CECS 456: Machine Learning

By: Ryley Benavides, David Shamis, Noah Daniels, Dhruv Salva, Pranik Pant

Professor: Mahshid Fardadi
Due: 5/1/2023

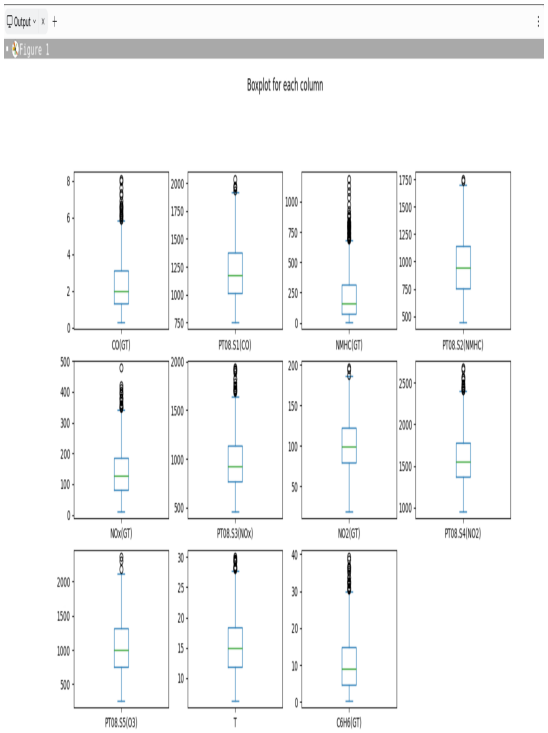
Group 14: Benzene Concentration Prediction in air quality dataset

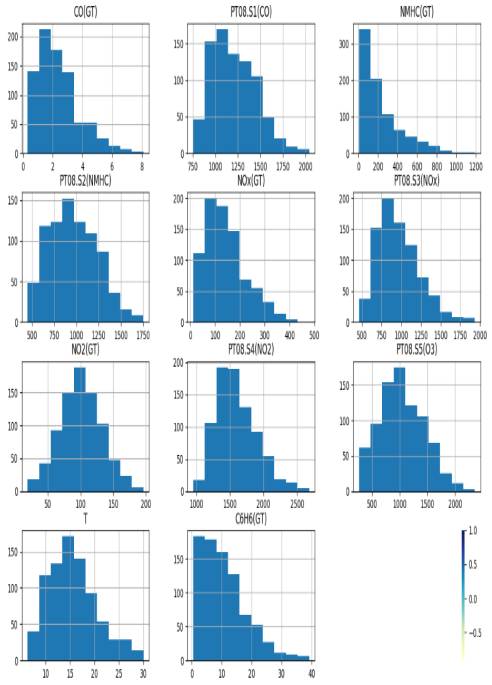
An important global issue that has been connected to a number of health issues is air pollution. Air pollution can be impactful in many different ways and one of the main pollutants in Benzene (C6H6). Benzene is a carcinogenic pollutant that studies have shown to lead to various cancers in people and is a pollutant of concern. In this project, we use regression models to predict the benzene concentration based on the other inputs. We will make use of the hourly air quality measurements from an unnamed Italian city found in the UCI Machine Learning Repository's Air Quality dataset. Hopefully this project will be able to provide insight as to the main culprits behind air pollution in that Italian city.

The dataset includes 9358 examples of hourly averaged responses from a group of five metal oxide chemical sensors that are built into an Air Quality Chemical Multi Sensor Device. The device was situated on a field at road level in a heavily polluted area of an Italian city. The longest publicly available recordings of responses from on-field deployed air quality chemical sensor devices were made from March 2004 to February 2005 (a period of one year). A co-located reference certified analyzer supplied Ground Truth hourly averaged readings for CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx), and Nitrogen Dioxide (NO2). Cross-sensitivities and concept and sensor drifts are evident, which ultimately impairs a sensor's ability to estimate concentration. The value -200 is assigned to missing values. With this type of data it is a good idea to use regression techniques as opposed to classification ones.

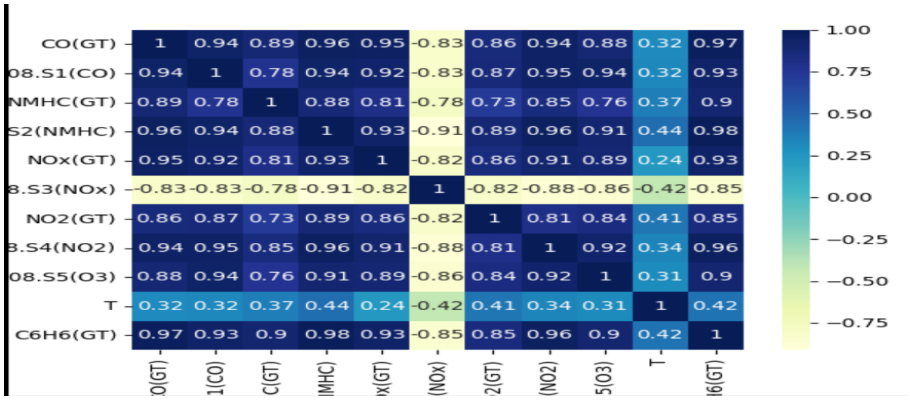
In order to better comprehend the dataset, we initially performed exploratory data analysis. We used box plots and histograms to visualize the distribution of the variable. We deduced from the box plots that certain variables had outliers, while others had medians and

ranges that were similar. We noticed from the histograms that some variables had a skewed distribution, which may have an impact on how well the model performs.

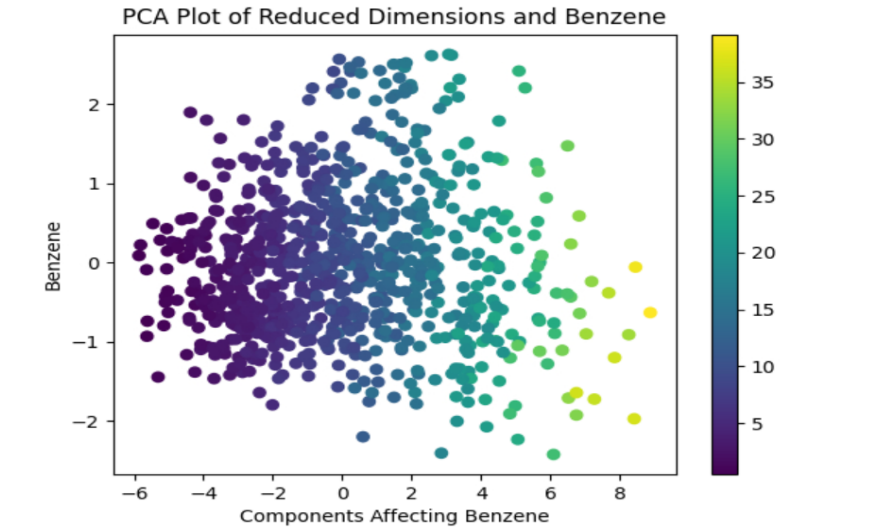




Next, we used SelectKBest to perform feature selection, which chooses the top k features based on their scores. Based on their rankings, we picked the top 10 features and included them in our study. To comprehend the connections between the variables, we produced the correlation matrix. We deduced from the correlation matrix that benzene correlated strongly with various variables, including NOx and CO. This suggests that these factors might be effective benzene concentration predictors.



We used Principal Component Analysis (PCA) to identify fresh features that effectively capture the majority of the data's volatility. We discovered that four primary components can account for 90% of the variance in the data. We tested five regression models: support vector regression (SVR), lasso regression (L1), polynomial regression, linear regression, and Bayes regression.



After we trained the model, we used 5-fold cross-validation to test the models to see how it would perform on new data. We also used k-fold cross validation to assess that SVM was the best regression model for this dataset. This technique is especially handy because it helped us

prevent overfitting and in our models and it did just that. Using this technique, we were primarily able to approximate the model's generalization error.

Throughout the process of creating the models we had to tune specific hyperparameters in order for each model to have adequate performance. Grid search was the best technique for this problem with the small amount of hyperparameters we are dealing with in these models. We tuned our L1 Regression model by defining a search grid to search over important parameters like lambda. We defined the range of lambda to be… and used k-fold validation to see if the model preformed better, which it did. For the Linear Regression model we search over hyperparameters such as alpha (regularization parameter) and the intercept to define our grid space and once again we used k-fold cross validation to test our models performance. For the Polynomial Regression we define our grid space with the hyperparameters degree of the polynomial and alpha. In this regression we create polynomial features for the input data based on the degree of the polynomial and again we use k-fold cross validation to test performance. For Bayes regression we also define the grid space on the hyperparameter alpha which is a regularization parameter. By now we can see how important regularization in regression models of machine learning is when we are constantly adjusting its hyperparameter representation. Its important to point out the kernel parameter for the SVM regression as well as it is an integral part of how the SVM algorithm works.

In this project, we used regression models to predict the benzene concentration with the other variables. We concluded after extensive hyperparameter tuning and k-fold cross validation performance estimation that the SVM regression had the best performance for this dataset. This would make sense as SVM uses The model’s performance can be enhanced by gathering more data, resolving outliers and skewness in the data, and utilizing more sophisticated approaches

like deep learning. The outcomes of this experiment may help with benzene concentration forecasting and urban air pollution control.