AIR QUALITY ANALYSIS MACHINE LEARNING ALGORITHMS

SEMINAR-1 REPORT

Submitted by

DARSHAN R (RA2111003020417)

SOWMITHIRAN S (RA2111003020431)

SHYAM SUNDAR S (RA2111003020435)

Under the guidance of

Dr.B.ABIRAMI, M.E., Ph.D.,

DR.C.SHANMUGANATHAN, M.E., Ph.D.,

(Assistant Professors, Department of Computer Science and Engineering)

In partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

of

COLLEGE OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

RAMAPURAM, CHENNAI-600089.

NOVEMBER 2023

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

(Deemed to be University Under Section 3 of UGC Act, 1956)

BONAFIDE CERTIFICATE

Certified that the Seminar-I report titled "**AIR QUALITY ANALYSIS USING MACHINE LEARNING ALGORITHMS**" is the bonafide work of "**DARSHAN R [RA2111003020417], SOWMITHIRAN S [RA2111003020431], SHYAM SUNDAR S [RA2111003020435]**" submitted for the course 18CSP103L Seminar – I. This report is a record of successful completion of the specified course evaluated based on literature reviews and the supervisor. No part of the Seminar Report has been submitted for any degree, diploma, title, or recognition before.

SIGNATURE

Dr.B.ABIRAMI, M.E., Ph.D.,
Assistant Professor
Dept. of Computer Science & Engineering
SRM Institute of Science and Technology
Ramapuram, Chennai.

SIGNATURE

Dr. K. RAJA, M.E., Ph.D.,
Professor and Head
Dept. of Computer Science & Engineering
SRM Institute of Science and Technology
Ramapuram, Chennai.

Submitted for the Seminar-1 Viva Voce Examination held on 16-03-2023 at SRM Institute of Science and Technology, Ramapuram, Chennai-600089.

EXAMINER 1

EXAMINER 2

ABSTRACT

Monitoring and preserving air quality has become one of the most essential activities in many industrial and urban areas today. The quality of air is adversely affected due to various forms of pollution caused by transportation, electricity, fuel uses etc. The deposition of harmful gases is creating a serious threat for the quality of life in smart cities. With increasing air pollution, we need to implement efficient air quality monitoring models which collect information about the concentration of air pollutants and provide assessment of air pollution in each area. Hence, air quality evaluation and prediction has become an important research area. The quality of air is affected by multi-dimensional factors including location, time, and uncertain variables. Recently, many researchers began to use the big data analytics approach due to advancements in big data applications and availability of environmental sensing networks and sensor data. The aim of this research paper is to investigate various big-data and machine learning based techniques for air quality forecasting. This paper reviews the published research results relating to air quality evaluation using methods of artificial intelligence, decision trees, deep learning etc. Furthermore, it throws light on some of the challenges and future research needs.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ANCRONYMS AND ABBREVATIONS

1. AQI - Air Quality Index
2. PM - Particulate Matter
3. PM2.5 - Fine Particulate Matter with a diameter of 2.5 micrometers or smaller
4. PM10 - Particulate Matter with a diameter of 10 micrometers or smaller
5. CO - Carbon Monoxide
6. NO2 - Nitrogen Dioxide
7. SO2 - Sulfur Dioxide
8. O3 - Ozone
9. VOCs - Volatile Organic Compounds
10. ML - Machine Learning
11. AI - Artificial Intelligence
12. RNN - Recurrent Neural Network
13. CNN - Convolutional Neural Network
14. SVM - Support Vector Machine
15. ANN - Artificial Neural Network
16. IoT - Internet of Things
17. GIS - Geographic Information System
18. EPA - Environmental Protection Agency

Chapter-1
INTRODUCTION

Air is one of the most essential natural resources for the existence and survival of the entire life on this planet. All forms of life including plants and animals depend on air for their basic survival. Thus, all living organisms need good quality of air which is free of harmful gases to continue their life. According to the world's worst polluted places by Blacksmith Institute in 2008 [1], two of the worst pollution problems in the world are urban air quality and indoor air pollution. The increasing population, its automobiles and industries are polluting all the air at an alarming rate. Air pollution can cause long-term and short-term health effects. It's found that the elderly and young children are more affected by air pollution. Short-term health effects include eye, nose, and throat irritation, headaches, allergic reactions, and upper respiratory infections. Some long-term health effects are lung cancer, brain damage, liver damage, kidney damage, heart disease, and respiratory disease. It also contributes to the depletion of the ozone layer, which protects the Earth from sun's UV rays. Another negative effect of air pollution is the formation of acid rain, which harms trees, soils, rivers, and wildlife. Some of the other environmental effects of air pollution are haze, eutrophication, and global climate change. Hence, air pollution is one of the most alarming concerns for us today. Addressing this concern, in the past decades, many researchers have spent lots of time on studying and developing different models and methods in air quality analysis and evaluation. Air quality evaluation has been conducted using conventional approaches in all these years. These approaches involve manual collection and assessment of raw data, with the advancement in technology and research, alternatives to traditional methods have been proposed which use big-data and machine learning approaches. In recent times, many researchers have developed or used big data analytics models and machine learning based models to conduct air quality evaluation to achieve better accuracy in evaluation and perdition. This paper is written based on our recent literature survey and study on the existing publications which focused on air quality evaluation and prediction using these approaches.

18.1 AIM of the project

The project aims to develop an integrated system that leverages Machine Learning (ML) and the Internet of Things (IoT) to address these challenges. The system will consist of a network of IoT sensors deployed across a geographic area to continuously collect air quality data. Machine Learning algorithms will be employed to analyze this data, predict pollution levels, and provide early warnings to the public, primary goal is to improve air quality monitoring and management by harnessing the capabilities of ML and IoT, thereby contributing to a healthier and more sustainable environment.

18.2 Objective of the Project

To monitor, assess, and improve air quality in various environments. This approach leverages IoT devices for data collection and transmission and ML algorithms for data analysis, prediction, and decision-making. The primary goals of such systems are to use Real-Time Monitoring: IoT sensors collect real-time data on various air pollutants such as PM2.5, PM10 CO , NO2, SO2, O3, VOCs, and more. This allows for continuous monitoring of air quality. The objective of IOT devices is to collect, aggregate, and store this data efficiently, making it accessible for analysis and ML models can be trained to detect patterns and anomalies in air quality data. When pollution levels exceed predefined thresholds, the system can trigger alerts and warnings to the relevant authorities and the public. This helps in taking timely actions to mitigate air quality issues. Providing the public with access to real-time air quality data and alerts can raise awareness about pollution issues and encourage behavioral changes that reduce individual contributions to pollution. Ultimately, the objective is to empower decision-makers with actionable insights for implementing effective air quality management strategies, reduce health risks, and enhance the overall quality of life in urban area.

18.3 Problem Statement

Air quality analysis using ML and IoT seeks to address the pressing issue of deteriorating air quality in urban environments. Rising pollution levels pose severe health and environmental hazards, necessitating real-time monitoring and effective mitigation strategies. The problem statement revolves around developing a comprehensive system that integrates IoT sensors to collect data on various air pollutants, including particulate matter (PM2.5 and PM10), VOCs and gases like NO2 and SO2. Assessing air quality in real-time is essential to mitigate these risks. However, traditional monitoring methods have limitations in terms of coverage and real-time data dissemination. These sensors are strategically deployed across the city to provide localized data.

Air pollution is a growing global concern with detrimental impacts on public health and the environment. Accurate and timely assessment of air quality is essential for mitigating these consequences. However, traditional monitoring systems are often limited in terms of coverage and real-time data availability. This project seeks to address this issue by developing an advanced air quality analysis framework.

18.4 Project Domain – ML/IoT

Machine Learning (ML) is transforming the world with research breakthroughs that are leading to the progress of every field. We are living in an era of data explosion. This further improves the output as data that can be fed to the models is more than it has ever been. Therefore, prediction algorithms are now capable of solving many of the complex problems that we face by leveraging the power of data. The models are capable of correlating a dataset and its features with an accuracy that humans fail to achieve. Bearing this in mind, this research takes an in-depth look into the of the problem- solving potential of air quality analysis **ML with IoT** enables accurate air quality forecasting, allowing authorities and the public to make informed decisions regarding outdoor activities, health measures, and pollution reduction strategies. These systems can identify pollution sources, track changes in air quality, and offer early warnings for events like smog or high pollen levels. Additionally, by integrating historical and real-time data, ML models can create adaptive solutions, optimizing the operation of pollution control systems and even influencing urban planning and transportation management for reduced emissions.

In summary, ML with IoT revolutionizes air quality analysis by enhancing data accuracy, prediction capabilities, and real-time accessibility. It empowers communities to actively address air pollution and contributes to cleaner, healthier environments.

In essence, ML with IoT is a catalyst for innovation, efficiency, and data-driven decision-making across sectors, paving the way for a smarter, more connected, and more sustainable future.

18.5 Scope of the Project

The scope of air quality analysis using Machine Learning (ML) and the Internet of Things (IoT) is vast and holds tremendous potential for addressing critical environmental and public health challenges. This integrated approach offers a multi-faceted range of opportunities, spanning several domains. Real-time air quality information can be made readily available to the general public through user-friendly interfaces and mobile applications. This empowers individuals to make informed decisions about outdoor activities and health measures, fostering a sense of responsibility and community involvement

In the **field of environmental science**, it provides a means to comprehensively monitor and assess air quality parameters, including particulate matter (PM2.5, PM10), gases (CO, NO2, SO2, O3), and VOCs. This data can then be analyzed to identify trends, sources of pollution, and predict air quality fluctuations.

For public health, it offers the potential for early warning systems that can notify individuals about poor air quality, allowing them to take precautions and make informed decisions to protect their health. Additionally, it enables healthcare professionals to track air quality-related health issues more accurately and develop targeted intervention strategies.

In urban planning, ML and IoT can influence city development, transportation management, and energy use to reduce pollution and enhance the quality of life. Moreover, it facilitates compliance with environmental regulations, helping authorities monitor and manage air quality more effectively.

The scope of this approach is not limited to a specific industry or domain but extends to various sectors. This technology can revolutionize various sectors, including transportation, urban planning, healthcare, and energy management. For instance, it can optimize traffic flow to reduce emissions, inform land-use planning to minimize pollution exposure, aid in early disease detection through health data correlation, and enhance energy efficiency through real-time data on energy consumption and emissions. Making it a critical tool for fostering cleaner, healthier environments, data-driven decision-making, and a more sustainable future.

Chapter 2
PROJECT DESCRIPTION

2.1 Existing System

Existing systems for air quality analysis use a combination of monitoring stations, remote sensing technologies, and computer modeling. Ground-based monitoring stations measure pollutants like particulate matter, ozone, and nitrogen oxides. Satellite-based sensors provide a broader perspective, enabling the monitoring of air quality over large areas. These data sources are integrated with computer models that simulate the dispersion and concentration of pollutants in the atmosphere. Together, these systems allow for real-time monitoring and forecasting of air quality, helping to assess pollution levels, support public health initiatives, and inform regulatory measures to improve air quality and reduce environmental and health impacts.

2.2 Literature Review

SNO.	TITLE	AUTHOR	YEAR	INFERENCE	LIMITATION
1.	Deep Learning Architecture for air quality predictions	[1] Xiang Li [2] Ling Peng [3] Yuvan Hu [4] Jing Shao [5] Tianhe Chi	**2016**	The paper introduces a Spatiotemporal Deep Learning (STDL) method for air quality prediction. It employs a stacked autoencoder model to extract air quality features, considering both spatial and temporal correlations. This approach outperforms traditional time series models and shows stability across all seasons, offering a more effective solution for air prediction	
2.	Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities		**2019**	The paper evaluates four regression techniques for air pollution prediction. It concludes that Random Forest outperforms the others, with the lowest error rate and faster processing times. Boosting, exhibited the worst performance with longer processing times and high error rates. Random Forest regression emerges as the most effective.	The study is that it primarily focused on traditional machine learning techniques and did not explore more advanced deep learning models, which have shown promise in air pollution prediction. This limitation may have overlooked potentially superior methods. .

3.	Deep Learning methods evaluation to predict air quality based on computational fluid dynamics	Xavier Jurado	**2022**	The study applied various architectures to pollutant dispersion modeling, finding that the multiResUnet architecture with J3D loss outperformed others. It achieved promising results in minutes compared to computational methods taking hours, offering potential for real-time urban air pollution assessment with Computational Fluid Dynamics (CFD) accuracy, particularly from traffic sources	This study include the lack of specific details on the metrics used, dataset characteristics, and environmental conditions, making it challenging to assess the model's robustness. Furthermore, real-world generalizability and potential issues with diverse urban settings and pollutant sources are not addressed.
4.	IMAGEBASED AIR QUALITY ANALYSIS USING DEEP CONVOLUTIONAL NEURAL NETWORK	[1] Avijoy Chakma [2] Tingting Cao [3] Jerry Lin [4] Ben Vizena [5] Jing Zhang	**2017**	This study introduces a CNN-based method for estimating PM2.5 concentration in natural images and offers a dataset for public use,testing various CNN models for better classification accuracyy.	The absence of specific details about the dataset's representativeness, the methodology used in image classification, and the evaluation metrics employed, making it challenging to assess the method's reliability and generalizability.

5.	**Air Quality Prediction by Machine Learning**	[1] Ritik Sharma, [2] Gaurav Shilimka **[3]** Shivam Pisal	**2021**	This paper anticipates the air quality list by utilizing distinctive calculations like direct relapse, Decision Tree and Random Forest. From the outcomes, we reasoned that the Random Forest calculation gives better expectation of air quality list	The limitation of this model lies in the lack of specific details regarding the machine learning algorithms used and their performance evaluation. Additionally, the paper doesn't address the challenges related to data accuracy, sensor calibration, or the potential need for real-time data validation
6	Air pollution prediction with machine learning: a case study of Indian cities	[1] Kumar **[2]** BP Pande	**2022**	The study addresses air quality prediction for 23 Indian cities using machine learning. XGBoost outperformed other models. The research contributes to Indian air quality analysis, highlighting a need for further exploration using deep learning techniques**.**	The paper doesn't delve into the specific deep learning techniques to be employed, potentially leaving a gap in the discussion of advanced methodologies. Additionally, further validation on diverse Indian cities is needed for comprehensive generalizability
7.	Air Quality Prediction using Machine Learning Algorithms	[1] Pooja Bhalgat [2] Sejal Pitale [3] Sachin Bhoite	**2019**	The study concludes that several cities exhibit high pollution levels and immediate attention is required. It utilized AR and ARIMA models for SO2 predictions, identified key air quality parameters, and discussed safe SO2 levels	The model's inability to account for data sequence and city-level predictions is a limitation, highlighting the need for AQI calculations and classification models. Additionally, the passage briefly mentions research challenges like

					PM2.5 and AQI, which require more exploration.
8	**Air Quality Prediction: Big Data and Machine Learning Approaches**	[1] Gaganjot Kaur Kang [2] Jerry Zeyu Gao [3] Sen Chiao [4] Shengqiang Lu [5] Gang Xie	**2018**	The paper discusses the potential of real-time air quality monitoring in smart cities using IoT, big data, and machine learning. It reviews current research in this area, aiming to identify trends and future research directions.	The paper lacks specific details about the findings or observations made in the literature study, making it difficult to assess the depth of the analysis. Additionally, it doesn't mention any specific challenges or needs in air quality evaluation
9	**Advances in air quality research – current and emerging challenges**	[1] Ranjeet S. Sokhi [2] Rainer Friedrich [3] Peter Suppan [4] Steve Hanna	**2022**	The paper highlights the advantages of multi-model ensemble modeling in air quality forecasting, particularly the potential for improved performance when combining top- and low-ranking models. It also discusses the value of including outliers in the ensemble.	The paper lacks specific details or examples of the models used, making it challenging to assess the practical implications of these findings. Additionally, it does not address potential challenges or limitations of ensemble m odeling.
10	**Air Quality Monitoring and Forecasting Services**	**Dr. M. Mohapatra Director General of Meteorology**	**2021**	The IMD deals the environmental monitoring and air quality forecasting in India. IMD provide pollution impacts. The SOP highlights the importance of air quality forecasts as a management	The paper doesn't provide specific examples or data on the effectiveness of the air quality forecasting system. Additionally, it doesn't discuss challenges or limitations faced in air quality forecasting in In dia

11	**Air Quality Index – A Comparative Study for Assessing the Status of Air Quality**	[1] Shivangi Nigam [2] B.P.S. Rao [3] N. Kumar [4] V. A. Mhaisalkar	**2016**	This paper highlights the significance of the Air Quality Index (AQI) in assessing ambient air quality, particularly the role of PM10 as a major pollutant. It emphasizes the need for pollution control measures, including green initiatives, for public health benefits.	The paper acknowledges that while the AQI is a valuable tool for decision-making, it doesn't fully account for temporal variations due to meteorological and ecological factors. It also raises concerns about health impacts on underprivileged populations and the need for inclusive development strategies, but doesn't provide specific recommendations or solutions.
12	**An integrated analysis of air pollution and meteorological conditions in Jakarta**	Teny Handhayani	**2023**	The integration analysis in Jakarta links air pollution and meteorological conditions. LSTM and GRU models effectively forecast various air quality parameters, but are less accurate .The COVID-19 outbreak's impact on human activities and air quality is less accurate for SO2,NO.	The paper doesn't delve into the specific machine learning techniques used, limiting insights into model performance. It mentions future work on broader factors but doesn't provide a detailed roadmap for this research. It is not generalized and is limited for a particular state
13	**Air quality analysis and PM2.5 modelling using machine learning techniques: A study of Hyderabad city in India**	[1] Aneesh Mathew, [2] P R Gokul, [3] Padala Raja Shekar [4] K. S. Arunab [5] Hazem Ghassan Abdo	**2023**	The study in Hyderabad employs machine learning models to investigate air quality, emphasizing PM2.5 levels' adverse impact on public health and the environment. The	The paper doesn't specify the dataset's size or detailed model performance metrics, making it challenging to gauge the study's scope and results in more detail. Additionally, it doesn't discuss

		[6] Hussein Almohamad [7] Ahmed Abdullah Al Dughair		findings underscore the importance of collaborative efforts among stakeholders to combat air pollution effectively and implement targeted measures	potential challenges or obstacles in implementing these measures. And is limited to one city not a generalized model
14	**Migration and hedonic valuation: The case of air quality**	[1] **Patrick Bayer** [2] **Nathaniel Keohane** [3] **Christopher Timmins**	**2014**	**The paper introduces a discrete-choice model to estimate the willingness to pay for improved air quality in the US, accounting for mobility constraints and residential patterns.**	**The paper lacks details on the model's specific methodology and data sources, making it challenging to assess the study's accuracy and the applicability of the results.**
15	**Cost-effective control of air quality and greenhouse gases in Europe: Modeling and policy applications**	[1] **Imrich Bertok** [2] **Jens Borken-Kleefeld** [3] **Janusz Cofala** [4] **Chris Heyes** [5] **Lena Höglund-Isaksson** [6] **Zbigniew Klimont** [7] **Binh Nguyen** [8] **Maximilian Posch Peter** [9] **Rafaj**	**2011**	The GAINS model informs negotiations under the Convention on Long-range Transboundary Air Pollution, providing baseline projections for emissions and air quality in European countries. Saliency, credibility, and legitimacy are vital for decision-makers' acceptance of scientific assessments and models, highlighting the model's close interaction with policymakers	The paper mentions the application of the GAINS model for informing negotiations on air pollution protocols. However, it lacks specific details about the model's performance, its methodologies, and the extent of interaction with decision-makers. This makes it challenging to fully understand the model's effectiveness and the scope of its influence on policy decisions.
16	***openair* — An R package for air quality data analysis**	[1] **David C. Carslaw** [2] **Karl Ropkins**	**2012**	The paper discusses the underutilization of air pollution data collected for compliance and research purposes, highlighting the need	The paper does not elaborate on the specific functionalities of the R-based tools, making it difficult to assess their

				for more insightful analysis. It introduces a set of air pollution analysis tools developed using the R statistical software and provides examples of their usage to draw meaningful inferences from the data.	effectiveness. It mentions potential benefits but doesn't provide concrete results or validation of the tools. Additionally, the paper primarily focuses on the UK, limiting the scope of international applica bility.
17	**the analysis of air quality data and total atmospheric ozone**	**Y.-S. Chung**	**1967**	The paper discusses the significant environmental impact of man-caused forest fires in Canada, including the release of substantial emissions like CO, TSP, HC, and NO. It highlights the reduced visibility due to these fires and suggests a potential correlation with surface ozone depletion in the lower troposphere.	The paper lacks specific data and details about the extent of environmental damage, making it challenging to assess the full scope and consequences of these forest fires. Additionally, it does not provide a comprehensive analysis of the methods used to measure these impacts
18	**THE IMPACT OF AIR QUALITY CONDITIONED BY EMISSION OF POLUTTANTS TO THE DEVELOPMENT OF RURAL TOURISM AND POTENTIA**	[1] **Drago Cvijanović** [2] **Jelena Matijašević – Obradović** [3] **Sanja Škorić**	**2017**	The paper discusses the impact of air quality on the development of rural areas in Serbia, classifying regions based on air pollution levels. It highlights that areas with cleaner air, like "Novi Sad," have potential for rural development. In contrast, areas with excessive pollution, such as "Bor" and	The paper provides a detailed classification of various regions in Serbia based on air quality. However, it lacks specific quantitative data on air pollution levels and their impact on rural development. The passage would benefit from statistical figures to support its claims, and it does not delve

	LS OF RURAL AREAS			"Belgrade," face challenges in promoting rural development and require significant improvements in reducing pollutant emissions.	into the specific strategies or measures that need to be implemented to address air quality issues and promote rural development.

2.3 Issues in Existing Systems

Existing air quality analysis systems may have various issues, some of which include

Limited Spatial Coverage: Many air quality monitoring systems are concentrated in urban areas, leaving rural and remote regions underserved. This can lead to an incomplete understanding of air quality on a larger scale.

Data Gaps: Monitoring networks can have gaps in coverage, both spatially and temporally. These gaps can lead to inaccurate assessments of air quality and hinder the ability to detect short-term pollution events or changes over time.

Outdated Technology: Some air quality monitoring systems still rely on outdated technology, which may result in less accurate measurements and slower data reporting. Modern sensor technology offers greater precision and real-time data collection.

Limited Pollutants Monitored: Many monitoring systems primarily focus on a few key pollutants, such as particulate matter (PM2.5 and PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO), and ozone (O3). However, there are many other pollutants and air quality parameters that can impact health and the environment.

Lack of Transparency: Access to air quality data can sometimes be limited, and the data may not always be presented in a user-friendly manner. This lack of transparency can hinder public awareness and involvement in air quality management.

Calibration and Maintenance: Air quality monitoring equipment requires regular calibration and maintenance to ensure accurate measurements. Neglecting these tasks can lead to measurement errors and unreliable data.

Data Quality Assurance: Quality control and quality assurance procedures are crucial in maintaining the accuracy and reliability of air quality data. Inadequate QA/QC processes can result in erroneous data.

Limited Integration: Many air quality monitoring systems operate in isolation, with data not integrated into a broader environmental monitoring framework. This can limit the ability to assess the interactions between air quality and other environmental factors.

Emerging Pollutants: New pollutants and sources of pollution may emerge over time (e.g., microplastics, pharmaceutical residues). Existing systems may not be equipped to monitor these emerging threats.

2.4 Software Requirements

To perform air quality analysis using machine learning (ML) and the Internet of Things (IoT), you'll need a combination of hardware and

software. Here are the software requirements for such a system

IoT Platforms	AWS IoT, Azure IoT, or Google Cloud IoT Core for device management and data ingestion.
Sensor Drivers	air quality sensors (e.g., PM2.5, CO2, NO2) for data collection
Database Management System	PostgreSQL, MongoDB, or Influx DB to store time-series data collected from IoT devices.
Data Warehousing	Amazon Redshift or Google Big Query for large-scale data storage and analysis
Data Cleaning Tools	Python libraries like Pandas, Numpy, and scikit-learn & raw sensor data
Machine learning libraries	TensorFlow, PyTorch, or scikit-learn for developing predictive models& frameworks like Flask or FastAPI
Data Visualization Tools	Matplotlib, Seaborn, or Plotly can create informative visualizations to present air quality insights.
Dashboard Tools	custom dashboards using frameworks like Django, Flask, or Node.js with libraries like React or Angular, Tableau, Power BI, or custom PDF
Feature Engineering	open-source libraries like scikit-learn and feature selection tools within R.
Cross-Validation Tools	K-fold cross-validation methods available in scikit-learn or R.
Evaluation Metrics	to calculate appropriate metrics like mean squared error (MSE), root mean square error (RMSE), or other domain-specific metrics.
Visualization Libraries	Matplotlib, Seaborn, Plotly, or ggplot2 for creating charts and graphs to visualize the results and Tools like Folium for mapping air quality data geospatially

3.1 Proposed System

The methodology involves obtaining and cleaning a suitable dataset, applying Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset, and then splitting it into training and testing data. The selected algorithms are used to make predictions, and their accuracy is assessed using metrics such as RMSE and R-SQUARE. The study aims to determine which algorithm provides the most accurate AQI predictions for both balanced and imbalanced datasets. The results will be presented with visual aids to aid in understanding and potentially guide future research. Overall, the research intends to contribute to improving AQI prediction methods in highly populated Indian cities. The primary objective is to determine the most accurate and efficient algorithm for predicting AQI values. The choice of these algorithms is based on their demonstrated high accuracy in previous studies .These cities were chosen due to their high population density and their representation of significant pollution sources in South Asia

The study will present the results in a clear and understandable manner through figures, graphs, and charts. The process is outlined in a flowchart, detailing the steps involved in the research.

This research can provide valuable insights into air quality prediction methods and help guide future studies in this area.

3.2 Architecture diagram

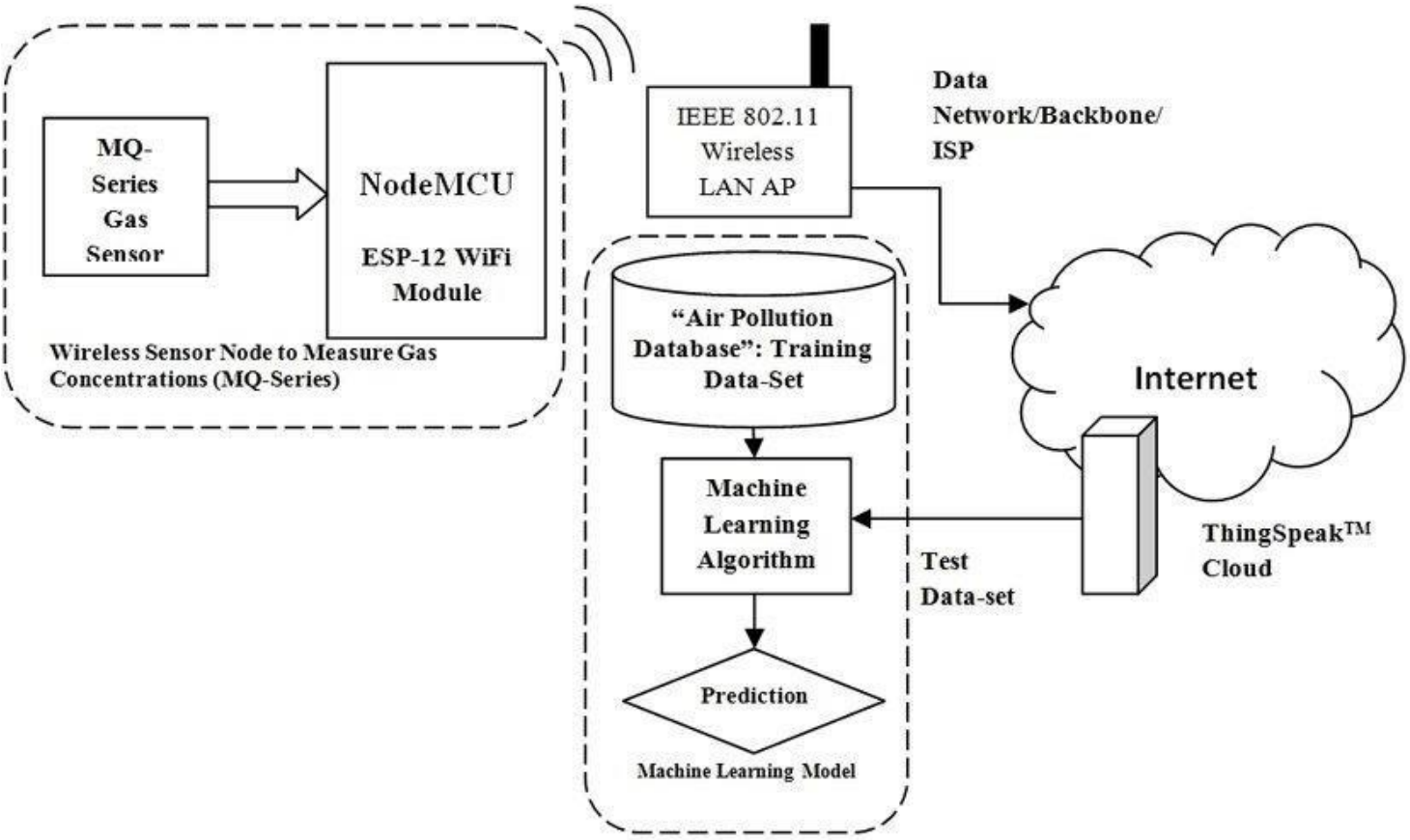


Figure 3.1

The above Fig 3.1 illustrates. An architecture diagram is a visual representation of the structure, components, relationships, and interactions within a system or application. It provides a high-level view of the system's design and helps stakeholders, such as developers, architects, and project managers, understand how the various parts of the system work together. This architecture diagram outlines the key components of an ML-based air quality analysis system, from data ingestion to model deployment and user interaction. It emphasizes the importance of data preprocessing and model evaluation, as well as real-time predictions and historical data analysis to support decision-making regarding air quality.

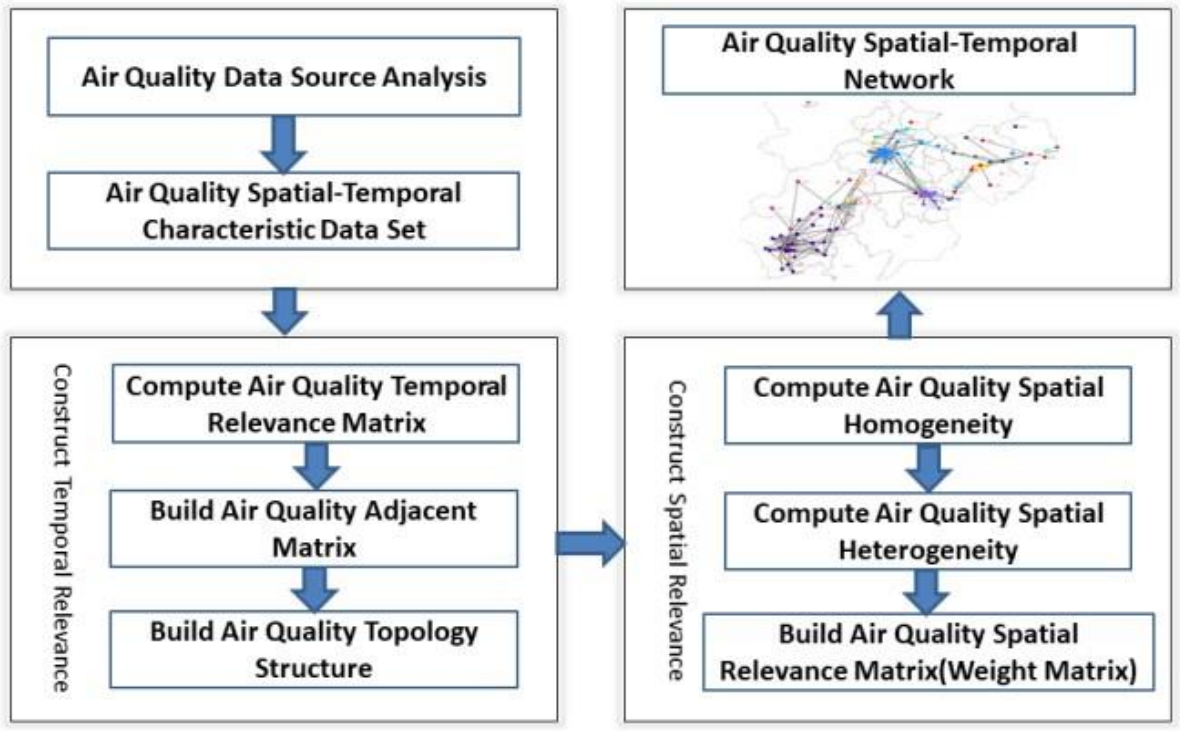


Figure 3.2 Air Quality Analysis Industry Market Architecture

The above figure 3.2 illustrates about the air quality analysis Market architecture. The air quality analysis market encompasses various segments, including ambient air quality monitoring, indoor air quality assessment, source emission monitoring, and modeling for predicting future air quality conditions. The market includes hardware, software, and services, providing a comprehensive solution for air quality analysis.

Overall, the air quality analysis market is poised for continued growth, with an increasing focus on environmental sustainability and public health, making it a vital and evolving sector in environmental science and technology

3.3 Design Phase

The Design Phase consists of the UML diagrams to design and construct the project.

1. Use Case Diagram
2. Data flow Diagram
3. Deployment Diagram

3.4 Use Case Diagram

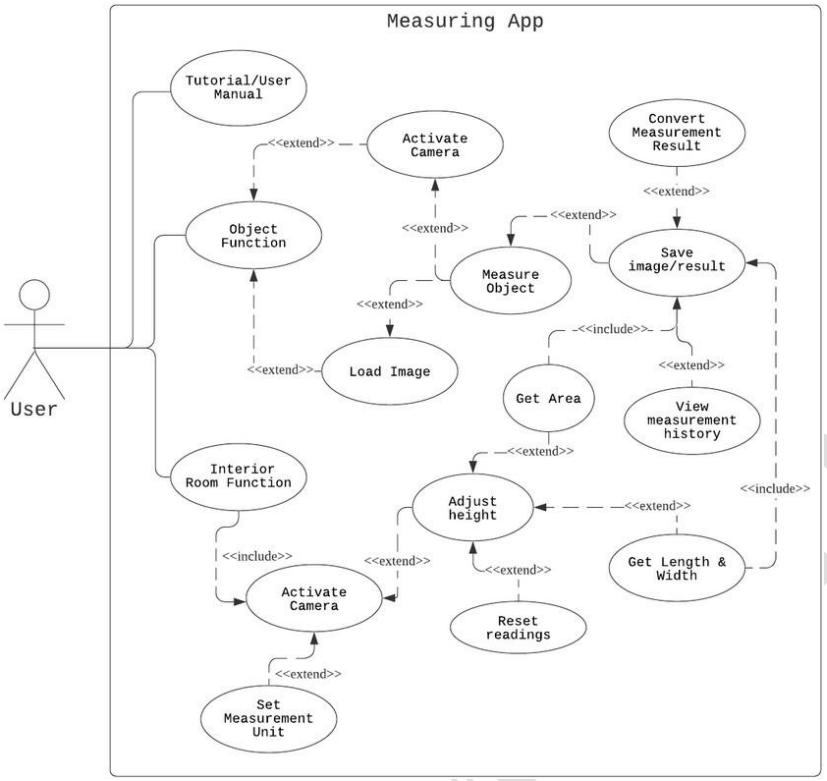


Figure 3.3 Use Case Diagram

The above figure 3.3 illustrates the Use case diagram of the project. A Use Case Diagram is a visual representation of the functional requirements of a system, showcasing how users or external systems interact with the system to achieve specific goals or tasks. It's a key tool in the Unified Modeling Language (UML) for understanding and documenting system functionality.

Actor: Represented as stick figures, typically with a label that describes their role or function.
Use Case: Represented as ovals with a label that describes the specific function or task.
Association Line: A line connecting an actor to a use case to show their interaction.

3.5 Data Flow Diagram

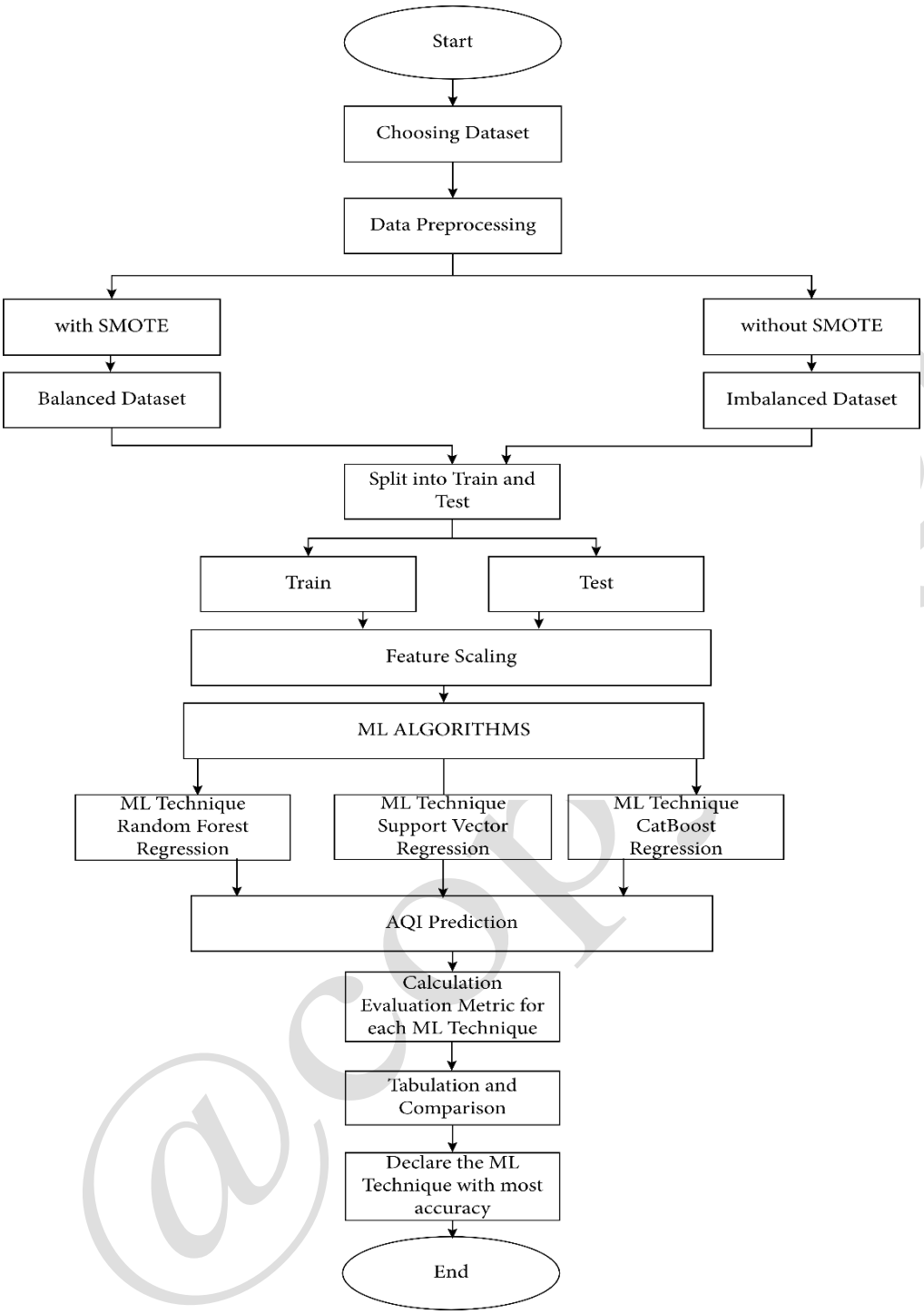


Figure 3.4 Data Flow

The above figure 3.4 illustrates the basic data flow of the project. A Data Flow Diagram (DFD) is a graphical representation of how data moves within a system. It is a visual tool used to depict the flow of data, processes that transform data, data stores, and data sources or sinks. DFDs are commonly used in systems analysis and design to understand, document, and communicate the flow of data in a system

3.5 Deployment Diagram

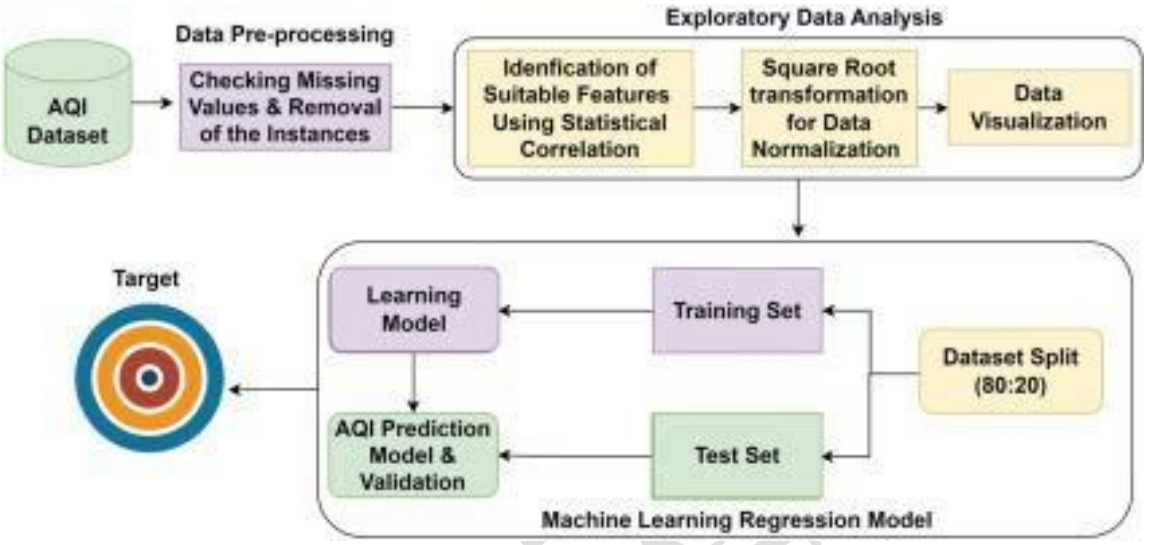


Figure 3.5 Air Quality Analysis

The above figure 3.5 illustrates the deployment diagram of the project. Deployment diagrams are typically used to visualize the physical hardware and software of a system.Using it you can understand how the system will be physically deployed on the hardware. Deployment diagrams help model the hardware topology of a system compared to otherUML diagram types which mostly outline the logical components of a system.

3.6 Module Description

3.7 Dataset Description and Sample Data:

The dataset used in this study comprises hourly and daily air quality data, including the Air Quality Index (AQI), from multiple monitoring stations across 26 cities in India for the years 2015 to 2020. These cities encompass major urban areas with varying levels of population density, making them representative of pollution dynamics in India. The dataset was initially extensive, with 16 columns, but it was refined for analysis, and data from New Delhi, Bangalore, Kolkata, and Hyderabad were extracted.
Link to Dataset : https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india

Each city's dataset contains information about parameters such as PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, AQI, and AQI_Bucket, which indicates air quality categories. Xylene data was omitted due to missing values. The original dataset was imbalanced, which could affect model accuracy, so the authors employed the Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset, a technique that augments underrepresented classes to create balance.

This dataset, with balanced representations, will be used to evaluate and compare the performance of three predictive algorithms in assessing air quality and AQI in the selected Indian cities. The study aims to identify the most accurate method for AQI prediction, considering the unique characteristics of each city's pollution profile.

Table 1:
Some of the existing algorithm accuracy in percentage from the literature survey

Name of the algorithm	Accuracy in percentage (%)	Comments
Naive Bayes (NB)	86.663	-
Support vector machine (SVM)	92.4	-
Artificial neural network (ANN)	84—93	After simulating a lot of models, ANN gives within the range.
Gradient boost (GB)	96	-
Decision tree (DT)	91.9978	Predicting the PM2_s with a near 89% accuracy rate.
Enhanced k-means	71.28	The K - means clustering method is 40% more efficient than the PFCM algorithm based on the speed of execution and accuracy.
Support vector regression (SVR)	99.4	-
Random forest regression (RFR)	99.985	Least MSE of 0.00013 and MAE of 0.00373.
Cat Boost regression (CR)	99.88	Predicting PM2.5 readings with an inaccuracy of just 0.0006 and a 99-88% accuracy-

Table 2:

Sample dataset for New Delhi city.

City	Date	PM2.5	PMIO	NO	N02	NOX	NH3	co	S02	O3	Benzene	Toluene	AQI	AQl_bucket
Delhi	02/01/2015	186.18	269.55	62.09	32.87	88.14	31.83	9.54	6.65	29.97	10.55	20.09	454	Servere
Delhi	03/01/2015	87.18	131.9	25.73	30.31	47.95	69.55	10.61	2.65	19.71	3.91	10.23	143	Moderate
Delhi	04/01/2015	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
Delhi	05/01/2015	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
Delhi	06/01/2015	149.58	252.1	17.21	37.84	42.46	134.97	9.44	36.6	26.83	3.63	7.35	318	Very poor
Delhi	07/01/2015	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
Delhi	08/01/2015	229.9	360.95	2334	43.16	51.21	138.13	11.01	3.31	30.51	5.8	11.4	383	Very poor
Delhi	09/01/2015	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
Delhi	10/01/2015	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
Delhi	11/01/2015	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

Table 3:

Sample dataset for Bangalore city

City	Date	PM2.5	PMIO	NO	N02	NOX	NH3	co	S02	O3	Benzene	Toluene	AQI	AQl_bucket
Bangalore	14/11/2015	42.42	156.8	7.3	29.94	31.78	21.9	1.6	2.23	31	1.82	4.65	130	Moderate
Bangalore	15/11/2015	21.99	39.86	7.1	16.44	19.51	42	1.7	2.95	10	1.52	2.38	103	Moderate
Bangalore	16/11/2015	13.89	31.44	6.8	12.14	15.35	23.9	1.7	2.5	4.4	0.74	1.48	74	Satisfactory
Bangalore	17/11/2015	19.66	36.84	6.5	16.37	20.87	24	1.4	2.83	4.1	1.18	2.17	75	Satisfactory
Bangalore	18/11/2015	20.35	33.97	7.8	20.64	24.75	27	1.4	2.59	7.8	1.02	1.9	85	Satisfactory
Bangalore	20/11/2015	34.39	36.29	8.4	28.8	32.28	32.8	2.5	3.76	15	1.32	3.17	141	Moderate
Bangalore	21/11/2015	43.91	43.65	12	29.33	32.78	55.4	1.5	3.44	15	1.53	3.59	90	Satisfactory
Bangalore	22/11/2015	44.14	112.8	7.1	26.64	27.06	32.3	2.2	4.3	26	1.69	3.36	126	Moderate
Bangalore	24/11/2015	44.94	114.3	8.5	28. I	29.37	32.8	2.3	4.7	29	I.56	2.38	147	Moderate
Bangalore	25/11/2015	29.35	75.79	5.7	21.21	21.4	19.1	1.6	4.55	29	1.01	1.15	87	Satisfactory

Table 4: Sample dataset for Kolkata city.

City	Date	PM2.5	PMIO	NO	N02	NOX	NH3	co	S02	O3	Benzene	Toluene	AQI	AQI_bucket
Kolkata	16/06/2018	47.55	128.66	6.01	24.89	24.51	7.4	0.72	7.3	27.24	2.14	0.81	119	Moderate
Kolkata	18/06/2018	50.1	105.68	3.23	33.28	36.5	8.55	I.47	3.02	72.28	1.97	2.62	107	Moderate
Kolkata	19/06/2018	39.25	87.24	2.6	30.86	33.45	12.1	1.35	1.93	81.12	1.59	2.47	148	Moderate
Kolkata	20/06/2018	24.44	53.19	5.77	38.03	43.79	9.14	1.7	6.88	49.58	2.02	3.13	94	Satisfactory
Kolkata	21/06/2018	31.68	60.16	4.46	38.39	43.04	6.52	1.42	1.31	13.47	3.76	5.52	100	Satisfactory
Kolkata	22/06/2018	25.22	48.96	0.99	28. I	29.07	6.53	0.39	2.31	30.32	1.62	2.65	60	Satisfactory
Kolkata	23/06/2018	22.95	44.58	1.14	25.76	26.85	5.38	0.38	1.06	22.84	1.67	2.63	47	Good
Kolkata	24/06/2018	24.61	46.54	0.86	25.49	26.32	3.96	0.4	1.1	23.13	1.51	2.28	48	Good
Kolkata	25/06/2018	28.6	45.36	1.95	43.45	45.37	3.62	0.41	1.11	13.56	2.58	4.17	50	Good
Kolkata	26/06/2018	30.5	46.08	1.27	37.12	38.33	3.19	0.38	2.29	34.84	2.05	4.41	61	Satisfactory

Table 5: Sample dataset for Hyderabad city

City	Date	PM2.5	PMIO	NO	N02	NOX	NH3	co	S02	O3	Benzene	Toluene	AQI	AQI_bucket
Hyderabad	08/09/2015	91.82	32.9	5.4	29	23.37	24.9	0.48	8	27.04	1.01	5.74	179	Moderate
Hyderabad	09/09/2015	35.56	40.8	4	31	24.31	24.8	0.57	4.9	22.48	1.41	7.61	162	Moderate
Hyderabad	10/09/2015	45.64	44.9	7.1	29	25.58	24.8	0.73	5.3	24.69	1.25	7.84	76	Satisfactory
Hyderabad	11/09/2015	60.88	51.3	5.2	31	24.22	25.9	0.53	5.2	24.11	1.09	5.42	140	Moderate
Hyderabad	12/09/2015	65.61	41.3	3.4	26	20.37	24.8	0.57	5.4	25.47	0.83	4.39	128	Moderate
Hyderabad	13/09/2015	60.02	36.7	2.4	20	14.51	21.7	0.49	4	37.7	0.79	4.07	164	Moderate
Hyderabad	14/09/2015	73.21	35.3	2.8	20	15.4	21.4	0.57	6	34.11	0.52	2.44	169	Moderate
Hyderabad	01/10/2015	120.8	92.3	1.9	22	15.87	27.7	0.64	2.7	15.85	1.21	5.95	340	Very poor
Hyderabad	02/10/2015	29.66	76	2	26	16.02	20.5	0.6	3.8	17.4	1.2	5.62	125	Moderate
Hyderabad	03/10/2015	36.56	63.1	3.1	20	15.07	18.1	0.64	7.6	19.16	1.2	6.4	75	Satisfactory

3.7.1

The aim is to analyze and present it in an efficient way. It would help us discover interesting and insightful information. These particular cities have a higher population density and give a good estimate of the pollution in a major South Asian city. More cities have not been added due to the fact that it makes the research paper way too lengthy. Hence, the major cities of India have been chosen to analyze the pollution levels in different

urban cities of India as they are the major contributors to pollution.

Figure 1: New minority class instances added.

IMBALANCED DATASET
NEW DELHI:

Selected attribute
Name: AQI_Bucket Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 0 (0%)

No.	Label	Count	Weight
1	Severe	239	239
2	Moderate	485	485
3	Very Poor	514	514
4	Poor	534	534
5	Satisfactory	108	108

BALANCED DATASET

Selected attribute
Name: AQI_Bucket Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 0 (0%)

No.	Label	Count	Weight
1	Severe	478	478
2	Moderate	485	485
3	Very Poor	514	514
4	Poor	534	534
5	Satisfactory	432	432

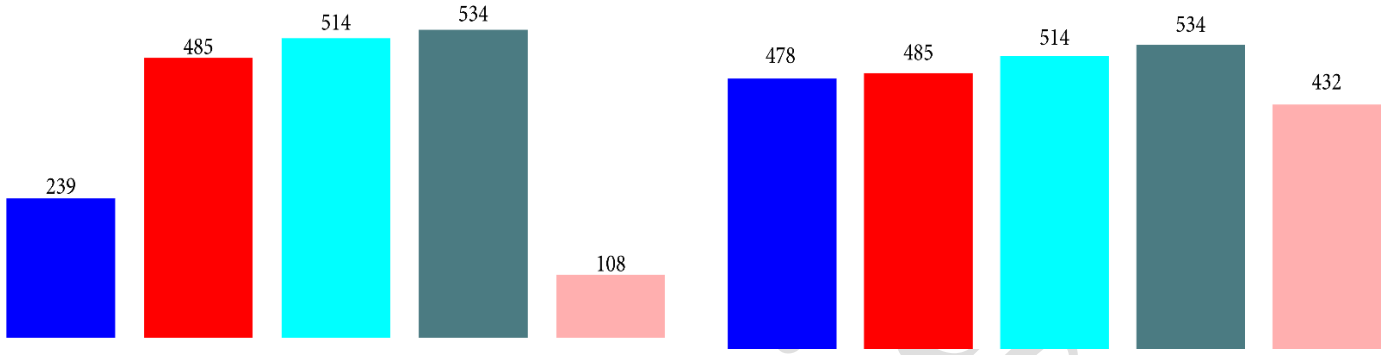


Figure 2: Balanced and imbalanced data values for New Delhi city.

IMBALANCED DATASET
HYDERABAD:

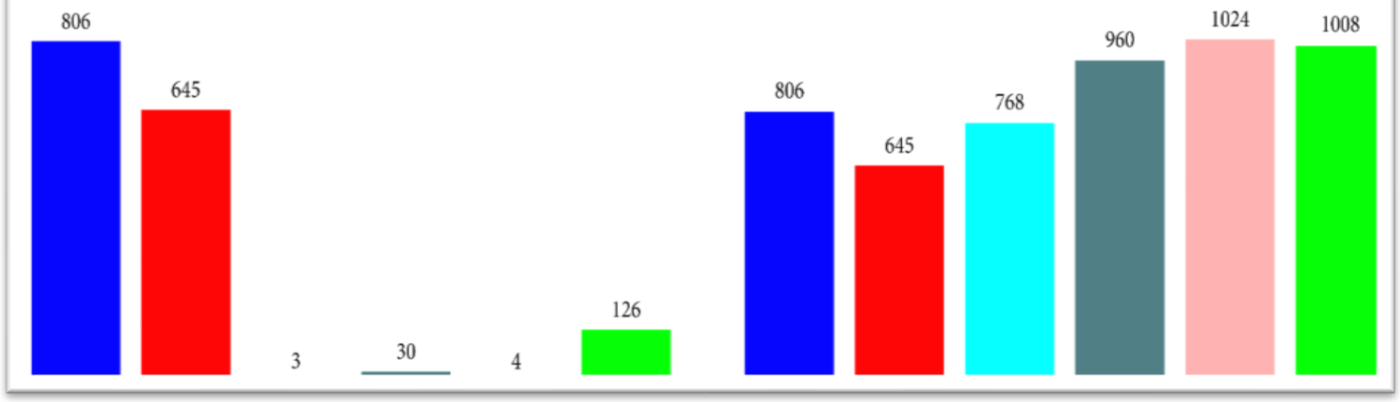
Selected attribute
Name: AQI_Bucket Type: Nominal
Missing: 0 (0%) Distinct: 6 Unique: 0 (0%)

No.	Label	Count	Weight
1	Moderate	806	806
2	Satisfactory	645	645
3	Very Poor	3	3
4	Poor	30	30
5	Severe	4	4
6	Good	126	126

BALANCED DATASET

Selected attribute
Name: AQI_Bucket Type: Nominal
Missing: 0 (0%) Distinct: 6 Unique: 0 (0%)

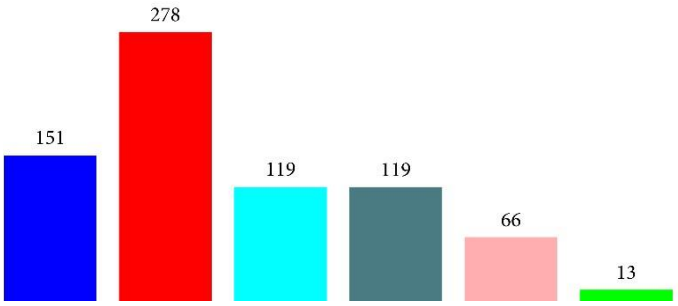
No.	Label	Count	Weight
1	Moderate	806	806
2	Satisfactory	645	645
3	Very Poor	768	768
4	Poor	960	960
5	Severe	1024	1024
6	Good	1008	1008



IMBALANCED DATASET
KOLKATA:

Selected attribute
Name: AQI_Bucket Type: Nominal
Missing: 0 (0%) Distinct: 6 Unique: 1 (0%)

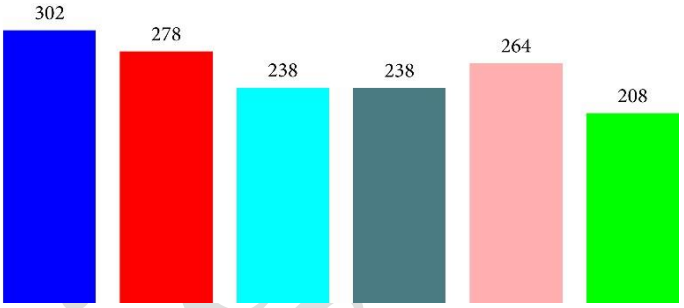
No.	Label	Count	Weight
1	Moderate	151	151
2	Satisfactory	278	278
3	Good	119	119
4	Poor	119	119
5	Very Poor	66	66
6	Severe	13	13



BALANCED DATASET

Selected attribute
Name: AQI_Bucket Type: Nominal
Missing: 0 (0%) Distinct: 6 Unique: 0 (0%)

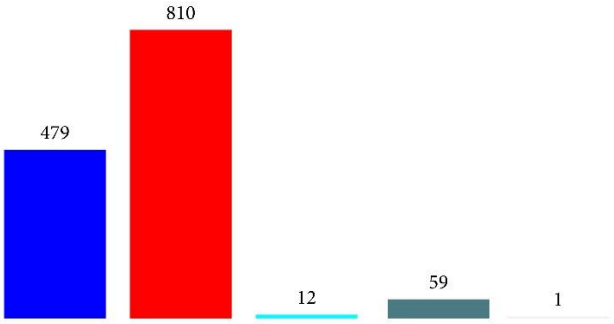
No.	Label	Count	Weight
1	Moderate	302	302
2	Satisfactory	278	278
3	Good	238	238
4	Poor	238	238
5	Very Poor	264	264
6	Severe	208	208



IMBALANCED DATASET
BANGALORE:

Selected attribute
Name: AQI_Bucket Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 1 (0%)

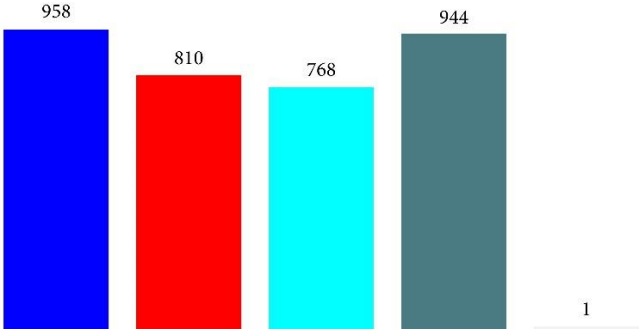
No.	Label	Count	Weight
1	Moderate	479	479
2	Satisfactory	810	810
3	Poor	12	12
4	Good	59	59
5	Very Poor	1	1



BALANCED DATASET

Selected attribute
Name: AQI_Bucket Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 1 (0%)

No.	Label	Count	Weight
1	Moderate	958	958
2	Satisfactory	810	810
3	Poor	768	768
4	Good	944	944
5	Very Poor	1	1



3.7.2 Process Module

Data Selection: The researchers chose a dataset from Kaggle and downloaded its CSV file to work with.

Data Preprocessing: They cleaned the dataset and extracted data for major Indian cities, including New Delhi, Bangalore, Kolkata, and Hyderabad. These cities were selected due to their significant contributions to pollution levels in India, given their high population density.

SMOTE Application: The Synthetic Minority Oversampling Technique (SMOTE) was used to address class imbalances in AQI_Bucket values, making the dataset more balanced.

Data Splitting: The datasets were divided into training and test sets in an 80:20 ratio for model training and testing.

Feature Scaling: Data normalization was performed using StandardScaler from the Scikit-Learn library to ensure uniformity and flexibility in the dataset.

Machine Learning Techniques: Various machine learning algorithms, including CatBoost regression, random forest regression, and support vector regression, were employed to predict the Air Quality Index (AQI) for each city. These algorithms were later compared to determine the best-performing one for each city.

AQI Prediction: Machine learning techniques were used to predict AQI values for each city, and their accuracy levels were assessed and compared.

Evaluation Metrics: Metrics such as R-SQUARE, MSE, RMSE, MAE, and accuracy (1-MAE) were calculated for each machine learning technique to evaluate their performance.

Tabulation and Comparison: The metric values and accuracy were tabulated, and various graphs, including line graphs, density plots, and scatter plots, were used for analysis. The accuracy and metric values were compared city-wise and algorithm-wise.

Final Comparative Results: After comparing all metric values, the researchers identified the best-performing machine learning techniques. In this study, random forest and CatBoost regression exhibited the highest accuracy and performance in predicting AQI values for specific cities.

These steps outline the systematic approach taken to analyze and predict air quality using machine learning techniques, ultimately identifying the algorithms with the highest accuracy for the selected cities.

3.7.3 Discussion on Metrics Used

The metrics used in the proposed work are R-SQUARE, mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and accuracy.

(i) **R-SQUARE** indicates to what extent the regression model is in line with the observed data. A higher R square value denotes a better model ft, the R Square equation is shown by equation

$$R - \text{SQUARE} = \frac{\text{SSregr}}{SStt}.$$

The sum of squares due to regression is denoted by *SS*regr (explained sum of squares), while the sum of squares overall is denoted by *SStt*. The degree to which the regression model fits the data well is shown by the sum of squares due to regression. The total sum of squares is used to determine how much the observed data has changed (data utilized in regression modeling)

.(ii)**MSE** is a parameter that measures how closely a fitted line resembles a set of data points. The lower the value, the closer it is to the line, and hence the better. If the MSE value = 0, the model is perfect. It is shown in equation

$$MSE = \sum_{i-1}^{n} \frac{(X_i - X_i^{\wedge})^2}{n},$$

where $A = \pi r^2$,

(a) x_i = The i^{th} observed value

(b) $x_i^{\wedge}$ = The corresponding predicted value

(c) n = The number of observations

(iii) RMSE indicates how densely the data are distributed along the line of best fit. RMSE values in the range of 0.2–0.5 demonstrate that the model can reasonably predict the data. It is shown in the equation

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(X_i - X_i^{\wedge})^2}{m}},$$

where

(a) x_i = The i^{th} observed value

(b) $x_i^{\wedge}$ = The corresponding predicted value

(c) n = The number of observations

(iv) **MAE evaluates** the absolute distance of the observations to the predictions on the regression line. It is shown in the equation

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{n} |X_i - X|,$$

where

(a) n is the number of errors

(b) Σ is the summation symbol (which means "add them all up")

(c) $|x_i - x|$ is the absolute errors

(v) **Accuracy** is used as a measurement to calculate how well a model is finding patterns and identifying relations in the dataset and it is shown in the equation

$$\text{Accuracy} = (1 - MAE) * 100.$$

Table 7

Accuracy results comparison of the imbalanced dataset for four cities and methods used.

	Cities			
Method	New Delhi	Bangalore (%)	Kolkata (%)	Hyderabad (%)
	Accuracy (%)			

Support vector regression	78.4867	66.4564	89.1656	76.6786
Random forest regression	79.4764	67.7038	90.9700	78.3672
Cat Boost regression	79.8622	68.6860	89.9766	77.8991

Table 9
The result of performance metrics used for Bangalore city imbalanced dataset, without using the SMOTE algorithm.

Algorithm name	R-square	MSE	RMSE	MAE
Support vector regression	0.6525	0.3772	0.6142	0.3354
Random forest regression	0.7035	0.3219	0.5674	0.3229
CatBoost regression	0.6877	0.3391	0.5823	0.3131

Results and Discussion

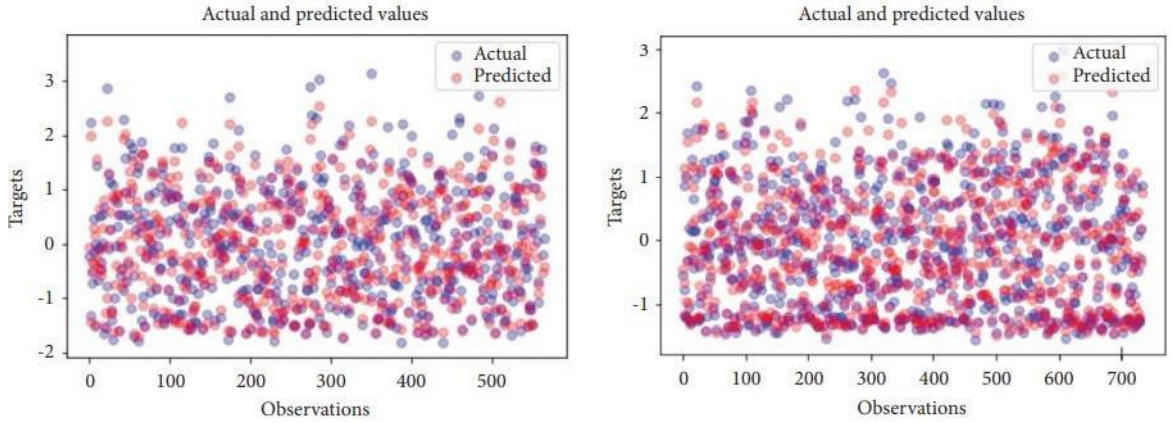


Figure 15: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for New Delhi-SVR\

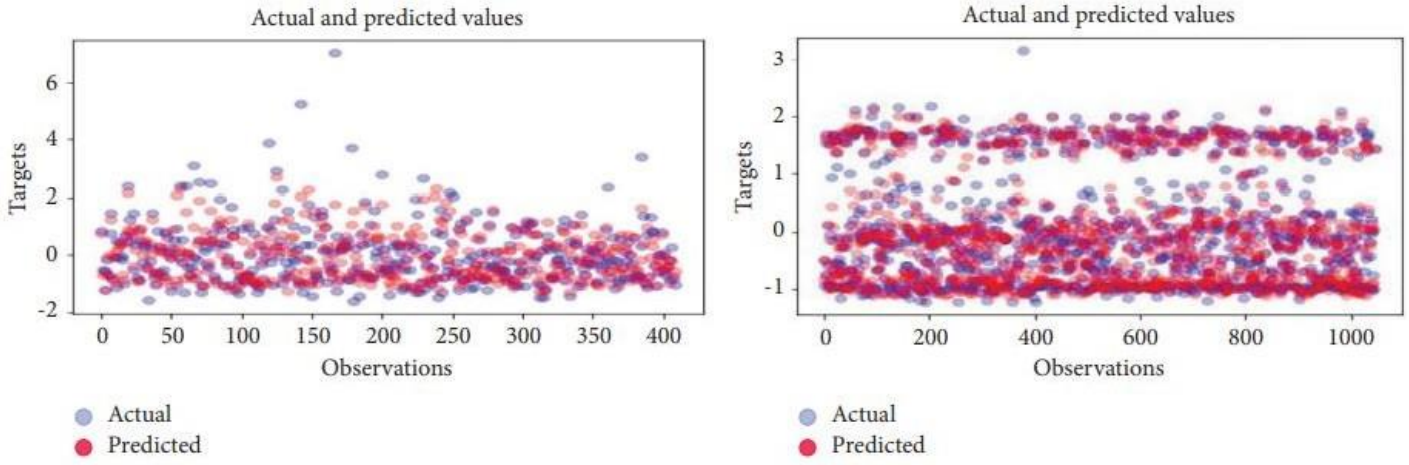


Figure 16: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for Bangalore-SVR.

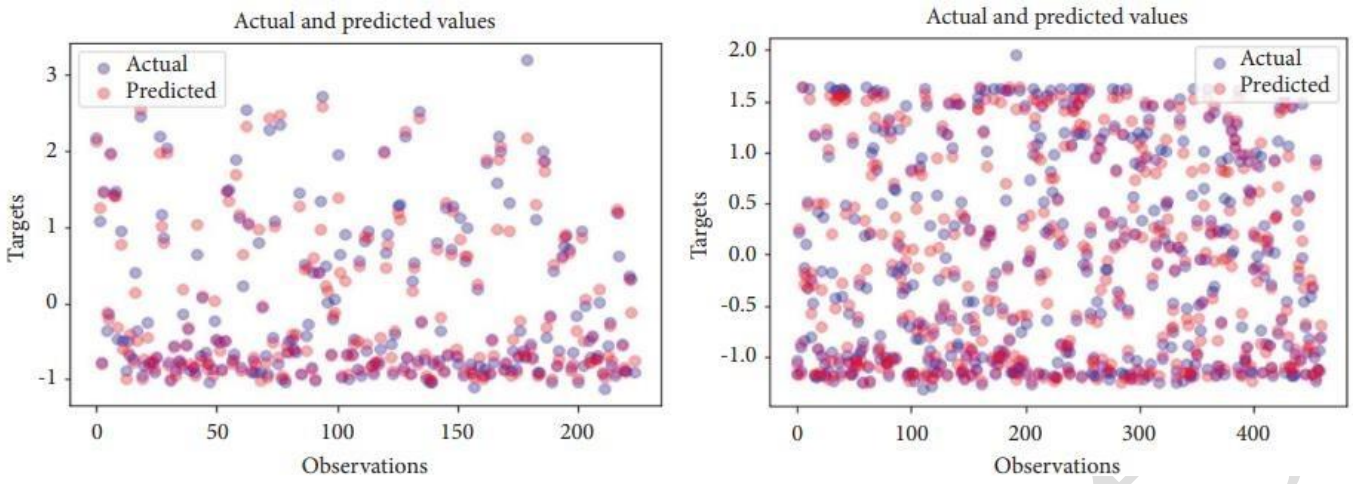


Figure 17: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for Kolkata-SVR.

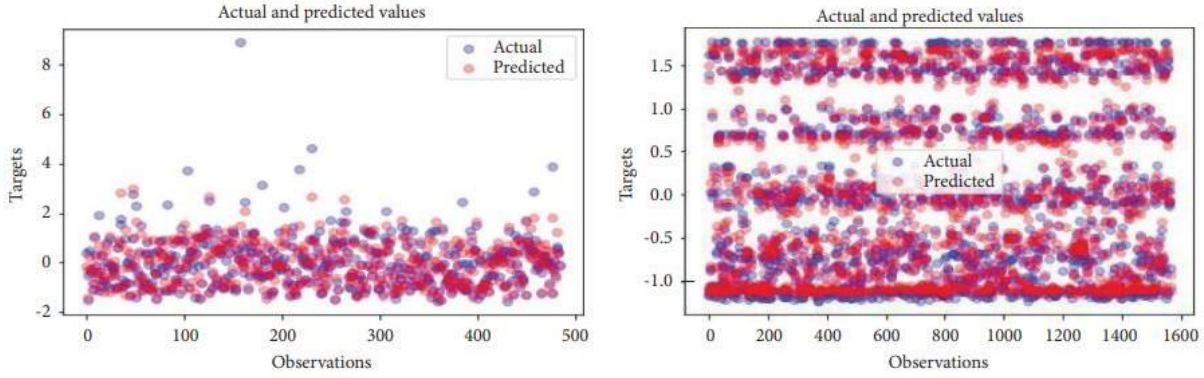


Figure 18: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for Hyderabad-SVR.

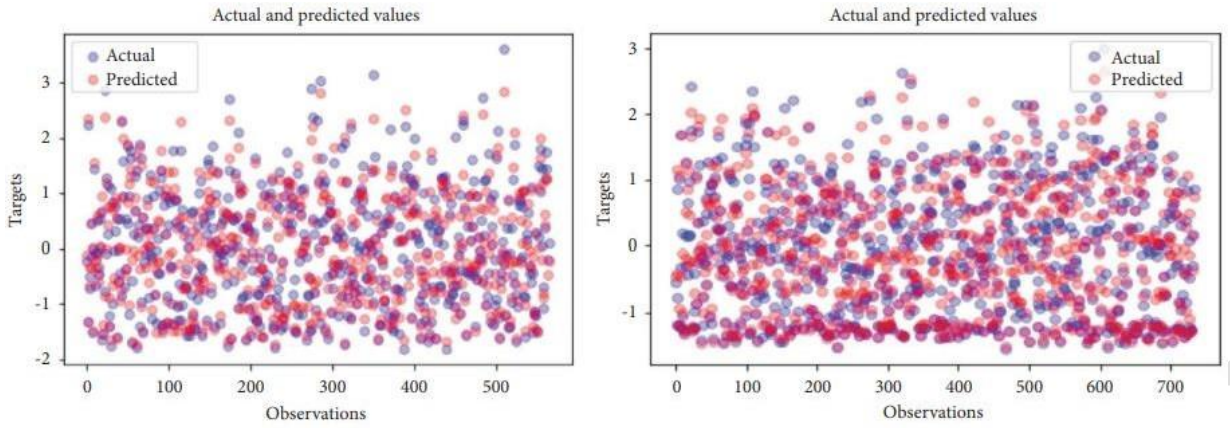


Figure 19: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for New Delhi-RFR.

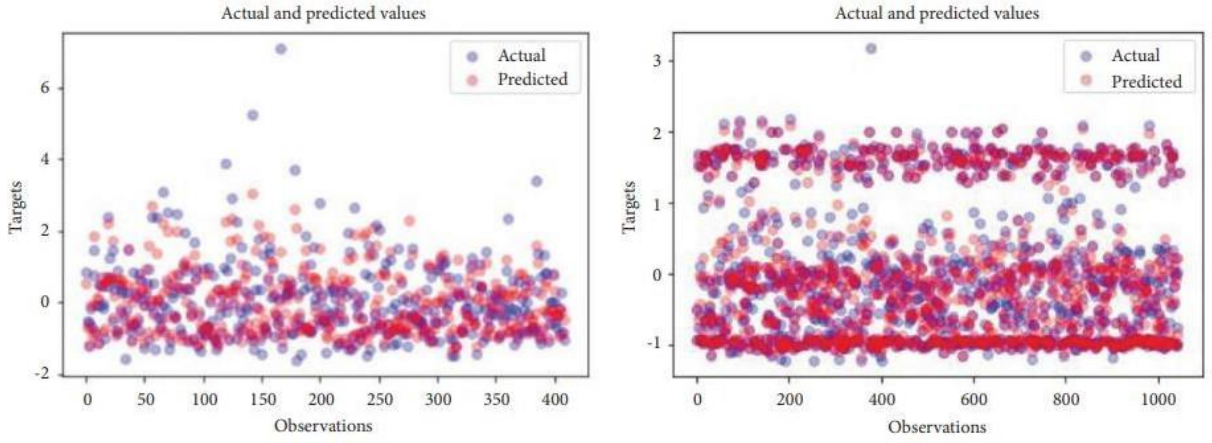
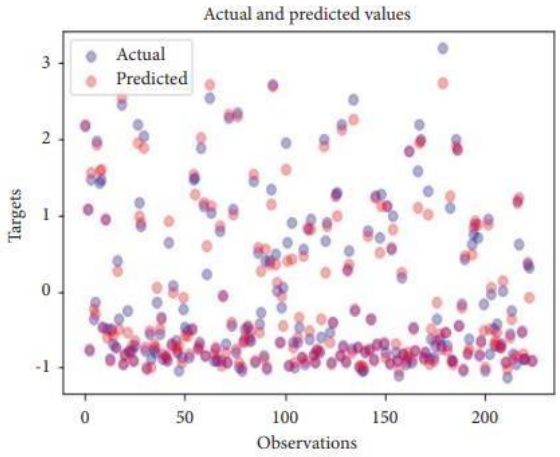


Figure 20: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for Bangalore-RFR



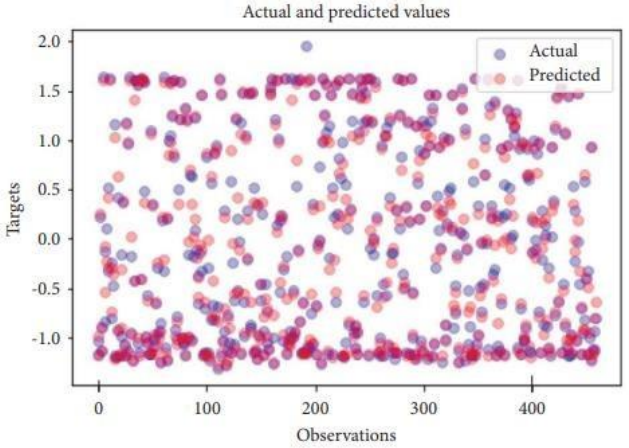


Figure 21: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for Kolkata–RF

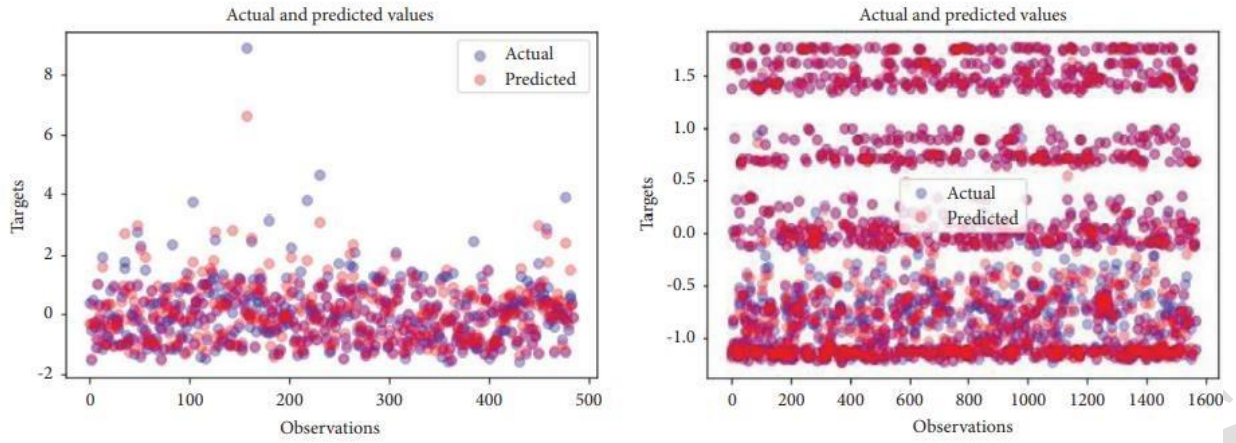


Figure 22: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for Hyderabad–RFR

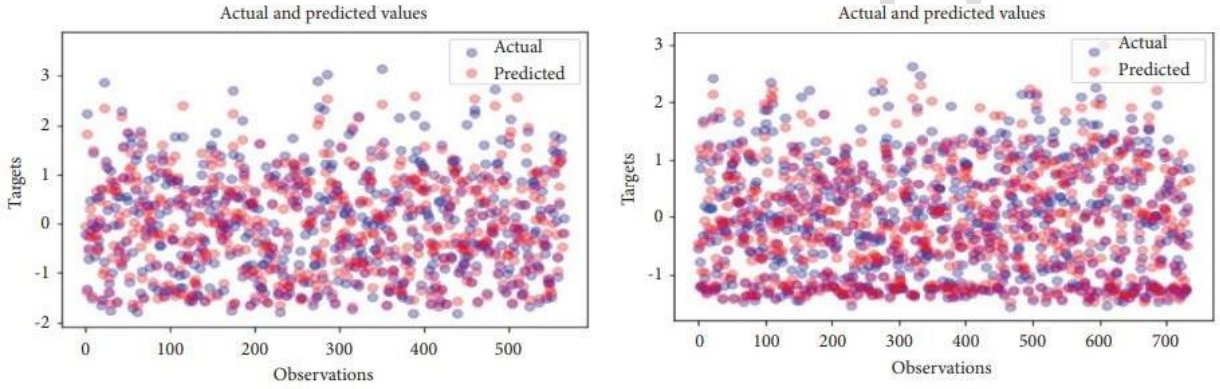


Figure 23: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for New Delhi–CR

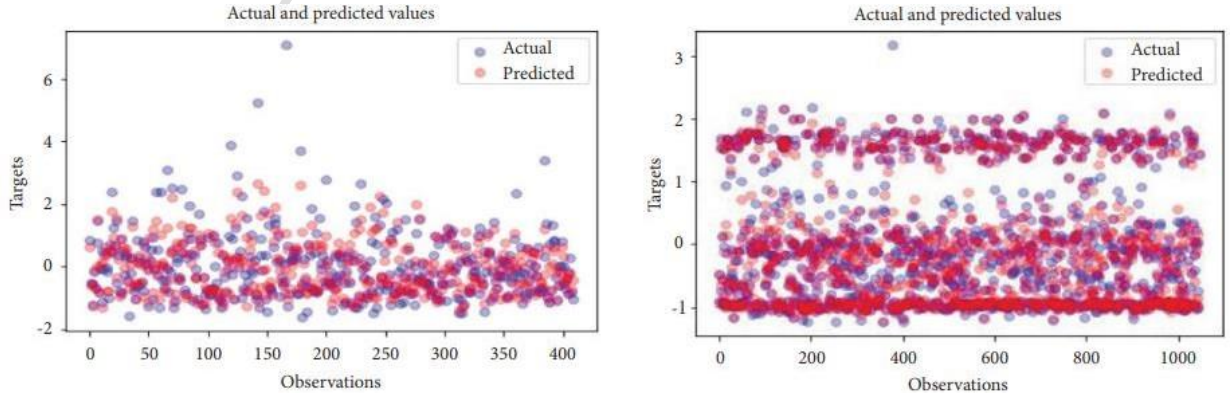


Figure 24: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for Bangalore–CR.

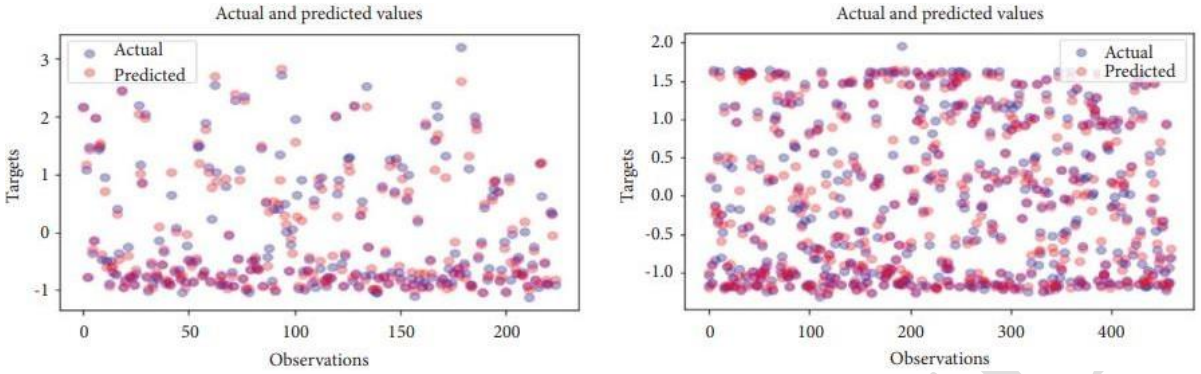


Figure 25: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for Kolkata-CR.

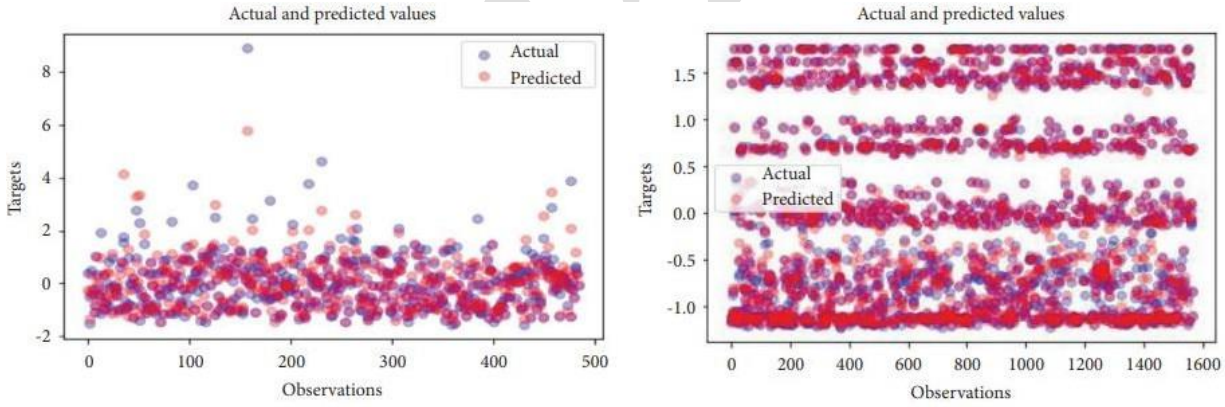


Figure 26: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for Hyderabad–CR.

TABLE 6:
Comparison of dataset size with and without the SMOTE algorithm

Imbalanced dataset size (not using the SMOTE				Balanced dataset size (using the SMOTE algorithm) algorithm)				
	Cities							
AQI_bucket values	Delhi	Bangalore	Kolkata	Hyderabad	Delhi	Bangalore	Kolkata	Hyderabad
	Size							
Moderate	485	479		806	485	958	302	806
Satisfactory	108	810	278	645	432	810	278	645
Good			119	126			238	1008
Poor	534	12	119	30	534	768	238	960
Very poor	514			20			264	768
Severe	239		13	0	478		208	1024

Table 7

Accuracy results comparison of the imbalanced dataset for four cities and methods used.

Method	Cities			
	New Delhi	Bangalore (%)	Kolkata (%)	Hyderabad (%)
	Accuracy (%)			
Support vector regression Random	78.4867	66.4564	89.1656	76.6786
forest regression	79.4764	67.7038	90.9700	78.3672
CatBoost regression	79.8622	68.6860	89.9766	77.8991

Table 8:

The result of performance metrics used for New Delhi city imbalanced dataset without using the SMOTE algorithm.

Algorithm name	R-square	MSE	RMSE	MAE
Support vector regression	0.9177	0.0908	0.3013	0.2151
Random forest regression	0.9265	0.0810	0.2846	0.2052
Cat Boost regression	0.9293	0.0779	0.2792	0.2013

Table 9:

The result of performance metrics used for Bangalore city imbalanced dataset, without using the SMOTE algorithm.

Algorithm name	R-square	MSE	RMSE	MAE
Support vector regression	0.6525	0.3772	0.6142	0.3354
Random forest regression	0.7035	0.3219	0.5674	0.3229
Cat Boost regression	0.6877	0.3391	0.5823	0.3131

TABLE 10:

The result of performance metrics used for Kolkata city imbalanced dataset, without using the SMOTE algorithm.

Algorithm name	R -square	MSE	RMSE	MAE
Support vector regression	0.9714	0.2942	0.1715	0.1083
Random forest regression	0.9808	0.0197	0.1403	0.0902
Cat Boost regression	0.9732	0.0235	0.1597	0.1002

Table 11:

The result of performance metrics used for the Hyderabad city imbalanced dataset, without using the SMOTE algorithm,

Algorithm	R-square	MSE	RMSE	MAE
Support vector regression	0.7399	0.2512	0.5012	0.2332
Random forest regression	0.8600	0.1464	0.3826	0.2163
Cat Boost regression	0 .8474	0.1596	0.3995	0.2210

Table 12:

Accuracy results comparison of the balanced dataset using SMOTE algorithm for four cities and methods used.

Method	Cities			
	New Delhi	Bangalore	Kolkata	Hyderabad
	Accuracy (%)			
Support vector regression (SVR)	84.8332	87.1756	91.5624	93.5658
Random forest regression (RFR)	84.7284	90.3071	93.7438	97.6080
Cat Boost regression (CR)	85.0847	90.3343	93.1656	96.7529

TABLE 13:

Comparison of SVR accuracy with and without SMOTE algorithm of four cities.

Cities	SVR accuracy (not using SMOTE algorithm-imbalanced dataset) (%)	SVR accuracy (using SMOTE algorithm-balanced dataset) (%)
New Delhi	78.4867	84.8332
Bangalore	66.4564	87.1756
Kolkata	89.1656	91.5624
Hyderabad	76.6786	93.5658

TABLE 14:

Comparison of RFR accuracy with and without the SMOTE algorithm of four cities.

Cities	RFR accuracy (not using SMOTE algorithm, imbalanced dataset)	RFR accuracy (using SMOTE algorithm, balanced dataset)
New Delhi	79.4764	84.7284
Bangalore	67.7038	90.3071
Kolkata	90.9700	93.7438
Hyderabad	78.3672	97.6080

Table15:

Comparison of CR accuracy with and without the SMOTE algorithm of four cities.

Cities	CR accuracy (not using SMOTE algorithm, imbalanced dataset)	CR accuracy (using SMOTE algorithm, balanced dataset)
New Delhi	79.8622	85.0847
Bangalore	68.6860	90.3343
Kolkata	89.9766	93.1656
Hyderabad	77.8991	96.7529

TABLE 16:

Overall comparison between accuracy values of the dataset with and without SMOTE algorithm Of four cities.

Method	Cities							
	Delhi	Bangalore	Kolkata	Hyderabad (%) Accuracy of the imbalanced dataset (without SMOTE algorithm) (%)	Delhi	Bangalore	Kolkata	Hyderabad Accuracy of the balanced dataset (with SMOTE algorithm) (%)
SVR	78.4867	66.4564	89.1656	76.6786	84.8332	87.1756	91.5624	93.5658
	79.4764	67.7038	90.9700	78.3672	84.7284	90.3071	93.7438	97.6080
Cat Boost	79.8622	68.6860	89.9766	77.8991	85.0847	90.3343	93.1656	96.7529

It is observed that when SMOTE is applied, the accuracy for New Delhi with SVR goes from 78.4867% to 84.8332%, with RFR it goes from 79.4764% to 84.7284%, and with CatBoost regression, it goes from 79.8622% to 85.0847%. In the Bangalore dataset again, it is noticed that once the SMOTE algorithm is applied to the dataset, those datasets help achieve that accuracies are considerably higher when models are applied to them than those with imbalanced datasets (without SMOTE). When SMOTE is applied, the accuracy for Bangalore with SVR goes from 66.4564% to 87.1756%, with RFR goes from 67.7038% to 90.3071%, and with CatBoost regression goes from 68.6860% to 90.3343%. It is noticed that when SMOTE is applied, accuracy for Kolkata with SVR jumps from 89.1656% to 91.5624%, with RFR from 90.9700% to 93.7438%, and with CatBoost Regression from 89.9766% to 93.1656%. To establish the trend more, even Hyderabad shows increased accuracies from models when SMOTE is applied, like when it is used with SVR, the accuracy goes from 76.6786% to 93.5658%, with RFR, 93.5658% to 97.6080%, and with CatBoost Regression, 77.8991% to 96.7529%. So, this gives quite a clear picture of the importance of balanced datasets. Having a dataset properly balanced can give more equal importance to each class. If there is too much of a gap between the number of values present for each class, it does not give an accurate portrayal of the actual scenario, and hence, the model fails. SMOTE creates multiple synthetic examples for the minority class and brings about a balance to the dataset. Tis makes the models work to the best of their ability, hence bringing better accuracy. Tis paper, hence makes clear about the importance of using SMOTE.

3.7.1 Database Module

The Database Module for air quality analysis is a critical component of any system or application that deals with the storage, management, and retrieval of air quality data. This module is responsible for handling various aspects related to data storage, including data acquisition, data storage, data retrieval, and data maintenance. Here are the key functionalities and components of a Database Module for air quality analysis:

[1] **Data Acquisition**: This component is responsible for collecting data from various sources, such as air quality monitoring stations, sensors, weather databases, or external data providers. It ensures that real-time or historical data is efficiently gathered and processed for analysis.

[2] **Data Storage**: The data storage component stores air quality data in a structured format. This may involve using relational databases (e.g., SQL databases like MySQL, PostgreSQL) or NoSQL databases (e.g., MongoDB, Cassandra) based on the specific requirements of the system**.**

[3] **Data Schema**: A well-defined data schema is essential to organize and store data accurately. This schema defines the structure of tables or collections in the database, specifying data types, relationships, and constraints.

[4] **Data Retrieval**: The database module provides methods for retrieving data based on user queries or system requirements. This includes querying historical data, retrieving real-time measurements, and aggregating data for analysis.

[5] **Data Maintenance**: Data maintenance involves tasks like data cleaning, data archiving, and ensuring data consistency. It also includes the removal of duplicate or outdated records and optimizing database performance.

[6] **Data Security**: Ensuring the security and privacy of air quality data is crucial. Implementing access controls, encryption, and authentication mechanisms to protect the data is a vital aspect of the module.

[7] **Scalability**: As the volume of air quality data may grow over time, the database module should be designed to scale horizontally or vertically to accommodate increasing data loads.

[8] **Data Integration**: This component facilitates the integration of data from various sources and formats. It ensures that data from multiple monitoring stations or sensors can be unified and stored in a consistent manner.

[9] **API for Access**: A well-defined API (Application Programming Interface) allows other system components, such as analysis modules or user interfaces, to interact with the database for data retrieval and storage.

[10] **Backup and Recovery**: Implementing regular data backups and a disaster recovery plan is essential to prevent data loss in case of hardware failure or other unforeseen events.

[11] **Data Logging**: Keeping a record of database operations, errors, and maintenance activities is crucial for monitoring and auditing purposes.

[12] **Data Quality Assurance**: Implementing data quality checks and validation mechanisms to ensure that the data stored in the database is accurate and reliable.

[13] **Reporting and Visualization Integration**: Integration with reporting and visualization modules to generate reports, graphs, and visual representations of air quality data for users and stakeholders.

3.7.2 Bot Module Management Module

The Bot Module Management Module for air quality analysis is responsible for handling automated or bot-based processes within the air quality analysis system. Bots in this context are typically software programs or scripts that perform specific tasks autonomously or semi-autonomously. These bots can be used for various purposes, such as data retrieval, data processing, notifications, and interactions with users. Here are the key functionalities and components of the Bot Module Management Module:

Bot Configuration: This component allows administrators or users to configure and manage bots. It includes settings for bot behavior, triggers, and scheduling.

Bot Development and Integration: The module may provide tools for developing and integrating custom bots. This can include scripting interfaces, APIs, or graphical bot development environments.

Bot Execution: Bots are executed to perform specific tasks or workflows. The module manages the execution of bots, including task scheduling and monitoring.

Data Retrieval Bots: Some bots may be responsible for retrieving air quality data from external sources, such as monitoring stations, weather services, or government databases. These bots can automate the process of data acquisition.

The Bot Module Management Module enhances the capabilities of the air quality analysis system by automating tasks, streamlining processes, and enabling interactions with users and external data sources. Proper management and configuration of bots are essential for optimizing the system's performance and efficiency

3.7.3 Music Recommendation Module

The "Better Air Quality Place Recommendation Module" is a component within an air quality analysis system that provides recommendations for locations or places with better air quality. This module is designed to help users, residents, or travelers make informed decisions about where to go based on air quality data. It can be a valuable tool for individuals seeking to avoid areas with poor air quality, especially for those with respiratory conditions or health concerns. Here are the key functionalities and components of this module:

Air Quality Data Integration: The module integrates real-time or historical air quality data from various sources, including monitoring stations, sensors, and government databases.

Location-Based Services: It utilizes location-based services and data to determine the user's current or intended location.

User Preferences: Users may have preferences or criteria for air quality, such as acceptable pollutant levels, air quality index (AQI) categories, or specific health considerations.

Data Analysis and Processing: The module processes the air quality data to evaluate the air quality at different locations. This may involve data aggregation, normalization, and quality assessment.

3.8 Module Description

Following are the main Modules of this Music Application

1. Login Module - At Admin Side
2. Process Module- At Admin Side
3. Database Module - AT Admin Side
4. Bot Module Management Module - At Admin Side
5. Music Recommendation Module – At Admin side (Database)
6. Music Genre Identification Module – At Admin Side (ML)
7. User Management Module - At Admin Side
8. Register Module– At User Side
9. Login Module – At User Side

Features:

Air quality analysis in machine learning encompasses several key features that enable comprehensive and accurate assessments of air quality parameters. These features include data integration from various sources such as monitoring stations, IoT sensors, and satellite imagery, facilitating a holistic understanding of air quality dynamics. Machine learning models utilize advanced algorithms to process and analyze this data, enabling the detection of complex patterns and trends in air pollution. Additionally, the integration of meteorological and geographical data allows for the contextualization of air quality assessments, considering weather patterns and geographical influe$_{4}$n$_{6}$ces. Real-time monitoring capabilities enable prompt

alerts and notifications, ensuring timely interventions and preventive measures. Furthermore, the incorporation of explainable AI (XAI) techniques enhances the interpretability of results, fostering transparency and trust in the analysis outcomes. By leveraging these features, air quality analysis in machine learning offers a robust framework for understanding, monitoring, and mitigating the impacts of air pollution on public health and the environment**.**

Chapter 4

Results and Discussion

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. Te present study assessed the performance of the three best data mining models (SVR, RFR, and CR) for predicting the accurate AQI data in some of India's most populous and polluted cities. The synthetic minority oversampling technique (SMOTE) was used to equalize the class data to get better and consistent results.

Chapter 5

CONCLUSION AND FUTURE ENHANCEMENT

5.1 Conclusion

In this study, machine learning techniques were employed to forecast accurate Air Quality Index (AQI) data in highly populated and polluted cities in India. The research compared the performance of three prominent data mining models (SVR, RFR, and CR) and introduced a unique approach of balancing datasets using the synthetic minority oversampling technique (SMOTE) to improve results.

The findings revealed a substantial increase in accuracy when using balanced datasets compared to unbalanced ones, with accuracy levels reaching as high as 97.6% in some cities. Random forest regression and CatBoost regression consistently outperformed SVR, both before and after applying SMOTE.

Various metrics, including R-SQUARE, MSE, MAE, and RMSE, were used to evaluate the models' performance, and the results indicated that CatBoost and random forest regression, in conjunction with SMOTE, offered the most promising outcomes for estimating air quality. This innovative approach holds significant potential for informing local and national authorities about air quality conditions, thereby prompting regulatory actions to improve air quality in highly polluted regions.

5.2 Future Enhancement

The field of air quality analysis using machine learning is continually evolving, and there are several areas where future enhancements and developments can be anticipated. Some of the potential avenues for improvement and innovation include:

The proliferation of Internet of Things (IoT) devices, such as low-cost air quality sensors, can enhance data collection and coverage. Integrating data from these devices into machine learning models can provide real-time, high-resolution air quality insights. Machine learning models can benefit from finer spatial resolution, allowing for more localized and precise air quality predictions. This can be achieved by incorporating data from a denser network of monitoring stations or IoT sensors. Combining air quality data with other environmental, meteorological, and geospatial data can lead to more comprehensive analyses. Machine learning models that fuse data from multiple sources can provide a holistic view of air quality determinants. The development of more advanced sensor technologies, such as hyperspectral imaging and remote sensing, can offer new ways to assess air quality from a distance and provide data for machine learning models.

REFERENCES

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

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

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

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

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

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

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

[8] J. Kleine Deters, R. Zalakeviciute, M. Gonzalez, and Y. Rybarczyk, “Modeling PM2. 5 urban pollution using machine learning and selected meteorological parameters,” Journal of Electrical and Computer Engineering, vol. 2017, Article ID 5106045, 14 pages, 2017.

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

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

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

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

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

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

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

[16] 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.

[17] R. Janarthanan, P. Partheeban, K. Somasundaram, and P Navin Elamparithi, “A deep learning approach for prediction of air quality index in a metropolitan city,” Sustainable Cities and Society, vol. 67, no. 2021, Article ID 102720, 2021.

[18] M. Londhe, “Data mining and machine learning approach for air quality index prediction,” International Journal of Engineering and Applied Physics, vol. 1, no. 2, pp. 136–153, May 2023