.5	Μετρικές αξιολόγησης αλγορίθμων ταξινόμησης	94
4.5.1	Πείραμα A	95
4.5.2	Πείραμα B	95
4.5.3	Πείραμα Γ	96
4.6	Αποτελέσματα ανάλυσης SHAP	101
5	Συζήτηση	105
5.1	Μορφή του συνόλου δεδομένων.....	105
5.2	Τεχνικές προ-επεξεργασίας δεδομένων	106
5.3	Απόδοση αλγορίθμων ταξινόμησης.....	108
5.3.1	Σύγκριση για διαφορετικά μεγέθη συνόλου δεδομένων	108

5.3.2	Συνολική επίδραση του μεγέθους του δείγματος.....	110
5.4	Ανάλυση SHAP.....	111
5.5	Εναλλακτικές Προσεγγίσεις.....	114
6	Συμπεράσματα.....	115
7	Βιβλιογραφία	116

ΠΕΡΙΛΗΨΗ

Η πανδημία COVID-19 έχει παρουσιάσει μια άνευ προηγουμένου παγκόσμια πρόκληση για την υγεία μέχρι και σήμερα, που οδήγησε στην εξαιρετικά ταχεία ανάπτυξη και έκτακτη χρήση εμβολίων κατά της νόσου. Παρά την αποτελεσματικότητά τους, τα εμβόλια μπορεί να προκαλέσουν ανεπιθύμητες αντιδράσεις, απαιτώντας ισχυρά εργαλεία για την πρόβλεψη και τη διαχείριση της σοβαρότητάς τους. Σε αυτή τη διπλωματική εργασία, αναπτύξαμε έναν αλγόριθμο ταξινόμησης μηχανικής μάθησης που στοχεύει στην πρόβλεψη του επιπέδου σοβαρότητας των παρενέργειών των εμβολίων κατά της COVID-19 χρησιμοποιώντας δεδομένα που ανακτήθηκαν από τη βάση δεδομένων EudraVigilance. Η μεθοδολογία μας περιελάμβανε την εκπαίδευση διαφόρων αλγορίθμων ταξινόμησης σε υποσύνολα δεδομένων με αυξανόμενα μεγέθη για την αξιολόγηση της απόδοσης σε διαφορετικά μεγέθη δειγμάτων. Μεταξύ αυτών των αλγορίθμων, ο Random Forest και ο XGBoost παρουσίασαν την πιο υποσχόμενη απόδοση, με τον XGBoost να επιδεικνύει ένα μικρό πλεονέκτημα στην ακρίβεια πρόβλεψης. Στη συνέχεια, εφαρμόσαμε ανάλυση SHAP (SHapley Additive Explanations) στο εκπαιδευμένο μοντέλο XGBoost για να αποσαφηνίσουμε τη σημασία συγκεκριμένων γνωρισμάτων. Τα ευρήματά μας αποκαλύπτουν την ηλικία ως τον πιο κρίσιμο προγνωστικό παράγοντα, υποδεικνύοντας ότι τα ηλικιαμένα άτομα είναι πιο επιρρεπή στο να εμφανίσουν σοβαρές παρενέργειες μετά τον εμβολιασμό. Επιπλέον, συγκεκριμένα συμπτώματα όπως πόνος στο στήθος, υπερευαισθησία, έμετος, δύσπνοια και επιληπτικές κρίσεις, μαζί με το εμβόλιο Moderna, εμφανίστηκαν ως σημαντικοί παράγοντες που σχετίζονται με την αυξημένη σοβαρότητα των αντιδράσεων. Το συγκεκριμένο μοντέλο δύναται να προσφέρει πληροφορίες για δημογραφικούς και συμπτωματικούς παράγοντες που επηρεάζουν τις παρενέργειες των εμβολίων. Με τον εντοπισμό ομάδων υψηλού κινδύνου και συναφών συμπτωμάτων, οι επαγγελματίες υγείας μπορούν να δώσουν προτεραιότητα στις στρατηγικές παρακολούθησης και παρέμβασης, δυνητικά μετριάζοντας τις δυσμενείς εκβάσεις. Επιπρόσθετα, η ενσωμάτωση τεχνικών μηχανικής μάθησης με βάσεις δεδομένων φαρμακοεπαγρύπνησης, όπως η EudraVigilance, αποτελεί ένα ισχυρό εργαλείο για παρακολούθηση και αξιολόγηση κινδύνου σε πραγματικό χρόνο σε μια προσπάθεια παρακολούθησης της ασφάλειας των εμβολίων. Συμπερασματικά, η μελέτη μας υπογραμμίζει τη σημασία της αξιοποίησης των αλγορίθμων μηχανικής μάθησης για να βελτιώσουμε την κατανόησή μας για τα προφίλ ασφάλειας των εμβολίων εν μέσω της πανδημίας COVID-19, και συμβάλλει στις συνεχείς προσπάθειες για τη βελτιστοποίηση των στρατηγικών εμβολιασμού και τη διασφάλιση της ευημερίας των πληθυσμών παγκοσμίως.

ABSTRACT

The COVID-19 pandemic has presented an unprecedented global health challenge, leading to the rapid development and emergency use of vaccines against the disease. Despite their efficacy, vaccines may induce adverse reactions, necessitating robust tools for predicting and managing their severity. In this study, we developed a machine learning classification algorithm aimed at predicting the level of severity of COVID-19 vaccine reactions using data retrieved from the EudraVigilance database. Our methodology involved training various classification algorithms on subsets of data with increasing sizes to evaluate performance across different sample sizes. Among these algorithms, Random Forest and XGBoost exhibited the most promising performance, with XGBoost demonstrating a slight advantage in predictive accuracy. Subsequently, we applied SHAP (SHapley Additive exPlanations) analysis to the trained XGBoost model to elucidate feature importance. Our findings reveal age as the most critical predictor, indicating that older individuals are more prone to experiencing severe vaccine reactions. Furthermore, specific symptoms such as chest pain, hypersensitivity, vomiting, dyspnoea, and seizures, along with the Moderna vaccine, emerged as significant factors associated with heightened severity of reactions. The implications of our model are profound, offering insights into demographic and symptomatic factors influencing vaccine reactions. By identifying high-risk groups and associated symptoms, healthcare professionals can prioritize monitoring and intervention strategies, potentially mitigating adverse outcomes. Moreover, the integration of machine learning techniques with pharmacovigilance databases like EudraVigilance presents a powerful tool for real-time surveillance and risk assessment in vaccine safety monitoring efforts. In conclusion, our study underscores the importance of leveraging machine learning algorithms to enhance our understanding of vaccine safety profiles amidst the COVID-19 pandemic and contributes to the ongoing efforts to optimize vaccination strategies and ensure the well-being of populations worldwide.

1 ΕΙΣΑΓΩΓΗ

1.1 Ιοι

1.1.1 Γενικές πληροφορίες

Στα βιολογικά συστήματα, στοιχειώδη μονάδα οργάνωσης αποτελεί το κύτταρο που επιτελεί τις βασικές βιολογικές λειτουργίες. Όπως τα ζώα και τα φυτά, έτσι και οι μικροοργανισμοί είναι δυνατό να είναι είτε μονοκύτταροι είτε πολυκύτταροι και τα κύτταρά τους είναι ικανά να επιτελούν όλες τις βασικές βιολογικές λειτουργίες, όπως ακριβώς των φυτών και των ζώων. Παρ' όλα αυτά πρόκειται για απλούστερους οργανισμούς με διαφορετική αρχιτεκτονική δομή. Με βάση αυτή, οι μικροοργανισμοί κατατάσσονται σε προκαρυωτικούς (αρχαία, βακτήρια) και ευκαρυωτικούς (μύκητες, φύκη, πρωτόζωα) με βασική διαφορά πως το γενετικό υλικό των ευκαρυωτικών οργανισμών βρίσκεται σε ξεχωριστό διαμέρισμα, τον πυρήνα (Αγγελής, 2007).

Οι ιοί θεωρούνται παραδοσιακά μικροοργανισμοί, παρ' όλα αυτά δεν κατατάσσονται σε καμία από τις δύο κατηγορίες που αναφέρθηκαν, καθώς δε διαθέτουν κυτταρική οργάνωση. Συγκεκριμένα, ένας ιός αποκτά υπόσταση ζωντανού οργανισμού μόνο όταν βρεθεί στο εσωτερικό του κυττάρου ενός άλλου οργανισμού (Αγγελής, 2007).

1.1.2 Ιστορική αναδρομή

Η πρωταρχική ανακάλυψη των ιών έγινε μέσω της νόσου μωσαϊκού του καπνού που μεταδίδεται στα φυτά. Ο ιός του μωσαϊκού του καπνού (*tobacco mosaic virus*, TMV) ανακαλύφθηκε αρχικά από τα πειράματα του Γερμανού επιστήμονα Adolf Mayer (1843 – 1942), διευθυντή του Agricultural Experiment Station στο Wageningen της Ολλανδίας, το 1879 ενώ τα ευρήματα του δημοσιεύτηκαν το 1886 περιγράφοντας την ασθένεια που προκαλεί ο ιός και τα συμπτώματα της (Zaitlin, 1998). Παρ' όλα αυτά, τα εύσημα για την ανακάλυψη του ιού αποδίδονται συχνότερα στο Ρώσο επιστήμονα Dimitrii Iwanowski (1864 – 1920) που παρουσίασε τα ευρήματα του το 1892 στο Academy of Science στην Αγία Πετρούπολη της Ρωσίας (Iwanowski, 1968; Zaitlin, 1998). 6 χρόνια αργότερα, το 1898, ένας ακόμα επιστήμονας ο Ολλανδός Martinus Beijerinck (1851 – 1931), μη γνωρίζοντας την παράλληλη μελέτη του Iwanowsky (Lustig & Levine, 1992), δημοσίευσε ένα άρθρο με μια λεπτομερή περιγραφή της ασθένειας του μωσαϊκού του καπνού καθώς και του παράγοντα που την προκαλεί (Beijerinck, 1898; Lwoff, 1957; Zaitlin, 1998).

Έτσι λοιπόν, παρά το γεγονός ότι η σύλληψη της ιδέας των μολυσματικών παραγόντων και των ιδιοτήτων τους υπήρχε ήδη από πολύ νωρίτερα το 1840, όπως για παράδειγμα από τον ανατομιστή Jacob Henle, αυτή η ιδέα δεν κατάφερε να έχει ιδιαίτερη απήχηση λόγω έλλειψης πειραματικής τεκμηρίωσης. Η πειραματική επιβεβαίωση ήρθε με τα πειράματα των Adolf Mayer, Dimitrii Iwanowski και Martinus Beijernick, όπως προαναφέρθηκε (Lustig & Levine, 1992).

Ο τελικός καθαρισμός και απομόνωση του ιού έγινε το 1935 από τον Wendell Meredith Stanley (1904 – 1971) με τα πειράματα κρυστάλλωσης που πραγματοποίησε (Stanley, 1935).

1.1.3 Ορισμός

Σε μια δημοσίευση του, ο Γάλλος μικροβιολόγος André Michel Lwoff (1902 – 1994) έδωσε έναν ορισμό για τους ιούς: οι ιοί είναι μολυσματικές, δυνητικά παθογόνες, νουκλεοπρωτεΐνικές οντότητες που διαθέτουν μόνο έναν τύπο νουκλεϊκού οξέος, που αναπαράγονται από το γενετικό τους υλικό, που δεν μπορούν να αναπτυχθούν και να υποστούν δυαδική σχάση και στερούνται του συστήματος Lipmann (Lwoff, 1957).

1.1.4 Δομή

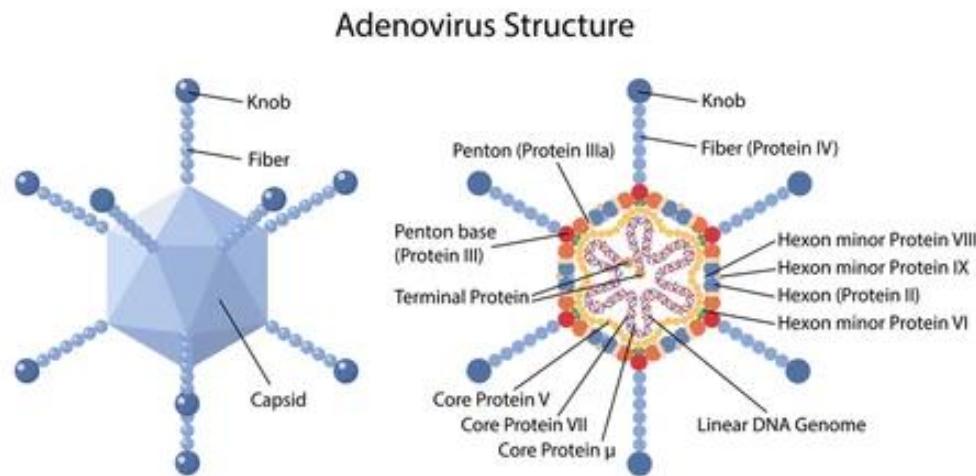
Οι ιοί αποτελούνται από ένα μόριο DNA ή RNA το οποίο περιβάλλεται από ένα πρωτεΐνικό περίβλημα, το καψίδιο. Το μόριο DNA ή RNA αποτελεί το γενετικό τους υλικό και περιλαμβάνει όλες τις απαραίτητες πληροφορίες για την αναπαραγωγή τους, όμως καθώς δε διαθέτουν στοιχειώδη κυτταρική δομή, δεν είναι δυνατό να υποστηρίξουν τις βιολογικές λειτουργίες των υπόλοιπων ζωντανών οργανισμών (Αγγελής, 2007).

1.1.5 Ταξινόμηση

Ουσιαστικά, δεν υπάρχει απόλυτα ικανοποιητικό σύστημα ταξινόμησης των ιών, παρ' όλα αυτά στο σύστημα Lwoff η ταξινόμηση των ιών βασίζεται στον τύπο του γονιδιώματος, στη μορφολογία και το μέγεθος του καψιδίου, στον αριθμό των καψομερών και στην ύπαρξη ή όχι ελύτρου. Παράλληλα, ένα δευτερεύοντα ρόλο στην ταξινόμηση παίζουν και οι ξενιστές των ιών καθώς η παθολογία και η συμπτωματολογία των ιών εξαρτάται και από αυτούς. Ανάλογα με τον ξενιστή που προσβάλουν, οι ιοί ομαδοποιούνται σε τρεις μεγάλες κατηγορίες: σε ιούς ζώων, σε ιούς φυτών και σε βακτηριοφάγους ιούς. Αυτές οι ομάδες ταξινομούνται περεταίρω με βάση τα μορφολογικά, βιοχημικά και ανοσοβιολογικά χαρακτηριστικά (Αγγελής, 2007).

1.1.5.1 Ιοί ζώων

- Αδενοϊό: είναι ιοί με δίκλωνο DNA (dsDNA) με καψίδιο εικοσαεδρικής συμμετρίας με 252 καψομερή. Οι κορυφές του εικοσαέδρου φέρουν χαρακτηριστικές ακίδες που εμπλέκονται στην πρόσφυση του ιοσωματίου στην επιφάνεια του κυττάρου-ξενιστή. Μια τυπική δομή ενός αδενοϊού φαίνεται στην παρακάτω εικόνα.



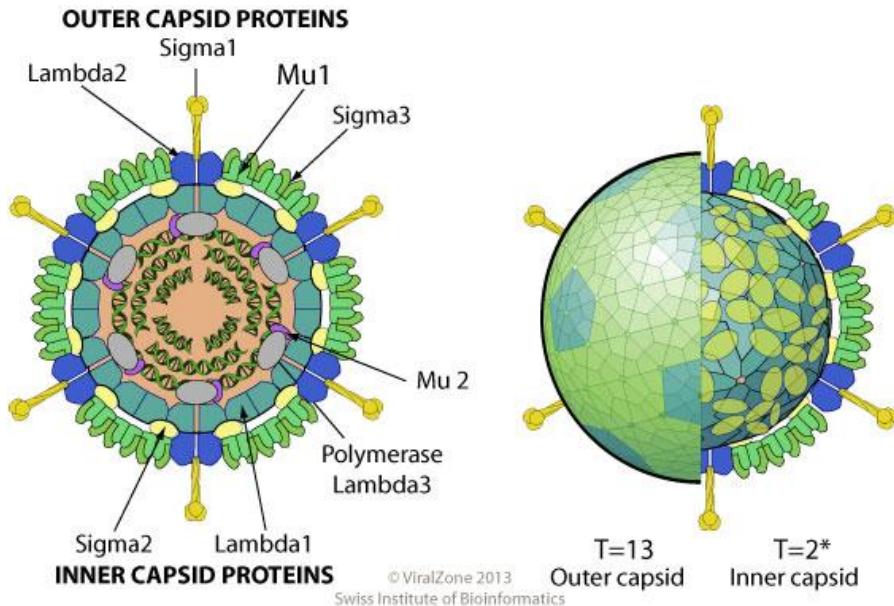
shutterstock.com · 1736109233

Εικόνα 1. Δομή αδενοϊού. Πηγή: <https://www.shutterstock.com>

Οι ιοί αυτής της κατηγορίας προσβάλλουν τόσο τον άνθρωπο όσο και τα ζώα. Συγκεκριμένα, μετά τη μόλυνση του κυττάρου-ξενιστή με τον αδενοϊό, αυτό συνεχίζει το φυσιολογικό μεταβολισμό του για μια χρονική περίοδο στην οποία το νουκλεϊκό οξύ του ιού ελευθερώνεται από το καψίδιο και μεταναστεύει στον πυρήνα του κυττάρου μέσω ενός πυρηνικού πόρου. Μέσα στον πυρήνα εκμεταλλεύεται τα απαραίτητα βιολογικά στοιχεία και πολλαπλασιάζεται, ενώ ταυτόχρονα εμποδίζει τη βιοσύνθεση των φυσιολογικών μακρομορίων του κυττάρου. Όλα τα γονίδια του ιού μεταγράφονται σε mRNA που μεταφράζεται στα ριβοσώματα του ξενιστή προς πρωτεΐνες. Η τελική αναδόμηση του ιοσωματίου πραγματοποιείται στον πυρήνα του κυττάρου. Τέλος, το κύτταρο λύεται και τα ιοσωμάτια απελευθερώνονται για να μολύνουν εκ νέου άλλα κύτταρα.

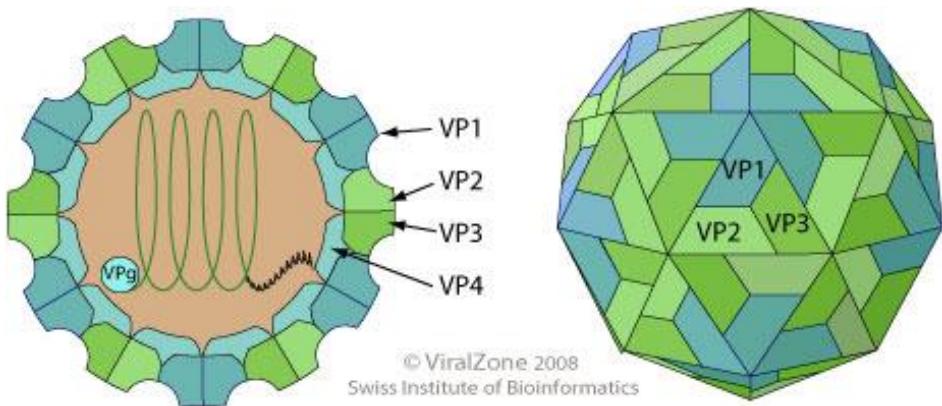
- Ρεοϊό: είναι ιοί με δίκλωνο RNA (dsRNA) που προσβάλλουν διάφορα είδη εντόμων και ανώτερων ζώων. Το γονίδιωμα τους αποτελείται από δέκα μόρια dsRNA, ενώ το καψίδιο τους είναι εικοσαεδρικής συμμετρίας όπως και στους αδενοϊούς. Οι ιοί αυτοί προσβάλλουν και προσφύονται στα κύτταρα μέσω αλληλεπίδρασης

συγκεκριμένων πρωτεΐνών με την επιφάνεια του κυττάρου. Μετά την είσοδο του ιοσωματίου στο κύτταρο, το καψίδιο απομακρύνεται μερικώς ενώ το γονιδίωμα εκφράζεται πλήρως. Μια τυπική δομή ρεοϊού φαίνεται στην παρακάτω εικόνα.



Εικόνα 2. Δομή ρεοϊού. Πηγή: <https://viralzone.expasy.org/>

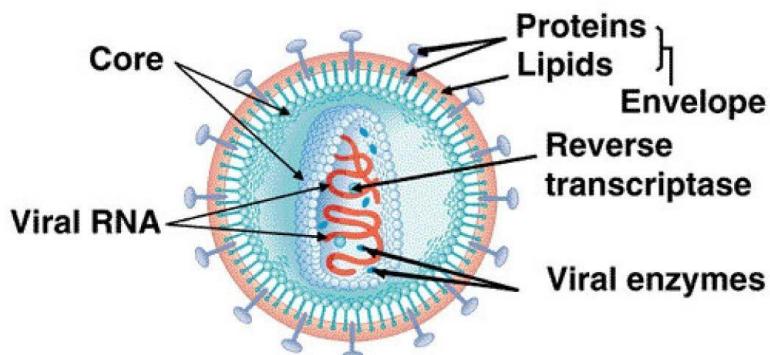
- Πικορναϊοί: οι ιοί πικορνά είναι μικροί ιοί με μονόκλωνο RNA (ss(+)RNA) με εικοσαεδρικό καψίδιο και χωρίς έλυτρο. Οι ιοί αυτοί προκαλούν σοβαρές ασθένειες στον άνθρωπο και τα ζώα, όπως ο ιός της πολιομυελίτιδας και της ηπατίτιδας A. Η πρόσφυση αυτών των ιών στα κύτταρα γίνεται με τη βοήθεια μιας πρωτεΐνης (VP1) της οποίας η βιοσύνθεση είναι κωδικοποιημένη στο γονιδίωμα του ιού. Το γονιδίωμα περιέχει επίσης την πληροφορία για τη βιοσύνθεση ενός μεγάλου πεπτιδίου που συμμετέχει με την παραγωγή μιας RNA πολυμεράσης και άλλων πρωτεΐνών στην αναδόμηση του ιοσωματίου. Αξίζει να σημειωθεί πως οι κορωνοϊοί, που αφορούν και κεντρικό αντικείμενο αυτής της μελέτης, ανήκουν σε αυτήν την κατηγορία, όπως και ο ιός της γρίπης και ο ιός Ebola. Παρακάτω φαίνεται μια τυπική δομή πικορναϊού.



Εικόνα 3. Δομή πικορναϊού. Πηγή: <https://viralzone.expasy.org/>

- Ρετροϊοί: αποτελούνται από διπλοειδές ssRNA και χρησιμοποιούν το ένζυμο 'αντίστροφη μεταγραφάση' για την παραγωγή DNA. Σε αυτήν τη κατηγορία ανήκουν και οι HIVs (human immunodeficiency virus). Το DNA που παράγεται με τη διαδικασία της αντίστροφης μεταγραφής περιλαμβάνει την πληροφορία για τη βιοσύνθεση των κατάλληλων συστατικών του ιοσωματίου. Παρακάτω φαίνεται μια χαρακτηριστική δομή ρετροϊού.

Structure of a retrovirus



courtesy www.andrew.cmu.edu

Εικόνα 4. Δομή ρετροϊού. Πηγή: Asamoah, G. (2018). *HIV/AIDS as a developmental problem in Cameroon: Issues, impacts & way forward.*

1.1.5.2 Ιοί φυτών

Ιός μωσαϊκωσης του καπνού: όπως αναφέρθηκε και νωρίτερα, ο ιός αυτός είναι από τους πιο μελετημένους ιούς των φυτών που προσβάλλει εκτός από τον καπνό και άλλα φυτά. Γενικότερα, οι ιοί που προσβάλλουν φυτά είναι συνήθως ss(+)RNA ιοί. Το ss(+)RNA έχει την ικανότητα να μεταφράζεται στα ριβοσώματα του ξενιστή σε πρωτεΐνες χρήσιμες για την αναδόμηση του καψιδίου του ιού. Η προσβολή του ξενιστή από αυτόν τον ιό έχει ως

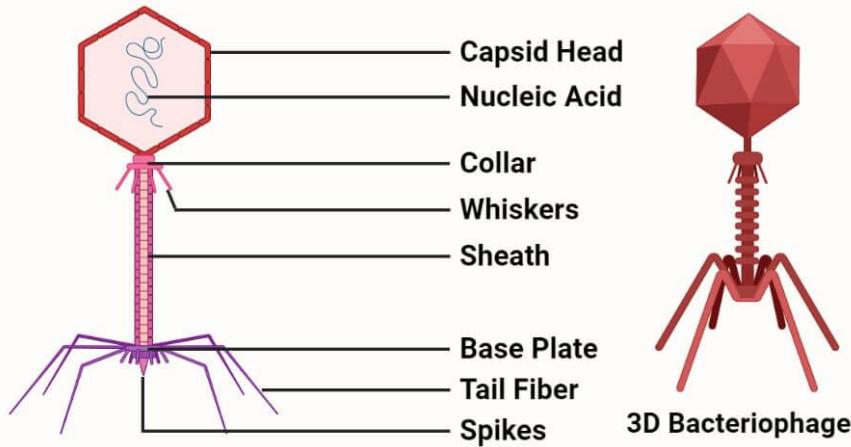
αποτέλεσμα την καταστροφή των χλωροπλαστών των κυττάρων του καθώς και των φύλλων τα οποία καθίστανται χλωρωτικά. Παρακάτω φαίνεται χαρακτηριστική εικόνα ενός προσβεβλημένου με αυτόν τον ιό φυτού.



Εικόνα 5. Χαρακτηριστική εικόνα φυτού που έχει προσβληθεί από τον ιό του μωσαϊκού του καπνού. Πηγή: <https://www.planetnatural.com/>

1.1.5.3 Ιοί βακτηρίων (βακτηριοφάγοι)

Μέσα στο βακτηριακό κύτταρο οι βακτηριοφάγοι χρησιμοποιούν τα βιομόρια και τις δομές του ξενιστή για να αναπαραχθούν αφού πρώτα προκαλούν λύση των βακτηριακών κυττάρων. Τα κύτταρα στην πορεία απελευθερώνουν πολυάριθμα ιοσωμάτια φάγων που μολύνουν νέα βακτηριακά κύτταρα. Οι φάγοι που φονεύουν τα κύτταρα καλούνται 'λυτικοί' ενώ υπάρχουν και φάγοι που δημιουργούν 'λυσιγόνο' σχέση με τα κύτταρα των ξενιστών, δηλαδή μια σχέση αρμονικής συνύπαρξης. Σε αυτή τη σχέση το γονιδίωμα του ιού ενσωματώνεται συνήθως στο χρωμόσωμα του βακτηρίου και επομένως πολλαπλασιάζεται συγχρόνως. Παρακάτω απεικονίζεται μια τυπική δομή ενός βακτηριοφάγου:



Εικόνα 6. Τυπική δομή βακτηριοφάγου. Πηγή: <https://microbenotes.com/>

1.2 ΚΟΡΩΝΟΪΟΙ

1.2.1 Ιστορική αναδρομή

Οι πρώτοι κορονοϊοί του ανθρώπου χαρακτηρίστηκαν στα μέσα της δεκαετίας του 1960. Συγκεκριμένα, το πρώτο στέλεχος περιεγράφηκε από τους Tyrrell και Bynoe το 1965 (Tyrrell & Bynoe, 1965) και στην πορεία ακολούθησαν άλλα ανθρώπινα στελέχη όπως το στέλεχος HCoV-229E, το οποίο απομονώθηκε από την αναπνευστική οδό από τους Hamre και Procknow το 1966 (Hamre & Procknow, 1966), καθώς και το στέλεχος HCoV-OC43 που απομονώθηκε από καλλιέργειες οργάνων από τους McIntosh et. al το 1967 (McIntosh et al., 1967).

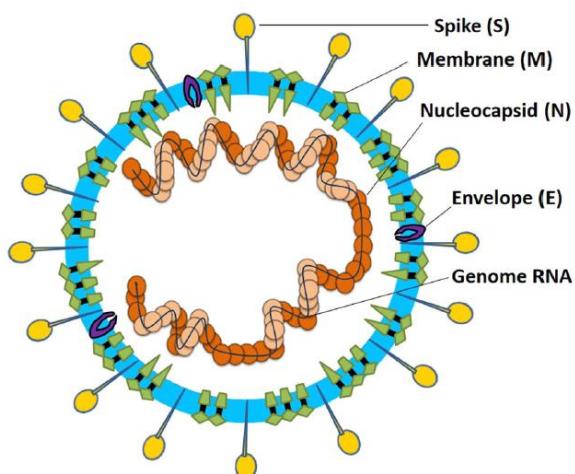
Μετά τις πρώτες ανακαλύψεις ενός ορισμένου αριθμού στελεχών κορονοϊών, το 1968 έγινε μια αρχική ανεπίσημη προσπάθεια ένταξης τους σε μια διακριτή οικογένεια από τους Tyrrell et. al, την οικογένεια *Coronaviridae* (Tyrell, 1968). Αργότερα το 1975 πραγματοποιήθηκε μια περισσότερο εμπεριστατωμένη μελέτη ως προς την ταξινόμηση αυτών των ιών και την τελική δημιουργία της οικογένειας *Coronaviridae* (Tyrrell et al., 1975).

Με τα δεδομένα και τη γνώση εκείνης της περιόδου, η συγκεκριμένη οικογένεια ιών ήταν συνδεδεμένη μόνο με μερικές ασθένειες σε σχέση με όσα γνωρίζουμε σήμερα. Συγκεκριμένα, ήταν συνδεδεμένη με τη λοιμώδη βρογχίτιδα στα κοτόπουλα, το κοινό κρυολόγημα στον άνθρωπο, την ηπατίτιδα και την εγκεφαλίτιδα στα ποντίκια, τη γαστρεντερίτιδα και την εγκεφαλίτιδα στους χοίρους, τις πνευμονικές μολύνσεις στα ποντίκια και τη σιελοδακρυοαδενίτιδα στα ποντίκια (Tyrrell et al., 1975).

1.2.2 Ταξινομική και γενικά χαρακτηριστικά κορονοϊών

Οι κορονοϊοί ανήκουν στην υποοικογένεια *Coronavirinae* της οικογένειας *Coronaviridae*, τάξη *Nidovirales* και προκαλούν αναπνευστικές ασθένειες, πεπτικές ασθένειες και ασθένειες του νευρικού συστήματος στον άνθρωπο και σε πολλά άλλα ζώα. Σύμφωνα με το ICTV (The International Committee for Taxonomy of Viruses) οι κορονοϊοί ταξινομούνται περαιτέρω σε τέσσερα γένη: Alpha-, Beta-, Gamma- και Deltacoronaviruses (McBride & Fielding, 2012).

Οι κορονοϊοί αποτελούνται από καψίδιο και ενιαίο θετικής πολικότητας μονόκλωνo RNA γονιδίωμα με μέγεθος που ποικίλει από 26 μέχρι 32 kb. Μάλιστα, αυτό είναι και το μεγαλύτερο γνωστό ιικό γονιδίωμα. Το ισωμάτιο αποτελείται από ένα νουκλεοκαψίδιο που αποτελείται από γενωμικό RNA και μια φωσφορυλιωμένη πρωτεΐνη (N) του νουκλεοκαψιδίου που βρίσκεται ανάμεσα από τις διπλοστιβάδες φωσφολιπιδίων και καλύπτεται από δύο ειδών πρωτεΐνες: τις «αγκαθωτές» γλυκοπρωτεΐνες (spike proteins) (S) που συναντώνται σε όλους τους κορονοϊούς και την πρωτεΐνη αιμοσυγκολλητίνης - εστεράσης (HE) που συναντάται σε κάποιους. Η πρωτεΐνη της μεμβράνης (M) και η πρωτεΐνη του φακέλου βρίσκονται μεταξύ των πρωτεΐνων S στον πρωτεϊνικό φάκελο. Το όνομα των κορονοϊών δόθηκε με βάση τη χαρακτηριστική εμφάνιση τους που προσομοιάζει μια κορώνα (G. Li et al., 2020; Lu et al., 2020). Παρακάτω φαίνεται μια εικόνα με τη χαρακτηριστική δομή του κορονοϊού.



Εικόνα 7. Τυπική δομή κορονοϊού. Πηγή: Li, G., Fan, Y., Lai, Y., Han, T., Li, Z., Zhou, P., . . . Wu, J. (2020). *Coronavirus infections and immune responses*. *J Med Virol*, 92(4), 424-432. doi:10.1002/jmv.25685

1.2.3 Αλληλεπίδραση ιού – ξενιστή

Οι κορωνοϊοί μπορούν να προσβάλλουν μια ποικιλία ξενιστών, συμπεριλαμβανομένων των πτηνών, χοίρων και ανθρώπων (Lim et al., 2016). Η μόλυνση από τον κορωνοϊό ξεκινά με την προσκόλληση σε συγκεκριμένους κυτταρικούς υποδοχείς του ξενιστή μέσω της πρωτεΐνης της ακίδας (S). Ο υποδοχέας του ξενιστή είναι καθοριστικός παράγοντας της παθογένειας, του τροπισμού των ιστών και του εύρους των ξενιστών του ιού. Η πρωτεΐνη S αποτελείται από δύο τομείς: S1 και S2. Η αλληλεπίδραση μεταξύ του τομέα S1 και του συγγενούς του υποδοχέα πυροδοτεί μια διαμορφωτική αλλαγή στην πρωτεΐνη S, η οποία στη συνέχεια προάγει τη σύντηξη μεμβρανών μεταξύ της ικής και της κυτταρικής μεμβράνης μέσω της περιοχής S2. Σήμερα, οι κύριοι υποδοχείς των κυττάρων-ξενιστών που χρησιμοποιούνται από όλους τους HCoV είναι γνωστοί ως: η αμινοπεπτιδάση N από τον HCoV-229E (Yeager et al., 1992), το ένζυμο μετατροπεας της αγγειοτενσίνης 2 (ACE2) από τον SARS-CoV (W. Li et al., 2003) και τον HCoV-NL63 (W. Li et al., 2007, p. 63; K. Wu et al., 2009), η διπεπτιδυλική πεπτιδάση 4 (DPP4) από τον MERS-CoV (van Doremalen et al., 2014) και το 9-O-ακετυλιωμένο σιαλικό οξύ από τον HCoV-OC43 και τον HCoV-HKU1 (Butler et al., 2006; X. Huang et al., 2015).

Εκτός από τη συμβατική ενδοσωμική οδό, ορισμένοι κορωνοϊοί μπορεί επίσης να εισέλθουν το κύτταρο μέσω μη-ενδοσωμικής οδού ή συνδυασμού και των δύο (Zumla et al., 2016).

Μετά την απελευθέρωση και την αποκάλυψη του ικού νουκλεοκαψιδίου στο κυτταρόπλασμα, αρχίζει η αντιγραφή του κορωνοϊού με τη μετάφραση των ORF1a και 1b σε πολυπρωτεΐνες pp1a (4382 αμινοξέα) και pp1ab (7073 αμινοξέα οξέα). Η περιοχή μετά το ORF1b μεταφράζεται μέσω ριβοσωμικού μηχανισμού μετατόπισης πλαισίου, στο το οποίο ένα μεταφραστικό ριβόσωμα μετατοπίζει ένα νουκλεοτίδιο προς την -1 κατεύθυνση, από το πλαισίο ανάγνωσης ORF1a στο πλαισίο ανάγνωσης ORF1b. Αυτή η επανατοποθέτηση ενεργοποιείται από δύο στοιχεία RNA - μια 5'-UUUAAAAC-3' επτανουκλεοτιδική «ολισθηρή» αλληλουχία και μια δομή «ψευδοκόμπων» (pseudoknot) RNA. Στη συνέχεια, οι πολυπρωτεΐνες pp1a και το pp1ab διασπώνται σε τουλάχιστον 15 nsp, τα οποία συναρμολογούνται και σχηματίζουν το σύμπλεγμα αντιγραφής-μεταγραφής. Με τη συναρμολόγηση της ρεπλικάσης-πολυμεράσης, το πλήρες μήκος του θετικού κλώνου του γονιδιωματικού RNA μεταγράφεται για να σχηματίσει ένα πλήρους μήκους πρότυπο αρνητικού κλώνου για τη σύνθεση νέων γονιδιωματικών RNAs καθώς και επικαλυπτόμενα υπογονιδιωματικά πρότυπα αρνητικού κλώνου. Αυτά τα υπογονιδιωματικά mRNAs έπειτα μεταγράφονται και μεταφράζονται για την παραγωγή των δομικών και βοηθητικών πρωτεΐνων. Αρκετά μέλη της οικογένειας των ετερόλογων πυρηνικών

ριβονουκλεοπρωτεΐνών (hnRNA, hnRNPA1, PTB, SYN-CRYP) έχουν βρεθεί ότι είναι απαραίτητα για την αποτελεσματική αντιγραφή του RNA (Luo et al., 2005). Άλλες πρωτεΐνες δέσμευσης RNA έχουν επίσης προταθεί ότι παίζουν ρόλο στην αντιγραφή του κορωνοϊού, όπως η m-ακονιτάση και η πρωτεΐνη δέσμευσης poly-A (PABP), DDX1, PCBP1/2, το δάκτυλο φυεδεργύρου (zinc finger) τύπου CCHC και μοτίβο δέσμευσης RNA 1 (MADP1) (Nanda & Leibowitz, 2001; Tan et al., 2012; C.-H. Wu et al., 2014).

1.2.4 SARS - CoV

Οι κορονοϊοί δεν είχαν συσχετιστεί με το σύνδρομο SARS (Severe Acute Respiratory Syndrome) που γνωρίζουμε σήμερα μέχρι και το 2003, όταν το Φεβρουάριο εκείνης της χρονιάς ξέσπασε επιδημία στην πόλη Guangdong της Κίνας (WHO, 2003). Μάλιστα, μέχρι εκείνη την περίοδο, η συγκεκριμένη ομάδα ιών δε θεωρούνταν ιδιαίτερα παθογόνα για τους ανθρώπους, καθώς οι κορονοϊοί που επικρατούσαν προκαλούσαν μόνο ήπια συμπτώματα σε ανοσοεπαρκείς ανθρώπους, παρόλα αυτά η έξαρση της επιδημίας SARS κατέστησε σοβαρές τις επιπτώσεις του κορονοϊού που σχετιζόταν με αυτή (Cui et al., 2019).

Η διαπίστωση πως ένα νέο στέλεχος κορονοϊού, που δεν ήταν γνωστό μέχρι τότε, έπαιξε αιτιολογικό ρόλο στην επιδημία, έγινε λίγο μετά το ξέσπασμα της επιδημίας όταν μελετήθηκαν εκτεταμένα ασθενείς που παρουσίαζαν συμπτώματα του SARS. Για την ταυτοποίηση του ιού που σχετιζόταν με το σύνδρομο SARS πραγματοποιήθηκαν πολλαπλά πειράματα PCR, ανοσολογικά, φυλογενετικής ανάλυσης καθώς και επιδημιολογικές μελέτες από διαφορετικά εργαστήρια (Drosten et al., 2003; Ksiazek et al., 2003; Peiris et al., 2003; Zhong et al., 2003).

Συγκεκριμένα, για τον πολλαπλασιασμό των δειγμάτων των ασθενών χρησιμοποιήθηκαν εκκινητές για πολλά είδη ιών, μέσα στα οποία συμπεριλαμβάνονταν και τα τότε γνωστά στελέχη κορονοϊών, ενώ μετά πραγματοποιήθηκε φυλογενετική ανάλυση με το εργαλείο BLAST για την ανίχνευση ομολογιών. Τα αποτελέσματα συνέβαλαν στην ταυτοποίηση ενός νέου στελέχους κορονοϊού σε ασθενείς με SARS το οποίο παρουσίαζε υψηλή μεταδοτικότητα καθώς και υψηλό χρόνο επώασης στον οργανισμό. Παράλληλα, τα πειράματα έδειξαν πως η αναπνευστική οδός αποτελούσε τη βασική δίοδο μετάδοσης (Drosten et al., 2003).

Συνολικά, ο συγκεκριμένος ιός έπληξε πάνω από 8.000 ανθρώπους και εξαπλώθηκε σε 29 χώρες, ενώ ο βαθμός θνησιμότητας ανήλθε στο 10% (Cheng et al., 2007; Lim et al., 2016; Ludwig & Zarbock, 2020).

1.2.5 SARS – CoV – 2

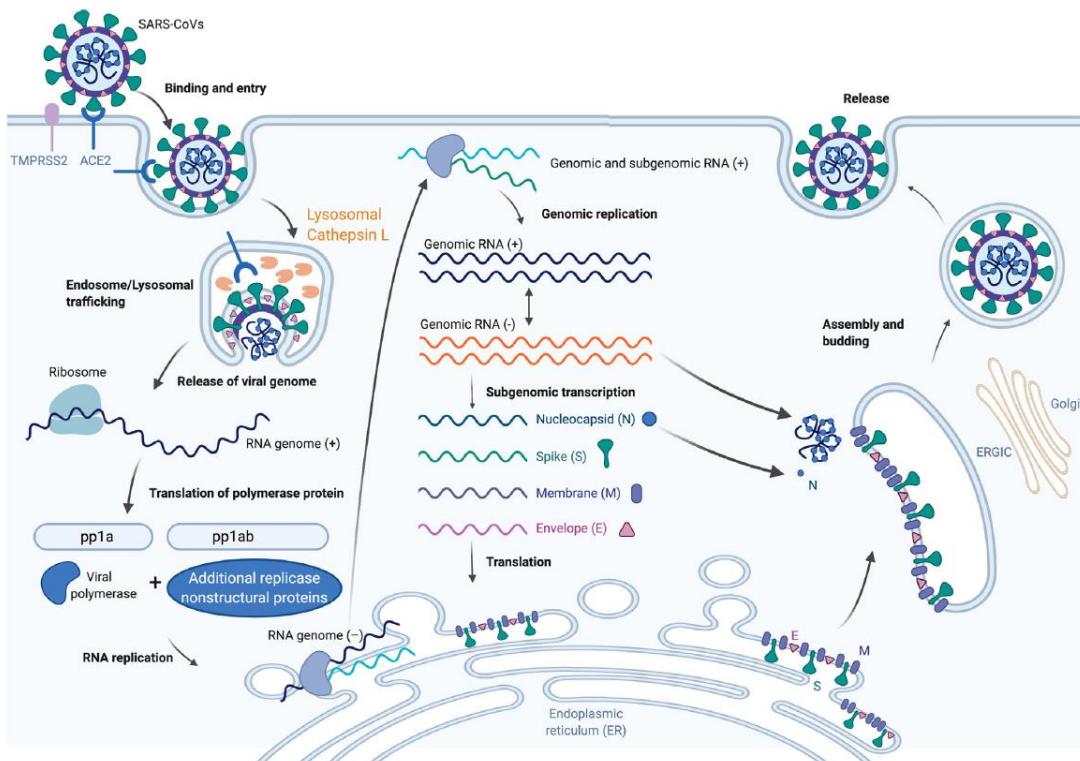
Μέχρι και πριν το Δεκέμβριο του 2019, μόλις έξι στελέχη κορονοϊών ήταν γνωστά ότι προσέβαλαν τον άνθρωπο και προκαλούσαν αναπνευστικά σύνδρομα. Τα ενδημικά στελέχη HCoV-229E, HCoV-OC43, HCoV-NL63 και HKU1 προκαλούν πιο ήπια συμπτώματα και έχουν σπάνια σοβαρές επιπτώσεις σε βρέφη, παιδιά και ηλικιωμένους. Παράλληλα, έχουν καταγραφεί και περιπτώσεις ασυμπτωματικών μολύνσεων (Ludwig & Zarbock, 2020; McIntosh & Peiris, 2009). Περισσότερο επικίνδυνα βέβαια είναι τα στελέχη SARS-CoV και MERS-CoV που προκαλούν σοβαρότερα προβλήματα υγείας στον άνθρωπο.

Στα τέλη του Δεκεμβρίου του 2019, έγινε αναφορά πολλών περιπτώσεων πνευμονιών άγνωστης προέλευσης και αργότερα τον Ιανουάριο του 2020 ανακοινώθηκε πως αυτά τα περιστατικά ήταν αποτέλεσμα της μόλυνσης με ένα νέο ιικό στέλεχος κορονοϊού. Το στέλεχος αυτό αργότερα προσδιορίστηκε ως Severe Acute Respiratory Syndrome coronavirus 2 (SARS-CoV-2) και ορίστηκε ως ο αιτιολογικός παράγοντας αυτού του είδους πνευμονίας άγνωστης προέλευσης που ονομάστηκε Coronavirus Disease 2019 (COVID-19) (Ludwig & Zarbock, 2020).

Η αλληλουχία του γονιδιώματος του SARS-CoV-2 είναι κατά ~80% ταυτόσημη με την αλληλουχία του SARS-CoV και κατά ~50% με την αλληλουχία του MERS-CoV. Το γονιδίωμα του αποτελείται από 14 ανοιχτά πλαίσια ανάγνωσης (ORFs), δύο τρίτα εκ των οποίων κωδικοποιούν 16 μη-δομικές πρωτεΐνες (nsp 1-16) που συνιστούν το σύμπλεγμα ρεπλικάσης. Το υπόλοιπο ένα τρίτο κωδικοποιεί 9 βιοθητικές πρωτεΐνες (ORF) και τέσσερις δομικές: 1. Ακίδας (spike (S)), 2. Φακέλου (envelope (E)), 3. Μεμβράνης (membrane (M)), 4. Νουκλεοκαψιδίου (nucleocapsid (N)), από τις οποίες η πρωτεΐνη ακίδας είναι αυτή που μεσολαβεί για την είσοδο του SARS-CoV στα κύτταρα του ξενιστή (Perlman & Netland, 2009). Το γονίδιο S του SARS-CoV-2 έχει μεγάλη μεταβλητότητα σε σχέση με τον SARS-CoV έχοντας πάνω από 75% νουκλεοτιδική ομοιότητα. Η πρωτεΐνη ακίδας διαθέτει μια περιοχή δέσμευσης υποδοχέα – receptor binding domain (RBD), που πυροδοτεί την άμεση επαφή με ένα κυτταρικό υποδοχέα, το ένζυμο ACE2 (angiotensin-converting enzyme 2) και μια S1/S2 πολυβασική περιοχή διάσπασης που διασπάται πρωτεολυτικά από την κυτταρική καθεψίνη L και τη διαμεμβρανική πρωτεάση σερίνη 2 – transmembrane protease serine 2 (TMPRSS2). Η TMPRSS2 διευκολύνει την είσοδο του ιού στην επιφάνεια της πλασματικής μεμβράνης, ενώ η καθεψίνη L ενεργοποιεί την πρωτεΐνη ακίδας του SARS-CoV-2 στα ενδοσώματα και αντισταθμίζει την είσοδο σε κύτταρα που δεν έχουν TMPRSS2 (Hoffmann et al., 2020). Όταν το γονιδίωμα απελευθερωθεί στο κυτοσόλιο του ξενιστή, τα πλαίσια ανάγνωσης ORF1a και ORF1b μεταφράζονται σε ιικές πρωτεΐνες ρεπλικάσης που διασπώνται σε ξεχωριστές μη-

δομικές πρωτεΐνες (nsp) που με τη σειρά τους σχηματίζουν την RNA-εξαρτημένη RNA πολυμεράση (RNA-dependent RNA polymerase) (το nsp12 προέρχεται από το πλαίσιο ανάγνωσης ORF1b) (Perlman & Netland, 2009). Με αυτόν τον τρόπο, τα συστατικά της ρεπλικάσης αναδιαρθρώνουν το ενδοπλασματικό δίκτυο σε διπλο-μεμβρανικά κυστίδια (double-membrane vesicles (DMVs)) που διευκολύνουν τον ιικό πολλαπλασιασμό των γενωμικών και υπογενωμικών RNAs (sgRNA). Το τελευταίο μεταφράζεται σε δομικές πρωτεΐνες του ιού και διευκολύνουν τη δημιουργία ιικών σωματιδίων (Harrison et al., 2020).

Παρακάτω φαίνεται σχηματικά ο κύκλος ζωής του SARS-CoV-2:

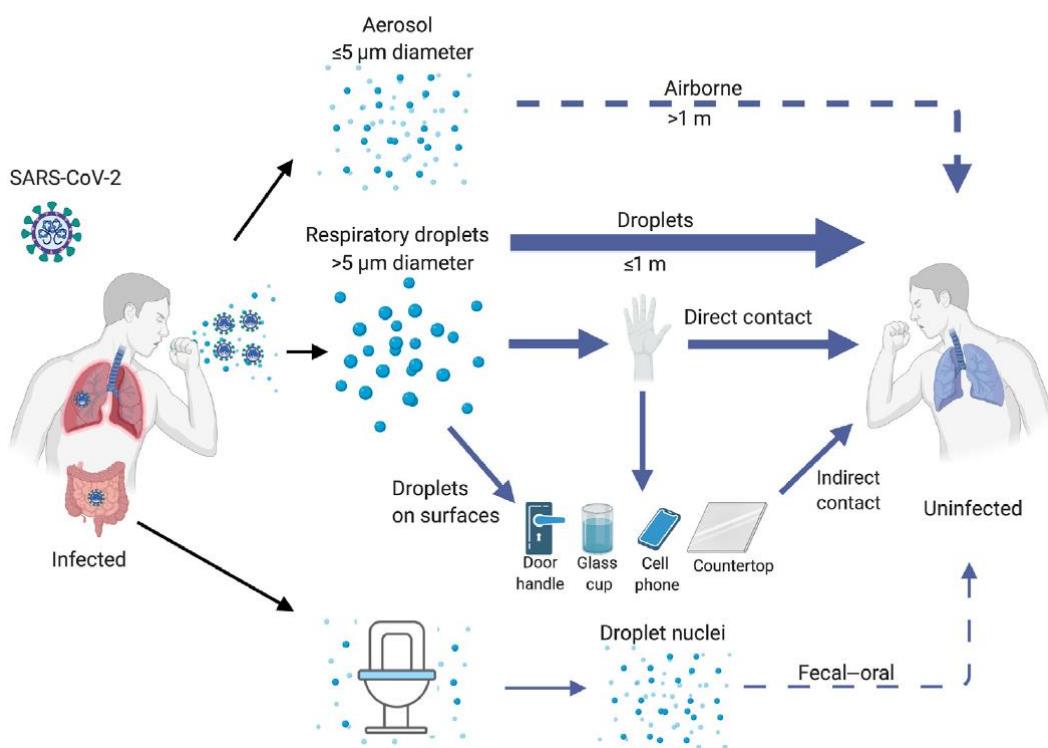


Εικόνα 8. Ο κύκλος ζωής του SARS-CoV-2, Πηγή: (Harrison et al., 2020)

Οι κορονοϊοί του ανθρώπου μεταδίδονται κυρίως μέσω αναπνευστικών σταγονιδίων, αλλά κατά τη διάρκεια της επιδημίας του SARS καταγράφηκαν και άλλοι τρόποι μετάδοσης, όπως η άμεση επαφή με μολυσμένες επιφάνειες, το αερόλυμα και η κοπρανοστοματική οδός. Αναφορές από ασθενείς με βήχα, εικόνα θολής υάλου (ground-glass) στους πνεύμονες και συμπτώματα που εξελίσσονται σε σοβαρή πνευμονία, προδίδουν τη διάδοση του SARS-CoV-2 μέσω της αναπνευστικής οδού (C. Huang et al., 2020; Peiris et al., 2003; Zhou et al., 2020). Η άμεση διάδοση των αναπνευστικών σταγονιδίων πυροδοτείται από τον παραγωγικό πολλαπλασιασμό του SARS-CoV-2 στην ανώτερη και κατώτερη αναπνευστική οδό, καθώς και τον αυξανόμενο αριθμό αναφορών που υποδεικνύουν εξάπλωση μεταξύ των ανθρώπων

με ενεργό βήχα που πραγματοποιούν στενές επαφές (Chan et al., 2020; Q. Li et al., 2020; The COVID-19 Investigation Team, 2020; Wang et al., 2020). Μέχρι πρόσφατα, ο αριθμός αναπαραγωγής (reproduction number (R_0)) ήταν ~2.2 με ρυθμό διπλασιασμού τις 5 μέρες (Ferretti et al., 2020; Q. Li et al., 2020). Επιπλέον, υπάρχουν στοιχεία για μη-συμπτωματική/προ-συμπτωματική εξάπλωση του SARS-CoV-2 που έρχονται σε αντίθεση με τη δυναμική διάδοσης του SARS-CoV (Arons et al., 2020). Αυτό το εύρημα υπογραμμίζει την ικανότητα του SARS-CoV-2 να εποικίζει και να πολλαπλασιάζεται στην περιοχή του λαιμού κατά τη διάρκεια της αρχικής μόλυνσης (Pan et al., 2020; Wölfel et al., 2020; Zou et al., 2020).

Παρακάτω απεικονίζονται οι προτεινόμενες οδοί διάδοσης του SARS-CoV-2.



Εικόνα 9. Οδοί διάδοσης του SARS-CoV-2, Πηγή: (Harrison et al., 2020)

Για τον SARS-CoV-2 έχουν προταθεί διάφορα μοντέλα διάδοσης συμπεριλαμβανομένου του αερολύματος, της μόλυνσης από επιφάνειες και την κοπρανοστοματική οδό, όπως φαίνεται και στην παραπάνω εικόνα, και η σχετική σημασία τους είναι υπό διερεύνηση. Η διάδοση μέσω του αερολύματος (εξάπλωση > 1 m) προτάθηκε πρώτη φόρα στην περύπτωση των Amoy Gardens κατά τη διάρκεια της επιδημίας SARS, παρόλα αυτά η ασυνοχή αυτών των ευρημάτων υποδεικνύει πως ο SARS-CoV πρόκειται για μια ευκαιριακή αερομεταφερόμενη λοιμωξη (Tomlinson & Cockram, 2003; I. T. S. Yu et al., 2004). Παράλληλα, δεν έχουν απομονωθεί μολυσματικά ιοσωμάτια παρόλο που ίικό RNA ήταν ανιχνεύσιμο στον αέρα των

νοσοκομειακών θαλάμων της COVID-19 (Liu et al., 2020). Η παραγωγή πειραματικών αερολυμάτων που φέρουν SARS-CoV-2 (συγκρίσιμα με αυτά που μπορεί να παράγονται από τον άνθρωπο), υποστηρίζουν τη θεωρία της αερομεταφερόμενης διάδοσης, αλλά τα αεροδυναμικά χαρακτηριστικά του SARS-CoV-2 κατά τη διάρκεια μιας φυσικής πορείας μόλυνσης είναι ακόμα υπό διερεύνηση (van Doremalen et al., 2014). Παρόλα αυτά, η εναπόθεση αερολυμάτων επιβαρυμένων με ιούς μπορεί να μολύνει αντικείμενα (π.χ. μικροβιοφόρες ουσίες (*fomites*)) και να συμβάλλει κατά συνέπεια στη μετάδοση στον άνθρωπο (Liu et al., 2020; Ong et al., 2020). Τέλος, η κοπρανοστοματική διάδοση έχει επίσης θεωρηθεί ως πιθανή οδός εξάπλωσης στον άνθρωπο, αλλά παραμένει ένα αίνιγμα παρά τις ενδείξεις αερολυμάτων επιβαρυμένων με RNA που έχουν βρεθεί κοντά σε λεκάνες τουαλέτας, μαζί με ανιχνεύσιμο RNA του SARS-CoV-2 σε επιχρίσματα από το ορόφο κατά τη διάρκεια της πρώιμης επιδημίας της COVID-19 στην Κίνα (Liu et al., 2020; Xiao et al., 2020; Xu et al., 2020).

1.2.5.1 Παθογένεση και κλινική εικόνα του SARS-CoV-2

Γενικά, οι κορονοϊοί του κοινού κρυολογήματος τείνουν να προκαλούν ήπια συμπτώματα λοίμωξης του ανώτερου αναπνευστικού συστήματος και κάποιες φορές εμπλέκεται και το γαστρεντερικό σύστημα. Αντίθετα, η μόλυνση με εξαιρετικά παθογόνους κορονοϊούς, συμπεριλαμβανομένου του SARS-CoV-2, προκαλεί σοβαρά συμπτώματα που μοιάζουν με γρίπη και μπορεί να εξελιχθούν σε οξεία αναπνευστική δυσχέρεια (acute respiratory distress (ARDS)), πνευμονία, νεφρική ανεπάρκεια, μέχρι και θάνατο (Guery et al., 2013; Ksiazek et al., 2003; Q. Li et al., 2020; Wang et al., 2020). Τα πιο κοινά συμπτώματα είναι ο πυρετός, ο βήχας και η δύσπνοια, που αντιτροσωπεύουν το 83%, το 82% και το 31% των ασθενών με COVID-19 ($N = 99$), αντίστοιχα, σε μία επιδημιολογική μελέτη (N. Chen et al., 2020). Η περίοδος επώασης στον COVID-19 είναι ταχεία: ~5–6 ημέρες έναντι 2–11 ημερών σε λοιμώξεις από SARS-CoV (Chan et al., 2020; Q. Li et al., 2020; Su et al., 2016). Καθώς η πανδημία εξελίσσεται, έχει γίνει ολοένα και πιο σαφές ότι η COVID-19 δεν περιλαμβάνει μόνο ταχέως αναπνευστικές/ γαστρεντερικές παθήσεις, αλλά μπορεί επίσης να προκαλέσει μακροχρόνιες επιπτώσεις, όπως για παράδειγμα φλεγμονή του μυοκαρδίου (Puntmann et al., 2020). Επιπλέον, η σοβαρή COVID-19 δεν περιορίζεται στον ηλικιωμένο πληθυσμό όπως αρχικά είχε αναφερθεί, αλλά παιδιά και νεαροί ενήλικες κινδυνεύουν επίσης (Chao et al., 2020). Από διαγνωστικής απόψεως, η COVID-19 παρουσιάζει ορισμένους «σήμα κατατεθέν» εργαστηριακούς και ακτινολογικούς δείκτες, που μπορεί να είναι χρήσιμοι στην αξιολόγηση της εξέλιξης της νόσου. Συνολικά, η COVID-19 εμφανίζεται αρχικά με συμπτώματα που προσομοιάζουν απλή γρίπη και μπορεί αργότερα να εξελιχθεί σε

απειλητική για τη ζωή συστηματική φλεγμονή και πολυοργανική δυσλειτουργία (Harrison et al., 2020).

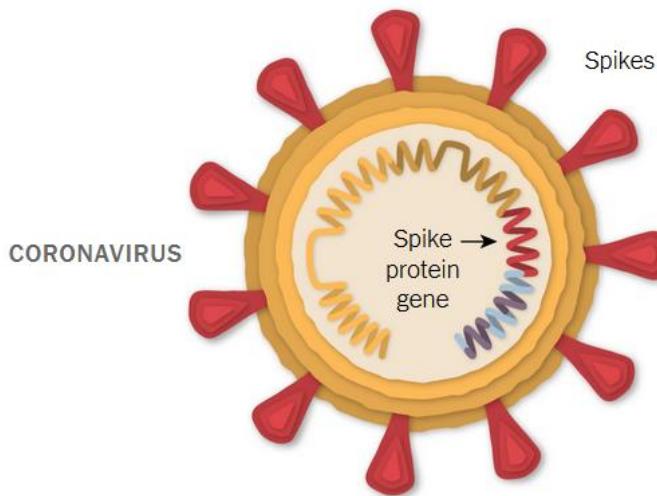
1.3 ΑΣΘΕΝΕΙΑ CoVID-19

Τα συμπτώματα της CoVID-19 είναι μη ειδικά και η εικόνα της ασθένειας μπορεί να κυμαίνεται από καθόλου συμπτώματα (ασυμπτωματική) έως σοβαρή πνευμονία και θάνατος. Μια μελέτη 41 ασθενών (C. Huang et al., 2020) που είχαν αρχικά διαγνωστεί με το ξέσπασμα, κατέληξε στο ότι τα πιο κοινά συμπτώματα ήταν ο πυρετός (98%), ο βήχας (76%), η μυαλγία ή κόπωση (44%) και τα άτυπα συμπτώματα περιλάμβαναν πτύελο (28%), πονοκέφαλο (8%), αιμόπτυση (5%) και διάρροια (3%). Περίπου οι μισοί ασθενείς είχαν δύσπνοια (η διάμεσος από την έναρξη έως τη δύσπνοια ήταν 8 ημέρες). Λεμφοκυτταροπενία παρατηρήθηκε στο 63% των ασθενών, ενώ όλοι οι ασθενείς είχαν πνευμονία. Οι επιπλοκές περιλάμβαναν σύνδρομο οξείας αναπνευστικής δυσχέρειας (29%), οξεία καρδιακή βλάβη (12%) και δευτερογενείς λοιμώξεις (10%), ενώ το 32% των ασθενών χρειάστηκε θεραπεία στη ΜΕΘ. Μια ανάλυση 1,099 επιβεβαιωμένων κρουσμάτων που διενήργησε η ομάδα του NanShan Zhong (Guan et al., 2020) διαπίστωσε ότι τα πιο κοινά συμπτώματα ήταν επίσης ο πυρετός (87,9%), ο βήχας (67,7%), η διάρροια (3,7%) και ο έμετος (5,0%). Το 25,2% των ασθενών είχαν τουλάχιστον ένα υποκείμενο νόσημα (όπως υπέρταση ή χρόνια αποφρακτική πνευμονοπάθεια). Λεμφοκυτταροπενία παρατηρήθηκε στο 82,1% των ασθενών. Κατά την εισαγωγή τους, το 50% των ασθενών παρουσίασε εικόνα θολής υάλου (ground-glass shadow) στην αξονική τομογραφία θώρακος. Μια αναδρομική μελέτη (Wang et al., 2020) με 138 νοσηλευόμενους ασθενείς που διεξάχθηκε από 1 έως 28 Ιανουαρίου, κατέληξε στο ότι οι ασθενείς που λάμβαναν θεραπεία στη ΜΕΘ ήταν μεγαλύτερης ηλικίας, πιο πιθανό να είχαν υποκείμενα νοσήματα, και πιο πιθανό να έχουν δύσπνοια, παράλληλα διάμεσος της διάρκειας παραμονής ήταν 10 ημέρες. Άλλες μελέτες δείχνουν ότι ασθενείς ηλικίας 60 ετών και άνω διατρέχουν υψηλότερο κίνδυνο από τα παιδιά που μπορεί να είναι λιγότερο πιθανό να μολυνθούν ή, αν μολυνθούν, μπορεί να εμφανίσουν ηπιότερα συμπτώματα ή ακόμα και ασυμπτωματική λοίμωξη (Q. Li et al., 2020). Σύμφωνα με την ομάδα Epidemiology Working Group (The Novel Coronavirus Pneumonia Emergency Response Epidemiology Team, 2020) για συνολικά 72.314 ασθενείς, ανέφεραν ότι υπήρξαν 44.672 (61,8%) επιβεβαιωμένα κρούσματα και 889 ασυμπτωματικές περιπτώσεις (1,2%) στο σύνολο των ασθενών. Μεταξύ των επιβεβαιωμένων περιπτώσεων, οι περισσότερες ήταν ηλικίας 30-79 ετών (86,6%) και θεωρήθηκαν ήπιες/ ήπια πνευμονία (80,9%) (D. Wu et al., 2020).

1.4 ΕΜΒΟΛΙΑ ΚΑΤΑ ΤΗΣ COVID-19

Η τρέχουσα πανδημία της νόσου του κορωνοϊού (COVID-19), που προκλήθηκε όπως αναφέρθηκε από τον κορωνοϊό σοβαρού οξείου αναπνευστικού συνδρόμου (SARS-CoV-2), προκάλεσε αναπόφευκτα εξαιρετικά ταχεία παραγωγή εμβολίων κατά της νόσου. Σήμερα, μετράμε συνολικά 13.57 δισεκατομμύρια δόσεις εμβολίων σε περίπου 70.6% του συνολικού πληθυσμού, σύμφωνα με δεδομένα του *Our World Data*, ένα πρότζεκτ του Global Change Data Lab (Mathieu et al., 2020). Ο Ευρωπαϊκός Οργανισμός Φαρμάκων (ΕΟΦ) είχε εγκρίνει κατά τη διάρκεια συλλογής δεδομένων της συγκεκριμένης διπλωματικής εργασίας τέσσερα εμβόλια κατά της COVID-19, συγκεκριμένα, το mRNA εμβόλιο Tozinameran (των εταιρειών Pfizer-Biontech), το mRNA εμβόλιο CX-024414 (της εταιρείας Moderna), το φορέα αδενοϊού του χιμπατζή CHADOX1 NCOV-19 (της εταιρείας AstraZeneca) και τέλος τον αδενοϊκό φορέα τύπου 26, AD26.COV2.S (της εταιρείας Janssen) (Abbattista et al., 2021).

Ο ιός SARS-CoV-2, όπως αναφέρθηκε, είναι γεμάτος πρωτεΐνες που χρησιμοποιεί για να εισέλθει στα ανθρώπινα κύτταρα. Αυτές οι λεγόμενες πρωτεΐνες ακίδας αποτελούν τον κατάλληλο στόχο για πιθανά εμβόλια και θεραπείες.



Εικόνα 10. Δομή ιού SARS-CoV-2, Πηγή: <https://www.nytimes.com/interactive/2020/health/pfizer-biontech-covid-19-vaccine.html>

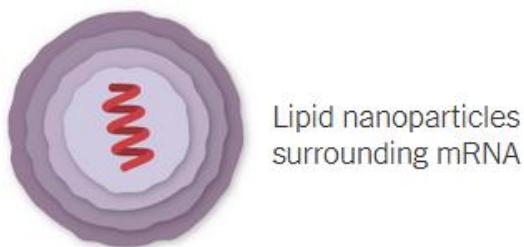
Παρακάτω παρουσιάζονται με περισσότερες λεπτομέρειες τα χαρακτηριστικά των συγκεκριμένων εμβολίων.

1.4.1 Εμβόλιο των Pfizer–BioNTech mRNA BNT162b2

Το εμβόλιο COVID-19 της Pfizer ήταν το εμβόλιο που αναπτύχθηκε ταχύτερα από όλα, αφού χρειάστηκε μόλις 7 μήνες μετά τη δοκιμή φάσης I/II που πραγματοποιήθηκε τον Μάιο του 2020,

προκειμένου ο FDA να επιτρέψει την επείγουσα χρήση του τον Δεκέμβριο του 2020 (Mulligan et al., 2021). Στις 11 Δεκεμβρίου 2020, ο FDA ενέκρινε χρήση έκτακτης ανάγκης του BNT162b28 και στις 23 Αυγούστου 2021, ο FDA των ΗΠΑ ενέκρινε το εμβόλιο Pfizer καθιστώντας το το πρώτο εγκεκριμένο εμβόλιο κατά της COVID-19 (FDA, 2021). Το εμβόλιο είναι επίσης γνωστό και ως BNT162b2, το γενικό όνομα tozinameran ή το εμπορικό όνομα Cominarty.

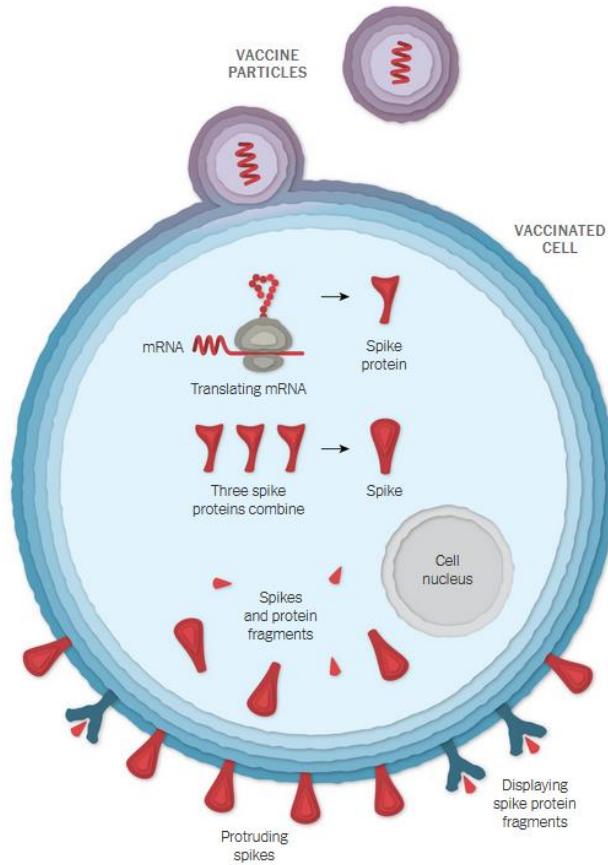
Το εμβόλιο της Pfizer (όπως και της Moderna) περιέχει ένα mRNA με τροποποιημένα με νουκλεοσίδια που κωδικοποιεί τη γλυκοπρωτεΐνη ακίδας του SARS-CoV-2 και χορηγείται σε νανοσωματίδια λιπιδίων για πιο αποτελεσματική χορήγηση στα κύτταρα ξενιστές .



Εικόνα 11. Μόριο mRNA μέσα σε νανοσωματίδιο λιπιδίου, Πηγή: <https://www.nytimes.com/interactive/2020/health/pfizer-biontech-covid-19-vaccine.html>

Το mRNA κωδικοποιεί ειδικά για το αντιγόνο S2-P, που αποτελείται από τη γλυκοπρωτεΐνη του SARS-CoV-2 με μια διαμεμβρανική άγκυρα. Οι αποκρίσεις των αντισωμάτων που δεσμεύονται στο S2-P χρησιμοποιήθηκαν ως μέθοδος αξιολόγησης της αποτελεσματικότητας του εμβολίου. Ο στόχος του εμβολίου είναι να προκαλέσει αποκρίσεις τόσο B- όσο και T-κυττάρων κατά της πρωτεΐνης ακίδας. Σύμφωνα με τα δημοσιευμένα δεδομένα, το εμβόλιο ήταν επιτυχές στην πρόκληση απόκρισης των αντισωμάτων τόσο σε πλήρους μήκους S2-P όσο και σε περιοχές δέσμευσης υποδοχέων (receptor-binding domains). Το ισχυρό σύστημα παροχής λιπιδίων-νανοσωματίδιων που χρησιμοποιείται από το εμβόλιο σε συνδυασμό με τη χρήση τροποποιημένων νουκλεοτιδίων που αποφεύγουν την πρώιμη ενεργοποίηση των γονιδίων που σχετίζονται με τις ιντερφερόνες, είναι μοναδικά χαρακτηριστικά που συμβάλλουν στην αποτελεσματικότητά του (Jackson et al., 2020). Το εμβόλιο mRNA παράγει παρατεταμένη έκφραση της πρωτεΐνης, επαγωγή ειδικών για αντιγόνο T-θυλακιωδών βιοηθητικών κυττάρων καθώς και ενεργοποίηση B-κυττάρων του βλαστικού κέντρου (Pardi et al., 2018).

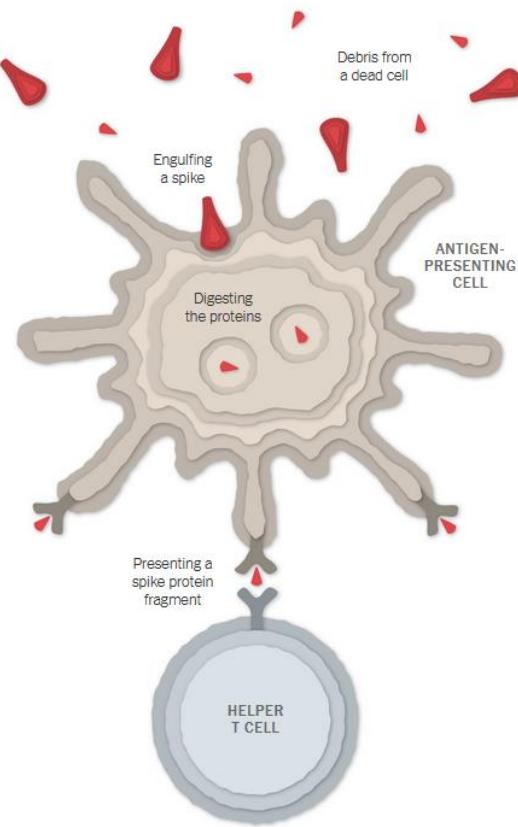
Παρακάτω απεικονίζεται η διαδικασία εισαγωγής των νανοσωματιδιών σε ένα ανθρώπινο κύτταρο:



Εικόνα 12. Διαδικασία εισαγωγής νανοσωματίδιου λιπιδίου σε ανθρώπινο κύτταρο, Πηγή: <https://www.nytimes.com/interactive/2020/health/pfizer-biontech-covid-19-vaccine.html>

Μετά την ένεση, τα σωματίδια του εμβολίου προσκρούουν στα ανθρώπινα κύτταρα και συγχωνεύονται σε αυτά, απελευθερώνοντας το mRNA που μεταφέρουν. Τα μόρια του κυττάρου διαβάζουν την αλληλουχία του και δημιουργούν τις πρωτεΐνες ακίδας. Το mRNA από το εμβόλιο καταστρέφεται τελικά από το κύτταρο, χωρίς να αφήνει μόνιμο ίχνος. Μερικές από τις πρωτεΐνες ακίδας σχηματίζουν «αιχμές» (spikes) που μεταναστεύουν στην επιφάνεια του κυττάρου και προεξέχουν τις άκρες τους. Τα εμβολιασμένα κύτταρα διασπούν επίσης ορισμένες από τις πρωτεΐνες σε θραύσματα, τα οποία και παρουσιάζουν στην επιφάνεια τους. Αυτές οι προεξέχουσες αιχμές και τα θραύσματα της πρωτεΐνης ακίδας μπορούν στη συνέχεια να αναγνωριστούν από το ανοσοποιητικό μας σύστημα.

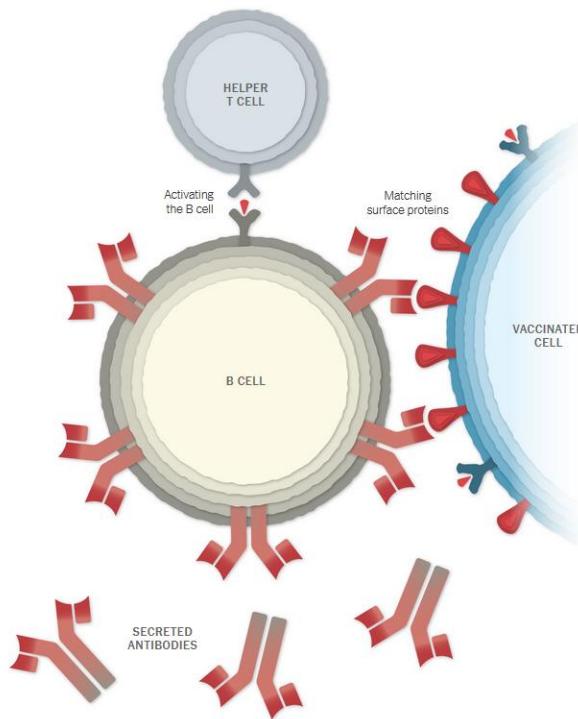
Στη συνέχεια απεικονίζεται η διαδικασία παρουσίασης των πρωτεϊνών ακίδας από τα αντιγονοπαρουσιαστικά κύτταρα του οργανισμού.



Εικόνα 13. Διαδικασία αντιγονοπαρουσίασης πρωτεΐνης ακίδας, Πηγή: <https://www.nytimes.com/interactive/2020/health/pfizer-biontech-covid-19-vaccine.html>

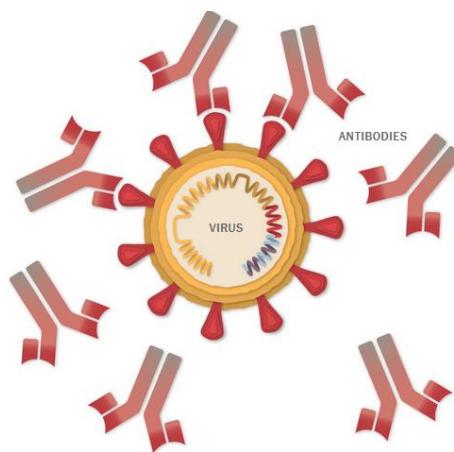
Όταν ένα εμβολιασμένο κύτταρο πεθάνει, τα υπολείμματα περιέχουν πολλές πρωτεΐνες ακίδας και θραύσματα πρωτεΐνης, τα οποία στη συνέχεια μπορούν να ληφθούν από έναν τύπο κυττάρου του ανοσοποιητικού που ονομάζεται αντιγονοπαρουσιαστικό κύτταρο. Το κύτταρο παρουσιάζει τα θραύσματα της πρωτεΐνης ακίδας στην επιφάνειά του. Όταν τα T-βοηθητικά κύτταρα ανιχνεύουν αυτά τα θραύσματα, τα βοηθητικά T-λεμφοκύτταρα μπορούν να προειδοποιήσουν και να βοηθήσουν στη συγκέντρωση άλλων κυττάρων του ανοσοποιητικού για την καταπολέμηση της λοίμωξης.

Έπειτα, παρουσιάζεται εικονικά η διαδικασία δημιουργίας αντισωμάτων:



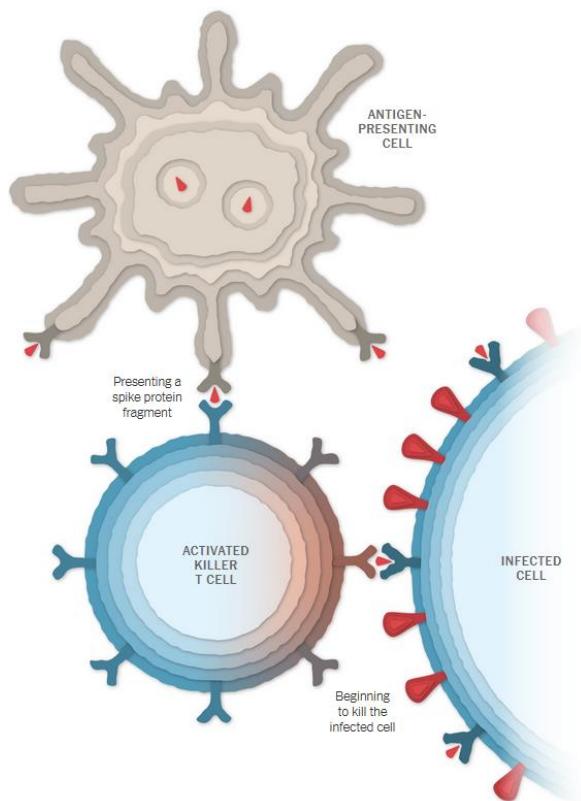
Εικόνα 14. Διαδικασία δημιουργίας αντισωμάτων, Πηγή: <https://www.nytimes.com/interactive/2020/health/pfizer-biontech-covid-19-vaccine.html>,

Τα B-κύτταρα, μπορεί να προσκρούσουν στις αιχμές του κορωνοϊού στην επιφάνεια των εμβολιασμένων κυττάρων ή σε θραύσματα πρωτεΐνης ακίδας που υπάρχουν στο χώρο ελεύθερα. Μερικά από τα B-κύτταρα μπορεί να είναι σε θέση να «κλειδώσουν» πάνω στις πρωτεΐνες ακίδας. Εάν αυτά τα B-κύτταρα ενεργοποιηθούν στη συνέχεια από τα βοηθητικά T-λεμφοκύτταρα, θα αρχίσουν να πολλαπλασιάζονται και να παράγουν αντισώματα που στοχεύουν την πρωτεΐνη ακίδας.



Εικόνα 15. Διαδικασία καταπολέμησης του ιού, Πηγή: <https://www.nytimes.com/interactive/2020/health/pfizer-biontech-covid-19-vaccine.html>

Τα αντισώματα μπορούν να δεσμευτούν στις αιχμές του κορωνοϊού, να σημάνουν τον ιό για καταστροφή και να αποτρέψουν τη μόλυνση εμποδίζοντας τις αιχμές να προσκολληθούν σε άλλα κύτταρα.



Εικόνα 16. Διαδικασία καταστροφής μολυσμένων κυττάρων, Πηγή: <https://www.nytimes.com/interactive/2020/health/pfizer-biontech-covid-19-vaccine.html>

Τα αντιγονοπαρουσιαστικά κύτταρα ενεργοποιούν τα *T*-φρονικά κύτταρα για να αναζητήσουν και να καταστρέψουν τυχόν κύτταρα μολυσμένα από κορονοϊό που εμφανίζουν τα θραύσματα πρωτεΐνης ακίδας στις επιφάνειές τους.

1.4.2 Εμβόλιο της Moderna (mRNA-1273)

Η Moderna ήταν μια από τις πρώτες φαρμακευτικές εταιρείες που δεσμεύτηκαν να αναπτύξουν ένα εμβόλιο κατά της COVID-19. Στις 7 Φεβρουαρίου, η πρώτη παρτίδα του εμβολίου Moderna αναπτύχθηκε και ήταν έτοιμη για χρήση σε δοκιμές ανάλυσης. Ο Διεθνής Οργανισμός Υγείας (National Institute of Health (NIH)) ήταν πρώιμος υποστηρικτής του εμβολίου mRNA-1273, δεδομένου ότι συνεργάστηκε με τη Moderna στη διερεύνηση της αλληλουχίας mRNA για τον ιό της COVID-19. Στις 30 Νοεμβρίου 2020, η Moderna υπέβαλε αίτηση εξουσιοδότησης χρήσης έκτακτης ανάγκης στον FDA των ΗΠΑ και έλαβε έγκριση λίγο μετά το Μάϊο του 2021 (Patel et al., 2022).

Το εμβόλιο Pfizer και Moderna χρησιμοποιούν και τα δύο τεχνολογία mRNA και θεωρείται ότι έχουν παρόμοιους μηχανισμούς δράσης (Fessenden, 2020), όπως αυτόν που περιγράφηκε παραπάνω.

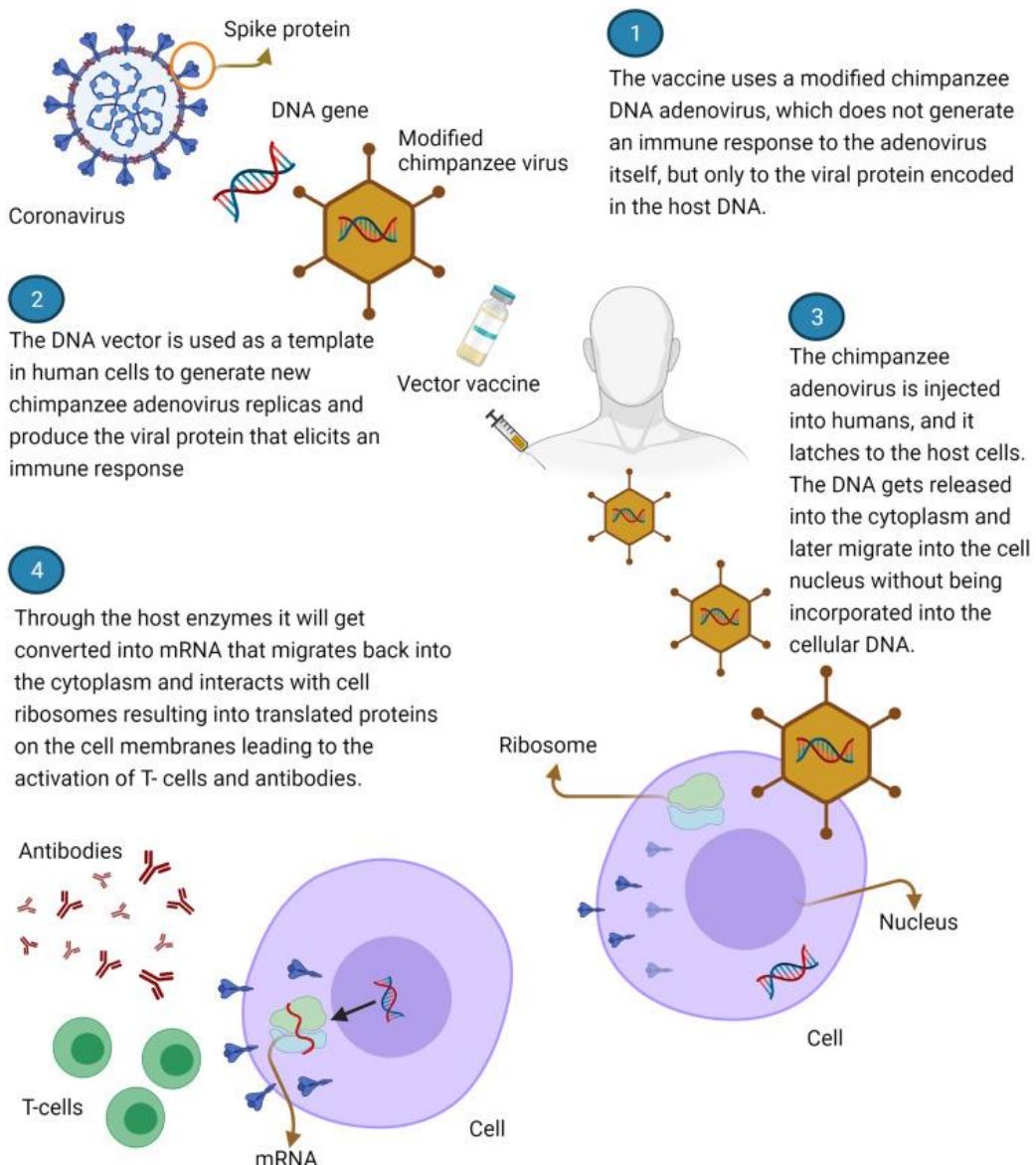
1.4.3 Εμβόλιο της AstraZeneca

Η AstraZeneca παρήγαγε ένα εμβόλιο ιικού φορέα χρησιμοποιώντας έναν γενετικά τροποποιημένο ιό που δεν μπορεί να προκαλέσει ασθένεια, αλλά που κωδικοποιεί τις πρωτεΐνες του κοροναϊού για να δημιουργήσει με ασφάλεια μια ανοσολογική απόκριση. Αυτό το εμβόλιο χρησιμοποιήθηκε αρχικά στο Ηνωμένο Βασίλειο και ήταν διαθέσιμο στην Ιταλία και την Πολωνία από τις 9 Φεβρουαρίου 2021 (Mascellino et al., 2021).

Το εμβόλιο AstraZeneca χρησιμοποιεί έναν τροποποιημένο DNA αδενοϊό χιμπατζή, ο οποίος δεν έχει εκτεθεί σε ανθρώπινους πληθυσμούς και δεν δημιουργεί ανοσολογική απόκριση στον ίδιο τον αδενοϊό, αλλά μόνο στην ιική πρωτεΐνη που κωδικοποιείται στο DNA του ξενιστή.

Ο φορέας DNA κωδικοποιεί μια πρωτεΐνη παρόμοια με το ιικό S-πεπτίδιο για να δημιουργήσει μια ανοσοαπόκριση εναντίον του. Ο φορέας DNA χρησιμοποιείται ως πρότυπο σε ανθρώπινα κύτταρα για τη δημιουργία νέων αντιγράφων αδενοϊού χιμπατζή και την παραγωγή της ιικής πρωτεΐνης που προκαλεί μια ανοσολογική απόκριση. Εν συντομίᾳ, ο αδενοϊός του χιμπατζή εγχέεται στον άνθρωπο και δεσμεύεται στα κύτταρα-ξενιστές. Το DNA απελευθερώνεται στο κυτταρόπλασμα και αργότερα μεταναστεύει στον κυτταρικό πυρήνα. Δεν ενσωματώνεται στο κυτταρικό DNA, αλλά χρησιμοποιεί τα ένζυμα του ξενιστή για να μετατραπεί σε mRNA που μεταναστεύει πίσω στο κυτταρόπλασμα και αλληλεπιδρά με τα ριβοσώματα του κυττάρου ξενιστή (ελεύθερα ή συνδεδεμένα με το ενδοπλασματικό δίκτυο), καταλήγοντας σε μεταφρασμένες πρωτεΐνες. Οι πρωτεΐνες εκφράζονται στις κυτταρικές μεμβράνες σχηματίζοντας σύμπλοκα MHC1 και MHC2. Σε αυτό το σημείο, οι μηχανισμοί των εμβολίων RNA και DNA είναι παρόμοιοι και οδηγούν στην ενεργοποίηση T-, B- και πλασματοκυττάρων και αντισωμάτων (Sette & Crotty, 2021).

Παρακάτω απεικονίζεται συνολικά ο μηχανισμός δράσης του εμβολίου:



Εικόνα 17. Σχηματική αναπαράσταση του μηχανισμού δράσεις του εμβολίου AstraZeneca, Πηγή: (Mascellino et al., 2021)

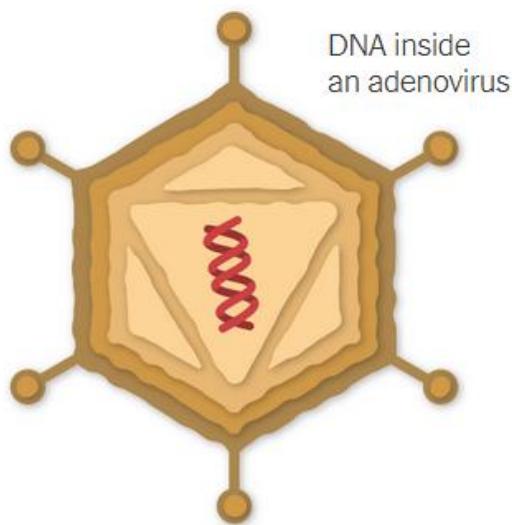
1.4.4 Εμβόλιο της Johnson & Johnson (Janssen)

Η Johnson & Johnson ανακοίνωσε επίσημα το βασικό υποψήφιο για το εμβόλιο τους χρησιμοποιώντας ένα φορέα αδενοϊού. Οι Φάσεις 1 και 2 των κλινικών μελετών ξεκίνησαν τον Ιούλιο του 2020 και συνδυάστηκαν για τον προσδιορισμό της ασφάλειας και της δοσολογίας στις Ηνωμένες Πολιτείες και το Βέλγιο. Η Φάση 3 ξεκίνησε στις 7 Σεπτεμβρίου 2020, με στόχο τον προσδιορισμό της αποτελεσματικότητας του εμβολίου. Η Φάση 3 ήταν η κλινική δοκιμή Ensemble και συμμετείχαν περίπου 44.000 άτομα. Τον Φεβρουάριο του 2021, ο FDA των ΗΠΑ εξέδωσε άδεια χρήσης έκτακτης ανάγκης της Johnson & Johnson και ξεκίνησε η χορήγηση εμβολίων (Patel et al., 2022). Το εμβόλιο είναι επίσης γνωστό ως JNJ-78436735 ή Ad26.COV2.S.

Το εμβόλιο Johnson & Johnson (Janssen) είναι προϊόν ενός φορέα ανθρώπινου αδενοϊού τύπου 26 ο οποίος είναι ανίκανος να ανασυνδυαστεί και να αναπαραχθεί και εκφράζει το αντιγόνο της πρωτεΐνης ακίδας SARS-CoV-2. Ο αδενοϊός είναι συνήθως υπεύθυνος για συμπτώματα παρόμοια με το κοινό κρυολόγημα και χρησιμεύει ως ο ικός φορέας. Ο αδενοϊός τύπου 26 είναι ένας ιός που απαντάται στη φύση και εμφανίζεται σε χαμηλό επιπολασμό στους ανθρώπους. Χωρίς τη διαδικασία της αντιγραφής, ο ιός δεν μπορεί να εξαπλωθεί στον ξενιστή του. Ένα γονίδιο της αντιγραφής διαγράφεται στο εμβόλιο, επομένως δεν μπορεί να αναπαραχθεί στον άνθρωπο και να προκαλέσει μόλυνση. Ως αποτέλεσμα, μπορούμε να πούμε ότι αποτελεί κατάλληλη μέθοδο για την παροχή γενετικού υλικού που κωδικοποιεί το αντιγόνο ακίδας στα ανθρώπινα κύτταρα. Το αντιγόνο ακίδας (S) είναι αυτό που είναι υπεύθυνο για την ανοσολογική προστασία που προσφέρει ο εμβολιασμός. Τα προστατευτικά αντισώματα παράγονται και προστατεύουν από μελλοντική μόλυνση (Wan et al., 2022). Έτσι λοιπόν, δημιουργείται μια ανοσολογική απόκριση έναντι του αντιγόνου S που έχει ως αποτέλεσμα μια επαγόμενη από το εμβόλιο απόκριση αντισωμάτων κατά της πρωτεΐνης ακίδας στον SARS-CoV-2 (Patel et al., 2022).

Σε γενικές γραμμές, τα εμβόλια για την CoViD-19 που βασίζονται σε αδενοϊούς είναι πιο ανθεκτικά από τα εμβόλια mRNA από την Pfizer και τη Moderna. Το DNA δεν είναι τόσο εύθραυστο όσο το RNA και το σκληρό πρωτεΐνικό περίβλημα του αδενοϊού βοηθά στην προστασία του γενετικού υλικού στο εσωτερικό.

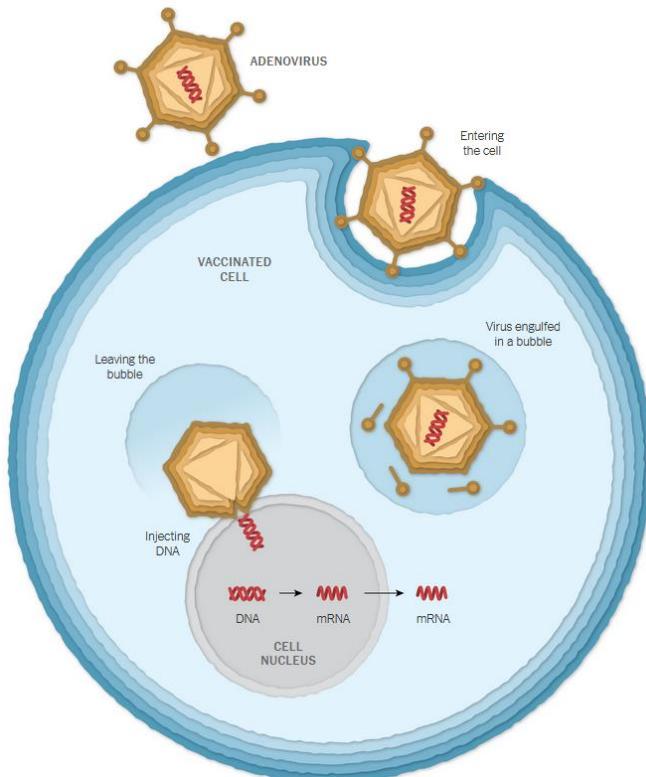
Παρακάτω παρουσιάζεται μια απεικόνιση του αδενοϊού 26 που φέρει το γενετικό υλικό με το γονίδιο της πρωτεΐνης ακίδας:



Εικόνα 18. DNA εντός του αδενοϊού 26, Πηγή:

<https://www.nytimes.com/interactive/2020/health/johnson-johnson-covid-19-vaccine.html>

Στη συνέχεια βλέπουμε τη διαδικασία εισαγωγής του αδενοϊού σε ένα ανθρώπινο κύτταρο:



Εικόνα 19. Διαδικασία εισαγωγής αδενοϊού σε ανθρώπινο κύτταρο, Πηγή: <https://www.nytimes.com/interactive/2020/health/johnson-johnson-covid-19-vaccine.html>

Μετά τη χορήγηση του εμβολίου, οι αδενοϊοί προσκρούουν στα κύτταρα και δεσμεύονται στις πρωτεΐνες στην επιφάνειά τους. Το κύτταρο «καταπίνει» τον ιό σε μια φυσαλίδα και τον απορροφάει στο εσωτερικό. Μόλις εισέλθει, ο αδενοϊός διαφεύγει από τη φυσαλίδα και ταξιδεύει στον πυρήνα, όπου βρίσκεται το DNA του κυττάρου. Ο αδενοϊός απελευθερώνει το DNA του στον πυρήνα. Σημειώνεται πως ο αδενοϊός είναι κατασκευασμένος έτσι ώστε να μην μπορεί να δημιουργήσει αντίγραφα του εαυτού του, αλλά το γονίδιο για την πρωτεΐνη ακίδας του κορωνοϊού μπορεί να διαβαστεί από το κύτταρο και να αντιγραφεί μέσω του mRNA.

1.4.5 Σύνοψη χαρακτηριστικών εμβολίων

Στον παρακάτω πίνακα μπορούμε να διακρίνουμε συγκεντρωτικά κάποια από τα χαρακτηριστικά των τεσσάρων εμβολίων που προαναφέρθηκαν:

Πίνακας 1. Σύνοψη χαρακτηριστικών εμβολίων

	Pfizer/BioNTtech	Moderna	Astra – Zeneca Oxford	Janssen Johnson & Johnson
Μηχανισμός δράσης	mRNA	mRNA	Ιϊκός φορέας αδενοϊού (αδενοϊός χιμπατζή με αδυναμία αντιγραφής)	Adenovirus Viral vector (ανθρώπινος αδενοϊός 26 με αδυναμία αντιγραφής)
Αντιγόνο	Συνολικό μήκος πρωτεΐνης ακίδας	Συνολικό μήκος πρωτεΐνης ακίδας	Πρωτεΐνη ακίδας	Πρωτεΐνη ακίδας
Δόσεις	2 δόσεις με διαφορά 21 ημερών	2 δόσεις με διαφορά 28 ημερών	2 δόσεις με διαφορά 12 εβομάδων	Μόνο μια δόση
Παρενέργειες	Σπάνιες αλλεργίες και αναφυλαξία	Σπάνια παράλυση προσώπου (Bell's Palsy)	Σπάνια επεισόδια θρομβοεμβολής, σπάνιες περιπτώσεις θρόμβων, πνευμονική εμβολή, θρομβοπενία	Σπάνιες περιπτώσεις θρόμβων, θρομβοπενία, σύνδρομο Guillain-Barré
Συνολική αποτελεσματικότητα	95% για την ασθένεια 87.5% για σοβαρή ασθένεια	94% για την ασθένεια 100% για σοβαρή ασθένεια	70% (64% μετά την πρώτη δόση) (70.4% μετά τις 2 δόσεις)	72% στις ΗΠΑ 66% στη Λατινική Αμερική 57% στη Νότια Αφρική

1.5 ΒΑΣΗ ΔΕΔΟΜΕΝΩΝ EUDRAVIGILANCE

1.5.1 Εισαγωγή

Η φαρμακοεπαγρύπνηση είναι μια βασική αρχή για την προαγωγή και την προστασία των ασθενών και της δημόσιας υγείας. Η βάση δεδομένων EudraVigilance είναι ένα σύστημα συλλογής, διαχείρισης και ανάλυσης ύποπτων ανεπιθύμητων παρενεργειών των φαρμάκων (Adverse Drug Reactions (ADRs)) σε φάρμακα που έχουν εγκριθεί στην Ευρωπαϊκή Ένωση. Το σύστημα διαχειρίζεται ο Ευρωπαϊκός Οργανισμός Φαρμάκων (European Medicines Agency (EMA)) και τέθηκε πρώτη φορά σε λειτουργία το Δεκέμβρη του 2001, εκ μέρους του ρυθμιστικού δικτύου φαρμάκων της Ευρωπαϊκής Ένωσης (Postigo et al., 2018).

Από το Νοέμβριο του 2005, η ηλεκτρονική καταγραφή των ύποπτων ανεπιθύμητων παρενεργειών των φαρμάκων είναι υποχρεωτική στην Ευρωπαϊκή Ένωση (European Comission, 2001, 2004). Η υποβολή των ADRs τόσο από εμπορικά προϊόντα όσο και από προϊόντα που χρησιμοποιούνται σε κλινικές μελέτες βασίζεται στα πρότυπα που έχουν συμφωνηθεί στο επίπεδο του Διεθνούς Συμβουλίου για την εναρμόνιση των τεχνικών απαιτήσεων για φαρμακευτικά προϊόντα που προορίζονται για ανθρώπινη χρήση (International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use) και του Διεθνούς Οργανισμού Τυποποίησης (International Organization for Standardization), ώστε οι πληροφορίες που αφορούν στην ασφάλεια να συλλέγονται και να ανταλλάσσονται με ένα δομημένο και τυποποιημένο τρόπο (Postigo et al., 2018).

Μέχρι πρόσφατα, ενώ ο Ευρωπαϊκός Οργανισμός Φαρμάκων και οι Αρμόδιες Εθνικές Αρχές (National Competent Authorities (NCAs)) είχαν πλήρη πρόσβαση στην EudraVigilance, τα λοιπά ενδιαφερόμενα μέρη είχαν περιορισμένη πρόσβαση. Η πιο πρόσφατη έκδοση της EudraVigilance παρέχει μια εύκολα προσβάσιμη πλατφόρμα σε όλα τα ενδιαφερόμενα μέρη δίνοντας τους τη δυνατότητα να χρησιμοποιήσουν τα διαθέσιμα δεδομένα ανάλογα με τις ανάγκες τους, τα ενδιαφέροντά τους και τους περιορισμούς που τίθενται λόγω φαρμακοεπαγρύπνησης (Postigo et al., 2018).

1.5.2 Δεδομένα της EudraVigilance

Τα δεδομένα στην EudraVigilance υποβάλλονται ηλεκτρονικά από τις εθνικές ρυθμιστικές αρχές φαρμάκων και από φαρμακευτικές εταιρείες που κατέχουν άδεια κυκλοφορίας για φάρμακα. Τα δεδομένα της EudraVigilance δημοσιεύονται στην Ευρωπαϊκή βάση δεδομένων αναφορών ύποπτων ανεπιθύμητων ενεργειών φαρμάκων, την πύλη adrreports.eu, σε 26 γλώσσες. Αυτή η

πύλη επιτρέπει στους χρήστες να έχουν πρόσβαση στον συνολικό αριθμό μεμονωμένων αναφορών με ύποπτες παρενέργειες (γνωστές και ως Individual Case Safety Reports ή ICSR) που υποβάλλονται στην EudraVigilance για φάρμακα που έχουν εγκριθεί στον Ευρωπαϊκό Οικονομικό Χώρο. Ο Ευρωπαϊκός Οργανισμός Φαρμάκων δημοσιεύει τα δεδομένα που είναι διαθέσιμα στην πύλη [adrreports.eu](#) ώστε τα ενδιαφερόμενα μέρη της, συμπεριλαμβανομένου του ευρύτερου κοινού, να έχουν πρόσβαση σε πληροφορίες ώστε οι Ευρωπαϊκές ρυθμιστικές αρχές να μπορούν να χρησιμοποιήσουν για να αξιολογήσουν την ασφάλεια ενός φαρμάκου ή μιας δραστικής ουσίας.

Τα δεδομένα που είναι διαθέσιμα στην πύλη βασίζονται σε ανεπιθύμητες ενέργειες που έχουν αναφερθεί αυθόρμητα από ασθενείς, επαγγελματίες υγείας ή άλλες πηγές, οι οποίες στη συνέχεια υποβάλλονται ηλεκτρονικά στη EudraVigilance με τη μορφή ICSR από εθνικές ρυθμιστικές αρχές φαρμάκων ή φαρμακευτικές εταιρείες. Η πύλη [adrreports.eu](#) παρέχει πρόσβαση σε συγκεντρωτικά δεδομένα που βασίζονται σε προκαθορισμένα ερωτήματα (queries). Αυτά στη συνέχεια διατίθενται με τη μορφή διαδικτυακών αναφορών που αποτελούνται από έναν αριθμό καρτελών (tabs), καθεμία από τις οποίες επιτρέπει στους χρήστες να αναζητούν, να φίλτραρουν και να έχουν πρόσβαση στα δεδομένα με διαφορετικούς τρόπους. Επιπλέον, παρέχεται πρόσβαση σε λίστα μεμονωμένων περιπτώσεων και τα έντυπα αναφοράς μεμονωμένων περιπτώσεων (ICSR) που συμμορφώνονται με το νόμο προστασίας προσωπικών δεδομένων της Ευρωπαϊκής Ένωσης.

Προτού υποβληθεί ένα ICSR στην EudraVigilance, ο ενδιαφερόμενος συμπληρώνει τα στοιχεία δεδομένων παρέχοντας πληροφορίες για τις ύποπτες ανεπιθύμητες ενέργειες που έχουν παρατηρηθεί μετά τη χρήση ενός ή περισσότερων φαρμάκων. Βέβαια, αυτές οι ύποπτες ανεπιθύμητες ενέργειες δεν σχετίζονται απαραίτητα με ή προκαλούνται από το φάρμακο. Οι διαδικτυακές αναφορές στις οποίες είναι δυνατή η πρόσβαση μέσω της πύλης [adrreports.eu](#) παρέχουν διαφορετικές οπτικές των δεδομένων σχετικά με τα ICSRs, που αποτελούν μέρος κάθε μεμονωμένης περίπτωσης που υποβάλλεται στην EudraVigilance. Τα στοιχεία δεδομένων που είναι διαθέσιμα στους χρήστες της πύλης καθορίζονται από την [Πολιτική Πρόσβασης EudraVigilance](#) (European Medicines Agency (EMA), 2017).

Σχετικά με τις διαδικτυακές αναφορές, ισχύουν τα ακόλουθα:

- Η **ηλικιακή ομάδα (Age)** και το **φύλο (Sex)** παρέχουν πληροφορίες για το άτομο που βίωσε την ύποπτη ανεπιθύμητη παρενέργεια.
- Ο **τύπος αναφοράς (Report Type)** παρέχει πληροφορίες σχετικά με την ταξινόμηση μιας αναφοράς από τον αποστολέα (π.χ. αυθόρμητη αναφορά).

- Η **σοβαρότητα (Seriousness)** παρέχει πληροφορίες σχετικά με την πιθανολογούμενη ανεπιθύμητη ενέργεια και μπορεί να ταξινομηθεί ως «σοβαρή» εάν αντιστοιχεί σε ιατρικό περιστατικό που οδηγεί σε θάνατο, είναι απειλητική για τη ζωή, απαιτεί νοσηλεία σε νοσοκομείο, οδηγεί σε άλλη σημαντική ιατρικά κατάσταση ή παράταση της ήδη υπάρχουσας νοσηλείας, οδηγεί σε επίμονη ή σημαντική αναπηρία ή ανικανότητα ή είναι συγγενής ανωμαλία/εκ γενετής ελάττωμα. Μπορεί επίσης να αναφέρεται σε άλλα σημαντικά ιατρικά συμβάντα που μπορεί να μην είναι άμεσα απειλητικά για τη ζωή ή οδηγούν σε θάνατο ή νοσηλεία αλλά μπορούν να θέσουν σε κίνδυνο τον ασθενή ή μπορεί να απαιτήσουν παρέμβαση (θεραπεία) για την πρόληψη κάποιου από τα άλλα αποτελέσματα που αναφέρονται παραπάνω. Παραδείγματα τέτοιων συμβάντων είναι ο αλλεργικός βρογχόσπασμος (ένα σοβαρό πρόβλημα με την αναπνοή) που απαιτείται θεραπεία σε δωμάτιο έκτακτης ανάγκης ή στο σπίτι καθώς και επιληπτικές κρίσεις/σπασμοί και σοβαρές δυσκρασίες του αίματος (διαταραχές του αίματος) που δεν οδηγούν σε νοσηλεία.
- Η **Γεωγραφική Προέλευση (Geographic Origin)** παρέχει πληροφορίες για την τοποθεσία του αναφέροντος.
- Η **Ομάδα Αναφέροντος (Reporter Group)** παρέχει πληροφορίες για την κατάρτιση του αναφέροντος.
- Η **Έκβαση (Outcome)** παρέχει πληροφορίες σχετικά με την τελευταία αναφερόμενη κατάσταση της ύποπτης ανεπιθύμητης ενέργειας.
- Η **Αναφερόμενη Ύποπτη Αντίδραση (Reported Suspected Reaction)** παρέχει πληροφορίες σχετικά με τις ανεπιθύμητες ενέργειες που εμφανίζει ένας ασθενής σύμφωνα με αυτόν που τις αναφέρει.

Στον παρακάτω πίνακα παρουσιάζονται με λεπτομέρεια οι κατηγορίες που αναγράφονται σε αυτές τις αναφορές καθώς και οι πιθανές τιμές που μπορούν να πάρουν.

Πίνακας 2. Περιγραφή κατηγοριών που εμφανίζονται στις αναφορές της EudraVigilance

Data element	Possible Values
Age group (based on the reported patient age or calculated based on difference between “Date of Birth” and “First Reaction Date” (if available in a valid date format dd/mm/yyyy))	Not specified 0-1 Month 2 Months – 2 Years 3 – 11 Years 12 – 17 Years 18 – 64 Years 65 – 85 Years More than 85 Years
Sex	Female Male Not specified
Report Type	Spontaneous
Seriousness	Not specified Serious Non-serious
Geographic Origin	European Economic Area (EEA) Non-European Economic Area (Non-EEA) Not Specified
Reporter Group	Healthcare Professional (Physician, Pharmacist or other Health Professional) Non-Healthcare Professional (Lawer, Consumer, or other non-Health Professional) Not specified
Outcome	Recovered/Resolved Recovering/Resolving Not recovered/Not resolved Recovered/Resolved with Sequelae Fatal Unknown Not specified
Reported Suspected Reaction	Any undesirable effect (suspected adverse reaction) reported by the reporter. Undesirable effect terms are coded in line with a dictionary of medicinal terms used to classify clinical information
Reaction Groups	Any undesirable effect group based on the classification reported by the reporter. Undesirable effect terms are coded in line with a dictionary of medicinal terms used to classify clinical information and are categorized into groups based on the clinical signification
Number of individual cases	Running total count of individual cases submitted to EudraVigilance

Η Αναφερόμενη Ύποπτη Αντίδραση (Reported Suspected Reaction) και η Ομάδα Αντιδράσεων (Reaction Groups) για μια αναφορά προέρχονται από το λεξικό ιατρικών όρων που χρησιμοποιείται για την ταξινόμηση κλινικής πληροφορίας. Το εν λόγω λεξικό είναι το Ιατρικό Λεξικό για Ρυθμιστικές Δραστηριότητες (Medical Dictionary for Regulatory Activities (MedDRA)). Η **Αναφερόμενη Ύποπτη Αντίδραση** αντιστοιχεί στην αντίδραση MedDRA «Προτιμώμενος όρος (Preferred Term (PT))» και οι Ομάδες Αντίδρασης αντιστοιχούν στην αντίδραση MedDRA «Κατηγορία Οργανικού Συστήματος (System Organ Class (SOC))».

1.5.3 Αναζήτηση στη βάση

Η πρόσβαση στη βάση πραγματοποιείται ακολουθώντας το σύνδεσμο <https://www.adrreports.eu/en/index.html>. Στη συνέχεια έχουμε τη δυνατότητα να πραγματοποιήσουμε αναζήτηση των δραστικών ουσιών ή των εμβολίων που μας ενδιαφέρουν με αλφαριθμητική σειρά. Έστω ότι θέλουμε να ανακτήσουμε πληροφορίες για τα εμβόλια της COVID-19 που είναι και το θέμα της παρούσας διπλωματικής. Επιλέγουμε το αρχικό γράμμα 'C' και περιηγούμαστε στη λίστα με τις διαθέσιμες δραστικές ουσίες μέχρι να βρούμε αυτή που μας ενδιαφέρει. Η εικόνα που παίρνουμε είναι η παρακάτω:

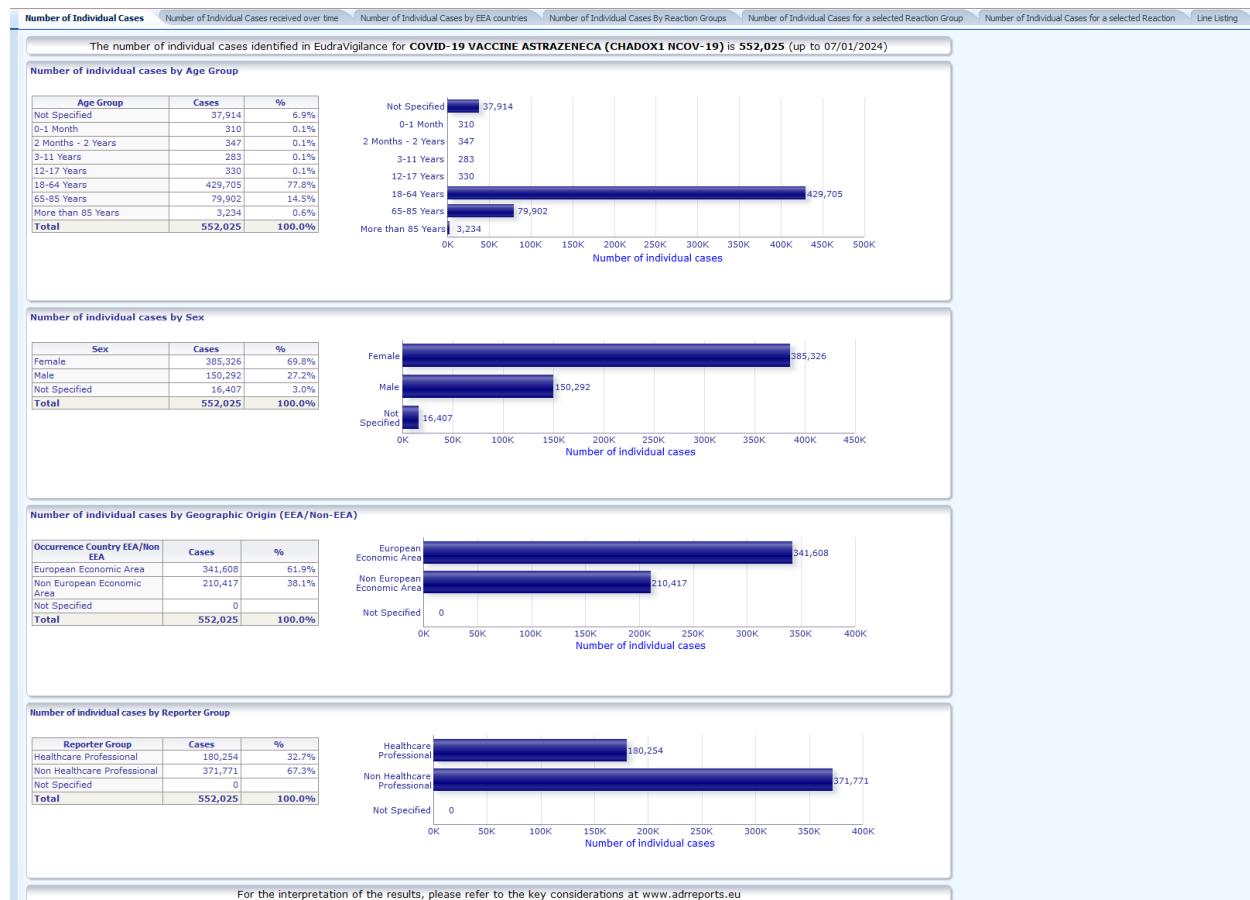
The screenshot shows the EudraVigilance homepage with a search bar at the top containing the term "COVID-19". Below the search bar, a list of search results is displayed, including:

- COLLAGENASE CLOSTRIDIUM HISTOLYTICUM
- COLLAGENASE CLOSTRIDIUM HISTOLYTICUM, PROTEASE*
- COLLOIDAL BISMUTH PECTIN*
- COMPLEMENT C1 ESTERASE INHIBITOR
- CONCENTRATE OF PROTEOLYTIC ENZYMES ENRICHED IN BROMELAIN
- CONCENTRATED SOLUTION OF SODIUM HYPOCHLORITE*
- CONCIZUMAB*
- COPPER*
- COPPER GLUCONATE, MANGANESE GLUCONATE*
- COPPER, IRON, MANGANESE*
- CORIFOLLITROPIN ALFA
- CORTICORELIN*
- CORTICOTROPIN*
- CORTISONE*
- CORTIVAZOL*
- COVID-19 MRNA VACCINE MODERNA (ELASOMERAN)
- COVID-19 MRNA VACCINE MODERNA OMICRON XBB.1.5 (ANDUSOMERAN)
- COVID-19 MRNA VACCINE MODERNA ORIGINAL/OMICRON BA.1 (ELASOMERAN, IMELASOMERAN)
- COVID-19 MRNA VACCINE MODERNA ORIGINAL/OMICRON BA.4-5 (ELASOMERAN, DAVESOMERAN)
- COVID-19 MRNA VACCINE PFIZER-BIONTECH (TOZINAMERAN)
- COVID-19 MRNA VACCINE PFIZER-BIONTECH OMICRON XBB.1.5 (RAXTOZINAMERAN)
- COVID-19 MRNA VACCINE PFIZER-BIONTECH ORIGINAL/OMICRON BA.1 (TOZINAMERAN, RILTOZINAMERAN)
- COVID-19 MRNA VACCINE PFIZER-BIONTECH ORIGINAL/OMICRON BA.4-5 (TOZINAMERAN, FAMTOZINAMERAN)
- COVID-19 VACCINE ASTRazeneca (ChAdOx1 NCov-19)
- COVID-19 VACCINE JANSSEN (AD26.COV2.S)
- COVID-19 VACCINE NOVAVAX (NVX-COV2373)
- COVID-19 VACCINE NOVAVAX XBB.1.5 (NVX-COV2373)
- COVID-19 VACCINE VALNEVA
- COVID-19 VACCINE VIDPREVTYN BETA
- CRISABOROLE
- CRIZANLIZUMAB
- CRIZOTINIB

Εικόνα 20. Περιβάλλον αναζήτησης δραστικών ουσιών της EudraVigilance

Επιλέγοντας την επιθυμητή δραστική ουσία ή εμβόλιο μεταβαίνουμε σε ένα νέο περιβάλλον που περιλαμβάνει το λεγόμενο 'web report'. Ένα web report αποτελείται συνολικά από επτά καρτέλες

(tabs) που περιέχουν στατιστικές πληροφορίες για την υπό συζήτηση ουσία/εμβόλιο και το αποτέλεσμα της αναζήτησης είναι το παρακάτω:



Εικόνα 21. Μορφή web report της EudraVigilance μετά από αναζήτηση δεδομένης δραστικής ουσίας

Οι καρτέλες (tabs) που εμφανίζονται είναι κατά σειρά:

1. **Number of Individual Cases:** Η καρτέλα αυτή παρέχει το τρέχον σύνολο των μεμονωμένων περιπτώσεων (individual cases) που εντοπίζονται στη EudraVigilance μέχρι το τέλος του προηγούμενου μήνα. Η καρτέλα παρουσιάζει πληροφορίες για τον αριθμό των μεμονωμένων περιπτώσεων ανά Ηλικιακή Ομάδα, Φύλο και Γεωγραφική Προέλευση.
2. **Number of Individual Cases received over time:** Η καρτέλα αυτή παρέχει τον αριθμό των μεμονωμένων υποθέσεων που ελήφθησαν τους τελευταίους 12 μήνες χωρισμένος κατά γεωγραφική προέλευση, δηλαδή περιπτώσεις που προκύπτουν σε χώρες του Ευρωπαϊκού Οικονομικού Χώρου (EOX) σε σχέση με εκείνες που προκύπτουν εκτός του EOX. Η καρτέλα παρέχει επίσης ένα γράφημα με μια γραμμή τάσης που υποδεικνύει τον συνολικό αριθμό των μεμονωμένων περιπτώσεων στο πέρασμα του χρόνου.

3. Number of Individual Cases by EEA countries: Η καρτέλα αυτή εμφανίζει τον αριθμό των μεμονωμένων περιπτώσεων στις χώρες του ΕΟΧ για το επιλεγμένο φάρμακο προϊόν/ουσία. Περιλαμβάνει ουσιαστικά ένα χάρτη που εμφανίζει το ποσοστό των συνολικών κρουσμάτων σε κάθε χώρα του ΕΟΧ. Επίσης περιλαμβάνει ένα γράφημα που εμφανίζει τον συνολικό αριθμό μεμονωμένων περιπτώσεων σε κάθε χώρα.
4. Number of Individual Cases by Reaction Groups: Η καρτέλα αυτή εμφανίζει ένα γράφημα που απεικονίζει τον αριθμό των μεμονωμένων περιπτώσεων ανά ομάδα αντίδρασης. Διατίθενται πέντε διακριτές προβολές, οι οποίες επιτρέπουν στους χρήστες να διαχωρίζουν τα δεδομένα της Ομάδας αντίδρασης σε αυτήν την καρτέλα κατά Ηλικία Ομάδα, Φύλο, Σοβαρότητα, Ομάδα Ρεπόρτερ και Γεωγραφική Προέλευση.
5. Number of Individual Cases for a selected Reaction Group: Η καρτέλα αυτή εμφανίζει τον αριθμό των μεμονωμένων περιπτώσεων για μια επιλεγμένη Ομάδα Αντίδρασης που έχει ορίσει ο χρήστης. Διατίθενται τρία είδη web reports για μια επιλεγμένη Ομάδα Αντίδρασης. Το πρώτο παρουσιάζει τα δεδομένα κατά Ηλικιακή Ομάδα & Φύλο (Age Group & Sex), το δεύτερο κατά Ομάδα Αναφέροντος (Reporter Group) και το τρίτο κατά Γεωγραφική Προέλευση (Geographic Origin).
6. Number of Individual Cases for a selected reaction: Η καρτέλα αυτή εμφανίζει τον αριθμό των μεμονωμένων περιπτώσεων για μια επιλεγμένη Αντίδραση που έχει ορίσει ο χρήστης. Διατίθενται τρία είδη web reports, όπως και στην προηγούμενη καρτέλα, για μια επιλεγμένη Αντίδραση. Το πρώτο παρουσιάζει τα δεδομένα κατά Ηλικιακή Ομάδα & Φύλο (Age Group & Sex), το δεύτερο κατά Ομάδα Αναφέροντος (Reporter Group) και το τρίτο κατά Γεωγραφική Προέλευση (Geographic Origin).
7. Line Listing: Η καρτέλα αυτή εμφανίζει μια λίστα μεμονωμένων περιπτώσεων που έχουν αναφερθεί στην EudraVigilance για ένα καθορισμένο προϊόν ή ουσία. Τα στοιχεία δεδομένων εμφανίζονται σύμφωνα με το επίπεδο πρόσβασης που παρέχεται στο κοινό κατά την Πολιτική Πρόσβασης της EudraVigilance. Τα στοιχεία δεδομένων που παρατίθενται παρακάτω μπορούν να χρησιμοποιηθούν για το φίλτραρισμα αυτής της λίστας:
 - Σοβαρότητα (Seriousness)
 - Γεωγραφική Προέλευση (Geographic Origin)
 - Ομάδα Αναφέροντος (Reporter Group)
 - Φύλο (Sex)
 - Ηλικιακή Ομάδα (Age Group)
 - Ομάδα Αντίδρασης (Reaction Group)

- Ύποπτη Αντίδραση του Αναφέροντος (Reporter Suspected Reaction)
- Ημερομηνία εγγραφής (Gateway date)

Τα στοιχεία δεδομένων που αναφέρονται σε αυτή τη λίστα συνοψίζονται στον παρακάτω πίνακα:

Πίνακας 3. Στοιχεία δεδομένων που εμφανίζονται στις λίστες της EudraVigilance

Line Listing Data Elements	ICH E2B (R3) Element Reference	Description	Example
EU Local Number	N/A	EudraVigilance local number, which is an identifier assigned to the ICSR in EudraVigilance	EU-EC-12345
EV Gateway Receipt Date	N/A	EudraVigilance Gateway Date, which is the date of receipt of the ICSR in EudraVigilance	01/01/2014
Report Type	C.1.3	Type of Report	Spontaneous
Primary source qualification	C.2.r.4	Primary source Qualification: grouped as Healthcare Professional or Non-Healthcare Professional	Healthcare Professional
Primary source country for regulatory purposes	C.2.r.5	Primary Source for Regulatory Purposes, displayed as EEA/non EEA	EEA
Literature Reference(s)	C.4.r.1	The literature reference(s) for suspected adverse reactions described in the literature and the corresponding ICSRs in EudraVigilance	Tolerable pain reduces gastric fundal accommodation and gastric motility in healthy subjects: a crossover ultrasonographic study. Hauo H1, Kusunoki H2, Kanbara K1, Abe T1, Yunoki N3, Haruma K2, Fukunaqa M1. Biopsychosoc Med. 2015 Feb
Patient age group	D.2.2a	Mapped against the 'Age at Time of Onset of Reaction/Event' based on the reported patient age or calculated based on difference between 'Date of Birth' and 'First Reaction Start Date' (if	18 – 64 years

		available in a valid date format dd/mm/yyyy)	
	D.2.2b	'Age at time of onset of Reaction/Event (unit)	
Patient sex	D.5	'Sex' (gender of the patient)	Female
Parent/Child	N/A	To indicated if this is a report that relates to a parent and a child	Yes
Reaction List PT (Duration – Outcome – seriousness criteria)	E.i.2.1b	'Reaction/Event MedDRA Preferred term' description	Rash (3d – Resolved – Life Threatening, Caused/Prolonged Hospitalisation)
	E.i.6a/b	'Duration of Reaction/Event'	Nausea (1d – Resolved)
	E.i.7	'Outcome of Reaction/Event at the Time of Last Observation'	Headache (3d – Not resolved)
	E.i.3.2a, E.i.3.2b, E.i.3.2c, E.i.3.2d, E.i.3.2e, E.i.3.2f	The seriousness criteria of the reported reaction, e.g. Results in Death, Life Threatening, Caused/Prolonged Hospitalisation, Disabling/Incapacitating, Congenital Anomaly/Birth Defect, Other Medically Important Condition	
Drug List (Drug Char – Indication PT – Action taken – [Duration – Dose – Route]) or Drug List (Drug Char – Indication PT – Action taken – [Duration – Dose – Route – More in ICSR])	G.k.1	Characterisation of 'Drug Role', defined as suspect, interacting, concomitant or drug not administered. Based on this data element, 2 different- 'Drug' (medicines) lists will be created: -for suspect and interacting drugs -for concomitant or drug not administered	PRODUCT [Substance] (S – Dental pain, Headache – Drug withdrawn – [1d – 0.5 mg – oral]) or PRODUCT [Substance] (S – Detnal pain, Headache – Drug withdrawn – [1d – 0.5mg – oral – More in ICSR])
	G.k.2.2	Reported medicinal product, displayed as recoded against the Extended EudraVigilance Medicinal Product Dictionary for centrally authorized products (for non-centrally authorized	

		products, only the recoded substance will be displayed where reported)	
G.k.2.3.r.1		Substance/Specified Substance Name, displayed as recoded against the Extended EudraVigilance Medicinal Product Dictionary (if not, it will be displayed as reported)	
G.k.7.r.2b		Indication of the medicinal product described as MedDRA Preferred Term	
G.k.4.r.6a		'Duration of Drug Administration', as reported or based on 'Drug Administration Start Date' and 'End Date'	
G.k.4.r.1a/b		Dose of the medicine	
G.k.4.r.10.2		Route of administration of the medicine	

Οι λίστες που παράγονται με βάση τα φίλτρα αναζήτησης που επιλέγει ο χρήστης μπορούν να εξαχθούν σε διάφορες μορφές αρχείων (όπως .csv, .pdf, .txt κλπ.) ώστε να χρησιμοποιηθούν και να αξιοποιηθούν ανάλογα με τις ανάγκες του. Αυτός ήταν και ο τρόπος που ανακτήθηκε το σύνολο δεδομένων για την εκπόνηση της συγκεκριμένης διπλωματικής, ο οποίος θα αναλυθεί παρακάτω στην ενότητα 'Μέθοδοι'.

1.6 ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

1.6.1 Ορισμός

Η Μηχανική Μάθηση (Machine Learning) είναι η τεχνολογία ανάπτυξης αλγορίθμων υπολογιστών που μπορούν να μιμηθούν την ανθρώπινη νοημοσύνη. Ως κλάδος της επιστήμης των υπολογιστών η μηχανική μάθηση προέρχεται από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην τεχνητή νοημοσύνη. Είναι ένα νέο πεδίο το οποίο εξελίσσεται συνεχώς και βρίσκει εφαρμογές σε πολλούς τομείς όπως η θιοπληροφορική, η οικονομία, το μάρκετινγκ, η χημειοπληροφορική κ.ά. Αυτοί οι αλγόριθμοι έχουν κατασκευαστεί για να μπορούν να

βελτιώνονται αυτόματα μέσω της εμπειρίας και με τη χρήση δεδομένων, γνωστών ως δεδομένα εκπαίδευσης. Οι αλγόριθμοι μηχανικής μάθησης μπορούν να ταξινομηθούν σε ξεχωριστές κατηγορίες ανάλογα με τη φύση των δεδομένων, τη διαδικασία εκμάθησης και τον τύπο μοντέλου (El Naqa & Murphy, 2015). Η Μηχανική Μάθηση θεωρείται ως μέρος της Τεχνητής Νοημοσύνης και χρησιμοποιείται για τη λήψη προβλέψεων ή αποφάσεων με βάση τη μαθησιακή εμπειρία που απέκτησε το μοντέλο μέσω της εκπαίδευσης. Μέχρι σήμερα η τεχνολογία μηχανικής μάθησης έχει εφαρμοστεί σε διαφορετικά πεδία όπως η αναγνώριση προτύπων (Bishop, 2006), υπολογιστική όραση (Apolloni et al., 2005), μηχανική διαστημικών σκαφών (Ao et al., 2010), χρηματοδότηση (Gyorfi et al., 2012), ψυχαγωγία (Gong & Xu, 2007; J. Yu & Tao, 2013), οικολογία (Stockwell & Fielding, 1999), υπολογιστική βιολογία (Mitra et al., 2008; Yang, 2010) και βιοϊατρικές και ιατρικές εφαρμογές (Cleophas et al., 2013; Malley et al., 2011).

1.6.2 Διαχωρισμός συνόλου δεδομένων

Στη μηχανική μάθηση η σημασία της αξιόπιστης αξιολόγησης μοντέλων είναι κομβική. Κεντρικό στοιχείο αυτής της διαδικασίας αξιολόγησης είναι ο διαχωρισμός των διαθέσιμων δεδομένων σε διακριτά υποσύνολα: ένα σύνολο εκπαίδευσης (training set) και ένα σύνολο δοκιμής (test set). Αυτή η πρακτική χρησιμεύει ως ακρογωνιαίος λίθος για τη διασφάλιση της αξιοπιστίας, της γενίκευσης και της αποτελεσματικότητας των μοντέλων μηχανικής μάθησης.

Η έννοια του διαχωρισμού των δεδομένων σε σύνολο εκπαίδευσης και δοκιμής συνεπάγεται τη διαίρεση του συνόλου δεδομένων σε δύο ξεχωριστά υποσύνολα, το καθένα από τα οποία εξυπηρετεί ένα διαφορετικό σκοπό κατά την ανάπτυξη του μοντέλου. Το σύνολο εκπαίδευσης, που αποτελεί το σημαντικότερο μέρος του συνόλου δεδομένων, χρησιμοποιείται για την εκπαίδευση του μοντέλου μηχανικής μάθησης. Εδώ, ο αλγόριθμος «μαθαίνει» τα υποκείμενα μοτίβα, τις σχέσεις και τα χαρακτηριστικά που υπάρχουν στα δεδομένα μέσω διαφόρων τεχνικών όπως η παλινδρόμηση, η ταξινόμηση ή η ομαδοποίηση, όπως θα εξηγηθούν παρακάτω. Αντίθετα, το σύνολο δοκιμής, το οποίο δε χρησιμοποιεί το μοντέλο κατά τη φάση εκπαίδευσης, λειτουργεί ως ανεξάρτητο σύνολο δεδομένων που χρησιμοποιείται αποκλειστικά για την αξιολόγηση του μοντέλου. Αυτός ο διαχωρισμός είναι ζωτικής σημασίας για την αξιολόγηση της απόδοσης του μοντέλου σε νέα δεδομένα, παρέχοντας έτσι πληροφορίες για τις δυνατότητες γενίκευσής του καθώς και τις δυνατότητες εφαρμογής στον πραγματικό κόσμο. Η λογική πίσω από αυτό το διαχωρισμό έγκειται στον μετριασμό του κινδύνου υπερπροσαρμογής, ένα σύνηθες πρόβλημα στη μηχανική μάθηση όπου το μοντέλο μαθαίνει να απομνημονεύει τα δεδομένα εκπαίδευσης αντί να καταγράφει υποκείμενα μοτίβα. Αξιολογώντας το μοντέλο σε ένα ξεχωριστό σύνολο δοκιμής, μπορούμε να μετρήσουμε και να αξιολογήσουμε την ικανότητά του να γενικεύει σε νέα

δεδομένα, διασφαλίζοντας έτσι την αξιοπιστία και την αποτελεσματικότητά του σε σενάρια πραγματικού κόσμου (Goodfellow et al., 2016; Hastie, Tibshirani, & Friedman, 2009a).

1.6.3 Τύποι Μηχανικής Μάθησης

Υπάρχουν τρεις τύποι μηχανικής μάθησης: η **επιβλεπόμενη μάθηση** (supervised learning), η **μάθηση χωρίς επίβλεψη** (unsupervised learning) και η **ενισχυτική μάθηση** (reinforcement learning). Σε αυτήν τη διπλωματική εργασία χρησιμοποιήθηκε η επιβλεπόμενη μάθηση για την ταξινόμηση, παρόλα αυτά θα σχολιαστούν συνοπτικά και οι υπόλοιποι τύποι μηχανικής μάθησης.

1.6.3.1 Επιβλεπόμενη Μάθηση

Το βασικό χαρακτηριστικό της επιβλεπόμενης μάθησης είναι η διαθεσιμότητα σχολιασμένων δεδομένων. Πιο συγκεκριμένα, η επιβλεπόμενη μάθηση συνεπάγεται την εκμάθηση μιας αντιστοίχισης μεταξύ ενός συνόλου μεταβλητών εισόδου (input variables) και μιας μεταβλητής εξόδου (output variable), που ονομάζεται label, και στη συνέχεια αυτή η αντιστοίχιση εφαρμόζεται για να προβλέψει τις τιμές για νέα δεδομένα που δεν έχουν ξαναϊδωθεί (Cunningham et al., 2008). Έχοντας επισημάνει τα δεδομένα, δηλαδή γνωρίζοντας τη σωστή έξοδο (output) για την κάθε είσοδο (input), το μοντέλο θα εκπαιδευτεί με την πάροδο του χρόνου, μετρώντας την ακρίβεια (accuracy) μέσω μιας συνάρτησης απώλειας (loss function) η οποία εκτελεί συνεχείς προσαρμογές μέχρι το σφάλμα ελαχιστοποιηθεί επαρκώς.

Υπάρχουν δύο τύποι τεχνικών επιβλεπόμενης μάθησης στη μηχανική μάθηση: **παλινδρόμηση** και **ταξινόμηση**.

- **Παλινδρόμηση (Regression):** χρησιμοποιείται για την πρόβλεψη μιας συνεχούς μεταβλητής που βασίζεται στη σχέση μεταξύ των μεταβλητών εισόδου και της μεταβλητής εξόδου που έχουν μαθευτεί κατά τη διάρκεια της εκπαίδευσης. Για παράδειγμα, η παλινδρόμηση μπορεί να είναι χρήσιμη για την πρόβλεψη τιμών κατοικιών, με την τιμή του σπιτιού ως μεταβλητή εξόδου, ενώ ως μεταβλητές εισόδου θα μπορούσαν να είναι μεταβλητές όπως π.χ. η τοποθεσία, το μέγεθος σπιτιού κ.λ.π.
- **Ταξινόμηση (Classification):** χρησιμοποιείται όταν η μεταβλητή εξόδου είναι κατηγορική. Έτσι, είναι πιο χρήσιμο να ομαδοποιείται η μεταβλητή εξόδου σε μια κλάση (class). Εάν ο αλγόριθμος προσπαθεί να κατηγοριοποιήσει την είσοδο σε μόνο δύο διακριτές κλάσεις, τότε ονομάζεται δυαδική ταξινόμηση (binary classification). Η επιλογή μεταξύ

περισσότερων από δύο κλάσεων αναφέρεται ως ταξινόμηση πολλαπλών τάξεων (multi-class classification), όπως αυτή που χρησιμοποιείται στην παρούσα διπλωματική εργασία.

1.6.3.2 Μη-Επιβλεπόμενη Μάθηση

Σε αντίθεση με την παραπάνω κατηγορία μάθησης, υπάρχουν περιπτώσεις στις οποίες δεν είναι δυνατή η επισήμανση δεδομένων ή είναι εξαιρετικά δύσκολο να πραγματοποιηθεί. Για την προσέγγιση τέτοιου είδους περιπτώσεων, χρησιμοποιούνται τεχνικές εκμάθησης χωρίς επίβλεψη για την εύρεση «κρυφών» μοτίβων από το σύνολο δεδομένων. Αυτό το είδος μάθησης μπορεί να συγκριθεί με τη διαδικασία που λαμβάνει χώρα στον ανθρώπινο εγκέφαλο όταν καλείται να μάθει καινούριες πληροφορίες. Καθώς δεν υπάρχει ακριβώς αντιστοίχιση των δεδομένων εξόδου με τα δεδομένα εισόδου, η μη-επιβλεπόμενη μάθηση δεν έχει τη δυνατότητα να εφαρμοστεί άμεσα στην παλινδρόμηση ή στα προβλήματα ταξινόμησης. Ο στόχος της μη-επιβλεπόμενης μάθησης είναι να βρει την υποκείμενη δομή του συνόλου δεδομένων, να ομαδοποιήσει αυτά τα δεδομένα σύμφωνα με τις ομοιότητες που πιθανόν να έχουν και να αναπαραστήσει αυτό το σύνολο δεδομένων σε μια πιο συμπιεσμένη μορφή. Οι αλγόριθμοι μη-επιβλεπόμενης μάθησης μπορούν να κατηγοριοποιηθούν περαιτέρω σε προβλήματα **ομαδοποίησης** (clustering) και προβλήματα **συσχέτισης** (association) (Hastie, Tibshirani, & Friedman, 2009b):

- **Ομαδοποίηση (Clustering):** Η ομαδοποίηση είναι μια μέθοδος που επιχειρεί να ομαδοποιήσει αντικείμενα με βάση ομοιότητες που έχουν μεταξύ τους με τρόπο τέτοιο ώστε τα αντικείμενα με τις περισσότερες ομοιότητες να κατατάσσονται στην ίδια ομάδα και να έχουν λίγες ή και καθόλου ομοιότητες με αντικείμενα που ανήκουν σε μια άλλη ομάδα.
- **Συσχέτιση (Association):** Η συσχέτιση χρησιμοποιείται για την ανίχνευση σχέσεων μεταξύ μεταβλητών που υπάρχουν σε μια μεγάλη βάση δεδομένων. Χρησιμοποιείται ευρέως για στρατηγικό μάρκετινγκ, για παράδειγμα όταν μια ομάδα ανθρώπων αγοράζει ένα αντικείμενο X, είναι πολύ πιθανό να αγοράσει ένα αντικείμενο Y.

1.6.3.3 Ενισχυτική Μάθηση

Η ενισχυτική μάθηση είναι ένας υπο-τομέας της μηχανικής μάθησης που ασχολείται με το πρόβλημα της εκπαίδευσης ενός παράγοντα με σκοπό τη μεγιστοποίηση ενός σήματος ανταμοιβής (reward signal) ενώ ταυτόχρονα ενεργεί σε ένα περιβάλλον. Αυτή η μέθοδος βασίζεται στην «επιβράβευση» επιθυμητών συμπεριφορών ή/και την «τιμωρία» ανεπιθύμητων συμπεριφορών, έτσι, ένας παράγοντας ενισχυτικής μάθησης έχει την ικανότητα να μάθει μέσω δοκιμής και λάθους (trial and error). Ο κύριος στόχος της ενισχυτικής μάθησης είναι να

προσδιορίσει την καλύτερη αλληλουχία αποφάσεων που πρέπει να ακολουθήσει ο παράγοντας για να λύσει ένα πρόβλημα μεγιστοποιώντας ταυτόχρονα μια μακροπρόθεσμη ανταμοιβή (reward). Αυτός είναι ο λόγος που είναι πρωτίστως εφαρμόζεται για σχεδιασμό κίνησης, δυναμικό σχεδιασμό διαδρομής, βελτιστοποίηση φωτεινών σηματοδοτών, πολιτικές εκμάθησης βάσει δεδομένων σεναρίων για αυτοκινητόδρομους κ.λ.π. Χαρακτηριστικό παράδειγμα επάρκειας αυτής της μεθόδου είναι η χρήση της για στάθμευση που μπορεί να επιτευχθεί με την εκμάθηση πολιτικών αυτόματης στάθμευσης.

1.6.4 SHAP (SHapley Additive exPlanations)

Το SHAP (SHapley Additive exPlanations) είναι μια μέθοδος που χρησιμοποιείται στη μηχανική μάθηση για να εξηγήσει το ουτρυτ ενός μοντέλου αποδίδοντας το αποτέλεσμα πρόβλεψης σε κάθε διαφορετικό χαρακτηριστικό. Βασίζεται στη θεωρία παιγνίων και τις τιμές Shapley, οι οποίες αποδίδουν μια τιμή σε κάθε χαρακτηριστικό λαμβάνοντας υπόψη τη συμβολή του στην πρόβλεψη όταν συνδυάζεται με άλλα χαρακτηριστικά. Οι τιμές SHAP παρέχουν μια σαφή κατανόηση του τρόπου με τον οποίο κάθε χαρακτηριστικό επηρεάζει τις προβλέψεις του μοντέλου και μπορούν να βοηθήσουν στην ερμηνεία πολύπλοκων μοντέλων όπως οι μέθοδοι ensemble, τα νευρωνικά δίκτυα και τα gradient boosting machines. Η ανάλυση SHAP είναι χρήσιμη για τους εξής λόγους (Lundberg et al., 2020; Lundberg & Lee, 2017; Štrumbelj & Kononenko, 2014):

- **Ερμηνευσιμότητα:** Η ανάλυση SHAP παρέχει «διαισθητικές» εξηγήσεις για μεμονωμένες προβλέψεις, βοηθώντας στην κατανόηση της συμπεριφοράς του μοντέλου τόσο σε τοπικό όσο και σε συνολικό (global) επίπεδο.
- **Σημασία χαρακτηριστικών (feature importance):** Με την ποσοτικοποίηση της συνεισφοράς κάθε χαρακτηριστικού στο ουτρυτ του μοντέλου, η ανάλυση SHAP βοηθά στον εντοπισμό των χαρακτηριστικών με τη μεγαλύτερη επιρροή για την πραγματοποίηση προβλέψεων. Αυτές οι πληροφορίες μπορούν να καθοδηγήσουν την επιλογή και την επεξεργασία των χαρακτηριστικών, καθώς και τη βελτιστοποίηση του μοντέλου.
- **Εντοπισμός σφαλμάτων μοντέλου (model debugging):** Οι τιμές SHAP μπορούν να αποκαλύψουν απροσδόκητες σχέσεις μεταξύ χαρακτηριστικών και προβλέψεων, βοηθώντας στον εντοπισμό προκαταλήψεων, σφαλμάτων ή περιοχών βελτίωσης του μοντέλου.
- **Αξιοπιστία:** Οι αμερόληπτες εξηγήσεις που παρέχονται από την ανάλυση SHAP ενισχύουν την αξιοπιστία των μοντέλων μηχανικής μάθησης, καθιστώντας τα πιο αποδεκτά για χρήση σε κρίσιμες εφαρμογές όπως η υγειονομική περίθαλψη.

Αξίζει να σημειωθεί πως οι τιμές SHAP δεν αλλάζουν όταν αλλάζει το μοντέλο εκτός εάν αλλάξει η συνεισφορά ενός χαρακτηριστικού. Αυτό σημαίνει ότι οι τιμές SHAP παρέχουν μια σταθερή ερμηνεία της συμπεριφοράς του μοντέλου, ακόμη και όταν αλλάζουν η αρχιτεκτονική ή οι παράμετροι του μοντέλου. Συνολικά, οι τιμές SHAP παρέχουν έναν συνεπή και αντικειμενικό τρόπο απόκτησης πληροφοριών σχετικά με τον τρόπο με τον οποίο ένα μοντέλο μηχανικής μάθησης κάνει προβλέψεις και ποια χαρακτηριστικά έχουν τη μεγαλύτερη επιρροή (Awan, 2023).

2 ΣΚΟΠΟΣ

Όπως αναφέρθηκε, στην Ευρωπαϊκή Ένωση την περίοδο εκπόνησης της εργασίας ήταν εγκεκριμένα 4 εμβόλια κατά του κορωνοϊού, το εμβόλιο της Johnson & Johnson, το εμβόλιο της Pfizer-BioNTech, το εμβόλιο της AstraZeneca και το εμβόλιο της Moderna. Η χορήγηση αυτών των εμβολίων στο κοινό έχει προκαλέσει σημαντικό αριθμό παρενεργειών μεταξύ των οποίων είναι πονοκέφαλος, τοπικό οίδημα και πόνος, κόπωση, μυοκαρδίτιδα (Mevorach et al., 2021; Witberg et al., 2021), διάφορα θρομβοεμβολικά επεισόδια (Bhattacharjee & Banerjee, 2020; Greinacher et al., 2021; Suvvari et al., 2021), κ.ά. Παρ' όλα αυτά συγκεντρωτικές μελέτες που εξετάζουν και αξιολογούν την ασφάλεια των εμβολίων δείχνουν πως το προφίλ τους είναι ασφαλές βραχυπρόθεσμα (Q. Wu et al., 2021). Από τις αρχές του 2020, δεδομένα που αφορούν τις παρενέργειες αυτών των εμβολίων κατά της Covid-19 ξεκίνησαν να καταγράφονται και να συγκεντρώνονται στη βάση δεδομένων EudraVigilance, η οποία όπως αναφέρθηκε περιλαμβάνει δεδομένα φαρμακοεπαγρύπνησης για τη διαχείριση και ανάλυση ύποπτων παρενεργειών ποικίλων φαρμάκων και εμβολίων (Postigo et al., 2018).

Στόχος λοιπόν της συγκεκριμένης διπλωματικής εργασίας είναι η ανάπτυξη ενός αλγορίθμου μηχανικής μάθησης που θα προβλέπει τη σοβαρότητα των παρενεργειών των εμβολίων που βρίσκονται σε χρήση με βάση πρωτογενή δεδομένα που αντλήθηκαν από τη βάση δεδομένων EudraVigilance. Η ανάπτυξη ενός αποτελεσματικού αλγορίθμου ταξινόμησης για την αξιολόγηση της σοβαρότητας των παρενεργειών που σχετίζονται με τα εμβόλια κατά της COVID-19 μπορεί να συνεισφέρει στην έγκαιρη ανίχνευση και παρακολούθηση, καθώς δύναται να εντοπίσει ανεπιθύμητες ενέργειες, επιτρέποντας στους επαγγελματίες υγείας να επέμβουν άμεσα. Παράλληλα, με την ταξινόμηση της σοβαρότητας, οι κλινικοί γιατροί θα μπορούν να δώσουν προτεραιότητα στην παρακολούθηση και τη φροντίδα των ασθενών που παρουσιάζουν πιο σοβαρά συμπτώματα, βελτιστοποιώντας την κατανομή των πόρων. Τα συγκεντρωτικά δεδομένα αυτής της ταξινόμησης σοβαρότητας μπορούν να φανούν χρήσιμα για τη δημόσια υγεία και την αξιολόγηση του προφίλ ασφαλείας των εμβολίων. Συνοπτικά, ένας καλά σχεδιασμένος αλγόριθμος ταξινόμησης θα μπορούσε να συμβάλλει στην καλύτερη φροντίδα του ασθενούς, στη λήψη ενημερωμένων αποφάσεων και στη συνολική ασφάλεια του εμβολίου.

3 ΜΕΘΟΔΟΙ

3.1 ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ – ΓΕΝΙΚΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ

Τα δεδομένα που χρησιμοποιήθηκαν στην παρούσα δυπλωματική αντλήθηκαν από τη βάση δεδομένων EudraVigilance, η οποία αναλύθηκε στην εισαγωγική ενότητα. Συγκεκριμένα, αντλήθηκαν δεδομένα για τα εμβόλια κατά της CoViD-19 που ήταν εγκεκριμένα στην Ευρωπαϊκή Ένωση το 2021, δηλαδή το εμβόλιο της Moderna, το εμβόλιο της Pfizer, το εμβόλιο της AstraZeneca και το εμβόλιο της Janssen. Η αναζήτηση στη EudraVigilance πραγματοποιήθηκε με τον τρόπο που περιεγράφηκε παραπάνω, δηλαδή επιλέγοντας από το μενού αναζήτησης τη δραστική ουσία ή το εμβόλιο που μας ενδιαφέρει, στην προκειμένη τα τέσσερα εμβόλια που αναφέρθηκαν. Καθώς ο όγκος κάθε υποσυνόλου δεδομένων για κάθε εμβόλιο ήταν αρκετά μεγάλος, η λήψη των δεδομένων πραγματοποιήθηκε σε τμήματα (μικρότερα υποσύνολα) έχοντας εκμεταλλευτεί τις δυνατότητες που παρέχει η τελευταία καρτέλα του web report της βάσης, όπως αναλύθηκε νωρίτερα, δηλαδή τη χρήση διάφορων φίλτρων. Στη συνέχεια πραγματοποιήθηκε συγχώνευση όλων των υποσυνόλων δεδομένων ώστε να φτιαχτεί το τελικό σύνολο δεδομένων που χρησιμοποιήθηκε στην εκπαίδευση του αλγορίθμου.

Παραδειγματικά, τρέχοντας την αναζήτηση δεδομένων για το εμβόλιο της Pfizer, η εικόνα που εμφανίζει η τελευταία καρτέλα «Line Listing» του web report της EudraVigilance, είναι η εξής:

Line Listing Report Time run: 20/01/2024 16:15:14													
EU Local Number	EV Gateway Report Date	Report Type	Primary Source Qualification	Primary Source Country for Reporting Purposes	Literature Reference	Patient Age Group (as per reporter)	Patient Sex	Patient Child Report	Reaction/Int PT Duration - Outcome - Seriousness Criteria	Suspect/Interacting Drug List (Drug Char - Indication PT - Action taken - Duration - Dose - Route)	Concomitant/Not Administered Drug List (Drug Char - Indication PT - Action taken - Duration - Dose - Route)	ICSR Part	
EU-EC-10016622686	15/01/2024	Spontaneous	Healthcare Professional	Non-European Economic Area	Not available	18-64 Years	Female	Yes	Immunization failure (n/a - Unknown - [1d - n/a - n/a])	[TOZINAMERAN] (S - COVID-19 immunisation - Not applicable - [1d - n/a - n/a])	Not reported	ICSR	
EU-EC-10016622776	15/01/2024	Spontaneous	Healthcare Professional	Non-European Economic Area	Not available	18-64 Years	Not Specified	Female	No	Subarachnoid hemorrhage (n/a - Unknown - [1d - n/a - n/a])	[COMBINATY TOZINAMERAN] (S - COVID-19 immunisation - Not applicable - [1d - n/a - n/a])	Not reported	ICSR
EU-EC-10016622796	15/01/2024	Spontaneous	Non Healthcare Professional	Non-European Economic Area	Not available	55-65 Years	Male	No	COVID-19 (2d - Unknwon - Other Medically Important Condition), Recovered/Resolved, Caused/Associated/Contributed/Hospitalisation, Other Medically important Condition), Immunization failure (2d - Recovered/Resolved - Other Medically Important Condition), Caused/Prlonged Hospitalisation, Other Medically Important Condition)	[TOZINAMERAN] (S - COVID-19 immunisation - Not applicable - [1d - n/a - n/a])	Not reported	ICSR	
EU-EC-10016622805	15/01/2024	Spontaneous	Non Healthcare Professional	Non-European Economic Area	Not available	65-85 Years	Not Specified	Female	No	COVID-19 (n/a - Unknown - Other Medically Important Condition), Drug ineffective (n/a - Recovered/Resolved - Other Medically Important Condition), Interactions with other products (n/a - Associated/Resolved - Other Medically Important Condition)	[SPERIFAX] (B - COVID-19 immunisation - Not applicable - [n/a - n/a - n/a]), [TOZINAMERAN] (S - COVID-19 immunisation - Not applicable - [n/a - n/a - n/a]), [FAMTOZINAMERAN TOZINAMERAN, FAMTOZINAMERAN TOZINAMERAN, FAMTOZINAMERAN] (S - COVID-19 immunisation - Not applicable - [n/a - n/a - n/a])	Not reported	ICSR
EU-EC-10016622819	15/01/2024	Spontaneous	Non Healthcare Professional	Non-European Economic Area	Not available	Not Specified	Not Specified	Female	No	Immunotherapy (n/a - Unknown - Other Medically Important Condition)	[TOZINAMERAN] (TOZINAMERAN) (S - COVID-19 immunisation - Not applicable - [n/a - n/a - n/a])	Not reported	ICSR
EU-EC-10016622820	15/01/2024	Spontaneous	Non Healthcare Professional	Non-European Economic Area	Not available	18-64 Years	Not Specified	Male	No	COVID-19 (n/a - Unknown - Other Medically Important Condition), Drug ineffective (n/a - Unknown - Other Medically Important Condition), Interactions with vaccine products (n/a - Unknown - Other Medically Important Condition)	[COMBINATY TOZINAMERAN] (S - COVID-19 immunisation - Not applicable - [1d - n/a - n/a]), [COV-19 VACCINE ASTRAZENECA CHADOLIX NOV-19] (B - COVID-19 VACCINE ASTRAZENECA CHADOLIX NOV-19), [COVID-19 VACCINE ASTRAZENECA CHADOLIX NOV-19] (S - COVID-19 immunisation - Not applicable - [1d - n/a - n/a]), [SPERIEVA ELASOMERAN] (S - COVID-19 immunisation - Not applicable - [1d - n/a - n/a]), [TOZINAMERAN] (S - COVID-19 immunisation - Not applicable - [1d - n/a - n/a])	Not reported	ICSR
EU-EC-10016622825	15/01/2024	Spontaneous	Non Healthcare Professional	Non-European Economic Area	Not available	65-85 Years	Not Specified	Female	No	Drug ineffective (n/a - Unknown - Other Medically Important Condition)	[TOZINAMERAN] (S - COVID-19 immunisation - Not applicable - [n/a - n/a - n/a])	Not reported	ICSR
EU-EC-10016622842	15/01/2024	Spontaneous	Non Healthcare Professional	Non-European Economic Area	Not available	65-85 Years	Not Specified	Male	No	COVID-19 (n/a - Unknown - Other Medically Important Condition)	[LUTANOPROST] (LATANOPROST) (C - n/a - n/a - [n/a - n/a - n/a]), [LOSARTAN, LOSARTAN POTASSIUM] (C - n/a - n/a - [n/a - n/a - n/a])	METHODS	ICSR
EU-EC-10016622842	15/01/2024	Spontaneous	Non Healthcare Professional	Non-European Economic Area	Not available	65-85 Years	Not Specified	Female	No	COVID-19 (n/a - Unknown - Other Medically Important Condition)	[TOZINAMERAN] (TOZINAMERAN) (S - COVID-19 immunisation - Not applicable - [1d - n/a - n/a]), [FAMTOZINAMERAN, FAMTOZINAMERAN] (S - COVID-19 immunisation - Not applicable - [1d - n/a - n/a]), [COLECALCIPEROL] (C - n/a - n/a - [n/a - n/a - n/a]), [POLICE ACTID] (C - Rheumatoid arthritis - n/a - [n/a - Img - Oral use])	METHODS	ICSR

Εικόνα 22. Line Listing Report στη EudraVigilance για δεδομένα του εμβολίου της Pfizer

Μετά το αποτέλεσμα της αναζήτησης προκειμένου να αποθηκεύσουμε τα δεδομένα επιλέγουμε εξαγωγή (export) και τη μορφή του αρχείου της επιλογής μας. Τα δεδομένα που χρησιμοποιήθηκαν εξήχθησαν σε μορφή .tsv.

Έτσι λοιπόν η μορφή των δεδομένων αν τα ανοίξουμε για παράδειγμα με Microsoft Excel, είναι η εξής:

	A	B	C	D	E	F	G	H	I	J	K	
1	EU Local N Report Ty: EV Gateway Receipt Primary Si:Primary St:Literature Patient As:Patient Ch:Patient Se:Reaction List PT:(Duration: "4C" Outcome: - Seriousness Criteria)											
2	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Vertigo (n/a - Unknown -)					
3	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Fibrin D dimer increased (n/a - Not Recovered/Not Resolved -), Hypertension (n/a - Not Recovered/Not Resolved -), Telangiectasia (n/a - Not Recovered/Not Resolved -)					
4	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Vertigo (n/a - Unknown -)					
5	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Bone pain (1d - Recovered/Resolved -), Fatigue (1d - Recovered/Resolved -), Headache (1d - Recovered/Resolved -), Pyrexia (1d - Recovered/Resolved -)					
6	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Back pain (n/a - Recovering/Resolving -), Chills (n/a - Recovering/Resolving -), Headache (n/a - Recovering/Resolving -), Hyperhidrosis (n/a - Recovering/Resolving -)					
7	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Pain in extremity (n/a - Recovering/Resolving -), Urinary tract infection (n/a - Recovering/Resolving -)					
8	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Menstrual disorder (n/a - Not Recovered/Not Resolved -)					
9	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Injection site pain (n/a - Recovered/Resolved/With Sequela -), Injection site papule (n/a - Recovered/Resolved/With Sequela -), Lymphadenopathy (n/a - Recovered/Resolved -)					
10	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Lymphadenopathy (n/a - Recovered/Resolved -)					
11	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Vertigo (n/a - Unknown -), Fatigue (n/a - Not Recovered/Not Resolved -)					
12	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Vertigo (n/a - Not Recovered/Not Resolved -)					
13	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Arthralgia (n/a - Not Recovered/Not Resolved -), Headache (n/a - Not Recovered/Not Resolved -), Musculoskeletal pain (n/a - Not Recovered/Not Resolved -)					
14	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Axillary pain (n/a - Not Recovered/Not Resolved -), Pain in extremity (n/a - Not Recovered/Not Resolved -)					
15	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Arthralgia (n/a - Unknown -), Asthenia (n/a - Unknown -), Pyrexia (n/a - Unknown -)					
16	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Cutaneous vasculitis (n/a - Not Recovered/Not Resolved -), Pruritus (n/a - Not Recovered/Not Resolved -), Urticaria (n/a - Not Recovered/Not Resolved -)					
17	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Pyrexia (n/a - Unknown -), Injection site pain (n/a - Unknown -)					
18	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Chest pain (n/a - Unknown -), Deep vein thrombosis (n/a - Unknown -), Hypertension (n/a - Unknown -), Paraesthesia (n/a - Unknown -), Tachycardia (n/a - Unknown -), Vertigo (n/a - Not Recovered/Not Resolved -)					
19	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Lymphadenopathy (n/a - Recovering/Resolving -), Nausea (n/a - Recovering/Resolving -)					
20	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Pyrexia (n/a - Unknown -)					
21	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Pyrexia (n/a - Unknown -), Pyrexia (n/a - Unknown -)					
22	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Dyspnoea (n/a - Unknown -), Pyrexia (n/a - Unknown -)					
23	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Cystitis (n/a - Recovering/Resolving -), Lymphadenopathy (n/a - Recovering/Resolving -), Pyrexia (n/a - Recovering/Resolving -)					
24	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Asymmetry (2d - Recovered/Resolved -), Fatigue (2d - Recovered/Resolved -), Nausea (2d - Recovered/Resolved -), Pyrexia (2d - Recovered/Resolved -)					
25	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Asymmetry (2d - Recovered/Resolved -), Fatigue (2d - Recovered/Resolved -), Nausea (2d - Recovered/Resolved -), Pyrexia (2d - Recovered/Resolved -)					
26	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Pyrexia (n/a - Unknown -), Vertigo (n/a - Recovering/Resolving -), Myalgia (n/a - Recovering/Resolving -), Pyrexia (n/a - Recovering/Resolving -), Vaccination site erythema (n/a - Recovery -)					
27	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Erythema (n/a - Not Recovered/Not Resolved -)					
28	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Electric shock sensation (n/a - Recovering/Resolving -)					
29	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Dermatitis bullous (n/a - Recovering/Resolving -), Fatigue (n/a - Recovering/Resolving -), Musculoskeletal pain (n/a - Recovering/Resolving -)					
30	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	More than 65 years Not Specified No	Female	Myalgia (2d - Recovered/Resolved -), Vaccination failure (2d - Recovered/Resolved -)					
31	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Asymmetry (2d - Recovered/Resolved -), Fatigue (2d - Recovered/Resolved -), Pyrexia (2d - Recovered/Resolved -)					
32	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Eructation (n/a - Recovered/Resolved -), Vertigo (n/a - Recovered/Resolved -)					
33	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Epigastric discomfort (n/a - Not Recovered/Not Resolved -)					
34	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Musculoskeletal pain (7d - Recovered/Resolved -)					
35	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Fatigue (n/a - Recovering/Resolving -)					
36	EU-EC-100 Spontane-	12/11/2021 0:00	Healthcare European	Not available	18-64 Year Not Specified No	Female	Arthralgia (n/a - Not Recovered/Not Resolved -), Asthenia (n/a - Not Recovered/Not Resolved -), Body temperature increased (n/a - Not Recovered/Not Resolved -), Urticaria (2d - Recovered/Resolved -)					

Εικόνα 23. Μορφή δεδομένων όπως λαμβάνονται από τη EudraVigilance όταν διαβάζονται με Microsoft Excel

Παρατηρούμε πως τα δεδομένα σε αυτή τη μορφή είναι εξαιρετικά δύσκολο να «διαβαστούν» από έναν αλγόριθμο, καθώς δεν υπάρχει σαφής διαχωρισμός των παραμέτρων. Έτσι λοιπόν αποφασίσαμε να πραγματοποιήσουμε parsing των δεδομένων με τη χρήση της γλώσσας προγραμματισμού PHP ώστε να μετατραπούν σε μια περισσότερο αναγνώσιμη και αναγνωρίσιμη μορφή από έναν αλγόριθμο.

3.2 PARSING ΤΟΥ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ

3.2.1 Γλώσσα PHP

Η PHP είναι μια ανοιχτού κώδικα γλώσσα που χρησιμοποιούν πολλοί προγραμματιστές για ανάπτυξη ιστοσελίδων. Είναι επίσης μια γλώσσα γενικής χρήσης για διάφορων ειδών εργασίες, συμπεριλαμβανομένης και της δημιουργίας Graphical User Interfaces (GUI). Το ακρωνύμιο PHP αρχικά σήμαινε 'Personal Homepage' αλλά στη συνέχεια διαμορφώθηκε στο 'Hypertext Preprocessor', ενώ η πρώτη της έκδοση ξεκίνησε περίπου 30 χρόνια πριν. Κατά κύριο λόγο η PHP

χρησιμοποιείται για τη δημιουργία διακομιστών παγκόσμιου ιστού (web servers), τρέχει στο λογισμικό πλοιόγησης αλλά μπορεί να τρέξει και στη γραμμή εντολών. Κάποια πλεονεκτήματα της χρήσης της PHP είναι πως δεν απαιτείται συγκεκριμένο Λειτουργικό Σύστημα γιατί μπορεί να τρέξει σε οποιαδήποτε πλατφόρμα Mac, Windows ή και Linux, είναι ανοιχτού κώδικα επομένως μπορεί οποιοσδήποτε να χτίσει περαιτέρω σε αυτόν αν θέλει, είναι εύκολη στη χρήση και το βασικότερο για εμάς ότι συγχρονίζεται με όλες τις βάσεις δεδομένων, σχεσιακές και μη (Kolade, 2021).

Στην παρούσα διπλωματική η PHP γλώσσα αποτέλεσε σημαντικό εργαλείο για την κατασκευή του αρχείου input που χρησιμοποιήθηκε για την εκπαίδευση του αλγορίθμου, καθώς αυτό προέκυψε από τον συνδυασμό κατάλληλων queries από την τοπική βάση δεδομένων που είχε δημιουργηθεί με τη χρήση της MySQL, όπως θα αναλυθεί παρακάτω.

3.2.2 Parser

Το πρώτο βήμα για τη ‘μετάφραση’ των δεδομένων που αντλήθηκαν κατευθείαν από τη EudraVigilance ήταν η δημιουργία ενός parser με τη χρήση της γλώσσας PHP, όπως αναφέρθηκε. Η λογική με την οποία πραγματοποιήθηκε το parsing των δεδομένων ήταν προσανατολισμένη στη δημιουργία μιας τοπικής βάσης δεδομένων (database) με χρήση της γλώσσας προγραμματισμού SQL στην οποία θα αποθηκεύονταν τα δεδομένα σε πίνακες (tables). Έτσι λοιπόν, ο parser που φτιάχτηκε ουσιαστικά είχε ως αποτέλεσμα τη δημιουργία τόσων αρχείων όσων και των πινάκων που θα αποτελούσαν στη συνέχεια τη βάση δεδομένων.

Πιο συγκεκριμένα, το αποτέλεσμα του parsing δημιούργησε τα εξής αρχεία που στη συνέχεια θα αποτελούσαν τους πίνακες της βάσης δεδομένων:

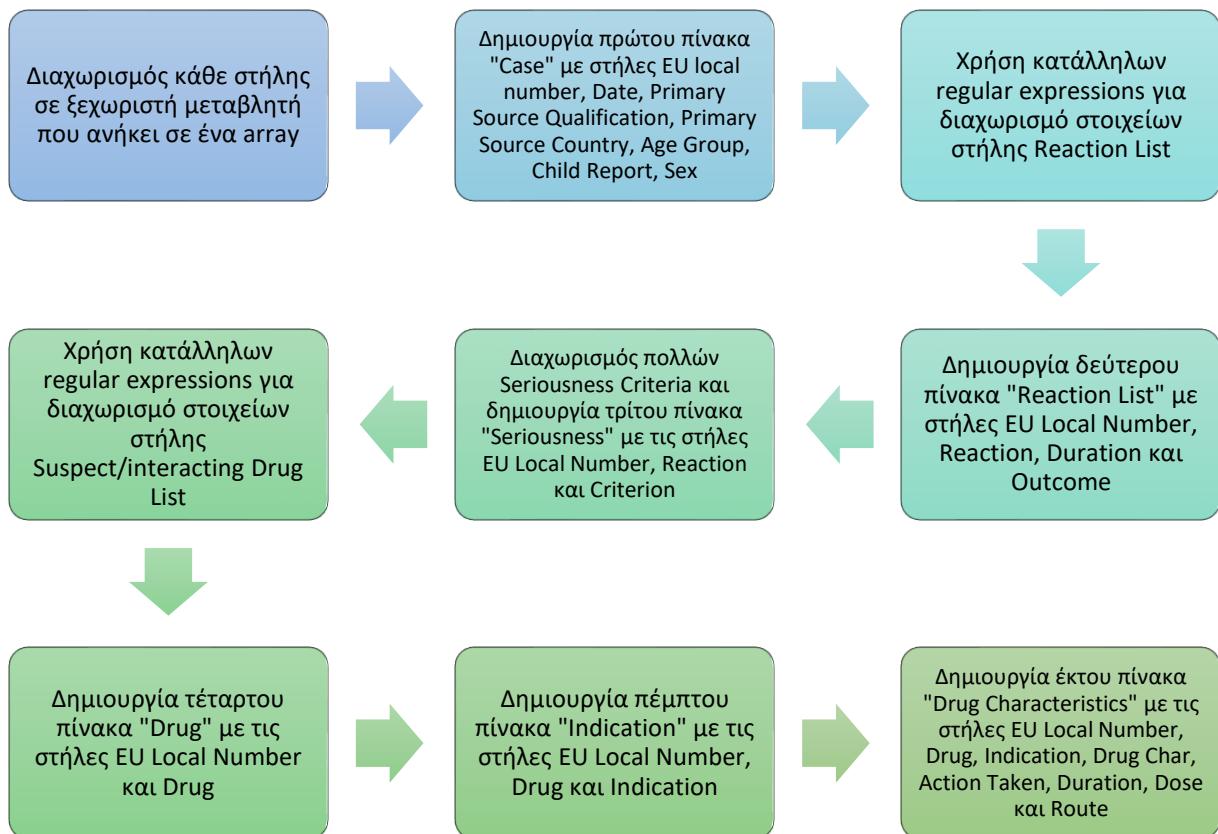
1. Case: περιέχει τις πληροφορίες που σχετίζονται με τον ασθενή όπως αριθμό ταυτότητας (ID number), φύλο, ηλικία κλπ.
2. Drug: περιέχει πληροφορίες σχετικά με τη δραστική ουσία που χορηγήθηκε στον ασθενή, δηλαδή το ένα εκ των τεσσάρων εμβολίων υπό διερεύνηση, καθώς και πιθανά φάρμακα που χορηγήθηκαν παράλληλα με σκοπό την αντιμετώπιση κάποιας αντίδρασης στο εμβόλιο ή ακόμα και φάρμακα που μπορεί να χορηγούνταν ως αγωγή σε χρόνια νοσήματα.
3. Drug Characteristics: περιέχει πληροφορίες για τις δραστικές ουσίες ή τα εμβόλια της προηγούμενης κατηγορίας συνοδευόμενες από περισσότερες λεπτομέρειες όπως την οδό χορήγησης, τη διάρκεια χορήγησης, το αν η ουσία αυτή αξιολογείται ως ύποπτη για την

αναφερόμενη αντίδραση ή απλά συμπληρωματική για κάποιο ενδεχόμενο υποκείμενο νόσημα κ.ά.

4. Indication: περιέχει πληροφορίες σχετικά με την ένδειξη για την οποία έχει χορηγηθεί είτε ένα από τα τέσσερα υπό μελέτη εμβόλια, είτε για φάρμακα που χορηγήθηκαν παράλληλα.
5. Reaction List: περιέχει πληροφορίες σχετικά με την αντίδραση ή τις αντιδράσεις που είχε ο εκάστοτε ασθενής μετά τη χορήγηση ενός δεδομένου εμβολίου. Συγκεκριμένα, περιέχει την πληροφορία του αριθμού ταυτότητας του ασθενούς, ώστε να μπορεί να γίνει η αντιστοίχιση αργότερα, την αντίδραση ή τις αντιδράσεις για τις οποίες έχει γίνει και η αναφορά στη βάση, τη διάρκεια της αντίδρασης (αν υπάρχει) και τέλος μια αξιολόγηση της σοβαρότητας της κάθε αναφερόμενης αντίδρασης π.χ. αν αντιμετωπίστηκε ή αν προκάλεσε κάποια αναπηρία.
6. Seriousness: αυτό το υποσύνολο δεδομένων είναι και το πιο ουσιώδες για το σκοπό αυτής της διπλωματικής καθώς περιλαμβάνει πληροφορίες που συσχετίζουν το ID του ασθενούς, την αναφερόμενη αντίδραση και πως αυτή αξιολογήθηκε με κριτήρια σοβαρότητας σύμφωνα με την ορολογία MedDRA. Οι τιμές αξιολόγησης της σοβαρότητας που περιέχονται σε αυτό το σύνολο δεδομένων αποτελούν και τις τιμές της τελική κλάσης στην οποία κατηγοριοποιούνται τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση του αλγορίθμου κατηγοριοποίησης.

Παρακάτω παρουσιάζεται σχηματικά η ροή εργασίας του parser:

1. Δημιουργία και άνοιγμα αρχείων που θα αποτελέσουν τους πίνακες της Βάσης Δεδομένων
2. Για κάθε αποθηκευμένο αρχείο από τη βάση:
 - Για κάθε γραμμή του αρχείου:



Εικόνα 24. Συνοπτικό σχεδιάγραμμα ροής εργασίας του parser

3.3 ΔΗΜΙΟΥΡΓΙΑ ΒΑΣΗΣ ΔΕΔΟΜΕΝΩΝ

Για την αποτελεσματικότερη διαχείριση και επεξεργασία των δεδομένων αυτής της διπλωματικής, κατασκευάστηκε τοπική βάση δεδομένων με τη χρήση της γλώσσας SQL. Σημειώνεται πως οι πίνακες αυτής της βάσης, συμπληρώθηκαν με το περιεχόμενο των αρχείων που δημιουργήθηκαν μέσω του parser που περιγράφηκε στην προηγούμενη ενότητα. Η τοπική βάση δεδομένων δημιουργήθηκε και συμπληρώθηκε μέσω λογισμικού Linux και με τα κατάλληλα queries που

αντλήθηκαν απευθείας από το MySQL Workbench. Παρακάτω αναλύονται περαιτέρω αυτές οι έννοιες.

3.3.1 Γλώσσα SQL

Η γλώσσα SQL, και τα συστήματα σχεσιακών βάσεων δεδομένων που βασίζονται σε αυτήν, είναι από τις πιο θεμελιώδεις τεχνολογίες στη βιομηχανία των υπολογιστών σήμερα. Κατά τις τελευταίες δεκαετίες, η δημοτικότητα της SQL έχει αυξηθεί κατακόρυφα και παραμένει σήμερα ως η τυπική υπολογιστική γλώσσα βάσης δεδομένων (Groff et al., 2002). Συγκεκριμένες ιστοσελίδες – εφαρμογές που χρησιμοποιούν την SQL για την αποθήκευση και επεξεργασία των δεδομένων τους είναι για παράδειγμα το Facebook, το Spotify και η Revolut (developNET, 2023).

Η SQL πρόκειται για ένα εργαλείο για την οργάνωση, τη διαχείριση καθώς και την ανάκτηση δεδομένων που είναι αποθηκευμένα σε μια βάση δεδομένων ενός υπολογιστή. Το όνομα "SQL" είναι συντομογραφία του Structured Query Language. Όπως υποδηλώνει και το όνομα, η SQL είναι μια γλώσσα υπολογιστή που χρησιμοποιείται για να μπορεί ο χρήστης να αλληλοεπιδράσει με μια βάση δεδομένων. Στην πραγματικότητα, η SQL λειτουργεί με έναν συγκεκριμένο τύπο βάσης δεδομένων, που ονομάζεται σχεσιακή βάση δεδομένων. Το πρόγραμμα του υπολογιστή που ελέγχει τη βάση δεδομένων καλείται Σύστημα Διαχείρισης Βάσεων Δεδομένων (Database Management System (DBMS)). Όταν απαιτείται η ανάκτηση δεδομένων από μια βάση δεδομένων, χρησιμοποιείται η γλώσσα SQL για να δημιουργηθεί ένα αίτημα. Το DBMS επεξεργάζεται αυτό το αίτημα SQL, ανακτά τα ζητούμενα δεδομένα και επιστρέφει την απάντηση. Αυτή η διαδικασία αίτησης δεδομένων από μια βάση δεδομένων και της επιστροφής των αποτελεσμάτων ονομάζεται database query —εξ ου και το όνομα Structured Query Language (Groff et al., 2002).

Η SQL χρησιμοποιείται για τον έλεγχο όλων των λειτουργιών που εν δυνάμει παρέχει ένα DBMS στους χρήστες του, συμπεριλαμβανομένων των παρακάτω (Groff et al., 2002):

- Ορισμός δεδομένων (data definition): Η SQL επιτρέπει σε έναν χρήστη να ορίσει τη δομή και την οργάνωση των αποθηκευμένων δεδομένων, καθώς και τις σχέσεις μεταξύ των αποθηκευμένων στοιχείων των δεδομένων.
- Ανάκτηση δεδομένων (data retrieval): Η SQL επιτρέπει σε έναν χρήστη ή σε ένα πρόγραμμα εφαρμογής να ανακτήσει αποθηκευμένα δεδομένα από τη βάση δεδομένων και να τα χρησιμοποιήσει.
- Διαχείριση δεδομένων (data manipulation): Η SQL επιτρέπει σε έναν χρήστη ή σε ένα πρόγραμμα εφαρμογής να ενημερώσει τη βάση δεδομένων προσθέτοντας νέα δεδομένα,

αφαιρώντας παλιά δεδομένα, καθώς και τροποποιώντας προηγουμένως αποθηκευμένα δεδομένα.

- Έλεγχος πρόσβασης (access control): Η SQL μπορεί να χρησιμοποιηθεί για να περιορίσει την ικανότητα ενός χρήστη να ανακτά, να προσθέτει και να τροποποιεί δεδομένα, προστατεύοντας τα αποθηκευμένα δεδομένα από μη εξουσιοδοτημένη πρόσβαση.
- Κοινή χρήση δεδομένων (data sharing): Η SQL χρησιμοποιείται για τον συντονισμό της κοινής χρήσης δεδομένων από ταυτόχρονους χρήστες, διασφαλίζοντας ότι δεν παρεμβαίνουν μεταξύ τους.
- Ακεραιότητα δεδομένων (data integrity): Η SQL θέτει όρια όσον αφορά την ακεραιότητα στη βάση δεδομένων, προστατεύοντάς την από καταστροφή λόγω ασυνεπών ενημερώσεων ή πιθανά λάθη του συστήματος.

Η SQL είναι επομένως μια ολοκληρωμένη γλώσσα για τον έλεγχο και την αλληλεπίδραση με ένα σύστημα Διαχείρισης Βάσεων Δεδομένων.

Κατά δεύτερον, η SQL δεν είναι στην πραγματικότητα μια πλήρης γλώσσα υπολογιστή όπως για παράδειγμα η COBOL, η C, η C++ ή η Java. Η SQL δεν περιέχει για παράδειγμα if statements για σενάρια δοκιμής και δεν περιέχει do ή for statements για τον έλεγχο ροής προγράμματος. Αντίθετα, η SQL είναι περισσότερο μια *υπο-γλώσσα* βάσης δεδομένων, που αποτελείται από περίπου σαράντα statements εξειδικευμένων για λειτουργίες διαχείρισης βάσεων δεδομένων. Αυτά τα SQL statements μπορούν να ενσωματωθούν και σε άλλη γλώσσα, όπως η C ή η PHP που χρησιμοποιήθηκε στην παρούσα διπλωματική, ώστε να επεκτείνει τη χρήση αυτής της γλώσσας με σκοπό την πρόσβαση στη βάση δεδομένων. Εναλλακτικά, μπορούν να αποσταλούν σε ένα σύστημα Διαχείρισης Βάσης Δεδομένων για επεξεργασία, μέσω ενός cell level interface από μια γλώσσα όπως η C, η C++ ή η Java. Τέλος, η SQL δεν είναι μια ιδιαίτερα δομημένη γλώσσα, ειδικά όταν συγκρίνεται με υψηλά δομημένες γλώσσες όπως η C, η Pascal ή η Java (Groff et al., 2002).

3.3.2 MySQL DBMS

Όπως αναφέρθηκε και νωρίτερα, ένα σύστημα διαχείρισης βάσεων δεδομένων (DBMS), είναι ένα εργαλείο λογισμικού που βοηθά στην οργάνωση, ανάκτηση, και διασταύρωση πληροφοριών. Αυτή τη στιγμή υπάρχει μεγάλος αριθμός τέτοιων συστημάτων διαθέσιμα, μερικά από τα οποία είναι: το Oracle, το Sybase, το Microsoft Access και το PostgreSQL. Αυτά τα συστήματα βάσης δεδομένων είναι ισχυρές, πλούσιες σε χαρακτηριστικά εφαρμογές λογισμικού, ικανές για οργάνωση και αναζήτηση εκατομμυρίων εγγραφών σε πολύ υψηλές ταχύτητες. Σε αυτήν την κατηγορία συστημάτων εντάσσεται και η MySQL. Η MySQL είναι ένα σχεσιακό σύστημα

διαχείρισης βάσεων δεδομένων (Relational Database Management System (RDBMS)) υψηλής απόδοσης, με δυνατότητα παράλληλης επεξεργασίας, χτισμένο γύρω από μια αρχιτεκτονική πελάτη-διακομιστή (client - server architecture) (Vaswani, 2009).

Η MySQL ήταν ένα από τα πρώτα RDBMS ανοιχτού κώδικα που αναπτύχθηκε και κυκλοφόρησε ποτέ. Επί του παρόντος, υπάρχουν πολλές παραλλαγές της MySQL. Ωστόσο, η βασική σύνταξη όλων των παραλλαγών παραμένει η ίδια. Σχεδιασμένη και γραμμένη σε γλώσσες προγραμματισμού C και C++, η MySQL είναι συμβατή με όλα τα κύρια OS (Operating Systems (λειτουργικά συστήματα)). Είναι ένα βασικό στοιχείο ενός ευρέως γνωστού πακέτου λογισμικού εφαρμογών ιστού ανοιχτού κώδικα που ονομάζεται LAMP, το οποίο σημαίνει Linux, Apache, MySQL, PHP/Perl/Python (Ravikiran, 2021).

Η MySQL βασίζεται σε μια κλιμακωτή αρχιτεκτονική, που αποτελείται από δύο κύρια υποσυστήματα και υποστηρικτικά στοιχεία που αλληλεπιδρούν μεταξύ τους για την ανάγνωση, ανάλυση και εκτέλεση ερωτημάτων (queries), και για την προσωρινή αποθήκευση και επιστροφή των αποτελεσμάτων των ερωτημάτων (Vaswani, 2009).

Η MySQL παρέχει μια βιβλιοθήκη πελάτη που είναι γραμμένη στη γλώσσα προγραμματισμού C, καθώς και ένα σύνολο APIs (Application Programming Interfaces) που παρέχουν ένα κατανοητό σύνολο κανόνων, μέσω του οποίου οι γλώσσες κεντρικού υπολογιστή (host languages) μπορούν να συνδεθούν στη MySQL και να στείλουν εντολές. Χρησιμοποιώντας ένα API προστατεύονται τα προγράμματα των πελατών από τυχόν υποκείμενες αλλαγές στη MySQL που θα μπορούσαν να επηρεάσουν τη συνδεσιμότητα. Επί του παρόντος, η MySQL παρέχει hooks σε C, C++, Eiffel, Java, Perl, PHP, Python, Ruby, και Tcl και οι connectors είναι επίσης διαθέσιμοι για εφαρμογές JDBC, ODBC και .NET (Vaswani, 2009).

Η βασική διαφορά της MySQL με την SQL είναι ότι η MySQL, είναι ένα λογισμικό που όπως φαίνεται και από την ονομασία του, χρησιμοποιεί την SQL, η οποία είναι αντίστοιχα γλώσσα προγραμματισμού, ως εργαλείο της. Πιο συγκεκριμένα, η SQL συμβάλλει στην διαχείριση των δεδομένων που βρίσκονται στις σχεσιακές βάσεις δεδομένων, ενώ η MySQL, έχει ως ρόλο την διαχείριση των σχεσιακών βάσεων δεδομένων, κάτι το οποίο πετυχαίνει χρησιμοποιώντας την SQL. Επιπλέον, ως γλώσσα προγραμματισμού, η SQL, χαρακτηρίζεται από αρκετή ασφάλεια. Αντίθετα, η πρόσβαση και η διαχείριση δεδομένων στο λογισμικό MySQL, μπορεί να γίνει σχετικά εύκολα (developNET, 2023).

Για την αλληλεπίδραση με τη MySQL χρησιμοποιώντας SQL, μπορούμε να χρησιμοποιήσουμε τη γραμμή εντολών του πελάτη MySQL (command-line client) ή ένα εργαλείο graphical user interface (GUI) όπως το MySQL Workbench.

Στον παρακάτω πίνακα συνοψίζονται οι βασικές διαφορές της SQL και της MySQL:

Πίνακας 4. Βασικές διαφορές SQL και MySQL, Πηγή: (Ravikiran, 2021)

SQL	MySQL
Η SQL είναι μια γλώσσα προγραμματισμού που χρησιμοποιεί ερωτήματα (queries) και διαχειρίζεται ένα σχεσιακό σύστημα διαχείρισης βάσης δεδομένων (RDBMS).	Η MySQL είναι ένα σχεσιακό σύστημα διαχείρισης βάσης δεδομένων που χρησιμοποιεί την SQL.
Η SQL χρησιμοποιείται κυρίως για ερωτήματα και για να λειτουργεί συστήματα βάσεων δεδομένων	Η MySQL επιτρέπει στο χρήστη να διαχειρίζεται, να αποθηκεύει, να επεξεργάζεται και να διαγράφει δεδομένα με έναν οργανωμένο τρόπο
Η SQL δεν υποστηρίζει καμία σύνδεση.	Η MySQL συνοδεύεται από ένα ενσωματωμένο εργαλείο γνωστό ως MySQL Workbench που διευκολύνει τη δημιουργία, το σχεδιασμό και τη δημιουργία βάσεων δεδομένων.
Η SQL ακολουθεί ένα απλό τυπικό format χωρίς πολλές ή τακτικές ενημερώσεις.	Η MySQL έχει πολλές παραλλαγές και λαμβάνει συχνές ενημερώσεις.
Η SQL υποστηρίζει μόνο μία μηχανή αποθήκευσης.	Η MySQL προσφέρει υποστήριξη για πολλαπλές μηχανές αποθήκευσης μαζί με ενσωματωμένη αποθήκευση, καθιστώντας την πιο ευέλικτη.
Η SQL δεν επιτρέπει σε άλλους επεξεργαστές ή ακόμα και στα δικά της δυαδικά αρχεία να χειρίζονται δεδομένα κατά την εκτέλεση.	Η MySQL είναι λιγότερο ασφαλής από την SQL, καθώς επιτρέπει σε τρίτους επεξεργαστές να χειρίζονται αρχεία δεδομένων κατά την εκτέλεση.

3.3.3 MySQL Workbench

Το MySQL Workbench είναι το γραφικό περιβάλλον χρήστη της Oracle για την αναζήτηση και τη διαχείριση του MySQL Server. Το κύριο χαρακτηριστικό του MySQL Workbench είναι η λειτουργία ερωτήματος (query mode) όπου ο χρήστης μπορεί να θέσει τα απαραίτητα ερωτήματα. Ωστόσο, υπάρχουν επίσης πολλά άλλα χαρακτηριστικά, όπως οι αναφορές απόδοσης (performance reports), το Visual Explain, η δυνατότητα διαχείρισης της διαμόρφωσης και επιθεώρησης σχημάτων και πολλά άλλα. (Krogh, 2020).

Στη συγκεκριμένη διπλωματική, κατασκευάστηκε πρώτα σχηματικά διάγραμμα οντοτήτων - συσχετίσεων με τη χρήση του MySQL Workbench και στη συνέχεια αντλήθηκαν οι κατάλληλες

εντολές, όπως αναφέρθηκε, με τη βοήθεια λειτουργιών που προσφέρει το MySQL Workbench. Με αυτόν τον τρόπο κατασκευάστηκε η τοπική βάση δεδομένων με το σύνολο δεδομένων μας. Παρακάτω αναλύονται με περισσότερες λεπτομέρειες τα διαγράμματα οντοτήτων - συσχετίσεων.

3.3.4 Διάγραμμα οντοτήτων - συσχετίσεων (ERD)

Ένα διάγραμμα οντοτήτων - συσχετίσεων (ER) είναι ένας τύπος διαγράμματος ροής (flow chart) που απεικονίζει πώς ορισμένες «οντότητες» όπως άνθρωποι, αντικείμενα ή έννοιες σχετίζονται μεταξύ τους μέσα σε ένα σύστημα. Τα διαγράμματα ER χρησιμοποιούνται συχνότερα για το σχεδιασμό ή τον εντοπισμό σφαλμάτων σχεσιακών βάσεων δεδομένων στους τομείς του software engineering, των συστημάτων πληροφοριών επιχειρήσεων, της εκπαίδευσης και της έρευνας. Επίσης γνωστά ως ERDs ή ER Models, χρησιμοποιούν ένα καθορισμένο σύνολο συμβόλων όπως ορθογώνια, διαμάντια, οβάλ και γραμμές σύνδεσης για να απεικονίσουν τη διασύνδεση των οντοτήτων, των σχέσεων και των ιδιοτήτων τους. Αντικατοπτρίζουν τη γραμματική δομή, με οντότητες ως ουσιαστικά και σχέσεις ως ρήματα (Lucidchart, 2024).

Τα διαγράμματα ER αποτελούνται από οντότητες (entities), σχέσεις (relationships) και χαρακτηριστικά (attributes). Απεικονίζουν επίσης την πολλαπλότητα (cardinality), η οποία ορίζει τις σχέσεις με αριθμητικούς όρους. Πιο συγκεκριμένα:

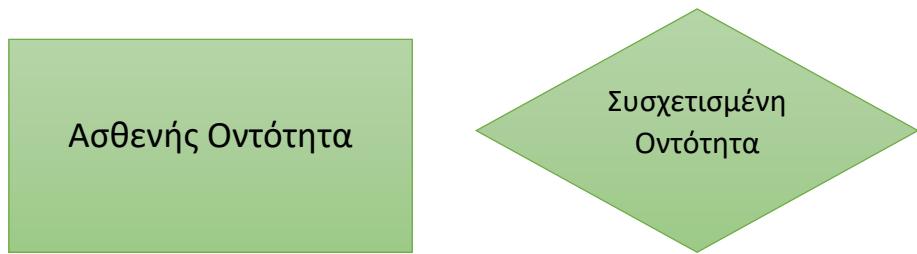
- **Οντότητα:** Θεωρείται μια προσδιορίσιμη και αυθύπαρκτη ομάδα - όπως ένα άτομο, ένα αντικείμενο, μια έννοια ή ένα γεγονός - που μπορεί να έχει αποθηκευμένα δεδομένα που σχετίζονται με αυτό. Μπορούμε να θεωρήσουμε τις οντότητες ως ουσιαστικά, για παράδειγμα: πελάτης, μαθητής, αυτοκίνητο ή προϊόν. Συνήθως εμφανίζεται ως ορθογώνιο.

Οντότητα

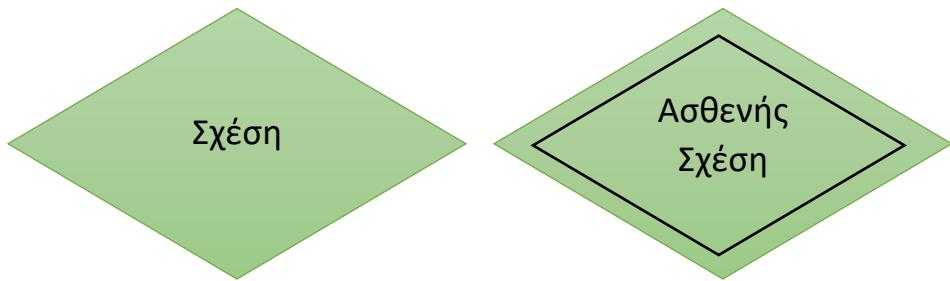
- Τύπος οντότητας: Πρόκειται για μια ομάδα προσδιορίσιμων πραγμάτων, όπως για παράδειγμα μαθητές ή αθλητές, ενώ η οντότητα θα ήταν ένας συγκεκριμένος μαθητής ή αθλητής.
- Σύνολο οντοτήτων: Είναι παρόμοιο με τον τύπο οντότητας, αλλά ορίζεται σε μια συγκεκριμένη χρονική στιγμή, όπως οι μαθητές που εγγράφηκαν σε μια τάξη την

πρώτη ημέρα ή για παράδειγμα πελάτες που αγόρασαν τον περασμένο μήνα αυτοκίνητα που είναι επί του παρόντος εγγεγραμμένα στην Ελλάδα. Ένας σχετικός όρος είναι το **στιγμιότυπο** (instance), στον οποίο το συγκεκριμένο άτομο ή αυτοκίνητο θα ήταν ένα στιγμιότυπο του συνόλου οντοτήτων.

- **Κατηγορίες οντοτήτων:** Οι οντότητες κατηγοριοποιούνται ως ισχυρές, αδύναμες ή συσχετισμένες. Μια ισχυρή οντότητα μπορεί να οριστεί αποκλειστικά από τα δικά της χαρακτηριστικά, ενώ μια αδύναμη οντότητα όχι. Μια συσχετισμένη οντότητα συσχετίζει οντότητες (ή στοιχεία) μέσα σε ένα σύνολο οντοτήτων.



- **Κλειδιά οντοτήτων:** Αναφέρεται σε ένα χαρακτηριστικό που ορίζει μοναδικά μια οντότητα σε ένα σύνολο οντοτήτων. Τα κλειδιά οντοτήτων μπορεί να είναι super key, candidate key ή primary key (πρωτεύον κλειδί). **Super key:** Ένα σύνολο χαρακτηριστικών (ένα ή περισσότερα) που μαζί ορίζουν μια οντότητα σε ένα σύνολο οντοτήτων. **Candidate key:** Ένα απλοποιημένο super key, που σημαίνει ότι έχει τον μικρότερο δυνατό αριθμό χαρακτηριστικών για να εξακολουθεί να είναι super key. Ένα σύνολο οντοτήτων μπορεί να έχει περισσότερα από ένα candidate keys. **Primary key (Πρωτεύον κλειδί):** Ένα candidate key που επιλέγεται από τον σχεδιαστή της βάσης δεδομένων για τον μοναδικό προσδιορισμό του συνόλου οντοτήτων. **Foreign key (Ξένο κλειδί):** Προσδιορίζει τη σχέση μεταξύ οντοτήτων.
- **Σχέση:** Περιγράφει το πώς οι οντότητες ενεργούν μεταξύ τους ή συσχετίζονται μεταξύ τους. Μπορούμε να θεωρήσουμε τις σχέσεις ως ρήματα. Για παράδειγμα, ο κατονομαζόμενος φοιτητής μπορεί να εγγραφεί σε ένα μάθημα. Οι δύο οντότητες θα είναι ο μαθητής και το μάθημα, και η σχέση που απεικονίζεται είναι η πράξη εγγραφής, συνδέοντας τις δύο οντότητες με αυτόν τον τρόπο. Οι σχέσεις εμφανίζονται συνήθως ως «διαμάντια» ή ετικέτες (labels) απευθείας στις γραμμές σύνδεσης.



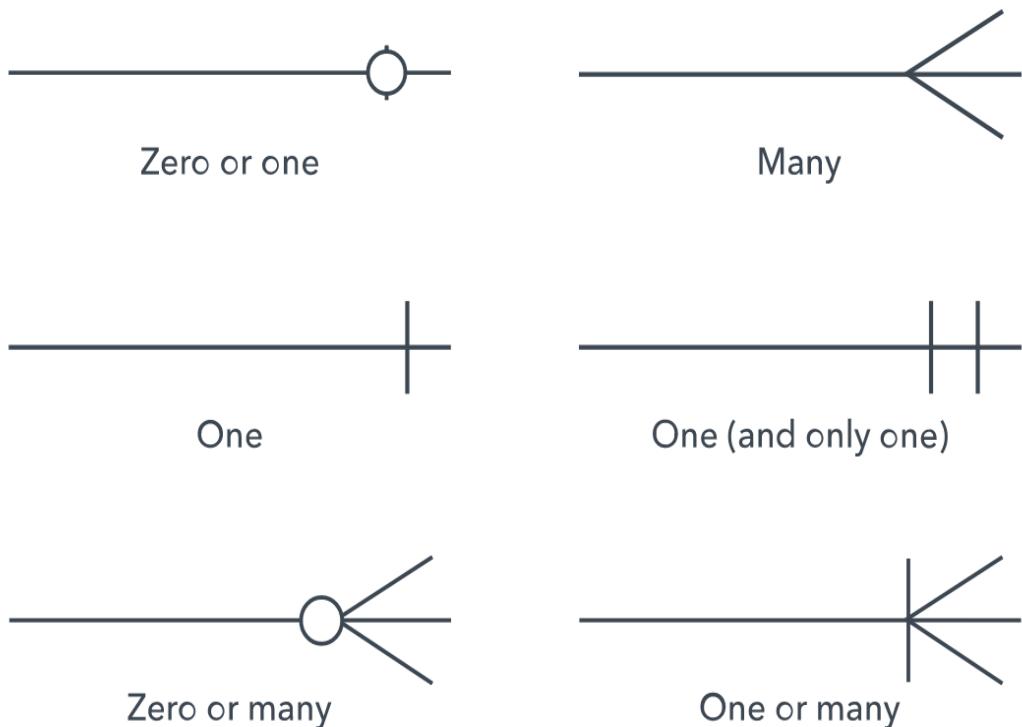
- Αμφίδρομη σχέση: Όταν η ίδια οντότητα συμμετέχει περισσότερες από μία φορές στη σχέση.
- **Χαρακτηριστικό:** Μια ιδιότητα ή χαρακτηριστικό μιας οντότητας. Συχνά εμφανίζεται ως οβάλ ή κύκλος.



- Περιγραφικό χαρακτηριστικό: Μια ιδιότητα ή χαρακτηριστικό μιας σχέσης (αντί για οντότητας)
- Κατηγορίες χαρακτηριστικών: Τα χαρακτηριστικά κατηγοριοποιούνται ως απλά, σύνθετα, παράγωγα, καθώς και ως μονής τιμής ή πολλαπλών τιμών. **Απλό:** Σημαίνει ότι η τιμή του χαρακτηριστικού είναι ατομική και δεν μπορεί να διαιρεθεί περαιτέρω, όπως για παράδειγμα ένας αριθμός τηλεφώνου. **Σύνθετο:** Τα δευτερεύοντα χαρακτηριστικά πηγάζουν από ένα χαρακτηριστικό. **Παράγωγο:** Το χαρακτηριστικό υπολογίζεται ή αλλιώς προέρχεται από ένα άλλο χαρακτηριστικό, όπως για παράδειγμα η ηλικία από μια ημερομηνία γέννησης.

Παράγωγο χαρακτηριστικό

- Πολλαπλών τιμών (multi-value): Υποδηλώνονται περισσότερες από μία τιμές χαρακτηριστικού, όπως πολλοί αριθμοί τηλεφώνου που αντιστοιχούν σε ένα άτομο.
- Μονής τιμής (single-value): Μόνο μία τιμή χαρακτηριστικού. Οι τύποι μπορούν να συνδυαστούν, όπως: απλά χαρακτηριστικά μιας τιμής ή σύνθετα χαρακτηριστικά πολλαπλών τιμών.
- **Πολλαπλότητα (cardinality):** Καθορίζει τα αριθμητικά χαρακτηριστικά της σχέσης μεταξύ δύο οντοτήτων ή συνόλων οντοτήτων. Οι τρεις κύριες βασικές σχέσεις πολλαπλότητας είναι one-to-one, one-to-many και many-to-many. Ένα παράδειγμα one-to-one θα ήταν ένας μαθητής που σχετίζεται με μία ταχυδρομική διεύθυνση. Ένα παράδειγμα one-to-many (ή many-to-oneα, ανάλογα με την κατεύθυνση της σχέσης): Ένας μαθητής εγγράφεται για πολλά μαθήματα, αλλά όλα αυτά τα μαθήματα περιλαμβάνουν μια γραμμή που επιστρέφει πίσω σε αυτόν τον έναν μαθητή. Παράδειγμα many-to-many: Οι φοιτητές ως ομάδα συνδέονται με πολλά μέλη ΔΕΠ και τα μέλη ΔΕΠ με τη σειρά τους συνδέονται με πολλούς φοιτητές. Παρακάτω απεικονίζονται οι πιθανοί τύποι γραμμών που δύνανται να συνδέουν τις οντότητες μεταξύ τους σε ένα διάγραμμα οντοτήτων-συσχετίσεων:



Εικόνα 25. Σχέσεις πολλαπλότητας σε ένα διάγραμμα οντοτήτων-συσχετίσεων, Πηγή: (Lucidchart, 2024)

- Όψεις πολλαπλότητας (cardinality views): Η πολλαπλότητα μπορεί να εμφανιστεί ως αμφίπλευρη ή στην ίδια πλευρά, ανάλογα με το που εμφανίζονται τα σύμβολα.
- Περιορισμοί πολλαπλότητας: Οι μέγιστοι και οι ελάχιστοι αριθμοί που ισχύουν για μια σχέση.

Με βάση λοιπόν τους κανόνες που διέπουν τη δημιουργία διαγραμμάτων οντοτήτων συσχετίσεων, κατασκευάσαμε αντίστοιχο κατάλληλο διάγραμμα ER με τη χρήση του MySQL Workbench που απεικονίζει τις σχέσεις των χαρακτηριστικών του συνόλου δεδομένων που χρησιμοποιήσαμε και μας βοήθησε στην πορεία στην κατασκευή του κατάλληλου αρχείου εισόδου που χρειάστηκε για την εκπαίδευση του μοντέλου κατηγοριοποίησης.

3.4 ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

3.4.1 Δημιουργία αρχείου input για την εκπαίδευση του μοντέλου

Για την αποτελεσματικότερη απόδοση της εκπαίδευσης των μοντέλων μηχανικής μάθησης, έπρεπε να κατασκευαστεί κατάλληλο αρχείο σε μορφή εύκολα επεξεργάσιμη και διαχειρίσιμη από έναν αλγόριθμο. Για αυτό το λόγο, κατασκευάστηκε εκ νέου συγκεντρωτικό αρχείο με τη χρήση της γλώσσας PHP, καθώς και της δυνατότητας της να επικοινωνεί άμεσα και αποτελεσματικά με τις βάσεις δεδομένων MySQL μέσω του αντικειμενοστρεφούς MySQLi extension. Σημειώνεται πως ένα σύνολο δεδομένων που μέλλεται να χρησιμοποιηθεί για την εκπαίδευση ενός αλγορίθμου μηχανικής μάθησης περιλαμβάνει τις εγγραφές (ή καταχωρήσεις) και τα χαρακτηριστικά (ή γνωρίσματα). Κάθε εγγραφή έχει το πολύ μια τιμή για κάθε διαφορετικό γνώρισμα.

Παρακάτω παρουσιάζονται επιγραμματικά τα βήματα ροής εργασίας του script που χρησιμοποιήθηκε για τη δημιουργία αυτού του αρχείου:

1. Δημιουργία σύνδεσης με τοπική ΒΔ με χρήση κατάλληλου servername, username, password, dbname
2. Ανάκτηση από τη ΒΔ όλων των μοναδικών IDs που να έχουν απαραίτητα αντίστοιχη εγγραφή στο πεδίο της κλάσης
3. Εισαγωγή IDs κατά πλάτος της πρώτης στήλης
4. Ανάκτηση από τη ΒΔ όλων των μοναδικών Drug, Indication, Reaction και εισαγωγή κατά μήκος της πρώτης γραμμής ως τίτλοι
5. Εισαγωγή Age, Sex, Country, Class κατά μήκος της πρώτης γραμμής ως τίτλοι
6. Δημιουργία όλων των πιθανών συνδέσεων κάθε ID με κάθε ξεχωριστό feature
7. Γέμισμα του πίνακα με βάση τα εξής κριτήρια:
 - Σε περίπτωση ύπαρξης σύνδεσης ενός ID με ένα χαρακτηριστικό από τα Drug, Indication, Reaction, τότε εισαγωγή του αριθμού 1, διαφορετικά εισαγωγή του αριθμού 0
 - Στη στήλη της ηλικίας, οι ηλικιακές κλάσεις μετατράπηκαν στο μέσο όρο της κλάσης, τύπου float

Έτσι λοιπόν, μια τυπική αναπαράσταση της μορφής του συνόλου δεδομένων που αποτέλεσε τελικά το αρχείο εισόδου για την εκπαίδευση του μοντέλου ήταν η εξής:

Πίνακας 5. Μορφή συνόλου δεδομένων που χρησιμοποιήθηκε ως είσοδος στο μοντέλο ταξινόμησης

EU_Local_Numb er	TOZINAMERAN	MODERNA	...	Dumping syndrome	COVID-19 immunizati on	Algesic therapy	...	Headache	1p36 deletion syndrome	...	Age	Sex	Country	Class
EU-EC-10007291936	1	0	...	1	1	0	...	0	0	...	41.000	Female	EEA	2
EU-EC-10007296035	0	1	...	0	1	0	...	0	1	...	41.000	Male	EEA	2
EU-EC-10007307981	0	0	...	0	0	1	...	1	0	...	75.000	Male	NEEA	0
EU-EC-10007314324	1	0	...	0	1	0	...	1	0	...	1.083	Female	EEA	4

3.4.2 Γλώσσα Python

Η Python είναι γλώσσα προγραμματισμού που χρησιμοποιεί interpreter, υψηλού επιπέδου με δυναμική σημασιολογία και διαθέτει μια πολύ αποτελεσματική προσέγγιση στον αντικειμενοστρεφή προγραμματισμό. Οι ενσωματωμένες δομές δεδομένων υψηλού επιπέδου, σε συνδυασμό με τη δυναμική πληκτρολόγηση και τη δυναμική σύνδεση, την καθιστούν εξαιρετικά εύχρηστη για γρήγορη ανάπτυξη εφαρμογών. Η απλή και εύκολη στην εκμάθηση σύνταξη της Python δίνει έμφαση στην αναγνωσιμότητα, και ως εκ τούτου μειώνει το κόστος συντήρησης του προγράμματος. Η Python υποστηρίζει λειτουργικές μονάδες και πακέτα, τα οποία ενθαρρύνουν την παραμετροποίηση των προγραμμάτων και την επαναχρησιμοποίηση του κώδικα. Ο Python interpreter και η εκτεταμένη βασική της βιβλιοθήκη είναι διαθέσιμα σε source ή δυαδική μορφή χωρίς χρέωση για όλες τις μεγάλες πλατφόρμες (Van Rossum & Drake, 2003).

Η Python έχει εδραιώσει τη θέση της ως η πλέον προτιμώμενη γλώσσα για τη μηχανική μάθηση και το πακέτο Scikit-learn είναι που παίζει καθοριστικό ρόλο σε αυτό. Πιο συγκεκριμένα, η «διαισθητική» σύνταξη της Python την καθιστά προσβάσιμη τόσο σε αρχάριους όσο και σε έμπειρους προγραμματιστές. Η βιβλιοθήκη Scikit-learn τηρεί αυτήν την απλότητα, επιτρέποντας στο χρήστη να εφαρμόσει τη μηχανική μάθηση χωρίς απαραίτητα τη χρήση πολύπλοκου κώδικα (Karl, 2023). Η Scikit-learn παρέχει εργαλεία ταξινόμησης, παλινδρόμησης, ομαδοποίησης, μείωσης διαστάσεων και πολλά άλλα. Ο συνεπής σχεδιασμός API και η ολοκληρωμένη τεκμηρίωση απλοποιούν την ανάπτυξη μοντέλων ML (CFI team, 2024; Karl, 2023). Παράλληλα, η ενεργή κοινότητα προγραμματιστών της Python συμβάλλει σημαντικά στην επιτυχία της στο ML. Το Scikit-learn επωφελείται από αυτό το συνεργατικό περιβάλλον, διασφαλίζοντας τακτικές ενημερώσεις, διορθώσεις σφαλμάτων και βελτιώσεις. Τα community forums και τα σεμινάρια διευκολύνουν την ανταλλαγή γνώσεων και την επίλυση προβλημάτων. Συνοπτικά, η αναγνωσιμότητα της Python, οι δυνατότητες του Scikit-learn, η υποστήριξη της κοινότητας και η

συμβατότητα την καθιστούν την ιδανική γλώσσα για δοκιμασίες μηχανικής μάθησης (CFI team, 2024; Karl, 2023).

Σε αυτή τη μελέτη, τα μοντέλα μηχανικής μάθησης αναπτύχθηκαν σε γλώσσα Python έκδοσης 3.11, αξιοποιώντας τις εκτεταμένες βιβλιοθήκες και τα frameworks που παρέχει για την προ-επεξεργασία δεδομένων, την εκπαίδευση μοντέλων και την αξιολόγηση τους.

3.4.3 Data pre-processing (προ-επεξεργασία δεδομένων)

3.4.3.1 Αφαίρεση δεδομένων

Κατά την διαδικασία προ-επεξεργασίας των δεδομένων, αφαιρέσαμε δεδομένα που δε θα έπαιζαν κάποιο ουσιαστικό ρόλο στην εκπαίδευση του μοντέλου. Αυτά ήταν τα εξής:

- Το γνώρισμα «EU_Local_Number» , δηλαδή ο αριθμός ταυτοποίησης κάθε εγγραφής, είναι διαφορετικό για κάθε ασθενή και έχει να κάνει αποκλειστικά με την καταγραφή τους στο σύστημα, επομένως δεν κρίθηκε απαραίτητο για τη μελέτη.
- Μετά τη δημιουργία του input αρχείου μέσω των κατάλληλων queries, όπως εξηγήθηκε παραπάνω, διαπιστώθηκε πως πολλά γνωρίσματα είχαν αποκλειστικά την τιμή 0 καθ' όλο το πλάτος της στήλης, κάτι το οποίο προέκυψε λόγω της επιλογής αποκλειστικά των εγγραφών που είχαν αντίστοιχη τιμή στο πεδίο της κλάσης, κάτι το οποίο θα αναλυθεί περεταίρω στην ενότητα ‘Αποτελέσματα’. Αυτά τα γνωρίσματα αφαιρέθηκαν καθώς δε θα συνέβαλαν με κάποιο τρόπο στην ισχύ του μοντέλου.

3.4.3.2 Imputation

Όταν καλούμαστε να διαχειριστούμε πραγματικά δεδομένα, όπως αυτά που αντλήθηκαν από τη βάση EudraVigilance, είναι αρκετά σύνηθες να συναντάμε «κενές» τιμές που λείπουν για ποικίλους λόγους, όπως σφάλματα εισαγωγής δεδομένων ή ελλιπείς εγγραφές. Γι' αυτό το λόγο κατά τη διαδικασία της προ-επεξεργασίας του συνόλου δεδομένων στη μηχανική μάθηση χρησιμοποιούμε τους λεγόμενους imputers, οι οποίοι διαδραματίζουν κρίσιμο ρόλο στο χειρισμό των «κενών» τιμών σε σύνολα δεδομένων προγνωστικών μοντέλων. Οι imputers μας βοηθούν να καλύψουμε αυτά τα κενά, διασφαλίζοντας ότι τα μοντέλα μας μπορούν να μάθουν αποτελεσματικά.

Υπάρχουν διαφορετικοί imputers, όπως IterativeImputer, KNNImputer, TimeSeriesImputer και CustomImputer που ο καθένας μπορεί να είναι καταλληλότερος για διαφορετικό τύπο συνόλου δεδομένων. Στην παρούσα διπλωματική εργασία χρησιμοποιήσαμε τον SimpleImputer που

παρέχεται από τη βιβλιοθήκη Scikit-learn της Python, όπως αναφέρθηκε νωρίτερα (GeeksforGeeks, 2019). Συγκεκριμένα, ο imputer αυτός χρησιμοποιήθηκε με τη μέθοδο ‘mean’, δηλαδή αντικαθιστώντας τις τιμές που λείπουν με τη μέση τιμή της εν λόγω στήλης. Σημειώνεται πως η μοναδική στήλη που περιείχε κενές τιμές ήταν η ηλικία (‘Age’) που περιείχε δεκαδικούς αριθμούς, επομένως κρίναμε ιδανικότερη τη χρήση του συγκεκριμένου imputer.

3.4.3.3 *Encoding* (κωδικοποίηση)

Στη σφαίρα της μηχανικής μάθησης, οι encoders (κωδικοποιητές) διαδραματίζουν κεντρικό ρόλο στη μετατροπή των δεδομένων εισόδου σε καταλληλότερες αναπαραστάσεις για την ανάλυση και τη μοντελοποίηση. Αυτοί οι μετασχηματισμοί απλοποιούν τη διαδικασία εκμάθησης για τους αλγόριθμους μηχανικής μάθησης και βελτιώνουν την απόδοσή τους, καθώς μετατρέπουν τα δεδομένα σε μια «συμπιεσμένη» μορφή που όμως διατηρεί τα ουσιώδη χαρακτηριστικά που χρειάζονται για την ανάλυση. Οι encoders είναι ιδιαίτερα χρήσιμοι όταν δουλεύουμε με αλγόριθμους που απαιτούν αριθμητική εισαγωγή, καθώς τα περισσότερα μοντέλα μηχανικής εκμάθησης λειτουργούν αποκλειστικά σε αριθμητικά δεδομένα (Schmitz, 2023).

Υπάρχουν διάφοροι τύποι encoders όπως οι label encoder, one-hot encoder, target encoder, binary encoder, count encoder, hashing encoder. Καθώς τα δεδομένα μας, όπως αναφέρθηκε, αποτελούνταν κυρίως από δυαδικές τιμές 0 και 1, η κωδικοποίηση χρειάστηκε μόνο σε λίγες στήλες, συγκεκριμένα στις στήλες ‘Country’ και ‘Sex’ που είχαν κατηγορικές τιμές. Στην παρούσα διπλωματική χρησιμοποιήθηκε ο Label Encoder. Ο Label Encoder είναι μια θεμελιώδης τεχνική που χρησιμοποιείται για τη μετατροπή κατηγορικών μεταβλητών σε αριθμητική μορφή. Ο πρωταρχικός σκοπός του είναι να αντιστοιχίσει διακριτές ετικέτες (όπως ονόματα κλάσεων ή κατηγορίες) σε μοναδικές ακέραιες τιμές, για παράδειγμα για ένα χαρακτηριστικό ‘Άψος’ με πιθανές τιμές ‘Κοντός’, ‘Μεσαίος’, ‘Ψηλός’, ο Label Encoder μπορεί να δώσει τις τιμές ‘0’, ‘1’ και ‘2’ αντίστοιχα (GeeksforGeeks, 2018).

3.4.3.4 *Feature selection*

Η επιλογή χαρακτηριστικών, το λεγόμενο feature selection, αποτελεί κρίσιμο βήμα στην ανάπτυξη ενός μοντέλου. Ο πρωταρχικός του στόχος είναι να επιλέξει ένα υποσύνολο των μεταβλητών εισόδου που είναι περισσότερο σχετικές από άλλες για την πρόβλεψη της μεταβλητής στόχου. Με αυτόν τον τρόπο, επιτυγχάνονται τα εξής οφέλη:

- Μειωμένο υπολογιστικό κόστος: Λιγότερες δυνατότητες σημαίνουν ταχύτερη εκπαίδευση και εξαγωγή συμπερασμάτων.

- Βελτιωμένη απόδοση μοντέλου: Η κατάργηση άσχετων ή περιττών χαρακτηριστικών μπορεί να βελτιώσει την ακρίβεια της πρόβλεψης.

Υπάρχουν τρεις βασικές κλάσεις- αλγορίθμων feature selection:

1. Μέθοδοι φίλτρου (Filter Methods): Αυτές οι μέθοδοι χρησιμοποιούν στατιστικά μέτρα για να βαθμολογήσουν τη συσχέτιση ή την εξάρτηση μεταξύ των μεταβλητών εισόδου. Τα πιο κοινά στατιστικά μέτρα περιλαμβάνουν τα:
 - Information Gain
 - Chi-Square Test
 - Fisher's Score
2. Μέθοδοι Wrapper: Αυτές οι μέθοδοι αξιολογούν υποσύνολα χαρακτηριστικών εκπαιδεύοντας και δοκιμάζοντας το μοντέλο επαναληπτικά. Κάποια παραδείγματα περιλαμβάνουν:
 - Forward Selection: Προσθέτει χαρακτηριστικά ένα προς ένα
 - Backward Elimination: Αφαιρεί επαναληπτικά χαρακτηριστικά.
 - Bi-directional Elimination: Συνδυάζει τους δύο παραπάνω τρόπους.
 - Recursive Elimination: Αφαιρεί αναδρομικά χαρακτηριστικά που βασίζονται στην απόδοση του μοντέλου.
3. Ενσωματωμένες μέθοδοι (Embedded Methods): Αυτές οι μέθοδοι ενσωματώνουν την επιλογή χαρακτηριστικών ως μέρος του ίδιου του αλγορίθμου εκμάθησης. Παραδείγματα:
 - LASSO (Least Absolute Shrinkage and Selection Operator)
 - Random Forest Importance

Το πακέτο Skicit-learn της Python που αναφέρθηκε νωρίτερα προσφέρει τη δυνατότητα εφαρμογής αλγορίθμων feature selection όπως οι SelectKBest, SelectPercentile, RFE (Recursive Feature Elimination) και SelectFromModel.

Στην παρούσα διπλωματική επιλέχθηκε ο αλγόριθμος SelectPercentile. Ο αλγόριθμος αυτός ανήκει στην πρώτη κλάση που αναφέρθηκε και ουσιαστικά ταξινομεί τα χαρακτηριστικά ανάλογα με την στατιστική τους σημαντικότητα και διατηρεί το κορυφαίο ποσοστό. Πιο αναλυτικά, χρησιμοποιεί τη μετρική F-value από την ANOVA για δοκιμές παλινδρόμησης, ενώ το τεστ χ-τετράγωνο για δοκιμές κατηγοριοποίησης (Brownlee, 2016).

3.4.3.5 Class Balancing

Η εξισορρόπηση κλάσεων , το λεγόμενο Class Balancing, αναφέρεται στην αντιμετώπιση του πιθανού προβλήματος της μη ισορροπημένης κατανομής της κλάσεις μέσα σε ένα σύνολο δεδομένων. Όταν μια κατηγορία υπερτερεί σημαντικά των άλλων, μπορεί να οδηγήσει σε μεροληπτική απόδοση του μοντέλου. Οι πιο συνήθεις περιπτώσεις περιλαμβάνουν τον εντοπισμό απάτης (fraud detection), τον εντοπισμό ανωμαλιών (anomaly detection) και την αναγνώριση προσώπου (facial recognition). Κάποιες τεχνικές για την εξισορρόπηση των δεδομένων είναι οι εξής:

- Υπερδειγματοληψία (oversampling):
 - SMOTE (Synthetic Minority Oversampling Technique): Δημιουργεί τεχνητές περιπτώσεις που αποτελούν μειοψηφία παρεμβάλλοντας μεταξύ των ήδη υπαρχόντων δειγμάτων μειοψηφίας. Στοχεύει στην εξισορρόπηση της κατανομής της κλάσης αυξάνοντας με τυχαίο τρόπο τον αριθμό των παρατηρήσεων που αποτελούν τη μειοψηφία.
 - Random Oversampling: Αναπαράγει τις παρατηρήσεις που ανήκουν στην κλάση μειοψηφίας με τυχαίο τρόπο για να επιτευχθεί ισορροπία.
- Υποδειγματοληψία (undersampling):
 - Αλγόριθμος NearMiss: Καταργεί τυχαία τις παρατηρήσεις της κλάσης πλειοψηφίας για να αυξήσει τον διαχωρισμό μεταξύ των κλάσεων. Αποτρέπει την απώλεια πληροφορίας χρησιμοποιώντας μεθόδους near-neighbor.
- Υβριδικές μέθοδοι: συνδυάζονται τεχνικές υπερδειγματοληψίας και υποδειγματοληψίας για καλύτερα αποτελέσματα.

Για την παρούσα διπλωματική επιλέχθηκε ο balancer TomekLinks της βιβλιοθήκης Scikit-learn της Python που αποτελεί τεχνική υποδειγματοληψίας. Αυτή η τεχνική θεωρήσαμε ότι είναι η πιο κατάλληλη καθώς το σύνολο δεδομένων ήταν ήδη πολύ μεγάλο και μπορούσε να υποστηρίξει την αφαίρεση κάποιων εγγραφών, και προτιμήθηκε σε σχέση με την κατασκευή τεχνητών εγγραφών που ενδέχεται να οδηγούσε σε υπερπροσαρμογή. Ο αλγόριθμος αυτός συγκεκριμένα στοχεύει στην αφαίρεση παρατηρήσεων που είναι κοντά στο ‘όριο απόφασης’. Αρχικά ταυτοποιεί ένα ζεύγος παρατηρήσεων που είναι ‘κοντινοί γείτονες’. Αν το ζεύγος αυτό ανήκει σε διαφορετικές κλάσεις και η απόσταση μεταξύ τους είναι μικρή, τότε η παρατήρηση με την κλάση πλειοψηφίας αφαιρείται. Η διαδικασία αυτή επαναλαμβάνεται αυξάνοντας το διαχωρισμό των κλάσεων μεταξύ τους (EliteDataScience, 2022).

3.4.4 Διαχωρισμός συνόλου δεδομένων

Στην προηγούμενη ενότητα εξηγήσαμε την έννοια του διαχωρισμού του συνόλου δεδομένων σε σύνολο εκπαίδευσης και σύνολο δοκιμής. Στη μηχανική εκμάθηση, τα πιο συνήθη ποσοστά για το διαχωρισμό του συνόλου δεδομένων σε σύνολα εκπαίδευσης και δοκιμής είναι συνήθως περίπου 70% έως 80% για δεδομένα εκπαίδευσης και 20% έως 30% για δεδομένα δοκιμής. Αυτή η διαίρεση επιτυγχάνει μια ισορροπία μεταξύ της παροχής επαρκών δεδομένων από τα οποία μπορεί να μάθει το μοντέλο, διασφαλίζοντας παράλληλα επαρκή ποσότητα νεών δεδομένων για αξιολόγηση. Ωστόσο, τα ακριβή ποσοστά μπορεί να διαφέρουν ανάλογα με παράγοντες όπως το μέγεθος του συνόλου δεδομένων, η πολυπλοκότητα του προβλήματος και οι ειδικές απαιτήσεις της εργασίας (Hastie, Tibshirani, & Friedman, 2009a).

Στην παρούσα διπλωματική το 80% του συνόλου δεδομένων αποτέλεσε το σύνολο εκπαίδευσης και το υπόλοιπο 20% αποτέλεσε το σύνολο δοκιμής.

3.4.5 Αλγόριθμοι κατηγοριοποίησης

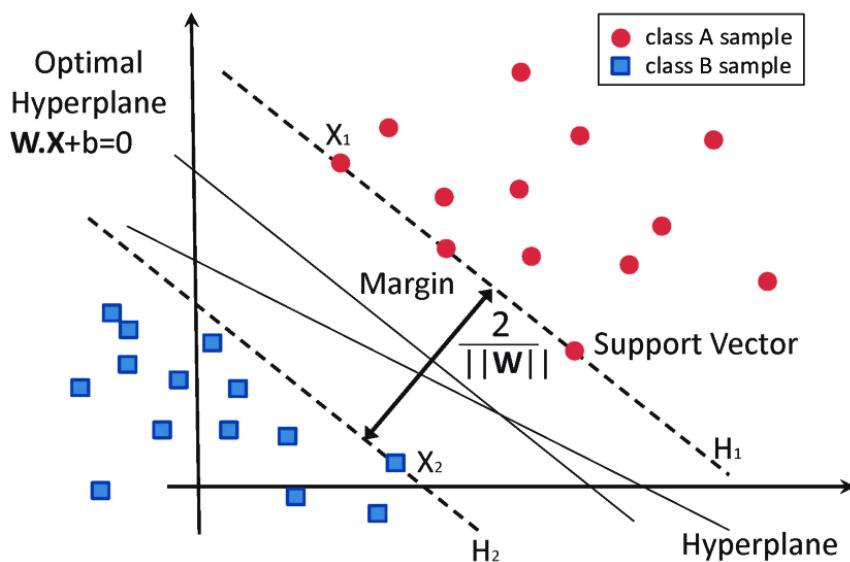
3.4.5.1 *Support Vector Machine (SVM) and Support Vector Classification (SVC)* – Μηχανές Διανυσμάτων Υποστήριξης και Ταξινόμηση Διανυσμάτων Υποστήριξης

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machine (SVM)) είναι ένας τύπος γραμμικών ταξινομητών (linear classifiers) που βασίζονται στην αρχή μεγιστοποίησης του περιθωρίου (margin maximization principle). Χρησιμοποιούνται για επιβλεπόμενη μάθηση και μπορούν να εφαρμοστούν τόσο σε προβλήματα ταξινόμησης όσο και σε προβλήματα παλινδρόμησης (Adankon & Cheriet, 2009; Shmilovici, 2005). Οι Μηχανές Διανυσμάτων Υποστήριξης εκτελούν ελαχιστοποίηση δομικού κινδύνου (structural risk minimization), η οποία βελτιώνει την πολυπλοκότητα του ταξινομητή με στόχο την επίτευξη της βέλτιστης απόδοσης για γενίκευση.

Οι Μηχανές Διανυσμάτων Υποστήριξης επιτυγχάνουν το σκοπό της ταξινόμησης κατασκευάζοντας, σε ένα χώρο ανώτερης διάστασης, το υπερεπίπεδο που διαχωρίζει με το βέλτιστο τρόπο τα δεδομένα σε δύο κατηγορίες. Ο αλγόριθμος SVM επιχειρεί να βρει το μέγιστο περιθώριο μεταξύ των δύο κατηγοριών των δεδομένων και στη συνέχεια προσδιορίζει το υπερεπίπεδο που βρίσκεται στο μέσο του μέγιστου περιθωρίου. Με αυτόν τον τρόπο, τα σημεία που είναι πλησιέστερα στο όριο απόφασης (decision boundary) βρίσκονται στην ίδια απόσταση από το βέλτιστο υπερεπίπεδο.

Το πρόβλημα βελτιστοποίησης που λύνουν οι Μηχανές Διανυσμάτων Υποστήριξης είναι να βρουν τις παραμέτρους w και b μιας συνάρτησης γραμμικής απόφασης (linear decision function) $f(x) = w^*x + b$ που ορίζουν το βέλτιστο υπερεπίπεδο. Τα σημεία πλησίον του ορίου απόφασης ορίζουν ουσιαστικά το περιθώριο. Θεωρώντας δύο σημεία x_1, x_2 στις απέναντι πλευρές του περιθωρίου με $f(x_1) = 1$ και $f(x_2) = -1$, το περιθώριο ισούται με $[f(x_1) - f(x_2)] / ||w|| = 2 / ||w||$. Έτσι, η μεγιστοποίηση του περιθωρίου ισοδυναμεί με την ελαχιστοποίηση $||w||^2 / 2$ (Adankon & Cheriet, 2009).

Παρακάτω απεικονίζεται η διαδικασία που περιγράφηκε για την ταξινόμηση δεδομένων σε δύο κλάσεις A και B:



Εικόνα 26. Support Vector Machines, Πηγή: <https://www.researchgate.net>

Σύμφωνα με τους Qi et al. (Qi et al., 2005), η Ταξινόμηση Διανυσμάτων Υποστήριξης (Support Vector Classification) (SVC) είναι μια επέκταση των Μηχανών Διανυσμάτων Υποστήριξης (SVM) που μπορεί να χρησιμοποιηθεί για προβλήματα ταξινόμησης πολλών κλάσεων. Στην περίπτωση της ταξινόμησης πολλαπλών κλάσεων, τα SVM μπορούν να χρησιμοποιηθούν για την κατασκευή πολλαπλών δυαδικών ταξινομητών, καθένας από τους οποίους διακρίνει μεταξύ δύο κατηγοριών δεδομένων. Ωστόσο, αυτή η προσέγγιση μπορεί να απαιτεί μεγάλη υπολογιστική ισχύ και ενδεχομένως να μην είναι πρακτική για μεγάλα σύνολα δεδομένων. Το SVC, από την άλλη πλευρά, έχει σχεδιαστεί ειδικά για προβλήματα ταξινόμησης πολλαπλών κλάσεων και μπορεί να χρησιμοποιηθεί για την ταξινόμηση δεδομένων σε περισσότερες από δύο κατηγορίες. Συγκεκριμένα, το SVC χρησιμοποιεί μια προσέγγιση one-vs-one, όπου κατασκευάζει έναν δυαδικό ταξινομητή για κάθε ζεύγος κλάσεων στο σύνολο δεδομένων. Κατά την ταξινόμηση νέων

δεδομένων, το SVC εφαρμόζει κάθε δυαδικό ταξινομητή στα δεδομένα και επιλέγει την κλάση που προβλέπεται πιο συχνά από τους δυαδικούς ταξινομητές (Baeldung, 2023; Bogawar & Bhoyar, 2018; Qi et al., 2005).

3.4.5.2 Αλγόριθμος *K-Nearest Neighbors* (*k*-πλησιέστεροι γείτονες)

Ο αλγόριθμος *k*-πλησιέστερων γειτόνων (*k*-nearest neighbors (*k*-NN)) είναι ένας απλός αλλά ισχυρός μη-παραμετρικός ταξινομητής που είναι ανθεκτικός σε δεδομένα με «θόρυβο» (noisy data) και εύκολος στην εφαρμογή. Ο αλγόριθμος λειτουργεί προσδιορίζοντας τους *k*-πλησιέστερους γείτονες σε ένα ερώτημα (query) και χρησιμοποιώντας αυτούς τους γείτονες για να καθορίσει την κλάση αυτού του ερωτήματος. Η μετρική της απόστασης που χρησιμοποιείται για τον προσδιορισμό των πλησιέστερων γειτόνων μπορεί να ποικίλει ανάλογα με τον τύπο δεδομένων και το πρόβλημα που καλούμαστε να αντιμετωπίσουμε. Ο αλγόριθμος μπορεί να φανεί ιδιαίτερα χρήσιμος όταν δεν υπάρχουν εμπόδια ανεπαρκούς απόδοσης στο χρόνο, δεδομένης της υπολογιστικής ισχύος που είναι διαθέσιμη. Συνοπτικά, ο αλγόριθμος *k*-NN περιγράφεται ως εξής (Cunningham & Delany, 2021):

- Υπολογισμός της απόστασης μεταξύ του παραδείγματος-ερωτήματος και όλων των παραδειγμάτων στο σύνολο δεδομένων.
- Ταξινόμηση των αποστάσεων σε αύξουσα σειρά και επιλογή των *k*-πλησιέστερων γειτόνων.
- Ανάθεση του παραδείγματος-ερωτήματος στην κατάλληλη κλάση με βάση την κλάση πλειοψηφίας των *k*-πλησιέστερων γειτόνων.

Αποτελέσματα πειραμάτων δείχνουν ότι η απόδοση του αλγορίθμου KNN εξαρτάται σημαντικά από την απόσταση που χρησιμοποιείται και τα αποτελέσματα δείχνουν μεγάλες διαφορές μεταξύ της απόδοσης με τη χρήση διαφορετικών αποστάσεων (Abu Alfeilat et al., 2019).

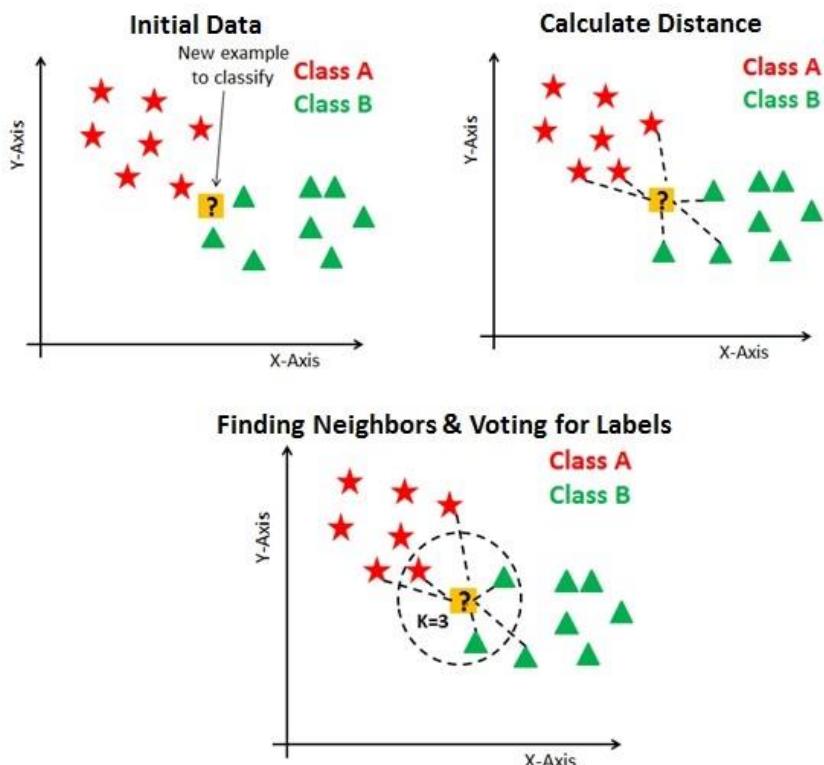
Υπάρχουν πολλά μέτρα απόστασης διαθέσιμα για χρήση στον ταξινομητή *K*-πλησιέστερων γειτόνων (KNN), ο οποίος θεωρείται ένας από τους απλούστερους και πιο συνηθισμένους ταξινομητές (Agrawal, 2018).

Μερικά από τα πιο δημοφιλή μέτρα απόστασης περιλαμβάνουν:

- Ευκλείδεια απόσταση (Euclidean distance): Αυτό είναι το πιο ευρέως χρησιμοποιούμενο μέτρο απόστασης και είναι η προεπιλεγμένη μέτρηση που χρησιμοποιείται από τη βιβλιοθήκη Scikit-learn της Python για τον αλγόριθμο KNN. Είναι το μέτρο της πραγματικής ευθείας απόστασης μεταξύ δύο σημείων στον Ευκλείδειο χώρο (Hu et al., 2016).

- Απόσταση Μανχάταν (Manhattan distance): Γνωστή και ως απόσταση ταξί (taxicab distance), είναι το άθροισμα των απόλυτων διαφορών των συντεταγμένων δύο σημείων στο χώρο (Agrawal, 2018).
- Απόσταση Μινκόφσκι (Minkowski distance): Αυτή είναι μια γενίκευση τόσο της Ευκλείδιας απόστασης όσο και της απόστασης Μανχάταν και ορίζεται ως η p -οστή ρίζα του αθροίσματος των απόλυτων διαφορών των συντεταγμένων δύο σημείων στο χώρο που υψώνονται στην δύναμη του p (Agrawal, 2018).

Παρακάτω αναπαρίσταται η λειτουργία του αλγορίθμου KNN απλοποιημένη:



Εικόνα 27. K-nearest neighbors algorithm

3.4.5.3 Αλγόριθμος Decision trees (Δέντρα Απόφασης)

Ένα Δέντρο Απόφασης είναι μια ιεραρχική δομή που αντιπροσωπεύει αποφάσεις ή ταξινομήσεις με βάση χαρακτηριστικά των δεδομένων υπό ανάλυση. Αποτελείται από κόμβους (που αντιπροσωπεύουν τα χαρακτηριστικά), ακμές (που συνδέουν τους κόμβους) και φύλλα (που αντιπροσωπεύουν τις ετικέτες (labels) των κλάσεων). Ο τρόπος που λειτουργούν είναι ο εξής:

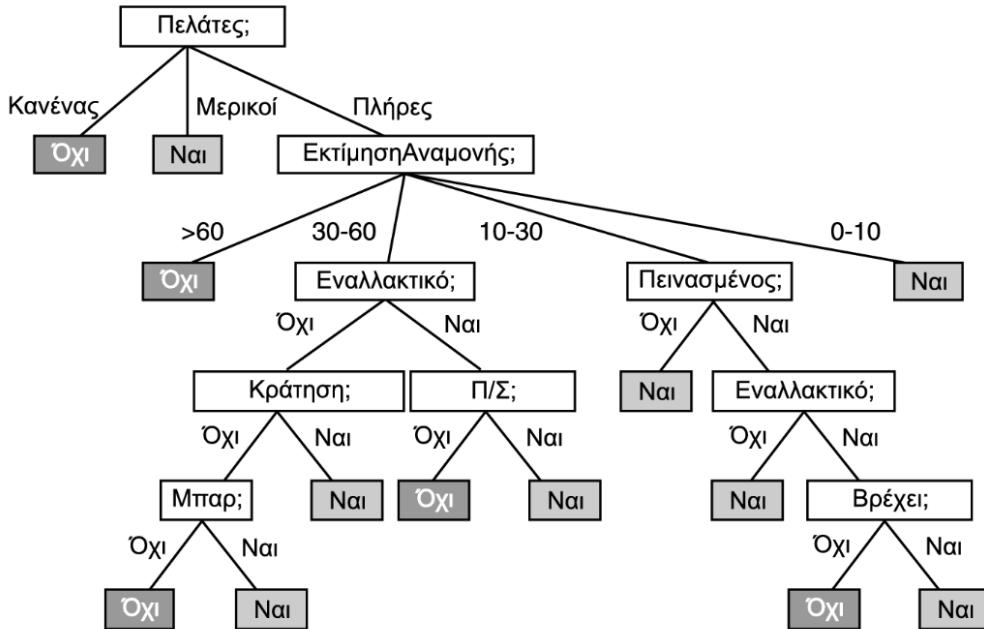
- Κριτήρια διαχωρισμού: Τα δέντρα αποφάσεων διαχωρίζουν αναδρομικά τα δεδομένα με βάση τις τιμές που παίρνουν τα χαρακτηριστικά. Ο στόχος είναι να δημιουργηθούν ομοιογενή υποσύνολα (καθαροί κόμβοι) σε σχέση με τη μεταβλητή-στόχο.

- Εντροπία και κέρδος πληροφορίας (information gain): Τα κοινά κριτήρια διαχωρισμού περιλαμβάνουν την εντροπία και το κέρδος πληροφορίας. Η εντροπία μετρά το πόσο «καθαρός» είναι ένας κόμβος του δέντρου, ενώ το κέρδος πληροφορίας ποσοτικοποιεί τη μείωση της αβεβαιότητας μετά από μια διάσπαση.
- Αναδρομικός διαχωρισμός (recursive splitting): Το δέντρο αναπτύσσεται επιλέγοντας αναδρομικά το εκάστοτε καλύτερο χαρακτηριστικό για το διαχωρισμό των δεδομένων. Η διαδικασία συνεχίζεται μέχρι να ικανοποιηθεί ένα κριτήριο διακοπής (π.χ. μέγιστο βάθος, ελάχιστα δείγματα ανά φύλλο).
- «Κλάδεμα» (pruning): Για να αποφευχθεί η υπερ-προσαρμογή των δεδομένων (overfitting), τεχνικές «κλαδέματος» αφαιρούν περιττά κλαδιά από το δέντρο.
- Πρόβλεψη: Για να ταξινομηθεί ένα νέο στιγμιότυπο (instant), το δέντρο διασχίζεται από τη ρίζα σε ένα φύλλο, ακολουθώντας τους κανόνες απόφασης σε κάθε κόμβο.

Αξίζει να αναφερθούν κάποιοι τύπου Δέντρων Απόφασης:

- Μέθοδοι συνόλου (ensemble methods): Οι ερευνητές έχουν εξερευνήσει μεθόδους συνόλου όπως το bagging και τα Random Forests που συνδυάζουν πολλαπλά δέντρα απόφασης για τη βελτίωση της ακρίβειας (Breiman, 1996, 2001).
- Δεδομένα υψηλής διάστασης (high-dimensional data): Έχουν διερευνηθεί τεχνικές για το χειρισμό των δεδομένων υψηλής διάστασης καθώς και των κατανεμημένων δεδομένων (distributed data) (Bar-Or et al., 2005; Caragea et al., 2004).
- «Ασαφή» Δέντρα Απόφασης (Fuzzy Decision Trees): Τα «Ασαφή» Δέντρα Απόφασης βασίζονται σε δείκτες Gini) μπορούν να προσφέρουν μεγαλύτερη ισχύ και ακρίβεια (Chandra & Varghese, 2009).
- Τακτικά Δεδομένα (Ordinal Data): έχουν μελετηθεί επίσης Δέντρα Απόφασης για την εκμάθηση τακτικών δεδομένων που εκφράζονται ως συγκρίσεις ανά ζεύγη (Qomariyah et al., 2020).

Μια τυπική εικόνα ενός Δέντρου Απόφασης είναι η παρακάτω:



Εικόνα 28. Δέντρο απόφασης σχετικά με το αν ένας πελάτης που εισέρχεται σε ένα εστιατόριο έχει πιθανότητα να παραμείνει στην αναμονή, Πηγή: www.cs.uoi.gr

3.4.5.4 Αλγόριθμος Random Forest

Ο αλγόριθμος random forest είναι μια εξαιρετικά ισχυρή μέθοδος συνόλου (ensemble method) για την πρόβλεψη μοντέλων. Συγκεκριμένα, τα random forests συνδυάζουν πολλαπλά Δέντρα Απόφασης, που αναλύθηκαν παραπάνω, ώστε να δημιουργήσουν ένα ισχυρό και ακριβές μοντέλο. Κάθε δέντρο στο δάσος εκπαιδεύεται σε ένα τυχαίο υπο-σύνολο δεδομένων των αρχικών δεδομένων και χαρακτηριστικών. Η τελική πρόβλεψη είναι ουσιαστικά η μέση τιμή μια ψήφου της πλειοψηφίας (majority vote) των μεμονωμένων προβλέψεων κάθε δέντρου.

Αξίζει να αναφερθεί η μελέτη του Louppe (Louppe, 2014) για τα random forests, η οποία παρέχει μια ολοκληρωμένη ανάλυση τους. Δίνει ιδιαίτερη έμφαση στην κατανόηση των μηχανισμών και των περιορισμών του αλγορίθμου. Σύμφωνα με αυτήν την ανάλυση, καταλήγει στα εξής:

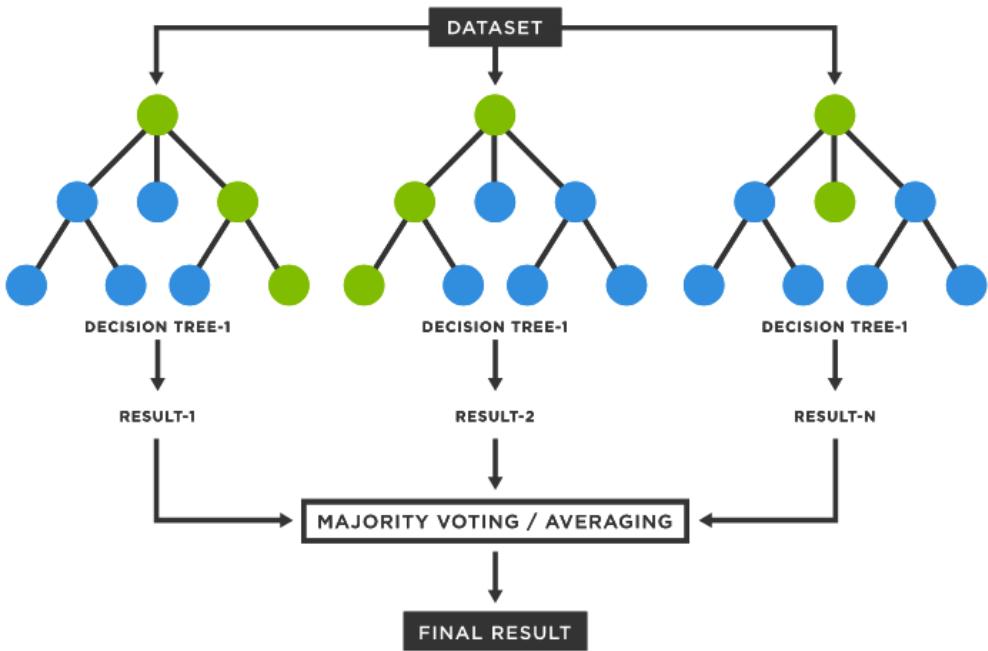
- **Δέντρα απόφασης:** Τα random forests δημιουργούνται από δέντρα απόφασης. Αυτά τα δέντρα κατασκευάζονται με αναδρομικό διαχωρισμό δεδομένων που βασίζονται στις τιμές των χαρακτηριστικών (features).
- **Σχεδιασμός συνόλου (ensemble design):** Τα random forest δημιουργούν σύνολα τυχαιοποιημένων δέντρων. Η τυχαιότητα αυτή διασφαλίζει την ποικιλομορφία μεταξύ των δέντρων, οδηγώντας σε ακόμα καλύτερη γενίκευση.

- Ανάλυση πολυπλοκότητας: Η έρευνα του Louppre περιλαμβάνει μια ανάλυση πολυπλοκότητας, που αποδεικνύει ότι τα τυχαία δάση αποδίδουν καλά υπολογιστικά και κλιμακώνονται αποτελεσματικά. Συζητά επίσης λεπτομέρειες εφαρμογής τους στη βιβλιοθήκη Scikit-Learn της Python.
- Σημασία μεταβλητής: Ο Louppre εμβαθύνει επίσης σε μέτρα σημασίας μεταβλητής. Σημειωτέον, χαρακτηρίζει το μέτρο «Mean Decrease of Impurity», αποκαλύπτοντας τις ιδιότητές του υπό διαφορετικές συνθήκες.
- Περιορισμοί: Ο Louppre έρχεται αντιμέτωπος με τους περιορισμούς που σχετίζονται με μεγάλα σύνολα δεδομένων, προτείνοντας την υποδειγματοληψία ως λύση.

Η αρχική δημοσίευση του Breiman (Breiman, 2001), που αναφέρθηκε και νωρίτερα, εισήγαγε την έννοια των random forests. Τα βασικά σημεία περιλαμβάνουν:

- Bootstrap Aggregating (Bagging): Τα τυχαία δάση χρησιμοποιούν το bagging για να δημιουργήσουν διαφορετικά σετ εκπαίδευσης για κάθε δέντρο.
- Τυχαία επιλογή χαρακτηριστικών (Random Feature Selection): Σε κάθε διαχωρισμό, λαμβάνεται υπόψη μόνο ένα τυχαίο υπο-σύνολο χαρακτηριστικών. Αυτή η τυχαιότητα μειώνει σε μεγάλο βαθμό την υπερπροσαρμογή (overfitting) των δεδομένων.
- Μηχανισμός Ψηφοφορίας (voting mechanism): Η τελική πρόβλεψη λαμβάνεται με ψήφο πλειοψηφίας σε όλα τα δέντρα.
- Ανθεκτικότητα: Τα τυχαία δάση χειρίζονται αποτελεσματικά τα δεδομένα με «θόρυβο» (noisy data) και τις ακραίες τιμές.

Παρακάτω η σχηματική αναπαράσταση ενός αλγορίθμου Random Forest:



Εικόνα 29. Σχηματική αναπαράσταση αλγορίθμου Random Forest, Πηγή: <https://www.spotfire.com/glossary/what-is-a-random-forest>

3.4.5.5 Αλγόριθμος Logistic Regression (Λογιστική Παλινδρόμηση)

Η Λογιστική Παλινδρόμηση είναι μια στατιστική μέθοδος που χρησιμοποιείται συνήθως για προβλήματα δυαδικής ταξινόμησης (binary classification). Αναλυτικότερα:

- Ορισμός και σκοπός: Η λογιστική παλινδρόμηση ασχολείται με σύνολα δεδομένων, αναλύοντας τα με βάση μία ή περισσότερες ανεξάρτητες μεταβλητές προκειμένου να προβλέψει τα επιθυμητά αποτελέσματα. Το αποτέλεσμα ποσοτικοποιείται χρησιμοποιώντας μια διωνυμική μεταβλητή απόκρισης, καθώς υπάρχουν μόνο δύο πιθανά αποτελέσματα (π.χ., ναι/όχι, επιτυχία/αποτυχία). Στοχεύει στην εύρεση ενός μοντέλου που εξηγεί καλύτερα τη σχέση μεταξύ των ανεξάρτητων παραγόντων και ενός δυαδικού χαρακτηριστικού ενδιαφέροντος.
- Ομοιότητες με τη Γραμμική παλινδρόμηση: Με μια διωνυμική μεταβλητή απόκρισης, η Γραμμική παλινδρόμηση και η Λογιστική Παλινδρόμηση συμπεριφέρονται παρόμοια. Ωστόσο, η Λογιστική Παλινδρόμηση έχει πλεονεκτήματα: Μπορεί να χειριστεί περισσότερες από δύο επεξηγηματικές μεταβλητές ταυτόχρονα και περιλαμβάνει μη συνεχείς επεξηγηματικές μεταβλητές.
- Δημιουργία και πρόβλεψη του μοντέλου: Η Λογιστική Παλινδρόμηση δημιουργεί έναν τύπο που βασίζεται σε συγκεκριμένους συντελεστές. Αυτός ο τύπος προβλέπει το

λεγόμενο logit transformation, υποδεικνύοντας την πιθανότητα να υπάρχει το χαρακτηριστικό ενδιαφέροντος. Προσαρμόζεται καλά σε διάφορες ερευνητικές περιπτώσεις, καθιστώντας το μια συνήθη επιθυμητή ερευνητική προσέγγιση.

- **Εφαρμογές:** Η Λογιστική Παλινδρόμηση χρησιμοποιείται ευρέως στην εξόρυξη δεδομένων (data mining) και στην κατηγοριοποίηση δυαδικών δεδομένων. Ισχύει για πολλαπλές εξαρτημένες μεταβλητές, όχι μόνο διχοτομικές. Οι ερευνητές το χρησιμοποιούν για να διερευνήσουν τις σχέσεις μεταξύ των ανεξάρτητων παραγόντων και των αποτελεσμάτων. (Kaur & Himanshu, 2023).

Σχετικά με τη χρήση της Λογιστικής Παλινδρόμησης για την ταξινόμηση σε πολλαπλές κλάσεις, που είναι και αυτή που χρησιμοποιήθηκε στην παρούσα διπλωματική εργασία, μπορούμε να αναφέρουμε τα εξής (Abramovich et al., 2021):

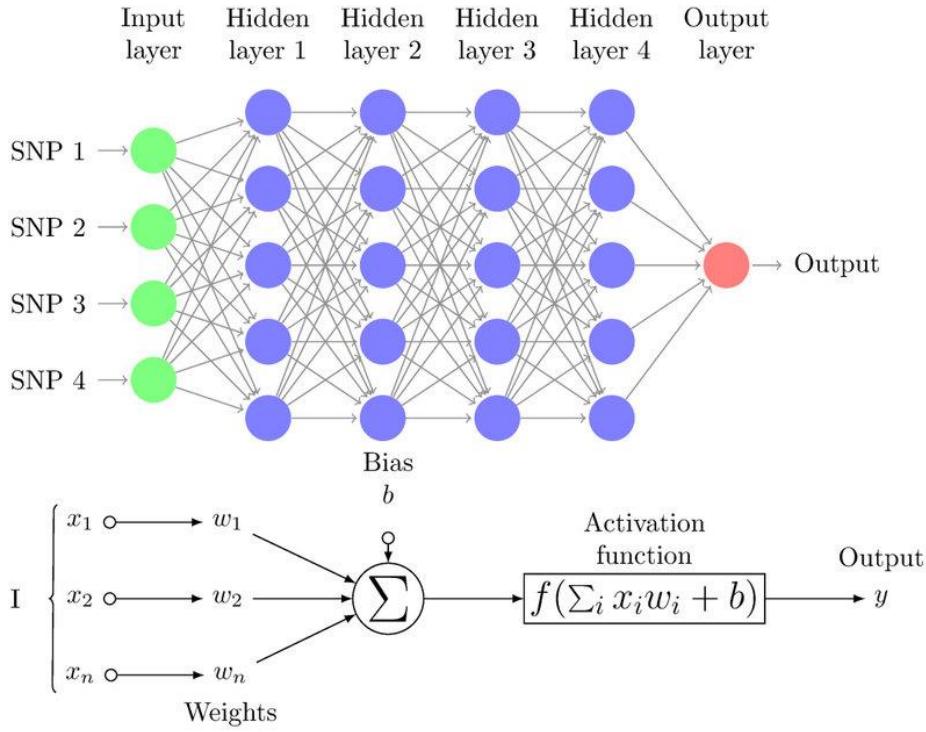
- **Σποραδική πολυωνυμική λογιστική παλινδρόμηση (sparse multinomial logistic regression):** Στην ταξινόμηση πολλαπλών κλάσεων υψηλής διάστασης, οι ερευνητές έχουν προτείνει μια σποραδική πολυωνυμική προσέγγιση λογιστικής παλινδρόμησης. Η μέθοδος περιλαμβάνει την επιλογή χαρακτηριστικών χρησιμοποιώντας μια ποινικοποιημένη μέγιστη πιθανότητα (penalized maximum likelihood) με ποινή πολυπλοκότητας στο συνολικό μέγεθος του μοντέλου. Τα μη συμπτωτικά όρια για τον κίνδυνο λανθασμένης ταξινόμησης προκύπτουν από τον τελικό ταξινομητή που επιλέγεται. Αυτά τα όρια φαίνεται να είναι στενά, με μια μετάβαση φάσης μεταξύ μικρών και μεγάλων αριθμών κλάσεων. Η προσέγγιση μπορεί να βελτιστοποιηθεί περαιτέρω σε συνθήκες χαμηλού «θορύβου».
- **Υπολογιστική σκοπιμότητα (computational feasibility):** Η εύρεση μιας ποινικοποιημένης λύσης μέγιστης πιθανότητας (penalized maximum likelihood solution) με ποινές πολυπλοκότητας, απαιτεί μια συνδυαστική αναζήτηση για όλα τα πιθανά μοντέλα. Για να αντιμετωπιστεί αυτό το πρόβλημα, οι ερευνητές προτείνουν ταξινομητές πολυωνυμικής λογιστικής παλινδρόμησης Group Lasso και ταξινομητές Slope. Αυτοί οι ταξινομητές επιτυχάνουν το minimax order (όταν ο «αντίπαλος» του αλγόριθμου προσπαθεί να ελαχιστοποιήσει οποιαδήποτε τιμή προσπαθεί να μεγιστοποιήσει ο αλγόριθμος) ενώ παραμένουν υπολογιστικά εφικτοί για δεδομένα υψηλής διάστασης.
- **Πρακτικές εφαρμογές:** Η σποραδική πολυωνυμική λογιστική παλινδρόμηση είναι πολύτιμη σε διάφορους τομείς:
 - Εξόρυξη δεδομένων: Βοηθά στην κατηγοριοποίηση δεδομένων με πολλαπλές εξαρτημένες μεταβλητές.

- Έρευνα: Διερευνά τις σχέσεις μεταξύ των ανεξάρτητων παραγόντων και των αποτελεσμάτων.
- Μηχανική μάθηση: Είναι ένα ισχυρό εργαλείο για την ταξινόμηση πολλαπλών τάξεων

3.4.5.6 Αλγόριθμος *Multi-Layer Perceptron (MLP)*

Το multilayer perceptron είναι ένας τύπος τεχνητού εμπροσθοτροφοδοτούμενου (feedforward) νευρωνικού δικτύου. Η αρχική του τοποθέτηση ήταν από τον Rosenblatt στα τέλη της δεκαετίας του '50 (Rosenblatt, 1958). Αποτελείται από πλήρως συνδεδεμένους νευρώνες οργανωμένους σε τουλάχιστον τρία επίπεδα: ένα στρώμα εισόδου (input layer), ένα ή περισσότερα κρυφά στρώματα (hidden layer) και ένα στρώμα εξόδου (output layer). Όσον αφορά τη δομή του, το στρώμα εισόδου λαμβάνει τα χαρακτηριστικά (features) εισόδου. Το κρυφό στρώμα περιέχει ένα ή περισσότερα στρώματα διασυνδεδεμένων νευρώνων που επιτρέπουν στο νευρωνικό δίκτυο να μάθει περίπλοκες αναπαραστάσεις, ενώ το στρώμα εξόδου παράγει τις τελικές προβλέψεις ή ταξινομήσεις. Όσον αφορά τον τρόπο ενεργοποίησης του, σε αντίθεση με το πρωτότυπο perceptron (το οποίο χρησιμοποιούσε μια συνάρτηση βήματος Heaviside), τα MLPs χρησιμοποιούν μια μη-γραμμική συνάρτηση ενεργοποίησης. Οι πιο συνήθεις επιλογές περιλαμβάνουν σιγμοειδείς συναρτήσεις ή rectified linear units (ReLU). Η συνάρτηση ενεργοποίησης εισάγει τη μη-γραμμικότητα, επιτρέποντας στο δίκτυο να μοντελοποιεί σύνθετες σχέσεις στα δεδομένα. Σχετικά με τον τρόπο εκπαίδευσης, τα MLP εκπαιδεύονται χρησιμοποιώντας τον αλγόριθμο ανάστροφης διάδοσης (backpropagation). Η ανάστροφη διάδοση περιλαμβάνει την προσαρμογή των «βαρών» των συνδέσεων με βάση το σφάλμα μεταξύ των προβλεπόμενων εξόδων και των πραγματικών στόχων. Η μέθοδος βελτιστοποίησης gradient descent χρησιμοποιείται συνήθως για την επαναπροσαρμογή των βαρών κατά τη διάρκεια της εκπαίδευσης. Τα MLPs είναι universal approximators, που σημαίνει ότι μπορούν να προσεγγίσουν οποιαδήποτε συνεχή συνάρτηση όταν διαθέτουν επαρκή αριθμό κρυφών νευρώνων. Με τη στοίβαξη πολλαπλών στρωμάτων, τα MLPs μπορούν να μάθουν ιεραρχικά χαρακτηριστικά από ακατέργαστα δεδομένα εισόδου (Murtagh, 1991).

Παρακάτω απεικονίζεται σχηματικά ένα Multi-layer Perceptron:



Εικόνα 30. Σχηματική αναπαράσταση Multi-Layer Perceptron, Πηγή: (Pérez-Enciso & Zingaretti, 2019)

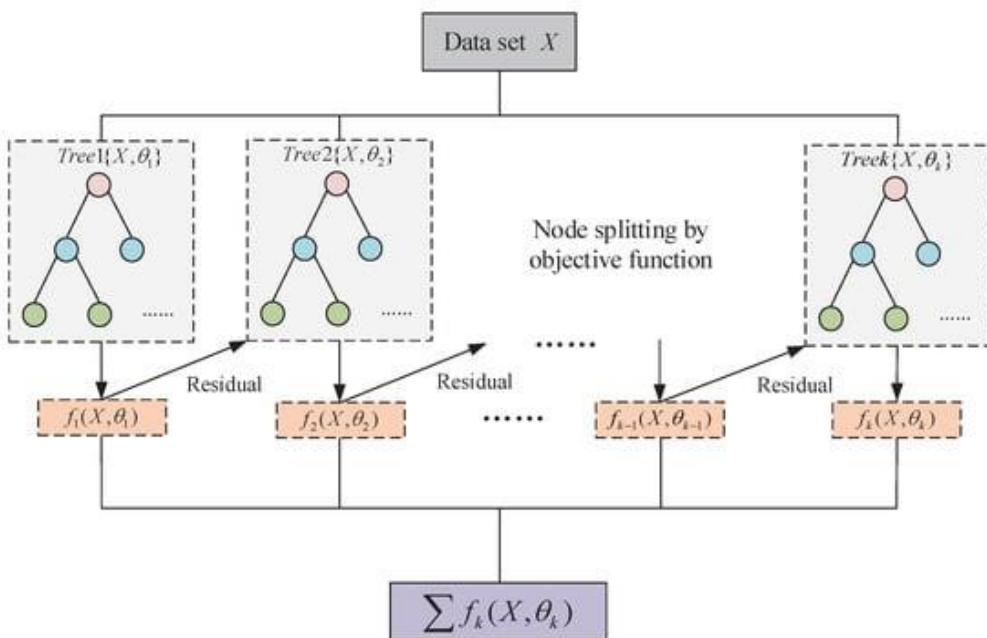
3.4.5.7 Αλγόριθμος XGBoost

Ο αλγόριθμος XGBoost (eXtreme Gradient Boosting) πρόκειται για μια αρκετά ισχυρή βιβλιοθήκη μηχανικής μάθησης που έχει σχεδιαστεί για την αποτελεσματικότερη καθώς και επεκτάσιμη εκπαίδευση μοντέλων μηχανικής εκμάθησης. Συνδυάζει τις προβλέψεις πολλών «αδύναμων» μοντέλων (συνήθως δέντρα αποφάσεων) με σκοπό να παράγει μια ισχυρότερη πρόβλεψη. Τα βασικά χαρακτηριστικά του περιλαμβάνουν την αποτελεσματική διαχείριση τιμών που λείπουν (missing values) και την ενσωματωμένη υποστήριξη για παράλληλη επεξεργασία (parallel processing). Ο τρόπος λειτουργίας του είναι ο εξής ("XGBoost," 2021):

- **Δέντρα απόφασης:** Για να κατανοήσουμε το XGBoost, θα πρέπει να θυμηθούμε τα Δέντρα Απόφασης. Ένα Δέντρο Απόφασης, όπως αναφέρθηκε νωρίτερα, είναι μια δομή που μοιάζει με διάγραμμα ροής όπου κάθε εσωτερικός κόμβος αντιπροσωπεύει μια δοκιμή σε ένα χαρακτηριστικό, οι κλάδοι αντιπροσωπεύουν τα αποτελέσματα της δοκιμής και οι κόμβοι φύλλων αντιπροσωπεύουν τις κλάσεις.
- **Gradient Boosting:** Ο XGBoost χρησιμοποιεί το λεγόμενο gradient boosting, το οποίο δημιουργεί διαδοχικά δέντρα απόφασεων. Έτσι, βελτιστοποιεί τη διαδικασία εκπαίδευσης τη συνάρτηση απώλειας (loss function).

- Εκμάθηση συνόλου (ensemble learning): Το XGBoost συνδυάζει πολλαπλά δέντρα αποφάσεων («αδύναμα» μοντέλα) για να δημιουργήσει ένα ισχυρό σύνολο. Κάθε δέντρο ουσιαστικά διορθώνει τα λάθη των προηγούμενων.
- Bagging: Το XGBoost χρησιμοποιεί τη μέθοδο bagging που αναφέρθηκε νωρίτερα, όπου οι βασικοί ταξινομητές (δέντρα) εκπαιδεύονται σε τυχαία υπο-σύνολα του συνόλου δεδομένων. Οι μεμονωμένες προβλέψεις συγκεντρώνονται με βάση το μέσο όρο για να σχηματίσουν την τελική πρόβλεψη.
- Στάθμιση: Η στάθμιση παίζει κρίσιμο ρόλο στον αλγόριθμο XGBoost, καθώς δίνει έμφαση στη σημασία ορισμένων δειγμάτων κατά τη διάρκεια της εκπαίδευσης.

Συνολικά, ο αποτελεσματικός χειρισμός των missing values, η παράλληλη επεξεργασία και η προσέγγιση συνόλου που προσφέρει ο αλγόριθμος XGBoost, τον καθιστούν την ιδανική επιλογή για ποικίλα προβλήματα μηχανικής μάθησης (Brownlee, 2016; Simplilearn, 2022; “XGBoost,” 2021).



Εικόνα 31. Σχηματική αναπαράσταση του αλγορίθμου XGBoost. Πηγή: (Guo et al., 2020)

3.4.6 SHAP (SHapley Additive exPlanations)

Στην προηγούμενη ενότητα εξηγήσαμε πως μπορεί η ανάλυση SHAP να συνεισφέρει στην καλύτερη κατανόηση της σχέσης των χαρακτηριστικών ενός μοντέλου. Πιο συγκεκριμένα, στο πλαίσιο της πρόβλεψης της σοβαρότητας των παρενεργειών των εμβολίων κατά της COVID-19 χρησιμοποιώντας αλγόριθμους μηχανικής μάθησης, η ανάλυση SHAP μπορεί να προσφέρει πολλά οφέλη, μεταξύ των οποίων είναι:

- Κατανόηση των προγνωστικών παραγόντων: Η ανάλυση SHAP μπορεί να προσδιορίσει ποια χαρακτηριστικά εμβολίου (π.χ. τύπος εμβολίου, δοσολογία) καθώς και μεμονωμένα χαρακτηριστικά του ασθενούς (π.χ. ηλικία, φύλο, υποκείμενα νοσήματα) συμβάλλουν περισσότερο στην πρόβλεψη της σοβαρότητας των παρενεργειών.
- Εξατομικευμένη εκτίμηση κινδύνου: Παρέχοντας απαντήσεις για μεμονωμένες προβλέψεις, η ανάλυση SHAP θα μπορεί να οδηγήσει στην εξατομικευμένη εκτίμηση κινδύνου, επιτρέποντας στους επαγγελματίες υγείας να προσαρμόσουν τη φροντίδα και την
- Πολιτική Δημόσιας Υγείας: Τα συγκεντρωτικά αποτελέσματα της ανάλυσης SHAP θα μπορούσαν να παρέχουν πληροφορίες για τις τάσεις σε πληθυσμιακό επίπεδο σχετικά με τη σοβαρότητα των παρενεργειών των εμβολίων, την ενημέρωση των πολιτικών δημόσιας υγείας και των στρατηγικών εμβολιασμού.

Στην παρούσα διπλωματική εξετάσαμε πρώτα τα αποτελέσματα ακρίβειας για κάθε αλγόριθμο ταξινόμησης σε διάφορα υποσύνολα δεδομένων και στη συνέχει εφαρμόσαμε την ανάλυση SHAP στο μοντέλο που είχε εκπαιδευτεί με το μεγαλύτερο σύνολο δεδομένων και με τα καλύτερα αποτελέσματα, ώστε να διαπιστώσουμε και να αποκαλύψουμε πιθανές σχέσεις μεταξύ των χαρακτηριστικών. Η εφαρμογή πραγματοποιήθηκε με τη βοήθεια της βιβλιοθήκης shap που είναι διαθέσιμη για τη γλώσσα Python (Lundberg, 2018).

3.5 ΜΕΤΡΙΚΕΣ ΑΞΙΟΛΟΓΗΣΗΣ

Μετά την ανάπτυξη του μοντέλου, η επόμενη φάση είναι να υπολογιστεί η απόδοση του αναπτυγμένου αυτού μοντέλου χρησιμοποιώντας ορισμένες μετρικές αξιολόγησης. Για να υπολογιστούν αυτές οι μετρικές, ουσιαστικά συγκρίνονται οι τιμές πρόβλεψης κάθε αλγορίθμου για κάθε εγγραφή με την πραγματική τιμή της κλάσης της εγγραφής. Τα αποτελέσματα συνοψίζονται αρχικά σε ένα πίνακα που λέγεται «πίνακας ταξινόμησης» (classification matrix). Ο πίνακας ταξινόμησης είναι ένα εργαλείο για τον προσδιορισμό της απόδοσης του ταξινομητή. Περιέχει πληροφορίες σχετικά με πραγματικές και προβλεπόμενες ταξινομήσεις. Παρακάτω φαίνεται η δομή ενός απλού πίνακα ταξινόμησης για ταξινόμηση τύπου «θετικό/αρνητικό», μαζί με κάποιες σημαντικές μετρικές αξιολόγησης:

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
	Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$	

Εικόνα 32. Πίνακας ταξινόμησης (confusion matrix), Πηγή: <https://encord.com/glossary/confusion-matrix/>

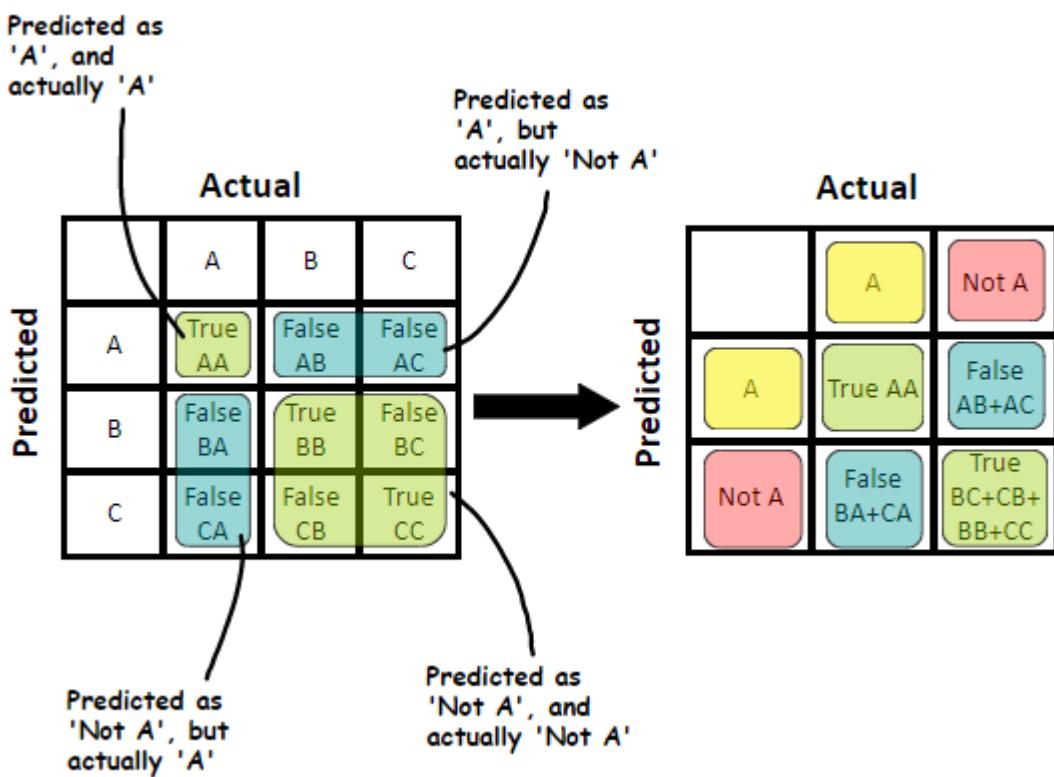
Ας εμβαθύνουμε περισσότερο στα δομικά στοιχεία αυτού του πίνακα εξηγώντας τις βασικές έννοιες (Encord, 2023):

- True Positives (TP): Είναι οι περιπτώσεις όπου το μοντέλο προβλέπει σωστά μια θετική κλάση όταν είναι όντως θετική. Αν σκεφτούμε για παράδειγμα ένα διαγνωστικό μοντέλο καρκίνου, ένα αληθινό θετικό θα προέκυπτε όταν το μοντέλο προσδιορίσει σωστά έναν ασθενή με καρκίνο ως πάσχοντα από τη νόσο. Το TP είναι ένα ζωτικό μέτρο της ικανότητας του μοντέλου να αναγνωρίζει θετικές περιπτώσεις με ακρίβεια.
- True Negatives (TN): Είναι οι περιπτώσεις όπου το μοντέλο προβλέπει σωστά μια αρνητική κλάση όταν είναι όντως αρνητική. Συνεχίζοντας την ιατρική αναλογία, ένα πραγματικό αρνητικό θα ήταν όταν το μοντέλο προσδιορίζει σωστά έναν υγιή ασθενή ότι δεν έχει τη νόσο. Το TN αντικατοπτρίζει την ικανότητα του μοντέλου στην αναγνώριση αρνητικών περιπτώσεων.
- False Positives (FP): Περιπτώσεις όπου το μοντέλο προβλέπει λανθασμένα μια θετική κλάση όταν θα έπρεπε να ήταν αρνητική. Στο παραπάνω ιατρικό σενάριο, ένα ψευδών θετικό θα σήμαινε ότι το μοντέλο υποδεικνύει λανθασμένα ότι ένας ασθενής έχει τη νόσο όταν είναι στην πραγματικότητα υγιής. Το FP απεικονίζει περιπτώσεις όπου το μοντέλο εμφανίζει υπερβολική εμπιστοσύνη στην πρόβλεψη θετικών αποτελεσμάτων.
- False Negatives (FN): Περιπτώσεις όπου το μοντέλο προβλέπει λανθασμένα μια αρνητική κλάση όταν θα έπρεπε να ήταν θετική. Στο ιατρικό πλαίσιο, ένα ψευδών αρνητικό θα ήταν

όταν το μοντέλο αποτυγχάνει να εντοπίσει μια ασθένεια σε έναν ασθενή που την έχει πραγματικά. Το FN υπογραμμίζει καταστάσεις όπου το μοντέλο αποτυγχάνει να συλλάβει πραγματικές θετικές περιπτώσεις.

Ο πίνακας ταξινόμησης λοιπόν χρησιμεύει ως βάση για τον υπολογισμό βασικών μετρικών αξιολόγησης που προσφέρουν ποικίλες πληροφορίες για την απόδοση ενός μοντέλου.

Σημειώνεται πως η διαδικασία αυτή είναι δυνατό να γενικευθεί και για κλάσεις με παραπάνω από δύο τιμές. Καθώς μπορεί να γίνει πιο περίπλοκο όταν υπάρχουν πάρα πολλές κλάσεις, ουσιαστικά κάθε κλάση μπορεί να διαχωριστεί σε έναν ενιαίο πίνακα ταξινόμησης, ως εξής:

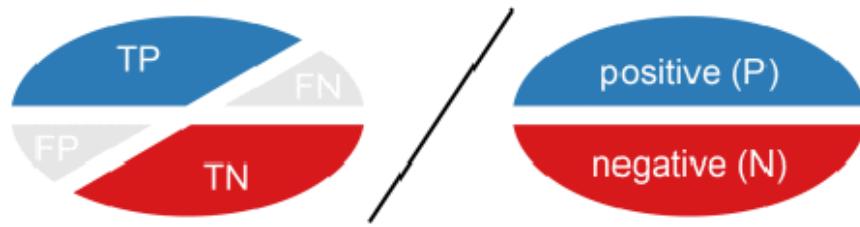


Εικόνα 33. Πίνακας ταξινόμησης για πολυωνυμική κλάση. Πηγή: <https://bassantqz30.medium.com/performance-metrics-for-classification-models-in-machine-learning-part-ii-9303a1c7cadd>

3.5.1 Accuracy

Η ακρίβεια (accuracy) υπολογίζεται ως το άθροισμα δύο ακριβών προβλέψεων (TP + TN) διαιρούμενο με τον συνολικό αριθμό συνόλων δεδομένων (P + N). Ουσιαστικά, η καλύτερη ακρίβεια είναι 1 και η χειρότερη είναι 0 (Tharwat, 2020). Η μετρική αυτή θα μπορούσαμε να πούμε ότι αντιπροσωπεύει την εγκυρότητα των συνολικών προβλέψεων, δηλαδή πόσες από τις τιμές προβλέφθηκαν σωστά.

Accuracy: $(TP + TN) / (P + N)$



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

Εικόνα 34. Σχηματική αναπαράσταση και τύπος της μετρικής αξιολόγησης Ακρίβεια (Accuracy), Πηγή: (Vujović, 2021)

3.5.2 Recall

TP Rate – True Positive Rate (Sensitivity or Recall) υπολογίζεται ως ο αριθμός των ακριβών θετικών προβλέψεων (TP) διαιρούμενο με τον συνολικό αριθμό των θετικών (P). Το καλύτερο TP Rate είναι 1 και το χειρότερο 0, όπως και προηγουμένως (Tharwat, 2020). Η μετρική αυτή ουσιαστικά αντιπροσωπεύει το πόσες από τις θετικές τιμές προβλέφθηκαν σωστά από τον αλγόριθμο.

Sensitivity: TP / P



$$SN = \frac{TP}{TP + FN} = \frac{TP}{P}$$

Εικόνα 35. Σχηματική αναπαράσταση και τύπος της μετρικής αξιολόγησης Ανάκληση ή Ευαισθησία (Recall or Sensitivity), Πηγή: (Vujović, 2021)

3.5.3 Precision

Η πιστότητα (precision) υπολογίζεται ως ο αριθμός των σωστών θετικών προβλέψεων (TP), διαιρεμένες με τον συνολικό αριθμό των θετικών προβλέψεων (TP + FP). Η καλύτερη ακρίβεια είναι 1 και η χειρότερη 0 (Tharwat, 2020). Η μετρική αυτή αντιπροσωπεύει την εγκυρότητα της θετικής πρόβλεψης, δηλαδή πόσες από τις προβλεπόμενες ως θετικές τιμές είναι όντως θετικές.

Precision: $TP / (TP + FP)$



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

Εικόνα 36. Σχηματική αναπαράσταση και τύπος της μετρικής αξιολόγησης Πιστότητα (Precision), Πηγή: (Vujović, 2021)

3.5.4 F1 score

Το F-Measure ή F-score ή F1 είναι και αυτό ένα μέτρο της ακρίβειας του τεστ. Υπολογίζεται, με βάση την πιστότητα, από τον παρακάτω τύπο:

$$2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

Θα μπορούσαμε να πούμε πως αυτή η μετρική αποτελεί μια ισορροπία μεταξύ πιστότητας και ευαισθησίας. Λαμβάνει επίσης τιμές από 0 έως 1.

Το ποια μετρική είναι καταλληλότερη να επιλέξουμε για να αξιολογήσουμε την απόδοση του εκάστοτε μοντέλου, εξαρτάται σε σημαντικό βαθμό από το πρόβλημα ταξινόμησης και τη βαρύτητα της κάθε κλάσης. Αν για παράδειγμα μας ενδιαφέρει περισσότερο να λάβουμε αξιόπιστα αποτελέσματα για μια συγκεκριμένη κλάση έναντι των άλλων, τότε θέλουμε για αυτήν την κλάση να έχουμε υψηλή ευαισθησία (recall). Αντίθετα, αν θέλουμε να διατηρήσουμε την αξιοπιστία των αποτελεσμάτων για όλες τις κλάσεις, τότε σε αυτήν την περίπτωση θα ήταν προτιμότερο να αυξήσουμε το F1-score για μια δεδομένη κλάση, αντί για την ευαισθησία. Συνηθέστερα, θέλουμε το μοντέλο μας να δίνει αξιόπιστα αποτελέσματα για όλες τις κλάσεις, ακόμα και αν εμείς ενδιαφερόμαστε περισσότερο για μια συγκεκριμένη κλάση, πρέπει να βρούμε μια μετρική που να αντιπροσωπεύει τη συνολική απόδοση του μοντέλου. Υπάρχουν δύο μετρικές που εξυπηρετούν αυτό το σκοπό και είναι οι εξής (Gamal, 2022):

1. Macro Average: ουσιαστικά υπολογίζεται η μέση τιμή για κάθε μετρική στο συνολικό αριθμό των κλάσεων.

- Weighted average: κάθε κλάση λαμβάνει ένα συντελεστή στάθμισης που είναι ανάλογη με τον αριθμό των δειγμάτων που αντιστοιχούν στη συγκεκριμένη κλάση και στη συνέχεια υπολογίζεται η μέση τιμή κάθε μετρικής στο συνολικό αριθμό των δειγμάτων.

3.6 ΓΕΝΙΚΕΣ ΤΕΧΝΙΚΕΣ ΠΛΗΡΟΦΟΡΙΕΣ

Πρέπει να σημειωθεί πως για το parsing του συνόλου δεδομένων, τη δημιουργία της βάσης, καθώς και για τη δημιουργία του αρχείου input για την εκπαίδευση του μοντέλου, χρησιμοποιήθηκε υπερυπολογιστής με λειτουργικό σύστημα Ubuntu 20.04, 14 πυρήνες Intel Xeon Processor (Skylake, IBRS) pc-i440fx-4.2 CPU @2GHz και 128GB RAM. Ο υπερυπολογιστής υποστηρίζεται από τη βιοπληροφορική υποδομή ELIXIR-GR.

4 ΑΠΟΤΕΛΕΣΜΑΤΑ

4.1 ΠΕΡΙΓΡΑΦΗ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ

4.1.1 Συνολικός αριθμός εγγραφών

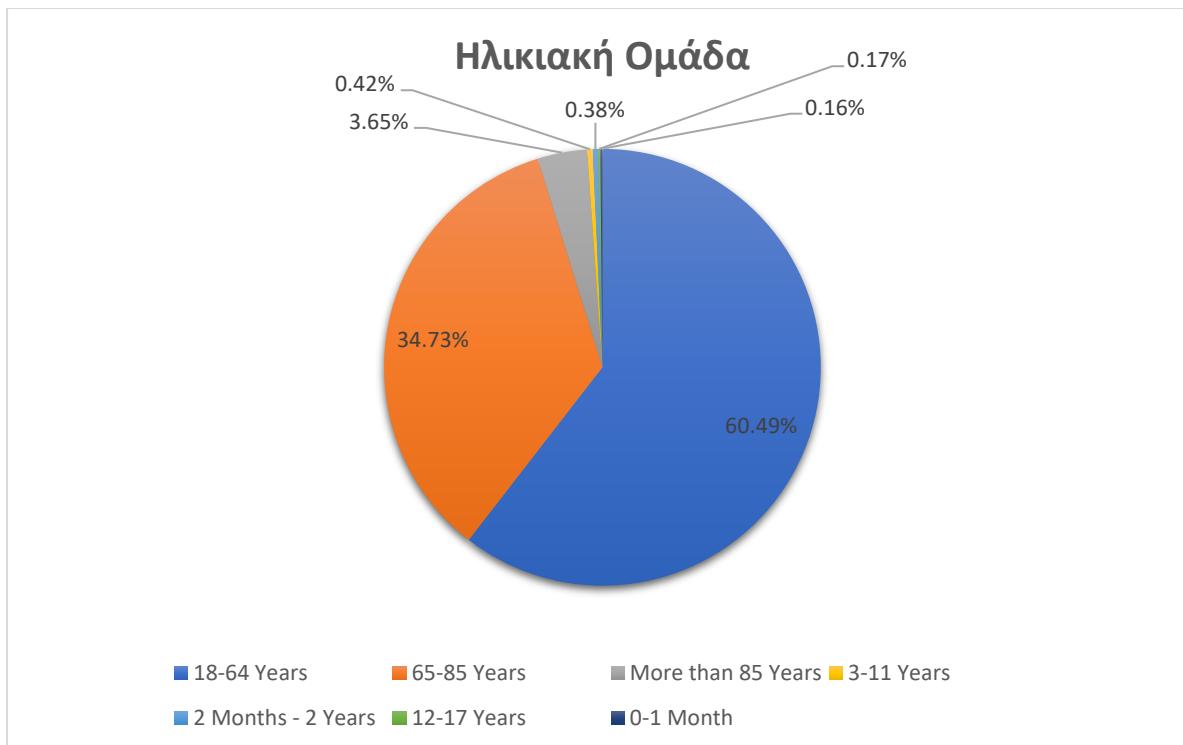
Το αρχικό σύνολο δεδομένων που είχε ανακτηθεί από τη βάση EudraVigilance, είχε συνολικά 977016 εγγραφές, δηλαδή μοναδικούς αριθμούς ταυτοποίησης EU_Local_Number. Παρόλα αυτά, μετά τη δημιουργία της τοπικής βάσης δεδομένων διαπιστώθηκε πως δεν είχαν όλες οι εγγραφές συμπληρωμένο το πεδίο αξιολόγησης της σοβαρότητας των αναφερόμενων παρενεργειών, επομένως αυτές οι εγγραφές δεν έπρεπε να ληφθούν υπόψη στο τελικό σύνολο δεδομένων. Αυτό διότι σκοπός της εργασίας ήταν η κατηγοριοποίηση της σοβαρότητας των παρενεργειών, επομένως αυτή η πληροφορία είναι απαραίτητο να υπάρχει για να πραγματοποιηθεί η εκπαίδευση. Έτσι λοιπόν, με την επιλογή κατάλληλων SQL queries και τη χρήση της γλώσσας PHP επιλέξαμε μόνο τις εγγραφές που είχαν αντίστοιχη τιμή κριτηρίου σοβαρότητας. Το σύνολο αυτών των εγγραφών ήταν 473333, δηλαδή 48.45% (περίπου το μισό) του πρωταρχικού συνόλου δεδομένων.

Ιδανικά, θα θέλαμε να εκπαιδεύσουμε το μοντέλο κατηγοριοποίησης με αυτό το σύνολο δεδομένων, κάτι το οποίο όμως δεν ήταν εφικτό λόγω τεράστιας υπολογιστικής ισχύος που απαιτούνταν εξ αιτίας του μεγάλου όγκου των δεδομένων. Όπως θα αναφερθεί και στην πορεία, το μοντέλο εκπαιδεύτηκε ενδεικτικά για 200, 1000, 10000, 20000, και το μέγιστο 40000 εγγραφές. Στη συνέχεια θα αναφερθούν τα στατιστικά περιγραφικά στοιχεία για το τελικό υποσύνολο δεδομένων που χρησιμοποιήθηκε για την εκπαίδευση, δηλαδή τις 40000 εγγραφές, που αποτελούν περίπου το 8.5% του επιθυμητού συνόλου δεδομένων.

4.1.2 Περιγραφικά στοιχεία

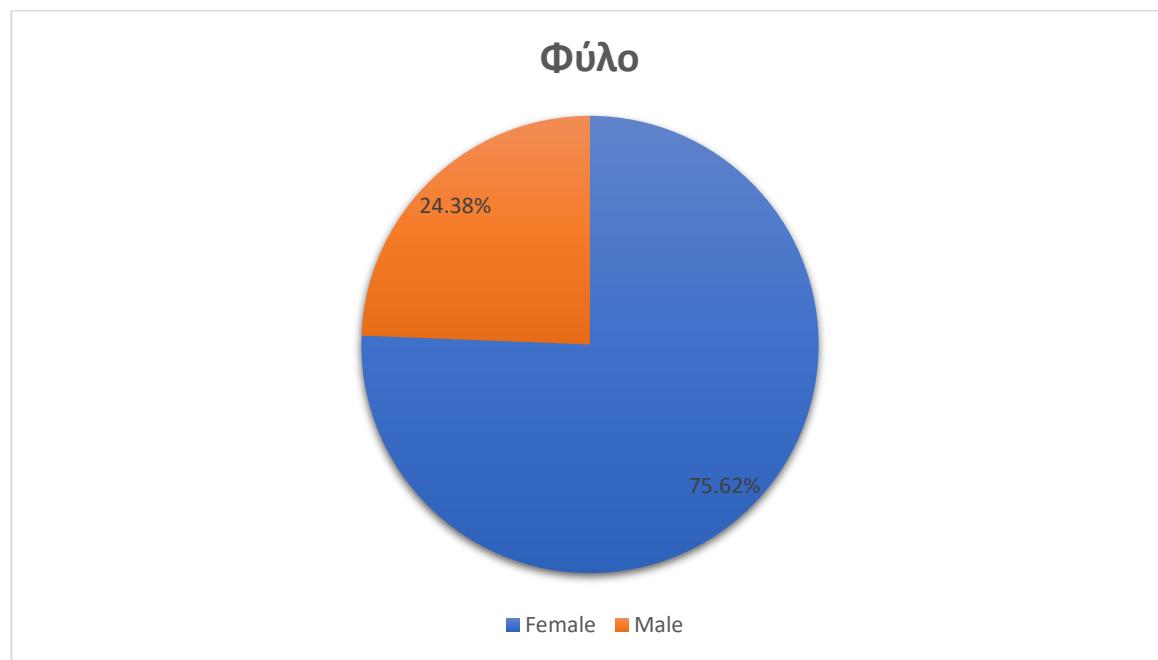
4.1.2.1 Ηλικία και Φύλο

Για το τελικό σύνολο δεδομένων των 40000 εγγραφών, το ποσοστό κάθε ξεχωριστής τιμής για τα χαρακτηριστικά 'Age' και 'Sex' φαίνονται στα παρακάτω διαγράμματα πίτας:



Εικόνα 37. Διάγραμμα πίτας κατανομής της ηλικίας στο σύνολο δεδομένων

Παρατηρούμε πως η ηλικιακή ομάδα 18 – 64 χρόνια είναι και η πολυπληθέστερη στο σύνολο δεδομένων με ποσοστό 60.49% επί του συνόλου και σημαντική διαφορά από τις υπόλοιπες. Ακολουθεί η ηλικιακή ομάδα των 65 – 85 χρονών με ποσοστό 34.73% και οι υπόλοιπες με μικρότερα ποσοστά που κυμαίνονται από 0.16 – 3.65 %.



Εικόνα 38. Διάγραμμα πίτας κατανομής φύλου στο σύνολο των δεδομένων

Παρατηρούμε πως υπάρχει μια υπερεκπροσώπηση στο ποσοστό των γυναικών που ξεπερνάει κατά περίπου 50% το ποσοστό των ανδρών.

4.1.2.2 Drugs, Indications και Reactions

Σχετικά με τον αριθμό των Indications, αυτά που κατέληξαν να είναι συνολικά στο σύνολο δεδομένων ήταν 3904, ο αριθμός των Reactions ήταν 9516 και ο αριθμός των Drugs ήταν 6640 εκ των οποίων τα 4 ήταν τα 4 εμβόλια. Στους παρακάτω πίνακες εμφανίζονται τα 10 indications και τα 10 reactions με το μεγαλύτερο ποσοστό στο σύνολο δεδομένων και που ενδεχομένως έπαιξαν το μεγαλύτερο ρόλο στην εκπαίδευση του μοντέλου:

Πίνακας 6. Η κατανομή των 10 indications με το μεγαλύτερο ποσοστό εμφάνισης στο σύνολο δεδομένων

Όνομα Indication	Άθροισμα	Ποσοστό
COVID-19 immunisation	22853	57.13%
Hypertension	1373	3.43%
Immunisation	1196	2.99%
Asthma	989	2.47%
Depression	710	1.78%
Product used for unknown indication	689	1.72%
Pain	517	1.29%
Hypothyroidism	508	1.27%
Anxiety	489	1.22%
Gastrooesophageal reflux disease	410	1.03%

Παρατηρούμε πως το μεγαλύτερο ποσοστό έχει η ένδειξη 'COVID-19 immunization' που είναι και αναμενόμενο καθώς αναφέρεται στο εκάστοτε εμβόλιο και αμέσως μετά ακολουθεί το 'Hypertension', δηλαδή η υπέρταση που αφορά το 3.43%, το 'Asthma', δηλαδή το άσθμα, που αφορά το 2.47% των ασθενών και το 'Depression', δηλαδή την κατάθλιψη που αφορά το 1.78% των ασθενών.

Πίνακας 7. Η κατανομή των 10 reactions με το μεγαλύτερο ποσοστό εμφάνισης στο σύνολο δεδομένων

Όνομα Reaction	Άθροισμα	Ποσοστό
Headache	13362	33.41%
Pyrexia	11030	27.58%
Fatigue	8404	21.01%
Chills	7694	19.24%
Nausea	6934	17.34%
Myalgia	5873	14.68%
Arthralgia	4228	10.57%
Dizziness	4095	10.24%
Malaise	3091	7.73%
Pain in extremity	3036	7.59%

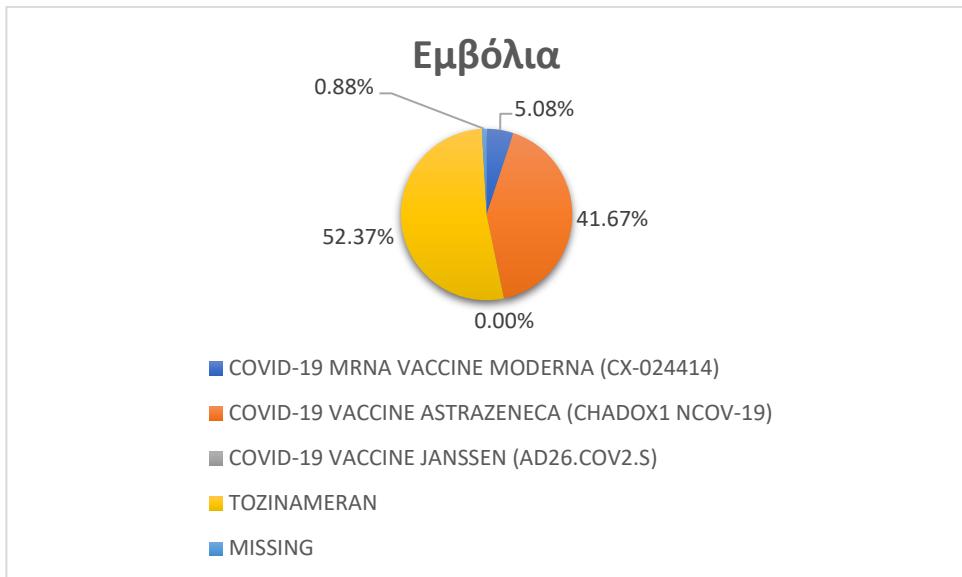
Παρατηρούμε πως το μεγαλύτερο ποσοστό εμφανίζεται στο Reaction ‘Headache’, δηλαδή τον πονοκέφαλο με ποσοστό 33.41% επί του συνόλου, αμέσως μετά ‘Pyrexia’, δηλαδή ο πυρετός με ποσοστό 27.58% επί του συνόλου και ακολουθούν η κόπωση, ρίγος, ναυτία μυαλγία με ποσοστά 21.01%, 19.24% και 17.34% αντίστοιχα.

Πίνακας 8. Η κατανομή των 10 Drugs με το μεγαλύτερο ποσοστό εμφάνισης στο σύνολο δεδομένων

Όνομα Drug	Άθροισμα	Ποσοστό
PARACETAMOL	1997	4.99%
INFLUENZA VIRUS	1769	4.42%
LEVOTHYROXINE SODIUM	1214	3.04%
ATORVASTATIN	1169	2.92%
OMEПRAZOLE	1072	2.68%
AMLODIPINE BESILATE	956	2.39%
AMLODIPINE	937	2.34%
ACETYLSALICYLIC ACID	892	2.23%
LEVOTHYROXINE	882	2.21%
AMLODIPINE MALEATE	850	2.13%

Σημειώνεται πως στον παραπάνω πίνακα δεν έχουν συμπεριληφθεί τα 4 εμβόλια, των οποίων η κατανομή θα αναφερθεί ξεχωριστά. Παρατηρούμε πως το μεγαλύτερο ποσοστό εμφάνισης των συγχορηγούμενων δραστικών ουσιών έχει η παρακεταμόλη με ποσοστό 4.99% επί του συνολικού

πληθυσμού και ακολουθούν διάφορα άλλα όπως λεβιθυροξύνη, η ατορβαστατίνη, η ομερπραζόλη, η αμλοδιπίνη κλπ.

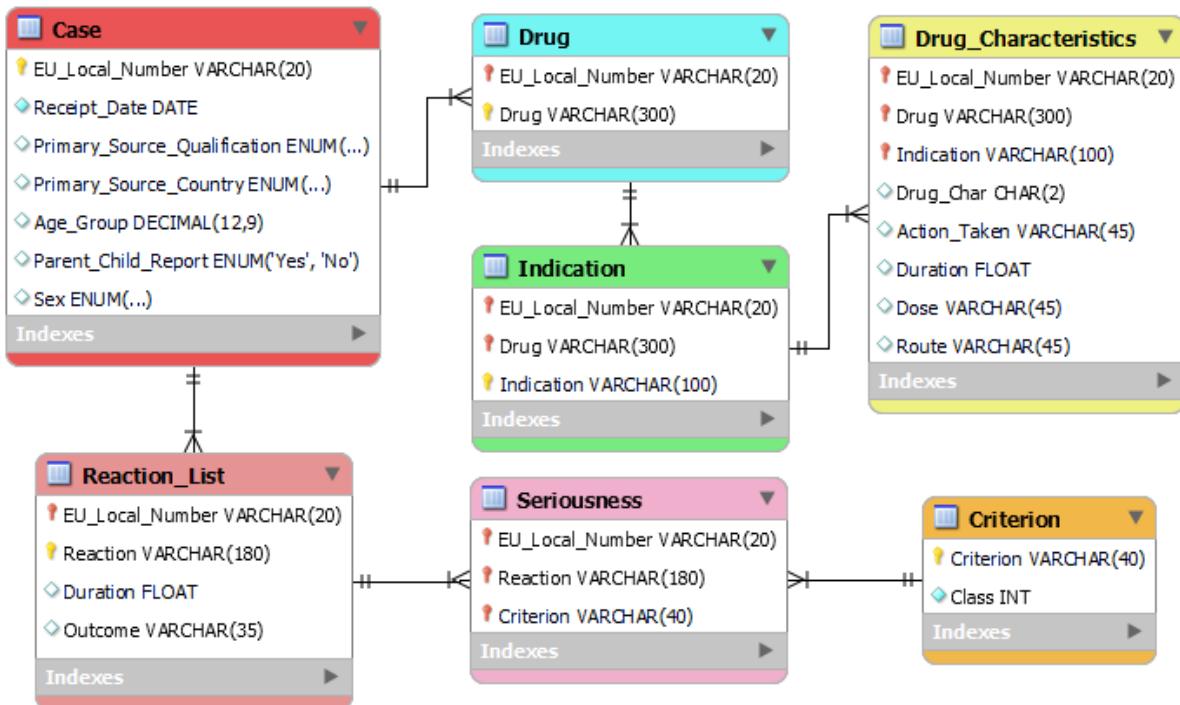


Εικόνα 39. Κατανομή εμβολίων στο σύνολο δεδομένων

Στο παραπάνω διάγραμμα πίτας φαίνεται η κατανομή των 4 εμβολίων στο σύνολο δεδομένων που χρησιμοποιήθηκαν για την εκπαίδευση. Τα εμβόλια της Pfizer-BioNTech και της AstraZeneca είναι αυτά που καταλαμβάνουν το μεγαλύτερο ποσοστό του πληθυσμού με ποσοστά 52.37% και 41.67% αντίστοιχα, ενώ το εμβόλιο της Moderna καταλαμβάνει μόλις το 5.08%. Αξίζει να παρατηρήσουμε πως στο συγκεκριμένο υποσύνολο δεδομένων δεν συμπεριλαμβάνεται το εμβόλιο της Janssen καθώς και ότι υπάρχουν κάποιες εγγραφές που δεν είχαν τιμή στο πεδίο του εμβολίου.

4.2 ΔΙΑΓΡΑΜΜΑ ΟΝΤΟΤΗΤΩΝ – ΣΥΣΧΕΤΙΣΕΩΝ

Όπως αναφέρθηκε νωρίτερα, κατασκευάσαμε κατάλληλο διάγραμμα οντοτήτων συσχετίσεων για τον ορισμό σχέσεων των χαρακτηριστικών του συνόλου δεδομένων με τη χρήση του MySQL Workbench. Το διάγραμμα αυτό απεικονίζεται παρακάτω:



Εικόνα 40. Διάγραμμα Οντοτήτων - Συσχετίσεων του συνόλου δεδομένων όπως απεικονίζεται στο MySQL Workbench

Σημειώνεται πως οι πίνακες σχεδιάστηκαν με τέτοιο τρόπο ώστε να συμπληρωθούν με το περιεχόμενο των αρχείων που είχαν προκύψει μετά το parsing των αρχικών δεδομένων από τη βάση EudraVigilance, όπως περιγράφηκε στην ενότητα 'Μέθοδοι'. Αναλυτικότερα αναφέρονται τα χαρακτηριστικά κάθε πίνακα:

- Πίνακας Case
 1. EU_Local_Number: ένας αριθμός ταυτοποίησης 20 στοιχείων, τύπου VARCHAR
 2. Receipt_Date: η ημερομηνία εισαγωγής της αναφοράς, με μορφή DATE
 3. Primary_Source_Qualification: η ιδιότητα του χρήστη που έκανε την αναφορά, τύπου ENUM με πιθανές τιμές 'Healthcare Professional', 'Non-healthcare Professional', 'Not specified'

4. Primary_Source_Country: η γεωγραφική περιοχή του υποκειμένου με την παρενέργεια, τύπου ENUM με πιθανές τιμές 'European Economic Area', 'Non-european Economic Area', 'Not specified'
5. Age Group: η ηλικιακή ομάδα που ανήκει το υποκείμενο με τις παρενέργειες σε μορφή δεκαδικού αριθμού με το πολύ 12 Ψηφία και 9 δεκαδικά Ψηφία μετά την υποδιαστολή, τύπου DECIMAL (12,9)
6. Parent_Child_Support: αν η αναφορά έχει γίνει για παιδί εκ μέρους του γονέα, τύπου ENUM με πιθανές τιμές 'Yes' ή 'No'
7. Sex: το φύλο του υποκειμένου με τις παρενέργειες, τύπου ENUM με πιθανές τιμές 'Male', 'Female', 'Not specified'

Σημειώνεται πως το χαρακτηριστικό EU_Local_Number αποτελεί πρωτεύον κλειδί (όπως περιγράφηκε στην ενότητα 3.3.4).

- Πίνακας Reaction List
 1. EU_Local_Number: είναι το χαρακτηριστικό που περιγράφηκε στον πίνακα Case το οποίο μεταφέρεται ως ξένο κλειδί στον πίνακα Reaction_List
 2. Reaction: μια αντίδραση που εμφανίστηκε στο υποκείμενο της αναφοράς, τύπου VARCHAR. Σημειώνεται πως το Reaction αποτελεί πρωτεύον κλειδί αυτού του πίνακα συνδέοντας μοναδικά ένα EU_Local_Number με ένα Reaction, ακόμα και στην περίπτωση που σε ένα EU_Local_Number αντιστοιχούν πολλά Reaction.
 3. Duration: η διάρκεια της εκάστοτε αντίδρασης, τύπου FLOAT
 4. Outcome: το αποτέλεσμα που καταγράφηκε για την αντίδραση (π.χ. 'Recovered/Resolved', 'Recovered/Resolved with sequelae' κλπ.), τύπου VARCHAR
- Πίνακας Seriousness
 1. EU_Local_Number: το χαρακτηριστικό αυτό μεταφέρεται ως ξένο κλειδί από τον πίνακα Reaction_List
 2. Reaction: το χαρακτηριστικό αυτό μεταφέρεται επίσης ως ξένο κλειδί από τον πίνακα Reaction_List
 3. Criterion: αποτελεί το κριτήριο σοβαρότητας (πχ. Life threatening, caused/prolonged hospitalisation κλπ.) για την εκάστοτε αντίδραση και αποτελεί ξένο κλειδί που προέρχεται από τον πίνακα Criterion, συνδέοντας μοναδικά μια αντίδραση με ένα Criterion, τύπου VARCHAR.
- Πίνακας Criterion
 1. Criterion: αποτελεί πρωτεύον κλειδί του πίνακα με την παραπάνω περιγραφή

2. Class: το χαρακτηριστικό αυτό αποτελεί την κλίμακα αξιολόγησης που δημιουργήσαμε εμείς προκειμένου να πραγματοποιηθεί η κατηγοριοποίηση των εγγραφών και πρόκειται για διαδοχικά αυξανόμενους ακέραιους αριθμούς από 0 έως 4 ανάλογα με τη σοβαρότητα που εκφράζεται από το χαρακτηριστικό Criterion, τύπου INT. Σημειώνεται πως δύο διαφορετικές τιμές του Criterion μπορούν να ανήκουν στο ίδιο Class, εξού και η σχέση many-to-one.
 - Πίνακας Drug
 1. EU_Local_Number: το χαρακτηριστικό αυτό μεταφέρεται ως ξένο κλειδί από τον πίνακα Case
 2. Drug: το χαρακτηριστικό αυτό περιλαμβάνει τα τέσσερα πιθανά εμβόλια που αναλύθηκαν νωρίτερα, καθώς και οποιοδήποτε άλλο φάρμακο είχε χορηγηθεί παράλληλα με αυτά, τύπου VARCHAR. Αποτελεί πρωτεύον κλειδί του πίνακα συνδέοντας μοναδικά ένα EU_Local_Number με ένα Drug
 - Πίνακας Indication
 1. EU_Local_Number: το χαρακτηριστικό αυτό μεταφέρεται ως ξένο κλειδί από τον πίνακα Drug
 2. Drug: το χαρακτηριστικό αυτό μεταφέρεται επίσης ως ξένο κλειδί από τον πίνακα Drug
 3. Indication: πρόκειται για το πρωτεύον κλειδί του πίνακα και αποτελεί την ένδειξη για την οποία χορηγήθηκε το εκάστοτε φάρμακο στο υποκείμενο (πχ. Pfizer vaccine for COVID-19 immunisation ή paracetamol for headache), τύπου VARCHAR
 - Πίνακας Drug Characteristics
 1. EU_Local_Number: το χαρακτηριστικό αυτό μεταφέρεται ως ξένο κλειδί από τον πίνακα Indication
 2. Drug: το χαρακτηριστικό αυτό μεταφέρεται επίσης ως ξένο κλειδί από τον πίνακα Indication
 3. Indication: το χαρακτηριστικό αυτό μεταφέρεται επίσης ως ξένο κλειδί από τον πίνακα Indication
 4. Drug Char: είναι ένας χαρακτήρας που λαμβάνει τις τιμές 'S' ή 'C', ανάλογα με το αν το εκάστοτε φάρμακο/εμβόλιο θεωρείται ύποπτο (Suspect) για την εκάστοτε παρενέγεια (Reaction) ή απλά συγχορηγούμενο (Concomitant), τύπου CHAR
 5. Action Taken: πρόκειται για πιθανές ενέργειες που ελήφθησαν για την αντιμετώπιση της εκάστοτε παρενέργειας (πχ. Drug withdrawn), τύπου VARCHAR
 6. Duration: πρόκειται για τη διάρκεια παραμονής του φαρμάκου στον οργανισμό, τύπου FLOAT

7. Dose: πρόκειται για τη δόση του φαρμάκου που χορηγήθηκε, τύπου VARCHAR
8. Πρόκειται για τη μέθοδο χορήγησης του φαρμάκου (πχ. Oral), τύπου VARCHAR

4.3 ΠΡΟ-ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Για να αποκτήσουμε μια πιο ολοκληρωμένη άποψη σχετικά με το βαθμό επίδρασης που είχαν οι συγκεκριμένοι feature selectors και balancers που χρησιμοποιήθηκαν κατά την προ-επεξεργασία των δεδομένων, θα πρέπει να γνωρίζουμε το πλήθος των γνωρισμάτων και των εγγραφών που διατηρήθηκαν αντίστοιχα. Εφαρμόσαμε τη συγκεκριμένη άσκηση για το υποσύνολο των 40000 δειγμάτων και λάβαμε τα εξής αποτελέσματα:

- Μετά την εφαρμογή του feature selector SelectPercentile, από τα αρχικά 19868 γνωρίσματα, το σύνολο εκπαιδεύτηκε τελικά με 1986 γνωρίσματα. Αυτό πρακτικά σημαίνει πως το σύνολο των γνωρισμάτων μειώθηκε κατά 90%.
- Μετά την εφαρμογή του balancer TomekLinks, το σύνολο δεδομένων εκπαίδευσης (`x_train`) από το 80% των 40000 εγγραφών, δηλαδή 32000 εγγραφές, κατέληξε να μετρά 31631 εγγραφές, δηλαδή το σύνολο των εγγραφών μειώθηκε κατά ~1.15%.

4.4 ΚΛΑΣΗ ΤΟΥ ΜΟΝΤΕΛΟΥ ΤΑΞΙΝΟΜΗΣΗΣ

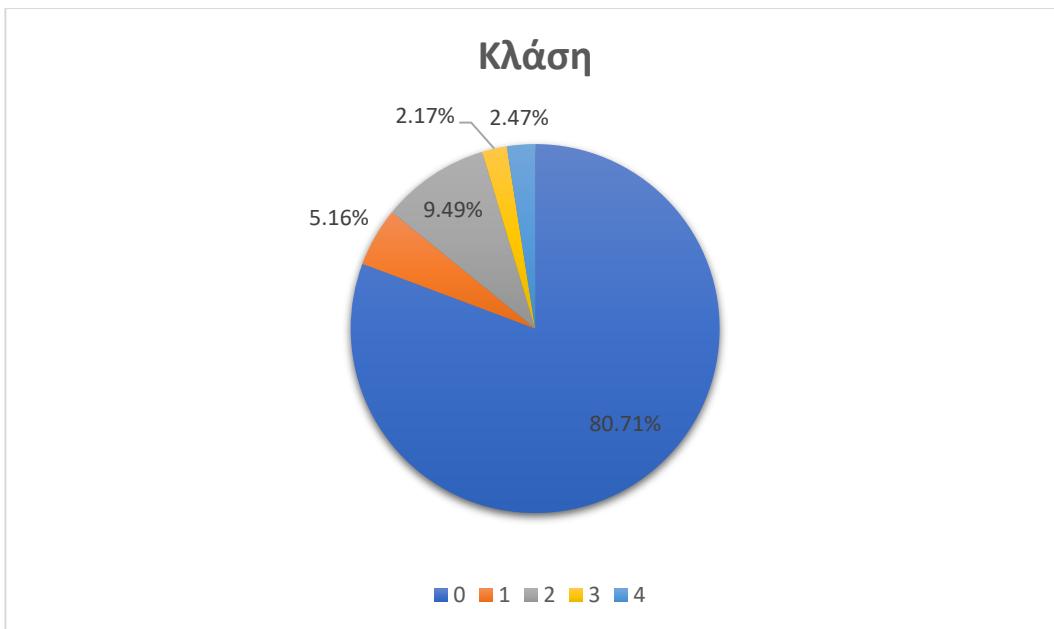
Ένα από τα πιο ουσιώδη ζητήματα που κληθήκαμε να αντιμετωπίσουμε είναι η επιλογή της κλάσης που θα αποτελεί και τον απώτερο σκοπό της κατασκευής του μοντέλου ταξινόμησης. Σκοπός της παρούσας διπλωματικής εργασίας ήταν η πρόβλεψη της σοβαρότητας των παρενεργειών των εμβολίων υπό μελέτη, επομένως η κλάση θα έπρεπε να περιλαμβάνει τιμές που να αντικατοπτρίζουν αυτή τη σοβαρότητα. Στο σύνολο δεδομένων που είχαμε ανακτήσει από τη βάση EudraVigilance, υπήρχαν διαφορετικά πεδία που περιείχαν πληροφορία σοβαρότητας, επομένως έπρεπε να επιλέξουμε το καταλληλότερο και πιο αντιπροσωπευτικό. Επιλέχθηκε το πεδίο Seriousness Criteria της στήλης Reaction (με βάση τη δομή του πρωτότυπου αρχείου που λαμβάνουμε από τη βάση) που ουσιαστικά αξιολογεί τη σοβαρότητα μιας συγκεκριμένης αντίδρασης. Αντιστοιχίσαμε τις κατηγορικές τιμές που υπήρχαν σε αυτήν την κλάση με ακέραιους αριθμούς από 0 έως 4. Όσο μεγαλύτερος ο ακέραιος αριθμός, τόσο μεγαλύτερη η σοβαρότητα. Η αντιστοίχιση έγινε ως εξής:

Πίνακας 9. Αντιστοίχιση τιμών κλάσης σοβαρότητας με ακέραιο αριθμό

Criterion	Rank
Other Medically Important Condition	0
Caused/Prolonged Hospitalisation	1
Disabling	2
Congenital Anomaly	2
Life Threatening	3
Results in Death	4

Οφείλουμε να σημειώσουμε πως καθώς μια εγγραφή, ένας ασθενής δηλαδή, δύναται να έχει πολλά διαφορετικά Reactions και κάθε ένα από αυτά έχει τη δική του αξιολόγηση. Προκειμένου να μειώσουμε το σύνολο εκπαίδευσης, αλλά και επειδή όπως παρατηρήσαμε πολλές φορές για το ίδιο άτομο κάποια Reactions συνδέονταν με το αποτέλεσμα θανάτου ενώ άλλα συνδέονταν με χαμηλότερα επίπεδα σοβαρότητας που προφανώς είχαν επέλθει πριν το θάνατο, αποφασίσαμε χρησιμοποιήσουμε τη μέθοδο του μεγίστου για την αντιστοίχιση μιας κλάσης στο εκάστοτε άτομο. Πιο συγκεκριμένα, για μια εγγραφή επιλέξαμε να κρατήσουμε το κριτήριο σοβαρότητας με το μέγιστο αριθμό. Αν δηλαδή για παράδειγμα μια εγγραφή είχε τρία διαφορετικά Reactions με Seriousness Criteria 'Disabling', 'Disabling' και 'Life Threatening', δηλαδή τιμές 2,2 και 3 αντίστοιχα για την κλάση, επιλέγονταν η τιμή 3 που είναι το μέγιστο από αυτό το σετ.

Έτσι λοιπόν, στο τελικό υποσύνολο δεδομένων που χρησιμοποιήσαμε (40000 εγγραφές), η αρχική κατανομή των κλάσεων πριν την προεπεξεργασία ήταν η εξής:



Εικόνα 41. Κατανομή των τιμών της κλάσης στο σύνολο δεδομένων

Παρατηρούμε πως η κλάση με τις περισσότερες παρατηρήσεις ήταν η 0, δηλαδή, όπως εξηγήθηκε νωρίτερα, το ‘Other medically important condition’ με ποσοστό 80.71% επί του συνόλου, στη συνέχεια η κλάση 2, δηλαδή το ‘Disabling’ ή το ‘Congenital Anomaly’ με ποσοστό 9.49% επί του συνόλου, έπειτα η κλάση 1, δηλαδή το ‘Caused/Prolonge Hospitalization’ με ποσοστό 5.16% επί του συνόλου, εν συνεχείᾳ η 4, δηλαδή το ‘Results in Death’ με ποσοστό 2.47% επί του συνόλου, και τέλος η 3, δηλαδή το ‘Life threatening’ με ποσοστό 2.17% επί του συνόλου.

Μετά την εφαρμογή της εξισορρόπησης των κλάσεων με τη βοήθεια του balancer TomekLinks, η τελική κατανομή των κλάσεων για το σύνολο δεδομένων εκπαίδευσης ήταν η εξής:



Εικόνα 42. Κατανομή των κλάσεων στο σύνολο εκπαίδευσης μετά την εφαρμογή εξισορρόπησης των κλάσεων.

Από το συγκεκριμένο διάγραμμα πίτας παρατηρούμε πως η σχετική κατανομή των κλάσεων έχει βελτιωθεί σε ένα μικρό βαθμό, παρόλα αυτά η κλάση με την τιμή 0 εξακολουθεί να είναι αυτή που συγκεντρώνει τις περισσότερες παρατηρήσεις σε σχέση με τις υπόλοιπες.

4.5 ΜΕΤΡΙΚΕΣ ΑΞΙΟΛΟΓΗΣΗΣ ΑΛΓΟΡΙΘΜΩΝ ΤΑΞΙΝΟΜΗΣΗΣ

Σε μια σειρά δοκιμών, αξιολογήσαμε την απόδοση ποικίλων αλγορίθμων ταξινόμησης σε ένα σύνολο δεδομένων που περιλαμβάνει πληροφορίες για τις παρενέργειες των εμβολίων κατά της COVID-19. Τα πειράματα διεξήχθησαν τοπικά και λόγω περιορισμών που θα αναφερθούν αργότερα, χρησιμοποιήθηκαν υπο-σύνολα δεδομένων για την ανάλυση, επομένως τα σύνολα δεδομένων που χρησιμοποιήθηκαν για την εκπαίδευση αποτελούνταν από διαφορετικά μεγέθη, που κυμαίνονται από 200 έως 40000 δείγματα - εγγραφές με συνολικά 19868 χαρακτηριστικά (features). Ενδεικτικά, παρουσιάζονται τα αποτελέσματα αξιολόγησης των αλγορίθμων για τα πειράματα A, B και Γ με $N = 200, 10000$ και 40000 δείγματα αντίστοιχα.

Παρακάτω συνοψίζονται οι μετρικές αξιολόγησης της απόδοσης κάθε αλγορίθμου ταξινόμησης, ανάλογα με το μέγεθος N του δείγματος που χρησιμοποιήθηκε για την εκπαίδευση:

4.5.1 Πείραμα Α

Μέγεθος X_train: 154 x 19867

Μέγεθος X_test: 40 x 19867

Πίνακας 10. Μετρικές αξιολόγησης αλγορίθμων ταξινόμησης για $N = 200$

	F1 score	MCC	Precision	Recall
Decision Tree	0.773	0.296	0.817	0.800
KNN	0.763	0.324	0.781	0.825
Logistic Regression	0.763	0.324	0.781	0.825
Multi-Layer Perceptron	0.763	0.324	0.781	0.825
Random Forest	0.763	0.324	0.781	0.825
Support Vector Machine	0.711	0.000	0.640	0.800
XGBoost	0.750	0.232	0.712	0.800

Αξίζει να παρατηρήσουμε πως οι αλγόριθμοι Logistic Regression, Multi-Layer Perceptron, Random Forest και Support Vector Machine δίνουν τα ίδια ποσοστά απόδοσης για όλες τις μετρικές.

Παρατηρούμε επίσης πως ο αλγόριθμος Decision Tree έχει το μεγαλύτερο F1-score και τη μεγαλύτερη πιστότητα, ενώ οι αλγόριθμοι Logistic Regression, Multi-Layer Perceptron, Random Forest και Support Vector Machine έχουν την μεγαλύτερη ευαισθησία.

4.5.2 Πείραμα Β

Μέγεθος X_train: 7826 x 19867

Μέγεθος X_test: 2000 x 19867

Πίνακας 4. Μετρικές αξιολόγησης αλγορίθμων ταξινόμησης για $N = 10000$

	F1 score	MCC	Precision	Recall
Decision Tree	0.678	0.267	0.669	0.690
KNN	0.674	0.264	0.654	0.734
Logistic Regression	0.686	0.327	0.700	0.756
Multi-Layer Perceptron	0.680	0.267	0.660	0.711
Random Forest	0.689	0.316	0.688	0.755
Support Vector Machine	0.611	0.000	0.527	0.726

XGBoost	0.686	0.325	0.662	0.757
---------	-------	-------	-------	-------

Με βάση τα παραπάνω αποτελέσματα παρατηρούμε πως ο αλγόριθμος Random Forest έχει το μεγαλύτερο F1-score και τη μεγαλύτερη πιστότητα, ενώ ο αλγόριθμος XGBoost έχει τη μεγαλύτερη ευαισθησία.

4.5.3 Πείραμα Γ

Μέγεθος X_train: 31549 x 19867

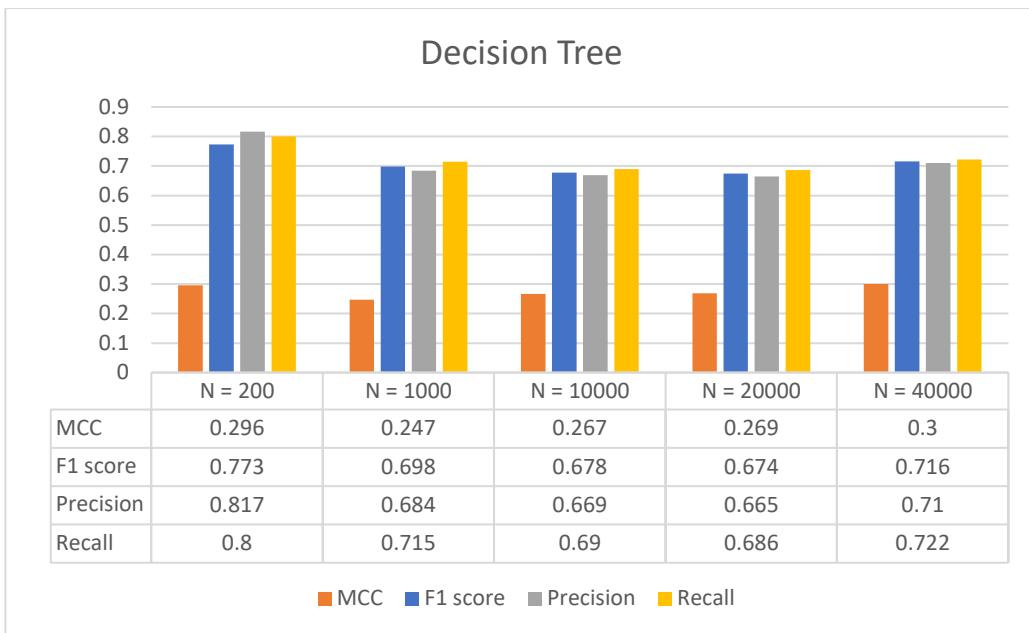
Μέγεθος X_test: 8000 x 19867

Πίνακας 5. Μετρικές αξιολόγησης αλγορίθμων ταξινόμησης για N = 40000

	F1 score	MCC	Precision	Recall
Decision Tree	0.716	0.300	0.710	0.722
KNN	0.721	0.300	0.709	0.770
Logistic Regression	0.706	0.279	0.723	0.771
Multi-Layer Perceptron	0.721	0.300	0.708	0.737
Random Forest	0.737	0.358	0.737	0.788
Support Vector Machine	0.652	0.000	0.573	0.757
XGBoost	0.739	0.379	0.758	0.794

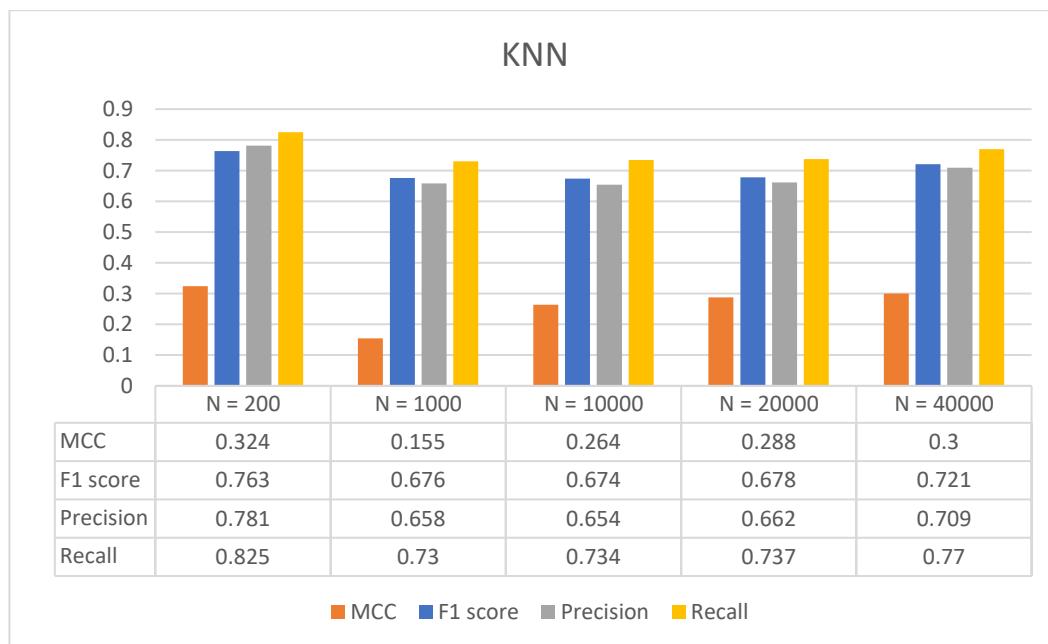
Από αυτό το σύνολο δεδομένων μπορούμε να παρατηρήσουμε πως ο αλγόριθμος XGBoost έχει το μεγαλύτερο F1-score, πιστότητα καθώς και ευαισθησία.

Παρακάτω απεικονίζονται γραφήματα που συνοψίζουν την απόδοση του κάθε αλγορίθμου ξεχωριστά για κάθε διαφορετικό διαδοχικά αυξανόμενο σύνολο εγγραφών που χρησιμοποιήθηκε για την εκπαίδευση:



Εικόνα 43. Ραβδόγραμμα ποσοστών απόδοσης του αλγορίθμου Decision Tree για αυξανόμενο αριθμό εγγραφών

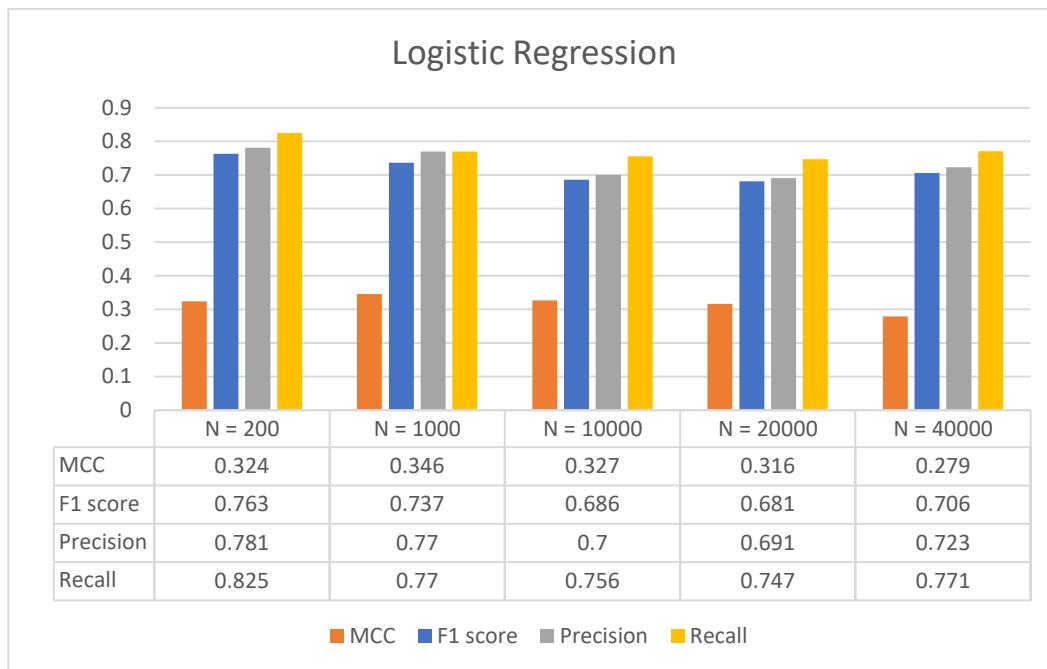
Παρατηρούμε πως η μετρική MCC έχει το χαμηλότερο σκορ έναντι όλων των μετρικών αξιολόγησης. Οι υπόλοιπες μετρικές αξιολόγησης συμπεριφέρονται σχεδόν με τον ίδιο τρόπο όσο αυξάνεται το πλήθος εγγραφών για την εκπαίδευση καθώς δεν εντοπίζονται σημαντικές διαφορές. Το μεγαλύτερο σκορ εντοπίζεται για $N = 200$ όσον αφορά την πιστότητα και την ευαισθησία.



Εικόνα 44. Ραβδόγραμμα ποσοστών απόδοσης του αλγορίθμου KNN για αυξανόμενο αριθμό εγγραφών

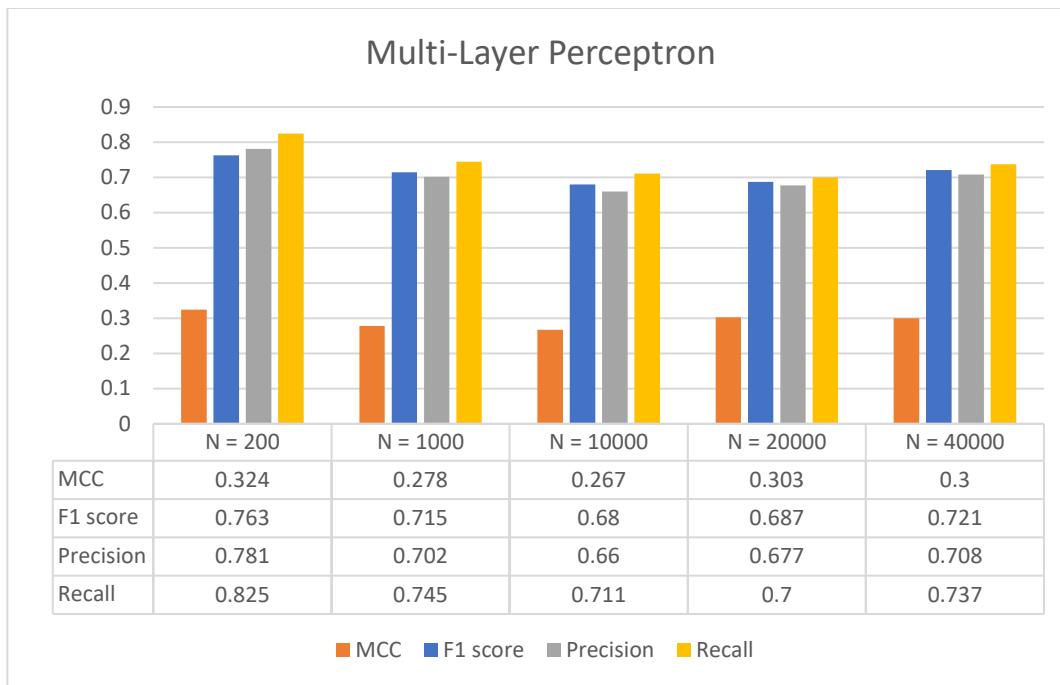
Παρατηρούμε όπως και στον αλγόριθμο Decision Tree πως η μετρική MCC έχει το χαμηλότερο σκορ για όλα τα μεγέθη εκπαίδευσης, ειδικά για $N = 1000$. Οι υπόλοιπες μετρικές

συμπεριφέρονται πάλι με παρόμοιο τρόπο έχοντας καλύτερη ευαισθησία (recall) σε όλες τις δοκιμές μεγεθών.



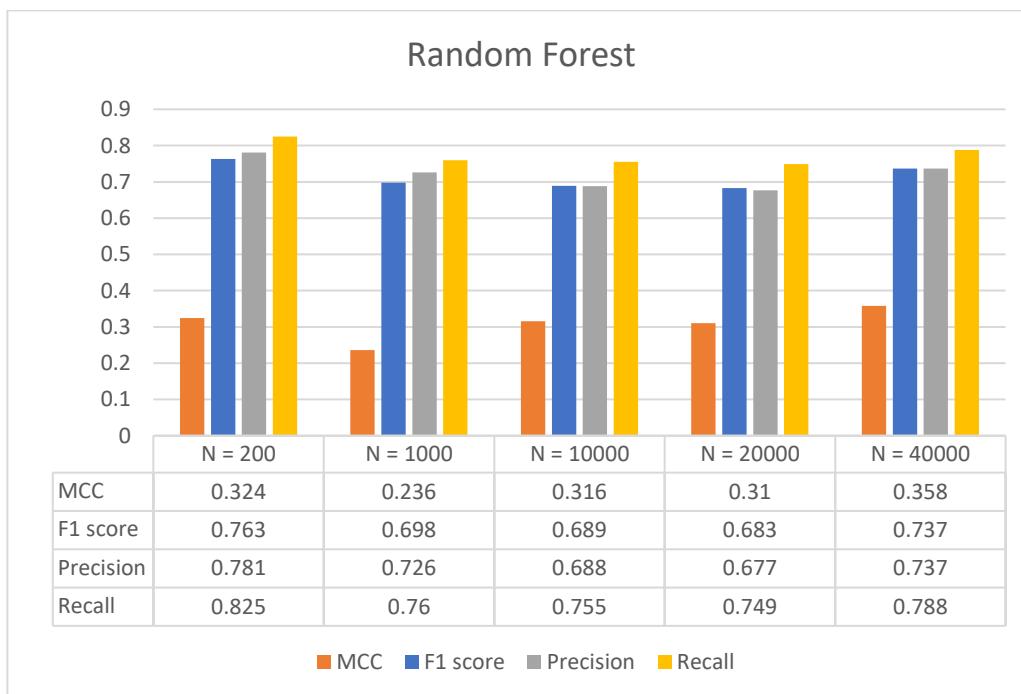
Εικόνα 45. Ραβδόγραμμα ποσοστών απόδοσης του αλγορίθμου Logistic Regression για αυξανόμενο αριθμό εγγραφών

Για τη Λογιστική Παλινδρόμηση, παρατηρούμε ξανά πως η μετρική MCC έχει τα χαμηλότερα ποσοστά, ενώ οι υπόλοιπες συμπεριφέρονται με παρόμοιο τρόπο για διαφορετικό πλήθος εγγραφών που χρησιμοποιούνται για την εκπαίδευση.



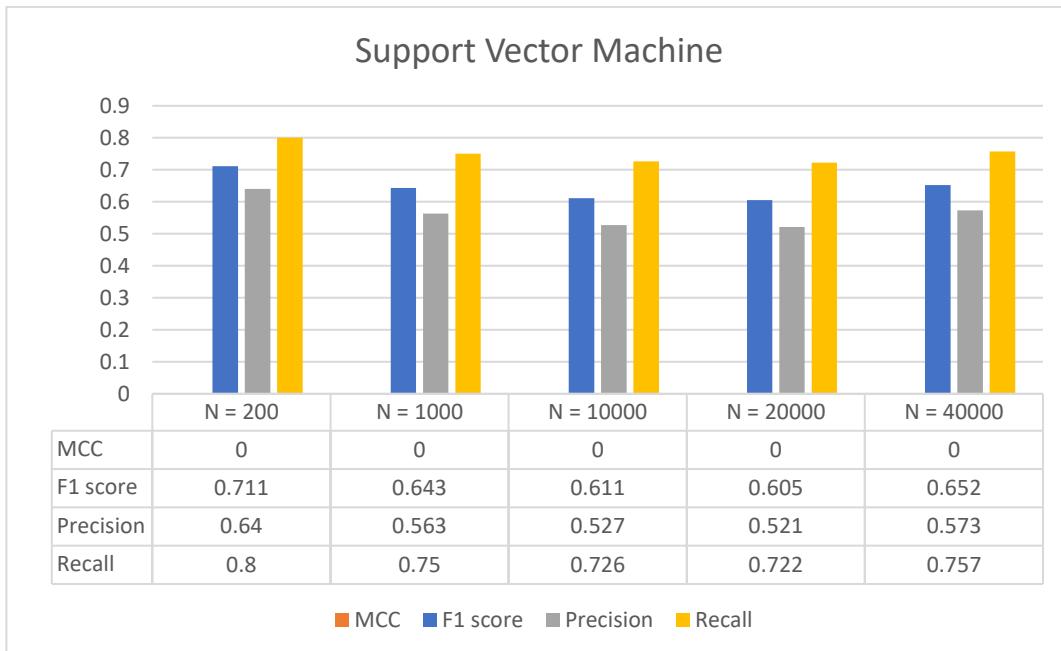
Εικόνα 46. Ραβδόγραμμα ποσοστών απόδοσης του αλγορίθμου *Multi-layer Perceptron* για αυξανόμενο αριθμό εγγραφών

Σχετικά με τον αλγόριθμο Multi-Layer Perceptron, παρατηρούμε πως η μετρική MCC έχει τη χαμηλότερη απόδοση, ενώ οι υπόλοιπες συμπεριφέρονται με τον ίδιο τρόπο για διαφορετικά μεγέθη δειγμάτων με καλύτερα σκορ να εντοπίζονται στην ευαισθησία του αλγορίθμου.



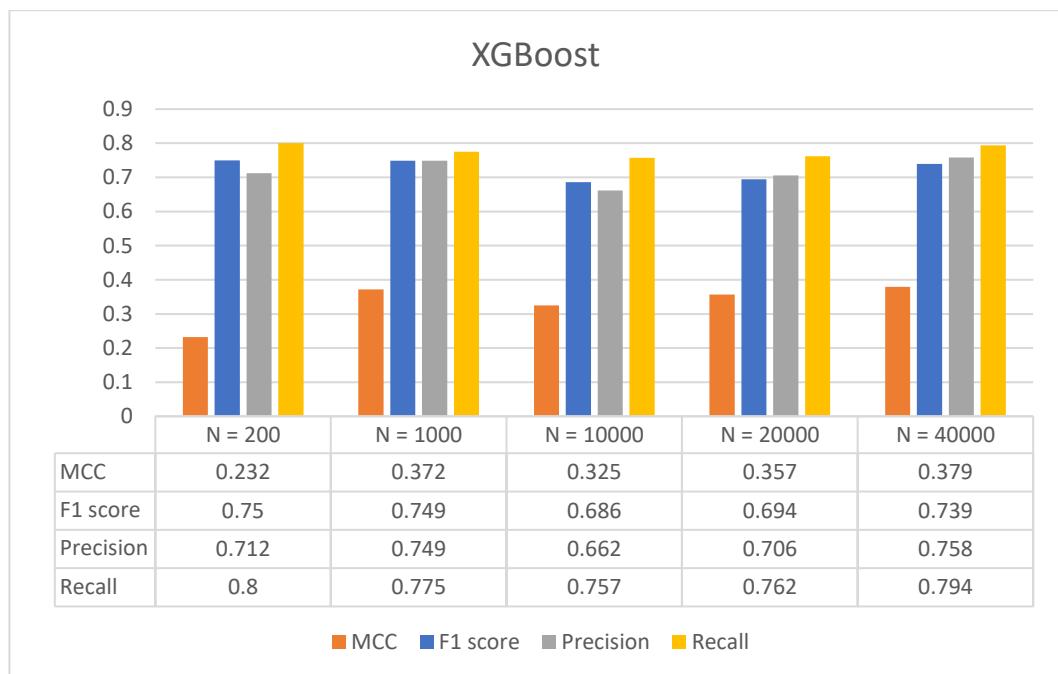
Εικόνα 47. Ραβδόγραμμα ποσοστών απόδοσης του αλγορίθμου *Random Forest* για αυξανόμενο αριθμό εγγραφών

Όσον αφορά τον αλγόριθμο Random Forest, η εικόνα που παρουσιάζει είναι παραπλήσια με την εικόνα του Multi-Layer Perceptron.



Εικόνα 48. Ραβδόγραμμα ποσοστών απόδοσης του αλγορίθμου *Support Vector Machine* για αυξανόμενο αριθμό εγγραφών

Για τον αλγόριθμο Support Vector Machine, παρατηρούμε πως η τιμή του MCC για όλες τις δοκιμές ισούται με το 0. Όσον αφορά τις υπόλοιπες μετρικές συμπεριφέρονται με παρόμοιο τρόπο κατά μήκος του γραφήματος με σαφές προβάδισμα στην ευαισθησία σε σύγκριση με τις άλλες μετρικές.

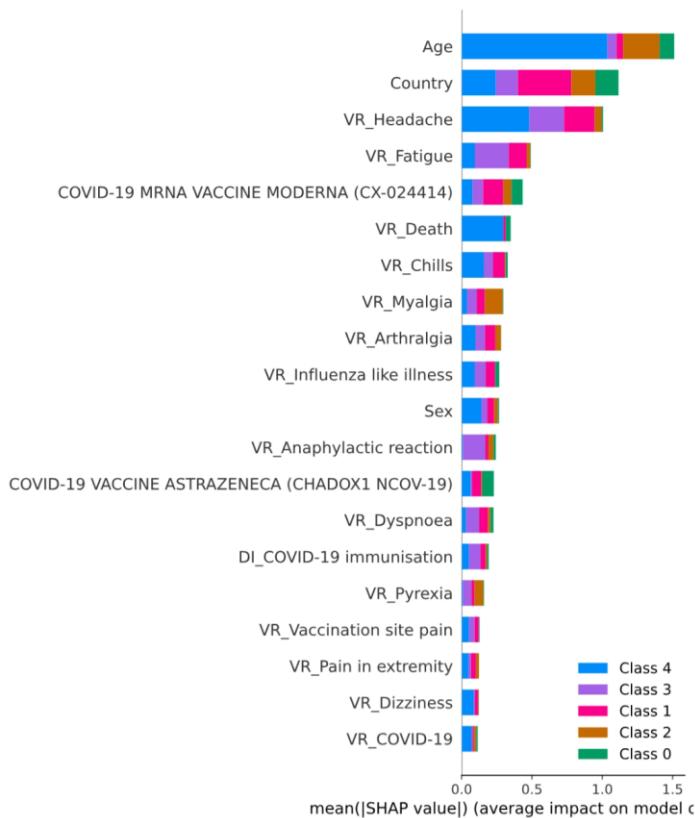


Εικόνα 49. Ραβδόγραμμα ποσοστών απόδοσης του αλγορίθμου *XGBoost* για αυξανόμενο αριθμό εγγραφών

Τέλος, για τον αλγόριθμο XGBoost παρατηρούμε πως η μετρική MCC είναι πάλι στα χαμηλότερα επίπεδα, ενώ οι υπόλοιπες μετρικές συμπεριφέρονται με παρόμοιο τρόπο με την ευαισθησία να έχει τα καλύτερα αποτελέσματα σε όλα τα πειράματα.

4.6 ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΝΑΛΥΣΗΣ SHAP

Όπως συζητήθηκε στις προηγούμενες ενότητες, εφαρμόσαμε ανάλυση SHAP στο μοντέλο που εκπαιδεύτηκε με το μεγαλύτερο υποσύνολο δεδομένων και με τα καλύτερα αποτελέσματα ακρίβειας, δηλαδή στο μοντέλο με τον ταξινομητή XGBoost. Μετά την εφαρμογή της ανάλυσης SHAP και με τη χρήση της προαναφερθείσας βιβλιοθήκης shap, λαμβάνουμε το εξής ραβδόγραμμα:



Εικόνα 50. Ραβδόγραμμα αναπαράστασης επίδρασης των γνωρισμάτων στην πρόβλεψη του μοντέλου

Ας αναλύσουμε περαιτέρω τα συστατικά του συγκεκριμένου γραφήματος:

- Ο άξονας X αναπαριστά τις απόλυτες μέσες τιμές SHAP που έχουν υπολογιστεί με τις μεθόδους που παρέχει το πακέτο shap στην Python. Οι τιμές αυτές ποσοτικοποιούν την επίδραση κάθε χαρακτηριστικού σε μεμονωμένες προβλέψεις και ουσιαστικά υποδεικνύουν κατά πόσο συμβάλλει κάθε χαρακτηριστικό στο να «σπρώξει» το

αποτέλεσμα του μοντέλου από μια βασική (baseline) πρόβλεψη (μέσο output του μοντέλου) στην πραγματική πρόβλεψη για μια δεδομένη περίπτωση. Οι τιμές SHAP μπορεί να κυμαίνονται από αρνητικές έως θετικές. Μια αρνητική τιμή SHAP υποδηλώνει ότι η παρουσία του χαρακτηριστικού μειώνει την πρόβλεψη του μοντέλου, ενώ μια θετική τιμή SHAP υποδηλώνει ότι η παρουσία του χαρακτηριστικού αυξάνει την πρόβλεψη. Τα χαρακτηριστικά με υψηλότερες απόλυτες τιμές SHAP έχουν μεγαλύτερη επιφροή στις προβλέψεις του μοντέλου. Στο παρόν γράφημα, οι τιμές SHAP είναι απόλυτες που σημαίνει πως δεν μπορούμε να βγάλουμε σαφές συμπέρασμα σχετικά με το αν η παρουσία του εκάστοτε χαρακτηριστικού αυξάνει ή μειώνει την πρόβλεψη, παρά μόνο ότι παίζει πιο σημαντικό ρόλο στην πρόβλεψη σε σχέση με άλλα χαρακτηριστικά.

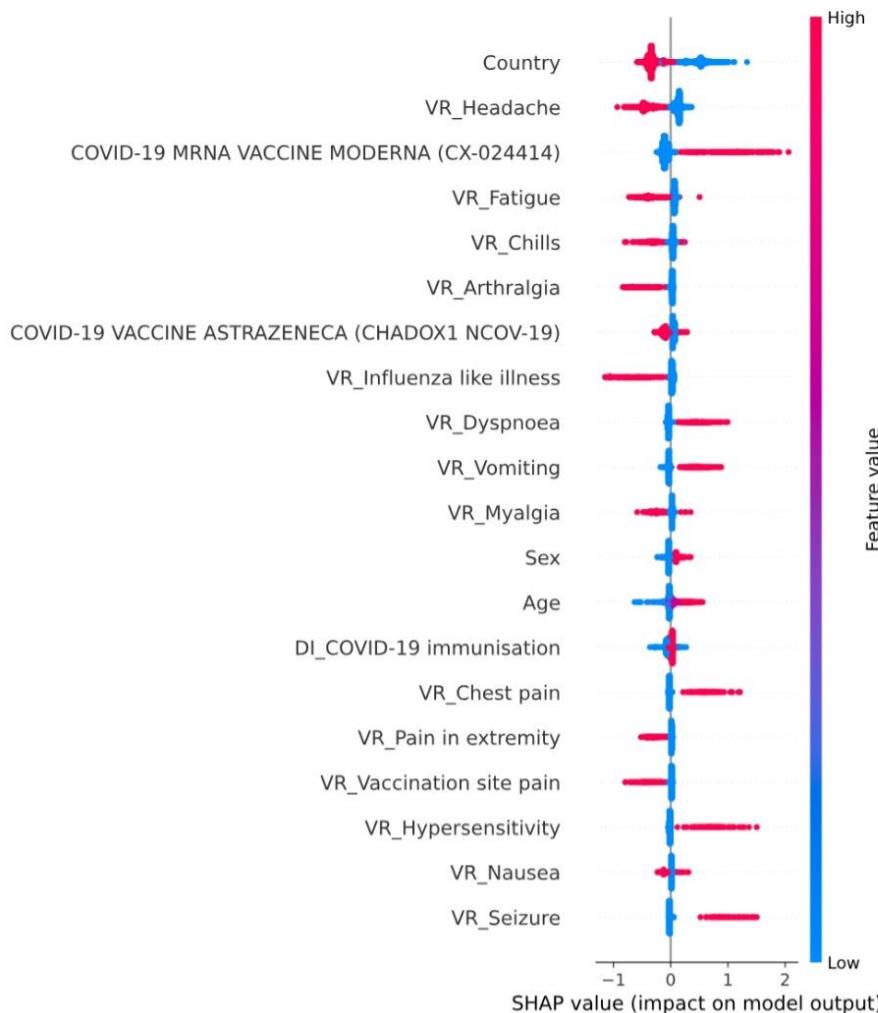
2. Όσον αφορά τον άξονα Y, αναπαριστά τα γνωρίσματα που περιλαμβάνονται στην ανάλυση. Αυτά τα γνωρίσματα αντιπροσωπεύουν διάφορα χαρακτηριστικά που σχετίζονται είτε με το εμβόλιο COVID-19 είτε με τα μεμονωμένα χαρακτηριστικά του ασθενούς.
3. Το μήκος κάθε ράβδου αντιστοιχεί στο μέγεθος της τιμής SHAP που σχετίζεται με το αντίστοιχο χαρακτηριστικό. Οι μεγαλύτερες ράβδοι υποδεικνύουν χαρακτηριστικά με υψηλότερες απόλυτες τιμές SHAP και επομένως μεγαλύτερη σημασία στον επηρεασμό των προβλέψεων του μοντέλου. Μέσα σε κάθε ράβδο (που αντιστοιχεί σε ένα χαρακτηριστικό), υπάρχουν 5 τιμήματα με διαφορετικά χρώματα, καθένα από τα οποία αντιπροσωπεύει τη συμβολή του χαρακτηριστικού σε μια συγκεκριμένη κλάση. Πιο συγκεκριμένα, το μήκος κάθε τιμήματος υποδεικνύει το μέγεθος της επίδρασης του χαρακτηριστικού στην αντίστοιχη κλάση σοβαρότητας. Τα χαρακτηριστικά με μακρύτερα τιμήματα συμβάλλουν πιο σημαντικά στην πρόβλεψη της σοβαρότητας των παρενεργειών σε μια δεδομένη κλάση.

Ένα τέτοιο γράφημα εμφανίζει συνήθως τα κορυφαία N πιο σημαντικά χαρακτηριστικά, όπου το N καθορίζεται από τον χρήστη ή τις προεπιλεγμένες ρυθμίσεις της βιβλιοθήκης SHAP. Σε αυτήν την περίπτωση, το γράφημα δείχνει τα κορυφαία 20 πιο σημαντικά χαρακτηριστικά. Αυτά τα κορυφαία χαρακτηριστικά έχουν την πιο σημαντική επίδραση στην προβλεπόμενη σοβαρότητα των παρενεργειών.

Από το συγκεκριμένο ραβδόγραμμα λοιπόν μπορούμε να παρατηρήσουμε πως σημαντικότερη επίδραση στην πρόβλεψη έχει η ηλικία, και μάλιστα σημαντικότερη επίδραση έχει στην πρόβλεψη της κλάσης 4, δηλαδή του θανάτου (βλ. ενότητα 4.4). Ακολουθεί η χώρα, που όμως στην προκειμένη περίπτωση γνωρίζουμε πως επειδή το συγκεκριμένο γνώρισμα μπορεί να πάρει μόνο τις τιμές 0 και 1 για τις χώρες της ΕΕ και αυτές εκτός ΕΕ, δεν παίζει στην πραγματικότητα κάποιον

ουσιαστικό ρόλο. Ακολουθεί η παρενέργεια του πονοκεφάλου, με μεγαλύτερη επίδραση στην πρόβλεψη της κλάσης 4, η κόπωση, το εμβόλιο της Moderna, ο θάνατος, το ρίγος, η μυαλγία κλπ.

Για περεταίρω κατανόηση των αποτελεσμάτων της ανάλυσης SHAP, μπορούμε με τη χρήση της ίδιας βιβλιοθήκης να παράξουμε και εναλλακτικό γράφημα που αναπαριστά σημεία αντί για ράβδους και ονομάζεται beeswarm plot. Παρακάτω παρατίθεται το εν λόγω γράφημα:



Εικόνα 51. Beeswarm plot που αναπαριστά τα γνωρίσματα με τη μεγαλύτερη σημασία στην πρόβλεψη του μοντέλου

Το συγκεκριμένο γράφημα αναπαριστά τα γνωρίσματα του μοντέλου στον άξονα Y, όπως και στο ραβδόγραμμα που εξηγήθηκε προηγουμένως, ενώ στον άξονα X καταγράφονται οι τιμές SHAP που έχουν υπολογιστεί. Σε αυτήν την περίπτωση οι τιμές SHAP δεν αναφέρονται κατά απόλυτη τιμή, επομένως μπορούμε να αντλήσουμε διαφορετικό είδος πληροφορίας από το προηγούμενο γράφημα και συνδυαστικά να καταλήξουμε σε πιο ολοκληρωμένα συμπεράσματα. Συγκεκριμένα, στο beeswarm plot κάθε σημείο αντιπροσωπεύει μια εγγραφή. Ο άξονας X αντιστοιχεί στην τιμή SHAP, και στα σημεία που εντοπίζεται κατακόρυφη διασπορά ή «θόρυβος» αυτό υποδηλώνει

υψηλή πυκνότητα σημείων. Η χρωματική διαβάθμιση υποδεικνύει το σχετικό μέγεθος κάθε χαρακτηριστικού (όχι τις τιμές SHAP) με το κόκκινο να υποδεικνύει υψηλές τιμές του χαρακτηριστικού (π.χ. μεγαλύτερη ηλικία) και το μπλε (π.χ. μικρότερη ηλικία) το αντίθετο. Στην προκειμένη περίπτωση που τα περισσότερα γνωρίσματα παίρνουν τιμές 0 και 1, το κόκκινο χρώμα ουσιαστικά αναπαριστά την τιμή 1 ενώ το μπλε την τιμή 0.

Στο συγκεκριμένο γράφημα παρατηρούμε ότι τα περισσότερα μπλε σημεία (δηλαδή κυρίως οι τιμές 0, άρα η μη-ύπαρξη του χαρακτηριστικού) συγκεντρώνονται κοντά στην τιμή 0 του άξονα X που αντιπροσωπεύει τις τιμές SHAP. Όσον αφορά τα κόκκινα σημεία (δηλαδή κυρίως τις τιμές 1, άρα την ύπαρξη του χαρακτηριστικού), αυτά εμφανίζουν μεγάλη διασπορά στον άξονα X, λαμβάνοντας και θετικές και αρνητικές τιμές SHAP για διαφορετικά χαρακτηριστικά. Περισσότερες υψηλές θετικές τιμές SHAP, δηλαδή που να υποδεικνύουν μεγάλη θετική επίδραση στην πρόβλεψη του μοντέλου, μπορούμε να παρατηρήσουμε για το εμβόλιο της Moderna, τον πόνο στο στήθος, την υπερευαισθησία, την επιληπτική διαταραχή, τη δύσπνοια και την έμεση, ενώ περισσότερες υψηλές αρνητικές τιμές SHAP, δηλαδή που συμβάλλουν αρνητικά στην πρόβλεψη του μοντέλου, εντοπίζουμε στον πονοκέφαλο, την κόπωση, το ρίγος, την αρθραλγία, τη γριπώδη συνδρομή, τη μυαλγία και τον τοπικό πόνο στην περιοχή εμβολιασμού.

5 ΣΥΖΗΤΗΣΗ

Σε αυτή τη μελέτη, επιχειρήσαμε να αναπτύξουμε και να εκπαιδεύσουμε έναν αλγόριθμο ταξινόμησης μηχανικής μάθησης που θα προβλέπει τη σοβαρότητα των παρενεργειών των εμβολίων με βάση πρωτογενή δεδομένα που αντλήθηκαν από τη βάση δεδομένων EudraVigilance.

5.1 ΜΟΡΦΗ ΤΟΥ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ

Η μορφή του συνόλου δεδομένων που χρησιμοποιήθηκαν σε αυτή τη διπλωματική χρήζει συζήτησης, ιδιαίτερα όσον αφορά την αναπαράσταση των ηλικιακών κλάσεων και την κατανομή του φύλου. Η ηλικία ήταν κατηγοριοποιημένη σε αρκετά ευρεία άρια, όπως για παράδειγμα 18-64 ετών, τα οποία μπορεί να μην συλλαμβάνουν επαρκώς τις πραγματικές ηλικίες των ασθενών και θα μπορούσαν ενδεχομένως να προκαλέσουν προβλήματα στην πρόβλεψη επιπτώσεων που σχετίζονται με την ηλικία στις παρενέργειες του εμβολίου. Παράλληλα, ένα σημαντικό ποσοστό παρατηρήσεων συγκεντρώθηκαν στην ηλικιακή κατηγορία 18-64 ετών, αποτελώντας περίπου το 60% του συνόλου δεδομένων, περιορίζοντας ενδεχομένως τη γενίκευση των ευρημάτων σε μεγαλύτερα ηλικιακά εύρη.

Επιπλέον, η ανισορροπία στην κατανομή του φύλου, με αναλογία γυναικών προς άντρες περίπου 3:1, μπορεί να εισάγει μεροληψία στην ανάλυση και την ερμηνεία των αποτελεσμάτων, καθώς ορισμένες παρενέργειες μπορεί να εκδηλωθούν διαφορετικά μεταξύ των δύο φύλων.

Επιτρόσθετα, είναι σημαντικό να αναγνωριστεί η εγγενής υποκειμενικότητα που σχετίζεται με την αναφορά παρενεργειών του εμβολίου, ιδίως όσον αφορά την αξιολόγηση της σοβαρότητας. Η αντίληψη της σοβαρότητας μπορεί να ποικίλει μεταξύ των ατόμων και μπορεί να επηρεάζεται από παράγοντες όπως η ανοχή στον πόνο, οι προηγούμενες εμπειρίες και οι πολιτισμικές διαφορές. Κατά συνέπεια, τα δεδομένα σοβαρότητας που αναλύθηκαν σε αυτή τη μελέτη ενδέχεται να μην αντικατοπτρίζουν πλήρως την αντικειμενική κλινική σοβαρότητα των παρενεργειών αλλά μάλλον την υποκειμενική ερμηνεία των ατόμων που τα αναφέρουν.

Τέλος, θα πρέπει να αναφερθεί πως η δομή του αρχείου που χρησιμοποιήσαμε ως είσοδο για την εκπαίδευση των μοντέλων βασιζόταν στη δυαδική λογική 0 - 1 για τη μη ύπαρξη και ύπαρξη αντίστοιχα σχέσης μεταξύ μιας εγγραφής και ενός γνωρίσματος από τα διαφορετικά Drugs, Indications και Reactions, κυρίως λόγω χαμηλότερης υπολογιστικής πολυπλοκότητας. Αυτό βέβαια δε σημαίνει απαραίτητα πως αυτή είναι και η βέλτιστη δομή των δεδομένων που

χρειάζεται ένας αλγόριθμος ταξινόμησης, καθώς θα μπορούσε να γίνει και με διαφορετικούς τρόπους με διαφορετική επεξεργασία.

5.2 ΤΕΧΝΙΚΕΣ ΠΡΟ-ΕΠΕΞΕΡΓΑΣΙΑΣ ΔΕΔΟΜΕΝΩΝ

Η χρήση του balancer TomekLinks, η κωδικοποίηση με τον label encoder για τις κατηγορικές μεταβλητές και η επιλογή χαρακτηριστικών μέσω του feature selector SelectPercentile πιθανότατα συνέβαλαν στη βελτιωμένη απόδοση των αλγορίθμων. Ο TomekLinks βοήθησε στην αντιμετώπιση ανισορροπιών της κλάσης, βελτιώνοντας την ικανότητα των αλγορίθμων να μαθαίνουν από τις τάξεις μειοψηφίας. Το encoding απλοποίησε την αναπαράσταση κατηγορικών μεταβλητών, ενώ ο SelectPercentile βοήθησε στην επιλογή των χαρακτηριστικών που μπορούν να προσφέρουν πιο χρήσιμη πληροφορία, μειώνοντας τις διαστάσεις και δυνητικά μετριάζοντας την υπερπροσαρμογή.

Πιο αναλυτικά, η εφαρμογή του feature selection σε έναν αλγόριθμο μηχανικής μάθησης μπορεί να επηρεάσει σημαντικά το αποτέλεσμα του μοντέλου, ιδιαίτερα όταν υπάρχει σημαντική ποσοστιαία μείωση του αριθμού των χαρακτηριστικών. Στο μοντέλο μηχανικής μάθησης που αναπτύχθηκε στην παρούσα διπλωματική, 90% των χαρακτηριστικών καταργήθηκε, μεταβαίνοντας σχεδόν από 20000 σε περίπου 2000 χαρακτηριστικά, κάτι το οποίο μπορεί να αποφέρει αρκετά πλεονεκτήματα. Κατά κύριο λόγο, μια τέτοια μείωση ενισχύει την ερμηνευτικότητα του μοντέλου και την κατανόηση των εγγενών σχέσεων μεταξύ των προγνωστικών παραγόντων και των αποτελεσμάτων(Guyon & Elisseeff, 2003). Επιπλέον, η εξάλειψη μη-σχετικών ή περιττών χαρακτηριστικών μπορεί να μετριάσει αποτελεσματικά τον κίνδυνο υπερ-προσαρμογής, όπου τα μοντέλα καταγράφουν λανθασμένα «θόρυβο» μέσα στα δεδομένα και όχι πραγματικά μοτίβα(Hastie, Tibshirani, Friedman, et al., 2009). Κατά συνέπεια, τα μοντέλα που προκύπτουν χαρακτηρίζονται από μεγαλύτερη γενίκευση και σταθερότητα, καλύτερα εκπαιδευμένα για βέλτιστη απόδοση σε νέα δεδομένα. Τέλος, η υπολογιστική αποτελεσματικότητα ενός αλγορίθμου ενισχύεται σημαντικά, καθώς ο μειωμένος «χώρος» των χαρακτηριστικών μειώνει τον υπολογιστικό φόρτο κατά τη διάρκεια των φάσεων εκπαίδευσης και συμπερασμάτων (Kira & Rendell, 1992). Αυτή η ταχεία επεξεργασία φέρει ιδιαίτερα πλεονεκτήματα για σύνολα δεδομένων μεγάλης κλίμακας, όπως της παρούσας διπλωματικής, διασφαλίζοντας την έγκαιρη ανάπτυξη του μοντέλου και τη λειτουργική του αποτελεσματικότητα. Έτσι, παρά τη σημαντική μείωση στη διάσταση των χαρακτηριστικών, η συνετή εφαρμογή των τεχνικών feature selection προωθεί τη δημιουργία ερμηνεύσιμων και αποτελεσματικών μοντέλων,

ενισχύοντας έτσι την αποτελεσματικότητα και την πρακτικότητα των εφαρμογών μηχανικής μάθησης.

Η εφαρμογή τεχνικών *balancing*, όπως ο TomekLinks, σε έναν αλγόριθμο μηχανικής μάθησης μπορεί να αποφέρει αξιοσημείωτες βελτιώσεις στην απόδοση του μοντέλου, ιδιαίτερα σε σύνολα δεδομένων που παρουσιάζουν σημαντικές ανισορροπίες στην κλάση. Στο σενάριο που περιγράφηκε, όπου η κλάση 0 περιλαμβάνει αρχικά περίπου το 80% των δειγμάτων, το *balancing* με τη χρήση του TomekLinks μειώνει την κυριαρχία αυτής της κατηγορίας σε περίπου 75%, αντιτροσωπεύοντας μια ναι μεν μέτρια αλλά ουσιαστική προσαρμογή. Παρά την παραμονή της συνολικής ανισορροπίας των κλάσεων, ο αλγόριθμος μετριάζει αποτελεσματικά την επιρροή της πλειοψηφικής τάξης διατηρώντας παράλληλα τις περιπτώσεις μειοψηφίας που είναι κρίσιμες για τη εκμάθηση μοντέλων (Tomek, 1976). Αυτή η διαδικασία ενισχύει την ικανότητα ενός ταξινομητή να διακρίνει τα όρια μεταξύ των κλάσεων, βελτιώνοντας έτσι την ακρίβεια ταξινόμησης και μειώνοντας την προκατάληψη προς την πλειοψηφική τάξη (Batista et al., 2004). Επιπλέον, η εξισορρόπηση με TomekLinks βοηθά στην πρόληψη της υπερ-προσαρμογής, οδηγώντας έτσι σε μια πιο ισορροπημένη αναπαράσταση των κλάσεων, προωθώντας έτσι καλύτερη απόδοση στη γενίκευση σε νέα δεδομένα (Chawla et al., 2002). Ωστόσο, είναι σημαντικό να αναγνωρίζονται οι εγγενείς περιορισμοί του συνόλου δεδομένων, όπως η πιθανή απώλεια πληροφοριών κατά τη διαδικασία εξισορρόπησης και η ανάγκη προσεκτικής εξέτασης των σχετικών επιπτώσεων.

Τέλος, αξίζει να σημειωθεί πως ο *label encoder* χρησιμοποιείται συνήθως για δεδομένα που έχουν μια τακτική (ordinal) σχέση μεταξύ τους. Επειδή ο *encoder* αυτός αντιστοιχεί μοναδικούς ακέραιους αριθμούς (ξεκινώντας από το 0) σε κάθε τάξη, αυτό μπορεί να οδηγήσει σε προβλήματα προτεραιότητας κατά την εκπαίδευση μοντέλων, δηλαδή οι υψηλότερα αριθμητικές τιμές ενδέχεται να συνεπάγονται κατά λάθος υψηλότερη προτεραιότητα, ακόμη και αν δεν υπάρχει εγγενής σειρά μεταξύ των κατηγοριών (Brownlee, 2016). Για κατηγορικές μεταβλητές όπου δεν υπάρχει τακτική σχέση, η κωδικοποίηση με ακεραίους μπορεί να μην είναι αρκετή, στην καλύτερη περίπτωση, ή παραπλανητική για το μοντέλο στη χειρότερη. Η επιβολή μιας τακτικής σχέσης μέσω μιας τακτικής κωδικοποίησης και η δυνατότητα στο μοντέλο να υποθέσει μια φυσική σειρά μεταξύ των κατηγοριών μπορεί να οδηγήσει σε κακή απόδοση ή απροσδόκητα αποτελέσματα (Brownlee, 2020). Παρ' όλα αυτά, καθώς τα χαρακτηριστικά που υπέστησαν κωδικοποίηση, είχαν μόνο δύο τιμές, η διαφορά στην απόδοση σε σχέση με έναν άλλον *encoder* (όπως τον one-hot *encoder*) θα ήταν αμελητέα.

5.3 ΑΠΟΔΟΣΗ ΑΛΓΟΡΙΘΜΩΝ ΤΑΞΙΝΟΜΗΣΗΣ

Το αρχικό σύνολο δεδομένων ήταν αρκετά ογκώδες, περιλαμβάνοντας περίπου 470000 παρατηρήσεις και 20000 χαρακτηριστικά. Λόγω υπολογιστικών περιορισμών όπως αναφέρθηκε, πραγματοποιήσαμε πειράματα σε υποσύνολα δεδομένων που περιείχαν διαδοχικά 200, 1000, 10000, 20000 και 40000 παρατηρήσεις. Στη συνέχεια σχολιάζονται τα αποτελέσματα της απόδοσης των αλγορίθμων σε σχέση με το μέγεθος του συνόλου εκπαίδευσης:

5.3.1 Σύγκριση για διαφορετικά μεγέθη συνόλου δεδομένων

Πείραμα Α

- Οι αλγόριθμοι Decision Tree, KNN, Logistic Regression, Multi-Layer Perceptron, Random Forest και XGBoost έδειξαν συγκρίσιμες επιδόσεις, με F1 score που κυμαίνονται από 0.750 έως 0.773. Ωστόσο, οι βαθμολογίες του MCC διέφεραν, υποδεικνύοντας διαφορές στην ικανότητά τους να χειρίζονται τις ταξικές ανισορροπίες και τη συνολική προγνωστική ισχύ.
- Ο αλγόριθμος Support Vector Machine παρουσίασε τη χαμηλότερη απόδοση σε σύγκριση με άλλους αλγόριθμους, ιδιαίτερα όσον αφορά το MCC, το οποίο ήταν 0. Αυτό υποδηλώνει ότι ο αλγόριθμος SVM αντιμετώπισε δυσκολίες στο να καταγράψει την πολυπλοκότητα του συνόλου δεδομένων και πιθανώς να απαιτούσε περαιτέρω συντονισμό ή καλύτερη επεξεργασία των χαρακτηριστικών για τη βελτίωση της αποτελεσματικότητάς του.

Πείραμα Β

- Οι αλγόριθμοι Logistic Regression, Random Forest και XGBoost ξεπέρασαν τους άλλους, επιτυγχάνοντας F1 score που κυμαίνονται από 0.686 έως 0.689. Συγκεκριμένα, ο αλγόριθμος Logistic Regression έδειξε βελτιωμένη απόδοση σε σύγκριση με το μικρότερο μέγεθος δείγματος, υποδεικνύοντας την ικανότητά του να αξιοποιεί μεγαλύτερα σύνολα δεδομένων για καλύτερη ταξινόμηση.
- Η απόδοση του αλγορίθμου SVM παρέμεινε σχετικά χαμηλή, με βαθμολογία F1 score 0.611, υπογραμμίζοντας περεταίρω τη δυσκολία του να ταξινομήσει αποτελεσματικά τις κατηγορίες σοβαρότητας σε αυτό το μοντέλο.

Πείραμα Γ

- Οι αλγόριθμοι Random Forest και XGBoost είχαν σταθερά την καλύτερη απόδοση σε όλες τις μετρήσεις, με σκορ F1 = 0.737 και 0.739, αντίστοιχα. Η σταθερότητα τους στον χειρισμό τόσο μικρών όσο και μεγάλων συνόλων δεδομένων, μαζί με την ικανότητά τους να

ταξινομούν αποτελεσματικά τις κατηγορίες σοβαρότητας, τους καθιστά ίσως τις καταλληλότερες επιλογές για αυτό το μοντέλο ταξινόμησης (Breiman, 2001; T. Chen & Guestrin, 2016).

- Ενώ ο αλγόριθμος Logistic Regression έδειξε επίσης ανταγωνιστική απόδοση, η βαθμολογία της F1 υστερούσε ελαφρώς σε σχέση με το Random Forest και το XGBoost. Ωστόσο, η απλότητα και η ικανότητα εύκολης ερμηνείας του μπορεί να το καθιστούν μια αρκετά καλή επιλογή.

Μια μελέτη που δημοσιεύτηκε στο περιοδικό Vaccines το 2021, εξέτασε την εφαρμογή μεθόδων μηχανικής μάθησης για την πρόβλεψη της σοβαρότητας των παρενεργειών των εμβολίων κατά της COVID-19 σε πληθυσμό της Ιορδανίας (Hatmal et al., 2021). Συγκεκριμένα, για τη συλλογή δεδομένων χρησιμοποιήθηκαν ερωτηματολόγια που απευθύνονταν στους κατοίκους της Ιορδανίας που έλαβαν οποιοδήποτε εμβόλιο COVID-19. Τα δεδομένα αναλύθηκαν στατιστικά και ορισμένα εργαλεία μηχανικής μάθησης, συμπεριλαμβανομένων των Multi-Layer Perceptron (MLP), eXtreme gradient boosting (XGBoost), Random Forest (RF) και K-star χρησιμοποιήθηκαν για την πρόβλεψη της σοβαρότητας των παρενεργειών. Σύμφωνα με τη μελέτη, συμμετείχαν 2,213 άτομα που είχαν λάβει δόσεις από τα εμβόλια Sinopharm, AstraZeneca, Pfizer-BioNTech κ.ά (38.2%, 31%, 27.3%, και 3.5% αντίστοιχα). Με βάση τον τύπο του εμβολίου, τα δημογραφικά δεδομένα και τις παρενέργειες, οι αλγόριθμοι RF, XGBoost και MLP έδωσαν και οι δύο υψηλά ποσοστά ακρίβειας (0.80, 0.79 και 0.70, αντίστοιχα). Ενώ το σύνολο δεδομένων που χρησιμοποιήθηκε για την εκπαίδευση των αλγορίθμων είχε διαφορετική δομή από αυτό που χρησιμοποιήθηκε σε αυτή τη διπλωματική, μπορούμε να εντοπίσουμε κάποιες ομοιότητες, αλλά και κάποιες διαφορές.

Όσον αφορά τις μετρικές απόδοσης, και η μελέτη των Hatmal et al., αλλά και τα δεδομένα αυτής της διπλωματικής καταλήγουν πως ο αλγόριθμος Random Forest και ο XGBoost πέτυχαν γενικά υψηλότερα ποσοστά ακρίβειας σε σύγκριση με το MLP. Ωστόσο, οι συγκεκριμένες μετρήσεις απόδοσης διαφέρουν μεταξύ των δύο μελετών, πιθανότατα λόγω διαφοροποιήσεων στα χαρακτηριστικά των δεδομένων, στις τεχνικές προεπεξεργασίας καθώς και στις υπερπαραμέτρους του μοντέλου. Η δημοσιευμένη μελέτη ανέφερε υψηλότερες ακρίβειες για το Random Forest (0.80) και το XGBoost (0.79) σε σύγκριση με τα δεδομένα της δικής μας έρευνας, όπου οι βαθμολογίες F1 κυμαίνονταν από 0.678 έως 0.737 για το, Random Forest και από 0.686 έως 0.750 για το XGBoost σε διαφορετικά μεγέθη δειγμάτων. Επιπλέον, ενώ η δημοσιευμένη μελέτη ανέφερε ακρίβεια 0.70 για το MLP, ενώ εμείς υπολογίσαμε F1 score που κυμαίνονται από 0.680 έως 0.763 σε διαφορετικά μεγέθη δειγμάτων. Αυτό υποδηλώνει ότι η απόδοση MLP στη μελέτη μας μπορεί να ήταν ελαφρώς πιο συνεπής σε διαφορετικά μεγέθη δειγμάτων σε σύγκριση με τη δημοσιευμένη μελέτη. Αυτές οι διαφορές θα μπορούσαν να αποδοθούν σε παράγοντες όπως το

μέγεθος δεδομένων, τη σύνθεση του δείγματος και την επιλογή χαρακτηριστικών. Πιο αναλυτικά, παρόλο που τα δεδομένα μας αποτελούνταν από περισσότερες παρατηρήσεις, επειδή ανακτήθηκαν από μια διαδικτυακή βάση δεδομένων με αρκετές ελλείψεις σε διάφορα πεδία, ενδεχομένως να μην είχαν την ίδια ισχύ με δεδομένα που θα μπορούσαν να προκύψουν από ένα ελεγχόμενο ερωτηματολόγιο. Συνοπτικά, ενώ και οι δύο μελέτες χρησιμοποίησαν παρόμοιους αλγόριθμους ταξινόμησης, υπάρχουν παραλλαγές στις αναφερόμενες μετρήσεις απόδοσης.

5.3.2 Συνολική επίδραση του μεγέθους του δείγματος

Σε όλους τους αλγόριθμους ταξινόμησης, παρατηρήσαμε μια σταθερή τάση βελτιωμένης απόδοσης με μεγαλύτερα μεγέθη δειγμάτων. Αυτό έρχεται σε συμφωνία με τη γενική αντίληψη ότι τα πιο εκτεταμένα σύνολα δεδομένων παρέχουν πλουσιότερες πληροφορίες για την εκμάθηση μοντέλων και οδηγούν σε καλύτερη γενίκευση. Συγκεκριμένα, καθώς το μέγεθος του δείγματος αυξήθηκε από 200 σε 40000 παρατηρήσεις, παρατηρήσαμε βελτιώσεις σε διάφορες μετρικές απόδοσης, συμπεριλαμβανομένου του F1 score, του MCC, της πιστότητας και της ευαισθησίας (Hastie, Tibshirani, & Friedman, 2009a; Kohavi, 2001).

Αξίζει να παρατηρήσουμε πως με βάση αυτά τα αποτελέσματα, είναι δύσκολο να διεξαχθεί ένα οριστικό συμπέρασμα σχετικά με το πώς θα κλιμακωνόταν η απόδοση των αλγορίθμων αν συμπεριλαμβάνονταν στην εκπαίδευση περισσότερες παρατηρήσεις, ειδικά δεδομένης της σημαντικής διαφοράς μεταξύ του αρχικού μεγέθους δεδομένων (473333 παρατηρήσεις) και των μεγεθών των υποσυνόλων που χρησιμοποιήθηκαν για την εκπαίδευση (που κυμαίνονται από 200 έως 40000 παρατηρήσεις). Ωστόσο, μπορούμε να επισημάνουμε τα εξής: Ενώ γενικότερα όπως αναφέρθηκε, η αύξηση του αριθμού των παρατηρήσεων στο σύνολο δεδομένων τείνει να βελτιώνει την απόδοση των αλγορίθμων μηχανικής μάθησης, η έκταση της βελτίωσης μπορεί να ποικίλλει ανάλογα με την πολυπλοκότητα των δεδομένων και τους αλγόριθμους που χρησιμοποιούνται. Συγκεκριμένα, ενώ η απόδοση μπορεί να βελτιωθεί με περισσότερα δεδομένα, ο ρυθμός βελτίωσης μπορεί να μειωθεί καθώς το σύνολο δεδομένων γίνεται μεγαλύτερο. Αυτό συμβαίνει επειδή οι αλγόριθμοι ενδέχεται να αρχίσουν να καταγράφουν συγκεκριμένα μοτίβα στα δεδομένα πιο αποτελεσματικά με ένα μικρότερο υποσύνολο και οι επιπλέον παρατηρήσεις μπορεί να μην συμβάλλουν σημαντικά στη βελτίωση της απόδοσης. Επίσης, διαφορετικοί αλγόριθμοι έχουν διαφορετικούς βαθμούς ευαισθησίας στο μέγεθος δεδομένων. Ορισμένοι αλγόριθμοι, μπορεί να ωφεληθούν σημαντικά από μεγαλύτερα σύνολα δεδομένων λόγω της ικανότητάς τους να μαθαίνουν πολύπλοκα μοτίβα. Από την άλλη πλευρά, απλούστεροι αλγόριθμοι όπως τα Δέντρα Απόφασης ή οι κ-πλησιέστεροι γείτονες μπορεί να φτάσουν σε ένα πλατό απόδοσης με μικρότερα σύνολα δεδομένων. Η πολυπλοκότητα των δεδομένων,

συμπεριλαμβανομένου του αριθμού των χαρακτηριστικών και των σχέσεων τους, επηρεάζει επίσης την απόδοση των αλγορίθμων με διαφορετικά μεγέθη δεδομένων. Σε εξαιρετικά πολύπλοκα σύνολα δεδομένων με πολλά χαρακτηριστικά, μπορεί να απαιτούνται μεγαλύτερα μεγέθη δειγμάτων για την ακριβή καταγραφή των υποκείμενων μοτίβων. Λαμβάνοντας υπόψη αυτές τις εκτιμήσεις λοιπόν, είναι δύσκολο να προσδιοριστεί οριστικά πώς θα άλλαζε η απόδοση των αλγορίθμων με περισσότερες παρατηρήσεις χωρίς πρόσθετους πειραματισμούς.

5.4 ΑΝΑΛΥΣΗ SHAP

Όπως αναφέρθηκε νωρίτερα, πραγματοποιήσαμε ανάλυση SHAP κατά την εκπαίδευση του μοντέλου ώστε να διαπιστώσουμε ποια συγκεκριμένα γνωρίσματα έχουν πιθανώς κάποιο σημαντικότερο ρόλο στην πρόβλεψη του μοντέλου και με ποιο τρόπο.

Στο πρώτο γράφημα, το ραβδόγραμμα, παρατηρούμε πως σημαντικότερο ρόλο στην πρόβλεψη του μοντέλου, ανεξαρτήτως κατεύθυνσης, παίζει κατά κύριο λόγο η ηλικία. Αν συνδυάσουμε την πληροφορία που μας δίνει το ραβδόγραμμα με αυτή του beeswarm plot όπου το γνώρισμα της ηλικίας εμφανίζει μια σχετικά ισορροπημένη κατανομή με τις μικρότερες τιμές του γνωρίσματος (μικρότερες ηλικίες) να λαμβάνουν αρνητικές τιμές SHAP, ενώ οι μεγαλύτερες τιμές του γνωρίσματος (μεγαλύτερες ηλικίες) να λαμβάνουν θετικές τιμές SHAP, μπορούμε λογικά να συμπεράνουμε πως όσο αυξάνεται η ηλικία, το μοντέλο «σπρώχνει» την πρόβλεψη προς τη μεγαλύτερη τιμή κλάσης, δηλαδή το μέγιστο επίπεδο σοβαρότητας 4, όπως το έχουμε ορίσει, το θάνατο. Σύμφωνα με έρευνες σχετικά με την αποτελεσματικότητα των εμβολίων, έχει διαπιστωθεί πως τα ηλικιωμένα άτομα μπορεί να εμφανίσουν μειωμένες ανοσολογικές αποκρίσεις (Lee & Linterman, 2022), αλλά παρόλα αυτά σοβαρότερες παρενέργειες. Συγκεκριμένα, σχετικά με τα εμβόλια κατά της COVID-19, οι ηλικιωμένοι άνω των 60 έχουν βρεθεί ότι παρουσιάζουν πιο συχνές και σοβαρές παρενέργειες μετά τον εμβολιασμό σε σύγκριση με τα νεότερα άτομα (Renia et al., 2022). Για παράδειγμα, μια μελέτη των Zhang et al. αναφέρει υψηλότερη συχνότητα ανεπιθύμητων ενεργειών, συμπεριλαμβανομένου πυρετού και κόπωσης, μεταξύ των ηλικιωμένων, υποδεικνύοντας διακυμάνσεις στην ανεκτικότητα του εμβολίου εξαρτώμενες από την ηλικία (Zhang et al., 2021). Επίσης, σε μελέτη ανοσοαπόκρισης μετά τον εμβολιασμό με το εμβόλιο της Pfizer-BioNTech, βρέθηκε πως η παραγωγή ιντερφερόνης-γ και ιντερλευκίνης-2 των ειδικών στην πρωτεΐνη ακίδας του SARS-CoV-2 T-κυττάρων ήταν χαμηλότερη σε άτομα μεγάλης ηλικίας σε σχέση με νεότερα (Collier et al., 2021). Σε μια άλλη μελέτη απόκρισης στο προαναφερθέν εμβόλιο που πραγματοποιήθηκε σε δύο διαφορετικές ηλικιακές ομάδες (κάτω των 60 και άνω των 80), οι ερευνητές διαπίστωσαν πως μετά τον δεύτερο εμβολιασμό, το 31.3% των

ηλικιωμένων δεν είχε ανιχνεύσιμα εξουδετερωτικά αντισώματα σε αντίθεση με τη νεότερη ομάδα, στην οποία μόνο το 2.2% δεν είχε ανιχνεύσιμα εξουδετερωτικά αντισώματα (Müller et al., 2021).

Όσον αφορά τα γνωρίσματα που αναφέρθηκαν ότι η ύπαρξή τους ωθεί το μοντέλο σε χαμηλότερες τιμές πρόβλεψης (στην προκειμένη χαμηλότερη σοβαρότητα), δηλαδή τα γνωρίσματα πονοκέφαλος, κόπωση, ρίγος, μυαλγία, αρθραλγία, πόνος στην περιοχή εμβολισμού και γριπώδης συνδρομή, ισχύουν τα εξής: το γεγονός ότι εμφανίζονται ως από τα σημαντικότερα γνωρίσματα για την πρόβλεψη σημαίνει πολύ απλά πως το σχετικό πλήθος τους στο σύνολο δεδομένων είναι μεγαλύτερο από άλλα, κάτι που επιβεβαιώνεται και από τα περιγραφικά στοιχεία των γνωρισμάτων που αναφέρθηκαν στο κεφάλαιο 4.1.2, με τον πονοκέφαλο να αποτελεί το γνώρισμα με το μεγαλύτερο ποσοστό. Δεύτερον, το γεγονός ότι τα συγκεκριμένα γνωρίσματα «σπρώχνουν» το μοντέλο προς την πρόβλεψη χαμηλότερης σοβαρότητας, εξηγείται από το ότι αυτά τα συμπτώματα εντοπίζονται ούτως ή άλλως στο μεγαλύτερο μέρος του πληθυσμού και αντιμετωπίζονται σύντομα στις περισσότερες περιπτώσεις. Πιο συγκεκριμένα, αρκετές μελέτες αναφέρουν τον πονοκέφαλο, τη μυαλγία και την κόπωση ως συχνές παρενέργειες μετά τον εμβολιασμό κατά της COVID-19, ιδιαίτερα με εμβόλια mRNA όπως το Pfizer-BioNTech και το Moderna (Rabail et al., 2022). Για παράδειγμα, μια συστηματική ανασκόπηση από τους Folegatti et al. τονίζει την κόπωση και τον πονοκέφαλο ως από τις πιο συχνά αναφερόμενες ανεπιθύμητες ενέργειες σε δοκιμές εμβολίων COVID-19, με την πλειοψηφία να είναι ήπιας έως μέτριας βαρύτητας (Folegatti et al., 2020). Το ρίγος και τα συμπτώματα της γριπώδους συνδρομής, όπως πυρετός και πόνοι στο σώμα, έχουν επίσης αναφερθεί ως συχνές αντιδράσεις στον εμβολιασμό κατά της COVID-19. Μελέτες έχουν βρει ότι αυτά τα συμπτώματα είναι πιο διαδεδομένα μετά τη δεύτερη δόση των εμβολίων mRNA και συχνά είναι ενδεικτικά της ανοσολογικής απόκρισης του οργανισμού στα αντιγόνα του εμβολίου (Baden et al., 2021; WHO, 2021). Ο εντοπισμένος πόνος στο σημείο της ένεσης είναι μια συχνά αναφερόμενη παρενέργεια πολλών εμβολίων, συμπεριλαμβανομένων των εμβολίων COVID-19. Κλινικές δοκιμές και δεδομένα έχουν δείξει σταθερά πόνο, ερυθρότητα και πρήξιμο στο σημείο της ένεσης ως κοινές, παροδικές αντιδράσεις μετά τον εμβολιασμό. Αυτές οι τοπικές αντιδράσεις είναι γενικά ήπιες και υποχωρούν μέσα σε λίγες ημέρες χωρίς σημαντικές επιπλοκές (Polack et al., 2020; Rabail et al., 2022).

Σχετικά με τα γνωρίσματα που ωθούν το μοντέλο σε πρόβλεψη μεγαλύτερης σοβαρότητας, δηλαδή η δύσπνοια, ο πόνος στο στήθος, η έμεση, η υπερευαισθησία και τα επιληπτικά επεισόδια αυτά είναι δυνητικά ενδεικτικά πιο σοβαρών ανεπιθύμητων ενεργειών. Η δύσπνοια και ο πόνος στο στήθος είναι συμπτώματα που μπορεί να υποδηλώνουν αναπνευστικές ή καρδιαγγειακές επιπλοκές μετά τον εμβολιασμό. Αν και σπάνια, σοβαρές ανεπιθύμητες ενέργειες όπως

μυοκαρδίτιδα και περικαρδίτιδα έχουν αναφερθεί μετά τον εμβολιασμό κατά της COVID-19, ιδιαίτερα σε νεότερα άτομα (Gentry et al., 2023) και αρκετές μελέτες έχουν τεκμηριώσει περιπτώσεις πόνου στο στήθος και δύσπνοιας που σχετίζεται με μυοκαρδίτιδα και περικαρδίτιδα μετά από εμβολιασμό με mRNA COVID-19 εμβόλιο (Diaz et al., 2021; Pirzada et al., 2020; Schroth et al., 2023). Η έμεση και οι αντιδράσεις υπερευαισθησίας, συμπεριλαμβανομένης της αναφυλαξίας, είναι σοβαρές ανεπιθύμητες ενέργειες που έχουν αναφερθεί μετά τον εμβολιασμό κατά της COVID-19. Η αναφυλαξία είναι μια σοβαρή, δυνητικά απειλητική για τη ζωή αλλεργική αντίδραση που χαρακτηρίζεται από συμπτώματα όπως έμετος, δυσκολία στην αναπνοή και εξάνθημα. Αν και η αναφυλαξία είναι σπάνια, απαιτεί άμεση ιατρική φροντίδα και μπορεί να εμφανιστεί λίγο μετά τον εμβολιασμό (CDCMMWR, 2021; Shimabukuro & Nair, 2021). Μάλιστα, η αναφερόμενη συχνότητα εμφάνισης άμεσων αντιδράσεων υπερευαισθησίας συμπεριλαμβανομένης της αναφυλαξίας μετά τον εμβολιασμό κατά της COVID-19 είναι 10 φορές υψηλότερη από ό,τι για άλλα εμβόλια (Ieven et al., 2022). Τέλος, οι επιληπτικές κρίσεις είναι νευρολογικά συμβάντα που μπορεί να εμφανιστούν μετά τον εμβολιασμό, αν και είναι σπάνια. Αναφορές περιστατικών και μελέτες έχουν τεκμηριώσει τις κρίσεις ως πιθανά ανεπιθύμητα συμβάντα μετά τον εμβολιασμό κατά της COVID-19, ιδιαίτερα σε άτομα με ιστορικό επιληπτικών διαταραχών ή άλλων νευρολογικών καταστάσεων (Pang et al., 2023; Shah et al., 2022).

Σχετικά με την εμφάνιση περισσότερων θετικών τιμών SHAP για το εμβόλιο της Moderna, αυτό μπορεί να εξηγηθεί με το ότι ενώ τα συνολικά προφίλ παρενεργειών για τα διαφορετικά εμβόλια μπορεί να είναι παρόμοια, ορισμένα ειδικά χαρακτηριστικά του εμβολίου θα μπορούσαν να επηρεάσουν τη συχνότητα ή τη σοβαρότητα ορισμένων αντιδράσεων. Για παράδειγμα, τα εμβόλια mRNA όπως το Moderna και το Pfizer-BioNTech έχουν συσχετιστεί με ελαφρώς υψηλότερη συχνότητα τοπικών και συστημικών αντιδράσεων σε σύγκριση με τα εμβόλια ΙΙκών φορέων όπως το εμβόλιο Janssen της Johnson & Johnson. Μάλιστα, σε μια μελέτη που εξέταζε την αντιδραστικότητα των mRNA εμβολίων, το εμβόλιο της Moderna φάνηκε να έχει υψηλότερο ποσοστό παρενεργειών, με περίπου 74% των να αναφέρουν περιστατικά μετά από μία δόση σε αντίθεση με το 65.4% που ανέφερε επεισόδια μετά από μια δόση εμβολίου της Pfizer. Αυτό το μοτίβο ενισχύθηκε με τη δεύτερη δόση, όπου το 82% των ληπτών του εμβολίου της Moderna ανέφεραν ανεπιθύμητη ενέργεια. Οι αποδέκτες της Pfizer είδαν επίσης μια ελαφρά αύξηση στις αντιδράσεις, με το 68.6% να αναφέρει αντίδραση μετά από μια δεύτερη λήψη (Chapin-Bardales et al., 2021).

5.5 ΕΝΑΛΛΑΚΤΙΚΕΣ ΠΡΟΣΕΓΓΙΣΕΙΣ

Σε αυτήν τη μελέτη, χρησιμοποιήσαμε επιβλεπόμενες τεχνικές μηχανικής μάθησης για να ταξινομήσουμε τη σοβαρότητα των παρενεργειών που σχετίζονται με τα εμβόλια κατά της COVID-19. Ωστόσο, οφείλουμε να αναγνωρίσουμε ότι ένα σημαντικό μέρος του αρχικού συνόλου δεδομένων αποκλείστηκε λόγω έλλειψης τιμών στο πεδίο της σοβαρότητας, οι οποίες ανήλθαν σε πάνω από το 50% των παρατηρήσεων. Αυτή η εξαίρεση οδήγησε δυνητικά στην απώλεια πολύτιμων πληροφοριών και μπορεί να προκατέλαβε την ανάλυσή μας. Για την αντιμετώπιση τέτοιων προκλήσεων, θα μπορούσαν να είχαν εξεταστεί εναλλακτικές προσεγγίσεις όπως η μη-επιβλεπόμενη ή η ημι-επιβλεπόμενη μηχανική εκμάθηση, όπως αναφέρθηκαν στην ενότητα Εισαγωγή. Μέθοδοι μη-επιβλεπόμενης μάθησης, όπως αλγόριθμοι ομαδοποίησης, θα μπορούσαν να είχαν χρησιμοποιηθεί για τον εντοπισμό φυσικών ομαδοποιήσεων ή μοτίβων εντός των δεδομένων χωρίς να υπάρχει ανάγκη για παρατηρήσεις με αντίστοιχη κλάση. Ομοίως, οι ημι-επιβλεπόμενες τεχνικές ενσωματώνουν δεδομένα τόσο με την πληροφορία της κλάσης όσο και χωρίς αυτή για να βελτιώσουν την απόδοση του μοντέλου, κάτι που θα μπορούσε να ήταν ιδιαίτερα επωφελές στο σενάριο μας όπου ένα σημαντικό μέρος των δεδομένων δεν είχε αντίστοιχη κλάση. Ενώ η επιβλεπόμενη μάθηση προσφέρει το πλεονέκτημα της χρήσης επισημασμένων δεδομένων για την εκπαίδευση μοντέλων που προσφέρουν μεγαλύτερη ακρίβεια, βασίζεται σε μεγάλο βαθμό στη διαθεσιμότητα επισημασμένων παρατηρήσεων, κάτι το οποίο μπορεί να μην είναι πάντα εφικτό, όπως αποδεικνύεται και από το σύνολο δεδομένων μας. Από την άλλη πλευρά, οι μέθοδοι μη-επιβλεπόμενης μάθησης δεν απαιτούν επισημασμένα δεδομένα, αλλά μπορεί να δυσκολεύονται με την ερμηνευσιμότητα και μπορεί να μην μοντελοποιούν ρητά τις κλάσεις σοβαρότητας ενδιαφέροντος. Η ημι-επιβλεπόμενη μάθηση επιτυγχάνει μια ισορροπία μεταξύ των δύο, χρησιμοποιώντας δεδομένα τόσο επισημασμένα όσο και μη-επισημασμένα, αλλά μπορεί να δημιουργήσει πρόσθετη πολυπλοκότητα στην ανάπτυξη και τη ρύθμιση των μοντέλων (Jain et al., 1999; Olivier, 2006; Zhu & Goldberg, 2022).

6 ΣΥΜΠΕΡΑΣΜΑΤΑ

Η τρέχουσα πανδημία της νόσου του κορωνοϊού (COVID-19), οδήγησε σε ιδιαίτερα ταχεία παραγωγή εμβολίων κατά της νόσου, γεγονός που προκάλεσε και αυξημένες αναφορές παρενεργειών από τα αντίστοιχα εμβόλια. Η βάση δεδομένων EudraVigilance αποτελεί ένα σύστημα συλλογής, διαχείρισης και ανάλυσης ύποπτων ανεπιθύμητων παρενεργειών φαρμάκων και εμβολίων που έχουν εγκριθεί στην Ευρωπαϊκή Ένωση, συμπεριλαμβανομένων και των παρενεργειών των εμβολίων κατά της COVID-19 και αποτέλεσε την πρωτογενή βάση συλλογής του συνόλου δεδομένων της παρούσας διπλωματικής εργασίας.

Στόχος της συγκεκριμένης διπλωματικής εργασίας ήταν η ανάπτυξη ενός μοντέλου ταξινόμησης μηχανικής μάθησης που στοχεύει στην πρόβλεψη της σοβαρότητας των παρενεργειών των εμβολίων κατά της COVID-19. Έπειτα από εκτενή ανάλυση και πειραματισμό, καταλήξαμε πως οι αλγόριθμοι Random Forest και XGBoost αναδείχθηκαν ως οι αλγόριθμοι με τις κορυφαίες επιδόσεις για την ταξινόμηση των τάξεων σοβαρότητας, επιδεικνύοντας σταθερά ανώτερη απόδοση στα διάφορα μεγέθη δειγμάτων. Ωστόσο, η περαιτέρω διερεύνηση με πιθανές μεθόδους ensemble και ο καλύτερος συντονισμός των υπερπαραμέτρων θα μπορούσε ενδεχομένως να βελτιώσει ακόμη περισσότερο την απόδοση της ταξινόμησης. Επιπλέον, η διερεύνηση του αντίκτυπου των διαφορετικών τεχνικών προεπεξεργασίας και των στρατηγικών αξιολόγησης μοντέλων θα μπορούσε να προσφέρει περεταίρω γνώσεις για τη βελτιστοποίηση της απόδοσης του μοντέλου για αυτήν τη συγκεκριμένη εργασία.

Με τον υπολογισμό των τιμών SHAP για το συγκεκριμένο μοντέλο, μπορέσαμε να ξεχωρίσουμε τα γνωρίσματα που έπαιξαν το σημαντικότερο ρόλο στην προβλεψιμότητά του, με την ηλικία να κατέχει την πρώτη θέση.

Συμπερασματικά, η βελτιστοποίηση ενός τέτοιου μοντέλου θα μπορούσε να συμβάλλει σημαντικά στην ενίσχυση των στρατηγικών εμβολιασμού και της διαχείρισης της υγειονομικής περίθαλψης στη μάχη κατά της πανδημίας, την προώθηση της τεκμηριωμένης λήψης αποφάσεων και, εν τέλει, τη διασφάλιση της δημόσιας υγείας.

7 ΒΙΒΛΙΟΓΡΑΦΙΑ

- Abbattista, M., Martinelli, I., & Peyvandi, F. (2021). Comparison of adverse drug reactions among four COVID-19 vaccines in Europe using the EudraVigilance database: Thrombosis at unusual sites. *J Thromb Haemost*. <https://doi.org/10.1111/jth.15493>
- Abramovich, F., Grinshtein, V., & Levy, T. (2021). Multiclass Classification by Sparse Multinomial Logistic Regression. *IEEE Transactions on Information Theory*, 67(7), 4637–4646. <https://doi.org/10.1109/TIT.2021.3075137>
- Abu Alfeilat, H. A., Hassanat, A. B. A., Lasassmeh, O., Tarawneh, A. S., Alhasanat, M. B., Eyal Salman, H. S., & Prasath, V. B. S. (2019). Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review. *Big Data*, 7(4), 221–248. <https://doi.org/10.1089/big.2018.0175>
- Adankon, M. M., & Cheriet, M. (2009). Support Vector Machine. In S. Z. Li & A. Jain (Eds.), *Encyclopedia of Biometrics* (pp. 1303–1308). Springer US. https://doi.org/10.1007/978-0-387-73003-5_299
- Agrawal, R. (2018). Integrated Effect of Nearest Neighbors and Distance Measures in k-NN Algorithm. In V. B. Aggarwal, V. Bhatnagar, & D. K. Mishra (Eds.), *Big Data Analytics* (pp. 759–766). Springer Singapore.
- Ao, S.-I., Rieger, B. B., & Amouzegar, M. (2010). *Machine learning and systems engineering* (Vol. 68). Springer Science & Business Media.
- Apolloni, B., Ghosh, A., Alpaslan, F., & Patnaik, S. (2005). *Machine learning and robot perception* (Vol. 7). Springer Science & Business Media.
- Arons, M. M., Hatfield, K. M., Reddy, S. C., Kimball, A., James, A., Jacobs, J. R., Taylor, J., Spicer, K., Bardossy, A. C., Oakley, L. P., Tanwar, S., Dyal, J. W., Harney, J., Chisty, Z., Bell, J. M., Methner, M., Paul, P., Carlson, C. M., McLaughlin, H. P., ... Jernigan, J. A. (2020). Presymptomatic SARS-CoV-2 Infections and Transmission in a Skilled Nursing Facility. *The New England Journal of Medicine*. <https://doi.org/10.1056/NEJMoa2008457>
- Awan, A. A. (2023). *An Introduction to SHAP Values and Machine Learning Interpretability*. <https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability>
- Baden, L. R., El Sahly, H. M., Essink, B., Kotloff, K., Frey, S., Novak, R., Diemert, D., Spector, S. A., Rouphael, N., Creech, C. B., McGgettigan, J., Khetan, S., Segall, N., Solis, J., Brosz, A., Fierro, C., Schwartz, H., Neuzil, K., Corey, L., ... Cove Study Group. (2021). Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *N Engl J Med*, 384, 403–416. <https://doi.org/10.1056/NEJMoa2035389>
- Baeldung. (2023). *Multiclass Classification Using Support Vector Machines*. <https://www.baeldung.com/cs/svm-multiclass-classification>
- Bar-Or, A., Schuster, A., Wolff, R., & Keren, D. (2005). Decision Tree Induction in High Dimensional, Hierarchically Distributed Databases. In *Proceedings of the 2005 SIAM International Conference on Data Mining (SDM)* (1–0, pp. 466–470). Society for Industrial and Applied Mathematics. <https://doi.org/10.1137/1.978161972757.42>
- Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1), 20–29.

- Beijerinck, M. W. (1898). *Ueber ein Contagium vivum fluidum als Ursache der Fleckenkrankheit der Tabaksblätter*.
- Bhattacharjee, S., & Banerjee, M. (2020). Immune Thrombocytopenia Secondary to COVID-19: a Systematic Review. *SN Compr Clin Med*, 1–11. <https://doi.org/10.1007/s42399-020-00521-8>
- Bishop, C. (2006). *Pattern Recognition and Machine Learning* (Vol. 4). <https://link.springer.com/book/9780387310732>
- Bogawar, P. S., & Bhoyar, K. K. (2018). An improved multiclass support vector machine classifier using reduced hyper-plane with skewed binary tree. *Applied Intelligence*, 48(11), 4382–4391. <https://doi.org/10.1007/s10489-018-1218-y>
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140. <https://doi.org/10.1007/BF00058655>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Brownlee, J. (2016, August 16). A Gentle Introduction to XGBoost for Applied Machine Learning. *MachineLearningMastery.Com*. <https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/>
- Brownlee, J. (2020, June 11). Ordinal and One-Hot Encodings for Categorical Data. *MachineLearningMastery.Com*. <https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/>
- Butler, N., Pewe, L., Trandem, K., & Perlman, S. (2006). Murine encephalitis caused by HCoV-OC43, a human coronavirus with broad species specificity, is partly immune-mediated. *Virology*, 347(2), 410–421. <https://doi.org/10.1016/j.virol.2005.11.044>
- Caragea, D., Silvescu, A., & Honavar, V. (2004). A Framework for Learning from Distributed Data Using Sufficient Statistics and its Application to Learning Decision Trees. *International Journal of Hybrid Intelligent Systems*, 1(1–2), 80–89. <https://doi.org/10.3233/HIS-2004-11-210>
- CDCMMWR. (2021). Allergic Reactions Including Anaphylaxis After Receipt of the First Dose of Pfizer-BioNTech COVID-19 Vaccine — United States, December 14–23, 2020. *MMWR. Morbidity and Mortality Weekly Report*, 70. <https://doi.org/10.15585/mmwr.mm7002e1>
- CFI team. (2024). *Python (in Machine Learning)*. Corporate Finance Institute. <https://corporatefinanceinstitute.com/resources/data-science/python-in-machine-learning/>
- Chan, J. F.-W., Yuan, S., Kok, K.-H., To, K. K.-W., Chu, H., Yang, J., Xing, F., Liu, J., Yip, C. C.-Y., Poon, R. W.-S., Tsoi, H.-W., Lo, S. K.-F., Chan, K.-H., Poon, V. K.-M., Chan, W.-M., Ip, J. D., Cai, J.-P., Cheng, V. C.-C., Chen, H., ... Yuen, K.-Y. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet (London, England)*, 395(10223), 514–523. [https://doi.org/10.1016/S0140-6736\(20\)30154-9](https://doi.org/10.1016/S0140-6736(20)30154-9)
- Chandra, B., & Varghese, P. P. (2009). Fuzzifying Gini Index based decision trees. *Expert Systems with Applications*, 36(4), 8549–8559. <https://doi.org/https://doi.org/10.1016/j.eswa.2008.10.053>
- Chao, J. Y., Derespina, K. R., Herold, B. C., Goldman, D. L., Aldrich, M., Weingarten, J., Ushay, H. M., Cabana, M. D., & Medar, S. S. (2020). Clinical Characteristics and Outcomes of Hospitalized and Critically Ill Children and Adolescents with Coronavirus Disease 2019 at a Tertiary Care

- Medical Center in New York City. *The Journal of Pediatrics*, 223, 14–19.e2. <https://doi.org/10.1016/j.jpeds.2020.05.006>
- Chapin-Bardales, J., Gee, J., & Myers, T. (2021). Reactogenicity Following Receipt of mRNA-Based COVID-19 Vaccines. *JAMA*, 325(21), 2201–2202. <https://doi.org/10.1001/jama.2021.5374>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
- Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., Wei, Y., Xia, J., Yu, T., Zhang, X., & Zhang, L. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet*, 395, 507–513. [https://doi.org/10.1016/S0140-6736\(20\)30211-7](https://doi.org/10.1016/S0140-6736(20)30211-7)
- Chen, T., & Guestrin, C. (2016, March 9). *XGBoost: A Scalable Tree Boosting System*. ArXiv.Org. <https://doi.org/10.1145/2939672.2939785>
- Cheng, V. C., Lau, S. K., Woo, P. C., & Yuen, K. Y. (2007). Severe acute respiratory syndrome coronavirus as an agent of emerging and reemerging infection. *Clin Microbiol Rev*, 20, 660–694. <https://doi.org/10.1128/CMR.00023-07>
- Cleophas, T. J., Zwinderman, A. H., & Cleophas-Allers, H. I. (2013). *Machine learning in medicine* (Vol. 9). Springer.
- Collier, D. A., Ferreira, I. A. T. M., Kotagiri, P., Datir, R. P., Lim, E. Y., Touizer, E., Meng, B., Abdullahi, A., Elmer, A., Kingston, N., Graves, B., Le Gresley, E., Caputo, D., Bergamaschi, L., Smith, K. G. C., Bradley, J. R., Ceron-Gutierrez, L., Cortes-Acevedo, P., Barcenas-Morales, G., ... Gupta, R. K. (2021). Age-related immune response heterogeneity to SARS-CoV-2 vaccine BNT162b2. *Nature*, 596(7872), 417–422. <https://doi.org/10.1038/s41586-021-03739-1>
- Cui, J., Li, F., & Shi, Z. L. (2019). Origin and evolution of pathogenic coronaviruses. *Nat Rev Microbiol*, 17, 181–192. <https://doi.org/10.1038/s41579-018-0118-9>
- Cunningham, P., Cord, M., & Delany, S. J. (2008). Supervised Learning. In M. Cord & P. Cunningham (Eds.), *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval* (pp. 21–49). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-75171-7_2
- Cunningham, P., & Delany, S. J. (2021). k-Nearest Neighbour Classifiers - A Tutorial. *ACM Comput. Surv.*, 54(6), Article 128.
- developNET. (2023). Τι Είναι η SQL και Πώς Λειτουργεί; Big Blue Data Academy. <https://bigblue.academy/gr/ti-einai-i-sql>
- Diaz, G. A., Parsons, G. T., Gering, S. K., Meier, A. R., Hutchinson, I. V., & Robicsek, A. (2021). Myocarditis and Pericarditis After Vaccination for COVID-19. *JAMA*, 326(12), 1210–1212. <https://doi.org/10.1001/jama.2021.13443>
- Drosten, C., Gunther, S., Preiser, W., van der Werf, S., Brodt, H. R., Becker, S., Rabenau, H., Panning, M., Kolesnikova, L., Fouchier, R. A., Berger, A., Burguiere, A. M., Cinatl, J., Eickmann, M., Escriou, N., Grywna, K., Kramme, S., Manuguerra, J. C., Muller, S., ... Doerr, H. W. (2003). Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N Engl J Med*, 348, 1967–1976. <https://doi.org/10.1056/NEJMoa030747>
- El Naqa, I., & Murphy, M. J. (2015). What Is Machine Learning? In I. El Naqa, R. Li, & M. J. Murphy (Eds.), *Machine Learning in Radiation Oncology: Theory and Applications* (pp. 3–11). Springer International Publishing. https://doi.org/10.1007/978-3-319-18305-3_1
- EliteDataScience. (2022). *How to Handle Imbalanced Classes in Machine Learning*. EliteDataScience. <https://elitedatascience.com/imbalanced-classes>

- Encord. (2023). *What is a Confusion Matrix? / Machine Learning Glossary / Encord.* <https://encord.com/glossary/confusion-matrix/>
- European Comission. (2001). *Directive 2001/83/EC of the European Parliament and of the Council of 6 November 2001 on the Community code relating to medicinal products for human use.* https://health.ec.europa.eu/medicinal-products/eudralex/eudralex-volume-1_en
- European Comission. (2004). *Regulation (EC) No 726/2004 of the European Parliament and of the Council of 31 March 2004 laying down Community procedures for the authorisation and supervision of medicinal products for human and veterinary use and establishing a European Medicines Agency.*
- European Medicines Agency (EMA). (2017). *EudraVigilance - European database of suspected adverse reactions related to medicines: User Manual for online access via the adrreports.eu portal.*
- FDA. (2021). *Pfizer-BioNTech COVID-19 vaccine EUA amendment review memorandum (EUA 27034–amendment 132).*
- Ferretti, L., Wymant, C., Kendall, M., Zhao, L., Nurtay, A., Abeler-Dörner, L., Parker, M., Bonsall, D., & Fraser, C. (2020). Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science (New York, N.Y.),* 368(6491). <https://doi.org/10.1126/science.abb6936>
- Fessenden, J. (2020, December 18). *Inside the new mRNA vaccines for COVID-19.* UMass Chan Medical School. <https://www.umassmed.edu/news/news-archives/2020/12/inside-the-new-mrna-vaccines-for-covid-19/>
- Folegatti, P. M., Ewer, K. J., Aley, P. K., Angus, B., Becker, S., Belij-Rammerstorfer, S., Bellamy, D., Bibi, S., Bittaye, M., Clutterbuck, E. A., Dold, C., Faust, S. N., Finn, A., Flaxman, A. L., Hallis, B., Heath, P., Jenkin, D., Lazarus, R., Makinson, R., ... Yau, Y. (2020). Safety and immunogenicity of the ChAdOx1 nCoV-19 vaccine against SARS-CoV-2: a preliminary report of a phase 1/2, single-blind, randomised controlled trial. *The Lancet,* 396(10249), 467–478. [https://doi.org/10.1016/S0140-6736\(20\)31604-4](https://doi.org/10.1016/S0140-6736(20)31604-4)
- Gamal, B. (2022, January 14). Performance Metrics for Classification Models in Machine Learning: Part II. *Medium.* <https://bassantgz30.medium.com/performance-metrics-for-classification-models-in-machine-learning-part-ii-9303a1c7cadd>
- GeeksforGeeks. (2018, October 15). Label Encoding in Python. *GeeksforGeeks.* <https://www.geeksforgeeks.org/ml-label-encoding-of-datasets-in-python/>
- GeeksforGeeks. (2019, September 23). ML | Handle Missing Data with Simple Imputer. *GeeksforGeeks.* <https://www.geeksforgeeks.org/ml-handle-missing-data-with-simple-imputer/>
- Gentry, V., Brown, N., LaTour, D., Ware, C., Cuevas, A., & Hamra, S. (2023). Chest Pain in a 15-Year-Old Boy Following Administration of Second COVID-19 Vaccine Dose. *Clinical Pediatrics,* 62(1), 73–76. <https://doi.org/10.1177/00099228221111637>
- Gong, Y., & Xu, W. (2007). *Machine learning for multimedia content analysis* (Vol. 30). Springer Science & Business Media.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* MIT press.
- Greinacher, A., Thiele, T., Warkentin, T. E., Weisser, K., Kytle, P. A., & Eichinger, S. (2021). Thrombotic Thrombocytopenia after ChAdOx1 nCov-19 Vaccination. *N Engl J Med,* 384, 2092–2101. <https://doi.org/10.1056/NEJMoa2104840>

- Groff, J. R., Weinberg, P. N., & Oppel, A. J. (2002). *SQL: the complete reference* (Vol. 2). McGraw-Hill/Osborne.
- Guan, W., Ni, Z., Hu, Y., Liang, W., Ou, C., He, J., Liu, L., Shan, H., Lei, C., Hui, D. S. C., Du, B., Li, L., Zeng, G., Yuen, K.-Y., Chen, R., Tang, C., Wang, T., Chen, P., Xiang, J., ... for 2019-nCoV, on behalf of C. M. T. E. G. (2020). Clinical characteristics of 2019 novel coronavirus infection in China. *MedRxiv*, 2020.02.06.20020974. <https://doi.org/10.1101/2020.02.06.20020974>
- Guery, B., Poissy, J., el Mansouf, L., Séjourné, C., Ettahir, N., Lemaire, X., Vuotto, F., Goffard, A., Behillil, S., Enouf, V., Caro, V., Mailles, A., Che, D., Manuguerra, J.-C., Mathieu, D., Fontanet, A., & van der Werf, S. (2013). Clinical features and viral diagnosis of two cases of infection with Middle East Respiratory Syndrome coronavirus: a report of nosocomial transmission. *Lancet (London, England)*, 381(9885), 2265–2272. [https://doi.org/10.1016/S0140-6736\(13\)60982-4](https://doi.org/10.1016/S0140-6736(13)60982-4)
- Guo, R., Zhao, Z., Wang, T., Liu, G., Zhao, J., & Gao, D. (2020). Degradation State Recognition of Piston Pump Based on ICEEMDAN and XGBoost. *Applied Sciences*, 10(18). <https://doi.org/10.3390/app10186593>
- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3(Mar), 1157–1182.
- Gyorfi, L., Ottucsak, G., & Walk, H. (2012). *Machine learning for financial engineering* (Vol. 8). World Scientific.
- Hamre, D., & Procknow, J. J. (1966). A new virus isolated from the human respiratory tract. *Proc Soc Exp Biol Med*, 121, 190–193. <https://doi.org/10.3181/00379727-121-30734>
- Harrison, A. G., Lin, T., & Wang, P. (2020). Mechanisms of SARS-CoV-2 Transmission and Pathogenesis. *Trends Immunol*, 41, 1100–1115. <https://doi.org/10.1016/j.it.2020.10.004>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009a). *The Elements of Statistical Learning* (2nd ed.). Springer New York, NY. <https://link.springer.com/book/10.1007/978-0-387-84858-7>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009b). Unsupervised Learning. In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (pp. 485–585). Springer New York. https://doi.org/10.1007/978-0-387-84858-7_14
- Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). Springer.
- Hatmal, M. M., Al-Hatamleh, M. A. I., Olaimat, A. N., Hatmal, M., Alhaj-Qasem, D. M., Olaimat, T. M., & Mohamud, R. (2021). Side Effects and Perceptions Following COVID-19 Vaccination in Jordan: A Randomized, Cross-Sectional Study Implementing Machine Learning for Predicting Severity of Side Effects. *Vaccines (Basel)*, 9. <https://doi.org/10.3390/vaccines9060556>
- Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., Schiergens, T. S., Herrler, G., Wu, N.-H., Nitsche, A., Müller, M. A., Drosten, C., & Pöhlmann, S. (2020). SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell*, 181(2), 271–280.e8. <https://doi.org/10.1016/j.cell.2020.02.052>
- Hu, L.-Y., Huang, M.-W., Ke, S.-W., & Tsai, C.-F. (2016). The distance function effect on k-nearest neighbor classification for medical datasets. *SpringerPlus*, 5(1), 1304. <https://doi.org/10.1186/s40064-016-2941-7>
- Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., ... Cao, B. (2020). Clinical features of

- patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet (London, England)*, 395(10223), 497–506. [https://doi.org/10.1016/S0140-6736\(20\)30183-5](https://doi.org/10.1016/S0140-6736(20)30183-5)
- Huang, X., Dong, W., Milewska, A., Golda, A., Qi, Y., Zhu, Q. K., Marasco, W. A., Baric, R. S., Sims, A. C., Pyrc, K., Li, W., & Sui, J. (2015). Human Coronavirus HKU1 Spike Protein Uses O-Acetylated Sialic Acid as an Attachment Receptor Determinant and Employs Hemagglutinin-Esterase Protein as a Receptor-Destroying Enzyme. *Journal of Virology*, 89(14), 7202–7213. <https://doi.org/10.1128/JVI.00854-15>
- leven, T., Vandeboterm, M., Nuyttens, L., Devolder, D., Vandenberghe, P., Bullens, D., & Schrijvers, R. (2022). COVID-19 Vaccination Safety and Tolerability in Patients Allegedly at High Risk for Immediate Hypersensitivity Reactions. *Vaccines*, 10(2). <https://doi.org/10.3390/vaccines10020286>
- Iwanowski, D. (1968). *Concerning the mosaic disease of the tobacco plant*.
- Jackson, L. A., Anderson, E. J., Rouphael, N. G., Roberts, P. C., Makhene, M., Coler, R. N., McCullough, M. P., Chappell, J. D., Denison, M. R., & Stevens, L. J. (2020). An mRNA vaccine against SARS-CoV-2—preliminary report. *New England Journal of Medicine*, 383(20), 1920–1931.
- Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys (CSUR)*, 31(3), 264–323.
- Karl, T. (2023). *6 Reasons Why Is Python Used for Machine Learning*. New Horizons. <https://www.newhorizons.com/resources/blog/why-is-python-used-for-machine-learning>
- Kaur, N., & Himanshu. (2023). Logistic Regression: A Basic Approach. In A. Joshi, M. Mahmud, & R. G. Ragel (Eds.), *Information and Communication Technology for Competitive Strategies (ICTCS 2022)* (pp. 481–488). Springer Nature Singapore.
- Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In *Machine learning proceedings 1992* (pp. 249–256). Elsevier.
- Kohavi, R. (2001). *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*. 14.
- Kolade, C. (2021, August 30). *What is PHP? The PHP Programming Language Meaning Explained*. FreeCodeCamp.Org. <https://www.freecodecamp.org/news/what-is-php-the-php-programming-language-meaning-explained/>
- Krogh, J. W. (2020). MySQL Workbench. In *MySQL 8 Query Performance Tuning* (pp. 199–226). Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-5584-1_11
- Ksiazek, T. G., Erdman, D., Goldsmith, C. S., Zaki, S. R., Peret, T., Emery, S., Tong, S., Urbani, C., Comer, J. A., Lim, W., Rollin, P. E., Dowell, S. F., Ling, A. E., Humphrey, C. D., Shieh, W. J., Guarner, J., Paddock, C. D., Rota, P., Fields, B., ... Sars Working Group. (2003). A novel coronavirus associated with severe acute respiratory syndrome. *N Engl J Med*, 348, 1953–1966. <https://doi.org/10.1056/NEJMoa030781>
- Lee, J. L., & Linterman, M. A. (2022). Mechanisms underpinning poor antibody responses to vaccines in ageing. *Immunology Letters*, 241, 1. <https://doi.org/10.1016/j.imlet.2021.11.001>
- Li, G., Fan, Y., Lai, Y., Han, T., Li, Z., Zhou, P., Pan, P., Wang, W., Hu, D., Liu, X., Zhang, Q., & Wu, J. (2020). Coronavirus infections and immune responses. *J Med Virol*, 92, 424–432. <https://doi.org/10.1002/jmv.25685>
- Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S. M., Lau, E. H. Y., Wong, J. Y., Xing, X., Xiang, N., Wu, Y., Li, C., Chen, Q., Li, D., Liu, T., Zhao, J., Liu, M., ... Feng, Z. (2020).

- Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *The New England Journal of Medicine*, 382(13), 1199–1207. <https://doi.org/10.1056/NEJMoa2001316>
- Li, W., Moore, M. J., Vasilieva, N., Sui, J., Wong, S. K., Berne, M. A., Somasundaran, M., Sullivan, J. L., Luzuriaga, K., Greenough, T. C., Choe, H., & Farzan, M. (2003). Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature*, 426(6965), 450–454. <https://doi.org/10.1038/nature02145>
- Li, W., Sui, J., Huang, I.-C., Kuhn, J. H., Radoshitzky, S. R., Marasco, W. A., Choe, H., & Farzan, M. (2007). The S proteins of human coronavirus NL63 and severe acute respiratory syndrome coronavirus bind overlapping regions of ACE2. *Virology*, 367(2), 367–374. <https://doi.org/10.1016/j.virol.2007.04.035>
- Lim, Y. X., Ng, Y. L., Tam, J. P., & Liu, D. X. (2016). Human Coronaviruses: A Review of Virus-Host Interactions. *Diseases*, 4. <https://doi.org/10.3390/diseases4030026>
- Liu, Y., Ning, Z., Chen, Y., Guo, M., Liu, Y., Gali, N. K., Sun, L., Duan, Y., Cai, J., Westerdahl, D., Liu, X., Xu, K., Ho, K., Kan, H., Fu, Q., & Lan, K. (2020). Aerodynamic analysis of SARS-CoV-2 in two Wuhan hospitals. *Nature*, 582(7813), 557–560. <https://doi.org/10.1038/s41586-020-2271-3>
- Louppe, G. (2014, July 28). *Understanding Random Forests: From Theory to Practice*. ArXiv.Org. <https://arxiv.org/abs/1407.7502v3>
- Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., Bi, Y., Ma, X., Zhan, F., Wang, L., Hu, T., Zhou, H., Hu, Z., Zhou, W., Zhao, L., ... Tan, W. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet*, 395, 565–574. [https://doi.org/10.1016/S0140-6736\(20\)30251-8](https://doi.org/10.1016/S0140-6736(20)30251-8)
- Lucidchart. (2024). *What is an Entity Relationship Diagram (ERD)?* Lucidchart. <https://www.lucidchart.com/pages/er-diagrams>
- Ludwig, S., & Zarbock, A. (2020). Coronaviruses and SARS-CoV-2: A Brief Overview. *Anesth Analg*, 131, 93–96. <https://doi.org/10.1213/ANE.0000000000004845>
- Lundberg, S. M. (2018). *Welcome to the SHAP documentation — SHAP latest documentation*. <https://shap.readthedocs.io/en/latest/>
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56–67.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
- Luo, H., Chen, Q., Chen, J., Chen, K., Shen, X., & Jiang, H. (2005). The nucleocapsid protein of SARS coronavirus has a high binding affinity to the human cellular heterogeneous nuclear ribonucleoprotein A1. *FEBS Letters*, 579(12), 2623–2628. <https://doi.org/10.1016/j.febslet.2005.03.080>
- Lustig, A., & Levine, A. J. (1992). One hundred years of virology. *Journal of Virology*, 66, 4629–4631.
- Lwoff, A. (1957). The concept of virus. *Microbiology*, 17, 239–253.
- Malley, J. D., Malley, K. G., & Pajevic, S. (2011). *Statistical learning for biomedical data*. Cambridge University Press.

- Mascellino, M. T., Di Timoteo, F., De Angelis, M., & Oliva, A. (2021). Overview of the Main Anti-SARS-CoV-2 Vaccines: Mechanism of Action, Efficacy and Safety. *Infection and Drug Resistance*, 14, 3459–3476. <https://doi.org/10.2147/IDR.S315727>
- Mathieu, E., Ritchie, H., Rodés-Guirao, L., Appel, C., Giattino, C., Hasell, J., Macdonald, B., Dattani, S., Beltekian, D., Ortiz-Ospina, E., & Roser, M. (2020). Coronavirus Pandemic (COVID-19). *Our World in Data*. <https://ourworldindata.org/covid-vaccinations>
- McBride, R., & Fielding, B. C. (2012). The role of severe acute respiratory syndrome (SARS)-coronavirus accessory proteins in virus pathogenesis. *Viruses*, 4, 2902–2923. <https://doi.org/10.3390/v4112902>
- McIntosh, K., Dees, J. H., Becker, W. B., Kapikian, A. Z., & Chanock, R. M. (1967). Recovery in tracheal organ cultures of novel viruses from patients with respiratory disease. *Proc Natl Acad Sci U S A*, 57, 933–940. <https://doi.org/10.1073/pnas.57.4.933>
- McIntosh, K., & Peiris, J. S. M. (2009). Coronaviruses. *Clinical Virology*, 1155–1171.
- Mevorach, D., Anis, E., Cedar, N., Bromberg, M., Haas, E. J., Nadir, E., Olsha-Castell, S., Arad, D., Hasin, T., Levi, N., Asleh, R., Amir, O., Meir, K., Cohen, D., Dichtiar, R., Novick, D., Hershkovitz, Y., Dagan, R., Leitersdorf, I., ... Alroy-Preis, S. (2021). Myocarditis after BNT162b2 mRNA Vaccine against Covid-19 in Israel. *N Engl J Med*. <https://doi.org/10.1056/NEJMoa2109730>
- Mitra, S., Datta, S., Perkins, T., & Michailidis, G. (2008). *Introduction to machine learning and bioinformatics*. CRC Press.
- Müller, L., Andrée, M., Moskorz, W., Drexler, I., Walotka, L., Grothmann, R., Ptok, J., Hillebrandt, J., Ritchie, A., Rabl, D., Ostermann, P. N., Robitzsch, R., Hauka, S., Walker, A., Menne, C., Grutza, R., Timm, J., Adams, O., & Schaal, H. (2021). Age-dependent immune response to the Biontech/Pfizer BNT162b2 COVID-19 vaccination. *MedRxiv*, 2021.03.03.21251066. <https://doi.org/10.1101/2021.03.03.21251066>
- Mulligan, M. J., Lyke, K. E., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Neuzil, K., Raabe, V., Bailey, R., & Swanson, K. A. (2021). Publisher Correction: Phase I/II study of COVID-19 RNA vaccine BNT162b1 in adults. *Nature*, 590(7844), E26.
- Murtagh, F. (1991). Multilayer perceptrons for classification and regression. *Neurocomputing*, 2(5), 183–197. [https://doi.org/10.1016/0925-2312\(91\)90023-5](https://doi.org/10.1016/0925-2312(91)90023-5)
- Nanda, S. K., & Leibowitz, J. L. (2001). Mitochondrial aconitase binds to the 3' untranslated region of the mouse hepatitis virus genome. *Journal of Virology*, 75(7), 3352–3362. <https://doi.org/10.1128/JVI.75.7.3352-3362.2001>
- Olivier, C. (2006). Semi-supervised learning (adaptive computation and machine learning). *Mit Pr*, 2006.
- Ong, S. W. X., Tan, Y. K., Chia, P. Y., Lee, T. H., Ng, O. T., Wong, M. S. Y., & Marimuthu, K. (2020). Air, Surface Environmental, and Personal Protective Equipment Contamination by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) From a Symptomatic Patient. *JAMA*, 323(16), 1610–1612. <https://doi.org/10.1001/jama.2020.3227>
- Pan, Y., Zhang, D., Yang, P., Poon, L. L. M., & Wang, Q. (2020). Viral load of SARS-CoV-2 in clinical samples. *The Lancet. Infectious Diseases*, 20(4), 411–412. [https://doi.org/10.1016/S1473-3099\(20\)30113-4](https://doi.org/10.1016/S1473-3099(20)30113-4)
- Pang, E. W., Lawn, N. D., Chan, J., Lee, J., & Dunne, J. W. (2023). COVID-19 vaccination-related exacerbation of seizures in persons with epilepsy. *Epilepsy & Behavior*, 138, 109024. <https://doi.org/10.1016/j.yebeh.2022.109024>

- Pardi, N., Hogan, M. J., Naradikian, M. S., Parkhouse, K., Cain, D. W., Jones, L., Moody, M. A., Verkerke, H. P., Myles, A., & Willis, E. (2018). Nucleoside-modified mRNA vaccines induce potent T follicular helper and germinal center B cell responses. *Journal of Experimental Medicine*, 215(6), 1571–1588.
- Patel, R., Kaki, M., Potluri, V. S., Kahar, P., & Khanna, D. (2022). A comprehensive review of SARS-CoV-2 vaccines: Pfizer, Moderna & Johnson & Johnson. *Human Vaccines & Immunotherapeutics*, 18(1). <https://doi.org/10.1080/21645515.2021.2002083>
- Peiris, J. S. M., Lai, S. T., Poon, L. L. M., Guan, Y., Yam, L. Y. C., Lim, W., Nicholls, J., Yee, W. K. S., Yan, W. W., Cheung, M. T., Cheng, V. C. C., Chan, K. H., Tsang, D. N. C., Yung, R. W. H., Ng, T. K., & Yuen, K. Y. (2003). Coronavirus as a possible cause of severe acute respiratory syndrome. *The Lancet*, 361, 1319–1325. [https://doi.org/https://doi.org/10.1016/S0140-6736\(03\)13077-2](https://doi.org/https://doi.org/10.1016/S0140-6736(03)13077-2)
- Pérez-Enciso, & Zingaretti, L. (2019). A Guide for Using Deep Learning for Complex Trait Genomic Prediction. *Genes*, 10, 553. <https://doi.org/10.3390/genes10070553>
- Perlman, S., & Netland, J. (2009). Coronaviruses post-SARS: update on replication and pathogenesis. *Nature Reviews Microbiology*, 7(6), 439–450. <https://doi.org/10.1038/nrmicro2147>
- Pirzada, A., Mokhtar, A. T., & Moeller, A. D. (2020). COVID-19 and Myocarditis: What Do We Know So Far? *CJC Open*, 2(4), 278–285. <https://doi.org/10.1016/j.cjco.2020.05.005>
- Polack, F. P., Thomas, S. J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Perez, J. L., Perez Marc, G., Moreira, E. D., Zerbini, C., Bailey, R., Swanson, K. A., Roychoudhury, S., Koury, K., Li, P., Kalina, W. V., Cooper, D., Frenck, R. W., Hammitt, L. L., ... C. Clinical Trial Group. (2020). Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *N Engl J Med*, 383, 2603–2615. <https://doi.org/10.1056/NEJMoa2034577>
- Postigo, R., Brosch, S., Slattery, J., van Haren, A., Dogne, J. M., Kurz, X., Candore, G., Domergue, F., & Arlett, P. (2018). EudraVigilance Medicines Safety Database: Publicly Accessible Data for Research and Public Health Protection. *Drug Saf*, 41, 665–675. <https://doi.org/10.1007/s40264-018-0647-1>
- Puntmann, V. O., Carerj, M. L., Wieters, I., Fahim, M., Arendt, C., Hoffmann, J., Shchendrygina, A., Escher, F., Vasa-Nicotera, M., Zeiher, A. M., Vehreschild, M., & Nagel, E. (2020). Outcomes of Cardiovascular Magnetic Resonance Imaging in Patients Recently Recovered From Coronavirus Disease 2019 (COVID-19). *JAMA Cardiology*, 5(11), 1265–1273. <https://doi.org/10.1001/jamacardio.2020.3557>
- Qi, Z., Tian, Y., & Deng, N. (2005). A New Support Vector Machine for Multi-class Classification. In Y. Hao, J. Liu, Y. Wang, Y. Cheung, H. Yin, L. Jiao, J. Ma, & Y.-C. Jiao (Eds.), *Computational Intelligence and Security* (pp. 580–585). Springer Berlin Heidelberg.
- Qomariyah, N. N., Heriyanni, E., Fajar, A. N., & Kazakov, D. (2020). Comparative Analysis of Decision Tree Algorithm for Learning Ordinal Data Expressed as Pairwise Comparisons. *2020 8th International Conference on Information and Communication Technology (ICoICT)*, 1–4. <https://doi.org/10.1109/ICoICT49345.2020.9166341>
- Rabail, R., Ahmed, W., Ilyas, M., Rajoka, M. S. R., Hassoun, A., Khalid, A. R., Khan, M. R., & Aadil, R. M. (2022). The Side Effects and Adverse Clinical Cases Reported after COVID-19 Immunization. *Vaccines*, 10(4). <https://doi.org/10.3390/vaccines10040488>

- Ravikiran, A. S. (2021, April 27). *Differentiating SQL and MySQL: A Comprehensive Guide*. Simplilearn.Com. <https://www.simplilearn.com/tutorials/sql-tutorial/difference-between-sql-and-mysql>
- Renia, L., Goh, Y. S., Rouers, A., Le Bert, N., Chia, W. N., Chavatte, J.-M., Fong, S., Chang, Z. W., Zhuo, N. Z., Tay, M. Z., Chan, Y.-H., Tan, C. W., Yeo, N. K., Amrun, S. N., Huang, Y., Wong, J. X. E., Hor, P. X., Loh, C. Y., Wang, B., ... Ng, L. F. P. (2022). Lower vaccine-acquired immunity in the elderly population following two-dose BNT162b2 vaccination is alleviated by a third vaccine dose. *Nature Communications*, 13. <https://doi.org/10.1038/s41467-022-32312-1>
- Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386.
- Schmitz, L. (2023, November 16). What Is An Encoder In Machine Learning. *Robots.Net*. <https://robots.net/fintech/what-is-an-encoder-in-machine-learning/>
- Schroth, D., Garg, R., Bocova, X., Hansmann, J., Haass, M., Yan, A., Fernando, C., Chacko, B., Oikonomou, A., White, J., Alhussein, M. M., Giusca, S., Ochs, A., Korosoglou, G., André, F., Friedrich, M. G., & Ochs, M. (2023). Predictors of persistent symptoms after mRNA SARS-CoV-2 vaccine-related myocarditis (myovacc registry). *Frontiers in Cardiovascular Medicine*, 10. <https://doi.org/10.3389/fcvm.2023.1204232>
- Sette, A., & Crotty, S. (2021). Adaptive immunity to SARS-CoV-2 and COVID-19. *Cell*, 184(4), 861–880.
- Shah, T., Figuracion, K. C., Schteiden, B., & Gruber, J. (2022). Breakthrough Seizures after COVID-19 Vaccines in Patients with Glioma. *Neurology*. <https://pesquisa.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov/resource/pt/covidwho-1925499>
- Shimabukuro, T., & Nair, N. (2021). Allergic Reactions Including Anaphylaxis After Receipt of the First Dose of Pfizer-BioNTech COVID-19 Vaccine. *JAMA*, 325(8), 780–781. <https://doi.org/10.1001/jama.2021.0600>
- Shmilovici, A. (2005). Support Vector Machines. In O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 257–276). Springer US. https://doi.org/10.1007/0-387-25465-X_12
- Simplilearn. (2022, November 22). *What is XGBoost? An Introduction to XGBoost Algorithm in Machine Learning* / Simplilearn. Simplilearn.Com. <https://www.simplilearn.com/what-is-xgboost-algorithm-in-machine-learning-article>
- Stanley, W. M. (1935). Isolation of a Crystalline Protein Possessing the Properties of Tobacco-Mosaic Virus. *Science*, 81, 644–645. <https://doi.org/10.1126/science.81.2113.644>
- Stockwell, D. R. B., & Fielding, A. H. (1999). *Machine learning methods for ecological applications*.
- Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41, 647–665.
- Su, S., Wong, G., Shi, W., Liu, J., Lai, A. C. K., Zhou, J., Liu, W., Bi, Y., & Gao, G. F. (2016). Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. *Trends in Microbiology*, 24(6), 490–502. <https://doi.org/10.1016/j.tim.2016.03.003>
- Suvvari, T. K., Rajesh, E., D. Silva RG, Corriero, A. C., & Kutikuppala, L. V. S. (2021). SARS-CoV-2 vaccine-induced prothrombotic immune thrombocytopenia: Promoting awareness to improve patient-doctor trust. *J Med Virol*, 93, 5721–5723. <https://doi.org/10.1002/jmv.27176>
- Tan, Y. W., Hong, W., & Liu, D. X. (2012). Binding of the 5'-untranslated region of coronavirus RNA to zinc finger CCHC-type and RNA-binding motif 1 enhances viral replication and

- transcription. *Nucleic Acids Research*, 40(11), 5065–5077. <https://doi.org/10.1093/nar/gks165>
- Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*, 17(1), 168–192. <https://doi.org/10.1016/j.aci.2018.08.003>
- The COVID-19 Investigation Team. (2020). Clinical and virologic characteristics of the first 12 patients with coronavirus disease 2019 (COVID-19) in the United States. *Nature Medicine*, 26(6), 861–868. <https://doi.org/10.1038/s41591-020-0877-5>
- The Novel Coronavirus Pneumonia Emergency Response Epidemiology Team. (2020). The Epidemiological Characteristics of an Outbreak of 2019 Novel Coronavirus Diseases (COVID-19) — China, 2020. *China CDC Weekly*, 2(8), 113–122.
- Tomek, I. (1976). *AN EXPERIMENT WITH THE EDITED NEAREST-NIEGHBOR RULE*.
- Tomlinson, B., & Cockram, C. (2003). SARS: experience at Prince of Wales Hospital, Hong Kong. *Lancet (London, England)*, 361(9368), 1486–1487. [https://doi.org/10.1016/S0140-6736\(03\)13218-7](https://doi.org/10.1016/S0140-6736(03)13218-7)
- Tyrrell, D. A. (1968). Coronaviruses. *Nature (London)*, 220, 650.
- Tyrrell, D. A., Almeida, J. D., Cunningham, C. H., Dowdle, W. R., Hofstad, M. S., McIntosh, K., Tajima, M., Zakstelskaya, L. Y., Easterday, B. C., Kapikian, A., & Bingham, R. W. (1975). Coronaviridae. *Intervirology*, 5, 76–82. <https://doi.org/10.1159/000149883>
- Tyrrell, D. A., & Bynoe, M. L. (1965). Cultivation of a Novel Type of Common-Cold Virus in Organ Cultures. *Br Med J*, 1, 1467–1470. <https://doi.org/10.1136/bmj.1.5448.1467>
- van Doremalen, N., Miazgowicz, K. L., Milne-Price, S., Bushmaker, T., Robertson, S., Scott, D., Kinne, J., McLellan, J. S., Zhu, J., & Munster, V. J. (2014). Host Species Restriction of Middle East Respiratory Syndrome Coronavirus through Its Receptor, Dipeptidyl Peptidase 4. *Journal of Virology*, 88(16), 9220–9232. <https://doi.org/10.1128/JVI.00676-14>
- Van Rossum, G., & Drake, F. L. (2003). *An introduction to Python*. Network Theory Ltd. Bristol.
- Vaswani, V. (2009). *MySQL database usage & administration*. McGraw Hill Professional.
- Vujović, Ž. (2021). Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications*, 12(6), 599–606.
- Wan, E. Y. F., Chui, C. S. L., Lai, F. T. T., Chan, E. W. Y., Li, X., Yan, V. K. C., Gao, L., Yu, Q., Lam, I. C. H., & Chun, R. K. C. (2022). Bell's palsy following vaccination with mRNA (BNT162b2) and inactivated (CoronaVac) SARS-CoV-2 vaccines: a case series and nested case-control study. *The Lancet Infectious Diseases*, 22(1), 64–72.
- Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., Wang, B., Xiang, H., Cheng, Z., Xiong, Y., Zhao, Y., Li, Y., Wang, X., & Peng, Z. (2020). Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus–Infected Pneumonia in Wuhan, China. *JAMA*, 323(11), 1061–1069. <https://doi.org/10.1001/jama.2020.1585>
- WHO. (2021). Reports of flu-like symptoms after COVID-19 vaccination. *Reactions Weekly*, 1847(1), 4. <https://doi.org/10.1007/s40278-021-92561-3>
- Witberg, G., Barda, N., Hoss, S., Richter, I., Wiessman, M., Aviv, Y., Grinberg, T., Auster, O., Dagan, N., Balicer, R. D., & Kornowski, R. (2021). Myocarditis after Covid-19 Vaccination in a Large Health Care Organization. *N Engl J Med*. <https://doi.org/10.1056/NEJMoa2110737>
- Wölfel, R., Corman, V. M., Guggemos, W., Seilmaier, M., Zange, S., Müller, M. A., Niemeyer, D., Jones, T. C., Vollmar, P., Rothe, C., Hoelscher, M., Bleicker, T., Brünink, S., Schneider, J., Ehmann, R., Zwirglmaier, K., Drosten, C., & Wendtner, C. (2020). Virological assessment of

- hospitalized patients with COVID-2019. *Nature*, 581(7809), 465–469. <https://doi.org/10.1038/s41586-020-2196-x>
- Wu, C.-H., Chen, P.-J., & Yeh, S.-H. (2014). Nucleocapsid phosphorylation and RNA helicase DDX1 recruitment enables coronavirus transition from discontinuous to continuous transcription. *Cell Host & Microbe*, 16(4), 462–472. <https://doi.org/10.1016/j.chom.2014.09.009>
- Wu, D., Wu, T., Liu, Q., & Yang, Z. (2020). The SARS-CoV-2 outbreak: What we know. *Int J Infect Dis*, 94, 44–48. <https://doi.org/10.1016/j.ijid.2020.03.004>
- Wu, K., Li, W., Peng, G., & Li, F. (2009). Crystal structure of NL63 respiratory coronavirus receptor-binding domain complexed with its human receptor. *Proceedings of the National Academy of Sciences of the United States of America*, 106(47), 19970–19974. <https://doi.org/10.1073/pnas.0908837106>
- Wu, Q., Dudley, M. Z., Chen, X., Bai, X., Dong, K., Zhuang, T., Salmon, D., & Yu, H. (2021). Evaluation of the safety profile of COVID-19 vaccines: a rapid review. *BMC Med*, 19, 173. <https://doi.org/10.1186/s12916-021-02059-5>
- XGBoost. (2021, September 18). GeeksforGeeks. <https://www.geeksforgeeks.org/xgboost/>
- Xiao, F., Tang, M., Zheng, X., Liu, Y., Li, X., & Shan, H. (2020). Evidence for Gastrointestinal Infection of SARS-CoV-2. *Gastroenterology*, 158(6), 1831–1833.e3. <https://doi.org/10.1053/j.gastro.2020.02.055>
- Xu, Y., Li, X., Zhu, B., Liang, H., Fang, C., Gong, Y., Guo, Q., Sun, X., Zhao, D., Shen, J., Zhang, H., Liu, H., Xia, H., Tang, J., Zhang, K., & Gong, S. (2020). Characteristics of pediatric SARS-CoV-2 infection and potential evidence for persistent fecal viral shedding. *Nature Medicine*, 26(4), 502–505. <https://doi.org/10.1038/s41591-020-0817-4>
- Yang, Z. R. (2010). *Machine learning approaches to bioinformatics* (Vol. 4). World scientific.
- Yeager, C. L., Ashmun, R. A., Williams, R. K., Cardellicchio, C. B., Shapiro, L. H., Look, A. T., & Holmes, K. V. (1992). Human aminopeptidase N is a receptor for human coronavirus 229E. *Nature*, 357(6377), 420–422. <https://doi.org/10.1038/357420a0>
- Yu, I. T. S., Li, Y., Wong, T. W., Tam, W., Chan, A. T., Lee, J. H. W., Leung, D. Y. C., & Ho, T. (2004). Evidence of Airborne Transmission of the Severe Acute Respiratory Syndrome Virus. *New England Journal of Medicine*, 350(17), 1731–1739. <https://doi.org/10.1056/NEJMoa032867>
- Yu, J., & Tao, D. (2013). *Modern machine learning techniques and their applications in cartoon animation research*. John Wiley & Sons.
- Zaitlin, M. (1998). The Discovery of the Causal Agent of the Tobacco Mosaic Disease. *Discoveries in Plant Biology*, 105–110.
- Zhang, Y., Zeng, G., Pan, H., Li, C., Hu, Y., Chu, K., Han, W., Chen, Z., Tang, R., Yin, W., Chen, X., Hu, Y., Liu, X., Jiang, C., Li, J., Yang, M., Song, Y., Wang, X., Gao, Q., & Zhu, F. (2021). Safety, tolerability, and immunogenicity of an inactivated SARS-CoV-2 vaccine in healthy adults aged 18–59 years: a randomised, double-blind, placebo-controlled, phase 1/2 clinical trial. *The Lancet. Infectious Diseases*, 21(2), 181–192. [https://doi.org/10.1016/S1473-3099\(20\)30843-4](https://doi.org/10.1016/S1473-3099(20)30843-4)
- Zhong, N. S., Zheng, B. J., Li, Y. M., Poon, Xie, Z. H., Chan, K. H., Li, P. H., Tan, S. Y., Chang, Q., Xie, J. P., Liu, X. Q., Xu, J., Li, D. X., Yuen, K. Y., Peiris, & Guan, Y. (2003). Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003. *Lancet*, 362, 1353–1358. [https://doi.org/10.1016/s0140-6736\(03\)14630-2](https://doi.org/10.1016/s0140-6736(03)14630-2)

- Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., Chen, H.-D., Chen, J., Luo, Y., Guo, H., Jiang, R.-D., Liu, M.-Q., Chen, Y., Shen, X.-R., Wang, X., ... Shi, Z.-L. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 579(7798), 270–273. <https://doi.org/10.1038/s41586-020-2012-7>
- Zhu, X., & Goldberg, A. B. (2022). *Introduction to semi-supervised learning*. Springer Nature.
- Zou, L., Ruan, F., Huang, M., Liang, L., Huang, H., Hong, Z., Yu, J., Kang, M., Song, Y., Xia, J., Guo, Q., Song, T., He, J., Yen, H.-L., Peiris, M., & Wu, J. (2020). SARS-CoV-2 Viral Load in Upper Respiratory Specimens of Infected Patients. *The New England Journal of Medicine*, 382(12), 1177–1179. <https://doi.org/10.1056/NEJMc2001737>
- Zumla, A., Chan, J. F. W., Azhar, E. I., Hui, D. S. C., & Yuen, K.-Y. (2016). Coronaviruses - drug discovery and therapeutic options. *Nature Reviews. Drug Discovery*, 15(5), 327–347. <https://doi.org/10.1038/nrd.2015.37>
- Αγγελής, Γ. (2007). *Μικροβιολογία & Μικροβιακή Τεχνολογία*. Εκδόσεις Σταμούλη Α.Ε.