ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΕΥΦΥΕΙΣ ΤΕΧΝΟΛΟΓΙΕΣ ΔΙΑΔΙΚΤΥΟΥ – WEB INTELLIGENCE

Πρόβλεψη μη προσέλευσης ασθενών σε ιατρικά ραντεβού σε νοσοκομεία με αλγορίθμους κατηγοριοποίησης

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΕΙΡΗΝΗΣ Σ. ΛΑΦΑΤΖΗ

Επιβλέπων : Στέφανος Ουγιάρογλου
Επίκουρος Καθηγητής

Θεσσαλονίκη, Σεπτέμβριος 2023

Υπόδειγμα φύλλου τίτλου (μπροστινή σελίδα) του αντιτύπου ΜΔΕ

ΔΙΕΘΝΕΣ ΠΑΝΕΠΙΣΤΗΜΙΟ ΤΗΣ ΕΛΛΑΔΟΣ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΕΥΦΥΕΙΣ ΤΕΧΝΟΛΟΓΙΕΣ ΔΙΑΔΙΚΤΥΟΥ – WEB INTELLIGENCE

Πρόβλεψη μη προσέλευσης ασθενών σε ιατρικά ραντεβού σε νοσοκομεία με αλγορίθμους κατηγοριοποίησης

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΕΙΡΗΝΗΣ Σ. ΛΑΦΑΤΖΗ

Επιβλέπων : Στέφανος Ουγιάρογλου
Επίκουρος Καθηγητής

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή στις 30 Σεπτεμβρίου 2023.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Όνομα Επώνυμο	Όνομα Επώνυμο	Στέφανος Ουγιάρογλου
Καθηγητής ΔΙ.ΠΑ.Ε.	Καθηγητής ΔΙ.ΠΑ.Ε.	Επίκουρος Καθηγητής

Θεσσαλονίκη, Σεπτέμβριος 2023

(Υπογραφή)

.....

Ειρήνη Λαφατζή

Μηχανικός Πληροφορικής & Επικοινωνιών

Περίληψη

Ένα από τα μεγαλύτερα προβλήματα στον τομέα της υγείας είναι η μη-εμφάνιση των ασθενών σε προγραμματισμένα ραντεβού. Το φαινόμενο αυτό έχει σημαντικό αντίκτυπο, τόσο στους ασθενείς λόγω της χαμηλής ποιότητας παροχής υπηρεσιών από μέρους των δομών υγείας, όσο και στη λειτουργία των δομών σε οργανωτικό, διοικητικό, οικονομικό επίπεδο. Για την αντιμετώπιση του προβλήματος ένας τρόπος είναι να γίνει πρόβλεψη των ασθενών που δεν θα εμφανισθούν στα προγραμματισμένα ραντεβού. Έτσι θα ληφθούν τα κατάλληλα εξατομικευμένα μέτρα για κάθε ασθενή, όπως για παράδειγμα υπενθύμιση του ραντεβού. Αυτό μπορεί να επιτευχθεί με την χρήση μεθόδων εξόρυξης γνώσης πάνω σε δεδομένα ιατρικών ραντεβού.

Στην παρούσα διπλωματική εργασία σκοπός είναι η σύγκριση μεθόδων εξόρυξης γνώσης και συγκεκριμένα η σύγκριση αλγορίθμων κατηγοριοποίησης μέσω πειραματικής διαδικασίας. Για την εκτέλεση των πειραμάτων χρησιμοποιήθηκαν οι αλγόριθμοι k-NN, Naïve Bayes και δέντρα αποφάσεων (C4.5), οι οποίοι εφαρμόστηκαν πάνω σε 2 ελεύθερα διαθέσιμα σύνολα δεδομένων με ιατρικά ραντεβού. Επιπλέον διερευνήθηκε η πιθανότητα βελτίωσης των αποτελεσμάτων κατηγοριοποίησης με την εφαρμογή της τεχνικής υπερδειγματοληψίας SMOTE στα παραπάνω σύνολα δεδομένων. Για την προεπεξεργασία των δεδομένων και την εκτέλεση των πειραμάτων χρησιμοποιήθηκε το ελεύθερο λογισμικό εξόρυξης γνώσης WEKA. Μετά την πειραματική διαδικασία έγινε σύγκριση των αποτελεσμάτων και εξαγωγή συμπερασμάτων σχετικά με την επίδοση των αλγορίθμων κατηγοριοποίησης, καθώς και τον βαθμό επίδρασης της τεχνικής SMOTE στις επιδόσεις τους. Η πειραματική διαδικασία απέδειξε ότι οι αλγόριθμοι γνώσης μπορούν να βοηθήσουν στην πρόβλεψη μη εμφάνισης ασθενών σε προγραμματισμένα ραντεβού.

Λέξεις Κλειδιά: Εξόρυξη Γνώσης από Ιατρικά Δεδομένα, Κατηγοριοποίηση, SMOTE, Δέντρα αποφάσεων (C4.5), Naïve Bayes, k-Nearest Neighbors, Oversampling, Precision, Recall, F-measure, Accuracy, WEKA.

Abstract

One of the biggest problems in the health sector is patients not showing up for scheduled appointments. This phenomenon has a significant impact, both on patients due to the low quality of service provided by health structures, and on the functioning of the structures at an organizational, administrative, financial level. To deal with the problem, one way is to predict the patients who will not show up for the scheduled appointments. In this way, appropriate personalized measures will be taken for each patient, such as a reminder of the appointment. This can be achieved by using knowledge mining methods on medical appointment data.

In this thesis, the purpose is to compare knowledge mining methods and specifically to compare classification algorithms through an experimental process. To perform the experiments, k-NN, Naïve Bayes and decision trees (C4.5) algorithms were used, which were applied on 2 freely available datasets with medical appointments. In addition, the possibility of improving the classification results by applying the SMOTE oversampling technique to the above datasets was investigated. WEKA free knowledge mining software was used to pre-process the data and run the experiments. After the experimental process, the results were compared and conclusions were drawn regarding the performance of the classification algorithms, as well as the degree of influence of the SMOTE technique on their performance.

Keywords: Knowledge mining on medical data, Classification, SMOTE, Decision Trees (C4.5), Naïve Bayes, k-Nearest Neighbors, Oversampling, Precision, Recall, F-measure, Accuracy, WEKA.

Στον αδερφό μου…

Πίνακας περιεχομένων

1

Εισαγωγή

1.1 Εξόρυξη γνώσης σε δεδομένα του χώρου υγείας

Η ραγδαία ανάπτυξη των τεχνολογιών της πληροφορικής και των επικοινωνιών δημιουργεί την ολοένα και μεγαλύτερη ανάγκη ροής και επεξεργασίας τεράστιου όγκου δεδομένων σε νευραλγικούς και όχι μόνο τομείς (οικονομίας, υγείας, εκπαίδευσης κ. τ. λ) της σύγχρονης κοινωνίας. Παράλληλα όμως δημιουργείται η ανάγκη εξεύρεσης ολοένα και μεγαλύτερων αποθηκευτικών χώρων για την φιλοξενία αυτών των δεδομένων.

Προκειμένου να αντιμετωπιστούν τα παραπάνω ζητήματα αναπτύσσονται διάφορες μέθοδοι και εργαλεία ανάκτησης, επεξεργασίας και αποθήκευσης δεδομένων. Έτσι λοιπόν δίνεται η δυνατότητα σε διάφορους οργανισμούς, επιχειρήσεις και ιδρύματα να διατηρούν τον πολύ μεγάλο όγκο των δεδομένων τους σε αντίστοιχα πολύ μεγάλες βάσεις δεδομένων, ικανές να ανταπεξέλθουν στις απαιτήσεις τους. Επιπρόσθετα επιτυγχάνεται η πλέον κατάλληλη ανάλυση και επεξεργασία, η εύστοχη ερμηνευσιμότητα καθώς και η σωστή αξιολόγηση των παραπάνω δεδομένων με αποτέλεσμα την ορθή εξαγωγή συμπερασμάτων.

Απόρροια των παραπάνω είναι η κατεύθυνση στη λήψη έξυπνων αποφάσεων, η άμεση επίλυση προβλημάτων άρα και η βελτίωση των απαιτούμενων διαδικασιών ακόμα και η ανακάλυψη νέας αξιοποιήσιμης γνώσης. Κατά συνέπεια η διαδικασία της έρευνας, εξαγωγής και ανάλυσης αυτών των δεδομένων δεν κρίνεται μόνο απαραίτητη αλλά και άρτια

συνυφασμένη με την βελτίωση της ποιότητας ζωής. Ο κλάδος που ασχολείται με αυτό το πεδίο έρευνας είναι ευρέως διαδεδομένος ως «Εξόρυξη Γνώσης».

Η εξόρυξη γνώσης είναι πολύ χρήσιμη στον κλάδο της υγείας λόγω του τεράστιου όγκου δεδομένων που παράγεται στις διάφορες δομές και μονάδες υγείας (κέντρα υγείας, νοσοκομεία κ. λ. π) που είναι αρμόδιες για την ιατρική περίθαλψη ασθενών. Οι δομές αυτές, είτε είναι δημόσιες είτε ιδιωτικές, καλούνται να διαχειριστούν ένα μεγάλο πλήθος ασθενών ετησίως σχετικά με επισκέψεις, νοσηλείες, εξετάσεις κ.τ.λ. Για την καλύτερη οργάνωση αυτού του μεγάλου όγκου ιατρικών πράξεων χρησιμοποιείται η πρακτική των ιατρικών ραντεβού. Ένας ασθενής δηλαδή πρέπει να κλείσει ραντεβού σε κάποια μονάδα ανάλογα με την περιοχή που διαμένει ή ανάλογα με το ιατρικό πρόβλημα, αν χρειάζεται εξειδικευμένη περίθαλψη. Τα ραντεβού μπορεί να είναι διαφόρων τύπων όπως για παράδειγμα ραντεβού ρουτίνας (προληπτικοί έλεγχοι υγείας), ραντεβού για επανεξέταση μετά από περίοδο θεραπείας ή ραντεβού για εξέταση σε περίπτωση ασθενείας. Τα ραντεβού μπορούν να προγραμματιστούν είτε αυτοπροσώπως, αλλά και εξ αποστάσεως (τηλεφωνικά ή ηλεκτρονικά) που είναι και το πιο συνηθισμένο. Επιπλέον, ανάλογα με τις συνθήκες και διάφορους παράγοντες το ραντεβού μπορεί να προγραμματίζεται άμεσα (εντός μερικών ημερών). Από την άλλη πλευρά υπάρχουν και περιπτώσεις που ο χρόνος αναμονής είναι αρκετά μεγάλος, δηλαδή εβδομάδες ή μήνες.

Προφανώς, υπάρχουν και περιπτώσεις όπου ένας ασθενής αδυνατεί να προσέλθει στο ραντεβού και το ακυρώνει. Όσο νωρίτερα γίνει η ακύρωση τόσο πιο εύκολα μπορεί το κενό να αναπληρωθεί και να προγραμματιστεί κάποιο άλλο ραντεβού στη θέση του. Υπάρχουν όμως και οι περιπτώσεις ασθενών που δεν εμφανίζονται στο ραντεβού χωρίς να το ακυρώσουν. Αυτό είναι το λεγόμενο πρόβλημα της μη εμφάνισης ασθενών σε προγραμματισμένα ραντεβού.

1.2 Κίνητρο εκπόνησης της διπλωματικής εργασίας

Ένα αρκετά συχνό φαινόμενο που απαντάται τα τελευταία χρόνια στις δομές υγείας (Νοσοκομεία, ιδιωτικές κλινικές κτλ.) είναι η μη εμφάνιση των ασθενών σε προγραμματισμένα ραντεβού. Το φαινόμενο αυτό αποτελεί πρόβλημα το οποίο έχει σημαντικό αντίκτυπο τόσο στην ποιότητα παροχής υπηρεσιών στους ασθενείς, όσο και στην καλή οργάνωση και εύρυθμη λειτουργία των δομών υγείας.

Συγκεκριμένα, ο βαθμός απόδοσης μιας δομής υγείας μειώνεται καθώς εμφανίζονται περίοδοι «νεκρού χρόνου» αυξάνοντας παράλληλα το λειτουργικό κόστος (Dantas et al., 2018). Επιπλέον στην περίπτωση ιδιωτικών δομών εμφανίζεται και το πρόβλημα των μειωμένων

εσόδων. Σε έρευνες που έγιναν στην αρχή της δεκαετίας του 2000 στην Αγγλία, όπου το ποσοστό της μη-εμφάνισης βρέθηκε μεταξύ 6.5 – 7.7%, υπολογίστηκε ότι το κόστος τους ανήλθε σε περίπου 150 εκατομμύρια λίρες Αγγλίας (George & Rubin, 2003).

Όσον αφορά τον ασθενή που «χάνει» ένα προγραμματισμένο ραντεβού, ουσιαστικά καθυστερεί την διάγνωση και θεραπεία κάποιας ασθένειας, η οποία μπορεί να είναι ακόμα δυσκολότερο να αντιμετωπιστεί σε μεταγενέστερο χρόνο.

Ταυτόχρονα, ο ασθενής που δεν εμφανίζεται στο ραντεβού στερεί από κάποιον άλλο την ευκαιρία να εξεταστεί νωρίτερα και να βελτιώσει έτσι τις πιθανότητες της θεραπείας του (Elvira et al., 2018).

Το φαινόμενο της μη-εμφάνισης σε ραντεβού, εκτός από σοβαρό από πλευράς επιπτώσεων παρουσιάζεται και με αρκετά μεγάλη συχνότητα. Σύμφωνα με τα στατιστικά στοιχεία του Εθνικού Συστήματος Υγείας της Αγγλίας (NHS), από τον Απρίλιο του 2020 μέχρι τον Απρίλιο του 2021 κλείστηκαν 101,9 εκατομμύρια ραντεβού. Από αυτά, τα 23,5 εκατομμύρια ραντεβού δεν πραγματοποιήθηκαν επειδή οι ασθενείς δεν εμφανίστηκαν. Έτσι, το ποσοστό μη εμφάνισης ανέρχεται σε περίπου 23%. Επιπλέον, την δεκαετία 2010-2020 ο συνολικός μέσος όρος μη εμφάνισης σε ραντεβού σε νοσοκομεία της Αγγλίας ήταν 20,8% (με μέγιστη και ελάχιστη ετήσια τιμή 19,4% και 22,8% αντίστοιχα) (NHS Digital, 2021). Παρά το γεγονός όμως ότι έχουν γίνει κάποιες προσπάθειες αντιμετώπισης του προβλήματος, το ποσοστό της μη-εμφάνισης παρέμεινε πρακτικά αμετάβλητο την τελευταία δεκαετία στην Αγγλία. Αυτό δείχνει ότι το συγκεκριμένο πρόβλημα είναι πολύπλοκο και η αντιμετώπιση του δεν είναι εύκολη υπόθεση.

Το πρόβλημα αυτό όμως δεν παρουσιάζεται μόνο στην Αγγλία αλλά και σε παγκόσμια κλίμακα πράγμα που υποδηλώνει την επιτακτική ανάγκη επίλυσης του. Σε μια μελέτη 105 άρθρων που δημοσιεύτηκαν από το 1980 μέχρι το 2016 σχετικά με αυτό το πρόβλημα και προέρχονται από διάφορες περιοχές του κόσμου, αναφέρεται ότι ο μέσος όρος του ποσοστού μη εμφάνισης σε όλες αυτές τις έρευνες είναι επίσης 23%. Το ποσοστό αυτό όμως παρουσιάζει αρκετά μεγάλη διακύμανση, δηλαδή από 4% έως και 79,2% ανάλογα με το είδος της δομής υγείας. Επίσης, υπάρχει και μεγάλη διακύμανση ανάλογα με την ήπειρο που ανήκει η χώρα που έγινε η έρευνα. Τα υψηλότερα κατά μέσο όρο ποσοστά παρατηρούνται σε μελέτες για χώρες της Αφρικής (43%), στη συνέχεια της Νότιας Αμερικής (27,8%), της Ασίας (25,1%), της Ευρώπης (19,3%) και τέλος της Ωκεανίας (13,2%) (Dantas et al., 2018).

Για την αντιμετώπιση των επιπτώσεων της μη-εμφάνισης σε ραντεβού, οι δομές υγείας χρησιμοποιούν διάφορες στρατηγικές. Η πιο συνηθισμένη είναι η υπενθύμιση του ραντεβού τηλεφωνικά ή ηλεκτρονικά (π.χ. με SMS). Αυτή όμως η πρακτική συνοδεύεται από κάποιο κόστος σε εργατοώρες και πόρους γενικότερα. Συνεπώς είναι ασύμφορη όταν εφαρμόζεται σε μεγάλη κλίμακα. Μια άλλη προσέγγιση είναι η επιβολή κυρώσεων (οικονομικές ή άλλες) σε

όσους ασθενείς δεν παρουσιάζονται στο ραντεβού. Αυτό όμως μπορεί να περιορίσει την δυνατότητα πρόσβασης στην περίθαλψη σε κάποιες οικονομικά αδύναμες κοινωνικές ομάδες. Τα τελευταία λοιπόν χρόνια, άρχισαν να αναπτύσσονται συστήματα με στόχο να βελτιώσουν την διαχείριση των ραντεβού (έγκαιρη προσέλευση, κατάργηση λίστας αναμονής, μηδενικός νεκρός χρόνος) με βάση τους διαθέσιμους πόρους. Για να γίνει αυτό προσπαθούν να προβλέψουν έγκαιρα και με ακρίβεια τους ασθενείς που δεν θα εμφανισθούν στο προγραμματισμένο ραντεβού, λαμβάνοντας έτσι άμεσα τα κατάλληλα εξατομικευμένα μέτρα (Carreras-García et al., 2020).

Τα συστήματα όμως αυτά έχουν αποθηκευμένο ένα μεγάλο όγκο δεδομένων σχετικά με ασθενείς και ραντεβού. Για να καταστεί δυνατή η πρόβλεψη της μη εμφάνισης ασθενών σε ραντεβού είναι απαραίτητο να γίνει κατάλληλη επεξεργασία αυτών των δεδομένων. Για το σκοπό αυτό μπορούν να χρησιμοποιηθούν διάφορες μέθοδοι εξόρυξης γνώσης όπως η μέθοδος της κατηγοριοποίησης η οποία εφαρμόζεται στην συγκεκριμένη διπλωματική εργασία. Για παράδειγμα λοιπόν από το πληροφοριακό σύστημα ενός νοσοκομείου μπορεί να ληφθεί ένα σύνολο δεδομένων με τα ιατρικά ραντεβού των ασθενών. Αφού γίνει κατάλληλη επεξεργασία αυτών των δεδομένων, το ζητούμενο είναι ο κατηγοριοποιητής να προβλέπει αν ο εκάστοτε ασθενής θα προσέλθει ή όχι στο ραντεβού που είχε κλείσει για να εξεταστεί από κάποιο ιατρό του νοσοκομείου. Στην περίπτωση που ο κατηγοριοποιητής θα προβλέψει ότι κάποιος ασθενής δεν θα προσέλθει στο προγραμματισμένο ραντεβού, θα ενημερώνει αυτόματα το πληροφοριακό σύστημα ώστε να γίνουν οι απαραίτητες ενέργειες, είτε για υπενθύμιση του ραντεβού στον ασθενή, είτε για να εξεταστεί κάποιος άλλος ασθενής στη θέση του. Με αυτό τον τρόπο επιτυγχάνεται από την μια πλευρά η βέλτιστη παροχή υπηρεσιών προς τους ασθενείς οι οποίοι δεν θα ταλαιπωρούνται σε ουρές αναμονής και δεν θα υπάρχει "νεκρός χρόνος" για τον ιατρό αφού θα εξετάζει τον αμέσως επόμενο ασθενή. Από την άλλη πλευρά θα βελτιωθεί η οργάνωση και λειτουργεία του νοσοκομείου καθώς θα εξοικονομούνται πόροι.

1.3 Συνεισφορά

Η συνεισφορά της παρούσας διπλωματικής εργασίας είναι μία εκτεταμένη πειραματική μελέτη όπου γίνεται σύγκριση μεθόδων εξόρυξης γνώσης, οι οποίες εφαρμόζονται πάνω σε σύνολα δεδομένων που αφορούν ιατρικά ραντεβού. Στην πειραματική μελέτη χρησιμοποιήθηκαν τρεις αλγόριθμοι κατηγοριοποίησης, ο αλγόριθμος των κοντινότερων γειτόνων, ο Naïve Bayes και ο αλγόριθμος C4.5 που βασίζεται σε δέντρα απόφασης. Στο θεωρητικό μέρος της διπλωματικής εργασίας παρουσιάζεται η αρχή λειτουργίας των παραπάνω αλγορίθμων με αντίστοιχα παραδείγματα εφαρμογής. Στη συνέχεια γίνεται

λεπτομερής περιγραφή και ανάλυση του προβλήματος της μη-εμφάνισης των ασθενών σε προγραμματισμένα ραντεβού και αναφέρονται πιθανές ενέργειες για την επίλυση του. Έπειτα περιγράφονται τεχνικές υπερδειγματοληψίας για την αντιμετώπιση μη ισορροπημένων συνόλων δεδομένων.

Κατόπιν, στο πρακτικό μέρος παρουσιάζονται αναλυτικά δύο σύνολα δεδομένων που χρησιμοποιήθηκαν για την εκπόνηση της συγκεκριμένης διπλωματικής εργασίας. Τα προαναφερθέντα σύνολα δεδομένων είναι το "Medical Appointment" του Alvaro Flores και το "Medical Appointment No Shows" του JoniHoppen, τα οποία είναι ελεύθερα διαθέσιμα στην πλατφόρμα μηχανικής μάθησης Kaggle. Μετά από έρευνα σε σχετική βιβλιογραφία δεν βρέθηκαν εργασίες στις οποίες να έχει χρησιμοποιηθεί το σύνολο δεδομένων "Medical Appointment" του Alvaro Flores. Από την άλλη πλευρά, το σύνολο δεδομένων "Medical Appointment No Shows" του JoniHoppen βρέθηκε ότι έχει χρησιμοποιηθεί σε 4 εργασίες που αναφέρονται στην σχετική έρευνα. Για την προεπεξεργασία των δεδομένων και την εκτέλεση των πειραμάτων χρησιμοποιήθηκε το λογισμικό ανοιχτού κώδικα WEKA. Αρχικά έγινε προεπεξεργασία των δεδομένων και στα 2 σύνολα δεδομένων. Για παράδειγμα πραγματοποιήθηκε έλεγχος για ελλιπείς τιμές, καθαρισμός προβληματικών δεδομένων, μετατροπή των δεδομένων στην κατάλληλη μορφή, ώστε να μπορεί να τα διαβάσει το WEKA, αφαίρεση ή δημιουργία νέων χαρακτηριστικών κ.τ.λ.

Στη συνέχεια εκτελέστηκαν πειράματα με εφαρμογή των αλγορίθμων κατηγοριοποίησης Naïve Bayes, κοντινότερων γειτόνων (KNN) και C 4.5. Σε κάποια από τα πειράματα εφαρμόστηκε επιπλέον η τεχνική υπερδειγματοληψίας SMOTE για εξισορρόπηση των συνόλων δεδομένων. Σε άλλα πειράματα εκτελέστηκαν μόνο οι παραπάνω αλγόριθμοι χωρίς την τεχνική SMOTE. Στόχος ήταν να διερευνηθεί αν η προσθήκη της τεχνικής SMOTE βελτιώνει ή όχι τα αποτελέσματα. Κατά την διάρκεια των πειραμάτων, προκειμένου να επιτευχθεί η καλύτερη δυνατή απόδοση των αλγορίθμων, έγινε προσπάθεια βελτιστοποίησης των παραμέτρων τους. Συνολικά εκτελέστηκαν 412 πειράματα. Τέλος έγινε αξιολόγηση των αποτελεσμάτων με βάση τις μετρικές precision-recall που οδήγησαν σε χρήσιμα συμπεράσματα.

1.4 Σχετική έρευνα

Για την διερεύνηση του θέματος της συγκεκριμένης διπλωματικής εργασίας αναζητήθηκαν διάφορες πηγές. Με βάση την αναζήτηση διαπιστώθηκε ότι το πρόβλημα της μη εμφάνισης των ασθενών σε προγραμματισμένα ραντεβού ερευνάται από την αρχή της δεκαετίας του 1980. Οι εργασίες όμως ήταν σποραδικές. Συστηματικότερη έρευνα άρχισε να γίνεται από το 2000 και έπειτα. Μάλιστα φαίνεται ότι το θέμα έγινε ιδιαίτερα δημοφιλές την τελευταία

δεκαετία, καθώς ένας σχετικά μεγάλος αριθμός άρθρων έχει δημοσιευτεί από το 2013 και μετά. Στην βιβλιογραφική ανασκόπηση (Carreras-García et al., 2020) έχει γίνει μια επιλογή 50 άρθρων σχετικών με την πρόβλεψη της μη εμφάνισης ασθενών σε ραντεβού και περίπου το 80% από αυτά δημοσιεύτηκαν την τελευταία δεκαετία. Στα άρθρα αναφέρεται ότι χρησιμοποιείται μια ποικιλία μεθόδων και αλγορίθμων πρόβλεψης. Η πιο συνηθισμένη μέθοδος ειδικά στις παλαιότερες εργασίες είναι η λογιστική παλινδρόμηση (logistic regression) και στη συνέχεια τα μοντέλα που βασίζονται σε δέντρα αποφάσεων. Υπάρχουν επίσης και κάποιες εργασίες που χρησιμοποίησαν Νευρωνικά Δίκτυα, Μπεϋζιανά μοντέλα κ.α. Τα δεδομένα που χρησιμοποιήθηκαν ποικίλουν επίσης τόσο σε μέγεθος όσο και σε χαρακτηριστικά, με ποιο συνηθισμένα τα δημογραφικά στοιχεία των ασθενών όπως ηλικία, φύλο, είδος ασφάλισης και κάποια βασικά στοιχεία του ραντεβού όπως απόσταση τόπου κατοικίας, ημέρα εβδομάδας και χρόνος αναμονής. Επιπλέον χρησιμοποιούνται διαφορετικές μετρικές απόδοσης της κατηγοριοποίησης, οπότε δεν είναι δυνατή η άμεση σύγκριση μεταξύ των αποτελεσμάτων της κάθε εργασίας. Παρόλα αυτά τα αποτελέσματα ποικίλουν και γενικά οι επιδόσεις δεν είναι πολύ υψηλές, υποδεικνύοντας την δυσκολία που παρουσιάζει το συγκεκριμένο πρόβλημα και την ανάγκη περαιτέρω έρευνας. Ενδεικτικά, στη συνέχεια θα αναφερθούν κάποιες από τις εργασίες που περιλαμβάνονται στην παραπάνω ανασκόπηση.

Το 2014 η εργασία (Dravenstott et al., 2014) αναφέρει ότι χρησιμοποιήθηκε Νευρωνικό Δίκτυο για να προβλέψει ποιοι ασθενείς είχαν μεγαλύτερη πιθανότητα να μην εμφανιστούν σε ραντεβού. Το μοντέλο εκπαιδεύτηκε σε μια βάση δεδομένων με περίπου 3 εκατομμύρια ιατρικά ραντεβού προερχόμενα από μονάδες Πρωτοβάθμιας φροντίδας και υπηρεσίες Ενδοκρινολογίας. Τα δεδομένα αφορούσαν μια περίοδο 2 ετών και περιείχαν δημογραφικά, κλινικά και δεδομένα σχετικά με τα ραντεβού.

Το 2016 στην εργασία (Huang & Hanauer, 2016) οι συγγραφείς αναφέρουν ότι ένας συνηθισμένος τρόπος αντιμετώπισης των χαμένων ραντεβού είναι να επαναπρογραμματίζονται αργότερα ως επιπλέον ραντεβού (overbooking). Αυτό δημιουργεί διάφορα προβλήματα και γι' αυτό προσπάθησαν να μειώσουν τις επιπτώσεις του overbooking προβλέποντας πιθανές μη εμφανίσεις και βάζοντας τα χαμένα ραντεβού στη θέση αυτών. Οι συγγραφείς εφάρμοσαν λογιστική παλινδρόμηση σε δεδομένα που προέρχονταν από μια παιδιατρική κλινική που ανήκει στο Πανεπιστήμιο του Μίσιγκαν ΗΠΑ. Αφορούσαν ραντεβού από 7291 ασθενείς σε χρονικό διάστημα 10 ετών (Ιανουάριος 2002 έως Δεκέμβριος 2011) και το βασικό κριτήριο της επιλογής των ασθενών ήταν να έχουν επισκεφτεί την κλινική τουλάχιστον 2 φορές μέσα σε αυτό το χρονικό διάστημα. Τα δεδομένα περιλάμβαναν 17 χαρακτηριστικά (πληροφορίες σχετικά με ραντεβού, δημογραφικά και ασφάλισης) και επιπλέον προστέθηκε ως χαρακτηριστικό το ποσοστό των προηγούμενων μη εμφανίσεων σε ραντεβού.

Το 2017 στην εργασία (Lee et al., 2017) χρησιμοποιήθηκε ένα σύνολο δεδομένων με 1 εκατομμύριο ραντεβού (2015-2016) προερχόμενα από διάφορα νοσοκομεία της Σιγκαπούρης. Τα δεδομένα περιγράφονται από 42 χαρακτηριστικά (δημογραφικά, κλινικά και λεπτομέρειες ραντεβού) και το ποσοστό των περιπτώσεων μη εμφάνισης ήταν 24,4% των συνολικών ραντεβού. Στα δεδομένα αυτά εφαρμόστηκαν αρκετοί αλγόριθμοι όπως δέντρα αποφάσεων, λογιστική παλινδρόμηση, gradient boosting trees, random forests, elastic-net και XGBoost.

Στην εργασία (Goffman et al., 2017) του 2017 προσπάθησαν να υπολογίσουν την πιθανότητα μη εμφάνισης ασθενών σε ραντεβού χρησιμοποιώντας μοντέλα βασισμένα στη λογιστική παλινδρόμηση, σε δέντρα αποφάσεων και νευρωνικά δίκτυα. Τα δεδομένα που χρησιμοποιήθηκαν προέρχονται από διάφορα στρατιωτικά νοσοκομεία των ΗΠΑ και αποτελούνται από περίπου 41 εκατομμύρια ραντεβού που πραγματοποιήθηκαν μεταξύ 2005 και 2012 και περιγράφονται από 49 χαρακτηριστικά. Ο αλγόριθμος της λογιστικής παλινδρόμησης επιλέχθηκε λόγω της ευκολίας εφαρμογής και μάλιστα για να ελεγχθεί η αποτελεσματικότητα του τελικού μοντέλου, δοκιμάστηκε πιλοτικά σε πραγματικά δεδομένα 3 εβδομάδων όπου εντοπίστηκαν οι ασθενείς με υψηλή πιθανότητα μη εμφάνισης και έγιναν ενέργειες υπενθύμισης μέσω τηλεφώνου λίγο πριν το ραντεβού. Το αποτέλεσμα ήταν ότι το ποσοστό μη εμφάνισης μειώθηκε από 35% σε 12,16%.

Το 2018 η εργασία (Elvira et al., 2018) χρησιμοποίησε δεδομένα από ένα πανεπιστημιακό νοσοκομείο της Μαδρίτης, που αφορούσαν ραντεβού ασθενών από τον Ιανουάριο του 2015 έως Σεπτέμβριο του 2016. Το σύνολο δεδομένων αποτελούνταν από περίπου 2,3 εκατομμύρια ραντεβού και από τους αλγορίθμους που δοκιμάστηκαν υπερίσχυσε ο Gradient Boosting Machine (GBM). Επιπλέον δοκίμασαν να προσθέσουν στα αρχικά δεδομένα που αποτελούνταν από 12 χαρακτηριστικά, επιπλέον δεδομένα που αφορούσαν το ιστορικό εμφάνισης των ασθενών σε προηγούμενα ραντεβού.

Στην εργασία (Srinivas & Ravindran, 2018) τα δεδομένα προέρχονται από νοσοκομείο της Πενσυλβάνια ΗΠΑ και περιλαμβάνουν 76.285 ραντεβού ασθενών από το Σεπτέμβριο του 2014 έως τον Αύγουστο του 2016. Χρησιμοποιήθηκαν 18 χαρακτηριστικά τα οποία περιγράφουν δημογραφικά δεδομένα των ασθενών αλλά και δεδομένα σχετικά με τα ραντεβού. Σε αυτό το σύνολο δεδομένων εφαρμόστηκαν αρκετοί αλγόριθμοι όπως Λογιστική Παλινδρόμηση, Νευρωνικά δίκτυα, Random Forests, Gradient Boosting και Stacking.

Στην εργασία (Nelson et al., 2019) επιλέχθηκαν δεδομένα από ραντεβού ασθενών για μαγνητική τομογραφία (MRI) από τα νοσοκομεία University College Hospital και National Hospital for Neurology and Neurosurgery για την περίοδο από Ιανουάριο 2014 έως Δεκέμβριο 2016. Το σύνολο δεδομένων περιείχε περίπου 22.300 ραντεβού για 17.295 ασθενείς. Εφαρμόστηκαν αρκετοί αλγόριθμοι κατηγοριοποίησης όπως Λογιστική παλινδρόμηση, Μηχανές Διανυσμάτων Υποστήριξης (SVM), Random Forests, AdaBoost και

Gradient Boosting Machine (GBM). Επιπλέον, λήφθηκε υπόψη πως το πλήθος των ασθενών που εμφανίστηκαν στο ραντεβού ήταν περίπου 10 φορές μεγαλύτερο από αυτών που δεν εμφανίστηκαν. Δηλαδή το σύνολο δεδομένων δεν ήταν ισορροπημένο. Για να αντιμετωπίσουν αυτό το θέμα δοκίμασαν τεχνικές υπερδειγματοληψίας όπως Synthetic Minority Over-sampling Technique (SMOTE).

Να σημειωθεί εδώ ότι κανένα από τα σύνολα δεδομένων που χρησιμοποιήθηκαν στις παραπάνω εργασίες δεν είναι δημόσια διαθέσιμο. Συνεπώς τα αποτελέσματα τους δεν μπορούν να επαληθευτούν.

Μια πιο πρόσφατη βιβλιογραφική ανασκόπηση (Salazar et al., 2022) περιλαμβάνει 24 άρθρα που δημοσιεύτηκαν μεταξύ 2017 και 2021. Μάλιστα τα 11 από αυτά, δηλαδή σχεδόν τα μισά, δημοσιεύτηκαν το 2021. Αυτό δείχνει πως η χρήση μεθόδων μηχανικής μάθησης για πρόβλεψη της μη εμφάνισης ασθενών σε ιατρικά ραντεβού έχει γίνει αρκετά δημοφιλής τα τελευταία χρόνια. Η ανασκόπηση των άρθρων έδειξε ότι χρησιμοποιήθηκαν δεδομένα από ποικίλες πηγές. Συγκεκριμένα χρησιμοποιήθηκαν 18 διαφορετικά σύνολα δεδομένων. Από αυτά, τα 5 ήταν αρκετά μεγάλα με πάνω από 1 εκατομμύριο εγγραφές, ενώ στα υπόλοιπα σύνολα δεδομένων ο όγκος των δεδομένων ήταν κατά μέσο όρο 120 χιλιάδες εγγραφές.

Το μεγαλύτερο σύνολο δεδομένων από όλα ήταν της τάξης των 33 εκατομμυρίων εγγραφών.

Ένα μεγάλο πλήθος μεθόδων μηχανικής μάθησης αναφέρεται για την πρόβλεψη των ασθενών που δεν εμφανίστηκαν, αλλά οι πιο συχνά χρησιμοποιούμενες ήταν τα Δέντρα Αποφάσεων, τα Τυχαία Δάση (Random Forests) και η Λογιστική Παλινδρόμηση. Για την μέτρηση των επιδόσεων χρησιμοποιούνται διάφορες μετρικές, όπως Precision, Recall, AUC-ROC, F-Score, AUROC κτλ., οπότε δεν είναι δυνατή η άμεση σύγκριση μεταξύ των αποτελεσμάτων των μελετών. Σε γενικές γραμμές οι αλγόριθμοι που έδωσαν τα καλύτερα αποτελέσματα ήταν τα Δέντρα Αποφάσεων, τα Τυχαία Δάση και ο αλγόριθμος Gradient Boosting. Σε πολλές από αυτές τις εργασίες μελετήθηκε ποια ήταν τα πιο σημαντικά χαρακτηριστικά για την πρόβλεψη της μη εμφάνισης των ασθενών στα ραντεβού. Στις περισσότερες περιπτώσεις αναφέρθηκε η ηλικία του ασθενή, αν ο ασθενής είχε προηγούμενες περιπτώσεις μη εμφάνισης και ο χρόνος αναμονής του ραντεβού. Σε κάποιες από τις εργασίες χρησιμοποιήθηκε το ελεύθερα διαθέσιμο σύνολο δεδομένων JoniHoppen, το οποίο περιέχει 14 χαρακτηριστικά και 110.528 ιατρικά ραντεβού προερχόμενα από νοσοκομεία της Βραζιλίας. Στην συνέχεια θα αναφερθούν κάποιες από αυτές τις εργασίες.

Στην εργασία (Alshaya et al., 2019) χρησιμοποιήθηκε το σύνολο δεδομένων JoniHoppen, στο οποίο προστέθηκαν επιπλέον χαρακτηριστικά όπως χρόνος αναμονής, ημέρα της εβδομάδας, πλήθος επισκέψεων του ασθενή κλπ. Επίσης, συνδυάστηκαν και πρόσθετα δεδομένα από κάποιο σύνολο δεδομένων καιρικών συνθηκών για τις συγκεκριμένες περιοχές των νοσοκομείων. Το σύνολο δεδομένων JoniHoppen είναι μη ισορροπημένο ως προς το πλήθος

των περιπτώσεων εμφάνισης και μη-εμφάνισης ασθενών. Για την αντιμετώπιση αυτού του προβλήματος, οι συγγραφείς δοκίμασαν αρκετές μεθόδους δειγματοληψίας. Συγκεκριμένα αναφέρονται οι μέθοδοι υπερδειγματοληψίας Random oversampling, ADASYN, SMOTE και οι μέθοδοι υποδειγματοληψίας Random undersampling, AllKNN και Edited Nearest Neighbors. Επίσης δοκιμάστηκαν και κάποιες υβριδικές τεχνικές όπως SMOTEENN και SMOTETomek. Για την πρόβλεψη χρησιμοποιήθηκαν οι αλγόριθμοι της Λογιστικής Παλινδρόμησης, Τυχαία Δάση, k-NN, Μηχανές Διανυσμάτων Υποστήριξης (SVM) και Στοχαστική Κατάβαση Δυναμικού (Stochastic Gradient Descent-SGD). Όσον αφορά την επίδοση οι αλγόριθμοι SGD και SVM ισοψήφησαν, παρουσιάζοντας τα καλύτερα αποτελέσματα σε συνδυασμό με διάφορες μεθόδους δειγματοληψίας. Τα αποτελέσματα αυτά προέκυψαν τόσο σε περιπτώσεις εφαρμογής των παραπάνω αλγορίθμων με τις ίδιες, όσο και με διαφορετικές μεθόδους δειγματοληψίας.

Η εργασία (Nasir et al., 2020) προτείνει την ανάπτυξη ενός εργαλείου, το οποίο δημιουργεί μοντέλα πρόβλεψης ασθενών που δε θα προσέλθουν σε ραντεβού. Τα μοντέλα αυτά στη συνέχεια θα μπορούν να χρησιμοποιηθούν ως εργαλεία υποστήριξης αποφάσεων σχετικά με ενέργειες που μπορούν να γίνουν για την αντιμετώπιση αυτών των περιπτώσεων. Όσον αφορά το σύνολο δεδομένων JoniHoppen, οι συγγραφείς δημιούργησαν 12 επιπλέον χαρακτηριστικά όπως χρόνος αναμονής, ημέρα εβδομάδας του ραντεβού, αριθμός προηγούμενων ραντεβού στα οποία δεν εμφανίστηκε ο ασθενής, εάν έγινε επαναπρογραμματισμός του ραντεβού κλπ. Μετά την δημιουργία αυτών των χαρακτηριστικών διέγραψαν όσα πρότυπα αφορούσαν ραντεβού τα οποία τελικά επαναπρογραμματίστηκαν. Επίσης στο χαρακτηριστικό age θεώρησαν τις μηδενικές τιμές ως ελλιπείς και αφαίρεσαν τις αντίστοιχες εγγραφές που αντιστοιχούσαν στο 3% των αρχικών δεδομένων. Μετά τις παραπάνω αλλαγές το σύνολο δεδομένων που χρησιμοποίησαν περιείχε 59.710 εγγραφές. Οι αλγόριθμοι που χρησιμοποιήθηκαν ήταν Τεχνητά Νευρωνικά Δίκτυα (ANN), Μηχανές Διανυσμάτων Υποστήριξης (SVM), Τυχαία Δάση (Random Forests) και Λογιστική Παλινδρόμηση. Επιπλέον, για την εξισορρόπηση του συνόλου δεδομένων δοκιμάστηκαν 4 μέθοδοι. Συγκεκριμένα εφαρμόστηκε η μέθοδος της τυχαίας υποδειγματοληψίας καθώς και οι μέθοδοι υπερδειγματοληψίας SMOTE, ADASYN και MWMOTE. Η μετρική Recall ήταν μια από τις μετρικές επίδοσης που χρησιμοποιήθηκε. Ο αλγόριθμος Random Forest σε συνδυασμό με τυχαία υποδειγματοληψία έδωσε τα καλύτερα αποτελέσματα.

Σε γενικές γραμμές όμως τα περισσότερα μοντέλα έδωσαν σχετικά χαμηλά αποτελέσματα για την μετρική Recall και οι τεχνικές εξισορρόπησης που χρησιμοποιήθηκαν δεν επηρέασαν ιδιαίτερα τα αποτελέσματα.

Το 2021 στην εργασία (Batool et al., 2021) προτάθηκε η δημιουργία ενός συστήματος διαχείρισης ραντεβού με όνομα ASIM (Appointment Scheduling and Intuitive Management). Το συγκεκριμένο σύστημα επιτρέπει τον εύκολο και γρήγορο προγραμματισμό ιατρικών ραντεβού μέσω εφαρμογής κινητής συσκευής για τους ασθενείς και μέσω διαδικτυακής εφαρμογής web για τον διαχειριστή του συστήματος. Επιπλέον, μπορεί να προβλέψει αν ένας ασθενής είναι πιθανόν να μην προσέλθει στο ραντεβού, δίνοντας τη δυνατότητα αυτοματοποιημένων ενεργειών για την αντιμετώπιση αυτών των περιπτώσεων. Για την δημιουργία του κατηγοριοποιητή του συστήματος χρησιμοποιήθηκε το σύνολο δεδομένων JoniHoppen. Στο σύνολο δεδομένων προστέθηκε το νέο χαρακτηριστικό Lead time, δηλαδή ο χρόνος αναμονής για το ραντεβού. Για την εξισορρόπηση του συνόλου δεδομένων οι συγγραφείς επέλεξαν τον αλγόριθμο υποδειγματοληψίας Instance Hardness Threshold (IHT), η εφαρμογή του οποίου βελτίωσε τις επιδόσεις των αλγορίθμων κατηγοριοποίησης κατά 10%. Οι αλγόριθμοι που χρησιμοποιήθηκαν ήταν Δέντρα Αποφάσεων, Naïve Bayes, k-NN και Μηχανές Διανυσμάτων Υποστήριξης (SVM) με γραμμικό και μη-γραμμικό πυρήνα αντίστοιχα. Για την αξιολόγηση των επιδόσεων χρησιμοποιήθηκαν οι μετρικές Precision και Recall και έδωσαν αρκετά υψηλότερα αποτελέσματα ως προς την πρόβλεψη σε σχέση με άλλες μελέτες.

Τέλος στην εργασία (Alshammari et al., 2021) οι συγγραφείς χρησιμοποίησαν το σύνολο δεδομένων JoniHoppen, αλλά κράτησαν μόνο τα πρότυπα με τα ραντεβού από ενήλικες ασθενείς. Έτσι το τελικό σύνολο δεδομένων που χρησιμοποιήθηκε περιείχε 82.992 πρότυπα. Επίσης, πρόσθεσαν νέα χαρακτηριστικά όπως το Lead time, δηλαδή το χρόνο αναμονής του ραντεβού και το AM, δηλαδή αν το ραντεβού προγραμματίστηκε για πριν ή μετά το μεσημέρι. Επιπλέον αφαίρεσαν τελείως κάποια χαρακτηριστικά όπως π.χ. τη διεύθυνση του Νοσοκομείου. Οι αλγόριθμοι που χρησιμοποιήθηκαν ήταν τα Δέντρα Αποφάσεων και ο αλγόριθμος AdaBoost, ενώ δεν χρησιμοποιήθηκε καμία μέθοδος εξισορρόπησης των δεδομένων. Για την αξιολόγηση των επιδόσεων χρησιμοποιήθηκαν αρκετές μετρικές όπως Precision, Recall, ROC κλπ. Η σύγκριση των αποτελεσμάτων έδειξε ότι τα Δέντρα Αποφάσεων είχαν καλύτερες επιδόσεις από τον αλγόριθμο AdaBoost.

Αξίζει να σημειωθεί ότι το σύνολο δεδομένων JoniHoppen που προαναφέρθηκε στις παραπάνω εργασίες, χρησιμοποιήθηκε και στην παρούσα διπλωματική εργασία. Παρόλα αυτά χρησιμοποιήθηκε με διαφορετικό τρόπο ως προς την προεπεξεργασία των δεδομένων, την μεθοδολογία εκτέλεσης των πειραμάτων καθώς και την μέτρηση της επίδοσης των κατηγοριοποιητών. Επιπλέον σε κάποιες περιπτώσεις εφαρμόστηκε ο ίδιος αλγόριθμος, αλλά σε συνδυασμό με διαφορετική τεχνική δειγματοληψίας. Σε άλλες πάλι περιπτώσεις εφαρμόστηκε η ίδια τεχνική δειγματοληψίας αλλά σε συνδυασμό με διαφορετικούς

αλγορίθμους. Κατά συνέπεια η σύγκριση των αποτελεσμάτων μεταξύ των προαναφερθέντων εργασιών δεν είναι άμεσα εφικτή.

1.5 Οργάνωση διπλωματικής εργασίας

Η συγκεκριμένη διπλωματική εργασία είναι δομημένη στα παρακάτω 8 Κεφάλαια.

Στο Κεφάλαιο 1 λοιπόν περιγράφεται το αντικείμενο της διπλωματικής εργασίας και η συνεισφορά της στο χώρο της εξόρυξης γνώσης. Επίσης παρουσιάζονται άλλες σχετικές με το θέμα εργασίες. Ακολουθεί το Κεφάλαιο 2, στο οποίο γίνεται αναφορά σε βασικές έννοιες, όπως εξόρυξη γνώσης, μηχανική μάθηση, κατηγοριοποίηση. Στη συνέχεια περιγράφονται αναλυτικά οι αλγόριθμοι κατηγοριοποίησης που χρησιμοποιήθηκαν με σχετικά παραδείγματα. Έπειτα στο Κεφάλαιο 3 παρουσιάζονται μέτρα για την εκτίμηση της επίδοσης των αλγορίθμων κατηγοριοποίησης, τα οποία ερμηνεύονται με τη βοήθεια διαγραμμάτων Venn. Στη συνέχεια γίνεται αναφορά στις μεθόδους εκτίμησης της επίδοσης των αλγορίθμων κατηγοριοποίησης και αναλυτική περιγραφή της μεθόδου επικυρωμένης διασταύρωσης (Cross Validation). Στο Κεφάλαιο 4 παρουσιάζεται διεξοδικά το πρόβλημα της μη εμφάνισης των ασθενών σε προγραμματισμένα ραντεβού και οι επιπτώσεις που προκύπτουν από αυτό. Έπειτα προτείνονται λύσεις για την εξάλειψη του φαινομένου. Το Κεφάλαιο 5 πραγματεύεται το πρόβλημα της ανισοκατανομής κλάσεων σε σύνολα δεδομένων. Για την επίλυση του αναφέρονται διάφορες τεχνικές δειγματοληψίας. Ακολουθεί αναλυτική επεξήγηση της τεχνικής υπερδειγματοληψίας SMOTE, η οποία χρησιμοποιείται στην παρούσα διπλωματική εργασία. Στο Κεφάλαιο 6 παρουσιάζεται το λογισμικό WEKA, στο οποίο πραγματοποιήθηκε η εκτέλεση των πειραμάτων. Στη συνέχεια αναφέρονται ιδιαιτερότητες των ιατρικών συνόλων δεδομένων, όπως θέματα ανωνυμοποίησης και δυσκολίας απόκτησης τους. Έπειτα παρουσιάζονται τα 2 σύνολα δεδομένων που χρησιμοποιήθηκαν, καθώς και η διαδικασία προεπεξεργασία τους. Στο Κεφάλαιο 7 γίνεται λόγος για τη μεθοδολογία που ακολουθήθηκε για την εκτέλεση των πειραμάτων. Στο Κεφάλαιο 8 παρουσιάζονται τα αποτελέσματα των πειραμάτων με αντίστοιχα διαγράμματα. Στο Κεφάλαιο 9 συνοψίζονται τα συμπεράσματα που προέκυψαν καθώς και προτάσεις για μελλοντική έρευνα.

2

Θεωρητικό υπόβαθρο

2.1 Εξόρυξη Γνώσης και Μηχανική Μάθηση

2.1.1 Εξόρυξη Γνώσης

Η Εξόρυξη Γνώσης είναι το σύνολο των μεθόδων και τεχνικών που επιτρέπουν την ανάλυση δεδομένων πολύ μεγάλου όγκου για την εξαγωγή και ανακάλυψη προηγουμένως άγνωστων προτύπων και σχέσεων που ενδεχομένως να υπάρχουν μέσα σε αυτά. Αυτές οι πληροφορίες φιλτράρονται, προετοιμάζονται και ταξινομούνται έτσι ώστε να αξιοποιηθούν κατάλληλα στο μέλλον για αποφάσεις και στρατηγικές (Grabmeier & Rudolph, n.d.).

Ειδικότερα στον τομέα της υγείας ο μεγάλος όγκος δεδομένων που παράγεται από τις απαιτούμενες διαδικασίες είναι αρκετά πολύπλοκος, ώστε η επεξεργασία και ανάλυση του να μην είναι δυνατή με τις παραδοσιακές μεθόδους, παρά μόνο με μεθόδους εξόρυξης γνώσης. Η εξόρυξη γνώσης σε ιατρικά και οικονομικά δεδομένα είναι πλέον αναγκαία, έτσι ώστε οι οργανισμοί υγείας να μπορούν να πάρουν τις καλύτερες κατά το δυνατόν αποφάσεις που απώτερο σκοπό έχουν από την μία πλευρά την μείωση του κόστους για τους οργανισμούς και από την άλλη, την καλύτερη ποιότητα παροχής υπηρεσιών προς τους πολίτες (Koh & Tan, 2011).

2.1.2 Μηχανική Μάθηση

Ως Μηχανική Μάθηση μπορεί να οριστεί η διαδικασία κατασκευής συστημάτων λογισμικού τα οποία μπορούν αυτόματα να βελτιώνονται μέσω της εμπειρίας και ενσωματώνουν κάποιο είδος διαδικασίας μάθησης. Η μάθηση συνήθως επιτελείται με την βοήθεια δεδομένων τα οποία είναι διαθέσιμα για το σκοπό αυτό. Ο σκοπός της διαδικασίας είναι η επίλυση διαφόρων τύπων προβλημάτων όπως κατηγοριοποίηση, συσταδοποίηση, πρόβλεψη τιμών κλπ. (Ayodele, 2010)**.**

2.1.3 Τύποι μάθησης

Ανάλογα με την φύση των δεδομένων και την μορφή του προβλήματος οι αλγόριθμοι μπορούν να διαχωριστούν στις παρακάτω κατηγορίες:

1) Επιβλεπόμενη μάθηση (supervised learning)
2) Μη Επιβλεπόμενη μάθηση (unsupervised learning)
3) Μάθηση με ημιεπίβλεψη (Semi-supervised Learning)
4) Ενισχυτική μάθηση (Reinforcement Learning)

Στην **Επιβλεπόμενη μάθηση (supervised learning)** σκοπός είναι ένα σύστημα να «μάθει» μια συνάρτηση η οποία συσχετίζει ομάδες δεδομένων που δίνονται ως είσοδος, με δεδομένα που είναι οι επιθυμητές έξοδοι. Το σύνολο όλων αυτών των δεδομένων ονομάζεται σύνολο εκπαίδευσης και αποτελείται από τιμές χαρακτηριστικών (είσοδοι) και κάποια τιμή «στόχο» (έξοδος). Η εκπαίδευση πραγματοποιείται με τη βοήθεια ενός αλγορίθμου εκπαίδευσης και των δεδομένων εκπαίδευσης. Οι δύο δημοφιλέστερες διαδικασίες Επιβλεπόμενης μάθησης είναι η κατηγοριοποίηση (classification) και η παλινδρόμηση (regression). Στην κατηγοριοποίηση η έξοδος μπορεί να πάρει συγκεκριμένες (διακριτές) τιμές που αντιστοιχούν στην κλάση στην οποία ανήκουν τα αντίστοιχα δεδομένα. Αντίθετα στην παλινδρόμηση η έξοδος μπορεί να πάρει συνεχείς αριθμητικές τιμές. Αφού το σύστημα εκπαιδευτεί με τα δεδομένα εκπαίδευσης, αποκτά μια ικανότητα «γενίκευσης», δηλαδή μπορεί να χρησιμοποιηθεί για την πρόβλεψη της τιμής εξόδου όταν δεχθεί ως είσοδο «άγνωστα» δεδομένα, δηλαδή δεδομένα διαφορετικά από αυτά πάνω στα οποία εκπαιδεύτηκε (Ayodele, 2010)**.**

Στην **Μη Επιβλεπόμενη μάθηση (unsupervised learning)** ένας αλγόριθμος δέχεται ένα σύνολο δεδομένων με σκοπό την ανακάλυψη συσχετίσεων, δομών και μοτίβων μεταξύ των δεδομένων αυτών. Η χαρακτηριστική διαφορά σε σχέση με την Επιβλεπόμενη μάθηση είναι

ότι εδώ τα δεδομένα εισόδου δεν περιέχουν κάποια τιμή «στόχο» ως έξοδο. Η πιο συνηθισμένη εφαρμογή Μη επιβλεπόμενης μάθησης είναι η συσταδοποίηση, δηλαδή ο χωρισμός των δεδομένων σε ομάδες με βάση κάποιο χαρακτηριστικό τους (Mohri et al., 2018).

Στην **Μάθηση με ημιεπίβλεψη (Semi-supervised Learning)** δίνεται στον αλγόριθμο μάθησης ένα σύνολο δεδομένων, στο οποίο κάποια από τα δεδομένα περιέχουν τιμή «στόχο» και τα υπόλοιπα δεδομένα (συνήθως τα περισσότερα) δεν περιέχουν καμία πληροφορία για τον «στόχο». Πρόκειται για συνδυασμό Επιβλεπόμενης μάθησης και Μη επιβλεπόμενης μάθησης και χρησιμοποιείται όταν η εύρεση δεδομένων που περιέχουν την αντίστοιχη τιμή «στόχο» είναι δύσκολη ή έχει μεγάλο κόστος. Αντίθετα η εύρεση δεδομένων χωρίς τιμή «στόχο» είναι ευκολότερη και φθηνότερη (Mohri et al., 2018)**.**

Στην **Ενισχυτική Μάθηση** ένα σύστημα εκπαιδεύεται λαμβάνοντας δεδομένα μέσω της αλληλεπίδρασης με το περιβάλλον του και λαμβάνοντας κάποιου είδους ανταμοιβή ανάλογα με το αποτέλεσμα της κάθε ενέργειας που εκτελεί μέσα σε αυτό. Ο σκοπός του αλγορίθμου είναι να «μάθει» την ακολουθία ενεργειών που δίνει συνολικά την καλύτερη ανταμοιβή, άρα θα πετυχαίνει και την καλύτερη δυνατή λύση στο πρόβλημα πάνω στο οποίο εκπαιδεύτηκε (Mohri et al., 2018)**.**

2.2 Κατηγοριοποίηση

Η κατηγοριοποίηση είναι η διαδικασία κατά την οποία προσδιορίζεται η κλάση στην οποία ανήκει ένα πρότυπο. Η "κλάση" στην κατηγοριοποίηση, μέσα σε ένα σύνολο δεδομένων είναι το χαρακτηριστικό για το οποίο ενδιαφέρονται περισσότερο οι χρήστες. Στη στατιστική η κλάση ορίζεται ως η εξαρτώμενη μεταβλητή. Ένας αλγόριθμος κατηγοριοποίησης παίρνει σαν είσοδο ένα σύνολο δεδομένων και παράγει στην έξοδο την κλάση στην οποία ανήκει το κάθε πρότυπο (Fan & Li, 1998). Για την κατηγοριοποίηση των δεδομένων κάποιοι αλγόριθμοι χρησιμοποιούν κανόνες κατηγοριοποίησης, όπως για παράδειγμα ο αλγόριθμος δέντρων αποφάσεων C4.5. Για παράδειγμα **IF LungCancerFamilyHistory = ναι και κάπνισμα = ναι, THEN CT_Scan = απαιτείται.** Ο συγκεκριμένος κανόνας ερμηνεύεται ως εξής: Εάν στο οικογενειακό ιστορικό ενός ασθενούς απαντάται καρκίνος του πνεύμονα και ο ασθενής καπνίζει, τότε είναι αναγκαίο να βγάλει ακτινογραφία θώρακος**,** προφανώς γιατί υπάρχει η πιθανότητα να νοσεί και ο ίδιος (Yoo et al., 2012).

Άλλοι αλγόριθμοι χρησιμοποιούν μαθηματικούς υπολογισμούς όπως οι Naive Bayes (N. Friedman et al., 1997) και KNN (Cover & Hart, 1967).

Η κατηγοριοποίηση είναι μια διαδικασία που περιλαμβάνει δύο φάσεις:

1η Φάση: Εκπαίδευση μοντέλου

2η Φάση: Έλεγχος επίδοσης μοντέλου

Στην πρώτη φάση ο αλγόριθμος επεξεργάζεται τα δεδομένα του συνόλου εκπαίδευσης και κατασκευάζει ένα μοντέλο. Σ' αυτή τη φάση δίνονται στον αλγόριθμο ως είσοδος τα δεδομένα εκπαίδευσης (training set) και ο αλγόριθμος με βάση αυτά τα δεδομένα εκπαιδεύεται ώστε να κατηγοριοποιεί κάθε πρότυπο σε μία κλάση. Δηλαδή παράγει ως έξοδο ένα συγκεκριμένο μοντέλο κατηγοριοποίησης, ανάλογα με τον αλγόριθμο που χρησιμοποιεί κάθε φορά. Ο στόχος στο στάδιο της εκπαίδευσης είναι το μοντέλο να μπορεί να κατηγοριοποιεί τα δεδομένα εκπαίδευσης με όσο το δυνατόν μεγαλύτερη ακρίβεια. Για παράδειγμα αν χρησιμοποιηθεί ο αλγόριθμός δέντρων απόφασης C4.5 για την κατηγοριοποίηση δεδομένων, δίνονται στον αλγόριθμο τα δεδομένα εκπαίδευσης, ο αλγόριθμος κάνει τους απαραίτητους υπολογισμούς και στη συνέχεια κατασκευάζει ένα δέντρο απόφασης το οποίο ακολούθως θα χρησιμοποιηθεί για κατηγοριοποίηση νέων δεδομένων.

Στη δεύτερη φάση της διαδικασίας κατηγοριοποίησης (φάση ελέγχου) ελέγχεται η επίδοση του μοντέλου ως προς την πρόβλεψη της κλάσης άγνωστων δεδομένων. Σε αυτή τη φάση ως είσοδος δίνονται στον αλγόριθμο άγνωστα δεδομένα τα οποία ονομάζονται δεδομένα ελέγχου (test set) και ο στόχος είναι να μπορεί να προβλέψει ο αλγόριθμος την κλάση στην οποία θα πρέπει να ανήκουν τα δεδομένα αυτά, δηλαδή να τα κατηγοριοποιήσει στη σωστή κλάση. Η διαδικασία ελέγχου είναι πολύ απλή και υπολογιστικά ανέξοδη σε σύγκριση με τη φάση της εκπαίδευσης, η οποία είναι σύνθετη και απαιτεί σημαντικούς υπολογιστικούς πόρους. Εξαίρεση αποτελούν οι οκνηροί αλγόριθμοι κατηγοριοποίησης όπως ο KNN, οι οποίοι κατά την διαδικασία ελέγχου (στάδιο κατηγοριοποίησης) παρουσιάζουν υψηλό υπολογιστικό κόστος. Οι αλγόριθμοι αυτοί περιγράφονται αναλυτικότερα σε επόμενη παράγραφο.

Μετά το τέλος της παραπάνω διαδικασίας εφόσον το μοντέλο που κατασκευάστηκε (στην πρώτη φάση) παρουσιάζει ικανοποιητική επίδοση τότε χρησιμοποιείται για την διατύπωση προβλέψεων (Han & Kamber, 2006).

Η κατηγοριοποίηση βρίσκει εφαρμογή σε διάφορους τομείς της καθημερινής ζωής όπως τα Οικονομικά και την ιατρική. Ένα παράδειγμα οικονομικής φύσεως είναι αυτό των τραπεζών οι οποίες έχουν καταρτίσει μοντέλα ταξινόμησης για να κατηγοριοποιήσουν τραπεζικά δάνεια σε επικίνδυνα ή ασφαλή (Yoo et al., 2012). Από την άλλη στον ιατρικό τομέα, η

κατηγοριοποίηση μπορεί να χρησιμοποιείται για να βοηθήσει στον καθορισμό ιατρικής διάγνωσης και πρόγνωσης με βάση τα συμπτώματα και τις συνθήκες υγείας ενός ασθενούς. Ένα τέτοιο παράδειγμα αναφέρεται στο άρθρο (Rashid Ahmed Ahmed et al., 2019) που αφορά τον τρόπο με τον οποίο διαγιγνώσκεται ο καρκίνος του πνεύμονα σε έναν ασθενή μέσω της ακτινογραφίας θώρακος.

Κατά την εκπαίδευση ενός μοντέλου υπάρχει ο κίνδυνος της απομνημόνευσης του συγκεκριμένου συνόλου εκπαίδευσης από τον αλγόριθμο. Το φαινόμενο αυτό ονομάζεται υπερπροσαρμογή (overfitting) και μειώνει την ικανότητα γενίκευσης του μοντέλου, δηλαδή την επίδοση του κατά την πρόβλεψη άγνωστων δεδομένων.

2.3 Αλγόριθμοι κατηγοριοποίησης

Οι αλγόριθμοι μάθησης μπορούν να χωριστούν σε δύο κατηγορίες, τους lazy (οκνηρούς) και τους eager (πρόθυμους). Αντίστοιχα lazy και eager ονομάζονται και οι κατηγοριοποιητές που δημιουργούνται από τους αλγορίθμους αυτούς. Οι χαρακτηρισμοί προήλθαν με βάση τη χρονική στιγμή που ο κάθε κατηγοριοποιητής εκτελεί τη διαδικασία της γενίκευσης.

Οι eager αλγόριθμοι κατασκευάζουν κατηγοριοποιητές που περιέχουν μια σαφή υπόθεση που συνδέει τα μη επισημασμένα δεδομένα με τις προβλεπόμενες ετικέτες τους, δηλαδή την πρόβλεψη για την κλάση στην οποία ανήκουν. Ουσιαστικά κατασκευάζεται ένα μαθηματικό μοντέλο με βάση τα δεδομένα εκπαίδευσης που στη συνέχεια χρησιμοποιείται για την παραγωγή προβλέψεων, χωρίς να χρειάζεται πλέον τα δεδομένα εκπαίδευσης από τα οποία δημιουργήθηκε. Για το λόγο αυτό οι eager κατηγοριοποιητές ονομάζονται επίσης και model-based (βασισμένοι σε μοντέλο).

Αντίθετα, οι lazy αλγόριθμοι κατά τη φάση της εκπαίδευσης δεν δημιουργούν μοντέλο, αλλά αποθηκεύουν απλώς τα δεδομένα εκπαίδευσης. Τα αποθηκευμένα δεδομένα εκπαίδευσης θα τα χρησιμοποιήσουν αργότερα, όταν χρειαστεί να παραγάγουν προβλέψεις, δηλαδή να κατηγοριοποιήσουν νέα μη επισημασμένα δεδομένα. Αυτός είναι και ο λόγος για τον οποίο ονομάζονται lazy, αφού δεν κάνουν τίποτα μέχρι την στιγμή που θα χρειαστεί να παραγάγουν προβλέψεις. Επίσης οι συγκεκριμένοι αλγόριθμοι επειδή αποθηκεύουν τα δεδομένα εκπαίδευσης ονομάζονται και instance-based (βασισμένοι σε στιγμιότυπα) (J. H. Friedman et al., 1996).

Οι lazy αλγόριθμοι απαιτούν λίγο χρόνο υπολογισμού κατά τη διάρκεια της φάσης εκπαίδευσης, ουσιαστικά όσο χρόνο χρειάζεται για την αποθήκευση των δεδομένων εκπαίδευσης. Αντίθετα, η κατασκευή του μοντέλου ενός eager αλγορίθμου μπορεί να είναι πολύ χρονοβόρα και εξαρτάται τόσο από τον ίδιο τον αλγόριθμο, όσο και από την φύση και

τον όγκο των δεδομένων εκπαίδευσης. Όμως κατά τη φάση της πρόβλεψης τα πράγματα είναι διαφορετικά. Οι lazy αλγόριθμοι εκτελούν τους υπολογισμούς εκείνη τη στιγμή, οπότε ο απαιτούμενος υπολογιστικός χρόνος μπορεί να είναι υψηλός, ειδικά στην περίπτωση που το πλήθος των αποθηκευμένων δεδομένων εκπαίδευσης είναι μεγάλο. Ωστόσο, στην περίπτωση των eager αλγορίθμων που για την κατηγοριοποίηση χρησιμοποιούν το ήδη κατασκευασμένο μοντέλο, ο απαιτούμενος χρόνος είναι πάρα πολύ μικρός.

Μια ακόμα διαφορά των δύο κατηγοριών αλγορίθμων είναι ως προς τον απαιτούμενο χώρο αποθήκευσης. Οι lazy αλγόριθμοι πρέπει να έχουν πάντα διαθέσιμο ολόκληρο το σύνολο των δεδομένων εκπαίδευσης για να παραγάγουν προβλέψεις, πράγμα που αυξάνει τον απαιτούμενο αποθηκευτικό χώρο. Αντίθετα, οι eager αλγόριθμοι χρειάζονται τα δεδομένα εκπαίδευσης μόνο για την κατασκευή του μοντέλου. Κατά τη φάση της πρόβλεψης τα δεδομένα εκπαίδευσης δεν χρειάζονται πλέον. Επίσης το ίδιο το μοντέλο συνήθως απαιτεί πολύ λιγότερο αποθηκευτικό χώρο (Phyu, 2009).

Ο γνωστότερος ίσως lazy κατηγοριοποιητής είναι ο κατηγοριοποιητής k-Κοντινότερων Γειτόνων (k-Nearest Neighbors – k-NN). Χρησιμοποιεί τον αλγόριθμο του Κοντινότερου Γείτονα για να εντοπίσει τους κοντινότερους γείτονες ενός νέου αγνώστου προτύπου και να το κατηγοριοποιήσει με βάση την κλάση της πλειοψηφίας αυτών (Cover & Hart, 1967).

Αντίστοιχα, υπάρχει μεγάλη ποικιλία eager κατηγοριοποιητών με γνωστότερο ίσως τα Δέντρα Απόφασης (Decision Trees). Τα Δέντρα Απόφασης με βάση τα δεδομένα εκπαίδευσης κατασκευάζουν δεντρικές δομές, ικανές να κατηγοριοποιήσουν νέα άγνωστα πρότυπα (Rokach & Maimon, 2015).

Επίσης, τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Nets – ANN) και οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVM) που κατά την εκπαίδευση τους δημιουργούν μαθηματικά μοντέλα, με τα οποία στη συνέχεια μπορούν να εντάξουν νέα πρότυπα σε κλάσεις (Maimon & Rokach, 2010).

Μια άλλη κατηγορία eager κατηγοριοποιητών κατασκευάζει μοντέλα που βασίζονται σε υπολογισμούς πιθανοτήτων, όπως για παράδειγμα ο κατηγοριοποιητής Naïve Bayes (N. Friedman et al., 1997). Στην παρούσα διπλωματική χρησιμοποιούνται οι κατηγοριοποιητές k-NN, C4.5 και Naïve Bayes που θα αναλυθούν στη συνέχεια.

2.4 Αλγόριθμοι κατηγοριοποίησης

2.4.1 Δέντρα Αποφάσεων (Decision Trees)

Μια δημοφιλής μέθοδος κατηγοριοποίησης είναι τα Δέντρα Αποφάσεων (Decision Trees). Πρόκειται για δενδρικές δομές που κατασκευάζονται χρησιμοποιώντας τα διαθέσιμα δεδομένα, τα οποία προέρχονται από κάποιο πεδίο εφαρμογής. Αφού κατασκευαστεί το δέντρο απόφασης, μπορεί να χρησιμοποιηθεί για να αποφασιστεί σε ποια κλάση θα ανήκει ένα νέο πρότυπο, με βάση τα χαρακτηριστικά του, ακολουθώντας μια διαδρομή μέσα στο δέντρο. Γενικά λοιπόν η κατηγοριοποίηση με Δέντρα Απόφασης χωρίζεται σε δύο φάσεις: την φάση κατασκευής του Δέντρου Απόφασης (εκπαίδευση του μοντέλου) και τη φάση της κατηγοριοποίησης νέων δεδομένων με τη βοήθεια του Δέντρου Απόφασης που κατασκευάστηκε στην προηγούμενη φάση.

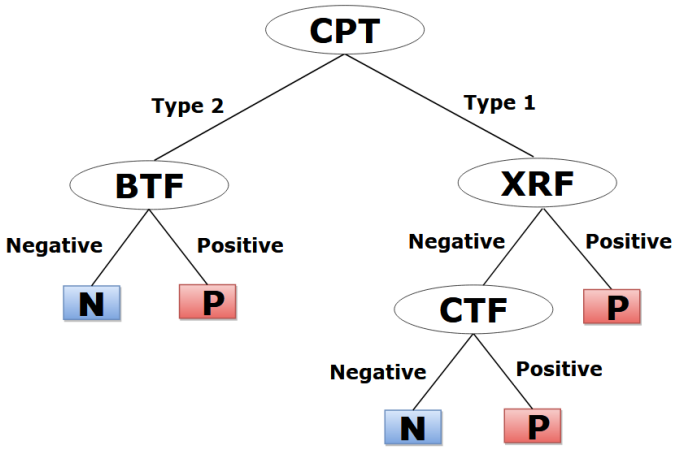
Η διαδικασία κατασκευής των δέντρων απόφασης βασίζεται στον αναδρομικό διαχωρισμό του συνόλου δεδομένων σε υποσύνολα με βάση τις τιμές κατάλληλα επιλεγμένων χαρακτηριστικών. Κάθε διαχωρισμός με βάση κάποιο χαρακτηριστικό, συμβολίζεται ως ένας κόμβος του δέντρου. Οι κόμβοι αυτοί στη συνέχεια συνδέονται μεταξύ τους με ακμές ανάλογα με τη σειρά του διαχωρισμού, δημιουργώντας έτσι μια δομή ανεστραμμένου δέντρου το οποίο ξεκινάει από ένα κόμβο-ρίζα και συνδέεται με άλλους κόμβους με κατεύθυνση από πάνω προς τα κάτω. Κάθε κόμβος μπορεί να έχει μόνο μια εισερχόμενη ακμή και καμία ή περισσότερες εξερχόμενες ακμές. Εξαίρεση αποτελεί ο κόμβος-ρίζα ο οποίος δεν έχει καμία εισερχόμενη ακμή. Όσοι κόμβοι έχουν εξερχόμενες ακμές ονομάζονται «εσωτερικοί», ενώ αυτοί που δεν έχουν ονομάζονται «τερματικοί ή απόφασης». Οι τερματικοί κόμβοι είναι επίσης γνωστοί και ως «φύλλα» του δέντρου γιατί βρίσκονται στα άκρα των κλαδιών. Η διαδικασία επιλογής χαρακτηριστικών και διαχωρισμού επαναλαμβάνεται μέχρι ο αλγόριθμος να φτάσει σε κόμβο που δεν μπορεί να διαχωρίσει, οπότε ο κόμβος αυτός αποτελεί φύλλο του δέντρου. Ο τερματισμός του αλγορίθμου εκπαίδευσης μπορεί να γίνει και με άλλα κριτήρια που περιγράφονται σε επόμενη παράγραφο. Τέλος, σε κάθε φύλλο αντιστοιχείται και η τιμή μίας κλάσης (Rokach & Maimon, 2015).

Η κατηγοριοποίηση των νέων προτύπων δεδομένων γίνεται ξεκινώντας από τη ρίζα του δέντρου. Στη συνέχεια διανύοντας μια διαδρομή μέσα στο δέντρο στόχος είναι ο τερματισμός σε κάποιο φύλλο-κλάση. Αρχικά λοιπόν, στη ρίζα του δέντρου συγκρίνεται η τιμή του επιλεγμένου χαρακτηριστικού με το κριτήριο διαχωρισμού, οδηγώντας σε κάποιο από τα κλαδιά που ξεκινούν από τον κόμβο-ρίζα. Αν το κλαδί συνδέεται σε ενδιάμεσο κόμβο, η διαδικασία που περιγράφηκε επαναλαμβάνεται για το χαρακτηριστικό που αντιστοιχεί στον

κόμβο αυτό. Ο αλγόριθμος συνεχίζεται μέχρι να φτάσει σε κόμβο-φύλλο, η τιμή του οποίου θα είναι το αποτέλεσμα της κατηγοριοποίησης (Rokach & Maimon, 2015).

Η δομή ενός δέντρου κατηγοριοποίησης μπορεί εύκολα να μετατραπεί σε ένα σύνολο κανόνων της μορφής IF-THEN, πράγμα που βοηθά ιδιαίτερα την ερμηνευσιμότητα (comprehensibility) του μοντέλου. Αυτό βέβαια ισχύει με την προϋπόθεση ότι το δέντρο παρουσιάζει σχετικά χαμηλή πολυπλοκότητα, άρα και οι αντίστοιχοι κανόνες κατηγοριοποίησης θα είναι σύντομοι και εύκολα κατανοητοί. Η πολυπλοκότητα ενός δέντρου συνδέεται άμεσα με το μέγεθος του και μπορεί να μετρηθεί με βάση κάποιες μετρικές όπως: τον συνολικό αριθμό κόμβων και φύλλων, τον αριθμό των χαρακτηριστικών που χρησιμοποιούνται και το βάθος του δέντρου.

Ένα παράδειγμα απλού δέντρου κατηγοριοποίησης φαίνεται στην Εικόνα 1 το οποίο χρησιμοποιείται για την διάγνωση ασθενών που πάσχουν από κάποια ασθένεια του αναπνευστικού συστήματος. Τα χαρακτηριστικά που χρησιμοποιούνται είναι: CPT (Chest Pain Type – Τύπος πόνου στο στήθος), CTF (CT Findings – Αποτέλεσμα Αξονικής Τομογραφίας), XRF (X Ray Findings – Αποτελέσματα Ακτινογραφίας), BTF (Blood Test Findings – Αποτέλεσμα Αιματολογικής Εξέτασης).



Εικόνα 1: Δέντρο αποφάσεων για ιατρικές εφαρμογές

Είναι προφανές ότι αν διαφοροποιηθεί η σειρά των χαρακτηριστικών με βάση τα οποία γίνεται ο διαχωρισμός, θα προκύψουν δέντρα διαφορετικής μορφής, τα οποία κατά συνέπεια θα δίνουν διαφορετικά αποτελέσματα ως προς την κατηγοριοποίηση. Για το λόγο αυτό είναι σημαντικό να επιλέγεται κάθε φορά το χαρακτηριστικό που μπορεί να διαχωρίσει καλύτερα τις τελικές κλάσεις. Υπάρχουν πολλές τεχνικές επιλογής του «καλύτερου» χαρακτηριστικού οι οποίες βασίζονται σε κάποιο μέτρο ποιότητας και χρησιμοποιούνται ανάλογα με τον αλγόριθμο κατασκευής δέντρου που εφαρμόζεται κάθε φορά (Murthy, 1998). Πολλές από τις

μεθόδους μέτρησης της ποιότητας ενός χαρακτηριστικού βασίζονται στην έννοια της εντροπίας της πληροφορίας.

Κριτήρια τερματισμού εκπαίδευσης δέντρου

Η φάση κατασκευής του δέντρου συνεχίζεται μέχρις ότου ενεργοποιηθεί κάποιο κριτήριο τερματισμού. Τα πιο συνηθισμένα κριτήρια τερματισμού είναι τα παρακάτω:

1. Η επέκταση ενός κόμβου τερματίζεται όταν όλες οι εγγραφές του ανήκουν στην ίδια κλάση.

2. Έχει επιτευχθεί το μέγιστο βάθος δέντρου. Βάθος ενός δέντρου απόφασης είναι το μήκος της μεγαλύτερης διαδρομής από τη ρίζα σε κάποιο φύλλο.

3. Αν έχει οριστεί από πριν, ότι θα πρέπει να υπάρχει ένας ελάχιστος αριθμός προτύπων, ώστε να δημιουργηθεί τερματικός κόμβος κλάσης, αλλά ο αριθμός των προτύπων που αντιστοιχούν σε αυτό τον κόμβο, είναι κάτω από αυτό το όριο, δεν δημιουργείται τερματικός κόμβος και η επέκταση του δέντρου σταματάει στον προηγούμενο κόμβο.

4. Η επέκταση ενός κόμβου τερματίζεται όταν δεν υπάρχει άλλο χαρακτηριστικό προς διαχωρισμό. Στη συνέχεια δημιουργείται φύλλο με κλάση εκείνη που εμφανίζεται με τη μεγαλύτερη συχνότητα.

5. Η επέκταση ενός κόμβου τερματίζεται όταν η τιμή του κριτηρίου διαχωρισμού (πχ InfoGain, GainRatio) κάθε χαρακτηριστικού είναι κάτω από ένα συγκεκριμένο όριο, το οποίο έχει οριστεί από πριν.

Μέθοδοι κλαδέματος

Η επιλογή του κριτηρίου τερματισμού εκπαίδευσης επηρεάζει το μέγεθος και την πολυπλοκότητα του δέντρου που θα δημιουργηθεί. Η χρησιμοποίηση αυστηρών κριτηρίων τερματισμού τείνει να δημιουργεί μικρά και υποπροσαρμοσμένα (underfitted) δέντρα. Από την άλλη πλευρά, η χρήση χαλαρών κριτηρίων τερματισμού τείνει να δημιουργεί μεγάλα δέντρα που είναι υπερπροσαρμοσμένα (overfitted) στο σύνολο εκπαίδευσης. Και στις δύο αυτές περιπτώσεις η χρήση του δέντρου απόφασης για κατηγοριοποίηση νέων δεδομένων δίνει μη-βέλτιστα αποτελέσματα. Για την επίλυση αυτού του προβλήματος προτάθηκε από τους (Breiman et al., 1984) μια τεχνική που ονόμασαν κλάδεμα (pruning) και συγκεκριμένα cost-complexity pruning. Για την εφαρμογή της τεχνικής αυτής, κατά την φάση εκπαίδευσης χρησιμοποιείται ένα χαλαρό κριτήριο τερματισμού, έτσι ώστε το δέντρο που θα κατασκευαστεί να υπερπροσαρμοστεί στο σύνολο εκπαίδευσης. Στη συνέχεια, το υπερπροσαρμοσμένο αυτό δέντρο μικραίνει, αφαιρώντας (κλάδεμα) τα κλαδιά (υποδέντρα) που δεν συμβάλλουν στην ακρίβεια της γενίκευσης.

Το κλάδεμα όμως μπορεί να γίνει σε διάφορα επίπεδα και σε μεγάλο ή μικρό βαθμό, οπότε το να βρεθεί το βέλτιστο κλαδεμένο δέντρο δεν είναι εύκολη διαδικασία. Καταρχάς για να ελεγχθεί η επίπτωση του κάθε πιθανού τρόπου κλαδέματος πρέπει να ελεγχθεί η ακρίβεια της κατηγοριοποίησης σε άγνωστα δεδομένα. Για το λόγο αυτό η τεχνική cost-complexity pruning, αλλά και άλλες που προτάθηκαν αργότερα όπως Reduced Error Pruning και Minimum Error Pruning χρησιμοποιούν ένα τμήμα των δεδομένων εκπαίδευσης για το έλεγχο του κλαδέματος. Οι τεχνικές αυτές έχουν καλά αποτελέσματα σε μεγάλα σύνολα δεδομένων. Αντίθετα σε σχετικά μικρά σύνολα δεδομένων τα αποτελέσματα δεν είναι τόσο καλά γιατί τα δεδομένα εκπαίδευσης που απομένουν για την κατασκευή του δέντρου είναι ακόμα λιγότερα.

Για την αποφυγή του παραπάνω προβλήματος έχουν προταθεί και τεχνικές οι οποίες δεν χρειάζονται επιπλέον δεδομένα εκτός από τα δεδομένα εκπαίδευσης. Για να το πετύχουν αυτό χρησιμοποιούν τεχνικές στατιστικής ανάλυσης για να εκτιμήσουν το σφάλμα που είναι πιθανό να προκύψει όταν θα κατηγοριοποιηθούν άγνωστα δεδομένα. Οι τεχνικές αυτές είναι η Pessimistic Pruning και Error-Based Pruning από τον ίδιο ερευνητή. Η τεχνική Error-Based Pruning είναι μετεξέλιξη της Pessimistic και χρησιμοποιείται από τον αλγόριθμο C4.5 που θα περιγραφεί στη συνέχεια.

Διάφορες μελέτες έδειξαν ότι η χρήση τεχνικών κλαδέματος μπορεί να βελτιώσει την ικανότητα γενίκευσης ενός δέντρου αποφάσεων, ειδικά σε περιπτώσεις που τα δεδομένα περιέχουν θόρυβο. Παρόλα αυτά, αποδείχτηκε ότι δεν υπάρχει κάποια τεχνική κλαδέματος που να έχει καλύτερα αποτελέσματα από τις υπόλοιπες σε όλες τις περιπτώσεις προβλημάτων και συνόλων δεδομένων (Rokach & Maimon, 2015)**.**

2.4.1.1 Αλγόριθμος ID3

Ο αλγόριθμος ID3 (Iterative Dichotomiser 3) θεωρείται ένας πολύ απλός αλγόριθμος κατασκευής δέντρων αποφάσεων που προτάθηκε από τον (Quinlan, 1986). Ως κριτήριο διαχωρισμού χρησιμοποιεί το κέρδος πληροφορίας (Information Gain) και σταματάει την κατασκευή του δέντρου (κριτήριο τερματισμού) όταν όλα τα πρότυπα ανήκουν σε μία μόνο τιμή κλάσης ή όταν το καλύτερο κέρδος πληροφορίας δεν είναι μεγαλύτερο από το μηδέν.

Το κέρδος πληροφορίας είναι μια μετρική μη καθαρότητας κόμβου (Node impurity metric) που βασίζεται στην έννοια της εντροπίας. Η εντροπία στη «Θεωρία Πληροφορίας» εκφράζει ένα «μέτρο αβεβαιότητας» που διακατέχει ένα σύστημα. Στη συγκεκριμένη περίπτωση η εντροπία είναι το μέτρο της ανομοιογένειας ενός συνόλου δεδομένων S ανάλογα με την κλάση στην οποία ανήκουν τα πρότυπα που περιέχει.

Αν για παράδειγμα ένα σύνολο δεδομένων S περιέχει s πρότυπα που ανήκουν σε δύο κλάσεις (1 και 2) τότε το s1 είναι το πλήθος των προτύπων της κλάσης 1, ενώ το s2 είναι το πλήθος των προτύπων της κλάσης 2.

Έστω p_1, p_2 είναι τα ποσοστά των προτύπων της κλάσης 1 και 2 αντίστοιχα και υπολογίζονται ως:

$$p_1 = \frac{s1}{s} \; και \; p_2 = \frac{s2}{s}$$

Η εντροπία για το σύνολο δεδομένων S δίνεται από τον παρακάτω τύπο:

$$Entropy(S) = -p_1 \, log_2(p_1) - p_2 \, log_2(p_2)$$

Ισχύει η παραδοχή $\log_2 0 = 0$ και η μονάδα μέτρησης της εντροπίας είναι το bit.

Εάν υπάρχουν περισσότερες από δύο κλάσεις (n > 2) ο τύπος γενικεύεται ως εξής:

$$Entropy(S) = -\sum_{i=1}^{n} p_i \, log_2(p_i)$$

Ας θεωρηθεί ότι ένα χαρακτηριστικό A του συνόλου δεδομένων μπορεί να πάρει κ δυνατές διακριτές τιμές (δηλαδή α_1, α_2, …, α_κ). Με βάση το χαρακτηριστικό A μπορεί το σύνολο S να χωριστεί σε κ υποσύνολα (S1, S2, …., Sκ), όπου σε κάθε υποσύνολο τα πρότυπα θα έχουν την ίδια τιμή στο χαρακτηριστικό A. Για κάθε υποσύνολο μπορεί να υπολογιστεί η αντίστοιχη εντροπία $Entropy(S_j)$ και τελικά να υπολογιστεί και η μέση εντροπία του διαχωρισμού του συνόλου S σε υποσύνολα ανάλογα με τις τιμές του χαρακτηριστικού A, με τον τύπο:

$$Entropy(S, A) = \sum_{j=1}^{k} p_j \, Entropy(S_j)$$

Όπου κ το πλήθος των δυνατών τιμών του χαρακτηριστικού A, S_j το υποσύνολο των προτύπων που έχουν τιμή a_j στο χαρακτηριστικό A, p_j το ποσοστό των προτύπων του S που ανήκουν στο S_j και Entropy(S_j) είναι η εντροπία του S_j που υπολογίζεται σύμφωνα με τον προηγούμενο τύπο.

Τελικά το κέρδος πληροφορίας (Information Gain) είναι η μείωση της αρχικής εντροπίας όταν γίνεται ο διαχωρισμός με βάση το χαρακτηριστικό A και δίνεται από τον τύπο:

$$InfoGain(S, A) = \; Entropy(S) - \; Entropy(S, A)$$

Έτσι όσο περισσότερο μειώνεται η εντροπία κατά το διαχωρισμό με βάση κάποιο χαρακτηριστικό, τόσο μεγαλύτερο είναι το κέρδος πληροφορίας. Ο αλγόριθμος ID3 υπολογίζει το κέρδος πληροφορίας για κάθε χαρακτηριστικό και επιλέγει κάθε φορά εκείνο που παρουσιάζει την μεγαλύτερη τιμή κέρδους πληροφορίας.

Ο αλγόριθμος ID3 δεν εφαρμόζει καμία τεχνική κλαδέματος και επίσης μπορεί να χειριστεί μόνο κατηγορικά χαρακτηριστικά. Αυτό σημαίνει ότι αν υπάρχουν αριθμητικά χαρακτηριστικά πρέπει πρώτα να μετατραπούν σε κατηγορικά. Τέλος, δεν μπορεί να χειριστεί τυχόν ελλιπείς τιμές (missing values). Το κύριο πλεονέκτημα του ID3 είναι η απλότητά του και για αυτόν τον λόγο χρησιμοποιείται κυρίως για διδακτικούς σκοπούς.

2.4.1.2 Αλγόριθμος C4.5

Ο αλγόριθμος C4.5, είναι μια εξέλιξη του ID3, που παρουσιάστηκε από τον ίδιο ερευνητή (Quinlan, 1993).

Ο C4.5 χρησιμοποιεί τον λόγο κέρδους (gain ratio) ως κριτήριο διαχωρισμού και η κατασκευή σταματά (κριτήριο τερματισμού) όταν ο αριθμός των προτύπων που πρόκειται να διαχωριστούν είναι κάτω από ένα ορισμένο όριο.

Η επιλογή χαρακτηριστικού με βάση το κέρδος πληροφορίας δεν λειτουργεί καλά στις περιπτώσεις χαρακτηριστικών με μεγάλο πλήθος δυνατών τιμών. Αυτό συμβαίνει γιατί προκύπτουν πολλά μικρά και ομοιογενή σύνολα τα οποία φαινομενικά έχουν μεγάλο κέρδος πληροφορίας χωρίς όμως να περιέχουν πάντα ουσιαστικά χρήσιμη πληροφορία. Για την αντιμετώπιση αυτού του προβλήματος ο C4.5 χρησιμοποιεί τον λόγο κέρδους (gain ratio) που ουσιαστικά είναι το κανονικοποιημένο κέρδος πληροφορίας ως προς την εντροπία και η χρήση του βελτιώνει την ακρίβεια κατηγοριοποίησης.

$$GainRatio(S, A) = \frac{InfoGain(S, A)}{Entropy(S, A)}$$

Ο C4.5 μπορεί να χειριστεί εκτός από κατηγορικά και αριθμητικά χαρακτηριστικά, καθώς και τυχόν ελλιπείς τιμές χαρακτηριστικών. Μετά την φάση κατασκευής του δέντρου εφαρμόζει κλάδεμα βάσει σφαλμάτων (error-based pruning).

Error-based pruning

Η τεχνική του κλαδέματος βάσει σφαλμάτων (error-based pruning) βασίζεται σε στατιστικές τεχνικές για να προβλέψει το ποσοστό των λανθασμένων κατηγοριοποιήσεων που είναι πιθανό να γίνουν σε άγνωστα δεδομένα. Η πρόβλεψη όμως αυτή γίνεται με βάση τα δεδομένα εκπαίδευσης οπότε δεν μπορεί να θεωρηθεί ακριβής, αλλά μπορεί να εκτιμηθεί με κάποια πιθανότητα μέσα σε συγκεκριμένα όρια εμπιστοσύνης (confidence limits).

Συγκεκριμένα, αν υποτεθεί ότι σε ένα κόμβο του δέντρου κατηγοριοποιηθούν N πρότυπα και τα E από αυτά είναι λανθασμένα, τότε το E / N που είναι το ποσοστό του σφάλματος για τον κόμβο αυτό, είναι αντιπροσωπευτικό για όλα τα πρότυπα που μπορεί να φτάσουν σε αυτό τον

κόμβο. Αυτό φυσικά δεν ισχύει απόλυτα, αλλά με βάση αυτό το ποσοστό σφάλματος μπορεί να υπολογιστεί το πιθανό αναμενόμενο σφάλμα για όλα τα άγνωστα πρότυπα. Δηλαδή, χρησιμοποιώντας το διωνυμικό θεώρημα και για ένα δεδομένο επίπεδο εμπιστοσύνης υπολογίζονται τα αντίστοιχα όρια εμπιστοσύνης μέσα στα οποία κυμαίνεται το πιθανό εκτιμώμενο σφάλμα. Μάλιστα επειδή αυτή η εκτίμηση βασίζεται μόνο στα δεδομένα εκπαίδευσης λαμβάνεται υπόψιν το άνω όριο εμπιστοσύνης, δηλαδή η μεγαλύτερη πιθανότητα σφάλματος.

Στη συνέχεια με βάση αυτήν την πιθανότητα μπορεί να υπολογιστεί το αναμενόμενο σφάλμα κατηγοριοποίησης σε ένα κόμβο ή υποδέντρο και επίσης το αναμενόμενο σφάλμα αν ο συγκεκριμένος κόμβος ή υποδέντρο κλαδευτεί και αντικατασταθεί με φύλλο. Αν το σφάλμα στη δεύτερη περίπτωση είναι χαμηλότερο, το κλάδεμα εκτελείται.

Το επίπεδο εμπιστοσύνης εκφράζει το πόσο σίγουρο είναι ότι τα όρια εμπιστοσύνης που θα προκύψουν στους υπολογισμούς θα περιέχουν την πραγματική πιθανότητα του σφάλματος. Έτσι, για μικρότερο επίπεδο εμπιστοσύνης, τα όρια εμπιστοσύνης θα είναι ευρύτερα, οπότε και το άνω όριο εμπιστοσύνης θα είναι μεγαλύτερο. Αυτό έχει ως αποτέλεσμα η αναμενόμενη πιθανότητα σφάλματος να είναι μεγαλύτερη, δηλαδή κάποιο πρότυπο είναι πιο πιθανό να κατηγοριοποιηθεί λανθασμένα, άρα θα γίνει περισσότερο κλάδεμα. Για μεγαλύτερες τιμές επιπέδου εμπιστοσύνης ισχύει το αντίθετο οπότε εκτελείται λιγότερο κλάδεμα. Το επίπεδο εμπιστοσύνης για τον υπολογισμό της αναμενόμενης πιθανότητας σφάλματος αποτελεί παράμετρο του αλγορίθμου C4.5 που ονομάζεται Confidence Factor (CF). Το Confidence Factor στον C4.5 έχει προεπιλεγμένη τιμή 0.25, δηλαδή αναφέρεται σε επίπεδο εμπιστοσύνης 25% (Hall et al., 2003).

Σύγκριση αλγορίθμων ID3 και C4.5

Ο αλγόριθμος C4.5 παρέχει αρκετές βελτιώσεις σε σύγκριση με τον ID3. Οι σημαντικότερες βελτιώσεις είναι:

(1) Ο C4.5 χρησιμοποιεί μια τεχνική κλαδέματος που αφαιρεί τα κλαδιά που δεν συμβάλλουν στην βελτίωση της ακρίβειας κατηγοριοποίησης και τα αντικαταστεί με φύλλα.

(2) Ο C4.5 επιτρέπει να υπάρχουν ελλιπείς τιμές χαρακτηριστικών.

(3) Ο C4.5 μπορεί να χειριστεί συνεχή χαρακτηριστικά. Αυτό το κάνει διαχωρίζοντας τις τιμές του χαρακτηριστικού σε δύο υποσύνολα (δυαδικός διαχωρισμός). Συγκεκριμένα, αναζητά το καλύτερο όριο διαχωρισμού (κατώφλι), δηλαδή αυτό που μεγιστοποιεί το κριτήριο του λόγου κέρδους (gain ratio). Στη συνέχεια τοποθετεί όλες τις τιμές που είναι μεγαλύτερες από το κατώφλι στο πρώτο υποσύνολο και όλες τις άλλες τιμές στο δεύτερο υποσύνολο.

2.4.1.3 Αλγόριθμος J48

Ο αλγόριθμος J48 είναι μια υλοποίηση ανοικτού κώδικα σε Java του αλγορίθμου C4.5 και υπάρχει στο εργαλείο εξόρυξης δεδομένων Weka (Rokach & Maimon, 2015).

Παράδειγμα υπολογισμού GainRatio με τον αλγόριθμο C4.5

Στη συνέχεια περιγράφεται ένα παράδειγμα υπολογισμού του GainRatio για την κατασκευή δέντρου αποφάσεων με τον αλγόριθμο C4.5. Τα δεδομένα εκπαίδευσης αποτελούν ένα δείγμα 24 εγγραφών από το σύνολο δεδομένων **Medical Appointments (Alvaro Flores)** όπως φαίνεται στον παρακάτω Πίνακα (Πίνακας 1).

No.	**age**	**sex**	**appointment_ hour_d**	**waiting_ days**	**Commu- nication_ channel**	**no_ show (class)**
1	11.0	1	14.0	1.0	3	No
2	2.0	2	15.0	0.0	1	No
3	11.0	2	19.0	8.0	3	Yes
4	61.0	2	14.0	8.0	1	No
5	6.0	2	19.0	16.0	3	No
6	9.0	2	15.0	29.0	1	No
7	66.0	1	18.0	0.0	2	No
8	13.0	1	19.0	5.0	3	Yes
9	47.0	2	9.0	4.0	3	No
10	9.0	2	17.0	0.0	1	No
11	49.0	2	12.0	1.0	3	No
12	51.0	1	14.0	2.0	1	No
13	68.0	1	8.0	6.0	1	No
14	29.0	1	18.0	2.0	2	No
15	70.0	1	16.0	2.0	1	No
16	20.0	2	9.0	1.0	3	No
17	3.0	2	17.0	10.0	1	No
18	8.0	1	16.0	2.0	1	Yes
19	54.0	1	16.0	6.0	1	Yes

20	3.0	1	17.0	8.0	3	No
21	24.0	2	16.0	1.0	1	No
22	50.0	2	11.0	14.0	1	No
23	45.0	2	17.0	4.0	2	No
24	6.0	1	16.0	0.0	3	No

Πίνακας 1: Σύνολο δεδομένων εκπαίδευσης S

Στα παραπάνω δεδομένα το χαρακτηριστικό της κλάσης είναι το no_show που μπορεί να πάρει τιμές Yes και No. Αρχικά υπολογίζονται τα ποσοστά των προτύπων των κλάσεων Yes και No. Έστω s το συνολικό πλήθος προτύπων και s_{No} και s_{Yes} το πλήθος προτύπων των κλάσεων No και Yes αντίστοιχα.

$$p_{No} = \frac{s_{No}}{s} = \frac{20}{24} = 0{,}8333$$

$$p_{Yes} = \frac{s_{Yes}}{s} = \frac{4}{24} = 0{,}1667$$

Η αρχική εντροπία για το σύνολο S δίνεται από τον τύπο:

$$Entropy(S) = -\,p_{No}\,log_2(p_{No}) - p_{Yes}\,log_2(p_{Yes})$$

$$Entropy(S) = -\frac{20}{24}\,log_2\left(\frac{20}{24}\right) - \frac{4}{24}log_2\left(\frac{4}{24}\right) =$$

$$= 0{,}2192 + 0{,}4308 = 0{,}65 \text{ bits}$$

Στη συνέχεια πρέπει να υπολογιστεί το κέρδος πληροφορίας για κάθε χαρακτηριστικό του συνόλου S. Έτσι για παράδειγμα για το κατηγορικό χαρακτηριστικό sex που παίρνει τιμές 1 και 2, τα αποτελέσματα φαίνονται στον παρακάτω πίνακα (Πίνακας 2):

	Σύνολο Προτύπων	**Πρότυπα κλάσης No**	**P(No)**	**Πρότυπα κλάσης Yes**	**P(Yes)**
Sex = 1	11	8	8/11	3	3/11
Sex = 2	13	12	12/13	1	1/13

Πίνακας 2: Κατηγορικό χαρακτηριστικό Sex

Υπολογίζεται η εντροπία για κάθε τιμή του χαρακτηριστικού sex

$$Entropy(S, sex = 1) = -\,p_{No}\,log_2(p_{No}) - p_{Yes}\,log_2(p_{Yes}) =$$

$$-\frac{8}{11}log_2\left(\frac{8}{11}\right) - \frac{3}{11}log_2\left(\frac{3}{11}\right) = = 0{,}3341 + 0{,}5112 = 0{,}8453\ bits$$

$$Entropy(S, sex = 2) = -\,p_{No}\,log_2(p_{No}) - p_{Yes}\,log_2(p_{Yes}) =$$

$$-\frac{12}{13}log_2\left(\frac{12}{13}\right) - \frac{1}{13}log_2\left(\frac{1}{13}\right) == 0{,}1066 + 0{,}2846 = 0{,}3912\ bits$$

Αν επιλεγεί το χαρακτηριστικό sex για τον διαχωρισμό του συνόλου, τότε για να υπολογιστεί η μέση εντροπία χρησιμοποιείται ο τύπος:

$$Entropy(S, sex) = \sum_{j=1}^{2} p_j \, Entropy(S_j) =$$

$$p_{sex=1} \cdot Entropy\ (sex = 1) + p_{sex=2} \cdot Entropy(sex = 2) =$$

$$= \frac{11}{24} \cdot 0{,}8453 + \frac{13}{24} \cdot 0{,}3912 = 0{,}5993\ bits$$

Για να βρεθεί το κέρδος πληροφορίας για το συγκεκριμένο διαχωρισμό:

$$InfoGain(S, sex) = Entropy(S) - Entropy(S, sex) = 0{,}65 - 0{,}5993 = 0{,}0507\ bits$$

Τελικά, αν επιλεγεί το χαρακτηριστικό sex, τότε ο λόγος κέρδους (gain ratio) είναι:

$$GainRatio(S, sex) = \frac{InfoGain(S, sex)}{Entropy(S, sex)} \Longrightarrow$$

$$GainRatio(S, sex) = \frac{0{,}0507}{0{,}5993} = 0{,}084$$

Η παραπάνω διαδικασία υπολογισμού επαναλαμβάνεται για όλα τα υπόλοιπα χαρακτηριστικά. Έτσι, αν επιλεγεί το κατηγορικό χαρακτηριστικό communication_channel τα αποτελέσματα φαίνονται στον παρακάτω πίνακα (Πίνακας 3):

	Σύνολο Προτύπων	**Πρότυπα κλάσης No**	**P(No)**	**Πρότυπα κλάσης Yes**	**P(Yes)**
communication_channel=1	12	10	10/12	2	2/12
communication_channel=2	3	3	3/3	0	0/3
communication_channel=3	9	7	7/9	2	2/9

Πίνακας 3: Κατηγορικό χαρακτηριστικό communication_channel

Υπολογίζεται η εντροπία για κάθε τιμή του χαρακτηριστικού communication_channel

$$Entropy(S, communication\_channel = 1) = -\ p_{No}\ log_2(p_{No}) - p_{Yes}\ log_2(p_{Yes}) =$$

$$-\frac{10}{12} log_2\left(\frac{10}{12}\right) - \frac{2}{12} log_2\left(\frac{2}{12}\right) = 0{,}2191 + 0{,}4308 = 0{,}6499\ bits$$

$$Entropy(S, communication\_channel = 2) = -\ p_{No}\ log_2(p_{No}) - p_{Yes}\ log_2(p_{Yes}) =$$

$$-\frac{3}{3} log_2\left(\frac{3}{3}\right) - \frac{0}{3} log_2(0) = 0 + 0 = 0\ bits$$

$$Entropy(S, communication\_channel = 3) = -\ p_{No}\ log_2(p_{No}) - p_{Yes}\ log_2(p_{Yes}) =$$

$$-\frac{7}{9} log_2\left(\frac{7}{9}\right) - \frac{2}{9} log_2\left(\frac{2}{9}\right) = 0{,}282 + 0{,}4822 = 0{,}7642\ bits$$

Η μέση εντροπία θα είναι:

$$Entropy(S, communication\_channel) = \sum_{j=1}^{3} p_j \, Entropy(S_j) =$$

$$p_{communication\_channel=1} \cdot Entropy\ (communication\_channel = 1)$$
$$+ p_{communication\_channel=2} \cdot Entropy(communication\_channel = 2)$$
$$+ p_{communication\_channel=3} \cdot Entropy(communication\_channel = 3) =$$
$$= \frac{12}{24} \cdot 0{,}6499 + \frac{3}{24} \cdot 0 + \frac{9}{24} \cdot 0{,}7642 = 0{,}6115\ bits$$

Για να βρεθεί το κέρδος πληροφορίας για το συγκεκριμένο διαχωρισμό:

$$InfoGain(S, communication\_channel)$$
$$= Entropy(S) - Entropy(S, communication\_channel)$$
$$= 0{,}65 - 0{,}6115 = 0{,}0385\ bits$$

Τελικά, αν επιλεγεί το χαρακτηριστικό communication_channel, τότε ο λόγος κέρδους (gain ratio) θα είναι:

$$GainRatio(S, communication\_channel) = \frac{InfoGain(S, communication\_channel)}{Entropy(S, communication\_channel)}$$
$$= \frac{0{,}0385}{0{,}6115} = 0{,}063$$

Για τα συνεχή χαρακτηριστικά όπως το appointment_hour_d, η προσέγγιση είναι λίγο διαφορετική σε σχέση με τα κατηγορικά χαρακτηριστικά. Ουσιαστικά, πρέπει να γίνει μετατροπή του συνεχούς χαρακτηριστικού σε κατηγορικό χρησιμοποιώντας τεχνική διακριτοποίησης. Έτσι, λαμβάνοντας υπόψιν την ελάχιστη και μέγιστη δυνατή τιμή του συνεχούς χαρακτηριστικού χωρίζεται το διάστημα τιμών σε δύο υποδιαστήματα S_1 και S_2 με βάση κάποια ενδιάμεση τιμή διαχωρισμού *c* έτσι ώστε κάθε τιμή του S_1 να είναι μικρότερη ή ίση του *c* (<=c) και κάθε τιμή του S_2 να είναι μεγαλύτερη του *c* (>c).

Η βέλτιστη επιλογή της τιμής διαχωρισμού θα είναι αυτή που προκαλεί το μέγιστο κέρδος πληροφορίας, οπότε για να βρεθεί πρέπει να γίνει ο υπολογισμός της εντροπίας για κάθε δυνατή τιμή διαχωρισμού.

Για παράδειγμα το συνεχές χαρακτηριστικό appointment_hour_d έχει ελάχιστη τιμή 8, μέγιστη τιμή 19 και ενδιάμεσες τιμές 9,11,12,14,15,16,17,18. Οι ενδιάμεσες τιμές είναι πιθανά σημεία διαχωρισμού και διαχωρίζουν το εύρος τιμών σε δύο τμήματα κάθε φορά. Η ελάχιστη τιμή και μέγιστη τιμή δεν μπορεί να είναι σημεία διαχωρισμού γιατί δεν υπάρχουν τιμές μικρότερες από 8 και μεγαλύτερες από 18 αντίστοιχα. Για κάθε διαχωριστική τιμή θα πρέπει να υπολογιστεί η εντροπία και να βρεθεί το αντίστοιχο κέρδος πληροφορίας. Για παράδειγμα, αν επιλεγεί ως τιμή διαχωρισμού το 15 θα πρέπει να γίνουν οι παρακάτω υπολογισμοί (Πίνακας 4):

	Σύνολο Προτύπων	Πρότυπα κλάσης No	P(No)	Πρότυπα κλάσης Yes	P(Yes)
appointment_ hour_d <= 15	10	10	10/10	0	0/10
appointment_ hour_d > 15	14	10	10/14	4	4/14

Πίνακας 4: Επιλογή τιμής διαχωρισμού το 15 για το συνεχές χαρακτηριστικό appointment_hour_d

Υπολογίζεται η εντροπία για κάθε υποδιάστημα του χαρακτηριστικού appointment_hour_d:

$$Entropy(S, appointment\_hour\_d \leq 15) = -\,p_{No}\,log_2(p_{No}) - p_{Yes}\,log_2(p_{Yes}) =$$

$$-\frac{10}{10}log_2\left(\frac{10}{10}\right) - \frac{0}{10}log_2\left(\frac{0}{10}\right) = -1\,log_2(1) - 0\,log_2(0) = 0 + 0 = 0\ bits$$

$$Entropy(S, appointment\_hour\_d > 15) = -\,p_{No}\,log_2(p_{No}) - p_{Yes}\,log_2(p_{Yes}) =$$

$$-\frac{10}{14}log_2\left(\frac{10}{14}\right) - \frac{4}{14}log_2\left(\frac{4}{14}\right) = 0{,}3467 + 0{,}5164 = 0{,}8631\ bits$$

Για να υπολογιστεί η μέση εντροπία αν επιλεγεί το χαρακτηριστικό appointment_hour_για τον διαχωρισμό του συνόλου, χρησιμοποιείται ο τύπος:

$$Entropy(S, appointment\_hour\_d(15)) = \sum_{j=1}^{2} p_j\,Entropy(S_j) =$$

$$p_{\text{appointment\_hour\_d}\leq 15} \cdot Entropy\ (appointment\_hour\_d \leq 15) + p_{\text{appointment\_hour\_d}>15}$$

$$\cdot Entropy(appointment\_hour\_d\ > 15) = \frac{10}{24}\cdot 0 + \frac{14}{24}\cdot 0{,}8631$$

$$= 0{,}5035\ bits$$

Για να βρεθεί το κέρδος πληροφορίας για το συγκεκριμένο διαχωρισμό:

$$InfoGain(S, appointment\_hour\_d(15))$$

$$= Entropy(S) - Entropy(S, appointment\_hour\_d(15)) =$$

$$InfoGain(S, appointment\_hour\_d) = 0{,}65 - 0{,}5035 = 0{,}1465\ bits$$

Τελικά, ο λόγος κέρδους (gain ratio) για την επιλογή του χαρακτηριστικού appointment_hour_d θα είναι:

$$GainRatio(S, appointment\_hour\_d(15)) = \frac{InfoGain(S, appointment\_hour\_d(15))}{Entropy(S, appointment\_hour\_d(15))}$$

$$= \frac{0{,}1465}{0{,}5035} = 0{,}291$$

Αντίστοιχα, αν γίνουν οι υπολογισμοί και για τις υπόλοιπες πιθανές διαχωριστικές τιμές θα προκύψουν τα αποτελέσματα που φαίνονται στον παρακάτω πίνακα (Πίνακας 5):

Διαχωριστική τιμή c	$\boldsymbol{GainRatio(S, appointment\_hour\_d(c))}$
9	0,057
11	0,080
12	0,106
14	0,202
15	**0,291**
16	0,015
17	0,108
18	0,270

Πίνακας 5: Αποτελέσματα GainRatio για όλες τις πιθανές τιμές διαχωρισμού για το συνεχές χαρακτηριστικό appointment_hour_d

Από τον πίνακα φαίνεται ότι η μέγιστη τιμή του λόγου κέρδους για το χαρακτηριστικό appointment_hour_d, επιτυγχάνεται αν χρησιμοποιηθεί ως διαχωριστική τιμή το 15, το οποίο και χρησιμοποιήθηκε στο παραπάνω παράδειγμα υπολογισμού.

Παρομοίως, το χαρακτηριστικό waiting_days έχει ελάχιστη τιμή 0, μέγιστη τιμή 29 και ενδιάμεσες τιμές 1, 2, 4, 5, 6, 8, 10, 14 και 16. Μετά τους αντίστοιχους υπολογισμούς για το λόγο κέρδους για κάθε πιθανή διαχωριστική τιμή τα αποτελέσματα φαίνονται στον παρακάτω πίνακα (Πίνακας 6):

Διαχωριστική τιμή c	$\boldsymbol{GainRatio(S, waiting\_days(c))}$
1	**0,202**
2	0,061
4	0,113
5	0,015
6	0,002
8	0,080
10	0,057
14	0,036
16	0,017

Πίνακας 6: Αποτελέσματα GainRatio για όλες τις πιθανές τιμές διαχωρισμού για το χαρακτηριστικό waiting_days

Από τον παραπάνω πίνακα φαίνεται ότι η μέγιστη τιμή του λόγου κέρδους για το χαρακτηριστικό waiting_days, επιτυγχάνεται αν χρησιμοποιηθεί ως διαχωριστική τιμή το 1.

Τέλος, για το χαρακτηριστικό age, με τον ίδιο τρόπο, προκύπτουν τα παρακάτω αποτελέσματα (Πίνακας 7):

Διαχωριστική τιμή c	$\boldsymbol{GainRatio(S, age(c))}$
3	0,058
6	**0,106**
8	0

9	0,007
11	0,006
13	0,084
20	0,061
24	0,042
29	0,027
45	0,016
47	0,007
49	0,002
50	0
51	0,002
54	0,080
61	0,058
66	0,037
68	0,018

Πίνακας 7: Αποτελέσματα GainRatio για όλες τις πιθανές τιμές διαχωρισμού για το χαρακτηριστικό age

Από τον παραπάνω πίνακα φαίνεται ότι η μέγιστη τιμή του λόγου κέρδους για το χαρακτηριστικό age, επιτυγχάνεται αν χρησιμοποιηθεί ως διαχωριστική τιμή το 6.

Τελικά, προκύπτει ένας πίνακας που περιέχει τα GainRatio όλων των χαρακτηριστικών του συνόλου δεδομένων (Πίνακας 8).

Χαρακτηριστικό	**GainRatio**
sex	0,084
communication_channel	0,063
appointment_hour_d(c=15)	**0,291**
age(c=6)	0,106
waiting_days(c=1)	0,202

Πίνακας 8: GainRatio όλων των χαρακτηριστικών του συνόλου δεδομένων

Συγκρίνοντας τους λόγους κέρδους συμπεραίνεται ότι το ιδανικότερο χαρακτηριστικό για τον διαχωρισμό του συνόλου δεδομένων είναι το appointment_hour_d με βάση την τιμή 15, καθώς δίνει τον μέγιστο λόγο κέρδους. Έτσι βάσει αυτού του διαχωρισμού, θα προκύψει το πρώτο υποσύνολο S_1 με τα πρότυπα που έχουν appointment_hour_d <= 15, όπως φαίνεται παρακάτω (Πίνακας 9):

No.	**age**	**sex**	***appointment_ hour_d***	**waiting_ days**	**Commu- nication_ channel**	**no_ show (class)**
1	11.0	1	*14.0*	1.0	3	No
2	2.0	2	*15.0*	0.0	1	No
4	61.0	2	*14.0*	8.0	1	No

6	9.0	2	*15.0*	29.0	1	No
9	47.0	2	*9.0*	4.0	3	No
11	49.0	2	*12.0*	1.0	3	No
12	51.0	1	*14.0*	2.0	1	No
13	68.0	1	*8.0*	6.0	1	No
16	20.0	2	*9.0*	1.0	3	No
22	50.0	2	*11.0*	14.0	1	No

Πίνακας 9: Υποσύνολο S_1 (appointment_hour_d <= 15)

Εδώ φαίνεται ότι όλα τα πρότυπα ανήκουν στην κλάση no_show = No, οπότε δεν χρειάζεται να διαμεριστεί περαιτέρω το υποσύνολο S_1. Έτσι, προκύπτει ο κανόνας:

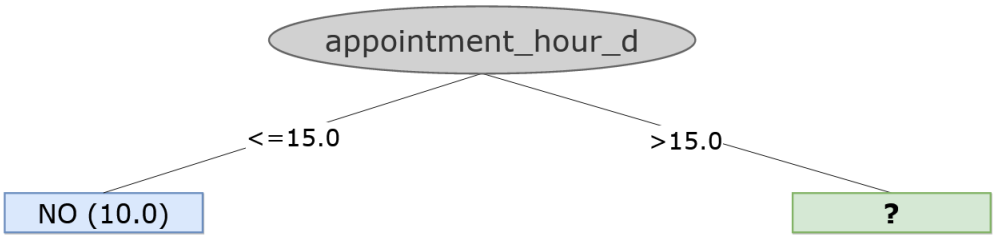
Αν appointment_hour_d <= 15, τότε no_show = No

Στον επόμενο πίνακα φαίνεται το δεύτερο υποσύνολο S_2 που είναι τα πρότυπα που έχουν appointment_hour_d > 15 (Πίνακας 10):

No.	**age**	**sex**	***appointment_hour_d***	**waiting_ days**	**Commu- nication_ channel**	**no_ show (class)**
3	11.0	2	*19.0*	8.0	3	Yes
5	6.0	2	*19.0*	16.0	3	No
7	66.0	1	*18.0*	0.0	2	No
8	13.0	1	*19.0*	5.0	3	Yes
10	9.0	2	*17.0*	0.0	1	No
14	29.0	1	*18.0*	2.0	2	No
15	70.0	1	*16.0*	2.0	1	No
17	3.0	2	*17.0*	10.0	1	No
18	8.0	1	*16.0*	2.0	1	Yes
19	54.0	1	*16.0*	6.0	1	Yes
20	3.0	1	*17.0*	8.0	3	No
21	24.0	2	*16.0*	1.0	1	No
23	45.0	2	*17.0*	4.0	2	No
24	6.0	1	*16.0*	0.0	3	No

Πίνακας 10: Υποσύνολο S_2 (appointment_hour_d > 15)

Εδώ παρατηρείται ότι υπάρχουν πρότυπα που ανήκουν και στις δύο κλάσεις, οπότε στη συνέχεια θα διαμεριστεί περαιτέρω το υποσύνολο S_2. Το δέντρο λοιπόν θα έχει ως πρώτο κόμβο το χαρακτηριστικό appointment_hour_d (Εικόνα 2).



Εικόνα 2: Ρίζα δέντρου appointment_hour_d (Ιδανικότερο χαρακτηριστικό για τον διαχωρισμό του συνόλου δεδομένων το χαρακτηριστικό appointment_hour_d)

Για το διαμερισμό του υποσυνόλου S_2, πρέπει να υπολογιστεί ο λόγος κέρδους για κάθε χαρακτηριστικό. Αρχικά υπολογίζονται τα ποσοστά των προτύπων των κλάσεων Yes και No. Έστω s το συνολικό πλήθος προτύπων και s_{No} και s_{Yes} το πλήθος προτύπων των κλάσεων No και Yes αντίστοιχα.

$$p_{No} = \frac{s_{No}}{s} = \frac{10}{14} = 0{,}7143$$

$$p_{Yes} = \frac{s_{Yes}}{s} = \frac{4}{14} = 0{,}2857$$

Η αρχική εντροπία για το υποσύνολο S_2 δίνεται από τον τύπο:

$$Entropy(S_2) = -\,p_{No}\,log_2(p_{No}) - p_{Yes}\,log_2(p_{Yes})$$

$$Entropy(S_2) = -\frac{10}{14}\,log_2\left(\frac{10}{14}\right) - \frac{4}{14}log_2\left(\frac{4}{14}\right) =$$

$$= 0{,}3467 + 0{,}5164 = 0{,}8631 \text{ bits}$$

Στη συνέχεια πρέπει να υπολογιστεί το κέρδος πληροφορίας για κάθε χαρακτηριστικό του υποσυνόλου S_2. Έτσι για το κατηγορικό χαρακτηριστικό sex θα έχουμε (Πίνακας 11):

	Σύνολο Προτύπων	**Πρότυπα κλάσης No**	**P(No)**	**Πρότυπα κλάσης Yes**	**P(Yes)**
Sex = 1	8	5	5/8	3	3/8
Sex = 2	6	5	5/6	1	1/6

Πίνακας 11: Κατηγορικό χαρακτηριστικό Sex του Υποσυνόλου S2

Κάνοντας τους υπολογισμούς που περιεγράφηκαν παραπάνω, βρίσκουμε ότι:

$$Entropy(S_2, sex) = 0{,}8240\ bits$$

Για να βρεθεί το κέρδος πληροφορίας για το συγκεκριμένο διαχωρισμό:

$$InfoGain(S, sex) = Entropy(S_2) - Entropy(S_2, sex) =$$

$$InfoGain(S_2, sex) = 0{,}8631 - 0{,}8240 = 0{,}0391\ bits$$

Τελικά, ο λόγος κέρδους (gain ratio) για την επιλογή του χαρακτηριστικού sex είναι:

$$GainRatio(S, sex) = \frac{InfoGain(S_2, sex)}{Entropy(S_2, sex)} = \frac{0{,}0391}{0{,}8240} = 0{,}0474$$

Για το κατηγορικό χαρακτηριστικό communication_channel προκύπτει ο παρακάτω πίνακας (Πίνακας 12):

	Σύνολο Προτύπων	**Πρότυπα κλάσης No**	**P(No)**	**Πρότυπα κλάσης Yes**	**P(Yes)**
communication_channel=1	6	4	4/6	2	2/6
communication_channel=2	3	3	3/3	0	0/3
communication_channel=3	5	3	3/5	2	2/5

Πίνακας 12: Κατηγορικό χαρακτηριστικό communication_channel του Υποσυνόλου S2

Κάνοντας τους υπολογισμούς που περιεγράφηκαν παραπάνω, προκύπτει ότι:

$$Entropy(S_2, communication\_channel) = 0{,}7403\ bits$$

Για να βρεθεί το κέρδος πληροφορίας για το συγκεκριμένο διαχωρισμό:

$$InfoGain(S_2, communication\_channel)$$
$$= Entropy(S_2) - Entropy(S_2, communication\_channel)$$
$$= 0{,}8631 - 0{,}7403 = 0{,}1228\ bits$$

Τελικά, ο λόγος κέρδους (gain ratio) για την επιλογή του χαρακτηριστικού communication_channel είναι:

$$GainRatio(S_2, communication\_channel) = \frac{InfoGain(S_2, communication\_channel)}{Entropy(S_2, communication\_channel)}$$
$$= \frac{0{,}1228}{0{,}7403} = 0{,}1659$$

Το συνεχές χαρακτηριστικό appointment_hour_d έχει ελάχιστη τιμή 16, μέγιστη τιμή 19 και ενδιάμεσες τιμές 17 και 18. Αν γίνουν οι παραπάνω υπολογισμοί για όλες τις πιθανές διαχωριστικές τιμές θα προκύψουν στα αποτελέσματα που φαίνονται στον παρακάτω πίνακα (Πίνακας 13):

Διαχωριστική τιμή c	$\boldsymbol{GainRatio(S_2, appointment\_hour\_d(c))}$
17	0,0299
18	**0,1755**

Πίνακας 13: Αποτελέσματα GainRatio για όλες τις πιθανές τιμές διαχωρισμού του Υποσυνόλου S2 για το συνεχές χαρακτηριστικό appointment_hour_d

Από τον πίνακα φαίνεται ότι η μέγιστη τιμή του λόγου κέρδους για το χαρακτηριστικό appointment_hour_d, επιτυγχάνεται αν χρησιμοποιηθεί ως διαχωριστική τιμή το 18.

Παρομοίως, το χαρακτηριστικό waiting_days έχει ελάχιστη τιμή 0, μέγιστη τιμή 16 και ενδιάμεσες τιμές 1, 2, 4, 5, 6, 8 και 10. Μετά τους αντίστοιχους υπολογισμούς για το λόγο κέρδους για κάθε πιθανή διαχωριστική τιμή προκύπτουν τα παρακάτω αποτελέσματα (Πίνακας 14):

Διαχωριστική τιμή c	$\boldsymbol{GainRatio(S_2, waiting\_days(c))}$
1	**0,2445**
2	0,0947
4	0,1677
5	0,0299
6	0,0021
8	0,0965
10	0,0438

Πίνακας 14: Αποτελέσματα GainRatio για όλες τις πιθανές τιμές διαχωρισμού του Υποσυνόλου S2 για το χαρακτηριστικό waiting_days

Από τον παραπάνω πίνακα φαίνεται ότι η μέγιστη τιμή του λόγου κέρδους για το χαρακτηριστικό waiting_days, επιτυγχάνεται αν χρησιμοποιηθεί ως διαχωριστική τιμή το 1.

Τέλος, για το χαρακτηριστικό age, με τον ίδιο τρόπο, θα προκύψουν τα παρακάτω αποτελέσματα (Πίνακας 15):

Διαχωριστική τιμή c ***age***	$\boldsymbol{GainRatio(S, age(c))}$
6	**0,2445**
8	0,0176
9	0,0475
11	0
13	0,0475
24	0,0176
29	0,0021
45	0,0025
54	0,0966
66	0,0438

Πίνακας 15: Αποτελέσματα GainRatio για όλες τις πιθανές τιμές διαχωρισμού του Υποσυνόλου S2 για το χαρακτηριστικό age

Από τον πίνακα φαίνεται ότι η μέγιστη τιμή του λόγου κέρδους για το χαρακτηριστικό age, επιτυγχάνεται αν χρησιμοποιηθεί ως διαχωριστική τιμή το 6.

Τελικά, προκύπτει ένας πίνακα που περιέχει τα GainRatio όλων των χαρακτηριστικών του υποσυνόλου S_2 (Πίνακας 16).

Χαρακτηριστικό	**GainRatio**
sex	0,0474
communication_channel	0,1659
appointment_hour_d(c=18)	0,1755
age(c=6)	**0,2445**
waiting_days(c=1)	**0,2445**

Πίνακας 16: GainRatio όλων των χαρακτηριστικών του Υποσυνόλου S_2

Στην συγκεκριμένη περίπτωση συγκρίνοντας τους λόγους κέρδους παρατηρείται ότι δύο χαρακτηριστικά παρουσιάζουν τον μέγιστο λόγο κέρδους, οπότε μπορεί να επιλεγεί οποιοδήποτε από τα δύο για τον δεύτερο διαμερισμό. Με τυχαίο τρόπο επιλέγεται το χαρακτηριστικό age με τιμή διαχωρισμού το 6, χωρίζοντας το υποσύνολο S_2 στα υποσύνολα S_3 και S_4. Έτσι λοιπόν, θα προκύψει το πρώτο υποσύνολο S_3 στο οποίο ανήκουν τα πρότυπα που έχουν age <= 6, όπως φαίνεται παρακάτω (Πίνακας 17):

No.	**age**	**sex**	***appointment_ hour_d***	**waiting_ days**	**Commu- nication_ channel**	**no_ show (class)**
5	6.0	2	*19.0*	16.0	3	No
17	3.0	2	*17.0*	10.0	1	No
20	3.0	1	*17.0*	8.0	3	No
24	6.0	1	*16.0*	0.0	3	No

Πίνακας 17: Υποσύνολο S_3 (age <= 6)

Εδώ φαίνεται ότι όλα τα πρότυπα ανήκουν στην κλάση no_show = No, οπότε δεν χρειάζεται να διαμεριστεί περαιτέρω το υποσύνολο S_3. Έτσι, προκύπτει ο κανόνας:

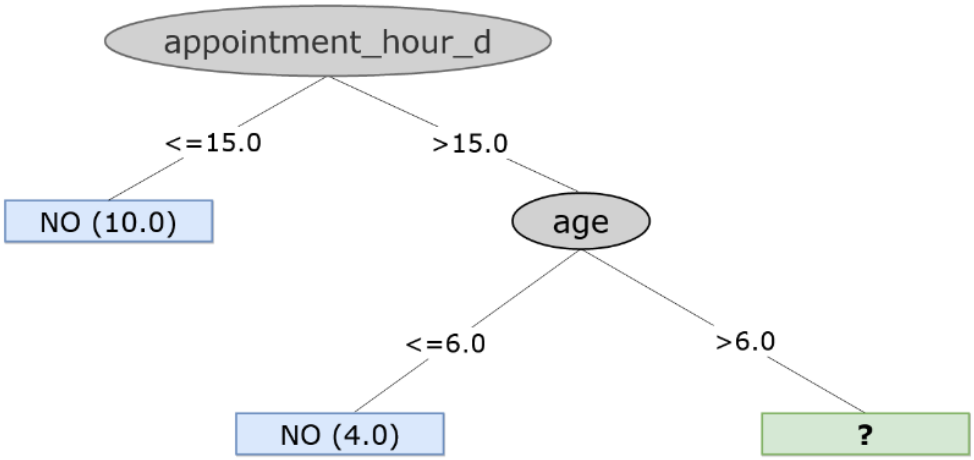
Αν appointment_hour_d > 15 και age <= 6, τότε no_show = No

Στον επόμενο πίνακα φαίνεται το δεύτερο υποσύνολο S_4 που είναι τα πρότυπα που έχουν age > 6 (Πίνακας 18):

No.	**age**	**sex**	***appointment_ hour_d***	**waiting_ days**	**Commu- nication_ channel**	**no_ show (class)**
3	11.0	2	*19.0*	8.0	3	Yes
7	66.0	1	*18.0*	0.0	2	No
8	13.0	1	*19.0*	5.0	3	Yes
10	9.0	2	*17.0*	0.0	1	No
14	29.0	1	*18.0*	2.0	2	No
15	70.0	1	*16.0*	2.0	1	No
18	8.0	1	*16.0*	2.0	1	Yes
19	54.0	1	*16.0*	6.0	1	Yes
21	24.0	2	*16.0*	1.0	1	No
23	45.0	2	*17.0*	4.0	2	No

Πίνακας 18: Υποσύνολο S_4 (age > 6)

Εδώ παρατηρείται ότι υπάρχουν πρότυπα που ανήκουν και στις δύο κλάσεις, οπότε στη συνέχεια θα διαμεριστεί περαιτέρω το υποσύνολο S_4. Ο δεύτερος κόμβος του δέντρου θα είναι το χαρακτηριστικό age (Εικόνα 3).



Εικόνα 3: Δεύτερος κόμβος δέντρου - χαρακτηριστικό age

Για το διαμερισμό του υποσυνόλου S_4 πρέπει να υπολογιστεί ο λόγος κέρδους για κάθε χαρακτηριστικό. Αρχικά υπολογίζονται τα ποσοστά των προτύπων των κλάσεων Yes και No. Έστω s το συνολικό πλήθος προτύπων και s_{No} και s_{Yes} το πλήθος προτύπων των κλάσεων No και Yes αντίστοιχα.

$$p_{No} = \frac{s_{No}}{s} = \frac{6}{10} = 0{,}6$$

$$p_{Yes} = \frac{s_{Yes}}{s} = \frac{4}{10} = 0{,}4$$

Η αρχική εντροπία για το σύνολο S_4 δίνεται από τον τύπο:

$$Entropy(S_4) = -\,p_{No}\,log_2(p_{No}) - p_{Yes}\,log_2(p_{Yes})$$

$$Entropy(S_4) = -\frac{6}{10}\,log_2\left(\frac{6}{10}\right) - \frac{4}{10}log_2\left(\frac{4}{10}\right) =$$

$$= 0{,}4422 + 0{,}5288 = 0{,}971 \text{ bits}$$

Στη συνέχεια πρέπει να υπολογιστεί το κέρδος πληροφορίας για κάθε χαρακτηριστικό του υποσυνόλου S_4. Έτσι για το κατηγορικό χαρακτηριστικό sex θα προκύψει ο παρακάτω πίνακας (Πίνακας 19):

	Σύνολο Προτύπων	**Πρότυπα κλάσης No**	**P(No)**	**Πρότυπα κλάσης Yes**	**P(Yes)**
Sex = 1	6	3	3/6	3	3/6
Sex = 2	4	3	3/4	1	1/4

Πίνακας 19: Κατηγορικό χαρακτηριστικό Sex (Υποσύνολο S4)

Κάνοντας τους υπολογισμούς που περιεγράφηκαν παραπάνω, προκύπτει ότι:

$$Entropy(S_4, sex) = 0{,}9245\ bits$$

Για να βρεθεί το κέρδος πληροφορίας για το συγκεκριμένο διαχωρισμό:

$$InfoGain(S_4, sex) = Entropy(S_4) - Entropy(S_4, sex) = 0{,}971 - 0{,}9245 = 0{,}0464\ bits$$

Τελικά, ο λόγος κέρδους (gain ratio) για την επιλογή του χαρακτηριστικού sex είναι:

$$GainRatio(S_4, sex) = \frac{InfoGain(S_4, sex)}{Entropy(S_4, sex)} = \frac{0{,}0464}{0{,}9245} = 0{,}0502$$

Για το κατηγορικό χαρακτηριστικό communication_channel θα προκύψουν τα παρακάτω αποτελέσματα (Πίνακας 20):

	Σύνολο Προτύπων	**Πρότυπα κλάσης No**	**P(No)**	**Πρότυπα κλάσης Yes**	**P(Yes)**
communication_ channel=1	5	3	3/5	2	2/5
communication_ channel=2	3	3	3/3	0	0/3
communication_ channel=3	2	0	0/2	2	2/2

Πίνακας 20: Κατηγορικό χαρακτηριστικό communication_channel (Υποσύνολο S4)

Κάνοντας τους υπολογισμούς που περιεγράφηκαν παραπάνω, προκύπτει ότι:

$$Entropy(S_4, communication\_channel) = 0{,}4855\ bits$$

Για να βρεθεί το κέρδος πληροφορίας για το συγκεκριμένο διαχωρισμό:

$$InfoGain(S_4, communication\_channel)$$
$$= Entropy(S_4) - Entropy(S_4, communication\_channel)$$
$$= 0{,}971 - 0{,}4855 = 0{,}4855\ bits$$

Τελικά, ο λόγος κέρδους (gain ratio) για την επιλογή του χαρακτηριστικού communication_channel είναι:

$$GainRatio(S_4, communication\_channel) = \frac{InfoGain(S_4, communication\_channel)}{Entropy(S_4, communication\_channel)}$$
$$= \frac{0{,}4855}{0{,}4855} = 1$$

Το συνεχές χαρακτηριστικό appointment_hour_d έχει ελάχιστη τιμή 16, μέγιστη τιμή 19 και ενδιάμεσες τιμές 17 και 18. Αν γίνουν οι παραπάνω υπολογισμοί για όλες τις πιθανές διαχωριστικές τιμές θα προκύψουν τα αποτελέσματα που φαίνονται στον παρακάτω πίνακα (Πίνακας 21):

Διαχωριστική τιμή c	$\boldsymbol{GainRatio(S_2, appointment\_hour\_d(c))}$
17	0,0210
18	**0,4960**

Πίνακας 21: Αποτελέσματα GainRatio για όλες τις πιθανές τιμές διαχωρισμού του Υποσυνόλου S4 για το συνεχές χαρακτηριστικό appointment_hour_d

Από τον πίνακα φαίνεται ότι η μέγιστη τιμή του λόγου κέρδους για το χαρακτηριστικό appointment_hour_d, επιτυγχάνεται αν χρησιμοποιηθεί ως διαχωριστική τιμή το 18.

Παρομοίως, το χαρακτηριστικό waiting_days έχει ελάχιστη τιμή 0, μέγιστη τιμή 8 και ενδιάμεσες τιμές 1, 2, 4, 5 και 6. Μετά τους αντίστοιχους υπολογισμούς για το λόγο κέρδους για κάθε πιθανή διαχωριστική τιμή προκύπτει ο παρακάτω πίνακας (Πίνακας 22):

Διαχωριστική τιμή c	$\boldsymbol{GainRatio(S_4, waiting\_days(c))}$
1	0,4079
2	0,3589
4	**1,3443**
5	0,4960
6	0,1748

Πίνακας 22: Αποτελέσματα GainRatio για όλες τις πιθανές τιμές διαχωρισμού του Υποσυνόλου S4 για το κατηγορικό χαρακτηριστικό waiting_days

Από τον πίνακα φαίνεται ότι η μέγιστη τιμή του λόγου κέρδους για το χαρακτηριστικό waiting_days, επιτυγχάνεται αν χρησιμοποιηθεί ως διαχωριστική τιμή το 4.

Τέλος, για το χαρακτηριστικό age, με τον ίδιο τρόπο, θα προκύψουν τα παρακάτω αποτελέσματα (Πίνακας 23):

Διαχωριστική τιμή c	$\boldsymbol{GainRatio(S_4, age(c))}$
9	0,0077
11	0,1038
13	**0,3589**
24	0,1471
29	0,0502
45	0,0060
54	0,2137
66	0,0885

Πίνακας 23: Αποτελέσματα GainRatio για όλες τις πιθανές τιμές διαχωρισμού του Υποσυνόλου S4 για το χαρακτηριστικό age

Από τον πίνακα (Πίνακας 23) φαίνεται ότι η μέγιστη τιμή του λόγου κέρδους για το χαρακτηριστικό age, επιτυγχάνεται αν χρησιμοποιηθεί ως διαχωριστική τιμή το 13.

Τελικά, προκύπτει ένας πίνακας που περιέχει τα GainRatio όλων των χαρακτηριστικών του υποσυνόλου S_4 (Πίνακας 24).

Χαρακτηριστικό	GainRatio
sex	0,0502
communication_channel	1
appointment_hour_d(c=18)	0,4960
age(c=13)	0,3589
waiting_days(c=4)	**1,3443**

Πίνακας 24: GainRatio όλων των χαρακτηριστικών του Υποσυνόλου S_4

Συγκρίνοντας τους λόγους κέρδους εξάγεται το συμπέρασμα ότι το ιδανικότερο χαρακτηριστικό για τον διαχωρισμό του υποσυνόλου S_4 είναι το waiting_days με βάση την τιμή 4, χωρίζοντας το υποσύνολο S_4 στα υποσύνολα S_5 και S_6. Έτσι λοιπόν, θα προκύψει το πρώτο υποσύνολο S_5 που είναι τα πρότυπα που έχουν waiting_days <= 4, όπως φαίνεται παρακάτω (Πίνακας 25):

No.	age	sex	*appointment_ hour_d*	waiting_ days	Commu- nication_ channel	no_ show (class)
7	66.0	1	*18.0*	0.0	2	No
10	9.0	2	*17.0*	0.0	1	No
14	29.0	1	*18.0*	2.0	2	No
15	70.0	1	*16.0*	2.0	1	No
18	8.0	1	*16.0*	2.0	1	Yes
21	24.0	2	*16.0*	1.0	1	No
23	45.0	2	*17.0*	4.0	2	No

Πίνακας 25: Υποσύνολο S_5 (waiting_days <= 4)

Παρατηρείται ότι στο υποσύνολο S_5 τα περισσότερα πρότυπα ανήκουν στη κλάση No. Εδώ μπορεί να συνεχιστεί ο διαμερισμός, αλλά επειδή υπάρχει μόνο ένα πρότυπο που ανήκει στην κλάση Yes, μπορεί να σταματήσει η διαδικασία του διαμερισμού ώστε να περιοριστεί τυχόν φαινόμενο υπερπροσαρμογής.

Έτσι, προκύπτει ο κανόνας:

Αν appointment_hour_d > 15 και age > 6 και waiting_days <= 4, τότε no_show = No

Ο κανόνας αυτός θα κατηγοριοποιήσει λανθασμένα ως No το πρότυπο υπ' αριθμόν 18.

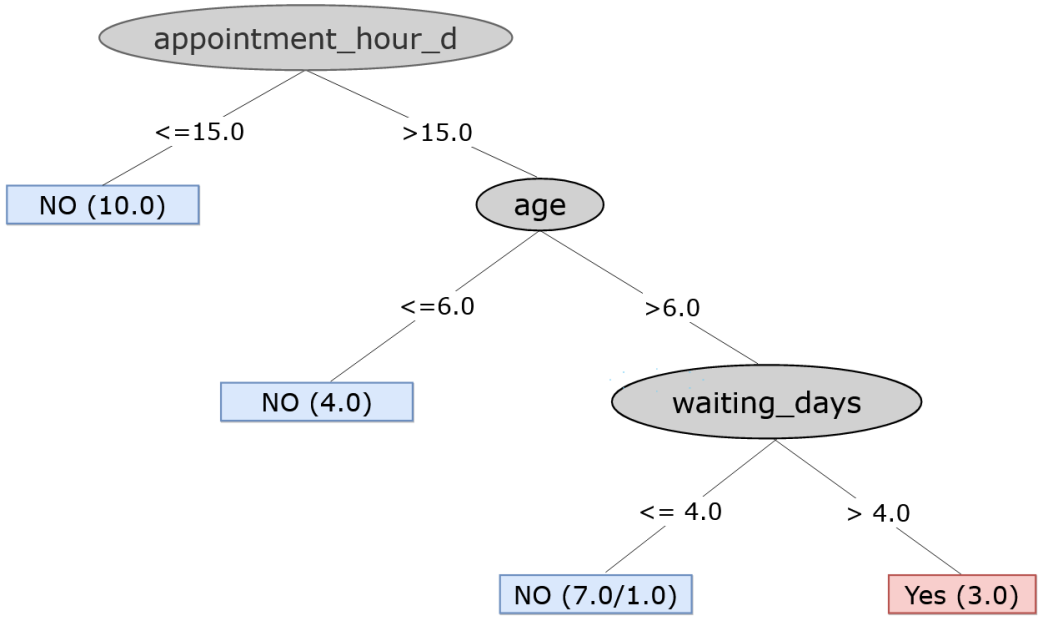
No.	age	sex	*appointment_ hour_d*	waiting_ days	Commu- nication_ channel	no_ show (class)
3	11.0	2	*19.0*	8.0	3	Yes
8	13.0	1	*19.0*	5.0	3	Yes
19	54.0	1	*16.0*	6.0	1	Yes

Πίνακας 26: Υποσύνολο S_6 (waiting_days > 4)

Στο δεύτερο υποσύνολο S_6 παρατηρείται ότι όλα τα πρότυπα ανήκουν στην κλάση Yes (Πίνακας 26). Έτσι, προκύπτει ο κανόνας:

Αν appointment_hour_d > 15 και age > 6 και waiting_days > 4, τότε no_show = Yes

Τελικά ο τρίτος κόμβος του δέντρου θα είναι το χαρακτηριστικό waiting_days και το ολοκληρωμένο δέντρο θα έχει τη μορφή που φαίνεται παρακάτω (Εικόνα 4).



Εικόνα 4: Ολοκληρωμένο δέντρο με εφαρμογή του αλγορίθμου ID3

2.4.2 Naïve Bayes

Ο Naive Bayes είναι ένας πιθανοτικός κατηγοριοποιητής που ανήκει στην οικογένεια των μπεϋζιανών κατηγοριοποιητών. Οι μπεϋζιανοί κατηγοριοποιητές βασίζονται στο θεώρημα του Bayes και για την κατηγοριοποίηση προτύπων σε ένα αριθμό από κλάσεις, υπολογίζουν την πιθανότητα να ανήκει το πρότυπο σε κάθε μία από τις κλάσεις.

Ο Naive Bayes είναι ο πιο απλός κατηγοριοποιητής της οικογένειας. Η απλότητα (naive) οφείλεται στην παραδοχή πώς η τιμή κάθε χαρακτηριστικού κάποιου προτύπου που ανήκει σε μια συγκεκριμένη κλάση, είναι ανεξάρτητη από τις τιμές των υπόλοιπων χαρακτηριστικών του ιδίου προτύπου. Η συγκεκριμένη παραδοχή ονομάζεται στατιστική ανεξαρτησία ως προς την κλάση (class conditional independence). Συνήθως δεν ισχύει για δεδομένα που προέρχονται από τον πραγματικό κόσμο, όμως απλοποιεί σε μεγάλο βαθμό τους υπολογισμούς που απαιτούνται για την κατηγοριοποίηση (N. Friedman et al., 1997).

Παρόλο το γεγονός ότι ο αλγόριθμος βασίζεται σε μια μη-ρεαλιστική παραδοχή, διάφορες έρευνες έδειξαν ότι σε πολλές περιπτώσεις ο κατηγοριοποιητής Naive Bayes έχει αντίστοιχη

απόδοση με κατηγοριοποιητές που βασίζονται σε δέντρα απόφασης και νευρωνικά δίκτυα, έχοντας όμως το πλεονέκτημα της ταχύτητας εκπαίδευσης, ειδικά όταν εφαρμόζεται σε μεγάλα σύνολα δεδομένων (N. Friedman et al., 1997), (Domingos & Pazzani, 1997).

Ένα άλλο σημαντικό πλεονέκτημα του Naive Bayes είναι ότι μπορεί να χειριστεί μεγάλο αριθμό χαρακτηριστικών, με κατηγορικές αλλά και συνεχείς τιμές, χωρίς μεγάλη επιβάρυνση υπολογιστικού χρόνου. Επιπλέον μπορεί να επιτύχει υψηλή απόδοση, ειδικά στην περίπτωση που τα χαρακτηριστικά είναι όντως στατιστικά ανεξάρτητα μεταξύ τους. Επιπρόσθετα, ο Naive Bayes είναι online αλγόριθμος, δηλαδή μπορεί να εκπαιδευτεί με επιπλέον δεδομένα και να βελτιωθεί, χωρίς να χρειάζεται να επαναληφθεί η διαδικασία εκπαίδευσης από την αρχή.

Λόγω των παραπάνω πλεονεκτημάτων, ο Naive Bayes είναι συνήθως ο πρώτος αλγόριθμος κατηγοριοποίησης που επιλέγεται για να δοκιμαστεί σε νέα προβλήματα μηχανικής μάθησης. Επιπλέον, τα αποτελέσματα του μπορούν να ερμηνευτούν πιο εύκολα σε σχέση με τα αποτελέσματα πιο πολύπλοκων αλγορίθμων όπως π.χ. τα Νευρωνικά Δίκτυα και οι Μηχανές Διανυσμάτων Υποστήριξης (SVM) (Buczak & Guven, 2016). Για την εφαρμογή του θεωρήματος του Bayes (Han & Kamber, 2006) θεωρείται ότι είναι γνωστό ένα δεδομένο X που αφορά κάποιο γεγονός και μια υπόθεση H σχετικά με κάποιο άλλο γεγονός. Με τη βοήθεια του θεωρήματος Bayes μπορεί να υπολογιστεί η πιθανότητα να ισχύει η υπόθεση H δεδομένου ότι είναι γνωστό το X. Η πιθανότητα αυτή ονομάζεται εκ των υστέρων πιθανότητα (posterior probability) και συμβολίζεται ως P(H|X). Η εξίσωση του θεωρήματος Bayes που αλλιώς ονομάζεται και κανόνας Bayes δίνεται στη συνέχεια:

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)}$$

όπου:

$P(H)$: η εκ των προτέρων (prior) πιθανότητα να ισχύει η υπόθεση H

$P(X)$: η εκ των προτέρων (prior) πιθανότητα παρατήρησης του X

$P(X|H)$: η δεσμευμένη (conditional) πιθανότητα που έχει το ενδεχόμενο παρατήρησης του X, αν είναι γνωστό ότι ισχύει η υπόθεση H. Ονομάζεται επίσης και πιθανοφάνεια (likelihood)

$P(H|X)$: η ζητούμενη εκ των υστέρων (posterior) πιθανότητα που έχει το ενδεχόμενο να ισχύει η υπόθεση H δεδομένου ότι γνωρίζουμε το X

Οι παραπάνω πιθανότητες μπορούν να υπολογιστούν από τα διαθέσιμα δεδομένα. Σε περίπτωση που υπάρχουν περισσότερες από μία υποθέσεις, αρχικά γίνεται ο υπολογισμός της εκ των υστέρων πιθανότητας για κάθε υπόθεση ξεχωριστά. Στη συνέχεια επιλέγεται η υπόθεση με την μέγιστη εκ των υστέρων πιθανότητα (maximum a-posteriori hypothesis ή MAP hypothesis), ως αυτή που είναι περισσότερο πιθανό να ισχύει.

Στα προβλήματα κατηγοριοποίησης συνήθως το X είναι το διάνυσμα χαρακτηριστικών ενός προτύπου, δηλαδή ένα σύνολο τιμών από n χαρακτηριστικά. Το H είναι η υπόθεση ότι το πρότυπο X ανήκει σε κάποια συγκεκριμένη κλάση C.

Για να περιγραφεί ο τρόπος λειτουργίας του κατηγοριοποιητή Naive Bayes, ας υποτεθεί ότι είναι διαθέσιμο ένα σύνολο δεδομένων D που περιέχει πρότυπα με την αντίστοιχη κλάση στην οποία ανήκει το καθένα. Κάθε ένα από τα πρότυπα του συνόλου D μπορεί να αναπαρασταθεί από ένα διάνυσμα n-διαστάσεων $X = (x_1, x_2, \dots, x_n)$ που οι τιμές του αντιστοιχούν στις μετρήσεις n χαρακτηριστικών $A_1, A_2, \dots, A_n$. Επίσης έστω ότι υπάρχουν m κλάσεις $C_1, C_2, \dots, C_m$. Το σύνολο δεδομένων D μπορεί να θεωρηθεί σύνολο δεδομένων εκπαίδευσης, γιατί ο κατηγοριοποιητής Naive Bayes θα βασιστεί σε αυτό για να κάνει στη συνέχεια προβλέψεις για άγνωστα πρότυπα.

Για να προβλέψει λοιπόν ο κατηγοριοποιητής, σε ποια από τις m κλάσεις θα ανήκει ένα άγνωστο πρότυπο X, αρκεί να βρει την κλάση με την μέγιστη εκ των υστέρων πιθανότητα δεδομένου του X. Δηλαδή να υπολογίσει την εκ των υστέρων πιθανότητα να ανήκει το X σε κάθε μία από τις m κλάσεις και στη συνέχεια να επιλέξει αυτήν με την μέγιστη τιμή:

$$\underset{i=1,2,..,m}{\operatorname{argmax}} \{ P(C_i|X) \}$$

Όπως προαναφέρθηκε, η εκ των υστέρων πιθανότητα $P(C_i|X)$ μπορεί να υπολογιστεί με τη βοήθεια του κανόνα του Bayes:

$$P(C_i|X) = \frac{P(X|C_i)\, P(C_i)}{P(X)}$$

Επιπλέον, η εκ των προτέρων πιθανότητα παρατήρησης του X, δηλαδή η P(X), είναι ανεξάρτητη από την κλάση στην οποία ανήκει το πρότυπο, άρα δεν παίζει ρόλο στην εύρεση της μέγιστης εκ των υστέρων πιθανότητας. Επομένως μπορεί να παραλειφθεί για απλοποίηση των υπολογισμών. Συνδυάζοντας λοιπόν τις δύο παραπάνω σχέσεις και παραλείποντας το P(X) θα ισχύει:

$$\underset{i=1,2,..,m}{\operatorname{argmax}} \{ P(C_i|X) \} = \underset{i=1,2,..,m}{\operatorname{argmax}} \left\{ \frac{P(X|C_i)\, P(C_i)}{P(X)} \right\} = \underset{i=1,2,..,m}{\operatorname{argmax}} \{ P(X|C_i)\, P(C_i) \}$$

Ο όρος $P(C_i)$ μπορεί να υπολογιστεί με βάση τις συχνότητες εμφάνισης της κλάσης C_i στα πρότυπα του συνόλου δεδομένων D, ως εξής:

$$P(C_i) = \frac{|C_{i,D}|}{|D|}, i = 1,2, \dots, m$$

όπου:

$|C_{i,D}|$ είναι το πλήθος των προτύπων του συνόλου δεδομένων D, που ανήκουν στην κλάση C_i.

$|D|$ είναι το συνολικό πλήθος των προτύπων του συνόλου δεδομένων D.

Ο υπολογισμός του όρου $P(X|C_i)$, δηλαδή της δεσμευμένης πιθανότητας του διανύσματος X αν είναι γνωστό ότι το X ανήκει στην κλάση C_i, υπό κανονικές συνθήκες θα ήταν αρκετά πολύπλοκος γιατί θα έπρεπε να ληφθούν υπ' όψιν οι τυχόν υπάρχουσες εξαρτήσεις μεταξύ των χαρακτηριστικών των προτύπων. Όμως στην προκειμένη περίπτωση ο κατηγοριοποιητής Naive Bayes κάνει την απλοϊκή (naive) παραδοχή ότι οι τιμές των χαρακτηριστικών είναι στατιστικά ανεξάρτητες μεταξύ τους, δεδομένης της κλάσης στην οποία ανήκει το πρότυπο X. Όπως αναφέρθηκε και πιο πάνω, η παραδοχή αυτή δεν ισχύει συνήθως, αλλά δίνει μια καλή προσέγγιση και το σημαντικότερο, απλοποιεί πάρα πολύ τον τρόπο υπολογισμού. Συγκεκριμένα, βάσει της παραδοχής, η ζητούμενη δεσμευμένη πιθανότητα μπορεί να υπολογιστεί ως ένα γινόμενο δεσμευμένων πιθανοτήτων:

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \cdot P(x_2|C_i) \cdot \dots \cdot P(x_n|C_i)$$

όπου n είναι το πλήθος των χαρακτηριστικών του διανύσματος X.

Οι δεσμευμένες πιθανότητες $P(x_1|C_i), P(x_2|C_i), \dots, P(x_n|C_i)$ μπορούν να υπολογιστούν εύκολα από το σύνολο των προτύπων του συνόλου δεδομένων D, λαμβάνοντας υπόψη το είδος του κάθε χαρακτηριστικού. Δηλαδή αν το χαρακτηριστικό είναι κατηγορικό τότε παίρνει διακριτές τιμές διαφορετικά παίρνει συνεχείς τιμές.

Αν το χαρακτηριστικό A_k είναι κατηγορικό, τότε το $P(x_k|C_i)$ θα είναι ο λόγος του πλήθους των προτύπων του D που ανήκουν στην κλάση C_i και έχουν την τιμή x_k στο χαρακτηριστικό A_k, προς το συνολικό πλήθος των προτύπων του D που ανήκουν στην κλάση C_i. Δηλαδή:

$$P(x_k|C_i) = \frac{|C_{i,D,A_k=x_k}|}{|C_{i,D}|}$$

Αν το χαρακτηριστικό A_k παίρνει συνεχείς τιμές (John & Langley, 1995), τότε γίνεται συνήθως η υπόθεση ότι οι τιμές αυτές ακολουθούν την κανονική ή αλλιώς Γκαουσιανή (Gaussian) κατανομή, με μέση τιμή μ και τυπική απόκλιση σ, η οποία δίνεται από τον τύπο:

$$g(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Έτσι, τελικά η δεσμευμένη πιθανότητα $P(x_k|C_i)$ του συνεχούς χαρακτηριστικού A_k θα είναι:

$$P(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i})$$

όπου μ_{C_i} και σ_{C_i} είναι ο μέσος όρος και η τυπική απόκλιση, αντίστοιχα, όλων των τιμών του χαρακτηριστικού A_k όλων των προτύπων που ανήκουν στην κλάση C_i μέσα στο σύνολο δεδομένων D. Η τυπική απόκλιση n τιμών με μέσο όρο μ, δίνεται από τον τύπο:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2}$$

Οι τιμές των μ_{C_i} και σ_{C_i} για κάθε χαρακτηριστικό και κάθε κλάση, μπορούν να υπολογιστούν εύκολα από τα δεδομένα του συνόλου D.

Έτσι τελικά για κάθε κλάση C_i με βάση τα δεδομένα του συνόλου D, μπορεί να υπολογιστεί εύκολα το γινόμενο $P(X|C_i)\,P(C_i)$. Στη συνέχεια βάσει αυτών των αποτελεσμάτων, μπορεί να βρεθεί η κλάση C_i, η οποία ικανοποιεί την υπόθεση με την μέγιστη εκ των υστέρων πιθανότητα. Η συγκεκριμένη κλάση θα αποτελεί και την πρόβλεψη της κλάσης στην οποία θα ανήκει το πρότυπο X.

Παράδειγμα κατηγοριοποίησης με Naive Bayes

Για καλύτερη κατανόηση του αλγορίθμου κατηγοριοποίησης Naive Bayes, δίνεται ένα αναλυτικό παράδειγμα. Τα δεδομένα που χρησιμοποιήθηκαν είναι ένα δείγμα 16 εγγραφών από το σύνολο δεδομένων Medical Appointments (Alvaro Flores). Τα δεδομένα αυτά αποτελούν το σύνολο δεδομένων εκπαίδευσης, όπως φαίνεται στον παρακάτω πίνακα (Πίνακας 27).

No.	age	sex	appointment_hour_d	waiting_days	communication_channel	no_show (class)
1	11	1	14	1	3	No
2	2	2	15	0	1	No
3	11	2	19	8	3	Yes
4	61	2	14	8	1	No
5	6	2	19	16	3	No
6	9	2	15	29	1	No
7	66	1	18	0	2	No
8	13	1	19	5	3	Yes
9	20	2	9	1	3	No
10	3	2	17	10	1	No
11	8	1	16	2	1	Yes
12	54	1	16	6	1	Yes
13	3	1	17	8	3	No
14	45	2	17	4	2	No
15	6	1	16	0	3	No
16	32	2	12	7	2	Yes

Πίνακας 27: Σύνολο δεδομένων εκπαίδευσης

Στα παραπάνω δεδομένα το χαρακτηριστικό της κλάσης είναι το no_show που μπορεί να πάρει τιμές Yes και No. Αρχικά υπολογίζεται η εκ των προτέρων πιθανότητα για κάθε κλάση, η οποία ουσιαστικά είναι τα ποσοστά των προτύπων των κλάσεων Yes και No ως προς το σύνολο των προτύπων του συνόλου δεδομένων εκπαίδευσης. Έστω s το συνολικό πλήθος προτύπων και s_{No} και s_{Yes} το πλήθος προτύπων των κλάσεων No και Yes αντίστοιχα.

$$P(\text{No}) = \frac{s_{\text{No}}}{s} = \frac{11}{16} = 0{,}6875$$

$$P(\text{Yes}) = \frac{s_{\text{Yes}}}{s} = \frac{5}{16} = 0{,}3125$$

Στη συνέχεια πρέπει να υπολογιστούν οι δεσμευμένες πιθανότητες P(X|No) και P(X|Yes) για κάθε ένα από τα χαρακτηριστικά του συνόλου εκπαίδευσης. Τα χαρακτηριστικά sex και communication_channel είναι κατηγορικά χαρακτηριστικά, ενώ τα age, appointment_hour_d και waiting_days είναι χαρακτηριστικά τα οποία παίρνουν συνεχείς τιμές.

Ξεκινώντας από τα κατηγορικά χαρακτηριστικά, για το sex που παίρνει τιμές 1 και 2, παρατηρείται ότι από τα 11 πρότυπα που ανήκουν στην κλάση No, τα 4 έχουν τιμή sex = 1 και τα υπόλοιπα 7 έχουν τιμή sex = 2. Έτσι προκύπτει ότι:

$$P(\text{sex} = 1|\text{No}) = \frac{s_{\text{sex}=1,\text{No}}}{S_{\text{No}}} = \frac{4}{11} = 0{,}3636$$

$$P(\text{sex} = 2|\text{No}) = \frac{s_{\text{sex}=2,\text{No}}}{S_{\text{No}}} = \frac{7}{11} = 0{,}6364$$

Επίσης, από τα 5 πρότυπα που ανήκουν στην κλάση Yes, τα 3 έχουν τιμή sex = 1 και τα υπόλοιπα 2 έχουν τιμή sex = 2. Έτσι:

$$P(\text{sex} = 1|\text{Yes}) = \frac{s_{\text{sex}=1,\text{Yes}}}{S_{\text{Yes}}} = \frac{3}{5} = 0{,}6$$

$$P(\text{sex} = 2|\text{Yes}) = \frac{s_{\text{sex}=2,\text{Yes}}}{S_{\text{Yes}}} = \frac{2}{5} = 0{,}4$$

Αντίστοιχα, για το χαρακτηριστικό communication_channel, που παίρνει τιμές 1, 2 ή 3, παρατηρείται ότι από τα 11 πρότυπα που ανήκουν στην κλάση No, τα 4 έχουν τιμή communication_channel = 1, τα 2 έχουν τιμή communication_channel = 2 και τα υπόλοιπα 5 έχουν τιμή communication_channel = 3. Με βάση τα παραπάνω δεδομένα υπολογίζονται οι παρακάτω δεσμευμένες πιθανότητες:

$$P(\text{communication\_channel} = 1|\text{No}) = \frac{s_{\text{communication\_channel}=1,\text{No}}}{S_{\text{No}}} = \frac{4}{11} = 0{,}3636$$

$$P(\text{communication\_channel} = 2|\text{No}) = \frac{s_{\text{communication\_channel}=2,\text{No}}}{S_{\text{No}}} = \frac{2}{11} = 0{,}1818$$

$$P(\text{communication\_channel} = 3|\text{No}) = \frac{s_{\text{communication\_channel}=3,\text{No}}}{S_{\text{No}}} = \frac{5}{11} = 0{,}4545$$

Από τα 5 πρότυπα της κλάσης Yes, παρατηρείται ότι τα 2 πρότυπα έχουν τιμή communication_channel = 1, το 1 πρότυπο έχει τιμή communication_channel = 2 και τα υπόλοιπα 2 πρότυπα έχουν τιμή communication_channel = 3. Έτσι προκύπτει ότι:

$$P(\text{communication\_channel} = 1|\text{Yes}) = \frac{S_{\text{communication\_channel=1,Yes}}}{S_{\text{Yes}}} = \frac{2}{5} = 0{,}4$$

$$P(\text{communication\_channel} = 2|\text{Yes}) = \frac{S_{\text{communication\_channel=2,Yes}}}{S_{\text{Yes}}} = \frac{1}{5} = 0{,}2$$

$$P(\text{communication\_channel} = 3|\text{Yes}) = \frac{S_{\text{communication\_channel=3,Yes}}}{S_{\text{Yes}}} = \frac{2}{5} = 0{,}4$$

Για τα χαρακτηριστικά που παίρνουν συνεχείς τιμές πρέπει πρώτα να υπολογιστεί ο μέσος όρος μ και η τυπική απόκλιση σ των τιμών τους ανάλογα με την κλάση στην οποία ανήκουν τα πρότυπα.

Για τα πρότυπα που ανήκουν στην κλάση No, οι τιμές του χαρακτηριστικού age έχουν άθροισμα $\Sigma_{\text{age,No}} = 232$ οπότε ο μέσος όρος $\mu_{\text{age,No}}$ θα είναι:

$$\mu_{\text{age,No}} = \frac{\Sigma_{\text{age,No}}}{S_{\text{No}}} = \frac{232}{11} = 21{,}0909$$

Η τυπική απόκλιση των τιμών του χαρακτηριστικού age, όπου x_i η κάθε τιμή ηλικίας, θα είναι:

$$\sigma_{\text{age,No}} = \sqrt{\frac{1}{S_{\text{No}} - 1}\sum_{i=1}^{S_{\text{No}}}\left(x_i - \mu_{\text{age,No}}\right)^2} = \sqrt{\frac{1}{11-1}\,5904{,}9091} = \sqrt{590{,}4909} = 24{,}3$$

Αντίστοιχα για τα πρότυπα που ανήκουν στην κλάση Yes, οι τιμές του χαρακτηριστικού age έχουν άθροισμα $\Sigma_{\text{age,Yes}} = 118$ οπότε ο μέσος όρος $\mu_{\text{age,Yes}}$ θα είναι:

$$\mu_{\text{age,Yes}} = \frac{\Sigma_{\text{age,Yes}}}{S_{\text{Yes}}} = \frac{118}{5} = 23{,}6$$

Η τυπική απόκλιση των τιμών του age, όπου x_i η κάθε τιμή ηλικίας, θα είναι:

$$\sigma_{\text{age,Yes}} = \sqrt{\frac{1}{S_{\text{Yes}} - 1}\sum_{i=1}^{S_{\text{Yes}}}\left(x_i - \mu_{\text{age,Yes}}\right)^2} = \sqrt{\frac{1}{5-1}1509{,}2} = \sqrt{377{,}3} = 19{,}4242$$

Κάνοντας τους αντίστοιχους υπολογισμούς και για τα χαρακτηριστικά appointment_hour_d και waiting_days προκύπτουν τα παρακάτω αποτελέσματα (Πίνακας 28):

	No		Yes	
	μ	**σ**	**μ**	**σ**
age	21,0909	24,3	23,6	19,4242
appointment_hour_d	15,5455	2,6968	16,4	2,8810
waiting_days	7	8,9666	5,6	2,3022

Πίνακας 28: Μέση τιμή και τυπική απόκλιση κάθε χαρακτηριστικού

Με τα παραπάνω δεδομένα μπορεί τώρα να υπολογιστεί η δεσμευμένη πιθανότητα για κάθε ένα από αυτά τα χαρακτηριστικά και δίνεται από τον τύπο:

$$P(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i})$$

$$\text{όπου } g(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Έτσι, για το χαρακτηριστικό age προκύπτουν οι δεσμευμένες πιθανότητες:

$$P(\text{age} = x|\text{No}) = \frac{1}{\sigma_{\text{age,No}}\sqrt{2\pi}} e^{-\frac{(x-\mu_{\text{age,No}})^2}{2\,\sigma_{\text{age,No}}{}^2}}$$

$$P(\text{age} = x|\text{Yes}) = \frac{1}{\sigma_{\text{age,Yes}}\sqrt{2\pi}} e^{-\frac{(x-\mu_{\text{age,Yes}})^2}{2\,\sigma_{\text{age,Yes}}{}^2}}$$

Για παράδειγμα η δεσμευμένη πιθανότητα του age = 2 όταν η κλάση είναι No $P(\text{age} = 2|\text{No})$, θα είναι:

$$P(\text{age} = 2|\text{No}) = \frac{1}{24{,}3\sqrt{2\pi}} e^{-\frac{(2-21{,}0909)^2}{2\cdot 24{,}3^2}} = \frac{1}{60{,}9104} e^{-\frac{364{,}4625}{1180{,}98}} = 0{,}0164\, e^{-0{,}3086}$$
$$= 0{,}0164 \cdot 2{,}7183^{-0{,}3086} = 0{,}0164 \cdot 0{,}7336 = 0{,}012$$

Αντίστοιχα, η δεσμευμένη πιθανότητα του age = 2 όταν η κλάση είναι Yes $P(\text{age} = 2|\text{Yes})$, θα είναι:

$$P(\text{age} = 2|\text{Yes}) = \frac{1}{19{,}4242\sqrt{2\pi}} e^{-\frac{(2-23{,}6)^2}{2\cdot 19{,}4242^2}} = 0{,}011$$

Με τον ίδιο τρόπο μπορεί να υπολογιστεί η δεσμευμένη πιθανότητα για κάθε χαρακτηριστικό με συνεχείς τιμές και για οποιαδήποτε τιμή του. Για λόγους συντομίας οι υπολογισμοί παραλείπονται και στη συνέχεια παρουσιάζονται όποια από τα αποτελέσματα είναι απαραίτητα.

Για να κατηγοριοποιηθεί ένα πρότυπο σε κάποια από τις δύο κλάσεις, θα υπολογιστεί η πιθανότητα το συγκεκριμένο πρότυπο να ανήκει σε κάθε μία από τις δύο κλάσεις χρησιμοποιώντας τον τύπο :

$$P(C_i)\ P(X|C_i) = P(C_i) \prod_{k=1}^{n} P(x_k|C_i)$$

Οι όροι $P(x_k|C_i)$ είναι οι δεσμευμένες πιθανότητες για την τιμή κάθε χαρακτηριστικού του συγκεκριμένου προτύπου και υπολογίζονται ανάλογα με το είδος του χαρακτηριστικού με έναν από τους τρόπους που αναφέρθηκαν πιο πάνω. Έτσι αν για παράδειγμα επιλεγεί προς κατηγοριοποίηση το πρότυπο No.2 του συνόλου δεδομένων (Πίνακας 29):

No.	age	sex	appointment_hour_d	waiting_days	communication_channel	no_show (class)
2	2	2	15	0	1	No

Πίνακας 29: Τιμές χαρακτηριστικών ενός τυχαίου προτύπου από το σύνολο εκπαίδευσης

Οι τιμές δεσμευμένης πιθανότητας κάθε χαρακτηριστικού για την κλάση No και Yes που χρειάζονται για να υπολογιστεί το παραπάνω γινόμενο πιθανοτήτων φαίνονται στον παρακάτω πίνακα (Πίνακας 30).:

Κλάση (Ci)	P(Ci)	P(age = 2|Ci)	P(sex = 2|Ci)	P(appointment_hour_d = 15|Ci)	P(waiting_days = 0|Ci)	P(communication_channel = 1|Ci)	P(Ci) P(X|Ci)
No	0,6875	0,0121	0,6364	0,1449	0,0328	0,3636	9,1486 x 10-6
Yes	0,3125	0,0111	0,4	0,3135	0,0090	0,4	1,5659 x 10-6

Πίνακας 30: Υπολογισμοί δεσμευμένης πιθανότητας του προτύπου για κάθε κλάση.

Το τελικό γινόμενο πιθανότητας για κάθε κλάση προκύπτει πολλαπλασιάζοντας όλες τις επιμέρους πιθανότητες του παραπάνω πίνακα.

$$P(\text{No})\ P(X|\text{No}) = 0{,}6875 \text{ x } 0{,}0121 \text{ x } 0{,}6364 \text{ x } 0{,}1449 \text{ x } 0{,}0328 \text{ x } 0{,}3636 = 9{,}1486 \text{ x } 10^{-6}$$

$$P(\text{Yes})\ P(X|\text{Yes}) = 0{,}3125 \text{ x } 0{,}0111 \text{ x } 0{,}4 \text{ x } 0{,}3135 \text{ x } 0{,}0090 \text{ x } 0{,}4 = 1{,}5659 \text{ x } 10^{-6}$$

Συγκρίνοντας τις δύο πιθανότητες προκύπτει το συμπέρασμα ότι η κλάση στην οποία αντιστοιχεί η μεγαλύτερη πιθανότητα, δηλαδή η No, είναι η κλάση στην οποία κατηγοριοποιείται το συγκεκριμένο πρότυπο. Επιπλέον η πρόβλεψη αυτή είναι σωστή.

Στη συνέχεια θα επαναληφθεί η ίδια διαδικασία για ένα άγνωστο πρότυπο αυτή τη φορά. Δηλαδή πρότυπο που δεν συμπεριλαμβάνεται στα δεδομένα εκπαίδευσης. Συγκεκριμένα για το παρακάτω πρότυπο που είναι το υπ' αριθμόν 19638 του αρχικού συνόλου δεδομένων Medical Appointments (Alvaro Flores) (Πίνακας 31):

No.	age	sex	appointment_hour_d	waiting_days	communication_channel	no_show (class)
19638	1	1	12	3	2	Yes

Πίνακας 31: Τιμές χαρακτηριστικών ενός τυχαίου προτύπου από το σύνολο ελέγχου

Οι τιμές δεσμευμένης πιθανότητας κάθε χαρακτηριστικού για την κλάση No και Yes που χρειάζονται για να υπολογιστεί το παραπάνω γινόμενο πιθανοτήτων φαίνονται στον παρακάτω πίνακα (Πίνακας 32):

Κλάση (Ci)	P(Ci)	P(age = 1|Ci)	P(sex = 1|Ci)	P(appointment_hour_d = 12|Ci)	P(waiting_days = 3|Ci)	P(communication_channel = 2|Ci)	P(Ci) P(X|Ci)
No	0,6875	0,0117	0,3636	0,0623	0,0403	0,1818	1,335 x 10-6
Yes	0,3125	0,0104	0,6	0,0633	0,0916	0,2	2,2613 x 10-6

Πίνακας 32: Υπολογισμοί δεσμευμένης πιθανότητας του άγνωστου προτύπου για κάθε κλάση.

Έτσι, για το πρότυπο 19638 προκύπτει ότι:

$$P(\text{No})\ P(X|\text{No}) = 1{,}335 \times 10^{-6}$$

$$P(\text{Yes})\ P(X|\text{Yes}) = 2{,}2613 \times 10^{-6}$$

Άρα εξάγεται το συμπέρασμα ότι θα κατηγοριοποιηθεί στην κλάση Yes και η συγκεκριμένη πρόβλεψη είναι σωστή.

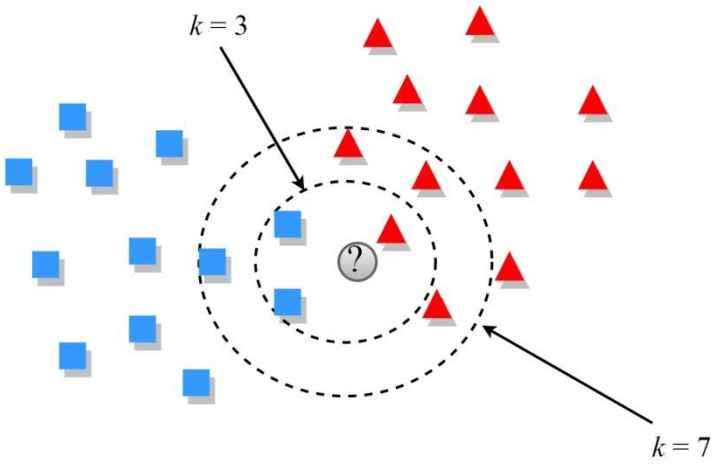
2.4.3 k-NN

Ο κατηγοριοποιητής k-NN είναι ένας από τους παλαιότερους αλγορίθμους κατηγοριοποίησης που προτάθηκε την δεκαετία του 1950 και βελτιώθηκε την δεκαετία του 1960 (Cover & Hart, 1967).

Είναι ένας από τους πιο γνωστούς αλγόριθμους κατηγοριοποίησης, έχοντας χρησιμοποιηθεί σε πάρα πολλά είδη προβλημάτων. Παρά την απλότητα του δίνει αρκετά καλά αποτελέσματα. Ανήκει στην κατηγορία των αλγορίθμων μάθησης βασισμένης στα στιγμιότυπα (instance based learning – IBL), αλλιώς γνωστοί και ως «οκνηροί» (lazy) αλγόριθμοι). Ο k-NN δηλαδή δεν κατασκευάζει κάποιο μοντέλο, αλλά χρησιμοποιεί τα δεδομένα εκπαίδευσης για να κατηγοριοποιήσει νέα πρότυπα.

Ουσιαστικά, ο κατηγοριοποιητής k-NN κατηγοριοποιεί τα νέα πρότυπα συγκρίνοντας τα με τα πρότυπα του συνόλου εκπαίδευσης και εντοπίζοντας αυτά που είναι παρόμοια. Τα δεδομένα εκπαίδευσης αναπαρίστανται ως σημεία σε ένα n-διάστατο χώρο, όπου n είναι το πλήθος των χαρακτηριστικών των δεδομένων. Έτσι, η εύρεση των προτύπων που είναι παρόμοια μπορεί να γίνει υπολογίζοντας την απόσταση μεταξύ τους. Όσο μικρότερη είναι η απόσταση μεταξύ δύο σημείων, τόσο μεγαλύτερη είναι η μεταξύ τους ομοιότητα. Τελικά ο

κατηγοριοποιητής, βρίσκοντας τα k κοντινότερα σημεία, δηλαδή τους k κοντινότερους γείτονες, μπορεί να κατηγοριοποιήσει το νέο πρότυπο στην κλάση στην οποία ανήκει η πλειοψηφία των γειτόνων (επικρατούσα κλάση).



Εικόνα 5: Αλγόριθμος k-NN

Ένα απλό παράδειγμα κατηγοριοποίησης ενός νέου προτύπου με τον αλγόριθμο k-NN φαίνεται στην Εικόνα 5. Τα πρότυπα απεικονίζονται ως σημεία 2-διαστάσεων και ανήκουν στις κλάσεις τετράγωνο και τρίγωνο. Το γκρι σημείο με ερωτηματικό είναι το νέο πρότυπο που πρόκειται να κατηγοριοποιηθεί. Από το σχήμα φαίνεται ότι ο κοντινότερος γείτονας του νέου προτύπου είναι ένα τρίγωνο και έτσι αν επιλεγεί k=1, το νέο πρότυπο θα κατηγοριοποιηθεί ως τρίγωνο. Όμως αν επιλεγεί k=3 παρατηρείται ότι οι 3 κοντινότεροι γείτονες είναι δύο τετράγωνα και ένα τρίγωνο, οπότε βάσει πλειοψηφίας θα κατηγοριοποιηθεί ως τετράγωνο. Από την άλλη για μεγαλύτερες τιμές του k το αποτέλεσμα μπορεί να είναι διαφορετικό όπως για παράδειγμα για k=7, όπου η πλειοψηφία των κοντινότερων γειτόνων ανήκει στην κλάση των τριγώνων. Από το παράδειγμα αντιλαμβάνεται κανείς ότι η επιλογή του k επηρεάζει το αποτέλεσμα της κατηγοριοποίησης. Επίσης εξάγεται το συμπέρασμα ότι η βέλτιστη επιλογή του k δεν είναι εύκολη υπόθεση γιατί εξαρτάται από την φύση των δεδομένων, την κατανομή τους κλπ.

Μάλιστα, ειδικά για προβλήματα δυαδικής κατηγοριοποίησης, δηλαδή όταν τα δεδομένα περιέχουν δύο κλάσεις, η επιλογή άρτιας τιμής k μπορεί να οδηγήσει σε ισοβαθμία αν οι μισοί γείτονες ανήκουν στην πρώτη κλάση και οι άλλοι μισοί στην δεύτερη κλάση. Σε αυτή την περίπτωση η επικρατούσα κλάση θα πρέπει να επιλεγεί τυχαία ή με κάποιο άλλο τρόπο όπως για παράδειγμα με την κατ' εξαίρεση επιλογή της κλάσης του κοντινότερου γείτονα (Ougiaroglou, 2014).

Για την αποφυγή τέτοιων περιπτώσεων σε αυτού του τύπου προβλήματα λοιπόν, καλό είναι η τιμή του k να παίρνει περιττές τιμές. Η βέλτιστη τιμή του k για κάποιο σύνολο δεδομένων

πρέπει να διερευνηθεί και συνήθως προσδιορίζεται ως η τιμή που δίνει το μικρότερο σφάλμα κατηγοριοποίησης κατά την εφαρμογή της τεχνικής cross validation (Lavrač & Zupan, 2010).

Η ομοιότητα ή αλλιώς η απόσταση μεταξύ των σημείων συνήθως υπολογίζεται χρησιμοποιώντας την μετρική της Ευκλείδειας απόστασης. Η Ευκλείδεια απόσταση μεταξύ δύο σημείων $X = (x_1, x_2, \dots, x_n)$ και $Y = (y_1, y_2, \dots, y_n)$ ενός n-διάστατου χώρου είναι:

$$d(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

Όπου n είναι ο αριθμός των διαστάσεων (δηλαδή των χαρακτηριστικών) και x_k , y_k είναι το κ-οστό χαρακτηριστικό του X και Y αντίστοιχα. Μπορούν επίσης να χρησιμοποιηθούν και άλλες μετρικές όπως για παράδειγμα η απόσταση Minkowski η οποία αποτελεί γενίκευση της Ευκλείδειας απόστασης και δίνεται από τον τύπο:

$$d(X,Y) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{\frac{1}{r}}$$

όπου το r είναι παράμετρος που ορίζει τον βαθμό της απόστασης. Ανάλογα με την τιμή του ο γενικός τύπος παίρνει συγκεκριμένες μορφές. Είναι δυνατόν να χρησιμοποιηθούν διάφορες τιμές του r, αλλά ειδικά για r = 2 προκύπτει ο τύπος της Ευκλείδειας απόστασης, ενώ για r = 1 προκύπτει ο τύπος της απόστασης *Manhattan:*

$$d(X,Y) = \sum_{k=1}^{n} |x_k - y_k|$$

Επιπλέον, έχουν προταθεί και πολλές άλλες μετρικές όπως για παράδειγμα απόσταση Chebyshev, Mahalanobis κ.α., αλλά και διάφορες μετρικές ομοιότητας που μπορούν να λάβουν υπ' όψιν και κατηγορικά χαρακτηριστικά (Deza & Deza, 2009).

Η πιο απλή περίπτωση μετρικής ομοιότητας μεταξύ κατηγορικών χαρακτηριστικών είναι να ορίζεται η απόσταση ως 0 αν οι τιμές του χαρακτηριστικού είναι ίδιες και 1 αν δεν είναι ίδιες, δηλαδή:

$$d(x,y) = \begin{cases} 0, & x = y \\ 1, & x \neq y \end{cases}$$

(Tan et al., 2014).

Ένα άλλο σημαντικό θέμα που έχει σχέση με τον υπολογισμό της απόστασης είναι η τάξη μεγέθους και το εύρος τιμών κάθε χαρακτηριστικού που παίρνει συνεχείς αριθμητικές τιμές. Αν λοιπόν υπάρχουν χαρακτηριστικά που οι τιμές τους έχουν διαφορετικές τάξεις μεγέθους,

δεν θα συνεισφέρουν στον ίδιο βαθμό στον υπολογισμό της απόστασης και αν η διαφορά είναι επαρκώς μεγάλη, το χαρακτηριστικό που παίρνει τιμές με την μικρότερη τάξη μεγέθους μπορεί να έχει αμελητέα ή μηδενική συμβολή στο τελικό αποτέλεσμα. Για την αποφυγή αυτού του προβλήματος μία λύση είναι να γίνει κανονικοποίηση στις τιμές των δεδομένων. Μια διαδεδομένη τεχνική κανονικοποίησης λοιπόν είναι για παράδειγμα η προσαρμογή των τιμών, ώστε να βρίσκονται στο διάστημα [0, 1]

Ο κατηγοριοποιητής k-NN είναι πολύ δημοφιλής λόγω της απλότητας του και της ευκολίας στην υλοποίηση του. Επιπρόσθετα λόγω του ότι είναι αλγόριθμος lazy, είναι βολικός σε περιπτώσεις που τα δεδομένα μεταβάλλονται καθώς δεν χρειάζεται φάση εκπαίδευσης, αλλά μπορεί να χρησιμοποιηθεί άμεσα για παραγωγή προβλέψεων.

Από την άλλη πλευρά το κόστος υπολογισμού όλων των αποστάσεων μεταξύ ενός νέου προτύπου και όλων των προτύπων των δεδομένων εκπαίδευσης, είναι ιδιαίτερα υψηλό, ειδικά για μεγάλα σύνολα δεδομένων. Επιπλέον το γεγονός ότι τα δεδομένα εκπαίδευσης πρέπει να είναι συνεχώς διαθέσιμα μπορεί να προκαλέσει δυσκολίες σε μεγάλα σύνολα δεδομένων καθώς αυξάνει τις απαιτήσεις σε αποθηκευτικό χώρο.

Παράδειγμα κατηγοριοποίησης με k-NN

Για την κατανόηση του αλγορίθμου κατηγοριοποίησης k-NN, δίνεται παρακάτω ένα αναλυτικό παράδειγμα. Τα δεδομένα που χρησιμοποιήθηκαν είναι ένα δείγμα 16 εγγραφών από το σύνολο δεδομένων Medical Appointments (Alvaro Flores). Τα δεδομένα αυτά αποτελούν το σύνολο δεδομένων εκπαίδευσης, όπως φαίνεται στον παρακάτω πίνακα (Πίνακας 33).

No.	age	sex	appointment_hour_d	waiting_days	communication_channel	no_show (class)
1	11	1	14	1	3	No
2	2	2	15	0	1	No
3	11	2	19	8	3	Yes
4	61	2	14	8	1	No
5	6	2	19	16	3	No
6	9	2	15	29	1	No
7	66	1	18	0	2	No
8	13	1	19	5	3	Yes
9	20	2	9	1	3	No
10	3	2	17	10	1	No
11	8	1	16	2	1	Yes

12	54	1	16	6	1	Yes
13	3	1	17	8	3	No
14	45	2	17	4	2	No
15	6	1	16	0	3	No
16	32	2	12	7	2	Yes

Πίνακας 33: Σύνολο δεδομένων εκπαίδευσης

Στα παραπάνω δεδομένα το χαρακτηριστικό της κλάσης είναι το no_show που μπορεί να πάρει τιμές Yes και No.

Για τον έλεγχο της επίδοσης του αλγορίθμου k-NN θα χρησιμοποιηθούν κάποια άγνωστα πρότυπα. Δηλαδή πρότυπα που δεν συμπεριλαμβάνονται στα δεδομένα εκπαίδευσης. Συγκεκριμένα τα πρότυπα που φαίνονται στον παρακάτω πίνακα και είναι τα υπ' αριθμόν 7962, 19638 και 25956, τα οποία επιλέχθηκαν τυχαία από το αρχικό σύνολο δεδομένων Medical Appointments (Alvaro Flores) (Πίνακας 34):

No.	**age**	**sex**	**appointment_hour_d**	**waiting_days**	**communication_channel**	**no_show (class)**
7962	79	1	9	2	1	No
19638	1	1	12	3	2	Yes
25956	46	2	13	14	1	Yes

Πίνακας 34: Άγνωστα πρότυπα για κατηγοριοποίηση

Για το παράδειγμα αυτό θα δοκιμαστούν οι τιμές k=1, 3 και 5. Για την κατηγοριοποίηση κάθε αγνώστου προτύπου πρέπει να υπολογιστεί η απόσταση του από κάθε ένα πρότυπο του συνόλου εκπαίδευσης. Στη συνέχεια μπορεί να επιλεγεί μια τιμή k, δηλαδή το πλήθος των κοντινότερων γειτόνων που θα ληφθεί υπ' όψιν και με βάση τις αποστάσεις θα βρεθεί η κλάση της πλειοψηφίας των k κοντινότερων γειτόνων.

Ξεκινώντας με το πρώτο άγνωστο πρότυπο (No. 7962) εφαρμόζεται ο τύπος υπολογισμού της Ευκλείδειας απόστασης. Μάλιστα για ευκολία μπορεί να υπολογιστεί αρχικά η διαφορά των τιμών για κάθε χαρακτηριστικό υψωμένη στο τετράγωνο. Στα κατηγορικά χαρακτηριστικά (sex και communication_channel) θα χρησιμοποιηθεί ο υπολογισμός ομοιότητας, όπου η απόσταση θα είναι 0 για ίδιες τιμές και 1 για διαφορετικές. Η τετραγωνική ρίζα του αθροίσματος των διαφορών αυτών θα δώσει την ζητούμενη απόσταση.

Η απόσταση του άγνωστου προτύπου (No. 7962) από το πρώτο πρότυπο του συνόλου εκπαίδευσης λοιπόν θα υπολογιστεί ως εξής (Πίνακας 35):

Χαρακτηριστικό	Τιμή στο άγνωστο πρότυπο (x)	Τιμή στο 1ο πρότυπο (y)	$(x-y)^2$
age	79	11	$(79-11)^2 = 4624$
sex	1	1	$1 = 1 \implies 0$
appointment_hour_d	9	14	$(9-14)^2 = 25$
waiting_days	2	1	$(2-1)^2 = 1$
communication_channel	1	3	$1 \neq 3 \implies 1$

Πίνακας 35: Υπολογισμοί για πρότυπα No. 7962 και No. 1

Άρα τελικά η απόσταση d του άγνωστου προτύπου (No. 7962) από το 1ο πρότυπο του συνόλου εκπαίδευσης θα είναι:

$$d = \sqrt{4624 + 0 + 25 + 1 + 1} = \sqrt{4651} = 68{,}198$$

Κάνοντας τους αντίστοιχους υπολογισμούς και για τα υπόλοιπα πρότυπα του συνόλου εκπαίδευσης, προκύπτουν τα αποτελέσματα του παρακάτω πίνακα (Πίνακας 36):

Πρότυπο No.	**age** $(x-y)^2$	**sex** $(x-y)^2$	**appointment_hour_d** $(x-y)^2$	**waiting_days** $(x-y)^2$	**communication_channel** $(x-y)^2$	**απόσταση d**	**Class no_show**
1	4624	0	25	1	1	68,198	No
2	5929	1	36	4	0	77,266	No
3	4624	1	100	36	1	69,007	Yes
4	324	1	25	36	0	19,647	No
5	5329	1	100	196	1	75,013	No
6	4900	1	36	729	0	75,273	No
7	169	0	81	4	1	15,969	No
8	4356	0	100	9	1	66,828	Yes
9	3481	1	0	1	1	59,025	No
10	5776	1	64	64	0	76,844	No
11	5041	0	49	0	0	71,344	Yes
12	625	0	49	16	0	26,268	Yes
13	5776	0	64	36	1	76,662	No
14	1156	1	64	4	1	35,014	No
15	5329	0	49	4	1	73,369	No
16	2209	1	9	25	1	47,381	Yes

Πίνακας 36: Αποστάσεις από πρότυπο No. 7962

Για την εύρεση των κοντινότερων γειτόνων του άγνωστου προτύπου No. 7962, θα πρέπει να εντοπιστούν οι γείτονες με την μικρότερη απόσταση από το πρότυπο No. 7962. Επομένως ο παραπάνω πίνακας θα πρέπει να ταξινομηθεί κατά αύξουσα απόσταση d. Μάλιστα χρειάζονται μόνο τα 5 πρότυπα με την μικρότερη απόσταση, επειδή η μέγιστη τιμή του k που θα δοκιμαστεί είναι k = 5 (Πίνακας 37).

Πρότυπο No.	απόσταση d	Class no_show	A/A
7	15,969	No	1
4	19,647	No	2
12	26,268	Yes	3
14	35,014	No	4
16	47,381	Yes	5

Πίνακας 37: Αποστάσεις των 5 κοντινότερων γειτόνων του προτύπου No. 7962

Για k = 1 λοιπόν το άγνωστο πρότυπο No. 7962 θα κατηγοριοποιηθεί στην κλάση No. Για k = 3 η πλειοψηφία των κοντινότερων γειτόνων ανήκει στην κλάση No, άρα θα κατηγοριοποιηθεί στην κλάση No. Τέλος, για k = 5, η πλειοψηφία των κοντινότερων γειτόνων ανήκει στην κλάση No, άρα και πάλι το άγνωστο πρότυπο No. 7962 θα κατηγοριοποιηθεί στην κλάση No. Η πραγματική κλάση του άγνωστου προτύπου No. 7962 είναι η κλάση No, άρα η κατηγοριοποίηση και στις 3 περιπτώσεις είναι σωστή.

Κάνοντας τους παραπάνω υπολογισμούς για το 2ο άγνωστο πρότυπο (No. 19638), προκύπτουν τα παρακάτω αποτελέσματα (Πίνακας 38):

Πρότυπο No.	age $(x-y)^2$	sex $(x-y)^2$	appointment_hour_d $(x-y)^2$	waiting_days $(x-y)^2$	communication_channel $(x-y)^2$	απόσταση d	Class no_show
1	100	0	4	4	1	10,440	No
2	1	1	9	9	0	4,472	No
3	100	1	49	25	1	13,266	Yes
4	3600	1	4	25	0	60,249	No
5	25	1	49	169	1	15,652	No
6	64	1	9	676	0	27,386	No
7	4225	0	36	9	1	65,353	No
8	144	0	49	4	1	14,071	Yes
9	361	1	9	4	1	19,391	No
10	4	1	25	49	0	8,888	No
11	49	0	16	1	0	8,124	Yes
12	2809	0	16	9	0	53,235	Yes
13	4	0	25	25	1	7,416	No
14	1936	1	25	1	1	44,317	No
15	25	0	16	9	1	7,141	No
16	961	1	0	16	1	31,289	Yes

Πίνακας 38: Αποστάσεις από πρότυπο No. 19638

Μετά από τον υπολογισμό των αποστάσεων προκύπτει ο παρακάτω πίνακας στον οποίο φαίνονται οι 5 κοντινότεροι γείτονες του άγνωστου προτύπου No. 19638 ταξινομημένοι κατά την απόσταση d (Πίνακας 39):

Πρότυπο No.	απόσταση d	Class no_show	A/A
2	4,472	No	1
15	7,141	No	2
13	7,416	No	3
11	8,124	Yes	4
10	8,888	No	5

Πίνακας 39: Αποστάσεις των 5 κοντινότερων γειτόνων του προτύπου No. 19638

Για k = 1 λοιπόν το άγνωστο πρότυπο No. 19638 θα κατηγοριοποιηθεί στην κλάση No. Επίσης, για τιμές k=3 και k=5, η πλειοψηφία των κοντινότερων γειτόνων ανήκει στην κλάση No, άρα και πάλι θα κατηγοριοποιηθεί στην κλάση No. Η πραγματική κλάση του άγνωστου προτύπου No. 19638 είναι η κλάση Yes, άρα η κατηγοριοποίηση και στις 3 περιπτώσεις είναι λανθασμένη.

Τέλος, για το 3ο άγνωστο πρότυπο (No. 25956), προκύπτουν τα παρακάτω αποτελέσματα (Πίνακας 40):

Πρότυπο No.	age $(x-y)^2$	sex $(x-y)^2$	appointment_hour_d $(x-y)^2$	waiting_days $(x-y)^2$	communication_channel $(x-y)^2$	απόσταση d	Class no_show
1	1225	1	1	169	1	37,376	No
2	1936	0	4	196	0	46,217	No
3	1225	0	36	36	1	36,028	Yes
4	225	0	1	36	0	16,186	No
5	1600	0	36	4	1	40,509	No
6	1369	0	4	225	0	39,975	No
7	400	1	25	196	1	24,960	No
8	1089	1	36	81	1	34,756	Yes
9	676	0	16	169	1	29,360	No
10	1849	0	16	16	0	43,370	No
11	1444	1	9	144	0	39,975	Yes
12	64	1	9	64	0	11,747	Yes
13	1849	1	16	36	1	43,623	No
14	1	0	16	100	1	10,863	No
15	1600	1	9	196	1	42,509	No
16	196	0	1	49	1	15,716	Yes

Πίνακας 40: Αποστάσεις από πρότυπο No. 25956

Μετά από τον υπολογισμό των αποστάσεων προκύπτει ο παρακάτω πίνακας στον οποίο φαίνονται οι 5 κοντινότεροι γείτονες του άγνωστου προτύπου No. 25956 ταξινομημένοι κατά την απόσταση d (Πίνακας 41):

Πρότυπο No.	απόσταση d	Class no_show	A/A
14	10,863	No	1
12	11,747	Yes	2
16	15,716	Yes	3
4	16,186	No	4
7	24,960	No	5

Πίνακας 41: Αποστάσεις των 5 κοντινότερων γειτόνων του προτύπου No. 25956

Για το συγκεκριμένο άγνωστο πρότυπο προκύπτει ότι για k = 1 το πρότυπο αυτό θα κατηγοριοποιηθεί στην κλάση No. Για k=3 η πλειοψηφία των κοντινότερων γειτόνων ανήκει στην κλάση Yes, άρα θα κατηγοριοποιηθεί στην κλάση Yes. Αντίθετα, για k=5 η πλειοψηφία των κοντινότερων γειτόνων ανήκει στην κλάση No, άρα θα κατηγοριοποιηθεί στην κλάση No. Η πραγματική κλάση του άγνωστου προτύπου No. 25956 είναι η κλάση Yes, άρα η κατηγοριοποίηση είναι σωστή σε μία περίπτωση και λανθασμένη στις άλλες δύο.

Συνοψίζοντας τα παραπάνω αποτελέσματα κατηγοριοποίησης για όλα τα άγνωστα πρότυπα και τις τιμές του k, προκύπτει ο παρακάτω πίνακας (Πίνακας 42):

No.	no_show (class)	k	no_show (πρόβλεψη)	Αποτέλεσμα
7962	No	1	No	Σωστή
		3	No	Σωστή
		5	No	Σωστή
19638	Yes	1	No	Λανθασμένη
		3	No	Λανθασμένη
		5	No	Λανθασμένη
25956	Yes	1	No	Λανθασμένη
		3	Yes	Σωστή
		5	No	Λανθασμένη

Πίνακας 42: Αποτελέσματα κατηγοριοποίησης των άγνωστων προτύπων με k-NN

3

Μέτρηση της επίδοσης κατηγοριοποιητών

Για την αξιολόγηση της επίδοσης των μεθόδων κατηγοριοποίησης υπάρχει πληθώρα μετρικών που χρησιμοποιούνται και χωρίζονται σε ποσοτικές και ποιοτικές. Ορισμένες από τις ποσοτικές μετρικές είναι η ακρίβεια πρόβλεψης (Accuracy), η ορθότητα (Precision) και η ευαισθησία (Recall). Οι συγκεκριμένες τεχνικές αναλύονται σε επόμενο κεφάλαιο.

Από την άλλη πλευρά μερικές από τις ποιοτικές μετρικές που υπάρχουν είναι η ερμηνευσιμότητα (interpretability - comprehensibility), η επεκτασιμότητα (scalability), η ανθεκτικότητα (robustness) και η υπολογιστική ικανότητα (computational complexity) που αναλύονται διεξοδικά παρακάτω.

3.1 Ποιοτικές μετρικές

3.1.1 Ερμηνευσιμότητα (Interpretability - Comprehensibility)

Η έννοια της ερμηνευσιμότητας αναφέρεται στο πόσο καλά κατανοούν οι άνθρωποι τα μοντέλα που παράγει ένας αλγόριθμος εξόρυξης γνώσης. Πολλοί αλγόριθμοι όπως νευρωνικά δίκτυα και μηχανές διανυσμάτων υποστήριξης (SVM), παράγουν κατηγοριοποιητές οι οποίοι αναπαρίστανται από μεγάλα σύνολα αριθμητικών παραμέτρων. Έτσι όμως δεν είναι εύκολα κατανοητά και ερμηνεύσιμα από τους ανθρώπους. Για το λόγο αυτό τα συγκεκριμένα μοντέλα αναφέρονται συχνά ως “μαύρα κουτιά”. Δηλαδή δέχονται

δεδομένα εισόδου και παράγουν αποτελέσματα χωρίς να είναι γνωστός ο τρόπος υπολογισμού τους. Αντίθετα κάποιοι άλλοι αλγόριθμοι κατασκευάζουν μοντέλα, που είναι πιο κατανοητά και ερμηνεύσιμα από τον άνθρωπο, όπως τα δέντρα αποφάσεων. Αυτό συμβαίνει γιατί τα δέντρα αποφάσεων μπορούν να μετατραπούν άμεσα σε κανόνες κατηγοριοποίησης της μορφής EAN-TOTE, οι οποίοι προσομοιάζουν τη φυσική γλώσσα (Maimon & Rokach, 2010).

Η ερμηνευσιμότητα είναι σημαντική για διάφορους λόγους. Για παράδειγμα για να εμπιστευθεί κάποιος το αποτέλεσμα ενός μοντέλου, ώστε να λάβει κάποια κρίσιμη απόφαση (όπως για ιατρικές διαγνώσεις), θα πρέπει προηγουμένως να κατανοεί πλήρως με βάση ποια κριτήρια παράχθηκε το συγκεκριμένο αποτέλεσμα. Ένα άλλο πλεονέκτημα της ερμηνευσιμότητας είναι το ακόλουθο. Ένας αλγόριθμος μπορεί να ανακαλύπτει κρυμμένα χαρακτηριστικά, σχέσεις και μοτίβα μέσα στα δεδομένα εισόδου, δηλαδή νέες πληροφορίες, η σημασία των οποίων ήταν προηγουμένως άγνωστη. Εάν αυτές οι νέες πληροφορίες που ανακαλύφθηκαν από τον αλγόριθμο είναι κατανοητές, τότε μπορούν να γίνουν επεξεργάσιμες από τον άνθρωπο. Κάτι εξίσου σημαντικό είναι η δυνατότητα βελτίωσης της απόδοσης ενός μοντέλου που παρέχεται μέσω της ερμηνευσιμότητας. Δηλαδή αν ο τρόπος αναπαράστασης των χαρακτηριστικών είναι ερμηνεύσιμος, μπορεί να βελτιωθεί ακόμη περισσότερο. Αυτό επιτυγχάνεται τροποποιώντας στοχευμένα συγκεκριμένα χαρακτηριστικά του συνόλου δεδομένων ή παραμέτρους του αλγορίθμου, με στόχο την δημιουργία ακόμη καλύτερων αναπαραστάσεων χαρακτηριστικών. Επιπρόσθετα οι αλγόριθμοι εξόρυξης γνώσης μπορούν να χρησιμοποιηθούν για να τελειοποιήσουν υπάρχουσες επιστημονικές θεωρίες. Η διαδικασία αυτή ονομάζεται “theory-refinement” (τελειοποίηση θεωρίας). Για να ολοκληρωθεί σωστά αυτή η διαδικασία, είναι σημαντικό να μπορεί ο αλγόριθμος να εκφράσει με τρόπο κατανοητό τις μεταβολές που πραγματοποιήθηκαν στην επιστημονική θεωρία κατά την διάρκεια της εξόρυξης γνώσης. Εν κατακλείδι οι αλγόριθμοι εξόρυξης γνώσης μπορούν να χρησιμοποιηθούν ως εργαλείο ανάλυσης, ικανό να προσφέρει καινοτόμα γνώση. (Zhou, 2005).

3.1.2 Επεκτασιμότητα (Scalability)

Επεκτασιμότητα ορίζεται ως η ικανότητα μιας μεθόδου να κατασκευάζει αποδοτικά ένα μοντέλο κατηγοριοποίησης παρά το γεγονός ότι, ο όγκος των δεδομένων που πρέπει να διαχειριστεί και να επεξεργαστεί, είναι πολύ μεγάλος.

Πιο συγκεκριμένα η έννοια της “επεκτασιμότητας” αναφέρεται συνήθως σε σύνολα δεδομένων που πληρούν τουλάχιστον μία από τις ακόλουθες ιδιότητες: μεγάλος αριθμός εγγραφών ή υψηλή διαστατικότητα (high dimensionality). Όταν ένα σύνολο δεδομένων έχει

υψηλή διαστατικότητα σημαίνει ότι έχει μεγάλο αριθμό διαστάσεων, δηλαδή χαρακτηριστικών.

Γενικά οι μεγάλες βάσεις δεδομένων έχουν γίνει ο κανόνας σε πολλούς κλάδους, συμπεριλαμβανομένης της αστρονομίας, της μοριακής βιολογίας, των οικονομικών, του εμπορίου, της υγειονομικής περίθαλψης και πολλών άλλων. Για παράδειγμα οργανισμοί με πολύ μεγάλο όγκο πληροφοριών όπως εταιρείες τηλεπικοινωνιών και τράπεζες θεωρείται ότι συσσωρεύουν αρκετά terabyte πρωτογενών δεδομένων κάθε ένα έως δύο χρόνια.

Το παράδοξο είναι ότι, αν και κάποτε το όνειρο ενός αναλυτή δεδομένων, ήταν η δυνατότητα να έχει στη διάθεση του έναν πολύ μεγάλο όγκο δεδομένων για επεξεργασία και εξαγωγή συμπερασμάτων, σήμερα το συνώνυμο του "πολύ μεγάλου" αφορά δεδομένα της τάξης των "terabyte". Δηλαδή έναν πολύ δύσκολα διαχειρίσιμο όγκο πληροφοριών που τελικά δυσχεραίνει το έργο του αναλυτή. Επιπλέον η διαχείριση και η ανάλυση τεράστιων αποθηκών δεδομένων απαιτεί εξειδικευμένο και πολύ ακριβό υλικό και λογισμικό. Το γεγονός αυτό αποτελεί σημαντικό μειονέκτημα, καθώς αναγκάζει συχνά μια εταιρεία να εκμεταλλεύεται μόνο ένα μικρό μέρος των αποθηκευμένων δεδομένων που έχει στη διάθεση της, χάνοντας έτσι ευκαιρίες για ανακάλυψη πολύτιμων ίσως πληροφοριών.

Όσον αφορά την χρήση αλγορίθμων εξόρυξης γνώσης, οι "κλασικοί" αλγόριθμοι επαγωγής έχουν εφαρμοστεί με πρακτική επιτυχία σε πολλά σχετικά απλά και μικρής κλίμακας προβλήματα. Ωστόσο η εφαρμογή τους και η προσπάθεια ανακάλυψης γνώσης στην πραγματική ζωή και σε μεγάλες βάσεις δεδομένων, εμφανίζει προβλήματα χρόνου και μνήμης (Maimon & Rokach, 2010).

Για να αντιμετωπιστούν τέτοιου είδους προβλήματα, χρησιμοποιούνται πιο αποτελεσματικοί αλγόριθμοί, όπως αλγόριθμοι δειγματοληψίας, πρόβλεψης και μαζικής παράλληλης επεξεργασίας (Fayyad, 1996).

3.1.3 Υπολογιστική πολυπλοκότητα (Computational complexity)

Ένα άλλο χρήσιμο κριτήριο για τη σύγκριση επαγωγικών αλγορίθμων και κατηγοριοποιητών είναι η υπολογιστική τους πολυπλοκότητα. Γενικότερα η υπολογιστική πολυπλοκότητα αφορά τους πόρους που απαιτεί ένας αλγόριθμος για την εκτέλεση του και κυρίως τον χρόνο που χρειάζεται για να διεκπεραιώσει μία διαδικασία αλλά και την μνήμη που απαιτεί για την ολοκλήρωση της.

Η υπολογιστική πολυπλοκότητα διακρίνεται σε 3 τύπους:

a) *Υπολογιστική πολυπλοκότητα για τη δημιουργία ενός νέου κατηγοριοποιητή:* Αυτή είναι η πιο σημαντική μετρική. Ειδικά δε, όταν είναι ανάγκη ο αλγόριθμος εξόρυξης

γνώσης να κλιμακωθεί σε σύνολα δεδομένων πολύ μεγαλύτερου όγκου, απ' ότι αυτά στα οποία εφαρμοζόταν μέχρι πρότινος. Σε μια τέτοια περίπτωση, επειδή η υπολογιστική πολυπλοκότητα χρόνου που έχουν οι περισσότεροι αλγόριθμοι, είναι χειρότερη από την γραμμική, ως προς τον αριθμό των χαρακτηριστικών ή εγγραφών, η διαδικασία εξόρυξης γνώσης, μπορεί να είναι "απαγορευτικά δαπανηρή" ως προς τον χρόνο που χρειάζεται για να ολοκληρωθεί.

b) *Υπολογιστική πολυπλοκότητα για την ενημέρωση ενός κατηγοριοποιητή*: Η υπολογιστική πολυπλοκότητα αυτού του τύπου έχει επίσης σχέση με τον χρόνο εκτέλεσης του αλγορίθμου. Δηλαδή δίνονται στον αλγόριθμο νέα δεδομένα εκπαίδευσης και το ζητούμενο είναι να υπολογιστεί ποιος είναι ο επιπλέον χρόνος που απαιτείται, ώστε να ενημερωθεί ο τρέχον κατηγοριοποιητής με τα νέα δεδομένα.

c) *Υπολογιστική πολυπλοκότητα για την κατηγοριοποίηση ενός νέου άγνωστου προτύπου:* Γενικά σε αυτό τον τύπο υπολογιστικής πολυπλοκότητας δεν δίνεται ιδιαίτερη σημασία, επειδή είναι μικρότερος, συγκριτικά με τους προηγούμενους. Ωστόσο, σε ορισμένες μεθόδους, όπως για παράδειγμα στον αλγόριθμο του κοντινότερου γείτονα (K - NN), ο χρόνος μπορεί να είναι πολύ σημαντικός. Αυτό συμβαίνει γιατί ο αλγόριθμος K - NN δεν κατασκευάζει κάποιο μοντέλο. Δηλαδή δεν κάνει κάποιου είδους εκπαίδευση, αλλά πραγματοποιεί όλους τους υπολογισμούς κατά την φάση της κατηγοριοποίησης. Κατά συνέπεια ο χρόνος εκπαίδευσης είναι μηδενικός ενώ ο χρόνος κατηγοριοποίησης μπορεί να είναι πολύ μεγάλος ανάλογα με το σύνολο δεδομένων. Σε αυτή την περίπτωση ή σε ορισμένες εφαρμογές πραγματικού χρόνου, όπως, αντιπυραυλικά συστήματα, αυτός ο τύπος υπολογιστικής πολυπλοκότητας, μπορεί να είναι κρίσιμος. Σκεφτείτε για παράδειγμα τι θα γίνει αν ο αλγόριθμος δεν έχει την ικανότητα να προβλέψει έγκαιρα την εκτόξευση πυραύλων και τελικά αυτοί φτάσουν στο στόχο τους (Maimon & Rokach, 2010)**.**

Συμπερασματικά προκύπτει ότι με το πέρασμα των χρόνων και την αύξηση του όγκου πληροφοριών προς επεξεργασία και ανάλυση, είναι πλέον εξίσου σημαντικές , τόσο η υπολογιστική πολυπλοκότητα ως προς τον χρόνο, όσο και η υπολογιστική πολυπλοκότητα ως προς το χώρο (μνήμη).

3.1.4 Ανθεκτικότητα (Robustness)

Μια άλλη σημαντική μετρική ενός κατηγοριοποιητή στην οποία συχνά δεν δίνεται η πρέπουσα σημασία, είναι η ανθεκτικότητα του (robustness) σε περιβάλλοντα με θόρυβο, δηλαδή όταν τα πρότυπα περιέχουν θόρυβο. Ως ανθεκτικότητα ορίζεται η ικανότητα ενός κατηγοριοποιητή να χειρίζεται θόρυβο ή δεδομένα με ελλιπείς τιμές και να πραγματοποιεί

σωστές προβλέψεις (Maimon & Rokach, 2010). Η ανθεκτικότητα λοιπόν ενός κατηγοριοποιητή είναι ιδιαίτερα σημαντική. Ειδικά δε όταν χρησιμοποιείται στον πραγματικό κόσμο, όπου δεν υπάρχει η δυνατότητα ελέγχου της ορθότητας των δεδομένων εισόδου και γενικότερα σε περιβάλλοντα όπου η παρουσία θορύβου είναι αναπόφευκτη. Για παράδειγμα σε ένα σύνολο δεδομένων που περιέχει χαρακτηριστικά ασθενών όπως την ηλικία, το ύψος, το βάρος κ. α., με στόχο την διάγνωση μίας ασθένειας, δεν μπορεί να διασφαλιστεί, ότι οι τιμές που έχουν καταχωρηθεί δεν έχουν κανένα σφάλμα μέτρησης. Πιο συγκεκριμένα ως προς το βάρος ενός ασθενή, ο αλγόριθμος δεν μπορεί να γνωρίζει με σιγουριά το πραγματικό βάρος του. Αυτό συμβαίνει γιατί στην πραγματικότητα το βάρος του ασθενή παρουσιάζει διακυμάνσεις. Δηλαδή θα υπάρχει μια μικρή απόκλιση σε αυτό, ανάλογα με την ώρα της ημέρας που ζυγίστηκε ο ασθενής. Σε αυτές τις περιπτώσεις, είναι ζωτικής σημασίας ο κατηγοριοποιητής να παρουσιάζει μεγάλη ανθεκτικότητα. Με άλλα λόγια, μια αρκετά μικρή απόκλιση λόγω θορύβου στις τιμές ορισμένων χαρακτηριστικών ενός προτύπου, δεν θα πρέπει να έχει ως αποτέλεσμα την αλλαγή της πρόβλεψης που κάνει ο κατηγοριοποιητής για την κλάση του προτύπου. Διαφορετικά θεωρείται ότι δεν έχει μεγάλη ανθεκτικότητα στο θόρυβο (Fawzi et al., 2016).

3.2 Ποσοτικές μετρικές

Για να αξιολογηθεί η επίδοση ενός κατηγοριοποιητή στον διαχωρισμό δεδομένων σε δύο κλάσεις πρέπει να συγκριθούν οι προβλέψεις του κατηγοριοποιητή με τα αντίστοιχα πραγματικά δεδομένα και να καταγραφεί τελικά πόσες από τις προβλέψεις που έκανε ήταν σωστές και πόσες λανθασμένες.

Στην απλούστερη περίπτωση που το πρόβλημα έχει δύο κλάσεις, η μία κλάση θεωρείται ως η Θετική (Positive) και η άλλη κλάση ως η Αρνητική (Negative). Οι σωστές προβλέψεις για την Θετική κλάση ονομάζονται True Positives (TP) και αντίστοιχα οι σωστές προβλέψεις για την Αρνητική κλάση ονομάζονται True Negatives (TN). Οι λανθασμένες προβλέψεις είναι επίσης δύο ειδών. Οι λανθασμένες προβλέψεις για την Θετική κλάση (δηλαδή οι προβλέψεις για τα πρότυπα που πραγματικά ανήκουν στην Αρνητική κλάση αλλά κατηγοριοποιήθηκαν λανθασμένα στην Θετική) ονομάζονται False Positives (FP). Οι λανθασμένες προβλέψεις για την Αρνητική κλάση (δηλαδή οι προβλέψεις για τα πρότυπα που πραγματικά ανήκουν στην Θετική κλάση αλλά κατηγοριοποιήθηκαν λανθασμένα στην Αρνητική) ονομάζονται False Negatives (FN).

Οι παραπάνω τιμές καταγράφονται συνοπτικά σε ένα πίνακα που ονομάζεται πίνακας σύγχυσης (confusion matrix). Ο πίνακας σύγχυσης (Πίνακας 43) για δύο κλάσεις αποτελείται από δύο γραμμές και δύο στήλες. Οι γραμμές αντιστοιχούν στις πραγματικές τιμές κλάσεων

που θα έπρεπε να προβλεφθούν (Γεγονός), ενώ οι στήλες αντιστοιχούν στις τιμές κλάσεων που προέβλεψε ο κατηγοριοποιητής (Πρόβλεψη) (Davis & Goadrich, 2006).

		ΠΡΟΒΛΕΨΗ	
		ΘΕΤΙΚΗ ΚΛΑΣΗ	**ΑΡΝΗΤΙΚΗ ΚΛΑΣΗ**
ΓΕΓΟΝΟΣ	**ΘΕΤΙΚΗ ΚΛΑΣΗ**	**TP**	**FN**
	ΑΡΝΗΤΙΚΗ ΚΛΑΣΗ	**FP**	**TN**

Πίνακας 43: Πίνακας Σύγχυσης (Confusion Matrix)

3.2.1 Ακρίβεια (Accuracy)

Η ακρίβεια (accuracy) (Metz, 1978) είναι ο λόγος του πλήθους των σωστών προβλέψεων προς το συνολικό πλήθος των προβλέψεων, δηλαδή:

$$Ακρίβεια\ (Accuracy) = \frac{Πλήθος\ σωστών\ προβλέψεων}{Συνολικό\ πλήθος\ προβλέψεων}$$

Σύμφωνα με τον πίνακα σύγχυσης το πλήθος των σωστών προβλέψεων είναι το TP+TN, ενώ το συνολικό πλήθος των προβλέψεων είναι TP+TN+FP+FN, άρα τελικά η ακρίβεια υπολογίζεται ως εξής:

$$Ακρίβεια\ (Accuracy) = \frac{TP + TN}{TP + TN + FP + FN}$$

3.2.2 Ορθότητα (Precision ή Positive Predicted Value)

Η ορθότητα (Precision) υπολογίζεται για κάθε κλάση ξεχωριστά (Davis & Goadrich, 2006).

Η ορθότητα για την Θετική κλάση (P) είναι ο λόγος του πλήθους των σωστών προβλέψεων για την Θετική κλάση, προς το συνολικό πλήθος προβλέψεων για την Θετική κλάση, δηλαδή:

$$Ορθότητα(Θ) = Precision(P) = \frac{TP}{TP + FP}$$

Η ορθότητα για την Αρνητική κλάση (N) είναι ο λόγος του πλήθους των σωστών προβλέψεων για την Αρνητική κλάση, προς το συνολικό πλήθος προβλέψεων για την Αρνητική κλάση, δηλαδή:

$$Ορθότητα(A) = Presicion(N) = \frac{TN}{TN + FN}$$

3.2.3 Ευαισθησία (Recall ή Sensitivity ή True Positive Rate)

Η Ευαισθησία (Recall) υπολογίζεται για κάθε κλάση ξεχωριστά (Davis & Goadrich, 2006). Η Ευαισθησία για την Θετική κλάση (P) είναι ο λόγος του πλήθους των σωστών προβλέψεων για την Θετική κλάση, προς το συνολικό πλήθος προτύπων που ανήκουν πραγματικά στη Θετική κλάση, δηλαδή:

$$Ευαισθησία(Θ) = Recall(P) = \frac{TP}{TP + FN}$$

Η Ευαισθησία για την Αρνητική κλάση (N) είναι ο λόγος του πλήθους των σωστών προβλέψεων για την Αρνητική κλάση, προς το συνολικό πλήθος προτύπων που ανήκουν πραγματικά στην Αρνητική κλάση, δηλαδή:

$$Ευαισθησία(A) = Recall(N) = \frac{TN}{TN + FP}$$

3.2.4 F-measure

Η μετρική F-measure αποτελεί συνδυασμό των μετρικών Precision και Recall. Συγκεκριμένα είναι ο αρμονικός μέσος μεταξύ των 2 πιο πάνω μετρικών.

Για την θετική κλάση (P) δίνεται από τον τύπο:

$$F - measure_{(Yes)} = \frac{2 \cdot Recall_{(Yes)} \cdot Precision_{(Yes)}}{Recall_{(Yes)} + Precision_{(Yes)}}$$

Για την αρνητική κλάση (N) δίνεται από τον τύπο:

$$F - measure_{(No)} = \frac{2 \cdot Recall_{(No)} \cdot Precision_{(No)}}{Recall_{(No)} + Precision_{(No)}}$$

Η τιμή της μετρικής F-measure είναι ανάμεσα στις τιμές των μετρικών Precision και Recall, αλλά τείνει να είναι πιο κοντά στην χαμηλότερη μεταξύ των 2 παραπάνω τιμών. Στην πραγματικότητα η μετρική F-measure είναι ο αρμονικός μέσος της τιμής των μετρικών Precision και Recall. Αναλυτικότερα στην περίπτωση που η μετρική F-measure έχει χαμηλή τιμή, αυτό μπορεί να σημαίνει 2 πράγματα. Το πρώτο είναι ότι αντίστοιχα και οι 2 τιμές των μετρικών Precision και Recall θα είναι εξίσου χαμηλές. Το δεύτερο είναι ότι μία από τις 2

τιμές (Precision ή Recall) θα είναι πολύ χαμηλή. Από την άλλη πλευρά μια υψηλή τιμή στην μετρική F-measure εξασφαλίζει ότι και οι 2 τιμές των μετρικών Precision και Recall θα είναι επίσης υψηλές (Sun et al., 2009).

Παράδειγμα με Precision / Recall Με Ασθενείς

Αφού έγινε αναλυτική επεξήγηση των εννοιών της ακρίβειας, ορθότητας και ευαισθησίας, θα περιγραφεί στην πράξη μέσω παραδείγματος, πως αυτές εφαρμόζονται. Στο παράδειγμα λοιπόν που παρατίθεται, υπάρχουν δύο κλάσεις. Οι ασθενείς που δεν θα εμφανιστούν στο ραντεβού θεωρούνται ως η θετική κλάση (Positive), άρα **Class P=Yes (No Show).** Αντίθετα οι ασθενείς που θα εμφανιστούν στο ραντεβού θεωρούνται ως η αρνητική κλάση, άρα **Class N=No (No Show)**.

Αφού εφαρμοστεί κάποιος αλγόριθμος κατηγοριοποίησης παράγεται ο παρακάτω πίνακας σύγχυσης (Confusion Matrix) (Πίνακας 44).

		ΠΡΟΒΛΕΨΗ		
		ΘΕΤΙΚΗ ΚΛΑΣΗ Yes (No Show)	**ΑΡΝΗΤΙΚΗ ΚΛΑΣΗ** No (No Show)	**ΣΥΝΟΛΟ**
ΓΕΓΟΝΟΣ	**ΘΕΤΙΚΗ ΚΛΑΣΗ** Yes (No Show)	**TP** 100	**FN** 30	**130**
	ΑΡΝΗΤΙΚΗ ΚΛΑΣΗ No (No Show)	**FP** 300	**TN** 80	**380**
	ΣΥΝΟΛΟ	**400**	**110**	**510**

Πίνακας 44: Πίνακας Σύγχυσης (Confusion Matrix)

Στη συνέχεια περιγράφονται τα περιεχόμενα του παραπάνω πίνακα σύγχυσης:

Ο συνολικός αριθμός ασθενών είναι **510 (TP+FN+FP+TN)**.

TP = 100: Ο κατηγοριοποιητής προέβλεψε ότι 100 ασθενείς δεν θα εμφανιστούν στο ραντεβού και όντως αυτοί δεν εμφανίστηκαν, άρα η **πρόβλεψη ήταν σωστή.**

FP = 300: Ο κατηγοριοποιητής προέβλεψε ότι 300 ασθενείς δεν θα εμφανιστούν στο ραντεβού, αλλά τελικά στην πραγματικότητα οι ασθενείς αυτοί εμφανίστηκαν, άρα η **πρόβλεψη ήταν λανθασμένη**.

FN = 30: Ο κατηγοριοποιητής προέβλεψε, ότι 30 ασθενείς θα εμφανιστούν στο ραντεβού, αλλά τελικά στην πραγματικότητα οι ασθενείς αυτοί δεν εμφανίστηκαν, άρα η **πρόβλεψη ήταν λανθασμένη**.

TN = 80: Ο κατηγοριοποιητής προέβλεψε, ότι 80 ασθενείς θα εμφανιστούν στο ραντεβού και όντως, στην πραγματικότητα οι ασθενείς αυτοί εμφανίστηκαν, άρα η **πρόβλεψη ήταν σωστή**.

Καταρχάς θα υπολογιστεί η μετρική Accuracy για τον συγκεκριμένο κατηγοριοποιητή:

$$Accuracy = \frac{Πλήθος\ σωστών\ προβλέψεων}{Συνολικό\ πλήθος\ προβλέψεων}$$

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

$$\frac{100 + 80}{100 + 80 + 300 + 30} = \frac{180}{510} = 0{,}352941 \text{ ή } (0{,}352941 * 100) = 35{,}2941\%$$

Άρα από τις προβλέψεις που έκανε ο κατηγοριοποιητής και για τις δύο κλάσεις, το 35,2941% ήταν σωστές.

Έπειτα θα υπολογιστεί το Precision για την **ΘΕΤΙΚΗ ΚΛΑΣΗ (No Show = Yes).** Το Precision (Yes) είναι ο λόγος των σωστών προβλέψεων για τους ασθενείς που δεν θα εμφανιστούν στο ραντεβού προς το συνολικό πλήθος προβλέψεων για τους ασθενείς που δεν θα εμφανιστούν στο ραντεβού. Δηλαδή, από τους ασθενείς που ο κατηγοριοποιητής προέβλεψε ότι δεν θα εμφανιστούν στο ραντεβού, τι ποσοστό από αυτούς δεν εμφανίστηκε τελικά στο ραντεβού.

$$Precision(Yes) = \frac{TP}{TP + FP} \frac{(οι\ σωστές\ προβλέψεις\ για\ την\ κλάση\ Yes)}{(όλες\ οι\ προβλέψεις\ για\ την\ κλάση\ Yes)} =$$

$$\frac{100}{100 + 300} = \frac{100}{400} = 0{,}25 \text{ ή } (0{,}25 * 100) = 25\%$$

Ο κατηγοριοποιητής προέβλεψε ότι συνολικά δεν θα εμφανισθούν στο ραντεβού 400 ασθενείς. Από τους 400 ασθενείς προέβλεψε σωστά ότι οι 100 ασθενείς ή διαφορετικά το 25% δεν θα εμφανισθούν στο ραντεβού.

Αντίθετα προέβλεψε λανθασμένα ότι οι 300 ασθενείς ή διαφορετικά το 75% δεν θα εμφανισθούν στο ραντεβού, ενώ αυτοί τελικά εμφανίσθηκαν.

Αντίστοιχα θα υπολογιστεί και το Precision για την **ΑΡΝΗΤΙΚΗ ΚΛΑΣΗ (No Show = No).** Το Precision (No) είναι ο λόγος των σωστών προβλέψεων για τους ασθενείς που θα εμφανιστούν στο ραντεβού προς το συνολικό πλήθος προβλέψεων για τους ασθενείς που θα

εμφανιστούν στο ραντεβού. Δηλαδή, από τους ασθενείς που ο κατηγοριοποιητής προέβλεψε ότι θα εμφανιστούν στο ραντεβού, τι ποσοστό από αυτούς εμφανίστηκε τελικά στο ραντεβού.

$$Precision(No) = \frac{TN}{TN + FN}\ \frac{(οι\ σωστές\ προβλέψεις\ για\ την\ κλάση\ No)}{(όλες\ οι\ προβλέψεις\ για\ την\ κλάση\ No)} =$$

$$\frac{80}{80 + 30} = \frac{80}{110} = 0{,}727273 \text{ ή } (0{,}727273 * 100) = 72{,}7273\%$$

Ο κατηγοριοποιητής προέβλεψε ότι συνολικά θα εμφανισθούν στο ραντεβού 110 ασθενείς. Από τους 110 ασθενείς προέβλεψε σωστά ότι οι 80 ασθενείς ή διαφορετικά το 72,7273% θα εμφανισθούν στο ραντεβού.

Αντίθετα προέβλεψε λανθασμένα ότι οι 30 ασθενείς ή διαφορετικά το 27,2727% θα εμφανισθούν στο ραντεβού, ενώ αυτοί τελικά δεν εμφανίσθηκαν.

Έπειτα θα υπολογιστεί το Recall για την **ΘΕΤΙΚΗ ΚΛΑΣΗ (No Show = Yes).** Το Recall (Yes) είναι ο λόγος των ασθενών που προέβλεψε σωστά ο κατηγοριοποιητής ότι δεν θα εμφανιστούν στο ραντεβού προς το συνολικό αριθμό ασθενών που πραγματικά δεν εμφανίστηκαν. Δηλαδή από τους ασθενείς που πραγματικά δεν εμφανίστηκαν στο ραντεβού τι ποσοστό από αυτούς προέβλεψε σωστά ο κατηγοριοποιητής ότι δεν θα εμφανιστούν στο ραντεβού.

$$Recall(Yes) = \frac{TP}{TP + FN}\ \frac{(οι\ σωστές\ προβλέψεις\ για\ την\ κλάση\ Yes)}{(Συνολικές\ πραγματικές\ τιμές\ για\ την\ κλάση\ Yes)} =$$

$$\frac{100}{100 + 30} = \frac{100}{130} = 0{,}769231 \text{ ή } (0{,}769231 * 100) = 76{,}9231\%$$

Άρα συνολικά 130 είναι οι ασθενείς που στην πραγματικότητα δεν εμφανίσθηκαν στο ραντεβού. Ο κατηγοριοποιητής προέβλεψε σωστά ότι οι 100 από αυτούς ή διαφορετικά το 76,9231% δεν θα εμφανισθούν στο ραντεβού και όντως δεν εμφανίσθηκαν.

Αντίθετα προέβλεψε λανθασμένα ότι οι υπόλοιποι 30 ασθενείς ή διαφορετικά το 23,0769% θα εμφανιστούν στο ραντεβού, ενώ στην πραγματικότητα δεν εμφανίσθηκαν.

Τέλος θα υπολογιστεί το Recall για την **ΑΡΝΗΤΙΚΗ ΚΛΑΣΗ (No Show = No).** Το Recall (No) είναι ο λόγος των ασθενών που προέβλεψε σωστά ο κατηγοριοποιητής ότι θα εμφανιστούν στο ραντεβού προς το συνολικό αριθμό ασθενών που πραγματικά εμφανίστηκαν. Δηλαδή από τους ασθενείς που πραγματικά εμφανίστηκαν στο ραντεβού τι ποσοστό από αυτούς προέβλεψε σωστά ο κατηγοριοποιητής ότι θα εμφανιστούν στο ραντεβού.

$$Recall(No) = \frac{TN}{TN + FP}\ \frac{(οι\ σωστές\ προβλέψεις\ για\ την\ κλάση\ No)}{(Συνολικές\ πραγματικές\ τιμές\ για\ την\ κλάση\ No)} =$$

$$\frac{80}{80+300} = \frac{80}{380} = 0{,}210526 \text{ ή } (0{,}210526 * 100) = 21{,}0526\%$$

Τέλος θα υπολογιστεί το Recall για την **ΑΡΝΗΤΙΚΗ ΚΛΑΣΗ (No Show = No).** Το Recall (No) είναι οι σωστές προβλέψεις που έκανε ο κατηγοριοποιητής για ασθενείς που θα εμφανισθούν στο ραντεβού προς το συνολικό αριθμό ασθενών που στην πραγματικότητα εμφανίσθηκαν στο ραντεβού.

Οι 380 είναι οι ασθενείς που στην πραγματικότητα εμφανίσθηκαν στο ραντεβού. Ο κατηγοριοποιητής προέβλεψε σωστά ότι από αυτούς θα εμφανισθούν στο ραντεβού οι 80 (TN) ασθενείς ή διαφορετικά το 21,0526%. Αντίθετα για τους υπόλοιπους 300 (FP) ασθενείς ή 78,9474% προέβλεψε λανθασμένα ότι δεν θα εμφανισθούν στο ραντεβού, ενώ στην πραγματικότητα εμφανίσθηκαν.

3.3 Διαγράμματα Venn

3.3.1 Τι είναι διαγράμματα Venn

Το διάγραμμα Venn είναι μια διαγραμματική τεχνική που απεικονίζει λογικές σχέσεις ανάμεσα σε κάποια σύνολα στοιχείων. Το διάγραμμα αποτελείται από κύκλους. Συνήθως χρησιμοποιούνται δύο κύκλοι που αλληλεπικαλύπτονται και η κοινή περιοχή περιλαμβάνει όσα στοιχεία των συνόλων είναι κοινά. Τα διαγράμματα Venn δημιουργήθηκαν το 1880 από τον John Venn (Venn, 1880).

3.3.2 Ερμηνεία Precision / Recall με Διαγράμματα Venn

Η επίδοση ενός κατηγοριοποιητή στον διαχωρισμό δεδομένων σε δύο κλάσεις αξιολογείται με την βοήθεια μετρικών όπως το Precision και το Recall. Οι μετρικές αυτές δεν είναι ανεξάρτητες μεταξύ τους, καθώς και οι δύο προκύπτουν από κάποιο συνδυασμό των τιμών TP, TN, FP, FN. Έτσι για κάθε κλάση μπορούν να διακριθούν κάποιες γενικές περιπτώσεις επίδοσης ανάλογα με το αν ο κατηγοριοποιητής παρουσιάζει χαμηλό ή υψηλό precision και χαμηλό ή υψηλό recall, καθώς και τους επιμέρους συνδυασμούς. Με το σκεπτικό αυτό προκύπτουν τέσσερις περιπτώσεις: χαμηλό Precision και χαμηλό Recall, υψηλό Precision και υψηλό Recall, χαμηλό Precision και υψηλό Recall και τέλος υψηλό Precision και χαμηλό Recall.

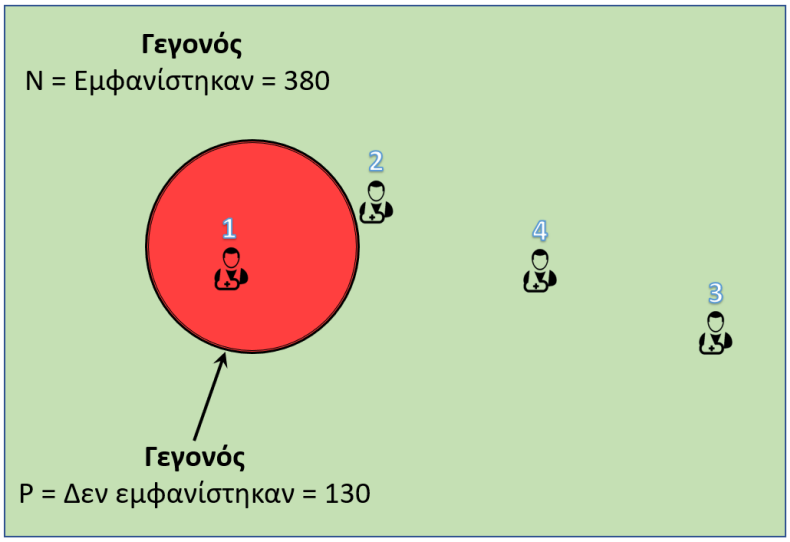
Προκειμένου να ερμηνευτεί το αποτέλεσμα ενός κατηγοριοποιητή για κάθε περίπτωση μπορούν να χρησιμοποιηθούν διαγράμματα Venn, στα οποία να φαίνεται η σχέση μεταξύ των TP, TN, FP, FN και κατά συνέπεια και η συσχέτιση μεταξύ Precision και Recall. Στη

συνέχεια μάλιστα γίνεται μία προσπάθεια ερμηνείας αποτελεσμάτων με την βοήθεια τέτοιων διαγραμμάτων. Για το σκοπό αυτό θα χρησιμοποιηθεί το ίδιο παράδειγμα με πιο πάνω. Δηλαδή με ασθενείς που εμφανίζονται ή όχι σε ραντεβού. Στο παράδειγμα λοιπόν υπάρχουν 2 κλάσεις. Οι ασθενείς που δεν θα εμφανιστούν στο ραντεβού και θεωρούνται, ως η θετική κλάση (Positive), άρα no-show = Yes και οι ασθενείς που θα εμφανιστούν στο ραντεβού και θεωρούνται, ως η αρνητική κλάση, άρα no-show = No.

Σε όλα τα διαγράμματα η ορθογώνια περιοχή συμβολίζει όλους τους ασθενείς που περιλαμβάνονται στο σύνολο δεδομένων, οι οποίοι υποχρεωτικά ανήκουν σε κάποια από τις κλάσεις που περιεγράφηκαν προηγουμένως. Τα 3 πρώτα διαγράμματα βασίζονται στο παράδειγμα της προηγούμενης παραγράφου και τα υπόλοιπα διαφοροποιούνται κατά περίπτωση.

Η Εικόνα 6 (Διάγραμμα 1) είναι το διάγραμμα Venn για τα πραγματικά δεδομένα (ground truth). Η κλάση P θεωρείται ότι είναι το σύνολο που περικλείεται στον κόκκινο κύκλο με συνεχή γραμμή και περιλαμβάνει τους 130 ασθενείς που δεν εμφανίστηκαν σε ραντεβού. Η κλάση N είναι το σύνολο που είναι εκτός του κόκκινου κύκλου, δηλαδή η πράσινη περιοχή και περιλαμβάνει τους 380 ασθενείς που εμφανίστηκαν σε ραντεβού. Το σύνολο όλων των ασθενών του παραδείγματος είναι το άθροισμα της πράσινης και κόκκινης περιοχής, δηλαδή 380+130=510.

Ενδεικτικά στην Εικόνα 6 (Διάγραμμα 1) απεικονίζονται τέσσερις ασθενείς από το σύνολο δεδομένων. Ο ασθενής No.1 ανήκει στην κλάση P (ασθενείς που δεν εμφανίστηκαν σε ραντεβού), ενώ οι ασθενείς No.2, No.4 και No.3 ανήκουν στην κλάση N (ασθενείς που εμφανίστηκαν σε ραντεβού).



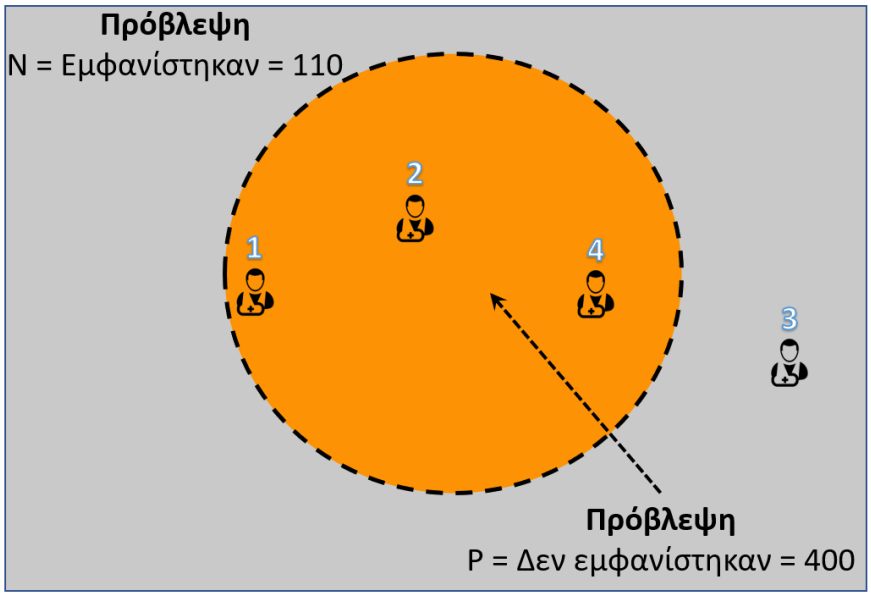
Διάγραμμα 1
Πραγματικά δεδομένα

Εικόνα 6: Διάγραμμα Venn - Σύνολο Δεδομένων (dataset) με πραγματικά δεδομένα

Αν κάποιος κατηγοριοποιητής εφαρμοστεί στα δεδομένα της Εικόνα 6 (Διάγραμμα 1), θα δώσει ως αποτέλεσμα κάποιες προβλέψεις. Οι προβλέψεις του κατηγοριοποιητή

απεικονίζονται στην Εικόνα 7 (Διάγραμμα 2). Η κλάση P μπορεί να θεωρηθεί ότι είναι το σύνολο των προβλέψεων που περικλείεται στον πορτοκαλί κύκλο με διακεκομμένη γραμμή και περιλαμβάνει τους 400 ασθενείς που ο κατηγοριοποιητής προέβλεψε, ότι δεν θα εμφανιστούν σε ραντεβού. Η κλάση N είναι το σύνολο των προβλέψεων που βρίσκεται εκτός του πορτοκαλί κύκλου, δηλαδή η γκρι περιοχή και περιλαμβάνει τους 110 ασθενείς που ο κατηγοριοποιητής προέβλεψε, ότι θα εμφανιστούν σε ραντεβού.

Στη συγκεκριμένη εικόνα ο κατηγοριοποιητής ταξινόμησε τους ενδεικτικούς ασθενείς No.1, No.2 και No.4 στην κλάση P και τον ασθενή No.3 στην κλάση N.

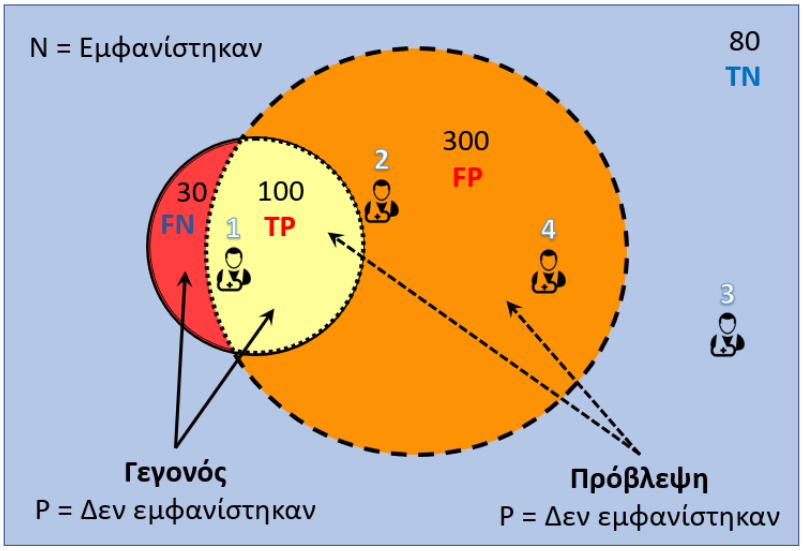


Διάγραμμα 2
Προβλέψεις

Εικόνα 7: Διάγραμμα Venn - Πρόβλεψη Κατηγοριοποιητή με βάση τα πραγματικά δεδομένα

Για να αξιολογηθεί η απόδοση του κατηγοριοποιητή θα πρέπει οι προβλέψεις που έκανε, οι οποίες φαίνονται στην Εικόνα 7 (Διάγραμμα 2), να συγκριθούν με τα πραγματικά δεδομένα που φαίνονται στην Εικόνα 6 (Διάγραμμα 1). Στη συνέχεια απεικονίζονται διαγραμματικά οι τέσσερις περιπτώσεις συνδυασμών precision και recall, καθώς και ο τρόπος που οι ενδεικτικοί ασθενείς κατηγοριοποιούνται σε κάθε μία από αυτές.

3.3.2.1 Περίπτωση 1: Χαμηλό Precision και υψηλό Recall



Διάγραμμα 3
Χαμηλό Precision
Υψηλό Recall

TP = 100
TN = 80
FP = 300
FN = 30

Κλάση P = Δεν εμφανίστηκαν
Precision = TP/(TP+FP) = 0,25
Recall = TP/(TP+FN) =0,77

Γεγονός P = Δεν Εμφανίστηκαν Κόκκινο + Κίτρινο = Κόκκινο (του 1ου διαγράμματος)	**Πρόβλεψη** P = Δεν Εμφανίστηκαν Πορτοκαλί + Κίτρινο = Πορτοκαλί (του 2ου διαγράμματος)
Γεγονός N = Εμφανίστηκαν Γαλάζιο + Πορτοκαλί = Πράσινο (του 1ου διαγράμματος)	**Πρόβλεψη** N = Εμφανίστηκαν Γαλάζιο + Κόκκινο = Γκρι (του 2ου διαγράμματος)

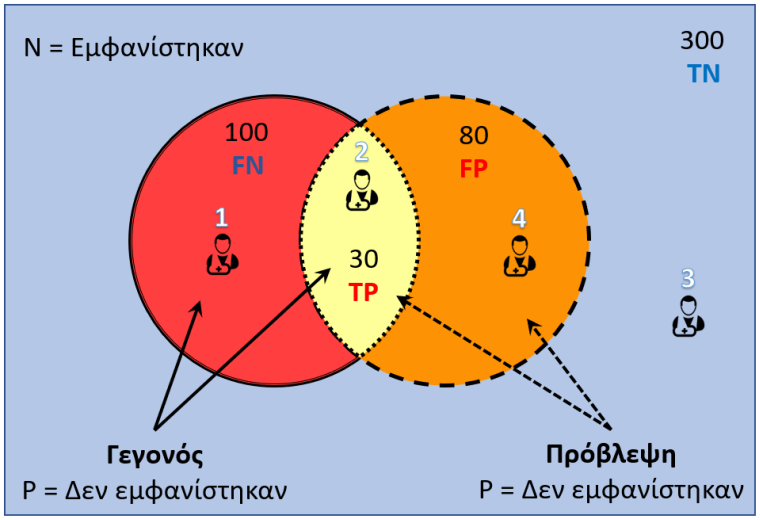
Εικόνα 8: Διάγραμμα Venn 1ης Περίπτωσης - Χαμηλό Precision / Υψηλό Recall

Στην Εικόνα 8 (Διάγραμμα 3) ο κατηγοριοποιητής παρουσιάζει χαμηλό precision=0,25, γιατί κάνει πολλά λάθη στην πρόβλεψη της θετικής κλάσης, δηλαδή FP > TP. Κατά συνέπεια ήταν σωστές μόνο το 25% από τις προβλέψεις που έκανε για ασθενείς που δεν θα εμφανισθούν στο ραντεβού.

Αντίθετα προβλέπει σωστά το μεγαλύτερο μέρος από τους ασθενείς που πραγματικά δεν εμφανίστηκαν στο ραντεβού, δηλαδή TP > FN, οπότε παρουσιάζει υψηλό Recall=0,769231. Αυτό σημαίνει ότι προέβλεψε σωστά το 76,9231% από τους ασθενείς που στην πραγματικότητα δεν εμφανίσθηκαν στο ραντεβού.

Στην συγκεκριμένη περίπτωση από τους ενδεικτικούς ασθενείς ο No.1 (TP) κατηγοριοποιείται σωστά, ενώ οι υπόλοιποι λανθασμένα (οι No.2 και No.4 είναι (FP), ενώ ο No.3 είναι (TN).

3.3.2.2 Περίπτωση 2: Χαμηλό Precision και χαμηλό Recall



Διάγραμμα 4
Χαμηλό Precision
Χαμηλό Recall

TP = 30
TN = 300
FP = 80
FN = 100

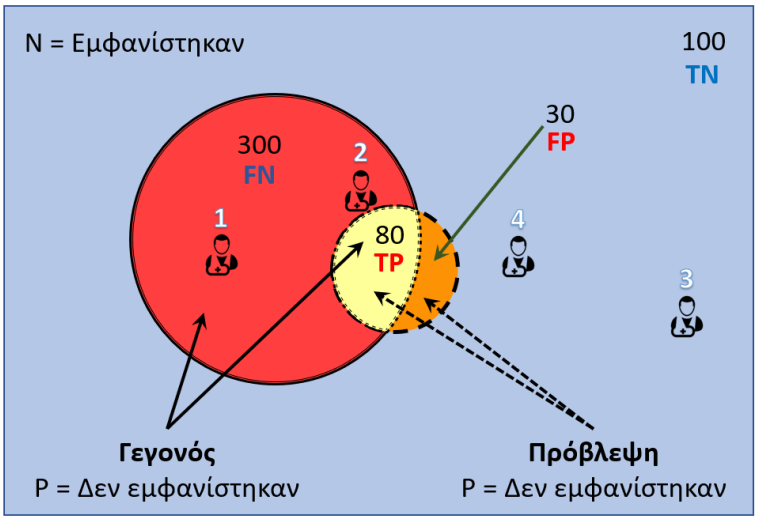
Κλάση P = Δεν εμφανίστηκαν
Precision = TP/(TP+FP) = 0,27
Recall = TP/(TP+FN) =0,23

Εικόνα 9: Διάγραμμα Venn 2ης Περίπτωσης - Χαμηλό Precision / Χαμηλό Recall

Όπως φαίνεται στην Εικόνα 9 (Διάγραμμα 4) ο κατηγοριοποιητής παρουσιάζει χαμηλό precision=0,272727 γιατί κάνει αρκετά λάθη ως προς την πρόβλεψη των ασθενών που δεν θα εμφανιστούν σε ραντεβού, δηλαδή FP > TP. Κατά συνέπεια ήταν σωστές το 27,2727% από τις προβλέψεις που έκανε για ασθενείς που δεν θα εμφανισθούν στο ραντεβού. Ταυτόχρονα προβλέπει σωστά μικρό αριθμό ασθενών που πραγματικά δεν εμφανίστηκαν σε ραντεβού, δηλαδή TP < FN, οπότε παρουσιάζει χαμηλό recall= 0,230769. Αυτό σημαίνει ότι προέβλεψε σωστά μόνο το 23,0769% από τους ασθενείς που στην πραγματικότητα δεν εμφανίσθηκαν στο ραντεβού.

Στην Εικόνα 9 (Διάγραμμα 4) φαίνεται επίσης ότι κατηγοριοποιούνται σωστά οι ασθενείς No.2 (TP) και No.3 (TN). Οι ασθενείς No.1 (FN) και No.4 (FP) κατηγοριοποιούνται λανθασμένα. Σε αυτήν την περίπτωση παρατηρείται ότι η πρόβλεψη για τους ασθενείς No.1 και No.2 είναι διαφορετική από την προηγούμενη περίπτωση Εικόνα 8 (Διάγραμμα 3), ενώ για τους ασθενείς No.3 (TN) και No.4 (FP) είναι ίδια. Ο ασθενής No.1 είχε κατηγοριοποιηθεί σωστά ως TP, ενώ τώρα κατηγοριοποιείται λανθασμένα ως FN Εικόνα 9 (Διάγραμμα 4). Αντίθετα ο ασθενής No.2 ενώ στην Εικόνα 8 (Διάγραμμα 3) είχε κατηγοριοποιηθεί λανθασμένα ως FP, παρατηρείται ότι στην Εικόνα 9 (Διάγραμμα 4) κατηγοριοποιείται σωστά ως TP.

3.3.2.3 Περίπτωση 3: Υψηλό Precision και χαμηλό Recall



Διάγραμμα 5
Υψηλό Precision
Χαμηλό Recall

TP = 80
TN = 100
FP = 30
FN = 300

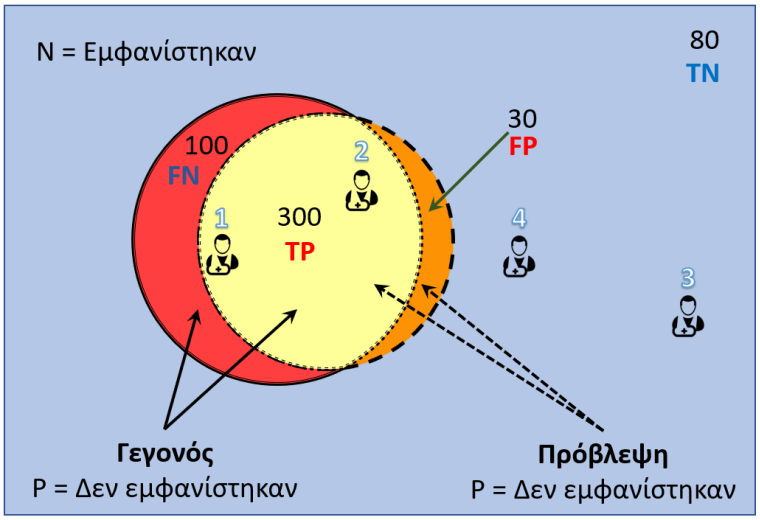
Κλάση P = Δεν εμφανίστηκαν
Precision = TP/(TP+FP) = 0,73
Recall = TP/(TP+FN) =0,21

Εικόνα 10: Διάγραμμα Venn 3ης Περίπτωσης - Υψηλό Precision / Χαμηλό Recall

Στην Εικόνα 10 (Διάγραμμα 5) ο κατηγοριοποιητής παρουσιάζει υψηλό precision=0,727273, γιατί από τις προβλέψεις που έκανε για ασθενείς που δεν θα εμφανιστούν σε ραντεβού οι περισσότερες ήταν σωστές, δηλαδή TP > FP. Κατά συνέπεια ήταν σωστές το 72,7273% από τις προβλέψεις που έκανε για ασθενείς που δεν θα εμφανισθούν στο ραντεβού. Αντίθετα προέβλεψε μικρό μέρος από τους ασθενείς που πραγματικά δεν εμφανίστηκαν, δηλαδή TP < FN, οπότε παρουσιάζει χαμηλό recall=0,210526. Αυτό σημαίνει ότι προέβλεψε σωστά μόνο το 21,0526% από τους ασθενείς που στην πραγματικότητα δεν εμφανίσθηκαν στο ραντεβού.

Στην Εικόνα 10 (Διάγραμμα 5) παρατηρείται ότι οι ασθενείς No.1 (FN) και No.2 (FN) κατηγοριοποιούνται λανθασμένα, ενώ οι ασθενείς No.4 και No.3 σωστά (TN). Συγκρίνοντας αυτή την περίπτωση με την περίπτωση που φαίνεται στην Εικόνα 9 (Διάγραμμα 4) παρατηρεί κανείς ότι η πρόβλεψη για τον ασθενή No.1 (FN) και No.3 (TN) παραμένει ίδια. Αντίθετα για τους υπόλοιπους ασθενείς είναι διαφορετική Εικόνα 9 και Εικόνα 10 (Διαγράμματα 4 και 5). Δηλαδή ο ασθενής No.2 ενώ στην Εικόνα 9 (Διάγραμμα 4) έχει κατηγοριοποιηθεί σωστά ως TP, στην Εικόνα 10 (Διάγραμμα 5) κατηγοριοποιείται λανθασμένα ως FN. Από την άλλη πλευρά ο ασθενής No.4 ενώ στην Εικόνα 9 (Διάγραμμα 4) έχει κατηγοριοποιηθεί λανθασμένα ως FP, στην Εικόνα 10 (Διάγραμμα 5) παρατηρείται ότι κατηγοριοποιείται σωστά, δηλαδή ως TN.

3.3.2.4 Περίπτωση 4: Υψηλό Precision και υψηλό Recall



Διάγραμμα 6
Υψηλό Precision
Υψηλό Recall

TP = 300
TN = 80
FP = 30
FN = 100

Κλάση P = Δεν εμφανίστηκαν
Precision = TP/(TP+FP) = 0,91
Recall = TP/(TP+FN) =0,75

Εικόνα 11: Διάγραμμα Venn - 4ης Περίπτωσης Υψηλό Precision / Υψηλό Recall

Στην Εικόνα 11 (Διάγραμμα 6) είναι η περίπτωση κατά την οποία φαίνεται ότι οι δύο κύκλοι συγκλίνουν. Ο κατηγοριοποιητής παρουσιάζει υψηλό precision=0,909091, γιατί από τις προβλέψεις που έκανε για ασθενείς που δεν θα εμφανιστούν σε ραντεβού οι περισσότερες ήταν σωστές, δηλαδή TP > FP. Κατά συνέπεια ήταν σωστές το 90,9091% από τις προβλέψεις που έκανε για ασθενείς που δεν θα εμφανισθούν στο ραντεβού. Ταυτόχρονα προέβλεψε σωστά το μεγαλύτερο αριθμό των ασθενών που πραγματικά δεν εμφανίστηκαν, δηλαδή TP > FN, οπότε παρουσιάζει υψηλό recall=0,75. Αυτό σημαίνει ότι προέβλεψε σωστά το 75% από τους ασθενείς που στην πραγματικότητα δεν εμφανίσθηκαν στο ραντεβού.

Από την Εικόνα 11 (Διάγραμμα 6) προκύπτει επίσης ότι όλοι οι ασθενείς κατηγοριοποιούνται σωστά. Σε σχέση με την Εικόνα 10 (Διάγραμμα 5) οι ασθενείς No.1 και No.2 κατηγοριοποιούνται διαφορετικά. Δηλαδή και ο ασθενής No.1 και ο ασθενής No.2, ενώ στην Εικόνα 10 (Διάγραμμα 5) κατηγοριοποιούνται λανθασμένα ως FN, αντίθετα στην Εικόνα 11 (Διάγραμμα 6) παρατηρείται ότι κατηγοριοποιούνται σωστά ως TP. Όσον αφορά τους ασθενείς No.4 και No.3 παρατηρείται ότι και στην Εικόνα 10 (Διάγραμμα 5) αλλά και στην Εικόνα 11 (Διάγραμμα 6) κατηγοριοποιούνται σωστά ως TN.

3.4 Μέθοδοι εκτίμησης της επίδοσης

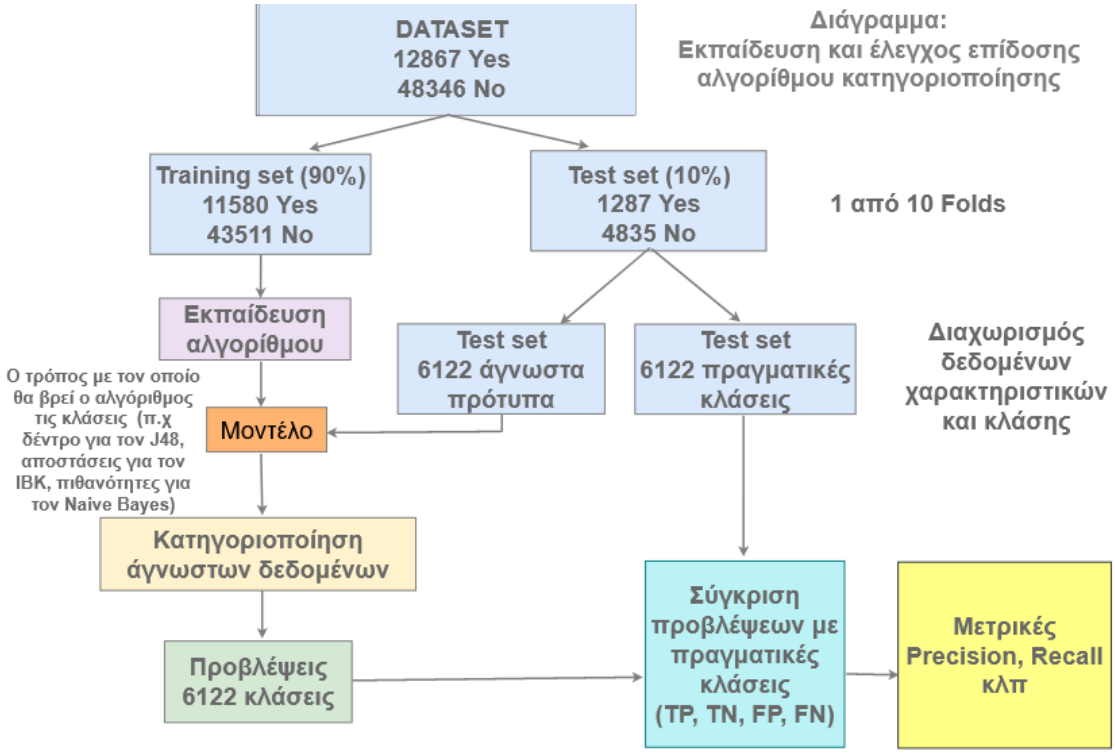
3.4.1 Διαχωρισμός συνόλου δεδομένων σε σύνολο εκπαίδευσης και σύνολο ελέγχου

Για να εκτιμηθεί η δυνατότητα γενίκευσης ενός μοντέλου κατηγοριοποιητή πρέπει να δοκιμαστεί η επίδοση του στην κατηγοριοποίηση άγνωστων δεδομένων. Επειδή η εύρεση νέων άγνωστων δεδομένων συνήθως δεν είναι εύκολη, για την εκτέλεση πειραμάτων χρησιμοποιείται η τεχνική διαχωρισμού του διαθέσιμου συνόλου δεδομένων σε σύνολα εκπαίδευσης (training set) και ελέγχου (test set). Ο διαχωρισμός γίνεται με τυχαίο τρόπο και το σύνολο εκπαίδευσης χρησιμοποιείται αποκλειστικά για την εκπαίδευση του μοντέλου, ενώ το σύνολο ελέγχου για τον έλεγχο της επίδοσης του με βάση κάποια μετρική επίδοσης. Στην προκειμένη περίπτωση το σύνολο ελέγχου αποτελείται από δεδομένα τα οποία είναι άγνωστα για τον κατηγοριοποιητή διότι είναι διαφορετικά από αυτά πάνω στα οποία εκπαιδεύτηκε.

Αρχικά λοιπόν ας θεωρηθεί ότι είναι διαθέσιμο ένα σύνολο δεδομένων πάνω στο οποίο θα εφαρμοστούν κάποιοι αλγόριθμοι κατηγοριοποίησης. Η διαδικασία αυτή έχει ως σκοπό την σύγκριση της επίδοσης των συγκεκριμένων αλγορίθμων ανάλογα με τις τιμές των παραμέτρων που δίνονται σε κάθε έναν από αυτούς.

Σε κάθε πείραμα το σύνολο δεδομένων διαχωρίζεται με τυχαίο τρόπο σε σύνολο εκπαίδευσης και σύνολο ελέγχου. Σε ένα τυπικό διαχωρισμό το σύνολο εκπαίδευσης περιέχει συνήθως το 90% των αρχικών προτύπων και το σύνολο ελέγχου περιέχει το υπόλοιπο 10%. Το σύνολο εκπαίδευσης χρησιμοποιείται για την εκπαίδευση του αλγορίθμου κατηγοριοποίησης και παράγεται το μοντέλο πρόβλεψης του κατηγοριοποιητή (Kohavi, 1995).

Μετά την δημιουργία του μοντέλου πρέπει να ελεγχθεί η επίδοση του σε άγνωστα δεδομένα. Για αυτό το σκοπό χρησιμοποιείται το σύνολο ελέγχου. Αρχικά λοιπόν δίνεται στο μοντέλο το σύνολο ελέγχου χωρίς τα δεδομένα της κλάσης στην οποία ανήκουν τα πρότυπα του και το μοντέλο προβλέπει την κλάση στην οποία θα τα κατηγοριοποιήσει. Στη συνέχεια για να εκτιμηθεί η επίδοση του μοντέλου συγκρίνονται οι κλάσεις που προέβλεψε το μοντέλο με τα πραγματικά δεδομένα των κλάσεων που περιέχονται στο σύνολο ελέγχου. Με βάση αυτή τη σύγκριση υπολογίζονται οι τιμές των True Positives (TP), False Positives (FP), True Negatives (TN) και False Negatives (FN). Στη συνέχεια χρησιμοποιώντας αυτές τις τιμές υπολογίζονται και οι υπόλοιπες μετρικές επίδοσης π.χ. Precision, Recall, Accuracy κλπ. Η παραπάνω διαδικασία παρουσιάζεται συνοπτικά στην Εικόνα 12.

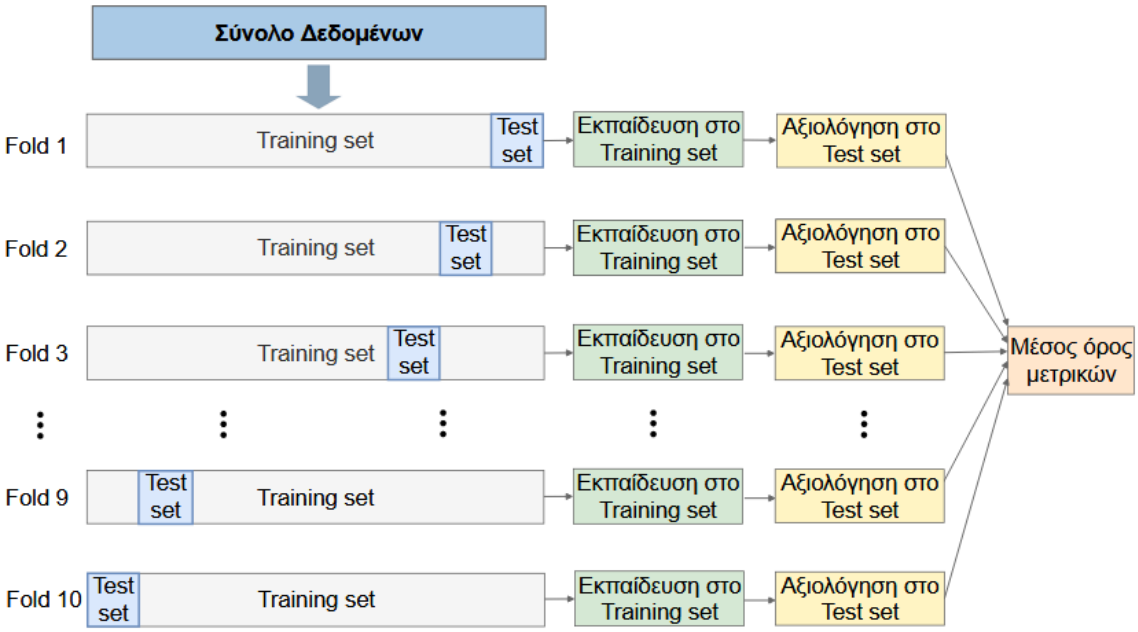


Εικόνα 12: Διαδικασία εκπαίδευσης και ελέγχου επίδοσης αλγορίθμου

Ένας συγκεκριμένος τυχαίος διαχωρισμός ενδέχεται να μην δώσει ασφαλή αποτελέσματα, ιδιαίτερα σε περιπτώσεις που το σύνολο δεδομένων δεν είναι μεγάλο. Για το λόγο αυτό ενδείκνυται η εκτέλεση πολλών πειραμάτων με διαφορετικό διαχωρισμό κάθε φορά και στο τέλος να υπολογίζεται ο μέσος όρος των μετρικών επίδοσης. Η συνήθης πρακτική είναι η χρήση της μεθόδου k-fold cross-validation όπου εκτελούνται k folds, δηλαδή k πειράματα αλλάζοντας κάθε φορά τα σύνολα εκπαίδευσης και ελέγχου και η τελική τιμή επίδοσης είναι η μέση τιμή των μετρικών των k folds. Στην τεχνική αυτή ο διαχωρισμός γίνεται με τέτοιο τρόπο ώστε κάθε πρότυπο να συμμετάσχει μόνο μια φορά σε σύνολο ελέγχου και τις υπόλοιπες σε σύνολο εκπαίδευσης.

Για να γίνει αυτό το αρχικό σύνολο δεδομένων χωρίζεται τυχαία σε k μη επικαλυπτόμενα τμήματα. Στη συνέχεια ένα από τα k τμήματα χρησιμοποιείται ως σύνολο ελέγχου και τα υπόλοιπα k-1 τμήματα ως σύνολο εκπαίδευσης. Αυτός ο διαχωρισμός ονομάζεται fold 1. Ο κατηγοριοποιητής εκπαιδεύεται στο σύνολο εκπαίδευσης και η επίδοση του αξιολογείται στο αντίστοιχο σύνολο ελέγχου του fold 1. Στη συνέχεια δημιουργείται το fold 2, επιλέγοντας ένα άλλο τμήμα από τα k ως σύνολο ελέγχου και τα υπόλοιπα k-1 ως σύνολο εκπαίδευσης. Ο κατηγοριοποιητής εκπαιδεύεται εκ νέου και αξιολογείται στο fold 2. Η παραπάνω διαδικασία επαναλαμβάνεται μέχρι τη δημιουργία του fold k, δηλαδή συνολικά k φορές εναλλάσσοντας τα τμήματα που συμμετέχουν κάθε φορά στα σύνολα εκπαίδευσης και ελέγχου. Η τελική επίδοση του κατηγοριοποιητή προκύπτει από τον μέσο όρο των μετρήσεων αξιολόγησης που

προήλθαν από κάθε fold. Τυπικές τιμές του k είναι 5 και 10 αλλά όσο μεγαλύτερο είναι το k τόσο ασφαλέστερη είναι και η συνολική εκτίμηση του κατηγοριοποιητή. Η διαδικασία k-fold cross validation για k = 10 απεικονίζεται στην Εικόνα 13 (Kohavi, 1995).



Εικόνα 13: Μέθοδος επικυρωμένης διασταύρωσης (Cross Validation)

4

Το πρόβλημα της μη εμφάνισης ασθενών στα ραντεβού

4.1 Περιγραφή του προβλήματος

Τα προβλήματα που απαντώνται σήμερα στα νοσοκομεία είναι πολλά και βαρύνουσας σημασίας. Ένα από αυτά είναι η μεγάλη λίστα αναμονής που δημιουργείται προκειμένου ένας ασθενής να κλείσει ραντεβού για εξέταση από κάποιον ιατρό. Η λίστα μεγαλώνει ακόμα περισσότερο λόγω του ότι πολλοί ασθενείς που έχουν προγραμματισμένο ραντεβού, είτε ξεχνούν την ημερομηνία που έχουν το ραντεβού και δεν προσέρχονται για εξέταση, είτε γιατί φοβούνται το αποτέλεσμα της εξέτασης ή για διάφορους άλλους λόγους. Για όποιον λόγο κι αν συμβαίνει αυτό, σημασία έχει ότι προκύπτουν προβλήματα που εμποδίζουν την εύρυθμη λειτουργία του νοσοκομείου και κατ' επέκταση την βέλτιστη ποιότητα παροχής υπηρεσιών προς τους πολίτες. Τα προβλήματα αυτά αναλύονται διεξοδικά παρακάτω.

4.2 Επιπτώσεις

Το φαινόμενο της μη εμφάνισης των ασθενών έχει σοβαρές επιπτώσεις τόσο στην ομαλή και εύρυθμη λειτουργία του νοσοκομείου, όσο και στους ίδιους τους ασθενείς. Μια σημαντική επίπτωση που παρατηρείται είναι ότι, όταν οι ασθενείς χάσουν το προγραμματισμένο

ραντεβού θα χρειαστεί να κλείσουν εκ νέου ραντεβού, με αποτέλεσμα να αυξηθεί η λίστα αναμονής για ασθενείς που θέλουν να κλείσουν ραντεβού για εξέταση από ιατρό. Σκεφτείτε όμως την περίπτωση ο ασθενής που δεν θα εξεταστεί, να πάσχει από μία σοβαρή ασθένεια και μάλιστα χωρίς να το γνωρίζει. Από τη στιγμή που ο συγκεκριμένος ασθενής θα χάσει το προγραμματισμένο του ραντεβού, δεν θα εξεταστεί από τον ιατρό, δεν θα του παρασχεθεί έγκαιρα διάγνωση και δεν θα πάρει την κατάλληλη φαρμακευτική αγωγή. Κατ' επέκταση είναι πολύ πιθανό να επιδεινωθεί η κατάσταση της υγείας του ή ακόμη και να εκτεθεί η ζωή του σε κίνδυνο.

Επιπρόσθετα παρατηρείται συχνά το φαινόμενο της ταυτόχρονης εμφάνισης των ασθενών στο ίδιο ραντεβού για εξέταση από τον ίδιο ιατρό. Δηλαδή είναι πολύ πιθανόν να προσέλθουν ταυτόχρονα την ίδια ώρα στο νοσοκομείο για εξέταση τόσο οι ασθενείς που έχουν κανονικά προγραμματισμένο ραντεβού όσο και οι ασθενείς που είχαν χάσει το προγραμματισμένο τους ραντεβού. Οι ασθενείς που είχαν χάσει το ραντεβού ζητούν συνήθως να εξεταστούν ως υπεράριθμοι ή στη θέση άλλων ασθενών. Αυτό πρακτικά σημαίνει δύο πράγματα. Το πρώτο είναι ότι ο ένας από τους δύο ασθενείς δεν θα προλάβει να εξεταστεί, οπότε, όπως ειπώθηκε προηγουμένως, θα χρειαστεί να κλείσει εκ νέου ραντεβού για εξέταση από τον ιατρό. Το δεύτερο πράγμα που θα συμβεί δεδομένης της κατάστασης είναι ο ιατρός να κάνει υπερωρία προκειμένου να εξεταστούν όλοι οι ασθενείς. Αυτό όμως θα είναι εις βάρος τόσο του ιατρού, όσο και των ασθενών. Ο μεν ιατρός θα αναγκαστεί να εξετάσει περισσότερους ασθενείς σε μία ημέρα, επομένως θα κουραστεί περισσότερο απ' ότι συνήθως. Κατά συνέπεια είναι πολύ πιθανό να μην έχει διαύγεια ώστε να πάρει κρίσιμες αποφάσεις και επομένως θα οδηγηθεί σε λανθασμένες διαγνώσεις. Κάτι τέτοιο όμως θα έχει άμεσο αντίκτυπο στους ασθενείς (ταλαιπωρία, πολύωρη αναμονή, αναποτελεσματικότητα, επικινδυνότητα). Έτσι λοιπόν διαταράσσεται η σχέση ασφάλειας και εμπιστοσύνης που θα έπρεπε να υπάρχει μεταξύ νοσοκομείου – ιατρού -ασθενή. Ο μεν ασθενής από την μια πλευρά εκλαμβάνει ως ψυχρή και αδιάφορη την στάση του νοσοκομείου ως προς την παροχή ικανοποιητικής ιατροφαρμακευτικής περίθαλψης. Ταυτόχρονα δεν νιώθει ικανοποιημένος από τον ιατρό, ούτε τον βλέπει ως αρωγό στο πρόβλημα του. Ο δε ιατρός αντιλαμβάνεται πόσο ανεπαρκής μπορεί να γίνει λόγω των συνθηκών, δεν νιώθει χρήσιμος και αποδοτικός παρά την υπερπροσπάθεια που καταβάλει. Επιπλέον κατακλύζεται από αισθήματα απογοήτευσης και απόγνωσης που αποτελούν τροχοπέδη ώστε να αποκτήσει κίνητρο για καλύτερη παροχή υπηρεσιών.

Τέλος στις περιπτώσεις όπου απαιτείται ο επαναπρογραμματισμός των ραντεβού όπως περιεγράφηκαν παραπάνω είναι πολύ πιθανό να δημιουργηθεί επιπρόσθετος φόρτος εργασίας για το διοικητικό προσωπικό. Η πιθανή ανάγκη απασχόλησης επιπλέον υπαλλήλου

για την εξυπηρέτηση των πολιτών σημαίνει αύξηση κόστους σε ανθρώπινο δυναμικό για το νοσοκομείο (Elvira et al., 2018).

4.3 Λύση προβλήματος μη εμφάνισης ασθενών στα ραντεβού

Σκοπός της παρούσας διπλωματικής είναι, η πρόβλεψη ασθενών που δεν θα εμφανιστούν σε προγραμματισμένα ραντεβού, ώστε στη συνέχεια να γίνουν κάποιες ενέργειες, για να αποφευχθεί αυτό το πρόβλημα.

4.3.1 1η Ενέργεια: Υπενθύμιση του ραντεβού

Μία πιθανή ενέργεια που θα μπορούσε να γίνει είναι η υπενθύμιση του ραντεβού στον ασθενή με κάποιο τρόπο, όπως για παράδειγμα μέσω e-mail, μέσω τηλεφωνικής επικοινωνίας ή μέσω SMS. Με αυτό τον τρόπο μπορεί να αυξάνεται ως ένα βαθμό το κόστος για το νοσοκομείο (λόγω αποστολής μηνυμάτων, ή τηλεφωνικής υπενθύμισης κ. τ. λ.). Το κόστος όμως αυτό δεν είναι τόσο σημαντικό σε σχέση με τα προβλήματα που προκύπτουν από τη μη εμφάνιση των ασθενών στα ραντεβού, όπως αυτά περιεγράφηκαν παραπάνω. Επιπλέον το κόστος μπορεί ακόμη και να μηδενιστεί στην περίπτωση που αντί για SMS ή τηλεφωνικής επικοινωνίας, η υπενθύμιση στους ασθενείς πραγματοποιείται μέσω e-mail. Το e-mail όμως δεν είναι τόσο αποτελεσματικός τρόπος υπενθύμισης όσο οι 2 προηγούμενοι καθώς οι ασθενείς μπορεί να μην το δουν. Επίσης στην περίπτωση του e-mail δεν είναι απαραίτητη η πρόβλεψη γιατί θα μπορούσε να στέλνεται σε όλους τους ασθενείς και μάλιστα χωρίς κόστος. Τα οφέλη που προκύπτουν από αυτή την ενέργεια είναι πολλά. Με την υπενθύμιση μειώνεται το φαινόμενο της μη προσέλευσης των ασθενών στα ραντεβού, οι οποίοι τείνουν να τα ξεχνούν. Κατά συνέπεια προάγεται η καλή οργάνωση και λειτουργία του νοσοκομείου, με κανονική ροή, χωρίς κενά ραντεβού, με την κατά το δυνατόν λιγότερη ταλαιπωρία για τους ασθενείς, χωρίς πολύωρη αναμονή για εξέταση από τον ιατρό. Επιπλέον δεν υπάρχει «νεκρός» χρόνος για τον ιατρό, ο οποίος έτσι παρέχει τις υπηρεσίες του σε πολύ περισσότερους ασθενείς. Κατ' επέκταση μειώνεται η λίστα αναμονής για ασθενείς που θέλουν να κλείσουν ραντεβού και παρέχεται σε αυτούς καλύτερη και ταχύτερη ποιότητα υπηρεσιών.

Επιπρόσθετα με αυτή την πρακτική αποφεύγεται το φαινόμενο οι ασθενείς που χάνουν το ραντεβού να προσέρχονται την επόμενη μέρα στο Νοσοκομείο, κάτι που συμβαίνει πολύ συχνά, ζητώντας να εξεταστούν ως υπεράριθμοι. Το όφελος από την αποφυγή αυτού του φαινομένου είναι ότι ο ιατρός δεν θα αναγκαστεί να εξετάσει περισσότερους ασθενείς σε μία ημέρα, άρα θα κουραστεί λιγότερο και πολύ πιθανόν να εκδίδει πιο έγκυρες διαγνώσεις (αποδοτικότητα, καλός συντονισμός, ελάχιστη αναμονή, αποτελεσματικότητα).

Ένα επιπλέον όφελος και ίσως το πιο σημαντικό είναι στην περίπτωση που κάποιος ασθενής μπορεί να νοσεί από μία σοβαρή ασθένειά χωρίς να το γνωρίζει. Από τη στιγμή που ο ασθενής θα προσέλθει στο ραντεβού, θα εξεταστεί από τον ιατρό, άρα θα υπάρξει έγκαιρη διάγνωση, θα πάρει την κατάλληλη φαρμακευτική αγωγή και είναι πολύ πιθανό να αποφευχθεί ο κίνδυνος για τη ζωή του, να μην επιδεινωθεί η κατάσταση της υγείας του ή ακόμη και να βελτιωθεί άμεσα (Elvira et al., 2018).

4.3.2 2η Ενέργεια: Κάλυψη των πιθανών κενών που θα προκύψουν

Μια άλλη ενέργεια που μπορεί να γίνει είναι η κάλυψη των πιθανών κενών που θα προκύψουν από μη εμφάνιση των ασθενών στα ραντεβού. Δηλαδή να υπάρχει η δυνατότητα, αν ο κατηγοριοποιητής προβλέψει ότι θα προκύψουν κενά σε μία συγκεκριμένη ημέρα, από μη εμφανίσεις ασθενών στα ραντεβού, τότε να μπορούν να προγραμματιστούν νέα ραντεβού σε αυτή την ημέρα, ώστε να καλυφθούν τα κενά.

Τα οφέλη που προκύπτουν από αυτή την ενέργεια είναι πολλά. Ένα βασικό όφελος είναι ότι δεν θα μείνει ανεκμετάλλευτος ο χρόνος του προγραμματισμένου ραντεβού που έμεινε κενό, αλλά στη θέση του συγκεκριμένου ασθενή, θα εξεταστεί άλλος ασθενής.

Αυτό είναι πολύ σημαντικό ιδιαιτέρως δε, αν σκεφτούμε την πιθανότητα ο ασθενής που εξετάστηκε σε αντικατάσταση του ασθενή που δεν προσήλθε στο προγραμματισμένο ραντεβού που είχε, να πάσχει από μία σοβαρή ασθένεια και μάλιστα χωρίς να το γνωρίζει. Σε αυτή την περίπτωση δεν θα εκτεθεί η ζωή του συγκεκριμένου ασθενή σε κίνδυνο, αφού θα εξεταστεί από τον ιατρό νωρίτερα, ο οποίος θα διαγνώσει έγκαιρά την κατάσταση της υγείας του και θα του συνταγογραφήσει την κατάλληλη φαρμακευτική αγωγή. Επιπλέον δεν θα υπάρχει καθόλου νεκρός χρόνος για τον ιατρό αφού δεν θα υπάρχουν κενά στα ραντεβού, ενώ θα μειωθεί αρκετά και η λίστα αναμονής για ασθενείς που θέλουν να κλείσουν νέο ραντεβού.

Το συμπέρασμα που προκύπτει είναι ότι οι παραπάνω δύο ενέργειες οδηγούν σε καλύτερη οργάνωση και λειτουργία του νοσοκομείου καθώς και καλύτερη και ταχύτερη ποιότητα παροχής υπηρεσιών προς τους πολίτες. Κατά συνέπεια οικοδομείται μία διαφορετική σχέση ασφάλειας και εμπιστοσύνης μεταξύ νοσοκομείου – ιατρού – ασθενή. Ο μεν ασθενής από την μια πλευρά αντιλαμβάνεται την προσπάθεια του νοσοκομείου να του παρέχει άμεσα την καλύτερη κατά το δυνατόν ιατροφαρμακευτική περίθαλψη παρόλες τις αντιξοότητες. Ο δε ιατρός αντιλαμβάνεται την σπουδαιότητα του λειτουργήματος που επιτελεί, νιώθει χρήσιμος, αποδοτικός και αποκτά κίνητρο για ακόμη καλύτερη παροχή υπηρεσιών προς τους ασθενείς (Huang & Hanauer, 2016).

4.4 Σημαντικότητα των μετρικών Precision, Recall και F-measure με βάση τις ενέργειες επίλυσης

Λαμβάνοντας υπόψη την 1η πιθανή ενέργεια που είναι η υπενθύμιση του ραντεβού στον ασθενή με κάποιο τρόπο όπως για παράδειγμα μέσω τηλεφωνικής επικοινωνίας ή μέσω SMS τότε για την περίπτωση του precision παρατηρείται ότι:

1. Αν ο κατηγοριοποιητής παρουσιάζει υψηλό precision (πολλές σωστές προβλέψεις) σημαίνει ότι, από τους ασθενείς που προέβλεψε ότι δεν θα εμφανιστούν στο ραντεβού (θετική κλάση) και τους έγινε υπενθύμιση με κάποιον τρόπο, οι περισσότεροι χρειαζόταν να λάβουν υπενθύμιση (σωστές προβλέψεις γιατί και στην πραγματικότητα ανήκουν στην θετική κλάση). Αντίστοιχα, λίγοι έλαβαν υπενθύμιση χωρίς λόγο (δηλαδή αυτοί που στην πραγματικότητα ανήκουν στην αρνητική κλάση) γιατί θα εμφανίζονταν έτσι κι αλλιώς (λανθασμένες προβλέψεις) στο ραντεβού.
2. Αντίθετα αν ο κατηγοριοποιητής παρουσιάζει χαμηλό precision (πολλές λανθασμένες προβλέψεις) σημαίνει ότι, από τους ασθενείς που προέβλεψε ότι δεν θα εμφανιστούν στο ραντεβού (θετική κλάση) και τους έγινε υπενθύμιση με κάποιον από τους παραπάνω τρόπους, πολλοί έλαβαν υπενθύμιση χωρίς λόγο (δηλαδή αυτοί που στην πραγματικότητα ανήκουν στην αρνητική κλάση) γιατί θα εμφανίζονταν έτσι κι αλλιώς (λανθασμένες προβλέψεις) στο ραντεβού. Αντίστοιχα, λίγοι χρειαζόταν να λάβουν υπενθύμιση και όντως την έλαβαν (σωστές προβλέψεις γιατί και στην πραγματικότητα ανήκουν στην θετική κλάση).

Σύμφωνα με τα παραπάνω προκύπτει το συμπέρασμα ότι ακόμη και αν το precision είναι χαμηλό, άρα πολλές άσκοπες υπενθυμίσεις, αυτό δεν αποτελεί πρόβλημα εάν το κόστος υπενθύμισης δεν είναι πολύ υψηλό. Για τη μείωση του κόστους υπάρχουν διάφοροι τρόποι. Ένας από αυτούς είναι το νοσοκομείο να συνάψει σύμβαση με πάροχο τηλεφωνίας, ώστε τα τηλεφωνήματα να χρεώνονται αρκετά χαμηλά. Αντίστοιχα στην περίπτωση αποστολής μηνυμάτων μπορεί να στέλνονται στους ασθενείς αυτοματοποιημένα μηνύματα μέσω πακέτων μηνυμάτων που μπορεί να κοστίζουν λιγότερο. Εάν το κόστος είναι υπολογίσιμο, είναι σημαντικό το precision να μην είναι πολύ χαμηλό. Επομένως θα επιλεγεί ένας κατηγοριοποιητής με υψηλό precision, αλλά χωρίς να μειωθεί ιδιαίτερα το recall.

Για την περίπτωση του recall παρατηρούνται τα παρακάτω:

1. Αντίστοιχα, αν ο κατηγοριοποιητής παρουσιάζει υψηλό recall σημαίνει ότι από τους ασθενείς που δεν θα εμφανίζονταν στο ραντεβού (θετική κλάση) και έπρεπε να τους γίνει υπενθύμιση με κάποιο τρόπο, προέβλεψε πολλούς σωστά και όντως έλαβαν υπενθύμιση. Αντίθετα προέβλεψε λανθασμένα ότι λίγοι ασθενείς θα εμφανιστούν στο

ραντεβού (λανθασμένες προβλέψεις), οπότε αν και θα έπρεπε, δεν έλαβαν υπενθύμιση. (π.χ. 0,503 ή 50% στο παράδειγμα).

2. Αντίστοιχα, αν ο κατηγοριοποιητής παρουσιάζει χαμηλό recall σημαίνει ότι από τους ασθενείς που δεν θα εμφανίζονταν στο ραντεβού (θετική κλάση) και έπρεπε να τους γίνει υπενθύμιση με κάποιο τρόπο, προέβλεψε λίγους σωστά που όντως έλαβαν υπενθύμιση. Αντίθετα προέβλεψε λανθασμένα ότι πολλοί ασθενείς θα εμφανιστούν στο ραντεβού (λανθασμένες προβλέψεις), οπότε αν και θα έπρεπε δεν έλαβαν υπενθύμιση.

Από τα παραπάνω εξάγεται το συμπέρασμα ότι το υψηλό recall είναι πολύ σημαντικό για αυτή την ενέργεια, με στόχο να γίνει υπενθύμιση σε όσο το δυνατόν περισσότερους ασθενείς, ώστε να μειωθούν τα προβλήματα που απορρέουν από τη μη εμφάνιση των ασθενών στα προγραμματισμένα ραντεβού.

Στην περίπτωση λοιπόν που το επιθυμητό είναι να γίνει υπενθύμιση του ραντεβού στους ασθενείς, τότε το recall είναι πιο σημαντικό από το precision. Αυτό συμβαίνει γιατί στην περίπτωση του precision εξετάζεται το ποσοστό της άσκοπης υπενθύμισης ραντεβού στους ασθενείς, ενώ στην περίπτωση του recall εξετάζεται αν θα λάβουν υπενθύμιση οι περισσότεροι από τους ασθενείς που δεν θα εμφανιστούν στο ραντεβού. Επομένως το να λάβει υπενθύμιση κάποιος ασθενής που δεν θα εμφανιστεί στο ραντεβού, είναι πιο σημαντικό από το να λάβει άσκοπη υπενθύμιση κάποιος ασθενής που ούτως ή άλλως θα εμφανιστεί στο ραντεβού.

Λαμβάνοντας υπόψη την $2^{η}$ πιθανή ενέργεια που είναι η κάλυψη των πιθανών κενών τότε για την περίπτωση του precision παρατηρείται ότι:

1. Αν ο κατηγοριοποιητής παρουσιάζει υψηλό precision αυτό σημαίνει ότι από τις συνολικές προβλέψεις που έκανε για τα κενά που θα προκύψουν και τα αντικατέστησε με νέα ραντεβού, οι περισσότερες προβλέψεις είναι σωστές γιατί οι ασθενείς δεν εμφανίστηκαν στο ραντεβού (θετική κλάση). Αντίστοιχα λίγα κενά ραντεβού από αυτά αντικαταστάθηκαν λανθασμένα γιατί προέβλεψε ότι τα ραντεβού αυτά θα ήταν κενά (Δηλαδή αυτά που στην πραγματικότητα ανήκουν στην αρνητική κλάση). Αυτά όμως δεν θα έπρεπε να αντικατασταθούν γιατί οι ασθενείς στην πραγματικότητα εμφανίσθηκαν στο ραντεβού. Κατά συνέπεια αν καλυφθούν τα πιθανά κενά ραντεβού από τους ασθενείς που προβλέφθηκε ότι δεν θα εμφανιστούν στο ραντεβού, με νέους ασθενείς, τότε στις περισσότερες περιπτώσεις δεν θα υπάρξει πρόβλημα ταυτόχρονης εμφάνισης των ασθενών στο ραντεβού.
2. Αν ο κατηγοριοποιητής παρουσιάζει χαμηλό precision αυτό σημαίνει ότι από τις συνολικές προβλέψεις που έκανε για τα κενά και τα αντικατέστησε με νέα ραντεβού, οι λιγότερες προβλέψεις είναι σωστές γιατί λίγοι ασθενείς όντως δεν εμφανίστηκαν

στο ραντεβού (θετική κλάση). Αντίστοιχα πολλά κενά από αυτά αντικαταστάθηκαν λανθασμένα γιατί προέβλεψε ότι τα ραντεβού αυτά θα ήταν κενά (Δηλαδή αυτά που στην πραγματικότητα ανήκουν στην αρνητική κλάση). Αυτά όμως δεν θα έπρεπε να αντικατασταθούν γιατί οι ασθενείς στην πραγματικότητα εμφανίσθηκαν στο ραντεβού. Κατά συνέπεια αν καλυφθούν τα πιθανά κενά ραντεβού από τους ασθενείς που προβλέφθηκε ότι δεν θα εμφανιστούν στο ραντεβού, με νέους ασθενείς, τότε στις περισσότερες περιπτώσεις θα υπάρξει πρόβλημα ταυτόχρονης εμφάνισης των ασθενών στο ραντεβού. Σύμφωνα με τα παραπάνω εξάγεται το συμπέρασμα, ότι το υψηλό precision είναι σημαντικό για την ελαχιστοποίηση των προβλημάτων ταυτόχρονης εμφάνισης ασθενών σε ραντεβού. Αντίθετα το χαμηλό precision δημιουργεί προβλήματα λόγω ταυτόχρονης εμφάνισης των ασθενών στο ίδιο ραντεβού.

Για την περίπτωση του recall παρατηρείται ότι:

1. Αν ο κατηγοριοποιητής παρουσιάζει υψηλό recall σημαίνει ότι από τους ασθενείς που δεν θα εμφανίζονταν στο ραντεβού (θετική κλάση) και θα έπρεπε να αντικαταστήσει τα ραντεβού τους με νέα ραντεβού, προέβλεψε πολλούς σωστά και όντως αντικατέστησε τα ραντεβού τους. Αντίθετα προέβλεψε λανθασμένα ότι λίγοι ασθενείς από αυτούς θα εμφανιστούν στο ραντεβού (λανθασμένες προβλέψεις γιατί αυτοί στην πραγματικότητα δεν εμφανίσθηκαν), οπότε δεν αντικατέστησε τα ραντεβού τους, ενώ θα έπρεπε.
2. Αν ο κατηγοριοποιητής παρουσιάζει χαμηλό recall σημαίνει ότι από τους ασθενείς που δεν θα εμφανίζονταν στο ραντεβού (θετική κλάση) και θα έπρεπε να αντικαταστήσει τα ραντεβού τους με νέα ραντεβού, προέβλεψε λίγους σωστά και όντως αντικατέστησε τα ραντεβού τους. Αντίθετα προέβλεψε λανθασμένα ότι πολλοί ασθενείς από αυτούς θα εμφανιστούν στο ραντεβού (λανθασμένες προβλέψεις γιατί αυτοί στην πραγματικότητα δεν εμφανίσθηκαν), οπότε δεν αντικατέστησε τα ραντεβού τους, ενώ θα έπρεπε.

Από τα παραπάνω συμπεραίνεται ότι το υψηλό recall είναι πολύ σημαντικό για αυτή την ενέργεια, με στόχο να αντικατασταθούν όσο το δυνατόν περισσότερα κενά ραντεβού που προκύπτουν λόγω μη εμφάνισης των ασθενών στα προγραμματισμένα ραντεβού.

Συνεπώς στην ενέργεια της αντικατάστασης κενών είναι σημαντικά τόσο το precision όσο και το recall. Η ιδανική κατάσταση βέβαια είναι να υπάρχει ταυτόχρονα υψηλό precision και υψηλό recall. Αυτό όμως συνήθως δεν είναι εφικτό, οπότε σε αυτήν την περίπτωση θεωρείται σημαντικότερο το υψηλό precision, ώστε να μην υπάρχουν προβλήματα ταυτόχρονης εμφάνισης ασθενών, ακόμα και αν το recall είναι χαμηλό. Δηλαδή ακόμα κι αν ο κατηγοριοποιητής προβλέψει σωστά πολύ λίγα κενά από αυτά που υπάρχουν στην

πραγματικότητα. Στην περίπτωση όμως που ένας κατηγοριοποιητής παρουσιάζει πολύ υψηλό Precision αλλά και πολύ χαμηλό Recall, ιδανικότερο θα ήταν να επιλεγεί κάποιος άλλος κατηγοριοποιητής με υψηλότερο Recall αλλά χωρίς να μειωθεί ιδιαίτερα το Precision. Δηλαδή να προβλέψει και να αντικαταστήσει σωστά περισσότερα κενά από αυτά που υπήρχαν στην πραγματικότητα, αλλά χωρίς να αυξηθούν ιδιαίτερα οι λάνθασμένες προβλέψεις. Το ζητούμενο είναι να μην αυξηθούν οι λανθασμένες προβλέψεις γιατί εάν αυξηθούν σημαίνει ότι ο επιλεγμένος κατηγοριοποιητής θα προβλέψει ότι υπάρχουν περισσότερα κενά και θα τα αντικαταστήσει λανθασμένα, ενώ δεν θα έπρεπε, αυξάνοντας έτσι το πρόβλημα της ταυτόχρονης εμφάνισης ασθενών στο προγραμματισμένο ραντεβού.

Οι πιθανές ενέργειες που περιγράφηκαν παραπάνω συμβάλλουν και οι δύο στη μείωση του προβλήματος των κενών που θα προκύψουν λόγω της μη εμφάνισης των ασθενών στα προγραμματισμένα ραντεβού. Με βάση αυτόν τον συλλογισμό το ιδανικό για ένα νοσοκομείο θα ήταν να υπάρχει η δυνατότητα να πραγματοποιούνται ταυτόχρονα και οι 2 παραπάνω ενέργειες. Το κίνητρο αυτής της προσέγγισης είναι η υπόθεση ότι ο συνδυασμός των 2 ενεργειών θα μπορούσε να επιφέρει ακόμα καλύτερα αποτελέσματα στη λύση του προβλήματος της μη εμφάνισης των ασθενών στα προγραμματισμένα ραντεβού. Παρόλα αυτά, ενώ θεωρητικά αυτή η προσέγγιση φαντάζει ιδανική, στην πράξη πιθανόν να μην επιφέρει τα αναμενόμενα αποτελέσματα. Οι λόγοι αναλύονται διεξοδικά παρακάτω.

Στην 1η ενέργεια το ζητούμενο είναι το υψηλό Recall ενώ στην 2η ενέργεια το ζητούμενο είναι το υψηλό Precision. Αυτό σημαίνει ότι για να πραγματοποιηθούν ταυτόχρονα και οι 2 ενέργειες ικανοποιητικά, απαιτείται ένας κατηγοριοποιητής που να δίνει ταυτόχρονα και υψηλό Recall αλλά και υψηλό Precision. Αυτό όμως είναι πολύ δύσκολο να συμβεί, γιατί συνήθως όταν αυξάνεται το Recall, μειώνεται το αντίστοιχο Precision και το αντίστροφο. Συνεπώς ο βέλτιστος κατηγοριοποιητής και για τις 2 ενέργειες είναι αυτός που παρουσιάζει ταυτόχρονα αξιόλογη τιμή και στις 2 μετρικές. Από την στιγμή λοιπόν που η επιλογή του βέλτιστου κατηγοριοποιητή είναι απαραίτητο να γίνεται λαμβάνοντας υπόψη συνδυαστικά και τις 2 μετρικές, ο καλύτερος τρόπος είναι να χρησιμοποιηθεί η μετρική F-measure. Η μετρική F-measure δίνει μία ισορροπημένη τιμή που είναι ο συνδυασμός των μετρικών Recall και Precision. Αυτό όμως πρακτικά σημαίνει ότι ο επιλεγμένος κατηγοριοποιητής πιθανότατα θα έχει χαμηλότερο Recall σε σχέση με τον κατηγοριοποιητή που θα επιλεγόταν, εάν γινόταν μόνο η 1η ενέργεια και αντίστοιχα χαμηλότερο Precision σε σχέση με τον κατηγοριοποιητή που θα επιλεγόταν, εάν γινόταν μόνο η 2η ενέργεια. Με απλά λόγια με χρήση της μετρικής F-measure η πραγματοποίηση των 2 ενεργειών ταυτόχρονα, πιθανότατα θα παρουσιάζει χειρότερα αποτελέσματα σε σχέση με το να γίνει κάθε ενέργεια μεμονωμένα.

Επιπλέον η ταυτόχρονη πραγματοποίηση και των 2 ενεργειών είναι πιθανό να επιφέρει και ένα πρόσθετο πρόβλημα. Με βάση την 1η ενέργεια, όταν ο κατηγοριοποιητής προβλέπει ότι

κάποιος ασθενής δεν θα εμφανισθεί στο προγραμματισμένο ραντεβού, τότε με κάποιο τρόπο θα γίνει υπενθύμιση του ραντεβού στον ασθενή. Κατά συνέπεια θα αυξηθεί η πιθανότητα ο ασθενής να μην ξεχάσει το ραντεβού του και να προσέλθει στο νοσοκομείο για εξέταση. Στην συνέχεια με βάση την 2η ενέργεια, θα γίνει αντικατάσταση του προγραμματισμένου ραντεβού του συγκεκριμένου ασθενή με νέο ραντεβού. Αυτό λοιπόν σημαίνει ότι ενώ από τη μια πλευρά γίνεται στον ασθενή υπενθύμιση για το προγραμματισμένο ραντεβού του, ώστε να προσέλθει στο νοσοκομείο, από την άλλη πλευρά γίνεται αντικατάσταση του ίδιου προγραμματισμένου ραντεβού με νέο ραντεβού. Άρα σε αυτή την περίπτωση θα προσέλθουν στο νοσοκομείο την ίδια ώρα, για το ίδιο ραντεβού και οι 2 ασθενείς. Έτσι όμως αυξάνεται το φαινόμενο της ταυτόχρονης εμφάνισης των ασθενών σε ραντεβού. Επομένως ενώ με την 1η ενέργεια γίνεται προσπάθεια επίλυσης του προβλήματος της μη εμφάνισης των ασθενών στα ραντεβού, πραγματοποιώντας στη συνέχεια και την 2η ενέργεια αυτόματα αυξάνεται το φαινόμενο της ταυτόχρονης εμφάνισης των ασθενών στα ραντεβού.

5

Σύνολα δεδομένων με ανισοκατανομή κλάσεων

5.1 Ανισοκατανομή κλάσεων

Σε κάθε σύνολο δεδομένων (dataset) η κατανομή των κλάσεων (class distribution) είναι η αναλογία μεταξύ των κλάσεων που περιέχει. Με άλλα λόγια κατανομή είναι η σχέση μεταξύ των ποσοστών των προτύπων που αντιστοιχούν σε κάθε κλάση. Αν ένα σύνολο δεδομένων έχει 2 κλάσεις (K1, K2) και το 50% των προτύπων ανήκει στην κλάση K1 και το υπόλοιπο 50% των προτύπων ανήκει στην κλάση K2, αυτό σημαίνει ότι το σύνολο δεδομένων έχει ίση αναλογία κλάσεων και χαρακτηρίζεται ως **ισορροπημένο (balanced)**. Στην πράξη ισορροπημένο σύνολο δεδομένων είναι αυτό που έχει περίπου ίσα ποσοστά προτύπων μεταξύ των κλάσεων. Δηλαδή, αν σε μία κλάση ανήκει το 48% των προτύπων και η άλλη κλάση έχει το 52% των προτύπων, τότε το σύνολο δεδομένων πάλι θεωρείται ισορροπημένο, εφόσον αυτή η μικρή απόκλιση που υπάρχει δεν επηρεάζει ουσιαστικά την απόδοση της κατηγοριοποίησης.

Αντίθετα, αν το ποσοστό των προτύπων που ανήκει στην κλάση K1 είναι πολύ μεγαλύτερο (ή πολύ μικρότερο) από το ποσοστό των προτύπων που ανήκει στην κλάση K2, τότε το σύνολο δεδομένων θεωρείται **μη-ισορροπημένο (imbalanced)**. Παράδειγμα στην κλάση K1 ανήκει το 80% των προτύπων και στην κλάση K2 ανήκει το 20% των προτύπων. Από την άλλη πλευρά υπάρχουν και σύνολα δεδομένων στα οποία η ανισοκατανομή μεταξύ των κλάσεων μπορεί να είναι πολύ μεγάλη, δηλαδή της τάξης του 100:1, 1000:1 ή ακόμα και 10.000:1. Σε ένα μη-ισορροπημένο σύνολο δεδομένων, η κλάση με τα λιγότερα πρότυπα αναφέρεται ως

μειοψηφική κλάση (minority class), ενώ η κλάση με τα περισσότερα πρότυπα ως πλειοψηφική κλάση (majority class).

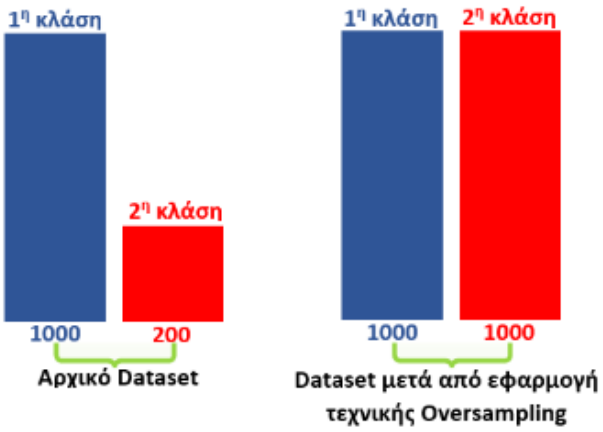
Κατά την εφαρμογή αλγορίθμων εξόρυξης γνώσης σε μη-ισορροπημένο σύνολο δεδομένων, μπορεί να δημιουργηθούν προβλήματα. Αυτό συμβαίνει γιατί τα πρότυπα της πλειοψηφικής κλάσης υπερισχύουν έναντι των προτύπων της μειοψηφικής κλάσης. Αυτό έχει ως αποτέλεσμα ο αλγόριθμος να εκπαιδεύεται περισσότερο με βάση τα πρότυπα που ανήκουν στην πλειοψηφική κλάση. Κατά συνέπεια να μην μπορεί να εκπαιδευτεί επαρκώς στην μειοψηφική κλάση, δηλαδή την κλάση που έχει τα λιγότερα πρότυπα. Άρα δεν θα μπορεί να κατηγοριοποιήσει σωστά νέα δεδομένα που ανήκουν στην μειοψηφική κλάση. Αντίθετα όταν το σύνολο δεδομένων είναι ισορροπημένο, η εφαρμογή αλγορίθμων εξόρυξης γνώσης δεν παρουσιάζει τα προαναφερθέντα προβλήματα. Επειδή όμως τα σύνολα δεδομένων που προκύπτουν από δεδομένα που χρησιμοποιούνται στον πραγματικό κόσμο είναι συνήθως μη-ισορροπημένα, για να αποφευχθούν κατά την εκπαίδευση του αλγορίθμου τα παραπάνω προβλήματα, ένας τρόπος είναι να μετατραπεί το σύνολο δεδομένων σε ισορροπημένο (Haibo He & Garcia, 2009).

5.2 Τεχνικές δειγματοληψίας

Με τη χρήση τεχνικών δειγματοληψίας μπορεί να ρυθμιστεί η κατανομή των κλάσεων σε ένα σύνολο δεδομένων, ώστε να μετατραπεί από μη-ισορροπημένο σε ισορροπημένο. Οι δύο βασικές τεχνικές δειγματοληψίας που εφαρμόζονται ευρέως είναι το Oversampling (υπερδειγματοληψία) και το Undersampling (υποδειγματοληψία).

5.2.1 Τεχνική Υπερδειγματοληψίας (Oversampling)

Υπερδειγματοληψία (Oversampling) είναι η τεχνική με την οποία προστίθενται στο αρχικό σύνολο δεδομένων καινούργια πρότυπα ώστε να αυξηθεί το πλήθος των προτύπων της μειοψηφικής κλάσης, για να γίνει ίσο με το πλήθος των προτύπων της πλειοψηφικής κλάσης. Τα νέα πρότυπα που προστίθενται στο αρχικό σύνολο δεδομένων προκύπτουν, με κάποιο τρόπο, από τα πρότυπα που ήδη υπήρχαν μέσα σε αυτό. Μετά την εφαρμογή αυτή της τεχνικής το σύνολο δεδομένων θα είναι πλέον ισορροπημένο. Για παράδειγμα αν η 1η κλάση έχει 1000 πρότυπα και η 2η κλάση έχει 200 πρότυπα, τότε μετά την εφαρμογή υπερδειγματοληψίας στο σύνολο δεδομένων, η 1η κλάση θα εξακολουθεί να έχει 1000 πρότυπα, ενώ το πλήθος των προτύπων της 2ης κλάσης θα αυξηθεί κατά 800 πρότυπα ώστε να αποτελείται τελικά από 1000 πρότυπα όπως και η 1η κλάση (Εικόνα 14) (Rahman & Davis, 2013).



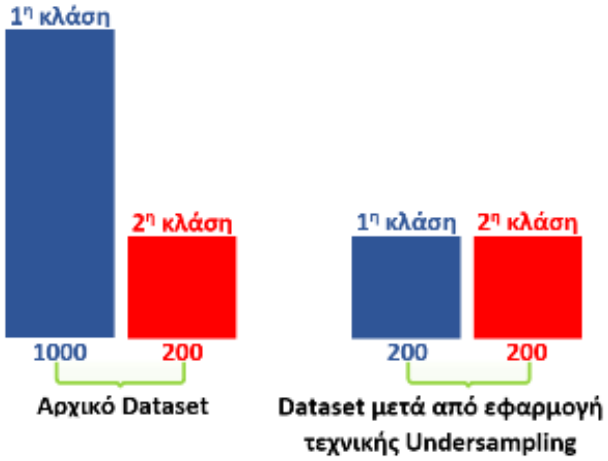
Εικόνα 14: Τεχνική Υπερδειγματοληψίας (Oversampling)

Υπάρχουν διάφορες τεχνικές υπερδειγματοληψίας όπως Random oversampling (Japkowicz & Stephen, 2002) , SMOTE (Chawla et al., 2002), ADASYN (Haibo He et al., 2008) και cluster-based oversampling (Jo & Japkowicz, 2004).

Η πιο απλή τεχνική υπερδειγματοληψίας είναι η Random Oversampling. Στη συγκεκριμένη τεχνική επιλέγονται τυχαία κάποια πρότυπα της μειοψηφικής κλάσης. Στη συνέχεια δημιουργούνται αντίγραφα τους, τα οποία προστίθενται στο αρχικό σύνολο δεδομένων εκπαίδευσης.

5.2.2 Τεχνική Υποδειγματοληψίας (Undersampling)

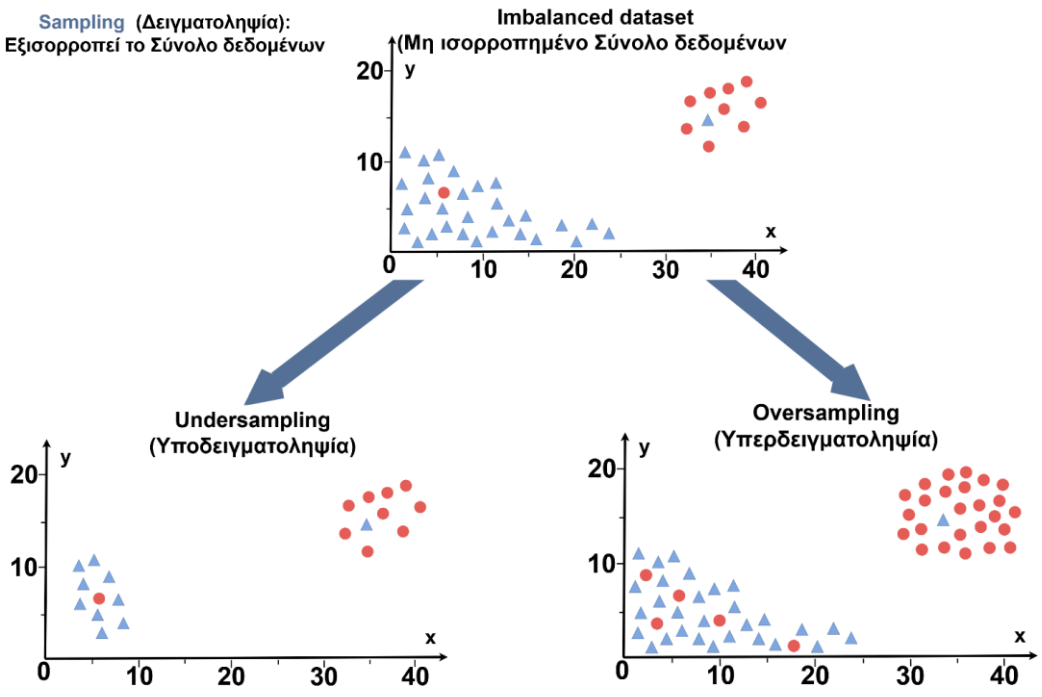
Υποδειγματοληψία (Undersampling) είναι η τεχνική με την οποία αφαιρούνται από το αρχικό σύνολο δεδομένων κάποια πρότυπα, ώστε να μειωθεί το πλήθος των προτύπων της πλειοψηφικής κλάσης, για να γίνει ίσο με το πλήθος των προτύπων της μειοψηφικής κλάσης. Μετά την εφαρμογή αυτή της τεχνικής το τελικό σύνολο δεδομένων θα είναι πλέον ισορροπημένο. Για παράδειγμα αν η 1η κλάση έχει 1000 πρότυπα και η 2η κλάση έχει 200 πρότυπα, τότε μετά την εφαρμογή υποδειγματοληψίας στο σύνολο δεδομένων, η 1η κλάση θα μειωθεί κατά 800 πρότυπα, ώστε να αποτελείται από 200 πρότυπα, όπως και η 1η κλάση, ενώ το πλήθος των προτύπων της 2ης κλάσης θα εξακολουθεί να έχει 200 πρότυπα (Εικόνα 15) (Rahman & Davis, 2013).



Εικόνα 15: Τεχνική Υποδειγματοληψίας (Undersampling)

5.3 Σύγκριση τεχνικών Υπερδειγματοληψίας και Υποδειγματοληψίας

Τόσο η τεχνική υπερδειγματοληψίας, όσο και η τεχνική υποδειγματοληψίας μπορούν να αποδώσουν εξίσου καλά. Η επιλογή της μιας ή της άλλης τεχνικής εξαρτάται κατά κύριο λόγο από τις συνθήκες και το πρόβλημα που καλούνται να λύσουν. Η βασική διαφορά τους όπως φαίνεται στην Εικόνα 16 είναι ότι, ενώ η υπερδειγματοληψία προσθέτει δεδομένα, η υποδειγματοληψία αφαιρεί δεδομένα από το αρχικό σύνολο δεδομένων.



Εικόνα 16: Σύγκριση τεχνικών Υπερδειγματοληψίας (Oversampling) – Υποδειγματοληψίας (Undersampling)

Στην περίπτωση της υποδειγματοληψίας, το πρόβλημα είναι σχετικά προφανές: Με την αφαίρεση προτύπων από την πλειοψηφική κλάση, είναι πιθανό ο κατηγοριοποιητής να χάσει σημαντικές πληροφορίες που αφορούν την πλειοψηφική κλάση. Παρόλα αυτά η μείωση του μεγέθους του συνόλου δεδομένων έχει δύο βασικά πλεονεκτήματα. Καταρχήν μειώνει τον απαιτούμενο χρόνο εκπαίδευσης, ο οποίος μπορεί να είναι πολύ μεγάλος σε περιπτώσεις υπερβολικά μεγάλων συνόλων δεδομένων. Επιπλέον η μείωση του μεγέθους του συνόλου δεδομένων είναι εξαιρετικά χρήσιμη σε περιπτώσεις μεγάλων συνόλων τα οποία μπορεί να μην χωράνε στη μνήμη του υπολογιστή (Batista et al., 2004).

Από την άλλη πλευρά, στην περίπτωση της υπερδειγματοληψίας το πλεονέκτημα είναι ότι, δεν χάνεται καμία πληροφορία από το σύνολο δεδομένων εκπαίδευσης, αφού διατηρούνται όλα τα πρότυπα και στις δύο κλάσεις. Ωστόσο, το μειονέκτημα είναι ότι, αυξάνεται σημαντικά το μέγεθος του συνόλου δεδομένων εκπαίδευσης. Κάτι τέτοιο έχει επιπτώσεις τόσο στο χρόνο που χρειάζεται για να εκπαιδευτεί το μοντέλο, όσο και στη μνήμη που απαιτείται κατά την διάρκεια της εκπαίδευσης (Rahman & Davis, 2013).

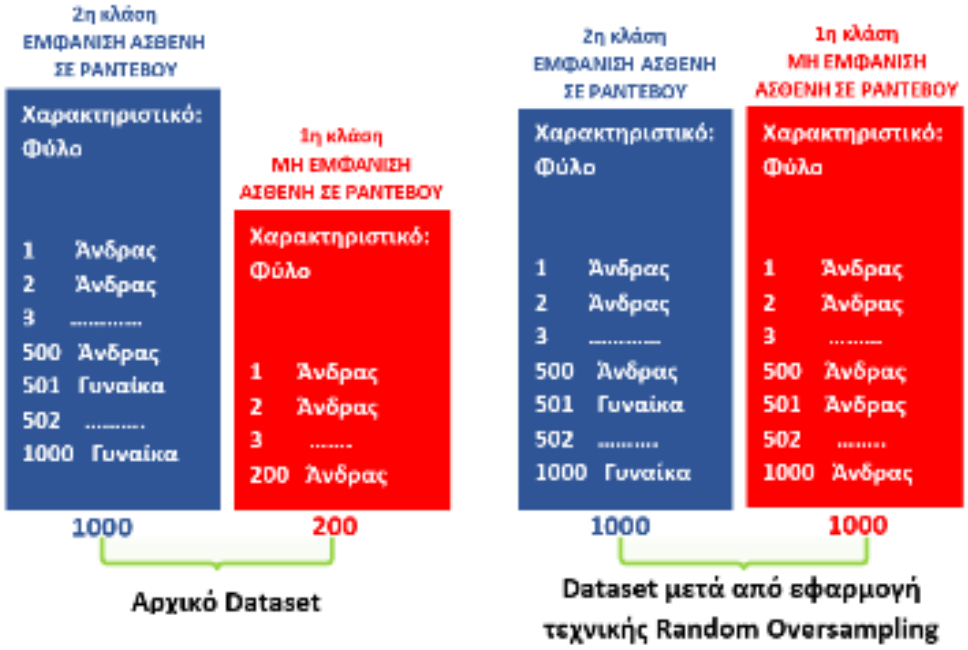
Ειδικά στην περίπτωση της τυχαίας υπερδειγματοληψίας προκύπτει ένα ακόμα μειονέκτημα. Δεδομένου ότι η τυχαία υπερδειγματοληψία απλώς προσθέτει πιστά αντίγραφα από τα πρότυπα που υπήρχαν ήδη στο αρχικό σύνολο δεδομένων, οι κανόνες κατηγοριοποίησης που παράγονται θα είναι πολύ εξειδικευμένοι. Αυτό συμβαίνει γιατί ένας τέτοιος κανόνας προκύπτει με βάση τα πολλαπλά αντίγραφα του ίδιου προτύπου. Δηλαδή ενώ φαινομενικά ο κανόνας αντιστοιχεί σε πολλά πρότυπα, στην πραγματικότητα αντιστοιχεί μόνο σε ένα πρότυπο. Σ' αυτή την περίπτωση αυτό που μπορεί να συμβεί είναι να δώσει ο κατηγοριοποιητής μεγαλύτερη βαρύτητα σε χαρακτηριστικά που στην πραγματικότητα δεν είναι πολύ σημαντικά. Αυτή η πρακτική μπορεί να οδηγήσει σε υπερμοντελοποίηση (overfitting) και κατά συνέπεια σε μειωμένη ικανότητα γενίκευσης του κατηγοριοποιητή (Haibo He & Garcia, 2009). Ένα παράδειγμα υπερμοντελοποίησης (overfitting) μετά από την εφαρμογή της τεχνικής Random Oversampling περιγράφεται στη συνέχεια.

Δίνεται ένα σύνολο δεδομένων που περιέχει δεδομένα ασθενών σχετικά με ιατρικά ραντεβού. Τα δεδομένα μπορεί να περιλαμβάνουν χαρακτηριστικά όπως ηλικία, φύλο, χρόνιες ασθένειες και γενικότερα πληροφοριές σχετικά με ιατρικά ραντεβού. Τα δεδομένα μπορεί να ανήκουν σε μία από τις παρακάτω 2 κλάσεις:

- 1η κλάση – Μη εμφάνιση ασθενή σε ραντεβού (Ο ασθενής για κάποιο λόγο δεν πήγε στο ραντεβού)
- 2η κλάση – Εμφάνιση ασθενή σε ραντεβού

Το ζητούμενο είναι, να μπορεί ο κατηγοριοποιητής να προβλέψει σε ποια κλάση θα κατηγοριοποιηθεί ένας νέος ασθενής (άγνωστο δεδομένο). Με άλλα λόγια, να μπορεί ο αλγόριθμος να εκτιμήσει, εάν ένα άγνωστο δεδομένο είναι πιο πιθανό να μην εμφανιστεί στο

ραντεβού, οπότε να το κατηγοριοποιήσει στην 1η κλάση ή αν είναι πιθανότερο να εμφανιστεί στο ραντεβού, οπότε να το κατηγοριοποιήσει στην 2η κλάση.



Εικόνα 17: Περίπτωση υπερμοντελοποίησης (Overfitting) μετά από εφαρμογή τεχνικής Random Oversampling

Έστω ότι η 2η κλάση που είναι η «Εμφάνιση ασθενή σε ραντεβού» περιέχει π.χ. 1000 πρότυπα από τα οποία τα μισά στο χαρακτηριστικό «Φύλο» έχουν τιμή «Άνδρας» και τα υπόλοιπα μισά έχουν τιμή «Γυναίκα». Η 1η κλάση που είναι η «Μη εμφάνιση ασθενή σε ραντεβού» περιέχει π.χ. 200 πρότυπα που όμως αφορούν ΜΟΝΟ ασθενείς που είναι «Άνδρες», δηλαδή προφανώς δεν περιέχει αντιπροσωπευτικό δείγμα απ' όλους τους ασθενείς.

Αν εφαρμοστεί η τεχνική Random Oversampling στην 1η κλάση, θα προστεθούν σ' αυτή τόσα «εικονικά» πρότυπα, όσα χρειάζεται ώστε οι δύο κλάσεις να έχουν ίσο αριθμό προτύπων. Στη συγκεκριμένη περίπτωση προστίθενται 800 «εικονικά» πρότυπα, που είναι ακριβή αντίγραφα των αρχικών προτύπων. Συνεπώς στο τελικό σύνολο δεδομένων η 1η κλάση θα περιέχει 1000 πρότυπα που αφορούν ασθενείς ΜΟΝΟ «Άνδρες». Αυτό σημαίνει ότι ο κατηγοριοποιητής θα εκπαιδευτεί πολύ καλά όσον αφορά την 1η κλάση, ΜΟΝΟ σε ασθενείς που είναι «Άνδρες», ενώ στη 2η κλάση θα εκπαιδευτεί πολύ καλά τόσο σε «Άνδρες», όσο και σε «Γυναίκες».

Αν όμως δοθούν στον αλγόριθμο καινούργια (άγνωστα) δεδομένα προς κατηγοριοποίηση (δηλαδή ένα καινούργιο σύνολο δεδομένων) με ασθενείς τόσο «Άνδρες» όσο και «Γυναίκες», ο αλγόριθμος θα μπορέσει να κατηγοριοποιήσει σωστά ΜΟΝΟ τους «Άνδρες» ασθενείς που δεν εμφανίστηκαν στο ραντεβού, γιατί έχει εκπαιδευτεί ΜΟΝΟ σε αυτούς. Άρα έχει γίνει υπερμοντελοποίηση (overfitting) στους «Άνδρες» ασθενείς της 1ης κλάσης – Μη εμφάνιση ασθενή σε ραντεβού (Εικόνα 17).

5.4 SMOTE (Synthetic Minority Over-sampling Technique)

Ο αλγόριθμος SMOTE είναι μια τεχνική υπερδειγματοληψίας η οποία δημιουργεί συνθετικά πρότυπα. Η συγκεκριμένη τεχνική θεωρείται μία από τις δημοφιλέστερες τεχνικές υπερδειγματοληψίας. Αυτή η προσέγγιση εμπνεύστηκε από μια τεχνική που αποδείχθηκε επιτυχής στην αναγνώριση χειρόγραφων χαρακτήρων με την οποία ασχολήθηκαν οι (Ha & Bunke, 1997) (Ha & Bunke, 1997).

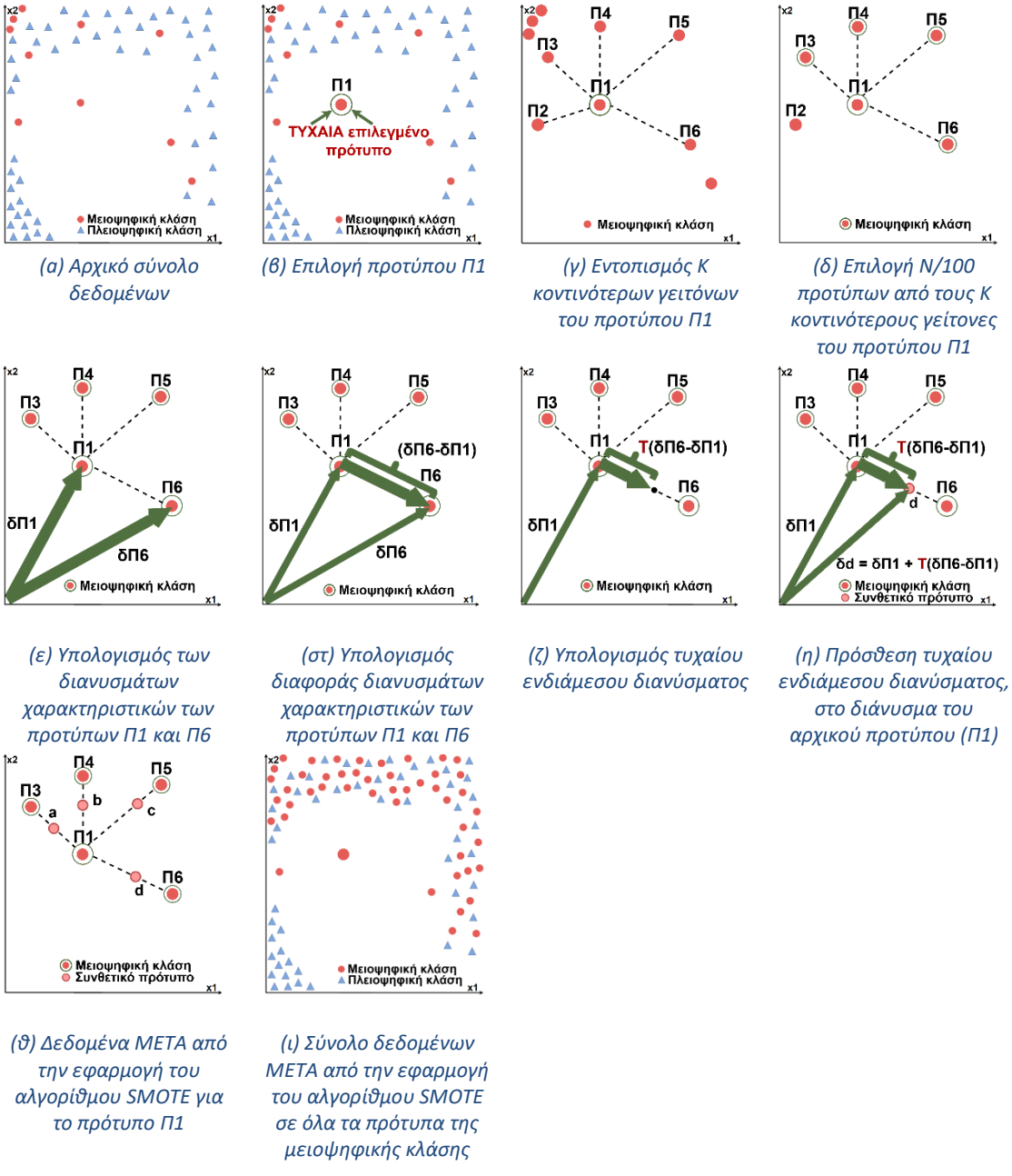
Αυτοί ουσιαστικά δημιούργησαν επιπλέον δεδομένα εκπαίδευσης, εκτελώντας συγκεκριμένες λειτουργίες στα πραγματικά δεδομένα που υπήρχαν στο αρχικό σύνολο δεδομένων. Στην περίπτωσή τους, λειτουργίες όπως η περιστροφή και η παραμόρφωση των χειρόγραφων χαρακτήρων, ήταν φυσικοί τρόποι για να τροποποιήσουν το σύνολο δεδομένων εκπαίδευσης. Αντίθετα, το SMOTE δημιουργεί συνθετικά πρότυπα με τέτοιον τρόπο, ώστε να μπορεί να χρησιμοποιηθεί σε ευρύ πεδίο εφαρμογών καθώς ο τρόπος λειτουργίας του είναι πιο γενικός και δεν εξαρτάται από τον τύπο των δεδομένων.

Η υπερδειγματοληψία της μειοψηφικής κλάσης γίνεται με την δημιουργία συνθετικών προτύπων κατά μήκος των ευθυγράμμων τμημάτων που ενώνουν το κάθε πρότυπο της μειοψηφικής κλάσης με κάποιους ή όλους τους k κοντινότερους γείτονες του (Εικόνα 18 - β, γ). Με τον τρόπο αυτό, τα νέα συνθετικά πρότυπα θα είναι πρότυπα τα οποία θα μοιάζουν αρκετά, τόσο με τα αρχικά πρότυπα της μειοψηφικής κλάσης, όσο και με τους αντίστοιχους γείτονες των αρχικών προτύπων, βάσει των οποίων δημιουργήθηκαν τα συνθετικά. Η προεπιλογή για το k είναι να χρησιμοποιεί τους πέντε κοντινότερους γείτονες.

Ανάλογα με το ποσοστό υπερδειγματοληψίας που απαιτείται να γίνει (δηλαδή ανάλογα με τον αριθμό των συνθετικών προτύπων που πρέπει να προστεθούν, ώστε να αυξηθεί το σύνολο των προτύπων της μειοψηφικής κλάσης), επιλέγονται τυχαία κάποιοι από τους k κοντινότερους γείτονες του εξεταζόμενου προτύπου (Εικόνα 18 - δ). Για παράδειγμα, αν το ποσοστό της απαιτούμενης υπερδειγματοληψίας είναι 400%, επιλέγονται μόνο τέσσερις από τους πέντε κοντινότερους γείτονες του εξεταζόμενου προτύπου και παράγεται ένα συνθετικό πρότυπο για τον κάθε ένα από τους τέσσερις κοντινότερους αυτούς γείτονες.

Τα συνθετικά πρότυπα παράγονται με τον ακόλουθο τρόπο: Υπολογίζεται η διαφορά μεταξύ του διανύσματος χαρακτηριστικών του προτύπου που εξετάζεται (Π1) και του διανύσματος του κοντινότερου γείτονα του (Π6) (Εικόνα 18 - ε, στ). Η διαφορά αυτή (δηλαδή το καινούργιο διάνυσμα προτύπου που προέκυψε) πολλαπλασιάζεται με έναν τυχαίο αριθμό (Τ) μεταξύ 0 και 1 και προκύπτει ένα τυχαίο ενδιάμεσο διάνυσμα (Εικόνα 18 - ζ). Το τυχαίο ενδιάμεσο διάνυσμα που προκύπτει προστίθεται στο διάνυσμα χαρακτηριστικών του προτύπου που εξετάζεται (δηλαδή στο αρχικό πρότυπο) και έτσι παράγεται το διάνυσμα ενός νέου συνθετικού προτύπου (d) (Εικόνα 18 - η). Αυτό έχει ως αποτέλεσμα την επιλογή ενός

τυχαίου σημείου (d) κατά μήκος του ευθύγραμμου τμήματος μεταξύ των δύο παραπάνω προτύπων, καθώς ο τυχαίος αριθμός μεταξύ 0 και 1, αντιπροσωπεύει το ποσοστό της απόστασης μεταξύ τους. Αυτό σημαίνει ότι η τιμή κάθε χαρακτηριστικού του νέου συνθετικού προτύπου (d) θα είναι ανάμεσα στις τιμές χαρακτηριστικών των δύο προαναφερθέντων προτύπων. Σε αντίθετη περίπτωση αν ο τυχαίος αριθμός ήταν μεγαλύτερος του 1, το νέο συνθετικό πρότυπο (d) δεν θα τοποθετούνταν ανάμεσα στο αρχικό πρότυπο και τον γείτονα του, οπότε θα είχε μικρό βαθμό ομοιότητας και με τους δύο. Στην Εικόνα 18 περιγράφεται αναλυτικά η διαδικασία δημιουργίας συνθετικών προτύπων με εφαρμογή του αλγορίθμου SMOTE.



(α) Αρχικό σύνολο δεδομένων

(β) Επιλογή προτύπου Π1

(γ) Εντοπισμός Κ κοντινότερων γειτόνων του προτύπου Π1

(δ) Επιλογή Ν/100 προτύπων από τους Κ κοντινότερους γείτονες του προτύπου Π1

(ε) Υπολογισμός των διανυσμάτων χαρακτηριστικών των προτύπων Π1 και Π6

(στ) Υπολογισμός διαφοράς διανυσμάτων χαρακτηριστικών των προτύπων Π1 και Π6

(ζ) Υπολογισμός τυχαίου ενδιάμεσου διανύσματος

(η) Πρόσθεση τυχαίου ενδιάμεσου διανύσματος, στο διάνυσμα του αρχικού προτύπου (Π1)

(θ) Δεδομένα ΜΕΤΑ από την εφαρμογή του αλγορίθμου SMOTE για το πρότυπο Π1

(ι) Σύνολο δεδομένων ΜΕΤΑ από την εφαρμογή του αλγορίθμου SMOTE σε όλα τα πρότυπα της μειοψηφικής κλάσης

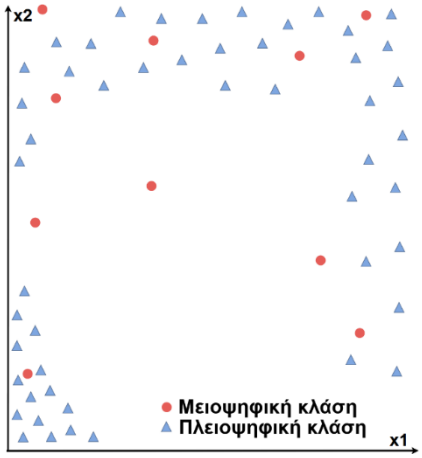
Εικόνα 18: Αλγόριθμος SMOTE

Με τον ίδιο τρόπο παράγονται και τα νέα συνθετικά πρότυπα για τους υπόλοιπους επιλεγμένους γείτονες (Εικόνα 18 - θ). Η παραπάνω διαδικασία επαναλαμβάνεται και για τα υπόλοιπα πρότυπα της μειοψηφικής κλάσης. Έτσι πλέον οι δύο κλάσεις είναι ισορροπημένες (Εικόνα 18 - ι) (Chawla et al., 2002).

Το πλεονέκτημα της τεχνικής SMOTE είναι ότι το πλήθος των προτύπων της μειοψηφικής κλάσης αυξάνεται με τη δημιουργία "συνθετικών" προτύπων και όχι με απλή αντιγραφή προτύπων στο σύνολο δεδομένων. Αυτό έχει ως αποτέλεσμα την μείωση του φαινομένου της υπερμοντελοποίησης κατά την εκπαίδευση του κατηγοριοποιητή, γιατί τα νέα συνθετικά πρότυπα που προστίθενται στο σύνολο δεδομένων είναι διαφορετικά από τα αρχικά και όχι απλά αντίγραφα τους. Με αυτό τον τρόπο οι κανόνες που θα δημιουργήσει ο κατηγοριοποιητής δεν θα είναι εξειδικευμένοι για τα συνθετικά πρότυπα αλλά πιο γενικοί, αυξάνοντας έτσι την ικανότητα γενίκευσης του.

5.4.1 Βήματα αλγορίθμου SMOTE

Ο αλγόριθμος **SMOTE** εκτελείται σε τέσσερα (4) βήματα, τα οποία επεξηγούνται διεξοδικά παρακάτω. Για την καλύτερη κατανόηση του αλγορίθμου χρησιμοποιείται ένα σύνολο δεδομένων με ενδεικτικά πρότυπα, όπως φαίνονται στην Εικόνα 19. Το σύνολο δεδομένων (60 πρότυπα) αποτελείται από πρότυπα που ανήκουν σε 2 κλάσεις: ΚΟΚΚΙΝΟΣ ΚΥΚΛΟΣ (10 πρότυπα) και ΜΠΛΕ ΤΡΙΓΩΝΟ (50 πρότυπα). Από την Εικόνα 19 φαίνεται ότι η μειοψηφική κλάση, είναι η κλάση ΚΟΚΚΙΝΟΣ ΚΥΚΛΟΣ και η πλειοψηφική κλάση είναι η κλάση ΜΠΛΕ ΤΡΙΓΩΝΟ. Ο στόχος είναι να αυξηθεί το πλήθος των προτύπων της κλάσης ΚΟΚΚΙΝΟΣ ΚΥΚΛΟΣ, δηλαδή να εφαρμοστεί τεχνική υπερδειγματοληψίας στην κλάση ΚΟΚΚΙΝΟΣ ΚΥΚΛΟΣ έτσι ώστε το σύνολο δεδομένων να γίνει πλέον ισορροπημένο.



Εικόνα 19: Δεδομένα ΠΡΙΝ το SMOTE

Ο αλγόριθμος SMOTE έχει 2 βασικές παραμέτρους, το K και το N. Το K είναι ο αριθμός των κοντινότερων γειτόνων των προτύπων της μειοψηφικής κλάσης που θα εξεταστούν, έτσι ώστε στη συνέχεια ο αλγόριθμος να δημιουργήσει συνθετικά πρότυπα. Η παράμετρος K επιλέγεται από τον χρήστη και μπορεί να πάρει ακέραιες τιμές K>=1. Για μικρές τιμές του K ο αλγόριθμος θα παράγει συνθετικά πρότυπα που θα είναι αρκετά όμοια με το αρχικά επιλεγμένο πρότυπο. Όσο αυξάνεται το K, αυξάνεται και η πιθανότητα να δημιουργηθούν και συνθετικά πρότυπα που θα είναι λιγότερο όμοια με το αρχικά επιλεγμένο πρότυπο. Γενικά η τιμή που χρησιμοποιείται συνήθως είναι K=5.

Η δεύτερη παράμετρος που είναι το N εκφράζει το απαιτούμενο ποσοστό αύξησης της μειοψηφικής κλάσης, ώστε το σύνολο δεδομένων να γίνει πλέον ισορροπημένο. Για παράδειγμα στο παραπάνω σύνολο δεδομένων Εικόνα 19, στο οποίο η πλειοψηφική κλάση περιλαμβάνει 50 πρότυπα και η μειοψηφική 10, πρέπει να αυξηθεί η μειοψηφική κλάση κατά 40 πρότυπα. Άρα το ποσοστό αύξησης θα είναι:

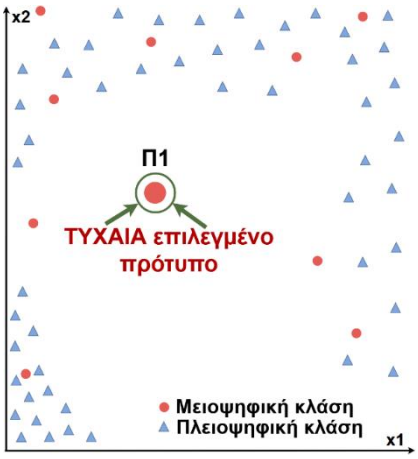
$$N = \frac{Αύξηση\ μειοψηφικής\ κλάσης}{Αρχικό\ πλήθος\ μειοψηφικής\ κλάσης} \cdot 100 = \frac{50-10}{10} \cdot 100 = 400\%$$

Για το συγκεκριμένο λοιπόν σύνολο δεδομένων, επιλέχθηκαν οι τιμές K=5 και N= 400%. Στη συνέχεια με βάση το παραπάνω παράδειγμα εφαρμόζονται τα 5 βήματα του αλγορίθμου SMOTE.

1ο ΒΗΜΑ

Επιλέγεται **ΤΥΧΑΙΑ** ένα από τα πρότυπα της μειοψηφικής κλάσης.

Για παράδειγμα όπως φαίνεται στην Εικόνα 20, επιλέγεται το πρότυπο Π1.

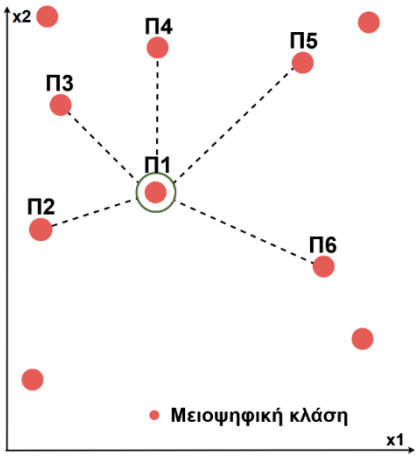


Εικόνα 20: Επιλογή προτύπου

2ο ΒΗΜΑ

Εντοπίζονται οι Κ κοντινότεροι γείτονες του προτύπου που επιλέχθηκε στο προηγούμενο βήμα. Ο εντοπισμός των κοντινότερων γειτόνων γίνεται με υπολογισμό της Ευκλείδειας απόστασης μεταξύ του αρχικά επιλεγμένου προτύπου (Π1) και των υπολοίπων προτύπων που ανήκουν στην μειοψηφική κλάση ξεχωριστά.

Οι 5 κοντινότεροι γείτονες για το πρότυπο Π1 είναι τα πρότυπα Π2, Π3, Π4, Π5, Π6 (Εικόνα 21).



Εικόνα 21: Εντοπισμός Κ κοντινότερων γειτόνων του Π1

3ο ΒΗΜΑ

Από τους Κ κοντινότερους γείτονες που εντοπίστηκαν, επιλέγονται με τυχαίο τρόπο οι Ν/100 γείτονες. Υπάρχει πιθανότητα ο ίδιος γείτονας να επιλεγεί περισσότερες από μία (1) φορές.

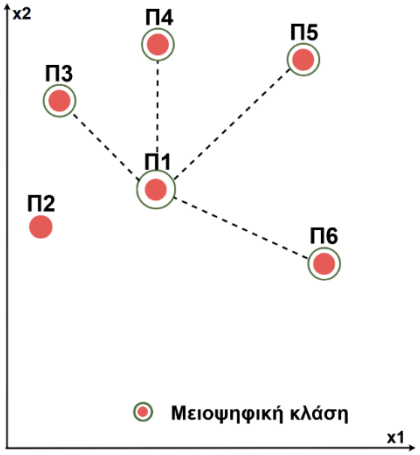
Για κάθε κοντινότερο γείτονα που θα επιλεγεί, θα δημιουργηθεί ένα νέο συνθετικό πρότυπο. Άρα για κάθε πρότυπο της μειοψηφικής κλάσης, θα δημιουργηθούν Ν/100 νέα συνθετικά πρότυπα. Αυτό ισχύει για τιμές του N>=100. Αντίθετα για τιμές του N<100 αναγκαστικά θα πρέπει να επιλεγούν με τυχαίο τρόπο, ορισμένα μόνο από τα πρότυπα της μειοψηφικής κλάσης και να δημιουργηθεί ένα νέο συνθετικό πρότυπο μόνο για το καθένα από αυτά τα πρότυπα και όχι για όλα τα πρότυπα που ανήκουν στη μειοψηφική κλάση.

Παράδειγμα μεταβολής της παραμέτρου Ν και πώς αυτή μεταβάλει το μέγεθος της μειοψηφικής κλάσης παρουσιάζεται στη συνέχεια (Πίνακας 45).

ΠΑΡΑΜΕΤΡΟΣ Ν	ΑΡΙΘΜΟΣ ΓΕΙΤΟΝΩΝ ΠΟΥ ΕΠΙΛΕΓΟΝΤΑΙ (Ν/100 γείτονες)	ΤΕΛΙΚΟ ΠΛΗΘΟΣ ΜΕΙΟΨΗΦΙΚΗΣ ΚΛΑΣΗΣ
100%	100/100 = 1	Διπλασιάζεται (x2)
200%	200/100 = 2	Τριπλασιάζεται (x3)
300%	300/100 = 3	Τετραπλασιάζεται (x4)
400%	400/100 = 4	Πενταπλασιάζεται (x5)
500%	500/100 = 5	Εξαπλασιάζεται (x6)

Πίνακας 45: Μεταβολή μεγέθους μειοψηφικής κλάσης ανάλογα με την παράμετρο Ν

Στο παράδειγμα που φαίνεται στην Εικόνα 19, στόχος είναι να αυξηθεί το πλήθος της μειοψηφικής κλάσης (ΚΟΚΚΙΝΟΣ ΚΥΚΛΟΣ) κατά 400% και για το λόγο αυτό επιλέγεται η τιμή N = 400. Αυτό έχει ως αποτέλεσμα να επιλεχθούν 400/100 = 4 κοντινότεροι γείτονες (Π3, Π4, Π5, Π6) του προτύπου Π1 στο βήμα αυτό, όπως φαίνεται στην Εικόνα 22.

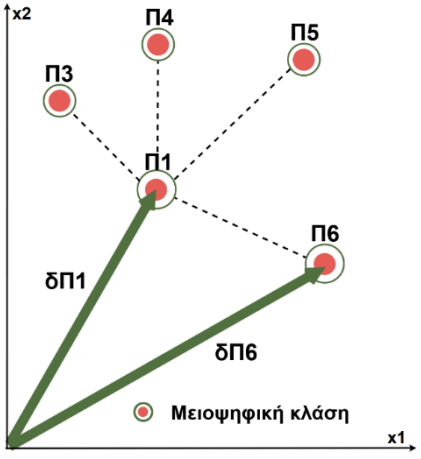


Εικόνα 22: Επιλογή N/100 προτύπων από τους K κοντινότερους γείτονες του Π1

4ο ΒΗΜΑ

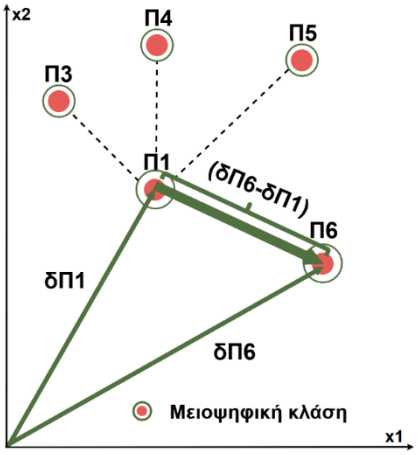
Για κάθε ένα γείτονα από αυτούς που επιλέχθηκαν στο **3ο ΒΗΜΑ**:

Α) Υπολογίζεται το διάνυσμα χαρακτηριστικών (feature vector) του αρχικού προτύπου και το διάνυσμα χαρακτηριστικών του κοντινότερου γείτονα που επιλέχθηκε (Εικόνα 23).



Εικόνα 23: Υπολογισμός των διανυσμάτων χαρακτηριστικών των προτύπων Π1 και Π6

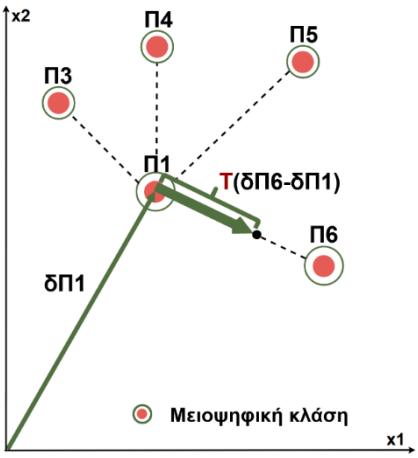
Β) Έπειτα υπολογίζεται η διαφορά των διανυσμάτων χαρακτηριστικών τους (Εικόνα 24).



Εικόνα 24: Υπολογισμός διαφοράς διανυσμάτων χαρακτηριστικών των προτύπων Π1 και Π6

Γ) Παράγεται (από το σύστημα) ένας τυχαίος αριθμός Τ, μεταξύ 0 και 1 (π.χ. 0,328) (Εικόνα 25).

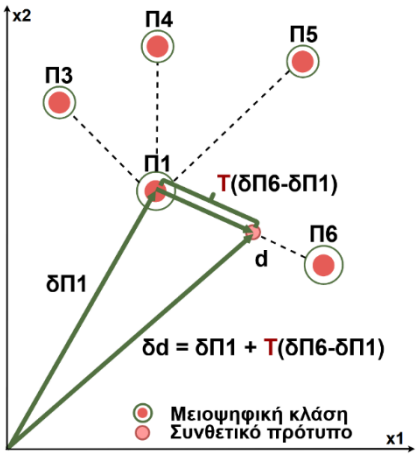
Δ) Ο τυχαίος αριθμός Τ πολλαπλασιάζεται με την διαφορά των διανυσμάτων χαρακτηριστικών και προκύπτει ένα καινούργιο τυχαίο ενδιάμεσο διάνυσμα με μήκος μικρότερο ή ίσο από την αρχική διαφορά των διανυσμάτων (Εικόνα 25).



Εικόνα 25: Υπολογισμός τυχαίου ενδιάμεσου διανύσματος

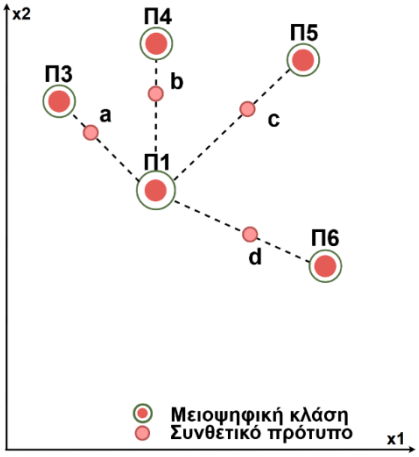
Όπως φαίνεται στην Εικόνα 25 το καινούργιο τυχαίο ενδιάμεσο διάνυσμα θα έχει ως αρχή του το αρχικά επιλεγμένο πρότυπο Π1, ενώ το τέλος του θα είναι ένα τυχαίο σημείο ανάμεσα στα πρότυπα Π1 και Π6.

Ε) Το καινούργιο τυχαίο ενδιάμεσο διάνυσμα που προέκυψε στο προηγούμενο βήμα, προστίθεται στο διάνυσμα του αρχικού προτύπου. Το αποτέλεσμα που προκύπτει, είναι το διάνυσμα ενός νέου συνθετικού (δηλαδή εικονικού) προτύπου d όπως φαίνεται στην Εικόνα 26 .



Εικόνα 26: Πρόσθεση τυχαίου ενδιάμεσου διανύσματος στο διάνυσμα του αρχικού προτύπου

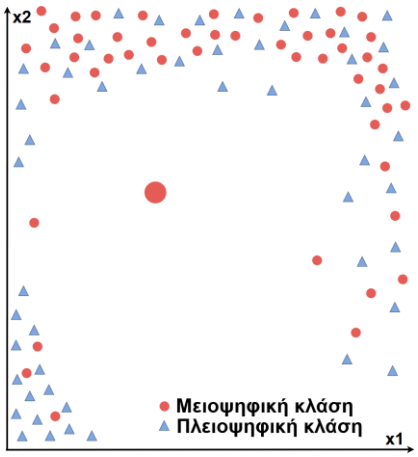
Το 4ο ΒΗΜΑ επαναλαμβάνεται και για τους υπόλοιπους 3 από τους 4 γείτονες που επιλέχθηκαν στο 3ο ΒΗΜΑ. Στην Εικόνα 27 απεικονίζονται διαγραμματικά τα νέα συνθετικά πρότυπα που δημιουργήθηκαν. Μεταξύ του προτύπου Π1 και Π3 δημιουργήθηκε το νέο συνθετικό πρότυπο a. Αντίστοιχα μεταξύ του Π1 και των Π4, Π5 και Π6 δημιουργήθηκαν τα νέα συνθετικά πρότυπα b, c, d.



Εικόνα 27: Δεδομένα ΜΕΤΑ από την εφαρμογή του αλγορίθμου SMOTE για το πρότυπο Π1

5ο ΒΗΜΑ

Τα ΒΗΜΑΤΑ 1 έως 4 επαναλαμβάνονται μέχρις ότου η διαδικασία να εκτελεστεί για όλα τα πρότυπα της μειοψηφικής κλάσης (εκτός αν N<100, οπότε σε αυτή την περίπτωση θα εκτελεστεί για όσα πρότυπα απαιτείται με βάση το N). Με την ολοκλήρωση του 5ου βήματος, το αρχικό σύνολο δεδομένων είναι πλέον ισορροπημένο. Κατά συνέπεια και οι δύο κλάσεις (ΤΡΙΓΩΝΑ-ΚΥΚΛΟΙ) περιέχουν ισάριθμα πρότυπα, όπως φαίνεται στην Εικόνα 28.



Εικόνα 28: Σύνολο δεδομένων ΜΕΤΑ από την εφαρμογή του αλγορίθμου SMOTE σε όλα τα πρότυπα της μειοψηφικής κλάσης

Στη συνέχεια παρατίθεται ένα απλό παράδειγμα για την αμεσότερη κατανόηση λειτουργίας του συγκεκριμένου αλγορίθμου.

5.4.2 Εφαρμογή της τεχνικής SMOTE σε Numeric χαρακτηριστικά

Θα εφαρμοστούν τα βήματα του αλγορίθμου SMOTE που περιεγράφηκαν προηγουμένως σε 16 ενδεικτικά πρότυπα του συνόλου δεδομένων Alvaro Flores (1.Alvaro Flores_ARXIKO_mono_d). Τα πρότυπα επιλέχτηκαν τυχαία και φαίνονται στον πίνακα που ακολουθεί (Πίνακας 46). Επίσης, χάριν παραδείγματος, επιλέχθηκαν τυχαία 4 από τα χαρακτηριστικά του συνόλου δεδομένων για να μειωθεί το πλήθος των απαιτούμενων υπολογισμών. Όλα τα χαρακτηριστικά που επιλέχθηκαν είναι τύπου numeric.

Τα 6 από τα πρότυπα αυτού του συνόλου δεδομένων ανήκουν στην κλάση Yes, ενώ τα υπόλοιπα 10 στην κλάση No. Άρα η μειοψηφική κλάση σε αυτή την περίπτωση είναι η κλάση Yes και θα πρέπει να αυξηθεί το πλήθος των προτύπων της εφαρμόζοντας τον αλγόριθμο SMOTE.

		Numeric	*Numeric*	*Numeric*	*Numeric*	*Κλάση*
A /A	**Πρότυπο**	**age**	**appointment_hour_d**	**scheduled_hour_d**	**waiting_days**	**No_show**
1226		34.0	8.0	10.0	5.0	No
4557		26.0	9.0	1.0	1.0	No
5568		37.0	9.0	11.0	10.0	No
14455		6.0	11.0	23.0	3.0	No
16684		37.0	11.0	12.0	9.0	No
23178		57.0	12.0	10.0	19.0	No
24899		18.0	13.0	10.0	25.0	No
26918		1.0	14.0	9.0	0.0	No
33773		21.0	15.0	11.0	18.0	No

36089		46.0	15.0	18.0	5.0	No
1000	Α (Αρχικό)	30.0	8.0	17.0	2.0	Yes
5428	Β	36.0	9.0	12.0	4.0	Yes
5157	Γ	34.0	9.0	18.0	1.0	Yes
1497	Δ	38.0	8.0	21.0	1.0	Yes
6365	Ε	46.0	9.0	16.0	4.0	Yes
4469	ΣΤ	25.0	10.0	9.0	0.0	Yes

Πίνακας 46: Πραγματικά δεδομένα μόνο με Numeric χαρακτηριστικά

Για να γίνει λοιπόν το σύνολο ισορροπημένο θα πρέπει να δημιουργηθούν 10 – 6 = 4 συνθετικά πρότυπα της κλάσης Yes. Άρα το ποσοστό αύξησης θα είναι:

$$N = \frac{\text{αύξηση μειοψηφικής κλάσης}}{\text{αρχικό πλήθος μειοψηφικής κλάσης}} \cdot 100 = \frac{4}{6} \cdot 100 = 66{,}7\%$$

Σε αυτήν την περίπτωση που το ποσοστό αύξησης είναι λιγότερο από 100%, σημαίνει ότι δεν θα πρέπει να δημιουργηθούν συνθετικά πρότυπα για κάθε ένα από τα 6 πρότυπα της μειοψηφικής κλάσης, αλλά μόνο για τα 4 από αυτά με τυχαία επιλογή.

Επίσης για το συγκεκριμένο παράδειγμα η τιμή που επιλέγεται για το K είναι K= 3, δηλαδή ο αλγόριθμος SMOTE για κάθε πρότυπο της μειοψηφικής κλάσης θα εντοπίσει τους 3 κοντινότερους γείτονες του. Από τους 3 κοντινότερους γείτονες θα επιλέξει $\frac{N}{100} = \frac{66{,}7}{100} = 0{,}667$, δηλαδή με στρογγυλοποίηση 1 γείτονα από αυτούς για να δημιουργήσει ένα νέο συνθετικό πρότυπο κάθε φορά. Τα βήματα του αλγορίθμου SMOTE που περιγράφονται στη συνέχεια θα επαναληφθούν 4 φορές.

Βήμα 1: Επιλογή προτύπου με τυχαίο τρόπο

Έστω ότι επιλέγεται με τυχαίο τρόπο το πρότυπο Α.

Βήμα 2: Εύρεση κοντινότερων γειτόνων του επιλεγμένου προτύπου

Υπολογισμός της Ευκλείδειας απόστασης μεταξύ του προτύπου Α και των υπόλοιπων προτύπων.

Η Ευκλείδεια απόσταση μεταξύ δύο προτύπων x και y, με διανύσματα χαρακτηριστικών $\boldsymbol{x} = \{x_1, x_2, \cdots, x_n\}$ και $\boldsymbol{y} = \{y_1, y_2, \cdots, y_n\}$, όπου n είναι το πλήθος των χαρακτηριστικών, δίνεται από τον τύπο:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2} = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2 + \cdots + (y_n - x_n)^2}$$

Έτσι προκύπτουν οι παρακάτω αποστάσεις:

$$d_{(A,B)} = \sqrt{(36-30)^2 + (9-8)^2 + (12-17)^2 + (4-2)^2} = \sqrt{6^2 + 1^2 + (-5)^2 + 2^2}$$
$$= \sqrt{36 + 1 + 25 + 4} = \sqrt{66} \cong 8{,}12$$

$$d_{(A,\Gamma)} = \sqrt{(34-30)^2 + (9-8)^2 + (18-17)^2 + (1-2)^2} = \sqrt{4^2 + 1^2 + 1^2 + (-1)^2}$$
$$= \sqrt{16 + 1 + 1 + 1} = \sqrt{19} \cong 4{,}36$$

$$d_{(A,\Delta)} = \sqrt{(38-30)^2 + (8-8)^2 + (21-17)^2 + (1-2)^2} = \sqrt{8^2 + 0^2 + 4^2 + (-1)^2}$$
$$= \sqrt{64 + 16 + 1} = \sqrt{81} = 9$$

$$d_{(A,E)} = \sqrt{(46-30)^2 + (9-8)^2 + (16-17)^2 + (4-2)^2} = \sqrt{16^2 + 1^2 + (-1)^2 + 2^2}$$
$$= \sqrt{256 + 1 + 1 + 4} = \sqrt{262} \cong 16{,}19$$

$$d_{(A,\Sigma T)} = \sqrt{(25-30)^2 + (10-8)^2 + (9-17)^2 + (0-2)^2} = \sqrt{(-5)^2 + 2^2 + (-8)^2 + (-2)^2}$$
$$= \sqrt{25 + 4 + 64 + 4} = \sqrt{97} \cong 9{,}85$$

Συγκρίνοντας τις αποστάσεις μεταξύ τους εντοπίζονται οι 3 μικρότερες που αντιστοιχούν στους 3 κοντινότερους γείτονες του προτύπου Α (Πίνακας 47)

Αριθμός Γραμμής	Πρότυπο	Απόσταση από πρότυπο Α	Κοντινότεροι Γείτονες στο Α
5157	Γ	4,36	1ος
5428	Β	8,12	2ος
1497	Δ	9	3ος
4469	ΣΤ	9,85	-
6365	Ε	16,19	-

Πίνακας 47: Αποστάσεις προτύπων από το Α (κατά αύξουσα σειρά)

Βήμα 3: Επιλογή κοντινότερων γειτόνων

Από τους 3 κοντινότερους γείτονες του προτύπου Α που εντοπίστηκαν στο προηγούμενο βήμα, επιλέγονται με τυχαίο τρόπο οι Ν/100, δηλαδή 66,7/100=0,667 ≅ 1 κοντινότερος γείτονας. Έστω λοιπόν ότι ο 1 κοντινότερος γείτονας που επιλέγεται από τους 3 και μάλιστα με τυχαίο τρόπο, είναι το πρότυπο Β.

Βήμα 4: **Δημιουργία συνθετικού προτύπου**

α) Υπολογισμός διανυσμάτων χαρακτηριστικών των προτύπων Α και Β. Το πρότυπο Α είναι το αρχικό πρότυπο που εξετάζεται, ενώ το πρότυπο Β είναι ο κοντινότερος γείτονας που επιλέχθηκε με τυχαίο τρόπο στο προηγούμενο βήμα.

Διάνυσμα A $= [30, 8, 17, 2]$

Διάνυσμα B $= [36, 9, 12, 4]$

β) Υπολογισμός διαφοράς διανυσμάτων χαρακτηριστικών (Διάνυσμα γείτονα – αρχικό διάνυσμα, άρα διάνυσμα B – διάνυσμα A)

Διαφορά Διανυσμάτων $=$ διάνυσμα B – διάνυσμα A $= [36 - 30, 9 - 8, 12 - 17, 4 - 2]$

$= [6, 1, -5, 2]$

γ) παραγωγή ενός τυχαίου αριθμού μεταξύ 0 και 1, π.χ.:

T $= 0.31415$ (παράγεται από το σύστημα)

δ) πολλαπλασιασμός του τυχαίου αριθμού T με την διαφορά των διανυσμάτων που υπολογίστηκε προηγουμένως. Προκύπτει ένα τυχαίο ενδιάμεσο διάνυσμα, το οποίο στη συνέχεια προστίθεται στο διάνυσμα του αρχικού προτύπου (δηλαδή στο διάνυσμα A) και το αποτέλεσμα είναι το διάνυσμα του νέου συνθετικού προτύπου

Πολλαπλασιασμός με τυχαίο αριθμό T

τυχαίο ενδιάμεσο Διάνυσμα $=$ T x διαφορά Διανυσμάτων $\Rightarrow$

τυχαίο ενδιάμεσο Διάνυσμα $= 0.31415 \cdot [6, \quad 1, \quad -5, \quad 2] =$

$[0.31415 \text{ x } 6, \quad 0.31415 \text{ x } 1, \quad 0.31415 \text{ x } (-5), \quad 0.31415 \text{ x } 2] =$

$[1.8849, \quad 0.31415, \quad -1.57075, \quad 0.6283]$

Πρόσθεση διανυσμάτων

Διάνυσμα νέου συνθετικού προτύπου $=$ διάνυσμα A $+$ τυχαίο ενδιάμεσο Διάνυσμα $=$

$[30, 8, 17, 2] + [1.8849, \quad 0.31415, \quad -1.57075, \quad 0.6283] =$

$[30 + 1.8849, 8 + 0.31415, \quad 17 - 1.57075, \quad 2 + 0.6283]$

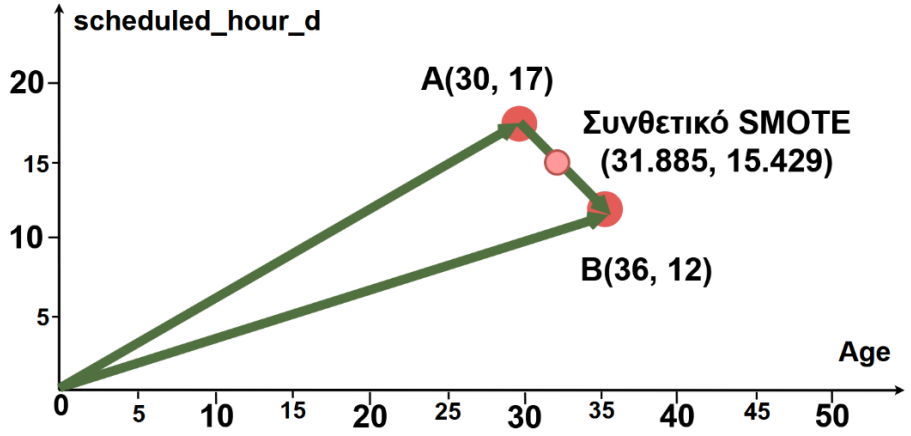
$= [31.8849, 8.31415, 15.42925, 2.6283]$

Άρα τελικά προκύπτουν τα παρακάτω αποτελέσματα (Πίνακας 48):

		Numeric	*Numeric*	*Numeric*	*Numeric*
A / A	**Πρότυπο**	**age**	**appointment_hour_d**	**scheduled_hour_d**	**waiting_days**
1000	A (Αρχικό)	30.0	8.0	17.0	2.0
5428	B (Κοντινότερος Γείτονας)	36.0	8.0	12.0	4.0
	Νέο Συνθετικό πρότυπο SMOTE	31.8849	8.31415	15.42925	2.6283

Πίνακας 48: Πραγματικά και συνθετικά δεδομένα μετά την εφαρμογή SMOTE

Το νέο (συνθετικό) πρότυπο μπορεί να αναπαρασταθεί γραφικά αν 2 από τα χαρακτηριστικά του χρησιμοποιηθούν ως συντεταγμένες. Για παράδειγμα στον οριζόντιο άξονα χρησιμοποιείται το χαρακτηριστικό age και στον κάθετο άξονα το scheduled_hour_d (Εικόνα 29).



Εικόνα 29: Γραφική αναπαράσταση της δημιουργίας του νέου συνθετικού προτύπου κατά την εφαρμογή του SMOTE

5.5 Επεκτάσεις του SMOTE για Nominal δεδομένα

Ο αλγόριθμος SMOTE που περιγράφηκε προηγουμένως δεν μπορεί να χειριστεί κατηγορικά χαρακτηριστικά, αφού αυτά δεν αποτελούν αριθμητικά δεδομένα ώστε να συμπεριληφθούν στον υπολογισμό της ευκλείδειας απόστασης. Με άλλα λόγια, ο κλασικός αλγόριθμος SMOTE αγνοεί τελείως τα κατηγορικά χαρακτηριστικά. Για το λόγο αυτό αναπτύχθηκαν επεκτάσεις του αλγορίθμου με δυνατότητα επεξεργασίας κατηγορικών χαρακτηριστικών, όπως το SMOTE-N. Το SMOTE-N χρησιμοποιείται από το WEKA και περιγράφεται αναλυτικά στη συνέχεια.

5.5.1 SMOTE – N

Το SMOTE-N είναι επέκταση του αλγορίθμου SMOTE και χρησιμοποιείται για να υποστηρίζει κατηγορικά χαρακτηριστικά (SMOTE-Nominal) (Chawla et al., 2002). Η διαφορά του με το SMOTE είναι ότι για τον υπολογισμό των κοντινότερων γειτόνων χρησιμοποιεί την τροποποιημένη έκδοση της μετρικής διαφοράς τιμών Value Difference Metric (Stanfill & Waltz, 1986) που προτάθηκε από τους (Cost & Salzberg, 1993).

Η μετρική VDM μετράει την απόσταση μεταξύ δύο τιμών ενός κατηγορικού χαρακτηριστικού. Για να το πετύχει αυτό, υπολογίζει την συχνότητα εμφάνισης αυτών των δύο τιμών σε κάθε κλάση. Όταν οι δύο τιμές παρουσιάζουν παρόμοιες συχνότητες εμφάνισης (παρόμοια ποσοστά), τότε η VDM υποθέτει ότι οι δύο αυτές τιμές είναι πιο κοντά μεταξύ τους. Επομένως έχουν μεγάλη ομοιότητα. Αντίθετα αν οι δύο τιμές έχουν μεγάλη διαφορά

στις συχνότητες εμφάνισης τους (μεγάλη διαφορά στα ποσοστά τους), η VDM υποθέτει ότι έχουν μεγάλη απόσταση μεταξύ τους. Κατά συνέπεια δεν έχουν μεγάλη ομοιότητα.

Χρησιμοποιώντας την VDM δημιουργείται ένας πίνακας που περιέχει την απόσταση μεταξύ των αντίστοιχων τιμών κατηγορικών χαρακτηριστικών για όλα τα διανύσματα κατηγορικών χαρακτηριστικών. Η απόσταση δ, δηλαδή η απόσταση VDM, μεταξύ δύο αντίστοιχων τιμών ενός κατηγορικού χαρακτηριστικού ορίζεται ως εξής:

$$\delta(V_1, V_2) = \sum_{i=1}^{n} \left| \frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2} \right|^k \qquad (1)$$

Στην παραπάνω εξίσωση, τα V_1 και V_2 είναι οι δύο αντίστοιχες τιμές ενός κατηγορικού χαρακτηριστικού. Το C_1 είναι ο συνολικός αριθμός εμφανίσεων της τιμής V_1 και C_{1i} είναι ο αριθμός των εμφανίσεων της τιμής V_1 για την κλάση i. Μια παρόμοια σύμβαση μπορεί επίσης να εφαρμοστεί για τα C_{2i} και C_2. Το k είναι μια σταθερά που συνήθως έχει τιμή ίση με 1. Το n είναι ο αριθμός των κλάσεων των δεδομένων. Αυτή η εξίσωση χρησιμοποιείται για τον υπολογισμό του πίνακα των διαφορών των τιμών για κάθε κατηγορικό χαρακτηριστικό μέσα στο σύνολο των διανυσμάτων χαρακτηριστικών. Τελικά, η απόσταση μεταξύ δύο διανυσμάτων χαρακτηριστικών X και Y, δίνεται από την εξίσωση:

$$\Delta(X, Y) = w_x w_y \sum_{i=1}^{N} d(x_i, y_i)^r \qquad (2)$$

Το r παίρνει την τιμή 1 για την απόσταση Μανχάταν και την τιμή 2 για την ευκλείδεια απόσταση. Τα w_x και w_y είναι τα βάρη των προτύπων στην τροποποιημένη VDM. Όμως για την περίπτωση του SMOTE-N αυτά τα βάρη στην εξίσωση (2) μπορούν να αγνοηθούν, καθώς το SMOTE-N δεν χρησιμοποιείται άμεσα για σκοπούς κατηγοριοποίησης.

Για την δημιουργία των νέων συνθετικών διανυσμάτων χαρακτηριστικών για τη μειοψηφική κλάση, ως τιμή για κάθε κατηγορικό χαρακτηριστικό τίθεται η τιμή που έχει η πλειοψηφία του αρχικού προτύπου και των k-κοντινότερων γειτόνων του. Ο Πίνακας 49 δείχνει ένα παράδειγμα δημιουργίας ενός συνθετικού διανύσματος κατηγορικών χαρακτηριστικών.

Έστω ότι F1 = [A, B, C, D, E] είναι το διάνυσμα χαρακτηριστικών του εξεταζόμενου προτύπου και τα διανύσματα χαρακτηριστικών των 2 κοντινότερων γειτόνων του είναι:

F2 = [A, F, C, G, N]

F3 = [H, B, C, D, N]

Η σύγκριση των χαρακτηριστικών μεταξύ τους δίνει το νέο συνθετικό διάνυσμα χαρακτηριστικών που θα δημιουργηθεί.

F1	A	B	C	D	E
F2	A	F	C	G	N
F3	H	B	C	D	N
FS (πλειοψηφία)	A	B	C	D	N

Πίνακας 49: Δημιουργία συνθετικού προτύπου με τεχνική SMOTE-N

Άρα το νέο διάνυσμα θα είναι FS=[A, B, C, D, N]

5.5.2 Εφαρμογή σε Numeric και Nominal χαρακτηριστικά (SMOTE-N)

Για το παράδειγμα αυτό χρησιμοποιείται το σύνολο δεδομένων με 16 πρότυπα του προηγούμενου παραδείγματος (*5.4.2 Εφαρμογή της τεχνικής SMOTE σε Numeric* χαρακτηριστικά*)* από το σύνολο δεδομένων του Alvaro Flores (1.Alvaro Flores_ARXIKO_mono_d) (Πίνακας 50). Η διαφορά εδώ είναι ότι λαμβάνονται υπόψη και δύο κατηγορικά χαρακτηριστικά, το φύλο (sex) και το κανάλι επικοινωνίας (communication_channel). Όπως και στο προηγούμενο παράδειγμα εφαρμόζεται ο αλγόριθμος SMOTE με k =3 και N=66,7%.

		Numeric	*Numeric*	*Numeric*	*Numeric*	*Nominal*	*Nominal*	*Κλάση*
A /A	**Πρότυπο**	**age**	**appointment_hour_d**	**scheduled_hour_d**	**waiting_days**	**sex**	**communication_channel**	**No_show**
1226		34.0	8.0	10.0	5.0	1	1	No
4557		26.0	9.0	1.0	1.0	2	3	No
5568		37.0	9.0	11.0	10.0	2	1	No
14455		6.0	11.0	23.0	3.0	2	3	No
16684		37.0	11.0	12.0	9.0	1	2	No
23178		57.0	12.0	10.0	19.0	1	1	No
24899		18.0	13.0	10.0	25.0	2	2	No
26918		1.0	14.0	9.0	0.0	2	3	No
33773		21.0	15.0	11.0	18.0	1	1	No
36089		46.0	15.0	18.0	5.0	2	2	No
1000	Α (Αρχικό)	30.0	8.0	17.0	2.0	1	1	Yes
5428	Β	36.0	9.0	12.0	4.0	2	3	Yes
5157	Γ	34.0	9.0	18.0	1.0	2	1	Yes
1497	Δ	38.0	8.0	21.0	1.0	1	3	Yes
6365	Ε	46.0	9.0	16.0	4.0	1	1	Yes
4469	ΣΤ	25.0	10.0	9.0	0.0	1	1	Yes

Πίνακας 50: Πραγματικά δεδομένα για εφαρμογή με Numeric και Nominal χαρακτηριστικά (Πριν το SMOTE)

Για τον υπολογισμό των αποστάσεων με βάση τα κατηγορικά χαρακτηριστικά θα πρέπει να κατασκευαστεί ένας πίνακας συχνοτήτων εμφάνισης κάθε δυνατής τιμής των χαρακτηριστικών ανά κλάση. Άρα για το χαρακτηριστικό sex κατασκευάζεται ο Πίνακας 51.

Χαρακτηριστικό	**Συχνότητες εμφάνισης**			**Ποσοστά**	
sex	**Σύνολο δεδομένων (C)**	**Κλάση No (C_{No})**	**Κλάση Yes (C_{Yes})**	**Κλάση No $\frac{C_{No}}{C}$**	**Κλάση Yes $\frac{C_{Yes}}{C}$**
sex = 1	8	4	4	$\frac{4}{8} = 0{,}5$	$\frac{4}{8} = 0{,}5$
sex = 2	8	6	2	$\frac{6}{8} = 0{,}75$	$\frac{2}{8} = 0{,}25$

Πίνακας 51: Συχνότητες εμφάνισης χαρακτηριστικού sex

Αντίστοιχα για το χαρακτηριστικό communication_channel κατασκευάζεται ο Πίνακας 52.

Χαρακτηριστικό	**Συχνότητες εμφάνισης**			**Ποσοστά**	
communication_channel	**Σύνολο δεδομένων (C)**	**Κλάση No (C_{No})**	**Κλάση Yes (C_{Yes})**	**Κλάση No $\frac{C_{No}}{C}$**	**Κλάση Yes $\frac{C_{Yes}}{C}$**
communication_channel = 1	8	4	4	$\frac{4}{8} = 0{,}5$	$\frac{4}{8} = 0{,}5$
communication_channel = 2	3	3	0	$\frac{3}{3} = 1$	$\frac{0}{3} = 0$
communication_channel = 3	5	3	2	$\frac{3}{5} = 0{,}6$	$\frac{2}{5} = 0{,}4$

Πίνακας 52: Συχνότητες εμφάνισης χαρακτηριστικού communication_channel

Βήμα 1: Επιλογή προτύπου με τυχαίο τρόπο

Έστω ότι επιλέγεται με τυχαίο τρόπο το πρότυπο Α.

Βήμα 2: Εύρεση κοντινότερων γειτόνων του επιλεγμένου προτύπου

Ο υπολογισμός της απόστασης γίνεται με βάση τον τύπο της Ευκλείδειας απόστασης για τα αριθμητικά χαρακτηριστικά με τη διαφορά ότι για τα κατηγορικά χαρακτηριστικά χρησιμοποιείται η μετρική VDM (Value Distance Metric) που υπολογίζεται με την βοήθεια των πινάκων συχνοτήτων. Έτσι για δύο τιμές ενός κατηγορικού χαρακτηριστικού V_1 και V_2 η VDM μεταξύ τους υπολογίζεται ως εξής:

$$\delta(V_1, V_2) = \sum_{i=1}^{n} \left| \frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2} \right|^k = \left| \frac{C_{1,No}}{C_1} - \frac{C_{2,No}}{C_2} \right|^1 + \left| \frac{C_{1,Yes}}{C_1} - \frac{C_{2,Yes}}{C_2} \right|^1$$

Τελικά η απόσταση μεταξύ των προτύπων θα υπολογιστεί από τον τύπο της Ευκλείδειας απόστασης συμπεριλαμβάνοντας και τις τιμές VDM των κατηγορικών χαρακτηριστικών. Η Ευκλείδεια απόσταση μεταξύ δύο προτύπων x και y, με διανύσματα χαρακτηριστικών $\boldsymbol{x} =$

$\{x_1, x_2, \cdots, x_n, x_{V1}, x_{V2}, \ldots x_{Vk}\}$ και $\boldsymbol{y} = \{y_1, y_2, \cdots, y_n, y_{V1}, y_{V2}, \ldots y_{Vk}\}$, όπου n είναι το πλήθος των αριθμητικών χαρακτηριστικών και k το πλήθος των κατηγορικών χαρακτηριστικών, δίνεται από τον τύπο:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2 + \sum_{i=1}^{k}\delta_k}$$

$$= \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2 + \cdots + (y_n - x_n)^2 + \delta_1 + \delta_2 + \cdots \delta_k}$$

όπου δ είναι η μετρική VDM για κάθε κατηγορικό χαρακτηριστικό που υπολογίστηκε με τον προηγούμενο τύπο.

Έτσι η VDM κάθε κατηγορικού χαρακτηριστικού για τα πρότυπα A και B θα είναι:

$$\delta_{A,B}\left(V_{A,sex}, V_{B,sex}\right) = \delta_{A,B}(sex = 1, sex = 2) = \left|\frac{4}{8} - \frac{6}{8}\right| + \left|\frac{4}{8} - \frac{2}{8}\right|$$

$$= |0{,}5 - 0{,}75| + |0{,}5 - 0{,}25| = 0{,}25 + 0{,}25 = 0{,}5$$

$$\delta_{A,B}\left(V_{A,com\_channel}, V_{B,com\_channel}\right) = \delta_{A,B}(com\_channel = 1, com\_channel = 3)$$

$$= \left|\frac{4}{8} - \frac{3}{5}\right| + \left|\frac{4}{8} - \frac{2}{5}\right| = |0{,}5 - 0{,}6| + |0{,}5 - 0{,}4| = 0{,}1 + 0{,}1 = 0{,}2$$

Η τελική Ευκλείδεια απόσταση μεταξύ των προτύπων A και B θα είναι:

$$d_{(A,B)} = \sqrt{(36-30)^2 + (9-8)^2 + (12-17)^2 + (4-2)^2 + 0{,}5 + 0{,}2}$$

$$= \sqrt{6^2 + 1^2 + (-5)^2 + 2^2 + 0{,}5 + 0{,}2} = \sqrt{36 + 1 + 25 + 4 + 0{,}5 + 0{,}2} = \sqrt{66{,}7}$$

$$\cong 8{,}17$$

Κάνοντας τον υπολογισμό της μετρικής VDM μεταξύ του προτύπου A και των υπόλοιπων προτύπων της κλάσης Yes, προκύπτουν τα παρακάτω αποτελέσματα (Πίνακας 53):

		Nominal	*VDM*	*Nominal*	*VDM*
A / A	**Πρότυπο**	**sex**	$\boldsymbol{\delta_{sex}}$	**communication_channel**	$\boldsymbol{\delta_{communication\_channel}}$
1000	A (Αρχικό)	1		1	
5428	B	2	0,5	3	0,2
5157	Γ	2	0,5	1	0
1497	Δ	1	0	3	0,2
6365	E	1	0	1	0
4469	ΣΤ	1	0	1	0

Πίνακας 53: Αποστάσεις VDM των κατηγορικών χαρακτηριστικών

Ολοκληρώνοντας τους υπολογισμούς για την Ευκλείδεια απόσταση παράγονται τα παρακάτω αποτελέσματα (Πίνακας 54):

Αριθμός Γραμμής	Πρότυπο	Απόσταση από πρότυπο Α	Κοντινότεροι Γείτονες στο Α
5157	Γ	4,42	1ος
5428	Β	8,17	2ος
1497	Δ	9,01	3ος
4469	ΣΤ	9,85	-
6365	Ε	16,19	-

Πίνακας 54: Αποστάσεις προτύπων από το Α (κατά αύξουσα σειρά)

Βήμα 3: Επιλογή κοντινότερων γειτόνων

Από τους 3 κοντινότερους γείτονες που εντοπίστηκαν στο προηγούμενο βήμα, επιλέγονται με τυχαίο τρόπο οι N/100, δηλαδή 66,7/100=0,667 ≅ 1 κοντινότερος γείτονας. Έστω λοιπόν ότι ο 1 κοντινότερος γείτονας που επιλέγεται από τους 3 και μάλιστα με τυχαίο τρόπο, είναι το πρότυπο Β.

Βήμα 4: **Δημιουργία συνθετικού προτύπου**

α) Υπολογισμός διανυσμάτων χαρακτηριστικών (ΜΟΝΟ ΑΠΟ NUMERIC ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ)

$\text{Διάνυσμα Α} = [30, 8, 17, 2]$

$\text{Διάνυσμα Β} = [36, 9, 12, 4]$

β) Υπολογισμός διαφοράς διανυσμάτων χαρακτηριστικών (γείτονας – αρχικό, άρα διάνυσμα Β – διάνυσμα Α)

$$\text{Διαφορά Διανυσμάτων} = \text{διάνυσμα Β} - \text{διάνυσμα Α} = [36-30, 9-8, 12-17, 4-2]$$
$$= [6, 1, -5, 2]$$

γ) παραγωγή ενός τυχαίου αριθμού μεταξύ 0 και 1, π.χ. Τ = 0.71647

δ) πολλαπλασιασμός του τυχαίου αριθμού Τ με την διαφορά των διανυσμάτων που υπολογίστηκε προηγουμένως. Προκύπτει ένα τυχαίο ενδιάμεσο διάνυσμα, το οποίο στη συνέχεια προστίθεται στο διάνυσμα του αρχικού προτύπου (δηλαδή στο διάνυσμα Α) και το αποτέλεσμα είναι το διάνυσμα του νέου συνθετικού προτύπου

<u>Πολλαπλασιασμός με τυχαίο αριθμό Τ</u>

$\text{τυχαίο ενδιάμεσο Διάνυσμα} = \text{Τ x διαφορά Διανυσμάτων} \Rightarrow$

$\text{τυχαίο ενδιάμεσο Διάνυσμα} = 0.71647 \cdot [6, \quad 1, \quad -5, \quad 2] =$

$[0.71647 \text{ x } 6, \quad 0.71647 \text{ x } 1, \quad 0.71647 \text{ x } (-5), \quad 0.71647 \text{ x } 2] =$

$[4.29882, \quad 0.71647, \quad -3.58235, \quad 1.43294]$

Πρόσθεση διανυσμάτων

Διάνυσμα νέου συνθετικού προτύπου $=$ διάνυσμα Α $+$ τυχαίο ενδιάμεσο Διάνυσμα $=$

$[30, 8, 17, 2] + [4.29882, \quad 0.71647, \quad -3.58235, \quad 1.43294] =$

$[30 + 4.29882, 8 + 0.71647, \quad 17 - 3.58235, \quad 2 + 1.43294]$

$= [34.29882, 8.71647, 13.417665, 3.43294]$

Άρα Διάνυσμα νέου συνθετικού προτύπου (μόνο από numeric χαρακτηριστικά) $= [34.29882, 8.71647, 13.417665, 3.43294]$

ΕΠΙΠΛΕΟΝ ΒΗΜΑ ΓΙΑ ΤΑ NOMINAL ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ

Για κάθε nominal χαρακτηριστικό υπολογίζεται η συχνότητα εμφάνισης των τιμών μεταξύ των k γειτόνων και του αρχικού προτύπου, οπότε προκύπτει ο Πίνακας 55.

		Nominal	*Nominal*
Α / Α	**Πρότυπο**	**sex**	**communication_channel**
1000	Α (Αρχικό)	1	1
5428	Β (γείτονας)	2	3
5157	Γ (γείτονας)	2	1
4469	ΣΤ (γείτονας)	1	1
		Συχνότητα 1: 2 Συχνότητα 2: 2 **Ισοβαθμία**	Συχνότητα 1: 3 Συχνότητα 2: 0 Συχνότητα 3: 1 **Άρα Συχνότερη τιμή το 1**

Πίνακας 55: Υπολογισμός συχνότητας εμφάνισης των Nominal χαρακτηριστικών (SMOTE - N)

Στο διάνυσμα του νέου συνθετικού προτύπου που υπολογίστηκε πριν, ως νέες τιμές κάθε nominal χαρακτηριστικού ενσωματώνονται οι συχνότερες τιμές που προέκυψαν από τον παραπάνω πίνακα (Πίνακας 55). Στην περίπτωση ισοβαθμίας στις συχνότητες εμφάνισης επιλέγεται μια από τις τιμές με τυχαίο τρόπο. Στην συγκεκριμένη περίπτωση έστω ότι επιλέγεται με τυχαίο τρόπο η τιμή 2.

Διάνυσμα νέου συνθετικού προτύπου (μόνο από numeric χαρακτηριστικά) $=$

$= [34.29882, 8.71647, 13.417665, 3.43294]$

Σε αυτό ενσωματώνονται οι υπόλοιπες nominal τιμές, άρα το διάνυσμα νέου συνθετικού προτύπου γίνεται τελικά:

Διάνυσμα νέου συνθετικού προτύπου (numeric + nominal χαρακτηριστικά) =

$$= [34.29882, 8.71647, 13.417665, 3.43294, 2, 1]$$

Άρα τελικά τα αποτελέσματα παρουσιάζονται στη συνέχεια (Πίνακας 56):

		Numeric	*Numeric*	*Numeric*	*Numeric*	*Nominal*	*Nominal*
A / A	**Πρότυπο**	**age**	**appointment_hour_d**	**scheduled_hour_d**	**waiting_days**	**sex**	**communication_channel**
1000	Α (Αρχικό)	30.0	8.0	17.0	2.0	1	1
5428	Β (επιλεγμένος γείτονας)	36.0	9.0	12.0	4.0	2	3
5157	Γ (γείτονας)	34.0	9.0	18.0	1.0	2	1
1497	Δ (γείτονας)	38.0	8.0	21.0	1.0	1	3
4469	ΣΤ (γείτονας)	25.0	10.0	9.0	0.0	1	1
	Νέο Συνθετικό πρότυπο SMOTE	**34.29882**	**8.71647**	**13.417665**	**3.43294**	**2**	**1**

Πίνακας 56: Πραγματικά και συνθετικά δεδομένα της κλάσης Yes μετά από εφαρμογή SMOTE - N

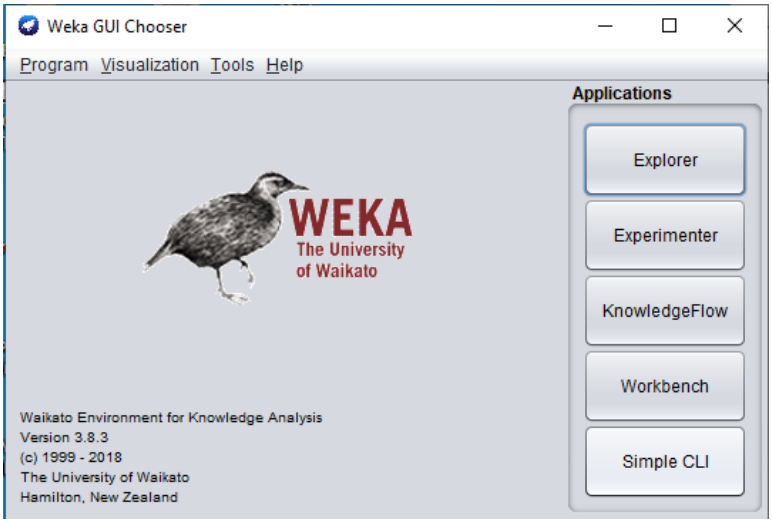
Η διαδικασία που περιγράφηκε επαναλαμβάνεται και για τα υπόλοιπα πρότυπα με βάση τα οποία θα δημιουργηθούν νέα συνθετικά πρότυπα.

6

Σύνολα Δεδομένων ιατρικών ραντεβού και περιβάλλον πειραμάτων

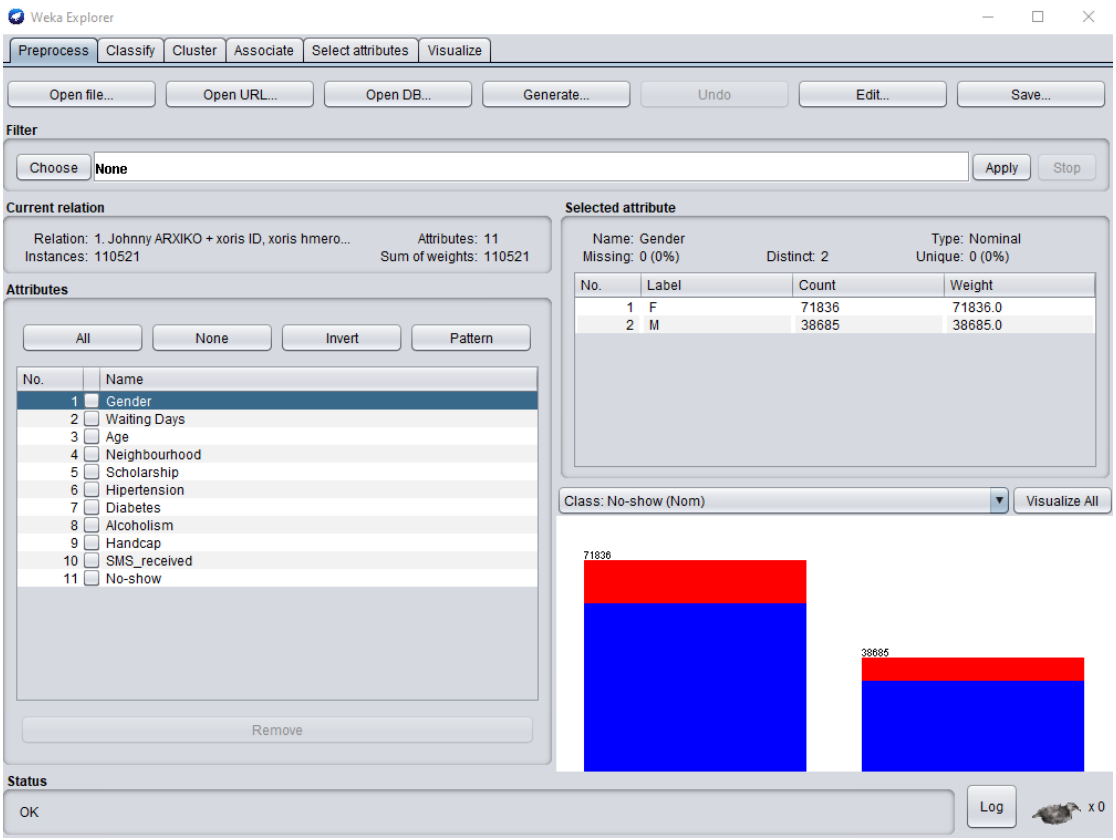
6.1 Το λογισμικό WEKA

Το WEKA είναι ένα πακέτο λογισμικού που αναπτύχθηκε στο Πανεπιστήμιο του Waikato της Ν. Ζηλανδίας και χρησιμοποιείται για Εξόρυξη Δεδομένων και Μηχανική Μάθηση (Eibe et al., 2016). Το όνομα του προέρχεται από τα αρχικά των λέξεων Waikato Environment for Knowledge Analysis (WEKA). Παράλληλα το όνομα Weka, αναφέρεται και σε ένα μικρό πουλί της Ν. Ζηλανδίας που είναι υπό εξαφάνιση. Το λογισμικό WEKA είναι γραμμένο σε γλώσσα προγραμματισμού Java και ανήκει στην κατηγορία του Ελεύθερου Λογισμικού. Ουσιαστικά αποτελεί μια συλλογή εργαλείων προεπεξεργασίας δεδομένων και αλγορίθμων μηχανικής μάθησης και Εξόρυξης δεδομένων. Στην Εικόνα 30 παρουσιάζονται οι αρχικές επιλογές του λογισμικού WEKA.



Εικόνα 30:Λογισμικό WEKA

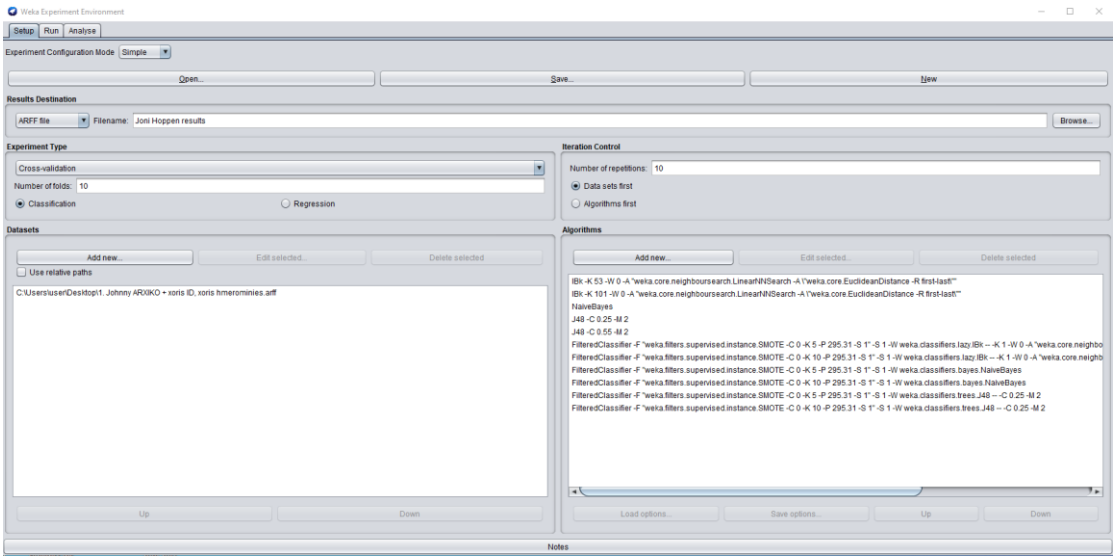
Είναι αρκετά διαδεδομένο καθώς παρέχει την δυνατότητα χρήσης όλων των λειτουργιών του μέσω γραφικού περιβάλλοντος. Η βασική διεπαφή του WEKA ονομάζεται Explorer. Μέσα από τη συγκεκριμένη διεπαφή μπορούν να χρησιμοποιηθούν εργαλεία που είναι διαθέσιμα στις κατηγορίες Προεπεξεργασία (Preprocess), Κατηγοριοποίηση (Classify), Συσταδοποίηση (Cluster), Κανόνες Συσχέτισης (Associate), Επιλογή χαρακτηριστικών (Select Attributes), Οπτικοποίηση (Visualize). Στην Εικόνα 31 φαίνονται οι παραπάνω κατηγορίες.



Εικόνα 31: Η διεπαφή Explorer του WEKA

Ειδικά για την κατηγοριοποίηση περιλαμβάνονται πολλές μέθοδοι όπως Δέντρα Αποφάσεων, Νευρωνικά Δίκτυα, Μηχανές Διανυσμάτων Υποστήριξης, Μπαϋεσιανοί κατηγοριοποιητές, k-κοντινότεροι γείτονες κλπ. με δυνατότητα παραμετροποίησης τους.

Μια άλλη χρήσιμη διεπαφή που περιλαμβάνει το λογισμικό WEKA είναι ο Experimenter (Πειραματιστής). Η συγκεκριμένη διεπαφή επιτρέπει τον προγραμματισμό μεγάλου πλήθους πειραμάτων, τα οποία μπορούν να εκτελεστούν μαζικά. Τα αποτελέσματα αυτών των πειραμάτων αποθηκεύονται αυτόματα σε ένα αρχείο, πράγμα που διευκολύνει την επιλογή των καλύτερων μοντέλων. Επομένως η διεπαφή experimenter βοηθάει στην αυτοματοποίηση της πειραματικής διαδικασίας σε περίπτωση που χρειάζεται να εκτελεστεί μεγάλο πλήθος πειραμάτων. Στην Εικόνα 32 φαίνεται το περιβάλλον του Experimenter.



Εικόνα 32: Διεπαφή Experimenter

Στην παρούσα διπλωματική εργασία για την προεπεξεργασία των συνόλων δεδομένων αλλά και την εκτέλεση πειραμάτων χρησιμοποιήθηκε το λογισμικό μηχανικής μάθησης WEKA, το οποίο είναι ελεύθερα διαθέσιμο στον ιστότοπο https://www.cs.waikato.ac.nz/ml/weka/.

Ως προς τα σύνολα δεδομένων το WEKA χρησιμοποιεί μια μορφή αρχείων δεδομένων που ονομάζεται ARFF (Attribute Relation File Format). Το ARFF αρχείο είναι ένα αρχείο κειμένου με ειδική μορφή και έχει αναπτυχθεί ειδικά για χρήση με το WEKA (Eibe et al., 2016). Φυσικά το WEKA μπορεί να διαχειριστεί και άλλους τύπους αρχείων όπως το csv. Ουσιαστικά όμως στη συνέχεια τα αρχεία μετατρέπονται και αποθηκεύονται σε μορφή ARFF. Ο τύπος ARFF υποστηρίζει 4 τύπους δεδομένων για τα χαρακτηριστικά: αριθμητικά (numeric), κατηγορικά (nominal), αλφαριθμητικά (string) και ημερομηνίες (date). Ο τύπος date περιέχει υποχρεωτικά ημερομηνία και ώρα μαζί. Οι παραπάνω πληροφορίες παρέχονται και στον ιστότοπο Weka Wiki https://waikato.github.io/weka-wiki/formats_and_processing/arff_stable/.

6.1.1 Φίλτρα προεπεξεργασίας που χρησιμοποιήθηκαν

Κατά την διάρκεια της προεπεξεργασίας των συνόλων δεδομένων χρησιμοποιήθηκαν τα παρακάτω φίλτρα προεπεξεργασίας.

Φίλτρο ChangeDateFormat

Το φίλτρο **ChangeDateFormat** μετατρέπει την μορφή ημερομηνίας ενός χαρακτηριστικού τύπου date σε κάποια άλλη μορφή ημερομηνίας. Παράδειγμα από μορφή **yyyy-MM-dd'THH:mm:ss** (ημερομηνία και ώρα) σε μορφή **MM** (μόνο μήνας). Το συγκεκριμένο φίλτρο μπορεί να εφαρμοστεί μέσα στο WEKA, ακολουθώντας τη διαδρομή: Preprocess -> Choose -> weka -> filters -> unsupervised -> ChangeDateFormat.

Φίλτρο AddExpression

Το φίλτρο **AddExpression** δημιουργεί ένα καινούργιο χαρακτηριστικό εφαρμόζοντας μια μαθηματική έκφραση σε υπάρχοντα χαρακτηριστικά. Για παράδειγμα έστω ότι στόχος είναι να δημιουργηθεί ένα νέο χαρακτηριστικό που να προκύπτει από την διαφορά 2 άλλων χαρακτηριστικών που βρίσκονται στις στήλες α8 και α4. Στο φίλτρο λοιπόν **AddExpression** στο πεδίο name μπορεί να οριστεί το όνομα για το νέο χαρακτηριστικό και στο πεδίο expression η μαθηματική έκφραση (α8-α4). Το συγκεκριμένο φίλτρο μπορεί να εφαρμοστεί μέσα στο WEKA, ακολουθώντας τη διαδρομή: Preprocess -> Choose -> weka -> filters -> unsupervised ->attribute -> AddExpression.

Φίλτρο SMOTE

Το φίλτρο SMOTE είναι η υλοποίηση της τεχνικής υπερδειγματοληψίας SMOTE. Σκοπός του συγκεκριμένου φίλτρου είναι να μετατρέψει ένα μη ισορροπημένο σύνολο δεδομένων σε ισορροπημένο. Το συγκεκριμένο φίλτρο δεν υπάρχει ενσωματωμένο στο λογισμικό WEKA. Επομένως απαιτείται η εγκατάσταση του. Η διαδικασία της εγκατάστασης γίνεται μέσω του διαχειριστή πακέτων (Package manager). Ο προαναφερόμενος διαχειριστής βρίσκεται στην αρχική διεπαφή του WEKA στο μενού Tools (Tools -> Package manager).

Το φίλτρο SMOTE μπορεί να χρησιμοποιηθεί με 2 τρόπους. Ο βασικός τρόπος είναι να εφαρμοστεί σε ολόκληρο το σύνολο δεδομένων. Αυτό γίνεται μέσα από την καρτέλα Preprocess, ακολουθώντας την διαδρομή: Preprocess -> Choose -> weka -> filters -> supervised ->Instance ->SMOTE. Ένας άλλος τρόπος είναι να εφαρμοστεί μόνο στο σύνολο εκπαίδευσης, πριν από την εκπαίδευση ενός κατηγοριοποιητή. Αυτό μπορεί να γίνει με την

βοήθεια του FilteredClassifier από την διαδρομή: Classify ->Choose ->weka ->classifiers ->meta ->FilteredClassifier.

6.1.2 Αλγόριθμοι κατηγοριοποίησης που χρησιμοποιήθηκαν

Στην παρούσα διπλωματική εργασία χρησιμοποιήθηκαν 3 αλγόριθμοι με διαφορετική μεθοδολογία κατηγοριοποίησης. Ένας αλγόριθμος είναι ο C4.5 (ή J48 στο WEKA), ο οποίος ανήκει στην κατηγορία των Δέντρων Απόφασης. Ο επόμενος είναι ο αλγόριθμος Naïve Bayes, ο οποίος ανήκει στην κατηγορία των αλγορίθμων που εφαρμόζουν το θεώρημα Bayes. Τέλος ο αλγόριθμος k-NN (ή Ibk στο WEKA) που ανήκει στην κατηγορία των αλγορίθμων που χρησιμοποιούν την απόσταση ως μέτρο κατηγοριοποίησης άγνωστων δεδομένων.

Η εφαρμογή των παραπάνω αλγορίθμων μέσα στο WEKA γίνεται με 2 τρόπους. Ο 1ος τρόπος γίνεται με απλή εφαρμογή του εκάστοτε αλγορίθμου στο σύνολο δεδομένων, όπου επιλέγεται η αντίστοιχη διαδρομή. Δηλαδή για τον αλγόριθμο C4.5 επιλέγεται η διαδρομή: Classify -> Choose -> weka -> classifiers -> trees -> J48. Για τον αλγόριθμο Naive Bayes η διαδρομή: Classify -> Choose -> weka -> classifiers -> bayes -> Naive Bayes. Τέλος για τον αλγόριθμο k – NN επιλέγεται η διαδρομή: Classify -> Choose -> weka -> classifiers -> lazy -> Ibk. Ο 2ος τρόπος είναι μέσω του FilteredClassifier, που αναλύεται παρακάτω.

Ειδικός κατηγοριοποιητής FilteredClassifier

Εκτός από τα φίλτρα το λογισμικό WEKA περιλαμβάνει επίσης ειδικού τύπου κατηγοριοποιητές οι οποίοι ονομάζονται metalearners. Οι συγκεκριμένοι κατηγοριοποιητές συνδυάζουν λειτουργίες άλλων κατηγοριοποιητών και φίλτρων, με σκοπό την κατά το δυνατόν βελτίωση των αποτελεσμάτων. Στην παρούσα διπλωματική εργασία χρησιμοποιείται ο metalearner κατηγοριοποιητής FilteredClassifier.

Ο κατηγοριοποιητής FilteredClassifier πρώτα εφαρμόζει ένα φίλτρο (πχ το SMOTE) στο σύνολο δεδομένων και στη συνέχεια έναν κατηγοριοποιητή (πχ k -NN). Αυτό που τον κάνει ιδιαίτερα χρήσιμο είναι ότι έχει την ιδιότητα να εφαρμόζει το εκάστοτε φίλτρο μόνο στο σύνολο εκπαίδευσης, αλλά πριν από την εκπαίδευση. Η εφαρμογή του γίνεται ακολουθώντας τη διαδρομή: Classify -> Choose -> weka -> classifiers -> meta -> FilteredClassifier (Eibe et al., 2016).

6.2 Ιδιαιτερότητες Ιατρικών Συνόλων Δεδομένων

Στην βιβλιογραφία που ασχολείται με παρόμοια προβλήματα χρησιμοποιούνται αρκετά σύνολα δεδομένων, τα οποία όμως δυστυχώς δεν διατίθενται ελεύθερα. Αυτό συμβαίνει κυρίως λόγω του ότι περιέχουν ευαίσθητα προσωπικά και ιατρικά δεδομένα ασθενών και για να δημοσιευθούν πρέπει να γίνει ανωνυμοποίηση των δεδομένων, έτσι ώστε να είναι σύμφωνα με τον Γενικό Κανονισμό Προστασίας Προσωπικών Δεδομένων (GDPR).

Η ανωνυμοποίηση δεδομένων (data anonymization) είναι η διαδικασία που απαιτείται, ώστε να είναι θεωρητικά αδύνατο να προσδιοριστεί η ταυτότητα ενός υποκειμένου με βάση τα δεδομένα. Για να επιτευχθεί αυτό πρέπει να αφαιρεθούν από τα δεδομένα όλες οι πληροφορίες που μπορούν να χρησιμοποιηθούν για ταυτοποίηση (όπως ονοματεπώνυμο, ημερομηνίες γέννησης κ.α.), διατηρώντας όμως τα υπόλοιπα στοιχεία, ώστε τα δεδομένα να μπορούν να χρησιμοποιηθούν για έρευνα. Παρ' όλα αυτά, ακόμα και μετά την αφαίρεση αυτών των στοιχείων, υπάρχει περίπτωση να επιτευχθεί ταυτοποίηση κάνοντας σύγκριση με άλλα διαθέσιμα σύνολα δεδομένων που περιέχουν κάποια από τα εν λόγω στοιχεία. Για το λόγο αυτό λοιπόν απαιτείται επιπλέον επεξεργασία των δεδομένων για να ελαχιστοποιηθεί και αυτή η πιθανότητα.

Από τα παραπάνω είναι σαφές ότι απαιτείται αρκετή επιπλέον εργασία για την ανωνυμοποίηση, ειδικά σε μεγάλα σύνολα δεδομένων και μάλιστα χωρίς να είναι σίγουρο το αποτέλεσμα. Ένας άλλος παράγοντας που δυσκολεύει την απόκτηση ιατρικών δεδομένων είναι οι απαιτούμενες γραφειοκρατικές διαδικασίες, οι οποίες είναι συνήθως αρκετά χρονοβόρες. Αυτοί είναι πιθανώς οι πιο σημαντικοί λόγοι για την μη δημοσίευση τέτοιων συνόλων δεδομένων.

Στα πλαίσια της συγκεκριμένης διπλωματικής εργασίας έγιναν προσπάθειες για την απόκτηση δεδομένων για ιατρικά ραντεβού από ελληνικά νοσοκομεία. Ωστόσο λόγω των ιδιαιτεροτήτων που έχουν αυτά τα σύνολα δεδομένων, αυτό δεν κατέστη εφικτό. Για το λόγο αυτό τελικά χρησιμοποιήθηκαν δύο (2) σύνολα δεδομένων που υπάρχουν ελεύθερα διαθέσιμα στην on-line πλατφόρμα μηχανικής μάθησης kaggle.com (απαιτείται όμως εγγραφή).

Τα συγκεκριμένα σύνολα δεδομένων είναι μικρού και μετρίου μεγέθους και ανωνυμοποιημένα. Πρόκειται για τα σύνολα δεδομένων "Medical Appointment No Shows" και "Medical Appointment", τα οποία θα περιγραφούν στη συνέχεια. Επειδή τα δύο σύνολα δεδομένων έχουν παραπλήσια ονόματα, για αποφυγή σύγχυσης και για λόγους ευκολίας στην παρούσα εργασία, θα τους δοθεί μια κωδική ονομασία με βάση το όνομα του χρήστη που ανέβασε το κάθε σύνολο στην πλατφόρμα. Έτσι, το πρώτο σύνολο (Medical Appointment No

Shows) θα αναφέρεται με το όνομα “Joni Hoppen” και το δεύτερο (Medical Appointment) με το όνομα “Alvaro Flores”.

6.3 Σύνολο δεδομένων Joni Hoppen

Το σύνολο δεδομένων έχει τίτλο “Medical Appointment No Shows” και το έχει ανεβάσει στην πλατφόρμα μηχανικής μάθησης Kaggle, ένας χρήστης με όνομα JoniHoppen το 2017. Συγκεκριμένα χρησιμοποιήθηκε η έκδοση 5 των δεδομένων. Τα δεδομένα είναι ελεύθερα προσβάσιμα στην δικτυακή διεύθυνση https://www.kaggle.com/joniarroba/noshowappointments. Το αρχείο δεδομένων είναι σε μορφή csv, με όνομα «KaggleV2-May-2016.csv» μεγέθους περίπου 10 MB.

Τα δεδομένα του συνόλου αφορούν στοιχεία ιατρικών ραντεβού ασθενών από διάφορα κρατικά νοσοκομεία της Βραζιλίας για μια περίοδο περίπου 3 μηνών του 2016 (Από 29/04/2016 έως και 08/06/2016). Ειδικότερα το συγκεκριμένο σύνολο δεδομένων περιλαμβάνει 110.527 ραντεβού με 14 χαρακτηριστικά για το καθένα. Τα χαρακτηριστικά αφορούν κάποια δημογραφικά στοιχεία όπως το φύλο και την ηλικία του ασθενούς, στοιχεία σχετικά με το ραντεβού όπως ημερομηνίες και τοποθεσία νοσοκομείου και κάποια ιατρικά στοιχεία όπως χρόνιες παθήσεις κλπ. Τα ονόματα, ο τύπος δεδομένων και μια συνοπτική περιγραφή των χαρακτηριστικών απεικονίζονται στον πίνακα που ακολουθεί (Πίνακας 57):

α/α	Όνομα	Τιμή	Περιγραφή
1	PatientID	αριθμός	ID ασθενή
2	AppointmentID	αριθμός	ID ραντεβού
3	Gender	F ή M	F: γυναίκα και M: Άνδρας
4	ScheduledDay	Ημερομηνία και ώρα	Ημερομηνία και ώρα καταχώρησης ραντεβού
5	AppointmentDay	Ημερομηνία και ώρα	Ημερομηνία και ώρα που πρέπει να γίνει το ραντεβού
6	Age	αριθμός	Ηλικία ασθενούς
7	Neighbourhood	string	Τοποθεσία του νοσοκομείου του ραντεβού
8	Scholarship	0 ή 1	Αν ο ασθενής υπάγεται σε πρόγραμμα κοινωνικής ασφάλισης
9	Hipertension	0 ή 1	Αν ο ασθενής έχει υπέρταση
10	Diabetes	0 ή 1	Αν ο ασθενής έχει διαβήτη
11	Alcoholism	0 ή 1	Αν ο ασθενής είναι αλκοολικός
12	Handcap	0 έως 4	Αν ο ασθενής έχει αναπηρία (1-3) ή όχι (0)
13	SMS_received	0 ή 1	Αν έχει σταλεί στον ασθενή SMS υπενθύμισης
14	No-show	Yes ή No	Yes: αν ο ασθενής ΔΕΝ παρουσιάστηκε στο ραντεβού No: αν ο ασθενής παρουσιάστηκε στο ραντεβού

Πίνακας 57: Συνοπτική περιγραφή των χαρακτηριστικών του συνόλου δεδομένων JoniHoppen (πριν την προεπεξεργασία)

6.3.1 Προεπεξεργασία δεδομένων

1) **Διόρθωση Πορτογαλικών χαρακτήρων:** Κατά το άνοιγμα του αρχείου με το WEKA παρουσιάστηκε πρόβλημα στην ανάγνωση κάποιων δεδομένων της στήλης Neighbourhood. Αυτό οφείλεται στην χρήση του πορτογαλικού αλφαβήτου που περιέχει κάποιους τονιζόμενους χαρακτήρες, τους οποίους δεν μπορεί να διαβάσει το WEKA. Το πρόβλημα διορθώνεται με άνοιγμα του αρχείου ως κείμενο (π.χ. με Notepad οπότε οι πορτογαλικοί χαρακτήρες εμφανίζονται κανονικά) και μαζική αντικατάσταση των προβληματικών χαρακτήρων με τους αντίστοιχους λατινικούς (π.χ. τονιζόμενο Ι σε κανονικό Ι).
2) **Έλεγχος για ελλιπείς τιμές (missing values):** Έγινε έλεγχος τιμών σε όλα τα χαρακτηριστικά. Το συγκεκριμένο σύνολο δεδομένων δεν έχει ελλιπείς τιμές.
3) **Καθαρισμός προβληματικών δεδομένων:** Ελέγχοντας για προβληματικές τιμές χαρακτηριστικών, διαπιστώθηκε ότι στην στήλη age υπάρχει μία εγγραφή με ηλικία -1. Καθώς η ηλικία δεν μπορεί να είναι αρνητική θεωρήθηκε ότι είναι καλύτερο να διαγραφεί η συγκεκριμένη εγγραφή.

   Επίσης παρατηρήθηκε ότι υπάρχουν 3539 εγγραφές με ηλικία 0. Η συγκεκριμένη τιμή φαίνεται περίεργη, αλλά είναι πιθανόν να πρόκειται για παιδιά κάτω του ενός έτους ή νεογέννητα. Για να επιβεβαιωθεί η παραπάνω υπόθεση έγινε έλεγχος στις στήλες Hipertension και Alcoolism στις συγκεκριμένες 3539 εγγραφές και διαπιστώθηκε ότι και στις 2 στήλες όλες οι τιμές ήταν 0. Καθώς τα νεογέννητα παιδιά είναι σπάνιο να εμφανίσουν υπέρταση και αδύνατο να πάσχουν από αλκοολισμό προέκυψε το συμπέρασμα ότι το πιθανότερο είναι οι συγκεκριμένες εγγραφές να αφορούν νεογέννητα παιδιά και όχι λανθασμένες τιμές ηλικίας. Κατά συνέπεια θεωρήθηκαν αποδεκτές.
4) **Έλεγχος τύπων δεδομένων:** Κατά το άνοιγμα του αρχείου το WEKA αναγνώριζε τα χαρακτηριστικά ScheduledDay και AppointmentDay ως nominal και όχι ως Date επειδή περιείχαν χαρακτήρες. Για να μπορέσει λοιπόν το WEKA να αναγνωρίσει τα παραπάνω χαρακτηριστικά έγινε τροποποίηση των δεδομένων με τέτοιο τρόπο ώστε να έρθουν στην κατάλληλη μορφή ημερομηνίας και στη συνέχεια να μπορούν μετατραπούν σε τύπο Date**.**
5) **Τροποποίηση στηλών ημερομηνίας και ώρας ScheduledDay και AppointmentDay:**

   Διαπιστώθηκε ότι οι στήλες ScheduledDay και AppointmentDay περιλάμβαναν την ημερομηνία και ώρα κλεισίματος καθώς και την ημερομηνία και ώρα πραγματοποίησης του ραντεβού σε 1 στήλη. Όμως στην στήλη AppointmentDay η ώρα πραγματοποίησης του ραντεβού σε όλες τις γραμμές φαινόταν ότι είναι η ίδια,

δηλαδή 00:00:00. Επειδή δεν μπορεί όλα τα ραντεβού να πραγματοποιήθηκαν την ίδια ώρα και μάλιστα 12 τα μεσάνυχτα θεωρήθηκε ότι η ώρα πραγματοποίησης του ραντεβού ουσιαστικά δεν υπάρχει οπότε θα έπρεπε να διαγραφεί. Αυτό όμως θα δημιουργούσε στις στήλες ScheduledDay και AppointmentDay πεδία ημερομηνίας με διαφορετικές μορφές, δηλαδή datetime και date αντίστοιχα που δεν θα μπορούσε να τις διαχειριστεί το WEKA. Στο WEKA όλα τα δεδομένα σχετικά με ημερομηνίες θα πρέπει υποχρεωτικά να είναι της ίδιας μορφής (datetime με datetime ή date με date).

Συνεπώς η στήλη ScheduledDay που είναι Nominal διαχωρίστηκε σε ξεχωριστές στήλες ημερομηνίας και ώρας (πχ ScheduledDay μόνο για την ημερομηνία και ScheduledDayTime μόνο για την ώρα). Το πεδίο της ημερομηνίας (στήλη ScheduledDay) που προέκυψε και εξακολουθεί να είναι Nominal, στη συνέχεια μετατράπηκε σε τύπο date.

Αντίθετα το πεδίο της ώρας (ScheduledDayTime) που προέκυψε από τον διαχωρισμό της στήλης ScheduledDay ήταν Nominal αλλά δεν μπορούσε να μετατραπεί σε τύπο date ή datetime αφού δεν περιείχε ημερομηνία. Μία λύση λοιπόν ήταν να παραμείνει το πεδίο της ώρας ως Nominal. Έτσι όμως το WEKA θα διάβαζε την ώρα σαν κείμενο κάτι που μάλλον δεν θα είχε νόημα. Μία δεύτερη λύση και προτιμότερη, η οποία και επιλέχθηκε ήταν το πεδίο της ώρας να διαχωριστεί επιπλέον σε 3 ξεχωριστές στήλες για ώρες, λεπτά και δευτερόλεπτα, οι οποίες θα ήταν πλέον Numeric.

Ως προς τη στήλη AppointmentDay διατηρήθηκε το τμήμα της ημερομηνίας και διαγράφτηκε εντελώς το τμήμα της ώρας. Τέλος η στήλη AppointmentDay μετατράπηκε σε τύπο date, έτσι ώστε να είναι της ίδιας μορφής με τη στήλη ScheduledDay. Σκοπός της συγκεκριμένης μετατροπής ήταν και οι δύο στήλες (AppointmentDay και ScheduledDay) να γίνουν τύπου δεδομένων date, ώστε να είναι εφικτός μεταξύ τους ο υπολογισμός των ημερών αναμονής (waitingDays) που περιμένει ένας ασθενής μέχρι την ημέρα πραγματοποίησης του προγραμματισμένου του ραντεβού.

Συνεπώς έγιναν οι παρακάτω μετασχηματισμοί:

i. **Τροποποίηση στήλης:** Διαγραφή του τμήματος της ώρας (00:00:00) από τη στήλη AppointmentDay.

ii. **Διαχωρισμός στήλης ScheduledDay:** Διαχωρισμός της στήλης ScheduledDay σε ξεχωριστές στήλες ημερομηνίας και ώρας π.χ. ScheduledDay και ScheduledDayTime.

iii. **Διαχωρισμός στήλης ScheduledDayTime:** Περαιτέρω διαχωρισμός της στήλης ScheduledDayTime του προηγούμενου βήματος σε ScheduledHour, ScheduledMin και ScheduledSec.

Οι παραπάνω μετασχηματισμοί ήταν πιο εύκολο να γίνουν στο αρχικό csv αρχείο με μαζική αντικατάσταση κειμένου, όπως και έγινε.

Μετά τις παραπάνω αλλαγές το αρχείο μπόρεσε να ανοίξει με το Weka και οι στήλες ημερομηνίας αναγνωρίστηκαν ως τύπος Date.

6) **Δημιουργία νέας στήλης waitingDays:** Από τις στήλες ScheduledDay και AppointmentDay μπορεί να υπολογιστεί ο χρόνος αναμονής του ραντεβού, δηλαδή ο αριθμός των ημερών ανάμεσα στην ημερομηνία που κλείστηκε και την ημερομηνία που πραγματοποιήθηκε το ραντεβού. Αυτό έγινε κάνοντας αφαίρεση των τιμών AppointmentDay – ScheduledDay και το αποτέλεσμα ήταν μια νέα στήλη waitingDays. Η πιο πάνω αφαίρεση έγινε με το φίλτρο AddExpression του WEKA. Συγκεκριμένα στο πεδίο name δόθηκε το όνομα waitingDays και στο πεδίο expression η μαθηματική έκφραση (α8-α4) / 86.400.000. Στην παραπάνω έκφραση το α8 είναι η στήλη AppointmentDay, το α4 η στήλη ScheduledDay και το 86.400.000 είναι τα milliseconds μίας ημέρας. Η διαφορά των ημερομηνιών δίνει αποτέλεσμα σε milliseconds, ενώ το ζητούμενο είναι ημέρες αναμονής. Για το λόγο αυτό λοιπόν γίνεται η διαίρεση με το 86.400.000.

   Επιπλέον, ελέγχοντας τις τιμές που προέκυψαν στη νέα στήλη waitingDays έγινε αντιληπτό ότι υπάρχουν 5 περιπτώσεις με αρνητικές τιμές που οφείλονται στο γεγονός ότι η ημερομηνία του ScheduledDay είναι μεταγενέστερη του AppointmentDay. Οι παραπάνω 5 περιπτώσεις θεωρήθηκαν λανθασμένες. Επομένως διαγράφηκαν.

7) **Δημιουργία νέων στηλών για ημερομηνίες:** Κάποιοι αλγόριθμοι όπως π.χ. ο Naïve Bayes δεν μπορούν να επεξεργαστούν δεδομένα τύπου Date (Στο WEKA αν υπάρχουν δεδομένα τύπου date ο αλγόριθμος Naïve Bayes δεν τρέχει). Επομένως μία λύση η οποία και επιλέχθηκε ήταν τα δεδομένα αυτά να μετατραπούν σε Numeric χωρίζοντας τα τμήματα της ημερομηνίας σε ξεχωριστές στήλες Ημέρα, Μήνας και Έτος. Στο συγκεκριμένο σύνολο δεδομένων παρατηρήθηκε ότι όλες οι ημερομηνίες της στήλης AppointmentDay ανήκουν στο ίδιο έτος (2016) ενώ στη στήλη ScheduledDay υπάρχουν ημερομηνίες του 2016 αλλά και κάποιες του 2015. Οι ημερομηνίες όμως του έτους 2015 είναι σχετικά πολύ λίγες, μόλις 62 εγγραφές από συνολικά 110.521 πρότυπα. Δηλαδή αποτελούν περίπου το 0,05% του συνόλου δεδομένων. Επομένως από την στήλη AppointmentDay διαγράφηκε το έτος εφόσον όλες οι εγγραφές είχαν την ίδια τιμή (2016). Αντίστοιχα το έτος διαγράφηκε και από

τη στήλη ScheduledDay επειδή οι εγγραφές του 2015 ήταν σχετικά πολύ λίγες και θεωρήθηκε ότι δεν θα επηρεάσουν σημαντικά το αποτέλεσμα της κατηγοριοποίησης.

Συνεπώς έγιναν οι παρακάτω μετασχηματισμοί:

i. **Διαχωρισμός στήλης ScheduledDay:** Μετατροπή της αρχικής στήλης ScheduledDay σε ScheduledDay (που περιέχει μόνο την ημέρα) και ScheduledDayMonth (μόνο μήνας). Με το που θα γίνει ο διαχωρισμός της στήλης ScheduledDay τα δεδομένα θα μετατραπούν αυτόματα σε τύπο Numeric.
ii. **Διαχωρισμός στήλης AppointmentDay:** Μετατροπή της αρχικής στήλης AppointmentDay σε AppointmentDay (που περιέχει μόνο την ημέρα) και AppointmentDayMonth (μόνο μήνας).

Με το που θα γίνει ο διαχωρισμός των παραπάνω στηλών τα δεδομένα θα μετατραπούν αυτόματα σε τύπο Numeric. Για τον διαχωρισμό χρησιμοποιήθηκε το φίλτρο ChangeDateFormat (yyyy-MM-dd'THH:mm:ss) του WEKA.

8) **Διαγραφή στηλών PatientID και AppointmentID:** Οι στήλες PatientID και AppointmentID περιέχουν αριθμητικές τιμές για ταυτοποίηση των εγγραφών και δεν προσφέρουν κάποια πληροφορία σχετικά με τα ιατρικά ραντεβού. Συνεπώς διαγράφηκαν.
9) **Διαγραφή στηλών ημερομηνίας και ώρας.**
Με το σκεπτικό ότι από τις ημερομηνίες (**ScheduledDay και AppointmentDay)** παράχθηκε καινούργια πληροφορία, δηλαδή η στήλη WaitingDays, στη συνέχεια αφαιρέθηκαν οι στήλες της ημερομηνίας και ώρας, αφού πλέον θεωρήθηκαν περιττές. Συγκεκριμένα αφαιρέθηκαν οι στήλες ScheduledDay, ScheduledDayMonth, ScheduledHour, ScheduledMin, ScheduledSec, AppointmentDay, AppointmentDayMonth.

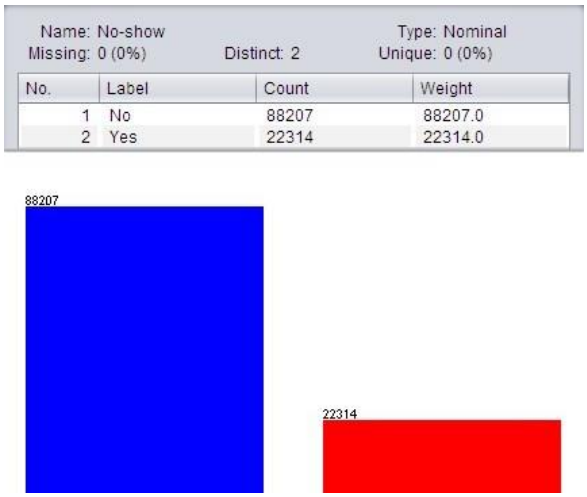
Μετά από την προεπεξεργασία των δεδομένων το τελικό αρχείο που προέκυψε είναι το αρχείο: 1. Johnny ARXIKO + xoris ID, xoris hmerominies και πλέον περιέχει συνολικά 11 χαρακτηριστικά και 110.521 στιγμιότυπα (Πίνακας 58).

α/α	Όνομα	τύπος	Περιγραφή
1	Gender	nominal	F: γυναίκα και M: Άνδρας
2	Waiting Days	numeric	Ημέρες αναμονής για ραντεβού
3	Age	numeric	Ηλικία ασθενούς
4	Neighbourhood	nominal	Τοποθεσία του νοσοκομείου του ραντεβού
5	Scholarship	numeric	Αν ο ασθενής υπάγεται σε πρόγραμμα κοινωνικής ασφάλισης
6	Hipertension	numeric	Αν ο ασθενής έχει υπέρταση

7	Diabetes	numeric	Αν ο ασθενής έχει διαβήτη
8	Alcoholism	numeric	Αν ο ασθενής είναι αλκοολικός
9	Handcap	numeric	Αν ο ασθενής έχει αναπηρία (1-3) ή όχι (0)
10	SMS_received	numeric	Αν έχει σταλεί στον ασθενή SMS υπενθύμισης
11	No-show	nominal	Yes: αν ο ασθενής ΔΕΝ παρουσιάστηκε στο ραντεβού No: αν ο ασθενής παρουσιάστηκε στο ραντεβού

Πίνακας 58: 1. Johnny ARXIKO + xoris ID, xoris hmeromhnies

Το σύνολο δεδομένων είναι μη-ισορροπημένο καθώς όπως φαίνεται στην Εικόνα 33, από τα 110.521 στιγμιότυπα, τα 88.207 (79,8%) ανήκουν στην κλάση No (πλειοψηφική κλάση), ενώ τα υπόλοιπα 22.314 (20,2%) ανήκουν στην κλάση Yes (Μειοψηφική κλάση).



Εικόνα 33: Κατανομή κλάσεων του συνόλου δεδομένων Joni Hoppen

6.4 Σύνολο δεδομένων Alvaro Flores

Το σύνολο δεδομένων έχει τίτλο "Medical Appointment" και το έχει ανεβάσει στην πλατφόρμα μηχανικής μάθησης Kaggle, ένας χρήστης με όνομα Alvaro Flores το 2017 (έκδοση 1). Τα δεδομένα είναι ελεύθερα προσβάσιμα στην δικτυακή διεύθυνση https://www.kaggle.com/afflores/medical-appointment.

Το αρχείο δεδομένων είναι σε μορφή csv, με όνομα «2017.csv» μεγέθους περίπου 4 MB.

Το σύνολο δεδομένων περιέχει πληροφορίες που αφορούν ιατρικά ραντεβού ασθενών από κάποιο νοσοκομείο στο Σαντιάγκο της Χιλής. Ειδικότερα το συγκεκριμένο σύνολο περιλαμβάνει 61.214 ραντεβού, τα οποία πραγματοποιήθηκαν την χρονική περίοδο από 1/1/2017 έως και 30/4/2017 (4 μήνες). Κάθε ραντεβού αποτελείται από 19 χαρακτηριστικά. Τα χαρακτηριστικά αυτά αφορούν δημογραφικά στοιχεία όπως το φύλο και η ηλικία του ασθενούς, καθώς και στοιχεία σχετικά με το ραντεβού όπως η ημέρα και ο μήνας

καταχώρησης και πραγματοποίησης ραντεβού, η ιατρική ειδικότητα, οι ημέρες αναμονής κλπ. Τα ονόματα, ο τύπος δεδομένων και μια συνοπτική περιγραφή των χαρακτηριστικών φαίνεται στον πίνακα που ακολουθεί (Πίνακας 59):

α/α	Όνομα	Τιμή	Περιγραφή
1	especialidad	αριθμός	Ειδικότητα ιατρού
2	edad	αριθμός	Ηλικία ασθενούς
3	sexo	1 ή 2	1: Άνδρας και 2: Γυναίκα
4	reserva_mes_d	αριθμός	Τιμή που αντιστοιχεί στο μήνα του ραντεβού (από 1 έως 12)
5	reserva_mes_c	αριθμός	Συνεχής τιμή του reserva_mes_d τύπος: cos(2*reserva_mes_d*Pi/12)
6	reserva_dia_d	αριθμός	Τιμή που αντιστοιχεί στην ημέρα του ραντεβού (1=Δευτέρα,….7=Κυριακή)
7	reserva_dia_c	αριθμός	Συνεχής τιμή του reserva_dia_d τύπος: cos(2*reserva_dia_d*Pi/7)
8	reserva_hora_d	αριθμός	Τιμή που αντιστοιχεί στην ώρα του ραντεβού
9	reserva_hora_c	αριθμός	Συνεχής τιμή του reserva_hora_d τύπος: cos(2*reserva_hora_d*Pi/24)
10	creacion_mes_d	αριθμός	Τιμή μήνα που αντιστοιχεί στο μήνα καταχώρησης του ραντεβού (από 1 έως 12)
11	creacion_mes_c	αριθμός	Συνεχής τιμή του creacion_mes_d τύπος: cos(2*creacion_mes_d*Pi/12)
12	creacion_dia_d	αριθμός	Τιμή που αντιστοιχεί στην ημέρα καταχώρησης του ραντεβού (1=Δευτέρα,….7=Κυριακή)
13	creacion_dia_c	αριθμός	Συνεχής τιμή του creacion_dia_d τύπος: cos(2*creacion_dia_d*Pi/7)
14	creacion_hora_d	αριθμός	Τιμή που αντιστοιχεί στην ώρα καταχώρησης του ραντεβού
15	creacion_hora_c	αριθμός	Συνεχής τιμή του creacion_hora_d τύπος: cos(2*creacion_hora_d*Pi/24)
16	latencia	αριθμός	Αριθμός ημερών μεταξύ της ημερομηνίας καταχώρησης του ραντεβού και της ημερομηνίας του ραντεβού
17	canal	1 ή 2 ή 3	Τρόπος καταχώρησης ραντεβού (1: τηλεφωνικό κέντρο, 2:Προσωπικά, 3:Web)
18	tipo	1 ή 2	Τύπος ραντεβού (1: Ιατρικό, 2: Διαδικαστικό)
19	show	0 ή 1	0: ο ασθενής ΔΕΝ παρουσιάστηκε στο ραντεβού 1: αν ο ασθενής παρουσιάστηκε στο ραντεβού

Πίνακας 59: Συνοπτική περιγραφή των χαρακτηριστικών του συνόλου δεδομένων Alvaro Flores (πριν την προεπεξεργασία)

6.4.1 Προεπεξεργασία δεδομένων

1) **Μετάφραση τίτλων στηλών:** Οι τίτλοι των στηλών είναι στην Ισπανική γλώσσα, οπότε για ευκολία μεταφράστηκαν στα αγγλικά.
2) **Έλεγχος για ελλιπείς τιμές (missing values):** Το συγκεκριμένο σύνολο δεδομένων δεν έχει ελλιπείς τιμές.

3) **Καθαρισμός προβληματικών δεδομένων:** Το πεδίο reserva_hora_d που περιέχει την ώρα του ραντεβού, παίρνει τιμές από 8 έως 21. Δηλαδή από 08:00 το πρωί έως 21:00 το βράδυ που είναι και το ωράριο εργασίας. Παρατηρήθηκε ότι υπάρχει ένα ραντεβού με τιμή 0. Αυτό σημαίνει ότι το συγκεκριμένο ραντεβού φαίνεται ότι πραγματοποιήθηκε εκτός του προβλεπόμενου ωραρίου εργασίας και είναι πολύ πιθανόν ότι έγινε λανθασμένη καταχώρηση (θόρυβος). Κατά συνέπεια η συγκεκριμένη εγγραφή διαγράφηκε.
4) **Διαγραφή περιττών στηλών:** Σύμφωνα με την περιγραφή του δημιουργού του συνόλου δεδομένων, υπάρχουν οι 6 στήλες reserva_mes_c, reserva_dia_c, creacion_mes_c, creacion_dia_c, creacion_dia_c, creacion_hora_c που προέκυψαν από τις αντίστοιχες στήλες που τελειώνουν με το γράμμα d. Ο λόγος της δημιουργίας τους δεν αναφέρεται. Επομένως θεωρήθηκε ότι πρόκειται για τις ίδιες πληροφορίες σε διαφορετική μορφή. Κατά συνέπεια διαγράφηκαν.
5) **Τροποποίηση χαρακτηριστικού κλάσης:** Το χαρακτηριστικό show είναι αυτό που ξεχωρίζει την κλάση στο συγκεκριμένο σύνολο δεδομένων. Αντίθετα στο προηγούμενο σύνολο το χαρακτηριστικό που ξεχωρίζει την κλάση είναι το no_show. Δηλαδή το αντίστροφο. Για να υπάρχει λοιπόν ομοιομορφία στο συλλογισμό και ευκολότερη κατανόηση χωρίς λάθη και παρερμηνείες το χαρακτηριστικό show μετονομάζεται σε no_show. Αντίστοιχα γίνεται αντικατάσταση και στις τιμές. Δηλαδή η τιμή 0 μετατρέπεται σε Yes και η τιμή 1 μετατρέπεται σε No (0 → Yes και 1 → No).

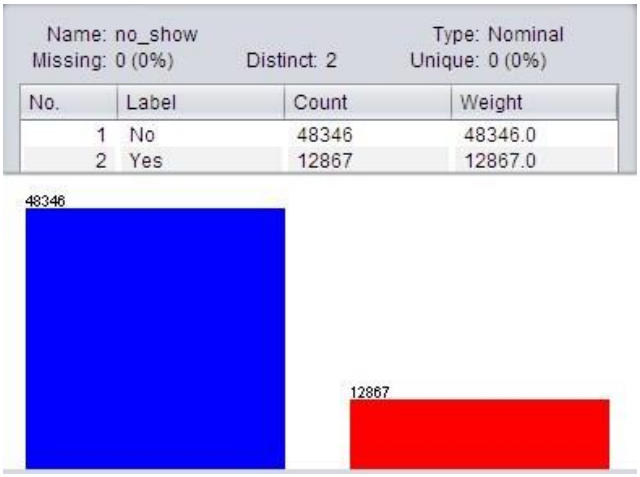
Μετά από την προεπεξεργασία των δεδομένων το τελικό αρχείο που προέκυψε είναι το αρχείο: 1.Alvaro Flores_ARXIKO_mono_d., το οποίο αποτελείται πλέον από 13 χαρακτηριστικά και 61.213 ραντεβού. Η περιγραφή του τελικού συνόλου δεδομένων μετά την προεπεξεργασία φαίνεται στον παρακάτω πίνακα (Πίνακας 60):

α/α	Όνομα	τύπος	Περιγραφή
1	medical_specialty	nominal	Ειδικότητα ιατρού
2	age	numeric	Ηλικία ασθενούς
3	sex	nominal	1: Άνδρας και 2: Γυναίκα
4	appointment_month_d	numeric	Τιμή που αντιστοιχεί στο μήνα του ραντεβού (από 1 έως 12)
5	appointment_weekday_d	numeric	Τιμή που αντιστοιχεί στην ημέρα του ραντεβού (1=Δευτέρα,….7=Κυριακή)
6	appointment_hour_d	numeric	Τιμή που αντιστοιχεί στην ώρα του ραντεβού
7	scheduled_month_d	numeric	Τιμή που αντιστοιχεί στο μήνα καταχώρησης του ραντεβού (από 1 έως 12)
8	scheduled _weekday_d	numeric	Τιμή που αντιστοιχεί στην ημέρα καταχώρησης του ραντεβού (1=Δευτέρα,….7=Κυριακή)
9	scheduled _hour_d	numeric	Τιμή που αντιστοιχεί στην ώρα καταχώρησης του

			ραντεβού
10	waiting_days	numeric	Αριθμός ημερών μεταξύ της ημερομηνίας καταχώρησης του ραντεβού και της ημερομηνίας του ραντεβού
11	communication_channel	nominal	Τρόπος καταχώρησης ραντεβού (1: τηλεφ. κέντρο, 2:Προσωπικά, 3:Web)
12	appointment_type	nominal	Τύπος ραντεβού (1: Ιατρικό, 2: Διαδικαστικό)
13	no_show	nominal	Yes: αν ο ασθενής ΔΕΝ παρουσιάστηκε στο ραντεβού No: αν ο ασθενής παρουσιάστηκε στο ραντεβού

Πίνακας 60: Συνοπτική περιγραφή των χαρακτηριστικών του συνόλου δεδομένων Alvaro Flores (μετά την προεπεξεργασία)

Το σύνολο δεδομένων είναι μη-ισορροπημένο καθώς όπως φαίνεται στην Εικόνα 34, από τα 61.213 στιγμιότυπα, τα 48.346 (79%) ανήκουν στην κλάση No (πλειοψηφική κλάση), ενώ τα υπόλοιπα 12.867 (21%) ανήκουν στην κλάση Yes (Μειοψηφική κλάση).



Εικόνα 34: Κατανομή κλάσεων του συνόλου δεδομένων Alvaro Flores

7

Πειραματική μελέτη

7.1 Πειραματική Διαδικασία

Η διαδικασία που ακολουθήθηκε είναι:

- Προεπεξεργασία των δεδομένων
- Επιλογή σημαντικότερων χαρακτηριστικών
- Οργάνωση – εκτέλεση πειραμάτων
    - Διαχωρισμός του συνόλου δεδομένων σε τμήματα εκπαίδευσης και ελέγχου (k-fold Cross validation)
    - Εφαρμογή αλγορίθμου υπερδειγματοληψίας SMOTE (SMOTE-N για τα nominal χαρακτηριστικά).
    - Εφαρμογή αλγορίθμου κατηγοριοποίησης και βελτιστοποίηση παραμέτρων
- Αξιολόγηση αποτελεσμάτων με βάση τις μετρικές precision – recall και F-measure.

7.1.1 Προεπεξεργασία των δεδομένων

Ο έλεγχος και η προεπεξεργασία των δεδομένων που περιλαμβάνονται στα σύνολα δεδομένων (Alvaro Flores, Joni Hoppen), περιγράφηκε αναλυτικά στο Κεφάλαιο 6.

7.1.2 Επιλογή σημαντικότερων χαρακτηριστικών (Select attributes)

Για τον εντοπισμό των σημαντικότερων χαρακτηριστικών χρησιμοποιήθηκε η τεχνική Select attributes στο WEKA. Συγκεκριμένα εφαρμόστηκαν οι αλγόριθμοι InfoGainAttributeEval και GainRatioAttributeEval, οι οποίοι έδωσαν για κάθε σύνολο δεδομένων μία λίστα όλων των χαρακτηριστικών (Jovic et al., 2015). Στη λίστα αυτή τα χαρακτηριστικά κατατάχτηκαν με βάση τη σημαντικότητα τους από το περισσότερο σημαντικό προς το λιγότερο σημαντικό. Αναλυτικότερα στο σύνολο δεδομένων Alvaro Flores τα λιγότερο σημαντικά χαρακτηριστικά ήταν το scheduled_month_d και το scheduled_weekday_d. Τα πιο πάνω χαρακτηριστικά αφαιρέθηκαν από το σύνολο δεδομένων και στη συνέχεια εκτελέστηκαν δειγματοληπτικά κάποια πειράματα για να διαπιστωθεί, εάν η αλλαγή αυτή βελτιώνει τα αποτελέσματα. Η ίδια διαδικασία ακολουθήθηκε και για το σύνολο δεδομένων Joni Hoppen, όπου τα λιγότερο σημαντικά χαρακτηριστικά ήταν το Handicap και Alcoholism.

Τα πειράματα έδειξαν ότι και στα δύο σύνολα δεδομένων τα αποτελέσματα δεν βελτιώθηκαν μετά από την αφαίρεση των παραπάνω χαρακτηριστικών. Για το λόγο αυτό, τα πειράματα συνεχίστηκαν με τα αρχικά σύνολα δεδομένων.

7.1.3 Οργάνωση πειραμάτων - εκτέλεση πειραμάτων

7.1.3.1 Εφαρμογή SMOTE

Επειδή τα διαθέσιμα σύνολα δεδομένων είναι μη-ισορροπημένα, χρησιμοποιήθηκε η μέθοδος υπερδειγματοληψίας SMOTE ώστε να γίνουν ισορροπημένα. Επιπλέον εκτελέστηκαν και πειράματα χωρίς την εφαρμογή SMOTE για σύγκριση των αποτελεσμάτων.

Ο αλγόριθμος SMOTE έχει δύο παραμέτρους, το πλήθος κοντινότερων γειτόνων k και το ποσοστό αύξησης της μειοψηφικής κλάσης N. Για το k η προεπιλεγμένη τιμή είναι k=5, αλλά δοκιμάστηκαν και άλλες τιμές για να ελεγχθεί αν η επιλογή του k επηρεάζει το αποτέλεσμα της κατηγοριοποίησης. Οι τιμές που επιλέχθηκαν για τα πειράματα είναι οι 2, 5, 7 και 10.

Όσον αφορά το N, η τιμή του εξαρτάται από την κατανομή των στιγμιότυπων στις δύο κλάσεις, οπότε είναι διαφορετική για κάθε σύνολο δεδομένων. Στο άρθρο του SMOTE αναφέρεται ότι το N θεωρείται ότι παίρνει τιμές που είναι ακέραια πολλαπλάσια του 100, δηλαδή 100, 200, 300 κλπ. Συνεπώς πρέπει να επιλεγεί μία συγκεκριμένη τιμή για το ποσοστό αύξησης, που να έχει ως αποτέλεσμα το πλήθος της μειοψηφικής κλάσης να φτάσει όσο γίνεται πιο κοντά στο πλήθος της πλειοψηφικής. Με αυτό τον τρόπο όμως είναι πολύ πιθανόν και πάλι να υπάρχει μια ανισοκατανομή στα πλήθη των στιγμιότυπων των δύο κλάσεων. Παρ' όλα αυτά, το εργαλείο WEKA επιτρέπει τη χρήση οποιασδήποτε

πραγματικής τιμής για το N, έτσι ώστε να μπορεί να γίνουν οι δύο κλάσεις σχεδόν τέλεια ισοκατανεμημένες.

7.1.3.2 Cross Validation και συνδυασμός με SMOTE

Για κάθε ένα από τα σύνολα δεδομένων που περιγράφηκαν παραπάνω, γίνεται διαχωρισμός των δεδομένων σε τμήματα εκπαίδευσης και ελέγχου σύμφωνα με την μέθοδο διασταυρούμενης επικύρωσης k-fold Cross Validation. Συγκεκριμένα χρησιμοποιήθηκε η μέθοδος 10-fold Cross Validation, καθώς είναι από τις πιο συχνά χρησιμοποιούμενες στην βιβλιογραφία.

Ο συνδυασμός υπερδειγματοληψίας με SMOTE και μεθόδου Cross Validation πρέπει να γίνει με σωστό τρόπο ώστε τα αποτελέσματα να είναι αξιόπιστα. Ένα συνηθισμένο σφάλμα είναι να εφαρμόζεται πρώτα υπερδειγματοληψία σε ολόκληρο το σύνολο δεδομένων και στη συνέχεια να χωρίζονται τα δεδομένα σε σύνολα εκπαίδευσης και ελέγχου μέσω του Cross Validation (Santos et al., 2018).

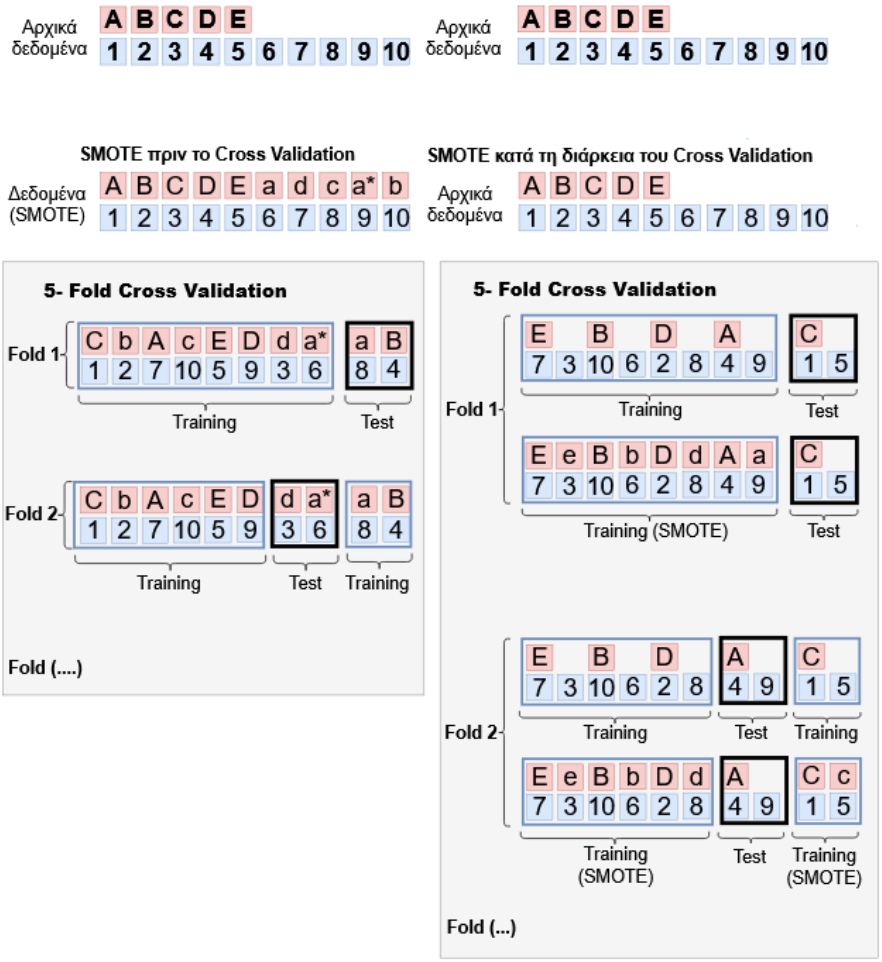
Αυτή η προσέγγιση έχει ως αποτέλεσμα να υπάρχει μεγάλη πιθανότητα να εμφανιστούν και στο σύνολο εκπαίδευσης και στο σύνολο ελέγχου στιγμιότυπα τα οποία να προέρχονται από τα ίδια αρχικά στιγμιότυπα προ της υπερδειγματοληψίας. Αν συμβεί αυτό, τα δεδομένα του συνόλου ελέγχου δεν θα είναι πραγματικά άγνωστα όπως θα έπρεπε. Κατά συνέπεια ο κατηγοριοποιητής θα εμφανίσει λιγότερα σφάλματα στο σύνολο ελέγχου επειδή θα είναι εκπαιδευμένος να κατηγοριοποιεί σχεδόν «ίδια» δεδομένα, δίνοντας έτσι μια εσφαλμένη υπερεκτιμημένη ικανότητα γενίκευσης.

Η λύση σε αυτή την κατάσταση είναι να μην εφαρμόζεται η υπερδειγματοληψία SMOTE πριν από το Cross Validation, αλλά ταυτόχρονα με αυτό. Στην προσέγγιση αυτή, πρώτα χωρίζεται το σύνολο δεδομένων σε k ίσα τμήματα αλλά χρησιμοποιώντας την παραλλαγή της στρωματοποιημένης διασταυρούμενης επικύρωσης (Stratified Cross Validation). Με αυτή την παραλλαγή του Cross Validation τα τμήματα που θα προκύψουν θα περιέχουν μεν στιγμιότυπα επιλεγμένα με τυχαίο τρόπο, αλλά η κατανομή τους στις δύο κλάσεις θα είναι ίδια με αυτή του αρχικού συνόλου δεδομένων. Από τα συνολικά k τμήματα του αρχικού συνόλου δεδομένων, τα k-1 τμήματα θα αποτελούν πλέον το σύνολο εκπαίδευσης και το 1 τμήμα θα αποτελεί το σύνολο ελέγχου. Στη συνέχεια ο αλγόριθμος SMOTE εφαρμόζεται μόνο στο σύνολο εκπαίδευσης και έτσι το σύνολο ελέγχου δεν τροποποιείται. (Δηλαδή περιέχει άγνωστα για τον κατηγοριοποιητή δεδομένα). Ταυτόχρονα με την εφαρμογή του SMOTE το σύνολο εκπαίδευσης είναι πλέον ισορροπημένο. Με αυτό τον τρόπο ο αλγόριθμος κατηγοριοποίησης θα εκπαιδευτεί σε ισορροπημένο σύνολο εκπαίδευσης και στη συνέχεια θα αξιολογηθεί ως προς την επίδοση του στο σύνολο ελέγχου. Το συγκεκριμένο σύνολο ελέγχου

θα περιέχει άγνωστα δεδομένα που προέρχονται από το αρχικό σύνολο δεδομένων και δεν έχουν πειραχτεί με κανένα τρόπο.

Οι δύο προσεγγίσεις για την εφαρμογή του SMOTE σε συνδυασμό με 5-fold Cross Validation απεικονίζονται σε απλοποιημένο παράδειγμα στην Εικόνα 35. Έστω ότι υπάρχουν 15 στιγμιότυπα που ανήκουν σε δύο κλάσεις: γράμματα και αριθμοί. Στην πρώτη προσέγγιση όπου το SMOTE γίνεται πριν το Cross Validation, ο αλγόριθμος εφαρμόζεται σε ολόκληρο το σύνολο δεδομένων. Η μειοψηφική κλάση (γράμματα) υπόκειται σε υπερδειγματοληψία και έτσι παράγονται κάποια συνθετικά στιγμιότυπα (συμβολίζονται με μικρά γράμματα). Για παράδειγμα από το στιγμιότυπο A παράγεται το a, από το D το d κοκ. Επίσης παρατηρείται ότι το στιγμιότυπο A επιλέχθηκε δύο φορές και παρήγαγε δύο συνθετικά πρότυπα a και a* που είναι παρόμοια μεταξύ τους.

Στη συνέχεια ξεκινώντας τη διαδικασία του Cross Validation, το ισορροπημένο πλέον σύνολο δεδομένων χωρίζεται τυχαία σε 5 τμήματα και το ένα από αυτά επιλέγεται ως σύνολο ελέγχου (Test set), ενώ τα υπόλοιπα ως σύνολο εκπαίδευσης (Training set). Παρατηρείται ότι κατά τον τυχαίο διαχωρισμό, το συνθετικό πρότυπο a τοποθετήθηκε στο Test set, ενώ το συνθετικό πρότυπο a* μαζί με το αρχικό του A παρέμεινε στο Training set. Σε αυτή λοιπόν την περίπτωση ο αλγόριθμος κατηγοριοποίησης που θα εκπαιδευτεί στο Training set και θα γνωρίζει τα παρόμοια μεταξύ τους πρότυπα A και a*, θα κληθεί στη συνέχει να προβλέψει την κλάση για το επίσης παρόμοιο πρότυπο a. Αυτό φυσικά θα του δώσει λανθασμένα ένα επιπλέον πλεονέκτημα ως προς την ορθή κατηγοριοποίηση του προτύπου a. Το ίδιο φαινόμενο παρατηρείται και στο Fold 2 όπου το πρότυπο a* επιλέγεται να συμμετέχει στο Test set, ενώ το αρχικό πρότυπο A και το παρόμοιο συνθετικό a είναι στο Training set. Εδώ επιπλέον στο Test set συμμετέχει και το συνθετικό πρότυπο d, ενώ το αρχικό του παρόμοιο πρότυπο D συμμετέχει στο Training set, δίνοντας και πάλι άδικο πλεονέκτημα. Φαίνεται λοιπόν ότι με αυτή την προσέγγιση σε κάθε Fold υπάρχει μεγάλη πιθανότητα να υπάρχουν παρόμοια δεδομένα τόσο στο Training set όσο και στο Test set, δίνοντας έτσι την ψευδαίσθηση της καλύτερης επίδοσης κατά την κατηγοριοποίηση.



Εικόνα 35: Παράδειγμα συνδυασμού Cross Validation και SMOTE

Στην δεύτερη προσέγγιση, ξεκινάει πρώτα η διαδικασία του Cross Validation οπότε το αρχικό σύνολο δεδομένων χωρίζεται σε 5 τμήματα που είναι stratified, δηλαδή η αναλογία των προτύπων του αρχικού συνόλου διατηρείται και στα τμήματα. Για το Fold 1 επιλέγονται τέσσερα τμήματα ως Training set και ένα ως Test set. Στη συνέχεια μόνο στο Training set του Fold 1 εφαρμόζεται ο αλγόριθμος SMOTE, ενώ το Test set διατηρείται ως έχει. Το τελικό Training set του Fold 1 περιέχει πλέον και τα συνθετικά πρότυπα a, b, d και e και είναι ισορροπημένο. Στο Test set υπάρχουν μόνο πρότυπα που ανήκουν στα αρχικά δεδομένα και σίγουρα δεν υπάρχουν στο Training set παρόμοια τους που προέκυψαν με υπερδειγματοληψία. Η ίδια διαδικασία ακολουθείται και για τα υπόλοιπα Folds.

Με την προσέγγιση της εφαρμογής του αλγορίθμου SMOTE κατά τη διάρκεια του Cross Validation επιτυγχάνονται πιο ρεαλιστικά αποτελέσματα, καθώς έτσι προσομοιάζεται η χρήση των αλγορίθμων κατηγοριοποίησης σε δεδομένα του «πραγματικού κόσμου» με βάση την ικανότητα γενίκευσης, δηλαδή την ικανότητα να πραγματοποιούν σωστή πρόβλεψη σε «νέα άγνωστα» δεδομένα.

Στο λογισμικό WEKA η παραπάνω διαδικασία μπορεί να υλοποιηθεί με την βοήθεια του metalearner κατηγοριοποιητή FilteredClassifier σε συνδυασμό με k-fold Cross Validation.

Στην συγκεκριμένη διπλωματική εργασία για τα πειράματα χρησιμοποιήθηκε ο metalearner κατηγοριοποιητής FilteredClassifier σε συνδυασμό με 10-fold cross validation. Καθώς χρησιμοποιήθηκε η τεχνική 10-fold cross validation, σε κάθε fold το Training set αποτελούνταν από το 90% του αρχικού συνόλου δεδομένων και το υπόλοιπο 10% ήταν το Test set.

Έτσι, ειδικά για το SMOTE κατά τη διάρκεια του Cross Validation, επειδή εφαρμόζεται μόνο στο Training set, θα πρέπει να υπολογιστούν τα ποσοστά αύξησης της μειοψηφικής κλάσης μόνο για το 90% των αρχικών δεδομένων.

Συνεπώς, στο σύνολο δεδομένων Joni Hoppen η κλάση Yes περιλαμβάνει 22.314 στιγμιότυπα (μειοψηφική κλάση) και η κλάση No 88.207 στιγμιότυπα (πλειοψηφική κλάση).

Κρατώντας το 90% από αυτά για το Training set στην κλάση Yes θα υπάρχουν πλέον 20.082 στιγμιότυπα και στην κλάση No 79.387 στιγμιότυπα.

Έτσι, στο Training set τα στιγμιότυπα της κλάσης Yes θα πρέπει να αυξηθούν μέσω υπερδειγματοληψίας κατά 79.387-20.082 = 59.305 πρότυπα. Αυτή η αύξηση σε σχέση με το αρχικό πλήθος της κλάσης Yes αντιστοιχεί σε ποσοστό:

$$N_{Joni\ Hoppen(CV-90\%)} = \frac{79.387 - 20.082}{20.082} = \frac{59.305}{20.082} = 2{,}9531 \text{ ή } 295{,}31\%$$

Αντίστοιχα, στο σύνολο δεδομένων Alvaro Flores η κλάση Yes περιλαμβάνει 12.867 στιγμιότυπα (μειοψηφική κλάση) και η κλάση No 48.346 στιγμιότυπα (πλειοψηφική κλάση).

Κρατώντας το 90% από αυτά για το Training set στην κλάση Yes θα υπάρχουν πλέον 11.580 στιγμιότυπα και στην κλάση No 43.512 στιγμιότυπα.

Έτσι, στο Training set τα στιγμιότυπα της κλάσης Yes θα πρέπει να αυξηθούν μέσω υπερδειγματοληψίας κατά 43.512-11.580 =31.932. Αυτή η αύξηση σε σχέση με το αρχικό πλήθος της κλάσης Yes αντιστοιχεί σε ποσοστό:

$$N_{Alvaro\ Flores(CV-90\%)} = \frac{43.512 - 11.580}{11.580} = \frac{31.932}{11.580} = 2{,}7575 \text{ ή } 275{,}75\%$$

7.1.3.3 Εκτέλεση πειραμάτων κατηγοριοποίησης

Σύμφωνα με όσα περιεγράφηκαν παραπάνω, για κάθε ένα από τα σύνολα των δεδομένων, εκτός από την αρχική μορφή υπάρχουν και άλλες 4 που προέκυψαν από την εφαρμογή του

SMOTE για k=2, 5, 7, 10 γείτονες. Συνολικά λοιπόν για κάθε σύνολο δεδομένων προέκυψαν 5 μορφές.

Σε κάθε μία από αυτές εφαρμόστηκαν οι αλγόριθμοι κατηγοριοποίησης k-NN, J48 και Naive Bayes. Για κάθε αλγόριθμο έγιναν πειράματα για διάφορες τιμές των παραμέτρων τους με σκοπό να βρεθούν οι βέλτιστες τιμές τους (Parameter Tuning). Συγκεκριμένα για τον αλγόριθμο k-NN δοκιμάστηκαν για την παράμετρο k οι τιμές 1, 5, 9, 13, 17, 21, 25, 29, 33, 37, 41, 45, 49, 53, 57, 61, 65, 69, 81, 101, 201, 301, 401, 501, 601, 701, 801 και 1201. Για τον αλγόριθμο J48 δοκιμάστηκαν τιμές της παραμέτρου CF από 0,05 έως 0,55 με βήματα ανά 0,05, δηλαδή συνολικά 11 περιπτώσεις. Τέλος ο αλγόριθμος Naive Bayes δεν έχει παραμέτρους οπότε κατ' εξαίρεση μόνο για αυτόν έγιναν περισσότερα πειράματα για επιπλέον τιμές του k στον αλγόριθμο SMOTE. Συγκεκριμένα δοκιμάστηκαν οι τιμές από 1 έως 10. Σε όλα τα πειράματα καταγράφηκαν διάφορες μετρικές όπως Precision, Recall, F-measure, Accuracy κλπ.

7.1.3.4 Συνολικά πειράματα που πραγματοποιήθηκαν

1. **Dataset "Medical Appointment" του Alvaro Flores.**

   - Έγιναν 140 πειράματα με τον k-NN, δηλαδή τον IBK
   - Έγιναν 55 πειράματα με τον J48, δηλαδή τον C4.5
   - Έγιναν 11 πειράματα με τον Naive Bayes

   ΣΥΝΟΛΟ ΠΕΙΡΑΜΑΤΩΝ: 206

2. **Dataset "Medical Appointment No Shows" του Joni Hoppen**
   - Έγιναν 140 πειράματα με τον k-NN, δηλαδή τον IBK
   - Έγιναν 55 πειράματα με τον J48, δηλαδή τον C4.5
   - Έγιναν 11 πειράματα με τον Naive Bayes

   ΣΥΝΟΛΟ ΠΕΙΡΑΜΑΤΩΝ: 206

Συνολικά λοιπόν πραγματοποιήθηκαν 412 πειράματα.

8

Αποτελέσματα πειραμάτων

8.1 Πειραματική μελέτη για την 1η ενέργεια

8.1.1 Σύνολο δεδομένων Alvaro Flores - Πειράματα με SMOTE και χωρίς SMOTE

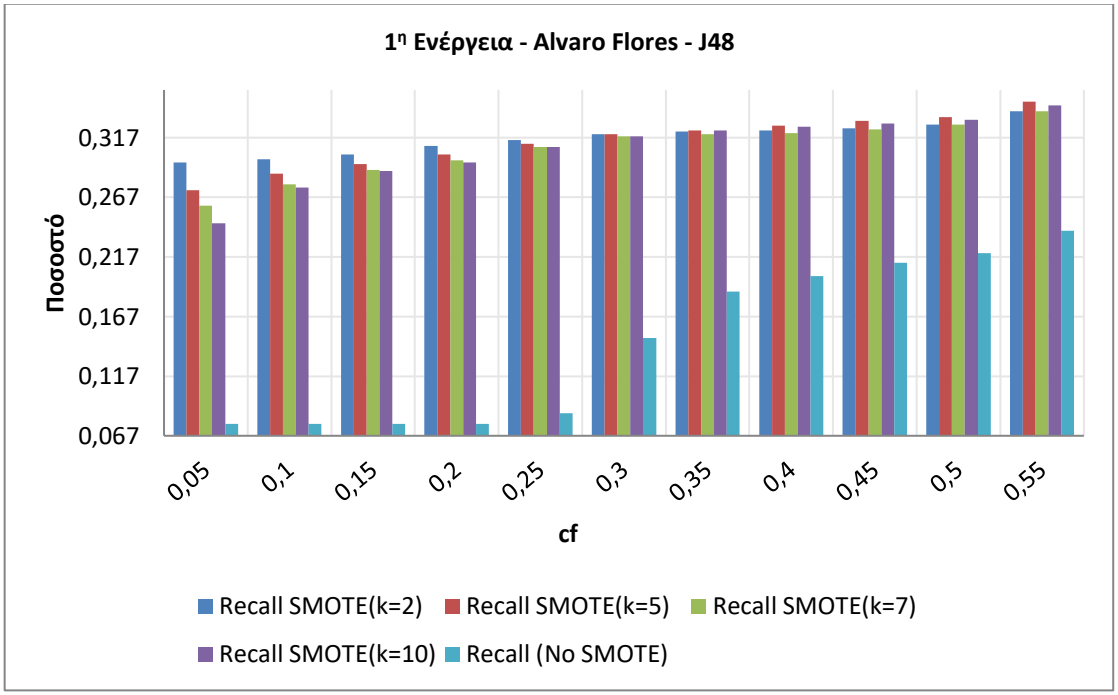
8.1.1.1 Εύρεση βέλτιστων παραμέτρων (Parameter Tuning)

Για την εφαρμογή της 1ης ενέργειας στόχος είναι η μεγιστοποίηση της μετρικής Recall όπως αναλύθηκε διεξοδικά στην παράγραφο 4.3.1 (1η Ενέργεια: Υπενθύμιση του ραντεβού). Για τον εντοπισμό των βέλτιστων παραμέτρων για κάθε αλγόριθμο ακολουθήθηκε η εξής διαδικασία. Με βάση τα πειράματα που πραγματοποιήθηκαν για διάφορες τιμές των παραμέτρων του κάθε αλγορίθμου και της παραμέτρου κ της τεχνικής SMOTE και αφού καταγράφηκαν οι αντίστοιχες τιμές τους για την μετρική Recall, στη συνέχεια έγινε σύγκριση των τιμών της μετρικής Recall από όλα τα πειράματα και εντοπίστηκε η μέγιστη τιμή. Στη συνέχεια καταγράφηκαν οι τιμές των παραμέτρων των αλγορίθμων στις οποίες παρουσιάστηκε η συγκεκριμένη μέγιστη τιμή Recall.

Στο επόμενο βήμα έγινε έλεγχος της μέγιστης τιμής της μετρικής Recall, ώστε να διαπιστωθεί εάν η αντίστοιχη τιμή της μετρικής Precision είναι πολύ χαμηλή. Εάν λοιπόν η τιμή

παραμέτρου του αλγορίθμου που δίνει την μέγιστη τιμή στην μετρική Recall παρουσιάσει πολύ χαμηλή τιμή στην μετρική Precision, τότε η συγκεκριμένη τιμή της μετρικής Precision θεωρείται μη αποδεκτή. Κατά συνέπεια επιλέγεται κάποια άλλη τιμή για την παράμετρο του αλγορίθμου (πχ άλλο k για τον K-NN), η οποία μπορεί να μην δώσει την μέγιστη τιμή , αλλά μία εξίσου καλή τιμή Recall και ταυτόχρονα αποδεκτή τιμή και για την μετρική Precision. Μετά από την παραπάνω διαδικασία η βέλτιστη τιμή παραμέτρου για κάθε έναν από τους 3 αλγορίθμους (K-NN, J48, Naïve Bayes) μαζί με την αντίστοιχη τιμή των μετρικών Precision, Recall, F-measure και Accuracy καταγράφηκαν συγκεντρωτικά σε ένα διάγραμμα (Διάγραμμα 6), ώστε να διαπιστωθεί ποιος αλγόριθμος δίνει τελικά καλύτερα αποτελέσματα.

Αλγόριθμος J48 (C4.5)



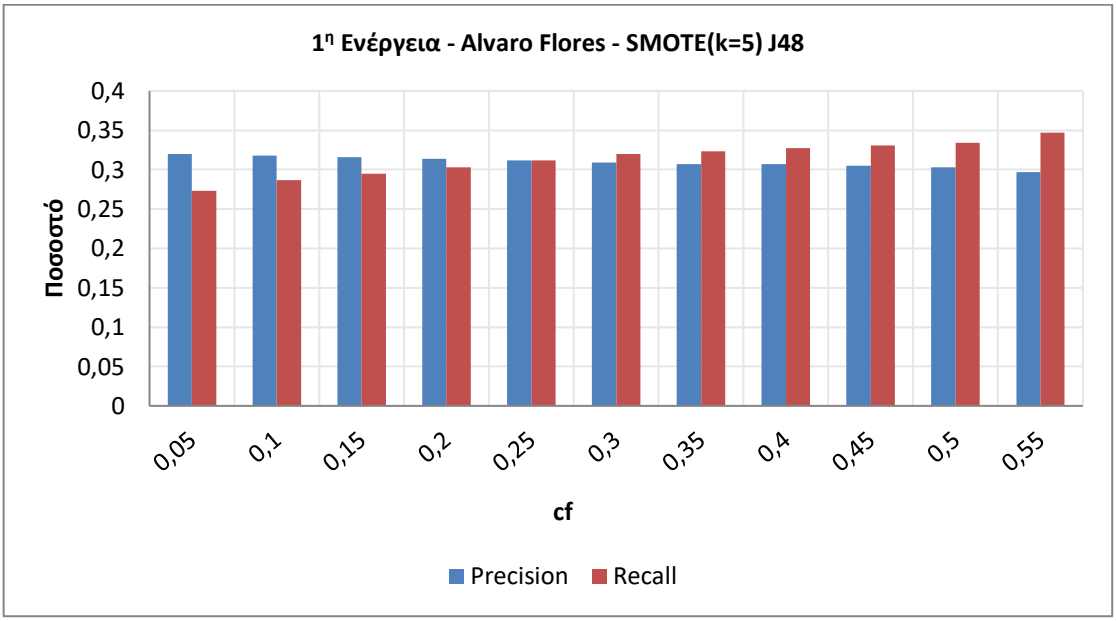
Διάγραμμα 1: Αποτελέσματα αλγορίθμου J48 για 1η ενέργεια. Σύγκριση Recall (χωρίς SMOTE) και Recall (με SMOTE K=2,5,7,10 γείτονες) για όλες τις τιμές CF.

Στο Διάγραμμα 1 παρουσιάζονται τα αποτελέσματα του Recall για τον αλγόριθμο J48 (C4.5) χωρίς SMOTE και με SMOTE με K=2,5,7,10 γείτονες. Με βάση το διάγραμμα επιλέγεται η υψηλότερη τιμή Recall (max Recall= 0,347), η οποία προκύπτει όταν εφαρμοστεί SMOTE με K= 5 γείτονες και Confidence Factor (CF)= 0,55.

Επιπλέον με βάση και πάλι το παραπάνω διάγραμμα αυτό που παρατηρεί κανείς είναι ότι όσο οι τιμές του CF μειώνονται, αντίστοιχα μειώνονται και οι τιμές της μετρικής Recall σε όλες τις περιπτώσεις, δηλαδή και με SMOTE και χωρίς SMOTE. Το CF όπως προαναφέρθηκε στη θεωρία ελέγχει τον βαθμό κλαδέματος του δέντρου. Επομένως μικρότερη τιμή CF σημαίνει μεγαλύτερο κλάδεμα, ενώ μεγαλύτερη τιμή CF σημαίνει μικρότερο κλάδεμα. Στη

συγκεκριμένη λοιπόν περίπτωση προκύπτει το συμπέρασμα ότι το περισσότερο κλάδεμα, μειώνει τα αποτελέσματα, δηλαδή την τιμή της μετρικής Recall.

Αυτό που επίσης αξίζει να σημειωθεί, είναι ότι στη συγκεκριμένη περίπτωση η εφαρμογή της τεχνικής SMOTE στο σύνολο δεδομένων, πριν την κατηγοριοποίηση με τον αλγόριθμο J48, βελτιώνει αισθητά την μετρική Recall σε όλες τις τιμές της παραμέτρου CF.

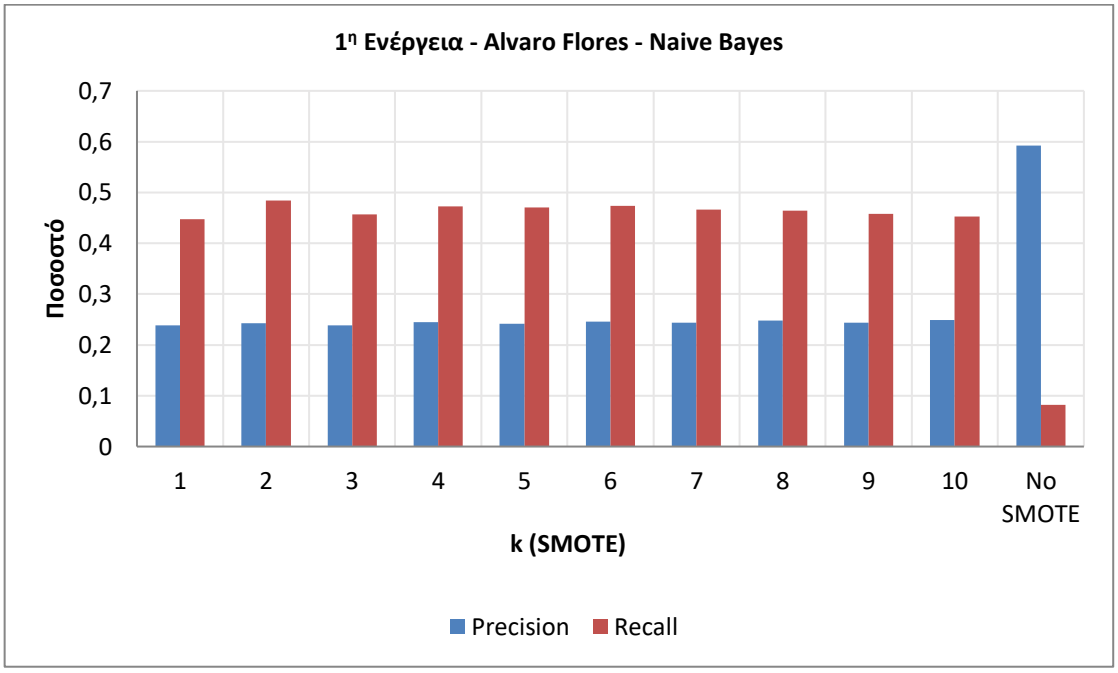


Διάγραμμα 2: Αποτελέσματα αλγορίθμου J48 για 1η ενέργεια. Σύγκριση Precision-Recall με SMOTE K=2 γείτονες για όλες τις τιμές CF.

Αφού επιλέχθηκε η υψηλότερη τιμή Recall (max Recall) με βάση το Διάγραμμα 1, στη συνέχεια θα πρέπει να ελεγχθεί και η αντίστοιχη τιμή του Precision, ώστε να εξασφαλιστεί, ότι αυτή δεν είναι πολύ χαμηλή. Στο Διάγραμμα 2 λοιπόν παρατηρείται ότι στην τιμή CF=0,55 με μέγιστη τιμή Recall=0,347 η αντίστοιχη τιμή της μετρικής Precision = 0,297 δεν είναι πολύ χαμηλή. Κατά συνέπεια θεωρείται αποδεκτή.

Επομένως οι βέλτιστες τιμές παραμέτρων για τον αλγόριθμο J48 είναι CF=0,55 με SMOTE K= 5 γείτονες που δίνουν Recall= 0,347 και Precision= 0,297.

Αλγόριθμος Naïve Bayes

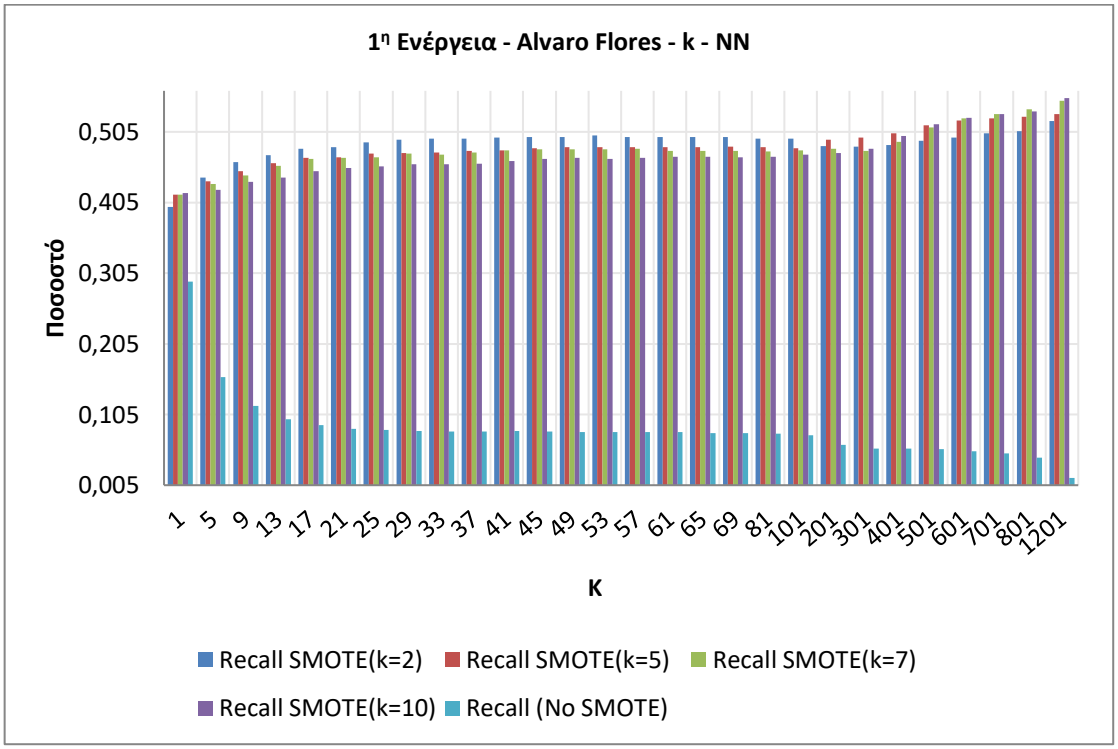


Διάγραμμα 3: Αποτελέσματα αλγορίθμου Naïve Bayes για 1η ενέργεια. Σύγκριση Precision-Recall χωρίς SMOTE και με SMOTE (K=1 έως 10 γείτονες).

Στο Διάγραμμα 3 παρουσιάζονται τα αποτελέσματα των μετρικών Recall-Precision για τον αλγόριθμο Naïve Bayes χωρίς SMOTE και με SMOTE με K= 1 έως 10 γείτονες. Με βάση το παραπάνω διάγραμμα επιλέγεται η υψηλότερη τιμή Recall (max Recall= 0,484), η οποία προκύπτει όταν εφαρμοστεί SMOTE με K= 2 γείτονες. Η αντίστοιχη τιμή της μετρικής Precision = 0,243 δεν θεωρείται πολύ χαμηλή. Κατά συνέπεια θεωρείται αποδεκτή. Επομένως η βέλτιστή τιμή της παραμέτρου k του SMOTE όταν εφαρμόζεται ο αλγόριθμος Naïve Bayes, είναι K= 2 γείτονες δίνοντας Recall= 0,484 και Precision= 0,243.

Σε αυτή την περίπτωση η εφαρμογή της τεχνικής SMOTE στο σύνολο δεδομένων, πριν την κατηγοριοποίηση με τον αλγόριθμο Naïve Bayes, βελτιώνει αισθητά την μετρική του Recall σε όλες τις τιμές k του SMOTE.

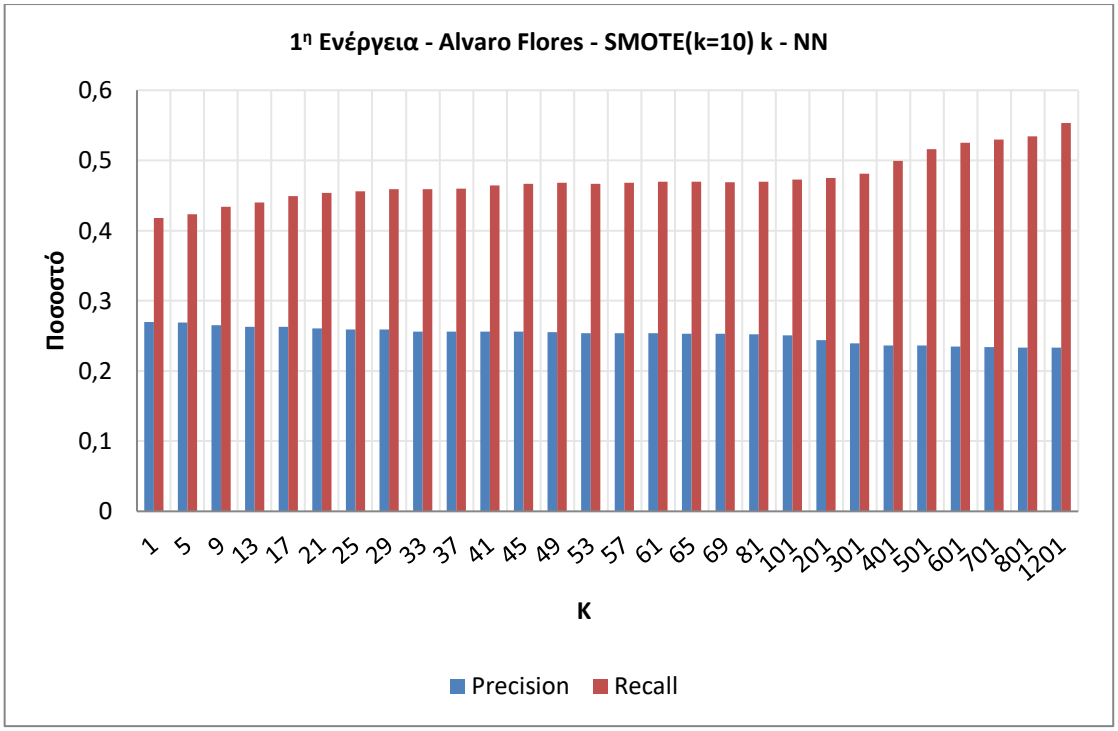
Αλγόριθμος k-NN



Διάγραμμα 4: Αποτελέσματα αλγορίθμου k-NN για 1η ενέργεια. Σύγκριση Recall (χωρίς SMOTE) και Recall (με SMOTE K=2,5,7,10 γείτονες) για όλες τις τιμές του k (k-NN).

Στο Διάγραμμα 4 παρουσιάζονται τα αποτελέσματα του Recall για τον αλγόριθμο k-NN χωρίς SMOTE και με SMOTE με K=2,5,7,10 γείτονες. Σύμφωνα με το διάγραμμα επιλέγεται η υψηλότερη τιμή Recall (max Recall= 0,553), η οποία προκύπτει όταν εφαρμοστεί SMOTE με K= 10 γείτονες και k-NN με k=1201 γείτονες.

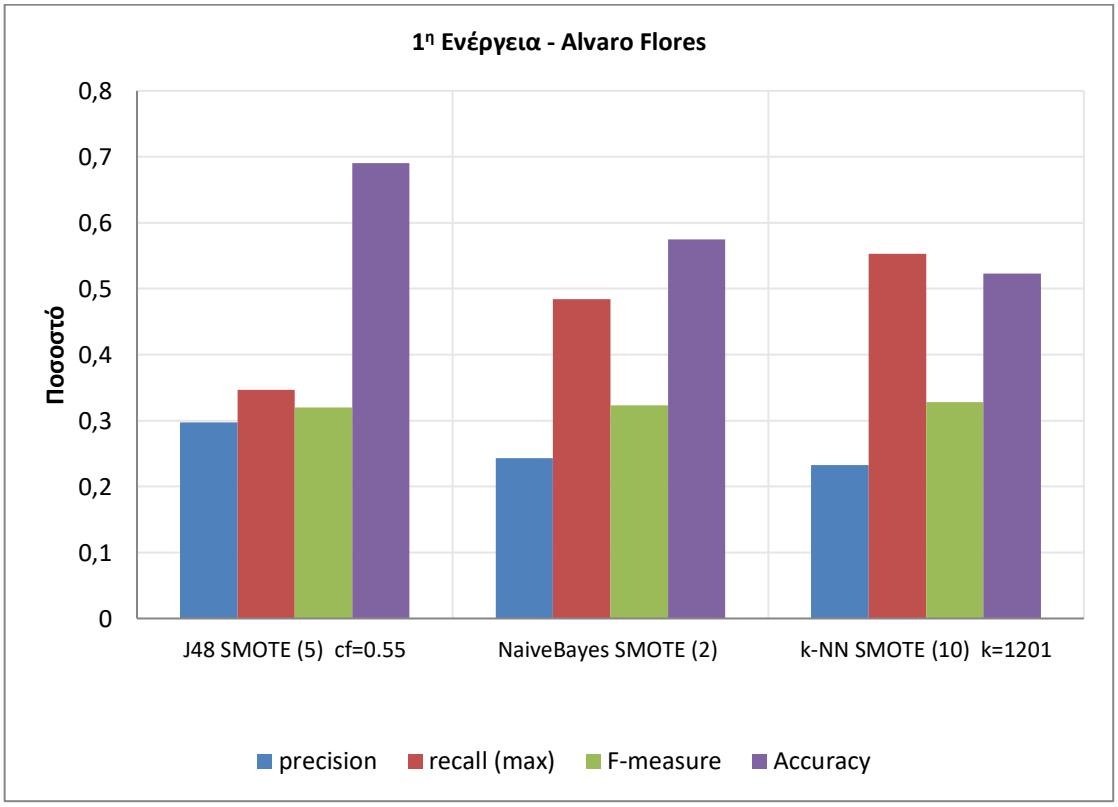
Αυτό που αξίζει να σημειωθεί, είναι ότι στη συγκεκριμένη περίπτωση η εφαρμογή της τεχνικής SMOTE στο σύνολο δεδομένων, πριν την κατηγοριοποίηση με τον αλγόριθμο k-NN, βελτιώνει αισθητά την μετρική Recall σε όλες τις τιμές της παραμέτρου k του αλγορίθμου k-NN .



Διάγραμμα 5: Αποτελέσματα αλγορίθμου k-NN για 1η ενέργεια . Σύγκριση Precision-Recall με SMOTE K=10 γείτονες για όλες τις τιμές του k (k-NN).

Αφού επιλέχθηκε η υψηλότερη τιμή Recall (max Recall) με βάση το Διάγραμμα 4, στη συνέχεια θα πρέπει να ελεγχθεί και η αντίστοιχη τιμή του Precision, ώστε να εξασφαλιστεί, ότι αυτή δεν είναι πολύ χαμηλή. Στο Διάγραμμα 5 λοιπόν παρατηρείται ότι στην τιμή k=1201 γείτονες του k-NN και SMOTE με k=10 γείτονες με μέγιστη τιμή Recall= 0,553, η αντίστοιχη τιμή της μετρικής Precision = 0,233 δεν είναι πολύ χαμηλή. Κατά συνέπεια θεωρείται αποδεκτή. Επομένως οι βέλτιστες τιμές παραμέτρων για τον αλγόριθμο k-NN είναι k (k-NN)= 1201 γείτονες με k (SMOTE)= 10 γείτονες δίνοντας Recall= 0,553 και Precision= 0,233.

8.1.1.2 Τελικά αποτελέσματα



Διάγραμμα 6: Τελικά αποτελέσματα αλγορίθμων k-NN, J48, Naïve Bayes για 1η ενέργεια. Σύγκριση μετρικών Recall (max), Precision, F-measure, Accuracy.

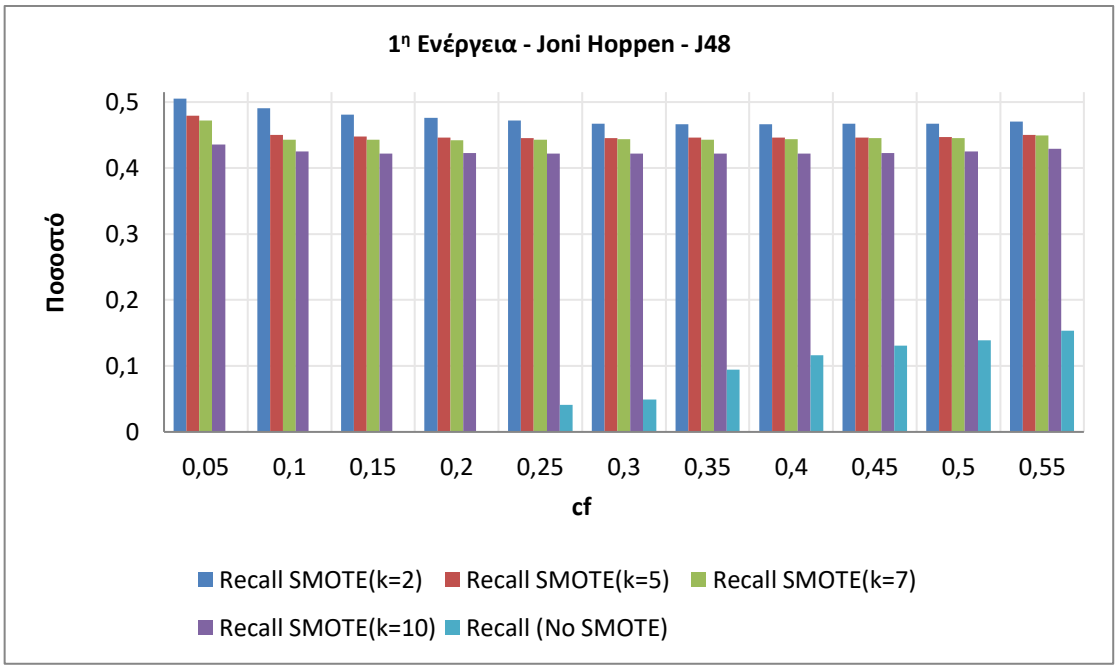
Στο Διάγραμμα 6 παρουσιάζονται τα αποτελέσματα των αλγορίθμων k-NN, J48 και Naïve Bayes με την βέλτιστη τιμή για την παράμετρο κάθε αλγορίθμου (k του k-NN, CF του J48 και k του SMOTE). Η βέλτιστη τιμή για κάθε παράμετρο προέκυψε από τη διαδικασία Εύρεσης βέλτιστων παραμέτρων (παράγραφος 8.1.1.1) και επιλέχθηκε με βάση την μέγιστη τιμή στην μετρική Recall (recall max), ενώ παράλληλα το Precision παραμένει σε αποδεκτά επίπεδα. Επιπλέον στο παραπάνω διάγραμμα παρουσιάζονται και οι αντίστοιχες τιμές των μετρικών Precision, F-measure και Accuracy. Η μέγιστη τιμή για την μετρική Recall=0,553 παρουσιάζεται στον αλγόριθμο k-NN με τιμή παραμέτρου k=1201 γείτονες σε συνδυασμό με εφαρμογή της τεχνικής SMOTE με k=10 γείτονες. Το δεύτερο καλύτερο αποτέλεσμα το δίνει ο αλγόριθμος Naïve Bayes και πάλι σε συνδυασμό με εφαρμογή της τεχνικής SMOTE με k=2 γείτονες με μέγιστη τιμή για την μετρική Recall=0,484. Από τα παραπάνω προκύπτει ότι η διαφορά μεταξύ των δύο αλγορίθμων ως προς τη μετρική Recall είναι 0,069 ή 6,9%. Δηλαδή σχετικά μικρή.

8.1.2 Σύνολο δεδομένων Joni Hoppen – Πειράματα με SMOTE και χωρίς SMOTE

8.1.2.1 Εύρεση βέλτιστων παραμέτρων (Parameter Tuning)

Η διαδικασία για την Εύρεση βέλτιστων παραμέτρων (παράγραφος 8.1.1.1) όσον αφορά το σύνολο δεδομένων Joni Hoppen είναι η ίδια που ακολουθήθηκε στο σύνολο δεδομένων Alvaro Flores και παρουσιάζεται αναλυτικά παρακάτω.

Αλγόριθμος J48 (C4.5)



Διάγραμμα 7: Αποτελέσματα αλγορίθμου J48 για 1η ενέργεια. Σύγκριση Recall (χωρίς SMOTE) και Recall (με SMOTE K=2,5,7,10 γείτονες) για όλες τις τιμές CF.

Στο Διάγραμμα 7 παρουσιάζονται τα αποτελέσματα του Recall για τον αλγόριθμο J48 (C4.5) χωρίς SMOTE και με SMOTE με K=2,5,7,10 γείτονες. Με βάση το διάγραμμα επιλέγεται η υψηλότερη τιμή Recall (max Recall= 0,505), η οποία προκύπτει όταν εφαρμοστεί SMOTE με K= 2 γείτονες και Confidence Factor (CF)= 0,05.

Αυτό που αξίζει να σημειωθεί, είναι ότι στη συγκεκριμένη περίπτωση η εφαρμογή της τεχνικής SMOTE στο σύνολο δεδομένων, πριν την κατηγοριοποίηση με τον αλγόριθμο J48, βελτιώνει αισθητά την μετρική του Recall σε όλες τις τιμές της παραμέτρου CF. Ειδικότερα δε, στην περιοχή τιμών CF=0,05 έως 0,20 παρατηρείται ότι ο αλγόριθμος J48 χωρίς SMOTE δίνει μηδενικά αποτελέσματα για την μετρική Recall.

Αντίθετα στην ίδια περιοχή τιμών μετά από την εφαρμογή του SMOTE παρουσιάζει Recall από 0,505 έως 0,476. Δηλαδή κατά την εφαρμογή του αλγορίθμου J48 με SMOTE

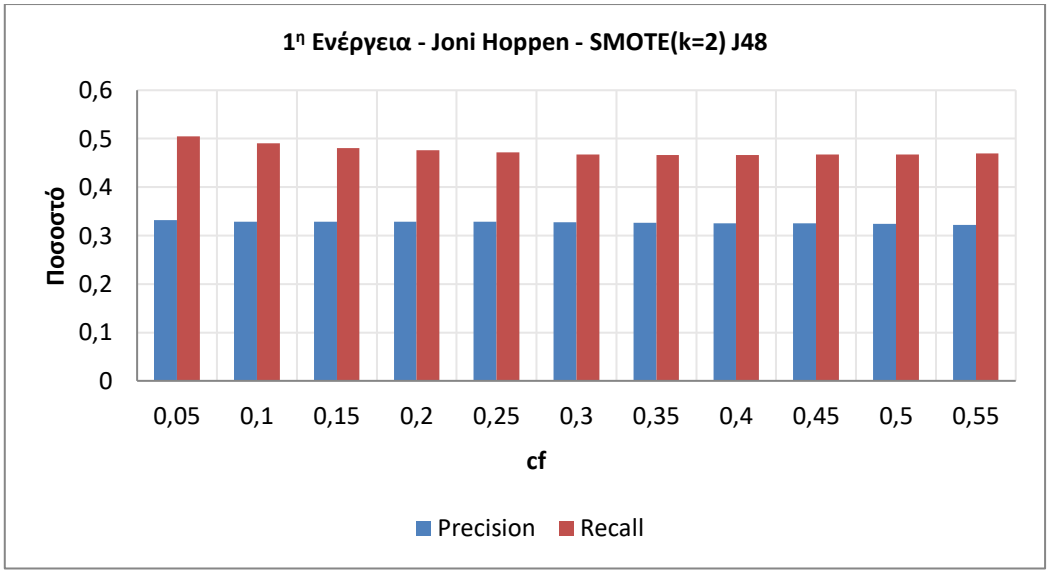
παρατηρείται βελτίωση κατά 50,5% στην μετρική Recall σε σύγκριση με τον αν θα εφαρμοζόταν ο αλγόριθμος χωρίς SMOTE.

Όσον αφορά τις μηδενικές τιμές της μετρικής Recall στις τιμές CF=0,05 έως 0,2 οι οποίες αναφέρθηκαν παραπάνω, το αποτέλεσμα αυτό προκύπτει γιατί σε αυτές τις περιπτώσεις ο αλγόριθμος J48 δημιουργεί δέντρο με ένα μόνο φύλλο που ανήκει στην κλάση No. Δηλαδή προβλέπει ότι όλοι οι ασθενείς θα εμφανισθούν. Κατά συνέπεια το Recall είναι μηδέν γιατί δεν προέβλεψε σωστά κανέναν από τους ασθενείς που στην πραγματικότητα δεν εμφανίσθηκαν.

Αντίστοιχα η τιμή της μετρικής Precision αν και δεν φαίνεται στο παραπάνω διάγραμμα, είναι ακαθόριστη, διότι έγιναν 0 προβλέψεις για την κλάση Yes. Δηλαδή ο κατηγοριοποιητής δεν προέβλεψε κανέναν ασθενή ότι δεν θα εμφανισθεί στο ραντεβού. Σε αυτή λοιπόν την περίπτωση ο υπολογισμός του Precision θα οδηγούσε σε διαίρεση με το μηδέν, πράγμα που δεν γίνεται. (Recall=TP/TP+FN=0/0 + κάποια τιμή=0, Precision=TP/TP+FP=0/0+0= ακαθόριστο).

Τέλος με βάση και πάλι το παραπάνω διάγραμμα αυτό που παρατηρεί κανείς είναι ότι, στην περίπτωση που δεν εφαρμόζεται τεχνική SMOTE, όσο οι τιμές του CF μειώνονται, αντίστοιχα μειώνονται και οι τιμές της μετρικής Recall. Στην περίπτωση όμως που εφαρμόζεται τεχνική SMOTE, όσο οι τιμές του CF μειώνονται, αντίθετα οι τιμές της μετρικής Recall αυξάνονται. Το CF όπως προαναφέρθηκε στη θεωρία ελέγχει τον βαθμό κλαδέματος του δέντρου. Επομένως μικρότερη τιμή CF σημαίνει μεγαλύτερο κλάδεμα, ενώ μεγαλύτερη τιμή CF σημαίνει μικρότερο κλάδεμα.

Άρα λοιπόν στην περίπτωση εφαρμογής του αλγορίθμου J48 χωρίς SMOTE προκύπτει το συμπέρασμα, ότι το περισσότερο κλάδεμα δίνει χειρότερα αποτελέσματα. Αντίθετα στην περίπτωση εφαρμογής με SMOTE το περισσότερο κλάδεμα βελτιώνει τα αποτελέσματα σε μικρό βαθμό.

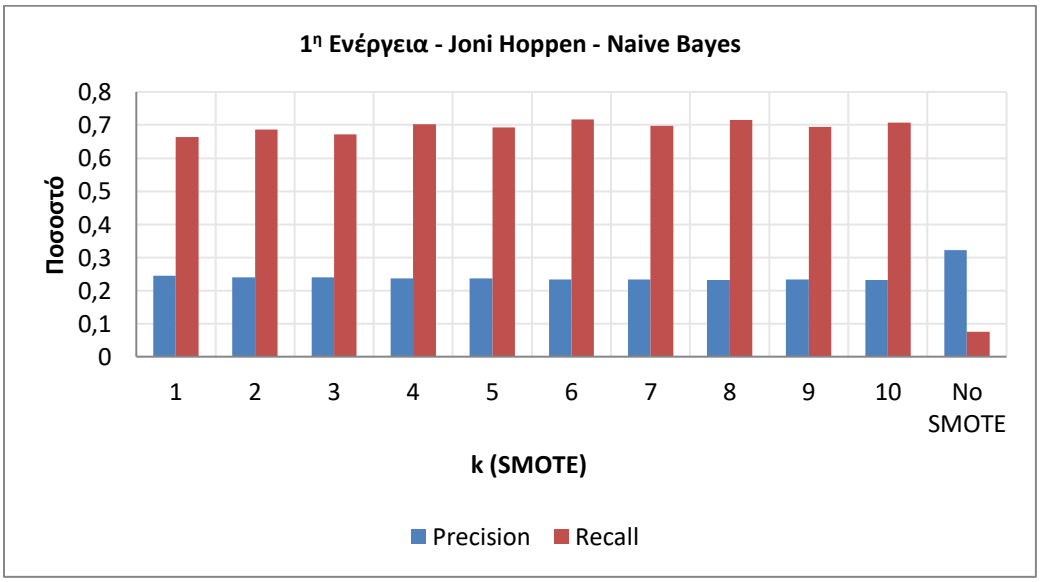


Διάγραμμα 8: Αποτελέσματα αλγορίθμου J48 για 1η ενέργεια. Σύγκριση Precision-Recall με SMOTE K=2 γείτονες για όλες τις τιμές CF

Αφού επιλέχθηκε η υψηλότερη τιμή Recall (max Recall) με βάση το Διάγραμμα 7, στη συνέχεια θα πρέπει να ελεγχθεί και η αντίστοιχη τιμή του Precision, ώστε να εξασφαλιστεί, ότι αυτή δεν είναι πολύ χαμηλή. Στο Διάγραμμα 8 λοιπόν παρατηρείται ότι στην τιμή CF=0,05 με μέγιστη τιμή Recall=0,505 η αντίστοιχη τιμή της μετρικής Precision = 0,332 δεν είναι πολύ χαμηλή. Κατά συνέπεια θεωρείται αποδεκτή.

Επομένως οι βέλτιστες τιμές παραμέτρων για τον αλγόριθμο J48 είναι CF=0,05 με SMOTE K= 2 γείτονες που δίνουν Recall= 0,505 και Precision= 0,332.

Αλγόριθμος Naïve Bayes

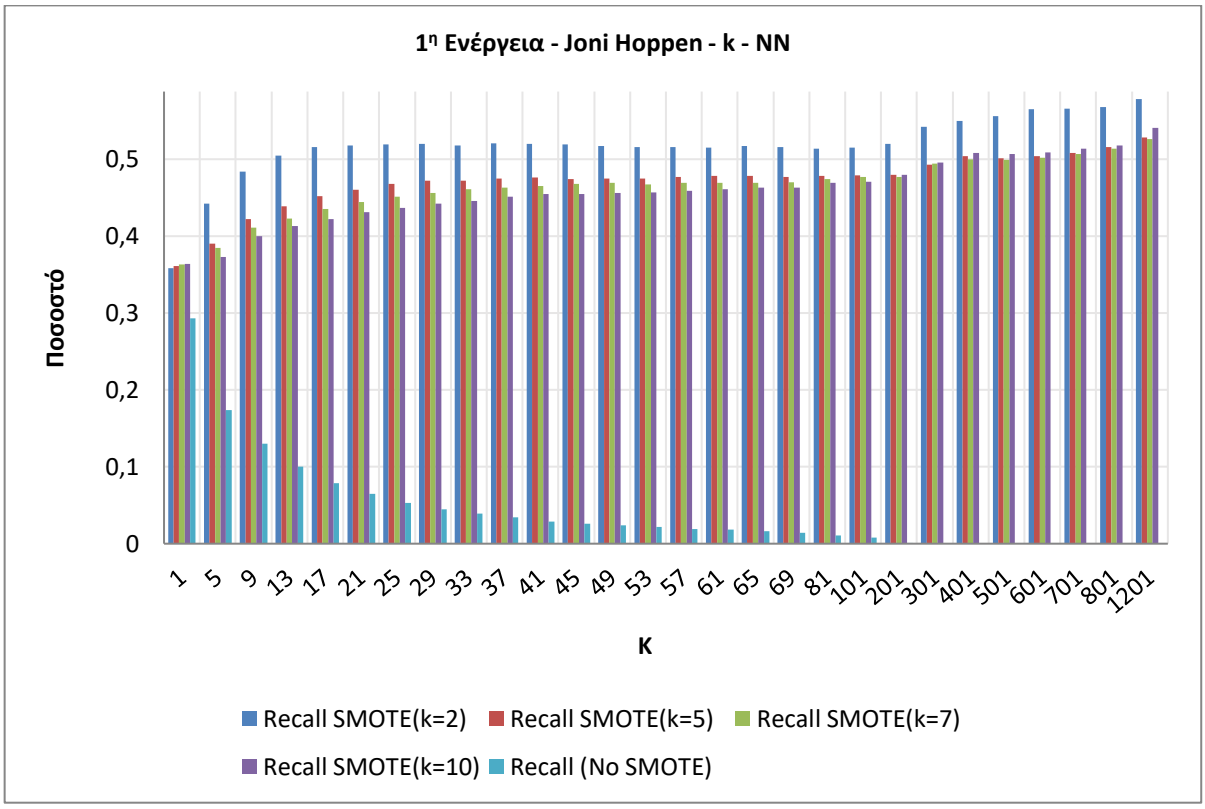


Διάγραμμα 9: Αποτελέσματα αλγορίθμου Naïve Bayes για 1η ενέργεια. Σύγκριση Precision-Recall χωρίς SMOTE και με SMOTE (K=1 έως 10 γείτονες).

Στο Διάγραμμα 9 παρουσιάζονται τα αποτελέσματα των μετρικών Recall-Precision για τον αλγόριθμο Naïve Bayes χωρίς SMOTE και με SMOTE με K= 1 έως 10 γείτονες. Με βάση το παραπάνω διάγραμμα επιλέγεται η υψηλότερη τιμή Recall (max Recall= 0,717), η οποία προκύπτει όταν εφαρμοστεί SMOTE με K= 6 γείτονες. Η αντίστοιχη τιμή της μετρικής Precision = 0,233 δεν θεωρείται πολύ χαμηλή. Κατά συνέπεια θεωρείται αποδεκτή. Επομένως η βέλτιστή τιμή της παραμέτρου k του SMOTE όταν εφαρμόζεται ο αλγόριθμος Naïve Bayes, είναι K= 6 γείτονες δίνοντας Recall= 0,717 και Precision= 0,233.

Από τα παραπάνω προκύπτει το συμπέρασμα ότι η εφαρμογή της τεχνικής SMOTE στο σύνολο δεδομένων, πριν την κατηγοριοποίηση με τον αλγόριθμο Naïve Bayes, βελτιώνει αισθητά την μετρική του Recall σε όλες τις τιμές k του SMOTE.

Αλγόριθμος k-NN



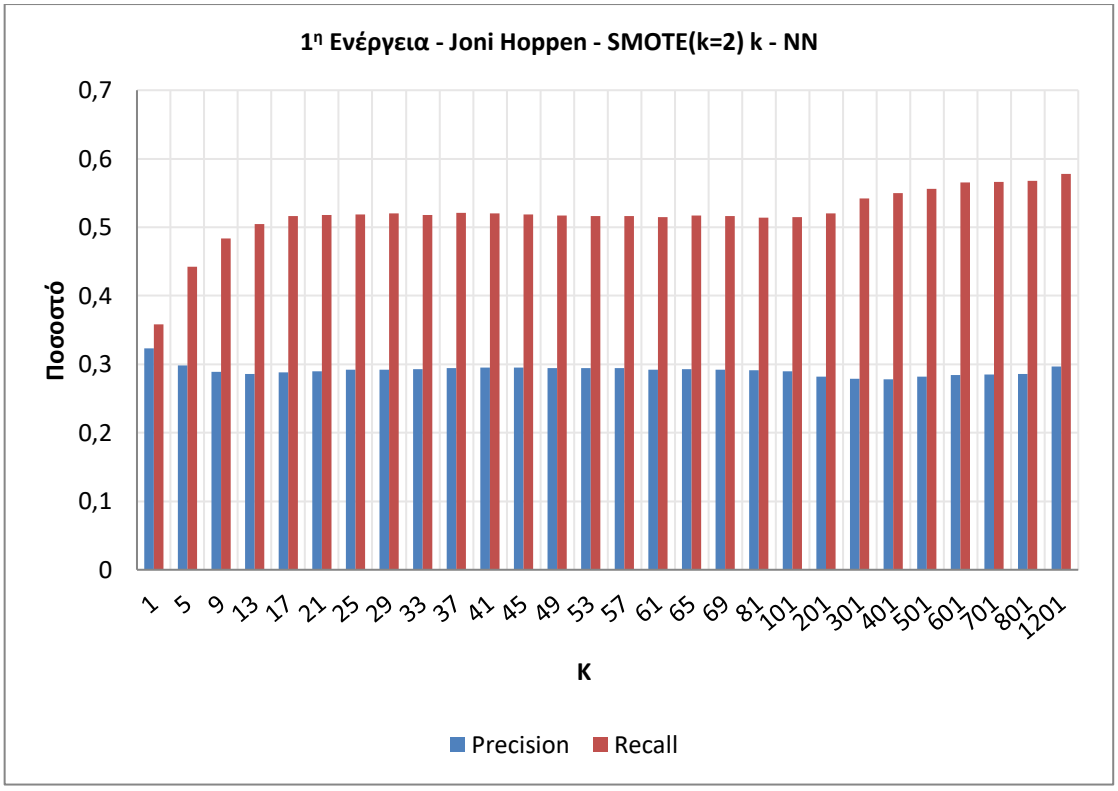
Διάγραμμα 10: Αποτελέσματα αλγορίθμου k-NN για 1η ενέργεια. Σύγκριση Recall (χωρίς SMOTE) και Recall (με SMOTE K=2,5,7,10 γείτονες) για όλες τις τιμές του k (k-NN).

Στο Διάγραμμα 10 παρουσιάζονται τα αποτελέσματα του Recall για τον αλγόριθμο k-NN χωρίς SMOTE και με SMOTE με K=2,5,7,10 γείτονες. Σύμφωνα με το διάγραμμα επιλέγεται η υψηλότερη τιμή Recall (max Recall= 0,578), η οποία προκύπτει όταν εφαρμοστεί SMOTE με K= 2 γείτονες και k-NN με k=1201 γείτονες.

Αυτό που αξίζει να σημειωθεί, είναι ότι στη συγκεκριμένη περίπτωση η εφαρμογή της τεχνικής SMOTE στο σύνολο δεδομένων, πριν την κατηγοριοποίηση με τον αλγόριθμο k-

NN, βελτιώνει αισθητά την μετρική του Recall σε όλες τις τιμές της παραμέτρου k του αλγορίθμου k-NN . Ειδικότερα δε, για τις τιμές της παραμέτρου κ=301 έως 1201 γείτονες του k-NN, παρατηρείται ότι ο αλγόριθμος k-NN χωρίς SMOTE δίνει μηδενικά αποτελέσματα για την μετρική Recall. Αντίθετα στην ίδια περιοχή τιμών μετά από την εφαρμογή του SMOTE και συγκεκριμένα όταν εφαρμόζεται SMOTE με k=2 γείτονες, ο αλγόριθμος k-NN παρουσιάζει Recall από 0,542 έως 0,578. Δηλαδή κατά την εφαρμογή του αλγορίθμου k-NN με SMOTE παρατηρείται βελτίωση κατά 57,8% στην μετρική Recall σε σύγκριση με τον αν θα εφαρμοζόταν ο αλγόριθμος χωρίς SMOTE.

Όσον αφορά τις μηδενικές τιμές της μετρικής Recall για τιμές k (k-NN)= 301 έως 1201 γείτονες, οι οποίες αναφέρθηκαν παραπάνω, ισχύει ότι περιγράφηκε και στην περίπτωση του J48. Δηλαδή ο κατηγοριοποιητής προβλέπει, ότι όλοι οι ασθενείς θα εμφανισθούν στο ραντεβού, άρα ότι όλα τα πρότυπα ανήκουν στην κλάση No. Κατά συνέπεια το Recall είναι μηδέν γιατί δεν προέβλεψε σωστά κανέναν από τους ασθενείς που στην πραγματικότητα δεν εμφανίσθηκαν. Επομένως στην περίπτωση του αλγορίθμου k-NN οι μηδενικές τιμές μπορεί να οφείλονται στο γεγονός ότι το σύνολο δεδομένων Joni Hoppen είναι μη ισορροπημένο και τα περισσότερα πρότυπα ανήκουν στην κλάση No. Επομένως στις μεγάλες τιμές k (k-NN) το πιθανότερο είναι η πλειοψηφία των γειτόνων να ανήκει πάντα στην κλάση No.

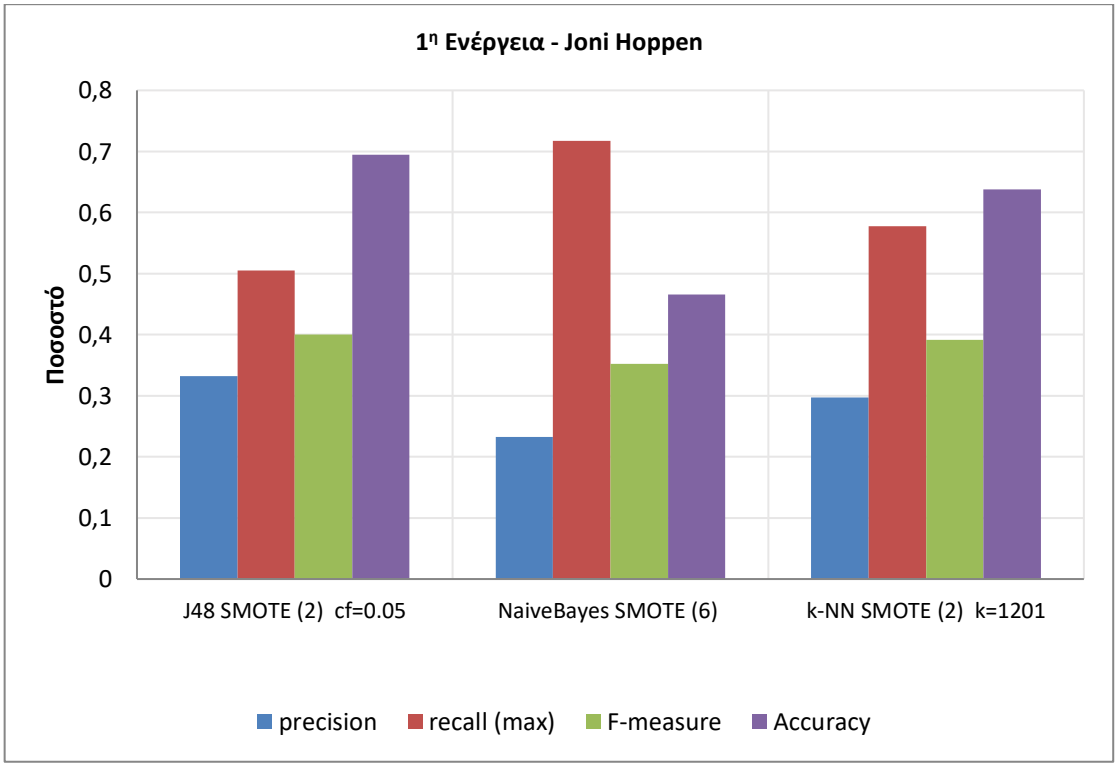


Διάγραμμα 11: Αποτελέσματα αλγορίθμου k-NN για 1η ενέργεια. Σύγκριση Precision-Recall με SMOTE K=2 γείτονες για όλες τις τιμές του k (k-NN).

Αφού επιλέχθηκε η υψηλότερη τιμή Recall (max Recall) με βάση το Διάγραμμα 10, στη συνέχεια θα πρέπει να ελεγχθεί και η αντίστοιχη τιμή του Precision, ώστε να εξασφαλιστεί, ότι αυτή δεν είναι πολύ χαμηλή. Στο Διάγραμμα 11 λοιπόν παρατηρείται ότι στην τιμή k=1201 γείτονες του k-NN και SMOTE με k=2 γείτονες με μέγιστη τιμή Recall= 0,578, η αντίστοιχη τιμή της μετρικής Precision = 0,297 δεν είναι πολύ χαμηλή. Κατά συνέπεια θεωρείται αποδεκτή.

Επομένως οι βέλτιστες τιμές παραμέτρων για τον αλγόριθμο k-NN είναι k (k-NN)= 1201 γείτονες με k (SMOTE)= 2 γείτονες δίνοντας Recall= 0,578 και Precision= 0,297.

8.1.2.2 Τελικά αποτελέσματα

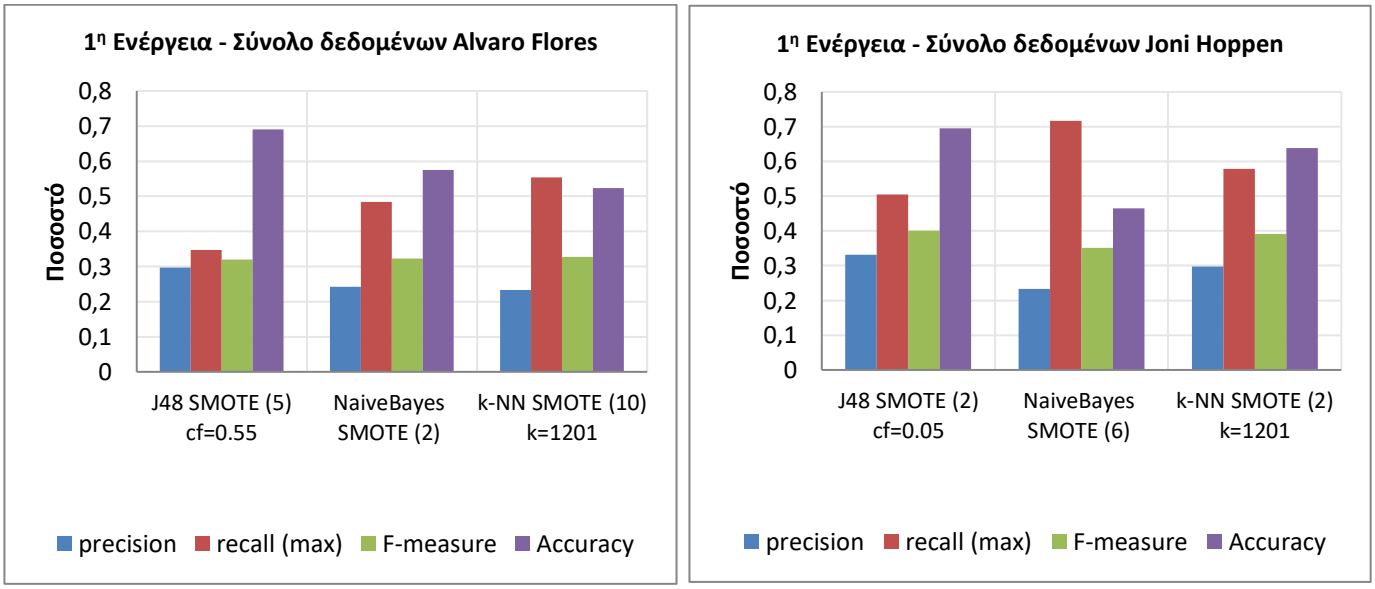


Διάγραμμα 12: Τελικά αποτελέσματα αλγορίθμων k-NN, J48, Naïve Bayes για 1η ενέργεια. Σύγκριση μετρικών Recall (max), Precision, F-measure, Accuracy.

Στο Διάγραμμα 12 παρουσιάζονται τα αποτελέσματα των αλγορίθμων k-NN, J48 και Naïve Bayes με την βέλτιστη τιμή για την παράμετρο κάθε αλγορίθμου (k του k-NN, CF του J48 και k του SMOTE). Η βέλτιστη τιμή για κάθε παράμετρο προέκυψε από τη διαδικασία Εύρεσης βέλτιστων παραμέτρων (παράγραφος 8.1.2.1) και επιλέχθηκε με βάση την μέγιστη τιμή στην μετρική Recall (recall max), ενώ παράλληλα το Precision παραμένει σε αποδεκτά επίπεδα. Επιπλέον στο παραπάνω διάγραμμα παρουσιάζονται και οι αντίστοιχες τιμές των μετρικών Precision, F-measure και Accuracy. Η μέγιστη τιμή για την μετρική Recall=0,717 παρουσιάζεται στον αλγόριθμο Naïve Bayes όταν εφαρμόζεται σε αυτόν τεχνική SMOTE με

k= 6 γείτονες. Το δεύτερο καλύτερο αποτέλεσμα το δίνει ο αλγόριθμος k-NN όταν εφαρμόζεται σε αυτόν τιμή παραμέτρου k=1201 γείτονες και τεχνική SMOTE με k=2 γείτονες, δίνοντας Recall=0,578. Από τα παραπάνω προκύπτει ότι η διαφορά μεταξύ των δύο αλγορίθμων ως προς τη μετρική Recall είναι 0,139 ή 13,9%. Δηλαδή σχετικά μικρή.

8.1.3 Συμπεράσματα για την 1η ενέργεια



Διάγραμμα 13: 1η Ενέργεια (Υπενθύμιση ραντεβού): Συγκεντρωτικά αποτελέσματα για τα σύνολα δεδομένων Alvaro Flores και Joni Hoppen

Με γνώμονα αυτά που ειπώθηκαν στις προηγούμενες παραγράφους, εξάγονται τα παρακάτω συμπεράσματα. Αρχικά θα πρέπει να αναφερθεί ότι το μέγιστο Recall που είναι και το ζητούμενο για την 1η ενέργεια (4.3.1) παρουσιάζεται στα 2 σύνολα δεδομένων σε διαφορετικούς αλγορίθμους. Στο σύνολο δεδομένων Joni Hoppen το μέγιστο Recall=0,717 σημειώνεται στον αλγόριθμο Naïve Bayes με SMOTE με k=6 γείτονες. Αντίθετα στο σύνολο δεδομένων Alvaro Flores το μέγιστο Recall=0,553 καταγράφεται στον αλγόριθμο k-NN με k=1201 γείτονες και SMOTE με k=10 γείτονες.

Ο Αλγόριθμος J48 ως προς την μετρική Recall εμφανίζει την χειρότερη επίδοση και στα 2 σύνολα δεδομένων. Σε γενικές γραμμές με βάση το Διάγραμμα 13 παρατηρείται ότι και οι 3 αλγόριθμοι παρουσιάζουν καλύτερο Recall στο σύνολο δεδομένων Joni Hoppen. Αναλυτικότερα, όπως προαναφέρθηκε, ο αλγόριθμος Naïve Bayes εμφανίζει Recall=0,717 στο σύνολο δεδομένων Joni Hoppen και Recall=0,484 στο σύνολο δεδομένων Alvaro Flores. Δηλαδή η διαφορά μεταξύ τους είναι 0,233 ή 23,3%. Από την άλλη πλευρά ο αλγόριθμος k-NN σημειώνει Recall=0,578 στο σύνολο δεδομένων Joni Hoppen και Recall=0,553 στο σύνολο δεδομένων Alvaro Flores. Επομένως η διαφορά τους είναι 0,025 ή 2,5%. Τέλος ο αλγόριθμος J48 παρουσιάζει Recall=0,505 στο σύνολο δεδομένων Joni Hoppen και

Recall=0,347 στο σύνολο δεδομένων Alvaro Flores. Κατά συνέπεια η διαφορά μεταξύ τους είναι ίση με 0,158 ή 15,8%. Επομένως θα ήταν εύλογο να ισχυριστεί κανείς ότι η 1η ενέργεια θα μπορούσε να εφαρμοστεί αποδοτικότερα στο σύνολο δεδομένων Joni Hoppen.

Σημειώνεται ότι αν και το ζητούμενο στην 1η ενέργεια είναι το υψηλότερο Recall, το Precision δεν πρέπει να είναι σε πολύ χαμηλά επίπεδα. Συνεπώς σε αυτή την παράγραφο, παρουσιάζονται αποτελέσματα πειραμάτων όπου η τιμή της μετρικής Recall είναι η υψηλότερη δυνατή, διατηρώντας παράλληλα την τιμή της μετρικής Precision σε αποδεκτό επίπεδο.

Στον παρακάτω πίνακα φαίνονται επίσης συγκεντρωτικά οι μέγιστες και βέλτιστες τιμές που προέκυψαν από τη διαδικασία του grid search για την 1η ενέργεια και για τους 3 αλγορίθμους κατηγοριοποίησης και για τα 2 σύνολα δεδομένων (Πίνακας 61).

Σύνολο δεδομένων	Αλγόριθμος	Παράμετρος	Τεχνική (SMOTE)	MAX Recall	MAX Precision	Optimum Recall	Optimum Precision
Alvaro Flores	J48	Cf=0,55	SMOTE (K=5)	0,347	0,297	0,347	0,297
	Naïve Bayes	-	SMOTE (K=2)	0,484	0,243	0,484	0,243
	k-NN	**K=1201**	**SMOTE (K=10)**	**0,553**	**0,233**	**0,553**	**0,233**
Joni Hoppen	J48	Cf=0,05	SMOTE (K=2)	0,505	0,332	0,505	0,332
	Naïve Bayes	**-**	**SMOTE (K=6)**	**0,717**	**0,233**	**0,717**	**0,233**
	k-NN	K=1201	SMOTE (K=2)	0,578	0,297	0,578	0,297

Πίνακας 61: Τελικά αποτελέσματα για 1η ενέργεια (Μέγιστες και βέλτιστες τιμές)

8.2 Πειραματική μελέτη για την 2η ενέργεια

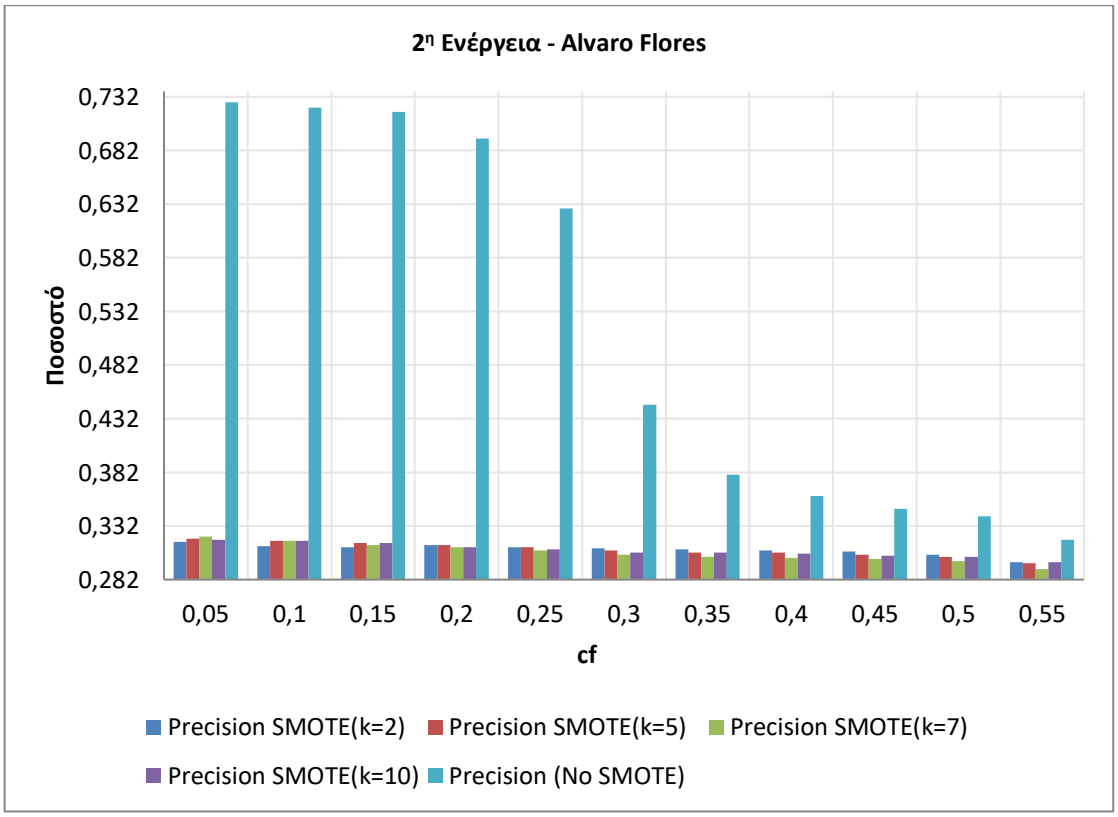
8.2.1 Σύνολο δεδομένων Alvaro Flores - Πειράματα με SMOTE και χωρίς SMOTE

8.2.1.1 Εύρεση βέλτιστων παραμέτρων (Parameter Tuning)

Για την εφαρμογή της 2ης ενέργειας στόχος είναι η μεγιστοποίηση της μετρικής Precision όπως αναλύθηκε διεξοδικά στην παράγραφο 4.3.2 (2η Ενέργεια: Κάλυψη των πιθανών κενών που θα προκύψουν). Για τον εντοπισμό των βέλτιστων παραμέτρων για κάθε αλγόριθμο ακολουθήθηκε η εξής διαδικασία. Με βάση τα πειράματα που πραγματοποιήθηκαν για διάφορες τιμές των παραμέτρων του κάθε αλγορίθμου και της παραμέτρου κ της τεχνικής SMOTE και αφού καταγράφηκαν οι αντίστοιχες τιμές τους για την μετρική Precision, στη συνέχεια έγινε σύγκριση των τιμών της μετρικής Precision από όλα τα πειράματα και

εντοπίστηκε η μέγιστη τιμή. Στη συνέχεια καταγράφηκαν οι τιμές των παραμέτρων των αλγορίθμων στις οποίες παρουσιάστηκε η συγκεκριμένη μέγιστη τιμή Precision. Στο επόμενο βήμα έγινε έλεγχος της μέγιστης τιμής της μετρικής Precision, ώστε να διαπιστωθεί εάν η αντίστοιχη τιμή της μετρικής Recall είναι πολύ χαμηλή. Εάν λοιπόν η τιμή παραμέτρου του αλγορίθμου που δίνει την μέγιστη τιμή στην μετρική Precision παρουσιάσει πολύ χαμηλή τιμή στην μετρική Recall, τότε η συγκεκριμένη τιμή της μετρικής Recall θεωρείται μη αποδεκτή. Κατά συνέπεια επιλέγεται κάποια άλλη τιμή για την παράμετρο του αλγορίθμου (πχ άλλο k για τον K-NN), η οποία μπορεί να μην δώσει την μέγιστη τιμή , αλλά μία εξίσου καλή τιμή Precision και ταυτόχρονα αποδεκτή τιμή και για την μετρική Recall. Μετά από την παραπάνω διαδικασία η βέλτιστη τιμή παραμέτρου για κάθε έναν από τους 3 αλγορίθμους (K-NN, J48, Naïve Bayes) μαζί με την αντίστοιχη τιμή των μετρικών Precision, Recall, F-measure και Accuracy καταγράφηκαν συγκεντρωτικά σε ένα διάγραμμα, ώστε να διαπιστωθεί ποιος αλγόριθμος δίνει τελικά καλύτερα αποτελέσματα.

Αλγόριθμος J48 (C4.5)



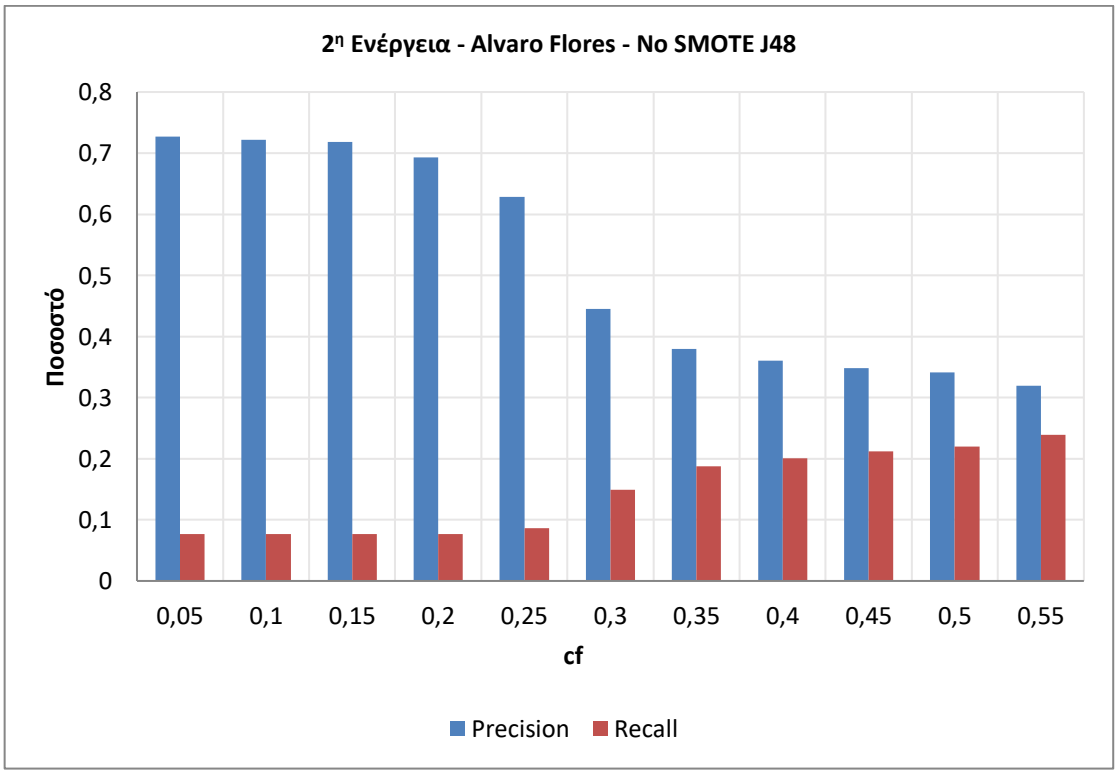
Διάγραμμα 14: Αποτελέσματα αλγορίθμου J48 για 2η ενέργεια. Σύγκριση Precision (χωρίς SMOTE) και Precision (με SMOTE K=2,5,7,10 γείτονες) για όλες τις τιμές CF.

Στο Διάγραμμα 14 παρουσιάζονται τα αποτελέσματα του Precision για τον αλγόριθμο J48 (C4.5) χωρίς SMOTE και με SMOTE με K=2,5,7,10 γείτονες. Με βάση το διάγραμμα

επιλέγεται η υψηλότερη τιμή Precision (max Precision= 0,727), η οποία προκύπτει όταν ο αλγόριθμος J48 εφαρμοστεί με Confidence Factor (CF)= 0,05 χωρίς SMOTE.

Επιπλέον με βάση και πάλι το παραπάνω διάγραμμα αυτό που παρατηρεί κανείς είναι ότι, όσο οι τιμές του CF μειώνονται, αντίθετα αυξάνονται οι αντίστοιχες τιμές της μετρικής Precision σε όλες τις περιπτώσεις, δηλαδή και με SMOTE και χωρίς SMOTE. Το CF όπως προαναφέρθηκε στη θεωρία ελέγχει τον βαθμό κλαδέματος του δέντρου. Επομένως μικρότερη τιμή CF σημαίνει μεγαλύτερο κλάδεμα, ενώ μεγαλύτερη τιμή CF σημαίνει μικρότερο κλάδεμα. Στη συγκεκριμένη λοιπόν περίπτωση κατά την εφαρμογή με SMOTE αλλά και χωρίς SMOTE, προκύπτει το συμπέρασμα, ότι το περισσότερο κλάδεμα βελτιώνει τα αποτελέσματα, δηλαδή την τιμή της μετρικής Recall.

Αυτό που αξίζει επίσης να σημειωθεί, είναι ότι στη συγκεκριμένη περίπτωση η εφαρμογή της τεχνικής SMOTE στο σύνολο δεδομένων, πριν την κατηγοριοποίηση με τον αλγόριθμο J48, μειώνει αισθητά την μετρική Precision σε όλες τις τιμές της παραμέτρου CF. Για το λόγο αυτό τα αποτελέσματα με την εφαρμογή της τεχνικής SMOTE δεν θα παρουσιαστούν αναλυτικά.



Διάγραμμα 15: Αποτελέσματα αλγορίθμου J48 για 2η ενέργεια. Σύγκριση Precision-Recall για όλες τις τιμές CF χωρίς SMOTE.

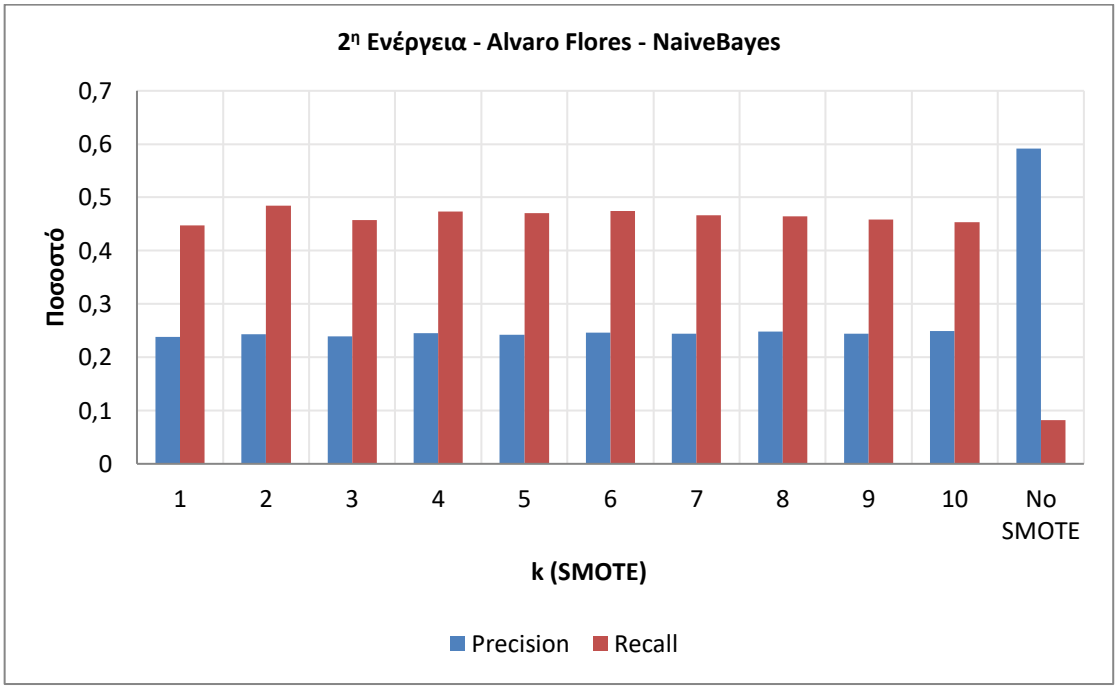
Αφού επιλέχθηκε η υψηλότερη τιμή Precision (max Precision) με βάση το Διάγραμμα 14, στη συνέχεια θα πρέπει να ελεγχθεί και η αντίστοιχη τιμή του Recall, ώστε να εξασφαλιστεί, ότι αυτή δεν είναι πολύ χαμηλή. Στο Διάγραμμα 15 λοιπόν παρατηρείται ότι στην τιμή CF=0,05

με μέγιστη τιμή Precision=0,727 η αντίστοιχη τιμή της μετρικής Recall = 0,077 είναι πολύ χαμηλή. Επομένως θεωρείται μη αποδεκτή και επιλέγεται κάποια άλλη. Επίσης παρατηρείται ότι και για τις τιμές CF= 0,1 έως 0,25 οι τιμές της μετρικής Recall είναι εξίσου πολύ χαμηλές, συνεπώς κι αυτές δεν επιλέγονται.

Στην τιμή CF=0,3 παρουσιάζεται αξιόλογη αύξηση του Recall=0,149 το οποίο σχεδόν διπλασιάζεται σε σχέση με το Recall στις προηγούμενες τιμές CF (0,05 έως 0,25). Κατά συνέπεια μπορεί να θεωρηθεί αποδεκτό, παρόλο που μειώνεται το αντίστοιχο Precision (από CF=0,05 σε CF=0,3 το precision μειώνεται κατά 0,282 ή 28%).

Στις επόμενες τιμές CF= 0,35 έως 0,55 το Recall παρουσιάζει πολύ μικρή αύξηση. Αντίθετα το Precision στις αντίστοιχες τιμές συνεχίζει να μειώνεται πράγμα που δεν είναι επιθυμητό. Επομένως η τιμή CF=0,3 με Recall=0,149 θα μπορούσε να θεωρηθεί μία σχετικά καλή επιλογή σε σύγκριση με τις υπόλοιπες τιμές. Για το λόγο αυτό επιλέγεται ως βέλτιστη.

Αλγόριθμος Naïve Bayes



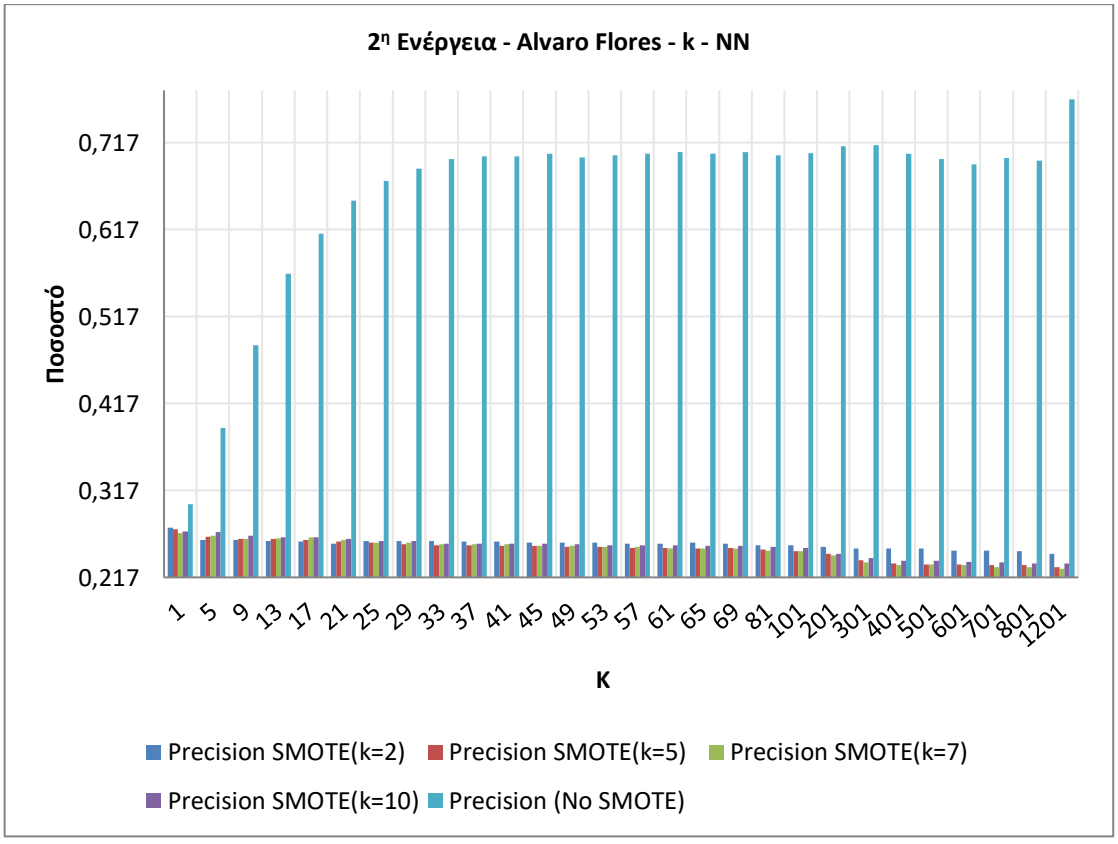
Διάγραμμα 16: Αποτελέσματα αλγορίθμου Naïve Bayes για 2η ενέργεια. Σύγκριση Precision-Recall χωρίς SMOTE και με SMOTE (K=1 έως 10 γείτονες).

Στο Διάγραμμα 16 παρουσιάζονται τα αποτελέσματα των μετρικών Recall-Precision για τον αλγόριθμο Naïve Bayes χωρίς SMOTE και με SMOTE με K= 1 έως 10 γείτονες. Με βάση το παραπάνω διάγραμμα επιλέγεται η υψηλότερη τιμή Precision (max Precision= 0,592), η οποία προκύπτει, όταν ο αλγόριθμος Naïve Bayes εφαρμοστεί χωρίς SMOTE. Η αντίστοιχη τιμή της μετρικής Recall = 0,082 είναι πολύ χαμηλή. Κατά συνέπεια θεωρείται μη αποδεκτή και επιλέγεται κάποια άλλη. Όταν ο αλγόριθμος Naïve Bayes εφαρμοστεί με την τεχνική

SMOTE η μετρική Recall παρουσιάζει μεγάλη αύξηση σε όλες τις τιμές k του SMOTE. Επομένως όλες οι αντίστοιχες τιμές του Recall είναι αποδεκτές. Κατά συνέπεια ως βέλτιστη τιμή της παραμέτρου k του SMOTE επιλέγεται η τιμή k=10 που παρουσιάζει το μεγαλύτερο precision σε σχέση με τις υπόλοιπες. Δηλαδή SMOTE με k=10 γείτονες δίνοντας Precision= 0,249 και Recall=0,453.

Από τα παραπάνω προκύπτει το συμπέρασμα ότι η εφαρμογή της τεχνικής SMOTE στο σύνολο δεδομένων, πριν την κατηγοριοποίηση με τον αλγόριθμο Naïve Bayes, μειώνει αισθητά την μετρική του Precision σε όλες τις τιμές k του SMOTE.

Αλγόριθμος k-NN

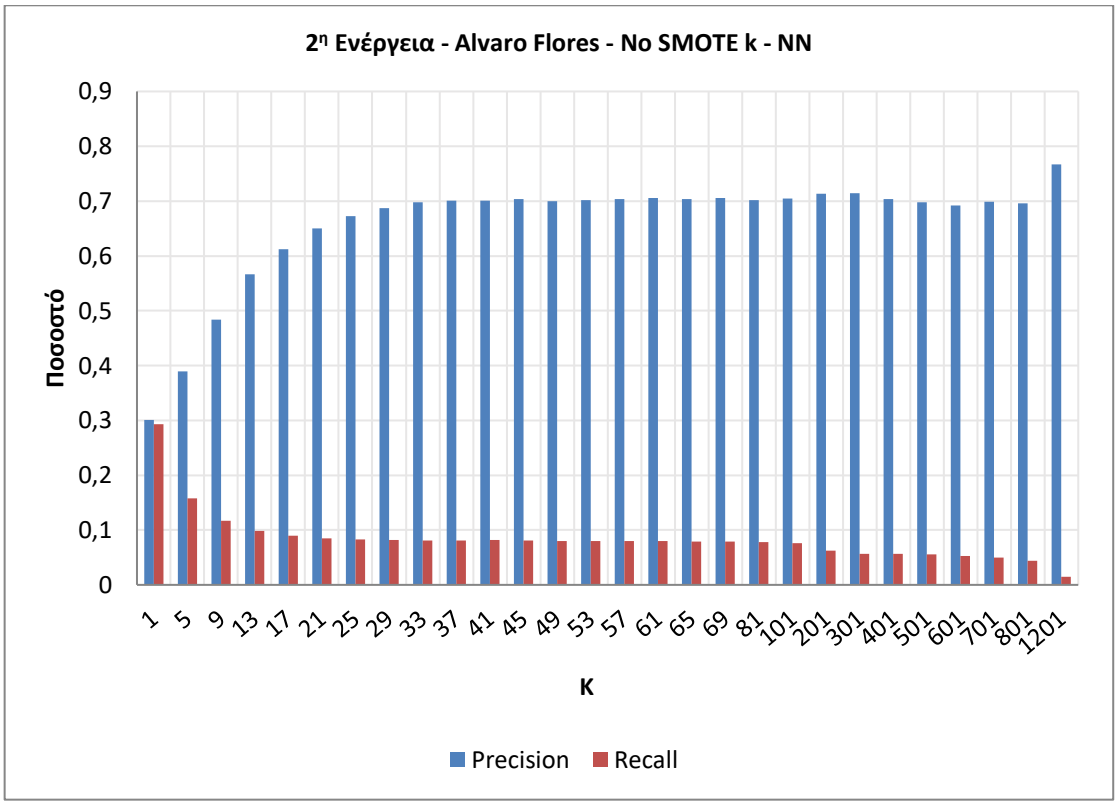


Διάγραμμα 17: Αποτελέσματα αλγορίθμου k-NN για 2η ενέργεια. Σύγκριση Precision (χωρίς SMOTE) και Precision (με SMOTE K=2,5,7,10 γείτονες) για όλες τις τιμές του k (k-NN).

Στο Διάγραμμα 17 παρουσιάζονται τα αποτελέσματα του Precision για τον αλγόριθμο k-NN χωρίς SMOTE και με SMOTE με K=2,5,7,10 γείτονες. Σύμφωνα με το διάγραμμα επιλέγεται η υψηλότερη τιμή Precision (max Precision= 0,767), η οποία προκύπτει όταν ο αλγόριθμος k-NN εφαρμοστεί με k=1201 γείτονες χωρίς SMOTE.

Αυτό που αξίζει να σημειωθεί, είναι ότι στη συγκεκριμένη περίπτωση η εφαρμογή της τεχνικής SMOTE στο σύνολο δεδομένων, πριν την κατηγοριοποίηση με τον αλγόριθμο k-NN, μειώνει αισθητά την μετρική Precision σε όλες τις τιμές της παραμέτρου k του

αλγορίθμου k-NN. Για το λόγο αυτό τα αποτελέσματα με την εφαρμογή της τεχνικής SMOTE δεν θα παρουσιαστούν αναλυτικά.



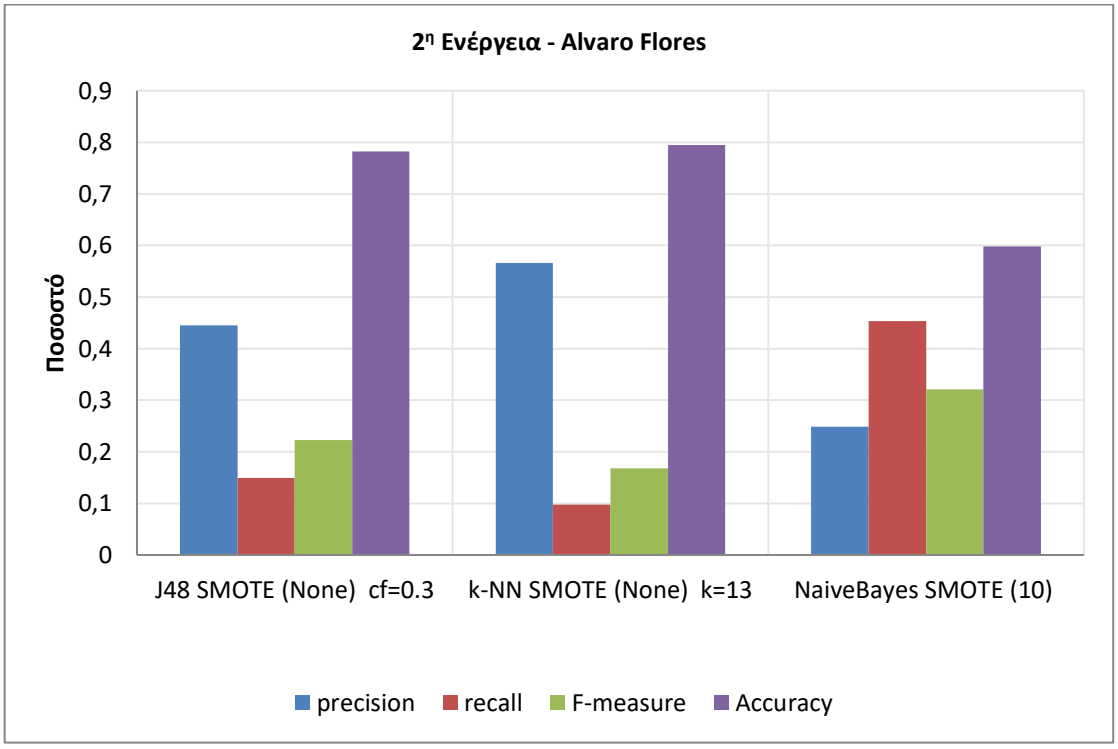
Διάγραμμα 18: Αποτελέσματα αλγορίθμου k-NN για 2η ενέργεια. Σύγκριση Precision-Recall για όλες τις τιμές του k (k-NN) χωρίς SMOTE

Αφού επιλέχθηκε η υψηλότερη τιμή Precision (max Precision) με βάση το Διάγραμμα 17, στη συνέχεια θα πρέπει να ελεγχθεί και η αντίστοιχη τιμή του Recall, ώστε να εξασφαλιστεί, ότι αυτή δεν είναι πολύ χαμηλή. Στο Διάγραμμα 18 λοιπόν παρατηρείται ότι στην τιμή k=1201 γείτονες του k-NN με μέγιστη τιμή Precision = 0,767, η αντίστοιχη τιμή της μετρικής Recall = 0,015 είναι πολύ χαμηλή. Κατά συνέπεια θεωρείται μη αποδεκτή και επιλέγεται κάποια άλλη.

Στο διάγραμμα φαίνεται ότι για τις τιμές του k από 801 έως και 33 γείτονες η τιμή της μετρικής Recall παρουσιάζει μικρή αύξηση μεν, αλλά εξακολουθεί να είναι πολύ χαμηλή. Για τις ίδιες τιμές k η τιμή της μετρικής Precision δεν μεταβάλλεται πολύ. Δηλαδή παραμένει κατά μέσο όρο περίπου ίση με 0,7. Στη συνέχεια παρατηρείται ότι στις τιμές του k από 29 έως και 13 γείτονες το Recall συνεχίζει να παρουσιάζει μικρή αύξηση αλλά πλέον το Precision αρχίζει να μειώνεται. Τέλος, για τις τιμές του k από 9 έως και 1 γείτονες το Recall συνεχίζει να αυξάνεται, αλλά το precision παρουσιάζει πλέον απότομη πτώση, κάτι που δεν είναι επιθυμητό. Συνεπώς η τιμή k=13 μπορεί να θεωρηθεί ως σχετικά καλή καθώς στις μικρότερες τιμές k (9 έως 1) η υπερβολική μείωση του precision είναι μη αποδεκτή για την 2η ενέργεια (Αντικατάσταση των κενών).

Επομένως οι βέλτιστες τιμές παραμέτρων για τον αλγόριθμο k-NN είναι k=13 γείτονες χωρίς SMOTE με Recall= 0,098 και αντίστοιχο Precision= 0,566.

8.2.1.2 Τελικά αποτελέσματα



Διάγραμμα 19: Τελικά αποτελέσματα αλγορίθμων k-NN, J48, Naïve Bayes για 2η ενέργεια. Σύγκριση μετρικών Precision (max), Recall, F-measure, Accuracy.

Στο Διάγραμμα 19 παρουσιάζονται τα αποτελέσματα των αλγορίθμων k-NN, J48 και Naïve Bayes με την βέλτιστη τιμή για την παράμετρο κάθε αλγορίθμου (k του k-NN, CF του J48 και k του SMOTE). Η βέλτιστη τιμή για κάθε παράμετρο προέκυψε από τη διαδικασία Εύρεσης βέλτιστων παραμέτρων (παράγραφος 8.2.1.1) και επιλέχθηκε με βάση την κατά το δυνατόν καλύτερη τιμή στην μετρική Precision και αντίστοιχα μια αξιόλογη τιμή στην μετρική Recall. Επιπλέον στο παραπάνω διάγραμμα παρουσιάζονται και οι αντίστοιχες τιμές των μετρικών Recall, F-measure και Accuracy.

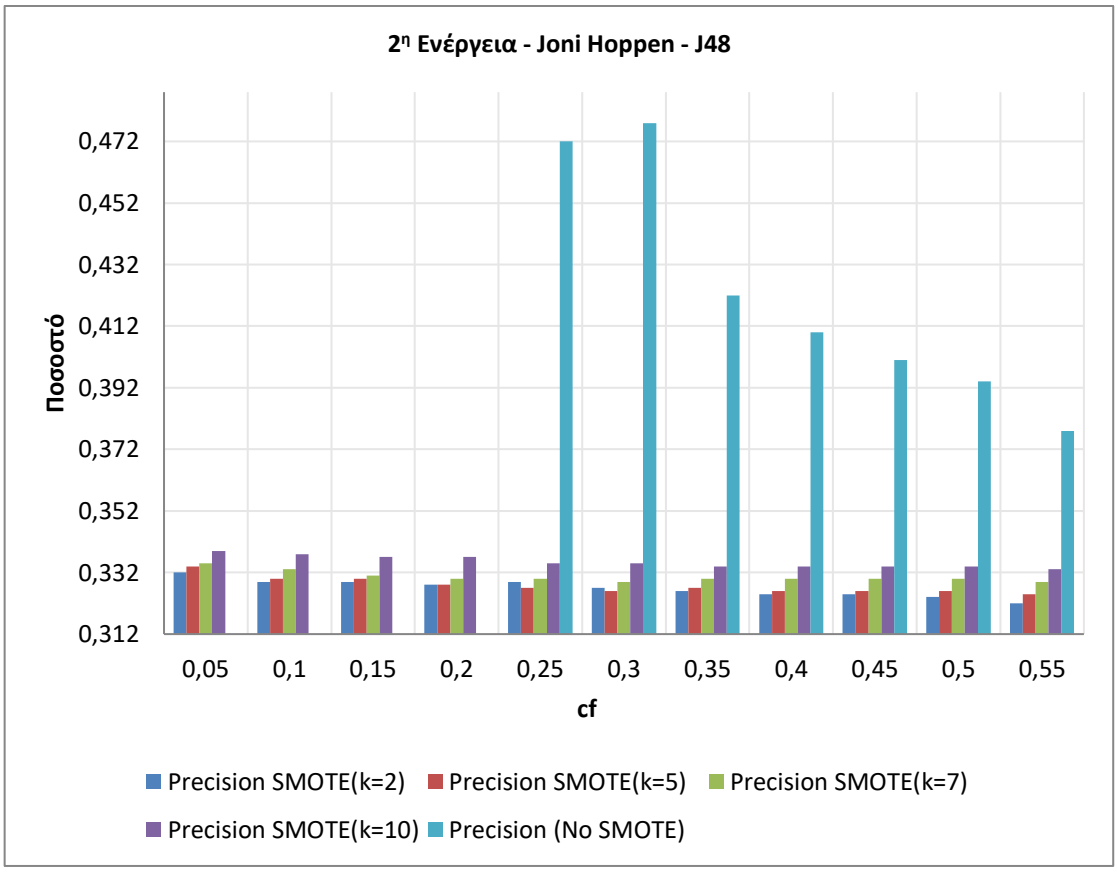
Η βέλτιστη τιμή για την μετρική Precision=0,566 παρουσιάζεται στον αλγόριθμο k-NN με τιμή παραμέτρου k=13 γείτονες χωρίς SMOTE. Το δεύτερο καλύτερο αποτέλεσμα το δίνει ο αλγόριθμος J48 και πάλι χωρίς την τεχνική SMOTE με βέλτιστη τιμή για την μετρική Precision=0,445. Από τα παραπάνω προκύπτει ότι η διαφορά μεταξύ των δύο αλγορίθμων ως προς τη μετρική Precision είναι 0,121 ή 12,1% η οποία θα μπορούσε να θεωρηθεί σχετικά αξιόλογη.

8.2.2 Σύνολο δεδομένων Joni Hoppen – Πειράματα με SMOTE και χωρίς SMOTE

8.2.2.1 Εύρεση βέλτιστων παραμέτρων (Parameter Tuning)

Η διαδικασία για την Εύρεση βέλτιστων παραμέτρων (παράγραφος 8.2.1.1) όσον αφορά το σύνολο δεδομένων Joni Hoppen είναι η ίδια που ακολουθήθηκε στο σύνολο δεδομένων Alvaro Flores και παρουσιάζεται αναλυτικά παρακάτω.

Αλγόριθμος J48 (C4.5)



Διάγραμμα 20: Αποτελέσματα αλγορίθμου J48 για 2η ενέργεια. Σύγκριση Precision (χωρίς SMOTE) και Precision (με SMOTE K=2,5,7,10 γείτονες) για όλες τις τιμές CF.

Στο Διάγραμμα 20 παρουσιάζονται τα αποτελέσματα του Precision για τον αλγόριθμο J48 (C4.5) χωρίς SMOTE και με SMOTE με K=2,5,7,10 γείτονες. Με βάση το διάγραμμα επιλέγεται η υψηλότερη τιμή Precision (max Precision= 0,478), η οποία προκύπτει όταν ο αλγόριθμος J48 εφαρμοστεί με Confidence Factor (CF)= 0,3 χωρίς SMOTE.

Επιπλέον με βάση και πάλι το παραπάνω διάγραμμα αυτό που παρατηρεί κανείς είναι ότι, στην περίπτωση που εφαρμόζεται τεχνική SMOTE, όσο οι τιμές του CF μειώνονται, αντίθετα αυξάνονται οι αντίστοιχες τιμές της μετρικής Precision. Στην περίπτωση όμως που δεν

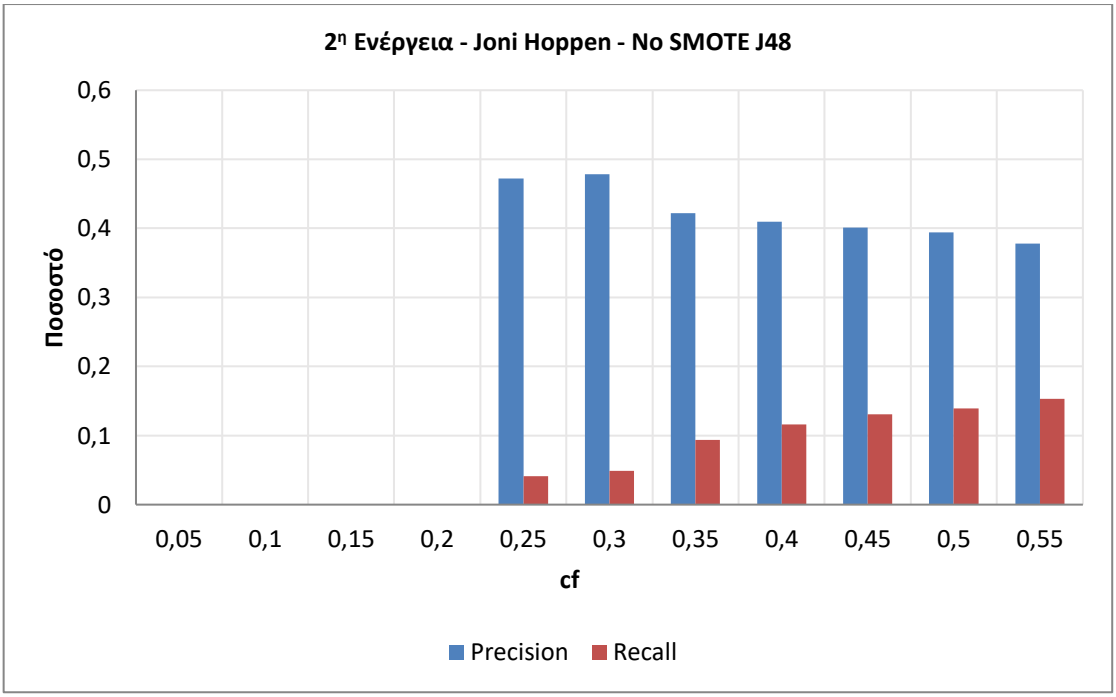
εφαρμόζεται τεχνική SMOTE, αυτό που παρατηρείται είναι ότι στις τιμές του CF= 0,55 έως 0,3 η τιμή της μετρικής Precision αυξάνεται. Αμέσως μετά όμως στην τιμή CF=0,25 η αντίστοιχη τιμή της μετρικής Precision έχει κάποια πτώση. Αντίθετα στις επόμενες τιμές, δηλαδή όταν CF=0,2 έως 0,05 (χωρίς SMOTE) παρατηρείται ότι δεν υπάρχουν καν αποτελέσματα για την μετρική Precision γιατί η τιμή της είναι ακαθόριστη. Αυτό προκύπτει από το γεγονός ότι έγιναν 0 συνολικές προβλέψεις (TP + FP=0) για την κλάση Yes, άρα δεν έγινε καμία σωστή πρόβλεψη (TP) ως προς τα πραγματικά δεδομένα (TP + FN=0). Επομένως και η τιμή της μετρικής Recall θα είναι 0, αν και δεν φαίνεται στο διάγραμμα.

Το CF όπως προαναφέρθηκε στη θεωρία ελέγχει τον βαθμό κλαδέματος του δέντρου. Επομένως μικρότερη τιμή CF σημαίνει μεγαλύτερο κλάδεμα, ενώ μεγαλύτερη τιμή CF σημαίνει μικρότερο κλάδεμα.

Άρα λοιπόν στην περίπτωση εφαρμογής του αλγορίθμου J48 χωρίς SMOTE προκύπτει το συμπέρασμα, ότι το κλάδεμα αρχικά βελτιώνει τα αποτελέσματα (τιμές 0,55 έως 0,3), ως προς την μετρική Precision. Αντίθετα με περισσότερο κλάδεμα (τιμές 0,25 έως 0,05) ο αλγόριθμος δίνει χειρότερα ή ακαθόριστα αποτελέσματα.

Στην περίπτωση εφαρμογής του αλγορίθμου J48 με SMOTE, προκύπτει ότι το περισσότερο κλάδεμα βελτιώνει τα αποτελέσματα σε μικρό βαθμό.

Τέλος αυτό που επίσης αξίζει να σημειωθεί, είναι ότι στη συγκεκριμένη περίπτωση η εφαρμογή της τεχνικής SMOTE στο σύνολο δεδομένων, πριν την κατηγοριοποίηση με τον αλγόριθμο J48, μειώνει αισθητά την μετρική Precision σε όλες τις τιμές της παραμέτρου CF. Για το λόγο αυτό τα αποτελέσματα με την εφαρμογή της τεχνικής SMOTE δεν θα παρουσιαστούν αναλυτικά.



Διάγραμμα 21: Αποτελέσματα αλγορίθμου J48 για 2η ενέργεια. Σύγκριση Precision-Recall για όλες τις τιμές CF χωρίς SMOTE.

Αφού επιλέχθηκε η υψηλότερη τιμή Precision (max Precision) με βάση το Διάγραμμα 20, στη συνέχεια θα πρέπει να ελεγχθεί και η αντίστοιχη τιμή του Recall, ώστε να εξασφαλιστεί, ότι αυτή δεν είναι πολύ χαμηλή. Στο Διάγραμμα 21 λοιπόν παρατηρείται ότι στην τιμή CF=0,3 με μέγιστη τιμή Precision=0,478 η αντίστοιχη τιμή της μετρικής Recall = 0,049 είναι πολύ χαμηλή. Επομένως θεωρείται μη αποδεκτή και επιλέγεται κάποια άλλη.

Στις τιμές CF=0,05 έως 0,2 δεν εμφανίζονται αποτελέσματα γιατί σε αυτές τις περιπτώσεις η τιμή της μετρικής Recall είναι μηδέν (0) και η τιμή της μετρικής Precision είναι ακαθόριστη. Αυτό συμβαίνει γιατί ο αλγόριθμος J48 δημιουργεί δέντρο με ένα μόνο φύλλο που ανήκει στην κλάση No. Δηλαδή προβλέπει ότι όλοι οι ασθενείς θα εμφανισθούν. Κατά συνέπεια το Recall είναι μηδέν γιατί δεν προέβλεψε σωστά κανέναν από τους ασθενείς που στην πραγματικότητα δεν εμφανίσθηκαν. Αντίστοιχα το Precision είναι ακαθόριστο διότι έγιναν 0 προβλέψεις για την κλάση Yes. Δηλαδή ο κατηγοριοποιητής δεν προέβλεψε κανέναν ασθενή ότι δεν θα εμφανισθεί στο ραντεβού. Σε αυτή λοιπόν την περίπτωση ο υπολογισμός του Precision θα οδηγούσε σε διαίρεση με το μηδέν, πράγμα που δεν γίνεται. (Recall=TP/TP+FN=0/0 + κάποια τιμή=0, Precision=TP/TP+FP=0/0+0= ακαθόριστο).

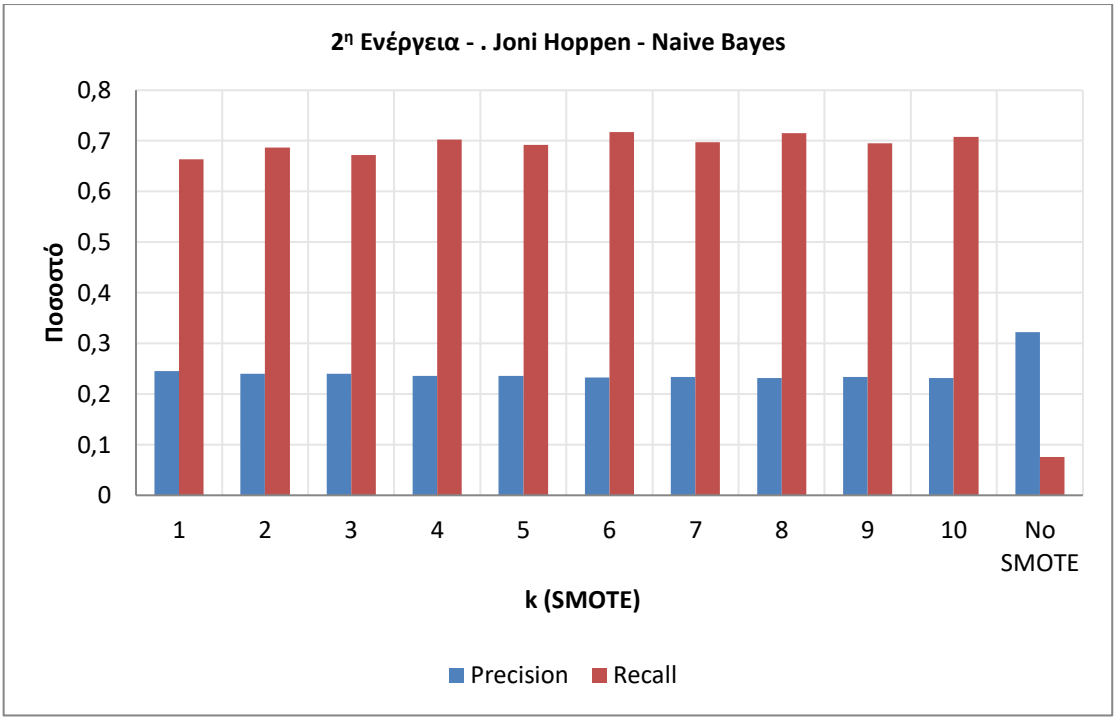
Από την άλλη πλευρά στην τιμή CF=0,25, σε σχέση με την τιμή CF=0,3 που δίνει την υψηλότερη τιμή στην μετρική Precision, παρατηρείται ταυτόχρονη πτώση και στην τιμή της μετρικής Recall=0,041 αλλά και στην τιμή της μετρικής Precision=0,472, επομένως και η τιμή CF=0,25 απορρίπτεται. Στη συνέχεια εξετάζονται οι υπόλοιπες τιμές CF σε σχέση με την τιμή CF=0,3 που δίνει την υψηλότερη τιμή για την μετρική Precision. Στην τιμή λοιπόν CF=0,35 παρουσιάζεται αύξηση στην τιμή της μετρικής Recall=0,094 της τάξης του 0,045 ή

4,5% αλλά και πάλι η συγκεκριμένη τιμή θεωρείται χαμηλή, ενώ ταυτόχρονά σημειώνεται πτώση και στην τιμή της μετρικής Precision=0,422 (Μέγιστο Precision=0,478) της τάξης του 0,056 ή 5,6%. Επομένως και η τιμή CF=0,35 δεν θεωρείται κατάλληλη. Μέχρι τώρα λοιπόν με βάση το παραπάνω διάγραμμα παρατηρείται ότι για να επιτευχθεί μία αξιόλογη τιμή στην μετρική Recall, θα πρέπει αναγκαστικά, αν και δεν είναι επιθυμητό, να επιλεγεί μία χαμηλότερη τιμή και για την μετρική Precision. Με βάση αυτό το συλλογισμό εξετάζονται οι υπόλοιπες τιμές CF του διαγράμματος. Οι τιμές CF=0,50 και CF=0,55 παρουσιάζουν την υψηλότερη τιμή στην μετρική Recall σε σύγκριση με τις προηγούμενες τιμές CF (0,25 έως 0,35) αλλά και την χαμηλότερη τιμή στην μετρική Precision, οπότε δεν θεωρούνται αποδεκτές και αποκλείονται. Επομένως το ενδιαφέρον εστιάζεται στην περιοχή τιμών CF=0,40 και CF=0,45, όπου και στις 2 τιμές CF παρατηρείται ικανοποιητική αύξηση της τιμής της μετρικής Recall με την κατά το δυνατόν μικρότερη πτώση στην τιμή της μετρικής Precision σε σύγκριση με τις τιμές CF=0,3 έως 0,35. Σε γενικότερο πλαίσιο λοιπόν και η τιμή CF=0,40 αλλά και η τιμή CF=0,45 θεωρούνται αποδεκτές. Ειδικότερα όμως θα μπορούσε να ισχυριστεί κανείς ότι η τιμή CF=0,45 πιθανόν είναι καταλληλότερη. Αυτό ισχύει διότι στην τιμή CF=0,45 η τιμή της μετρικής Precision μειώνεται ελάχιστα, δηλαδή κατά 0,009 ή 0,9% σε σχέση με την τιμή της μετρικής Precision στην τιμή CF=0,40, ενώ ταυτόχρονα παρουσιάζει μεγαλύτερη αύξηση στην τιμή της μετρικής Recall της τάξης του 0,015 ή 1,5%. Με βάση τα παραπάνω ως βέλτιστη τιμή παραμέτρου για τον αλγόριθμο J48 θεωρείται η τιμή CF=0,45 με Recall= 0,131 και Precision= 0,401.

Αλγόριθμος Naïve Bayes

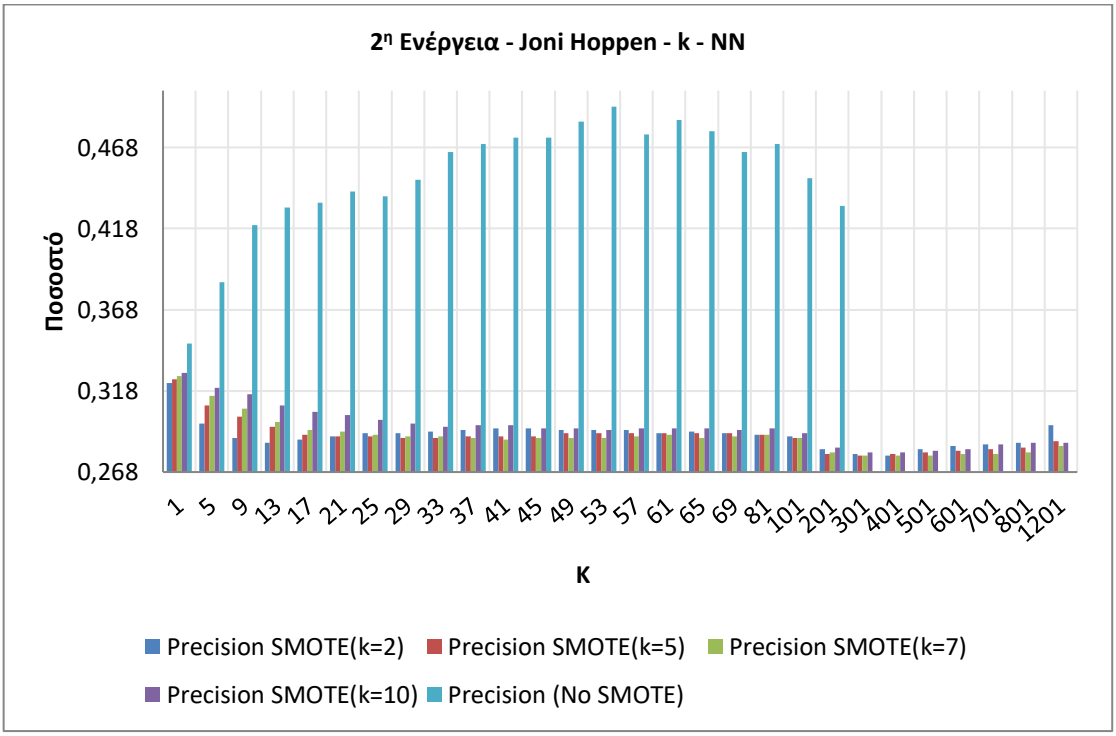
Στο Διάγραμμα 22 παρουσιάζονται τα αποτελέσματα των μετρικών Recall-Precision για τον αλγόριθμο Naïve Bayes χωρίς SMOTE και με SMOTE με K= 1 έως 10 γείτονες. Με βάση το παραπάνω διάγραμμα επιλέγεται η υψηλότερη τιμή Precision (max Precision= 0,322), η οποία προκύπτει όταν εφαρμοστεί ο αλγόριθμος Naïve Bayes χωρίς SMOTE. Η αντίστοιχη τιμή της μετρικής Recall = 0,076 είναι πολύ χαμηλή. Κατά συνέπεια θεωρείται μη αποδεκτή και επιλέγεται κάποια άλλη τιμή. Όταν ο αλγόριθμος Naïve Bayes εφαρμοστεί με την τεχνική SMOTE η μετρική Recall παρουσιάζει μεγάλη αύξηση σε όλες τις τιμές k του SMOTE. Επομένως όλες οι αντίστοιχες τιμές του Recall είναι αποδεκτές. Κατά συνέπεια ως βέλτιστη τιμή της παραμέτρου k του SMOTE επιλέγεται η τιμή k=1 γείτονας που παρουσιάζει το μεγαλύτερο precision σε σχέση με τις υπόλοιπες τιμές. Δηλαδή SMOTE με k=1 γείτονα δίνοντας Precision= 0,245 και Recall=0,664.

Από τα παραπάνω προκύπτει το συμπέρασμα ότι η εφαρμογή της τεχνικής SMOTE στο σύνολο δεδομένων, πριν την κατηγοριοποίηση με τον αλγόριθμο Naïve Bayes, μειώνει αισθητά την μετρική Precision σε όλες τις τιμές k του SMOTE.



Διάγραμμα 22: Αποτελέσματα αλγορίθμου Naïve Bayes για 2η ενέργεια. Σύγκριση Precision-Recall χωρίς SMOTE και με SMOTE (K=1 έως 10 γείτονες).

Αλγόριθμος k-NN

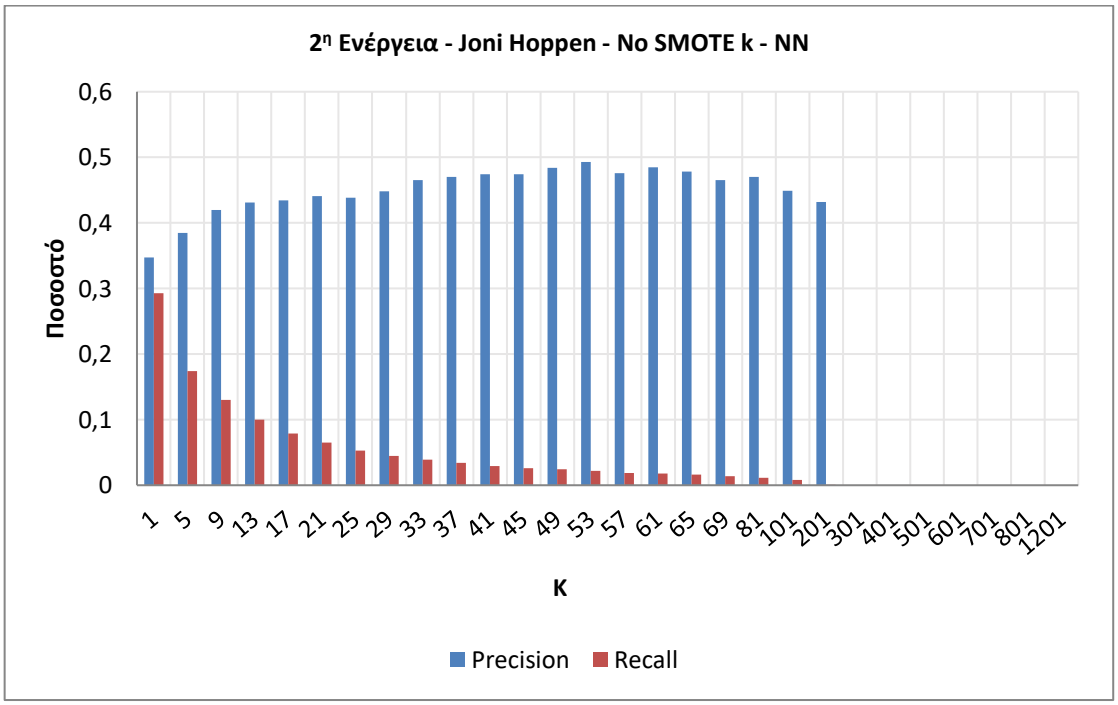


Διάγραμμα 23: Αποτελέσματα αλγορίθμου k-NN για 2η ενέργεια. Σύγκριση Precision (χωρίς SMOTE) και Precision (με SMOTE K=2,5,7,10 γείτονες) για όλες τις τιμές του k (k-NN).

Στο Διάγραμμα 23 παρουσιάζονται τα αποτελέσματα του Precision για τον αλγόριθμο k-NN χωρίς SMOTE και με SMOTE με K=2,5,7,10 γείτονες. Σύμφωνα με το διάγραμμα επιλέγεται

η υψηλότερη τιμή Precision (max Precision= 0,493), η οποία προκύπτει όταν ο αλγόριθμος k-NN εφαρμοστεί με k=53 γείτονες χωρίς SMOTE. Επίσης παρατηρείται ότι στις τιμές k (k-NN)= 301 έως 1201 (χωρίς SMOTE) δεν υπάρχουν αποτελέσματα για την μετρική Precision γιατί η τιμή της είναι ακαθόριστη.

Αυτό που αξίζει να σημειωθεί, είναι ότι στη συγκεκριμένη περίπτωση η εφαρμογή της τεχνικής SMOTE στο σύνολο δεδομένων, πριν την κατηγοριοποίηση με τον αλγόριθμο k-NN, μειώνει αισθητά την μετρική του Precision σε όλες τις τιμές της παραμέτρου k του αλγορίθμου k-NN. Για το λόγο αυτό τα αποτελέσματα με την εφαρμογή της τεχνικής SMOTE δεν θα παρουσιαστούν αναλυτικά.



Διάγραμμα 24: Αποτελέσματα αλγορίθμου k-NN για 2η ενέργεια. Σύγκριση Precision-Recall για όλες τις τιμές του k (k-NN) χωρίς SMOTE.

Αφού επιλέχθηκε η υψηλότερη τιμή Precision (max Precision) με βάση το Διάγραμμα 23, στη συνέχεια θα πρέπει να ελεγχθεί και η αντίστοιχη τιμή του Recall, ώστε να εξασφαλιστεί, ότι αυτή δεν είναι πολύ χαμηλή. Στο Διάγραμμα 24 λοιπόν παρατηρείται ότι στην τιμή k=53 γείτονες του k-NN με μέγιστη τιμή Precision = 0,493, η αντίστοιχη τιμή της μετρικής Recall = 0,022 είναι πολύ χαμηλή. Κατά συνέπεια θεωρείται μη αποδεκτή και επιλέγεται κάποια άλλη. Στο διάγραμμα φαίνεται ότι για τις τιμές του k από 49 έως και 33 γείτονες η τιμή της μετρικής Recall παρουσιάζει μικρή αύξηση μεν, αλλά εξακολουθεί να είναι πολύ χαμηλή. Για τις ίδιες τιμές k η τιμή της μετρικής Precision δεν μεταβάλλεται πολύ. Δηλαδή παραμένει κατά μέσο όρο περίπου ίση με 0,47. Στη συνέχεια παρατηρείται ότι στις τιμές του k από 29 έως και 25 γείτονες το Recall συνεχίζει να παρουσιάζει μικρή αύξηση αλλά πλέον το Precision αρχίζει να μειώνεται.

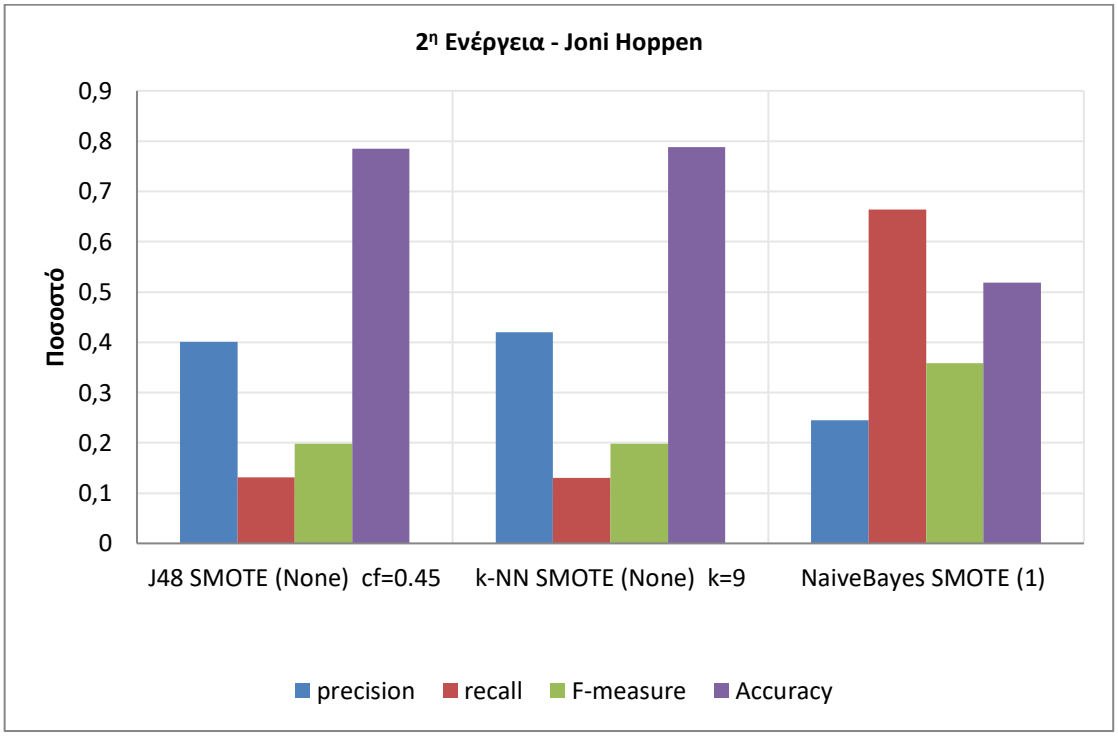
Έπειτά όμως, δηλαδή στις τιμές του k από 21 έως και 9 γείτονες η τιμή της μετρικής Precision παρουσιάζει μικρές διακυμάνσεις και είναι κατά μέσο όρο περίπου ίση με 0,431. Αντίθετα η τιμή της μετρικής Recall σημειώνει μεγαλύτερη άνοδο σε σχέση με τις προηγούμενες τιμές k (k-NN).

Επιπλέον, για τις τιμές του k από 5 έως και 1 γείτονες το Recall συνεχίζει να αυξάνεται και μάλιστα αισθητά, αλλά το precision παρουσιάζει πλέον απότομη πτώση, κάτι που δεν είναι επιθυμητό. Συνεπώς η τιμή k=9 μπορεί να θεωρηθεί ως σχετικά καλή καθώς στις μικρότερες τιμές k (5 έως 1) η υπερβολική μείωση του precision είναι μη αποδεκτή για την $2^{η}$ ενέργεια (Αντικατάσταση των κενών). Τέλος σημειώνεται ότι για τις τιμές k (k-NN)= 301 έως 1201 γείτονες οι τιμές της μετρικής Recall είναι μηδενικές και οι τιμές της μετρικής Precision ακαθόριστες. Σε αυτή λοιπόν την περίπτωση, ισχύει ότι περιγράφηκε και στην περίπτωση του J48. Δηλαδή ο κατηγοριοποιητής προβλέπει, ότι όλοι οι ασθενείς θα εμφανισθούν στο ραντεβού, άρα ότι όλα τα πρότυπα ανήκουν στην κλάση No. Κατά συνέπεια το Recall είναι μηδέν γιατί δεν προέβλεψε σωστά κανέναν από τους ασθενείς που στην πραγματικότητα δεν εμφανίσθηκαν.

Αντίστοιχα το Precision είναι ακαθόριστο διότι έγιναν 0 προβλέψεις για την κλάση Yes. Δηλαδή ο κατηγοριοποιητής δεν προέβλεψε κανέναν ασθενή, ότι δεν θα εμφανισθεί στο ραντεβού. Σε αυτή λοιπόν την περίπτωση ο υπολογισμός του Precision θα οδηγούσε σε διαίρεση με το μηδέν, πράγμα που δεν γίνεται. (Recall=TP/TP+FN=0/0 + κάποια τιμή=0, Precision=TP/TP+FP=0/0+0= ακαθόριστο). Συνοψίζοντας με βάση όσα ειπώθηκαν παραπάνω, οι βέλτιστες τιμές παραμέτρων για τον αλγόριθμο k-NN είναι k=9 γείτονες χωρίς SMOTE με Recall= 0,13 και αντίστοιχο Precision= 0,42.

Τέλος υποθετικά μιλώντας, θα μπορούσε να ισχυριστεί κανείς, ότι στην περίπτωση του αλγορίθμου k-NN οι μηδενικές τιμές μπορεί να οφείλονται στο γεγονός ότι το σύνολο δεδομένων Joni Hoppen είναι μη ισορροπημένο και τα περισσότερα πρότυπα ανήκουν στην κλάση No. Επομένως στις μεγάλες τιμές k (k-NN) το πιθανότερο είναι η πλειοψηφία των γειτόνων να ανήκει πάντα στην κλάση No.

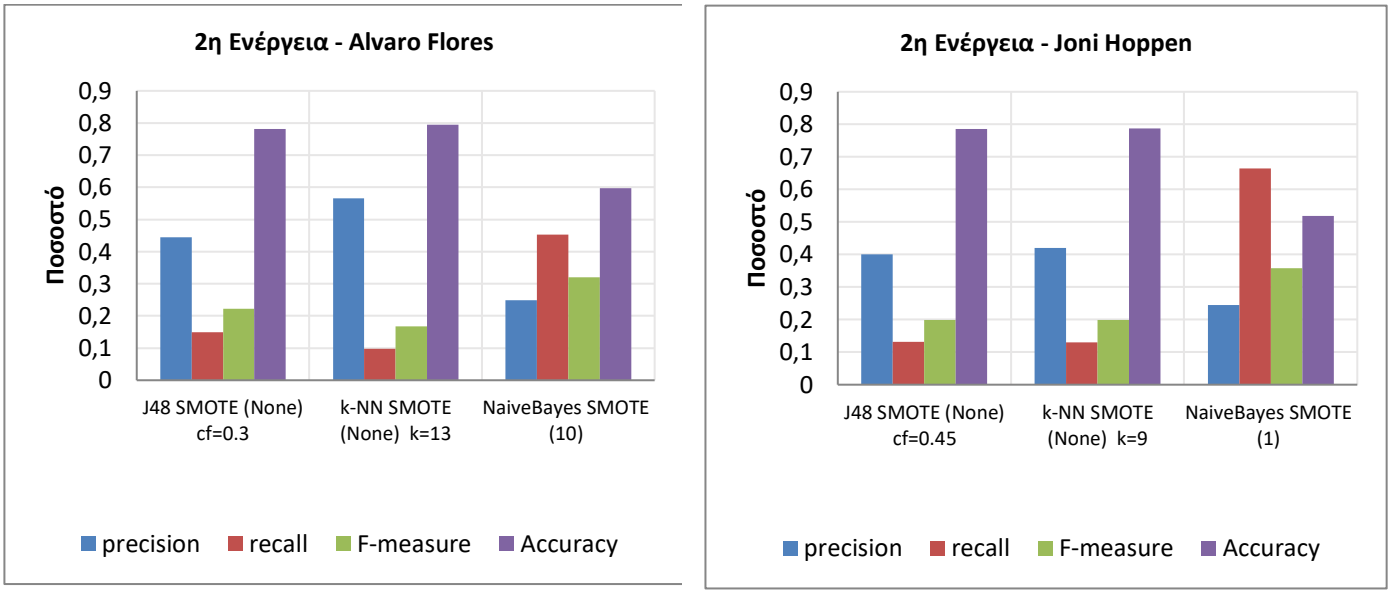
8.2.2.2 *Τελικά αποτελέσματα*



Διάγραμμα 25: Τελικά αποτελέσματα αλγορίθμων k-NN, J48, Naïve Bayes για 2η ενέργεια. Σύγκριση μετρικών Precision (max), Recall, F-measure, Accuracy.

Στο Διάγραμμα 25 παρουσιάζονται τα αποτελέσματα των αλγορίθμων k-NN, J48 και Naïve Bayes με την βέλτιστη τιμή για την παράμετρο κάθε αλγορίθμου (k του k-NN, CF του J48 και k του SMOTE). Η βέλτιστη τιμή για κάθε παράμετρο προέκυψε από τη διαδικασία Εύρεσης βέλτιστων παραμέτρων (παράγραφος 8.2.1.1**)** και επιλέχθηκε με βάση την κατά το δυνατόν καλύτερη τιμή στην μετρική Precision (Precision) και αντίστοιχα μια αξιόλογη τιμή στην μετρική Recall. Επιπλέον στο παραπάνω διάγραμμα παρουσιάζονται και οι αντίστοιχες τιμές των μετρικών Recall, F-measure και Accuracy. Η βέλτιστη τιμή για την μετρική Precision=0,42 παρουσιάζεται στον αλγόριθμο k-NN όταν εφαρμόζεται σε αυτόν τεχνική SMOTE με k=9 γείτονες. Το δεύτερο καλύτερο αποτέλεσμα το δίνει ο αλγόριθμος J48 όταν εφαρμόζεται σε αυτόν τιμή παραμέτρου CF=0,45 χωρίς SMOTE, δίνοντας Precision=0,401. Από τα παραπάνω προκύπτει ότι η διαφορά μεταξύ των δύο αλγορίθμων ως προς τη μετρική Precision είναι 0,019 ή 1,9%. Δηλαδή σχετικά μικρή.

8.2.3 Συμπεράσματα για την 2η ενέργεια



Διάγραμμα 26: 2η Ενέργεια (Αντικατάσταση κενών): Συγκεντρωτικά αποτελέσματα για τα σύνολα δεδομένων Alvaro Flores και Joni Hoppen

Με βάση τα αποτελέσματα στο Διάγραμμα 26, εξάγονται τα παρακάτω συμπεράσματα. Αρχικά υπενθυμίζεται ότι το υψηλό Precision είναι το ζητούμενο για την 2η ενέργεια (4.3.2). Με βάση αυτό ως δεδομένο παρατηρείται ότι ο αλγόριθμος k-NN είναι αυτός που παρουσιάζει το υψηλότερο Precision και στα δύο σύνολα δεδομένων. Ακολουθούν αντίστοιχα οι αλγόριθμοι J48 και Naïve Bayes. Αναλυτικότερα στο σύνολο δεδομένων Alvaro Flores, ο αλγόριθμος k-NN παρουσιάζει το υψηλότερο Precision=0,566, όταν εφαρμοστεί με k(k-NN)=13 γείτονες χωρίς SMOTE. Αντίστοιχα στο σύνολο δεδομένων Joni Hoppen και πάλι ο αλγόριθμος k-NN σημειώνει το υψηλότερο Precision=0,42, αλλά αυτή τη φορά όταν εφαρμοστεί με k(k-NN)=9 γείτονες χωρίς SMOTE. Σύμφωνα με τα παραπάνω προκύπτει το συμπέρασμα ότι ο αλγόριθμος k-NN αποδίδει κατά 0,146 ή 14,6% καλύτερα στο σύνολο δεδομένων Alvaro Flores, απ' ότι στο σύνολο δεδομένων Joni Hoppen.

Το δεύτερο καλύτερο αποτέλεσμα και στα δύο σύνολα δεδομένων, όπως προαναφέρθηκε, παρατηρείται στον αλγόριθμο J48, ο οποίος παρουσιάζει Precision=0,445 στο σύνολο δεδομένων Alvaro Flores και Precision=0,401 στο σύνολο δεδομένων Joni Hoppen. Επομένως θα μπορούσε να ειπωθεί, ότι και ο αλγόριθμος J48 αποδίδει κατά 0,044 ή 4,4% καλύτερα στο σύνολο δεδομένων Alvaro Flores, απ' ότι στο σύνολο δεδομένων Joni Hoppen.

Το τρίτο καλύτερο αποτέλεσμα και στα δύο σύνολα δεδομένων καταγράφεται στον αλγόριθμο Naïve Bayes, ο οποίος παρουσιάζει Precision=0,249 στο σύνολο δεδομένων Alvaro Flores και Precision=0,245 στο σύνολο δεδομένων Joni Hoppen. Με βάση τα παραπάνω προκύπτει ότι και ο αλγόριθμος Naïve Bayes δείχνει να προσαρμόζεται κατά 0,004 ή 0,4% καλύτερα στο σύνολο δεδομένων Alvaro Flores, απ' ότι στο σύνολο δεδομένων Joni

Hoppen. Γενικότερα λοιπόν παρατηρείται ότι και οι 3 αλγόριθμοι παρουσιάζουν υψηλότερο Precision στο σύνολο δεδομένων Alvaro Flores απ' ότι στο σύνολο δεδομένων Joni Hoppen. Επομένως θα ήταν εύλογο να ισχυριστεί κανείς, ότι η 2η ενέργεια θα μπορούσε να εφαρμοστεί αποδοτικότερα στο σύνολο δεδομένων Alvaro Flores.

Σημειώνεται ότι αν και το ζητούμενο στην 2η ενέργεια είναι το υψηλότερο Precision, το Recall δεν πρέπει να είναι σε πολύ χαμηλά επίπεδα. Συνεπώς στο υποκεφάλαιο (8.2.3) παρουσιάζονται αποτελέσματα πειραμάτων, όπου η τιμή της μετρικής Precision είναι η υψηλότερη δυνατή, διατηρώντας παράλληλα την τιμή της μετρικής Recall σε αποδεκτό επίπεδο.

Στον παρακάτω πίνακα φαίνονται επίσης συγκεντρωτικά οι μέγιστες και βέλτιστες τιμές που προέκυψαν από τη διαδικασία του grid search για την 2η ενέργεια και για τους 3 αλγορίθμους κατηγοριοποίησης και για τα 2 σύνολα δεδομένων (Πίνακας 62).

Σύνολο Δεδομένων	**Αλγόριθμος**	**Παράμετρος**	**Τεχνική (SMOTE)**	**MAX Recall**	**MAX Precision**	**Optimum Recall**	**Optimum Precision**
Alvaro Flores	J48	Cf=0,05	Χωρίς SMOTE	0,077	0,727		
	J48	Cf=0,3	Χωρίς SMOTE			0,149	0,445
	Naïve Bayes	-	Χωρίς SMOTE	0,082	0,592		
	Naïve Bayes	-	SMOTE (K=10)			0,453	0,249
	k-NN	K=1201	Χωρίς SMOTE	0,015	0,767		
	k-NN	**K=13**	**Χωρίς SMOTE**			**0,098**	**0,566**
Joni Hoppen	J48	Cf=0,3	Χωρίς SMOTE	0,049	0,478		
	J48	Cf=0,45	Χωρίς SMOTE			0,131	0,401
	Naïve Bayes	-	Χωρίς SMOTE	0,076	0,322		
	Naïve Bayes	-	SMOTE (K=1)			0,664	0,245
	k-NN	K=53	Χωρίς SMOTE	0,022	0,493		
	k-NN	**K=9**	**Χωρίς SMOTE**			**0,13**	**0,42**

Πίνακας 62: Τελικά αποτελέσματα για 2η ενέργεια (Μέγιστες και βέλτιστες τιμές)

9

Συμπεράσματα και μελλοντική έρευνα

Στο κεφάλαιο αυτό ακολουθούν τα συμπεράσματα που προέκυψαν από την πειραματική μελέτη που πραγματοποιήθηκε και θα αναφερθούν ιδέες για μελλοντική έρευνα.

9.1 Συμπεράσματα

Στην παρούσα διπλωματική εργασία έγινε μία συγκριτική μελέτη αλγορίθμων κατηγοριοποίησης, οι οποίοι εφαρμόστηκαν σε σύνολα δεδομένων που αφορούν ιατρικά ραντεβού. Συγκεκριμένα χρησιμοποιήθηκαν οι αλγόριθμοι C4.5 (ή J48 στο WEKA), k-NN και Naïve Bayes.

Στο θεωρητικό μέρος της διπλωματικής εργασίας παρουσιάστηκαν οι αλγόριθμοι κατηγοριοποίησης, περιγράφηκε το πρόβλημα της μη εμφάνισης των ασθενών σε προγραμματισμένο ραντεβού και αναφέρθηκαν 2 πιθανές ενέργειες για την επίλυση του. Τέλος περιγράφηκαν τεχνικές υπερδειγματοληψίας για την αντιμετώπιση μη ισορροπημένων συνόλων δεδομένων.

Στο πρακτικό μέρος έγινε παρουσίαση των δύο δημόσια διαθέσιμων συνόλων δεδομένων που χρησιμοποιήθηκαν. Επειδή όμως τα 2 παραπάνω σύνολα δεδομένων είναι μη ισορροπημένα, εφαρμόστηκε ο αλγόριθμος υπερδειγματοληψίας SMOTE για την εξισορρόπηση τους. Τέλος εκτελέστηκαν πειράματα με τους παραπάνω 3 αλγορίθμους κατηγοριοποίησης και έγινε σύγκριση των αποτελεσμάτων τους.

Η σύγκριση των αποτελεσμάτων ανέδειξε τα ακόλουθα γενικά συμπεράσματα. Για την καλύτερη όμως κατανόηση των συμπερασμάτων υπενθυμίζεται ότι για την 1η ενέργεια (Υπενθύμιση του ραντεβού) το ζητούμενο είναι το υψηλό Recall. Τα πειράματα για την 1η ενέργεια εκτελέστηκαν στο σύνολο δεδομένων Alvaro Flores (Περίπτωση 1) και στο σύνολο δεδομένων Joni Hoppen (Περίπτωση 2). Αντίστοιχα για την 2η ενέργεια (αντικατάσταση κενών) το ζητούμενο είναι το υψηλό Precision και έγιναν πειράματα στο σύνολο δεδομένων Alvaro Flores (Περίπτωση 3) και στο σύνολο δεδομένων Joni Hoppen (Περίπτωση 4).

Στην 1η ενέργεια το υψηλότερο Recall στο σύνολο δεδομένων Alvaro Flores (Περίπτωση 1) το παρουσίασε ο αλγόριθμος k-NN με SMOTE ενώ στο σύνολο δεδομένων Joni Hoppen (Περίπτωση 2) ο αλγόριθμος Naïve Bayes με SMOTE. Στη 2η ενέργεια το υψηλότερο Precision και στα 2 σύνολα δεδομένων (Περίπτωση 3 και 4) το παρουσίασε ο αλγόριθμος k-NN αλλά χωρίς SMOTE.

Αν δει λοιπόν κανείς συνολικά τις 2 ενέργειες, αυτό που παρατηρεί είναι ότι γενικά στις 3 από τις 4 περιπτώσεις (Περίπτωση 1, 3, 4) ο αλγόριθμος k-NN είναι αυτός που δίνει τα καλύτερα αποτελέσματα. Ακόμα και στην Περίπτωση 2 στην οποία είναι καλύτερος ο αλγόριθμος Naïve Bayes, παρατηρείται ότι ο k-NN έχει επίσης σχετικά καλή απόδοση αφού παρουσιάζει τα δεύτερα καλύτερα αποτελέσματα.

Το συμπέρασμα λοιπόν που απορρέει είναι ότι και στις 2 ενέργειες, ο αλγόριθμος k-NN παρουσιάζει γενικά καλά αποτελέσματα σε σύγκριση με τον αλγόριθμο J48 και Naïve Bayes.

Όσον αφορά την επίδραση της τεχνικής SMOTE στα αποτελέσματα, παρατηρείται ότι στην 1η ενέργεια τα καλύτερα αποτελέσματα σημειώνονται όταν εφαρμόζεται η τεχνική SMOTE. Αντίθετα στην 2η ενέργεια η καλύτερη επίδοση προκύπτει χωρίς εφαρμογή της τεχνικής SMOTE.

Με βάση λοιπόν όσα ειπώθηκαν παραπάνω, εξάγεται το συμπέρασμα, ότι η εφαρμογή της τεχνικής SMOTE βελτιώνει τα αποτελέσματα στην 1η ενέργεια. Αντίθετα στην 2η ενέργεια η προσθήκη της τεχνικής SMOTE δίνει χαμηλότερα αποτελέσματα συγκριτικά με τα αποτελέσματα που προκύπτουν από εφαρμογή των αλγορίθμων χωρίς SMOTE.

Όσον αφορά τις βέλτιστες τιμές των μετρικών Precision και Recall, στην 1η ενέργεια στο σύνολο δεδομένων Alvaro Flores ο αλγόριθμος k-NN με SMOTE παρουσιάζει Recall=0,553 και Precision= 0,233. Επίσης στην 1η ενέργεια αλλά στο σύνολο δεδομένων Joni Hoppen ο Naïve Bayes με SMOTE σημειώνει Recall=0,717 και Precision= 0,233. Στη 2η ενέργεια ο αλγόριθμος k-NN χωρίς SMOTE στο σύνολο δεδομένων Alvaro Flores δίνει Precision=0,566 και Recall=0,098 και στο σύνολο δεδομένων Joni Hoppen παρουσιάζει Precision=0,42 και Recall=0,13. Σε σχέση λοιπόν με έναν ιδανικό κατηγοριοποιητή ο οποίος θα παρουσίαζε τιμές Recall και Precision κοντά στο 1, παρατηρείται ότι οι παραπάνω τιμές είναι σε γενικές

γραμμές μέτριες προς χαμηλές. Εξαίρεση αποτελεί ο αλγόριθμος Naïve Bayes με SMOTE που παρουσιάζει σχετικά υψηλό Recall (Recall=0,717).

Εν κατακλείδι, σύμφωνα με όσα αναλύθηκαν παραπάνω το γενικό συμπέρασμα που προκύπτει, είναι ότι τα αποτελέσματα ποικίλουν, ανάλογα με το σύνολο δεδομένων που χρησιμοποιείται κάθε φορά. Δεν υπάρχει δηλαδή ένας συγκεκριμένος αλγόριθμος, ο οποίος να θεωρηθεί ότι είναι ο καλύτερος από όλους. Η καλύτερη ή χειρότερη επίδοση των αλγορίθμων που εφαρμόστηκαν στα 2 σύνολα δεδομένων εξαρτάται από τα χαρακτηριστικά που περιέχει το κάθε σύνολο δεδομένων. Επίσης η εφαρμογή της τεχνικής SMOTE δεν μπορεί να θεωρηθεί η καταλληλότερη, καθώς δεν βελτιώνει τα αποτελέσματα σε όλες τις περιπτώσεις. Τέλος σε γενικές γραμμές φαίνεται ότι ο αλγόριθμος k-NN έδωσε τις καλύτερες επιδόσεις, οι οποίες όμως εξαρτώνται από την ταυτόχρονη ή όχι εφαρμογή της τεχνικής SMOTE. Δηλαδή στην 1η ενέργεια έδωσε καλά αποτελέσματα όταν εφαρμόστηκε σε συνδυασμό με την τεχνική SMOTE. Αντίθετα στην 2η ενέργεια έδωσε τα καλύτερα αποτελέσματα όταν εφαρμόστηκε χωρίς SMOTE.

Στον παρακάτω πίνακα παρουσιάζονται συγκεντρωτικά τα βέλτιστα αποτελέσματα που προέκυψαν από τη διαδικασία του grid search και για τις 2 ενέργειες και για τα 2 σύνολα δεδομένων και για τους 3 αλγορίθμους κατηγοριοποίησης (Πίνακας 63).

Ενέργεια	Σύνολο Δεδομένων	Αλγόριθμος	Παράμετρος	Τεχνική (SMOTE)	Optimum Recall	Optimum Precision
1η	Alvaro Flores	k-NN	K=1201	SMOTE (K=10)	0,553	0,233
1η	**JoniHoppen**	**Naïve Bayes**	**-**	**SMOTE (K=6)**	**0,717**	**0,233**
2η	Alvaro Flores	k-NN	K=13	Χωρίς SMOTE	0,098	0,566
2η	JoniHoppen	k-NN	K=9	Χωρίς SMOTE	0,13	0,42

Πίνακας 63: Βέλτιστα αποτελέσματα ανά ενέργεια και σύνολο δεδομένων

9.2 Μελλοντικές επεκτάσεις

Στην παρούσα εργασία το σκεπτικό ήταν να δοκιμαστούν διαφορετικά είδη αλγορίθμων κατηγοριοποίησης. Αρχικά λοιπόν εκτελέστηκαν πειράματα με τον αλγόριθμο J48 που ανήκει στην κατηγορία των δέντρων απόφασης, έπειτα με τον Naïve Bayes που ανήκει στην κατηγορία των αλγορίθμων που εφαρμόζουν το θεώρημα Bayes και κατόπιν με τον αλγόριθμο k-NN που ανήκει στην κατηγορία των αλγορίθμων που χρησιμοποιούν την απόσταση ως μέτρο κατηγοριοποίησης άγνωστων δεδομένων. Κατά την εκτέλεση των

πειραμάτων όμως διαπιστώθηκε ότι γενικά χρειάζεται αρκετός χρόνος για την περάτωση της παραπάνω διαδικασίας. Έτσι λοιπόν ενώ το επιθυμητό ήταν να εκτελεστούν πειράματα και με άλλα είδη αλγορίθμων όπως ο αλγόριθμος MultilayerPerceptron που ανήκει στην κατηγορία των Νευρωνικών δικτύων και ο αλγόριθμος SVM (Μηχανές Διανυσμάτων Υποστήριξης), η ιδέα εγκαταλείφθηκε άμεσα, καθώς οι συγκεκριμένοι αλγόριθμοι χρειαζόταν, συγκριτικά με τους υπόλοιπους αλγορίθμους, πολύ περισσότερο χρόνο για την εκτέλεση πειραμάτων. Αρκεί να αναφερθεί ότι για την εκτέλεση ενός πειράματος 10-fold Cross Validation και μάλιστα χωρίς SMOTE στο σύνολο δεδομένων Alvaro Flores που είναι και το μικρότερο από τα δύο που χρησιμοποιήθηκαν στη διπλωματική, ο MultilayerPerceptron χρειάστηκε 30171,6 δευτερόλεπτα (ή 8,381 ώρες) και ο SVM με Linear Kernel χρειάστηκε 70576,1 δευτερόλεπτα (ή 19,6 ώρες). Αντίθετα ο J48 χρειάστηκε 3972,3 δευτερόλεπτα (ή 1,1 ώρες), ο Naïve Bayes χρειάστηκε 4,5 δευτερόλεπτα (ή 0,00125 ώρες) και ο k-NN 780 δευτερόλεπτα (ή 13 λεπτά). Για την εκτέλεση λοιπόν όλων των πειραμάτων ο συνολικός χρόνος που θα χρειαζόταν θα ήταν πολλαπλάσιος των προαναφερθέντων χρόνων. Επομένως οι αλγόριθμοι MultilayerPerceptron και SVM με Linear Kernel απορρίφθηκαν λόγω μεγάλου χρόνου εκτέλεσης των πειραμάτων.

Μελλοντικά λοιπόν, δεδομένης της ύπαρξης περισσότερου χρόνου, ο στόχος είναι να εκτελεστούν πειράματα και με τους αλγόριθμους MultilayerPerceptron και SVM με Linear Kernel, ώστε να διαπιστωθεί αν δίνουν καλύτερα αποτελέσματα. Επιπλέον, θα μπορούσε να δοκιμαστεί κάποιο άλλο εργαλείο εξόρυξης γνώσης, το οποίο ενδεχομένως να περιέχει ταχύτερες υλοποιήσεις των παραπάνω αλγορίθμων. Επιπρόσθετα κάποιο άλλο εργαλείο θα περιέχει μεγαλύτερη ποικιλία αλγορίθμων.

Σχετικά με τις τεχνικές υπερδειγματοληψίας, το εργαλείο WEKA υποστηρίζει μόνο τις τεχνικές υπερδειγματοληψίας SMOTE και Random oversampling. Στην παρούσα διπλωματική χρησιμοποιήθηκε μόνο η τεχνική υπερδειγματοληψίας SMOTE. Εκτός από αυτή, θα μπορούσε να δοκιμαστεί και η τεχνική Random oversampling. Επιπλέον θα μπορούσαν να δοκιμαστούν και τεχνικές υποδειγματοληψίας.

Ως προς τα δεδομένα ο αρχικός στόχος ήταν να χρησιμοποιηθούν δεδομένα από κάποιο ελληνικό νοσοκομείο και συγκεκριμένα από το Γενικό Νοσοκομείο Κατερίνης, ώστε τα δεδομένα να ακολουθούν τις προδιαγραφές του Ελληνικού Συστήματος Υγείας. Αυτό δυστυχώς δεν κατέστη δυνατό για διάφορους λόγους, οπότε χρησιμοποιήθηκαν δύο ελεύθερα διαθέσιμα σύνολα δεδομένων, τα οποία όμως προέρχονται από νοσοκομεία άλλων χωρών. Συγκεκριμένα από νοσοκομεία της Βραζιλίας και της Χιλής. Πέρα από αυτό όμως, σε αυτά τα σύνολα δεδομένων τα προγραμματισμένα ραντεβού προέρχονται από μία σχετικά μικρή χρονική περίοδο, (3-4 μήνες) που ίσως δεν είναι ικανοποιητική, ώστε να επιτευχθούν καλά αποτελέσματα στην κατηγοριοποίηση άγνωστων δεδομένων.

Επομένως το ιδανικό θα ήταν να υπάρχει η δυνατότητα λήψης δεδομένων από ελληνικά νοσοκομεία, ώστε να εξασφαλιστεί, ότι αυτά θα προέρχονται από μεγαλύτερη χρονική περίοδο, για παράδειγμα δύο ή και περισσότερων ετών. Αυτό ενδεχομένως να βοηθήσει στην επίτευξη καλύτερων αποτελεσμάτων όσον αφορά την κατηγοριοποίηση.

Όσον αφορά τις πιθανές ενέργειες που μπορούν να γίνουν για την επίλυση του προβλήματος της μη εμφάνισης των ασθενών σε προγραμματισμένα ραντεβού, η ταυτόχρονη εφαρμογή της $1^{ης}$ και $2^{ης}$ ενέργειας, όπως αναφέρθηκε σε προηγούμενο κεφάλαιο, δημιουργεί επιπλέον προβλήματα.

Ο συνδυασμός της $1^{ης}$ και $2^{ης}$ ενέργειας μπορεί λοιπόν να γίνει ικανοποιητικά αν μετά από την εφαρμογή της $1^{ης}$ ενέργειας καταγραφούν οι υπενθυμίσεις που έγιναν στους ασθενείς και αν τελικά αυτοί οι ασθενείς προσήλθαν ή όχι στο ραντεβού. Αυτό θα έχει ως αποτέλεσμα να εμπλουτιστεί το σύνολο δεδομένων με επιπλέον δεδομένα, τα οποία μπορούν να βοηθήσουν ώστε η $2^{η}$ ενέργεια να εφαρμόζεται σε $2^{η}$ φάση μόνο για τους ασθενείς που παρά το ότι λαμβάνουν υπενθύμιση, τελικά και πάλι δεν εμφανίζονται στο ραντεβού.

Για παράδειγμα στο σύνολο δεδομένων Joni Hoppen υπάρχει ήδη το χαρακτηριστικό SMS_received (υπενθύμιση μέσω SMS). Με βάση αυτό το χαρακτηριστικό, θα μπορούσε λοιπόν να δημιουργηθεί ένα νέο σύνολο δεδομένων που θα περιέχει μόνο τους ασθενείς που έχουν λάβει υπενθύμιση για ραντεβού (Σε αυτούς ουσιαστικά έχει γίνει η $1^{η}$ ενέργεια). Στη συνέχεια σε αυτό το σύνολο δεδομένων να εφαρμοστεί η παραπάνω μεθοδολογία ώστε να προβλεφθεί ποιοι ασθενείς δεν θα προσέλθουν στο ραντεβού τους, παρόλο που τους έχει γίνει υπενθύμιση. Ο σκοπός αυτής της πρόβλεψης είναι η αντικατάσταση των προγραμματισμένων ραντεβού με νέα ραντεβού (δηλαδή η $2^{η}$ ενέργεια).

Μια άλλη σκέψη είναι να δημιουργηθεί μία εφαρμογή που θα χρησιμοποιεί το καλύτερο μοντέλο με βάση την μελέτη που πραγματοποιήθηκε, ώστε να προβλέπει σε πραγματικό χρόνο (real time), αν ένας ασθενής θα προσέλθει ή όχι στο προγραμματισμένο ραντεβού. Στη συνέχεια με βάση την $1^{η}$ ενέργεια, το σύστημα θα κάνει υπενθύμιση του ραντεβού στον ασθενή ή με βάση την $2^{η}$ ενέργεια, θα κάνει αντικατάσταση του προγραμματισμένου ραντεβού με νέο ραντεβού.

10

Βιβλιογραφία

Alshammari, A., Almalki, R., & Alshammari, R. (2021). Developing a Predictive Model of Predicting Appointment No-Show by Using Machine Learning Algorithms. *Journal of Advances in Information Technology*, *12*(3). https://doi.org/10.12720/jait.12.3.234-239

Alshaya, S., McCarren, A., & Al-Rasheed, A. (2019). Predicting No-show Medical Appointments Using Machine Learning. In A. Alfaries, H. Mengash, A. Yasar, & E. Shakshuki (Eds.), *Advances in Data Science, Cyber Security and IT Applications* (Vol. 1097, pp. 211–223). Springer International Publishing. https://doi.org/10.1007/978-3-030-36365-9_18

Ayodele, T. O. (2010). Machine Learning Overview. In *New Advances in Machine Learning*.

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20–29. https://doi.org/10.1145/1007730.1007735

Batool, T., Abuelnoor, M., El Boutari, O., Aloul, F., & Sagahyroon, A. (2021). Predicting Hospital No-Shows Using Machine Learning. *2020 IEEE International Conference*

on Internet of Things and Intelligence System (IoTaIS), 142–148. https://doi.org/10.1109/IoTaIS50849.2021.9359692

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Wadsworth.

Buczak, A. L., & Guven, E. (2016). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Communications Surveys & Tutorials*, *18*(2), Article 2. https://doi.org/10.1109/COMST.2015.2494502

Carreras-García, D., Delgado-Gómez, D., Llorente-Fernández, F., & Arribas-Gil, A. (2020). Patient No-Show Prediction: A Systematic Literature Review. *Entropy*, *22*(6), Article 6. https://doi.org/10.3390/e22060675

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Cost, S., & Salzberg, S. (1993). A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning*, *10*(1), 57–78. https://doi.org/10.1007/BF00993481

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Dantas, L. F., Fleck, J. L., Cyrino Oliveira, F. L., & Hamacher, S. (2018). No-shows in appointment scheduling – a systematic literature review. *Health Policy*, *122*(4), Article 4. https://doi.org/10.1016/j.healthpol.2018.02.002

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*, 233–240. https://doi.org/10.1145/1143844.1143874

Deza, E., & Deza, M. M. (2009). *Encyclopedia of Distances*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-00234-2

Domingos, P., & Pazzani, M. (1997). On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning*, *29*(2), 103–130.

Dravenstott, R., Kirchner, H. L., Strömblad, C., Boris, D., Leader, J., & Devapriya, P. (2014). *Applying Predictive Modeling to Identify Patients at Risk to No-Show*. 1.

Eibe, F., Hall, M. A., & Witten, I. H. (2016). *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques.* (Fourth Edition). Morgan Kaufmann.

Elvira, C., Ochoa, A., Gonzalvez, J. C., & Mochon, F. (2018). Machine-Learning-Based No Show Prediction in Outpatient Visits. *International Journal of Interactive Multimedia and Artificial Intelligence*, *4*(7), 29. https://doi.org/10.9781/ijimai.2017.03.004

Fan, J., & Li, D. (1998). An overview of data mining and knowledge discovery. *Journal of Computer Science and Technology*, *13*(4), 348–368. https://doi.org/10.1007/BF02946624

Fawzi, A., Moosavi-Dezfooli, S.-M., & Frossard, P. (2016). *Robustness of classifiers: From adversarial to random noise*. 9.

Fayyad, U. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37–54.

Friedman, J. H., Kohavi, R., & Yun, Y. (1996). Lazy decision trees. *AAAI/IAAI, Vol. 1*, 717–724.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian Network Classifiers. *Machine Learning*, *29*(2), 131–163. https://doi.org/10.1023/A:1007465528199

George, A., & Rubin, G. (2003). Non-attendance in general practice: A systematic review and its implications for access to primary health care. *Family Practice*, *20*(2), Article 2. https://doi.org/10.1093/fampra/20.2.178

Goffman, R. M., Harris, S. L., May, J. H., Milicevic, A. S., Monte, R. J., Myaskovsky, L., Rodriguez, K. L., Tjader, Y. C., & Vargas, D. L. (2017). Modeling Patient No-Show History and Predicting Future Outpatient Appointment Behavior in the Veterans Health Administration. *Military Medicine*, *182*(5), Article 5. https://doi.org/10.7205/MILMED-D-16-00345

Grabmeier, J., & Rudolph, A. (n.d.). *Techniques of Cluster Algorithms in Data Mining*. 58.

Ha, T. M., & Bunke, H. (1997). Off-line, handwritten numeral recognition by perturbation method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(5), 535–539.

Haibo He, & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Haibo He, Yang Bai, Garcia, E. A., & Shutao Li. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

Hall, L. O., Bowyer, K. W., Banfield, R. E., Eschrich, S., & Collins, R. (2003). Is Error-Based Pruning Redeemable? *International Journal on Artificial Intelligence Tools*, *12*(03), Article 03. https://doi.org/10.1142/S0218213003001228

Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques* (2nd ed). Elsevier ; Morgan Kaufmann.

Huang, Y.-L., & Hanauer, D. A. (2016). Time dependent patient no-show predictive modelling development. *International Journal of Health Care Quality Assurance*, *29*(4), Article 4. https://doi.org/10.1108/IJHCQA-06-2015-0077

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, *6*(5), 429–449.

Jo, T., & Japkowicz, N. (2004). Class imbalances versus small disjuncts. *ACM SIGKDD Explorations Newsletter*, *6*(1), 40–49. https://doi.org/10.1145/1007730.1007737

John, G. H., & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 338–345.

Jovic, A., Brkic, K., & Bogunovic, N. (2015). A review of feature selection methods with applications. *2015 38th International Convention on Information and Communication*

Technology, Electronics and Microelectronics (MIPRO), 1200–1205. https://doi.org/10.1109/MIPRO.2015.7160458

Koh, H. C., & Tan, G. (2011). Data Mining Applications in Healthcare. *Journal of Healthcare Information Management*, *19*(2), 9.

Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection*. *14*, 1137–1145.

Lavrač, N., & Zupan, B. (2010). Data mining in medicine. In *Data Mining and Knowledge Discovery Handbook* (Second Edition, pp. 1111–1136). Springer.

Lee, G., Wang, S., Dipuro, F., Hou, J., Grover, P., Low, L. L., Liu, N., & Loke, C. Y. (2017). Leveraging on Predictive Analytics to Manage Clinic No Show and Improve Accessibility of Care. *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 429–438. https://doi.org/10.1109/DSAA.2017.25

Maimon, O., & Rokach, L. (Eds.). (2010). *Data Mining and Knowledge Discovery Handbook* (Second Edition). Springer US. https://doi.org/10.1007/978-0-387-09823-4

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, *8*(4), 283–298. https://doi.org/10.1016/S0001-2998(78)80014-2

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of Machine Learning (second edition)*. MIT Press.

Murthy, S. K. (1998). Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. *Data Mining and Knowledge Discovery*, *2*(4), 345–389.

Nasir, M., Summerfield, N., Dag, A., & Oztekin, A. (2020). A service analytic approach to studying patient no-shows. *Service Business*, *14*(2), 287–313. https://doi.org/10.1007/s11628-020-00415-8

Nelson, A., Herron, D., Rees, G., & Nachev, P. (2019). Predicting scheduled hospital attendance with artificial intelligence. *Npj Digital Medicine*, *2*(1), 26. https://doi.org/10.1038/s41746-019-0103-3

NHS Digital. (2021, September 23). *Hospital Outpatient Activity 2020-21*. NHS Digital. https://digital.nhs.uk/data-and-information/publications/statistical/hospital-outpatient-activity/2020-21

Ougiaroglou, S. (2014). *Algorithms and Techniqes for Efficient and Effective Nearest Neighbours Classification*. University of Macedonia.

Phyu, T. N. (2009). Survey of Classification Techniques in Data Mining. *Hong Kong*, 5.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. https://doi.org/10.1007/BF00116251

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc.

Rahman, M. M., & Davis, D. N. (2013). Addressing the Class Imbalance Problem in Medical Datasets. *International Journal of Machine Learning and Computing*, 224–228. https://doi.org/10.7763/IJMLC.2013.V3.307

Rashid Ahmed Ahmed, S., Al Barazanchi, I., Mhana, A., & Abdulshaheed, H. R. (2019). Lung cancer classification using data mining and supervised learning algorithms on multi-dimensional data set. *Periodicals of Engineering and Natural Sciences (PEN)*, *7*(2), 438. https://doi.org/10.21533/pen.v7i2.483

Rokach, L., & Maimon, O. (2015). *Data mining with decision trees: Theory and applications* (Second edition). World Scientific.

Salazar, L. H. A., Parreira, W. D., Fernandes, A. M. da R., & Leithardt, V. R. Q. (2022). No-Show in Medical Appointments with Machine Learning Techniques: A Systematic Literature Review. *Information*, *13*(11), 507. https://doi.org/10.3390/info13110507

Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H., & Santos, J. (2018). Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches [Research Frontier]. *IEEE Computational Intelligence Magazine*, *13*(4), 59–76. https://doi.org/10.1109/MCI.2018.2866730

Srinivas, S., & Ravindran, A. R. (2018). Optimizing outpatient appointment system using machine learning algorithms and scheduling rules: A prescriptive analytics

framework. *Expert Systems with Applications*, *102*, 245–261. https://doi.org/10.1016/j.eswa.2018.02.022

Stanfill, C., & Waltz, D. (1986). Toward memory-based reasoning. *Communications of the ACM*, *29*(12), 1213–1228. https://doi.org/10.1145/7902.7906

Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). CLASSIFICATION OF IMBALANCED DATA: A REVIEW. *International Journal of Pattern Recognition and Artificial Intelligence*, *23*(04), 687–719. https://doi.org/10.1142/S0218001409007326

Tan, P.-N., Steinbach, M., & Kumar, V. (2014). *Introduction to data mining* (New internat. edition). Pearson.

Venn, J. (1880). On the diagrammatic and mechanical representation of propositions and reasonings. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *10*(59), 1–18. https://doi.org/10.1080/14786448008626877

Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J.-F., & Hua, L. (2012). Data Mining in Healthcare and Biomedicine: A Survey of the Literature. *Journal of Medical Systems*, *36*(4), 2431–2448. https://doi.org/10.1007/s10916-011-9710-5

Zhou, Z.-H. (2005). Comprehensibility of Data Mining Algorithms. *Journal of Computer Science and Technology - JCST*, 8.