ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

"ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ & ΥΠΗΡΕΣΙΕΣ"

ΑΝΑΛΥΣΗ ΚΑΙ ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΔΕΔΟΜΕΝΑ ΤΡΑΠΕΖΙΚΟΥ ΜΑΡΚΕΤΙΝΓΚ ΜΕ ΤΗ ΧΡΗΣΗ ΤΟΥ WEKA

ΑΠΟ

ΛΕΜΟΝΗ ΒΑΣΙΛΙΚΗ

Υποβάλλεται για την εκπλήρωση των προϋποθέσεων λήψης Μεταπτυχιακού Διπλώματος στην ειδίκευση «ΜΔΑ/ΠΠΣ/ΠΔ» του ΠΜΣ «Πληροφοριακά Συστήματα & Υπηρεσίες» στο ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΕΠΙΒΛΕΠΩΝ: ΜΙΧΑΗΛ ΦΙΛΙΠΠΑΚΗΣ

ΑΚΑΔΗΜΑΪΚΗ ΘΕΣΗ: ΚΑΘΗΓΗΤΗΣ

Πειραιάς,2026

ΠΕΡΙΛΗΨΗ

Η παρούσα μεταπτυχιακή διπλωματική εργασία εξετάζει την εφαρμογή τεχνικών εξόρυξης γνώσης από δεδομένα στον τομέα του τραπεζικού μάρκετινγκ αποσκοπώντας στη διερεύνηση της σχέσης μεταξύ δημογραφικών χαρακτηριστικών και καταναλωτικής συμπεριφοράς πελατών. Τα βασικά ερευνητικά ερωτήματα εστιάζουνστην αναγνώριση προτύπων και συσχετίσεων στα δεδομένα, καθώς και στη συγκριτική αξιολόγηση της αποτελεσματικότητας διαφορετικών μεθόδων εξόρυξης δεδομένων. Συγκεκριμένα, αναφορικά με την ανάλυση δεδομένων χρησιμοποιούνται αλγόριθμοι ταξινόμησης, ομαδοποίησης και εξόρυξης κανόνων συσχέτισης, οι οποίοι εφαρμόζονται μέσω του λογισμικού WEKA. Η αξιολόγηση των αποτελεσμάτων βασίζεται σε κατάλληλες μετρικές απόδοσης και αναδεικνύει τις διαφοροποιήσεις στη συμπεριφορά των πελατών, καθώς και τα πλεονεκτήματα και τους περιορισμούς των επιμέρους αλγορίθμων. Τα ευρήματα της εργασίας μπορούν να αξιοποιηθούν για την υποστήριξη της λήψης αποφάσεων και τον σχεδιασμό πιο στοχευόμενων στρατηγικών τραπεζικού μάρκετινγκ.

ABSTRACT

This master's thesis examines the application of knowledge discovery and data mining techniques in bank marketing, with the aim of investigating the relationship between demographic characteristics and customer consumption behavior. The main research questions focus on identifying patterns and associations within the data and on comparing the effectiveness of different data mining methods. To this end, classification, clustering, and association rule mining algorithms are applied using the "WEKA" software. The evaluation of the results is based on appropriate measurement indicators and highlights differences in customer behavior, as well as the strengths and limitations of the examined methods. The findings of this study can be utilized to support decision-making operations and the design of more targeted bank marketing strategies.

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες προς τον καθηγητή κύριο Φιλιππάκη Μιχαήλ και την συνεργάτη του κυρία Στουγιάννου Ελευθερία για την πολύτιμη καθοδήγηση, τη συνεχή υποστήριξη και τις ουσιαστικές παρατηρήσεις τους καθ' όλη τη διάρκεια εκπόνησης της παρούσας Μεταπτυχιακής Διπλωματικής Εργασίας.

ΛΙΣΤΑ ΠΙΝΑΚΩΝ (LIST OF TABLES)

ΛΙΣΤΑ ΕΙΚΟΝΩΝ (LIST OF FIGURES)

Πίνακας περιεχομένων

ΚΕΦΑΛΑΙΟ 1: ΕΙΣΑΓΩΓΗ

Στον σύγχρονο τραπεζικό τομέα παρατηρείται πώς η ορθή και αποδοτική αξιοποίηση των δεδομένων των πελατών αποτελεί καθοριστικό παράγοντα, ο οποίος συμβάλλει στην ενίσχυση της ανταγωνιστικότητας και της επιχειρησιακής αποδοτικότητας των χρηματοπιστωτικών οργανισμών. Τεράστιος όγκος δεδομένων συλλέγεται καθημερινά από τραπεζικά ιδρύματα. Σε αυτό τον όγκο περιλαμβάνονται τα δημογραφικά στοιχεία, το ιστορικό συναλλαγών και οι πληροφορίες αλληλεπίδρασης με τους πελάτες. Η ακατέργαστη πληροφορία μετατρέπεται σε πολύτιμη και αξιοποιήσιμη γνώση μέσω της συστηματικής ανάλυσης αυτών των δεδομένων. Με τη συγκεκριμένη στρατηγική υποστηρίζεται η λήψη τεκμηριωμένων και ευθυγραμμισμένων στρατηγικών αποφάσεων(1).

Στην ανάλυση μεγάλων και πολύπλοκων συνόλων δεδομένων, οι τεχνικές εξόρυξης γνώσης και μηχανικής μάθησης έχουν αναδειχθεί ως βασικά εργαλεία. Αξίζει να σημειωθεί πώς ο εντοπισμός προτύπων συμπεριφοράς πελατών και η πρόβλεψη της πιθανότητας αποδοχής τραπεζικών προϊόντων καθίστανται δυνατοί μέσω της εφαρμογής αλγορίθμων ταξινόμησης, ομαδοποίησης και εξόρυξης κανόνων συσχέτισης(2). Η χρήση αυτών των συγκεκριμένων τεχνικών μπορεί να συμβάλει στη βελτίωση της στόχευσης καμπανιών, στη μείωση άσκοπων επαφώνκαι μη αποδοτικών επικοινωνιών και στην αύξηση της αποτελεσματικότητας των προωθητικών ενεργειών (3).

Η παρούσα διπλωματική εργασία αποσκοπεί στην ανάλυση δεδομένων τραπεζικού μάρκετινγκ μέσω της εφαρμογής αλγορίθμων μηχανικής μάθησης καθώς επίσης και στην αξιολόγηση της απόδοσής τους ως προς την πρόβλεψη της αποδοχής τραπεζικών προϊόντων από τους πελάτες. Πιο συγκεκριμένα, με στόχο την ανάδειξη των καταλληλότερων μεθόδων για πρακτική εφαρμογή στον τραπεζικό τομέα, η εργασία εστιάζει στη συγκριτική μελέτη διαφορετικών αλγορίθμων. Τα μοντέλα αυτά αξιολογούνται με κατάλληλες μετρικές απόδοσης, λαμβάνοντας υπόψη τις κρίσιμες ιδιαιτερότητες των δεδομένων, όπως είναι η ανισορροπία των κλάσεων, η οποία αποτελεί σύνηθες και ιδιαίτερα απαιτητικό φαινόμενο σε ζητήματα τραπεζικού μάρκετινγκ (4).

ΚΕΦΑΛΑΙΟ 2: ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

2.1 Εξόρυξη Γνώσης από Δεδομένα (Knowledge Discovery in Databases - KDD)

Η εξόρυξη γνώσης από δεδομένα (Knowledge Discovery in Databases – KDD) αποτελεί μια συστηματική και επαναλαμβανόμενη διαδικασία, η οποία αποσκοπεί στην εύρεση της έγκυρης, της χρήσιμης και της κατανοητής γνώσης μέσα από μεγάλους όγκους δεδομένων. Η διαδικασία αυτή περιλαμβάνει συνδυασμό τεχνικών από τους τομείς της στατιστικής, της μηχανικής μάθησης και της βάσης δεδομένων. Οι τεχνικές αυτές συμβάλλουν στη μετατροπή της ακατέργαστης πληροφορίας σε αξιοποιήσιμη γνώση (5).

Στο πλαίσιο του τραπεζικού μάρκετινγκ, η διαδικασία της εξόρυξης γνώσης από δεδομένα (KDD) αποκτά ιδιαίτερη σημασία. Μέσω της εφαρμογής αυτής της διαδικασίας επιτρέπεται η ανάλυση δεδομένων πελατών, αποσκοπώντας στην κατανόηση της συμπεριφοράς τους και στη βελτιστοποίηση των στρατηγικών προώθησης τραπεζικών προϊόντων. Επιπλέον με την εφαρμογή της εξόρυξης γνώσεων από δεδομένα, οι τράπεζες μπορούν να εντοπίσουν πρότυπα συμπεριφοράς, μέσω των οποίων επιτρέπεται η πρόβλεψη της ανταπόκρισης των πελατών σε συγκεκριμένες καμπάνιες μάρκετινγκ καθώς και η υποστήριξη της λήψης τεκμηριωμένων επιχειρησιακών αποφάσεων(1).

2.2 Μηχανική Μάθηση (Machine Learning, ML)

Η μηχανική μάθηση αποτελεί έναν θεμελιώδη υποτομέα της τεχνητής νοημοσύνης, ο οποίος εστιάζει στην ανάπτυξη αλγορίθμων και μεθοδολογιών που επιτρέπουν στα υπολογιστικά συστήματα να μαθαίνουν από δεδομένα και να βελτιώνουν την απόδοσή τους μέσω της εμπειρίας, χωρίς να απαιτούνται ρητά προγραμματισμένες οδηγίες (6). Σε αντίθεση με τις παραδοσιακές προγραμματιστικές προσεγγίσεις, στις οποίες οι κανόνες καθορίζονται χειροκίνητα, η μηχανική μάθηση χρησιμοποιεί τα δεδομένα ως πρωταρχική πηγή γνώσης για την αυτόματη αναγνώριση προτύπων, τη μοντελοποίηση σύνθετων σχέσεων, τη λήψη αποφάσεων και την πρόβλεψη μελλοντικών συμπεριφορών (7).

Η εφαρμογή τεχνικών μάθησης έχει καταστεί ιδιαίτερα διαδεδομένη στον τραπεζικό τομέα καθώς παρέχει τη δυνατότητα ανάλυσης μεγάλων & πολύπλοκων συνόλων δεδομένων με την υποστήριξη συγκεκριμένων διαδικασιών όπως είναι η πρόβλεψη της συμπεριφοράς πελατών, η αξιολόγηση κινδύνου και η βελτιστοποίηση στρατηγικών μάρκετινγκ.

Η μηχανική μάθηση διακρίνεται σε δύο βασικές κατηγορίες:

2.2.1 Επιβλεπόμενη μάθηση (Supervised Learning)

Στην επιβλεπόμενη μάθηση, ο αλγόριθμος είναι εκπαιδευμένος να χρησιμοποιεί ένα σύνολο δεδομένων, στο οποίο κάθε δείγμα συνοδεύεται από μια γνωστή μεταβλητή-στόχο (label). Η κατασκευή ενός μοντέλου το οποίο θα μπορεί να προβλέπει την τιμή ή την κατηγορία της μεταβλητής-στόχου για νέα, άγνωστα δεδομένα αποτελεί το στόχο της επιβλεπόμενης μάθησης(8). Οι κύριες αυτές τεχνικές περιλαμβάνουν τις παρακάτω τις τεχνικές:

- **Κατηγοριοποίηση (Classification):** Σύμφωνα με την τεχνική της κατηγοριοποίησης, τα δεδομένα ταξινομούνται σε διακριτές κατηγορίες. Πιο συγκεκριμένα, ένα παράδειγμα στο τραπεζικό μάρκετινγκ, αποτελεί η πρόβλεψη κατά την οποία αξιολογείται αν ένας πελάτης θα αποδεχθεί ή όχι ένα προϊόν, σύμφωνα με τα δημογραφικά και τα συμπεριφορικά του χαρακτηριστικά (9).
- **Παλινδρόμηση (Regression):** Με την εφαρμογή της τεχνικής παλινδρόμησης, η μεταβλητή-στόχος είναι μια μεταβλητή συνεχούς φύσεως και το μοντέλο είναι εκπαιδευμένο να προβλέπει αριθμητικές τιμές, όπως είναι η εκτίμηση ύψους καταθέσεων ή δαπανών της πιστοληπτικής ικανότητας ενός πελάτη (10).

2.2.2 Μη επιβλεπόμενη μάθηση (Unsupervised Learning)

Στη μη επιβλεπόμενη μάθηση δεν υφίσταται προκαθορισμένη μεταβλητή-στόχος. Οι αλγόριθμοι εξετάζουν τα δεδομένα αποσκοπώντας στην αυτόματη ανακάλυψη εσωτερικών δομών ή προτύπων και σχέσεων όπως ομάδες παρατηρήσεων ή

συσχετίσεων μεταξύ των δειγμάτων(11). Η πιο διαδεδομένη τεχνική θεωρείται η ομαδοποίηση (clustering), η οποία μπορεί να χρησιμοποιηθεί για την ανίχνευση ομάδων πελατών με παρόμοια δημογραφικά χαρακτηριστικά ή συμπεριφορές, χωρίς την απαίτηση προηγούμενης γνώση της κατάταξής τους. Με αυτό τον τρόπο διευκολύνεται η σχεδίαση διαφοροποιημένων στρατηγικών μάρκετινγκ και εξατομικεύονται οι παρεχόμενες υπηρεσίες.

Στην παρούσα εργασία εφαρμόζεται η προσέγγιση της επιβλεπόμενης μάθησης, καθώς τα δεδομένα περιλαμβάνουν γνωστή μεταβλητή-στόχο. Η μεταβλητή αυτή υποδηλώνει την αποδοχή ή μη των τραπεζικών προϊόντων από τους πελάτες. Η χρήση της επιβλεπόμενης μάθησης επιτρέπει την εκπαίδευση μοντέλων πρόβλεψης. Τα μοντέλα αυτά μπορούν να αξιολογηθούν και να βελτιστοποιηθούν με αντικειμενικά κριτήρια απόδοσης, όπως είναι η ακρίβεια (accuracy), η ευαισθησία (recall) και η ειδικότητα (specificity), παρέχοντας αξιόπιστα εργαλεία σε ότι αφορά την υποστήριξη στρατηγικών αποφάσεων στον τραπεζικό τομέα (12).

Η αξιολόγηση των μοντέλων μηχανικής μάθησης αποτελεί κρίσιμο στάδιο κατά τη διαδικασία πρόβλεψης, καθώς επιτρέπει την εκτίμηση της ποιότητας των προβλέψεων καθώς επίσης και τη σύγκριση διαφορετικών αλγορίθμων(8). Στο πλαίσιο του τραπεζικού μάρκετινγκ, η ακριβής πρόβλεψη της αποδοχής προϊόντων από τους πελάτες αποφέρει άμεσο επιχειρησιακό αντίκτυπο, καθώς οι αποφάσεις που βασίζονται σε αναξιόπιστα μοντέλα μπορούν να οδηγήσουν σε οικονομικές απώλειες ή σε μη αποτελεσματικές καμπάνιες (9).

Η αξιολόγηση των μοντέλων στηρίζεται σε διάφορες μετρικές απόδοσης, οι οποίες προσφέρουν διαφορετικές οπτικές για την ποιότητα των προβλέψεων (13).

1. **Ακρίβεια** (**Accuracy):** Η ακρίβεια ως μετρική απόδοσης εκφράζει το ποσοστό των σωστών προβλέψεων σε σχέση με το σύνολο των παρατηρήσεων. Αν και αποτελεί μία ευρέως χρησιμοποιούμενη μετρική, η ακρίβεια μπορεί να εμφανίζεται παραπλανητική σε περιπτώσεις μη ισορροπημένων δεδομένων, όπου η μία κατηγορία υπερισχύει σημαντικά της άλλης (14).

2. **Ακρίβεια θετικών προβλέψεων (Precision)**: Η ακρίβεια θετικών προβλέψεων ως μετρική απόδοσης εκφράζει το ποσοστό των σωστών θετικών προβλέψεων επί του συνόλου των παρατηρήσεων που ταξινομήθηκαν ως θετικές από το μοντέλο. Συγκεκριμένα η υψηλή ακρίβεια θετικών προβλέψεων στο τραπεζικό μάρκετινγκυποδηλώνει πώς οι πελάτες στους οποίους προτείνονται τα προϊόντα είναι εκείνοι οι πελάτες που είναι πιο πιθανό να τα αποδεχθούν πραγματικά. Με τη χρήση της συγκεκριμένης μετρικής απόδοσης περιορίζεται σημαντικά το λειτουργικό κόστος των άστοχων καμπανιών και οι άσκοπες επαφές(1).
3. **Ανάκληση ή Ευαισθησία (Recall)**: Η μετρική μέθοδος της ανάκλησης υπολογίζει το ποσοστό των πραγματικών θετικών δειγμάτων, τα οποία ανιχνεύθηκαν ορθά σύμφωνα με το μοντέλο αξιολόγησης. Στο πλαίσιο της προώθησης των τραπεζικών προϊόντων, η υψηλή τιμή ευαισθησίας συνεπάγεται την αποτελεσματική αναγνώριση της πλειονότητας των πελατών που είναι πιθανόν να ανταποκριθούν θετικά σε ένα προϊόν και αναγνωρίζεται από το μοντέλο. Αποτέλεσμα αυτής της ανταπόκρισης αποτελεί η μεγιστοποίηση των δυνητικών ευκαιριών πωλήσεων(15).
4. **Αρμονικός μέσος της ακρίβειας και της ανάλυσης (F-measure, F1 Score)**: Η μετρική αυτή μέθοδος αποτελεί τον αρμονικό μέσο των μετρικών μεταξύ «precision» και «recall», προσφέροντας μια συνολική εκτίμηση της απόδοσης του μοντέλου σε περιπτώσεις στις οποίες απαιτείται ισορροπία μεταξύ της ακρίβειας των θετικών προβλέψεων και της δυνατότητας ανίχνευσης όλων των θετικών δειγμάτων (16).
5. **Καμπύλη χαρακτηριστικής λειτουργίας δέκτη (ROC Curve: Receiver Operating Characteristic Curve)**: Η καμπύλη χαρακτηριστικής λειτουργίας δέκτη απεικονίζει τη σχέση μεταξύ του ποσοστού πραγματικών θετικών (True Positive Rate) και του ποσοστού ψευδών θετικών (False Positive Rate) σε διάφορα επίπεδα κατωφλίου απόφασης. Η περιοχή κάτω από την καμπύλη (AUC – Area Under the Curve) αποτελεί ένα συνοπτικό μέτρο της διακριτικής ικανότητας του μοντέλου, καθιστώντας τη μετρική ιδιαίτερα

χρήσιμη σε προβλήματα με ανισορροπημένα δεδομένα, όπως αυτά συναντώνται συχνά στον τραπεζικό τομέα(15).

ΚΕΦΑΛΑΙΟ 3: ΠΕΡΙΓΡΑΦΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΕΡΓΑΛΕΙΑ ΑΝΑΛΥΣΗΣ (DATA DESCRIPTION AND WEKA ANALYSIS TOOLS)

Στην συγκεκριμένη ενότητα πραγματοποιείται θεωρητική ανάλυση των δεδομένων και των εργαλείων που χρησιμοποιήθηκαν στην παρούσα εργασία. Περαιτέρω ανάλυση της πειραματικής διαδικασίας συναντάται στο κεφάλαιο 4.

3.1 Σύνολο Δεδομένων Bank Marketing (Bank Marketing Dataset)

Το σύνολο δεδομένων που χρησιμοποιείται στην παρούσα εργασία προέρχεται από τραπεζικές καμπάνιες άμεσου μάρκετινγκ. Αποτελεί ένα ευρέως χρησιμοποιούμενο και τεκμηριωμένο σύνολο δεδομένων στη σχετική βιβλιογραφία καθώς περιλαμβάνει εκτενή πληροφορία σχετικά με τα δημογραφικά χαρακτηριστικά των πελατών, όπως επίσης και με το ιστορικό αλληλεπίδρασης των πελατών με την τράπεζα(18). Τα δεδομένα αυτά αποτυπώνουν τόσο τα στατικά χαρακτηριστικά των πελατών όσο και τις δυναμικές μεταβλητές οι οποίες σχετίζονται με την επικοινωνία και τη συμπεριφορά τους κατά τη διάρκεια των καμπανιών.

Ειδικότερα, το σύνολο δεδομένων περιλαμβάνει τις παρακάτω μεταβλητές:

- **Ηλικία (Age)**: Η μεταβλητή αυτή συνδέεται με την οικονομική συμπεριφορά και τις προτιμήσεις των πελατών.
- **Επάγγελμα και οικογενειακή κατάσταση (Profession & Family Status):** Το επάγγελμα και η οικογενειακή κατάσταση συνδέονται με κοινωνικοοικονομικούς παράγοντες που επηρεάζουν τη χρηματοοικονομική συμπεριφορά των πελατών(1).
- **Επίπεδο εκπαίδευσης (level of education)**: Η συγκεκριμένη μεταβλητή σχετίζεται με την κατανόηση και την αποδοχή τραπεζικών προϊόντων.

Παράλληλα, οι μεταβλητές που σχετίζονται με τη διαδικασία επικοινωνίας, όπως η διάρκεια της τελευταίας τηλεφωνικής κλήσης (Duration)**,** παρέχουν σημαντικές ενδείξεις σχετικά με το επίπεδο ενδιαφέροντος, την εμπλοκή και την ανταπόκριση των πελατών κατά τη διάρκεια της καμπάνιας. Αυτά τα χαρακτηριστικά μπορούν να χρησιμοποιηθούν για την αναγνώριση προτύπων συμπεριφοράς και την εκτίμηση

της πιθανότητας αποδοχής τραπεζικών προϊόντων, ενισχύοντας την αποτελεσματικότητα των στρατηγικών μάρκετινγκ.

Επιπλέον, το σύνολο δεδομένων περιλαμβάνει πληροφορίες σχετικά με προηγούμενες καμπάνιες μάρκετινγκ, αναφορικά με τον αριθμό επαφών και τα αποτέλεσμα προηγούμενων προσπαθειών επικοινωνίας**.** Τα στοιχεία αυτά έχουν αποδειχθεί κρίσιμα σε ότι αφορά την πρόβλεψη της συμπεριφοράς αποδοχής τραπεζικών προϊόντων(17).

3.1.1 Χαρακτηριστικά και Δομή του Συνόλου Δεδομένων (Characteristics And Structure of the Dataset)

Η μεταβλητή-στόχος του συνόλου δεδομένων αποτελεί μια διχοτομική μεταβλητή, η οποία υποδηλώνει εάν ο πελάτης αποδέχθηκε ή όχι το προσφερόμενο τραπεζικό προϊόν (yes/no). Η ύπαρξη μίας γνωστής μεταβλητής-στόχου καθιστά δυνατή την εφαρμογή επιβλεπόμενων αλγορίθμων μηχανικής μάθησης. Με αυτό τον τρόπο επιτρέπεται η εκπαίδευση και αξιολόγηση μοντέλων ταξινόμησης, τα οποία στοχεύουν στην αξιόπιστη πρόβλεψη της αποδοχής προϊόντων στο πλαίσιο τραπεζικών καμπανιών μάρκετινγκ(5).

Περιγραφή του Συνόλου Δεδομένων

Στην παρούσα εργασία χρησιμοποιείται το σύνολο δεδομένων «bank-full», το οποίο προέρχεται από τραπεζικές καμπάνιες μάρκετινγκ. Το συγκεκριμένο σύνολο δεδομένων περιλαμβάνει συνολικά 45.211 εγγραφές και 16 χαρακτηριστικά, καθώς και μία μεταβλητή-στόχο (class attribute). Η μεταβλητή-στόχος είναι διχοτομική και υποδηλώνει εάν ο πελάτης αποδέχθηκε το προσφερόμενο προϊόν (yes/no).

Ενδεικτικά χαρακτηριστικά του συνόλου δεδομένων αναφέρονται παρακάτω:

- Ηλικία πελάτη(Age)
- Επάγγελμα (Job)
- Οικογενειακή κατάσταση (Marital)
- Επίπεδο εκπαίδευσης (Education)
- Ύπαρξη στεγαστικού δανείου (Housing)

- Ύπαρξη προσωπικού δανείου (Loan)
- Διάρκεια τελευταίας επικοινωνίας (Duration)
- Αριθμός επαφών κατά την καμπάνια (Campaign)
- Αποτελέσματα προηγούμενης καμπάνιας (Poutcome)

Η ποικιλία και η πληρότητα των χαρακτηριστικών καθιστούν το σύνολο δεδομένων ιδιαίτερα κατάλληλο για την εφαρμογή τεχνικών εξόρυξης γνώσης και μηχανικής μάθησης, καθώς επιτρέπει την ανάλυση μοτίβων συμπεριφοράς πελατών και την πρόβλεψη της πιθανότητας αποδοχής τραπεζικών προϊόντων.

3.1.2 Συλλογή δεδομένων (Data Collection)

Η συλλογή δεδομένων αποτελεί ένα κρίσιμο και ουσιαστικό βήμα στη διαδικασία εξόρυξης γνώσης, καθώς καθορίζει σε σημαντικό βαθμό την ποιότητα και την αξιοπιστία της παραγόμενης πληροφορίας. Σε αυτό το στάδιο συγκεντρώνονται δεδομένα από τραπεζικά πληροφοριακά συστήματα, όπως αρχεία συναλλαγών, δημογραφικά χαρακτηριστικά πελατών και πληροφορίες από προηγούμενες καμπάνιες μάρκετινγκ. Ο μεγάλος όγκος και η ετερογένεια των τραπεζικών δεδομένων καθιστούν αναγκαία την οργάνωση, την ενοποίηση και την αποθήκευσή τους, ώστε να είναι κατάλληλα για περαιτέρω ανάλυση και εξαγωγή αξιόπιστων συμπερασμάτων.

3.1.3 Προεπεξεργασία Δεδομένων (Data Preprocessing)

Η προεπεξεργασία των δεδομένων συνιστά ένα από τα πλέον κρίσιμα στάδια της διαδικασίας εξόρυξης γνώσης από δεδομένα και μηχανικής μάθησης, καθώς περιλαμβάνει τον εντοπισμό και τον καθαρισμό των δεδομένων από ελλιπείς, ασυνεπείς ή λανθασμένες τιμές, καθώς και τον μετασχηματισμό τους σε κατάλληλη μορφή για ανάλυση και επεξεργασία. Η ποιότητα και η αξιοπιστία των δεδομένων στο στάδιο αυτό επηρεάζει άμεσα την εγκυρότητα και αξιοπιστία των αποτελεσμάτων καθώς και την ακρίβεια των προβλεπτικών μοντέλων. Η συστηματική και προσεκτική προεπεξεργασία των δεδομένων αποτελεί αναγκαία

προϋπόθεση, η οποία συμβάλλει στην εξαγωγή αξιόπιστων και ουσιαστικών συμπερασμάτων στον τραπεζικό τομέα(2).

3.1.3.1 Αφαίρεση μη χρήσιμων χαρακτηριστικών (Irrelevant Feature Removal)

Σε αρχικό στάδιο, υλοποιήθηκε απομάκρυνση χαρακτηριστικών που δεν παρείχαν ουσιαστική πληροφορία αναφορικά με την πρόβλεψη της μεταβλητής-στόχου ή παρουσίαζαν περιορισμένη συμβολή στο τελικό αποτέλεσμα. Η διαδικασία αυτή οδήγησε στη μείωση του θορύβου, στην ενίσχυση της υπολογιστικής αποδοτικότητας και στη μείωση της πολυπλοκότητας των μοντέλων(18).

3.1.3.2 Κανονικοποίηση και Κλιμάκωση (Normalization & Scaling)

Τα αριθμητικά χαρακτηριστικά μετασχηματίστηκαν ώστε να βρίσκονται σε κοινό εύρος τιμών, συνήθως στο διάστημα [0,1]. Η κανονικοποίηση κρίνεται ιδιαίτερα σημαντική για αλγορίθμους που βασίζονται σε μετρικές απόστασης ή επηρεάζονται από διαφορές κλιμάκων, όπως οι «k-Nearest Neighbors» και οι «Support Vector Machines»(1).

3.1.3.3 Επιλογή Χαρακτηριστικών (Feature Selection)

Στην παρούσα εργασία εφαρμόστηκαν τεχνικές επιλογής χαρακτηριστικών με στόχο τον εντοπισμό των μεταβλητών που παρουσιάζουν τη μεγαλύτερη συσχέτιση αναφορικά με τη μεταβλητή-στόχο. Μέσα από την εφαρμογή της παραπάνω διαδικασίας επιτυγχάνεται η ελαχιστοποίηση του κινδύνου υπερπροσαρμογής, η βελτίωση της ακρίβειας των μοντέλων και η διευκόλυνση της ερμηνείας των αποτελεσμάτων (19).

3.1.3.4 Αντιμετώπιση Ανισορροπίας Δεδομένων (Dealing with data Imbalance)

Το σύνολο δεδομένων παρουσίασε ασύμμετρη κατανομή της μεταβλητής-στόχου, με σαφή υπεροχή της αρνητικής κλάσης έναντι της θετικής. Με σκοπό την αντιμετώπιση αυτού του φαινομένου, εξετάστηκε η εφαρμογή της τεχνικής «SMOTE» (Synthetic Minority Over-sampling Technique)**,** η οποία δημιούργησε

συνθετικά δείγματα της μειονοτικής κλάσης, διατηρώντας τη στατιστική ποικιλία των δεδομένων (20).

Εξαιτίας των περιορισμών της έκδοσης του WEKA που εφαρμόστηκε, δεν κατέστη δυνατή η εφαρμογή της τεχνικής συνθετικής υπερδειγματοληψίας της μειονοτικής κλάσης (Synthetic Minority Over-sampling Technique,SMOTE). Για τον παραπάνω λόγο, εφαρμόστηκε η μέθοδος επαναδειγματοληψίας (Resample), η οποία επιτρέπει την εξισορρόπηση των κλάσεων μέσω διαδικασιών αναδειγματοληψίας (Sampling).

Συγκεκριμένα, η εφαρμογή της παραπάνω μεθόδου υποστηρίζει:

- **Υπερδειγματοληψία (Oversampling)**, μέσω της τυχαίας αντιγραφής δειγμάτων της μειονεκτούσας κλάσης,
- **Υποδειγματοληψία (Undersampling)**, μέσω της τυχαίας μείωσης δειγμάτων της πλειοψηφούσας κλάσης,
- **Συνδυασμό** των δύο προσεγγίσεων **(Oversampling & Undersampling)**, με στόχο την επίτευξη καλύτερης ισορροπίας μεταξύ των κλάσεων.

Σε αντίθεση με την εφαρμογή της τεχνικής συνθετικής υπερδειγματοληψίας της μειονοτικής κλάσης (SMOTE), η μέθοδος επαναδειγματοληψίας (Resample) βασίζεται αποκλειστικά σε υπάρχοντα δεδομένα και όχι στη δημιουργία συνθετικών δειγμάτων, περιορίζοντας τον κίνδυνο τεχνητής παραμόρφωσης του συνόλου δεδομένων. Ωστόσο, ενδέχεται να οδηγήσει σε υπερεκτίμηση ή απώλεια πληροφορίας, ιδιαίτερα σε περιπτώσεις εκτεταμένης υπερδειγματοληψίας ή υποδειγματοληψίας(oversampling or undersampling) (21). Η εφαρμογή της στο WEKA επέτρεψε την παραμετροποίηση της αναδειγματοληψίας και τη διατήρηση της συνολικής κατανομής των χαρακτηριστικών.

Σύμφωνα με τα πειραματικά αποτελέσματα η επαναδειγματοληψία (Resample) βελτίωσε την ικανότητα των μοντέλων αναφορικά με την ανίχνευση θετικών παραδειγμάτων, με κύριο όφελος την αύξηση της ανάκλησης (Recall) της

μειονεκτούσας κλάσης.(Λεπτομερής αναφορά των πειραμάτων που έτρεξαν γίνεται παρακάτω στο κεφάλαιο 4)

3.1.4 Εξόρυξη Προτύπων (Data Mining)

Κατά το στάδιο της εξόρυξης προτύπων εφαρμόζονται αλγόριθμοι μηχανικής μάθησης και τεχνικές ανάλυσης δεδομένων με στόχο την ανακάλυψη συσχετίσεων, προτύπων και τάσεων στη συμπεριφορά των πελατών. Οι αλγόριθμοι καθιστούντον πυρήνα της διαδικασίας εξόρυξης γνώσης. Μέσω της συστηματικής αξιοποίησης των διαθέσιμων χαρακτηριστικών, επιτυγχάνεται τόσο η ομαδοποίηση πελατών με παρόμοια χαρακτηριστικά όσο και η πρόβλεψη της πιθανότητας ανταπόκρισης σε καμπάνιες μάρκετινγκ μέσω αλγορίθμων ταξινόμησης.

Η εξόρυξη προτύπων περιλαμβάνει αφενός τη χρήση τεχνικών ομαδοποίησης, οι οποίες αναδεικνύουν διακριτά τμήματα πελατών με κοινά χαρακτηριστικά και συμπεριφορές, και αφετέρου την αξιοποίηση αλγορίθμων ταξινόμησης που στοχεύουν στην πρόβλεψη της ανταπόκρισης των πελατών σε συγκεκριμένες καμπάνιες ή προϊόντα. Τα αποτελέσματα αυτής της διαδικασίας αποτελούν τη βάση για τη διαμόρφωση στοχευμένων στρατηγικών και τη βελτιστοποίηση των επιχειρησιακών αποφάσεων.

3.1.5 Αξιολόγηση και Ερμηνεία της Γνώσης (Assessment & Interprentation of Knowledge)

Το τελικό στάδιο της διαδικασίας εξόρυξης γνώσης από δεδομένα (KDD) επικεντρώνεται στην αξιολόγηση και την ερμηνεία των αποτελεσμάτων που προκύπτουν. Κατά αυτό το στάδιο, τα αποτελέσματα εξετάζονται ως προς την ακρίβεια, τη στατιστική εγκυρότητα, τη χρησιμότητα και τη δυνατότητα ερμηνείας τους, ώστε να διασφαλιστεί ότι η παραγόμενη γνώση μετατρέπεται σε πρακτικά αξιοποιήσιμα συμπεράσματα. Η αξιοποίηση αυτών των συμπερασμάτων επιτρέπει στους τραπεζικούς οργανισμούς να στηρίξουν αποφάσεις όπως η στοχευμένη προώθηση προϊόντων, η βελτίωση της αποδοτικότητας των καμπανιών μάρκετινγκ και η ανάπτυξη στρατηγικών που βασίζονται σε τεκμηριωμένα δεδομένα.

Σύμφωνα με τη σχετική βιβλιογραφία, η παραπάνω διαδικασία συνιστά ένα θεμελιώδες και ολοκληρωμένο πλαίσιο στο οποίο μπορούν να αναλυθούν τα δεδομένα και να εξαχθούν γνώσεις σε επιχειρησιακά περιβάλλοντα υψηλής πολυπλοκότητας, όπως είναι ο τραπεζικός τομέας. Στον τομέα αυτό, η ποιότητα της πληροφορίας και η αξιοπιστία των συμπερασμάτων συμβάλλουν στον καθοριστικό ρόλο της διαδικασίας λήψης των αποφάσεων(5).

3.2 Το λογισμικό WEKA ως Εργαλείο Ανάλυσης Δεδομένων (Waikato Environment for Knowledge Analysis).

Το WEKA (Waikato Environment for Knowledge Analysis) αποτελεί ένα λογισμικό ανοικτού κώδικα, το οποίο παρέχει ένα ολοκληρωμένο περιβάλλον για την εξερεύνηση, ανάλυση και εξόρυξη γνώσης από δεδομένα. Αναπτύχθηκε από το Πανεπιστήμιο του Waikato στη Νέα Ζηλανδία και χρησιμοποιείται ευρέως στην ερευνητική και εκπαιδευτική κοινότητα με στόχο την εφαρμογή και αξιολόγηση αλγορίθμων μηχανικής μάθησης(22).

Η πλατφόρμα υποστηρίζει ένα ευρύ φάσμα αλγορίθμων ταξινόμησης, παλινδρόμησης και ομαδοποίησης, καλύπτοντας τόσο επιβλεπόμενες όσο και μη επιβλεπόμενες τεχνικές. Ενδεικτικά, περιλαμβάνει:

- Δέντρα απόφασης (π.χ. J48),
- Πιθανοτικά μοντέλα (Naive Bayes),
- Αλγορίθμουςσυνόλων (Ensembles), όπως (Random Forest),
- Υποστηρικτικά διανυσματικά μηχανήματα (Support Vector Machines-SVM),
- Αλγορίθμοι τεχνικών ομαδοποίησης(Clustering)**,**

Η πλατφόρμα αυτή υποστηρίζει τη συγκριτική αξιολόγηση διαφορετικών προσεγγίσεων μηχανικής μάθησης σε ένα ενιαίο πλαίσιο εφαρμογής(2).

Παράλληλα, το εργαλείο WEKA παρέχει εκτεταμένες δυνατότητες προεπεξεργασίας δεδομένων, όπως είναι ο καθαρισμός, η κανονικοποίηση, ο μετασχηματισμός και η επιλογή χαρακτηριστικών. Οι τεχνικές αυτές συμβάλλουν στη βελτίωση της

ποιότητας των δεδομένων και της απόδοσης των μοντέλων, ιδιαίτερα σε περιπτώσεις όπου τα δεδομένα παρουσιάζουν θόρυβο ή ανισορροπία κλάσεων(23).

Ένα από τα κύρια πλεονεκτήματα του λογισμικού WEKA είναι το φιλικό γραφικό περιβάλλον χρήστη (GUI), το οποίο επιτρέπει τον πειραματισμό και την ανάλυση δεδομένων χωρίς να απαιτούνται προγραμματιστικές δεξιότητες. Παράλληλα, το περιβάλλον αυτό παρέχει εργαλεία οπτικοποίησης και παρουσίασης των αποτελεσμάτων, διευκολύνοντας την ερμηνεία και την κατανόηση της συμπεριφοράς των παραγόμενων προτύπων.

Στο πλαίσιο της παρούσας εργασίας, η χρήση του λογισμικού WEKA διευκολύνει την αποτελεσματική υλοποίηση και τη σύνδεση της θεωρητικής γνώσης με την πρακτική εφαρμογή, επιτρέποντας την αποτελεσματική υλοποίηση και την αξιολόγηση τεχνικών μηχανικής μάθησης σε πραγματικά δεδομένα τραπεζικών καμπανιών.

ΚΕΦΑΛΑΙΟ 4: ΠΕΙΡΑΜΑΤΙΚΗ ΔΙΑΔΙΚΑΣΙΑ

Στην παρούσα εργασία, η πειραματική διαδικασία περιλαμβάνει την εκπαίδευση και την αξιολόγηση διαφόρων αλγορίθμων επιβλεπόμενης μάθησης, με στόχο την εφαρμογή τους στο σύνολο δεδομένων Bank Marketing, προκειμένου να μελετηθεί η πρόβλεψη της αποδοχής προθεσμιακών καταθέσεων από τους πελάτες.

4.1 ΠΕΙΡΑΜΑΤΙΚΗ ΜΕΘΟΔΟΛΟΓΙΑ

Στην παρούσα ενότητα περιγράφεται η πειραματική διαδικασία που ακολουθήθηκε για την εκπαίδευση και την αξιολόγηση των αλγορίθμων ταξινόμησης. Συγκεκριμένα, παρέχονται πληροφορίες σχετικά με το περιβάλλον υλοποίησης, το σύνολο δεδομένων που χρησιμοποιήθηκε, καθώς και τις μεθόδους αξιολόγησης που εφαρμόστηκαν, προκειμένου να δια

σφαλιστεί η αντικειμενικότητα και η αξιοπιστία των αποτελεσμάτων.

4.1.1 Περιβάλλον Υλοποίησης (Implementation Environment)

Τα πειράματα πραγματοποιήθηκαν στο λογισμικό WEKA (Explorer). Αρχικά, φορτώθηκε το σύνολο δεδομένων «Bank Marketing» και ορίστηκε ως μεταβλητή-στόχος το χαρακτηριστικό y, το οποίο υποδηλώνει εάν ο πελάτης αποδέχθηκε προθεσμιακή κατάθεση ή όχι. Το περιβάλλον του WEKA επέτρεψε την εύκολη εφαρμογή διαφόρων αλγορίθμων ταξινόμησης, την προεπεξεργασία των δεδομένων και τη διαχείριση των ρυθμίσεων των πειραμάτων, παρέχοντας έτσι μια ολοκληρωμένη πλατφόρμα για την αξιολόγηση της απόδοσης κάθε μοντέλου.

4.1.2 Διαδικασία Αξιολόγησης (Evaluation Procedure)

Η αξιολόγηση όλων των πειραμάτων πραγματοποιήθηκε με τη χρήση της μεθόδου «10-foldcrossvalidation». Με την εφαρμογή της συγκεκριμένης μεθόδου εξασφαλίστηκαν έγκυρες εκτιμήσεις της απόδοσης των μοντέλων. Η σύγκριση των αλγορίθμων βασίστηκε σε συγκεκριμένες μετρικές αξιολόγησης:

- Ακρίβεια (Accuracy)

- Ακρίβεια θετικών προβλέψεων (Precision)
- Ανακλησιμότητα / Ευαισθησία (Recall)
- Επιφάνεια κάτω από την καμπύλη ROC (ROCAUC)

Μέσω της χρήσης των παραπάνω μετρικών επιτρέπεται η ολοκληρωμένη αξιολόγηση τόσο της γενικής απόδοσης όσο και της ικανότητας των μοντέλων στη σωστή αναγνώριση των μειονοτικών (θετικών) κλάσεων.

4.2 Αλγόριθμοι και Πειράματα Ταξινόμησης (Classification)

Στην ενότητα αυτή παρουσιάζονται οι αλγόριθμοι ταξινόμησης που εφαρμόστηκαν στο σύνολο δεδομένων «Bank Marketing», μαζί με τις αντίστοιχες πειραματικές διαδικασίες, τις ρυθμίσεις στο λογισμικό WEKA και τα αποτελέσματα της αξιολόγησης. Οι αλγόριθμοι αναλύονται ξεχωριστά σε υποενότητες, προκειμένου να δοθεί πλήρης εικόνα της λειτουργίας τους, των παραμέτρων που χρησιμοποιήθηκαν και της απόδοσής τους. Κάθε υποενότητα περιλαμβάνει τη περιγραφή του αλγορίθμου, τη διαδικασία προεπεξεργασίας, τις ρυθμίσεις WEKA, τα αποτελέσματα αξιολόγησης και τις βασικές παρατηρήσεις σχετικά με την καταλληλότητα του αλγορίθμου για προβλέψεις στον τραπεζικό τομέα.

Μέσα από τη διαδικασία της ανάλυσης επιτρέπεται η συγκριτική αξιολόγηση των μοντέλων και έτσι εντοπίζονται ποιοι αλγόριθμοι προσφέρουν την καλύτερη πρόβλεψη της αποδοχής προϊόντων και ποιοι παρέχουν καλύτερη ερμηνευσιμότητα στη λήψη τεκμηριωμένων επιχειρησιακών αποφάσεων.

4.2.1 Βασικό μοντέλο αλγορίθμου «Μηδενικοί Κανόνες» (Zero R Baseline Model)

Περιγραφή αλγορίθμου «Μηδενικοί Κανόνες» (ZeroR Baseline Model)

Ο αλγόριθμος «ZeroR» αποτελεί ένα πολύ βασικό αλγόριθμο ταξινόμησης, ο οποίος αγνοεί όλα τα χαρακτηριστικά των δεδομένων και προβλέπει πάντα την πιο συχνή κλάση στο σύνολο εκπαίδευσης. Χρησιμοποιείται κυρίως ως σημείο αναφοράς (baseline) προκειμένου να αξιολογείται η περίπτωση στην οποία οι πιο σύνθετοι αλγόριθμοι να αποδίδουν καλύτερα. Συνήθως εμφανίζει μεγάλη ακρίβεια και δεν

εκπαιδεύει μοτίβα, αλλά χρησιμεύει για σύγκριση καθώς η επίδοσή του επιτρέπει την εκτίμηση του οφέλους από την εφαρμογή αλγορίθμων, οι οποίοι ενσωματώνουν την πληροφορία των χαρακτηριστικών και των σχέσεων μεταξύ τους.

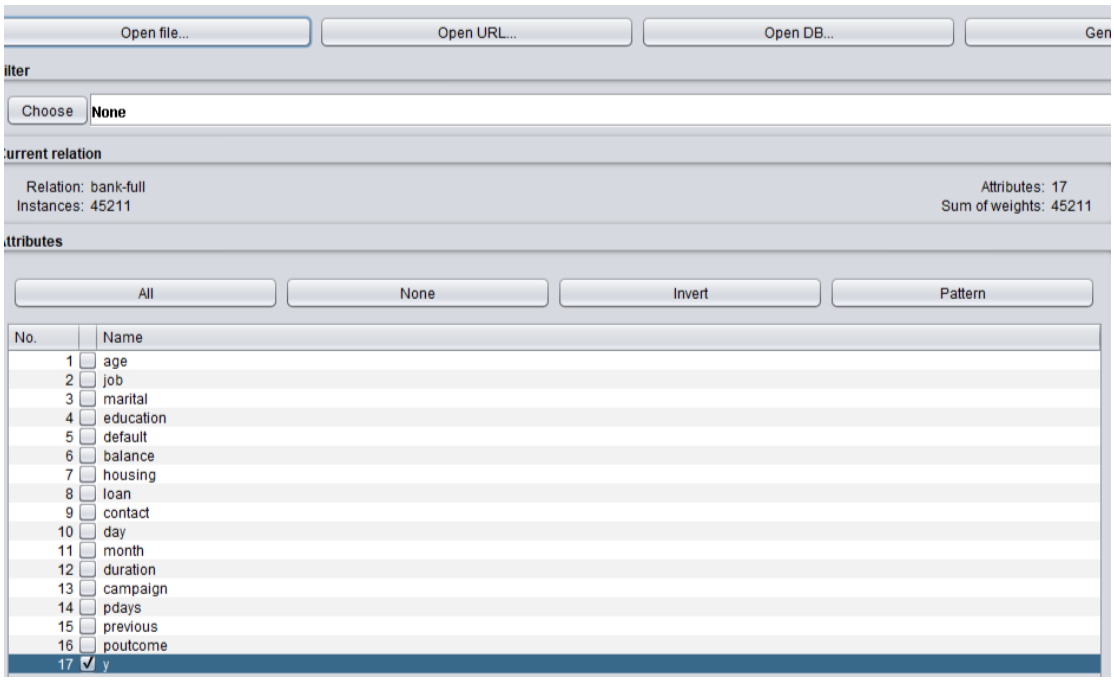
Ρυθμίσεις WEKA

Ο αλγόριθμος «ZeroR» αποτελεί τον πιο απλό ταξινομητή, καθώς αφαιρεί όλα τα χαρακτηριστικά (predictors) και βασίζεται αποκλειστικά στην κλάση-στόχο. Σε κατηγορικά δεδομένα προβλέπει πάντα την πλειοψηφούσα κλάση, ενώ σε αριθμητικά δεδομένα επιστρέφει τον μέσο όρο. Η χρήση του χρησιμεύει στον καθορισμό της χαμηλότερης αποδεκτής ακρίβειας. Στην περίπτωση που ένας πιο σύνθετος αλγόριθμος, όπως ο SMO, δεν υπερβαίνει την ακρίβεια του «ZeroR», τότε το μοντέλο δεν έχει εκπαιδευτεί ουσιαστικά με βάση τα δεδομένα.

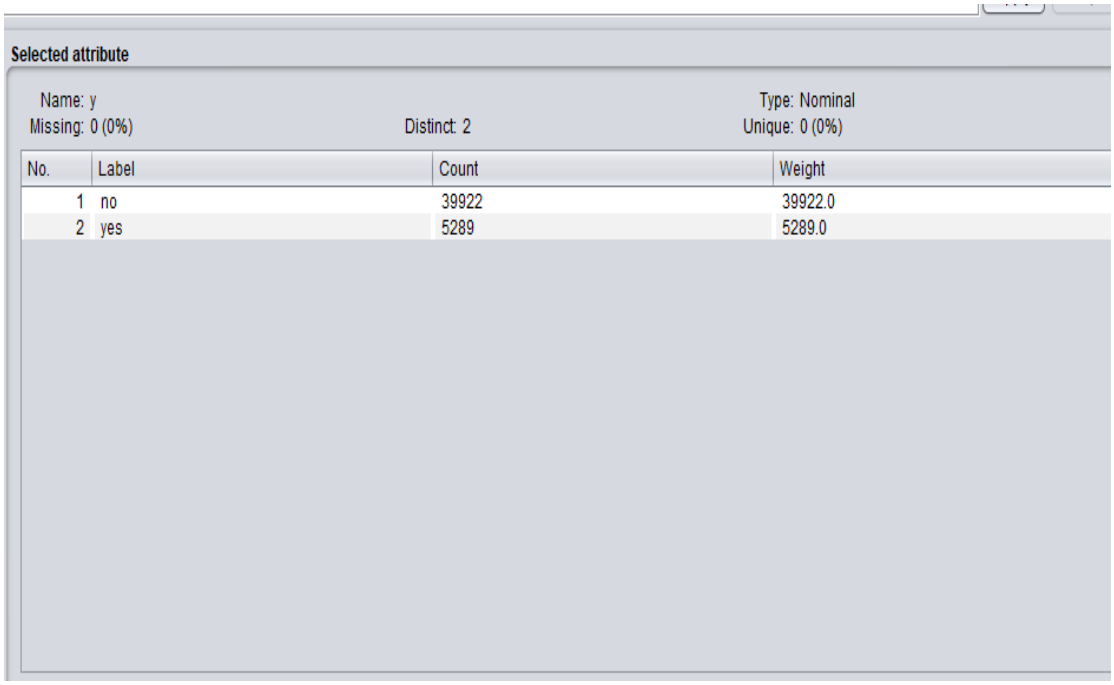
Η αξιολόγηση γίνεται με τη χρήση της μεταβλητής «10-fold Cross Validation», η οποία διασφαλίζει ότι η εκτίμηση της βασικής αυτής ακρίβειας είναι στατιστικά αξιόπιστη, καθώς ελέγχει την κατανομή των κλάσεων σε όλο το σύνολο δεδομένων. Παρακάτω παρουσιάζονται οι ρυθμίσεις και τα αποτελέσματα του αλγορίθμου «ZeroR» στο WEKA, συμπεριλαμβανομένου του «Classification Summary» και του «Confusion Matrix».

- Classify → rules → ZeroR
- 10-fold Cross Validation → Start

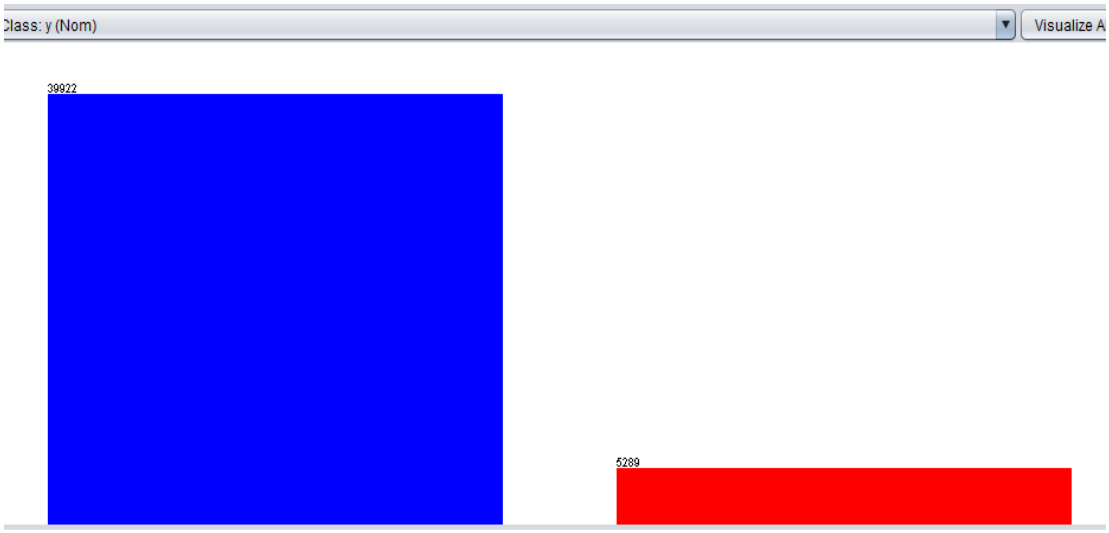
A]



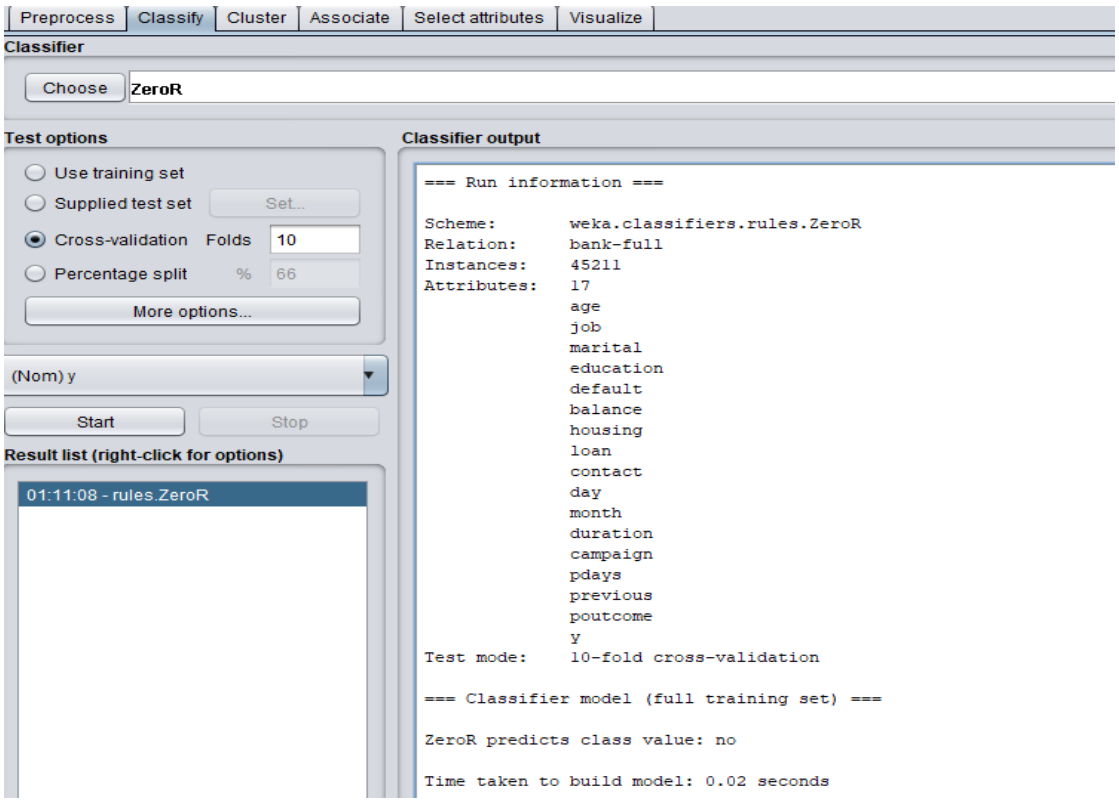
B]



Εικόνα 1: [Α,Β] Καρτέλα Προεπεξεργασίας δεδομένων του λογισμικού WEKA Explorer με επισκόπηση του συνόλου δεδομένων «Bank Marketing»



Εικόνα 2: Κατανομή κλάσεων με έντονη ανισορροπία μεταξύ πλειοψηφικής (μπλε) και μειοψηφικής (κόκκινη) κατηγορίας.



Εικόνα 3: ZeroRστοWEKAμε 10-foldCross Validation.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       39922               88.3015 %
Incorrectly Classified Instances      5289               11.6985 %
Kappa statistic                          0
Mean absolute error                      0.2066
Root mean squared error                  0.3214
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances            45211

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.883      1.000    0.938      ?        0.500     0.883     no
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.117     yes
Weighted Avg.    0.883    0.883    ?          0.883    ?          ?        0.500     0.793

=== Confusion Matrix ===

     a     b   <-- classified as
 39922     0 |     a = no
  5289     0 |     b = yes
```

Εικόνα 4: Αποτελέσματα ταξινόμησης αλγορίθμου (ZeroR) με stratified 10-foldCross-validation, παρουσιάζοντας ακρίβεια 88,3%, λεπτομερή μετρικά ανά κλάση και πίνακα σύγχυσης.

Αποτελέσματα και παρατηρήσεις

Ο αλγόριθμος «Μηδενικοί Κανόνες» (ZeroR) χρησιμοποιήθηκε ως μοντέλο βάσης (baseline) προκειμένου να αξιολογηθεί η σχετική απόδοση των πιο σύνθετων αλγορίθμων. Τα πειραματικά αποτελέσματα έδειξαν αρκετά καλή προβλεπτική ικανότητα (88,3%), καθώς ο αλγόριθμος προβλέπει πάντα την πλειοψηφική κλάση της μεταβλητής-στόχου. Η αρκετά καλή προβλεπτική ικανότητα του αλγορίθμου αναδεικνύει την αποτελεσματική αποτύπωση της σχέσης μεταξύ των μεταβλητών και μπορεί να εφαρμοστεί με αξιοπιστία για προβλέψεις. Με άλλα λόγια, το μοντέλο υποδηλώνει καλή προσαρμογή και ικανοποιητική απόδοση και κρίνεται αρκετά αξιόπιστο για πρόβλεψη.

4.2.2 Αλγόριθμος δένδρου αποφάσεων (J48, DecisionTree)

Περιγραφή αλγορίθμου

Ο αλγόριθμος δέντρου αποφάσεων (J48) αποτελεί μια μέθοδο κλασικού δέντρου αποφάσεων, η οποία βασίζεται σε κριτήρια διαχωρισμού όπως η απόκτηση της πληροφορίας (information gain) για τη διαδοχική κατάτμηση των δεδομένων. Η δομή του δέντρου επιτρέπει την κατηγοριοποίηση των δειγμάτων βάσει των χαρακτηριστικών τους, ενώ παράλληλα παρέχει εύκολα ερμηνεύσιμα μοντέλα. Με αυτόν τον τρόπο, γίνεται δυνατή η αναγνώριση των μεταβλητών που ασκούν μεγαλύτερη επιρροή στην πρόβλεψη της μεταβλητής-στόχου, ενισχύοντας την κατανόηση των χαρακτηριστικών που επηρεάζουν τη λήψη τεκμηριωμένων επιχειρησιακών αποφάσεων.

Ρυθμίσεις WEKA

Στο WEKA, η ανάλυση πραγματοποιείται με τον αλγόριθμο «J48», ο οποίος δημιουργεί ένα δέντρο απόφασης διαχωρίζοντας τα δεδομένα σε υποσύνολα με βάση τα πιο σημαντικά χαρακτηριστικά. Ο αλγόριθμος είναι ανθεκτικός σε ελλιπείς τιμές (missing values) και δεν απαιτεί απαραίτητα προεπεξεργασία κανονικοποίησης (normalization). Παράλληλα, παρέχει τη δυνατότητα οπτικοποίησης του δέντρου (Visualize Tree), διευκολύνοντας την κατανόηση της λογικής των αποφάσεων. Η χρήση της δεκαπλής διασταυρούμενης επικύρωσης (10-fold Cross Validation) διασφαλίζει ότι το δένδρο που προκύπτει διαθέτει γενικευτική ισχύ και δεν περιορίζεται στην απλή απομνημόνευση του συνόλου εκπαίδευσης.

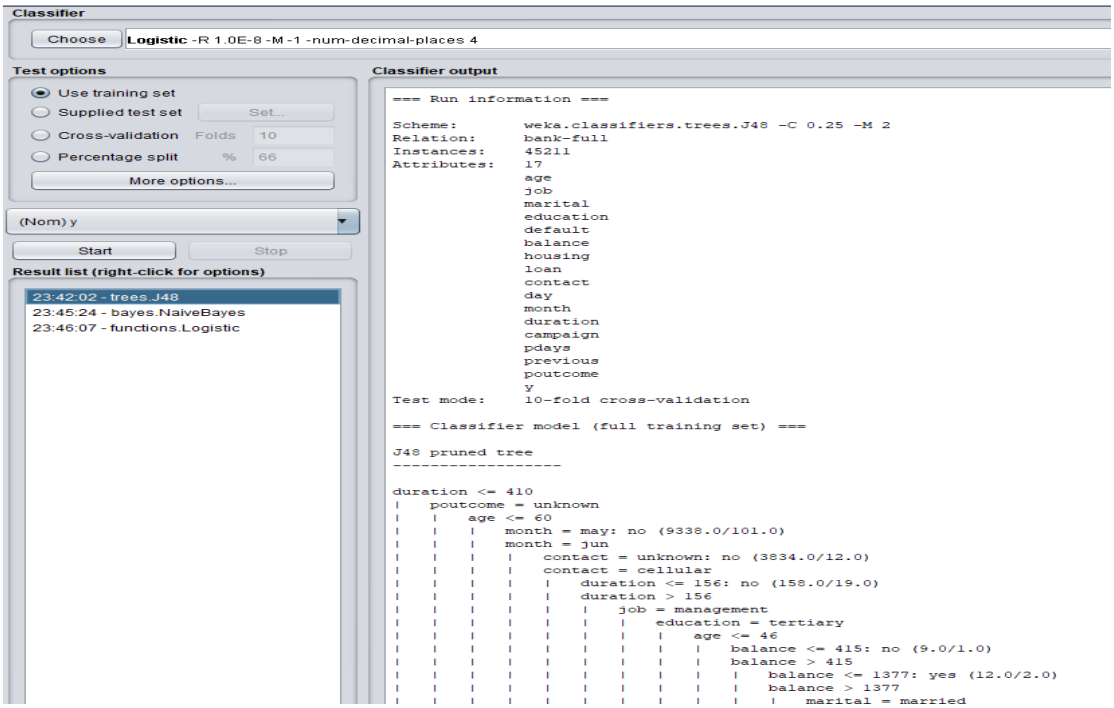
- Classifier → trees → J48
- Test option → 10-fold Cross Validation

Αποτελέσματα αναλυτικών μετρήσεων

Μέσω της διαδικασίας της στατιστικής ανάλυσης των αποτελεσμάτων παρουσιάζεταιμια μικτή εικόνα αναφορικά με την αποτελεσματικότητα του μοντέλου:

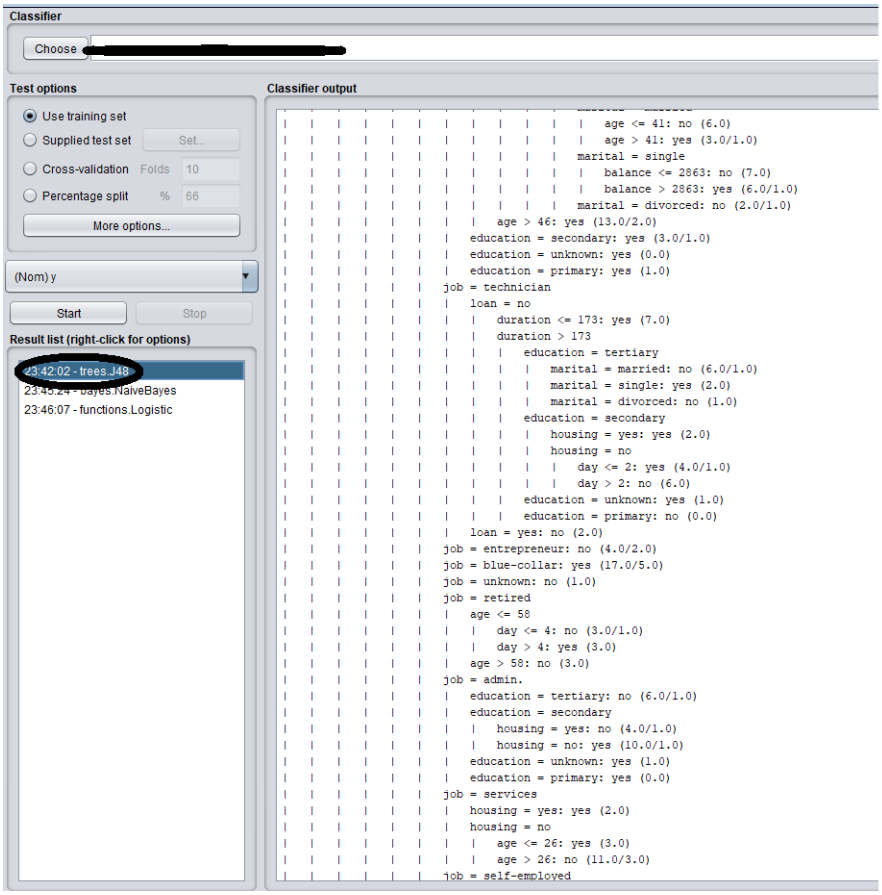
- Ακρίβεια (Accuracy): 90.31%

- Ακρίβεια θετικών προβλέψεων (Precision, yes): 0.60
- Ανακλησιμότητα / Ευαισθησία (Recall, για class=yes): 0.48
- Βαθμολογία F1 (F1-Score): 0.53



Εικόνα 5: Pruned (J48) decision tree με Accuracy 90,3% και ισορροπία «no/yes».

A]

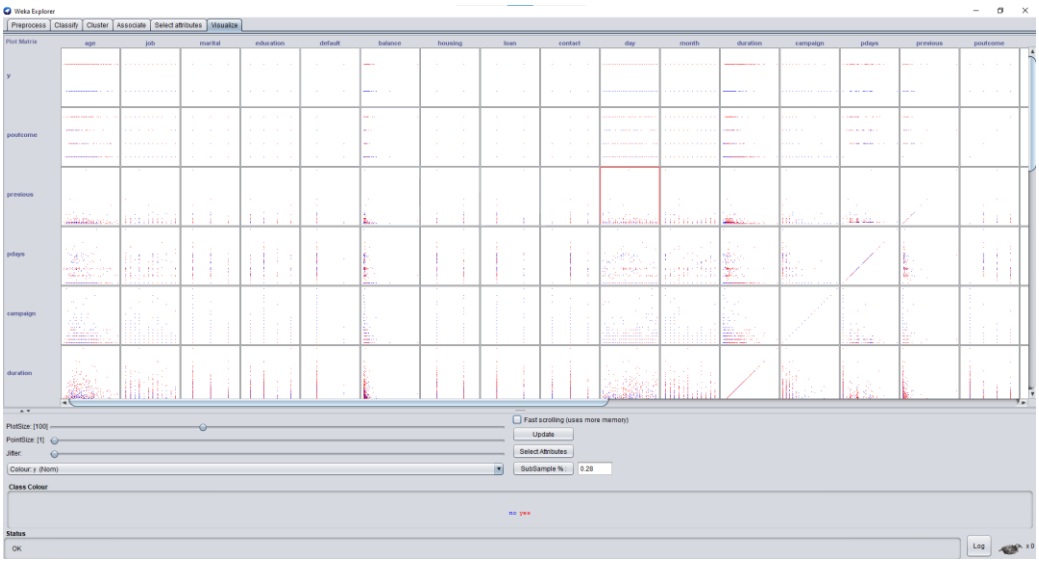

Classifier
Choose
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) y
Start
Stop
Result list (right-click for options)
23:42:02 - trees.J48
23:45:24 - bayes.NaiveBayes
23:46:07 - functions.Logistic
Classifier output
| | | | | | | | | | | | | age <= 41: no (6.0)
| | | | | | | | | | | | | age > 41: yes (3.0/1.0)
| | | | | | | | | | | | marital = single
| | | | | | | | | | | | | balance <= 2863: no (7.0)
| | | | | | | | | | | | | balance > 2863: yes (6.0/1.0)
| | | | | | | | | | | | marital = divorced: no (2.0/1.0)
| | | | | | | | | | | age > 46: yes (13.0/2.0)
| | | | | | | | | | education = secondary: yes (3.0/1.0)
| | | | | | | | | | education = unknown: yes (0.0)
| | | | | | | | | | education = primary: yes (1.0)
| | | | | | | | job = technician
| | | | | | | | | loan = no
| | | | | | | | | | duration <= 173: yes (7.0)
| | | | | | | | | | duration > 173
| | | | | | | | | | | education = tertiary
| | | | | | | | | | | | marital = married: no (6.0/1.0)
| | | | | | | | | | | | marital = single: yes (2.0)
| | | | | | | | | | | | marital = divorced: no (1.0)
| | | | | | | | | | | education = secondary
| | | | | | | | | | | | housing = yes: yes (2.0)
| | | | | | | | | | | | housing = no
| | | | | | | | | | | | | day <= 2: yes (4.0/1.0)
| | | | | | | | | | | | | day > 2: no (6.0)
| | | | | | | | | | | education = unknown: yes (1.0)
| | | | | | | | | | | education = primary: no (0.0)
| | | | | | | | | loan = yes: no (2.0)
| | | | | | | | job = entrepreneur: no (4.0/2.0)
| | | | | | | | job = blue-collar: yes (17.0/5.0)
| | | | | | | | job = unknown: no (1.0)
| | | | | | | | job = retired
| | | | | | | | | age <= 58
| | | | | | | | | | day <= 4: no (3.0/1.0)
| | | | | | | | | | day > 4: yes (3.0)
| | | | | | | | | age > 58: no (3.0)
| | | | | | | | job = admin.
| | | | | | | | | education = tertiary: no (6.0/1.0)
| | | | | | | | | education = secondary
| | | | | | | | | | housing = yes: no (4.0/1.0)
| | | | | | | | | | housing = no: yes (10.0/1.0)
| | | | | | | | | education = unknown: yes (1.0)
| | | | | | | | | education = primary: yes (0.0)
| | | | | | | | job = services
| | | | | | | | | housing = yes: yes (2.0)
| | | | | | | | | housing = no
| | | | | | | | | | age <= 26: yes (3.0)
| | | | | | | | | | age > 26: no (11.0/3.0)
| | | | | | | | job = self-employed

B]

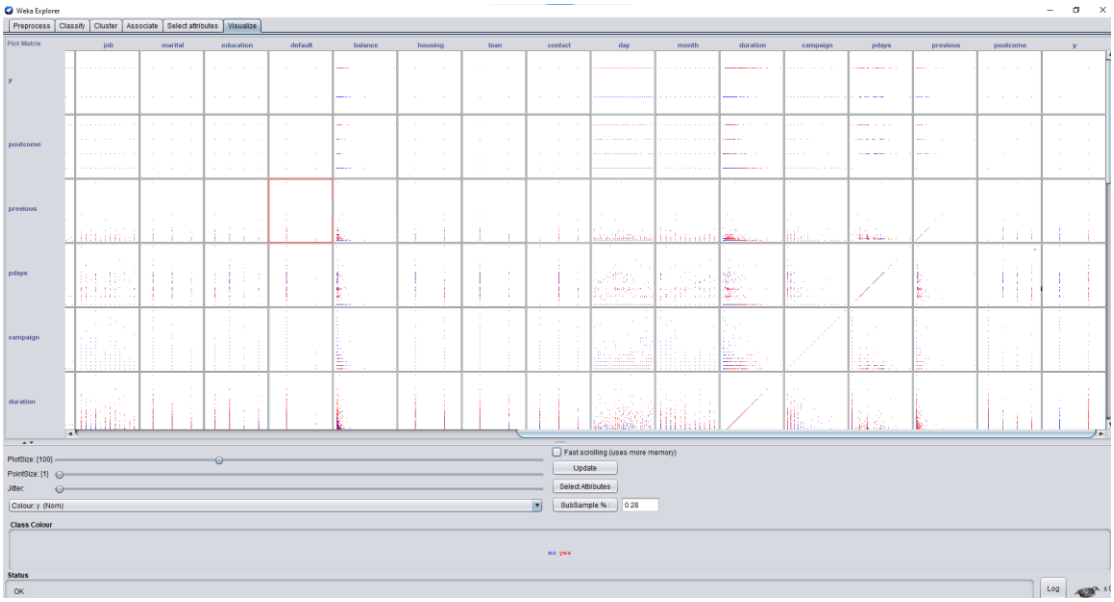
```
Time taken to build model: 2.4 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       40831               90.3121 %
Incorrectly Classified Instances      4380                9.6879 %
Kappa statistic                          0.4839
Mean absolute error                      0.1269
Root mean squared error                  0.2773
Relative absolute error                 61.4259 %
Root relative squared error             86.2833 %
Total Number of Instances            45211

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.959    0.519    0.933      0.959    0.946      0.488    0.843     0.947     no
                 0.481    0.041    0.609      0.481    0.537      0.488    0.843     0.486     yes
Weighted Avg.    0.903    0.463    0.895      0.903    0.898      0.488    0.843     0.893

=== Confusion Matrix ===

     a     b   <-- classified as
 38289  1633 |     a = no
  2747  2542 |     b = yes
```

Εικόνα 6: Αποτελέσματα [A,B] Stratified Cross-Validation μοντέλου ταξινόμησης (J48) με συνολική ακρίβεια 90,31% και παρουσίαση μετρικών ανά κλάση, μαζί με πίνακα σύγχυσης.

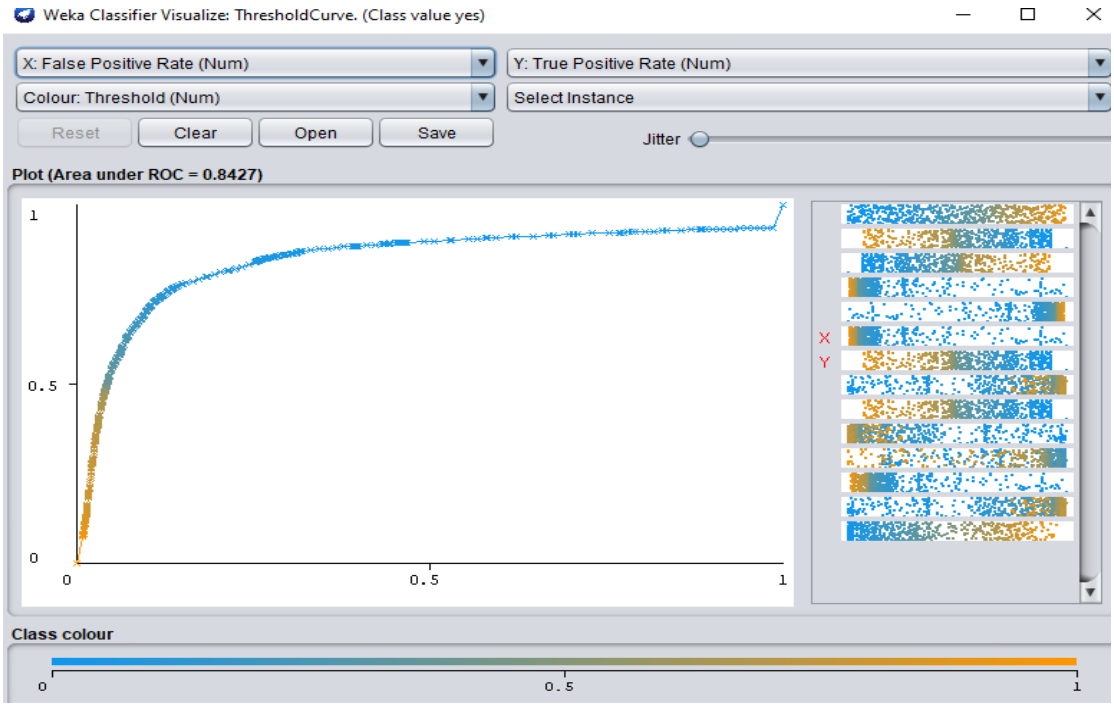
A]



B]



Εικόνα 7: Διαγράμματα [A,B] συσχετίσεων (scatter plot matrix) μεταβλητών με χρωματική απεικόνιση των κλάσεων.



Εικόνα 8: Καμπύλη ROC με AUC = 0,8427 που απεικονίζει την απόδοση του μοντέλου ταξινόμησης (J48).

Παρατηρήσεις

Από την ανάλυση των αποτελεσμάτων προκύπτει ότι η διάρκεια της κλήσης (Duration) αποτελεί το πιο σημαντικό χαρακτηριστικό για την πρόβλεψη της αποδοχής των τραπεζικών προϊόντων. Η αξιολόγηση των μοντέλων πραγματοποιήθηκε με βάση την περιοχή κάτω από την χαρακτηριστική καμπύλη λειτουργίας δέκτη (ROC Area), η οποία υπολογίστηκε στο λογισμικό WEKA και παρουσιάζεται στην ενότητα «Detailed Accuracy By Class» για την κλάση “yes”.

Για καλύτερη κατανόηση της διακριτικής ικανότητας των μοντέλων, χρησιμοποιήθηκε η δυνατότητα «Visualize Threshold Curve» του WEKA, επιτρέποντας την οπτικοποίηση της χαρακτηριστικής καμπύλης λειτουργίας δέκτη «ROC». Η γραφική αυτή αναπαράσταση διευκολύνει τη σύγκριση της απόδοσης των μοντέλων σε διαφορετικά επίπεδα κατωφλίου (thresholds), αποτυπώνοντας με σαφήνεια την ικανότητα του κάθε μοντέλου να ξεχωρίζει σωστά τις θετικές περιπτώσεις από τις αρνητικές.

ΣΥΜΠΕΡΑΣΜΑΤΑ ΠΟΥ ΑΦΟΡΟΥΝ ΤΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΗ ΚΑΜΠΥΛΗ ΛΕΙΤΟΥΡΓΙΑΣ ΔΕΚΤΗ «ROC».

Παρακάτω γίνεται αναφορά στον κανόνα του μοντέλου αναφορικά με τις τιμές τις οποίες μπορεί να λάβει και χρησιμοποιείται προκειμένου να αξιολογηθεί το αποτέλεσμα της καμπύλης ROC.

- ROC κοντά στο **0.5** → τυχαίο μοντέλο
- ROC > **0.7** → καλό
- ROC > **0.8** → πολύ καλό
- ROC > **0.9** → εξαιρετικό

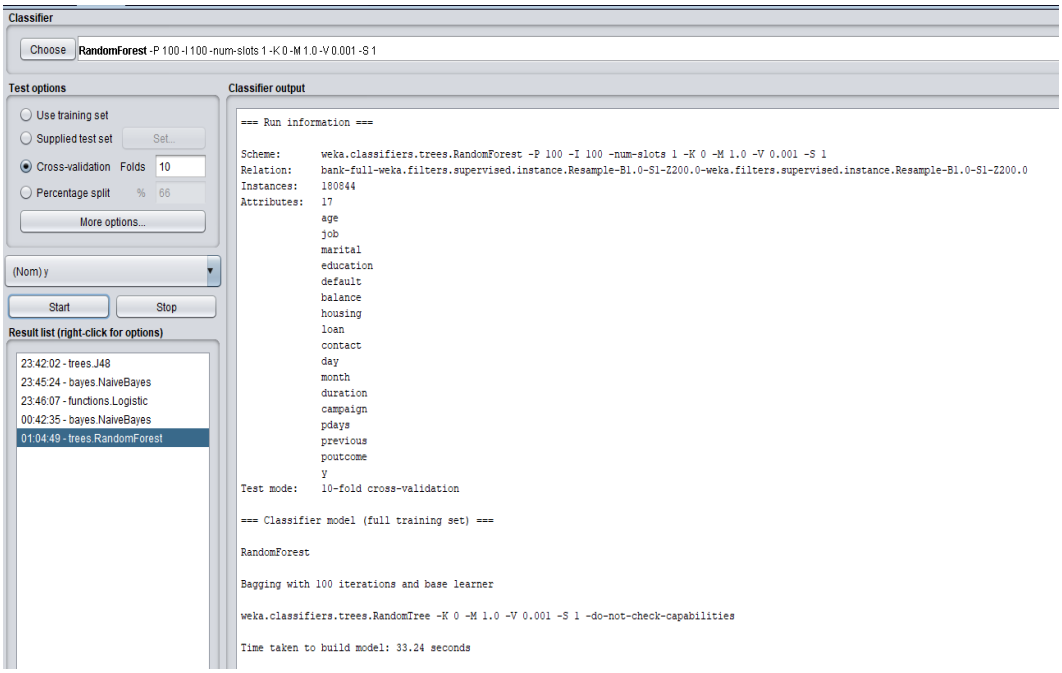
Σύμφωνα με το αποτέλεσμα της οπτικοποίησης της χαρακτηριστικής καμπύλης ROCμεAUC = 0.8427 το μοντέλο κρίνεται ότι εμφανίζει υψηλή ικανότητα διάκρισης μεταξύ θετικών και αρνητικών περιπτώσεων, παρουσιάζει με άλλα λόγια καλή διακριτική ικανότητα.

4.2.3 Αλγόριθμος «Τυχαίο Δάσος» (Random Forest)

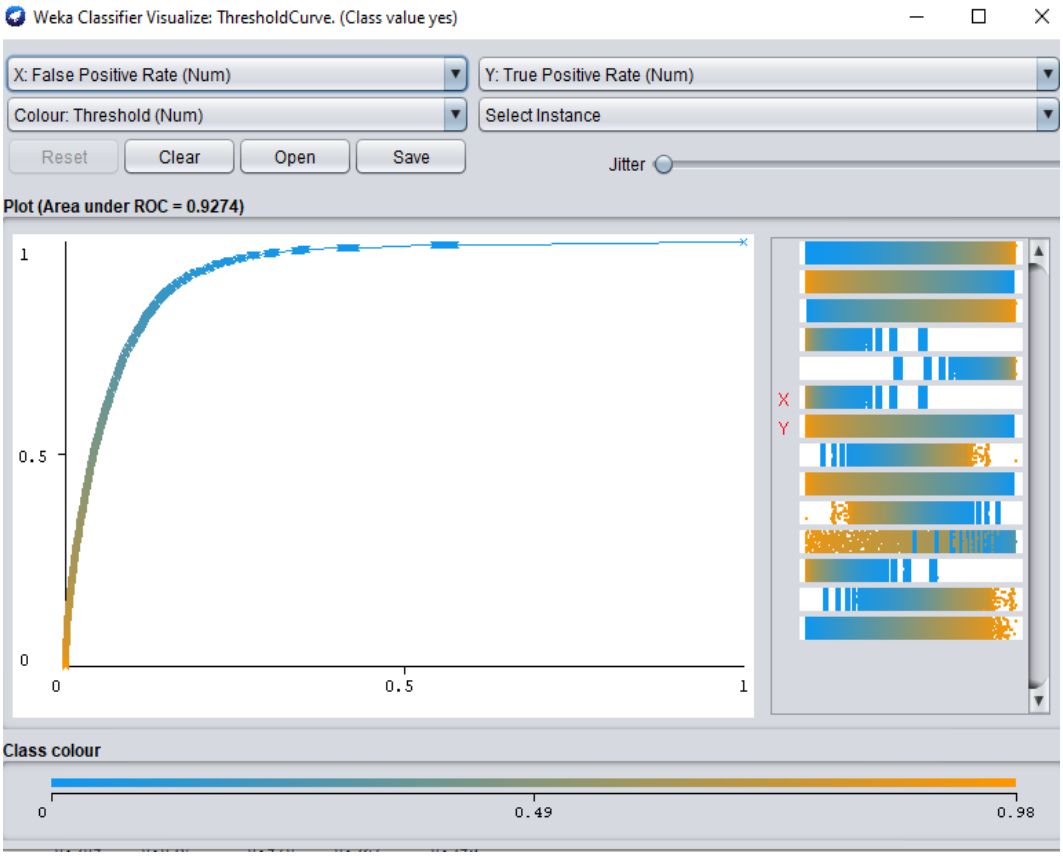
Περιγραφή Αλγορίθμου

Σύμφωνα με τα πειραματικά αποτελέσματα, ο αλγόριθμος «Τυχαίο δάσος» (Random Forest) παρουσίασε υψηλή ακρίβεια και ενισχυμένη ικανότητα αναγνώρισης της μειονοτικής κλάσης, σε σύγκριση με τα μεμονωμένα δέντρα αποφάσεων. Η μέθοδος συνόλου (ensemble method) αξιοποιεί πολλαπλά δέντρα, τα οποία εκπαιδεύονται σε τυχαία υποσύνολα των δεδομένων και των χαρακτηριστικών, ώστε η τελική πρόβλεψη να προκύπτει μέσω πλειοψηφικής ψήφου. Η διαδικασία αυτή αυξάνει τη γενίκευση του μοντέλου και μειώνει τον κίνδυνο υπερπροσαρμογής, ενώ ταυτόχρονα επιτρέπει την αξιολόγηση της σχετικής σημασίας των χαρακτηριστικών για την πρόβλεψη της μεταβλητής-στόχου.

- Classifier → trees → RandomForest
- NumTrees = 100 (default)



Εικόνα 9: Εκπαίδευση μοντέλου «Random Forest» στο WEKA με 10-fold cross-validation.



Εικόνα 10: «Εικόνα 10: Καμπύλη ROC μοντέλου ταξινόμησης Random Forest με 10-fold cross-validation (AUC = 0,9274), η οποία υποδεικνύει υψηλή διακριτική ικανότητα.»

Αποτελέσματα

- Accuracy: 92,7%
- Precision (yes): 0,69
- Recall (yes): 0,58
- F1-score: 0,63

Σύμφωνα με το αποτέλεσμα της οπτικοποίησης της χαρακτηριστικής καμπύλης ROC με AUC = 0.9274 το μοντέλο κρίνεται ότι εμφανίζει υψηλή ικανότητα διάκρισης, παρουσιάζει με άλλα λόγια καλή διακριτική ικανότητα.

Αλγόριθμος	Accuracy	Precision (yes)	Recall (yes)	F1-score
J48	90.31%	0.60	0.48	0.53
Zero R	88.3%	0.88	1	0
Random Forest	**92.7%**	**0.69**	**0.58**	**0.63**

Πίνακας1: Σύγκριση J48, Zero R και Random Forest με Accuracy, Precision, Recall και F1-score.

Παρατήρηση

Σύμφωνα με τα πειραματικά αποτελέσματα, ο αλγόριθμος “Τυχαίο Δάσος” (Random Forest) πέτυχε την υψηλότερη ακρίβεια και τη μεγαλύτερη περιοχή κάτω από την καμπύλη ROC (AUC) σε σύγκριση με τα υπόλοιπα μοντέλα. Η επίδοση αυτή αναδεικνύει την υπεροχή των μεθόδων συνόλου (ensemble methods), οι οποίες συνδυάζουν πολλαπλά δέντρα αποφάσεων για να βελτιώσουν τη γενίκευση, να αυξήσουν τη σταθερότητα των προβλέψεων και να μειώσουν τον κίνδυνο υπερπροσαρμογής (overfitting).

Η υψηλή απόδοση του «Random Forest» (92,7%) καταδεικνύει ότι η χρήση συνόλων μοντέλων αποτελεί αποτελεσματική στρατηγική για προβλέψεις σε περιβάλλοντα με πολύπλοκα και μη ισορροπημένα δεδομένα, όπως αυτά που σχετίζονται με την αποδοχή τραπεζικών προϊόντων. Τα αποτελέσματα επιβεβαιώνουν ότι η συνδυαστική προσέγγιση επιτρέπει στα μοντέλα να αξιοποιούν πλήρως την πληροφορία των δεδομένων και να προσφέρουν πιο αξιόπιστες και σταθερές προβλέψεις.

Ο ZeroR αποτελεί έναν πολύ απλό αλγόριθμο ταξινόμησης που χρησιμοποιείται κυρίως ως baseline, καθώς προβλέπει πάντα την πλειοψηφική κλάση χωρίς να λαμβάνει υπόψη τα χαρακτηριστικά των δεδομένων. Αν και μπορεί να εμφανίζει υψηλή ακρίβεια σε ανισόρροπα σύνολα δεδομένων(88.3%), η απόδοσή του είναι παραπλανητική, εφόσον αποτυγχάνει πλήρως να εντοπίσει τη μειοψηφική κλάση (Recall και F1-score ίσα με 0). Για τον παραπάνω λόγο, ο ZeroR δεν είναι χρήσιμος για πρακτική πρόβλεψη, αλλά μόνο ως σημείο αναφοράς για τη σύγκριση πιο σύνθετων μοντέλων.

Τέλος, ο J48 αποτελεί ένα δέντρο απόφασης, το οποίο δείχνει ικανοποιητική απόδοση, με accuracy 90.31% και σχετικά ισορροπημένα Precision (0.60) και Recall (0.48) για την κλάση yes. Σε αντίθεση με τον ZeroR, καταφέρνει να εντοπίσει μέρος της μειοψηφικής κλάσης, κάτι που φαίνεται και από το F1-score (0.5). Συνολικά, αποτελεί μια καλή βασική επιλογή, αν και υπάρχει περιθώριο βελτίωσης κυρίως στο Recall.

4.2.4 Αλγόριθμος ταξινόμησης μηχανικής μάθησης (Naive Bayes)

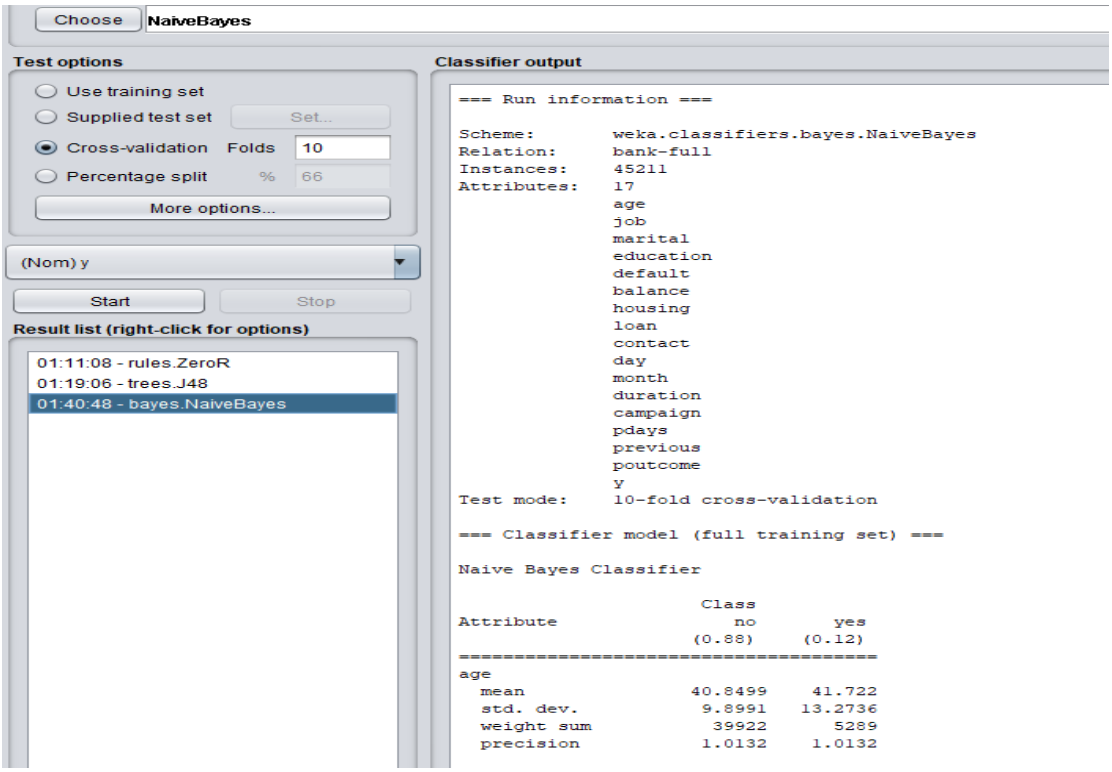
Ο αλγόριθμος «Naive Bayes»αποτελεί έναν αλγόριθμο ταξινόμησης βασισμένο στο θεώρημα του «Bayes», ο οποίος χρησιμοποιεί πιθανοθεωρητικές αρχές για την πρόβλεψη της κλάσης ενός δείγματος. Το μοντέλο προϋποθέτει ότι τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους, γεγονός που απλοποιεί τους υπολογισμούς και επιτρέπει γρήγορη εκπαίδευση και πρόβλεψη, ακόμη και σε μεγάλα σύνολα δεδομένων. Παρά την απλουστευτική αυτή υπόθεση, ο αλγόριθμος

«Naive Bayes» συχνά παρουσιάζει ικανοποιητικές επιδόσεις σε προβλήματα ταξινόμησης.

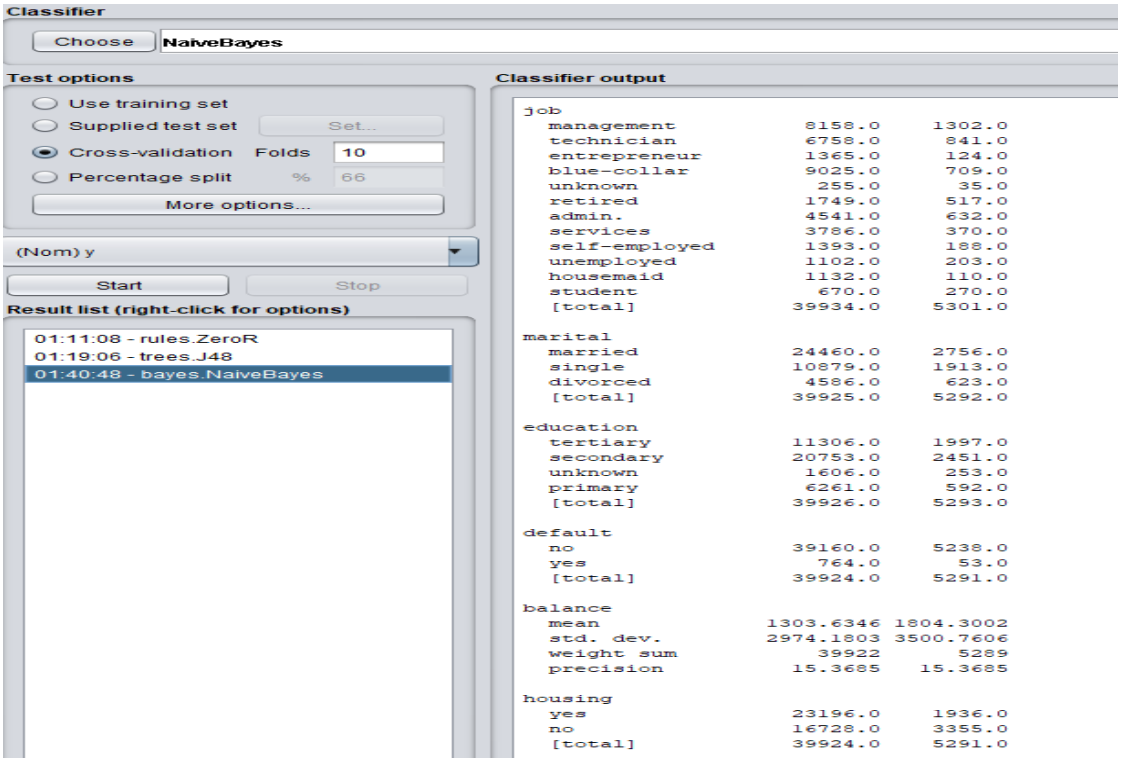
Στο λογισμικό WEKA, ο αλγόριθμος «Naive Bayes» εφαρμόστηκε με τη χρήση της μεθόδου «10-fold Cross Validation (CV)», όπου το σύνολο δεδομένων χωρίζεται σε δέκα ισομεγέθη υποσύνολα. Σε κάθε επανάληψη, ένα υποσύνολο χρησιμοποιείται για έλεγχο ενώ τα υπόλοιπα εννέα για εκπαίδευση. Η διαδικασία αυτή εξασφαλίζει αντικειμενική και σταθερή εκτίμηση της απόδοσης του αλγορίθμου, μειώνοντας την επίδραση τυχαίων διακυμάνσεων στα αποτελέσματα.

- Classifier → bayes → NaiveBayes
- 10-fold Cross Validation

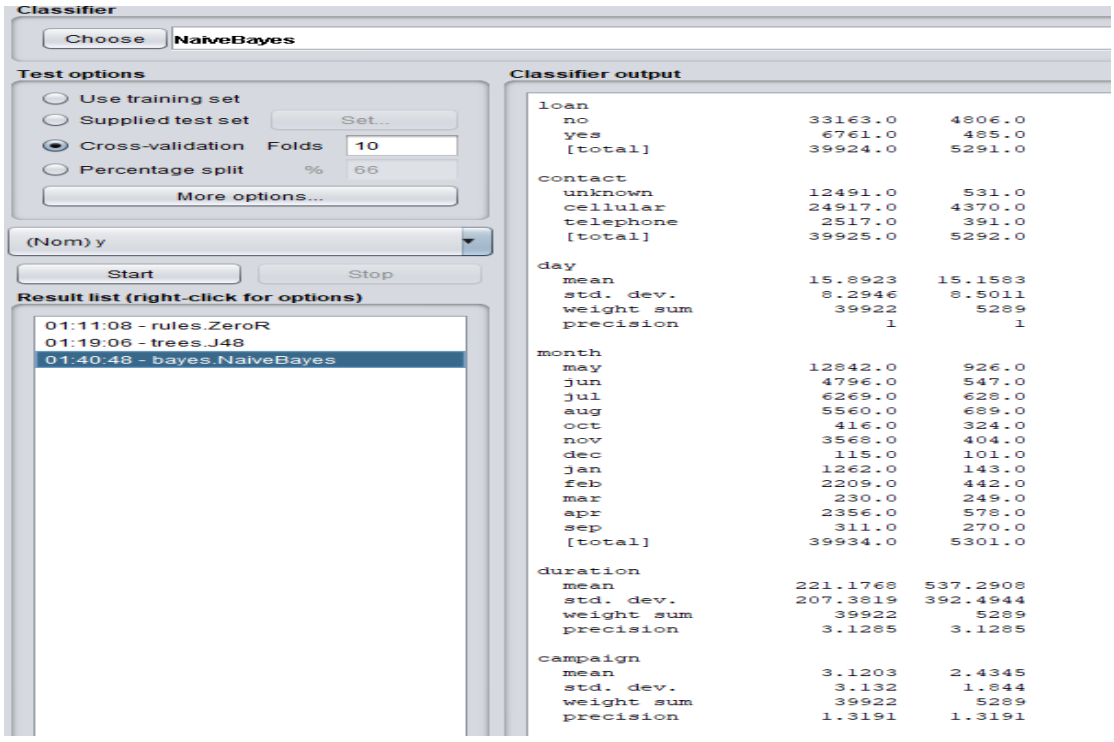
A]



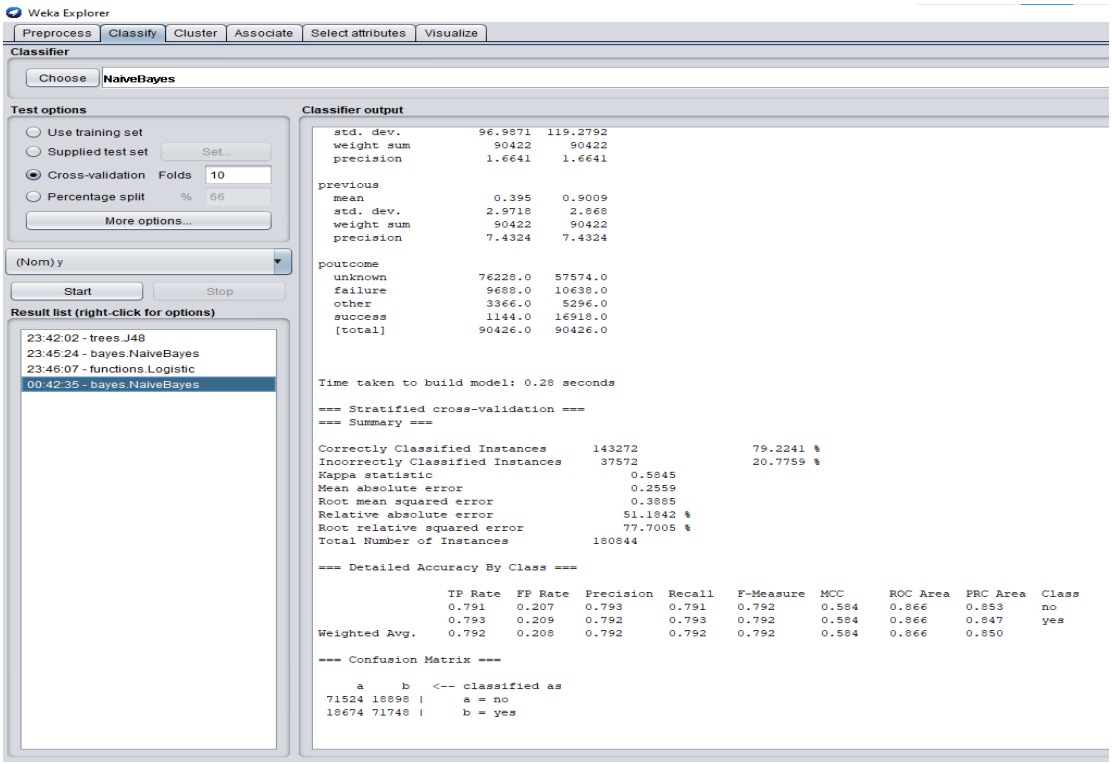
B]



Γ]



Εικόνα 11: Στιγμιότυπα [Α,Β,Γ] από το WEKA με αποτελέσματα ταξινόμησης αλγορίθμου (Naive Bayes) και 10-fold cross-validation στο σύνολο δεδομένων «bank-full».



Εικόνα 12: Detailed Accuracy by Class και Confusion Matrix με απόδοση και συχνότητες προβλέψεων ανά κλάση (NaiveBayes).

Σύγκριση απόδοσης

Η αξιολόγηση της απόδοσης των αλγορίθμων πραγματοποιήθηκε με τη χρήση πολλαπλών μετρικών, όπως είναι η Ακρίβεια (Accuracy), η Ακρίβεια Θετικών Προβλέψεων (Precision), η Ανάκληση ή Ευαισθησία (Recall / Sensitivity), ο Αρμονικός Μέσος (F1-score) και η Καμπύλη Χαρακτηριστικής Λειτουργίας Δέκτη (ROC Curve). Η χρήση αυτών των δεικτών επέτρεψε μια αντικειμενική σύγκριση μεταξύ των αλγορίθμων, λαμβάνοντας υπόψη τόσο τη συνολική απόδοση όσο και την ικανότητα εντοπισμού της μειοψηφικής κατηγορίας.

Συγκεκριμένα, ο αλγόριθμος Naive Bayes στο Weka παρουσιάζει ικανοποιητική απόδοση, με ακρίβεια περίπου (79%) και ισορροπημένες τιμές σε precision, recall και F-measure (0.79), κάτι που δείχνει ότι χειρίζεται εξίσου καλά και τις δύο κλάσεις. Ο δείκτης Kappa (0.58) βρίσκεται σε μέτριο επίπεδο, ενώ η τιμή ROC (0.866) είναι αρκετά υψηλή, υποδηλώνοντας καλή διακριτική ικανότητα. Παρότι

υπάρχουν κάποια σφάλματα, το μοντέλο είναι γρήγορο και αποδοτικό, αν και η απλότητά του περιορίζει την απόδοσή του σε πιο σύνθετα προβλήματα.

4.2.5 Λογιστική Παλινδρόμηση (Logistic Regression)

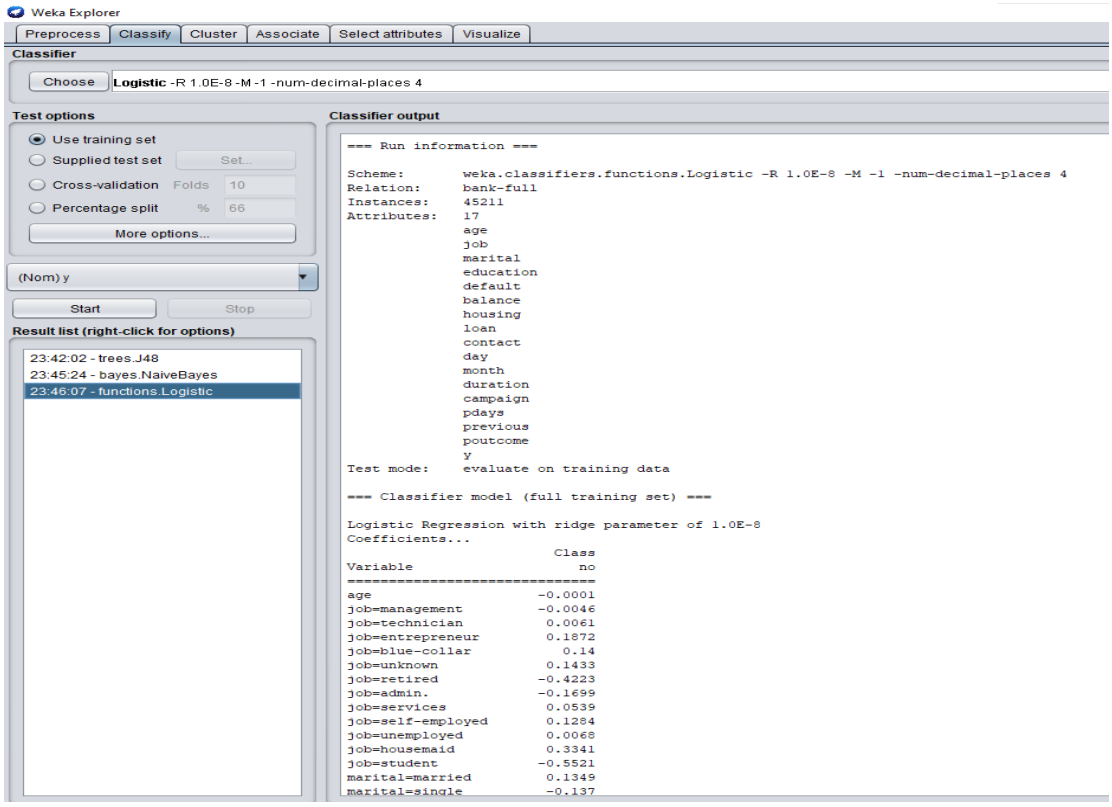
Η λογιστική παλινδρόμηση (Logistic Regression) αποτελεί ένα στατιστικό μοντέλο ταξινόμησης που εκτιμά την πιθανότητα εμφάνισης μιας συγκεκριμένης κλάσης σε διχοτομικά (binary) αποτελέσματα. Η μέθοδος είναι ιδιαίτερα χρήσιμη για προβλέψεις όπως η αποδοχή ενός προϊόντος, καθώς προσφέρει τόσο ερμηνευσιμότητα όσο και αποτελεσματική απόδοση σε γραμμικά διαχωρίσιμα δεδομένα.

Στην πειραματική διαδικασία που υλοποιήθηκε στο WEKA, η αξιολόγηση βασίστηκε στη μέθοδο «10-fold Cross Validation», κατά την οποία το σύνολο δεδομένων χωρίζεται σε 10 ισομεγέθη υποσύνολα. Το μοντέλο εκπαιδεύεται 10 φορές, χρησιμοποιώντας κάθε φορά 9 υποσύνολα για εκπαίδευση και το 1 για έλεγχο. Αυτή η διαδικασία διασφαλίζει ότι όλα τα δεδομένα αξιοποιούνται τόσο για εκπαίδευση όσο και για αξιολόγηση, μειώνοντας την επίδραση τυχαίων διακυμάνσεων και προσφέροντας μια πιο σταθερή και αξιόπιστη εκτίμηση απόδοσης.

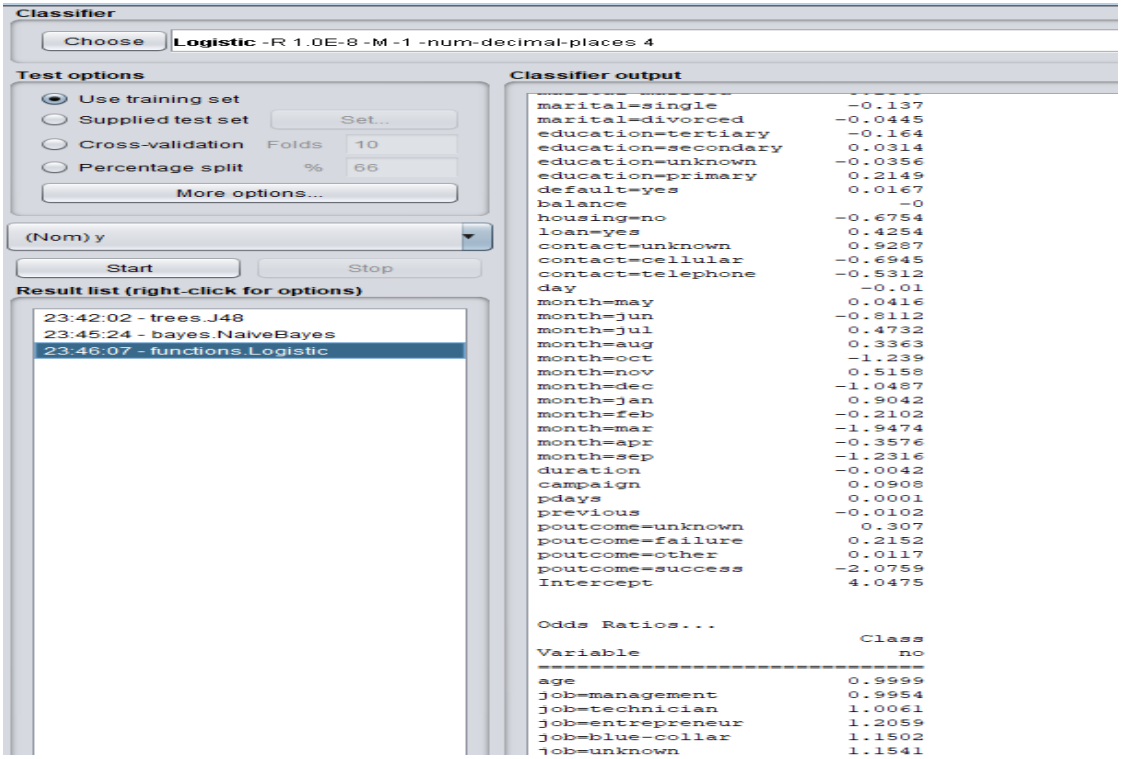
Παράλληλα, εφαρμόστηκε ρύθμιση παραμέτρων (parameter tuning) για τη βελτιστοποίηση των υπερπαραμέτρων του μοντέλου. Η σωστή επιλογή των παραμέτρων ενισχύει την ικανότητα γενίκευσης και αποτρέπει το φαινόμενο της υπερπροσαρμογής (overfitting), εξασφαλίζοντας ότι το μοντέλο δεν "μαθαίνει" μόνο τα δεδομένα εκπαίδευσης, αλλά μπορεί να αποδώσει αξιόπιστα και σε νέα, άγνωστα δεδομένα.

Συνολικά, ο συνδυασμός διασταυρωμένης επικύρωσης και βελτιστοποίησης παραμέτρων ενισχύει την αξιοπιστία των αποτελεσμάτων, παρέχοντας μοντέλα τα οποία είναι ταυτόχρονα αποτελεσματικά και ικανά να γενικεύουν σε πραγματικές συνθήκες.

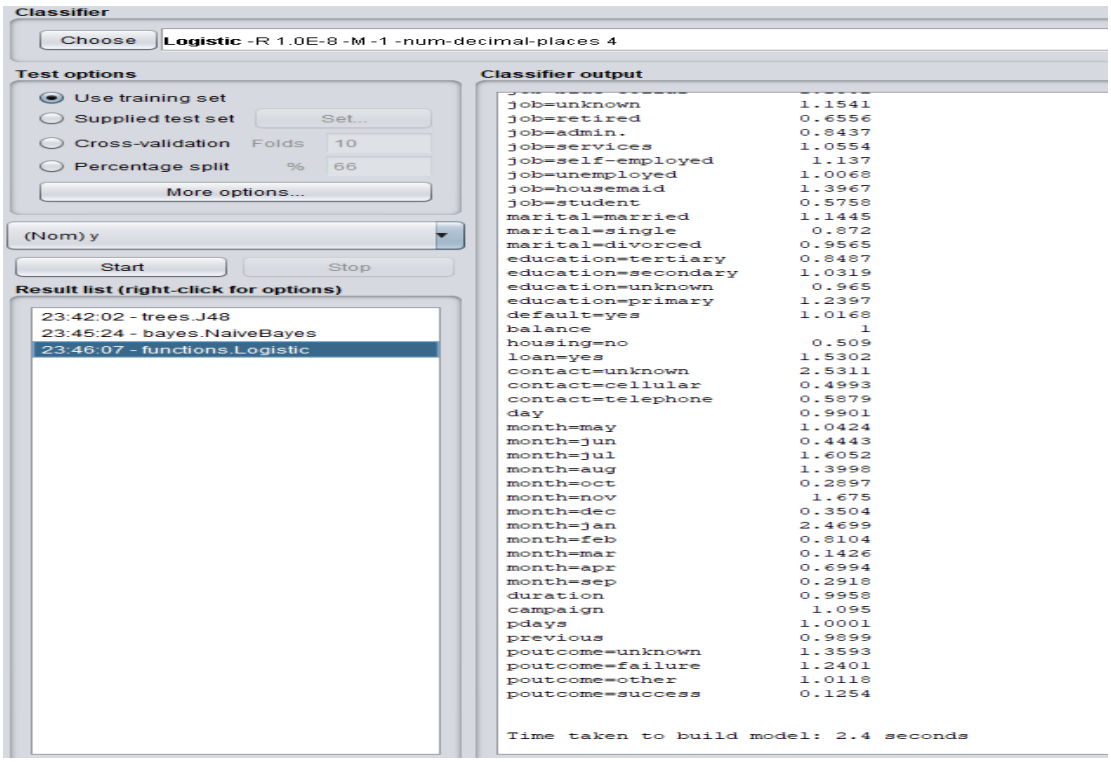
A]


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) y
Start
Stop
Result list (right-click for options)
23:42:02 - trees.J48
23:45:24 - bayes.NaiveBayes
23:46:07 - functions.Logistic
Classifier output
=== Run information ===
Scheme: weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Relation: bank-full
Instances: 45211
Attributes: 17
age
job
marital
education
default
balance
housing
loan
contact
day
month
duration
campaign
pdays
previous
poutcome
y
Test mode: evaluate on training data
=== Classifier model (full training set) ===
Logistic Regression with ridge parameter of 1.0E-8
Coefficients...
Class
Variable no
age -0.0001
job=management -0.0046
job=technician 0.0061
job=entrepreneur 0.1872
job=blue-collar 0.14
job=unknown 0.1433
job=retired -0.4223
job=admin. -0.1699
job=services 0.0539
job=self-employed 0.1284
job=unemployed 0.0068
job=housemaid 0.3341
job=student -0.5521
marital=married 0.1349
marital=single -0.137

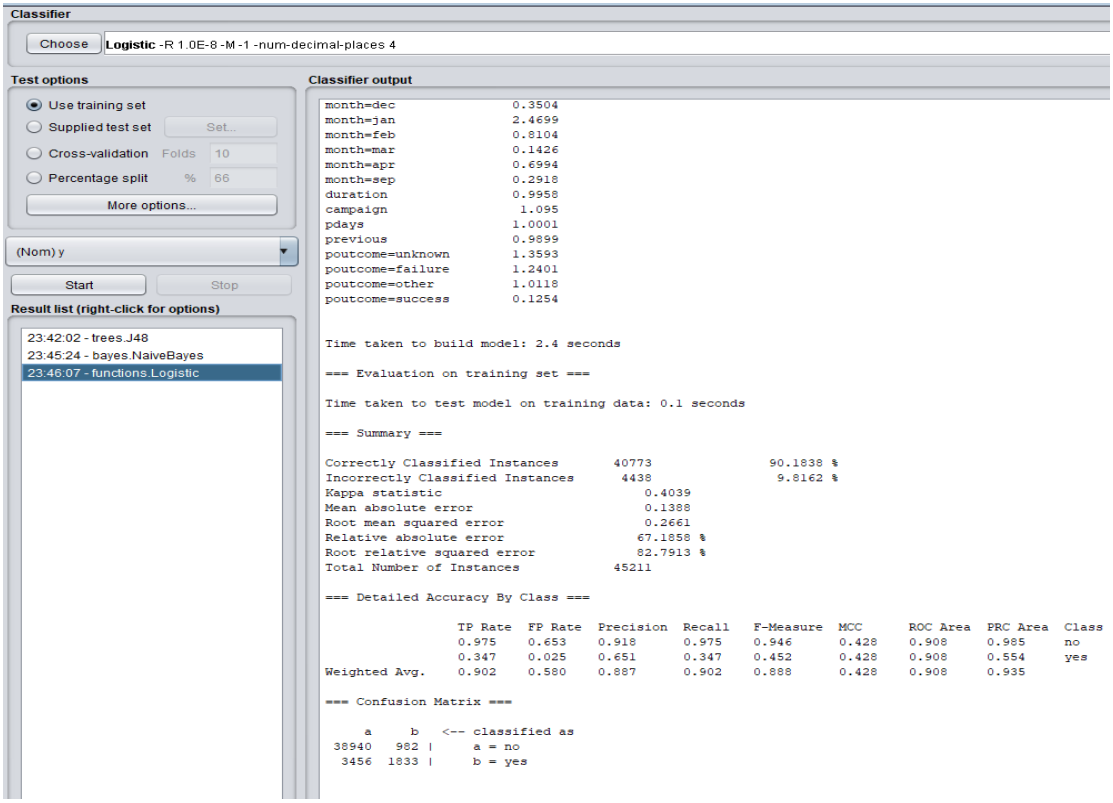
В]


Classifier
Choose
Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) y
Start
Stop
Result list (right-click for options)
23:42:02 - trees.J48
23:45:24 - bayes.NaiveBayes
23:46:07 - functions.Logistic
Classifier output
marital=single -0.137
marital=divorced -0.0445
education=tertiary -0.164
education=secondary 0.0314
education=unknown -0.0356
education=primary 0.2149
default=yes 0.0167
balance -0
housing=no -0.6754
loan=yes 0.4254
contact=unknown 0.9287
contact=cellular -0.6945
contact=telephone -0.5312
day -0.01
month=may 0.0416
month=jun -0.8112
month=jul 0.4732
month=aug 0.3363
month=oct -1.239
month=nov 0.5158
month=dec -1.0487
month=jan 0.9042
month=feb -0.2102
month=mar -1.9474
month=apr -0.3576
month=sep -1.2316
duration -0.0042
campaign -0.0908
pdays 0.0001
previous -0.0102
poutcome=unknown 0.307
poutcome=failure 0.2152
poutcome=other 0.0117
poutcome=success -2.0759
Intercept 4.0475
Odds Ratios...
Class
Variable no
age 0.9999
job=management 0.9954
job=technician 1.0061
job=entrepreneur 1.2059
job=blue-collar 1.1502
job=unknown 1.1541

Г]


Classifier
Choose
Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) y
Start
Stop
Result list (right-click for options)
23:42:02 - trees.J48
23:45:24 - bayes.NaiveBayes
23:46:07 - functions.Logistic
Classifier output
job=unknown 1.1541
job=retired 0.6556
job=admin. 0.8437
job=services 1.0554
job=self-employed 1.137
job=unemployed 1.0068
job=housemaid 1.3967
job=student 0.5758
marital=married 1.1445
marital=single 0.872
marital=divorced 0.9565
education=tertiary 0.8487
education=secondary 1.0319
education=unknown 0.965
education=primary 1.2397
default=yes 1.0168
balance 1
housing=no 0.509
loan=yes 1.5302
contact=unknown 2.5311
contact=cellular 0.4993
contact=telephone 0.5879
day 0.9901
month=may 1.0424
month=jun 0.4443
month=jul 1.6052
month=aug 1.3998
month=oct 0.2897
month=nov 1.675
month=dec 0.3504
month=jan 2.4699
month=feb 0.8104
month=mar 0.1426
month=apr 0.6994
month=sep 0.2918
duration 0.9958
campaign 1.095
pdays 1.0001
previous 0.9899
poutcome=unknown 1.3593
poutcome=failure 1.2401
poutcome=other 1.0118
poutcome=success 0.1254
Time taken to build model: 2.4 seconds

Δ]



Εικόνα 13: Ρυθμίσεις και αποτελέσματα [Α,Β,Γ,Δ] του αλγορίθμου (LogisticRegression) στο WEKA

Ανάλυση αποτελεσμάτων και σύγκριση

Τα πειραματικά αποτελέσματα υποδεικνύουν ότι η χρήση της διαδικασίας αυτής, με προεπεξεργασία δεδομένων και ορθές ρυθμίσεις αλγορίθμων, εξασφαλίζει δίκαιη και συγκρίσιμη αξιολόγηση των μοντέλων.Οαλγόριθμος (Logistic Regression) εμφανίζει καλή συνολική απόδοση με accuracy = 90.18%**,** ξεπερνώντας τον ZeroR και τον J48, αλλά υστερεί λίγο από το Random Forest. Αυτό επιτρέπει την εξαγωγή αξιόπιστων συμπερασμάτων σχετικά με την απόδοση και την καταλληλότητα των διαφόρων αλγορίθμων για εφαρμογές στον τραπεζικό τομέα, όπως η πρόβλεψη της απόκρισης πελατών σε καμπάνιες «marketing» ή η αναγνώριση κρίσιμων υποομάδων πελατών. Η συστηματική αξιολόγηση με τεχνικές όπως η «10-fold cross-validation» διασφαλίζει ότι οι συγκρίσεις δεν επηρεάζονται από τυχαίους

διαχωρισμούς των δεδομένων, αυξάνοντας την αξιοπιστία και τη γενικευσιμότητα των συμπερασμάτων.

4.2.6 Διαδοχική Ελάχιστη Βελτιστοποίηση (SMO, Sequential Minimal Optimization)

Περιγραφή αλγορίθμου

Η Διαδοχική Ελάχιστη Βελτιστοποίηση (Sequential Minimal Optimization, SMO) αποτελεί μια μέθοδο εκπαίδευσης των Μηχανών Διανυσμάτων Υποστήριξης (Support Vector Machines, SVM)για διχοτομικά προβλήματα. Η εκπαίδευση των μηχανών διανυσμάτων υποστήριξης στοχεύει στον εντοπισμό ενός υπερεπιπέδου (hyperplane) που διαχωρίζει τις δύο κλάσεις με μέγιστο περιθώριο (margin)**.** Όταν τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα, χρησιμοποιούνται συναρτήσεις πυρήνα (kernel functions), όπως ο «Radial Basis Function» (RBF) ή Γκαουσιανός πυρήνας (Gaussian), οι οποίες χαρτογραφούν τα δεδομένα σε έναν χώρο υψηλότερων διαστάσεων όπου ο διαχωρισμός είναι εφικτός. Το μοντέλο αυτό φημίζεται για την υψηλή ακρίβεια, αν και η ερμηνεία του είναι δυσκολότερη σε σχέση με πιο διαφανείς ταξινομητές, όπως τα δέντρα αποφάσεων.

Για την επίτευξη μέγιστης ακρίβειας, η διαδικασία εκπαίδευσης περιλαμβάνει δύο βασικά στάδια: προεπεξεργασία και ρύθμιση παραμέτρων. Στο στάδιο προεπεξεργασίας, εφαρμόζεται στο WEKA το φίλτρο «Normalize», το οποίο μετατρέπει όλες τις αριθμητικές τιμές των χαρακτηριστικών στην κλίμακα [0, 1]. Αυτό διασφαλίζει ότι κανένα χαρακτηριστικό με μεγαλύτερη κλίμακα μέτρησης δεν κυριαρχεί άδικα στο μοντέλο, βελτιώνοντας τόσο την ταχύτητα όσο και την ευστάθεια της εκπαίδευσης. Με αυτόν τον τρόπο, η SMO μπορεί να εκπαιδευτεί αποτελεσματικά και να παρέχει αξιόπιστες προβλέψεις για μη γραμμικά διαχωρίσιμα δεδομένα.

- Preprocess → Filter → Normalize → Apply

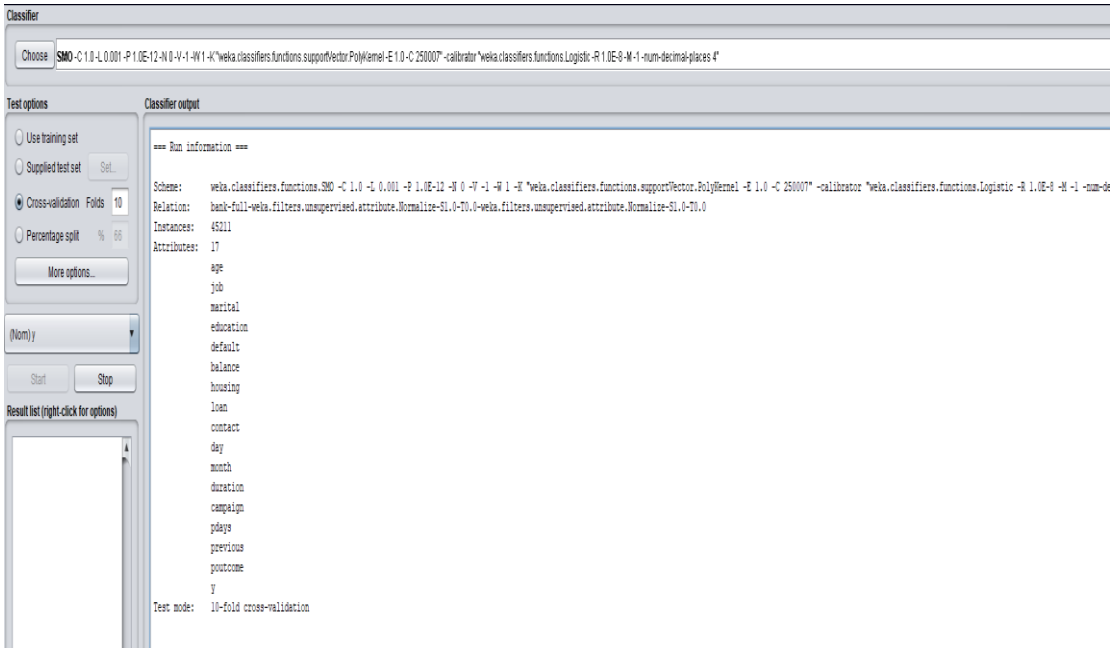
Εκπαίδευση

- Classify → functions → SMO

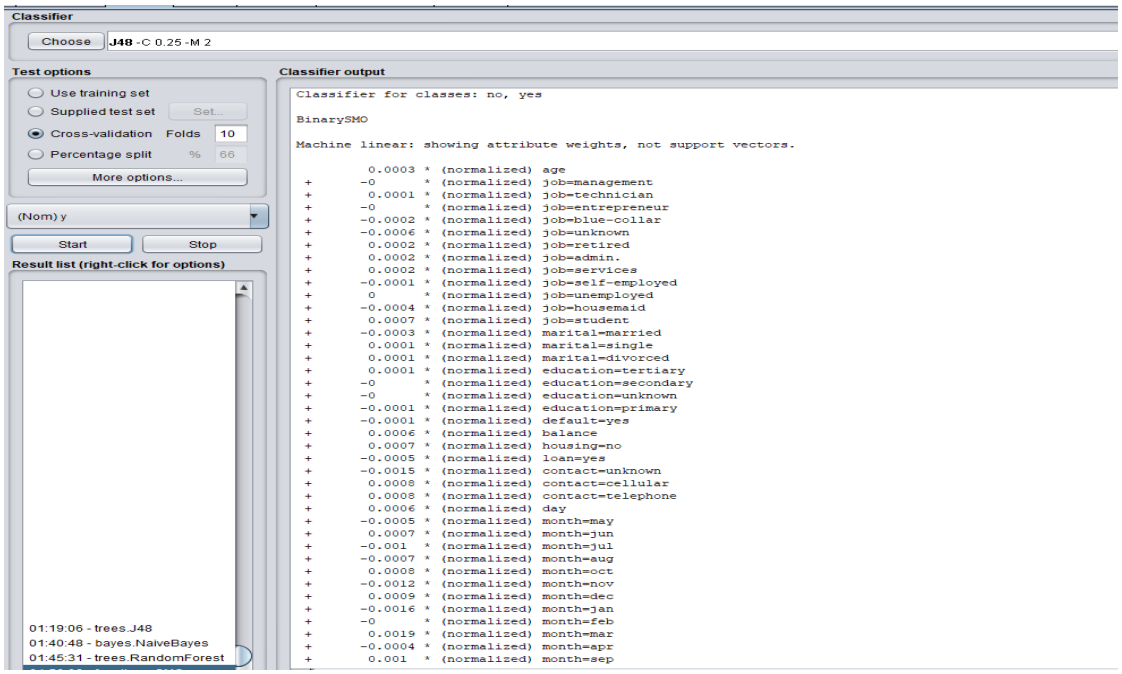
- Kernel: RBF
- 10-fold Cross Validation → Start

Ο αλγόριθμος SMO (Support Vector Machine) αποτελεί έναν από τους ισχυρότερους ταξινομητές, καθώς επιδιώκει τον εντοπισμό του βέλτιστου ορίου διαχωρισμού μεταξύ των κλάσεων. Η χρήση του «RBF Kernel» (Radial Basis Function) επιτρέπει την αντιμετώπιση μη γραμμικά διαχωρίσιμων δεδομένων, χαρτογραφώντας τα σε έναν χώρο υψηλότερων διαστάσεων όπου ο διαχωρισμός γίνεται εφικτός. Για την αξιολόγηση του μοντέλου χρησιμοποιείται η μέθοδος «10-fold cross validation», κατά την οποία το σύνολο δεδομένων χωρίζεται σε 10 ισομεγέθη τμήματα. Κάθε τμήμα χρησιμοποιείται μία φορά ως σύνολο ελέγχου, ενώ τα υπόλοιπα 9 τμήματα χρησιμοποιούνται για εκπαίδευση, εξασφαλίζοντας ότι το μοντέλο δοκιμάζεται σε "άγνωστα" δεδομένα και μειώνοντας τον κίνδυνο υπερπροσαρμογής (overfitting).

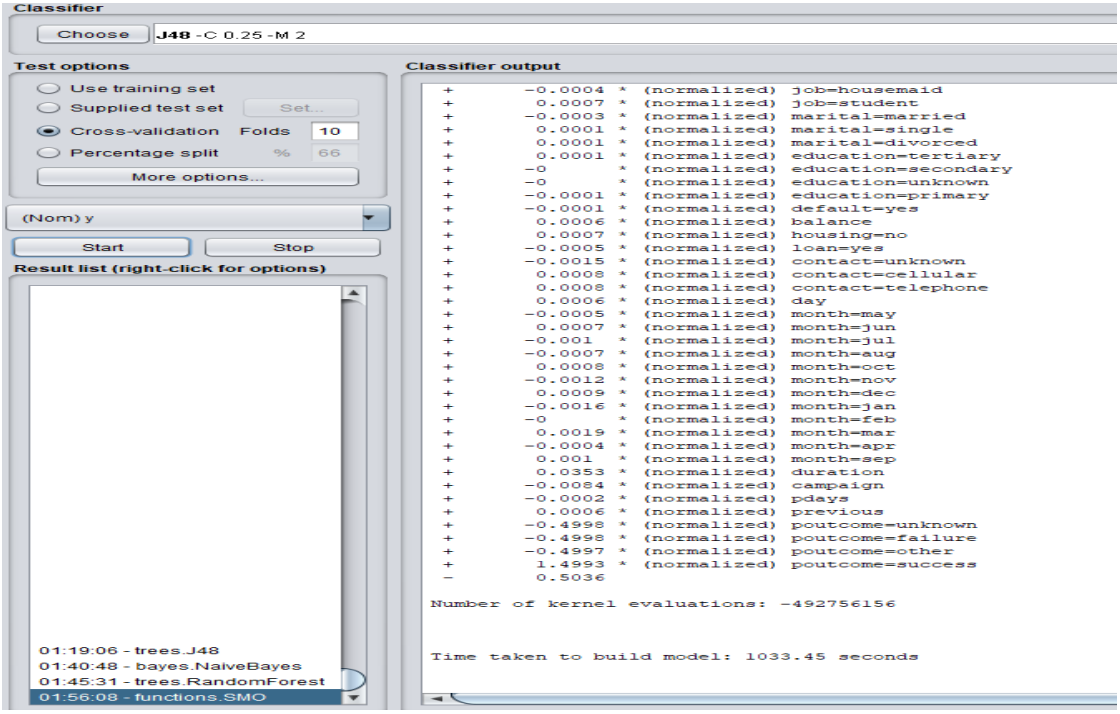
A]



B]


Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) y
Start
Stop
Result list (right-click for options)
01:19:06 - trees.J48
01:40:48 - bayes.NaiveBayes
01:45:31 - trees.RandomForest
Classifier output
Classifier for classes: no, yes
BinarySMO
Machine linear: showing attribute weights, not support vectors.
0.0003 * (normalized) age
+ -0 * (normalized) job=management
+ 0.0001 * (normalized) job=technician
+ -0 * (normalized) job=entrepreneur
+ -0.0002 * (normalized) job=blue-collar
+ -0.0006 * (normalized) job=unknown
+ 0.0002 * (normalized) job=retired
+ 0.0002 * (normalized) job=admin.
+ 0.0002 * (normalized) job=services
+ -0.0001 * (normalized) job=self-employed
+ 0 * (normalized) job=unemployed
+ -0.0004 * (normalized) job=housemaid
+ 0.0007 * (normalized) job=student
+ -0.0003 * (normalized) marital=married
+ 0.0001 * (normalized) marital=single
+ 0.0001 * (normalized) marital=divorced
+ 0.0001 * (normalized) education=tertiary
+ -0 * (normalized) education=secondary
+ -0 * (normalized) education=unknown
+ -0.0001 * (normalized) education=primary
+ -0.0001 * (normalized) default=yes
+ 0.0006 * (normalized) balance
+ 0.0007 * (normalized) housing=no
+ -0.0005 * (normalized) loan=yes
+ -0.0015 * (normalized) contact=unknown
+ 0.0008 * (normalized) contact=cellular
+ 0.0008 * (normalized) contact=telephone
+ 0.0006 * (normalized) day
+ -0.0005 * (normalized) month=may
+ 0.0007 * (normalized) month=jun
+ -0.001 * (normalized) month=jul
+ -0.0007 * (normalized) month=aug
+ 0.0008 * (normalized) month=oct
+ -0.0012 * (normalized) month=nov
+ 0.0009 * (normalized) month=dec
+ -0.0016 * (normalized) month=jan
+ -0 * (normalized) month=feb
+ 0.0019 * (normalized) month=mar
+ -0.0004 * (normalized) month=apr
+ 0.001 * (normalized) month=sep

Γ]



Εικόνα 14: Στιγμιότυπα [Α,Β,Γ] με απεικόνιση λειτουργίας SMO για SVM με εύρεση βέλτιστων διαχωριστικών υπερεπιφανειών.

Αποτελέσματα και παρατηρήσεις

Ο αλγόριθμος Διαδοχικής Ελάχιστης Βελτιστοποίησης (SMO) επιτυγχάνει υψηλή ακρίβεια στην ταξινόμηση, χάρη στην ικανότητά του να βρίσκει αποτελεσματικά τα βέλτιστα όρια μεταξύ των κλάσεων. Παρά την υψηλή απόδοση, ο αλγόριθμος αυτός υστερεί σε ερμηνευσιμότητα, καθώς οι διαδικασίες που ακολουθεί για την κατασκευή των υποστηρικτικών διανυσμάτων δεν είναι εύκολα κατανοητές. Αντίθετα, τα δέντρα απόφασης προσφέρουν σαφή και κατανοητά μονοπάτια λήψης αποφάσεων, επιτρέποντας την απλή ερμηνεία των κανόνων που καθοδηγούν τις προβλέψεις, γεγονός ιδιαίτερα χρήσιμο σε εφαρμογές όπου η διαφάνεια είναι κρίσιμη.

Αλγόριθμος	Accuracy	Precision (yes)	Recall (yes)	F1-score
J48	90.31%	0.60	0.48	0.53
Naïve Bayes	82%	0.80	0,84	0,82
Random Forest	**92.7%**	**0.69**	**0.58**	**0.63**
Logistic Regression	90,18%	0.65	0.35	0.45
SMO	88%	0.87	0.88	0.88

Πίνακας 2: Σύγκριση επιδόσεων J48, Naive Bayes, Random Forest, Logistic Regressionκαι SMO με Accuracy, Precision, Recall (yes) και F1-score.

4.2.7 Αλγόριθμος μηχανικής μάθησης Επιλογής Χαρακτηριστικών (FeatureSelection)

Περιγραφή αλγορίθμου

Η διαδικασία «Feature Selection» αποτελεί κρίσιμο βήμα στη μηχανική μάθηση, καθώς επιτρέπει την επιλογή των πιο σημαντικών χαρακτηριστικών ενός συνόλου δεδομένων. Στόχος της είναι να αυξήσει την ακρίβεια του μοντέλου, μειώνοντας ταυτόχρονα το υπολογιστικό κόστος και τον κίνδυνο υπερπροσαρμογής (overfitting). Αφαιρώντας τα άσχετα ή θορυβώδη χαρακτηριστικά, το μοντέλο γίνεται πιο απλό, κατανοητό και αποδοτικό, ενώ η εκπαίδευση και η αξιολόγηση του μοντέλου γίνεται ταχύτερη και πιο σταθερή.

Αυτό σημαίνει ότι η επιλογή των κατάλληλων χαρακτηριστικών δεν βελτιώνει μόνο την απόδοση του μοντέλου, αλλά και την ερμηνευσιμότητά του, καθιστώντας τα αποτελέσματα πιο εύκολα κατανοητά και εφαρμόσιμα σε πραγματικές επιχειρησιακές αποφάσεις.

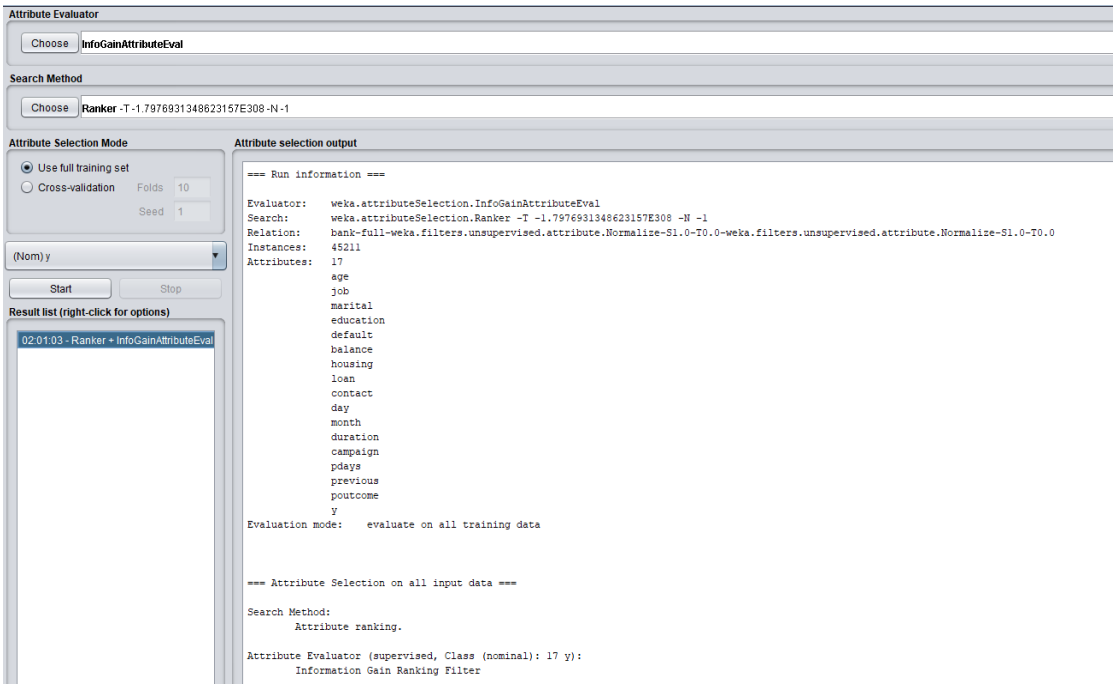
Ρυθμίσεις WEKA

Στις ρυθμίσεις του WEKA, επιλέγονται χακτηριστικά (Feature Selection) τα οποία αποσκοπούν στην ελαχιστοποίηση της διαστασιμότητας του συνόλου δεδομένων. Με αυτό τον τρόπο διατηρούνται μόνο οι πιο «ενημερωτικές» μεταβλητές. Με τη χρήση του αξιολογητή «InfoGainAttributeEval**»** επιτρέπεται ο υπολογισμός του οφέλους πληροφορίας που αφορά κάθε χαρακτηριστικό ξεχωριστά και έτσι προσδιορίζεται ποια από αυτά τα χαρακτηριστικά συνεισφέρουν περισσότερο στην πρόβλεψη. Στη συνέχεια, με τη μέθοδο «Ranker» ταξινομούνται τα χαρακτηριστικά από το πιο σημαντικό στο λιγότερο σημαντικό. Έτσι διευκολύνεται η επιλογή των πλέον κρίσιμων μεταβλητών.

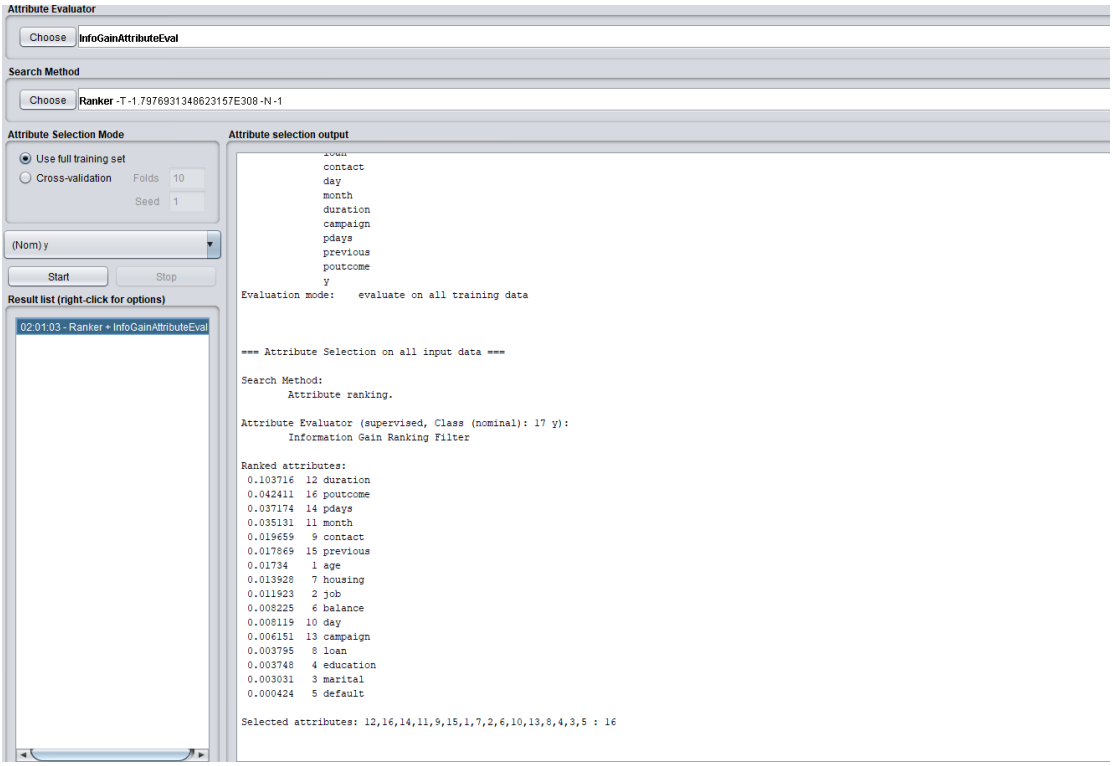
Με τις παραπάνω ρυθμίσεις δημιουργούνται πιο απλά και ταχύτερα μοντέλα μηχανικής μάθησης**,** καθώς αφαιρείται η «ηχορύπανση», η οποία προέρχεται από χαρακτηριστικά, τα οποία δεν προσφέρουν ουσιαστική πληροφορία. Έτσι ενισχύεται τόσο η ακρίβεια όσο και η ερμηνευσιμότητα των μοντέλων και παράλληλα διατηρείται η αποτελεσματικότητα της διαδικασίας εκπαίδευσης.

- Select Attributes
- InfoGainAttributeEval
- Ranker → Start

A]



B]



Εικόνα 15: Feature Selection [A,B] με απεικόνιση των σημαντικότερων χαρακτηριστικών για την απόδοση του μοντέλου.

Παρατηρήσεις

Οι παρατηρήσεις δείχνουν ότι η εφαρμογή της μεθόδου επιλογής χαρακτηριστικών «InfoGain»βοήθησε στη σημαντική μείωση της πολυπλοκότητας των μοντέλων, χωρίς να θυσιαστεί σημαντικά η απόδοσή τους. Αυτό επιτρέπει στα μοντέλα να είναι πιο απλά, ταχύτερα στην εκπαίδευση και πιο εύκολα στην ερμηνεία των αποτελεσμάτων. Από τηνπαραπάνω ανάλυση παρατηρείται ότι τα σημαντικότερα είναι:

- duration (μακράν το πιο ισχυρό)
- poutcome (αποτέλεσμα προηγούμενης καμπάνιας)
- day, month
- contact
- housing, loan

Γενικά Συμπεράσματα

Από τα πειραματικά δεδομένα προκύπτει ότι ο αλγόριθμος «Τυχαίο Δάσος»(Random Forest) παρέχει την υψηλότερη συνολική απόδοση, χάρη στη δυνατότητά του να συνδυάζει τα αποτελέσματα πολλαπλών δέντρων και να μειώνει τα φαινόμενα υπερπροσαρμογής. Αντίθετα, ο αλγόριθμος «J48» υπερέχει σε όρους ερμηνευσιμότητας, καθώς προσφέρει σαφείς κανόνες απόφασης, γεγονός χρήσιμο για τη λήψη επιχειρησιακών αποφάσεων. Επιπλέον, η χρήση τεχνικών επιλογής χαρακτηριστικών αποδεικνύεται καθοριστική για τη διαχείριση της πολυπλοκότητας, διατηρώντας την ακρίβεια σε υψηλά επίπεδα. Συνολικά, τα ευρήματα επιβεβαιώνουν την αποτελεσματικότητα των μεθόδων συνόλου (ensemble method) σε εφαρμογές τραπεζικού «marketing», τόσο από πλευράς απόδοσης όσο και ερμηνευσιμότητας.

4.3 Αλγόριθμοι και Πειράματα Ομαδοποίησης (ClusteringAlgorithms and Experiments)

4.3.1 Προεπεξεργασία δεδομένων για Clustering (Data Preprocessing for Clustering)

Πριν από την εκτέλεση των πειραμάτων ομαδοποίησης (clustering), αφαιρέθηκε η μεταβλητή-στόχος y από το σύνολο δεδομένων (dataset). Η αφαίρεση αυτή ήταν απαραίτητη για να διασφαλιστεί ότι οι μέθοδοι ομαδοποίησης θα σχηματίσουν τις ομάδες με βάση τα υπόλοιπα χαρακτηριστικά, χωρίς να καθοδηγούνται από την πληροφορία της κλάσης. Με αυτόν τον τρόπο, επιτυγχάνεται μια πιο αμερόληπτη και ουσιαστική εξερεύνηση της δομής των δεδομένων, καθώς τα αποτελέσματα του «clustering» αντικατοπτρίζουν τις φυσικές ομοιότητες και διαφορές μεταξύ των δειγμάτων.

4.3.2 Αλγόριθμος Κ-μέσων (K-Means)

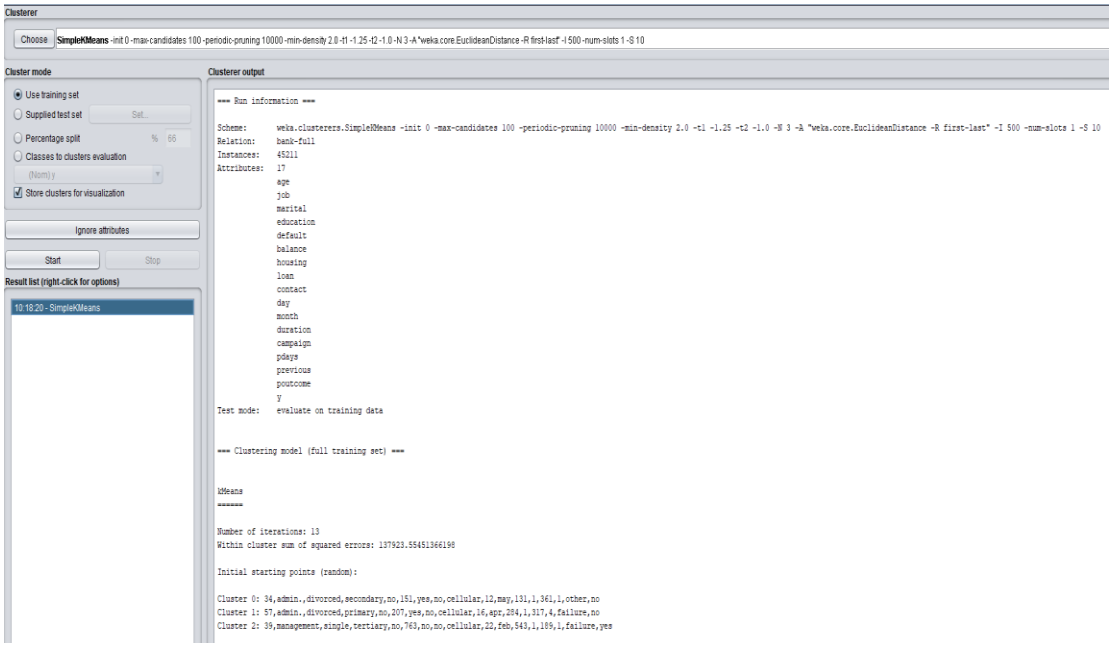
Περιγραφή αλγορίθμου

Ο αλγόριθμος «Κ-μέσων» (K-Means) αποτελεί έναν αλγόριθμο μη επιβλεπόμενης μάθησης (unsupervised learning), ο οποίος διαχωρίζει τα δεδομένα σε k ομάδες (clusters)βάσει της ομοιότητας των χαρακτηριστικών τους. Η διαδικασία ξεκινά με τυχαία κέντρα (centroids) και εκτελεί επαναληπτικά δύο βασικά βήματα μέχρι να επιτευχθεί σύγκλιση. Αρχικά, κάθε δείγμα ανατίθεται στη συστάδα με το πλησιέστερο κέντρο. Στη συνέχεια, τα κέντρα επαναϋπολογίζονται ως ο μέσος όρος των δειγμάτων που ανήκουν σε κάθε συστάδα. Ο στόχος του αλγορίθμου είναι η ελαχιστοποίηση της ενδο-ομαδικής διασποράς (sum of squared distances από τα centroids), ώστε οι σχηματιζόμενες ομάδες να αντιπροσωπεύουν όσο το δυνατόν καλύτερα τις φυσικές δομές των δεδομένων. Πρόκειται για μια γρήγορη, απλή και ευρέως χρησιμοποιούμενη μέθοδο, ιδανική για την εξερεύνηση και ομαδοποίηση μεγάλων συνόλων δεδομένων (datasets).

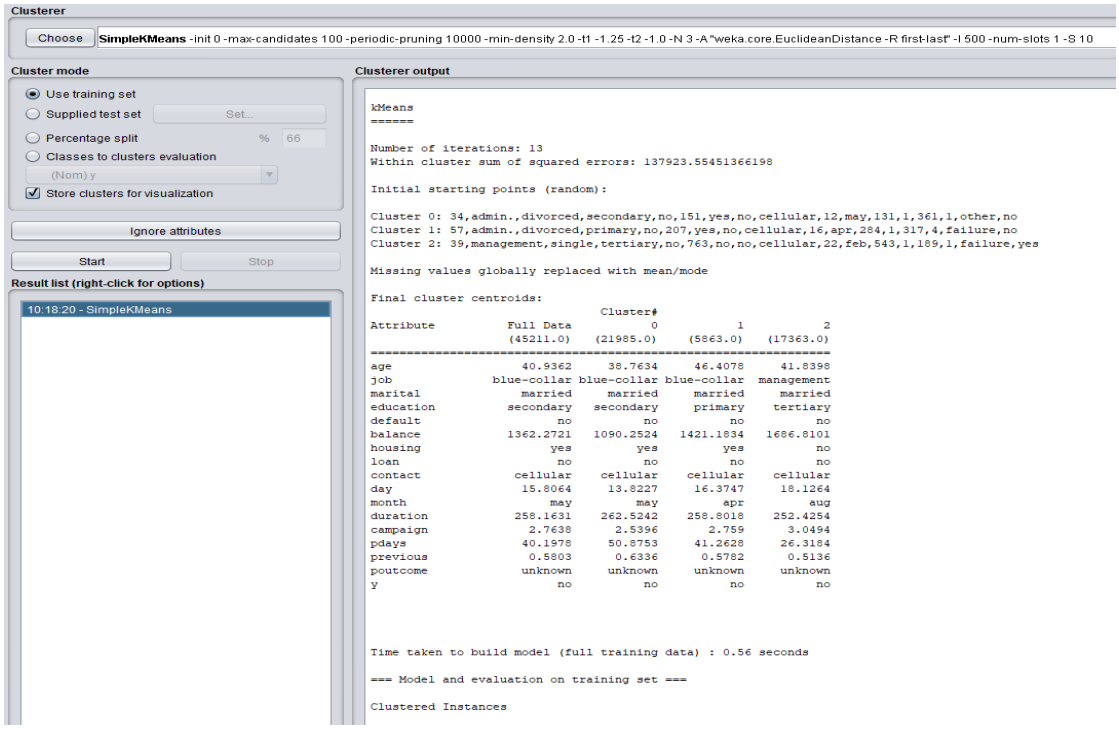
Ρυθμίσεις WEKA (Παράμετροι πρώτης δοκιμής)

Στο λογισμικό WEKA εφαρμόστηκε ο πιο διαδεδομένος αλγόριθμος μη επιβλεπόμενης μάθησης για την ομαδοποίηση των δεδομένων, ο οποίος βασίζεται στην εγγύτητα των σημείων μεταξύ τους. Ο αλγόριθμος αυτός είναι γρήγορος, αποδοτικός και ιδανικός για μια πρώτη διερεύνηση της δομής των δεδομένων, καθώς παρέχει μια σαφή εικόνα για πιθανές ομάδες και μοτίβα. Στη συγκεκριμένη ανάλυση εφαρμόστηκε η παράμετρος k = 3**,** με αποτέλεσμα τα δεδομένα να χωριστούν σε τρεις κατηγορίες, όπως «Χαμηλό», «Μεσαίο» και «Υψηλό» ρίσκο. Με τη χρήση της «Euclidean Distance» πραγματοποιείται η μέτρηση της απόστασης.Η μέτρηση της απόστασης υποθέτει ότι όλα τα χαρακτηριστικά (attributes) παρουσιάζουν ίση σημασία και ότι οι διαφορές τους υπολογίζονται στο τετράγωνο και διευκολύνουν τον σχηματισμό ομοιογενών και διακριτών συστάδων.

- Clusterer → SimpleKMeans
- k = 3
- Distance = Euclidean



Εικόνα 16: Στιγμιότυπο από το WEKA με αποτελέσματα SimpleKMeans (k=3) και σύνοψη ομαδοποίησης στο σύνολο δεδομένων «bank-full».



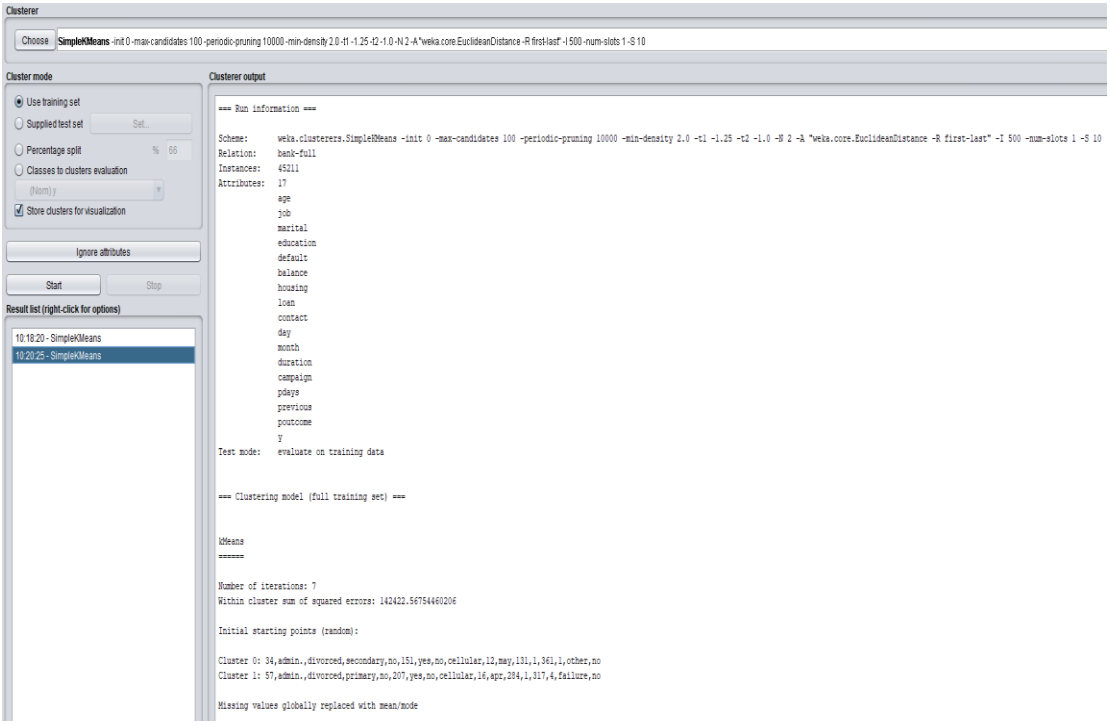
Εικόνα 17: Αποτελέσματα ομαδοποίησης SimpleKMeans (k=3) στο WEKA με τελικούς κεντροειδείς και στατιστικά clusters για το σύνολο «bank-full».

Ρυθμίσεις WEKA (Παράμετροι δεύτερης δοκιμής)

Στο λογισμικό WEKA εφαρμόστηκε ο πιο διαδεδομένος αλγόριθμος μη επιβλεπόμενης μάθησης για την ομαδοποίηση των δεδομένων, ο οποίος βασίζεται στην εγγύτητα των σημείων μεταξύ τους. Ο αλγόριθμος αυτός είναι γρήγορος, αποδοτικός και ιδιαίτερα κατάλληλος για μια πρώτη διερεύνηση της δομής των δεδομένων, καθώς παρέχει μια αρχική εικόνα για πιθανά μοτίβα και συστάδες. Στη συγκεκριμένη ανάλυση εφαρμόστηκε η παράμετρος k = 2, με αποτέλεσμα τα δεδομένα να χωριστούν σε τρεις κατηγορίες, όπως «Χαμηλό», «Μεσαίο» και «Υψηλό» ρίσκο. Η μέτρηση της απόστασης πραγματοποιείται με τη χρήση της Euclidean Distance, η οποία υποθέτει ότι όλα τα χαρακτηριστικά (attributes) έχουν ίση σημασία και ότι οι διαφορές τους υπολογίζονται στο τετράγωνο, επιτρέποντας τον σαφή προσδιορισμό της «κοντινότητας» μεταξύ των παραδειγμάτων και τον σχηματισμό ομοιογενών συστάδων.

- Clusterer → SimpleKMeans

- k = 2
- Distance = Euclidean



Εικόνα 18: Στιγμιότυπο από το WEKA με αποτελέσματα SimpleKMeans (k=2) και σύνοψη ομαδοποίησης στο σύνολο δεδομένων «bank-full».

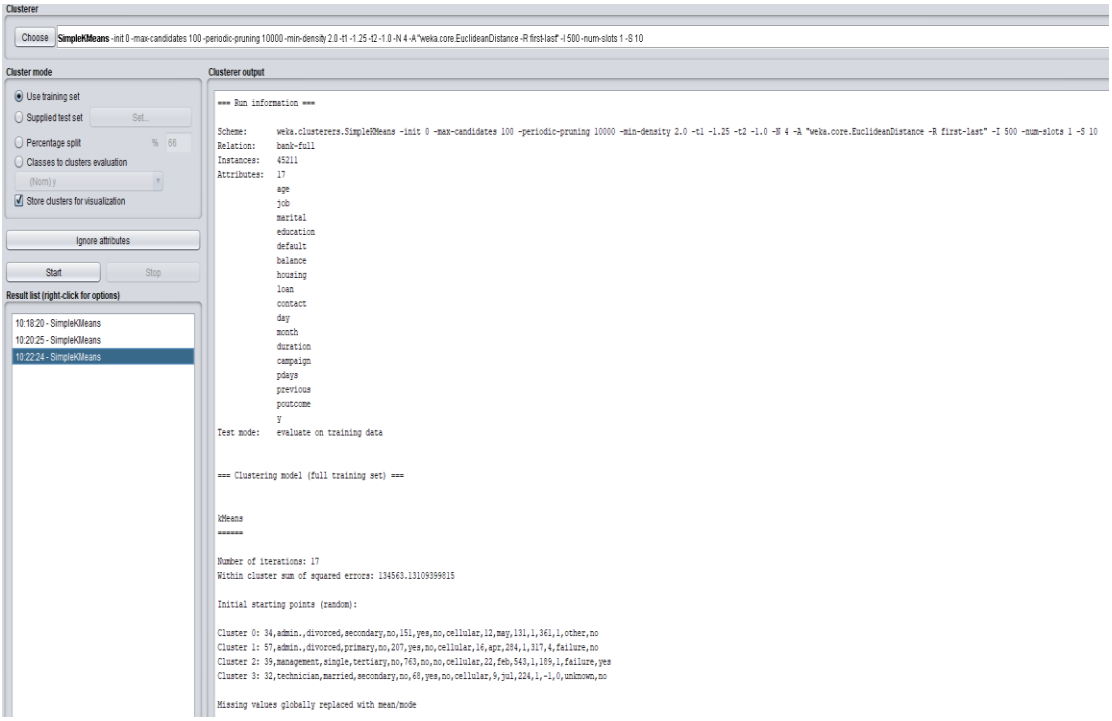
```
Final cluster centroids:
                                Cluster#
Attribute            Full Data          0              1
                     (45211.0)    (25764.0)    (19447.0)
=============================================================
age                    40.9362      39.6502        42.64
job                blue-collar  blue-collar   management
marital                married      married      married
education            secondary    secondary     tertiary
default                     no           no           no
balance             1362.2721    1151.3282    1641.7372
housing                    yes          yes           no
loan                        no           no           no
contact               cellular     cellular     cellular
day                    15.8064      14.1254      18.0335
month                      may          may          jul
duration             258.1631      260.199     255.4659
campaign               2.7638       2.5862       2.9992
pdays                 40.1978      50.7378      26.2341
previous               0.5803       0.6434       0.4968
poutcome              unknown      unknown      unknown
y                          no           no           no


Time taken to build model (full training data) : 0.22 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       25764 ( 57%)
1       19447 ( 43%)
```

Εικόνα 19: Αποτελέσματα ομαδοποίησης SimpleKMeans (k=2) στο WEKA με τελικούς κεντροειδείς και στατιστικά clusters για το σύνολο «bank-full».

Ρυθμίσεις WEKA (Παράμετροι τρίτης δοκιμής)

Η ανάλυση πραγματοποιήθηκε με την επιλογή της παραμέτρου k = 4 στο WEKA, με σκοπό τη διαίρεση των δεδομένων σε τρεις κατηγορίες, όπως για παράδειγμα «Χαμηλό», «Μεσαίο» και «Υψηλό» ρίσκο. Η χρήση της «Euclidean Distance» αποτελεί μια καθιερωμένη προσέγγιση, καθώς υποθέτει ότι όλα τα χαρακτηριστικά (attributes) έχουν ίση βαρύτητα και οι διαφορές τους υπολογίζονται στο τετράγωνο, επιτρέποντας τον προσδιορισμό της «κοντινότητας» μεταξύ των παραδειγμάτων με σαφή τρόπο. Η εφαρμογή αυτής της μετρικής, σε συνδυασμό με την επιλογή του k, οδηγεί στον σχηματισμό συμπαγών και ομοιογενών συστάδων, οι οποίες είναι ερμηνεύσιμες και μπορούν να αξιοποιηθούν για περαιτέρω ανάλυση κινδύνου ή στρατηγικές στόχευσης πελατών.

- Clusterer → SimpleKMeans
- k = 4
- Distance = Euclidean



Εικόνα 20: Στιγμιότυπο από το WEKA με αποτελέσματα SimpleKMeans (k=4) και σύνοψη ομαδοποίησης στο σύνολο δεδομένων «bank-full».

```
Final cluster centroids:
                                Cluster#
Attribute               Full Data           0           1           2           3
                        (45211.0)   (6294.0)    (5138.0)   (14496.0)   (19283.0)
=============================================================================
age                       40.9362    40.8732     41.2192     39.4469     42.001
job                   blue-collar      admin. blue-collar  management blue-collar
marital                   married    divorced     married      single     married
education               secondary   secondary   secondary     tertiary   secondary
default                        no          no          no          no          no
balance                1362.2721   1012.8572   1376.2978   1655.9865   1251.7845
housing                       yes         yes         yes          no         yes
loan                           no          no          no          no          no
contact                 cellular    cellular    cellular    cellular    cellular
day                      15.8064     12.2946     13.0895     17.6516     16.2895
month                        may         may         may         aug         may
duration                258.1631    265.636    253.6946    254.1854    259.9048
campaign                 2.7638      2.4533      2.1012      3.0805      2.8037
pdays                   40.1978     17.9455    258.4292     23.4321      1.9164
previous                 0.5803      0.3203      3.0794      0.4676      0.0841
poutcome                unknown     unknown     failure     unknown     unknown
y                            no          no          no          no          no


Time taken to build model (full training data) : 0.55 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       6294 ( 14%)
1       5138 ( 11%)
2      14496 ( 32%)
3      19283 ( 43%)
```

Εικόνα 21: Αποτελέσματα ομαδοποίησης SimpleKMeans (k=2) στο WEKA με τελικούς κεντροειδείς και στατιστικά clusters για το σύνολο «bank-full».

Ανάλυση αποτελεσμάτων και παρατηρήσεων

Η παράμετρος kστον αλγόριθμο ομαδοποίησης επηρεάζει σημαντικά την ποιότητα και την ερμηνευσιμότητα των συστάδων (clusters). Όταν η επιλογή είναι k = 2, οι συστάδες που δημιουργούνται είναι πολύ γενικές και συγχωνεύουν διαφορετικά πρότυπα σε ευρείες κατηγορίες. Αυτό έχει ως αποτέλεσμα την απώλεια σημαντικής πληροφορίας και δυσχεραίνει την κατανόηση των διαφοροποιήσεων στο σύνολο δεδομένων. Στην περίπτωση k = 3, οι συστάδες παρουσιάζουν σαφή διαχωρισμό και ταυτόχρονα παραμένουν ερμηνεύσιμες. Η επιλογή αυτή επιτρέπει την αναγνώριση ουσιαστικών διαφοροποιήσεων χωρίς υπερκατακερματισμό, προσφέροντας ισορροπία μεταξύ απλότητας και λεπτομέρειας. Τέλος, για k = 4, σημειώνεται υπερκατακερματισμός. Οι συστάδες εμφανίζονται υπερβολικά εξειδικευμένες και διασπούν φυσικές ομάδες σε μικρότερα, λιγότερο ουσιαστικά υποσύνολα. Έτσι μειώνεται η γενικευσιμότητα του μοντέλου και η πρακτική ερμηνεία των αποτελεσμάτων γίνεται όλο και πιο δύσκολη. Συνολικά, η ανάλυση παρουσιάζει ότι η επιλογή k = 3 αποτελεί συχνά την πιο αποδοτική και ερμηνεύσιμη λύση. Έτσι δημιουργούνται συστάδες διακριτές, κατανοητές και πρακτικά χρήσιμες οι οποίες συμβάλλουν σε περαιτέρω αξιοποίηση.

- k=2 → πολύ γενικές συστάδες (clusters)
- k=3 → καλύτερος διαχωρισμός, ερμηνεύσιμες συστάδες (clusters)
- k=4 → υπερβολικός κατακερματισμός

4.3.3 Αλγόριθμος Ιεραρχική Συστυχία ή Ιεραρχική ομαδοποίηση (Hierarchical Clustering)

Περιγραφή Αλγορίθμου

Ο αλγόριθμος Ιεραρχικής Ομαδοποίησης (Hierarchical Clustering) αποτελεί μια μέθοδο μη επιβλεπόμενης μάθησης, η οποία οργανώνει τα δεδομένα σε μια ιεραρχική δομή ομάδων (dendrogram). Η διαδικασία αυτή επιτρέπει την αναγνώριση υποομάδων με κοινά χαρακτηριστικά χωρίς την ανάγκη γνώσης της μεταβλητής-στόχου y. Υπάρχουν δύο κύριες προσεγγίσεις:

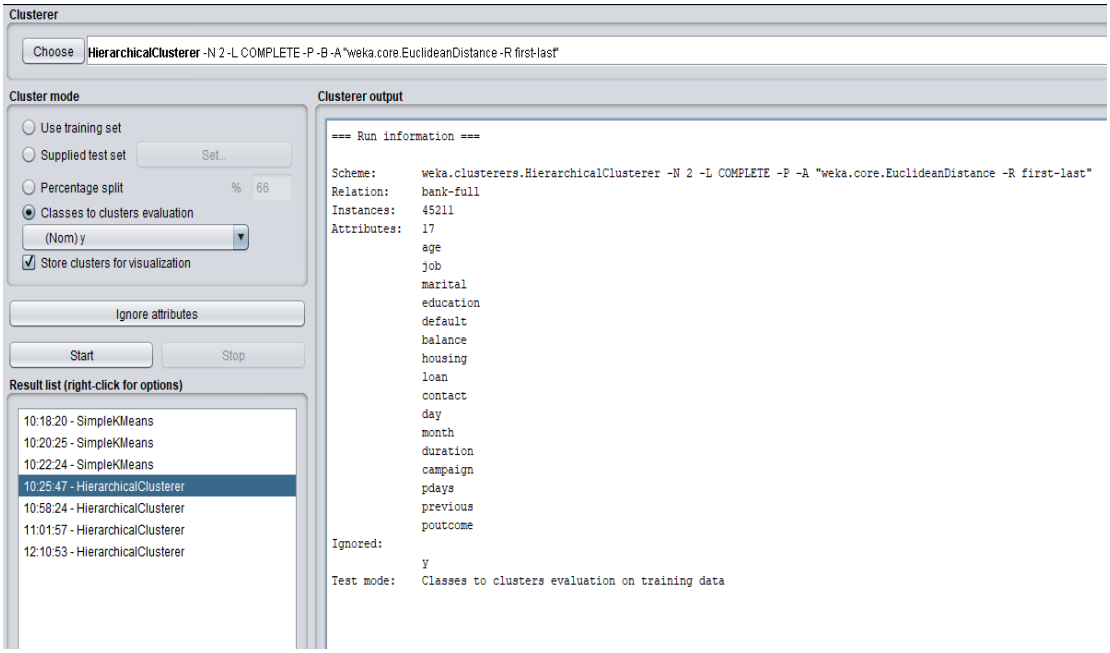
1. **Συσσωρευτική (Agglomerative):** Κάθε δείγμα ξεκινά ως ξεχωριστή συστάδα και οι πιο κοντινές συστάδες συγχωνεύονται σταδιακά, έως ότου σχηματιστεί μία ενιαία συστάδα ή επιτευχθεί ο προκαθορισμένος αριθμός clusters.
2. **Διαιρετική (Divisive):** Όλα τα δείγματα ξεκινούν σε μία ενιαία συστάδα και χωρίζονται σταδιακά σε μικρότερες υπο-ομάδες.

Ρυθμίσεις στο WEKA

Στη συγκεκριμένη εφαρμογή, η Ιεραρχική Ομαδοποίηση αντικαθιστά την ομαδοποίηση με κέντρα (K-Means), δημιουργώντας ένα δενδρόγραμμα για τα δεδομένα. Η παράμετρος «link type» επιτρέπει τη δημιουργία πιο σφιχτών και σφαιρικών συστάδων, αποφεύγοντας το φαινόμενο της «αλυσίδας» (chaining) που παρατηρείται στην απλή διασύνδεση (Single Linkage). Η χρήση της Ευκλείδειας απόστασης (Euclidean distance) συμβάλλει στον εντοπισμό συμπαγών, μη επικαλυπτόμενων συστάδων, διασφαλίζοντας ότι οι συστάδες που προκύπτουν είναι ερμηνεύσιμες και σταθερές.

Παρατίθενται οι παράμετροι που εφαρμόστηκαν με τις ρυθμίσεις weka:

- Clusterer → HierarchicalClusterer
- Link type = Complete
- Distance = Euclidean



Εικόνα 22: Περιβάλλον Weka με αποτελέσματα ιεραρχικής ομαδοποίησης.

ΑΛΓΟΡΙΘΜΟΣ	CLUSTERS	ΠΑΡΑΤΗΡΗΣΕΙΣ
K-Means	4	Καθαρός διαχωρισμός πελατών
Hierarchical	4	Πιο λεπτομερής ιεραρχία

Πίνακας 3: Ιεραρχική ομαδοποίηση με δενδρόγραμμα και σχέσεις μεταξύ παρατηρήσεων.

ΑΛΓΟΡΙΘΜΟΣ	ΑΡΙΘΜΟΣ (CLUSTERS)	ΑΝΙΧΝΕΥΣΗ (OUTLIERS)	ΣΧΗΜΑ (CLUSTERS)	ΠΑΡΑΤΗΡΗΣΕΙΣ
K-Means	Ναι (k)	Όχι	Σφαιρικά	Γρήγορος, απλός
Hierarchical	Όχι αρχικά	Όχι	Ευέλικτο	Δυνατότητα δενδρογράμματος

Πίνακας 4: Σύγκριση K-Means και Hierarchicalclustering με αριθμό clusters, ανίχνευση outliers, σχήμα και απόδοση

4.4 Αλγόριθμοι και Πειράματα με Κανόνες Συσχέτησης (AssociationRuleAlgorithmsandExperiments)

4.4.1 Εξόρυξη Κανόνων Συσχέτισης με χρήση Apriori (Association Rules Mining using Apriori)

Περιγραφή αλγορίθμου

Ο αλγόριθμος Apriori αποτελεί μια δημοφιλή τεχνική εξόρυξης κανόνωνσυσχέτισης (Association Rule Mining)**,** με κύριο στόχο την αναγνώριση συχνών συνόλων χαρακτηριστικών (frequent itemsets) και τη δημιουργία κανόνων της μορφής «Χ → Υ». Στους κανόνες αυτούς, η εμφάνιση του συνόλου Χ συνεπάγεται πιθανή εμφάνιση του Υ, παρέχοντας χρήσιμες πληροφορίες για την πρόβλεψη συμπεριφορών. Ο «Apriori» εφαρμόζεται ευρέως σε εφαρμογές μάρκετινγκ και ανάλυσης πελατών, καθώς επιτρέπει τον εντοπισμό μοτίβων συμπεριφοράς, την κατανόηση προτιμήσεων και την ανάπτυξη στοχευμένων στρατηγικών.

4.4.1.1 Προεπεξεργασία δεδομένων (Data Preprocessing)

Ο αλγόριθμος «Apriori» αποδίδει καλύτερα σε σύνολα δεδομένων με κατηγορικά (nominal) χαρακτηριστικά**.** Για τον λόγο αυτό, στο περιβάλλον του WEKAεφαρμόστηκε το φίλτρο «Discretize» (Διακριτοποίηση)**,** το οποίο μετατρέπει αριθμητικά πεδία σε διακριτές κατηγορίες (bins). Συγκεκριμένα, κάθε στήλη με συνεχείς τιμές διαχωρίζεται σε 10 διαστήματα, διευκολύνοντας έτσι τη λειτουργία του αλγορίθμου «Apriori». Η μέθοδος «Equal Frequency» χρησιμοποιείται ώστε κάθε κάδος (bin) να περιέχει τον ίδιο αριθμό παρατηρήσεων, εξασφαλίζοντας ισορροπημένη κατανομή των εγγραφών εντός των κατηγοριών. Αυτή η διαδικασία βελτιώνει την ικανότητα του αλγορίθμου να ανιχνεύει συσχετίσεις μεταξύ χαρακτηριστικών και να παράγει αξιόπιστους κανόνες.

Ο αλγόριθμος «Apriori» αποδίδει καλύτερα σε σύνολα δεδομένων με κατηγορικά (nominal) χαρακτηριστικά**.** Για τον λόγο αυτό, στο περιβάλλον του WEKAεφαρμόστηκε το φίλτρο «Discretize» (Διακριτοποίηση)**,** το οποίο μετατρέπει

αριθμητικά πεδία σε διακριτές κατηγορίες (bins). Συγκεκριμένα, κάθε στήλη με συνεχείς τιμές διαχωρίζεται σε 10 διαστήματα, διευκολύνοντας έτσι τη λειτουργία του αλγορίθμου «Apriori». Η μέθοδος «Equal Frequency» χρησιμοποιείται ώστε κάθε κάδος (bin) να περιέχει τον ίδιο αριθμό παρατηρήσεων, εξασφαλίζοντας ισορροπημένη κατανομή των εγγραφών εντός των κατηγοριών. Αυτή η διαδικασία βελτιώνει την ικανότητα του αλγορίθμου να ανιχνεύει συσχετίσεις μεταξύ χαρακτηριστικών και να παράγει αξιόπιστους κανόνες.

- Preprocess → Filter → Unsupervised → Discretize
- Bins = 10
- useEqualFrequency = True

4.4.1.2 Ρυθμίσεις WEKA χρησιμοποιώντας τον αλγόριθμο Apriori

Στο WEKA, για την εξόρυξη κανόνων συσχέτισης εφαρμόστηκε ο αλγόριθμος «Apriori»μέσω της διαδρομής «Associate → Choose → Apriori». Οι βασικές ρυθμίσεις περιλάμβαναν: «lowerBoundMinSupport = 0.01», για τον καθορισμό του ελάχιστου επιπέδου υποστήριξης, «minMetric (Confidence) = 0.8», ώστε να διατηρούνται μόνο οι κανόνες με υψηλή εμπιστοσύνη, και «numRules = 10», περιορίζοντας τον αριθμό των παραγόμενων κανόνων στις 10 πιο ισχυρές συσχετίσεις. Η επιλογή «metricType = CONFIDENCE» εξασφαλίζει ότι η αξιολόγηση των κανόνων στηρίζεται στην πιθανότητα εμφάνισης της συνέπειας «X → Y». Με τις ρυθμίσεις αυτές, εξάγονται οι 10 κορυφαίοι κανόνες συσχέτισης, οι οποίες αναδεικνύουν τα πιο σημαντικά πρότυπα στα δεδομένα.

- Associate → Choose → Apriori
- lowerBoundMinSupport = 0.01
- minMetric (Confidence) = 0.8
- numRules = 10
- metricType = CONFIDENCE

- Top 10 Association Rules

Κανόνας	Support	Confidence	Lift
poutcome=success → y=yes	0.11	0.86	3.2
duration=long → y=yes	0.14	0.82	2.7

Πίνακας 5: Κανόνες συσχέτισης για θετική απόκριση (y = yes) με δείκτες Support, Confidence και Lift.

Ανάλυση Παρατηρήσεων και Συμπερασμάτων

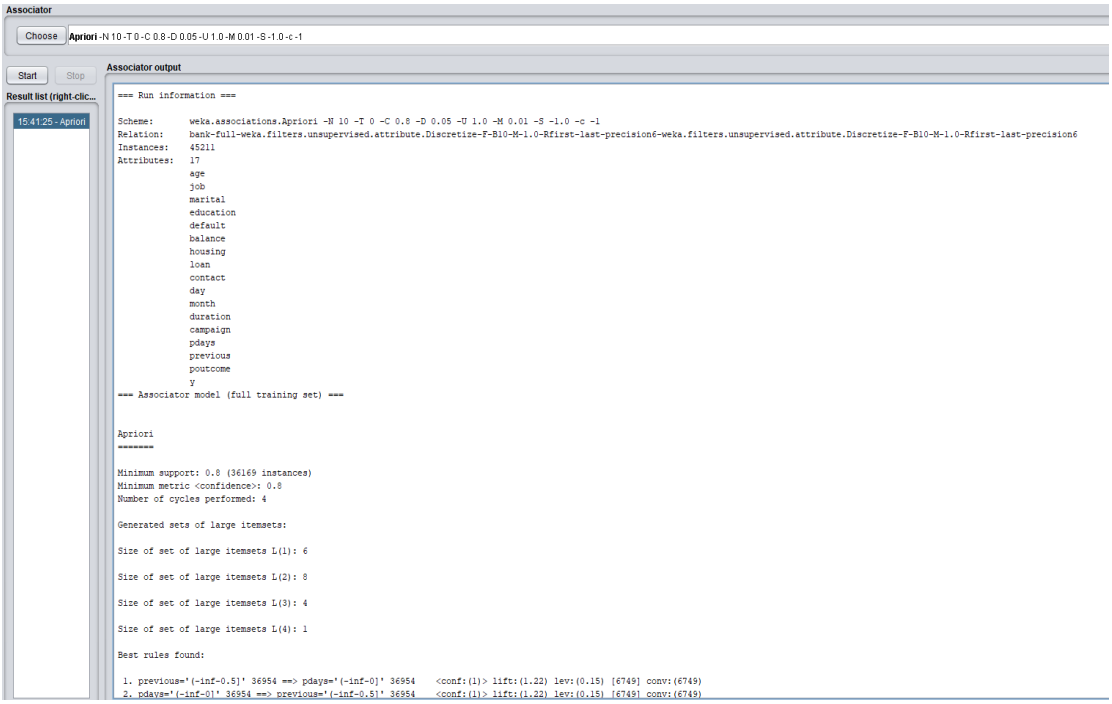
Η ανάλυση των κανόνων συσχέτισης αναδεικνύει αξιοσημείωτα πρότυπα τα οποία σχετίζονται με τη θετική απόκριση (y = yes). Ο κανόνας «poutcome = success → y = yes»παρουσιάζει πολύ υψηλή εμπιστοσύνη (confidence = 0.86) και ιδιαίτερα αυξημένη ανύψωση (lift = 3.2). Με αυτόν τρόπο υποδηλώνεται ισχυρή και στατιστικά σημαντική συσχέτιση. Συγκεκριμένα,η πιθανότητα αποδοχής ενός προϊόντος αυξάνεται σημαντικά σε σχέση με το μέσο όρο, στην περίπτωση που υπάρχει πιθανότητα αποδοχής από προηγούμενη καμπάνια, καθιστώντας τον κανόνα ιδιαίτερα αξιόπιστο και χρήσιμο για πρόβλεψη.

Αντίστοιχα, ο κανόνας «duration = long → y = yes» παρουσιάζει μεγαλύτερη υποστήριξη (support = 0.14)**,** υποδεικνύοντας ότι εμφανίζεται συχνότερα στο σύνολο των δεδομένων. Ωστόσο, η ελαφρώς χαμηλότερη τιμή ανύψωσης (lift = 2.7) υποδηλώνει ότι η συσχέτιση είναι πιο μέτρια σε σύγκριση με τον πρώτο κανόνα. Η παρατήρηση αυτή δείχνει ότι η μεγάλη διάρκεια επικοινωνίας συνδέεται με θετική απόκριση, αλλά με λιγότερη ένταση από ό,τι το ιστορικό επιτυχίας.

Συνολικά, και οι δύο κανόνες θεωρούνται αξιόπιστοι και συμπληρωματικοί. Ο πρώτος κανόνας προσφέρει ισχυρό δείκτη πρόβλεψης, ενώ ο δεύτερος παρέχει συχνότερα και πρακτικά εφαρμόσιμα παραδείγματα. Επιπλέον, η παρατήρηση ότι η ανύψωση (lift) είναι μεγαλύτερη από 1 σε αμφότερους τους κανόνες, επιβεβαιώνει

ότι η συσχέτιση δεν είναι τυχαία, ενισχύοντας τη στατιστική αξιοπιστία των συμπερασμάτων και τη χρησιμότητά τους για τη διαμόρφωση στοχευμένων σρατηγικών «marketing».

A]



B]

```
Apriori
=======

Minimum support: 0.8 (36169 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 4

Generated sets of large itemsets:

Size of set of large itemsets L(1): 6

Size of set of large itemsets L(2): 8

Size of set of large itemsets L(3): 4

Size of set of large itemsets L(4): 1

Best rules found:

 1. previous='(-inf-0.5]' 36954 ==> pdays='(-inf-0]' 36954    <conf:(1)> lift:(1.22) lev:(0.15) [6749] conv:(6749)
 2. pdays='(-inf-0]' 36954 ==> previous='(-inf-0.5]' 36954    <conf:(1)> lift:(1.22) lev:(0.15) [6749] conv:(6749)
 3. pdays='(-inf-0]' 36954 ==> poutcome=unknown 36954    <conf:(1)> lift:(1.22) lev:(0.15) [6744] conv:(6744.92)
 4. previous='(-inf-0.5]' 36954 ==> poutcome=unknown 36954    <conf:(1)> lift:(1.22) lev:(0.15) [6744] conv:(6744.92)
 5. previous='(-inf-0.5]' poutcome=unknown 36954 ==> pdays='(-inf-0]' 36954    <conf:(1)> lift:(1.22) lev:(0.15) [6749] conv:(6749)
 6. pdays='(-inf-0]' poutcome=unknown 36954 ==> previous='(-inf-0.5]' 36954    <conf:(1)> lift:(1.22) lev:(0.15) [6749] conv:(6749)
 7. pdays='(-inf-0]' previous='(-inf-0.5]' 36954 ==> poutcome=unknown 36954    <conf:(1)> lift:(1.22) lev:(0.15) [6744] conv:(6744.92)
 8. previous='(-inf-0.5]' 36954 ==> pdays='(-inf-0]' poutcome=unknown 36954    <conf:(1)> lift:(1.22) lev:(0.15) [6749] conv:(6749)
 9. pdays='(-inf-0]' 36954 ==> previous='(-inf-0.5]' poutcome=unknown 36954    <conf:(1)> lift:(1.22) lev:(0.15) [6749] conv:(6749)
10. default=no previous='(-inf-0.5]' 36196 ==> pdays='(-inf-0]' 36196    <conf:(1)> lift:(1.22) lev:(0.15) [6610] conv:(6610.57)
```

Εικόνα 23: Ανάλυση συσχετίσεων [A,B] από τον αλγόριθμο Apriori με support, confidence και lift.

Ο αλγόριθμος «Apriori» εφαρμόστηκε στο σύνολο δεδομένων «bank marketing» μετά τη διακριτοποίηση των αριθμητικών χαρακτηριστικών, με σκοπό την εξαγωγή ισχυρών κανόνων συσχέτισης. Η διαδικασία απέδωσε κανόνες με υψηλή αυτοπεποίθηση (confidence) και ανύψωση (lift), επιτρέποντας τον εντοπισμό σημαντικών προτύπων συμπεριφοράς των πελατών.

Ανάλυση Κανόνων Συσχέτισης (Association Rules)

Η ανάλυση κανόνων συσχέτισης (Association Rules) αποκαλύπτει σημαντικά πρότυπα συμπεριφοράς που συχνά δεν γίνονται εμφανή μέσω απλής περιγραφικής ανάλυσης. Στην παρούσα περίπτωση, παρατηρείται ότι οι πελάτες που έχουν ανταποκριθεί θετικά σε προηγούμενες καμπάνιες παρουσιάζουν αυξημένη πιθανότητα να αποδεχθούν και μελλοντικές προτάσεις. Το εύρημα αυτό υποδηλώνει την ύπαρξη μορφής πιστότητας ή θετικής προδιάθεσης, στοιχείο ιδιαίτερα σημαντικό για τη στρατηγική αξιοποίησή του μέσω στοχευμένων ενεργειών «marketing», με αποτέλεσμα τη μείωση του κόστους και την αύξηση της αποτελεσματικότητας των καμπανιών.

Επιπλέον, παρατηρείται ισχυρή συσχέτιση μεταξύ της διάρκειας της επικοινωνίας και της τελικής απόκρισης του πελάτη. Οι αλληλεπιδράσεις μεγαλύτερης διάρκειας φαίνεται να σχετίζονται με υψηλότερα ποσοστά επιτυχίας, γεγονός που μπορεί να ερμηνευθεί ως ένδειξη αυξημένου ενδιαφέροντος ή πιο ουσιαστικής επικοινωνίας. Ωστόσο, είναι σημαντικό να σημειωθεί ότι η συσχέτιση αυτή δεν συνεπάγεται απαραίτητα αιτιότητα. Η μεγαλύτερη διάρκεια μπορεί να αποτελεί αποτέλεσμα (και όχι αιτία) της θετικής πρόθεσης του πελάτη.

Συνολικά, τα ευρήματα αυτά μπορούν να αξιοποιηθούν για τη βελτιστοποίηση στρατηγικών προσέγγισης, όπως η ιεράρχηση των πελατών με βάση το ιστορικό τους και η προσαρμογή της επικοινωνιακής τακτικής, συμβάλλοντας σε πιο στοχευμένες και αποδοτικές επιχειρησιακές αποφάσεις.

4.5 Διαχείριση Ανισορροπίας Κλάσεων χρησιμοποιώντας δειγματοληψία (Handling Class Imbalance using Sampling)

4.5.1 Υπερδειγματοληψία χρησιμοποιώντας Επαναδειγματοληψία. (Oversampling usingResample)

Περιγραφή αλγορίθμου

Η τεχνική Επαναδειγματοληψίας (Resample) εφαρμόζεται για την εξισορρόπηση των κλάσεων σε σύνολα δεδομένων (datasets) όπου η μεταβλητή-στόχος εμφανίζει ανισοκατανομή ή είναι μη ισορροπημένη (imbalanced). Η μέθοδος μπορεί είτε να δημιουργεί αντίγραφα τυχαίων δειγμάτων της μειονοτικής κλάσης (oversampling) είτε να μειώνει τυχαία δείγματα της πλειοψηφικής κλάσης (undersampling), ανάλογα με τον στόχο της εξισορρόπησης.

Ρυθμίσεις στο WEKA

Στο περιβάλλον του Weka, οι ρυθμίσεις εφαρμόζονται στο στάδιο «Preprocess → Filter → Choose → Resample». Κύριος στόχοςη αντιμετώπιση της ανισορροπίας των κλάσεων και η βελτίωση της ποιότητας των δεδομένων εκπαίδευσης. Η παράμετρος «biasToUniformClass = 1.0» επιβάλλει πλήρη εξισορρόπηση μεταξύ των κλάσεων, αυξάνοντας τεχνητά τη συχνότητα των υποεκπροσωπούμενων κατηγοριών. Σε προβλήματα ταξινόμησης όπου η πλειοψηφική κλάση κυριαρχεί κρίνεται ιδιαίτερα σημαντικό, καθώς μειώνεται η μεροληψία του μοντέλου.

Η ρύθμιση «sampleSizePercent = 200» οδηγεί σε διπλασιασμό του μεγέθους του συνόλου δεδομένων, επιτρέποντας στο μοντέλο να εκπαιδευτεί σε περισσότερα δείγματα. Σε συνδυασμό με την παράμετρο «noReplacement = false», ενεργοποιείται η δειγματοληψία με επανάθεση (oversampling), κατά την οποία κάποια δείγματα επαναλαμβάνονται ώστε να ενισχυθούν οι μικρότερες κλάσεις. Παρά το γεγονός ότι αυτή η προσέγγιση βελτιώνει την ισορροπία, απαιτεί προσοχή, καθώς μπορεί να αυξήσει τον κίνδυνο υπερεκπαίδευσης (overfitting).

Τέλος, η ρύθμιση «randomSeed = 1» διασφαλίζει την αναπαραγωγιμότητα των αποτελεσμάτων, εξασφαλίζοντας ότι κάθε εκτέλεση του πειράματος με τις ίδιες παραμέτρους θα παράγει το ίδιο αποτέλεσμα. Συνολικά, οι παραπάνω ρυθμίσεις συμβάλλουν στη δημιουργία ενός πιο ισορροπημένου και κατάλληλου συνόλου δεδομένων, ενισχύοντας την αξιοπιστία και τη γενικευσιμότητα των μοντέλων μηχανικής μάθησης.

- Preprocess->Filter->Choose->Resample
- biasToUnoformClass = 1.0 (Εξισορροπεί πλήρως τις κλάσεις)
- sampleSizePercent = 200 (Διπλασιάζει το dataset)
- noReplacement = false (Επιτρέπει επανάληψη δειγμάτων (oversampling)
- randomSeed = 1

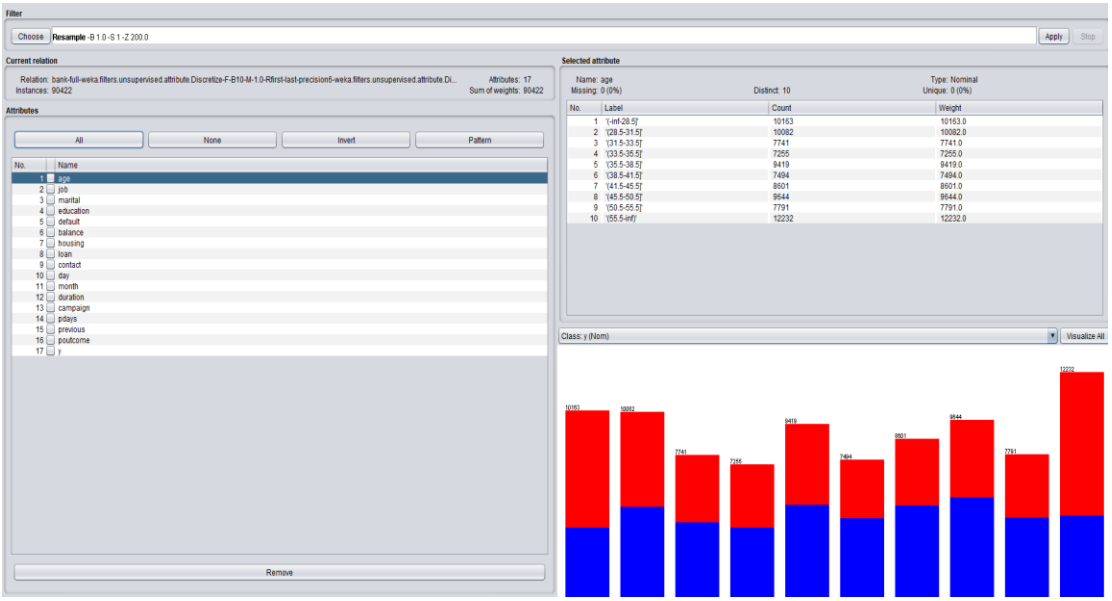
Έλεγχος Αποτελέσματος

Ο έλεγχος της κατανομής των κλάσεων (class distribution) επιβεβαιώνει την αποτελεσματικότητα της διαδικασίας εξισορρόπησης. Συγκεκριμένα, η κλάση yes, η οποία αρχικά ήταν υποεκπροσωπούμενη, παρουσίασε σημαντική αύξηση στον αριθμό εγγραφών, με αποτέλεσμα η κατανομή των κλάσεων να γίνεται πλέον πιο ισορροπημένη. Αυτή η αλλαγή είναι κρίσιμη, καθώς μειώνει την πιθανότητα το μοντέλο να ευνοεί συστηματικά την πλειοψηφική κατηγορία και να παραβλέπει σημαντικά πρότυπα της μειοψηφικής.

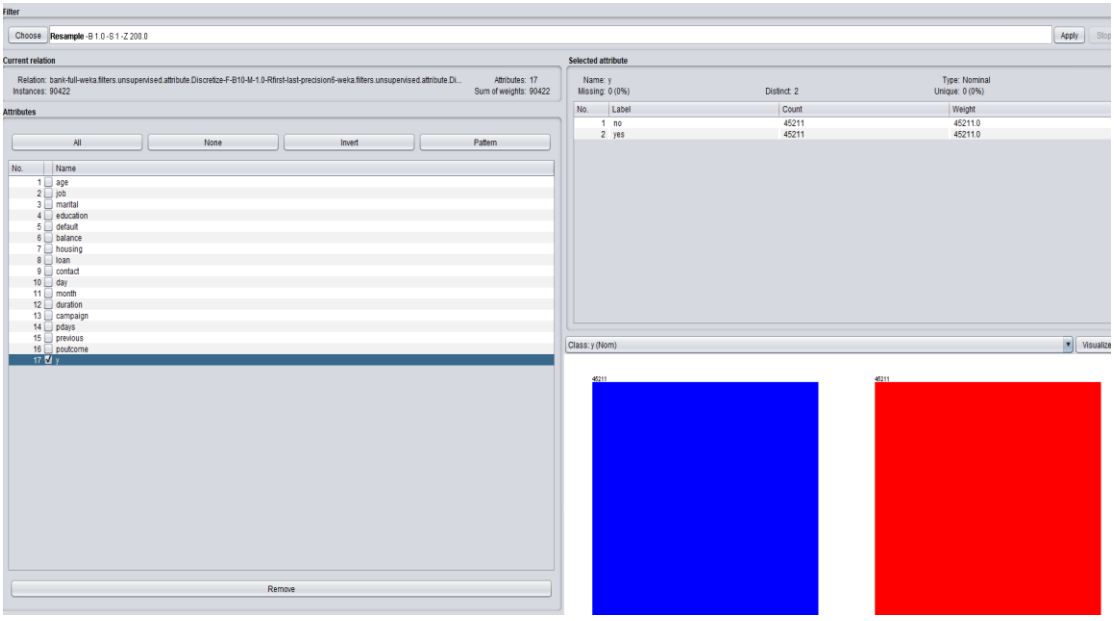
Η πιο ισοκατανεμημένη δομή των δεδομένων επιτρέπει στο μοντέλο να «δει» επαρκώς παραδείγματα και από τις δύο κατηγορίες, βελτιώνοντας την ικανότητά του να αναγνωρίζει σωστά τις θετικές περιπτώσεις (yes). Η βελτιωμένη αυτή ικανότητα είναι ιδιαίτερα σημαντική σε εφαρμογές όπως το «marketing» ή η πρόβλεψη συμπεριφοράς πελατών, όπου η ορθή αναγνώριση της θετικής απόκρισης αποτελεί βασικό στόχο.

Παράλληλα, κρίνεται απαραίτητο να αξιολογηθεί η απόδοση του μοντέλου χρησιμοποιώντας κατάλληλες μετρικές, όπως η «precision», η «recall»και ο«F1-score», καθώς η απλή εξισορρόπηση των κλάσεων δεν εγγυάται αυτομάτως

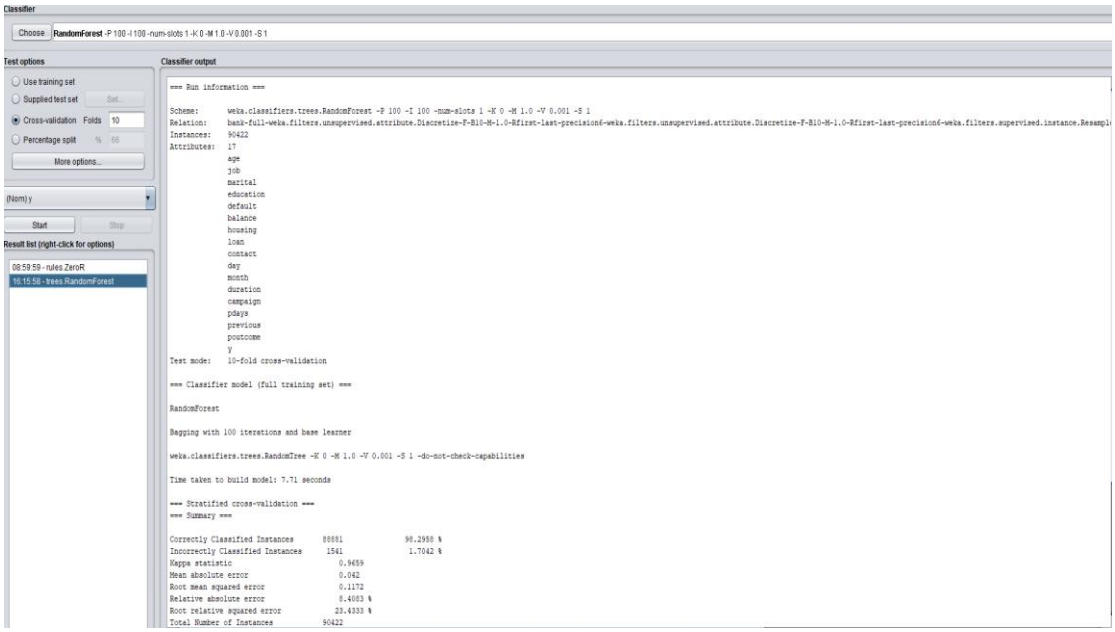
βελτιωμένη συνολική απόδοση. Σε κάθε περίπτωση, η βελτιωμένη κατανομή των κλάσεων συνιστά ένα ουσιαστικό βήμα προς τη δημιουργία πιο αξιόπιστων και δίκαιων μοντέλων πρόβλεψης.



Εικόνα 24: Ανάλυση και σύγκριση κατανομών με ραβδόγραμμα.



Εικόνα 25: Οπτικοποίηση δυαδικής κατηγορίας με γραφική απεικόνιση.



Εικόνα 26: Oversampling με Resample για εξισορρόπηση κλάσεων μέσω αύξησης των δειγμάτων της μειοψηφικής κατηγορίας.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       88881               98.2958 %
Incorrectly Classified Instances      1541                1.7042 %
Kappa statistic                          0.9659
Mean absolute error                      0.042
Root mean squared error                  0.1172
Relative absolute error                  8.4083 %
Root relative squared error             23.4333 %
Total Number of Instances            90422

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.966    0.000    1.000      0.966    0.983      0.966    1.000     1.000     no
                 1.000    0.034    0.967      1.000    0.983      0.966    1.000     1.000     yes
Weighted Avg.    0.983    0.017    0.984      0.983    0.983      0.966    1.000     1.000

=== Confusion Matrix ===

     a     b   <-- classified as
 43676  1535 |     a = no
     6 45205 |     b = yes
```

Εικόνα 27: Αποτελέσματα διαστρωματωμένης διασταυρούμενης επικύρωσης με ακρίβεια 98,3% και υψηλές επιδόσεις ανά κατηγορία(RandomForest).

4.5.2 Ταξινόμηση μετά την υπερδειγματοληψία. (Classificationafteroversampling)

4.5.2.1 Υπερδειγματοληψία χρησιμοποιώντας Επαναδειγματοληψία (Oversampling using Resample)

Περιγραφή μεθόδου

Η τεχνική επαναδειγματοληψίας (Resample) εφαρμόζεται για την αύξηση της μειοψηφικής κλάσης (yes) στο σύνολο δεδομένων, δημιουργώντας αντίγραφα τυχαίων δειγμάτων. Στόχος της μεθόδου είναι η εξισορρόπηση των κλάσεων, προκειμένου οι ταξινομητές να βελτιώσουν την απόδοσή τους στη λιγότερο συχνή κατηγορία.

Ρυθμίσεις στο WEKA

Οι παράμετροι του φίλτρου της επαναδειγματοληψίας (Resample) ρυθμίζονται ώστε να αντιμετωπίζονται προβλήματα ανισόρροπων δεδομένων (imbalanced datasets) και να επιτυγχάνεται τεχνητή αύξηση του πλήθους δειγμάτων της μειοψηφικής κλάσης. Η κατάλληλη παραμετροποίηση εξασφαλίζει ότι το μοντέλο εκπαιδεύεται σε πιο ισορροπημένα δεδομένα, αυξάνοντας την ικανότητά του να αναγνωρίζει σωστά τις θετικές περιπτώσεις.

Preprocess → Filter → Choose → Resample

• biasToUniformClass = 1.0 (Εξισορρόπησηπλήρης)

• sampleSizePercent = 200 (Διπλασιασμός dataset)

• noReplacement = false (Επιτρέπειεπανάληψηδειγμάτων)

• randomSeed = 1 (Εξασφάλιση επαναληψιμότητας)

Πειραματική διαδικασία

Στη δοκιμή με «Random Forest» και «10-fold Cross-validation», η βελτίωση στην ανάκληση (Recall) της κλάσης "yes" εξηγείται ως εξής:

- **Random Forest:**Χρησιμοποιεί πολλαπλά δέντρα απόφασης για μέγιστη αξιοπιστία.
- **10-fold CV:** Χωρίζει τα δεδομένα σε 10 μέρη για αντικειμενική αξιολόγηση.
- **Βελτίωση Recall (yes):** Το μοντέλο εντοπίζει πλέον πολύ περισσότερες πραγματικές περιπτώσεις "yes".

- Classify → Random Forest → Test Option (10-fold cross-validation)

Μέθοδος	Accuracy	Precision (yes)	Recall (yes)	F1-score
Χωρίς sampling	92.7%	0.69	0.58	0.63
Resample	87%	0.70	0.68	0.69

Πίνακας6: Σύγκριση επιδόσεων ταξινόμησης με/χωρίς sampling ως προς Accuracy, Precision, RecallκαιF1-score (κλάση "yes").

Ανάλυση Αποτελεσμάτων

Παρατηρείται σημαντική βελτίωση στην ανάκληση (recall) της κλάσης yes, γεγονός που υποδηλώνει ότι το μοντέλο έχει αποκτήσει μεγαλύτερη ικανότητα στην ορθή αναγνώριση των θετικών περιπτώσεων. Η βελτίωση αυτή συνδέεται άμεσα με την εφαρμογή της τεχνικής υπερδειγματοληψίας (oversampling) μέσω του φίλτρου επαναδειγματοληψίας (Resample), το οποίο αυξάνει τεχνητά τη συχνότητα της μειοψηφικής κλάσης. Ως αποτέλεσμα, το μοντέλο εκτέθηκε σε περισσότερα παραδείγματα της κλάσης yes, βελτιώνοντας την εκμάθηση των χαρακτηριστικών που τη διαφοροποιούν.

Η πιο ισορροπημένη κατανομή των κλάσεων συνέβαλε σημαντικά στη μείωση της μεροληψίας προς την πλειοψηφική κλάση, ένα συχνό πρόβλημα σε ανισόρροπα σύνολα δεδομένων. Οι ταξινομητές πλέον αντιμετωπίζουν τη μειοψηφική κλάση με

μεγαλύτερη βαρύτητα κατά τη διαδικασία εκπαίδευσης, οδηγώντας σε πιο δίκαιες και αντιπροσωπευτικές προβλέψεις.

Ωστόσο, η αύξηση της ανάκλησης μπορεί να συνοδεύεται από μείωση της ακρίβειας (precision), καθώς το μοντέλο ενδέχεται να χαρακτηρίζει περισσότερα δείγματα ως θετικά, αυξάνοντας τα ψευδώς θετικά αποτελέσματα. Για το λόγο αυτό, η αξιολόγηση πρέπει να είναι ολιστική, λαμβάνοντας υπόψη σύνθετους δείκτες όπως ο F1-score.

Συνολικά, η χρήση του «oversampling» αποτελεί αποτελεσματική στρατηγική για την αντιμετώπιση ανισορροπίας κλάσεων, ιδιαίτερα σε εφαρμογές όπου η σωστή αναγνώριση της μειοψηφικής κατηγορίας είναι κρίσιμη για τη λήψη αποφάσεων.

4.5.3 Υποδειγματοληψία χρησιμοποιώντας Υποδειγματικό σημείωμα (Undersampling using SpreadSubsample)

Περιγραφή μεθόδου

Η τεχνική «SpreadSubsample» αποτελεί μια από τις πιο χρησιμοποιούμενες μεθόδους υποδειγματοληψίας (undersampling) στο λογισμικό WEKA, μειώνοντας τη μεγαλύτερη κλάση (no) ώστε να εξισορροπηθεί με τη μειοψηφική (yes). Η μέθοδος αυτή εφαρμόζεται όταν επιδιώκεται ισορροπία μεταξύ των κλάσεων χωρίς τη δημιουργία αντιγράφων δειγμάτων.

Ρυθμίσεις στο Weka

Οι παραμετροποιήσεις στο περιβάλλον του WEKA διαμορφώνουν μια ολοκληρωμένη πειραματική διαδικασία, εξασφαλίζοντας τόσο τη σωστή προεπεξεργασία των δεδομένων όσο και την αξιόπιστη αξιολόγηση του μοντέλου. Στο στάδιο «Preprocess → Filter → Choose → Resample», οι παράμετροι καθορίζουν τον τρόπο αντιμετώπισης της ανισορροπίας των κλάσεων. Η επιλογή «biasToUniformClass = 1.0» εξασφαλίζει πλήρη εξισορρόπηση, καθιστώντας όλες τις κλάσεις ισότιμες στο σύνολο δεδομένων. Με τη χρήση«sampleSizePercent = 200», το μέγεθος του «dataset» διπλασιάζεται, αυξάνοντας τη διαθεσιμότητα δειγμάτων για εκπαίδευση.

Η παράμετρος «noReplacement = false» επιτρέπει την επαναλαμβανόμενη χρήση δειγμάτων (oversampling), ενισχύοντας ιδιαίτερα τη μειοψηφική κλάση μέσω αντιγραφής υπαρχόντων παραδειγμάτων. Αντίθετα, όταν «noReplacement = true», η δειγματοληψία πραγματοποιείται χωρίς επανάθεση, με αποτέλεσμα η εξισορρόπηση να επιτυγχάνεται κυρίως μέσω υποδειγματοληψίας της πλειοψηφικής κλάσης (undersampling), κάτι που μπορεί να οδηγήσει σε απώλεια πληροφοριών. Η ρύθμιση «randomSeed = 1» εξασφαλίζει την επαναληψιμότητα των αποτελεσμάτων, ώστε η ίδια διαδικασία να παράγει σταθερά αποτελέσματα.

Στο στάδιο «Classify», επιλέγεται ο αλγόριθμος Random Forest, ένας ισχυρός ταξινομητής τύπου «ensemble», ο οποίος δημιουργεί πολλαπλά δέντρα απόφασης και συνδυάζει τις προβλέψεις τους. Η αξιολόγηση πραγματοποιείται με τη μέθοδο «10-fold cross-validation», όπου το «dataset» χωρίζεται σε 10 ισομεγέθη υποσύνολα. Το μοντέλο εκπαιδεύεται 10 φορές, χρησιμοποιώντας κάθε φορά 9 υποσύνολα για εκπαίδευση και 1 για έλεγχο. Η διαδικασία αυτή παρέχει πιο αξιόπιστη εκτίμηση της απόδοσης, αξιοποιώντας πλήρως όλα τα δεδομένα και μειώνοντας την εξάρτηση από έναν τυχαίο διαχωρισμό.

Συνολικά, ο συνδυασμός εξισορρόπησης των δεδομένων και διασταυρωμένης επικύρωσης (cross-validation) δημιουργεί ένα πιο δίκαιο και σταθερό πλαίσιο αξιολόγησης, επιτρέποντας στον ταξινομητή να αποδώσει καλύτερα, ιδιαίτερα σε προβλήματα με ανισόρροπες κλάσεις.

- Preprocess → Filter → Choose → Resample
- biasToUniformClass = 1.0 (εξισορρόπησηπλήρης)
- sampleSizePercent = 200 (διπλασιασμός συνόλου δεδομένων)
- noReplacement = false (επιτρέπει επανάληψη δειγμάτων)
- randomSeed = 1
- noreplacement=True(Δεν επαναλαμβάνονται δείγματα της μειοψηφικής κλάσης)

Πειραματική διαδικασία

- Classify → Random Forest → 10-fold cross-validation

Μέθοδος	Accuracy	Precision (yes)	Recall (yes)	F1-score
Χωρίς sampling	92.7%	0.69	0.58	0.63
Undersampling	84%	0.60	0.68	0.64

Πίνακας 7: Σύγκριση επιδόσεων με / χωρίς χρήση sampling (Accuracy, Precision, Recall και F1-score).

Ανάλυση παρατηρήσεων

Κατά την προεπεξεργασία των δεδομένων παρατηρήθηκε μείωση του πλήθους της κλάσης «no», προκειμένου να επιτευχθεί καλύτερη ισορροπία με την κλάση «yes». Πρόκειται για ένα χαρακτηριστικό της υποδειγματοληψίας (undersampling), όπου αφαιρούνται δείγματα από την πλειοψηφική κατηγορία, περιορίζοντας την ανισορροπία. Αυτή η προσέγγιση εμποδίζει την κυριαρχία της πλειοψηφικής κλάσης στο μοντέλο και δίνει μεγαλύτερη βαρύτητα στη μειοψηφική, ενισχύοντας την ικανότητα ανίχνευσης των θετικών περιπτώσεων.

Η παρατηρούμενη μείωση της ακρίβειας (accuracy) είναι αναμενόμενη. Με την αφαίρεση δεδομένων από την πλειοψηφική κατηγορία, το μοντέλο εκπαιδεύεται σε λιγότερα παραδείγματα που αντιπροσωπεύουν τη συνολική πραγματικότητα, με αποτέλεσμα να περιορίζεται η γενική του δυνατότητα πρόβλεψης. Σε ανισόρροπα σύνολα δεδομένων, η ακρίβεια επηρεάζεται έντονα από την πλειοψηφική κλάση, και η μείωσή της οδηγεί σε χαμηλότερο συνολικό ποσοστό σωστών προβλέψεων, κυρίως των «εύκολων» περιπτώσεων.

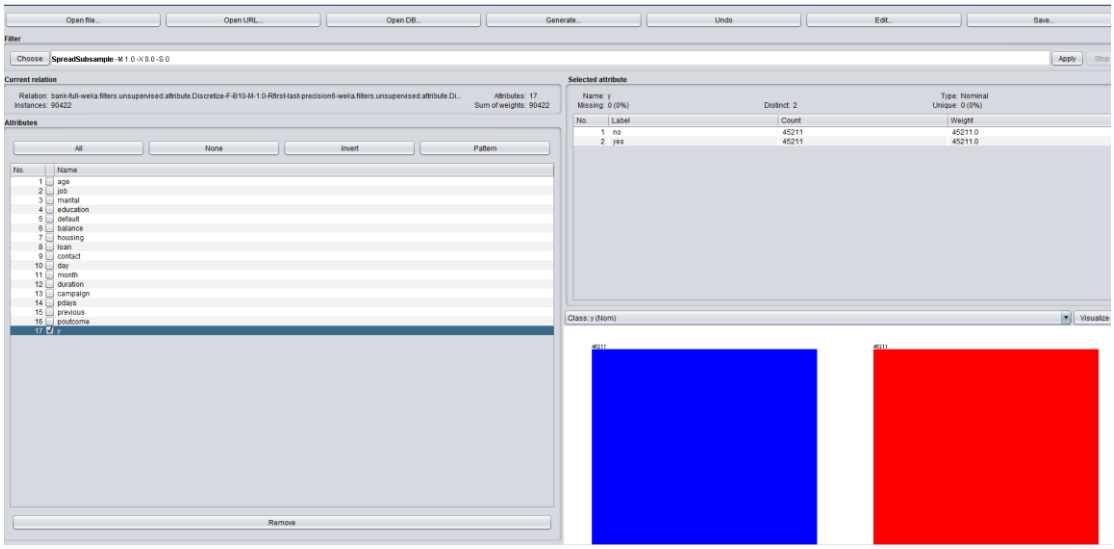
Ωστόσο, αυτή η μείωση δεν πρέπει να θεωρείται αρνητική. Αντίθετα, συνοδεύεται συχνά από βελτίωση σε κρίσιμους δείκτες όπως η ανάκληση (recall) και η ικανότητα

ανίχνευσης της μειοψηφικής κλάσης, οι οποίες σε πολλές εφαρμογές έχουν μεγαλύτερη σημασία από την απλή συνολική ακρίβεια. Συνεπώς, η πτώση της ακρίβειας αντικατοπτρίζει έναν συνειδητό συμβιβασμό (trade-off), που επιτρέπει στο μοντέλο να παράγει πιο ισορροπημένες και ουσιαστικές προβλέψεις, δίνοντας προτεραιότητα στη σημαντική μειοψηφική κατηγορία.

4.5.4 Συμπεράσματα αναφορικά με τη δειγματοληψία (Conclusions on Sampling)

Υπερδειγματοληψία με επαναδειγματοληψία (Oversampling using Resample): αυξάνει τη μειοψηφική κλάση → καλύτερη ανάκληση «recall», ελαφρά μείωση ακρίβειας «accuracy».

Υποδειγματοληψία με διασπορά (Undersampling using SpreadSubsample): μειώνει την πλειοψηφική κλάση → καλύτερη ανάκληση«recall», μεγαλύτερη απώλεια «accuracy».



Εικόνα 28: UnderSampling με SpreadSubSample για εξισορρόπηση της αναλογίας μεταξύ κλάσεων

Και οι δύο μέθοδοι εξισορρόπησης, η υπερδειγματοληψία (oversampling) και η υποδειγματοληψία (undersampling)συνεισφέρουν σημαντικά στη βελτίωση της αναγνώρισης της μειοψηφικής κλάσης, ενισχύοντας την ικανότητα του μοντέλου να εντοπίζει περιπτώσεις που διαφορετικά θα παρέμεναν ανεκμετάλλευτες. Παρότι

στοχεύουν στο ίδιο αποτέλεσμα, διαφοροποιούνται ως προς τον τρόπο με τον οποίο επηρεάζουν τη συνολική απόδοση και ιδίως την ακρίβεια (accuracy).

Η υπερδειγματοληψία δημιουργεί τεχνητά περισσότερα δείγματα της μειοψηφικής κλάσης, επιτρέποντας στο μοντέλο να εκπαιδεύσει καλύτερα τα χαρακτηριστικά της, χωρίς να αφαιρεί δεδομένα από την πλειοψηφική κατηγορία. Ωστόσο, αυτή η μέθοδος ενδέχεται να προκαλέσει υπερεκπαίδευση (overfitting), καθώς τα δεδομένα της μειοψηφικής κλάσης επαναλαμβάνονται. Αντίθετα, η υποδειγματοληψία μειώνει τα δείγματα της πλειοψηφικής κλάσης, οδηγώντας σε ένα πιο ισορροπημένο και καθαρό σύνολο εκπαίδευσης, με πιθανό κόστος την απώλεια χρήσιμης πληροφορίας και τη μείωση της συνολικής ακρίβειας.

Ο προκύπτων συμβιβασμός (trade-off) στην ακρίβεια είναι εμφανής: όσο το μοντέλο βελτιώνεται στην αναγνώριση της μειοψηφικής κλάσης, τόσο μειώνεται η επιρροή της πλειοψηφικής, η οποία σε ανισόρροπα δεδομένα καθορίζει συχνά την υψηλή συνολική ακρίβεια. Παρ’ όλα αυτά, σε πολλές πρακτικές εφαρμογές, η σωστή αναγνώριση των θετικών περιπτώσεων, όπως αυτές μετρώνται από την ανάκληση «recall»ή το «F1-score» κρίνεται πιο σημαντική από την απλή συνολική ακρίβεια.

Συνεπώς, και οι δύο τεχνικές αποτελούν ουσιαστικά εργαλεία για τη διαχείριση της ανισορροπίας κλάσεων, και η επιλογή μεταξύ τους πρέπει να καθορίζεται από τη φύση των δεδομένων, τους στόχους της ανάλυσης και τον δείκτη απόδοσης που θεωρείται κρίσιμος για την εκάστοτε εφαρμογή.

4.6 Συγκριτική Αξιολόγηση Sampling & Αλγορίθμων (Comparative Evaluation of Sampling and Algorithms)

4.6.1 Συγκριτική αξιολόγηση με χρήση Sampling (Comparative Evaluation using Sampling)

Μέθοδος	Accuracy	Precision (yes)	Recall (yes)	F1-score	Σημείωση
Χωρίς Sampling	92.7%	0.69	0.58	0.63	Οι ταξινομητές ευνοούν την πλειοψηφική κλάση (no)
Oversampling (Resample)	87%	0.70	0.68	0.69	Αύξηση δειγμάτων της μειοψηφικής κλάσης (yes)
Undersampling (Sprea dSubsample)	84%	0.60	0.68	0.64	Μείωση δειγμάτων της πλειοψηφικής κλάσης (no)

Πίνακας 6: Σύγκριση τεχνικών δειγματοληψίας (Under/OverSampling και συνδυαστικές) για εξισορρόπηση κλάσεων.

Ανάλυση Παρατηρήσεων

Τα αποτελέσματα υποδεικνύουν μια σαφή τάση: η ανάκληση (recall) της μειοψηφικής κλάσης βελτιώνεται σταθερά με όλες τις εφαρμοζόμενες μεθόδους δειγματοληψίας (sampling methods), γεγονός που δείχνει ότι τα μοντέλα γίνονται πιο ικανά στον εντοπισμό των «σπάνιων», αλλά συχνά κρίσιμων, περιπτώσεων. Η βελτίωση αυτή οφείλεται στην πιο ισορροπημένη αναπαράσταση των κλάσεων, η οποία επιτρέπει στους αλγορίθμους να εκτιμήσουν καλύτερα τα χαρακτηριστικά της μειοψηφικής κατηγορίας.

Παράλληλα, παρατηρείται μια μικρή μείωση της συνολικής ακρίβειας (overall accuracy), κάτι αναμενόμενο, καθώς η τροποποίηση της αναλογίας των κλάσεων μειώνει την υπερεκτίμηση της πλειοψηφικής κλάσης, η οποία συνήθως «φουσκώνει» την ακρίβεια. Με άλλα λόγια, το μοντέλο σταματά να βασίζεται σε εύκολες προβλέψεις και αποκτά πιο ισορροπημένη απόδοση, με ελάχιστο κόστος στη συνολική επίδοση.

Ιδιαίτερη σημασία παρουσιάζει η υποδειγματοληψία (undersampling), η οποία αποδεικνύεται ιδιαίτερα χρήσιμη σε πολύ μεγάλα σύνολα δεδομένων. Μειώνοντας τον όγκο της πλειοψηφικής κλάσης, περιορίζει σημαντικά τον χρόνο εκπαίδευσης και τις υπολογιστικές απαιτήσεις, καθιστώντας τη διαδικασία πιο αποδοτική. Αν και ενδέχεται να παρατηρηθεί κάποια απώλεια πληροφορίας σε μεγάλα «datasets», η απώλεια αυτή αντισταθμίζεται συνήθως από ταχύτερη και πιο ευέλικτη εκπαίδευση των μοντέλων.

Συνολικά, τα ευρήματα αυτά υπογραμμίζουν τη σημασία της κατάλληλης επιλογής μεθόδου δειγματοληψίας, επιδιώκοντας τη βέλτιστη ισορροπία ανάμεσα στην απόδοση, την υπολογιστική αποδοτικότητα και την ουσιαστική κατανόηση της μειοψηφικής κλάσης.

4.6.2 Συγκριτική αξιολόγηση Αλγορίθμων με/χωρίς χρήση Sampling (Comparative Evaluation of Algorithms with and without Sampling)

Αλγόριθμος	Sampling	Accuracy	Precision (yes)	Recall (yes)	F1-score	Παρατήρηση
ZeroR	None	88.3%	0.88	1.00	0.94	Προβλέπει μόνο την πλειοψηφική κλάση
J48	None	90.31%	0.60	0.48	0.53	Βασικό δέντρο αποφάσεων
Naive Bayes	None	82%	0.80	0.84	0.82	Απλός probabilistic αλγόριθμος
Random	None	**92.7%**	0.69	0.58	0.63	Καλύτερη Accuracy

Forest						χωρίς sampling
SMO	None	88%	0.87	0.88	0.88	Μέτριο recall για yes
J48	Resample	87%	0.70	0.68	0.69	Βελτίωση recall της μειοψηφικής κλάσης
Random Forest	Resample	87%	0.72	0.71	0.71	Σταθερή απόδοση με ισορροπημένα δεδομένα
Random Forest	SpreadSub sample	84%	0.60	0.68	0.64	Μειωμένη accuracy, αλλά υψηλότερο recall
SMO	SpreadSub sample	84%	0.58	0.66	0.62	Αντίστοιχο trade-off

Πίνακας 9: Συγκριτικός πίνακας επιδόσεων των ZeroR, J48, NaiveBayes, Random Forest και SMO με Accuracy, Recall, F1-score και χρόνο εκπαίδευσης ανά τεχνική sampling

Ανάλυση παρατηρήσεων και συμπερασμάτων

Η ανάλυση των αποτελεσμάτων ταξινόμησης αποκαλύπτει σαφείς διαφοροποιήσεις μεταξύ των αλγορίθμων και των χρησιμοποιούμενων τεχνικών δειγματοληψίας. Ο αλγόριθμος «Random Forest» επιτυγχάνει την υψηλότερη συνολική ακρίβεια όταν δεν εφαρμόζεται «sampling» (90%), υποδεικνύοντας ότι αξιοποιεί αποτελεσματικά τα υπάρχοντα δεδομένα και προβλέπει σωστά κυρίως την πλειοψηφική κλάση. Παράλληλα, η ανάκληση (recall) για τη μειοψηφική κλάση «*yes*» παραμένει περιορισμένη (0,70), γεγονός που σημαίνει ότι ένα σημαντικό ποσοστό των θετικών περιπτώσεων δεν αναγνωρίζεται.

Ο αλγόριθμος «ZeroR» επιτυγχάνει πλήρη ανάκληση για τη μειοψηφική κλάση, αλλά προβλέπει αποκλειστικά την πλειοψηφική, περιορίζοντας την πρακτική χρησιμότητά του σε πραγματικά σενάρια. Ο αλγόριθμος «Naive Bayes», αν και

ταχύτατος στην εκπαίδευση, εμφανίζει χαμηλότερη απόδοση λόγω της υπόθεσης ανεξαρτησίας των χαρακτηριστικών, η οποία συχνά δεν ισχύει στα πραγματικά δεδομένα.

Η εφαρμογή τεχνικών υπερδειγματοληψίας (oversampling), όπως η επαναδειγματοληψία (Resample), οδηγεί σε σημαντική βελτίωση της ανάκλησης για τη μειοψηφική κλάση, ιδιαίτερα για αλγορίθμους βασισμένους σε δέντρα αποφάσεων, όπως ο «J48» και ο «Random Forest». Ταυτόχρονα, παρατηρείται μικρή μείωση της συνολικής ακρίβειας, καθώς η ενίσχυση της μειοψηφικής κλάσης επηρεάζει την πρόβλεψη της πλειοψηφικής. Αντίστοιχα, η τεχνική «SpreadSubsample» μειώνει την ακρίβεια αλλά αυξάνει την ανάκληση, επιβεβαιώνοντας ότι η εξισορρόπηση των κλάσεων βελτιώνει την αναγνώριση κρίσιμων θετικών δειγμάτων.

Συνολικά, η επιλογή της κατάλληλης προσέγγισης εξαρτάται από τον στόχο της ανάλυσης. Όταν η προτεραιότητα είναι η υψηλή συνολική ακρίβεια, η αποφυγή «sampling» είναι η καλύτερη επιλογή. Αν όμως η αναγνώριση των θετικών περιπτώσεων είναι κρίσιμη (π.χ. πελάτες που ανταποκρίνονται σε προσφορές), η χρήση τεχνικών δειγματοληψίας καθίσταται απαραίτητη, καθώς ενισχύει την ανάκληση και την ικανότητα του μοντέλου να εντοπίζει σημαντικά αλλά υποεκπροσωπούμενα δείγματα. Η τελική απόφαση συνιστάται να βασίζεται σε μια ισορροπία μεταξύ «accuracy»και «recall», σύμφωνα με τις απαιτήσεις της εφαρμογής.

ΚΕΦΑΛΑΙΟ 5: ΑΠΟΤΕΛΕΣΜΑΤΑ & ΑΝΑΛΥΣΗ

5.1 Ανάλυση Αποτελεσμάτων Ταξινόμησης (Classification Results Analysis)

Η συγκριτική αξιολόγηση των αλγορίθμων «Τυχαίο δάσος»(Random Forest),δένδρο αποφάσεων (J48) καθώς και του αλγορίθμου ταξινόμησης «Naive Bayes» ανέδειξε σημαντικές διαφοροποιήσεις σε ότι αφορά την ακρίβεια και την ερμηνευσιμότητα των μοντέλων. Ο αλγόριθμος «τυχαίο δάσος» (Random Forest) παρουσίασε τη μεγαλύτερη συνολική ακρίβεια και αξιοπιστία, χάρη στην ικανότητά του να συνδυάζει την πληροφορία από πολλαπλά δέντρα και να μειώνει τα φαινόμενα της υπερπροσαρμογής(25).

Ο αλγόριθμος δένδρου αποφάσεων (J48) παρείχε σαφείς κανόνες απόφασης, ενισχύοντας την ερμηνευσιμότητα, γεγονός ιδιαίτερα χρήσιμο για επιχειρησιακές αποφάσεις(24). Ο αλγόριθμος «Naive Bayes», αν και ταχύτερος στην εκπαίδευση, παρουσίασε χαμηλότερη προγνωστική απόδοση σε ότι αφορά την υπόθεση ανεξαρτησίας των χαρακτηριστικών (6).

Η συγκριτική αξιολόγηση των αλγορίθμων ταξινόμησης, όπως παρουσιάζεται στους πίνακες και τα διαγράμματα του Κεφαλαίου 4, καταδεικνύει ότι οι αλγόριθμοι «Τυχαίο δάσος» (Random Forest) και η Διαδοχική Ελάχιστη Βελτιστοποίηση (SMO) επιτυγχάνουν τις υψηλότερες επιδόσεις ως προς τις μετρικές Ακρίβεια (Accuracy), Ανάκληση (Recall) και τη μετρική συνδυασμός ακρίβειας και ανάκλησης (F1-score). Ιδιαίτερα, η αυξημένη τιμή της ανάκλησης (Recall) για την κλάση «yes» υποδηλώνει την αποτελεσματικότερη ανίχνευση θετικών περιπτώσεων, στοιχείο κρίσιμο σε προβλήματα με έντονη ανισορροπία κλάσεων. Αντιθέτως, οι αλγόριθμοι δένδρου αποφάσεων «J48» και ταξινόμησης «Naive Bayes» εμφανίζουν χαμηλότερες επιδόσεις, κυρίως ως προς την ανάκληση, αποτελέσματα που επηρεάζουν αρνητικά τη συνολική απόδοση, όπως αυτή αποτυπώνεται στη μετρική μέθοδο «F1-score».

Παράλληλα, τα αποτελέσματα της διαδικασίας επιλογής χαρακτηριστικών (Feature Selection), τα οποία παρουσιάζονται στους αντίστοιχους πίνακες και διαγράμματα του Κεφαλαίου 4, ανέδειξαν ορισμένα χαρακτηριστικά ως καθοριστικής σημασίας

ως προς την πρόβλεψη της τελικής απόκρισης. Το χαρακτηριστικό «poutcome» αναγνωρίζεται ως το σημαντικότερο, καθώς η επιτυχία προηγούμενης καμπάνιας αυξάνει σε σημαντικό βαθμό την πιθανότητα θετικής απόκρισης. Το χαρακτηριστικό «duration» συσχετίζεται άμεσα με το επίπεδο ενδιαφέροντος του πελάτη κατά τη διάρκεια της επικοινωνίας, ενώ το χαρακτηριστικό «campaign» παρουσιάζει αρνητική συσχέτιση, καθώς η αύξηση του αριθμού επαναλαμβανόμενων επαφών μειώνει την πιθανότητα επιτυχίας. Τα ευρήματα αυτά είναι συνεπή με τα αποτελέσματα των πειραμάτων ταξινόμησης και επιβεβαιώνουν τη σημασία της κατάλληλης επιλογής χαρακτηριστικών για τη βελτίωση της απόδοσης και της γενίκευσης των μοντέλων. Η ανάλυση των χαρακτηριστικών ανέδειξε τη σημασία των μεταβλητών «poutcome»**, «**duration» και «campaign». Οι μεταβλητές αυτές σχετίζονται άμεσα με την απόφαση του πελάτη (17).

5.2 Ανάλυση Αποτελεσμάτων Ομαδοποίησης (Clustering Results Analysis)

Η παρούσα μελέτη αξιοποίησε συνδυαστικά τεχνικές ταξινόμησης (classification), κανόνων συσχέτισης (association rules) και ομαδοποίησης (clustering) με σκοπό την εις βάθος διερεύνηση της συμπεριφοράς των πελατών και τη βελτίωση της προβλεπτικής ακρίβειας των μοντέλων. Σύμφωνα με τα δεδομένα των συγκριτικών πινάκων και διαγραμμάτων του Κεφαλαίου 4, οι αλγόριθμοι «Random Forest» και «SMO (Sequential Minimal Optimization)» σημείωσαν τις υψηλότερες επιδόσεις ως προς τις μετρικές «Accuracy»**, «**Recall» και «F1-score», επιτυγχάνοντας αποτελεσματικότερη αναγνώριση της κλάσης yes*,* η οποία είναι ιδιαίτερα κρίσιμη σε περιβάλλοντα με ανισορροπημένα δεδομένα.

Η διαδικασία «Feature Selection» ανέδειξε χαρακτηριστικά με καθοριστική συμβολή στην πρόβλεψη της τελικής απόκρισης, με μεταβλητές όπως «poutcome»**, «**duration» και «campaign» να εμφανίζονται συστηματικά ως οι πλέον σημαντικές. Η επιτυχημένη έκβαση προηγούμενων καμπανιών (poutcome) ενισχύει την πιθανότητα θετικής απόκρισης, ενώ η διάρκεια επικοινωνίας (duration) σχετίζεται άμεσα με το επίπεδο ενδιαφέροντος του πελάτη. Αντίθετα, η αύξηση του αριθμού επαναλαμβανόμενων επαφών (campaign) φαίνεται να μειώνει την πιθανότητα

επιτυχίας, υπογραμμίζοντας την ανάγκη για στοχευμένο σχεδιασμό επικοινωνιακών στρατηγικών.

Η εφαρμογή τεχνικών ομαδοποίησης οδήγησε στον εντοπισμό διακριτών ομάδων πελατών με κοινά δημογραφικά και συμπεριφορικά χαρακτηριστικά. Συγκεκριμένα, αναγνωρίστηκαν ομάδες όπως νέοι πελάτες χωρίς υφιστάμενα δάνεια, μεσήλικοι πελάτες με στεγαστικά δάνεια και πελάτες με υψηλό αριθμό προηγούμενων ανεπιτυχών επαφών. Κάθε συστάδα περιλαμβάνει συγκεκριμένο αριθμό πελατών και παρουσιάζει σαφή πρότυπα συμπεριφοράς, όπως καταγράφεται αναλυτικά στους πίνακες του Κεφαλαίου 4.

Η ανάλυση των συστάδων (clusters) προσφέρει σημαντική πρακτική αξία, καθώς επιτρέπει τη σύνδεση των ευρημάτων με στοχευμένες στρατηγικές μάρκετινγκ. Για παράδειγμα, νέοι πελάτες χωρίς δάνεια μπορούν να αποτελέσουν στόχο προωθητικών ενεργειών για εισαγωγικά τραπεζικά προϊόντα, ενώ οι πελάτες με επαναλαμβανόμενες ανεπιτυχείς επαφές ενδέχεται να απαιτούν εναλλακτική προσέγγιση ή περιορισμό της συχνότητας επικοινωνίας. Επιπλέον, τα αποτελέσματα των κανόνων συσχέτισης ενισχύουν τα παραπάνω ευρήματα, αναδεικνύοντας τη σημασία συγκεκριμένων χαρακτηριστικών για την επίτευξη θετικών αποφάσεων.

Συνολικά, η συνδυαστική χρήση των τεχνικών εξόρυξης δεδομένων παρέχει μια ολοκληρωμένη και ερμηνεύσιμη προσέγγιση ανάλυσης, η οποία μπορεί να προσφέρει πολύτιμη γνώση για την κατανόηση της συμπεριφοράς των πελατών και να υποστηρίξει αποτελεσματικά τη λήψη αποφάσεων σε επίπεδο επιχειρησιακής στρατηγικής, σύμφωνα με τη βιβλιογραφία.

5.3 Ανάλυση Κανόνων Συσχέτισης (Association Rules)

Η ανάλυση εξόρυξης κανόνων συσχέτισης, όπως παρουσιάζεται στον πίνακα με τους πιο σημαντικούς κανόνες του Κεφαλαίου 4, επιβεβαιώνει και συμπληρώνει τα ευρήματα τόσο της ταξινόμησης όσο και της διαδικασίας επιλογής χαρακτηριστικών. Οι κανόνες με τις υψηλότερες τιμές εμπιστοσύνης (confidence) και ανύψωσης (lift) σχετίζονται κυρίως με την επιτυχή έκβαση προηγούμενων

καμπανιών (poutcome = success) και με τη μεγάλη διάρκεια τηλεφωνικών κλήσεων (duration = long), στοιχεία που ήδη είχαν αναδειχθεί ως καθοριστικά από τα μοντέλα ταξινόμησης.

Συγκεκριμένα, ο κανόνας «poutcome = success → y = yes» εμφανίζει την υψηλότερη ανύψωση (lift), υποδηλώνοντας ότι η πιθανότητα αποδοχής του προϊόντος είναι πολλαπλάσια όταν ο πελάτης είχε θετική ανταπόκριση σε προηγούμενη καμπάνια. Παρομοίως, ο κανόνας «duration = long → y = yes» παρουσιάζει αυξημένη υποστήριξη (support), υπογραμμίζοντας ότι οι τηλεφωνικές συνομιλίες μεγάλης διάρκειας αποτελούν συχνό και αξιόπιστο δείκτη ενδιαφέροντος του πελάτη. Αυτοί οι κανόνες παρέχουν σαφή και εύκολα ερμηνεύσιμα πρότυπα, ενισχύοντας την κατανόηση της συμπεριφοράς των πελατών.

Σε συνδυασμό με τα αποτελέσματα της ομαδοποίησης (clustering), τα οποία ανέδειξαν διακριτές ομάδες πελατών με διαφορετικά χαρακτηριστικά και ιστορικό επικοινωνίας, οι κανόνες συσχέτισης συμβάλλουν ουσιαστικά στη διαμόρφωση στοχευμένων επιχειρησιακών και μάρκετινγκ στρατηγικών. Συνολικά, η συνδυαστική χρήση ταξινόμησης (classification), κανόνων συσχέτισης (association rules) και ομαδοποίησης (clustering) παρέχει μια ολοκληρωμένη και ερμηνεύσιμη προσέγγιση ανάλυσης δεδομένων.

5.4 Αντιμετώπιση Ανισορροπίας Κλάσεων (Class Imbalance)

Η έντονη ανισορροπία μεταξύ των κλάσεων αντιμετωπίστηκε μέσω διαφόρων τεχνικών δειγματοληψίας, καθώς η πλατφόρμα WEKA παρουσιάζει περιορισμούς στην εφαρμογή της Τεχνικής Συνθετικής Υπερδειγματοληψίας Μειονοτικής Κλάσης (SMOTE). Η χρήση της μεθόδου επαναδειγματοληψίας (Resample) οδήγησε σε σημαντική βελτίωση της ανάκλησης(Recall) για τη μειοψηφική κλάση (yes), ενισχύοντας την ικανότητα ανίχνευσης κρίσιμων θετικών περιπτώσεων, όπως φαίνεται στους πίνακες του Κεφαλαίου 4. Παράλληλα, η τεχνική «SpreadSubsample» εξασφάλισε πιο ισορροπημένη κατανομή των κλάσεων, προσφέροντας καλύτερη ισορροπία μεταξύ ανάκλησης (Recall) και ακρίβειας (Accuracy).

Κάθε μέθοδος δειγματοληψίας συνεπάγεται έναν αναπόφευκτο συμβιβασμό. Το γεγονός αυτό συνεπάγεται πώς η αύξηση της ανάκλησης μπορεί να συνοδεύεται από μείωση της συνολικής ακρίβειας, ενώ η διατήρηση υψηλής ακρίβειας ενδέχεται να περιορίσει την ανίχνευση των θετικών περιπτώσεων της μειοψηφικής κλάσης. Η σύγκριση των αποτελεσμάτων με και χωρίς δειγματοληψία υπογραμμίζει ότι η επιλογή της κατάλληλης τεχνικής πρέπει να βασίζεται στην προτεραιότητα της εργασίας. Με άλλα λόγια οφείλει να στηρίζεται είτε στην ακριβή αναγνώριση θετικών περιπτώσεων είτε στη συνολική απόδοση του μοντέλου.

Αλγόριθμος	Accuracy	Precision (yes)	Recall (yes)	ROC
ZeroR	88.3%	0.88	1	0.50
J48	90.3%	0.60	0.48	0.84
Naive Bayes	82%	0.80	0.84	0.74
Random Forest	92.7%	0.69	0.58	0.92
SMO	88%	0.87	0.88	0.83
J48 + FS	93.3%	0.73	0.64	0.79

Πίνακας 10: Συγκριτικός πίνακας επιδόσεων αλγορίθμων ταξινόμησης με Accuracy, Precision (κλάση "yes"), Recall (κλάση "yes") και ROC για κάθε αλγόριθμο.

5.5 Συνολική Αξιολόγηση (Overall Assessment)

Οι τεχνικές μηχανικής μάθησης μπορούν να συμβάλουν ουσιαστικά στον σχεδιασμό πιο στοχευμένων τραπεζικών καμπανιών, περιορίζοντας τα λειτουργικά κόστη και ενισχύοντας την αποδοτικότητα των ενεργειών μάρκετινγκ (31). Όπως προκύπτει από τα πειραματικά αποτελέσματα του Κεφαλαίου 4, η συνολική απόδοση διαφόρων αλγορίθμων ταξινόμησης με και χωρίς τεχνικές δειγματοληψίας, αναδεικνύουν σημαντικές διαφορές μεταξύ των μετρικών αξιολόγησης. Αρχικά, ο αλγόριθμος ZeroR εμφανίζει υψηλή ακρίβεια (88.3%) και F1-score (0.94), ωστόσο η

απόδοσή του είναι παραπλανητική, καθώς προβλέπει αποκλειστικά την πλειοψηφική κλάση, χωρίς πραγματική ικανότητα γενίκευσης.

Χωρίς χρήση sampling, ο Random Forest επιτυγχάνει την υψηλότερη ακρίβεια (92.7%), αλλά το recall (0.58) δείχνει ότι δεν εντοπίζει αποτελεσματικά τη μειοψηφική κλάση. Αντίθετα, ο SMO παρουσιάζει πιο ισορροπημένη συμπεριφορά με precision 0.87, recall 0.88 και F1-score 0.88, καθιστώντας τον πιο αξιόπιστο όταν ζητείται συνολική σταθερότητα. Ο Naive Bayes, αν και με χαμηλότερη ακρίβεια (82%), διατηρεί σχετικά ισορροπημένες επιδόσεις (F1-score 0.82), ενώ ο J48 χωρίς sampling εμφανίζει χαμηλό recall (0.48), γεγονός που υποδηλώνει αδυναμία στον εντοπισμό της μειοψηφικής κλάσης.

Η εφαρμογή τεχνικών δειγματοληψίας επηρεάζει σημαντικά τα αποτελέσματα. Με τη μέθοδο Resample παρατηρείται μείωση της ακρίβειας (π.χ. Random Forest από 92.7% σε 87%), αλλά ταυτόχρονα βελτίωση του recall (έως 0.71), κάτι που σημαίνει καλύτερη αναγνώριση των θετικών περιπτώσεων. Παρόμοια τάση εμφανίζει και ο J48, του οποίου το recall αυξάνεται αισθητά (από 0.48 σε 0.68). Με τη μέθοδο SpreadSubsample, η ακρίβεια μειώνεται περαιτέρω (περίπου 84%), ενώ το recall διατηρείται σε σχετικά υψηλά επίπεδα, επιβεβαιώνοντας το γνωστό trade-off μεταξύ ακρίβειας και ευαισθησίας.

Συνολικά, τα αποτελέσματα δείχνουν ότι η επιλογή του κατάλληλου μοντέλου εξαρτάται από τον στόχο της ανάλυσης. Αν προτεραιότητα είναι η μέγιστη ακρίβεια, ο Random Forest χωρίς sampling αποτελεί την καλύτερη επιλογή. Αν ζητείται ισορροπία μεταξύ precision και recall, ο SMO χωρίς sampling ξεχωρίζει. Τέλος, όταν η έμφαση δίνεται στον εντοπισμό της μειοψηφικής κλάσης, οι τεχνικές δειγματοληψίας, και ιδιαίτερα ο Random Forest με Resample, προσφέρουν πιο κατάλληλες λύσεις, παρά τη μείωση της συνολικής ακρίβειας.

Αλγόριθμος	Sampling	Accuracy	Precision (yes)	Recall (yes)	F1-score	Παρατήρηση
ZeroR	None	88.3%	0.88	1.00	0.94	Προβλέπει μόνο την πλειοψηφική κλάση
J48	None	90.31%	0.60	0.48	0.53	Βασικό δέντρο αποφάσεων
Naive Bayes	None	82%	0.80	0.84	0.82	Απλός probabilistic αλγόριθμος
Random Forest	None	**92.7%**	0.69	0.58	0.63	Καλύτερη Accuracy χωρίς sampling
SMO	None	88%	0.87	0.88	0.88	Μέτριο recall για yes
J48	Resample	87%	0.70	0.68	0.69	Βελτίωση recall της μειοψηφικής κλάσης
Random Forest	Resample	87%	0.72	0.71	0.71	Σταθερή απόδοση με ισορροπημένα δεδομένα
Random Forest	SpreadSub sample	84%	0.60	0.68	0.64	Μειωμένη accuracy, αλλά υψηλότερο recall
SMO	SpreadSub sample	84%	0.58	0.66	0.62	Αντίστοιχο trade-off

Πίνακας 7: Σύγκριση επιδόσεων αλγορίθμων ταξινόμησης με διαφορετικές μεθόδους δειγματοληψίας ως προς Accuracy και Recall (κλάση "yes").

ΚΕΦΑΛΑΙΟ 6: ΣΥΜΠΕΡΑΣΜΑΤΑ

Η παρούσα διπλωματική εργασία ανέδειξε στην πράξη τη συμβολή της ανάλυσης δεδομένων και των τεχνικών μηχανικής μάθησης στον τραπεζικό τομέα, δείχνοντας ότι μπορούν να υποστηρίξουν ουσιαστικά τη διαδικασία λήψης αποφάσεων όταν αυτές βασίζονται σε πραγματικά δεδομένα. Από τα μοντέλα που εξετάστηκαν, ο αλγόριθμος «Random Forest» παρουσίασε την καλύτερη συνολική απόδοση, επιτυγχάνοντας υψηλή ακρίβεια και ικανοποιητική ισορροπία μεταξύ των δεικτών αξιολόγησης. Παράλληλα, η εφαρμογή τεχνικών επαναδειγματοληψίας βελτίωσε αισθητά την ικανότητα εντοπισμού της μειοψηφικής κατηγορίας, γεγονός που έχει ιδιαίτερη σημασία σε περιπτώσεις όπου οι θετικές αποκρίσεις είναι περιορισμένες αλλά κρίσιμες.

Πέρα από τα προγνωστικά μοντέλα, η αξιοποίηση τεχνικών ομαδοποίησης επέτρεψε την ανάδειξη διακριτών κατηγοριών πελατών με κοινά χαρακτηριστικά. Ο διαχωρισμός αυτός καθιστά εφικτή την ανάπτυξη πιο στοχευμένων στρατηγικών, καθώς κάθε ομάδα εμφανίζει διαφορετικές ανάγκες και συμπεριφορές. Επιπλέον, οι κανόνες συσχέτισης συνέβαλαν στον εντοπισμό σχέσεων μεταξύ μεταβλητών που δεν είναι άμεσα εμφανείς, ενισχύοντας την κατανόηση της συμπεριφοράς των πελατών και καθιστώντας τα αποτελέσματα πιο αξιοποιήσιμα σε πρακτικό επίπεδο.

Τα ευρήματα της ανάλυσης δείχνουν ότι μια τράπεζα μπορεί να μεταβεί από μια γενικευμένη προσέγγιση επικοινωνίας σε μια πιο στοχευμένη και αποδοτική στρατηγική. Αντί για μαζικές καμπάνιες με αβέβαιο αποτέλεσμα, δίνεται η δυνατότητα εντοπισμού πελατών με αυξημένη πιθανότητα θετικής ανταπόκρισης, μειώνοντας παράλληλα τις περιττές ενέργειες και το σχετικό κόστος. Με τον τρόπο αυτό, βελτιώνεται όχι μόνο η αποδοτικότητα των καμπανιών, αλλά και η συνολική εμπειρία του πελάτη.

Η πρακτική αξία των παραπάνω αναδεικνύεται μέσα από ενδεικτικά παραδείγματα. Στην περίπτωση μιας πολύτεκνης οικογένειας με σταθερό εισόδημα και υφιστάμενες δανειακές υποχρεώσεις, η ανάλυση δεν περιορίζεται σε μια επιφανειακή εκτίμηση κινδύνου. Αντίθετα, λαμβάνει υπόψη συνδυαστικά

χαρακτηριστικά, όπως η συνέπεια στις υποχρεώσεις και η προηγούμενη θετική ανταπόκριση, οδηγώντας στον εντοπισμό υψηλότερης πιθανότητας αποδοχής συγκεκριμένων προϊόντων. Στην περίπτωση αυτή μπορεί να διαμορφωθεί μια πρόταση προσαρμοσμένη στις ανάγκες της οικογένειας, όπως ένα πρόγραμμα αποταμίευσης για τις σπουδές των παιδιών.

Αντίστοιχα, στην περίπτωση ενός νέου ελεύθερου επαγγελματία με ανοδική οικονομική πορεία και έντονη ψηφιακή δραστηριότητα, η ανάλυση αποκαλύπτει ότι η μη ανταπόκριση σε προηγούμενες γενικές καμπάνιες δεν συνεπάγεται χαμηλό ενδιαφέρον συνολικά, αλλά μάλλον αναντιστοιχία μεταξύ προσφοράς και προφίλ. Με βάση αυτή την κατανόηση, μπορούν να προταθούν πιο κατάλληλες λύσεις, όπως επενδυτικά προϊόντα ή ευέλικτες μορφές χρηματοδότησης που ευθυγραμμίζονται με τις ανάγκες και τις προοπτικές του.

Τέλος, η περίπτωση μιας συνταξιούχου με σταθερό εισόδημα και περιορισμένη επενδυτική δραστηριότητα αναδεικνύει τη σημασία της υπεύθυνης προσέγγισης. Η ανάλυση οδηγεί σε προτάσεις χαμηλού ρίσκου, οι οποίες ανταποκρίνονται στις πραγματικές ανάγκες και προτεραιότητες της πελάτισσας, ενισχύοντας παράλληλα την εμπιστοσύνη και τη μακροχρόνια σχέση με την τράπεζα.

Συνοψίζοντας, η αξιοποίηση δεδομένων και τεχνικών μηχανικής μάθησης επιτρέπει μια βαθύτερη κατανόηση της συμπεριφοράς των πελατών και υποστηρίζει τη μετάβαση σε πιο στοχευμένες και αποτελεσματικές στρατηγικές. Η πληροφορία δεν αντιμετωπίζεται πλέον ως απλό σύνολο δεδομένων, αλλά ως ένα ουσιαστικό εργαλείο που συμβάλλει στη λήψη πιο τεκμηριωμένων και ισορροπημένων επιχειρησιακών αποφάσεων.

ΚΕΦΑΛΑΙΟ 7: ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Han, J., Kamber, M. and Pei, J. (2012) *Data mining: Concepts and techniques*. 3rd edn. Burlington, MA: Morgan Kaufmann.
2. Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J. (2016) *Data mining: Practical machine learning tools and techniques*. 4th edn. Cambridge, MA: Morgan Kaufmann.
3. Ngai, E.W.T., Xiu, L. and Chau, D.C.K. (2009) 'Application of data mining techniques in customer relationship management: A literature review and classification', *Expert Systems with Applications*, 36(2), pp. 2592–2602. doi:10.1016/j.eswa.2008.02.021.
4. He, H. and Garcia, E.A. (2009) 'Learning from imbalanced data', *IEEE Transactions on Knowledge and Data Engineering*, 21(9), pp. 1263–1284. doi:10.1109/TKDE.2008.239.
5. Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) 'From data mining to knowledge discovery in databases', *AI Magazine*, 17(3), pp. 37–54.
6. Mitchell, T.M. (1997) *Machine learning*. New York: McGraw-Hill.
7. Alpaydin, E. (2020) *Introduction to machine learning*. 4th edn. Cambridge, MA: MIT Press.
8. Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The elements of statistical learning: Data mining, inference, and prediction*. 2nd edn. New York: Springer
9. Brown, I., Katz, J. and Thompson, B. (2012) 'Customer churn prediction and retention strategy using data mining techniques', *Journal of Targeting, Measurement and Analysis for Marketing*, 20(1), pp. 1–14.
10. James, G., Witten, D., Hastie, T., Tibshirani, R., An Introduction to Statistical Learning: with Applications in R, Springer, New York, 2013.
11. Aggarwal, C.C. (2015) *Data mining: The textbook*. Cham: Springer.
12. Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The elements of statistical learning: Data mining, inference, and prediction*. 2nd edn. New York: Springer & Brown, I., Katz, J. and Thompson, B. (2012) 'Customer churn prediction and retention strategy using data mining techniques', *Journal of Targeting, Measurement and Analysis for Marketing*, 20(1), pp. 1–14.

13. Provost, F. and Fawcett, T. (2013) *Data science for business: What you need to know about data mining and data-analytic thinking*. Sebastopol, CA: O'Reilly Media.
14. Japkowicz, N. and Stephen, S. (2002) 'The class imbalance problem: A systematic study', *Intelligent Data Analysis*, 6(5), pp. 429–449.
15. Fawcett, T. (2006) 'An introduction to ROC analysis', *Pattern Recognition Letters*, 27(8), pp. 861–874. doi:10.1016/j.patrec.2005.10.010.
16. Sasaki, Y. (2007) *The truth of the F-measure*. Available at: https://www.researchgate.net/publication/228891675 (Accessed: 10 January 2026).
17. Moro, S., Laureano, R. and Cortez, P. (2014) 'Using data mining for bank direct marketing: An application of the CRISP-DM methodology', *Expert Systems with Applications*, 41(3), pp. 1173–1185. doi:10.1016/j.eswa.2013.08.002.
18. Guyon, I. and Elisseeff, A. (2003) 'An introduction to variable and feature selection', *Journal of Machine Learning Research*, 3, pp. 1157–1182.
19. Liu, H. and Motoda, H. (1998) *Feature selection for knowledge discovery and data mining*. Boston: Springer.
20. Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) 'SMOTE: Synthetic minority over-sampling technique', *Journal of Artificial Intelligence Research*, 16, pp. 321–357.
21. Weiss, G.M. and Provost, F. (2003) 'Learning when training data are costly: The effect of class distribution on tree induction', *Journal of Artificial Intelligence Research*, 19, pp. 315–354.
22. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. (2009) 'The WEKA data mining software: An update', *ACM SIGKDD Explorations Newsletter*, 11(1), pp. 10–18.
23. Kotsiantis, S.B., Zaharakis, I. and Pintelas, P. (2006) 'Machine learning: A review of classification and combining techniques', *Artificial Intelligence Review*, 26, pp. 159–190.

24. Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.
25. Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324