جامعة بيرزيت
BIRZEIT UNIVERSITY

COMP4388: Machine Learning
Fall 2023/2024

Deadline: Monday 22 December 2023

In this (individual) project you will build a model that predicts if a given person has diabetes or not.

The dataset can be found under this link:
https://www.dropbox.com/scl/fi/ahlg01iial19mfl7wrjsy/Diabetes.csv?rlkey=7vwl95ly3hcdvqmwo7t3ply4j&dl=0

You have to perform the following tasks:

1. Following the steps we have learnt in Exploratory Data Analysis (EDA), print the summary statistics of all attributes in the dataset.
2. Show the distribution of the class label (Diabetic) and indicate any highlights in the distribution of the class label.
3. For each age group, draw a histogram detailing the amount of diabetics in each sub-group.
4. Show the density plot for the age.
5. Show the density plot for the BMI.
6. Visualise the correlation between all features and explain them in your own words.
7. Split the dataset into training (80%) and test (20%).

As for the correlation, please detail the correlation between features and make sure to have your features that will be input to the machine learning models being clean. Based on the correlation, you have to decide which features to stay for the learning stage and which can be deleted.

Data dictionary: NPG: number of pregnancies; PLG: Plasma glucose concentration; DIA: Diastolic blood pressure; TSF: Triceps skin fold thickness; INS: 2-Hour serum insulin; BMI: body mass index; DPF: Diabetes pedigree function; AGE: age; and Diabetic: 1 for positive (diabetic) and 0 otherwise.

Regression tasks:

1. Apply linear regression to learn the attribute "Age" using all independent attributes (call this model LR1).
2. Apply linear regression using the most important feature (based on the correlation matrix); and explain why did you use this single attribute (call this model LR2).
3. Apply linear regression using the set of 3-most important features (based on the correlation matrix); and explain why did you use these 3 attributes (call this model LR3).
4. Compare the performance of these models using adequate performance metrics and explain the difference between them.

Classification tasks:

1. Run k-Nearest Neighbours classifier to predict (the "Diabetic" feature) using the test set.
2. Generate kNN classifier using different values of *k*. You should have at least **4 models** for the different values of *k* and compare their performance in an appropriate results section. Make sure to use the appropriate performance metrics and you should include the ROC/AUC score and the Confusion Matrix. Report the results in an appropriate table and explain in your own words why one model outperforms the other.

You have to turn in a softcopy of your Python code and a Word document containing the information required as specified above. The document should be on a paper-format. Please send your submissions as a reply to the message sent on Ritaj only with the files named "COMP4388-XXXXX.docx/pdf" and "COMP4388-XXXXX.py" where XXXXX is your BZU-student ID number.

If you have any questions, please feel free to contact me via Ritaj or email: rjarrar@birzeit.edu