Lab assignment 1: Classification

Statement

Datasets:

This assignment will analyze the **Diabetes.csv** dataset which contains the following information:

- The data contains information about patients in a diabetes study. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.
- Source: https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes.
- Attribute Information:
    1. **PREGNANT**: Number of times pregnant
    2. **GLUCOSE**: Plasma glucose concentration in an oral glucose tolerance test
    3. **BLOODPRESS**: Diastolic blood pressure (mm Hg)
    4. **BODYMASSINDEX**: Body mass index (weight in kg/(height in m)^2)
    5. **INSULIN**: 2-Hour serum insulin (mu U/ml)
    6. **SKINTHICKNESS**: Triceps skin fold thickness (mm)
    7. **PEDIGREEFUNC**: A synthesis of the diabetes mellitus history in relatives and the genetic relationship of those relatives to the subject.
    8. **AGE**: Age (years)
- Output:
    9. **DIABETES**: Class variable (0 or 1)

Deliverables:

You will have to submit two files through **Moodlerooms** before October 19th:

- **A report in PDF format** that contains the developed code --screen captures of the code are not allowed--, justified answers to the proposed questions, and analyses of the results. The report does not need to be long, but should demonstrate that you worked through the whole statement. Do not include figures or code without a comment about it. Remember to include a conclusion section.
- **A compressed folder**, in .zip or .7z format, with all your code files and any additional file[1] that you might want to attach (for example, a model which takes too long to train).
- **Quality of the code will be assessed and may penalize the final grade of the assignment.**
- **Format of the report will be assessed and may penalize the final grade of the assignment.**

Questions:

The objective of this practice is to compare different classification algorithms with a real dataset.
Load the dataset **Diabetes.csv** and:

1. Exploratory Data Analysis (EDA)
2. Identification and fitting process of classification models
3. Comparative analysis of the fitted models
4. Creativity and innovation

   In this section you should look for resources on the internet that are applicable to this assignment and were not taught in class. These resources can be concepts, techniques, packages, etc... that are applicable in this exercise. For example, you could use packages that aid in dataset exploration or classification model analysis. Extra effort on the other sections would also be taken into account in this section.
5. Conclusions

[1] https://pythonbasics.org/pickle/