# Machine Learning
# Project Report

---

**Τίτλος:**
"Air Quality Prediction from environmental and demographic factors using Machine Learning Classification"

**Συντάκτες:**
Ομιλιάδης Σωκράτης (ics22054)
Καραμάνης Απόστολος (ics22009)

**Μάθημα:**
Μηχανική Μάθηση, 7$^{ο}$ Εξάμηνο

**Collab Link:** https://colab.research.google.com/drive/1dCB_Z40PBRcX24Ug3m4QVo8mhw9yFgt-?usp=sharing

# Περιεχόμενα

# 1. Πληροφορίες Προβλήματος

Η ανάλυση της ποιότητας του ατμοσφαιρικού αέρα συνιστά ζήτημα υψίστης σημασίας για τη δημόσια υγεία και την περιβαλλοντική διαχείριση στη σύγχρονη εποχή. Η παρούσα μελέτη επικεντρώνεται στη διερεύνηση των συσχετίσεων μεταξύ περιβαλλοντικών και δημογραφικών παραμέτρων αναφορικά με την ποιότητα του ατμοσφαιρικού αέρα, αξιοποιώντας ένα εκτενές σύνολο δεδομένων. Συγκεκριμένα, το σύνολο δεδομένων περιλαμβάνει μετεωρολογικές παραμέτρους (θερμοκρασία, σχετική υγρασία), συγκεντρώσεις ατμοσφαιρικών ρύπων (αιωρούμενα σωματίδια PM2.5 και PM10, διοξείδιο του αζώτου NO2, διοξείδιο του θείου SO2, μονοξείδιο του άνθρακα CO), καθώς και χωροταξικά χαρακτηριστικά (εγγύτητα σε βιομηχανικές ζώνες και πληθυσμιακή πυκνότητα).

Η μεθοδολογική προσέγγιση της έρευνας βασίζεται στην εφαρμογή τεχνικών ταξινόμησης (Classification), με απώτερο στόχο την ανάπτυξη ενός αξιόπιστου μοντέλου πρόβλεψης της ποιότητας του ατμοσφαιρικού αέρα. Το προτεινόμενο μοντέλο αναμένεται να αποτελέσει ένα χρήσιμο εργαλείο για την κατανόηση, την παρακολούθηση και την αποτελεσματική διαχείριση της ατμοσφαιρικής ρύπανσης σε αστικά περιβάλλοντα.

Η διεξαγωγή της παρούσας έρευνας κρίνεται ιδιαίτερα επίκαιρη και αναγκαία, καθώς η ατμοσφαιρική ρύπανση αποτελεί μία από τις σημαντικότερες περιβαλλοντικές προκλήσεις του 21ου αιώνα. Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας, η έκθεση σε ατμοσφαιρικούς ρύπους συνδέεται άμεσα με την εμφάνιση καρδιαγγειακών και αναπνευστικών παθήσεων, ενώ εκτιμάται ότι προκαλεί περίπου 7 εκατομμύρια πρόωρους θανάτους ετησίως παγκοσμίως. Η ανάπτυξη ενός αποτελεσματικού μοντέλου ταξινόμησης και πρόβλεψης της ποιότητας του αέρα θα συμβάλλει καθοριστικά στην έγκαιρη λήψη προληπτικών μέτρων από τις αρμόδιες αρχές, στη βελτίωση των συστημάτων προειδοποίησης του πληθυσμού, καθώς και στον αποτελεσματικότερο σχεδιασμό περιβαλλοντικών πολιτικών για την προστασία της δημόσιας υγείας.

# 2. Σχετικές Έρευνες

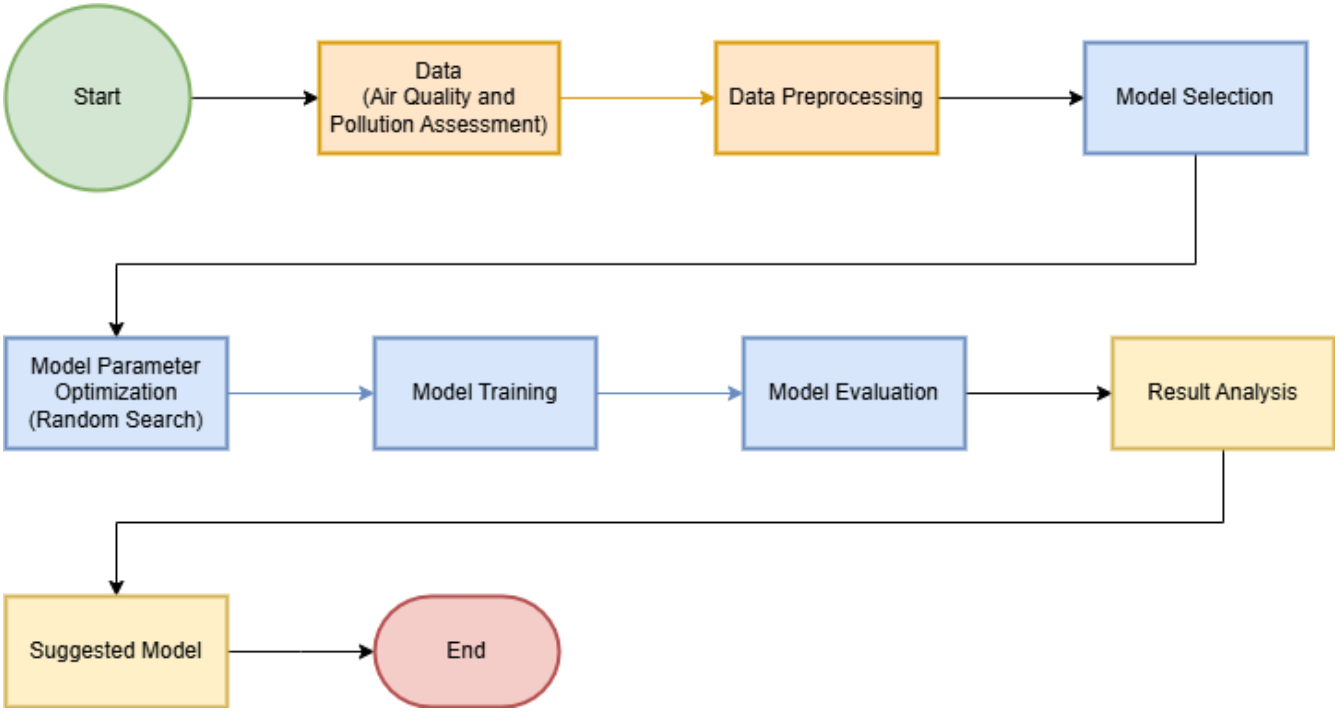
Η μελέτη των Mahesh Chhagan Patil και Snehal Mahesh Katke, από το S.E.S. Polytechnic του Solapur στην Ινδία, εξετάζει την ταξινόμηση και την πρόβλεψη της ποιότητας του αέρα μέσω της χρήσης δεδομένων από πολλαπλές πηγές περιβαλλοντικής πληροφορίας. Στόχος της έρευνας είναι η βελτίωση της ακρίβειας και της έγκαιρης πρόβλεψης της ποιότητας του αέρα. Το dataset που χρησιμοποιήθηκε περιλαμβάνει πληροφορίες για τα επίπεδα ρύπων, τις μετεωρολογικές συνθήκες και τα κυκλοφοριακά πρότυπα, τα οποία ενσωματώθηκαν και αναλύθηκαν με τη χρήση αλγορίθμων μηχανικής μάθησης. Η έρευνα στοχεύει στην ταξινόμηση της ποιότητας του αέρα σε διακριτά επίπεδα ρύπανσης και στην πρόβλεψη μελλοντικών τάσεων για την καλύτερη κατανόηση και διαχείριση των επιπτώσεων της ατμοσφαιρικής ρύπανσης.

Αξιολογήθηκαν τέσσερα μοντέλα μηχανικής μάθησης—XGBoost, GradientBoosting, CatBoost και LightGBM—για την αποτελεσματικότητά τους στην πρόβλεψη της ποιότητας του αέρα σε αστικά και αγροτικά περιβάλλοντα. Μεταξύ αυτών, το XGBoost αναδείχθηκε ως το πιο ακριβές μοντέλο με ακρίβεια 95,70%, ακολουθούμενο στενά από τα GradientBoosting και CatBoost, που είχαν ακρίβεια 95,60%, ενώ το LightGBM πέτυχε ακρίβεια 94,00%.

Τα ευρήματα παρέχουν κρίσιμες πληροφορίες για τη βελτίωση της διαχείρισης της ποιότητας του αέρα. Αυτές οι πληροφορίες στοχεύουν να υποστηρίξουν τους υπεύθυνους χάραξης πολιτικής και τους περιβαλλοντικούς φορείς στην ανάπτυξη στρατηγικών για τη μείωση των κινδύνων για την υγεία που συνδέονται με την ατμοσφαιρική ρύπανση. Η έρευνα υπογραμμίζει επίσης τη σημασία της αξιοποίησης προηγμένων τεχνικών μηχανικής μάθησης για την αντιμετώπιση περιβαλλοντικών προκλήσεων και την εξασφάλιση καλύτερων αποτελεσμάτων για τη δημόσια υγεία.

# 3. Προτεινόμενη Υλοποίηση

Η εργασία αφορά στην ταξινόμηση του ατμοσφαιρικού αέρα ως Good, Moderate, Poor, Hazardous. Το πρόβλημα αφορά στην ανάπτυξη κατάλληλων μοντέλων ταξινόμησης, τα οποία δέχονται ως πληροφορία περιβαλλοντικούς και δημογραφικούς παράγοντες. Η μεθοδολογία που ακολουθήθηκε απεικονίζεται από το παρακάτω διάγραμμα ροής




Παρακάτω απαριθμούνται τα μοντέλα τα οποία χρησιμοποιήθηκαν:

### 1. Logistic Regression

Ένα γραμμικό μοντέλο που χρησιμοποιείται για δυαδική και πολυκατηγορική ταξινόμηση. Υπολογίζει την πιθανότητα ένα δείγμα να ανήκει σε μια συγκεκριμένη κατηγορία χρησιμοποιώντας τη λογιστική συνάρτηση. Είναι απλό, κατανοητό και αποτελεσματικό για δεδομένα που διαχωρίζονται γραμμικά.

### 2. K-Nearest Neighbors (KNN)

Ένας μη παραμετρικός αλγόριθμος που ταξινομεί δεδομένα με βάση την πλειοψηφία των κατηγοριών των **k** πιο κοντινών γειτόνων. Υπολογίζει αποστάσεις

(π.χ., Ευκλείδεια) για να βρει τους κοντινότερους γείτονες. Είναι διαισθητικός αλλά απαιτητικός υπολογιστικά για μεγάλα σύνολα δεδομένων.

### 3. Random Forest

Μια μέθοδος συνόλου που δημιουργεί πολλαπλά δέντρα αποφάσεων κατά τη διάρκεια της εκπαίδευσης και συνδυάζει τα αποτελέσματά τους (πλειοψηφία για την ταξινόμηση) για να βελτιώσει την απόδοση και να μειώσει την υπερπροσαρμογή. Είναι αξιόπιστη, ευέλικτη και αποδίδει καλά σε ποικίλα σύνολα δεδομένων.

### 4. Gradient Boosting

Μια τεχνική συνόλου που δημιουργεί μοντέλα διαδοχικά, με κάθε νέο μοντέλο να διορθώνει τα λάθη του προηγούμενου. Χρησιμοποιεί δέντρα αποφάσεων ως βασικούς εκπαιδευτές και βελτιστοποιεί μια συνάρτηση απώλειας για μεγαλύτερη ακρίβεια. Είναι ιδιαίτερα αποτελεσματικό, αλλά μπορεί να υπερπροσαρμόζεται αν δεν ρυθμιστεί σωστά.

### 5. Support Vector Classification (SVC)

Ένα μοντέλο που βρίσκει το βέλτιστο υπερεπίπεδο για να διαχωρίσει τα δεδομένα σε διαφορετικές κατηγορίες. Χρησιμοποιεί συναρτήσεις πυρήνα για να διαχειριστεί μη γραμμικά όρια και είναι ιδιαίτερα αποτελεσματικό σε δεδομένα με πολλές διαστάσεις, αλλά μπορεί να είναι υπολογιστικά απαιτητικό.

### 6. XGBoost

Μια βελτιστοποιημένη υλοποίηση του **gradient boosting**, που είναι ταχύτερη και πιο αποδοτική χάρη σε τεχνικές όπως η παράλληλη επεξεργασία, η κανονικοποίηση και το κλάδεμα των δέντρων. Το XGBoost χρησιμοποιείται ευρέως σε διαγωνισμούς μηχανικής μάθησης για την ταχύτητα και την ακρίβειά του.

# 4. Πειραματικά Αποτελέσματα

## Περίληψη και Περιγραφή Δεδομένων - Dataset

Παρακάτω περιγράφονται τα βασικά χαρακτηριστικά του dataset, ενώ στην συνέχεια αναλύεται η χρήση και η σημασία αυτών των χαρακτηριστικών.
Το σύνολο δεδομένων που χρησιμοποιείται στην παρούσα μελέτη αποτελείται από **5000** καταγραφές, συμπεριλαμβάνοντας εννέα μεταβλητές εισόδου και μία μεταβλητή-στόχο.

Οι μετεωρολογικές παράμετροι διαδραματίζουν καθοριστικό ρόλο στη διασπορά και συγκέντρωση των ατμοσφαιρικών ρύπων. Η θερμοκρασία, μετρούμενη σε βαθμούς Κελσίου (°C), επηρεάζει τις χημικές αντιδράσεις των ρύπων στην ατμόσφαιρα και τη δημιουργία δευτερογενών ρύπων, ενώ η σχετική υγρασία (%) συμβάλλει στο σχηματισμό και τη μεταφορά των αιωρούμενων σωματιδίων.

Η **ατμοσφαιρική ρύπανση** χαρακτηρίζεται από πέντε βασικούς ρύπους: τα αιωρούμενα σωματίδια PM2.5 και PM10, των οποίων οι συγκεντρώσεις μετρώνται σε μικρογραμμάρια ανά κυβικό μέτρο (μg/m³). Τα σωματίδια PM2.5 είναι ιδιαίτερα επικίνδυνα καθώς μπορούν να εισχωρήσουν βαθιά στους πνεύμονες και την κυκλοφορία του αίματος. Οι αέριοι ρύποι περιλαμβάνουν το διοξείδιο του αζώτου (NO2), το διοξείδιο του θείου (SO2) και το μονοξείδιο του άνθρακα (CO). Το NO2 και το SO2, μετρούμενα σε μέρη ανά δισεκατομμύριο (ppb), προέρχονται κυρίως από την καύση ορυκτών καυσίμων και βιομηχανικές δραστηριότητες, ενώ το CO, μετρούμενο σε μέρη ανά εκατομμύριο (ppm), εκπέμπεται κυρίως από οχήματα και βιομηχανικές διεργασίες.

Οι **χωροταξικές παράμετροι** παρέχουν σημαντικές πληροφορίες για τις πηγές ρύπανσης και την έκθεση του πληθυσμού. Η απόσταση από βιομηχανικές περιοχές, μετρούμενη σε χιλιόμετρα (km), αποτελεί δείκτη της πιθανής επίδρασης βιομηχανικών εκπομπών στην τοπική ποιότητα του αέρα. Η πληθυσμιακή πυκνότητα, εκφρασμένη σε κατοίκους ανά τετραγωνικό χιλιόμετρο (άτομα/km²), σχετίζεται με την ένταση των ανθρωπογενών δραστηριοτήτων και, κατά συνέπεια,

με τις εκπομπές ρύπων από πηγές όπως η κυκλοφορία οχημάτων και τα συστήματα θέρμανσης.

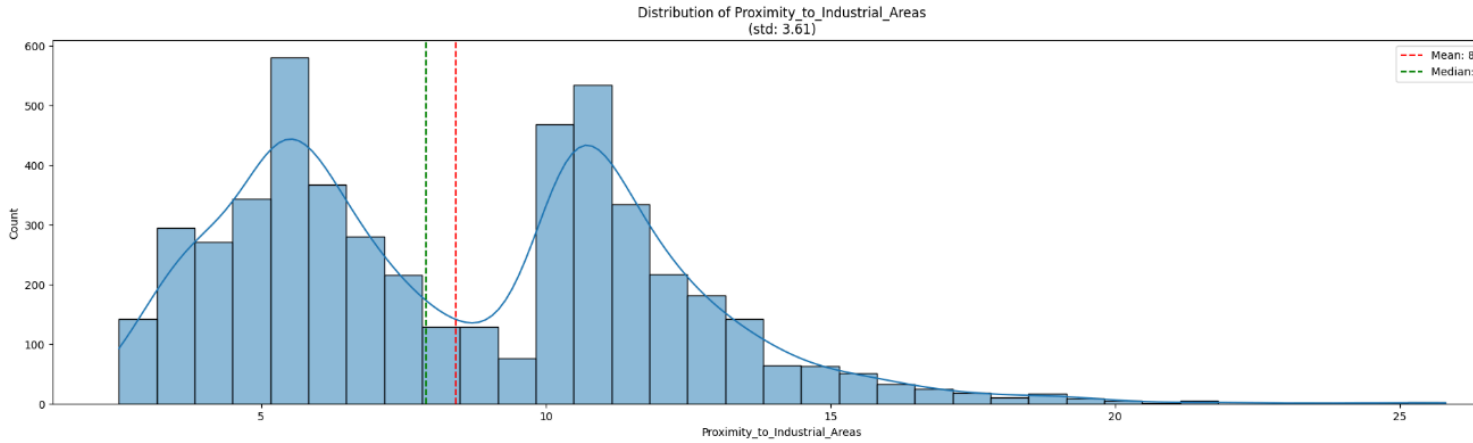
Η **μεταβλητή-στόχος**, που αντιπροσωπεύει τα επίπεδα ποιότητας του αέρα, κατηγοριοποιείται σε τέσσερις διακριτές κλάσεις: "Καλή" (Good), υποδηλώνοντας καθαρό αέρα με χαμηλά επίπεδα ρύπανσης, "Μέτρια" (Moderate), αντιπροσωπεύοντας αποδεκτή ποιότητα αέρα με την παρουσία ορισμένων ρύπων, "Κακή" (Poor), υποδεικνύοντας αξιοσημείωτη ρύπανση που ενδέχεται να προκαλέσει προβλήματα υγείας σε ευαίσθητες ομάδες του πληθυσμού, και "Επικίνδυνη" (Hazardous), χαρακτηρίζοντας ιδιαίτερα μολυσμένο αέρα που ενέχει σοβαρούς κινδύνους για τη δημόσια υγεία.

Η συνδυαστική ανάλυση αυτών των παραμέτρων επιτρέπει την ολοκληρωμένη κατανόηση των παραγόντων που επηρεάζουν την ποιότητα του αέρα και τη δημιουργία ενός αξιόπιστου μοντέλου ταξινόμησης.

## Κατανομή Δεδομένων και Διαγράμματα

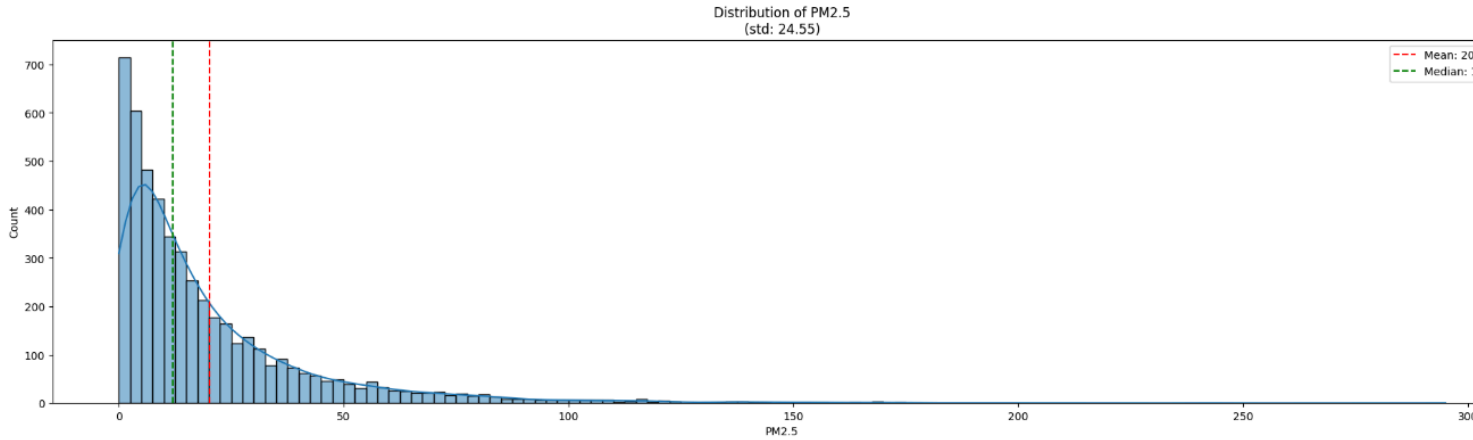
Οι παρακάτω εικόνες απεικονίζουν τις εξής κατανομές:

- Κατανομή της **απόστασης από βιομηχανικές περιοχές** σε χιλιόμετρα (km)



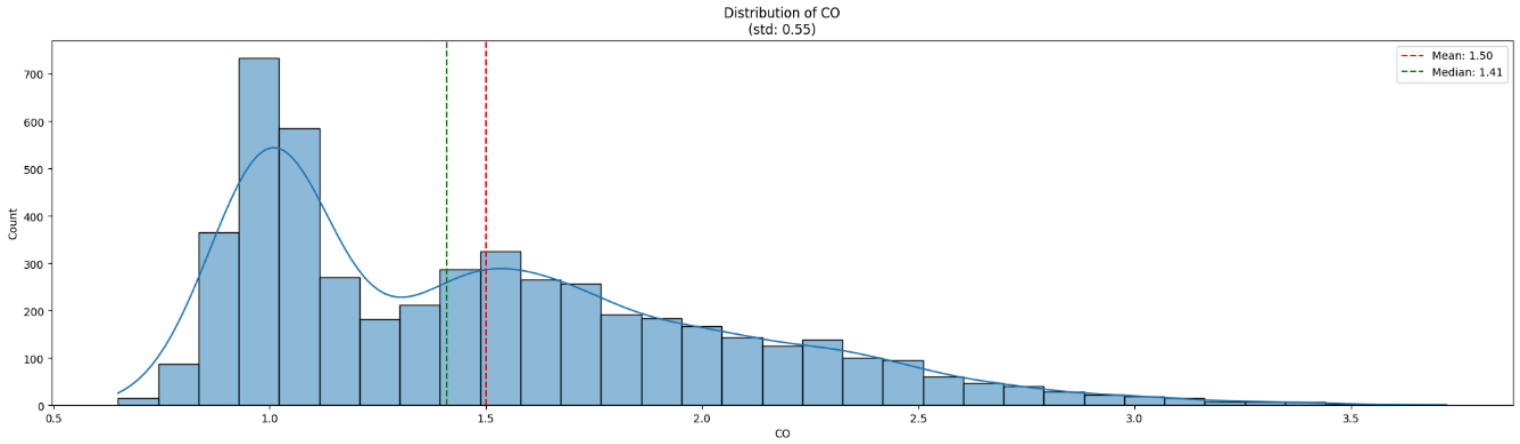

Εικόνα 4.1

- Κατανομή **σωματιδίων PM2.5** σε μικρογραμμάρια ανά κυβικό μέτρο (μg/m³)



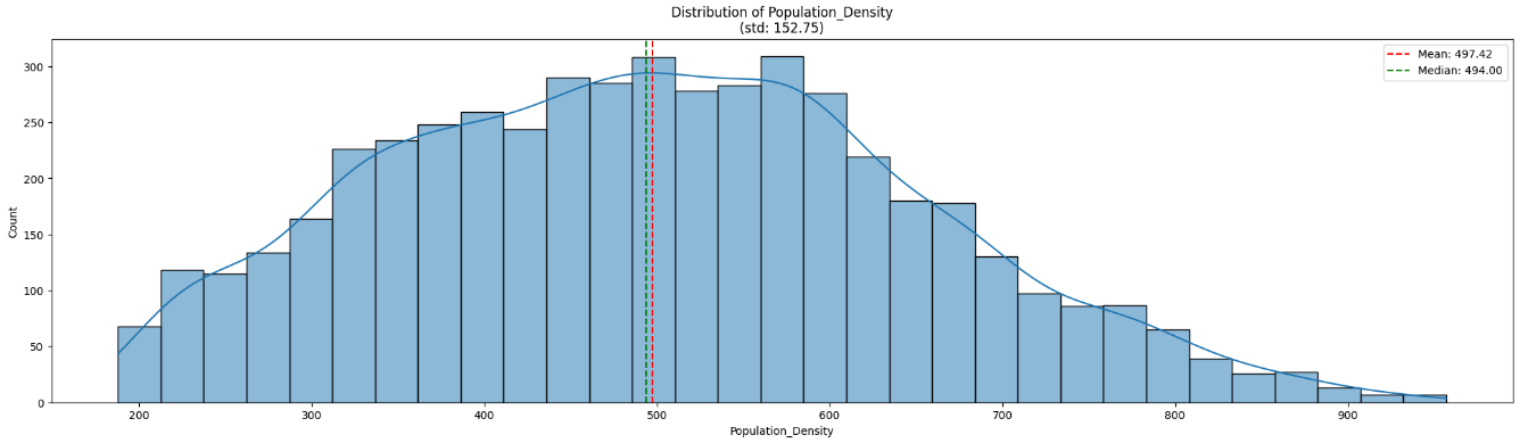

Εικόνα 4.2

- Κατανομή **μονοξείδιο του άνθρακα** (CO) σε μέρη ανά εκατομμύριο (ppm)



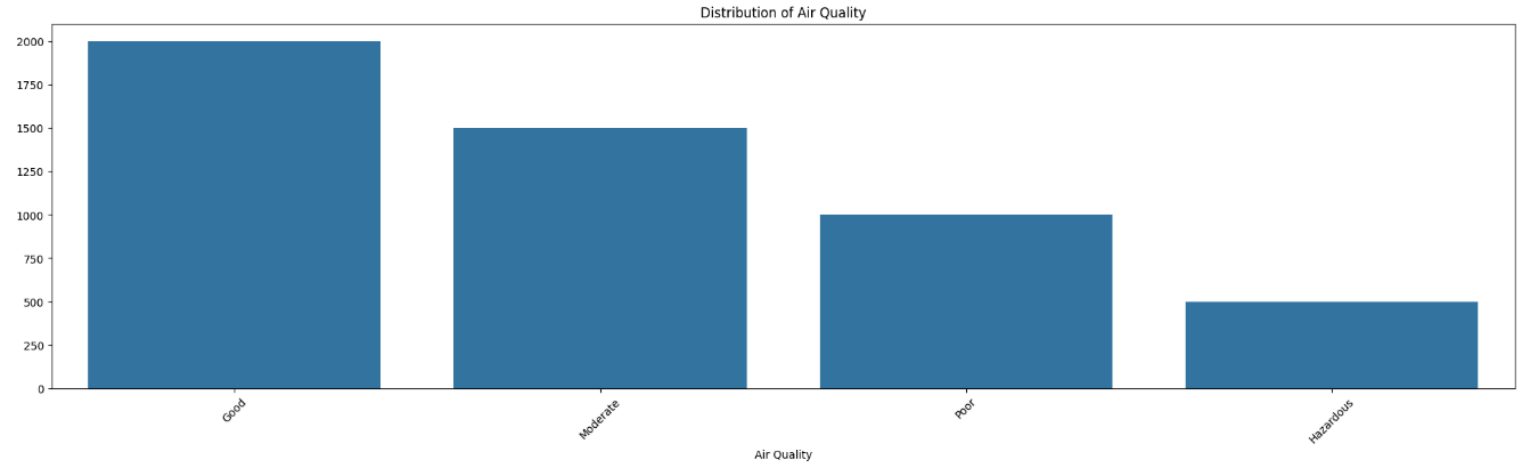

Εικόνα 4.3

- Κατανομή της **πληθυσμιακής πυκνότητας** σε κατοίκους ανά τετραγωνικό χιλιόμετρο (άτομα/km²).




Εικόνα 4.4

- Κατανομή τελικών τιμών για την **ποιότητα του αέρα**




Εικόνα 4.5

## Προεπεξεργασία

Πριν από την εκπαίδευση του μοντέλου, εφαρμόστηκαν τα ακόλουθα βήματα προεπεξεργασίας στο σύνολο δεδομένων: I

**1. Κανονικοποίηση:** Το column Air Quality κανονικοποιήθηκε με τον LabelEncoder και την μέθοδο fit_transform() για να φέρει τις τιμές στο εύρος [0,3].

## Μετρικές Επίδοσης

Στο συγκεκριμένο τμήμα της εργασίας θα περιγραφούν οι μετρικές που χρησιμοποιούμε στα πλαίσια της εργασίας για την αξιολόγηση και σύγκριση των επιμέρους προαναφερόμενων μοντέλων.

1. **Accuracy**: Η μετρική accuracy μετράει τη συχνότητα με την οποία το μοντέλο κάνει σωστές εκτιμήσεις. Υπολογίζεται ως ο λόγος των σωστών προβλέψεων του μοντέλου προς όλα τα δεδομένα του dataset. Αποτελεί την πιο απλή και κοινή μετρική με μεγάλη συχνότητα χρήσης, κυρίως σε προβλήματα classification.

2. **Precision**: Η μετρική precision αντικατοπτρίζει την ακρίβεια των true positives (TP) προβλέψεων του μοντέλου. Υπολογίζεται ως ο λόγος των TP προβλέψεων προς το πλήθος των positive προβλέψεων. Η σημαντικότητα της μετρικής precision αναδεικνύεται σε περιπτώσεις όπου οι false positives (FP) προβλέψεις έχουν μεγάλο κόστος, για παράδειγμα σε ιατρικές διαγνώσεις. Διαισθητικά το precision ποσοτικοποιεί την ικανότητα του μοντέλου να βρίσκει positive περιπτώσεις.

3. **Recall**: Η μετρική recall, ή αλλιώς sensitivity, αντικατροπτρίζει την ικανότητα του μοντέλου να αναγνωρίζει τις σχετικές περιπτώσεις ενός dataset.Υπολογίζεται ως ο λόγος των TP προβλέψεων πως το πλήθος όλων των positive περιπτώσεων στο dataset. Η χρησιμότητα του recall φαίνεται καλύτερα σε προβλήματα classification που σκοπός είναι η ελαχιστοποίηση των false negatives (FN).

4. **F1-Score**: Η μετρική f1-score παράγει μία μέτρηση συνδυάζοντας τις μετρικές precision και recall το οποίο είναι χρήσιμο σε περιπτώσεις στις

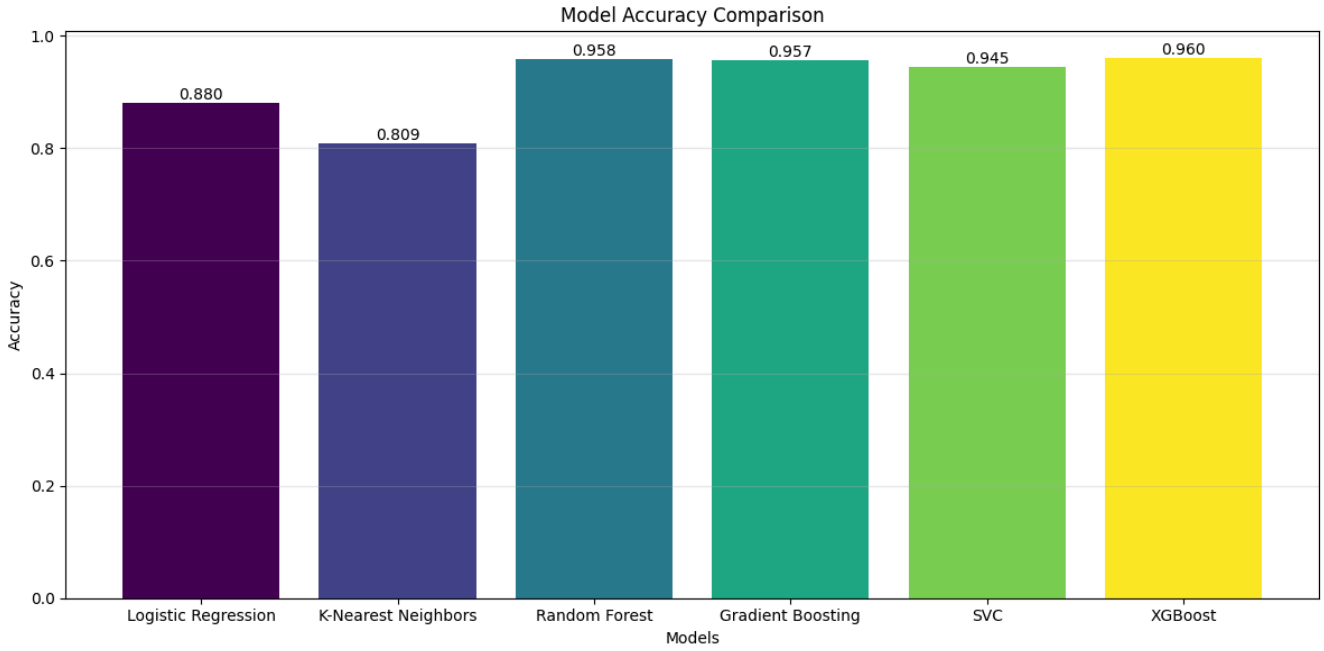
οποίες τα FP και FN είναι εξίσου σημαντικά. Υπολογίζεται ως το διπλάσιο του λόγου του γινομένου των precision και recall προς το άθροισμά τους. Η μέτρηση που παράγει κυμαίνεται μεταξύ 0 και 1, με το 1 να αντιστοιχεί στην καλύτερη επίδοση και το 0 στη χειρότερη. Το f1-score είναι μία πολύ χρήσιμη μετρική για προβλήματα classification με μη ισορροπημένο (unbalanced) dataset και παρέχει καλό μέτρο σύγκρισης μεταξύ διαφορετικών μοντέλων.

Οι μετρήσεις των παραπάνω μετρικών, μη συμπεριλαμβανομένου του accuracy, υπολογίζονται για κάθε κλάση ξεχωριστά. Για την απόδοση ακριβώς μίας μέτρησης για κάθε μετρική στο εκάστοτε μοντέλο χρησιμοποιήσαμε τις παρακάτω τεχνικές.

1. **Macro Average**: Η τεχνική macro average θεωρεί πως όλες οι κλάσεις είναι εξίσου σημαντικές μεταξύ τους, ανεξαρτήτως των συχνοτήτων τους στο dataset. Πρακτικά, υπολογίζεται η εκάστοτε μετρική για κάθε κλάση και στη συνέχεια υπολογίζεται το μέσο των μετρήσεων. Με αυτόν τον τρόπο υπολογίζεται η επίδοση του μοντέλου μη λαμβάνοντας υπόψη πιθανές διαφορές στην κατανομή των κλάσεων.
2. **Weighted Average**: Η τεχνική weighted average λαμβάνει υπόψη το support, δηλαδή το πραγματικό πλήθος περιπτώσεων, κάθε κλάσης. Το βάρος (weight) κάθε κλάσης είναι ίσο με το support της. Πρακτικά, υπολογίζεται η εκάστοτε μετρική για κάθε κλάση και στη συνέχεια υπολογίζεται ο λόγος του αθροίσματος των μετρήσεων πολλαπλασιασμένες με τα βάρη τους προς το άθροισμα όλων των βαρών. Η συγκεκριμένη τεχνική είναι χρήσιμη κυρίως σε περιπτώσεις που θεωρούμε πως η ανισότητα (imbalance) των κλάσεων επηρεάζει τη συνολική απόδοση.

## Σύγκριση και Αξιολόγηση Μοντέλων

Στο παρακάτω κεφάλαιο παρουσιάζεται μια εκτενής συγκριτική ανάλυση των υπό εξέταση μοντέλων, μέσω πολλαπλών γραφικών απεικονίσεων. Κάθε ένα από τα γραφήματα επικεντρώνεται σε διακριτό σύνολο μετρικών αξιολόγησης, παρέχοντας μια ολοκληρωμένη και συστηματική αποτίμηση της απόδοσης των μοντέλων που αποτέλεσαν αντικείμενο της ερευνητικής μας μελέτης.

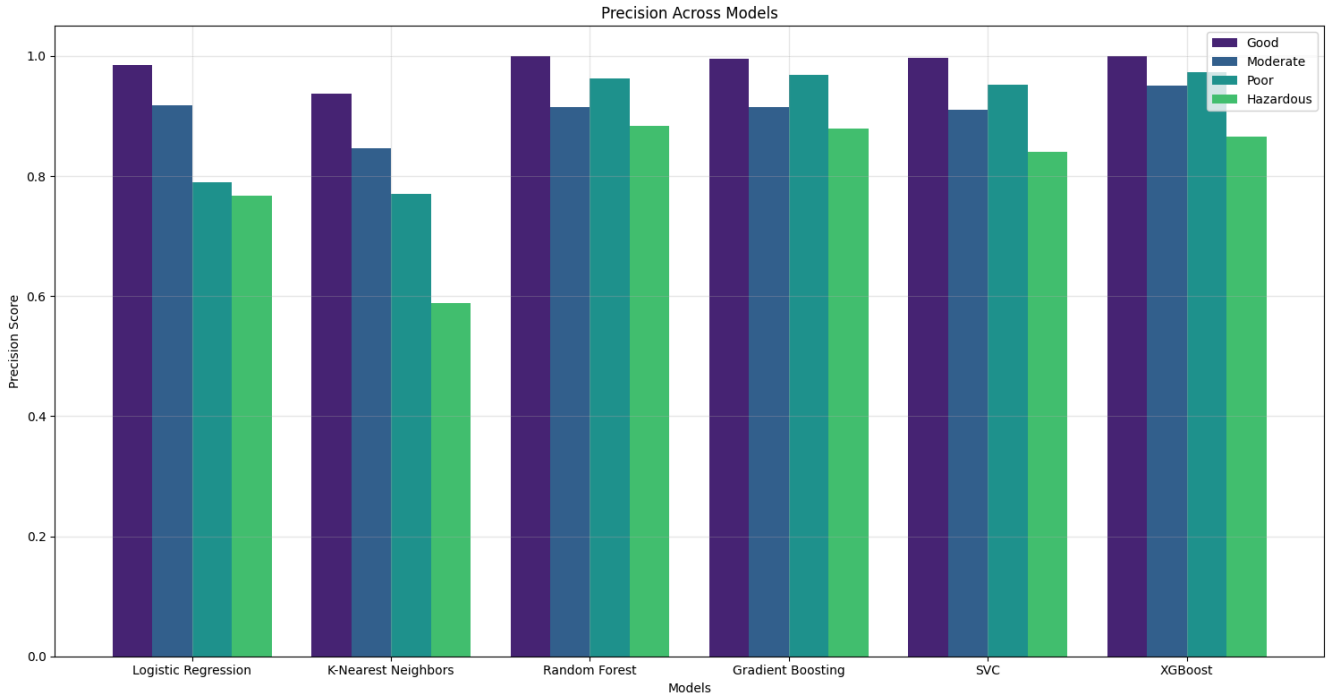



Εικόνα 4.6

Η σύγκριση των μοντέλων δείχνει τις βαθμολογίες Accuracy για τα έξι διαφορετικά μοντέλα μηχανικής μάθησης, με το XGBoost να προηγείται ελαφρώς με ακρίβεια 96,0%. Το Random Forest και το Gradient Boosting ακολουθούν πολύ κοντά με ακρίβεια 95,8% και 95,7% αντίστοιχα. Το Support Vector Classifier (SVC) αποδίδει επίσης καλά με ακρίβεια 94,5%. Υπάρχει μια αισθητή πτώση στην απόδοση με τα απλούστερα μοντέλα - το Logistic Regression επιτυγχάνει ακρίβεια 88,0%, ενώ ο αλγόριθμος K-Nearest Neighbors δείχνει τη χαμηλότερη απόδοση στο 80,9%.

Αυτή η σύγκριση υποδεικνύει ότι οι μέθοδοι XGBoost, Random Forest και Gradient Boosting είναι ιδιαίτερα κατάλληλες για αυτή τη συγκεκριμένη εργασία ταξινόμησης, επιτυγχάνοντας σταθερά ακρίβεια πάνω από 95%, ενώ οι απλούστεροι αλγόριθμοι αποδίδουν επαρκώς αλλά όχι τόσο εντυπωσιακά.

Precision Across Models

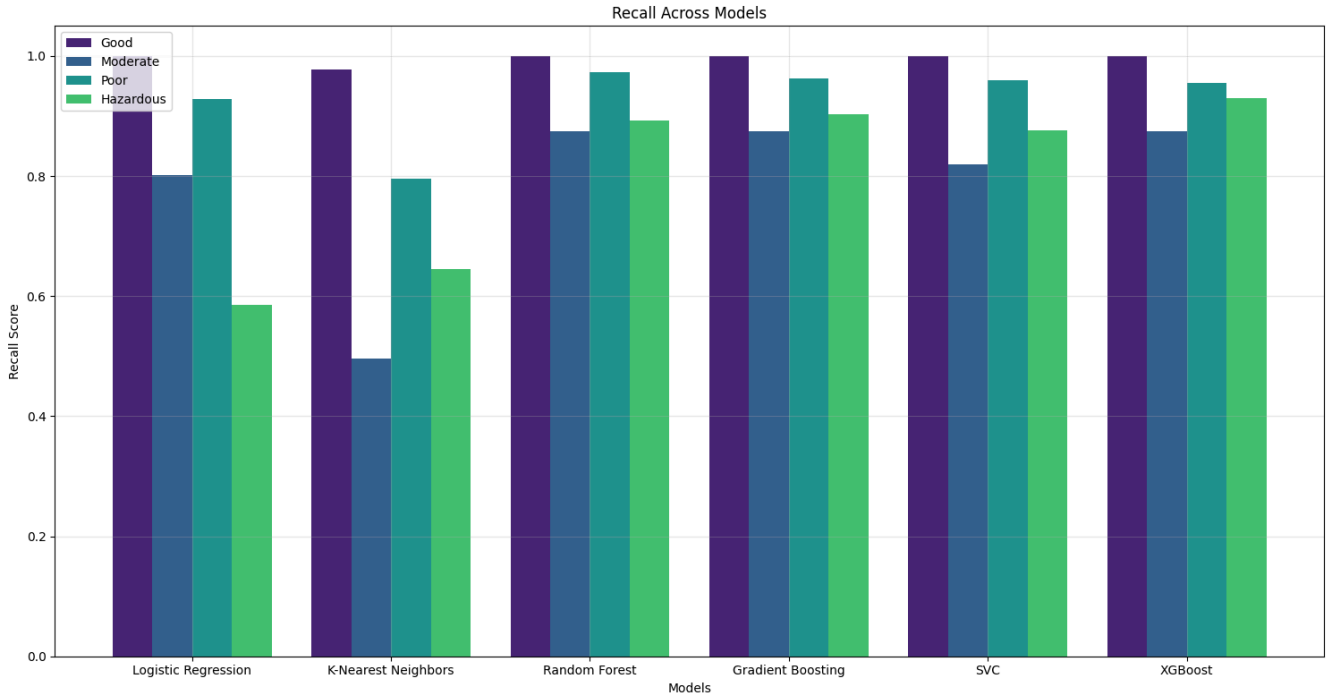



Εικόνα 4.7

Το γράφημα απεικονίζει την precision διαφόρων των μοντέλων μηχανικής μάθησης σε τέσσερις διαφορετικές κατηγορίες: Καλή (Good), Μέτρια (Moderate), Κακή (Poor) και Επικίνδυνη (Hazardous). Το XGBoost και το Random Forest παρουσιάζουν την καλύτερη απόδοση, με πολύ υψηλή ακρίβεια στην κατηγορία "Καλή" (σχεδόν 1.0) και καλή ισορροπία στις υπόλοιπες κατηγορίες. Τα μοντέλα Gradient Boosting και SVC δείχνουν παρόμοια απόδοση, ενώ το Logistic Regression και ο αλγόριθμος K-Nearest Neighbors εμφανίζουν χαμηλότερη απόδοση, ιδιαίτερα στην κατηγορία "Επικίνδυνη".

Αξιοσημείωτο είναι ότι όλα τα μοντέλα πετυχαίνουν την υψηλότερη ακρίβεια στην κατηγορία "Καλή", με σταδιακή μείωση της απόδοσης στις υπόλοιπες κατηγορίες, υποδεικνύοντας ότι είναι πιο αξιόπιστα στην αναγνώριση των καλών περιπτώσεων.

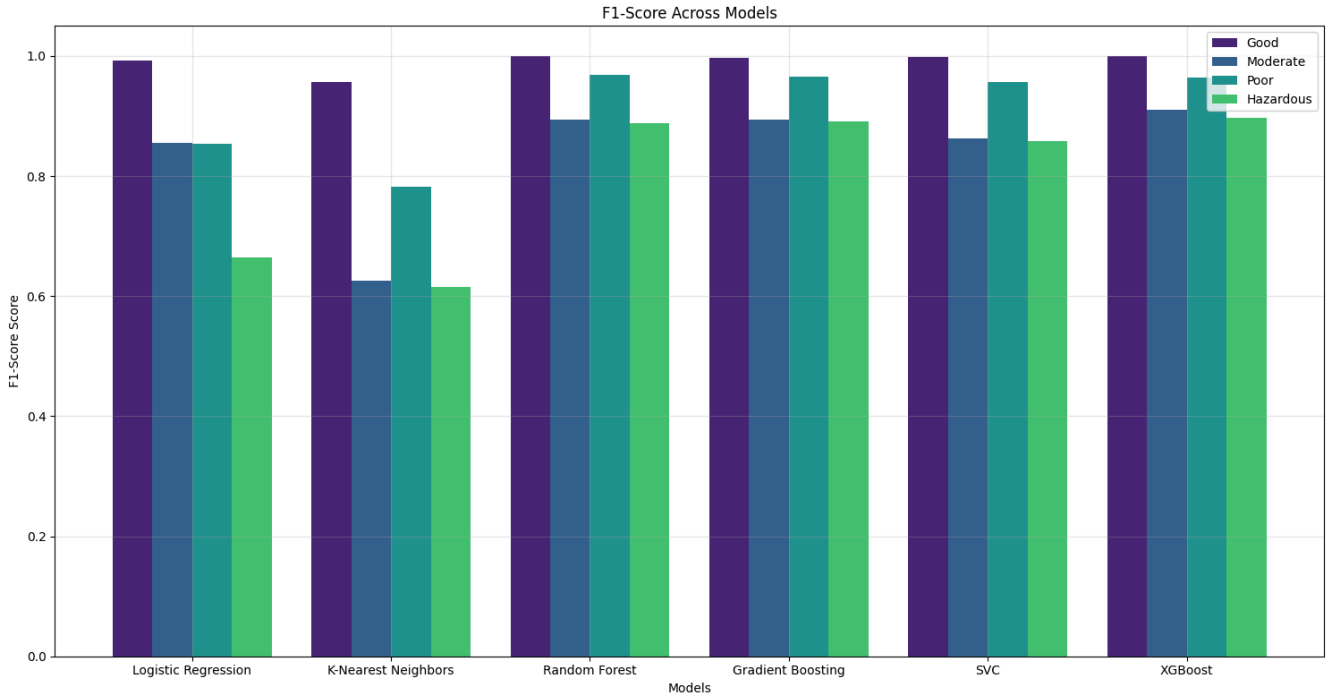
## Recall Across Models




Εικόνα 4.8

Το γράφημα παρουσιάζει το Recall διαφόρων μοντέλων μηχανικής μάθησης για τις τέσσερις κατηγορίες: Καλή (Good), Μέτρια (Moderate), Κακή (Poor) και Επικίνδυνη (Hazardous). Το XGBoost εμφανίζει την πιο ισορροπημένη απόδοση σε όλες τις κατηγορίες, με υψηλή ανάκληση τόσο στην κατηγορία "Καλή" όσο και στις υπόλοιπες κατηγορίες. Τα μοντέλα Random Forest και Gradient Boosting παρουσιάζουν παρόμοια απόδοση, με ελαφρώς χαμηλότερο recall στην κατηγορία "Moderate". Το Logistic Regression και το K-Nearest Neighbors δείχνουν τη χαμηλότερη απόδοση, ιδιαίτερα στην κατηγορία "Hazardous", με το K-Nearest Neighbors να έχει ιδιαίτερα χαμηλή ανάκληση στην κατηγορία "Moderate". Το SVC διατηρεί καλή απόδοση, αλλά με μικρή πτώση στην κατηγορία "Moderate" σε σύγκριση με τα καλύτερα μοντέλα.

## F1-Score

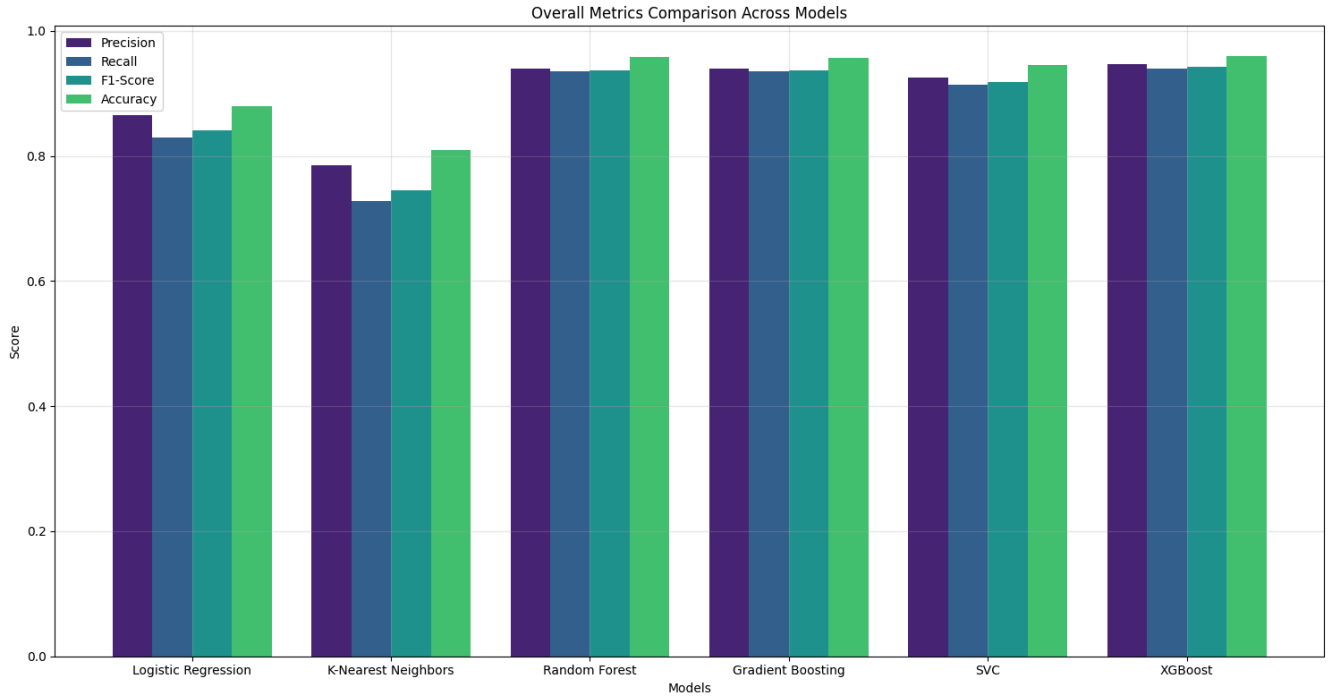



Εικόνα 4.9

Το διάγραμμα του F1-Score μας αποκαλύπτει μια ενδιαφέρουσα σύγκριση μεταξύ των μοντέλων μηχανικής μάθησης. Στην κορυφή της κατάταξης βρίσκεται το XGBoost, επιδεικνύοντας εξαιρετική σταθερότητα και υψηλές επιδόσεις σε όλες τις κατηγορίες (Good, Moderate, Poor, Hazardous). Αξιοσημείωτη είναι η επίδοση των Random Forest και Gradient Boosting, που ακολουθούν πολύ κοντά, με ιδιαίτερη επιτυχία στην αναγνώριση της κατηγορίας "Good". Το SVC δείχνει αξιοπρεπή απόδοση, αν και υπολείπεται στην κατηγορία "Moderate". Στο κάτω άκρο της κατάταξης συναντάμε το Logistic Regression και το K-Nearest Neighbors, με το τελευταίο να παρουσιάζει αξιοσημείωτη αδυναμία στην κατηγορία "Moderate".

Ένα κοινό μοτίβο που παρατηρείται σε όλα τα μοντέλα είναι η υψηλότερη επίδοσή τους στην κατηγορία "Good", υποδηλώνοντας μια γενική τάση προς την αποτελεσματικότερη αναγνώριση των θετικών περιπτώσεων.

## Overall Metrics Comparison across Models

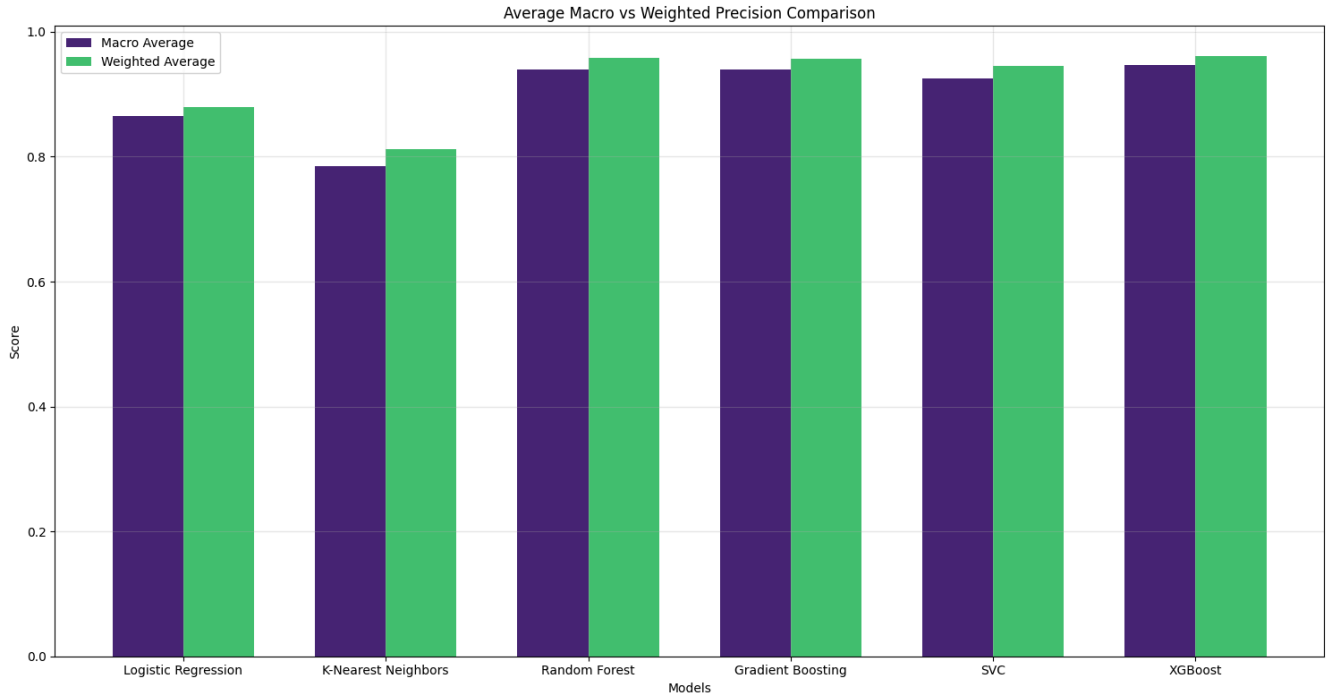



Εικόνα 4.10

Η συνολική σύγκριση των μετρικών (Precision, Recall, F1-Score και Accuracy) αποκαλύπτει μια ξεκάθαρη διαστρωμάτωση στην απόδοση των μοντέλων μηχανικής μάθησης. Στην κορυφή της ιεραρχίας βλέπουμε μια τριάδα υψηλής απόδοσης: το XGBoost, Random Forest και Gradient Boosting, με σχεδόν πανομοιότυπες επιδόσεις γύρω στο 0.95 σε όλες τις μετρικές. Το SVC ακολουθεί με ελαφρώς χαμηλότερες αλλά σταθερές επιδόσεις. Στον αντίποδα, το Logistic Regression και ιδιαίτερα το K-Nearest Neighbors παρουσιάζουν αισθητά χαμηλότερες επιδόσεις, με το τελευταίο να εμφανίζει τη μεγαλύτερη διακύμανση μεταξύ των διαφορετικών μετρικών.

Αξίζει να σημειωθεί το γεγονός  ότι σε όλα τα μοντέλα η Accuracy τείνει να είναι ελαφρώς υψηλότερη από τις υπόλοιπες μετρικές, υποδηλώνοντας μια γενικά καλή ισορροπία στην ταξινόμηση όλων των κατηγοριών.

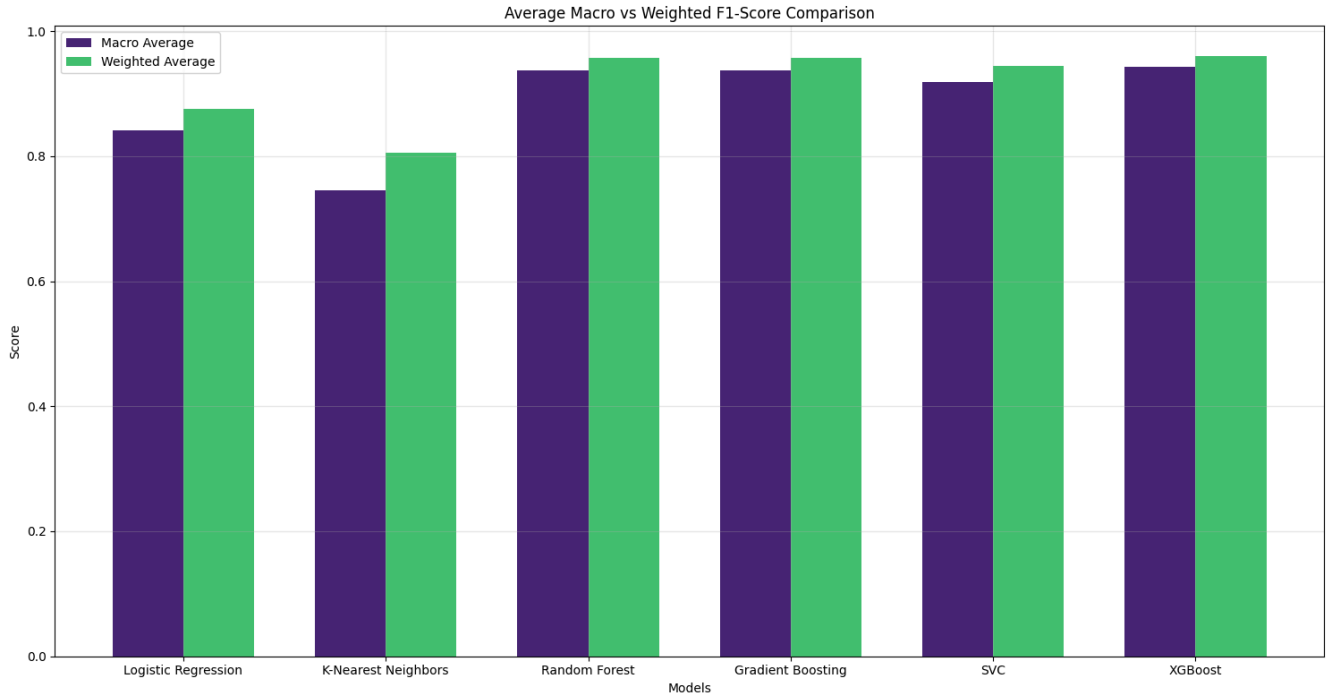
Average Macro vs Weighted Precision Comparison




Εικόνα 4.11

Σε αυτό το διάγραμμα παρατηρούμε μια ενδιαφέρουσα σύγκριση μεταξύ του Macro Average και του Weighted Average για κάθε μοντέλο. Ένα ενδιαφέρον μοτίβο είναι ότι το Weighted Average είναι συστηματικά υψηλότερο από το Macro Average σε όλα τα μοντέλα, υποδηλώνοντας την επίδραση του μεγέθους των κλάσεων στην απόδοση. Το XGBoost και το Random Forest παρουσιάζουν τις υψηλότερες τιμές και στους δύο μέσους όρους, με σκορ που πλησιάζουν το 0.95. Το Gradient Boosting ακολουθεί με σχεδόν ταυτόσημη απόδοση. Το SVC διατηρεί αξιοπρεπή επίπεδα και στους δύο μέσους όρους, ενώ το Logistic Regression και το K-Nearest Neighbors εμφανίζουν τη μεγαλύτερη διαφορά μεταξύ των δύο μετρικών, με το K-Nearest Neighbors να παρουσιάζει τη χαμηλότερη συνολική απόδοση.

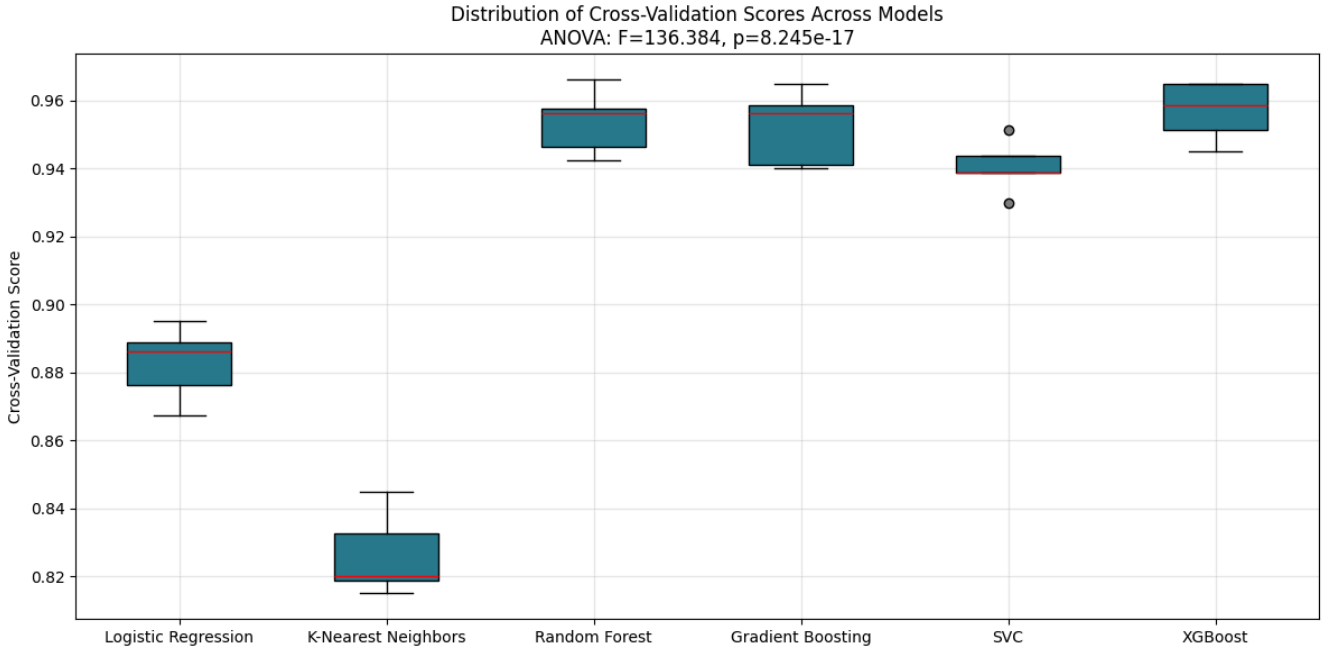
## Average Macro vs Weighted F1-Score Comparison




Εικόνα 4.11

Η σύγκριση των F1-Score μεταξύ Macro Average και Weighted Average μας προσφέρει μια διαφορετική οπτική της απόδοσης των μοντέλων. Παρατηρούμε ότι τα προηγμένα μοντέλα - XGBoost, Random Forest και Gradient Boosting - παρουσιάζουν εντυπωσιακή συνέπεια μεταξύ των δύο μετρικών, με το Weighted Average να είναι ελαφρώς υψηλότερο, κυμαινόμενο γύρω στο 0.95. Το SVC, αν και με μικρότερη απόδοση, διατηρεί μια αξιοπρεπή ισορροπία. Ιδιαίτερο ενδιαφέρον παρουσιάζει η συμπεριφορά του Logistic Regression και του K-Nearest Neighbors, όπου παρατηρείται η μεγαλύτερη απόκλιση μεταξύ των δύο μέσων όρων, υποδεικνύοντας ότι η απόδοσή τους επηρεάζεται σημαντικά από την κατανομή των κλάσεων στο σύνολο δεδομένων.

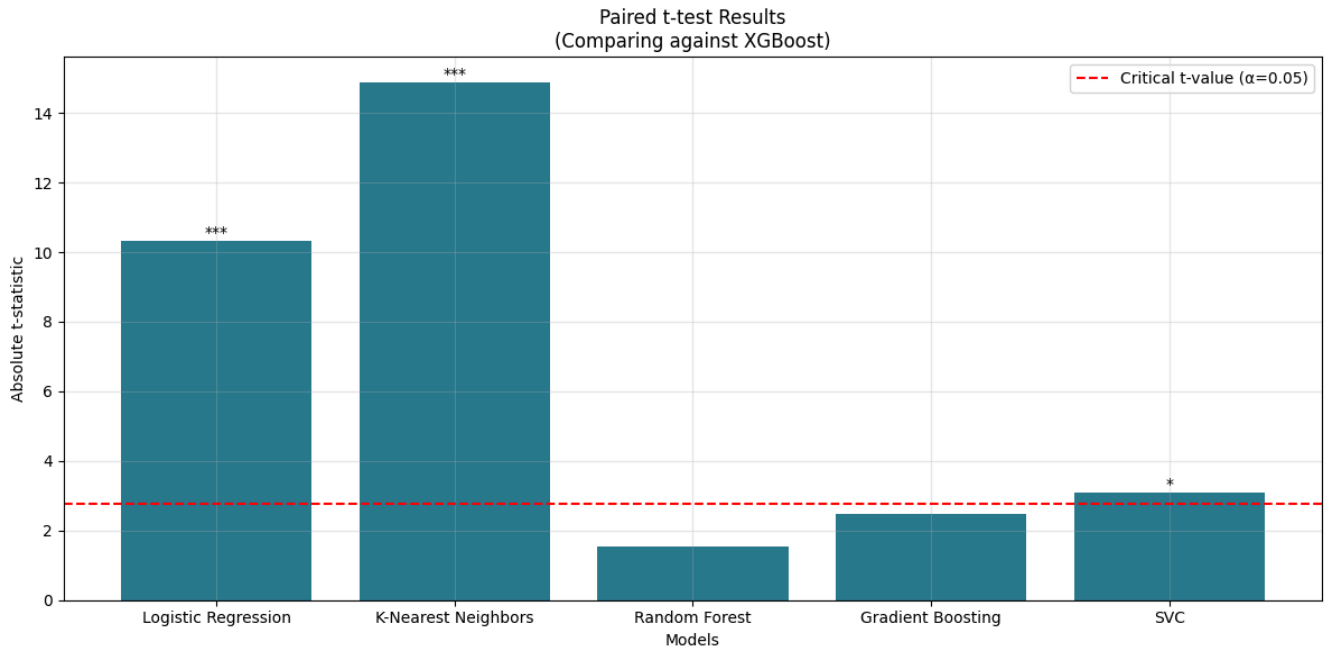
## Distribution of Cross-Validation Scores




Εικόνα 4.12

Το συγκεκριμένο διάγραμμα box plot παρουσιάζει την κατανομή των cross-validation scores για κάθε μοντέλο, με το ANOVA να δείχνει σημαντική στατιστική διαφορά μεταξύ των μοντέλων (F=136.384, p=8.245e-17). Το XGBoost εμφανίζει την υψηλότερη διάμεσο και τη μικρότερη διασπορά, με σκορ γύρω στο 0.96. Τα Random Forest και Gradient Boosting ακολουθούν με παρόμοια απόδοση γύρω στο 0.95, ενώ το SVC δείχνει ελαφρώς χαμηλότερη απόδοση με μερικές ακραίες τιμές. Τα Logistic Regression και K-Nearest Neighbors παρουσιάζουν σημαντικά χαμηλότερη απόδοση, με το K-Nearest Neighbors να έχει τη χαμηλότερη διάμεσο γύρω στο 0.82.

Paired t-test Results




Εικόνα 4.13

Το παραπάνω διάγραμμα παρουσιάζει τα αποτελέσματα των paired t-tests, συγκρίνοντας κάθε μοντέλο με το XGBoost. Η κόκκινη διακεκομμένη γραμμή δείχνει την κρίσιμη τιμή t για α=0.05. Τα Logistic Regression και K-Nearest Neighbors εμφανίζουν εξαιρετικά υψηλές τιμές t-statistic (περίπου 10 και 15 αντίστοιχα) και τρεις αστερίσκους (***), υποδεικνύοντας πολύ σημαντική στατιστική διαφορά από το XGBoost. Το SVC δείχνει μικρότερη αλλά ακόμα στατιστικά σημαντική διαφορά, ενώ τα Random Forest και Gradient Boosting έχουν τιμές t-statistic κάτω από το critical value, υποδηλώνοντας ότι δεν διαφέρουν σημαντικά από το XGBoost.

# 5. Συμπεράσματα

Τα αποτελέσματα της εργασίας μας παρέχουν μια ολοκληρωμένη ανάλυση της απόδοσης και της βελτιστοποίησης διάφορων μοντέλων ταξινόμησης. Με βάση τις μετρήσεις αξιολόγησης και τις στατιστικές αναλύσεις, προκύπτουν τα εξής συμπεράσματα:

1. **Συνολική Απόδοση των Μοντέλων:**
    - Το μοντέλο XGBoost πέτυχε το μεγαλύτερο precision (96,0%) από όλα τα μοντέλα που δοκιμάστηκαν, ακολουθούμενο από το Random Forest (95,8%) και το Gradient Boosting (95,7%).
    - Τα Logistic Regression και K-Nearest Neighbors (KNN) είχαν χαμηλότερη απόδοση, με precision 88,0% και 80,9% αντίστοιχα.
2. **Απόδοση ανά Class:**
    - Τα μοντέλα XGBoost και Random Forest παρουσίασαν εξαιρετική απόδοση σε όλες τις κατηγορίες, με σχεδόν τέλεια αποτελέσματα σε precision, recall και F1-score για την Class 0 και σταθερά υψηλές βαθμολογίες για τις υπόλοιπες κατηγορίες.
    - Το KNN αντιμετώπισε δυσκολίες με τις ανισόρροπες κατανομές κατηγοριών, παρουσιάζοντας χαμηλότερη ανάκληση για την Class 1 (50%) και σχετικά μέτρια F1-score για την Class 3 (62%).
3. **Στατιστική Σημασία:**
    - Το ANOVA τεστ επιβεβαίωσε ότι υπάρχει στατιστικά σημαντική διαφορά ($p$-value < 0.0001) στην απόδοση των μοντέλων, δείχνοντας ότι η επιλογή του μοντέλου επηρεάζει σημαντικά την ακρίβεια της ταξινόμησης.
    - Τα ζεύγη t-test έδειξαν ότι το XGBoost υπερτερεί σημαντικά των Logistic Regression, KNN και SVC ($p$-value < 0.05). Ωστόσο, δεν παρατηρήθηκε στατιστικά σημαντική διαφορά μεταξύ XGBoost και Random Forest ($p$-value = 0.199) ή Gradient Boosting ($p$-value = 0.0679).

4. **Ανθεκτικότητα Μοντέλων και Βελτιστοποίηση Παραμέτρων:**
    - Η διαδικασία ρύθμισης των παραμέτρων διαδραμάτισε καθοριστικό ρόλο στη βελτίωση της απόδοσης όλων των μοντέλων. Για παράδειγμα:
        - Το Logistic Regression πέτυχε τα καλύτερα αποτελέσματα με τον αλγόριθμο `liblinear` και L1 κανονικοποίηση.
        - Το KNN ωφελήθηκε από τη χρήση της μετρικής Manhattan και της απόστασης για τη στάθμιση.
        - Οι αλγόριθμοι συνόλου, όπως τα Random Forest, Gradient Boosting και XGBoost, πέτυχαν υψηλή ακρίβεια με προσεκτικά επιλεγμένες παραμέτρους, όπως το βάθος, ο ρυθμός εκμάθησης και η κανονικοποίηση.
5. **Σύσταση:**
    - Το XGBoost προτείνεται ως το καταλληλότερο μοντέλο για αυτό το πρόβλημα ταξινόμησης, λόγω της εξαιρετικής ακρίβειας, της σταθερής απόδοσης στις κατηγορίες και της στατιστικής υπεροχής έναντι των περισσότερων εναλλακτικών.
    - Τα Random Forest και Gradient Boosting αποτελούν επίσης ισχυρές επιλογές, ειδικά όταν η υπολογιστική αποδοτικότητα ή η ερμηνευσιμότητα είναι προτεραιότητα.

Συνοψίζοντας, το XGBoost αποδείχθηκε το μοντέλο με την καλύτερη απόδοση σε αυτή τη μελέτη, επιτυγχάνοντας έναν αποτελεσματικό συνδυασμό precision, recall και F1-score σε όλες τις κατηγορίες, ενώ διατηρεί στατιστική υπεροχή έναντι των περισσότερων άλλων μοντέλων.