ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ - ΤΜΗΜΑ ΜΗΧΑΝΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΑΚΑΔΗΜΑΪΚΟ ΕΤΟΣ 2023-24

ΕΞΑΜΗΝΟ: 8ο

ΜΑΘΗΜΑ: ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

ΔΙΔΑΣΚΟΥΣΑ: ΣΟΦΙΑ ΠΑΝΑΓΙΩΤΙΔΟΥ, Αν. Καθηγήτρια

3η Ε Ρ Γ Α Σ Ι Α

Ονοματεπώνυμο φοιτητή : _____

Α.Ε.Μ. : _____

Θεσσαλονίκη, Μάιος 2024

3η Εργασία στο μάθημα "Ανάλυση Δεδομένων" 2023-24

Για τα ερωτήματα που έπονται να χρησιμοποιηθούν τα δεδομένα που σας δόθηκαν για τα 4 τελευταία ερωτήματα της 2ης εργασίας (και να εξαιρεθούν όσα δεδομένα είχατε εξαιρέσει και στη 2η εργασία). Επίσης, να γίνει και πάλι διαχωρισμός των δεδομένων σε train και test set με το 80% των δεδομένων να αξιοποιηθεί για την εκπαίδευση των μοντέλων (training) και το 20% για τον έλεγχο αυτών (testing).

Να γίνει ταξινόμηση των δεδομένων σε υψηλό και χαμηλό life expectancy με εφαρμογή των παρακάτω μεθόδων. Σε κάθε περίπτωση, να υπολογιστεί η ακρίβεια που επιτυγχάνεται (test accuracy) με την εκάστοτε μέθοδο.

1. Με δέντρο ταξινόμησης μέγιστου βάθους 3 (max depth).
2. Με τη μέθοδο Bagging χρησιμοποιώντας εκτιμήσεις από 200 δειγματοληψίες (n estimators).
3. Με τη μέθοδο Random Forest χρησιμοποιώντας εκτιμήσεις από 200 δειγματοληψίες (n estimators) και τυχαίο πλήθος υποψήφιων χαρακτηριστικών (features/predictors) $m=\sqrt{p}$, όπου p το σύνολο των χαρακτηριστικών του προβλήματος.
4. Με τη μέθοδο Boosting χρησιμοποιώντας εκτιμήσεις από 200 επιμέρους μοντέλα (n estimators), συντελεστή μάθησης 1.0 (learning rate) και μέγιστο βάθος 1 (max depth).
5. Να παρασταθεί γραφικά το test error της μεθόδου του ερωτήματος (1) συναρτήσει του βάθους του δέντρου (για όσες τιμές βάθους κρίνετε απαραίτητο και λογικό). Τι παρατηρείτε; Σχολιάστε.
6. Να παρασταθεί γραφικά το test error των μεθόδων των ερωτημάτων (2), (3) και (4) συναρτήσει της μεταβλητής «n estimators». Τι παρατηρείτε σε κάθε περίπτωση; Σχολιάστε.

Ποια από τις μεθόδους ταξινόμησης παρουσιάζει την καλύτερη ακρίβεια; Σχολιάστε παρουσιάζοντας τα σχετικά αριθμητικά αποτελέσματα σε έναν συνοπτικό πίνακα.