**Course Project Milestone 4- Preliminary Analysis- Air Quality Prediction**

Guruprasad Velikadu Krishnamoorthy and Hemalatha Subbiah

College of Science and Technology, Bellevue University

[DSC630-T302: Predictive Analysis](https://cyberactive.bellevue.edu/webapps/blackboard/execute/courseMain?course_id=_512542_1)

Professor. Fadi Alsaleem

February 9, 2024

**Air Quality Prediction**

**Introduction**

Air pollution is a significant concern in the modern world. With the rise of the Industrial Revolution in the 19th century and the invention of new machinery and technologies, economies worldwide saw explosive growth in the following decades, accompanied by the migration of farmers from the countryside to fast-growing, crowded cities in search of factory jobs. (Kiger, 2021) While the industrialists were amazed by the profits being made, the governments were astonished by the progress made in the economy; little did they all know the side effects that changed the landscape of the earth forever, for the worse. The emission of gases from factories and the pollutants from the automobiles had an everlasting impact on the face of the earth. In this project, the effects of air pollution and the ways to combat it by being able to predict it using machine learning algorithms will be discussed in detail.

**Problem statement:**

On average, it is estimated that a person inhales about 2000 gallons of air daily. Hence, it is essential that the air we breathe is of good quality. The polluted air, when inhaled, gets straight into our lungs, then enters the bloodstream, and can cause more damage to internal organs such as the brain, heart, etc., and young children are the most affected ones.

The aim of the project is to develop a machine-learning model capable of predicting the Air Quality Index(AQI) based on various environmental parameters and pollutant concentrations. The model should be capable of providing accurate AQI predictions for future time points, allowing for early detection of potential air quality issues..

**Visualizations can be useful in explaining the data.**

As most of the features used in the project are numerical, scatter plots to show the relationship between the pollutants or harmful gases versus the AQI (Air Quality Index) can be a useful visualization to understand the relationship between the two.

A graph with a line and a line

Description automatically generatedThe scatter plot of AQI vs Ozone indicates that there is a positive correlation between and most of the AQI is between 20 and 50.

A graph of a graph

Description automatically generated(Figure 1)

Also, a histogram is plotted on the AQI to identify the outliers, and based on the nature of the outlier, a decision can be made to either eliminate them or impute the data.

(Figure 2)

A graph with red squares

Description automatically generated A Bar plot showing the distribution of the data across the US states has been plotted (Figure 3). The plot indicates that most observations in the dataset are from California and Texas followed by Utah and Wyoming.

(Figure 3)

**Process of the Data Preparation:**

The data for this project was collected by combining multiple CSV files each containing information about the pollutant concentration, meteorological data, etc. Each row represents information about the pollutant concentration and the Air Quality Index for a location in the US on a given day and has additional details such as state, county, pollutant description, etc. After creating data frames for each dataset, data conversion was performed by converting some character columns to Factors and Dates

.Also, the field names were renamed for easier computations by removing spaces in the column names and replacing them with underscores.

To prepare the final data, the Dataset containing NO2 data was joined with all other datasets based on common columns such as State ID, County ID, Site ID, and the Date of Observation.

The nulls were then handled by replacing them with Median values for the column. The dataset was then checked for duplicates in the key columns and only the distinct values were retained.

Model **Development:**

The prepared dataset is involved for

* Feature engineering techniques
* Experimenting with different algorithms,
* Hyperparameters to optimize model performance.

As the project is about predicting the Air quality index, which is a continuous Numeric variable, the Regression algorithm is the appropriate machine learning algorithm(s) based on the nature of the problem, the type of data,

Feature engineering:

I leveraged the correlation matrix for analyzing and to identify pairs of features with high correlation coefficients (close to 1 or -1) and consider removing one of the features from each highly correlated pair to reduce redundancy. use the correlation matrix as a basis for feature selection by selecting features that are highly correlated with the target variable.

# **Feature Reduction by Correlation Coefficient:**

A screenshot of a computer

Description automatically generated

Feature is reduced as below:

A screenshot of a computer program

Description automatically generated

# **Feature Reduction by PCA :**

I explored PCA as another tool for dimensionality reduction as it preserves most of the variability in the data.I set for 80% here.

A math equation with numbers

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

With two types of different feature engineering methods dimensionality reduction done ,started exploring different regression algorithms as it is an essential part of model development in MLOps projects I experimented with below set of regression algorithms.

* Linear Regression
* Decision Tree Regression
* Random Forest Regression
* Gradient Boosting

Before starting the experimentation, I split the dataset into training and testing sets. The training set will be used to train the models, while the testing set will be used to evaluate their performance.

For each regression algorithm chosen, follow these steps:

* Train the model on the training data.
* Use the trained model to make predictions on the testing data
* Evaluate the performance of the model using appropriate regression evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared (R2), etc. d. Repeat steps a-c for each algorithm.
* Repeat steps a-c for each algorithm.

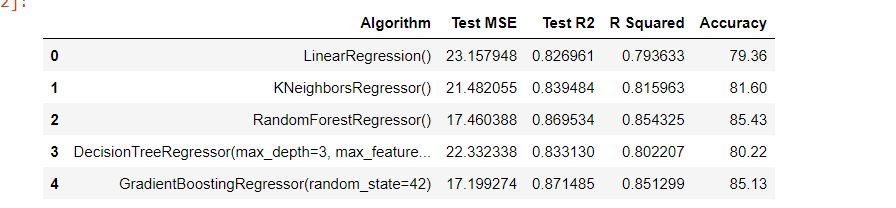
Built a function incorporated all the above steps and snippet below.

A computer code with text

Description automatically generated with medium confidence

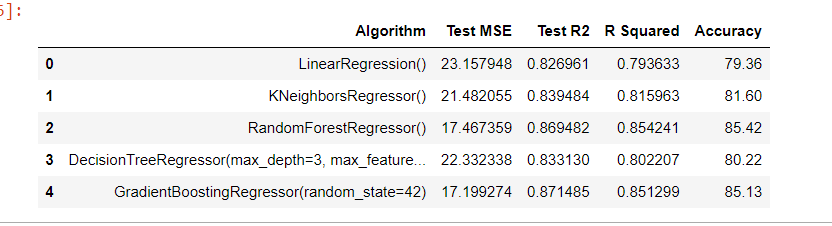
**Interpreting the Model results:**

**Model Results with No feature reduction:**

****

I observed that Gradient Boosting Regressor and Random Forest Regressor outperform the other algorithms based on the given metrics. They have lower MSE, higher R2, and better accuracy compared to Linear Regression, K Nearest Neighbors Regressor, and Decision Tree Regressor. Among these two, Gradient Boosting Regressor has a slightly lower MSE and higher R2, indicating slightly better performance overall.

**Model Results with Pearson’s method feature reduction:**

****

From above observation,Random Forest Regressor and Gradient Boosting Regressor demonstrate superior performance across all metrics compared to the other algorithms, with Random Forest Regressor having a slight edge in terms of accuracy. These algorithms are well-suited for regression tasks where accurate prediction is crucial.

**Model Results with PCA feature reduction:**

**A screenshot of a computer

Description automatically generated**

I see that K Nearest Neighbors Regressor performs the best based on the given metrics, having the lowest MSE, highest R2, and accuracy among the algorithms listed. Linear Regression has the lowest accuracy and relatively poor performance compared to the other algorithms. Random Forest Regressor and Decision Tree Regressor have similar performance but are outperformed by K Nearest Neighbors Regressor and Gradient Boosting Regressor.

**From all the analysis model performance is better without** any feature reduction technique c leading to better model results compared to employing feature reduction methods like correlation coefficient-based feature selection or Principal Component Analysis (PCA).

The results of the linear regression model in R is shown in the figure below.

One of the things that stands out in the model results is that the model yields a high R-squared value of 0.83 which indicates the proportion of the variance in the AQI that is explained by the features of the model. The R-squared and adjusted R-squared values being very close indicates that the additional factors in the model are not being penalized and the model is a good representation of the data. A screenshot of a computer

Description automatically generated

The Root mean square Error (RMSE) is one of the performance indicators for the model and is consistent for both the training and test datasets. The RMSE, which measures the average difference between the AQI values predicted by the model versus the actual AQI, indicates that we can expect an error of up to 4.77 while using the model predictions.

A screenshot of a computer program

Description automatically generated

**A graph of a graph showing a line of dots

Description automatically generated with medium confidenceOther Observations/ Recommendations:**

The results of the linear regression model are plotted in the scatter plot (Figure 4) that compares the AQI versus the predicted AQI values. It indicates the predicted values are not too different from the actual values of AQI in the dataset.

(Figure 4)

The combined data set contains the concentration of pollutants and metrological parameters such as temperature, pressure, wind speed, etc. for a location on a given day. Many graphs are plotted on the final dataset as discussed in this section.

A graph of number and number of states

Description automatically generated with medium confidenceFigure 5 represents a bar plot that contains the most polluted counties in California in 2022 and their average AQI values. San Bernardino was the most polluted followed by Los Angeles counties based on the NO2 concentration levels.

(Figure 5)

**A graph of a number of levels

Description automatically generated with medium confidence** Figure 6 represents multiple scatter plots of NO2 concentration versus AQI colored based on ozone values. Each subplot represents a state in the Southwest region. All 4 states represent a similar trend with ozone concentration uniform, though Oklahoma and Texas had higher Ozone levels between 20-40 AQI. (Figure 6)

A map of the united states

Description automatically generatedFigure 7 represents a Tree Map of the number of ozone observations in each state. As expected, California has the highest level followed by Texas and Utah.

(Figure 7)

A graph with a bar chart

Description automatically generated

Figure 8 represents the comparison of Median values of the national average of NO2 versus the top 20 states in the US. Surprisingly, the states of Georgia, Arizona, and Illinois are in the top states above the national average, while Kansas, Iowa, Colorado, and Maine are the lowest.

(Figure 8)

**Ethical Implications:**

Though we don’t have control over choosing the air that we breathe, several ethical impacts must be considered while analyzing the impacts of air pollution. In many cases, it is hard to identify the source of the origin of air pollution, reasonable measures can be taken to control it while it cannot be avoided or stopped completely. Some of the ethical implications are listed below:

* Though gas-powered vehicle emissions and industrial smoke have played a significant role in air pollution, they cannot be entirely replaced by sustainable solutions, as they can lead to many job losses affecting many families employed by the manufacturing industries. Care must be taken while publishing the results, keeping in mind the impact it can have on families.
* While determining the acceptable levels of greenhouse gases and pollutants for humans, careful assessment should be made while determining the values, as the acceptable levels for humans may cause significant damage to other ecosystems and species. (Brown, 2001)
* The acceptable levels should also be carefully assessed with international considerations in mind, as the gas emissions and pollutants from the developed countries are no longer a local issue. These impacts are already seen on the other side of the world, with extreme floods and drought conditions that were not seen in the past.

**References**:

Kiger, P. (2021, November 9). Seven negative effects of the Industrial Revolution. History. <https://www.history.com/news/industrial-revolution-negative-effects>

‌Air quality. (n.d.). Transport & Environment. <https://www.transportenvironment.org/challenges/air-quality/>

‌10 Things You Never Knew Could Cause Lung Cancer. (n.d.). WebMD. <https://www.webmd.com/cancer/ss/slideshow-lung-cancer-surprising-causes>

“Like breathing poison”: Children in India’s Delhi hit hard by smog. (n.d.). Al Jazeera. Retrieved December 8, 2023, from <https://www.aljazeera.com/gallery/2023/11/10/photos-like-breathing-poison-delhi-children-hardest-hit-by-smog#:~:text=Many%20have%20asthma%20or%20pneumonia>

‌Download Files | AirData | US EPA. (2015). Epa.gov. <https://aqs.epa.gov/aqsweb/airdata/download_files.html>

‌Air Quality Dataset. (n.d.). Www.kaggle.com. <https://www.kaggle.com/datasets/fedesoriano/air-quality-data-set>

UCI Machine Learning Repository: Air Quality Data Set. (n.d.). Archive.ics.uci.edu. <https://archive.ics.uci.edu/ml/datasets/Air+Quality>

Brown, D. (2001). The Ethical Dimensions of Global Environmental Issues. American Academy of Arts & Sciences. <https://www.amacad.org/publication/ethical-dimensions-global-environmental-issues>

US EPA, O. (2016, August 30). Air Data Basic Information. US EPA. <https://www.epa.gov/outdoor-air-quality-data/air-data-basic-information>

‌USA States to region. (n.d.). Www.kaggle.com. Retrieved January 13, 2024, from https://www.kaggle.com/datasets/omer2040/usa-states-to-region