First Classwork/Homework

data mining introduction

25/10/2024

DATASETS

The libary scikit-learn contains a suites of toy datasets for algorithms testing. Use the following ones for this classwork:

- **Diabetes**,
- **Wine recognition**
- **Breast cancer wisconsin in scikit-learn library**

URL: **https://scikit-learn.org/1.5/datasets/toy_dataset.html**

TASKS

1. Compute the **correlations** between the features of each dataset. Therefore, if a dataset is composed of 4 features for example, then we need to compute the correlation between: (feature 1, feature 2), (feature 1, feature 3) and so on. After that, consider the top three and more correlated pair of features and plot the second element of the tuple on the basis of the first one.
   Use the following correlation techiques:
   - **Pearson**
   - **Spearman**
   - **Kendall**

2. Compute the **PCA** and **SVM** on each dataset. Then saved the reduced dataset into a csv file.

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.

DEADLINE

29/10/2024
Send an email to antonio.dimaria1@unict.it
Object: name surname – university id
Append .zip or .tar.gz file