Term Project

Forensics investigation from fingerprint microbes

Recently it was discovered that DNA samples from human fingerprints are unique to individuals. Therefore, it is possible to get samples from computer keyboards and identify who is using the computer. This task provides very strong patterns and the recognition rate is high. However, the harder task is to detect which hand (left or right) the samples are gathered from. In this project, the task is to identify which hand of an individual touched the computer keyboard. This project involves assessing classification performance of clinical data gathered from DNA data on computer keyboards. The task is to perform supervised learning on the dataset and report the classification performance.

Data

There are 271 samples (first 136 left, second 135 right hands). Each sample contains 3302 features. Therefore each file contains a table of 3302 x 271 entries. 136 of the samples are gathered from right-hand, and 135 are from left-hand. The dataset *otu.csv* (in an alternative format *otu.xlsx*) is provided. The first row of the files is the sample names, and the second row indicates if they are collected from left or right hands. Thay all have the same data in different formats.

Goal

With this project it is expected to have the highest possible correct classification percentage. In order to achieve that you are expected to perform attribute selection (note: cross-validation in attribute selection is also required), and then go for classification with the selected attributes.

Classification Algorithms

Project Groups (should be between 1 to 3 people) are free to use **ANY** classification algorithm/technology that can be found in the literature. **ANY** programming language and platforms including machine learning packages **except WEKA** can be used. If a group programs the project, the executable build is requested.

Performance Measures

Sensitivity, *specificity* and AUC is requested as the output of the program performance. If a group programs the project AUC is not necessary.
Sensitivity: $= \frac{correct\ number\ of\ prediction\ of\ the\ first\ class}{totalnumberofelementsinthefirstclass}$
Specificity: $= \frac{correct\ number\ of\ prediction\ of\ the\ second\ class}{totalnumberofelementsinthesecondclass}$
AUC = Area under the ROC curve.
DEADLINE IS DECEMBER 29TH, 2022. Return your reports to: nalbantoglu@odev.erciyes.edu.tr. If you compress your files please use zip but not rar format. For all your questions and unclear points please e-mail: nalbantoglu@odev.erciyes.edu.tr Good Luck!