UNIVERSITY OF LIVERPOOL

COMP534 – APPLIED ARTIFICIAL INTELLIGENCE

FIRST ASSIGNMENT

Performance analysis of multiple supervised learning methods for solving a binary classification problem

Dr Pamela Bezerra

28 of February 2022

PROJECT GOAL: Write a python program to compare the performance of three different classification methods selected among the list below:

- Decision Tree
- Random Forest
- KNN
- SVM
- Naïve Bayes
- Logistic Regression

For each method selected, evaluate the performance of different parameters or different version of the algorithm considering the metrics studied in our lectures (precision, recall, accuracy, etc). Based on the obtained results, compare the best version of each method with the remaining ones.

Any python library (scikit-learn, keras, tensorflow, matplotlib, seaborn, etc) can be used in this project, however the essential steps your code should include are:

- Load the dataset
- Split and clean the dataset
- Load and train the chosen classification methods and parameters
- Plot the results for each method and parameter (using tables or graphs)

Please include comments describing each of these key steps in your code. You are free to include any extra steps (such as data visualization) that you like (optional – this won't result in extra points).

DATASET: The dataset used in this project is the Pima Indians Diabetes Database - https://www.kaggle.com/uciml/pima-indians-diabetes-database

This database contains medical data of different female patients describing several health conditions (features) and if the patient has diabetes or not (label). It consists of 768 observations, of which 268 are positive for diabetes (label equal 1) and 500 are negative for diabetes (label equal 0).

REQUIREMENTS: This assignment should be completed in pairs and consists of two deliverables:

- The .py file containing the code (40%)
- A short report of maximum of three pages (not considering cover page and references, in case you want to include) (60%). This report should contain:
    a. An introduction Section explaining the development process (20%):
        i. The libraries used
        ii. The classification methods used, the parameters selected for each (and if any search was used to find these parameters)
        iii. The training and testing process (data split ratio, if cross-validation was used, etc)
    b. An evaluation section describing your results (60%):
        i. The confusion matrix of the best version of each method
        ii. A table comparing the Precision, Recall, F1-Score, and accuracy of all methods
        iii. Your conclusions on the results (e.g., the reasons for method A have better performance than B) based on your AI knowledge.
    c. Final conclusions (20%)
        i. The challenges of the project
        ii. The project task allocation (what each member did)

OBS: PLEASE INCLUDE THE NAME AND ID OF BOTH MEMBERS IN THE REPORT. Both members need to submit on Canvas a zip folder containing the code and report.

DEADLINE: WEDNESDAY, 16TH OF MARCH 2022 AT 17:00

GUIDANCE: Please consult the following materials for further guidance on implementing Supervised Machine Learning methods with python:

1. **Python Data Science Handbook -** https://jakevdp.github.io/PythonDataScienceHandbook/
2. **Hands-on Machine Learning with scikit-learn, keras and tensorflow -** https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/
3. **Scikit-learn tutorials -** https://scikit-learn.org/stable/tutorial/index.html

Good luck to everyone,

Best Regards,

Pamela Bezerra