AIR QUALITY MONITORING AND ANALYSIS BASED PREDICTIVE SYSTEM

2024-078

Project Proposal Report

Mohamed Ismail Mohamed Inthikhaff

B.Sc. (Hons) Degree in Information Technology Specialized in Information Technology

Department of Information Technology

Sri Lanka Institute of Information Technology
Sri Lanka

Auguest 2024

AIR QUALITY MONITORING AND ANALYSIS BASED PREDICTIVE SYSTEM

2024-078

Project Proposal Report

Mohamed Ismail Mohamed Inthikhaff – IT21058028

Supervisor: Chathurangika Kahandawaarachchi

B.Sc. (Hons) Degree in Information Technology Specialized in Information Technology

Department of Information Technology

Sri Lanka Institute of Information Technology
Sri Lanka

Auguest 2024

DECLARATON

I declare that this is my own work, and this proposal does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of our knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. Also, I hereby grant to Sri Lanka Institute of Information Technology, the nonexclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Name	**Student ID**	**Signature**
M. I. M Inthikhaff	IT21058028	

The supervisors should certify the proposal report with the following declaration.

The above candidates are carrying out research for the undergraduate Dissertation under my supervision.

Signature of the Supervisor   Date
(Ms. Chathurangika Kahandawaarachchi)

…………………………….   ..……………………….

ABSTRACT

Generally, air pollution is the release of substances into the air that are harmful to human beings and the earth as a whole. It can be said to be one of worst threats faced by humanity. It harms animals, crops, forests etc. In Sri Lanka's cities such as Kandy and Colombo, air pollution effects are dangerous to daily life hence endangering the well-being of the population. To avoid this problem in transport sectors have predicted pollutants from machine learning techniques for air quality. As such, air quality evaluation and prediction has become an important area of study. The intention is to examine machine learning based methods for predicting air quality with focus on best accuracy predictions using forecasting outcomes. The entire provided dataset will be subjected to analysis of variables identification, uni-variate analysis, bi- variate and multi-variate analysis, missing value treatments and analyze validation data cleaning/preparing and data visualization by supervised machine learning technique (SMLT). Our study provides a comprehensive guide on how model parameters vary with performance in terms of prediction of pollution levels by accuracy calculation. To propose a Machine Learning Based method which accurately predicts Air Quality Index (AQI) value through super vised classification machine learning algorithms comparison results in terms of best accuracy from prediction results. Furthermore, we aim to compare and analyze the effectiveness of machine learning algorithms using data from the Central Environmental Authority (CEA) and National Building Resource Organization (NBRO) datasets. Our goal is to create a prediction system through a user graphical interface for forecasting air quality based on various attributes. The objective is not to assess air quality but also to predict future levels ultimately improving living conditions, for the people of Sri Lanka and reducing health issues caused by air pollution.

Keywords: Machine Learning, Ambient Air Quality Index, Model, Correlation , Graphical User Interface

ACKNOWLEDGEMENT

First and foremost, it is a genuine pleasure to express my deep sense of gratitude to my Research Supervisors, Ms. Chathurangika Kahandawaarachchi and Ms. Pipuni Wijesiri for her fullest support, dedication, enthusiasm, and overwhelming desire his expressed throughout the project to make this research a success. I gratefully acknowledge his patience at the time of long piece of writing this report.

My sincere thanks go to seniors who helped to do this research well.

Finally, Thanks for everyone who support me in various way to convert my project to successful project.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

PM	Particulate Matters
CEA	Central Environmental Authority
NBRO	National Building Resource Organization
AQI	Air Quality Indexing
API	Application Program Interface
DB	Database
AI	Artificial Intelligence
IT	Information Technology

1. INTRODUCTION

1.1 Background Literature

Poor air quality affects about two thirds of the world's population. Air quality, or the quality of the air we breathe, is a highly valued commodity that everyone should be able to buy. The majority of people on the planet live in heavily air-polluted urban areas—over half, according to the most recent reports from the World Health Organization—but almost none of them are aware of how clean the air is where they live. The quality of the air in Colombo indicates that air pollution in the city is increasing at a very rapid annual rate. In the rush hours of the morning and evening, The Colombo Air Quality Index is higher above the 4th category standard set by the World Health Organization. Sadly, most people don't know what the local air quality index (AQI) level is, and as a result, especially in big cities like Colombo, they've become indifferent to the possible health risks that come with air pollution [1].

This air pollution level prediction mainly focuses on air quality index prediction using various machine learning algorithms such as regression, support vector machines, k-nearest neighbor, and random forest to predict the air quality index. Many machine learning models have been tried and implemented in order to identify the best accurate model. Improving methods for forecasting the air quality index, deepening our knowledge of the index, and appreciating the effects of low air quality are the main objectives of this study.

In Colombo, the average unhealthy level is always the AQI score. With the aid of some reliable and precise AQI category predictions, the people of Colombo can take the necessary preventive measures, such as increasing indoor activities and decreasing outside activities, to protect themselves from the detrimental impacts of the city's poor air quality. This chapter begins with a brief explanation of the AQI, the rationale behind the inquiry, and some background data on the state of the air in Colombo before presenting the research's goal [2].

We must take action to lessen air pollution and ensure that everyone has access to clean air because human activity has led to an increase in global air pollution. The main causes of air pollution on Earth are human-caused revolutions like industrialization. The rising number of cars and other gear that emits carbon dioxide into the sky is another factor causing air pollution. Broadly speaking, air pollution is the result of large-scale releases of dangerous gases and chemicals into the atmosphere that contaminate the air in a particular geographic area. Emissions from production, transportation, and industry are the main sources of poor air quality. When compared to other rural areas, densely populated urban places like Colombo usually have the lowest air quality, mostly because of human activity. Threats to human health, including both immediate and long-term adverse effects on the health of living things and their surroundings, are strongly correlated with the value of the air quality index. When exposed to contaminated air, those who have previously had illnesses like asthma and pneumonia are more likely to develop heart- and lung-related conditions.

It has been noted that after breathing in contaminated air, the human immune system finds it extremely difficult, if not practically impossible, to self-purify PM2.5 and PM10 particles that have entered the body [3]. As several studies have shown, the main problem in Colombo is that people are unaware of the bad quality of the air. As a result, the air quality index (AQI) has become increasingly important in predicting and educating the public about the possible effects that air pollution (AQI levels) may have on human health and the environment, especially in light of the recent substantial increase in air pollution [4]. Furthermore, the negative consequences of air pollution on the ecosystem and human health can be significantly reduced or eliminated by anticipating the AQI value. Finding the best approach to predict the AQI value among the many approaches and the dataset that can be utilized to provide the most accurate forecast are two of the study's main accomplishments. The basis of human existence is air.

The Air Quality Index is the index value that most accurately represents the state of the air right now. The atmosphere and the wellbeing of all living things will be even more at risk from a substantial AQI. For this study, the Central Environment (CEA) Authority provided a historical air concentration dataset with hourly concentration levels of various air pollution factors & weather factors, such as PM10, PM2.5, SO2, NO2, CO, O3, wind speed, wind direction, average temperature, relative humidity, and solar radiation.

Current research on variations in air quality indicates that there have been some recent trends in Battaramulla and Kandy that negatively impact human life. The data indicates that the recent infrastructure and transportation work has had a major influence in both locations. Tiny particles are added to the local atmosphere as a result. As a result, it is evident that PM10PM10 is a significant factor that brings up several concerns that have a detrimental effect on quality of life.

This study was initiated as a first step in aiding the process of preserving air quality for the proper level of both natural and human quality of life. In preparation for a review of earlier studies on air quality forecasting, this work is the first to provide a comparative analysis of PM10 forecasting utilizing cutting-edge models [5].

Thus, this study analyzes the analysis of several machine learning models to forecast the PM10 concentration of two urban sites, namely Battaramulla in Colombo and Kandy. The concentration levels of PM10 were predicted using meteorological variables such as ambient temperature, relative humidity, solar radiation, rainfall, wind speed, and wind direction in addition to other air pollution components like the concentrations of O3, CO, NO2, SO2, and PM2.5. The comparison analysis was carried out by evaluating a number of performance indicators, including Mean Absolute Relative Error (MARE), coefficient of determination (R2), mean absolute error (MAE), mean squared error (MSE), and relative absolute error (RAE). This study thus extends our understanding of machine learning techniques for PM10 concentration forecasting with an emphasis on Sri Lanka (Mampitiya et al., 2023). The outcomes would be taken into consideration for formulating policies that ensure greater living conditions in urban areas [5].

The method for developing a machine-learning technique for PM10 value forecasting was suggested by this study. To put it briefly, this study uses eight machine learning models in conjunction with a special technique. The extensive literature review makes it evident that a specific methodology is needed for this study of environmental factors because a variety of factors are involved. This study recommended appropriate methods for processing data, such as data cleaning, correlation detection, and the use of machine learning models with loop-based hyperparameter optimization, which should be followed by a comprehensive comparative analysis.

Numerous techniques for data preparation were employed in order to ensure the accuracy of the forecast outcome. Using 20% of the preprocessed dataset for testing and the remaining 80% for model training allowed the cross-validation method to be used. Some of the machine learning approaches that have been applied as prediction models include Support Vector Machine, Random Forest, K Nearest Neighbors, and Multiple Linear Regression. On the basis of accuracy and performance, the optimal machine learning model for AQI prediction is selected.

'Air Pollution Level Prediction and Air Quality Indexing in Colombo' is becoming more and more necessary, according to the readings described above, in order to fill in the gaps in the current processes and systems. Predicting air pollution levels is beneficial for schoolchildren, those who are impacted by air pollution, extension workers, and others since it allows for real-time notifications when the Air Quality Index (AQI) is high. Moreover, the capacity to visualize air pollution levels through the comparison of past, present, and future situations is beneficial and provides a more intelligent way to deal with current problems.

1.2 Research Gap

For AQI prediction, there are several choices. Few of these forecasts, which depend on the atmospheric concentrations of PM10 and PM2.5 fine particles, provide comprehensive suggestions for reducing the detrimental impacts of air pollution on human health. A detailed examination of the literature was carried out in order to provide a concise overview of the research on AQI predictions made using different machine learning models, as there was no precise approach for AQI prediction within the study area. The study topic differs not only in terms of methodologies and strategies but also in terms of the datasets that are accessible, many of which are unique because of the traffic, climate, and environmental factors of the selected geographic area. The primary source of poor air quality is PM particles. For example, air pollutants like PM2.5 and PM10 are the primary cause of air pollution in certain cities, while COx, SOx, and NOx are the main causes of air pollution in other places [6].

Due to these limitations, a thorough review of the literature is done for this study in an effort to comprehend the breadth of AQI prediction research and identify relevant studies that achieve the same objective as this one. The literature review section of this paper compiles the most recent and relevant studies on AQI predictions. Some of the most recent techniques for AQI prediction are looked at here.

It is essential to consider the best approach for air pollution forecasting. The deep learning approach is one of the most often used techniques among the models that are currently in use for AQI prediction. The technique used most commonly to forecast AQI is machine learning. Machine learning techniques use large data sets and algorithms based on machine learning to train the model. Neural network methodology based on deep learning is another approach to AQI forecasting. A basic neural network can be used to provide accurate AQI predictions. The model can then be further tuned by varying the testing settings and input parameters [7].

The literature reviewed for this work highlights a few flaws in the methods currently used to forecast the air quality index, such as issues with dataset collecting. The low forecast accuracy of the air quality index prediction models that are currently available in Sri Lanka is caused by incomplete and inaccurate data. Data preparation is another element that makes the accuracy decline. Considering these elements, it is evident that the majority of Sri Lanka's present systems are unable to generate accurate projections. In light of the aforementioned in relation to other countries, they have succeeded in overcoming these obstacles and achieving exceptional precision.

Table 1.1 Summary of Literature review

Author	**Application**	**Techniques**	**Remark**
Min Lee and others	Air pollution prediction	• Deep Learning	• Predict against PM 2.5, PM 10 particulars. • Accuracy based on PM 10 is very low. • Accuracy based on PM2.5 is very high.
Usha Mahalingam and others	Air quality prediction	• Neural Networks • Support Vector Machine	• Accuracy of 91.62% for neural network. • Accuracy of 97.3% for support vector machine
S. Silva and others	Air quality prediction for smart cities	• Support Vector Regression	• Predict PM 2.5 levels variability. • Model is suitable for predict hourly air pollution. • Obtain an accuracy of 94.1%
Timothy M. A. and others	Development of Air Quality Monitoring model	• Naive Bayesian • KNN • Support Vector Machines • Neural Networks • Random Forest	• Highest accuracy was obtained through Neural Networks. • Sometimes Neural Network leads slower response.
C. Zhao and others	Air Quality Index Prediction	• Linear regression	• AQI Prediction based on a year data of PM2.5, PM10 etc. • There is a deviation between predicted results and actual date.
Ismail Ahmadi	Air pollution prediction	• Data mining • Decision Tree	• Used clementine software for data clustering. • Data sample include climate data of 53 years.

Colin Belinger and others	A systematic review based on Machine Learning and data mining for Air Pollution.	• Machine Learning Algorithms • Data Mining • Big Data	• Refer 400 research papers & Reduce to 47 after the inclusion/exclusion criteria's. • Divided research papers into three categories • End of the Literature survey that highest accuracy levels always obtain in Machine Learning Algorithms based approaches.

1.3 Research Problem

A major gap in the field of air quality prediction in Sri Lanka is the focus of the research challenge addressed in this work. This gap is particularly related to shortcomings in data preprocessing techniques and the general accuracy of the current Air Quality Index (AQI) prediction systems. The problems include shortcomings in the way that data preprocessing is currently done, such as how to deal with incomplete and erroneous data, which has a big effect on how reliable AQI forecasts are. Furthermore, difficulties in obtaining data pose a barrier to the high accuracy that the current prediction systems are capable of. Solving this research issue is critical to improving the accuracy of AQI forecasts, which is necessary for efficient air quality control. Additionally, in order to ensure accuracy on par with worldwide benchmarks, it is imperative that the current systems be improved in order to bring Sri Lanka's AQI prediction algorithms into compliance with international standards. It becomes necessary to close this gap in order to create reliable air quality forecasts that benefit human health as well as the environment. The main objective of the research is to create more sophisticated data preprocessing techniques and enhance current AQI prediction algorithms in order to provide a more sustainable and healthful living environment in Sri Lanka.

The aim is to create a machine learning model that can forecast air quality in time potentially replacing the updatable supervised machine learning classification models by predicting results, with the highest accuracy when compared to supervised algorithms. Description of the Issue/Problem Statement; Monitoring and maintaining air quality has become a task in industrial and urban areas today. Air quality is negatively impacted by sources of pollution such as transportation, electricity and fuel usage. The accumulation of gases poses a threat to the quality of life in smart cities. As air pollution continues to rise there is a need, for air quality monitoring models that can gather data efficiently.

2. OBJECTIVES

2.1 Main Objectives

This component's main goal is to provide a reliable model for estimating pollution levels. To effectively anticipate the concentration of contaminants in the air, this requires examining historical data, present environmental conditions, and other pertinent elements. By means of a thorough investigation, the project team determines which variables have the most influence. Then, feature engineering methods are implemented to improve the predictive power of the model. To extract more useful information and raise the predictive model's overall accuracy, this entails modifying and adjusting the selected features. Dynamic predictions are a novel element of the predictive model that allows for real-time modifications based on shifting environmental parameters.

The objective of this study is to investigate a dataset of air pollutants records for Colombo meteorological sector using machine learning technique. To identifying air quality is more difficult. We try to reduce this risk factor behind predicting from Air Quality Index (AQI) of Colombo to safe human so as to save lots of meteorological efforts and assets and to predict whether assigning the air quality is bad or good.

2.2 Specific Objectives

1. Identification of dataset, cleaning, and preprocessing:

    - Gather and verify relevant datasets containing air pollutant records for the Colombo meteorological sector.
    - Conduct thorough data cleaning processes to address missing values, outliers, and inconsistencies, ensuring data integrity and reliability for subsequent analysis.

2. Training dataset using different models:

    - Employ various machine learning algorithms such as linear regression, decision trees, and neural networks to train on the preprocessed dataset.
    - Utilize cross-validation techniques to assess model performance and generalization ability, ensuring robustness and reliability in predicting air pollution levels.

3. Finding the best model which gives high-level accuracy:

    - Evaluate the performance metrics of trained models, such as accuracy, precision, recall, and F1-score, to identify the model with the highest predictive accuracy.
    - Fine-tune hyperparameters and optimize model architectures to enhance performance and achieve the desired level of accuracy in predicting air quality levels.

4. Data visualization and comparison of the degree of air pollution in Colombo areas:

    - Develop interactive visualization tools to display air pollution levels across different regions of Colombo, facilitating easy comparison and interpretation.
    - Implement features for authorized users to access historical, current, and predicted air quality data, enabling informed decision-making and resource allocation to mitigate pollution risks.

3. METHODOLOGY

3.1 System Architecture Diagram

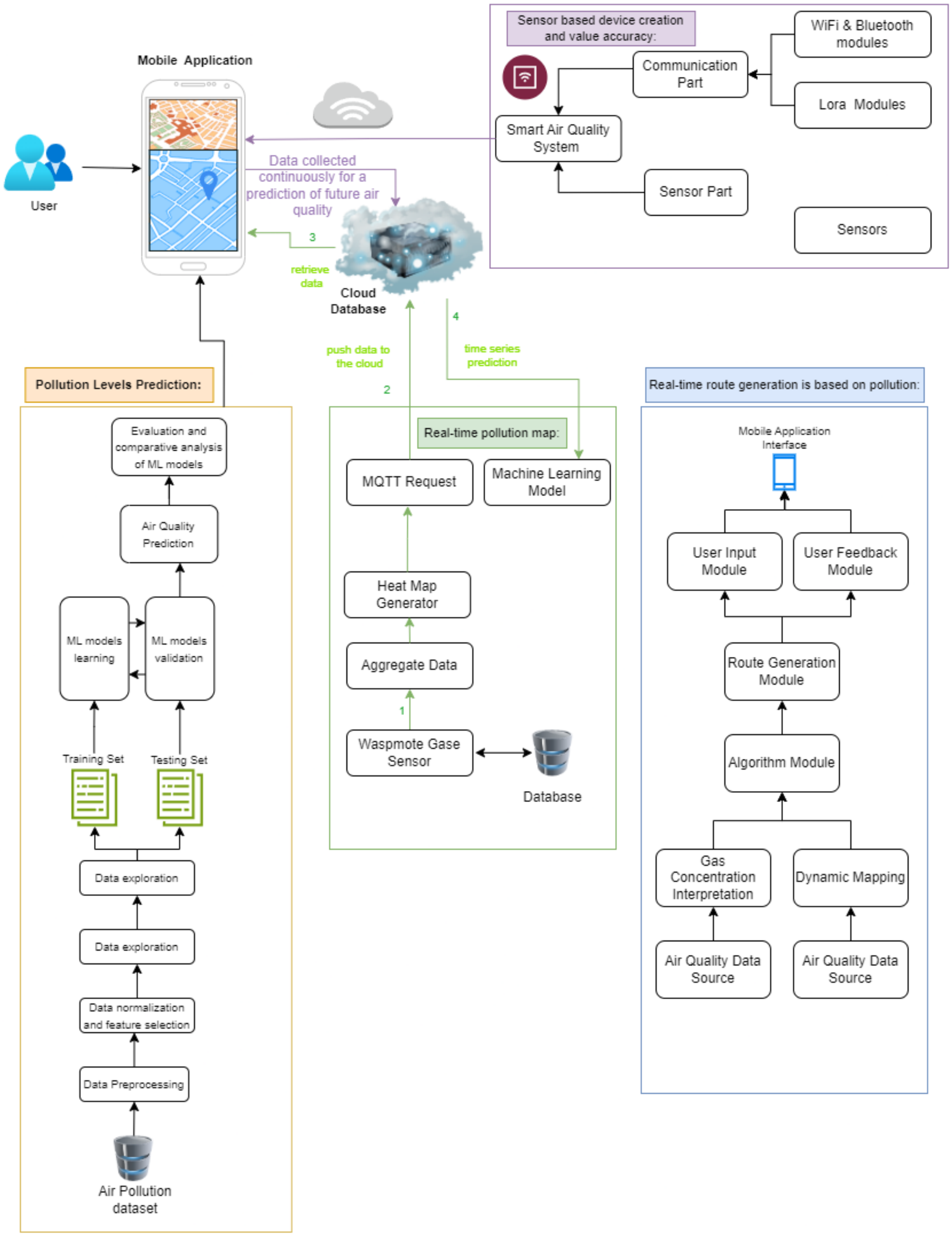


Figure 3. 1 System Architecture Diagram

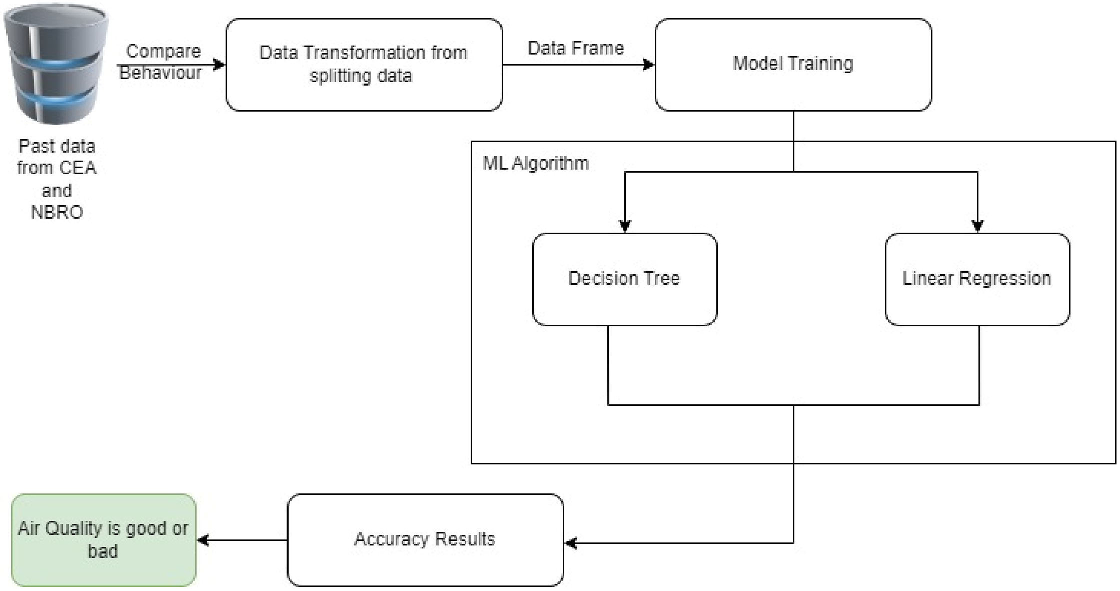


Figure 3. 2 Pollution level Prediction Architecture Diagram

The suggested method for estimating the amount of air pollution is a methodical procedure. The Central Environmental Authority is the original source of the dataset, which is carefully pre-processed to guarantee data integrity. The investigation then applies a variety of machine learning techniques with the goal of revealing relationships between the numerous components that contribute to air pollution. Choosing the best machine learning approach that produces precise predictions is the main objective. This methodical approach is in keeping with the system architecture, which outlines the internal parts that come together to create the finished system. The architecture shows how the system's many components communicate with one another, offering insight into how the system operates. Within this all-encompassing system structure, the machine learning algorithms that have been developed are essential in producing accurate findings that indicate whether the air quality is classified as good or bad.

The primary objective of this component is to develop a robust model for predicting pollution levels. This involves analyzing historical data, current environmental conditions, and other relevant factors to accurately forecast the concentration of pollutants in the air. Through rigorous analysis, the project team pinpoints the variables with the most substantial impact. Subsequently, feature engineering techniques are applied to enhance the model's predictive capabilities. This involves transforming and manipulating the chosen features to extract more valuable information and improve the overall accuracy of the predictive model. The original air pollution data, including variables such as CH4, CO, CO2, NO2, and SO2, are sourced from the passive sampling stations of the National Building Research Organization and Central Environmental Authority. These data are collected on an hourly basis over a period of 5 years. However, there are some missing data in the original dataset due to power failures, A/C failures, system failures, and calibration failures. This type of missing data is categorized as Missing Completely At Random (MCAR), meaning the probability of a missing value depends neither on the variable itself nor on another variable in the database. To improve prediction accuracy, the obtained dataset was pre-processed. The dataset was then cross validated, with 80% used for training and 20% for testing the prediction model.

Class	**Test**	**Train**
CO2 (5702)	4562	1140
PM2.5 (3057)	2446	611
PM10 (5703)	4563	1140

Table 2.1 Summary of Data Set Train and Test

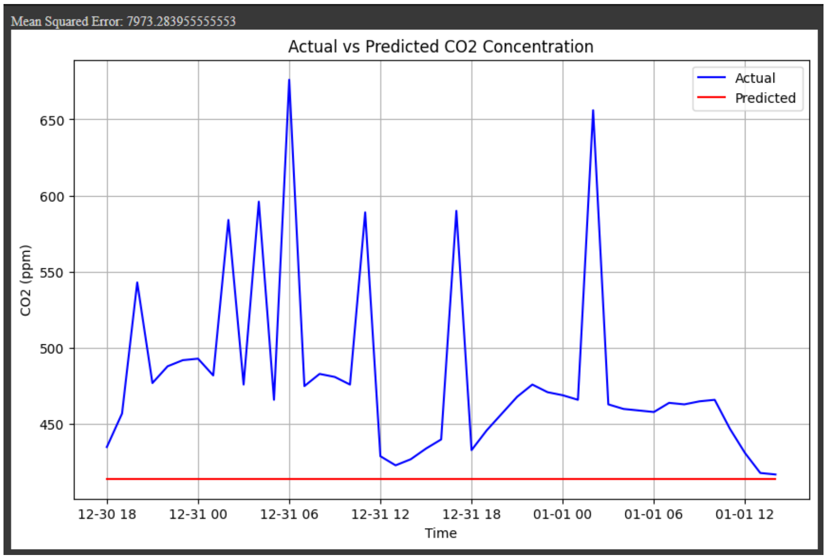
In the model development process, the first step is time series identification by plotting sample autocorrelations (SAC) and sample partial autocorrelations (SPAC) based on the original data. The stationarity of the data is then identified, and outliers are removed. Observations from correlograms indicate that more spikes are found in the regular differencing and seasonal differencing. As a result, regular and seasonal differ ences are chosen for the model development. The suggested method for estimating the amount of air pollution follows a methodical procedure, with the dataset carefully preprocessed to ensure data integrity.

The research applies various machine learning techniques to reveal relationships between the components contributing to air pollution. The main objective is to choose the best machine learning approach that produces precise predictions. This systematic approach aligns with the system architecture, which outlines the internal components that come together to form the final system. The architecture demonstrates how the system's various components communicate with one an other, providing insight into the system's operation. Within this comprehensive system structure, the developed machine learning algorithms are crucial for producing accurate results, indicating whether the air quality is classified as good or bad.

CO2 Prediction Models

In the prediction of CO2 levels, four different models were employed: Random Forest, Linear Regression, XGBoost, and Prophet. Each model was evaluated based on its Mean Squared Error (MSE), which serves as a metric to determine the accuracy of the model's predictions.

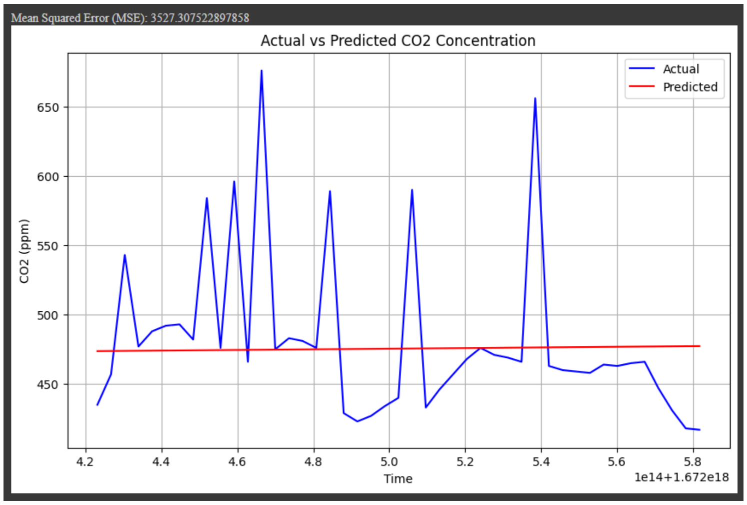
Random Forest is an ensemble learning method that operates by constructing multiple decision trees and aggregating their predictions. Despite its robustness and general high performance, the Random Forest model yielded an MSE of 7973.28, indicating that it was not the most accurate in predicting CO2 levels.



Mean Squared Error: 7973.283955555553

Figure 3.3: Random Forest Model for Predicting CO_2 Levels

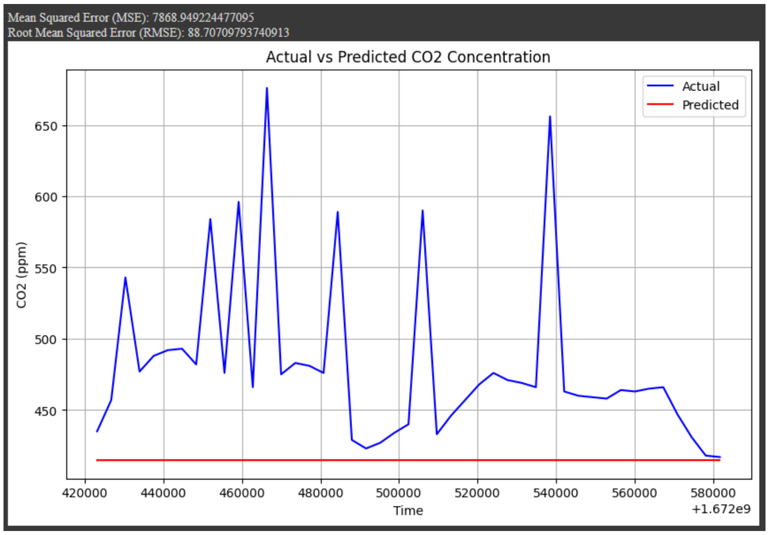
Linear Regression is a widely used model that attempts to capture the linear relationship between the input variables and the target variable. In this study, the Linear Regression model showed better performance than Random Forest, with a lower MSE of 3527.31.



```
Mean Squared Error (MSE): 3527.307522897858
```

Figure 3.4: Linear Regression Model for Predicting CO₂ Levels

XGBoost (eXtreme Gradient Boosting) is a more sophisticated model known for its efficiency and accuracy in handling complex datasets. This model performed significantly better than both Random Forest and Linear Regression, achieving an MSE of 88.71, which suggests a much higher accuracy in predicting CO2 levels.



```
Mean Squared Error (MSE): 7868.949224477095
Root Mean Squared Error (RMSE): 88.70709793740913
```

Figure 3.5: XGBhoost Model for Predicting CO₂ Levels

Finally, **Prophet**, a model designed for time series forecasting, particularly excelled in this task. It recorded the lowest MSE of 54.45 among all the models, making it the most accurate model for predicting CO2 levels in this study.

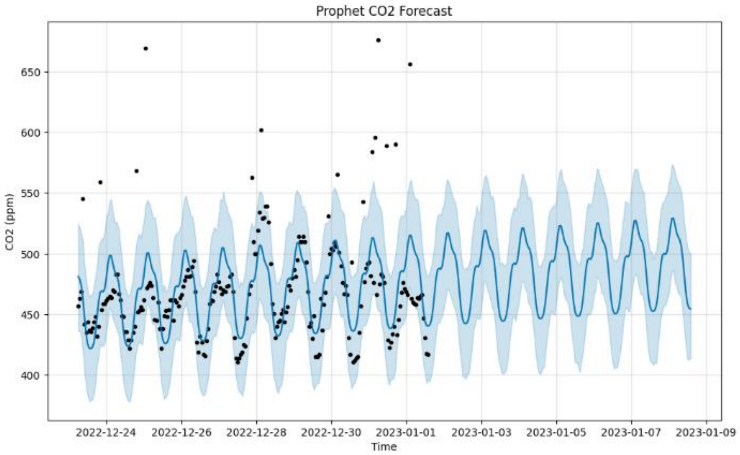


Figure 3.6: Prophet Model for Predicting CO₂ Levels

Based on the MSE values, **Prophet** emerged as the best model for CO2 prediction, followed by **XGBoost**, **Linear Regression**, and **Random Forest** in descending order of accuracy.

PM2.5 Prediction Models

For predicting PM2.5 levels, two models were utilized: Random Forest and Support Vector Regressor (SVR).

The **Random Forest** model, as previously mentioned, is a robust ensemble method known for its ability to handle large datasets with high dimensionality. In this context, Random Forest produced an MSE of 123.26, indicating a reasonably good performance in predicting PM2.5 levels.

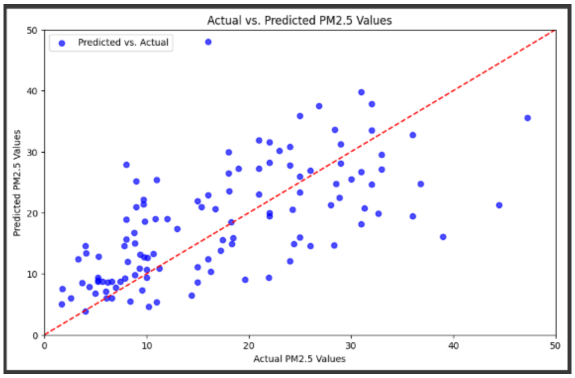


Figure 3.7: Random Forest Model for Predicting PM2.5 Levels

The **Support Vector Regressor (SVR)**, a regression model that leverages the principles of Support Vector Machines, was also tested. The SVR model achieved an MSE of 132.065, slightly higher than that of Random Forest, suggesting that its predictions were marginally less accurate for PM2.5 levels.

Between the two models, **Random Forest** was found to be the superior model for predicting PM2.5 levels, outperforming the Support Vector Regressor.

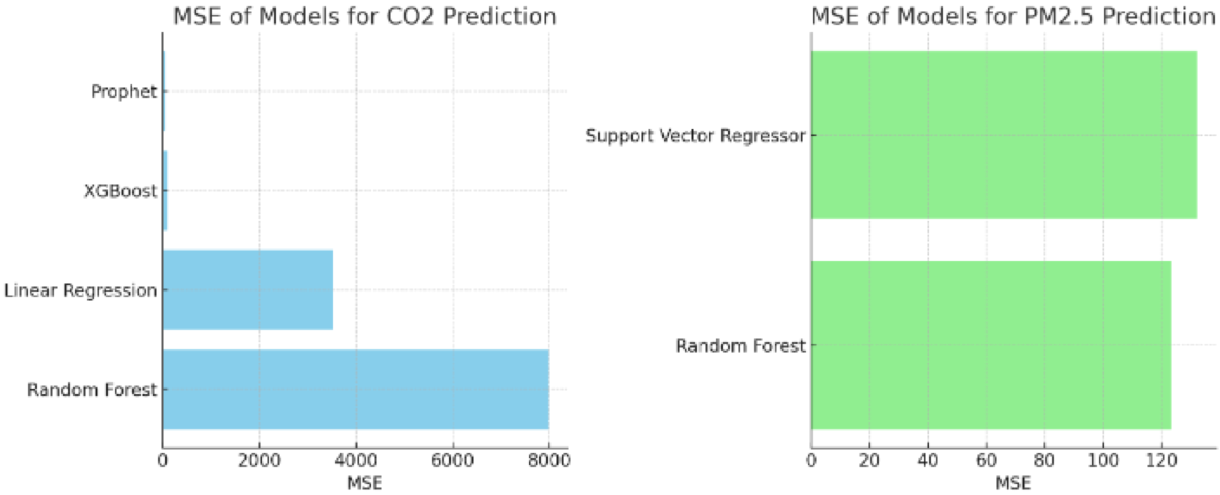


Figure 3.8: Model Comparison for Predicting CO₂ and PM2.5 Levels

The bar charts presented compare the performance of various models used for predicting CO2 and PM2.5 levels based on their Mean Squared Error (MSE).

- **CO2 Prediction**: Prophet is the best-performing model, followed by XGBoost, Linear Regression, and Random Forest.
- **PM2.5 Prediction**: Random Forest outperforms Support Vector Regressor, making it the more accurate model for PM2.5 prediction.

3.2 Project requirements

3.2.1 Functional Requirements and Non-Functional Requirements

Table 3. 1 Functional and Non-Functional Requirements

Functional Requirements	**Non-Functional Requirements**
1. Identification of dataset, cleaning, and preprocessing o Develop a module to gather and verify relevant datasets containing air pollutant records for the Colombo meteorological sector. o Implement a data cleaning process that addresses missing values, outliers, and inconsistencies to ensure data integrity for subsequent analysis.	System should respond real-time
2. Training dataset using different models o Integrate a module to employ various machine learning algorithms, such as linear regression, decision trees, and neural networks, for training on the preprocessed dataset. o Implement cross-validation techniques within the system to assess model performance and generalization ability, ensuring robustness in predicting air pollution levels.	The application should be reliable.
3. Finding the best model which gives high-level **accuracy** o Develop a module to evaluate performance metrics of trained models, including accuracy, precision, recall, and F1-score, to identify the model with the highest predictive accuracy. o Integrate functionality to fine-tune hyperparameters and optimize model architectures, enhancing performance to achieve the desired level of accuracy in predicting air quality levels.	Higher accuracy of results and more efficient results

4. Data visualization and comparison of the degree of air pollution in Colombo areas o Design and implement interactive visualization tools that display air pollution levels across different regions of Colombo. o Develop features allowing authorized users to access historical, current, and predicted air quality data, facilitating informed decision-making and resource allocation to mitigate pollution risks.	Interfaces should be user-friendly.
5. Alerting the affected users who have registered once a high pollution level is discovered. o Internal Server will be implemented to capture the event whenever an air quality index found high using the application.	Should properly work for android and IOS devices.

3.2.2 Software Requirements

Software requirements are:

i. Python
ii. Anaconda with Jupyter Notebook
iii. Google Collab
iv. TensorFlow
v. React Native
vi. Flutter
vii. Node JS
viii. Firebase
ix. MongoDB
x. Fast API

3.2.3 Personnel Requirements

The following people are needed in order to improve the research's integrity, quality, and knowledge.

Resources and Dataset Air pollution level in Colombo

- Central Environment Authority (CEA)
- The National Building Research Organization (NBRO).

3.3 Software Solution

The Anaconda Data Science Framework is a robust software solution that harnesses the power of the Anaconda distribution, Python, and R programming languages. It is designed for scientific computing, data science, machine learning applications, and large-scale data processing. The framework simplifies package management and deployment, making it accessible for over 12 million users on Windows, Linux, and MacOS. With more than 1,400 popular data science packages, the Anaconda Data Science Framework includes the Conda package and the Anaconda Navigator, eliminating the need for independent library installations.

Components:

(i) Jupyter Notebook:

An open-source web application enabling the creation and sharing of documents with live code, equations, visualizations, and narrative text. Used for diverse applications like data cleaning, statistical modeling, and machine learning.

(ii) Notebook Document:

Produced by the Jupyter Notebook App, these documents contain both computer code and rich text elements. They serve as both human-readable descriptions and executable documents for data analysis.

(iii) Jupyter Notebook App:

A server-client application facilitating the editing and running of notebook documents through a web browser. It operates locally or on a remote server, providing a dashboard for file management and kernel control.

(iv) Google Colab: A cloud-based Jupyter Notebook environment that allows you to write and execute Python code directly in your browser. Google Colab provides free access to computing resources, including GPUs and TPUs, making it an excellent platform for tasks such as machine learning, data analysis, and deep learning. It supports the collaboration and sharing of notebooks with others and integrates seamlessly with Google Drive for easy storage and access to your projects.

A computational engine executing code in Notebook documents. The ipython kernel, as referenced, executes Python code. Kernels for various languages exist, automatically launching when a Notebook document is opened.

Phases of the Proposed Approach:

A. Data Collection & Pre-processing:
Historical datasets on air pollution factors in Colombo, obtained from CEA and NBRO, are pre-processed to enhance accuracy and reliability.

B. Data Analysis:
Utilizing RStudio, correlation matrix, and distribution charts to identify correlations and factors affected by PM2.5.

C. Evaluation:
Applying Train-Test-Split for cross-validation, dividing the pre-processed dataset into training and evaluation sets.

D. Training the Model:
Implementing machine learning algorithms - Random Forest, Multiple Linear Regression, Support Vector Machine, and K Nearest Neighbors using Python libraries like pandas, scikit-learn, and PyCharm IDE.

E. Model Evaluation:
Predicting the Air Quality Index (AQI) based on pre-processed data and identifying the most suitable algorithm by calculating accuracy. AQI categories include Good, Satisfactory, Moderate, Poor, Very Poor, and Severe.

The Anaconda Data Science Framework offers an integrated solution for air quality prediction, encompassing data collection, pre-processing, analysis, and model training and evaluation. The software enhances accessibility and efficiency in environmental data science applications.

User Interface (UI) Development: The UI will be created to collect essential air pollution related parameters, starting with the React-based frontend. Input of demographic data will be handled using dynamic interfaces that are designed to be intuitive and easy to use.

Backend Development: Flask or Django will work with Python to power the backend. APIs for data processing, machine learning, and creating customized exercise regimens will be housed in this component. There will be a strong communication link created between the frontend and backend.

Development of the Machine Learning Model: The machine learning model will be built using the robust libraries provided by Python. Several algorithms will be examined and put into practice in order to precisely forecast health percentages. Hyperparameters will be adjusted during an optimization phase to achieve peak performance.

Database Management: MongoDB will be employed for efficient data storage and retrieval. The database will manage preprocessed data, model parameters, and user information, ensuring seamless access for the entire system.

Agile Scrum approach is the software development life cycle that is being suggested. Based on the agile principles of the Agile Manifesto, Scrum is an iterative methodology for software development [8]. Another definition of scrum is a lightweight development approach that offers complete transparency and quick adaptation [8]. Because of the aforementioned tendencies, agile allows for rapid component modifications during the implementation phase in response to changing needs. Agile scrum technique will be used to facilitate the changes in an effective manner. As a result, the suggested solution will be put into practice in accordance with the framework by facilitating quick adaptation and ongoing adjustments.

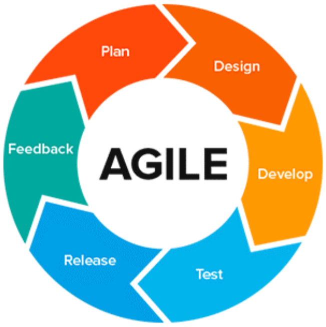


Figure 3. 9 Agile Methodology [9]

3.3.1 Requirement Gathering and Analysis

- **Gathering requirements from National Building Research Organization (NBRO) and Central Environmental Authority (CEA).**

  By requesting authority people who worked at above mention workplace collecting dataset of minimum of 5 years data of pollution level in Colombo areas. Discuss the problem in the datasets.

3.3.2 Feasibility Study

Feasibility Analysis is the process of determination of whether or not a project is worth doing.

- **Technical Feasibility**

  Members should have extensive knowledge of node.js, JavaScript, React Native, using external libraries, retrieving data from external APIs, event-driven architectures, and CPU optimization techniques in order to effectively finish the project. Project contributors also need to have specialized knowledge in software architectures and frameworks, mobile app development, and web app development.

- **Economic Feasibility**

  The limitations of the project required the employment of more dependable and reasonably priced resources in order to guarantee cost-effectiveness and dependability. As a result, budgetary constraints affected the choice of materials and essential elements. It is anticipated that the proposed sub-component will function flawlessly, with no problems or faults, stressing dependability, high performance, and affordability. The project strives for low costs related to component requirements and resources.

- **Schedule Feasibility**

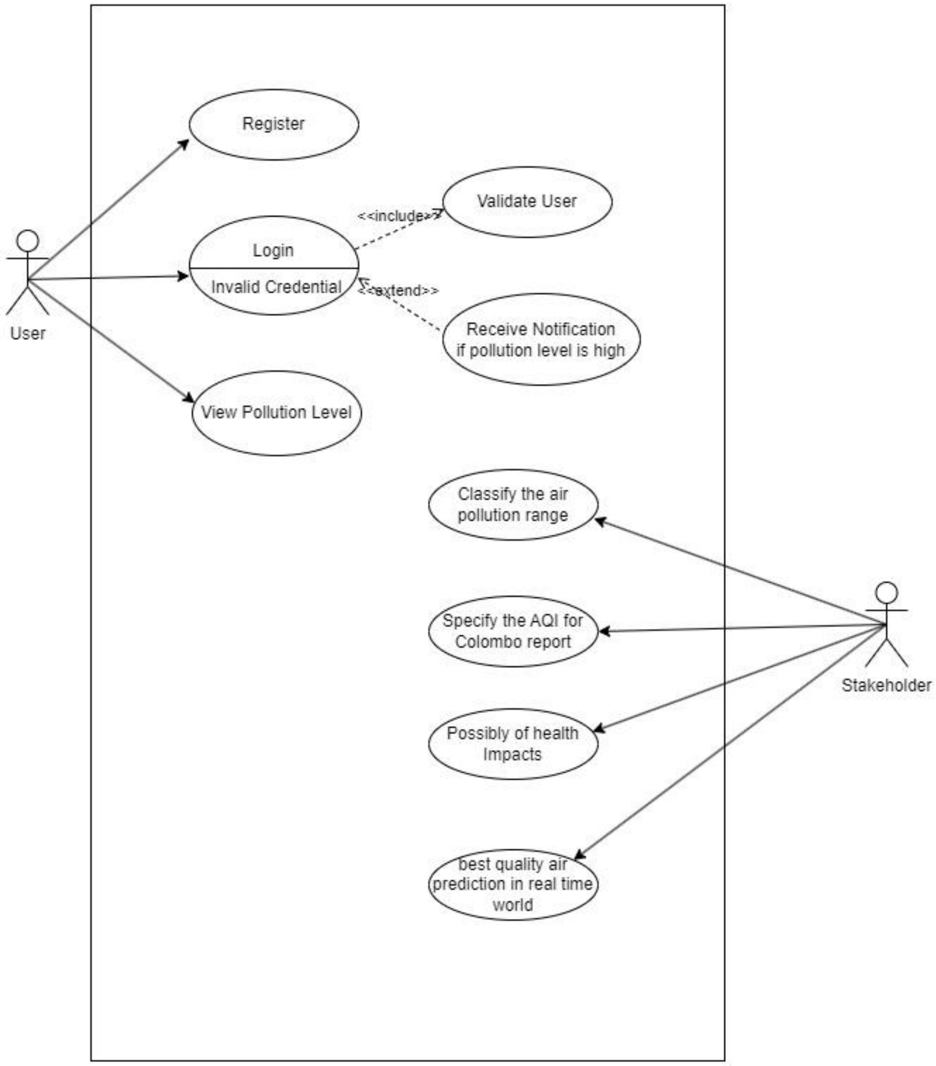
The proposed subcomponent needs to be completed in the allotted time frame, with higher precision in the results, and strictly adhering to the timeline that has been set forth. Last but not least, the final product presentation ought to happen on the scheduled deadline.

- **Operation Feasibility**

A designated member is in charge of each stage of the software life cycle, with an emphasis on the requirement analysis phase in particular. Ensuring that the finished product meets the needs specified by the consumers is the ultimate goal, with a focus on user-centric design.

3.3.3 Design

3.3.3.1 Use Case Diagram



The Figure 3.10 overall use case diagram indicates the external and internal actor functional and technical overview with the node server system.

3.3.3.2 Sequence Diagram

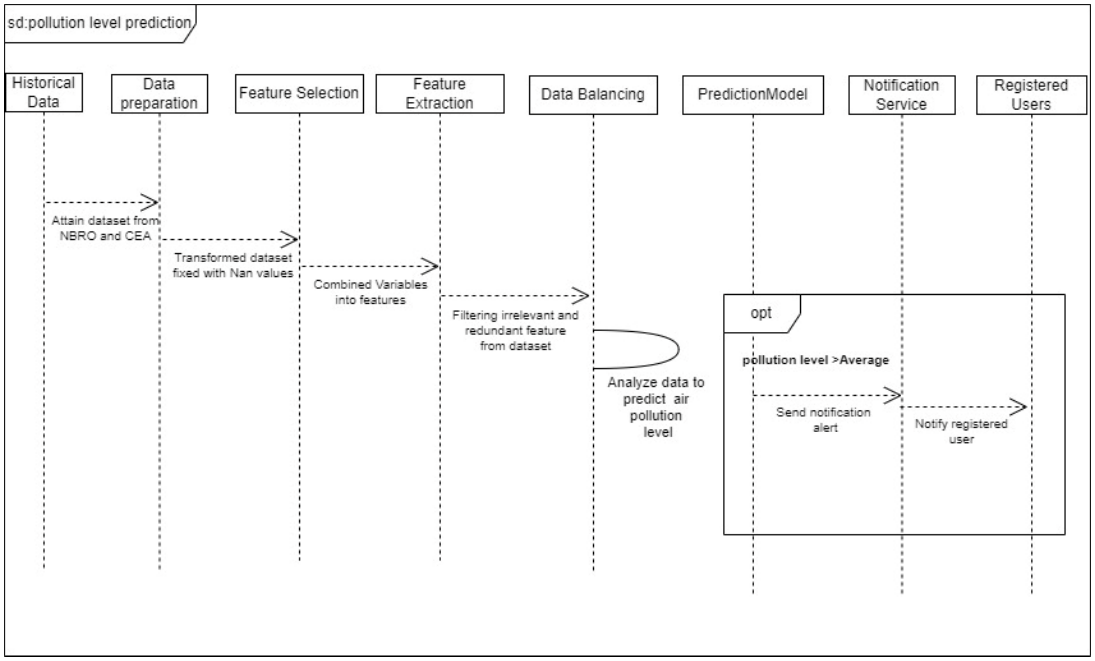

sd:pollution level prediction
Historical Data
Data preparation
Feature Selection
Feature Extraction
Data Balancing
PredictionModel
Notification Service
Registered Users
Attain dataset from NBRO and CEA
Transformed dataset fixed with Nan values
Combined Variables into features
Filtering irrelevant and redundant feature from dataset
Analyze data to predict air pollution level
opt
pollution level >Average
Send notification alert
Notify registered user

Figure 3. 11 Sequence Diagram

The figure 3.4 sequence diagram depicts the technical interactions with the system.

3.3.3.4 Dataset

The Central Environment Authority (CEA) and the National Building Research Organization (NBRO) have historical datasets with data on the hourly concentration levels of air pollution components in Colombo. The dataset includes average concentrations of air and meteorological components, including humidity, CO, SO2, NO2, PM10, and PM2.5, for the period of January 2020 to December 2023. Preprocessing techniques are being applied to the generated dataset in order to increase prediction accuracy and ensure the validity of the results. The distribution and type of the dataset, as well as the relationships between the variables related to air pollution, are ascertained using distribution charts and correlation matrices.

3.3.3.5 Work Breakdown Structure

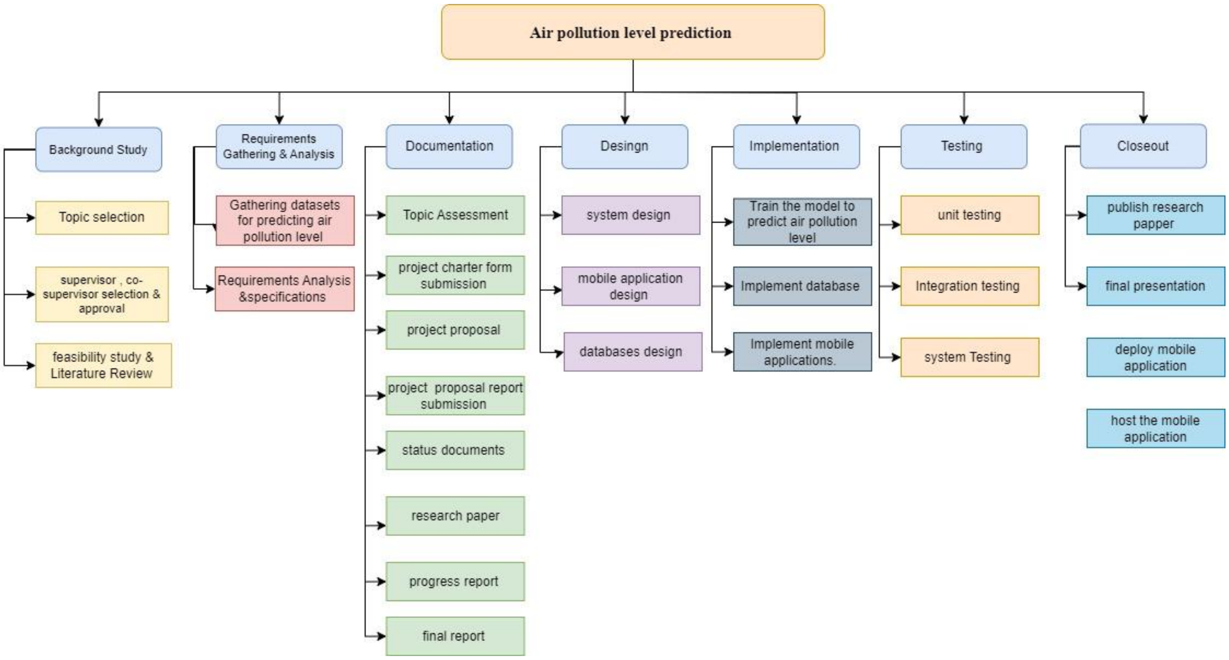

Air pollution level prediction
Background Study
Topic selection
supervisor , co-supervisor selection & approval
feasibility study & Literature Review
Requirements Gathering & Analysis
Gathering datasets for predicting air pollution level
Requirements Analysis &specifications
Documentation
Topic Assessment
project charter form submission
project proposal
project proposal report submission
status documents
research paper
progress report
final report
Desingn
system design
mobile application design
databases design
Implementation
Train the model to predict air pollution level
Implement database
Implement mobile applications.
Testing
unit testing
Integration testing
system Testing
Closeout
publish research papper
final presentation
deploy mobile application
host the mobile application

Figure 3. 12 Work Breakdown Structure

3.3.3.6 High Fidelity Diagram

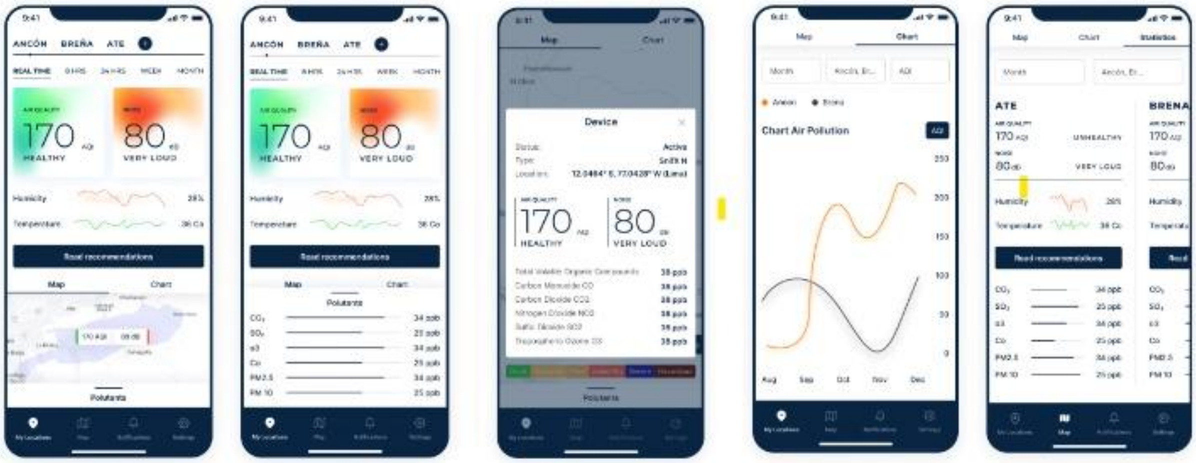


Figure 3. 13 High fidelity diagrams for information sharing to crowd.

3.3.4 Implementation

The proposed approach consists of a sequence of phases for predicting the AQI. The sequence of phases includes collecting the dataset from Central Environmental Authority, pre-processing the collected dataset, analyzing the collected dataset to identifying the correlations among air pollution factors, applying appropriate ml algorithms, & ultimately selecting the most suitable machine learning approach
& analyzing the prediction results. as given below,

- Collecting polluted air data from the emission standard publication websites and Preprocessing, Cleaning, Transforming.
- Feature Selection and Analysis.
- Model Development Implement machine learning algorithms for pollution level prediction. Explore and select the most suitable predictive model based on performance metrics.
- Dynamic Predictions for Real-time Adjustments.
- Real-time Pollution Data and Comparison and Visualization.

We're bringing all these features together so that everyone involved can easily see and use them through a mobile app.

- **Mobile Application**

The envisioned mobile application for predicting air pollution levels offers support in the English language. The application will be available in two versions: a commodity version and a premium version. These versions provide users with the flexibility to choose based on their preferences and needs, offering enhanced features in the premium version. To ensure users stay informed, the application is designed to send push notifications, delivering timely updates on air quality even when the app is not actively in use. This feature contributes to a seamless and user-friendly experience, keeping individuals updated on crucial information related to air pollution levels.

- **Sever Back-End**

The server back-end is the most important part of the research, which is responsible for data storing and processing. MongoDB is the database of the system, it is a cross-platform, open source and document-oriented database.

3.3.5 Testing (Track and Monitor)

Every function in the application is tested for errors during this step. We have a plan to test it in different ways, like checking if users like it, making sure it works with other parts, and checking each part on its own. This step is not just about fixing mistakes; it's also about making sure the app is really good and works well.

4. GANTT CHART

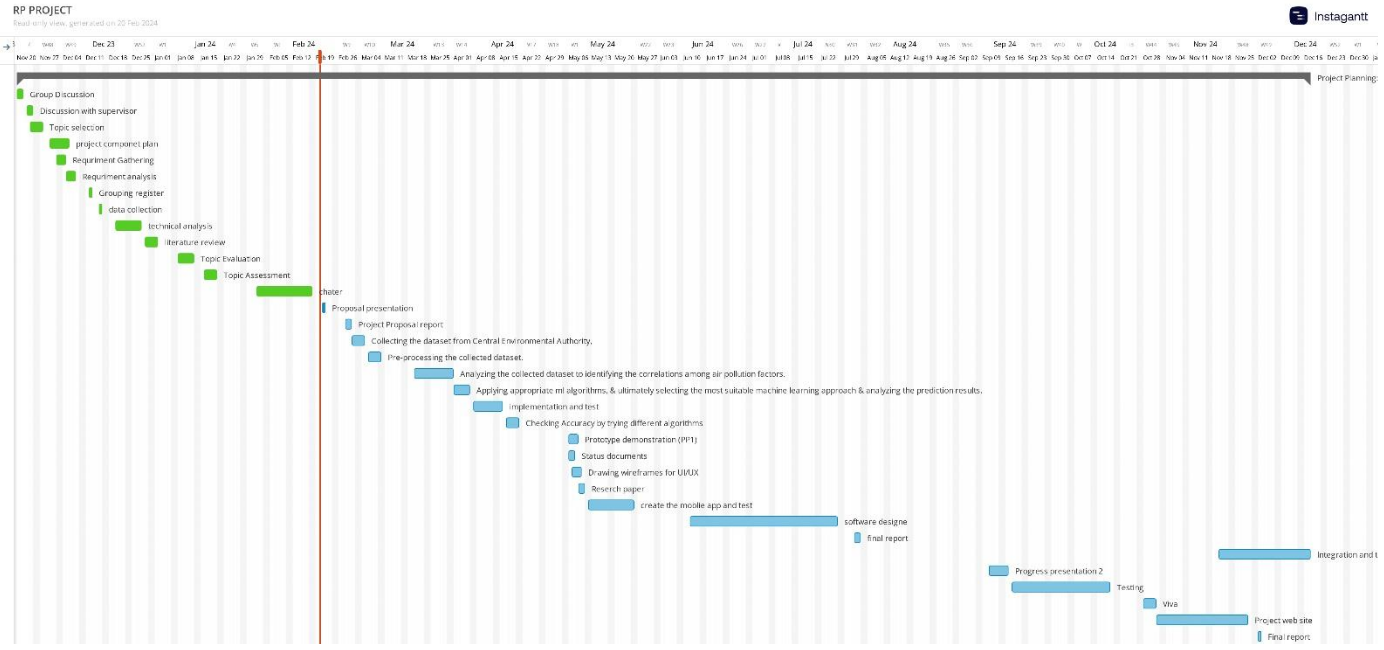


Figure 3. 14 Gantt Chart

5. DESCRIPTION OF PERSONAL AND FACILITIES

Facilitators:

Ms. Chathurangika Kahandawarachi - Sri Lanka Institute of Information Technology (SLIIT)

Ms. Pipuni Wijesera - Sri Lanka Institute of Information Technology (SLIIT)

Facilities:

National Building Resource Management (NBRO)

Central Environment Authority

6. BUDGET AND BUDGET JUSTIFICATION

The estimated budget contains subscription costs, deployment costs, database costs and hosting costs.

Table 5. 1 Budget of the crowdsourcing component

Feature	**Price**
Travelling cost for collecting data	Rs. 4850.00
Deployment Cost	Rs. 5080.00 / month
MongoDB	Free
Mobile App -Hosting on Play Store	Rs. 4996.00
Mobile App -Hosting on App Store	Rs.18000.00 /annual
Total	Rs. 32926.00

6. COMERCIALIZATION

The proposed software solution for air pollution level prediction has significant business potential for market analysis and a well-rounded marketing plan. Here's an assessment based on the provided information:

1. SaaS Model:

The Software as a Service (SaaS) approach emerges as a powerful model for predicting air pollution levels. Its advantages lie in scalability, accessibility, and cost-effectiveness, aligning with the contemporary trend of delivering applications over the internet. Eliminating maintenance costs and hardware expenses, the SaaS model provides a steady revenue stream through one-time costs or monthly subscriptions. The appeal is further heightened by customization and integration features, making it an attractive choice for businesses seeking reliable and scalable air pollution prediction services.

2. Mobile Application with Ad-Based Revenue:

The development of a mobile application with integrated advertisements offers a lucrative avenue, especially considering the real-time nature of air pollution predictions. This model benefits advertisers by engaging users actively day and night. Ad-Based revenue becomes a significant income source, particularly with a large user base. Introducing a subscription model to disable advertisements enriches the revenue stream, creating a mutually beneficial scenario for both consumers and advertisers.

3. Freemium Model:

The freemium model, proven in various industries, is a strategic approach offering basic services for free while introducing premium features for a fee. Alert notifications, positioned as a compelling premium service, aim to attract users to opt for the paid version. This model's attractiveness lies in building a broad user base with essential free services, creating opportunities to upsell premium features. The freemium model's flexibility allows users to tailor their experience, contributing to a diversified and sustainable revenue stream.

6.1 Target Audience and Market Space

- **Target Audience:**
  - Environmental agencies
  - Government bodies
  - Individual concerned about air quality.

- **Marketing Channels:** Online platforms, social media, environmental forums, and collaborations with government and environmental agencies.

- **Pricing Strategy:** Tailored pricing plans for different user categories (individuals, businesses, government agencies) with clear benefits for each tier.

- **Promotions:** Launch promotions, free trial periods, and partnerships with environmental organizations for wider visibility.

- **Customer Engagement:** Regular updates, educational content on air quality, and responsive customer support.

- **Market Space**
  - No need of prior knowledge in technology.
  - No age limitation for users.
  - No need of prior knowledge regarding Air Pollution.

REFERENCES

[1] Mahanta, S., Ramakrishnudu, T., Jha, R. R., & Tailor, N. (2019, October). Urban air quality prediction using regression analysis. In *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)* (pp. 1118-1123). IEEE.

[2] Castelli, M., Clemente, F. M., Popovič, A., Silva, S., & Vanneschi, L. (2020). A machine learning approach to predict air quality in California. *Complexity*, *2020*.

[3] Zhong, S., Yu, Z., & Zhu, W. (2019). Study of the effects of air pollutants on human health based on Baidu indices of disease symptoms and air quality monitoring data in Beijing, China. *International journal of environmental research and public health*, *16*(6), 1014.

[4] Nandasena, Y. L. S., Wickremasinghe, A. R., & Sathiakumar, N. (2010). Air pollution and health in Sri Lanka: a review of epidemiologic studies. *BMC Public Health*, *10*, 1-14.

[5] Xayasouk, T., & Lee, H. (2018). Air pollution prediction system using deep learning. *WIT Trans. Ecol. Environ*, *230*, 71-79.

[6] Iskandaryan, D., Ramos, F., & Trilles, S. (2020). Air quality prediction in smart cities using machine learning technologies based on sensor data: a review. *Applied Sciences*, *10*(7), 2401.

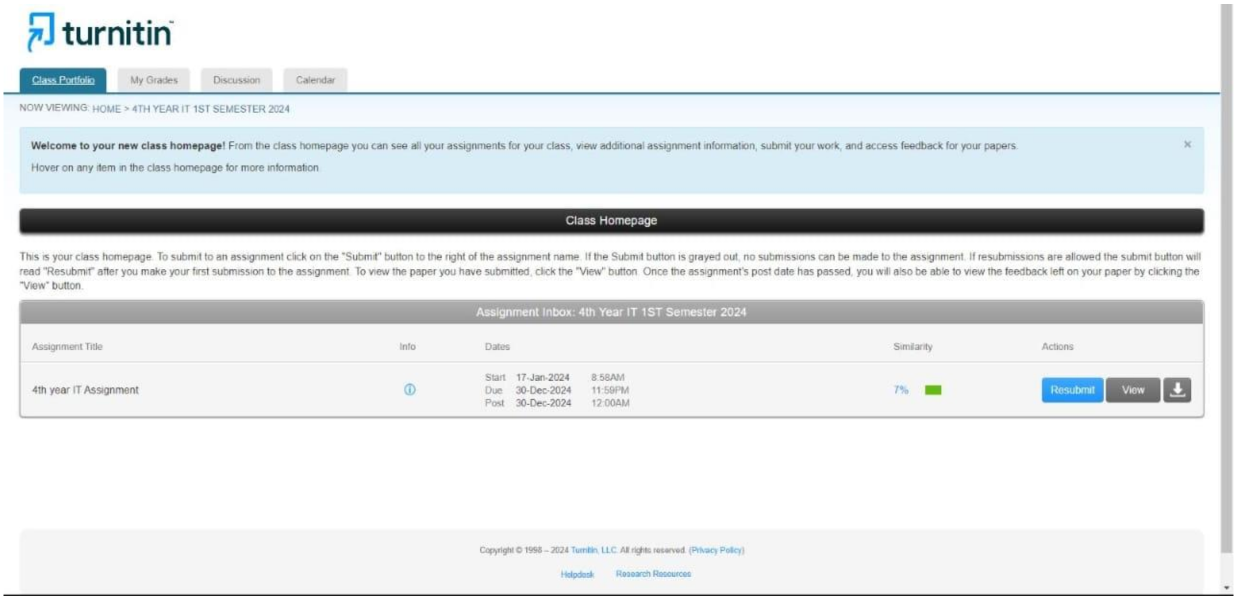
[7] Samad, A., Garuda, S., Vogt, U., & Yang, B. (2023). Air pollution prediction using machine learning techniques–an approach to replace existing monitoring stations with virtual monitoring stations. *Atmospheric Environment*, *310*, 119987.

[8] Anum, A., Rehman, M., & Anjum, M. (2017). Framework For Applicability Of Agile Scrum Methodology: A Perspective Of Software Industry. *International journal of advanced computer science and applications*, *8*(9).

[9] D. Iddo, “Agile Development,” Medium, 04-Oct-2019. [Online]. Available: https://medium.com/moodah-pos/agile-development-95cad3573abf. [Accessed: 22-Mar-2021].

APPENDICES

- Plagiarism Report



NOW VIEWING: HOME > 4TH YEAR IT 1ST SEMESTER 2024

Welcome to your new class homepage! From the class homepage you can see all your assignments for your class, view additional assignment information, submit your work, and access feedback for your papers.

Hover on any item in the class homepage for more information.

Class Homepage

This is your class homepage. To submit to an assignment click on the "Submit" button to the right of the assignment name. If the Submit button is grayed out, no submissions can be made to the assignment. If resubmissions are allowed the submit button will read "Resubmit" after you make your first submission to the assignment. To view the paper you have submitted, click the "View" button. Once the assignment's post date has passed, you will also be able to view the feedback left on your paper by clicking the "View" button.

Assignment Inbox: 4th Year IT 1ST Semester 2024

Assignment Title	Info	Dates	Similarity	Actions
4th year IT Assignment	ⓘ	Start 17-Jan-2024 8:58AM Due 30-Dec-2024 11:59PM Post 30-Dec-2024 12:00AM	7% 🟩	Resubmit View ⬇

Copyright © 1998 – 2024 Turnitin, LLC. All rights reserved. (Privacy Policy)

Helpdesk Research Resources

Figure 3.15: Random Forest Model for Predicting PM2.5 Levels