Εργασία 2 – Δέντρα Απόφασης & Τυχαία Δάση

Χημική εταιρεία θέλει να κατασκευάσει φορητή συσκευή ανάλυσης νερού, που μέσω χημικών αναλύσεων θα εξάγει διάφορα χαρακτηριστικά από το νερό και θα αποφασίζει αν το νερό είναι πόσιμο. Διαθέτετε στη διάθεση σας 3276 παραδείγματα πόσιμου και μη πόσιμου νερού που μπορείτε να κατεβάσετε εδώ: https://www.kaggle.com/datasets/adityakadiwal/water-potability

Ερωτήματα:

1. Να φορτώσετε τα δεδομένα σε DataFrame. Στη συνέχεια, να τα περιγράψετε (describe) και να δημιουργήσετε το ιστόγραμμα για κάθε χαρακτηριστικό, καθώς και ραβδόγραμμα για τη μεταβλητή Potability. Να αναφέρετε το πλήθος των ελλειπών τιμών για κάθε χαρακτηριστικό και τα ποσοστά πόσιμου και μη-πόσιμου νερού των παραδειγμάτων. Θεωρείτε η ποιότητα των δεδομένων είναι ικανοποιητική? Απαντήστε.
2. Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας (ΠΟΥ), τα προτεινόμενα επίπεδα pH του νερού είναι 6.5 και 8.5. Υπολογίστε τα ποσοστά πόσιμου και μη-πόσιμου νερού των παραδειγμάτων για α) $pH < 6.5$, β) $6.5 \leq pH \leq 8.5$ και η) $8.5 < pH$. Σε τι βαθμό επαληθεύεται η ιδιότητα αυτή στα δεδομένα σας?
3. Σύμφωνα με τον ΠΟΥ, τα προτεινόμενα επίπεδα χλωραμίνης είναι ως 4 ppm. Δημιουργήστε διάγραμμα διασποράς (scatter plot) μεταξύ x: pH και y: Chloramine, στο οποίο θα χρωματίσετε τα πόσιμα παραδείγματα με μπλε και τα μη-πόσιμα με κόκκινο. Τι διαπιστώνετε για τη διαχωρισιμότητα των παραδειγμάτων?
4. Συμπληρώστε τις ελλειπής τιμές (αν υπάρχουν). Μπορείτε να χρησιμοποιήσετε τη συνάρτηση fillna() του pandas. Για απλότητα, μπορείτε να χρησιμοποιήσετε κάποια σταθερά (πχ df['x'] = df['x'].fillna(c) όπου c 0 ή -1 αν δεν υπάρχουν άλλες τέτοιες τιμές στη στήλη). Εναλλακτικά μπορείτε να συμπληρώσετε με τη μέση τιμή (πχ df['x'] = df['x'].fillna(df['x'].mean()). **Το βήμα αυτό είναι απαραίτητο καθώς δε γίνεται να προχωρήσετε στην εκπαίδευση του δέντρου (η scikit-learn δε χειρίζεται αυτόματα τις ελλειπής τιμές στα δέντρα).**
5. Δημιουργήστε numpy arrays με κατάλληλα inputs (x) και targets (y), όπου target το potability. Χωρίστε τα δεδομένα σε train-test σε ποσοστό 70-30 αντίστοιχα με 0 seed.
6. Εκπαιδεύστε ταξινομητή Δέντρο Απόφασης (Decision Tree) στο train set και ύστερα μετρήστε την ακρίβεια του (accuracy) στα train, test σετ. Συμβουλευτείτε το documentation: https://scikit-learn.org/dev/modules/generated/sklearn.tree.DecisionTreeClassifier.html
7. Επαναλάβετε το ερώτημα 5, δοκιμάζοντας τους παρακάτω συνδυασμούς: criterion (gini, entropy), max-depth (None, 3, 5), min-samples-split (2, 5), min-samples-leaf (1, 2), max-features (None, sqrt), cost-complexity-pruning (0, 0.01). Δημιουργήστε πινακάκι ακρίβειας σε train-test για κάθε συνδυασμό με pandas, όπου οι στήλες θα είναι οι τιμές κάθε χαρακτηριστικό, καθώς και train acc, test acc.
8. Επιλέξτε τον συνδυασμό με τη μεγαλύτερη ακρίβεια στο test όταν max-depth = 3. Στη συνέχεια, συνέχεια, εκπαιδεύστε πάλι το δέντρο αυτό και απεικονίστε τη δομή του https://scikit-learn.org/1.5/modules/generated/sklearn.tree.plot_tree.html. Να περιγράψετε τους κανόνες που εξήγαγε το δέντρο.
9. Επιλέξτε το συνδυασμό με τη μεγαλύτερη ακρίβεια στο test, επανεκπαιδεύστε το δέντρο και δημιουργήστε ραβδόγραμμα με τη σημαντικότητα κάθε χαρακτηριστικού (feature

importance) σύμφωνα με το δέντρο αυτό. Αν η συσκευή μπορεί να υποστηρίζει μέχρι 5 χημικές αναλύσεις, ποια χαρακτηριστικά θα έπρεπε να εξάγει από το νερό? Αιτιολογήστε.

10. Επιλέξτε το συνδυασμό με τη μεγαλύτερη ακρίβεια στο test, επανεκπαιδεύστε το δέντρο και δημιουργήστε ραβδόγραμμα με τη σημαντικότητα κάθε χαρακτηριστικού (feature importance) σύμφωνα με το δέντρο αυτό. Αν η συσκευή μπορεί να υποστηρίζει μέχρι 5 χημικές αναλύσεις, ποια χαρακτηριστικά θα έπρεπε να εξάγει από το νερό? Αιτιολογήστε.

11. Εξηγήστε τους λόγους για τους οποίους ένα Τυχαίο Δάσος (Random Forest) ενδεχομένως να πετύχαινε καλύτερη ακρίβεια από το Δέντρο Απόφασης.
12. Επαναλάβετε το ερώτημα 6 χρησιμοποιώντας Random Forest https://scikit-learn.org/1.5/modules/generated/sklearn.ensemble.RandomForestClassifier.html . Ορίστε το 0 ως seed. Επιπλέον, στους συνδυασμούς να προστεθεί και το πλήθος των δέντρων (n_estimators) για 50, 100 και 200 δέντρα.
13. Τι είναι πιο σημαντικό για το μοντέλο, να προβλέπει καλά το πόσιμο νερό, αλλά χάνοντας ακρίβεια από το μη-πόσιμο νερό ή να προβλέπει καλύτερα το μη-πόσιμο νερό, χάνοντας ακρίβεια από το πόσιμο? Αιτιολογήστε.
14. Τι είναι νομικά ασφαλέστερο για την εταιρία, η χρήση του καλύτερου δέντρου ή του καλύτερου τυχαίου δάσους? Αιτιολογήστε.

Οδηγίες

- **Χρησιμοποιείστε την πλατφόρμα Google Colab για την υλοποίηση της άσκησης.**
- **Τα plots/πινακάκια να εμφανίζονται επάνω στο Colab. Επίσης, μπορείτε να εισάγετε κελιά για κείμενο και πινακάκια όπως φαίνεται στον παρακάτω σύνδεσμο:**
  https://colab.research.google.com/notebooks/markdown_guide.ipynb
- **Δώστε έμφαση στην παρουσίαση της εργασίας. Copy-Paste από το ChatGPT θα αγνοούνται.**
- **Στο elearning θα υποβάλλετε το link της εργασίας σας στο Github.**