جامعة بيرزيت
BIRZEIT UNIVERSITY

COMP4388: Machine Learning
Fall 2023/2024
Project 2

Deadline: Monday 25 January 2024 by 11:00 PM

In this team project of twos, you will build a model that predicts if a person is a Smoker or not using data that is collected from their Health Insurance data.

The dataset can be found under this link:
https://www.dropbox.com/scl/fi/8et6xuwh9luvfg03hhji3/Data.csv?rlkey=10s6tu2sgw5z3ft43qk3wey79&dl=0

You have to perform the following tasks:

1. Show the distribution of the class label (Smoker) and indicate any highlights in the distribution of the class label.
2. Show the density plot for the age.
3. Show the density plot for the BMI.
4. Visualise the scatterplot of data and split based on Region attribute.
5. Split the dataset into training (80%) and test (20%).

Data dictionary: Age: The age of a person; Gender: the gender of a person; BMI: body mass index; Region: north or south; No. Children: number of children; Insurance Charges: the amount paid for the insurance company by the person; and Smoker: yes for positive (smoker) and no otherwise.

Tasks to do:

1. Compare the following Machine Learning algorithms: KNN (using 3 different values of K), Decision Trees (C4.5), NB, ANN (with a single hidden layer, number of epochs = 500, sigmoid activation function).
2. 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.P2.STUDENT_ID.docx/pdf" and "COMP4388.P2.STUDENT_ID.py".

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