A Comparative Analysis of Machine Learning Models for Air Quality and Pollution Assessment

Amiya Ranjan Panda
School of computer engineering
KIIT Deemed to be University
Bhubaneswar, Odisha
amiya.pandafcs@kiit.ac.in

Aritra Bera
School of computer engineering
KIIT Deemed to be University
Bhubaneswar, Odisha
2205541@kiit.ac.in

Ankan Banerjee
Schoolof computer engineering
KIIT Deemed to be University
Bhubaneswar, Odisha
2205534@kiit.ac.in

Abhisek Mishra
Schoolof computer engineering
KIIT Deemed to be University
Bhubaneswar, Odisha
2205523@kiit.ac.in

Anurag Mukherjee
School of computer engineering
KIIT Deemed to be University
Bhubaneswar, Odisha
2205361@kiit.ac.in

Abstract— Air pollution remains a critical environmental and public health challenge, necessitating accurate classification of air quality levels. This study presents a comparative analysis of multiple machine learning models distinguishing air quality into several categories: Good, Moderate, Poor, and Hazardous. The dataset downloaded from Kaggle comprises 5000 samples with key environmental and demographic attributes such as temperature, humidity, pollutant concentrations and population density. A variety of machine learning classifiers, including Logistic Regression, Decision Tree, Random Forest, SVM, Gradient Boosting (GBM, XGBoost, CatBoost), and ensemble methods, were evaluated based on accuracy, precision, recall, and F1-score. The results indicate that ensemble methods like Random Forest, Extra Trees Classifier, and XGBoost achieved the highest accuracy (96.10% - 96.20%) and overall classification performance. Conversely, Bernoulli Naive Bayes exhibited the lowest accuracy (40.90%), highlighting its limitations in handling this dataset. Our study highlights the efficiency of machine learning in air quality assessment and its potential application in environmental monitoring and policy-making.

Keywords— Air Pollution, Data Science, Ensemble Learning, Multiclass Classification, XGBoost.

I. Introduction

Air pollution is a significant environmental and public health challenge, contributing to respiratory diseases, cardiovascular complications, and reduced life expectancy. With rapid urbanization and industrialization, monitoring air quality has become essential for implementing effective pollution control strategies and issuing timely health advisories. Traditional air quality monitoring relies on physical sensors and manual observations, which, while reliable, are often expensive, resource-intensive, and geographically limited. Machine learning [1] provides a scalable and efficient alternative by leveraging historical data to predict pollution levels accurately.

This study focuses on classifying air quality using machine learning techniques based on a dataset containing 5000 samples. The dataset captures key environmental and demographic factors that influence pollution levels. These include temperature, which affects the chemical behavior of pollutants in the atmosphere, and humidity, which impacts pollutant dispersion and concentration. The dataset also includes critical air pollutants which are coarse particulate matter responsible for respiratory illnesses, as well as gaseous pollutants like nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and carbon monoxide (CO), which originate primarily from combustion processes and industrial emissions. Additionally, proximity to industrial areas is considered a crucial factor, as regions closer to major industrial zones tend to experience higher pollution levels. Population density is also included, as human activities such as transportation and energy consumption contribute to air quality variations.

The classification problem [2] involves four air quality levels: Good, Moderate, Poor, and Hazardous, determined based on pollutant concentrations and environmental conditions. To address this, we have used traditional classifiers like Logistic Regression, Decision Tree, and Support Vector Machine (SVM), Random Forest, Gradient Boosting, XGBoost, and CatBoost, and probabilistic methods like Naïve Bayes, among others. Models are evaluated using accuracy, precision, recall, and F1-score to determine their effectiveness in predicting air quality levels. The paper is divided into sub-sections such as Literature Survey, Methodology, Assessment of Results, and Conclusion.

II. Literature Survey

Air pollution remains a vital challenge with severe consequences for people's health and the environment. Air quality monitoring systems rely on sensor-based networks and manual reporting, which, although effective, are costly and limited in coverage. In recent years, machine learning has been widely applied in air pollution prediction and classification, offering more efficient and scalable solutions. Various studies have explored different algorithms and data-driven methodologies to improve air quality assessment and forecasting.

One study examines the application of machine learning in predicting air pollution levels based on meteorological parameters and pollutant concentrations [3]. The researchers employ decision tree-based ensemble methods such as Random Forest and Gradient Boosting to enhance predictive accuracy. Their findings highlight the effectiveness of

machine learning models in capturing complex relationships between atmospheric conditions and pollution levels.

Another study integrates deep learning techniques for real time monitoring using IoT-based sensor networks [4]. The research demonstrates how deep learning can effectively process temporal and spatial air quality data, improving prediction accuracy compared to traditional statistical models.

In a separate study, the authors propose a hybrid approach integrating important feature choosing and classifiers to improve air quality classification [5]. The study explores the impact of different parameters, such as temperature, humidity and pollutant concentrations, on classification performance. Support Vector Machines (SVM) and Naïve Bayes classifiers are used to differentiate air quality levels, with SVM achieving higher accuracy due to its capability to handle complex data.

Additionally, an IoT-driven framework for air quality prediction is developed, where data from multiple sensors is processed using machine learning models such as Decision Trees and K-Nearest Neighbors (KNN) [6]. The authors emphasize the significance of selecting appropriate in improving performance of the model and reducing computation time.

Another research study discusses the use of ensemble learning techniques, including XGBoost and CatBoost, in air pollution forecasting [7]. The study shows that boosting-based models significantly outperform traditional classifiers due to their power to deal with unbalanced data and capture nonlinear relationships between environmental conditions.

To conclude, the usage of machine learning in air quality classification and prediction has gained significant traction. Studies have explored various approaches, including decision trees, ensemble learning and SVM, demonstrating their effectiveness in improving classification accuracy and prediction reliability. These advancements contribute to the development of robust, scalable air quality monitoring systems, ultimately aiding in pollution control efforts and public health protection.

III. Basic Concepts

A. Logistic Regression

Logistic Regression a model using concepts of statistics widely used for problems having two classes. It determines the likelihood of an event occurring by passing data to a function, also known as the sigmoid curve. This algorithm is useful for modeling relationships between a dependent variable and one or more independent variables, ensuring interpretability and efficiency.

B. K-Nearest Neighbors (KNN)

KNN does not use any parameters: it is a lazy learner-based learning algorithm that distinguishes data points based on the closest *k*-nearest neighbors. It uses distance calculation to measure similarity and is effective in scenarios with well-separated classes, though it may struggle with high-dimensional data.

C. Decision Tree Classifier

A Decision Tree is one of the best learning algorithms that repeatedly divides the data into sub-parts based on columns values. It follows a tree-like structure, with nodes representing decisions and leaves indicating class labels. The model selects the best split using metrics like Gini impurity or entropy, making it interpretable but prone to generating a complex algorithm.

D. Random Forest Classifier

Random Forest is a type of combinational learning method that builds a few decision trees and combines their outputs to improve performance and robustness. By averaging predictions or taking majority votes, it reduces overfitting compared to a single tree and increases generalization on unseen data.

E. Naive Bayes Classifier

Naive Bayes is a classifier based on Bayes' theorem, assuming non-dependence among variables. Despite this assumption, its performance is good in high-dimensional data and text distinguishing tasks. It is efficient in calculation and requires less training data.

F. *Support Vector Machine (SVM)*

SVM is a classification algorithm that finds the hyperplane to separate different categories. It maximizes the distance between support vectors, leading to better generalization. It also employs kernel functions to handle non-linearly separable data effectively.

G. AdaBoost Classifier

Adaptive Boosting (AdaBoost) is a combining method that combines non-strong learners, typically decision trees, to create a better classifier. It gives larger weights to wrongly classified samples, forcing next learners to focus on hard cases, thereby enhancing overall classification performance.

H. Gradient Boosting Machine (GBM)

GBM is a boosting algorithm that creates models sequentially, rectifying errors from previous work. Unlike AdaBoost, GBM optimizes the total loss occurred through calculus techniques, making it effective for both regression and classification tasks but computationally expensive.

I. Extra Trees Classifier

Extra Trees (Extremely Randomized Trees) is a variation of Random Forest that introduces additional randomness in feature selection and splitting criteria. This leads to more diverse trees and reduces variance, resulting in an efficient and robust model for classification tasks.

J. CatBoost Classifier

CatBoost is a boosting algorithm designed for non-numeric data. It efficiently handles categorical features without extensive preprocessing, improving performance and reducing model complexity. It is often used in organized data classification problems.

K. XGBoost Classifier

XGBoost (Extreme Gradient Boosting) [8] is a highly optimized implementation of gradient boosting that uses parallel computing, regularization techniques, and tree pruning to enhance efficiency and accuracy. It is known for its superior performance in structured data classification challenges.

L. Bagging Classifier

Bagging is a combinational [9] learning method that improves model stability by training multiple instances of a weak learner on different bootstrap samples. It reduces variance and enhances generalization, often using decision trees as base learners.

M. SGD Classifier

SGD is a looping optimizing algorithm [10] used for large-scale classification problems. It updates model parameters incrementally with each training instance, making it efficient for high-dimensional data but sensitive to hyperparameter tuning and learning rate adjustments.

N. Ridge Classifier

Ridge Classifier is a linear model that uses L2 ridge regularization [11] to prevent complexities. It works well in high-dimensional spaces by penalizing large coefficients, ensuring better generalization compared to simple logistic regression.

O. Bernoulli Naive Bayes

Bernoulli Naive Bayes is a specifically designed for two-variable classes or data having a Bernoulli distribution [12]. It is commonly used for distinction, where a column's presence or absence plays a critical role in decision-making.

P. Passive Aggressive Classifier

The Passive Aggressive Classifier is an algorithm that updates variables only when a misclassification occurs. It is efficient in handling large-scale, streaming data but requires careful tuning of its aggressiveness parameter to prevent overfitting or underfitting.

IV. METHODOLOGY

A. *Exploratory Data Analysis:*

Using the Pandas_*read_csv* method, we first read the dataset. We begin by loading the dataset and then carry out the following fundamental data exploration:

1. Analyzing the dataset's size:

   (5000 samples × 10 columns)

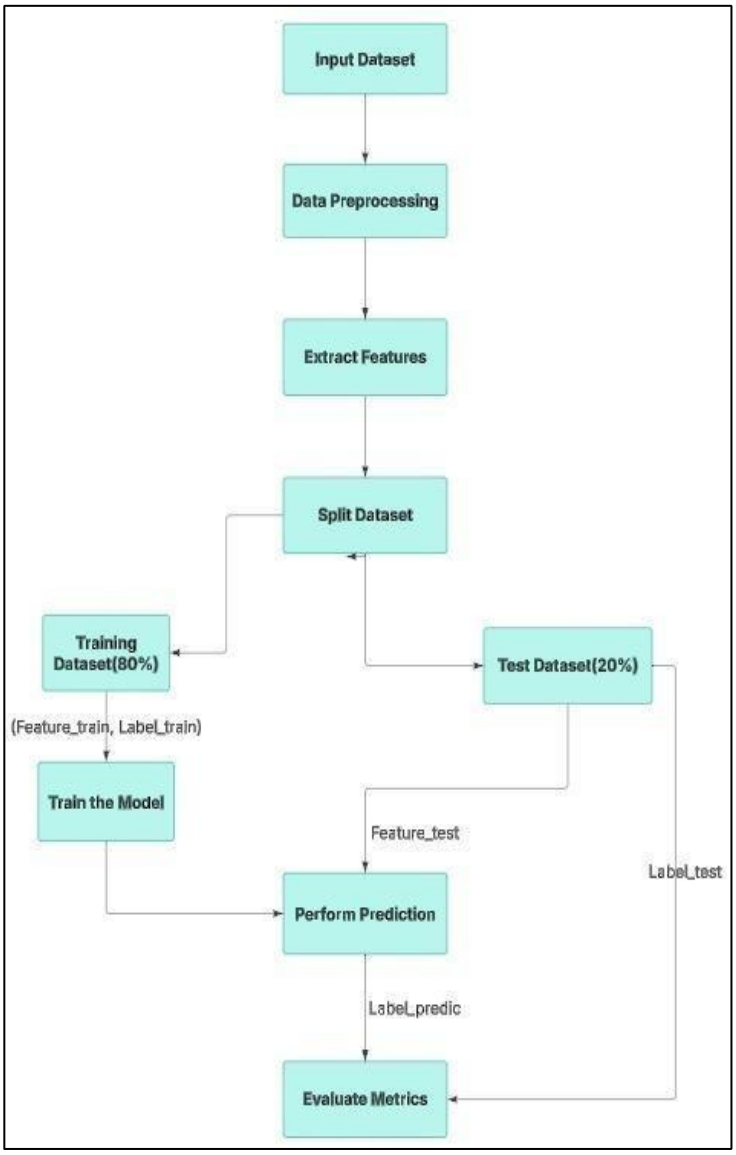

Fig.1 Flowchart for the whole process of the model implementation

2. Temperature (°C): The average air temperature recorded in the region, measured in degrees Celsius. Temperature influences chemical reactions in the atmosphere and can affect the formation of air pollutants such as ozone.
3. Humidity (%): The relative humidity in the region, expressed in terms of within hundred. It shows how much water vapour is present in the air compared to the absolute capacity it can contain at a given condition. High moisture can impact pollutant dispersion and influence the formation of secondary pollutants.
4. PM2.5 Concentration (µg/m³): The concentration of fine particles in the air, with diameters less than 2.5 micrometers, measured in mg/m³. These particles are small enough to penetrate deep into the lungs, posing serious health risks.
5. PM10 Concentration (µg/m³): The concentration of particulate matter with diameters less than 10 micrometers, measured in micrograms per cubic meter. Although larger than PM2.5, these particles can still cause respiratory irritation and health issues.
6. NO_2 Concentration (ppb): The concentration of nitrogen dioxide in parts per billion. NO_2 is a harmful

gas produced primarily by vehicle emissions and industrial processes, contributing to respiratory issues and air pollution.

7. SO_2 Concentration (ppb): The quantity of sulfur dioxide in the air, measured in ppb. SO_2 is emitted from burning fossil fuels and industrial activities, contributing to acid rain and respiratory problems.
8. CO Concentration (ppm): The concentration of carbon monoxide in parts per million. CO is a toxic gas produced by incomplete combustion of fossil fuels, which can impair oxygen delivery to the body when inhaled in high amounts.
9. Proximity to Industrial Areas (km): The distance between the measurement location and the nearest industrial area, measured in kilometers. Closer proximity generally indicates a higher likelihood of exposure to industrial pollutants.
10. Population Density : The number of people living per square kilometer of the place. Higher population density often correlates with increased pollution from human activities such as vehicle emissions, industrial operations, and domestic fuel combustion.

• Analyzing the first few rows of the dataset:

A. Data Preprocessing:

Converting the UTC to datetime format makes it easily interpretable to viewers.

B. Feature Engineering:

Generating temporal features using functions from the *feature_engine* package.

C. Data Optimization:

Iterate through all the columns of a dataframe and modify the datatype to reduce memory usage

```
Initial Memory (MB): 0.3815956115722656
Optimized Memory (MB): 0.19086074829101562
Training Data Memory (MB): 0.213623046875
Testing Data Memory (MB): 0.05340576171875
```

.Fig.2 Memory usage before and after optimization of training and testing data

D. EDA :

From the correlation matrix [13], we analyze which features have the strongest correlation with Air Quality. The most significant feature is CO (Carbon Monoxide), with a correlation of -0.91, indicating that higher CO levels strongly degrade air quality. The second strongest correlation is with NO2 (Nitrogen Dioxide) at -0.79, showing that higher NO_2 concentrations also lead to poor air quality. Additionally, Industrial_Areas shows a correlation of 0.77, suggesting that proximity to industrial zones contributes significantly to air pollution. Similarly, PM2.5 (Particulate Matter) exhibits a correlation of -0.74, reinforcing its impact on air quality deterioration. The negative correlations indicate that as pollutant levels rise, air quality declines significantly. These findings highlight the importance of monitoring emissions, especially in urban and industrial regions, to mitigate air pollution.Therefore, stringent regulations and effective pollution control measures are essential to improving air quality and safeguarding public health.

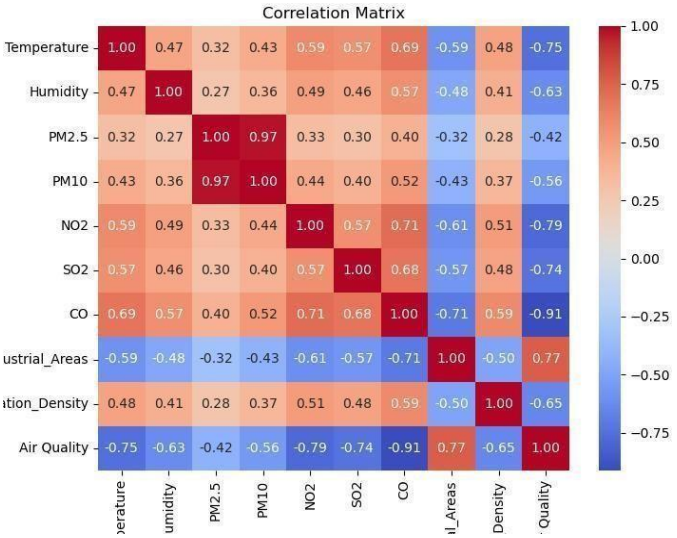

Fig. 3 Correlation matrix visualization of all the feature types in the dataset

E. Feature Importance:

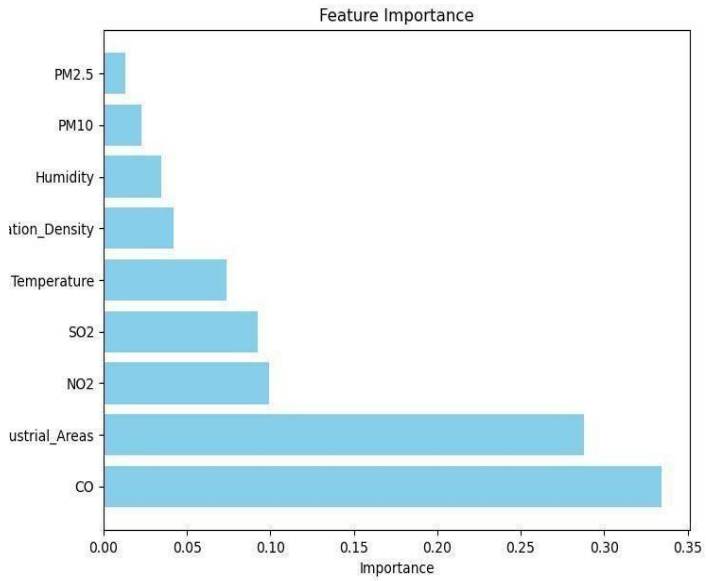

Fig.4 Bar Graph showing the feature importance of dataset attributes

F. t-SNE :

The visualization [14] reveals that the categories of the air quality are mostly separable, indicating that the features in the dataset hold valuable information that can distinguish between these classes. This separability suggests that machine learning models can potentially learn to predict the air quality status with reasonable accuracy.

An interesting observation from the visualization is the formation of meandering lines by the data points. This pattern occurs because the dataset has a time-series structure, meaning that each data point is influenced by the temporal order of events. In time-series data, consecutive samples often exhibit similarities, leading to clusters and continuous trajectories in the t-SNE plot. The proximity of similar data points in the low-dimensional space indicates that t-SNE effectively captures and visualizes these temporal relationships.

Table 1 Comparative Analysis Of Models

Model Name	Accuracy	Class Labels											
		Good			Moderate			Poor			Hazardous		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Logistic Regression	94.80%	1.00	1.00	1.00	0.96	0.96	0.96	0.84	0.89	0.86	0.92	0.87	0.86
KNN	75.10%	0.87	0.97	0.92	0.70	0.71	0.71	0.53	0.59	0.56	0.88	0.32	0.47
Decision Tree	92.10%	0.99	1.00	1.00	0.96	0.91	0.94	0.81	0.80	0.80	0.76	0.86	0.81
Random Forest	96.10%	1.00	1.00	1.00	0.97	0.97	0.97	0.88	0.91	0.90	0.92	0.88	0.90
Naive Bayes Classifier	92.60%	1.00	0.99	0.99	0.94	0.94	0.94	0.79	0.87	0.83	0.88	0.77	0.82
SVM	78.00%	0.88	0.97	0.92	0.69	0.74	0.72	0.68	0.58	0.63	0.81	0.45	0.58
AdaBoost	85.90%	1.00	1.00	1.00	0.92	0.81	0.86	0.62	0.76	0.60	0.70	0.75	0.72
GBM	95.50%	0.99	1.00	1.00	0.97	0.96	0.96	0.88	0.89	0.90	0.91	0.86	0.89
Extra Trees Classifier	96.20%	1.00	1.00	1.00	0.98	0.97	0.97	0.88	0.91	0.93	0.84	0.91	0.86
CatBoost	95.60%	1.00	1.00	1.00	0.96	0.97	0.94	0.86	0.91	0.88	0.92	0.86	0.89
XGBoost	96.10%	1.00	1.00	1.00	0.97	0.95	0.96	0.87	0.93	0.90	0.95	0.92	0.90
Bagging Classifier	94.50%	1.00	1.00	1.00	0.97	0.95	0.96	0.85	0.87	0.86	0.87	0.87	0.87
SGD Classifier	63.30%	0.82	1.00	0.90	0.44	0.72	0.54	1.00	0.00	0.00	0.68	0.12	0.20
Ridge Classifier	73.70%	0.85	1.00	0.92	0.75	0.72	0.73	0.49	0.46	0.48	0.66	0.19	0.29
Bernoulli Naive Bayes	40.90%	0.41	1.00	0.58	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Passive Aggressive Classifier	51.70%	0.98	0.81	0.89	0.00	0.00	0.00	0.28	0.99	0.44	0.00	0.00	0.00

Moreover, the t-SNE visualization provides insights into the data's inherent structure, helping identify potential challenges such as overlapping regions between classes, noise, or outliers. If distinct clusters represent the air quality classes, it confirms the potential feasibility of training classification models. Conversely, significant overlap or poor separability may suggest a need for additional feature engineering, data cleaning, or alternative approaches for modeling.

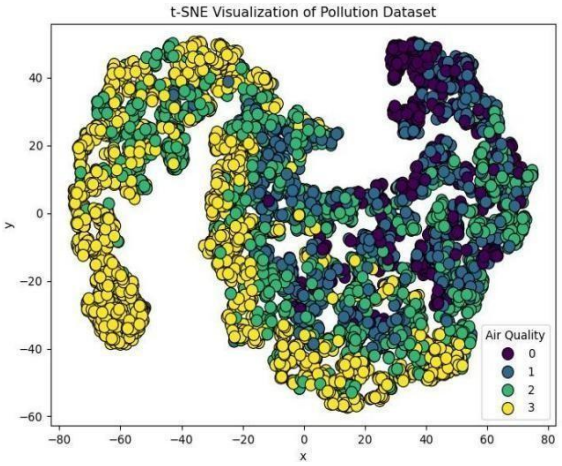

Fig. 5 t-SNE graphical representation

G. Defining Functions:

The functions listed below are what our models will eventually use to examine the outcomes.

1. Confusion Matrix:

The confusion matrix [15] is used for understanding how well our classification model performs. It is a grid with rows and columns representing actual categories in our data. Each box showcases how many data points fell into different prediction buckets. High numbers along the diagonal, which indicate accurate classifications (True Positives and True Negatives), are what we ideally want to see. Off-diagonal values, however, reveal errors. False Positives are mistakenly assigned a positive class, while False Negatives represent missed positive cases [17].

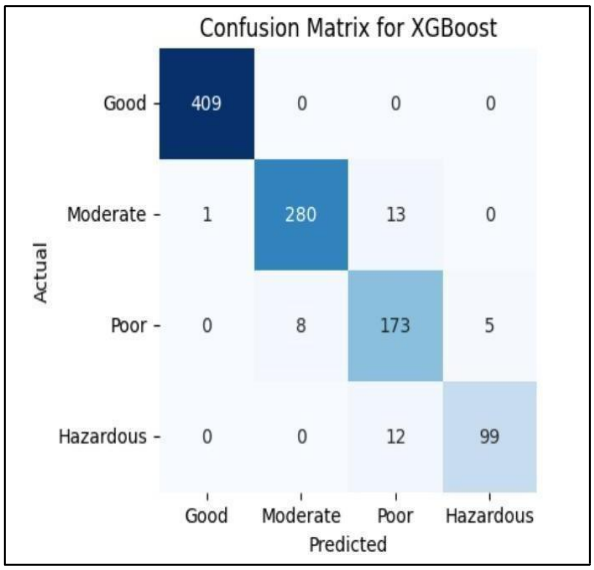

Fig. 6 Confusion matrix of the XGBoost model having accuracy 96.10%

H. Models :

Now that the aforementioned procedures have been completed, we are ready to train and evaluate the machine learning models: Logistic Regression, KNN, Decision Tree, Random Forest, Naive Bayes, Support Vector Machine (SVM), AdaBoost, Gradient Boosting (GBM), Extra Trees Classifier, CatBoost, XGBoost, Bagging Classifier, SGD Classifier, Ridge Classifier, Bernoulli Naive Bayes, and Passive Aggressive Classifier.

For each model accuracy, precision, recall, and F1-score are calculated.

V. ASSESSMENT OF RESULTS

Table 1 presents the evaluation outcomes of different models on the air pollution data. The results indicate that ensemble methods like Random Forest, XGBoost, and Extra Trees Classifier achieved the highest accuracy (96.10% - 96.20%) and overall classification performance. Conversely, Bernoulli Naive Bayes exhibited the lowest accuracy (40.90%), highlighting its limitations in handling this dataset.

However, it is important to note that accuracy alone might not be a sufficient metric to capture the performance of a model, especially when datasets are unbalanced or when the costs associated with classification differ from class to class.

VI. CONCLUSION AND FUTURE SCOPE

In this study, we conducted an analysis of models for air quality and pollution assessment. We find that the Extra Trees Classifier emerged as the highly effective model, achieving an correctness of 96.20%. This performance indicate the model's robustness in handling complex datasets and its potential for real-world applications in air quality monitoring.

Further studies could explore the combination of additional environmental factors, like meteorological data and geographical information, to enhance predictive accuracy further. Additionally, implementing ensemble methods or hybrid models may yield even better performance. Expanding the scope to include real-time data analysis and developing user-friendly applications for stakeholders could also amplify the impact of this research on public health and environmental policy.

Moreover, integrating deep learning techniques, such as recurrent neural networks or transformers, could further enhance the model's ability to capture temporal dependencies in air quality data. Incorporating domain-specific feature engineering and explainable AI methods may also improve interpretability and trust in the model's predictions. Future work could focus on cross-regional generalizability by testing the model on diverse datasets from different geographical locations. Finally, collaboration with policymakers and environmental agencies could facilitate the deployment of these models in decision-making frameworks for proactive pollution control strategies.

VII. REFERENCES

[1] Pattnayak P, Panda AR. Innovation on machine learning in healthcare services—An introduction. Technical Advancements of Machine Learning in Healthcare. 2021:1-30.

[2] Mishra S, Tripathy HK, Panda AR. An improved and adaptive attribute selection technique to optimize dengue fever prediction. Int J Eng Technol. 2018;7:480-6.

[3] Larkin A, Hystad P. Towards personal exposures: how technology is changing air pollution and health research. Current environmental health reports. 2017 Dec;4:463-71.

[4] Devi M, Dhaya R, Kanthavel R, Algarni F, Dixikha P. Data Science for Internet of Things (IoT). InSecond international conference on computer networks and communication technologies: ICCNCT 2019 2020 (pp. 60-70). Springer International Publishing.

[5] Zhao X, Zhang R, Wu JL, Chang PC. A Deep Recurrent Neural Network for Air Quality Classification. J. Inf. Hiding Multim. Signal Process.. 2018 Mar;9(2):346-54.

[6] Guo, G., Wang, H., Bell, D., Bi, Y. and Greer, K., 2003. KNN model-based approach in classification. In On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings (pp. 986-996). Springer Berlin Heidelberg.

[7] Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. Journal of big data. 2020 Nov 4;7(1):94.

[8] Chen T, Guestrin C. Xgboost: A scalable tree boosting system.InProceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 785-794).

[9] Breiman L. Bagging predictors. Machine learning. 1996 Aug;24:123-40.

[10] Woodworth B, Patel KK, Stich S, Dai Z, Bullins B, Mcmahan B, Shamir O, Srebro N. Is local SGD better than minibatch SGD?. InInternational Conference on Machine Learning 2020 Nov 21 (pp. 10334-10343). PMLR.

[11] Hastie T. Ridge regularization: An essential concept in data science. Technometrics. 2020 Oct 1;62(4):426-33.

[12] Dai B, Ding S, Wahba G. Multivariate bernoulli distribution.

[13] Steiger JH. Tests for comparing elements of a correlation matrix. Psychological bulletin. 1980 Mar;87(2):245.

[14] Van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008 Nov 1;9(11).

[15] Townsend JT. Theoretical analysis of an alphabetic confusion matrix. Perception & Psychophysics. 1971 Jan;9:40-50.