

**VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF  
TECHNOLOGY**  
**Department of Computer Engineering**



Project Report on

**PradushanCheck: Comprehensive Urban Air Quality Forecasting and  
Monitoring**

In partial fulfilment of the Fourth Year, Bachelor of Engineering (B.E.) Degree in  
Computer Engineering at the University of Mumbai  
Academic Year 2023-2024

**Submitted by**

Ashutosh Mishra - D17C - 36

Muskan Chhabria - D17C-13

Vanshika Thakur - D17C - 57

Nikhil Haswani - D17A - 22

**Project Mentor**

Dr. Gresha Bhatia

**VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF  
TECHNOLOGY**  
**Department of Computer Engineering**



## Certificate

This is to certify that **Ashutosh Mishra (D17C, 36)**, **Muskan Chhabria (D17C, 13)**, **Vanshika Thakur (D17C, 57)** and **Nikhil Haswani (D17A, 22)** of Fourth Year Computer Engineering studying under the University of Mumbai have satisfactorily completed the project on “**PradushanCheck: Comprehensive Urban Air Quality Forecasting and Monitoring**” as a part of their coursework of PROJECT-II for Semester-VIII under the guidance of their mentor **Dr. Gresha Bhatia** in the year 2023-24.

This project report entitled **PradushanCheck: Comprehensive Urban Air Quality Forecasting and Monitoring** by *Ashutosh Mishra, Muskan Chhabria, Vanshika Thakur and Nikhil Haswani* is approved for the degree of **B.E. Computer Engineering**.

Programme Outcomes	Grade
PO1,PO2,PO3,PO4,PO5,PO6,P O7, PO8, PO9, PO10, PO11, PO12 PSO1, PSO2	

Date:

Project Guide:

-----

# **Project Report Approval**

## **For**

### **B. E (Computer Engineering)**

This project report entitled **PradushanCheck: Comprehensive Urban Air Quality Forecasting and Monitoring** by *Ashutosh Mishra, Muskan Chhabria, Vanshika Thakur and Nikhil Haswani* is approved for the degree of **B.E. Computer Engineering**.

Internal Examiner

-----

External Examiner

-----

Dr. Nupur Giri (Head of Department)

-----

Dr. (Mrs) JM Nair (Principal)

-----

Date:  
Place: Mumbai

# Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

-----  
Ashutosh Mishra (36)

-----  
Muskan Chhabria(13)

-----  
Vanshika Thakur(57)

-----  
Nikhil Haswani(22)

Date:

# Acknowledgement

We are thankful to our college Vivekanand Education Society's Institute of Technology for considering our project and extending help at all stages needed during our work of collecting information regarding the project.

It gives us immense pleasure to express our deep and sincere gratitude to Deputy HOD **Dr. Gresha Bhatia** (Project Guide) for her kind help and valuable advice during the development of project synopsis and for her guidance and suggestions.

We are deeply indebted to Head of the Computer Department **Dr. (Mrs.) Nupur Giri** and our Principal **Dr. (Mrs.) J. M. Nair** , for giving us this valuable opportunity to do this project.

We express our hearty thanks to them for their assistance without which it would have been difficult in finishing this project synopsis and project review successfully.

We convey our deep sense of gratitude to all teaching and non-teaching staff for their constant encouragement, support and selfless help throughout the project work. It is a great pleasure to acknowledge the help and suggestion, which we received from the Department of Computer Engineering.

We wish to express our profound thanks to all those who helped us in gathering information about the project. Our families too have provided moral support and encouragement several times.

## Computer Engineering Department COURSE OUTCOMES FOR B.E PROJECT

Learners will be to,

Course Outcome	Description of the Course Outcome
CO 1	Able to apply the relevant engineering concepts, knowledge and skills towards the project.
CO2	Able to identify, formulate and interpret the various relevant research papers and to determine the problem.
CO 3	Able to apply the engineering concepts towards designing solutions for the problem.
CO 4	Able to interpret the data and datasets to be utilised.
CO 5	Able to create, select and apply appropriate technologies, techniques, resources and tools for the project.
CO 6	Able to apply ethical, professional policies and principles towards societal, environmental, safety and cultural benefit.
CO 7	Able to function effectively as an individual, and as a member of a team, allocating roles with clear lines of responsibility and accountability.
CO 8	Able to write effective reports, design documents and make effective presentations.
CO 9	Able to apply engineering and management principles to the project as a team member.
CO 10	Able to apply the project domain knowledge to sharpen one's competency.
CO 11	Able to develop a professional, presentational, balanced and structured approach towards project development.
CO 12	Able to adopt skills, languages, environment and platforms for creating innovative solutions for the project.

# Index

Chapter No.	Title	Page No.
<b>1</b>	<b>Introduction</b>	
1.1	Introduction to the project	12
1.2	Motivation for the project	13
1.3	Problem Definition	14
1.4	Existing Systems	15
1.5	Lacuna of the Existing Systems	15
1.6	Relevance of the Project	16
<b>2</b>	<b>Literature Survey</b>	
A	Brief overview of Literature Survey	17
2.1	Research Papers a. Abstract of the research paper b. Inference drawn from the paper	17
2.3	Inference Drawn	27
2.4	Comparison with the Existing Systems	28
<b>3.</b>	<b>Requirement Gathering for the proposed System</b>	
3.1	Introduction to Requirement Gathering	29
3.2	Functional Requirements	30
3.3	Non-Functional Requirements	32
3.4	Hardware, Software, Technology and tools utilised	34
<b>4.</b>	<b>Proposed Design</b>	
4.1	Block diagram of the system	35
4.2	Modular design of the system	39

4.3	System Design	41
4.4	Detailed Design (Flowchart)	43
4.5	Project Scheduling & Tracking using Timeline / Gantt Chart	46
<b>5.</b>	<b>Implementation of the Proposed System</b>	
5.1	Methodology employed for development	47
5.2	Algorithms and flowcharts for the respective modules developed	47
5.3	Datasets source and utilization	50
<b>6.</b>	<b>Testing of the Proposed System</b>	
6.1	Introduction to testing	52
6.2	Types of tests Considered	52
6.3	Various test case scenarios considered	54
<b>7.</b>	<b>Results and Discussions</b>	
7.1	Screenshots of User Interface (UI) for the respective module	57
7.2	Performance Evaluation measures	61
7.3	Input Parameters / Features considered	63
7.4	Comparison of results with existing systems	64
7.5	Inference drawn	64
<b>8.</b>	<b>Conclusion</b>	
8.1	Limitations	65
8.2	Conclusion	65
8.3	Future Scope	66
	<b>References</b>	67
	<b>Appendix</b>	70
<b>1</b>	<b>Paper I</b>	



a	Paper I	70
b	Plagiarism Report of Paper I	70
c	Project review sheet i. Review 1 (10th February, 2024) ii. Review 2 (9th March, 2024)	71

## LIST OF FIGURES

Figure No	Heading	Page No
4.1	Block diagram of the System	26
4.2	Modular Diagram of the System	30
4.3	System Diagram	32
4.4	Flowchart	34
4.5	Gantt chart	37
5.1	CSV of Bengaluru	54
5.2	New Delhi CSV	55
7.1	Screenshot of AQI at Bengaluru	56
7.2	Screenshot of AQI at Hyderabad	56
7.3	Screenshot of AQI at Kolkata	57
7.4	Comparison of Air Quality of New Delhi, Bengaluru, Kolkata and Hyderabad	57
7.5	Visualization of data via dashboard	58
7.6	Temporal AQI Heatmap	58
7.7	PM 2.5 Heatmap Comparison of the	59

	four major cities	
7.8	PM10 Heatmap Comparison of the four major cities	59

## LIST OF TABLES

<b>Figure No</b>	<b>Heading</b>	<b>Page No</b>
2.3	Comparison of Results with Existing Systems	23
5.4	Banglore Dataset in tabular format	51
5.5	New Delhi Dataset in tabular format	51

# Abstract

“Increasing urbanization and industrialization have led to a pressing concern for urban air quality and its impact on public health. To address this challenge, we propose a comprehensive solution leveraging advanced technologies. This project aims to establish a real-time air quality monitoring system for urban areas by harnessing the power of big data from air quality sensors and applying cutting-edge machine learning algorithms. By continuously collecting and analyzing air quality data, the system will provide up-to-the-minute updates on pollution levels, enabling residents and authorities to make informed decisions. Furthermore, the integration of predictive modeling will allow the system to forecast potential pollution events, enhancing proactive measures and public awareness. The proposed approach aspires to contribute to healthier and more sustainable urban environments, fostering a better quality of life for all.”

**Keywords:** increasing urbanization, industrialization, urban air quality, public health, advanced technologies, real-time air quality monitoring, big data analytics, machine learning, pollution levels, informed decisions, predictive modeling, historical data patterns, current trends, potential pollution spikes, hazardous conditions, proactive measures, cutting-edge technology, societal need, healthier urban living, air quality dynamics, public awareness, improved quality of life, cleaner environment, safer environment, sustainable urban environment, environmental insights, pollution control measures.

# **Chapter 1: Introduction**

## **1.1. Introduction**

The quality of the air we breathe is a critical determinant of public health and environmental well-being, particularly in urban areas where industrial activities, vehicular emissions, and other factors can lead to elevated pollution levels. As cities continue to grow and develop, the need for effective air quality monitoring and management becomes increasingly urgent. The advent of advanced technologies, including big data analytics and machine learning, offers a transformative opportunity to address this challenge.

This project centers around the development of a real-time air quality monitoring system tailored for urban environments. By harnessing the vast amount of data generated by air quality sensors strategically positioned throughout the city, coupled with sophisticated machine learning algorithms, the system aims to provide instantaneous updates on the state of the air. This information empowers both residents and relevant authorities to make informed decisions that can safeguard public health.

Going beyond real-time monitoring, this project also embraces the power of predictive modeling. By analyzing historical data patterns and current trends, the system can anticipate potential pollution spikes or hazardous conditions. This predictive capability not only allows for timely interventions but also facilitates proactive measures to mitigate pollution impacts.

In this pursuit, the project merges cutting-edge technology with a critical societal need, paving the way for healthier urban living. By enhancing our understanding of air quality dynamics and fostering greater public awareness, the system aspires to contribute to improved quality of life, creating a cleaner, safer, and more sustainable urban environment for present and future generations.

## 1.2. Motivation

The escalating levels of air pollution, particularly in urban areas, have become a pressing concern worldwide. The detrimental impact of poor air quality on human health and the environment underscores the urgency to address this issue. This urgency serves as the primary motivation for our project, “PradushanCheck: Comprehensive Urban Air Quality Monitoring and Forecasting”.

Air pollution has been linked to a multitude of serious health conditions, including bronchitis, heart disease, pneumonia, and lung cancer. Furthermore, it contributes to other environmental issues such as global warming, acid rain, reduced visibility, smog formation, climate change, and even premature deaths.

The ability to predict the Air Quality Index (AQI) in advance is crucial in mitigating the effects of air pollution. This is especially true in urban areas where rapid industrial and vehicular developments have exacerbated air quality issues. Machine learning techniques have emerged as efficient tools for predicting air pollution levels. These techniques can analyze vast amounts of data and provide high-accuracy predictions.

In developing countries like India, the need for air quality prediction is even more critical due to the rapid pace of development across various sectors. The adverse effects of this development on air quality necessitate effective monitoring and prediction mechanisms.

Our project aims to leverage machine learning algorithms to monitor and predict urban air quality accurately. By doing so, we hope to enable timely preventive measures and informed decision-making regarding air quality management.

In conclusion, the motivation for “PradushanCheck” lies in addressing the urgent need for comprehensive urban air quality monitoring and forecasting. Through this project, we aim to contribute to healthier living conditions and a better environment.

### **1.3. Problem Definition**

The problem statement, “Applying Machine Learning techniques to forecast the possibility of Air Pollution in an Urban Environment,” involves the development and application of machine learning models to predict future levels of air pollution in urban settings. The task requires the use of historical and current air quality data, which includes various pollutants like PM2.5, PM10, NO2, SO2, CO, O3, etc., to train machine learning algorithms. These algorithms could range from simple linear regression models to more complex deep learning models. The goal is to accurately forecast air pollution levels, which could be represented as the Air Quality Index (AQI) or concentrations of specific pollutants. The focus on urban environments is crucial as these areas typically have different pollution sources and patterns compared to rural areas and tend to have more air quality monitoring stations providing high-resolution spatio-temporal data. Accurate predictions can aid in timely decision-making and planning for pollution control measures, contributing significantly to improving public health and environmental conditions.

## **1.4. Existing Systems**

There are several existing systems for monitoring and forecasting air quality. Some of the examples are:

1. System for Integrated modelLing of Atmospheric coMposition (SILAM): This is an air quality forecast model that aims to predict dust, particulate matter (PM<sub>2.5</sub> and PM<sub>10</sub>), gaseous pollutants (CO, O<sub>3</sub>, SO<sub>2</sub> and NO<sub>2</sub>) and air quality index.
2. Environmental information FUsion SERvice (ENFUSER): This is a very high-resolution city-scale air quality model for Delhi.
3. Internet GIS-Based Air Quality Monitoring and Forecast: This system promotes the dissemination of datasets related to air quality through intuitive GUI and provides real-time information support for stakeholders like academicians, scientists, decision-makers, and the general public.

There are also various monitoring networks based on low-cost sensors that have been deployed around the world to aid or supplement air monitoring. Some representative examples are the IVAN project in Imperial County—California, the Lufdaten project, and the Opensense project.

## **1.5. Lacuna of the Existing System**

1. The existing systems have calculated the Air Quality Index (AQI) for a particular city for the entire day. Keeping the changing weather conditions, its hour-wise prediction should be done for a broad understanding.
2. The data we use may have some outliers / anomalies, i.e. the values which are dissimilar with the other values which can affect the ML model in some way or the other.

## **1.6. Relevance of the Project:**

Air quality monitoring systems are essential for several reasons. They help in reducing exposure to pollutants, viruses, and bacteria, provide cleaner and healthier indoor environments for homes, offices, schools, etc., protect from harmful emissions from industries, vehicles, power plants, etc., conserve the environment by preventing air pollution and its effects on ecosystems, provide data for research and analysis to understand air quality trends and patterns, enable efficient industrial operations by optimizing energy use and reducing emissions, and mitigate climate change by reducing greenhouse gasses and improving air quality.



# Chapter 2: Literature Survey

## A. Overview of literature survey:

In the field of environmental research, literature on comprehensive air quality forecasting and monitoring is fundamental. With growing concerns about the impacts of air pollution, researchers explore techniques and systems to forecast and monitor air quality effectively. These papers examine current science, innovative methodologies, and policy implications, providing insights for informed decision-making and sustainable solutions.

## B. Related Works

### 2.1. Research Papers :

1. **Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis** N. Srinivasa Gupta,<sup>1</sup>Yashvi Mohta,<sup>2</sup>Khyati Heda,<sup>2</sup>Raahil Armaan,<sup>2</sup>B. Valarmathi,<sup>2</sup>and G. Arulkumaran, Hindawi Journal of Environmental and Public Health, Volume 2023, Article ID 4916267

a) **Abstract of the research paper:** An index for reporting air quality is called the air quality index (AQI). It measures the impact of air pollution on a person's health over a short period of time. The purpose of the AQI is to educate the public on the negative health effects of local air pollution. The amount of air pollution in Indian cities has significantly increased. There are several ways to create a mathematical formula to determine the air quality index. Numerous studies have found a link between air pollution exposure and adverse health impacts in the population. Data mining techniques are one of the most interesting approaches to forecast AQI and analyze it. The aim of this paper is to find the most effective way for AQI prediction to assist in climate control. The most effective method can be improved upon to find the most optimal solution. Hence, the work in this paper involves

intensive research and the addition of novel techniques such as SMOTE to make sure that the best possible solution to the air quality problem is obtained. Another important goal is to demonstrate and display the exact metrics involved in our work in such a way that it is educational and insightful and hence provides proper comparisons and assists future researchers. In the proposed work, three distinct methods—support vector regression (SVR), random forest regression (RFR), and CatBoost regression (CR)—have been utilized to determine the AQI of New Delhi, Bangalore, Kolkata, and Hyderabad.

**b) Inference drawn:** Researchers from all across the world are attempting to find a solution to the worldwide problem of air pollution. In order to predict the AQI with accuracy, machine learning techniques were looked into. The current study evaluated the three top data mining methods' performance (SVR, RFR, and CR) for forecasting the precise AQI data in a few of the most populated and contaminated cities in India. In order to obtain more accurate and consistent findings, the class data was equalized using the synthetic minority oversampling method (SMOTE). This novel strategy involved balancing the datasets, utilizing them, and then closely comparing the outcomes of the balanced and unbalanced datasets to ensure that the findings were very accurate. The better results were then confirmed using statistical techniques including RMSE, MAE, MSE, and R-SQUARE.

## **2. Soubhik Mahanta; T. Ramakrishnudu; Rajat Raj Jha; Niraj Tailor.**

### **Urban Air Quality Prediction using Regression Analysis**

**a) Abstract of the research paper:** In the last several years, air pollution has risen steadily in urban environments. Cities like Gurugram, Faisalabad, Delhi, Beijing are few of the world's most polluted cities and have seen a dangerous rise in air pollution levels. Forecasting is important because of the human, ecologic and economic toll of pollution, and is a useful investment at individual and

community levels. Accurate forecasting will help us plan in advance, decreasing the effects on health and the costs associated. Local weather conditions strongly affect air pollution levels. Generating deterministic models to study air pollutant behavior in environmental science research is often not very accurate because they are complex and need simulation at the molecular interaction level. Here comes machine learning to the rescue with high computing facilities to predict air pollution. This paper investigates how effective some available prediction models are in predicting the Air Quality Index(AQI) values given some input data, based on the pollution and meteorological information in New Delhi, India. We perform regression analysis on the dataset, and our results show which meteorological factors affect the AQI values more and how useful the predictive models are to help in air quality forecasting.

**b) Inference drawn:** This paper underscores the significance of regression models available in the sklearn library for air quality forecasting. Notably, the findings reveal that the majority of these models exhibit a commendable accuracy of nearly 85%, with the Extra Trees regression model emerging as the most accurate predictor. This underscores the practical utility of these models in predicting air quality levels. Moreover, to further enhance predictive accuracy, integrating real-time and historical traffic data alongside weather data holds promise for refining AQI predictions. This synthesis of diverse datasets presents an avenue for advancing the precision and reliability of air quality forecasts, thereby bolstering efforts towards proactive pollution management strategies.

### **3. N. Srinivasa Gupta,Yashvi Mohta,Khyati Heda,Raahil Armaan,B. Valarmathi,and G. Arulkumaran: Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis**

**a) Abstract of the research paper:** The Air Quality Index (AQI) serves to inform the public about the health effects of local air pollution, particularly prevalent in Indian cities. Various mathematical formulas have been explored to calculate AQI, with data mining techniques emerging as a promising approach. This paper aims to identify the

most effective method for AQI prediction, leveraging techniques like SMOTE to optimize results. Three methods—Support Vector Regression (SVR), Random Forest Regression (RFR), and CatBoost Regression (CR)—are applied to cities like New Delhi, Bangalore, Kolkata, and Hyderabad. Results show that RFR consistently outperforms others in terms of RMSE values and accuracy, particularly after applying SMOTE. The paper's novelty lies in meticulously selecting the best regression models and employing dataset balancing through SMOTE, all supported by comprehensive documentation through graphs and metrics.

**b) Inference drawn:** Air pollution is a global problem; researchers from all around the world are working to discover a solution. To accurately forecast the AQI, machine learning techniques were investigated. The present study assessed the performance of the three best data mining models (Support Vector Regression, Random Forest Regressor, and Catboost Regressor) for predicting the accurate AQI data in some of India's most populous and polluted cities. For future work, there are plans to use satellite imagery and more extensive data to provide estimations for individual areas of a city as well. Another avenue to explore would be artificial intelligence (AI) to make the models more effective and innovative.

#### **4. Hongna He<sup>1</sup> and Fei Luo<sup>1</sup>: Study of LSTM Air Quality Index Prediction Based on Forecasting Timeliness**

**a) Abstract of the research paper:** The change of urban air quality is affected by pollutant emission, meteorological conditions and other factors, so air quality prediction is a multi-variable, nonlinear and time-series problem, which is difficult to be predicted by traditional methods. To solve this problem, a circular neural network based on Long Short Time Memory (LSTM) is proposed to predict Air Quality Index (AQI) by considering the pollution sources, meteorological conditions and time series. The transfer entropy is used to select the

meteorological factors that affect the strong change of AQI. Combined with the prediction time, the prediction accuracy of this algorithm in the future 0~48 hours within different forecast time is studied and compared with the traditional BP neural network and the Gated Recurrent Unit (GRU), and then the Root Mean Square Error (RMSE) was used for evaluation. Taking the measured data of hourly air quality index of Chengdu from January 1, 2018 to September 15, 2019 and the measured data of meteorological factors in the same period as experimental examples, the experimental results show that LSTM has better prediction accuracy and robustness than traditional neural networks in the aging of 0~48h forecasting, and has the advantages of temporary forecasting and short-term forecasting. At the same time, it is verified that GRU has no obvious advantage in air quality index prediction applications compared with LSTM.

**b) Inference drawn:** By using LSTM cycle neural network to predict AQI based on 9 parameters. Based on the characteristics of time series, we can solve the problem of multiple input time variables very well. This experiment based on the LSTM time series model to predict the Shanghai AQI has high precision, long range prediction and strong adaptive ability. It can approximate nonlinear mapping well. This mode can also be used in other multivariable input time series prediction problems, also has been widely used in life.

## **5. K. Kumar & B. P. Pande: Air pollution prediction with machine learning: a case study of Indian cities**

**a) Abstract of the research paper:** Air quality monitoring and prediction have become imperative in the face of increasing pollution from industrial, transport, and domestic activities. Utilizing machine learning techniques, this study investigates six years of air pollution data from 23 Indian cities. Through preprocessing and correlation analysis, key features are identified, and an exploratory data analysis reveals insights into pollutant trends. Despite a significant decrease in

pollutants during the pandemic year of 2020, predicting air quality remains crucial. Addressing data imbalance with resampling techniques, five machine learning models are employed, with Gaussian Naive Bayes achieving the highest accuracy. XGBoost emerges as the most effective model, demonstrating strong alignment between predicted and actual data.

**b) Inference drawn:** The paper proposes that their innovative approach harnessing machine learning models can offer precise and dependable predictions of air pollution levels in Delhi. By leveraging these models, the research aims not only to forecast pollution levels but also to delve into the intricate factors influencing them. This comprehensive analysis is anticipated to yield valuable insights into the dynamics of air pollution in the region, facilitating the design and implementation of effective policies and strategies for mitigation. By elucidating the underlying determinants of air pollution, the research endeavors to empower policymakers with the necessary tools and knowledge to enact targeted interventions, thereby contributing to the improvement of air quality and the overall well-being of the population. Through its interdisciplinary approach, combining machine learning techniques with environmental science, the paper endeavors to bridge the gap between data-driven insights and actionable measures, ultimately striving towards a cleaner and healthier environment for the residents of Delhi.

## **6. Manuel Méndez, Mercedes G. Merayo & Manuel Núñez: Machine learning algorithms to forecast air quality: a survey**

**a) Abstract of the research paper:** Air pollution is a risk factor for many diseases that can lead to death. Therefore, it is important to develop forecasting mechanisms that can be used by the authorities, so that they can anticipate measures when high concentrations of certain pollutants are expected in the near future. Machine Learning models, in particular, Deep Learning models, have been widely used to

forecast air quality. In this paper we present a comprehensive review of the main contributions in the field during the period 2011–2021. We have searched the main scientific publications databases and, after a careful selection, we have considered a total of 155 papers. The papers are classified in terms of geographical distribution, predicted values, predictor variables, evaluation metrics and Machine Learning model.

**b) Inference drawn:** It provides a comprehensive and updated overview of the state-of-the-art machine learning techniques for air quality forecasting. - It helps researchers and practitioners to understand the strengths and limitations of different models and to choose the most suitable ones for their problems. - It inspires new research ideas and innovations to improve the accuracy and reliability of air quality forecasting.

## **7. K. Nandini; G. Fathima: Urban Air Quality Analysis and Prediction Using Machine Learning**

**a) Abstract of the research paper:** Air pollution is one of the influential factors that can affect the quality of every living being in the environment. Monitoring the air pollution is a scathing issue. In this work, air pollutant prediction is done using Machine learning techniques. K Means algorithm is used for clustering and different classifiers such as Multinomial Logistic Regression and Decision Tree algorithms are used to analyze the results based on available data in the R programming language. The results obtained using classifiers are compared based on error rate and accuracy. The multinomial logistic regression model has given high accuracy compared to the decision tree model.

**b) Inference drawn:** It uses a large and diverse data set that covers various pollutants and their concentrations in urban areas. - It applies four different machine learning algorithms that can handle non-linear and complex relationships between the variables. - It achieves a high accuracy in predicting the air quality level, which can help in alerting

the authorities and the public about the air pollution situation.

**8. Raquel Espinosa a, José Palma a, Fernando Jiménez a, Joanna Kamińska b, Guido Sciavicco c, Estrella Lucena-Sánchez: A time series forecasting based multi-criteria methodology for air quality prediction.**

**a) Abstract of the research paper:** There is a very extensive literature on the design and test of models of environmental pollution, especially in the atmosphere. Current and recent models, however, are focused on explaining the causes and their temporal relationships, but do not explore, in full detail, the performances of pure forecasting models. We consider here three years of data that contain hourly nitrogen oxides concentrations in the air; exposure to high concentrations of these pollutants has been indicated as a potential cause of numerous respiratory, circulatory, and even nervous diseases. Nitrogen oxides concentrations are paired with meteorological and vehicle traffic data for each measure. We propose a methodology based on exactness and robustness criteria to compare different pollutant forecasting models and their characteristics. 1DCNN, GRU and LSTM deep learning models, along with Random Forest, Lasso Regression and Support Vector Machines regression models, are analyzed with different window sizes. As a result, our best models offer a 24-hours ahead, very reliable prediction of the concentration of pollutants in the air in the considered area, which can be used to plan, and implement, different kinds of interventions and measures to mitigate the effects on the population.

**b) Inference drawn:** This study presents several notable advantages. Foremost, it addresses a pressing global issue: the escalating levels of air pollution, stemming from diverse sources such as industrial activities and transportation. By introducing a robust methodology for evaluating and comparing deep learning models, this research not only sheds light on the complex dynamics of air quality but also makes substantial contributions to the fields of environmental science and



data analysis. Moreover, by delving into the nuances of deep learning techniques, the study offers insights that can inform policymakers and stakeholders in implementing more effective measures to combat air pollution and safeguard public health. Additionally, the systematic approach adopted in this research, coupled with its rigorous analysis of deep learning models, lays a solid foundation for future studies seeking to refine air quality prediction methodologies and develop targeted interventions to mitigate pollution levels. Thus, this study stands as a significant step forward in our collective efforts to address the urgent challenge of air pollution on a global scale.

## **9. Shengdong Du; Tianrui Li; Yan Yang; Shi-Jinn Horng: Deep Air Quality Forecasting Using Hybrid Deep Learning Framework**

**a) Abstract of the research paper:** Air quality forecasting has been regarded as the key problem of air pollution early warning and control management. In this article, we propose a novel deep learning model for air quality (mainly PM<sub>2.5</sub>) forecasting, which learns the spatial-temporal correlation features and interdependence of multivariate air quality related time series data by hybrid deep learning architecture. Due to the nonlinear and dynamic characteristics of multivariate air quality time series data, the base modules of our model include one-dimensional Convolutional Neural Networks (1D-CNNs) and Bi-directional Long Short-term Memory networks (Bi-LSTM). The former is to extract the local trend features and spatial correlation features, and the latter is to learn spatial-temporal dependencies. Then we design a jointly hybrid deep learning framework based on one-dimensional CNNs and Bi-LSTM for shared representation features learning of multivariate air quality related time series data. We conduct extensive experimental evaluations using two real-world datasets, and the results show that our model is capable of dealing with PM<sub>2.5</sub> air pollution forecasting with satisfied accuracy.

**b) Inference drawn:** The proposed air quality forecasting framework

(DAQFF) presents a novel and promising approach by combining the strengths of one-dimensional Convolutional Neural Networks (CNNs) and Bi-directional Long Short-Term Memory (Bi-LSTM) networks. This hybrid architecture capitalizes on the unique capabilities of each component, leveraging the CNN's proficiency in feature extraction from sequential data and the Bi-LSTM's ability to capture long-term dependencies and intricate relationships within multivariate air quality datasets. By integrating these two powerful techniques, DAQFF achieves a more comprehensive understanding of the complex correlations inherent in air quality data. This enhanced capacity enables the model to provide more accurate and reliable forecasts compared to conventional methods. Thus, DAQFF stands as a promising advancement in air quality forecasting, offering significant advantages in prediction accuracy and performance.

#### **10. Varsha Hable-Khandekar; Pravin Srinath: Machine Learning Techniques for Air Quality Forecasting and Study on Real-Time Air Quality Monitoring**

**c) Abstract of the research paper:** Air Pollution has become a major, serious problem worldwide. Because of its close relation to human health, it has gained a lot of attention from many researchers. People are becoming more cautious about better ways of monitoring air quality information and it has become important to protect human health from serious health problems caused by air pollution. Many researchers are working on real-time air quality monitoring and forecasting for getting accurate results which will help in implementing various government policies related to the environment or air pollution and for taking crucial decisions. There are many recent advancements in the air quality forecasting and monitoring techniques. Most of the techniques are Machine Learning (ML) based as it has become a popular analysis tool because of its various distinctive features. This paper summarizes air quality forecasting

models as well as realtime monitoring tools and techniques based on real-time and historical data. It has discussed the merits and demerits of every methodology used for air quality forecasting and monitoring used in recent research along with their comparative analysis, limitations, and challenges. This paper will be useful to understand current status, past work done and future research questions which need to be addressed.

**d) Inference drawn:** The dynamics of air pollution are multifaceted, with meteorological conditions, traffic density, industrial emissions, and the combustion of fossil fuels among the pivotal factors contributing to its complexity. However, among the myriad of pollutants, Particulate Matter (PM 2.5) emerges as a particularly concerning agent due to its minute size and profound impact on human health. When elevated levels of PM 2.5 permeate the air, they pose serious health risks, penetrating deep into the respiratory system and exacerbating conditions such as asthma, cardiovascular diseases, and even contributing to premature death. Thus, understanding and mitigating the sources and concentrations of PM 2.5 have become paramount in efforts to safeguard public health and mitigate the adverse effects of air pollution.

## **2.2 Inferences drawn**

The papers collectively highlight the use of machine learning techniques to predict Air Quality Index (AQI) in an effort to address global air pollution. Techniques such as Support Vector Regression, Random Forest Regressor, Catboost Regressor, and LSTM cycle neural networks have been employed to forecast AQI with high precision. The integration of real-time and historical traffic data alongside weather data is suggested to refine AQI predictions. Future work includes the use of satellite imagery and more extensive data for area-specific estimations, and the exploration of artificial intelligence for more effective and innovative models. The research aims to provide insights into the dynamics of air pollution, facilitating the

design of effective mitigation policies and strategies, and contributing to the improvement of air quality and overall well-being of the population. The papers also provide a comprehensive overview of state-of-the-art machine learning techniques for air quality forecasting, aiding researchers and practitioners in understanding the strengths and limitations of different models.

## 2.3 Comparison with the existing systems

Other Systems	Our System
Only predicts day-wise AQI	More focus on hour-wise AQI
May contain unbalanced data, leading to less accuracy	First balances the data, followed by the model creation
Does not focus on visualization of data	Visualizes data via interactive dashboard

# Chapter 3: Requirement Gathering for the Proposed System

## 3.1. Introduction to Requirement Gathering

This project aims to develop a comprehensive system for urban air quality forecasting and monitoring. To achieve accurate predictions, the system will utilize historical and real-time air quality data. However, this data may be imbalanced, with certain pollution levels being less frequent than others. To address this, we will employ SMOTE (Synthetic Minority Oversampling Technique) to balance the data before training the forecasting models. This ensures the models are not biased towards the majority class and can accurately predict even less frequent pollution events.

This project aims to develop a comprehensive system for urban air quality forecasting and monitoring. Here are some key points:

- **Data Acquisition:** The system will collect real-time and historical air quality data from various sources, which includes the concentrations of air pollutants like PM2.5, NO2, NOx etc.
- **Data Preprocessing and Balancing:** The collected data may be imbalanced, with certain pollution levels being less frequent. To address this, we will employ SMOTE (Synthetic Minority Oversampling Technique) to create synthetic data points for under-represented pollution classes. This ensures the models are trained on a balanced dataset and can accurately predict even infrequent pollution events.
- **Forecasting Models:** Machine learning algorithms, such as Random Forest, Support Vector Machines and CatBoost Regressors will be employed to analyze the preprocessed data and generate forecasts for future air quality conditions.
- **Dissemination and Visualization:** The predicted air quality information will be disseminated through various channels, including mobile applications, web portals, and public displays. User-friendly visualizations will be developed to communicate complex air quality data in a clear and actionable way.
- **Evaluation and Improvement:** The system's performance will be

continuously monitored and evaluated. We will use metrics like Root Mean Square Error (RMSE) to assess the accuracy of the forecasts. Based on the evaluation results, the system will be iteratively improved and refined.

## **3.2. Functional Requirements**

**1. Data Collection and Monitoring:** The system must continuously collect real-time air quality data, including information on various pollutants like PM2.5, PM10, O3, CO, SO2, NO2, and more. It should interface with air quality monitoring sensor networks and stations to gather data, ensuring data accuracy through validation processes.

**2. Data Analysis and Modeling:** The system should be capable of calculating and reporting the Air Quality Index (AQI) based on the collected data. It must also have the capability to forecast pollutant concentrations for both short-term and long-term periods, using statistical analysis and modeling techniques. Additionally, the system must incorporate meteorological data into its forecasting models.

**3. Alerts and Notifications:** To inform the public and relevant authorities, the system should trigger alerts and notifications when air quality exceeds predefined threshold values. It should also provide special alerts during severe air quality events, such as smog, wildfires, or industrial accidents.

**4. User Interface:** The system's user interface should be user-friendly, with web and mobile interfaces offering clear access to air quality information. Interactive maps and data visualizations should be available to enhance user understanding, and historical data access should be provided for research and trend analysis.

**5. Geospatial Information:** The system must provide location-specific air quality data, allowing users to search for information in their areas of interest. It should also

incorporate Geographic Information Systems (GIS) for spatial analysis, enhancing the relevance of data.

**6. Forecasting:** The system must provide both short-term (daily) and long-term (weekly, monthly) air quality forecasts. It should use historical data, real-time measurements, and meteorological information to make these forecasts as accurate as possible.

**7. Data Sharing and APIs:** The system should facilitate data sharing with government agencies, researchers, and the public. It must provide well-documented APIs for developers to access air quality data for integration into external applications.

**8. Data Archiving and Storage:** The system must maintain historical air quality data for reference and analysis. It should include data backup and redundancy mechanisms to ensure data integrity, even in the event of hardware failures or data corruption.

**9. Customization:** The system should allow users to customize their experience, including their preferred units (e.g.,  $\mu\text{g}/\text{m}^3$  or ppm) and alert thresholds. It should also support multiple languages for broader accessibility.

**10. Compliance and Reporting:** The system must adhere to local, national, and international air quality monitoring and reporting regulations. Additionally, it should enable the export of air quality data reports for compliance and research purposes.

These functional requirements collectively ensure that air quality monitoring and forecasting systems provide accurate, timely, and reliable information to support public health, environmental protection, and research efforts effectively.

### 3.3. Non-Functional Requirements

**1. Performance:** The system must exhibit responsive behavior, with defined acceptable response times for data retrieval and user interactions. It should be able to scale efficiently as data volume and user load increase, and the system's throughput, or the number of concurrent users and data transactions it can handle, should be well-defined.

**2. Reliability:** Ensuring the system's reliability is paramount. This involves specifying the expected uptime and defining maintenance processes and downtime schedules. The system should exhibit fault tolerance, allowing it to continue providing critical functions in the presence of hardware or software failures. Additionally, data integrity measures must be in place to guarantee that collected and stored data remain accurate and untampered.

**3. Security:** Stringent security measures are vital. The system should protect sensitive air quality data from unauthorized access and data breaches, including robust data encryption for both transmission and storage. Secure user authentication and authorization mechanisms are necessary to control user access.

**4. Usability:** Usability requirements focus on the user experience. The system should offer a user-friendly interface with intuitive navigation and clear information presentation. Accessibility is also crucial, requiring compliance with accessibility standards to ensure that individuals with disabilities can use the system. User training or documentation should be available to assist users in understanding and effectively navigating the system.

**5. Compatibility:** The system should be compatible with various web browsers, ensuring a consistent experience. Additionally, it should be optimized for mobile



devices, accommodating different screen sizes and resolutions.

**6. Scalability:** Scalability considerations encompass database scalability to accommodate growing data volumes, as well as load balancing mechanisms to distribute user requests evenly and prevent performance bottlenecks.

**7. Compliance:** The system must comply with relevant air quality monitoring regulations and standards. This includes defining data retention policies in accordance with legal requirements to ensure that data is stored and managed appropriately.

**8. Interoperability:** To foster interoperability, the system should support industry-standard data formats for seamless interaction with other systems and tools. Well-documented APIs should also be provided to facilitate integration with external applications and services.

**9. Maintainability:** The system must be designed for easy maintenance and updates, allowing for enhancements without causing significant downtime. Modularity in the system's architecture facilitates easier troubleshooting and maintenance.

**10. Performance Monitoring:** Implementing performance monitoring tools and metrics is essential to track system health, detect anomalies, and trigger alerts when necessary. Monitoring system health and performance metrics such as server response time and data retrieval speed helps maintain optimal functionality.

### 3.4. Hardware, Software, Technology and Tools Utilised

#### A) Hardware Used:

Processor: Intel i3 or AMD equivalent

Disk Space: 4GB

RAM: 8GB

GPU: NVIDIA GPU

#### B) Software and Algorithms:

**Google Colaboratory & GitHub** - Google Colab allows you to use and share notebooks with others without having to download, install, or run anything.

**Ensemble Learning:** Ensemble Learning are machine learning methods which combine several base models in order to produce one optimal model. We are

**SVM:** Support Vector Regression (SVR) is a type of support vector machine (SVM) that is used for regression tasks. It tries to find a function that best predicts the continuous output value for a given input value.

**Tableau** - Tableau is an interactive data visualization software product developed by Microsoft with a primary focus on business intelligence.

**Matplotlib , seaborn, pandas , numpy** -Matplotlib is a widely-used plotting library for creating static, animated, and interactive visualizations in Python.

**Seaborn** is a statistical data visualization library built on top of Matplotlib, designed for creating attractive and informative visualizations.

**Pandas** is a powerful library for data manipulation and analysis, providing data structures and functions to efficiently work with structured data.

# Chapter 4: Proposed Design

## 4.1. Block Diagram of the proposed system

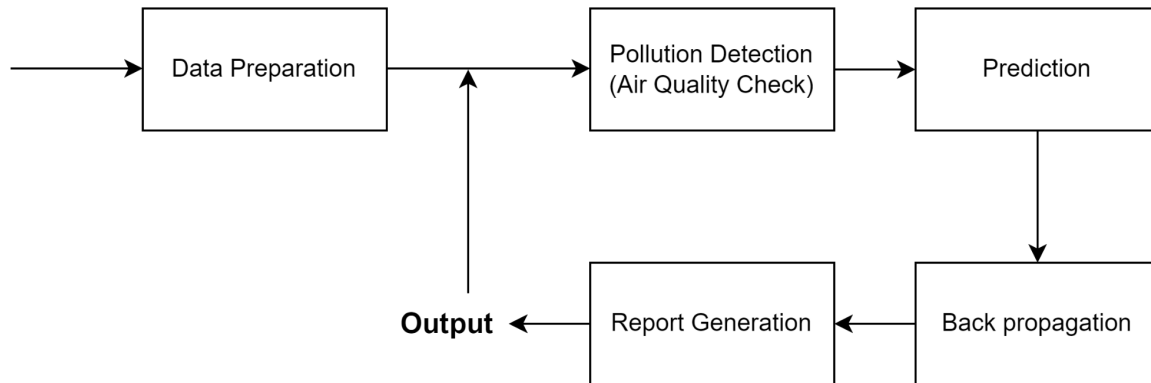


Fig 4.1: Block Diagram

### 1. Data Preparation:

**Data Collection:** This stage gathers air quality data from various sources. This can include:

- **Government Agencies:** Environmental protection agencies often maintain networks of air quality monitoring stations that collect real-time data on various pollutants.
- **Private Organizations:** Research institutions or private companies may operate their own air quality monitoring stations, providing additional data points.
- **Citizen Science Projects:** Crowdsourced air quality data collected by individuals using low-cost sensors can contribute valuable information.

**Data Preprocessing:** Once collected, the data needs to be cleaned and prepared for machine learning. This may involve:

- **Handling Missing Values:** Techniques like interpolation or deletion can address missing data points.
- **Feature Scaling:** Standardizing the data ensures all features contribute equally during model training.
- **Outlier Detection and Removal:** Identifying and removing outliers can

improve model performance.

## 2. Pollution Detection (Air Quality Check):

- This stage analyzes the prepared data to assess the current air quality.
- It often involves comparing the pollutant concentrations to predefined thresholds set by regulatory bodies like the Environmental Protection Agency (EPA).
- These thresholds define different AQI levels (e.g., Good, Moderate, Unhealthy for Sensitive Groups).
- By comparing the data to these thresholds, the system can determine if the current air quality is acceptable or if pollution levels are a cause for concern.

## 3. Prediction:

This is the core of the system, where machine learning models are used to forecast future air quality levels.

- **Model Training:** The models are trained on historical air quality data. The training data includes pollutant concentrations, meteorological data (temperature, wind speed, humidity), and potentially other relevant factors like traffic data or industrial emissions.
- **Model Selection:** Different machine learning algorithms are explored to identify the one that performs best for the specific task. Common options include:
  - **Linear Regression:** This is a simple and interpretable model that can capture linear relationships between pollutants and other factors.
  - **Support Vector Regression (SVR):** This technique is effective for handling complex nonlinear relationships.
  - **Random Forest:** This ensemble method combines multiple decision trees, leading to robust and accurate predictions.

- **Deep Learning Models (e.g., Long Short-Term Memory Networks - LSTMs):** These powerful architectures can capture complex temporal dependencies in the data, leading to more accurate long-term forecasts.

#### **4. Report Generation:**

This stage generates reports based on the system's findings.

The reports typically include:

- **Current Air Quality:** This section presents the current AQI level or pollutant concentrations.
- **Air Quality Forecast:** This section provides predictions for future air quality levels for a specific timeframe (e.g., hourly, daily).
- **Visualizations:** Charts and graphs can be used to present the information clearly and concisely.

#### **5. Backpropagation (Optional):**

- While not a separate block in real-world applications, backpropagation deserves mention.
- It is a training algorithm used in artificial neural networks.
- During training, the model's predictions are compared to actual values. Backpropagation allows the model to adjust its internal parameters to minimize the error between predictions and actual values.
- This process is iterative, gradually improving the model's performance over time.

#### **Benefits of using Machine Learning for Air Quality Forecasting:**

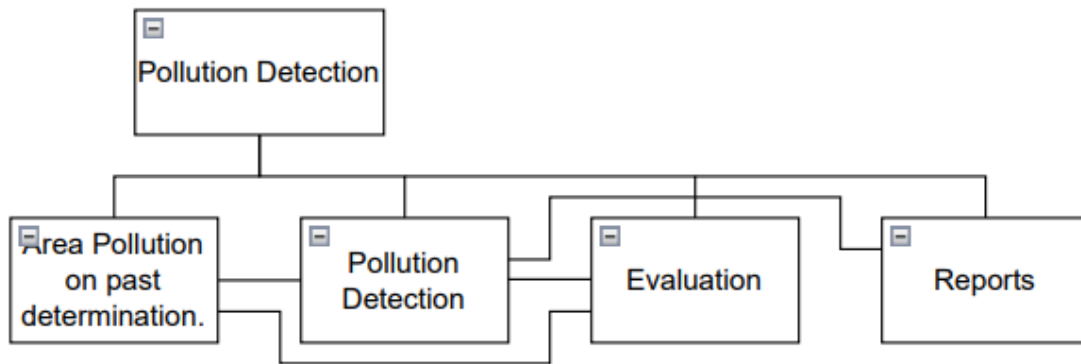
- **Improved Accuracy:** Machine learning models can learn complex patterns from historical data, leading to more accurate air quality forecasts compared to traditional

statistical methods.

- **Timely Decision-Making:** Accurate forecasts provide valuable information for individuals and authorities to make informed decisions.
- Individuals can adjust their activities to minimize exposure to harmful pollutants (e.g., avoiding outdoor exercise on high pollution days).
- Authorities can implement temporary measures to reduce pollution levels (e.g., traffic restrictions for high-polluting vehicles, temporary closures of industrial facilities).
- **Enhanced Public Health:** By enabling proactive measures, air quality forecasting can help reduce respiratory problems and other health issues associated with air pollution.
- **Environmental Sustainability:** Improved air quality forecasts can guide long-term environmental policies and strategies for sustainable development in urban areas.

Machine learning offers a powerful tool for forecasting air pollution in urban environments. By leveraging historical and real-time data, this approach can generate accurate predictions, empowering individuals and authorities to make informed decisions that safeguard public health and promote environmental well-being.

## 4.2. Modular diagram of the system



**Fig 4.2: Modular Diagram**

The block diagram of a system that uses machine learning to forecast air pollution. Let's walk through the blocks, although the terminology used may differ slightly from the explanation you provided:

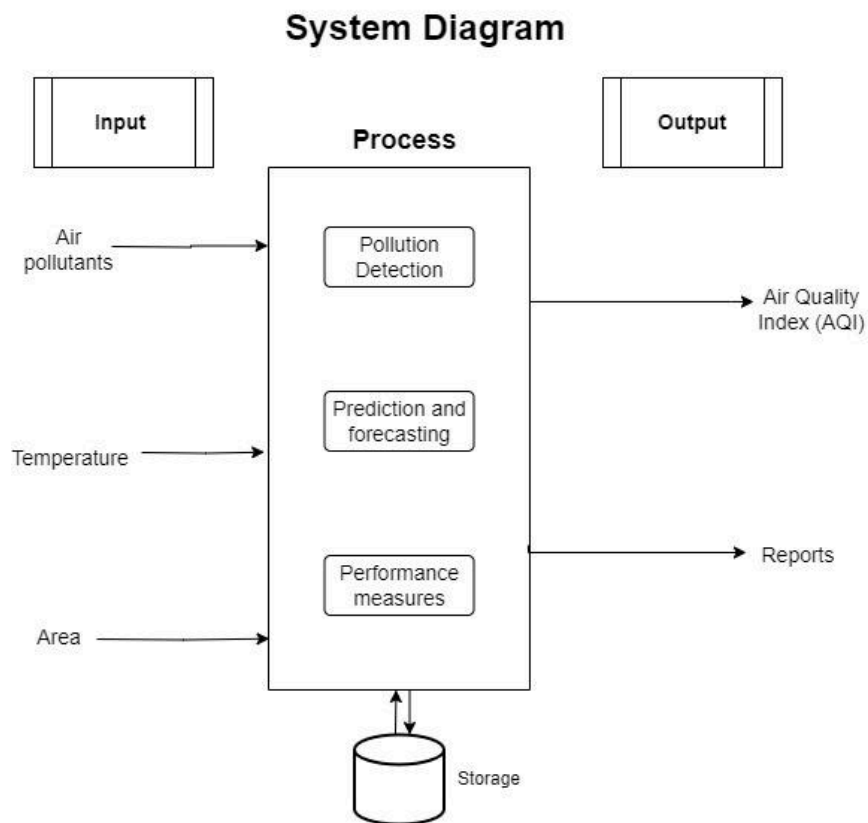
- **Area Pollution Determination:** This block represents the collection of data from various sources. This data likely includes historical and current air quality data on pollutants like PM2.5, PM10, NO2, SO2, CO, and O3 from urban environments, which tend to have denser networks of air quality monitoring stations.
- **Pollution Detection (Air Quality Check):** This block likely compares the collected data to predefined thresholds to determine if pollution levels are acceptable. The thresholds might be set by regulatory bodies to define air quality levels (e.g., Good, Moderate, Unhealthy for Sensitive Groups).
- **Prediction:** This block uses machine learning models to forecast future air quality levels. The model is likely trained on the historical data collected in the Area Pollution Determination block.
- **Evaluation Reports:** This block generates reports based on the findings from the Pollution Detection and Prediction blocks. These reports could include information on current air quality levels and forecasts for future air quality.

The system you described focuses on using machine learning to forecast air pollution in urban environments. This is important because urban environments tend to have more air quality monitoring stations than rural areas, which provides more data to train the machine learning models.

By accurately forecasting air pollution levels, this system can aid in decision-making and planning for pollution control measures, which can improve public health and environmental conditions.



### 4.3. System Design



**Fig 4.3: System Diagram**

The diagram which is using machine learning to forecast air quality in urban environments. Here's a breakdown of the system's components, along with some additional details based on the image:

#### Data Input

- **Air Pollutants:** This block represents the collection of air quality data on various pollutants like PM2.5, PM10, NO2, SO2, CO, and O3.
- **Temperature:** This block signifies the inclusion of temperature data, which can influence air pollution patterns.
- **Area:** This likely refers to incorporating data specific to the urban environment under study. This could include factors like building density, traffic patterns, or proximity to industrial areas.

## Process

- **Prediction and Forecasting:** This block represents the core function of the system – using machine learning models to predict future air quality levels. The model is likely trained on the historical data from the Data Input blocks.
- **Performance Measures:** This block indicates that the system evaluates the performance of the machine learning model. Common metrics include accuracy, root mean squared error (RMSE), or mean absolute error (MAE). These metrics help assess how well the model's predictions align with actual air quality data.

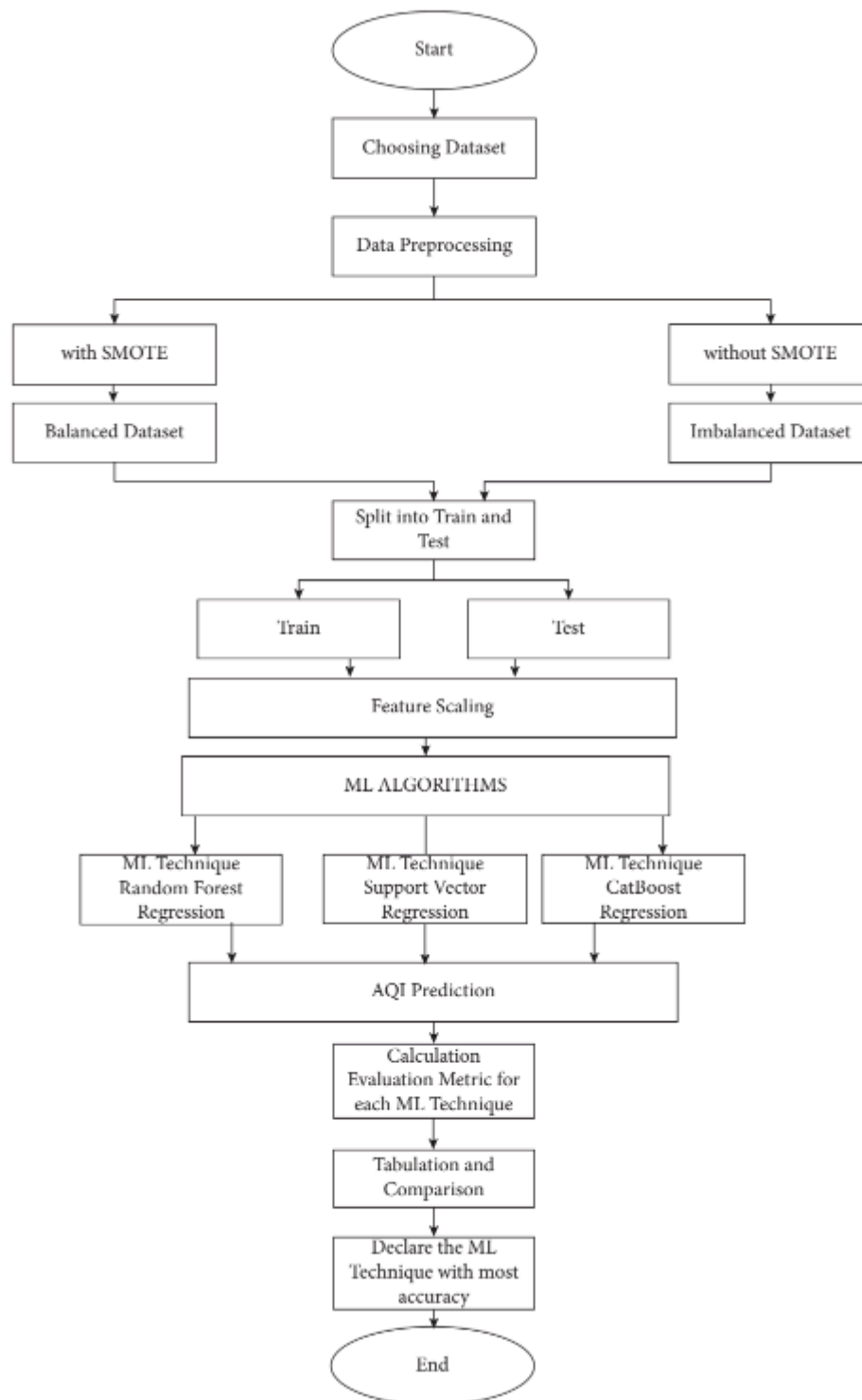
## Output

- **Air Quality Index (AQI):** This represents the predicted Air Quality Index, a standardized metric for reporting air quality.
- **Reports:** This block refers to generating reports that likely include predicted AQI values, performance measures, and potentially visualizations of the data or forecasts.

## Additional Notes:

- While the block labeled "Backpropagation" is not typically a separate step in real-world systems, it represents a training algorithm used in artificial neural networks (a type of machine learning model). It's included here to depict how the model is continuously refined for better accuracy.
- The focus on urban environments is important because these areas have distinct pollution sources and patterns compared to rural areas. Additionally, they typically have denser networks of air quality monitoring stations, providing the system with more data for training and improving the model's performance.

#### 4.4 Flowchart for the proposed system



**Fig 4.4 : Flowchart of Model**

## Data Preprocessing

- **Choosing Dataset:** This block refers to selecting the data to be used for training the machine learning model. There may be options to choose between a balanced dataset (where the amount of data for each pollution level is similar) or an imbalanced dataset (where there is more data for some levels than others). Techniques like SMOTE (Synthetic Minority Oversampling Technique) can be used to create a balanced dataset from imbalanced data.
- **Data Preprocessing (with/without SMOTE):** This block refers to cleaning and preparing the chosen data for machine learning. This may involve:
  - **Handling Missing Values:** Techniques like interpolation or deletion can address missing data points.
  - **Feature Scaling:** Standardizing the data ensures all features contribute equally during model training.
  - **Outlier Detection and Removal:** Identifying and removing outliers can improve model performance.

## Machine Learning Algorithms

- This block represents the core function of the system – using machine learning models to predict future air quality levels. The text mentions three specific Machine Learning (ML) techniques: Random Forest Regression, Support Vector Regression (SVR), and CatBoost Regression. These are all examples of supervised learning algorithms, where the model learns from labeled data (i.e., data where the desired outcome - air quality level - is already known) to make predictions on new, unseen data.

## Evaluation

- **Calculation of Evaluation Metric:** This block refers to evaluating the performance of each machine learning model on a separate test dataset. This dataset is not used to train the model but to assess its generalizability - how well it performs on unseen data. Common metrics used for evaluating

regression models include accuracy, root mean squared error (RMSE), or mean absolute error (MAE). These metrics help assess how well the model's predictions align with actual air quality data.

## Model Selection

- **Declare the ML Technique with most accuracy:** This block refers to selecting the machine learning model that achieved the best performance on the test dataset based on the evaluation metrics. This model will then be used to make predictions on new data.

## Reporting

- **Tabulation and Comparison:** This block likely refers to creating a table or report that compares the performance of the different machine learning models on the test dataset. This can be helpful for understanding the strengths and weaknesses of each model.
- **AQI Prediction:** This block refers to using the chosen machine learning model to predict future air quality levels, likely expressed as AQI (Air Quality Index).

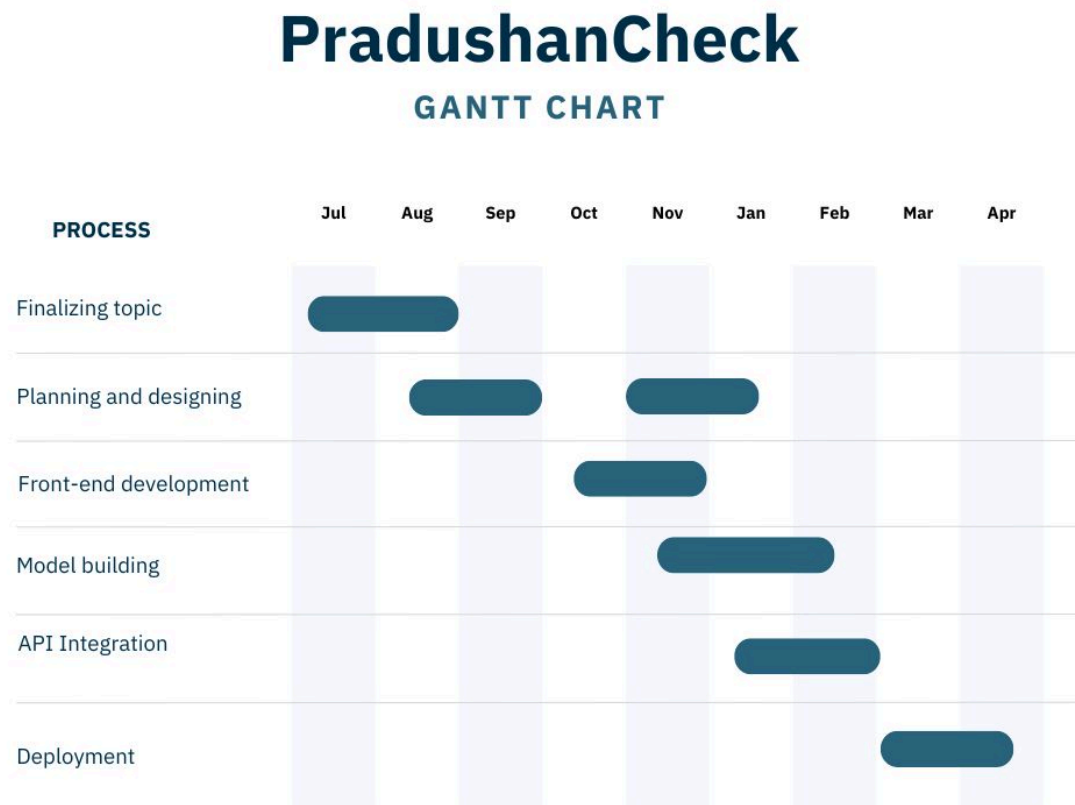
## Additional Notes:

- The concept of "Backpropagation" is not explicitly included in the blocks, but it's a common technique used to train artificial neural networks (one type of machine learning model). It's included in some depictions of this process to show how the model is continuously refined for better accuracy.
- The focus on urban environments is important because these areas have distinct pollution sources and patterns compared to rural areas. Additionally, they typically have denser networks of air quality monitoring stations, providing the system with more data for training and improving the model's performance.

Overall, the diagram depicts a system for using machine learning to forecast air quality in urban environments. By leveraging historical and current data, the system can predict future air quality levels, empowering individuals and authorities to make informed decisions for safeguarding public health and the environment.

## 4.5 Project Scheduling & Tracking using Time line / Gantt Chart

The Gantt chart of our project where we worked for the whole semester to create this model is shown in a timeline pattern. It is the most important part to think and design the planning of your topic and so we planned our work like the gantt chart shown.



**Fig 4.5 : Gantt chart**

# Chapter 5: Implementation of the Proposed System

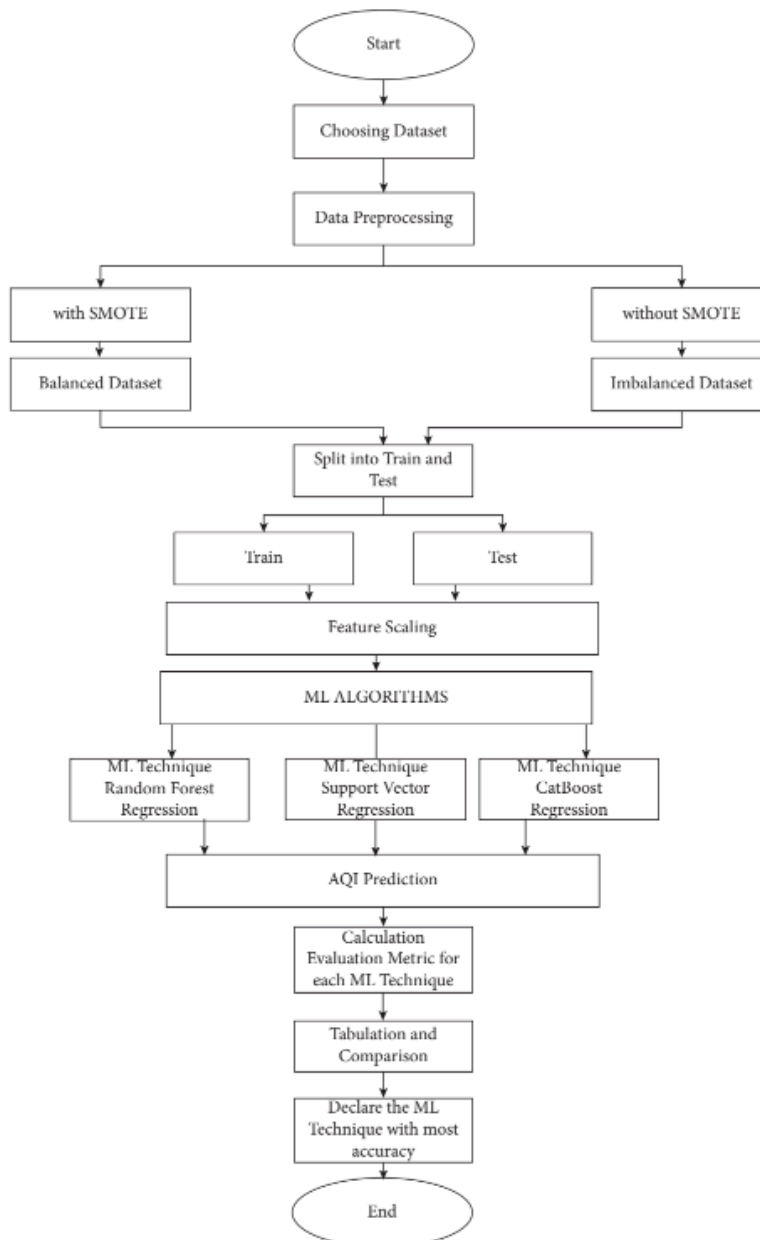
## 5.1. Methodology employed for development

In this project, we aim to predict air quality index (AQI) as the target variable, utilizing various input features including PM2.5 , PM10, oxides of the Nitrogen. Data collection involves gathering information from relevant sources such as air quality monitoring stations, satellite images, and weather stations. Preprocessing the data includes handling missing values, normalization, and feature engineering to enhance the model's performance. Different preprocessing methods, including SMOTE and Binning were implemented to address class imbalances within the dataset, resulting in a more balanced dataset and improved model performance. Following this, we conducted a comparative analysis of the predictive performance of three distinct algorithms—Random Forest Regression, Support Vector Regression, and CatBoost Regression—in forecasting air quality indices (AQI). Additionally, we visually represented the dataset through an interactive dashboard using Tableau software. Moreover, we integrated the 'Air Quality Programmatic API' to provide up-to-date information on air pollution levels on our website. Notably, this API facilitates the retrieval of air pollution data for cities worldwide. Our observations emphasize the effectiveness of PradushanCheck in furnishing real-time insights conducive to efficient pollution management and enhancing public awareness regarding air quality.

## 5.2. Algorithms and Flowcharts for the respective modules developed

After Research, we aim to predict air quality index (AQI) as the target variable, utilizing various input features including PM2.5 , PM10, oxides of the Nitrogen. Data collection involves gathering information from relevant sources such as air quality monitoring stations, satellite images, and weather stations. Preprocessing the data includes handling missing values, normalization, and feature engineering to enhance the model's performance. Different preprocessing methods, including SMOTE and Binning were implemented to address class imbalances within the dataset, resulting

in a more balanced dataset and improved model performance. Following this, we conducted a comparative analysis of the predictive performance of three distinct algorithms—Random Forest Regression, Support Vector Regression, and CatBoost Regression—in forecasting air quality indices (AQI). Additionally, we visually represented the dataset through an interactive dashboard using Tableau software. Moreover, we integrated the 'Air Quality Programmatic API' to provide up-to-date information on air pollution levels on our website. Notably, this API facilitates the retrieval of air pollution data for cities worldwide. Our observations emphasize the effectiveness of PradushanCheck in furnishing real-time insights conducive to efficient pollution management and enhancing public awareness regarding air quality. We have compared 3 algorithms from the flowchart below.





### **a) Random Forest Regressor :**

The Random Forest Regressor operates as an ensemble learning technique rooted in the decision tree paradigm, wherein multiple decision trees are created and amalgamated to yield more precise outcomes. Assessment of the model's efficacy on the testing set involves the application of suitable metrics, such as Mean Squared Error and R-squared, to gauge its predictive performance accurately. Hyperparameters may be adjusted as deemed necessary to refine the model's accuracy. Following model evaluation, the trained Random Forest Regressor is utilized for AQI prediction based on prevailing environmental conditions. Subsequent monitoring of the model's predictions over time serves to validate its aptitude in reflecting fluctuations in air quality conditions effectively.

### **b) Support Vector Regressor :**

Support Vector Regressor (SVR) is an example of a supervised learning algorithm, which finds the hyperplane that maximizes the margin between the data and the hyperplane and fits the data the best for regression tasks. With its capacity to manage non-linear connections and adjust to different data distributions, SVR can provide precise AQI monitoring and offer information for well-informed decision-making about the management of air quality in urban settings. It seeks to fit the data inside the tolerance margin while minimizing error.

### **c) Catboost Regressor :**

CatBoost Regressor is a machine learning algorithm specifically designed for regression tasks, making it suitable for predicting continuous target variables such as the Air Quality Index (AQI). CatBoost Regressor operates by building an ensemble of decision trees and combining their predictions to produce an accurate forecast of AQI levels.

By training the CatBoost regression model on relevant datasets, incorporating parameters specific to categorical boosting, one can achieve accurate predictions of

AQI values. This approach allows the model to adapt to changing conditions and improve its forecasting capabilities, contributing to effective air quality monitoring and management in urban environments.

CatBoost Regressor employs several optimizations to enhance the training process and improve predictive performance. It makes use of a cutting-edge feature importance calculation approach that sheds light on the relative significance of various information in forecasting the target variable.

### **5.3. Datasets source and utilization:**

We utilized a dataset sourced from Kaggle, encompassing air quality data from four major cities in India: New Delhi, Bengaluru, Kolkata, and Hyderabad, to forecast the Air Quality Index (AQI) for each city. The dataset comprises diverse pollution indicators, including nitrogen oxides (NO, NO<sub>2</sub>) and sulfur compounds (SO, SO<sub>2</sub>), among others. Each entry in the dataset includes the AQI value for a particular day, categorized into distinct AQI Buckets such as Poor, Satisfactory, Moderate or Good.

The following are the columns in the dataset :

- ID
- College Name
- City
- Date
- PM 2.5
- PM 10
- NO
- NO<sub>2</sub>
- NO<sub>x</sub>
- NH<sub>3</sub>
- CO
- SO<sub>2</sub>
- O<sub>3</sub>
- Benzene
- Toluene

- AQI
- AQI Bucket

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	AQI	AQI_Bucket			
2	4604 Bengaluru	#####	31.44	70.46	3.03	15.85	10.27	27.73	1.22	1.94	17.98	2.41	17.84	60	Satisfactory			
3	4607 Bengaluru	#####	12.63	43.64	1.46	6.81	4.75	12.63	1.48	2.05	41.66	1.12	2.07	51	Satisfactory			
4	4610 Bengaluru	#####	28.22	153.3	5.8	21.5	21.78	27.79	2.09	2.69	34.32	2.08	4.5	61	Satisfactory			
5	4611 Bengaluru	#####	42.42	156.84	7.25	29.94	31.78	21.94	1.56	2.23	31.35	1.82	4.65	130	Moderate			
6	4616 Bengaluru	#####	21.99	39.86	7.08	16.44	19.51	41.96	1.73	2.95	9.98	1.52	2.38	103	Moderate			
7	4617 Bengaluru	#####	13.89	31.44	6.84	12.14	15.35	23.93	1.72	2.5	4.56	0.74	1.48	74	Satisfactory			
8	4620 Bengaluru	#####	19.66	36.84	6.47	16.37	20.87	24.04	1.35	2.83	4.09	1.18	2.17	75	Satisfactory			
9	4621 Bengaluru	#####	20.35	33.97	7.76	20.64	24.75	26.98	1.36	2.59	7.77	1.02	1.9	85	Satisfactory			
10	4622 Bengaluru	#####	34.39	36.29	8.38	28.8	32.28	32.75	2.48	3.76	14.63	1.32	3.17	141	Moderate			
11	4623 Bengaluru	#####	43.91	43.65	11.74	29.33	32.78	55.4	1.52	3.44	14.8	1.53	3.59	90	Satisfactory			
12	4624 Bengaluru	#####	44.14	112.78	7.05	26.64	27.06	32.33	2.18	4.3	25.57	1.69	3.36	126	Moderate			
13	4625 Bengaluru	#####	44.94	114.34	8.47	28.1	29.37	32.75	2.3	4.7	29.1	1.56	2.38	147	Moderate			
14	4626 Bengaluru	#####	29.35	75.79	5.72	21.21	21.4	19.08	1.55	4.55	29.03	1.01	1.15	87	Satisfactory			
15	4627 Bengaluru	#####	15.34	53.7	8.46	22.22	24.6	21.4	1.13	4.45	8.84	1.11	2.53	56	Satisfactory			
16	4628 Bengaluru	#####	14.18	55.71	12.9	25.86	27.06	21.81	1.63	5.14	9.89	0.97	2.15	88	Satisfactory			
17	4629 Bengaluru	#####	13.36	33.82	8.24	24.04	17.46	22	1.7	11.07	8.15	0.73	1.66	67	Satisfactory			
18	4630 Bengaluru	#####	19.86	72.72	11.73	33.56	26.85	37.08	1.47	8.32	9.51	0.98	2.6	74	Satisfactory			
19	4631 Bengaluru	#####	28.35	102.15	13.06	32.95	24.78	52.45	2.03	6.08	21.38	1	3.46	105	Moderate			
20	4632 Bengaluru	#####	25.04	89.26	9.79	33.02	22.96	51.77	1.72	5.43	26.44	0.77	1.47	103	Moderate			
21	4634 Bengaluru	#####	41.16	85.25	10.04	34.42	23.91	46.07	1.88	5.24	27.26	0.88	1.7	65	Satisfactory			
22	4635 Bengaluru	#####	46.33	82.86	17.24	44.41	34.36	45.43	1.81	5.81	19.1	1.01	1.92	102	Moderate			
23	4636 Bengaluru	#####	48.01	78.58	16.65	39.78	31.61	46.15	1.84	4.86	6.32	1.16	1.49	94	Satisfactory			
24	4637 Bengaluru	#####	35.48	58.04	25.04	35.58	35.53	51.46	1.92	5.27	3.36	0.99	1.46	86	Satisfactory			
25	4638 Bengaluru	#####	32.95	79.07	23.41	41.05	36.25	26.38	2.07	5.94	2.3	1.43	1.63	89	Satisfactory			
26	4639 Bengaluru	#####	50.19	73.43	18.73	57.62	31.41	20.93	1.78	5.34	2.33	1.29	1.39	99	Satisfactory			
27	4640 Bengaluru	#####	34.57	64.09	15.68	42.88	37.39	12.64	1.68	4.79	2.16	0.91	1.07	92	Satisfactory			
28	4641 Bengaluru	#####	51.06	56.71	12.51	27.66	31.45	11.08	1.22	4.28	1.87	1.30	1.6	85	Satisfactory			

Fig 5.4: Bangalore CSV

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	AQI	AQI_Bucket		
2	10229 Delhi	#####	313.22	607.98	69.16	36.39	110.59	33.85	15.2	9.25	41.68	14.36	24.86	472	Severe		
3	10230 Delhi	#####	186.18	269.55	62.09	32.87	88.14	31.83	9.54	6.65	29.97	10.55	20.09	454	Severe		
4	10231 Delhi	#####	87.18	131.9	25.73	30.31	47.95	69.55	10.61	2.65	19.71	3.91	10.23	143	Moderate		
5	10232 Delhi	#####	151.84	241.84	25.01	36.91	48.62	130.36	11.54	4.63	25.36	4.26	9.71	319	Very Poor		
6	10233 Delhi	#####	146.6	219.13	14.01	34.92	38.25	122.88	9.2	3.33	23.2	2.8	6.21	325	Very Poor		
7	10234 Delhi	#####	149.58	252.1	17.21	37.84	42.46	134.97	9.44	3.66	26.83	3.63	7.35	318	Very Poor		
8	10235 Delhi	#####	217.87	376.51	26.99	40.15	52.41	134.82	9.78	5.82	28.96	4.93	9.42	353	Very Poor		
9	10236 Delhi	#####	229.9	360.95	23.34	43.16	51.21	138.13	11.01	3.31	30.51	5.8	11.4	383	Very Poor		
10	10237 Delhi	#####	201.66	397.43	19.18	38.56	45.6	140.6	11.09	3.48	32.94	5.25	11.12	375	Very Poor		
11	10238 Delhi	#####	221.02	361.74	24.79	46.39	55.19	134.06	9.7	5.91	34.12	4.87	9.44	376	Very Poor		
12	10239 Delhi	#####	205.41	393.2	28.46	47.29	57.88	131.1	10.98	5.54	50.37	5.93	10.59	379	Very Poor		
13	10240 Delhi	#####	212.41	345.63	24.77	44.71	54.66	148.51	9.3	5.17	40.08	6.2	10.68	375	Very Poor		
14	10241 Delhi	#####	197.61	301.04	34.28	45.88	62.95	145.48	11.52	5.95	39.54	5.18	9.5	366	Very Poor		
15	10242 Delhi	#####	164.39	227.38	20.78	41.17	48.09	166.7	10.32	4.85	42.05	4.57	10.72	353	Very Poor		
16	10243 Delhi	#####	166.19	283.93	47.91	56.7	79.73	124.64	10.81	6.94	43.25	6.98	20.55	340	Very Poor		
17	10244 Delhi	#####	174.98	309.99	42.33	49.41	70.6	136.4	11.46	4.69	34.7	7.75	15.83	356	Very Poor		
18	10245 Delhi	#####	196.46	356.2	42.18	53.04	80.18	125.49	8.87	5.13	25.77	9.92	17.59	360	Very Poor		
19	10246 Delhi	#####	201.51	359.75	45.51	53.77	75.83	120.45	8.99	5.43	23.22	9	20.95	370	Very Poor		
20	10247 Delhi	#####	183.35	305.13	27.32	39.88	55.67	148.49	9.02	4.21	25.34	5.57	12.93	362	Very Poor		
21	10248 Delhi	#####	165.63	257.04	16.47	34.23	37.68	146.31	8.03	4.11	22.69	3.42	6.76	340	Very Poor		
22	10249 Delhi	#####	159.54	235.27	12.27	22.56	33.02	63.23	9.01	7.05	18.56	4.1	8.26	338	Very Poor		
23	10250 Delhi	#####	143.68	183.89	14.75	21.82	34.88	39.65	10.58	8.64	6.94	4.15	9.54	332	Very Poor		
24	10251 Delhi	#####	128.73	136.07	11.99	17.35	29.46	39.98	8.74	6.77	10.91	2.51	7.64	254	Poor		
25	10252 Delhi	#####	164.98	200.02	10.38	15.77	25.49	38.77	9.39	8.89	9.35	2.6	5.5	324	Very Poor		
26	10253 Delhi	#####	148.59	203.81	11.82	20.83	28.29	45.62	11.44	5.96	16.65	3.37	6.98	333	Very Poor		
27	10254 Delhi	#####	129.34	137.86	12.46	25.66	29.48	60.96	10.19	6.78	18.18	3	5.87	292	Poor		
28	10255 Delhi	#####	154.2	217.2	14.41	20.20	24.62	40.05	10.26	6.42	17.11	2.31	6.21	318	Very Poor		

Fig 5.5: New Delhi CSV

# **Chapter 6: Testing of the Proposed System**

## **6.1. Introduction to Testing**

Software testing is the sequence of activities that happen during software testing. By employing a sane software testing life cycle, an organization ends up with a quality strategy more likely to produce better results. Project Testing Phase means a group of activities designated for investigating and examining progress of a given project to provide stakeholders with information about actual levels of performance and quality of the project. It is an attempt to get an independent view of the project to allow stakeholders to evaluate and understand potential risks of project failure or mismatch. The purpose of the testing phase is to evaluate and test declared requirements, features, and expectations regarding the project prior to its delivery in order to ensure the project matches initial requirements stated in specification documents.

## **6.2. Types of tests Considered**

### **A. Pre testing phase**

Alpha testing in the context of air quality monitoring refers to the initial phase of testing conducted on a newly developed air quality monitoring system or device. This phase involves testing the system in a controlled environment by the developers or a select group of users before it is released to a wider audience.

During alpha testing for air quality monitoring systems, developers typically focus on the following aspects:

1. **Functionality:** Testing the basic functions of the monitoring device or system, such as data collection, sensor accuracy, and real-time monitoring capabilities.
2. **Performance:** Assessing the performance of the system under various conditions, including different levels of air pollution, temperature, and humidity.
3. **Reliability:** Checking the reliability of the system by monitoring its ability to provide accurate and consistent readings over an extended period.

4. User Interface: Evaluating the user interface for ease of use, clarity of information, and accessibility of features.
5. Compatibility: Ensuring compatibility with different devices, operating systems, and data management platforms.
6. Data Accuracy: Verifying the accuracy of the data collected by comparing it with reference standards or other established monitoring systems.
7. Fault Tolerance: Testing the system's ability to handle errors or malfunctions gracefully, including sensor failures or communication issues.
8. Feedback Collection: Gathering feedback from alpha testers regarding their experience with the system, any issues encountered, and suggestions for improvement.

## **B. Beta-Testing Phase**

Beta testing in the context of air quality monitoring involves deploying the monitoring system or device to a larger group of users or customers for testing in real-world environments. Unlike alpha testing, which is conducted by developers or a select group of users, beta testing involves a broader audience and aims to gather feedback on the product's performance, usability, and reliability under normal operating conditions.

1. Selection of Beta Testers: Developers select a diverse group of beta testers, including individuals, organizations, or communities interested in air quality monitoring. These testers may include environmental agencies, researchers, businesses, or concerned citizens.
2. Deployment of Monitoring Devices: Beta testers receive the air quality monitoring devices or access to the monitoring system. They install the devices in various locations, such as urban areas, industrial sites, residential neighborhoods, or rural regions, depending on the objectives of the testing.
3. Testing Period: The beta testing phase typically lasts for a defined period, during which testers use the monitoring devices or system to collect data on air quality. This period allows testers to assess the performance of the system over time and under different environmental conditions.

4. **Feedback Collection:** Beta testers provide feedback on their experience with the monitoring system, including its accuracy, reliability, ease of use, and any issues encountered during operation. Developers may collect feedback through surveys, interviews, online forums, or dedicated feedback channels.
5. **Bug Reporting:** Testers report any bugs, glitches, or technical issues they encounter while using the monitoring system. Developers use this information to identify and address software or hardware problems to improve the overall performance and reliability of the product.

### **6.3. Various test case scenarios considered**

When considering test case scenarios for air quality monitoring, it's essential to cover a wide range of potential conditions and scenarios to ensure the system's accuracy, reliability, and usability. Here are various test case scenarios that may be considered:

#### **1. Baseline Testing:**

- Verify that the monitoring system provides accurate readings in an environment with known air quality conditions.
- Test the system's ability to detect normal background levels of pollutants, such as carbon dioxide (CO<sub>2</sub>), particulate matter (PM<sub>2.5</sub> and PM<sub>10</sub>), nitrogen dioxide (NO<sub>2</sub>), sulfur dioxide (SO<sub>2</sub>), and ozone (O<sub>3</sub>).

#### **2. Pollution Spike Simulation:**

- Simulate sudden increases in pollution levels to test the system's responsiveness.
- Assess how quickly the system detects and responds to elevated levels of pollutants, such as during traffic congestion or industrial emissions.

#### **3. Environmental Conditions:**

- Test the system's performance under different environmental conditions, including variations in temperature, humidity, and atmospheric pressure.

- Evaluate the system's ability to maintain accuracy and reliability in extreme weather conditions, such as heatwaves, cold snaps, or high winds.

#### 4. Indoor vs. Outdoor Monitoring:

- Compare the system's performance in indoor and outdoor environments.
- Test the accuracy of indoor air quality measurements, considering factors like ventilation, occupancy, and indoor pollutant sources (e.g., cooking, cleaning, smoking).

#### 5. Spatial Variability:

- Assess the spatial variability of air quality within a defined area.
- Deploy multiple monitoring units in different locations to compare readings and identify spatial patterns in pollutant concentrations.

#### 6. Long-Term Monitoring:

- Conduct extended monitoring over days, weeks, or months to assess the system's stability and reliability over time.
- Evaluate the system's ability to maintain accuracy and consistency during prolonged operation.

#### 7. Sensor Calibration:

- Test the accuracy of sensor calibration by comparing the monitoring system's readings with reference instruments or standardized calibration gases.
- Verify that the system can be calibrated accurately and reliably to maintain measurement accuracy over time.

#### 8. Data Transmission and Connectivity:

- Evaluate the system's ability to transmit data reliably to a central server or cloud-based platform.
- Test the system's connectivity under different network conditions, including Wi-Fi, cellular, and Ethernet connections.

#### 9. Battery Life and Power Management:

- Assess the system's battery life under typical operating conditions.
- Test power-saving features and automatic shutdown mechanisms to optimize battery usage without sacrificing performance.

#### 10. User Interface and Alerts:

- Evaluate the user interface for clarity, ease of use, and accessibility.
- Test the effectiveness of alert notifications for informing users about changes in air quality levels and potential health risk.



# Chapter 7: Results and Discussions

## 7.1. Screenshot of Use Interface(UI) for the system:

### Know Air Quality Index(AQI)

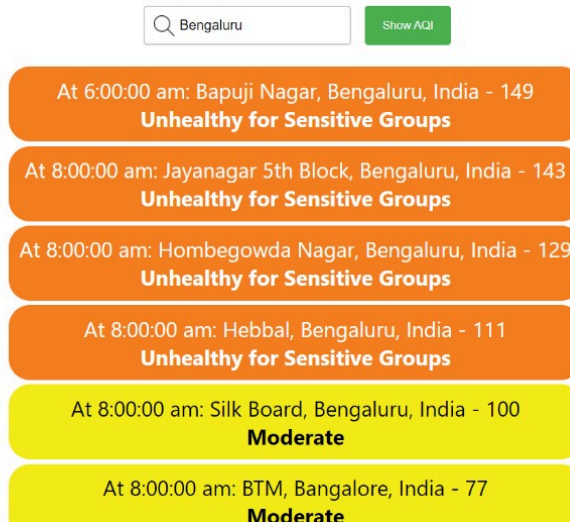


Fig 7.1: Screenshot for AQI at Bengaluru

### Know Air Quality Index(AQI)

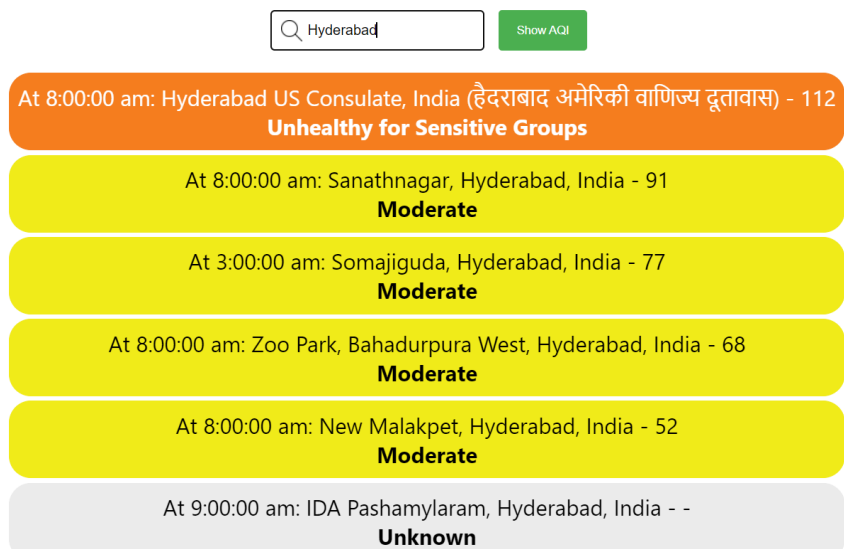
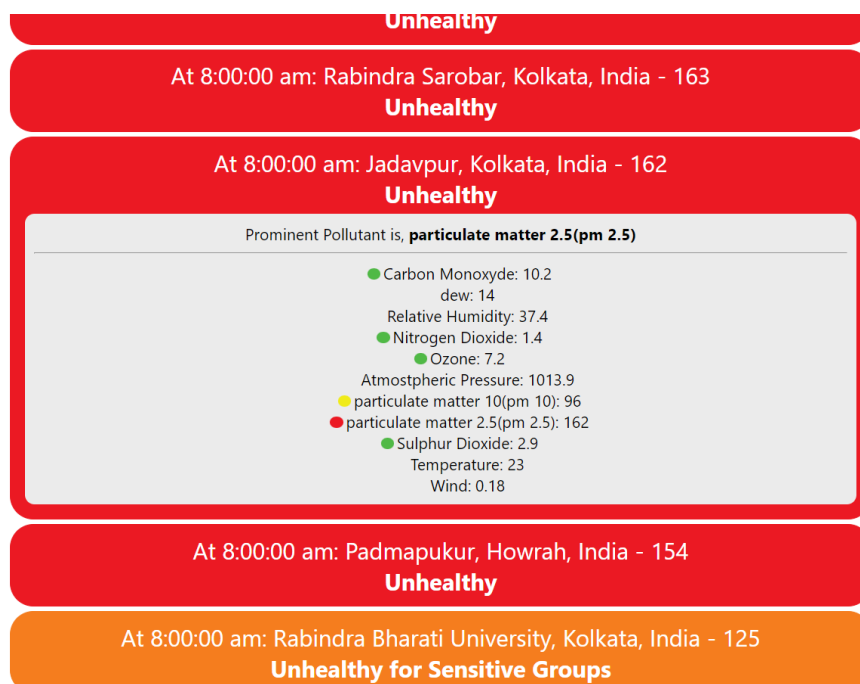
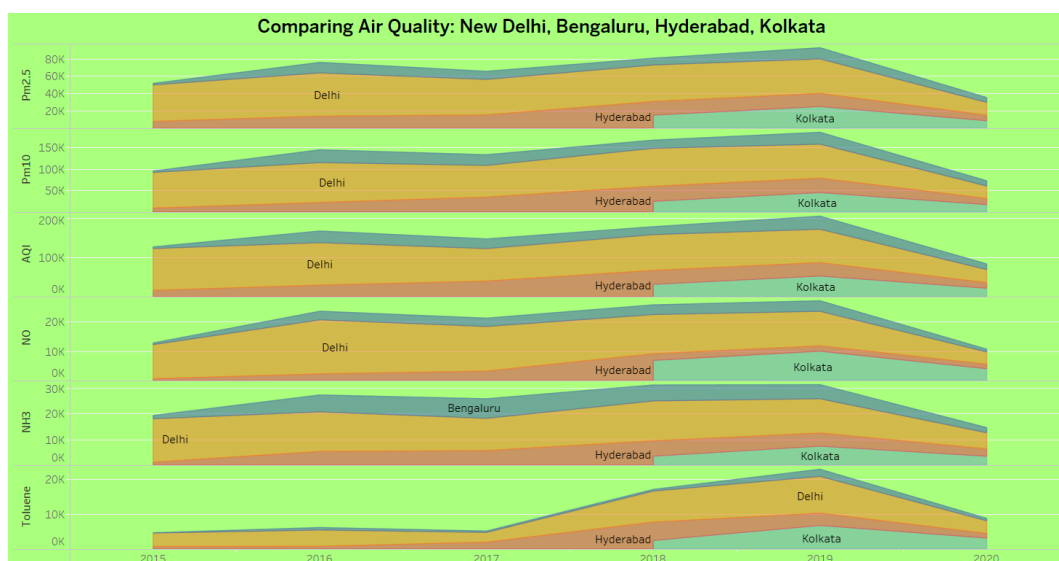


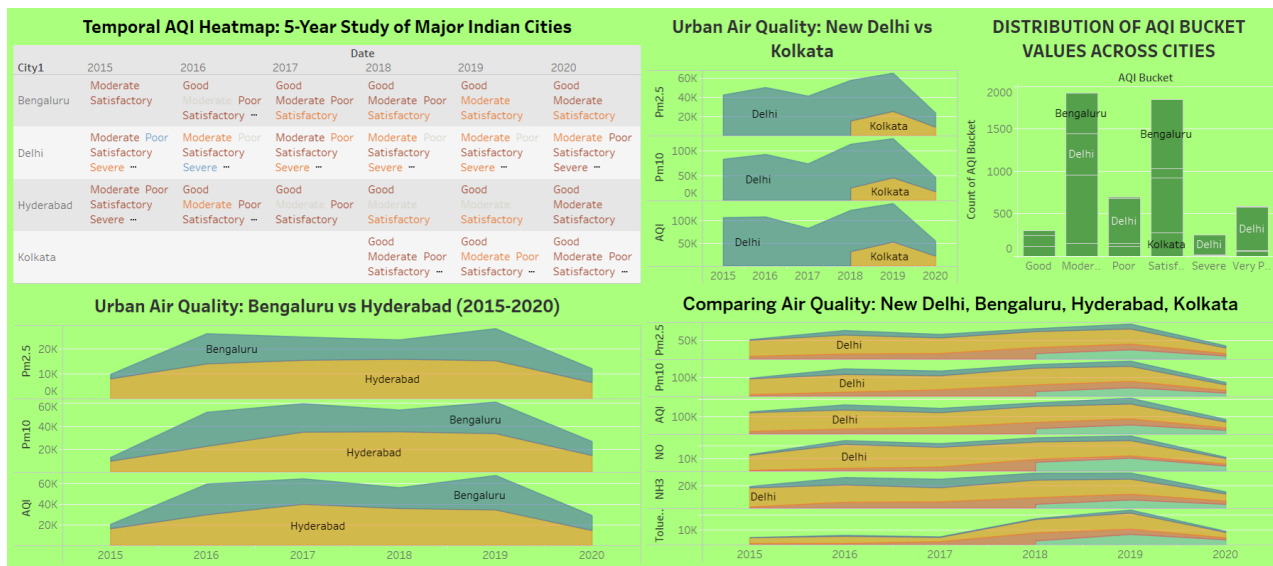
Fig 7.2: Screenshot for AQI at Hyderabad



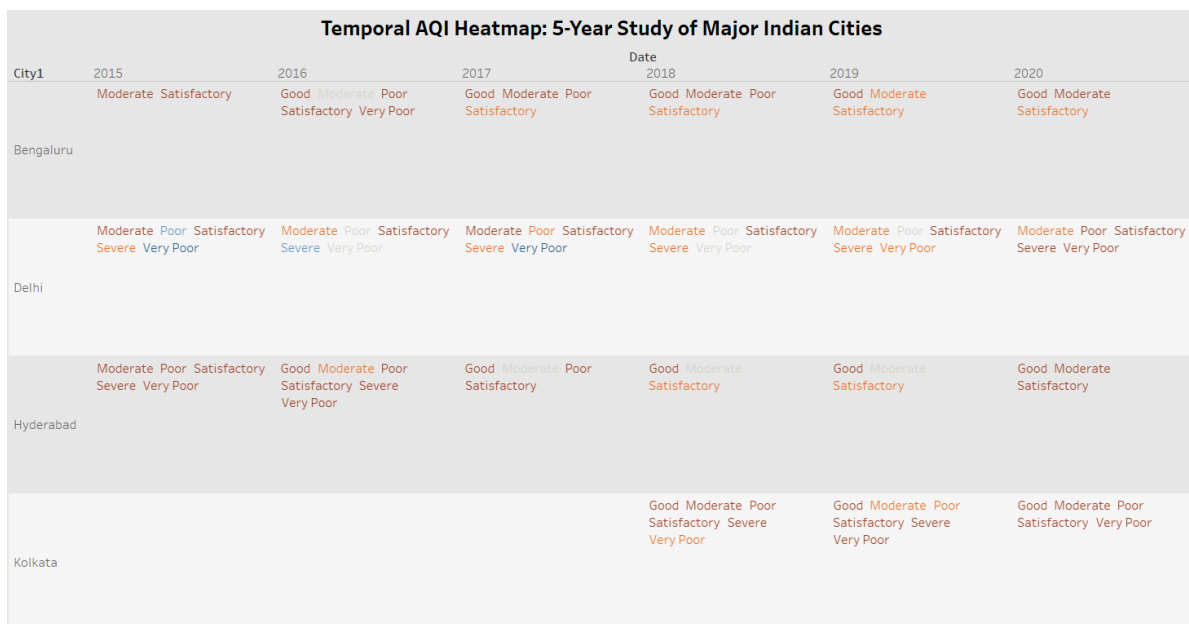
**Fig 7.3: Screenshot for AQI at Kolkata**



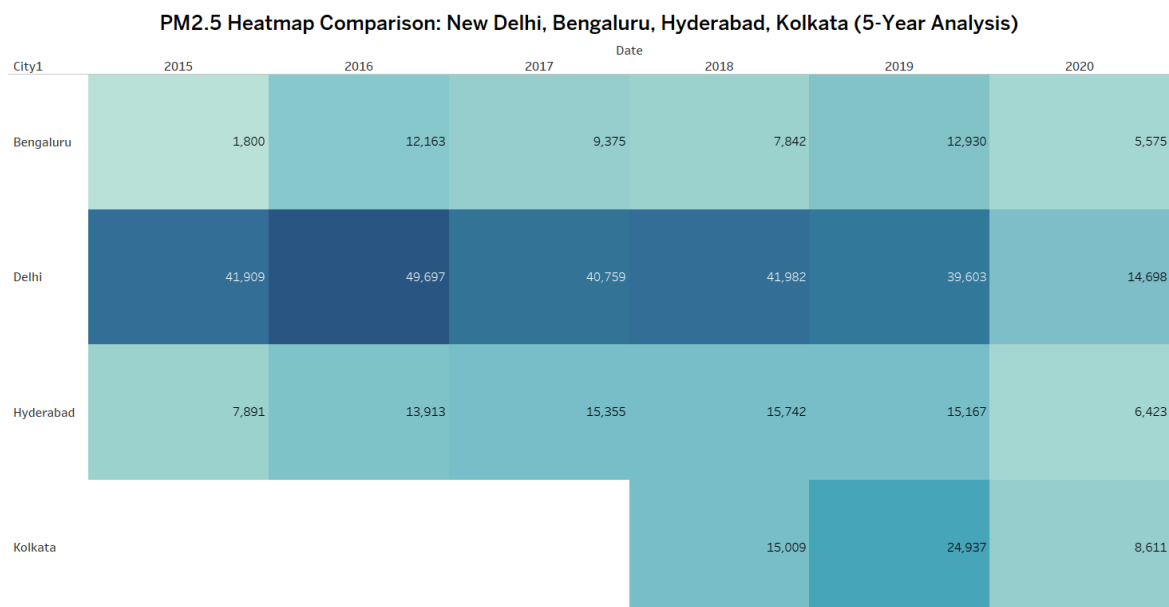
**Fig 7.4: Comparison of Air Quality of New Delhi, Bengaluru,, Kolkata and Hyderabad**



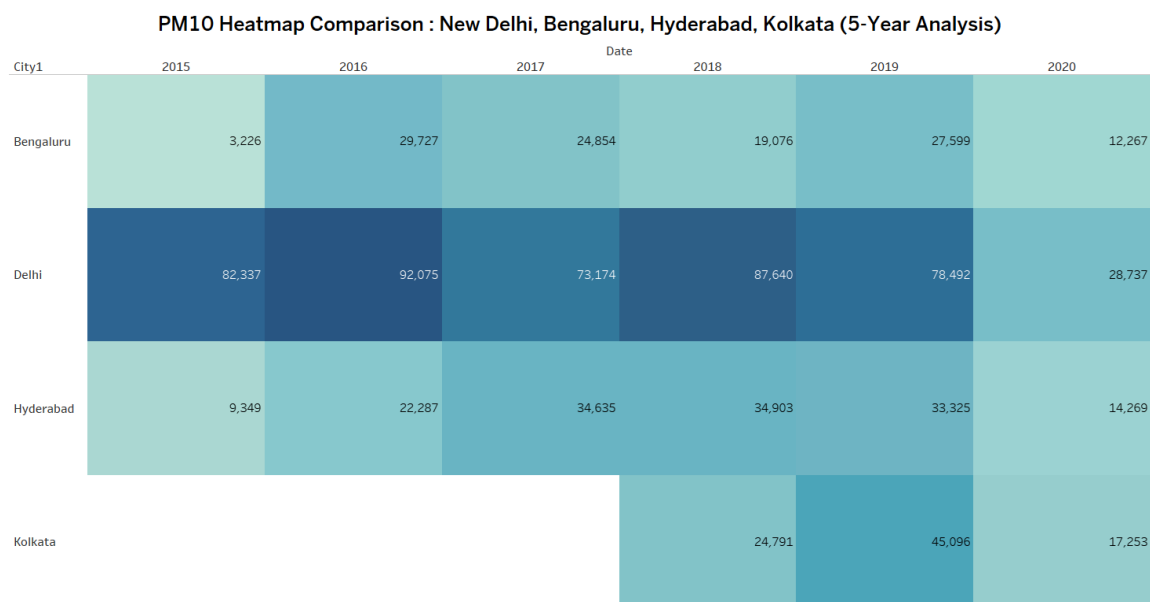
**Fig 7.5: Visualization of data via dashboard**



**Fig 7.6: Temporal AQI Heatmap**



**Fig 7.7: PM 2.5 Heatmap Comparison of the four major cities**



**Fig 7.8: PM10 Heatmap Comparison of the four major cities**

## 7.2. Performance Evaluation Measures:

1. **R2 Score (Coefficient of Determination):** R2 Score measures the proportion of the variance in the dependent variable (target) that is predictable from the independent variables (features). It provides an indication of the goodness of fit of the model. R2 score ranges from 0 to 1, where 1 indicates a perfect fit and 0 indicates that the model does not explain any of the variance in the target variable better than the mean of the target values.

- Formula:

$$R2 = 1 - (\text{Sum of Squared Residuals} / \text{Total Sum of Squares})$$

where,

- Sum of Squared Residuals: Sum of squared differences between actual and predicted values.
- Total Sum of Squares: Sum of squared differences between actual values and mean of actual values.

- Interpretation: Higher R2 score indicates a better fit of the model to the data. A score closer to 1 implies that the model explains a large proportion of the variance in the target variable.

2. **Root Mean Squared Error (RMSE):** RMSE measures the average deviation of the predicted values from the actual values. It gives an estimate of how much error the model makes in its predictions, with a lower value indicating better performance.

- Formula:

$$RMSE = \sqrt{1/n * \sum(\text{predicted\_i} - \text{actual\_i})^2}$$

where,

- n is the number of data points,
- predicted\_i is the predicted value,
- actual\_i is the actual value.

- Interpretation: RMSE is in the same units as the target variable, and lower values indicate better model performance.

**3. Mean Squared Error (MSE):** MSE is similar to RMSE but lacks the square root operation, so it penalizes large errors more than smaller ones. It measures the average squared difference between the predicted values and the actual values.

- Formula:

$$\text{MSE} = 1/n * \sum (\text{predicted\_i} - \text{actual\_i})^2$$

where,

- n is the number of data points,
- predicted\_i is the predicted value,
- actual\_i is the actual value.

- Interpretation: Like RMSE, lower values of MSE indicate better model performance. However, since MSE is not in the same units as the target variable, it may be harder to interpret directly.

**4. Mean Absolute Error (MAE):** MAE measures the average absolute difference between the predicted values and the actual values. It provides an indication of the average magnitude of errors made by the model.

- Formula:

$$\text{MAE} = 1/n * \sum |\text{predicted\_i} - \text{actual\_i}|$$

where,

- n is the number of data points,
- predicted\_i is the predicted value,
- actual\_i is the actual value.

- Interpretation: MAE is in the same units as the target variable and provides a straightforward interpretation of the average magnitude of errors made by the model. Lower values indicate better model performance.

### 7.3. Input Parameters/Features considered

We utilized a dataset sourced from Kaggle, encompassing air quality data from four major cities in India: New Delhi, Bengaluru, Kolkata, and Hyderabad, to forecast the Air Quality Index (AQI) for each city. The dataset comprises diverse pollution indicators, including nitrogen oxides (NO, NO<sub>2</sub>) and sulfur compounds (SO, SO<sub>2</sub>), among others. Each entry in the dataset includes the AQI value for a particular day, categorized into distinct AQI Buckets such as Poor, Satisfactory, Moderate or Good.

The following are the columns in the dataset :

- ID
- College Name
- City
- Date
- PM 2.5
- PM 10
- NO
- NO<sub>2</sub>
- NO<sub>x</sub>
- NH<sub>3</sub>
- CO
- SO<sub>2</sub>
- O<sub>3</sub>
- Benzene
- Toluene
- AQI
- AQI Bucket

#### 7.4. Comparison of Results with Existing System

Other System	Our System
Only predicts day-wise AQI	More focus on hour-wise AQI
May contain unbalanced data, leading to less accuracy	First balances the data, followed by the model creation
Does not focus on visualization of data	Visualizes data via interactive dashboard

#### 7.5. Inference Drawn:

This research prioritizes predicting hourly air quality (AQI) with high accuracy. It acknowledges that existing methods focus on daily predictions and may have unbalanced data, leading to inaccuracies. To address this, the approach involves balancing the data before creating a model. While data visualization isn't a primary focus, the final product will be an interactive dashboard for users to explore the AQI data.



# Chapter 8: Conclusion

## 8.1. Limitations

- a) Relies on accurate and sufficient air quality data, which may be limited in some regions.
- b) Forecasting models may not perfectly predict complex air quality dynamics.
- c) The system can't directly control pollution sources, relying on external actions for mitigation.

## 8.2. Conclusion

Leveraging the power of CatBoost on our AQI monitoring website opens doors to exciting possibilities beyond basic prediction. By integrating predicted AQI values with weather forecasts and historical data, we can create a sophisticated air quality information hub. Users won't just see a single number - they'll have access to a dynamic picture of how air quality might evolve throughout the day, week, or even longer. Imagine being able to plan outdoor exercise routines around predicted good air quality days or proactively reschedule picnics based on potential spikes in pollution. Furthermore, the website can become an interactive platform. Real-time sensor data can be fed back into the CatBoost model, creating a self-learning loop. This allows the model to constantly adapt to changing environmental conditions, ensuring the predictions remain razor-sharp. Ultimately, this user-centric approach empowers individuals to make informed health decisions. By providing a clear understanding of air quality trends, the website becomes a valuable tool for safeguarding public health and promoting a more proactive approach to well-being.

### 8.3. Future Scope:

- a) **Improved Data Integration:** Integrating air quality data with other environmental datasets (e.g., weather, traffic) will lead to a more comprehensive understanding of pollution dynamics.
- b) **Machine Learning Applications:** Machine learning can further enhance air quality forecasting, allowing for more accurate predictions and earlier warnings.
- c) **Real-time Decision Support Systems:** Combining air quality data with real-time traffic management, energy production, and industrial emission control systems could lead to dynamic pollution mitigation strategies.

# References

[1] Analyzed Research Papers:-

[https://docs.google.com/document/d/1egy-c7UJd\\_XG-ZWv1WJ9uycaseoE6W2NtPNbWoNb9p0/edit?usp=sharing](https://docs.google.com/document/d/1egy-c7UJd_XG-ZWv1WJ9uycaseoE6W2NtPNbWoNb9p0/edit?usp=sharing)

[2] Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis N. Srinivasa Gupta,<sup>1</sup>Yashvi Mohta,<sup>2</sup>Khyati Heda,<sup>2</sup>Raahil Armaan,<sup>2</sup>B. Valarmathi,<sup>2</sup>and G. Arulkumaran, Hindawi Journal of Environmental and Public Health, Volume 2023, Article ID 4916267

[3] A time series forecasting based multi-criteria methodology for air quality prediction, Raquel Espinosa, José Palma, Fernando Jiménez , Joanna Kamińska , Guido Sciavicco, Estrella Lucena-Sánchez.

[4] Deep Air Quality Forecasting Using Hybrid Deep Learning Framework, IEEE Transactions on Knowledge and Data Engineering ( Volume: 33, Issue: 6, 01 June 2021)

[5] Urban Air Quality Prediction Using Regression Analysis, Soubhik Mahanta; T. Ramakrishnu; Rajat Raj Jha; Niraj Tailor, TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)

[6] Air Quality Prediction using Machine Learning Algorithms –A Review, Tanisha Madan; Shreddha Sagar; Deepali Virmani, 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)

[7] 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.

[8] 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.

[9] M. Bansal, “Air quality index prediction of Delhi using LSTM,” Int. J. Emerg.

Trends Technol. Comput. Sci, vol. 8, pp. 59–68, 2019.

[10] 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*.

[11] N. Srinivasa Gupta, Yashvi Mohta, Khyati Heda, Raahil Armaan, Valarmathi and G. Arulkumaran, “Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis” *Journal of Environmental and Public Health*

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] U. Mahalingam, K. Elangovan, H. Dobhal, C. Valliappa, S. Shrestha, and G. Kedam, “A machine learning model for air quality prediction for smart cities,” in

Proceedings of the 2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET), pp. 452–457, Chennai, India, March 2019.



[19] S. Hansun and M. Bonar Kristanda, “AQI measurement and prediction using B-wema method,” *International Journal of Engineering Research and Technology*, vol. 12, pp. 1621–1625, 2019.

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


# Appendix

## 1. Paper I Details

### a. Paper published

5th International **Conference** on Data Science and Applications : Submission (223) has been created.  

**External** **Inbox x**

 **Microsoft CMT** <email@msr-cmt.org> Wed, Apr 3, 12:40 AM (8 days ago) ☆ ↶ ⋮  
to me ▾

Hello,

The following submission has been created.

Track Name: ICDSA2024

Paper ID: 223

Paper Title: PradushanCheck - Air Quality Monitoring

Abstract:  
PradushanCheck offers a comprehensive approach to air quality monitoring, addressing the critical issue of urban air pollution. Utilizing datasets from major cities including New Delhi, Hyderabad, Kolkata, and Bangalore, we employed advanced techniques such as SMOTE and binning for data preprocessing. Subsequently, we compared the performance of three algorithms—Random Forest Regression, Support Vector Regression, and CatBoost Regression—in predicting air quality indices (AQI) for these cities. Furthermore, we visualized the data on an interactive dashboard using Tableau and integrated the real-time API 'Air Quality Programmatic API'

Authors:  
- [2020.ashutosh.mishra@ves.ac.in](mailto:2020.ashutosh.mishra@ves.ac.in) (Primary)  
- [2020.muskan.chhabria@ves.ac.in](mailto:2020.muskan.chhabria@ves.ac.in)  
- [2020.nikhil.haswani@ves.ac.in](mailto:2020.nikhil.haswani@ves.ac.in)  
- [2020.vanshika.thakur@ves.ac.in](mailto:2020.vanshika.thakur@ves.ac.in)  
- [gresha.bhatia@ves.ac.in](mailto:gresha.bhatia@ves.ac.in)

Primary Subject Area: Data Science Applications

Secondary Subject Areas: Data Science

Submission Files: PradushanCheck Monitoring.pdf (779 Kb, Tue, 02 Apr 2024 18:52:18 GMT)

Submission Questions Response:  
1. Conflict of interest  
Agreement accepted  
2. Status of using third-party material in your article.  
I am not using third-party material for which formal permission is required.  
3. Certificate of originality  
Agreement accepted  
4. Corresponding Author's Contact number  
+918669301882  
5. Country Name  
India  
6. Authors Contributions  
Author\_1, Author\_2, Author\_3, Author\_4: Comparison of different algorithms  
Author\_5: Mentoring and guidance  
Author\_1, Author\_3: Content  
Author\_3: Diagrams

### b. Plagiarism report



## c. Project review sheet

### i. Review 1 (10th February, 2024)

Inhouse/ Industry_Innovation/Research:														Class: D17 A/B/C			
Sustainable Goal: <u>Good Health &amp; Well Being</u>														Project Evaluation Sheet 2023 - 24		Group No.: 8	
Title of Project: <u>PradushanCheck- Comprehensive Air Quality Monitoring and Forecasting</u>																	
Group Members: <u>Ashutosh Mishra (36), Varshika Thakur (57), Nikhil Hamsari (D17A-22), Nikhil Chhabria (D17C), Nikhil Chhabria (D17C-12)</u>																	
Engineering Concepts & Knowledge	Interpretation of Problem & Analysis	Design / Prototype	Interpretation of Data & Dataset	Modern Tool Usage	Societal Benefit, Safety Consideration	Environment Friendly	Ethics	Team work	Presentation Skills	Applied Engg&Mgmt principles	Life - long learning	Professional Skills	Innovative Approach	Research Paper	Total Marks		
(5)	(5)	(5)	(3)	(5)	(2)	(2)	(2)	(2)	(2)	(3)	(3)	(3)	(3)	(5)	(50)		
04	04	04	03	04	02	02	02	02	02	02	03	02	02	04	42		

Comments: \_\_\_\_\_

Dr. G. Bhatia  
Name & Signature Reviewer1

Engineering Concepts & Knowledge	Interpretation of Problem & Analysis	Design / Prototype	Interpretation of Data & Dataset	Modern Tool Usage	Societal Benefit, Safety Consideration	Environment Friendly	Ethics	Team work	Presentation Skills	Applied Engg&Mgmt principles	Life - long learning	Professional Skills	Innovative Approach	Research Paper	Total Marks
(5)	(5)	(5)	(3)	(5)	(2)	(2)	(2)	(2)	(2)	(3)	(3)	(3)	(3)	(5)	(50)
04	04	04	03	04	02	02	02	02	02	02	03	02	02	04	42

Comments: \_\_\_\_\_

Date: 10th february, 2024

Priyanka P  
Name & Signature Reviewer 2

### ii. Review 2 (9th March, 2024)

Inhouse/ Industry_Innovation/Research:														Class: D17 A/B/C			
Sustainable Goal: <u>Good Health &amp; Well Being</u>														Project Evaluation Sheet 2023 - 24		Group No.: 8	
Title of Project: <u>PradushanCheck: Comprehensive Air Quality Forecasting &amp; Monitoring</u>																	
Group Members: <u>Ashutosh Mishra, Nikhil Hamsari, Nikhil Chhabria, Varshika Thakur</u>																	
Engineering Concepts & Knowledge	Interpretation of Problem & Analysis	Design / Prototype	Interpretation of Data & Dataset	Modern Tool Usage	Societal Benefit, Safety Consideration	Environment Friendly	Ethics	Team work	Presentation Skills	Applied Engg&Mgmt principles	Life - long learning	Professional Skills	Innovative Approach	Research Paper	Total Marks		
(5)	(5)	(5)	(3)	(5)	(2)	(2)	(2)	(2)	(2)	(3)	(3)	(3)	(3)	(5)	(50)		
05	04	04	02	04	02	02	02	02	02	02	03	03	03	04	44		

Comments: Forecasting module is pending.

Dr. G. Bhatia  
Name & Signature Reviewer1

Engineering Concepts & Knowledge	Interpretation of Problem & Analysis	Design / Prototype	Interpretation of Data & Dataset	Modern Tool Usage	Societal Benefit, Safety Consideration	Environment Friendly	Ethics	Team work	Presentation Skills	Applied Engg&Mgmt principles	Life - long learning	Professional Skills	Innovative Approach	Research Paper	Total Marks
(5)	(5)	(5)	(3)	(5)	(2)	(2)	(2)	(2)	(2)	(3)	(3)	(3)	(3)	(5)	(50)
05	04	03	02	04	02	02	02	02	02	02	03	03	03	04	43

Comments: \_\_\_\_\_

Date: 9th March, 2024

Priyanka Thakur  
Name & Signature Reviewer 2