# DHAKA CITY AIR POLLUTION PREDICTION USING MACHINE LEARNING TECHNIQUES

BY

SYED FAHIM AL SHARIAR

ID: 191-15-2501

## FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

SHAYLA SHARMIN
Senior Lecturer
Department of Computer Science and Engineering
Daffodil International University


DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

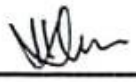
JANURAY 2025

# APPROVAL

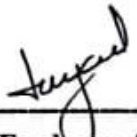
This Project titled **"Dhaka City Air Pollution Prediction Using Machine Learning Techniques"**, submitted by Name: **Syed Fahim Al Shariar** , ID: **191-15-2501** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 13 January, 2025.

## BOARD OF EXAMINERS

**Dr. Sheak Rashed Haider Noori** — **Chairman**
**Professor and Head**
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



**Most. Hasna Hena** — **Internal Examiner**
**Assistant Professor**
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



**Md. Ferdouse Ahmed Foysal** — **Internal Examiner**
**Lecturer**
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

**Dr. Md. Arshad Ali** — **External Examiner**
Professor
Department of Computer Science and Engineering
Hajee Mohammad Danesh Science and Technology University

## DECLARATION

We so certify that we, **Shayla Sharmin, Senior Lecturer, Department of Computer Science and Engineering,** Faculty of Science and Information Technology, **Mr. Abdul Muntakim, Lecturer, Department of Computer Science and Engineering,** Faculty of Science and Information Technology, Daffodil International University, have supervised this study. We further declare that no portion of my study, nor any portion of it, has been submitted for consideration for a degree elsewhere.

**Supervised by:**

Shayla Sharmin

**Shayla Sharmin**
Senior Lecturer
Department of CSE
Daffodil International University

Muntakim

**Mr. Abdul Muntakim**
Lecturer
Department of CSE
Daffodil International University

**Submitted by:**

Syed Fahim Al Shariar

**Syed Fahim Al Shariar**
ID: 191-15-2501
Department of CSE
Daffodil International University

# ACKNOWLEDGEMENT

Initially, we would want to sincerely thank and feel grateful to the Almighty God for His divine favor, which has enabled us to successfully finish the final year thesis.

To Supervisor **Shayla Sharmin,** Senior Lecturer in the CSE Department of Daffodil International University in Dhaka's Faculty of Science and Information Technology, we sincerely thank you and express my great gratitude. In the end, we finished my study on "**Dhaka City Air Pollution Prediction Using Machine Learning Techniques**" because to his immense knowledge, eager interest, and encouraging guidance. It has been made possible to finish this assignment by their unending patience, academic direction, support, encouragement, active supervision, constructive criticism, insightful counsel, reading several subpar versions, and persistent correction at every level.

For his invaluable support and guidance in seeing my project through to completion, we would like to extend my sincere gratitude to **Shayla Sharmin**, Senior Lecturer and __, Professor and Head, Department of Computer Science and Engineering, Faculty of Science and Information Technology, DIU; and to other faculty members and the staff of the CSE department at Daffodil International University.

Lastly, once again, we would want to express my gratitude to all of my supporters, friends, family, and elders for their encouragement and support. Work hard and thanks to everyone who inspired and helped with this study.

Finally, we would want to respectfully thank my parents for their unwavering support and patience.

# ABSTRACT

Clean air is crucial for animal life as well as human health since it is associated with a number of deadly illnesses, including cancer. However, due to the world's rapid urbanization and population growth, activities including housing, industries, ships, and farming contribute to air pollution. As a result, pollutants in the air have become a severe problem in many cities, especially in developing countries like Bangladesh. Maintaining indoor air quality requires frequent monitoring and forecasting of air pollution. As such, ML has demonstrated potential in predicting the air quality index (AQI) more accurately than conventional methods. An indicator of the condition of the air is the index for air quality (AQI). It computes the short-term impact of moderate exposure on an individual's health. The AQI's mission is to raise public awareness of the harmful effects that nearby contaminants have on health. The quantity of pollutants in the environment has significantly increased in Indian cities. By using the AQI for Bangladesh's capital, Dhaka, we are focusing on a few variables, starting with PM2.5 in 2017 and going up to 2022. The goal of the study is to find out how successfully NLP techniques identify and classify activity in AQI categories. Using labeled data, controlled instruction teaches an algorithm how to accurately forecast outcomes and classify AQI data. For this purpose, a variety of machine learning models were employed, including XG Boost, a Random Forest, K-Nearest Neighbors, Naive Bayes, and Linear Regression. Following data analysis, the most accurate classifier, with a 99.81% classification accuracy, was the Random Forest classifier for AQI values that fell into six categories: Hazardous, Unhealthy, Very Unhealthy, Good, Moderate, and Unhealthy for Sensitive Groups. To produce a web prototype, the AQI category is finally classified using a Random Forest model.

# TABLE OF CONTENTS

# LIST OF FIGURES

# List of Tables

# CHAPTER 1
# Introduction

## 1.1 Introduction

Human life is dependent entirely on air. Its caliber has to be observed and understood for our own wellbeing. Due to airborne pollutants, a large number of people globally suffer from respiratory diseases and other health problems. The single greatest environmental threat, according to actual statistics, is air pollution. Population growth has been accelerated by the toxic gasses produced by rapid development. The air pollution that poses a serious threat to human health is highly detrimental. There has been a significant decline in air quality due to uncontrolled pollution. The Air Quality Index, also known as the AQI, is a tool used to quantify and report pollution levels. Twelve factors, or airborne pollutants, make up the ethereal quality index (AQI). These factors include the following: the amount of oxides of nitrogen (NO2), acid rain (SO2), carbon monoxide (CO), the layer of ozone (O3), PM10 particles (which are substantial particles with a dimension of ten tiny particles or less), PM2.5 particles (which are significant particles with a length that is two microns or less), potassium hydroxide (ammonia), and a mixture called at The six contaminants—PM10, PM2.5, SO2, nitrogen dioxide, CO, and O3—are used to calculate the air pollution indicator (AQI) in various applications. However, the actual choice of pollutants is contingent upon the particular objective as well as other aspects, such as the availability of data, the methods of measurement, and the frequency of the surveillance. A high AQI indicates extremely dirty air, which may be harmful to health. You can monitor the current state of the air quality with the AQI. Numerous meteorological centers in our local region also monitor the AQI on a daily or weekly basis. Utilizing this data for the suggested project is the aim of the data mining and harvesting process.

The primary driver of global warming is the release of greenhouse emissions (GHGs), which also modify plant-soil dynamics and alter the temperature, both of which have negative impacts on agriculture, the environment, and human activities (Malhi et al.,

2021). utilizing information from 2010 to 2019, the World Health Assembly (WHO) published a research on worldwide air quality in 2022.

This research examined a wide range of air pollutants stated above, based on analyses of 6743 places in 117 countries worldwide. It revealed that PM2.5 was expanding globally, responsible for nearly two million yearly fatalities inside India alone.

An elevated AQI score indicates an area where people are most likely to die. Consequently, AQI surveillance and prediction have emerged as crucial tools for ensuring long-term prosperity on a worldwide scale (Rybarczyk and Zalakeviciute, 2021). Numerous scholars have devised techniques to forecast the AQI by utilizing mathematical, predictable, scientific, ml, and dl ideas. In complex circumstances, models of choice and statistics are not adaptable enough. According to recent advancements in sensors, it is now possible to quickly identify varying degrees of air pollution, thus the AQI is computed instantaneously. Making advantage of the readily available data sets makes the AQI prediction straightforward (Bekkar et al., 2021). The computer's learning algorithm predicts the AQI with high accuracy and reliability in all cases. Our ability to provide increasingly accurate AQI forecasts is made feasible by machine learning (ML), which is made possible by the growing quantity of historical data that is accessible for analysis. They are multiplying as they attempt to demonstrate that they constitute a good alternative to established statistical methodologies in time series forecasting. The dynamics of the highly unpredictable systems influencing pollution levels are not well understood, making it exceedingly difficult to create a statistical model to anticipate such events. A machine learning model is an excellent illustration of a proportional and nonlinear technique as it only requires historical data to determine the relationship between the independent variables. This enables us to create a prediction model that is more accurate.

## 1.2 Motivation

My goal in doing this research is to apply ML and AL to classify different types of air pollution according to AQI readings. After some consideration, we were reluctant to arrive at with an article design that would meet the requirements of the research. We thus requested assistance from some of my favorite teachers. Given that several bird species are supported by the current ecology, it was proposed that we look into a relevant concept in relation to this issue. We also see how society is progressing via fresh perspectives and how researchers are investigating these themes further in thoughts to better develop my own. This was the reasoning behind my decision to produce a paper titled "**Dhaka City Air Pollution Prediction Using Machine Learning Techniques**." These were the driving forces behind this kind of research-based work. This is why machine learning is so important: intelligent equipment has connected things inside of me.

## 1.3 Objective

One of the most effective methods for tackling problems is machine learning (ML). Relevant data may be categorized into air pollution using supervised learning, an artificially intelligent approach. First, train the study set (labeled data necessary) using efficient models to determine whether any AQI category is comparable to any other of the six groups. This is the main goal of guided machine learning. AQI categories can be identified using artificial intelligence approaches. Consequently, we had the ability to determine the following objectives:

- Predicting each of the six AQI subcategories using machine learning.
- To gather data in order to predict air pollution.
- Gaining a thorough grasp of the domains associated with machine learning.
- Using a variety of tactics to improve results.
- Addiction is categorized using AQI categories.

## 1.4 Methodology

The phases of our study methodology are displayed, which can be used to calculate the air quality category index (AQI) values for the contaminants that comprise each index. Since our online data was compiled from a variety of sources, it is true and accurate. A value of the AQI for every kind of air was used to construct the dataset. We have reviewed this data gathering process, removed any unnecessary numerals of null values, and tidied up the wording to ensure that the whole data set only contains accurate and relevant information. To investigate machine learning methods, we build and improve models using data that already exists. We also utilize permutation approaches to address the issue of class heterogeneity, which will guarantee equal inclusion and increase the overall efficacy of the model. Apart from giving a summary, our approach searches for ways to reduce the amount of erroneous and imprecise findings. By combining language understanding methods with machine learning technology, we can offer a comprehensive solution that manages the results in the air quality index numbers that are utilized to forecast pollution in the air.

## 1.5 Project Outcome

This analysis aimed to ascertain if phrases associated with the measure of the air quality (AQI) were present. The objective is to accurately determine the value of an AQI remark. The six categories into which we have categorized the numerous and varied kinds of statements that were communicated are "Hazardous," "Unhealthy," "Very Unhealthy," "Good," "Moderate," and "Unhealthy for Sensitive Groups". Given the relevance and size of the dataset utilized for this investigation, we were able to get results with a respectable level of precision. Utilize machine learning in this study to extract a select handful of the dataset' most accurate results. The instructional set's richness plus the oversight of the learning strategy's efficacy will determine the model's accuracy. The AQI category is found within the AQI supplied values by our data mining techniques. For the goal of identifying harmful material, we are able to apply algorithms to sets and anticipate outcomes 100% of the time. To produce the

best report possible, the accuracy, recall, error rates, and reliability of results of several algorithms were assessed.

### 1.6 Organization of the report

In Chapter 1, the study's objectives, worries, research questions, and anticipated results were described. The report's overall structure is also covered in this section.

Chapter 2 contains every previous study done in this area. In the portion that follows, they give an example of the scope that arises from their limiting of this study question. The main challenges or obstacles to this study were the subject of the conversation that concluded. This chapter includes parts on relevant research summaries and studies, as well as a discussion of the challenges experienced throughout the project's development.

Chapter 3 offers a theoretical evaluation of the research's findings. This chapter contains further information on the statistical techniques, especially those applied in the investigation's arithmetic portion. This section also offers examples of practical uses of machine learning techniques. The procedures for gathering stats and the system that is used to compile them are covered in the section that follows. A single-family matrix for confusion analysis is used in the last part to evaluate the model and provide a valid tag for the classifier. Application assessment is necessary while utilizing machine learning techniques to guarantee real correctness. The study subject and methodology, operational efficacy, data collecting strategy, handling data, recommended methodology, style of teaching, and requirements that must be satisfied in order for this project to move forward are all included in this part. Each machine learning method and classifier employed in this study gets a detailed explanation.

In Chapter 4, the study's results, an assessment of the results, and a discussion of the implications are presented. To help with the project's implementation, a few test shots are supplied in this chapter. This chapter ends with a summary of the findings

and an application of the machine learning techniques. Furthermore, provide an explanation of the web-based tool that uses AQI values to determine the air pollution index (AQI).

An outcome, an explanation of the intended course of action, and a summary of the study were provided in chapters 5 and 6. The next part contains a validated sample demonstrating that the report's construction conforms with all standards. Impacts on the Sustainable Development Goal, the environment, and society at large The chapter's conclusion highlights the restrictions on our work, which may have an impact on the subsequent generations of professionals in our sector.

# CHAPTER 2

# Background

## 2.1 Introduction

This option includes the investigation's findings, challenges faced, pertinent literature, and a research review. I will review research by various writers and discuss the relationships between their methodologies and conceptual correctness within the "related Works" section. This section on similar works will provide a review of the papers, methods, and reliability of other scholarly works which are relevant to my research. In the overview unit of the study, I will give a summary of my connected work. We outline our methods for resolving any difficulties that arose during the study and how we increased the accuracy of each layer throughout the demanding stage. All of this has previously been discussed. Datasets analysis is an integrated field that uses cycles, logical processes, calculations, and frameworks to extract information and insights from both organized and unstructured data, and then applies those insights and important facts to a wide range of applications. Hazardous substances like PM2.5, PM10, including so on may cause lung tissue cell death, ischemic cardiac disease, chronic pulmonary obstructive conditions (COPD), severe bronchial asthma, and pediatric pneumonia. Particulate pollutants in the air has been linked to strokes, which happen when the blood supply to a person's brain is cut off.

## 2.2 Literature review

They looked at the relationship between many air metrics at the outset of the research [1], such as the AQI, PM2.5 concentrations, total NOx (nitrogen oxide) levels, and others. After building forecasting models with random forest regression and support vector estimation, they used r, Rc-squared (R-Mean), and arithmetic methodologies to assess the performance of the two models. An effective machine learning method known as SVR is used to estimate the state of the air gauge and assess the levels of particles and contaminants [2]. The findings demonstrate that, for the entire state of

California, per hour air quality indexes (AQI) as well as the amount of various pollutants, such as dioxide of sulfur, nitrogen dioxide, carbon monoxide, acid sulfur, outside ozone, and particulate matter in fines 2.5, can be accurately anticipated using SVR using the RBF kernel. Ninety-one percent of the unprocessed information used for validation were properly categorized into one of the six AQI categories provided by the EPA at the request of the nation (dataset).

The AQI forecast using machine learning techniques, such as learning regression and time series analysis. To forecast the AQI, a supervised artificial intelligence technique known as MLR was used. Numerous quantitative criteria were used to assess the success. In addition, the AQI was projected into the future using an updated ARIMA series theory. Both models demonstrated exceptionally high precision and efficacy in AQI prediction [3]. A combination framework combining artificial neural networks and the Kriging method was used to evaluate the quantity of contaminants across the air at different places in Mumbai and Navi Mumbai in the country of India. It was found that the elevated R values were really responsible for achieving the necessary degree of variance between the observed and projected results. In terms of predicting and R value, ANN outperformed the standard regression model [4]. using variables like PM2.5, PM10, SO2, and NO2, to predict the author's AQI intensity. Ultimately, the randomized forest regression approach yielded the best results out of the selection tree, SVR, with RFR regression techniques, with a level of accuracy of 0.99985 on the experimental data, a median variance inaccuracy of 0.00013, and an average absolute mistake of 0.00373 altogether [5]. The AQI was anticipated by using data from the prior year and projecting over a certain future year as the slope lowering improved the multivariate regression issue. They outperformed typical regression models by optimizing their predictive ability and utilizing projected costs for the forecast problem. They also assessed the level of resemblance between the alternatives and the ideal solution to ascertain the priority order of requests employing the AHP the MCDM technique [6]. A logistic regression technique [7] was utilized to determine the contamination level of the provided data sample of daily atmospheric as well as

surrounding conditions in a specific city. With the use of historical PM2.5 data, this approach attempted to predict PM2.5 levels and determine air quality. According to the results, auto or logistic regression may be used to forecast future amounts of PM2.5 & air quality. An ML approach was presented in this article [8] that used six years of meteorological and pollutant data to anticipate the quantity of PM2.5 from breezy (the velocity & azimuth) and moisture levels. The findings showed significant accuracy of the filtering model in classifying the lowest (10g/m3) versus big (>25g/m3) levels of PM 2.5 as well as the smallest (10g/m3) compared moderate (10–25g/m3) PM2.5 quantities. An ANN and the Kriging procedure approach were used in a combined approach that integrated historical data from the meteorological division and the Air Quality Control Board in order to estimate the number of airborne particulate matter in Bombay and the Navi Mumbai region [9]. The recommended framework was then developed and assessed using the MATLAB software for ANN and the R programmed for Kriging. The vast amount of pollution data was evaluated and future pollution was forecasted with the use of technologies. Additionally, additional information sources for the environment's quality forecast were identified through the study of time series. The possibility of forecasting hourly pollution levels using a deep RNN with LSTM to predict Delhi's AQI was explored. Even in the weekly estimates, the results were accurate. Deep learning-based strategies beat conventional statistical approaches, according to the results [10] [11].

### 2.2.1 Related research

Table 2.1: Comparison with earlier research

| References | Purpose | Datasets and Parameter | Tools and classifier | Accuracy |
|---|---|---|---|---|
| [12] | To calculate the air quality index (AQI), as well as the quantity of pollutants and particles. separation into six AQI categories. | The ozone layer, PM2, temperature, humidity, wind, CO, and sulfur dioxide (SO2) NO2.5. Both May 1, 2018, and January 1, 2016. All in all, 102090 data were used. | Support vector classifier | 94.1% |
| [13] | This study looks at air pollution data from 23 Indian cities over a six-year period in order to anticipate and assess air quality. | The data collection was gathered between January 2015 and July 2020 and covers 23 different Indian locations. It contains 29,531 instances of 12 different attributes.CO, SO2, O3, NO2, PM2.5, PM10, and so forth. | KNN, GNB, SVM, RF, XG Boost | **Testing set:** KNN:85%, GNB:83%, SVM:78%, RF:86%, XG Boost: 90% |
| [14] | The investigation and forecasting of air quality by machine learning | The dataset has twelve features in total. | Random Forest, Support vector machine, Artificial | ANN: 90.4%, RF:93.5%, SVM: 99.4% |

| | | | | |
|---|---|---|---|---|
| | techniques is covered in this work. | | neural network | |
| [15] | This study aims to reduce pollution by tackling the issue of Air Quality Index (AQI) estimation. | PM 10, PM2.5, which are atmospheric carbon monoxide (CO), lead (Pb), ozone (O3), sulfur dioxide (SO2), nitrogen dioxide (NO2), and ammonia (NH3). | Neural Networks, SVM | Neural Networks:91.62%, SVM: 97.03% |
| [16] | The goal of this study is to find an alternative method for characterizing and monitoring air quality. | MQ2, MQ5, and MQ135 gas sensors, as well as the relative humidity sensor and AQI 5 categories e. | KNN, SVM, NB, RF and Neural Networks | KNN:98.67%, SVM:97.78%, NB:98.67%, RF:94.22%, Neural network: 99.56% |

## 2.3 Gap analysis

Several areas for development are highlighted by a gap analysis for machine learning (ML)-based air pollution prediction. The accuracy of predictions is impacted by the insufficient, inconsistent, or missing data in many of the models now in use. Additionally, existing models frequently underrepresent environmental elements such local terrain and weather. Understanding prediction processes can be tough due to the interpretability of machine learning models, particularly deep learning. Additionally, real-time decision-making tools for health warnings or policy changes are not integrated. Better data, more complicated models, and greater real-time capabilities are needed to close these gaps.

## 2.4 Summary

This section contains every previous study done in this area. In the portion that follows, they give an example of the scope that arises from their limiting of this study question. The main challenges or obstacles to this study were the subject of the conversation that concluded. This chapter includes parts on relevant research summaries and studies, as well as a discussion of the challenges experienced throughout the project's development.

# CHAPTER 3

# Research Methodology

## 3.1 Methodology/Requirement Analysis & Design Specification

### 3.1.1 Overview

The research approach is covered in the section that follows, along with instructions on how to gather datasets, carry out each test, and use each model to improve accuracy. Additionally, all-data research and a suggested approach were provided in this chapter. Consequently, in an attempt to enhance and streamline the data presented in this chapter. In addition to providing a thorough explanation of the entire procedure, this study part intends to highlight the methods utilized to determine the quality of air pollution in data. This section will detail the entire study methodology. There are several approaches to solve every given analysis. Selecting an ML strategy is the next stage. As we have already mentioned, since we are using five different machine learning techniques, building the framework and running the algorithm need first building a data store. After then, the collected data is used to train the model. This is the foundation for feature selection. Next, sets of training and testing data were created from the data. The term "training dataset" is frequently used interchangeably with "testing dataset." Only a significant portion of the data needed to assess our model is available to us once the input is extracted from the data set for training and matched into multiple ML method models. Following that, the model's reliability is assessed. Several of the methods will be explained using formulas and illustrations to help you grasp them better. We have been explaining with the help of our straightforward process flow graphic. This is a summary of the research endeavor's methodology as well as the whole study effort. Among the most important components are the data obtained for the examination, and the proposed model. With the right graph, explanation, layout, and formula, the idea is better clarified.

### 3.1.2 Proposed Methodology

This research probably used a number of different approaches or techniques to get the desired outcomes. In this experiment, the text and all values were cleaned up, the workflow, the process, gathering, and AQI level were selected, and the classifier's performance was assessed using the randomly chosen forest classification algorithm's output.

**Step 1: Collection of Data:** After gathering the data from internet resources, such as Kaggle, we evaluated it. There isn't a large, comprehensive dataset accessible in this industry since it's difficult to gather information for the specific air pollution content of high-quality analyzer and categorization type.

**Step 2: Processing of Data:** Each piece of data was examined separately once all practical means of data collection had been used. There are many instances of incorrect and imprecise language all around us. Before utilizing the chosen dataset, we are advised to go over its final part.

**Step 3: Prepare of Data**: When the dataset is set up, the "AQI" and "AQI category" continue to direct the data preparation and growth. Training requires organizing the data, eliminating null values, and showing it. There hasn't been much preparation done to get the information ready for separation.

**Step 4: Choosing of Models:** We choose a prediction approach, train it using my data, and then assess it to increase reliability. Filters are heavily used in machine learning. Several designs were used to enhance the materials design and enable the ML model to determine the type of air pollution, but ultimately, only one device was selected to evaluate the accuracy of the data.

**Step 5: Evaluation of Performance:** This phase's later sections address all the consequences. After training and testing, these approaches gave us a limited level of reliability for each of the two distinct air quality datasets label groups. To bolster the

confusion matrices, f1 ratings and accuracy data were generated. Applying machine learning algorithmic learning techniques, ascertain if all kinds of contaminants in the natural environment can be represented by AQI values.

**Step 6: Conclusion and Future work:** There will be a summary of the growth plan for this field given.

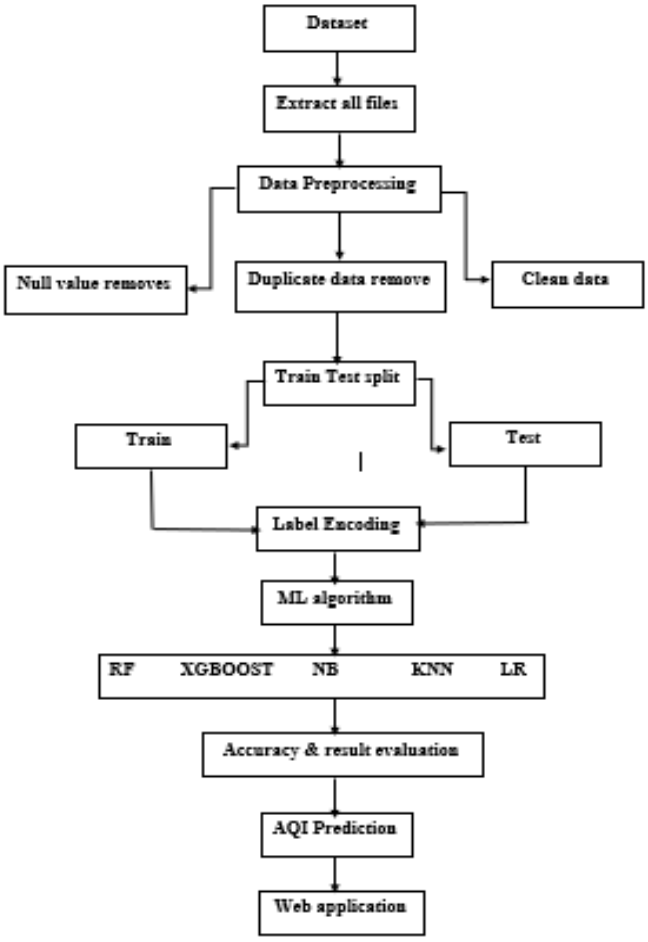



Fig 3.1: Procedure for the Whole Suggested Study.

The phases of our study methodology are displayed in Fig. 3.1, which can be used to calculate the air quality category index (AQI) values for the contaminants that comprise each index. Since our online data was compiled from a variety of sources, it is true and accurate. A value of the AQI for every kind of air was used to construct the dataset. We have reviewed this data gathering process, removed any unnecessary numerals of null values, and tidied up the wording to ensure that the whole data set only contains accurate and relevant information. To investigate machine learning methods, we build and improve models using data that already exists. We also utilize permutation approaches to address the issue of class heterogeneity, which will guarantee equal inclusion and increase the overall efficacy of the model. Apart from giving a summary, our approach searches for ways to reduce the amount of erroneous and imprecise findings. By combining language understanding methods with machine learning technology, we can offer a comprehensive solution that manages the results in the air quality index numbers that are utilized to forecast pollution in the air.

### 3.1.3 Implementation Requirements

An area of study is an inquiry field that is looked at and explored to explain concepts for handling modeling, data gathering, task fulfillment, and modeling education, in addition to execution. We talk about the instruments and methods we use as measurement specialists. We made use of the Windows operating system of the business, the language for programming, Python, and a few other tools including NumPy, SkLearn, and OpenCV. For all teaching and testing, Colab by Google served as the preferred platform. Python programmers may use Google Colab to write code for methods related to data sciences and machine learning. These methods use statistical methods associated with machine learning to determine clusters, i.e., the 6-the air quality indices type categorization.

Libraries used:

- **Matplotlib:** Plotting, scoring, organizing, and graphing processes are all facilitated by the Pyplot package of tools from Matplotlib. This might be employed in form-building to draw emphasis to specific viewpoints inside a story or to draw attention to the boundaries of the story's authority.
- **NumPy**: The NumPy toolkit in the language has simplified vector processing. This topic includes detailed coverage of matrix calculations, transformation indices, and a Fourier inverted transform. The NumPy is Python packages include many tools and methods for working with different kinds of matrices. Designing gadgets may be made more rational and realistic with the aid of NumPy. To put it quickly, NumPy is a Python library primarily intended for numerical evaluation. Another way to put this would be "estimates. Py."
- **Scikit-learn**: Sklearn is a powerful and intuitive tool for data analysis and modeling. Three Python tools were used in its design: NumPy, SciPy, and Matplotlib. These are open-source, publicly accessible tools that any individual may utilize.
- **Seaborn:** You may utilize matplotlib with the upcoming edition of this renowned the Python data visualization utility. This is a simple-to-use application for artistic data visualization.
- **H5py**: Utilizing the h5py library for Python, users can access unstructured HDF5 code. Much of the data, mostly integers, are handled by NumPy and then HDF5 is used for storage.
- **TensorFlow:** This AI technology collection seems to be publicly available. It provides a wide range of tools for building and using various machine learning models, including models based on neural networks. TensorFlow's versatility, efficiency, and ease of use make it a great choice for many intelligent applications.
- **Pandas:** For evaluating and working with language-specific data, Pandas provides a freeware toolbox. Particularly when managing summary data,

fundamental data types and methods for statistical analysis can support orderly data administration.

- OS: The Py System component offers a number of capabilities that enable workers to communicate with each other using the same lexicon.

The technology, software, and data capabilities utilized in this system for air pollution prediction allow for in-depth investigation. Computing systems with a processing core composed of strong CPUs and potent GPUs are known as high-performance computers. I understand that no invention can provide perfect results. In a similar way, we may adjust our model's parameters as it's being trained to increase accuracy.

1. Equipment and Software specifications

    - Operating System (Windows 7 or above)
    - Hard Disk (minimum 1 TB)
    - Ram (Minimum 4 GB)

2. Creating for Tools

    - Environment of python
    - PyCharm.
    - Google Collab.
    - VS code.

### 3.2 Detailed Methodology & Design

In addition to providing a thorough explanation of the entire procedure, this study part intends to highlight the methods utilized to determine the quality of air pollution in data. This section will detail the entire study methodology. There are several approaches to solve every given analysis. Selecting an ML strategy is the next stage. As we have already mentioned, since we are using five different machine learning

techniques, building the framework and running the algorithm need first building a data store. After then, the collected data is used to train the model. This is the foundation for feature selection. Next, sets of training and testing data were created from the data. The term "training dataset" is frequently used interchangeably with "testing dataset." Only a significant portion of the data needed to assess our model is available to us once the input is extracted from the data set for training and matched into multiple ML method models. Following that, the model's reliability is assessed. Several of the methods will be explained using formulas and illustrations to help you grasp them better. We have been explaining with the help of our straightforward process flow graphic. This is a summary of the research endeavor's methodology as well as the whole study effort. Among the most important components are the data obtained for the examination, and the proposed model. With the right graph, explanation, layout, and formula, the idea is better clarified.

### 3.2.1 Data Collection

Using the online platform Kaggle, the dataset was collected. This database includes the PM2.5 AQI assessment for Dhaka city's air pollution projection from 2017 to 2022. Six folders covering six years of Dhaka PM2.5 are included in this dataset. A total of 44,571 data points are gathered, and the rows titled "AQI" and "AQI Category" are recognized. The process of training & testing parts makes up the dataset's two sides. After the nulls and duplicates were removed, 44,160 data were still present. These findings were then used to categorize the air quality index into six groups: "Hazardous," "Very Unhealthy," "Unhealthy," "Unhealthy for Sensitive Groups," "Moderate," and "Good."

| | Date (LT) | Hour | NowCast Conc. | Raw Conc. | Conc. Unit | AQI | AQI Category | QC Name |
|---|---|---|---|---|---|---|---|---|
| 0 | 01/01/2017 01:00 | 1 | 287.8 | 287 | ug/m3 | 338 | Hazardous | Valid |
| 1 | 01/01/2017 02:00 | 2 | 297.4 | 307 | ug/m3 | 347 | Hazardous | Valid |
| 2 | 01/01/2017 03:00 | 3 | 300.2 | 303 | ug/m3 | 350 | Hazardous | Valid |
| 3 | 01/01/2017 04:00 | 4 | 306.1 | 312 | ug/m3 | 356 | Hazardous | Valid |
| 4 | 01/01/2017 05:00 | 5 | 313.8 | 322 | ug/m3 | 364 | Hazardous | Valid |
| 5 | 01/01/2017 06:00 | 6 | 290.3 | 248 | ug/m3 | 340 | Hazardous | Valid |
| 6 | 01/01/2017 07:00 | 7 | 269.3 | 224 | ug/m3 | 320 | Hazardous | Valid |
| 7 | 01/01/2017 08:00 | 8 | 262.5 | 247 | ug/m3 | 313 | Hazardous | Valid |
| 8 | 01/01/2017 09:00 | 9 | 264.0 | 268 | ug/m3 | 314 | Hazardous | Valid |
| 9 | 01/01/2017 10:00 | 10 | 259.4 | 249 | ug/m3 | 310 | Hazardous | Valid |

Fig 3.2: Sample PM2.5 data for Dhaka from 2017 to 202.

Table 3.1 lists all of the data entries included in each and every one of these files, organizing them into eight major groups:

Table 3.1: Dataset's Column of Description

| Column's Name | Description of the Column's |
|---|---|
| Date (LT) | Date |
| Hour | Hour 0-23 |
| Nowcast Conc. | Nowcast Concentration in micrograms/cubic meter. |
| Raw Conc. | Raw Concentration in micrograms/cubic meter. |
| Conc. Unit | Concentration Unit- micrograms/cubic meter |
| AQI | Air Quality Index |
| AQI Category | "Hazardous," "Very Unhealthy," "Unhealthy," "Unhealthy for Sensitive Groups," "Moderate," and "Good." |
| QC Name | Valid, Missing |

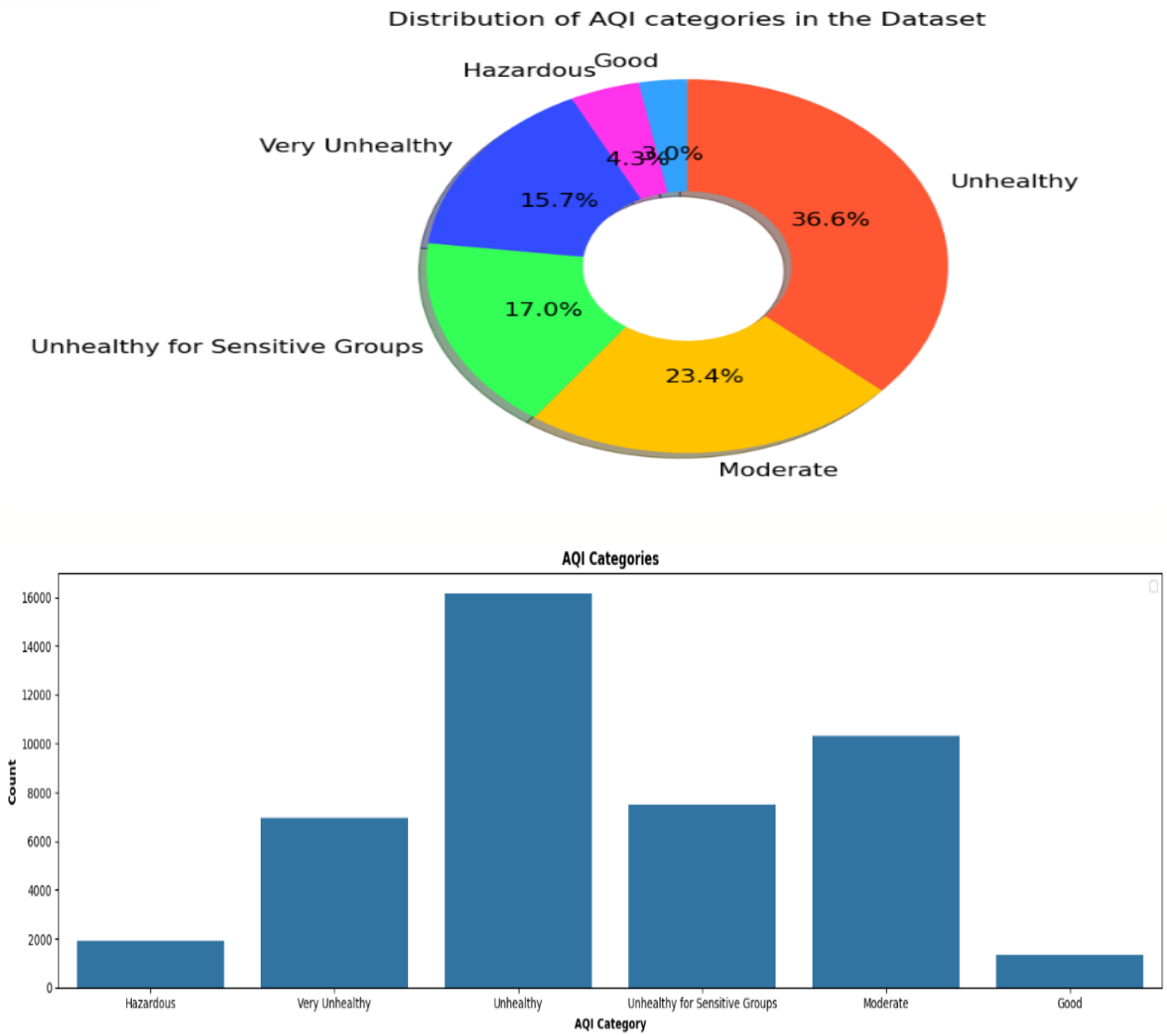



Fig 3.3: Six data classes in the AQI.

## Encoder label:

Regardless of the total amount of rows in the dataset, machine learning experts frequently work on dataset that contain several features. These unique IDs can contain both letters and integers. Keywords are frequently used in education to organize data in order to improve user comfort and comprehension.

Encoding the labels is one method of transforming signals into computer-readable text or numerical values. To create a structure that conveys numerical values,

identifiers must be translated within the framework of the previously described method. How these labels are used is ultimately up to the creators of the machine learning algorithms. After the data is supplied, a preliminary analysis has to be carried out while keeping an eye on the alterations that are being monitored.

### 3.2.2 Preprocessing data

Our goal for the data first processing stage is to transform numerical aqi data into a structure suitable for assessment and model training. In order to collect the requisite "AQI" and "AQI Category" features, data pertaining to the identification of air quality categories linked to air pollution must first be obtained. When several datasets are combined, data that may be utilized for testing and training is produced. To start fixing any mistakes, we removed any superfluous letters and symbols from the contents of the database that had null values. Another technique for turning words into vectors of numbers that may be easily concatenated and utilized in machine learning models is tokenization. Our models get increasingly skilled at using AQI values to determine AQI Category as a result of this meticulous data processing approach.

**About AQI:**

The air quality index (AQI) serves as a daily measure for evaluating the quality of the air. It is a gauge of how short-term air pollution affects human health. The purpose of the AQI is to educate people about how the air in their area affects their health. For the AQI, there are six categories. Each classification denotes a different level of health concern. Plus, every group has a different color. Because of the color, people can immediately and readily determine if the air quality within their communities is poor.

Think of the AQI as a scale, where 0 to 500 is the range. A greater AQI rating is associated with an increase in air pollution levels and the associated health risks. An AQI rating of Fifty or less, for example, indicates acceptable air quality, whereas an AQI number of 300 or above indicates dangerous air quality. Any specific pollutant

with an AQI score of 100 usually indicates a level that ambient air contamination that, for the short term, satisfies the national exterior air quality requirement for public health safety. AQI values of 100 or lower are frequently seen to be excellent. As, AQI readings soar beyond 100, air quality becomes dangerous, first for those who are susceptible and then for everyone.

Table 3.2: Category of the Air Quality Index (AQI).

| Air Quality Index | | |
|---|---|---|
| **AQI Category** | **Index Value** | **Air quality Description** |
| Good | 0-50 | Air quality is satisfactory |
| Moderate | 51-100 | Air quality is acceptable |
| Unhealthy for sensitive groups | 101-150 | Some sensitive groups have health effects. |
| Unhealthy | 151-200 | General people have health effects. |
| Very Unhealthy | 201-300 | Air quality is Health alert. |
| Hazardous | 300- Higher | Health warning of emergency conditions. |

### 3.2.3 Data Cleaning and Null value remove

Modifying numerical values is an essential stage in the creation of datasets. The two main strategies in this method are designed to raise the standard and relevance of written content. News is vetted before to publication, and articles are only kept for a certain hundred words. Our commitment to deliver top-notch, legally-compliant, and instructive information is protected by filtering techniques. To update the material methodically, a text correction strategy is also used to remove unneeded symbols such

groups, stop categories, emoji removers, special signs, etc. Several English letters, line breaks, and typical symbols have been eliminated from the data or numerical value including the null values. To ensure that everything we gather is ready for a detailed examination, we use an extensive preservation process. Including three rows for the aqi visualization of the data in the six categories.

### 3.2.3 Tokenization

Tokenization, also and padding are two key components of contemporary data use. We need to tokenize words—turn them into numerical sequences—in order for our model to understand English. Assigning a distinct number to every word establishes the link among the written word and symbolism numbers. In an arrangement training session, padding ensures that every sequence has a fixed time, which raises the likelihood of agreement. Our computers convert sentences into monetary units and verify that the lengths of the phrases vary so much that they can assess the many linguistic subtleties of the AQI score. The computing devices we use depend on this step to accurately evaluate the data and distinguish between the six distinct categories that are represented in the AQI.

### 3.2.5 Data Preparation

Even after removing redundant data and null values through the addition of four separate columns (date, month, aqi_le, and aqi_map), we did not randomly divide the information to test the model and training during the data preparation process. From the 44,160 records that comprise the air pollution quality dataset, we extract the most important "AQI" and "AQI Category" variables. The dataset consists of two sections: Train and Test. Following the elimination of duplicate entries, the data was classified as "Very Unhealthy," "Unhealthy," "Hazardous," "Unhealthy for Sensitive Groups," "Moderate," and "Good." 8,832 data are included in testing, whereas 35,328 data are included in training, which uses AQI values to analyze the air quality.

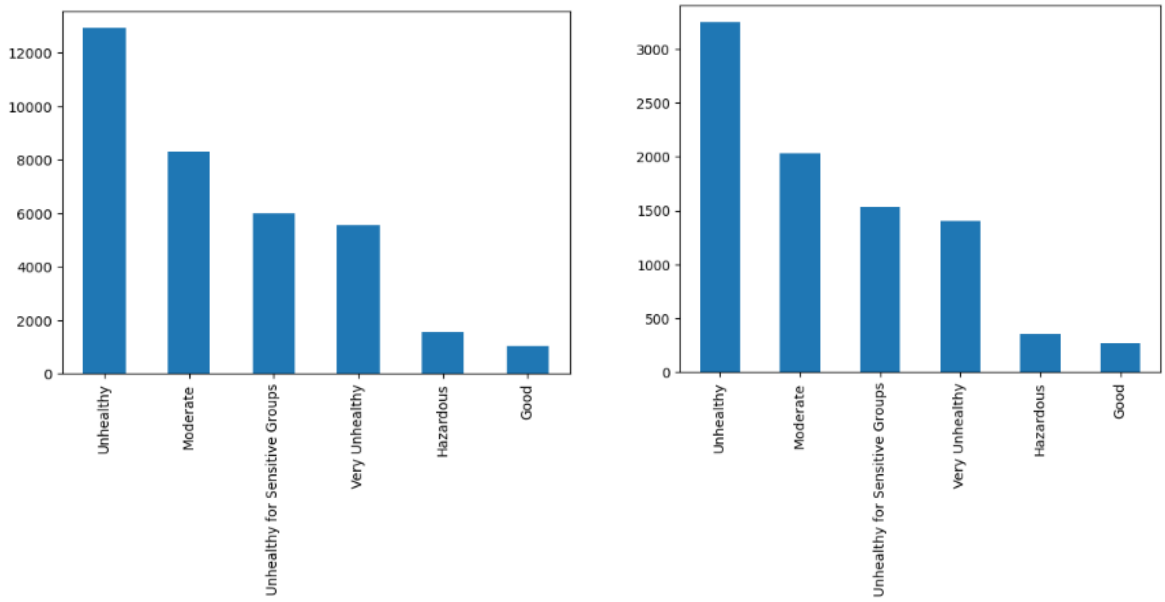



Fig 3.4: Data Count of Train and Test AQI category.

### 3.2.6 Models Applied

When the amount of PM2.5 in the air increases, it generates hazy air and decreases vision. This study computes the Dhaka PM2.5 using a variety of methods. We examined a number of machine learning methods, such as Naive Bayes, Random Forest classifier, XG Boost, KNN, and LR. We leverage the ability of our models to capture language variations, context, and patterns across a broad variety of models. Our goal of creating a reliable system for categorizing the degree of pollutants in the air based on AQI data is made possible by this careful examination.

1. **Random Forest:** For the two categories of classification, the tree-based method of the RF Classifier may be applied. A ML method creates a hierarchical tree. This method is used by artificial intelligence to create hierarchical "decision trees". Several decision trees are constructed using a combination process and then integrated in the random forest classification method. This makes the issues connected to overfitting clearer. Machine learning is one of the hottest subjects in business right now because of its adaptability and potential to be utilized anywhere there is a lot of data. Because of its many benefits, RF Encoder based

on machine learning is often chosen over other approaches. When the approach was initially presented in 1997, it was designed to work with very big datasets.

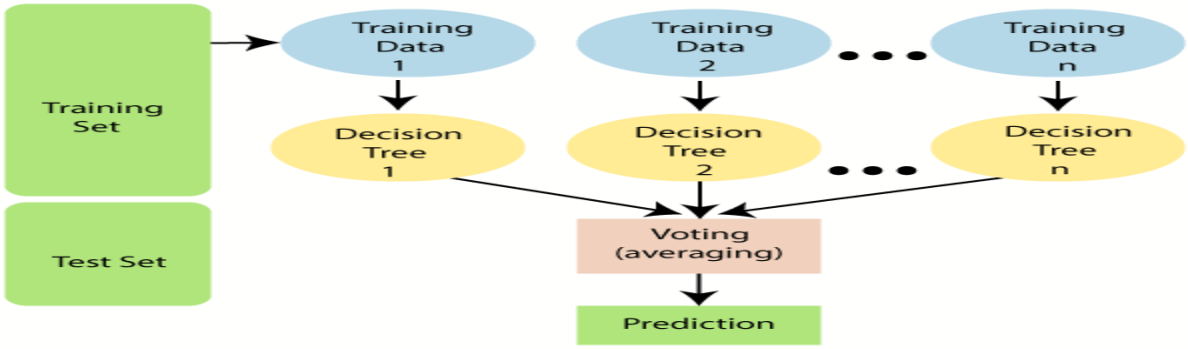



Fig 3.5: The Random Forest classifier's workflow.

2. **Naïve Bayes:** Even with thousands of records, it is recommended to use the naïve Bayes model for machine learning since it can handle large amounts of data. It excels in tasks requiring natural language processing (NLP), including text analysis for AQI categories. Filtering is an easy and quick operation. Prior to comprehending the naive Bayes classifier, we must grasp the Bayes theorem. The theory of Bayes is the topic of our first discussion. The basis of this thesis is the idea of conditional likelihood. Contingency probability is the chance that one event will occur given the possibility that another may occur. Using our prior information and the conditional probability, we may calculate the likelihood of an event. Naive Bayes methods find widespread use in systems for guidance, sentiment analysis, and spam removal, to name just a few. Notwithstanding their quickness and simplicity of use, their main disadvantage is the requirement for independent predictors. In the majority of real-world scenarios, where the predicting elements are interdependent, the classifier performs badly.

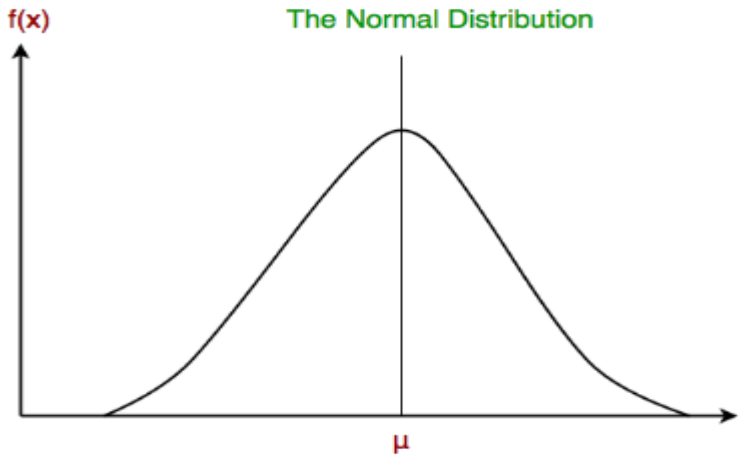



Fig 3.6: Functions of Naive Bayes.

3. **XG Boost:** Extreme Gradient Boost, or "XG Boost," is a well-liked and successful machine learning technique for resolving regression, clustering, and other problems. Producing slope-enhanced decision-tree structures is made simple and effective with this technology. XG Boost is a form of ensemble learning that gets a model stronger by aggregating the predictions of several weak learners, usually decision trees. Learning from the less fortunate students one by one, each weak student makes up for the shortcomings of the one before it. In order to prevent overfitting, XG Boost has many normalizing penalty measures. Because penalty-based regularizations yield effective training, the model is able to attain the appropriate degree of generality. What is non-linear? Unpredictable data may be used to identify and train structures using XG Boost. Verifying again: everything is built and ready to use right out of the box.

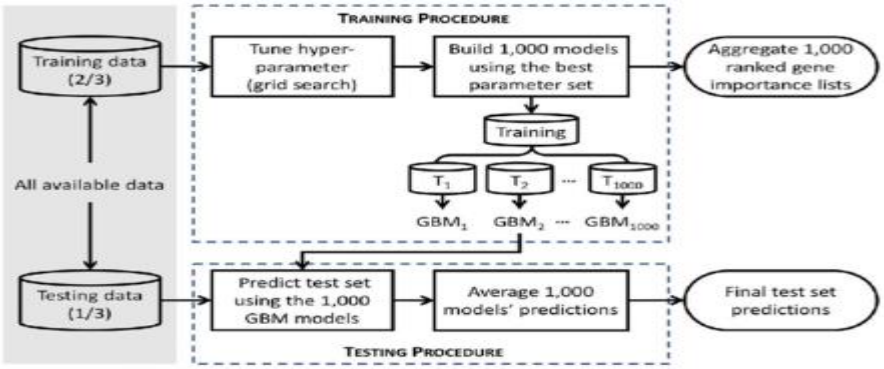



Fig 3.7: Construction of the XG Boost model.

4. **K-Nearest Neighbors:** The k-nearest-neighbors (KNN) method is a potent and approachable machine learning tool that is used to resolve regression and classification problems. The K neighboring neighbors of the most recent point of data in the set that is being trained are used by KNN, which employs the principle of similarity to predict the value or title of the most recent data point. This article will cover the k-Nearest Neighbors method, generally referred to as the algorithm used for supervised machine learning (KNN), and its user-friendly characteristics. The (KNN) technique is a well-liked and flexible predictive technology due to its simplicity and ease of use. There is no need to make any assumptions about how the underlying data will be distributed. Since it can handle both categorical and numerical data, it is a flexible alternative for a variety of statistic types in the fields of regression, classification, and various other applications. This unofficial method makes predictions based on the degree of similarity between data points in a given dataset. In contrast to other algorithms, K-NN is less vulnerable to outliers.

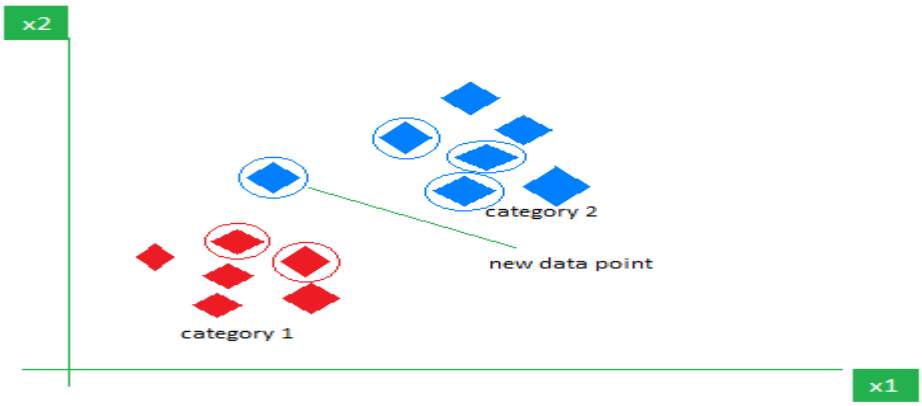



Fig 3.8: KNN Algorithm's Operational Visualization

5. **Logistic Regression**: For single type problems, which occur when results depend on the presence of particular categories of data but remain constant, the controlled forecasting technique of LR is typically used. The logistic regression method's fundamental operations are represented by a sigmoidal function for values between 0 and 1. Logic regression (LR) is essentially a technique for classification. According to one definition, an integer result can only have two potential results: both it will take place (1) or it will not occur at all (0). To put it briefly, distinct variables refer to factors that might affect the research's result but aren't a dependent component itself. Therefore, using a logistic regression is the optimal analytical technique for handling binary data. while dealing with data that determines whether a result or cause is classified into one or more categories. That acknowledges that you are dealing with information that is binary, but only if the data can be classified into either of the two groups.

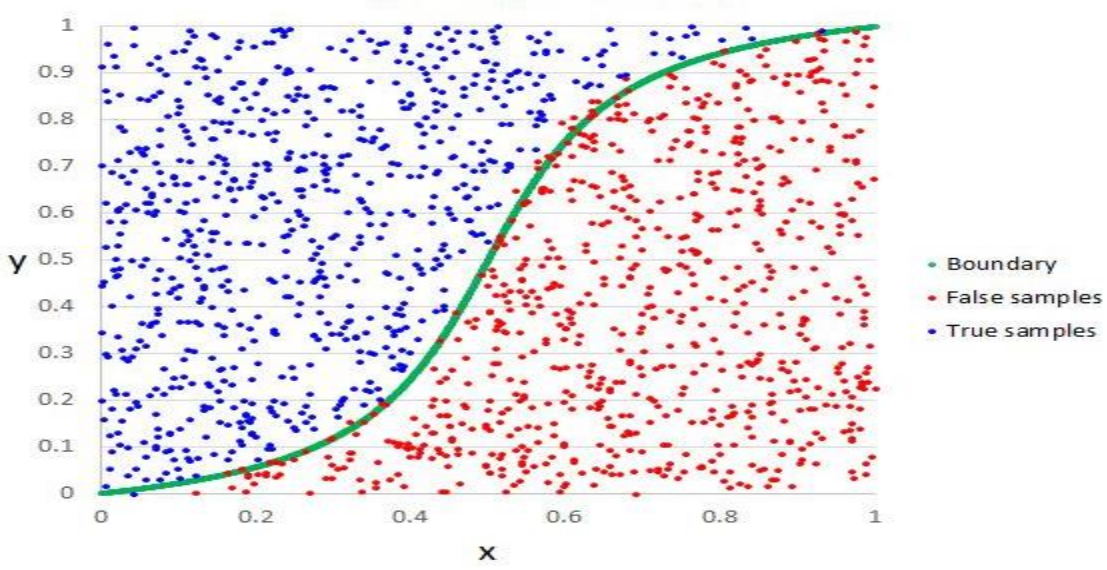



Fig 3.9: A logistic regression illustration

## 3.3 Project Plan

| S.No. | Next Task | Estimate completion time (MM-YY) |
|---|---|---|
| 1 | Data collection | 07-24 |
| 2 | Complete very well of data preprocessing. | 08-24 |
| 3 | Choosing machine learning models deploying. | 09-24 |
| 4 | Reaching the greatest accuracy levels of more than 90%, web application design | 10-24 |
| 5 | Report writing | 11-24 |

### 3.4 Task Allocation

| Tasks | Weeks | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
| Data Collection | | | | | | | | | | | | | | | | | | |
| Data Preprocessing | | | | | | | | | | | | | | | | | | |
| Deploy Models | | | | | | | | | | | | | | | | | | |
| Results | | | | | | | | | | | | | | | | | | |
| Develop web application for prediction | | | | | | | | | | | | | | | | | | |

| | |
|---|---|
| Estimated Work Period | |
| Actual Work Period | |

### 3.5 Summary

To ensure accuracy, details must be gathered after the preceding phases are finished. Ten components were required to finish our project's basic design. If our goal is to be accomplished, each and every one of these duties needs to be completed.

- Data gathering.
- Preprocessing of data.

- Data cleaning of null values.
- Adding 4 columns for Data Preparing.
- Tokenization.
- Models should be used for each of the five approaches.

I had to start coding the idea's code in order for it to function. I evaluated the accuracy of five different methods. We assessed the method's accuracy when it was complete. After evaluating the accuracy, we concluded that the design indicated above would be better appropriate for our requirements. Following a thorough examination of relevant theoretical and numerical approaches and concepts, a set of standards has been developed for any endeavor to categorize the quality of air pollution. Some of the potentially noteworthy results include the following:

# CHAPTER 4

# Implementation and Results

## 4.1 Environment Setup

In order to carry out thorough investigations, the setup used for this study into " **Dhaka City Air Pollution Prediction Using Machine Learning Techniques** " comprises a well-planned arrangement of hardware, software, and data resources. The computational core is made up of strong graphics processing units (GPUs) and a sturdy central processor unit (CPU) in outstanding performance equipment. I am conscious that no invention can lead to flawless outcomes. In a manner comparable to that, we can modify the parameters of our model during training to boost accuracy. But using several approaches, we find that the accuracy level is rather high. This represents a pictorial summary of the work we have been doing. The recalled, precision, f1 score, supports, heat maps, and other data are displayed in these visualizations. Here, we use our data to determine what of the 7 air classes classified by the AQI categories is most often predicted by the AQI measurements.

## 4.2 Performance & comparative analysis

Identifying writing on pollutants in the air which may be classified under a measure of air quality is the aim of this study, which makes use of machine learning methods. In the process of classification, each phrase in any field must be given significant weight. Finding the text for the AQI category has been the main focus of our effort. Data are among the most crucial elements of every study. The available data may cause findings from the same experiment to differ significantly. We knew that other people's conclusions may differ since we collected Dhaka PM2.5 statistics. By applying a number of algorithms based on machine learning to the fidelity ratings and averages, we were able to accomplish our aim. We employed five distinct algorithms in this study. Before we might go to work, they had to find a good deal of stuff. Yes, we did select the algorithm and proceed to work on it. Next, an assessment was made of each approach's accuracy. The RF classifier, which is used as a forecast for the other

five models, obtained the maximum accuracy of 99.81% using techniques. Again, an online application was eventually created to identify the air quality class associated with air pollution AQI readings.

**Precision:** Accuracy is defined as having the ability to distinguish real accomplishments from all expected ones. It is possible to derive the formula below:

$$Precision = \frac{TP}{TP + FP}$$

**Recall:** The term recall describes the model's capacity to discriminate between each true positive response and the expected positive response. The structure of the formula is given as follows:

$$Recall = \frac{TP}{TP + FN}$$

**F1-Score:** By integrating recall and reliability, as the F1 rating does, it is possible to assess the model's effectiveness. The equation you may use to calculate it is as follows:

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

**Accuracy:** The proportion of all observed occurrences in a collection of data that can be anticipated with accuracy and is therefore deemed accurate. calculated as follows:

$$Accuracy = \frac{TP + TN}{Total\ Instances}$$

## 4.3 Results & Discussion

The method we employed determined the disparate outcomes we obtained. We were able to accurately calculate the air pollution category label linked with Dhaka's air pollution thanks to five machine learning algorithms. Following the use of these methodologies to assess the relative efficacy of each component of the overall framework, many verification procedures were conducted to ascertain the outcome of the forecast. After selecting all of the data, each model utilized a single file containing information from publicly available internet sources in addition to information gathered from our own study. Following the completion of the data gathering process, we used Matlab and related pre-configured libraries to assess the algorithms' output. The next step involves utilizing a comparable dataset to determine if the item qualifies for the index of air quality (AQI). They fall into several categories, including Good, Moderate, Unhealthy for Sensitive Groups, Very Unhealthy, Hazardous, and Unhealthy. In this case, we conducted a thorough examination of many models using relevant performance standards. Metrics like as recall that accuracy, precision, and total F1 score offer a thorough picture of the strategies' effectiveness.

Table 4.1: Accuracy.

| Classifier | Accuracy Score (AUC) |
|---|---|
| XG Boost | 99.03% |
| Logistic Regression | 99.74% |
| Random Forest | 99.81% |
| Naïve Bayes | 99.73% |
| KNN | 99.68% |

In this part, several classifiers' output is displayed. PyCharm and CoLab were two complementary tools that were used throughout. Five classifiers in all were used. Naive Bayes, Random Forest, KNN, XG Boost, and Logistic Regression.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Good | 0.94 | 1.00 | 0.97 | 268 |
| Hazardous | 1.00 | 0.95 | 0.98 | 349 |
| Moderate | 1.00 | 1.00 | 1.00 | 2034 |
| Unhealthy | 1.00 | 1.00 | 1.00 | 3247 |
| Unhealthy for Sensitive Groups | 1.00 | 1.00 | 1.00 | 1531 |
| Very Unhealthy | 1.00 | 1.00 | 1.00 | 1403 |
| accuracy | | | 1.00 | 8832 |
| macro avg | 0.99 | 0.99 | 0.99 | 8832 |
| weighted avg | 1.00 | 1.00 | 1.00 | 8832 |

Fig 4.1: RF (Random Forest) classification reports.

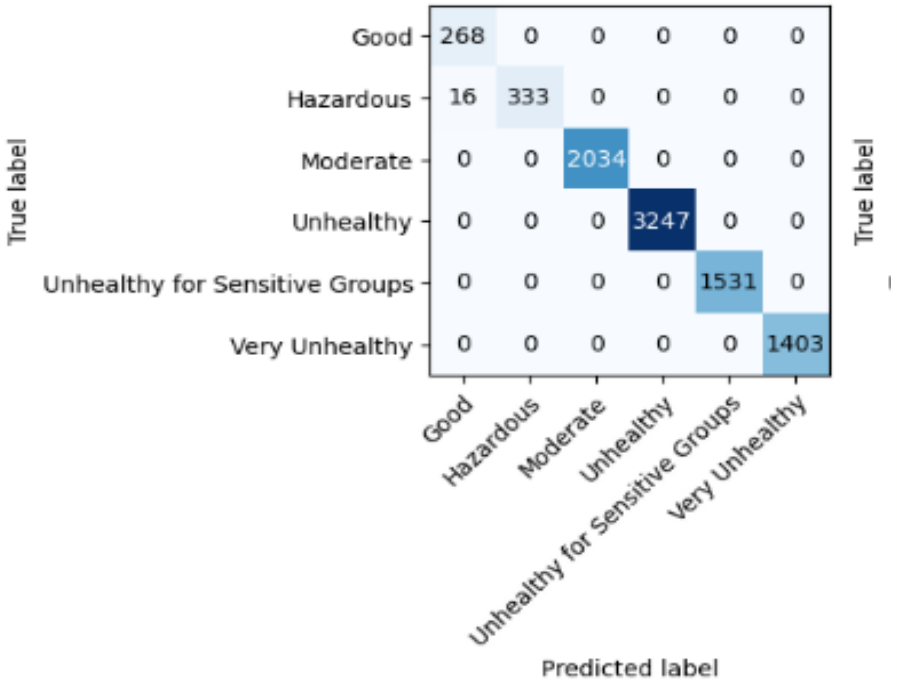



Fig 4.2: Confusion matrix of RF (Random Forest).

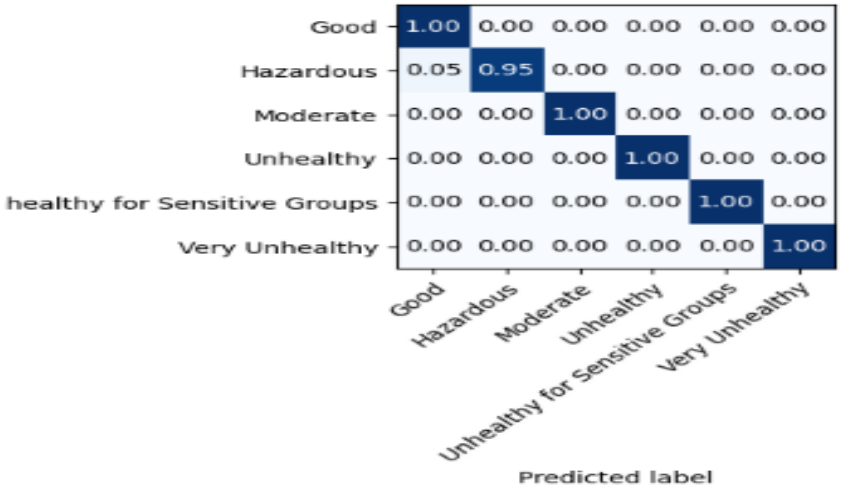



Fig 4.3: Normalized Confusion matrix RF (Random Forest).

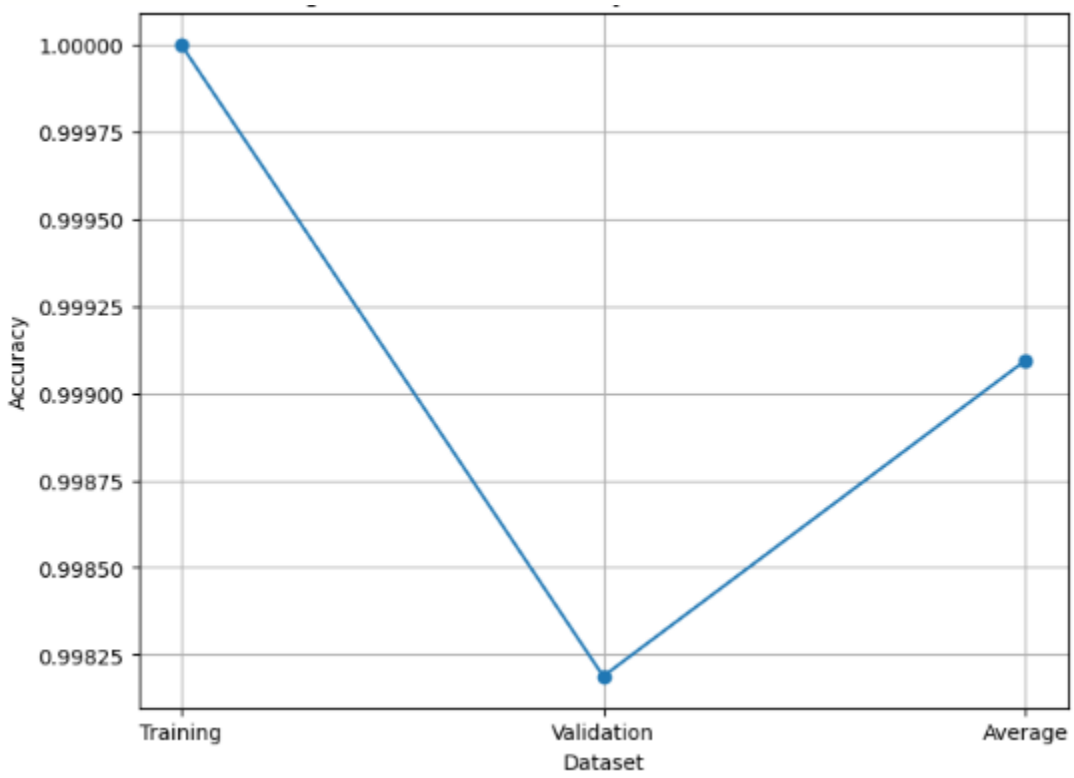



Fig 4.4: Training and Validation Accuracy of RF (Random Forest).

To get optimal accuracy, the classifier automatically uses Random Forests to classify all reports, with just the Classification report being shown.

As seen in Fig. 4.5 below, a working web-based application is created by classifying a particular kind of air pollution quality using the random forest's most accurate model. The six AQI Category categories (Figures 4.6, 4.7, 4.8, and 4.9) that an online application using the random forest "aqi.pkl" model correctly predicted are Hazardous, Unhealthy, Very Unhealthy, Good, Moderate, and Unhealthy for Sensitive Groups. Often, machine learning provides one of the most closely watched methods for AQI detection or forecast values.


Fig 4.5: A prototype web application called Air Quality Detector.


Fig 4.6: AQI category detection for "Good".

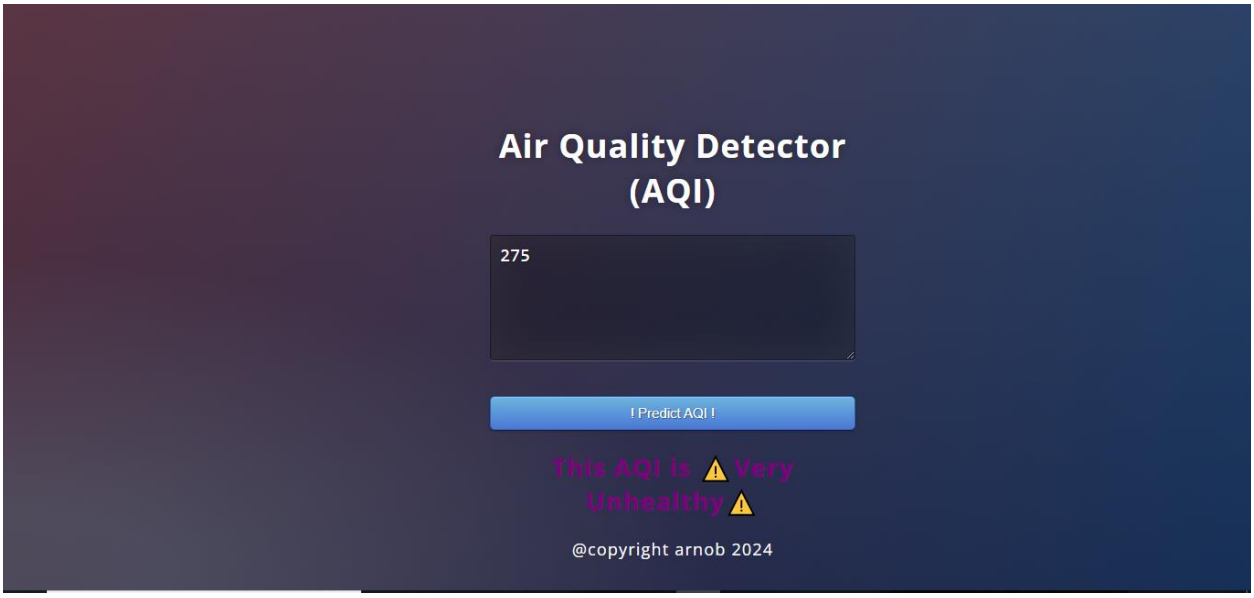



Fig 4.7: AQI category detection for "Very Unhealthy"


Fig 4.8: AQI category detection for "Hazardous".


Fig 4.9: AQI category detection for "Moderate".

As seen in Figs. 4.10 below, the anticipated air pollution quality features in the labeling of the gathered data, along with the ML RF classifiers methodology, explain the process of AQI Categories pollutants identification. The following table shows the two types of erroneous data that learners can detect in addition to the six types of

pollutants in the environment category. One of the best ways to learn unsupervised is to replace the more popular detection or forecasting methods with machine learning-based algorithms. It worked really effectively, especially when compared to older methods.

```
# Preprocess new text data (replace 'new_text_data' with your actual new text data)
new_text_data = ["275",
                 "183",
                 "350",
                 "75",
                 "26"]

new_text_data_processed = [clean_text(text, CONTRACTION_MAPPING) for text in new_text_data]
new_text_data_sequences = tokenizer.texts_to_sequences(new_text_data_processed)
new_text_data_padded = pad_sequences(new_text_data_sequences, maxlen=X_SEQ_LEN)

# Use the trained Random Forest model for predictions
new_text_predictions = text_clf_rf.predict(tokenizer.sequences_to_texts_generator(new_text_data_padded))

# Decode the predicted labels to class names
predicted_class_names = encoder.inverse_transform(new_text_predictions)

# Check the predicted class for each text
for text, predicted_class in zip(new_text_data, predicted_class_names):
    print(f"AQI Value: '{text}' so Air is '{predicted_class}'.")
```

```
AQI Value: '275' so Air is 'Very Unhealthy'.
AQI Value: '183' so Air is 'Unhealthy'.
AQI Value: '350' so Air is 'Hazardous'.
AQI Value: '75' so Air is 'Moderate'.
AQI Value: '26' so Air is 'Good'.
```

Fig 4.10: Predictor of Air Quality.

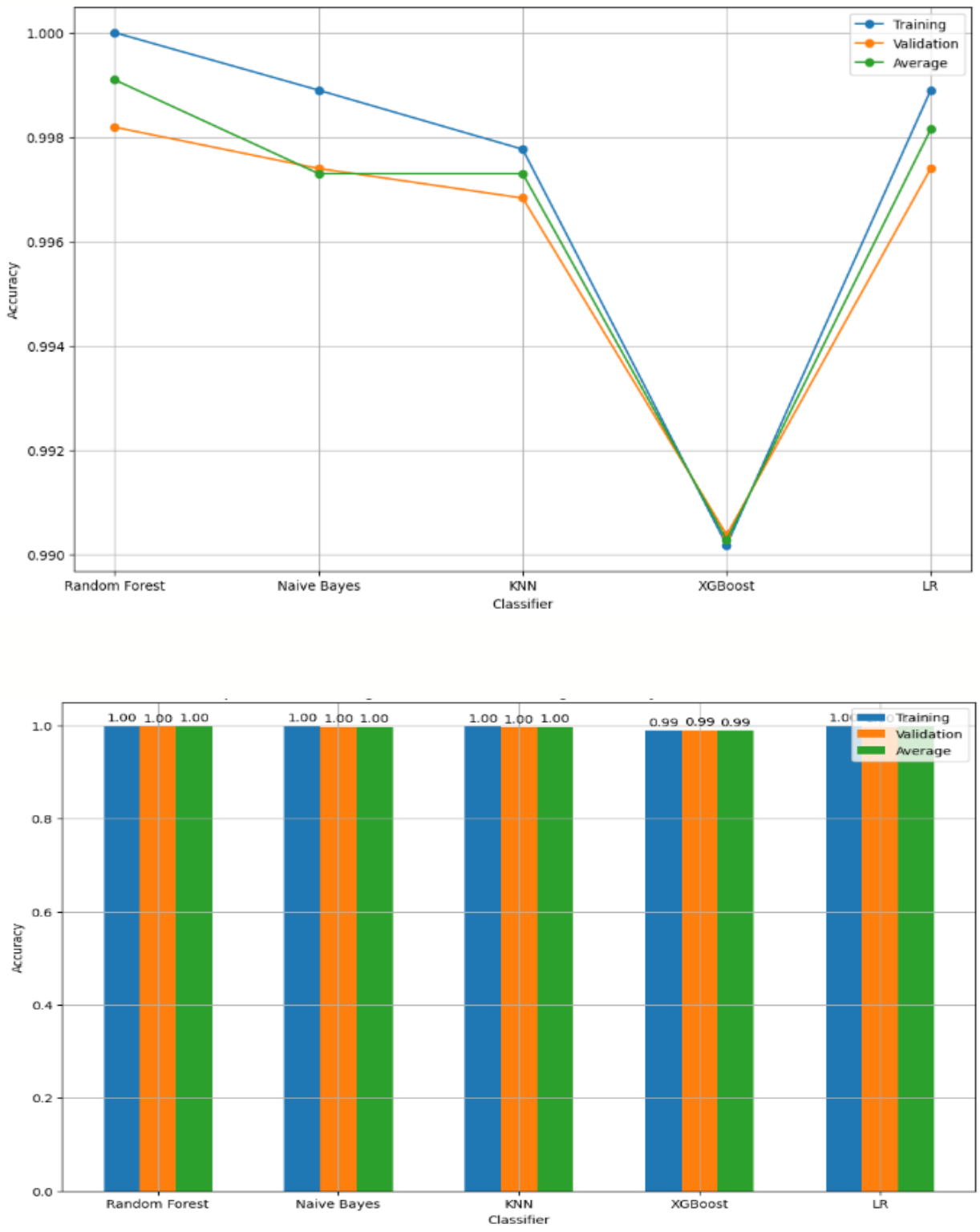



Fig 4.11: Training, validation, and average accuracy Comparison of five classifiers

## 4.4 Summary

In order to measure the quality of air pollution, an AQI Category is assigned. This process is covered in this section. Creating a model merely required selecting a model, collecting and analyzing data, removing null values, adding additional columns to purify the text, and assessing efficacy based on air pollution level detection results. This section of the article presents an analysis and demonstration of the outcomes of the experiment.

# CHAPTER 5

# Engineering standards and Design Challenges

## 5.1 Compliance with the standards

Following legal frameworks and industry best practices that guarantee precise, trustworthy, and moral forecasts is part of complying with requirements for machine learning (ML)-based air pollution prediction. This includes abiding with laws pertaining to the environment and data protection, such as the Air Quality Standards established by national and international organizations like the WHO or the EPA. To adhere to ethical and scientific standards, ML models must also guarantee data quality, open communication, and repeatability. Real-world data should be used to evaluate models and make sure they satisfy the accuracy requirements for making decisions. Important elements of ethical compliance also include preventing prejudice, embracing fairness, and guaranteeing the models' interpretability. Last but not least, incorporating forecasts into policy frameworks and real-time monitoring systems ensures prompt reactions to air pollution incidents while also helping to satisfy public health and safety requirements.

## 5.2 Impact on Society, Environment and Sustainability

### 5.2.1 Impact on Life

Applying machine learning to forecast PM2.5 levels of pollutants in Dhaka might have very significant societal ramifications. Particulate matter that is two millimeters or smaller in length is referred to as PM2.5. Due to their extensive penetration into lung tissue, these particles may have detrimental impacts on health. The following are a few potential impacts on society:

- **Promoting Public Health:** When high levels of PM2.5 are anticipated in advance, governments and communities can reduce exposure by adopting proactive steps to lower levels, especially for vulnerable groups such as kids, seniors, and those with respiratory issues.

- **Environmental Awareness:** By providing timely and accurate forecasts of the environmental impacts of various activities, one may encourage individuals and organizations to adopt sustainable practices.

- **Developing Policies and Urban Development:** Governments may make decisions about regulations and urban development with the assistance of PM2.5 forecasts. For example, the information may have an impact on the locations of schools, hospitals, and residential areas in order to limit access to elevated pollution levels.

While using machine learning to predict air quality can result in positive changes, it is important to consider data privacy, ethical issues, and equitable data access to ensure that the benefits are shared by all segments of society.

### 5.2.2 Impact on society & Environment

In addition to its social benefits, the application of machine learning (ML) to predict Dhaka's PM2.5 air quality might have negative environmental effects. Some possible consequences of machine learning (ML) quality of air forecast on the environment include the following:

- **Pollution mitigation techniques:** pollution from PM2.5 trends and sources may be identified using machine learning algorithms. Implementing targeted environmental protection and control strategies will be made easier with the usage of this knowledge. For instance, if specific activities are frequently connected to high pollution levels, rules might be enacted to decrease emissions from those sources, such as automobiles or specific businesses.

- **Adherence to environmental rules and mitigation of emissions:** The projections can be used to monitor and enforce compliance with environmental

regulations. Businesses and automobiles may be urged to adopt more environmentally friendly technologies and practices to reduce the amount of air pollution they contribute to, as high-pollution periods may be anticipated and managed.

- **Management of Changes in the Climate:** One form of pollution of the air that might worsen global warming is particulates. Machine learning-based forecasts of air quality can help reduce hazardous pollutants and enhance efforts to ameliorate the impacts of global warming when paired with focused preventative strategies.

While there are potential environmental benefits to machine learning (ML)-based pollutant predictions, it is crucial to ensure that these algorithms are incorporated within a comprehensive environmental strategy that includes sustainable development, legislation, and monitoring. Additionally, while choosing green solutions, ethical considerations must be made to prevent biases or unintended negative impacts.

### 5.2.3 Ethical Aspects

It is important to carefully analyze the ethical implications that arise when using machine learning (ML) to predict PM2.5 pollution levels in Dhaka. Important moral considerations include the following:

- **Data protection:** When collecting and utilizing data for air quality forecasts, strict confidentiality guidelines must be adhered to. Make sure that, in order to prevent identifying individuals, any private information is pooled and anonymized. It is crucial to obtain appropriate consent and to make sure that data collection procedures are transparent.

- **Equity and Fairness:** Recognize that machine learning models may contain inaccurate training data. Check to make sure the actions and projections are equal and do not adversely affect any one population. Consider the economic and social implications of air quality legislation and regulations to avoid exacerbating already-existing inequities.

- **Definability and Openness:** Machine learning algorithms need to be understandable and explicit in order to predict air quality. The public, decision-makers, and users should all be able to comprehend forecasts, as well as the factors that affect them. That transparency encourages ownership and faith in the machinery.

- **International Collaboration:** Recognize the interconnectedness of environmental concerns and engage in collaboration with regional and international institutions. In order to address global air quality challenges collectively, share knowledge, suggestions, and ideas.

From data collection and model training to installation and ongoing monitoring, air quality forecasting technologies should take ethical considerations into account at every stage of development. Securing a balance between ethical principles and technological advancements is necessary for the ethical and sustainable application of machine learning (ML) to environmental executives.

### 5.2.4 Sustainability Plan

Using machine learning to predict Dhaka's PM2.5 air quality necessitates developing a framework that addresses ethical issues, sustainability, and long-term efficacy. The following is a suggested ecological tactic:

- **Continuous Data Monitoring:** Establish a dependable system for recording air quality data on a regular basis. The machine learning model should be updated and improved often according to new data and emerging trends. This ensures that projections remain accurate and relevant throughout time.

- **Free Access to Data:** To encourage transparency and collaboration, make machine learning models and data on air quality available to the general public. This encourages communication between the general public, policymakers, and researchers using the data, which fosters accountability and cooperative problem-solving.

- **Research and Creation:** Constant expenditures in R&D are necessary to be at the forefront of ecology and machine learning. Examine innovations that improve the accuracy and efficacy of air quality predictions in order to support the continuous refinement of the sustainability plan.

Through the integration of these elements into a comprehensive sustainable strategy, the application of machine learning for PM2.5 pollution prediction in Dhaka has the potential to eventually advance social progress, environmental health, and fair growth. When the plan is regularly reviewed and modified, it will continue to be effective in addressing emerging problems.

### 5.3 Project Management and Financial Analysis

Good project management and financial supervision are essential to any endeavor's success and longevity, and they are especially important when it comes to the proposed study "**Dhaka City Air Pollution Prediction Using Machine Learning Techniques**" Management of projects includes all of the careful preparation, arranging, and carrying out of work, from gathering data to training and assessing models. A comprehensive schedule for the project will be developed to delineate significant checkpoints, distribute resources sensibly, and set deadlines, so promoting

an organized and effective workflow. Simultaneously, sound financial management—which includes budgetary allotments for hardware, computing resources, and possible research partnerships—will be critical to the project's sustainability. The implementation of an open and prudent financial plan would guarantee the best possible use of resources while upholding fiscal responsibility. These coordinated methods to project finance and leadership are crucial building blocks that will enable the researchers to successfully navigate the study's complexities, guarantee that the resources allotted are in line with the goals of the research, and enable the project to move forward smoothly toward its objectives.

Table 5.1: Estimated Cost for Air pollution prediciton

| SN | Components | Estimated Cost (BDT) |
|---|---|---|
| 01. | Visiting Stakeholders | 500-1000 |
| 02. | Software and Tools | 1500-2000 |
| 03. | Data Collection and Processing | 500-1000 |
| 04. | Documentation and Report Writing | 500-1000 |
| 05. | Contingency (10% of total) | 1500-2000 |
| **Total Estimated Cost** | | 4,500-7,500 |

## 5.4 Complex Engineering Problem

### 5.4.1 Complex Problem Solving

Create a category-based mapping to solve this area's problems. Provide subsections for each mapping to support your claims (see Table 5.1).

Table 5.2: Mapping with complex problem solving.

| EP1 Dept of Knowledge | EP2 Range of Con-flicting Requirements | EP3 Depth of Analysis | EP4 Familiarity of Issues | EP5 Extent of Applicable Codes | EP6 Extent of Sta ke-holder Involvement | EP7 Interdependence |
|---|---|---|---|---|---|---|
| √ | √ | √ | √ | √ | | √ |

**Mapping with Knowledge Profile for EP1**

This table 5.2) is designed to map the EP1 to the Knowledge Profile.

Table 5.3: Mapping with knowledge Profile.

| K3 Engineering Fundamentals | K4 Specialist Knowledge | K5 Engineering Design | K6 Engineering Practice | K8 Research Literature |
|---|---|---|---|---|
| √ | √ | √ | √ | √ |

### 5.4.2 Engineering Activities

Provide an engineering activity mapping in this section. To provide justification, provide subsections for every mapping (see Table 5.3).

Table 5.4: Mapping add subsections to put rationale.

| Engineering Activity | Rationale |
|---|---|
| 1. Data Collection | After gathering the data from internet resources, such as Kaggle, we evaluated it. There isn't a large, comprehensive dataset accessible in this industry since it's difficult to gather information for the specific air pollution content of high-quality analyzer and categorization type. |

| | |
|---|---|
| 2. Data Preprocessing | Each piece of data was examined separately once all practical means of data collection had been used. There are many instances of incorrect and imprecise language all around us. Before utilizing the chosen dataset, we are advised to go over its final part. |
| 3. Model Development | We choose a prediction approach, train it using my data, and then assess it to increase reliability. Filters are heavily used in machine learning. Several designs were used to enhance the materials design and enable the ML model to determine the type of air pollution, but ultimately, only one device was selected to evaluate the accuracy of the data. |
| 4. Model Evaluation and Testing | This phase's later sections address all the consequences. After training and testing, these approaches gave us a limited level of reliability for each of the two distinct air quality datasets label groups. To bolster the confusion matrices, f1 ratings and accuracy data were generated. Applying machine learning algorithmic learning techniques, ascertain if all kinds of contaminants in the natural environment can be represented by AQI values. |
| 5. Deployment and Integration | Predictions can be made in real time by integrating the model into an intuitive platform, such as a smartphone app or website. Protecting sensitive neural data requires security methods like access restriction and encryption. |
| 6. Continuous Monitoring and Maintenance | Continuous observation guarantees the model's accuracy over time. Maintaining the model's relevance and conformity to changing standards is facilitated by frequent updates based on fresh data and feedback regarding performance. |

## 5.5 Summary

A verified sample that shows the report's creation complies with all criteria is included in this section. Effects on the environment, society overall, and the Sustainable Development Goal The limitations on our research, which might affect future generations of experts in our field, are highlighted at the end of the chapter.

Table 5.5: Mapping with complex engineering activities.

| EA1<br>Range of re-sources | EA2<br>Level of Interaction | EA3<br>Innovation | EA4<br>Consequences for society and environment | EA5<br>Familiarity |
|---|---|---|---|---|
| √ | √ | √ | √ | √ |

# CHAPTER 6

# Conclusion

## 6.1 Summary

We now know a great deal more about the problem thanks to this inquiry. Air quality forecasts remain a contentious subject. This has a major impact on the yearly drop in patient usage of air pollution monitoring programs that employ the air quality indices. The seven fundamental categories of air danger that we were able to identify using machine learning are: hazardous, unhealthy, extremely unhealthy, good, moderate, and unhealthy for sensitive populations. Additionally, based just on looks, the ML approach has been utilized to evaluate the chance of conveying AQI values for contaminated air.

As we've already stated, the research's objective is to understand as much as possible about this subject. This was accomplished by combining data available online from Kaggle of Dhaka 2017–2022 PM2.5 with data from an online application that included six classifications: hazardous, unhealthy, severely unhealthy, excellent, moderate, and unhealthy for sensitive populations. With the help of the labeling that was extracted from the data datasets, we were able to analyze and train our five techniques thanks to the extraordinarily accurate identification of the precise "AQI" number values of the "AQI Category" by the Random Forest classifiers. Determining how much a message differs from another in regards to air pollution severity is made simpler by the model predicting method. An air quality index (AQI) is a means to report air quality on a daily basis. The short-term health impacts of air pollution are measured by this indicator. The purpose of the AQI is to help people understand how the air in their area affects their health. In Dhaka, environmental degradation is used as it is. It is critical to increase public awareness of the problem and move quickly to eliminate air pollution. These include promoting appropriate online conduct, urging people to report any incidents of air pollution, as well as aiding those who have been harmed by the environment in Dhaka. To lower Dhaka PM.25 2017–

2022, establish safe zones, and stop air pollution, according to the AQI values for each category. Together, we can reduce the negative impact of air pollution on our environment and create a polite and welcoming environment in Dhaka. Using supervised learning algorithms is one method; these algorithms are trained on a dataset that includes six books from the Air Pollution Quality Index (AQI) Category. An computerized air quality monitoring system's performance might be influenced by several things. The kind of algorithms used, their settings, and the amount and caliber of the initial training data comprise these. This document compiles our algorithm's problems for identifying air pollution. Five classifiers are provided to assist in identifying the AQI category that can be considered air pollution; a highly accurate prediction can be chosen. Using the ambient quality index as a guide, the top five classifiers identify air pollution with an accuracy value of 99.81%. Our model performs best when RF is applied. A web-based application that determines whether or not AQI values indicate air pollution is another method that makes use of the RF algorithm.

## 6.2 Limitation

We argue that our method's outcomes are affected by changing the observation place. As a result, we worked harder to do our duty accurately. We make sure that the content we provide is legitimate by adhering to our security requirements. As such, we decide to use an aggregated online dataset.ML models may be able to identify language and symptoms associated with an AQI category earlier than traditional diagnostic methods. Early diagnosis combined with prompt intervention to air pollution may enhance outcomes overall. Since machine learning techniques are widely used in a variety of settings, including websites and mobile applications, it makes perfect sense to stop air pollution from Dhaka. If we had appeared more personable, we might have helped those who might not have asked for help.

## 6.3 Future Works

Future work estimating PM2.5 air quality for Dhaka by Machine Learning (ML) may involve several advancements and enhancements. more research on the ethical concerns related to air quality predictions. To guarantee equitable distribution of anticipated benefits, develop strategies to mitigate biases in data collection and model results. Consider the impact of climate change on air quality patterns. Make models that account for changing weather patterns and assess how climate-related variables could affect PM2.5 concentrations in the future. Ultimately, further study and innovation in these areas may lead to healthier and more salubrious urban environments, therefore augmenting and amplifying the effectiveness of PM2.5 contaminants in Dhaka's air quality forecasts using machine learning.

# Reference


[1] H. Liu, Q. Li, D. Yu, and Y. Gu, "Air quality index and air pollutant concentration prediction based on machine learning algorithms," Applied Sciences, vol. 9, p. 4069, 2019.

[2] M. Castelli, F. M. Clemente, A. Popovic, S. Silva, and L. Vanneschi, "A machine learning approach to predict air quality in California," Complexity, vol. 2020, Article ID 8049504, 23 pages, 2020.

[3] G. Mani, J.K. Viswanadhapalli and A.A. Stonie, "Prediction and forecasting of air quality index in Chennai using regression and ARIMA time series models," Journal of Engineering Research, vol. 9, 2021.

[4] S. V. Kottur and S. S. Mantha, "An integrated model using Artificial NeuralNetwork (ANN) and Kriging for forecasting air pollutants using meteorological data," Int. J. Adv. Res. Comput. Commun. Eng, vol. 4, pp. 146–152, 2015.

[5] S. Halsana, "Air quality prediction model using supervised machine learning algorithms," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 8, pp. 190–201, 2020.

[6] A. G. Soundari, J. Gnana, and A. C. Akshaya, "Indian air quality prediction and analysis using machine learning," International Journal of Applied Engineering Research, vol.14, p. 11, 2019.

[7] C.R. Aditya, C.R. Deshmukh, N.DK, P. Gandhi and V. astu, "Detection and prediction of air pollution using machine learning models," International Journal of Engineering Trends and Technology, vol. 59, no. 4, pp. 204–207, 2018.

[8] J. Kleine Deters, R. Zalakeviciute, M. Gonzalez, and Y. Rybarczyk, "Modeling PM2. 5 urban pollution using machine learning and selected meteorological parameters,"

Journal of Electrical and Computer Engineering, vol. 2017, Article ID 5106045, 14 pages, 2017.

[9] P. Bhalgat, S. Pitale, and S. Bhoite, "Air quality prediction using machine learning algorithms," International Journal of Computer Applications Technology and Research, vol. 8, pp. 367–370, 2019

[10] M. Bansal, "Air quality index prediction of Delhi using LSTM," Int. J. Emerg. Trends Technol. Comput. Sci, vol. 8, pp. 59–68, 2019.

[11] A. Shishegaran, M. Saeedi, A. Kumar, and H. Ghiasinejad, "Prediction of air quality in Tehran by developing the nonlinear ensemble model," Journal of Cleaner Production, vol. 259, Article ID 120825, 2020.

[12] L. Tuan-Vinh, "Improving the awareness of sustainable smart cities by analyzing lifelog images and IoTair pollution data," in Proceedings of the 2021 IEEE International Conference on Big Data (Big Data), IEEE, Orlando, FL, USA, September 2021.

[13] R. Kumar, P. Kumar, and Y. Kumar, "Time series data prediction using IoT and machine learning technique," Procedia Computer Science, vol.167, no. 2020, pp. 373–381, 2020.

[14] H. Maleki, A. Sorooshian, G. Goudarzi, Z. Baboli, Y. Tahmasebi Birgani, and M. Rahmati, "Air pollution prediction by using an artificial neural network model," Clean Technologies and Environmental Policy, vol. 21, no. 6, pp. 1341–1352, 2019.

[15] K. P. Singh, S. Gupta, and P. Rai, "Identifying pollution sources and predicting urban air quality using ensemble learning methods," Atmospheric Environment, vol. 80, pp. 426–437, 2013.

DHAKA_CITY_AIR_POLLUTION_PREDICTION_USING_MACHI...
2

ORIGINALITY REPORT

| 20% | 11% | 6% | 12% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | Submitted to BRAC University<br>Student Paper | 4% |
| 2 | Submitted to Daffodil International University<br>Student Paper | 4% |
| 3 | dspace.daffodilvarsity.edu.bd:8080<br>Internet Source | 4% |
| 4 | www.hindawi.com<br>Internet Source | 1% |
| 5 | Gokulan Ravindiran, Gasim Hayder, Karthick Kanagarathinam, Avinash Alagumalai, Christian Sonne. "Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam", Chemosphere, 2023<br>Publication | 1% |
| 6 | Submitted to United International University<br>Student Paper | <1% |
| 7 | Submitted to University of Finance – Marketing | <1% |