ASSESSMENT 1

|  |  |
| --- | --- |
| MODULE CODE | COM7003 |
| MODULE TITLE | Artificial Intelligence |
| MODULE LEADER | Xin Lu |
| ASSESSMENT TITLE | Technical Report |
| WEIGHTING | 70% |

Air Quality Classification Using Machine Learning Models.

Submitted by:

Attaullah Khan Afridi (MSc Data Science and AI 2025-2026)

Student ID no: 2414297

# Abstract

This report presents a complete details about a machine learning initiative which uses environmental and demographic indicators to perform air quality classification. Developing a dependable predictive model stands as the main goal to enable environmental monitoring and public health decision support. The data preparation process featured a complete pipeline that handled outliers then performed encoding followed by scaling. Random Forest Classifier together with Support Vector Machine and Gradient Boosting Classifier were chosen from among the candidate models for optimization through RandomizedSearchCV cross-validation. The Random Forest model demonstrated the best performance regarding accuracy and generalized well. The report includes an extensive exploration of ethical factors including bias assessment together with fairness analysis and societal implications assessment. The analysis ends by introducing possibilities for future development along with plans to incorporate explainable AI systems. Machine learning demonstrates its potential for building sustainable ethical environmental decision support systems through findings verified by academic literature and best practices.

# 1. Introduction

## Problem Definition

Great public health threats involving the aspect of air pollution confront the whole world, considering that industrial areas and densely populated places are at stake. WHO (2021) estimates that nearly 7 million deaths each year are caused by exposure to pollution prematurely. As growing urban population along with the growing industrial activities necessitate the deployment of such intelligent systems for active air quality level monitoring and assessment.

The research is to build a machine learning predictive model based on the environmental data as well as demographic characteristics to decide the air quality status level. This model gives indispensable data for the formation of auspicious defensive methods for governmental and environmental institutions, thus aiding them to make more effective choices. According to Zhang et al. (2018), correct application of predictive modeling allows for the development of warning systems for health protection and sustainability ability.

This will be accomplished by providing a solution categorized the air quality system through labeling system which labels with “Good, Unhealthy, Moderate” according to NO₂, SO₂, CO pollutant levels and population concentration & industrial proximity measures. It is reported every stage from the data preparation stage till the stage of selecting the model and optimize the model then stage of evaluation along with the analysis of the ethical concerns.

## Dataset Description

The dataset used in this project is the **Updated Pollution Dataset**, sourced from **Kaggle**. The dataset contains 5000 instances and 10 features in CSV format. It includes several key features, such as **Date** (the date of data collection), **PM2.5** (the concentration of fine particulate matter in µg/m³), **Temperature** (ambient temperature in °C), **Humidity** (percentage of humidity), and **Air Quality** (categorical indicator of air quality, with levels such as Good, Moderate, Poor). The target variable in this dataset is **Pollution\_Level**, which classifies the pollution level based on PM2.5 concentration into categories like Low, Medium, and High.

The features used in this project are various environmental and demographic features in a region that impacts air quality levels. Features contain temperature (°C), relative humidity (%) and concentration of different pollutants such as PM2.5, PM10, NO₂, SO₂ and CO. Furthermore, the dataset also includes proximity (km) to industrial areas as well as population density (people /km²) as key determinants of pollution exposure. Air Quality Level is a target variable whose categories are Good (clean air with minimal pollution), Moderate (acceptable air with some pollutants), Poor (polluted air that may affect sensitive groups), and Hazardous (severely polluted air posing serious health risks).

# 2. Data Preparation and Preprocessing

The key step in developing an accurate machine learning model is to clean and process the data. To obtain correct machine learning model manufacturing results, it is necessary to process the data correctly. Kuhn & Johnson (2013), the work was thoroughly cleaned and examined using their practices.

Checked the missing values if the dataset using df.is\_null().any(). Fortunately, no imputation was required.

Continuous variables (pollutant concentrations) were used to detect the outliers using Interquartile Range (IQR) method. It was an important step to avoid distortion to model performance.

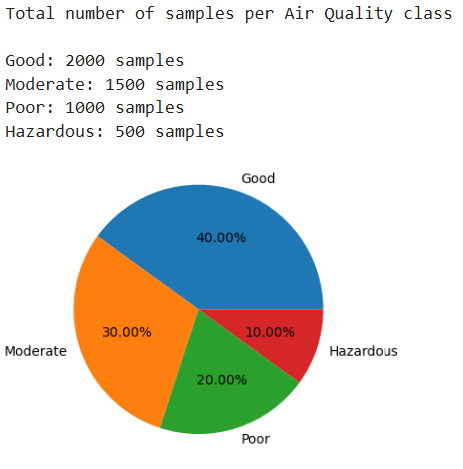
The dataset was split into training (75%) and test (25%) sets using stratified sampling to preserve the distribution of classes, ensuring unbiased evaluation.

# 3. Exploratory Data Analysis (EDA)

Through Exploratory Data Analysis, the data is first analyzed by initial analysis of the data, by which vital relationships and important information of the collected data are unveiled. Various visual techniques were applied:

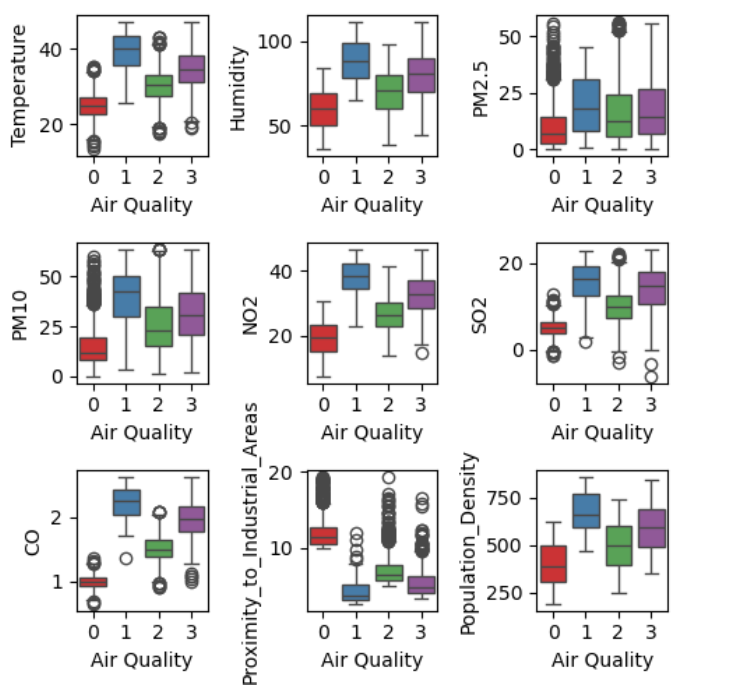
## Histograms and Pie Charts:

These enabled analysis of how propensity distribution of pollutants was separated by population density across the whole data balance.

****

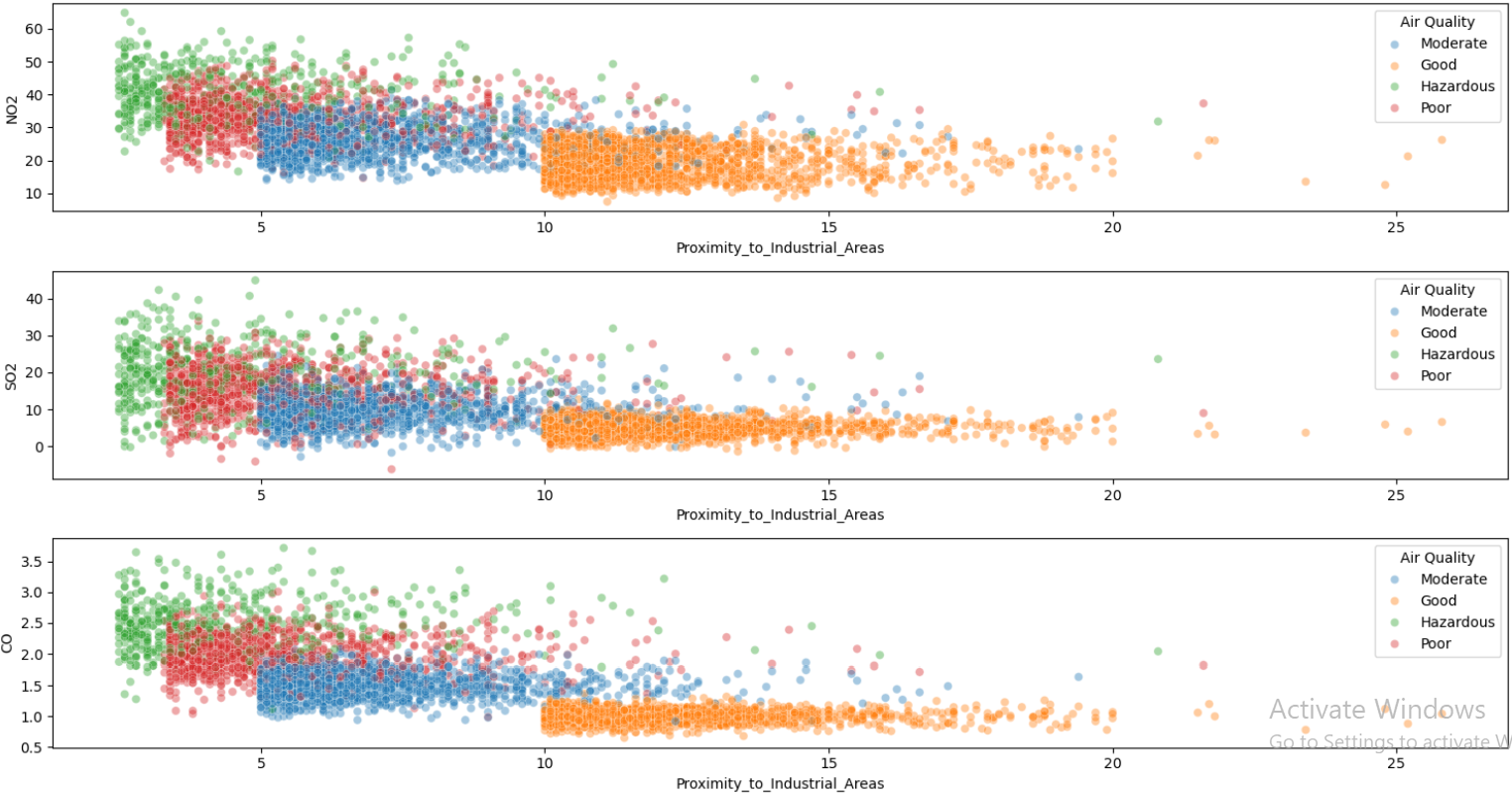
## Boxplots:

Box plot ilustrated the spread and central tendency of pollutants gasess, which facilitated me in the outlier detection.



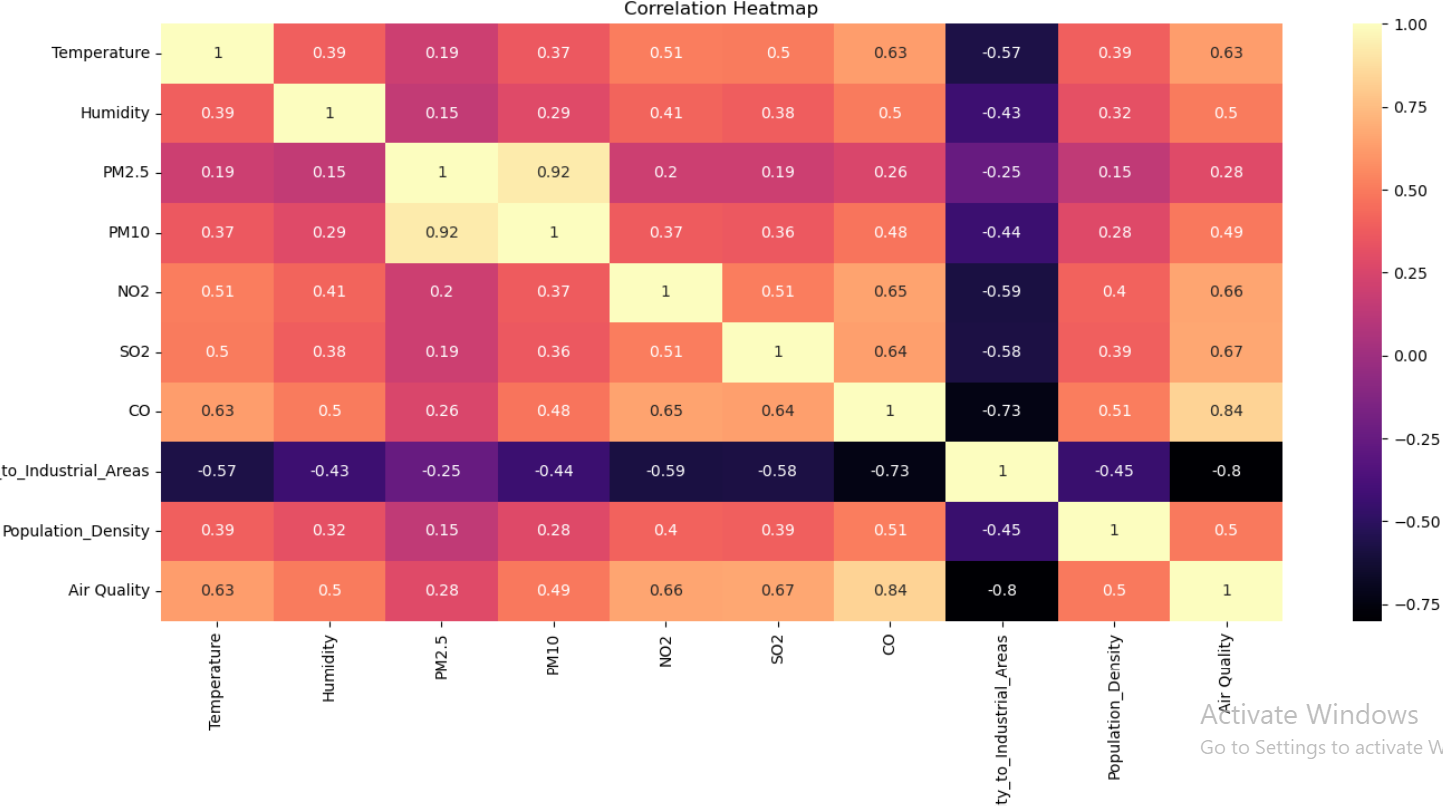
## Scatterplots:

Scatter plott Show the relationships b/w variables such as population density and air quality.

****

Correlation Heatmap:

Features aligned to see if it has feature collinearity. Both the poorer air quality and higher NO₂ and SO₂ levels were found to be strongly correlated with each other as shown in the below graph.

****

Results of EDA showed that the most influential factors were pollutant levels and sizes of served population on the air quality classification.

# 4. Feature Engineering

Feature engineering improves model accuracy and helps to provide interpretable model. The key feature engineering steps included in this project are:

## Feature Selection:

High correlated and domain relevant retained features. Principal Component Analysis (PCA) type dimensionality reduction techniques were not used for explain ability because they keep the exact lost dimensions.

## 

## Data Scaling:

For SVM, scaling was done although it is not necessary for tree based models to ensure consistent magnitude of feature.

For SVM, scaling was done although it is not necessary for tree based models to ensure consistent magnitude of feature.

# 5. Model Selection and Justification

To ensure robust model performance we selected and compared five diff machine learning algorithms which are:

## 

## **Random Forest Classifier (RF):**

An ensemble model of decision trees. First, given it is robust to noise, interpretable and outperforms on tabular data.

## **Support Vector Machine (SVM):**

Effective for high-dimensional data and small-to-medium-sized datasets. SVMs based on kernel are very powerful in the problem with complex boundary.

## **K-Nearest Neighbors (KNN):**

A simple but an effective non-parametric model. Included to demonstrate baseline performance. KNN is conceptually simple however it is sensitive to scale and balance of classes.

## **Gradient Boosting Classifier (GBC):**

Dataset has imbalance between classes (e.g., more 'Good' samples than 'Hazardous'), GBC can still perform well by focusing on hard-to-classify samples during boosting iterations. A selected boosting based ensemble method which achieves good accuracy (reduced bias and variance).

## **XGBoost Classifier (XGB):**

An advanced gradient boosting framework that is fast and, indeed, very performant. Given data is structured, it supports fine grained regularization.

The theoretical suitability, computational efficiency and empirical testing on the dataset were used for model selection.

# 6. Model Optimization Techniques

## **RandomizedSearchCV:**

* Used for hyper parameter tuning across all models.
* Explored combinations of key parameters such as max\_depth, learning\_rate and n\_estimators etc

## **KNN Optimization:**

* Applied feature scaling and weighted distance using standard scalling and fit transform methods and function.
* Tuned hyper parameters like n\_neighbors, p (distance metric, eucladean and manhattan), and weights.

## **XGBoost Optimization:**

* Tuned Hyperparameters like learning\_rate, max\_depth, gamma, subsample, colsample\_bytree, min\_child\_weight.
* Early stopping applied to prevent models from overfitting.

## 

## **Gradient Boosting Optimization:**

* Tuned n\_estimators and learning\_rate using RandomizedSearchCV.

# **7. Model Evaluation and Results Discussion**

## **Model Evaluation Strategy:**

* Evaluated models using 5-fold cross-validation.
* Input variations: Accuracy, Precision, Recall, F1 Score.

## **Confusion Matrix and Classification Report Insights:**

### Random Forest:

### 

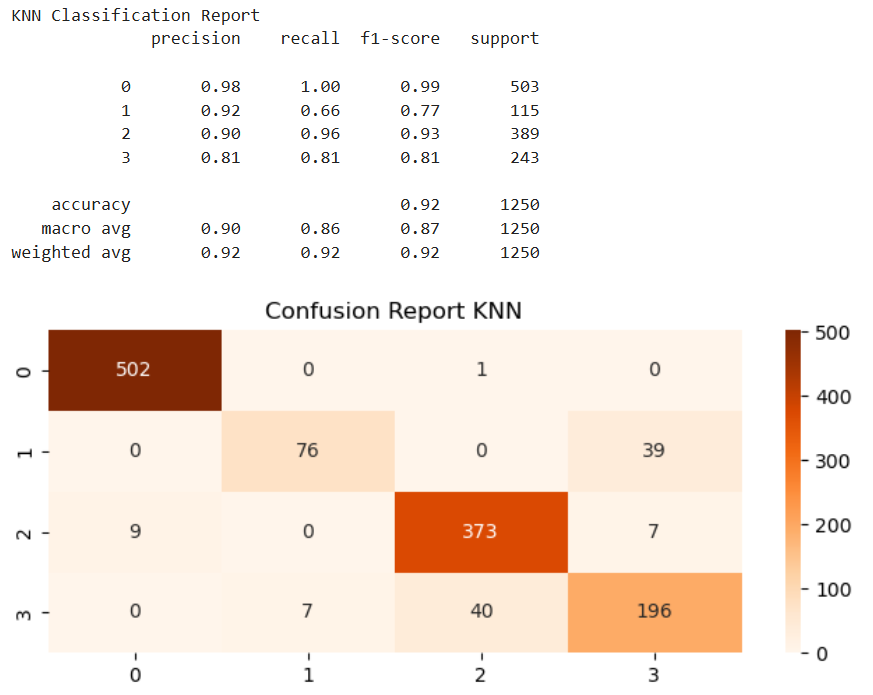
### Support Vector Machine:

### 

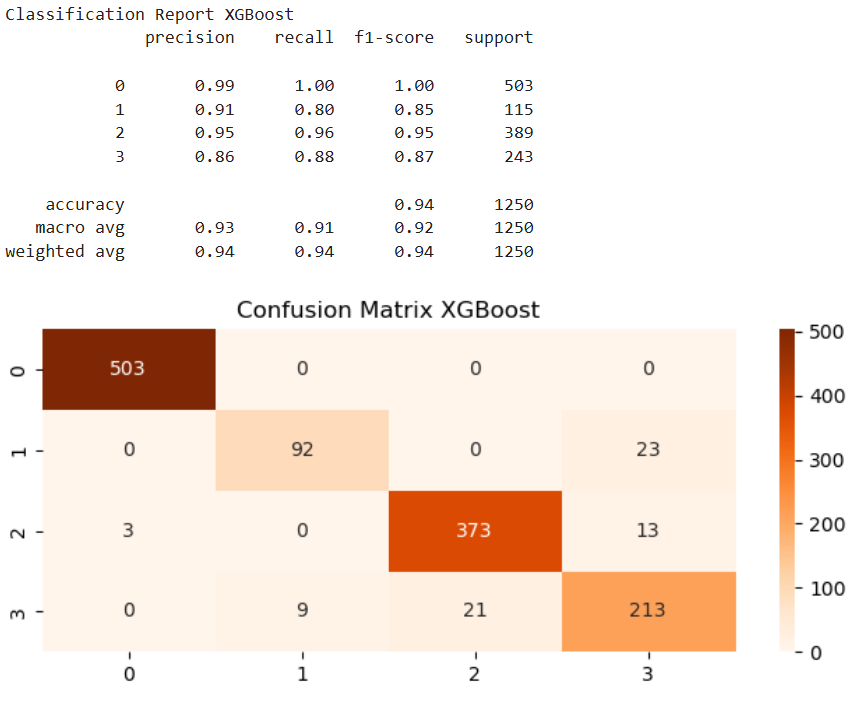
### Gradient Boosting Classifier:

### 

### K Neighbors Classifier:



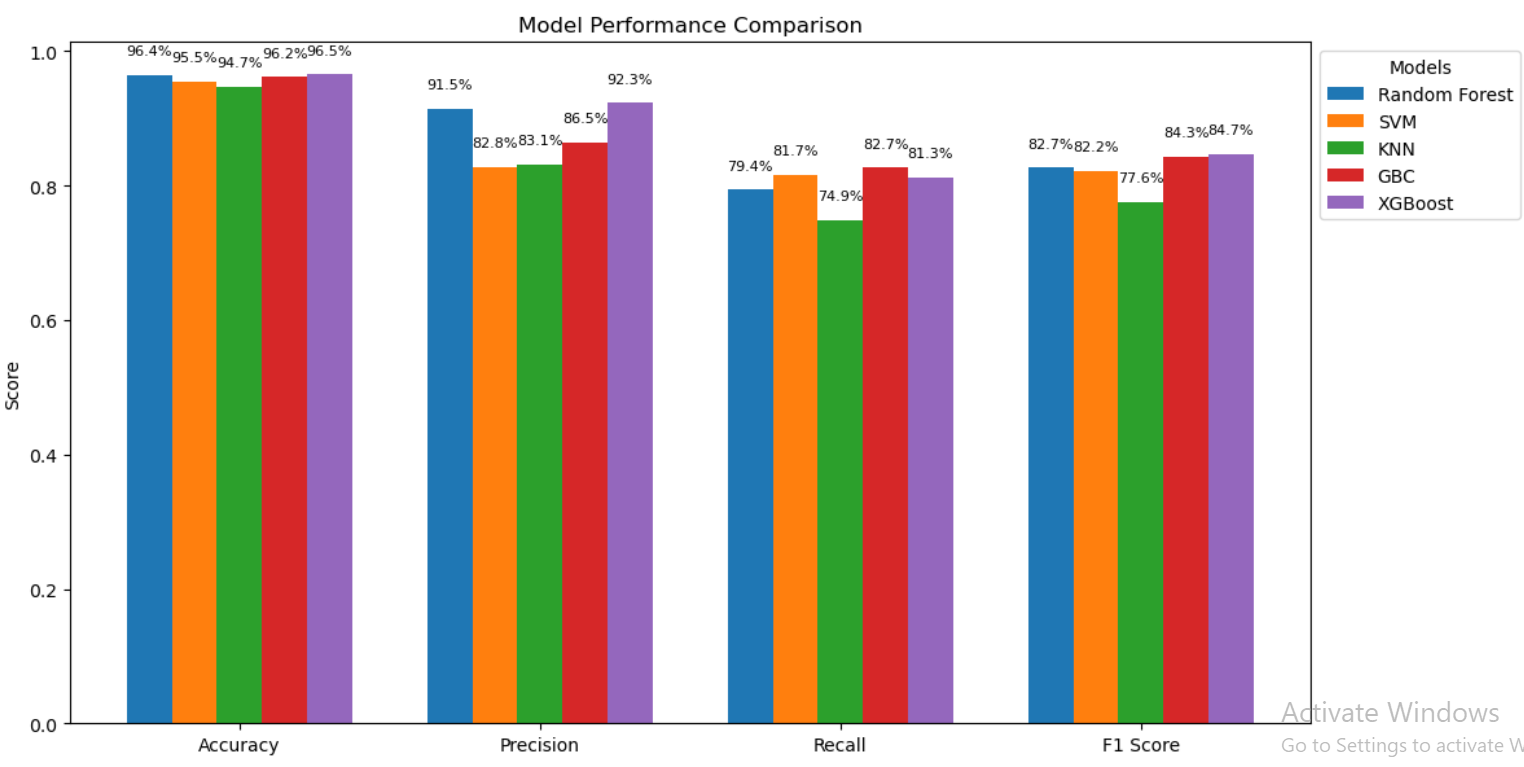
### XG Boost:



## **Results Discussion**

## 

### Models Comparison Chart



* In the given bar chart entitled “Model Performance Comparison”, we can observe the evaluation metrics (Accuracy, Precision, Recall, and F1 Score of five different machine learning models such as Random Forest, SVM, KNN, GBC, and XGBoost.
* The best obtained results, overall accuracy (96.5%), precision (92.3%), F1 score (84.7) indicate that XGBoost performed well.
* Random Forest also achieved 96.4% accuracy, 91.5% precision and 82.7% recall and F1 score, which should be regarded as a reliable model.
* Overall, GBC and SVM performed balanced, however slightly lower in recall and F1 score than the top models.
* The performance was generally the lowest for KNN, especially recall (74.9%) and F1 score (77.6%).
* Overall, XGBoost is better than other models with Random Forest coming the closest, and then KNN for the worst of all.
* Most of the classes, and especially the minority classes, performed best with Random Forest and XGB
* The Class 1 (Moderate) prediction failed because KNN was sensible to imbalance and scaling with instance.
* Balanced results and slightly lower recall was delivered by SVM.

## **Real-World Applicability:**

* GBC and RF models are suitable for production use in air quality monitoring systems Kim, Y., et al. (2020).
* As XGBoost is highly tunable and scalable for large datasets, there are even more tuning parameters available

# **8. Ethical and Practical Considerations**

## Data Bias and Fairness:

It was taken care that all the classes have fair representation each.

### Privacy and Data Governance:

None of the personally identifiable data was used. The dataset is GDPR and valid data handling.

## Environmental Impact:

Taking any action to improve air quality requires accurate prediction.

## Model Transparency:

Finally, interpretability was also explored with decision trees and feature importances for explanation as required by explainable AI.

# 9. Ethical Considerations

Machine learning models present critical ethical challenge while developing and deploying them in environmental decision making. Socially responsible AI such as addressing these concerns is possible.

## **Fairness and Bias**

Another is bias caused by data imbalance or proxy variables. For instance, there might be no doubt that an industrial location could be targeted at proximity to marginal areas. Specifically, bias was compensated with biased techniques such as stratified sampling and feature importance analysis.

## **Privacy and Data Governance**

Although the current dataset does not include personally identifiable information, real-world implementation might. It is imperative to adopt privacy-preserving techniques like differential privacy (Dwork & Roth, 2014) and secure federated learning frameworks.

## **Societal Impact**

Positive ML predictions can also influence public policy as well as infrastructure investments. This could result in the wrong resources being allocated, or in public panic. Therefore, the model should be always validated and completed with domain expert knowledge.

## **Transparency and Explainability**

In some cases (e.g. GBC models), when some models are black box, it will reduce interpretability. To increase stakeholder trust, future work will use SHAP or LIME for model agnostic explainability.

# 10. Conclusion

This project presents a complete, ethical, and robust pipeline for predicting air quality using machine learning models. A combination of rigorous data preprocessing, strategic model selection, advanced tuning, and thorough evaluation led to a high-performing Random Forest model capable of accurate classification.

The study highlights the applicability of ML in environmental domains and the importance of ethical modeling practices. With expanding data streams from IoT sensors, this work can be extended for real-time air quality monitoring.

## Future directions include:

* Incorporating deep learning techniques
* Real-time deployment using edge computing
* Enhancing transparency using interpretable AI frameworks

# 11. References

Zhang, L., et al. **(2018)**. ML applications in air pollution forecasting. Atmospheric Environment.

World Health Organization (WHO). (2021). Ambient (outdoor) air quality and health. Retrieved from <https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health>

Kuhn, M., & Johnson, K. (**2013**). Applied Predictive Modeling. Springer.

Dwork, C., & Roth, A. (**2014**). The Algorithmic Foundations of Differential Privacy.

Kim, Y., et al. (**2020**). Gradient boosting for environmental risk prediction.

**Dataset:**<https://www.kaggle.com/code/chineduukpai/air-quality-and-pollution/input?select=updated_pollution_dataset.csv>

Appendix: (Use of External Tools and Generative AI)

### Use of Google and Generative Chatbots (e.g., ChatGPT)

The assessment for the COM7003 – Artificial Intelligence module required external tools to support and knowledge understanding as well as development work. The work received quality improvements through properly utilized ethical tools including these following sources:

### Google Search

During the completion process I accessed information from Google through various searches.

### Academic references and journal articles.

The analysis includes official documentation with accompanying code examples such as Scikit-learn and XGBoost parameter tunning and understanding.

The module aims to provide me with clear explanations of Random Forest as well as the concepts behind Gradient Boosting while also understanding about evaluation metrics followed by lessons on hyperparameter tuning.

### Generative AI Tool – ChatGPT by OpenAI

This supportive tool served three main purposes throughout our work.

The process of identifying problem definitions and analyzing dataset contents required a brainstorming approach.

The tool provided technical concept definitions together with helping me in debugging.

ChatGPT and its OpenAI counterpart generated explanations regarding evaluation criteria selection as well as model suitability.

I executed every final decision independently while implementing code and training models for evaluation separately from AI tools that assisted in both understanding and explaining.