Fifth Classwork/Homework

data mining introduction

15/01/2026

DATASETS

The dataset is downloadable from the following google drive link:
https://docs.google.com/spreadsheets/d/1XdOqrtSZJ1iTzTOmSG_xFR7WrUln1V6_/edit?usp=drive_link&ouid=111650272931636936412&rtpof=true&sd=true

The dataset, provided in Excel format, contains a list of children divided into three categories (represented as three separate sheets): **ASD**, **GDD**, and **Controls**. These categories represent the classes:

- **ASD**: Children diagnosed with autism.
- **GDD**: Children with global developmental delay.
- **Controls**: Typically developing children.

Each patient is described by several features, with values ranging between 0, 1, or 2. However, the following five features should be removed as they are not relevant for the classification task:

- "Età cronologica (mesi)"
- "Scala B"
- "Scala D"
- "TOT."
- "Score di rischio"

Additionally, children with an "Età equivalente" (equivalent age) of less than 12 months should be excluded from the dataset. This is because such patients are only assessed using "Scala B primo anno" and "Scala D primo anno," while other features are defaulted to 0 (a value that could incorrectly represent valid data) introducing noisy.

TASKS

1. Plot the patients (based on their features) in a 2d scatter plot (use PCA with reduction equal to 2).
2. Split the dataset in train test and test set.
3. Implement a class that allows to use Decision Tree, Random Forest, SVC, K-NN Classifiers. Then, you have to train such models based on the dataset. The gool of each

classifier is to predict if a new patient is ASD, GDD or Controll. In addition, you need to use a 5-cross validation to select the best model avoiding overfitting. To do this use the library GridSearchCV of scikit-learn.

4. Use the boosting and bagging concepts to use more classifiers of the same type to improve the performance of classification.

NOTE

You have to do an object-oriented implementation. Therefore create a work directory, then put a README.md file in order to write the explanation of the implemented project, and finali creare a src directory (inside the work dir) containing the code. You need to create an open repository on github called IDM 2025, then you will create a sub-directory called first_classwork, and store the code inside such directory. You need to explain the results into a doc.

DEADLINE

Date when you complete
Send an email to antonio.dimaria1@unict.it when you complete the project, with the link to the repository