جامعة بيرزيت
BIRZEIT UNIVERSITY

COMP4388: Machine Learning
Fall 2021/2022

Deadline: Friday 23 November 2021

In this project you will deal with a dataset of Algerian Forest Fires. You are required to perform a series of tasks that are explained in this document.

The dataset using in this assignment can be found under this link:

UCI Machine Learning Repository: Algerian Forest Fires Dataset Data Set

The link contains a full description of the data.

You have to perform the following tasks:

1. Print the summary statistics of all attributes in the dataset (i.e., minimum, maximum, first quartile, second quartile, third quartile, and the mean)
2. Print the density plot of the entire dataset and split it into two curves (split by Classes, draw the Temperature)
3. Visualise the correlation between the independent features and explain them in your own words
4. Visualise correlation between the dependent feature (i.e., "Classes") and the independent variable and explain them in your own words
5. Split the data into 80% training and 20% test and then:
    a. From the correlation coefficient that resulted from the previous step, please decide on the one feature that is best used to predict the "Fire Weather Index (FWI)" and state the reason in your own words
    b. Apply Linear regression using the feature from the previous step (i.e., 5.a) to predict the values of "Fire Weather Index (FWI)" in the test set
    c. Apply Linear regression using the two features with the highest correlations to predict the value of "Fire Weather Index (FWI)" in the test set
    d. Apply Linear regression using all independent features to predict the value of "Fire Weather Index (FWI)" in the test set
    e. For each model, please provide analysis of the Linear regression output using performance measures for

regression and explain the difference in performance (if exists) in your own words

6. Run Logistic regression classifier to predict if there will be a fire or not (the "Classes" feature) using the training dataset
7. Run k-Nearest Neighbour classifier to predict if there will be a fire or not using the training set
8. Compare the performance of Logistic regression and kNN classifiers in an appropriate results section. Compare the Error rate, Precision, and Recall of the models 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

All the best of luck!
Dr. Radi Jarrar