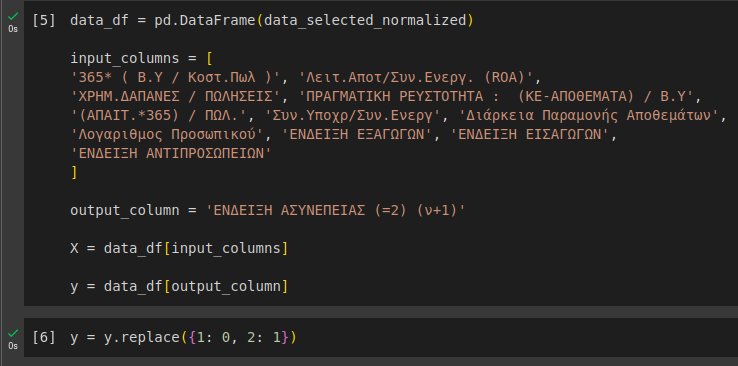
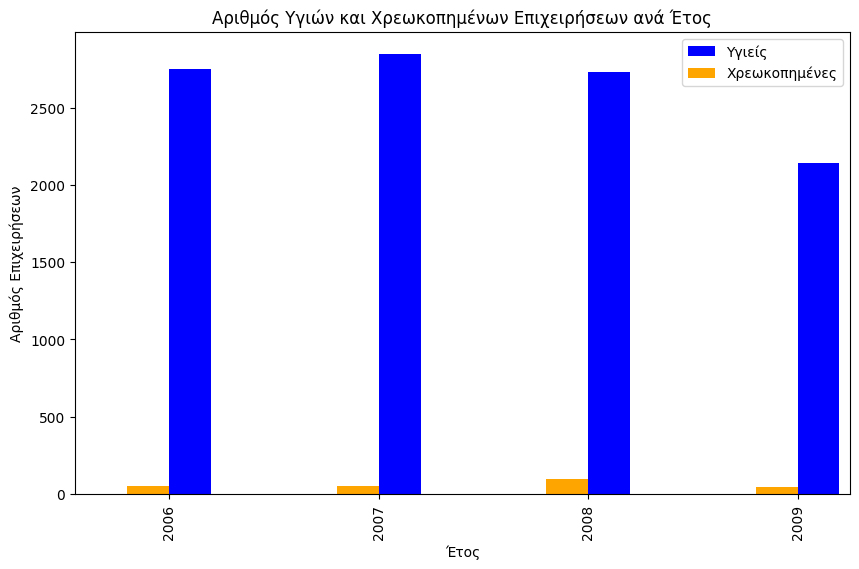
Η παρούσα αναφορά αναφέρεται στην εργασία που διενεργήθηκε στα πλαίσια του προγράμματος μεταπτυχιακών σπουδών 'Εφαρμοσμένη Πληροφορική' στο μάθημα Μέθοδοι και Εργαλεία της Τεχνητής Νοημοσύνης. Το πρόβλημα που αντιμετωπίσαμε εστιάζεται στον εντοπισμό εταιρειών που ενδέχεται να κηρύξουν χρεωκοπία, χρησιμοποιώντας δεδομένα που παρέχονται έτοιμα από κάποιον οργανισμό. Σκοπός της εργασίας μας ήταν να αναπτύξουμε ένα μοντέλο ταξινόμησης που θα μπορεί να προβλέπει με υψηλή ακρίβεια τις εταιρείες που ενδέχεται να κηρύξουν χρεωκοπία, λαμβάνοντας υπόψη διάφορους δείκτες απόδοσης και δραστηριοτήτων των εταιρειών.

Η μεθοδολογία που ακολουθήθηκε για την ανάπτυξη του μοντέλου περιλάμβανε τα εξής βήματα:

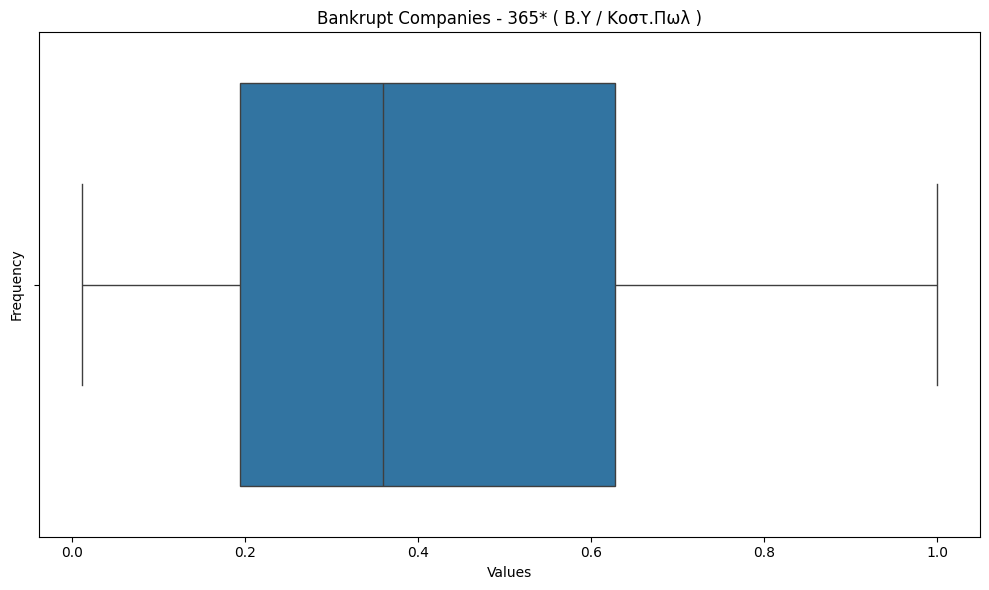
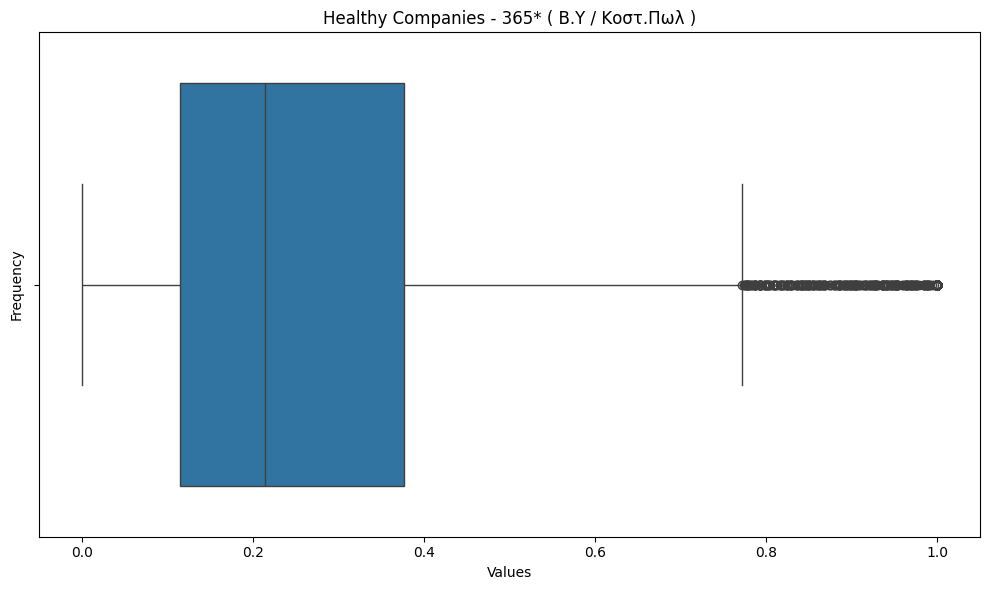
1. **Φόρτωση Δεδομένων**: Τα δεδομένα διαβάστηκαν από ένα αρχείο Excel, το οποίο περιείχε πληροφορίες σχετικά με την οικονομική κατάσταση εταιριών στις στήλες από Α μέχρι και Κ. Η τελευταία στήλη με το ΈΤΟΣ δεν τη χρειαζόμαστε, ενώ θα διαλέξουμε τη στήλη ΕΝΔΕΙΞΗ ΑΣΥΝΕΠΕΙΑΣ (=2) (ν+1) για τη μεταβλητή y. Επίσης στη μεταβλητή y



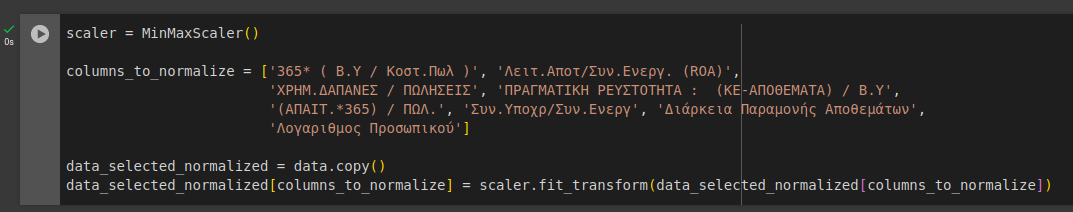
1. **Εξερεύνηση Δεδομένων**: Πραγματοποιήθηκε εξερεύνηση των δεδομένων με την εμφάνιση των πρώτων δειγμάτων και τη δημιουργία γραφημάτων για να αναλυθεί η κατάσταση των υγιών και των χρεωκοπημένων εταιριών ανά έτος όπου παρατηρείται ότι το dataset είναι unbalanced.



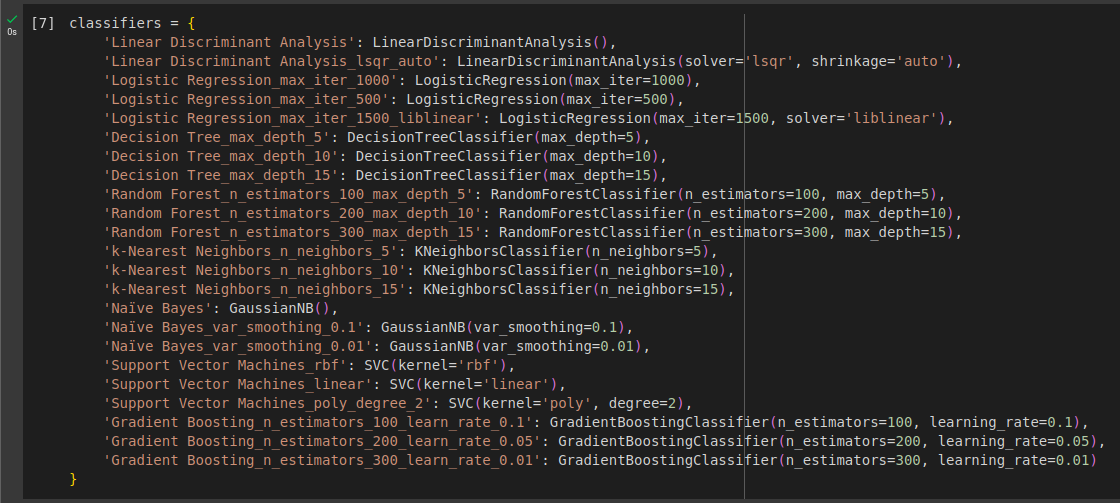
1. **Ανάλυση και Επεξεργασία Δεδομένων**: Εκτελέστηκε ανάλυση στατιστικών μεγεθών για κάθε δείκτη, συμπεριλαμβανομένων των μέγιστων, ελαχίστων και μέσων τιμών, ώστε να κατανοηθεί η κατανομή των δεδομένων. Ενδεικτικά ακολουθούν 2 γραφήματα για τις υγιείς και χρεωκοπημένες εταιρείες όσον αφορά τη μετρική “365\* ( Β.Υ / Κοστ.Πωλ )”.



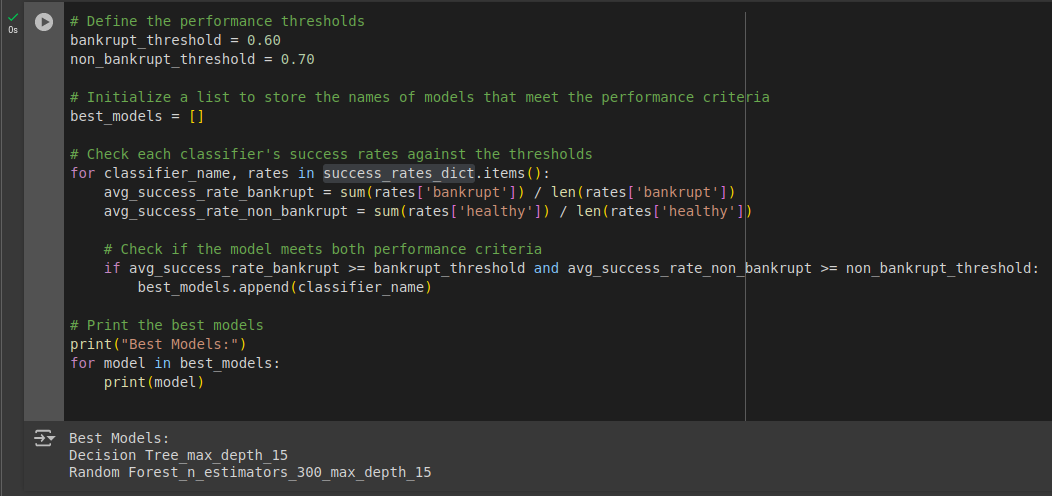
1. **Κανονικοποίηση Δεδομένων:** Εδώ χρησιμοποιείται ο MinMaxScaler από τη βιβλιοθήκη scikit-learn για την κανονικοποίηση συγκεκριμένων στηλών ενός πίνακα δεδομένων data. Οι στήλες που θα κανονικοποιηθούν αναφέρονται στη λίστα columns\_to\_normalize. Η κανονικοποίηση αυτή μετασχηματίζει τις τιμές των επιλεγμένων στηλών ώστε να βρίσκονται στην κλίμακα [0, 1]. Στη συνέχεια, το κανονικοποιημένο σύνολο δεδομένων αποθηκεύεται στη μεταβλητή data\_selected\_normalized. Αυτό βοηθά στη βελτίωση της απόδοσης αλγορίθμων μηχανικής μάθησης, οι οποίοι μπορεί να επηρεάζονται από την κλίμακα των δεδομένων.



1. **Επιλογή Μοντέλων**: Επιλέχθηκαν πολλά διαφορετικά μοντέλα μηχανικής μάθησης, συμπεριλαμβανομένων του LDA, Logistic Regression, Decision Tree, Random Forest, k-Nearest Neighbors, Naïve Bayes, Support Vector Machines και Gradient Boosting με διάφορες παραμέτρους έτσι ώστε να μπορούμε να συγκρίνουμε τα αποτελέσματα και να επιλέξουμε το καλύτερο μοντέλο. Ενδεικτικά στη παρακάτω φωτογραφία φαίνονται μαζικά ποιοι είναι οι classifiers που χρησιμοποιήσαμε:



1. **Εκπαίδευση και Αξιολόγηση Μοντέλων**: Τα μοντέλα εκπαιδεύτηκαν στα δεδομένα εκπαίδευσης και αξιολογήθηκαν χρησιμοποιώντας διάφορες μετρικές απόδοσης, όπως ο F1 Score, ο AUC ROC και οι επιτυχίες στις προβλέψεις των υγιών και των χρεωκοπημένων εταιριών. Στη συνέχεια μέσω της μέσης τιμής των success rate που υπολογίστηκαν βγαίνουν τα καλύτερα μοντέλα και στη περίπτωση του unbalanced dataset βλέπουμε ότι η επιλογή είναι Decision Tree με παραμέτρους max depth 15 και Random Forest με n\_estimators 300 και max depth 15.



1. **Εξαγωγή Αποτελεσμάτων**: Τα αποτελέσματα της ανάλυσης και της αξιολόγησης αποθηκεύτηκαν σε ένα αρχείο CSV για μελλοντική αναφορά.

Βάσει των δεδομένων απόδοσης για διάφορους ταξινομητές σε διαφορετικές πτυχές, πρέπει να γίνει μια λεπτομερής σύνοψη και σύγκριση:

**Linear Discriminant Analysis (LDA)**

Η LDA εμφανίζει υψηλή επιτυχία στην πρόβλεψη υγιών εταιρειών με ποσοστά επιτυχίας 0.99 τόσο στο σετ εκπαίδευσης όσο και στο σετ δοκιμής. Ωστόσο, παρουσιάζει πολύ χαμηλή απόδοση στην πρόβλεψη των πτωχευμένων εταιρειών, με ποσοστά που κυμαίνονται από 0.04 έως 0.07 στο σετ εκπαίδευσης και από 0.00 έως 0.10 στο σετ δοκιμής. Αυτό υποδηλώνει ότι ο LDA είναι αρκετά αποτελεσματικός για την αναγνώριση υγιών εταιρειών αλλά αποτυγχάνει σημαντικά στην πρόβλεψη των πτωχευμένων.

**Logistic Regression**

Η Logistic Regression με διαφορετικές παραμέτρους (max\_iter\_1000, max\_iter\_500, max\_iter\_1500\_liblinear) παρουσιάζει σχεδόν τέλεια ποσοστά επιτυχίας για τις υγιείς εταιρείες (1.00). Παρ' όλα αυτά, η απόδοσή της στην πρόβλεψη των πτωχευμένων εταιρειών είναι εξαιρετικά χαμηλή, κυμαινόμενη από 0.00 έως 0.02, ανεξαρτήτως του αριθμού επαναλήψεων. Αυτό δείχνει ότι, αν και αποτελεσματική για τις υγιείς, η Λογιστική Παλινδρόμηση δεν είναι κατάλληλη για την αναγνώριση των πτωχευμένων εταιρειών.

**Decision Tree**

Τα Decision Trees με βάθος 5 έχουν καλή απόδοση για τις υγιείς εταιρείες (1.00) αλλά πολύ χαμηλή για τις πτωχευμένες (0.06 έως 0.13). Με βάθος 10, παρουσιάζουν υψηλή μεταβλητότητα αλλά καλύτερη απόδοση για τις πτωχευμένες (0.48 έως 0.55). Τέλος, με βάθος 15, η απόδοση για τις πτωχευμένες βελτιώνεται αλλά παραμένει ασυνεπής (0.68 έως 0.76).

**Random Forest**

Τα Random Forest με 100 δέντρα και βάθος 5 έχουν τέλεια απόδοση για τις υγιείς (1.00) αλλά χαμηλή για τις πτωχευμένες (0.00). Με 200 δέντρα και βάθος 10, η απόδοση βελτιώνεται αλλά είναι ασυνεπής (0.19 έως 0.35). Με 300 δέντρα και βάθος 15, η απόδοση για τις πτωχευμένες βελτιώνεται σημαντικά (0.68 έως 0.79).

**k-Nearest Neighbors (k-NN)**

Ο k-NN με 5 γείτονες παρουσιάζει καλή απόδοση για τις υγιείς (1.00) αλλά ασυνεπή για τις πτωχευμένες (0.05 έως 0.13). Με 10 γείτονες, η απόδοση είναι πολύ υψηλή για τις υγιείς (1.00) αλλά μεικτή για τις πτωχευμένες (0.00 έως 0.03). Με 15 γείτονες, η απόδοση για τις υγιείς παραμένει υψηλή αλλά για τις πτωχευμένες είναι χαμηλή (0.00 έως 0.02).

**Naïve Bayes**

Ο Naïve Bayes έχει μέτρια απόδοση για τις υγιείς (0.94 έως 0.95) και χαμηλή για τις πτωχευμένες (0.21 έως 0.23). Με Var Smoothing 0.1, η απόδοση για τις υγιείς βελτιώνεται (0.97 έως 0.98) αλλά για τις πτωχευμένες παραμένει χαμηλή (0.08 έως 0.15). Με Var Smoothing 0.01, η απόδοση για τις πτωχευμένες βελτιώνεται (0.18 έως 0.32) αλλά παραμένει χαμηλή.

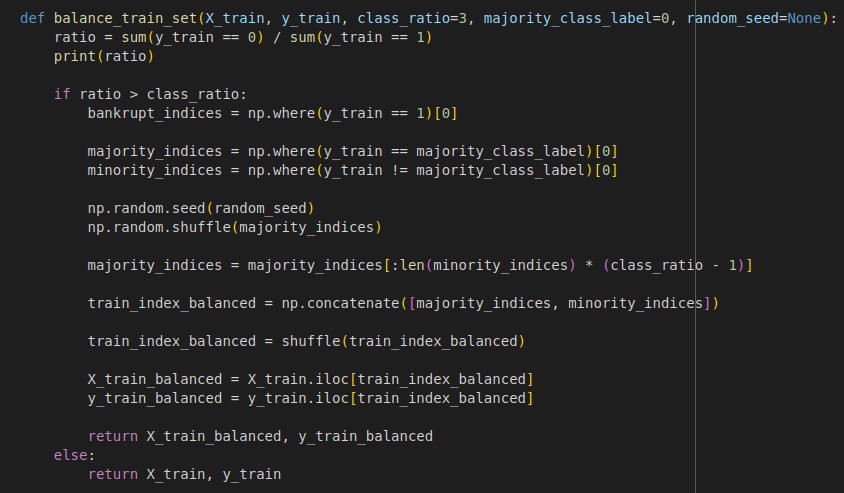
**Support Vector Machines (SVM)**

Οι SVMs με RBF και γραμμικό πυρήνα παρουσιάζουν τέλεια ποσοστά επιτυχίας για αμφότερες τις κατηγορίες (1.00). Η απόδοση του πολυωνυμικού πυρήνα (βαθμός 2) δεν παρέχεται λεπτομερώς, αλλά αναμένεται να είναι παρόμοια με τις άλλες δύο.

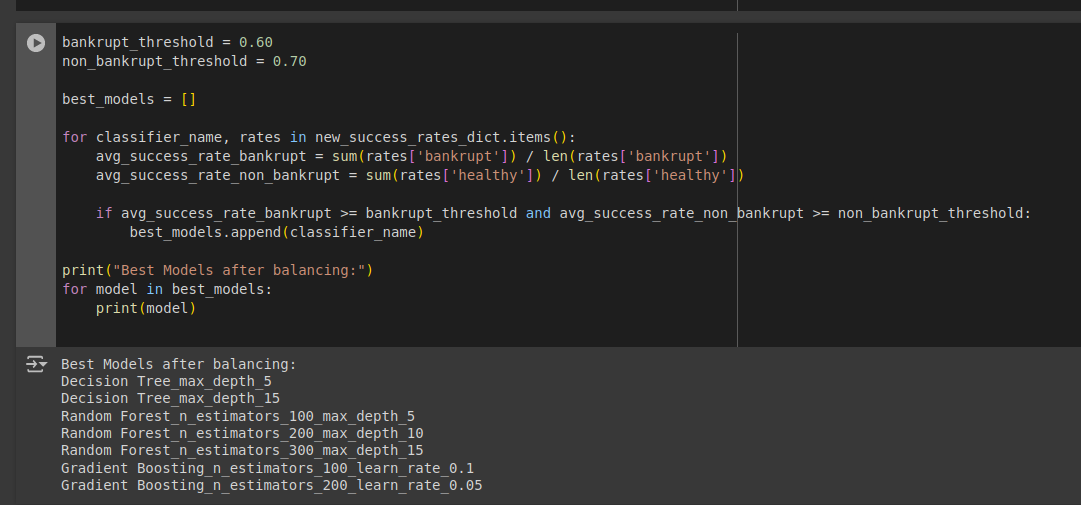
Η επιλογή των καλύτερων μοντέλων έγινε βάσει της αξιολόγησης των ποσοστών επιτυχίας για κάθε μοντέλο, τόσο για τις υγιείς όσο και για τις χρεοκοπημένες εταιρείες. Ορίστηκαν συγκεκριμένα κατώφλια απόδοσης: 0.60 για τις χρεοκοπημένες και 0.70 για τις υγιείς. Για κάθε μοντέλο, υπολογίστηκε ο μέσος όρος των ποσοστών επιτυχίας σε όλες τις πτυχές των δεδομένων. Μόνο τα μοντέλα που πληρούσαν και τα δύο αυτά κριτήρια επιλέχθηκαν ως τα καλύτερα. Από την ανάλυση αυτή, τα μοντέλα "Decision Tree με βάθος 15" και "Random Forest με 300 εκτιμητές και βάθος 15" ήταν αυτά που πληρούσαν τα κριτήρια και επιλέχθηκαν ως τα καλύτερα, λόγω της ισορροπημένης απόδοσής τους σε ανίχνευση και των δύο κατηγοριών.

Η διαδικασία της αξιολόγησης των μοντέλων μηχανικής μάθησης αντιμετώπισε πολλά προβλήματα, τα οποία απαιτούσαν σημαντική προσοχή και επεξεργασία. Αρχικά, η ανάγκη για καθαρισμό και προεπεξεργασία των δεδομένων ήταν κρίσιμη, καθώς ανάμεσα στα δεδομένα υπήρχαν πιθανώς ανεπιθύμητες τιμές ή απουσία δεδομένων που έπρεπε να διαχειριστούμε πριν από την εκπαίδευση των μοντέλων. Επιπλέον, η επιλογή του κατάλληλου μοντέλου μηχανικής μάθησης ήταν ένα σημαντικό βήμα. Κάθε μοντέλο έχει τα δικά του πλεονεκτήματα και μειονεκτήματα, και η επιλογή του σωστού μοντέλου εξαρτάται από τη φύση των δεδομένων και το πρόβλημα που επιχειρούμε να λύσουμε. Ένα ακόμα πρόβλημα ήταν η ανισορροπία στα δεδομένα, καθώς ο αριθμός των χρεωκοπημένων εταιρειών ήταν σημαντικά μικρότερος από τον αριθμό των υγιών εταιρειών. Αυτό μπορεί να οδηγήσει σε μη αξιόπιστες προβλέψεις, καθώς τα μοντέλα τείνουν να μην εκπαιδεύονται επαρκώς σε σπάνιες κατηγορίες.

Συνεπώς, ένα επόμενο βήμα είναι η εφαρμογή τεχνικών εξισορρόπησης των δεδομένων, προκειμένου να αντιμετωπιστεί η ανισορροπία αυτή. Αυτό μπορεί να γίνει με τη δημιουργία νέων δεδομένων ή με την προσαρμογή των υπαρχόντων, ώστε να δημιουργηθεί μια ισορροπημένη συλλογή που να περιλαμβάνει επαρκή πληροφορία για κάθε κλάση.



Θα εξισορροπήσουμε το σύνολο εκπαίδευσης για να βελτιώσουμε την απόδοση των μοντέλων μας στις χρεωκοπημένες εταιρείες. Αρχικά, υπολογίζουμε την αναλογία των υγιών προς τις χρεωκοπημένες εταιρείες. Αν η αναλογία αυτή είναι μεγαλύτερη από το καθορισμένο όριο (3:1), προχωρούμε στην εξισορρόπηση. Εντοπίζουμε τους δείκτες των υγιών και χρεωκοπημένων εταιρειών και επιλέγουμε τυχαία έναν υποσύνολο των υγιών εταιρειών ώστε να επιτύχουμε την επιθυμητή αναλογία. Συνδυάζουμε τους δείκτες των επιλεγμένων υγιών εταιρειών με αυτούς των χρεωκοπημένων και δημιουργούμε ένα νέο εξισορροπημένο σύνολο εκπαίδευσης. Τέλος, επιστρέφουμε το εξισορροπημένο σύνολο εκπαίδευσης, διασφαλίζοντας ότι τα δεδομένα μας είναι πιο ισορροπημένα, επιτρέποντας στα μοντέλα μας να εκπαιδευτούν καλύτερα και στις δύο κατηγορίες. Με αυτόν τον τρόπο, αντιμετωπίζουμε τα προβλήματα που προέκυψαν από την αρχική ανισορροπία των δεδομένων και επιδιώκουμε βελτιωμένα αποτελέσματα απόδοσης.



Αφού εκτελέσαμε το ίδιο πείραμα με εξισορροπημένα δεδομένα, τα καλύτερα μοντέλα μας είναι τα εξής: Decision Tree με μέγιστο βάθος 5, Decision Tree με μέγιστο βάθος 15, Random Forest με 100 δέντρα και μέγιστο βάθος 5, Random Forest με 200 δέντρα και μέγιστο βάθος 10, Random Forest με 300 δέντρα και μέγιστο βάθος 15, Gradient Boosting με 100 δέντρα και ρυθμό εκμάθησης 0.1, και Gradient Boosting με 200 δέντρα και ρυθμό εκμάθησης 0.05.

Η εξισορρόπηση των δεδομένων βελτίωσε την απόδοση των μοντέλων μας, επιτρέποντάς τους να εκπαιδευτούν αποτελεσματικότερα και να αποδώσουν καλύτερα στις χρεωκοπημένες εταιρείες. Χρησιμοποιώντας την τεχνική της εξισορρόπησης που περιγράφηκε, μειώσαμε την ανισορροπία μεταξύ των κατηγοριών, αυξάνοντας έτσι την ακρίβεια και την ευαισθησία των μοντέλων μας. Ως αποτέλεσμα, τα παραπάνω μοντέλα πληρούν τα κριτήρια απόδοσης και είναι τα πλέον κατάλληλα για την πρόβλεψη της χρεωκοπίας με βάση τα δεδομένα μας.