Programming Questions

The goal of this section is to create two regression-based models to assess air quality. The data (data train.csv) which includes the training data (features and labels) and (data test.csv) which includes the test data. Remember, the test data only contains the features we use for the training, but not the labels. For each row in the test data, you need to use the trained model to predict the corresponding labels. Each row of the data corresponds to a sample and the columns include the following information:

1. NMHC(GT): hourly averaged overall Non Metanic HydroCarbons concentration in microg/ m^3
2. C6H6(GT): hourly averaged Benzene concentration in microg/m^3
3. C6H6(GT): hourly averaged Benzene concentration in microg/m^3
4. PT08.S2(NMHC): hourly averaged sensor response to NMHC
5. NOx(GT): hourly averaged NOx concentration in ppb
6. PT08.S3(NOx): hourly averaged sensor response for NOx
7. NO2(GT): hourly averaged NO2 concentration in microg/m^3
8. PT08.S4(NO2): hourly averaged sensor response for NO2
9. PT08.S5(O3): hourly averaged sensor response for O3
10. T: Temperature in C
11. RH: Relative Humidity
12. AH: Absolute Humidity
13. PT08.S1(CO): TARGET VARIABLE - hourly averaged sensor response for CO

(a) Data Processing

1. Download and read the data. For Python, you may use *pandas* library and use *read csv* function.
2. Print the first 5 rows of the data using the command. (You may use *head()* function in *pandas* library). Print the shape of the training dataframe. Write a short description of the data.

3. Does the data have any missing values? How many are missing? Return the number of missing values. (In *pandas*, check out *isnul()* and *isnul()*.sum())
4. Drop all the rows with any missing data. (In *pandas*, check out *dropna()*. *dropna()* accepts an argument *inplace*, check out what it does and when it comes in handy.)
5. Extract the features and the label. The label is *PT08.S1(CO)*.

(b) Exploratory Data Analysis

1. Plot the histograms of all the features in the data. Do all the features have a normal distribution? Do you see any outlier values? Do you need to apply any normalization technique to these values? If so, you can transform your data in this step and explain your thought process in the corresponding markdown cell.
2. Pick 2 features and create a scatter plot to illustrate the correlation between these two features. Is there a high correlation between these features?
3. Compute the Pearson's correlation between all pairs of variables 1-12. Assign the resulting correlation values in a 12x12 matrix C, whose (i; j) element represents the correlation value between variables i and j, i.e., C(i; j) = corr(i; j). Visualize the resulting matrix C with a heatmap and discuss potential associations between the considered variables. Note: You can use the 'heatmap' function from 'seaborn'.

(c) Linear Regression Implementation Implement a linear regression model **from scratch** to regress the target variable, Carbon monoxide (CO). (Remember: You can not use any libraries for the linear regression model.)

(d) Logistic Regression Implementation Using the column PT08.S1(CO), create a binary label for this dataset where the values more than 1000 correspond to label 1 and the values less than or equal to 1000 correspond to label 0. Implement a logistic regression model **from scratch** to predict this binary label. (Remember: You can not use any libraries for the logistic regression model.)

(e) Result Analysis - Linear Regression Perform a 5-fold cross validation. Compute RMSE for each validation set across 5 folds. Report average and standard deviation of RMSE values. Do you see a big change across different folds? How can you use the coefficient of this model to find the most informative features?

(f) Result Analysis - Logistic Regression Perform a 5-fold cross validation. Compute accuracy, precision, recall, and F1 score for each validation set across 5 folds. Report the average and standard deviation of these metrics. Do you see a big change across different folds?

(g) ROC Curve - Logistic Regression Use the logistic regression model from sklearn and repeat 5-fold cross validation. Then using *roc curve* package from *sklearn.metrics*, plot the ROC curve for each fold and compute the area under the curve. Is this result consistent with the results you obtained using the logistic regression model you implemented?

(h) Inference - Linear and Logistic Regression

1. Use the trained linear regression model and predict the PT08.S1(CO) value for the test data.

2. Use the trained logistic regression model and predict the PT08.S1(CO) binary value (the same label you created in step d) for the test data.

3. Save the predictions in a csv file with two main columns. One for the linear regression predictions with the name pred_linear and one for the logistic regression predictions with the name pred_logistic.

4. Add this csv file to your submission.