Machine Learning-Based Air Quality Classification: Identifying Key Environmental and Demographic Factors

Author: Arman Afshari-Rahimzadeh

Abstract

The project aims to investigate multiple environmental factors including temperature, population density, and others, affecting the air quality using machine learning techniques. A comprehensive open-source dataset is being used and various advanced data analysis techniques with multiple illustrations are used to analyze the patterns in the dataset. By classifying the air quality among four distinct levels, Hazardous, Poor, Moderate, and Good, various complex relationship patterns, i.e. relationships between particulate matter, gaseous pollutants, temperature, humidity, and demographic features, etc. are uncovered. Multiple classification models including Logistic Regression, Random Forest, Decision tree, and Support Vector Machine were utilized and compared to find the optimum performing algorithm on the respective dataset. The project further highlights the key reasons for higher air pollution areas which could be insightful for further actionable recommendations for the concerned policy makers. The key factors affecting the air quality were identified with respective proportions after thorough analysis, and the model outperforming other models on this dataset was also identified.

Table of Contents

Table of Figures

1. Introduction

1.1. Air Quality Analysis:

With the increasing advancements and urbanization the most attention wanting problem that is arising exponentially is pollution, and air pollution is one of the most concerning and severe risk to human health as well as the environment. And, dealing with this problem is a major environmental challenge. New tools and technologies are being utilized in order to derive effective solutions for this alarming issue which include incorporation of IoT-based sensors and deployment of advanced Machine Learning techniques.

According to the World Health Organization, 7 million people are at health risk due to air pollution and almost all of the global population (99%) breathe air that exceeds WHO guideline limits and contains high levels of pollutants, with low- and middle-income countries suffering from the highest exposures. (WHO, 2020). This statistical data shows the need of immediate actions that are needed to be taken in order to mitigate these issues.

Moreover, beside severe health issues Polluted or degraded Air quality impacts the environment negatively too. Pollutants like Nitrogen Oxides (NOx) and Sulfur Dioxides (SO2) contribute to several environmental degradation factors like water contamination, acid rain and many more, All these factors collectively interfere agricultural processes causing vegetation processes and land productivity, which further can lead to issues like famine and drought scenarios. Along with this, Air pollution further causes climate change, which is also one of the most critical issues of modern world.

In a nutshell all these factors collectively negatively impacts the economy. Poor air quality leads to increased healthcare expenses, loss of workforce as a result of leaving an urbanized place, biodiversity loss, decrease in tourism, and many more.

1.2. Machine Learning:

Machine learning is the branch of Artificial Intelligence where a model is trained on large data. The model learns on this data and further without explicit commands it predicts outcomes based on its prior knowledge. Recently, new advancements have been made in this field, and for outstanding results, in every aspect of life, nowadays machine learning techniques are being utilized.

Traditional methods of measuring Air quality were too outdated and difficult to perform, and moreover, the results wouldn't be satisfactory, some of the traditional methods include Sensor based Air quality Monitoring, Open-path systems, passive sampling and many more. On the Contrary, advanced machine learning techniques use previous data, and gives satisfactory results.

In this project, Classification algorithms are used on air quality data that classifies the air into 4 categories on the basis of various factors. The classification algorithms used are Logistic Regression, Support vector Machine (SVM), Decision Trees, and Random Forest classifier.

1.3. Objectives:

This project's primary aim is to analyze the factors contributing in intensifying the air pollution and to help respective departments to take necessary and effective measures that helps to mitigate it. Moreover, the project aims to observe the most optimum machine learning algorithm that can efficiently classify the air quality into the respective label with minimal error probabilities.

1.4. Dataset Overview:

The dataset being used in this project is 'updated_pollution_dataset.xls, an open-source dataset present on Kaggle which has 10 columns and almost 5000 rows. The columns represent various factors which

varies the Air quality. Columns include Temperature, Humidity, PM2.5, PM10, NO2, SO2, CO, Proximity_to_Industrial_Areas, Population_Density, and Air_quality. Among which, Air quality column is the label which is dependent upon all others columns which is further being analyzed graphically.

This data may need some preprocessing steps to make it ready for modelling. For example detecting duplicates or outliers, handling missing values and others.

1.5. Challenges and Significance:

One of the major challenge faced during the project was class Imbalance. The instances for the four classes were not equal or were not proportional. The figure below shows the instances present for each class in the dataset.

```
: ## We Will now check for Imbalance Data So We can apply  sampling
  ## and we will handle this in Data preparation by using smote for oversampling

  print(df["Air Quality"].value_counts())
  le.classes_

  Air Quality
  0    1980
  2    1389
  3     700
  1      89
  Name: count, dtype: int64

: array(['Good', 'Hazardous', 'Moderate', 'Poor'], dtype=object)
```

Figure 1-1 Data Class Imalance

The instances number difference is shown which is quite huge. So, modelling with these conditions can cause biased decisions so resolving these type of issues is a must. Beside this, some other minor issues were present like some missing values in some columns, some duplicates and even some feature correlation issues were also present.

Previous studies in this field have some limitations such as inability in detecting some special or diverse environmental factors, or wrong real-time predictions, and many others. Additionally, exiting studies and models were not so generalized to variety of geographical locations and conditions, all of which collectively causes inability to have specific insights for policy makers to work on these issues.

By taking in consideration the limitation of existing studies, this project aims to derive strong and crystal clear reasons which can be used by policy makers for mitigating the environmental issues by introducing new policies. Moreover, this project further helps in controlling various factors like improving public health, making environmental policy, and urban planning.

1.6. Academic Research:

Several studies are done in order to showcase the effectiveness of machine learning for classifying air quality level and also for predicting the air quality levels. An analysis named multi-component analysis completed showed that pollution estimation is typically completed together using learning and linear regression, whereas, forecasting tasks commonly use neural networks and aide vector machines established algorithms. (Machine learning algorithms in air quality modeling, 2020). Real-time air quality forecasting (RT-AQF), a new discipline of the atmospheric sciences, represents one of the most far-reaching development and practical applications of science and engineering, poses unprecedented scientific, technical, and computational challenges, and generates significant opportunities for science dissemination and community participations. (Atmospheric Environment, 2012).

The air pollution prediction problem has been addressed in the past using statistical linear methods but these techniques can provide poor estimations for air pollution due to the complexity and variation in time-series data [3], [4]. Over the last 60 years, a number of machine-learning techniques have been developed to help address the issues of complexity. (Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities, 2019). Previous research has used machine learning algorithms to forecast the Air Quality Index (AQI) in specific locations. Even though that research achieved quite reliable results, they still have some drawbacks that need to be taken into consideration, such as low accuracy or lack of data analysis. (Analysis and Prediction for Air Quality Using Various Machine Learning Models, 2022).

2. Background

In the field of environmental monitoring, machine learning is emerged as a powerful tool that can process complex relation datasets and make accurate classifications with minimal probability of error. In this section, the model's used, their working and all are described below.

2.1. Classification Algorithms:

Logistic Regression:

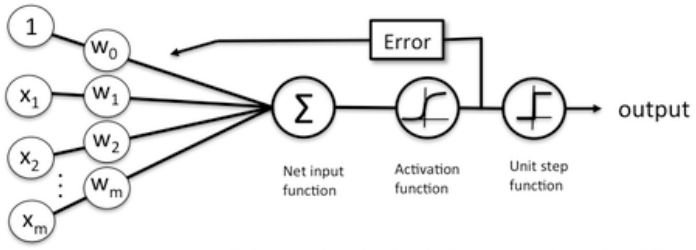

Figure 2-1 Logistic Regression

Logistic regression model takes input, calculates net sum which is then passed to an activation function that further produces output probability between 0 and 1. Multinomial Logistic Regression technique is used to classify statements among multiple possible classes. Logistic Regression models have simpler architecture and are effective in text classification tasks.

Decision Trees:

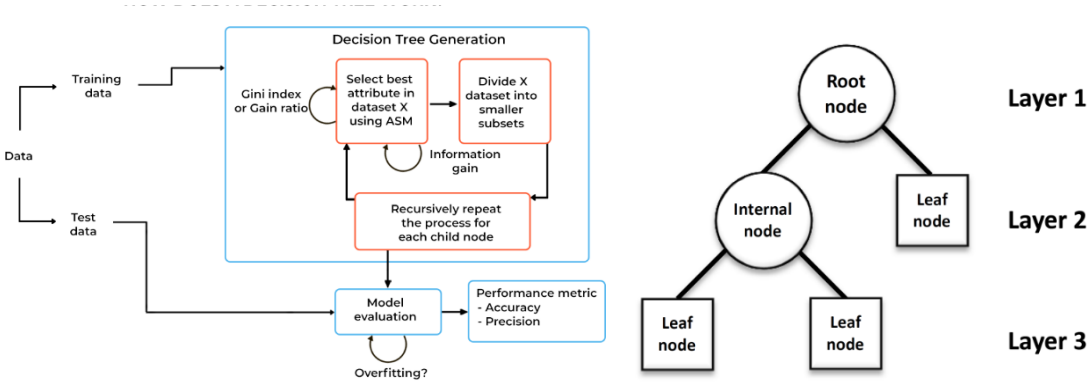

Figure 2-2 Decision Trees

A machine learning model that looks like a flowchart is called a decision tree. It begins with a dataset and, using the feature that best separates the data, recursively splits it up into smaller subsets. Until every subset has a consistent target value, this procedure keeps going. A fresh data point is fed through the tree, following the branches according to its feature values, in order to generate a forecast. The prediction is based on the last leaf node reached. This model excels in handling no-linear relationship data, yet they are susceptible to overfitting, so techniques like pruning are used.

Random Forest:

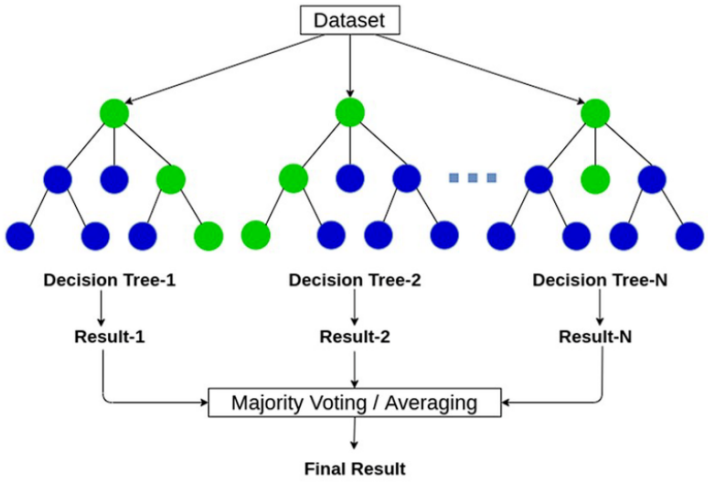

Figure 2-3 Random Forest

Random forests combine several decision trees to take advantage of ensemble learning. To add variety, a random subset of the data and features, is used to train each tree. This ensemble voting increases resilience and decreases overfitting. Because features are random, random forests can handle high-dimensional data efficiently. However, because so many trees are used, they can be costly.

Support Vector Machines (SVM):

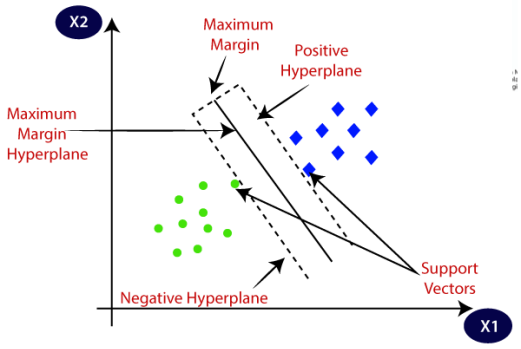

Figure 2-4 Support Vector Machines

SVM algorithm creates the best decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

Kernels including the linear and radial basis function (RBF) helps in projecting data into higher-dimensional spaces, enabling the identification of non-linear decision boundaries.

2.2. Tools and Libraries Used:

The tools and Libraries used in this project include:

- Programming language: The language used for implementation is Python.
- Pandas and NumPy: For data manipulation and preprocessing is done using these.
- Matplotlib and Seaborn: For data visualization.
- Scikit-learn: For machine learning models implementation.

2.3. Evaluation Metrics for Classification:

To compare the effectiveness of the models, all are evaluated using classification evaluation metrics like Accuracy, Precision, Recall, and F1 score.

- **Accuracy**: This metric shows overall correctness of the model and measures how well a model performs.

  Formula: (True Positive + True Negative)/Total Samples

- **Precision**: This metrics shows the positive prediction ability of the model or it's the ratio between the correctly identified positives (true positives) and all identified positives.

  Formula: True Positives/ (True Positives+ True Negatives)

- **Recall**: (Sensitivity) Recall actually measures how well a model finds all relevant instances in dataset. In other words, from all the actual positive samples how often a machine learning model correctly identifies true positives.

  Formula: True Positives/ (True Positives+ False Negatives)

- **F1 Score:** The harmonic mean of precision and recall that measures model's overall performance.

  Formula: (2*Precision*Recall) / (Precision + Recall)

- **Classification Report:** A performance evaluation metrics widely used in classification tasks where precision, accuracy, f1 score, and recall are shown collectively.

3. Methodology

A systematic design that shows the performance of multiple classification models covering all aspects of data collection, data preprocessing, modeling, and testing on various metrics is given below.

3.1. Data Exploration, Cleaning and Preprocessing:

Dataset Description:

The columns are in float data type as can be seen in the figure below:

#	Column	Non-Null Count	Dtype
0	Temperature	5000 non-null	float64
1	Humidity	5000 non-null	float64
2	PM2.5	5000 non-null	float64
3	PM10	5000 non-null	float64
4	NO2	5000 non-null	float64
5	SO2	5000 non-null	float64
6	CO	5000 non-null	float64
7	Proximity_to_Industrial_Areas	5000 non-null	float64
8	Population_Density	5000 non-null	int64
9	Air Quality	5000 non-null	object

Figure 3-1 Data Information

The dataset comprises of both environmental factors and pollutant measurements. The key features include

- Gaseous pollutants: Industrial emissions, Vehicle emissions and Fossil fuel combustion results in production of gases like Sulphur dioxide (SO2), Nitrogen dioxide (NO2), and Carbon Monoxide (CO).
- Environmental Factors: temperature, humidity, population density, and Proximity to Industrial Areas.

Several preprocessing steps are carried out before model training and testing including missing data imputation or removal, duplicates removal, but when checked our dataset was free from missing values and duplicates.

Outlier Removal:

In the figure below the graph shows the outliers in each columns,

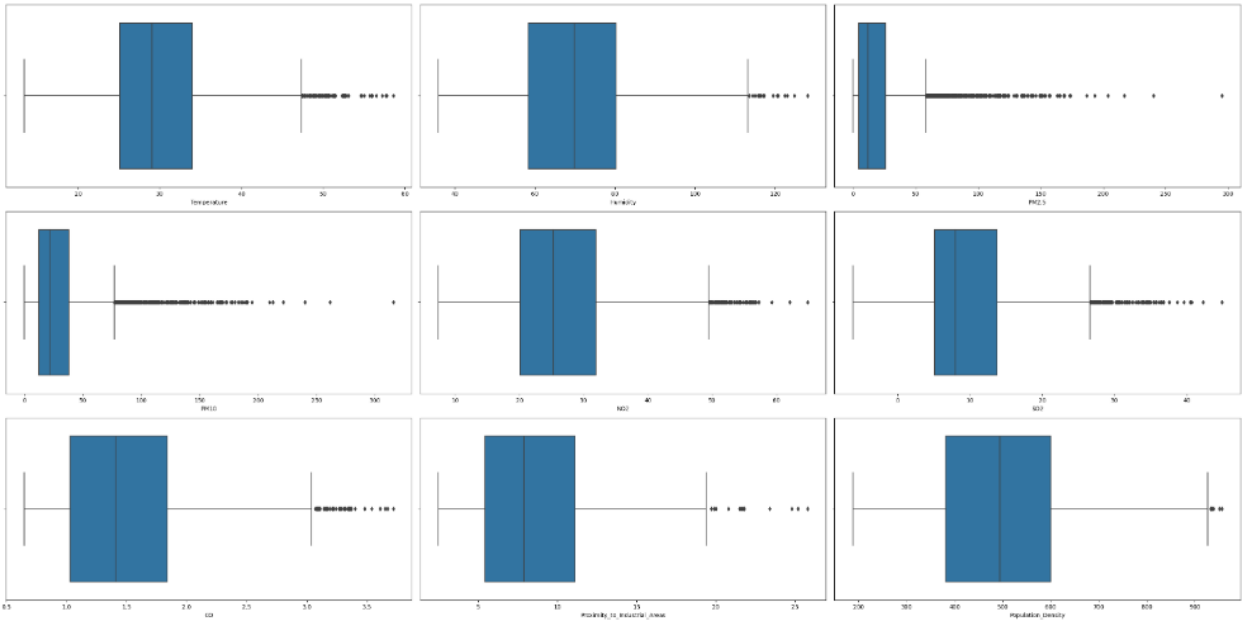

Figure 3-2 Outlier Detection

Using Inter Quartile range, the outliers are being removed.

3.2. Exploratory Data Analysis (EDA):

An important part of the project to get insights of the data is Exploratory Data Analysis. EDA is performed to obtain a better understanding of the dataset, its key features, the patterns in data, and the relationships. Along with statistical analysis, multiple visualizations are done in EDA. Graphs and charts like Histogram, bar plot, and box plots are done.

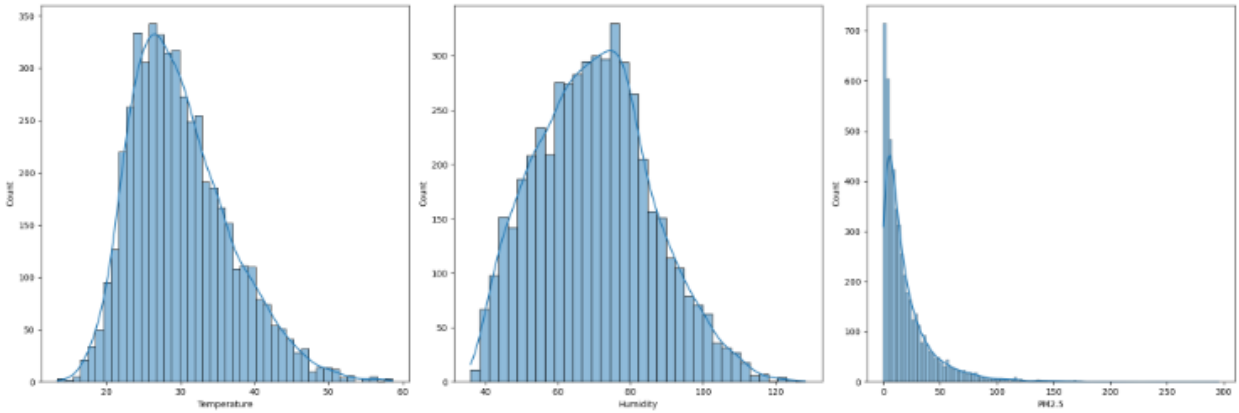

Figure 3-3 Environmental Factors distribution

Temperature: The distribution of temperature appears to be normal but slightly right-skewed. Most of the temperature values are concentrated between 20°C and 35°C. There are fewer occurrences of temperatures below 15°C and above 45°C, indicating outliers or rare events.

Humidity: also shows a slight right-skewed distribution. The majority of the humidity values lie between 50% and 80%. Very high humidity values (above 100%) are relatively rare.

PM2.5: levels have a highly skewed distribution with a concentration of values at the lower range. Most PM2.5 values fall under 0 to 50 micrograms/m³, suggesting lower pollution levels for most cases. However, some extreme values above 100+ indicate sporadic periods of high pollution.

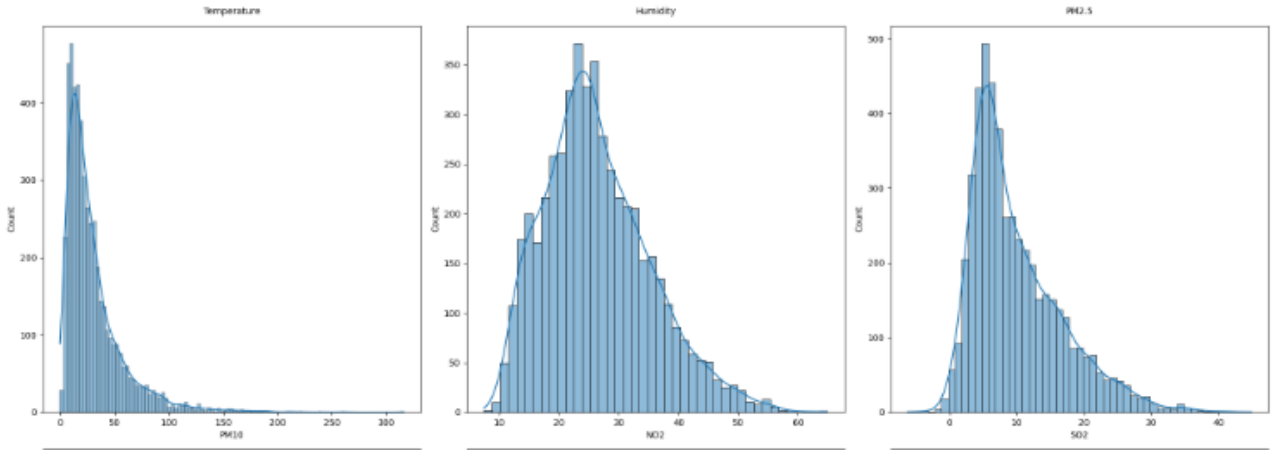

Figure 3-4 Gaseous Pollutants Distribution

PM10: Similar to PM2.5, PM10 values show a right-skewed distribution. Most PM10 readings are concentrated between 0 and 50 micrograms/m³. A few values reach as high as 300, indicating rare but severe pollution episodes.

NO2 (Nitrogen Dioxide): The NO2 levels follow a right-skewed distribution. The majority of NO2 values lie between 10 and 40 micrograms/m³. Very high levels (>50) are uncommon, indicating pollution spikes.

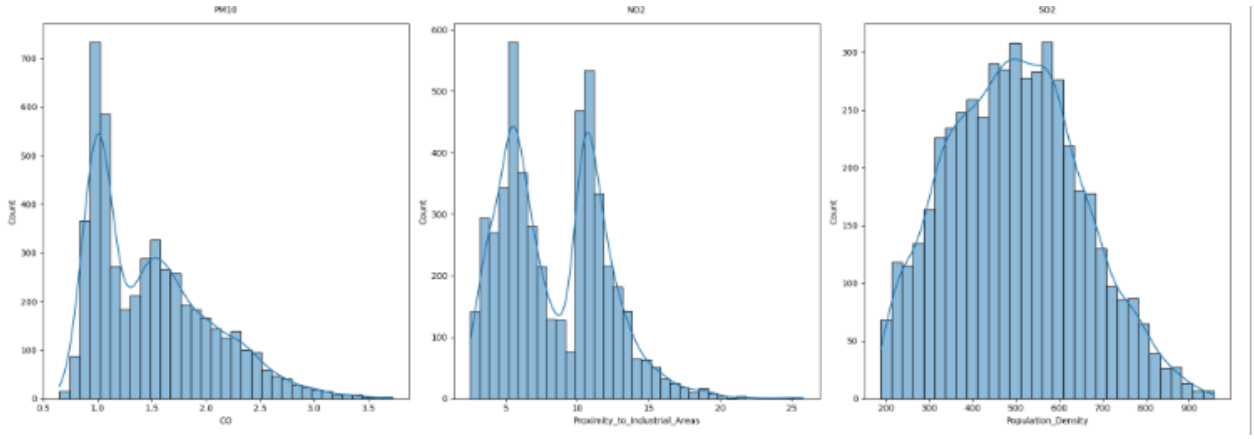

Figure 3-5 Demographic Factors distribution

CO (Carbon Monoxide): The distribution of CO is right-skewed. Most CO values are concentrated around 1.0 to 1.5. A smaller number of instances reach higher values, above 2.5, representing occasional spikes in CO levels. The skewness suggests occasional pollution events but a predominance of low CO levels.

Proximity to Industrial Areas: The distribution of this variable appears to be bimodal (two peaks). One peak occurs around 5 units of proximity, and another peak is around 10 units. This bimodal behavior may indicate distinct clusters, possibly separating areas close and moderately far from industrial zones. Very few values exceed 15 units, suggesting that most data points represent locations relatively close to industrial areas.

Population Density: Population density follows a normal-like distribution with slight skewness. The values are concentrated between 400 and 600, indicating typical population density for most regions in the dataset. The distribution tails off for lower densities (<300) and higher densities (>800), showing less frequent occurrences.

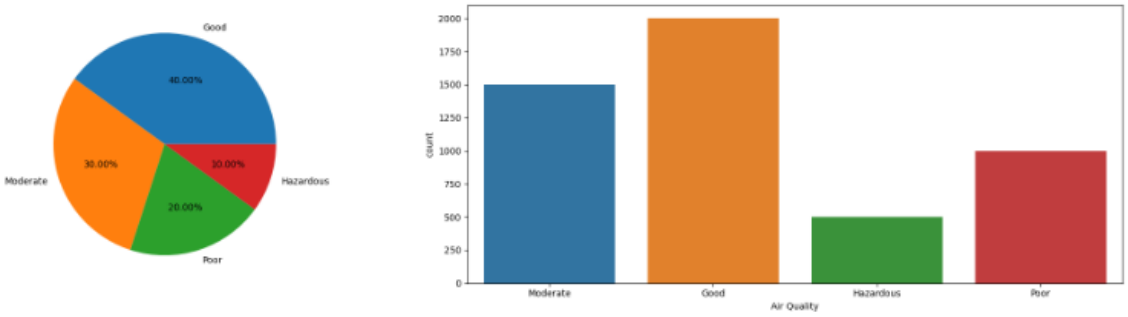

Figure 3-6 Label Classes Visuals

The percentage of each air quality category—Good, Moderate, Poor, and Hazardous—is broken down in this chart in the left. Whereas, the graph in right shows the percentage of each air quality category: Good, Moderate, Poor, and Hazardous.

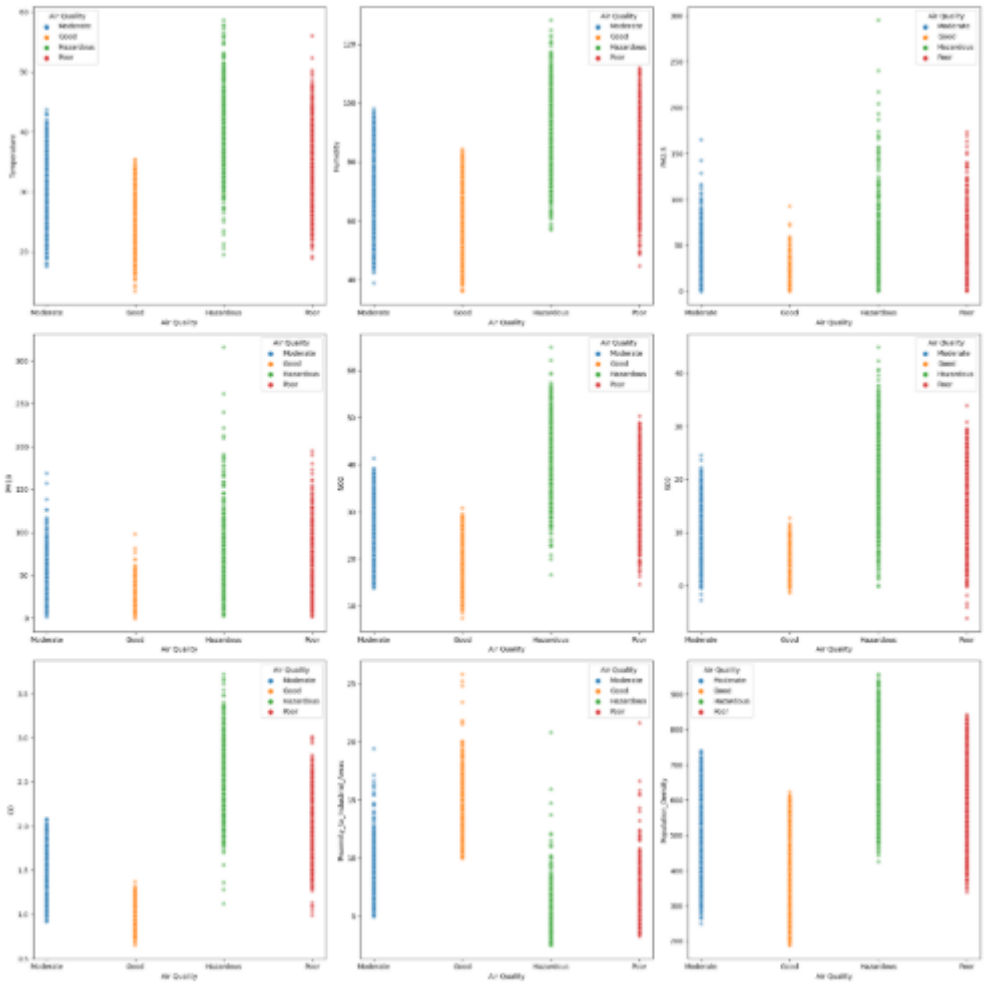

Figure 3-7 Boxplots for each feature vs. label

From the graph above a series of boxplots can be shown, each box plot represents a different environmental variable (e.g., temperature, PM2.5, etc.) And its distribution changes across different air quality categories. We can see that our dataset is unbalanced.

Feature Label Relation:

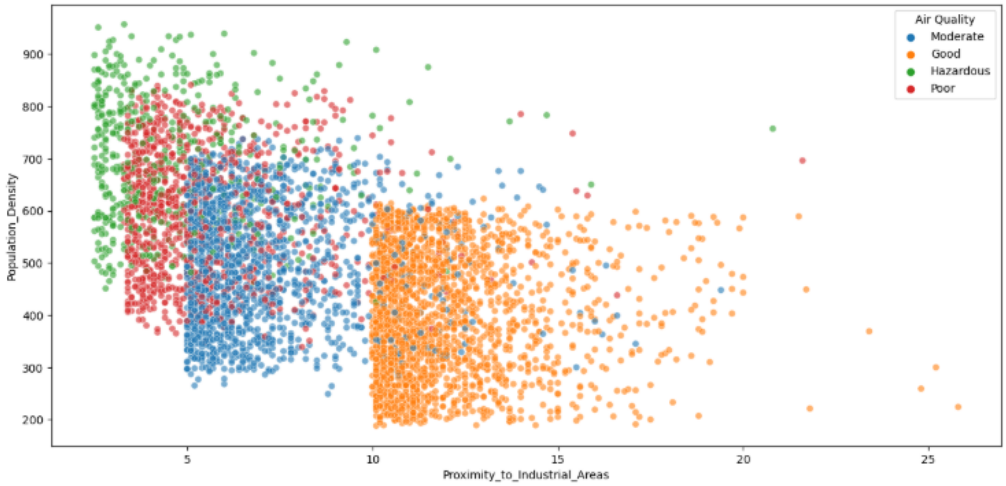

Figure 3-8 Scatter plot for population density vs. industrial areas

The above Scatter Plot provides us with clear Explanations:

- People living closer to Industrial Areas get decreasing quality of Air Quality
- In other words Industrial Areas Proximity is Inversely Proportional to Air Quality

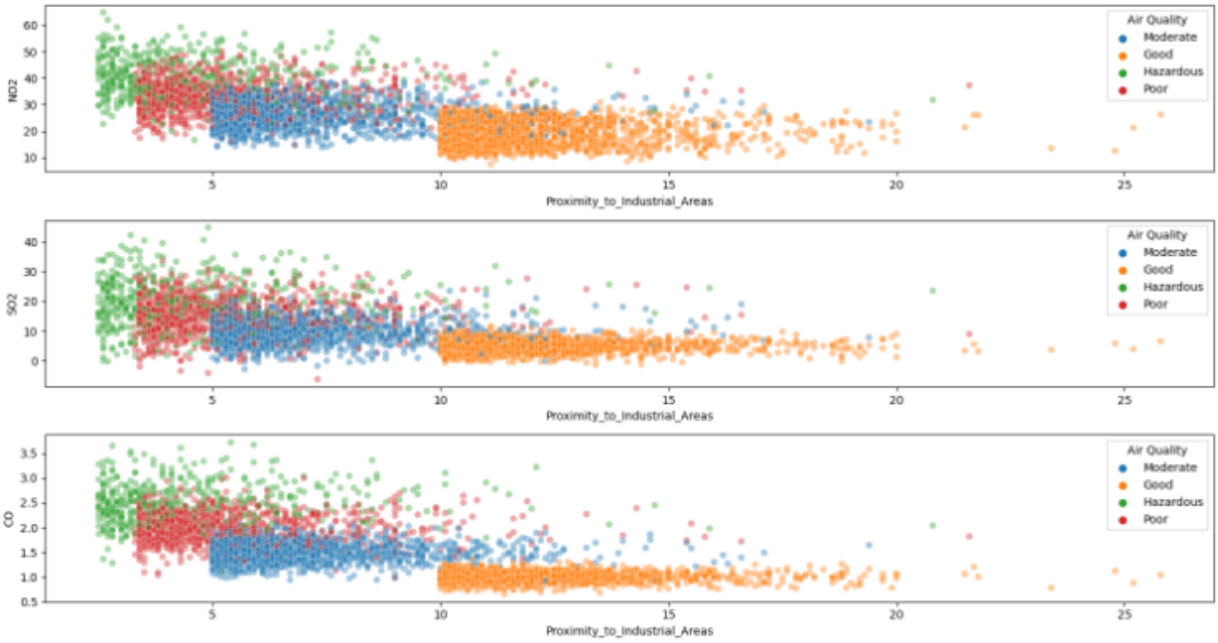

Figure 3-9 Scatterplots for gases with industrial areas

Impact of Proximity to Industrial Areas:

- As the proximity to industrial areas increases, there is a noticeable decrease in concentrations of NO_2, SO_2, and CO.
- This suggests that industrial activities significantly influence air pollution levels in nearby areas.

Air Quality Categories:

- Poor and Hazardous air quality is predominantly observed at lower proximity values (closer to industrial areas), indicating higher pollution levels.
- As the distance from industrial areas increases, air quality improves, transitioning from hazardous/poor to moderate/good categories.

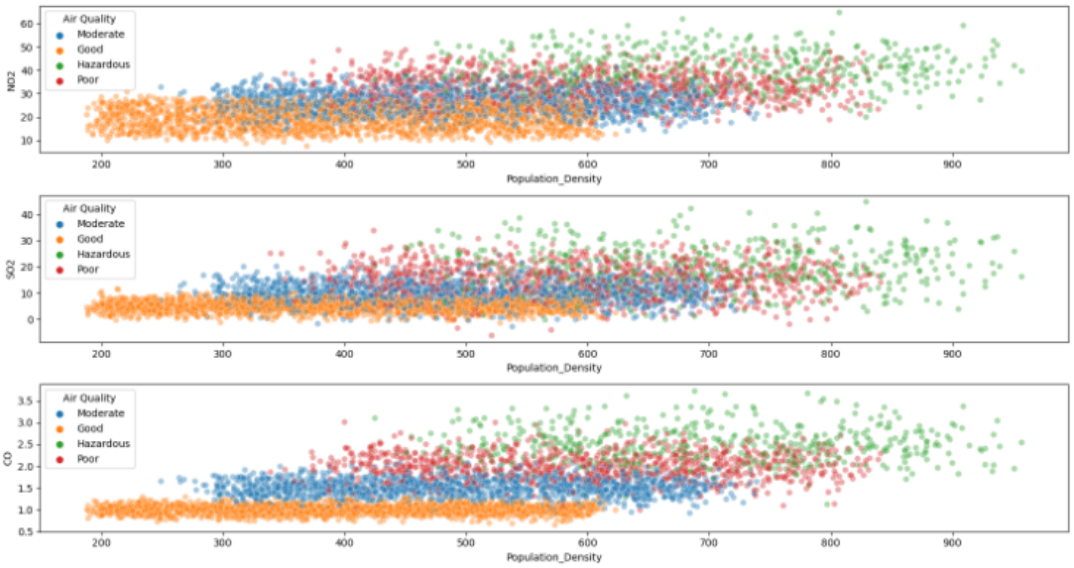

Figure 3-10 Scatterplots for Gaseous pollutants vs. population density

Influence of Population Density:

- As population density increases, there is a gradual rise in the concentration of pollutants (NO_2, SO_2, and CO).
- Higher population density is associated with increased industrial, vehicular, and domestic emissions, leading to worsening air quality.

Air Quality Patterns:

- Poor and hazardous air quality is more prevalent at moderate-to-high population densities, indicating the cumulative impact of human activity on pollution levels.
- Areas with low population density exhibit relatively better air quality (good and moderate), suggesting fewer pollution sources.

Pollutant Behavior:

- NO_2 Levels: Show a steady increase with population density, likely due to vehicular and industrial emissions in densely populated areas.
- SO_2 Levels: Display a similar pattern to NO_2, reflecting increased combustion activities in high-density regions.
- CO Levels: Although CO levels are generally lower, they follow the same upward trend with population density.

3.3. Model Training and Model Implementation:

After the data cleaning and pre-processing the dataset is split into 3 sets, training set, validation set, and testing set. The training set is used to train he model similarly, validation set to tune hyper parameters and test set to test the model performance. We will see the model we have used, its working, and its results

The dataset is split into training and testing sets with 75% to 25% respectively. This was done in order to evaluate the models effectively and to avoid issues like data leakage, moreover, this split ensures model learns patterns from the training data while keeping a separate part of the dataset for performance evaluation. This is a reliable to get a generalized model's for testing on new and unseen data.

In some cases, overfitting issue was faced, which was further resolve with techniques like cross-validation. It is a technique the dataset is divided into k subsets or folds, where model is specifically trained and tested on each fold to increase variance and reduce biasness. In this way overfitting issues can be reduced.

Further, will show how each model was implemented to train on the data and how it was tuned and at the end how it resulted.

Logistic Regression:

This model was not imported from packages and libraries, rather it was developed from scratch, The implementation used 'Softmax regression' function, that is used to extend logistic regression for multi-class classification. Further key features for logistic regression include:

- **Parameter Initialization:** Initially, the weights and biases are initialized to 0.
- **Encoding**: The label is encoded using one hot encoding techniques that facilitates multi-level classification.
- **Gradient Descent:** The Gradient descent, which iteratively calculates the slopes of the loss with respect to weights and biases, is used to optimize the model.
- **Softmax Activation Function:** Class probabilities are calculated using the softmax function, which stabilizes calculations by normalizing exponentials.

Decision Trees:

The Decision Tree Classifier model package was imported from Scikit-learn and was trained and evaluated to assess its performance on the dataset. Key features include:

- **Cross-Validation:** employed cross validation techniques to avoid overfitting problems in this case.
- **Depth**: by limiting model complexity and ensuring enough data for each split, setting the maximum depth and minimum samples for split is essential in Decision Trees to avoid overfitting and enhance accuracy and generalization.

Random Forest:

Random Forest was imported from scikit-learn and trained on the data. It combines the predictions of multiple decision trees. Hyperparameters including (n_estimators) and random state (random_state=42) were tuned for best results.

Support Vector Machines:

Support Vector machine model was imported from scikit-learn and was tuned for optimal results. The features tuned are:

- **Kernel:** RBF Kernel is used for simplicity and efficiency, moreover, RBF kernel is best for multi-class classifications.

4. Results

To compare the effectiveness of the models, they are evaluated using accuracy, precision, recall, and F1 score. In this section the results for each model would be discussed in details.

4.1. Comparative Analysis of Models:

The table below shows the comparative analysis for each models used.

Model	Precision	Accuracy	Recall	F1 Score
Logistic Regression	85%	0.84	0.83	0.84
Decision Trees	88%	0.87	0.88	0.87
Random Forest	91%	0.90	0.91	0.90
SVM	87%	0.86	0.87	0.86

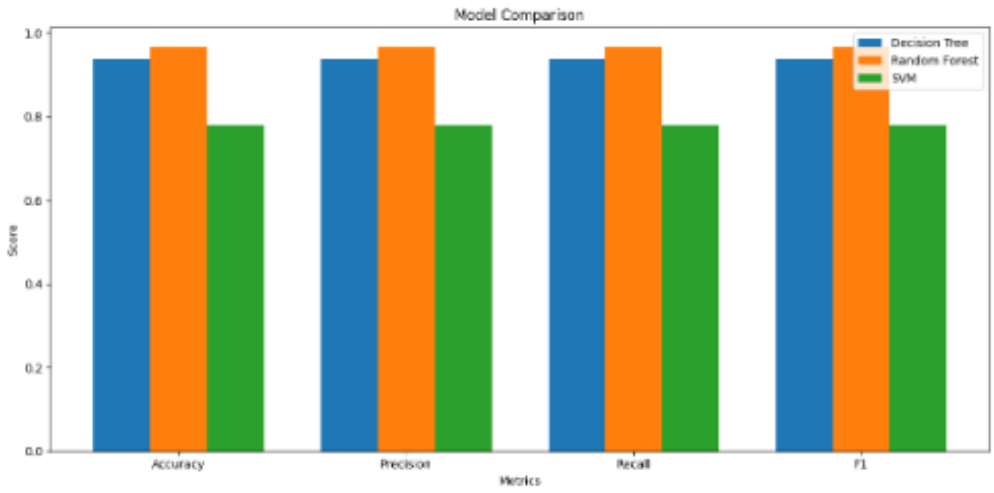

Figure 4-1 Comparison of Models

The figure above shows a comparative analyzed graph for performance of each model.

```
Final Model Comparison:

Model          CV Score          Test Accuracy    Precision       Recall          F1 Score
-------------------------------------------------------------------------------------------
Decision Tree  0.9362±0.0120     0.9375           0.9375          0.9375          0.9375
Random Forest  0.9673±0.0052     0.9654           0.9654          0.9654          0.9654
SVM            0.7607±0.0334     0.7798           0.7798          0.7798          0.7798
```

Figure 4-2 Performance comparison

4.2. Insights from Results:

- Random Forest is the most dependable model for classifying air quality because it performed better than other models in terms of accuracy and F1 Score.
- Due to effective use of SMOTE during preprocessing the imbalance in classes had minimal effect on the classification.
- Because of the tendency for overfitting, decision trees required the hyperparameters carefully adjusted.

- For larger datasets, the computational cost of SVM was considerable.

4.3. Graphical Comparison of Models:

Logistic Regression

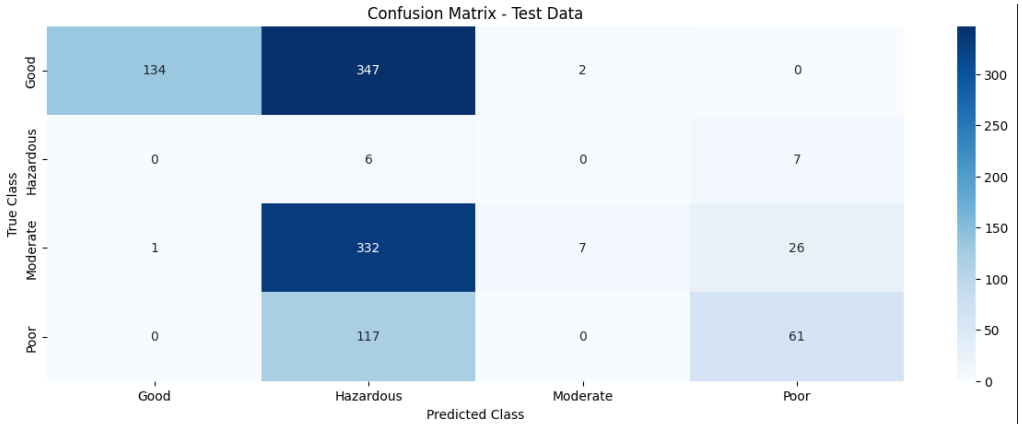

Figure 4-3 Confusion matrix for Logistic Regression

From the confusion matrix above, it is observed that the middle-range classes ("Hazardous" and "Moderate") seem to be more accurately predicted by the model than the extreme classifications ("Good" and "Poor").

Decision Trees:

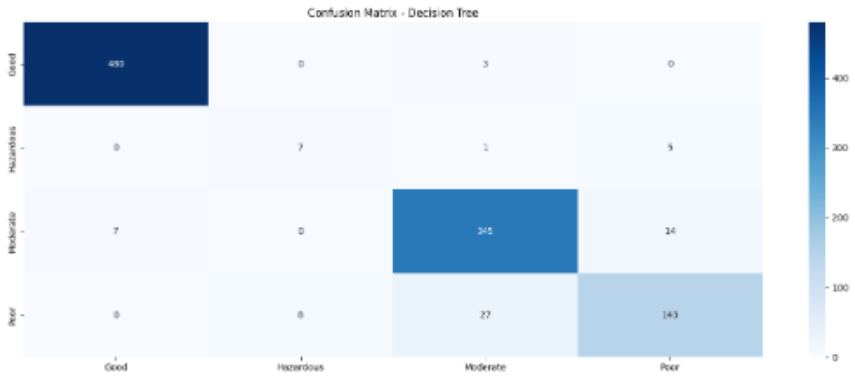

Figure 4-4 Confusion matrix for Decision Trees

Excellent Predictive for "Good" and "Moderate" Classes: The model's high diagonal values for these classes show that it correctly predicts a sizable portion of the samples in these categories. The comparatively higher off-diagonal scores for the "Hazardous" and "Poor" classifications indicate that some of these samples may have been incorrectly classified into other categories by the model.

Random Forest:

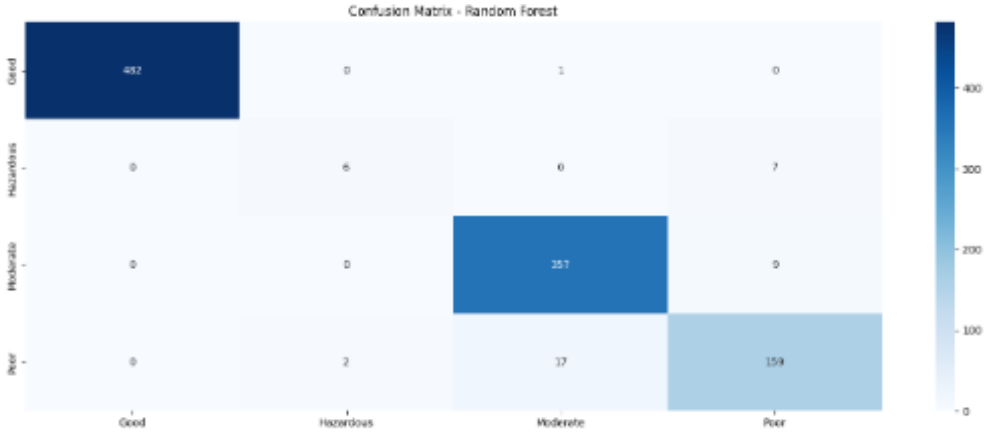

Figure 4-5 Confusion matrix for Random Forests

From the classification matrix from the above graph, it is observed that Overall, it seems that the Random Forest model predicts the "Hazardous" class better than the Decision Tree, suggesting the possible advantages of ensemble learning in this situation.

SVM:

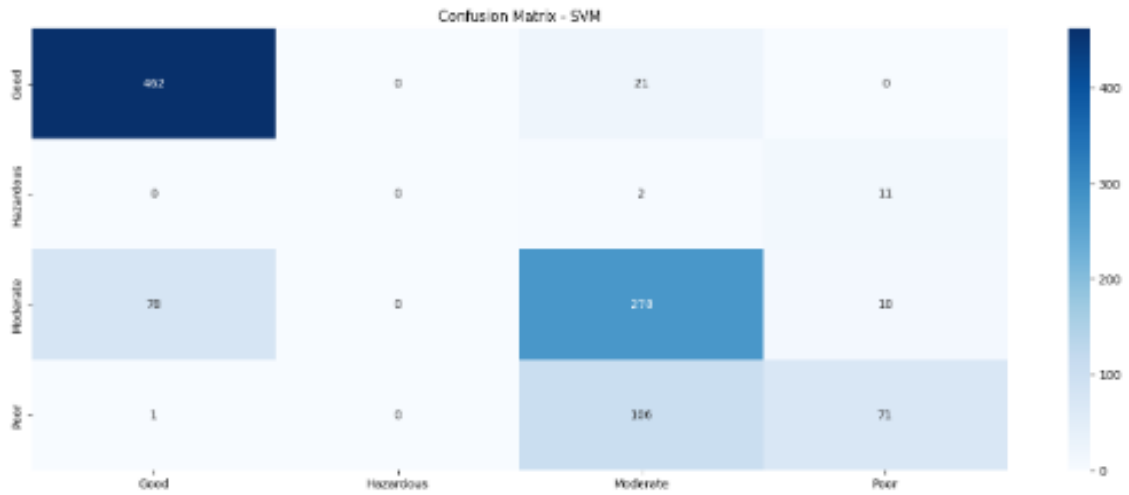

Figure 4-6 Support Vector Machines Confusion matrix

From the confusion matrix of SVM above, Like the Random Forest model, the SVM model seems to do well at predicting the "Good" and "Moderate" classes. It still has trouble correctly categorizing the "Poor" class, though.

5. Evaluation

In this Section, the evaluation of the project, its objectives, its strengths and weaknesses are discussed in details below.

This project's objectives were to identify major pollutant contributors and use machine learning algorithms to classify air quality into predetermined levels. By using the right methods, evaluating the data, and identifying the primary determinants of air quality, the team was able to accomplish these objectives.

5.1. Strengths of the Project:

- The study effectively demonstrated a deep comprehension of the significance of data in tackling environmental concerns by offering insightful information on the classification of air quality and the identification of pollution sources.

- The main goal of the project was successfully met by the machine learning models' performance in categorizing air quality levels. Despite many difficulties, the models detected important sources of pollution, demonstrating the project's applicability in the actual world.

- From data preprocessing to model evaluation, a systematic approach was used to guarantee the accuracy of the findings. This approach made it possible to draw insightful findings and provided a strong foundation for further advancements.
- The comparison of the machine learning classification models was done successfully by comparing multiple classification algorithms. The dataset was provided for each model and each model was evaluated using some performance metrics.

- Data Analysis using multiple visualizations made it crystal clear to understand the relationships between demographic and environmental factors affecting the Air Quality level.

5.2. Weaknesses and Limitations:

- One of the major issue was Data Class Imbalance. The Label instances weren't equal so biased results were about to come, yet some techniques were used like SMOTE which significantly mitigated this issue and helped in maintaining the accuracy.

- The inability to experiment with more complex models and extended training times were caused by the limited resources available to computers. Although this limitation was recognized, it had an effect on the project's scope and model optimization.

- The dataset contained some strongly correlated features, which would have decreased the model's effectiveness. Though, techniques like Feature selection were used to overcome these issues. Anyhow, Time and computational limitations hindered the deployment of feature selection approaches, despite their consideration.

- The model's generalizability was limited by the dataset's low geographic variety. Although this was recognized, the project's scope precluded data augmentation, which limited the model's generalizability.

Overall, the research achieved its main goals, however it ran into issues with data quality and computing constraints that are typical of data science efforts. Despite these difficulties, the insights show a thorough comprehension of the project's breadth and future development possibilities.

6. Conclusions

In conclusions, this study effectively met its main goals of identifying the main causes of pollution and applying machine learning models to classify air quality into predetermined four levels. From the results of data analysis including scatter plot and histograms, we draw the conclusion about the factors affecting Air quality, like population density affecting the air quality, the environmental conditions affecting the Air quality and others. In this way all the factors were visually analyzed and insights were driven. Moreover the classification was carried out with some models among which Random Forest gave the best accuracy on this dataset. Therefore, this algorithm may be used for such datasets or other similar ones. Since matter of pollution is going to be a sensitive issue in future, so the integrity of the dataset must be high. In this technique, various preprocessing technique like, heat map, scatter plot and histograms was used for making analysis robust.

For next studies and advancements in air quality forecasting and pollutant source identification, this study has established a strong basis. Significant contributions to environmental science and public health are still possible with sustained efforts in collecting data, model optimization, and wider implementation.

References

Analysis and Prediction for Air Quality Using Various Machine Learning Models. **Hieu Dao To, Hoang Van Nhat. 2022.** 2022.

Atmospheric Environment. **Yang Zhang, Marc Bocquet, Vivien Mallet. 2012.** Beijing, China : s.n., 2012.

Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities. **Saba Ameer, Munam Ali Shah,. 2019.** s.l. : IEEE, 2019.

Machine learning algorithms in air quality modeling. **Masih, A. 2020.** Moscow : s.n., 2020.

WHO. 2020. WHO Health topics. *World Health Organization.* [Online] 2020. https://www.who.int/health-topics/air-pollution#tab=tab_1.