Machine Learning Assignment

| Group 145 | PS - 10 | Part - B |

Group Members:

Name	Email	Contribution
THANGADURAI C	2021sc04117@wilp.bits-pilani.ac.in	100%
B VENKATA NARAHARA SESHA SAI PAVAN KUMAR	2021sc04115@wilp.bits-pilani.ac.in	100%

Documentation:

EDA:

After performing exploratory data analysis on the given global air pollution dataset, and plotting necessary graphs such as boxplot, correlation heatmap and bivariate analysis, we have inferred the following insights.

1. There is more data related to big countries such as the United States of America (2454 rows), India (1747 rows) and so on. Smaller countries such as Hakodate, Dharmadam, and Sungairaya have only 1 row. This will affect the result of the analysis as the amount of data with respect to smaller countries is very very small.
2. The PM2.5 AQI value is highly correlated with the AQI Value for any given country with a positive correlation value of 0.98.
3. The Ozone AQI Value is negatively correlated with the NO2 AQI Value with a correlation value of -0.24.

Classification:

1. *Preprocessing and Feature Engineering:*
    a. The data had a significant amount of outliers which were processed using the Inter-Quartile Range method.
    b. The values for AQIs of various types are of various ranges. This will lead to inaccuracies in data modelling. Therefore all the numeric columns are re-scaled to have similar scales.
    c. The categorical columns such as AQI Category have been converted to numerical by mapping the classes into their respective labels ranging from 0 to 3.
2. *Model Development:*
    a. AQI Category is the chosen target attribute as it is evident from the data analysis that this value is a mathematical model with respect to other AQI values. AQI

Value column is excluded from the training features as it is the directly correlated feature.
    b. This classification problem will be a multi-class classification problem, as the number of classes are 4.
    c. Selected model is a Decision Tree Classifier. Model is built using Scikit Learn library. As the data has multiple classes, logistic regression is not possible. Support Vector Machines are weaker with classification when compared to Decision Tree. Decision Tree also handles collinearity better than SVM. Therefore, a Decision Tree Classifier is modelled.
3. *Model Evaluation:*
    a. The model is trained on 17,786 rows and the remaining 1,977 rows are left for testing. Upon training, the training accuracy is 1.00.
    b. The model is evaluated with a testing data set which is sampled randomly from the complete dataset. Upon testing, the testing accuracy is 1.00.
4. *Final Comments:*
    a. Based on the classification problem, all the other AQI Values contribute to a mathematical equation to give the AQI Value which is a direct correlative of the AQI Category.
    b. With training and testing accuracies of 1.00 and 1.00 respectively, this must be the best model for this data. Also the model is easy to build and deploy.

Regression:

1. *Preprocessing and Feature Engineering:*
    a. The data had a significant amount of outliers which were processed using the Inter-Quartile Range method.
    b. The values for AQIs of various types are of various ranges. This will lead to inaccuracies in data modelling. Therefore all the numeric columns are re-scaled to have similar scales.
    c. The categorical columns such as AQI Category have been converted to numerical by mapping the classes into their respective labels ranging from 0 to 3.
2. *Model Development:*
    a. AQI Value is the chosen target attribute as it is evident from the data analysis that this value is a mathematical model with respect to other AQI values. AQI Category column is excluded from the training features as it is the directly correlated feature.
    b. This problem will be a regression problem, as the target value is a continuous variable.
    c. Selected model is a Linear Regressor. Model is built using Scikit Learn library. As the data is tabular and seems to have a mathematical model with respect to other AQI values, and absence of multi-collinearity between features, it is better to use.
3. *Model Evaluation:*
    a. The model is trained on 17,786 rows and the remaining 1,977 rows are left for testing. Upon training, the training accuracy is 0.98.

    b. The model is evaluated with a testing data set which is sampled randomly from the complete dataset. Upon testing, the testing accuracy is 0.98.
4. *Final Comments*:
    a. Based on the regression problem, all the other AQI Values contribute to a mathematical equation to give the AQI Value.
    b. With training and testing accuracies of 0.98 and 0.98 respectively, this must be the best model for this data, as the implementation complexity and accuracy tradeoff is low. We can achieve high accuracy with simplest implementation.

Ensemble Modeling Classification:

1. *Preprocessing and Feature Engineering:*
    a. The data had a significant amount of outliers which were processed using the Inter-Quartile Range method.
    b. The values for AQIs of various types are of various ranges. This will lead to inaccuracies in data modelling. Therefore all the numeric columns are re-scaled to have similar scales.
    c. The categorical columns such as AQI Category have been converted to numerical by mapping the classes into their respective labels ranging from 0 to 3
2. *Model Development:*
    a. AQI Category is the chosen target attribute as it is evident from the data analysis that this value is a mathematical model with respect to other AQI values. AQI Value column is excluded from the training features as it is the directly correlated feature.
    b. This classification problem will be a multi-class classification problem, as the number of classes are 4.
    c. Selected model is a Random Forest Classifier. Model is built using Scikit Learn library. As the data has multiple classes, logistic regression is not possible. Random forest is better on bigger datasets. Random forests work better with classification problems, Therefore, Random Forest classifier is chosen to model.
3. *Model Evaluation:*
    a. The model is trained on 17,786 rows and the remaining 1,977 rows are left for testing. Upon training, the training accuracy is 1.00.
    b. The model is evaluated with a testing data set which is sampled randomly from the complete dataset. Upon testing, the testing accuracy is 1.00.
4. *Final Comments:*
    a. Based on the classification problem, all the other AQI Values contribute to a mathematical equation to give the AQI Value which is a direct correlative of the AQI Category.
    b. With training and testing accuracies of 1.00 and 1.00 respectively, this must be the best model for this data. Also the model is easy to build and deploy.