

Bankruptcy Prediction Using Supervised Machine Learning

Coursework Project

2025

Πίνακας περιεχομένων

1. Εισαγωγή
2. Περιγραφή dataset
3. Προτεινόμενη Μεθοδολογία
4. Πειραματικά Αποτελέσματα
5. Συμπεράσματα
6. Παράρτημα
7. Βιβλιογραφία

Πίνακας εικόνων

Αριθμός εικόνας	Περιγραφή	Σελίδα
Figure 1	Κατανομή υγιών και πτωχευμένων επιχειρήσεων ανά έτος.	4
Figure 2	Σύγκριση βασικών χρηματοοικονομικών δεικτών (min, max, mean) μεταξύ υγιών και πτωχευμένων επιχειρήσεων, με χρήση λογαριθμικής κλίμακας.	5
Figure 3	Μέσος όρος της F1 Score για κάθε μοντέλο ταξινόμησης στο test set.	9
Figure 4	Ραβδόγραμμα που παρουσιάζει τον μέσο όρο της τιμής ROC-AUC για κάθε αλγόριθμο ταξινόμησης, βασισμένο στην αξιολόγηση του test set.	10
Figure 5	Διάγραμμα που συγκρίνει τον μέσο όρο της μετρικής Balanced Accuracy για καθένα από τα μοντέλα	10
Figure 6	Πίνακας με τις τιμές των μετρικών Recall και Specificity για κάθε μοντέλο ταξινόμησης.	12
Figure 7	Διάγραμμα με τις τιμές των μετρικών Recall και Specificity για κάθε μοντέλο ταξινόμησης.	13

Figure 8	Confusion Matrices των Random Forest και SVM για τα folds 1 και 2.	14
Figure 9	Confusion Matrices των Random Forest και SVM για τα folds 3 και 4.	15

1. Εισαγωγή

Στο πλαίσιο της παρούσας εργασίας, εξετάζεται το πρόβλημα της ταξινόμησης επιχειρήσεων ως υγιείς (μη πτωχευμένες) ή πτωχευμένες, με χρήση τεχνικών επιβλεπόμενης μάθησης. Η ικανότητα πρόβλεψης της οικονομικής σταθερότητας μιας επιχείρησης αποτελεί σημαντικό εργαλείο τόσο για επενδυτές όσο και για χρηματοπιστωτικούς οργανισμούς.

Τα δεδομένα που αξιοποιούνται περιλαμβάνουν μια σειρά από οικονομικούς δείκτες και χαρακτηριστικά για ένα πλήθος επιχειρήσεων και αφορούν την κατάστασή τους (ενεργή ή πτωχευμένη). Η πρόκληση εντοπίζεται τόσο στην επιλογή κατάλληλων χαρακτηριστικών όσο και στην αξιοποίηση μεθόδων ταξινόμησης που μπορούν να αποδώσουν με αξιοπιστία σε δεδομένα με ανισομερή κατανομή τάξεων (δηλαδή πολλές υγιείς και λίγες πτωχευμένες επιχειρήσεις).

Στόχος της εργασίας είναι:

- Να υλοποιηθεί και να αξιολογηθεί ένα πλήθος ταξινομητικών μοντέλων (π.χ. SVM, Random Forest, LDA κ.ά.)
- Να εξεταστεί η απόδοσή τους με βάση μετρικών αξιολόγησης (Accuracy, Recall, F1 Score, AUC)
- Να διερευνηθεί εάν κάποια από τα μοντέλα ικανοποιούν συγκεκριμένους επιχειρησιακούς περιορισμούς (π.χ. εντοπισμός πτωχευμένων εταιρειών με ακρίβεια $\geq 60\%$)
- Τελικά, να προταθεί το πιο κατάλληλο μοντέλο για την επίλυση του προβλήματος ταξινόμησης επιχειρήσεων με βάση τα οικονομικά τους χαρακτηριστικά.

Η εργασία δομείται με τρόπο ώστε να παρουσιάζει τα πειραματικά αποτελέσματα, και να συνοδεύεται από αναλυτική παρουσίαση γραφημάτων και συγκριτικών αξιολογήσεων.

2. Περιγραφή dataset

Το dataset που χρησιμοποιήθηκε σε αυτή την εργασία παρέχεται σε μορφή Excel και περιλαμβάνει χρηματοοικονομικές και λειτουργικές μεταβλητές για κάθε επιχείρηση. Κάθε γραμμή του αρχείου αντιστοιχεί σε μία εταιρεία για ένα συγκεκριμένο έτος. Οι στήλες του dataset καλύπτουν ένα εύρος από ποσοτικούς δείκτες απόδοσης,

δυναμικούς δείκτες δραστηριότητας, καθώς και την κατηγορία αν η επιχείρηση πτώχευσε ή όχι.

Δομή και Περιεχόμενο

Οι στήλες του αρχικού dataset διακρίνονται στις εξής κατηγορίες:

- **Αριθμητικοί δείκτες απόδοσης** (στήλες A έως H):
 - Π.χ.Αποδοτικότητα Ενεργητικού (ROA), Ρευστότητα, Υποχρεώσεις/Ενεργητικό κ.ά.
 - Οι τιμές κυμαίνονται σε διαφορετικές κλίμακες και περιέχουν θετικούς και αρνητικούς αριθμούς.
- **Δυναμικοί δείκτες δραστηριότητας** (στήλες I έως K):
 - Ένδειξη Εξαγωγών, Ένδειξη Εισαγωγών, Ένδειξη Αντιπροσωπειών
 - Οι τιμές είναι είτε **0** (όχι) είτε **1** (ναι)
- **Μεταβλητή στόχος** (στήλη L):
 - ΕΝΔΕΙΞΗ ΑΣΥΝΕΠΕΙΑΣ (=2) (v+1) — επαναορίστηκε ως target
 - Τιμή **1**: Υγιής επιχείρηση
 - Τιμή **2**: Πτωχευμένη επιχείρηση
- **Χρονική πληροφορία** (στήλη M):
 - Το πεδίο ΕΤΟΣ δείχνει το οικονομικό έτος στο οποίο αφορούν τα δεδομένα

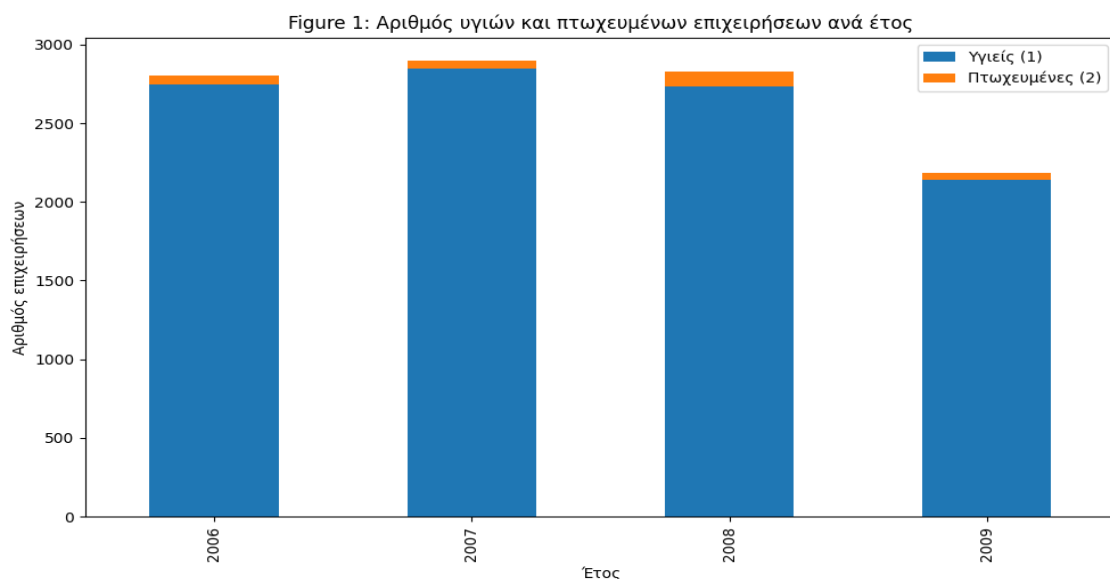


Figure 1: Κατανομή υγιών και πτωχευμένων επιχειρήσεων ανά έτος.

Η Εικόνα 1 απεικονίζει τον αριθμό των υγιών και πτωχευμένων επιχειρήσεων για κάθε έτος από το 2006 έως και το 2009. Παρατηρείται ότι το πλήθος των υγιών επιχειρήσεων παραμένει σταθερά υψηλό σε όλα τα έτη, ενώ ο αριθμός των πτωχευμένων εταιρειών είναι σημαντικά μικρότερος, γεγονός που επιβεβαιώνει την ανισορροπία του dataset. Το 2007 καταγράφεται το μέγιστο πλήθος καταχωρήσεων,

ενώ το 2009 το ελάχιστο, πιθανότατα λόγω περιορισμένης διαθεσιμότητας δεδομένων. Η σταθερή αυτή απόκλιση μεταξύ των δύο κατηγοριών καθιστά απαραίτητη τη χρήση τεχνικών εξισορρόπησης των τάξεων (όπως εφαρμόστηκε στο training set), ώστε να διασφαλιστεί η σωστή εκπαίδευση των μοντέλων ταξινόμησης.

Figure 2: Min, Max, Mean ανά δείκτη (Log scale)

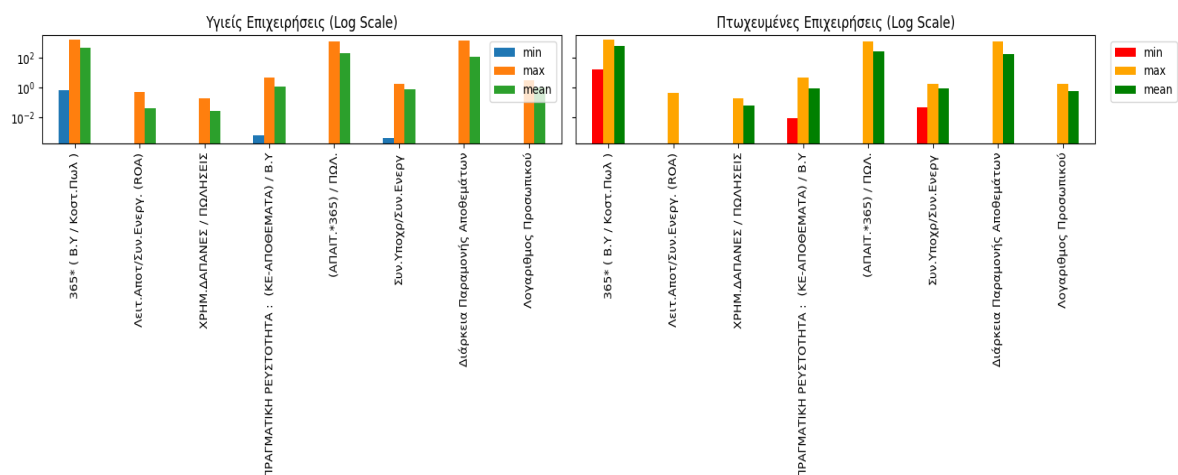


Figure 2: Σύγκριση στατιστικών (min, max, mean) για τους βασικούς οικονομικούς δείκτες υγιών και πτωχευμένων επιχειρήσεων.

Η Εικόνα 2 παρουσιάζει τις στατιστικές τιμές (ελάχιστη, μέγιστη και μέση τιμή) για οκτώ βασικούς χρηματοοικονομικούς δείκτες, διαχωρισμένους σε δύο ομάδες: υγιείς και πτωχευμένες επιχειρήσεις. Η χρήση λογαριθμικής κλίμακας καθιστά ευδιάκριτες τις σημαντικές διαφοροποιήσεις μεταξύ των ομάδων. Οι υγιείς επιχειρήσεις εμφανίζουν πιο σταθερές τιμές, ενώ οι πτωχευμένες χαρακτηρίζονται από μεγαλύτερη διασπορά, κάτι που αντανακλά τη χρηματοοικονομική αστάθεια που τις διακρίνει.

3. Προτεινόμενη Μεθοδολογία

Η προτεινόμενη μεθοδολογία για την αντιμετώπιση του προβλήματος ταξινόμησης επιχειρήσεων ως υγιείς ή πτωχευμένες βασίζεται σε τεχνικές επιβλεπόμενης μάθησης και περιλαμβάνει μια σειρά από βήματα επεξεργασίας, ανάλυσης και αξιολόγησης.

A) Διαδικασία Επεξεργασίας Δεδομένων

Τα αρχικά δεδομένα περιείχαν χρηματοοικονομικούς δείκτες και χαρακτηριστικά δραστηριότητας για κάθε επιχείρηση. Πραγματοποιήθηκε εξεργασία ως εξής :

- Η στήλη 'ΕΝΔΕΙΞΗ ΑΣΥΝΕΠΕΙΑΣ (=2) (v+1)' μετονομάστηκε σε 'target', ενώ η στήλη του έτους 'ΕΤΟΣ' μετονομάστηκε σε 'year'.
- Υπολογίστηκαν και απεικονίστηκαν σε stacked bar chart οι αριθμοί υγιών και πτωχευμένων επιχειρήσεων ανά έτος.

- Για τις δύο κατηγορίες (target = 1 και target = 2), υπολογίστηκαν οι τιμές ελάχιστης, μέγιστης και μέσης τιμής για κάθε αριθμητικό δείκτη.
- Οι τιμές αυτές παρουσιάστηκαν γραφικά σε **λογαριθμική κλίμακα** για λόγους οπτικής ευκρίνειας.
- Πραγματοποιήθηκε έλεγχος για NaN τιμές με χρήση της βιβλιοθήκης **Pandas**. Το dataset δεν περιείχε καμία ελλιπή εγγραφή, επομένως δεν απαιτήθηκε περαιτέρω διαχείριση.
- Οι οκτώ πρώτοι δείκτες απόδοσης κανονικοποιήθηκαν στο διάστημα [0, 1] με χρήση της τεχνικής **Min-Max Scaling**, με στόχο την ομοιομορφία κλίμακας πριν την εφαρμογή των αλγορίθμων ταξινόμησης.

B) Εφαρμογή Τεχνικών Ταξινόμησης

Για την εκπαίδευση και αξιολόγηση των μοντέλων, εφαρμόστηκε **Stratified K-Fold Cross-Validation** με 4 διαμερίσεις (folds). Σε κάθε fold:

- Το training set ισορροπήθηκε ώστε η αναλογία υγιών/πτωχευμένων να είναι 3:1
- Δεν μεταφέρθηκαν οι “περισσευούμενες” υγιείς στο test set

Τα μοντέλα που εκπαιδεύτηκαν ήταν τα εξής:

- Linear Discriminant Analysis (LDA)
- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- k-Nearest Neighbors (k-NN)
- Naïve Bayes
- Support Vector Machines (SVM)
- Multi-Layer Perceptron (MLP – Νευρωνικό Δίκτυο)

Αιτιολόγηση Επιλογής MLP για το παρόν πρόβλημα

Η επιλογή του Multi-Layer Perceptron (MLP) ως ταξινομητικό μοντέλο βασίστηκε σε συγκεκριμένα χαρακτηριστικά του προβλήματος.

Το πρόβλημα χαρακτηρίζεται από έντονη ανισορροπία μεταξύ των τάξεων (πολύ περισσότερες υγιείς από πτωχευμένες επιχειρήσεις). Παρόλο που εφαρμόστηκε τεχνική εξισορρόπησης του training set, η ανάγκη για μοντέλο που μπορεί να διαχειριστεί μειοψηφούσες κλάσεις παραμένει κρίσιμη. Το MLP, όταν συνδυαστεί με κατάλληλη κανονικοποίηση των δεδομένων και παραμετροποίηση (π.χ. επιλογή αριθμού νευρώνων, learning rate), μπορεί να μάθει να εντοπίζει patterns που χαρακτηρίζουν τις πτωχευμένες επιχειρήσεις, ακόμη και με περιορισμένη παρουσία τέτοιων περιπτώσεων.

Επίσης, το dataset περιλαμβάνει πλήθος χρηματοοικονομικών δεικτών και δυαδικών μεταβλητών που σχετίζονται με τη λειτουργία κάθε επιχείρησης. Οι σχέσεις μεταξύ

αυτών των μεταβλητών και της πιθανότητας πτώχευσης είναι πιθανό να είναι μη γραμμικές και να περιλαμβάνουν πολύπλοκες αλληλεπιδράσεις. Το MLP, έχει τη δυνατότητα να μοντελοποιεί τέτοιες σχέσεις.

Για τους παραπάνω λόγους, κρίθηκε σκόπιμο να συμπεριληφθεί το MLP στο σύνολο των ταξινομητών που χρησιμοποιήθηκαν στο πλαίσιο της εργασίας.

Γ) Υπολογισμός Μετρικών Απόδοσης

Για την αξιολόγηση κάθε μοντέλου, τόσο στο training όσο και στο test set, υπολογίστηκαν οι εξής μετρικές:

- **Accuracy:** Ποσοστό σωστών προβλέψεων
- **Precision:** Ακρίβεια στην πρόβλεψη των πτωχευμένων
- **Recall:** Ποσοστό εντοπισμένων πτωχευμένων επιχειρήσεων
- **F1 Score:** Αρμονικός μέσος Precision και Recall
- **ROC-AUC:** Ικανότητα διαχωρισμού των δύο τάξεων
- **Specificity (Μετρική 1):** Ικανότητα εντοπισμού υγιών εταιρειών
- **Balanced Accuracy (Μετρική 2):** Μέσος όρος Recall και Specificity

Επιλογή Μετρικών:

Για την αξιολόγηση των μοντέλων, εκτός από τις βασικές μετρικές (Accuracy, Precision, Recall, F1 score), χρησιμοποιήθηκαν και δύο πρόσθετες μετρικές:

- **Specificity (Μετρική 1):** Αντιπροσωπεύει την ικανότητα του μοντέλου να αναγνωρίζει σωστά τις υγιείς επιχειρήσεις. Αυτό είναι κρίσιμο, καθώς μια λανθασμένη εκτίμηση υγιούς επιχείρησης ως πτωχευμένη μπορεί να οδηγήσει σε άστοχη επιχειρηματική απόφαση (π.χ. απόρριψη συνεργασίας, άρνηση χρηματοδότησης).
- **Balanced Accuracy (Μετρική 2):** Υπολογίζεται ως ο μέσος όρος μεταξύ Recall (για τις πτωχευμένες εταιρείες) και Specificity (για τις υγιείς). Είναι ιδιαίτερα χρήσιμη σε προβλήματα ανισόρροπων τάξεων, όπως το παρόν, γιατί αξιολογεί την απόδοση εξίσου και για τις δύο κατηγορίες, παρέχοντας μια πιο συνολική εικόνα.

Όλα τα αποτελέσματα καταγράφηκαν ανά fold και μοντέλο, και αποθηκεύτηκαν σε μορφή Excel για περαιτέρω ανάλυση και δημιουργία γραφημάτων.

4. Πειραματικά Αποτελέσματα

Η αξιολόγηση των μοντέλων ταξινόμησης πραγματοποιήθηκε μέσω Stratified K-Fold με 4 folds. Για κάθε fold εκπαιδεύτηκαν τα ακόλουθα μοντέλα, τόσο σε balanced όσο και σε unbalanced training set:

- Linear Discriminant Analysis (LDA)
- Logistic Regression

- Decision Tree
- Random Forest
- k-Nearest Neighbors (k-NN)
- Naïve Bayes
- Support Vector Machines (SVM)
- Multi-Layer Perceptron (MLP – Νευρωνικό Δίκτυο)

Αξιολόγηση Απόδοσης

Για κάθε μοντέλο υπολογίστηκαν οι ακόλουθες μετρικές απόδοσης, τόσο στο σύνολο εκπαίδευσης όσο και στο σύνολο δοκιμής:

- Accuracy
- Precision
- Recall
- F1 Score
- ROC-AUC
- Specificity (Μετρική 1)
- Balanced Accuracy (Μετρική 2)

Τα αποτελέσματα συγκεντρώθηκαν στο αρχείο **balancedDataOutcomes.csv** και στη συνέχεια μετατράπηκαν σε Excel αρχείο, προκειμένου να παραχθούν πίνακες σύνοψης και διαγράμματα σύγκρισης.

Συγκριτική Ανάλυση Μοντέλων

Η σύγκριση των μοντέλων βασίστηκε κυρίως στην **F1 Score**, καθώς πρόκειται για μια μετρική που λαμβάνει υπόψη τόσο την Precision όσο και την Recall. Η επιλογή της στο πλαίσιο της παρούσας εργασίας έγινε διότι το dataset είναι unbalanced καθώς περιλαμβάνει πολύ περισσότερες υγιείς επιχειρήσεις σε σχέση με τις πτωχευμένες. Σε τέτοιες περιπτώσεις, το **Accuracy** μπορεί να είναι παραπλανητικό, καθώς ένα μοντέλο που προβλέπει σχεδόν όλες τις επιχειρήσεις ως υγιείς μπορεί να εμφανίζεται ως ακριβές, ενώ στην πραγματικότητα αποτυγχάνει να εντοπίσει τις εταιρείες που τελικά θα πτωχεύσουν.

Η χρησιμότητα του F1 Score έγκειται στο ότι εστιάζει στην ικανότητα του μοντέλου να αναγνωρίζει σωστά τις πτωχευμένες εταιρείες, χωρίς όμως να θυσιάζει την ακρίβεια της πρόβλεψης. Αυτό έχει πρακτική σημασία σε επιχειρησιακά περιβάλλοντα, όπου το να προβλεφθεί έγκαιρα μια πτώχευση μπορεί να αποτρέψει επενδυτικούς κινδύνους ή εμπορικές συνεργασίες με υψηλό ρίσκο.

Από την ανάλυση προέκυψαν τα εξής:

- Το **Random Forest** έχει τον υψηλότερο μέσο όρο F1 Score (0.21) στο test set ωστόσο το **SVM** είναι πολύ κοντά με μέσο όρο F1 Score (0.20).
- Τα **MLP, LDA, Logistic Regression** ακολουθούν με μέσο όρο F1 Score (0.18).

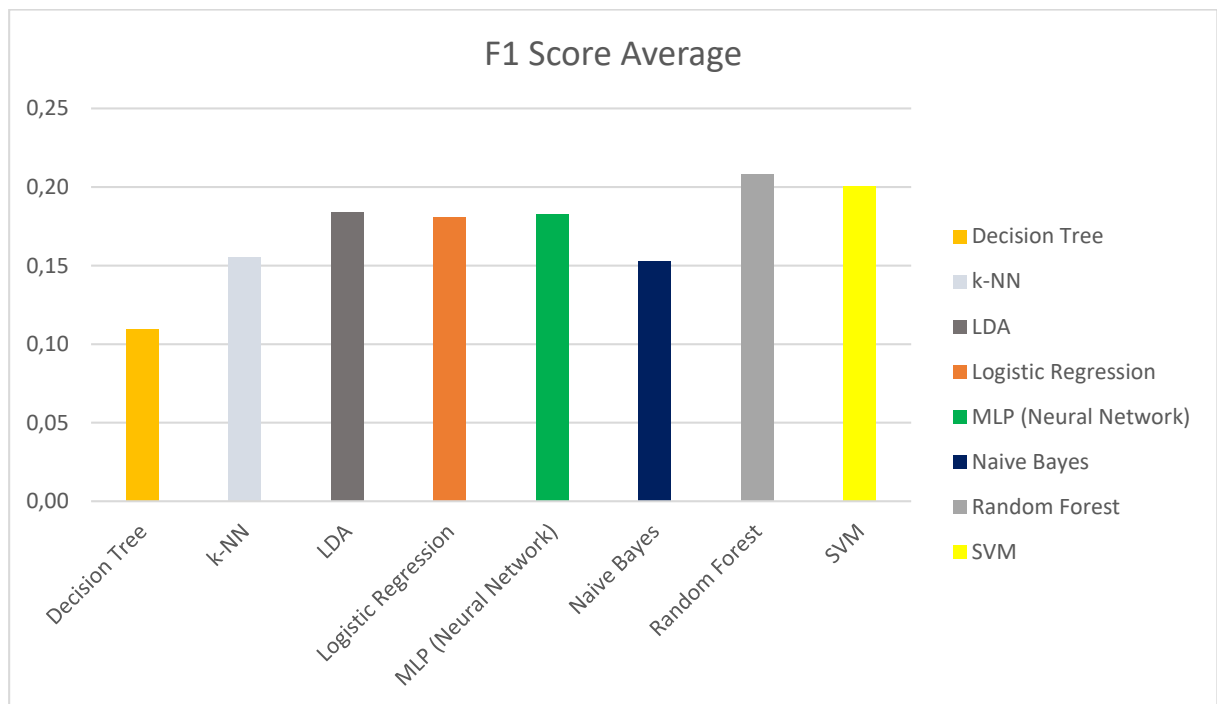


Figure 3: Μέσος όρος της F1 Score για κάθε μοντέλο ταξινόμησης στο test set.

Η Εικόνα 3 παρουσιάζει τη σύγκριση των μοντέλων ταξινόμησης με βάση τον μέσο όρο της F1 Score στο test set, όπως αυτός υπολογίστηκε μέσω Stratified K-Fold. Το διάγραμμα αυτό προσφέρει μια σαφή οπτική αναπαράσταση της ικανότητας κάθε μοντέλου να διαχειριστεί την ασυμμετρία του dataset και να προβλέψει τις κρίσιμες περιπτώσεις.

Εκτός από το F1 Score, στην παρούσα εργασία αξιολογήθηκαν τα μοντέλα και με βάση τις μετρικές **ROC-AUC** και **Balanced Accuracy**, καθώς προσφέρουν επιπλέον οπτικές για την ποιότητα των προβλέψεων. Η ROC-AUC αποτυπώνει την ικανότητα του μοντέλου να διαχωρίζει σωστά τις δύο κατηγορίες ανεξαρτήτως threshold. Αντίστοιχα, η Balanced Accuracy, ως μέσος όρος Recall και Specificity, επιτρέπει τη δίκαιη αξιολόγηση της απόδοσης και στις δύο τάξεις, αποφεύγοντας την αυξημένη ακρίβεια λόγω της κυρίαρχης τάξης (υγιείς επιχειρήσεις).

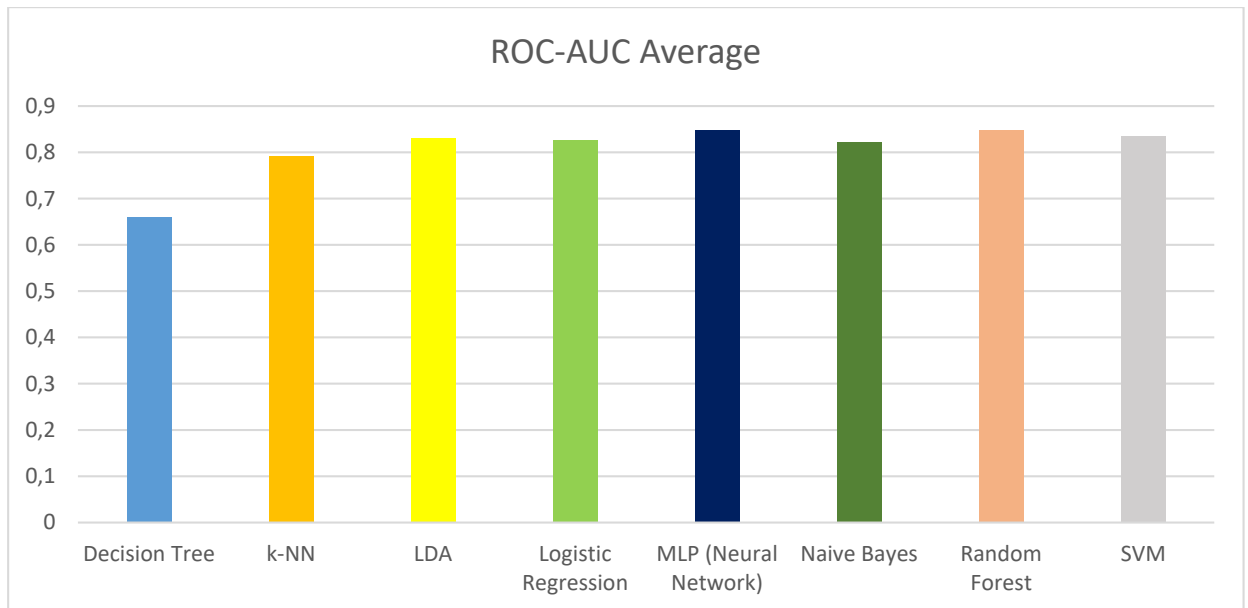


Figure 4: Μέση απόδοση των μοντέλων βάσει της μετρικής ROC-AUC στο test set.

Το παραπάνω διάγραμμα παρουσιάζει τη μέση τιμή της μετρικής ROC-AUC για όλα τα εξεταζόμενα μοντέλα, όπως υπολογίστηκε από τα test sets. Παρατηρούμε ότι όλα τα μοντέλα σημειώνουν υψηλές επιδόσεις, με τιμές ROC-AUC κυμαινόμενες μεταξύ ~ 0.68 και ~ 0.84 , υποδεικνύοντας καλή ικανότητα διαχωρισμού μεταξύ υγιών και πτωχευμένων επιχειρήσεων. Το MLP (Neural Network) και το Random Forest φαίνεται να επιτυγχάνουν τις υψηλότερες επιδόσεις, ξεπερνώντας το όριο του 0.82, κάτι που ενισχύει την αξιοπιστία τους για χρήση σε προβλήματα με μη εξισορροπημένες κατηγορίες.

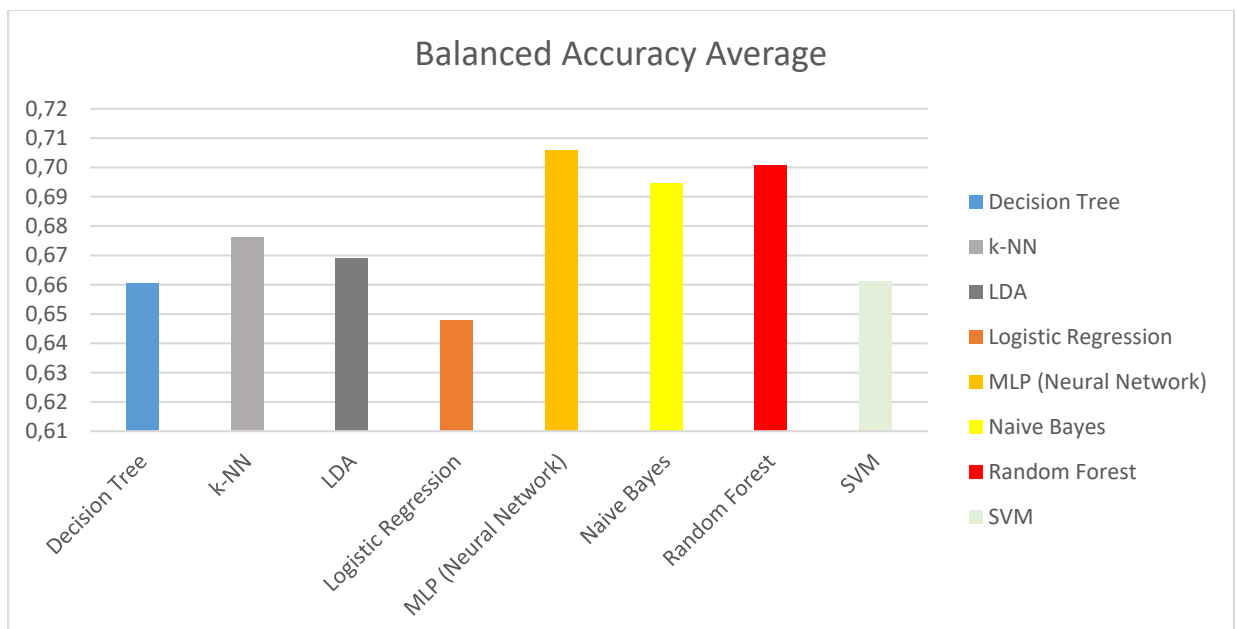


Figure 5: Μέση απόδοση των μοντέλων ταξινόμησης με βάση τη Balanced Accuracy στο test set.

Το γράφημα αποτυπώνει τη μέση **Balanced Accuracy** των μοντέλων στο **test set**, προσφέροντας μια εξισορροπημένη εικόνα της απόδοσής τους τόσο στην ανίχνευση των πτωχευμένων όσο και των υγιών επιχειρήσεων. Παρατηρούμε πως το **MLP (Neural Network)**, το **Random Forest** και το **Naive Bayes** ξεχωρίζουν με τιμές άνω του 0.69, ενώ τα υπόλοιπα μοντέλα κυμαίνονται σε χαμηλότερα επίπεδα. Η **Balanced Accuracy** είναι ιδιαίτερα κρίσιμη για την αξιολόγηση του παρόντος προβλήματος, καθώς αποτρέπει την παραπλανητική υπερεκτίμηση της ακρίβειας λόγω της κυρίαρχης τάξης (υγιείς εταιρείες), δίνοντας ίσο βάρος στη σωστή πρόβλεψη και των δύο κατηγοριών. Έτσι, μοντέλα που εμφανίζονται ισορροπημένα μεταξύ **Recall** και **Specificity** αποκτούν προβάδισμα ως πιο αξιόπιστες επιλογές.

5. Συμπεράσματα

Η παρούσα εργασία ανέλυσε το πρόβλημα της πρόβλεψης πτωχεύσεων επιχειρήσεων χρησιμοποιώντας πληθώρα αλγορίθμων ταξινόμησης και πειραματιζόμενη με διάφορες μετρικές απόδοσης σε συνθήκες ανισορροπίας των κλάσεων. Η αξιολόγηση των μοντέλων βασίστηκε κυρίως στο **F1 Score**, ενώ συμπληρώθηκε από τις μετρικές **ROC-AUC** και **Balanced Accuracy**, για πληρέστερη εικόνα των αποτελεσμάτων σε σχέση με την ταξινόμηση και των δύο κατηγοριών (υγιείς και πτωχευμένες επιχειρήσεις).

Με βάση τα αποτελέσματα των δοκιμών σε 4 διαφορετικά **folds**:

- Το καλύτερο συνολικά μοντέλο, βάσει του μέσου **F1 Score** στο **test set**, ήταν το **Random Forest**, το οποίο πέτυχε την υψηλότερη τιμή μεταξύ όλων των ταξινομητών. Παράλληλα, το **Random Forest** εμφάνισε και εξαιρετική επίδοση σε **ROC-AUC** και **Balanced Accuracy**, γεγονός που ενισχύει την αξιοπιστία του.
- Εξετάστηκε αν υπάρχει μοντέλο που πληροί ταυτόχρονα τους δύο περιορισμούς απόδοσης:
 - **Recall $\geq 60\%$** για τις πτωχευμένες επιχειρήσεις (ικανότητα εντοπισμού των θετικών περιπτώσεων).
 - **Specificity $\geq 70\%$** για τις υγιείς επιχειρήσεις (ικανότητα αποφυγής ψευδώς θετικών).

Για να εντοπιστεί αν υπάρχει μοντέλο που να ικανοποιεί τις παραπάνω απαιτήσεις δημιουργήθηκε ο παρακάτω συγκεντρωτικός πίνακας:

Ετικέτες γραμμής	Μέσος όρος από Recall	Μέσος όρος από Specificity
Decision Tree	0,75	0,91
test	0,50	0,82
train	1	1
k-NN	0,54	0,91
test	0,46	0,90
train	0,62	0,93
LDA	0,41	0,93
test	0,41	0,93
train	0,40	0,93
Logistic Regression	0,36	0,94
test	0,36	0,94
train	0,36	0,94
MLP (Neural Network)	0,58	0,92
test	0,51	0,90
train	0,64	0,94
Naive Bayes	0,51	0,88
test	0,51	0,88
train	0,51	0,88
Random Forest	0,74	0,96
test	0,48	0,93
train	1	1
SVM	0,40	0,95
test	0,38	0,94
train	0,43	0,96
Γενικό Άθροισμα	0,54	0,93

Figure 6: Μέσοι όροι Recall και Specificity στο test set για κάθε μοντέλο (balanced train set)

Ο πίνακας απεικονίζει τους μέσους όρους των μετρικών Recall και Specificity για κάθε μοντέλο στο test set, όταν η εκπαίδευση έγινε σε ισορροπημένο (balanced) σύνολο. Η μετρική Recall αξιολογεί την ικανότητα του μοντέλου να εντοπίζει σωστά τις πτωχευμένες επιχειρήσεις, ενώ η Specificity αξιολογεί την ακρίβεια εντοπισμού των υγιών επιχειρήσεων.

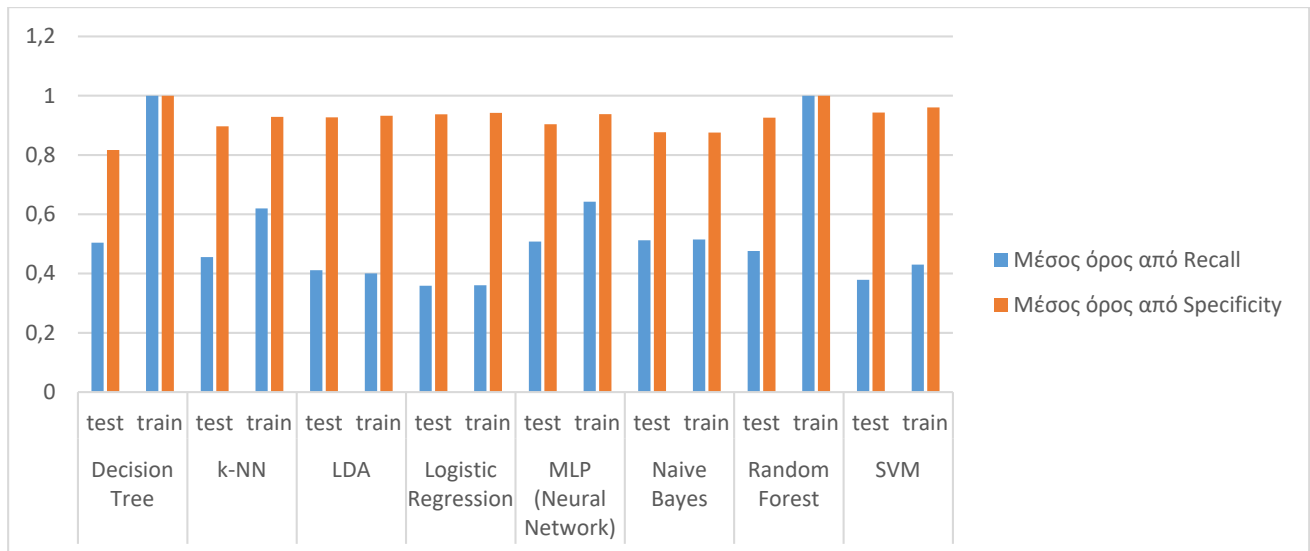


Figure 7: Διάγραμμα με τον μέσο όρο Recall και Specificity για κάθε μοντέλο στο train και test set (balanced).

Το γράφημα απεικονίζει συγκριτικά τον μέσο όρο των μετρικών Recall και Specificity για κάθε μοντέλο, τόσο στο training όσο και στο test set, όταν η εκπαίδευση έγινε σε ισορροπημένο σύνολο δεδομένων.

Με βάση τα μέσα ποσοστά ανά μοντέλο, προκύπτει ότι:

- Κανένα μοντέλο στο test set δεν πληροί και τις δύο απαιτήσεις ταυτόχρονα.
- Το **Random Forest** είχε υψηλή **Specificity (0.93)**, αλλά το **Recall (0.48)** ήταν κάτω από το όριο.
- Το **Decision Tree** είχε το υψηλότερο Recall (0.50), αλλά και αυτό **δεν έφτασε το 60%**.

Για να αξιολογηθεί αν κάποιο μοντέλο πληροί τους δύο ποσοτικούς περιορισμούς της εκφώνησης — δηλαδή την ικανότητα εντοπισμού τουλάχιστον 60% των πτωχευμένων επιχειρήσεων και τουλάχιστον 70% των υγιών — κρίθηκε απαραίτητη η χρήση των κατάλληλων ταξινομικών μετρικών. Συγκεκριμένα, ο πρώτος περιορισμός αντιστοιχεί στη μετρική Recall για την πτωχευμένη τάξη (τάξη 2), η οποία μετρά το ποσοστό των πραγματικά πτωχευμένων εταιρειών που το μοντέλο εντόπισε σωστά. Ο δεύτερος περιορισμός σχετίζεται με τη μετρική Specificity, η οποία μετρά την ικανότητα του μοντέλου να αναγνωρίζει σωστά τις υγιείς εταιρείες (τάξη 1), αποφεύγοντας ψευδώς θετικές προβλέψεις.

Η παρούσα μελέτη ανέδειξε τη χρησιμότητα της επιβλεπόμενης μάθησης για την πρόβλεψη πτωχεύσεων επιχειρήσεων με τη χρήση χρηματοοικονομικών δεδομένων. Μέσα από τη σύγκριση πλήθους ταξινομητικών μοντέλων και την εφαρμογή τεχνικών εξισορρόπησης δεδομένων, διαπιστώθηκε ότι μοντέλα όπως το Random Forest μπορούν να επιτύχουν αξιόπιστη απόδοση, ακόμα και σε περιβάλλοντα με έντονη ανισορροπία κατηγοριών. Επιπλέον, η ανάλυση βάσει των απαιτήσεων ($\text{Recall} \geq 60\%$,

Specificity $\geq 70\%$) επιτρέπει μια πιο ρεαλιστική αποτίμηση της πρακτικής αξίας κάθε αλγορίθμου.

6. Παράρτημα

Στο παρόν παράρτημα παρατίθενται ενδεικτικά τα Confusion Matrices για τα μοντέλα που είχαν την καλύτερη επίδοση F1 Score δηλαδή τα Random Forest και SVM. Παρουσιάζεται η απόδοσή τους στο test set για καθένα από τα τέσσερα folds.

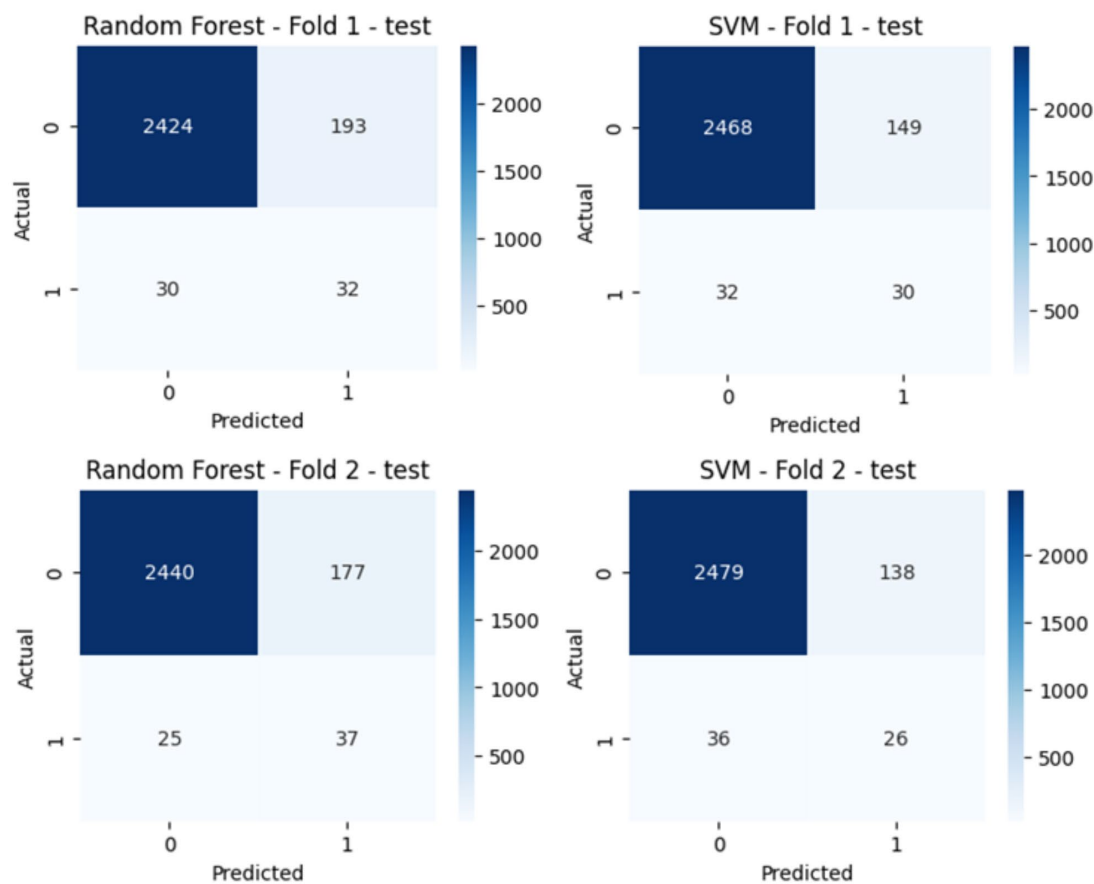


Figure 8 : Confusion Matrices των Random Forest και SVM για τα folds 1 και 2.

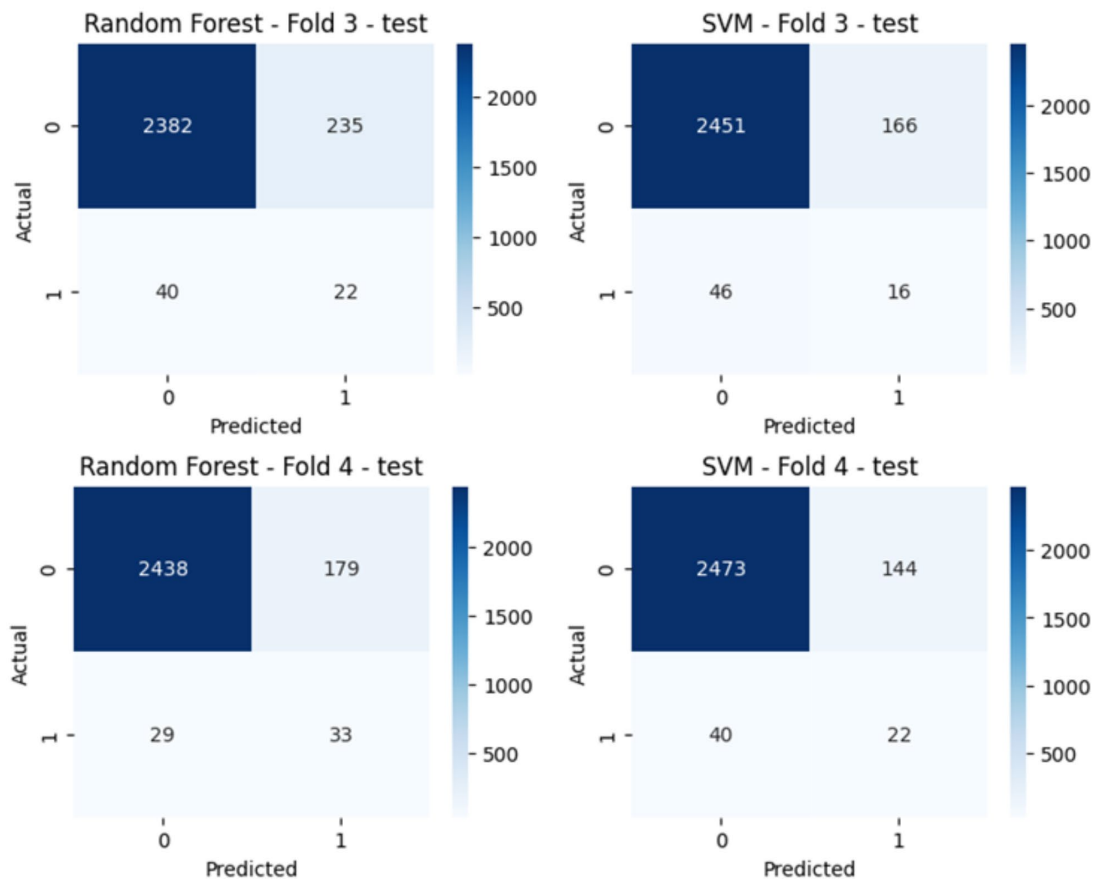


Figure 9 : Confusion Matrices των Random Forest και SVM για τα folds 3 και 4.

Οι υπόλοιποι confusion matrices και οι αναλυτικές μετρικές ανά fold βρίσκονται στο συνοδευτικό `Code.py`.

7. Βιβλιογραφία

- [1] Scikit-learn: Machine Learning in Python. Επίσημη τεκμηρίωση:
<https://scikit-learn.org/stable/>
- [2] pandas – Data Analysis Library. Εργαλείο για ανάλυση και επεξεργασία δεδομένων:
<https://pandas.pydata.org/>
- [3] NumPy – Scientific Computing with Python. Βασική βιβλιοθήκη για αριθμητικούς υπολογισμούς:
<https://numpy.org/>
- [4] Python Documentation – Επίσημη τεκμηρίωση της γλώσσας Python:
<https://docs.python.org/3/>
- [5] Matplotlib – Visualization with Python. Εργαλείο για δημιουργία γραφημάτων:
<https://matplotlib.org/>

[6] Seaborn – Statistical Data Visualization. Επέκταση της matplotlib για στατιστικά διαγράμματα:

<https://seaborn.pydata.org/>

[7] Confusion Matrix – Wikipedia: Formulas for evaluation metrics (Accuracy, Precision, Recall, F1 Score, Specificity):

https://en.wikipedia.org/wiki/Confusion_matrix

[8] MLPClassifier – Scikit-learn Neural Network Models (Supervised):

https://scikit-learn.org/stable/modules/neural_networks_supervised.html