ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ

ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ & ΥΠΗΡΕΣΙΕΣ

Κατεύθυνση: Μεγάλα Δεδομένα και Αναλυτική

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αλγόριθμοι Μηχανικής Μάθησης σε Ανομοιογενή Δεδομένα:

Πρόβλεψη της HIV λοίμωξης σε Χρήστες Ενδοφλέβιων Ναρκωτικών της Αθήνας

Εκπόνηση: Σωτήριος Ρούσσος, ΜΕ1942

Επιβλέπων: Ηλίας Μαγκλογιάννης

Πειραιάς

Φεβρουάριος 2022

Στον αδερφό μου, Γιάννη

Ευχαριστίες

Πρώτα απ’ όλα, θα ήθελα να ευχαριστήσω θερμά τον Καθηγητή κ. Ηλία Μαγκλογιάννη, πρόεδρο του Τμήματος Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς και επιβλέποντά μου, για την πολύτιμη βοήθεια και καθοδήγησή του, αλλά και για την υπομονή του κατά τη διάρκεια εκπόνησης της διπλωματικής μου.

Ευχαριστώ ιδιαιτέρως τους κ. Χρήστο Δουλκερίδη, Αναπληρωτή Καθηγητή και κ. Ορέστη Τελέλη, Επίκουρο Καθηγητή του Τμήματος Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς, μέλη της εξεταστικής επιτροπής, για την προθυμία τους να απαντήσουν σε οποιοδήποτε ερευνητικό ερώτημα παρουσιαζόταν κατά τη διάρκεια εκπόνησης της διπλωματικής μου. Επιπλέον, ευχαριστώ θερμά την κα. Βάνα Σύψα, Αναπληρώτρια Καθηγήτρια Επιδημιολογίας και Ιατρικής Στατιστικής της Ιατρικής Σχολής του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών για τη διάθεση των δεδομένων αλλά κυρίως για τη στήριξή της κατά τη διάρκεια των σπουδών μου.

Τέλος, θα ήθελα να ευχαριστήσω όλους τους φίλους(ες) μου, τόσο για την ηθική υποστήριξη, όσο και για την κατανόησή τους, που έδειξαν μέχρι την ολοκλήρωση των σπουδών μου. Πάνω απ’ όλα, είμαι ευγνώμων στους γονείς μου, Ηλία και Μαρία για την ολόψυχη αγάπη και υποστήριξή τους όλα αυτά τα χρόνια.

Σωτήριος Ρούσσος

Περίληψη

Η μηχανική μάθηση διανύει μια περίοδο συνεχούς ανάπτυξης. Τα τελευταία χρόνια, όλο και περισσότερο, οι τεχνικές μηχανικής μάθησης χρησιμοποιούνται στην Ιατρική για διάφορα νοσήματα μεταξύ άλλων και των λοιμωδών νοσημάτων, όπως η HIV λοίμωξη. Στις αρχές του 2011 σημειώθηκε επιδημική έκρηξη της HIV λοίμωξης στον πληθυσμό των χρηστών ενδοφλέβιων ναρκωτικών (ΧΕΝ) της Αθήνας. Το Πανεπιστήμιο Αθηνών σε συνεργασία με τον Οργανισμό Κατά των Ναρκωτικών υλοποίησαν το πρόγραμμα ΑΡΙΣΤΟΤΕΛΗΣ, με σκοπό τόσο τον έλεγχο όσο και τη διασύνδεση σε φροντίδα των ΧΕΝ με HIV λοίμωξη. Σκοπός της παρούσας διπλωματικής εργασίας είναι η εύρεση βέλτιστου ταξινομητή για την HIV λοίμωξη σε ΧΕΝ της Αθήνας.

Τα δεδομένα που χρησιμοποιήθηκαν προήλθαν από το πρόγραμμα ΑΡΙΣΤΟΤΕΛΗΣ και αφορούσαν στους 3.320 μοναδικούς ΧΕΝ. Επιπροσθέτως, περιείχαν πληροφορίες για τα δημογραφικά χαρακτηριστικά, τη χρήση ουσιών, τις σεξουαλικές συμπεριφορές και τα προγράμματα μείωσης της βλάβης (προγράμματα υποκατάστασης με οπιοειδή, λήψη δωρεάν συρίγγων και άλλα). Εφαρμόστηκαν πέντε αλγόριθμοι ταξινόμησης (Logistic Regression, Random Forest, Support Vector Machines, k-Εγγύτεροι Γείτονες και Decision Tree) χρησιμοποιώντας τα δεδομένα: 1) χωρίς επαναδειγματοληψία, 2) με υποδειγματοληψία, 3) με τυχαία υπερδειγματοληψία, 4) με τεχνική υπερδειγματοληψίας συνθετικής μειονότητας και 5) με προσαρμοστική συνθετική μέθοδο δειγματοληψίας. Οι προαναφερθείσες περιπτώσεις εφαρμόστηκαν στο σύνολο των χαρακτηριστικών, ύστερα από την επιλογή μέρους αυτών καθώς και έπειτα από ανάλυση σε κύριες συνιστώσες.

Την καλύτερη επίδοση την είχε ο αλγόριθμος Random forest όταν εφαρμόστηκε σε τυχαία υπερδειγματοληψία. Η ευαισθησία, η ορθότητα καθώς και το AUC score ήταν 0.9929, 0.9805 και 0.9967, αντίστοιχα. Επιλέγοντας 34 από τα 112 χαρακτηριστικά η ευαισθησία, η ορθότητα καθώς και το AUC score ήταν 0.9929, 0.9751 και 0.9967, αντίστοιχα.

Συμπερασματικά, το αποτέλεσμα που προέκυψε από τον έλεγχο για την HIV λοίμωξη σε ΧΕΝ της Αθήνας προβλέφθηκε ορθά σε υψηλά ποσοστά, καθιστώντας

τους αλγορίθμους ως ένα επιπλέον εργαλείο για την έγκαιρη ανίχνευση των οροθετικών ΧΕΝ, προκειμένου να αποφευχθεί μια νέα επιδημική έκρηξη.

Λέξεις Κλειδιά

Μη ισορροπημένα δεδομένα, Αλγόριθμοι μηχανικής μάθησης, HIV, Χρήστες ενδοφλέβιων ναρκωτικών.

Abstract

Machine learning is going through a period of continuous development. In recent years, more and more, machine learning techniques are being used in medicine for numerous diseases including infectious diseases, such as HIV infection. At the beginning of 2011, there was an HIV outbreak in people who inject drugs (PWID) in the metropolitan area of Athens. The University of Athens, in collaboration with the Organization Against Drugs, implemented the ARISTOTLE program with the aim of both testing and linking to HIV care. The aim of this thesis is to find the best classifier for HIV infection in PWID.

Data from the ARISTOTLE program was used and concerned 3320 unique PWID. The data included information on demographic characteristics, substance use, sexual behavior, and information about harm reduction programs (opioid substitution therapy, free syringes, etc.). Five classification algorithms (Logistic Regression, Random Forest, Support Vector Machines, k-Nearest Neighbors, and Decision Tree) were used to the data: 1) without resampling; 2) by random undersampling; 3) by random oversampling; 4) by synthetic minority oversampling technique and 5) by adaptive synthetic sampling method. These cases were applied to all features, after feature selection and after principal components analysis.

The Random Forest algorithm performed best when random oversampling was used. Sensitivity, accuracy, and AUC score were 0.9929, 0.9805 and 0.9967, respectively. Selecting 34 of the 112 characteristics, the sensitivity, accuracy, and AUC score were 0.9929, 0.9751 and 0.9967, respectively.

In conclusion, the status of HIV infection in the sample of PWID in Athens was correctly predicted at high rates, making algorithms an additional tool for early diagnosis in HIV cases, in order to avoid a new HIV outbreak.

Keywords

Imbalanced data, Machine learning algorithms, HIV, People who inject drugs.

Περιεχόμενα

Λίστα Πινάκων

Λίστα Εικόνων

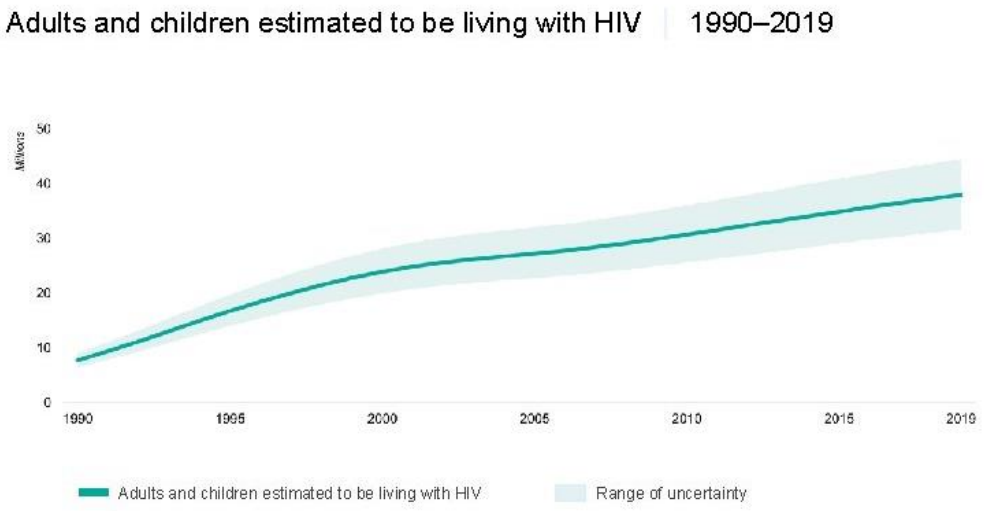
1. Εισαγωγή

1.1 Ο ιός HIV και η νόσος AIDS

Στις αρχές του 1981 καταγράφονται τα πρώτα κλινικά περιστατικά της νόσου, που λίγο αργότερα ονομάστηκε σύνδρομο επίκτητης ανοσολογικής ανεπάρκειας (AIDS). Σε νέους, οι οποίοι έως τότε ήταν υγιείς, παρατηρήθηκε αύξηση στην επίπτωση ορισμένων σπάνιων νοσημάτων, όπως η πνευμονία από πνευμοκύστη, μια δυνητικά θανατηφόρος ευκαιριακή λοίμωξη και το σάρκωμα Kaposi, μια ιογενής μορφή καρκίνου του δέρματος. Διαπιστώθηκε ότι το κοινό χαρακτηριστικό των ασθενών αυτών ήταν ο χαμηλός αριθμός των Τ4 λεμφοκυττάρων. Το Κέντρο Ελέγχου και Πρόληψης Νοσημάτων (CDC) των ΗΠΑ τον Ιούνιο του 1981 δημοσίευσε στην εβδομαδιαία περιοδική του έκδοση, Morbidity and Mortality Weekly Report, μια κλινική περιγραφή των πέντε πρώτων ασθενών, με σκοπό να ενθαρρύνει τους γιατρούς να αναφέρουν παρόμοια περιστατικά. Μέχρι το τέλος του έτους είχαν αναφερθεί 337 περιστατικά [1]. Το CDC αρχικά δεν είχε κάποια επίσημη ονομασία για την νόσο αυτή, όμως τον Ιούλιο του 1982 εισήχθη ο όρος AIDS. Το 1983 η επιστημονική κοινότητα ανακάλυψε τον ιό, που ήταν υπεύθυνος για το AIDS και το 1986 ο ιός πήρε την τελική του ονομασία, ιός της ανθρώπινης ανοσοανεπάρκειας (HIV) [2]. Υπάρχουν δύο τύποι του ιού, οι HIV-1 και HIV-2. Οι περισσότερες μολύνσεις από τον ιό οφείλονται στον HIV-1, ο οποίος κατανέμεται σε παγκόσμιο επίπεδο ενώ, ο HIV-2 εμφανίζεται κυρίως στη Δυτική Αφρική και τη Νοτιοδυτική Ινδία. Ωστόσο, έχουν αναφερθεί σποραδικά περιστατικά του HIV-2 και σε ευρωπαϊκές χώρες [3].

Η πανδημία HIV αποτελεί ένα από τα σημαντικότερα προβλήματα δημόσιας υγείας, σε παγκόσμιο επίπεδο. Τα άτομα που ζουν με τον ιό HIV αυξάνονται συνεχώς από το 1990 (Εικόνα 1). Έως και το τέλος του 2020 εκτιμάται πως τα άτομα αυτά, παγκοσμίως ανέρχονται στα 38 εκατομμύρια. Οι νέες μολύνσεις, καθώς και οι θάνατοι οι οποίοι σχετίζονται με το AIDS για το 2019 σε παγκόσμιο επίπεδο ήταν 1,7 και 690 χιλιάδες αντιστοίχως [4]. Το Δεκέμβριο του 2013 το Διεθνές τμήμα του Οργανισμού Ηνωμένων Εθνών για την καταπολέμηση του AIDS (UNAIDS) έθεσε νέους

στόχους μέχρι το 2020, όπως να γνωρίζει το 90% των ανθρώπων που έχουν μολυνθεί με τον ιό HIV, την κατάστασή τους, να έχει πρόσβαση σε αντιρετροϊκή θεραπεία το 90% των ανθρώπων που είναι HIV θετικοί και τέλος το 90% των ανθρώπων αυτών να έχουν μη ανιχνεύσιμο ιικό φορτίο [5].

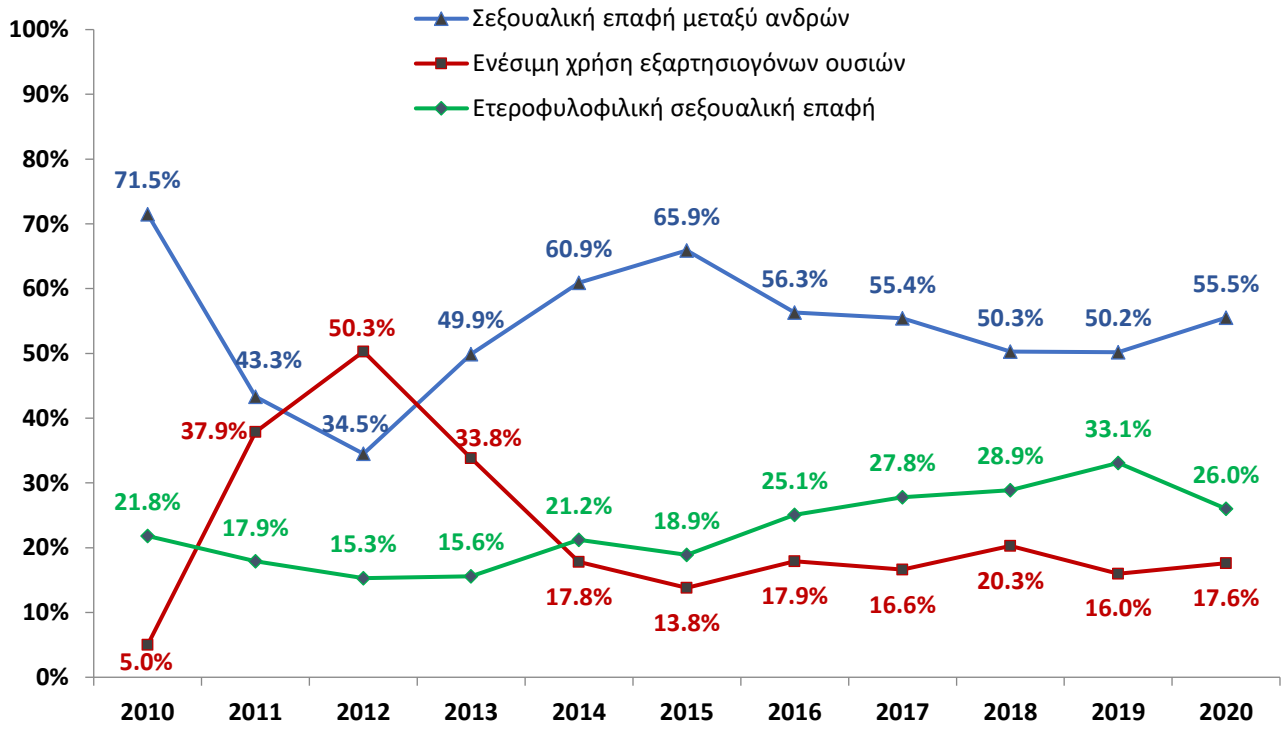


Εικόνα 1. Εκτίμηση του αριθμού των ενηλίκων και των παιδιών που ζουν παγκοσμίως με HIV από το 1990 έως και το 2019 [4].

Στην Ελλάδα, η πρώτη περίπτωση AIDS εμφανίστηκε το 1981 και δύο χρόνια αργότερα καταγράφηκε ο πρώτος θάνατος από τη νόσο. Η δήλωση των περιπτώσεων AIDS στην Ελλάδα ξεκίνησε το 1984, ενώ των HIV οροθετικών ατόμων το 1998 και σε ευρωπαϊκό επίπεδο τον Ιανουάριο του 1999. Είναι ανώνυμη, απόρρητη και υποχρεωτική και γίνεται με τη χρήση των αρχικών ονοματεπωνύμου και της ημερομηνίας γέννησης ως αναγνωριστικών προσωπικών στοιχείων και με τον τρόπο αυτό επιτυγχάνεται ο έλεγχος για πιθανές διπλοεγγραφές. Παρ' όλα αυτά, η εγγενής αδυναμία της σωστής και ακριβούς καταγραφής αυτών των αναγνωριστικών οδηγεί στην ύπαρξη κάποιων διπλοεγγραφών, που δεν μπορούν ν' ανιχνευθούν. Σύμφωνα με την ετήσια έκθεση του ΚΕΕΛΠΝΟ, έως το 2020 στη χώρα μας, έχουν καταγραφεί συνολικά 18.710 περιστατικά HIV οροθετικών ατόμων (συμπεριλαμβανομένων και των περιπτώσεων AIDS), ενώ οι νέες διαγνώσεις HIV, για το ίδιο έτος, ανήλθαν στις 601 και οι θάνατοι οφειλόμενοι στο AIDS σε 41 [6].

Σε αρκετές χώρες, η επιδημία του HIV επικεντρώνεται σε ευάλωτες ομάδες πληθυσμού, όπως για παράδειγμα σε άνδρες που έχουν σεξουαλικές επαφές με

άνδρες, σε χρήστες ενέσιμων ναρκωτικών, σε εργαζόμενους στο σεξ και σε φυλακισμένους. Στην Ελλάδα, μέχρι και το 2011 ο κύριος τρόπος μετάδοσης του ιού ήταν η σεξουαλική επαφή μεταξύ ανδρών. Ωστόσο, στις αρχές του 2011 παρατηρήθηκε μια σημαντική αύξηση στα δηλωθέντα HIV οροθετικά άτομα που ανήκαν στην κατηγορία των χρηστών ενδοφλέβιων ναρκωτικών ουσιών, με αποτέλεσμα ο κύριος τρόπος μετάδοσης το 2012 να ήταν η ενέσιμη χρήση εξαρτησιογόνων ουσιών (Εικόνα 2) [6].



Εικόνα 2. Ποσοστιαία αναλογία διαγνώσεων HIV με γνωστό τρόπο μετάδοσης κατά κατηγορία μετάδοσης και έτος διάγνωσης (Ελλάδα, 2010-2020). (Το γράφημα δημιουργήθηκε βάσει των στοιχείων [6]).

1.1.1 HIV σε χρήστες ενδοφλέβιων ναρκωτικών

Οι χρήστες ενδοφλέβιων ναρκωτικών (ΧΕΝ) αποτελούν πληθυσμό με υψηλό φορτίο HIV λοίμωξης [7]. Σύμφωνα με το UNAIDS εκτιμάται πως οι ΧΕΝ έχουν 29 φορές μεγαλύτερο κίνδυνο να μολυνθούν με HIV σε σύγκριση με τον γενικό πληθυσμό. Σε παγκόσμιο επίπεδο οι ΧΕΝ εκτιμάται πως ήταν 11,2 εκατομμύρια το 2019, από τους

οποίους τα 1,4 εκατομμύρια έχουν μολυνθεί με HIV, 5,6 εκατομμύρια ζουν με HCV και 1,2 εκατομμύρια ζουν με HCV και HIV [8].

Οι πρώτες επιδημικές εκρήξεις στον πληθυσμό των ΧΕΝ, τόσο στην Ευρώπη, όσο και στην Αμερική, καταγράφηκαν την περίοδο 1980-1990 [9]. Οι επιδημικές εκρήξεις αυτές σχετίστηκαν με την ανταλλαγή συρίγγων αλλά και με την ύπαρξη πυκνού δικτύου ενέσιμης χρήσης μεταξύ των ΧΕΝ. Από τα μέσα της δεκαετίας του 90, δεν σημειώθηκαν επιδημίες σε ευρωπαϊκές πόλεις με εξαίρεση τις χώρες της πρώην Σοβιετικής Ένωσης. Ωστόσο, μετά το 2010, η κατάσταση άλλαξε. Σημειώθηκε μια σειρά από νέες HIV επιδημίες στην Ευρώπη και τη Βόρεια Αμερική [10]. Η πρώτη και μεγαλύτερη ήταν εκείνη της Αθήνας, την περίοδο 2011-2013 [10]. Ο επιπολασμός HIV από λιγότερο του 1% που ήταν πριν το 2010, αυξήθηκε στο 16,5% το 2013 [11].

1.1.2 Πρόγραμμα ΑΡΙΣΤΟΤΕΛΗΣ

Από τις αρχές του 2011 παρατηρήθηκε αύξηση στις νέες δηλώσεις κρουσμάτων HIV λοίμωξης στην Αθήνα σε χρήστες ενδοφλέβιων ναρκωτικών. Τον Ιούλιο του 2011 δηλώθηκε η επιδημία αυτή στο Σύστημα Έγκαιρης Προειδοποίησης (Early Warning System) του Ευρωπαϊκού Κέντρου Πρόληψης και Ελέγχου Νόσων (European Centre for Disease Prevention and Control, ECDC).

Τον Αύγουστο του 2012 ξεκίνησε το πρόγραμμα «ΑΡΙΣΤΟΤΕΛΗΣ» το οποίο ήταν μία πρωτοβουλία του Πανεπιστημίου Αθηνών σε συνεργασία με τον ΟΚΑΝΑ. Ο σκοπός του ΑΡΙΣΤΟΤΕΛΗ ήταν τριπλός. α) Να γίνει ένας ταχύς έλεγχος σε όλους τους χρήστες ενδοφλέβιων ναρκωτικών, που κατοικούσαν στην ευρύτερη περιοχή της Αθήνας, β) όλα τα άτομα που συμμετείχαν στο πρόγραμμα να λάβουν όλη εκείνη τη φροντίδα όπως ορίζεται από τον Παγκόσμιο Οργανισμό Υγείας (ΠΟΥ), UNODC και UNAIDS με τελικό στόχο να συμβάλει στη μείωση της επίπτωσης του HIV-1 στους χρήστες ενδοφλέβιων ναρκωτικών. Οι δευτερεύοντες σκοποί του προγράμματος ήταν η εκτίμηση του επιπολασμού HIV στους ΧΕΝ, κατά τη διάρκεια του προγράμματος, η αποσαφήνιση των επιδημιολογικών, δημογραφικών, συμπεριφορικών, κοινωνικών και ιολογικών χαρακτηριστικών της επιδημίας, η περιγραφή φυλογενετικών και κοινωνικών δικτύων των χρηστών, η βελτίωση της

διασύνδεσής τους με κλινικές που παρέχουν φροντίδα και θεραπεία για τον HIV, καθώς και με προγράμματα υποκατάστασης οπιοειδών και τέλος η επίτευξη της παραμονής τους σε αυτά τα προγράμματα, καθώς η ομάδα αυτή των χρηστών συχνά αποχωρεί από αντίστοιχα προγράμματα.

Ο πληθυσμός-στόχος του ΑΡΙΣΤΟΤΕΛΗ ήταν όλα τα άτομα ηλικίας από 18 ετών και άνω, που διέμεναν σε μια ευρύτερη περιοχή της Αθήνας και έκαναν χρήση ενδοφλέβιων ναρκωτικών τους τελευταίους 12 μήνες. Το πρόγραμμα αυτό, διήρκεσε 16 μήνες (Αύγουστος 2012 – Δεκέμβριος 2013), κατά τους οποίους πραγματοποιήθηκαν πέντε διαδοχικοί κύκλοι κατευθυνόμενης από τους συμμετέχοντες δειγματοληψίας (Respondent Driven Sampling – RDS) [12].

Σε κάθε κύκλο υπήρχαν από πέντε έως δέκα «σπόροι» (seeds), όπου κάθε ένας έπαιρνε έως και 3 κουπόνια, τα οποία με τη σειρά τους τα διένεμαν σε άλλους χρήστες ενδοφλέβιων ναρκωτικών. Κατά τη διάρκεια της μελέτης, συμμετείχαν 3.320 μοναδικά άτομα, ενώ το συνολικό δείγμα σε κάθε κύκλο ξεπερνούσε τα 1.400 άτομα, αφού ο κάθε χρήστης είχε τη δυνατότητα να συμμετάσχει σε πολλαπλούς κύκλους, αλλά μόνο μια φορά σε κάθε κύκλο. Συνολικά, συλλέχθηκαν 7.110 ερωτηματολόγια και δείγματα αίματος, από τα οποία βρέθηκε πως 547 άτομα (16.5%) είχαν μολυνθεί με τον HIV. Στον πρώτο κύκλο το ποσοστό των αδιάγνωστων HIV(+), δηλαδή εκείνων που δεν γνώριζαν πως είχαν μολυνθεί ήταν 84,3%, ενώ στον τελευταίο κύκλο 15,0% [11].

1.2 Σκοπός της διπλωματικής

Σκοπός της παρούσας διπλωματικής εργασία είναι η εύρεση ταξινομητή στον πληθυσμό των ΧΕΝ για τη μόλυνση ή μη με HIV.

1.3 Δομή της διπλωματικής

Η παρούσα διπλωματική εργασία αποτελείται από 8 κεφάλαια. Στο 1ο κεφάλαιο γίνεται μια εισαγωγή αναφορικά με την HIV λοίμωξη και τον πληθυσμό των ΧΕΝ. Επίσης, παρουσιάζεται ο σκοπός της διπλωματικής εργασίας. Το 2ο κεφάλαιο αφορά

στη μηχανική μάθηση και δίνεται έμφαση σε προβλήματα τόσο από το χώρο της Ιατρικής που αυτή έχει χρησιμοποιηθεί, όσο και της HIV λοίμωξης. Στο 3ο κεφάλαιο παρουσιάζονται οι αλγόριθμοι που χρησιμοποιούνται για την ταξινόμηση καθώς και οι μετρικές μέσω των οποίων αξιολογείται η επίδοση ενός ταξινομητή. Στο 4ο κεφάλαιο γίνεται μια σύντομη επισκόπηση στα μη ισορροπημένα δεδομένα και παρουσιάζονται οι διαθέσιμες τεχνικές για την αντιμετώπιση της ανισορροπίας. Στο 5ο κεφάλαιο παρουσιάζονται οι τεχνικές προεπεξεργασίας των δεδομένων που χρησιμοποιήθηκαν στα πλαίσια της διπλωματικής εργασίας. Το κεφάλαιο 6 αφορά το πειραματικό κομμάτι της μελέτης και γίνεται περιγραφή των δεδομένων και παρουσίαση των αποτελεσμάτων βάσει των διάφορων σεναρίων που χρησιμοποιήθηκαν. Τέλος, στο κεφάλαιο 7 αποτυπώνονται τα συμπεράσματα της μελέτης, ενώ στο κεφάλαιο 8, γίνεται πρόταση για μελλοντικές εργασίες που μπορεί να γίνουν πάνω στην HIV λοίμωξη και την μηχανική μάθηση.

2. Θεωρητικό υπόβαθρο – Βιβλιογραφική ανασκόπηση

Οι υπολογιστές πριν το 1949 είχαν τη δυνατότητα να εκτελέσουν εντολές αλλά δεν μπορούσαν να τις αποθηκεύσουν, δηλαδή δεν είχαν τη δυνατότητα να «θυμηθούν» τι έκαναν. Το κόστος για την αγορά ενός υπολογιστή ήταν πολύ υψηλό και ως αποτέλεσμα υπολογιστές υπήρχαν μόνο σε μεγάλα πανεπιστήμια ή σε μεγάλες εταιρείες τεχνολογίας. Για να αλλάξει αυτό θα έπρεπε επιφανείς επιστήμονες να πείσουν αυτούς που αποτελούσαν πηγές χρηματοδότησης ότι άξιζε να προσπαθήσουν για την ύπαρξη «νοημοσύνης» στους υπολογιστές.

Πέντε χρόνια αργότερα, οι Allen Newell, Cliff Shaw, και Herbert Simon κατασκεύασαν ένα πρόγραμμα που σχεδιάστηκε για να μιμείται τις ανθρώπινες δεξιότητες επίλυσης προβλημάτων. Το 1956, στο συνέδριο Dartmouth Summer Research Project on Artificial Intelligence (DSRPAI), παρουσιάστηκε το πρόγραμμα αυτό και από πολλούς θεωρείται πως είναι το πρώτο πρόγραμμα τεχνητής νοημοσύνης (Artificial Intelligence – AI), όρος που επινοήθηκε για πρώτη φορά στο συνέδριο [13, 14].

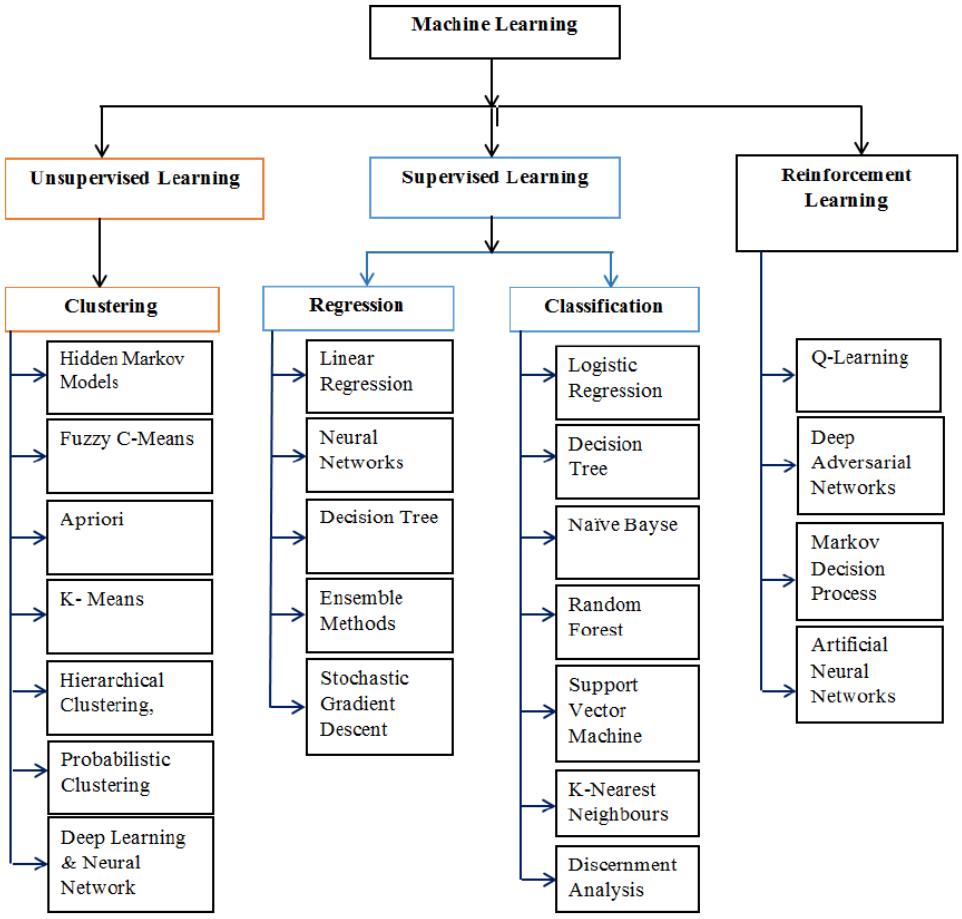
Τεχνητή νοημοσύνη είναι ο κλάδος της επιστήμης των υπολογιστών που ασχολείται με τη σχεδίαση και την υλοποίηση ευφυών (νοημόνων) υπολογιστικών συστημάτων που μιμούνται στοιχεία της ανθρώπινης συμπεριφοράς [15].

2.1 Μηχανική μάθηση

Η μηχανική μάθηση αποτελεί έναν κλάδο της τεχνητής νοημοσύνης και έχουν δοθεί αρκετοί ορισμοί έως σήμερα. Ένας γενικός ορισμός για τη μηχανική μάθηση, δόθηκε από τον Mitchell το 1997: *«Ένα πρόγραμμα υπολογιστή λέμε ότι μαθαίνει από την εμπειρία E ως προς κάποια κλάση εργασιών T και μέτρο απόδοσης P, αν η απόδοσή του σε εργασίες από το T, όπως μετριέται από το P, βελτιώνεται μέσω της εμπειρίας E.»* [16]. Ο Mehryar Mohri, ορίζει τη μηχανική μάθηση ως τις υπολογιστικές μεθόδους (αλγορίθμους) που χρησιμοποιούν εμπειρία/υπάρχουσα γνώση (δεδομένα) για να βελτιώσουν την απόδοση ενός συστήματος ή να πραγματοποιήσουν ακριβείς

προβλέψεις. Η ακρίβεια των προβλέψεων είναι άρρηκτα συνδεδεμένη με την ποιότητα και το μέγεθος των παρεχόμενων δεδομένων [17].

Οι διάφοροι τύποι μηχανικής μάθησης που έχουν αναπτυχθεί, χρησιμοποιούνται ανάλογα με τη φύση του προβλήματος και διαφέρουν ως προς α) τον τύπο δεδομένων που είναι διαθέσιμα για την εκπαίδευση, β) τη σειρά και τη μέθοδο με την οποία λαμβάνονται τα δεδομένα εκπαίδευσης και γ) τα δεδομένα ελέγχου που χρησιμοποιούνται για την αξιολόγηση του αλγορίθμου μάθησης [15, 17]. Υπάρχουν πολλοί διαφορετικοί τύποι μάθησης, ωστόσο μπορούν να ταξινομηθούν σε τρεις μεγάλες κατηγορίες: τη μάθηση χωρίς επίβλεψη, την εποπτευόμενη μάθηση και την πρόσφατα αναδυόμενη και πιο δημοφιλή ενισχυτική μάθηση [18]. Κάθε μια από τις τρεις κατηγορίες περιέχει πληθώρα αλγορίθμων, ωστόσο οι πιο συχνά χρησιμοποιούμενοι παρουσιάζονται στην Εικόνα 3 [19].



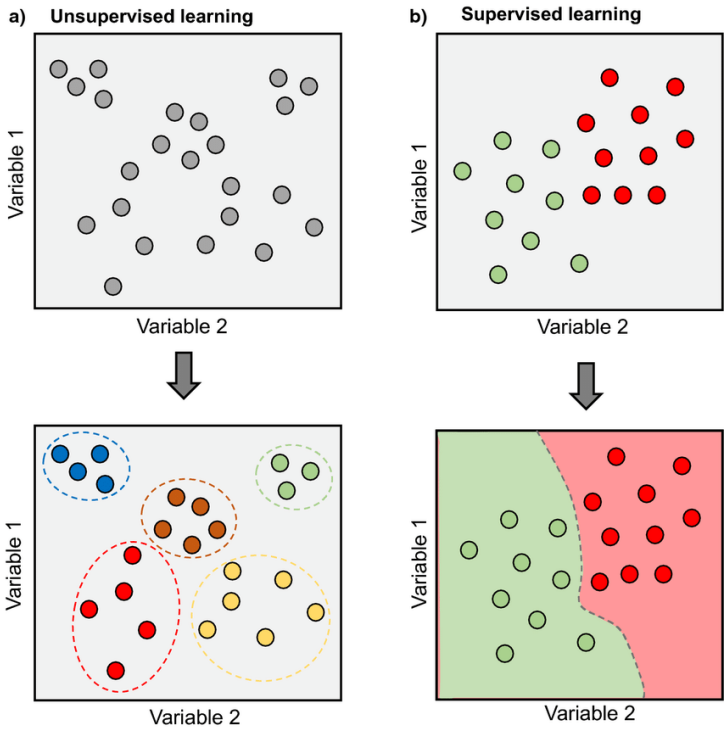
Εικόνα 3. Ευρέως χρησιμοποιούμενοι αλγόριθμοι μηχανικής μάθησης ανά κατηγορία μάθησης [19].

2.1.1 Μάθηση χωρίς επίβλεψη (Unsupervised learning)

Στην μάθηση χωρίς επίβλεψη ο αλγόριθμος κατασκευάζει ένα μοντέλο χωρίς να υπάρχει κάποια προϋπάρχουσα γνώση. Στην πραγματικότητα καλείται ο αλγόριθμος να βρει τη δομή των δεδομένων που εισέρχονται, ώστε με βάση κάποιο κριτήριο το κάθε σύνολο να είναι διαχωρίσιμο από το άλλο. Συνήθως, χρησιμοποιείται σε προβλήματα ομαδοποίησης (clustering) και ελάττωσης διαστάσεων (dimensionality reduction) [20] (Εικόνα 4).

2.1.2 Μάθηση με επίβλεψη (Supervised learning)

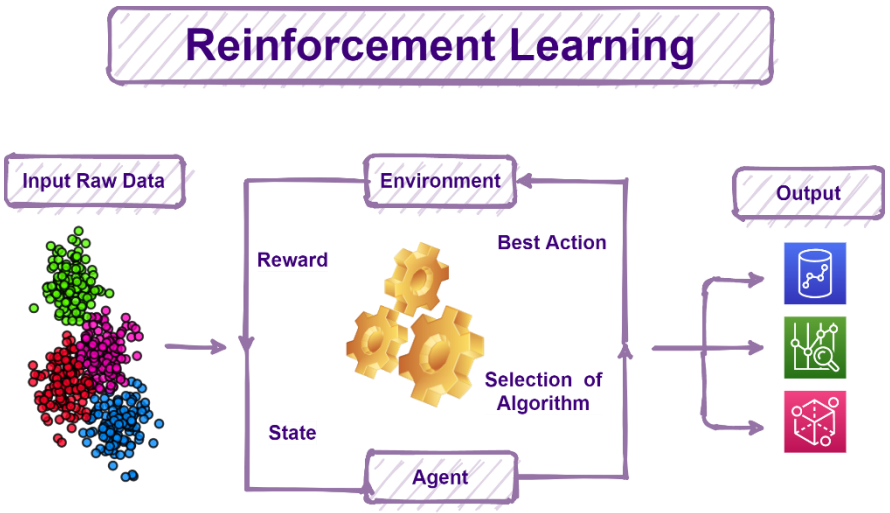
Είναι η διαδικασία κατά την οποία ο αλγόριθμος εκπαιδεύεται σε ένα σύνολο δεδομένων (παραδείγματα), των οποίων είναι γνωστή η κατηγορία που ανήκουν. Κάθε παράδειγμα αποτελείται από ένα σύνολο εισόδου (διάνυσμα χαρακτηριστικών) και μια τιμή εξόδου. Από την εκπαίδευση προκύπτει ένα μοντέλο βάσει του οποίου γίνεται ο χαρακτηρισμός νέων παραδειγμάτων [21]. Η μάθηση με επίβλεψη χρησιμοποιείται κυρίως σε προβλήματα ταξινόμησης και πρόγνωσης (Εικόνα 4).



Εικόνα 4. Σχηματική αναπαράσταση ενός μοντέλου μάθησης α) χωρίς επίβλεψη και β) με επίβλεψη [22].

2.1.3 Μάθηση με ενίσχυση (Reinforcement learning)

Στην μάθηση με ενίσχυση, ο αλγόριθμος εκπαιδεύεται χωρίς κάποιο σταθερό σύνολο δεδομένων εισόδου. Η μάθηση προέρχεται από την αλληλεπίδραση του μοντέλου με το περιβάλλον. Στόχος είναι να μεγιστοποιηθούν οι επιβραβεύσεις για τις ενέργειες που πραγματοποιεί κατά την αλληλεπίδρασή του με το περιβάλλον. Δεν υπάρχει ενημέρωση ως προς το ποιες ενέργειες πρέπει να κάνει ο αλγόριθμος αλλά αντίθετα θα πρέπει να ανακαλύψει τις ενέργειες εκείνες που θα αποφέρουν τη μεγαλύτερη ανταμοιβή (Εικόνα 5) [21].



Εικόνα 5. Σχηματική αναπαράσταση ενός μοντέλου ενισχυτικής μάθησης [23].

2.2 Μηχανική μάθηση στην Ιατρική

Η μηχανική μάθηση διανύει μια περίοδο συνεχούς ανάπτυξης. Στην καθημερινή μας ζωή χρησιμοποιούμε πληθώρα εφαρμογών, που τις περισσότερες φορές αγνοούμε πως προέρχονται από το χώρο της μηχανικής μάθησης. Η αναγνώριση εικόνας, η αναγνώριση λόγου, η πρόβλεψη της κυκλοφοριακής συμφόρησης, η πρόταση προϊόντων, τα αυτόματα αυτοκίνητα, η ιατρική διάγνωση αποτελούν μερικές από τις εφαρμογές της μηχανικής μάθησης.

Τα τελευταία χρόνια, όλο και περισσότερο, οι τεχνικές μηχανικής μάθησης βρίσκουν εφαρμογή σε διάφορες ειδικότητες της Ιατρικής όπως στην καρδιολογία

[24, 25], στη Γαστρεντερολογία, στην Οφθαλμολογία, στη Δερματολογία, στη Νευρολογία και την Παθολογία [26]. Επιπλέον, έχουν εφαρμοστεί σε διάφορα νοσήματα όπως ο διαβήτης [25, 27], ο καρκίνος [28, 29], τα λοιμώδη νοσήματα [30-33], τα οποία απασχολούν την Ιατρική κοινότητα και σύμφωνα με τον ΠΟΥ βρίσκονται στις 10 πρώτες αιτίες θανάτου παγκοσμίως για το 2019 [34].

2.3 Μηχανική μάθηση και HIV – Ανασκόπηση εργασιών

Ο Παγκόσμιος Οργανισμός Υγείας (WHO) έχει θέσει ως προτεραιότητα την εξάλειψη της HIV λοίμωξης έως το 2030 [35]. Η επίτευξη του στόχου αυτού εξαρτάται τόσο από την πρόσβαση σε αντιρετροϊκή θεραπεία και τη λήψη προφύλαξης πριν την επαφή (preexposure prophylaxis – PrEP) σε όσους αποτελούν ομάδες υψηλού κινδύνου [36] όσο και από την έγκαιρη ανίχνευση των HIV λοιμώξεων [37]. Στη συνέχεια παρουσιάζονται μόνο μελέτες που έχουν ασχοληθεί με την πρόβλεψη του αποτελέσματος της HIV λοίμωξης, ο οποίος είναι και ο κύριος σκοπός της παρούσας διπλωματικής εργασίας.

Το 1990 θέλοντας να εξετάσουν εναλλακτικές λύσεις πέραν του ορολογικού ελέγχου για την HIV λοίμωξη σε έγκυες γυναίκες στην Κινσάσα του Ζαΐρ, οι ερευνητές εφάρμοσαν λογιστική παλινδρόμηση με διάφορους παράγοντες κινδύνου που μπορεί να σχετίζονται με την HIV λοίμωξη. Ωστόσο, μόνο το μοντέλο που συνδύαζε πληροφορίες για συμπτώματα του AIDS/HIV, όπως για παράδειγμα χρόνιος πυρετός, διάρροια ή έντονη απώλεια βάρους ήταν προβλεπτικό για το αποτέλεσμα του HIV [38]. Λίγο αργότερα, το 1992 οι ερευνητές εφαρμόζοντας λογιστική παλινδρόμηση, σε μη ορολογικά δεδομένα, για να επιτευχθεί 80% σωστή ταξινόμηση στους οροθετικούς, προέκυπτε λανθασμένη ταξινόμηση στο 50% των αρνητικών [39].

Το 2001, εφαρμόστηκαν τεχνητά νευρωνικά δίκτυα για την πρόγνωση του HIV/AIDS χρησιμοποιώντας κλινικά και δημογραφικά χαρακτηριστικά. Η μέγιστη ακρίβεια που επιτεύχθηκε ήταν 88% [40]. Το 2018 χρησιμοποιήθηκαν δεδομένα από την Γκάνα, τα οποία αφορούσαν σε γυναίκες εργαζόμενες στο σεξ για το έτος 2015. Έγινε χρήση πέντε αλγορίθμων μηχανικής μάθησης: τυχαίο δέντρο, νευρωνικά δίκτυα, αλγόριθμοι J48, λογιστική παλινδρόμηση και Naive Bayes με σκοπό να

προβλέψουν εάν μια γυναίκα, εργαζόμενη στο σεξ, θα ήταν αρνητική ή θετική για την HIV λοίμωξη. Οι αλγόριθμοι ως δεδομένα εισόδου είχαν ορισμένους κοινωνικοδημογραφικούς και συμπεριφορικούς παράγοντες. Η ακρίβεια των παραπάνω αλγορίθμων ήταν 98,9%, 97,41%, 93,18%, 91,12% και 89,97% αντίστοιχα [41]. Το 2019, χρησιμοποιήθηκαν δεδομένα από το μητρώο της Δανίας προκειμένου να προβλέψουν το HIV αποτέλεσμα. Το πλήρες μοντέλο που περιείχε δημογραφικά χαρακτηριστικά αλλά και πληροφορίες αναφορικά με το ιατρικό ιστορικό είχε την υψηλότερη περιοχή κάτω από την καμπύλη λειτουργικού χαρακτηριστικού δέκτη (Area Under the Curve – AUC) 88.4% [95% Διάστημα εμπιστοσύνης (95% ΔΕ): 87.5%-89.4%] [37].

Παρά τις εφαρμογές μεθόδων μηχανικής μάθησης τόσο στον γενικό πληθυσμό όσο σε συγκεκριμένες ομάδες, όπως για παράδειγμα στις εγκύους γυναίκες, οι προαναφερθείσες μέθοδοι δεν έχουν εφαρμοστεί στον πληθυσμό των χρηστών ενδοφλέβιων ναρκωτικών. Η εύρεση ενός αποδοτικού αλγορίθμου θα βοηθούσε στην προτεραιοποίηση του HIV ελέγχου και την ανάδειξη των παραγόντων υψηλού κινδύνου.

3. Το πρόβλημα ταξινόμησης και οι αλγόριθμοι

Μια υποκατηγορία προβλημάτων τα οποία πραγματεύεται η Μηχανική Μάθηση είναι τα προβλήματα ταξινόμησης (classification problems). Οι αλγόριθμοι που χρησιμοποιούνται για την επίλυση τέτοιων προβλημάτων μαθαίνουν μέσα από ένα σύνολο δεδομένων (δεδομένα εκπαίδευσης), τα οποία γνωρίζουμε εξαρχής σε ποια κλάση ανήκουν. Ο αλγόριθμος μέσα από τις γνωστές περιπτώσεις καλείται να αναπτύξει τη δυνατότητα ταξινόμησης νέων δεδομένων τα οποία δεν ανήκουν στο σύνολο εκπαίδευσης και επομένως, δεν είναι γνωστή η κλάση στην οποία ταξινομούνται [42]. Τις περισσότερες φορές το πρόβλημα της ταξινόμησης χρησιμοποιεί μάθηση με επίβλεψη, δεδομένου ότι υπάρχουν εκ των προτέρων κάποια δεδομένα τα οποία ήδη ανήκουν σε κάποια κλάση [43]. Μετά την εφαρμογή ενός αλγορίθμου ταξινόμησης σε ένα σύνολο δεδομένων διανυσμάτων χαρακτηριστικών, το μοντέλο που προκύπτει ονομάζεται και ταξινομητής (classifier). Με τον όρο κλάσεις εννοούμε τις κατηγορίες που μπορεί να ανήκουν τα δεδομένα. Υπάρχουν 4 διακεκριμένες κατηγορίες προβλημάτων ταξινόμησης [44, 45]:

1. Δυαδικά προβλήματα ταξινόμησης (Binary Classification)
2. Προβλήματα ταξινόμησης πολλαπλών κλάσεων (Multi-Class Classification)
3. Προβλήματα ταξινόμησης δεδομένων που ανήκουν σε πολλαπλές κλάσεις (Multi-Label Classification)
4. Πρόβλημα ανομοιογενούς ταξινόμησης (Imbalanced Classification)

3.1 Αλγόριθμοι ταξινόμησης

Στη βιβλιογραφία υπάρχουν αρκετοί αλγόριθμοι που μπορούν να χρησιμοποιηθούν για τα προβλήματα ταξινόμησης στη μηχανική μάθηση, μεταξύ των οποίων είναι η λογιστική παλινδρόμηση, τα τυχαία δάση, οι μηχανές διανυσμάτων στήριξης, οι κ-Πλησιέστεροι γείτονες και τα δέντρα απόφασης. Στη συνέχεια της παρούσας μελέτης δίνεται μια σύντομη περιγραφή των αλγορίθμων.

3.1.1 Λογιστική Παλινδρόμηση (Logistic Regression)

Η λογιστική παλινδρόμηση (Logistic regression) αποτελεί ένα μοντέλο ταξινόμησης των τιμών μιας μεταβλητής απόκρισης Y με βάση τη θεωρία των πιθανοτήτων. Η μεταβλητή Y είναι δίτιμη και τις περισσότερες φορές παίρνει τις τιμές 0 και 1 που δηλώνουν την πραγματοποίηση ή μη ενός γεγονότος που εκφράζει η μεταβλητή Y.

Αν υποθέσουμε ότι $P(Y_i = 1) = \pi_i$, τότε $P(Y_i = 0) = 1 - \pi_i$. Για να συσχετιστεί η πιθανότητα π_i με τη γραμμική έκφραση $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$, όπου x_j είναι ανεξάρτητες μεταβλητές, χρησιμοποιείται μια συνάρτηση σύνδεσης (link function) $g(\pi)$, με πεδίο ορισμού το διάστημα $(0,1)$ και πεδίο τιμών το διάστημα $(-\infty, +\infty)$. Η συνάρτηση αυτή είναι η $g(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$, η οποία ονομάζεται συνάρτηση λογαρίθμου συμπληρωματικών πιθανοτήτων (logit). Επομένως προκύπτει ότι:

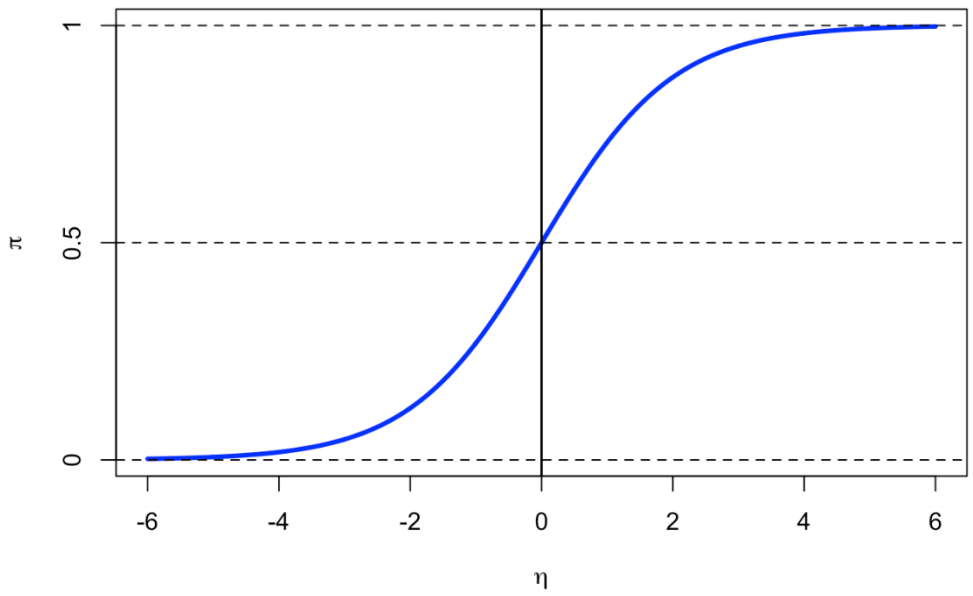
$$\eta = \text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

Η μαθηματική έκφραση του μοντέλου της λογιστικής παλινδρόμησης είναι:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p .$$

Η αντίστροφη συνάρτηση της logit είναι η λογιστική ή σιγμοειδής συνάρτηση (logistic ή sigmoid) με τύπο $S(t) = \frac{1}{1+e^{-t}}$ (Εικόνα 6).

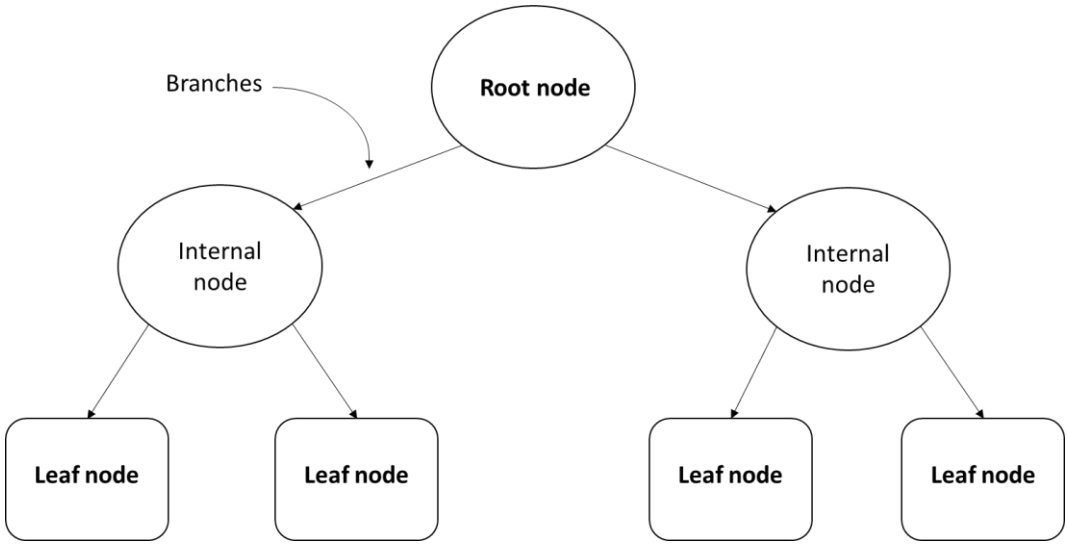
$$\pi = \text{logit}^{-1}(\eta) = \frac{e^{\eta}}{1+e^{\eta}} = \frac{1}{1+e^{-\eta}} = \frac{1}{1+\exp\left[-\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\right)\right]}$$



Εικόνα 6. Η σιγμοειδής συνάρτηση για τιμές στο διάστημα [-6, 6].

3.1.2. Δέντρα Απόφασης (Decision Trees)

Ένα από τα πιο δημοφιλή αλλά και απλά μοντέλα κατηγοριοποίησης είναι τα δένδρα απόφασης. Ένα δέντρο απόφασης αποτελείται από α) έναν ριζικό κόμβο (root node), β) τα κλαδιά (branches), γ) ένα σύνολο από εσωτερικούς (ή μη τερματικούς) κόμβους διαχωρισμού (split nodes) και δ) τους τερματικούς κόμβους ή αλλιώς φύλλα (terminal nodes – leaves) (Εικόνα 7).



Εικόνα 7. Σχηματική αναπαράσταση ενός δέντρου απόφασης.

Κάθε εσωτερικός κόμβος του δέντρου αντιστοιχεί σε ένα χαρακτηριστικό x_i (μεταβλητή διαχωρισμού), δηλαδή σε μια από τις μεταβλητές εισόδου. Κάθε κλαδί, που ενώνει δυο κόμβους αντιστοιχεί σε μια συνθήκη ανάμεσα στο x_i και μια τιμή θ_i του γονικού κόμβου, ενώ κάθε φύλλο αντιστοιχεί σε μια κλάση. Επομένως, γίνεται αντιληπτό πως η επιλογή του χαρακτηριστικού x_i αλλά και της παραμέτρου θ_i είναι ιδιαίτερα σημαντική. Η επιλογή δύναται να γίνει με διάφορα κριτήρια όπως με τη χρήση του Gini index, της εντροπίας και του κέρδους πληροφορίας ή του λάθους ταξινόμησης (misclassification error).

Η δυσκολία με τα δέντρα απόφασης είναι ότι επεκτείνονται πολύ, με την έννοια ότι υπάρχουν αρκετοί κόμβοι, ειδικά αν υπάρχουν διαθέσιμα πολλά χαρακτηριστικά. Προκειμένου να αντιμετωπιστεί αυτή η υπερφόρτωση του δέντρου, ακολουθείται μια διαδικασία γνωστή και ως κλάδεμα (pruning) του δέντρου. Στο κλάδεμα του δέντρου αφαιρούνται κόμβοι ξεκινώντας από τα φύλλα του, προκειμένου να μη μεταβληθεί η συνολική ακρίβεια του μοντέλου.

3.1.3. Τυχαία Δάση (Random Forest)

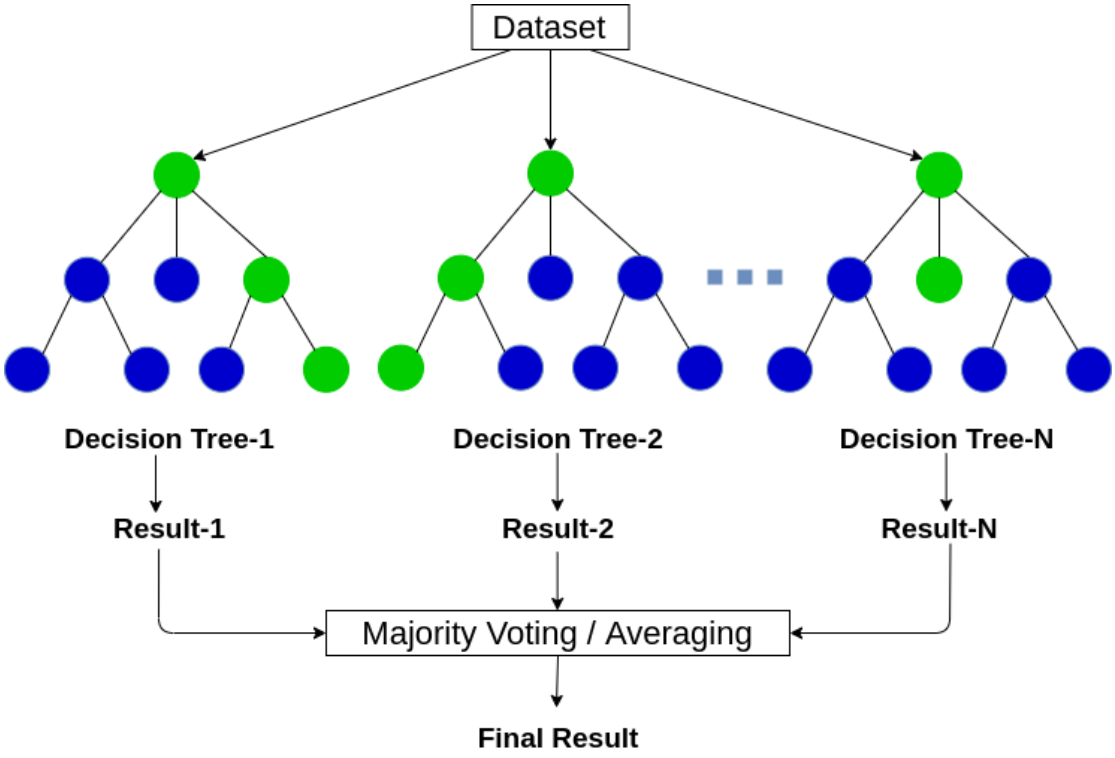
Τα τυχαία δάση αποτελούν μια ειδική κατηγορία των συνδυαστικών μεθόδων ταξινόμησης, που χρησιμοποιεί για ταξινομητές ένα μεγάλο αριθμό ανεξάρτητων δέντρων απόφασης, τα οποία δεν έχουν υποστεί κλάδεμα [46].

Για την κατασκευή ενός δέντρου απόφασης, αρχικά επιλέγεται ένα τυχαίο υποσύνολο των χαρακτηριστικών, τα οποία χρησιμοποιούνται με τυχαίο τρόπο σε κάθε κόμβο μέχρι να κατασκευαστεί το προαναφερθέν δέντρο απόφασης. Η διαδικασία αυτή επαναλαμβάνεται μέχρι να φτιαχτούν όλα τα δέντρα απόφασης που έχουν οριστεί (Εικόνα 8).

Η λογική για τη δημιουργία ενός τυχαίου δάσους είναι ότι για το k-οστό δέντρο απόφασης δημιουργείται ένα τυχαίο διάνυσμα Θ_k, το οποίο είναι ανεξάρτητο από τα προηγούμενα τυχαία διανύσματα $\Theta_1, \Theta_2, ..., \Theta_{k-1}$. Το k-οστό δέντρο αναπτύσσεται από το σύνολο εκπαίδευσης και το διάνυσμα Θ_k, καταλήγοντας σε έναν ταξινομητή

$h = (x, \Theta_{\kappa})$, όπου x το διάνυσμα εισόδου. Έτσι, γίνεται εύκολα αντιληπτό πως το τυχαίο δάσος είναι ένας ταξινομητής αποτελούμενος από k δέντρα απόφασης, τα οποία καταλήγουν στην προτίμηση-ψήφο μιας κλάσης. Η κλάση που θα συγκεντρώσει στις περισσότερες προτιμήσεις-ψήφους είναι εκείνη που προκύπτει από τον αλγόριθμου του τυχαίου δάσους.

Στην πραγματικότητα η ανάπτυξη των δέντρων απόφασης παίζει σημαντικό ρόλο στη λειτουργία των τυχαίων δασών. Αν κάθε δέντρο απόφασης είναι ένας καλός ταξινομητής, δηλαδή έχουν μικρό ρυθμό εμφάνισης λάθος ταξινομήσεων, τότε όσο περισσότερα τέτοια δέντρα διαθέτει ένα τυχαίο δάσος τόσο μικρότερη είναι η πιθανότητα για λάθος ταξινόμηση.



Εικόνα 8. Γραφική αναπαράσταση ενός τυχαίου δάσους [47].

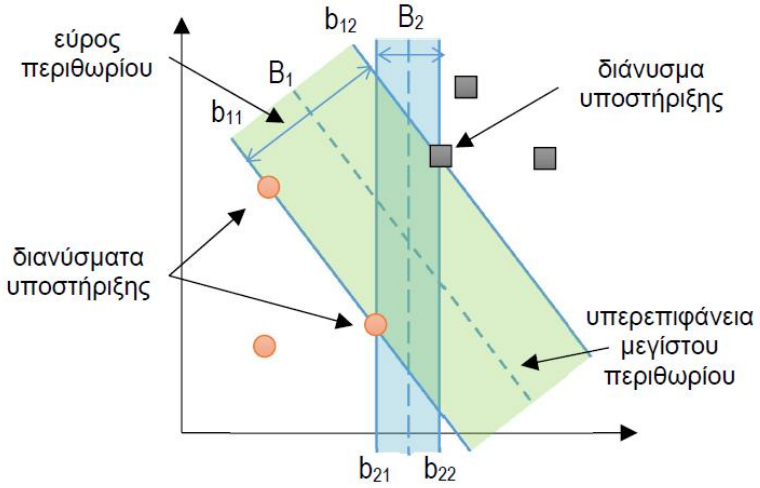
3.1.4. Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)

Οι μηχανές διανυσμάτων υποστήριξης προτάθηκαν το 1963 ως γραμμικοί ταξινομητές από τον Vladimir Vapnik [15]. Ωστόσο έγιναν ιδιαίτερα γνωστοί μετά το

1992, όταν ενισχύθηκαν με το κόλπο του πυρήνα (kernel trick), με το οποίο έγινε εφικτή η εφαρμογή τους και σε μη γραμμικώς διαχωρίσιμα προβλήματα [15].

Για να γίνει αντιληπτή η βασική ιδέα των μηχανών διανυσμάτων υποστήριξης, δίνεται ένα παράδειγμα ταξινόμησης δύο κλάσεων (Εικόνα 9). Σε μια τέτοια περίπτωση, μπορεί να μην υπάρχει μοναδική γραμμή (σύνορο) που να χωρίζει τα δεδομένα στις δύο κλάσεις. Όπως φαίνεται στην Εικόνα 9, υπάρχουν δύο τέτοιες ευθείες B_1 και B_2. Σκοπός των μηχανών διανυσμάτων υποστήριξης είναι να βρουν την ευθεία εκείνη που απέχει όσο το δυνατόν περισσότερο από τα παραδείγματα των κλάσεων. Η ευθεία αυτή ονομάζεται σύνορο μέγιστου περιθωρίου και σε γραμμικώς διαχωρίσιμα προβλήματα ορίζεται από έναν πεπερασμένο αριθμό παραδειγμάτων του συνόλου εκπαίδευσης που ονομάζονται διανύσματα υποστήριξης (support vectors) [15].

Στις περιπτώσεις που υπάρχουν μη γραμμικώς διαχωρίσιμα προβλήματα (δηλαδή τα χαρακτηριστικά των παρατηρήσεων του συνόλου δεν είναι γραμμικά διαχωρίσιμα), οι μηχανές διανυσμάτων υποστήριξης μπορούν να μετασχηματίσουν τον αρχικό χώρο υποθέσεων, με τη βοήθεια των συναρτήσεων πυρήνα, έτσι ώστε να μετατραπούν σε προβλήματα γραμμικώς διαχωρίσιμα. Στην πραγματικότητα, γίνεται προβολή των παρατηρήσεων σε ένα διανυσματικό χώρο περισσότερων διαστάσεων που είναι γραμμικά διαχωρίσιμος. Η έννοια της ευθείας, σε μη δισδιάστατους χώρους αντικαθίσταται από εκείνη της υπερεπιφάνειας [15].



Εικόνα 9. Γραφική αναπαράσταση δύο συνόρων απόφασης (B_1 και B_2) σε πρόβλημα ταξινόμησης δύο τάξεων [15].

3.1.5. Κ-Πλησιέστεροι Γείτονες (K-Nearest Neighbors)

Ένας μη παραμετρικός (δεν υποθέτει κάποια κατανομή για τα δεδομένα) και μη γραμμικός ταξινομητής είναι ο ταξινομητής Κ-Πλησιέστερων Γειτόνων, ο οποίος στηρίζεται στην έννοια της εγγύτητας. Οι παρατηρήσεις του χώρου των χαρακτηριστικών ταξινομούνται στην κλάση που είναι η πιο κοινή μεταξύ των k πλησιέστερων παρατηρήσεων εκπαίδευσης. Με δεδομένο τα διανύσματα εκπαίδευσης στο χώρο χαρακτηριστικών, κάθε νέα παρατήρηση, καταχωρείται στην κλάση που πλειοψηφεί μεταξύ των k πλησιέστερων βάσει κάποιας μετρικής.

Μια βελτίωση είναι να μην λαμβάνονται ισότιμα οι k πλησιέστεροι γείτονες αλλά να συνεισφέρουν περισσότερο τα σημεία που είναι εγγύτερα στη νέα παρατήρηση, μέσω συντελεστών βαρύτητας, οι οποίοι θα είναι ίσοι με 1/d, όπου d απόσταση του εκάστοτε σημείου από τη νέα παρατήρηση.

Στο σημείο αυτό, κρίνεται αναγκαίο να αναφερθεί ο ορισμός της μετρικής καθώς και να αναφερθούν οι πιο δημοφιλείς αναφορικά με τον ταξινομητή Κ-Πλησιέστερων Γειτόνων.

Έστω X ένα μη κενό σύνολο. Μετρική στο X λέγεται κάθε συνάρτηση $\rho : \mathrm{X} \times \mathrm{X} \to \mathbb{R}$ με τις παρακάτω ιδιότητες:

(i) $\rho(\mathrm{x},\mathrm{y}) \geq 0$ για κάθε $\mathrm{x},\mathrm{y} \in \mathrm{X}$ και $\rho(\mathrm{x},\mathrm{y}) = 0$ αν και μόνο αν $\mathrm{x} = \mathrm{y}$ (η ρ είναι μη αρνητική).

(ii) $\rho(\mathrm{x},\mathrm{y}) = \rho(\mathrm{y},\mathrm{x})$ για κάθε $\mathrm{x},\mathrm{y} \in \mathrm{X}$ (συμμετρική ιδιότητα).

(iii) $\rho(\mathrm{x},\mathrm{z}) \leq \rho(\mathrm{x},\mathrm{y}) + \rho(\mathrm{y},\mathrm{z})$ για κάθε $\mathrm{x},\mathrm{y},\mathrm{z} \in \mathrm{X}$ (τριγωνική ανισότητα).

Οι πιο δημοφιλείς μετρικές είναι η Ευκλείδεια απόσταση, η απόσταση Minkowski, η απόσταση Mahalanobis και η απόσταση Manhattan. Έστω $\mathrm{x} = (\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_\mathrm{m})$ και $\mathrm{y} = (\mathrm{y}_1, \mathrm{y}_2, ..., \mathrm{y}_\mathrm{m}) \in \mathbb{R}^\mathrm{m}$, τότε ορίζονται οι παρακάτω αποστάσεις:

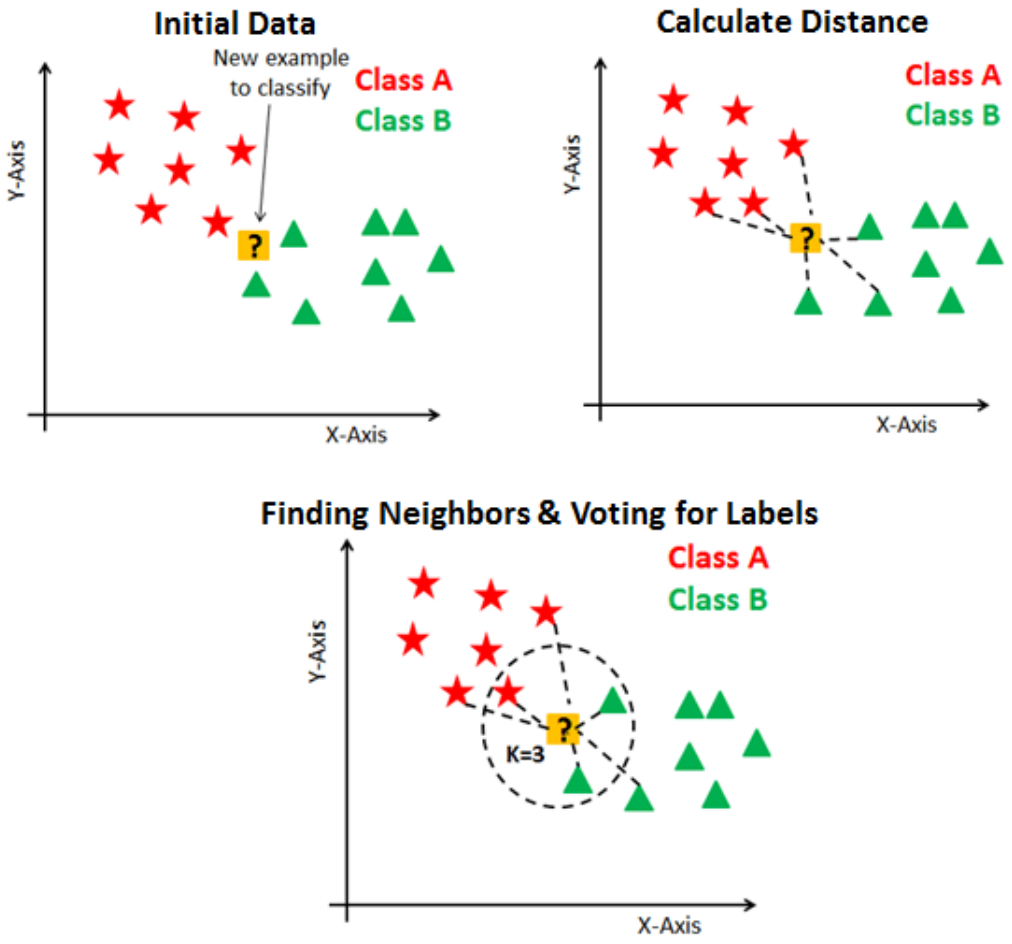
α) Απόσταση Minkowski: $\mathrm{D}(\mathrm{x},\mathrm{y}) = \mathrm{L}_\mathrm{k}(\mathrm{x},\mathrm{y}) = \left(\sum_{\mathrm{i}=1}^{\mathrm{m}} |\mathrm{x}_\mathrm{i} - \mathrm{y}_\mathrm{i}|^\mathrm{k} \right)^{1/\mathrm{k}}$

Ειδικές περιπτώσεις της απόστασης Minkowski είναι η Ευκλείδεια απόσταση (k = 2) και η απόσταση Manhattan (k = 1).

β) Ευκλείδεια απόσταση: $D(x,y) = L_2(x,y) = \left(\sum_{i=1}^{m} (x_i - y_i)^2 \right)^{1/2}$

γ) Απόσταση Manhattan: $D(x,y) = L_1(x,y) = \left(\sum_{i=1}^{m} |x_i - y_i|^1 \right)^{1/1} = \sum_{i=1}^{m} (x_i - y_i)$

δ) Απόσταση Mahalanobis: $D(x,y,S) = \left[(x-y)^T S^{-1} (x-y) \right]^{1/2}$, όπου S είναι ο πίνακας διακύμανσης-συνδιακύμανσης (variance-covariance matrix), που είναι τετραγωνικός $(N \times N)$. Στην ειδική περίπτωση όπου $S = I$, προκύπτει η Ευκλείδεια απόσταση.



Εικόνα 10. Σχηματική αναπαράσταση του ταξινομητή Κ-Πλησιέστερων Γειτόνων [48].

3.2 Αξιολόγηση ταξινομητή

3.2.1 Πίνακας σύγχυσης

Η επίδοση ενός ταξινομητή μπορεί εύκολα να αναπαρασταθεί με τη βοήθεια του Πίνακα Σύγχυση (confusion Matrix). Πρόκειται για έναν πίνακα $M \times M$ διαστάσεων, όπου το στοιχείο $a_{i,j}$ ισούται με το πλήθος των περιπτώσεων που, ενώ ανήκουν στην κλάση i, ταξινομούνται στην κλάση j.

$$\mathrm{A} = \begin{pmatrix} a_{1,1} & \cdots & a_{1,M} \\ \vdots & \ddots & \vdots \\ a_{M,1} & \cdots & a_{M,M} \end{pmatrix}$$

Συνεπώς, τα διαγώνια στοιχεία του πίνακα αντιστοιχούν στις περιπτώσεις που έχουν ταξινομηθεί σωστά στην κλάση i, με $1 \le i \le M$ [49].

Η πιο απλή περίπτωση είναι όταν υπάρχουν δύο κλάσεις. Έτσι, ο πίνακας Α είναι 2×2 διαστάσεων, που ως γραμμές έχει τις πραγματικές τιμές, ενώ ως στήλες τις προβλέψεις [50, 51]. Ο πίνακας Α μπορεί να πάρει την ακόλουθη μορφή [52]:

		Πρόβλεψη κλάσης		
		Θετική	Αρνητική	Σύνολο
Πραγματική κλάση	Θετική	$a_{1,1}$ **TP**	$a_{1,2}$ **FN**	**Positive**
	Αρνητική	$a_{2,1}$ **FP**	$a_{2,2}$ **TN**	**Negative**

όπου:

TP: όσα παραδείγματα ανήκουν στην κλάση 1 και ταξινομήθηκαν στην 1

FN: όσα παραδείγματα ανήκουν στην κλάση 1, αλλά ταξινομήθηκαν στην 2

FP: όσα παραδείγματα ανήκουν στην κλάση 2, αλλά ταξινομήθηκαν στην 1

TN: όσα παραδείγματα ανήκουν στην κλάση 2 και ταξινομήθηκαν στην 2

3.2.2 Μετρικές αξιολόγησης

α) Ορθότητα (accuracy): Ένα από τα πιο συχνά μέτρα αξιολόγησης ενός ταξινομητή είναι η ορθότητα, που εκφράζει το συνολικό ποσοστό των ορθών ταξινομήσεων. η οποία υπολογίζεται βάσει του ακόλουθου τύπου [53-55]:

$$\text{Ορθότητα} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

Πρόκειται για μια απλοϊκή μετρική που δύναται να δώσει καλή εικόνα για τον ταξινομητή μόνο στην περίπτωση κλάσεων με ισάριθμα στοιχεία (balanced data), ενώ μπορεί να οδηγήσει σε εσφαλμένα συμπεράσματα όταν τα δεδομένα είναι μη ισορροπημένα (imbalanced data) [52], δηλαδή σε δεδομένα που υπάρχει διαφορά στον αριθμό των αρνητικών και θετικών περιπτώσεων, με τις αρνητικές, τις περισσότερες, φορές να υπερτερούν [56].

Την αδυναμία της «ορθότητας», καλύπτουν άλλες μετρικές, οι οποίες παρέχουν μια πιο πλήρη και αντικειμενική εικόνα για τον ταξινομητή.

β) Ακρίβεια (precision) ή θετική προγνωστική αξία (positive predictive value – PPV): Η ακρίβεια απαντάει στο ερώτημα: *«Δοθέντος ενός δείγματος που ταξινομήθηκε στην κλάση i, πόσο πιθανό είναι η ταξινόμηση αυτή να είναι σωστή;»*. Η ακρίβεια υπολογίζεται βάσει του ακόλουθου τύπου [53-55, 57]:

$$\text{Ακρίβεια} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

γ) Ανάκληση (recall) ή ευαισθησία (sensitivity) ή ποσοστό αληθώς θετικών (true positive rate): Η ανάκληση απαντάει στο ερώτημα: *«Δοθέντος ενός δείγματος που προέρχεται από την κλάση i, πόσο πιθανό είναι να ταξινομηθεί σωστά;»*. Η ανάκληση υπολογίζεται βάσει του ακόλουθου τύπου [55, 57]:

$$\text{Ανάκληση} = \text{Ευαισθησία} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

δ) Ειδικότητα (specificity) ή ποσοστό αληθώς αρνητικών (true negative rate): Η ειδικότητα απαντάει στο ερώτημα: *«Δοθέντος ενός δείγματος που προέρχεται από την κλάση i, πόσο πιθανό είναι να ταξινομηθεί λανθασμένα;»*. Η ειδικότητα υπολογίζεται από τον ακόλουθο τύπο [53-55, 57]:

$$\text{Ειδικότητα} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

ε) F-measure (F1-score): Είναι ο αρμονικός μέσος της ακρίβειας και της ανάκλησης και παρέχει μια συνολική εκτίμηση του ταξινομητή. Η μετρική F-measure δίνεται από τον ακόλουθο τύπο [53, 55]:

$$\text{F1} - \text{score} = \frac{2}{\frac{1}{\text{Ακρίβεια}} + \frac{1}{\text{Ανάκληση}}} = 2 \cdot \frac{\text{Ακρίβεια} \cdot \text{Ανάκληση}}{\text{Ακρίβεια} + \text{Ανάκληση}}$$

στ) Αρνητική προγνωστική αξία (negative predictive value – NPV): Η ακρίβεια απαντάει στο ερώτημα: *«Δοθέντος ενός δείγματος που ταξινομήθηκε στην κλάση i, πόσο πιθανό είναι η ταξινόμηση αυτή να είναι λανθασμένη;»*. Η αρνητική προγνωστική αξία υπολογίζεται βάσει του ακόλουθου τύπου [53-55]:

$$\text{Αρνητική προγνωστική αξία} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

ζ) Ισορροπημένη ορθότητα (balanced accuracy): Η ισορροπημένη ορθότητα είναι μια μετρική αξιολόγησης που χρησιμοποιείται κυρίως σε ανομοιογενή δεδομένα και υπολογίζεται από τον τύπο [55]:

$$\text{Ισορροπημένη ορθότητα} = \frac{\text{sensitivity} + \text{specificity}}{2}$$

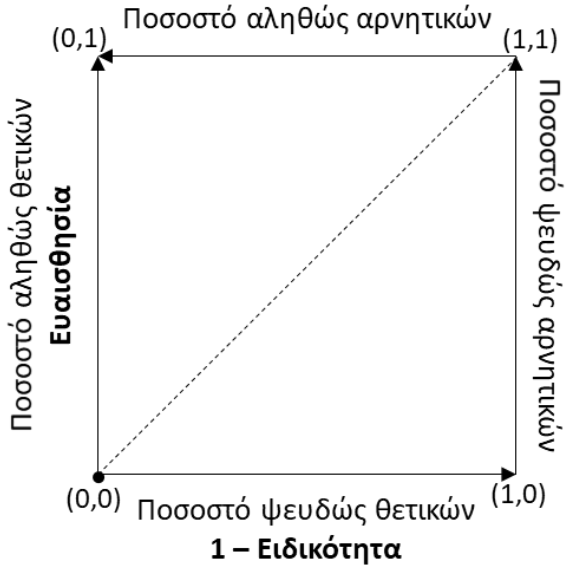
3.2.3 Καμπύλη λειτουργικού χαρακτηριστικού δέκτη

Η καμπύλη λειτουργικού χαρακτηριστικού δέκτη (Receiver Operating Characteristic Curve – ROC) αναπαριστά τη σχέση του ποσοστού των αληθώς θετικών (TP) και των ψευδώς θετικών περιπτώσεων για όλες τις πιθανές οριακές τιμές του διαχωρισμού.

Η γραφική παράσταση της καμπύλης βρίσκεται σε ένα τετράγωνο του το οποίο έχει πλευρά μήκους 1 (Σχήμα 1).

Η διαγώνιος του τετραγώνου που ξεκινάει από το σημείο $(0,0)$ και καταλήγει στο σημείο $(1,1)$ εκφράζει έναν ταξινομητή που προβλέπει τυχαία την κλάση. Όσοι ταξινομητές βρίσκονται πάνω από την διαγώνιο του τετραγώνου είναι καλύτεροι σε σχέση με την τυχαία πρόβλεψη. Όσο πιο μετατοπισμένο βρίσκεται κάποιο σημείο της καμπύλης προς την πάνω αριστερή γωνία του τετραγώνου, δηλαδή στο σημείο $(0,1)$, τόσο πιο καλός είναι ο ταξινομητής.

Το εμβαδόν κάτω από την καμπύλη, χρησιμοποιείται ως δείκτης διαχωρισμού των κατανομών εκείνων που εμφάνισαν το συμβάν και εκείνων που δεν το εμφάνισαν. Η δυνατή τιμή του εμβαδού κάτω από την καμπύλη λειτουργικών χαρακτηριστικών, μπορεί να πάρει τιμές από 0,5 (όταν πρόκειται για τυχαίο ταξινομητή) έως 1 (τιμή που υποδηλώνει την ύπαρξη ενός άριστου ταξινομητή καθώς οι δύο κατανομές δεν συμπίπτουν πουθενά). Για τη σύγκριση δύο ή περισσοτέρων ταξινομητών χρησιμοποιείται το μέτρο «Περιοχή κάτω από την καμπύλη ROC» (Area under ROC Curve – AUC) [58].



Εικόνα 11. Τετράγωνο που απεικονίζεται η καμπύλη λειτουργικών χαρακτηριστικών.

4. Ανομοιογενή δεδομένα

Με τον όρο ανομοιογενή ή μη ισορροπημένα δεδομένα (imbalanced data) εννοούμε τα δεδομένα εκείνα τα οποία κατανέμονται δυσανάλογα στις κλάσεις. Έτσι, ο αριθμός των περιπτώσεων σε μια κλάση είναι πολύ μικρότερος από ότι είναι στην άλλη κλάση. Στην περίπτωση των προβλημάτων ταξινόμησης με δύο κλάσεις, μια κλάση είναι εκείνη με τα περισσότερες περιπτώσεις (majority class), ενώ η δεύτερη κλάση περιέχει λιγότερα περιπτώσεις (minority class) [59]. Ιδιαίτερη σημασία παρουσιάζει η ταξινόμηση των θετικών περιπτώσεων επειδή το κόστος μιας λάθος ταξινόμησης μπορεί να είναι ιδιαίτερα μεγάλο [60, 61]. Η ύπαρξη ανομοιογενών δεδομένων συναντάται σε διάφορα επιστημονικά πεδία και σε αρκετές περιπτώσεις. Για παράδειγμα, από το χώρο της Ιατρικής ανομοιογενή δεδομένα έχουμε στους ασθενείς με καρκίνο (ίαση έναντι όχι ίαση), στον πληθυσμό των ΧΕΝ η ύπαρξη HIV λοίμωξης (ναι έναντι όχι), από το χώρο των χρηματοπιστωτικών ιδρυμάτων η υποκλοπή κωδικών χρεωστικών καρτών (ναι έναντι όχι) και πολλά άλλα.

Συνήθως, οι αλγόριθμοι ταξινόμησης τείνουν να έχουν πολύ υψηλή ακρίβεια στην κλάση εκείνη με τις περισσότερες περιπτώσεις, σε αντίθεση με την χαμηλή ακρίβεια που πετυχαίνουν στην κλάση με τις λιγοστές περιπτώσεις, η οποία συνήθως είναι και η κλάση ενδιαφέροντος μας [62]. Ο λόγος που συμβαίνει αυτό είναι ότι οι αλγόριθμοι ταξινόμησης, στοχεύουν στην υψηλή ακρίβεια, στο σύνολο των δεδομένων και όχι στις επιμέρους κλάσεις [62, 63].

Τα τελευταία χρόνια έχει δοθεί ιδιαίτερη προσοχή στο πρόβλημα της ανισορροπίας μεταξύ των κλάσεων στο χώρο της Μηχανικής Μάθησης [64, 65] και έχουν προταθεί διάφορες μέθοδοι που εφαρμόζονται είτε πάνω στα δεδομένα είτε στους αλγορίθμους, προκειμένου να επιτευχθεί υψηλή ακρίβεια στη μειονοτική τάξη χωρίς όμως να θυσιάζεται η ακρίβεια της πλειοψηφικής κλάσης [62, 63, 66]. Στη βιβλιογραφία, οι διαθέσιμες μέθοδοι για την αντιμετώπιση του προβλήματος της ανισορροπίας μεταξύ των κλάσεων (class imbalance problem) μπορούν να ομαδοποιηθούν σε τρεις κατηγορίες [65]:

1. Επαναδειγματοληψία (resampling)

2. Μάθηση ευαισθησίας-κόστους (cost-sensitive learning)

3. Μάθηση συνόλου (ensemble learning)

4.1 Επαναδειγματοληψία

Παρακάτω δίνεται μια σύντομη περιγραφή των πιο διαδεδομένων μεθόδων επαναδειγματοληψίας. Οι λόγοι που δίνεται ιδιαίτερη έμφαση στην κατηγορία αυτή των μεθόδων για την αντιμετώπιση του προβλήματος της ανισορροπίας μεταξύ των κλάσεων ποικίλουν [67]. Μεταξύ άλλων είναι: α) η εύκολη εφαρμογή τους σε πραγματικά προβλήματα, β) εμπειρικά έχει φανεί πως τα αποτελέσματα από τη χρήση επαναδειγματοληψίας είναι ανταγωνιστικά ή ακόμη και πανομοιότυπα με εκείνα που προκύπτουν από τη μάθηση ευαισθησίας-κόστους και γ) με τη σωστή επαναδειγματοληψία οποιοσδήποτε αλγόριθμος ταξινόμησης μπορεί να γίνει ευαίσθητος στο κόστος χωρίς να αλλάξει η εσωτερική λειτουργία του ίδιου ταξινομητή [67].

4.2 Μέθοδοι Επαναδειγματοληψίας

4.2.1 Τυχαία υποδειγματοληψία (Random undersampling)

Η μέθοδος της τυχαίας υπερδειγματοληψίας εφαρμόζεται στην κλάση πλειοψηφίας, δηλαδή για εκείνες που υπερέχουν από αριθμό παραδειγμάτων. Στην πραγματικότητα αφαιρούνται παραδείγματα από την κλάση πλειοψηφίας, με τυχαίο τρόπο, μέχρις ότου να έχει το ίδιο πλήθος παραδειγμάτων με την κλάση μειοψηφίας (Εικόνα 12α). Το μειονέκτημα της μεθόδου αυτής είναι η τυχαία απόρριψη πληροφορίας από την πλειοψηφική κλάση μπορεί να επηρεάσει την επίδοση του ταξινομητή.

4.2.2 Τυχαία υπερδειγματοληψία (Random oversampling)

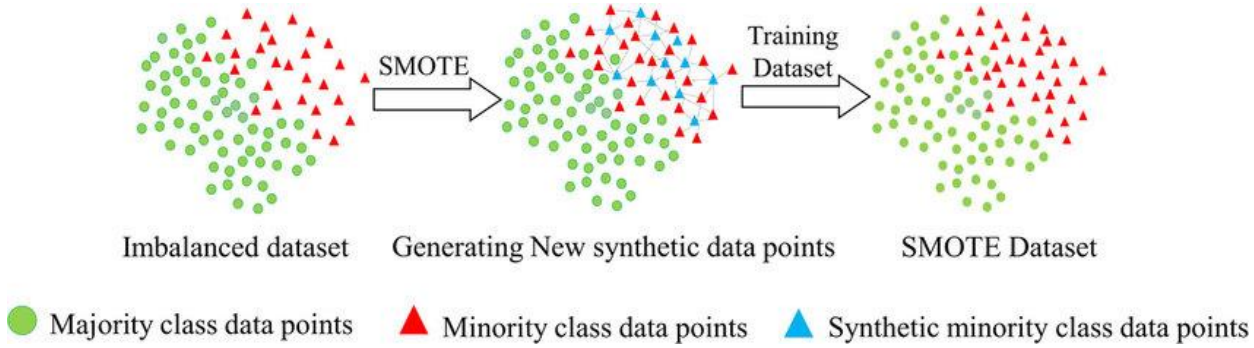
Η μέθοδος της τυχαίας υπερδειγματοληψίας εφαρμόζεται στην κλάση μειοψηφίας, δηλαδή με τα λιγότερα παραδείγματα. Στην πραγματικότητα προστίθενται επιπλέον δεδομένα εκπαίδευσης στην κλάση μειοψηφίας μέχρις ότου να έχει το ίδιο πλήθος παραδειγμάτων με την κλάση πλειοψηφίας (Εικόνα 12β). Η μέθοδος αυτή έχει το πλεονέκτημα πως δεν υπάρχει απώλεια πληροφορίας. Από την άλλη πλευρά, ασκείται κριτική στο ότι δεν προσθέτει πραγματικά δεδομένα στο σύνολο εκπαίδευσης [67], το οποίο μπορεί να αυξήσει την πιθανότητα υπερπροσαρμογής (overfitting) καθώς δημιουργεί ακριβή αντίγραφα των παραδειγμάτων της μειοψηφικής κλάσης [68].

Εικόνα 12. Σχηματική αναπαράσταση της μεθόδου α) υποδειγματοληψίας και β) υπερδειγματοληψίας [69].

4.2.3 Τεχνική υπερδειγματοληψίας συνθετικής μειονότητας (Synthetic minority oversampling technique - SMOTE)

Η τεχνική υπερδειγματοληψίας συνθετικής μειονότητας αντιμετωπίζει το ενδεχόμενο της υπερπροσαρμογής που μπορεί να προκληθεί από την τυχαία υπερδειγματοληψία. Αυτό επιτυγχάνεται με την προσθήκη νέων συνθετικών μειονοτικών παραδειγμάτων. Προκειμένου να δημιουργηθεί ένα νέο συνθετικό παράδειγμα η μέθοδος SMOTE βρίσκει τους k-πλησιέστερους γείτονες για κάθε περίπτωση της κλάσης μειοψηφίας, επιλέγει εκείνον τον πλησιέστερο από τους k, και στη συνέχεια πολλαπλασιάζει τη διαφορά των διανυσμάτων του κοντινότερου γείτονα και των χαρακτηριστικών με έναν τυχαίο αριθμό m με $\mathrm{m} \in (0,1)$ και προστίθεται στο διάνυσμα των χαρακτηριστικών του δείγματος. Το διάνυσμα που

προκύπτει είναι το νέο συνθετικό παράδειγμα [70]. Με αυτόν τον τρόπο, επιτυγχάνεται τόσο η αύξηση της μειονοτικής κλάσης όσο και η αύξηση της ποικιλίας των παραδειγμάτων.



Εικόνα 13. Σχηματική αναπαράστασης της υπερδειγματοληψίας με την τεχνική της συνθετικής μειονότητας.

4.2.4 Προσαρμοστική συνθετική μέθοδος δειγματοληψίας για μη ισορροπημένα δεδομένα (Adaptive Synthetic Sampling Method for Imbalanced Data - ADASYN)

Το 2008 παρουσιάστηκε μια νέα τεχνική υπερδειγματοληψίας που παράγει συνθετικά παραδείγματα [71]. Η βασική ιδέα της Προσαρμοστικής συνθετικής μεθόδου δειγματοληψίας για μη ισορροπημένα δεδομένα (Adaptive Synthetic Sampling Method for Imbalanced Data – ADASYN) έγκειται στη χρήση της κατανομής πιθανότητας των περιπτώσεων μειοψηφίας με σκοπό τον υπολογισμό του αριθμού των συνθετικών παραδειγμάτων που πρέπει να δημιουργηθούν για κάθε περίπτωση της κλάσης μειοψηφίας. Η μέθοδος ADASYN δημιουργεί μία νέα περίπτωση ακολουθώντας τα παρακάτω βήματα:

Βήμα 1: Υπολογίζει τον λόγο $d = \frac{m_s}{m_l}$, $d \in (0,1]$, όπου m_s και m_l ο αριθμός των παραδειγμάτων της κλάσης μειοψηφίας και της κλάσης πλειοψηφίας, αντίστοιχα.

Βήμα 2: Υπολογίζει το συνολικού αριθμού των συνθετικών παραδειγμάτων της κλάσης μειοψηφίας που θα δημιουργηθούν. Ο συνολικός αριθμός προκύπτει από τον

τύπο $G = (m_s - m_l) \times \beta$, όπου β είναι η επιθυμητή αναλογία που θέλουμε να υπάρχει μετά την υλοποίηση του ADASYN. Στην περίπτωση όπου $\beta = 1$, τότε θα υπάρχει τέλεια ισορροπία μεταξύ των δύο κλάσεων.

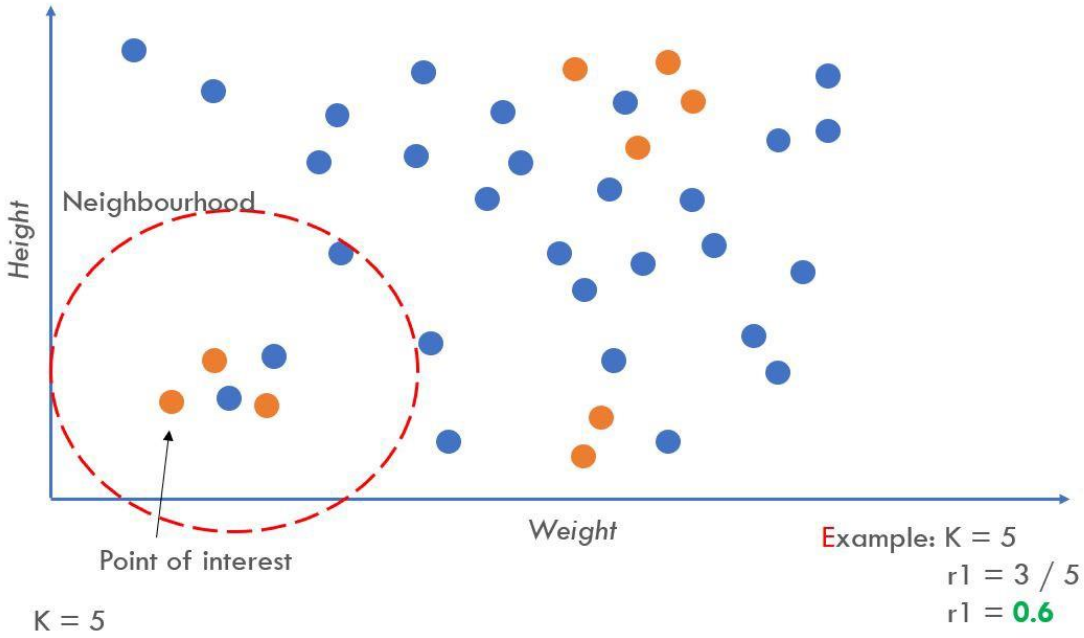
Βήμα 3: Βρίσκει τους k-πλησιέστερους γείτονες για κάθε ένα παράδειγμα της κλάσης μειοψηφίας και υπολογίζει την τιμή του r_i, όπου $r_i = \frac{\#\text{majority}}{k}$. Η τιμή του r_i υποδηλώνει την κυριαρχία της πλειοψηφικής τάξης σε κάθε συγκεκριμένη γειτονιά. Όσο πιο μεγάλη είναι η τιμή του r_i, τόσο η γειτονιά περιέχει περισσότερα παραδείγματα από την κλάση πλειοψηφίας και τόσο πιο δύσκολο είναι να μάθει (Εικόνα 14).

Βήμα 4: Κανονικοποιεί τις τιμές του r_i σύμφωνα με τον τύπο: $\hat{r}_i = \frac{r_i}{\sum_{i=1}^{m_s} r_i}$ έτσι ώστε το άθροισμα όλων των τιμών r_i να είναι ίσο με 1.

Βήμα 5: Υπολογίζει το πλήθος των συνθετικών παραδειγμάτων που χρειάζεται να δημιουργηθούν για κάθε παράδειγμα x_i της κλάσης μειοψηφίας: $g_i = \hat{r}_i \times G$, όπου G είναι ο συνολικός αριθμός των συνθετικών παραδειγμάτων της κλάσης μειοψηφίας που χρειάζεται να δημιουργηθούν για την κλάση (όπως ορίστηκε στο Βήμα 2).

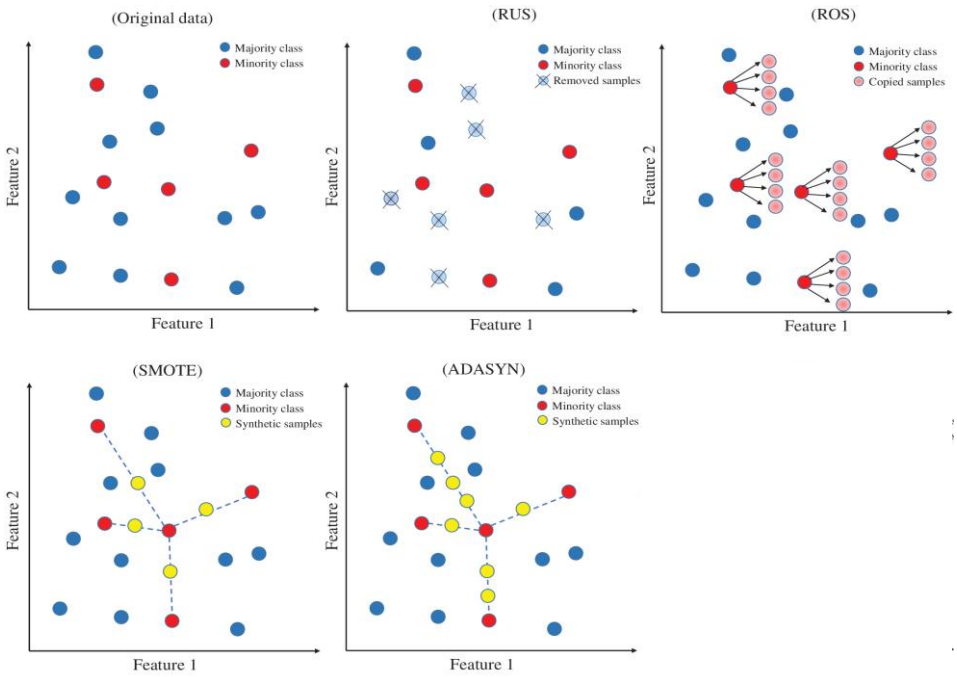
Βήμα 6: Δημιουργεί δεδομένα g_i για κάθε γειτονιά. Πρώτα, παίρνει ένα παράδειγμα από τη γειτονιά της κλάσης μειοψηφίας, x_i. Στη συνέχεια, επιλέγει τυχαία ένα άλλο παράδειγμα της κλάσης μειοψηφίας από τη ίδια γειτονιά, x_{zi}. Το νέο συνθετικό παράδειγμα μπορεί να υπολογιστεί χρησιμοποιώντας: $s_i = x_i + (x_{zi} - x_i)\lambda$.

Η τεχνική αυτή βελτιώνει τη μάθηση με δύο τρόπους: 1) μειώνοντας τη μεροληψία από τα μη ισορροπημένα δεδομένα και 2) μετατοπίζοντας το όριο απόφασης της ταξινόμησης προς τη μεριά των δύσκολων παραδειγμάτων. Η αποτελεσματικότητα της μεθόδου έχει επιβεβαιωθεί και από τη χρήση προσομοιώσεων [71].



Εικόνα 14. Γραφική αναπαράσταση της τεχνικής ADASYN για k = 5 [72].

Στην Εικόνα 15, παρουσιάζονται οι τεχνικές επαναδειγματοληψίας που αναλύθηκαν στην ενότητα αυτή.



Εικόνα 15. Γραφική αναπαράσταση των τεχνικών επαναδειγματοληψίας [73]

5. Προεπεξεργασία δεδομένων

Η προεπεξεργασία των δεδομένων είναι η διαδικασία κατά την οποία μετατρέπονται τα αδρά δεδομένα, που μπορεί να περιέχουν ελλείπουσες τιμές, ασυμφωνίες, θόρυβο, σε τέτοια μορφή, ώστε να είναι επεξεργάσιμα και συνεπή [74]. Επομένως, η προεπεξεργασία των δεδομένων τις περισσότερες φορές είναι αναγκαία προκειμένου να διασφαλιστεί η εγκυρότητα και η αξιοπιστία των αποτελεσμάτων της ανάλυσης των δεδομένων [75, 76]. Τα βασικά βήματα της προεπεξεργασίας είναι:

1) ο καθαρισμός δεδομένων (data cleaning),
2) η ενοποίηση δεδομένων (data integration),
3) ο μετασχηματισμός και η διακριτοποίηση των δεδομένων (data transformation και data discretization, αντίστοιχα) και
4) η μείωση διαστάσεων (dimensionality reduction) ή η μείωση δεδομένων (data reduction).

5.1 Καθαρισμός δεδομένων

Με τον όρο καθαρισμός δεδομένων εννοούμε όλες εκείνες τις διαδικασίες που απαιτούνται προκειμένου να διαχειριστούμε τις ελλείπουσες τιμές, τον θόρυβο και τις ασυμφωνίες που υπάρχουν στα δεδομένα. Οι ελλείπουσες τιμές μπορεί να επηρεάσουν τα αποτελέσματα της εκπαίδευσης ενός αλγορίθμου βάσει της κατηγορίας στην οποία ανήκουν. Οι ελλείπουσες τιμές κατηγοριοποιούνται σε μια από τις τρεις ακόλουθες κατηγορίες [77]:

α) Τελείως τυχαία (Completely at random): Η πιθανότητα να λείπει κάποια τιμή δεν εξαρτάται από την προηγούμενη τιμή που έχει καταγραφεί ούτε από το ποια ήταν τελικά η τιμή αυτή.

β) Σχεδόν τυχαία (At random): Η πιθανότητα να λείπει κάποια τιμή εξαρτάται από το ποιες τιμές υπήρχαν μέχρι εκείνη τη στιγμή αλλά δεν εξαρτάται από το ποια τελικά είναι η τιμή αυτή.

γ) Όχι τυχαία (Not at random): Η πιθανότητα να λείπει κάποια τιμή εξαρτάται από το πόσο είναι η τιμή που δεν μετρήθηκε.

Στη βιβλιογραφία έχουν αναφερθεί πληθώρα τεχνικών για τη διαχείριση των ελλειπουσών τιμών, οι οποίες ομαδοποιούνται σε πέντε κατηγορίες [77]: 1) Μέθοδοι που αγνοούν τις ελλείπουσες παρατηρήσεις, 2) Μέθοδοι μεμονωμένου υπολογισμού (Single imputation methods), 3) Μέθοδοι άλλων υπολογισμών (Other imputation methods), 4) Μέθοδοι βάσει πιθανοφάνειας (Likelihood-based methods), 5) Μέθοδοι δεικτών (Indicator methods).

5.2 Ενοποίηση δεδομένων

Στις περιπτώσεις που το σύνολο των δεδομένων βρίσκεται σε διαφορετικές βάσεις, η ενοποίησή τους σε μια ενιαία, συνεκτική βάση κρίνεται απαραίτητη για την περαιτέρω επεξεργασία και ανάλυση των δεδομένων. Στα δεδομένα που θα προκύψουν μετά την ενοποίηση (μεταδεδομένα) δύναται να υπάρχουν πιθανές συγκρούσεις ή ασυνέπειες ή/και πλεονάζουσα πληροφορία. Η ορθή ενοποίηση των βάσεων μπορεί να οδηγήσει σε λιγότερο απαιτούμενο χρόνο εκπαίδευσης αλλά και σε πιο ποιοτικά αποτελέσματα [78].

5.3 Μετασχηματισμός δεδομένων

Η δημιουργία ενός συνόλου δεδομένων σπάνια γίνεται μόνο για προβλέψεις. Τις περισσότερες φορές τα δεδομένα δεν είναι στη σωστή μορφή ή απαιτούν κάποιους μετασχηματισμούς για να γίνουν πιο χρήσιμα [79]. Οι τεχνικές μετασχηματισμού των δεδομένων περιλαμβάνουν την κατηγορική κωδικοποίηση (categorical encoding), την κλιμάκωση χαρακτηριστικών (feature scaling) και την διακριτοποίηση (discretization).

5.3.1 Κατηγορική κωδικοποίηση

Συχνά τα σύνολα δεδομένων περιλαμβάνουν και κατηγορικά δεδομένα, δηλαδή τα δεδομένα τα οποία δεν είναι αριθμητικά. Οι αλγόριθμοι μηχανικής μάθησης δεν

δέχονται χαρακτηριστικά τα οποία δεν είναι αριθμητικά. Έτσι, είναι αναγκαία η μετατροπή των κατηγορικών δεδομένων σε αριθμητικά δεδομένα. Οι δύο πιο συνηθισμένες μέθοδοι για την μετατροπή αυτή είναι η Διατάξιμη Κωδικοποίηση (Ordinal Encoding) και η Κωδικοποίηση One-Hot (One-Hot Encoding).

Υπάρχουν μεταβλητές των οποίων οι κατηγορίες τους σχετίζονται με κάποιου είδους διάταξη. Ένα παράδειγμα μιας τέτοιας μεταβλητής μπορεί να είναι η οικονομική κατάσταση του ερωτώμενου, η οποία να έχει τρεις κατηγορίες: Κακή, Μέτρια και Καλή. Οι μεταβλητές αυτές ονομάζονται διατάξιμες (ordinal) και σε αυτές εφαρμόζουμε τη Διατάξιμη Κωδικοποίηση, δηλαδή στο παράδειγμα που αναφέρθηκε οι κατηγορίες θα ήταν 1, 2, 3 αντί για Κακή, Μέτρια και Καλή.

Για τις μεταβλητές εκείνες που οι κατηγορίες τους δεν υπαινίσσονται κάποιου είδους διάταξη ονομάζονται ονομαστικές (nominal). Χαρακτηριστικό παράδειγμα μιας ονομαστικής μεταβλητής είναι το φύλο του συμμετέχοντα σε μια έρευνα ή ο τόπος κατοικίας. Στις ονομαστικές μεταβλητές εφαρμόζουμε την κωδικοποίηση One-Hot, εφόσον οι κατηγορίες της μεταβλητής δεν είναι πολλές.

5.3.2 Κλιμάκωση χαρακτηριστικών

Στα περισσότερα σύνολα δεδομένων υπάρχουν χαρακτηριστικά, τα οποία έχουν διαφορετικό εύρος τιμών. Οι αλγόριθμοι μηχανικής μάθησης που χρησιμοποιούν τη μέθοδο “Επικλινής κάθοδος” (gradient descent) ως τεχνική βελτιστοποίησης επηρεάζονται περισσότερο από τα χαρακτηριστικά εκείνα με το μεγαλύτερο εύρος τιμών σε σχέση με εκείνα που έχουν μικρότερα εύρη τιμών. Στόχος λοιπόν της κλιμάκωσης χαρακτηριστικών είναι να διασφαλίσει ότι τα χαρακτηριστικά είναι σχεδόν στην ίδια κλίμακα, ώστε κάθε χαρακτηριστικό να είναι εξίσου σημαντικό και να διευκολύνεται η επεξεργασία τους από τους αλγορίθμους μηχανικής μάθησης. Για την επίτευξη αυτού του στόχου μπορεί να εφαρμοστούν διάφοροι μετασχηματισμοί στα δεδομένα. Οι πιο δημοφιλείς μετασχηματισμοί είναι: η τυποποίηση, η μέγιστη-ελάχιστη κανονικοποίηση και η κανονικοποίηση δεκαδικής κλίμακας.

Στην τυποποίηση των δεδομένων, όλα τα χαρακτηριστικά παίρνουν τιμές που έχουν μέση τιμή 0 και τυπική απόκλιση ίση με 1. Για να επιτευχθεί αυτό, εφαρμόζεται στα χαρακτηριστικά ο παρακάτω μετασχηματισμός:

$$X' = \frac{X - \mu}{\sigma},$$

όπου μ είναι η μέση τιμή των τιμών του χαρακτηριστικού και σ είναι η τυπική απόκλιση.

Στην μέγιστη-ελάχιστη κανονικοποίηση των δεδομένων, όλα τα χαρακτηριστικά παίρνουν τιμές μεταξύ του ελάχιστου και του μέγιστου. Για να επιτευχθεί αυτό, εφαρμόζεται στα χαρακτηριστικά ο παρακάτω μετασχηματισμός:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}(new\_max - new\_min) + new\_min,$$

όπου X_{min} και X_{max} είναι η ελάχιστη και μέγιστη τιμή του χαρακτηριστικού X και new_max, new_min είναι οι νέες επιθυμητές τιμές για το μέγιστο και το ελάχιστο, αντίστοιχα. Στην ειδική περίπτωση που επιθυμούμε οι τιμές της μεταβλητής να κυμαίνονται στο διάστημα [0, 1], δηλαδή για $new\_max = 1$ και $new\_min = 0$, ο παραπάνω μετασχηματισμός παίρνει τη μορφή [80]:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}.$$

Τέλος, η κανονικοποίηση δεκαδικής κλίμακας πραγματοποιεί υποδεκαπλασιασμό των τιμών αφού τις διαιρεί με μια δύναμη του 10. Η δύναμη του 10 υπολογίζεται έτσι ώστε η απόλυτη η απόλυτη τιμή του νέου μεγίστου να είναι μικρότερη της μονάδας [80]. Ο μετασχηματισμός για την κανονικοποίηση δεκαδικής κλίμακας δίνεται από τον ακόλουθο τύπο:

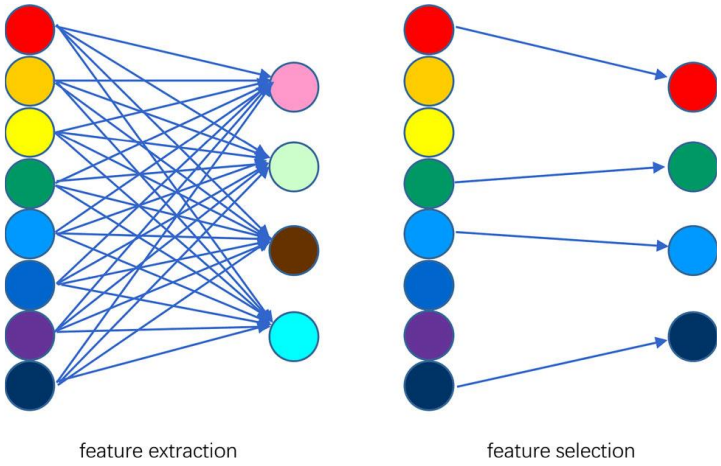
$$X' = \frac{X}{10^k}.$$

5.3.3 Διακριτοποίηση

Η διακριτοποίηση είναι μια τεχνική προεπεξεργασίας δεδομένων που μετατρέπει τα συνεχή χαρακτηριστικά σε διακριτά. Η εκπαίδευση ενός μοντέλου με διακριτά δεδομένα είναι ταχύτερη και πιο αποτελεσματική από ό,τι όταν επιχειρείται το ίδιο με συνεχή δεδομένα. Αν και τα συνεχή δεδομένα είναι πιο πληροφοριακά, σε επίπεδο μεγάλων δεδομένων μπορεί να επιβραδύνουν το μοντέλο.

5.4 Μείωση διαστάσεων

Με τον όρο διαστάσεις εννοούμε το πλήθος των στηλών που έχει το σύνολο των δεδομένων. Οι στήλες αναφέρονται στη βιβλιογραφία και ως διαστάσεις (dimensions), γνωρίσματα (attributes) και χαρακτηριστικά (features) [80]. Τα τελευταία χρόνια είναι συνηθισμένο να υπάρχουν μεγάλα σύνολα δεδομένων. Τα επιπλέον χαρακτηριστικά αντανακλούν σε επιπλέον πληροφορία. Ωστόσο, αυξάνοντας τον αριθμό των χαρακτηριστικών, εισάγεται θόρυβος και τις περισσότερες φορές υπάρχει πλεονασμός [81]. Αυτό συμβαίνει καθώς μπορεί να υπάρχουν αρκετές στήλες οι οποίες μπορεί να μην είναι χρήσιμες, δηλαδή να περιέχουν πληροφορία μη σχετική με την μεταβλητή ενδιαφέροντος, να υπάρχουν στήλες που είναι συσχετισμένες μεταξύ τους ή ακόμα να υπάρχουν στήλες με χαμηλή διακύμανση. Αναφορικά με τα δεδομένα μεγάλων διαστάσεων, η αύξηση της υπολογιστικής πολυπλοκότητας εκτός του ότι οδηγεί σε καθυστερήσεις του χρόνου εκπαίδευσης των μοντέλων, συνήθως οδηγεί και σε χαμηλά επίπεδα επίδοσης [80, 81]. Το πρόβλημα των μεγάλων διαστάσεων στη βιβλιογραφία αναφέρεται και ως «κατάρα διαστάσεων» [82, 83]. Οι τεχνικές για τη μείωση των διαστάσεων χωρίζονται σε δύο μεγάλες κατηγορίες και αφορούν στην επιλογή χαρακτηριστικών και στην εξαγωγή των χαρακτηριστικών (Εικόνα 16).



Εικόνα 16. Γραφική αναπαράσταση μεταξύ εξαγωγής και επιλογής χαρακτηριστικών [84].

5.4.1 Εξαγωγή χαρακτηριστικών (Feature extraction)

Σε αυτή την κατηγορία εντάσσονται όλες οι τεχνικές εκείνες που αποσκοπούν στη λήψη χαρακτηριστικών από τα υπάρχοντα δεδομένα. Οι τεχνικές αυτές μπορούν να διαχωριστούν σε μεθόδους Γραμμικές (Linear) και Μη γραμμικές (Non Linear). Μια από τις πιο διαδεδομένες γραμμικές μεθόδους είναι η Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis – PCA), η οποία ανακαλύφθηκε από τον Karl Pearson [85] το 1901, ενώ λίγα χρόνια αργότερα, το 1933, αναπτύχθηκε από τον Harold Hotelling [86].

Η ανάλυση κυρίων συνιστωσών είναι μια στατιστική διαδικασία μέσω της οποίας ένα σύνολο αρχικών μεταβλητών αναπαρίσταται από ένα διαφορετικό και συνήθως, μικρότερο σύνολο νέων μεταβλητών, οι οποίες προκύπτουν από τον γραμμικό συνδυασμό των αρχικών μεταβλητών. Οι νέες μεταβλητές ονομάζονται κύριες συνιστώσες. Η διαδικασία αυτή γίνεται με τέτοιο τρόπο, ώστε η πρώτη συνιστώσα να εξηγεί τη μέγιστη δυνατή διακύμανση των αρχικών μεταβλητών, η δεύτερη, μη συσχετιζόμενη με την πρώτη, να εξηγεί ένα σημαντικό μέρος αυτής αλλά πάντα μικρότερο από την πρώτη κοκ. [87]. Θεωρητικά, μπορούν να εξαχθούν τόσες συνιστώσες, όσες είναι και οι αρχικές μεταβλητές. Στην πράξη όμως, υπάρχουν κριτήρια που μας καθορίζουν τον αριθμό των συνιστωσών που θα πάρουμε. Με αυτόν τον τρόπο από ένα σύνολο συσχετισμένων μεταβλητών (αρχικές μεταβλητές)

καταλήγουμε σε ένα σύνολο ασυσχέτιστων μεταβλητών (κύριες συνιστώσες) που είναι ιδιαίτερα χρήσιμο σε διάφορες στατιστικές μεθόδους [88].

Τα κριτήρια για την επιλογή των k πρώτων κύριων συνιστωσών είναι διάφορα, ωστόσο τα επικρατέστερα είναι τα ακόλουθα τρία:

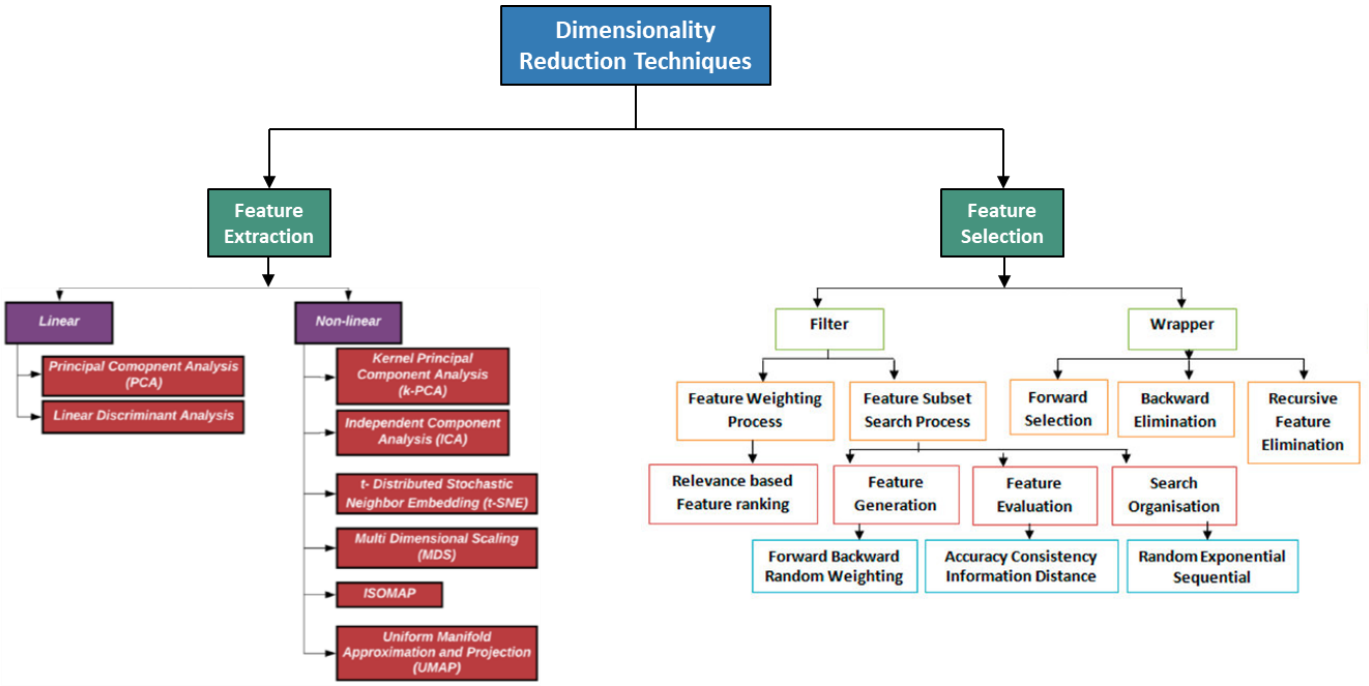
1. Ποσοστό συνολικής διακύμανσης που εξηγούν οι κύριες συνιστώσες.
2. Κριτήριο του Kaiser
3. Τεχνική του αγκώνα (scree plot).

Σύμφωνα με το πρώτο κριτήριο επιλέγουμε τόσες συνιστώσες, ώστε αθροιστικά να ερμηνεύεται ένα μεγάλο ποσοστό της διακύμανσης. Το κριτήριο Kaiser αναφέρεται στις ιδιοτιμές των κύριων συνιστωσών. Πιο συγκεκριμένα στην περίπτωση που χρησιμοποιηθεί ο πίνακας συνδιακύμανσης θα εξαχθούν τόσες συνιστώσες όσες η ιδιοτιμή τους είναι μεγαλύτερη από τη μέση τους τιμή, ενώ αν χρησιμοποιηθεί ο πίνακας συσχετίσεων, τόσες όσες έχουν ιδιοτιμή μεγαλύτερη της μονάδας. Το τρίτο κριτήριο αφορά μια γραφική μέθοδο μέσω της οποίας γίνεται η επιλογή του αριθμού των κύριων συνιστωσών. Στον άξονα των x είναι η σειρά των ιδιοτιμών, ενώ στον άξονα των y είναι οι τιμές των ιδιοτιμών. Επιλέγονται τόσες συνιστώσες μέχρι το γράφημα να αρχίσει να αλλάζει κλίση [88].

5.4.2 Επιλογή χαρακτηριστικών (Feature selection)

Σε αυτή την κατηγορία εντάσσονται όλες οι τεχνικές εκείνες που αποσκοπούν στην επιλογή ενός υποσυνόλου από το αρχικό σύνολο χαρακτηριστικών. Οι τεχνικές αυτές μπορούν να διαχωριστούν σε μεθόδους τύπου filter και μεθόδους τύπου wrapper [80].

- Οι μέθοδοι τύπου filter βασίζονται σε χαρακτηριστικά των δεδομένων και χρησιμοποιούν μεθόδους διαφορετικές από τους αλγορίθμους που θα εφαρμοστούν για την τελική εξόρυξη των προτύπων [80].
- Οι μέθοδοι τύπου wrapper χρησιμοποιούν τον ίδιο τον αλγόριθμο εξόρυξης για να αξιολογήσουν τα υποψήφια υποσύνολα χαρακτηριστικών [80].



Εικόνα 17. Ταξινόμηση τεχνικών μείωσης διαστάσεων [89, 90].

6. Πειραματική μελέτη

6.1 Σύνολο δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν για τις ανάγκες της παρούσας διπλωματικής εργασίας προήλθαν από το πρόγραμμα ΑΡΙΣΤΟΤΕΛΗΣ, που απευθυνόταν σε χρήστες ενδοφλέβιων ναρκωτικών της Αθήνας. Η συμμετοχή στο πρόγραμμα εκτός από αιματολογικές εξετάσεις περιελάμβανε και συνέντευξη με δομημένο ερωτηματολόγιο, το οποίο ήταν αρκετά εκτεταμένο. Το πρόγραμμα υλοποιήθηκε σε πέντε διαδοχικούς κύκλους, συνολικής διάρκειας 16 μηνών (Αύγουστος 2012 – Δεκέμβριος 2013).

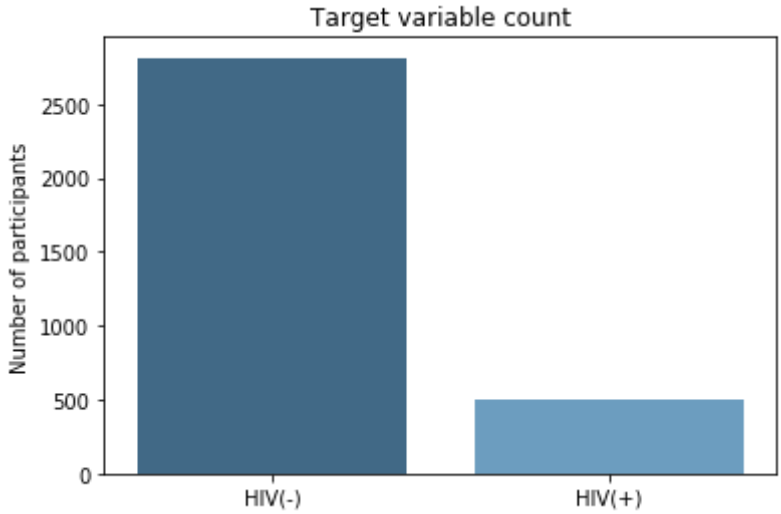
Συνολικά, συλλέχθηκαν 7.113 ερωτηματολόγια/βιολογικά δείγματα από 3.320 μοναδικούς ΧΕΝ με πολλαπλές συμμετοχές. Σε κάθε κύκλο μπορούσε κάποιος να συμμετάσχει μόνο μια φορά, ενώ υπήρχε η δυνατότητα συμμετοχής και σε επόμενους κύκλους [12]. Στη διάρκεια του προγράμματος προσήλθε το 88% των ΧΕΝ εκείνης της περιόδου, βάσει των επίσημων εκτιμήσεων για τον πληθυσμό των χρηστών στην Αθήνα από το Εθνικό Κέντρο Τεκμηρίωσης και Πληροφόρησης για τα Ναρκωτικά [11].

Το αρχικό σύνολο δεδομένων περιείχε 438 μεταβλητές που αφορούσαν μεταξύ άλλων, στα δημογραφικά χαρακτηριστικά, στη χρήση ουσιών, στις σεξουαλικές συμπεριφορές, και σε πληροφορίες σχετικά με τα προγράμματα μείωσης της βλάβης (προγράμματα υποκατάστασης, λήψη δωρεάν συρίγγων και άλλα).

Για την εύρεση ταξινομητή αναφορικά με τη μόλυνση με HIV χρησιμοποιήθηκαν τα δεδομένα από την πρώτη επίσκεψη των συμμετεχόντων στο πρόγραμμα ΑΡΙΣΤΟΤΕΛΗΣ, βάσει της οποίας ο επιπολασμός του HIV ήταν 15,2%. Πιο συγκεκριμένα, στο σύνολο των 3.320 ΧΕΝ υπήρχαν 506 HIV(+) (Εικόνα 18).

Για την ανάλυση των δεδομένων, χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python. Ένα από τα πλεονεκτήματα της Python είναι η ύπαρξη πληθώρας βιβλιοθηκών. Στο πλαίσιο εκπόνησης της παρούσας διπλωματικής εργασίας

χρησιμοποιήθηκαν μεταξύ άλλων οι βιβλιοθήκες Pandas και Numpy για τη στατιστική ανάλυση, η βιβλιοθήκη Matplotlib και Seaborn για τα γραφήματα, η βιβλιοθήκη scikit learn, για την υλοποίηση αλγορίθμων Μηχανικής Μάθησης, παρέχοντας τη δυνατότητα εύρεσης των κατάλληλων παραμέτρων με σκοπό την επίτευξη όσο το δυνατόν καλύτερης προσαρμογής στα δεδομένα, καθώς και η imbalanced-learn για την αντιμετώπιση μη ισορροπημένων δεδομένων προσφέροντας ένα μεγάλο αριθμό τεχνικών επαναδειγματοληψίας.



Εικόνα 18. Κατανομή της μεταβλητής στόχου (HIV).

6.2 Προεπεξεργασία δεδομένων

Το πρώτο στάδιο της προεπεξεργασίας αφορά στον καθαρισμό των δεδομένων. Για κάθε μεταβλητή υπολογίστηκε το ποσοστό των ελλειπουσών τιμών και στη συνέχεια εξαιρέθηκαν όσες μεταβλητές είχαν πάνω από 12% ελλείπουσες τιμές. Στο στάδιο αυτό αφαιρέθηκαν 296 μεταβλητές. Στη συνέχεια δημιουργήθηκαν 3 μεταβλητές, οι οποίες προήλθαν από 6 υπάρχουσες μεταβλητές, οι οποίες αφαιρέθηκαν από το σύνολο δεδομένων. Αυτό είχε σαν αποτέλεσμα ο αριθμός των μεταβλητών να είναι 139. Κατόπιν, αφαιρέθηκαν οι μεταβλητές εκείνες που ήταν ίδιες αλλά με διαφορετικό όνομα ή υπήρχε επικάλυψη, όπως η ηλικία και το έτος γέννησης και όσες δεν είχαν κάποιο «νόημα» στην ανάλυση, όπως για παράδειγμα ο μοναδικός κωδικός του συμμετέχοντα. Συνολικά, σε αυτό το βήμα αφαιρέθηκαν 64 μεταβλητές.

Τέλος, αφαιρέθηκαν όσες μεταβλητές σχετίζονταν με την έκβαση, όπως για παράδειγμα το αποτέλεσμα του αντιγόνου για την HIV λοίμωξη, η λήψη αντιρετροϊκής θεραπείας κοκ.. Στο στάδιο αυτό αφαιρέθηκαν συνολικά 6 μεταβλητές. Το σύνολο δεδομένων μετά την ολοκλήρωση των παραπάνω βημάτων περιείχε 69 μεταβλητές και 3.320 εγγραφές (Πίνακας 1).

Το επόμενο στάδιο ήταν η συμπλήρωση των ελλειπουσών τιμών. Στις συνεχείς μεταβλητές έγινε αντικατάσταση των ελλειπουσών τιμών με τη διάμεση τιμή της εκάστοτε μεταβλητής. Λόγω πλήθους παρατηρήσεων, οι κατανομές των συνεχών μεταβλητών θεωρείται πως ακολουθούν την κανονική κατανομή, δηλαδή η μέση και η διάμεση τιμή είναι αρκετά κοντά μεταξύ τους. Ωστόσο, αν κάποια μεταβλητή δεν ακολουθεί κανονική κατανομή, τότε η διάμεση τιμή είναι πιο αντιπροσωπευτική καθώς δεν επηρεάζεται από ακραίες παρατηρήσεις. Στις κατηγορικές μεταβλητές έγινε αντικατάσταση των ελλειπουσών τιμών με την επικρατούσα τιμή, δηλαδή την τιμή εκείνη που είχε τη μεγαλύτερη συχνότητα.

Το επόμενο στάδιο της προεπεξεργασίας δεδομένων ήταν η κωδικοποίηση των κατηγορικών μεταβλητών. Δημιουργήθηκε μια λίστα με όλες τις αλφαριθμητικές (string, str) μεταβλητές, προκειμένου να εντοπιστούν και να εφαρμοστεί είτε η Διατάξιμη Κωδικοποίηση (Ordinal Encoding) είτε η Κωδικοποίηση One-Hot. Όπως αναφέρθηκε και στο προηγούμενο κεφάλαιο στις ονομαστικές κατηγορικές εφαρμόστηκε η κωδικοποίηση One-Hot. Με αυτόν τον τρόπο τελικά προστέθηκαν 44 νέες δίτιμες μεταβλητές. Εφαρμόζοντας την διατάξιμη κωδικοποίηση δεν επηρεάστηκε, όπως ήταν αναμενόμενο το πλήθος των μεταβλητών. Συνεπώς, το τελικό σύνολο δεδομένων περιείχε 3.320 εγγραφές και 113 μεταβλητές.

Στο τελευταίο στάδιο έγινε η διάσπαση του συνόλου δεδομένων σε σύνολο εκπαίδευσης (X_train, y_train) και σε σύνολο ελέγχου (X_test, y_ test). Η διάσπαση έγινε με αναλογία 4:1 και στρωματοποιώντας ως προς τη μεταβλητή ενδιαφέροντος. Με αυτόν τον τρόπο διασφαλίζεται η τήρηση των αναλογιών ως προς τη μεταβλητή ενδιαφέροντος. Συνεπώς, το 80% των εγγραφών χρησιμοποιήθηκαν για την εκπαίδευση των αλγορίθμων, ενώ το υπόλοιπο 20% για το έλεγχο της απόδοσής τους. Με αυτόν τον διαχωρισμό το σύνολο εκπαίδευσης πλέον είχε 2.656 εγγραφές από τις

οποίες το 15,25% ήταν HIV(+), ενώ το σύνολο ελέγχου είχε 664 εγγραφές από τις οποίες το 15,21% ήταν HIV(+).

Πίνακας 1. Περιγραφή 69 χαρακτηριστικών του τελικού συνόλου δεδομένων

a/a	Μεταβλητή	Τύπος	Μοναδ. τιμές	Περιγραφή	Ελλείπουσες τιμές, ν (%)
1	hivstatus	Κατηγορική	2	HIV-status	0 (0,0)
2	sex	Κατηγορική	3	Φύλο	0 (0,0)
3	age	Συνεχής	2825	Ηλικία	0 (0,0)
4	newethnic	Κατηγορική	6	Εθνικότητα	0 (0,0)
5	ht1	Κατηγορική	3	Έχεις ποτέ εξεταστεί για τον ιό HIV;	0 (0,0)
6	lastshare_pos	Κατηγορική	2	Το τελευταίο άτομο με το οποίο έκανες κοινή ενδοφλέβια χρήση είχε HIV;	0 (0,0)
7	ns1	Κατηγορική	22	Ποιο άτομο σου έδωσε το κουπόνι;	7 (0,2)
8	id2	Κατηγορική	7	Συχνότητα ενδοφλέβιας χρήσης	7 (0,2)
9	dm1	Κατηγορική	2	Άστεγος τους τελ. 12 μήνες	8 (0,2)
10	dm1a	Κατηγορική	2	Άστεγος τώρα	8 (0,2)
11	hc2	Κατηγορική	5	Έχεις ποτέ εξεταστεί και διαγνωστεί με ηπατίτιδα;	8 (0,2)
12	home	Κατηγορική	3	Άστεγος (από dm1 και dm1a)	8 (0,2)
13	es13	Κατηγορική	58	Συνήθως σε ποιο σημείο κάνεις ενέσιμη χρήση;	9 (0,3)
14	id31	Κατηγορική	4	Συχνότητα ηρωίνης	9 (0,3)
15	pa1	Κατηγορική	2	Δωρεάν προφυλακτικά τους τελ. 12 μήνες	11 (0,3)
16	es112	Συνεχής	48	Πότε ήταν η τελευταία φορά που έκανες ενδοφλέβια χρήση (ημέρες);	10 (0,3)
17	idu_within_month	Κατηγορική	2	Χρήση τελευταίο μήνα	10 (0,3)
18	id32	Κατηγορική	4	Συχνότητα κοκαΐνης	11 (0,3)
19	id4	Κατηγορική	88	Από που έβρισκες σύριγγες τους τελ. 12 μήνες	11 (0,3)
20	id5	Κατηγορική	5	Χρήση καινούριας βελόνας (τελ. 12 μήνες)	12 (0,4)
21	idf3	Συνεχής	5	Την τελευταία σύριγγα πόσες φορές την χρησιμοποίησες μόνο εσύ;	14 (0,4)
22	share_cotton	Κατηγορική	4	Χρήση χρησιμοποιημένου βαμβακιού (τελ. 12 μήνες)	12 (0,4)
23	drugtreat	Κατηγορική	2	Πρόγραμμα απεξάρτησης από τα ναρκωτικά	13 (0,4)
24	prison	Κατηγορική	2	Ιστορικό φυλάκισης	13 (0,4)
25	prison_last_year	Κατηγορική	2	Φυλάκιση τους τελ. 12 μήνες	14 (0,4)
26	share_howmany	Κατηγορική	4	Από πόσα άτομα πήρες χρησιμοποιημένη σύριγγα (τελ. 30 ημέρες)	14 (0,4)
27	share_tasi	Κατηγορική	4	Χρήση χρησιμοποιημένου τάσι (τελ. 12 μήνες)	15 (0,5)
28	dm6	Κατηγορική	7	Επαγγελματική κατάσταση	17 (0,5)
29	dm8	Κατηγορική	2	Ασφάλιση	17 (0,5)

a/a	Μεταβλητή	Τύπος	Μοναδ. τιμές	Περιγραφή	Ελλείπουσες τιμές, ν (%)
30	nd1_b	Κατηγορική	2	Τους τελ. 12 μήνες, έχεις χρησιμοποιήσει άλλες ουσίες με μη ενέσιμο τρόπο;	17 (0,5)
31	share	Κατηγορική	5	Τους τελ. 12 μήνες, συχνότητα χρήσης χρησιμοποιημένης σύριγγας	17 (0,5)
32	id36	Κατηγορική	4	Συχνότητα speedball	19 (0,6)
33	share_water	Κατηγορική	4	Τους τελ. 12 μήνες, συχνότητα χρήσης χρησιμοποιημένου νερού	22 (0,7)
34	id34	Κατηγορική	4	Συχνότητα μορφίνης	23 (0,7)
35	es12	Κατηγορική	5	Ουσία ενδοφλέβιας χρήσης	24 (0,7)
36	hc113	Κατηγορική	3	Εξέταση για σύφιλη τους τελ. 12 μήνες	24 (0,7)
37	id33	Κατηγορική	4	Συχνότητα βουπρενορφίνης	24 (0,7)
38	hc112	Κατηγορική	3	Εξέταση για χλαμύδια τους τελ. 12 μήνες	25 (0,8)
39	hc111	Κατηγορική	3	Εξέταση για γονόρροια τους τελ. 12 μήνες	26 (0,8)
40	share_divide	Κατηγορική	4	Τους τελ. 12 μήνες, συχνότητα χρήσης ναρκωτικών που είχαν μοιραστεί με χρησιμοποιημένη σύριγγα	26 (0,8)
41	tx1	Κατηγορική	2	Συμμετοχή σε πρόγραμμα απεξάρτησης από το αλκοόλ	27 (0,8)
42	pa4	Κατηγορική	2	Ενημέρωση πρόληψης για HIV τους τελευταίος 12 μήνες	27 (0,8)
43	ns2m	Συνεχής	51	Δίκτυο. Πόσα άτομα γνωρίζεις	28 (0,8)
44	id35	Κατηγορική	4	Συχνότητα σίσα	29 (0,9)
45	dm5a	Κατηγορική	7	Εκπαιδευτικό επίπεδο	32 (1,0)
46	last_test	Κατηγορική	2	Πότε εξετάστηκες τελευταία φορά για τον HIV;	37 (1,1)
47	al3	Συνεχής	24	Ημέρες που ήπιε αλκοόλ στη διάρκεια του τελευταίου μήνα	47 (1,4)
48	nowost	Κατηγορική	4	Πρόγραμμα υποκατάστασης	49 (1,5)
49	money	Κατηγορική	3	Έλαβες κάποιο χρηματικό ποσό για sex ως αντάλλαγμα;	52 (1,6)
50	syringes_number_cat	Κατηγορική	4	Αριθμός συρίγγων των τελευταίο μήνα (ομαδοποιημένη)	56 (1,7)
51	id6m	Συνεχής	35	Αριθ. ατόμων για κοινή χρήση βελόνας, τάσι, σύνεργα που έχει χρησιμοποιηθεί τον τελ. χρόνο	74 (2,2)
52	dm4ma	Κατηγορική	3	Συγκατοικείς με χρήστη;	84 (2,5)
53	nd21	Κατηγορική	4	Χρήση ηρωίνης τον τελευταίο χρόνο με μη-ενέσιμο τρόπο	109 (3,3)
54	nd22	Κατηγορική	4	Χρήση κοκαΐνης τον τελευταίο χρόνο με μη-ενέσιμο τρόπο	113 (3,4)
55	nd25	Κατηγορική	4	Χρήση σίσα τον τελευταίο χρόνο με μη-ενέσιμο τρόπο	115 (3,5)
56	nd210	Κατηγορική	4	Χρήση βενζοδιαζεπίνες τον τελευταίο χρόνο με μη-ενέσιμο τρόπο	116 (3,5)
57	nd211	Κατηγορική	4	Χρήση κάναβη-χασις τον τελευταίο χρόνο με μη-ενέσιμο τρόπο	117 (3,5)

a/a	Μεταβλητή	Τύπος	Μοναδ. τιμές	Περιγραφή	Ελλείπουσες τιμές, ν (%)
58	nd212	Κατηγορική	4	Χρήση αμφεταμίνη τον τελευταίο χρόνο με μη-ενέσιμο τρόπο	121 (3,6)
59	nd26	Κατηγορική	4	Χρήση μεθαδόνη τον τελευταίο χρόνο με μη-ενέσιμο τρόπο	121 (3,6)
60	nd23	Κατηγορική	4	Χρήση βουπρενορφίνη τον τελευταίο χρόνο με μη-ενέσιμο τρόπο	128 (3,9)
61	nd27	Κατηγορική	4	Χρήση LSD τον τελευταίο χρόνο με μη-ενέσιμο τρόπο	129 (3,9)
62	nd213	Κατηγορική	4	Χρήση έκσταση, MDA τον τελευταίο χρόνο με μη-ενέσιμο τρόπο	133 (4,0)
63	nd28	Κατηγορική	4	Χρήση βαρβιτουρικά τον τελευταίο χρόνο με μη-ενέσιμο τρόπο	138 (4,2)
64	nd29	Κατηγορική	4	Χρήση κωδεΐνη τον τελευταίο χρόνο με μη-ενέσιμο τρόπο	147 (4,4)
65	nd24	Κατηγορική	4	Χρήση μορφίνη τον τελευταίο χρόνο με μη-ενέσιμο τρόπο	184 (5,5)
66	share_last	Κατηγορική	2	Τελευταία φορά χρήση βελόνας αφού την είχε χρησιμοποιήσει κάποιος άλλος	240 (7,2)
67	tx2b	Κατηγορική	2	Προσπάθησες να μπεις σε κάποιο πρόγραμμα απεξάρτησης αλλά διέκοψες τον τελευταίο χρόνο;	317 (9,5)
68	hc6m	Κατηγορική	3	Έχεις κάνει εμβόλιο για την ηπατίτιδα Β;	371 (11,2)
69	newinjector	Κατηγορική	2	Νέος χρήστης (<=2 έτη/>2έτη)	15 (0,5)

6.3 Αποτελέσματα

Αρχικά παρουσιάζονται κάποια από τα βασικά χαρακτηριστικά των συμμετεχόντων, προκειμένου να σκιαγραφηθεί το προφίλ των ΧΕΝ της Αθήνας το 2012. Στη συνέχεια, τα αποτελέσματα χωρίζονται σε τρία βασικά μέρη. Στην πρώτο μέρος παρουσιάζονται τα αποτελέσματα των αλγορίθμων μηχανικής μάθησης στο σύνολο των δεδομένων. Στο δεύτερο μέρος παρουσιάζονται τα αποτελέσματα των αλγορίθμων που προκύπτουν μετά την επιλογή χαρακτηριστικών, ενώ στο τρίτο μέρος τα αποτελέσματα μετά την ανάλυση σε κύριες συνιστώσες.

Κάθε μέρος χωρίζεται σε δύο ενότητες. Στην πρώτη ενότητα χρησιμοποιούνται τα δεδομένα χωρίς κάποια επαναδειγματοληψία, ενώ στη δεύτερη ενότητα γίνεται εφαρμογή διαφορετικών μεθόδων επαναδειγματοληψίας. Τέλος, η δεύτερη ενότητα χωρίζεται σε τέσσερις υποενότητες, στις οποίες παρουσιάζονται τα αποτελέσματα από την τυχαία υποδειγματοληψία, την τυχαία υπερδειγματοληψία με επανάθεση, την τεχνική υπερδειγματοληψίας συνθετικής μειονότητας και την προσαρμοστική συνθετική μέθοδο δειγματοληψίας για μη ισορροπημένα δεδομένα.

Οι αλγόριθμοι που χρησιμοποιήθηκαν σε όλα τα στάδια της ανάλυσης ήταν οι παρακάτω πέντε: Logistic Regression, Random Forest, Support Vector Machines, k-Εγγύτεροι Γείτονες και Decision Tree. Ως μετρικές για την αξιολόγηση των αλγορίθμων χρησιμοποιήθηκε η ορθότητα, η ακρίβεια, η ανάκλαση, το f1-score καθώς και το εμβαδόν κάτω από την καμπύλη λειτουργικών χαρακτηριστικών.

Στον Πίνακα 2, παρουσιάζονται κάποια από τα βασικά χαρακτηριστικά των ΧΕΝ, βάσει της πρώτης τους συμμετοχής στο πρόγραμμα ΑΡΙΣΤΟΤΕΛΗΣ. Η μέση ηλικία ήταν τα 36 έτη, ενώ στην πλειοψηφία τους ήταν άνδρες, Ελληνικής εθνικότητας, άνεργοι και ανασφάλιστοι. Το 23% των ΧΕΝ δήλωναν πως ήταν άστεγοι κατά την επίσκεψή τους στο πρόγραμμα, ενώ πάνω από 80% ήταν ενεργοί χρήστες, δηλαδή είχαν κάνει ενέσιμη χρήση τις τελευταίες 30 ημέρες. Σε πρόγραμμα υποκατάστασης ήταν ενταγμένοι κατά την επίσκεψή τους στο πρόγραμμα ΑΡΙΣΤΟΤΕΛΗΣ, το 13% των συμμετεχόντων. Οι μισοί περίπου δήλωσαν πως είχαν λάβει δωρεάν σύριγγες τους τελευταίους 12 μήνες.

Πίνακας 2. Βασικά χαρακτηριστικά του δείγματος βάσει της πρώτης τους συμμετοχής στο πρόγραμμα ΑΡΙΣΤΟΤΕΛΗΣ, Αθήνα, 2012.

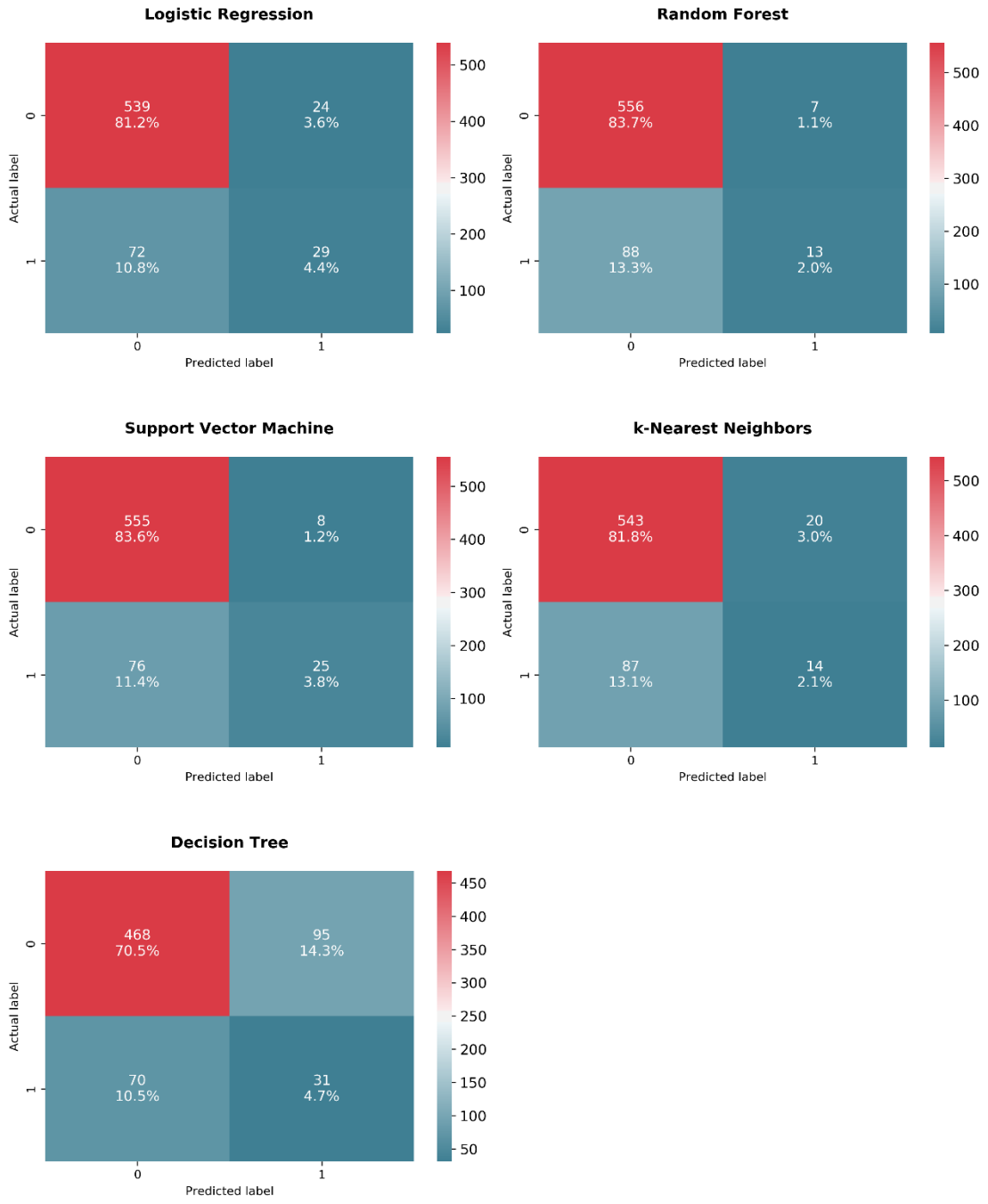
Χαρακτηριστικά	**N=3.320**
Άνδρες, ν (%)	2.806 (84,5)
Ηλικία (έτη), μ.τ. (τ.α.)	35,9 (8,4)
Ελληνική Εθνικότητα, ν (%)	2.793 (84,1)
Επίπεδο εκπαίδευσης, ν (%)	
Έως Δημοτικό	907 (27,6)
Γυμνάσιο	1.004 (30,5)
Λύκειο	972 (29,6)
Πάνω από Λύκειο	405 (12,3)
Απουσία στέγης (τώρα), ν (%)	765 (23,1)
Άνεργοι/Δεν μπορεί να εργαστεί για λόγους υγείας, ν (%)	2.861 (86,6)
Ανασφάλιστοι, ν (%)	2.117 (64,1)
Ενδοφλέβια χρήση τον τελευταίο μήνα, ν (%)	2.689 (81,2)
Σε πρόγραμμα υποκατάστασης (τώρα), ν (%)	409 (12,5)
Λήψη δωρεάν συρίγγων τους τελευταίους 12 μήνες, ν (%)	1.675 (50,6)

6.4 Αποτελέσματα στο σύνολο των δεδομένων

6.4.1 Αποτελέσματα χωρίς εφαρμογή μεθόδων επαναδειγματοληψίας

Στην Εικόνα 19, παρουσιάζονται οι πίνακες σύγχυσης από κάθε αλγόριθμό, ενώ στον Πίνακα 3, παρουσιάζονται οι μετρικές βάσει των οποίων αξιολογήθηκαν οι πέντε αλγόριθμοι στο σύνολο των δεδομένων χωρίς να εφαρμοστεί κάποια μέθοδος

επαναδειγματοληψίας. Στην Εικόνα 20, αποτυπώνονται τα AUC score από κάθε αλγόριθμο.

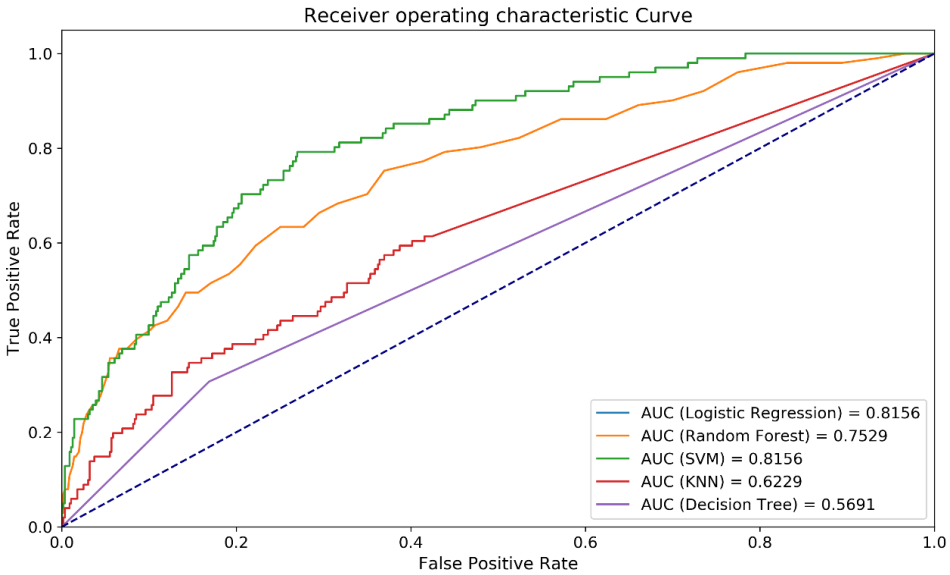


Εικόνα 19. Πίνακες σύγχυσης των αλγορίθμων Logistic Regression, Random Forest, Support Vector Machines, k-Εγγύτεροι Γείτονες και Decision Tree στο σύνολο των δεδομένων χωρίς την εφαρμογή κάποιας μεθόδου επαναδειγματοληψίας.

Πίνακας 3. Μετρικές αξιολόγησης των αλγορίθμων Logistic Regression, Random Forest, Support Vector Machines, k-Εγγύτεροι Γείτονες και Decision Tree στο σύνολο των δεδομένων χωρίς την εφαρμογή κάποιας μεθόδου επαναδειγματοληψίας.

	Accuracy	Precision	Recall	f1-score	AUC
Logistic Regression	0.8554	0.5472	0.2871	0.3766	0.8156
Random Forest	0.8569	0.6500	0.1287	0.2149	0.7529
Support Vector Machines	0.8735	0.7576	0.2475	0.3731	0.8156
k-Εγγύτεροι Γείτονες	0.8389	0.4118	0.1386	0.2074	0.6229
Decision Tree	0.7515	0.2460	0.3069	0.2731	0.5691

Από τον Πίνακα 3, φαίνεται πως η λογιστική παλινδρόμηση είχε την καλύτερη επίδοση ως προ τη μετρική AUC και ακολουθούν, με πολύ μικρή διαφορά, οι μηχανές διανυσμάτων υποστήριξης. Ωστόσο, όλοι οι αλγόριθμοι είχαν πολύ χαμηλή επίδοση ως προς τη μετρική recall (εύρος: 0.1287 – 0.3069), δηλαδή την ικανότητά τους να ανιχνεύουν περιπτώσεις που ανήκουν στην κλάση 1. Ο αλγόριθμος Decision tree, που είχε τη μεγαλύτερη ανάκλαση-ευαισθησία, προέβλεψε σωστά το 30.69% των παρατηρήσεων της κλάσης 1. Συνολικά, οι λάθος ταξινομήσεις κυμάνθηκαν από 14.4% έως 24.8%, ενώ για την κλάση 1 κυμάνθηκαν από 69.3% έως 87.1%.



Εικόνα 20. Εμβαδόν κάτω από την καμπύλη ανά αλγόριθμο, την εφαρμογή κάποιας μεθόδου επαναδειγματοληψίας.

6.4.2 Αποτελέσματα με εφαρμογή μεθόδων επαναδειγματοληψίας

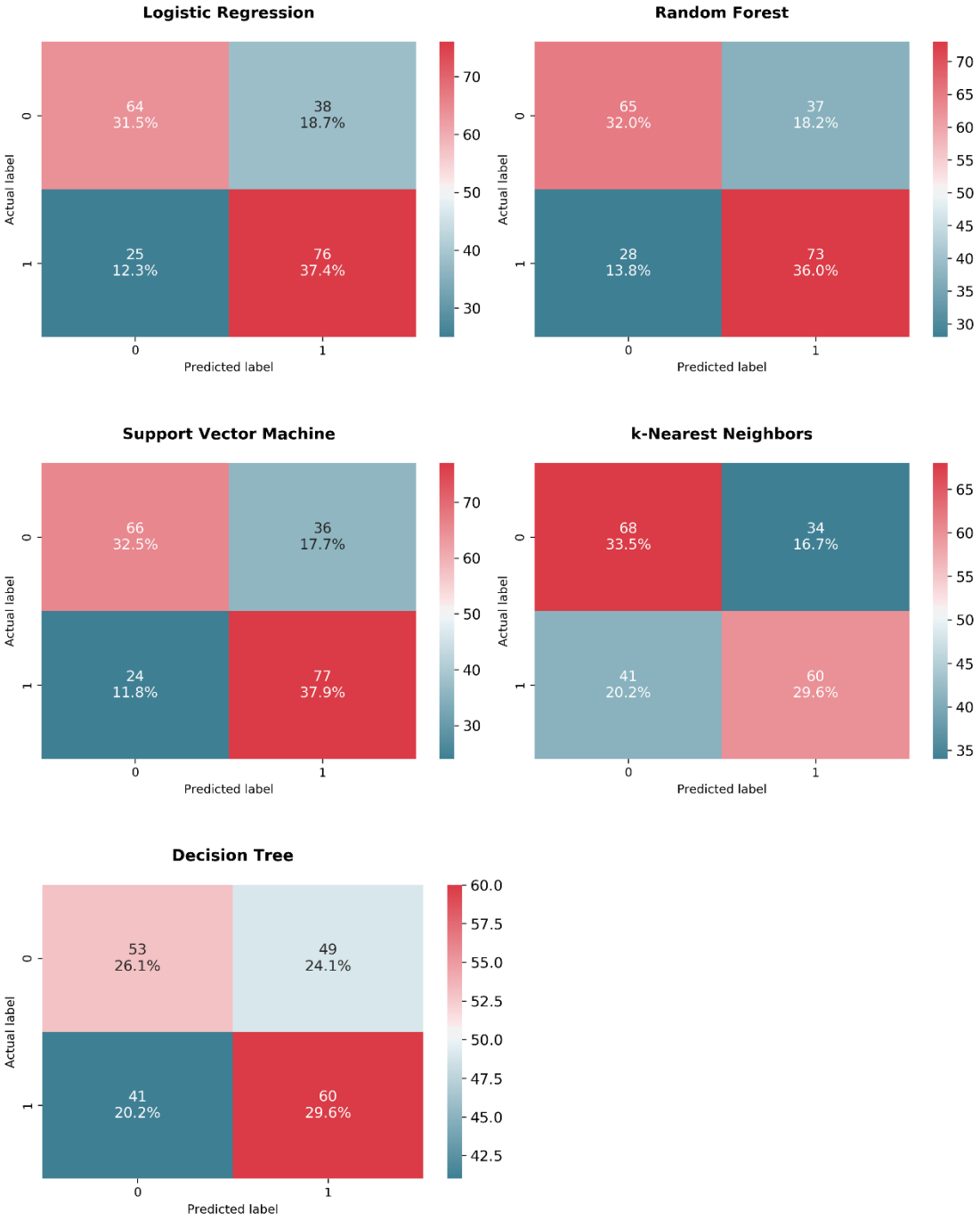
Μια βασική κατηγορία τεχνικών που μπορεί να εφαρμοστεί για την αντιμετώπιση των μη ισορροπημένων δεδομένων είναι οι μέθοδοι επαναδειγματοληψίας. Στη συνέχεια, παρουσιάζονται αποτελέσματα τόσο από την εφαρμογή τυχαίας υποδειγματοληψίας, όσο και από την τυχαία υπερδειγματοληψία.

6.4.2.1 Με τυχαία υποδειγματοληψία

Στη μέθοδο τυχαίας υποδειγματοληψίας, αφαιρούνται με τυχαίο τρόπο εγγραφές από την κλάση HIV(-) μέχρι να έχει το ίδιο πλήθος με την κλάση HIV(+). Στην περίπτωση αυτή τα σύνολα εκπαίδευσης και ελέγχου περιέχουν 809 και 203 εγγραφές, αντίστοιχα, που είναι το 80% και 20% του συνόλου δεδομένων έτσι όπως έχει διαμορφωθεί μετά την τυχαία υποδειγματοληψία.

Στην Εικόνα 21, παρουσιάζονται οι πίνακες σύγχυσης από κάθε αλγόριθμό, ενώ στον Πίνακα 4, παρουσιάζονται οι μετρικές βάσει των οποίων αξιολογήθηκαν οι πέντε αλγόριθμοι στο σύνολο των δεδομένων με τυχαία υπερδειγματοληψία. Στην Εικόνα 22, αποτυπώνονται τα AUC score από κάθε αλγόριθμο.

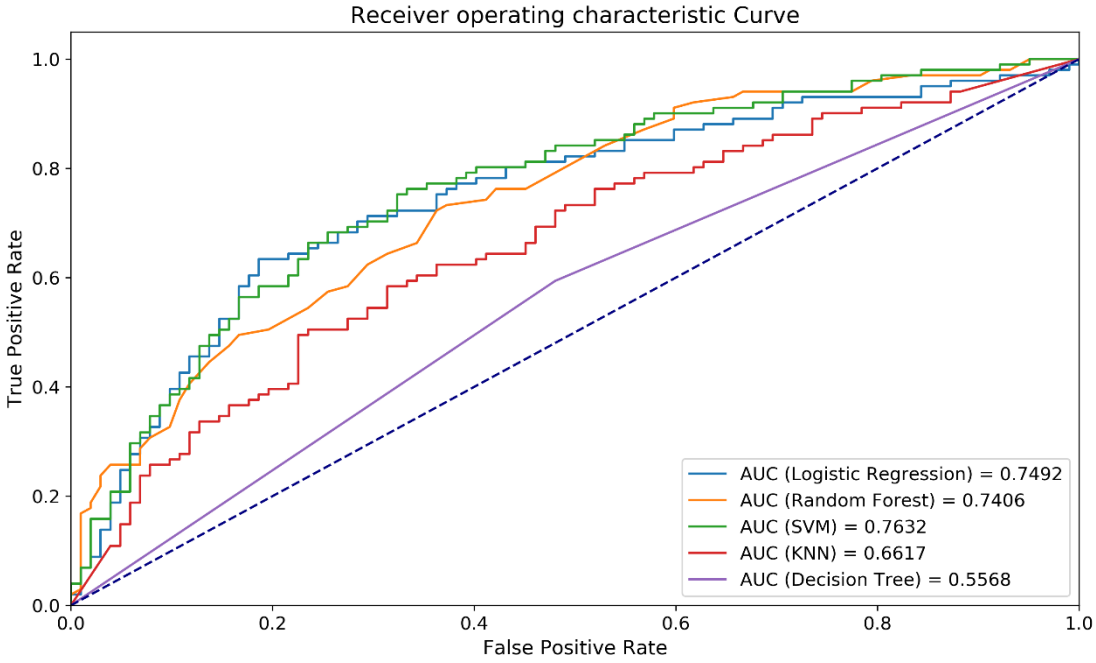
Εφαρμόζοντας τη μέθοδο τυχαίας υποδειγματοληψίας, φαίνεται πως μειώθηκε το AUC στην πλειοψηφία των αλγορίθμων, ωστόσο υπήρχε αύξηση τόσο στην ανάκλαση (εύρος: 0.5941 – 0.7624), όσο και στην ακρίβεια (εύρος: 0.5505 – 0.6814) (Πίνακας 4). Όπως ήταν αναμενόμενο, αυξήθηκε και το f1-score. Από τους πίνακες σύγχυσης (Εικόνα 21), συμπεραίνουμε πως ο αλγόριθμος SVM είχε το μεγαλύτερο ποσοστό σωστής ταξινόμησης για την κλάση 1 (76.2%), σε σχέση με τον kNN και το Decision tree που είχαν τα μικρότερα ποσοστά (59.4%).



Εικόνα 21. Πίνακες σύγχυσης των αλγορίθμων Logistic Regression, Random Forest, Support Vector Machines, k-Εγγύτεροι Γείτονες και Decision Tree στο σύνολο των δεδομένων με τυχαία υποδειγματοληψία.

Πίνακας 4. Μετρικές αξιολόγησης των αλγορίθμων Logistic Regression, Random Forest, Support Vector Machines, k-Εγγύτεροι Γείτονες και Decision Tree στο σύνολο των δεδομένων με τυχαία υποδειγματοληψία.

	Accuracy	Precision	Recall	f1-score	AUC
Logistic Regression	0.6897	0.6667	0.7525	0.7070	0.7492
Random Forest	0.6798	0.6636	0.7228	0.6919	0.7406
Support Vector Machines	0.7044	0.6814	0.7624	0.7196	0.7632
k-Εγγύτεροι Γείτονες	0.6305	0.6383	0.5941	0.6154	0.6617
Decision Tree	0.5567	0.5505	0.5941	0.5714	0.5568



Εικόνα 22. Εμβαδόν κάτω από την καμπύλη ανά αλγόριθμο, την εφαρμογή κάποιας μεθόδου επαναδειγματοληψίας.

6.4.2.2 Με τυχαία υπερδειγματοληψία

Στη μέθοδο τυχαίας υπερδειγματοληψίας, προστίθενται με τυχαίο τρόπο εγγραφές από την κλάση HIV(+) μέχρι να έχει το ίδιο πλήθος με την κλάση HIV(-). Η δειγματοληψία γίνεται με επανάθεση. Μετά την τυχαία υπερδειγματοληψία με

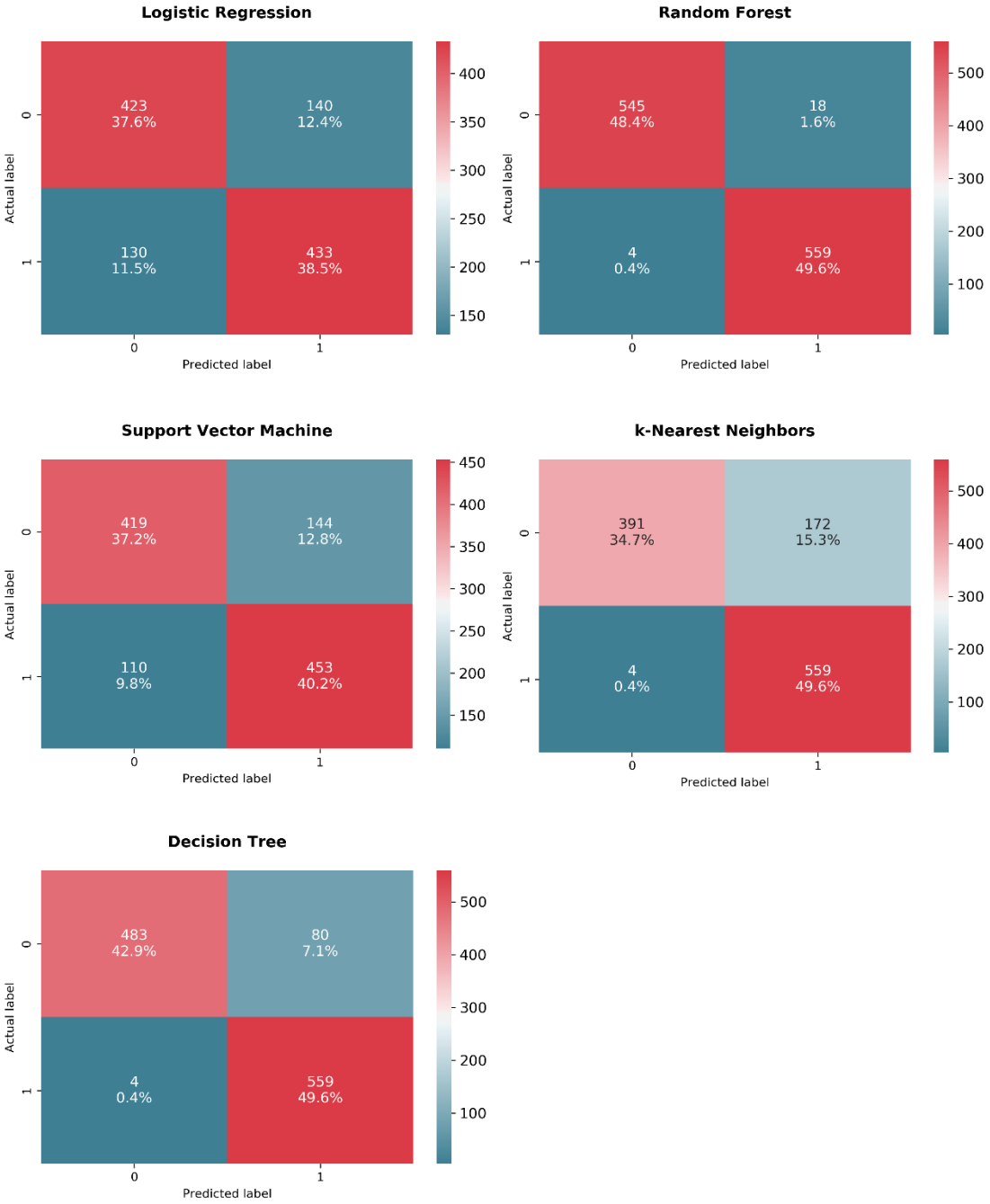
επανάθεση το σύνολο δεδομένων αριθμεί 5.628 εγγραφές, με δύο κλάσεις που έχουν τον ίδιο πληθικό αριθμό. Επομένως, τα σύνολα εκπαίδευσης και ελέγχου περιέχουν 4502 και 1126 εγγραφές, αντίστοιχα.

Στην Εικόνα 23, παρουσιάζονται οι πίνακες σύγχυσης από κάθε αλγόριθμό, ενώ στον Πίνακα 5, παρουσιάζονται οι μετρικές βάσει των οποίων αξιολογήθηκαν οι πέντε αλγόριθμοι στο σύνολο των δεδομένων με τυχαία υπερδειγματοληψία. Στην Εικόνα 24, αποτυπώνονται τα AUC score από κάθε αλγόριθμο.

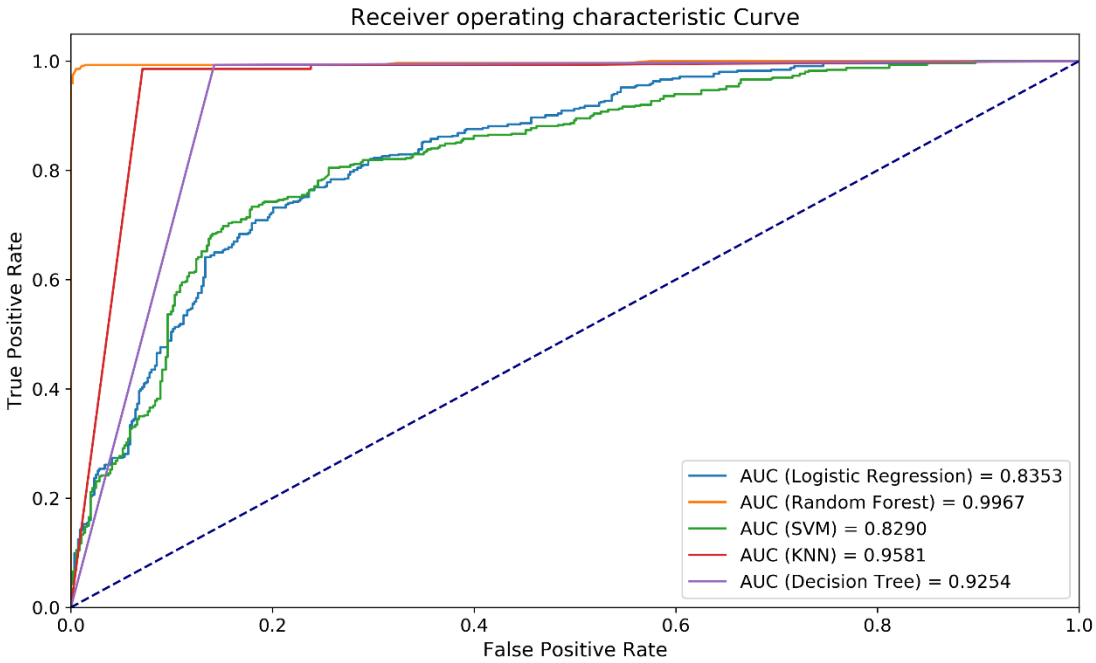
Από την Εικόνα 23, συμπεραίνουμε πως οι αλγόριθμοι Random forest, kNN και Decision tree είχαν τα μεγαλύτερα ποσοστά σωστής ταξινόμησης για την κλάση 1. Ωστόσο ο Random forest ήταν εκείνος με το μικρότερο ποσοστό λανθασμένης ταξινόμησης και για τις δύο κλάσεις, το οποίο ήταν 2%. Ο αλγόριθμος με το υψηλότερο AUC score, ήταν ο Random forest (0.9967, Εικόνα 24), ενώ η ανάκλαση ήταν 99.29%, η ορθότητα 98.05% και η ακρίβεια 96.9% (Πίνακας 5). Συνολικά, οι λάθος ταξινομήσεις κυμάνθηκαν από 2.0% έως 23.9%, ενώ για την κλάση 1 κυμάνθηκαν από 0.7% έως 23.1%.

Πίνακας 5. Μετρικές αξιολόγησης των αλγορίθμων Logistic Regression, Random Forest, Support Vector Machines, k-Εγγύτεροι Γείτονες και Decision Tree στο σύνολο των δεδομένων με τυχαία υπερδειγματοληψία.

	Accuracy	**Precision**	**Recall**	**f1-score**	**AUC**
Logistic Regression	0.7602	0.7557	0.7691	0.7623	0.8353
Random Forest	0.9805	0.9688	0.9929	0.9807	0.9967
Support Vector Machines	0.7744	0.7588	0.8046	0.7810	0.8290
k-Εγγύτεροι Γείτονες	0.8437	0.7647	0.9929	0.8640	0.9581
Decision Tree	0.9254	0.8748	0.9929	0.9301	0.9254



Εικόνα 23. Πίνακες σύγχυσης των αλγορίθμων Logistic Regression, Random Forest, Support Vector Machines, k-Εγγύτεροι Γείτονες και Decision Tree στο σύνολο των δεδομένων με τυχαία υπερδειγματοληψία.

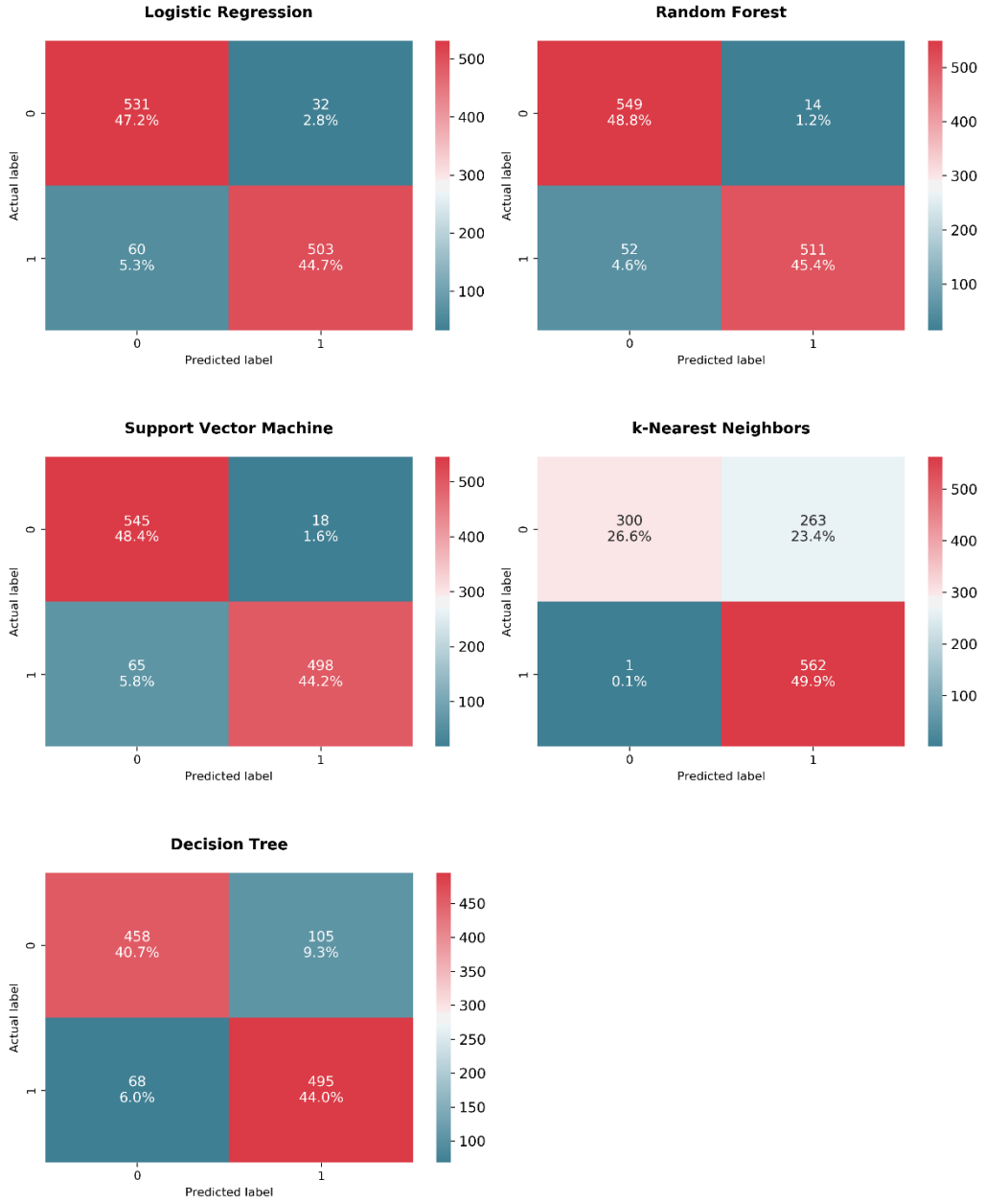


Εικόνα 24. Εμβαδόν κάτω από την καμπύλη ανά αλγόριθμο, την εφαρμογή κάποιας μεθόδου επαναδειγματοληψίας.

6.4.2.3 Με τεχνική υπερδειγματοληψίας συνθετικής μειονότητας (SMOTE)

Στην Εικόνα 25, παρουσιάζονται οι πίνακες σύγχυσης από κάθε αλγόριθμό, ενώ στον Πίνακα 6, παρουσιάζονται οι μετρικές βάσει των οποίων αξιολογήθηκαν οι πέντε αλγόριθμοι στο σύνολο των δεδομένων με τυχαία υπερδειγματοληψία συνθετικής μειονότητας. Στην Εικόνα 26, αποτυπώνονται τα AUC score από κάθε αλγόριθμο.

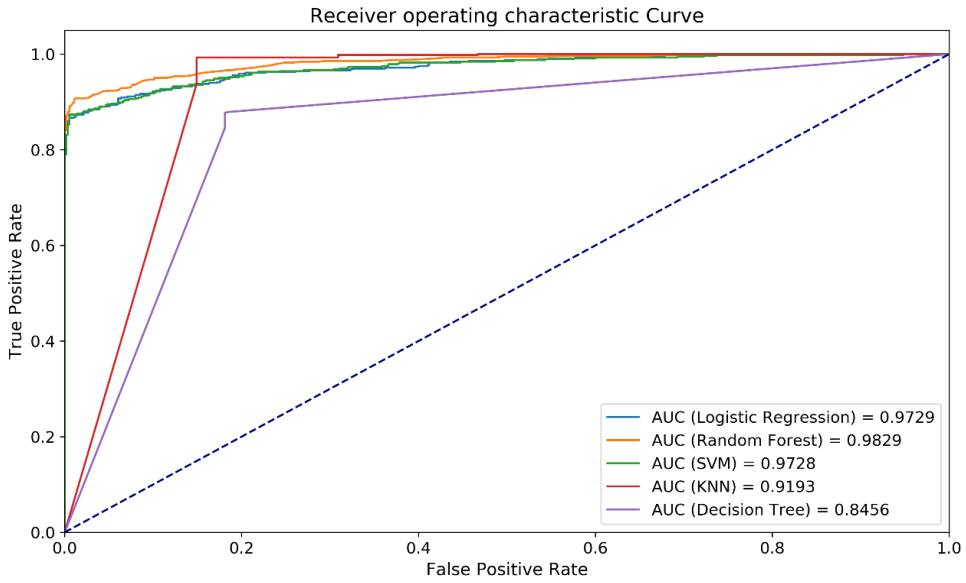
Ο αλγόριθμος με το μεγαλύτερο AUC-score ήταν ο Random forest (0.9829), με ευαισθησία 90.8%, ορθότητα 94.1%, ακρίβεια 97.3%. Συνολικά, οι λάθος ταξινομήσεις κυμάνθηκαν από 5.9% έως 23.4%, ενώ για την κλάση 1 κυμάνθηκαν από 0.2% έως 12.1%.



Εικόνα 25. Πίνακες σύγχυσης των αλγορίθμων Logistic Regression, Random Forest, Support Vector Machines, k-Εγγύτεροι Γείτονες και Decision Tree στο σύνολο των δεδομένων με τεχνική υπερδειγματοληψίας συνθετικής μειονότητας.

Πίνακας 6. Μετρικές αξιολόγησης των αλγορίθμων Logistic Regression, Random Forest, Support Vector Machines, k-Εγγύτεροι Γείτονες και Decision Tree στο σύνολο των δεδομένων με τεχνική υπερδειγματοληψίας συνθετικής μειονότητας.

	Accuracy	Precision	Recall	f1-score	AUC
Logistic Regression	0.9183	0.9402	0.8934	0.9162	0.9729
Random Forest	0.9414	0.9733	0.9076	0.9393	0.9829
Support Vector Machines	0.9263	0.9651	0.8845	0.9231	0.9728
k-Εγγύτεροι Γείτονες	0.7655	0.6812	0.9982	0.8098	0.9193
Decision Tree	0.8464	0.8250	0.8792	0.8512	0.8456



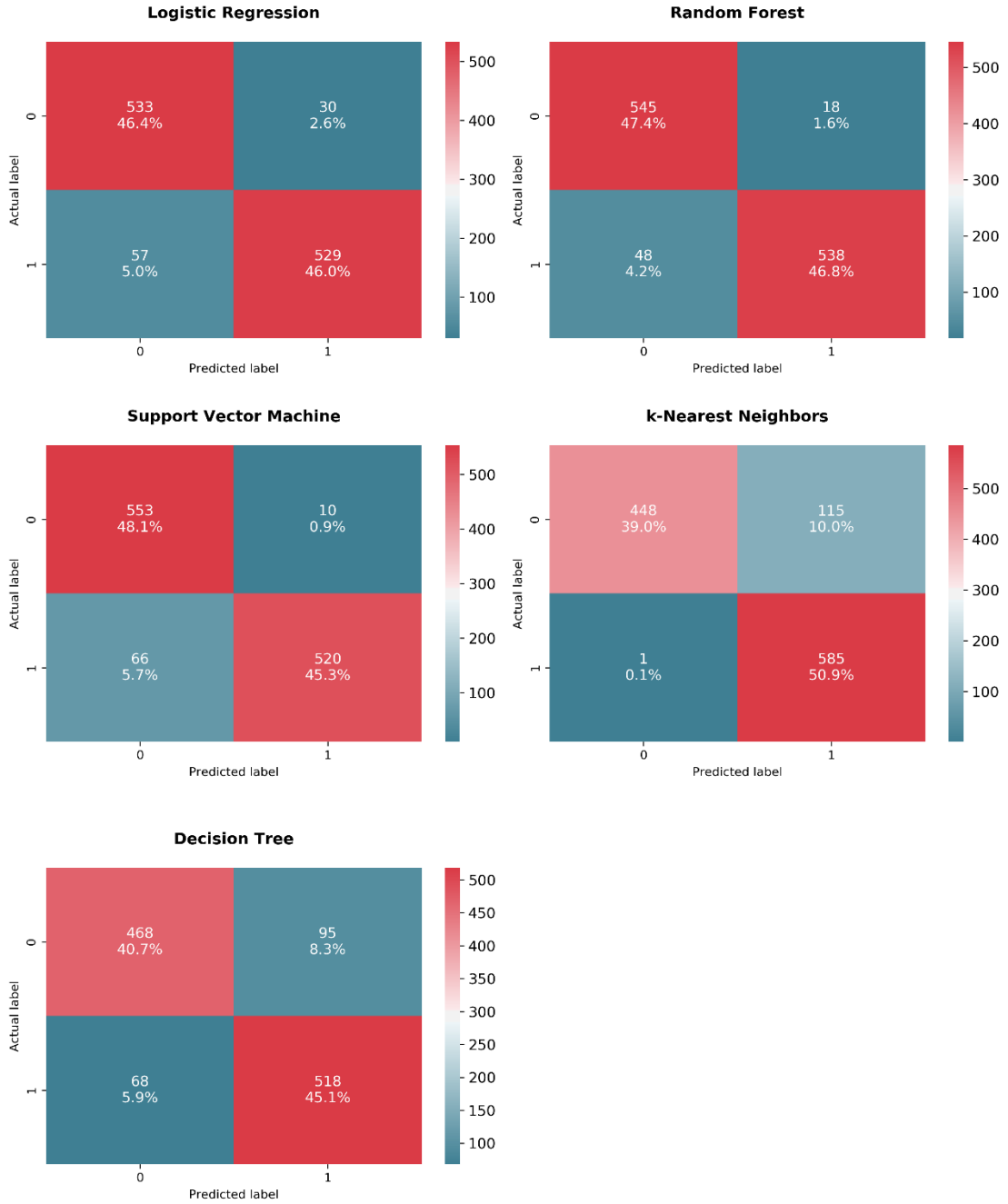
Εικόνα 26. Εμβαδόν κάτω από την καμπύλη ανά αλγόριθμο, την εφαρμογή κάποιας μεθόδου επαναδειγματοληψίας.

6.4.2.4 Με προσαρμοστική συνθετική μέθοδο δειγματοληψίας (ADASYN)

Στην Εικόνα 27, παρουσιάζονται οι πίνακες σύγχυσης από κάθε αλγόριθμό, ενώ στον Πίνακα 7, παρουσιάζονται οι μετρικές βάσει των οποίων αξιολογήθηκαν οι πέντε αλγόριθμοι στο σύνολο των δεδομένων με τυχαία υπερδειγματοληψία συνθετικής μειονότητας. Στην Εικόνα 28, αποτυπώνονται τα AUC score από κάθε αλγόριθμο.

Ο αλγόριθμος Random forest είχε την καλύτερη επίδοση ως προ τη μετρική AUC (0.9431) και ακολουθούν, με πολύ μικρή διαφορά, οι μηχανές διανυσμάτων

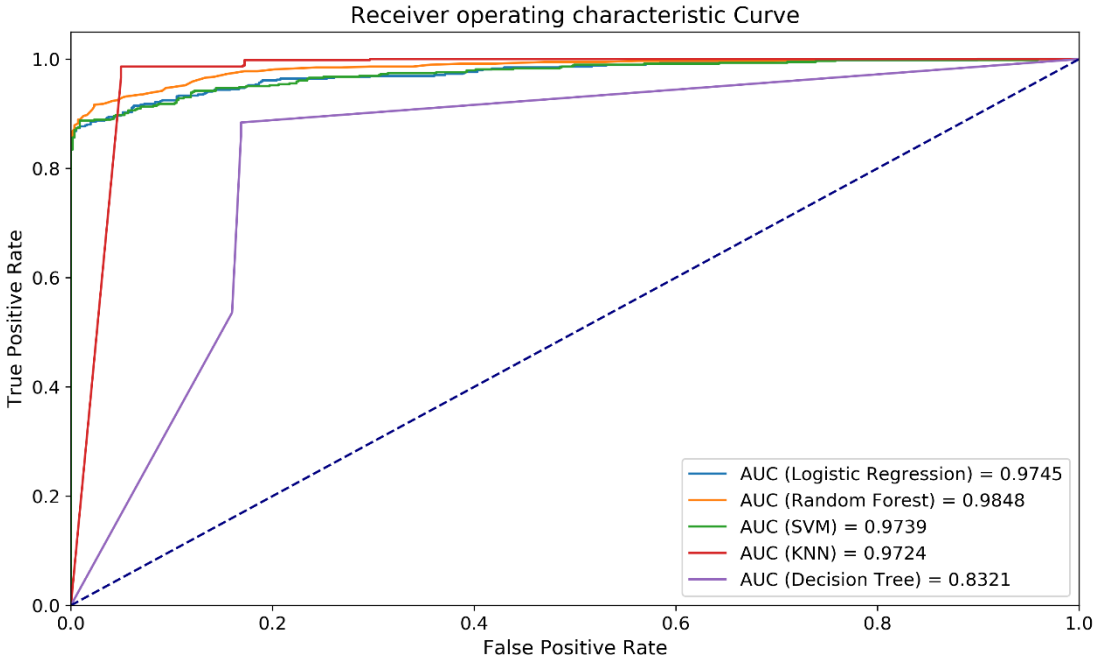
υποστήριξης (0.9348). Ωστόσο, ο kNN αλγόριθμος είχε την μεγαλύτερη ευαισθησία (0.9983) αλλά τη χαμηλότερη ακρίβεια (0.8357). Συνολικά, οι λάθος ταξινομήσεις κυμάνθηκαν από 5.7% έως 14.2%, ενώ για την κλάση 1 κυμάνθηκαν από 0.2% έως 11.6%.



Εικόνα 27. Πίνακες σύγχυσης των αλγορίθμων Logistic Regression, Random Forest, Support Vector Machines, k-Εγγύτεροι Γείτονες και Decision Tree στο σύνολο των δεδομένων με προσαρμοστική συνθετική μέθοδο δειγματοληψίας.

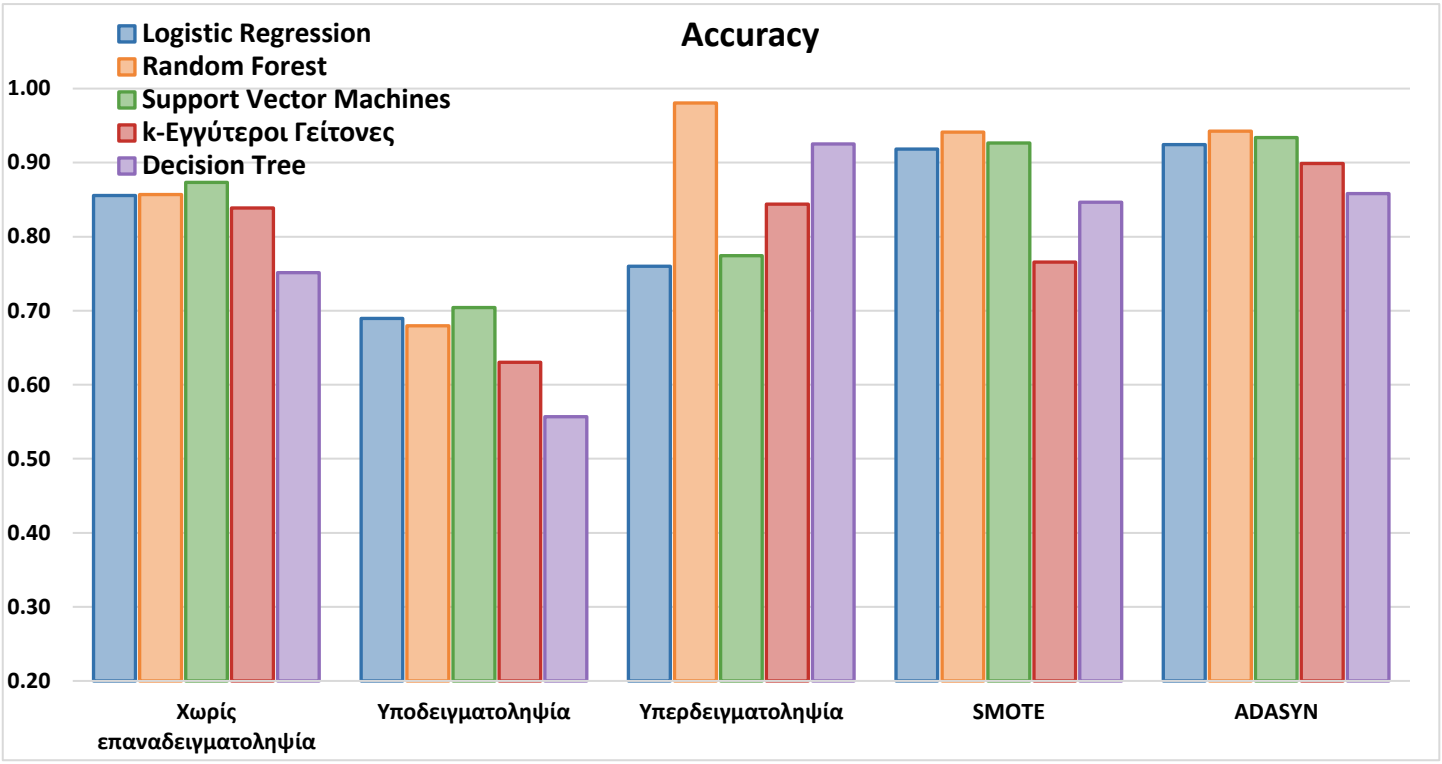
Πίνακας 7. Μετρικές αξιολόγησης των αλγορίθμων Logistic Regression, Random Forest, Support Vector Machines, k-Εγγύτεροι Γείτονες και Decision Tree στο σύνολο των δεδομένων με προσαρμοστική συνθετική μέθοδο δειγματοληψίας.

	Accuracy	Precision	Recall	f1-score	AUC
Logistic Regression	0.9243	0.9463	0.9027	0.9240	0.9745
Random Forest	0.9426	0.9676	0.9181	0.9422	0.9848
Support Vector Machines	0.9339	0.9811	0.8874	0.9319	0.9739
k-Εγγύτεροι Γείτονες	0.8990	0.8357	0.9983	0.9098	0.9724
Decision Tree	0.8581	0.8450	0.8840	0.8641	0.8321

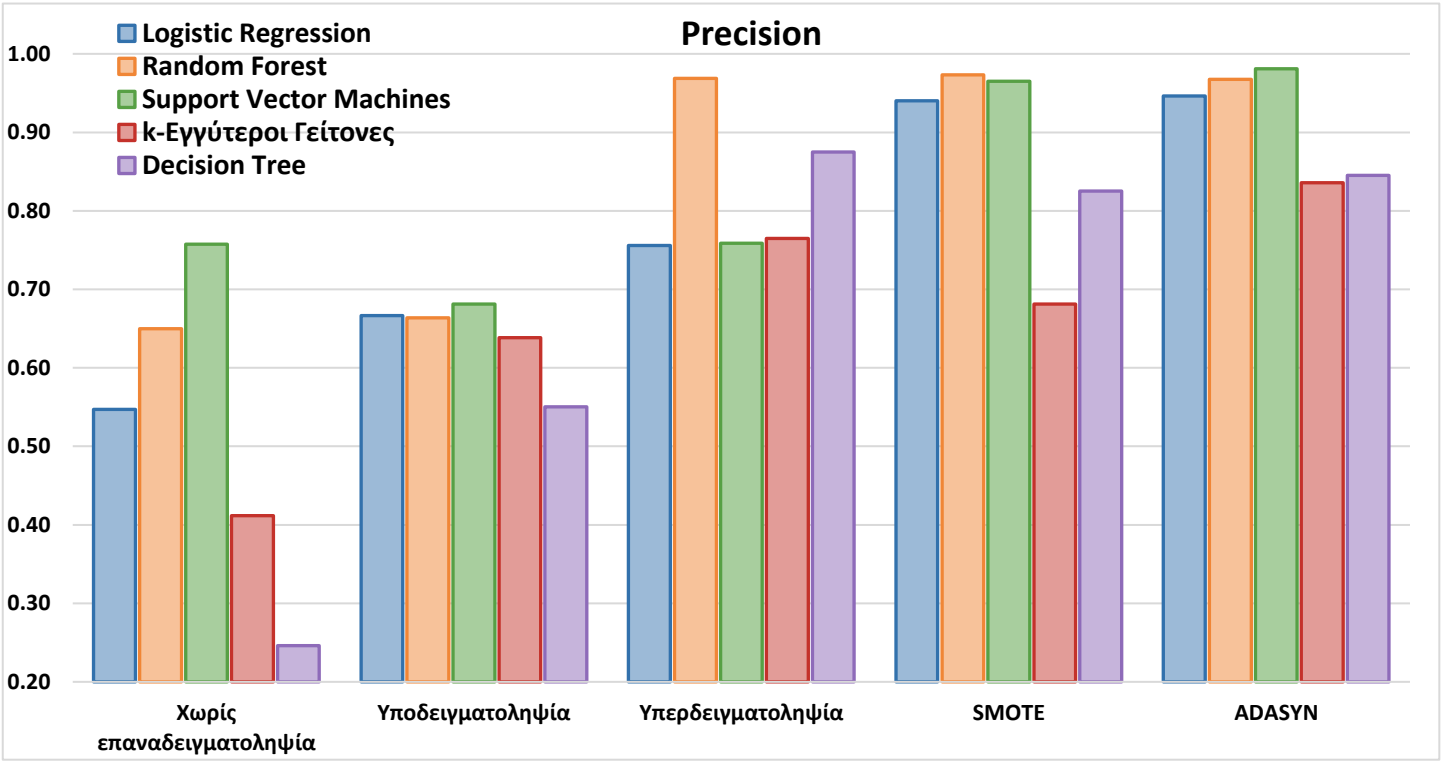


Εικόνα 28. Εμβαδόν κάτω από την καμπύλη ανά αλγόριθμο, την εφαρμογή κάποιας μεθόδου επαναδειγματοληψίας.

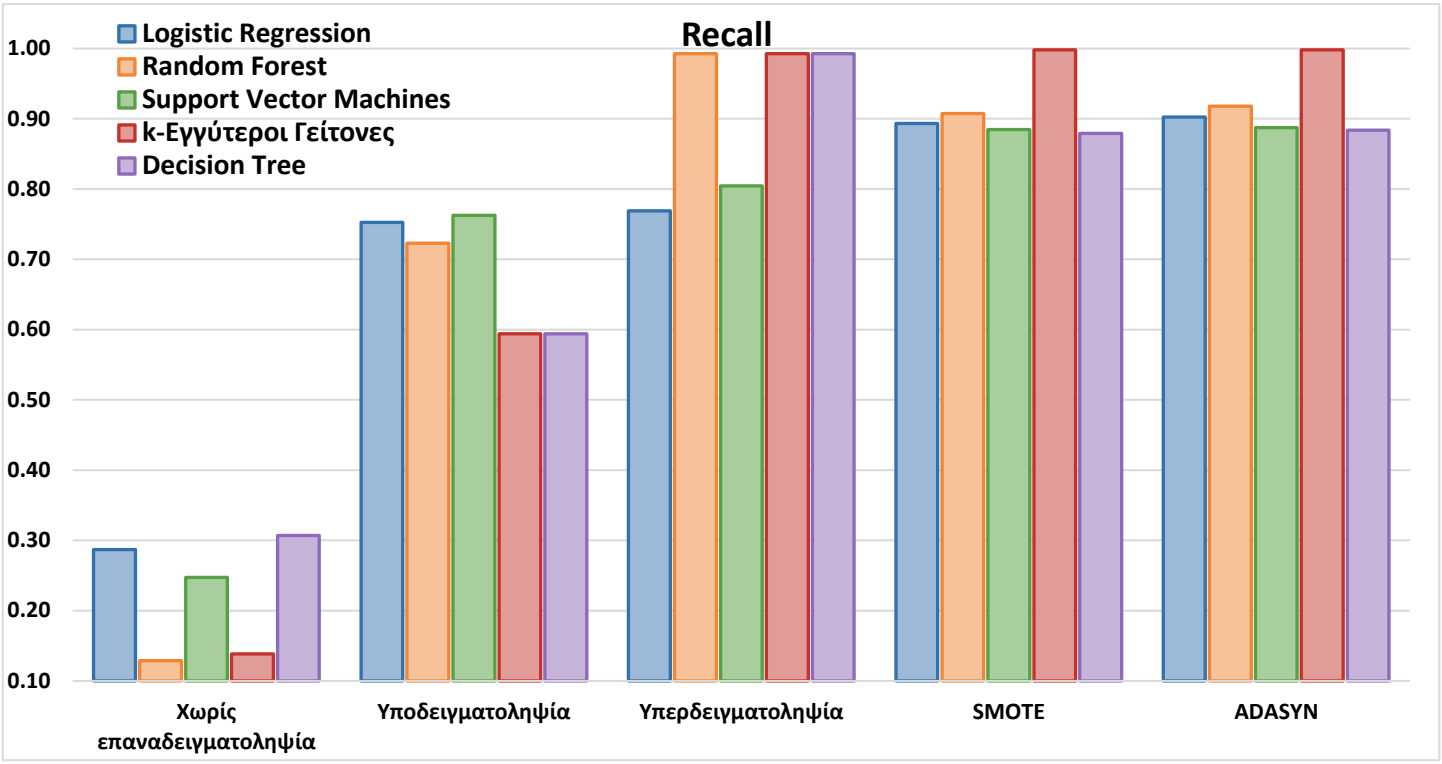
Για την καλύτερη παρουσίαση των αποτελεσμάτων, παρουσιάζονται οι μετρικές αξιολόγησης που χρησιμοποιήθηκαν σε όλες τις περιπτώσεις (Εικόνες 29-33).



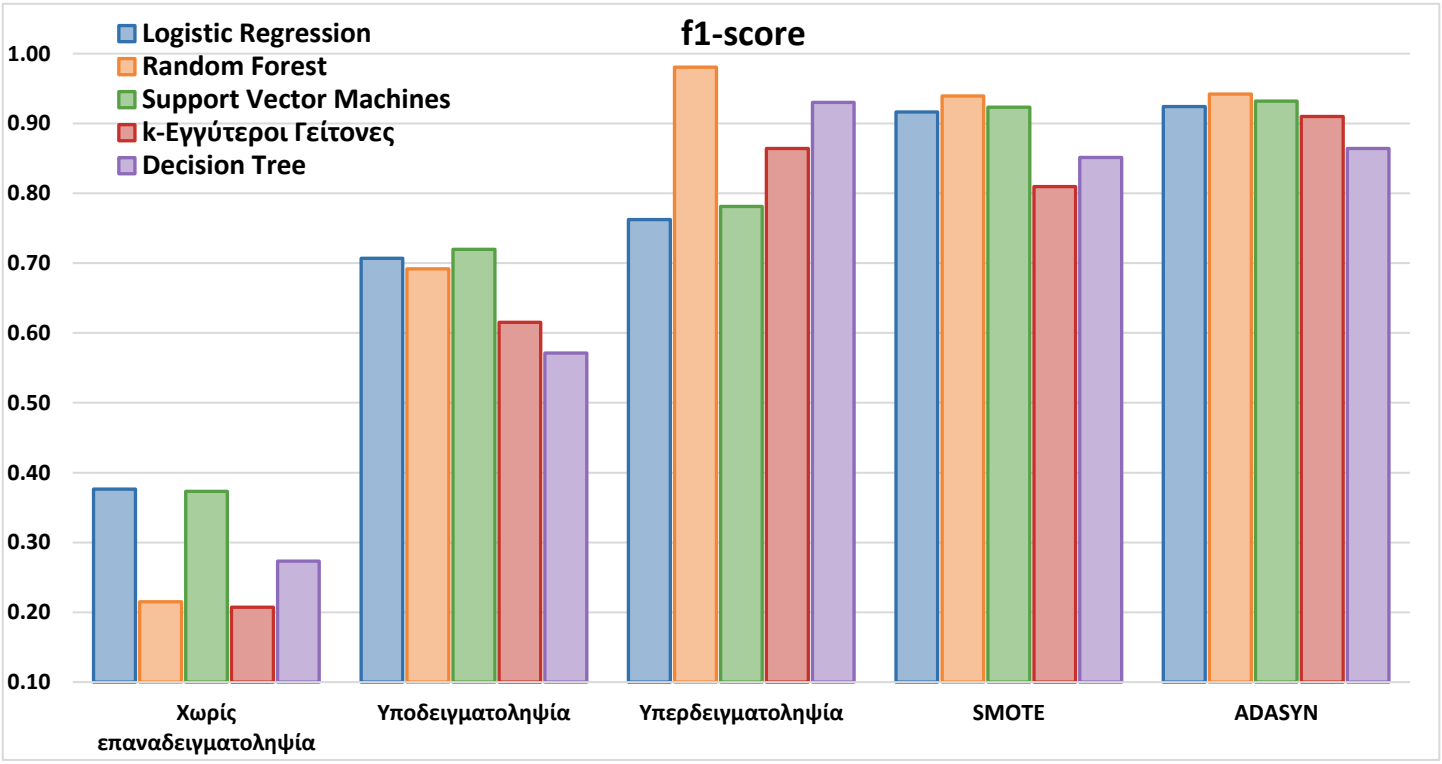
Εικόνα 29. Η ορθότητα των αλγορίθμων με ή χωρίς επαναδειγματοληψία.



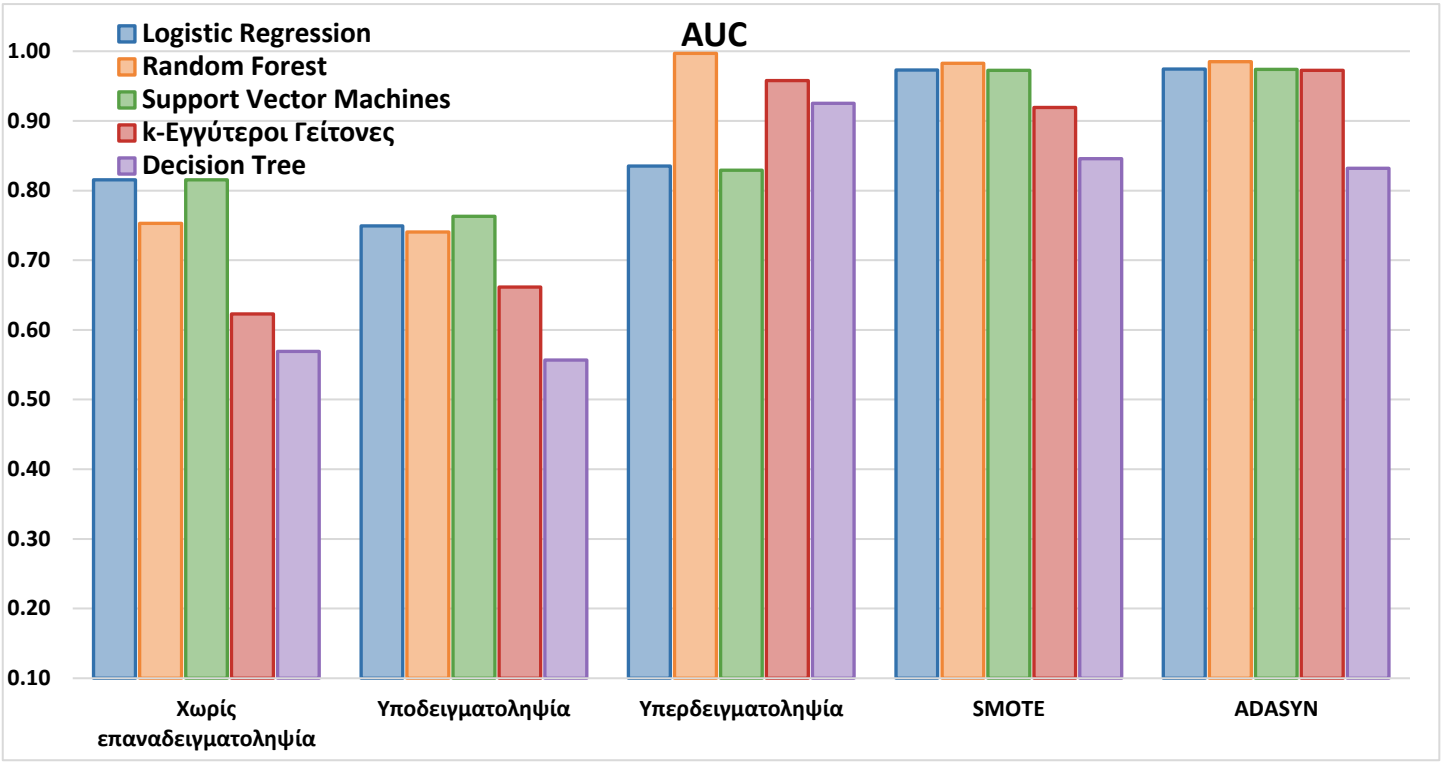
Εικόνα 30. Η ακρίβεια των αλγορίθμων με ή χωρίς επαναδειγματοληψία.



Εικόνα 31. Η ακρίβεια των αλγορίθμων με ή χωρίς επαναδειγματοληψία.



Εικόνα 32. f1-score των αλγορίθμων με ή χωρίς επαναδειγματοληψία.



Εικόνα 33. Εμβαδόν κάτω από την καμπύλη ROC των αλγορίθμων με ή χωρίς επαναδειγματοληψία.

6.5 Αποτελέσματα στο σύνολο των δεδομένων μετά την επιλογή χαρακτηριστικών

Στην παράγραφο αυτή παρουσιάζονται τα αποτελέσματα που προέκυψαν μετά την επιλογή χαρακτηριστικών, που έγινε με α) Univariate Method (SelectKBest), β) Wrapper Method (Backward Elimination) και γ) Embedded Method (Random Forest Importance).

6.5.1 Αποτελέσματα χωρίς εφαρμογή μεθόδων επαναδειγματοληψίας

Η ευαισθησία παρέμεινε σε χαμηλά επίπεδα και μετά την επιλογή χαρακτηριστικών, ανεξάρτητα από τη μέθοδο που ακολουθήθηκε. Ο αλγόριθμος Decision Tree είχε τη μέγιστη ευαισθησία: 0.3465, 0.3960 και 0.3366 με την επιλογή 20 καλύτερων χαρακτηριστικών, με τη μέθοδο της προς τα πίσω απαλοιφής και με τη μέθοδο

σημαντικότητας από τα τυχαία δάση, αντίστοιχα (Πίνακες 8-10). Ωστόσο και στις 3 περιπτώσεις ο αλγόριθμος Decision tree είχε το μικρότερο AUC score.

Πίνακας 8. Μετρικές αξιολόγησης των αλγορίθμων μετά την επιλογή των 20 καλύτερων χαρακτηριστικών.

	Accuracy	**Precision**	**Recall**	**f1-score**	**AUC**
Logistic Regression	0.8524	0.5429	0.1881	0.2794	0.7678
Random Forest	0.8358	0.3947	0.1485	0.2158	0.7110
Support Vector Machines	0.8479	0.0000	0.0000	0.0000	0.7659
k-Εγγύτεροι Γείτονες	0.8163	0.2941	0.1485	0.1974	0.6309
Decision Tree	0.7726	0.2917	0.3465	0.3167	0.5911

Πίνακας 9. Μετρικές αξιολόγησης των αλγορίθμων μετά την επιλογή χαρακτηριστικών με τη μέθοδος της προς τα πίσω απαλοιφής. Επιλέχθηκαν 34 χαρακτηριστικά.

	Accuracy	**Precision**	**Recall**	**f1-score**	**AUC**
Logistic Regression	0.8584	0.5686	0.2871	0.3816	0.8027
Random Forest	0.8539	0.5714	0.1584	0.2481	0.7506
Support Vector Machines	0.8614	0.8000	0.1188	0.2069	0.7983
k-Εγγύτεροι Γείτονες	0.8328	0.3214	0.0891	0.1395	0.6348
Decision Tree	0.7575	0.2857	0.3960	0.3320	0.6096

Πίνακας 10. Μετρικές αξιολόγησης των αλγορίθμων μετά την επιλογή χαρακτηριστικών με τη μέθοδος random forest importance. Επιλέχθηκαν 24 χαρακτηριστικά.

	Accuracy	**Precision**	**Recall**	**f1-score**	**AUC**
Logistic Regression	0.8494	0.5106	0.2376	0.3243	0.7788
Random Forest	0.8524	0.5652	0.1287	0.2097	0.7341
Support Vector Machines	0.8479	0.0000	0.0000	0.0000	0.7783
k-Εγγύτεροι Γείτονες	0.8449	0.4722	0.1683	0.2482	0.6195
Decision Tree	0.7485	0.2537	0.3366	0.2894	0.5836

6.5.2 Αποτελέσματα με εφαρμογή μεθόδων επαναδειγματοληψίας

6.5.2.1 Με τυχαία υποδειγματοληψία

Εφαρμόζοντας τυχαία υποδειγματοληψία, όπως ήταν αναμενόμενο βελτιώθηκαν οι επιδόσεις των αλγορίθμων. Η επιλογή χαρακτηριστικών δεν άλλαξε τη συνολική εικόνα των αλγορίθμων. Ο αλγόριθμος της λογιστικής παλινδρόμησης πέτυχε τη μεγαλύτερη ευαισθησία (0.7822) με τη μέθοδο της προς τα πίσω απαλοιφής χαρακτηριστικών, ενώ η ακρίβεια ήταν 0.6695 και το AUC score 0.7672.

Πίνακας 11. Μετρικές αξιολόγησης των αλγορίθμων μετά την επιλογή των 20 καλύτερων χαρακτηριστικών.

	Accuracy	**Precision**	**Recall**	**f1-score**	**AUC**
Logistic Regression	0.6798	0.6579	0.7426	0.6977	0.7459
Random Forest	0.6502	0.6364	0.6931	0.6635	0.7063
Support Vector Machines	0.6700	0.6574	0.7030	0.6794	0.7361
k-Εγγύτεροι Γείτονες	0.6453	0.6355	0.6733	0.6538	0.6735
Decision Tree	0.5764	0.5676	0.6238	0.5943	0.5729

Πίνακας 12. Μετρικές αξιολόγησης των αλγορίθμων μετά την επιλογή χαρακτηριστικών με τη μέθοδος της προς τα πίσω απαλοιφής. Επιλέχθηκαν 26 χαρακτηριστικά.

	Accuracy	**Precision**	**Recall**	**f1-score**	**AUC**
Logistic Regression	0.6995	0.6695	0.7822	0.7215	0.7672
RandoEφαm Forest	0.6798	0.6579	0.7426	0.6977	0.7565
Support Vector Machines	0.6847	0.6581	0.7624	0.7064	0.7658
k-Εγγύτεροι Γείτονες	0.6355	0.6392	0.6139	0.6263	0.6813
Decision Tree	0.6108	0.6078	0.6139	0.6108	0.6109

Πίνακας 13. Μετρικές αξιολόγησης των αλγορίθμων μετά την επιλογή χαρακτηριστικών με τη μέθοδος random forest importance. Επιλέχθηκαν 24 χαρακτηριστικά.

	Accuracy	Precision	Recall	f1-score	AUC
Logistic Regression	0.6946	0.6726	0.7525	0.7103	0.7288
Random Forest	0.6700	0.6518	0.7228	0.6854	0.7074
Support Vector Machines	0.6798	0.6698	0.7030	0.6860	0.7293
k-Εγγύτεροι Γείτονες	0.6158	0.6162	0.6040	0.6100	0.6593
Decision Tree	0.5320	0.5283	0.5545	0.5411	0.5321

6.5.2.2 Με τυχαία υπερδειγματοληψία

Με την επιλογή χαρακτηριστικών στην τυχαία υπερδειγματοληψία φάνηκε πως ο αλγόριθμός των τυχαίων δασών ξεχώρισε για τη συνολική του απόδοση, όταν εφαρμόστηκε η μέθοδος της προς τα πίσω απαλοιφής χαρακτηριστικών. Η μέθοδος επέλεξε 34 χαρακτηριστικά από τα 112, δηλαδή το 30.4% των χαρακτηριστικών. Η ευαισθησία του αλγορίθμου των τυχαίων δασών ήταν 0.9929, η ορθότητα 0.9751, η ακρίβεια 0.9588 και το AUC score 0.9967 (Πίνακας 14).

Πίνακας 14. Μετρικές αξιολόγησης των αλγορίθμων μετά την επιλογή των 20 καλύτερων χαρακτηριστικών.

	Accuracy	Precision	Recall	f1-score	AUC
Logistic Regression	0.7078	0.7082	0.7069	0.7076	0.7694
Random Forest	0.9254	0.8784	0.9876	0.9298	0.9902
Support Vector Machines	0.6972	0.7063	0.6750	0.6903	0.7663
k-Εγγύτεροι Γείτονες	0.8259	0.7503	0.9769	0.8488	0.9491
Decision Tree	0.8961	0.8369	0.9840	0.9045	0.9070

Πίνακας 15. Μετρικές αξιολόγησης των αλγορίθμων μετά την επιλογή χαρακτηριστικών με τη μέθοδος της προς τα πίσω απαλοιφής. Επιλέχθηκαν 34 χαρακτηριστικά.

	Accuracy	**Precision**	**Recall**	**f1-score**	**AUC**
Logistic Regression	0.7638	0.7610	0.7691	0.7650	0.8338
Random Forest	0.9751	0.9588	0.9929	0.9756	0.9967
Support Vector Machines	0.7682	0.7568	0.7904	0.7732	0.8301
k-Εγγύτεροι Γείτονες	0.8517	0.7727	0.9964	0.8704	0.9651
Decision Tree	0.9165	0.8693	0.9805	0.9215	0.9165

Πίνακας 16. Μετρικές αξιολόγησης των αλγορίθμων μετά την επιλογή χαρακτηριστικών με τη μέθοδος random forest importance. Επιλέχθηκαν 24 χαρακτηριστικά.

	Accuracy	**Precision**	**Recall**	**f1-score**	**AUC**
Logistic Regression	0.7043	0.7091	0.6927	0.7008	0.8338
Random Forest	0.9618	0.9348	0.9929	0.9630	0.9967
Support Vector Machines	0.6963	0.7097	0.6643	0.6862	0.8301
k-Εγγύτεροι Γείτονες	0.8295	0.7490	0.9911	0.8532	0.9651
Decision Tree	0.9174	0.8660	0.9876	0.9228	0.9165

6.5.2.3 Με τεχνική υπερδειγματοληψίας συνθετικής μειονότητας (SMOTE)

Οι αλγόριθμοι είχαν υψηλά ποσοστά ορθής ταξινόμησης, επιλέγοντας χαρακτηριστικά μετά την υπερδειγματοληψία συνθετικής μειονότητας. Για την επιλογή χαρακτηριστικών φάνηκε πως η μέθοδος της προς τα πίσω επιλογής χαρακτηριστικών είχε καλύτερα αποτελέσματα σε σχέση με την εφαρμογή των μεθόδων της επιλογής των 20 καλύτερων χαρακτηριστικών και της σημαντικότητας των χαρακτηριστικών από τα τυχαία δάση (Πίνακες 16-18). Η μέγιστη ευαισθησία επετεύχθη από τον αλγόριθμο KNN (0.9982), ενώ το καλύτερο AUC score το είχε ο αλγόριθμος Random forest (0.9819), έναντι 0.9484 που είχε ο αλγόριθμος KNN (Πίνακας 17).

Πίνακας 17. Μετρικές αξιολόγησης των αλγορίθμων μετά την επιλογή των 20 καλύτερων χαρακτηριστικών.

	Accuracy	**Precision**	**Recall**	**f1-score**	**AUC**
Logistic Regression	0.8073	0.8157	0.7940	0.8047	0.9042
Random Forest	0.9192	0.9370	0.8988	0.9175	0.9656
Support Vector Machines	0.8037	0.8076	0.7975	0.8025	0.8983
k-Εγγύτεροι Γείτονες	0.8393	0.7851	0.9343	0.8532	0.9143
Decision Tree	0.8481	0.8345	0.8686	0.8512	0.8470

Πίνακας 18. Μετρικές αξιολόγησης των αλγορίθμων μετά την επιλογή χαρακτηριστικών με τη μέθοδος της προς τα πίσω απαλοιφής. Επιλέχθηκαν 70 χαρακτηριστικά.

	Accuracy	**Precision**	**Recall**	**f1-score**	**AUC**
Logistic Regression	0.9174	0.9384	0.8934	0.9154	0.9718
Random Forest	0.9423	0.9734	0.9094	0.9403	0.9819
Support Vector Machines	0.9254	0.9687	0.8792	0.9218	0.9713
k-Εγγύτεροι Γείτονες	0.8020	0.7168	0.9982	0.8344	0.9484
Decision Tree	0.8419	0.8182	0.8792	0.8476	0.8412

Πίνακας 19. Μετρικές αξιολόγησης των αλγορίθμων μετά την επιλογή χαρακτηριστικών με τη μέθοδος random forest importance. Επιλέχθηκαν 24 χαρακτηριστικά.

	Accuracy	**Precision**	**Recall**	**f1-score**	**AUC**
Logistic Regression	0.7549	0.7461	0.7726	0.7592	0.8315
Random Forest	0.9352	0.9554	0.9130	0.9337	0.9785
Support Vector Machines	0.7513	0.7461	0.7620	0.7540	0.8274
k-Εγγύτεροι Γείτονες	0.7762	0.6917	0.9964	0.8166	0.9222
Decision Tree	0.8561	0.8370	0.8845	0.8601	0.8561

6.5.2.4 Με προσαρμοστική συνθετική μέθοδο δειγματοληψίας (ADASYN)

Οι αλγόριθμοι είχαν υψηλά ποσοστά ορθής ταξινόμησης, επιλέγοντας χαρακτηριστικά μετά την υπερδειγματοληψία συνθετικής μειονότητας με εύρος 0.7952 – 0.9949 (Πίνακες 19-21). Το μεγαλύτερο AUC και f1 score, το πέτυχε ο αλγόριθμος των τυχαίων δασών με τη μέθοδο της προς τα πίσω επιλογής χαρακτηριστικών, επιλέγοντας 68 από τα 112 χαρακτηριστικά.

Πίνακας 20. Μετρικές αξιολόγησης των αλγορίθμων μετά την επιλογή των 20 καλύτερων χαρακτηριστικών.

	Accuracy	**Precision**	**Recall**	**f1-score**	**AUC**
Logistic Regression	0.8146	0.8134	0.8259	0.8196	0.9120
Random Forest	0.9156	0.9312	0.9010	0.9159	0.9664
Support Vector Machines	0.8146	0.8166	0.8208	0.8187	0.9078
k-Εγγύτεροι Γείτονες	0.8964	0.8689	0.9386	0.9024	0.9493
Decision Tree	0.8573	0.8470	0.8788	0.8626	0.8570

Πίνακας 21. Μετρικές αξιολόγησης των αλγορίθμων μετά την επιλογή χαρακτηριστικών με τη μέθοδος της προς τα πίσω απαλοιφής. Επιλέχθηκαν 68 χαρακτηριστικά.

	Accuracy	**Precision**	**Recall**	**f1-score**	**AUC**
Logistic Regression	0.9234	0.9431	0.9044	0.9233	0.9718
Random Forest	0.9417	0.9659	0.9181	0.9414	0.9819
Support Vector Machines	0.9304	0.9738	0.8874	0.9286	0.9713
k-Εγγύτεροι Γείτονες	0.9034	0.8437	0.9949	0.9131	0.9713
Decision Tree	0.8616	0.8362	0.9061	0.8698	0.8607

Πίνακας 22. Μετρικές αξιολόγησης των αλγορίθμων μετά την επιλογή χαρακτηριστικών με τη μέθοδος random forest importance. Επιλέχθηκαν 24 χαρακτηριστικά.

	Accuracy	**Precision**	**Recall**	**f1-score**	**AUC**
Logistic Regression	0.7694	0.7610	0.7986	0.7794	0.8378
Random Forest	0.9321	0.9456	0.9198	0.9325	0.9801
Support Vector Machines	0.7685	0.7614	0.7952	0.7780	0.8371
k-Εγγύτεροι Γείτονες	0.9025	0.8425	0.9949	0.9124	0.9656
Decision Tree	0.8729	0.8526	0.9078	0.8793	0.8722

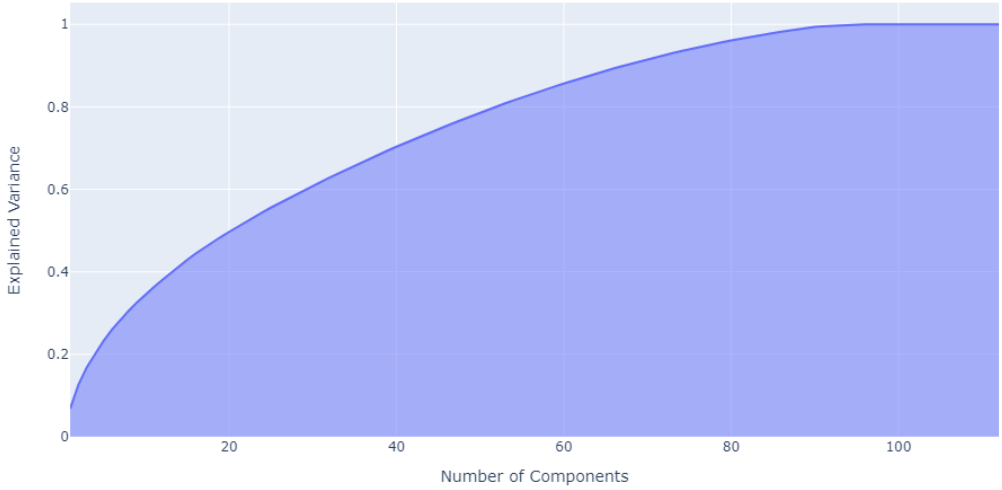
6.6 Αποτελέσματα από ανάλυση κύριων συνιστωσών

Στο τρίτο μέρος των αποτελεσμάτων παρουσιάζονται οι επιδόσεις των αλγορίθμων μετά από ανάλυση σε κύριες συνιστώσες. Η επιλογή του αριθμού των κύριων συνιστωσών έγινε με το κριτήριο του «Ποσοστού συνολικής διακύμανσης που εξηγούν οι κύριες συνιστώσες». Κάθε φορά επιλέγονταν k κύριες συνιστώσες, ώστε να εξηγούν το 96% της μεταβλητότητας. Ενδεικτικά όμως αναφέρεται και ο αριθμός των κύριων συνιστωσών που προκύπτουν και από το κριτήριο του Kaiser.

6.6.1 Αποτελέσματα χωρίς εφαρμογή μεθόδων επαναδειγματοληψίας

Για να μπορέσει να εξηγηθεί το 96% της μεταβλητότητας χρειάστηκε να συμπεριληφθούν 80 συνιστώσες (Εικόνα 34). Με βάση το κριτήριο του Kaiser, θα έπρεπε να επιλεχθούν οι 98 πρώτες κύριες συνιστώσες.

Η ευαισθησία των αλγορίθμων ήταν ιδιαίτερα χαμηλή και κυμάνθηκε από 5.9% έως 29.7% (Πίνακας 23).



Εικόνα 34. Γραφική αναπαράσταση του ποσοστού της διακύμανσης που εξηγείται ανάλογα με τον αριθμό των κύριων συνιστωσών χωρίς επαναδειγματοληψία.

Πίνακας 23. Μετρικές αξιολόγησης των αλγορίθμων μετά μείωση χαρακτηριστικών με τη μέθοδο PCA.

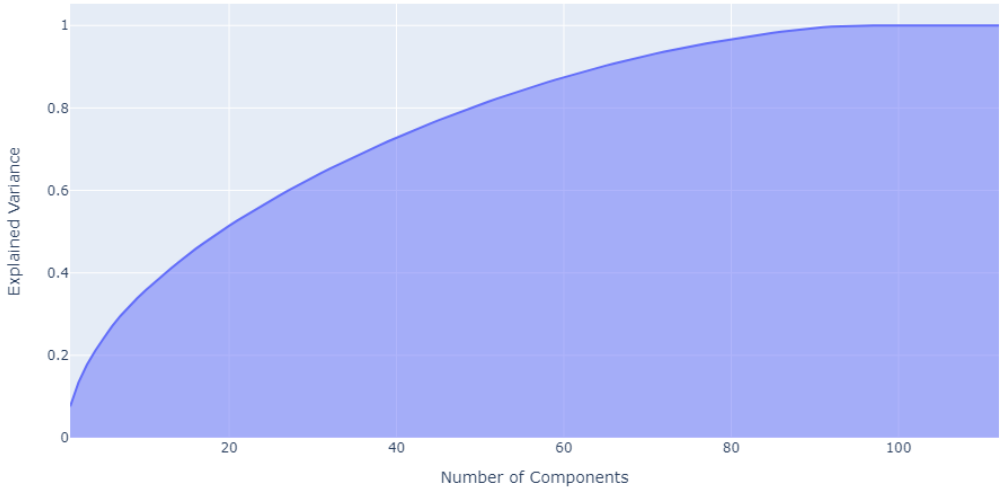
	Accuracy	Precision	Recall	f1-score	AUC
Logistic Regression	0.8569	0.5556	0.2970	0.3871	0.8108
Random Forest	0.8509	0.6000	0.0594	0.1081	0.7302
Support Vector Machines	0.8705	0.8000	0.1980	0.3175	0.8101
k-Εγγύτεροι Γείτονες	0.8373	0.4000	0.1386	0.2059	0.6493
Decision Tree	0.7636	0.2358	0.2475	0.2415	0.4789

6.6.2 Αποτελέσματα με εφαρμογή μεθόδων επαναδειγματοληψίας

6.6.2.1 Με τυχαία υποδειγματοληψία

Με τη μέθοδο της τυχαίας υποδειγματοληψίας, ο αριθμός των συνιστωσών που χρειάστηκε για να εξηγήσουν το 96% της μεταβλητότητας ήταν 79 (Εικόνα 35). Με βάση το κριτήριο του Kaiser, θα έπρεπε να επιλεχθούν οι 95 πρώτες κύριες συνιστώσες.

Η ευαισθησία των αλγορίθμων κυμάνθηκε από 49.5% έως 76.2% (Πίνακας 24), με τον αλγόριθμο SVM να υπερέχει έναντι των υπολοίπων (0.7624). Ακολουθεί η λογιστική παλινδρόμηση με ευαισθησία 0.7525 και AUC score 0.7540, που ήταν και το μεγαλύτερο μεταξύ των πέντε αλγορίθμων.



Εικόνα 35. Γραφική αναπαράσταση του ποσοστού της διακύμανσης που εξηγείται ανάλογα με τον αριθμό των κύριων συνιστωσών με τυχαία υποδειγματοληψία.

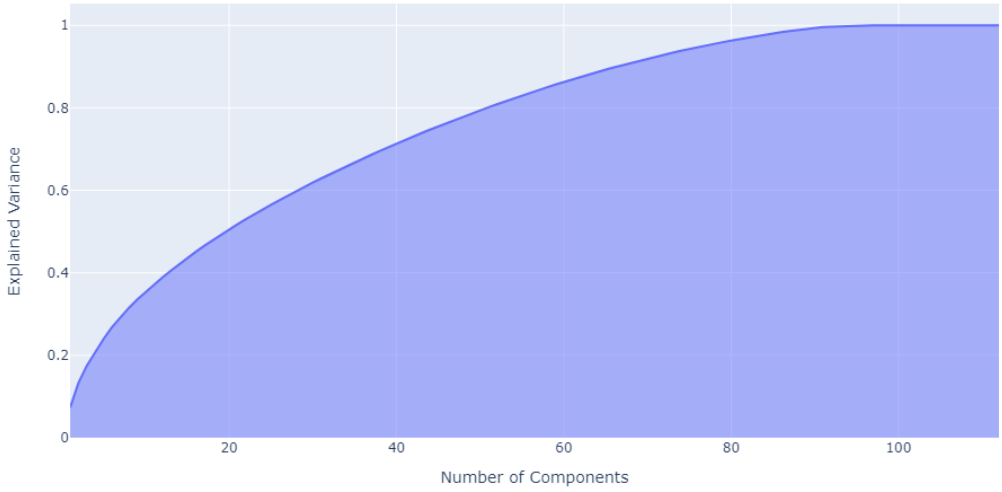
Πίνακας 24. Μετρικές αξιολόγησης των αλγορίθμων μετά μείωση χαρακτηριστικών με τη μέθοδο PCA.

	Accuracy	**Precision**	**Recall**	**f1-score**	**AUC**
Logistic Regression	0.6946	0.6726	0.7525	0.7103	0.7540
Random Forest	0.6946	0.6893	0.7030	0.6961	0.7252
Support Vector Machines	0.6946	0.6696	0.7624	0.7130	0.7349
k-Εγγύτεροι Γείτονες	0.6946	0.7010	0.6733	0.6869	0.7262
Decision Tree	0.5222	0.5208	0.4950	0.5076	0.5220

6.6.2.2 Με τυχαία υπερδειγματοληψία

Με τη μέθοδο της τυχαίας υπερδειγματοληψίας, ο αριθμός των συνιστωσών που χρειάστηκε για να εξηγήσουν το 96% της μεταβλητότητας ήταν 80 (Εικόνα 36), ενώ με το κριτήριο του Kaiser, θα έπρεπε να επιλεχθούν οι 98 πρώτες κύριες συνιστώσες.

Η ευαισθησία των αλγορίθμων ήταν πάνω από 74.8%, ενώ ο αλγόριθμος Decision tree είχε τη μεγαλύτερη ευαισθησία 98.8% (Πίνακας 25). Το υψηλότερο AUC score επετεύχθη από τον αλγόριθμο Random forest (0.9946).



Εικόνα 36. Γραφική αναπαράσταση του ποσοστού της διακύμανσης που εξηγείται ανάλογα με τον αριθμό των κύριων συνιστωσών με τυχαία υπερδειγματοληψία.

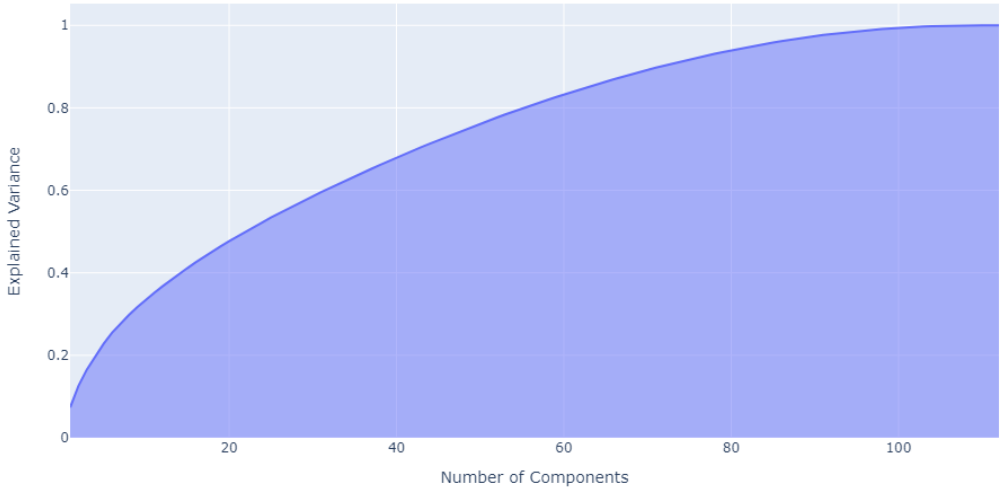
Πίνακας 25. Μετρικές αξιολόγησης των αλγορίθμων μετά μείωση χαρακτηριστικών με τη μέθοδο PCA.

	Accuracy	**Precision**	**Recall**	**f1-score**	**AUC**
Logistic Regression	0.7567	0.7604	0.7496	0.7549	0.8285
Random Forest	0.9849	0.9892	0.9805	0.9848	0.9946
Support Vector Machines	0.7540	0.7572	0.7478	0.7525	0.8258
k-Εγγύτεροι Γείτονες	0.8481	0.7753	0.9805	0.8659	0.9624
Decision Tree	0.8970	0.8361	0.9876	0.9055	0.8923

6.6.2.3 Με τεχνική υπερδειγματοληψίας συνθετικής μειονότητας (SMOTE)

Με τη μέθοδο της τυχαίας υπερδειγματοληψίας, ο αριθμός των συνιστωσών που χρειάστηκε για να εξηγήσουν το 96% της μεταβλητότητας ήταν 86 (Εικόνα 37), ενώ με το κριτήριο του Kaiser, θα έπρεπε να επιλεχθούν οι 109 πρώτες κύριες συνιστώσες.

Η ευαισθησία των αλγορίθμων ήταν πάνω από 83.7% με υψηλότερη εκείνη του k-NN (97.2%). Επιπλέον, το AUC-score κυμάνθηκε από 0.7461 έως και 0.9699 που είχε ο αλγόριθμος της λογιστικής παλινδρόμησης.



Εικόνα 37. Γραφική αναπαράσταση του ποσοστού της διακύμανσης που εξηγείται ανάλογα με τον αριθμό των κύριων συνιστωσών με τυχαία υπερδειγματοληψία συνθετικής μειονό9τητας.

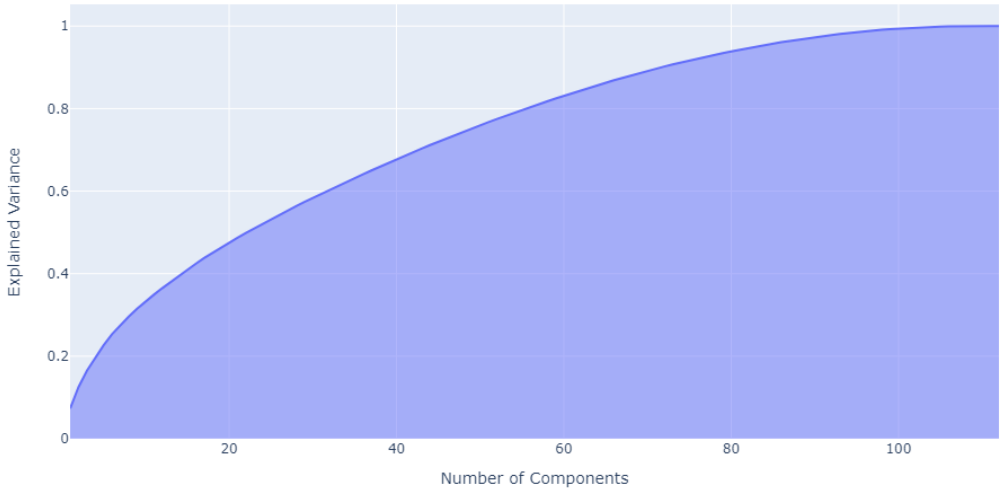
Πίνακας 26. Μετρικές αξιολόγησης των αλγορίθμων μετά μείωση χαρακτηριστικών με τη μέθοδο PCA.

	Accuracy	**Precision**	**Recall**	**f1-score**	**AUC**
Logistic Regression	0.9139	0.9299	0.8952	0.9122	0.9699
Random Forest	0.9174	0.9590	0.8721	0.9135	0.9612
Support Vector Machines	0.9165	0.9467	0.8828	0.9136	0.9695
k-Εγγύτεροι Γείτονες	0.8348	0.7629	0.9716	0.8547	0.9555
Decision Tree	0.7886	0.7634	0.8366	0.7983	0.7461

6.6.2.4 Με προσαρμοστική συνθετική μέθοδο δειγματοληψίας (ADASYN)

Με τη μέθοδο της τυχαίας υπερδειγματοληψίας, ο αριθμός των συνιστωσών που χρειάστηκε για να εξηγήσουν το 96% της μεταβλητότητας ήταν 86 (Εικόνα 38), ενώ με το κριτήριο του Kaiser, θα έπρεπε να επιλεχθούν οι 109 πρώτες κύριες συνιστώσες.

Η ευαισθησία των αλγορίθμων κυμάνθηκε από 0.8345 – 0.9744., ενώ το AUC-score από 0.7982 – 0.9250. Οι αλγόριθμοι με τη μεγαλύτερη ευαισθησία και AUC score ήταν ο KNN και η λογιστική παλινδρόμηση, αντίστοιχα.

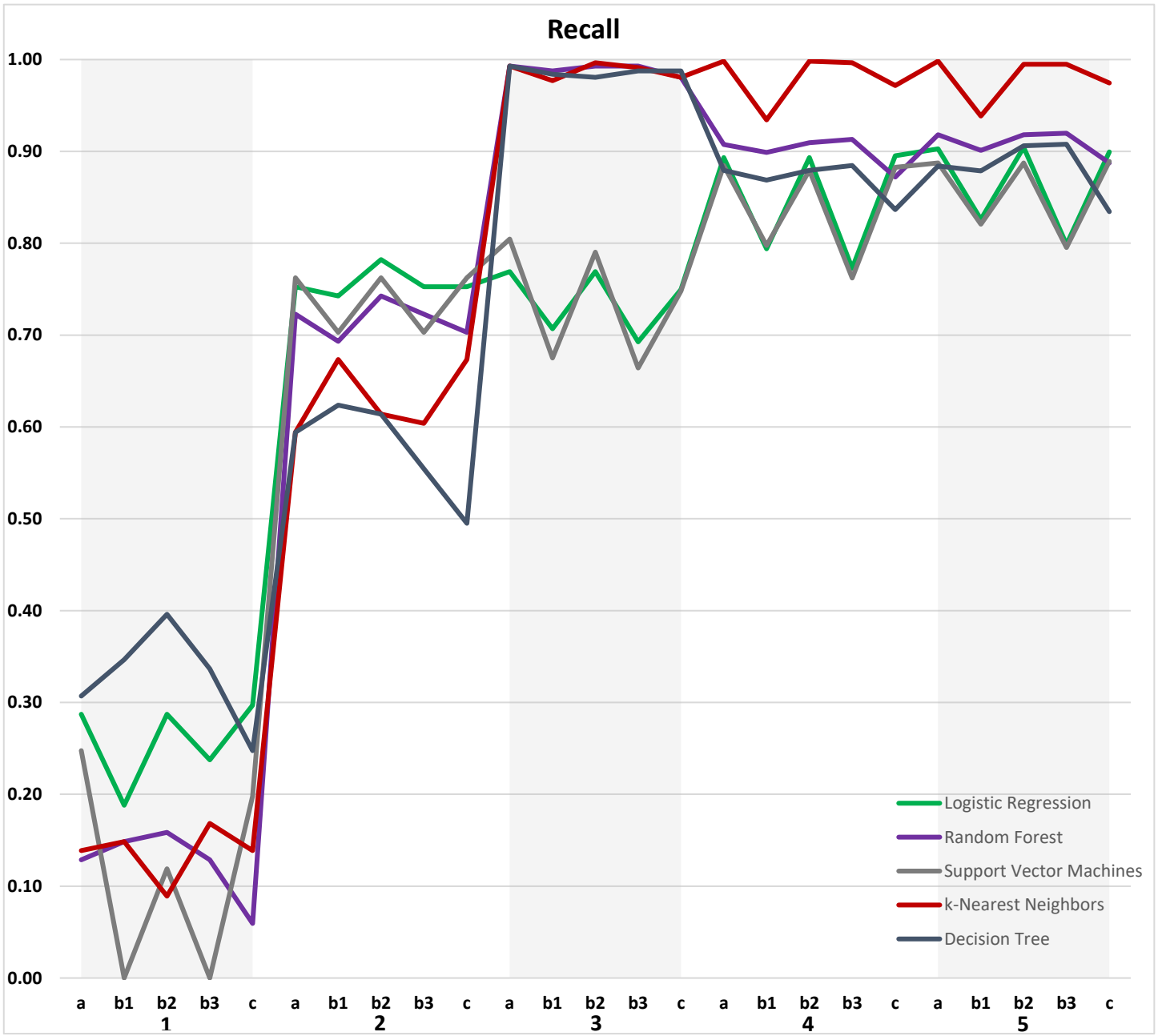


Εικόνα 38. Γραφική αναπαράσταση του ποσοστού της διακύμανσης που εξηγείται ανάλογα με τον αριθμό των κύριων συνιστωσών με προσαρμοστική συνθετική υπερδειγματοληψία.

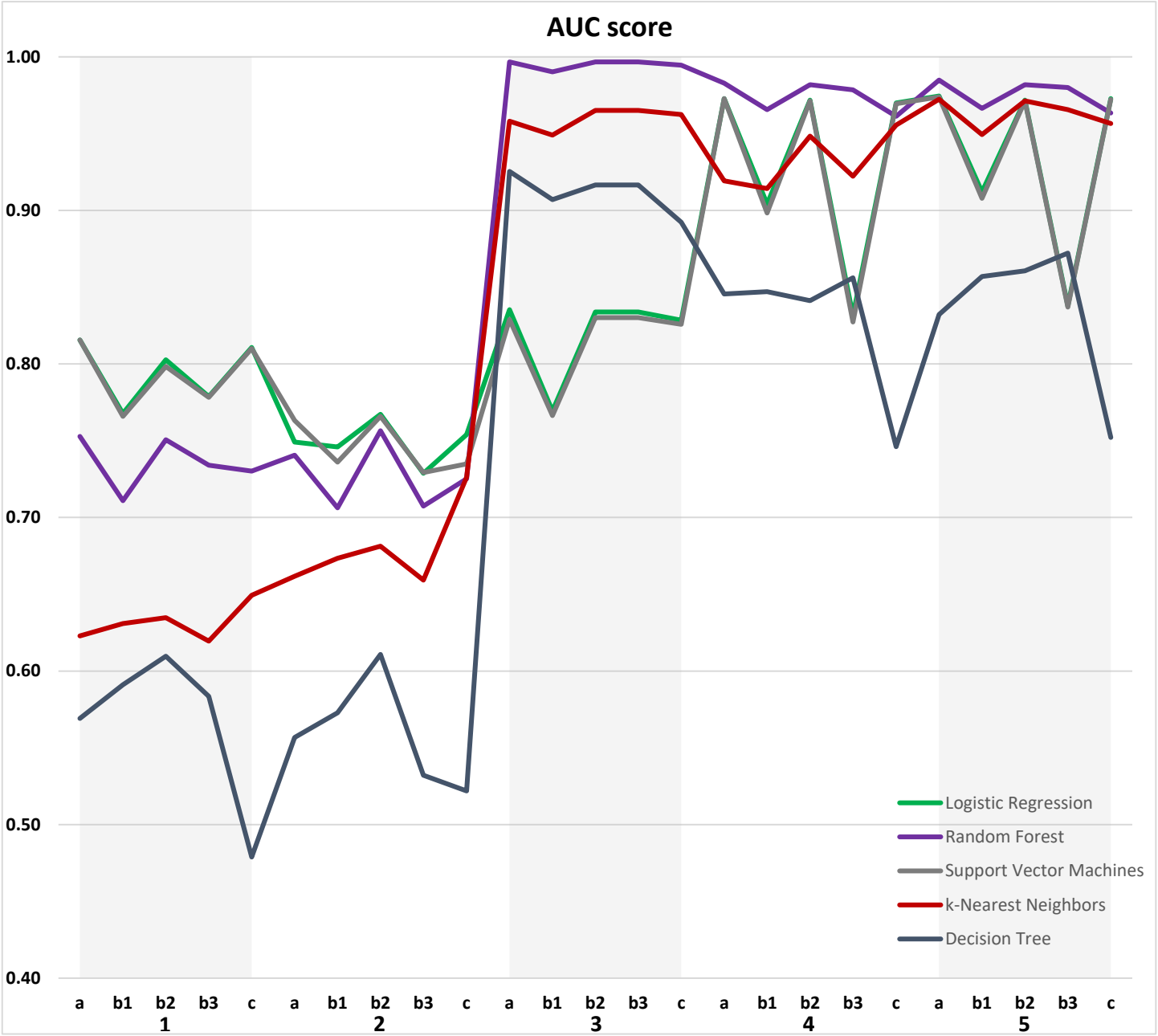
Πίνακας 27. Μετρικές αξιολόγησης των αλγορίθμων μετά μείωση χαρακτηριστικών με τη μέθοδο PCA.

	Accuracy	**Precision**	**Recall**	**f1-score**	**AUC**
Logistic Regression	0.9199	0.9411	0.8993	0.9197	0.9729
Random Forest	0.9243	0.9612	0.8874	0.9228	0.9634
Support Vector Machines	0.9243	0.9595	0.8891	0.9229	0.9723
k-Εγγύτεροι Γείτονες	0.8599	0.7964	0.9744	0.8764	0.9566
Decision Tree	0.7990	0.7849	0.8345	0.8089	0.7521

Τέλος, στις Εικόνες 39 και 40, παρουσιάζονται η ευαισθησία και το AUC score των ταξινομητών όλων των παραπάνω περιπτώσεων.

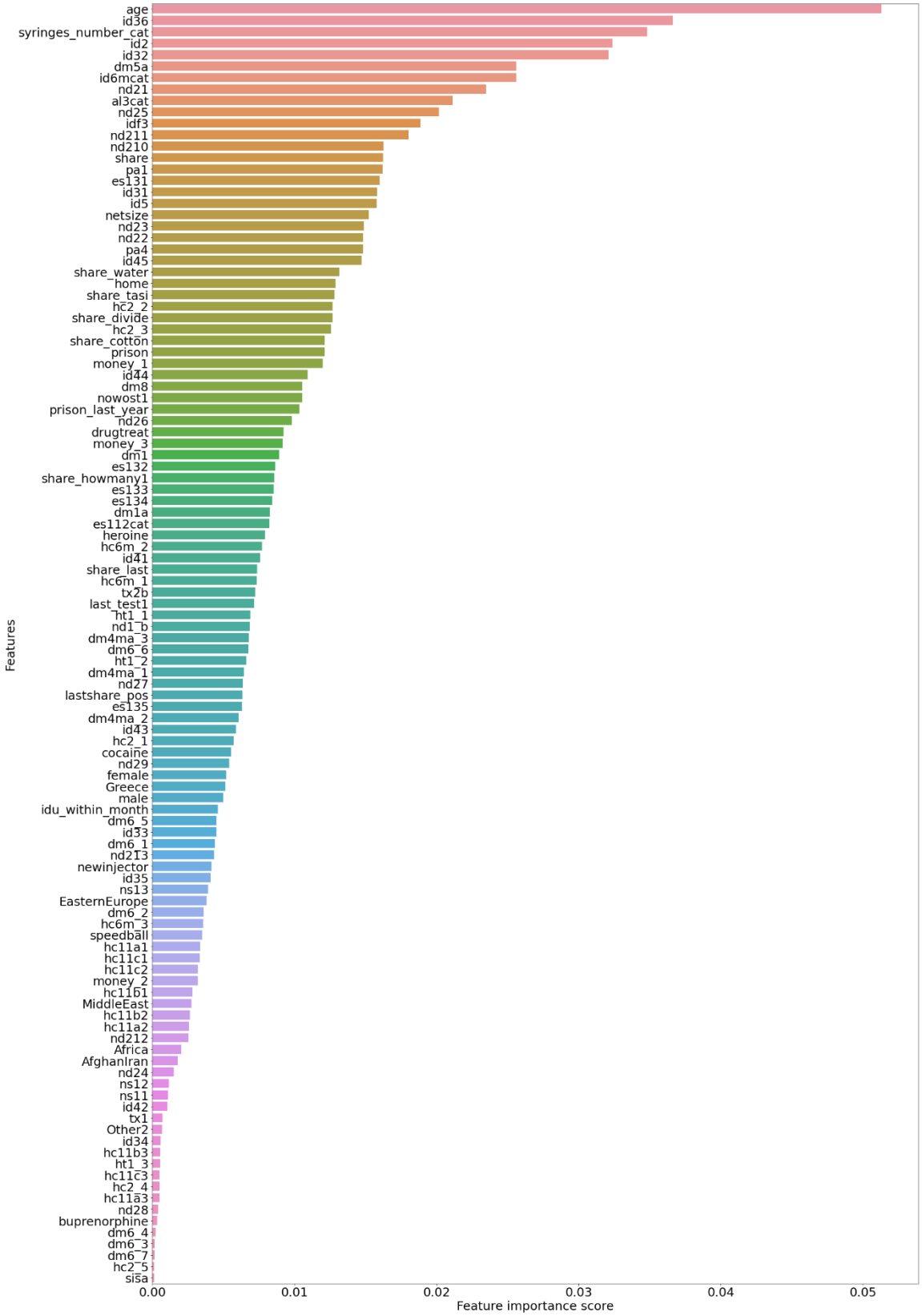


Εικόνα 39. Ευαισθησία των ταξινομητών στα τρία βασικά σενάρια: a) σε όλα τα χαρακτηριστικά, b) μετά την επιλογή των χαρακτηριστικών με τρεις διαφορετικές μεθόδους (1: επιλογή των 20 καλύτερων χαρακτηριστικών, 2: προς τα πίσω απαλοιφή χαρακτηριστικών και 3: random forest importance) και c) μετά από ανάλυση σε κύριες συνιστώσες στις πέντε περιπτώσεις: 1) Δεδομένα χωρίς επαναδειγματοληψία, 2) Δεδομένα με υποδειγματοληψία, 3) Δεδομένα με τυχαία υπερδειγματοληψία, 4) Δεδομένα με τεχνική υπερδειγματοληψίας συνθετικής μειονότητας και 5) Δεδομένα με προσαρμοστική συνθετική μέθοδο δειγματοληψίας.



Εικόνα 40. AUC score των ταξινομητών στα τρία βασικά σενάρια: a) σε όλα τα χαρακτηριστικά, b) μετά την επιλογή των χαρακτηριστικών με τρεις διαφορετικές μεθόδους (1: επιλογή των 20 καλύτερων χαρακτηριστικών, 2: προς τα πίσω απαλοιφή χαρακτηριστικών και 3: random forest importance) και c) μετά από ανάλυση σε κύριες συνιστώσες στις πέντε περιπτώσεις: 1) Δεδομένα χωρίς επαναδειγματοληψία, 2) Δεδομένα με υποδειγματοληψία, 3) Δεδομένα με τυχαία υπερδειγματοληψία, 4) Δεδομένα με τεχνική υπερδειγματοληψίας συνθετικής μειονότητας και 5) Δεδομένα με προσαρμοστική συνθετική μέθοδο δειγματοληψίας.

6.7 Κατάταξη χαρακτηριστικών



Εικόνα 41. Κατάταξη χαρακτηριστικών βάσει του αλγορίθμου Random Forest (Gini importance) στην περίπτωση της τυχαίας υπερδειγματοληψίας

7. Συμπεράσματα

Εφαρμόστηκαν για πρώτη φορά στην Ελλάδα, αλγόριθμοι μηχανικής μάθησης για την ταξινόμηση των χρηστών ενδοφλέβιων ναρκωτικών της Αθήνας ως προς την HIV λοίμωξη, την περίοδο 2012-2013, που υπήρχε επιδημική έκρηξη HIV στον πληθυσμό αυτόν. Χρησιμοποιήθηκαν πέντε αλγόριθμοι (Logistic Regression, Random Forest, Support Vector Machines, k-Εγγύτεροι Γείτονες και Decision Tree) κάτω από τρία βασικά σενάρια (1: όλα τα χαρακτηριστικά, 2: με επιλογή χαρακτηριστικών και 3: με ανάλυση σε κύριες συνιστώσες) με πέντε περιπτώσεις το καθένα (1: χωρίς μεθόδους επαναδειγματοληψίας, 2: με τυχαία υποδειγματοληψία, 3: με τυχαία υπερδειγματοληψία, 4: με την τεχνική υπερδειγματοληψίας συνθετικής μειονότητας και 5: με την προσαρμοστική συνθετική μέθοδο δειγματοληψίας).

Όταν εφαρμόστηκαν τα τρία σενάρια στα αρχικά δεδομένα (δηλαδή σε δεδομένα χωρίς επαναδειγματοληψία), η ορθότητα κυμάνθηκε σε ικανοποιητικά επίπεδα (0.7485 – 0.8735), ενώ το AUC score κυμάνθηκε από 0.4789 έως 0.8156. Η ευαισθησία ήταν ιδιαίτερα χαμηλή (0.0000 – 0.3960), που ήταν αναμενόμενο καθώς η κλάση με τους οροθετικούς χρήστες ήταν εκείνη με τα λιγότερα άτομα, οι αλγόριθμοι έτειναν να χαρακτηρίζουν τα άτομα ως οροαρνητικά. Λόγω της σοβαρότητας της HIV λοίμωξης γίνεται αντιληπτό πως ένας ταξινομητής θα πρέπει να έχει ιδιαίτερα υψηλή ευαισθησία, δηλαδή να χάνει όσο το δυνατόν λιγότερους οροθετικούς. Αυτό είχε ως αποτέλεσμα τη χρήση άλλων μεθόδων προκειμένου να ξεπεραστεί το πρόβλημα των μη ισορροπημένων δεδομένων.

Εφαρμόζοντας την τυχαία υποδειγματοληψία, παρατηρήθηκε μείωση στην ορθότητα (0.5222 – 0.7044), αντίθετα με την ευαισθησία, που αυξήθηκε, παίρνοντας τιμές στο διάστημα (0.4950 – 0.7822). Επίσης, παρατηρήθηκε μια μικρή μείωση στο εύρος του διαστήματος του AUC score (0.5220 – 0.7672). Η τυχαία δειγματοληψία λοιπόν φαίνεται πως διόρθωσε την ευαισθησία των ταξινομητών αλλά μειώθηκε η ορθότητα τους, το οποίο σημαίνει πως υπάρχει λανθασμένη ταξινόμηση των οροαρνητικών σε οροθετικούς χρήστες ενδοφλέβιων ναρκωτικών.

Η μέθοδοι υπερδειγματοληψίας φάνηκε να επιλύουν το παραπάνω πρόβλημα. Δηλαδή, αύξησαν τόσο την ευαισθησία, όσο και την ορθότητα των ταξινομητών. Ως προς την ευαισθησία φάνηκε πως τα μέγιστα επίπεδα επιτεύχθηκαν από τον αλγόριθμο KNN (0.9983) με την προσαρμοστική συνθετική μέθοδο δειγματοληψίας (ADASYN) στα δεδομένα που περιείχαν όλα τα χαρακτηριστικά. Το AUC score στην περίπτωση αυτή ήταν 0.9724. Ως προς το AUC score, τη μέγιστη επίδοση (0.9967) την είχε ο αλγόριθμος Random forest στην τυχαία υπερδειγματοληψία όταν εφαρμόστηκε είτε στα δεδομένα που περιείχαν όλα τα χαρακτηριστικά, είτε επιλέγοντας χαρακτηριστικά με τις μεθόδους της προς τα πίσω απαλοιφής χαρακτηριστικών και του random forest importance.

Η επιλογή του βέλτιστου ταξινομητή είναι άρρηκτα συνδεδεμένη με τη φύση του προβλήματος. Στην HIV λοίμωξη, όπως έχει αναφερθεί, μας ενδιαφέρει να ελαχιστοποιηθούν τα ψευδώς αρνητικά ταυτόχρονα όμως να είναι υψηλή η ορθότητα του ταξινομητή. Στην περίπτωση της τυχαίας υπερδειγματοληψίας ο αλγόριθμος Random forest είχε ευαισθησία 0.9929, ενώ η ορθότητα ήταν 0.9805 και το AUC score 0.9967. Στα ίδια επίπεδα κυμάνθηκε η απόδοση του αλγορίθμου και στην περίπτωση που έγινε επιλογή χαρακτηριστικών με την μέθοδο της προς τα πίσω απαλοιφής (ευαισθησία = 0.9929, accuracy = 0.9751 και AUC score = 0.9967). Με την μέθοδο αυτή χρησιμοποιήθηκε το 30.4% των χαρακτηριστικών. Δηλαδή χρησιμοποιήθηκαν μόνο τα 34 από τα 112 χαρακτηριστικά. Η επιλογή των χαρακτηριστικών είναι ιδιαίτερα χρήσιμη γιατί απαιτείται λιγότερη υπολογιστική ισχύς έχοντας την ίδια επίδοση με την περίπτωση που θα χρησιμοποιούνταν όλα τα χαρακτηριστικά.

Συμπερασματικά, θα λέγαμε πως ο αλγόριθμος Random Forest στην περίπτωση της τυχαίας υπερδειγματοληψίας είχε την καλύτερη επίδοση βάσει των μετρικών AUC, Recall και Accuracy. Ανάμεσα στα χαρακτηριστικά με τη μεγαλύτερη σημαντικότητα, μεταξύ άλλων ήταν η ηλικία, η συχνότητα χρήσης του speedball, ο αριθμός των δωρεάν συρίγγων που είχε λάβει τους τελευταίους 12 μήνες, η συχνότητα ενδοφλέβιας χρήσης καθώς και η χρήση της κοκαΐνης.

8. Μελλοντικές εργασίες

Στα πλαίσια μελλοντικών εργασιών θα μπορούσε να συμπεριληφθεί τόσο η υλοποίηση επιπλέον αλγορίθμων μηχανικής μάθησης, όσο και η διερεύνηση των παραμέτρων του εκάστοτε αλγόριθμου για να επιτευχθεί όσο το δυνατόν καλύτερη ταξινόμηση. Τέλος, θα είχε ιδιαίτερο ενδιαφέρον να αναπτυχθούν αντίστοιχοι αλγόριθμοι στον ίδιο πληθυσμό για την ύπαρξη χρόνιας ηπατίτιδας C καθώς και την πιθανότητα έναρξης και ολοκλήρωσης της θεραπείας με άμεσα δρώντα αντιικά φάρμακα. Αυτοί οι αλγόριθμοι θα επιτρέψουν την προτεραιοποίηση χρηστών σε αυξημένο κίνδυνο να πάσχουν από χρόνια ηπατίτιδα C και την εφαρμογή στοχευμένων παρεμβάσεων σε όσους έχουν αυξημένο κίνδυνο να μη διασυνδεθούν ή να εγκαταλείψουν τη θεραπεία.

Βιβλιογραφία

[1] "A Timeline of HIV and AIDS." HIV.gov. hiv.gov/hiv-basics/overview/history/hiv-and-aids-timeline (accessed September 15, 2021).

[2] "AIDS." Wikipedia. https://el.wikipedia.org/wiki/AIDS#.CE.99.CF.83.CF.84.CE.BF.CF.81.CE.AF.CE.B1 (accessed September 15, 2021).

[3] Y. Takebe, R. Uenishi, and X. Li, "Global molecular epidemiology of HIV: understanding the genesis of AIDS pandemic," *Advances in pharmacology,* vol. 56, pp. 1-25, 2008.

[4] "UNAIDS DATA 2020," 2020. [Online]. Available: https://www.unaids.org/sites/default/files/media_asset/2020_aids-data-book_en.pdf

[5] J. U. N. P. o. HIV/AIDS, "Ambitious treatment targets: writing the final chapter of the AIDS epidemic," *Geneva, Switzerland: UNAIDS,* 2014.

[6] "Εθνικός Οργανισμός Δημόσιας Υγείας. Δελτίο Επιδημιολογικής Επιτήρησης HIV/AIDS στην Ελλάδα, 31-12-2020 (Τεύχος 35). Αθήνα 2021. Πρόσβαση στο δικτυακό τόπο: https://eody.gov.gr."

[7] L. Degenhardt *et al.*, "Global prevalence of injecting drug use and sociodemographic characteristics and prevalence of HIV, HBV, and HCV in people who inject drugs: a multistage systematic review," *The Lancet Global Health,* vol. 5, no. 12, pp. e1192-e1207, 2017.

[8] "World Drug Report 2021," United Nations Office on Drugs and Crime, Vienna, Austria, 2021.

[9] D. C. Des Jarlais, T. Kerr, P. Carrieri, J. Feelemyer, and K. Arasteh, "HIV infection among persons who inject drugs: ending old epidemics and addressing new outbreaks," *AIDS (London, England),* vol. 30, no. 6, p. 815, 2016.

[10] D. C. Des Jarlais *et al.*, "HIV outbreaks among people who inject drugs in Europe, North America, and Israel," *The Lancet HIV,* vol. 7, no. 6, pp. e434-e442, 2020.

[11] V. Sypsa *et al.*, "Rapid decline in HIV incidence among persons who inject drugs during a fast-track combination prevention program after an HIV outbreak in Athens," *The Journal of infectious diseases,* vol. 215, no. 10, pp. 1496-1505, 2017.

[12] A. Hatzakis *et al.*, "Design and baseline findings of a large-scale rapid response to an HIV outbreak in people who inject drugs in Athens, Greece: the ARISTOTLE programme," *Addiction,* vol. 110, no. 9, pp. 1453-1467, 2015.

[13] J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon, "A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955," *AI magazine,* vol. 27, no. 4, pp. 12-12, 2006.

[14] R. Anyoha. "Sience in the News Special Edition Summer 2017 Artificial Intelligence: The History of Artificial Intelligence." Harvard University: The graduate school of arts and sciences. (accessed.

[15] Ι. Βλαχάβας, Π. Κεφαλάς, Ν. Βασιλειάδης, Φ. Κόκκορας, and Η. Σακελλαρίου, "Τεχνητή νοημοσύνη," *Δ Έκδοση. Γκιούρδας,* 2020.

[16] Α. Γεωργούλη, "Τεχνητή νοημοσύνη," 2015.

[17] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2018.

[18] A. Intelligence, "A modern approach," *by Stuart Russell and Peter Norvig,* 2003.

[19] S. R. Pandey, J. Ma, and C.-H. Lai, "A supervised machine learning approach to generate the auto rule for clinical decision support system," *Trends in Medicine,* vol. 20, no. 3, pp. 1-9, 2020.

[20] J. J. Oliver, R. A. Baxter, and C. S. Wallace, "Unsupervised learning using MML," in *ICML*, 1996: Citeseer, pp. 364-372.

[21] P. Cunningham, M. Cord, and S. J. Delany, "Supervised learning," in *Machine learning techniques for multimedia*: Springer, 2008, pp. 21-49.

[22] J. Morimoto and F. Ponton, "Virtual reality in biology: could we become virtual naturalists?," *Evolution: Education and Outreach,* vol. 14, no. 1, pp. 1-13, 2021.

[23] "Reinforcement learning algorithms and applications." https://techvidvan.com/tutorials/reinforcement-learning/ (accessed 19 January, 2022).

[24] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?," *PloS one,* vol. 12, no. 4, p. e0174944, 2017.

[25] F. D'Ascenzo *et al.*, "Machine learning-based prediction of adverse events following an acute coronary syndrome (PRAISE): a modelling study of pooled datasets," *The Lancet,* vol. 397, no. 10270, pp. 199-207, 2021.

[26] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature medicine,* vol. 25, no. 1, pp. 44-56, 2019.

[27] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Computational and structural biotechnology journal,* vol. 15, pp. 104-116, 2017.

[28] A. W. Forsyth *et al.*, "Machine learning methods to extract documentation of breast cancer symptoms from electronic health records," *Journal of pain and symptom management,* vol. 55, no. 6, pp. 1492-1499, 2018.

[29] O.-J. Skrede *et al.*, "Deep learning for prediction of colorectal cancer outcome: a discovery and validation study," *The Lancet,* vol. 395, no. 10221, pp. 350-360, 2020.

[30] N. Peiffer-Smadja *et al.*, "Machine learning for clinical decision support in infectious diseases: a narrative review of current applications," *Clinical Microbiology and Infection,* vol. 26, no. 5, pp. 584-595, 2020.

[31] G. A. Tadesse *et al.*, "Multi-modal diagnosis of infectious diseases in the developing world," *IEEE journal of biomedical and health informatics,* vol. 24, no. 7, pp. 2131-2141, 2020.

[32] S. Lim, C. S. Tucker, and S. Kumara, "An unsupervised machine learning model for discovering latent infectious diseases using social media data," *Journal of biomedical informatics,* vol. 66, pp. 82-94, 2017.

[33] R. Wei *et al.*, "Clinical prediction of HBV and HCV related hepatic fibrosis using machine learning," *EBioMedicine,* vol. 35, pp. 124-132, 2018.

[34] "The top 10 causes of death." World Health Organization. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death (accessed December 27, 2021).

[35] "Global Health Sector Strategy on Viral Hepatitis 2016-2021." World Health Organization. http://apps.who.int/iris/bitstream/10665/246177/1/WHO-HIV-2016.06-eng.pdf?ua=1 (accessed December 12, 2021).

[36] J. L. Marcus, W. C. Sewell, L. B. Balzer, and D. S. Krakower, "Artificial intelligence and machine learning for HIV prevention: emerging approaches to ending the epidemic," *Current HIV/AIDS Reports,* vol. 17, no. 3, pp. 171-179, 2020.

[37] M. G. Ahlström, A. Ronit, L. H. Omland, S. Vedel, and N. Obel, "Algorithmic prediction of HIV status using nation-wide electronic registry data," *EClinicalMedicine,* vol. 17, p. 100203, 2019.

[38] S. E. Hassig *et al.*, "Prevention of perinatal HIV transmission: are there alternatives to pre-pregnancy serological screening in Kinshasa, Zaire?," *AIDS (London, England),* vol. 4, no. 9, pp. 913-916, 1990.

[39] M. Lallemant *et al.*, "Characteristics associated with HIV-1 infection in pregnant women in Brazzaville, Congo," *Journal of acquired immune deficiency syndromes,* vol. 5, no. 3, pp. 279-285, 1992.

[40] C. W. Lee and J.-A. Park, "Assessment of HIV/AIDS-related health performance using an artificial neural network," *Information & Management,* vol. 38, no. 4, pp. 231-238, 2001.

[41] D. A. Annang, "Performance Comparison of Data Mining Techniques for Predicting Hiv Status Among Female Sex Workers in Ghana," Master of Science, School of Public Health, University Of Ghana, 2018. [Online]. Available: http://ugspace.ug.edu.gh/bitstream/handle/123456789/25817/Performance%20Comparison%20of%20Data%20Mining%20Techniques%20for%20Predicting%20Hiv%20Status%20Among%20Female%20Sex%20Workers%20in%20Ghana.pdf?sequence=1&isAllowed=y

[42] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering,* vol. 160, no. 1, pp. 3-24, 2007.

[43] R. Sathya and A. Abraham, "Comparison of supervised and unsupervised learning algorithms for pattern classification," *International Journal of Advanced Research in Artificial Intelligence,* vol. 2, no. 2, pp. 34-38, 2013.

[44] F. A. Thabtah, P. Cowling, and Y. Peng, "MMAC: A new multi-class, multi-label associative classification approach," in *Fourth IEEE International Conference on Data Mining (ICDM'04)*, 2004: IEEE, pp. 217-224.

[45] A. Mondal. "A Complete guide to Understand Classification in Machine Learning." Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/09/a-complete-guide-to-understand-classification-in-machine-learning/ (accessed December 12, 2021, 2021).

[46] L. Rokach, "Ensemble-based classifiers," *Artificial intelligence review,* vol. 33, no. 1, pp. 1-39, 2010.

[47] H. Ampadu. "Random Forests Understanding: Intuition and Implementation on a key algorithm to reduce overfitting in tree based algorithms." https://ai-pool.com/a/s/random-forests-understanding (accessed 22 January, 2022).

[48] A. Navlani. "KNN Classification using Scikit-learn." https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn (accessed 22 January, 2022).

[49] C. O. Freitas, J. M. De Carvalho, J. Oliveira, S. B. Aires, and R. Sabourin, "Confusion matrix disagreement for multiple classifiers," in *Iberoamerican Congress on Pattern Recognition*, 2007: Springer, pp. 387-396.

[50] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognition,* vol. 91, pp. 216-231, 2019.

[51] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC genomics,* vol. 21, no. 1, pp. 1-13, 2020.

[52] A. Kulkarni, D. Chong, and F. A. Batarseh, "Foundations of data imbalance and solutions for a data democracy," in *Data Democracy*: Elsevier, 2020, pp. 83-106.

[53] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International journal of data mining & knowledge management process,* vol. 5, no. 2, p. 1, 2015.

[54] H. Dalianis, "Evaluation metrics and evaluation," in *Clinical Text Mining*: Springer, 2018, pp. 45-53.

[55] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics,* 2020.

[56] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PloS one,* vol. 10, no. 3, p. e0118432, 2015.

[57] M. Vihinen, "How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis," in *BMC genomics*, 2012, vol. 13, no. 4: BioMed Central, pp. 1-10.

[58] K. Hajian-Tilaki, "Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation," *Caspian journal of internal medicine,* vol. 4, no. 2, p. 627, 2013.

[59] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Information sciences,* vol. 250, pp. 113-141, 2013.

[60] C. Elkan, "The foundations of cost-sensitive learning," in *International joint conference on artificial intelligence*, 2001, vol. 17, no. 1: Lawrence Erlbaum Associates Ltd, pp. 973-978.

[61] J. Grzyb, J. Klikowski, and M. Woźniak, "Hellinger Distance Weighted Ensemble for imbalanced data stream classification," *Journal of Computational Science,* vol. 51, p. 101314, 2021.

[62] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review," *GESTS International Transactions on Computer Science and Engineering,* vol. 30, no. 1, pp. 25-36, 2006.

[63] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, "On the class imbalance problem," in *2008 Fourth international conference on natural computation*, 2008, vol. 4: IEEE, pp. 192-201.

[64] E. Lin, Q. Chen, and X. Qi, "Deep reinforcement learning for imbalanced classification," *Applied Intelligence,* vol. 50, no. 8, pp. 2488-2502, 2020.

[65] S. Liu, J. Zhang, Y. Xiang, W. Zhou, and D. Xiang, "A study of data pre-processing techniques for imbalanced biomedical data classification," *International Journal of Bioinformatics Research and Applications,* vol. 16, no. 3, pp. 290-318, 2020.

[66] M. M. Rahman and D. N. Davis, "Addressing the class imbalance problem in medical datasets," *International Journal of Machine Learning and Computing,* vol. 3, no. 2, p. 224, 2013.

[67] A. Liu, J. Ghosh, and C. E. Martin, "Generative Oversampling for Mining Imbalanced Datasets," in *DMIN*, 2007, pp. 66-72.

[68] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from imbalanced data sets*. Springer, 2018, p. 83.

[69] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine learning with oversampling and undersampling techniques: overview study and experimental results," in *2020 11th International Conference on Information and Communication Systems (ICICS)*, 2020: IEEE, pp. 243-248.

[70] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research,* vol. 16, pp. 321-357, 2002.

[71] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, 2008: IEEE, pp. 1322-1328.

[72] R. Nian. "An Introduction to ADASYN." https://medium.com/@ruinian/an-introduction-to-adasyn-with-code-1383a5ece7aa (accessed 23 January, 2022).

[73] R. Taghizadeh-Mehrjardi *et al.*, "Synthetic resampling strategies and machine learning for digital soil mapping in Iran," *European Journal of Soil Science,* vol. 71, no. 3, pp. 352-368, 2020.

[74] V. Agarwal, "Research on data preprocessing and categorization technique for smartphone review analysis," *International Journal of Computer Applications,* vol. 131, no. 4, pp. 30-36, 2015.

[75] S. K. Dwivedi and B. Rawat, "A review paper on data preprocessing: A critical phase in web usage mining process," in *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*, 2015: IEEE, pp. 506-510.

[76] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. Front," *Energy Res,* vol. 9, p. 652801, 2021.

[77] D. A. Bennett, "How can I deal with missing data in my study?," *Australian and New Zealand journal of public health,* vol. 25, no. 5, pp. 464-469, 2001.

[78] Β. Βερύκιος, Β. Καγκλής, and Η. Σταυρόπουλος, "Τύποι, Ποιότητα και Προεπεξεργασία Δεδομένων," in *Η επιστήμη των δεδομένων μέσα από τη γλώσσα R*: Εκδόσεις Κάλλιπος, 2015, ch. 3.

[79] J. Boyer. "Prepare data for machine learning: Improve data quality with data cleansing, data transformation, and feature engineering." IBM. https://www.ibm.com/garage/method/practices/reason/prepare-data-for-machine-learning/ (accessed 27 January, 2022).

[80] Ε. Κύρκος, "Προεπεξεργασία Δεδομένων," *Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων*: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, 2015. [Online]. Available: http://hdl.handle.net/11419/1234

[81] X. Huang, L. Wu, and Y. Ye, "A review on dimensionality reduction techniques," *International Journal of Pattern Recognition and Artificial Intelligence,* vol. 33, no. 10, p. 1950017, 2019.

[82] F. Y. Kuo and I. H. Sloan, "Lifting the curse of dimensionality," *Notices of the AMS,* vol. 52, no. 11, pp. 1320-1328, 2005.

[83] R. Bellman, *Dynamic Programming*. Princeton University Press, 1958.

[84] Y. Ding, K. Zhou, and W. Bi, "Feature selection based on hybridization of genetic algorithm and competitive swarm optimizer," *Soft Computing,* pp. 1-10, 2020.

[85] K. Pearson, "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science,* vol. 2, no. 11, pp. 559-572, 1901, doi: 10.1080/14786440109462720.

[86] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of educational psychology,* vol. 24, no. 6, p. 417, 1933.

[87] Δ. Πετρίδης, "ΑΝΑΛΥΣΗ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ-ΠΑΡΑΓΟΝΤΙΚΗ ΑΝΑΛΥΣΗ," *Ανάλυση πολυμεταβλητών τεχνικών*: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, 2015. [Online]. Available: http://hdl.handle.net/11419/2129

[88] Κ. Φωκιανός and Χ. Χαραλάμπους, "Ανάλυση σε Κύριες Συνιστώσες και Διαχωριστική Ανάλυση," *Εισαγωγή στην R,* 2 ed., 2010, pp. 219-226. [Online]. Available: https://cran.r-project.org/doc/contrib/mainfokianoscharalambous.pdf

[89] P. Chhikara, N. Jain, R. Tekchandani, and N. Kumar, "Data dimensionality reduction techniques for Industry 4.0: Research results, challenges, and future research directions," *Software: Practice and Experience,* 2020.

[90] D. Elavarasan, D. R. Vincent PM, K. Srinivasan, and C.-Y. Chang, "A hybrid CFS filter and RF-RFE wrapper-based feature extraction for enhanced agricultural crop yield prediction modeling," *Agriculture,* vol. 10, no. 9, p. 400, 2020.