ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΝΟΣΗΛΕΥΤΙΚΗΣ

ΔΙΙΔΡΥΜΑΤΙΚΟ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΕΙΔΙΚΕΥΣΗ: ΠΛΗΡΟΦΟΡΙΚΗ ΤΗΣ ΥΓΕΙΑΣ

ΕΦΑΡΜΟΓΗ ΤΕΧΝΙΚΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΓΙΑ ΤΗΝ ΠΡΟΒΛΕΨΗ ΤΗΣ ΕΚΒΑΣΗΣ ΤΗΣ ΑΝΤΙΦΥΜΑΤΙΚΗΣ ΑΓΩΓΗΣ

ΕΙΡΗΝΗ ΑΡΓΥΡΟΠΟΥΛΟΥ

ΤΕΧΝΟΛΟΓΟΣ ΙΑΤΡΙΚΩΝ ΕΡΓΑΣΤΗΡΙΩΝ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΘΗΝΑ 2024

ΕΦΑΡΜΟΓΗ ΤΕΧΝΙΚΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΓΙΑ ΤΗΝ ΠΡΟΒΛΕΨΗ ΤΗΣ ΕΚΒΑΣΗΣ ΤΗΣ ΑΝΤΙΦΥΜΑΤΙΚΗΣ ΑΓΩΓΗΣ

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΝΟΣΗΛΕΥΤΙΚΗΣ

ΔΙΙΔΡΥΜΑΤΙΚΟ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΕΙΔΙΚΕΥΣΗ:ΠΛΗΡΟΦΟΡΙΚΗ ΤΗΣ ΥΓΕΙΑΣ

ΕΦΑΡΜΟΓΗ ΤΕΧΝΙΚΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΓΙΑ ΤΗΝ ΠΡΟΒΛΕΨΗ ΤΗΣ ΕΚΒΑΣΗΣ ΤΗΣ ΑΝΤΙΦΥΜΑΤΙΚΗΣ ΑΓΩΓΗΣ

ΕΙΡΗΝΗ ΑΡΓΥΡΟΠΟΥΛΟΥ

ΤΕΧΝΟΛΟΓΟΣ ΙΑΤΡΙΚΩΝ ΕΡΓΑΣΤΗΡΙΩΝ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΘΗΝΑ 2024

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

ΔΡ. ΙΩΣΗΦ ΛΙΑΣΚΟΣ, Ε.ΔΙ.Π (ΕΠΙΒΛΕΠΩΝ)

ΔΡ. ΕΜΜΑΝΟΥΗΛ ΖΟΥΛΙΑΣ, Ε.ΔΙ.Π

ΟΜ. ΚΑΘΗΓΗΤΗΣ ΙΩΑΝΝΗΣ ΜΑΝΤΑΣ

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΝΟΣΗΛΕΥΤΙΚΗΣ

ΔΙΙΔΡΥΜΑΤΙΚΟ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΕΙΔΙΚΕΥΣΗ: ΠΛΗΡΟΦΟΡΙΚΗ ΤΗΣ ΥΓΕΙΑΣ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΕΦΑΡΜΟΓΗ ΤΕΧΝΙΚΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΓΙΑ ΤΗΝ ΠΡΟΒΛΕΨΗ ΤΗΣ ΕΚΒΑΣΗΣ ΤΗΣ ΑΝΤΙΦΥΜΑΤΙΚΗΣ ΑΓΩΓΗΣ

ΤΗΣ ΕΙΡΗΝΗΣ ΑΡΓΥΡΟΠΟΥΛΟΥ

ΠΕΡΙΛΗΨΗ

Η φυματίωση είναι μια ιδιαίτερα μεταδοτική ασθένεια και παραμένει μια από τις κύριες αιτίες θανάτου από μολυσματικές ασθένειες σε παγκόσμιο επίπεδο. Παρόλο που έχει επιτευχθεί σημαντική πρόοδος στην πρόληψη, τη διάγνωση και τη θεραπεία της νόσου, τα αναφερόμενα περιστατικά παραμένουν αυξημένα, ιδιαίτερα σε περιοχές με περιορισμένη πρόσβαση σε υγειονομική περίθαλψη. Η ραγδαία εξέλιξη της Τεχνητής Νοημοσύνης θα μπορούσε να αποτελέσει ακρογωνιαίο λίθο στην ανάπτυξη και εφαρμογή νέων θεραπευτικών προσεγγίσεων μέσω της δημιουργίας μοντέλων Μηχανικής Μάθησης.

Η παρούσα εργασία στοχεύει στην δημιουργία ενός αποδοτικού μοντέλου Μηχανικής Μάθησης, το οποίο θα προβλέπει την έκβαση της θεραπευτικής αγωγής των ασθενών με φυματίωση. Το μοντέλο αυτό θα μπορούσε να συμβάλλει στην αύξηση των ποσοστών επιτυχίας της θεραπείας και κατ' επέκταση στον περιορισμό της θνησιμότητας της νόσου, διαμορφώνοντας

εξατομικευμένες θεραπευτικές προσεγγίσεις και βοηθώντας στην βελτιστοποίηση της διαχείρισης των διαθέσιμων πόρων.

Για την επίτευξη αυτού του στόχου, εφαρμόστηκαν οι αλγόριθμοι Random Forest (RF) και Support Vector Machines (SVM) σε ένα σύνολο δεδομένων που αποτελούνταν από τα κοινωνικά και δημογραφικά χαρακτηριστικά των ασθενών, καθώς και τις κλινικές πληροφορίες και τα εργαστηριακά δεδομένα που αφορούν την φυματίωση. Επιπρόσθετα, εξετάστηκε η συνεισφορά της τεχνικής Synthetic Minority Over-sampling Technique (SMOTE) για την αντιμετώπιση της ανισορροπίας των κλάσεων, καθώς και της τεχνικής Information Gain Attribute Evaluation για την εύρεση των βέλτιστων χαρακτηριστικών. Επιπλέον, διενεργήθηκε Supplied test set, χρησιμοποιώντας νέα ανεξάρτητα δεδομένα για την αξιολόγηση της ικανότητας γενίκευσης του μοντέλου που επιλέχθηκε ως πιο αποδοτικό.

Το πιο αποδοτικό μοντέλο ήταν αυτό του SVM, με τη χρήση της τεχνικής SMOTE και την επιλογή της πολυωνυμικής συνάρτησης πυρήνα, χωρίς την εφαρμογή της τεχνικής Information Gain Attribute Evaluation. Το μοντέλο αυτό ταξινόμησε σωστά το 98.21% των δειγμάτων. Για την κλάση του θανάτου, στην οποία επικεντρώνεται η εργασία, το μοντέλο σημείωσε TPR ή Recall 0,858, FPR 0,009, Precision 0,867 και F-Measure 0,862. Για την κλάση της ίασης, το μοντέλο σημείωσε TPR ή Recall 0,991, FPR 0,142, Precision 0,990 και F-Measure 0,990. Συνολικά, ο σταθμισμένος μέσος όρος (weighted average), ο οποίος λαμβάνει υπόψη τον αριθμό των δειγμάτων σε κάθε κλάση, εμφάνισε TPR ή Recall 0,982, FPR 0,134, Precision 0,982 και F-Measure 0,982. Το Supplied test είχε ως αποτέλεσμα τη μείωση της απόδοσης του μοντέλου στην κλάση του θανάτου, αλλά την αύξηση της απόδοσης στην κλάση της ίασης, διατηρώντας τη συνολική απόδοση του μοντέλου ιδιαίτερα ικανοποιητική για τον στόχο της μελέτης.

Συμπερασματικά, τα αποτελέσματα δείχνουν ότι το μοντέλο SVM, σε συνδυασμό με την τεχνική SMOTE και την πολυωνυμική συνάρτηση πυρήνα, αποτελεί την πιο αποτελεσματική προσέγγιση για την πρόβλεψη της έκβασης της θεραπείας της φυματίωσης. Αυτό το μοντέλο θα μπορούσε να διαδραματίσει καθοριστικό ρόλο στη βελτίωση των αποτελεσμάτων της θεραπείας και στον μετριασμό της θνησιμότητας που σχετίζεται με τη φυματίωση, προωθώντας έτσι τις εξατομικευμένες θεραπευτικές στρατηγικές και τη βέλτιστη διαχείριση των διαθέσιμων πόρων.

Λέξεις Κλειδιά: Μηχανική μάθηση, Κατηγοριοποίηση, Ανισορροπία κλάσεων, Επιλογή χαρακτηριστικών, Μηχανές Διανυσμάτων Υποστήριξης, Τυχαίο Δάσος, Φυματίωση, Πρόβλεψη έκβασης θεραπείας

NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

FACULTY OF NURSING

INTERUNIVERSITY POSTGRADUATE PROGRAM IN HEALTH CARE MANAGEMENT AND HEALTH CARE INFORMATICS

DISSERTATION

APPLYING MACHINE LEARNING TECHNIQUES TO PREDICT THE OUTCOME OF TUBERCULOSIS TREATMENT

BY EIRINI ARGYROPOULOU

SUMMARY

Tuberculosis (TB) is a highly contagious disease and remains one of the leading causes of death from infectious diseases worldwide. Although significant progress has been made in the prevention, diagnosis, and treatment of the disease, reported cases remain high, particularly in areas with limited access to healthcare. The rapid advancement of Artificial Intelligence (AI) holds the potential to serve as a cornerstone in the development and application of new therapeutic approaches through Machine Learning (ML) models.

This study aims to develop an efficient ML model that predicts the outcome of the treatment of TB patients. Such a model could contribute to increasing treatment success rates and, consequently, reducing TB mortality rates by tailoring personalized therapeutic approaches and optimizing the management of available resources.

In pursuit of this objective, the Random Forest (RF) and Support Vector Machines (SVM) algorithms were applied to a data set comprising the social

and demographic characteristics of patients, as well as clinical and laboratory information related to TB. In addition, the Synthetic Minority Over-sampling Technique (SMOTE) was examined for its contribution to addressing class imbalance, along with the Information Gain Attribute Evaluation technique for identifying the optimal features. Moreover, a Supplied test set was conducted using new independent data to assess the generalizability of the model.

The most efficient model was the SVM one, using the SMOTE technique and the polynomial kernel function, without the Information Gain Attribute Evaluation technique. This model correctly classified 98.21% of the samples. For the class of death, which is the primary concern of this study, the model achieved a TPR or Recall of 0.858, an FPR of 0.009, a Precision of 0.867, and an F-measure of 0.862. For the recovery class, the model achieved a TPR or Recall of 0.991, an FPR of 0.142, a Precision of 0.990, and an F-measure of 0.990. Overall, the weighted average, which considers the sample distribution across classes, indicated a TPR or Recall of 0.982, an FPR of 0.134, a Precision of 0.982, and an F-measure of 0.982. The supplied test resulted in a decrease in model performance in the death class, while showing an increase in performance in the recovery class. Nevertheless, the overall performance of the model remained highly satisfactory for the objective of this paper.

To conclude, the results indicate that the SVM model, combined with the SMOTE technique and using the polynomial kernel function, represents the most efficient approach for predicting the outcome of TB treatment. This model could play a pivotal role in improving treatment outcomes and mitigating TB-associated mortality, thus advancing personalized therapeutic strategies and resource management.

Keywords: Machine learning, Classification, Class-imbalanced data, Feature selection, Support Vector Machines, Random Forest, Tuberculosis, Treatment outcome prediction

Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο του Διιδρυματικού Προγράμματος Μεταπτυχιακών Σπουδών «Οργάνωση και Διοίκηση Υπηρεσιών Υγείας – Πληροφορική της Υγείας» στην ειδίκευση «Πληροφορική της Υγείας» του Τμήματος Νοσηλευτικής, Σχολής Επιστημών Υγείας, του Εθνικού Καποδιστριακού Πανεπιστημίου Αθηνών.

Αρχικά, θα ήθελα να εκφράσω τις ευχαριστίες και την ευγνωμοσύνη μου στον καθηγητή Δρ. Ιωσήφ Λιάσκο για την εμπιστοσύνη, την βοήθεια και τις πολύτιμες συμβουλές που μου παρείχε κατά τη διάρκεια της συγγραφής της διπλωματικής μου εργασίας, εφόδια τα οποία αποτέλεσαν τη βάση για την επίτευξη του στόχου μου.

Επιπρόσθετα, θα ήθελα να εκφράσω την ευγνωμοσύνη μου στο διδακτικό προσωπικό του μεταπτυχιακού προγράμματος για τις πολύτιμες γνώσεις που αποκόμισα κατά τη διάρκεια της φοίτησης μου, οι οποίες συνέβαλαν στην περάτωση της διπλωματικής μου εργασίας και στην διεύρυνση των γνώσεων μου.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια, τους συναδέλφους και το φιλικό περιβάλλον μου για την αμέριστη υποστήριξη και κατανόηση κατά τη διάρκεια αυτής της πορείας.

Πίνακας περιεχομένων

Κατάλογος εικόνων

Κατάλογος πινάκων

ΓΕΝΙΚΟ ΜΕΡΟΣ

1.Φυματίωση

1.1 Εισαγωγή

Η φυματίωση είναι μια θανατηφόρα λοιμώδης νόσος, ιδιαίτερα μεταδοτική και προκαλείται έπειτα από βακτηριακή λοίμωξη. Συγκεκριμένα, οι αιτιολογικοί παράγοντες της προέρχονται από το σύμπλεγμα του **Μυκοβακτηριδίου της φυματίωσης** (M.tuberculosis complex, MTBC). Η μετάδοση της γίνεται αερογενώς από άνθρωπο σε άνθρωπο. Κάθε χρόνο εκατομμύρια άνθρωποι προσβάλλονται παγκοσμίως, ωστόσο είναι μια ιάσιμη νόσος και μπορεί να προληφθεί. Η συνήθης εντόπιση της φυματίωσης γίνεται στους πνεύμονες, αλλά μπορεί να προσβάλει και άλλα όργανα ή κοιλότητες του σώματος.(1) Η έγκαιρη διάγνωση της είναι ζωτικής σημασίας, καθώς έτσι όχι μόνο αυξάνονται οι πιθανότητες ίασης του ασθενούς αλλά περιορίζεται και η μετάδοση της. Επίσης, η σωστή διαχείριση των πόρων και η κατάλληλη αντιμετώπιση των ασθενών που πάσχουν από φυματίωση εμφανίζει σημαντικό αντίκτυπο στην έκβαση της θεραπείας και συνεπώς στον περιορισμό της νόσου.

1.2 Ιστορική αναδρομή

Το γένος Mycobacterium πιστεύεται ότι εμφανίστηκε πριν από περισσότερα από 150 εκατομμύρια χρόνια. Από το 2.400 π.Χ. είχαν παρατηρηθεί χαρακτηριστικές αλλοιώσεις φυματίωσης σε οστά που προέρχονταν από αιγυπτιακές μούμιες.(2) Η φυματίωση κατά την αρχαία Ελλάδα ήταν γνωστή ως μια θανατηφόρα ασθένεια και την αποκαλούσαν “Φθίση”. Ο Ιπποκράτης (460-370 π.Χ.) κατάφερε να καταγράψει με μεγάλη ακρίβεια την κλινική εικόνα της φυματίωσης και τις αλλοιώσεις που προκαλεί στο αναπνευστικό σύστημα και παρατήρησε πως είναι ιδιαίτερα θανατηφόρα για τους νέους ενήλικες. Ο Ισοκράτης (436-338 π.Χ.) ήταν ο πρώτος που έκανε την υπόθεση πως η φυματίωση είναι πιθανά μεταδοτική νόσος.(3)

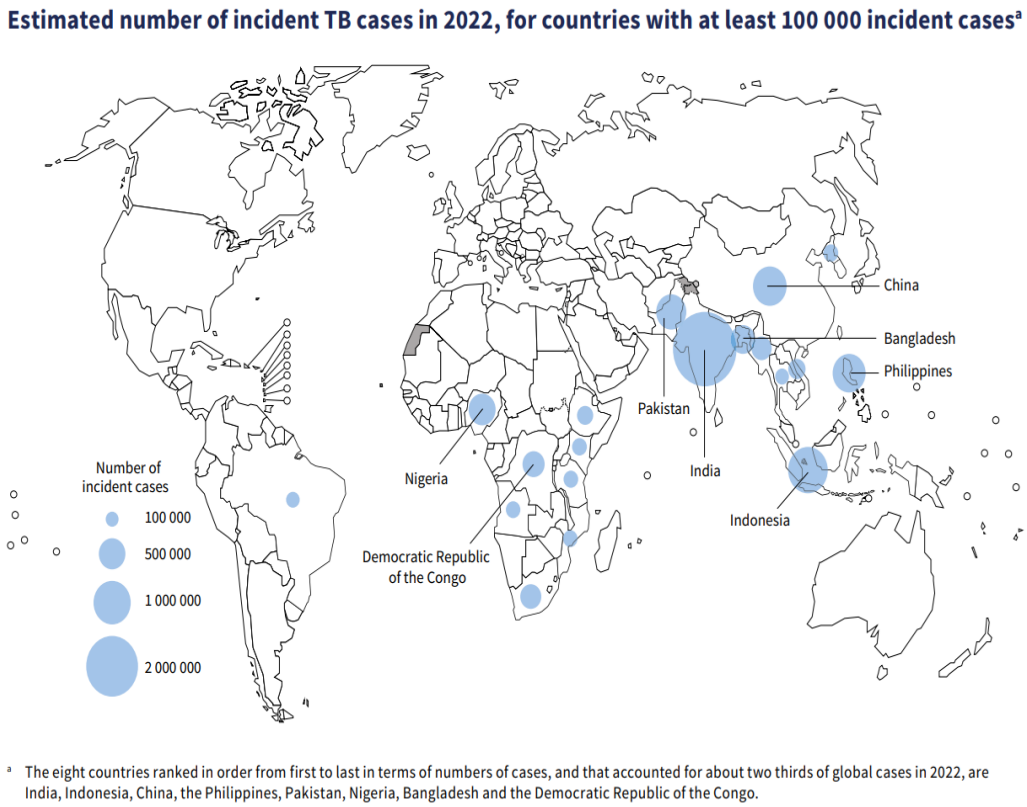
Η απομόνωση του βακίλου της φυματίωσης και η καλλιέργεια του το 1882 από τον Robert Koch (Εικόνα 1) αποτέλεσε ορόσημο στην ιστορία της ασθένειας.(4) Εντός των επόμενων δεκαετιών αναπτύχθηκε η δερμοαντίδραση Mantoux, το αντιφυματικό εμβόλιο BCG καθώς και κάποια αντιφυματικά φάρμακα όπως η στρεπτομυκίνη.(5)(6) Μέχρι και σήμερα, η φυματίωση αποτελεί σημαντική αιτία αύξησης της θνησιμότητας σε παγκόσμιο επίπεδο.

Εικόνα 1: Robert Koch

[Πηγή: Robert Koch: Fundador de la bacteriología - Hidden Nature (hidden-nature.com)]

1.3 Επιδημιολογία

Κάθε χρόνο περισσότερα από 10 εκατομμύρια άνθρωποι διεθνώς νοσούν από φυματίωση. Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας (ΠΟΥ), περίπου το ¼ του παγκόσμιου πληθυσμού έχει μολυνθεί από το Μυκοβακτηρίδιο της φυματίωσης. Ωστόσο, το μεγαλύτερο ποσοστό δεν θα αναπτύξει την νόσο της φυματίωσης. Το έτος 2019 καταγράφηκαν 7.1 εκατομμύρια νέες διαγνώσεις, ενώ το 2020 μόλις 5.8 εκατομμύρια. Αυτή η μείωση οφείλεται στην πανδημία της νόσου COVID-19, η οποία είχε αρνητικό αντίκτυπο όχι μόνο στη διάγνωση αλλά και στη θεραπεία της φυματίωσης, με αποτέλεσμα την επιβράδυνση του περιορισμού της νόσου. Έπειτα, το 2021 καταγράφηκαν 6.4 εκατομμύρια κρούσματα. Μεταξύ του έτους 2020 και 2021 σημειώθηκε 3,6% αύξηση στο ποσοστό επίπτωσης (νέα περιστατικά ανά 100.000 κατοίκους σε ένα χρόνο), με την αύξηση αυτή να υποδηλώνει μερική ανάκαμψη στην διάγνωση των κρουσμάτων της φυματίωσης.(7) Το έτος 2022 η φυματίωση συνέχισε να αποτελεί την δεύτερη κύρια αιτία θανάτου παγκοσμίως, μετά την νόσο COVID-19, με περίπου 1,3 εκατομμύρια θανάτους. Επιπλέον, η εκτίμηση των νέων κρουσμάτων φυματίωσης (incident cases) σε παγκόσμιο επίπεδο για το έτος 2022 ανέρχονταν σε 10,6 εκατομμύρια (Εικόνα 2), ενώ τα κρούσματα που καταγράφηκαν επίσημα και διαγνώστηκαν (notified cases) σε 7,5 εκατομμύρια. Το ίδιο έτος, το 55% των ατόμων που ανέπτυξαν την νόσο ήταν άνδρες, το 33% γυναίκες και το 12% παιδιά ηλικίας 0 έως 12 ετών.(8)



Εικόνα 2: Εκτιμώμενος αριθμός περιστατικών φυματίωσης το 2022, για χώρες με τουλάχιστον 100.000 περιστατικά.

[Πηγή: World Health Organization. Global tuberculosis report 2023. Geneva: World Health Organization; 2023. Licence: CC BY-NC-SA 3.0 IGO.]

Οι οκτώ χώρες στις οποίες διαγνώστηκαν τα περισσότερα κρούσματα παγκοσμίως το 2022 είναι η Ινδία, η Ινδονησία, η Κίνα, οι Φιλιππίνες, το Πακιστάν, η Νιγηρία, το Μπαγκλαντές και η Λαϊκή Δημοκρατία του Κονγκό. Η Ινδία ηγείται μεταξύ των χωρών αυτών και καταλαμβάνει το 27% των κρουσμάτων φυματίωσης παγκοσμίως, ακολουθούμενη από την Ινδονησία με 10%, την Κίνα με 7,1%, τις Φιλιππίνες με 7,0%, το Πακιστάν με 5,7%, τη Νιγηρία με 4,5%, το Μπαγκλαντές με 3,6% και τέλος τη Λαϊκή Δημοκρατία του Κονγκό με 3,0%.(8)

1.4 Χαρακτηριστικά του Μυκοβακτηριδίου της φυματίωσης

Το σύμπλεγμα του Μυκοβακτηριδίου της φυματίωσης (M.tuberculosis complex, MTBC) αναφέρεται σε μια ομάδα βακτηριδίων που αποτελείται από το Μυκοβακτηρίδιο της φυματίωσης (Mycobacterium tuberculosis, M.tuberculosis) και άλλα γενετικά συγγενικά του είδη όπως για παράδειγμα τα

Mycobacterium africanum, Mycobacterium bovis, Mycobacterium microti και Mycobacterium caprae. Όσον αφορά τα χαρακτηριστικά του, το Μυκοβακτηρίδιο της φυματίωσης είναι ένα ακίνητο, μη σπορογόνο, υποχρεωτικά αερόβιο, καταλάση αρνητικό ενδοκυττάριο βακτήριο και συγκεκριμένα λόγω της μορφολογίας του ανήκει στην κατηγορία των βακίλων. Είναι βραδέως αναπτυσσόμενο καθώς ο χρόνος αναδιπλασιασμού του ανέρχεται σε 12-24 ώρες όταν οι συνθήκες είναι βέλτιστες. Οι συμβατικές εργαστηριακές χρώσεις που χρησιμοποιούνται στην ρουτίνα των μικροβιολογικών εργαστηρίων, δηλαδή η χρώση Gram και η χρώση Giemsa δεν μπορούν να συμβάλουν στην εργαστηριακή διάγνωση του Μυκοβακτηριδίου της φυματίωσης. Αυτό οφείλεται στο γεγονός ότι η σύνθεση του κυτταρικού τοιχώματος τους είναι ιδιαίτερη, καθώς αυτό είναι πλούσιο σε λιπίδια όπως για παράδειγμα τα μυκολικά οξέα τα οποία προσδίδουν υδρόφοβες ιδιότητες στο κυτταρικό τοίχωμα με αποτέλεσμα να εμποδίζουν τις χρωστικές να διεισδύσουν σε αυτό. (1)(9)

1.5 Μετάδοση

Η μετάδοση των βακτηριδίων της φυματίωσης γίνεται αερογενώς από άτομο με ενεργό φυματίωση σε υγειές άτομο. Πιο συγκεκριμένα, τα αερολύματα που αποβάλλονται με τον βήχα, το γέλιο και το φτέρνισμα, οδηγούν στην εισπνοή των μικρών σωματιδίων, που εμπεριέχουν τα βακτηρίδια, από τα υγιή άτομα. Τα βακτηρίδια θα έρθουν σε επαφή με το βλεννογόνο του αναπνευστικού συστήματος στα βρογχικά δέντρα των πνευμόνων. Το αναπνευστικό σύστημα θα προσπαθήσει να εμποδίσει την διέλευση των μυκοβακτηριδίων μέσω διαφόρων παραγόντων όπως η βλέννα, τα πεπτίδια κατά των μυκοβακτηριδίων, οι ανοσοσφαιρινές, οι χημειοκίνες και οι κυτταροκίνες. (10)

1.6 Κλινικές εκδηλώσεις

Οι κλινικές εκδηλώσεις που μπορούν να προκύψουν από τη λοίμωξη του Μυκοβακτηριδίου της φυματίωσης ανήκουν σε ένα ευρύ φάσμα και μπορούν να κυμανθούν από υποκλινική λοίμωξη έως ήπια, μέτρια αλλά και σοβαρή ενεργό λοίμωξη. Το Μυκοβακτηρίδιο της φυματίωσης μπορεί να παραμείνει αδρανές στον ανθρώπινο οργανισμό, χωρίς να προκαλέσει κλινικές εκδηλώσεις για αρκετά χρόνια δημιουργώντας έτσι τους ασυμπτωματικούς φορείς της λανθάνουσας φυματίωσης, οι οποίοι ωστόσο μπορούν να εκδηλώσουν μελλοντικά ενεργό φυματίωση. Ανάλογα με το σύστημα που προσβάλλεται, η φυματίωση διακρίνεται σε πνευμονική (pulmonary

tuberculosis) σε ποσοστό περίπου 80 τοις εκατό των περιπτώσεων και εξωπνευμονική (extrapulmonary tuberculosis) σε ποσοστό περίπου 20 τοις εκατό. Η εντόπιση της πνευμονικής φυματίωσης γίνεται στους πνεύμονες και εμφανίζει διάφορες κλινικές εκδηλώσεις με κυριότερη τον χρόνιο βήχα. Επιπρόσθετα, οι ασθενείς πιθανά θα εμφανίσουν πόνο στο στήθος, πυρετό, αιμόπτυση και παραγωγή πτυέλων, νυχτερινές εφιδρώσεις και απώλεια όρεξης που οδηγεί στην απώλεια κιλών. Η εξωπνευμονική φυματίωση μπορεί να προσβάλει διαφορά όργανα γειτονικά ή και απομακρυσμένα από την εστία της πρωτογενούς μόλυνσης, και συνεπώς υπάρχουν οι ανάλογες κλινικές εκδηλώσεις. Οι εστίες αυτές δημιουργούνται είτε μέσω του λεμφικού συστήματος είτε αιματογενώς. Η εξωπνευμονική φυματίωση μπορεί να εντοπιστεί στο Κεντρικό Νευρικό Σύστημα (ΚΝΣ), στο δέρμα, στις αρθρώσεις και τα οστά, στο ουρογεννητικό και γαστρεντερικό σύστημα, αλλά και ως φυματιώδης λεμφαδενίτιδα, πλευρίτιδα, περικαρδίτιδα και περιτονίτιδα. (10)(11)

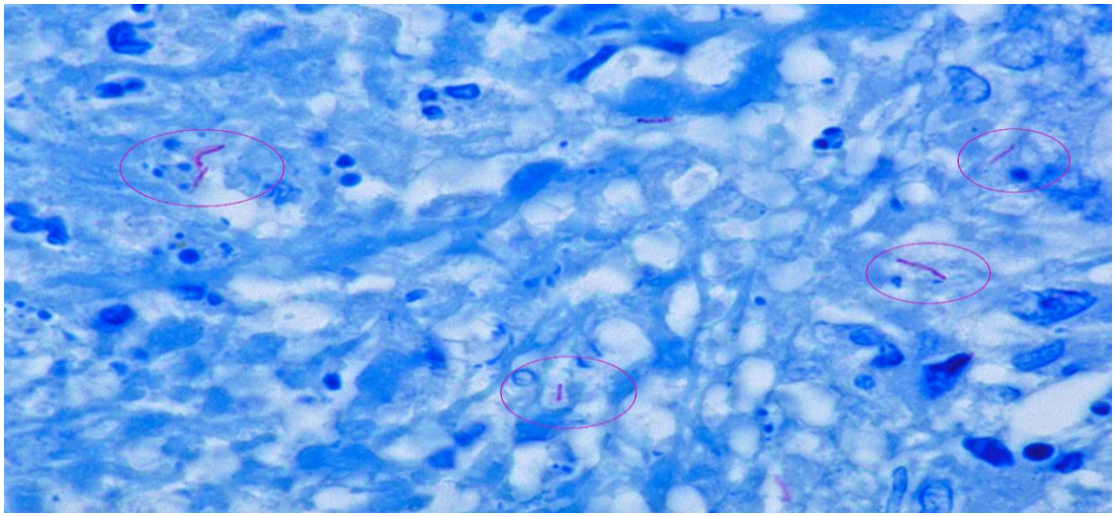
1.7 Διάγνωση

Η έγκαιρη διάγνωση της φυματίωσης είναι ιδιαίτερα σημαντική για την έκβαση της θεραπείας της νόσου, ωστόσο δεν είναι πάντα εύκολη. Για τη διάγνωση του Μυκοβακτηριδίου της φυματίωσης μπορούν να ελεγχθούν εργαστηριακά διάφορα είδη βιολογικών δειγμάτων, ανάλογα με την εκδήλωση των κλινικών συμπτωμάτων. Από το αναπνευστικό σύστημα μπορεί να γίνει δειγματοληψία πτυέλων, βρογχικών εκκρίσεων και βρογχοκυψελιδικού εκπλύματος για τη διάγνωση της πνευμονικής φυματίωσης. Για την διαπίστωση εξωπνευμονικής φυματίωσης μπορούν να ελεγχθούν υγρά από τραύματα και αποστήματα, αίμα, μυελός των οστών, ούρα, κόπρανα, ιστοτεμάχια, γαστρικό υγρό και υγρά παρακεντήσεων όπως πλευριτικό, περιτοναϊκό, περικαρδιακό και εγκεφαλονωτιαίο υγρό.

Μέθοδος αναφοράς (gold standard) για την εργαστηριακή διάγνωση του Μυκοβακτηριδίου της φυματίωσης αποτελεί η απομόνωση του μέσω καλλιέργειας. Παρ’ όλα αυτά, η μικροσκόπιση είναι η πιο συνηθισμένη μέθοδος και μπορεί να εφαρμοστεί σχεδόν σε κάθε εργαστηριακή υποδομή λόγω της ευκολίας και της ταχύτητας της. Επιπρόσθετα, εφικτή είναι πλέον και η ταυτοποίηση του γενετικού υλικού του M.Tuberculosis με τη χρήση μοριακών μεθόδων.(12)

1.7.1 Μικροσκόπηση

Η συνήθης χρώση που χρησιμοποιείται ευρέως για την διάγνωση της φυματίωσης μέσω μικροσκόπησης είναι η Ziehl-Neelsen. Είναι μια γρήγορη και οικονομική μέθοδος, ιδιότητες που την καθιστούν ιδανική ιδιαίτερα σε χώρες με περιορισμένους πόρους, όπου μπορεί να είναι και η μοναδική μέθοδος που διαθέτουν οι φορείς δημόσιας υγείας. Κατά τη χρώση αυτή, χρησιμοποιείται η χρωστική καρβολική φουξίνη. Έπειτα, γίνεται αποχρωματισμός με οξινισμένη αλκοόλη και στο τέλος χρησιμοποιείται η χρωστική μπλε του μεθυλενίου. Εξαιτίας της υψηλής συγκέντρωσης λιπιδίων στο κυτταρικό τοίχωμα, τα Μυκοβακτηρίδια δεν αποχρωματίζονται με την οξινισμένη αλκοόλη, οπότε διατηρούν το χρώμα από την φουξίνη (Εικόνα 3). Η ευαισθησία της μικροσκοπικής μεθόδου κυμαίνεται από 20-80% σύμφωνα με διάφορες μελέτες, παράγοντας που αποτελεί τον σημαντικότερο περιορισμό αυτής της μεθόδου.(13)(14)



Εικόνα 3: Μικροσκοπική εικόνα του Μυκοβακτηριδίου της φυματίωσης κατόπιν χρώσης Ziehl-Neelsen.

[Πηγή:https://commons.wikimedia.org/wiki/File:Mycobacterium_tuberculosis_Ziehl-Neelsen_stain.jpg]

1.7.2 Καλλιέργεια

Ο ΠΟΥ συνιστά την καλλιέργεια ως μέθοδο αναφοράς για την διάγνωση του Μυκοβακτηριδίου της φυματίωσης, η οποία συμβάλλει στην απομόνωση του και εν συνεχεία στην διάκριση των ανθεκτικών στα αντιβιοτικά στελεχών. Η καλλιέργεια μπορεί να πραγματοποιηθεί σε στερεά θρεπτικά υλικά όπως για παράδειγμα το Lowenstein-Jensen (Εικόνα 4) και σε υγρά θρεπτικά υλικά όπως το Middlebrook 7H9. Τα βιολογικά δείγματα που θα εξεταστούν και δεν είναι στείρα μικροβίων θα πρέπει να υποστούν ειδική επεξεργασία έτσι ώστε να εξαλειφθεί η φυσιολογική χλωρίδα τους, που αποτελείται από διάφορα μικρόβια, για την αποφυγή επιμόλυνσης της καλλιέργειας. Η καλλιέργεια είναι

μια ιδιαίτερα ευαίσθητη μέθοδος με το όριο ανίχνευσης να υπολογίζεται από 10 έως 100 ζωντανά Μυκοβακτηρίδια ανά ml. Ωστόσο, η καλλιέργεια υστερεί έναντι των άλλων μεθόδων λόγω της μεγάλης περιόδου επώασης (περίπου 2-8 εβδομάδες) που απαιτείται για την ανάπτυξη των μυκοβακτηριδίων.(13)(15)

Εικόνα 4: Θετική καλλιέργεια φυματίωσης σε στερεό θρεπτικό υλικό Lowenstein-Jensen.

[Πηγή: https://clinicare.gr/product/solinaria-lowenstein-jensen-medium/]

Ευρέως διαδεδομένο στην εργαστηριακή ρουτίνα είναι το αυτοματοποιημένο σύστημα BACTEC MGIT 960. Σε αυτό εισάγονται τα σωληνάρια της καλλιέργειας με υγρό θρεπτικό υλικό και έπειτα το σύστημα ανιχνεύει την κατανάλωση οξυγόνου από τα Μυκοβακτηρίδια διενεργώντας συνεχείς μετρήσεις.(16)

1.7.3 Μοριακές μέθοδοι

Τα τελευταία χρόνια, οι μοριακές μέθοδοι χρησιμοποιούνται όλο και περισσότερο στον τομέα της διάγνωσης. Η PCR (Polymerase Chain Reaction) είναι εξαιρετικά αποτελεσματική για την ταχύτατη διάγνωση όχι μόνο του Μυκοβακτηριδίου της φυματίωσης αλλά και της αντίστασης σε ορισμένα αντιβιοτικά. Ωστόσο, το μειονέκτημα της είναι το αυξημένο κόστος που απαιτείται σε σχέση με την μικροσκόπηση και την καλλιέργεια. (17)

Ευρύτατα χρησιμοποιούμενη σε παγκόσμιο επίπεδο για την ανίχνευση του Μυκοβακτηριδίου της φυματίωσης είναι η αυτοματοποιημένη μοριακή εξέταση Xpert MTB/RIF Ultra (Cepheid). Το 2017 ο ΠΟΥ συνέστησε την χρήση του

Xpert MTB/RIF Ultra (Εικόνα 5) ως αρχική διαγνωστική εξέταση για τη διάγνωση της φυματίωσης σε ενήλικες και παιδιά, όταν αυτό είναι εφικτό.(18)

Εικόνα 5: Αυτοματοποιημένη μοριακή εξέταση Xpert MTB/RIF Ultra (Cepheid).

[Πηγή: https://commons.wikimedia.org/wiki/File:GeneXpert_lightbox.jpg]

Το Xpert MTB/RIF Ultra ανιχνεύει τις αλληλουχίες των γονιδίων IS6110 και IS1081 προκειμένου να δώσει θετικό αποτέλεσμα για την ύπαρξη του γενετικού υλικού του Μυκοβακτηριδίου της φυματίωσης σε δείγματα αναπνευστικού. Επιπλέον, στοχεύει κυρίως σε συγκεκριμένες περιοχές του γονιδίου rpoB στο γονιδίωμα του Μυκοβακτηριδίου της φυματίωσης για την ανίχνευση μεταλλάξεων, για να διαγνώσει την αντοχή στο αντιβιοτικό ριφαμπικίνη. (19) Η μέθοδος αυτή καθίσταται εξαιρετικά ισχυρή καθώς δίνει αποτελέσματα εντός δυο ωρών, με ελάχιστη παρέμβαση στη διαδικασία από το εργαστηριακό προσωπικό. (20)

Η ευαισθησία της μεθόδου ανέρχεται σε ποσοστό 88% ενώ η ειδικότητα 96% σύμφωνα με τον μέσο όρο που προέκυψε από διάφορες μελέτες.(21)

1.7.4 Απεικονιστικές μέθοδοι

Οι απεικονιστικές τεχνικές παίζουν αρκετά καθοριστικό ρόλο στην διάγνωση της φυματίωσης. Συμβάλλουν στην πρωτογενή διάγνωση της φυματίωσης μέσω της ανίχνευσης ιστικών αλλαγών. Αν και οι ακτινογραφίες θώρακος παραμένουν η πιο συχνά χρησιμοποιούμενη απεικονιστική τεχνική για την

απεικόνιση της πνευμονικής φυματίωσης, η υπολογιστική ή αξονική τομογραφία (Computed Tomography, CT), η μαγνητική τομογραφία (Magnetic Resonance Imaging, MRI) και η ποζιτρονική τομογραφία σε συνδυασμό με την αξονική PET/CT, έχει αποδειχθεί ότι είναι εξαιρετικά χρήσιμες στην αξιολόγηση τόσο της πνευμονικής όσο και της εξωπνευμονικής φυματίωσης. Σε περιπτώσεις οπού έχει γίνει ήδη η διάγνωση της φυματίωσης οι απεικονιστικές τεχνικές μπορούν να χρησιμοποιηθούν για την εκτίμηση της έκτασης της νόσου, την αξιολόγηση της ανταπόκρισης στη θεραπεία ή την ανίχνευση λοίμωξης που παραμένει μετά την ολοκλήρωση της θεραπείας. Η αξονική τομογραφία λόγω της υψηλής ταχύτητας και της καλής ανάλυσης που προσφέρει, αποτελεί τη μέθοδο επιλογής για την αξιολόγηση της θωρακικής, γαστρεντερικής και ουρογεννητικής φυματίωσης. Ωστόσο, η νεφροτοξικότητα των σκιαγραφικών ουσιών που χρησιμοποιούνται για την αξονική τομογραφία αποτελεί μειονέκτημα, ιδίως σε ασθενείς με μειωμένη νεφρική λειτουργία. Επίσης, η δυνατότητα πολυεπίπεδης απεικόνισης και η βέλτιστη ανάλυση αντίθεσης μαλακών ιστών καθιστούν τη μαγνητική τομογραφία τη μέθοδο επιλογής για την αξιολόγηση της φυματίωσης του Κεντρικού Νευρικού Συστήματος (ΚΝΣ).(22)

Η πρωτογενής φυματίωση που αφορά άτομα εκτεθειμένα στο Μυκοβακτηρίδιο της φυματίωσης για πρώτη φορά, προσβάλλει συνήθως τα παιδιά σε ενδημικές περιοχές και ενήλικες σε μη ενδημικές περιοχές. Η πρωτογενής φυματίωση μπορεί να επηρεάσει οποιοδήποτε τμήμα του πνευμονικού παρεγχύματος, των λεμφαδένων, του τραχειοβρογχικού δέντρου και του υπεζωκότα. Η λεμφαδενοπάθεια είναι το πιο συνηθισμένο ακτινολογικό εύρημα που εμφανίζεται στην πρωτογενή φυματίωση και είναι πιο πιθανό να παρουσιαστεί σε παιδιά παρά σε ενήλικες, προκαλώντας πιθανώς ατελεκτασία των πνευμόνων.(23)

Η μεταπρωτοπαθής φυματίωση αφορά την επανενεργοποίηση ή την επαναμόλυνση από το M.tuberculosis, και παρατηρείται σχεδόν αποκλειστικά σε ενήλικες. Απεικονιστικά χαρακτηρίζεται από απουσία διόγκωσης των λεμφαδένων και εμφανίζεται κυρίως στους άνω λοβούς των πνευμόνων. Χαρακτηριστική είναι η δημιουργία κοιλοτήτων (σπηλαίωση) στους πνεύμονες οι οποίες σχηματίζονται από τον κατεστραμμένο πνευμονικό ιστό. Επίσης σε ποσοστό 6-18% μπορεί να ανευρεθεί πλευριτική συλλογή.(22)(24)

1.7.5 Φυματινοαντίδραση Mantoux

Η διαγνωστική μέθοδος Mantoux είναι μια δερματική δοκιμασία η οποία χρησιμοποιείται για να τον έλεγχο της έκθεσης στο M.tuberculosis. Είναι η συνήθης μέθοδος διάγνωσης για την λανθάνουσα λοίμωξη από φυματίωση. Ως λανθάνουσα λοίμωξη χαρακτηρίζεται η ύπαρξη εμμένουσας ανοσολογικής απόκρισης λόγω διέγερσης από έκθεση στα αντιγόνα του M.tuberculosis, χωρίς να υπάρχουν κλινικές εκδηλώσεις ενεργού φυματίωσης.(25) Τα πλεονεκτήματα της μεθόδου είναι το χαμηλό κόστος της και η σχετικά απλή

εφαρμογή της.(26) Με την Mantoux ελέγχεται η ευαισθησία του δέρματος στην φυματίνη (tuberculin) που είναι το πρωτεϊνικό τμήμα του βακίλου της φυματίωσης. Υπάρχουν δύο κύριοι τύποι φυματίνης, η παλαιά φυματίνη (old tuberculin, OT) και η κεκαθαρμένη φυματίνη (purified protein derivative, PPD). Πραγματοποιείται ενδοδερμική έγχυση 0,1 mL PPD που περιέχει 5 TU (μονάδες φυματίνης) στην οσφυϊκή επιφάνεια του αντιβραχίου (Εικόνα 6). Προκαλείται ερυθρότητα στο σημείο η οποία ωστόσο δεν έχει διαγνωστική αξία. Η δοκιμασία ερμηνεύεται μετά από 48-72 ώρες, με τη μέτρηση της εγκάρσιας διαμέτρου της ψηλαφητής σκληρίας.(27)

Εικόνα 6: Ενδοδερμική έγχυση φυματίνης στην οσφυϊκή επιφάνεια του αντιβραχίου.

[Πηγή: https://globalhealthnow.org/object/tb-skin-test]

Μια αντίδραση 5 mm ή μεγαλύτερη θεωρείται θετική όταν υπάρχει στενή επαφή με κρούσμα φυματίωσης, όταν το άτομο είναι ανοσοκατεσταλμένο, ιδίως σε άτομα με HIV λοίμωξη και σε άτομα με κλινικές ή ακτινογραφικές ενδείξεις ενεργού ή παρελθοντικής λοίμωξης. Μια αντίδραση ≥10 mm θεωρείται θετική σε άτομα με αυξημένο κίνδυνο λανθάνουσας φυματίωσης όπως για παράδειγμα άτομα που έχουν γεννηθεί σε χώρες με υψηλή επίπτωση φυματίωσης και άτομα με υψηλό κίνδυνο έκθεσης σε αυτήν λόγω επαγγέλματος. Μια αντίδραση 15 mm ή μεγαλύτερη θεωρείται θετική για οποιοδήποτε άτομο.(27) Η θετική φυματινοαντίδραση Mantoux έχει χαμηλή ειδικότητα και δεν μπορεί να διακρίνει αν το θετικό αποτέλεσμα οφείλεται στην μόλυνση από M.tuberculosis, σε εμβολιασμό με BCG (Bacillus Calmette-Guérin) ή σε μόλυνση ή έκθεση σε άτυπα Μυκοβακτηρίδια, καθώς τα αντιγόνα που θα εγχυθούν κατά τη διαδικασία του τεστ είναι κοινά για αυτές τις περιπτώσεις. Επίσης, έχει χαμηλή ευαισθησία σε ανοσοκατεσταλμένα άτομα, όπως για παράδειγμα άτομα που ζουν με HIV.(28)

1.8 Παράγοντες κινδύνου

Ορισμένοι παράγοντες και νοσήματα συσχετίζονται με την εμφάνιση, εξέλιξη και πορεία της θεραπείας της φυματίωσης. Οι πιο σημαντικοί παράγοντες κινδύνου είναι ο σακχαρώδης διαβήτης, το κάπνισμα, η συστηματική κατανάλωση αλκοόλ, ο υποσιτισμός, η χρήση ενδοφλέβιων ουσιών, η ανοσοκαταστολή και ιδιαίτερα η νόσηση από τον ιό HIV, κοινωνικοοικονομικοί παράγοντες και δημογραφικά χαρακτηριστικά όπως η ηλικία.(29)

1.8.1 Σακχαρώδης διαβήτης

Οι ασθενείς με σακχαρώδη διαβήτη διατρέχουν υψηλότερο κίνδυνο να εμφανίσουν ενεργό λοίμωξη όταν μολυνθούν από το Μυκοβακτηρίδιο της φυματίωσης. Σύμφωνα με διάφορες μελέτες, ο λόγος πιθανοτήτων ανάπτυξης φυματίωσης είναι 2,44 έως 8,33 φορές υψηλότερος σε διαβητικούς ασθενείς σε σχέση με τους μη διαβητικούς.(30) Σε συστηματική ανασκόπηση που περιλάμβανε 13 μελέτες διαπιστώθηκε ότι ο σακχαρώδης διαβήτης σχετίζεται με τριπλάσιο κίνδυνο ανάπτυξης της φυματίωσης (RR = 3.11, 95% CI: 2.27-4.26).(31) Σε άλλη μελέτη βρέθηκε ότι μετά από το πέρας της θεραπείας 6 μηνών, οι ασθενείς με φυματίωση και σακχαρώδη διαβήτη είχαν θετικό αποτέλεσμα στην μικροσκοπική εξέταση από καλλιέργεια πτυέλων σε ποσοστό 22,2% ενώ οι ασθενείς χωρίς σακχαρώδη διαβήτη σε ποσοστό 6,9%.(32) Για την καλύτερη αντιμετώπιση της φυματίωσης ο ΠΟΥ συνιστά να διενεργείται έλεγχος για τη διάγνωση του σακχαρώδους διαβήτη σε ασθενείς με φυματίωση και το αντίστροφο.(33) Οι ασθενείς με φυματίωση που είναι ταυτόχρονα διαβητικοί, έχουν περισσότερες πιθανότητες να εμφανίσουν ως έκβαση την υποτροπή ή τον θάνατο. Μάλιστα, σύμφωνα με δυο αναδρομικές μελέτες κοόρτης που πραγματοποιήθηκαν, οι ασθενείς με πνευμονική φυματίωση και σακχαρώδη διαβήτη διατρέχουν 6,5-6,7 φορές μεγαλύτερο κίνδυνο να αποβιώσουν σε σχέση με αυτούς που δεν έχουν σακχαρώδη διαβήτη.(34)(35) Σε άλλη μελέτη που πραγματοποιήθηκε στην Αίγυπτο, διαπιστώθηκε ότι η αποτυχία της θεραπευτικής αγωγής για την φυματίωση ήταν 3.9 φορές μεγαλύτερη στους ασθενείς με σακχαρώδη διαβήτη έναντι των μη διαβητικών.(36)

1.8.2 Καπνιστική συνήθεια

Όλο και περισσότερες ενδείξεις προκύπτουν σύμφωνα με μελέτες για την επίδραση του ενεργητικού και παθητικού καπνίσματος στην φυματίωση. Το κάπνισμα μπορεί να επιδράσει ως παράγοντας κινδύνου προκαλώντας λανθάνουσα λοίμωξη, εξέλιξη της πρωτογενούς λοίμωξης σε ενεργό φυματίωση, χαμηλότερα ποσοστά επιτυχίας στην θεραπεία και ως

αποτέλεσμα υψηλότερα ποσοστά θανάτων που σχετίζονται με το Mycobacterium tuberculosis. Στην μετα-ανάλυση που διενεργήθηκε από τον Bates και τους συναδέλφους του, στην οποία χρησιμοποιήθηκαν δεδομένα από 24 μελέτες σχετικά με την επίδραση του καπνίσματος στην φυματίωση, βρέθηκε ότι ο κίνδυνος ανάπτυξης φυματίωσης ήταν 2,3-2,7 φορές υψηλότερος στους καπνιστές σε σχέση με τους μη καπνιστές.(37) Σε μια μετα-ανάλυση η οποία βασίστηκε σε 20 μελέτες και είχε συνολικό δείγμα 47.770 συμμετέχοντες, έδειξε ότι υπάρχει 51% αυξημένη πιθανότητα κακής έκβασης στην θεραπεία της φυματίωσης (OR = 1.51; 95% CI = 1.30-1.75). Έγινε ανάλυση των δεδομένων βάση εισοδήματος σε υποομάδες κ βρέθηκε ότι το κάπνισμα είχε μεγαλύτερη επίδραση στις χώρες με χαμηλά και μεσαία εισοδήματα (OR = 1,74) και στις χώρες ανώτερου μεσαίου εισοδήματος (OR = 1,52) σε σύγκριση με τις χώρες με υψηλά εισοδήματα (OR = 1,34), αν και οι διαφορές δεν ήταν στατιστικά σημαντικές.(38) Σε μελέτη που πραγματοποιήθηκε στο Ιράν, τα αποτελέσματα έδειξαν ότι οι μη καπνιστές (83,4% επιτυχία), καθώς και τα άτομα που διέκοψαν το κάπνισμα στους δύο μήνες (80,8% επιτυχία), είχαν υψηλότερο ποσοστό επιτυχίας στην θεραπεία μετά το πέρας των 6 μηνών, από τα άτομα που συνέχισαν το κάπνισμα (67,6% επιτυχία).(39) Η μετα-ανάλυση που διεξήχθη από τον E. Y. Wang και τους συνεργάτες του είχε ως στόχο να διερευνήσει την επίδραση του καπνίσματος στο αποτέλεσμα της θεραπείας της φυματίωσης. Στην μετα-ανάλυση συμπεριλήφθηκαν 21 άρθρα που αφορούσαν την ενεργό πνευμονική φυματίωση και το κάπνισμα. Από τη μετα-ανάλυση προέκυψε ότι τα άτομα που καπνίζουν έχουν 1,23 φορές περισσότερες πιθανότητες να εμφανίσουν δυσμενείς εκβάσεις στη θεραπεία της φυματίωσης. Επιπρόσθετα, οι καπνιστές εμφανίζουν 1,55 φορές περισσότερες πιθανότητες να παρουσιάσουν καθυστέρηση στη αρνητικοποίηση του μικροσκοπικού επιχρίσματος ή της καλλιέργειας κατά τη διάρκεια της θεραπείας της φυματίωσης. Τέλος, οι καπνίζοντες ασθενείς βρέθηκε ότι έχουν 1,35 φορές περισσότερες πιθανότητες να διακόψουν τη θεραπεία της φυματίωσης πριν από την ολοκλήρωση της.(40)

1.8.3 Κατανάλωση αλκοόλ

Η συστηματική κατανάλωση αλκοόλ σε μεγάλες ποσότητες φαίνεται να είναι ένας ακόμη παράγοντας που αυξάνει τον κίνδυνο εκδήλωσης της φυματίωσης και δυσχεραίνει την έκβαση της θεραπευτικής αγωγής. Οι ασθενείς με φυματίωση που καταναλώνουν αλκοόλ πιθανά να αντιμετωπίζουν δυσκολίες στην ίαση τους είτε λόγω αδυναμίας συμμόρφωσης στην φαρμακευτική αγωγή και στην επανεξέταση-παρακολούθηση τους, είτε λόγω της επίδρασης της αιθανόλης στην ανοσολογική απόκριση.(41) Επιπλέον, τα ευρήματα ερευνών υποδηλώνουν ότι η κατανάλωση αλκοόλ μπορεί να επηρεάσει τον μεταβολισμό, την απορρόφηση και συνεπώς τις συγκεντρώσεις των φαρμάκων που χρησιμοποιούνται για τη θεραπεία όπως η ριφαμπικίνη και η ισονιαζίδη. Ως εκ τούτου, δεν θα επιτευχθεί η βέλτιστη συγκέντρωση τους, η

οποία είναι ζωτικής σημασίας για την έκβαση της θεραπείας.(42) Σε μια μετα-ανάλυση που πραγματοποιήθηκε για να εξεταστεί η συσχέτιση της κατανάλωσης αλκοόλ με την φυματίωση, οι ερευνητές συμπεριέλαβαν 36 μελέτες (κοόρτης και ασθενών-μαρτύρων) που προσφέρουν ποικιλομορφία στην γεωγραφική εκπροσώπηση των δειγμάτων. Η κατανάλωση αλκοόλ συσχετίστηκε με 35% υψηλότερο κίνδυνο εμφάνισης της φυματίωσης σε σύγκριση με τη μη κατανάλωση αλκοόλ (RR 1,35, 95% CI 1,09-1,68). Επιπλέον, εξετάστηκε αν η ποσότητα αιθανόλης σε γραμμάρια αυξάνει τον κίνδυνο ανάπτυξης φυματίωσης. Τα ευρήματα έδειξαν πως ο κίνδυνος αυξανόταν όσο αυξάνονταν και η πρόσληψη αιθανόλης. Συγκεκριμένα, πρόσληψη 25 g αιθανόλης την ημέρα συνεπάγεται σχετικό κίνδυνο 1,57 (RR 1,57, 95% CI 1,10-2,23), 50 g ημερησίως σχετικό κίνδυνο 2,46 (RR 2,46, 95% CI 1,21-4,98), 75 g ημερησίως σχετικό κίνδυνο 3,85 (RR 3,85, 95% CI 1,33-11,11) και 100 g ημερησίως (RR 6,03, 95% CI 1,47-24,81). Τα παραπάνω ευρήματα αναδεικνύουν μια σχέση δόσης-απόκρισης μεταξύ της κατανάλωσης αλκοόλ και του κινδύνου ανάπτυξης φυματίωσης.(43) Στην μετα-ανάλυση που πραγματοποιήθηκε από τον Ragan και τους συνεργάτες του, συμπεριλήφθηκαν 31 μελέτες με πολυανθεκτικά στελέχη του M.tuberculosis και 80 μελέτες με στελέχη που εμφάνιζαν ευαισθησία στα αντιβιοτικά. Τα αποτελέσματα, συμπεριλαμβανομένων και των περιπτώσεων που δεν ολοκλήρωσαν την θεραπευτική αγωγή, έδειξαν ότι οι ασθενείς που κατανάλωναν αλκοόλ είχαν υψηλότερες πιθανότητες κακής έκβασης της θεραπείας είτε εμφάνιζαν πολυανθεκτικότητα είτε ευαισθησία στα αντιβιοτικά με λόγο πιθανοτήτων 2 (OR 2.00, 95% CI 1.73–2.32) και 1,99 (OR 1,99, 95% CI 1,57-2,51) αντίστοιχα, σε σχέση με την ομάδα αναφοράς (καθόλου ή ελάχιστη κατανάλωση αλκοόλ). Όσον αφορά την έκβαση του θανάτου, η κατανάλωση αλκοόλ συσχετίστηκε με υψηλότερες πιθανότητες θανάτου σε ασθενείς που εμφάνιζαν πολυανθεκτικότητα είτε ευαισθησία στα αντιβιοτικά με λόγο πιθανοτήτων 1,38 (OR 1,38, 95% CI 1,04-1,83) και 1,58 (OR 1,58, 95% CI 1,24-2,00) αντίστοιχα.(41)

1.8.4 HIV

Ιδιαίτερη ανησυχία σε παγκόσμιο επίπεδο εμφανίζει η συλλοίμωξη του Μυκοβακτηριδίου της φυματίωσης με τον ιό της ανθρώπινης ανοσοανεπάρκειας (Human Immunodefiency Virus-HIV), καθώς έχει αποδειχθεί η κλινική σημασία της. Η λοίμωξη της φυματίωσης είναι ιδιαίτερα συχνή σε HIV θετικούς ασθενείς και αποτελεί κύρια αιτία θανάτου τους. Ο επιπολασμός της συλλοίμωξης HIV και φυματίωσης έχει μειωθεί σημαντικά, αλλά παραμένει υψηλότερος στις οικονομικά υποβαθμισμένες περιοχές. Η παρουσία του HIV συμβάλλει στην αυξημένη εμφάνιση της φυματίωσης και επηρεάζει την εξέλιξη της νόσου. Συγκεκριμένα, σύμφωνα με την παγκόσμια έκθεση του 2020 για την φυματίωση από τον ΠΟΥ, οι άνθρωποι που έχουν μολυνθεί με HIV έχουν 15-21 φορές περισσότερες πιθανότητες να αναπτύξουν ενεργό φυματίωση σε σχέση με άτομα που δεν έχουν τον ιό HIV.

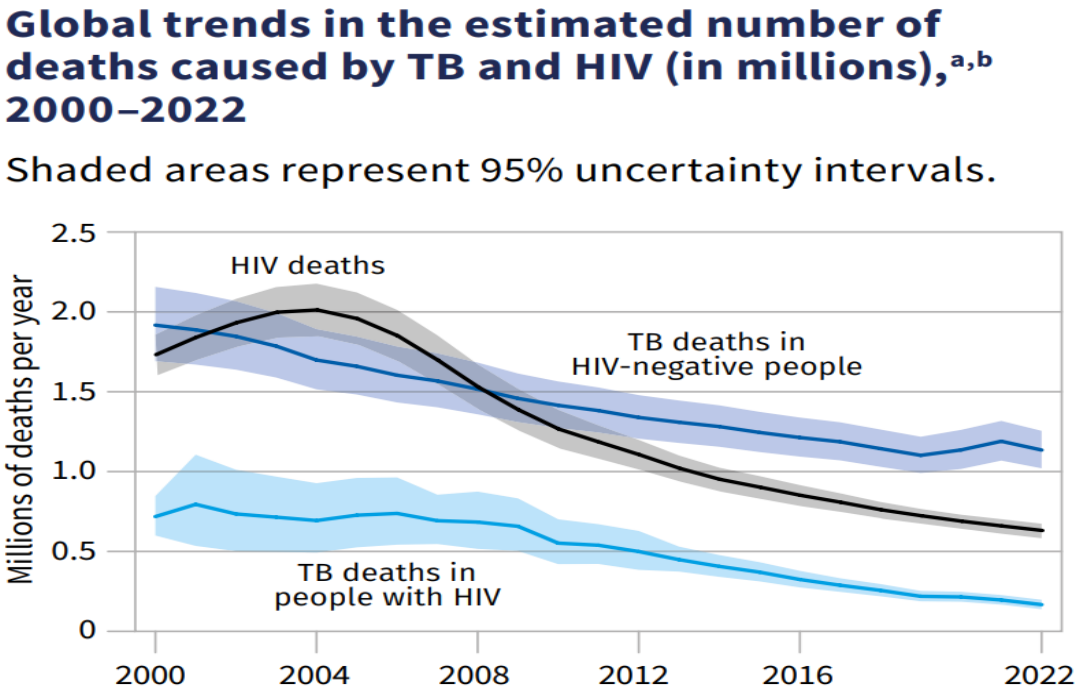
Επιπλέον, η λοίμωξη από τον ιό HIV μπορεί να προκαλέσει μετατροπή της λανθάνουσας φυματίωσης σε ενεργό. Ο κίνδυνος εμφάνισης της φυματίωσης είναι υψηλότερος σε ασθενείς με HIV λόγω του χαμηλότερου αριθμού CD4+ T λεμφοκυττάρων και της εξασθένισης της λειτουργείας των μακροφάγων, που έχει ως αποτέλεσμα την αποδυνάμωση του ανοσοποιητικού συστήματος και συνεπώς την αδυναμία περιορισμού της φυματίωσης.(44)(45)(46) Από την άλλη πλευρά, η φυματίωση ενισχύει τον πολλαπλασιασμό του ιού HIV και άρα την ανάπτυξη του, επηρεάζοντας έτσι την πορεία της HIV λοίμωξης.(47)

Σε μια αναδρομική μελέτη κοόρτης που πραγματοποιήθηκε στην Βραζιλία από τον Mauro Sanchez και τους συνεργάτες του, ανακτήθηκαν δεδομένα από το 2003 έως το 2008 με στόχο να αποτιμηθεί η έκβαση της θεραπείας της φυματίωσης. Τα αποτελέσματα της θεραπευτικής αγωγής της φυματίωσης ήταν καλύτερα για εκείνους που είχαν διαγνωστεί αρνητικοί στον HIV και χειρότερα για εκείνους που είχαν διαγνωστεί θετικοί στον HIV. Η πολυμεταβλητή ανάλυση έδειξε ότι οι θετικοί στον HIV ασθενείς με φυματίωση ήταν 3 φορές περισσότερο πιθανό (Risk ratio 3,09, 95% CI 3,02-3,16) να έχουν δυσμενή έκβαση, σε σύγκριση με τους αρνητικούς στον HIV ασθενείς με φυματίωση. Όσον αφορά την έκβαση του θανάτου βρέθηκε ότι οι HIV θετικοί ασθενείς με φυματίωση παρουσίασαν 9,24 φορές υψηλότερο κίνδυνο θανάτου (RR 9,24 95% CI 8.78–9.72) σε σύγκριση με τους HIV αρνητικούς.(48)

Σε μελέτη που πραγματοποιήθηκε σε δεδομένα ασθενών από εννιά Ευρωπαϊκές χώρες έγινε ανάλυση για την επίδρασης της HIV λοίμωξης στα αποτελέσματα της θεραπείας ασθενών με φυματίωση. Από τις 61.138 περιπτώσεις ασθενών οι 3.347 (5,5%) αφορούσαν συλλοίμωξη HIV και φυματίωσης. Τα αποτελέσματα της μελέτης έδειξαν ότι οι ασθενείς με συλλοίμωξη HIV και φυματίωσης είχαν χαμηλότερη επιτυχία (56,9%, P< 0,001) στη θεραπεία της φυματίωσης σε σύγκριση με τα HIV αρνητικά περιστατικά (78,7%, P< 0,001). Επιπλέον, τα ευρήματα ήταν περισσότερο ευνοϊκά για τους HIV αρνητικούς ασθενείς καθώς μεγαλύτερο ποσοστό ατόμων με συλλοίμωξη HIV πέθανε κατά τη διάρκεια της θεραπείας της φυματίωσης (13,5%, P < 0.001) σε σύγκριση με τα HIV αρνητικά άτομα (6,2%, P < 0.001).(49)

Επιπρόσθετα, στην Αιθιοπία ερευνητές διενέργησαν μια μελέτη κοόρτης για την διερεύνηση του αντίκτυπου της HIV λοίμωξης στην θεραπεία της φυματίωσης, συλλέγοντας αναδρομικά δεδομένα ασθενών. Τα αποτελέσματα της μελέτης έδειξαν ότι το ποσοστό θανάτου ήταν υψηλότερο στους ασθενείς με συλλοίμωξη HIV (19,4%) σε σύγκριση με τους HIV-αρνητικούς ασθενείς (2,9%). Οι HIV-αρνητικοί ασθενείς είχαν περίπου 10 φορές περισσότερες πιθανότητες επιτυχίας της αντιφυματικής θεραπείας (AOR = 10.3, 95% CI, 3.216–32.968, P < 0.001) σε σύγκριση με τους HIV-θετικούς ασθενείς με φυματίωση. Συνοπτικά, η συλλοίμωξη HIV με φυματίωση φαίνεται να σχετίζεται με χαμηλότερα ποσοστά ίασης, υψηλότερα ποσοστά θανάτου και μικρότερη πιθανότητα επιτυχίας της θεραπείας σε σύγκριση με άτομα χωρίς HIV λοίμωξη.(50)

Σύμφωνα με τον ΠΟΥ, οι θάνατοι από φυματίωση μεταξύ ατόμων με λοίμωξη HIV εμφανίζουν μια σταθερή πτωτική τάση τα τελευταία χρόνια (Εικόνα 7). Μάλιστα, το 2022 ο εκτιμώμενος αριθμός θανάτων από φυματίωση με HIV συλλοίμωξη ήταν 167.000 (95% UI: 139.000-198.000).(7)



Εικόνα 7: Παγκόσμιες τάσεις του εκτιμώμενου αριθμού θανάτων από φυματίωση και HIV (σε εκατομμύρια) το χρονικό διάστημα 2000-2022.

[Πηγή: World Health Organization. Global tuberculosis report 2023. Geneva: World Health Organization; 2023. Licence: CC BY-NC-SA 3.0 IGO.]

1.9 Φυματίωση και COVID-19

Πριν από την εμφάνιση της νόσου COVID-19, η φυματίωση ήταν η πιο θανατηφόρος μεταξύ των λοιμωδών ασθενειών. Ωστόσο, η εμφάνιση της νέας πανδημίας COVID-19 μετατόπισε την φυματίωση στην δεύτερη θέση όσον αφορά την κύρια αιτία θανάτου από μολυσματικές ασθένειες παγκοσμίως μέχρι και σήμερα σύμφωνα με τις στατιστικές αναλύσεις. Η συλλοίμωξη φυματίωσης με λοίμωξη από COVID-19 δημιούργησε σοβαρές επιπτώσεις στην δημόσια υγεία. Η πανδημία COVID-19 επηρέασε όλα τα προγράμματα εξάλειψης της φυματίωσης λόγω της σοβαρή επιβάρυνσης των συστημάτων υγειονομικής περίθαλψης για τον έλεγχο της πανδημίας. Συνεπώς, η διάγνωση νέων περιπτώσεων φυματίωσης, τα θεραπευτικά προγράμματα αλλά και οι εκστρατείες πρόληψης τέθηκαν σε δεύτερη μοίρα. Μάλιστα, σε

ενδημικές χώρες, όπως η Ινδία, η Κίνα, η Ινδονησία και οι Φιλιππίνες, παρατηρήθηκε μείωση κατά 25-30% των δηλωθέντων κρουσμάτων φυματίωσης κατά το πρώτο εξάμηνο του 2020. Έτσι, η μειωμένη δυνατότητα διάγνωσης των νέων κρουσμάτων οδήγησε σε αύξηση της μετάδοσης του M.tuberculosis.(51)

Η φυματίωση και η COVID-19 είναι δυο ασθένειες που προσβάλλουν το αναπνευστικό σύστημα και προκαλούν κάποια κοινά συμπτώματα όπως ο πυρετός, ο βήχας, η απώλεια όρεξης και η δύσπνοια. Αυτό μπορεί να δυσχεραίνει την διάγνωση ειδικά σε περιπτώσεις συλλοίμωξης. Η πιθανή καθυστέρηση στη διάγνωση κάποιας εκ των δυο ασθενειών και συνεπώς της θεραπείας, μπορεί να επηρεάσει σημαντικά την έκβαση της θεραπείας και την ίαση του ασθενούς.(52)

Η τήρηση των πρωτοκόλλων και η ολοκλήρωση της θεραπείας της φυματίωσης επηρεάζονται επίσης από την έλλειψη πόρων, φαρμάκων και ιατρικών προμηθειών λόγω της πανδημίας, γεγονός που μπορεί να οδηγήσει σε αποτυχία της θεραπείας και κατά συνέπεια σε αύξηση της επίπτωσης της πολυανθεκτικής φυματίωσης (Multidrug-resistant tuberculosis, MDR-TB) και της θνησιμότητας.(53)

Η φυματίωση και η COVID-19 επηρεάζουν πρωταρχικά τους πνεύμονες ως αρχική εντόπιση της μόλυνσης. Προκαλούν παρόμοιες ανοσολογικές αποκρίσεις, με παραγωγή ιδιαίτερα μεγάλης ποσότητας κυτταροκινών, η οποία επιδεινώνει την πνευμονική βλάβη προκαλώντας πνευμονικό οίδημα και μεγαλύτερη ευπάθεια σε δευτερογενείς λοιμώξεις.(54) Επιπλέον, τα ανοσοκατασταλτικά φάρμακα που χρησιμοποιούνται ως θεραπεία για την αντιμετώπιση της COVID-19, όπως τα κορτικοστεροειδή και οι αναστολείς κυτταροκινών ενέχουν τον κίνδυνο απόκτησης ευκαιριακών λοιμώξεων. Η ανοσοκατασταλτική αγωγή θα μπορούσε ενδεχομένως να οδηγήσει στην ενεργοποίηση της λανθάνουσας φυματίωσης, η οποία μπορεί να είναι επιζήμια σε περιοχές όπου η φυματίωση είναι ενδημική. (51)

Σε μετα-ανάλυση που πραγματοποιήθηκε το 2020, με σκοπό την αξιολόγηση των συνεπειών της COVID-19 λοίμωξης όταν υπάρχει ταυτόχρονη συλλοίμωξη με άλλες ασθένειες, βρέθηκε ότι παρατηρείται αρκετά αυξημένος κίνδυνος θνησιμότητας σε ασθενείς με φυματίωση που νοσούσαν ταυτόχρονα από COVID-19. Συγκεκριμένα, η μετα-ανάλυση έδειξε διπλάσια αύξηση της θνησιμότητας των ασθενών με συλλοίμωξη, (RR = 2.10, 95% CI, 1.75–2.51, I 2 = 0%).(55)

1.10 Εμβολιασμός

Ο εμβολιασμός κατά της φυματίωσης είναι ιδιαίτερα ουσιώδης και συμβάλλει στην πρόληψη της νόσου παρέχοντας μερική προστατευτική δράση κατά της ενεργού φυματίωσης. Το εμβόλιο Bacillus Calmette-Guérin (BCG) είναι το μόνο εγκεκριμένο εμβόλιο και περιορίζει τη θνησιμότητα σε ορισμένους πληθυσμούς. Η ανάπτυξη του ξεκίνησε το 1921, ενώ το 1974 καθιερώθηκε και συμπεριλήφθη από τον ΠΟΥ στο διευρυμένο πρόγραμμα ανοσοποίησης (Expanded Programme on Immunization, EPI). Ωστόσο, η αποτελεσματικότητα του είναι περιορισμένη και δεν μπορεί να θεωρηθεί επαρκής για τον περιορισμό της φυματίωσης. Το εμβόλιο BCG δεν προσφέρει προστασία έναντι της πρωτογενούς λοίμωξης, της επανενεργοποίησης της λανθάνουσας φυματίωσης και της μετάδοσης του Μυκοβακτηριδίου από άτομο σε άτομο. Έτσι, η ανάπτυξη ενός νέου εμβολίου πιο αποτελεσματικού είναι επιτακτικής ανάγκης.(56)(57)(58)

Οι χώρες που παρουσιάζουν υψηλή συχνότητα εμφάνισης νέων περιπτώσεων φυματίωσης εφαρμόζουν στρατηγικές καθολικού εμβολιασμού, ενώ οι χώρες με μέτρια ή χαμηλή εμφάνιση εφαρμόζουν τον εμβολιασμό επιλεκτικά σε ομάδες υψηλού κινδύνου. (59)

1.11 Θεραπεία

Η φυματίωση είναι μια ασθένεια η οποία μπορεί να θεραπευτεί. Η θεραπευτική αγωγή της φυματίωσης κρίνεται επιτυχημένη όταν ολοκληρωθεί και επέλθει η ίαση του ασθενούς.(60) Η ίαση συμβάλλει στον περιορισμό της εξάπλωσης της νόσου στον γενικό πληθυσμό. Επιπλέον, στόχος της είναι η αποφυγή ανάπτυξης ανθεκτικών στελεχών του M.Tuberculosis. Συχνά όμως, πιθανές παρενέργειες εμφανίζονται λόγω των αντιφυματικών φαρμάκων που χορηγούνται. Η θεραπεία της φυματίωσης βασίζεται σε συνδυασμό φαρμάκων.

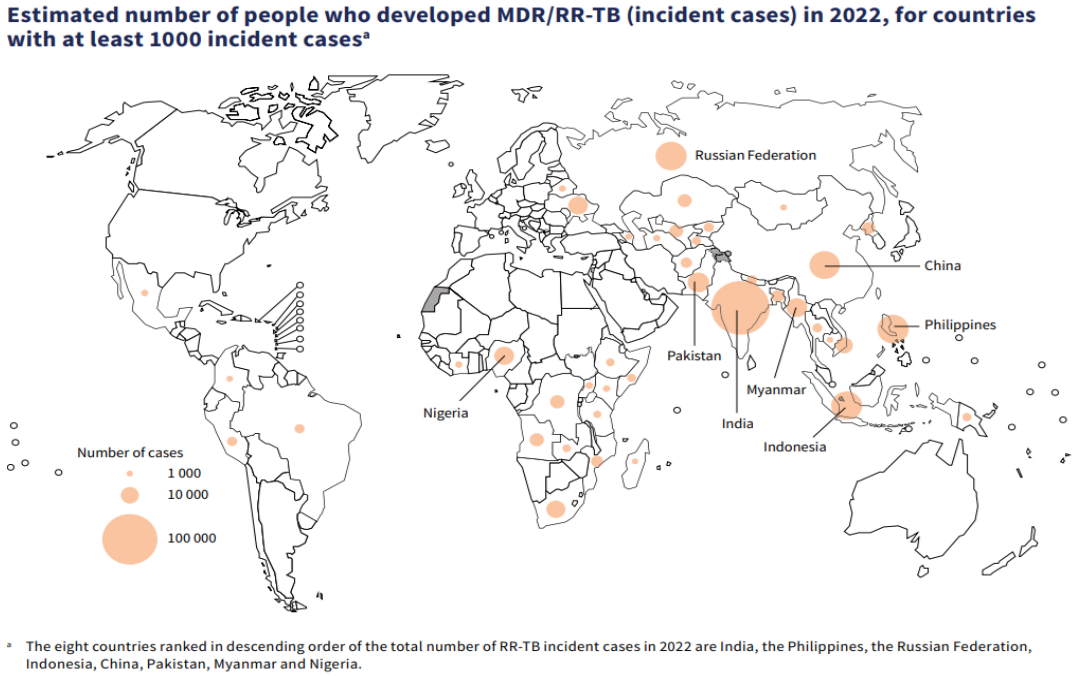
Η τρέχουσα θεραπεία που συνιστάται περιλαμβάνει τον συνδυασμό των τεσσάρων πρωτευόντων αντιβιοτικών ισονιαζίδη (INH), ριφαμπικίνη (RIF), πυραζιναμίδη (PZA) και εθαμβουτόλη (EMB). Ο συνδυασμός αυτός θα πρέπει να χορηγείται στους πάσχοντες για τουλάχιστον 6 μήνες στο πλαίσιο της θεραπείας υπό άμεση παρακολούθηση (Directly Observed Treatment, DOT) για να διασφαλίζονται υψηλά τα ποσοστά της ολοκλήρωσης της θεραπείας και της ίασης. Η θεραπεία διακρίνεται σε δύο φάσεις, την αρχική φάση η οποία περιλαμβάνει τη χορήγηση των προαναφερθέντων τεσσάρων φαρμάκων κατά την έναρξη της θεραπείας για τους πρώτους δύο μήνες, και τη φάση συνέχισης της θεραπείας με τα αντιβιοτικά INH και RIF για τους τελευταίους τέσσερις μήνες με στόχο να σκοτωθούν τα αδρανή βακτήρια. Οι μηχανισμοί

δράσης αυτών των φαρμάκων διαφέρουν. Για παράδειγμα, η ριφαμπικίνη ασκεί τη βακτηριοκτόνο δράση της αναστέλλοντας τα πρώιμα στάδια της γονιδιακής μεταγραφής όταν συνδέεται με την β-υπομονάδα της RNA πολυμεράσης.(61) Επιπλέον, η ριφαμπικίνη παρουσιάζει αποστειρωτική δράση κατά των βραδέως πολλαπλασιαζόμενων υποπληθυσμών του Μυκοβακτηριδίου. Η ισονιαζίδη εμφανίζει βακτηριοκτόνο δράση κατά των εξωκυττάριων και εσωκυττάριων βακίλων, ενώ η πυραζιναμίδη έχει βακτηριοκτόνο δράση κυρίως στους βραδέως πολλαπλασιαζόμενους βακίλους και αποστειρωτική δράση συνεργικά με την ισονιαζίδη και την ριφαμπικίνη. Τέλος, ανάλογα με την χορηγούμενη δοσολογία, η εθαμβουτόλη μπορεί να έχει βακτηριοκτόνο και βακτηριοστατική δράση. (62)

Στις περιπτώσεις που είναι εφικτό σύμφωνα με τον ΠΟΥ, η βέλτιστη συχνότητα χορήγησης δόσης σε νέους ασθενείς με πνευμονική φυματίωση είναι καθημερινά και καθ' όλη τη διάρκεια της θεραπείας. Επιπλέον, μια νέα σύσταση θεραπείας που αφορά τα παιδιά και τους εφήβους (ηλικίας 3 μηνών έως 16 ετών) που διαγιγνώσκονται με φυματίωση και δεν παρουσιάζουν υποψία ή ενδείξεις πολυανθεκτικής ή ανθεκτικής στη ριφαμπικίνη φυματίωσης (MDR/RR-TB), προτείνει τη χρήση ενός θεραπευτικού σχήματος 4 μηνών. Συγκεκριμένα στους δυο πρώτους μήνες συνδυασμό ισονιαζίδης, ριφαμπικίνης και πυραζιναμίδης, ακολουθούμενου από δύο επιπλέον μήνες με χορήγηση ισονιαζίδης και ριφαμπικίνης. (63)

Ιδιαίτερη δυσκολία στη θεραπεία παρουσιάζεται στους ασθενείς με πολυανθεκτικά στελέχη φυματίωσης (Multidrug-resistant tuberculosis, MDR-TB). Για την πρόληψη της δημιουργίας ανθεκτικών στα αντιβιοτικά στελεχών, κρίνεται αναγκαία η ανάπτυξη συντομότερων θεραπευτικών σχημάτων τα οποία θα εξαλείφουν ταχέως όλους τους πληθυσμούς των μυκοβακτηριδίων. Ταυτόχρονα, θα περιορίζονται οι βακτηριακές μεταβολικές διεργασίες που οδηγούν στην αντοχή στα αντιβιοτικά και στη μεταλλαξιγένεση.(64) Η πολυανθεκτική φυματίωση αφορά την αντοχή στην ισονιαζίδη και την ριφαμπικίνη (RR-TB), τα δύο ισχυρότερα φάρμακα πρώτης γραμμής για την αντιμετώπιση της φυματίωσης. Το 2021, εκτιμάται ότι υπήρχαν 450.000 περιστατικά MDR-TB. Μάλιστα, τα ποσοστά ίασης για την MDR-TB είναι συνήθως σημαντικά χαμηλότερα σε σχέση με τα ευαίσθητα στα αντιβιοτικά στελέχη.(7) Έτσι λοιπόν, για την πολυανθεκτική φυματίωση όπου τα φάρμακα πρώτης γραμμής δεν είναι αποτελεσματικά, χρησιμοποιούνται φάρμακα δεύτερης γραμμής όπως για παράδειγμα τα από του στόματος χορηγούμενα (παραμινοσαλικυλικό οξύ, κυκλοσερίνη, τεριζιδόνη) και ενέσιμα αντιφυματικά φάρμακα (στρεπτομυκίνη, καναμυκίνη, αμικασίνη και καπρεομυκίνη). Επίσης, μπορούν να χρησιμοποιηθούν οι κινολόνες όπως η λεβοφλοξασίνη και η μοξιφλοξασίνη.(62) Ένα φαρμακευτικό σχήμα που συνέστησε ο ΠΟΥ το 2020 σε ανθεκτική στην ισονιαζίδη, αλλά ευαίσθητη στη ριφαμπικίνη φυματίωση, είναι να αντιμετωπίζεται με ριφαμπικίνη, εθαμβουτόλη, πυραζιναμίδη και λεβοφλοξασίνη για 6 μήνες. Ακόμη, τα θεραπευτικά σχήματα για RR-TB και MDR-TB που πραγματοποιούνται σε μια εντατική φάση διάρκειας 4 μηνών αποτελούνται από λεβοφλοξασίνη/μοξιφλοξασίνη, κλοφαζιμίνη, εθιοναμίδη,

εθαμβουτόλη, ισονιαζίδη (σε υψηλή δόση), πυραζιναμίδη και μπεντακιλίνη (για 6 μήνες). Έπειτα, ακολουθεί η φάση συνέχισης με θεραπεία με λεβοφλοξασίνη/μοξιφλοξασίνη, κλοφαζιμίνη, εθαμβουτόλη και πυραζιναμίδη με διάρκεια 5 μηνών. Τα άτομα που έχουν μολυνθεί με φυματίωση ανθεκτική σε οποιοδήποτε άλλο αντιβιοτικό απαιτούν θεραπευτικά σχήματα που διαρκούν 18 μήνες και θα πρέπει να είναι εξατομικευμένα για κάθε περίπτωση.(61)(64)



Εικόνα 8: Εκτιμώμενος αριθμός ατόμων που ανέπτυξαν MDR/RR-TB (νέα περιστατικά) το 2022, για τις χώρες με τουλάχιστον 1000 περιστατικά.

[Πηγή: World Health Organization. Global tuberculosis report 2023. Geneva: World Health Organization; 2023. Licence: CC BY-NC-SA 3.0 IGO.]

Ως υπερανθεκτικότητα (Extensively Drug-Resistant Tuberculosis/XDR-TB) ορίζεται η εκτεταμένα ανθεκτική στα αντιβιοτικά φυματίωση. Τα στελέχη που είναι MDR με πρόσθετη αντοχή σε φάρμακα δεύτερης γραμμής, συγκεκριμένα σε μια φθοριοκινολόνη και είτε στην μπεντακιλίνη είτε στην λινεζολίδη.(60)

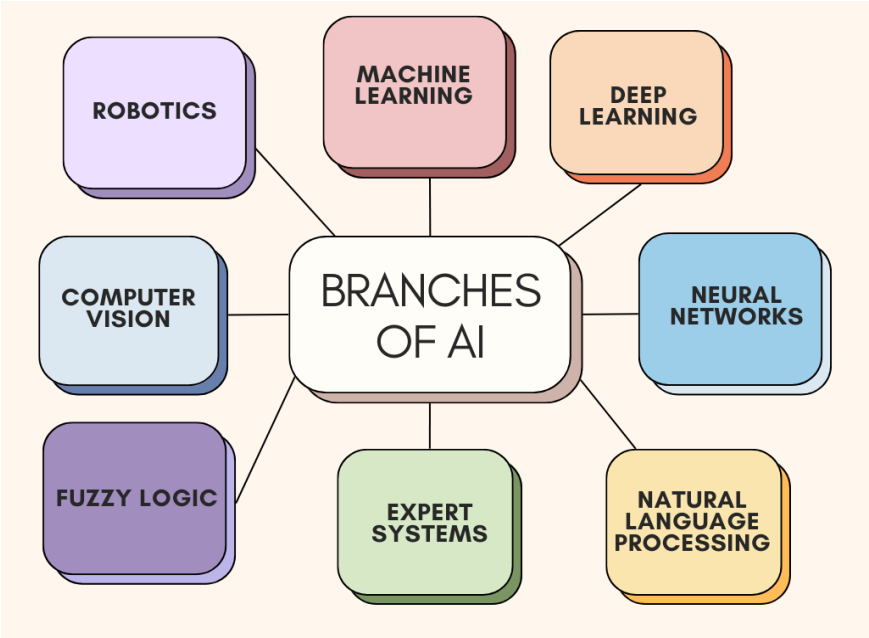
Όλα τα παραπάνω φαρμακευτικά σχήματα για την αντιμετώπιση της φυματίωσης είναι πιθανό να ενέχουν κάποιες ανεπιθύμητες ενέργειες σε ορισμένους ασθενείς. Αυτές μπορεί να είναι ήπιες αλλά και αρκετά σοβαρές. Επί παραδείγματι, μερικές από αυτές είναι η ανάπτυξη εξανθημάτων και οι γαστρεντερικές διαταραχές από οποιοδήποτε φάρμακο, η ηπατοτοξικότητα από συγκεκριμένα αντιβιοτικά όπως η ισονιαζίδη και η πυραζιναμίδη, η οπτική νευρίτιδα με εκδήλωση μειωμένης οπτικής οξύτητας και δυσχρωματοψία, η νεφροτοξικότητα, η περιφερική νευρίτιδα και η νεύροτοξικότητα. Σε καταστάσεις υπερευαισθησίας με συμπτώματα όπως ο πυρετός, η κνίδωση

και η δύσπνοια, απαραίτητη είναι η άμεση διακοπή της φαρμακευτικής αγωγής και η ανεύρεση του αιτιολογικού παράγοντα της υπερευαισθησίας. Τέλος, σε περίπτωση που η αγωγή κατά της φυματίωσης διακοπεί για χρονικό διάστημα μεγαλύτερο των 14 ημερών, κατά το αρχικό στάδιο της θεραπείας, τότε κρίνεται αναγκαίο να ξεκινήσει ο ασθενής εκ νέου την φαρμακευτική αγωγή.(62)

2.Τεχνητή Νοημοσύνη και Μηχανική Μάθηση

2.1 Τεχνητή Νοημοσύνη

Η Τεχνητή Νοημοσύνη (Artificial Intelligence, AI) σύμφωνα με τον ορισμό που παρέχει το Medical Subject Headings (MeSH) και εισήχθη το 1986, αναφέρεται στην θεωρία και την ανάπτυξη υπολογιστικών συστημάτων, που εκτελούν εργασίες που συνήθως απαιτούν ανθρώπινη νοημοσύνη. Τέτοιες εργασίες μπορεί να είναι η αναγνώριση ομιλίας, η μάθηση, η οπτική αντίληψη, οι μαθηματικοί υπολογισμοί, η συλλογιστική, η επίλυση προβλημάτων, η λήψη αποφάσεων και η μετάφραση της γλώσσας.(65) Η τεχνητή νοημοσύνη απαιτεί ελάχιστη ανθρώπινη παρέμβαση. Ορισμένοι από τους κυριότερους κλάδους (Εικόνα 9) της είναι η Μηχανική Μάθηση (Machine Learning), η Βαθιά Μάθηση (Deep Learning), τα Νευρωνικά Δίκτυα (Neural Networks), η Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing), η Ρομποτική (Robotics), η Υπολογιστική Όραση (Computer Vision), η Ασαφής Λογική (Fuzzy Logic) και τα Συστήματα Εμπειρογνωμόνων (Expert Systems).(66)



Εικόνα 9: Κλάδοι της Τεχνητής νοημοσύνης.

[Πηγή: Δημιουργήθηκε από την συγγραφέα με τη χρήση του https://www.canva.com/]

Η ραγδαία εξέλιξη στον τομέα της τεχνολογίας τα τελευταία χρόνια, έχει οδηγήσει στην ψηφιακή επανάσταση, με αποτέλεσμα την εξάπλωση της τεχνητής νοημοσύνης σε διάφορους τομείς της καθημερινότητας μας. Η συμβολή της τεχνητής νοημοσύνης στον τομέα της υγείας επαναπροσδιορίζει

την διαδικασία της πρόληψης και διάγνωσης ασθενειών και της περίθαλψης των ασθενών. Παρέχει την δυνατότητα εξατομικευμένης ιατρικής η οποία βασίζεται στα ατομικά χαρακτηριστικά και στις ανάγκες κάθε ασθενούς. Τα υπολογιστικά μοντέλα και οι αλγόριθμοι της τεχνητής νοημοσύνης, δύνανται να κατανοήσουν περίπλοκα και μεγάλα σε όγκο ιατρικά δεδομένα, αναβαθμίζοντας την παρεχόμενη υγειονομική φροντίδα.(67)

2.2 Μηχανική μάθηση

Η Μηχανική Μάθηση αποτελεί κλάδο της τεχνητής νοημοσύνης και εμφανίζει σημαντική άνθιση και ανάπτυξη τις τελευταίες δεκαετίες. Ο Άρθουρ Σάμιουελ επινόησε τη φράση "μηχανική μάθηση" το 1952, όταν ανέπτυξε ένα πρόγραμμα υπολογιστή για ένα παιχνίδι με σκοπό να προβλέπει τις πιθανότητες νίκης και την επόμενη κίνηση. (68) Ακόμη, σύμφωνα με το MeSH ένας ορισμός που παρέχεται και εισήχθη το 2016 είναι ο εξής: “Η μηχανική μάθηση είναι ένας τύπος τεχνητής νοημοσύνης που επιτρέπει στους υπολογιστές να προβαίνουν ανεξάρτητα σε εκκίνηση και εκτέλεση μάθησης, όταν έρχονται σε επαφή με νέα δεδομένα”.(69) Ένας άλλος γενικός ορισμός για την έννοια της μηχανικής μάθησης ο οποίος δίνεται από τον Mitchell το 1997 είναι ο παρακάτω: “ Ένα πρόγραμμα υπολογιστή λέμε ότι μαθαίνει από την εμπειρία E, όπου T η κλάση εργασιών και P το μέτρο απόδοσης, αν η απόδοση P στις εργασίες T βελτιώνεται μέσω της εμπειρίας E”.(70)

Η ανάλυση δεδομένων μας επιτρέπει να κατανοήσουμε διάφορα φαινόμενα, να μοντελοποιήσουμε συμπεριφορές και να κάνουμε προβλέψεις. Παλαιότερα, οι άνθρωποι ανέλυαν δεδομένα και κατασκεύαζαν αλγόριθμους, τους οποίους στη συνέχεια χρησιμοποιούσαν οι μηχανές για την επίλυση προβλημάτων. Στην εποχή μας, οι άνθρωποι παρέχουν δεδομένα στις μηχανές και επιτρέπουν σε αυτές να μαθαίνουν από μόνες τους χωρίς να προγραμματίζονται ρητά και με όσο το δυνατόν ελάχιστη ανθρώπινη παρέμβαση.(71)

2.2.1 Εφαρμογές μηχανικής μάθησης

Οι εφαρμογές της μηχανικής μάθησης αφορούν ένα όλο και πιο διευρυμένο σύνολο πεδίων τόσο στην καθημερινή ζωή και την επιστήμη όσο και στην βιομηχανία. Η ανάπτυξη που παρατηρείται σε αυτόν τον κλάδο είναι ραγδαία τα τελευταία χρόνια και πολλά υποσχόμενη για το μέλλον. Κάποιες από τις εφαρμογές αυτές αναφέρονται παρακάτω.

2.2.1.1 Κυβερνοασφάλεια

Η κυβερνοασφάλεια αφορά την πρακτική της προστασίας των δικτύων και των συστημάτων, του υλικού και των δεδομένων από ψηφιακές επιθέσεις. Η μηχανική μάθηση αναλύοντας συνεχώς δεδομένα, εντοπίζει μοτίβα και μαθαίνει να ανιχνεύει καλύτερα το κακόβουλο λογισμικό (malware), να βρίσκει εσωτερικές απειλές, να παρέχει ασφάλεια κατά την περιήγηση στο διαδίκτυο και να διασφαλίζει την ακεραιότητα των δεδομένων σε υπηρεσίες cloud, αποκαλύπτοντας ύποπτη δραστηριότητα. Επιπλέον, η μηχανική μάθηση συμβάλλει στην αναγνώριση ανεπιθύμητων δραστηριοτήτων (spam) όπως τα ανεπιθύμητα μηνύματα και οι ιστοσελίδες. (72)(73)

2.2.1.2 Εμπόριο και διαφήμιση

Στον τομέα του εμπορίου και της διαφήμισης, η μηχανική μάθηση συμβάλλει στην δημιουργία βέλτιστων στρατηγικών προώθησης με στόχο την αύξηση των πωλήσεων, την ικανοποίηση των ήδη υπάρχοντών πελατών και την προσέλκυση νέων πελατών. Αυτό επιτυγχάνεται μέσα από την ανάλυση των προτιμήσεων και της καταναλωτικής συμπεριφοράς του πελάτη αλλά και των δημογραφικών χαρακτηριστικών του όπως για παράδειγμα η ηλικία και το φύλο. Έτσι, μπορούν να γίνουν εξατομικευμένες προτάσεις για αγορά προϊόντων με βάση το προφίλ του κάθε καταναλωτή. Τα εργαλεία συστάσεων είναι ιδιαίτερα ισχυρά και προσοδοφόρα στον τομέα του ηλεκτρονικού εμπορίου. Επιπρόσθετα, οι εταιρείες ηλεκτρονικού εμπορίου μπορούν εύκολα να τοποθετήσουν προτάσεις προϊόντων και προσφορές αναλύοντας τις τάσεις περιήγησης και τα ποσοστά προβολών σε συγκεκριμένα είδη με στόχο την αύξηση των κερδών τους. Μια άλλη εφαρμογή των μοντέλων πρόβλεψης της μηχανικής μάθησης στο ηλεκτρονικό εμπόριο είναι η καλύτερη διαχείριση των αποθεμάτων και η αποτροπή εξάντλησης τους ώστε να βελτιστοποιηθεί η διαδικασία του εφοδιασμού και της αποθήκευσης.(72)(73)

2.2.1.3 Επεξεργασία φυσικής γλώσσας και ανάλυση συναισθήματος

Η επεξεργασία φυσικής γλώσσας (Natural Language Processing, NLP) περιλαμβάνει την ανάγνωση και κατανόηση προφορικού ή γραπτού λόγου μέσω υπολογιστή. Έτσι, παρέχεται η δυνατότητα στους υπολογιστές να διαβάζουν ένα κείμενο, να ακούν ομιλία, να την ερμηνεύουν, να αναλύουν το συναίσθημα και να αποφασίζουν ποιες πληροφορίες είναι σημαντικές, ώστε να χρησιμοποιηθούν σε αυτές τεχνικές μηχανικής μάθησης. Παραδείγματα που σχετίζονται με την NLP είναι ο εικονικός προσωπικός βοηθός (virtual

personal assistant), τα chatbot, η περιγραφή εγγράφων και η γλωσσική μετάφραση. Η ανάλυση συναισθήματος (sentiment analysis) ή εξόρυξη γνώμης (opinion mining) είναι ένα πεδίο της NLP που επιδιώκει να εντοπίσει και να κάνει εκτίμηση της διάθεσης και των απόψεων του κοινού μέσω των ιστολογίων, των μέσων κοινωνικής δικτύωσης, των κριτικών, των φόρουμ, κ.λπ. Για παράδειγμα, οι επιχειρήσεις χρησιμοποιούν την ανάλυση συναισθήματος για να κατανοήσουν την άποψη του κοινού για το εμπορικό σήμα, το προϊόν ή την παροχή υπηρεσίας. Αυτό μπορεί να επιτευχθεί από τις πλατφόρμες κοινωνικής δικτύωσης ή γενικότερα το διαδίκτυο. Τα συναισθήματα μπορούν να χαρακτηρίζονται ως "θετικό", "αρνητικό" ή "ουδέτερο" ή πιο περιγραφικά όπως πολύ χαρούμενος, ευτυχισμένος, λυπημένος, πολύ λυπημένος, θυμωμένος, έχω ενδιαφέρον ή δεν ενδιαφέρομαι κ.λπ. Ένα άλλο παράδειγμα είναι η αναγνώριση των ψευδών ειδήσεων μέσω αναγνώρισης των ψευδών κειμένων. (72)(73)

2.2.1.4 Αειφόρος γεωργία

Οι πρακτικές της αειφόρου γεωργίας συμβάλλουν στη βελτίωση της γεωργικής παραγωγικότητας, μειώνοντας παράλληλα τις αρνητικές επιπτώσεις στο περιβάλλον.(72) Η μηχανική μάθηση έχει εφαρμογές στις διάφορες φάσεις του τομέα της γεωργίας. Για παράδειγμα, οι αλγόριθμοι μηχανικής μάθησης μπορούν να φανούν χρήσιμοι στην αρχική φάση πριν την παραγωγή, για την πρόβλεψη των αποδόσεων της σοδειάς με βάση διάφορες παραμέτρους όπως ο καιρός, η ημερομηνία καλλιέργειας, ο τύπος του εδάφους, το pH του εδάφους, οι συγκεντρώσεις διάφορων χημικών στοιχείων στο έδαφος και άλλοι παράγοντες.(74) Επίσης, στην φάση της παραγωγής η μηχανική μάθηση μπορεί να χρησιμοποιηθεί για την πρόβλεψη των καιρικών συνθηκών, την ανίχνευση ασθενειών στην καλλιέργεια και τη διαχείριση των θρεπτικών στοιχείων του εδάφους. Στη φάση της μεταποίησης, για την εκτίμηση της ζήτησης, τον προγραμματισμό της παραγωγής κ.λπ. και στη φάση της διανομής, για τη διαχείριση των αποθεμάτων, την ανάλυση της συμπεριφοράς των καταναλωτών κ.λπ.(72) Ακόμη, για την διαχείριση της άρδευσης με στόχο την καλύτερη εξοικονόμηση του νερού όσον αφορά την επίτευξη μιας βιώσιμης φυτική παραγωγής, τα μοντέλα μηχανικής μάθησης μπορούν να βοηθήσουν στην βέλτιστη χρήση των υδάτινων πόρων μέσω της πρόβλεψης των αναγκών άρδευσης με βάση την κατάσταση του εδάφους, τις καιρικές συνθήκες και άλλες μεταβλητές.(75)

2.2.1.5 Υγεία

Η εφαρμογή της μηχανικής μάθησης στον τομέα της υγείας αποτελεί έναν σημαντικό παράγοντα εξέλιξης, δημιουργώντας νέες δυνατότητες για την πρόληψη, τη διάγνωση, τη θεραπεία, την παρακολούθηση της υγείας, τη δημιουργία νέων φαρμάκων και τη διαχείριση των ιατρικών πόρων. Οι αλγόριθμοι της μηχανικής μάθησης εκμεταλλεύονται την διαχρονική παραγωγή του τεράστιου όγκου ιατρικών δεδομένων. Παρακάτω αναφέρονται ενδεικτικά κάποιοι κλάδοι της υγείας και η συμβολή της μηχανικής μάθησης σε αυτούς.

Η μηχανική μάθηση έχει αναδειχθεί σε ένα κρίσιμο εργαλείο στην υγειονομική περίθαλψη, ιδίως κατά τη διάρκεια της πανδημίας **COVID-19**. Βοηθά στην ταξινόμηση των ασθενών σε υψηλού κινδύνου, προβλέπει τις εξάρσεις κρουσμάτων και βοηθά στη διάγνωση και τη θεραπεία της νόσου.(72)

Στην υποβοηθούμενη **διάγνωση ασθενειών** μέσω υπολογιστών (Computer Assisted Diagnosis, CAD) η μηχανική μάθηση συμβάλλει μέσω της ανάλυσης κλινικών και εργαστηριακών δεδομένων, αλλά και των δημογραφικών στοιχείων όπως η ηλικία και το φύλο. Τα εργαστηριακά δεδομένα που χρησιμοποιούνται για τη δημιουργία μοντέλων μηχανικής μάθησης προέρχονται από απεικονιστικές τεχνικές (ακτινογραφία, αξονική ή μαγνητική τομογραφία, κλπ), από ηλεκτροκαρδιογραφήματα, βιοχημικούς και αιματολογικούς δείκτες, τιμές ζωτικών σημείων, κλπ. (73) Για παράδειγμα, στην **ογκολογία** η μηχανική μάθηση έχει ποικίλες εφαρμογές, καθώς χρησιμοποιείται συχνά διάγνωση και ανίχνευση του καρκίνου. Τα μοντέλα μηχανικής μάθησης χρησιμοποιούνται για την ανίχνευση και την ταξινόμηση όγκων μέσω εικόνων ακτινογραφιών και την ταξινόμηση κακοηθειών από γονιδιωματικές δοκιμασίες. Εκτός από τη διάγνωση, η μηχανική μάθηση στην ογκολογία φαίνεται χρήσιμη και στην εκτίμηση του κινδύνου, προσδιορίζοντας την πιθανότητα εμφάνισης και επανεμφάνισης καρκίνου σε ένα άτομο, αλλά και την πιθανότητα επιβίωσης του ατόμου από αυτόν.(76)

Όσον αφορά τον τομέα της **καρδιολογίας**, τα σύγχρονα μοντέλα μηχανικής μάθησης μπορούν να προσδιορίσουν τις μορφολογίες των κυμάτων που προκύπτουν από το ηλεκτροκαρδιογράφημα με μεγάλη ακρίβεια, επιτρέποντας τον υπολογισμό κλινικά σημαντικών παραμέτρων όπως ο καρδιακός ρυθμός (heart rate) και η απόκλιση του άξονα (axis deviation). Εντοπίζουν επίσης κοινές διαταραχές του καρδιακού ρυθμού όπως η κολπική μαρμαρυγή και παθήσεις όπως η στένωση αορτικής βαλβίδας και η υπερτροφική μυοκαρδιοπάθεια. Τα τελευταία χρόνια γίνονται προσπάθειες με τη χρήση της μηχανικής μάθησης ώστε να διευκολυνθεί η διάγνωση της καρδιακής ανεπάρκειας, η εκτίμηση της σοβαρότητας της και να γίνει δυνατή η πρόβλεψη ανεπιθύμητων επιπλοκών.(77)(78)

Στον κλάδο της **οφθαλμολογίας**, η εξέλιξη της τεχνητής νοημοσύνης και συγκεκριμένα της μηχανικής μάθησης στην ανίχνευση και παρακολούθηση οφθαλμολογικών ασθενειών εξελίσσεται με ταχείς ρυθμούς. Μοντέλα

μηχανικής μάθησης θα μπορούσαν να ελαχιστοποιήσουν την υποκειμενικότητα της διάγνωσης ερμηνεύοντας και ποσοτικοποιώντας τη βλάβη σε εικόνες του αμφιβληστροειδούς και του οπτικού νεύρου. Χρησιμοποιώντας λοιπόν φωτογραφίες του βυθού του οφθαλμού μπορεί να υπάρξει βελτίωση στην διάγνωση και την πρόβλεψη του γλαυκώματος και άλλων οφθαλμολογικών παθήσεων.(79) Ένα άλλο παράδειγμα εφαρμογής της μηχανικής μάθησης είναι η χρήση ενός μοντέλου που αφορά τις επισκέψεις στα τμήματα επειγόντων οφθαλμολογικών περιστατικών το οποίο θα διακρίνει τα επείγοντα από τα μη επείγοντα περιστατικά βελτιστοποιώντας έτσι την διαλογή (triage) ασθενών. Δημογραφικά και κλινικά χαρακτηριστικά χρησιμοποιούνται ως είσοδοι, με το σύστημα να δίνει συστάσεις ανάλογα με την σοβαρότητα του περιστατικού ώστε είτε να γίνει άμεση αντιμετώπιση του, είτε εφόσον δεν επείγει το περιστατικό να πραγματοποιηθεί προγραμματισμένη εξέταση.(80)

Στον τομέα της **μικροβιολογίας**, τα μοντέλα μηχανικής μάθησης μπορούν να συμβάλουν στην πρόβλεψη όσον αφορά την ταυτοποίηση των μικροοργανισμών ανάλογα με τα διαφορετικά χαρακτηριστικά τους. Οι μικροοργανισμοί μπορούν να ταξινομηθούν είτε με βάση το βασίλειο, την συνομοταξία, την τάξη, την οικογένεια, το γένος και το είδος, είτε να προσδιοριστεί αν ανήκουν σε ένα συγκεκριμένο είδος ή όχι. Επιπλέον, η μηχανική μάθηση χρησιμοποιείται σε μελέτες που αφορούν το ανθρώπινο μικροβίωμα με στόχο να διερευνηθούν οι ανισορροπίες σε αυτό και να προβλεφθούν ασθένειες όπως η βακτηριακή κολπίτιδα και η φλεγμονώδης νόσος του εντέρου. (81) Σημαντικό ρόλο παίζει η χρήση των τεχνικών μηχανικής μάθησης στην πρόβλεψη της μικροβιακής αντοχής στα αντιβιοτικά, με πληθώρα μελετών να διενεργείται τα τελευταία χρόνια. Στόχος αυτών των μελετών είναι η δημιουργία μοντέλων που διακρίνουν τα ανθεκτικά μικροβιακά στελέχη γρηγορότερα σε σχέση με τις κλασικές εργαστηριακές μεθόδους, με αποτέλεσμα την ταχύτερη έναρξη σωστής αντιμικροβιακής θεραπείας. (82)(83)

Παράλληλα, η μηχανική μάθηση συμβάλλει στην αντιμετώπιση του **διαβήτη τύπου 2**. Για παράδειγμα, έχουν δημιουργηθεί εφαρμογές που συμβάλλουν στην αυτοφροντίδα των ασθενων για την καλύτερη διαχείριση του διαβήτη, όπως συστήματα εικονικών βοηθών ή εφαρμογές, τα οποία στοχεύουν στην σωστή πρόσληψη τροφής από τους ασθενείς ή προτείνουν κατάλληλες επιλογές τροφίμων για διαβητικούς ασθενείς με βάση τις γευστικές τους προτιμήσεις. Επιπλέον, η μηχανική μάθηση μπορεί να συμβάλλει στην πρόβλεψη της γλυκόζης του αίματος και των διακυμάνσεων της και την ανίχνευση της υπεργλυκαιμίας. Όσον αφορά τις επιπλοκές του διαβήτη, για την ανίχνευση της διαβητικής αμφιβληστροειδοπάθειας, οι αλγόριθμοι μπορούν να αναλύσουν εικόνες του αμφιβληστροειδούς ώστε να γίνει η διάγνωση της με μεγάλη ακρίβεια, επιτρέποντας την έγκαιρη παρέμβαση και την πρόληψη της απώλειας όρασης του διαβητικού ασθενούς.(84)

Μια άλλη εφαρμογή της μηχανικής μάθησης στον τομέα της υγείας είναι τα **συστήματα υποστήριξης κλινικών αποφάσεων (Clinical Decision**

Support System-CDSS), τα οποία λειτουργούν υποστηρικτικά για τους κλινικούς ιατρούς στην λήψη αποφάσεων για τη φροντίδα των ασθενών. Για παράδειγμα, έχουν δημιουργηθεί CDSS με σκοπό τη μείωση των σφαλμάτων κατά τη συνταγογράφηση φαρμάκων.(85)

Ακόμη, η εφαρμογή της μηχανικής μάθησης στην **ανάπτυξη νέων φαρμάκων** επιτρέπει την πιο αποτελεσματική, λιγότερο δαπανηρή και πιο γρήγορη ανάπτυξη νέων φαρμακευτικών προσεγγίσεων, βοηθώντας έτσι στην προώθηση της ιατρικής έρευνας. Οι αλγόριθμοι μηχανικής μάθησης χρησιμοποιούνται για την ανάπτυξη διαφόρων μοντέλων πρόβλεψης των χημικών, βιολογικών και φυσικών χαρακτηριστικών των ενώσεων (compounds) στην ανακάλυψη φαρμάκων και μπορούν να φανούν χρήσιμοι σε όλα τα στάδια της διαδικασίας ανάπτυξης νέων φαρμάκων.(86) Η σχεδίαση φαρμάκων μέσω υπολογιστών (Computer-Aided Drug Design, CADD) είναι μια διαδικασία η οποία χρησιμοποιεί υπολογιστικές τεχνικές για την ανάπτυξη νέων φαρμάκων ή την βελτίωση των υπαρχόντων και συνδυάζει την επιστήμη της βιολογίας με αυτή των υπολογιστών. Ο σχεδιασμός φαρμάκων με τη χρήση της διαδικασίας CADD παρέχει τη δυνατότητα εικονικής εξέτασης εκατομμυρίων ενώσεων. Αυτό μειώνει σημαντικά τον αριθμό των μορίων που πρέπει να δοκιμαστούν βιοχημικά. Ως εκ τούτου, η προσέγγιση αυτή μπορεί να μειώσει το κόστος και να επιταχύνει το προκαταρκτικό στάδιο της ανάπτυξης φαρμάκων.(87) Επίσης, η μηχανική μάθηση συμβάλλει στην επιλογή κατάλληλων βιοδεικτών (biomarkers) οι οποίοι θα είναι ασφαλείς και αποτελεσματικοί για χρήση στην κλινική έρευνα και την αξιολόγηση της ασφάλειας των φαρμάκων.(88) Τέλος, η μηχανική μάθηση μπορεί να βοηθήσει στην κατανόηση της φαρμάκοκινητικής συμπεριφοράς των ενώσεων προβλέποντας την απορρόφηση, την κατανομή, τον μεταβολισμό, την απέκκριση και την τοξικότητα τους.(89)

2.2.2 Κατηγορίες μηχανικής μάθησης

Τα μοντέλα μηχανικής μάθησης διακρίνονται σε κάποιες βασικές κατηγορίες. Συγκεκριμένα, οι 3 κύριες κατηγορίες είναι η επιβλεπόμενη μάθηση, η μη επιβλεπόμενη μάθηση και η ενισχυτική μάθηση.

2.2.2.1 Επιβλεπόμενη μάθηση (Supervised learning)

Η επιβλεπόμενη μάθηση χρησιμοποιείται για την πραγματοποίηση προβλέψεων σχετικά με μελλοντικές περιπτώσεις (νέα δεδομένα), μέσω της δημιουργίας μιας συνάρτησης από ένα σύνολο γνωστών εισόδων (input) και εξόδων (output) το οποίο ονομάζεται σύνολο εκπαίδευσης (training set). (90)

Τα δεδομένα εκπαίδευσης που εισάγονται στο μοντέλο, αποτελούνται από ένα σύνολο εγγραφών (instances). Πρωταρχικός στόχος στην επιβλεπόμενη μάθηση είναι η ικανότητα του μοντέλου να γενικεύει, να μπορεί δηλαδή να αντιστοιχίζει τις νέες εισόδους με τη μέγιστη δυνατή ακρίβεια στις αντίστοιχες εξόδους.(91)

Η ύπαρξη ευρείας ποικιλίας αλγόριθμων επιβλεπόμενης μάθησης καθιστά δυνατή την επίλυση περίπλοκων προβλημάτων. Ανάλογα με τη φύση του προβλήματος, γίνεται η επιλογή του καταλληλότερου αλγόριθμου για το εκάστοτε πρόβλημα. Κάποιοι από τους γνωστότερους αλγόριθμους επιβλεπόμενης μάθησης είναι οι εξής:

- Δέντρα Απόφασης (Decision Trees, DT)
- Κ-Κοντινότεροι γείτονες (K-Nearest Neighbors, KNN)
- Λογιστική παλινδρόμηση (Logistic Regression)
- Τεχνητά νευρωνικά δίκτυα (Artificial Neural Networks, ANN)
- Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines, SVM)
- Γραμμική παλινδρόμηση (Linear Regression)
- Τυχαίο δάσος (Random Forest, RF)

Η επιβλεπόμενη μάθηση διακρίνεται σε δυο μεγάλες κατηγορίες, την κατηγοριοποίηση (Classification) και την παλινδρόμηση (Regression).

Κατηγοριοποίηση

Η κατηγοριοποίηση είναι μια δημοφιλής μέθοδος της μηχανικής μάθησης, η οποία έχει ως απώτερο στόχο την ταξινόμηση δεδομένων σε διαφορετικές κατηγορίες-κλάσεις (classes) μέσω αναγνώρισης μοτίβων που εντοπίζονται στα χαρακτηριστικά (attributes). Μπορεί να διακριθεί σε δυαδική κατηγοριοποίηση (Binary Classification), όπου ο αλγόριθμος ταξινομεί τα νέα δεδομένα-εισόδους σε μια εκ των δύο κατηγοριών και σε πολλαπλή (Multiclass Classification), όπου η ταξινόμηση γίνεται σε περισσότερες από δύο κατηγορίες.(92) Αρχικά, το μοντέλο εκπαιδεύεται χρησιμοποιώντας ένα σύνολο δεδομένων οπού οι έξοδοι-κλάσεις είναι γνωστές. Κατόπιν, μετά το πέρας της εκπαίδευσης, ακολουθεί ο έλεγχος της ακρίβειας του μοντέλου τροφοδοτώντας το με ένα νέο σύνολο άγνωστων δεδομένων-εισόδων, όπου το μοντέλο καλείται να ταξινομήσει (test set).(93)

Παλινδρόμηση

Η παλινδρόμηση είναι μια μέθοδος πρόβλεψης, η οποία χρησιμοποιείται για να προβλέψουμε για τα δεδομένα εισόδου την πιθανή τιμή σε μια συνεχή έξοδο. Ουσιαστικά, στη μέθοδο αυτή γίνεται προσπάθεια κατανόησης της σχέσης μεταξύ μιας ή περισσότερων ανεξάρτητων μεταβλητών και μιας εξαρτημένης μεταβλητής που είναι συνεχής. Εφόσον ο αλγόριθμος παλινδρόμησης εκπαιδευτεί στο σύνολο δεδομένων εκπαίδευσης, το μοντέλο θα μπορεί να προβλέψει τιμές για τις εξαρτημένες μεταβλητές με βάση τα νέα δεδομένα. (94)

2.2.2.2 Μη επιβλεπόμενη μάθηση (Unsupervised learning)

Στην μη επιβλεπόμενη μάθηση ο αλγόριθμος κατασκευάζει ένα μοντέλο για κάποιο σύνολο εισόδων υπό τη μορφή παρατηρήσεων χωρίς να γνωρίζει τις επιθυμητές εξόδους. Δηλαδή, δεν υπάρχουν προκαθορισμένες κατηγορίες οι οποίες να υποδεικνύουν τις επιθυμητές σχέσεις μεταξύ των δεδομένων. Αναγνωρίζονται μοτίβα και σχέσεις στα δεδομένα εκπαίδευσης, χωρίς αυτά να έχουν κάποια ετικέτα (unlabeled data). Εφόσον τα δεδομένα δεν περιέχουν πληροφορίες για την έξοδο, δεν μπορεί να πραγματοποιηθεί αξιολόγηση της απόδοση του αλγορίθμου, όπως γίνεται στην επιβλεπόμενη μάθηση.(91)(92)

Κάποιοι από τους αλγόριθμους που ανήκουν στην μη επιβλεπόμενη μάθηση είναι ο K-Means, ο Density-Based Spatial Clustering of Applications with Noise (DBSCAN) και ο Balanced Iterative Reducing and Clustering Hierarchies (BIRCH).

Στην κατηγορία της μη επιβλεπόμενης μάθησης ανήκει η μέθοδος της συσταδοποίησης (Clustering) ή ομαδοποίησης.

Συσταδοποίηση

Η συσταδοποίηση είναι μια μέθοδος μη επιβλεπόμενης μάθησης. Κατά την συσταδοποίηση, τα δεδομένα οργανώνονται σε ομάδες βάση των κοινών χαρακτηριστικών τους. Οι ομάδες αυτές δεν είναι καθορισμένες από πριν. Οι ομάδες που θα δημιουργηθούν θα πρέπει να περιέχουν αντικείμενα τα οποία είναι περισσότερο όμοια μεταξύ τους και διαφορετικά με τα αντικείμενα από τις υπόλοιπες ομάδες. Στις ομάδες που δημιουργούνται θέλουμε τα δεδομένα να διαχωρίζονται όσο το δυνατόν πιο ορθά. Με αυτή την τεχνική μπορούμε όχι μόνο να εξερευνήσουμε τη δομή των δεδομένων αλλά και να εξάγουμε σημαντικές πληροφορίες χωρίς την καθοδήγηση ενός γνωστού αποτελέσματος.

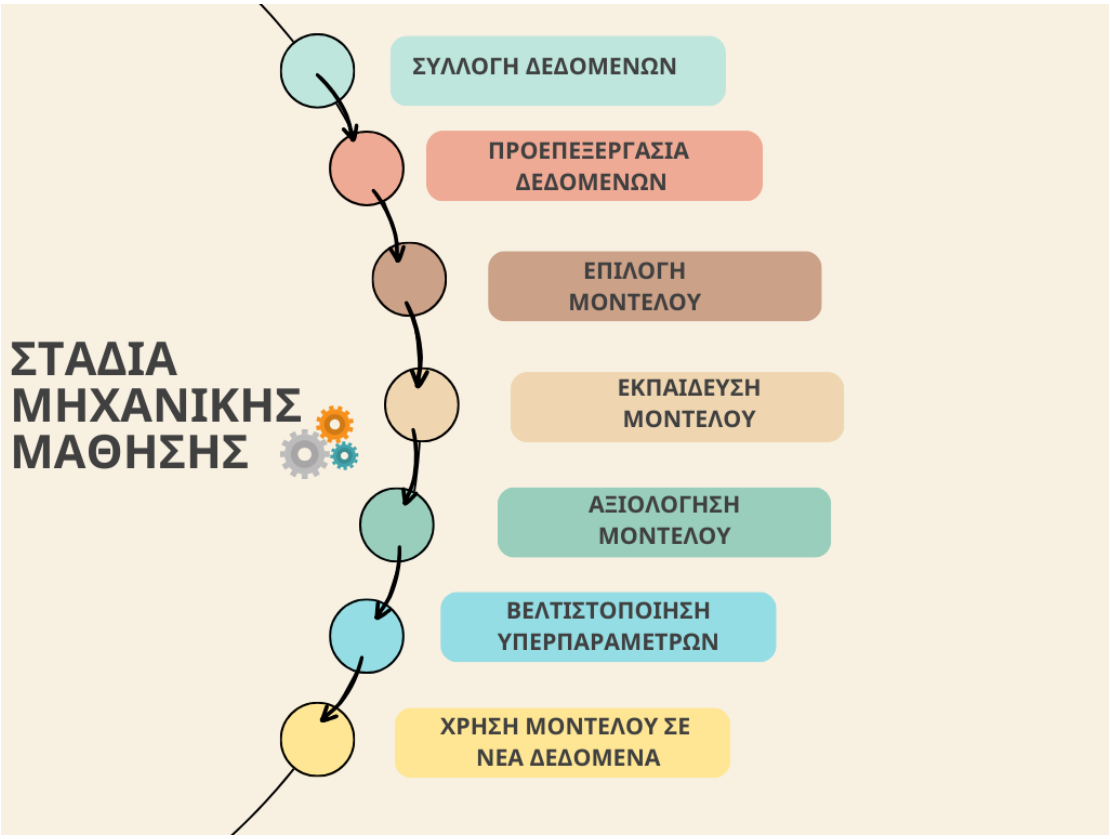
2.2.2.3 Ενισχυτική μάθηση (Reinforcement learning)

Κατά την ενισχυτική μάθηση, ένας πράκτορας (agent) αλληλεπιδρά μέσα σε ένα δυναμικό περιβάλλον (environment) προσπαθώντας να εκτελέσει με τον καλύτερο δυνατό τρόπο μια συγκεκριμένη εργασία. Ο πράκτορας λαμβάνει αποφάσεις βάσει μιας στρατηγικής (policy) και λαμβάνει αποτελέσματα (ανταμοιβές ή τιμωρίες) για τις ενέργειες (actions) του. Υπάρχει μια συνεχής αλληλεπίδραση μεταξύ του πράκτορα και του περιβάλλοντος και έτσι ο πράκτορας μαθαίνει εμπειρικά μέσω της επιβράβευσης ή της τιμωρίας. Στόχος είναι η εύρεση των ενεργειών που οδηγούν στα καλύτερα αποτελέσματα, δηλαδή στην μεγιστοποίηση της συνολικής ανταμοιβής.(70)

2.2.3 Στάδια μηχανικής μάθησης

Η διαδικασία της μηχανικής μάθησης περιλαμβάνει κάποια βασικά στάδια τα οποία είναι απαραίτητα για την δημιουργία ενός αξιόπιστου μοντέλου. Η σειρά των σταδίων αυτών μπορεί να διαφέρει ανάλογα με τις ανάγκες και τις ιδιαιτερότητες του προβλήματος. Επίσης, κάποια στάδια μπορεί να επαναληφθούν όπως για παράδειγμα το στάδιο της βελτιστοποίησης υπερπαραμέτρων και της εκπαίδευσης του μοντέλου. Πριν την έναρξη της διαδικασίας γίνεται ο καθορισμός του προβλήματος προς επίλυση και αποσαφηνίζονται οι στόχοι που πρέπει να επιτευχθούν. Τα βασικά στάδια (Εικόνα 10) της μηχανικής μάθησης είναι τα εξής(95):

1. Συλλογή δεδομένων
2. Προεπεξεργασία δεδομένων
3. Επιλογή μοντέλου
4. Εκπαίδευση μοντέλου
5. Αξιολόγηση μοντέλου
6. Βελτιστοποίηση υπερπαραμέτρων
7. Χρήση μοντέλου σε νέα δεδομένα



Εικόνα 10: Στάδια μηχανικής μάθησης.

[Πηγή: Δημιουργήθηκε από την συγγραφέα με τη χρήση του https://www.canva.com/]

2.2.3.1 Συλλογή δεδομένων

Η συλλογή των δεδομένων αποτελεί το πρώτο βήμα της διαδικασίας της μηχανικής μάθησης. Ακατέργαστα δεδομένα συλλέγονται για το προς επίλυση πρόβλημα. Τα δεδομένα αυτά μπορούν να συλλεχθούν από διάφορες πηγές όπως για παράδειγμα ερωτηματολόγια, βάσεις δεδομένων, κοινωνικά δίκτυα, το διαδίκτυο, αισθητήρες και συσκευές. Τα δεδομένα πρέπει να είναι αξιόπιστα ώστε το μοντέλο μηχανικής μάθησης που θα δημιουργηθεί να μπορέσει να εντοπίσει σωστά μοτίβα και να είναι αποτελεσματικό. Η ποιότητα και η ποσότητα των δεδομένων κρίνεται εξαιρετικά σημαντική, καθώς αυτή θα καθορίσει σε μεγάλο βαθμό την ακρίβεια του μοντέλου.(71)

2.2.3.2 Προεπεξεργασία δεδομένων

Τα ακατέργαστα δεδομένα (raw data) που έχουν συλλεχθεί από τις διάφορες πήγες συνήθως περιέχουν θόρυβο και μπορεί να είναι ελλιπή, ή να υπάρχουν διπλότυπες εγγραφές. Επίσης, συχνά τα δεδομένα πρέπει να μετατραπούν σε άλλη μορφή για να είναι χρήσιμα στη διαδικασία της μηχανικής μάθησης. Δεν

είναι αναγκαίο να χρησιμοποιηθεί το σύνολο των δεδομένων που έχουν συλλεχθεί. Μπορούν να επιλεχθούν συγκεκριμένα χαρακτηριστικά (feature selection) από το σύνολο δεδομένων, με τη χρήση ανάλογων τεχνικών, τα οποία θα αναλυθούν από το μοντέλο μηχανικής μάθησης, καθώς δεν είναι πάντοτε όλα χρήσιμα για το προς επίλυση πρόβλημα. Αντιθέτως, χαρακτηριστικά τα οποία δεν προσφέρουν κάποια γνώση κατά τη δημιουργία του μοντέλου, συντελούν στο να είναι λιγότερο αποτελεσματικό. Ο καθαρισμός και η προεπεξεργασία των δεδομένων κρίνεται ως ένα από τα βασικότερα στάδια για τη δημιουργία ενός ισχυρού μοντέλου μηχανικής μάθησης. Μετά το πέρας της προεπεξεργασίας προκύπτει ένα νέο σύνολο δεδομένων. (71)

Πριν την επιλογή του αλγόριθμου, γίνεται ο διαχωρισμός των δεδομένων και έτσι το νέο σύνολο δεδομένων διαιρείται σε υποσύνολα. Ως αποτέλεσμα, προκύπτει ένα σύνολο για την εκπαίδευση του μοντέλου το οποίο ονομάζεται "σύνολο εκπαίδευσης" και χρησιμοποιείται στο training set και ένα σύνολο για τον έλεγχο της απόδοσης του μοντέλου το οποίο ονομάζεται "σύνολο ελέγχου" και χρησιμοποιείται στο test set.(92)

2.2.3.3 Επιλογή μοντέλου

Εν συνεχεία, γίνεται η επιλογή του κατάλληλου μοντέλου (αλγόριθμου) μηχανικής μάθησης η οποία καθορίζεται από τη φύση του προς επίλυση προβλήματος και τη μορφή των πλέον καθαρών δεδομένων. Επί παραδείγματι, αν το πρόβλημα προς επίλυση είναι ένα πρόβλημα κατηγοριοποίησης, θα γινόταν η κατάλληλη επιλογή μεταξύ των διάφορων αλγορίθμων κατηγοριοποίησης πχ SVM, DT, κ.α. Επιπλέον, ανάλογα με την επιλογή του μοντέλου, καθορίζονται διάφορες υπερπαράμετροι όπως για παράδειγμα στον SVM η συνάρτηση πυρήνα (Kernel), η παράμετρος κόστους C, κλπ.

2.2.3.4 Εκπαίδευση μοντέλου

Το επόμενο στάδιο είναι η εκπαίδευση του μοντέλου μηχανικής μάθησης. Στο στάδιο αυτό χρησιμοποιείται το σύνολο εκπαίδευσης. Στόχος είναι η κατανόηση των δεδομένων και των διάφορων μοτίβων και κανόνων, ώστε το μοντέλο να παρουσιάζει υψηλή ικανότητα γενίκευσης σε νέα δεδομένα, που δεν έχουν χρησιμοποιηθεί κατά την εκπαίδευση.

2.2.3.5 Αξιολόγηση μοντέλου

Για την αξιολόγηση του μοντέλου δεν είναι δυνατό να χρησιμοποιηθούν τα δεδομένα που τροφοδοτήθηκαν κατά τη διάρκεια της εκπαίδευσης του μοντέλου μηχανικής μάθησης. Αυτό οφείλεται στο γεγονός ότι το μοντέλο το οποίο δημιουργείται, έχει στη μνήμη του ολόκληρο το σύνολο εκπαίδευσης. Συνεπώς, θα κατηγοριοποιεί σωστά τα δεδομένα από το σύνολο εκπαίδευσης, με αποτέλεσμα να μην παρέχονται πληροφορίες για την αξιολόγηση του και την ικανότητα γενίκευσης, δηλαδή αν θα έχει καλή απόδοση σε νέα δεδομένα.(92) Έτσι, μπορεί να χρησιμοποιηθεί ένα νέο σύνολο δεδομένων τα οποία θα είναι αθέατα για το μοντέλο που δημιουργήθηκε. Τα δεδομένα αυτά χρησιμοποιούνται αποκλειστικά για να διαπιστωθεί η απόδοση του μοντέλου που έχει ήδη εκπαιδευτεί ώστε να ελεγχθεί η ικανότητα γενίκευσης του, δηλαδή πόσο καλά αποδίδει σε δεδομένα που δεν έχει ξαναδεί.

Κατά την αξιολόγηση του μοντέλου μηχανικής μάθησης που δημιουργήθηκε η οποία έπεται της εκπαίδευσης του, χρησιμοποιούνται διάφορες μετρικές απόδοσης (metrics) με σκοπό να ποσοτικοποιηθεί η ακρίβεια, η αξιοπιστία και η αποτελεσματικότητα του. Κάποιες από αυτές είναι ο Πίνακας Σύγχυσης (Confusion Matrix), η ακρίβεια (Accuracy), η ακρίβεια (Precision), η Ευαισθησία-Ανάκληση (Recall) , η Ειδικότητα (Specificity), η Βαθμολογία F-Measure και η περιοχή κάτω από την καμπύλη ROC (Area Under Curve ROC, AUC_ROC)(96)

Αν τα αποτελέσματα που προκύψουν δεν είναι ικανοποιητικά, τότε μπορεί να επαναληφθεί το στάδιο της εκπαίδευσης με διαφορετικές παραμετροποιήσεις.

Πίνακας Σύγχυσης (Confusion Matrix)

Ο πίνακας σύγχυσης είναι χρήσιμος για προβλήματα κατηγοριοποίησης, όπου η έξοδος μπορεί να αποτελείται από δύο ή περισσότερες κλάσεις. Αποτελεί μια συνοπτική απεικόνιση του αριθμού των σωστών και λανθασμένων ταξινομήσεων σε κάθε κατηγορία με σκοπό την αξιολόγηση του μοντέλου μηχανικής μάθησης. Οι τέσσερις τιμές στον πίνακα σύγχυσης δηλαδή αληθώς θετικά (TP), ψευδώς θετικά (FP), αληθώς αρνητικά (TN) και ψευδώς αρνητικά (FN) μπορούν να χρησιμοποιηθούν για τον υπολογισμό διάφορων μετρικών απόδοσης των μοντέλων μηχανικής μάθησης, όπως η ακρίβεια, η βαθμολογία F1 και η ευαισθησία-ανάκληση.(97) Όταν το προς επίλυση πρόβλημα ταξινόμησης απαιτεί ταξινόμηση δύο κλάσεων ο πίνακας σύγχυσης θα είναι 2 x 2, ενώ σε πρόβλημα ταξινόμησης N κλάσεων τότε οι διαστάσεις του πίνακα σύγχυσης θα είναι N x N.

- Όπου True Positive (TP) για την πρόβλεψη που αντιστοιχεί στην θετική κλάση που ταξινομήθηκε σωστά.
- True Negative (TN) για την πρόβλεψη που αντιστοιχεί στην αρνητική κλάση και έχει ταξινομηθεί σωστά.

- False Positive (FP) για την πρόβλεψη που αντιστοιχεί στην αρνητική κλάση αλλά λανθασμένα ταξινομείται στην θετική.
- False Negative (FN) για την πρόβλεψη που αντιστοιχεί στην θετική κλάση αλλά λανθασμένα ταξινομείται στην αρνητική.

Ακρίβεια (Accuracy)

Η ακρίβεια (Accuracy) είναι μια βασική μετρική η οποία αφορά τον αριθμό των σωστών προβλέψεων στο σύνολο των προβλέψεων που πραγματοποιήθηκαν. Λαμβάνει τιμές 0 ≤ Accuracy ≤ 1 και υπολογίζεται σύμφωνα με την παρακάτω εξίσωση:(97)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{\Sigma ωστές\ προβλέψεις}{\Sigma ύνολο\ όλων\ των\ προβλέψεων}$$

Ακρίβεια (Precision)

Η μετρική (Precision) αφορά τον αριθμό των πραγματικά θετικών προβλέψεων που αναγνωρίστηκαν ως θετικές προς τον αριθμό όλων των προβλέψεων που αναγνωρίστηκαν ως θετικές. Υπολογίζεται από τον παρακάτω τύπο(98):

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

Ευαισθησία-Ανάκληση (Recall)

Η ευαισθησία-ανάκληση μετράει το πόσο καλά μπορεί το μοντέλο μηχανικής μάθησης που δημιουργήθηκε να αναγνωρίσει όλες τις θετικές τιμές σε ένα σύνολο δεδομένων. Ουσιαστικά, αντικατοπτρίζει το ποσοστό των πραγματικά θετικών τιμών που προβλέφθηκαν σωστά από το μοντέλο. Όσο μεγαλύτερη η τιμή, τόσο λιγότερα θετικά παραδείγματα έχουν ταξινομηθεί λάθος. Υπολογίζεται από τον τύπο:(97)

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

Ειδικότητα (Specificity)

Η ειδικότητα είναι μια μετρική που αξιολογεί την ικανότητα ενός ταξινομητή να αναγνωρίζει σωστά τις αρνητικές περιπτώσεις και υπολογίζεται από τον τύπο:(99)

$$Specificity = \frac{TN}{TN + FP}$$

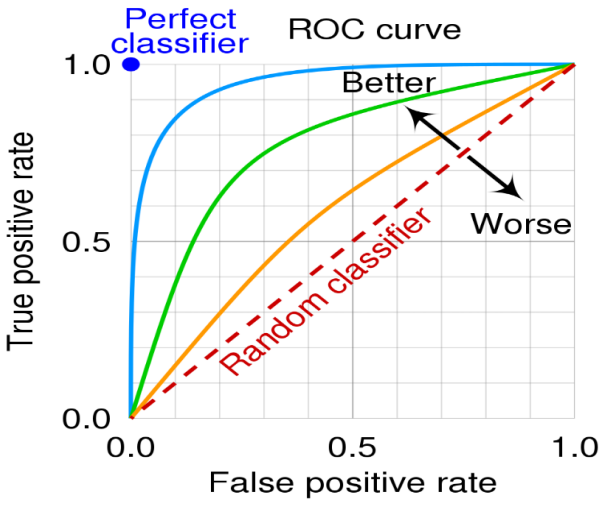
Βαθμολογία F-Measure

Το F-Measure μπορεί να κυμανθεί από 0 έως 1 και οι υψηλότερες τιμές υποδεικνύουν καλύτερη συνολική απόδοση του μοντέλου μηχανικής μάθησης. Καθίσταται ιδιαίτερα χρήσιμο όταν υπάρχει ανισορροπία μεταξύ των κλάσεων και συνεπώς πρέπει να ληφθεί υπόψιν τόσο η ακρίβεια όσο και η ανάκληση. Ως F-Measure ορίζεται ο αρμονικός μέσος της ακρίβειας και της ανάκλησης του και υπολογίζεται με τον παρακάτω τύπο:(100)

$$F - Measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Περιοχή κάτω από την καμπύλη ROC (Area Under Curve ROC, AUC_ROC)

Η καμπύλη ROC (Receiver Operating characteristic Curve) είναι ένα γράφημα (Εικόνα 11) που απεικονίζει το True Positive Rate (TPR) σε συνάρτηση με το False Positive Rate (FPR) για όλα τα πιθανά κατώφλια (threshold) ταξινόμησης. Έτσι, μπορεί να γίνει εκτίμηση της επίδοσης του μοντέλου ταξινόμησης.(97)(101)



Εικόνα 11: Καμπύλη ROC.

[Πηγή:https://www.researchgate.net/figure/Receiver-Operating-Characteristic-ROC-Curve_fig2_363218385]

Το TPR γνωστό και ως ευαισθησία (Sensitivity) είναι συνώνυμο του Recall και υπολογίζεται από τον τύπο: $TPR = \frac{TP}{TP+FN}$

Το FPR είναι το κλάσμα των ψευδώς θετικών επί του συνόλου των αρνητικών δειγμάτων και υπολογίζεται ως εξής: $FPR = \frac{FP}{FP+TN}$ (92)

Η περιοχή κάτω από την καμπύλη ROC, αξιολογεί την ικανότητα του μοντέλου να διαχωρίζει μεταξύ των κλάσεων σε ένα πρόβλημα δυαδικής ταξινόμησης (binary classification). Όσο μεγαλύτερο είναι το εμβαδόν που αντιστοιχεί στη περιοχή κάτω από την σχηματιζόμενη καμπύλη τόσο καλύτερη απόδοση εμφανίζει το μοντέλο πρόβλεψης. Όταν το AUC-ROC είναι ίσο με 1, αυτό υποδεικνύει τέλεια διάκριση μεταξύ των κλάσεων, ενώ ένα AUC-ROC ίσο με 0.5 υποδεικνύει τυχαία ταξινόμηση. Συνεπώς, όσο πιο κοντά στο 1 βρίσκεται το AUC-ROC, τόσο καλύτερη είναι και η απόδοση του μοντέλου.(102)

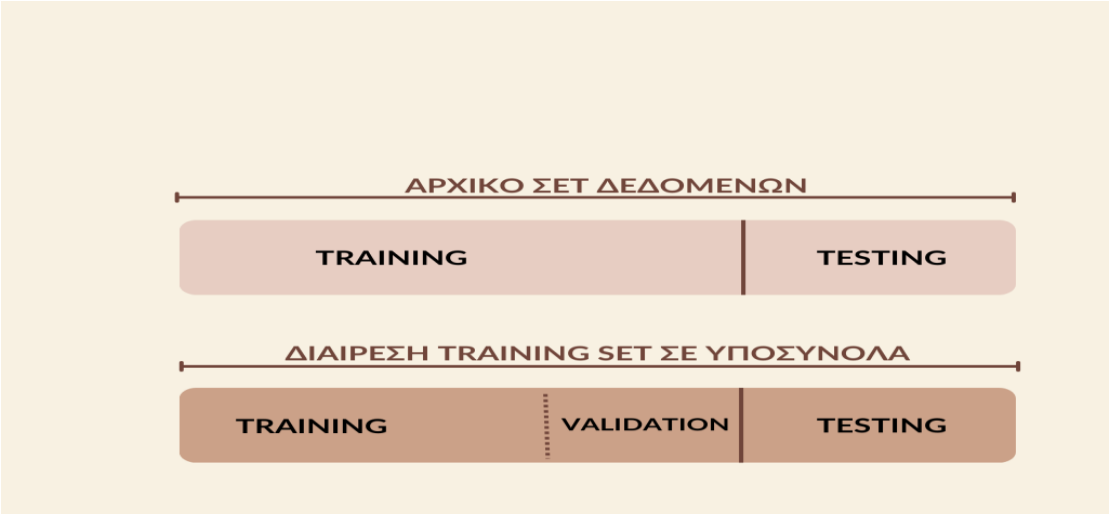
2.2.3.6 Βελτιστοποίηση υπερπαραμέτρων

Μετά την αξιολόγηση του μοντέλου, ενδέχεται να χρειαστεί να προσαρμοστούν οι υπερπαράμετροι του για να βελτιωθεί η απόδοσή του. Το στάδιο της βελτιστοποίησης υπερπαραμέτρων (hyperparameter optimization) αφορά την επιλογή των καλύτερων τιμών για τις υπερπαραμέτρους του μοντέλου. Είναι ιδιαίτερα σημαντικό καθώς οι λανθασμένες επιλογές παραμετροποίησης μπορεί να οδηγήσουν σε ένα μη αποδοτικό μοντέλο προκαλώντας για παράδειγμα υπερπροσαρμογή (overfitting) ή υποπροσαρμογή (underfitting) του μοντέλου στα δεδομένα. Η υπερπροσαρμογή είναι ένα συχνά παρατηρούμενο ζήτημα στην επιβλεπόμενη μηχανική μάθηση. Η υπερπροσαρμογή εμποδίζει τα μοντέλα που δημιουργούνται να γενικεύσουν σε αθέατα νέα δεδομένα, καθώς τα μοντέλα έχουν εκπαιδευτεί με τα διαθέσιμα δεδομένα σε τέτοιο βαθμό, που έχουν υπερεκπαιδευτεί στα δεδομένα εκπαίδευσης.(103) Αυτό σημαίνει ότι το μοντέλο τη συγκεκριμένη χρονική στιγμή κατά τη διάρκεια της περιόδου εκπαίδευσης δεν βελτιώνει πλέον την ικανότητά του να επιλύει το πρόβλημα.(104) Λόγω της ύπαρξης υπερπροσαρμογής, το μοντέλο αποδίδει τέλεια στο training set ενώ εμφανίζει κακή απόδοση στο test set. Η υπερπροσαρμογή μπορεί να οφείλεται στο γεγονός ότι το training set είναι πολύ μικρό σε μέγεθος ή περιέχει λιγότερο αντιπροσωπευτικά δεδομένα ή ιδιαίτερα θορυβώδη δεδομένα. Επιπλέον, η πολυπλοκότητα της μαθηματικής συνάρτησης (υπόθεσης) που χρησιμοποιείται από ένα μοντέλο μηχανικής μάθησης για να προβλέψει τα δεδομένα, μπορεί να επηρεάσει το φαινόμενο του overfitting. Ο στόχος είναι να βρεθεί μια "χρυσή τομή" όπου το μοντέλο είναι αρκετά πολύπλοκο ώστε να προβλέπει σωστά, αλλά όχι τόσο

πολύπλοκο ώστε να προσαρμόζεται στο θόρυβο των δεδομένων εκπαίδευσης.(103)

Η μέθοδος **Grid Search** η οποία είναι η πιο διαδεδομένη, χρησιμοποιείται κατά την αναζήτηση της βέλτιστης τιμής της κάθε υπερπαραμέτρου, μέσα από ένα πλέγμα τιμών (**grid**), το οποίο περιέχει όλα τα σύνολα των τιμών των διαφορετικών υπερπαραμέτρων. Ουσιαστικά, πραγματοποιείται αξιολόγηση της απόδοσης του μοντέλου για κάθε συνδυασμό υπερπαραμέτρων στο πλέγμα ώστε να γίνει επιλογή του συνδυασμού που παρουσιάζει την καλύτερη απόδοση. Έπειτα, το σύστημα πρέπει να δοκιμαστεί σε δεδομένα, τα οποία δεν έχουν χρησιμοποιηθεί κατά την εκπαίδευση. Για κάθε συνδυασμό παραμέτρων που δημιουργείται, το μοντέλο θα εκπαιδευτεί στο training set και η απόδοση του θα δοκιμαστεί σε ένα άλλο σετ. Οι τρόποι με τους οποίους μπορεί να επιτευχθεί αυτό, είναι είτε με τη μέθοδο Cross-validation, είτε με την χρήση του Validation set. (92)

Η διασταυρούμενη επικύρωση (Cross-validation) είναι μια μέθοδος που αξιοποιεί με βέλτιστο τρόπο τα διαθέσιμα δεδομένα καθώς γίνεται επαναλαμβανόμενη εκπαίδευση και έλεγχος του μοντέλου σε διαφορετικά υποσύνολα δεδομένων. Συνήθως χρησιμοποιείται ένα μεγάλο σύνολο εκπαίδευσης και ένα μικρό σύνολο επικύρωσης σε κάθε επανάληψη. Τα δεδομένα διαιρούνται σε k υποσύνολα, συνήθως 5 έως 10. Η εκπαίδευση γίνεται σε k-1 υποσύνολα, ενώ η αξιολόγηση στο 1 υπολειπόμενο. Για παράδειγμα αν οριστούν 5 υποσύνολα, το 80 τοις εκατό των δεδομένων θα χρησιμοποιηθεί για την εκπαίδευση ενώ το 20 τοις εκατό για την επικύρωση. Στην επόμενη επανάληψη θα επιλεχθεί ένα άλλο 20 τοις εκατό των δεδομένων για επικύρωση. Αυτή η διαδικασία θα επαναληφθεί πέντε φορές έως ότου όλα τα δεδομένα να έχουν χρησιμεύσει μία φορά ως δεδομένα επικύρωσης. Το μειονέκτημα αυτής της μεθόδου είναι η δαπάνη χρόνου και υπολογιστικής ισχύος. (105)(106)



Εικόνα 12: Διαίρεση του training set σε δυο υποσύνολα

[Πηγή: Δημιουργήθηκε από την συγγραφέα με τη χρήση του https://www.canva.com/]

Στην περίπτωση όπου θα χρησιμοποιηθεί η διαδικασία με το Validation set, θα πρέπει το training set να διαχωριστεί σε δύο υποσύνολα, ένα νέο training set και το validation set όπως φαίνεται στην Εικόνα 12.

Επομένως, για τους συνδυασμούς παραμέτρων που θα δημιουργηθούν, το μοντέλο εκπαιδεύεται στο νέο training set το οποίο είναι σαφώς μικρότερο από το αρχικό και δοκιμάζεται η απόδοση του στο validation set. Έπειτα, επιλέγεται για το μοντέλο ο συνδυασμός παραμέτρων που επιτυγχάνει την καλύτερη απόδοση στο validation set. Τελικά, το μοντέλο θα εκπαιδευτεί στο αρχικό training set (δηλαδή το training set πριν το διαχωρισμό) και η απόδοση του μοντέλου θα ελεγχθεί στο test set.(92)

2.2.3.7 Χρήση μοντέλου σε νέα δεδομένα

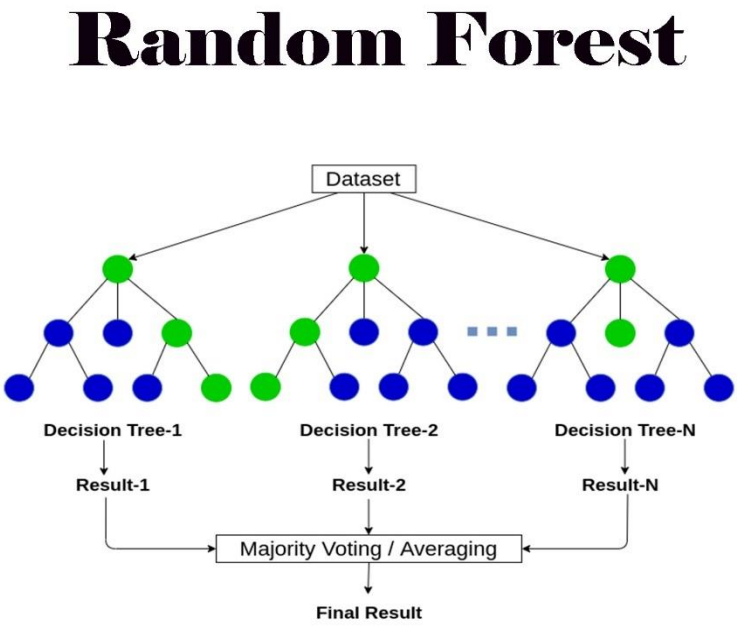
Εφόσον ολοκληρωθούν όλα τα παραπάνω στάδια, το τελικό μοντέλο είναι έτοιμο να χρησιμοποιηθεί σε νέα δεδομένα με άγνωστες εξόδους για την επίλυση του προβλήματος για το οποίο αναπτύχθηκε. Κατά την εξωτερική αξιολόγηση (external validation), το μοντέλο εφαρμόζεται σε ανεξάρτητα νέα δεδομένα που δεν έχουν χρησιμοποιηθεί κατά την εκπαίδευση. Αυτό είναι ένα κρίσιμο στάδιο που επιτρέπει την αξιολόγηση της απόδοσης και της ικανότητας γενίκευσης του μοντέλου. Εφόσον η απόδοση του μοντέλου παραμένει καλή σε αυτά τα νέα δεδομένα, τότε είναι πιθανό ότι θα είναι αποτελεσματικό και σε πραγματικές καταστάσεις.

2.2.4 Μοντέλα μηχανικής μάθησης

2.2.4.1 Random Forest (RF)

Το Τυχαίο δάσος (Random Forest-RF) αποτελεί μια μέθοδο εποπτευόμενης μηχανικής μάθησης και χρησιμοποιείται για προβλήματα ταξινόμησης και παλινδρόμησης. Ο αλγόριθμος αυτός κατασκευάζει ένα σύνολο τυχαίων δέντρων αποφάσεων. Κάθε δέντρο απόφασης κατασκευάζεται με τη χρήση της τεχνικής της τυχαίας δειγματοληψίας (bootstrap) από τα δεδομένα εκπαίδευσης. Δηλαδή, προκύπτει με τυχαία επιλογή δειγμάτων από τα δεδομένα εκπαίδευσης και τυχαίων χαρακτηριστικών. Με αυτόν τον τρόπο δημιουργούνται δέντρα αποφάσεων τα οποία δεν συσχετίζονται μεταξύ τους, καθώς κάθε δέντρο εστιάζει σε διαφορετικά δεδομένα. Όσον αφορά την απόφαση για την τελική πρόβλεψη για ένα δείγμα, αυτή λαμβάνεται από την πλειοψηφία για τις προβλέψεις κατηγορικών μεταβλητών ενώ λαμβάνεται από

το μέσο όρο των τιμών για τις προβλέψεις των αριθμητικών μεταβλητών.(93)(107)



Εικόνα 13: Random Forest. Γραφική απεικόνιση του σχηματισμού των δέντρων αποφάσεων.

[Πηγή: https://anasbrital98.github.io/blog/2021/Random-Forest/]

Ένα ισχυρό πλεονέκτημα του αλγόριθμου RF είναι ότι αποτελεί μία επέκταση του αλγόριθμου των δέντρων αποφάσεων (Decision Trees, DT), με αποτέλεσμα να αντιστέκεται στο φαινόμενο της υπερπροσαρμογής, διατηρώντας ταυτόχρονα χαμηλό το σφάλμα μεροληψίας (bias). (92)

Οι υπερπαράμετροι που μπορούν να επηρεάσουν την απόδοση του μοντέλου RF που θα δημιουργηθεί είναι οι παρακάτω(108):

- ***n estimators***: η παράμετρος αυτή καθορίζει το πλήθος των δέντρων αποφάσεων που θα δημιουργηθούν στο σύνολο. Ένα μεγαλύτερο πλήθος δέντρων μπορεί να οδηγήσει θεωρητικά σε πιο ακριβείς προβλέψεις. Ωστόσο, αυτό θα έχει ως συνέπεια πιθανά την αύξηση του χρόνου εκπαίδευσης.

- ***max depth***: η παράμετρος που ορίζει το μέγιστο επιτρεπόμενο βάθος που θα έχει κάθε δέντρο απόφασης. Όσο μεγαλύτερο είναι το βάθος του δέντρου, τόσο περισσότεροι θα είναι οι διαχωρισμοί των δεδομένων. Έτσι, μπορεί να επιτευχθεί καλύτερη προσαρμογή στα

δεδομένα εκπαίδευσης, με κίνδυνο όμως ο αλγόριθμος να οδηγηθεί σε υπερπροσαρμογή.

- ***min samples split****:* είναι η παράμετρος που αφορά το ελάχιστο πλήθος των δειγμάτων που απαιτούνται για τον διαχωρισμό ενός εσωτερικού κόμβου στο δέντρο αποφάσεων. Δηλαδή, στην περίπτωση που ένας κόμβος έχει λιγότερα δείγματα από το ορισμένο ελάχιστο πλήθος, τότε αυτός δεν θα διακλαδωθεί περαιτέρω.

- ***min samples leaf***: η συγκεκριμένη παράμετρος καθορίζει τον ελάχιστο αριθμό δειγμάτων που απαιτούνται για ένα φύλλο κατά την δημιουργία των δέντρων απόφασης. Μικρότερες τιμές ελλοχεύουν τον κίνδυνο υπερεκπαίδευσης του μοντέλου.

- ***max features***: αυτή η παράμετρος διαμορφώνει τον μέγιστο αριθμό των χαρακτηριστικών που λαμβάνονται υπόψη κατά τη διαδικασία διακλάδωσης σε κάθε δέντρο του τυχαίου δάσους.

- ***Bootstrap****:* η παράμετρος αυτή προσδιορίζει εάν θα γίνει τυχαία δειγματοληψία με επανάληψη κατά την εκπαίδευση των δέντρων στο τυχαίο δάσος. Σε αυτή την περίπτωση, βελτιώνεται η ικανότητα γενίκευσης του μοντέλου. Αν απενεργοποιηθεί αυτή η παράμετρος, απενεργοποιείται και η επαναλαμβανόμενη δειγματοληψία (bootstrap = False). Έτσι, θα γίνει χρήση όλων των δειγμάτων εκπαίδευσης για κάθε δέντρο.

Ο αλγόριθμος RF παρουσιάζει αρκετά πλεονεκτήματα αλλά και κάποια μειονεκτήματα που παρουσιάζονται παρακάτω:

Πλεονεκτήματα

- Λειτουργεί αρκετά αποδοτικά σε μεγάλα σετ δεδομένων.

- Παρέχει συνήθως υψηλή ακρίβεια στις προβλέψεις που πραγματοποιεί επειδή αυτές προέρχονται από τη δημιουργία πολλών δέντρων αποφάσεων.

- Είναι λιγότερο επιρρεπής στο φαινόμενο της υπερπροσαρμογής σε σχέση με άλλους αλγόριθμους. Αυτό συμβαίνει επειδή κάθε δέντρο εκπαιδεύεται σε ένα υποσύνολο των δεδομένων και χρησιμοποιεί μόνο ένα υποσύνολο των χαρακτηριστικών.

- Μπορεί να λειτουργήσει αποδοτικά σε περιπτώσεις που υπάρχουν αρκετές απουσιάζουσες τιμές (missing values) στα δεδομένα.

- Είναι ευέλικτο τόσο σε προβλήματα ταξινόμησης όσο και σε προβλήματα παλινδρόμησης.

- Λειτουργεί καλά τόσο με κατηγορικές όσο και με συνεχείς τιμές.

Μειονεκτήματα

- Συνήθως απαιτεί μεγάλη υπολογιστική ισχύ για την εφαρμογή του αλγορίθμου.

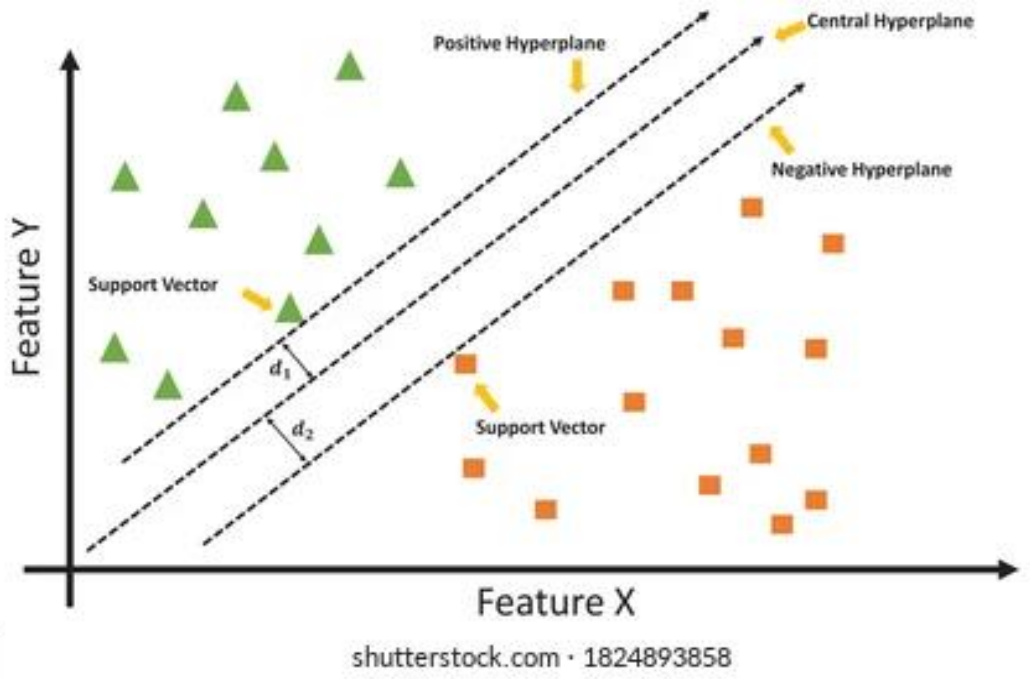
- Η κατασκευή τέτοιων μοντέλων απαιτεί συνήθως πολύ περισσότερο χρόνο και προσπάθεια σε σχέση με τα δέντρα αποφάσεων, ειδικά για μεγάλα σετ δεδομένων ή όταν έχει καθοριστεί μεγάλος αριθμός δέντρων.

- Είναι σχετικά πιο δύσκολη η ερμηνεία του σε σχέση με άλλους αλγόριθμους μηχανικής μάθησης.

2.2.4.2 Support Vector Machines (SVM)

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines-SVM) είναι μία ομάδα αλγορίθμων που ανήκουν στην επιτηρούμενη μηχανική μάθηση. Αρχικά χρησιμοποιήθηκαν σε προβλήματα κατηγοριοποίησης ενώ αργότερα εφαρμόστηκαν και σε προβλήματα παλινδρόμησης.(109)

Η κατηγοριοποίηση των δεδομένων στηρίζεται στην εύρεση ενός βέλτιστου υπερεπιπέδου (hyperplane) που διαχωρίζει τα δεδομένα σε δύο κλάσεις με όσο το δυνατόν μεγαλύτερη ακρίβεια, δημιουργώντας το μέγιστο περιθώριο (margin) μεταξύ των δυο κλάσεων. Ένα μεγαλύτερο περιθώριο οδηγεί σε μικρότερο σφάλμα γενίκευσης. Τα παραδείγματα (instances) του συνόλου εκπαίδευσης που είναι πιο κοντά στο βέλτιστο υπερεπίπεδο ονομάζονται "διανύσματα υποστήριξης" και είναι αυτά που χρησιμοποιούνται για να αποφασιστεί σε ποια κλάση θα ταξινομηθεί η έξοδος, δηλαδή το νέο δείγμα.(93)



Εικόνα 14: Support Vector Machines. Γραφική απεικόνιση του υπερεπιπέδου και των αντίστοιχων διανυσμάτων υποστήριξης .

[Πηγή: https://www.shutterstock.com/el/image-illustration/svm-overview-support-vector-hyperplanes-1824893858.]

Όταν τα δεδομένα εκπαίδευσης είναι γραμμικώς διαχωρίσιμα, τότε διαχωρίζονται με ένα υπερεπίπεδο και κάθε στοιχείο ταξινομείται σε μια πλευρά του υπερεπιπέδου. Ωστόσο σε περιπτώσεις μη γραμμικού διαχωρισμού μπορεί να χρησιμοποιηθεί το τέχνασμα του πυρήνα (kernel trick), όπου με τη χρήση μιας συνάρτησης πυρήνα (kernel function) τα δεδομένα απεικονίζονται σε ένα χώρο μεγαλύτερων διαστάσεων. Αυτό επιτρέπει τη δημιουργία υπερεπιπέδων που μπορούν να διαχωρίσουν μη γραμμικά διαχωρίσιμα δεδομένα. Οι πιο συνηθισμένες **συναρτήσεις πυρήνα** είναι η γραμμική (Linear), η πολυωνυμική (Polynomial), η σιγμοειδής (Sigmoid) και η RBF (Radial Basis Function).(102)

Κάποιες από τις κύριες υπερπαραμέτρους που τροποποιούνται στον SVM και επηρεάζουν την απόδοση και την ικανότητα γενίκευσης του μοντέλου είναι η παράμετρος C (Cost), η παράμετρος γ (gamma), η παράμετρος Degree και η συνάρτηση πυρήνα (kernel) που αναφέρθηκε παραπάνω. Ανάλογα με την συνάρτηση πυρήνα που επιλέγεται, ρυθμίζονται διαφορετικές παράμετροι.(110)

- Η **παράμετρος C** ή αλλιώς παράμετρος κόστους, καθορίζει το όριο μεταξύ της επίτευξης μέγιστου περιθωρίου (margin) και της ελαχιστοποίησης των σφαλμάτων ταξινόμησης. Ένα μεγαλύτερο C οδηγεί σε πιο πολύπλοκα μοντέλα τα οποία προσαρμόζονται καλύτερα

στα δεδομένα εκπαίδευσης, αλλά μπορεί να οδηγήσει στο φαινόμενο της υπερεκπαίδευσης. Αντίθετα, όταν η παράμετρος C έχει μικρότερη τιμή, δημιουργείται μεγαλύτερο περιθώριο το οποίο επιτρέπει περισσότερα λάθη στη ταξινόμηση.(102)

- Η **παράμετρος γ**, ορίζει την επίδραση που έχει το κάθε παράδειγμα της εκπαίδευσης. Μεγαλώνοντας την τιμή σε αυτή τη παράμετρο, το μοντέλο προσπαθεί να προσαρμοστεί όσο καλύτερα μπορεί στα δεδομένα εκπαίδευσης. Αυτό έχει ως αποτέλεσμα την λήψη πιο γενικευμένης απόφασης, αλλά μπορεί να οδηγήσει σε υπερβολική απλότητα του μοντέλου και κατ' επέκταση στο φαινόμενο της υποεκπαίδευσης.(109)

- Η παράμετρος **degree**, αφορά τον βαθμό της πολυωνυμικής συνάρτησης πυρήνα. Όσο αυξάνεται αυτή η παράμετρος, τόσο αυξάνεται και η πολυπλοκότητα του μοντέλου. Σε υψηλότερες τιμές της παραμέτρου το μοντέλο μπορεί να προσαρμοστεί καλύτερα στα δεδομένα εκπαίδευσης. Ωστόσο, μπορεί να γίνει πιο ευαίσθητο στο φαινόμενο της υπερπροσαρμογής.

Ο αλγόριθμος SVM εμφανίζει αρκετά πλεονεκτήματα και κάποια μειονεκτήματα όπως:

Πλεονεκτήματα

- Ο αλγόριθμος SVM παρουσιάζει επαρκή ικανότητα γενίκευσης. Αυτό οφείλεται στην αναζήτηση του βέλτιστου περιθωρίου, προσπαθώντας έτσι να ελαχιστοποιηθεί το σφάλμα πρόβλεψης σε νέα δεδομένα.

- Ο SVM είναι κατάλληλος για σύνολα δεδομένων με υψηλό αριθμό χαρακτηριστικών και μικρότερο αριθμό δειγμάτων όπως για παράδειγμα στην επεξεργασία εικόνας.

- Είναι λιγότερο ευάλωτος σε σχέση με άλλους αλγόριθμους στο φαινόμενο της υπερπροσαρμογής λόγω της προσπάθειας του για εύρεση του μέγιστου περιθωρίου μεταξύ των κλάσεων.

- Μπορεί να χρησιμοποιηθεί για την αντιμετώπιση μη γραμμικών προβλημάτων, με την κατάλληλη επιλογή συναρτήσεων πυρήνα. Οι πυρήνες μπορούν να προσαρμοστούν στις ανάγκες του προβλήματος με την προσαρμογή των παραμέτρων τους. Έτσι, ο αλγόριθμος SVM είναι ευέλικτος και μπορεί να προσαρμοστεί ακόμη και σε πολύπλοκα προβλήματα ταξινόμησης και παλινδρόμησης.

- Μπορεί να αποδώσει καλά και με μικρά σύνολα δεδομένων, καθώς απαιτεί μόνο ένα μικρό αριθμό διανυσμάτων υποστήριξης για να καθορίσει το όριο απόφασης.

Μειονεκτήματα

- Απαιτεί σχετικά μεγάλη υπολογιστική ισχύ και δαπάνη χρόνου για μεγάλα σύνολα δεδομένων και σύνολα δεδομένων με αρκετά χαρακτηριστικά καθώς ο αλγόριθμος SVM πρέπει να υπολογίσει το εσωτερικό γινόμενο μεταξύ κάθε δείγματος.

- Η παραμετροποίηση του κρίνεται αρκετά ευαίσθητη καθώς υπερπαράμετροι όπως η παράμετρος κόστους C και η παράμετρος γ, πρέπει να επιλεγούν προσεκτικά. Σε αντίθετη περίπτωση, οι λανθασμένες επιλογές μπορεί να οδηγήσουν σε χαμηλή απόδοση του μοντέλου.

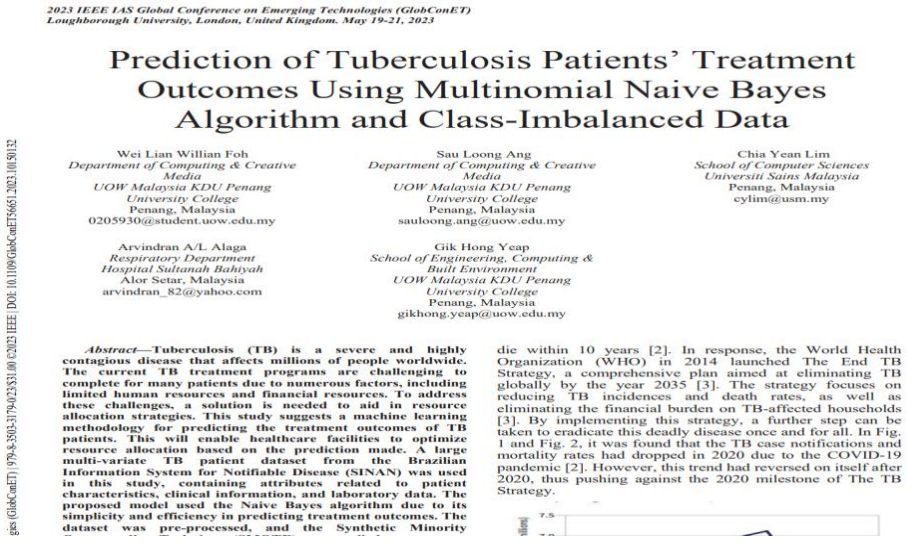
- Δεν είναι κατάλληλος για σύνολα δεδομένων που περιέχουν απουσιάζουσες τιμές (missing values) διότι αυτό μπορεί να οδηγήσει σε λανθασμένες αποφάσεις του μοντέλου. Ωστόσο, αυτό μπορεί να αντιμετωπιστεί στο στάδιο της προεπεξεργασίας των δεδομένων.

3.Μελέτες σχετικές με τη χρήση τεχνικών μηχανικής μάθησης στην καταπολέμηση της φυματίωσης

Η φυματίωση αποτελεί μία από τις παλαιότερες και πιο θανατηφόρες ασθένειες που επηρεάζουν τον άνθρωπο σε παγκόσμιο επίπεδο. Η παγκόσμια κοινότητα σε συνεργασία με τον ΠΟΥ δρουν με στόχο την καταπολέμηση της φυματίωσης, με προσπάθειες για τη βελτίωση της διάγνωσης, της θεραπείας, και της πρόληψης. Η μηχανική μάθηση συνιστά ένα ισχυρό εργαλείο που χρησιμοποιείται ραγδαία τα τελευταία χρόνια στην έρευνα για την αντιμετώπιση της φυματίωσης. Μέχρι σήμερα, αρκετές μελέτες έχουν διεξαχθεί. Κατόπιν αναζήτησης στις βάσεις δεδομένων PubMed και Scopus, στην μηχανή αναζήτησης Google Scholar και στην πλατφόρμα ResearchGate, ανευρέθηκαν ορισμένες μελέτες που παρουσιάζουν παρόμοιο ερευνητικό σκοπό με την παρούσα εργασία. Οι λέξεις κλειδιά που χρησιμοποιήθηκαν ήταν οι εξής: "tuberculosis", "treatment outcome", "prediction" και "machine learning". Παρακάτω αναφέρονται οι 5 πιο συναφείς μελέτες οι οποίες βασίζονται σε κλινικά και εργαστηριακά δεδομένα, καθώς και σε κάποια δημογραφικά χαρακτηριστικά και έχουν ως στόχο την αποτελεσματικότερη προσέγγιση στην θεραπεία της φυματίωσης.

Μελέτη 1

Wei Lian Willian Foh et al. (May 2023)



2023 IEEE IAS Global Conference on Emerging Technologies (GlobConET)
Loughborough University, London, United Kingdom. May 19-21, 2023

Prediction of Tuberculosis Patients' Treatment Outcomes Using Multinomial Naive Bayes Algorithm and Class-Imbalanced Data

Wei Lian Willian Foh
Department of Computing & Creative Media
UOW Malaysia KDU Penang University College
Penang, Malaysia
0205930@student.uow.edu.my

Sau Loong Ang
Department of Computing & Creative Media
UOW Malaysia KDU Penang University College
Penang, Malaysia
sauloong.ang@uow.edu.my

Chia Yean Lim
School of Computer Sciences
Universiti Sains Malaysia
Penang, Malaysia
cylim@usm.my

Arvindran A/L Alaga
Respiratory Department
Hospital Sultanah Bahiyah
Alor Setar, Malaysia
arvindran_82@yahoo.com

Gik Hong Yeap
School of Engineering, Computing & Built Environment
UOW Malaysia KDU Penang University College
Penang, Malaysia
gikhong.yeap@uow.edu.my

Abstract—**Tuberculosis (TB) is a severe and highly contagious disease that affects millions of people worldwide. The current TB treatment programs are challenging to complete for many patients due to numerous factors, including limited human resources and financial resources. To address these challenges, a solution is needed to aid in resource allocation strategies. This study suggests a machine learning methodology for predicting the treatment outcomes of TB patients. This will enable healthcare facilities to optimize resource allocation based on the prediction made. A large multi-variate TB patient dataset from the Brazilian Information System for Notifiable Disease (SINAN) was used in this study, containing attributes related to patient characteristics, clinical information, and laboratory data. The proposed model used the Naive Bayes algorithm due to its simplicity and efficiency in predicting treatment outcomes. The dataset was pre-processed, and the Synthetic Minority**

die within 10 years [2]. In response, the World Health Organization (WHO) in 2014 launched The End TB Strategy, a comprehensive plan aimed at eliminating TB globally by the year 2035 [3]. The strategy focuses on reducing TB incidences and death rates, as well as eliminating the financial burden on TB-affected households [3]. By implementing this strategy, a further step can be taken to eradicate this deadly disease once and for all. In Fig. 1 and Fig. 2, it was found that the TB case notifications and mortality rates had dropped in 2020 due to the COVID-19 pandemic [2]. However, this trend had reversed on itself after 2020, thus pushing against the 2020 milestone of The TB Strategy.

Εικόνα 15: Prediction of Tuberculosis Patients' Treatment Outcomes Using Multinomial Naive Bayes Algorithm and Class-Imbalanced Data.

[Πηγή: Foh WLW, Ang SL, Lim CY, Alaga AAL, Yeap GH. Prediction of Tuberculosis Patients' Treatment Outcomes Using Multinomial Naive Bayes Algorithm and Class-Imbalanced Data. 2023 IEEE IAS Glob Conf Emerg Technol GlobConET 2023. 2023;1–6.]

Στην μελέτη του Wei Lian Willian Foh και των συνεργατών του(111), που είχε ως σκοπό την πρόβλεψη του αποτελέσματος της θεραπείας έναντι της φυματίωσης, χρησιμοποιήθηκε ένα αρκετά μεγάλο σύνολο δεδομένων. Τα δεδομένα αυτά προήλθαν από το πληροφοριακό σύστημα της Βραζιλίας για ασθένειες οι οποίες βάση νόμου πρέπει να καταγράφονται.(112) Έγινε προσπάθεια δημιουργίας ενός μοντέλου μηχανικής μάθησης που θα επιτρέψει στις μονάδες υγειονομικής περίθαλψης να βελτιστοποιήσουν την κατανομή των πόρων τους με βάση την πρόβλεψη που γίνεται όσον αφορά την έκβαση. Οι μεταβλητές του συνόλου δεδομένων είναι 37, οι περισσότερες κατηγορικές και σχετίζονται με τα χαρακτηριστικά των ασθενών, τις κλινικές πληροφορίες καθώς και τα εργαστηριακά δεδομένα. Όλες οι εγγραφές ασθενών στο σύνολο δεδομένων χαρακτηρίζονται είτε ως "CURED " για τους ασθενείς που θεραπεύτηκαν είτε ως "DIED " για αυτούς που απεβίωσαν, υποδεικνύοντας την έκβαση της θεραπείας τους, η οποία αποτελεί τις κλάσεις σε αυτό το πρόβλημα ταξινόμησης.

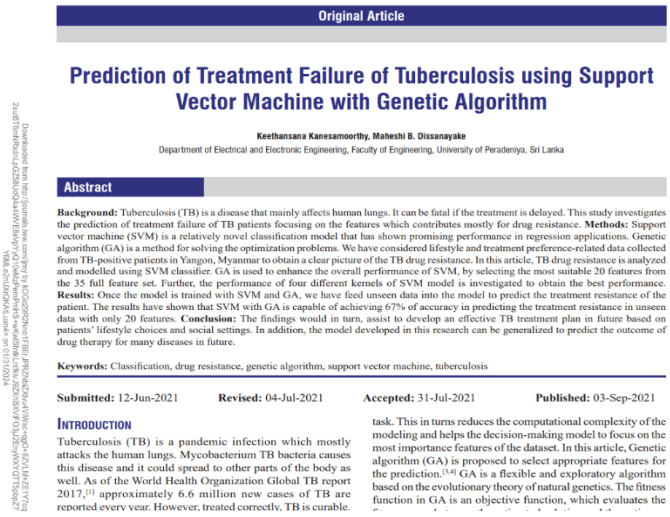
Ωστόσο, το σύνολο δεδομένων παρουσίασε άνιση κατανομή των κλάσεων, φαινόμενο το οποίο μπορεί να οδηγήσει σε λιγότερο αποδοτικό εντοπισμό της κλάσης που αποτελεί την μειοψηφία. Για την αντιμετώπιση της ανισορροπίας των κλάσεων στο στάδιο της προεπεξεργασίας, οι άγνωστες ή ελλιπείς τιμές αντικαταστάθηκαν από την τιμή "άγνωστο". Επιπλέον, για κάθε αριθμητική τιμή, υπολογίστηκαν η τυπική απόκλιση και το ενδοτεταρτημοριακό εύρος και οι μεγαλύτερες ακραίες τιμές αντικαταστάθηκαν με άγνωστες τιμές. Στο τέλος αυτού του σταδίου, οι εγγραφές με άγνωστες τιμές απορρίφθηκαν οδηγώντας σε ένα αρκετά μικρότερο σετ δεδομένων με 313,992 συνολικές εγγραφές θανούντων και μη.

Ο αλγόριθμος που χρησιμοποιήθηκε είναι ο Naive Bayes. Επίσης χρησιμοποιήθηκε η τεχνική SMOTE (Synthetic Minority Oversampling Technique), καθώς δεν λύθηκε το πρόβλημα της ανισορροπίας των κλάσεων. Η τεχνική SMOTE βοηθάει κατά το στάδιο της εκπαίδευσης εισάγοντας τυχαία καινούρια δείγματα στην κλάση που αποτελεί τη μειοψηφία. Η τεχνική αυτή λειτούργησε αποδοτικά και είχε ως αποτέλεσμα την αύξηση των μετρικών της ακρίβειας και της ειδικότητας. Συνέβαλλε στο να ταξινομούνται σωστά περισσότερες περιπτώσεις της μειοψηφικής κλάσης, δηλαδή των θανούντων, αλλά προκάλεσε μια μικρή αύξηση των εσφαλμένα ταξινομημένων περιπτώσεων που ανήκουν στην κλάση των θεραπευμένων, μειώνοντας έτσι ελαφρώς την ευαισθησία.

Το τελικό μοντέλο που δημιουργήθηκε χρειάστηκε λίγα λεπτά για να εκπαιδευτεί και πέτυχε ισορροπημένη ακρίβεια (Balanced Accuracy-BA) 91,9%. Τέλος, οι μετρικές απόδοσης που υπολογίστηκαν ήταν η ROC (0.91), η Ευαισθησία (0.96) και η Ειδικότητα (0.87).

Μελέτη 2

Keethansana Kanesamoorthy, Maheshi B. Dissanayake (July 2021)



Original Article

Prediction of Treatment Failure of Tuberculosis using Support Vector Machine with Genetic Algorithm

Keethansana Kanesamoorthy, Maheshi B. Dissanayake
Department of Electrical and Electronic Engineering, Faculty of Engineering, University of Peradeniya, Sri Lanka

Abstract

Background: Tuberculosis (TB) is a disease that mainly affects human lungs. It can be fatal if the treatment is delayed. This study investigates the prediction of treatment failure of TB patients focusing on the features which contributes mostly for drug resistance. **Methods:** Support vector machine (SVM) is a relatively novel classification model that has shown promising performance in regression applications. Genetic algorithm (GA) is a method for solving the optimization problems. We have considered lifestyle and treatment preference-related data collected from TB-positive patients in Yangon, Myanmar to obtain a clear picture of the TB drug resistance. In this article, TB drug resistance is analyzed and modelled using SVM classifier. GA is used to enhance the overall performance of SVM, by selecting the most suitable 20 features from the 35 full feature set. Further, the performance of four different kernels of SVM model is investigated to obtain the best performance. **Results:** Once the model is trained with SVM and GA, we have feed unseen data into the model to predict the treatment resistance of the patient. The results have shown that SVM with GA is capable of achieving 67% of accuracy in predicting the treatment resistance in unseen data with only 20 features. **Conclusion:** The findings would in turn, assist to develop an effective TB treatment plan in future based on patients' lifestyle choices and social settings. In addition, the model developed in this research can be generalized to predict the outcome of drug therapy for many diseases in future.

Keywords: Classification, drug resistance, genetic algorithm, support vector machine, tuberculosis

Submitted: 12-Jun-2021 **Revised:** 04-Jul-2021 **Accepted:** 31-Jul-2021 **Published:** 03-Sep-2021

Introduction

Tuberculosis (TB) is a pandemic infection which mostly attacks the human lungs. Mycobacterium TB bacteria causes this disease and it could spread to other parts of the body as well. As of the World Health Organization Global TB report 2017,[1] approximately 6.6 million new cases of TB are reported every year. However, treated correctly, TB is curable.

task. This in turns reduces the computational complexity of the modeling and helps the decision-making model to focus on the most importance features of the dataset. In this article, Genetic algorithm (GA) is proposed to select appropriate features for the prediction.[3,4] GA is a flexible and exploratory algorithm based on the evolutionary theory of natural genetics. The fitness function in GA is an objective function, which evaluates the

Εικόνα 16: Prediction of Treatment Failure of Tuberculosis using Support Vector Machine with Genetic Algorithm.

[Πηγή: Kanesamoorthy K, Dissanayake M. Prediction of treatment failure of tuberculosis using support vector machine with genetic algorithm. Int J Mycobacteriology [Internet]. 2021;10(3):279.]

Αυτή η μελέτη(113), αφορά την πρόβλεψη της αποτυχίας της θεραπείας των ασθενών με φυματίωση εστιάζοντας στα χαρακτηριστικά που συμβάλλουν στην αντοχή στα αντιβιοτικά που χρησιμοποιούνται κατά την θεραπεία. Χρησιμοποιήθηκε ένα δημόσια διαθέσιμο σύνολο δεδομένων από 356 ασθενείς με φυματίωση τα δεδομένα του οποίου συλλέχθηκαν από δέκα δημόσιους φορείς υγείας. Χρησιμοποιήθηκε ο γενετικός αλγόριθμος (Genetic Algorithm-GA) για την βελτιστοποίηση της επιλογής του κατάλληλου υποσυνόλου χαρακτηριστικών εξαλείφοντας τα περιττά και μη συναφή χαρακτηριστικά με στόχο τη βελτίωση της ακρίβειας πρόβλεψης του μοντέλου. Ο αλγόριθμος κατηγοριοποίησης που εφαρμόστηκε στα 20 κοινωνικοοικονομικά και κλινικά χαρακτηριστικά ήταν ο SVM. Το σύνολο δεδομένων εισόδου χωρίστηκε σε δύο σύνολα, το σύνολο εκπαίδευσης και το σύνολο δοκιμών με αναλογία 75% και 25% αντίστοιχα.

Τα αποτελέσματα έδειξαν ότι οι γραμμικοί πυρήνες (linear kernels) και οι πυρήνες RBF (Radial Basis Function) υπερτερούν έναντι των πολυωνυμικών (polynomial kernels) και σιγμοειδών πυρήνων (sigmoid kernels) με σαφές περιθώριο (margin), όταν χρησιμοποιήθηκαν και τα 35 χαρακτηριστικά στην κατηγοριοποίηση. Με τη χρήση του αλγόριθμου GA και έχοντας πλέον τα 20 βέλτιστα χαρακτηριστικά, βελτιώθηκε η απόδοση του πολυωνυμικού και του γραμμικού πυρήνα, ενώ οι πυρήνες RBF και οι σιγμοειδείς πυρήνες παρουσίασαν πολύ μικρή βελτίωση στην ακρίβεια ταξινόμησης. Κατά τη σύγκριση της ακρίβειας ταξινόμησης των τεσσάρων συναρτήσεων πυρήνα που αναφέρθηκαν παραπάνω, η συνάρτηση του πυρήνα RBF και του

γραμμικού πυρήνα κατάφερε να βρει υψηλό όριο απόφασης (decision boundary) που ταξινομούσε σχεδόν το 67% των δεδομένων σωστά. Τα αποτελέσματα των μετρικών απόδοσης για την κάθε συνάρτηση πυρήνα παρουσιάζονται στον Πίνακα 1.

	RBF	Linear	Poly	Sigmoid
Accuracy (%)	0.67	0.67	0.49	0.46
Precision	0.45	0.50	0.51	0.54
Recall	0.67	0.67	0.49	0.46
F1-score	0.54	0.57	0.50	0.47

RBF: Radial basis function

Πίνακας 1. Μετρικές αξιολόγησης των πυρήνων SVM του άρθρου “Prediction of Treatment Failure of Tuberculosis using Support Vector Machine with Genetic Algorithm.”

[Πηγή: Kanesamoorthy K, Dissanayake M. Prediction of treatment failure of tuberculosis using support vector machine with genetic algorithm. Int J Mycobacteriology [Internet]. 2021;10(3):279.]

Μελέτη 3

Christopher Martin Sauer et al. (October 2018)

PLOS ONE

RESEARCH ARTICLE

Feature selection and prediction of treatment failure in tuberculosis

Christopher Martin Sauer[1,2☯*], David Sasson[3☯], Kenneth E. Paik[2], Ned McCague[2], Leo Anthony Celi[2], Iván Sánchez Fernández[4‡], Ben M. W. Illigens[5‡]

1 Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, United States of America, 2 Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, United States of America, 3 Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, United States of America, 4 Division of Epilepsy and Clinical Neurophysiology, Department of Neurology, Boston Children's Hospital, Harvard Medical School, Boston, MA, United States of America, 5 Department of Neurology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, United States of America

☯ These authors contributed equally to this work.
‡ These authors also contributed equally to this work.
* csauer@hsph.harvard.edu

Check for updates

OPEN ACCESS

Citation: Sauer CM, Sasson D, Paik KE, McCague N, Celi LA, Sánchez Fernández I, et al. (2018) Feature selection and prediction of treatment failure in tuberculosis. PLoS ONE 13(11): e0207491. https://doi.org/10.1371/journal.pone.0207491

Editor: Zhengxing Huang, Zhejiang University, CHINA

Received: May 25, 2018

Accepted: October 28, 2018

Published: November 20, 2018

Abstract

Background

Tuberculosis is a major cause of morbidity and mortality in the developing world. Drug resistance, which is predicted to rise in many countries worldwide, threatens tuberculosis treatment and control.

Objective

To identify features associated with treatment failure and to predict which patients are at highest risk of treatment failure.

Εικόνα 17: Feature selection and prediction of treatment failure in tuberculosis.

[Πηγή: Sauer CM, Sasson D, Paik KE, McCague N, Celi LA, Fernández IS, et al. Feature selection and prediction of treatment failure in tuberculosis. PLoS One. 2018;13(11):1–14.]

Στην μελέτη του Christopher Martin Sauer και των συνεργατών του(114), στόχος ήταν να προσδιοριστούν τα χαρακτηριστικά που σχετίζονται με την αποτυχία της θεραπείας και να γίνει πρόβλεψη όσον αφορά το ποιοι ασθενείς βρίσκονται σε υψηλότερο κίνδυνο αποτυχίας της θεραπείας. Το σετ δεδομένων που χρησιμοποιήθηκε για τη δημιουργία των μοντέλων μηχανικής μάθησης που θα συμβάλουν στην ανακάλυψη των προγνωστικών παραγόντων αποτυχίας της θεραπείας προήλθε από το Εθνικό Ινστιτούτο Αλλεργιών και Λοιμωδών Νοσημάτων και περιείχε δεδομένα 587 ασθενών που έχουν συλλεχθεί από το Αζερμπαϊτζάν, την Μολδαβία, την Λευκορωσία, την Γεωργία και την Ρουμανία. Τα δεδομένα αφορούσαν κοινωνικοοικονομικά και δημογραφικά χαρακτηριστικά, παράγοντες κινδύνου και ιατρικά δεδομένα. Η αποτυχία της θεραπείας εμφανίστηκε στο ένα τέταρτο περίπου των περιπτώσεων.

Οι ερευνητές, χώρισαν τυχαία το αρχικό σύνολο δεδομένων στο 70% των ασθενών για το training set και στο 30% για το test set. Η απόδοση των μοντέλων βελτιστοποιήθηκε με cross-validation στο training set πριν από την εφαρμογή τους στο test set. Η προσέγγιση αυτή διασφαλίζει ότι η απόδοση των αλγορίθμων στο test set είναι παρόμοια με την απόδοσή τους σε άλλα δεδομένα στα οποία τα μοντέλα δεν έχουν εκτεθεί ποτέ.

Χρησιμοποιήθηκαν αρκετές τεχνικές μηχανικής μάθησης όπως οι μέθοδοι "stepwise forward selection," "stepwise backward elimination," "backward elimination," και "forward selection" που ανήκουν στην κατηγορία των τεχνικών παλινδρόμησης (regression techniques). Συγκεκριμένα, αυτές οι μέθοδοι εφαρμόζονται συνήθως όταν επιδιώκεται η πρόβλεψη μιας εξαρτημένης μεταβλητής με βάση ένα σύνολο ανεξάρτητων μεταβλητών. Επίσης χρησιμοποιήθηκαν το μοντέλο παλινδρόμησης LASSO (Least Absolute Shrinkage and Selection Operator), οι μηχανές διανυσμάτων υποστήριξης SVM με γραμμικό και πολυωνυμικό πυρήνα, καθώς και το τυχαίο δάσος RF. Η παλινδρόμηση LASSO επιλέγει τις μεταβλητές που έχουν τη σημαντικότερη επίδραση στην εξαρτημένη μεταβλητή και απομακρύνει ορισμένες με μηδαμινή επίδραση. Ο αλγόριθμος RF προβλέπει με βάση έναν αριθμό δέντρων αποφάσεων που εμπεριέχει τα οποία έχουν εκπαιδευτεί σε διαφορετικό υποσύνολο του συνόλου εκπαίδευσης. Η τελική πρόβλεψη λαμβάνεται από την πλειοψηφία όσον αφορά τις προβλέψεις κατηγορικών μεταβλητών ή από το μέσο όρο των τιμών για τις προβλέψεις των αριθμητικών μεταβλητών. Όσον αφορά τον SVM, προσπαθεί να βρει ένα μέγιστο οριακό υπερεπίπεδο (Maximum Margin Hyperplane) που διαιρεί καλύτερα το σύνολο δεδομένων σε κλάσεις.

Τα περισσότερα μοντέλα είχαν απόδοση με AUC ίσο ή μεγαλύτερο του 0,7. Σύμφωνα με τους ερευνητές, ανάλογα με τον σκοπό της πρόβλεψης θα μπορούσε να φανεί χρήσιμο κάποιο από τα μοντέλα μηχανικής μάθησης που δημιουργήθηκαν, για παράδειγμα για την έναρξη θεραπείας φυματίωσης ανθεκτικής στα αντιβιοτικά το μοντέλο πρόβλεψης που είναι καταλληλότερο είναι αυτό της παλινδρόμησης LASSO με ειδικότητα 0,96 και θετική προγνωστική αξία (Positive Predictive Value, PPV) ίση με 0,64.

Τα αποτελέσματα από τα μοντέλα μηχανικής μάθησης που δημιουργήθηκαν φαίνονται στον παρακάτω Πίνακα 2.

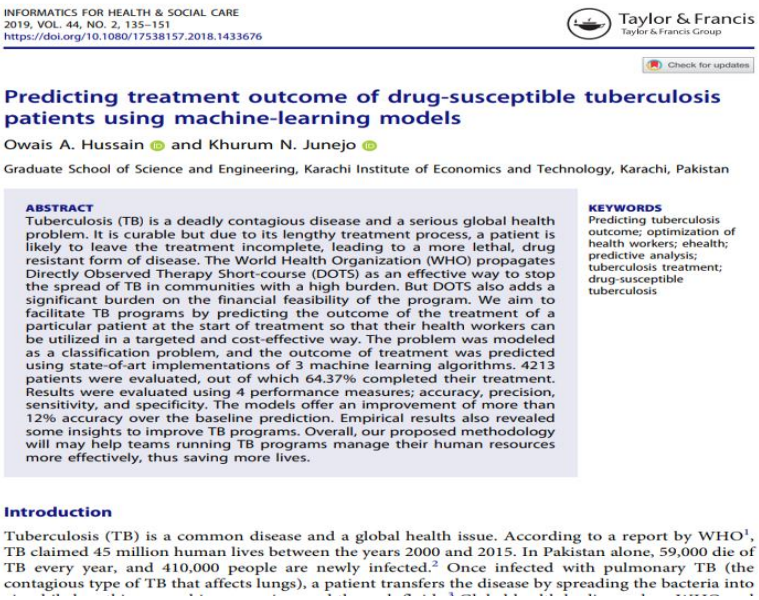
Method	AUC (95% CI)	Misclassi-fication	Sensitivity	Specificity	PPV	NPV
Forward stepwise selection	0.74 (0.66–0.82)	0.24	0.36	0.89	0.53	0.81
Backward stepwise elimination	0.73 (0.65–0.81)	0.27	0.3	0.88	0.45	0.79
Backward stepwise elimination & forward stepwise selection	0.73 (0.65–0.81)	0.27	0.30	0.88	0.45	0.79
LASSO	0.72 (0.64–0.80)	0.23	0.21	0.96	0.64	0.78
Random forest	0.70 (0.62–0.79)	0.24	0.30	0.91	0.52	0.80
SVM linear kernel	0.69 (0.60–0.77)	0.24	0.21	0.94	0.56	0.78
SVM polynomial kernel	0.69 (0.60–0.77)	0.25	0	1	NA	0.75

Πίνακας 2. Σύγκριση της απόδοσης πρόβλεψης των στατιστικών μοντέλων του άρθρου “Feature selection and prediction of treatment failure in tuberculosis”. Ως PPV ορίζεται η θετική προγνωστική αξία και ως NPV η αρνητική προγνωστική αξία.

[Πηγή: Πηγή: Sauer CM, Sasson D, Paik KE, McCague N, Celi LA, Fernández IS, et al. Feature selection and prediction of treatment failure in tuberculosis. PLoS One. 2018;13(11):1–14.]

Μελέτη 4

Owais A. Hussain, Khurum N. Junejo (February 2018)



INFORMATICS FOR HEALTH & SOCIAL CARE
2019, VOL. 44, NO. 2, 135–151
https://doi.org/10.1080/17538157.2018.1433676

Taylor & Francis
Taylor & Francis Group

Check for updates

Predicting treatment outcome of drug-susceptible tuberculosis patients using machine-learning models

Owais A. Hussain and Khurum N. Junejo

Graduate School of Science and Engineering, Karachi Institute of Economics and Technology, Karachi, Pakistan

ABSTRACT

Tuberculosis (TB) is a deadly contagious disease and a serious global health problem. It is curable but due to its lengthy treatment process, a patient is likely to leave the treatment incomplete, leading to a more lethal, drug resistant form of disease. The World Health Organization (WHO) propagates Directly Observed Therapy Short-course (DOTS) as an effective way to stop the spread of TB in communities with a high burden. But DOTS also adds a significant burden on the financial feasibility of the program. We aim to facilitate TB programs by predicting the outcome of the treatment of a particular patient at the start of treatment so that their health workers can be utilized in a targeted and cost-effective way. The problem was modeled as a classification problem, and the outcome of treatment was predicted using state-of-art implementations of 3 machine learning algorithms. 4213 patients were evaluated, out of which 64.37% completed their treatment. Results were evaluated using 4 performance measures; accuracy, precision, sensitivity, and specificity. The models offer an improvement of more than 12% accuracy over the baseline prediction. Empirical results also revealed some insights to improve TB programs. Overall, our proposed methodology will may help teams running TB programs manage their human resources more effectively, thus saving more lives.

KEYWORDS

Predicting tuberculosis outcome; optimization of health workers; ehealth; predictive analysis; tuberculosis treatment; drug-susceptible tuberculosis

Introduction

Tuberculosis (TB) is a common disease and a global health issue. According to a report by WHO[1], TB claimed 45 million human lives between the years 2000 and 2015. In Pakistan alone, 59,000 die of TB every year, and 410,000 people are newly infected.[2] Once infected with pulmonary TB (the contagious type of TB that affects lungs), a patient transfers the disease by spreading the bacteria into

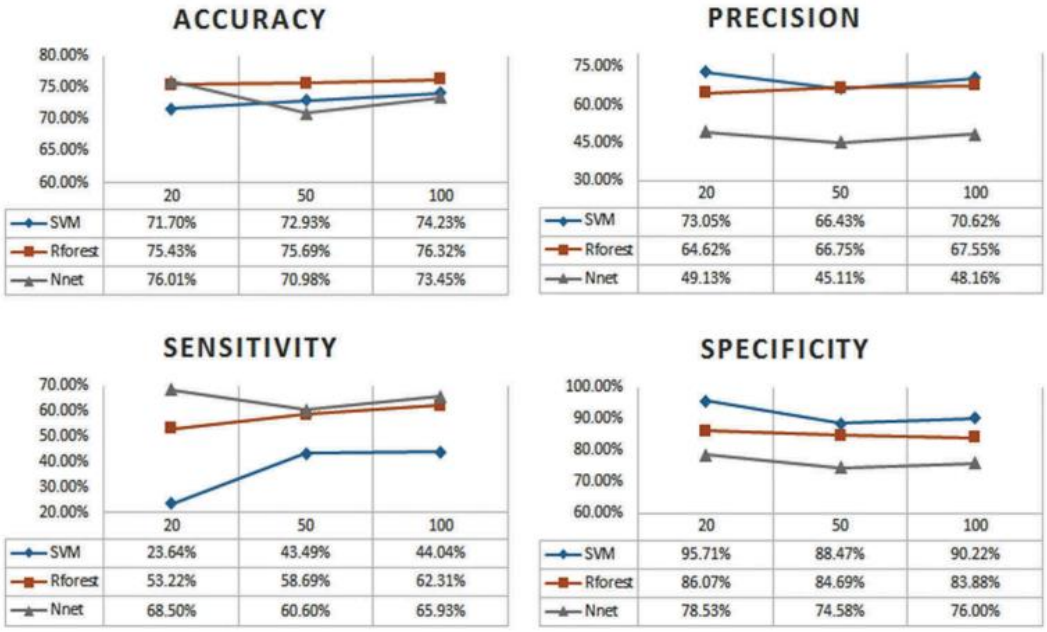
Εικόνα 18: Predicting treatment outcome of drug-susceptible tuberculosis patients using machine-learning models.

[Πηγή: Hussain OA, Junejo KN. Predicting treatment outcome of drug-susceptible tuberculosis patients using machine-learning models. Informatics Heal Soc Care [Internet]. 2019;44(2):135–51.]

Στην μελέτη των Owais A. Hussain και Khurum N. Junejo(115), στόχος ήταν η πρόβλεψη της έκβασης της θεραπείας των ασθενών κατά την έναρξη της θεραπείας, με απώτερο σκοπό την πιο εύρυθμη λειτουργεία των προγραμμάτων διαχείρισης ασθενών με φυματίωση μέσω της βέλτιστης και στοχευμένης διαχείρισης των πόρων από τους εργαζόμενους στον τομέα της υγείας. Σε αυτή τη μελέτη οι ερευνητές δημιούργησαν μοντέλα μηχανικής μάθησης που πραγματοποιούν πρόβλεψη για το αν κάποιος ασθενής θα ολοκληρώσει ή όχι τη θεραπευτική αγωγή. Αυτό αποτελεί πρόβλημα δυαδικής κατηγοριοποίησης.

Το σύνολο δεδομένων που χρησιμοποιήθηκε περιλάμβανε ιατρικούς φακέλους 4213 ασθενών (που συλλέχθηκαν μεταξύ των ετών 2011 και 2014), από την αρχική εξέταση έως το τέλος της θεραπευτικής αγωγής. Οι ασθενείς που παρουσίαζαν ήδη αντοχή στα αντιβιοτικά αποκλείστηκαν από την μελέτη. Από τα 84 χαρακτηριστικά του συνόλου δεδομένων, χρησιμοποιήθηκαν τα 52 που επιλέχθηκαν κατά το στάδιο της προεπεξεργασίας. Τα δεδομένα χωρίστηκαν σε training set, validation set και test set, ώστε να γίνει αξιολόγηση της απόδοσης των μοντέλων σε αθέατα δεδομένα. Λόγω της ανισορροπίας των κλάσεων χρησιμοποιήθηκε η τεχνική της στρωματοποιημένης δειγματοληψίας (stratified sampling), προκειμένου να διασφαλιστεί ότι η αναλογία των δύο κλάσεων παρέμεινε ίδια στα τρία σύνολα δεδομένων. Για την επιλογή των βέλτιστων χαρακτηριστικών, πραγματοποιήθηκε ο στατιστικός έλεγχος χ^2 (Chi-square test), όπου είχε ως αποτέλεσμα την αφαίρεση 32 χαρακτηριστικών. Επιπλέον, το δεύτερο στάδιο για την επιλογή χαρακτηριστικών περιλάμβανε την επιλογή χαρακτηριστικών για κάθε έναν από τους τρεις αλγόριθμους ξεχωριστά. Όταν η ακρίβεια της πρόβλεψης για ένα συγκεκριμένο χαρακτηριστικό ήταν μικρότερη από την πιθανότητα της πιο συχνά εμφανιζόμενης κλάσης (δηλαδή της ολοκληρωμένης θεραπείας), τότε το χαρακτηριστικό αφαιρούνταν, θεωρώντας ότι δεν αυξάνει την ακρίβεια του μοντέλου.

Οι αλγόριθμοι που χρησιμοποιήθηκαν ήταν οι RF, SVM και ANN. Όσον αφορά την πρόβλεψη των ασθενών πού δεν θα ολοκληρώσουν την θεραπεία, το μοντέλο SVM έδειξε υψηλότερη ακρίβεια αλλά χαμηλότερη ευαισθησία, υποδηλώνοντας ότι είναι πολύ συντηρητικό στην ταξινόμηση ενός ασθενούς στην κλάση που αφορά την μη ολοκληρωμένη θεραπευτική αγωγή. Ως εκ τούτου, έκανε λιγότερα σφάλματα τύπου I (ψευδώς θετικά) και περισσότερα σφάλματα τύπου II (ψευδώς αρνητικά). Σε αντίθεση με το μοντέλο SVM, το μοντέλο ANN έκανε λιγότερα σφάλματα τύπου II και περισσότερα σφάλματα τύπου I, δηλαδή είχε χαμηλή ακρίβεια αλλά υψηλή ευαισθησία. Το μοντέλο RF εμφάνισε την καλύτερη απόδοση καθώς παρουσίασε την μεγαλύτερη ακρίβεια (έως 76,3%) μεταξύ των τριών μοντέλων για την πρόβλεψη της έκβασης της θεραπείας. Ο SVM ξεπέρασε τους υπόλοιπους ταξινομητές μηχανικής μάθησης όσον αφορά την ειδικότητα (έως 95,71%). Από την άλλη πλευρά, ο ANN πέτυχε την υψηλότερη ευαισθησία (έως 68,5%). Τα αποτελέσματα των μετρικών απόδοσης για κάθε μοντέλο παρουσιάζονται στην Εικόνα 19.

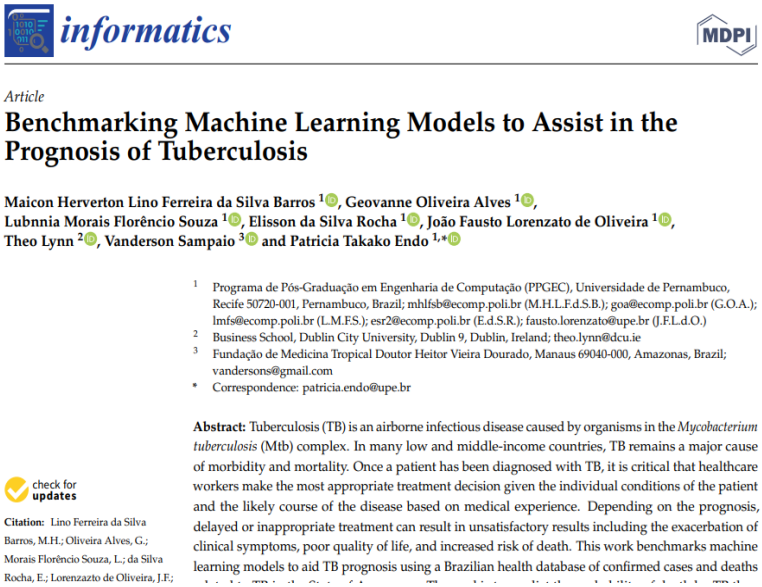


Εικόνα 19: Μετρικές αξιολόγησης της απόδοσης των τριών αλγορίθμων του άρθρου "Predicting treatment outcome of drug-susceptible tuberculosis patients using machine-learning models."

Στον άξονα x απεικονίζεται το μέγεθος του δείγματος (το ποσοστό των δεδομένων εκπαίδευσης) ενώ στον άξονα y η απόδοση των αλγορίθμων. [Πηγή: Hussain OA, Junejo KN. Predicting treatment outcome of drug-susceptible tuberculosis patients using machine-learning models. Informatics Heal Soc Care [Internet]. 2019;44(2):135–51.]

Μελέτη 5

Maicon Herverton Lino Ferreira da Silva Barros et al. (April 2021)



informatics

MDPI

Article

Benchmarking Machine Learning Models to Assist in the Prognosis of Tuberculosis

Maicon Herverton Lino Ferreira da Silva Barros [1], Geovanne Oliveira Alves [1], Lubnnia Morais Florêncio Souza [1], Elisson da Silva Rocha [1], João Fausto Lorenzato de Oliveira [1], Theo Lynn [2], Vanderson Sampaio [3] and Patricia Takako Endo [1,*]

1 Programa de Pós-Graduação em Engenharia de Computação (PPGEC), Universidade de Pernambuco, Recife 50720-001, Pernambuco, Brazil; mhlfsb@ecomp.poli.br (M.H.L.F.d.S.B.); goa@ecomp.poli.br (G.O.A.); lmfs@ecomp.poli.br (L.M.F.S.); esr2@ecomp.poli.br (E.d.S.R.); fausto.lorenzato@upe.br (J.F.L.d.O.)
2 Business School, Dublin City University, Dublin 9, Dublin, Ireland; theo.lynn@dcu.ie
3 Fundação de Medicina Tropical Doutor Heitor Vieira Dourado, Manaus 69040-000, Amazonas, Brazil; vandersons@gmail.com
* Correspondence: patricia.endo@upe.br

Abstract: Tuberculosis (TB) is an airborne infectious disease caused by organisms in the *Mycobacterium tuberculosis* (Mtb) complex. In many low and middle-income countries, TB remains a major cause of morbidity and mortality. Once a patient has been diagnosed with TB, it is critical that healthcare workers make the most appropriate treatment decision given the individual conditions of the patient and the likely course of the disease based on medical experience. Depending on the prognosis, delayed or inappropriate treatment can result in unsatisfactory results including the exacerbation of clinical symptoms, poor quality of life, and increased risk of death. This work benchmarks machine learning models to aid TB prognosis using a Brazilian health database of confirmed cases and deaths

check for updates

Citation: Lino Ferreira da Silva Barros, M.H.; Oliveira Alves, G.; Morais Florêncio Souza, L.; da Silva Rocha, E.; Lorenzazto de Oliveira, J.F.;

Εικόνα 20: Benchmarking Machine Learning Models to Assist in the Prognosis of Tuberculosis.

[Πηγή: Lino Ferreira da Silva Barros MH, Alves GO, Morais Florêncio Souza L, da Silva Rocha E, Lorenzato de Oliveira JF, Lynn T, et al. Benchmarking machine learning models to assist in the prognosis of tuberculosis. Informatics. 2021;8(2):1–17.]

Στην μελέτη του Maicon Herverton Lino Ferreira da Silva Barros και των συνεργατών του(99), δημιουργήθηκαν μοντέλα μηχανικής μάθησης χρησιμοποιώντας δεδομένα από μια βάση δεδομένων υγείας της Βραζιλίας με επιβεβαιωμένα κρούσματα και θανάτους που σχετίζονται με τη φυματίωση. Ο στόχος ήταν να προβλεφθεί η πιθανότητα θανάτου από φυματίωση έτσι ώστε να βελτιωθεί η πρόγνωση της φυματίωσης και η σχετική διαδικασία λήψης αποφάσεων για τη θεραπεία των ασθενών.

Η γλώσσα προγραμματισμού που χρησιμοποιήθηκε στη μελέτη αυτή ήταν η Python. Μετά την προεπεξεργασία, το σύνολο δεδομένων περιλάμβανε 24.015 εγγραφές ασθενών και 38 χαρακτηριστικά που αφορούσαν δημογραφικά, κλινικά και εργαστηριακά δεδομένα.

Για την επιλογή ενός υποσυνόλου χαρακτηριστικών, πραγματοποιήθηκε σύγκριση της απόδοση τεσσάρων διαφορετικών τεχνικών επιλογής χαρακτηριστικών. Η επιλογή των χαρακτηριστικών βασίστηκε στο F1-score. Η εκτέλεση των τεχνικών επιλογής χαρακτηριστικών έγινε υπό συνθήκες διασταυρούμενης επικύρωσης (cross-validation) με k=10. Η πρώτη μέθοδος Sequential Forward Selection (SFS), ξεκινά με ένα κενό σύνολο χαρακτηριστικών και επαναληπτικά επιλέγει ένα χαρακτηριστικό κάθε φορά, ξεκινώντας με το χαρακτηριστικό που δυνητικά θα αυξάνει την απόδοση του μοντέλου μέχρις ότου δεν μπορεί να επιτευχθεί βελτίωση της ακρίβειας ταξινόμησης. Ωστόσο, όταν προστεθεί ένα χαρακτηριστικό δεν αφαιρείται ποτέ. Σε αντίθεση με την SFS, μέθοδος Sequential Backward Selection (SBS) ξεκινά με το σύνολο όλων των χαρακτηριστικών και εξαλείφει προοδευτικά τα λιγότερο υποσχόμενα και σταματά εάν η απόδοση των αλγορίθμων μάθησης πέσει κάτω από ένα δεδομένο κατώτατο όριο (threshold) λόγω της αφαίρεσης όλων των εναπομεινάντων χαρακτηριστικών. Στη μέθοδο αυτή, η αφαίρεση των χαρακτηριστικών είναι οριστική. Επιπρόσθετα, οι μέθοδοι Sequential Forward Floating Selection (SFFS) και Sequential Backward Floating Selection (SBFS) λειτουργούν παρόμοια με τις SFS και SBS αντίστοιχα, με τη διαφορά ότι σε αυτές μπορεί να πραγματοποιηθεί αναστροφή στην επιλογή χαρακτηριστικών. (116) Η δυνατότητα αναστροφής της προσθήκης ή αφαίρεσης χαρακτηριστικών επιτρέπει στους αλγόριθμους να δοκιμάσουν περισσότερους συνδυασμούς και να λάβουν αποφάσεις με βάση τις βέλτιστες επιδόσεις στο σύνολο δεδομένων.

Οι αλγόριθμοι μηχανικής μάθησης που χρησιμοποιήθηκαν είναι οι εξής: Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Naive Bayes (NB), Decision Trees (DT), Support Vector Machine (SVM), Gradient Boosting (GB), Random Forest (RF) και Multilayer Perceptron (MLP). Από τα 38 χαρακτηριστικά επιλέχθηκαν 17 για καθένα από τα εννέα μοντέλα μηχανικής μάθησης.

Τα πειραματικά μοντέλα σχεδιάστηκαν για το σύνολο δεδομένων με ανισορροπία στις κλάσεις αλλά και για το σύνολο δεδομένων με ισορροπία στις κλάσεις που δημιουργήθηκε. Για τη δημιουργία ισορροπημένου συνόλου δεδομένων, εφαρμόστηκε η τεχνική της τυχαίας υποδειγματοληψίας (random

under-sampling technique). Το ισορροπημένο σύνολο δεδομένων περιλάμβανε 1139 εγγραφές ασθενών που θεραπεύτηκαν από φυματίωση και 1139 θανάτους. Στη συνέχεια, το σύνολο δεδομένων χωρίστηκε σε training set και validation set (70%) και test set (30%).

Όσον αφορά τις τεχνικές επιλογής χαρακτηριστικών, το μοντέλο DT παρουσίασε το καλύτερο F1-score (96,00%) κατά τη χρήση της τεχνικής SFS. Ενώ το μοντέλο LDA παρουσίασε το καλύτερο F1-score (95,31%) όταν χρησιμοποιήθηκε η τεχνική SBS. Τα μοντέλα KNN, NB, SVM και RF παρουσίασαν το καλύτερο F1-score, 95,40%, 94,39%, 95,23% και 94,84%, αντίστοιχα, όταν χρησιμοποιήθηκε η τεχνική SFFS. Τέλος τα μοντέλα LR, GB και MLP παρουσίασαν το καλύτερο F1-score, 95,31%, 96,30% και 95,72%, αντίστοιχα, όταν χρησιμοποιήθηκε η τεχνική SBFS. Τα αποτελέσματα αυτά φαίνονται στον Πίνακα 3.

Model	Feature Selection Techniques			
	SFS	SFFS	SBS	SBFS
LR	94.71 (±0.007)	94.88 (±0.007)	95.30 (±0.000)	95.31 (±0.000)
LDA	94.94 (±0.007)	95.13 (±0.006)	95.31 (±0.001)	95.30 (±0.001)
KNN	95.17 (±0.004)	95.40 (±0.002)	93.79 (±0.004)	93.89 (±0.005)
DT	96.00 (±0.002)	95.99 (±0.002)	95.71 (±0.001)	95.70 (±0.001)
NB	94.11 (±0.003)	94.39 (±0.001)	90.15 (±0.004)	90.15 (±0.004)
SVM	95.22 (±0.002)	95.23 (±0.002)	94.37 (±0.002)	94.38 (±0.002)
GB	96.04 (±0.003)	96.02 (±0.003)	96.29 (±0.000)	96.30 (±0.000)
RF	94.63 (±0.006)	94.84 (±0.006)	92.69 (±0.005)	92.74 (±0.005)
MLP	95.51 (±0.004)	95.55 (±0.003)	95.70 (±0.000)	95.72 (±0.000)

Πίνακας 3. Αποτελέσματα του F1-score (σε %) και της αντίστοιχης τυπικής απόκλισης των τεχνικών επιλογής χαρακτηριστικών για κάθε μοντέλο μηχανικής μάθησης του άρθρου “Benchmarking Machine Learning Models to Assist in the Prognosis of Tuberculosis”.

[Πηγή: Lino Ferreira da Silva Barros MH, Alves GO, Morais Florêncio Souza L, da Silva Rocha E, Lorenzato de Oliveira JF, Lynn T, et al. Benchmarking machine learning models to assist in the prognosis of tuberculosis. Informatics. 2021;8(2):1–17.]

Τα αποτελέσματα του test set έδειξαν ότι για το μη ισορροπημένο σύνολο δεδομένων (Πίνακας 4), το μοντέλο RF και το μοντέλο ensemble (συνδυασμός των αποτελεσμάτων των μοντέλων RF, GB και MLP) παρουσίασαν τον καλύτερο μέσο όρο για τρεις μετρικές. Το μοντέλο RF είχε καλύτερες επιδόσεις στην ακρίβεια (99,58%), την ευαισθησία (91,50%) και την AUC ROC (94,41%), ενώ το μοντέλο ensemble είχε καλύτερες επιδόσεις στην ακρίβεια (98,57%), το F1-score (99,25%) και F1-macro (91,46%). Κατά τη χρήση του ισορροπημένου συνόλου δεδομένων (Πίνακας 5), το μοντέλο GB είχε την καλύτερη απόδοση μεταξύ των 9 μοντέλων. Ωστόσο, τα μοντέλα DT, SVM και ensemble παρουσίασαν πολύ παρόμοια αποτελέσματα με το μοντέλο GB. Σε γενικές γραμμές, τα μοντέλα που εκπαιδεύτηκαν με το ισορροπημένο σύνολο δεδομένων πέτυχαν καλύτερα αποτελέσματα.

Imbalanced Data Set				
Metric	**GB**	**RF**	**MLP**	**Ensemble**
Accuracy	98.47 (±0.000)	97.05 (±0.000)	98.11 (±0.000)	98.57 (±0.000)
Precision	98.90 (±0.000)	99.58 (±0.000)	98.80 (±0.001)	99.02 (±0.000)
Sensitivity	77.12 (±0.008)	91.50 (±0.001)	75.05 (±0.021)	79.67 (±0.003)
Specificity	99.50 (±0.000)	97.32 (±0.000)	99.22 (±0.000)	99.48 (±0.000)
F1-score	99.20 (±0.000)	98.43 (±0.000)	99.01 (±0.000)	99.25 (±0.000)
AUC ROC	88.31 (±0.004)	94.41 (±0.000)	87.13 (±0.010)	89.57 (±0.001)
F1-macro	90.76 (±0.002)	86.65 (±0.002)	89.12 (±0.004)	91.46 (±0.001)

Πίνακας 4. Αποτελέσματα των μετρικών (σε %) και της σχετικής τυπικής απόκλισης για τη δοκιμή του μοντέλου με τη χρήση του μη ισορροπημένου συνόλου δεδομένων του άρθρου “Benchmarking Machine Learning Models to Assist in the Prognosis of Tuberculosis”.

[Πηγή: Lino Ferreira da Silva Barros MH, Alves GO, Morais Florêncio Souza L, da Silva Rocha E, Lorenzato de Oliveira JF, Lynn T, et al. Benchmarking machine learning models to assist in the prognosis of tuberculosis. Informatics. 2021;8(2):1–17.]

Balanced Data Set				
Metric	**DT**	**SVM**	**GB**	**Ensemble**
Accuracy	94.14 (± 0.017)	95.30 (± 0.000)	95.97 (± 0.001)	95.80 (± 0.004)
Precision	99.56 (± 0.001)	99.17 (± 0.000)	99.86 (± 0.000)	99.85 (± 0.000)
Sensitivity	91.54 (± 0.023)	83.38 (± 0.000)	97.22 (± 0.001)	97.12 (± 0.002)
Specificity	94.26 (± 0.018)	95.88 (± 0.000)	95.91 (± 0.001)	95.74 (± 0.004)
F1-score	96.83 (± 0.001)	97.50 (± 0.000)	97.84 (± 0.000)	97.75 (± 0.002)
AUC ROC	92.90 (± 0.016)	89.63 (± 0.000)	96.56 (± 0.000)	96.43 (± 0.002)
F1-macro	78.29 (± 0.039)	79.76 (± 0.000)	83.40 (± 0.003)	82.92 (± 0.011)

Πίνακας 5. Αποτελέσματα των μετρικών (σε %) και της σχετικής τυπικής απόκλισης για τη δοκιμή του μοντέλου με τη χρήση του ισορροπημένου συνόλου δεδομένων του άρθρου “Benchmarking Machine Learning Models to Assist in the Prognosis of Tuberculosis”.

[Πηγή: Lino Ferreira da Silva Barros MH, Alves GO, Morais Florêncio Souza L, da Silva Rocha E, Lorenzato de Oliveira JF, Lynn T, et al. Benchmarking machine learning models to assist in the prognosis of tuberculosis. Informatics. 2021;8(2):1–17.]

ΕΙΔΙΚΟ ΜΕΡΟΣ

4. Σκοπός

Για την καλύτερη διαχείριση της φυματίωσης σε παγκόσμιο επίπεδο, η ανάγκη για νέες προσεγγίσεις όσον αφορά την θεραπεία της νόσου κρίνεται επιτακτική. Αυτό προκύπτει από τα ανεπαρκή αποτελέσματα των θεραπευτικών προσεγγίσεων, τα οποία οφείλονται κυρίως στην αδυναμία συμμόρφωσης των ασθενών στα προγράμματα θεραπείας. Ως αποτέλεσμα της μη συμμόρφωσης, θα μπορούσαν να επέλθουν δυσμενείς επιπτώσεις όπως η αύξηση του κινδύνου του θανάτου αλλά και η δημιουργία στελεχών ανθεκτικών στα αντιβιοτικά. Η παρούσα εργασία στοχεύει στην δημιουργία ενός μοντέλου μηχανικής μάθησης το οποίο θα προβλέπει αποδοτικά και με ακρίβεια την έκβαση της θεραπευτικής αγωγής των ασθενών με φυματίωση. Το μοντέλο αυτό θα μπορούσε να συμβάλλει στην λήψη πιο στοχευμένων αποφάσεων για την επιλογή των ασθενών που χρειάζονται εντατικότερη παρακολούθηση, καθώς και στην βέλτιστη κατανομή των πόρων από τους φορείς παροχής υγείας. Έτσι, θα ήταν εφικτή η αύξηση των ποσοστών ολοκλήρωσης και επιτυχίας της θεραπείας για τους ασθενείς με φυματίωση.

5. Εργαλεία και μέθοδοι

5.1 Δεδομένα

Το σύνολο δεδομένων που χρησιμοποιήθηκε στην παρούσα εργασία, στο οποίο έχουν εφαρμοστεί τεχνικές μηχανικής μάθησης, ανακτήθηκε από το Mendeley Data το οποίο είναι αποθετήριο ερευνητικών δεδομένων. Τα δεδομένα ανακτήθηκαν υπό τη μορφή αρχείου CSV (Comma-Separated Values).(117)

Είναι ένα μεγάλο σύνολο δεδομένων που περιλαμβάνει πολλαπλές μεταβλητές για κάθε ασθενή. Τα δεδομένα αυτά συλλέχθηκαν από το Βραζιλιάνικο Σύστημα Πληροφόρησης για Μεταδιδόμενα Νοσήματα (SINAN) μεταξύ Ιανουαρίου του 2001 και Απριλίου του 2020. Το σετ δεδομένων αποτελείται από 37 χαρακτηριστικά (μεταβλητές), 964.099 εγγραφές και περιέχει τα κοινωνικά και δημογραφικά χαρακτηριστικά των ασθενών, καθώς και τις κλινικές πληροφορίες και τα εργαστηριακά δεδομένα που αφορούν την φυματίωση. Η έκβαση της θεραπείας, δηλαδή η ίαση των ασθενών ή ο θάνατος αποτελούν τις κλάσεις οι οποίες χαρακτηρίζονται ως "CURED" και ως "DIED" αντίστοιχα.

Το αρχείο που ανακτήθηκε από το αποθετήριο ερευνητικών δεδομένων Mendeley Data, τροποποιήθηκε ώστε να είναι κατάλληλο για χρήση με τη βοήθεια του επεξεργαστή κειμένου Notepad++ (έκδοση v8.6.2 64-bit) και αποθηκεύτηκε ως αρχείο CSV με το όνομα "data set". Από τις 964.099 εγγραφές ασθενών επιλέχθηκαν τυχαία οι πρώτες 20.000 εγγραφές, καθώς ο όγκος του συνόλου δεδομένων ήταν αρκετά μεγάλος.

5.2 Λογισμικό

Για την ανάπτυξη και την αξιολόγηση των μοντέλων μηχανικής μάθησης που δημιουργήθηκαν στην παρούσα εργασία, χρησιμοποιήθηκε το εργαλείο WEKA (Waikato Environment for Knowledge Analysis), το οποίο είναι ένα δημοφιλές και ευρέως αποδεκτό στην ακαδημαϊκή κοινότητα(118), ελεύθερο λογισμικό (freeware) μηχανικής μάθησης και εξόρυξης δεδομένων, ανοιχτού κώδικα γραμμένο σε Java. Περιέχει αρκετά μεγάλη ποικιλία μεθόδων για παλινδρόμηση, κατηγοριοποίηση και συσταδοποίηση. Επιπρόσθετα, παρέχει τη δυνατότητα για προεπεξεργασία των δεδομένων, καθώς και εργαλεία οπτικοποίησης.(119)(120) Η έκδοση του WEKA που χρησιμοποιήθηκε είναι η 3.8.6. Τα πειράματα διεξήχθησαν στο υπολογιστικό σύστημα Vostro 3500 της εταιρείας Dell με επεξεργαστή 11ης γενιάς Intel Core i3 (i3-1115G4 3.00 GHz), εγκατεστημένη μνήμη RAM 8GB και λειτουργικό σύστημα Windows 11 Home (64 bit).

5.3 Προεπεξεργασία και μετασχηματισμός

Βήματα

1. Το πρώτο βήμα στην προεπεξεργασία των δεδομένων ήταν η μετάφραση των ονομάτων των χαρακτηριστικών του συνόλου δεδομένων από την Πορτογαλική γλώσσα στην Αγγλική.

2. Έπειτα, το χαρακτηριστικό **«treatment outcome»**, που αφορά την έκβαση της θεραπείας της φυματίωσης ορίστηκε ως **κλάση**.

3. Κατόπιν, εξετάστηκε η ύπαρξη απουσιάζουσων τιμών (missing values) και διαπιστώθηκε ότι τα δεδομένα δεν ήταν ελλιπή.

4. Στην συνέχεια, αφαιρέθηκε από το σύνολο δεδομένων το χαρακτηριστικό «date of notification», δηλαδή η ημερομηνία που καταγράφηκε το κρούσμα φυματίωσης στο σύστημα SINAN, το οποίο δεν σχετίζεται με την πρόβλεψη της έκβασης της θεραπείας της φυματίωσης. Ως αποτέλεσμα, διατηρήθηκαν 36 από τα 37 αρχικά χαρακτηριστικά και οι κλάσεις, τα οποία παρουσιάζονται στον Πίνακα 6.

Χαρακτηριστικό	Περιγραφή	Εύρος τιμών
biological sex	Βιολογικό φύλο του ασθενούς	0 - Γυναίκα 1 - Άνδρας 2 - Απροσδιόριστο
Race	Φυλή που δηλώνει ο ασθενής	1 - Λευκός 2 - Μαύρος 3 - Κίτρινος 4 - Μικτή 5 – Ιθαγενής 9 - Απροσδιόριστο
patient situation	Κατάσταση του ασθενούς κατά την είσοδο στον φορέα υγείας	1 - Νέα περίπτωση 2 - Υποτροπή 3 - Επανεισαγωγή μετά από εγκατάλειψη θεραπείας 4 - Άγνοια 5 - Μεταφορά 6 - Μετά θάνατον 9 - Απροσδιόριστο

chest X-ray	Αποτέλεσμα ακτινογραφίας θώρακος κατά τη στιγμή της αναφοράς του περιστατικού	1 - Ύποπτο 2 - Φυσιολογικό, 3 - Άλλη παθολογία 4 - Όχι 9 - Απροσδιόριστο
tuberculin skin test	Αποτέλεσμα φυματινοαντίδρασης: μη αντιδρών (0-4mm), ασθενώς αντιδρών (5- 9mm), ισχυρώς αντιδρών (10mm ή περισσότερο)	1 - Μη αντιδρών 2 - Ασθενώς αντιδρών 3 - Ισχυρώς αντιδρών 4 - Δεν εκτελέστηκε 9 - Απροσδιόριστο
tuberculosis form	Κλινική μορφή φυματίωσης κατά τη στιγμή της αναφοράς του περιστατικού	1 - Πνευμονική 2 - Εξωπνευμονική, 3 - Πνευμονική+ Εξωπνευμονική 9 - Απροσδιόριστο
AIDS	AIDS που σχετίζεται με φυματίωση κατά τη στιγμή της αναφοράς του περιστατικού	1 - Ναι 2 - Όχι 9 - Απροσδιόριστο
alcohol consumption	Κατανάλωση αλκοόλ που σχετίζεται με φυματίωση κατά τη στιγμή της αναφοράς του περιστατικού	1 - Ναι 2 - Όχι 9 - Απροσδιόριστο
Diabetes	Διαβήτης που σχετίζεται με τη φυματίωση κατά τη στιγμή της αναφοράς του περιστατικού	1 - Ναι 2 - Όχι 3 - Δεν εκτελέστηκε, 9 - Απροσδιόριστο
mental disease	Ψυχική ασθένεια που σχετίζεται με τη φυματίωση κατά τη στιγμή της αναφοράς του περιστατικού	1 - Ναι 2 – Όχι 9 - Απροσδιόριστο
other diseases	Άλλες ασθένειες που σχετίζονται με τη φυματίωση κατά τη στιγμή της αναφοράς του περιστατικού	1 - Ναι 2 - Όχι 9 - Απροσδιόριστο
sputum smear 1st sample	Αποτέλεσμα του επιχρίσματος πτυέλων για εύρεση οξεάντοχων βακίλων. 1ο δείγμα	1 - Θετικό 2 - Αρνητικό 3 - Σε εξέλιξη 4 - Δεν έχει πραγματοποιηθεί 9 - Απροσδιόριστο
sputum smear 2nd sample	Αποτέλεσμα του επιχρίσματος πτυέλων για εύρεση οξεάντοχων βακίλων. 2ο δείγμα	1 - Θετικό 2 - Αρνητικό 3 - Δεν πραγματοποιήθηκε

smear of other material	Αποτέλεσμα του επιχρίσματος άλλου βιολογικού υλικού για εύρεση οξεάντοχων βακίλων	1 - Θετικό 2 - Αρνητικό 3 - Δεν πραγματοποιήθηκε
Culture	Αποτέλεσμα της καλλιέργειας πτυέλων για διάγνωση του M. tuberculosis	1 - Θετικό 2 - Αρνητικό 3 - Σε εξέλιξη 4 - Δεν έχει πραγματοποιηθεί 9 - Απροσδιόριστο
HIV serology	Αποτέλεσμα ορολογικής εξέτασης για τον ιό HIV	1 - Θετικό 2 - Αρνητικό 3 - Σε εξέλιξη, 4 - Δεν πραγματοποιήθηκε 9 - Απροσδιόριστο
Rifampicin	Χορήγηση Ριφαμπικίνης	1 - Ναι 2 - Όχι 9 - Απροσδιόριστο
Isoniazid	Χορήγηση Ισονιαζίδης	1 - Ναι 2 - Όχι 9 - Απροσδιόριστο
Ethambutol	Χορήγηση Εθαμβουτόλης	1 - Ναι 2 - Όχι 9 - Απροσδιόριστο
streptomycin	Χορήγηση Στρεπτομυκίνης	1 - Ναι 2 - Όχι 9 - Απροσδιόριστο
pyrazinamide	Χορήγηση Πυραζιναμίδης	1 - Ναι 2 - Όχι 9 - Απροσδιόριστο
ethionamide	Χορήγηση Εθειοναμίδης	1 - Ναι 2 - Όχι 9 - Απροσδιόριστο
other drugs	Άλλα φάρμακα	1 - Ναι 2 - Όχι 9 - Απροσδιόριστο
supervised treatment	Υπόδειξη για εποπτευόμενη θεραπεία κατά τη στιγμή της διάγνωσης	1 - Ναι 2 - Όχι 9 - Απροσδιόριστο
work acquired disease	Εάν ο ασθενής απέκτησε τη νόσο ως αποτέλεσμα της εργασιακών συνθηκών	1 - Ναι 2 - Όχι 9 - Απροσδιόριστο
sputum smear 1st month	Αποτέλεσμα της μικροσκοπικής εξέτασης επιχρίσματος πτυέλων για εύρεση οξεάντοχου βακίλου που πραγματοποιήθηκε σε δείγμα που συλλέχθηκε στο τέλος του 1ου μήνα της θεραπείας	1 - Θετικό 2 - Αρνητικό 3 - Δεν πραγματοποιήθηκε 4 - Δεν εφαρμόζεται 9 - Απροσδιόριστο

sputum smear 2nd month	Αποτέλεσμα της μικροσκοπικής εξέτασης επιχρίσματος πτυέλων για εύρεση οξεάντοχου βακίλου που πραγματοποιήθηκε σε δείγμα που συλλέχθηκε στο τέλος του 2ου μήνα της θεραπείας	1 - Θετικό 2 - Αρνητικό 3 - Δεν πραγματοποιήθηκε 4 - Δεν εφαρμόζεται 9 - Απροσδιόριστο
sputum smear 3rd month	Αποτέλεσμα της μικροσκοπικής εξέτασης επιχρίσματος πτυέλων για εύρεση οξεάντοχου βακίλου που πραγματοποιήθηκε σε δείγμα που συλλέχθηκε στο τέλος του 3ου μήνα της θεραπείας	1 - Θετικό 2 - Αρνητικό 3 - Δεν πραγματοποιήθηκε 4 - Δεν εφαρμόζεται 9 - Απροσδιόριστο
sputum smear 4th month	Αποτέλεσμα της μικροσκοπικής εξέτασης επιχρίσματος πτυέλων για εύρεση οξεάντοχου βακίλου που πραγματοποιήθηκε σε δείγμα που συλλέχθηκε στο τέλος του 4ου μήνα της θεραπείας	1 - Θετικό 2 - Αρνητικό 3 - Δεν πραγματοποιήθηκε 4 - Δεν εφαρμόζεται 9 - Απροσδιόριστο
sputum smear 5th month	Αποτέλεσμα της μικροσκοπικής εξέτασης επιχρίσματος πτυέλων για εύρεση οξεάντοχου βακίλου που πραγματοποιήθηκε σε δείγμα που συλλέχθηκε στο τέλος του 5ου μήνα της θεραπείας	1 - Θετικό 2 - Αρνητικό 3 - Δεν πραγματοποιήθηκε 4 - Δεν εφαρμόζεται 9 - Απροσδιόριστο
sputum smear 6th month	Αποτέλεσμα της μικροσκοπικής εξέτασης επιχρίσματος πτυέλων για εύρεση οξεάντοχου βακίλου που πραγματοποιήθηκε σε δείγμα που συλλέχθηκε στο τέλος του 6ου μήνα της θεραπείας	1 - Θετικό 2 - Αρνητικό 3 - Δεν πραγματοποιήθηκε 4 - Δεν εφαρμόζεται 9 - Απροσδιόριστο
drugs tuberculosis associated	Άλλα φάρμακα που σχετίζονταν με τη φυματίωση κατά τη στιγμή της αναφοράς του περιστατικού	1 - Ναι 2 - Όχι 9 - Απροσδιόριστο
tobacco consumption	Κατανάλωση καπνού που σχετίζεται με φυματίωση κατά τη στιγμή της αναφοράς του περιστατικού	1 - Ναι 2 - Όχι 9 - Απροσδιόριστο
state of Brazil	Πολιτεία της Βραζιλίας όπου διαμένει ο ασθενής	Αλφαριθμητικό

treatment(days)	Αριθμός ημερών που ο ασθενής βρισκόταν σε θεραπεία, μεταξύ της ημερομηνίας της διάγνωσης και της ημερομηνίας λήξης της θεραπείας	Αριθμητικό
Age	Ηλικία ασθενούς	Αριθμητικό
treatment outcome (κλάση)	Έκβαση της θεραπείας της φυματίωσης	1 - Ίαση 3 - Θάνατος

Πίνακας 6. Επεξήγηση των χαρακτηριστικών (attributes) και των τιμών που λαμβάνουν.

Βήματα (συνέχεια)

5. Το νέο σετ δεδομένων αποθηκεύτηκε υπό μορφή αρχείου ARFF με το όνομα «data set 2.arff».

6. Όσον αφορά τις μέρες νοσηλείας, έγινε διακριτοποίηση για το χαρακτηριστικό «treatment(days)» καθώς υπήρξε πολύ μεγάλος αριθμός διαφορετικών τιμών. Για την αντιμετώπιση της πολυπλοκότητας των δεδομένων, οι αριθμητικές τιμές μετατράπηκαν σε κατηγορίες χρησιμοποιώντας το κατάλληλο φίλτρο (Filter→ filters→ unsupervised→ attribute→ Discretize). Έτσι δημιουργήθηκαν 14 κατηγορίες στις οποίες χωρίστηκαν οι τιμές του χαρακτηριστικού «treatment(days)». Τα αποτελέσματα αποθηκεύτηκαν υπό μορφή αρχείου ARFF με το όνομα «data set 4.arff».

7. Όσον αφορά την ηλικία, έγινε διακριτοποίηση για το χαρακτηριστικό «age» με σκοπό να μειωθεί η πολυπλοκότητα των δεδομένων και να είναι πιο εύκολη η ανάλυση τους και έτσι οι αριθμητικές τιμές μετατράπηκαν σε κατηγορίες χρησιμοποιώντας το κατάλληλο φίλτρο (Filter→ filters→ unsupervised→ attribute→ Discretize). Ως αποτέλεσμα, δημιουργήθηκαν 10 κατηγορίες στις οποίες χωρίστηκαν οι τιμές του χαρακτηριστικού «age». Τα αποτελέσματα αποθηκεύτηκαν υπό μορφή αρχείου ARFF με το όνομα «data set 5.arff».

5.3.1 Τεχνική SMOTE

Στην συνέχεια, λόγω της ανισορροπίας των κλάσεων, κατά το στάδιο της προεπεξεργασίας των δεδομένων εφαρμόστηκε η τεχνική **Synthetic Minority Over-sampling Technique (SMOTE)**. Η τεχνική αυτή εφαρμόζεται πριν από

τη δημιουργία του μοντέλου. Η ανισορροπία των κλάσεων είναι ένα αρκετά συχνό πρόβλημα σε πολλά σετ δεδομένων, όπου μία κλάση εμφανίζεται συχνότερα από μια άλλη, με πιθανή συνέπεια την μείωση της απόδοσης του μοντέλου που θα δημιουργηθεί. Ο στόχος της τεχνικής αυτής είναι να δημιουργήσει επιπλέον στιγμιότυπα από την κλάση που εκπροσωπείται λιγότερο (minority class), με τρόπο που να αυξάνει την ισορροπία μεταξύ των κλάσεων, χωρίς ωστόσο να προσθέτει θόρυβο στα δεδομένα. (121)

Η τεχνική αυτή εφαρμόστηκε χρησιμοποιώντας το κατάλληλο φίλτρο (Filter→ filters→ supervised→ instance→ SMOTE). Δημιουργήθηκε ένας αριθμός παραδειγμάτων ίσος με τον αρχικό αριθμό των περιπτώσεων της κλάσης που εκπροσωπείται λιγότερο (μέγεθος υπερδειγματοληψίας 100%). Τα αποτελέσματα αποθηκεύτηκαν υπό μορφή αρχείου ARFF με το όνομα «data set 6.arff», και φαίνονται στον Πίνακα 7.

Μη ισορροπημένα δεδομένα	
Ίαση	19.325
Θάνατος	675
Σύνολο	20.000
Ισορροπημένα δεδομένα (τεχνική SMOTE)	
Ίαση	19.325
Θάνατος	1.350
Σύνολο	20.675

Πίνακας 7. Κατανομή δεδομένων στις κλάσεις πριν και μετά την τεχνική SMOTE.

5.3.2 Επιλογή χαρακτηριστικών

Όσον αφορά την επιλογή των χαρακτηριστικών (feature selection) στο στάδιο της προεπεξεργασίας δεδομένων, χρησιμοποιήθηκε ο αλγόριθμος **Information Gain Attribute Evaluation**. Αυτός ο αλγόριθμος λειτουργεί υπολογίζοντας την πληροφορία που παρέχεται από κάθε χαρακτηριστικό,

χρησιμοποιώντας το κέρδος πληροφορίας (Information Gain). Θέτει μια κατώτατη τιμή (threshold) και τα χαρακτηριστικά που ταξινομούνται πάνω από την τιμή αυτή, θα ληφθούν υπόψη στην δημιουργία του μοντέλου μηχανικής μάθησης. Τα χαρακτηριστικά ταξινομούνται από το πιο σημαντικό στο λιγότερο σημαντικό.(121)

Το κέρδος πληροφορίας δίνεται από την εξίσωση(122):

$$IG(Class, Attribute) = H(Class) - H(Class|Attribute)$$

Όπου H(Class) είναι η εντροπία της κλάσης πριν από την προσθήκη του χαρακτηριστικού και H(Class | Attribute) είναι η εντροπία της κατηγορίας υπό την προϋπόθεση ότι γνωρίζουμε την τιμή του χαρακτηριστικού.

Η τιμή "average rank" στα αποτελέσματα δείχνει το μέσο όρο της θέσης που καταλαμβάνει το κάθε χαρακτηριστικό στην αξιολόγηση, ενώ η τιμή "average merit" αφορά την μέση απόδοση του μοντέλου στα διάφορα υποσύνολα των δεδομένων. Τα χαρακτηριστικά με υψηλότερο average merit και average rank είναι πολύ σημαντικά για την πρόβλεψη της κατηγορίας του ασθενούς.

Τα αποτελέσματα αποθηκεύτηκαν υπό μορφή αρχείου κειμένου με το όνομα «Ranker+InforGainAttributeEval.txt» και δημιουργήθηκε ένα καινούριο αρχείο ARFF με το όνομα «data set 7.arff» το οποίο περιέχει τα πρώτα 20 χαρακτηριστικά της κατάταξης που προέκυψε, η οποία φαίνεται στον Πίνακα 8.

Χαρακτηριστικό	**Average merit**	**Average rank**
treatment(days)	0.193	1
sputum smear 4th month	0.106	2
sputum smear 6th month	0.104	3
sputum smear 2nd month	0.1	4
sputum smear 3rd month	0.042	5.1
sputum smear 5th month	0.041	5.9
sputum smear 1st month	0.027	7
age	0.021	8
sputum smear 2nd sample	0.019	9
HIV serology	0.014	10.4

other diseases	0.014	11.1
race	0.014	12.1
AIDS	0.013	12.4
alcohol consumption	0.011	14.4
sputum smear 1st sample	0.01	14.6
tuberculin skin test	0.008	16
chest X-ray	0.006	17
patient situation	0.005	18.4
state of Brazil	0.005	19
diabetes	0.004	21.1

Πίνακας 8. Κατάταξη χαρακτηριστικών με τη χρήση του Information Gain Attribute Evaluation.

5.4 Επιλογή αλγορίθμων

Στην παρούσα μελέτη εφαρμόστηκαν οι αλγόριθμοι ταξινόμησης **RF** και **SVM** οι οποίοι είναι ευρέως διαδεδομένοι και αναλύθηκαν στο κεφάλαιο 2.2.4. Οι αλγόριθμοι αυτοί είναι κατάλληλοι για προβλήματα δυαδικής ταξινόμησης, δηλαδή στην συγκεκριμένη περίπτωση την πρόβλεψη του θανάτου ή της ίασης. Επιπλέον, είναι αρκετά ευέλικτοι καθώς επιτρέπουν την ρύθμιση διαφόρων παραμέτρων, ιδιότητα που τους καθιστά ιδιαίτερα ισχυρούς στον τομέα της μηχανικής μάθησης. Ακόμη, είναι ικανοί στο να χειριστούν μεγάλα σύνολα δεδομένων, παρέχοντας καλή απόδοση. Επίσης, επιλέχθηκαν οι συγκεκριμένοι αλγόριθμοι επειδή παρέχουν μια διαφορετική προσέγγιση του συνόλου δεδομένων καθώς λειτουργούν με αρκετά διαφορετικό τρόπο συγκριτικά μεταξύ τους. Τέλος, οι αλγόριθμοι RF και SVM έχουν χρησιμοποιηθεί αποτελεσματικά σε προηγούμενες μελέτες. (113)(114)(115)

6. Αποτελέσματα

6.1 Random Forest

Μέρος Α

Στα πειράματα που παρουσιάζονται παρακάτω, έγινε εφαρμογή του RF στο σύνολο δεδομένων πριν και μετά τη χρήση της τεχνικής SMOTE.

1ο πείραμα

Το 1ο πείραμα αφορά την εφαρμογή του RF στο αρχείο με όνομα «data set 5.arff», δηλαδή στο σετ δεδομένων χωρίς την χρήση της τεχνικής υπερδειγματοληψίας SMOTE και χωρίς την επιλογή χαρακτηριστικών, με σύνολο 20.000 δείγματα. Το πλήθος των δέντρων ορίστηκε σε 100. (παράμετρος numIterations). Το μοντέλο εκπαιδεύτηκε και αξιολογήθηκε χρησιμοποιώντας την τεχνική της διασταυρούμενης επικύρωσης με k=10 (10-fold cross-validation).

Ο χρόνος που χρειάστηκε για να κατασκευαστεί το μοντέλο ήταν 2,99 δευτερόλεπτα. Το ποσοστό των σωστά ταξινομημένων δειγμάτων στο σύνολο δεδομένων ήταν 98.37% (19674 δείγματα), ενώ το ποσοστό των λάθος ταξινομημένων δειγμάτων ήταν 1.63% (326 δείγματα). Ο συντελεστής κάπα (Kappa statistic) ήταν ίσος με 0.7245. Το μέσο απόλυτο σφάλμα (Mean absolute error) ήταν ίσο με 0.0311, ενώ το σχετικό απόλυτο σφάλμα (Relative absolute error) ήταν ίσο με 47.6566%.

Στον Πίνακα 9, παρουσιάζονται τα αποτελέσματα από τις μετρικές απόδοσης του 1ου πειράματος για τις κλάσεις ξεχωριστά, καθώς και ο σταθμισμένος μέσος όρος (weighted average) ο οποίος λαμβάνει υπόψη τον αριθμό των δειγμάτων σε κάθε κλάση.

	TPR	FPR	Precision	Recall	F-measure	ROC Area	PRC Area	Κλάση
	0,995	0,338	0,988	0,995	0,992	0,970	0,998	1
	0,662	0,005	0,820	0,662	0,733	0,970	0,790	3
Weighted avg.	0,984	0,327	0,983	0,984	0,983	0,970	0,991	

Πίνακας 9. Μετρικές απόδοσης 1ου πειράματος του RF.

Cross-validation			
TARGET / OUTPUT	Ίαση	Θάνατος	SUM
Ίαση	19227 96.14%	228 1.14%	19455 98.83% 1.17%
Θάνατος	98 0.49%	447 2.23%	545 82.02% 17.98%
SUM	19325 99.49% 0.51%	675 66.22% 33.78%	19674 / 20000 98.37% 1.63%

Εικόνα 21: Πίνακας σύγχυσης 1ου πειράματος του RF.

Ως "target" αναφέρεται η πραγματική κλάση, ενώ ως "output" η κλάση πρόβλεψης. [Δημιουργήθηκε με τη χρήση του Confusion Matrix Generator (https://www.damianoperri.it/public/confusionMatrix/).]

2ο πείραμα

Το 2ο πείραμα αφορά την εφαρμογή του RF αλλά διαφέρει από το 1ο ως το πλήθος των δέντρων που ορίστηκαν 200 (παράμετρος numIterations). Το μοντέλο εκπαιδεύτηκε και αξιολογήθηκε χρησιμοποιώντας την τεχνική της διασταυρούμενης επικύρωσης με k=10 (10-fold cross-validation).

Ο χρόνος που χρειάστηκε για να κατασκευαστεί το μοντέλο ήταν 2 δευτερόλεπτα. Το ποσοστό των σωστά ταξινομημένων δειγμάτων στο σύνολο δεδομένων ήταν 98.31% (19663 δείγματα), ενώ το ποσοστό των λάθος ταξινομημένων δειγμάτων ήταν 1.68% (337 δείγματα). Ο συντελεστής κάπα (Kappa statistic) ήταν ίσος με 0.7145. Το μέσο απόλυτο σφάλμα (Mean absolute error) ήταν ίσο με 0.031, ενώ το σχετικό απόλυτο σφάλμα (Relative absolute error) ήταν ίσο με 47.5585%.

Στον Πίνακα 10, παρουσιάζονται τα αποτελέσματα από τις μετρικές απόδοσης του 2ου πειράματος για τις κλάσεις ξεχωριστά, καθώς και ο σταθμισμένος μέσος όρος (weighted average).

	TPR	FPR	Precision	Recall	F-measure	ROC Area	PRC Area	Κλάση
	0,995	0,348	0,988	0,995	0,991	0,970	0,998	1
	0,652	0,005	0,812	0,652	0,723	0,970	0,790	3
Weighted avg.	0,983	0,337	0,982	0,983	0,982	0,970	0,991	

Πίνακας 10. Μετρικές απόδοσης 2ου πειράματος του RF.

Cross-validation			
TARGET / OUTPUT	Ίαση	Θάνατος	SUM
Ίαση	19223 96.11%	235 1.18%	19458 98.79% 1.21%
Θάνατος	102 0.51%	440 2.20%	542 81.18% 18.82%
SUM	19325 99.47% 0.53%	675 65.19% 34.81%	19663 / 20000 98.31% 1.69%

Εικόνα 22: Πίνακας σύγχυσης 2ου πειράματος του RF.

Ως "target" αναφέρεται η πραγματική κλάση, ενώ ως "output" η κλάση πρόβλεψης. [Δημιουργήθηκε με τη χρήση του Confusion Matrix Generator (https://www.damianoperri.it/public/confusionMatrix/).]

3ο πείραμα

Το 3ο πείραμα αφορά την εφαρμογή του RF στο αρχείο με όνομα «data set 6.arff», δηλαδή στο σετ δεδομένων με την χρήση της τεχνικής υπερδειγματοληψίας SMOTE αλλά χωρίς την επιλογή χαρακτηριστικών, με σύνολο 20.675 δείγματα. Το πλήθος των δέντρων ορίστηκε σε 100. Το μοντέλο εκπαιδεύτηκε και αξιολογήθηκε χρησιμοποιώντας την τεχνική της διασταυρούμενης επικύρωσης με k=10 (10-fold cross-validation).

Ο χρόνος που χρειάστηκε για να κατασκευαστεί το μοντέλο ήταν 2,12 δευτερόλεπτα. Το ποσοστό των σωστά ταξινομημένων δειγμάτων στο σύνολο δεδομένων ήταν 98.17% (20297 δείγματα), ενώ το ποσοστό των λάθος ταξινομημένων δειγμάτων ήταν 1.82% (378 δείγματα). Ο συντελεστής κάπα (Kappa statistic) ήταν ίσος με 0.8463. Το μέσο απόλυτο σφάλμα (Mean absolute error) ήταν ίσο με 0.0372, ενώ το σχετικό απόλυτο σφάλμα (Relative absolute error) ήταν ίσο με 30.4822 %.

Στον Πίνακα 11, παρουσιάζονται τα αποτελέσματα από τις μετρικές απόδοσης του 3ου πειράματος για τις κλάσεις ξεχωριστά, καθώς και ο σταθμισμένος μέσος όρος (weighted average).

	TPR	FPR	Precision	Recall	F-measure	ROC Area	PRC Area	Κλάση
	0,992	0,167	0,988	0,992	0,990	0,989	0,999	1
	0,833	0,008	0,881	0,833	0,856	0,989	0,930	3
Weighted avg.	0,982	0,157	0,981	0,982	0,981	0,989	0,995	

Πίνακας 11. Μετρικές απόδοσης 3ου πειράματος του RF.

Cross-validation			
TARGET / OUTPUT	Ίαση	Θάνατος	SUM
Ίαση	19173 92.74%	226 1.09%	19399 98.83% 1.17%
Θάνατος	152 0.74%	1124 5.44%	1276 88.09% 11.91%
SUM	19325 99.21% 0.79%	1350 83.26% 16.74%	20297 / 20675 98.17% 1.83%

Εικόνα 23: Πίνακας σύγχυσης 3ου πειράματος του RF.

Ως "target" αναφέρεται η πραγματική κλάση, ενώ ως "output" η κλάση πρόβλεψης. [Δημιουργήθηκε με τη χρήση του Confusion Matrix Generator (https://www.damianoperri.it/public/confusionMatrix/).]

4ο πείραμα

Το 4ο πείραμα διαφοροποιήθηκε από το 3ο καθώς το πλήθος των δέντρων ορίστηκε σε 200 (παράμετρος numIterations).

Ο χρόνος που χρειάστηκε για να κατασκευαστεί το μοντέλο ήταν 4,02 δευτερόλεπτα. Το ποσοστό των σωστά ταξινομημένων δειγμάτων στο σύνολο δεδομένων ήταν 98.15% (20294 δείγματα), ενώ το ποσοστό των λάθος ταξινομημένων δειγμάτων ήταν 1.84% (381 δείγματα). Ο συντελεστής κάπα (Kappa statistic) ήταν ίσος με 0.845. Το μέσο απόλυτο σφάλμα (Mean absolute error) ήταν ίσο με 0.0373, ενώ το σχετικό απόλυτο σφάλμα (Relative absolute error) ήταν ίσο με 30.5177%.

Στον Πίνακα 12, παρουσιάζονται τα αποτελέσματα από τις μετρικές απόδοσης του 4ου πειράματος για τις κλάσεις ξεχωριστά, καθώς και ο σταθμισμένος μέσος όρος (weighted average).

	TPR	FPR	Precision	Recall	F-measure	ROC Area	PRC Area	Κλάση
	0,992	0,169	0,988	0,992	0,990	0,990	0,999	1
	0,831	0,008	0,880	0,831	0,855	0,990	0,931	3
Weighted avg.	0,982	0,158	0,981	0,982	0,981	0,990	0,995	

Πίνακας 12. Μετρικές απόδοσης 4ου πειράματος του RF.

Cross-validation			
TARGET / OUTPUT	Ίαση	Θάνατος	SUM
Ίαση	19172 92.73%	228 1.10%	19400 98.82% 1.18%
Θάνατος	153 0.74%	1122 5.43%	1275 88.00% 12.00%
SUM	19325 99.21% 0.79%	1350 83.11% 16.89%	20294 / 20675 98.16% 1.84%

Εικόνα 24: Πίνακας σύγχυσης 4ου πειράματος του RF.

Ως "target" αναφέρεται η πραγματική κλάση, ενώ ως "output" η κλάση πρόβλεψης. [Δημιουργήθηκε με τη χρήση του Confusion Matrix Generator (https://www.damianoperri.it/public/confusionMatrix/).]

Μέρος Β

Εν συνεχεία, έγινε εφαρμογή του RF στο σύνολο δεδομένων με τη χρήση της τεχνικής SMOTE και του Information Gain Attribute Evaluation.

5ο πείραμα

Το 5ο πείραμα αφορά την εφαρμογή του RF στο αρχείο με όνομα «data set 7.arff», δηλαδή στο σετ δεδομένων με την χρήση της τεχνικής υπερδειγματοληψίας SMOTE και με την επιλογή 20 χαρακτηριστικών βάση των αποτελεσμάτων του Information Gain Attribute Evaluation, σε σύνολο 20.675 δείγματα. Το πλήθος των επαναλήψεων ορίστηκε σε 100. (παράμετρος numIterations). Το μοντέλο εκπαιδεύτηκε και αξιολογήθηκε

χρησιμοποιώντας την τεχνική της διασταυρούμενης επικύρωσης με k=10 (10-fold cross-validation).

Ο χρόνος που χρειάστηκε για να κατασκευαστεί το μοντέλο ήταν 6,51 δευτερόλεπτα. Το ποσοστό των σωστά ταξινομημένων δειγμάτων στο σύνολο δεδομένων ήταν 98.0314 % (20268 δείγματα), ενώ το ποσοστό των λάθος ταξινομημένων δειγμάτων ήταν 1.9686% (407 δείγματα). Ο συντελεστής κάπα (Kappa statistic) ήταν ίσος με 0.8373. Το μέσο απόλυτο σφάλμα (Mean absolute error) ήταν ίσο με 0.0343, ενώ το σχετικό απόλυτο σφάλμα (Relative absolute error) ήταν ίσο με 28.1305%.

Στον Πίνακα 13, παρουσιάζονται τα αποτελέσματα από τις μετρικές απόδοσης του 5ου πειράματος για τις κλάσεις ξεχωριστά, καθώς και ο σταθμισμένος μέσος όρος (weighted average).

	TPR	FPR	Precision	Recall	F-measure	ROC Area	PRC Area	Κλάση
	0,990	0,160	0,989	0,990	0,989	0,987	0,999	1
	0,840	0,010	0,856	0,840	0,848	0,987	0,907	3
Weighted avg.	0,980	0,150	0,980	0,980	0,980	0,987	0,993	

Πίνακας 13. Μετρικές απόδοσης 5ου πειράματος του RF.

Cross-validation			
TARGET / OUTPUT	Ίαση	Θάνατος	SUM
Ίαση	19134 92.55%	216 1.04%	19350 98.88% 1.12%
Θάνατος	191 0.92%	1134 5.48%	1325 85.58% 14.42%
SUM	19325 99.01% 0.99%	1350 84.00% 16.00%	20268 / 20675 98.03% 1.97%

Εικόνα 25: Πίνακας σύγχυσης 5ου πειράματος του RF.

Ως "target" αναφέρεται η πραγματική κλάση, ενώ ως "output" η κλάση πρόβλεψης. [Δημιουργήθηκε με τη χρήση του Confusion Matrix Generator (https://www.damianoperri.it/public/confusionMatrix/).]

6ο πείραμα

Το 6ο πείραμα διαφέρει από το 5ο ως προς το πλήθος των δέντρων το οποίο ορίστηκε σε 200. (παράμετρος numIterations). Το μοντέλο εκπαιδεύτηκε και αξιολογήθηκε χρησιμοποιώντας την τεχνική της διασταυρούμενης επικύρωσης με k=10 (10-fold cross-validation).

Ο χρόνος που χρειάστηκε για να κατασκευαστεί το μοντέλο ήταν 14,48 δευτερόλεπτα. Το ποσοστό των σωστά ταξινομημένων δειγμάτων στο σύνολο δεδομένων ήταν 98.0121% (20264 δείγματα), ενώ το ποσοστό των λάθος ταξινομημένων δειγμάτων ήταν 1.9879% (411 δείγματα). Ο συντελεστής κάπα (Kappa statistic) ήταν ίσος με 0.8356. Το μέσο απόλυτο σφάλμα (Mean absolute error) ήταν ίσο με 0.0345, ενώ το σχετικό απόλυτο σφάλμα (Relative absolute error) ήταν ίσο με 28.2681%.

Στον Πίνακα 14, παρουσιάζονται τα αποτελέσματα από τις μετρικές απόδοσης του 6ου πειράματος για τις κλάσεις ξεχωριστά, καθώς και ο σταθμισμένος μέσος όρος (weighted average).

	TPR	FPR	Precision	Recall	F-measure	ROC Area	PRC Area	Κλάση
	0,990	0,162	0,989	0,990	0,989	0,987	0,999	1
	0,838	0,010	0,855	0,838	0,846	0,987	0,908	3
Weighted avg.	0,980	0,152	0,980	0,980	0,980	0,987	0,993	

Πίνακας 14. Μετρικές απόδοσης 6ου πειράματος του RF.

Cross-validation			
TARGET / OUTPUT	Ίαση	Θάνατος	SUM
Ίαση	19133 92.55%	219 1.06%	19352 98.87% 1.13%
Θάνατος	191 0.92%	1131 5.47%	1322 85.55% 14.45%
SUM	19324 99.01% 0.99%	1350 83.78% 16.22%	20264 / 20674 98.02% 1.98%

Εικόνα 26: Πίνακας σύγχυσης 6ου πειράματος του RF.

Ως "target" αναφέρεται η πραγματική κλάση, ενώ ως "output" η κλάση πρόβλεψης. [Δημιουργήθηκε με τη χρήση του Confusion Matrix Generator (https://www.damianoperri.it/public/confusionMatrix/).]

6.2 Support Vector Machines

Μέρος Α

Παρακάτω αναλύονται τα μοντέλα SVM με την καλύτερη απόδοση για κάθε συνάρτηση πυρήνα. Δημιουργήθηκαν χρησιμοποιώντας το σύνολο δεδομένων με και χωρίς την χρήση της τεχνικής SMOTE, ρυθμίζοντας διάφορες παραμέτρους. Τα αποτελέσματα των διαφορετικών παραμέτρων σε κάθε συνάρτηση πυρήνα είχαν μικρές αποκλίσεις μεταξύ τους. Έτσι, η καλύτερη απόδοση των μοντέλων επικεντρώθηκε κυρίως στην ικανότητα τους να προβλέπουν σωστά την κλάση που αφορά την έκβαση του θανάτου.

1ο πείραμα

Το 1ο πείραμα αφορά την εφαρμογή του **SVM** στο αρχείο με όνομα «data set 5.arff», δηλαδή στο σετ δεδομένων χωρίς την χρήση της τεχνικής υπερδειγματοληψίας SMOTE και χωρίς την επιλογή χαρακτηριστικών, με σύνολο 20.000 δείγματα. Ως συνάρτηση πυρήνα επιλέχθηκε η **γραμμική**. Το μοντέλο εκπαιδεύτηκε και αξιολογήθηκε χρησιμοποιώντας την τεχνική της διασταυρούμενης επικύρωσης με k=10 (10-fold cross-validation).

Πραγματοποιήθηκαν δοκιμές με διάφορες τιμές για την παράμετρο κόστους C. Καλύτερη απόδοση εμφάνισε το μοντέλο με C=50.

Ο χρόνος που χρειάστηκε για να κατασκευαστεί το μοντέλο ήταν 289,48 δευτερόλεπτα. Το ποσοστό των σωστά ταξινομημένων δειγμάτων στο σύνολο δεδομένων ήταν 98.355% (19671 δείγματα), ενώ το ποσοστό των λάθος ταξινομημένων δειγμάτων ήταν 1.645% (329 δείγματα). Ο συντελεστής κάπα (Kappa statistic) ήταν ίσος με 0.7462. Το μέσο απόλυτο σφάλμα (Mean absolute error) ήταν ίσο με 0.0164, ενώ το σχετικό απόλυτο σφάλμα (Relative absolute error) ήταν ίσο με 25.2029%.

Στον Πίνακα 15, παρουσιάζονται τα αποτελέσματα από τις μετρικές απόδοσης του 1ου πειράματος για τις κλάσεις ξεχωριστά, καθώς και ο σταθμισμένος μέσος όρος (weighted average).

	TPR	FPR	Precision	Recall	F-measure	ROC Area	PRC Area	Κλάση
	0,992	0,250	0,991	0,992	0,991	0,871	0,991	1
	0,750	0,008	0,760	0,750	0,755	0,871	0,578	3
Weighted avg.	0,984	0,242	0,983	0,984	0,983	0,871	0,997	

Πίνακας 15. Μετρικές απόδοσης 1ου πειράματος του SVM.

Cross-validation			
TARGET / OUTPUT	Ίαση	Θάνατος	SUM
Ίαση	19165 95.83%	169 0.84%	19334 99.13% 0.87%
Θάνατος	160 0.80%	506 2.53%	666 75.98% 24.02%
SUM	19325 99.17% 0.83%	675 74.96% 25.04%	19671 / 20000 98.36% 1.64%

Εικόνα 27: Πίνακας σύγχυσης 1ου πειράματος του SVM.

Ως "target" αναφέρεται η πραγματική κλάση, ενώ ως "output" η κλάση πρόβλεψης. [Δημιουργήθηκε με τη χρήση του Confusion Matrix Generator (https://www.damianoperri.it/public/confusionMatrix/).]

2ο πείραμα

Το 2ο πείραμα διαφοροποιείται από το 1ο, ως προς τη συνάρτηση πυρήνα που επιλέχθηκε, δηλαδή την **πολυωνυμική**. Το μοντέλο εκπαιδεύτηκε και αξιολογήθηκε χρησιμοποιώντας την τεχνική της διασταυρούμενης επικύρωσης με k=10 (10-fold cross-validation).

Πραγματοποιήθηκαν δοκιμές με διάφορες τιμές για τις παραμέτρους C, coef0 και degree. Καλύτερη απόδοση εμφάνισε το μοντέλο με C=1, coef0=1 και degree=3. Ο χρόνος που χρειάστηκε για να κατασκευαστεί το μοντέλο ήταν 5,54 δευτερόλεπτα. Το ποσοστό των σωστά ταξινομημένων δειγμάτων στο σύνολο δεδομένων ήταν 98.395% (19679 δείγματα), ενώ το ποσοστό των λάθος ταξινομημένων δειγμάτων ήταν 1.605% (321 δείγματα). Ο συντελεστής κάπα (Kappa statistic) ήταν ίσος με 0.7569. Το μέσο απόλυτο σφάλμα (Mean absolute error) ήταν ίσο με 0.016, ενώ το σχετικό απόλυτο σφάλμα (Relative absolute error) ήταν ίσο με 24.5901%.

Στον Πίνακα 16, παρουσιάζονται τα αποτελέσματα από τις μετρικές απόδοσης του 2ου πειράματος για τις κλάσεις ξεχωριστά, καθώς και ο σταθμισμένος μέσος όρος (weighted average).

	TPR	FPR	Precision	Recall	F-measure	ROC Area	PRC Area	Κλάση
	0,991	0,225	0,992	0,992	0,992	0,883	0,992	1
	0,775	0,009	0,756	0,775	0,765	0,883	0,593	3
Weighted avg.	0,984	0,218	0,984	0,984	0,984	0,883	0,978	

Πίνακας 16. Μετρικές απόδοσης 2ου πειράματος του SVM.

Cross-validation			
TARGET / OUTPUT	Ίαση	Θάνατος	SUM
Ίαση	19156 95.78%	152 0.76%	19308 99.21% 0.79%
Θάνατος	169 0.84%	523 2.62%	692 75.58% 24.42%
SUM	19325 99.13% 0.87%	675 77.48% 22.52%	19679 / 20000 98.39% 1.61%

Εικόνα 28: Πίνακας σύγχυσης 2ου πειράματος του SVM.

Ως "target" αναφέρεται η πραγματική κλάση, ενώ ως "output" η κλάση πρόβλεψης. [Δημιουργήθηκε με τη χρήση του Confusion Matrix Generator (https://www.damianoperri.it/public/confusionMatrix/).

3ο πείραμα

Το 3ο πείραμα διαφοροποιείται από το 1ο ως προς τη συνάρτηση πυρήνα που επιλέχθηκε, δηλαδή την **σιγμοειδή**. Το μοντέλο εκπαιδεύτηκε και αξιολογήθηκε χρησιμοποιώντας την τεχνική της διασταυρούμενης επικύρωσης με k=10 (10-fold cross-validation).

Πραγματοποιήθηκαν δοκιμές με διάφορες τιμές για τις παραμέτρους C, coef0 και gamma. Καλύτερη απόδοση εμφάνισε το μοντέλο με C=10, coef0=1 και gamma= 0. Ο χρόνος που χρειάστηκε για να κατασκευαστεί το μοντέλο ήταν 7 δευτερόλεπτα. Το ποσοστό των σωστά ταξινομημένων δειγμάτων στο σύνολο δεδομένων ήταν 98.395% (19679 δείγματα), ενώ το ποσοστό των λάθος ταξινομημένων δειγμάτων ήταν 1.605% (321 δείγματα). Ο συντελεστής κάπα (Kappa statistic) ήταν ίσος με 0.7544. Το μέσο απόλυτο σφάλμα (Mean absolute error) ήταν ίσο με 0.016, ενώ το σχετικό απόλυτο σφάλμα (Relative absolute error) ήταν ίσο με 24.5901%.

Στον Πίνακα 17, παρουσιάζονται τα αποτελέσματα από τις μετρικές απόδοσης του 3ου πειράματος για τις κλάσεις ξεχωριστά, καθώς και ο σταθμισμένος μέσος όρος (weighted average).

	TPR	FPR	Precision	Recall	F-measure	ROC Area	PRC Area	Κλάση
	0,992	0,236	0,992	0,992	0,992	0,878	0,992	1
	0,764	0,008	0,761	0,764	0,763	0,878	0,590	3
Weighted avg.	0,984	0,228	0,984	0,984	0,984	0,878	0,978	

Πίνακας 17. Μετρικές απόδοσης 3ου πειράματος του SVM.

Cross-validation			
TARGET / OUTPUT	Ίαση	Θάνατος	SUM
Ίαση	19163 95.81%	159 0.80%	19322 99.18% 0.82%
Θάνατος	162 0.81%	516 2.58%	678 76.11% 23.89%
SUM	19325 99.16% 0.84%	675 76.44% 23.56%	19679 / 20000 98.39% 1.61%

Εικόνα 29: Πίνακας σύγχυσης 3ου πειράματος του SVM.

Ως "target" αναφέρεται η πραγματική κλάση, ενώ ως "output" η κλάση πρόβλεψης. [Δημιουργήθηκε με τη χρήση του Confusion Matrix Generator (https://www.damianoperri.it/public/confusionMatrix/).]

4ο πείραμα

Το 4ο πείραμα διαφοροποιείται από το 1ο ως προς τη συνάρτηση πυρήνα που επιλέχθηκε, δηλαδή την **RBF**. Το μοντέλο εκπαιδεύτηκε και αξιολογήθηκε χρησιμοποιώντας την τεχνική της διασταυρούμενης επικύρωσης με k=10 (10-fold cross-validation).

Πραγματοποιήθηκαν δοκιμές με διάφορες τιμές για τις παραμέτρους C και gamma. Καλύτερη απόδοση εμφάνισε το μοντέλο με C=5 και gamma= 0. Ο χρόνος που χρειάστηκε για να κατασκευαστεί το μοντέλο ήταν 3,88 δευτερόλεπτα. Το ποσοστό των σωστά ταξινομημένων δειγμάτων στο σύνολο δεδομένων ήταν 98.445% (19689 δείγματα), ενώ το ποσοστό των λάθος

ταξινομημένων δειγμάτων ήταν 1.555% (311 δείγματα). Ο συντελεστής κάπα (Kappa statistic) ήταν ίσος με 0.7544. Το μέσο απόλυτο σφάλμα (Mean absolute error) ήταν ίσο με 0.0155, ενώ το σχετικό απόλυτο σφάλμα (Relative absolute error) ήταν ίσο με 23.824%.

Στον Πίνακα 18, παρουσιάζονται τα αποτελέσματα από τις μετρικές απόδοσης του 4ου πειράματος για τις κλάσεις ξεχωριστά, καθώς και ο σταθμισμένος μέσος όρος (weighted average).

	TPR	FPR	Precision	Recall	F-measure	ROC Area	PRC Area	Κλάση
	0,992	0,227	0,992	0,992	0,992	0,883	0,992	1
	0,773	0,008	0,768	0,773	0,770	0,883	0,601	3
Weighted avg.	0,984	0,219	0,985	0,984	0,984	0,883	0,979	

Πίνακας 18. Μετρικές απόδοσης 4ου πειράματος του SVM.

Cross-validation			
TARGET / OUTPUT	Ίαση	Θάνατος	SUM
Ίαση	19167 95.83%	153 0.77%	19320 99.21% 0.79%
Θάνατος	158 0.79%	522 2.61%	680 76.76% 23.24%
SUM	19325 99.18% 0.82%	675 77.33% 22.67%	19689 / 20000 98.45% 1.55%

Εικόνα 30: Πίνακας σύγχυσης 4ου πειράματος του SVM.

Ως "target" αναφέρεται η πραγματική κλάση, ενώ ως "output" η κλάση πρόβλεψης. [Δημιουργήθηκε με τη χρήση του Confusion Matrix Generator (https://www.damianoperri.it/public/confusionMatrix/).]

5ο πείραμα

Το 5ο πείραμα αφορά την εφαρμογή του **SVM** στο αρχείο με όνομα «data set 6.arff», δηλαδή στο σετ δεδομένων με την χρήση της τεχνικής υπερδειγματοληψίας SMOTE και χωρίς την επιλογή χαρακτηριστικών, με σύνολο 20.675 δείγματα. Ως συνάρτηση πυρήνα επιλέχθηκε η **γραμμική**. Το

μοντέλο εκπαιδεύτηκε και αξιολογήθηκε χρησιμοποιώντας την τεχνική της διασταυρούμενης επικύρωσης με k=10 (10-fold cross-validation).

Τα μοντέλα με C=10 και C=20 εμφάνισαν ακριβώς την ίδια απόδοση. Επιλέχθηκε ως πιο αποδοτικό το μοντέλο με C=10, καθώς χρειάστηκε λιγότερο χρόνο για να κατασκευαστεί (135,73 δευτερόλεπτα). Το ποσοστό των σωστά ταξινομημένων δειγμάτων στο σύνολο δεδομένων ήταν 97.8138% (20.223 δείγματα), ενώ το ποσοστό των λάθος ταξινομημένων δειγμάτων ήταν 2.1862 % (452 δείγματα). Ο συντελεστής κάπα (Kappa statistic) ήταν ίσος με 0.8186. Το μέσο απόλυτο σφάλμα (Mean absolute error) ήταν ίσο με 0.0219, ενώ το σχετικό απόλυτο σφάλμα (Relative absolute error) ήταν ίσο με 17.9043%

Στον πίνακα 19, παρουσιάζονται τα αποτελέσματα από τις μετρικές απόδοσης του 5ου πειράματος για τις κλάσεις ξεχωριστά, καθώς και ο σταθμισμένος μέσος όρος (weighted average).

	TPR	FPR	Precision	Recall	F-measure	ROC Area	PRC Area	Κλάση
	0,989	0,181	0,987	0,989	0,988	0,904	0,987	1
	0,819	0,011	0,842	0,819	0,830	0,904	0,701	3
Weighted avg.	0,978	0,170	0,978	0,978	0,978	0,904	0,968	

Πίνακας 19. Μετρικές απόδοσης 5ου πειράματος του SVM.

Cross-validation			
TARGET / OUTPUT	Ίαση	Θάνατος	SUM
Ίαση	19117 92.46%	244 1.18%	19361 98.74% 1.26%
Θάνατος	208 1.01%	1106 5.35%	1314 84.17% 15.83%
SUM	19325 98.92% 1.08%	1350 81.93% 18.07%	20223 / 20675 97.81% 2.19%

Εικόνα 31: Πίνακας σύγχυσης 5ου πειράματος του SVM.

Ως "target" αναφέρεται η πραγματική κλάση, ενώ ως "output" η κλάση πρόβλεψης. [Δημιουργήθηκε με τη χρήση του Confusion Matrix Generator (https://www.damianoperri.it/public/confusionMatrix/).]

6ο πείραμα

Το 6ο πείραμα διαφοροποιείται από το 5ο, καθώς επιλέχθηκε η **πολυωνυμική** συνάρτηση πυρήνα. Το μοντέλο εκπαιδεύτηκε και αξιολογήθηκε χρησιμοποιώντας την τεχνική της διασταυρούμενης επικύρωσης με k=10 (10-fold cross-validation).

Πραγματοποιήθηκαν δοκιμές με διάφορες τιμές για τις παραμέτρους C, coef0 και degree. Καλύτερη απόδοση εμφάνισε το μοντέλο με C=50, coef0=1 και degree= 3. Ο χρόνος που χρειάστηκε για να κατασκευαστεί το μοντέλο ήταν 23,79 δευτερόλεπτα.

Το ποσοστό των σωστά ταξινομημένων δειγμάτων στο σύνολο δεδομένων ήταν 98.2104% (20305 δείγματα), ενώ το ποσοστό των λάθος ταξινομημένων δειγμάτων ήταν 1.7896% (370 δείγματα). Ο συντελεστής κάπα (Kappa statistic) ήταν ίσος με 0.8527. Το μέσο απόλυτο σφάλμα (Mean absolute error) ήταν ίσο με 0.0179, ενώ το σχετικό απόλυτο σφάλμα (Relative absolute error) ήταν ίσο με 14.6561%.

Στον Πίνακα 20, παρουσιάζονται τα αποτελέσματα από τις μετρικές απόδοσης του 6ου πειράματος για τις κλάσεις ξεχωριστά, καθώς και ο σταθμισμένος μέσος όρος (weighted average).

	TPR	FPR	Precision	Recall	F-measure	ROC Area	PRC Area	Κλάση
	0,991	0,142	0,990	0,991	0,990	0,924	0,990	1
	0,858	0,009	0,867	0,858	0,862	0,924	0,753	3
Weighted avg.	0,982	0,134	0,982	0,982	0,982	0,924	0,974	

Πίνακας 20. Μετρικές απόδοσης 6ου πειράματος του SVM.

Cross-validation			
TARGET / OUTPUT	Ίαση	Θάνατος	SUM
Ίαση	19147 92.61%	192 0.93%	19339 99.01% 0.99%
Θάνατος	178 0.86%	1158 5.60%	1336 86.68% 13.32%
SUM	19325 99.08% 0.92%	1350 85.78% 14.22%	20305 / 20675 98.21% 1.79%

Εικόνα 32: Πίνακας σύγχυσης 6ου πειράματος του SVM.

Ως "target" αναφέρεται η πραγματική κλάση, ενώ ως "output" η κλάση πρόβλεψης. [Δημιουργήθηκε με τη χρήση του Confusion Matrix Generator (https://www.damianoperri.it/public/confusionMatrix/).]

7ο πείραμα

Το 7ο πείραμα διαφοροποιείται από το 5ο ως προς τη συνάρτηση πυρήνα που επιλέχθηκε, δηλαδή την **σιγμοειδή**. Το μοντέλο εκπαιδεύτηκε και αξιολογήθηκε χρησιμοποιώντας την τεχνική της διασταυρούμενης επικύρωσης με k=10 (10-fold cross-validation).

Πραγματοποιήθηκαν δοκιμές με διάφορες τιμές για τις παραμέτρους C, coef0 και gamma. Καλύτερη απόδοση εμφάνισε το μοντέλο με C=50, coef0=2 και gamma= 0. Ο χρόνος που χρειάστηκε για να κατασκευαστεί το μοντέλο ήταν 23,1 δευτερόλεπτα. Το ποσοστό των σωστά ταξινομημένων δειγμάτων στο σύνολο δεδομένων ήταν 97.7025% (20.200 δείγματα), ενώ το ποσοστό των λάθος ταξινομημένων δειγμάτων ήταν 2.2975% (475 δείγματα). Ο συντελεστής κάπα (Kappa statistic) ήταν ίσος με 0.8128. Το μέσο απόλυτο σφάλμα (Mean absolute error) ήταν ίσο με 0.023, ενώ το σχετικό απόλυτο σφάλμα (Relative absolute error) ήταν ίσο με 18.8153%.

Στον Πίνακα 21, παρουσιάζονται τα αποτελέσματα από τις μετρικές απόδοσης του 7ου πειράματος για τις κλάσεις ξεχωριστά, καθώς και ο σταθμισμένος μέσος όρος (weighted average).

	TPR	FPR	Precision	Recall	F-measure	ROC Area	PRC Area	Κλάση
	0,987	0,170	0,988	0,987	0,988	0,908	0,987	1
	0,830	0,013	0,821	0,830	0,825	0,908	0,692	3
Weighted avg.	0,977	0,160	0,977	0,977	0,977	0,908	0,968	

Πίνακας 21. Μετρικές απόδοσης 7ου πειράματος του SVM.

Cross-validation			
TARGET / OUTPUT	Ίαση	Θάνατος	SUM
Ίαση	19080 92.29%	230 1.11%	19310 98.81% 1.19%
Θάνατος	245 1.19%	1120 5.42%	1365 82.05% 17.95%
SUM	19325 98.73% 1.27%	1350 82.96% 17.04%	20200 / 20675 97.70% 2.30%

Εικόνα 33: Πίνακας σύγχυσης 7ου πειράματος του SVM.

Ως "target" αναφέρεται η πραγματική κλάση, ενώ ως "output" η κλάση πρόβλεψης. [Δημιουργήθηκε με τη χρήση του Confusion Matrix Generator (https://www.damianoperri.it/public/confusionMatrix/).]

8ο πείραμα

Το 8ο πείραμα διαφοροποιείται από το 5ο ως προς τη συνάρτηση πυρήνα που επιλέχθηκε, δηλαδή την **RBF**. Το μοντέλο εκπαιδεύτηκε και αξιολογήθηκε χρησιμοποιώντας την τεχνική της διασταυρούμενης επικύρωσης με k=10 (10-fold cross-validation).

Πραγματοποιήθηκαν δοκιμές με διάφορες τιμές για τις παραμέτρους C και gamma. Καλύτερη απόδοση εμφάνισε το μοντέλο με C=50 και gamma= 0. Ο χρόνος που χρειάστηκε για να κατασκευαστεί το μοντέλο ήταν 16,91 δευτερόλεπτα. Το ποσοστό των σωστά ταξινομημένων δειγμάτων στο σύνολο δεδομένων ήταν 98.1814% (20299 δείγματα), ενώ το ποσοστό των λάθος ταξινομημένων δειγμάτων ήταν 1.8186% (376 δείγματα). Ο συντελεστής κάπα (Kappa statistic) ήταν ίσος με 0.8503. Το μέσο απόλυτο σφάλμα (Mean absolute error) ήταν ίσο με 0.0182, ενώ το σχετικό απόλυτο σφάλμα (Relative absolute error) ήταν ίσο με 14.8938%.

Στον Πίνακα 22, παρουσιάζονται τα αποτελέσματα από τις μετρικές απόδοσης του 8ου πειράματος για τις κλάσεις ξεχωριστά, καθώς και ο σταθμισμένος μέσος όρος (weighted average).

	TPR	FPR	Precision	Recall	F-measure	ROC Area	PRC Area	Κλάση
	0,991	0,144	0,990	0,991	0,990	0,923	0,989	1
	0,856	0,009	0,865	0,856	0,860	0,923	0,749	3
Weighted avg.	0,982	0,136	0,982	0,982	0,982	0,923	0,974	

Πίνακας 22. Μετρικές απόδοσης 8ου πειράματος του SVM.

Cross-validation			
TARGET / OUTPUT	Ίαση	Θάνατος	SUM
Ίαση	19144 92.59%	195 0.94%	19339 98.99% 1.01%
Θάνατος	181 0.88%	1155 5.59%	1336 86.45% 13.55%
SUM	19325 99.06% 0.94%	1350 85.56% 14.44%	20299 / 20675 98.18% 1.82%

Εικόνα 34: Πίνακας σύγχυσης 8ου πειράματος του SVM.

Ως "target" αναφέρεται η πραγματική κλάση, ενώ ως "output" η κλάση πρόβλεψης. [Δημιουργήθηκε με τη χρήση του Confusion Matrix Generator (https://www.damianoperri.it/public/confusionMatrix/).]

Μέρος Β

Κατόπιν, έγινε εφαρμογή του SVM στο σύνολο δεδομένων με τη χρήση της τεχνικής SMOTE και του Information Gain Attribute Evaluation. Τα αποτελέσματα των διαφορετικών παραμέτρων σε κάθε συνάρτηση πυρήνα είχαν μικρές αποκλίσεις μεταξύ τους. Έτσι, η καλύτερη απόδοση των μοντέλων επικεντρώθηκε κυρίως στην ικανότητα τους να προβλέπουν σωστά την κλάση που αφορά την έκβαση του θανάτου.

9ο πείραμα

Το 9ο πείραμα αφορά την εφαρμογή του **SVM** στο αρχείο με όνομα «data set 7.arff», δηλαδή στο σετ δεδομένων με την χρήση της τεχνικής υπερδειγματοληψίας **SMOTE** και την επιλογή 20 χαρακτηριστικών βάση των αποτελεσμάτων του **Information Gain Attribute Evaluation**, σε σύνολο 20.675 δείγματα. Ως συνάρτηση πυρήνα επιλέχθηκε η **γραμμική**. Το μοντέλο εκπαιδεύτηκε και αξιολογήθηκε χρησιμοποιώντας την τεχνική της διασταυρούμενης επικύρωσης με k=10 (10-fold cross-validation).

Πραγματοποιήθηκαν δοκιμές με διάφορες τιμές για την παράμετρο κόστους C. Τα μοντέλα με C=50 και C=20 εμφάνισαν ακριβώς την ίδια απόδοση. Επιλέχθηκε ως πιο αποδοτικό το μοντέλο με C=20, καθώς χρειάστηκε λιγότερο χρόνο για να κατασκευαστεί. Ο χρόνος που χρειάστηκε για να κατασκευαστεί το μοντέλο ήταν 58,84 δευτερόλεπτα. Το ποσοστό των σωστά ταξινομημένων δειγμάτων στο σύνολο δεδομένων ήταν 97.6929% (20.198 δείγματα), ενώ το ποσοστό των λάθος ταξινομημένων δειγμάτων ήταν 2.3071% (477 δείγματα). Ο συντελεστής κάπα (Kappa statistic) ήταν ίσος με 0.8108. Το μέσο απόλυτο σφάλμα (Mean absolute error) ήταν ίσο με 0.8108, ενώ το σχετικό απόλυτο σφάλμα (Relative absolute error) ήταν ίσο με 18.8945%

Στον Πίνακα 23, παρουσιάζονται τα αποτελέσματα από τις μετρικές απόδοσης του 9ου πειράματος για τις κλάσεις ξεχωριστά, καθώς και ο σταθμισμένος μέσος όρος (weighted average).

	TPR	FPR	Precision	Recall	F-measure	ROC Area	PRC Area	Κλάση
	0,988	0,178	0,988	0,988	0,988	0,905	0,987	1
	0,822	0,012	0,824	0,822	0,823	0,905	0,689	3
Weighted avg.	0,977	0,167	0,977	0,977	0,977	0,905	0,967	

Πίνακας 23. Μετρικές απόδοσης 9ου πειράματος του SVM.

Cross-validation			
TARGET / OUTPUT	Ίαση	Θάνατος	SUM
Ίαση	19088 92.32%	237 1.15%	19325 98.77% 1.23%
Θάνατος	240 1.16%	1110 5.37%	1350 82.22% 17.78%
SUM	19328 98.76% 1.24%	1347 82.41% 17.59%	20198 / 20675 97.69% 2.31%

Εικόνα 35: Πίνακας σύγχυσης 9ου πειράματος του SVM.

Ως "target" αναφέρεται η πραγματική κλάση, ενώ ως "output" η κλάση πρόβλεψης. [Δημιουργήθηκε με τη χρήση του Confusion Matrix Generator (https://www.damianoperri.it/public/confusionMatrix/).]

10ο πείραμα

Το 10ο πείραμα διαφοροποιείται από το 9ο, ως προς τη συνάρτηση πυρήνα που επιλέχθηκε, δηλαδή την **πολυωνυμική**. Το μοντέλο εκπαιδεύτηκε και αξιολογήθηκε χρησιμοποιώντας την τεχνική της διασταυρούμενης επικύρωσης με k=10 (10-fold cross-validation).

Πραγματοποιήθηκαν δοκιμές με διάφορες τιμές για τις παραμέτρους C, coef0 και degree. Καλύτερη απόδοση εμφάνισε το μοντέλο με C=10, coef0=10 και degree= 3. Ο χρόνος που χρειάστηκε για να κατασκευαστεί το μοντέλο ήταν 87,76 δευτερόλεπτα. Το ποσοστό των σωστά ταξινομημένων δειγμάτων στο σύνολο δεδομένων ήταν 98.0411% (20.270δείγματα), ενώ το ποσοστό των λάθος ταξινομημένων δειγμάτων ήταν 1.9589% (405 δείγματα). Ο συντελεστής κάπα (Kappa statistic) ήταν ίσος με 0.8392. Το μέσο απόλυτο σφάλμα (Mean absolute error) ήταν ίσο με 0.0196, ενώ το σχετικό απόλυτο σφάλμα (Relative absolute error) ήταν ίσο με 16.0425%.

Στον Πίνακα 24, παρουσιάζονται τα αποτελέσματα από τις μετρικές απόδοσης του 10ου πειράματος για τις κλάσεις ξεχωριστά, καθώς και ο σταθμισμένος μέσος όρος (weighted average).

	TPR	FPR	Precision	Recall	F-measure	ROC Area	PRC Area	Κλάση
	0,990	0,152	0,989	0,990	0,990	0,919	0,989	1
	0,848	0,010	0,851	0,848	0,850	0,919	0,732	3
Weighted avg.	0,980	0,143	0,980	0,980	0,980	0,919	0,972	

Πίνακας 24. Μετρικές απόδοσης 10ου πειράματος του SVM.

Cross-validation			
TARGET / OUTPUT	Ίαση	Θάνατος	SUM
Ίαση	19125 92.50%	205 0.99%	19330 98.94% 1.06%
Θάνατος	200 0.97%	1145 5.54%	1345 85.13% 14.87%
SUM	19325 98.97% 1.03%	1350 84.81% 15.19%	20270 / 20675 98.04% 1.96%

Εικόνα 36: Πίνακας σύγχυσης 10ου πειράματος του SVM.

Ως "target" αναφέρεται η πραγματική κλάση, ενώ ως "output" η κλάση πρόβλεψης. [Δημιουργήθηκε με τη χρήση του Confusion Matrix Generator (https://www.damianoperri.it/public/confusionMatrix/).]

11ο πείραμα

Το 11ο πείραμα διαφοροποιείται από το 9ο ως προς τη συνάρτηση πυρήνα που επιλέχθηκε, δηλαδή την **σιγμοειδή**. Το μοντέλο εκπαιδεύτηκε και αξιολογήθηκε χρησιμοποιώντας την τεχνική της διασταυρούμενης επικύρωσης με k=10 (10-fold cross-validation).

Πραγματοποιήθηκαν δοκιμές με διάφορες τιμές για τις παραμέτρους C, coef0 και gamma. Καλύτερη απόδοση εμφάνισε το μοντέλο με C=10, coef0=1 και gamma= 0. Ο χρόνος που χρειάστηκε για να κατασκευαστεί το μοντέλο ήταν 11,71 δευτερόλεπτα. Το ποσοστό των σωστά ταξινομημένων δειγμάτων στο σύνολο δεδομένων ήταν 97.7025% (20.184 δείγματα), ενώ το ποσοστό των λάθος ταξινομημένων δειγμάτων ήταν 2.2975% (491 δείγματα). Ο συντελεστής κάπα (Kappa statistic) ήταν ίσος με 0.8082. Το μέσο απόλυτο

σφάλμα (Mean absolute error) ήταν ίσο με 0.0237, ενώ το σχετικό απόλυτο σφάλμα (Relative absolute error) ήταν ίσο με 19.4491%.

Στον Πίνακα 25, παρουσιάζονται τα αποτελέσματα από τις μετρικές απόδοσης του 11ου πειράματος για τις κλάσεις ξεχωριστά, καθώς και ο σταθμισμένος μέσος όρος (weighted average).

	TPR	FPR	Precision	Recall	F-measure	ROC Area	PRC Area	Κλάση
	0,986	0,167	0,988	0,986	0,987	0,910	0,988	1
	0,833	0,014	0,809	0,833	0,821	0,910	0,685	3
Weighted avg.	0,976	0,157	0,977	0,976	0,976	0,910	0,968	

Πίνακας 25. Μετρικές απόδοσης 11ου πειράματος του SVM.

Cross-validation			
TARGET / OUTPUT	Ίαση	Θάνατος	SUM
Ίαση	19059 92.18%	225 1.09%	19284 98.83% 1.17%
Θάνατος	266 1.29%	1125 5.44%	1391 80.88% 19.12%
SUM	19325 98.62% 1.38%	1350 83.33% 16.67%	20184 / 20675 97.63% 2.37%

Εικόνα 37: Πίνακας σύγχυσης 11ου πειράματος του SVM.

Ως "target" αναφέρεται η πραγματική κλάση, ενώ ως "output" η κλάση πρόβλεψης. [Δημιουργήθηκε με τη χρήση του Confusion Matrix Generator (https://www.damianoperri.it/public/confusionMatrix/).]

12ο πείραμα

Το 12ο πείραμα διαφοροποιείται από το 9ο ως προς τη συνάρτηση πυρήνα που επιλέχθηκε, δηλαδή την **RBF**. Το μοντέλο εκπαιδεύτηκε και αξιολογήθηκε χρησιμοποιώντας την τεχνική της διασταυρούμενης επικύρωσης με k=10 (10-fold cross-validation).

Πραγματοποιήθηκαν δοκιμές με διάφορες τιμές για τις παραμέτρους C και gamma. Καλύτερη απόδοση εμφάνισε το μοντέλο με C=50 και gamma= 0. Ο χρόνος που χρειάστηκε για να κατασκευαστεί το μοντέλο ήταν 5,14 δευτερόλεπτα.

Το ποσοστό των σωστά ταξινομημένων δειγμάτων στο σύνολο δεδομένων ήταν 97.9347% (20.248 δείγματα), ενώ το ποσοστό των λάθος ταξινομημένων δειγμάτων ήταν 2.0653% (427 δείγματα). Ο συντελεστής κάπα (Kappa statistic) ήταν ίσος με 0.8303. Το μέσο απόλυτο σφάλμα (Mean absolute error) ήταν ίσο με 0.0207, ενώ το σχετικό απόλυτο σφάλμα (Relative absolute error) ήταν ίσο με 16.914%.

Στον Πίνακα 26, παρουσιάζονται τα αποτελέσματα από τις μετρικές απόδοσης του 12ου πειράματος για τις κλάσεις ξεχωριστά, καθώς και ο σταθμισμένος μέσος όρος (weighted average).

	TPR	FPR	Precision	Recall	F-measure	ROC Area	PRC Area	Κλάση
	0,989	0,161	0,989	0,989	0,989	0,914	0,988	1
	0,839	0,011	0,844	0,839	0,841	0,914	0,718	3
Weighted avg.	0,979	0,152	0,979	0,979	0,979	0,914	0,971	

Πίνακας 26. Μετρικές απόδοσης 12ου πειράματος του SVM.

Cross-validation			
TARGET / OUTPUT	Ίαση	Θάνατος	SUM
Ίαση	19116 92.46%	218 1.05%	19334 98.87% 1.13%
Θάνατος	209 1.01%	1132 5.48%	1341 84.41% 15.59%
SUM	19325 98.92% 1.08%	1350 83.85% 16.15%	20248 / 20675 97.93% 2.07%

Εικόνα 38: Πίνακας σύγχυσης 12ου πειράματος του SVM.

Ως "target" αναφέρεται η πραγματική κλάση, ενώ ως "output" η κλάση πρόβλεψης. [Δημιουργήθηκε με τη χρήση του Confusion Matrix Generator (https://www.damianoperri.it/public/confusionMatrix/).]

6.3 Σύγκριση μοντέλων RF και SVM

Στην παρούσα εργασία, η επιλογή του βέλτιστου μοντέλου μηχανικής μάθησης όσον αφορά την πρόβλεψη της έκβασης της θεραπευτικής αγωγής της φυματίωσης επικεντρώνεται κυρίως στην κλάση που αφορά τον θάνατο των ασθενών. Αυτό οφείλεται στο γεγονός ότι η παρούσα μελέτη στοχεύει κυρίως στην πρόληψη της αποτυχίας της θεραπείας και κατ' επέκταση του θανάτου. Η ψευδώς θετική πρόβλεψη της ίασης ενός ασθενούς έχει ως αποτέλεσμα σοβαρές συνέπειες για τον περιορισμό της νόσου. Στόχος είναι η επιλογή ενός μοντέλου μηχανικής μάθησης το οποίο θα προβλέπει σωστά τα περιστατικά των θανάτων και δεν θα τα εντάσσει ψευδώς στην κατηγορία της ίασης. Συνεπώς, ένα αποδοτικό μοντέλο θα έχει υψηλό True Positive Rate ή αλλιώς Recall για την κλάση του θανάτου και ιδανικά θα έχει χαμηλή τιμή για το FPR της κλάσης αυτής.

6.3.1 Σύγκριση μοντέλων RF

Η πρώτη σύγκριση αφορά την επιλογή του πιο αποδοτικού μοντέλου του **RF** πριν και μετά τη χρήση της τεχνικής υπερδειγματοληψίας **SMOTE**, δηλαδή τα πειράματα 1 έως 4 του RF.

- Από τα πειράματα 1 και 2 που αφορούν την εφαρμογή του RF χωρίς τη χρήση της τεχνικής υπερδειγματοληψίας SMOTE, επιλέχθηκε ως αποδοτικότερο το 1ο με τις 100 επαναλήψεις (numIterations=100) συνολικά και για τις δυο κλάσεις σε σχέση με το 2ο με τις 200 επαναλήψεις (numIterations=200). Όσον αφορά την κλάση της ίασης, το 1ο πείραμα παρουσίασε υψηλότερη απόδοση από το 2ο καθώς είχε χαμηλότερο FPR (0,338 και 0,348 αντίστοιχα) και υψηλότερο F-Measure (0,992 και 0,991 αντίστοιχα). Οι μετρικές TPR ή Recall και Precision είχαν ίδιες τιμές για την κλάση της ίασης και στα δυο πειράματα. Όσον αφορά την κλάση του θανάτου, το 1ο πείραμα παρουσίασε υψηλότερη απόδοση από το 2ο καθώς είχε υψηλότερο TPR ή Recall (0,662 και 0,652 αντίστοιχα), υψηλότερο Precision (0,820 και 0,812 αντίστοιχα) και υψηλότερο F-Measure (0,733 και 0,723 αντίστοιχα), ενώ το FPR ήταν ίδιο. Συνολικά, το μοντέλο του 1ου πειράματος κρίθηκε αποδοτικότερο και για τις δυο κλάσεις.

- Από τα πειράματα 3 και 4 που αφορούν την εφαρμογή του RF με τη χρήση της τεχνικής υπερδειγματοληψίας SMOTE, επιλέχθηκε ως αποδοτικότερο το 3ο με τις 100 επαναλήψεις (numIterations=100) συνολικά και για τις δυο κλάσεις σε σχέση με το 4ο με τις 200 επαναλήψεις (numIterations=200). Όσον αφορά την κλάση της ίασης, το 3ο πείραμα παρουσίασε υψηλότερη απόδοση από το 4ο καθώς είχε

χαμηλότερο FPR (0,167 και 0,169 αντίστοιχα). Οι μετρικές TPR ή Recall Precision και F-Measure είχαν ίδιες τιμές για την κλάση της ίασης και στα δυο πειράματα. Όσον αφορά την κλάση του θανάτου, το 3ο πείραμα παρουσίασε υψηλότερη απόδοση από το 4ο καθώς είχε υψηλότερο TPR ή Recall (0,833 και 0,831 αντίστοιχα), υψηλότερο Precision (0,881 και 0,880 αντίστοιχα) και υψηλότερο F-Measure (0,856 και 0,855 αντίστοιχα), ενώ το FPR ήταν ίδιο. Συνολικά, το μοντέλο του 3ου πειράματος κρίθηκε αποδοτικότερο και για τις δυο κλάσεις.

- Εν συνεχεία, συγκρίνοντας το 1ο και το 3ο πείραμα, δηλαδή τα αποδοτικότερα χωρίς αλλά και με τη χρήση της τεχνικής SMOTE αντίστοιχα, διαπιστώθηκε ότι το μοντέλο του 3ου πειράματος παρουσίασε αρκετά υψηλότερη απόδοση από αυτό του 1ου όσον αφορά την σωστή ταξινόμηση των δειγμάτων της κλάσης του θανάτου που είναι ο πρωταρχικός στόχος αυτής της εργασίας. Συγκεκριμένα, το μοντέλο του 3ου πειράματος σε σχέση με το 1ο, για την κλάση του θανάτου, είχε υψηλότερο TPR ή Recall (0,833 και 0,662 αντίστοιχα), υψηλότερο Precision (0,881 και 0,820 αντίστοιχα) και υψηλότερο F-Measure (0,856 και 0,733 αντίστοιχα). Ωστόσο, το FPR ήταν υψηλότερο στο μοντέλο του 3ου πειράματος σε σχέση με αυτό του 1ου (0,008 και 0,005 αντίστοιχα), γεγονός που υποδηλώνει την ταξινόμηση δειγμάτων στην κλάση του θανάτου λανθασμένα. Όσον αφορά την κλάση της ίασης, το μοντέλο του 3ου πειράματος σε σχέση με του 1ου είχε χαμηλότερο FPR (0,167 και 0,338 αντίστοιχα), ενώ η τιμή του Precision ήταν ίδια. Ωστόσο, το μοντέλο του 3ου πειράματος σε σχέση με του 1ου είχε χαμηλότερο TPR ή Recall (0,992 και 0,995 αντίστοιχα) και χαμηλότερο F-Measure (0,990 και 0,992 αντίστοιχα).

  Συμπερασματικά, οι παραπάνω τιμές για τις μετρικές απόδοσης των δυο μοντέλων, υποδεικνύουν **υψηλότερη απόδοση** του μοντέλου του **3ου πειράματος** όσον αφορά την κλάση του θανάτου καθώς ταξινομούνται σωστά πολύ περισσότερα δείγματα σε αυτή. Συνεπώς, η τεχνική υπερδειγματοληψίας **SMOTE** που εφαρμόστηκε στο 3ο πείραμα ήταν ιδιαίτερα **αποτελεσματική** καθώς αντιμετωπίστηκε το φαινόμενο της ανισορροπίας των κλάσεων και έτσι η κλάση του θανάτου εκπροσωπήθηκε περισσότερο κατά την εκπαίδευση του μοντέλου. Ωστόσο, αυξήθηκε ελάχιστα το ποσοστό των δειγμάτων που ταξινομήθηκαν στην κλάση του θανάτου ψευδώς (FPR 0,008) σε σχέση με το μοντέλου του 1ου πειράματος (FPR 0,005), γεγονός που δεν επηρεάζει ιδιαίτερα την αποδοτικότητα του μοντέλου συνολικά.

Η δεύτερη σύγκριση αφορά την επιλογή του πιο αποδοτικού μοντέλου του **RF** με την χρήση της τεχνικής **SMOTE,** με το πιο αποδοτικό μετά τη χρήση του **Information Gain Attribute Evaluation** δηλαδή τα πειράματα 3,5,6 του RF.

- Από τα πειράματα 5 και 6 που αφορούν την εφαρμογή του RF με τη χρήση της τεχνικής υπερδειγματοληψίας SMOTE και την εφαρμογή του Information Gain Attribute Evaluation, επιλέχθηκε ως αποδοτικότερο το 5ο με τις 100 επαναλήψεις (numIterations=100) συνολικά και για τις δυο κλάσεις σε σχέση με το 6ο με τις 200 επαναλήψεις (numIterations=200). Όσον αφορά την κλάση της ίασης, το 5ο πείραμα παρουσίασε υψηλότερη απόδοση από το 6ο καθώς είχε χαμηλότερο FPR (0,160 και 0,162 αντίστοιχα). Οι μετρικές TPR ή Recall, Precision και F-Measure είχαν ίδιες τιμές για την κλάση της ίασης και στα δυο πειράματα. Όσον αφορά την κλάση του θανάτου, το 5ο πείραμα παρουσίασε υψηλότερη απόδοση από το 6ο καθώς είχε υψηλότερο TPR ή Recall (0,840 και 0,838 αντίστοιχα), υψηλότερο Precision (0,856 και 0,855 αντίστοιχα) και υψηλότερο F-Measure (0,848 και 0,846 αντίστοιχα), ενώ το FPR ήταν ίδιο. Συνολικά, το μοντέλο του 5ου πειράματος κρίθηκε αποδοτικότερο και για τις δυο κλάσεις.

- Έπειτα, πραγματοποιήθηκε σύγκριση για το 3ο (τεχνική SMOTE) και το 5ο πείραμα (τεχνική SMOTE και Information Gain Attribute Evaluation), δηλαδή τα πειράματα που κρίθηκαν αποδοτικότερα για τον RF. Για την κλάση της ίασης, το μοντέλο από το 3ο πείραμα παρουσιάζει υψηλότερο TPR ή Recall έναντι του 5ου (0,992 και 0,990 αντίστοιχα) και υψηλότερο F-Measure (0,990 και 0,989 αντίστοιχα). Ωστόσο, το FPR είναι υψηλότερο στο 3ο πείραμα σε σχέση με το 5ο (0,167 και 0,160 αντίστοιχα), υποδηλώνοντας περισσότερα δείγματα ψευδώς ταξινομημένα στην κλάση της ίασης. Επίσης, το Precision είναι χαμηλότερο στο 3ο πείραμα σε σχέση με το 5ο (0,988 και 0,989 αντίστοιχα). Όσον αφορά την κλάση του θανάτου, το 3ο πείραμα παρουσιάζει χαμηλότερο TPR ή Recall έναντι του 5ου (0,833 και 0,840 αντίστοιχα). Το FPR του 3ου πειράματος είναι χαμηλότερο από του 5ου (0,008 και 0,010 αντίστοιχα). Επίσης, το 3ο πείραμα παρουσιάζει υψηλότερο Precision έναντι του 5ου (0,881 και 0,856 αντίστοιχα) και υψηλότερο F-Measure (0,856 και 0,848 αντίστοιχα).

  Συμπερασματικά, οι παραπάνω τιμές για τις μετρικές απόδοσης των δυο μοντέλων, υποδεικνύουν **υψηλότερη απόδοση** του μοντέλου του **5ου πειράματος** όσον αφορά την **κλάση του θανάτου** στην οποία επικεντρώνεται η παρούσα εργασία, καθώς ταξινομούνται σωστά πολύ περισσότερα δείγματα σε αυτή και όχι ψευδώς στην κλάση της ίασης. Έτσι, η εφαρμογή του **Information Gain Attribute Evaluation** στο 5ο πείραμα κρίνεται αποτελεσματική καθώς το TPR ή Recall αυξήθηκε, που σημαίνει ότι το μοντέλο έχει την ικανότητά να αναγνωρίζει τα

δείγματα της κλάσης του θανάτου και να μην τα κατατάσσει λανθασμένα στην κλάση της ίασης. Ωστόσο, στο 5ο πείραμα αυξήθηκε ελάχιστα το ποσοστό των δειγμάτων που ταξινομήθηκαν στην κλάση του θανάτου ψευδώς (FPR 0,010) σε σχέση με το μοντέλου του 3ου πειράματος (FPR 0,008), γεγονός που δεν επηρεάζει ιδιαίτερα την αποδοτικότητα του μοντέλου συνολικά. Επίσης, το μοντέλο του 5ου πειράματος έχει σχετικά χαμηλότερο Precision και F-Measure, αλλά δεν επηρεάζεται σημαντικά η συνολική αποδοτικότητα του μοντέλου.

Έτσι, από τα πειράματα που πραγματοποιήθηκαν για τον **RF**, το πιο αποδοτικό ήταν το μοντέλο του **5ου πειράματος** (**SMOTE**, **Information Gain Attribute Evaluation**, **100 επαναλήψεις**) σύμφωνα με το στόχο της παρούσας εργασίας, παρόλο που οι μετρικές απόδοσης για την κλάση της ίασης ευνοούσαν την επιλογή του μοντέλου του 3ου πειράματος. Η επιλογή αυτή οφείλεται στο γεγονός ότι η εργασία αυτή επικεντρώνεται περισσότερο στην υψηλότερη απόδοση της κλάσης του θανάτου, όταν δεν είναι εφικτή η συνολική μέγιστη απόδοση του μοντέλου. Το μοντέλο του 5ου πειράματος είχε το υψηλότερο TPR ή Recall για την κλάση του θανάτου υποδεικνύοντας ότι έχει την ικανότητα να αναγνωρίζει τα δείγματα της κλάσης του θανάτου και να μην τα κατατάσσει λανθασμένα στην κλάση της ίασης.

6.3.2 Σύγκριση μοντέλων SVM

Η πρώτη σύγκριση αφορά την επιλογή του πιο αποδοτικού μοντέλου του **SVM** πριν και μετά τη χρήση της τεχνικής υπερδειγματοληψίας **SMOTE**, δηλαδή τα πειράματα 1 έως 8 του SVM.

- Από τα πειράματα 1 έως 4 που αφορούν την εφαρμογή του SVM χωρίς τη χρήση της τεχνικής υπερδειγματοληψίας SMOTE, όσον αφορά την κλάση της ίασης επιλέχθηκαν ως αποδοτικότερα τα μοντέλα των πειραμάτων 2 και 4, δηλαδή της πολυωνυμικής και της RBF συνάρτησης πυρήνα αντίστοιχα. Τα μοντέλα αυτά, με βάση τις μετρικές απόδοσης που παρουσιάζονται στον Πίνακα 27 ήταν πολύ κοντά στην απόδοση τους. Συγκεκριμένα, τα δυο μοντέλα είχαν ίδια τιμή στις μετρικές Precision και F-Measure. Επιπλέον, το μοντέλο του 2ου πειράματος (πολυωνυμική συνάρτηση πυρήνα) είχε ελάχιστα χαμηλότερο FPR από το μοντέλο του 4ου πειράματος (0,225 και 0,227 αντίστοιχα). Επίσης, παρόμοια τιμή παρουσίασαν τα δύο μοντέλα και στην μετρική TPR ή Recall, συγκεκριμένα 0,991 το 2ο και 0,992 το 4ο. Όσον αφορά την κλάση του θανάτου, επιλέχθηκαν πάλι ως αποδοτικότερα τα μοντέλα των πειραμάτων 2 και 4, δηλαδή της

πολυωνυμικής και της RBF συνάρτησης πυρήνα αντίστοιχα. Τα μοντέλα αυτά, με βάση τις μετρικές απόδοσης που παρουσιάζονται στον Πίνακα 28, είχαν παρόμοια απόδοση. Το μοντέλο του 2ου πειράματος είχε ελάχιστα υψηλότερο TPR ή Recall σε σχέση με αυτό του 4ου (0,775 και 0,773 αντίστοιχα), αλλά είχε και ελάχιστα υψηλότερο FPR (0,009 και 0,008 αντίστοιχα). Επιπλέον, το μοντέλο του 2ου πειράματος είχε χαμηλότερο Precision από αυτό του 4ου (0,756 και 0,768 αντίστοιχα) και χαμηλότερο F-Measure (0,765 και 0,770 αντίστοιχα). Συμπερασματικά, η πολυωνυμική συνάρτηση πυρήνα και η RBF εμφάνισαν παρόμοια και υψηλότερη απόδοση στα μοντέλα που δημιουργήθηκαν χωρίς τη χρήση της τεχνικής SMOTE.

Κλάση ίασης	Γραμμική	Πολυωνυμική	Σιγμοειδής	RBF
TPR/Recall	0,992	0,991	0,992	0,992
FPR	0,250	**0,225**	0,236	**0,227**
Precision	0,991	0,992	0,992	0,992
F-Measure	0,991	0,992	0,992	0,992

Πίνακας 27. Μετρικές απόδοσης πειραμάτων 1-4 του SVM για την κλάση της ίασης.

Κλάση θανάτου	Γραμμική	Πολυωνυμική	Σιγμοειδής	RBF
TPR/Recall	0,750	**0,775**	0,764	**0,773**
FPR	0,008	0,009	0,008	0,008
Precision	0,760	0,756	0,761	0,768
F-Measure	0,755	0,765	0,763	0,770

Πίνακας 28. Μετρικές απόδοσης πειραμάτων 1-4 του SVM για την κλάση του θανάτου.

- Από τα πειράματα 5 έως 8 που αφορούν την εφαρμογή του SVM με τη χρήση της τεχνικής υπερδειγματοληψίας SMOTE, όσον αφορά την κλάση της ίασης επιλέχθηκε ως αποδοτικότερο το μοντέλο του 6ου πειράματος, δηλαδή της πολυωνυμικής συνάρτησης πυρήνα με μικρή διαφορά από το μοντέλο του 8ου πειράματος , δηλαδή της συνάρτησης πυρήνα RBF, σύμφωνα με τις Μετρικές απόδοσης του Πίνακα 29. Συγκριτικά, τα δυο μοντέλο είχαν ίδιες τιμές για τις μετρικές TPR ή Recall, Precision και F-Measure. Το μοντέλο του 6ου πειράματος είχε ελάχιστα χαμηλότερο FPR σε σύγκριση με του 8ου (0,142 και 0,144 αντίστοιχα). Για την κλάση του θανάτου, επιλέχθηκε επίσης ως αποδοτικότερο το μοντέλο του 6ου πειράματος με μικρή διαφορά από το μοντέλο του 8ου πειράματος, σύμφωνα με τις μετρικές απόδοσης του Πίνακα 30. Πιο συγκεκριμένα, τα δυο μοντέλα παρουσίασαν ίδια τιμή στην μετρική FPR (0,009). Ωστόσο, το 6ο πείραμα είχε ελάχιστα υψηλότερο TPR ή Recall από το 8ο (0,858 και 0,856 αντίστοιχα),

ελάχιστα υψηλότερο Precision (0,867 και 0,865 αντίστοιχα), καθώς και ελάχιστα υψηλότερο F-Measure (0,862 και 0,860 αντίστοιχα). Συμπερασματικά, η πολυωνυμική συνάρτηση πυρήνα και η RBF είχαν παρόμοια απόδοση με τη χρήση της τεχνικής SMOTE, με ένα μικρό προβάδισμα της πολυωνυμικής και για τις δυο κλάσεις.

Κλάση ίασης	Γραμμική	Πολυωνυμική	Σιγμοειδής	RBF
TPR/Recall	0,989	0,991	0,987	0,991
FPR	0,181	**0,142**	0,170	0,144
Precision	0,987	0,990	0,988	0,990
F-Measure	0,988	0,990	0,988	0,990

Πίνακας 29. Μετρικές απόδοσης πειραμάτων 5-8 του SVM για την κλάση της ίασης.

Κλάση θανάτου	Γραμμική	Πολυωνυμική	Σιγμοειδής	RBF
TPR/Recall	0,819	**0,858**	0,830	0,856
FPR	0,011	0,009	0,013	0,009
Precision	0,842	**0,867**	0,821	0,865
F-Measure	0,830	**0,862**	0,825	0,860

Πίνακας 30. Μετρικές απόδοσης πειραμάτων 5-8 του SVM για την κλάση του θανάτου.

Συμπερασματικά, οι παραπάνω τιμές για τις μετρικές απόδοσης των μοντέλων που δημιουργήθηκαν στα πειράματα 1 έως 8, υποδεικνύουν **υψηλότερη απόδοση** της **πολυωνυμικής συνάρτησης πυρήνα** με μικρή διαφορά από την RBF πριν αλλά και μετά την χρήση της τεχνικής SMOTE. Επιπρόσθετα, στα μοντέλα 5 έως 8 όπου εφαρμόστηκε η τεχνική SMOTE, η απόδοση των μοντέλων παρουσίασε σαφή βελτίωση στην κλάση των θανούντων σύμφωνα με τον Πίνακα 28 και τον Πίνακα 30, καθώς ταξινομούνται σωστά πολύ περισσότερα δείγματα σε αυτή σε σχέση με τα μοντέλα των πειραμάτων 1 έως 4. Συνεπώς, η τεχνική υπερδειγματοληψίας **SMOTE** που εφαρμόστηκε στα πειράματα 5 έως 8 ήταν **ιδιαίτερα αποτελεσματική** καθώς αντιμετωπίστηκε το φαινόμενο της ανισορροπίας των κλάσεων και έτσι η κλάση του θανάτου εκπροσωπήθηκε περισσότερο κατά την εκπαίδευση του μοντέλου. Όμως, αυξήθηκε ελάχιστα το ποσοστό των δειγμάτων που ταξινομήθηκαν στην κλάση του θανάτου ψευδώς, δηλαδή η μετρική FPR, γεγονός που δεν επηρεάζει ιδιαίτερα την αποδοτικότητα του μοντέλου συνολικά.

Η δεύτερη σύγκριση αφορά την επιλογή του πιο αποδοτικού μοντέλου του **SVM** με την χρήση της τεχνικής **SMOTE,** με το πιο αποδοτικό μετά τη χρήση του **Information Gain Attribute Evaluation** δηλαδή τα πειράματα του 6,9,10,11 και 12 του SVM.

- Από τα πειράματα 9 έως 12 που αφορούν την εφαρμογή του SVM με τη χρήση της τεχνικής υπερδειγματοληψίας SMOTE και την εφαρμογή του Information Gain Attribute Evaluation, επιλέχθηκε ως αποδοτικότερο για την κλάση της ίασης το μοντέλο του 10ου πειράματος, δηλαδή της πολυωνυμικής συνάρτησης πυρήνα σύμφωνα με τις μετρικές απόδοσης του Πίνακα 31. Το μοντέλο αυτό παρουσίασε υψηλότερο TPR ή Recall (0,990), Precision (0,989), F-Measure (0,990) σε σχέση με τα υπόλοιπα. Επίσης, είχε το χαμηλότερο FPR (0,152) συγκριτικά με τα υπόλοιπα μοντέλα. Για την κλάση του θανάτου, ήταν επίσης το πιο αποδοτικό σύμφωνα με τις μετρικές απόδοσης του Πίνακα 32. Το μοντέλο του 10ου πειράματος παρουσίασε υψηλότερο TPR ή Recall (0,848), Precision (0,851), F-Measure (0,850) σε σχέση με τα υπόλοιπα, ενώ είχε το χαμηλότερο FPR (0,010). Συνεπώς, το μοντέλο του 10ου πειράματος με την πολυωνυμική συνάρτηση πυρήνα ήταν το πιο αποδοτικό από τα υπόλοιπα και για τις δυο κλάσεις για όλες τις μετρικές απόδοσης.

Κλάση ίασης	Γραμμική	Πολυωνυμική	Σιγμοειδής	RBF
TPR/Recall	0,988	**0,990**	0,986	0,989
FPR	0,178	**0,152**	0,167	0,161
Precision	0,988	**0,989**	0,988	0,989
F-Measure	0,988	**0,990**	0,987	0,989

Πίνακας 31. Μετρικές απόδοσης πειραμάτων 9-12 του SVM για την κλάση της ίασης.

Κλάση θανάτου	Γραμμική	Πολυωνυμική	Σιγμοειδής	RBF
TPR/Recall	0,822	**0,848**	0,833	0,839
FPR	0,012	**0,010**	0,014	0,011
Precision	0,824	**0,851**	0,809	0,844
F-Measure	0,823	**0,850**	0,821	0,841

Πίνακας 32. Μετρικές απόδοσης πειραμάτων 9-12 του SVM για την κλάση του θανάτου.

- Έπειτα, πραγματοποιήθηκε σύγκριση για το 6ο (τεχνική SMOTE με πολυωνυμική συνάρτηση πυρήνα) και το 10ο πείραμα (τεχνική SMOTE και Information Gain Attribute Evaluation με πολυωνυμική συνάρτηση πυρήνα), δηλαδή τα πειράματα που κρίθηκαν αποδοτικότερα για τον SVM. Για την κλάση της ίασης, το μοντέλο από το 6ο ήταν αποδοτικότερο σε σχέση με αυτό από το 10ο σύμφωνα με τις μετρικές απόδοσης του Πίνακα 29 και του Πίνακα 31. Συγκεκριμένα, το 6ο πείραμα είχε υψηλότερο TPR ή Recall σε σχέση με το 10ο (0,991 και 0,990 αντίστοιχα) και υψηλότερο Precision (0,990 και 0,989 αντίστοιχα). Η μετρική F-Measure είχε την ίδια τιμή και στα δυο πειράματα, ενώ το FPR ήταν χαμηλότερο στο 6ο πείραμα συγκριτικά με το 10ο (0,142 και 0,152 αντίστοιχα). Όσον αφορά την κλάση του θανάτου, επίσης το μοντέλο από το 6ο ήταν αποδοτικότερο σε σχέση με αυτό από το 10ο σύμφωνα με τις μετρικές απόδοσης του Πίνακα 30 και του Πίνακα 32. Αναλυτικότερα, το 6ο πείραμα είχε υψηλότερο TPR ή Recall σε σχέση με το 10ο (0,858 και 0,848 αντίστοιχα), υψηλότερο Precision (0,867 και 0,851 αντίστοιχα) και υψηλότερο F-Measure (0,862 και 0,850 αντίστοιχα). Επιπλέον, το FPR ήταν χαμηλότερο στο 6ο πείραμα συγκριτικά με το 10ο (0,009 και 0,010 αντίστοιχα).

  Συμπερασματικά, οι παραπάνω τιμές για τις μετρικές απόδοσης των δυο μοντέλων, υποδεικνύουν **υψηλότερη απόδοση** του μοντέλου του **6ου πειράματος** και για τις δυο κλάσεις. Συνεπώς, η εφαρμογή του **Information Gain Attribute Evaluation** στο 10ο πείραμα κρίνεται **μη αποτελεσματική** καθώς το TPR ή Recall μειώθηκε και στις δυο κλάσεις, που σημαίνει ότι μειώθηκε η ικανότητά του μοντέλου να αναγνωρίζει τα δείγματα της κάθε κλάσης σωστά και να μην τα κατατάσσει λανθασμένα στην άλλη κλάση. Αντίστοιχα, το FPR αυξήθηκε και στις δυο κλάσεις του 10ου πειράματος σε σχέση με το 6ο. Επίσης, το μοντέλο του 6ου πειράματος έχει σχετικά χαμηλότερο Precision και F-Measure, αλλά δεν επηρεάζεται σημαντικά η συνολική αποδοτικότητα του μοντέλου.

Έτσι, από τα πειράματα που πραγματοποιήθηκαν για τον **SVM**, το πιο αποδοτικό ήταν το μοντέλο του **6ου πειράματος** (**SMOTE**, **πολυωνυμική**). Το μοντέλο αυτό είχε το υψηλότερο TPR ή Recall και για τις δυο κλάσεις υποδεικνύοντας ότι έχει την ικανότητα να αναγνωρίζει τα δείγματα κάθε κλάσης καλύτερα από τα υπόλοιπα μοντέλα και να μην τα κατατάσσει λανθασμένα.

6.3.3 Σύγκριση αποδοτικότερου μοντέλου RF και SVM

Από τα 6 πειράματα που πραγματοποιήθηκαν για τον **RF**, το πιο αποδοτικό εστιάζοντας κυρίως στην απόδοση της κλάσης του θανάτου σύμφωνα με το στόχο της παρούσας εργασίας ήταν το μοντέλο του **5ου πειράματος**, δηλαδή αυτό με τη χρήση της τεχνικής **SMOTE** και του **Information Gain Attribute Evaluation** με **100 επαναλήψεις** (παράμετρος numIterations), όπως αναλύθηκε στο κεφάλαιο 5.3.1.

Επιπλέον, από τα 12 πειράματα που διενεργήθηκαν για τον **SVM**, το πιο αποδοτικό και για τις δυο κλάσεις ήταν το μοντέλο του **6ου πειράματος,** δηλαδή αυτό με τη χρήση της τεχνικής **SMOTE** και την επιλογή της **πολυωνυμικής συνάρτησης πυρήνα** χωρίς την εφαρμογή του Information Gain Attribute Evaluation όπως αναλύθηκε στο κεφάλαιο 5.3.2.

Για την κλάση της **ίασης**, σύμφωνα με τον Πίνακα 33, αποδοτικότερο είναι το μοντέλο του **SVM** καθώς εμφανίζει υψηλότερο TPR ή Recall σε σχέση με το μοντέλο του RF (0,991 και 0,990 αντίστοιχα), υψηλότερο Precision (0,990 και 0,989 αντίστοιχα) και υψηλότερο F-Measure (0,990 και 0,989 αντίστοιχα), ενώ εμφάνισε χαμηλότερο FPR (0,142 και 0,160 αντίστοιχα).

Κλάση ίασης	TPR/Recall	FPR	Precision	F-Measure
RF	0,990	0,160	0,989	0,989
SVM	**0,991**	**0,142**	**0,990**	**0,990**

Πίνακας 33. Μετρικές απόδοσης των πιο αποδοτικών πειραμάτων του RF και SVM για την κλάση της ίασης.

Παράλληλα, σύμφωνα με τον Πίνακα 34, για την κλάση του **θανάτου** αποδοτικότερο είναι επίσης το μοντέλο του **SVM** καθώς εμφανίζει υψηλότερο TPR ή Recall σε σχέση με το μοντέλο του RF (0,858και 0,840 αντίστοιχα), υψηλότερο Precision (0,867 και 0,856 αντίστοιχα) και υψηλότερο F-Measure (0,862 και 0,848 αντίστοιχα), ενώ εμφάνισε χαμηλότερο FPR (0,009 και 0,010 αντίστοιχα).

Κλάση θανάτου	TPR/Recall	FPR	Precision	F-Measure
RF	0,840	0,010	0,856	0,848
SVM	**0,858**	**0,009**	**0,867**	**0,862**

Πίνακας 34. Μετρικές απόδοσης των πιο αποδοτικών πειραμάτων του RF και SVM για την κλάση του θανάτου.

Εν κατακλείδι, **το πιο αποδοτικό μοντέλο** από το σύνολο των μοντέλων που δημιουργήθηκαν και για τους δυο αλγόριθμους σύμφωνα με τον στόχο της παρούσας εργασίας ήταν το μοντέλο του 6^{ου} πειράματος του **SVM**, δηλαδή αυτό με τη χρήση της τεχνικής **SMOTE** και την επιλογή της **πολυωνυμικής** συνάρτησης πυρήνα χωρίς την εφαρμογή του Information Gain Attribute Evaluation.

7. Αξιολόγηση αποδοτικότερου μοντέλου σε νέα δεδομένα

Το τελευταίο στάδιο που εφαρμόστηκε στην παρούσα εργασία στη διαδικασία της μηχανικής μάθησης αφορά την διενέργεια του Supplied test set. Στο Supplied test set φορτώθηκαν νέα ανεξάρτητα δεδομένα από ένα αρχείο με στόχο την αξιολόγηση της απόδοσης του μοντέλου που επιλέχθηκε ως αποδοτικότερο στο κεφάλαιο 6.3.3 και δημιουργήθηκε μετά την εκπαίδευση του με Cross-validation. Στόχος ήταν να αποτιμηθεί η ικανότητα γενίκευσης του μοντέλου σε νέα δεδομένα που δεν είχαν χρησιμοποιηθεί κατά την εκπαίδευση.

Συγκεκριμένα, το αρχείο περιείχε 5.000 νέες εγγραφές από το αρχικό σύνολο δεδομένων των 964.099. Στις 5.000 νέες εγγραφές ακολουθήθηκαν τα ίδια βήματα προεπεξεργασίας όπως και στο 6ο πείραμα του SVM που επιλέχθηκε ως το πιο αποδοτικό σύμφωνα με το στόχο της εργασίας, εκτός την εφαρμογή της τεχνικής SMOTE, καθώς η τεχνική αυτή ήταν απαραίτητη μόνο για την εκπαίδευση του μοντέλου. Τα αποτελέσματα του Supplied test set παρουσιάζονται στον Πίνακα 35.

Ο χρόνος που χρειάστηκε για να ελεγχθεί το μοντέλο ήταν 0,66 δευτερόλεπτα. Το ποσοστό των σωστά ταξινομημένων δειγμάτων στο σύνολο δεδομένων ήταν 99.4% (4970 δείγματα), ενώ το ποσοστό των λάθος ταξινομημένων δειγμάτων ήταν 0.6% (30 δείγματα). Ο συντελεστής κάπα (Kappa statistic) ήταν ίσος με 0.8593. Το μέσο απόλυτο σφάλμα (Mean absolute error) ήταν ίσο με 0.006, ενώ το σχετικό απόλυτο σφάλμα (Relative absolute error) ήταν ίσο με 12.968%

	TPR	FPR	Precision	Recall	F-measure	ROC Area	PRC Area	Κλάση
	0,999	0,203	0,995	0,999	0,997	0,898	0,995	1
	0,797	0,001	0,940	0,797	0,862	0,898	0,754	3
Weighted avg.	0,994	0,199	0,994	0,994	0,994	0,898	0,989	

Πίνακας 35. Μετρικές απόδοσης Supplied test set.

Supplied test set			
TARGET / OUTPUT	Ίαση	Θάνατος	SUM
Ίαση	4876 97.52%	24 0.48%	4900 99.51% 0.49%
Θάνατος	6 0.12%	94 1.88%	100 94.00% 6.00%
SUM	4882 99.88% 0.12%	118 79.66% 20.34%	4970 / 5000 99.40% 0.60%

Εικόνα 39: Πίνακας σύγχυσης Supplied test set.

Ως "target" αναφέρεται η πραγματική κλάση, ενώ ως "output" η κλάση πρόβλεψης. [Δημιουργήθηκε με τη χρήση του Confusion Matrix Generator (https://www.damianoperri.it/public/confusionMatrix/).]

Όσον αφορά την κλάση της ίασης (Πίνακας 36), το Supplied test set παρουσίασε υψηλότερη απόδοση σε σχέση με το Cross-validation, καθώς είχε υψηλότερο TPR ή Recall (0,999 και 0,991 αντίστοιχα), υψηλότερο Precision (0,995 και 0,990 αντίστοιχα) και υψηλότερο F-Measure (0,997 και 0,990 αντίστοιχα). Ωστόσο, το FPR ήταν υψηλότερο (0,203 και 0,142 αντίστοιχα).

Κλάση ίασης	TPR/Recall	FPR	Precision	F-Measure
Cross-validation	0,991	0,142	0,990	0,990
Supplied test set	0,999	0,203	0,995	0,997

Πίνακας 36. Μετρικές απόδοσης Cross-validation και Supplied test set για την κλάση της ίασης.

Όσον αφορά την κλάση του θανάτου (Πίνακας 37), το Supplied test set σε σχέση με το Cross-validation είχε μειωμένη απόδοση καθώς είχε χαμηλότερο TPR ή Recall (0,797 και 0,858 αντίστοιχα). Ωστόσο, το Supplied test set σε σχέση με το Cross-validation παρουσίασε χαμηλότερο FPR (0,001 και 0,009 αντίστοιχα) και υψηλότερο Precision (0,940 και 0,867 αντίστοιχα), ενώ το F-Measure ήταν ίδιο.

Κλάση θανάτου	TPR/Recall	FPR	Precision	F-Measure
Cross-validation	0,858	0,009	0,867	0,862
Supplied test set	0,797	0,001	0,940	0,862

Πίνακας 37. Μετρικές απόδοσης Cross-validation και Supplied test set για την κλάση του θανάτου.

Στο Supplied test set συγκριτικά με το Cross-validation, μειώθηκε η ικανότητα του μοντέλου να ταξινομεί σωστά τα δείγματα της κλάσης του θανάτου, δηλαδή περισσότερα δείγματα ταξινομήθηκαν λανθασμένα στην κλάση της ίασης ενώ ανήκουν στην κλάση του θανάτου. Αντίθετα, αυξήθηκε η ικανότητα του μοντέλου να ταξινομεί σωστά τα δείγματα στην κλάση της ίασης. Το γεγονός αυτό πιθανά οφείλεται στην τεχνική SMOTE, η οποία χρησιμοποιήθηκε στο Cross-validation καθώς πραγματοποιήθηκε εκπαίδευση και αξιολόγηση του μοντέλου, αλλά όχι στο Supplied test set όπου πραγματοποιήθηκε μόνο η αξιολόγηση του μοντέλου. Το μοντέλο έχει εκπαιδευτεί με δεδομένα που έχουν υποστεί υπερδειγματοληψία με την τεχνική SMOTE, δηλαδή έχουν επαυξηθεί οι παρατηρήσεις της κλάσης που εκπροσωπείται λιγότερο. Η απουσία της υπερδειγματοληψίας, στην κλάση του θανάτου που αποτελεί μειοψηφία, στο Supplied test set μπορεί να οδήγησε σε χαμηλότερη απόδοση του μοντέλου. Ωστόσο, η εφαρμογή της τεχνικής SMOTE στο Supplied test set δεν είναι συνήθης πρακτική, καθώς αυτό χρησιμοποιείται κυρίως για την αξιολόγηση της ικανότητας γενίκευσης του μοντέλου. Με την εφαρμογή της τεχνικής SMOTE στο Supplied test set, δεν θα μπορούσε να αντικατοπτριστεί η κατανομή των δεδομένων σε πραγματικές συνθήκες. Επιπλέον, επειδή τα δεδομένα στο Supplied test set ήταν τυχαία επιλεγμένα θα μπορούσαν να παρουσιαστούν κάποιες αποκλίσεις στην απόδοση του μοντέλου. Εν κατακλείδι, η απόδοση του μοντέλου παρέμεινε ιδιαίτερα ικανοποιητική για την κλάση του θανάτου και πολλά υποσχόμενη για την κλάση της ίασης.

8. Συζήτηση

Στην παρούσα εργασία διερευνήθηκε η δυνατότητα πρόβλεψης της έκβασης της θεραπείας της φυματίωσης με τη δημιουργία μοντέλων μηχανικής μάθησης με στόχο την αποτροπή της αποτυχίας της θεραπείας των ασθενών. Συγκεκριμένα, χρησιμοποιήθηκαν οι αλγόριθμοι **RF** και **SVM** σε ένα σύνολο με δεδομένα τα οποία συλλέχθηκαν από το Βραζιλιάνικο Σύστημα Πληροφόρησης για Μεταδιδόμενα Νοσήματα (SINAN). Από τα 37 χαρακτηριστικά (μεταβλητές), που αφορούσαν κοινωνικά και δημογραφικά χαρακτηριστικά των ασθενών, κλινικές πληροφορίες και εργαστηριακά δεδομένα που αφορούν την φυματίωση, διατηρήθηκαν τα 36 κατά το στάδιο της προεπεξεργασίας. Επίσης, έγινε διακριτοποίηση σε δυο χαρακτηριστικά, δηλαδή μετατράπηκαν από αριθμητικές σε κατηγορικές μεταβλητές.

Η εφαρμογή των αλγορίθμων πραγματοποιήθηκε σε 3 διαφορετικά σύνολα δεδομένων που δημιουργήθηκαν σε διάφορα στάδια της προεπεξεργασίας. Αρχικά, οι αλγόριθμοι εφαρμόστηκαν στο σετ δεδομένων με τα 36 χαρακτηριστικά και έπειτα από την διακριτοποίηση. Κατόπιν, οι αλγόριθμοι εφαρμόστηκαν στο νέο σύνολο δεδομένων που προέκυψε από τη χρήση της τεχνικής **SMOTE** για την αντιμετώπιση της ανισορροπίας των κλάσεων, με στόχο την καλύτερη εκπροσώπηση της κλάσης του θανάτου. Τέλος, εφαρμόστηκαν στο σύνολο δεδομένων που προέκυψε από την εφαρμογή του **Information Gain Attribute Evaluation** για την εύρεση των βέλτιστων χαρακτηριστικών και τη χρήση της τεχνικής SMOTE, το οποίο αποτελούνταν από 20 χαρακτηριστικά. Για την εκπαίδευση και την αξιολόγηση των αλγορίθμων εφαρμόστηκε η τεχνική K-fold Cross-Validation με K = 10.

Συνολικά, περιγράφηκαν τα **20 πειράματα** που υλοποιήθηκαν για τα 3 σύνολα δεδομένων και παρουσιάστηκαν οι μετρικές απόδοσης τους. Έπειτα, πραγματοποιήθηκε η σύγκριση τους και η αξιολόγηση της απόδοσης τους βάση των μετρικών απόδοσης **TPR** ή **Recall**, **FPR**, **Precision** και **F-Measure** για την **κλάση της ίασης** και του **θανάτου** ξεχωριστά. Η επιλογή του πιο αποδοτικού μοντέλου σύμφωνα με το στόχο της εργασίας επικεντρώνεται περισσότερο στην υψηλότερη απόδοση της κλάσης του θανάτου, όταν δεν είναι εφικτή η συνολική μέγιστη απόδοση του μοντέλου. Έτσι, μέγιστη σημασία δόθηκε στο υψηλότερο TPR ή Recall για την κλάση του θανάτου και στο χαμηλότερο FPR για την κλάση της ίασης , καθώς πρωταρχικός στόχος ήταν η δημιουργία ενός μοντέλου που έχει την ικανότητα να αναγνωρίζει τα δείγματα της κλάσης του θανάτου και να μην τα κατατάσσει λανθασμένα στην κλάση της ίασης.

Για τον αλγόριθμο **SVM,** η **πολυωνυμική** συνάρτηση πυρήνα καθώς και η **RBF**, εμφάνισαν **υψηλότερη απόδοσ**η για τον στόχο της εργασίας σε σχέση με την γραμμική και σιγμοειδή σε όλα τα πειράματα που υλοποιήθηκαν .

Η χρήση της τεχνικής **SMOTE** για την αντιμετώπιση της ανισορροπίας των κλάσεων ήταν ιδιαίτερα αποτελεσματική και για τους δυο αλγόριθμους, καθώς

αυξήθηκε το TPR ή Recall στην κλάση του θανάτου, συνεπώς ο αλγόριθμος κατάφερε να ανιχνεύσει περισσότερες περιπτώσεις θανάτου και να τις ταξινομήσει σωστά στην κλάση αυτή. Επιπρόσθετα, αυξημένες ήταν και οι μετρικές Precision και F-Measure. Ωστόσο, αυξήθηκε ελάχιστα η μετρική FPR, δηλαδή αυξήθηκαν τα δείγματα που ταξινομήθηκαν στην κλάση του θανάτου ψευδώς, γεγονός που δεν επηρεάζει ιδιαίτερα την αποδοτικότητα του μοντέλου συνολικά. Συμπερασματικά, η τεχνική **SMOTE** ήταν περισσότερο **ωφέλιμη** παρά επιζήμια για τον στόχο της εργασίας.

Όσον αφορά την εφαρμογή του **Information Gain Attribute Evaluation** για την εύρεση των βέλτιστων χαρακτηριστικών, για τον αλγόριθμο **RF** κρίνεται **αποτελεσματική** για την **κλάση του θανάτου** καθώς το TPR ή Recall αυξήθηκε, που σημαίνει ότι το μοντέλο έχει την ικανότητά να αναγνωρίζει τα δείγματα της κλάσης του θανάτου και να μην τα κατατάσσει λανθασμένα στην κλάση της ίασης. Ωστόσο, αυξήθηκε ελάχιστα το ποσοστό των δειγμάτων που ταξινομήθηκαν στην κλάση του θανάτου ψευδώς δηλαδή το FPR, γεγονός που δεν επηρεάζει ιδιαίτερα την αποδοτικότητα του μοντέλου συνολικά. Επίσης, οι μετρικές Precision και F-Measure ήταν ελαφρώς χαμηλότερες, αλλά δεν επηρεάζεται σημαντικά η συνολική αποδοτικότητα του μοντέλου σχετικά με το στόχο της εργασίας. Για τον αλγόριθμο **SVM,** η εφαρμογή του **Information Gain Attribute Evaluation** κρίνεται **μη αποτελεσματική** καθώς το TPR ή Recall μειώθηκε και για τις δυο κλάσεις, που σημαίνει ότι μειώθηκε η ικανότητά του μοντέλου να αναγνωρίζει τα δείγματα της κάθε κλάσης σωστά και να μην τα κατατάσσει λανθασμένα στην άλλη κλάση. Αντίστοιχα, το FPR αυξήθηκε και στις δυο κλάσεις.

Ως **πιο αποδοτικό μοντέλο** από το σύνολο των μοντέλων που δημιουργήθηκαν και για τους δυο αλγόριθμους σύμφωνα με τον στόχο της παρούσας εργασίας επιλέχθηκε το μοντέλο του **SVM** με τη χρήση της τεχνικής **SMOTE** και την επιλογή της **πολυωνυμικής** συνάρτησης πυρήνα χωρίς την εφαρμογή του Information Gain Attribute Evaluation. Δημιουργήθηκε ταχύτατα καθώς ο **χρόνος κατασκευής** του μοντέλου ήταν μόλις **23.79 δευτερόλεπτα**. Το μοντέλο αυτό ταξινόμησε σωστά το 98.21% των δειγμάτων. Για την **κλάση του θανάτου** στην οποία επικεντρώνεται η εργασία το μοντέλο σημείωσε **TPR** ή **Recall** 0,858, **FPR** 0,009, **Precision** 0,867 και **F-Measure** 0,862. Για την **κλάση της ίασης,** το μοντέλο σημείωσε **TPR** ή **Recall** 0,991, **FPR** 0,142, **Precision** 0,990 και **F-Measure** 0,990. Ακόμη, ο σταθμισμένος μέσος όρος (**weighted average**), ο οποίος λαμβάνει υπόψη τον αριθμό των δειγμάτων σε κάθε κλάση, εμφάνισε **TPR** ή **Recall** 0,982, **FPR** 0,134, **Precision** 0,982 και **F-Measure** 0,982. Συνολικά, τα αποτελέσματα φαίνονται ιδιαίτερα ενθαρρυντικά.

Τέλος, για την περαιτέρω αξιολόγηση της απόδοσης του μοντέλου του SVM που επιλέχθηκε ως αποδοτικότερο, διενεργήθηκε **Supplied test set** με **νέα ανεξάρτητα δεδομένα** με 5.000 εγγραφές (από το αρχικό σετ δεδομένων) ώστε να αποτιμηθεί η **ικανότητα γενίκευσης** του σε νέα δεδομένα που δεν είχαν χρησιμοποιηθεί κατά την εκπαίδευση. Για την **κλάση του θανάτου** στην οποία επικεντρώνεται η εργασία το μοντέλο σημείωσε **TPR** ή **Recall** 0,797,

FPR 0,001, **Precision** 0,940 και **F-Measure** 0,862. Για την **κλάση της ίασης,** το μοντέλο σημείωσε **TPR** ή **Recall** 0,999, **FPR** 0,203, **Precision** 0,990 και **F-Measure** 0,990. Ακόμη, ο σταθμισμένος μέσος όρος (**weighted average**), ο οποίος λαμβάνει υπόψη τον αριθμό των δειγμάτων σε κάθε κλάση, εμφάνισε **TPR** ή **Recall** 0,982, **FPR** 0,134, **Precision** 0,995 και **F-Measure** 0,997. Η μειωμένη απόδοση στο Supplied test set για την κλάση του θανάτου πιθανά οφείλεται στην απουσία υπερδειγματοληψίας με την τεχνική SMOTE που στόχο είχε την προσομοίωση της κατανομής των δεδομένων σε πραγματικές συνθήκες, ενώ ταυτόχρονα πιθανώς για τον ίδιο λόγο αυξήθηκε η απόδοση στην κλάση της ίασης.

Η τεχνική SMOTE που χρησιμοποιήθηκε στην παρούσα εργασία φάνηκε αποτελεσματική και σε παρόμοια μελέτη που είχε ως στόχο την πρόβλεψη της έκβασης της θεραπείας της φυματίωσης. Συγκεκριμένα, στην μελέτη του Wei Lian Willian Foh και των συνεργατών του(111), χρησιμοποιήθηκε το ίδιο σετ δεδομένων με την παρούσα εργασία αλλά διαφορετικός αλγόριθμος, δηλαδή ο Naïve Bayes. Το μοντέλο του Naïve Bayes μπόρεσε να ταξινομήσει σωστά περισσότερες περιπτώσεις ασθενών που απεβίωσαν, προκαλώντας μια μικρή αύξηση των εσφαλμένα ταξινομημένων περιπτώσεων ίασης. Τα αποτελέσματα αυτά έρχονται σε απόλυτη συμφωνία με τα αποτελέσματα στην παρούσα εργασία.

Στη μελέτη των Keethansana Kanesamoorthy και Maheshi B. Dissanayake (113), η οποία αφορά την πρόβλεψη της αποτυχίας της θεραπείας των ασθενών με φυματίωση εστιάζοντας στα χαρακτηριστικά που συμβάλλουν στην αντοχή στα αντιβιοτικά που χρησιμοποιούνται κατά την θεραπεία, χρησιμοποιήθηκε επίσης ο αλγόριθμος SVM όπως και στην παρούσα μελέτη. Κατά τη σύγκριση των τεσσάρων συναρτήσεων πυρήνα (γραμμική, πολυωνυμική, σιγμοειδής και RBF), η συνάρτηση πυρήνα RBF και η γραμμική συνάρτηση πυρήνα εμφάνισαν υψηλότερη ακρίβεια ταξινόμησης. Τα αποτελέσματα των Keethansana Kanesamoorthy και Maheshi B. Dissanayake έρχονται σε συμφωνία με αυτά της παρούσας εργασίας όσον αφορά την απόδοση της συνάρτησης RBF. Ωστόσο, η παρούσα εργασία έκρινε ιδιαίτερα αποτελεσματική και την πολυωνυμική συνάρτηση πυρήνα.

Επιπρόσθετα, οι αλγόριθμοι RF και SVM που επιλέχθηκαν στην παρούσα εργασία, χρησιμοποιήθηκαν και στην μελέτη των Owais A. Hussain και Khurum N. Junejo(115), η οποία είχε ως στόχο την πρόβλεψη της έκβασης της θεραπείας των ασθενών κατά την έναρξη της θεραπείας για την διασφάλιση της εύρυθμης λειτουργείας των προγραμμάτων διαχείρισης ασθενών με φυματίωση. Όσον αφορά τον SVM, στην μελέτη των Owais A. Hussain και Khurum N. Junejo, εμφάνισε υψηλές τιμές για τη μετρική Precision και χαμηλές για τη μετρική Sensitivity για την κλάση της αποτυχημένης θεραπείας. Συνεπώς, τα περισσότερα από τα δείγματα που ταξινόμησε ο SVM στην κλάση της αποτυχημένης θεραπείας, ανήκαν όντως στην κλάση αυτή, ωστόσο ταξινόμησε αρκετά δείγματα λανθασμένα στην κλάση της επιτυχημένης θεραπείας ενώ ανήκαν σε αυτή της αποτυχημένης θεραπείας. Επιπλέον, τα αποτελέσματα των ερευνητών έδειξαν ότι ο RF

παρουσίασε συνολικά καλύτερη απόδοση από τον SVM, καθώς εμφάνισε υψηλή απόδοση στην κλάση της αποτυχημένης θεραπείας, δηλαδή υψηλές τιμές για τις μετρικές Precision και Sensitivity, όπως επίσης και στην κλάση της επιτυχημένης θεραπείας, δηλαδή υψηλές τιμές για την μετρική Specificity. Τα αποτελέσματα των Owais A. Hussain και Khurum N. Junejo έρχονται σε αντίθεση με αυτά της παρούσας εργασίας, καθώς ο SVM αποδείχθηκε πιο αποτελεσματικός από τον RF στο να ταξινομεί σωστά τα περισσότερα από τα δείγματα που ανήκαν στην κλάση του θανάτου και όχι να τα ταξινομεί λανθασμένα στην κλάση της ίασης.

9. Συμπεράσματα

Συνοπτικά, στην παρούσα εργασία διερευνήθηκε η δυνατότητα πρόβλεψης της έκβασης της θεραπείας της φυματίωσης δημιουργώντας μοντέλα μηχανικής μάθησης, δηλαδή χρησιμοποιώντας τους αλγόριθμους **RF** και **SVM** σε σύνολα δεδομένων που συλλέχθηκαν από το Βραζιλιάνικο Σύστημα Πληροφόρησης για Μεταδιδόμενα Νοσήματα (SINAN). Έγινε χρήση δύο διαφορετικών συνόλων χαρακτηριστικών, καθώς και τεχνικών προεπεξεργασίας ώστε να εκτιμηθεί η απόδοση των μοντέλων.

Για τον SVM, η πολυωνυμική συνάρτηση πυρήνα και η RBF, εμφάνισαν υψηλότερη απόδοση. Η χρήση της τεχνικής **SMOTE** για την αντιμετώπιση της ανισορροπίας των κλάσεων κρίθηκε ιδιαίτερα αποτελεσματική και για τους δυο αλγόριθμους, σε αντίθεση με την εφαρμογή του **Information Gain Attribute Evaluation** για την εύρεση των βέλτιστων χαρακτηριστικών που συνολικά δεν ωφέλησε στον στόχο της εργασίας.

Ως **πιο αποδοτικό μοντέλο** επιλέχθηκε το μοντέλο του **SVM** με τη χρήση της τεχνικής **SMOTE** και την επιλογή της **πολυωνυμικής** συνάρτησης πυρήνα χωρίς την εφαρμογή του Information Gain Attribute Evaluation. Το μοντέλο που δημιουργήθηκε στην παρούσα εργασία και επιλέχθηκε ως το πιο αποδοτικό θα μπορούσε να χρησιμοποιηθεί επικουρικά στις κλινικές αποφάσεις που λαμβάνονται από το προσωπικό υγείας για την εξατομίκευση της θεραπευτικής αγωγής των ασθενών με φυματίωση και την καλύτερη διαχείριση των διαθέσιμων πόρων, μέσω της πρόβλεψης πιθανών επιπλοκών που μπορεί να αντιμετωπίσει ένας ασθενής. Συγκεκριμένα, θα μπορούσε να συνδράμει στη λήψη προληπτικών μέτρων από το ιατρικό προσωπικό ή στην προσαρμογή της θεραπευτικής προσέγγισης εγκαίρως, βοηθώντας έτσι στη μείωση του κινδύνου επιπλοκών και στη βελτίωση των αποτελεσμάτων της θεραπείας.

Ωστόσο, η παρούσα εργασία εμφανίζει κάποιους περιορισμούς. Επειδή τα δεδομένα προέρχονται από τη Βραζιλία, η απόδοση των αλγόριθμων θα παρουσίαζε κάποια απόκλιση σε δεδομένα από άλλες χώρες εξαιτίας της ετερογένειας των δεδομένων λόγω των πολιτισμικών διαφορών, των περιβαλλοντικών συνθηκών καθώς και των συνθηκών περίθαλψης των ασθενών. Επιπλέον, το μοντέλο που επιλέχθηκε ως το πιο αποδοτικό δεν εφαρμόστηκε στην ρουτίνα κάποιου φορέα υγείας έτσι ώστε να αποδειχθεί η αρκετά υψηλή απόδοση του. Για την περαιτέρω αξιολόγηση της απόδοσης του, διενεργήθηκε **Supplied test set** με **νέα ανεξάρτητα δεδομένα** που είχε ως αποτέλεσμα τη μείωση της απόδοσης του μοντέλου στην κλάση του θανάτου αλλά την αύξηση της απόδοσης στην κλάση της ίασης. Η απόδοση του μοντέλου συνολικά για τον στόχο της εργασίας παρέμεινε ιδιαίτερα ικανοποιητική.

Ως μελλοντική προέκταση, θα μπορούσαν να εφαρμοστούν περισσότεροι αλγόριθμοι στο σύνολο δεδομένων και κατ’ επέκταση να χρησιμοποιηθούν

μέθοδοι ensemble, οι οποίες συνιστούν έναν αποτελεσματικό τρόπο συνδυασμού διαφορετικών μοντέλων μηχανικής μάθησης με σκοπό τη βελτίωση της απόδοσής τους. Επιπλέον, θα μπορούσε να εξεταστεί η συνεισφορά περισσότερων τεχνικών στην επιλογή χαρακτηριστικών, όπως για παράδειγμα ο Γενετικός Αλγόριθμος (Genetic Algorithm) και η μέθοδος Greedy Stepwise.

Βιβλιογραφία

1. Delogu G, Sali M, Fadda G. The biology of mycobacterium tuberculosis infection. Mediterr J Hematol Infect Dis. 2013;5(1).
2. Frith J. History of tuberculosis. Part 1 - Phthisis, consumption and the white plague. J Mil Veterans Health. 2014;22(2):29–35.
3. Barberis I, Bragazzi NL, Galluzzo L, Martini M. The history of tuberculosis: From the first historical records to the isolation of Koch's bacillus. J Prev Med Hyg. 2017;58:E9–12.
4. Gradmann C. Robert Koch and the pressures of scientific research: tuberculosis and tuberculin. Med Hist. 2001;45(1):1–32.
5. Harries AD. Robert Koch and the discovery of the tubercle bacillus: The challenge of HIV and tuberculosis 125 years later. Int J Tuberc Lung Dis. 2008;12(3):241–9.
6. Sable SB, Posey JE, Scriba TJ. Tuberculosis vaccine development: Progress in clinical evaluation. Clin Microbiol Rev. 2020;33(1):1–30.
7. Global tuberculosis report 2022. World Health Organization. 2022.
8. Global tuberculosis report 2023. 2023.
9. Rotim A, Singh R. Tuberculosis [Internet]. StatPearls Publishing LLC; 2023. Available from: https://www.ncbi.nlm.nih.gov/books/NBK441916/
10. Carabalí-Isajar ML, Rodríguez-Bejarano OH, Amado T, Patarroyo MA, Izquierdo MA, Lutz JR, et al. Clinical manifestations and immune response to tuberculosis. World J Microbiol Biotechnol. 2023;39(8):1–26.
11. Loddenkemper R, Lipman M, Zumla A. Clinical aspects of adult tuberculosis. Cold Spring Harb Perspect Med. 2016;6(1):1–25.
12. Achkar JM, Lawn SD, Moosa MYS, Wright CA, Kasprowicz VO. Adjunctive Tests for Diagnosis of Tuberculosis: Serology, ELISPOT for Site-Specific Lymphocytes, Urinary Lipoarabinomannan, String Test, and Fine Needle Aspiration. J Infect Dis. 2011;204.
13. Huang Y, Ai L, Wang X, Sun Z, Wang F. Review and Updates on the Diagnosis of Tuberculosis. J Clin Med. 2022;11(19).
14. Vilchèze C, Kremer L. Acid-fast positive and acid-fast negative Mycobacterium tuberculosis: The Koch paradox. Tuberc Tuber Bacillus Second Ed. 2017;519–32.
15. Reed JL, Basu D, Butzler MA, McFall SM. XtracTB Assay, a Mycobacterium tuberculosis molecular screening test with sensitivity approaching culture. Sci Rep. 2017;7(1):1–12.
16. Tortoli E, Cichero P, Piersimoni C, Simonetti MT, Gesu G, Nista D. Use of BACTEC MGIT 960 for recovery of mycobacteria from clinical

specimens: Multicenter study. J Clin Microbiol. 1999;37(11):3578–82.

17. Phetsuksiri B, Rudeeaneksin J, Srisungngam S, Bunchoo S, Klayut W, Nakajima C, et al. Comparison of loop-mediated isothermal amplification, microscopy, culture, and PCR for diagnosis of pulmonary tuberculosis. Jpn J Infect Dis. 2020;73(4):272–7.

18. World Health Organization. WHO Meeting Report of a Technical Expert Consultation: Non-inferiority analysis of Xpert MTB/RIF Ultra compared to Xpert MTB/RIF. 2017;1–11.

19. Maclean E, Kohli M, Weber SF, Suresh A. Advances in Molecular Diagnosis of Tuberculosis. J Clin Microbiol. 2020;(September):1–13.

20. World Health Organization (WHO). Consolidated Guidelines on Tuberculosis Treatment. Who. 2020. 99 p.

21. Horne DJ, Kohli M, Zifodya JS, Schiller I, Dendukuri N, Tollefson D, et al. Xpert MTB/RIF and Xpert MTB/RIF Ultra for pulmonary tuberculosis and rifampicin resistance in adults. Cochrane Database Syst Rev. 2019;(6).

22. Bomanji JB, Gulati P, Das CJ, Gupta N. Imaging in Tuberculosis. Cold Spring Harb Perspect Med. 2015;5(6).

23. Alshoabi SA, Almas KM, Aldofri SA, Hamid AM, Alhazmi FH, Alsharif WM, et al. The Diagnostic Deceiver: Radiological Pictorial Review of Tuberculosis. Diagnostics. 2022;12(2).

24. Alshoabi SA, Almas KM, Aldofri SA, Hamid AM, Alhazmi FH, Alsharif WM, et al. The Diagnostic Deceiver: Radiological Pictorial Review of Tuberculosis. Diagnostics. 2022;12(2):1–28.

25. World Health Organization (WHO). Latent tuberculosis infection: updated and consolidated guidelines for programmatic management. In 2018.

26. Carranza C, Pedraza-Sanchez S, de Oyarzabal-Mendez E, Torres M. Diagnosis for Latent Tuberculosis Infection: New Alternatives. Front Immunol. 2020;11:1–13.

27. Lewinsohn DM, Leonard MK, Lobue PA, Cohn DL, Daley CL, Desmond E, et al. Official American Thoracic Society/Infectious Diseases Society of America/Centers for Disease Control and Prevention Clinical Practice Guidelines: Diagnosis of Tuberculosis in Adults and Children. Clin Infect Dis. 2017;64(2).

28. Gualano G, Mencarini P, Lauria FN, Palmieri F, Mfinanga S, Mwaba P, et al. Tuberculin skin test – Outdated or still useful for Latent TB infection screening? Int J Infect Dis. 2019;80:S20–2.

29. Narasimhan P, Wood J, Macintyre CR, Mathai D. Risk Factors for Tuberculosis. Pulm Med. 2013;

30. Silva DR, Muñoz-torrico M, Duarte R, Galvão T, Bonini EH, Arbex FF. Risk factors for tuberculosis : diabetes , smoking , alcohol use , and the

use of other drugs. J Bras Pneumol. 2018;44(2):145–52.

31. Jeon CY, Murray MB. Diabetes Mellitus Increases the Risk of Active Tuberculosis : A Systematic Review of 13 Observational Studies. PLoS Med. 2008;5(7).

32. Alisjahbana B, Sahiratmadja E, Nelwan EJ, Purwa AM, Ahmad Y, Ottenhoff THM, et al. The Effect of Type 2 Diabetes Mellitus on the Presentation and Treatment Response of Pulmonary Tuberculosis. Clin Infect Dis. 2007;45(4):428–35.

33. World Health Organization (WHO). Collaborative Framework for Care and Control ofTuberculosis and Diabetes. 2011.

34. Dooley KE, Tang T, Golub JE, Cronin W. Impact of Diabetes Mellitus on Treatment Outcomes of Patients with Active Tuberculosis. Am J Trop Med Hyg. 2009;80(4):634–9.

35. Oursler KK, Moore RD, Bishai WR, Harrington SM, Pope DS, Chaisson RE. Survival of Patients with Pulmonary Tuberculosis : Clinical and Molecular Epidemiologic Factors. Clin Infect Dis. 2002;34(6):752–759.

36. Chaisson RE, Dooley KE. Tuberculosis and diabetes mellitus: convergence of two epidemics. Lancet Infect Dis. 2009;9(12):737–46.

37. Bates MN, Khalakdina A, Pai M, Chang L, Lessa F, Smith KR. Risk of Tuberculosis From Exposure to Tobacco Smoke: A Systematic Review and Meta-analysis. Arch Intern Med. 2007 Feb 26;167(4):335.

38. Burusie A, Enquesilassie F, Addissie A, Dessalegn B, Lamaro T. Effect of smoking on tuberculosis treatment outcomes: A systematic review and meta-analysis. Glantz SA, editor. PLoS One. 2020 Sep 17;15(9).

39. Masjedi MR, Hosseini M, Aryanpur M, Mortaz E, Tabarsi P, Soori H, et al. The effects of smoking on treatment outcome in patients newly diagnosed with pulmonary tuberculosis. Int J Tuberc Lung Dis. 2017;21:351–6.

40. Wang EY, Arrazola RA, Mathema B, Ahluwalia IB, Mase SR. The impact of smoking on tuberculosis treatment outcomes: a meta-analysis. Int J Tuberc Lung Dis. 2020 Feb 1;24(2):170–5.

41. Ragan EJ, Kleinman MB, Sweigart B, Gnatienko N, Parry CD, Horsburgh CR, et al. The impact of alcohol use on tuberculosis treatment outcomes: a systematic review and meta-analysis. Int J Tuberc Lung Dis. 2020 Jan 1;24(1):73–82.

42. Wigger GW, Bouton TC, Jacobson KR, Auld SC, Yeligar SM, Staitieh BS. The Impact of Alcohol Use Disorder on Tuberculosis: A Review of the Epidemiology and Potential Immunologic Mechanisms. Front Immunol. 2022 Mar 31;13:13.

43. Imtiaz S, Shield KD, Roerecke M, Samokhvalov A V., Lönnroth K, Rehm J. Alcohol consumption as a risk factor for tuberculosis: Meta-analyses and burden of disease. Eur Respir J. 2017;50(1).

44. World Health Organization (WHO). Global tuberculosis report 2020. 2020.

45. Qi CC, Xu LR, Zhao CJ, Zhang HY, Li QY, Liu MJ, et al. Prevalence and risk factors of tuberculosis among people living with HIV/AIDS in China: a systematic review and meta-analysis. BMC Infect Dis. 2023;23(1):1–13.

46. Duarte R, Lönnroth K, Carvalho C, Lima F, Carvalho ACC, Muñoz-Torrico M, et al. Tuberculosis, social determinants and co-morbidities (including HIV). Pulmonology. 2018;24(2):115–9.

47. Diedrich CR, Flynn JAL. HIV-1/Mycobacterium tuberculosis coinfection immunology: How does HIV-1 exacerbate tuberculosis? Infect Immun. 2011;79(4):1407–17.

48. Sanchez M, Bartholomay P, Arakaki-sanchez D, Enarson D, Bissell K. Outcomes of TB Treatment by HIV Status in National Recording Systems in Brazil , 2003 – 2008. PLoS One. 2012;7(3).

49. Karo B, Krause G, Hollo V, Werf MJ van der, Castell S, Hamouda O, et al. Impact of HIV infection on treatment outcome of tuberculosis in Europe. AIDS. 2016;30(7):1089–98.

50. Fekadu G, Turi E, Kasu T, Bekele F, Chelkeba L, Tolossa T, et al. Impact of HIV status and predictors of successful treatment outcomes among tuberculosis patients: A six-year retrospective cohort study. Ann Med Surg. 2020;60:531–41.

51. Shariq M, Sheikh JA, Quadir N, Sharma N, Hasnain SE, Ehtesham NZ. COVID-19 and tuberculosis: the double whammy of respiratory pathogens. Eur Respir Rev. 2022;31.

52. Dass SA, Balakrishnan V, Arifin N, Lim CSY, Nordin F, Tye GJ. The COVID-19/Tuberculosis Syndemic and Potential Antibody Therapy for TB Based on the Lessons Learnt From the Pandemic. Front Immunol. 2022;13.

53. Tassi G, Peres A, Fiegenbaum M. Pathology of TB/COVID-19 Co-Infection: The phantom menace. Tuberculosis. 2021;126.

54. Zamparelli SS, Mormile M, Zamparelli AS, Guarino A, Parrella R, Bocchino M. Clinical impact of COVID-19 on tuberculosis. Infez Med. 2022;30(4):495–500.

55. Sarkar S, Khanna P, Singh AK. Impact of COVID-19 in patients with concurrent co-infections: A systematic review and meta-analyses. J Med Virol. 2021;93:2385–95.

56. Hawn TR, Day TA, Scriba TJ, Hatherill M, Hanekom WA, Evans TG, et al. Tuberculosis Vaccines and Prevention of Infection. ASM Journals Microbiol Mol Biol Rev. 2014;78(4):650–71.

57. Qu M, Zhou X, Li H. BCG vaccination strategies against tuberculosis : updates and perspectives. Hum Vaccin Immunother. 2021;17(12):5284–95.

58. Yamazaki-nakashimada MA, Unzueta A, Gámez-gonzález LB, González-saldaña N. BCG: a vaccine with multiple faces. Hum Vaccin Immunother. 2020;16(8):1841–50.

59. Cho T, Khatchadourian C, Nguyen H, Dara Y, Jung S, Venketaraman V. A review of the BCG vaccine and other approaches toward tuberculosis eradication ABSTRACT. Hum Vaccin Immunother. 2021;17(8):2454–70.

60. Linh NN, Viney K, Gegia M, Falzon D, Glaziou P, Floyd K, et al. World Health Organization treatment outcome definitions for tuberculosis: 2021 update. Eur Respir J. 2021;58(2).

61. Alsayed SSR, Gunosewoyo H. Tuberculosis: Pathogenesis, Current Treatment Regimens and New Drug Targets. Int J Mol Sci. 2023;24(6).

62. Κατευθυντήριες οδηγίες για τη θεραπεία της φυματίωσης στους ενήλικες. 2015;28(3).

63. WHO consolidated guidelines on tuberculosis. WHO. 2020.

64. Liebenberg D, Gordhan BG. Drug resistant tuberculosis : Implications for transmission , diagnosis , and disease management. Front Cell Infect Microbiol. 2022;(September):1–18.

65. MeSH. No Title [Internet]. [cited 2023 Dec 20]. Available from: https://www.ncbi.nlm.nih.gov/mesh/?term=ai+artificial+intelligence

66. Tyagi N. 6 Major Branches of Artificial Intelligence (AI) [Internet]. Analytic Steps. Available from: https://www.analyticssteps.com/blogs/6-major-branches-artificial-intelligence-ai

67. Davenport T, Kalakota R. The Potential for Artificial Intelligence in Healthcare. Futur Healthc J. 2019;6(2):94–8.

68. Foote KD. A Brief History of Machine Learning [Internet]. Dataversity. 2021. Available from: https://www.dataversity.net/a-brief-history-of-machine-learning/#:~:text=Samuel also designed a number,“machine learning” in 1952

69. MeSH. No Title [Internet]. 2016. Available from: https://www.ncbi.nlm.nih.gov/mesh/2010029

70. Mitchell TM. Machine Learning [Internet]. Vol. 45, Machine Learning. McGraw-Hill; 2017. Available from: https://books.google.ca/books?id=EoYBngEACAAJ&dq=mitchell+machine+learning+1997&hl=en&sa=X&ved=0ahUKEwiomdqfj8TkAhWGslkKHRCbAtoQ6AEIKjAA

71. Farhat R, Mourali Y, Jemni M, Ezzedine H. An overview of Machine Learning Technologies and their use in E-learning. In: 2020 International Multi-Conference on: “Organization of Knowledge and Advanced Technologies” (OCTA) [Internet]. IEEE; 2020. p. 1–4. Available from: https://ieeexplore.ieee.org/document/9151758/

72. Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Comput Sci [Internet]. 2021;2(3):1–21.

Available from: https://doi.org/10.1007/s42979-021-00592-x

73. Διαμαντάρας Κ, Μπότσης Δ. Μηχανική μάθηση. Κλειδάριθμος; 2019. 792 p.

74. Mishra S, Mishra D, Santra GH. Applications of machine learning techniques in agricultural crop production: A review paper. Indian J Sci Technol. 2016;9(38).

75. Benos L, Tagarakis AC, Dolias G, Berruto R, Kateris D, Bochtis D. Machine learning in agriculture: A comprehensive updated review. Sensors. 2021;21(11):1–55.

76. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Inform. 2006;2:59–77.

77. Al'Aref SJ, Anchouche K, Singh G, Slomka PJ, Kolli KK, Kumar A, et al. Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. Eur Heart J. 2019;40(24):1975–86.

78. Zaidan AM. The leading global health challenges in the artificial intelligence era. Front Public Heal. 2023;11:1–10.

79. Huang X, Islam MR, Akter S, Ahmed F, Kazami E, Serhan HA, et al. Artificial intelligence in glaucoma: opportunities, challenges, and future directions. Vol. 22, BioMedical Engineering Online. BioMed Central; 2023. 1–48 p.

80. Brandao-de-Resende C, Melo M, Lee E, Jindal A, Neo YN, Sanghi P, et al. A machine learning system to optimise triage in an adult ophthalmic emergency department: a model development and validation study. eClinicalMedicine. 2023;66.

81. Qu K, Guo F, Liu X, Lin Y, Zou Q. Application of machine learning in microbiology. Front Microbiol. 2019;10(APR):1–10.

82. Kong PH, Chiang CH, Lin TC, Kuo SC, Li CF, Hsiung CA, et al. Discrimination of Methicillin-Resistant Staphylococcus aureus by MALDI-TOF Mass Spectrometry with Machine Learning Techniques in Patients with Staphylococcus aureus Bacteremia. Pathogens. 2022;11(5).

83. Feucherolles M, Nennig M, Becker SL, Martiny D, Losch S, Penny C, et al. Combination of MALDI-TOF Mass Spectrometry and Machine Learning for Rapid Antimicrobial Resistance Screening: The Case of Campylobacter spp. Front Microbiol. 2022;12(February):1–16.

84. Tahir F, Farhan M. Exploring the progress of artificial intelligence in managing type 2 diabetes mellitus: a comprehensive review of present innovations and anticipated challenges ahead. Front Clin Diabetes Healthc. 2023;4(December):1–7.

85. Corny J, Rajkumar A, Martin O, Dode X, Lajonchère JP, Billuart O, et al. A machine learning-based clinical decision support system to identify prescriptions with a high risk of medication error. J Am Med Informatics Assoc. 2020;27(11):1688–94.

86. Patel L, Shukla T, Huang X, Ussery DW, Wang S. Machine Learning Methods in Drug Discovery. Molecules. 2020 Nov 12;25(22).

87. Niazi SK, Mariam Z. Computer-Aided Drug Design and Drug Discovery: A Prospective Analysis. Pharmaceuticals. 2023;17(1):22.

88. Dara S, Dhamercherla S, Jadav SS, Babu CM, Ahsan MJ. Machine Learning in Drug Discovery: A Review. Vol. 55, Artificial Intelligence Review. Springer Netherlands; 2022. 1947–1999 p.

89. Kunduru AR. Machine Learning in Drug Discovery: A Comprehendive Analysis of Applications, Challenges, and Future Directions. Int J Orange Technol. 2023;5(8).

90. MeSH. Supervised Machine Learning [Internet]. 2016. Available from: https://www.ncbi.nlm.nih.gov/mesh/2010032

91. Γεωργούλη Κ. Τεχνητη Νοημοσυνη - Μια Εισαγωγική Προσέγγιση [Internet]. 2015. Available from: http://repfiles.kallipos.gr/html_books/93/00e-introduction.html

92. Chopra D, Khurana R. Introduction to Machine Learning with Python. Introduction to Machine Learning with Python. O'Reilly Media; 2023. 392 p.

93. Aggarwal CC. Data classification: Algorithms and applications. Data Classification: Algorithms and Applications. 2014. 1–675 p.

94. Kassambara A. Machine Learnging Essentials. 2018;

95. Crabtree M. What is Machine Learning? Definition, Types, Tools & More [Internet]. DataCamp. 2023. Available from: https://www.datacamp.com/blog/what-is-machine-learning

96. Seliya N, Khoshgoftaar TM, Van Hulse J. A study on the relationships of classifier performance metrics. Proc - Int Conf Tools with Artif Intell ICTAI. 2009;59–66.

97. Erickson BJ, Kitamura F. Magician's corner: 9. performance metrics for machine learning models. Radiol Artif Intell. 2021;3(3):1–7.

98. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. ACM Int Conf Proceeding Ser. 2006;148:233–40.

99. Lino Ferreira da Silva Barros MH, Alves GO, Morais Florêncio Souza L, da Silva Rocha E, Lorenzato de Oliveira JF, Lynn T, et al. Benchmarking machine learning models to assist in the prognosis of tuberculosis. Informatics. 2021;8(2):1–17.

100. Zhang E, Zhang Y. F-Measure. In: Encyclopedia of Database Systems [Internet]. Boston, MA: Springer US; 2009. p. 1147–1147. Available from: http://link.springer.com/10.1007/978-0-387-39940-9_483

101. Naser MZ, Alavi AH. Error Metrics and Performance Fitness Indicators for Artificial Intelligence and Machine Learning in Engineering and Sciences. Archit Struct Constr. 2023;3(4):499–517.

102. Hastie T, Tibshirani R, James G, Witten D. An Introduction to Statistical Learning. Vol. 102, Springer Texts. 2006. 618 p.

103. Ying X. An Overview of Overfitting and its Solutions. J Phys Conf Ser. 2019;1168(2).

104. Jabbar HK, Khan RZ. Methods to Avoid Over-Fitting and Under-Fitting in Supervised Machine Learning (Comparative Study). In: Computer Science, Communication and Instrumentation Devices. Singapore: Research Publishing Services; 2014. p. 163–72.

105. Korjus K, Hebart MN, Vicente R. An efficient data partitioning to improve classification performance while keeping parameters interpretable. PLoS One. 2016;11(8):1–16.

106. Joseph VR. Optimal ratio for data splitting. Stat Anal Data Min. 2022;15(4):531–8.

107. Biau G, Scornet E. A random forest guided tour. Test. 2016;25(2):197–227.

108. Rigatti SJ. Random Forest. J Insur Med [Internet]. 2017 Jan 1;47(1):31–9. Available from: https://meridian.allenpress.com/jim/article/47/1/31/131479/Random-Forest

109. Al-Mejibli IS, Alwan JK, Abd DH. The effect of gamma value on support vector machine performance with different kernels. Int J Electr Comput Eng. 2020;10(5):5497–506.

110. Sheng L, Na J. SVM parameters optimization algorithm and its application. Proc 2008 IEEE Int Conf Mechatronics Autom ICMA 2008. 2008;509–13.

111. Foh WLW, Ang SL, Lim CY, Alaga AAL, Yeap GH. Prediction of Tuberculosis Patients' Treatment Outcomes Using Multinomial Naive Bayes Algorithm and Class-Imbalanced Data. 2023 IEEE IAS Glob Conf Emerg Technol GlobConET 2023. 2023;1–6.

112. Lino Ferreira da Silva Barros MH, Santos GL, de Almeida Rodrigues MG, Sampaio V, Lynn T, Endo PT. A Brazilian classified data set for prognosis of tuberculosis, between January 2001 and April 2020. Sci Data. 2022;9(1):1–8.

113. Kanesamoorthy K, Dissanayake M. Prediction of treatment failure of tuberculosis using support vector machine with genetic algorithm. Int J Mycobacteriology. 2021;10(3):279–84.

114. Sauer CM, Sasson D, Paik KE, McCague N, Celi LA, Fernández IS, et al. Feature selection and prediction of treatment failure in tuberculosis. PLoS One. 2018;13(11):1–14.

115. Hussain OA, Junejo KN. Predicting treatment outcome of drug-susceptible tuberculosis patients using machine-learning models. Informatics Heal Soc Care. 2019;44(2):135–51.

116. Gheyas IA, Smith LS. Feature subset selection in large dimensionality domains. Pattern Recognit. 2010;43(1):5–13.

117. Barros MHLF da S, Santos G, Sampaio V, Lynn T, Endo PT. A Brazilian classified dataset for prognosis of tuberculosis. Mendeley Data. 2022.

118. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software: An Update. ACM SIGKDD Explor Newsl. 2009 Nov 16;11(1):10–8.

119. Dimov R. Weka : Practical Machine Learning Tools and Techniques with Java Implementations. 2007;

120. Witten IH, Frank E. Data Mining: Practical machine learning tools and techniques [Internet]. 2nd ed. Elsevier Inc. Morgan Kaufmann Publishers; 2005. Available from: http://books.google.com/books?hl=en&lr=&id=QTnOcZJzlUoC&oi=fnd&pg=PR17&dq=Data+Mining+Practical+Machine+Learning+Tools+and+Techniques&ots=3gpDdrWiOc&sig=TZS7G8l1eXSa2SpAvfD6aBoJ2lw

121. Khadija MA, Setiawan NA. Detecting Liver Disease Diagnosis by Combining SMOTE, Information Gain Attribute Evaluation and Ranker. ITSMART J Teknol dan Inf. 2020;9(1):13–7.

122. C. SK, R.J. RS. Application of Ranking Based Attribute Selection Filters To Perform Automated Evaluation of Descriptive Answers Through Sequential Minimal Optimization Models. ICTACT J Soft Comput. 2014;05(01):860–8.