ΕΚΦΩΝΗΣΗ

Επιλέγοντας 1 σετ δεδομένων από τον σύνδεσμο https://www.ebi.ac.uk/gxa/sc/home, το οποίο να σχετίζεται με τις λέξεις κλειδιά «Covid-19», «SARS-CoV-2», θα εκτελέσετε τα εξής με τη χρήση ενός λογισμικού της επιλογής σας (πχ. MatLab, R, Python):

A1. **Να εκτελέσετε και να συγκρίνετε τους παρακάτω αλγορίθμους ταξινόμησης.**

- **CatBoost**
- **Light GBM**
- **XGBoost**

Σημειώσεις:

- Σε κάθε αλγόριθμο θα αναφέρετε τις παραμέτρους του και ποιες τιμές χρησιμοποιήσατε.
- Οι εκτελέσεις να γίνουν με 10-fold cross validation στην επιλογή των δειγμάτων
- Οι συγκρίσεις θα γίνουν χρησιμοποιώντας τις μετρικές ορθότητα (accuracy), εξειδίκευση (specificity), F1-σκορ.
- Θα πραγματοποιηθούν 10 ανεξάρτητες επαναλήψεις για κάθε πείραμα και τα αποτελέσματα θα είναι σε μορφή boxplot (x άξονας: αλγόριθμοι, y άξονας: σκορ μετρικής)

A2. **Να εξάγετε τις μετρικές «feature importance» ή «variable importance» των τριών τεχνικών του Α1 ερωτήματος και μέσω μιας μεθόδου συνάθροισης αποτελεσμάτων στην κατηγορία rank-based (Borda, Schulze, Condorcet loser), να εξάγεται την ιεραρχία των πιο «σημαντικών» γονιδίων ως προς τον διαχωρισμό των κατηγοριών του δεδομένου που έχετε επιλέξει.**

Καλή επιτυχία!

Οδηγίες για την εκπόνηση της εργασίας:

- Η εργασία σας θα πρέπει να έχει εξώφυλλο, να είναι σε μορφή εγγράφου Portable Document Format (PDF) ή Microsoft Word (.doc ή .docx). Προτεινόμενος τύπος και μέγεθος γραμματοσειράς: Times New Roman, 12. Προτεινόμενο διάστιχο: 1,5. Η αναφορά θα πρέπει να μην ξεπερνά τις 2000 λέξεις.
- Τηρήστε το πρότυπο APA στη γραφή των παραπομπών και βιβλιογραφικών αναφορών.