Classification Problems

Αμανατίδης Ιωάννης - ics23028

Ιωαννίδου Μαρία Μαρκέλλα - ics23082

Μαυρουδής Δημήτριος - ics23109

Μάθημα: Μηχανική Μάθηση

Ακαδημαϊκό έτος: 2025-2026

Πίνακας Περιεχομένων

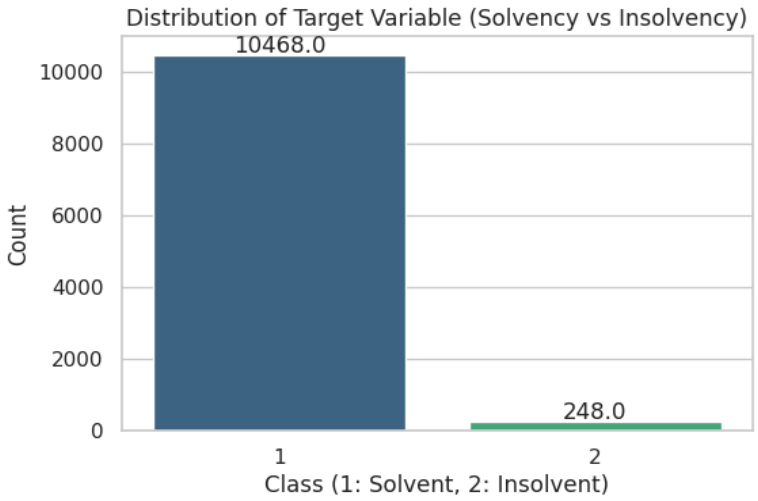
Εισαγωγή

Στην παρούσα εργασία αντιμετωπίζεται το πρόβλημα εντοπισμού εταιρειών που θα κηρύξουν χρεοκοπία. Στόχος της είναι μέσω των παρεχόμενων δεδομένων, τα οποία αφορούν ξεχωριστές εταιρείες, να γίνει μια συγκριτική ανάλυση και αξιολόγηση διαφορετικών τεχνικών ταξινόμησης σχετικά με την ικανότητα τους να διαχειριστούν το προαναφερθέν πρόβλημα και πόσο αποδοτικές μπορούν να είναι με βάση συγκεκριμένους περιορισμούς απόδοσης.

Περιγραφή Dataset

Το dataset που χρησιμοποιήθηκε στην εργασία περιλαμβάνει δεδομένα σχετικά με εταιρείες και, συγκεκριμένα, κάθε γραμμή αφορά διαφορετική εταιρεία. Παρέχεται ένα σύνολο χαρακτηριστικών που αποτυπώνουν την οικονομική κατάσταση και την δραστηριότητα κάθε εταιρείας:

1. **Στήλες A - H:** Αριθμητικοί δείκτες που υποδεικνύουν την οικονομική κατάσταση και την απόδοση.
2. **Στήλες I - K:** Δυαδικοί δείκτες (0 ή 1) οι οποίοι δηλώνουν τις δραστηριότητες που παρουσιάζει η κάθε εταιρεία (όπως εισαγωγές ή εξαγωγές).
3. **Στήλη L:** Αφορά την τωρινή κατάσταση της εταιρείας (1 όταν είναι υγιής, 2 όταν είναι χρεοκοπημένη), η οποία αποτελεί και τον στόχο της πρόβλεψης στα μοντέλα μηχανικής μάθησης που θα εκπαιδευτούν.
4. **Στήλη M:** Το έτος στο οποίο αναφέρονται τα προηγούμενα δεδομένα.



Γράφημα στο οποίο απεικονίζεται η αναλογία υγιών και χρεοκοπημένων εταιρειών. Είναι φανερό ότι υπάρχει αρκετή ανισορροπία.

Προτεινόμενη μεθοδολογία

Πριν αρχίσει η εκπαίδευση των μοντέλων, χρειάστηκε να ληφθούν κάποια απαραίτητα μέτρα. Αρχικά, έγινε έλεγχος για ελλιπείς εγγραφές στο dataset έτσι ώστε να αποφευχθούν πιθανά προβλήματα που αυτές θα προκαλούσαν αργότερα. Επίσης, εφαρμόστηκε κανονικοποίηση δεδομένων με χρήση τεχνικής τύπου Min-Max. Αυτό κρίθηκε αναγκαίο διότι τα δεδομένα του dataset παρουσιάζουν μεγάλες διαφορές κλίμακας. Δηλαδή, υπάρχουν δείκτες που φτάνουν τιμές εκατοντάδων ενώ άλλοι που δεν ξεπερνούν την τιμή «1». Χρησιμοποιήθηκε και η τεχνική Stratified k-fold έτσι ώστε τα μοντέλα να εκπαιδευτούν πάνω σε πολλά folds (συγκεκριμένα 4) και να μειωθεί ο παράγοντας της τύχης για ένα πιο αντικειμενικό αποτέλεσμα. Στην επιλογή της συγκεκριμένης τεχνικής συνέβαλε και το γεγονός της σημαντικής ανισορροπίας των δεδομένων, όπως παρατηρήθηκε και παραπάνω, η οποία βελτιώθηκε μετά από τυχαία ανακατανομή που εφαρμόστηκε με σκοπό η αναλογία να γίνει 3 υγιείς εταιρείες για κάθε χρεοκοπημένη.

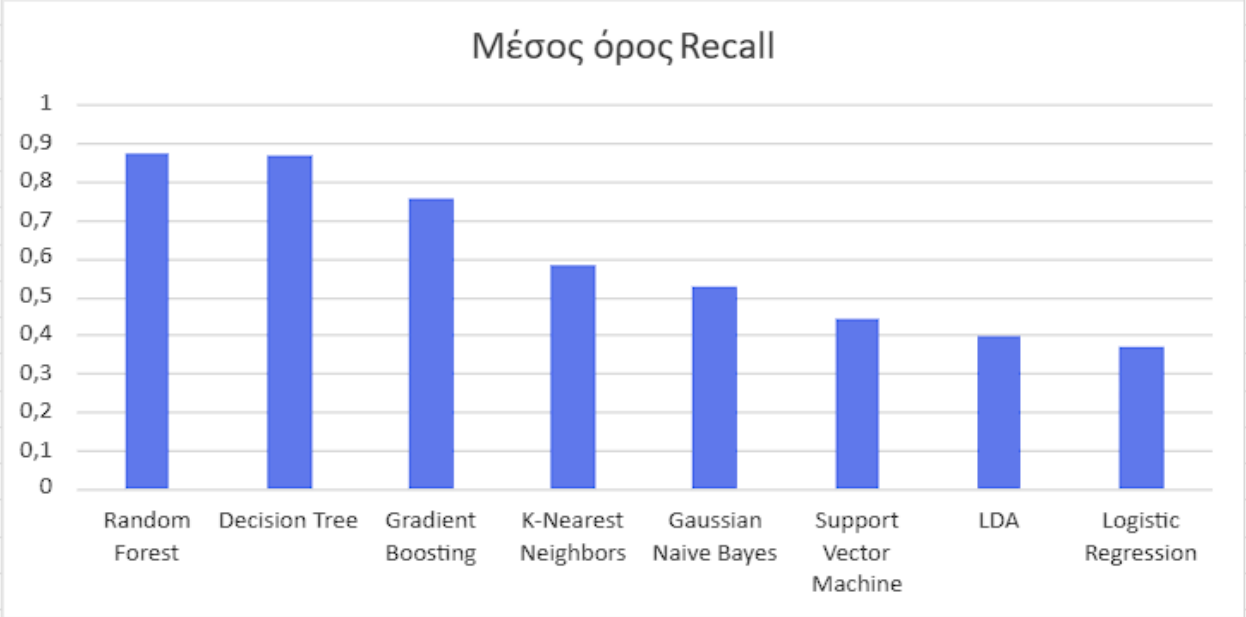
Τα μοντέλα που εκπαιδεύτηκαν είναι τα εξής:

1. **Linear Discriminant Analysis**
2. **Logistic Regression**
3. **Decision Trees**
4. **Random Forests**
5. **k-Nearest Neighbors**
6. **Naïve Bayes**
7. **Support Vector Machines**
8. **Gradient Boosting**

Τα πρώτα εφτά μοντέλα ήταν απαιτούμενα της εργασίας, ενώ το Gradient Boosting επιλέχθηκε από την ομάδα λόγω της υψηλής του ακρίβειας και της ικανότητας του να μειώνει τα λάθη του προηγούμενου μοντέλου σε κάθε επανάληψη. Για κάθε μοντέλο, υπολογίστηκαν διάφορες μετρικές οι οποίες θα αναλυθούν περαιτέρω παρακάτω.

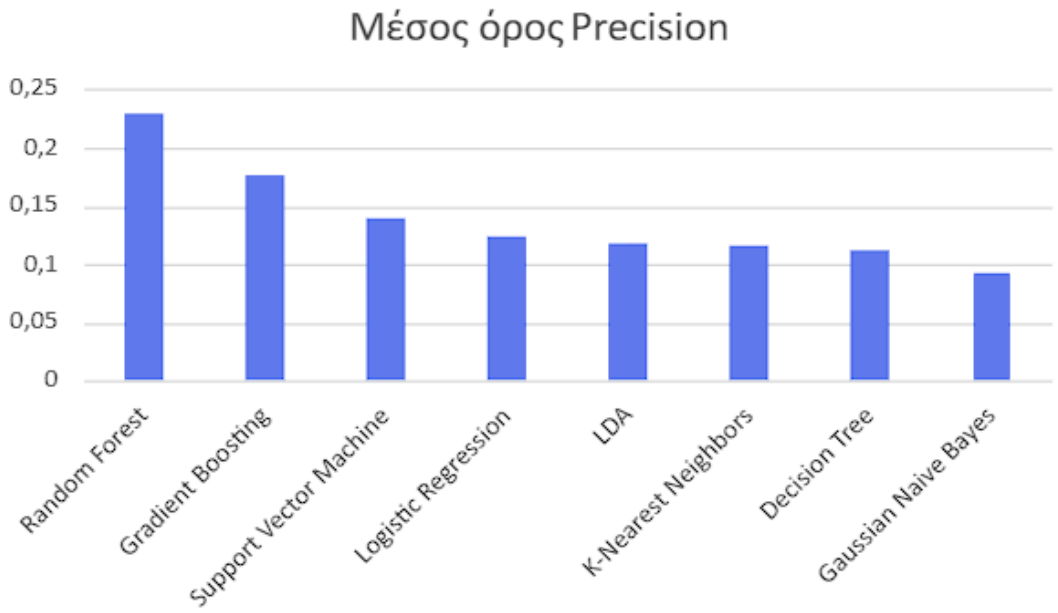
Πειραματικά Αποτελέσματα

Η επίδοση του κάθε εκπαιδευμένου μοντέλου εξετάστηκε μέσω των ακόλουθων μετρικών: **Accuracy, Precision, Recall, F1 score και AUC-ROC**. Τα αποτελέσματα που προέκυψαν είναι τα εξής:



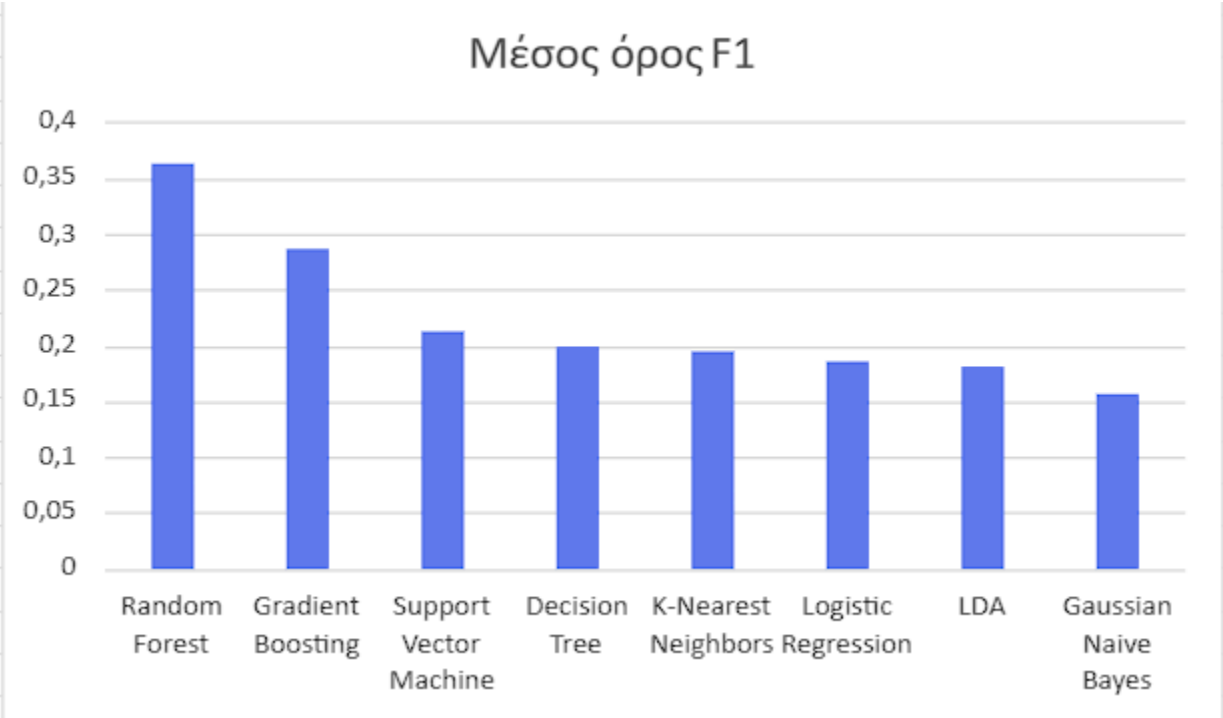
Μέσος όρος Recall μεταξύ όλων των folds για κάθε μοντέλο στο Test Set.

Το Recall ως μετρική, δείχνει πόσες από τις εταιρείες που πτώχευσαν πραγματικά, κατάφερε να εντοπίσει το μοντέλο. Όπως φαίνεται και στο παραπάνω γράφημα, τα Random Forest, Decision Tree και Gradient Boosting είχαν τις υψηλότερες τιμές recall με μέσους όρους 0.87, 0.86 και 0.75 αντίστοιχα.



Μέσος όρος Precision μεταξύ όλων των folds για κάθε μοντέλο στο Test Set.

Το precision δείχνει πόσες από τις εταιρείες που το μοντέλο προέβλεψε ότι θα χρεοκοπήσουν, το έκαναν πραγματικά. Από το γράφημα φαίνεται ότι τα Random Forest και Gradient Boosting έχουν πάλι τις υψηλότερες τιμές, οι οποίες είναι 0.22 και 0.17 αντίστοιχα.

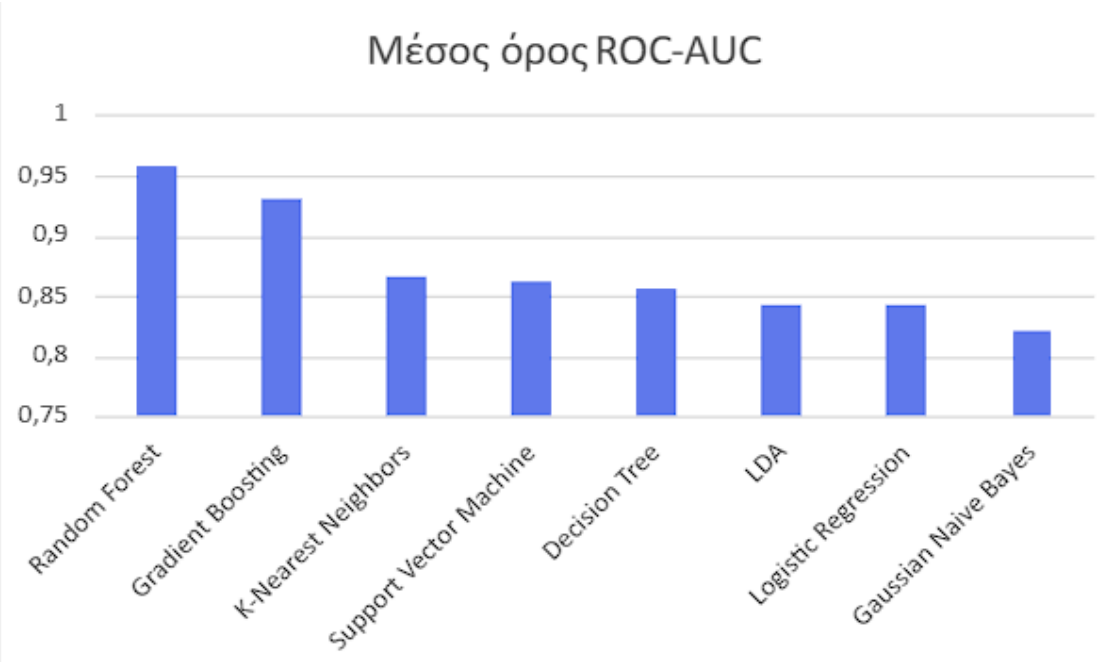


Μέσος όρος F1 μεταξύ όλων των folds για κάθε μοντέλο στο Test Set.

Το F1 συνδυάζει το precision και το recall σε μία μετρική. Είναι, ουσιαστικά, ο αρμονικός μέσος όρος τους. Για άλλη μια φορά, παρατηρείται ότι τα Random Forest και Gradient Boosting έχουν τις υψηλότερες τιμές, όπου είναι 0.36 και 0.28 αντίστοιχα.
Αξίζει να σημειωθεί ότι οι χαμηλές τιμές των F1 και Precision οφείλονται, πιθανότατα, στην μεγάλη ανισορροπία του dataset.

Μέσος όρος Accuracy μεταξύ όλων των folds για κάθε μοντέλο στο Test Set.

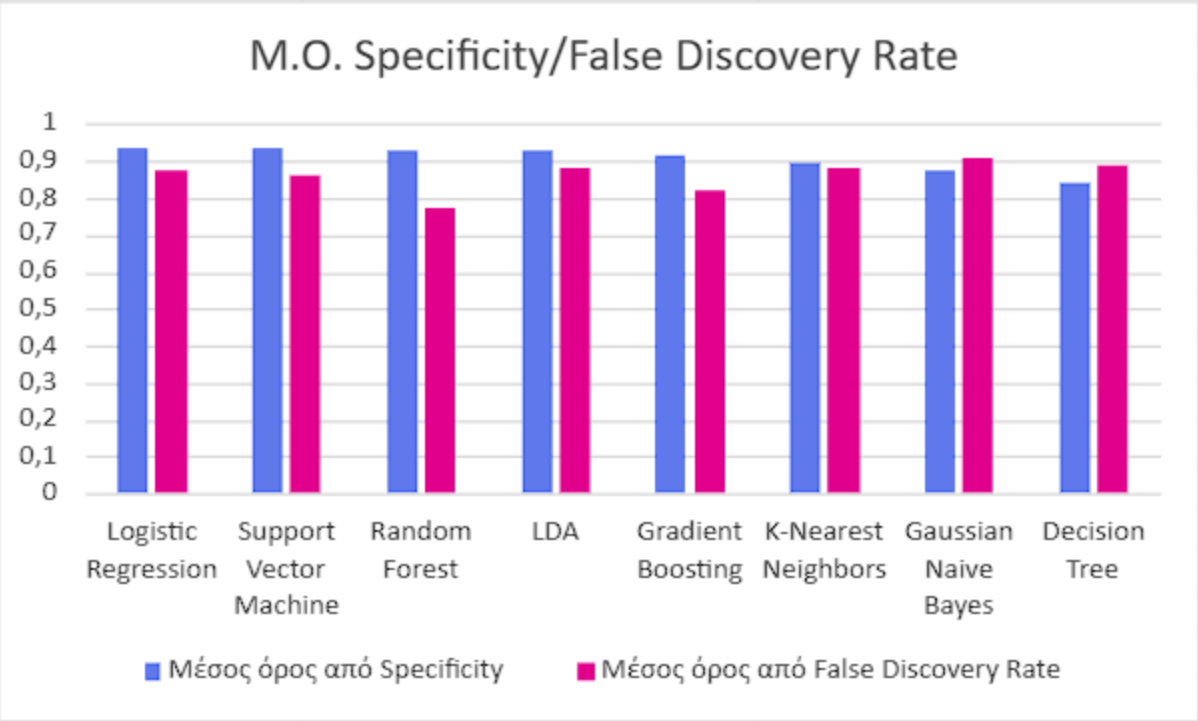
Το accuracy δείχνει το ποσοστό των σωστών προβλέψεων στο σύνολο όλων των προβλέψεων. Οι υψηλότερες τιμές στην συγκεκριμένη περίπτωση είναι 0.929, 0.925 και 0.923 για τα μοντέλα Random Forest, Logistic Regression και Support Vector Machine αντίστοιχα. Βέβαια, λόγω της ανισορροπίας του dataset, το accuracy χάνει σε μεγάλο βαθμό την αξιοπιστία του.



Μέσος όρος ROC-AUC μεταξύ όλων των folds για κάθε μοντέλο στο Test Set.

Η συγκεκριμένη μετρική δείχνει πόσο καλά ξεχωρίζει το κάθε μοντέλο τις δύο κλάσεις. Τα Random Forest και Gradient Boosting έχουν για άλλη μια φορά τις υψηλότερες τιμές οι οποίες είναι 0.95 και 0.93 αντίστοιχα. Από αυτές φαίνεται ότι τα δύο αυτά μοντέλα κάνουν πολύ καλή δουλειά στο να ξεχωρίζουν τις χρεοκοπημένες εταιρείες από τις υγιείς.

Πέρα από τις προηγούμενες μετρικές, η ομάδα επέλεξε ακόμα δύο για την ανάλυση της επίδοσης: **Specificity και False Positive Rate**.



Μέσος όρος Specificity/False Discovery Rate μεταξύ όλων των folds για κάθε μοντέλο στο Test Set.

Το Specificity, ως μετρική, δείχνει πόσες υγιείς εταιρείες κατάφερε να αναγνωρίσει το μοντέλο σωστά ως υγιείς. Τις μεγαλύτερες τιμές, οι οποίες είναι 0.938, 0.934 και 0.931, τις έχουν τα Logistic Regression, Support Vector Machine και Random Forest αντίστοιχα.
Από την άλλη, το False Discovery Rate, δείχνει πόσες από τις εταιρείες που το μοντέλο προέβλεψε ότι θα χρεοκοπήσουν, ήταν στην πραγματικότητα υγιείς. Ουσιαστικά, είναι το αντίθετο του Precision. Όπως ήταν αναμενόμενο, τα μοντέλα στα οποία παρατηρήθηκε χαμηλό Precision (Gaussian Naive Bayes, Decision Tree, K-Nearest Neighbors), έχουν υψηλότερο False Discovery Rate (0.90, 0.887, 0,883 αντίστοιχα). Από την άλλη, τα μοντέλα με υψηλό Precision (Random Forest, Gradient Boosting) έχουν χαμηλότερο False Discovery Rate (0.77, 0.82 αντίστοιχα). Βέβαια, οι τιμές παραμένουν σχετικά υψηλές καθώς και οι τιμές του Precision ήταν χαμηλές ανάλογα.

Συμπεράσματα

Σύμφωνα με τα παραπάνω αποτελέσματα, φαίνεται ότι το καλύτερο δυνατό μοντέλο ταξινόμησης είναι το Random Forest. Έχει τις υψηλότερες τιμές σχεδόν σε όλες τις μετρικές (με εξαίρεση το False Discovery Rate) που αποδεικνύει ότι μπορεί να ξεχωρίζει

με μεγάλη ακρίβεια τις δύο κλάσεις και να εντοπίζει επιτυχώς τις χρεοκοπημένες εταιρείες. Μάλιστα, πληροί ακόμα και τους δύο περιορισμούς απόδοσης που δίνονται στην εργασία:

1. Βρίσκει με ποσοστό επιτυχίας τουλάχιστον 60% τις εταιρείες που θα πτωχεύσουν (συγκεκριμένα, περίπου 87% επιτυχία όπως φαίνεται από το Recall).
2. Βρίσκει με ποσοστό επιτυχίας τουλάχιστον 70% τις εταιρείες που δεν θα πτωχεύσουν (συγκεκριμένα, περίπου 93% επιτυχία όπως φαίνεται από το Specificity).

Τους περιορισμούς αυτούς, τους πληρούν ακόμα δύο μοντέλα: το Decision Tree (με 86% Recall και 83% Specificity) και το Gradient Boosting (με 75% Recall και 91% Specificity). Ωστόσο, το Random Forest παραμένει το δυνατότερο μοντέλο μεταξύ τους καθώς υπερισχύει και από άποψη Recall αλλά και από άποψη Specificity.

Σχετική Βιβλιογραφία

GeeksforGeeks. (2025a, July 15). *Stratified K Fold Cross Validation*. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/stratified-k-fold-cross-validation/

GeeksforGeeks. (2025b, September 12). *K Fold Cross Validation in Machine Learning*. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/k-fold-cross-validation-in-machine-learning/

GeeksforGeeks. (2025c, December 3). *Gradient boosting in ML*. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/ml-gradient-boosting/

GeeksforGeeks. (2025d, December 17). *AUC ROC curve in machine learning*. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/auc-roc-curve/

Wikipedia contributors. (2025, December 15). *Confusion matrix*. Wikipedia. https://en.wikipedia.org/wiki/Confusion_matrix

Διαλέξεις του μαθήματος «Μηχανική Μάθηση».

Χρησιμοποιήθηκε, επίσης, το Gemini για υποστήριξη στην συγγραφή κώδικα και την κατασκευή των γραφημάτων, καθώς επίσης ως υποστηρικτικό εργαλείο κατανόησης διαφόρων εννοιών.