Πανεπιστήμιο Μακεδονίας
Πρόγραμμα Μεταπτυχιακών Σπουδών

Μάθημα:
Μέθοδοι και Εργαλεία Τεχνητής Νοημοσύνης

Επιβλεπόμενη μάθηση – Ταξινόμηση

ΣΥΓΓΡΑΦΕΙΣ:
Κωνσταντίνος Θωμασιάδης, mai25016
Ευστάθιος Ιωσηφίδης, mai25017
Αικατερίνη Κρότκα, mai25031

Θεσσαλονίκη, Μάιος 2025

Κατάλογος σχημάτων

1. Εισαγωγή

Η πρόβλεψη της οικονομικής κατάστασης μιας επιχείρησης αποτελεί κομβικό ζήτημα για πλήθος ενδιαφερόμενων φορέων, όπως τράπεζες, επενδυτές, πιστωτές και ρυθμιστικές αρχές. Ιδιαίτερα σημαντική είναι η δυνατότητα έγκαιρου εντοπισμού επιχειρήσεων που κινδυνεύουν να κηρύξουν χρεοκοπία, προκειμένου να ληφθούν προληπτικά μέτρα ή να περιοριστεί η έκθεση σε σχετικό κίνδυνο.

Στο πλαίσιο αυτής της μελέτης, επιχειρείται η επίλυση του προβλήματος ταξινόμησης επιχειρήσεων σε δύο κατηγορίες: (α) υγιείς και (β) χρεωκοπημένες, με χρήση δεδομένων που περιλαμβάνουν οικονομικούς δείκτες απόδοσης και δραστηριότητας. Η ανάλυση βασίζεται σε δεδομένα που καλύπτουν πλήθος εταιρειών και ετών, παρέχοντας μια κατάλληλη βάση για την εφαρμογή και αξιολόγηση αλγορίθμων μηχανικής μάθησης.

Σκοπός της παρούσας αναφοράς είναι η συγκριτική αξιολόγηση διαφορετικών τεχνικών ταξινόμησης ως προς την ικανότητά τους να προβλέπουν με ακρίβεια τη χρεοκοπία επιχειρήσεων. Για τον σκοπό αυτό, εφαρμόζονται και αναλύονται μοντέλα όπως Logistic Regression, Random Forest, Support Vector Machines και άλλα, με στόχο την εξαγωγή χρήσιμων συμπερασμάτων τόσο για την απόδοσή τους όσο και για την καταλληλότητά τους σε πραγματικά επιχειρησιακά περιβάλλοντα.

2. Περιγραφή Δεδομένων και Προεπεξεργασία

Η παρούσα μελέτη βασίζεται σε δεδομένα επιχειρήσεων, με στόχο την πρόβλεψη της πιθανότητας χρεοκοπίας. Κάθε γραμμή του πίνακα δεδομένων αντιστοιχεί σε μια συγκεκριμένη επιχείρηση κατά ένα οικονομικό έτος. Τα δεδομένα κρίνονται επαρκή για την εκπαίδευση και αξιολόγηση μοντέλων μηχανικής μάθησης.

2.1 Δομή Δεδομένων

Το σύνολο δεδομένων περιλαμβάνει τις εξής μεταβλητές:

- Οικονομικοί δείκτες (στήλες A έως H): 8 αριθμητικές μεταβλητές που αποτυπώνουν την οικονομική απόδοση των επιχειρήσεων
- Δείκτες δραστηριότητας (στήλες I, J, K): 3 δυαδικές μεταβλητές που αναπαριστούν χαρακτηριστικά ή ενέργειες της επιχείρησης
- Έτος: Περιέχει το οικονομικό έτος στο οποίο αντιστοιχούν οι υπόλοιπες μεταβλητές
- Κατάσταση επιχείρησης:
    1. Υγιής επιχείρηση
    2. Επιχείρηση που έχει κηρύξει χρεοκοπία

Για λόγους διευκόλυνσης της ανάλυσης, η κατάσταση της επιχείρησης μετατράπηκε σε δυαδική:

0: Υγιής (αρχικό 1)
1: Χρεωκοπημένη (αρχικό 2)

2.2 Έλεγχος Ποιότητας – Ελλείπουσες Τιμές

Πραγματοποιήθηκε έλεγχος πληρότητας του dataset, από τον οποίο **δεν προέκυψαν ελλιπείς τιμές**. Συνεπώς, δεν κρίθηκε απαραίτητη η εφαρμογή τεχνικών διαχείρισης ελλείψεων. (Protopapadakis, nd)

2.3 Κανονικοποίηση Δεδομένων

Οι αριθμητικές μεταβλητές (στήλες A έως H) παρουσίαζαν σημαντικές διαφοροποιήσεις στις κλίμακές τους. Για τον λόγο αυτό, εφαρμόστηκε κανονικοποίηση min-max, μεταφέροντας τις τιμές σε κοινό εύρος [0, 1]. Η ενέργεια αυτή είναι κρίσιμη για την ισότιμη συμμετοχή των μεταβλητών στην εκπαίδευση των μοντέλων. (Protopapadakis, nd)

Οι δυαδικοί δείκτες (I, J, K) διατηρήθηκαν ως είχαν, καθώς βρίσκονταν ήδη στην κατάλληλη μορφή.

2.4 Αντιμετώπιση Ανισορροπίας Κλάσεων

Η κατανομή των εγγραφών στο dataset είναι έντονα μη ισοβαρής, με τις υγιείς επιχειρήσεις να υπερτερούν σημαντικά αριθμητικά έναντι των χρεωκοπημένων. Η ανισορροπία αυτή μπορεί να επηρεάσει αρνητικά την εκπαίδευση, οδηγώντας τα μοντέλα σε μεροληψία υπέρ της κυρίαρχης κλάσης. (Protopapadakis, nd)

Για την αντιμετώπιση του προβλήματος εφαρμόστηκε στρατηγική υποδειγματοληψίας (undersampling) στο εκπαιδευτικό σύνολο, ώστε να επιτευχθεί αναλογία 3:1 μεταξύ υγιών και προβληματικών εταιρειών. Η εξισορρόπηση έγινε μόνο στο training set, διατηρώντας το test set ανεπηρέαστο για να αποτυπώνει ρεαλιστικά την απόδοση των μοντέλων σε συνθήκες πραγματικού κόσμου.

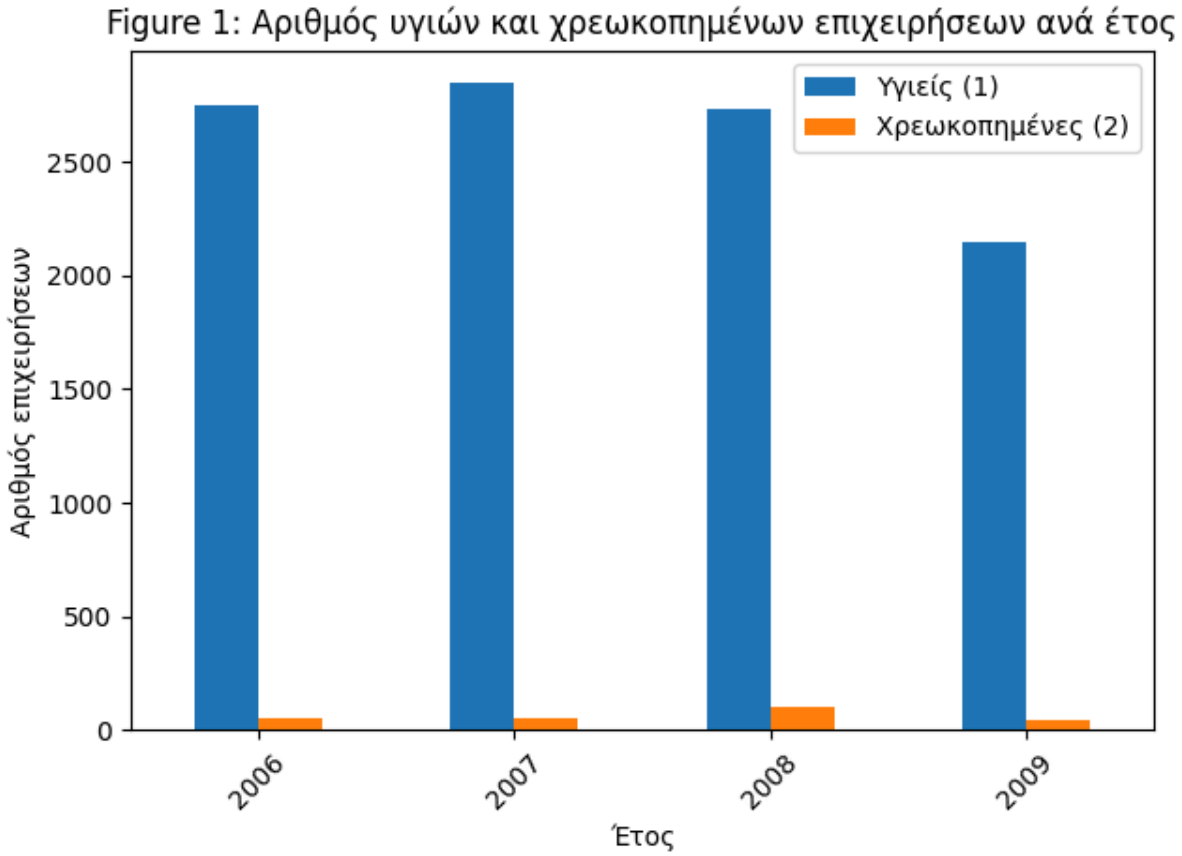
2.5 Διαχωρισμός Δεδομένων

Η αξιολόγηση των μοντέλων έγινε με 4-fold cross-validation, με stratified διαχωρισμό ώστε να διατηρηθούν οι αναλογίες των κλάσεων σε κάθε fold. Σε κάθε επανάληψη, το 75% των δεδομένων χρησιμοποιήθηκε για εκπαίδευση και το 25% για δοκιμή.

3. Μεθοδολογία

3.1 Κατανομή Υγιών και Χρεωκοπημένων Εταιρειών ανά Έτος

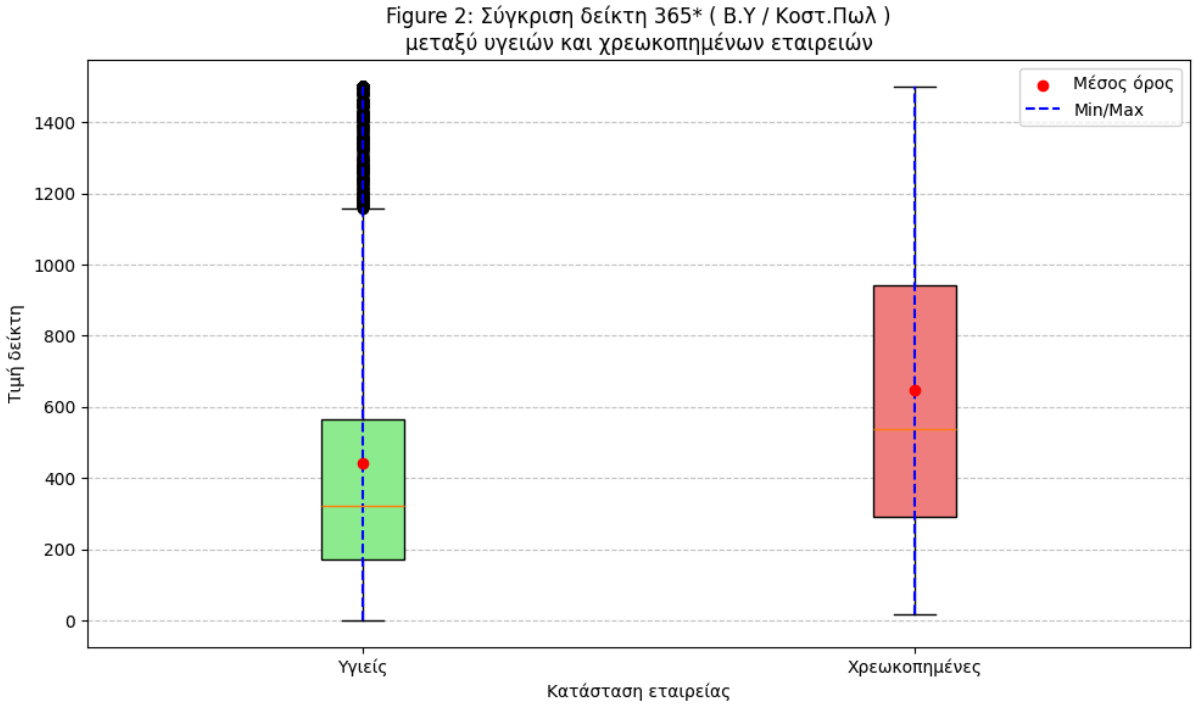
Στο Σχήμα 3.1 παρουσιάζεται η κατανομή των υγιών και χρεωκοπημένων εταιρειών ανά έτος. Παρατηρείται έντονη ασυμμετρία μεταξύ των δύο κατηγοριών, με τις υγιείς επιχειρήσεις να αποτελούν τη συντριπτική πλειοψηφία, γεγονός που αναδεικνύει τη φύση του προβλήματος ως πρόβλημα με μη ισορροπημένες κλάσεις. (Protopapadakis, nd)



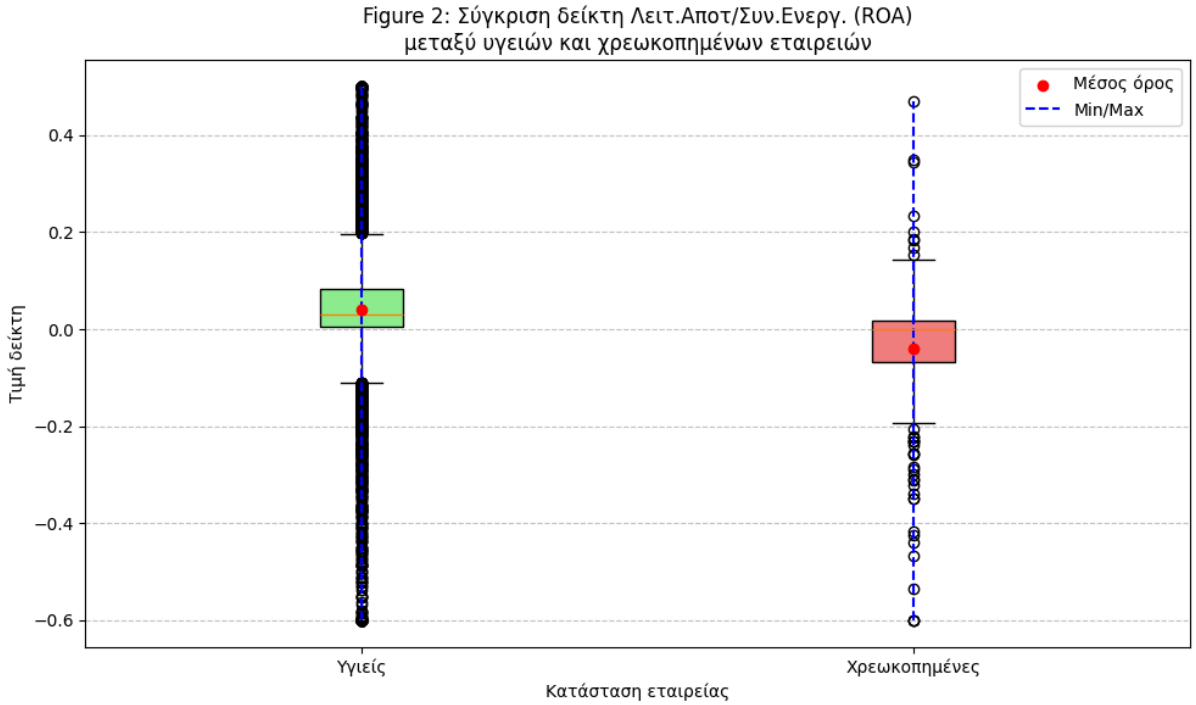
***Σχήμα 3.1**: Κατανομή Υγιών και Χρεωκοπημένων Εταιρειών ανά Έτος*

3.2 Σύγκριση Δεικτών Απόδοσης μεταξύ Υγιών και Χρεωκοπημένων Εταιρειών - Boxplots

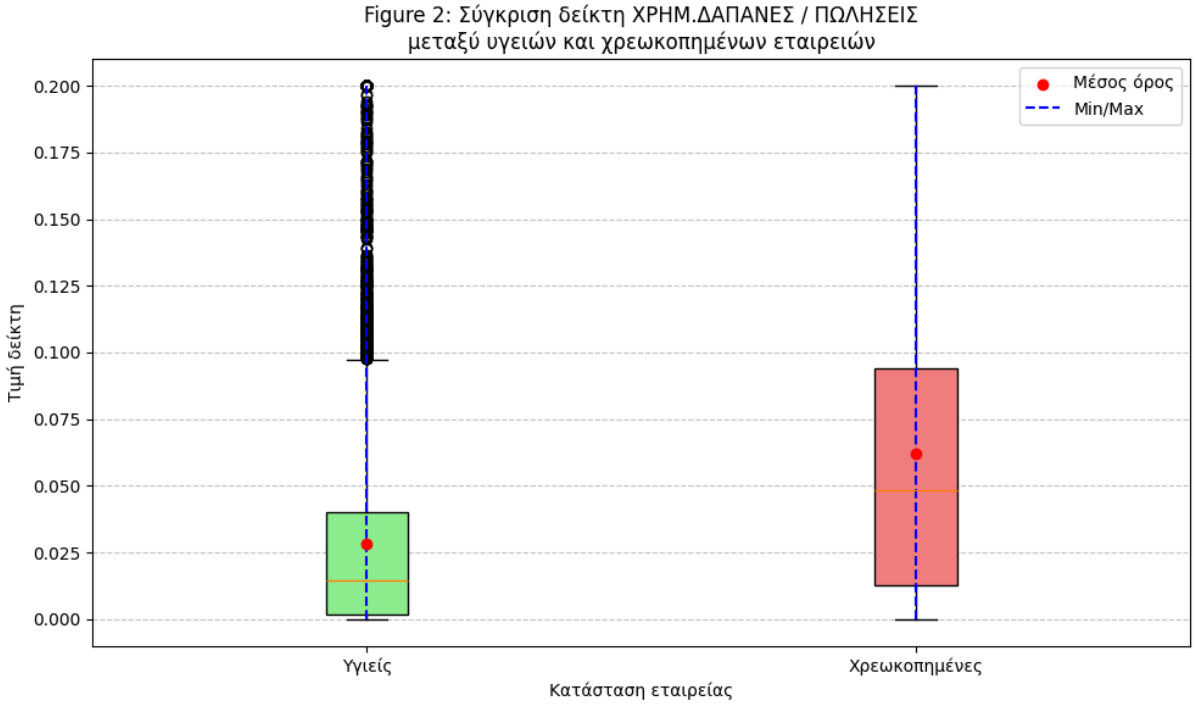
Παρουσιάζονται boxplots για κάθε χρηματοοικονομικό δείκτη, με σκοπό τη διερεύνηση πιθανών διαφορών στην κατανομή των τιμών μεταξύ υγιών και χρεωκοπημένων εταιρειών. Η χρήση διαγραμμάτων αυτού του τύπου επιτρέπει την απεικόνιση της κεντρικής τάσης, της διασποράς, καθώς και της παρουσίας ακραίων τιμών (outliers) για κάθε ομάδα. (Protopapadakis, nd)



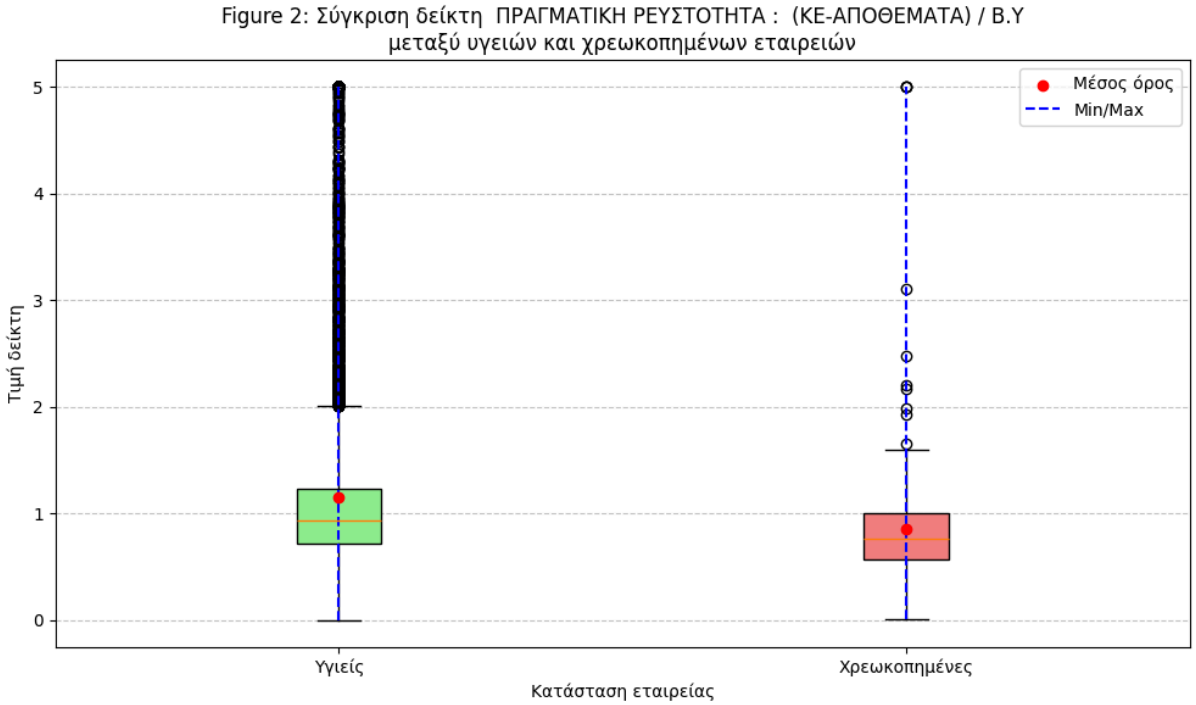
***Σχήμα 3.2.1**: Σύγκριση Δεικτών Απόδοσης μεταξύ Υγιών και Χρεωκοπημένων Εταιρειών - Boxplots: Σύγκριση δεικτών 365* (Β.Υ /Κοστ.Πωλ) μεταξύ υγειών και χρεωκοπημένων εταιρειών*



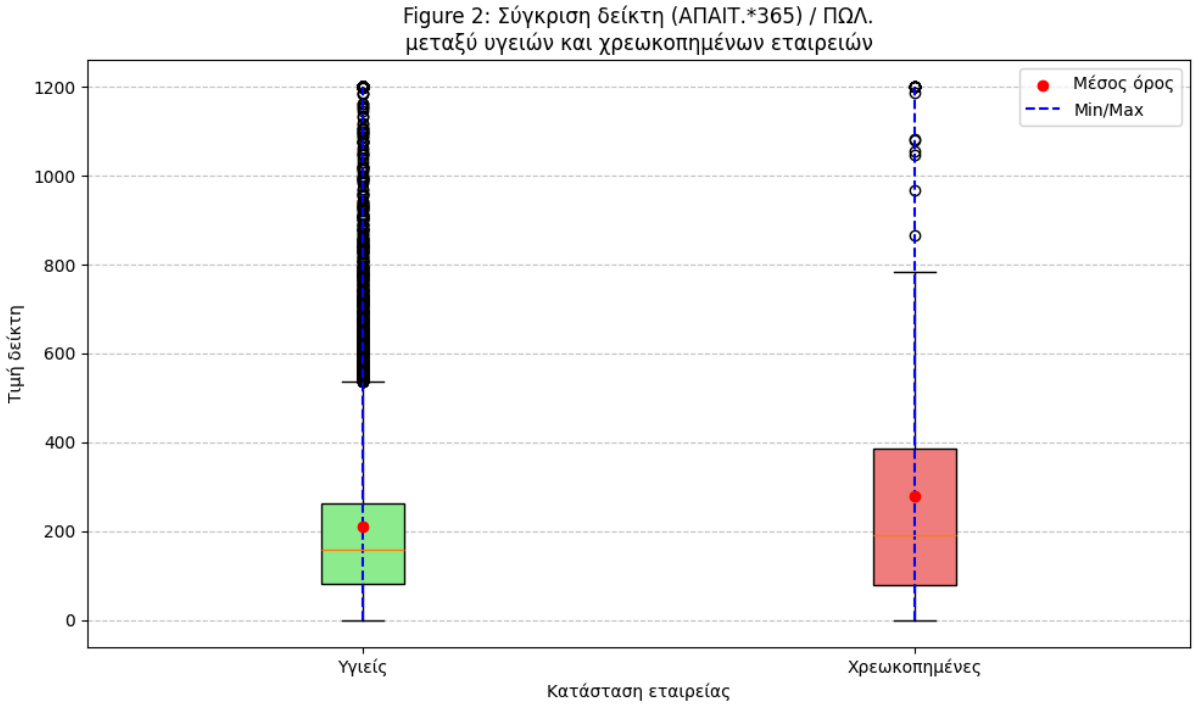
***Σχήμα 3.2.2**: Σύγκριση Δεικτών Απόδοσης μεταξύ Υγιών και Χρεωκοπημένων Εταιρειών - Boxplots: Σύγκριση δείκτη Λειτ.Αποστ/Συν.Ενεργ.(ROA) μεταξύ υγειών και χρεωκοπημένων εταιρειών*



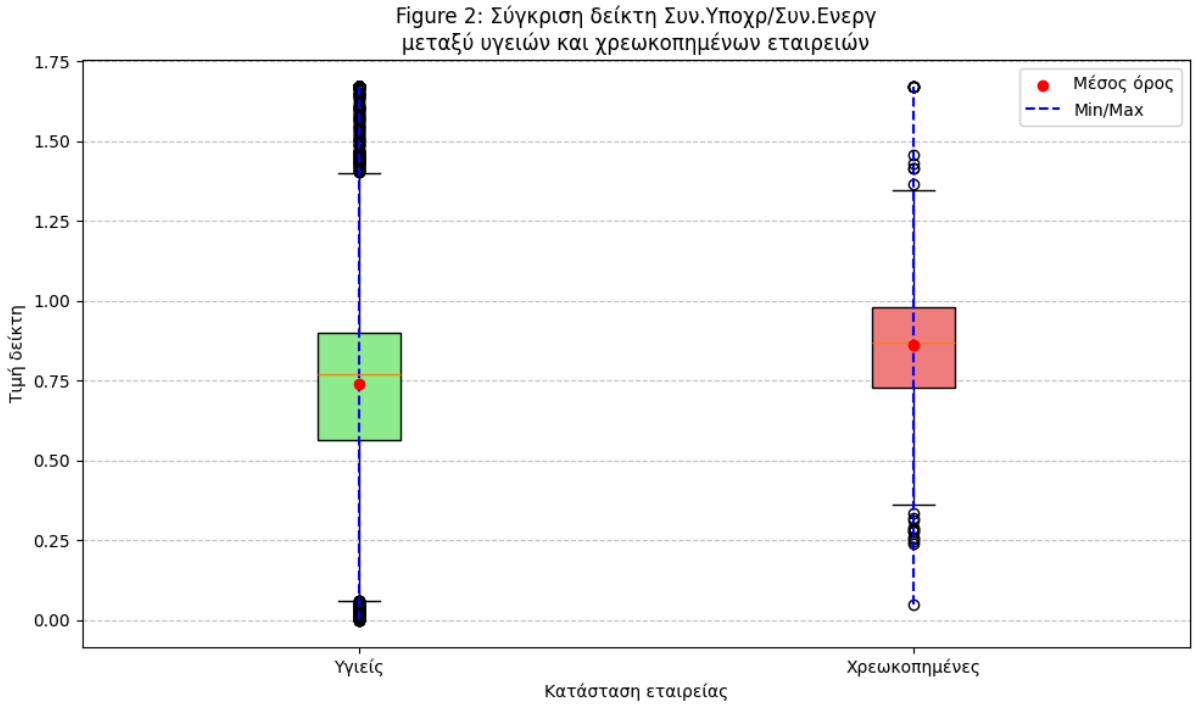
***Σχήμα 3.2.3**: Σύγκριση Δεικτών Απόδοσης μεταξύ Υγιών και Χρεωκοπημένων Εταιρειών - Boxplots: Σύγκριση δείκτη ΧΡΗΜ.ΔΑΠΑΝΕΣ / ΠΩΛΗΣΕΙΣ μεταξύ υγειών και χρεωκοπημένων εταιρειών*



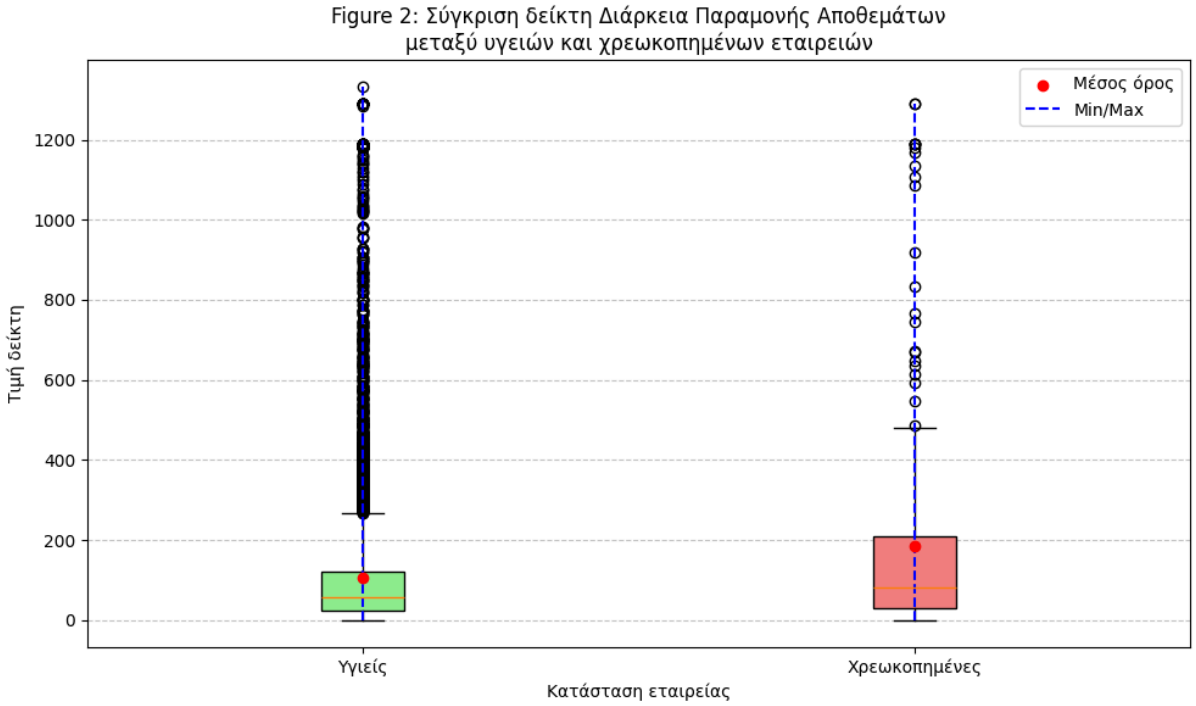
***Σχήμα 3.2.4**: Σύγκριση Δεικτών Απόδοσης μεταξύ Υγιών και Χρεωκοπημένων Εταιρειών - Boxplots: Σύγκριση δεικτών 365* (Β.Υ /Κοστ.Πωλ) μεταξύ υγειών και χρεωκοπημένων εταιρειών*



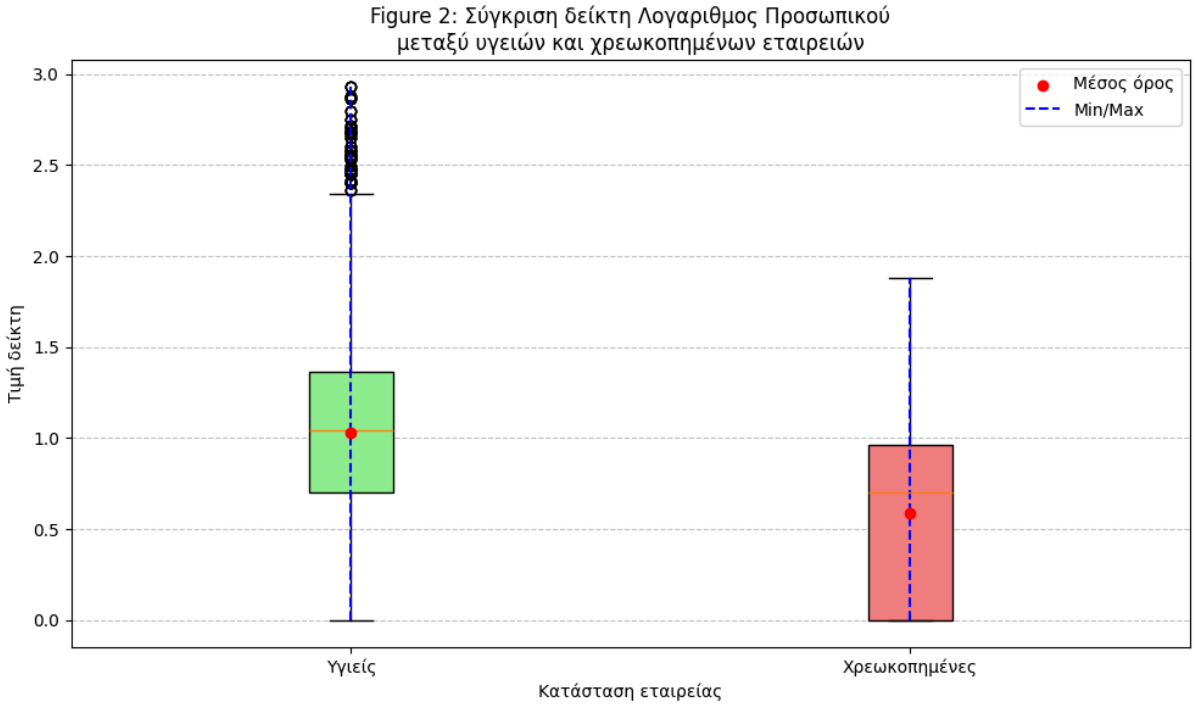
***Σχήμα 3.2.5**: Σύγκριση Δεικτών Απόδοσης μεταξύ Υγιών και Χρεωκοπημένων Εταιρειών - Boxplots: Σύγκριση δείκτη (ΑΠΑΙΤ.*365) / Πωλ μεταξύ υγειών και χρεωκοπημένων εταιρειών*



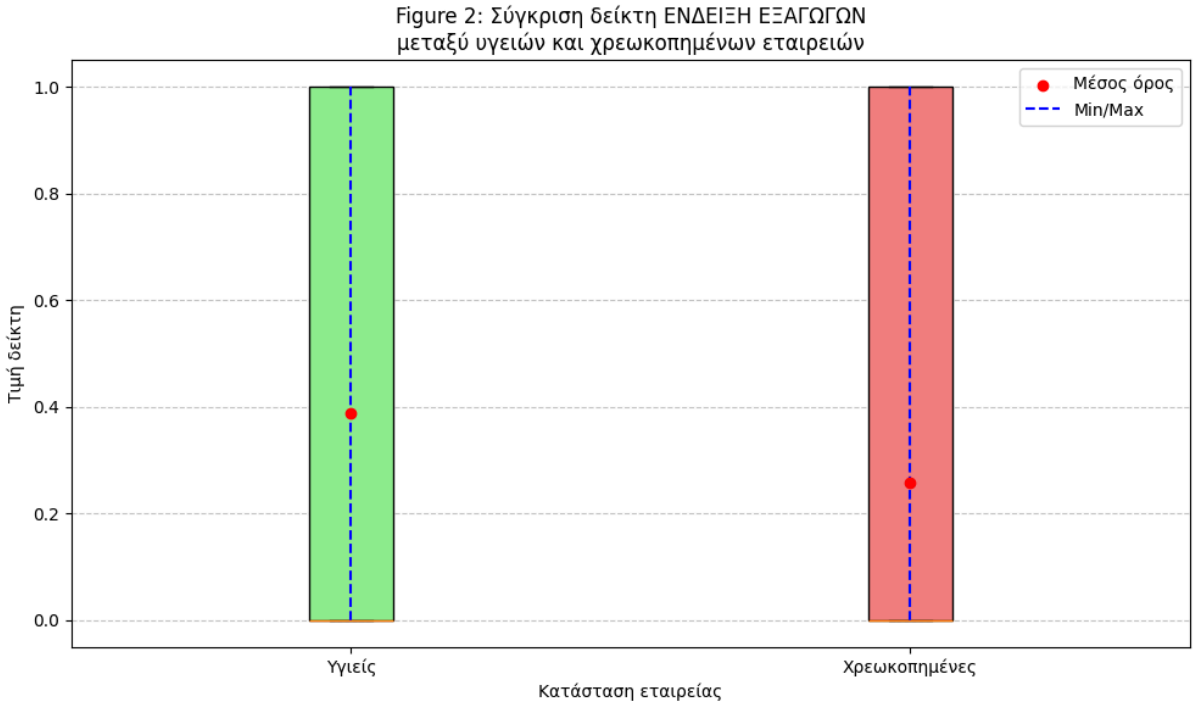
***Σχήμα 3.2.6**: Σύγκριση Δεικτών Απόδοσης μεταξύ Υγιών και Χρεωκοπημένων Εταιρειών - Boxplots: Σύγκριση δείκτη Συν.Υποχρ/Συν.Ενεργ μεταξύ υγειών και χρεωκοπημένων εταιρειών*



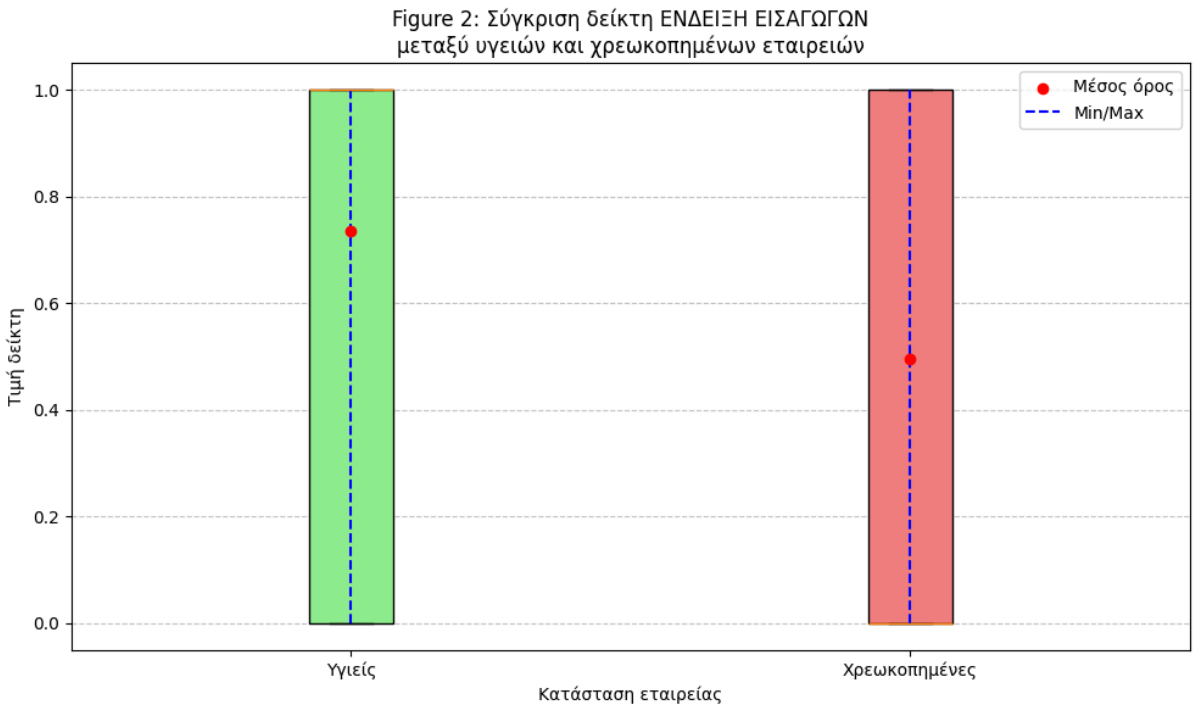
***Σχήμα 3.2.7**: 3.2 Σύγκριση Δεικτών Απόδοσης μεταξύ Υγιών και Χρεωκοπημένων Εταιρειών - Boxplots: Σύγκριση δείκτη Διάρκεια Παραμονής Αποθεμάτων μεταξύ υγειών και χρεωκοπημένων εταιρειών*



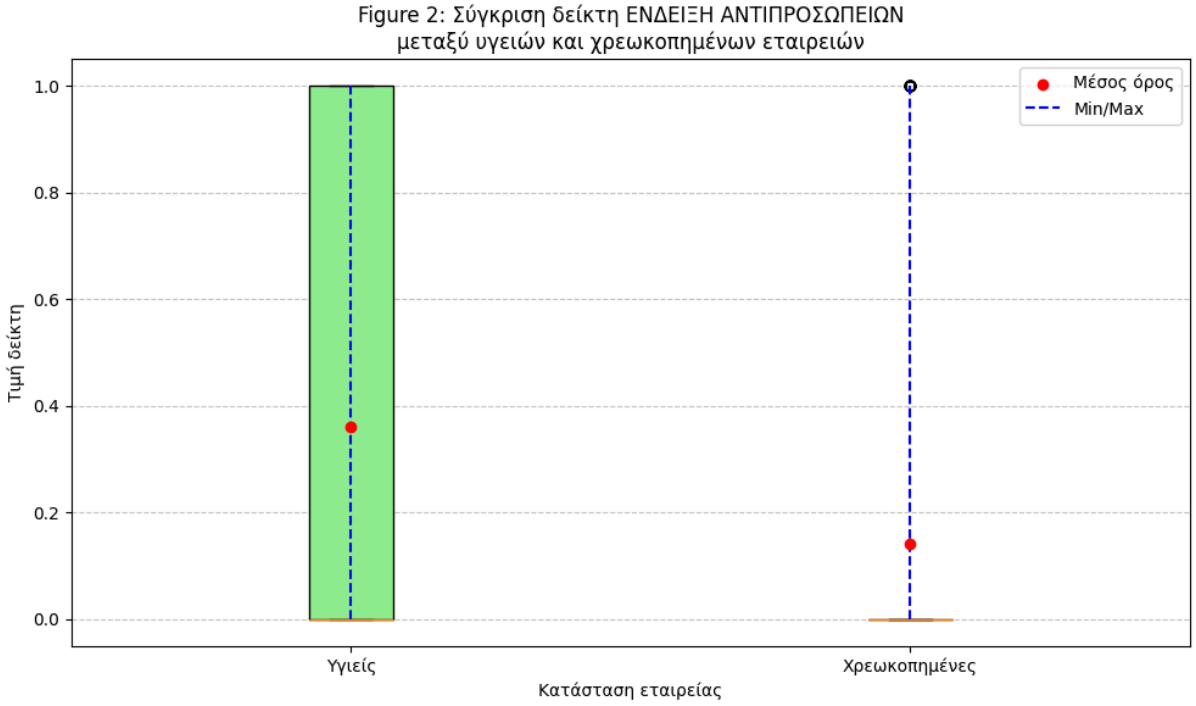
***Σχήμα 3.2.8**: Σύγκριση Δεικτών Απόδοσης μεταξύ Υγιών και Χρεωκοπημένων Εταιρειών - Boxplots: Σύγκριση δείκτη Λογάριθμος Προσωπικού μεταξύ υγειών και χρεωκοπημένων εταιρειών*



***Σχήμα 3.2.9**: Σύγκριση Δεικτών Απόδοσης μεταξύ Υγιών και Χρεωκοπημένων Εταιρειών - Boxplots: Σύγκριση δείκτη ΕΝΔΕΙΞΗ ΕΞΑΓΩΓΩΝ μεταξύ υγειών και χρεωκοπημένων εταιρειών*



***Σχήμα 3.2.10**: Σύγκριση Δεικτών Απόδοσης μεταξύ Υγιών και Χρεωκοπημένων Εταιρειών - Boxplots: Σύγκριση δείκτη ΈΝΔΕΙΞΗ ΕΙΣΑΓΩΓΩΝ μεταξύ υγειών και χρεωκοπημένων εταιρειών*



***Σχήμα 3.2.11**: Σύγκριση Δεικτών Απόδοσης μεταξύ Υγιών και Χρεωκοπημένων Εταιρειών - Boxplots: Σύγκριση δείκτη ΕΝΔΕΙΞΗ ΑΝΤΙΠΡΟΣΩΠΕΙΩΝ μεταξύ υγειών και χρεωκοπημένων εταιρειών*

Οι στατιστικές αναλύσεις των οικονομικών δεικτών αποκαλύπτουν σημαντικές διαφορές μεταξύ υγειών και χρεωκοπημένων εταιρειών. Στον δείκτη 365*(Β.Υ./Κόστος Πωλήσεων), οι υγιείς εταιρείες παρουσιάζουν μέση τιμή 441.65 (εύρος 0.65-1500), ενώ οι χρεωκοπημένες εμφανίζουν υψηλότερο μέσο όρο 646.92 (17.37-1500), γεγονός που μπορεί να υποδηλώνει λιγότερο αποδοτική διαχείριση αποθεμάτων.

Η απόδοση περιουσιακών στοιχείων (ROA) είναι θετική για τις υγιείς εταιρείες (μέσος όρος 0.04, -0.60 έως 0.50) ενώ οι χρεωκοπημένες εμφανίζουν αρνητική μέση απόδοση (-0.04). Ο λόγος χρηματοοικονομικών δαπανών προς πωλήσεις είναι διπλάσιος για τις χρεωκοπημένες εταιρείες (0.06 έναντι 0.03), υπογραμμίζοντας υψηλότερο λειτουργικό κόστος. Η πραγματική ρευστότητα διαφέρει σημαντικά (1.15 για υγιείς έναντι 0.85 για χρεωκοπημένες), ενώ οι χρεωκοπημένες εταιρείες χρειάζονται περισσότερο χρόνο για είσπραξη απαιτήσεων (278.52 ημέρες έναντι 211.91) και διαχείριση αποθεμάτων (183.98 ημέρες έναντι 107.06).

Ο δείκτης χρέους επιβεβαιώνει μεγαλύτερη χρηματοοικονομική πίεση στις χρεωκοπημένες εταιρείες (0.86 έναντι 0.74). Στον τομέα της διεθνούς δραστηριότητας, οι υγιείς εταιρείες εμφανίζουν υψηλότερες τιμές στους δείκτες εξαγωγών (0.39 έναντι 0.26), εισαγωγών (0.74 έναντι 0.50) και αντιπροσωπειών (0.36 έναντι 0.14). Τέλος, ο λογάριθμος του προσωπικού υποδηλώνει μεγαλύτερο μέγεθος για τις υγιείς εταιρείες (1.03 έναντι 0.59).

Αυτά τα ευρήματα καταδεικνύουν ότι οι χρεωκοπημένες εταιρείες αντιμετωπίζουν πολλαπλές δυσκολίες, από τη διαχείριση ρευστότητας και χρέους έως τον έλεγχο λειτουργικού κόστους και τη διεθνή παρουσία, ενώ οι υγιείς εταιρείες διατηρούν πιο ισορροπημένες χρηματοοικονομικές αναλογίες και αποτελεσματικότερες λειτουργικές διαδικασίες.

4. Πειραματικά Αποτελέσματα

4.1 Περιγραφή των Μοντέλων

Στην παρούσα μελέτη εφαρμόστηκαν διάφοροι αλγόριθμοι επιβλεπόμενης μάθησης με στόχο την ταξινόμηση των επιχειρήσεων σε υγιείς ή χρεωκοπημένες. Συγκεκριμένα, χρησιμοποιήθηκαν τα εξής μοντέλα: (Protopapadakis, nd)

- **Linear Discriminant Analysis (LDA**): Αλγόριθμος βασισμένος στη γραμμική διαχωριστική ανάλυση, που προσπαθεί να βρει το καλύτερο γραμμικό σύνορο μεταξύ των κλάσεων.
- **Logistic Regression**: Κλασικό στατιστικό μοντέλο που εκτιμά την πιθανότητα ανήκειν σε μία από τις δύο κατηγορίες μέσω της λογαριθμικής συνάρτησης.
- **Decision Trees**: Μοντέλο που χωρίζει τα δεδομένα σε υποσύνολα βάσει χαρακτηριστικών, δημιουργώντας δέντρο αποφάσεων.
- **Random Forests:** Σύνολο από δέντρα αποφάσεων που βελτιώνει τη σταθερότητα και την ακρίβεια μέσω της ψηφοφορίας μεταξύ πολλαπλών δένδρων.
- **k-Nearest Neighbors (k-NN)**: Αλγόριθμος βασισμένος στην εγγύτητα των δεδομένων στο χώρο χαρακτηριστικών, ταξινομώντας ένα δείγμα βάσει των κοντινότερων γειτόνων του.
- **Naïve Bayes:** Βασίζεται στον θεώρημα του Bayes, με απλοποιητική υπόθεση ανεξαρτησίας μεταξύ των χαρακτηριστικών.
- **Support Vector Machines (SVM)**: Προσπαθεί να βρει το υπερεπίπεδο που μεγιστοποιεί το περιθώριο διαχωρισμού μεταξύ των κατηγοριών.
- **Gradient Boosting:** Πρόκειται για μια ισχυρή μέθοδος που συνδυάζει πολλούς αδύναμους προβλεπτές, συνήθως δέντρα απόφασης, δημιουργώντας ένα ενιαίο, πιο ακριβές μοντέλο. Η βελτίωση της απόδοσης επιτυγχάνεται σταδιακά, καθώς κάθε επόμενο μοντέλο εκπαιδεύεται ώστε να διορθώσει τα σφάλματα του προηγούμενου.

4.2 Αποτελέσματα Ανάλυσης

Στην παρούσα ενότητα παρουσιάζονται αναλυτικά τα αποτελέσματα των διαφορετικών αλγορίθμων ταξινόμησης που εφαρμόστηκαν στα δεδομένα μας. Η ανάλυση καλύπτει τις μετρικές απόδοσης όπως ROC-AUC, Accuracy, precision, recall, F1-score. Επιπλέον, για να ληφθεί υπόψη η ισορροπία των κλάσεων και η συνολική ποιότητα της ταξινόμησης, χρησιμοποιήθηκαν και δύο επιπλέον μετρικές: η Balanced Accuracy (BA) και ο Matthews Correlation Coefficient (MCC). Οι μετρικές αυτές παρέχουν μια πιο ολοκληρωμένη εικόνα της απόδοσης των μοντέλων. (Protopapadakis, nd)

Classifier N	# Fold	Training or	Balanced o	Number of	Number of	True positi	True negati	False positi	False negat	ROC-AUC	Accuracy	Precision	Recall	F1-score	Balanced a	Matthews (
Linear Discrimin	1	test	balanced	744	186	31	2399	218	31	0,830334	0,90705	0,12450	0,50000	0,19936	0,7083492549	0,2157828093
Logistic Regress	1	test	balanced	744	186	30	2431	186	32	0,834870	0,91863	0,13889	0,48387	0,21583	0,7063986096	0,2279687866
Decision Trees	1	test	balanced	744	186	35	2149	468	27	0,692843	0,81523	0,06958	0,56452	0,12389	0,6928427034	0,1484969299
Random Forests	1	test	balanced	744	186	31	2382	235	31	0,868897	0,90071	0,11654	0,50000	0,18902	0,705101261	0,2062417347
k-Nearest Neigh	1	test	balanced	744	186	33	2311	306	29	0,820220	0,87495	0,09735	0,53226	0,16459	0,7076651423	0,1878380314
Naïve Bayes	1	test	balanced	744	186	30	2318	299	32	0,824011	0,87645	0,09119	0,48387	0,15345	0,6848090032	0,1693241664
Support Vector M	1	test	balanced	744	186	32	2422	195	30	0,847332	0,91601	0,14097	0,51613	0,22145	0,7208081157	0,2384347616
Gradient Boostir	1	test	balanced	744	186	37	2347	270	25	0,845871	0,88988	0,12052	0,59677	0,20054	0,7468013115	0,232996091
Linear Discrimin	2	test	balanced	744	186	31	2438	179	31	0,867701	0,92161	0,14762	0,50000	0,22794	0,715800535	0,2414404469
Logistic Regress	2	test	balanced	744	186	27	2473	144	35	0,858826	0,93318	0,15789	0,43548	0,23176	0,6902295167	0,234015048
Decision Trees	2	test	balanced	744	186	28	2148	469	34	0,636200	0,81224	0,05634	0,45161	0,10018	0,636200032	0,1053657212
Random Forests	2	test	balanced	744	186	37	2444	173	25	0,895035	0,92609	0,17619	0,59677	0,27206	0,7653339825	0,2968591127
k-Nearest Neigh	2	test	balanced	744	186	31	2354	263	31	0,845976	0,89026	0,10544	0,50000	0,17416	0,699751624	0,1921764487
Naïve Bayes	2	test	balanced	744	186	43	2265	352	19	0,843745	0,86152	0,10886	0,69355	0,18818	0,7795216143	0,2370808632
Support Vector M	2	test	balanced	744	186	28	2501	116	34	0,850352	0,94401	0,19444	0,45161	0,27184	0,7036436698	0,271536608
Gradient Boostir	2	test	balanced	744	186	35	2384	233	27	0,860577	0,90295	0,13060	0,56452	0,21212	0,7377414424	0,2382684792
Linear Discrimin	3	test	balanced	744	186	19	2435	182	43	0,796301	0,91601	0,09453	0,30645	0,14449	0,618453166	0,1352152561
Logistic Regress	3	test	balanced	744	186	14	2452	165	48	0,795530	0,92049	0,07821	0,22581	0,11618	0,5813785793	0,09800343219
Decision Trees	3	test	balanced	744	186	24	2235	382	38	0,620564	0,84323	0,05911	0,38710	0,10256	0,6205640539	0,101107369
Random Forests	3	test	balanced	744	186	22	2438	179	40	0,778896	0,91825	0,10945	0,35484	0,16730	0,6432198898	0,1634866735
k-Nearest Neigh	3	test	balanced	744	186	17	2371	246	45	0,727033	0,89138	0,06464	0,27419	0,10462	0,5900963921	0,09105604071
Naïve Bayes	3	test	balanced	744	186	22	2290	327	40	0,770964	0,86301	0,06304	0,35484	0,10706	0,6149432371	0,1026881788
Support Vector M	3	test	balanced	744	186	15	2477	140	47	0,794772	0,93020	0,09677	0,24194	0,13825	0,594219557	0,1213551567
Gradient Boostir	3	test	balanced	744	186	21	2411	206	41	0,784726	0,90780	0,09251	0,33871	0,14533	0,6299967951	0,140374165
Linear Discrimin	4	test	balanced	744	186	21	2422	195	41	0,850666	0,91191	0,09722	0,33871	0,15108	0,6320984383	0,1459036993
Logistic Regress	4	test	balanced	744	186	20	2445	172	42	0,846568	0,92012	0,10417	0,32258	0,15748	0,6284282668	0,1497268943
Decision Trees	4	test	balanced	744	186	29	2185	432	33	0,651334	0,82643	0,06291	0,46774	0,11090	0,6513337113	0,1205680262
Random Forests	4	test	balanced	744	186	30	2421	196	32	0,864348	0,91489	0,13274	0,48387	0,20833	0,7044880249	0,2212547425
k-Nearest Neigh	4	test	balanced	744	186	31	2337	280	31	0,807943	0,88391	0,09968	0,50000	0,16622	0,6965036301	0,1844706321
Naïve Bayes	4	test	balanced	744	186	32	2302	315	30	0,838771	0,87122	0,09222	0,51613	0,15648	0,6978811	0,1772160316
Support Vector M	4	test	balanced	744	186	24	2439	178	38	0,850229	0,91937	0,11881	0,38710	0,18182	0,6595399805	0,181701515
Gradient Boostir	4	test	balanced	744	186	31	2396	221	31	0,864392	0,90594	0,12302	0,50000	0,19745	0,7077760795	0,2140366231

***Σχήμα 4.2**: Αποτελέσματα Ανάλυσης*

4.5 Ανάλυση Confusion Matrices – 4-Fold Cross Validation

Για την αξιολόγηση της επίδοσης των μοντέλων εφαρμόστηκε η μέθοδος 4-Fold Cross Validation. Παρουσιάζονται τα confusion matrices όλων των folds και η διαδικασία αξιολόγησης τους. Τα αποτελέσματα παρουσίασαν παρόμοια σταθερότητα και επαναληψιμότητα. (Protopapadakis, nd)

Η ανάλυση επικεντρώνεται ιδιαίτερα στην απόδοση ως προς την κλάση “Bankrupt”, λόγω της σημασίας της ορθής πρόβλεψης χρεοκοπημένων επιχειρήσεων.

Το Linear Discriminant Analysis (LDA) εμφάνισε σχετικά καλή επίδοση στην ταξινόμηση υγιών επιχειρήσεων ("Healthy"), ωστόσο παρατηρήθηκε κάποια δυσκολία στην αναγνώριση περιπτώσεων "Bankrupt", με αυξημένο αριθμό των false negative τιμών, γεγονός που ενδέχεται να περιορίζει την ευαισθησία του μοντέλου στις περιπτώσεις υψηλού κινδύνου.

Το Logistic Regression σημείωσε ισορροπημένη επίδοση μεταξύ των δύο κατηγοριών, διατηρώντας σταθερή συμπεριφορά στα training και test sets. Το μοντέλο έδειξε ικανοποιητική ικανότητα γενίκευσης, χωρίς σημαντικές αποκλίσεις ανάμεσα στα διαφορετικά folds.

Τα Decision Trees παρουσίασαν πολύ καλή προσαρμογή στο training set, χωρίς λάθη, γεγονός που υποδεικνύει πιθανή υπερεκπαίδευση (overfitting). Στο test set, ωστόσο, η απόδοση μειώθηκε, ιδίως ως προς την κλάση "Bankrupt", κάτι που επαναλήφθηκε και στα υπόλοιπα folds.

Τα Random Forests παρουσίασαν σταθερά και παρόμοια χαρακτηριστικά με τα απλά decision trees, με βελτιωμένη συμπεριφορά στο test set. Παρ’ όλα αυτά, καταγράφηκε σχετική δυσκολία στην πλήρη αναγνώριση των "Bankrupt" επιχειρήσεων, με περιορισμένα false positives αλλά παρουσία false negative.

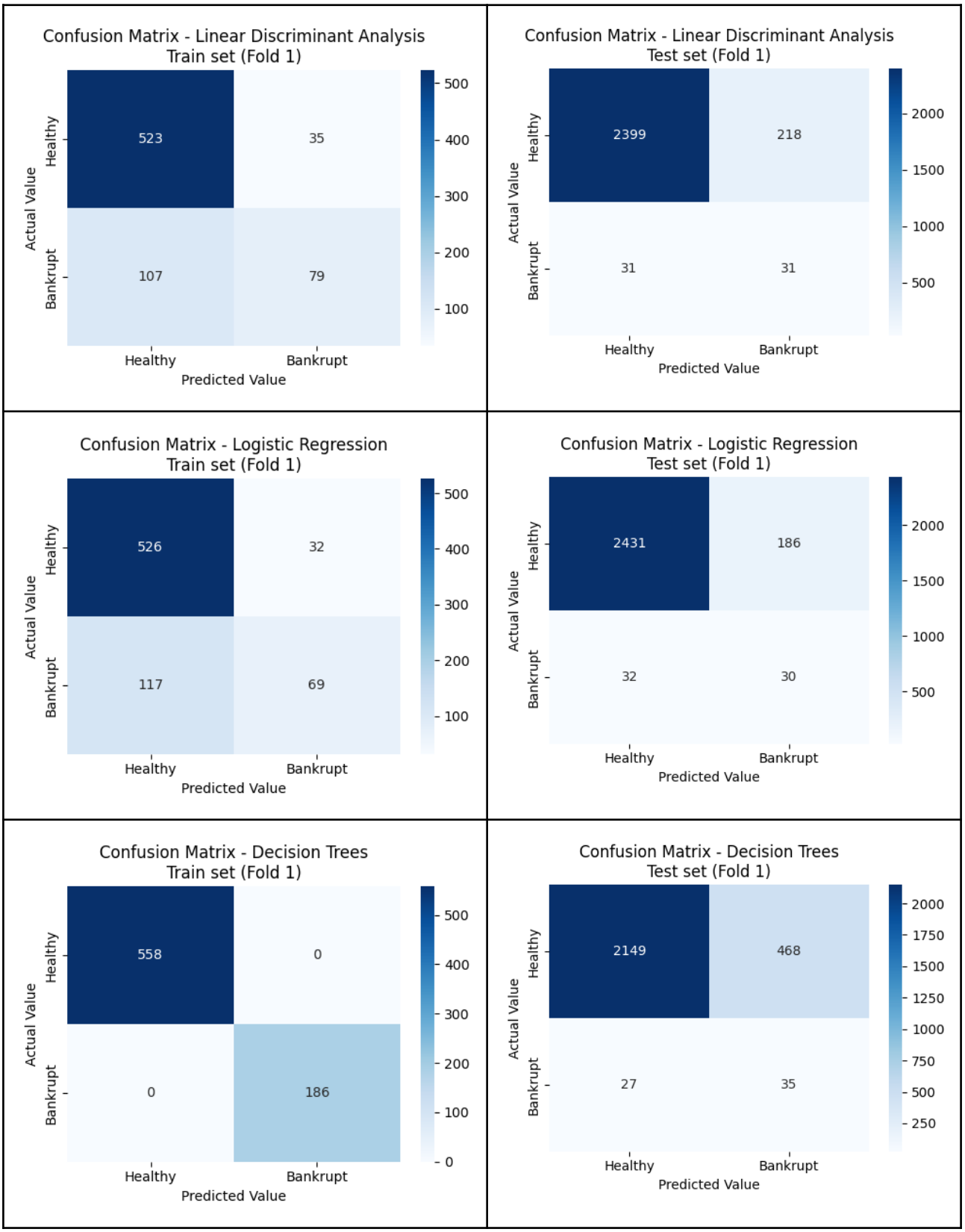
Το k-Nearest Neighbors (k-NN) είναι ευαίσθητο στην κατανομή των δεδομένων.Η απόδοσή του φαίνεται να επηρεάζεται περισσότερο από τη δομή των δεδομένων και τις ιδιαιτερότητες κάθε fold. Η συμπεριφορά του ήταν λιγότερο σταθερή σε σύγκριση με άλλους αλγορίθμους, γεγονός που σχετίζεται με την τοπική φύση του μοντέλου και την ευαισθησία του σε προβλήματα όπως η ανισορροπία των κατηγοριών.

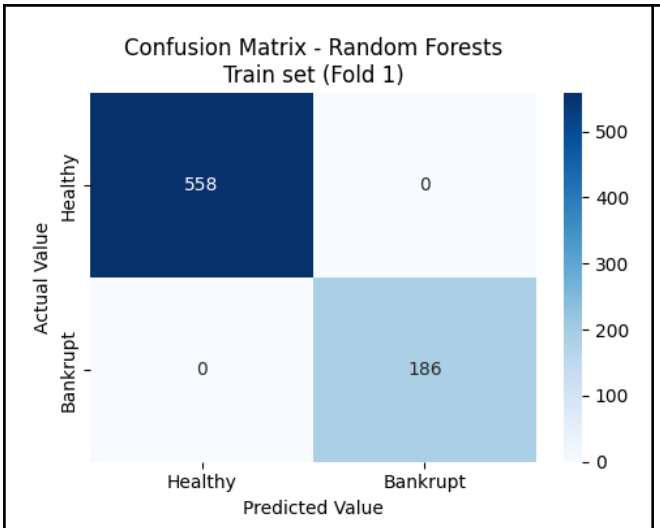
Το Naive Bayes παρουσίασε περιορισμένη ικανότητα πρόβλεψης της χρεοκοπίας, κυρίως λόγω της υπόθεσης ανεξαρτησίας μεταξύ των χαρακτηριστικών. Παρ' όλα αυτά, η συμπεριφορά του ήταν σταθερή σε όλα τα folds, χωρίς έντονη απόκλιση μεταξύ training και test sets.

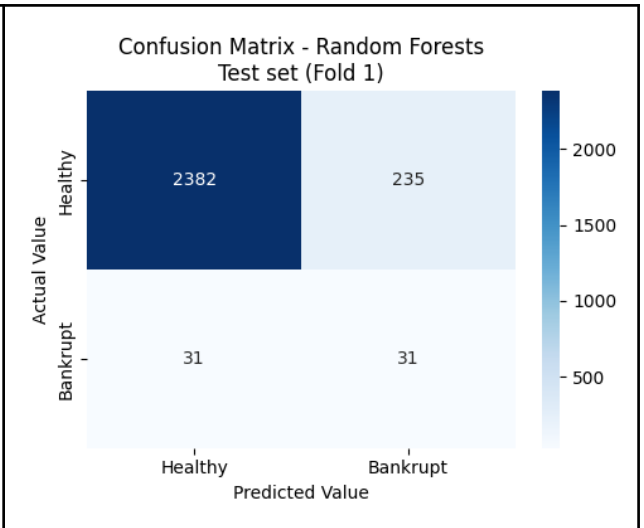
Το μοντέλο Support Vector Machines (SVM) κατέγραψε συνεπή και σχετικά ισορροπημένη απόδοση για τις δύο κατηγορίες, τόσο στο training όσο και στο test set. Τα αποτελέσματα ήταν σταθερά σε όλα τα folds, υποδεικνύοντας καλή ικανότητα γενίκευσης.

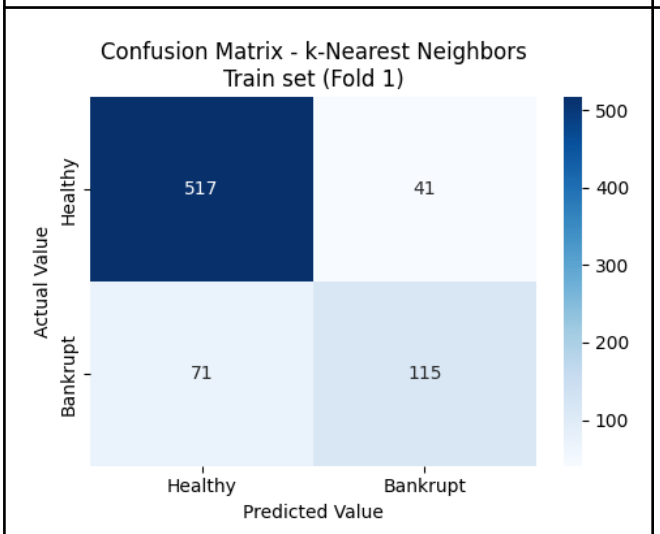
Τέλος, το Gradient Boosting εμφάνισε υψηλή ακρίβεια στο training set και αρκετά ικανοποιητική απόδοση στο test set. Η επίδοσή του ως προς την κλάση "Bankrupt" ήταν βελτιωμένη συγκριτικά με άλλα μοντέλα, διατηρώντας όμως και αυτό ορισμένα false negatives. Η συμπεριφορά του ήταν σταθερή σε όλα τα folds, χωρίς έντονα φαινόμενα υπερεκπαιδευσης.

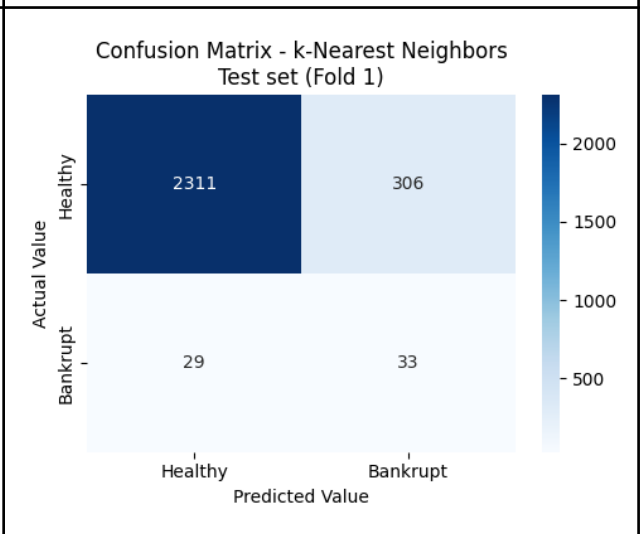
FOLD 1

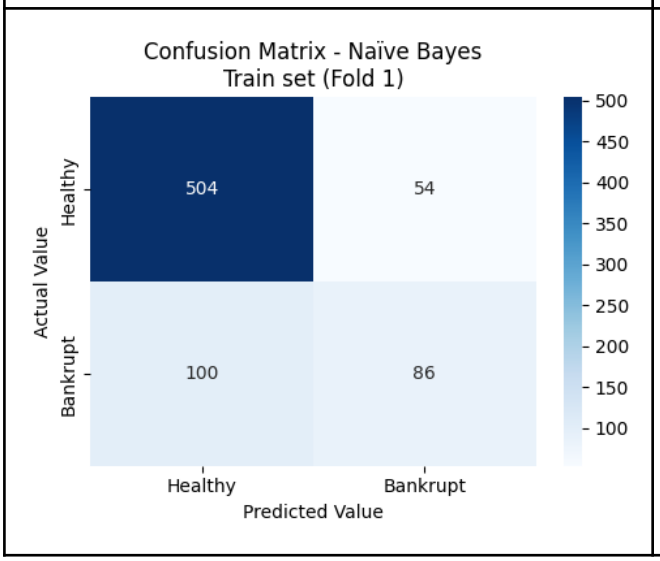

Confusion Matrix - Linear Discriminant Analysis
Train set (Fold 1)
Actual Value
Healthy
Bankrupt
523
35
107
79
Healthy
Bankrupt
Predicted Value
500
400
300
200
100
Confusion Matrix - Linear Discriminant Analysis
Test set (Fold 1)
Actual Value
Healthy
Bankrupt
2399
218
31
31
Healthy
Bankrupt
Predicted Value
2000
1500
1000
500
Confusion Matrix - Logistic Regression
Train set (Fold 1)
Actual Value
Healthy
Bankrupt
526
32
117
69
Healthy
Bankrupt
Predicted Value
500
400
300
200
100
Confusion Matrix - Logistic Regression
Test set (Fold 1)
Actual Value
Healthy
Bankrupt
2431
186
32
30
Healthy
Bankrupt
Predicted Value
2000
1500
1000
500
Confusion Matrix - Decision Trees
Train set (Fold 1)
Actual Value
Healthy
Bankrupt
558
0
0
186
Healthy
Bankrupt
Predicted Value
500
400
300
200
100
0
Confusion Matrix - Decision Trees
Test set (Fold 1)
Actual Value
Healthy
Bankrupt
2149
468
27
35
Healthy
Bankrupt
Predicted Value
2000
1750
1500
1250
1000
750
500
250

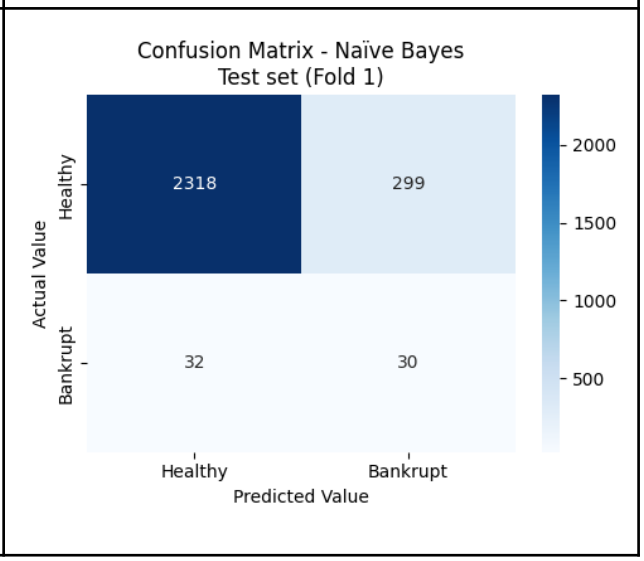

Confusion Matrix - Random Forests
Train set (Fold 1)
Actual Value
Healthy
Bankrupt
558
0
0
186
Healthy
Bankrupt
Predicted Value
500
400
300
200
100
0

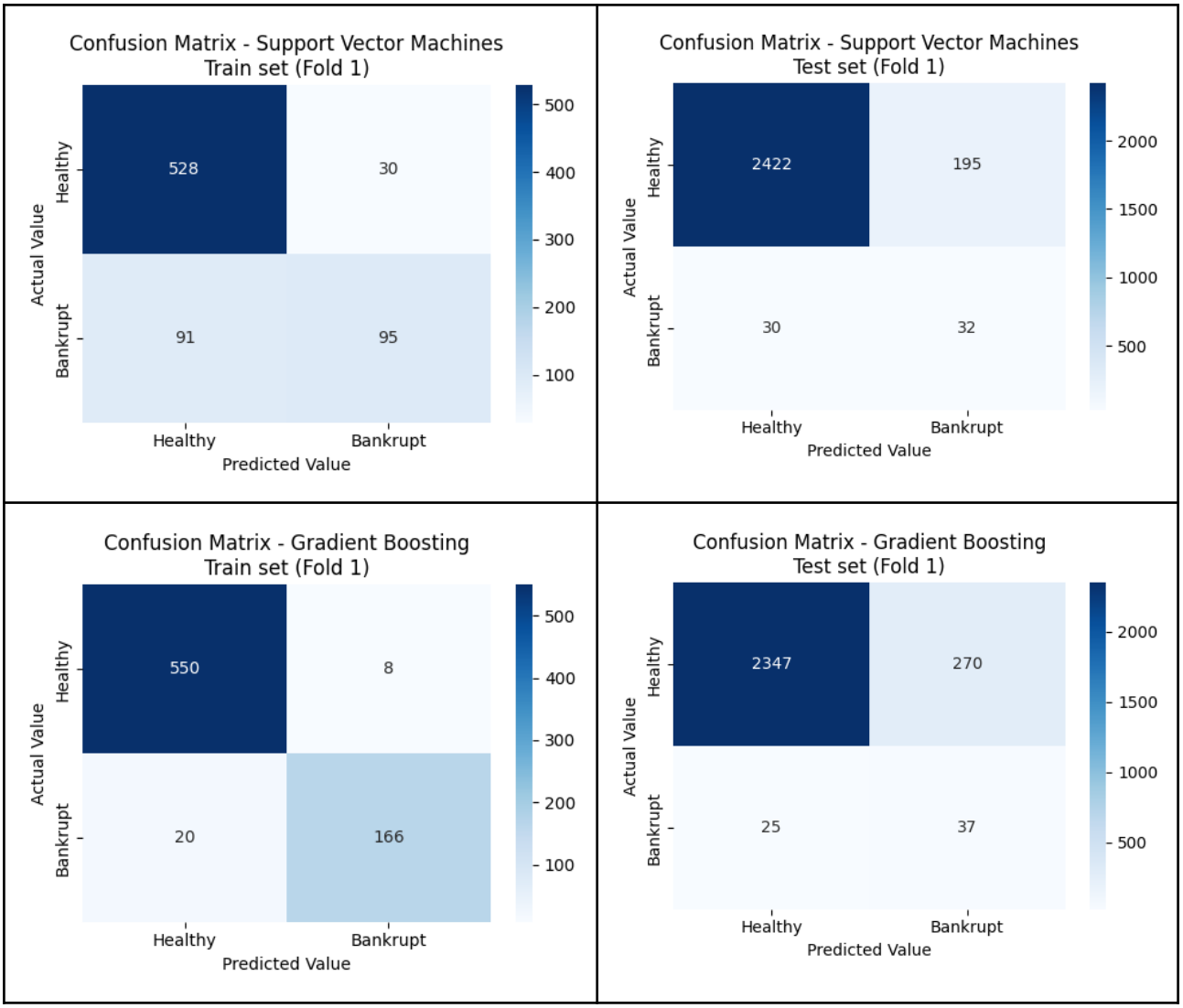

Confusion Matrix - Random Forests
Test set (Fold 1)
Actual Value
Healthy
Bankrupt
2382
235
31
31
Healthy
Bankrupt
Predicted Value
2000
1500
1000
500


Confusion Matrix - k-Nearest Neighbors
Train set (Fold 1)
Actual Value
Healthy
Bankrupt
517
41
71
115
Healthy
Bankrupt
Predicted Value
500
400
300
200
100


Confusion Matrix - k-Nearest Neighbors
Test set (Fold 1)
Actual Value
Healthy
Bankrupt
2311
306
29
33
Healthy
Bankrupt
Predicted Value
2000
1500
1000
500

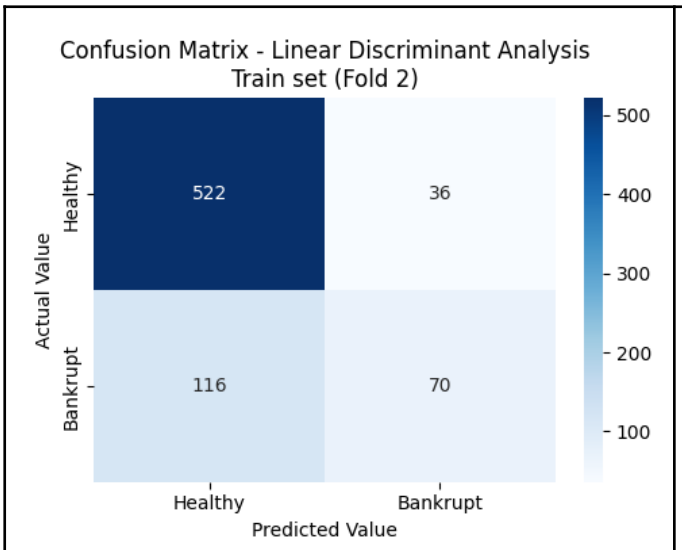

Confusion Matrix - Naïve Bayes
Train set (Fold 1)
Actual Value
Healthy
Bankrupt
504
54
100
86
Healthy
Bankrupt
Predicted Value
500
450
400
350
300
250
200
150
100

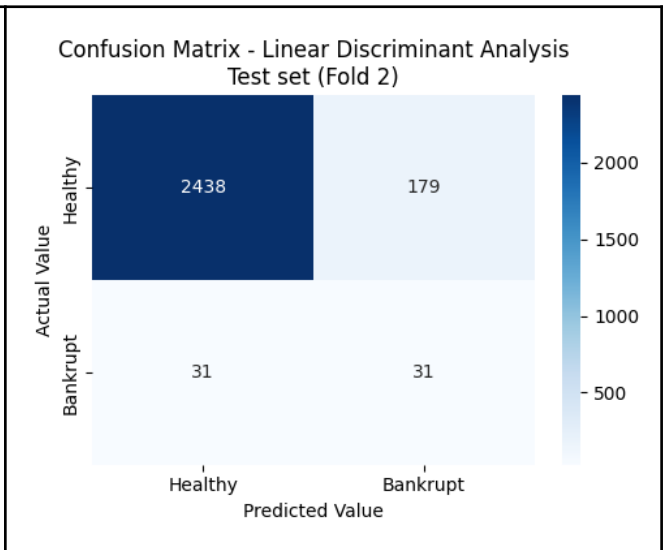

Confusion Matrix - Naïve Bayes
Test set (Fold 1)
Actual Value
Healthy
Bankrupt
2318
299
32
30
Healthy
Bankrupt
Predicted Value
2000
1500
1000
500

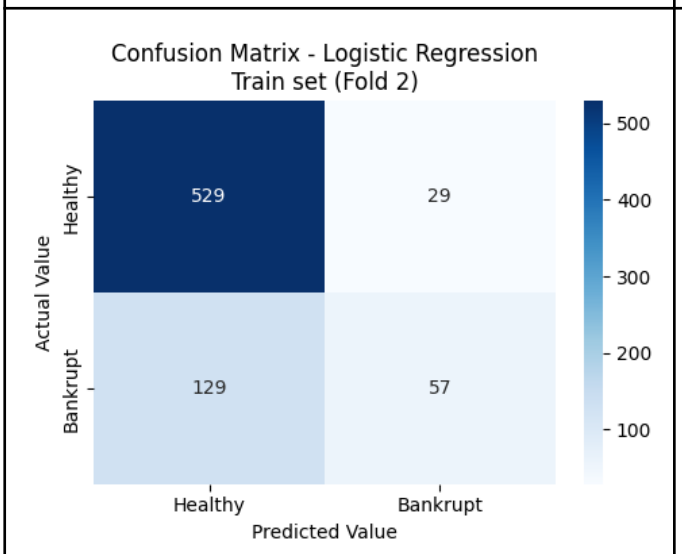


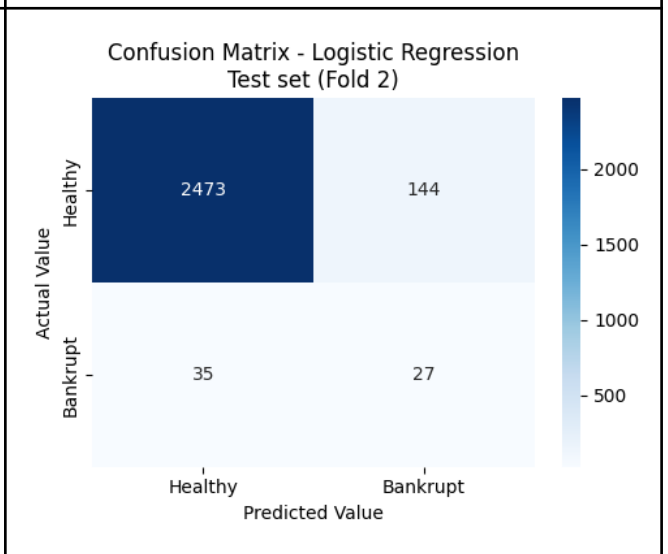
***Σχήμα 4.3.1**: Confusion Matrices – Fold 1*

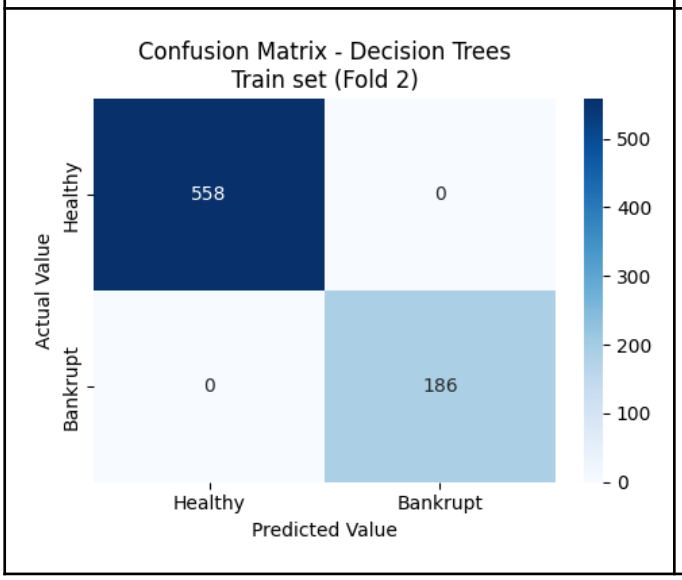
FOLD 2

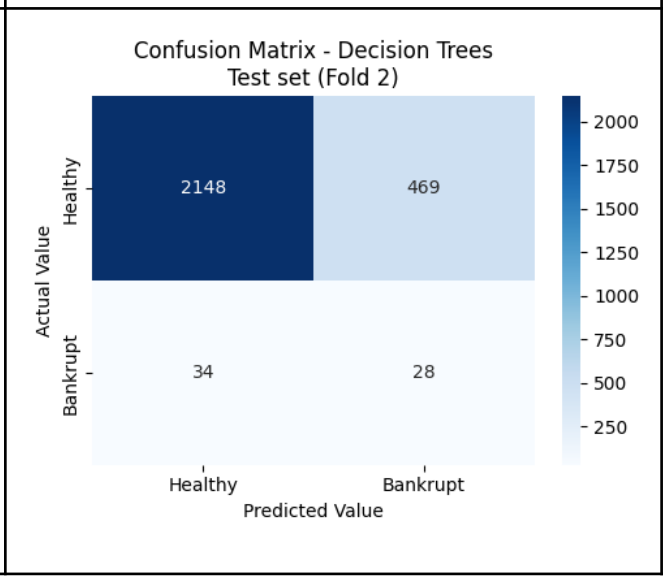

Confusion Matrix - Linear Discriminant Analysis
Train set (Fold 2)
Actual Value
Healthy
Bankrupt
522
36
116
70
Healthy
Bankrupt
Predicted Value
500
400
300
200
100

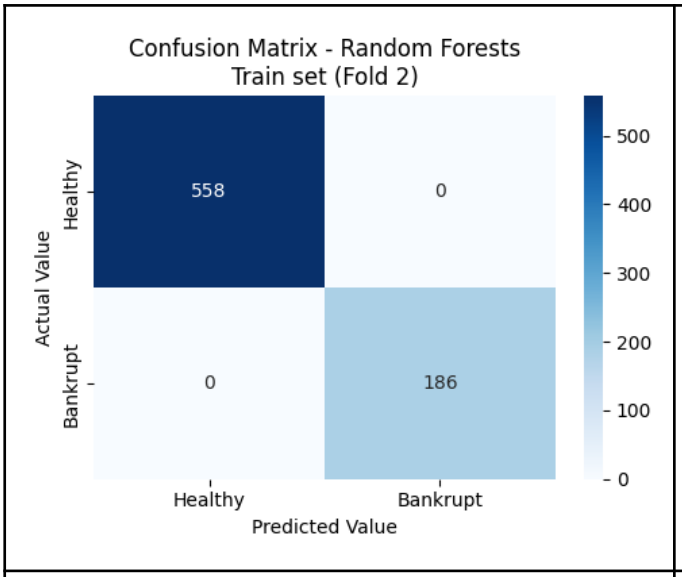

Confusion Matrix - Linear Discriminant Analysis
Test set (Fold 2)
Actual Value
Healthy
Bankrupt
2438
179
31
31
Healthy
Bankrupt
Predicted Value
2000
1500
1000
500

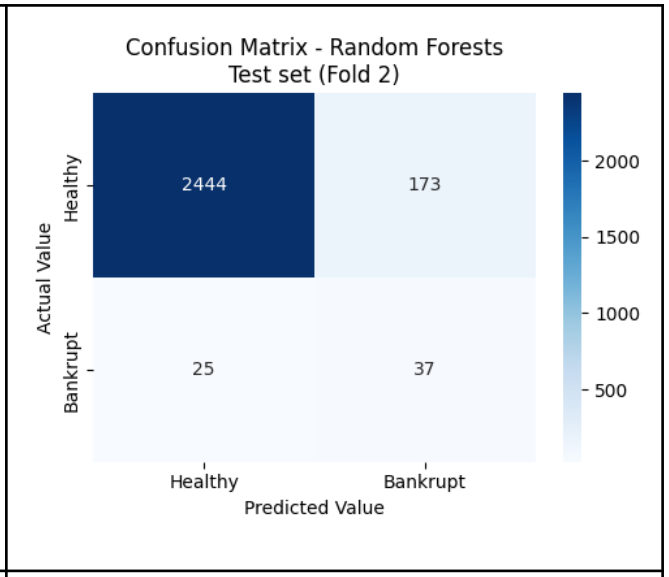

Confusion Matrix - Logistic Regression
Train set (Fold 2)
Actual Value
Healthy
Bankrupt
529
29
129
57
Healthy
Bankrupt
Predicted Value
500
400
300
200
100

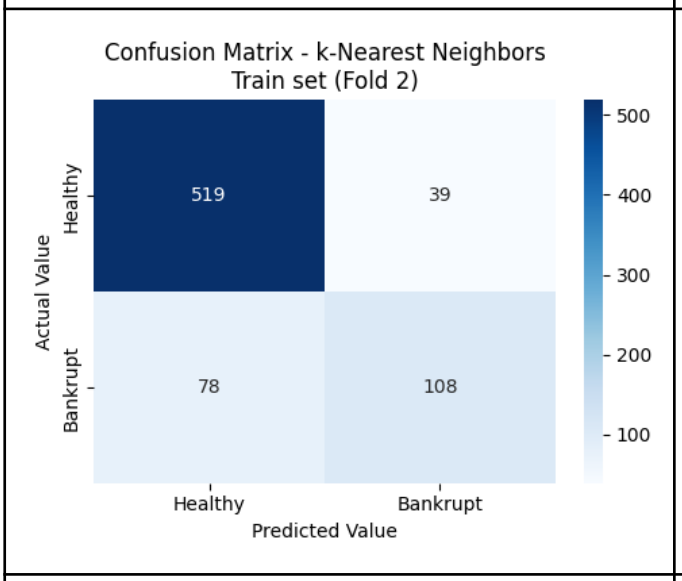

Confusion Matrix - Logistic Regression
Test set (Fold 2)
Actual Value
Healthy
Bankrupt
2473
144
35
27
Healthy
Bankrupt
Predicted Value
2000
1500
1000
500

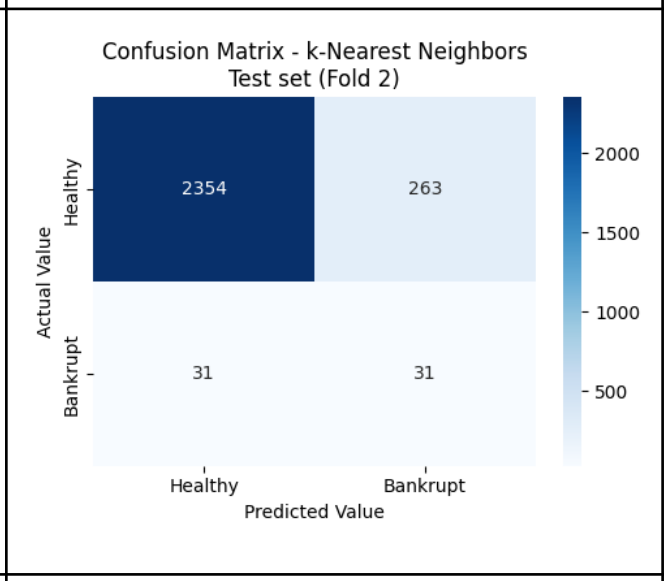

Confusion Matrix - Decision Trees
Train set (Fold 2)
Actual Value
Healthy
Bankrupt
558
0
0
186
Healthy
Bankrupt
Predicted Value
500
400
300
200
100
0

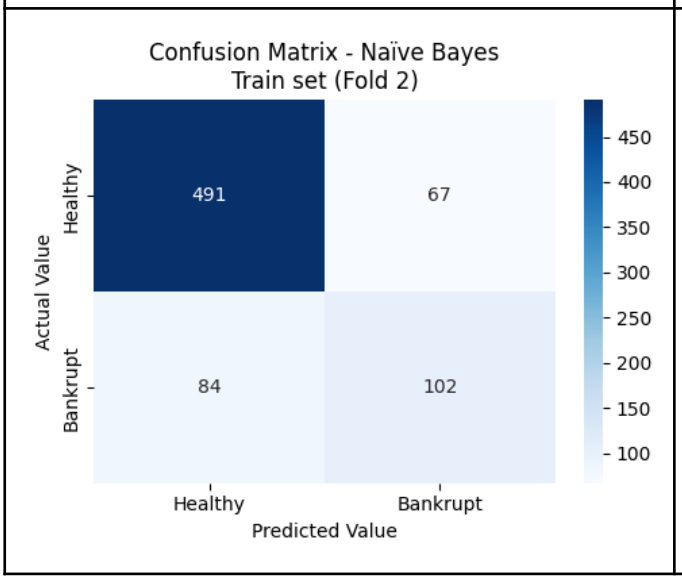

Confusion Matrix - Decision Trees
Test set (Fold 2)
Actual Value
Healthy
Bankrupt
2148
469
34
28
Healthy
Bankrupt
Predicted Value
2000
1750
1500
1250
1000
750
500
250

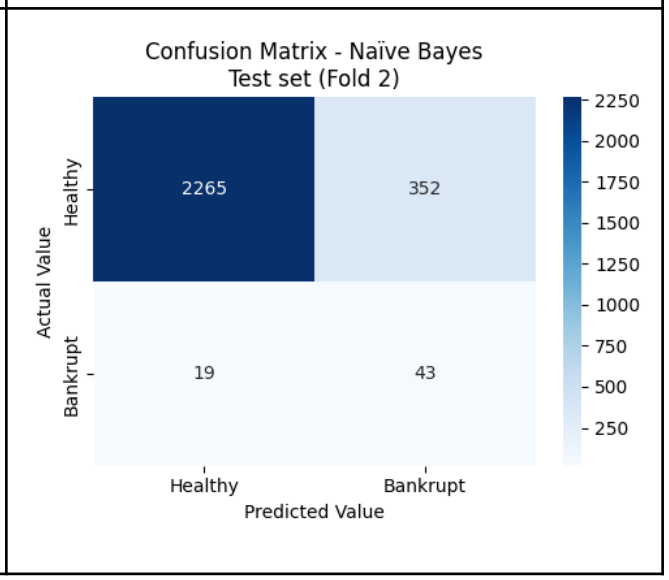

Confusion Matrix - Random Forests
Train set (Fold 2)
Actual Value
Healthy
Bankrupt
558
0
0
186
Healthy
Bankrupt
Predicted Value
500
400
300
200
100
0

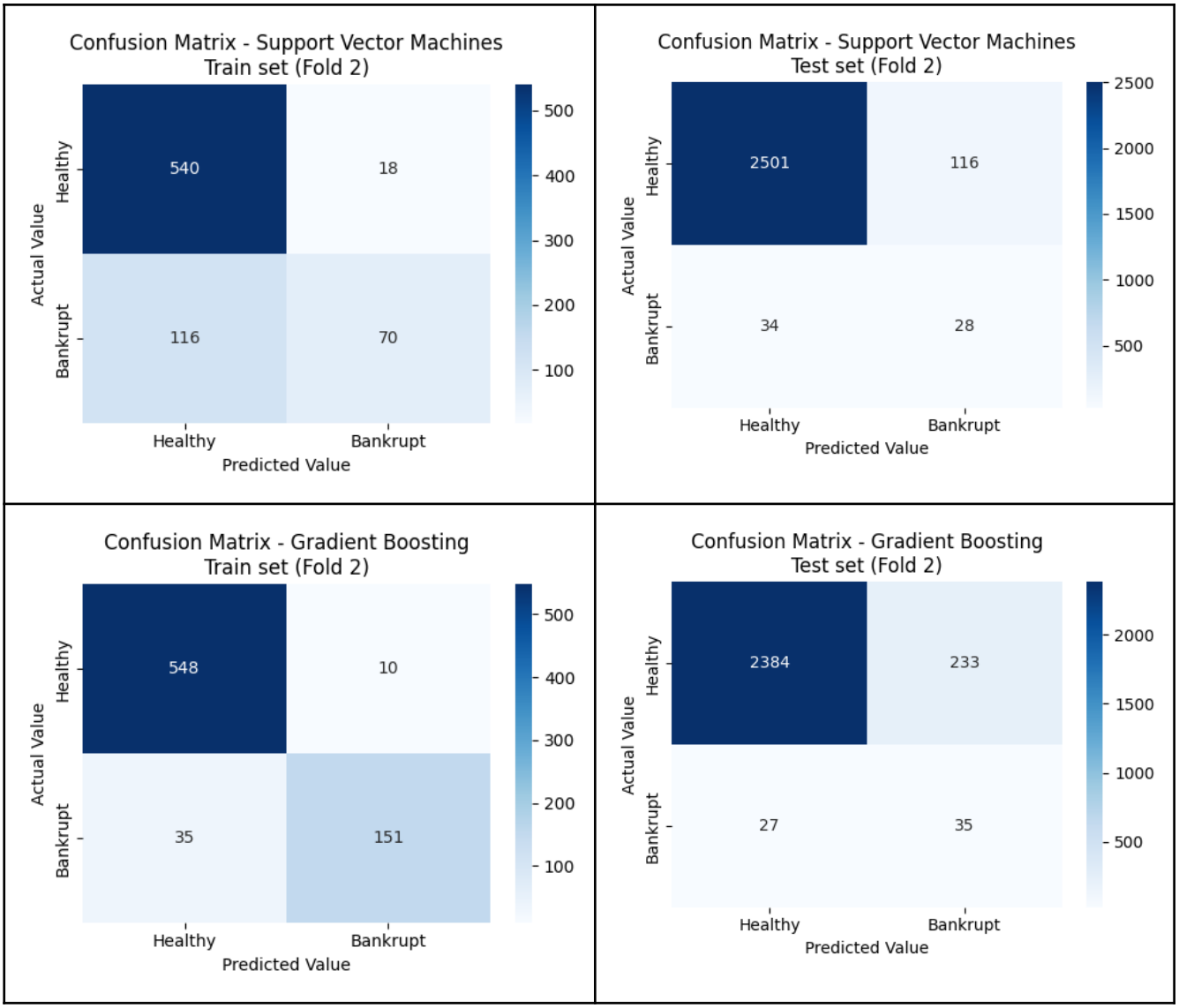

Confusion Matrix - Random Forests
Test set (Fold 2)
Actual Value
Healthy
Bankrupt
2444
173
25
37
Healthy
Bankrupt
Predicted Value
2000
1500
1000
500


Confusion Matrix - k-Nearest Neighbors
Train set (Fold 2)
Actual Value
Healthy
Bankrupt
519
39
78
108
Healthy
Bankrupt
Predicted Value
500
400
300
200
100


Confusion Matrix - k-Nearest Neighbors
Test set (Fold 2)
Actual Value
Healthy
Bankrupt
2354
263
31
31
Healthy
Bankrupt
Predicted Value
2000
1500
1000
500

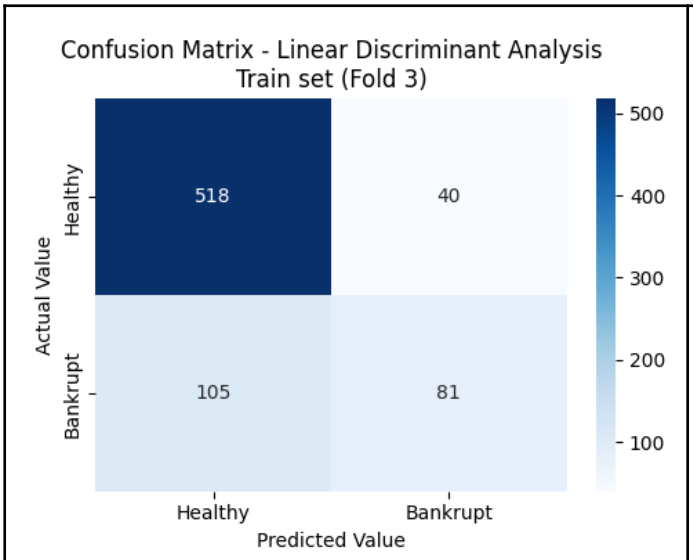

Confusion Matrix - Naïve Bayes
Train set (Fold 2)
Actual Value
Healthy
Bankrupt
491
67
84
102
Healthy
Bankrupt
Predicted Value
450
400
350
300
250
200
150
100

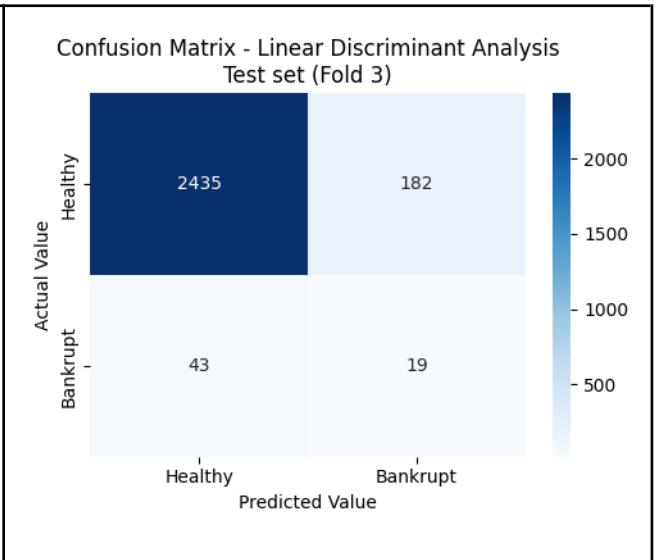

Confusion Matrix - Naïve Bayes
Test set (Fold 2)
Actual Value
Healthy
Bankrupt
2265
352
19
43
Healthy
Bankrupt
Predicted Value
2250
2000
1750
1500
1250
1000
750
500
250

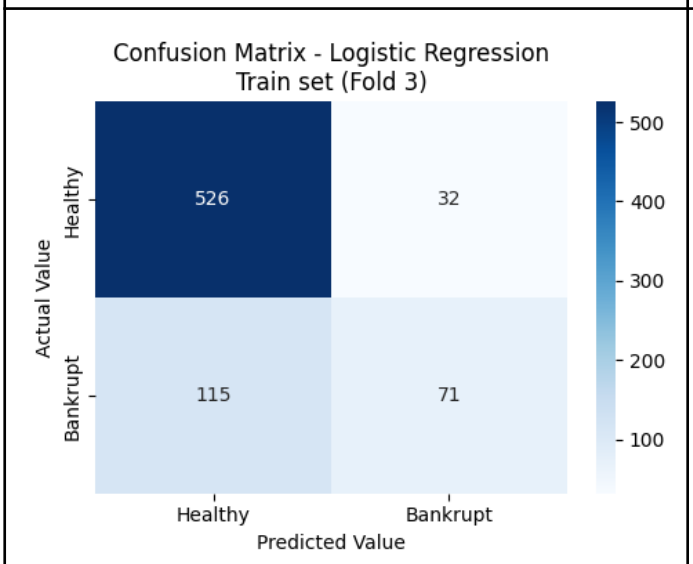


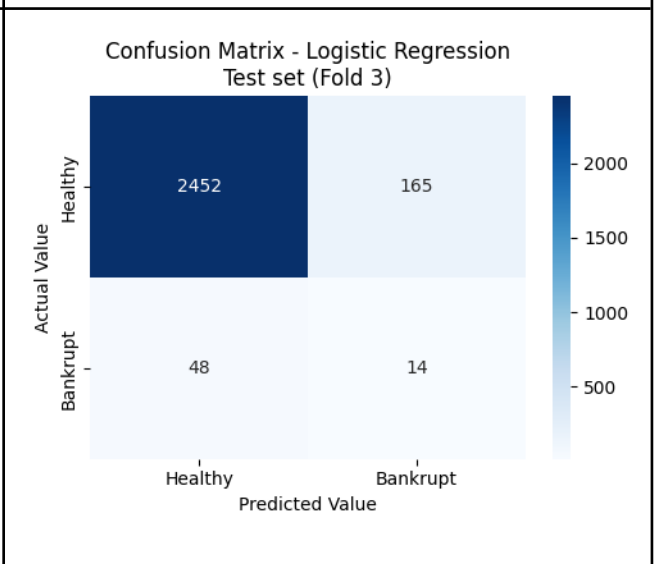
***Σχήμα 4.3.2**: Confusion Matrices – Fold 2*

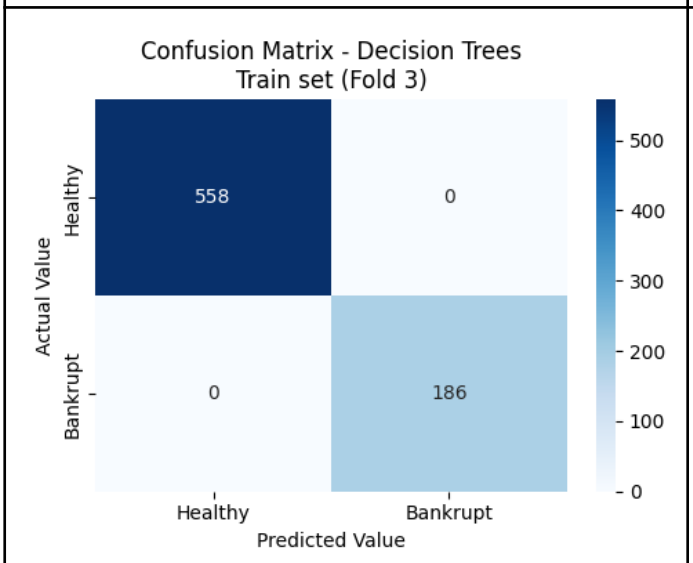
FOLD 3

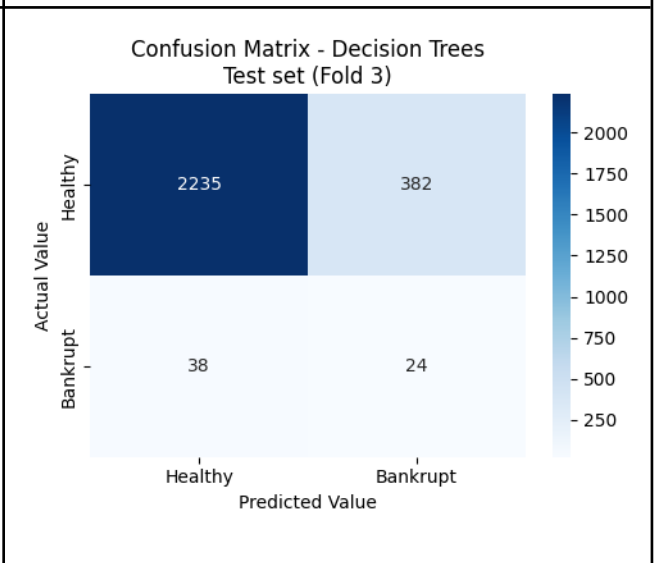

Confusion Matrix - Linear Discriminant Analysis
Train set (Fold 3)
Actual Value
Healthy
Bankrupt
518
40
105
81
Healthy
Bankrupt
Predicted Value
500
400
300
200
100

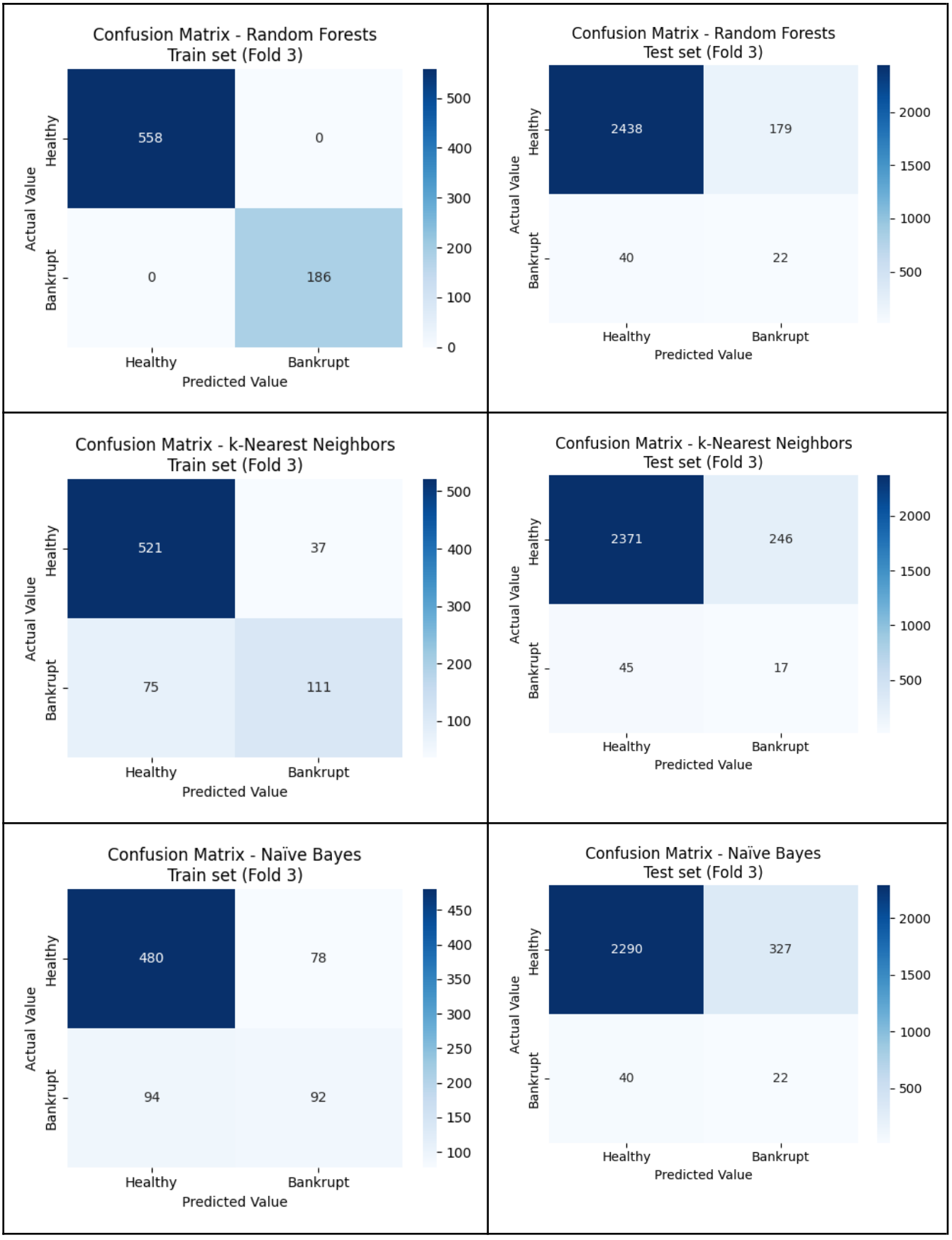

Confusion Matrix - Linear Discriminant Analysis
Test set (Fold 3)
Actual Value
Healthy
Bankrupt
2435
182
43
19
Healthy
Bankrupt
Predicted Value
2000
1500
1000
500

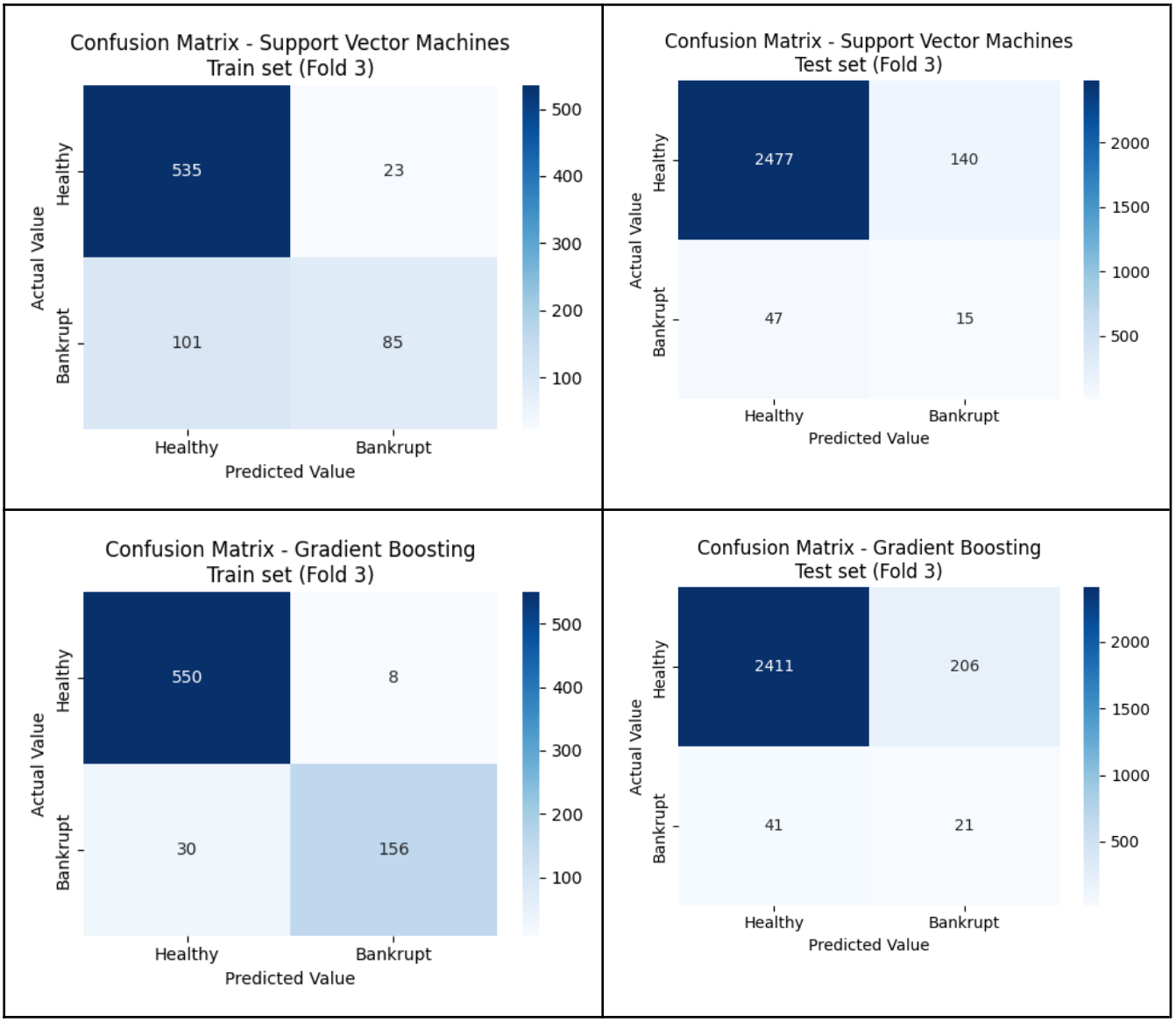

Confusion Matrix - Logistic Regression
Train set (Fold 3)
Actual Value
Healthy
Bankrupt
526
32
115
71
Healthy
Bankrupt
Predicted Value
500
400
300
200
100


Confusion Matrix - Logistic Regression
Test set (Fold 3)
Actual Value
Healthy
Bankrupt
2452
165
48
14
Healthy
Bankrupt
Predicted Value
2000
1500
1000
500


Confusion Matrix - Decision Trees
Train set (Fold 3)
Actual Value
Healthy
Bankrupt
558
0
0
186
Healthy
Bankrupt
Predicted Value
500
400
300
200
100
0

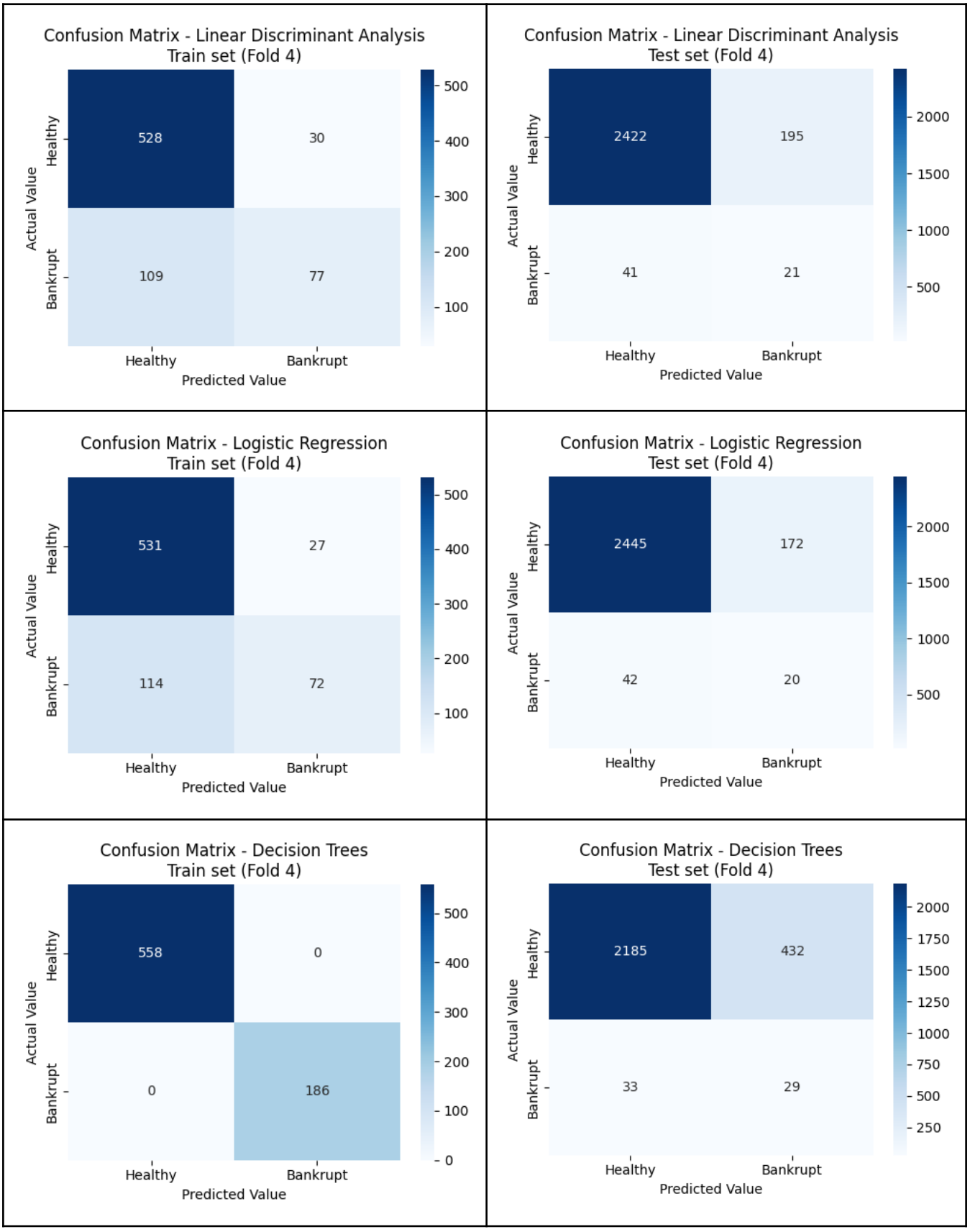

Confusion Matrix - Decision Trees
Test set (Fold 3)
Actual Value
Healthy
Bankrupt
2235
382
38
24
Healthy
Bankrupt
Predicted Value
2000
1750
1500
1250
1000
750
500
250

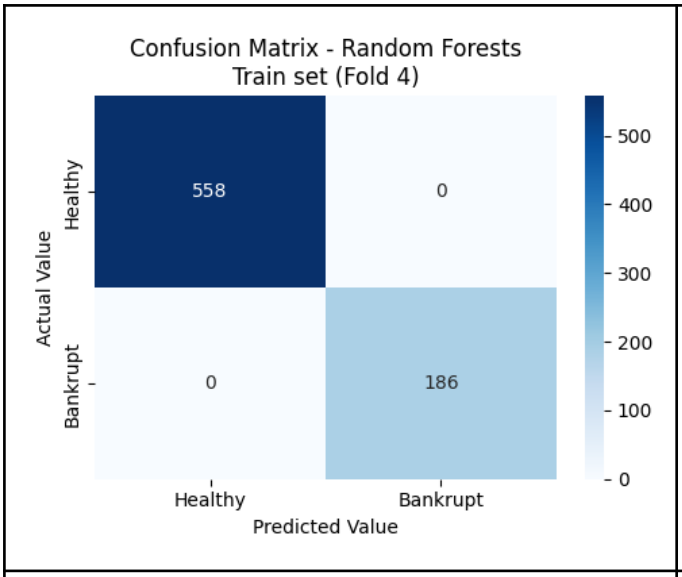

Confusion Matrix - Random Forests
Train set (Fold 3)
Actual Value
Healthy
Bankrupt
558
0
0
186
Healthy
Bankrupt
Predicted Value
500
400
300
200
100
0
Confusion Matrix - Random Forests
Test set (Fold 3)
Actual Value
Healthy
Bankrupt
2438
179
40
22
Healthy
Bankrupt
Predicted Value
2000
1500
1000
500
Confusion Matrix - k-Nearest Neighbors
Train set (Fold 3)
Actual Value
Healthy
Bankrupt
521
37
75
111
Healthy
Bankrupt
Predicted Value
500
400
300
200
100
Confusion Matrix - k-Nearest Neighbors
Test set (Fold 3)
Actual Value
Healthy
Bankrupt
2371
246
45
17
Healthy
Bankrupt
Predicted Value
2000
1500
1000
500
Confusion Matrix - Naïve Bayes
Train set (Fold 3)
Actual Value
Healthy
Bankrupt
480
78
94
92
Healthy
Bankrupt
Predicted Value
450
400
350
300
250
200
150
100
Confusion Matrix - Naïve Bayes
Test set (Fold 3)
Actual Value
Healthy
Bankrupt
2290
327
40
22
Healthy
Bankrupt
Predicted Value
2000
1500
1000
500

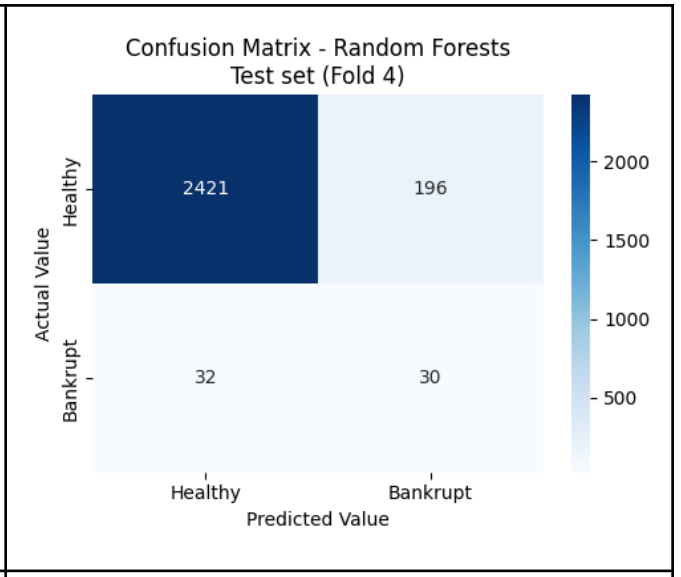


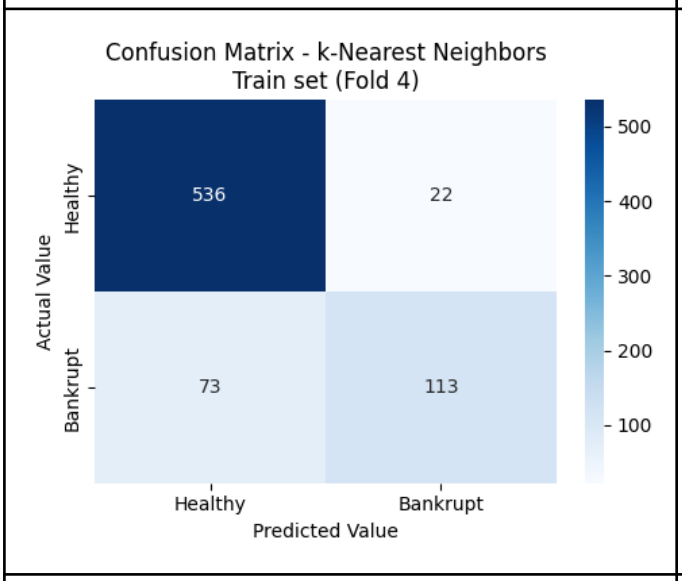
***Σχήμα 4.3.3**: Confusion Matrices – Fold 3*

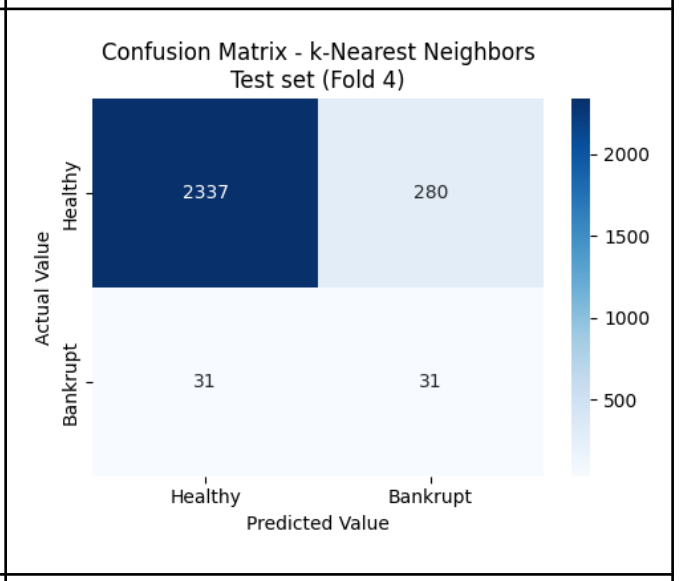
FOLD 4

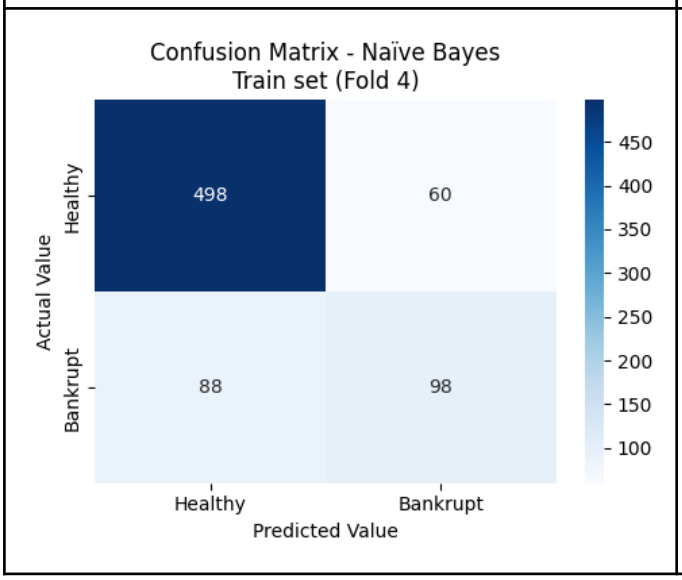

Confusion Matrix - Linear Discriminant Analysis
Train set (Fold 4)
528
30
109
77
Healthy
Bankrupt
Actual Value
Predicted Value
500
400
300
200
100
Confusion Matrix - Linear Discriminant Analysis
Test set (Fold 4)
2422
195
41
21
2000
1500
1000
500
Confusion Matrix - Logistic Regression
Train set (Fold 4)
531
27
114
72
Confusion Matrix - Logistic Regression
Test set (Fold 4)
2445
172
42
20
Confusion Matrix - Decision Trees
Train set (Fold 4)
558
0
186
Confusion Matrix - Decision Trees
Test set (Fold 4)
2185
432
33
29
1750
1250
750
250

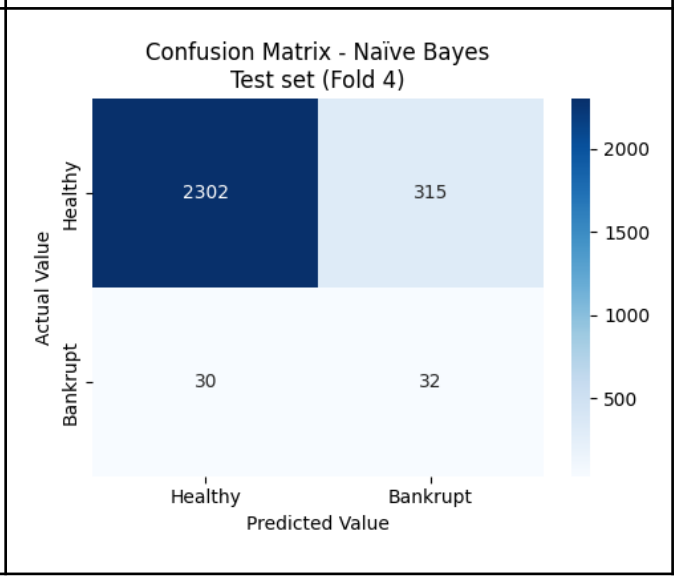

Confusion Matrix - Random Forests
Train set (Fold 4)
Actual Value
Healthy
Bankrupt
558
0
0
186
Healthy
Bankrupt
Predicted Value
500
400
300
200
100
0

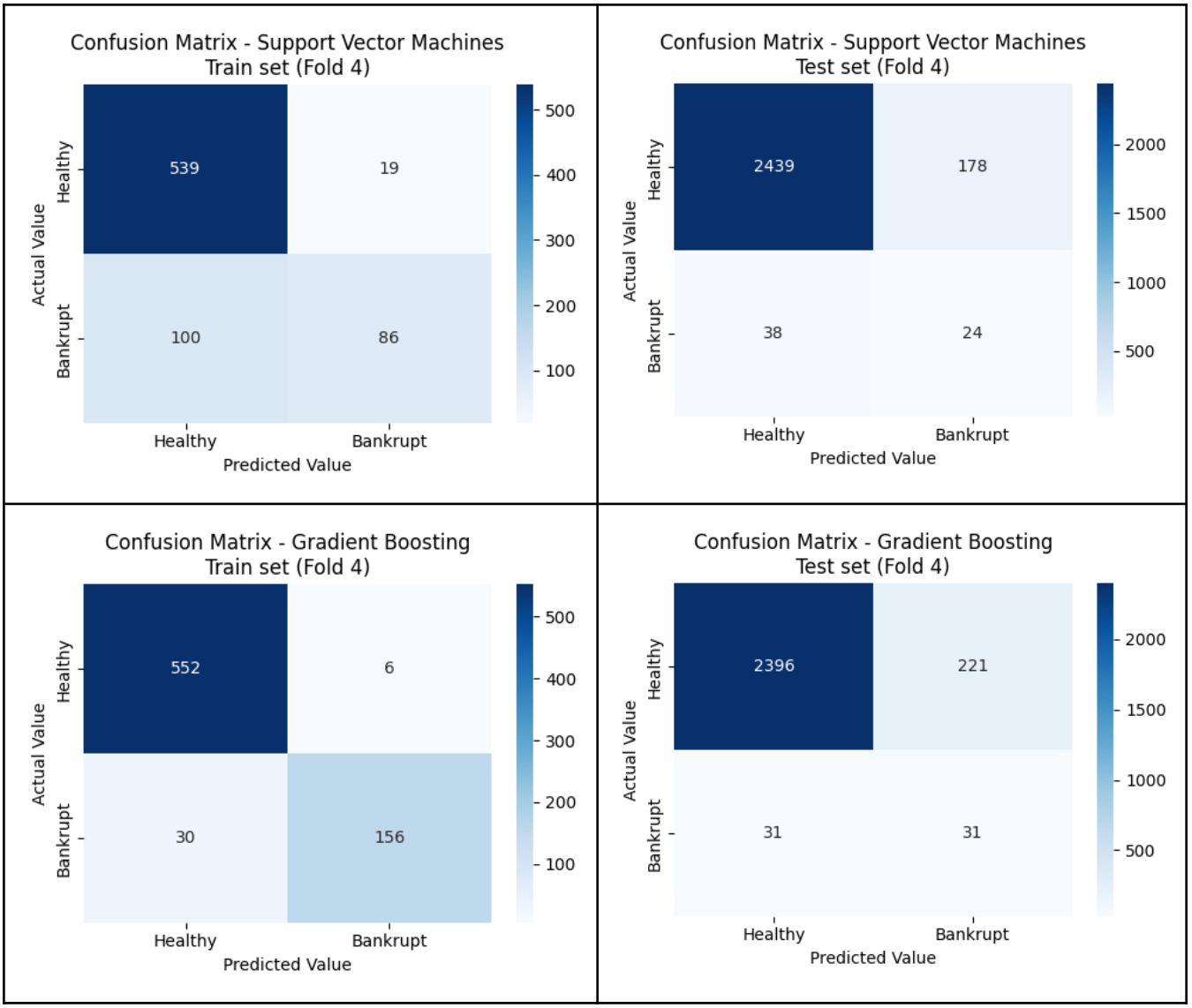

Confusion Matrix - Random Forests
Test set (Fold 4)
Actual Value
Healthy
Bankrupt
2421
196
32
30
Healthy
Bankrupt
Predicted Value
2000
1500
1000
500


Confusion Matrix - k-Nearest Neighbors
Train set (Fold 4)
Actual Value
Healthy
Bankrupt
536
22
73
113
Healthy
Bankrupt
Predicted Value
500
400
300
200
100


Confusion Matrix - k-Nearest Neighbors
Test set (Fold 4)
Actual Value
Healthy
Bankrupt
2337
280
31
31
Healthy
Bankrupt
Predicted Value
2000
1500
1000
500


Confusion Matrix - Naïve Bayes
Train set (Fold 4)
Actual Value
Healthy
Bankrupt
498
60
88
98
Healthy
Bankrupt
Predicted Value
450
400
350
300
250
200
150
100


Confusion Matrix - Naïve Bayes
Test set (Fold 4)
Actual Value
Healthy
Bankrupt
2302
315
30
32
Healthy
Bankrupt
Predicted Value
2000
1500
1000
500



***Σχήμα 4.3.4**: Confusion Matrices – Fold 4*

4.4 Ανάλυση Αποτελεσμάτων με βάση το F1-score

Στον παρακάτω πίνακα παρουσιάζονται οι μέσοι όροι του F1-score για κάθε μοντέλο ταξινόμησης, αναλυτικά για κάθε fold της διαδικασίας διασταύρωσης (cross-validation), καθώς και ο συνολικός μέσος όρος. (Protopapadakis, nd)

AVERAGE από	Fold				
Classifier Name	1	2	3	4	Γενικό σύνολο
Decision Trees	0,12389	0,10018	0,10256	0,11090	0,10938
Gradient Boostir	0,20054	0,21212	0,14533	0,19745	0,18886
k-Nearest Neigh	0,16459	0,17416	0,10462	0,16622	0,15240
Linear Discrimin	0,19936	0,22794	0,14449	0,15108	0,18072
Logistic Regress	0,21583	0,23176	0,11618	0,15748	0,18031
Naïve Bayes	0,15345	0,18818	0,10706	0,15648	0,15129
Random Forests	0,18902	0,27206	0,16730	0,20833	0,20918
Support Vector M	0,22145	0,27184	0,13825	0,18182	0,20334
Γενικό σύνολο	**0,18352**	**0,20978**	**0,12822**	**0,16622**	**0,17194**

***Σχήμα 4.3**: Ανάλυση Αποτελεσμάτων με βάση το F1-score*

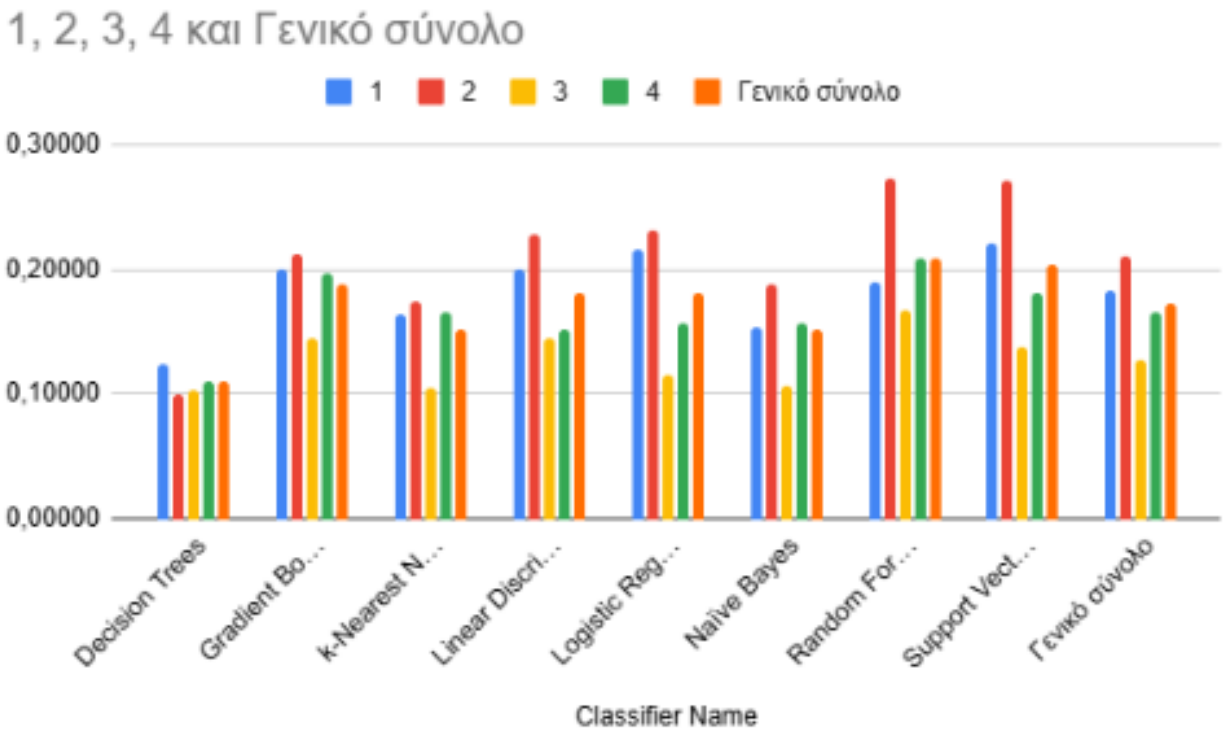
Από τον πίνακα διακρίνεται ότι το μοντέλο Random Forests έχει τον υψηλότερο συνολικό μέσο όρο F1-score (0,209), ακολουθούμενο από τα μοντέλα Support Vector Machines (0,203) και Gradient Boosting (0,189). Τα Decision Trees εμφανίζουν τη χαμηλότερη απόδοση (0,109).Παρατηρείται επίσης ότι το fold 2 εμφανίζει συνολικά καλύτερη απόδοση, ενώ το fold 3 έχει τις χαμηλότερες τιμές F1-score για τα περισσότερα μοντέλα.

5. Συμπεράσματα

5.1 Συμπέρασμα για το καλύτερο μοντέλο ταξινόμησης

Με βάση την παραπάνω ανάλυση, το **Random Forests** φαίνεται να είναι το καλύτερο μοντέλο ταξινόμησης, λαμβάνοντας υπόψη το F1-score, το οποίο είναι ιδιαίτερα χρήσιμο όταν υπάρχει ανισορροπία στις κλάσεις. Επιπλέον, η επιλογή των Balanced Accuracy και Matthews Correlation Coefficient ως επιπλέον μετρικών επιβεβαιώνει την καλή απόδοση του Random Forests, καθώς λαμβάνουν υπόψη και την ισορροπία των κλάσεων και την συνολική ποιότητα της ταξινόμησης. (Protopapadakis, nd)

Για την υποστήριξη της αξιολόγησης των μοντέλων, παρακάτω παρατίθεται γραφική απεικόνιση των μέσων τιμών του F1-score για κάθε αλγόριθμο ταξινόμησης. Όπως προκύπτει από τη σύγκριση, το μοντέλο **Random Forests** παρουσιάζει τη **βέλτιστη απόδοση**, επιτυγχάνοντας την υψηλότερη μέση τιμή F1-score μεταξύ όλων των υπολοίπων μεθόδων που δοκιμάστηκαν.



***Σχήμα 5.1**: Συμπέρασμα για το καλύτερο μοντέλο ταξινόμησης*

5.2 Έλεγχος Κριτηρίων Απόδοσης

Στόχος της παρούσας μελέτης είναι η αξιολόγηση των μοντέλων ταξινόμησης ως προς την ικανότητά τους να εντοπίζουν αποτελεσματικά τις επιχειρήσεις που πρόκειται να

χρεωκοπήσουν, υπό την προϋπόθεση ότι πληρούν συγκεκριμένα ποσοτικά κριτήρια απόδοσης. Συγκεκριμένα, απαιτείται τα μοντέλα να επιτυγχάνουν:

- Recall (Ανάκληση) ≥ 0.60 για τις πτωχευμένες επιχειρήσεις, διασφαλίζοντας την έγκαιρη και αξιόπιστη ανίχνευση των περιπτώσεων υψηλού κινδύνου, και
- Specificity (Ειδικότητα) ≥ 0.70 για τις υγιείς επιχειρήσεις, περιορίζοντας τα ψευδώς θετικά αποτελέσματα και ενισχύοντας την αξιοπιστία της πρόβλεψης.

Από την ανάλυση των αποτελεσμάτων προκύπτει ότι:

- Το μοντέλο Naïve Bayes υπερβαίνει το όριο του Recall, φθάνοντας σε τιμή 0.69355 στο δεύτερο fold της διασταύρωσης, υποδεικνύοντας ικανοποιητική ικανότητα εντοπισμού των πτωχευμένων επιχειρήσεων.
- Ωστόσο, το ίδιο μοντέλο δεν πληροί το κριτήριο Specificity ≥ 0.70, παρουσιάζοντας αυξημένο ποσοστό ψευδώς θετικών.
- Τα μοντέλα Random Forest και Gradient Boosting επιδεικνύουν υψηλότερες τιμές Specificity, όμως αποτυγχάνουν να διατηρήσουν Recall πάνω από το επιθυμητό όριο.
- Κανένα από τα υπόλοιπα μοντέλα που εξετάστηκαν δεν ικανοποιεί ταυτόχρονα τα δύο βασικά κριτήρια απόδοσης.

Συμπερασματικά, δεν αναγνωρίζεται κάποιο μοντέλο ταξινόμησης που να πληροί αμφότερα τα κριτήρια Recall ≥ 0.60 και Specificity ≥ 0.70 στο σύνολο της ανάλυσης, γεγονός που υποδηλώνει την ανάγκη για περαιτέρω βελτιστοποίηση μοντέλων ή την ανάπτυξη υβριδικών προσεγγίσεων προκειμένου να επιτευχθεί η επιθυμητή ισορροπία ανάμεσα στην ανίχνευση των πτωχευμένων και την ελαχιστοποίηση των ψευδώς θετικών αποτελεσμάτων.

6. Βιβλιογραφία

Protopapadakis, E. (n.d.). *A tutorial on classification. A breast cancer detection scenario* [Tutorial document]. Retrieved from eClass course website for Μέθοδοι και εγαλεία τεχνητής νοημοσύνης

Wikipedia contributors. (n.d.). Confusion matrix. Wikipedia. Retrieved May 20, 2025, from https://en.wikipedia.org/wiki/Confusion_matrix