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

ΕΞΑΜΗΝΟ: 8ο

ΜΑΘΗΜΑ: ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

ΔΙΔΑΣΚΟΥΣΑ: ΣΟΦΙΑ ΠΑΝΑΓΙΩΤΙΔΟΥ, Αν. Καθηγήτρια

2η Ε Ρ Γ Α Σ Ι Α

Ονοματεπώνυμο φοιτητή : _____

Α.Ε.Μ. : _____

Θεσσαλονίκη, Απρίλιος 2024

2η Εργασία στο μάθημα "Ανάλυση Δεδομένων" 2023-24

Σε όλη την εργασία να χρησιμοποιηθούν τα δεδομένα της 1ης εργασίας ως εξής: (α) να εξαιρεθούν οι μεταβλητές Country και Status από την ανάλυση και (β) να μην χρησιμοποιηθούν τα 2 έτη που έχουν εξαιρεθεί για κάθε έναν από εσάς σύμφωνα με το excel αντιστοίχισης φοιτητή – δεδομένων. Επίσης, για το Μέρος Β να μη χρησιμοποιηθεί καθόλου η στήλη Year (εφόσον έχουν εξαιρεθεί βέβαια τα 2 έτη για τον καθέναν από εσάς).

Μέρος Α

Χρησιμοποιώντας τα δεδομένα που σας δόθηκαν για την 1η εργασία, να εξεταστούν τα ακόλουθα:

1. Να εφαρμοστεί η μέθοδος της διασταυρούμενης επικύρωσης (k-fold cross-validation με k=5) για την εκτίμηση του σφάλματος πρόβλεψης (test MSE) στα μοντέλα παλινδρόμησης των ερωτημάτων 2 και 3 της 1ης εργασίας. Ποιο από τα δύο μοντέλα παρουσιάζει το μικρότερο σφάλμα;
2. Για το μοντέλο που παρουσίασε το ελάχιστο σφάλμα στο προηγούμενο ερώτημα να εφαρμοστεί επίσης η μέθοδος Leave-One-Out Cross-Validation και να συγκριθεί με το αντίστοιχο αποτέλεσμα του προηγούμενου ερωτήματος. Ποια από τις δύο εκτιμήσεις σφάλματος που προέκυψαν θεωρείτε εγκυρότερη και γιατί;

Μέρος Β

- *Η μεταβλητή Life Expectancy να μετατραπεί σε δυαδική ως ακολούθως: να της αντιστοιχιστεί η τιμή 1 οποτεδήποτε η τιμή της είναι υψηλότερη από τη μέση της τιμή (υψηλό life expectancy) και η τιμή 0 οποτεδήποτε η τιμή της είναι χαμηλότερη από τη μέση της τιμή (χαμηλό life expectancy).*
- *Τα δεδομένα να διαχωριστούν σε training (80%) και test (20%).*

Κατόπιν αυτών να απαντηθούν τα ακόλουθα ερωτήματα χρησιμοποιώντας αποκλειστικά τη δυαδική εκδοχή της μεταβλητής Life Expectancy και όχι την ακριβή τιμή της.

3. Να γίνει ταξινόμηση των δεδομένων σε υψηλό και χαμηλό life expectancy με εφαρμογή της μεθόδου της λογιστικής παλινδρόμησης (με οριακή τιμή της πιθανότητας κατάταξης – threshold - ίση με 0.5). Κατά την εφαρμογή της μεθόδου να γίνει εκτίμηση της ακρίβειας που επιτυγχάνεται στα test data. Να επαναληφθεί η ίδια διαδικασία για threshold = 0.4 και threshold = 0.6. Τι παρατηρείτε ως προς την ακρίβεια της μεθόδου με τις διαφορετικές τιμές threshold;
4. Να επαναληφθεί η διαδικασία του ερωτήματος 3 με εφαρμογή της μεθόδου Linear Discriminant Analysis.
5. Να επαναληφθεί η διαδικασία του ερωτήματος 3 με εφαρμογή της μεθόδου K-nearest-neighbors και K= 1, 3, 5, 7 και 9.
6. Ποια από τις μεθόδους ταξινόμησης και για ποια τιμή της παραμέτρου της παρουσιάζει την καλύτερη ακρίβεια; Σχολιάστε υπολογίζοντας τόσο το training error, όσο και το test error της κάθε μεθόδου.