# DHAKA AIR POLLUTION PREDICTION USING MACHINE LEARNING APPROACHES

**BY**

**KAZI AFSANA AKTER**

**ID: 213-15-14791**

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

Supervised By

**ABDUS SATTAR**
Assistant Professor & Coordinator MSc.
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised By

**RAJA TARIQUL HASAN TUSHER**
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University


**DAFFODIL INTERNATIONAL UNIVERSITY**

**DHAKA, BANGLADESH**

**JULY 2024**

## APPROVAL

The research paper titled " **Dhaka Air Pollution Prediction Using Machine Learning Approaches,**" which was turned in to the Department of Computer Science and Engineering at Daffodil International University by **Kazi Afsana Akter**, has been accepted as satisfactory for partially fulfilling the requirements for the B.Sc. in Computer Science and Engineering degree. Its style and contents have also been approved. On July 16, 2024, this presentation was given.

## BOARD OF EXAMINERS

**Narayan Ranjan Chakraborty (NRC)** — **Chairman**
**Associate Professor & Associate Head**
Chairman,Department of CSE
Faculty of Science & Information Technology
Daffodil International University

**Naznin Sultana (NS)** — **Internal Examiner 1**
**Associate Professor**
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

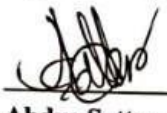
**Raja Tariqul Hasan Tusher (THT)** — **Internal Examiner 2**
**Assistant Professor**
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

**Dr. Ahmed Wasif Reza (DWR)** — **External Examiner**
**Professor**
Department of Computer Science and Engineering
East West University

## DECLARATION

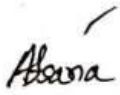
We so certify that we, **Abdus Sattar, Assistant Professor, Department of Computer Science and Engineering,** Faculty of Science and Information Technology, Daffodil International University, have supervised this study. We further declare that no portion of my study, nor any portion of it, has been submitted for consideration for a degree elsewhere.

**Supervised by:**



**Abdus Sattar**
**Assistant Professor**
Department of CSE
Daffodil International University

**Submitted by:**



**Kazi Afsana Akter**
ID: 213-15-14791
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 **Abdus Sattar,** Assistant Professor 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 Air Pollution Prediction Using Machine Learning Approaches** " 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 **Abdus Sattar,** Assistant Professor and **Dr.Sheak Rashed Haider Noori**, 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

Air pollution is crucial for the health of both humans and animals because it has been connected to a number of deadly illnesses, including cancer. Nonetheless, a lot of industries, ships, farms, and housing contribute to air pollution due to the world's swift urbanization and population expansion. As a result, air pollution has become a major problem in many cities, especially in developing countries like Bangladesh. Maintaining indoor air quality requires regular forecasting and monitoring of air pollution. As a result, machine learning (ML) has demonstrated potential in surpassing conventional methods in the prediction of the air quality index (AQI). An indicator of the condition of the atmosphere is the air quality index, or AQI. It estimates the short-term effects of modest exposure on an individual's health. The public is to be made aware of the harmful effects that ambient pollution has on health through the use of the AQI. The quantity of pollutants in the air has significantly increased in Indian cities. Using Dhaka's (the capital city of Bangladesh) AQI, we focus on a few parameters starting with PM2.5 in 2017 and going all the way through to 2022. The objective of the research is to ascertain how well NLP methods recognize and classify activity inside AQI categories. An algorithm is trained via managed instruction to classify AQI information using labeled data and forecast outcomes with accuracy. Amongst the models that machine learning applied to achieve this goal were XG Boost, a Random Forest, K-Nearest Neighbors, Naive Bayes, and Linear Regression. With a 99.81% classification accuracy, the Random Forest classifier was shown to be the most accurate after data analysis, correctly classifying AQI values into six different categories: hazardous, unhealthy, very unhealthy, good, moderate, and unhealthy for delicate groups. Finally, a web prototype is generated by classifying the AQI subcategory using the method known as Random Forest.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# CHAPTER 1

# Introduction

## 1.1 Overview

The only ingredient required for human life is air. We need to assess and understand its quality for our personal well-being. Airborne pollution causes many people to have respiratory ailments and other health problems all around the world. Official studies show that pollution in the air constitutes by much the most significant environmental hazard. Population growth has been accelerated by the toxic gasses produced by rapid development. Air pollution poses a serious threat to human health and is exceedingly detrimental. The air quality has significantly deteriorated as a result of unchecked pollution. An alternative term used to describe the Air Quality Index, or AQI in simple terms, is a pollution measurement and reporting instrument. The immaterial air purity index (AQI) is made up of twelve components, or airborne pollutants.

The amount of PM10 (large fragments with a circumference of ten a millimeter or less), PM2.5 (large fragments with an average length of two millimeters in or less), ammonium hydroxide (ammonia), oxides of nitrogen (NO2), sulfur dioxide (SO2), the gases carbon monoxide (CO), the oxygen layer (O3), and an amalgamation thereof known as at The six contaminants—PM10, PM2.5, sulfur dioxide (SO2) carbon monoxide, CO, and O3—are used to calculate the air quality indicator (AQI), that is utilized in several applications. However, the precise objective must be taken into consideration along with other elements like data accessibility, measurement methods, and monitoring frequency when choosing the contaminants. Extremely polluted air that could be harmful to health is indicated by a high AQI.

You are able to monitor the present state of the air quality thanks to the AQI. Numerous weather stations in our neighborhood monitor the AQI on a daily or weekly basis. This data will be used for the proposed project, which is the reason behind the data collection and gathering process. The primary driver of global warming is emissions of greenhouse gases (GHGs). They have negative consequences on farming, the environment, and human activities. They also alter plant-soil interactions and warmth (Malhi et al., 2021). In 2022, the World Medical Organization (WHO) published research on the quality of the air worldwide, utilizing information from 2010, to the spring of 2019.

As previously indicated, this study examined a range of air pollutants based on analyses of 6743 sites across 117 countries. It was found that PM2.5 was growing more commonplace globally and was responsible for around two million fatalities per year in India alone.
An elevated AQI score designates an area where deaths are more likely to occur. Because of this, AQI monitoring and prediction have emerged as essential tools for ensuring long-term prosperity worldwide (Rybarczyk and Zalakeviciute, 2021). Numerous academics have created techniques that use computational, anticipated, intellectual ml, and dl ideas to forecast the AQI. Existing theories and analytics are insufficiently adaptable in complex scenarios.

Thanks to recent developments in sensors, which allow for the quick identification of varying levels of air pollution, the AQI is determined instantly. The technique of AQI prediction is made easier by using the readily available data sets (Bekkar et al., 2021). A machine learning system determines the AQI extremely accurately and reliably in every situation. We can now produce AQI estimates which are more reliable than ever thanks to machine learning (ML). The growing amount of historical data that is accessible for analysis made machine learning (ML) possible. They are multiplying in an attempt to demonstrate that they can serve as a competitive alternative to widely recognized statistical methodologies in time series projection.

The lack of knowledge about the dynamics of the highly unpredictable systems regulating the amount of pollution makes it difficult to develop statistical models to anticipate pollution levels. A machine learning model is an ideal illustration of a homogeneous and asymmetric strategy, since it just requires historical data to create the relationship between the indistinguishable variables. Consequently, we may create a more accurate predicting model.

### 1.2 Problem Statement

Gathering and analyzing all of these little bits of data seems to be the major issue with this study, given how difficult it was to wade through a single enormous data file. We cleaned and standardized the data collecting process using a number of tools and techniques. Because of the large number of papers and the variety of qualities they encompassed from different historical periods, it required some time to arrive at the appropriate conclusions. There are more data sets in this field. We never complete the study, therefore always have to put in a lot of effort on any connected tasks, which makes it difficult for me to come up with the best answers quickly. The development of natural language processing (NLP) methods for automatically categorizing

instances of the AQI category has garnered significant attention because to the possible advantages of identifying and mitigating such behavior.

### 1.3 Objectives

Machine learning (ML) has shown to be an extremely successful tool for problem solving. Using supervised learning, an artificial intelligence technique, relevant data may be grouped into categories such as "Hazardous," "Unhealthy," "Very Unhealthy," "Good," "Moderate," and "Unhealthy for Sensitive Groups." To find out if an AQI classification is equivalent to each of the six classes, first train efficient models on an analysis set (which needs labeled data). The primary objective of directed automated learning is this. AQI Categories might be ascertained by AI techniques. Consequently, we were able to ascertain the following goals:

- Using machine learning to predict each of the six AQI subcategory classifications.
- To collect data with the goal to forecast the AQI.
- Acquiring a solid understanding of the machine learning-related fields.
- Applying a range of strategies to enhance outcomes.
- Dhaka's predicted PM2.5 AQI classification.

### 1.4 Motivation

My research uses artificial intelligence and machine learning to classify different types of air pollution according to AQI readings. After some consideration, we were reluctant regarding coming up with an investigation design that would meet the requirements of the study. So, we reached out to some of my favorite teachers for assistance. It was recommended that we look into a relevant concept in relation to this subject because the current environment air pollution situation. We can additionally observe that civilization is changing through the prism of fresh perspectives and the manner in which academics are going deeper into these topics in order to better develop my own opinions. Because of this, I made the decision to create a paper titled " **Dhaka Air Pollution Prediction Using Machine Learning Approaches**." These were the driving forces behind this kind of research-based practice. I am surrounded by intelligent equipment, showing why the use of AI is so important.

### 1.5 Thesis Outcome

Finding any words associated with the air quality index (AQI) was the aim of this investigation. The objective is to accurately determine the value of an AQI remark. Six categories—"Hazardous," "Unhealthy," "Very Unhealthy," "Good," "Moderate," and "Unhealthy for Sensitive Groups"—have been established from the diverse and varied remarks that were communicated. Thanks to the size and significance of the dataset utilized in this investigation, we were able to get results with a respectable level of accuracy. A subset of the dataset's most accurate results is extracted from it using machine learning in this study. The comprehensiveness of the training set and the monitoring of the efficiency of the learning method will determine how accurate the model is. The AQI values produced by our data mining techniques contain the AQI category. In order to identify potentially dangerous materials, we can program computers to anticipate outcomes 100% of the time. To produce the best report possible, a number of algorithms' output parameters, including accuracy, recall, error rates, and reliability, were assessed.

### 1.6 Report Organization

The goals, concerns, research questions, and expected outcomes of the study were outlined in Chapter 1. This section also covers the general organization of the report.

All past research in this field is included in Chapter 2. They provide an illustration of the breadth that results from their narrowing of this research topic in the section that follows. The topic of the last discussion was the primary difficulties or barriers to this investigation. This chapter discusses the difficulties encountered across the project's growth and contains sections on pertinent research abstracts and investigations.

The study findings are theoretically evaluated in Chapter 3. More details on the statistical methods are provided in this chapter, particularly those that were used in the numeracy section of the inquiry. Additionally, this section provides real-world applications of machine learning algorithms. The part of the article afterwards discusses the methods for obtaining data and the structure used to compile it. In the last step, which assesses the algorithm utilizing a single-family confusion evaluation matrix, a suitable tag is given for identifying the classifier. Application assessment is required when using machine learning algorithms in order to ensure true accuracy. This section covers the research topic and technique, operational efficiency, data collection plan,

data processing, suggested methodology, teaching style, and requirements that must be completed for the study to proceed. This research offers a thorough justification for every machine learning technique and classifier used.

Chapter 4 presents the study's findings, an evaluation of the findings, and a conversation on the implications. A few test pictures are included in this chapter to aid in the project's execution. An review of the results and an application of AI techniques round up this chapter. Explain the web-based protype that calculates an air pollution index (AQI) based on AQI measurements as well.

The last two chapters 5 and 6 had a result, a description of the planned sequence of behavior, and a research summary. A verified sample proving the report's structure complies with all standards is provided in the next section. Effects on the environment, society in general, and the objectives of sustainable development the limitations on our job are highlighted in the chapter's conclusion, and these may have an effect on next generations of specialists in our field.

### 1.7 Summary

This investigation has taught us a lot more about the issue. Forecasts for air quality are still up for debate. This leads to the annual decline in customer use of pollution surveillance systems that use the air circumstances parameters. By employing machine learning, we managed to categorize air threat into seven basic groups: hazardous, unhealthy, severely unhealthy, excellent, moderate, and unhealthy for people with specific needs. The air quality index, or AQI, measures for polluted air have also been evaluated for transmission using the machine learning technique, based solely on features.

# CHAPTER 2

# Background

## 2.1 Overview

This option includes the investigation's findings, challenges faced, relevant literature, or a summary of previous studies. In the "related Works" section, I will review research done by several writers and discuss how their methods relate to conceptual correctness. This linked works section will provide an overview of the papers, methods, and reliability of other scholarly works that are relevant to my study. The research's introductory unit will provide a summary of my relevant work. We explain how we overcome any obstacles that arose during the study and how we increased the accuracy of each layer during the demanding period. All of this has previously been discussed with me. The vast discipline of dataset analysis uses calculations, logical processes, cycles, and frameworks to extract ideas and understanding from both organized and unorganized info, and then it exploits those important details and revelations to a range of uses. PM2.5 and PM10 are harmful pollutants that can cause lung cell death, ischemic heart disease, severe asthma of the lungs, neonatal pneumonia, and chronically respiratory obstruction disorders (COPD). Particulate matter that exists in the atmosphere has been linked to strokes, which happen when the blood supply to the brain is cut off.

## 2.2 Related Works

First, a number of air indicators were analyzed for correlation, such as the Air Quality Index (AQI), PM2.5 particles amounts, total NOx (nitrogen oxides) levels, and more [1]. Following the construction of forecasting models utilizing random forest regression and support vector estimation, the effectiveness of the two systems was assessed using r, and Rc-squared (R-Mean), and further statistical approaches. Using a machine learning approach known as SVR, it is possible to determine the state of the atmospheric gauge and assess the amounts of contaminants and particles [2]. The findings demonstrate that, for the entire state of California, per hour air quality indexes (AQI) and values of various pollutants, such as nitrogen oxides, carbon monoxide, sulfur oxides sulfuric acid sulfur, besides ozone, and particle matter in fines 2.5, can be accurately calculated by combining SVR with the RBF kernel. Among the six AQI categories that the EPA

submitted at the point of the nation's investigation, ninety-one of the the initial information used for testing were correctly categorized (dataset).

Regression learning and time series analysis are two machine learning techniques used to predict the AQI. To anticipate the AQI, the directed artificial intelligence method known as MLR was utilized. A number of quantitative criteria were used to assess the result. In addition, the AQI was projected into the future using an updated version of the ARIMA series theory. When it came to AQI prediction, both models demonstrated exceptionally high efficacy and accuracy [3].Using a blended approach that combined neural network technology with the Kriging technique, the amount of contaminants in the air at several places in Mumbai as well as Navi Mumbai, which is located in India, was assessed. It was discovered that the higher R values were mostly responsible for the necessary degree of distinction among what happened and expected results. In terms of forecasting and R value, Announcement Networks (ANN) outperformed the standard regression model [4]. using variables like PM2.5, PM10, SO2, or NO2 to estimate an author's AQI intensity. The at- random woodland regression approach eventually resulted in the most accurate results when compared to selecting the tree, the power source SVR, with exposure to radiation regression techniques, with a degree for high accuracy of 0.99985 on the experiment's data, a mean variance erroneous value of 0.00013, and a mean unreserved misunderstanding of 0.00373 general [5]. Minimizing the slope of the multivariate regression issue allowed for the prediction of the AQI using data from the prior year and projecting to a specific future year. They outperformed traditional regression models by optimizing their predictive powers and employing expected costs for the prediction problem. The amount of similarity between all of the possibilities and the best solution was also assessed in order to rank the requests according to importance using both the AHP plus MCDM technique [6]. Using a logistic regression technique [7], it was possible to determine the level of contamination of a specific data set containing daily meteorological conditions and contextual elements in various cities. Using historical PM2.5 data, this approach attempted to anticipate PM2.5 levels and evaluate the quality of the air. Based on the results, automatic or logistic regression analysis may be used to forecast potential amounts of PM2.5 and air quality. This article [8] presents a machine learning algorithm that predicts PM2.5 levels based on six years of meteorological and pollution data, moisture levels, and breezy (velocity & azimuth). The results illustrated the categorization of the cheapest (10g/m3) against enormous (>25g/m3) and the most tiny (10g/m3) against intermediate (10–25g/m3) PM2.5 levels. They also

showed the significant accuracy of the filtering model. An artificial neural network (ANN) and the Kriging technique approaches were used in combination with historical data from the hydrology office and the Air Quality board to predict the total number of airborne allergens in Mumbai and the wider Mumbai region [9]. The recommended framework was then developed and assessed using the R coding language for Criticizing and its MATLAB tools for ANN. Predicting future contamination and evaluating a vast amount of poisoning data were made possible by technology. To locate other information sources on the environmental quality, anticipate, a time-series analysis was used as well. It was looked at if hourly pollution levels in Delhi might be predicted using a deep neural network and long short-term memory (LSTM). Even in line with the weekly estimates, the outcomes were accurate. According to the findings [10] and [11], deep learning-based strategies perform better than conventional statistical methods.

### 2.3 Research summary

Most of our current work is focused on the range of approaches that other communities may provide. We expanded our dataset with additional processes and employed a total of five distinct methods. This example's primary data source is the internet-based Kaggle collection, which is accessible to the public. As mentioned before, our compilation contains both recently collected and previously used data. Using the same source, we will be able to evaluate the impact of the additional data we gave and the efficacy of the five strategies we employed. It implies the existence of many comparable classes and labeling systems. Python was used to build the majority of machine learning techniques and related text classification algorithms that were used in the feature removal and data gathering procedures. Python proved to be the perfect environment for feature extraction techniques when we employed an RF classifier methodology for categorizing web applications.

### 2.4 Challenges

The primary issue with this study seems to be gathering and analyzing all of these little bits of data, given how difficult it was to dig through a single enormous data file. We cleaned and standardized the data collecting process using a number of tools and techniques. Because of the large number of papers and the variety of qualities they encompassed from different historical periods, it required some time to arrive at the appropriate conclusions. There are more data sets in

this field. We never complete the study, consequently we need to put in a lot of effort on any connected tasks, which makes it difficult for me to come up with the best answers quickly.

The development of natural language processing (NLP) methods for automatically categorizing instances within an AQI category has garnered significant attention because to the possible advantages of identifying and mitigating such behavior. However, there are still more difficulties involved with this project. The lack of a general term for air pollution is one of the main problems. One reason why developing a unified strategy might be challenging is because various organizations and researchers have very varied definitions of what constitutes a class. Since they harm and deteriorate their targets, their complex and AQI characteristics can additionally render it challenging for an NLP algorithm to determine air quality.

- Collecting data;
- Importing datasets;
- Preprocessing data.
- Achieving 90% accuracy at the greatest conceivable level.
- Choosing every machine learning model based on the results of the tests.

# CHAPTER 3

# Research Methodology

## 3.1 Overview

The study technique is covered in the section that follows, along with instructions for compiling datasets, running each test, and making use of each model to improve accuracy. Additionally, suggestions for an approach and all-data study were presented in this chapter. Thus, in an attempt to make the data presented in this chapter better and simpler. This research area focuses on the methods utilized to evaluate the quality of air pollution in research, in addition to providing a detailed explanation of the entire procedure. The study's whole methodology will be covered in this part. Numerous approaches can be used to solve any given analysis. The selection of an ML strategy is the next stage. As we have demonstrated, creating a repository for data is a necessary step before creating the framework and running the algorithm because we are using five different machine learning techniques. Next, the collected data is used to train the model. This serves as the foundation for feature selection. Next, sets for training and testing were created using the data. A widespread misinterpretation exists between "training data" and "testing data." Only a sizable portion of the data needed to assess our model is available to us once the input is extracted from the collection of data to be educated & fitted into different ML method models. Next, the reliability of the model is assessed. Many of the strategies will be explained with the use of formulas and examples to help you grasp them better. We proceeded to use our basic process flow diagram to explain things. This is a summary of the whole research endeavor, including the methodology employed. Two among the most important components are the suggested framework and the data collected for the study. With the right formula, explanation, arrangement, and graph, the idea is communicated more effectively.

## 3.2 Requirement Analysis

An area of research is a topic of study that is looked at and studied to provide insights into managing modeling, data gathering, job fulfillment, and showcasing education, in addition to execution. We talk about the instruments and methods we use as measurement experts. We made use of the company's Windows operating system, Python for our programming, and a few other tools including NumPy, Scikit Learn, and OpenCV. For all education and evaluation, Google

Colab was the preferred platform. Python programmers may create code for data science and machine learning methods by using Google Colab. These methods use statistical approaches associated with machine learning to detect clusters, i.e., the 6-an air quality score type categorization.

**Libraries used**:

- **Matplotlib:** Displaying, scoring, organizing, and graphing are made easier with the Pyplot set of Matplotlib tools. This might be applied to form-building to draw attention to certain narrative points of view or the story's boundaries of plausibility.
- **NumPy**: The language's NumPy library has simplified vector processing. This topic includes detailed descriptions of matrix calculations, conversion indices, and the inverted transform of a Fourier transform. The NumPy Python bundles include a number of tools and techniques for working with different kinds of matrices. The process of creating gadgets may be made more rational and useful with the aid of NumPy. In short, NumPy is a Python module primarily intended for numerical evaluation. Another way to say this would be "forecasts. Py."
- **Scikit learn:** A powerful and intuitive tool for data modeling and analysis is Sklearn. Three Python utilities—NumPy, SciPy, and Matplotlib—were used in the layout. These are freely accessible, open-source tools that anyone may use.
- **Seaborn**: The upcoming iteration of Seaborn, a popular Python data analysis tool, will incorporate matplotlib. This is a user-friendly tool for imaginative data visualization.
- **H5py**: Users can use the Python h5py library to retrieve fragmented HDF5 code. HDF5 is used for storage after a significant amount of the data—mostly integers—is handled by NumPy.
- **TensorFlow**: This group of AI technologies seems to be available to the general public. It provides a wide range of tools for developing and applying various machine learning theories, most notably models based on neural networks. TensorFlow's versatility, efficiency, and ease of use make it an excellent choice for a wide range of intelligent applications.

- **Pandas**: This freeware toolkit for examining and working with data that is particular to a language is provided by Pandas. Basic data kinds and statistical analysis methods can help ensure orderly data administration, particularly when working with summary data.
- **OS:** The Py Platform element offers a range of capabilities that enable staff members to communicate with one other using the same vernacular.

### 3.3 Proposed Methodology

To accomplish its objectives, this research probably used a number of different approaches or techniques. The written content and all values had to be cleaned up for the experiment. The workflow, manipulation, collection, and AQI level had to be chosen. The classifier's efficacy was then assessed using the results of a forest classification technique that was chosen at random. **Step 1: Gathering Data:** After gathering the data from internet resources such as Kaggle, we evaluated it. This company lacks a large, thorough dataset since it is hard to collect data regarding the accurate pollutant level from high-quality analyzer and categorization type.

**Step 2: Data processing**: Each item of information was looked at separately once all practical means of obtaining data had been used. Words that are poorly chosen or ambiguous are all around us. Before utilizing the dataset, we are advised to go over its last part.

**Step 3:** Get the data ready: Both the "AQI" and the "AQI category" continue to direct the creation and analysis of the data even after it has been compiled. Training requires not only showing the data but also sorting and deleting null values. A comprehensive preparation of the data for separation has not been made.

**Step 4: Model Selection**: We decide on a prediction approach, train it using my data, and then assess it to increase reliability. In the field of machine learning, several filters are used. Despite using many designs to enhance the component design and enable the ML model to detect specific types of air pollution, a single sensor was ultimately selected to evaluate the data's trustworthiness.

**Step 5: Evaluation of Performance**: This phase's latter stages address all the ramifications. These techniques offered us a restricted level of reliability for the label categories for the two distinct air quality datasets after the training and assessment phase. Accuracy data and f1 scores were generated to verify the confusion matrices. By applying machine learning and computational

learning techniques, ascertain if all forms of contaminants in the environment may be represented as AQI values.

**Step 6: summarize and Upcoming Projects**: This section will give an outline of the field's development plan.

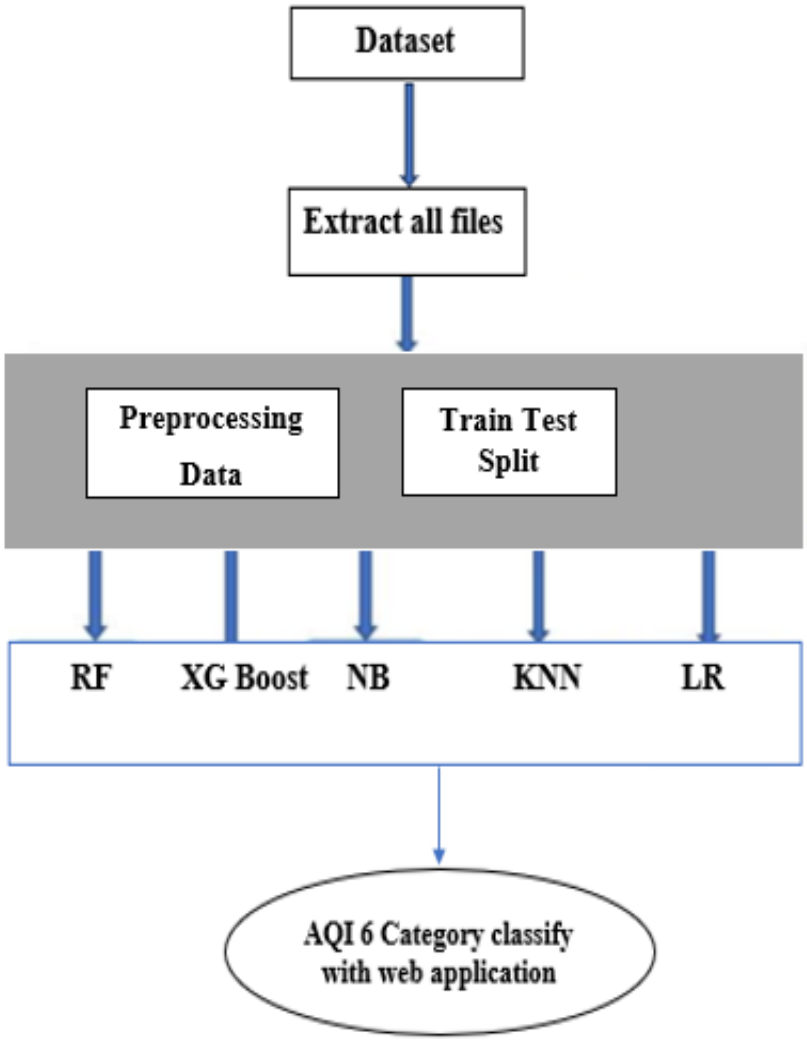



Figure 3.1: Total research procedure

The phases of our research methodology are depicted in Figure 3.1, which can be used to ascertain the state of the air classification index (AQI) values for each of the pollutants that comprise each index. Since our online data was compiled from multiple sources, it is accurate and trustworthy.

The AQI worth to every kind of air was used to build the dataset. We have reviewed the process utilized for gathering the data, removed any unnecessary null values, and polished up the phrasing to ensure that the complete data collection only contains accurate and relevant data. We develop and improve models to investigate machine learning methods using previously collected data. We also use permutation approaches to address the issue of class heterogeneity, which will guarantee fair inclusion and increase the overall efficacy of the model. Apart from giving a summary, our approach searches for ways to reduce the amount of imprecise and erroneous results. By combining machine learning technology with language comprehension techniques, we can offer an integrated strategy that regulates the results in the air quality index statistics that are utilized to predict air pollution.

### 3.4 Data Collection

The dataset was acquired through the use of the online tool Kaggle. This database includes the PM2.5 AQI assessment as well as the air pollution projection for Dhaka city during 2017 to 2022. Six files totaling six years' worth of Dhaka PM2 are included in this collection.5. The rows with the labels "AQI" and "AQI Category" become apparent, and a total of 44,571 data points are gathered. The methods for testing and training make up the two parts of the dataset. After removing duplicates and nulls, there remained 44,160 data points. The air quality index was then classified into six categories using this data: "Hazardous," "Very Unhealthy," "Unhealthy," "Unhealthy for Sensitive Groups," "Moderate," and "Good."

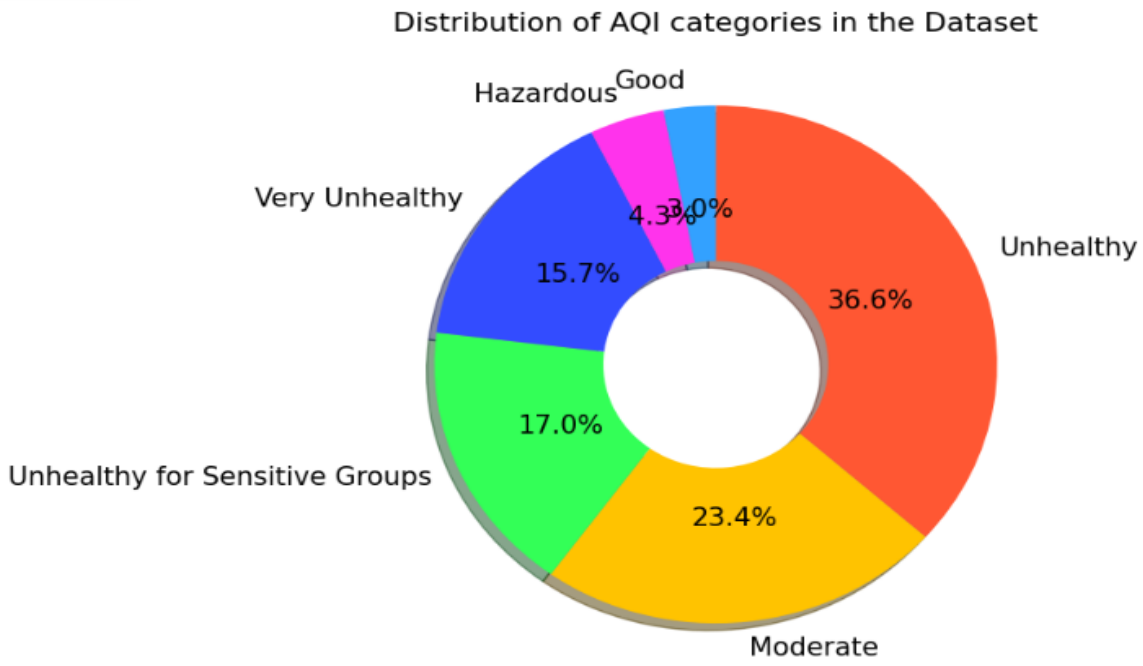
| | Date (LT) | Hour | NowCast Conc. | Raw Conc. | Conc. Unit | AQI | AQI Category | QC Name |
|---|---|---|---|---|---|---|---|---|
| 0 | 01/01/2017 01:00 | 1 | 287.8 | 287 | ug/m3 | 338 | Hazardous | Valid |
| 1 | 01/01/2017 02:00 | 2 | 297.4 | 307 | ug/m3 | 347 | Hazardous | Valid |
| 2 | 01/01/2017 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 |

Figure 3.2: An instance of PM2.5 data for Dhaka between 2017 and 2022.

Table 3.1 displays all of the data elements from each of these data files, classifying them according to eight primary categories:

Table 3.1: Description Columns in the Dataset

| 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 |



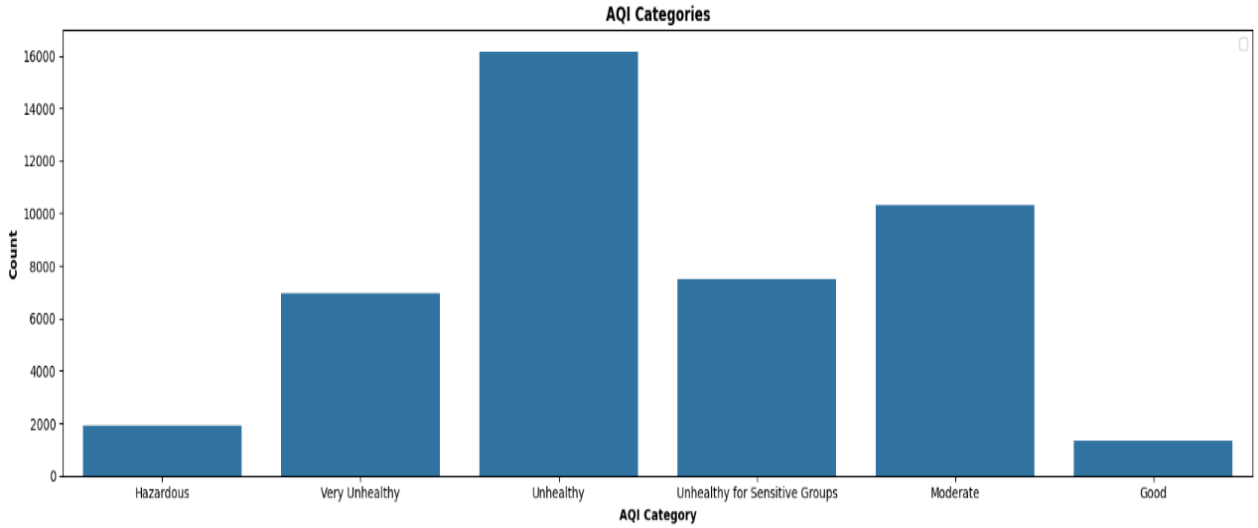



Figure 3.3: Six data classes are present in the AQI.

**Encoder label:**

Irrespective of the total number data entries in the dataset, machine learning experts frequently work with multi-feature datasets. These unique IDs can contain both letters and digits. Keywords are frequently used in education to organize facts in a way that improves user comfort and understanding.

Encoding the labels is one method of converting impulses into numbers or computer-readable language. A framework that conveys numerical values must be produced by transforming identifiers within the constraints of the previously described process. How these labels are used is ultimately up to the creators of the machine learning algorithms. As soon as the data is received, a preliminary analysis has to be done while keeping an eye on the alterations that are occurring.

## 3.5 Statistical Analysis

### 3.5.1 Analyzing the Data

The goal of the first stage of data processing is to transform the numerical aqi data into an arrangement that is able to be utilized for model training and assessment. To get the required "AQI" and "AQI Category" characteristics, information on the air quality categories linked to air pollution must be obtained. Data which could be utilized for testing and training can be obtained by combining many datasets. In order to begin fixing any mistakes, we began by removing any extraneous characters or symbols from the data stored in the database that had null values. Another

method that transforms words to sequences of numbers that may be easily combined and utilized in models for machine learning is tokenization. This exacting approach to data processing makes our models more skilled at using AQI values to determine AQI Category.

**About AQI:**

To evaluate the condition of the air every day, one uses the air quality index (AQI). It tracks the impact of sporadic air pollution on human health. Public education on the health impacts of the ambient air is the goal of the AQI. There are an overall of six AQI categories. Each classification denotes a different level of health risk. Every group also has a different color. Because of the color, people can determine quickly and readily if the air quality within their communities is bad.

The possible values of the AQI, assuming it were a scale, would range from 0 to 500. Higher air quality index (AQI) ratings are associated with increased levels of pollution and the associated health risks. For example, an AQI rating of fifties or below indicates excellent air quality, whereas an AQI rating of 300 or more indicates dangerous air pollution. For a given pollutant, an AQI score of 100 usually indicates ambient air pollution levels that, for public health safety, meet the national outdoor air quality standard in the near future. AQI ratings of 100 or below are often regarded as really good. When AQI readings surpass 100, air quality becomes a concern for everybody, first for those who are vulnerable.

Table 3.2: Air Quality Index (AQI) category range.

| **Air Quality Index** | | |
|---|---|---|
| **AQI Category** | **Index Value** | **Air quality Description** |
| Good | 0-50 | Air is satisfactory |
| Moderate | 51-100 | Air 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 is Health alert. |
| Hazardous | 300- Higher | Health warning of emergency conditions. |

### 3.5.2 Data Cleaning and Null value remove

Modifying numerical values is an essential stage in the creation of datasets. The two main strategies used in this strategy are meant to raise the standard and relevance of written content. News is edited before it is released, and it is only kept for a predetermined amount of words. Filtering procedures protect our commitment to offering instructive, legally-compliant, and high-quality content. To update the information gradually, extraneous symbols such as organizations, stop groupings, emoji solvents, specialized signs, etc. are also removed using a text correction technique. All numerical values, especially the empty values, have been cleaned of common symbols, line breaks, and some English characters. To ensure that all of the data we gather is ready for a detailed examination, we follow a rigorous preservation process. Each of the six categories' statistics is represented by three rows in an AQI presentation.

| | hour | nowcast_conc. | raw_conc. | aqi | aqi_category | month | day | aqi_map |
|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 287.8 | 287 | 338 | Hazardous | 1 | 1 | 5 |
| 1 | 2 | 297.4 | 307 | 347 | Hazardous | 1 | 1 | 5 |
| 2 | 3 | 300.2 | 303 | 350 | Hazardous | 1 | 1 | 5 |
| 3 | 4 | 306.1 | 312 | 356 | Hazardous | 1 | 1 | 5 |
| 4 | 5 | 313.8 | 322 | 364 | Hazardous | 1 | 1 | 5 |

Figure 3.4: Following data cleaning

### 3.5.3 Tokenization

Tokenization is fundamental components of contemporary data use. Words have to be tokenized and or translated into numerical sequences, in order for our system to understand English. Attributing a distinct number to every word establishes the link among written language and symbolic numerals. By giving each sequence, a certain amount of time during a planned training

session, padding helps to maximize the likelihood of agreement. To assess the many linguistic subtleties of the AQI outcome, our machines convert phrases into monetary units and verify that the overall word length varies appropriately. This phase is crucial in order for our computers to accurately assess the data and distinguish between the six distinct categories that are represented in the AQI.


Figure 3.5: Tokenization Example.

### 3.5.4 Data Preparation

In the data preparations stage, even after removing null values and duplicate data through the inclusion of four unique variables (date, month, aqi_le, and aqi_map), we did not randomly divide the data for testing the model and training. We take the most important "AQI" and "AQI Category" variables out of the 44,160 entries in the dataset on air pollution quality. Training and Testing comprise the two components of the dataset. Following the removal of duplicate entries, the data was classified as "Very Unhealthy," "Unhealthy," "Hazardous," "Unhealthy for Sensitive Groups," "Moderate," and "Good." The testing set has 8,832 data, while the training set contains 35,328 data that are used to analyze the air quality using AQI values.

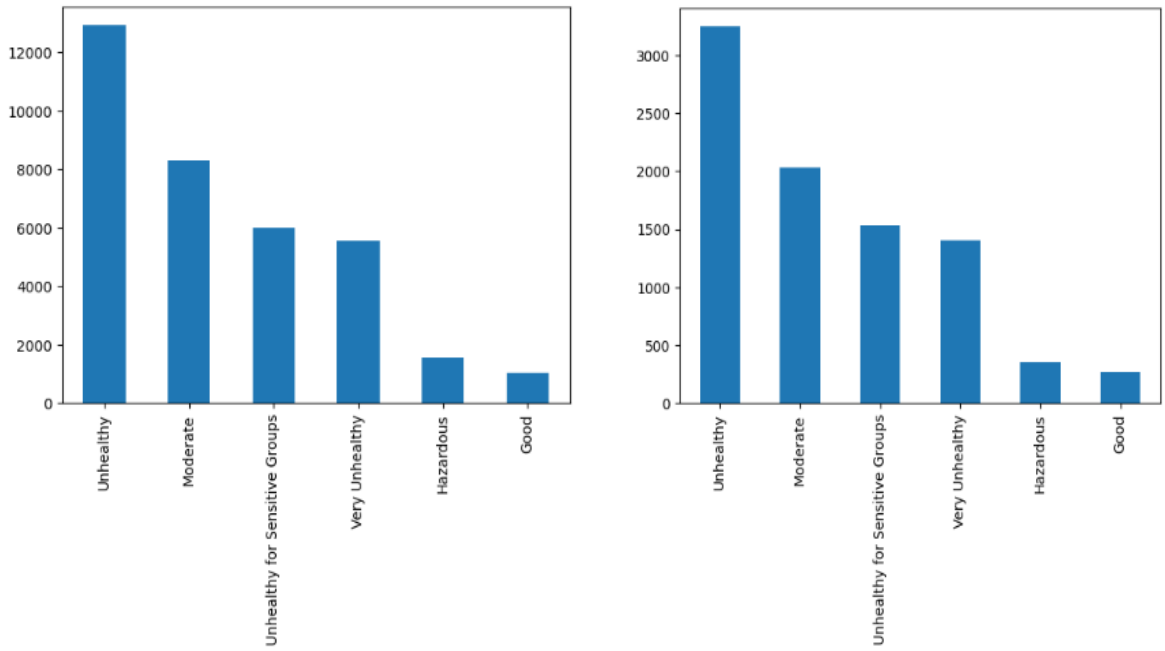



Figure 3.6: Data Count for the Train and Test AQI categories.

### 3.5.5 Models Applied

The air is murky and visibility decreases as rates of PM2.5 rise. Several methods are used in this study to compute the Dhaka PM2.5. Numerous machine learning approaches were examined, such as LR, KNN, XG Boost, Random Forest classifier, and Naive Bayes. We leverage the ability of our models to capture language variations, context, and patterns in a variety of models. We are able to accomplish our goal of creating a trustworthy system for categorizing the degree of air pollution according to AQI data thanks to this comprehensive investigation.

1. **Random Forest:** With the two classification categories, the RF Separator's tree-based method might be applied. An ML algorithm creates an organizational tree. This method is used by artificial intelligence to create hierarchical "decision trees". The stochastic forest classification technique builds several decision trees using a combination mechanism and then combines them all. This sheds lighter on the issues around overfitting. Machine learning is one of the hottest subjects in business right now because of its adaptability and ability to be utilized anywhere there is a lot of data. Owing to its many benefits, the machine learning-inspired RF Encoder is usually chosen over other approaches. The approach was first developed in 1997 and was initially designed to work with very big datasets [5].

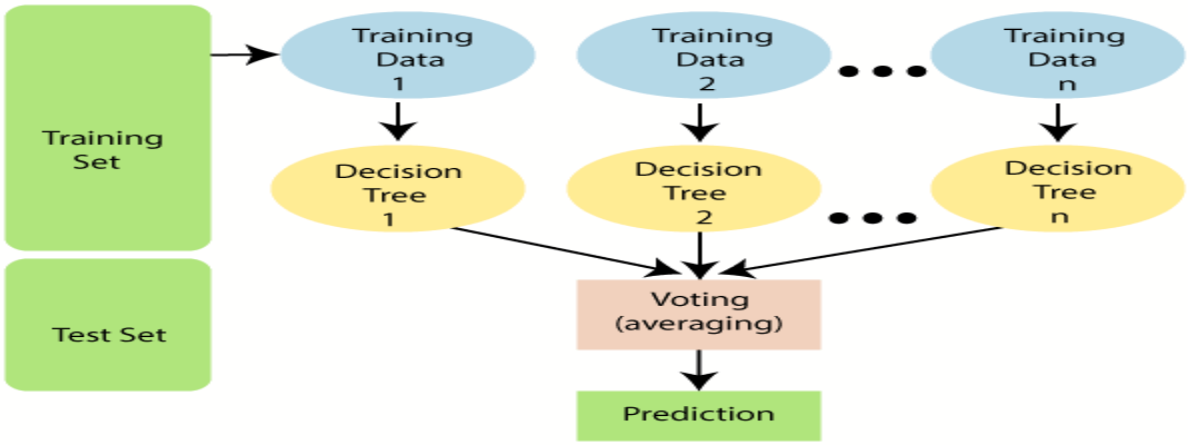



Figure 3.7: The Random Forest classification procedure [5].

2. **Naïve Bayes:** It is recommended to use the naïve Bayes method for learning algorithms even with a large number of records since it can handle large amounts of data. Among its strong points are tasks related to natural language processing, or NLP, such assessing texts for AQI classifications. The filtering procedure is easy to use and fast. Comprehending the Bayes classifier's naivety completely requires an understanding of the Bayes hypothesis. We talk about the Bayes theory in our first exchange. The basis of this argument is the idea of conditional probability. Contingency probability is the chance that one event will occur given the possibility of another. We may use our prior knowledge and the conditional probability to determine the likelihood of an event. Algorithms for analyzing sentiment, eliminating spam, and directing, to name a few uses, frequently employ naive Bayes approaches. This method's main flaw is the requirement for distinct models, even with its speed and simplicity of use. The classifier performs badly in most real-world scenarios due to the interdependence of the prediction components.[6]

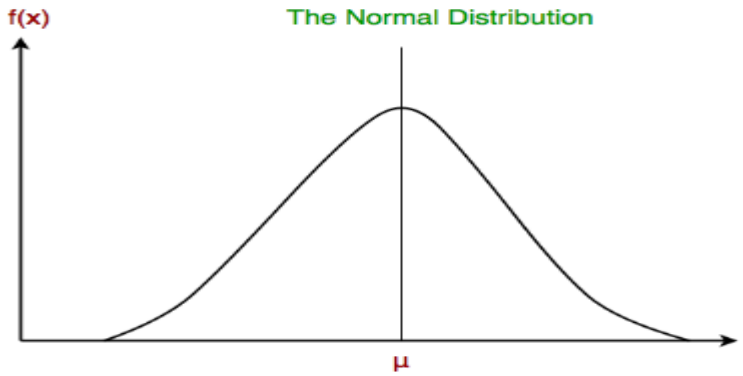



Figure 3.8: Bayesian models without refinement.[6]

3. **XG Boost:** A well-liked and successful machine learning technique is called Extreme Gradients Boost, or "XG Boost," and it may be used to address clustering problems such as regression. Slope-enhanced decision-tree topologies may be produced more quickly and easily with the help of this technique. XG Boost is an ensemble learning strategy that combines the forecasts from numerous weak learners (typically decision trees) to enhance a model. One by one, each failed student makes up for previous mistakes by learning from their less fortunate peers. XG Boost has many normalizing penalty techniques built in to guard against excessive fitting. Due to the effectiveness of training using penalty-based regularizations, the method is able to attain the appropriate degree of applicability. What is a non-linear system? Structures may be identified and trained on unexpected input using XG Boost. To make it clear, right out of the box, everything is built and functional [7].

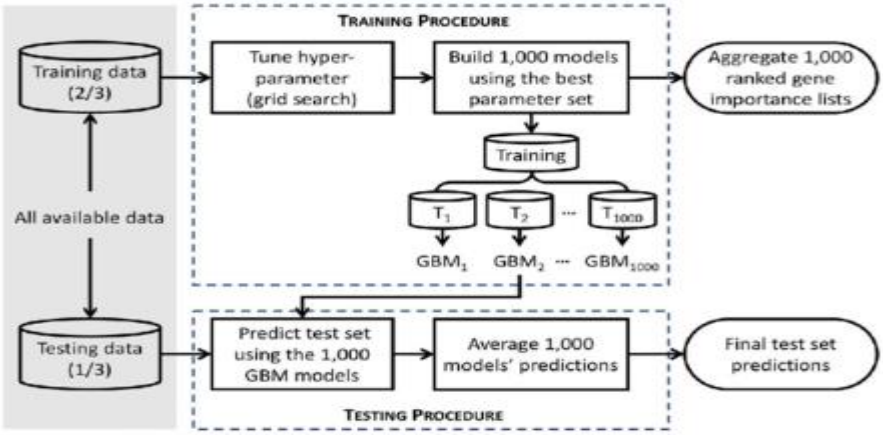



Figure 3.9: Building of the XG Boost model.[7]

4. **K-Nearest Neighbors:** Machine learning professionals employ the robust and intuitive k-nearest-neighbors (KNN) technique to solve regression and classification problems. KNN employs the K near neighbors of the most recent information item from the training set to apply the resemblance concept when estimating the value or title of a specific recent data point. This article will cover the supervising neural network (KNN) method, commonly referred to as the k-Nearest Neighbors approach, and its user-friendly features. Due to its simplicity and ease of use, the KNN technique is a widely used prediction technique. There is little to infer about the geographic distribution based on the data at hand. It is a flexible replacement for many other types of analytics in the areas of regression, classification, and

various other applications due to its capacity to handle data that is numerical as well as categorical. This unapproved method creates predictions by comparing the degree of similarity between data points in a certain batch of data. Outliers affect K-NN less than they do with other algorithms [9].

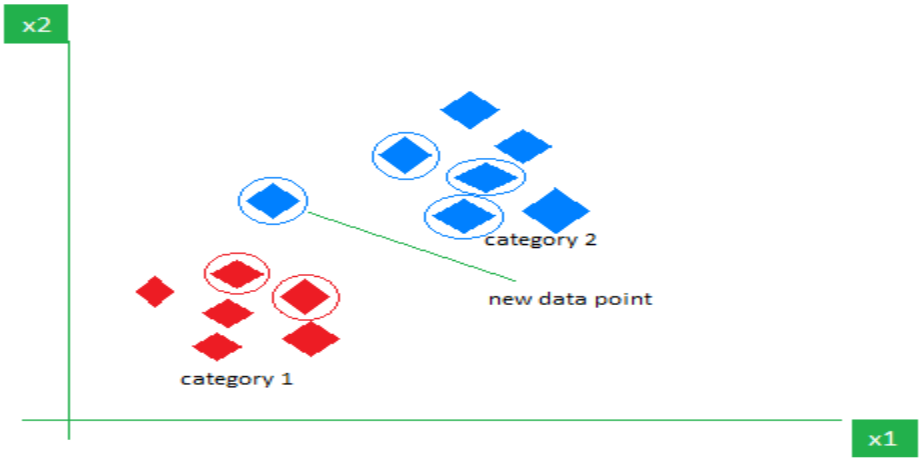



Figure 3.10: Visualizing the KNN Algorithm Operationally [9]

5. **Logistic Regression**: The LR controlled forecasting method is frequently used for single type problems, where the results are predictable yet dependent on the presence of certain data types. The fundamental functions of the logarithm regression approach are represented by a sinusoidal function for variables with values between zero and one. Logic rebound (LR) is essentially a classification technique. According to one definition, there are only two conceivable outcomes for an integer result: whether it will take place (1) or it will not occur at all (0). To put it briefly, distinct variables are elements that have no reliance on self but might nevertheless affect the results of the study. Therefore, using a logistic regression is the most effective analytical technique when working with binary data. while utilizing data that determines if a factor or an effect fits into several categories. That acknowledges that you are dealing with binary data, but only in the event that the data fits into either of the two groups [9].

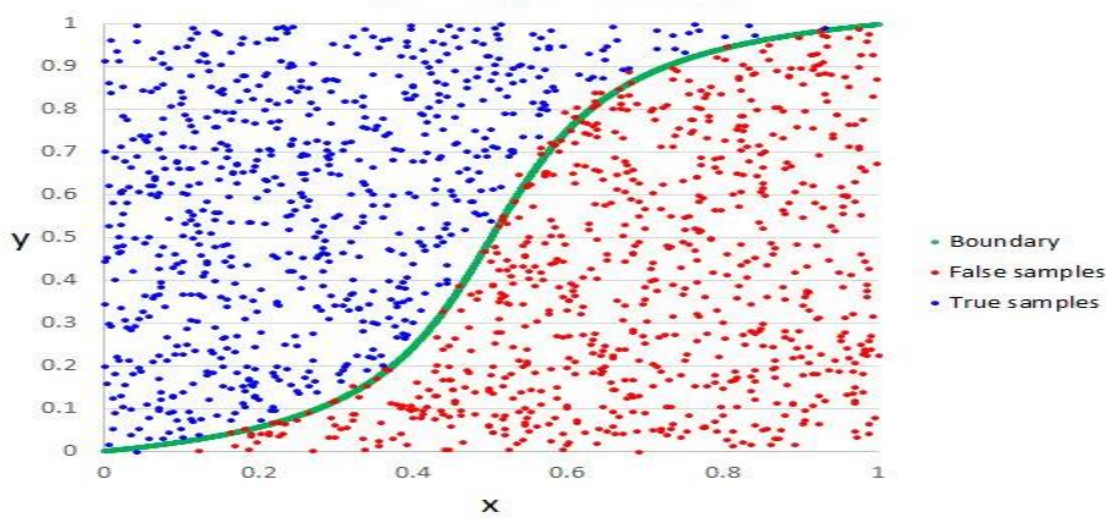



Figure 3.11: Logistic regression example [9]

### 3.6 Implementation

To ensure accuracy, details must be acquired after the earlier phases are finished. Our initiative's core idea required the completion of 10 components. If we are to reach our goal, all of these chores need to be completed.

- Compiling information.
- Preprocessing of data.
- Data cleansing for null values.
- Including four columns while preparing the data.
- The tokenization processes.
- Using models is advised for each of the five approaches.
- Building a functional prototype website for AQI classification.
- Examine and discuss the outcome.

The idea wouldn't function until we began coding it. The accuracy of five different techniques was examined. After it was finished, we assessed the method's accuracy. After examining the accuracy, we concluded that the design indicated above would be better appropriate for our requirements. After a thorough examination of relevant conceptual and numerical approaches and concepts, a set of criteria has been developed for each effort to categorize the quality of air pollution. Here are a few outcomes that could be noteworthy:

**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 Colab.
- VS code.

# CHAPTER 4

# Experiment Results and Discussion

## 4.1 Overview

In order to measure the amount of air pollution, an AQI Category is assigned. This procedure is covered in this section. Selecting a model, collecting and analyzing data, removing null values, adding more columns to improve the text, and assessing the model's performance in light of the levels of air pollution detection findings were all that was required to develop a model. This section of the article analyzes and presents the experiment's results.

## 4.2 Experimental Result

We are aware that nothing can ever yield flawless outcomes. In a manner similar to a component of we may modify the parameters of our model during training to improve accuracy. But using various approaches, we find that the accurate level is rather high. This provides as a graphic summary of the job that we are currently doing. The recalled, exactness, f1 score, allows, temperature maps, and more data are displayed in these graphics. Here, we use our data to determine which of the six AQI groups, based on AQI values, is the most often forecasted air class.

## 4.3 Classification and Descriptive Analysis

The approach we took determined the different outcomes we obtained. We were able to accurately predict the air quality classifications linked to the air pollution in Dhaka thanks to five machine learning algorithms. Following the application of these techniques to assess the relative efficacy of every component of the entire structure, a number of verification procedures were conducted to ascertain the outcome of the forecast. After each statistic was chosen, each model used a single file including information from both publicly available web sources and our own study. We used Matlab and related pre-configured modules to assess the algorithms' output when the data collecting was complete. The item is next assessed to see whether it satisfies the standards regarding the air quality index (AQI) using a comparable dataset. They fall into one of the following categories: Hazardous, Unhealthy, Very Unhealthy, Good, Moderate, and Unhealthy for Sensitive Groups. In this case, we extensively examined different models using relevant

performance requirements. Metrics including recall, precision, recall, accuracy, and total F1 score offer a thorough assessment of the strategies' effectiveness.

Table 4.1: Accuracy Table.

| **Classifier** | **Accuracy** |
|---|---|
| XG Boost | 99.03% |
| Logistic Regression | 99.74% |
| Random Forest | 99.81% |
| Naïve Bayes | 99.73% |
| KNN | 99.68% |

The following section displays the results of several classifiers. Throughout, PyCharm and CoLab were two complementary tools that performed well in tandem. There were five classifiers used in all. Naive Bayes, Logistic Regression, Random Forest, KNN, and XG Boost.

| | 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 |

Figure 4.1: Reports on the classification of Random Forest.

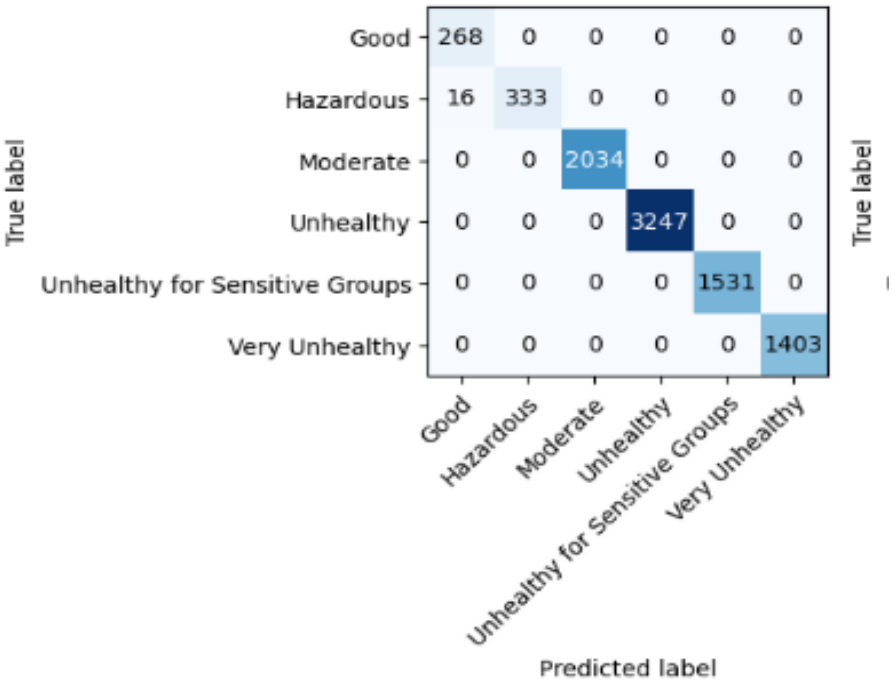



Figure 4.2: Confusion matrix of Random Forest

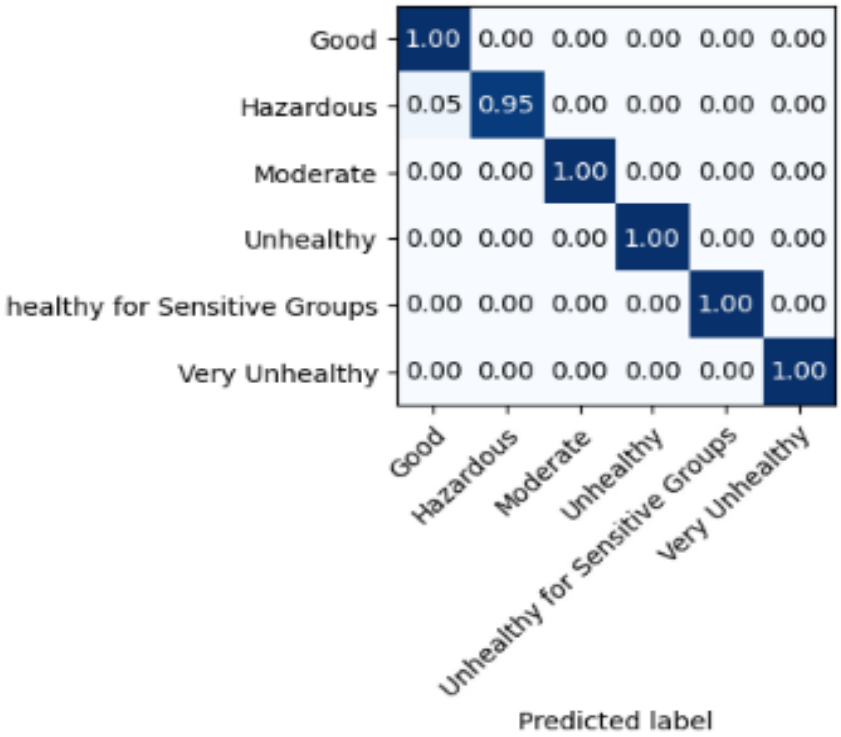



Figure 4.3: Normalization Confusion matrix of random forest.

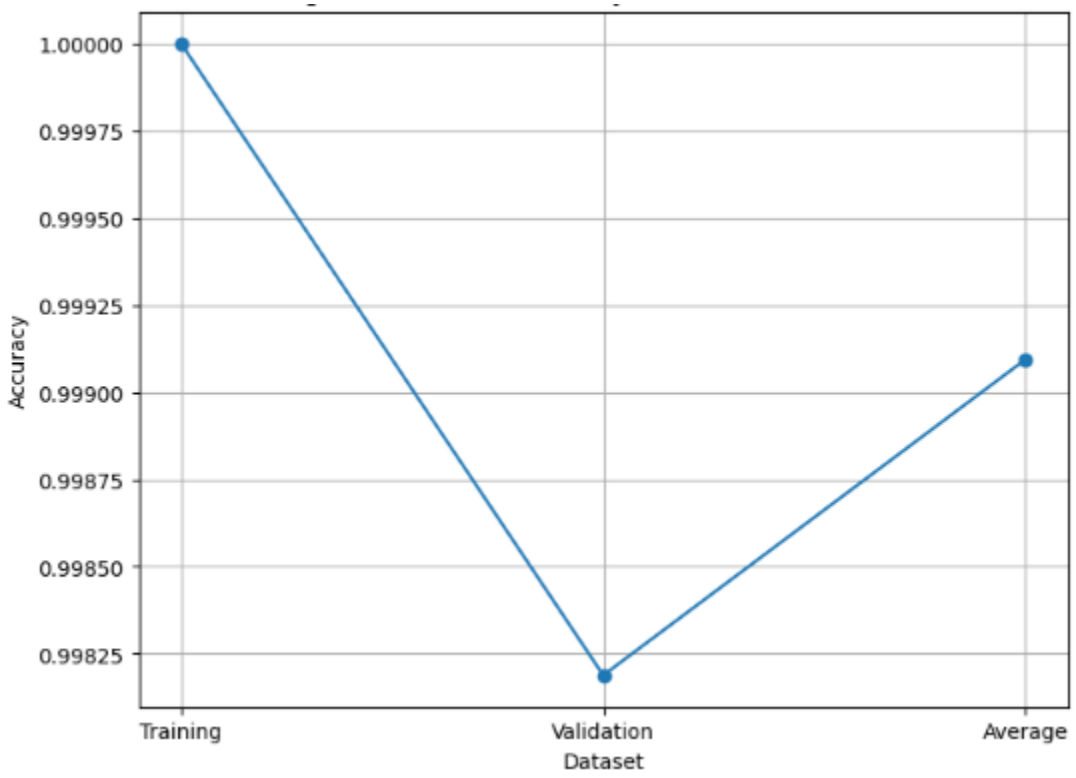



Figure 4.4: Training and Validation of RF Accuracy

As the classifier uses Random Forests to automatically classify all data to maximize accuracy, just the Classifier report is shown.

To create a useful web-based application, the most efficient model using the random forest approach is utilized to classify a particular kind of air pollution excellence, as seen in Fig. 4.5 below. Six AQI Category groups were correctly predicted by the online software using the random forest "aqi.pkl" model (Graphs 4.6, 4.7, 4.8, and 4.9): Hazardous, Unhealthy, Very Unhealthy, Good, Moderate, and Unhealthy for Sensitive Groups. Machine learning usually provides one of the most researched methods for AQI identification or predictions.


Figure 4.5: Air Quality Detector is a prototype web application.


Figure 4.6: Detecting the "Good" category in the AQI.

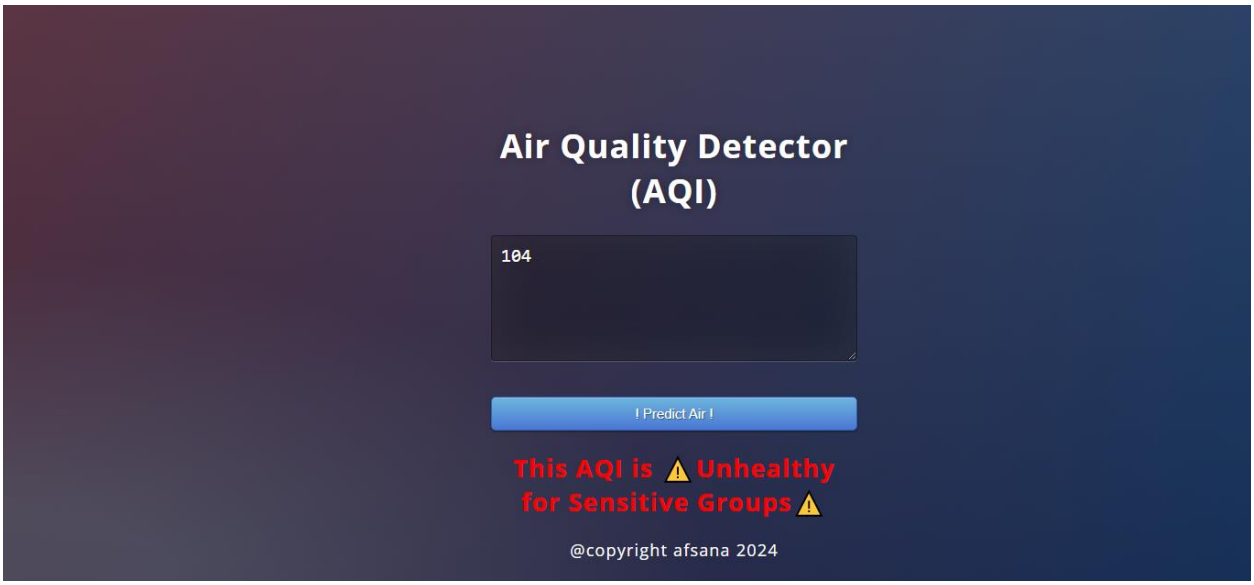



Figure 4.7: Detecting the " Unhealthy for Sensitive Groups" category in the AQI.

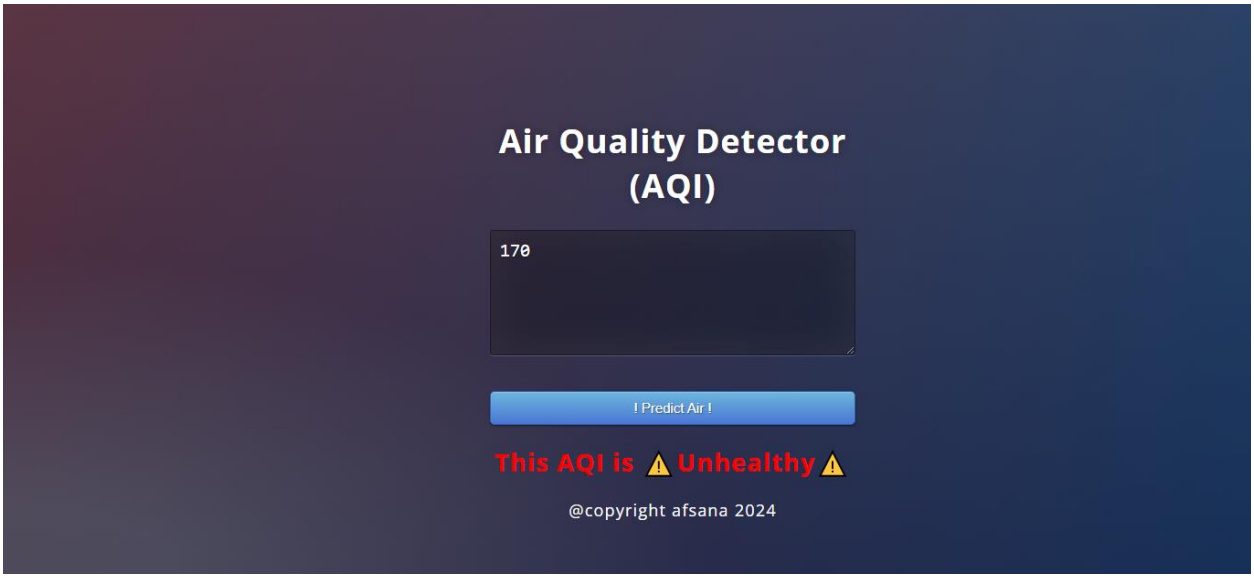



Figure 4.8: Detecting the "Unhealthy" category in the AQI.


Figure 4.9: Detecting the "Moderate" category in the AQI.

The predicted air pollution quality features in the data labeling and the ML RF classifiers approach explain the AQI Categories contaminants identification procedure, as may be seen in Figs. 4.10 afterward. The table below enumerates the two types of misleading information that pupils may distinguish, in addition to the six kinds of pollutants found in the environment's category. Machine learning-based algorithms become excellent replacements for the more popular detection or forecasting approaches since they can learn unsupervised. It worked amazingly well, especially when compared to more conventional methods.

```
# Preprocess new text data (replace 'new_text_data' with your actual new text data)
new_text_data = ["378",
                 "83",
                 "250",
                 "56",
                 "36"]

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: '378' so Air is 'Hazardous'.
AQI Value: '83' so Air is 'Moderate'.
AQI Value: '250' so Air is 'Very Unhealthy'.
AQI Value: '56' so Air is 'Moderate'.
AQI Value: '36' so Air is 'Good'.
```

Figure 4.10: Quality predictor for air pollution.

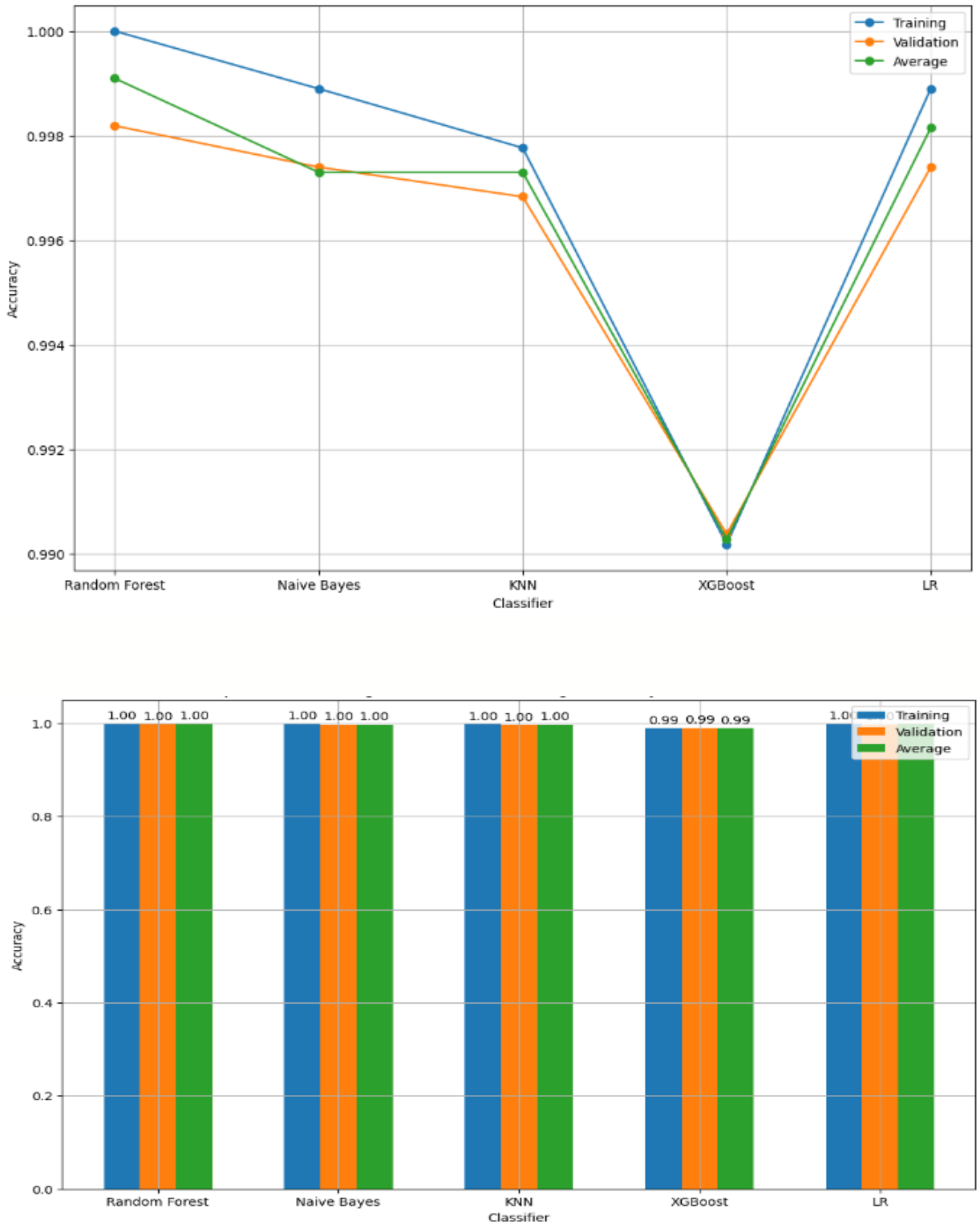



Figure 4.11: Average training, validation, and accuracy Comparing five classifiers is done.

### 4.4 Discussion

This study identifies material about air pollutants that may be grouped into an air quality indicator using machine learning techniques. Every phrase in every field ought to be accorded a lot of weight when categorizing. Our main focus has been trying to find the text beneath the AQI heading. Data is one of among the most crucial elements of any investigation. The findings of the same experiment might change significantly depending on the data presented. We were conscious that various individuals may arrive at different results as we collected Dhaka PM2.5 readings. By adding many algorithms made up of machine learning to the validity evaluations and averages, we were able to accomplish our aim. Five distinct algorithms were employed in this study. Before we could go to work, they had a lot of stuff to find. Yes, we chose to and begin working on the algorithm. Next, each strategy's accuracy was assessed. The five models' respective estimates are based on the RF classifier's final accuracy of 99.81%, which was attained through the use of techniques. Finally, an online application was created to identify the quality of the air category associated with air pollution AQI readings.

**Precision:** Reliability is defined as the ability to distinguish between all predicted and actual accomplishments. The following may be used to derive this formula:

$$Precision = \frac{TP}{TP+ FP} \text{ ---------- (1)}$$

**Recall:** Recall is the model's capacity to discriminate between each true favorable reaction and the expected positive response. The format of the formula is as follows:

$$Recall = \frac{TP}{TP+ FN} \text{ ---------- (2)}$$

**F1-Score:** As with the F1 score, the effectiveness of the model may be assessed through the combination of recall and dependability. The computation may be done using the following equation:

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall} \text{ ----------(3)}$$

**Accuracy:** Since the F1 score does, it is possible to assess the model's effectiveness via its recall and trustworthiness. The following equation may be used to calculate it:

$$Accuracy = \frac{TP + TN}{Total\ Instances}$$ ----------(4)

# CHAPTER 5

# Impact on Society, Environment and Sustainability

## 5.1 Society Impact

Relying on machine learning to forecast PM2.5 levels in Dhaka could result in far-reaching societal consequences. Particulate matter that has a maximum length of two millimeters is referred to as PM2.5. Considering how deeply these particles pierce lung tissue, they could be hazardous to health. These are some potential repercussions on society:

- **Improving Public Health**: When high PM2.5 concentrations are anticipated in advance, localities and governments can reduce exposure by proactively lowering levels, especially for vulnerable groups like children, older individuals, and those with lung conditions.

- **The surroundings Awareness**: Timely and accurate assessments of the environmental consequences of diverse activities can influence people and organizations to adopt sustainable practices.

- **Developing Regulations and Civic Growth**: Governments may make decisions on legislation and urban expansion using PM2.5 estimates. For example, the data may have an impact on the locations of hospitals, schools, and residential areas in order to limit access to areas with elevated pollution levels.

Even while using machine learning to predict air quality can result in positive changes, it's important to take into account data privacy, equitable access to data, and ethical issues to ensure that society as a whole gain equitably.

## 5.2 Environment Impact

The program that forecasts Dhaka's PM2.5 levels of pollution using algorithmic learning (ML) might have negative environmental effects as well as to its economic benefits. Some possible environmental impacts of the use of machine learning (ML) air quality forecasting include the following:

- **Emissions mitigation techniques**: to determine the origins and trends of PM2.5 pollution, machine learning methods may be applied. Applying this knowledge will facilitate the execution of targeted environmental management and protection strategies. Legislation, for instance, may be enacted to lower emissions from vehicles or certain businesses if these activities are frequently connected to elevated pollution levels.
- **Conformity to environmental requirements and decreased emissions**: The projections may be used to monitor and enforce conformity to environmental requirements. Businesses and automobiles may be urged to use more ecologically friendly technologies and practices in order to reduce the quantity of pollutants they contribute to the air, since high-pollution periods may be anticipated and managed.
- **Managing Climate Changes:** One form of pollution in the air that could contribute to global warming worse is particulates. Machine learning-based air quality predictions can help reduce harmful pollutants and boost efforts to mitigate the consequences of global warming when paired with focused preventative actions.

There could be some benefits to the environment from machine learning (ML)-based contamination projections, but it's crucial to make certain that such algorithms are a part of a larger environmental strategy that includes regulations, growth, and monitoring that is ecologically friendly. When choosing green solutions, ethical considerations must also be included in order to prevent biases or unintended negative impacts.

### 5.3 Ethical Aspects

It is important to carefully analyze the ethical considerations highlighted by using machine learning to anticipate the PM2.5 pollution levels in Dhaka. Among the morally significant elements are the following:

- **Data security:** When collecting and utilizing data for air quality forecasts, strict confidentiality guidelines must be adhered to. Make sure that any private information is aggregated and anonymized to avoid identifying particular people. Obtaining appropriate consent and making sure the procedures for collecting data are transparent are crucial.

- **Diversity and Fairness**: Recognize that inaccurate training data may be used in machine learning models. Make that the intended results and activities are fair and do not adversely affect any specific demographic. Consider the social and economical implications of air quality legislation and regulations to avoid exacerbating already-existing inequities.
- **Definability and Disclosure**: Machine learning algorithms need to be understandable and visible in order to anticipate air quality. Predictions and the factors influencing them should be understandable to users, legislators, and the general public. The feeling of responsibility and faith in the system are fostered by this transparency.

- **International Cooperation**: Recognize the interconnectedness of environmental concerns and collaborate with regional and international groups to share knowledge, solutions, and ideas in order to address challenges related to global air quality.

When developing technology for air quality prediction, ethical considerations should be considered at every stage, from data collection and model training to installation and ongoing monitoring. The ethical and sustainable application of machine learning (ML) for specialists in the environment depends on striking a balance between technological advancements and ethical principles.

### 5.4 Sustainability Plan

To use machine learning to anticipate PM2.5 air quality in Dhaka, an approach that takes environmental responsibility, long-term efficacy, and ethical factors into account must be developed. The following is an advised ecological strategy:

- **Continuous Data Collection:** Establish a reliable system for collecting air quality data on a regular basis. When updating and refining the machine learning model, emerging patterns and new data should be considered. This ensures that approximations will consistently be precise and relevant.
- **Free Researcher Access**: To promote transparency and collaboration, make information on air quality and automated learning models available to the general public. This facilitates contact between the general public, lawmakers, and academics who are

employing the data, which encourages transparency and cooperative problem-solving techniques.

- **Innovation and Development**: Constant expenditures in R&D are necessary to maintain the state-of-the-art in ecology and machine learning. In order to provide continuous enhancements to the sustainability plan, investigate technologies that boost the precision and effectiveness of air quality forecasts.

By combining these elements into a comprehensive sustainable strategy, the application of artificial intelligence to predict PM2.5 emissions in Dhaka has the potential to eventually enhance social advancement, health of the environment, and equitable growth. As long as the approach is regularly assessed and modified, it will continue to be effective in addressing new problems.

# CHAPTER 6

# Conclusion and Future Research

## 6.1 Summary of the Study

We now know a great deal more about the problem thanks to this inquiry. Air quality predictions continue to be debatable. The yearly decrease in patient usage of air pollution monitoring initiatives that employ the air quality criterion is mostly caused by this. Using machine learning, we were able to classify air danger into seven fundamental categories: hazardous, unhealthy, extremely unhealthy, good, moderate, and unhealthy for those with special needs. Based just on appearances, the machine learning approach has also been used to evaluate the possibility of transmitting air quality index (AQI) readings for contaminated air.

As it was already shown, the research's objective is to understand as much as possible about this subject. This was accomplished by combining data from Kaggle of Dhaka 2017–2022 PM2.5 with data from an online application that included six classifications: dangerous, unhealthy, extremely unhealthy, excellent, moderate, and unhealthy for sensitive people. With the use of the labeling extracted from the data files, we were able to analyze and educate our five techniques based on the randomized forest classifier's extremely accurate ability to recognize the precise "AQI" values that belonged to the "AQI Category". It is simpler to ascertain how much the intensity of air pollution differs across messages when using the model forecasting technique.

## 6.2 Conclusion

A daily summary on air quality can be created with an air quality index (AQI). This indicator calculates how air pollution affects health in the short term. The AQI was developed to inform the public about the impact of the regional air quality on health. The degradation of the surroundings is employed in the same way as it occurs in Dhaka. It is essential to increase public awareness of the problem and hasten the effort to end air pollution. Those that were affected by Dhaka's environment can be assisted, appropriate online conduct is encouraged, and people are urged to report any incidents of air pollution. Using the AQI data for each category, lower Dhaka PM.25 2017–2022, establish safe zones, and abolish air pollution. Join us in reducing the negative environmental consequences of air pollution and promoting a polite and welcoming environment in Dhaka. One method that might be used is supervised learning algorithms, which are trained on

a dataset made up of six resources that are part of the air quality Index (AQI) Classes. The effectiveness of a computerized monitoring of air quality system can be influenced by several factors. These consist of the kind of algorithms used, their configuration, and the size and caliber of the original data set. This report discusses the issues we encountered with our air pollution detection approach. A very accurate prediction can be used to assist decide whether among the six AQI groups can be considered air pollution. The top five methods of classification identify air pollution with an accuracy rate of 99.81% by comparing their results to the local air quality index. Our model performs better when RF is used. An online tool that determines if AQI values indicate air pollution is another application that makes use of the RF algorithm.

### 6.3 Possible impacts

We argue that changing the observation site has an effect on the outcomes of our method. As a result, we work harder to do our tasks accurately. By adhering to our security policies, we guarantee the authenticity of everything we provide. As a result, we decide to utilize a merged online dataset. Compared to traditional diagnostic methods, ML models may be able to identify language and symptoms associated with an AQI category more rapidly. Prompt effort to minimize air pollution and early diagnosis might enhance overall outcomes. It makes perfect sense to stop air pollution above Dhaka since machine learning techniques are widely used in a variety of different contexts, including mobile apps and websites. If we had seemed friendlier, we may have helped others who weren't as inclined to seek for help.

### 6.4 Future Work

Future work on estimating PM2.5 pollution levels in Dhaka utilizing machine learning (ML) may involve several advancements and enhancements. more research on the moral implications of air quality predictions. To guarantee an equitable distribution of anticipated benefits, develop strategies to minimize inefficiencies in data collection and model output. Consider how climate change affects air quality patterns. Make models that account for changing weather patterns and assess how climate-related factors could affect PM2.5 concentrations in the future. Ultimately, additional study and advancement in these areas may lead to more habitable and healthful urban environments, which would raise and amplify the effectiveness of PM2.5 contaminants in the city's predictive machine learning air quality forecasts.

# 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 Neural Network (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 AIR POLLUTION PREDICTION

ORIGINALITY REPORT

| 9% | 5% | 1% | 6% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | Submitted to BRAC University<br>Student Paper | 4% |
| 2 | dspace.daffodilvarsity.edu.bd:8080<br>Internet Source | 2% |
| 3 | Submitted to Daffodil International University<br>Student Paper | 1% |
| 4 | Submitted to Brown Mackie College<br>Student Paper | <1% |
| 5 | www.rockwool.com<br>Internet Source | <1% |
| 6 | decisionsciences.org<br>Internet Source | <1% |
| 7 | Submitted to Coventry University<br>Student Paper | <1% |
| 8 | Submitted to Le Moyne College<br>Student Paper | <1% |