BINDURA UNIVERSITY OF SCIENCE EDUCATION

FACULTY OF SCIENCE AND ENGINEERING

COMPUTER SCIENCE DEPARTMENT

APPLICATION OF DATA MINING TECHNIQUES FOR PREDICTING AIR POLLUTION FOR A SUNSHINE CITY (HARARE CASE STUDY)

EDSPARTIA A MUFANDAEDZA

B191605B

SUPERVISOR: MR O. MUZURURA

A RESEARCH PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE BACHELOR OF SCIENCE HONOURS DEGREE IN COMPUTER SCIENCE

APPROVAL FORM

The undersigned certify that they have supervised the student Edspartia A Mufandaedza's dissertation entitled the application of data mining techniques for predicting air pollution for a Sunshine City submitted in Partial fulfillment of the requirements for the Bachelor of Computer Science Honors Degree of Bindura University of Science Education.

…….	……..	……/……/…….
STUDENT	SIGNATURE	DATE
…….	…….....	……./……./…..
SUPERVISOR	SIGNATURE	DATE
…….	…….	……./……/……
CHAIRPERSON DATE	SIGNATURE	DATE
……..	…….	…..../……/…
EXTERNAL EXAMINER	SIGNATURE	DATE

DEDICATION

This paper is dedicated to my family and friends for their unconditional support towards the completion of my project. Through the hardships and pains I endured because of their endless support. I hereby extended my dedication and special thanks to my Supervisor Mr. O Muzurura for his support. Patience, endurance and passion are fundamentals they gave which have been taking me to different levels in life.

ACKNOWLEDGEMENTS

First and foremost, I would like to acknowledge the God of heaven for His continuous guidance and wisdom throughout the duration of this project. My second acknowledgement goes to my supervisor Mr. O Muzurura for his support in the research project and upon development of the system and its documentation. His contributions were valuable to this research project. Lastly I would like to acknowledge all the lecturers within the Computer Science department whom I consulted along the way.

ABSTRACT

Air pollution is a significant environmental hazard that poses a danger to all living organisms, as fresh and good quality air is vital for survival. Human activities such as automotive transportation, agricultural practices, industrialization, mining, and fossil fuel combustion contribute to air pollution by releasing harmful pollutants like sulfur dioxide, nitrogen dioxide, carbon monoxide, and particulate matter into the air. The polluted air we breathe can cause various health issues. Therefore, there is a need for an effective system to predict air pollution and improve environmental conditions. To address this issue, advanced techniques, such as data mining, can be used to predict air pollution in smart cities. In this study, a multivariate multistep Time Series data mining technique using the random forest algorithm was employed to predict air pollution levels. The model uses past data to make predictions, reducing complexity and improving effectiveness and practicality. This approach can provide more reliable and accurate decisions for environmental protection departments in smart cities. The software was able to forecast whether the weather will be good or bad on a certain day

Table of Contents

List of Figures

List of Tables

List of Screenshots

CHAPTER 1: INTRODUCTION

1.1 Introduction

United Nations heads of state reportedly attended the COOP27 meeting this year in Egypt. The ozone layer is believed to be being degraded, which is causing terrible weather and high temperatures. Since then, several strategies have been put out on how to effectively address the problem of the greenhouse effect and dangerous gases in the atmosphere. The author was inspired to do thorough study on the use of data mining, a branch of artificial intelligence, to forecast air pollution in the Sunshine City (Harare).

1.2 Background of the Study

Ensemble climate change projections were first released in 1992 by the National Centers for Environmental Prediction (NCEP) and the European Center for Medium-Range Weather projections (ECMWF). The other significant weather forecasting facilities from across the world quickly followed them. In order to cover the spectrum of potential future atmospheric conditions, ensemble forecasts aim to present a collection of equally plausible prediction realizations. Richardson (2000) presented the earliest evidence for the economic advantage of probabilistic predictions produced from the ECMWF ensemble compared to a single deterministic forecast. It is widely understood that ensemble predictions are crucial for capturing the probabilistic character of weather forecasting and for overcoming the inherent constraints to the predictability of the atmosphere. (Palmer, 2018).

To determine the status of the atmosphere at the moment and to anticipate how it will change in the future, climate change primarily relies on data assimilation and numerical weather prediction (NWP). The chaotic accumulation of modest beginning flaws and imperfections in our approximation models of the atmosphere usually limits the accuracy of such deterministic weather forecasts to two weeks or less. The atmosphere's interaction with slowly changing ocean-land forcing on considerably longer, multi-month time scales enables accurate seasonal projections of monthly or seasonally averaged conditions. Between these two extremes, it has proven particularly difficult to provide accurate one- or two-week averaged predictions at lead durations of around two weeks to two months (the sub seasonal-to-seasonal or S2S time frame); nonetheless, there are several socioeconomic sectors that can. (Vitart et al., 2017).

1.3 Statement of the Problem

Tropical Cyclone Idai made landfall in Zimbabwe on March 15, 2019, illustrating the necessity for quick and dependable technology intervention to concatenate the need to be alert of such catastrophes to the human race. Unexpected natural disasters occur on a daily basis. As a result of this awareness of human nerves, everyone needs to feel confident about the current and forecasted weather. There may be cyclones, strong winds, heavy rains, cold spells, fog, hot temperatures, or a combination of these that lead to weariness and road accidents. This has been revealed by the researcher to further explain the necessity for study on supervised learning-based weather forecasting.

1.4 Research Aim

To research the application of data mining techniques for predicting air pollution for a Sunshine City. Therefore, the author will use machine learning algorithm model to predict the outcome of the air pollution index.

1.5 Research Objectives

1. To design and develop a model to predict the air pollution index on a city.
2. To evaluate the system performance on predicting air pollution on a city.
3. To assess the accuracy and effectiveness of the model.

1.6 Research Questions

1. How the air pollution prediction application is designed and developed by the author?
2. How the author is going to analyze the application's performances?
3. How the researcher will evaluate the machine learning algorithm used for weather

1.7 Justification or significance of the research

Unexpected natural disaster occurring on our day-to-day basis, there is need for fast and reliable technological intervention to concatenate the need to be aware of such disasters to human race thereby the use of data mining techniques to predict air pollution levels can accurately predict air pollution levels with a high degree of accuracy, timely, and cost-effective solutions for reducing pollution levels whereby authorities can take preventive measures that are less expensive than reactive measures. and protecting public health by providing early warning systems and allowing authorities to take preventive measures to reduce pollution levels

1.8 RESEARCH LIMITATION

The researcher needs more accurate dataset pertaining the city to be used for testing.

1.9 Scope of the study

The focus of the research is to develop a weather forecasting application that can predicts if the weather fine or not for a particular day. This helps in determining further precautions before any danger or harm happened to human beings.

1.10 DEFINITION OF TERMS.

1. **Air pollution** – is the state of the atmosphere at a particular place and time as regards heat, cloudiness, dryness, sunshine, wind, rain, etc.
2. **Prediction**- is a technique that uses historical data as inputs to make informed estimates that are predictive in determining the direction of future trends.
3. **Supervised Learning**- known as supervised machine learning, is a subcategory of machine learning and artificial intelligence. It is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately
4. **Machine Learning** - the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data

1.11 SUMMARY

This script serves the different aspects on the objectives of the study that brought out the main objectives and research questions which also helps to formulate the problem statement. A detailed overview of the study to be carried was outlined in this chapter.

CHAPTER 2 LITERATURE REVIEW

2.1 INTRODUCTION

The researcher addresses the research topics in this chapter and tries to highlight current studies that have been conducted by other authors that are related to the current research endeavour. This will be extremely helpful to the author since it serves as a manual for the procedures and techniques that other writers have used to solve similar difficulties in the past. It serves as a tool that informs the researcher about whether the study endeavour is possible in light of past studies in that particular field.

2.2 MACHINE LEARNING IN AIR POLLUTION PREDICTION

Various authors have presented prediction of air pollution using Convolutional Neural Network (CNN) (Yan et al., 2021). If single image is to be used for air pollution prediction, then CNN is employed and if the air pollution prediction uses series of images the RNN (Recurrent Neural Network) is used. Dataset from Kaggle were used for training and testing the NN (Neural Networks). RF (Random Forest) classifiers was used to estimate pollution data from single image (Kumar, D., 2018). Henceforth other authors presented cooperative learning approach for classification of weather, and estimate whether the weather is sunny or cloudy in nature.

CNN assisted in doing this, and "Sun Dataset," "Labelme Dataset," and "Flickr" were utilized. Others employed "multi-feature texture analysis" to create a typical NNs system. This was subsequently utilized for weather analysis with the use of computer-controlled categorization of satellite cloud images. In a related research, other authors suggested using a weather estimation model based on cluster learning to help predict traffic dangers caused by various meteorological conditions. CCTVs stationed in various places helped to achieve this. In a different study, scientists described a method for super pixel segmentation (SPS)-based automated cloud identification.

2.2 PROBLEMS FACED IN AIR POLLUTION PREDICTION

A significant objective of any research endeavor is to predict correctly. In the hope of identifying some fundamental principles which can be used for predicting the outcomes of certain experiments based on these principles. In addition, most science laws only provide very detailed forecasts of the outcomes of certain types of experiments. However, few physicians face more nuanced or complicated forecast issues than the meteorologist. First and foremost, the meteorological laboratory encompasses the entire world, making it incredibly difficult to

calculate the actual condition of the atmosphere. Moreover, the Earth's surface is a mixture of soil and water that react differently to the source of energy. The atmosphere itself is, then too, a mixture of gasses, liquids and solid components, all of which have an effect on the energy balance of the planet. The circulations in the atmosphere are often incredibly broad, with a life span of just a few seconds, and can last for weeks or months or minute whirls. Apart from the above-mentioned issues we also face different challenges when we want to model weather attributes forecasting. (Simu et al., 2020).

Challenges involve tasks such as:

- Figuring out how various atmospheric characteristics relate to one another. This process gets difficult since it involves connecting non-linear elements for anticipating weather features. Because of the non-linearity in the data needed for meteorological attribute forecasting, many forecasts continue to be wrong today.
- Data obtained for weather attributes forecasting can turn out to be imperfect at sometimes. This may be the due to the data being misread, or data being corrupted, or sensor problem imposing data to be imperfect.
- When image processing related to ground based images and satellite images is involved in the weather attributes forecasting then it involves challenges for pre-processing the images for extracting the features from the image. Here we may face challenges in processing the images. For example, if we want to process the ground based image for cloud classification then the intensity present in the image may pose challenges in processing and classifying the clouds into low level, middle level, and high level clouds. Resolution of the image acquired also plays a vital role in such types of image classification

2.3 TYPES OF AIR POLLUTION ATTRIBUTES FORECASTING

Different methods of weather forecasting have been incorporated from the ancient times which relied in observing the sky to present day systems which are making use of artificial intelligence. Below are few methods used for weather attributes forecasting;

2.3.1 Persistence Air Pollution Forecasting

This is the simplest approach of anticipating weather attribute values. It makes the supposition that the weather will remain unchanged from now to tomorrow. Such a technique is appropriate

for anticipating very short-term weather characteristics predictions. (Russo, A. and Soares, A.O., 2014.).

2.3.2 Trend Air Pollution Forecasting

This approach to weather prediction looks for data on wind speed, wind direction, atmospheric pressure, etc., and predicts the locations where these features will be same in the near future. Such a form of weather prediction is often appropriate for currently casting. This approach of weather forecasting is based on numerical data. (Masood, A and Ahmad, K., 2021).

2.3.3 Analogue Air Pollution Forecasting

The information about the weather today is compared to the historically significant weather occurrence using this approach of anticipating weather qualities. The outcomes of this approach would provide insight into whether the weather will act similarly to how it had in the previous occurrence. (Shaban et al., 2016).

2.3.4 Statistical Air Pollution Forecasting

Utilizing historical data, statistical analysis is used to the data in this approach of statistical weather attribute forecasting in order to anticipate what will happen in the future.

2.3.5 Numerical Air Pollution Prediction (NWP)

Method The NWP technique use models created with the aid of computer languages to forecast weather features. The non-linearity between the weather parameters can be somewhat mitigated with this strategy. Today's Artificial Intelligence (AI) is used in conjunction with the aforementioned techniques to get improved weather attribute predicting outcomes.

The major goals of the AI projections are to employ methods like Multilayer Perceptron (MLP), Machine Learning (ML), Artificial Neural Networks (ANN), etc. This facilitates the use of several ideas to quickly digest complex information. We can get more precision thanks to this procedure. Due to ML's excellent ability to handle non-linear data, many forecasting approaches anticipate using ML principles. (Hong et al., 2022).

2.4 PREVIOUS STUDIES ON AIR POLLUTION PREDICTION

Air pollution prediction systems are computer-based models that use mathematical algorithms to predict future air quality levels based on historical data and current meteorological conditions. These models can be used to forecast the concentration of various pollutants such as particulate matter (PM), nitrogen oxides (NOx), sulfur dioxide (SO2), and ozone (O3). The accuracy of these models depends on the availability and quality of data used for modeling.

The most commonly used air pollution prediction systems are statistical model, dispersion model and hybrid models.

2.4.1 Statistical Models

These models use statistical techniques such as regression analysis, time-series analysis, and artificial neural networks (ANN) to predict future air quality levels based on historical data. Statistical models are widely used for predicting air quality levels due to their simplicity and ease of implementation. Regression analysis is a commonly used statistical technique that involves fitting a linear or nonlinear equation between pollutant concentrations and meteorological variables such as temperature, wind speed, humidity, etc. Time-series analysis involves analyzing historical data to identify patterns in pollutant concentrations over time. Artificial neural networks (ANN) are machine learning algorithms that can learn from historical data to predict future air quality levels. (Lee, J. 2019).

Air pollution prediction systems typically rely on statistical models to predict concentrations of pollutants in the atmosphere. There are many different statistical models that can be used, depending on the specific needs of the application. Here are a few examples:

- **Linear regression**: Linear regression models are often used to predict air pollutant concentrations based on historical data. These models can be used to identify trends and patterns in the data, and to make predictions about future concentrations based on changes in various factors, such as weather patterns, traffic volume, or industrial emissions.
- **Time series analysis**: Time series models are used to analyze data that is collected over time, such as hourly or daily air pollutant measurements. These models can be used to identify patterns and trends in the data, and to make predictions about future concentrations based on historical patterns.
- **Artificial neural networks**: Artificial neural networks are a type of machine learning model that can be used to predict air pollutant concentrations. These models are designed to mimic the way that the human brain works, and can be trained to make predictions based on historical data.
- **Gaussian dispersion models**: Gaussian dispersion models are used to predict how pollutants will disperse in the atmosphere. These models take into account factors such as wind speed, atmospheric stability, and topography to predict the concentration of pollutants at various locations.

2.4.2 Dispersion Models

These models simulate the dispersion of pollutants in the atmosphere using mathematical equations based on meteorological conditions such as wind speed, direction, temperature, and humidity. Dispersion models simulate the dispersion of pollutants in the atmosphere using mathematical equations based on meteorological conditions. These models are more complex than statistical models and require detailed information about the source of pollutants, emission rates, and meteorological conditions. The most commonly used dispersion models are Gaussian plume models and Lagrangian particle dispersion models. (Cimorelli et al., 2005)

Air pollution dispersion models are used to predict how air pollutants disperse in the atmosphere, and how they spread out over time and space. These models take into account a variety of factors, including meteorological conditions, topography, and emissions sources, to estimate the concentration of pollutants at various locations. Here are a few examples of dispersion models used in air pollution prediction:

- **The AERMOD model**: AERMOD (Atmospheric Dispersion Modeling System) is a widely used dispersion model that predicts air pollutant concentrations in complex terrain. The model takes into account factors such as building heights, terrain elevations, and vegetation cover, as well as meteorological conditions, to estimate pollutant concentrations at various locations. AERMOD is frequently used in regulatory applications, such as for permitting and compliance purposes.
- **The CALPUFF model**: CALPUFF (California Puff) is a dispersion model that is commonly used to simulate long-range transport of air pollutants. The model takes into account the effects of terrain, weather, and emissions sources to predict pollutant concentrations over large distances and long time periods.
- **The ISCST3 model**: ISCST3 (Industrial Source Complex Short Term) is a model that is used to predict air pollutant concentrations near industrial sources, such as factories and power plants. The model takes into account a variety of factors, including stack height, wind speed, and atmospheric stability, to estimate pollutant concentrations at various distances from the source.

2.4.3 Hybrid Models

These models combine statistical and dispersion models to improve the accuracy of predictions. Hybrid models combine statistical and dispersion models to improve the accuracy of predictions. These models use statistical techniques to predict pollutant concentrations at a specific location and then use dispersion models to simulate the transport of pollutants from the source to that location. Hybrid models are more accurate than statistical or dispersion models alone but require more data and computational resources. (Liu, Y., & Zhang, Y. 2021).

Hybrid models, which combine two or more different modeling techniques, are increasingly being used in air pollution prediction systems. These models can provide more accurate and reliable predictions by leveraging the strengths of different modeling approaches. Here are a few examples of hybrid models used in air pollution prediction:

- **Machine learning and physical models**: One common approach to hybrid modeling is to combine machine learning techniques, such as artificial neural networks or support vector machines, with physical models, such as dispersion models or atmospheric chemistry models. This allows the model to capture both the complex physical processes that determine pollutant concentrations and the nonlinear relationships between pollutant concentrations and various predictors, such as meteorological conditions or emissions.
- **Data assimilation and modeling**: Another approach to hybrid modeling is to combine data assimilation techniques with air pollution models. Data assimilation is a process by which observational data are incorporated into a model to improve its predictions. By combining data from ground-based sensors, satellite imagery, and other sources with air pollution models, hybrid models can provide more accurate and timely predictions of pollutant concentrations.
- **Ensemble models**: Ensemble models combine multiple different models, often of different types, to provide a more robust prediction. By averaging the results of different models, or selecting the best-performing model for a given situation, ensemble models can reduce the uncertainty and improve the accuracy of air pollution predictions.

Conclusion

Air pollution prediction systems play a crucial role in mitigating the adverse effects of air pollution on human health and the environment. Statistical, dispersion, and hybrid models are

commonly used for predicting air quality levels. The accuracy of these models depends on the availability and quality of data used for modeling. Hybrid models are more accurate than statistical or dispersion models alone but require more data and computational resources. Further research is needed to improve the accuracy of air pollution prediction systems by incorporating new technologies such as remote sensing, machine learning, and artificial intelligence.

2.5 MACHINE LEARNING ALGORITHMS

There are a few different ways to categorize the many kinds of machine learning algorithms, but the four primary categories are reinforcement learning, unsupervised learning, semi-supervised learning, and supervised learning.

2.5.1 Supervised Learning

In supervised learning, which is a form of function approximation, we simply train an algorithm and then select the function that best captures the input data—that is, the function that, given a given X, provides the best guess of y (X -> y)—as the final product. The algorithm relies on an assumption made by humans about how the computer should learn, and these assumptions introduce bias. I'll explain bias in another post. Most of the time, we are unable to identify the true function that always makes the correct predictions. Here, human specialists serve as the computer's teacher by providing training data that include input/predictors and correct responses (output). As a result of this data, the computer should be able to recognize patterns.

In order to forecast the values of the outputs for new data based on the associations that the supervised learning algorithms have learnt from the prior data sets, they attempt to model the relationships and dependencies between the target prediction output and the input characteristics. (Somvanshi. M. et al., 2016).

Draft

- Predictive Model.
- We have labeled data.
- The main types of supervised learning problems include regression and classification problems

List of Common Algorithms

- Nearest Neighbor
- Naive Bayes
- Decision Trees
- Linear Regression
- Support Vector Machines (SVM)
- Neural Networks

2.5.2 Unsupervised Learning

Data without labels is used to train the computer. In this situation, there isn't even a teacher; instead, the computer could be able to educate you when it discovers patterns in the data. These algorithms are especially helpful when a human expert is unsure of what to search for in the data. Are the family of machine learning algorithms that are mostly employed in descriptive modeling and pattern detection? The algorithm cannot attempt to model relationships because there are no output labels or categories present.

These algorithms make an effort to employ techniques on the input data to mine for rules, find patterns, summarize, and aggregate the data points, which enable users better understand the data and derive insightful conclusions. (Ghahramani. 2003)

Descriptive Model

Clustering algorithms and association rule learning algorithms are the two primary categories of unsupervised learning algorithms.

List of Common Algorithms

K-means clustering, Association Rules

Semi-supervised Learning

In the first two kinds, either all of the observations in the dataset have labels or all of the observations have no labels. This is where semi-supervised learning comes in. Given that labeling often involves hiring experienced human specialists, the expense to do so is relatively significant. Semi-supervised methods are therefore the best options for model development when labels are absent from the bulk of observations but present in a small number of them.

These techniques take use of the notion that, despite the fact that the group memberships of the unlabeled data are unknown, this data yet contains vital information on the group parameters. (Zhu et al., 2009).

Reinforcement Learning

This approach tries to perform behaviors that would maximize reward or reduce risk utilizing observations acquired through interactions with the environment. The agent, a reinforcement learning algorithm, iteratively continually learns from its surroundings. The agent gradually gains knowledge from its interactions with the environment until it has investigated every state that is conceivable. Machine learning, which includes reinforcement learning, is a subset of artificial intelligence. It enables software agents and machines to autonomously decide the best course of action within a certain situation in order to maximize performance. The reinforcement signal, or simple reward feedback, is necessary for the agent to learn its behavior. (Chang et al., 2019).

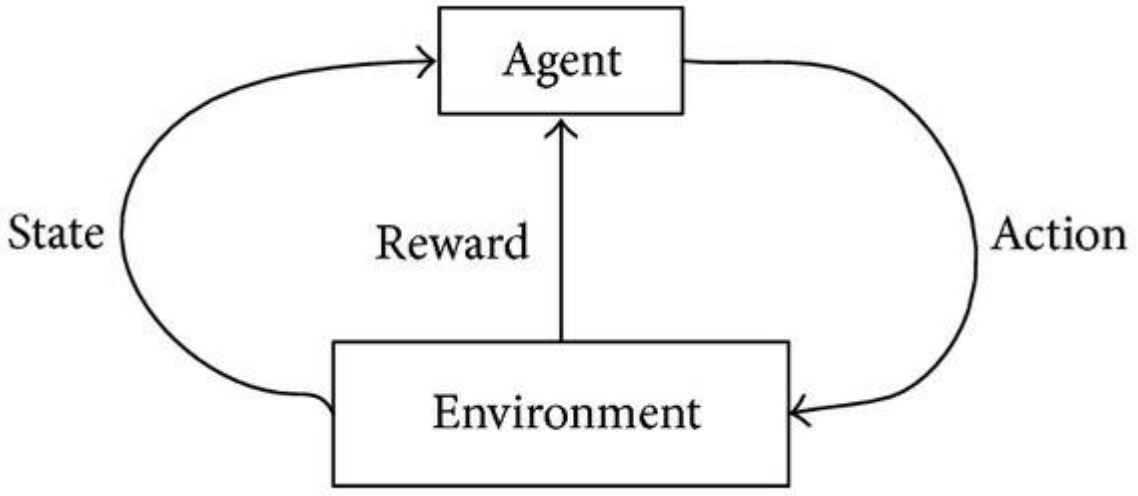


Figure 2.5: Reinforcement Algorithm

Algorithms of various varieties can be used to solve this problem. In actuality, a certain issue type defines reinforcement learning, and all of its solutions are categorized as reinforcement learning algorithms. An agent is required to choose the optimal course of action in the problem

based on his existing situation. The issue is referred to as a Markov Decision Process when this phase is repeated. Reinforcement learning follows these phases in order to create intelligent programs (also known as agents): The agent observes the input state. The agent is forced to take action using the decision-making function. The environment rewards or reinforces the agent when the activity is completed. Information about the reward's state-action pair is saved.

List of Common Algorithms

- Q-Learning
- Temporal Difference (TD)
- Deep Adversarial Networks

Use cases:

Some applications of the reinforcement learning algorithms are computer played board games (Chess, Go), robotic hands, and self-driving cars.

2.6 BENEFITS OF PROPOSED SYSTEM

The application will be a great use for the human beings and countries at large. This will help in forecast weather and predicts if the air pollution index is high or low.

2.7 THE PROPOSED SYSTEM

The system proposed used supervised machine learning technology. It is required to:

- Understand a certain air pollution pattern
- The application will use the air pollution knowledge it has been trained with to determine if the place is habitable or not
- The application will use Bayesian networks to make a best decision for the air pollution index condition.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 INTRODUCTION

The chapter seeks to describe the methods and equipment employed to carry out the suggested goals of the study and system. With the knowledge gained in the preceding chapter, the author will be able to choose among competing strategies to produce the desired outcomes for the study and construct the essential procedures to build a solution.

Many academics are delving into the fascinating field of machine learning in an effort to make life easier for people and safeguard them from dangerous weather patterns, among other things. The EMA department and the community will benefit greatly from the author's machine learning application. To assess if the air pollution index is high or low in a specific city, the application will be utilized. This aids in shielding the populace from harmful and hazardous weather patterns that might expose them to infections, floods, and other hazards.

3.2 RESEARCH DESIGN

Every stage of a project should involve a reflective approach for research design. The various system modules and their intended functions are developed during the design phase. The main goal of this phase is to make sure that a functional, competent, long-lasting, and dependable system is created. As a result, the researcher should create an application that will be utilized to illustrate the study topic of the author.

3.3 REQUIREMENTS ANALYSIS

The outcome of a requirements analysis will determine whether a project is successful or unsuccessful, and the requirements that are created must be relevant to stated business needs, practical, documented, tested, executable, traceable, and quantifiable. (Grady, J.O., 2010).

It is crucial to note all of the functional and non-functional specifications of the desired system at this stage. It is advisable to organize all incoming data, evaluate it while taking into account any potential customer limits, and then create a ready-to-use specification based on the demands of the consumer. The researcher also considered all potential obstacles, such as financial constraints that can obstruct the design technique.

3.4 FUNCTIONAL REQUIREMENTS

Specifies the interaction between inputs and outputs, with a function often being a definition of the function of a system or its modules (Fulton & Vandermolen, 2017). Functional requirements, then, describe the system service that is delivered upon completion of the tasks at hand by describing how the system reacts to a set of inputs, behavior, and output.

The proposed system must be able to meet the following requirements.

- User should select the day in which he/she want to predict the air pollution index
- The application should be able to determine if the provided data gives a high or low air pollution index prediction.

3.5 NON-FUNCTIONAL REQUIREMENTS

They are sometimes referred to as "quality requirements," and they are used to assess a system's performance as opposed to its intended behavior. The following requirements must be met by the suggested system:

- System should have very relatively small response time and decision time
- The system should be easy to assemble

3.6 TOOLS USED (Hardware and Software)

- Streamlit
- Vs Code
- Python 3.9
- Pandas and scikit-learn library

3.7 SYSTEM DEVELOPMENT

This is a summary of the system and how it was created to generate the findings. As a result, it lists every software tool and model that was employed during the system's development.

3.8 SYSTEM DEVELOPMENT TOOLS

Python is being utilized by the researcher as a programming language to create a testing application. This program acts as a tool for testing outcomes. The researcher utilized the Pandas and scikit-learn libraries to evaluate if the air pollution index was high or low. The user will

now enter the weather information for the research. As a result, the Pandas and scikit-learn library will improve their capacity for making decisions about whether the weather is favorable or not.

3.9 BUILD METHODOLOGY

- Prototyping Development- Evolutionary prototyping

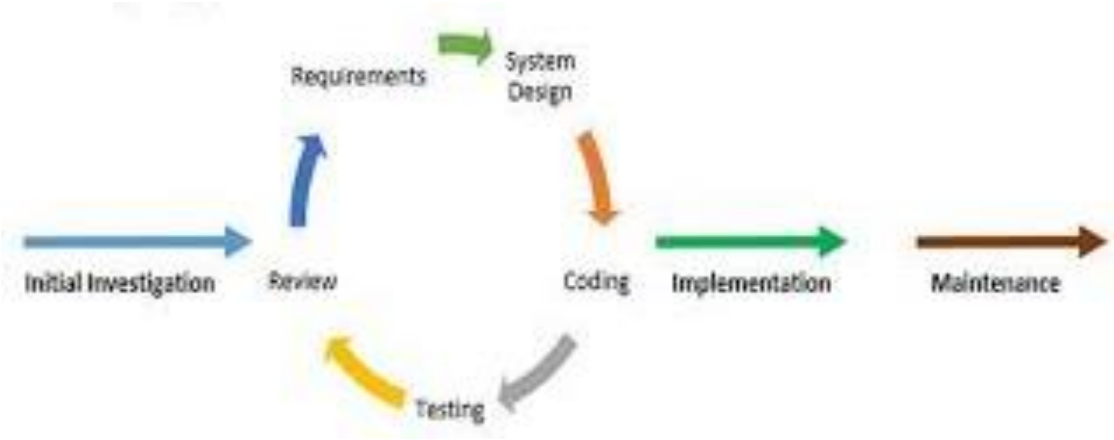


Figure 3.9 Prototype development

3.10 PROTOTYPE

The prototypes often lack many of the intricacies and are not fully functional systems. The objective is to give a system general functionality.

3.10.1 ADVANTAGES OF PROTOTYPE

- Users are actively involved in the development
- Since in this methodology a working model of the system is provided, the users get a better understanding of the system being developed.
- Errors can be detected much earlier.
- Quicker user feedback is available leading to better solutions.
- Missing functionality can be identified easily
- Confusing or difficult functions can be identified

3.10.2 DISADVANTAGES OF PROTOTYPE

- Leads to implementing and then repairing way of building systems.

- Practically, this methodology may increase the complexity of the system as scope of the system may expand beyond original plans.
- Incomplete application may cause application not to be used as the full system was designed
- Incomplete or inadequate problem analysis

3.11 TECHNOLOGY USED

- Python
- Python 3.9
- Pandas and sklearn library

3.12 ALGORITHMS USED

- Supervised Machine Learning
- Bayesian Networks

3.13 PANDAS AND SKLEARN LIBRARY

Pandas is an open-source toolkit for the Python programming language that makes it simple to utilize data structures and data analysis tools. Data Frame columns may be mapped to transformations using Sklearn, which are then concatenated to create features.

3.14 GENERAL OVERVIEW OF AIR POLLUTION INDEX PREDICTION APPLICATION USING SUPERVISED MACHINE LEARNING

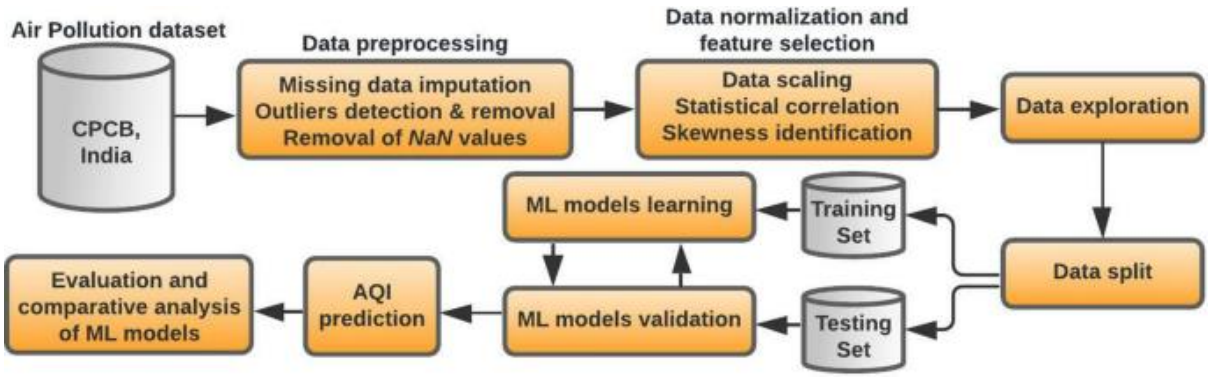


Figure 1.14: Overview of air pollution index application

3.15 PROPOSED SYSTEM FLOWCHART

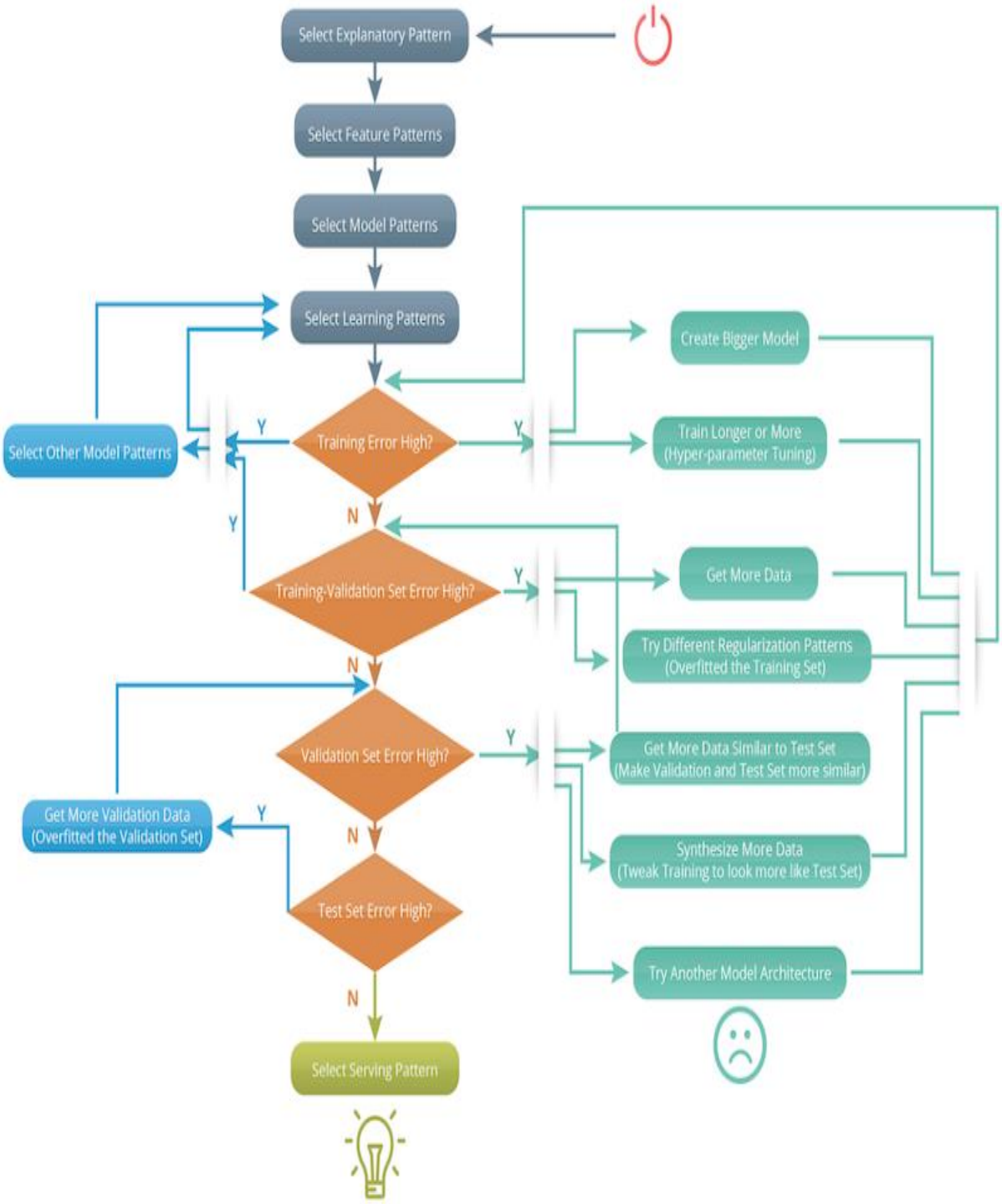


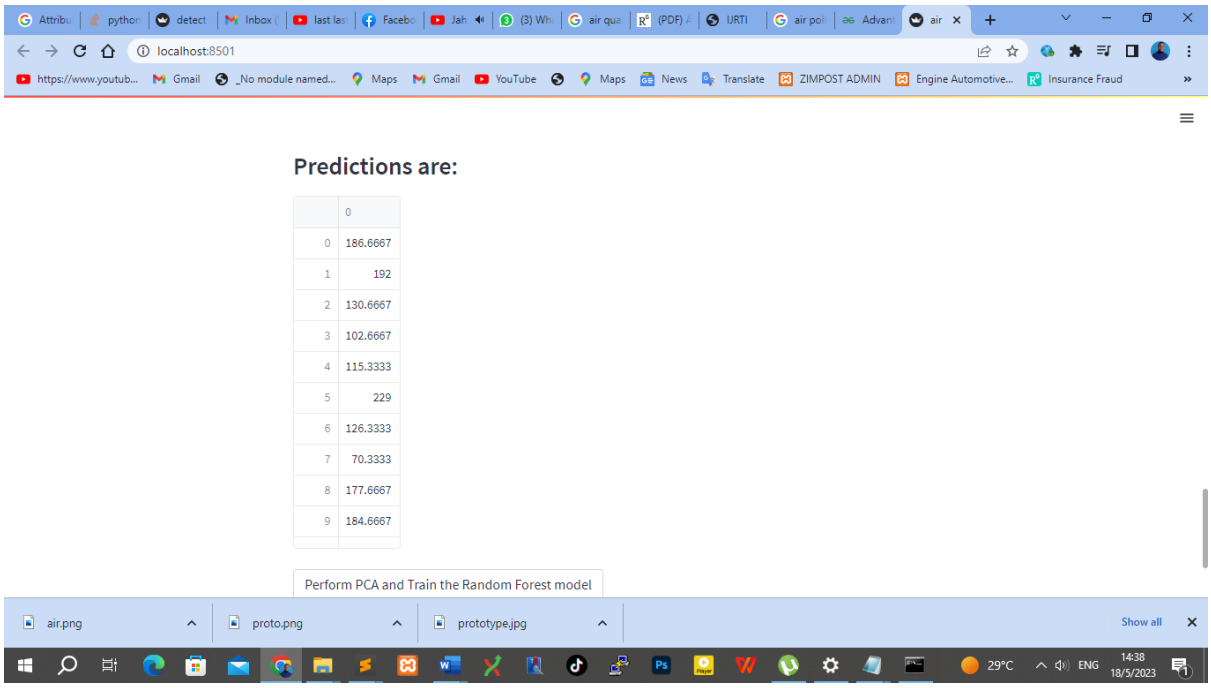
Figure 3.15: System flow chart

3.16 IMPLEMENTATION

The system must be put into operation in this part in order to coordinate and guide the resources developed in the previous chapter to achieve the goals of the research plan. As a result, the prior chapters' documentation is being finished in order to deploy the system.

A specific data set of historical weather data must be available for the application to be evaluated. In order to train the dataset, the system must use the Python libraries pandas and sklearn. Using the provided dataset, which contains information on temperature, humidity, and other variables, this library will assist the system in delivering accurate results. The user is permitted to submit accurate data, and the system will output a result that will indicate whether the provided data indicates a high or low air pollution index.

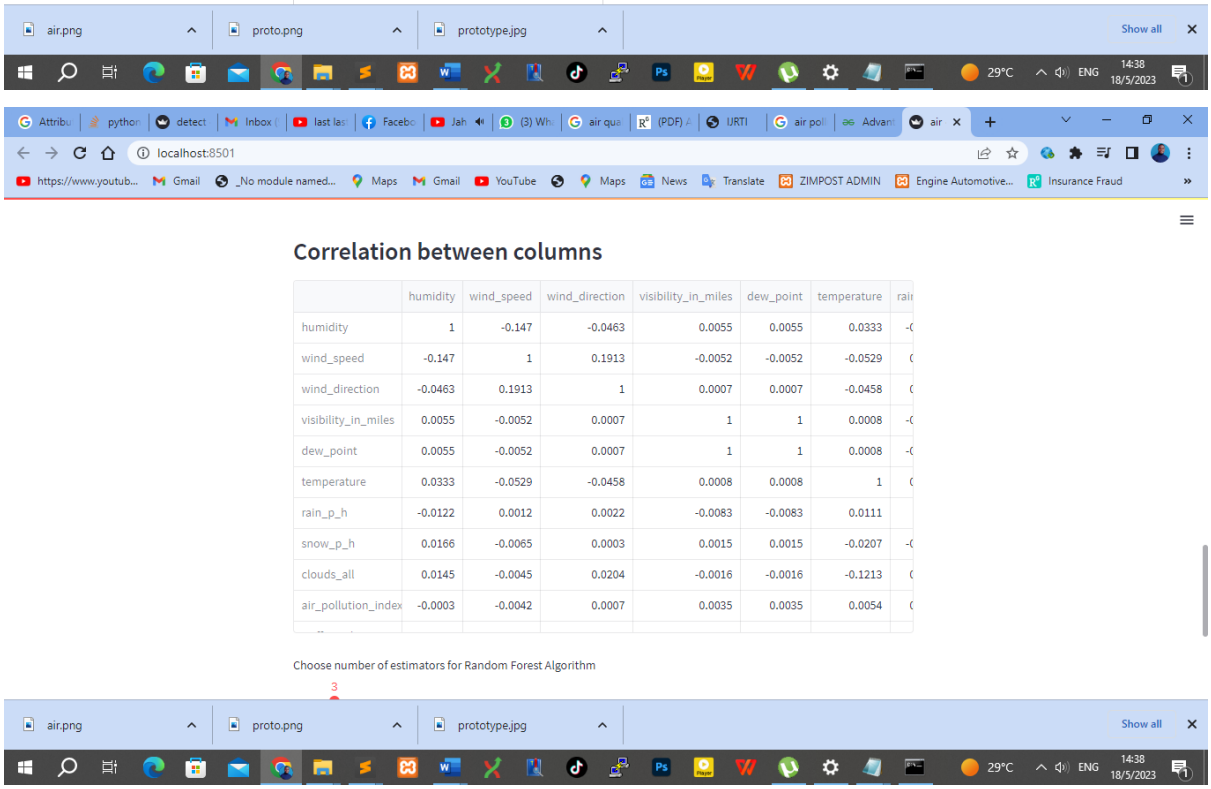
3.17 AIR POLLUTION INDEX PREDICTION APPLICATION



Predictions are:

	0
0	186.6667
1	192
2	130.6667
3	102.6667
4	115.3333
5	229
6	126.3333
7	70.3333
8	177.6667
9	184.6667

Perform PCA and Train the Random Forest model

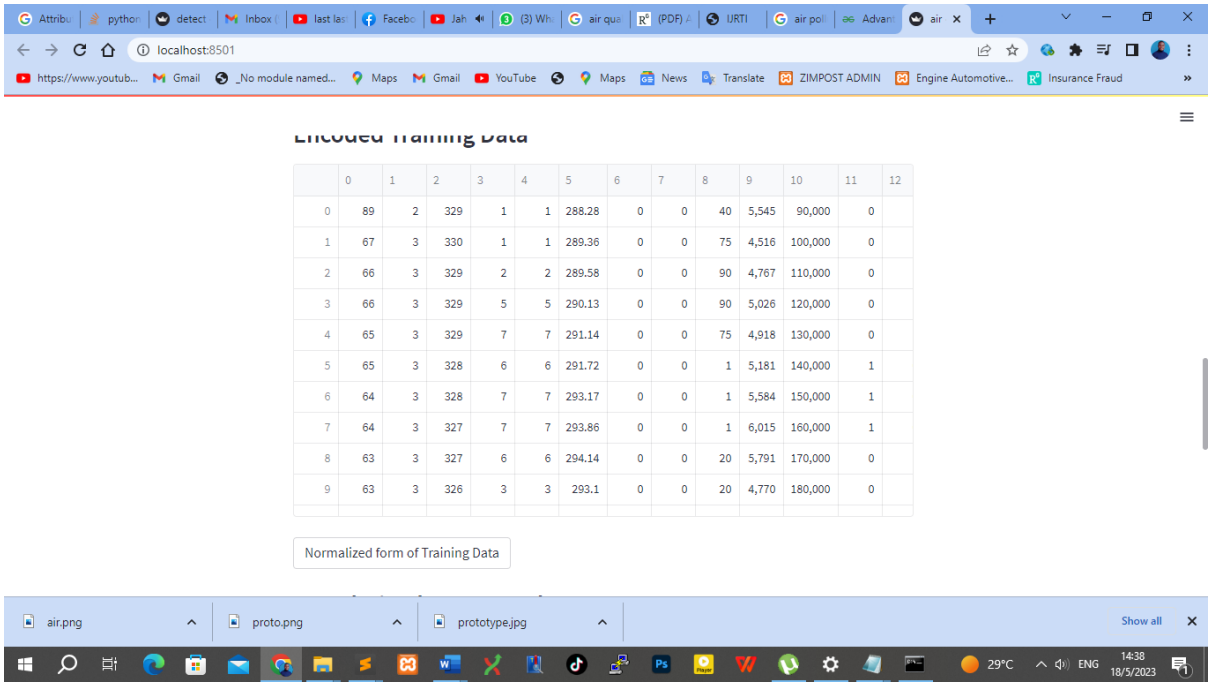


Correlation between columns

	humidity	wind_speed	wind_direction	visibility_in_miles	dew_point	temperature	rai
humidity	1	-0.147	-0.0463	0.0055	0.0055	0.0333	-(
wind_speed	-0.147	1	0.1913	-0.0052	-0.0052	-0.0529	(
wind_direction	-0.0463	0.1913	1	0.0007	0.0007	-0.0458	(
visibility_in_miles	0.0055	-0.0052	0.0007	1	1	0.0008	-(
dew_point	0.0055	-0.0052	0.0007	1	1	0.0008	-(
temperature	0.0333	-0.0529	-0.0458	0.0008	0.0008	1	(
rain_p_h	-0.0122	0.0012	0.0022	-0.0083	-0.0083	0.0111	
snow_p_h	0.0166	-0.0065	0.0003	0.0015	0.0015	-0.0207	-(
clouds_all	0.0145	-0.0045	0.0204	-0.0016	-0.0016	-0.1213	(
air_pollution_index	-0.0003	-0.0042	0.0007	0.0035	0.0035	0.0054	(

Choose number of estimators for Random Forest Algorithm

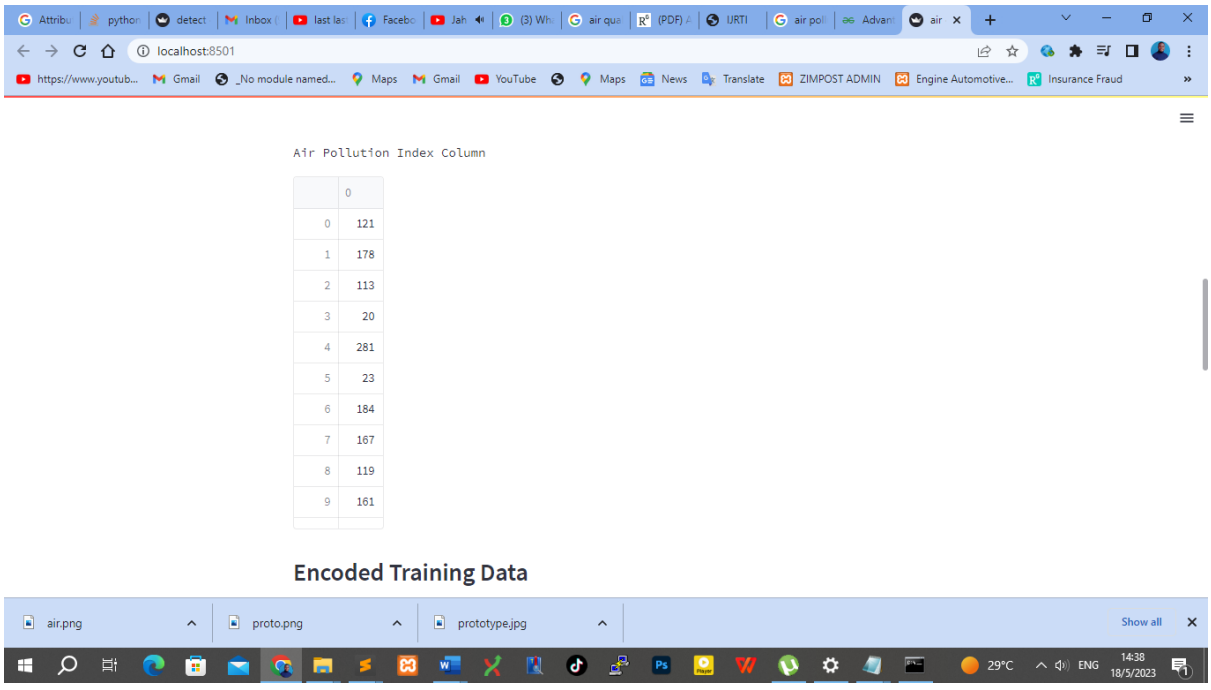
3



Encoded Training Data

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	89	2	329	1	1	288.28	0	0	40	5,545	90,000	0	
1	67	3	330	1	1	289.36	0	0	75	4,516	100,000	0	
2	66	3	329	2	2	289.58	0	0	90	4,767	110,000	0	
3	66	3	329	5	5	290.13	0	0	90	5,026	120,000	0	
4	65	3	329	7	7	291.14	0	0	75	4,918	130,000	0	
5	65	3	328	6	6	291.72	0	0	1	5,181	140,000	1	
6	64	3	328	7	7	293.17	0	0	1	5,584	150,000	1	
7	64	3	327	7	7	293.86	0	0	1	6,015	160,000	1	
8	63	3	327	6	6	294.14	0	0	20	5,791	170,000	0	
9	63	3	326	3	3	293.1	0	0	20	4,770	180,000	0	

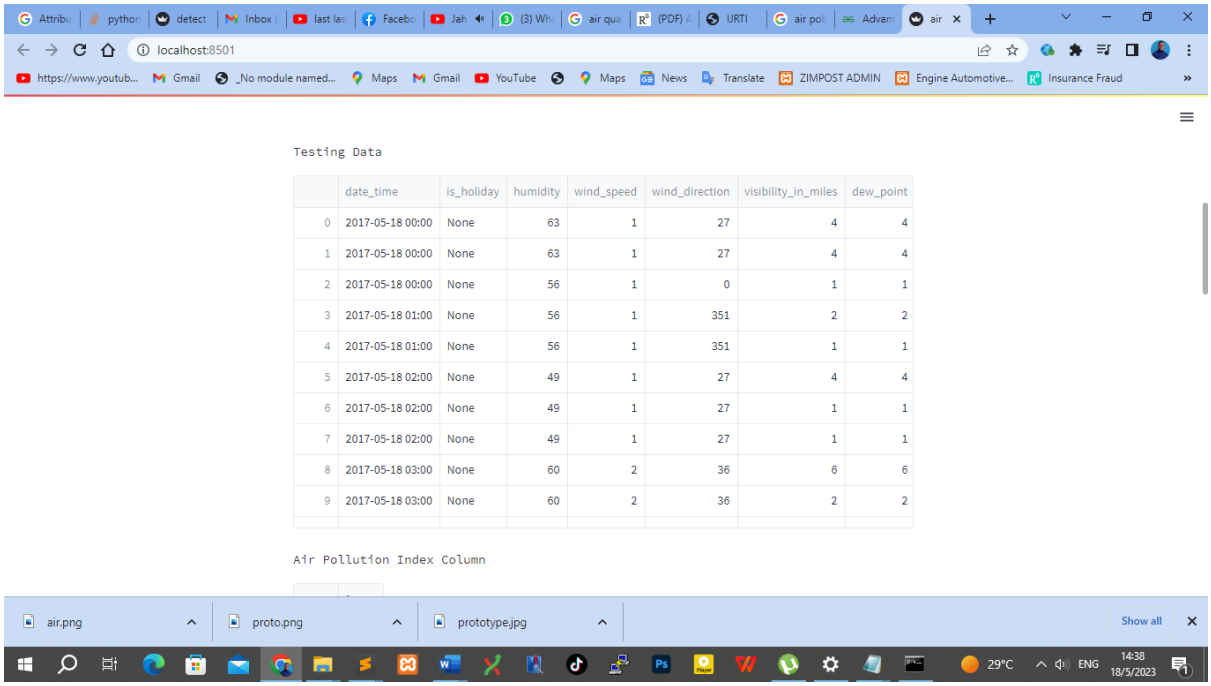
Normalized form of Training Data



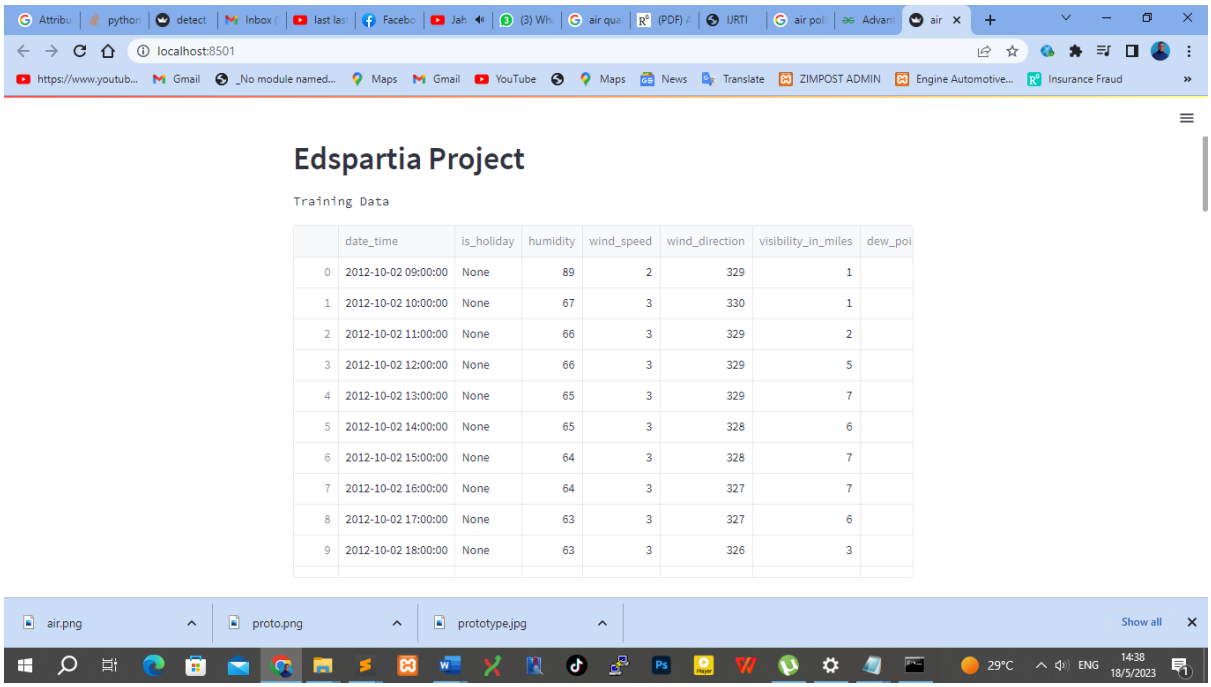
Air Pollution Index Column

	0
0	121
1	178
2	113
3	20
4	281
5	23
6	184
7	167
8	119
9	161

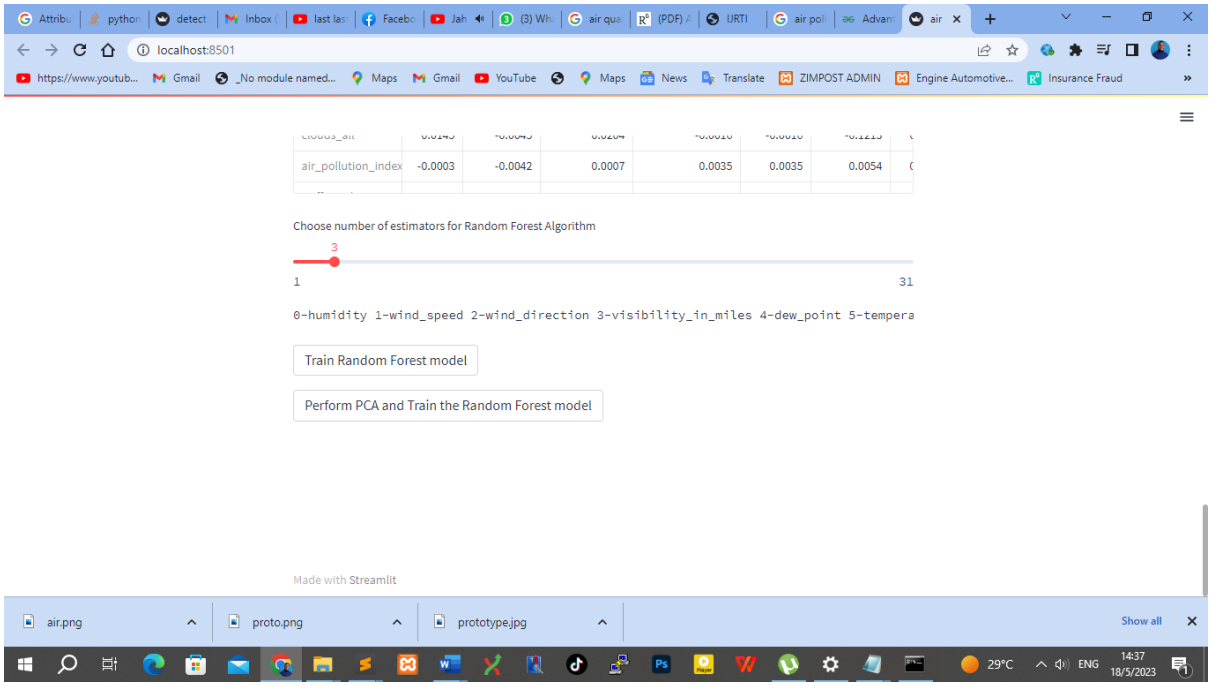
Encoded Training Data



	date_time	is_holiday	humidity	wind_speed	wind_direction	visibility_in_miles	dew_point
0	2017-05-18 00:00	None	63	1	27	4	4
1	2017-05-18 00:00	None	63	1	27	4	4
2	2017-05-18 00:00	None	56	1	0	1	1
3	2017-05-18 01:00	None	56	1	351	2	2
4	2017-05-18 01:00	None	56	1	351	1	1
5	2017-05-18 02:00	None	49	1	27	4	4
6	2017-05-18 02:00	None	49	1	27	1	1
7	2017-05-18 02:00	None	49	1	27	1	1
8	2017-05-18 03:00	None	60	2	36	6	6
9	2017-05-18 03:00	None	60	2	36	2	2



	date_time	is_holiday	humidity	wind_speed	wind_direction	visibility_in_miles	dew_poi
0	2012-10-02 09:00:00	None	89	2	329	1	
1	2012-10-02 10:00:00	None	67	3	330	1	
2	2012-10-02 11:00:00	None	66	3	329	2	
3	2012-10-02 12:00:00	None	66	3	329	5	
4	2012-10-02 13:00:00	None	65	3	329	7	
5	2012-10-02 14:00:00	None	65	3	328	6	
6	2012-10-02 15:00:00	None	64	3	328	7	
7	2012-10-02 16:00:00	None	64	3	327	7	
8	2012-10-02 17:00:00	None	63	3	327	6	
9	2012-10-02 18:00:00	None	63	3	326	3	



3.18 SUMMARY OF HOW THE SYSTEM WORKS

This chapter mainly focused on the methodology used in the development of the system and how it was designed as well as implemented. Different techniques were used to come up with the system, also different tools like the python and different Bayesian algorithms made it possible to come up with the proposed system.

CHAPTER 4: RESULTS AND ANALYSIS

4.0 INTRODUCTION

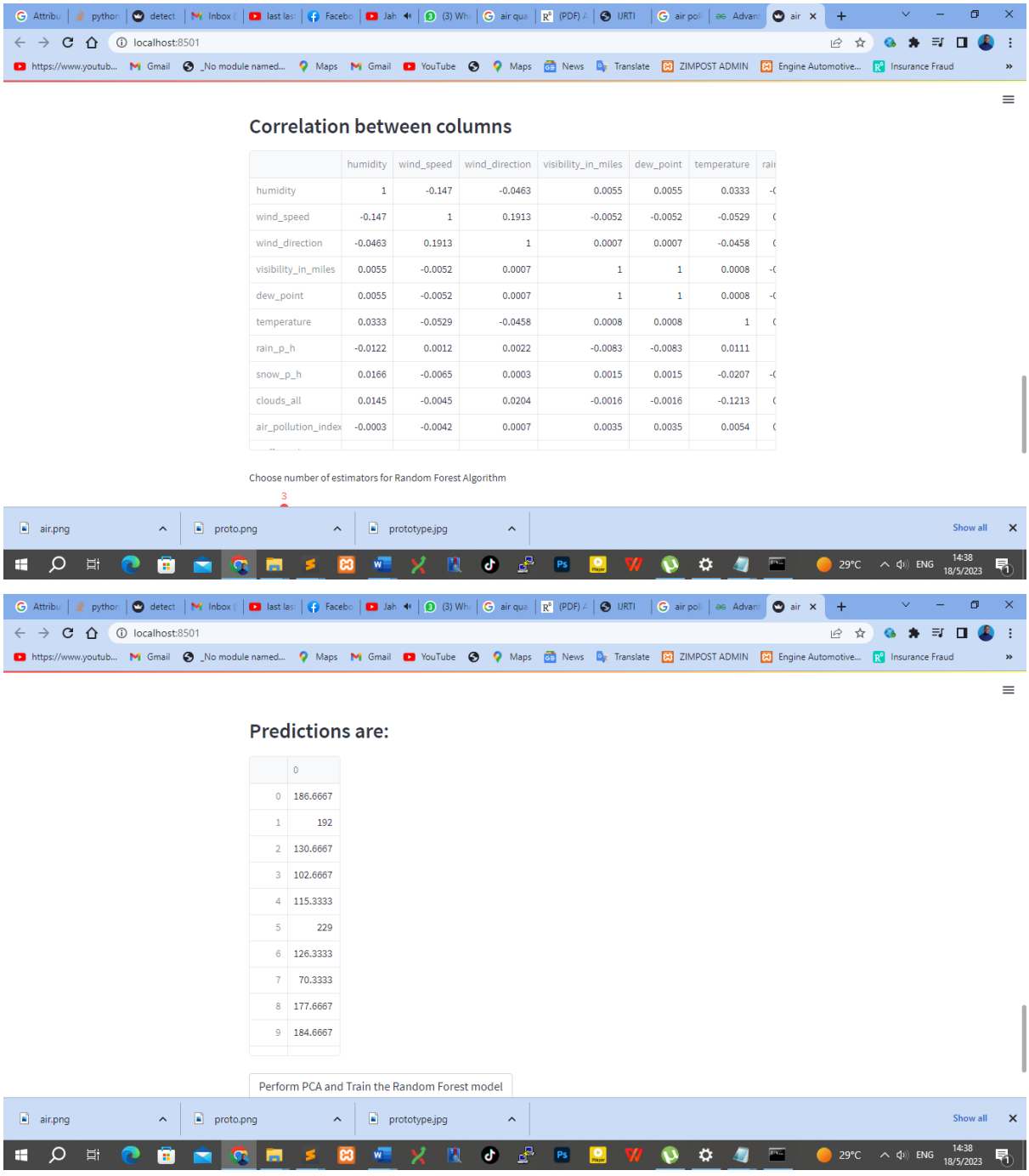
The necessity to assess the effectiveness of the created solution occurred after the author had successfully deployed the system. The matrices utilized to evaluate the efficiency and efficacy of the produced solution were accuracy, performance, and reaction time. The behavior of the generated solution was thoroughly examined at various periods, and the results were shown in a tabular manner.

4.1 TESTING

This chapter demonstrates the tests that were conducted and the outcomes they produced. Testing is an essential step in the development process. As a result, the testing is evaluated in relation to the functional and non-functional requirements listed in the preceding chapter.

4.1.1 BLACK BOX TESTING

Black box testing enables a user to evaluate a system against its functional and occasionally non-functional needs without having any knowledge of the system's underlying structure. It mainly served to remind the patient of the proper time to take their medications and the effects of taking them on time or not at all. Thus, the primary goal of black box testing was to determine if the system performed as predicted by the requirements. (Beizer, B.,1995).



Screenshot 1: Predictions

4.1.2 FUZZY TESTING

The researcher utilized fuzzy testing, a type of black box testing, on the weather forecasting application to see if the system properly responded and provided the right results for the specified coordinates. (Singh et al.,2023).

4.2 EVALUATION MEASURES AND RESULTS

A classifier's performance is measured by an evaluation metric (Hossin & Sulaiman, 2015). Additionally, Hossin & Sulaiman (2015) assert that there are three categories of model assessment metrics: threshold, probability, and ranking.

4.2.1 Measuring System Performance

The performance of the system is ranked according to its ability to give a real time feedback as per given dataset.

Machine 1(4 gig Ram,Core i3 500gb)

Test Runs	1	2	3	4	5
Time(s)	150	145	151	164	149

Mean Value for the performance of the system on machine 1

150+145+151+164+149=759/5

=151.8 seconds

Machine 2(8 gig Ram,Core i3 1terabyte)

Test Runs	1	2	3	4	5
Time(s)	69	78	76	67	80

Mean Value for the performance of the system on machine 1

69+78+76+67+80=370/5 =74 seconds

Measuring Supervised Machine Learning to previous algorithms

Algorithms	Linear Regression	Decision Tree	Random Forest
Accuracy	0.99	0.98	0.99

Table 1: Confusion Matric

Type	Bad Weather	Good Weather
Bad Weather	True Positive	False Negative
Good Weather	False Positive	True Negative

Three scenes and test environment were created for observation of the system. On each scene the system was observer on 40 occasions 20 days were good weather and 20 days were bad weather and the behavior of the system was observed. All the analysis on the scenes was carried out to test for the solution's accuracy and elimination of false prediction. The tables below show the observed results from the tests carried out.

Table 2: During Morning

Test cases	Low Air Pollution Index	Number of days	Correct readings	False Readings	Classification
1	Yes	20	16	4	True positive
2	No	20	18	2	True negative

Table 3: During Afternoon

Test cases	Low Air Pollution Index	Number of tests	Correct readings	False Readings	Classification
1	Yes	20	14	6	True positive
2	No	20	17	3	True negative

Table 4: During Evening

Test cases	Low Air Pollution Index	Number of tests	Correct readings	False Readings	Classification
1	Yes	20	18	2	True positive
2	No	20	17	3	True negative

4.2.2 Accuracy

Accuracy is calculated as the total number of forecasts in each category divided by the number of correct predictions. It is then multiplied by 100 to determine the accuracy percentage. It is computed via the following equation:

Equation 1: Accuracy calculation as adopted from Karl Pearson (1904)

Accuracy = (TP+TN)/ (TP+TN+FP+FN) *100

Accuracy during morning = (16+18)/ (20+20+0+0)

=0.85

=0.85*100= 85%

Accuracy in the afternoon = (14+17)/ (14+17+3+6) *100

= 76%

Accuracy during evening = (18+17)/ (20+20+0+0) *100

=0.88*100

=88%

Average Accuracy rate = Accuracy at (spring + winter + summer) /3

= (85+88+76)/3 *100 = 295/3 *100

=83%

4.3 Conclusion

The test results indicated the solution had a high level of accuracy since in 2 scenes it produced 87% and 85 % rate of accuracy respectively which was a result of the analysis of the confusion matrix. However, the solution had an eighty and eighty (88%) percent accuracy during the spring this was due to the high levels of wind and insufficient training data and proper environment exposure. The high levels of accuracy of the system indicate a reduction of false prediction on air pollution index.

CHAPTER 5: RECOMMENDATIONS AND FUTURE WORK

5.1 Introduction

The researcher concentrated on the presentation and analysis of the data in the preceding chapter. The investigation and development of the solution in accordance with the predetermined objectives are covered in this chapter. The research was successful, but the chapter also looked at the challenges the researcher had when planning and carrying out the study.

Town and City councils are greatly in need for proper and advanced digitization procedures with the help of Meteorological departments with problems associated with weather notifications and forecasting. Therefore, the author's research focused on supervised machine learning model to aid the meteorological department in determining whether the weather is fine or not on a particular day using air quality index. This was a success and therefore the researcher recommends the implementation and advancement of this research.

5.2 Aims and Objectives Realization

In summary, the objectives of this study was to evaluate the system performance on predicting air pollution on a city whether the weather is fine or not. This was by using different terminologies. The objective of the system was to use supervised machine learning and Bayesian networks to help achieve developing the application. The application uses a dataset. Therefore, the factors on dataset are to be used to predict if the weather is fine or not. The objectives were met and the system was performing well.

5.3 Conclusion

The combination of supervised machine learning and a Bayesian network aids the author in attaining the optimal outcome for the study endeavor. The software was able to forecast whether the weather will be good or bad on a certain day. The system demonstrated that it outperformed every benchmark established by the experimental control at a scaled-down version.

5.4 Recommendations

There is need for a greater coverage and quality dataset to cover a lot to predict air quality index which shows that the weather is fine or not. In this research the data used by the author is static which means the data remains the same after it's collected. However, some organizations update data in time bases. Further we can use real time data analysis using cloud to obtain better outcomes as the updates for every time interval. Furthermore, we can also use two or more machine learning algorithms and process large data to get more accurate results.

REFERENCE

1. Yan, R., Liao, J., Yang, J., Sun, W., Nong, M. and Li, F., 2021. Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering. Expert Systems with Applications.
2. Kumar, D., 2018. Evolving Differential evolution method with random forest for prediction of Air Pollution. Procedia computer science.
3. Simu, S., Turkar, V., Martires, R., Asolkar, V., Monteiro, S., Fernandes, V. and Salgaoncary, V., 2020, December. Air Pollution Prediction Using Machine Learning. In 2020 IEEE Bombay Section Signature Conference (IBSSC).
4. Russo, A. and Soares, A.O., 2014. Hybrid model for urban air pollution forecasting: A stochastic spatio-temporal approach. Mathematical Geosciences.
5. Masood, A. and Ahmad, K., 2021. A review on emerging artificial intelligence (AI) techniques for air pollution forecasting: Fundamentals, application and performance. Journal of Cleaner Production.
6. Shaban, K.B., Kadri, A. and Rezk, E., 2016. Urban air pollution monitoring system with forecasting models.
7. Hong, H., Choi, I., Jeon, H., Kim, Y., Lee, J.B., Park, C.H. and Kim, H.S., 2022. An Air Pollutants Prediction Method Integrating Numerical Models and Artificial Intelligence Models Targeting the Area around Busan Port in Korea. Atmosphere.
8. Somvanshi, M., Chavan, P., Tambade, S., and Shinde, S., 2016 "A review of machine learning techniques using decision tree and support vector machine," in 2016 international conference on computing communication control and automation (ICCUBEA).
9. Z. Ghahramani, Z., 2003 "Unsupervised learning," in Summer school on machine learning. Springer.
10. Zhu, X., and A. B. Goldberg, A.B., "Introduction to semi-supervised learning," Synthesis lectures on artificial intelligence and machine learning, vol. 3.
11. Chang, S.W., Chang, C.L., Li, L.T. and Liao, S.W., 2019. Reinforcement learning for improving the accuracy of pm2. 5 pollution forecast under the neural network framework.
12. Grady, J.O., 2010. System requirements analysis. Elsevier.

13. Beizer, B., 1995. Black-box testing: techniques for functional testing of software and systems. John Wiley & Sons.
14. Singh, V., Kumar, V. and Singh, V.B., 2023. A hybrid novel fuzzy AHP-Topsis technique for selecting parameter-influencing testing in software development.
15. Lee, J., 2019. Air pollution prediction using machine learning techniques: A review. International Journal of Environmental Research and Public Health.
16. Cimorelli, A. J., Perry, S. G., Venkatram, A., Weil, J. C., Paine, R. J., Wilson, R. B., ... & Lee, R. F., 2005. AERMOD: A dispersion model for industrial source applications. Part I: General model formulation and boundary layer characterization. Journal of Applied Meteorology.
17. 3. Liu, Y., & Zhang, Y., 2021. Ensemble modeling for air quality prediction: A review. Atmospheric Environment.