Βαθιά Μάθηση
Εργασία Μέρος 1

Χρήστος Πανουργιάς aivc22014

Περιεχόμενα

1 Εισαγωγή

Λόγω των δυνατοτήτων της σημερινής τεχνολογίας, οι υπεύθυνοι επιχειρήσεων έχουν την επιλογή να αξιοποιήσουν τα δεδομένα που συλλέγουν έτσι ώστε να βελτιστοποιήσουν τις λειτουργίες και τις αποφάσεις των επιχειρήσεων τους. Ένας πολύ σημαντικός παράγοντας για την επιτυχία μίας επιχείρησης είναι η οικονομική της υγεία, συνεπώς είναι κρίσιμης σημασίας ένας υπεύθυνος επιχείρησης να έχει υψηλής ποιότητας πληροφόρηση όσον αφορά την οικονομική πορεία της επιχείρησης του. Σκοπός της εργασίας αυτής είναι να αναλυθούν χρηματοπιστωτικοί δείκτες αλλά και άλλες πληροφορίες από μία σειρά δεδομένων ελληνικών εταιρειών, έτσι ώστε να πραγματοποιηθεί επιτυχής πρόβλεψη της πτώχευσης ή μη πτώχευσης μίας εταιρείας. Συνεπώς θα δημιουργηθεί ένα μοντέλο το οποίο λαμβάνοντας ως είσοδο τα κατάλληλα δεδομένα θα έχει την δυνατότητα να κατευθύνει τους υπεύθυνους επιχειρήσεων στην αποτροπή χρεωκοπίας της επιχείρησης τους.

Θα πραγματοποιηθεί προεπεξεργασία δεδομένων και θα συγκριθούν αρχιτεκτονικές μοντέλων όπως λογιστική παλινδρόμηση, δέντρα αποφάσεων, κ-πλησιέστεροι γείτονες, Support Vector Machines, Naive Bayes, Linear Discriminant Analysis και Multi-Layer Perceptron και τα αποτελέσματα θα αξιολογηθούν κυρίως βάση των μετρικών specificity, δηλαδή για πόσες εταιρείες έγινε σωστά η πρόβλεψη χρεωκοπίας και recall, δηλαδή για πόσες εταιρείες έγινε σωστά η πρόβλεψη μη χρεωκοπίας, θα υπολογισθούν και άλλες μετρικές αξιολόγησης ταξινομητών όπως F1-Score, Accuracy και Precision έτσι ώστε να δημιουργηθεί μια ολοκληρωμένη αξιολόγηση για τον κάθε ταξινομητή.

2 Μέθοδοι που Εφαρμόστηκαν

2.1 Χωρισμός Δεδομένων

Αρχικά από τα δεδομένα διαχωρίστηκαν τα χαρακτηριστικά, που θα αποτελούν τις εισόδους του μοντέλου, από τον στόχο που αποτελεί τις εξόδους το μοντέλου. Έπειτα, για να υλοποιηθεί αντικειμενική αξιολόγηση του μοντέλου, θα πρέπει να εξεταστεί σε δεδομένα τα οποία είναι ανεξάρτητα από τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση του. Συνεπώς πραγματοποιήθηκε χωρισμός δεδομένων σε δεδομένα εκπαίδευσης τα οποία αποτελούν το 80% των αρχικών δεδομένων και δεδομένα ελέγχου τα οποία αποτελουν το υπόλοιπο 20% των αρχικών δεδομένων.

2.2 Κανονικοποίηση

Τα χαρακτηριστικά έχουν διαφορετικό εύρος τιμών το κάθε ένα, και αυτό μπορεί να προκαλέσει προβλήματα στους υπολογισμούς, έαν κάποιες τιμές είναι πολύ μεγάλες ή πολύ μικρές, αλλά και στον χρόνο υπολογισμού των πράξεων. Με βάση αυτό το σκεπτικό υλοποιήθηκε κανονικοποίηση με την χρήση του MinMax Scaler έτσι ώστε όλα τα χαρακτηριστικά να έχουν εύρος τιμών $[0, 1]$. Αξίζει να σημειωθεί ότι η κανονικοποίηση στα χαρακτηριστικά του συνόλου ελέγχου πραγματοποιήθηκε βάση της μέγιστης και ελάχιστης τιμής του συνόλου εκπαίδευσης έτσι ώστε το σύνολο ελέγχου να μην επηρεαστεί από τα δεδομένα εκπαίδευσης.

2.3 Εκπαίδευση και Αξιολόγηση

Το κάθε μοντέλο εκπαιδεύεται με την χρήση της συνάρτησης .fit() της βιβλιοθήκης Scikit-Learn της Python. Εφόσον το εκάστοτε μοντέλο έχει εκπαιδευτεί με την χρήση του συνόλου εκπαίδευσης, χρησιμοποιείται η συνάρτηση .predict() για να πραγματοποιηθεί πρόβλεψη των τιμών του συνόλου εκπαίδευσης, έτσι ώστε να αξιολογηθεί το μοντέλο στις προβλέψεις δεδομένων επάνω στα οποία έχει εκπαιδευτεί, και προβλέψη των τιμών του συνόλου ελέγχου, έτσι ώστε να αξιολογηθεί στις προβλέψεις των τιμών που δεν του έχουν εμφανιστεί.

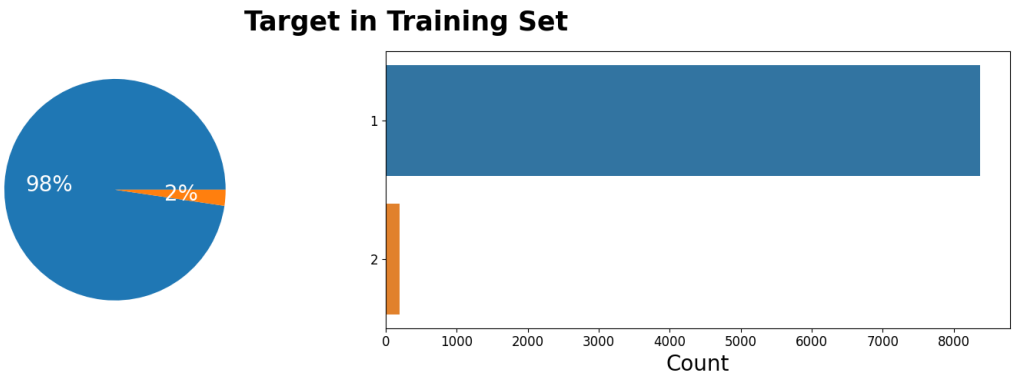
3 Πειραματικά Αποτελέσματα

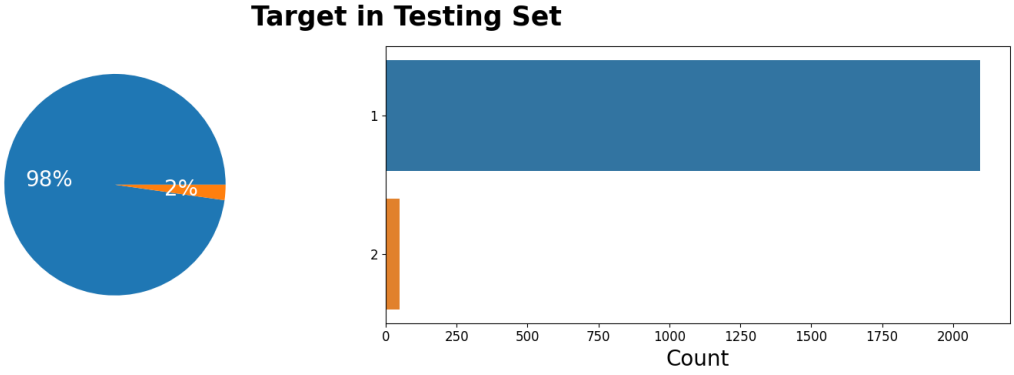
3.1 Εκαπαίδευση χωρίς ισορρόπηση των κλάσεων

Η ισορροπία των κλάσεων στα δεδομένα, είναι ένας σημαντικός παράγοντας εκπαίδευσης του μοντέλου διότι καθορίζει την αντίληψη που θα έχει το μοντέλο για την σημαντικότητα της κάθε κλάσης. Ιδανικά είναι επιθυμητό το μοντέλο να έχει ίσο αριθμό δεδομένων από κάθε κλάση, έτσι ώστε να μην κάνει προκατελειμμένες προβλέψεις ως προς την πλειοψηφούσα κλάση. Τα μειονεκτήματα ισορρόπησης των κλάσεων είναι ότι ενδέχεται είτε να χρειαστεί να μην χρησιμοποιηθούν δεδομένα από την πλειοψηφούσα κλάση, είτε να δημιουργηθούν τεχνητά δεδομένα τα οποία μπορεί να εισάγουν θόρυβο στο μοντέλο.

Για να γίνει αντιληπτή η ανάγκη ισορρόπησης των κλάσεων στα δεδομένα εκπαίδευσης, παρακάτω παρουσιάζεται η εκπαίδευση των μοντέλων στα μη ισορροπημένα δεδομένα εκπαίδευσης.

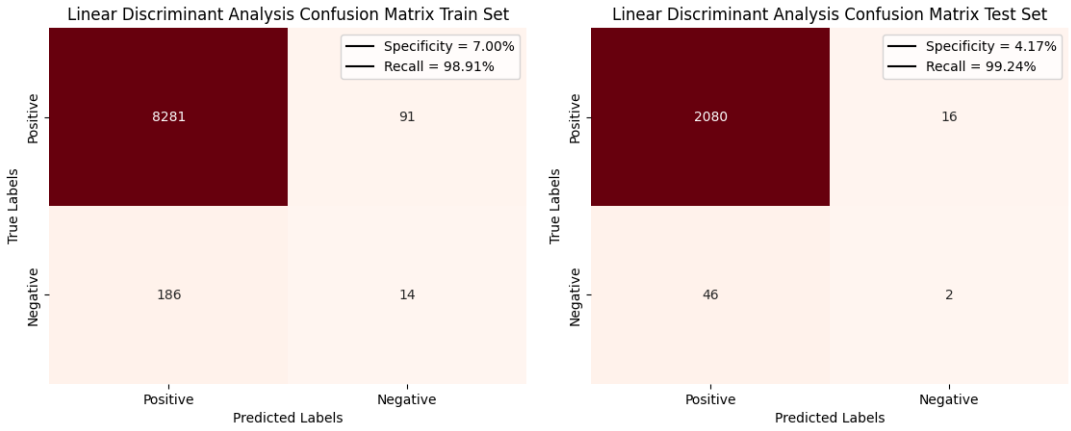
3.1.1 Περιεκτικότητα Κλάσεων



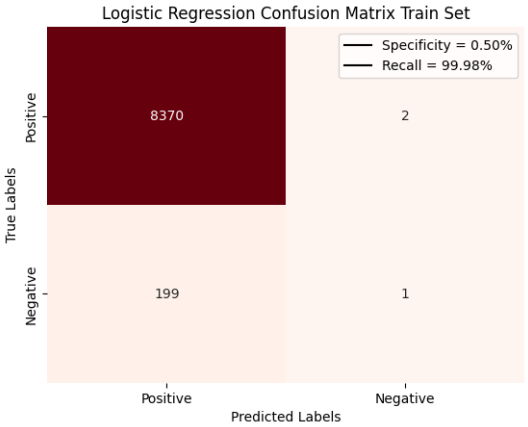


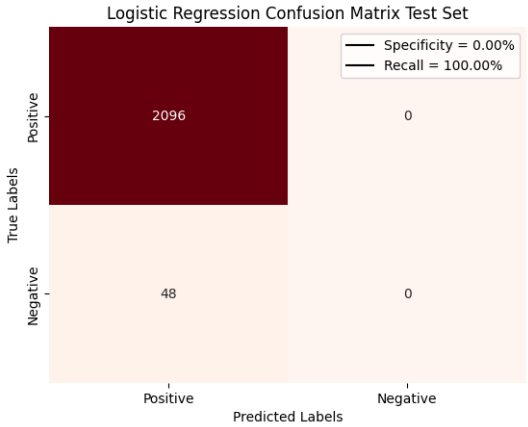
Από τα παραπάνω γραφήματα, παρατηρείται ότι η περιεκτικότητα της κλάσης των χρεωκοπημένων εταιρειών στα δεδομένα είναι 2%. Αν ο σκοπός είναι να προβλεφθούν οι χρεωκοπημένες εταιρείες τότε αυτό θα δυσκολέψει την μάθηση του μοντέλου. Αυτό επιβεβαιώνεται και από τα παρακάτω αποτελέσματα των πινάκων σύγχυσης.

3.1.2 Linear Discriminant Analysis

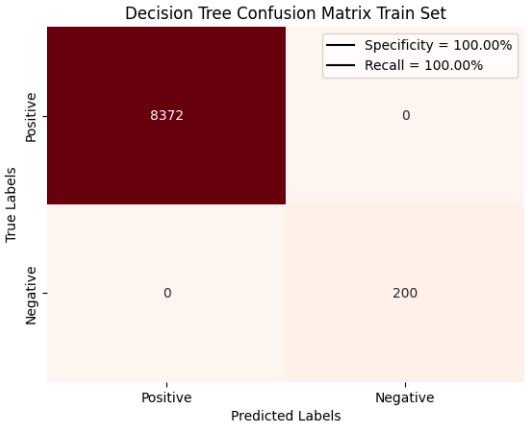


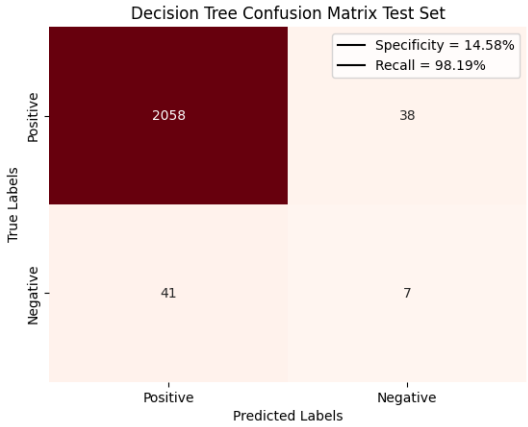
3.1.3 Logistic Regression



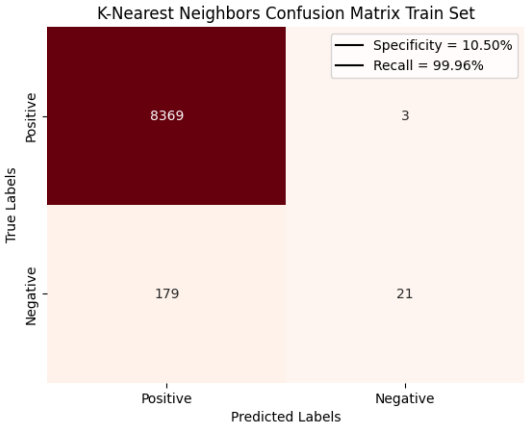


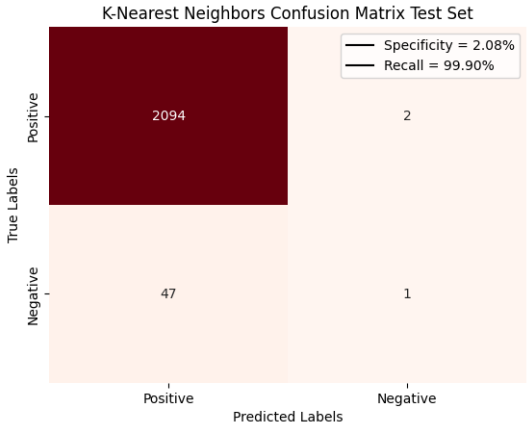
3.1.4 Decision Trees



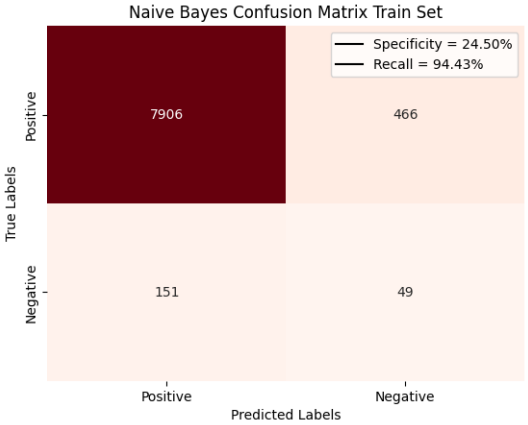


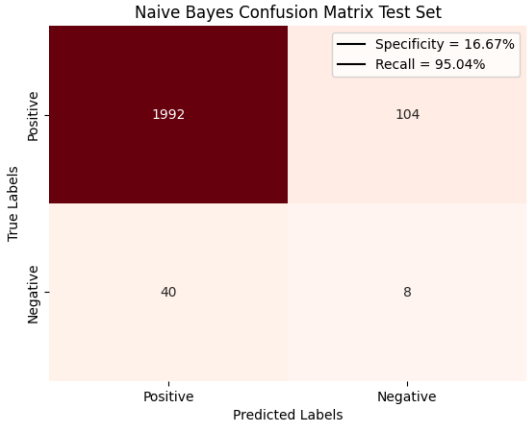
3.1.5 K-Nearest Neighbors



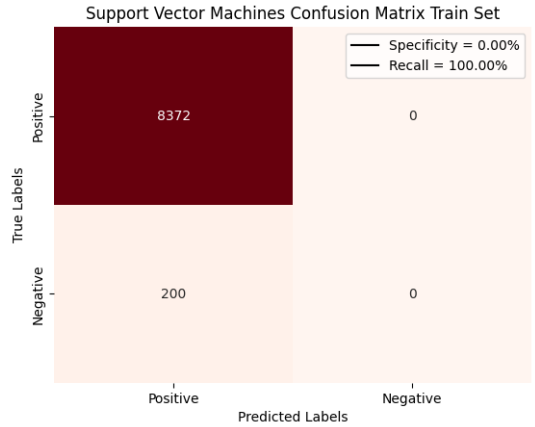


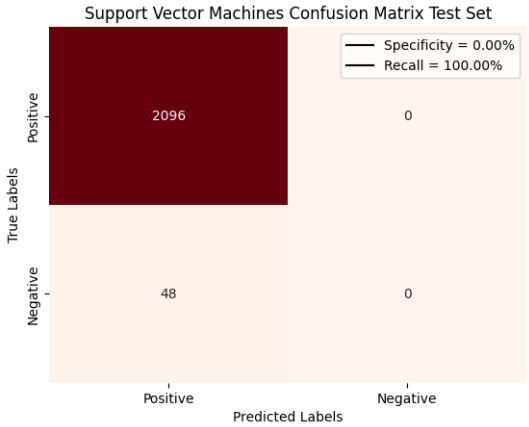
3.1.6 Naive Bayes



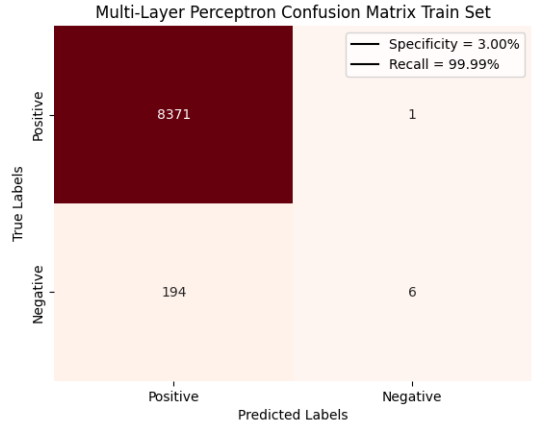


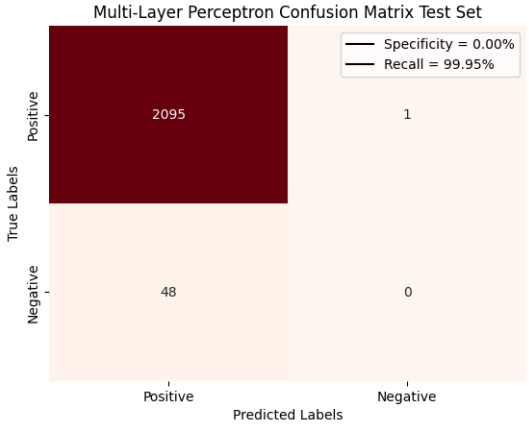
3.1.7 Support Vector Machines





3.1.8 Multi-Layer Perceptron





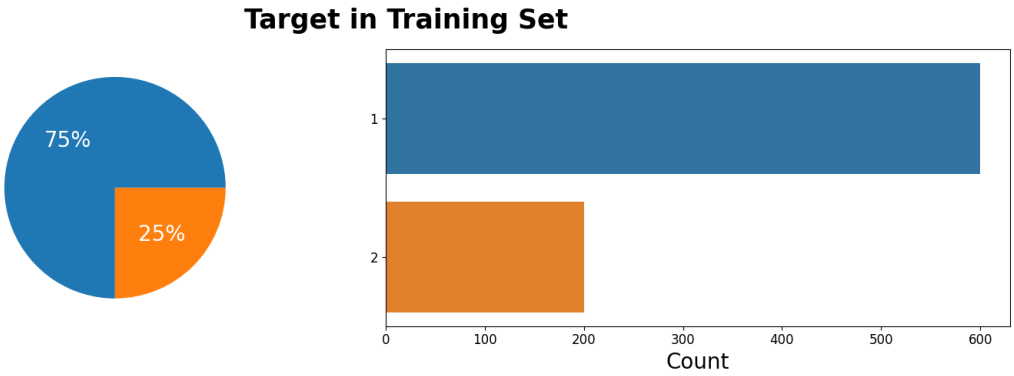
3.1.9 Συνολικά Αποτελέσματα

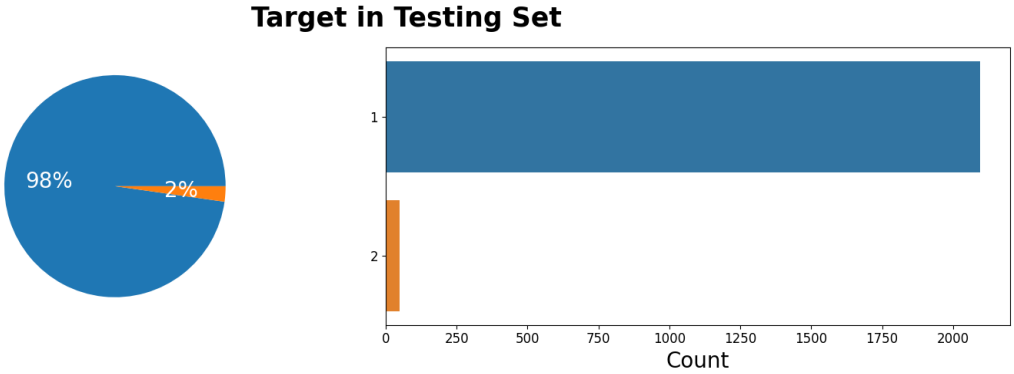
Classifier Name	Set type	Number of training samples	Number of non-healthy companies in training sample	TP	TN	FP	FN	Precision	Recall	Specificity	F1-Score	Accuracy
Linear Discriminant Analysis	Train	8572	200	8281	14	186	91	0.978	0.989	0.07	0.984	0.968
Linear Discriminant Analysis	Test	8572	200	2080	2	46	16	0.978	0.992	0.042	0.985	0.971
Logistic Regression	Train	8572	200	8370	1	199	2	0.977	1	0.005	0.988	0.977
Logistic Regression	Test	8572	200	2096	0	48	0	0.978	1	0	0.989	0.978
Decision Tree	Train	8572	200	8372	200	0	0	1	1	1	1	1
Decision Tree	Test	8572	200	2059	7	41	37	0.98	0.982	0.146	0.981	0.964
K-Nearest Neighbors	Train	8572	200	8369	21	179	3	0.979	1	0.105	0.989	0.979
K-Nearest Neighbors	Test	8572	200	2094	1	47	2	0.978	0.999	0.021	0.988	0.977
Naive Bayes	Train	8572	200	7906	49	151	466	0.981	0.944	0.245	0.962	0.928
Naive Bayes	Test	8572	200	1992	8	40	104	0.98	0.95	0.167	0.965	0.933
Support Vector Machines	Train	8572	200	8372	0	200	0	0.977	1	0	0.988	0.977
Support Vector Machines	Test	8572	200	2096	0	48	0	0.978	1	0	0.989	0.978
Multi-Layer Perceptron	Train	8572	200	8371	3	197	1	0.977	1	0.015	0.988	0.977
Multi-Layer Perceptron	Test	8572	200	2096	0	48	0	0.978	1	0	0.989	0.978

3.2 Εκπαίδευση με ισορρόπηση κλάσεων 3 προς 1

Από τα αποτελέσματα της εκπαίδευσης στα γνήσια δεδομένα, φαίνεται πως κανένα μοντέλο δεν έιχε $specificity > 62\%$ και $recall > 70\%$ όπως είναι και το ζητούμενο. Παρακάτω παρουσιάζονται τα αποτελέσματα από την εκπαίδευση των μοντέλων σε ισορροπημένα δεδομένα. Συγκεκριμένα αφαιρέθηκαν δεδομένα της κλάσης εταιρειών που ανήκουν στην κλάση των μη-χρεωκοπημένων έτσι ώστε για κάθε τρείς μη-χρεωκοπημένες εταιρείες να υπάρχει μία χρεωκοπημένη εταιρεία.

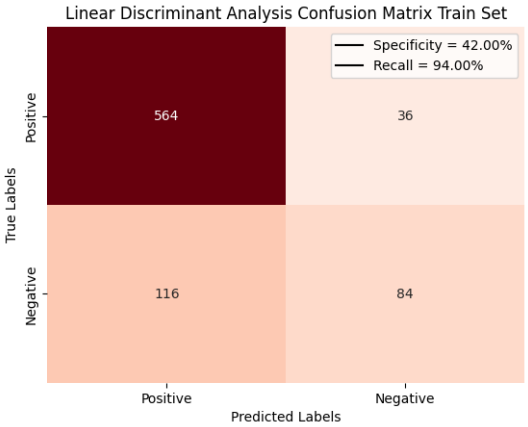
3.2.1 Περιεκτικότητα Κλάσεων

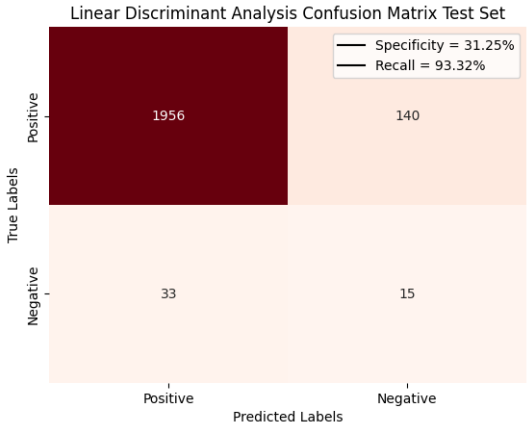




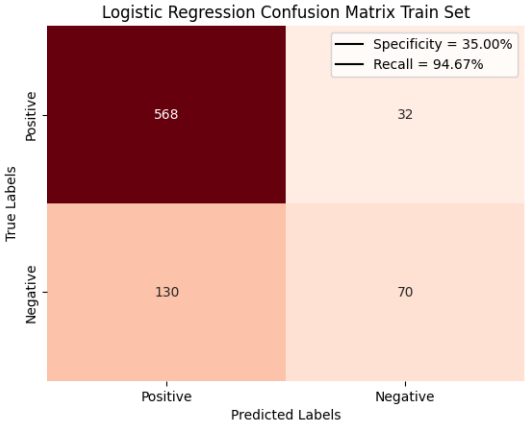
Όπως φαίνεται και παραπάνω το σύνολο δεδομένων εκπαίδευσης είναι πλέον ισορροπημένο 3 προς 1 και το σύνολο ελέγχου έχει αφεθεί ως έχει. Παρακάτω παρουσιάζονται τα αποτελέσματα της εκπαίδευσης των μοντέλων σε αυτό το σύνολο.

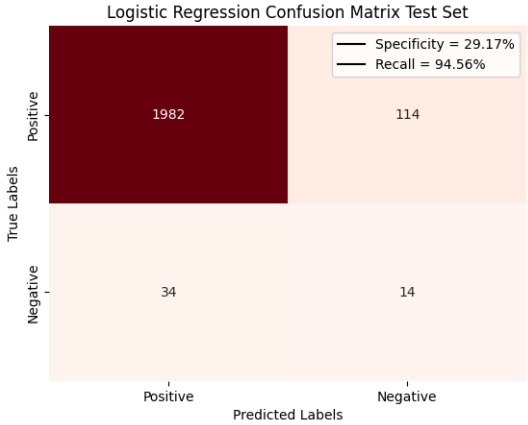
3.2.2 Linear Discriminant Analysis



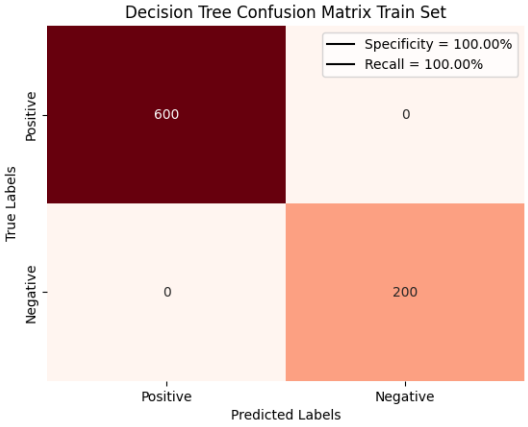


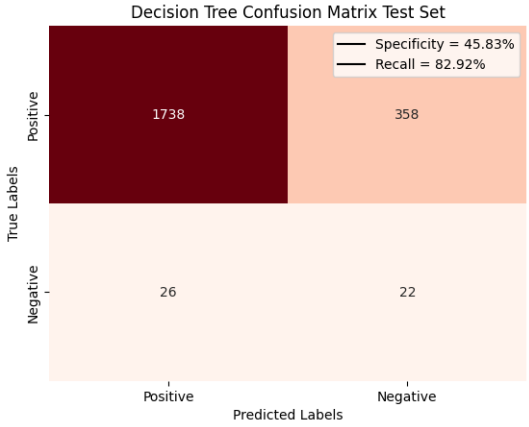
3.2.3 Logistic Regression



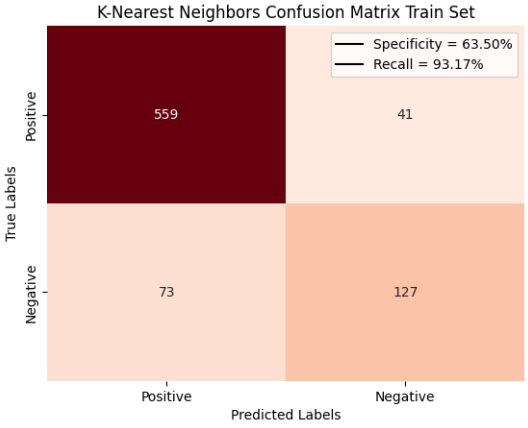


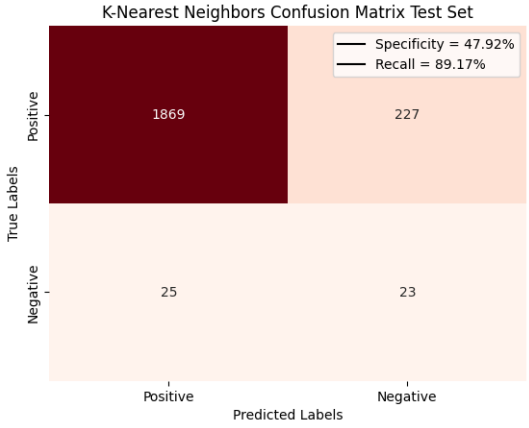
3.2.4 Decision Trees



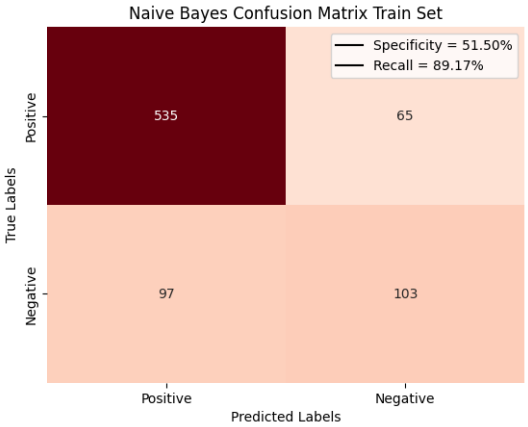


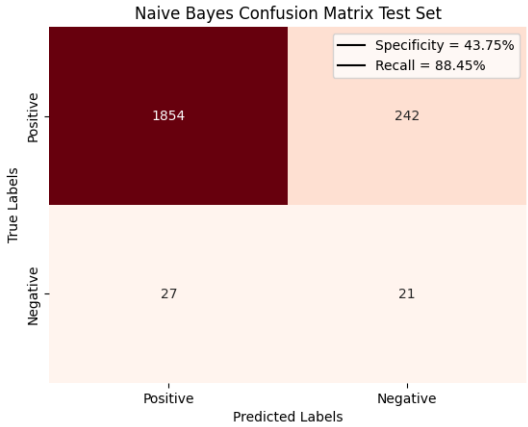
3.2.5 K-Nearest Neighbors



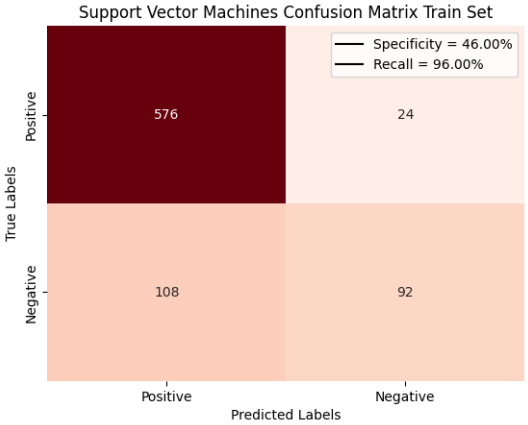


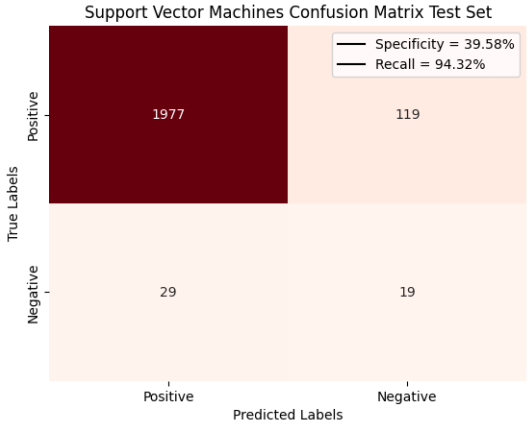
3.2.6 Naive Bayes



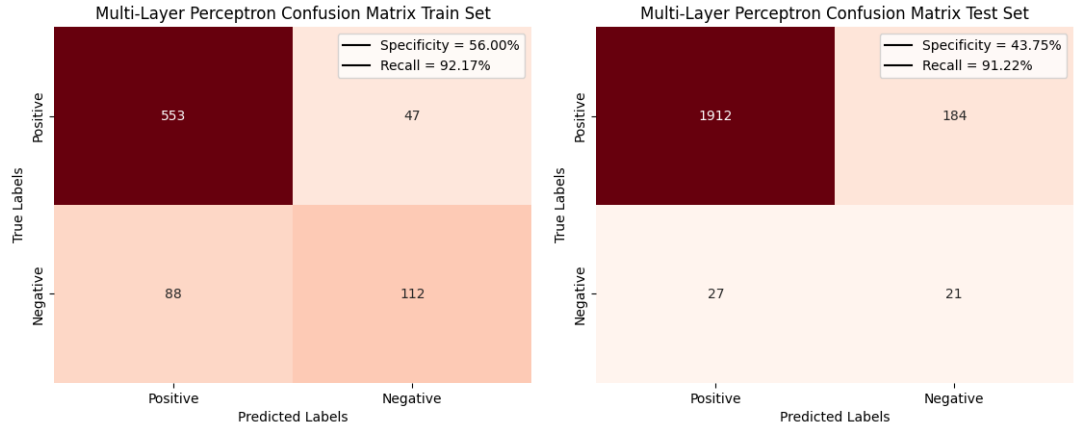


3.2.7 Support Vector Machines





3.2.8 Multi-Layer Perceptron



3.2.9 Συνολικά Αποτελέσματα

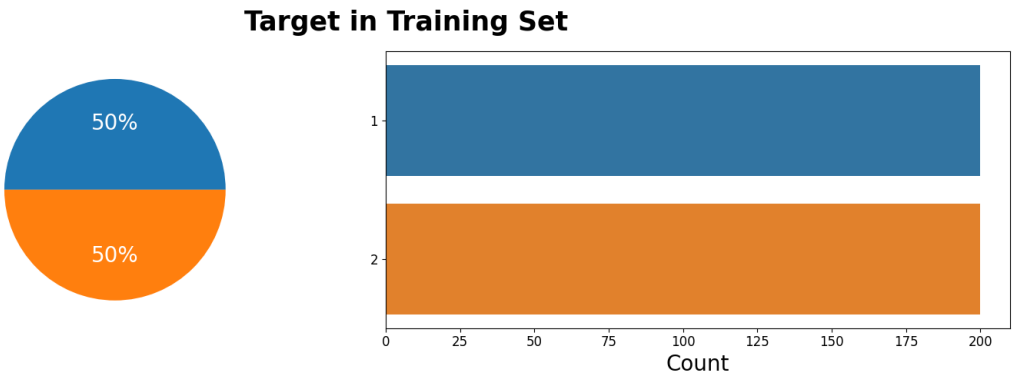
Classifier Name	Set type	Number of training samples	Number of non-healthy companies in training sample	TP	TN	FP	FN	Precision	Recall	Specificity	F1-Score	Accuracy
Linear Discriminant Analysis	Train	800	200	564	84	116	36	0.829	0.94	0.42	0.881	0.81
Linear Discriminant Analysis	Test	800	200	1956	15	33	140	0.983	0.933	0.312	0.958	0.919
Logistic Regression	Train	800	200	568	70	130	32	0.814	0.947	0.35	0.875	0.798
Logistic Regression	Test	800	200	1982	14	34	114	0.983	0.946	0.292	0.964	0.931
Decision Tree	Train	800	200	600	200	0	0	1	1	1	1	1
Decision Tree	Test	800	200	1742	23	25	354	0.986	0.831	0.479	0.902	0.823
K-Nearest Neighbors	Train	800	200	559	127	73	41	0.884	0.932	0.635	0.907	0.858
K-Nearest Neighbors	Test	800	200	1869	23	25	227	0.987	0.892	0.479	0.937	0.882
Naive Bayes	Train	800	200	535	103	97	65	0.847	0.892	0.515	0.869	0.798
Naive Bayes	Test	800	200	1854	21	27	242	0.986	0.885	0.438	0.932	0.875
Support Vector Machines	Train	800	200	576	92	108	24	0.842	0.96	0.46	0.897	0.835
Support Vector Machines	Test	800	200	1977	19	29	119	0.986	0.943	0.396	0.964	0.931
Multi-Layer Perceptron	Train	800	200	557	108	92	43	0.858	0.928	0.54	0.892	0.831
Multi-Layer Perceptron	Test	800	200	1923	21	27	173	0.986	0.917	0.438	0.951	0.907

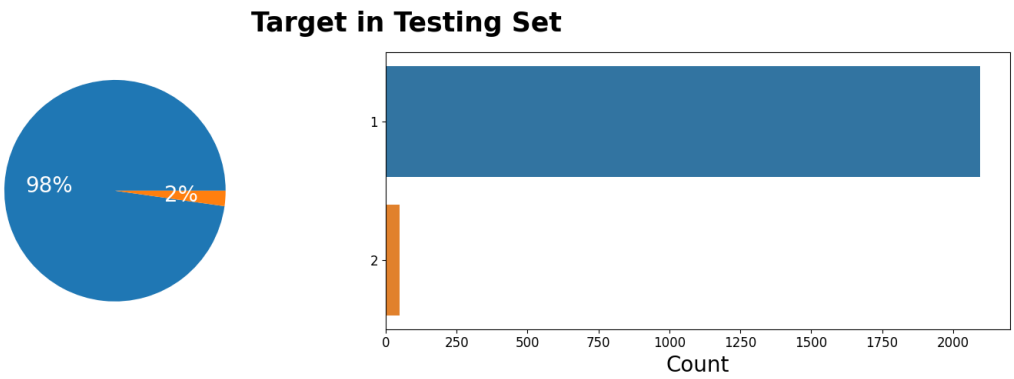
Παρατηρείται ότι κανένα από τα μοντέλα δεν ικανοποιεί τα ποσοστά επιτυχίας που απαιτούν οι περιορισμοί. Ωστόσο αξίζει να σημειωθεί ότι τα ποσοστά σωστής πρόβλεψης αλλά και οι υπόλοιπες μετρικές αξιολόγησης αυξήθηκαν για κάθε μοντέλο.

3.3 Εκπαίδευση με ισορρόπηση κλάσεων 1 προς 1

Η ισορρόπηση 3-1 βελτίωσε τα αποτελέσματα αλλά όχι αρκετά έτσι ώστε να μην παραμείνει προκατειλημμένο το μοντέλο ως προς την πλειοψηφούσα κλάση. Συνεπώς οι κλάσεις ισορροπήθηκαν 1-1 έτσι ώστε το μοντέλο να θεωρεί ίσης σημαντικότητας και τις δύο κλάσεις.

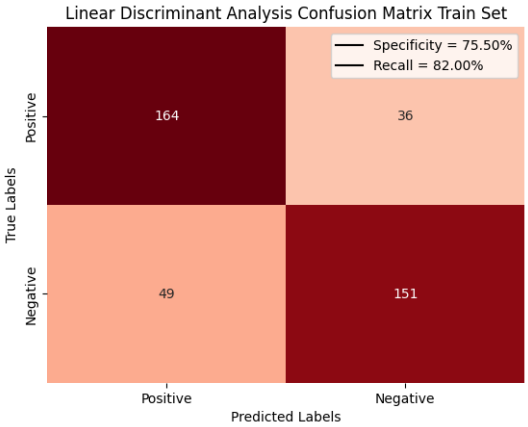
3.3.1 Περιεκτικότητα Κλάσεων

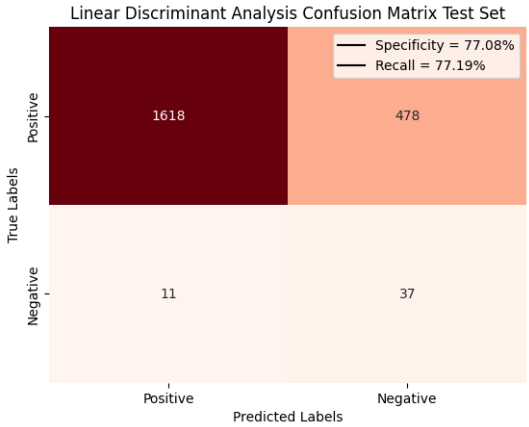




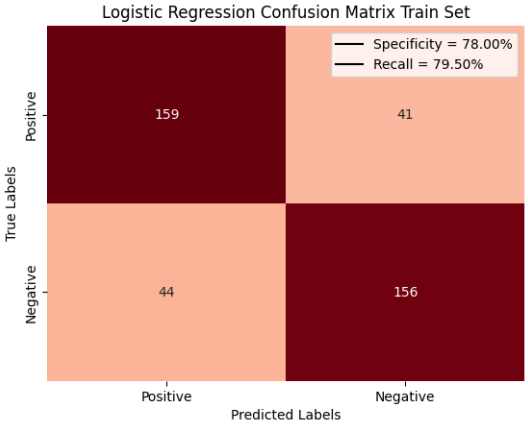
Παραπάνω απεικονίζεται η 1 προς 1 ισορρόπηση των κλάσεων στο σύνολο εκπαίδευσης, και παρακάτω παρουσιάζονται οι πίνακες σύγχυσης και τα αποτελέσματα της εκπαίδευσης των μοντέλων χρησιμοποιώντας το σύνολο αυτό.

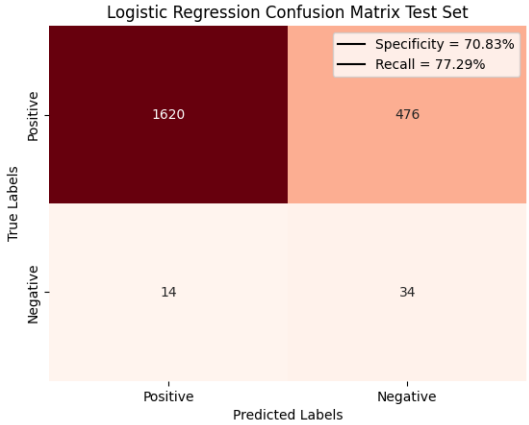
3.3.2 Linear Discriminant Analysis



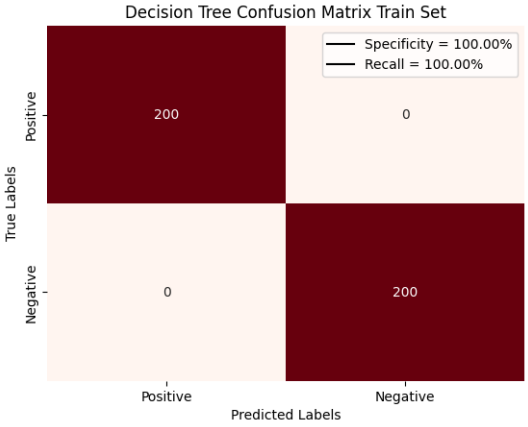


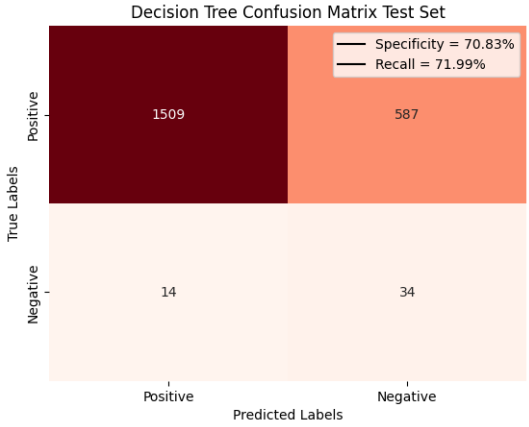
3.3.3 Logistic Regression



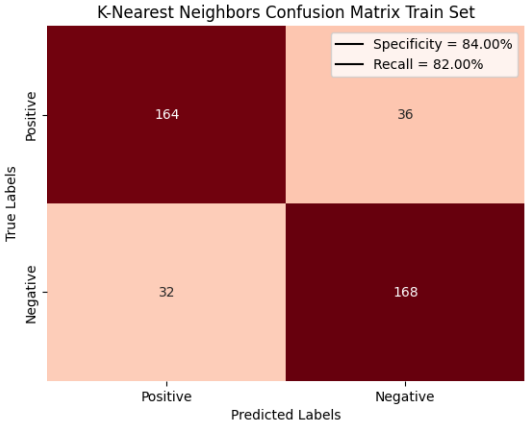


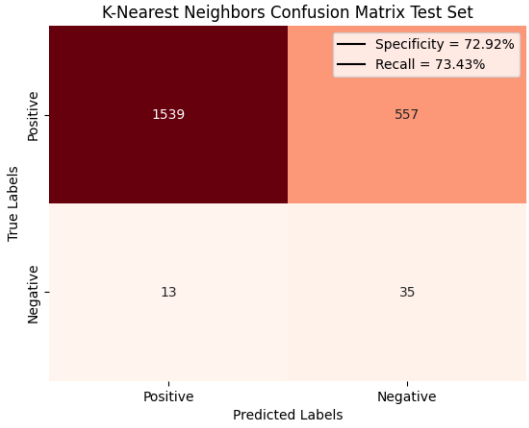
3.3.4 Decision Trees



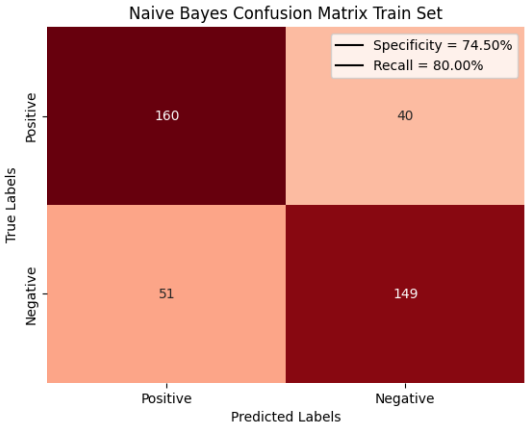


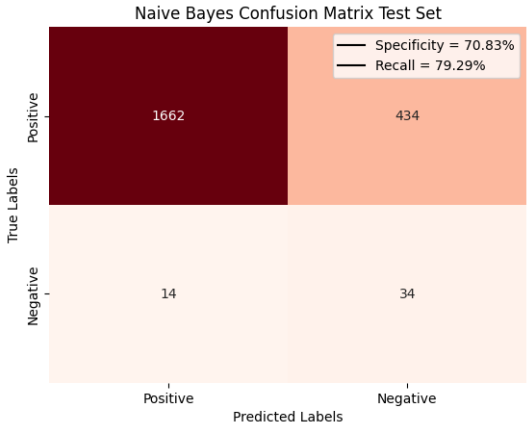
3.3.5 K-Nearest Neighbors



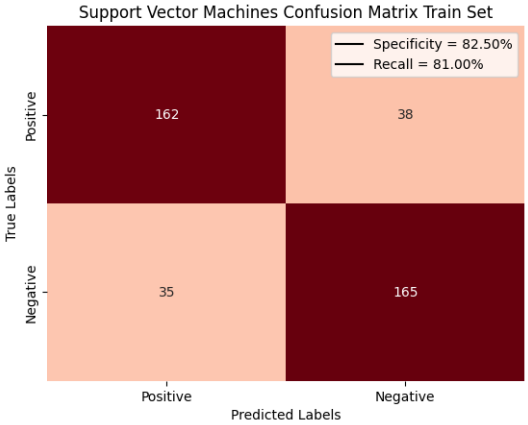


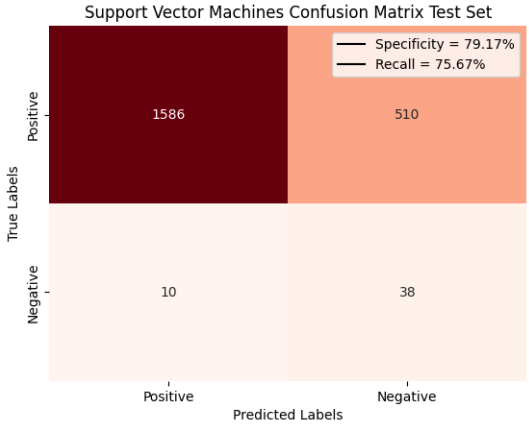
3.3.6 Naive Bayes



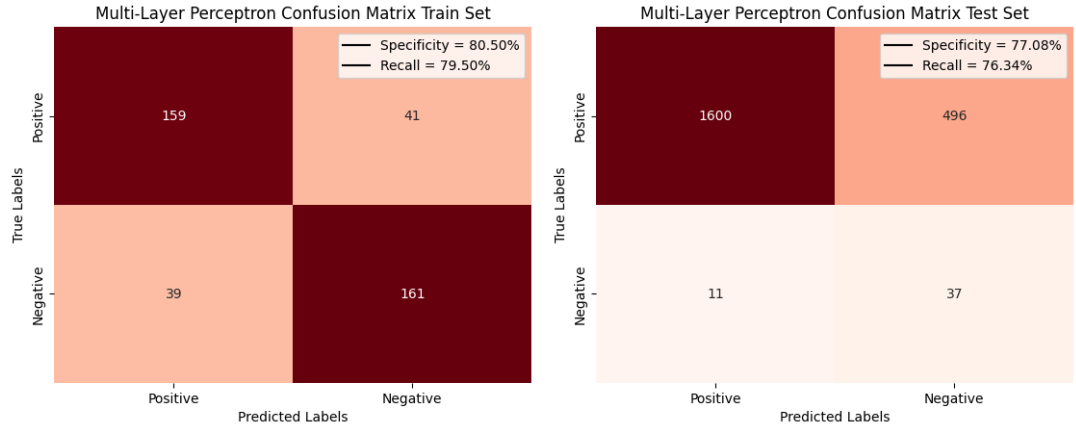


3.3.7 Support Vector Machines





3.3.8 Multi-Layer Perceptron



3.3.9 Συνολικά Αποτελέσματα

Classifier Name	Set type	Number of training samples	Number of non-healthy companies in training sample	TP	TN	FP	FN	Precision	Recall	Specificity	F1-Score	Accuracy
Linear Discriminant Analysis	Train	400	200	164	151	49	36	0.77	0.82	0.755	0.794	0.788
Linear Discriminant Analysis	Test	400	200	1618	37	11	478	0.993	0.772	0.771	0.869	0.772
Logistic Regression	Train	400	200	159	156	44	41	0.783	0.795	0.78	0.789	0.788
Logistic Regression	Test	400	200	1620	34	14	476	0.991	0.773	0.708	0.869	0.771
Decision Tree	Train	400	200	200	200	0	0	1	1	1	1	1
Decision Tree	Test	400	200	1469	32	16	627	0.989	0.701	0.667	0.82	0.7
K-Nearest Neighbors	Train	400	200	164	168	32	36	0.837	0.82	0.84	0.828	0.83
K-Nearest Neighbors	Test	400	200	1539	35	13	557	0.992	0.734	0.729	0.844	0.734
Naive Bayes	Train	400	200	160	149	51	40	0.758	0.8	0.745	0.779	0.772
Naive Bayes	Test	400	200	1662	34	14	434	0.992	0.793	0.708	0.881	0.791
Support Vector Machines	Train	400	200	162	165	35	38	0.822	0.81	0.825	0.816	0.818
Support Vector Machines	Test	400	200	1586	38	10	510	0.994	0.757	0.792	0.859	0.757
Multi-Layer Perceptron	Train	400	200	160	160	40	40	0.8	0.8	0.8	0.8	0.8
Multi-Layer Perceptron	Test	400	200	1598	36	12	498	0.993	0.762	0.75	0.862	0.762

Παρατηρείται ότι, με αυτήν την ισορρόπηση των δεδομένων εκπαίδευσης, όλα τα μοντέλα ικανοποιούν τα ποσοστά επιτυχίας των περιορισμών.

4 Συμπεράσματα

Με βάση τα παραπάνω αποτελέσματα, παρατηρείται ότι είναι απαραίτητο, το σύνολο εκπαίδευσης των εταιρειών, να ισορροπηθεί με αναλογία κλάσεων 1 προς 1 έτσι ώστε οι κλάσεις να θεωρηθούν ίσης σημαντικότητας από το εκάστοτε μοντέλο. Δεδομένου του ότι είναι πιό σημαντικό να πραγματοποιηθεί επιτυχής πρόβλεψη μίας εταιρείας που εμπρόκειτο να χρεωκοπήσει, αναζητείται το μοντέλο που θα εξάγει το υψηλότερο specificity. Το μοντέλο που αποδίδει καλύτερα σε αυτήν την μετρική και έχει εκπαιδευτεί στο 1 προς 1 σύνολο εκπαίδευσης είναι το Support Vector Machines με ποσοστό επιτυχίας πρόβλεψης εταιρειών που θα πτωχεύσουν 79.17% και με ποσοστό επιτυχίας πρόβλεψης εταιρειών που δεν θα πτωχεύσουν 75.67%.

Για περαιτέρω βελτίωση θα μπορούσαν να συλλεχθούν παραπάνω δεδομένα από χρεωκοπημένες εταιρείες έτσι ώστε να μπορεί να γίνει εκπαίδευση σε μεγαλύτερο σύνολο δεδομένων το οποίο όμως να είναι ταυτόχρονα και ισορροπημένο.