Πανεπιστήμιο Μακεδονίας
Εφαρμοσμένη Πληροφορική

Μηχανική Μάθηση
Classification Problem

Γιαγκούδη Δήμητρα

27/07/2025

Περιεχόμενα

Κατάλογος Πινάκων και Εικόνων

Εισαγωγή

Η ικανότητα πρόβλεψης οικονομικών καταρρεύσεων αποτελεί κρίσιμο εργαλείο για την έγκαιρη λήψη αποφάσεων στους τομείς της οικονομίας και των επενδύσεων. Στο πλαίσιο αυτό, η παρούσα εργασία εξετάζει την πρόβλεψη πιθανής πτώχευσης επιχειρήσεων με τη χρήση τεχνικών μηχανικής μάθησης. Η έγκαιρη αναγνώριση εταιρειών με υψηλό ρίσκο πτώχευσης μπορεί να προσφέρει σημαντικά οφέλη τόσο σε επενδυτές όσο και σε ελεγκτικούς φορείς, συμβάλλοντας στη μείωση οικονομικών απωλειών και στην ενίσχυση της σταθερότητας της αγοράς.

Στόχος της εργασίας είναι η σύγκριση διαφορετικών ταξινομητών και η αξιολόγηση της απόδοσής τους στο πρόβλημα δυαδικής ταξινόμησης εταιρειών σε υγιείς ή πτωχευμένες. Το σύνολο δεδομένων που χρησιμοποιήθηκε περιλαμβάνει οικονομικούς δείκτες και στοιχεία επιχειρηματικής δραστηριότητας για διάφορα έτη. Το πρόβλημα τίθεται ως δυαδική ταξινόμηση: η μεταβλητή εξόδου λαμβάνει την τιμή 0 για υγιείς εταιρείες και 1 για πτωχευμένες.

Η εργασία περιλαμβάνει στάδια προεπεξεργασίας δεδομένων, κανονικοποίηση χαρακτηριστικών, αντιμετώπιση ανισόρροπης κατανομής των τάξεων, εφαρμογή Stratified K-Fold Cross-Validation, καθώς και εκπαίδευση οκτώ διαφορετικών αλγορίθμων ταξινόμησης. Τα αποτελέσματα παρουσιάζονται μέσω μετρικών απόδοσης και γραφημάτων, ενώ η τελική αξιολόγηση γίνεται με βάση δύο συγκεκριμένα κριτήρια επιτυχίας, που σχετίζονται με την ακρίβεια αναγνώρισης υγιών και πτωχευμένων επιχειρήσεων.

Θεωρητικό Υπόβαθρο

Στο πλαίσιο της παρούσας μελέτης εφαρμόστηκαν οκτώ διαφορετικές μέθοδοι ταξινόμησης, οι οποίες ανήκουν σε διαφορετικές κατηγορίες αλγορίθμων επιβλεπόμενης μάθησης.

Η γραμμική διακριτική ανάλυση (Linear Discriminant Analysis) είναι μια μέθοδος που βασίζεται στη μείωση της ενδοκλασικής διασποράς και στη μεγιστοποίηση της απόστασης μεταξύ των κλάσεων. Πρόκειται για μία από τις πιο απλές και αποτελεσματικές γραμμικές μεθόδους, χωρίς ιδιαίτερα κρίσιμες υπερπαραμέτρους που να απαιτούν βελτιστοποίηση.

Η λογιστική παλινδρόμηση (Logistic Regression) αποτελεί επίσης ένα γραμμικό μοντέλο, κατάλληλο για δυαδικά προβλήματα ταξινόμησης. Από τις σημαντικότερες υπερπαραμέτρους της είναι η C, η οποία ρυθμίζει το βαθμό κανονικοποίησης (regularization) του μοντέλου, καθώς και ο solver, δηλαδή ο αλγόριθμος επίλυσης του μοντέλου, για τον οποίο επιλέχθηκε ο liblinear.

Ο αλγόριθμος δέντρου αποφάσεων (Decision Tree) είναι ένας ιεραρχικός ταξινομητής, βασισμένος σε διαδοχικούς κόμβους τύπου if-else. Η βασική του παράμετρος είναι το max_depth, το οποίο ελέγχει το βάθος του δέντρου και επομένως την πιθανότητα υπερεκπαίδευσης. Χρησιμοποιήθηκε επίσης το κριτήριο Gini για τον υπολογισμό της πληροφορίας.

Το Random Forest είναι μία μέθοδος συνόλου που συνδυάζει πλήθος δέντρων αποφάσεων, δημιουργώντας ένα πιο σταθερό και αξιόπιστο μοντέλο. Εφαρμόστηκε με τις προκαθορισμένες ρυθμίσεις του scikit-learn, όπως n_estimators=100 και max_depth=None.

Ο αλγόριθμος K-Nearest Neighbors (KNN) βασίζεται σε μετρικές απόστασης για την ταξινόμηση ενός δείγματος με βάση τους πλησιέστερους γείτονές του. Χρησιμοποιήθηκαν 5 γείτονες (n_neighbors=5) και η μετρική minkowski, η οποία αποτελεί γενίκευση της ευκλείδειας απόστασης.

Ο Naive Bayes βασίζεται στο θεώρημα του Bayes και υποθέτει την ανεξαρτησία μεταξύ των χαρακτηριστικών. Παρόλο που αυτή η υπόθεση σπάνια ισχύει πλήρως, ο αλγόριθμος είναι ταχύς και αποτελεσματικός, ιδιαίτερα σε προβλήματα με υψηλή διαστατικότητα. Για την εργασία αυτή, χρησιμοποιήθηκε ο GaussianNB.

Η υποστήριξη διανυσμάτων (Support Vector Machines) προσπαθεί να εντοπίσει το υπερεπίπεδο που διαχωρίζει τις δύο κλάσεις με το μέγιστο δυνατό περιθώριο. Χρησιμοποιήθηκε πυρήνας rbf και η τιμή της παραμέτρου C ορίστηκε στο 1.0.

Τέλος, εφαρμόστηκε ο αλγόριθμος XGBoost, μία ισχυρή υλοποίηση του gradient boosting. Ο συγκεκριμένος αλγόριθμος έχει αποδειχθεί ιδιαίτερα αποδοτικός σε πληθώρα προβλημάτων ταξινόμησης. Οι βασικές του παράμετροι επιλέχθηκαν ως εξής: αριθμός δέντρων n_estimators = 100, ρυθμός μάθησης learning_rate = 0.1, μέγιστο βάθος κάθε δέντρου max_depth = 3, και μετρική αξιολόγησης eval_metric = logloss.

Όλα τα παραπάνω μοντέλα ενσωματώθηκαν σε ένα ενιαίο πειραματικό σχήμα, στο οποίο εφαρμόστηκε κανονικοποίηση όλων των χαρακτηριστικών στην κλίμακα [0, 1], ενώ το σύνολο εκπαίδευσης εξισορροπήθηκε ως προς τις δύο τάξεις με αναλογία 3:1 υπέρ των υγιών επιχειρήσεων. Για την εκτίμηση της απόδοσης των μοντέλων χρησιμοποιήθηκε η τεχνική της διασταυρωμένης επικύρωσης τύπου Stratified K-Fold, με 4 πτυχές (folds).

Η διασταυρωμένη επικύρωση χωρίζει το σύνολο των δεδομένων σε K υποσύνολα ίσου μεγέθους, φροντίζοντας ώστε η κατανομή των τάξεων σε κάθε fold να είναι αντιπροσωπευτική της συνολικής κατανομής. Σε κάθε επανάληψη, ένα από τα folds χρησιμοποιείται ως σύνολο ελέγχου (test set), ενώ τα υπόλοιπα K–1 χρησιμοποιούνται για εκπαίδευση. Η διαδικασία επαναλαμβάνεται K φορές, ώστε κάθε fold να λειτουργήσει μία φορά ως test set. Η χρήση της στρατηγικής κατανομής (stratification) εξασφαλίζει δίκαιη σύγκριση μεταξύ των μοντέλων, περιορίζοντας την πιθανότητα μεροληψίας λόγω ανισόρροπων δεδομένων.

Πειραματικά Αποτελέσματα

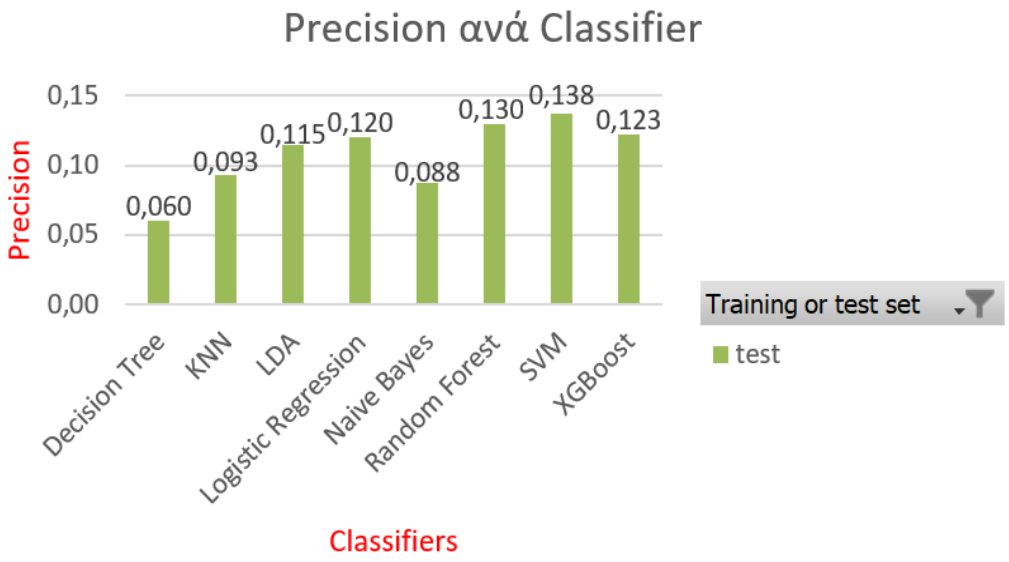
Η αξιολόγηση των ταξινομητών πραγματοποιήθηκε με χρήση 4-fold Stratified Cross-Validation, ώστε να διασφαλιστεί η σταθερότητα των αποτελεσμάτων σε διαφορετικά υποσύνολα του συνόλου δεδομένων. Για κάθε fold, υπολογίστηκαν οι βασικές μετρικές απόδοσης στο test set: Recall, Specificity, Precision, F1 Score, ROC-AUC και Balanced Accuracy. Στη συνέχεια, για κάθε αλγόριθμο υπολογίστηκε ο μέσος όρος των τιμών αυτών ανά μετρική.

Classifier	Precision	Recall	F1	Specificity	Balanced Accuracy	ROC-AUC
Decision Tree	0.060	0.460	0.105	0.832	0.646	0.645
KNN	0.093	0.450	0.150	0.895	0.674	0.802
LDA	0.115	0.412	0.180	0.926	0.669	0.837
Logistic Regression	0.120	0.367	0.183	0.936	0.652	0.835
Naive Bayes	0.088	0.510	0.153	0.876	0.694	0.817
Random Forest	0.130	0.470	0.200	0.924	0.698	**0.850**
SVM	**0.138**	0.400	**0.203**	**0.939**	0.670	0.835
XGBoost	0.123	**0.585**	0.200	0.898	**0.742**	0.835

Πίνακας 1: Μέσες τιμές μετρικών απόδοσης για κάθε αλγόριθμο ταξινόμησης

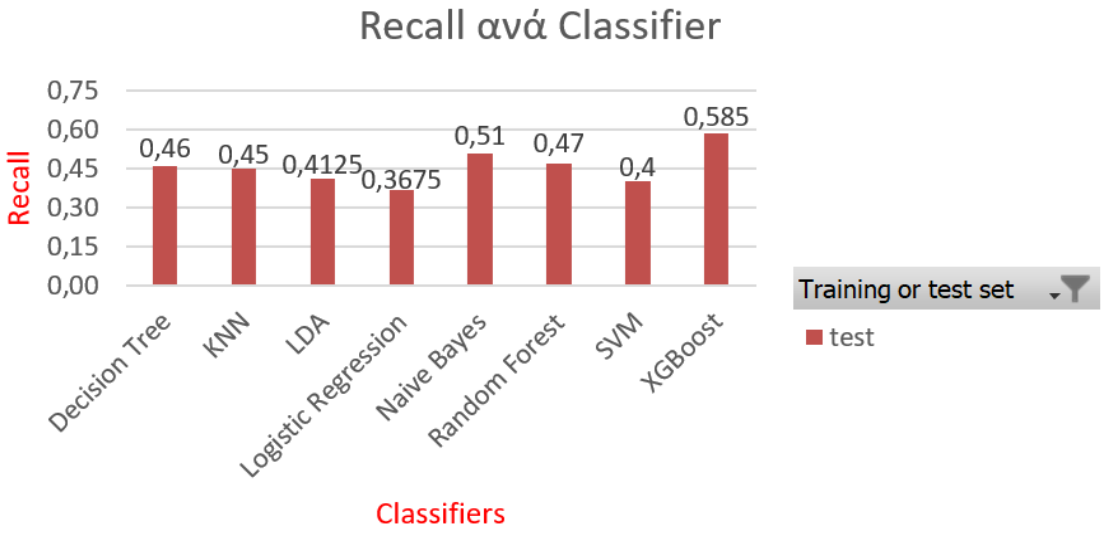
Ο πίνακας παρέχει μια συνολική επισκόπηση της απόδοσης των οκτώ ταξινομητών που εφαρμόστηκαν στο πρόβλημα δυαδικής ταξινόμησης των επιχειρήσεων σε υγιείς και πτωχευμένες. Οι τιμές που παρατίθενται αποτελούν μέσους όρους των μετρικών στο test set και επιτρέπουν άμεση σύγκριση μεταξύ των μοντέλων σε πολλαπλά κριτήρια. Από τον πίνακα παρατηρείται ότι ο XGBoost καταγράφει την υψηλότερη τιμή σε τρεις κρίσιμες μετρικές: Recall (0,585), ROC-AUC (0,835) και Balanced Accuracy (0,742) γεγονός που υποδεικνύει την ικανότητά του να ανιχνεύει επαρκώς τις πτωχευμένες επιχειρήσεις, ενώ διατηρεί παράλληλα ικανοποιητική ακρίβεια για τις υγιείς. Αντίστοιχα, ταξινομητές όπως ο LDA και η Logistic Regression εμφανίζουν υψηλή Specificity αλλά πολύ χαμηλότερο Recall, γεγονός που περιορίζει την πρακτική τους χρησιμότητα σε περιβάλλοντα όπου η πρόβλεψη πτώχευσης έχει κρίσιμη σημασία.

Η χρήση του πίνακα επιτρέπει την ταυτόχρονη θεώρηση όλων των μετρικών, βοηθώντας στην αξιολόγηση των trade-offs που κάνει κάθε μοντέλο μεταξύ ευαισθησίας (Recall) και ειδικότητας (Specificity), αλλά και στη συνολική του ισορροπία μέσω του δείκτη Balanced Accuracy.



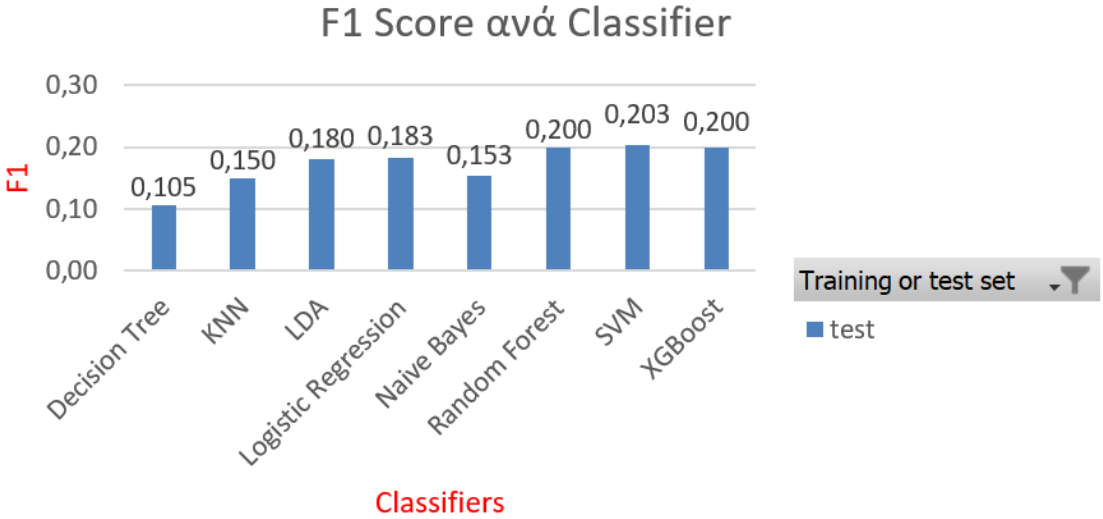
Εικόνα 1: Μέσος όρος Precision για κάθε αλγόριθμο ταξινόμησης

Το Precision εκφράζει το ποσοστό των θετικών προβλέψεων που ήταν πράγματι σωστές, επομένως είναι ιδιαίτερα σημαντικό όταν είναι κρίσιμο να αποφεύγονται τα ψευδώς θετικά (false positives). Στο συγκεκριμένο πρόβλημα, αντιστοιχεί στην ακρίβεια με την οποία οι ταξινομητές αναγνωρίζουν πτωχευμένες εταιρείες χωρίς να "παρασύρουν" υγιείς επιχειρήσεις ως προβληματικές. Από το γράφημα παρατηρείται ότι ο SVM πετυχαίνει τη μεγαλύτερη τιμή (0,138), με Random Forest (0,130) και XGBoost (0,123) να ακολουθούν. Αντίθετα, οι πιο απλοί ταξινομητές όπως το Decision Tree και ο Naive Bayes εμφανίζουν σημαντικά χαμηλότερες επιδόσεις (0,060 και 0,088 αντίστοιχα), κάτι που σημαίνει πως έχουν μεγαλύτερη τάση να κάνουν λανθασμένες θετικές προβλέψεις. Ο Logistic Regression και το LDA προσφέρουν μέτριες επιδόσεις (0,120 και 0,115), καταδεικνύοντας ότι η επιλογή πιο εξελιγμένων μοντέλων μπορεί να οδηγήσει σε πιο αξιόπιστες προβλέψεις με λιγότερα false positives.



Εικόνα 2: Μέσος όρος Recall για κάθε αλγόριθμο ταξινόμησης

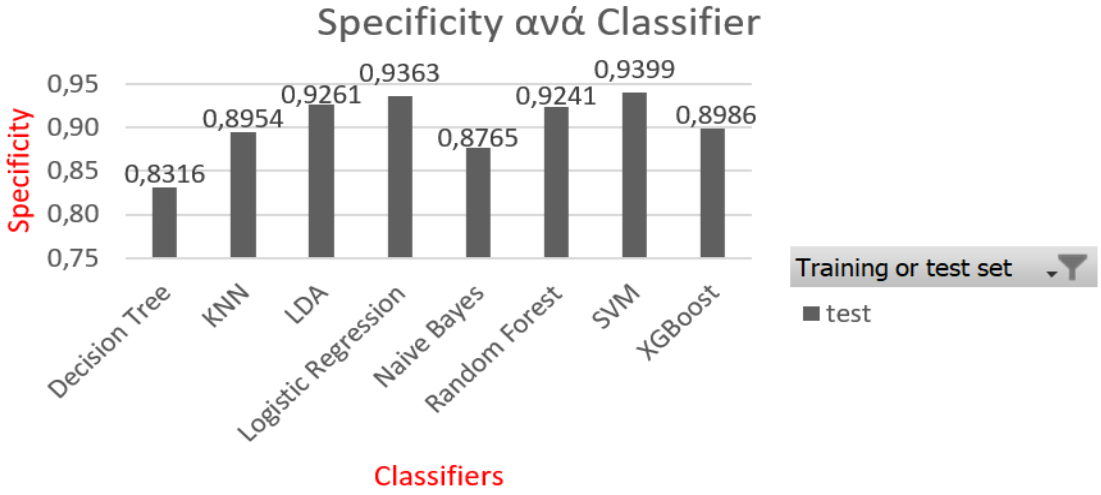
Το Recall αποτελεί κρίσιμη μετρική για το πρόβλημα πρόβλεψης πτώχευσης, καθώς εκφράζει την ικανότητα του μοντέλου να εντοπίζει σωστά τις πτωχευμένες επιχειρήσεις. Από το γράφημα προκύπτει ότι ο αλγόριθμος XGBoost παρουσιάζει την υψηλότερη τιμή (0,585), καταδεικνύοντας αυξημένη ικανότητα ανίχνευσης της θετικής κλάσης (πτώχευση). Ο Naive Bayes ακολουθεί με 0,51, ενώ σχετικά ικανοποιητικές τιμές παρατηρούνται και για τους Random Forest (0,47) και Decision Tree (0,46). Αντίθετα, ο Logistic Regression και ο LDA εμφανίζουν τις χαμηλότερες επιδόσεις (0,3675 και 0,4125 αντίστοιχα), υποδεικνύοντας δυσκολία στην αναγνώριση των επιχειρήσεων υψηλού κινδύνου. Συνεπώς, το XGBoost είναι ο πιο αποτελεσματικός ταξινομητής όσον αφορά την ευαισθησία στο θετικό κλάδο.



Εικόνα 3: Μέσος όρος F1 Score για κάθε αλγόριθμο ταξινόμησης

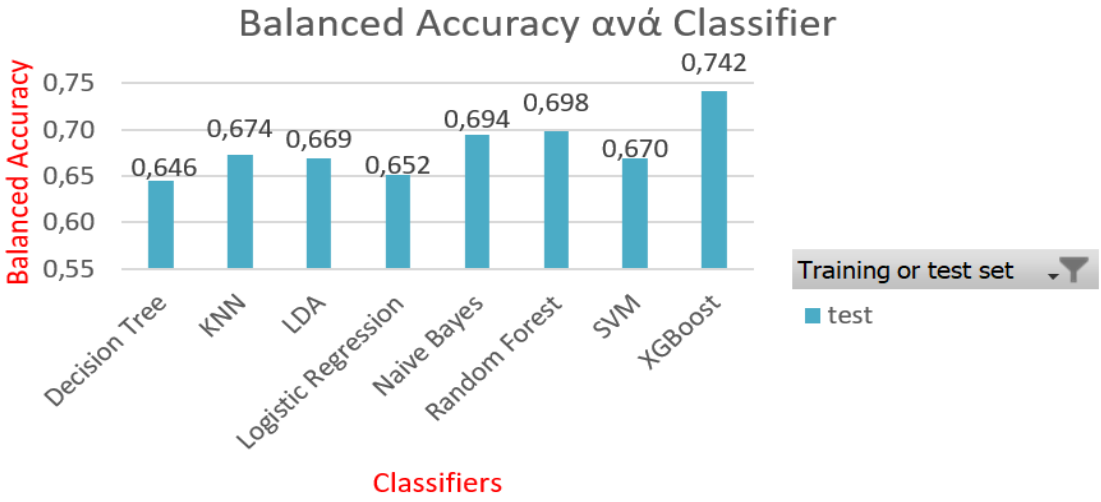
Το F1 Score αποτελεί σημαντική μετρική σε προβλήματα με ανισόρροπες κλάσεις, καθώς συνδυάζει την ακρίβεια (Precision) και την ανάκληση (Recall) σε μία συνολική τιμή. Στο παραπάνω σχήμα παρατηρούμε ότι οι αλγόριθμοι Random Forest**,** SVM και XGBoost επιτυγχάνουν τις υψηλότερες τιμές F1 (περίπου 0.20), γεγονός που υποδεικνύει καλύτερη ισορροπία μεταξύ ανίχνευσης πτωχευμένων εταιρειών και αποφυγής ψευδών θετικών. Αντίθετα, ο Decision Tree παρουσιάζει την χαμηλότερη επίδοση (0.105), κάτι που ενδέχεται να οφείλεται σε υπερεκπαίδευση ή σε ευαισθησία σε μικρές αλλαγές των

δεδομένων. Συνολικά, τα ensemble μοντέλα και οι πιο πολύπλοκοι ταξινομητές φαίνεται να ανταποκρίνονται καλύτερα στις απαιτήσεις του προβλήματος.



Εικόνα 4: Μέσος όρος Specificity για κάθε αλγόριθμο ταξινόμησης

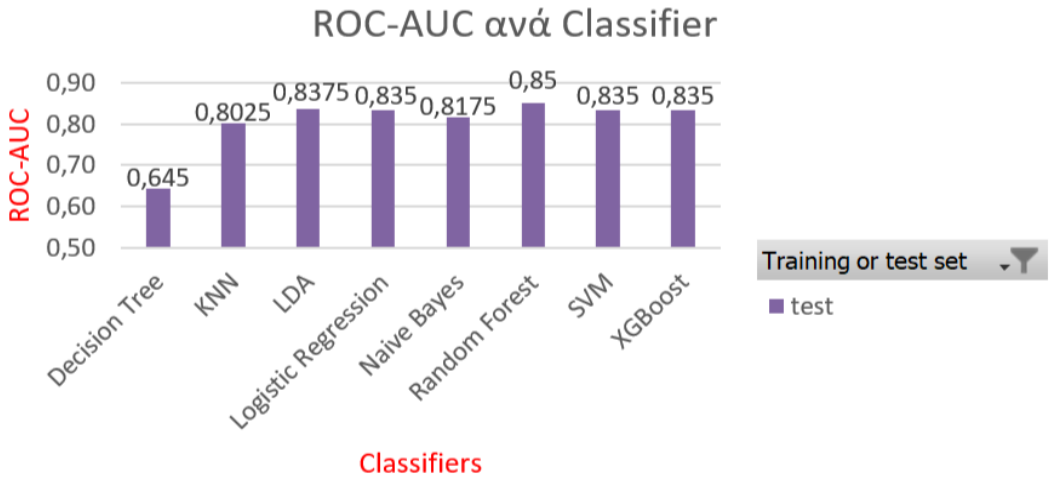
Η μετρική Specificity μετρά την ικανότητα ενός μοντέλου να εντοπίζει σωστά τις υγιείς εταιρείες, δηλαδή να αποφεύγει τα false positives. Στο πρόβλημα που μελετάται, υψηλό specificity σημαίνει ότι το μοντέλο σπάνια χαρακτηρίζει μια υγιή εταιρεία ως πτωχευμένη, κάτι που είναι ιδιαίτερα σημαντικό για επενδυτές και φορείς που δεν θέλουν να αποκλείσουν αδικαιολόγητα σταθερές επιχειρήσεις. Σύμφωνα με το γράφημα, το SVM παρουσιάζει τη μεγαλύτερη τιμή (0,9399), ενώ ακολουθεί η Logistic Regression (0,9363) και το LDA (0,9261). Ο Random Forest και το XGBoost επιτυγχάνουν επίσης υψηλές επιδόσεις (>0,89), αποδεικνύοντας ότι τα σύγχρονα μοντέλα είναι αξιόπιστα στη σωστή αναγνώριση των υγιών περιπτώσεων. Αντίθετα, ο Decision Tree έχει τη χαμηλότερη τιμή (0,8316), γεγονός που δείχνει τάση υπερταξινόμησης προς την πτώχευση.



Εικόνα 5: Μέσος όρος Balanced Accuracy για κάθε αλγόριθμο ταξινόμησης

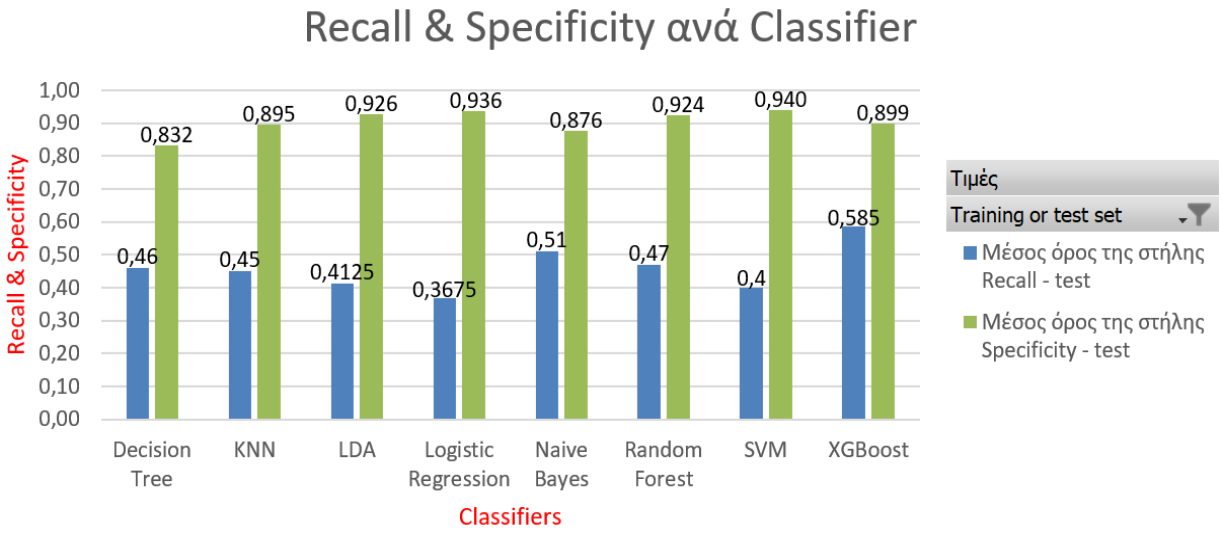
Η μετρική Balanced Accuracy λαμβάνει υπόψη τόσο το recall για τις πτωχευμένες εταιρείες όσο και το specificity για τις υγιείς, γεγονός που την καθιστά κατάλληλη για προβλήματα με ανισόρροπες κλάσεις. Στο παραπάνω γράφημα, ο XGBoost επιτυγχάνει την υψηλότερη balanced accuracy (0,742), ξεπερνώντας όλους τους υπόλοιπους ταξινομητές. Ακολουθούν οι Random Forest (0,698) και Naive Bayes (0,694), ενώ οι KNN και SVM βρίσκονται σε ενδιάμεσες θέσεις. Η Logistic Regression και το Decision Tree παρουσιάζουν τις χαμηλότερες τιμές (0,652 και 0,646 αντίστοιχα). Τα αποτελέσματα αυτά

επιβεβαιώνουν την υπεροχή του XGBoost όσον αφορά την ισορροπημένη απόδοση μεταξύ των δύο τάξεων, καθιστώντας τον ιδιαίτερα αξιόπιστο στο συγκεκριμένο πρόβλημα ταξινόμησης.



Εικόνα 6: Μέσος όρος ROC-AUC για κάθε αλγόριθμο ταξινόμησης

Ο δείκτης ROC-AUC αποτελεί μια συνολική μετρική αξιολόγησης για δυαδικούς ταξινομητές, καθώς εκτιμά την ικανότητα του μοντέλου να διαχωρίζει σωστά τις δύο κλάσεις (υγιείς και πτωχευμένες εταιρείες) σε όλα τα πιθανά κατώφλια ταξινόμησης. Όσο μεγαλύτερη είναι η τιμή του, τόσο καλύτερα το μοντέλο διαχωρίζει τις περιπτώσεις. Σύμφωνα με τα αποτελέσματα, το Random Forest σημειώνει την υψηλότερη τιμή (0,85), αποδεικνύοντας την ικανότητά του να διακρίνει τις δύο κατηγορίες με ακρίβεια. Ακολουθούν LDA, SVM και XGBoost με παρόμοιες επιδόσεις (0,835–0,8375), ενώ αξιοπρεπή αποτελέσματα εμφανίζουν επίσης τα Naive Bayes και KNN. Αντίθετα, το Decision Tree υπολείπεται αισθητά (0,645), γεγονός που δείχνει ότι αποτυγχάνει να αποδώσει έναν σταθερό διαχωρισμό μεταξύ των κλάσεων.

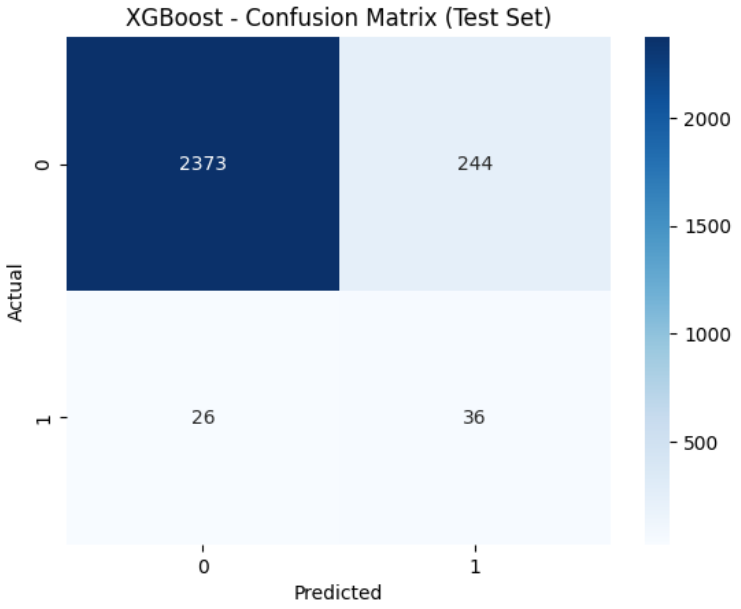


Εικόνα 7: Σύγκριση των μετρικών Recall και Specificity στο test set για κάθε αλγόριθμο ταξινόμησης

Το γράφημα συγκρίνει τις τιμές των μετρικών Recall και Specificity για κάθε αλγόριθμο ταξινόμησης Παρατηρούμε ότι οι περισσότεροι ταξινομητές επιτυγχάνουν υψηλές τιμές specificity, με τον Logistic Regression (0,936) και τον LDA (0,926) να

ξεχωρίζουν, παρουσιάζοντας υψηλή ικανότητα αναγνώρισης υγιών επιχειρήσεων. Από την άλλη πλευρά, οι τιμές Recall είναι σημαντικά χαμηλότερες για όλους τους αλγόριθμους, υποδεικνύοντας δυσκολία στην ανίχνευση πτωχευμένων εταιρειών. Το μοντέλο XGBoost εμφανίζει την καλύτερη ισορροπία ανάμεσα στις δύο μετρικές, με recall (0,585) και specificity (0,8986), πλησιάζοντας περισσότερο τα επιχειρησιακά όρια που τέθηκαν στην εκφώνηση. Αντιθέτως, μοντέλα όπως ο KNN (recall: 0,45 , specificity: 0,8954) και ο Naive Bayes (recall: 0,3675 , specificity: 0,8765) παρουσιάζουν μειωμένη απόδοση και στις δύο διαστάσεις.

Συνολικά, διαπιστώνεται ότι οι περισσότεροι ταξινομητές διακρίνονται περισσότερο στη σωστή πρόβλεψη υγιών εταιρειών (υψηλό specificity), αλλά υστερούν στην ανίχνευση των πτωχευμένων (χαμηλό recall). Η επίτευξη ισορροπίας μεταξύ recall και specificity είναι κρίσιμη σε τέτοια προβλήματα, και το XGBoost αναδεικνύεται ως το πιο αξιόπιστο και ισορροπημένο μοντέλο σε αυτό το πλαίσιο.



Εικόνα 8: Πίνακας σύγχυσης για τον αλγόριθμο ταξινόμησης XGBoost.

Η παραπάνω μήτρα σύγχυσης αποτυπώνει την απόδοση του ταξινομητή XGBoost στο test set του τέταρτου fold. Οι τιμές απεικονίζουν την επίδοση του μοντέλου στην κατηγοριοποίηση υγιών (0) και πτωχευμένων (1) εταιρειών. Το μοντέλο ταξινόμησε σωστά 2.373 από τις 2.617 υγιείς επιχειρήσεις (true negatives), επιτυγχάνοντας specificity περίπου 90,7%, ενώ εντόπισε επιτυχώς 36 από τις 62 πτωχευμένες (true positives), με recall περίπου 58%. Το συγκεκριμένο παράδειγμα επιβεβαιώνει την ανάγκη για περαιτέρω ενίσχυση των δεδομένων μειοψηφίας ή ρύθμιση υπερπαραμέτρων για βελτίωση της ανάκτησης πτωχευμένων περιπτώσεων.

Συμπεράσματα

Στο πλαίσιο αυτής της εργασίας συγκρίθηκαν οκτώ ταξινομητικοί αλγόριθμοι μηχανικής μάθησης για την πρόβλεψη πτώχευσης επιχειρήσεων, με στόχο την ταξινόμηση εταιρειών σε υγιείς ή πτωχευμένες, με βάση χρηματοοικονομικά και λειτουργικά χαρακτηριστικά. Μετά την ανάλυση των αποτελεσμάτων και τη σύγκριση των μετρικών

απόδοσης, προτείνεται ως πιο αποτελεσματικό το μοντέλο XGBoost, το οποίο εμφανίζει την υψηλότερη ισορροπία μεταξύ των βασικών μετρικών, με ιδιαίτερα υψηλές επιδόσεις σε ROC-AUC και Balanced Accuracy, και σημαντικά καλύτερο recall σε σχέση με τους υπόλοιπους ταξινομητές.

Ωστόσο, τίθενται δύο συγκεκριμένες προϋποθέσεις: πρώτον, το μοντέλο να εντοπίζει τουλάχιστον το 60% των πτωχευμένων εταιρειών (recall ≥ 60%), και δεύτερον, να εντοπίζει τουλάχιστον το 70% των υγιών εταιρειών (specificity ≥ 70%). Διαπιστώνεται ότι κανένας από τους αλγόριθμους ταξινόμησης δεν ικανοποιεί και τα δύο κριτήρια ταυτόχρονα. Ο XGBoost πλησιάζει περισσότερο, με recall περίπου 58,5% και specificity 89,9%, αλλά υπολείπεται οριακά στην πρώτη προϋπόθεση. Άλλοι ταξινομητές, όπως LDA ή Logistic Regression, επιτυγχάνουν υψηλό specificity, αλλά έχουν χαμηλό recall, και συνεπώς αποτυγχάνουν να ανιχνεύσουν επαρκώς τις πτωχευμένες περιπτώσεις.

Για τη βελτίωση της απόδοσης στο μέλλον, κρίσιμη είναι η εφαρμογή πιο αποτελεσματικών τεχνικών εξισορρόπησης του training set, όπως η χρήση του SMOTE ή του ADASYN, ώστε να αυξηθεί η παρουσία της μειονοτικής τάξης χωρίς υπερδειγματοληψία. Παράλληλα, θα μπορούσε να πραγματοποιηθεί επιλογή ή κατασκευή πιο διακριτών χαρακτηριστικών μέσω feature engineering, που θα επιτρέπουν καλύτερο διαχωρισμό των τάξεων. Η ρύθμιση των υπερπαραμέτρων μέσω grid search ή Bayesian optimization, ιδιαίτερα για τα ισχυρά μοντέλα όπως το XGBoost και το Random Forest, ενδέχεται να αυξήσει περαιτέρω την ακρίβεια. Τέλος, η ενσωμάτωση εξωτερικών μεταβλητών, όπως ο κλάδος δραστηριότητας, η γεωγραφική τοποθεσία ή το μέγεθος των εταιρειών, μπορεί να προσδώσει περισσότερη πληροφορία και να ενισχύσει τη συνολική προβλεπτική ικανότητα του συστήματος.

Συμπερασματικά, αν και δεν πληροί αυστηρά το ζητούμενο όριο recall, το μοντέλο XGBoost προτείνεται ως η πιο αξιόπιστη επιλογή, εφόσον συνδυάζει σταθερότητα, καλή γενίκευση και ικανοποιητική ισορροπία απόδοσης ανάμεσα στις δύο κατηγορίες, ενώ προσφέρει δυνατότητες βελτίωσης με περαιτέρω επεξεργασία δεδομένων και παραμετροποίηση.

Χρησιμοποιούμενα Εργαλεία και Βιβλιοθήκες

Για την υλοποίηση της παρούσας εργασίας χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python, μέσω περιβάλλοντος Google Colab. Οι βασικές βιβλιοθήκες που αξιοποιήθηκαν είναι οι εξής:

- Pandas και NumPy: Για την ανάγνωση, επεξεργασία και ανάλυση των δεδομένων.
- Matplotlib και Seaborn: Για την απεικόνιση γραφημάτων και τη δημιουργία διαγνωστικών παραστάσεων.
- Scikit-learn: Για την υλοποίηση των περισσότερων αλγορίθμων ταξινόμησης, την κανονικοποίηση χαρακτηριστικών, την εφαρμογή cross-validation και την εξαγωγή μετρικών απόδοσης.
- XGBoost: Για την υλοποίηση του αλγορίθμου XGBoost, με έμφαση στη βελτιστοποίηση απόδοσης.
- OpenPyXL: Για την εξαγωγή και επεξεργασία των αποτελεσμάτων σε μορφή Excel, καθώς και για τη δημιουργία pivot charts στο περιβάλλον του Excel.

Η επιλογή των εργαλείων έγινε με γνώμονα τη σταθερότητα, την ευκολία χρήσης και τη δυνατότητα αναπαραγωγής των αποτελεσμάτων.