Department of Electrical and Computer Engineering
North South University

CSE445 Report

Group-7

Air Quality Index Prediction

Md. Mushfiqur Rahman Mahin	**ID**	**2014299042**
Ratul Bhattarcharjee	**ID**	**2012996042**
Mohammad Olid Afzal	**ID**	**2011831042**
Ramisa Asad	**ID**	**1931532042**

Faculty:

Riasat Khan

Assistant Professor

ECE Department

Spring, 2024

Individual Contribution Table

Section	Contributing Member Name	
IEEE/LaTEX formatting	Mahin	
Turnitin check	Olid	
Grammarly check	Mahin	Grammarly Score:
Abstract	Ratul	100
Keywords	Ratul	
Introduction Motivation	Ramisa	99
Paper Review 1	Mahin	100
Paper Review 2	Ratul	100
Paper Review 3	Olid	100
Paper Review 4	Ramisa	97
Introduction Second-Last Paragraph	Mahin, Ramisa	100
Proposed System (Dataset and Preprocessing)	Ratul	91
Proposed System (Model description)	Random forest - Olid	98
	KNN - Ramisa	98
	XGBoost - Ratul	98
	Adaboost - Mahin	98
Results and Discussion	Mahin	97
Figure and Table Title Formatting	Mahin, Olid	
Conclusions	Ratul	95
Equations formatting	Ratul, Ramisa	
References Formatting in IEEE format	Olid	

Air Quality Index Prediction

Md. Mushfiqur Rahman Mahin
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
mushfiqur.mahin1@northsouth.edu

Mohammad Olid Afzal
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
olid.mohammad@northsouth.edu

Ratul Bhattarcharjee
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
ratul.bhattarcharjee@northsouth.edu

Ramisa Asad
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
ramisa.asad@northsouth.edu

Abstract— **Dhaka, Bangladesh's capital, has always been among the top 3 cities with the worst air quality worldwide. Dhaka's air quality is getting worse every day since it's one of the world's most densely inhabited cities and because of its overcrowding, transportation, and industrialization. Many deadly diseases, including respiratory, cardiovascular, cancer, neurological disorders, etc., are caused by air pollution. This project intends to predict the Air Quality Category to reduce these effects using machine learning. This paper aims to know the AQI Category based on some given pollutant (PM2.5) concentrations monitored in Dhaka for the past two years (2022,2023). The study utilized the Scikit-learn library to implement four machine-learning algorithms. We also used Explainable AI (LIME) to visualize the prediction results. We have also used smote to balance the class distribution. The evaluation metrics for these algorithms for the classification task were precision, recall, accuracy, and f1-score. The result shows that XGBoost and K-Nearest Neighbor achieved the highest cross-validation scores of 84.9% and 82.6%, respectively.**

Keywords—Air Quality Index, Machine Learning, Scikit-learn, PM2.5 Concentration, Smote, Evaluation Metrics.

I. INTRODUCTION

The Air Quality Index reports air quality. It detects if the air is clean or polluted and what type of health issue might arise. The Air Quality Index focuses on the health effects a person may encounter within a few hours or days after inhaling contaminated air.[1] On a global scale, air pollution is gradually becoming a severe matter. Air pollution happens when the climate gets infected or polluted by any chemical or physical substance or natural elements.[2]. Deaths from air pollution can be fatal since medical experts cannot connect both. Globally, air pollution is positioned fourth due to early causes of death. In 2021, IQAir reported that on an annual average of PM2.5, Bangladesh ranked first and became the most polluted country among other nations. Traditional techniques for predicting the Air Quality Index (AQI) employ mathematical models that assess many aspects of meteorological states, emissions from different sources, chemical reactions in the air, and geographical attributes.[3] Traditional methods become complicated for specific necessities and public resources. Also, statistical methods and practical connections from monitoring data should be relied on. [4] It integrates the machine learning model in the air quality index prediction to decrease complexity and make it an easy process for routine monitoring, air pollution predictions, and ambient air quality maintenance. Air quality monitoring is necessary for maintaining air quality, protecting health issues, and ensuring compliance with regulations.

Kothandaraman et al. [5] applied machine learning methods to predict the levels of pollutants and overall air quality in a specific area. The authors operated Meteorological and PM2.5 Datasets containing 49056 samples. However, there was some improper data with null values. These null values were restored using the mean value. The authors tested different machine learning models, where Adab models achieved the highest accuracy of 42.90%.

Natarajan and his team [6] attempted to predict the AQI (Air Quality Index) in various Indian cities using an unconventional machine learning model that combines Grey Wolf Optimization and a Decision Tree. They have considered the "Air Quality Data in India (2015-2020)" dataset from Kaggle. The verification measures used by the authors included RMSE, R-square, MAE, MSE, and accuracy. Compared to other classic machine learning techniques, a hybrid approach achieves an accuracy of up to 97.68%.

Gupta and his team [7] used machine learning to forecast the Air Quality Index in Indian cities like Delhi, Kolkata, Bangalore, and Hyderabad. The original dataset for this model had 29532 rows and 16 columns. The model was created using SMOTE, SVR, RFR, and CR algorithms. Hyderabad has the best accuracy at 90.97.

Ravindiran and his team [8] employed machine learning for routine monitoring, air pollution predictions, and ambient air quality maintenance. The authors used the Indian Central Control Room (CCR) for the Air dataset for their research. The authors employed several machine-learning models from which the Catboost model achieved a high prediction accuracy of 0.9998 and a low RMSE of 0.76.

Our project used four machine learning models to predict the air quality index (AQI). We collected the dataset from the AirNow website, and then we developed models to forecast the level of air pollution. We used four machine learning methods to find the best model for predicting air quality. Our research addresses the pressing need for state-of-the-art predictive models in environmental science, focusing on improving public health outcomes and giving decision-makers the tools they need to reduce air pollution. We are investigating more advanced techniques for controlling and observing air quality.

We have organized our findings to give a comprehensive view of what we’ve found through our research. Section II presents the elaborated system proposal with the help of tables, diagrams, or flowcharts to enable better

understanding. The most exciting findings of our study are unraveled in Section III, where we discuss some interesting insights obtained from performance measures analysis. We finish our work by summarizing Section IV and suggesting possible ways to advance it. Our primary target is to achieve the optimal performance.

II. Proposed System

In this section, we will describe the theory of all the software components- dataset, preprocessing and machine learning models.

A. *Dataset*

The dataset utilized in our study was sourced from the airnow.gov [5] website, comprising 17,130 samples collected over two years, each year contributing 8,413 samples. The dataset encompasses various features relevant to air quality monitoring, including Site, Parameter, Year, Month, Day, Hour, NowCast Conc., AQI, Raw Conc., Conc. Unit, Duration, QC Name, and Date. Among these features, the primary focus was predicting the AQI Category, which serves as the target variable for our analysis. Each sample in the dataset provides valuable information about air quality measurements, enabling us to explore relationships between different parameters and their impact on AQI categorization. Class distribution by various AQI categories is illustrated in Fig. 1 & . Through rigorous dataset analysis, we aim to develop robust predictive models for assessing and forecasting air quality conditions.

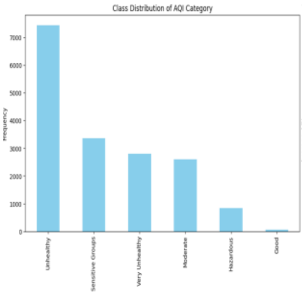
TABLE I. MAX, MIN, MEAN VALUE OF FEATURES OF THE EMPLOYED DATASET

index	Year	Month	Day	Hour	NowCast Conc.	AQI	Raw Conc.
count	17131	17131	17131	17131	17131	17131	17131
mean	2022.50	6.52	15.62	11.50	98.63	167.17	98.94
std	0.50	3.47	8.76	6.92	83.95	78.34	90.24
min	2022	1	1	0	-999	-999	-999
25%	2022	3	8	5	44.8	124	43
50%	2023	7	16	11	74.5	161	73
75%	2023	10	23	18	136.5	193	137
max	2024	12	31	23	648.5	598	985

TABLE I illustrates the max, mean value of features of the employed dataset

Fig. 1. Class Distribution by AQI category

Fig. 2.. Average AQI by month of 2022



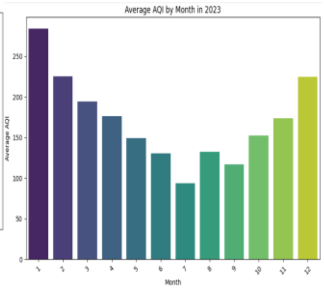


Fig. 1. shows the Class Distribution month of category

Fig. 2.. shows the Average AQI by 2022

Fig. 3. Box-Plot of Nowcast Concentration.

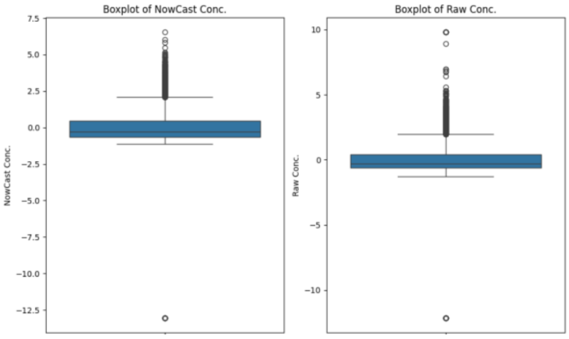


Fig. 3. shows the box plot of nowcast concentration and raw concentration

B. *Dataset Preprocessing*

To improve the computational efficiency we dropped the Date (Local Time) and Date features from the dataset as we can effectively derive this information from Year, Month, Day, and Hour features. We used the Label Encoding method for our target variable 'AQI Category'. We assigned 6 unique values to the classes Good: 0, Hazardous: 1, Moderate: 2, Unhealthy: 3, Unhealthy for Sensitive Groups: 4, Very Unhealthy: 5, nan: 6. We filled the most frequent AQI Category to impute the missing values of our target variable.

We used the standard scaler method for a balanced scale.

$$X = \frac{X - X_{mean}}{standard\ deviation} \tag{1}$$

X = previous value, X_{mean}= mean feature value.

We used the Boxplot method using the quartile Range for outlier detection. After removing outliers, we used the min-max scaler to scale our data between 0 to 1.

$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{2}$$

X_{new}= new value of features, X_{min}= minimum value of feature

X_{max} =maximum value of feature

We used Pearson's Correlation Coefficient method to find the linear relation of our features only on the training dataset of X. Then, we dropped both 'AQI' and 'Raw Conc.' features from the training and testing set as their threshold value is above 85%. X.

$$\mathrm{r} = \frac{n\,(\Sigma xy - (\Sigma x)(\Sigma y))}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}} \tag{3}$$

r= Pearson Correlation Coefficient, n= number of data point

Σx= sum of X scores, Σy= sum of Y scores

Σx^2= sum of squared X scores, Σy^2= sum of squared Y scores

Fig. 4. Pearson's Correlation Coefficient.

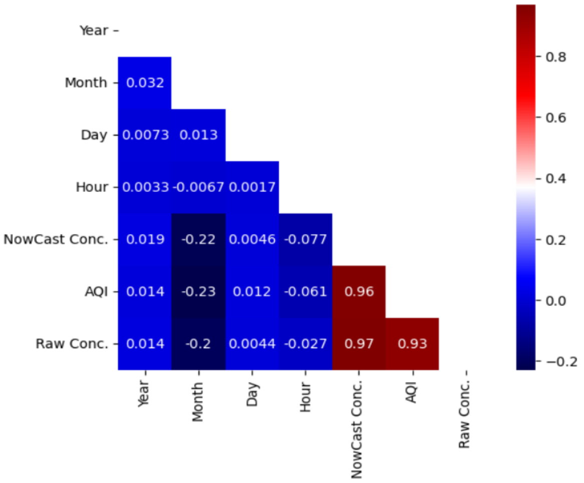


Fig. 4. illustrates Pearson's correlation coefficient for the various features of the dataset

C. *Machine Learning Models;*

We used four machine learning models K-Nearest Neighbors (KNN), Random Forest, XGBoost, and AdaBoost to predict the air index quality.

1) *RANDOM FOREST:* The Random Forest ensemble approach uses random data selections to create several independent decision trees. Combining predictions from each tree produces the final prediction, which lowers variance and enhances model generalization.

2) *K-Nearest Neighbours (KNN):* Using the training data as a starting point, this non-parametric classification technique groups data points according to how similar they are to their k nearest neighbors.

3) *XGBOOST:* Extreme Gradient Boosting, or XGBoost, is an ensemble learning method that builds a robust model by combining several weak decision trees. XGBoost is renowned for managing complicated datasets well and can stop overfitting.

4) *ADABOOST:* Adaptive Boosting, or AdaBoost, is a method that repeatedly refines a weak learner by emphasizing instances that the prior learner misclassifies. AdaBoost produces a more robust final model by adaptively increasing the weights of difficult occurrences.

We begin the process by collecting raw datasets. Then, we do some preprocessing to ensure data quality. We divide the data into 80% training and 20% testing data. Given a training set, machine learning algorithms are used to build predictive models while evaluating them using the testing set. The model with the best evaluation metrics has finally been chosen to be deployed. Our work procedure is shown in Fig. 5.

Fig. 5. Working sequences of the proposed diabetes prediction system.

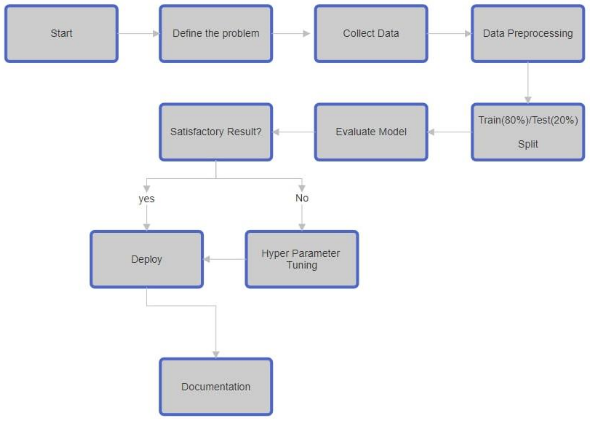


Fig. 5. shows the Working sequences of the proposed diabetes prediction system.

III. Results and Discussion

We fine-tune the settings of our machine-learning models to make them even better at predicting things. On top of that, we also look at what other researchers have done in this area, showing how our system takes things to the next level. Additionally, we use the LIME explainable AI library to understand how our models make predictions. This helps us see what factors are essential in determining air quality and makes our models more transparent and easily understood.

TABLE II. MODEL RESULTS IN DEFAULT SETTINGS

References	ML model	Accuracy	Precision	Recall	F1 Score
[1]	Random Forest	71%	0.72	0.71	0.72
[2]	KNN	65%	0.69	0.65	0.66
[3]	XGBoost	71%	0.70	0.71	0.71
[4]	AdaBoost	73%	0.73	0.73	0.73

TABLE II. shows the employed models results in default settings

We got low accuracy with default settings. To improve the accuracy, we used hyperparameters.

TABLE III. MODEL RESULTS IN HYPERPARAMETER SETTINGS

References	ML model	Accuracy	Precision	Recall	F1 Score
[1]	Random Forest	78%	0.79	0.76	0.77
[2]	KNN	82.63%	0.83	0.80	0.79
[3]	XGBoost	83.19%	0.82	0.83	0.83
[4]	AdaBoost	80.13%	0.81	0.80	0.80

TABLE III. shows the employed model results in hyperparameter settings

There are various hyperparameter values in a model. We trained with all possible hyperparameter combinations and found the optimized model to predict air index quality. The hyperparameter value range as well as optimized values are shown in TABLE IV.

TABLE IV. HYPERPARAMETER VALUES' RANGES FOR ALL THE ML MODELS

Model	Hyperparameter Value Range	Optimized value
Random Forest	'n_estimators': [200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000], 'max_features': [auto, sqrt, log2], 'max_depth': [10, 120, 230, 340, 450, 560, 670, 780, 890, 1000], 'min_samples_spli't': [2, 5, 10, 14],' min_samples_leaf': [1, 2, 4, 6, 8], 'criterion': [entropy, gini]	'n_estimators': 1800, 'min_samples_split': 2,' min_samples_leaf': 1, 'max_features': log2, 'max_depth': 560, 'criterion': entropy
KNN	'n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20], 'weights': [uniform, distance], 'metric': [euclidean, manhattan, chebyshev, minkowski],' n_iter': 50, scoring: accuracy,' cv': 3, verbose: 1,'n_job's: -1, random_state: 20	'n_neighbors': 2, 'weights': distance, 'metric': manhattan
XGBoost	'n_estimators': [200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000], 'learning_rate': [0.01, 0.1, 0.2, 0.3], 'max_depth': [3, 5, 7, 9], 'min_child_weight': [1, 5, 10], 'gamma': [0, 0.1], 'subsample': [0.8, 0.9, 1.0]	'subsample': 0.8, 'n_estimators': 800, 'min_child_weight': 1, 'max_depth': 9, 'learning_rate': 0.1, 'gamma': 0.1
AdaBoost	'n_estimators': [50, 100, 200, 300], 'learning_rate': [0.01, 0.1, 0.2, 0.3], 'base_estimator__max_depth': [3, 5, 7, 9], 'base_estimator__min_samples_split': [2, 5, 10], 'base_estimator__min_samples_leaf': [1, 2, 4]	'n_estimators': 300, learning_rate: 0.1, 'base_estimator__min_samples_split': 5, 'base_estimator__min_samples_leaf': 2, 'base_estimator__max_depth': 9

Table IV. shows the hyperparameter values' ranges for all the ML models.

Fig. 6. Machine learning model prediction interpretation by LIME explainable AI library.

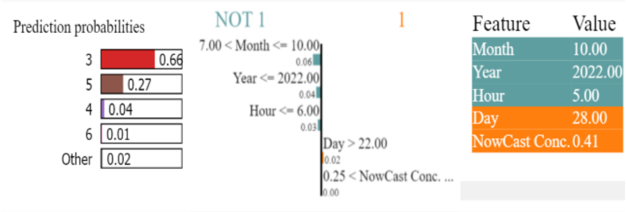


Fig. 6. shows the Machine learning model prediction interpretation by the LIME explainable AI library where 1 is positive and NOT 1 is negative.

TABLE V. COMPARISON OF OUR MODEL RESULTS WITH EXISTING WORKS

References	ML model	Accuracy
[1]	Random Forest	78%
[2]	KNN	82.63%
[3]	XGBoost	83.19%
[4]	AdaBoost	80.13%
[5]	Adab	42.90%
[6]	DT	97.68%
[7]	RFR	90.97%
[8]	Catboost	99.98%

TABLE V. shows the comparison of our model results with existing works

IV. Conclusions

The forecast for Dhaka's AQI and AQI categories between 2022 and 2023 was examined in this study. The winter season saw an increase in AQI levels beginning in October, followed by an abrupt drop in AQI starting in April. The most essential element for determining the AQI category was discovered to be Nowcast Concentration. Compared to other machine learning models, the results show that XGBoost produced the best cross-validation score of 84.9% following hyperparameter optimization. Adding more data for future models could increase their accuracy. Alternative feature engineering methods, such as one-hot encoding, can be applied to enhance the model's effectiveness.

References

[1] D. Iskandaryan, F. Ramos and S. Trilles, "Graph Neural Network for Air Quality Prediction: A Case Study in Madrid," IEEE Access, vol. 11, pp. 2729-2742, 2023.

[2] C. Liu, G. Pan, D. Song and H. Wei, "Air Quality Index Forecasting via Genetic Algorithm-Based Improved Extreme Learning Machine," IEEE Access, vol. 11, pp. 67086-67097, 2023.

[3] S. Al-Eidi, F. Amsaad, O. Darwish, Y. Tashtoush, A. Alqahtani and N. Niveshitha, "Comparative Analysis Study for Air Quality Prediction in Smart Cities Using Regression Techniques," IEEE Access, vol. 11, pp. 115140-115149, 2023.

[4] Y. Cao, D. Zhang, S. Ding, W. Zhong and C. Yan, "A Hybrid Air Quality Prediction Model Based on Empirical Mode Decomposition," Tsinghua Science and Technology, vol. 29, pp. 99-111, 2024.

[5] D. Kothandaraman, N. Praveena, K. Varadarajkumar, B. Madhav Rao, Dharmesh Dhabliya, Shivaprasad Satla and Worku Abera, "Intelligent Forecasting of Air Quality and Pollution Prediction Using Machine Learning," Adsorption Science & Technology, 2022.

[6] S. K. Natarajan, P. Shanmurthy, D. Arockiam, B. Balusamy, and S. Selvarajan, "Optimized machine learning model for air quality index prediction in major cities in India," Sci Rep, vol. 14, p. 6795, 2024.

[7] N. S. Gupta, Y. Mohta, K. Heda, R. Armaan, B. Valarmathi, and G. Arulkumaran, "Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis," Journal of Environmental and Public Health, vol. 2023, pp. 1–26, 2023.

[8] G. Ravindiran, G. Hayder, K. Kanagarathinam, A. Alagumalai, and C. Sonne, " Air quality prediction by machine learning models: A predictive study on the Indian coastal city of Visakhapatnam," Chemosphere, vol. 338, 2023.

[9] AirNow, "U.S. Embassies and Consulates: Bangladesh - Dhaka," available. at: https://www.airnow.gov/international/us-embassies-and-consulates/#Bangladesh$Dhaka. Accessed: Jun. 8, 2024.

[10] exploreASEAN, "Population Density," [Online]. Available: https://exploreasean.ch/tag/population-density/. Accessed: Jun. 8, 2024.

[11] Y. Zheng, S. Wang, R. Calhoun, and Z. Zou, "Air quality index prediction with a deep learning model: a case study in Beijing, China," Environmental Science and Pollution Research, vol. 27, no. 17, pp. 21008-21019, 2020.

[12] Z. Wang and J. Ma, "Air quality index prediction using long short-term memory and random forest models," International Journal of Environmental Research and Public Health, vol. 17, no. 19, p. 710.

[13] J. M. Gutiérrez, E. García-Gonzalo, M. Carreras, and E. Yubero, "Air quality index prediction with machine learning models: A case study in the Barcelona Metropolitan Area," Sustainability, vol. 13, no. 8, p. 4196, 2021.