University of New Orleans

Department of Computer Science

**Spr 2019: CSCI 6990**

**Programming Assignment # 1**

Machine Learning - I

Submitted to:

**Professor Dr. Tamjidul Hoque**

By

Name: **Astha Sharma**

Student ID: **2564173**

The programming assignment was done in Google Collab Notebook

Every program file in the main folder is defined below as per the functionality.

**PART (A)**

Features.ipynb:

At first, the original dataset was splitted into feature set (x) and output (y). The feature set was preprocessed and correlation coefficient was calculated for each one of those. Depending upon the value of the coefficient, the top 5 features were listed in a separate file top5.csv. Also the correlation graph for the top 5 features were plotted.

RandomizedData-12.ipynb:

For all 12 features, the data is normalized, shuffled, and saved as separate files for different models (M=1,..6, YC) inside the Processed Dataset folder. The using 10 folds cross validation technique to create test and train datasets using source and destination location of input and output files, respectively. The dataframe is used from initial optimized dataset.csv file.

ModelPrediction -12.ipynb:

For the 12 feature set, the value of beta is calculated for models of order 1,2,4,6, and 5 (my choice). Consecutively, the mean-absolute-error and root-mean-square-error is also calculated using the given formula. For each model, the input data is provided from a relevant input link. The obtained values are used to plot tables and graphs using Excel.

RandomizedData-5.ipynb:

For top 5 features, the data is normalized, shuffled, and saved as separate files for different models (M=1,..6, YC) inside the Processed Dataset folder. The using 10 folds cross validation technique to create test and train datasets using source and destination location of input and output files, respectively. The dataframe is used from the created top5.csv file.

ModelPrediction -5.ipynb:

For the top 5 feature set, the value of beta is calculated for models of order 1,2,4,6, and 5 (my choice). Consecutively, the mean-absolute-error and root-mean-square-error is also calculated using the given formula. For each model, the input data is provided from a relevant input link. The obtained values are used to plot tables and graphs using Excel.

**PART(B)**

RegularizationModel-12.ipynb

A new beta value is calculated using given equation and for varying values of lambda = 0, 0.5E-8, 1.5E-6, 2.0E-4, 1 and 2. The lambda vs MAE and lambda vs RMSE graph for Model 6 and 5 (my choice) is plotted from obtained values.

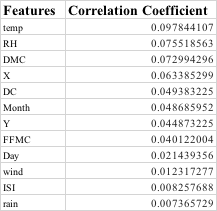
RegularizationModel-5.ipynb

Same thing is done for the top 5 input-features based model, the only difference is in the size of the identity matrix.

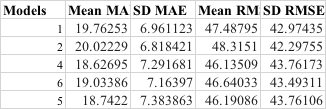


**TABLES**:

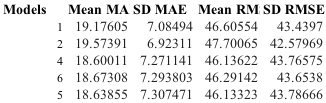
TABLE#1



TABLE#2



TABLE#3:



TABLE#4:

















GRAPHS:

