Analyzing Area-wise Air Pollution Level using Machine Learning for a Better Future

By

Sk. Atik Tajwar Sihan
17301109
Maisha Rabbani
19201123
Manish Agarwala
17301120
Sanjida Alam Maliha
20301453

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

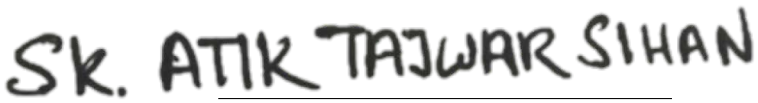
Department of Computer Science and Engineering
Brac University
September 2021

Declaration

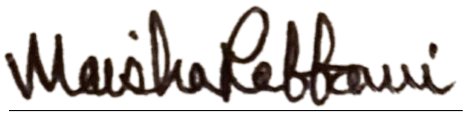
It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

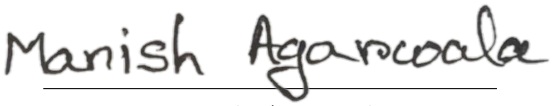
Student's Full Name & Signature:



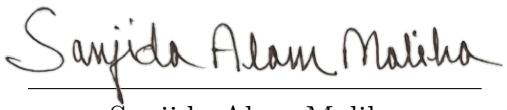
Sk. Atik Tajwar Sihan
17301109



Maisha Rabbani
19201123



Manish Agarwala
17301120



Sanjida Alam Maliha
20301453

Approval

The thesis/project titled "Analyzing Area-wise Air Pollution Level Using Machine-Learning for a Better Future" submitted by

1. Sk. Atik Tajwar Sihan (17301109)
2. Maisha Rabbani (19201123)
3. Manish Agarwala (17301120)
4. Sanjida Alam Maliha (20301453)

Of Summer, 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on October 02, 2021.

Examining Committee:

Supervisor:
(Member)

Md. Saiful Islam
Lecturer
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Ethics Statement

We have assured complete transparency of our evaluation process in addition to providing visual interpretation of the output generated by the model.

Abstract

Environment consists of nature and surroundings where all living beings co-exist. Harming the environment will in turn harm all living and non-living things alike. One of the major concerns of environment pollution is air pollution, which affects human health, vegetation and aquatic life. However, in developing countries like Bangladesh, air pollution is not considered a major issue. It is mostly caused by the release of harmful gases into the atmosphere. Our goal is to develop a model using machine learning which will determine the level of air pollution in a particular area, detect elements which cause air pollution and predict future pollution level. Algorithms such as Linear Regression, Facebook Prophet, RNN and ARIMA models have been used throughout the course of this study. From RNN we have used LSTM model for prediction which uses special units as well as standard units. With these models we have predicted the pollutant emission rate for analyzing the area-wise pollution rate. We have used different type of algorithms to successfully get the optimum result and to get the final result with less error. This will help to analyze the overall air pollution condition which will help to take necessary steps accordingly.

Keywords: Environment, Air Pollution, Pollutants, Machine Learning, Linear Regression, Facebook Prophet, RNN, LSTM, ARIMA

Dedication

We are extremely grateful to our loving parents for being our source of strength and motivation all the way.

Acknowledgement

We would like to acknowledge our supervisor, Md. Saiful Islam, and our co-supervisors, Nadia Rubaiyat and Rafeed Rahman, and express our sincere gratitude for their constant support throughout the year. Without their valuable guidance and continuous feedback, we would not be able to come this far.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

$ADFC$ Augmented Dicky-Fuller

$ARIMA$ Autoregressive Integrated Moving Average

CTM Chemical Transport Model

$FB\ Prophet$ Facebook Prophet

$LSTM$ Long Short Term Memory

MAE Mean Absolute Error

$MAPE$ Mean Absolute Percentage Error

$RMSE$ Root Mean Square Error

RNN Recurrent Neural Network

SVR Support Vector Regression

Chapter 1

Introduction

1.1 Overview

The presence of harmful substances in the atmosphere, emitted by both natural and human sources is known as air pollution. A country is required to have 25% area of forest to produce oxygen, which is essential for life. However, the total area of forests in Bangladesh has been reduced to 7-9% [12]. This has affected the quality of air in Bangladesh. According to the 2019 World Air Quality Report, Bangladesh has been found to have the worst air quality and its capital city, Dhaka, has the second-worst air quality among capital cities. The rapid growth of industrialization, private transportation, and deforestation are the major factors that contribute to an increase in the level of harmful particles in the air. The government of Bangladesh is encouraging the people not to cut trees unnecessarily as we all know trees play a key role to produce oxygen. Bangladesh is the 9th most populous country with a population of over 10 million. The Air Quality Index (AQI) by U.S Embassy in Bangladesh states that Bangladesh falls in the 'Very Unhealthy' level (285) in the charter of health concern, so it is very certain that people of Bangladesh face various kinds of heart and lungs disease because of the air pollution. It is a health warning that the entire population can be affected. So, in this study, we are working to create a model to predict the level of pollutants in the air by the year 2030. The model will use various datasets to obtain a high accuracy prediction that should help people to take necessary actions to prevent air pollution.

1.2 Research Objectives

One of the major dilemmas in modern times is air pollution. Air is important for all living beings including plants, animals, and other organisms. According to World Air Quality, Bangladesh is considered to have the most polluted air in terms of living for humans [21]. Average human lifespan has become 1.8 years shorter because of air pollution [19]. The atmosphere keeps the Earth at a habitable temperature for people to live. But due to air pollution, our ecosystem is facing dramatic change. Air pollution is the main cause of climate change which also leads to an increase in temperature. As the temperature increases, the glaciers melt causing the sea level to rise. There are six pollutants which are the main cause of air pollution, and those are carbon monoxide, nitrogen dioxide, sulfur dioxide, ozone, particulate matter, and lead. Air pollution damages the lungs, liver, kidneys, as well as causes

lung cancer, heart disease, and respiratory diseases. Thus, air pollution poses a deadly threat. The awareness of the dangers of air pollution is not widespread in Bangladesh. The motive behind this paper is to make people conscious of the effects of air pollution with proper knowledge and proof. When more people are aware of the risks, change is bound to happen. Our aim is to predict the level of pollutants in the air for different areas with the highest accuracy. For that, we have used machine learning algorithms such as Time series, linear regression method. If people are aware of the prediction of air pollution level of certain areas, they might consider staying away from those areas. This will lower the number of people affected due to air pollution. Moreover, seeing the prediction level, laws and regulations can be implemented which will regulate the amount of pollutants released in the air. In this way, we can help improve the air quality.

1.3 Problem Statement

Air pollution is a major concern which results in the loss of life. It is not possible to explain the level of damage due to air pollution. Air pollution is harmful for human health. It leads to respiratory and cardiovascular diseases as well as lung damage. A study by Pope III, C., [1] has shown that long-term exposure to fine particulate in the air leads to higher risks of both cardiopulmonary and lung cancer mortality. Presence of a pollutant such as PM2.5 increased the risk of cardiopulmonary and lung cancer mortality by a huge margin. Follow-up study claimed that the number of deaths tripled. Further studies by Andersen, Z., Hvidberg, M., Jensen, S., Ketzel, M., Loft, S., Sørensen, M., Tjønneland, A., Overvad, K. and Raaschou-Nielsen, O., [3] claim that exposure to NO2 and NOx in the air for more than thirty-five years lead to a rise in chronic obstructive pulmonary disease (COPD), especially in patients diagnosed with diabetes and asthma. In countries like Bangladesh, air pollution is caused by burning of fossil fuels, emission of harmful gases from factories and exhaust from vehicles. According to a study by Dasgupta, S., Huq, M., Khaliquzzaman, M., Pandey, K. and Wheeler, D., [2], young children and women from lower income households suffer from prolonged exposure to air pollution than men and women from higher income households with more education. This is due to household pollution caused by primitive cooking methods and lack of proper ventilation in lower income households. In this paper, machine learning algorithms such as linear regression and time series forecasting have been used to determine the level of air pollution.

Chapter 2

Related Work

To measure the extent of pollution we use machine learning and deep learning algorithms. We looked at existing research and found a lot of study has been carried out in this sector.
Rubal and Kumar, D., [14] have proposed a hybrid technique in order to predict pollutants in the air. This technique is the combination of differential evolution and random forest method. This proposed technique surpasses the more commonly used methods with higher area under curve, higher success index, higher accuracy, lower cost as well as higher correlation . The accuracy is 80%.
Siwek, K. and Osowski, S., [9] have used algoritms such as genetic and stepwise fit selection feature. They compared the outputs with the outputs of correlation of single feature. The experiment showed that using the two algorithms together had a higher accuracy of prediction. However, correlation of the single feature led to much lower level of accuracy.
Castelli, M., Clemente, F., Popovi˘c, A., Silva, S., and Vanneschi, L. [17] have created an Support Vector Regression (SVR) model which is a variant of Support Vector Machines (SVMs), to predict the extent of pollutants so as to predict the Air Quality Index (AQI). The accuracy was 94.1%.
Azar, A., Elshazly, H., Hassanien, A., Elkorany, A., [5] proposed a hybrid system which combines the feature selection phase with the classification phase. The feature selection phase is made up of the Genetic Algorithm method while the classification phase uses the Random Forest method. This hybrid model obtained 92.2% accuracy.
Shakerkhatibi et al., 2015 [8], in their paper used Artificial Neural Network and Conditional Logistic Regression in order to demonstrate the relationship between polluted air and cardiorespiratory diseases. Hospital admission data and air quality data were analyzed using ANN and LR which showed a notable connection between the two. The study showed that exposure to nitrogen oxides lead to cardiovascular diseases while exposure to particulate matter causes respiratory infections. The results of LR and ANN were compared which showed ANN gave better predictions.
Azid et al., 2014, [6] Principal Component Analysis (PCA) and Artificial Neural Network (ANN) to predict the level of air pollution. They used PCA to pinpoint the place of origin of the major air pollutants. A combination of PCA and ANN was created to predict the air pollution level. PCA-ANN produced better results than PCA.
Zhao et al., 2018 [16], in their paper, used Recurrent Neural Network (RNN), Random Forest (RF) and Support vector machines (SVM) to predict Air Quality Classi-

fication (AQC). They proposed building a better prediction model called RNN-AQC. As the RNN-AQC name suggests, they used RNN to measure AQC. RNN was selected as it can work on non-linear data resolving sequential problems. AQC was used instead of more commonly used AQI as AQC helps to determine the level of intensity of air pollution. The proposed RNN-AQC model produced a higher accuracy compared to SVM and RF.

The k-nearest neighbors (KNN) algorithm is a supervised machine learning algorithm which was used by Saad et al., 2017 [10] to measure the level of indoor air pollutants. Indoor air pollution is caused by tobacco smoke, cooking oil and other household cleaning products. Indoor air quality data was collected on which data preprocessing and feature selection was carried out before it was used as the input for the classification. The KNN model scored an accuracy of 97%.

Heydari, A., Nezhad, M. M., Garcia, D. A., Keynia, F., and De Santoli, L. [20] proposed a new hybrid intelligent model based on Mutual Information (MI), Elman neural network (ENN), long short-term memory (LSTM) and multi-verse optimization algorithm (MVO) that has been used to predict air pollution. The study shows that they used time series algorithm to predict which model performs well. The result of MI, LSTM, MVO gave better prediction.

Yeganeh, B., Motlagh, M. S. P., Rashidi, Y., and Kamalan, H. [4] used support vector machine (SVM) for the prediction and partial least square (PLS) for data selection tool for their dataset. The data had multiple parameters such as air pressure, temperature, wind speed, direction and air humidity. The results they got had good accuracy for both of the methods but shows even better accuracy for hybrid PLS–SVM.

In this article, Aditya C, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu [11] explain how PM2.5 has affected our health. They explain how it plays an important role in the pollutant index because it will be a major concern for human health if its level arises. Dan Wai has applied Naive Bayes classification and support vector machine algorithms (SVM) to ensure the highest accuracy in predicting the air condition in Beijing city. The main objective of their system was to measure the PM2.5 levels and make predictions on a particular date. They use Logistic regression to confirm if the sample is clean or not and they predict the future PM2.5 value using autoregression. Their dataset contains six futures and they are Temperature, Wind speed, Dew point, Pressure, PM2.5 Concentration, Result (data sample is classified either to be polluted or not polluted). They have found out that logistic regression is best fitted for their system. Moreover, they have acquired 0.998859 and 0.000612 for mean accuracy and standard deviation accuracy. To get this result they use autoregression on their time-series dataset and predict 7 days prior PM2.5 value. As a result, they come to the conclusion that logistic regression and autoregression both can be efficiently used to detect the quality of air and predict the level of PM2.5 in the future.

The issues of linear regression were explained in details in the article by Kerckhoffs, et al. In their article, they tell us that linear regression expected three linearities of predictor and they are potential interactions, pollution relationships, and using predictors which show vast correlation therefore may not identify the optimal model. Likewise, they expressed that neural networks11, irregular forests12, and other machine learning techniques offer prospects to make methods for pollutants of air by learning the basic connections in a train set, with no predefined constrictions. In

this article, the studied and looked at modeling techniques which are 11 in numbers for foreseeing spatiotemporal flexibility of PM 2.5 fixations in the course of rapidly spreading forest fire occasions and discovered that Random Forest, General Boosting, and SVM performed in a way that is better than linear regression modeling. However, Van den Bossche et al referenced in their article that to make black carbon LUR model observing project there are no critical differences between LASSO, linear regression, and SVR. They likewise analyzed a lot bigger number of algorithms and the various discoveries in their examination for short-term training data and to some degree conflicting discoveries in past studies foreclose the judgment on the empirical performance of models.

The findings above helped us understand how drastic the air is changing due to pollution and its effect on human health. Also, it encourages us to build our own model and a suitable dataset which will help us to understand the change in air particles and the effect it will have in our future.

Chapter 3

Methodology

We have selected a dataset containing values of different harmful gases and AQI values of different cities in India from 2015 to 2020. We will use this dataset to run our models for calculation and prediction. We will use the following algorithms for our prediction:

- Facebook Prophet
- Linear Regression
- RNN (LSTM)
- ARIMA

3.1 Facebook Prophet

We have used Facebook Prophet algorithm for time series forecasting. It uses Stan for optimization to fit a non-linear additive model and generating uncertainty interval [18]. Many linear and non-linear functions are used as parts in Prophet. It is an additive regression model. We need to find,

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{3.1}$$

where, piece-wise linear or logistic growth curve is denoted by $g(t)$, periodic changes and holidays effects are denoted by $s(t)$ and $h(t)$ accordingly, error terms denoted by ϵ_t. The logistic growth model is fit using the following statistical equation,

$$g(t) = \frac{C}{1 + e^{-k(t-m)}} \tag{3.2}$$

Where, carry capacity, growth rate, offset parameter is denoted by C, K and M accordingly. Using the following statistical equations Piece-wise linear model is fitted,

$$y = \begin{cases} \beta_0 + \beta_1 x & x \leq c \\ \beta_0 - \beta_2 c + (\beta_1 + \beta_2)\, x & x > c \end{cases} \tag{3.3}$$

By the following function Seasonal effects s(t) are approximated,

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right) \right) \tag{3.4}$$

The period is denoted by P and we need to estimate the Parameters such as a_1, b_1 for seasonality modelling.

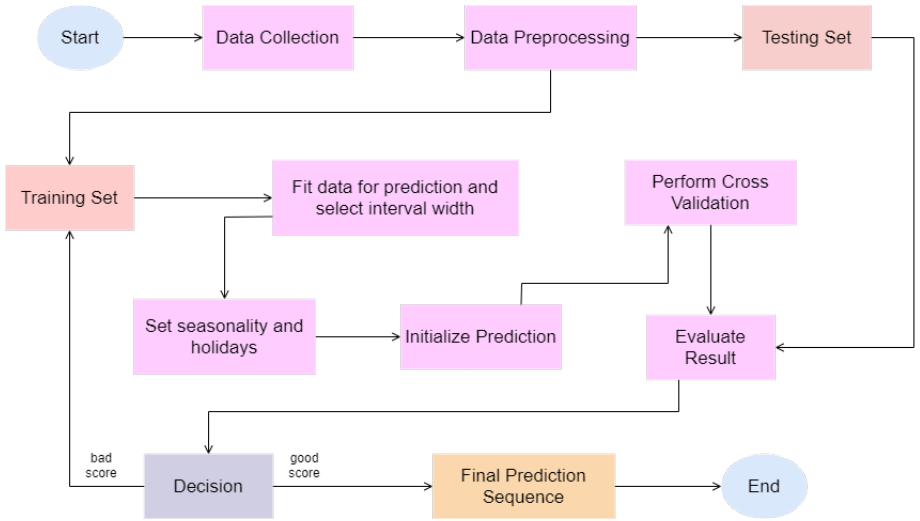


Figure 3.1: Workflow Diagram of Facebook Prophet

3.2 Linear Regression

Linear regression is a supervised machine learning algorithm using which we can determine the independent and dependent variables. The dependent variable is x which is plotted on the x-axis and the independent variable is y which is plotted on the y-axis. It demonstrates a relationship between the dependent and independent variables using the line of best fit. Regression analysis technique can be used to predict the concentration of different elements in the environment. The formula for linear regression can be expressed as:

$$\hat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x \tag{3.5}$$

Here, the y-intercept is denoted as $\widehat{\beta}_0$ and the slope is denoted as $\widehat{\beta}_1$.

This algorithm is used to find out the value of y based on the given value of x, using the linear regression equation. However, there is bound to be some error. So, we construct a line of best fit in order to minimize the error. We can calculate the error using the formula:

$$e_i = y_i - \hat{y}_i \tag{3.6}$$

where, y_i = actual value of y for the nth observation $\hat{y}_i$ = and predicted value of y for the nth observation.

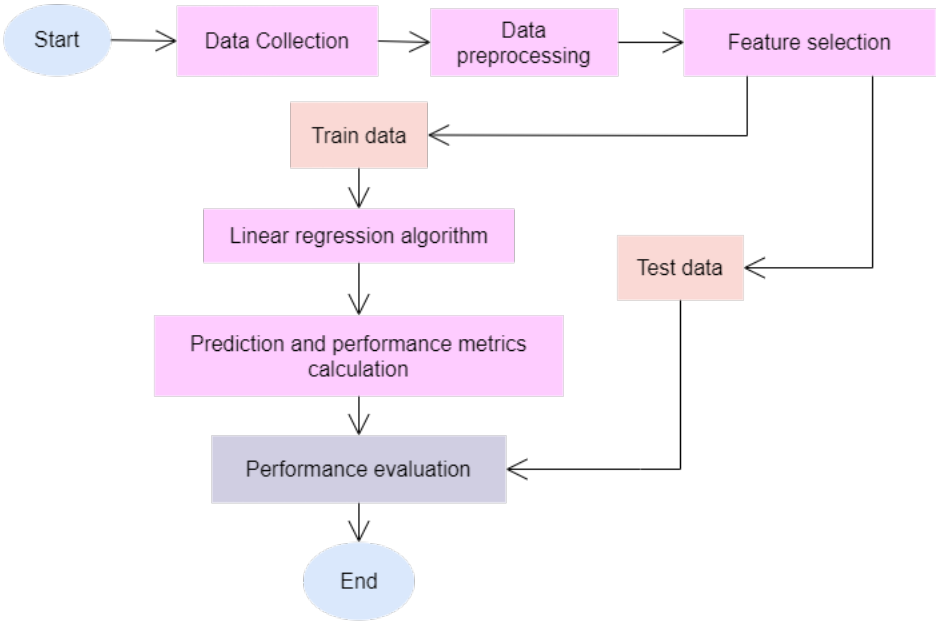


Figure 3.2: Workflow Diagram of Linear Regression

3.3 RNN (LSTM)

RNN is a networking technique that is able to model any sequence of information like the Time Series Algorithm or Natural Languages [15]. Long Short-Term Memory (LSTM) is extension of RNN which is in a position to extend memory. For RNN models, LSTM are used as the building blocks. Like several other deep learning algorithms that we all know, RNN are comparatively old but efficient. Nowadays, increasing computing power, large number of data and the support of LSTM, has brought RNN to the spotlight.
Firstly, in LSTM, we select the data we want to remove using the following formula,

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{3.7}$$

The sigmoid function is used to achieve this. It uses its previous state h_{t-1} and its current input x_t to determine the next function.

$$\begin{aligned} i_t &= \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \\ \tilde{C}_t &= \tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right) \end{aligned} \tag{3.8}$$

The next part consists of the sigmoid function we used previously and the tanh function. The sigmoid function is responsible to select which values to keep and which ones to discard. The tanh function is responsible to attach weight to the values which are not discarded depending on their level of importance.

$$\begin{aligned} o_t &= \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \\ h_t &= o_t * \tanh\left(C_t\right) \end{aligned} \tag{3.9}$$

In the last step, we determine the output of the sigmoid gate o_t. The final output h_t is the product of the output of the sigmoid gate and tanh function containing the cell value.

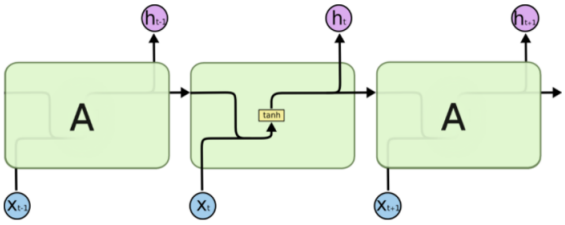


Figure 3.3: Workflow Diagram of RNN(LSTM)

3.4 ARIMA

For our time series analysis, we have used an autoregressive integrated moving average (ARIMA) model. ARIMA is used in order get more insight of the data or it is used for prediction.
In ARIMA, AR implies that the variable is relapsed on its own prior values. MA represents linear sum of error terms [7]. Here, I means integrated. The difference between the current value and the old values are represented by I.
The ARIMA hyperparameters defines this polynomial factorisation property with $p = p' - d$, and is given by:

$$\left(1 - \sum_{i=1}^{p} \varphi_i L^i\right) (1 - L)^d X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t \tag{3.10}$$

In this formula, t is an integer index and X_t is a real number. Also L is the lag operator, the θ_i are use as the parameters for the moving average step and ε_t stands for error terms.

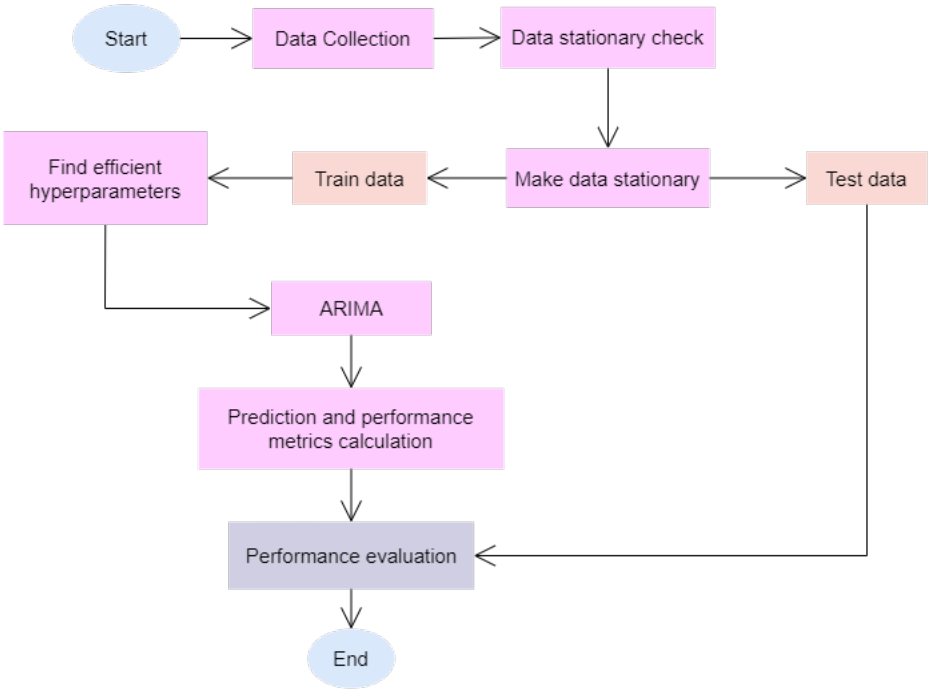


Figure 3.4: Workflow Diagram of ARIMA

Chapter 4

Work Plan

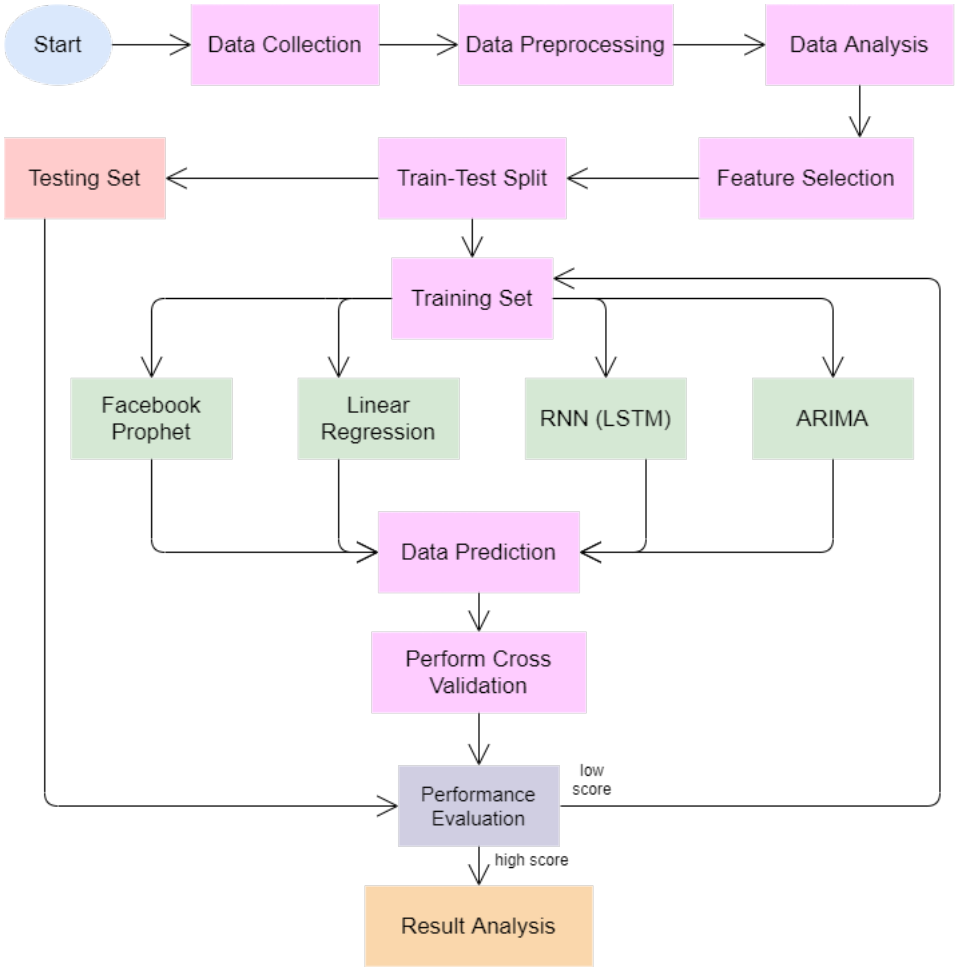


Figure 4.1: Workflow Diagram

Firstly, we acquire a dataset. The data is preprocessed and analyzed by removing null and redundant values. A feature is selected for prediction. Then we split our dataset into test and train. We input the training set into our models which then give us the prediction values. We perform cross-validation and evaluate the performance of the predictions. Then we run our models on the test set and check the performance. If the model has low scores, we go back to the training step and repeat. If it has high scores, we can say our model is suited for data prediction.

Chapter 5

Experiment and Analysis

5.1 Data Collection

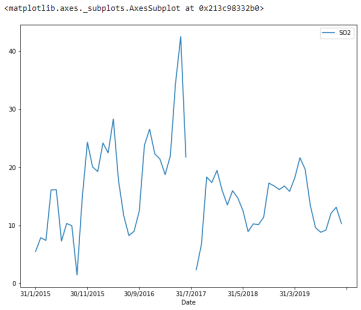
When it comes to data collection, we need to find a dataset which is appropriate for testing, is clearly labeled and has enough features. A single dataset [21] was acquired from Kaggle, containing daily air pollutant and air quality values of 26 cities in India from 2015 to 2020. The dataset contains more than 29000 instances and 16 features. The features are namely City, Date, NO, NO2, CO, SO2 etc. 13 features carry float values and the rest of the 3 are string values. The data has been collected from 24 cities in India.

#	Column	Non-Null Count	Dtype
0	City	29531 non-null	object
1	Date	29531 non-null	object
2	PM2.5	24933 non-null	float64
3	PM10	18391 non-null	float64
4	NO	25949 non-null	float64
5	NO2	25946 non-null	float64
6	NOx	25346 non-null	float64
7	NH3	19203 non-null	float64
8	CO	27472 non-null	float64
9	SO2	25677 non-null	float64
10	O3	25509 non-null	float64
11	Benzene	23908 non-null	float64
12	Toluene	21490 non-null	float64
13	Xylene	11422 non-null	float64
14	AQI	24850 non-null	float64
15	AQI_Bucket	24850 non-null	object

Figure 5.1: Dataset description

5.2 Data Preprocessing

The dataset was saved in Microsoft Excel Comma Separated Values File (.csv) format. From our dataset, we selected Delhi as our target city as it had fewer null values than the other cities. The data was in a daily format, meaning there were 30 iterations for every month. We averaged and converted the data from a daily-basis to a monthly-basis. By averaging, we did not need to omit the null rows. This helped us in giving better prediction results. The data from 2020 was incomplete. So, we ignored those and used the data from 2015 to 2019.
We tested our dataset using ADCF (Augmented Dickey–Fuller) test. We omitted the outliers from the datasets. After calculating the monthly average, there were few cells with null values. We used dropna() function to ignore those null values. A graphical representation is given below showing the change by ignoring null values.

[Before] 

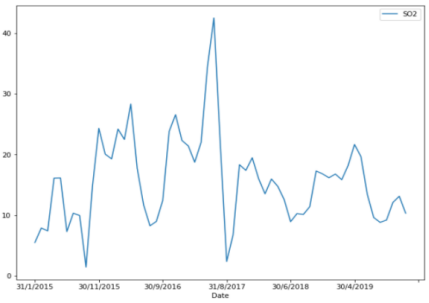
[After] 

Figure 5.2: Before and after ignoring null value

5.3 Data Visualization

As we are using such a large dataset, we tested it to find stationary and non-stationary data. We had to make our data stationary so that our model performed better. By observing the plot of the graphs, we can assume the trend and can detect if our prediction is in the right track.

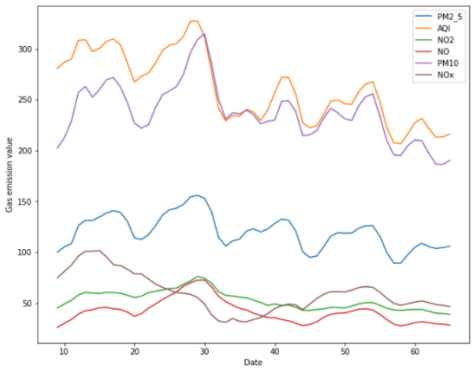


Figure 5.3: Graphical representation features values

Visualizing data helps us to see the trends in prediction. We can also be aware if we need to make further changes to our dataset.

5.4 Data Analysis

We divided the dataset into test and train in the ratio, 15:85. We did hyper-parameter tuning for optimizing our model to get better results.
For predicting with Prophet, we selected the frequency in monthly basis and set the interval width to 95%. We set seasonality='True' as emission rate of a gas increases or decreases seasonally. The accuracy of Prophet depends on the percentage of MAPE[19]. We got some MAPE values below 20% and some above 20%. So, the MAPE below 20% can be said to have a good result and the ones above need improvement. After prediction, we compared the forecast value with the training and testing dataset which gave good accuracy for most of the features. Then we generated a graph to visually compare the difference between the actual and predicted values.

Figure 5.4: Comparison of actual(y) and predicted(yhat) value using Facebook Prophet

For linear regression, we first labelled the feature we wanted to predict. Then we copied all the features into a new dataset except the labelled feature. We divided the new dataset into training and testing sets. Then we ran linear regression on the training set. Lastly, we fed the test set in order to find the predicted data. The predicted values were then stored in an array. The highest accuracy was 97% and the lowest was 5%. We plotted the prediction results in a graph as shown below.

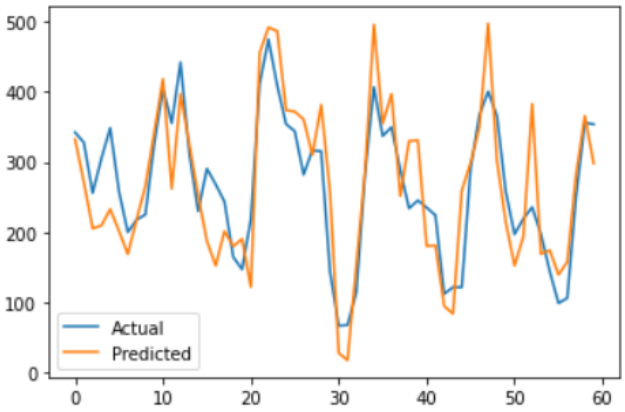


Figure 5.5: Comparison of actual and predicted value using Linear regression

For RNN model, we first divide our dataset into training and testing sets. There are 60 data points in our dataset. We used 32 data points for our training set and 28 data points for our testing set. Then we have to make an object that contains batches. We have to convert the series into a Numpy array, defining windows and the number of inputs and outputs. Subsequently, we have to write a function using TensorFlow to construct batches. For the batches, we take value that is 1 less than the time period. Therefore, the output we get must have three parts which are, quantity of batches, the size of the windows and the number of inputs. Then we need to divide the dataset into five equal batches in order to form objects with batches. To make a RNN model, we need RNN, variables that have tensors, losses together with optimization. To reduce RMSE, we use optimization for a continual variable. After building the architecture for RNN model, we have to train and test our model in order to predict data.

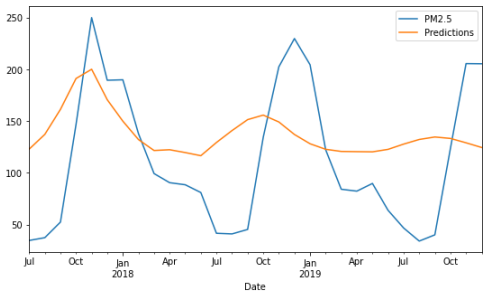


Figure 5.6: Test plot for PM2.5 using RNN

For ARIMA model, we check that if our data has any kind of trend and seasonality in it. If we find any kind of trend and seasonality then we separate that which allows a more accurate prediction of our values. Moreover, we split our total 60 data point into two parts train and test where train contains first 50 data point and test contain last 10 data point. Here the train is helping the model to understand. Moreover, we then find the p, d, and q value by using combination here we find the RMSE(Root mean square deviation) value for each p, d, and q value of total 128 combinations and sorted the lowest RMSE value and use it as the order to building and training our model. After training is complete, we predict our value and evaluate the model on the test data point. So, finally, our model successfully plots the original value with the new value that we found.

Figure 5.7: Graph of PM2.5 using ARIMA

Chapter 6

Result Analysis

Accuracy, MAE (mean absolute error) and MAPE (mean absolute percentage error) values are shown in the tables below.

6.1 Facebook Prophet

For Facebook Prophet, we have predicted all the features by splitting our dataset to training and testing dataset to 85% and 15% accordingly. It is a generalized additive model it has three main parts which are seasonality, trend, and holidays. We have kept the seasonality true as environment pollution varies in different time of the year. For tuning purpose we have nullified the outliers and kept the changing point to default value which is 0.05 to ensure that our data doesn't overfit or underfit. We have added holidays so that it can know the significant changes on that days. Frequency was set to monthly and historical forecast was set to true so that it will use only the last point of each historical forecast to compute error scores. The accuracy of Prophet model depends on the value of the MAPE[19].The formula of MAPE and MAE are given below,

$$\begin{aligned} \text{MAPE} &= \frac{1}{n}\sum_{i=0}^{n}\left|\frac{y^i \text{ forecast} - y^i \text{ true}}{y^i \text{ true}}\right| \cdot 100\% \\ \text{MAE} &= \frac{1}{n}\sum_{i=1}^{n}|x_i - x| \end{aligned} \tag{6.1}$$

Where, n = the number of errors, $\sum$= summation symbol, $|x_i - x|$ = the absolute errors.

For most of the cases the value was significantly good. But few of the cases it didn't perform well as the values of that feature was non-stationary and Prophet does not perform well on non-stationary data as it is difficult to find the actual trend and seasonality of inconsistent data patterns.

For monthly prediction Prophet outruns many other time-series models as it has built-in hyperparameters which allows to adjust seasonality. For non-stationary data it cannot comeup with a good solution though you try to change the data to stationary as much as possible with scaling and other methods but when it gets a stationary dataset we witnessed robust performance. Though seasonality and trend are difficult to quantify, FB Prophet does a great job capturing both from other

models. The ability to include holidays does significant effect while predicting the values.

Table 6.1: Error values of Facebook Prophet

Model	Facebook Prophet			
Features	Training data		Testing data	
	MAE	MAPE(%)	MAE	MAPE(%)
PM2.5	9.37	8.5	10.691	15.8
PM10	30.505	15.2	37.99	28
NO	5.7	19.5	12.04	45.8
NO2	8.33	16.7	6.139	14.1
NOx	16.7	23.4	23.45	38.75
CO	0.297	25	0.22	18
SO2	2.32	15.9	4.64	34.3
O3	0.074	4.5	0.094	5.9
NH3	3.64	9	3.76	10.3
Benzene	0.63	23.4	0.75	25.1
Toluene	7.887	52.3	4.181	15.9

6.2 Linear Regression

We have arranged the accuracy of prediction using linear regression in the table below.
In order to determine the accuracy in linear regression, we need coefficient of determination. The higher the coefficient of determination, the better the performance. The formula to find the coefficient of determination is given below,

$$R^2 = 1 - (\text{RSS}/\text{TSS}) \tag{6.2}$$

Where, the Coefficient of Determination is denoted by R^2, Residuals sum of squares is denoted by RSS and the Total sum of squares is denoted by TSS.
For linear regression, 85% data was selected for training and 15% for testing. From the table, we can see that PM2.5 has the highest training accuracy of 92.44% while NOx has the lowest testing accuracy of 4.95%. The 4 gases, NOx, CO, SO2 and O3 show a much lower accuracy compared to the others. This is because there was a huge number of null values in the dataset for these 4 gases. The other gases had comparatively lower number of null values. So they produced a better result with higher accuracy.
Another reason for the lower accuracy is that the values were not constant. The values changed considerably with very high variation. The values for these 4 gases showed high variation. However, the values of the other gases showed much less variation. This is one of the reasons the accuracy was higher.
The accuracy values of linear regression are shown in the table below:

Table 6.2: Accuracy values of Linear Regression

Model	Linear Regression	
Features	Training accuracy(%)	Testing accuracy(%)
PM2.5	92.44	85.52
PM10	90.27	77.48
NO	91.71	79.18
NO2	91.62	70
NOx	77.21	4.59
CO	75.19	18.76
SO2	60.72	32.78
O3	60.78	29.01
NH3	72.29	42.23
Benzene	88.11	61.99
Toluene	60.7	65.7

6.3 RNN (LSTM)

For Time Series Forecasting with RNN using LSTM we have predicted all the features by splitting our dataset into test and train. And to get better result we have divided our dataset into training and testing sets, 53% data was selected for training and 47% test. Unlike any other neural networking models, RNN uses its memory to run lengthy data. In the other neural networks, the inputs do not need to be related to each other. But in RNN model, we need all the inputs that are related to each other. Some features of the dataset we are working with has some disconnected inputs. Basically, the inputs are not related to each other for example we can look at training and testing accuracy of PM2.5. So, after splitting our dataset into test and train, we have scaled our dataset using MinMaxScaler to get a standard value. We have performed this to prevent features with wider range.
After finding error values we calculated the accuracy for RNN model. As we can see for most of the cases the value was not that good. But few of the cases it did perform well as the input values of that feature were related somehow and the data that are not related does not perform well for this model as it is difficult to get accuracy.

Table 6.3: Accuracy values of RNN (LSTM)

Model	RNN (LSTM)	
Features	Training accuracy(%)	Testing accuracy(%)
PM2.5	31	41
PM10	2	16
NO	72	82
NO2	76	86
NOx	66	77
CO	83	98
SO2	90	97
O3	75	89
NH3	85	91
Benzene	97	98
Toluene	64	81

RNN can only be used on related data which are dependent on each other [15]. Data with huge range can cause problem while finding accuracy. Moreover, it is quite hard to train RNN as it cannot process lengthy data. So that is why for some features this model has a large error value and is unable to predict accuracy. From the above description we can say that we are unable to use this dataset for this algorithm efficiently, as for some features it is unable to predict accuracy. Our dataset is huge and the values are not dependent on each other. Hence, the accuracy values are quite low.

6.4 ARIMA

For our time series analysis with ARIMA model, we predicted our values by splitting our dataset to training and testing dataset to 85% and 15% accordingly. Unlike other models, ARIMA models are used where the data is non-stationary. We can make the data stationary by using an initial differencing step [13]. We separate the trend and seasonality from the features until we where left with random errors. We calculate the hyperparameters of ARIMA model. After that we build our model and predict our values. For most of the cases the value was significantly good. But few of the cases it didn't perform well due to small dataset.

Table 6.4: Accuracy values of Arima model

Model	ARIMA	
Features	Training accuracy(%)	Testing accuracy(%)
PM2.5	57.92	82.65
PM10	40.53	70.27
NO	86.86	91.39
NO2	87.74	85.56
NOx	75.44	77.29
CO	97.45	98.98
SO2	92.26	94.93
O3	85.39	85.84
NH3	87.74	94.93
Benzene	98.92	99.25
Toluene	93.16	95.12

For time series analysis ARIMA model is used for more insight on the data and for prediction. ARIMA model has built-in hyperparameters which allows to adjust seasonality. For non-stationary data it come up with a good solution to we change the data to stationary. However, ARIMA is best fitted with large dataset but our dataset is a mid-range dataset for that our prediction accuracy sometime goes downwards. Anyhow, ARIMA perform great due to the ability to separate the seasonality and trend does great effect while predicting the values.

Chapter 7

Performance Evaluation

Figure 7.1: Graphical representation of testing accuracy

From the above diagrams, we can see that the results of ARIMA model are consistent and better than the rest. So, we can say ARIMA model has the best performance and is best suited for our dataset.
In ARIMA model, we split our data in such a way which can give us the best result possible. ARIMA is used for a much clearer understanding of the data and for prediction of future values. Furthermore, we separate seasonality and trend by converting the non-stationarity data to stationarity, which help us acquire more accurate prediction. Where Facebook Prophet in some cases didn't perform well as the values of that feature was non-stationary and Prophet does not perform well on non-stationary data as it is difficult to find the actual trend and seasonality. Again, in Linear Regression model the reason for the lower accuracy is that the values were non- stationary. The values changed considerably with very high variation. In RNN model we can see the same issue where RNN also give higher RMSE value due to

the fact that the dataset has seasonality in them. Anyhow, in ARIMA model we then find the p, d, and q value by using combination here we find the RMSE(Root mean square deviation) value for each p, d, and q value of total 128 combinations and sorted the lowest RMSE value by that we ensure that we can get the highest accuracy possible. For this reason. ARIMA performance is comparatively better than the others.

Chapter 8

Future Work and Conclusion

8.1 Future Work

In future we will predict air pollution level of different region of a country. Then we will compare those prediction and will create a visual mapping of threat level (such as mild, medium, high) over the country. Using models such as, logistic regression and Decision Tree we will try to predict the area-wise health threats which people may face. Lastly we will try to improve our prediction models by including new time-series models and by doing necessary tuning to the model.

8.2 Conclusion

Air Pollution is the biggest problem faced by the world today. A problem that needs to be solved at any cost as our human life depends on it. We humans depend on the air around us to live, without it we would die. As we have seen previously, there is considerable unreliability in estimating both detection, effects, and the data we can find. So many attempts were made to detect the sources, effects, and causes of air pollution. We are accountable for damaging our environment. Regulating all possible scenarios is not viable. So, by reducing exposure can prevent the possible health issues and also protection of the population is best achieved. Still, though we will try our best to provide a good model to utilize air pollution that can help us with possible outcomes. Applying machine learning algorithms on a detailed and organized datasets we can certainly generate enough information to estimate the air pollution level and related cause of a country. With the help of previous research papers and available datasets, we are certain to achieve our goal. Finally, we believe that this paper will contribute to the environmental science and society. Also, this model further makes us aware of the needs and challenges in the future.

Bibliography

[1] C. A. Pope Iii, R. T. Burnett, M. J. Thun, E. E. Calle, D. Krewski, K. Ito, and G. D. Thurston, "Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution," *Jama*, vol. 287, no. 9, pp. 1132–1141, 2002.

[2] S. Dasgupta *et al.*, *Indoor air quality for poor families: new evidence from Bangladesh*. World Bank Publications, 2004, vol. 3393.

[3] Z. J. Andersen, M. Hvidberg, S. S. Jensen, M. Ketzel, S. Loft, M. Sørensen, A. Tjønneland, K. Overvad, and O. Raaschou-Nielsen, "Chronic obstructive pulmonary disease and long-term exposure to traffic-related air pollution: A cohort study," *American journal of respiratory and critical care medicine*, vol. 183, no. 4, pp. 455–461, 2011.

[4] B. Yeganeh, M. S. P. Motlagh, Y. Rashidi, and H. Kamalan, "Prediction of co concentrations based on a hybrid partial least square and support vector machine model," *Atmospheric Environment*, vol. 55, pp. 357–365, 2012.

[5] A. T. Azar, H. I. Elshazly, A. E. Hassanien, and A. M. Elkorany, "A random forest classifier for lymph diseases," *Computer methods and programs in biomedicine*, vol. 113, no. 2, pp. 465–473, 2014.

[6] A. Azid, H. Juahir, M. E. Toriman, M. K. A. Kamarudin, A. S. M. Saudi, C. N. C. Hasnam, N. A. A. Aziz, F. Azaman, M. T. Latif, S. F. M. Zainuddin, *et al.*, "Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in malaysia," *Water, Air, & Soil Pollution*, vol. 225, no. 8, pp. 1–14, 2014.

[7] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[8] M. Shakerkhatibi, I. Dianat, M. A. Jafarabadi, R. Azak, and A. Kousha, "Air pollution and hospital admissions for cardiorespiratory diseases in iran: Artificial neural network versus conditional logistic regression," *International journal of environmental science and technology*, vol. 12, no. 11, pp. 3433–3442, 2015.

[9] K. Siwek and S. Osowski, "Data mining methods for prediction of air pollution," *International Journal of Applied Mathematics and Computer Science*, vol. 26, no. 2, pp. 467–478, 2016.

[10] S. Saad, A. Shakaff, M. Hussein, M. Mohamad, M. Dzahir, and Z. Ahmad, "Analysis of feature selection with k-nearest neighbour (knn) to classify indoor air pollutants," *MALAYSIAN JOURNAL OF INDUSTRIAL TECHNOLOGY (MJIT)*, 2017.

[11] C. Aditya, C. R. Deshmukh, D. Nayana, and P. G. Vidyavastu, "Detection and prediction of air pollution using machine learning models," in *International Journal of Engineering Trends and Technology (IJETT)*, vol. 59, 2018, pp. 204–207.

[12] A. Hussain, "Experts: Bangladesh's forest coverage under 10%— dhaka tribune," *Dhaka Tribune*, 2018.

[13] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. OTexts, 2018.

[14] D. Kumar *et al.*, "Evolving differential evolution method with random forest for prediction of air pollution," *Procedia computer science*, vol. 132, pp. 824–833, 2018.

[15] N. K. Manaswi, "Rnn and lstm," in *Deep Learning with Applications Using Python*, Springer, 2018, pp. 115–126.

[16] X. Zhao, R. Zhang, J.-L. Wu, and P.-C. Chang, "A deep recurrent neural network for air quality classification.," *J. Inf. Hiding Multim. Signal Process.*, vol. 9, no. 2, pp. 346–354, 2018.

[17] M. Castelli, F. M. Clemente, A. Popovič, S. Silva, and L. Vanneschi, "A machine learning approach to predict air quality in california," *Complexity*, vol. 2020, 2020.

[18] B. Vishwas and A. Patel, "Prophet," in *Hands-on Time Series Analysis with Python*, Springer, 2020.

[19] E. Zunic, K. Korjenic, K. Hodzic, and D. Donko, "Application of facebook's prophet algorithm for successful sales forecasting based on real-world data," *arXiv preprint arXiv:2005.07575*, 2020.

[20] A. Heydari, M. M. Nezhad, D. A. Garcia, F. Keynia, and L. De Santoli, "Air pollution forecasting application based on deep learning model and optimization algorithm," *Clean Technologies and Environmental Policy*, pp. 1–15, 2021.

[21] K. Tripathi and P. Pathak, "Deep learning techniques for air pollution," in *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, IEEE, 2021, pp. 1013–1020.