Άσκηση [Ανάπτυξη και εφαρμογή μοντέλων εποπτευόμενης μάθησης για την πρόβλεψη πολλαπλών επιπέδων παχυσαρκίας και εξαγωγή σημαντικών χαρακτηριστικών]

Εισαγωγή στο πρόβλημα

Σε αυτήν την άσκηση καλείστε να εφαρμόσετε τεχνικές εποπτευόμενης μηχανικής μάθησης στο προγραμματιστικό περιβάλλον της Python με την χρήση της βιβλιοθήκης scikit-learn (https://scikit-learn.org/stable/) για την ανάπτυξη και επαλήθευση μοντέλων πρόβλεψης πολλαπλών επιπέδων παχυσαρκίας. Το σύνολο δεδομένων που θα αναλύσετε βρίσκεται στον παρακάτω σύνδεσμο από το UCI (University of California, Irvine):

https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition (αρχείο: ObesityDataSet_raw_and_data_sinthetic.csv)

και περιλαμβάνει δεδομένα που αφορούν επίπεδα παχυσαρκίας με βάση τις διατροφικές συνήθειες και την φυσική κατάσταση. Τα δεδομένα φέρουν ετικέτα με την μεταβλητή κλάσης «NObesity» (Επίπεδο παχυσαρκίας) που καθιστά δυνατή την ταξινόμηση των δεδομένων σε 7 επίπεδα βάρους: (i) Ανεπαρκούς Βάρους (Insufficient weight), (ii) Κανονικού Βάρους (Normal weight), (iii) Υπερβολικού Βάρους Επιπέδου I (Overweight Level I), (iv) Υπερβολικού Βάρους Επιπέδου II (Overweight Level II), (v) Παχυσαρκίας Τύπου I (Obesity Type I), (vi) Παχυσαρκίας Τύπου II (Obesity Type II) και (vii) Παχυσαρκίας τύπου III (Obesity Type III).

Ερωτήματα

Αφού μελετήσετε το σύνολο δεδομένων απαντήστε τα ακόλουθα ερωτήματα:

1. Δώστε μια σύντομη παρουσίαση του dataset (τι περιγράφει, πλήθος χαρακτηριστικών).
2. Αναφέρετε ποια είναι τα χαρακτηριστικά εισόδου και ποιο το χαρακτηριστικό – στόχος.
3. Αναφέρετε τον τύπο δεδομένων των χαρακτηριστικών (π.χ., ακέραιος, αλφαριθμητικό).
4. Αναφέρετε τον τύπο των χαρακτηριστικών (π.χ., συνεχές, διακριτό).
5. Χρειάστηκε να κάνετε μετατροπές στις τιμές του αρχείου .csv (π.χ., στον διαχωρισμό των δεκαδικών, στην κωδικοποίηση των αλφαριθμητικών σε αριθμούς)? Αν ναι, ποιες?
6. Εντοπίσατε απουσιάζουσες τιμές? Αν ναι, αναφέρετε το ποσοστό επί του συνόλου.
7. Εφαρμόστε τεχνικές feature selection για να ταξινομήσετε τα χαρακτηριστικά εισόδου ως προς την σημαντικότητά τους με το χαρακτηριστικό-στόχο.
8. Εκπαιδεύστε και συγκρίνετε τους παρακάτω αλγορίθμους εποπτευόμενης μάθησης (supervised machine learning) τύπου bagging και boosting:
    a. Random Forests Classifier
    b. AdaBoost Classifier
    c. Gradient Boosting Classifier
9. Για την αξιολόγηση της απόδοσης των αλγορίθμων πρέπει να διεξάγετε την διαδικασία επαλήθευσης stratified k-fold cross validation. Ορίστε το πλήθος των folds (παράμετρος k) ώστε να έχετε όσο το δυνατόν ποιο ισορροπημένους πληθυσμούς ανά κλάση.
10. Υπολογίστε τις μετρικές απόδοσης accuracy, sensitivity, specificity (ανά κλάση).
11. Δώστε τον τελικό confusion matrix που θα αποτυπώνει τα συνολικά αποτελέσματα.
12. Ανεβάστε τον κώδικα σε ένα GitHub repository.
13. Σχολιάστε τα ευρήματά σας σε ξεχωριστή αναφορά (docx).

Σημείωση: Σε περίπτωση που χρησιμοποιηθούν τεχνολογίες generative AI για την παραγωγή

του κώδικα παρακαλώ αναφέρετε ποια τεχνολογία χρησιμοποιήθηκε και τον τρόπο χρήση της.

Σχετικές δημοσιεύσεις

1. Palechor, F. M., & de la Hoz Manotas, A. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. Data in Brief, 104344.
2. De-La-Hoz-Correa, E., Mendoza Palechor, F., De-La-Hoz-Manotas, A., Morales Ortega, R., & SÃ¡nchez HernÃ¡ndez, A. B. (2019). Obesity level estimation software based on decision trees.