https://www.overleaf.com/project/5e9c381eb912040001384284

A Project Report

On

# Development of Fractal Nature,Chaos and Self Similarity of Air Quality Data

Submitted in partial fulfilment of the requirement of

**University of Mumbai**

For the Degree of

**Bachelor of Engineering**

*in*

**COMPUTER ENGINEERING**

*Submitted by*

**Pragya Verma**
**Sai Reddy**
**Mithilesh Waghulade**

*Supervised by*

**Dr.Lata Ragha**

*&*

**Dr.Debrabata Datta(BARC)**


**Department of Computer Engineering**

**Fr. Conceicao Rodrigues Institute of Technology**
**Sector 9A, Vashi, Navi Mumbai - 400703**

**UNIVERSITY OF MUMBAI**

**2019-2020**

# APPROVAL SHEET

This is to certify that the project entitled

**" Development of Fractal Nature,Chaos and Self Similarity of Air Quality Data"**

**Submitted by**


| | |
|---|---|
| **Pragya Verma** | **101661** |
| **Sai Reddy** | **101647** |
| **Mithilesh Waghulade** | **101663** |


**Supervisors :** _____

**Project Coordinator :** _____

**Examiners : 1.** _____

**2.** _____

**Head of Department :** _____

**Date :**

**Place :**

# Declaration

We declare that this written submission for B.E. Declaration entitled "**Development of Fractal Nature,Chaos and Self Similarity of Air Quality Data**" represent our ideas in our own words and where others' ideas or words have been included. We have adequately cited and referenced the original sources. We also declared that we have adhere to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any ideas / data / fact / source in our submission. We understand that any violation of the above will cause for disciplinary action by institute and also evoke penal action from the sources which have thus not been properly cited or from whom paper permission have not been taken when needed.

**Project Group Members:**

1. Pragya Verma, 101661

_____

2. Sai Reddy, 101647

_____

3. Mithilesh Waghulade , 101663

_____

# Abstract

The air quality index (AQI) is an index for reporting the air quality. It tells how clean or polluted the air is and what associated health effects it has on people breathing it. Considering the four major pollutants viz, Nitrogen dioxide(NO2), Sulphur Dioxide(SO2), Suspended particulate Matter(SPM), and Residual Suspended particulate Matter(RSPM), this project concentrates on calculating AQI for air in Delhi to answer the related questions canvassing air quality. A plethora of statistical tests and analyses have been utilized on the relevant time-series data to produce the desired outcome. Further, this proposal encompasses an ARIMA forecasting model whose primary challenge is to accurately predict the daily concentration of pollutants of the upcoming days. The Chaos and bifurcation theory will be used to model this nonlinear dynamic system which is difficult to control and predict. This theory will help us understand whether the data is chaotic in nature or not. Then fractal and self-similarity analysis, a type of statistical analysis, for this time-series data will help to determine a particular pattern or hidden behavior in the data. Finally, an LSTM machine will be developed which would help in the prediction of future AQI.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Background

Air is humans' only source of oxygen for survival. But with the passage of time, the fresh and pure air is gradually getting contaminated due to the increase in air pollution. So, to determine whether the air we breathe is pure or not the government has taken an initiative to calculate AQI of air. This information is made available to general citizens and policymakers to make decisions and to prevent and minimise air pollution exposure and the ailments induced by the exposure.

## 1.2 Motivation

Apart from land and water, air is the prime resource for the sustenance of life. Air is an integral and essential necessity in everyday life. Whether it is agriculture, or pollination of various crops, or even basic survival of numerous living species, everything everywhere is dependent on air. The importance of air cannot be overemphasized and hence the rising levels of air pollution are a matter of serious concern.

Air Pollution is the inadequate change in physical, chemical or biological characteristics of air which hampers life as well as leads to potential health problems. Air pollution majorly affects the eyes, lungs, nose, and throat by causing irritation. It also creates respiratory problems and exacerbates existing conditions such as asthma and emphysema. The risk of cardiovascular diseases becomes much higher when humans are continually exposed to air pollution. In India, air pollution is the third highest cause of death among health risks and because of this, life expectancy has gone down by 2.6 years. Hence, it has become increasingly necessary to not only control the contamination but also enlighten the people affecting the quality of air effectively in a bid to maintain a healthy standard.

A reasonable way to analyze the amount of pollution is by determining the standard of air. With technological advancements, a vast amount of data on ambient air quality is generated which is used to establish the quality of air in different areas. The large monitoring data result has astronomical volumes of information that neither provides any useful insights to a decision-maker nor is intelligible to a common man who simply wants to understand how good or bad the air is. One way to describe air quality is to report the concentrations of all pollutants with acceptable levels(standards). As for the general public, they generally will not be satisfied with raw data, time series plots, statistical analysis, and other complex findings pertaining to air quality and hence people tend to lose

interest. They can neither appreciate the state of air quality nor the pollution alleviation efforts by regulatory agencies. Since awareness of daily levels of urban air pollution is important to those who suffer from illness caused by exposure to air pollution, the issue of air quality communication should be addressed in an effective manner. Further, the success of a nation to improve the air quality depends on the support of its citizens who are well informed about local and national air pollution problems and about the progress of mitigation efforts.

To address the aforementioned concerns, many developed countries over the past three decades have devised and utilized effectively, the concept of the Air Quality Index (AQI). An AQI is defined as a strategy that involves the transformation of weighted values of individual air pollution parameters (SO2, NO2, visibility, etc.) into a single number or set of numbers. AQI mostly considers eight pollutants PM10, PM2.5, NO2, SO2, CO, O3, NH3, and Pb for its determination.

The challenge of communicating with the people in a comprehensible manner has two dimensions: (i) Translate the complex scientific and medical information into simple and precise knowledge and (ii) communicate with citizens in the historical, current and futuristic sense. Addressing these challenges and thus developing an efficient and comprehensible AQI scale is required for citizens and policymakers to make decisions and to prevent and minimize air pollution exposure and the ailments induced by the exposure.

## 1.3 Aim and Objective

The main objective of the project is to calculate the Air Quality Index (AQI) of air of Delhi. Prior to calculating AQI, the dataset is analyzed using numerous data analytics algorithms and visualizing it with statistical concepts. Finally, a forecasting model will be developed which will help us gain an insight into the quality of air of immediate upcoming days.

## 1.4 Report Outline

The report elucidates an approach of tackling a grave social problem which is the rapid decline of the quality of air in the atmosphere using a universally used parameter known as the Air Quality Index(AQI). It provides a detailed description on the computation methodology of AQI and further explains the development of a forecasting model. This model provides predictions of the AQI and the concentrations of different pollutants involved.

The testing results are also included in the report which are important for the project as it quantifies the actual accuracy of the proposed system. This report aims to provide valuable insights in an effective manner to both the policymakers and the general citizens to encourage them to play their part to overcome this social challenge.

# Chapter 2

# Study Of the System

## 2.1 About the Technique

### 2.1.1 Handling missing data

Many real-world datasets may contain missing values for various reasons. They are often encoded as NaNs, blanks or any other placeholders. Training a model with a dataset that has a lot of missing values can drastically impact the machine learning model's quality. There are three main types of missing data:

- Missing at Random (MAR): Missing at random means that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data
- Missing Completely at Random (MCAR): The fact that a certain value is missing has nothing to do with its hypothetical value and with the values of other variables.
- Missing not at Random (MNAR): Two possible reasons are that the missing value depends on the hypothetical value (e.g. People with high salaries generally do not want to reveal their incomes in surveys) or missing value is dependent on some other variable's value (e.g. Let's assume that females generally don't want to reveal their ages! Here the missing value in age variable is impacted by gender variable). For our dataset specific steps to clean the dataset:
    1. Retrieve missing dates and filling those dates with the monthly mean.
    2. Combining multiple reading of a particular date with its mean.
    3. Then in the end filling missing values with monthly mean.
    4. Then even if there were any missing values they were imputed by yearly mean.

### 2.1.2 Shannon's Entropy

Shannon's Entropy is the quantification of the amount of uncertainty in the entire probability distribution[1]. In simpler words, this quantitative measure is used to determine the amount of information that any variable in the data withholds. The mathematical formula is as follows: H(x) = $\sum P(x) log \frac{1}{P(x)}$ where P(x) is the probability of occurrence event x. $\frac{1}{P(X)}$ represents the amount of information associated with event x The entropy is used to display the randomness or the likeliness of an event to transpire[2].

Hence, the greater value of entropy signifies that the occurrence of that particular event is less likely. To address this concept, Shannon introduced the reciprocal (ie; $\frac{1}{P(X)}$) in the amount of information in his computational formula.

### 2.1.3 Data Decomposition

The initial steps involved in a time series forecasting project is to visualise the data and then to decompose the data into trends and cyclic components for better insights. A time series data has a timestamp associated with it whether it is hourly, daily, monthly, yearly and so on. The first step involves the smoothening of the data in a temporal form where the rolling average method is the most widely used technique. This helps in removing the outliers whilst preserving the underlying trends and patterns existing in the dataset. Then, the classical seasonal decomposition technique is used to "detrend" the data. There are various models to serve this purpose. Firstly, the additive model where a static seasonal value that changes with each timestamp but repeats its value after a cycle is added to the original data value to create the model. This model is then compared with the actual data plot and the goodness of fit is determined. If the fitness of the model is erratic, the multiplicative model is applied. This is similar to its additive counterpart with the only difference being the static seasonal parameter is multiplied to the original data value ie; the multiplicative model scales the size of the seasonal cycle as the trend rises or falls. The best fit model is then selected on the basis of which, the data is decomposed to reveal the underlying trends and/or patterns present in it.[3]

### 2.1.4 QQ Plotting

A quantile-quantile plot (also known as a QQ-plot) is a method through which one can determine whether a dataset matches a specified probability distribution. Graphically, the horizontal and vertical axes of a QQ-plot are used to show quantiles. Generally, quartiles divide a dataset into four equal parts but this is only a preferred practice and one can perform this division in any number of parts in accordance to their predilection. With a QQ-plot, the quantiles of the sample data are on the vertical axis, and the quantiles of a specified probability distribution are on the horizontal axis. The plot consists of a series of points that show the relationship between the actual data and the specified probability distribution. If the elements of a dataset perfectly match the specified probability distribution, the points on the graph will form a 45 degree line. If the line plotted by the dataset

show any discrepancy with respect to the distribution, it indicates that the dataset does not conform with the specified probability distribution.[4]

### 2.1.5 Hurst exponent

According to the original proposition, Hurst exponent (H) = 0.5 would represent a self-determining process, in which the current value of the series would not be dependent on past values of the series. When the value of H lies between the range 0 <H <0.5, the series becomes anti-persistent. Anti-persistent series display 'mean-reverting' characteristics. If a value in the time series was high in the previous period, it is likely to reduce in the following period towards the mean value. The strength of the mean reverting behavior increases as Hurst exponent approaches to zero. When the range of H exponent varies between 0.5 <H <1, the values of the series rise and fall in upward and downward direction in a broader range than likely by pure random walk. Such series display seeming trends for some time, but these seeming trends are erratically interrupted by abrupt discontinuities. The power of the trend-reinforcing behavior increases as the value of the Hurst exponent increases to the upper ceiling value of one. When the observations from the series are independent of previous observations, the value of H is expected to be around 0.5.[5]

### 2.1.6 Fractal Dimension

Most of the objects found in nature possess irregular shapes that can not be quantified with the help of standard Euclidean geometry. In many cases these objects have a peculiar character of self similarity where the part of the object looks like the whole. Such objects are known as fractals and the associated degree of complexity of shape, structure and texture is quantified in terms of Fractal dimension. A similar parallel can be drawn for a time-series data. A time-series data may or may not display any patterns which can be determined using the fractal dimension. If a small part of the data is considered (similar to the part of the object in Euclidean geometry) and if this portion is showcases all the characteristics present in the entire dataset then it can labelled as a fractal for the dataset and the length of this portion, number and combination of dimensions and other relevant characteristic would be quantified to be termed as the Fractal dimension for the dataset. The fractal dimension (D) can be calculated using the Hurst Exponent (H) by the following formula:

$$D = 2 - H$$

### 2.1.7 Power Law

is a functional relationship between two quantities, where a relative change in one quantity results in a proportional relative change in the other quantity, independent of the initial size of those quantitiesPower law [6]. The color of noise refers to the power spectrum of a noise signal and different colors of noise have significantly different properties. Many of the color definitions assume a signal with components at all frequencies, with a power spectral density per unit of bandwidth proportional to $\frac{1}{f}$ or $\beta$ and hence they are examples of power-law noise. For instance, the spectral density of white noise is flat ($\beta = 0$), while flicker or pink noise has $\beta = 1$, and Brownian noise has $\beta = 2$. For a time series data however, this color of the noise or the power-law noise can be calculated using the Hurst Exponent by the following formula:

$$\beta = 2H + 1$$

where, if $\beta = 0$,white noise, uncorrelated and power spectrum independent of frequency.
if $\beta = 1$, flicker of 1/f noise systems, moderately correlated.
if $\beta = 2$, brownian noise like systems, strongly correlated.

### 2.1.8 Lyapunov Exponent

The Lyapunov Exponents of a system are a set of invariant geometric measures that describe the dynamical content of the system. In particular, they serve as a measure of how easy it is to perform prediction on the system under consideration. Lyapunov Exponents quantify the average rate of convergence or divergence of nearby trajectories in a global sense. A positive exponent implies divergence and a negative one implies convergence. Consequently, a system with positive exponents has positive entropy in that trajectories that are initially close together move apart over time. The more positive the exponent, the faster they move apart. Similarly, for negative exponents, the trajectories move together. A system with both positive and negative Lyapunov Exponents is said to be chaotic.[7]

### 2.1.9 Chaos and Bifurcation

Chaos theory is a branch of mathematics focusing on the study of chaos - states of dynamical systems whose apparently-random states of disorder and irregularities are often governed by deterministic laws that are highly sensitive to initia conditions[8] The analysis of chaotic systems is done

with the help of bifurcation diagrams and Lyapunov exponents . The qualitative changes in dynamics of the system are evaluated with the help of bifurcation diagrams. They provide models of transitions and instabilities as some control parameter is varied. The Bifurcation Diagram (BD) of a given dynamical system gives the idea of the behaviour of one of the outputs of that system with different values of one of the control input parameters keeping all the other input parameters constant. A bifurcation diagram is a plot that shows the value of the changing parameter, on one axis and the solution to the system on the other axis. In other words change in the qualitative character of a solution as a control parameter is varied is known as a bifurcation. This occurs where a linear stability analysis yields an instability which is characterized by a perturbation.[9]

### 2.1.10 Self Similarity

Fractal structures are said to be self–similar, when part of an object looks like the whole object under appropriate scaling i.e. the structure looks like a reduced copy of the full set on a different scale of magnification. However this scaling can not be indefinitely extended, after a certain stage the smaller pieces may not perfectly represent the original shape, this is the characteristic of natural fractals. In general this is termed as self-similarity or statistical self–similarity. Thus natural fractals exhibit self-similarity over a limited range and naturally occurring fractals usually exhibit statistical self similarity. Whereas, mathematical fractals exhibit self similarity at all length scales and thus are strictly self-similar. In other words, statistical fractals are self-affine, or statistically self-similar; they are composed of statistically equivalent replicas of the whole object.

### 2.1.11 Air Quality index calculation

An air quality index is defined as an overall scheme that transforms the weighted values of individual air pollution related parameters (for example, pollutant concentrations) into a single number or set of numbers. [10]The result is a set of rules (i.e. most set of equations) that translates parameter values into a more simple form by means of numerical manipulation.

Although various mathematical formulas have been devised for the calculation of the AQI, the Fenstock Air Quality Index is most suitable for the computations involved in this project for multiple reasons viz, (i) The formula is simple and effective with no complicated mathematical equations involved. (ii) All the data required for computation is easily available. (iii) It is generally used for estimation of the overall air pollution potential for

a metropolitan area (in this case, Delhi).[11]

The Fenstock Air Quality Index is computed as follows:

$$AQI = Wi \times Ii$$

where, Wi = weightages for the pollutants,
Ii= estimated sub-indices for pollutants

### 2.1.12 Autoregressive Integrated Moving Average

The Autoregressive Integrated Moving Average (ARIMA) method models the next step in the sequence as a linear function of the differenced observations and residual errors at prior time steps. It combines both Autoregression (AR) and Moving Average (MA) models as well as a differencing pre-processing step of the sequence to make the sequence stationary, called integration (I). The notation for the model involves specifying the order for the AR(p), I(d), and MA(q) models as parameters to an ARIMA function, e.g. ARIMA(p, d, q). An ARIMA model can also be used to develop AR, MA, and ARMA models and hence is highly effective.The method is suitable for univariate time series with trend and without seasonal components.[12]

### 2.1.13 Stationarity

A stationary time series is the one for which the properties (namely mean, variance and covariance) do not depend on time.[13] It is necessary to determine stationarity and convert the data (if non-stationary) into a stationary dataset as most statistical models require the series to be stationary to make effective and precise predictions. The different methods of making a dataset stationary as explained below help us to further determine which statistical test to apply to the given non-stationary dataset. (i) Differencing: difference of consecutive terms in the series is computed. (ii) Seasonal Differencing: Calculation of the difference between an observation and a previous observation from the same season. (iii) Transformation: used to stabilize the non-constant variance of a series. Among the various statistical methods available, the Augmented Dicky Fuller (ADF) test was found to be the most suited to convert the dataset under consideration into a stationary dataset. The Augmented Dickey Fuller test is one of the most popular statistical tests. It can be used to determine the presence of unit root in the series, and hence help us understand if the series is stationary or not. The test is explained below: Null Hypothesis: The series has a unit root (value of a = 1) Alternate Hypothesis: The series has no unit

root. If we fail to reject the null hypothesis, we can say that the series is non-stationary. Test for stationarity: If the test statistic is less than the critical value, we can reject the null hypothesis (aka the series is stationary). When the test statistic is greater than the critical value, we fail to reject the null hypothesis (which means the series is not stationary).

### 2.1.14 Recurrent Neural Network

-

A recurrent neural network (RNN) is a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence. This allows it to exhibit temporal dynamic behavior. Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture. It can not only process single data points (such as images), but also entire sequences of data (such as speech or video). LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series. LSTMs were developed to deal with the vanishing gradient problem that can be encountered when training traditional RNNs.[14] A common architecture is composed of a cell (the memory part of the LSTM unit) and three "regulators", usually called gates, of the flow of information inside the LSTM unit: an input gate, an output gate and a forget gate. Intuitively, the cell is responsible for keeping track of the dependencies between the elements in the input sequence. The input gate controls the extent to which a new value flows into the cell, the forget gate controls the extent to which a value remains in the cell and the output gate controls the extent to which the value in the cell is used to compute the output activation of the LSTM unit. The activation function of the LSTM gates is often the logistic sigmoid function. There are connections into and out of the LSTM gates, a few of which are recurrent. The weights of these connections, which need to be learned during training, determine how the gates operate.

## 2.2 Various Available Technique

### 2.2.1 Filling missing values

Occurrences of missing values in a dataset are a common feature. There are numerous ways in which these values can be tackled, a few of which are cited as follows:

#### 2.2.1.1 Deletion

In this, the rows with the missing values are completely discarded from the dataset and only the existing values are used in further analysis. This method whilst easy to implement generally leads to a huge loss of information especially if the missing values are scattered across the dataset. The usage of this technique is advisable only when the proportion of missing values negligible as compared to the number of data points. There are various techniques for deletion viz, listwise deletion (entire row with one or more missing value is discarded), dropping variables (a variable with a high percentage of missing values is rendered useless in analysis).[15]

#### 2.2.1.2 Imputation

Instead of discarding the missing values and thereby preventing the loss of information, this method aims to fill the missing data by using some mathematical formulas or computational algorithms.[16] A few of which are elaborated on as follows:

1. Mean, Median, Mode: As the name suggests, this method involves the calculation of one of the three statistical parameters for a variable using the data available and replaces all the missing values in that variable with the value of this statistical parameter.

2. K-Nearest Neighbors: In this method, k neighbors are chosen based on some distance measure and their average is used as an imputation estimate. The method requires the selection of the number of nearest neighbors, and a distance metric. KNN can predict both discrete attributes (the most frequent value among the k nearest neighbors) and continuous attributes (the mean among the k nearest neighbors).

### 2.2.2 AQI Calculation

Apart from the Fenstock Air Quality Index, there are numerous methods devised for the effective calculation of AQI depending on the location and the data available. Some of these are described below:

#### 2.2.2.1 AQI System of US EPA

U.S. EPA's AQI is defined with respect to the five main common pollutants: carbon monoxide (CO), nitrogen dioxide (NO2), ozone (O3), particulate matter (PM10 and PM2.5) and sulphur dioxide (SO2).

$$I_P = \frac{(I_{HI} - I_{LO})}{BP_{HI} - BP_{LO}}(C_P - BP_{LO}) + I_{LO}$$

where $I_P$=Index for pollutant P,
$C_P$=Rounded concentration of pollutant P,
$BP_{HI}$=Break point that is greater than or equal to $C_P$,
$BP_{LO}$=Breakpoint that is less than or equal to $C_P$,
$I_{HI}$=AQI value corresponding to $BP_{HI}$,
$I_{LO}$=AQI value corresponding to $BP_{LO}$ The highest individual pollutant index, $I_P$, represents the Air Quality Index (AQI) of the location. The above method does not have the flexibility to incorporate any number of air pollutants. The method also does not consider the pollutant aggregation and spatial aggregation. It can be used for determining the short term and long term air quality indices.

#### 2.2.2.2 New Air Quality Index (NAQI)

New Air Quality Index is based on Factor Analysis of the major pollutants. The concentration of each pollutant or their deviation from the mean or their standardized values are expressed as a linear combination of these factors. The first factor will cause the highest variance of AQI. The second will contribute less variance than first but more than the third factor and so on. NAQI=$\sum_{i=1}^{n}(P_iE_i)/\sum_{i=1}^{n}(E_i)$

where i=1 to n represent the 'n' principal components (or the pollutants) and E1, E2 and E3 are the initial eigen values ( $\geq$1) with respect to the percentage variance. The method can be applied to assess the relative air quality without facilitating or considering the spatial aggregation, health effects and uncertainty measures but considers pollutant aggregation.

#### 2.2.2.3 Air Quality Depreciation Index (AQDI)

It is used to measure the depreciation in air quality using the value function curves for individual pollutants. The method considers the pollutant aggregation to determine the depreciation in air quality with respect to standard air quality. The air quality depreciation is measured in a scale between 0 to -10. An index value of '0' represents most desirable air quality having no depreciation from the best possible air quality with respect to the pollutants under consideration, while an index value of -10 represents maximum depreciation or worst air quality.

$AQ_{dep}$ =$\sum_{i=1}^{n}(AQ_i * CW_i)$−$\sum_{i=1}^{n}(CW_i)$

where, AQi=Air Quality Index for the ith parameter and obtained from value function curve. In the value function curve 0 represents the worst

quality and 1 represents the best quality of air due to the pollutant under consideration,
$CW_i$ is composite weight for ith parameter, and
n is the total no of pollutants considered.

### 2.2.3 AQI Prediction

There are various statistical and machine learning algorithms that can be used in the prediction of AQI. Some of these are explained below:

#### 2.2.3.1 Support Vector Machines (SVMs)

Support vector machines (SVMs) are supervised learning models, with associated learning algorithms that analyze the data used for classification and regression analyses. In SVR, the set of training data includes predictor variables and observed response values. The goal is to find a function f(x) that deviates from yn (sample labels) by a value no greater than  (bias) for each training point x—that is, remain as flat as possible. Therefore, SVR is also known as tube regression.

#### 2.2.3.2 Random Forest (RF)

Random forests (RFs), or random decision forests, are an ensemble learning method for classification, regression, and other tasks. An RF operates by constructing multiple decision trees at different training times, and outputting the class representing the mode of classes (classification) or the mean prediction (regression) of individual trees [31].

The RF algorithm incorporates growing classification and regression trees (CARTs). Each CART is built using random vectors. For the RF-based classifier model, the main parameters were the number of decision trees, as well as the number of features (NF) in the random subset at each node in the growing trees. During model training, the number of decision trees was determined first. For the number of trees, a larger number is better, but takes longer to compute. A lower NF leads to a greater reduction in variance, but a larger increase in bias. NF can be defined using the empirical formula NF $= \sqrt{}M, where M denotes the total number of features.$

### 2.2.4 Time Series prediction - for pollutants

There are numerous machine learning algorithms that can be used for prediction in time series problems[17]. A few of these are described below:

#### 2.2.4.1 Autoregression (AR)

The autoregression (AR) method models the next step in the sequence as a linear function of the observations at prior time steps. The method is suitable for univariate time series without trend and seasonal components.

#### 2.2.4.2 Moving Average (MA)

The moving average (MA) method models the next step in the sequence as a linear function of the residual errors from a mean process at prior time steps. The method is suitable for univariate time series without trend and seasonal components.

#### 2.2.4.3 Autoregressive Moving Average (ARMA)

The Autoregressive Moving Average (ARMA) method models the next step in the sequence as a linear function of the observations and residual errors at prior time steps. It combines both Autoregression (AR) and Moving Average (MA) models. The method is suitable for univariate time series without t

Autoregressive Integrated Moving Average (ARIMA) The Autoregressive Integrated Moving Average (ARIMA) method models the next step in the sequence as a linear function of the differenced observations and residual errors at prior time steps. It combines both Autoregression (AR) and Moving Average (MA) models as well as a differencing pre-processing step of the sequence to make the sequence stationary, called integration (I). An ARIMA model can also be used to develop AR, MA, and ARMA models. The method is suitable for univariate time series with trend and without seasonal components.

#### 2.2.4.4 Simple Exponential Smoothing (SES)

This method models the next time step as an exponentially weighted linear function of observations at prior time steps. The method is suitable for univariate time series without trend and seasonal components.

#### 2.2.4.5 Holt Winter's Exponential Smoothing (HWES)

Also called the Triple Exponential Smoothing method, models the next time step as an exponentially weighted linear function of observations at prior time steps, taking trends and seasonality into account. The method is suitable for univariate time series with trend and/or seasonal components.

### 2.2.5 Stationarity

A stationary time series is the one for which the properties (namely mean, variance and covariance) do not depend on time. It is necessary to determine stationarity and convert the data (if non-stationary) into a stationary dataset as most statistical models require the series to be stationary to make effective and precise predictions. The different methods of making a dataset stationary as explained below help us to further determine which statistical test to apply to the given non-stationary dataset. (i) Differencing: difference of consecutive terms in the series is computed. (ii) Seasonal Differencing: Calculation of the difference between an observation and a previous observation from the same season. (iii) Transformation: used to stabilize the non-constant variance of a series. The KPSS (Kwiatkowski-Phillips-Schmidt-Shin) Test is one test that can be used to determine the stationarity of a data. The test is as explained below: Null Hypothesis: The process is trend stationary. Alternate Hypothesis: The series has a unit root (series is not stationary). Test for stationarity: If the test statistic is greater than the critical value, we reject the null hypothesis (series is not stationary). If the test statistic is less than the critical value, if fail to reject the null hypothesis (series is stationary). For the air passenger data, the value of the test statistic is greater than the critical value at all confidence intervals, and hence we can say that the series is not stationary[18].

## 2.3 Related Works

### 2.3.1 Estimation of missing values in air pollution data using single imputation techniques

Noor Norazian et al implemented various techniques to replace the missing values in air pollution data obtained using automated machines to avoid bias between observed and unobserved values. The interpolation technique involved the computation of missing values using a polynomial equation whose degree depended on the number of neighboring values present. Secondly, the mean imputation technique was deployed wherein the absent data was filled by calculating the mean of the neighboring values.[19]

### 2.3.2 Understanding Shannon's Entropy metric for Information

Shriram presented a thesis explaining the meaning of the "amount of information" and thereby Shannon's Entropy. The research included the comparison of the information present in a completely biased coin against

an entirely fair coin. Citing this, the metrics were defined and the mathematical formula was thus derived.[20]

### 2.3.3 QQ-plots for assessing distributions of biomarker measurements and generating defensible summary statistics

Pliel in his research generated various QQ-plots for assessing distributions of biomarker measurements and generating defensible summary statistics and further discussed the methodology for interpreting and evaluating data distributions using quartile-quartile plots (QQ-plots) and making decisions as to how to treat outliers, interpreting effects of mixed distributions, and identifying left-censored data. The research includes the plot of 200 hypothetical breath toluene and also involves the comparison of QQ-plots with the traditional statistical measurement distributions. The study concludes with the various applications (like comparing population and assessing relative risk) and different use-cases of QQ-plots viz, detecting outliers in data and determining the data distribution.[21]

### 2.3.4 Hurst Exponent and Financial Market Predictibility

Qion et al posited a study which delineated the association of the Hurst Exponent and the behavior of the financial market. The Hurst Exponent was calculated using the R/S analysis and time-series forecasting of the stock market was done using a single-layer neural network. The results obtained were thus compared which were found to be in alignment with the initial hypothesis and hence it was concluded that the Hurst Exponent provides a measure of predictability in time-series data.[22]

### 2.3.5 A Comparative Study of Air Quality Index Based on Factor Analysis and US-EPA Methods for an Urban Environment

[23] Bishoi et al posited the EPA method for the computation of AQI (EPAQI). This technique involved the calculation of the index value for each pollutant (SO2, NO2, CO, O3, PM). The EPAQI was then evaluated by determining the maximum index value of the single pollutant which provided a rough estimate of the impact on the quality of air on human health. Furthermore, the research involved the Factor Analysis method to calculate the New AQI (NAQI) encompassing the Principal Component Analysis (PCA), which was used to ascertain whether the air quality has worsened or improved over the months.

### 2.3.6 Forecasting of air quality in Delhi using principal component regression technique

Anikender proposed a forecasting model to predict the AQI value which implemented the technique of Multiple Linear Regression and Principal Component Regression model. This research model included the usage of the past days' AQI values. These values were computed using the EPA, 1999 formula.[24]

# Chapter 3

# Proposed System

## 3.1 Problem Statement

To develop an efficient system to calculate the Air Quality Index based on various statistical and mathematical concepts of Chaos Theory, Self-Similarity, Fractals and so on. Further, to create an effective and accurate forecasting model using Neural Networks to predict the AQI values based on the past data.

## 3.2 Scope

The scope of the project includes the development of a uniform AQI considering various objectives such as health impacts, air quality standards, existing and future monitoring scenario including parameters, methods and frequency of measurements and other relevant aspects. It further involves the qualitative description of air quality and associated likely health impacts for different AQI values thereby enabling the policy makers to establish required rules and regulations for the enhancement of the quality of air. Unfortunately, working with real time data and to provide any means to collect and scrape data is out of scope for this project.

## 3.3 Proposed System

In this system, various statistical models are used to get an insight into the data. The bifurcation and chaos theory is used to determine whether the data is chaotic in nature. If the data is found to be chaotic then there is no relationship among the data parameters and hence no predictions can be made. Similarly, if the data is found to be non-chaotic, then it is said to be suitable for making predictions. Fractals and self-similarity are used to better understand the underlying patterns in the dataset. The Hurst exponent calculation is used to determine the type of correlation among data dimensions and to determine the nature of the dataset. Furthermore, a detailed trend analysis of the air pollutants is carried out using Mann-Kendall test and data decomposition by the system provides an even better understanding of the dataset. Many data visualization techniques such as the QQ plot are applied to impart various inferences to the “non-technical” people. The system calculates the air quality index with the help of the fenstock index thus providing the air quality index of past, present and future days along with the concentrations of the different pollutants involved. It also provides location wise AQI value in different parts of Delhi.

The stationarity of the dataset is then checked using the ADF test

as this is a prerequisite for developing an ARIMA model. The dataset, if non-stationarity, is converted into one by using either the differencing or the rolling method. ARIMA is a forecasting model that is deployed for the prediction of the pollutants' concentration. Further, the LSTM (Long Short Term Memory) is also used in the forecasting of the AQI as an alternative. The final system is capable of providing predictions to help enhance the air quality. In addition, the citizens are provided with a set of general instructions that would assist them in inferring these aforementioned values and utilize this information to their advantage hence assisting and encouraging them to play their part in the pollution control.

The system is designed using the streamlit framework, which gives the user flexibility to change the parameters required for the trend analysis and provides users to select the pollutant that is to be scrutinized with the statistical concepts applied according to their convenience.

# Chapter 4

# Design Of the System

## 4.1 Requirement Engineering

### 4.1.1 Requirement Elicitation

The final system should be able to generate predictions of the concentrations of pollutants and the Air Quality Index values for the upcoming week which enables the citizens to use this information and plan their days accordingly. To achieve the required performance accuracy and efficiency, it is necessary that the dataset used contains a large number of data points spanning across multiple years or even upto a few decades. Finally, the forecasting models are optimized by altering the relevant parameters.

### 4.1.2 Software lifecycle model

The software lifecycle model used here is the Agile development model. Agile methodology is based on collaborative decision making between requirements and solutions teams, and a cyclical, iterative progression of producing working software.

Tasks were carried out in regularly iterated cycles, or sprints, that usually lasted three to four weeks.

Each sprint had a stack of new and old requirements known as the backlog. Regular Scrum meetings (weekly meetings) took place between the development team members during a sprint. The Scrum meetings were performed to share the status of the work being performed on the backlog of the sprint and to identify potential issues to be added to the backlog of the next sprint.

### 4.1.3 Requirement Analysis

#### 4.1.3.1 UML Diagrams

1. Use Case Diagram

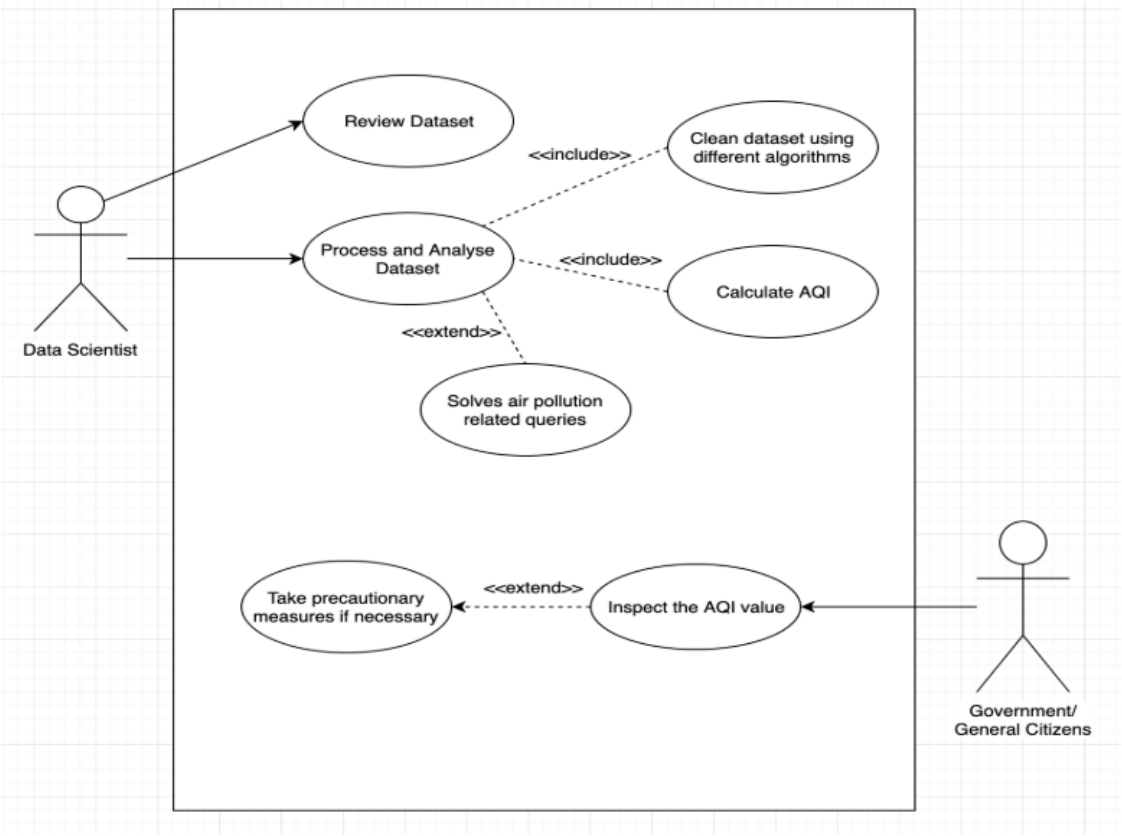



Figure 4.1: Use Case Diagram

2. Activity Diagram

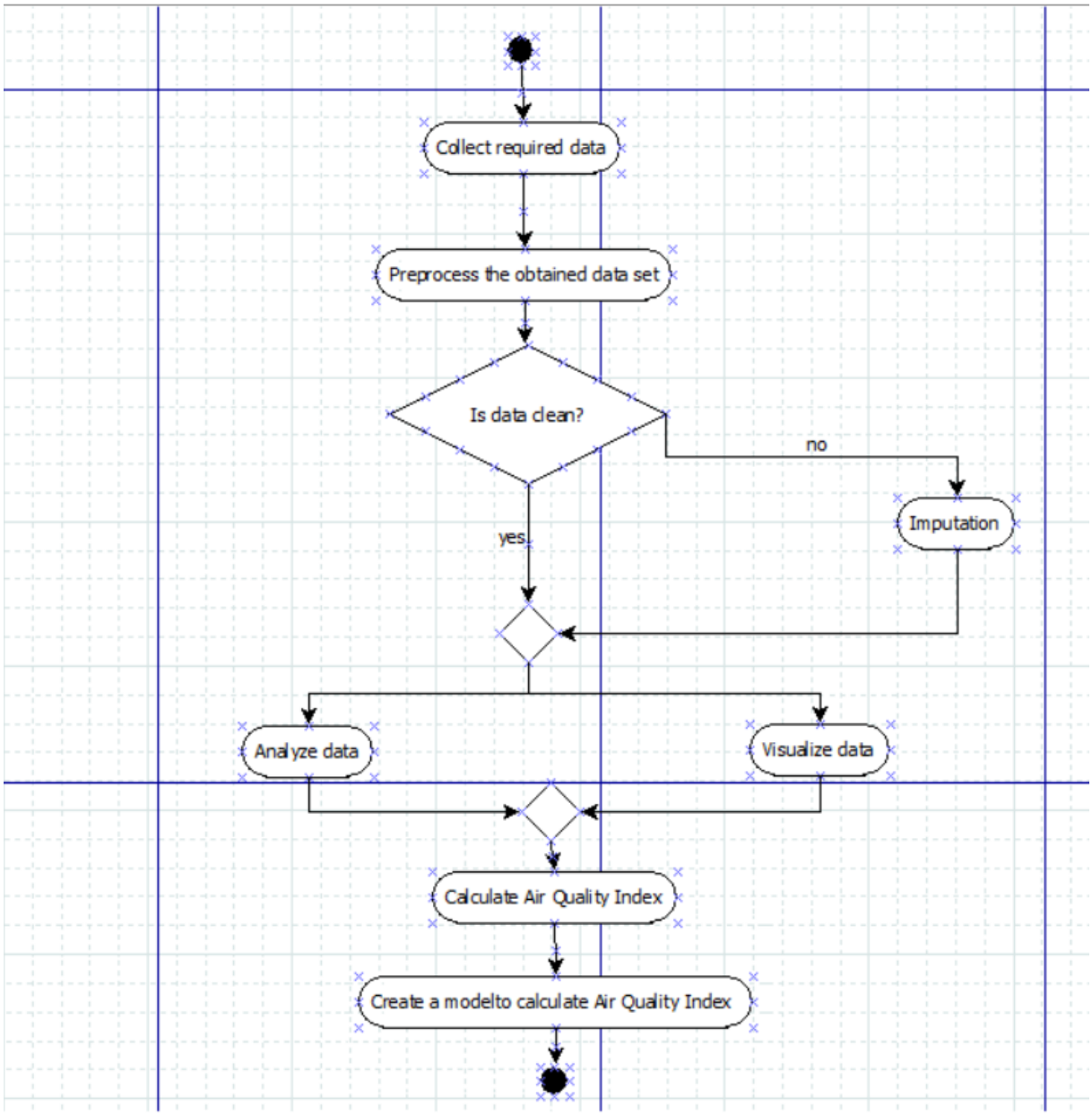



Figure 4.2: Activity Diagram

#### 4.1.3.2 Cost Analysis

The training of neural networks involves high configuration computers. Moreover, the anaconda environment requires a computer of high processing speed.

#### 4.1.3.3 Hardware Requirement

Computer with specifications:

1. Intel Pentium IV Processor or above

2. Internet connectivity
3. $\geq 4GBRam$

#### 4.1.3.4 Software Requirements

1. Operating system - Linux Ubuntu 18.04 or windows 8 and greater
2. Anaconda Environment
3. Jupyter Notebook
4. Python 3
5. Dependencies
    - Front end - Streamlit
    - Machine learning - Tensorflow and keras
    - Data Analysis - Pandas, Numpy, Scikit-learn, Hurst, Nolds, Statsmodel
    - Visualization - plotly, matplotlib, seaborn

## 4.2 Block Diagram

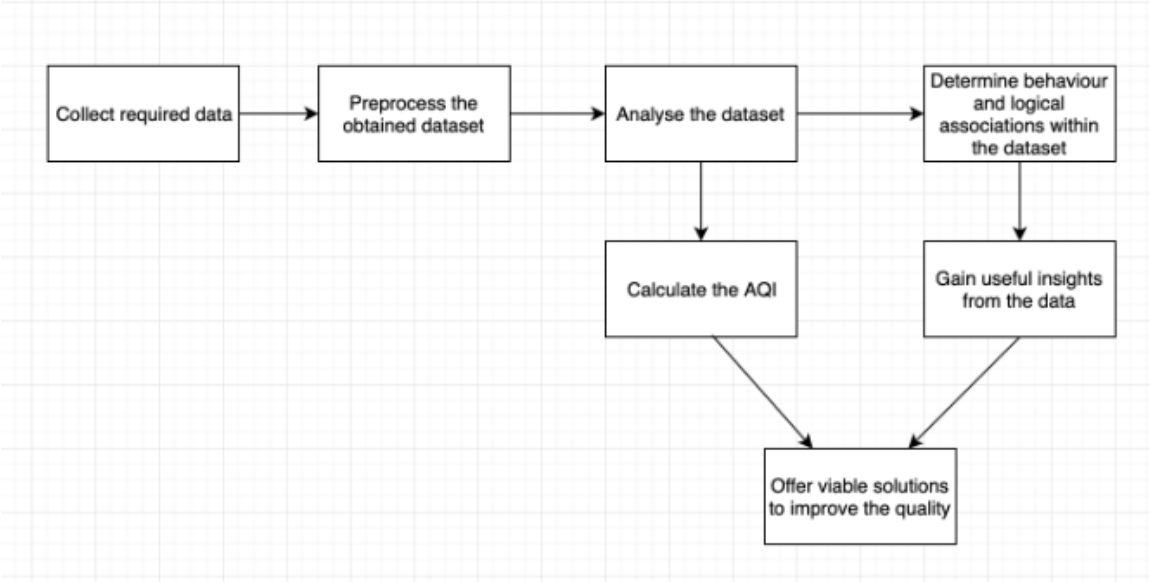



Figure 4.3: Block Diagram

# Chapter 5

# Result and Discussion

## 5.1 Screenshots of the System

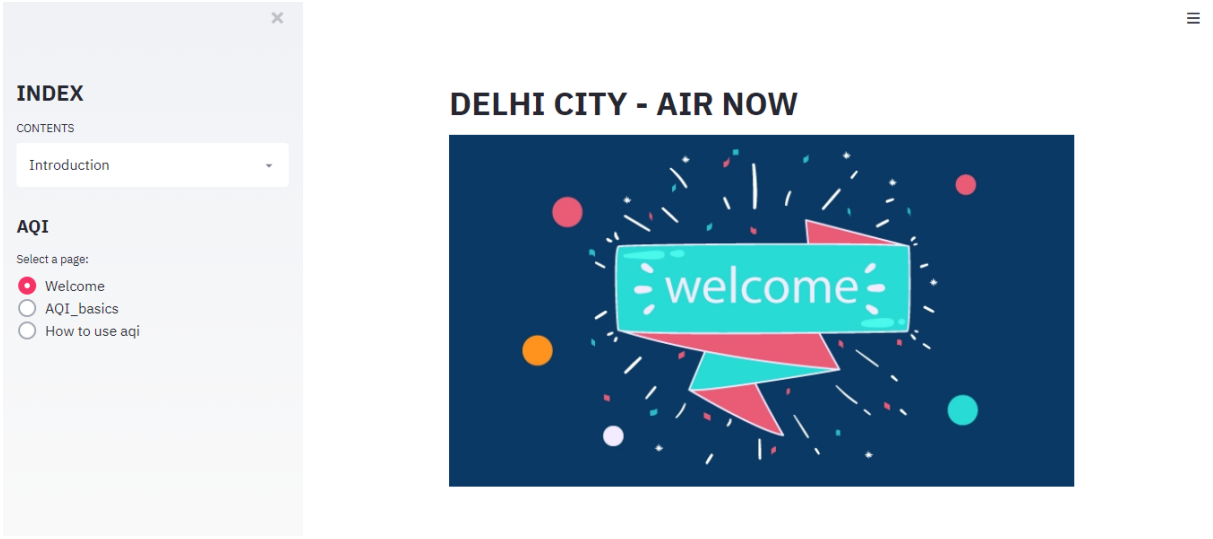



Figure 5.1: Welcome Page

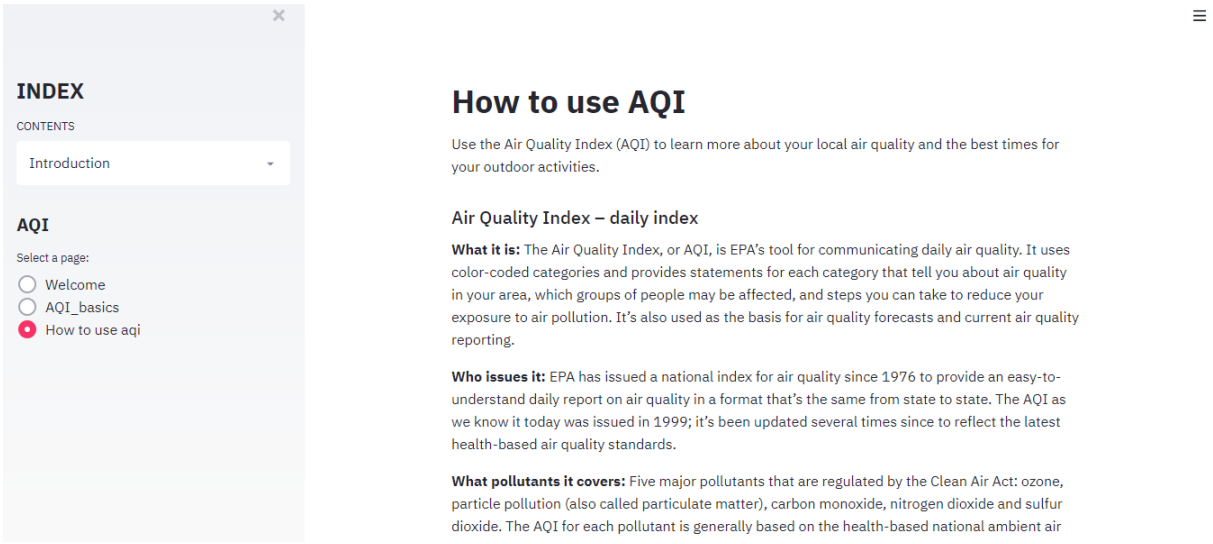



Figure 5.2: Instruction manual to understand AQI Values

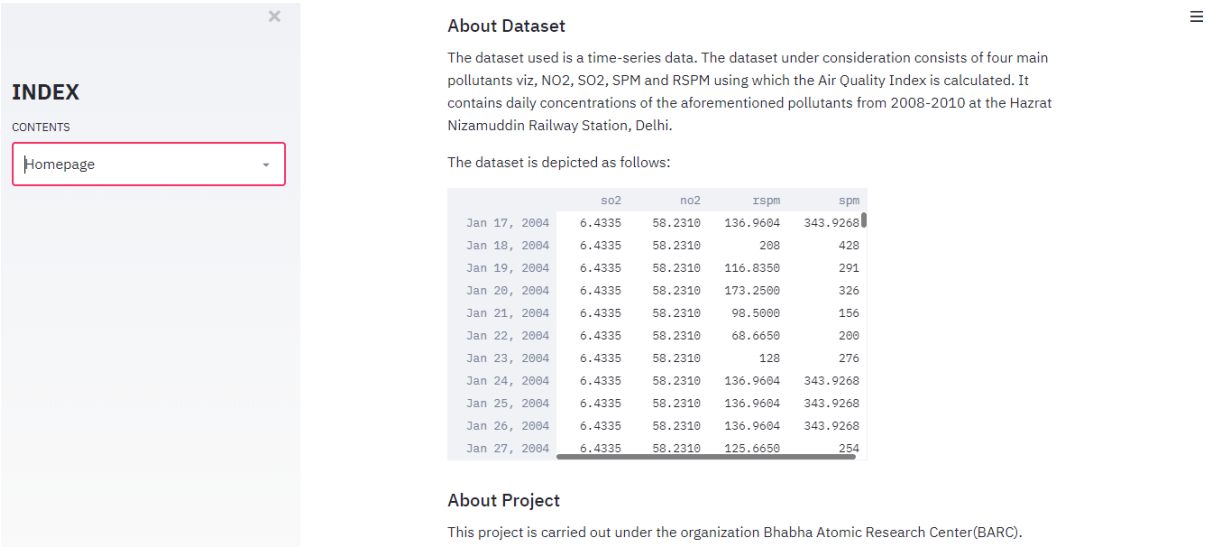



Figure 5.3: Homepage - A brief Description

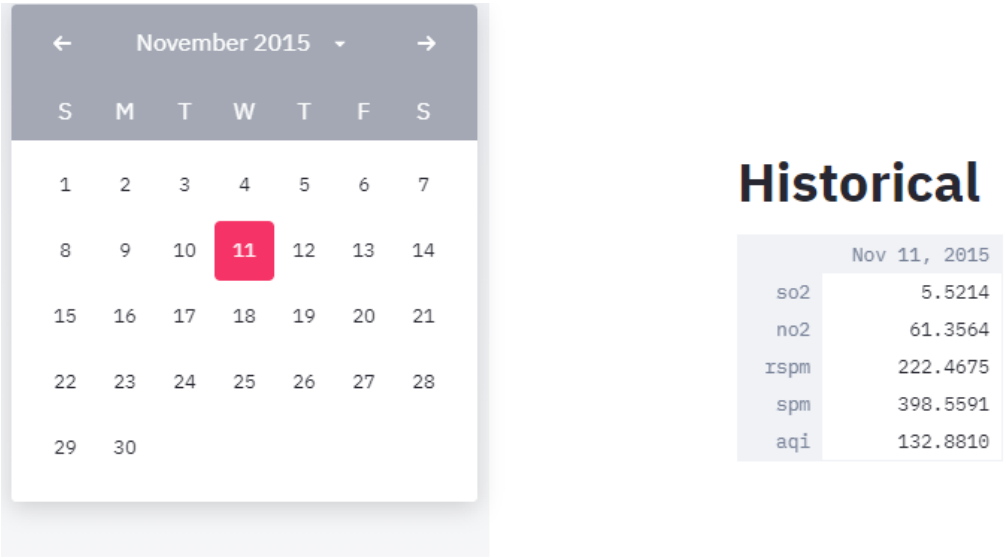
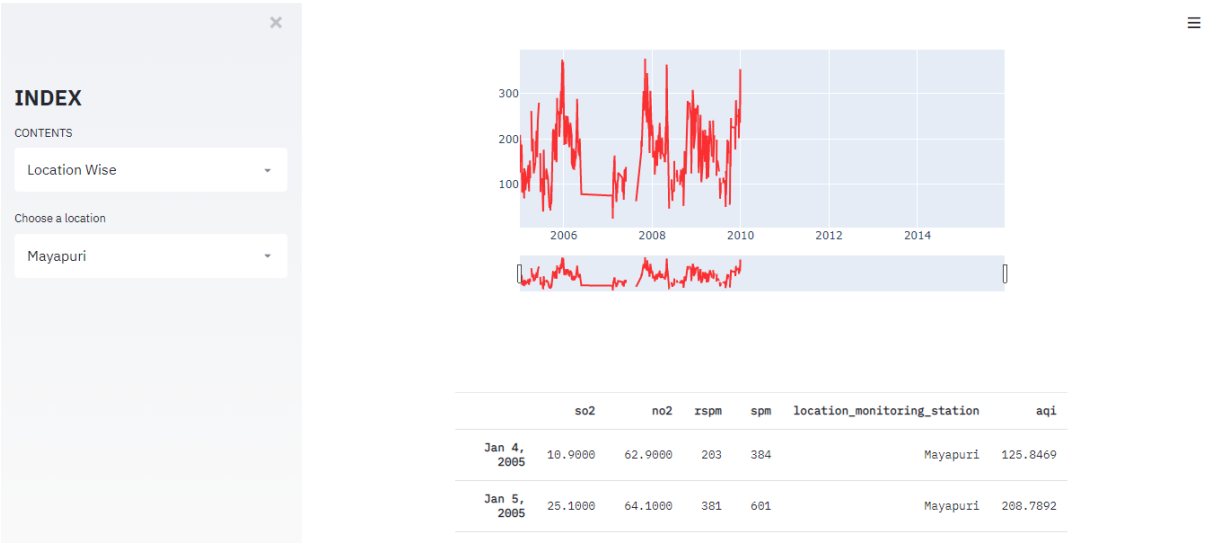
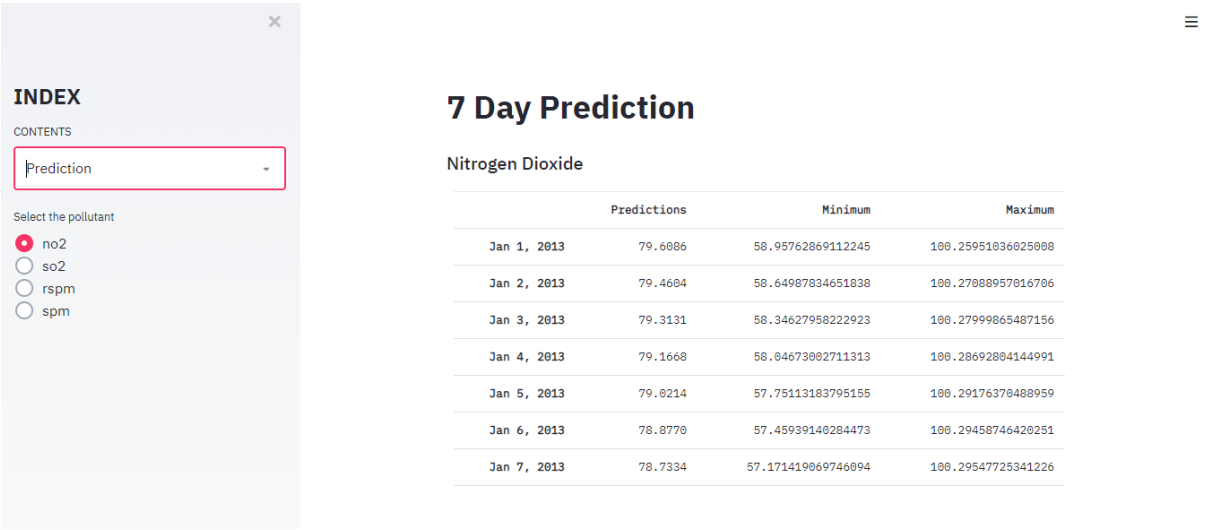



Figure 5.4: Showing historical information to user




| | so2 | no2 | rspm | spm | location_monitoring_station | aqi |
|---|---|---|---|---|---|---|
| Jan 4, 2005 | 10.9000 | 62.9000 | 203 | 384 | Mayapuri | 125.8469 |
| Jan 5, 2005 | 25.1000 | 64.1000 | 381 | 601 | Mayapuri | 208.7892 |

Figure 5.5: Depicting Location wise AQI and pollutant concentration values




| | Predictions | Minimum | Maximum |
|---|---|---|---|
| Jan 1, 2013 | 79.6086 | 58.95762869112245 | 100.25951036025008 |
| Jan 2, 2013 | 79.4604 | 58.64987834651838 | 100.27088957016706 |
| Jan 3, 2013 | 79.3131 | 58.34627958222923 | 100.27999865487156 |
| Jan 4, 2013 | 79.1668 | 58.04673002711313 | 100.28692804144991 |
| Jan 5, 2013 | 79.0214 | 57.75113183795155 | 100.29176370488959 |
| Jan 6, 2013 | 78.8770 | 57.45939140284473 | 100.29458746420251 |
| Jan 7, 2013 | 78.7334 | 57.171419069746094 | 100.29547725341226 |

Figure 5.6: Providing 7 days forecast

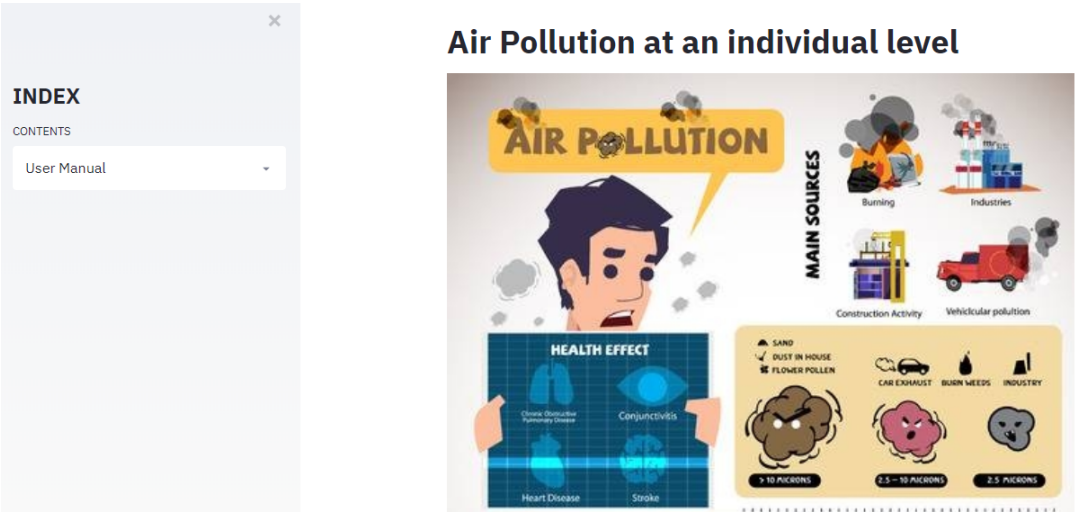



Figure 5.7: Providing basic information to user

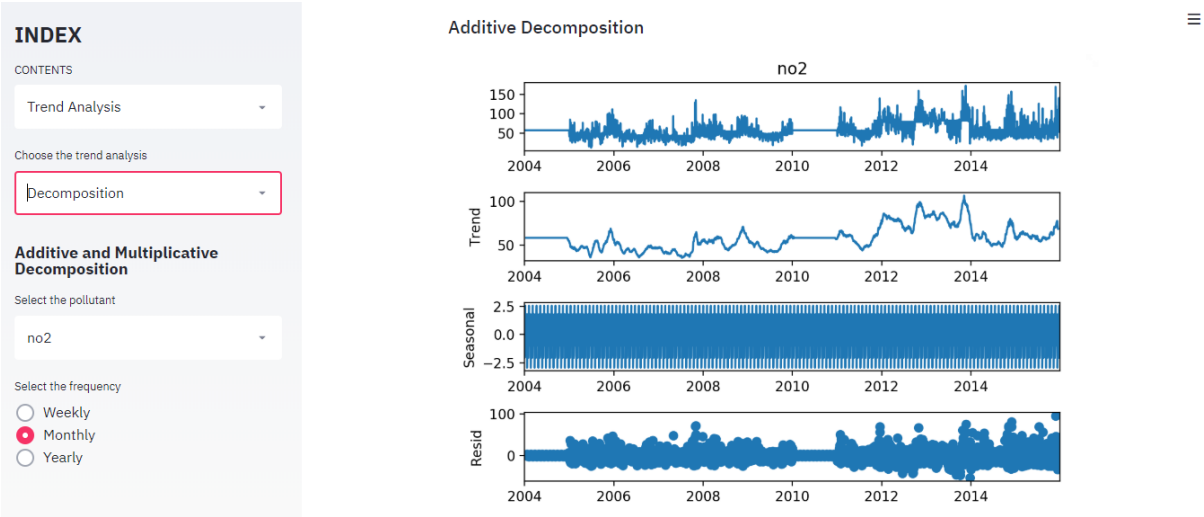



Figure 5.8: Data Decomposition

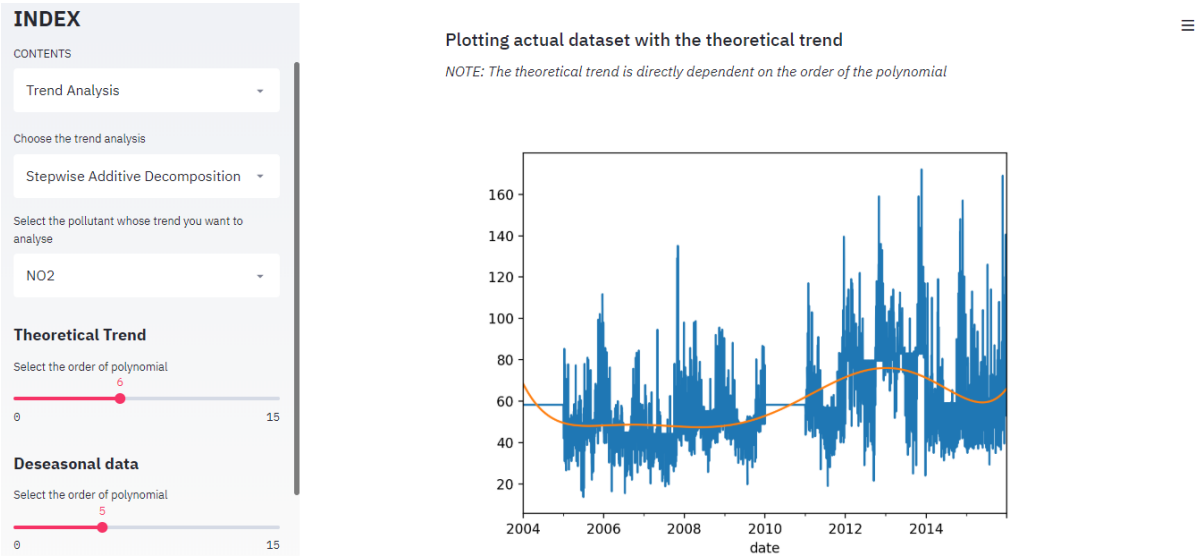



Figure 5.9: Page for interactive trend analysis

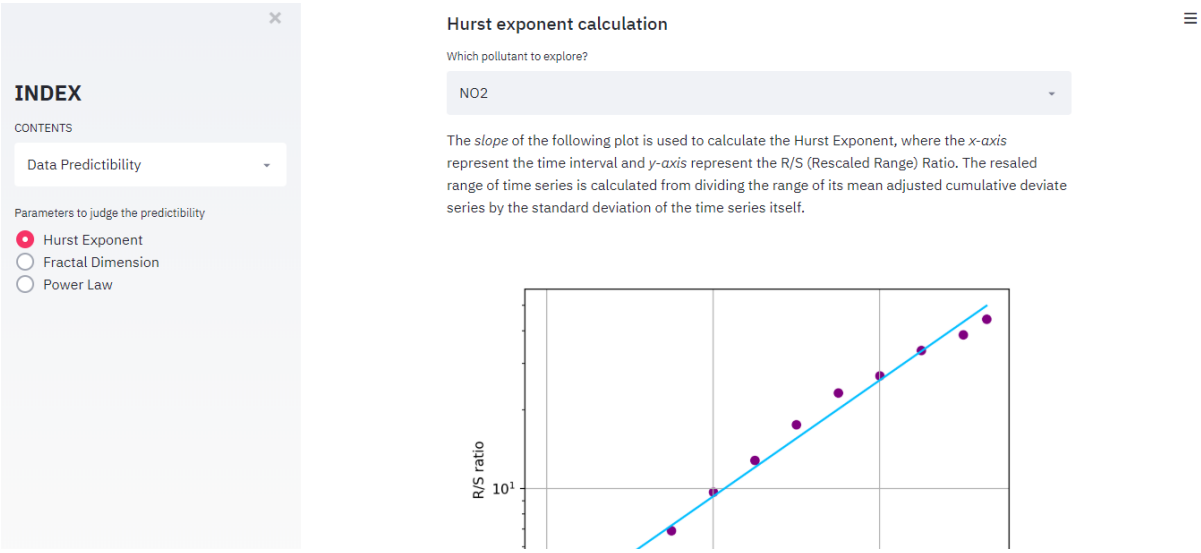



Figure 5.10: Tab for predictability analysis

## 5.2 Sample Code (of imp part/ main logic)

1. LSTM Model

```
EPOCHS = 50
BATCH_SIZE = 100
NAME = f"{SEQ_LEN}-SEQ-{FUTURE_PERIOD_PREDICT}-PRED-{int(time.time())}"
```

```
model = Sequential()
# input layer
model.add(LSTM(128, input_shape=(train_x.shape[1:]), return_sequences=True, activation='relu'))
model.add(Dropout(0.2))
model.add(BatchNormalization())

#hidden layer
model.add(LSTM(128, input_shape=(train_x.shape[1:]), return_sequences=True,activation='relu'))
model.add(Dropout(0.1))
model.add(BatchNormalization())

model.add(LSTM(128, input_shape=(train_x.shape[1:]), activation='relu'))
model.add(Dropout(0.2))
model.add(BatchNormalization())

model.add(Dense(32, activation='relu'))
model.add(Dropout(0.2))

#output layer
model.add(Dense(2,activation='softmax'))

opt = tf.keras.optimizers.Adam(lr=0.001, decay=1e-6)

model.compile(loss='sparse_categorical_crossentropy',
            optimizer=opt,
            metrics=['accuracy'])


history = model.fit(
                    train_x, train_y,
                    batch_size = BATCH_SIZE,
                    epochs=EPOCHS,
                    validation_data=(validation_x, validation_y))
                    #callbacks=[tensorboard,checkpoint])
```

Figure 5.11: Code snippet for LSTM

2. Air Quality Index Calculation

```
w = [5,43,133,201] #these values are taken as per the Indian standards
w_sum = 0
for i in range(0,len(w)):
    w_sum += w[i]

w_ratio = []
for i in range(len(w)):
    x = float(w[i])/float(w_sum)
    w_ratio.append(x)
print(w_ratio)
```

```
[0.013089005235602094, 0.112565445026178, 0.3481675392670157, 0.5261780104712042]
```

```
s = [50,40,60,40]
list_aqi = []
conc = []

for index,row in df.iterrows():
    # concentration (mean/median) for each pollutant
    conc = [ row['so2'], row['no2'], row['spm'], row['rspm'] ]

    # calculating pollution index
    q = []
    for i in range(0,4):
        z = s[i]*conc[i]/100
        q.append(z)

    #AQI calculation
    aqi = 0
    for i in range(0,4):
        aqi += q[i]*w_ratio[i]

    list_aqi.append(aqi)
```

Figure 5.12: Code to calculate Air Quality Index

3. ARIMA

```
model = ARIMA(train, order = (1,0,1))
results = model.fit()

#Printing the summary of ARIMA model
print(results.summary())
```

Figure 5.13: Code snippet to fit an ARIMA Model

```
warnings.filterwarnings('ignore')

param_df = pd.DataFrame(columns=['Parameter', 'AIC'])

i=0
for param in pdq:
    try:
        model_arima = ARIMA(train, order=param)
        model_arima_fit = model_arima.fit()
        param_df.loc[i] = [param, model_arima_fit.aic]
        i = i + 1
    except:
        continue
```

```
param_df[param_df['AIC']==param_df['AIC'].min()]
```

Figure 5.14: Code Snippet to find minimum AIC for ARIMA

## 5.3 Testing

The arima model requires p,q, d as its input parameters where p is order of Auto-Regressive model, q is order of Moving Average model, and d is the number of times differencing performed on the dataset to make it stationary. Since our dataset is stationary in nature the value of d is zero. Value of p and q is determined from the figure 5.15.

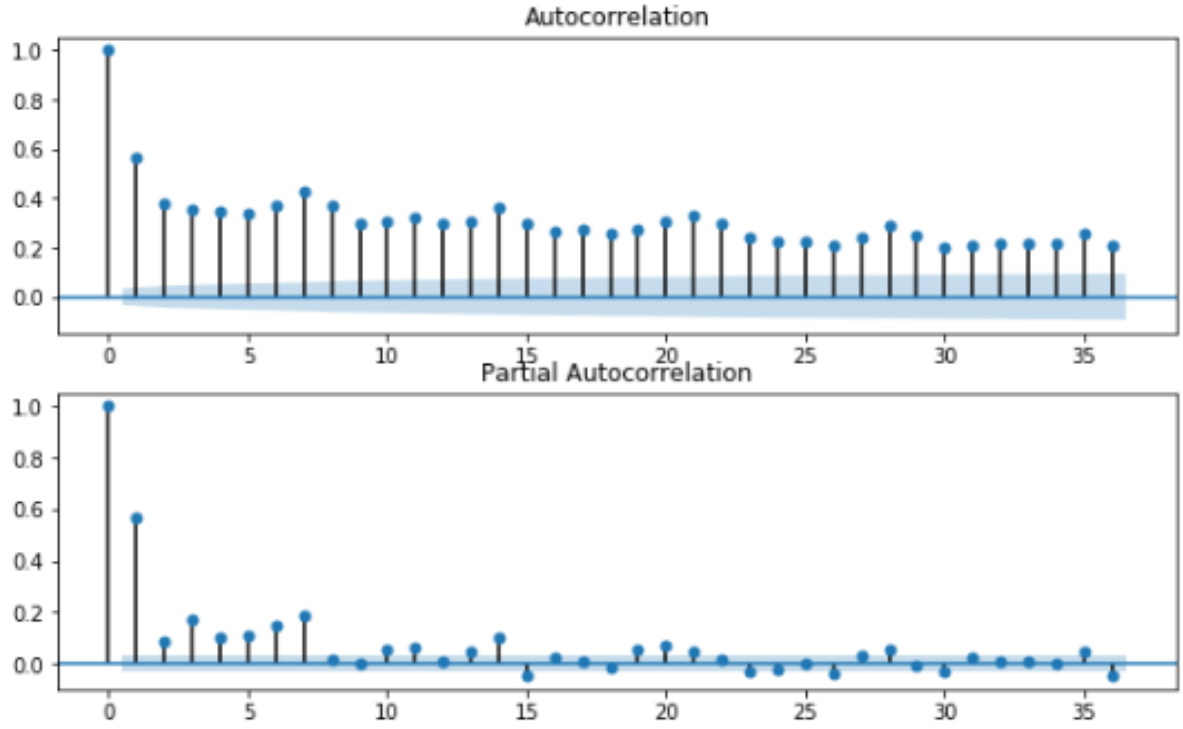



Figure 5.15: Autocorrelation and Partial Autocorrelation graph

Since we cannot infer the accurate values for p and q from this graph, an alternative approach is applied, called AIC.The Akaike Information Criteria (AIC) is a widely used measure of a statistical model. It basically quantifies 1) the goodness of fit, and 2) the simplicity/parsimony, of the model into a single statistic.

When comparing two models, the one with the lower AIC is generally "better". In this all the possible value is fed into the model input parameters and the model which produces the minimum AIC value is accepted for the fitting of the model to the dataset.

After using this value, the model is again cross referenced. The predictions made by the model are tested with the testing data and these predicted pollutant concentrations are plotted against the testing data as shown in figure 5.16 and 5.17.

The predictions are verified using the chi-square goodness of fit test. For

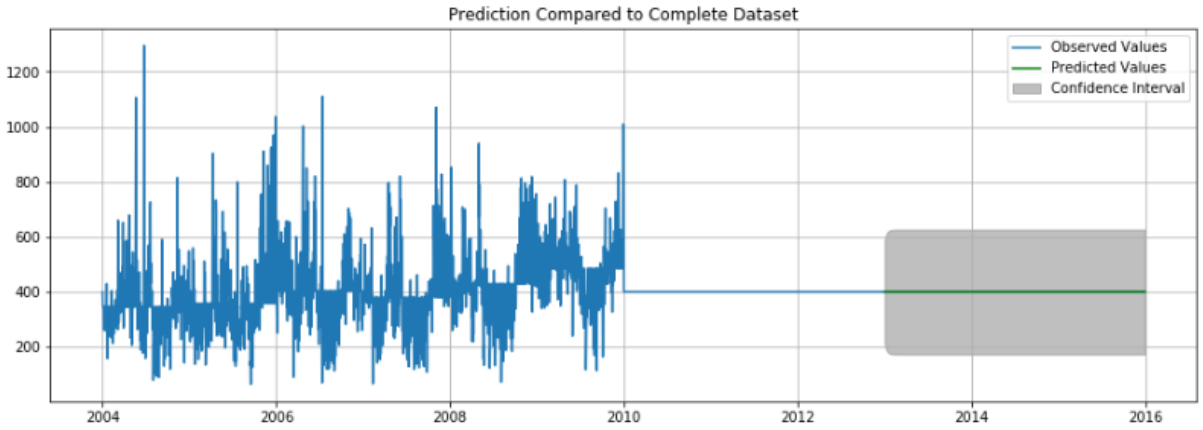



Figure 5.16: Prediction Compared to Complete Dataset

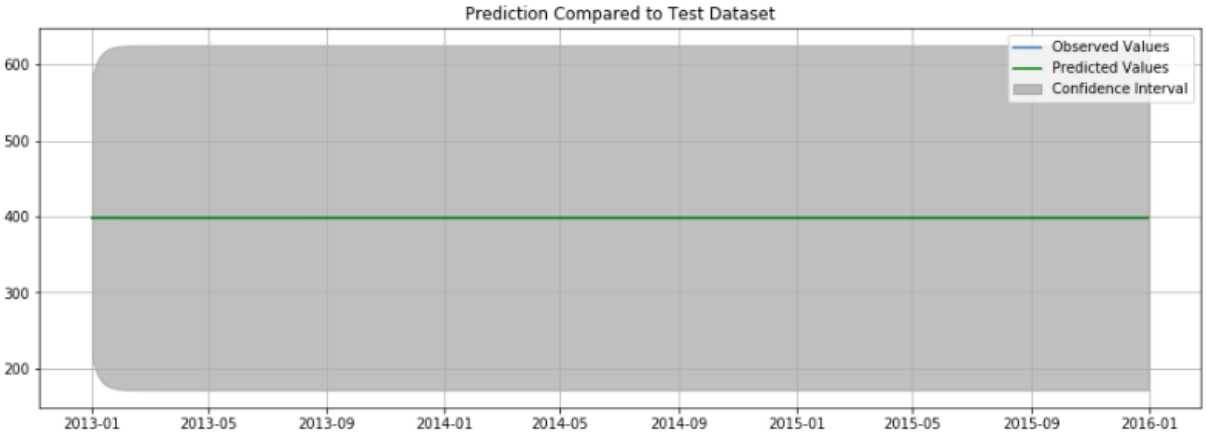



Figure 5.17: Prediction Compared to Test Dataset

a chi-square goodness of fit test, the hypotheses take the following form.

- Ho: The data are consistent with a specified distribution.
- Ha: The data are not consistent with a specified distribution.

The P-value is the probability of observing a sample statistic as extreme as the test statistic. It can be used to decide whether to accept or reject a hypothesis .

Since the P-value is 1.00 for this model which is more than the significance level (0.05), we can accept the null hypothesis. Thus the model passes the test.

## 5.4 Analysis Result

In this project a lot of statistical analysis were performed. The results of all these tests are enumerated below:

### 5.4.1 QQ Plotting

The points seem to fall about a straight line indicating that the sample space is normally distributed. However, the data points at the tail form a curve instead of a straight line owing to the skewness present in the sample data.

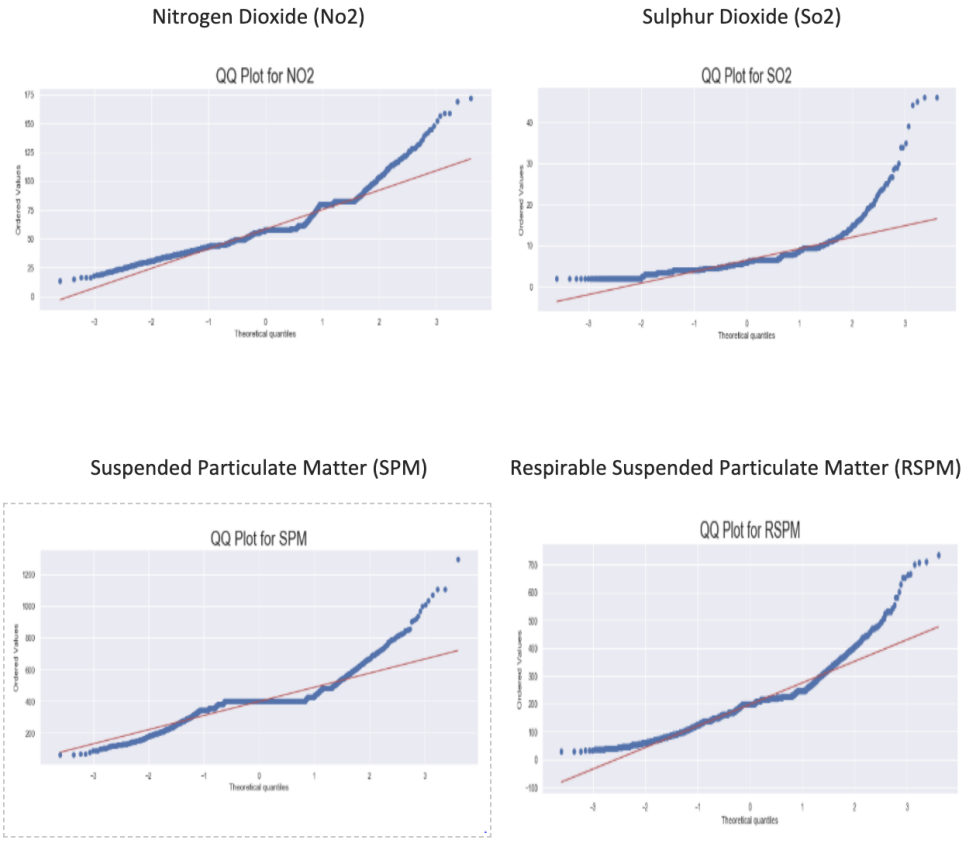



Figure 5.18: QQ Plot for all the pollutants

### 5.4.2 Decomposition

Additive decomposition is performed on our dataset which, as per the figure X, does not show any seasonality nor the presence of a defined trend. This in turn proves to be advantageous for upcoming forecasting models.


Figure 5.19: Additive Decomposition of Data

### 5.4.3 Shannon's Entropy

Thus the variable residual suspended particulate matter contains the maximum information while the suspended particulate matter contains very less. Nitrogen dioxide and sulphur dioxide contain enough information for us to not neglect them. Thus while calculating aqi we can omit the pollutant spm as it does not contribute much information to our dataset.

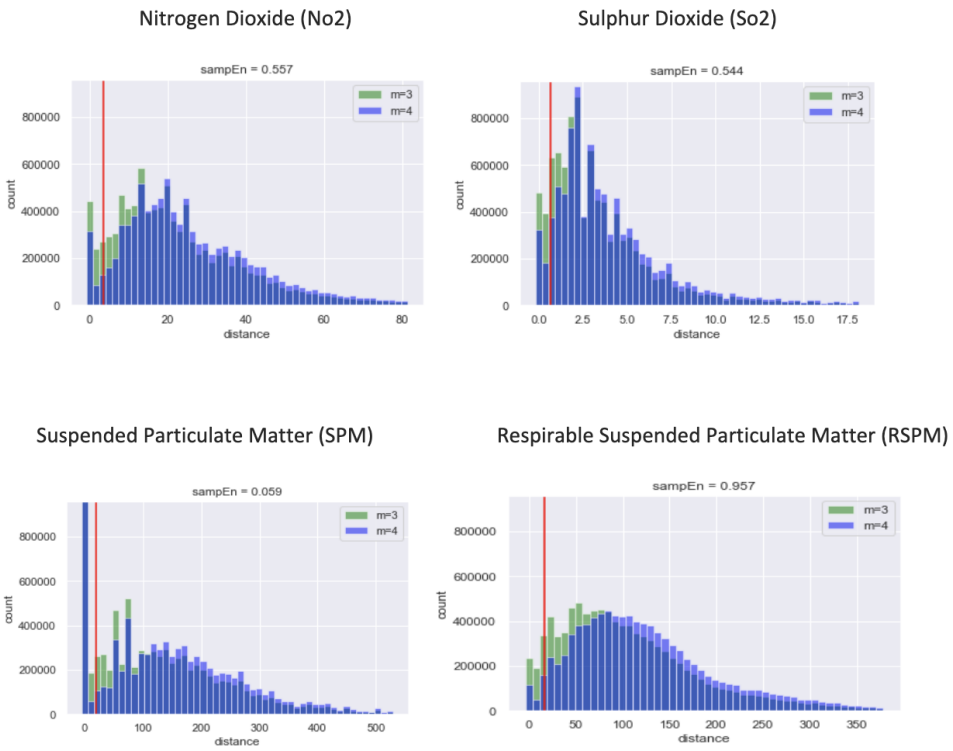



Figure 5.20: Shannon's Entropy for all the pollutants

### 5.4.4 Fractal Dimension

The fractal dimension for all the pollutants is close to 1.5 depicting that there is less complexity in our dataset.

| Pollutant | Fractal Dimension Value |
|---|---|
| Nitrogen Dioxide (NO2) | 1.5569 |
| Sulphur Dioxide (SO2) | 1.5651 |
| Suspended Particulate Matter (SPM) | 1.5167 |
| Residual Suspended Particulate Matter (RSPM) | 1.5469 |

Table 5.1: Fractal Dimension Values

### 5.4.5 Power law

Here for all the variables,

$$\beta \approx 2$$

Thus, we can conclude that our data have brownian noise in it and all the variables are strongly correlated.

| Pollutant | Power Law Value |
|---|---|
| Nitrogen Dioxide (NO2) | 1.8862183860251993 |
| Sulphur Dioxide (SO2) | 1.8697414987080596 |
| Suspended Particulate Matter (SPM) | 1.9062648883918447 |
| Residual Suspended Particulate Matter (RSPM) | 1.9666765264457275 |

Table 5.2: Power Law Values

### 5.4.6 Hurst Exponent

Thus the pollutant so2 and no2 are anti persistent in nature and rspm and spm are brownian time series.

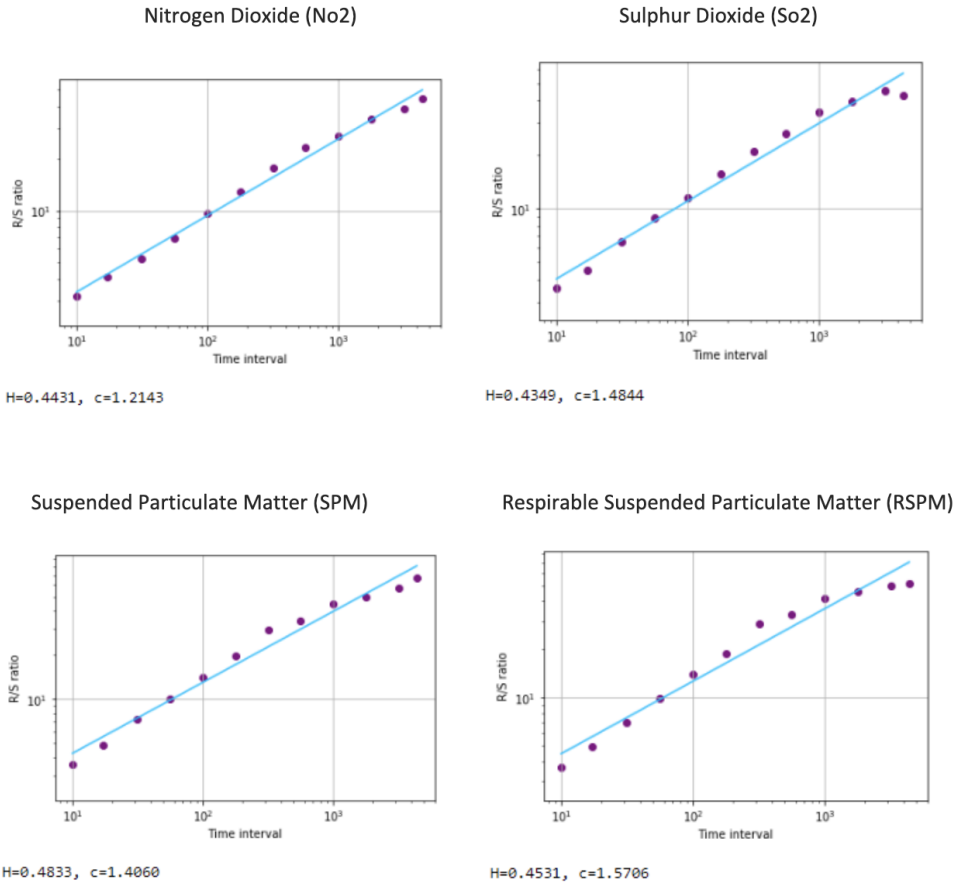



Figure 5.21: Hurst Exponent for all the pollutants

| Pollutant | Hurst Exponent Value |
|---|---|
| Nitrogen Dioxide (NO2) | 0.4344 |
| Sulphur Dioxide (SO2) | 0.4290 |
| Suspended Particulate Matter (SPM) | 0.4860 |
| Residual Suspended Particulate Matter (RSPM) | 0.4506 |

Table 5.3: Hurst Exponent Values

### 5.4.7 Lyapunov's Coefficient

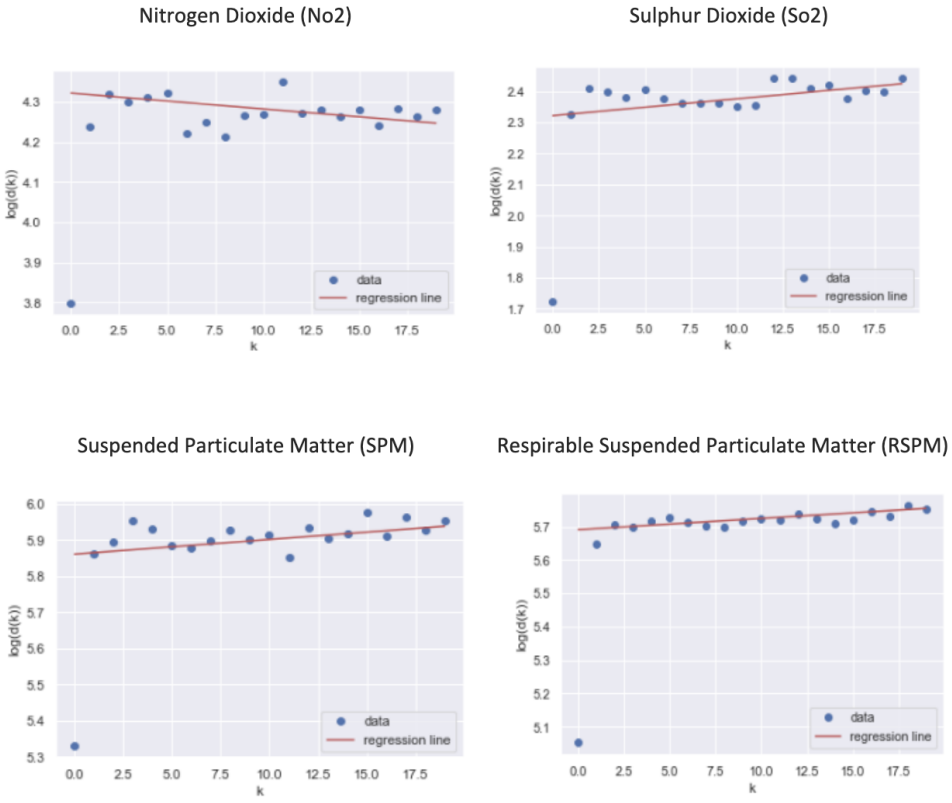



Figure 5.22: Lyapunov's Exponent for all the pollutants

As per table 5.4 all the values are approximately zero, the predictability is like a random walk, i.e, the information content levels off; we have neither information loss nor information increase. Moreover, zero value represents that the dataset is not chaotic in nature thus it will be convenient to make predictions.

### 5.4.8 Air Quality Index

The annual average aqi is found to be increasing with each year. With its value in the range of 100-200, it falls under the "Unhealthy" category according to the figure 5.23.

| Pollutant | Lyapunov's Coefficient Value |
|---|---|
| Nitrogen Dioxide (NO2) | -0.003930704727153612 |
| Sulphur Dioxide (SO2) | 6.259258366785646e-08 |
| Suspended Particulate Matter (SPM) | 4.7037278381366356e-08 |
| Residual Suspended Particulate Matter (RSPM) | 3.900087333178925e-08 |

Table 5.4: Lyapunov's Coefficient Values

| Air Quality Index Levels of Health Concern | Numerical Value | Meaning |
|---|---|---|
| Good | 0 to 50 | Air quality is considered satisfactory, and air pollution poses little or no risk. |
| Moderate | 51 to 100 | Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution. |
| Unhealthy for Sensitive Groups | 101 to 150 | Members of sensitive groups may experience health effects. The general public is not likely to be affected. |
| Unhealthy | 151 to 200 | Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects. |
| Very Unhealthy | 201 to 300 | Health alert: everyone may experience more serious health effects. |
| Hazardous | 301 to 500 | Health warnings of emergency conditions. The entire population is more likely to be affected. |

Figure 5.23: Shannon's Entropy for all the pollutants

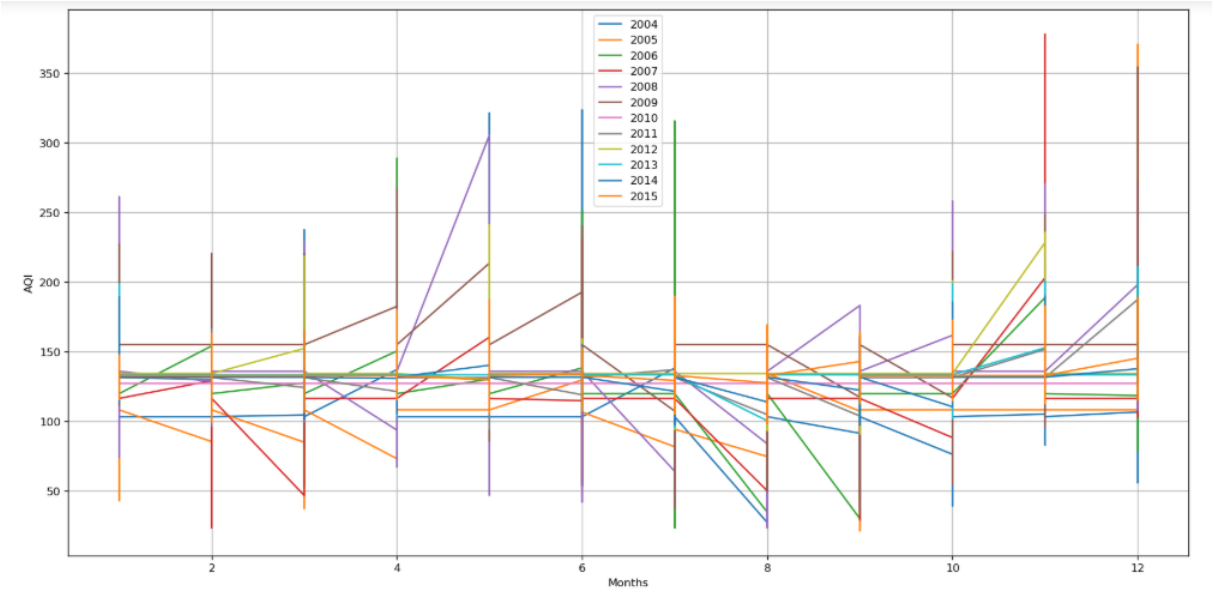



Figure 5.24: Yearly Air Quality Index

# Chapter 6

# Conclusion & Future Scope

The Air Quality Index System is an automated system that helps to determine the standard of the air existing in the atmosphere by computing the AQI of the region. The system requires only the concentration of the pollutants prevalent in the surroundings to determine the AQI. The system developed is simple, user-friendly keeping in mind its main purpose of being available for use by the general public, both technical and non-technical. Further, the analytics and the prediction system provide a great insight to the policy makers of the condition of this necessary life resource in the coming years. It includes a menu bar which makes it simpler to navigate through the results of the various statistical techniques used in the project. Albeit this system had to be burdened with a few technical jargons which were used in the analysis and forecasting models, it is a sophisticated platform to explore and examine the dataset under consideration. The LSTM Recurrent Neural Network renowned for its capability to "remember" data was deployed in the time-series predictive model which provided with an accuracy close to 75% .

As for the future scope of the project, the accuracy of the forecasting model can further be improved by training it even further against a larger dataset the lack of which was the difference between a great and a top-class model. Moreover, this model can prove to be the basis of similar applications like weather-forecasting, solar radiations prediction and other dimensions associated with climate prediction.

# References


[1] A. kim, "The intuition behind shannon's entropy," *towards data science*, 2018.

[2] S. Kwiatkowski, "Entropy is a measure of uncertainty," *towards data science*, 2018.

[3] M. Miller, "The basics: Time series and seasonal decomposition," *towards data science*, 2018.

[4] D. S. Alan Anderson, "Quantile-quantile (qq) plots: Graphical technique for statistical data," *Dummies :A Wiley Brand*.

[5] O. Rose, "Estimation of the hurst parameter of long-range dependent time series," *University of Wurzburg Institute of Computer Science Research Report Series*, 1996.

[6] G. D. G. C. Monica Cusenza1, Agostino Accardo, "Relationship between fractal dimension and power-law exponent of heart rate variability in normal and heart failure subjects," *1DEEI, University of Trieste, Trieste, Italy S. Maugeri Foundation, Rehabilitation Institute of Telese, Telese Terme, Italy Department of Health Sciences, University of Molise, Campobasso, Italy*.

[7] J. N. T. L. C. Louis M. Pecora, Linda Moniz, "A unified approach to attractor reconstruction," *Cornell University Journal*, 2006.

[8] A. U. Archana R and R. Gopikakumari, "Bifurcation analysis of chaotic systems using a model built on artificial neural networks," *2nd International Conference on Computational Techniques and Artificial Intelligence (ICCTAI'2013) March 17-18, 2013 Dubai (UAE)*, 2013.

[9] C. S. K. J Krishnaiah and M. Faruqi, "Constructing bifurcation diagram for a chaotic timeseries data through a recurrent neural network model," *Proceedings of the 9th International Conference on Neural Information Processing (ICONIP'OZ) , Vol. 5*.

[10] C. D.-. PR Division on behalf of Dr. A.B. Akolkar, Member Secretary, "National air quality index," *www.cpcb@nic.in.*, 2014.

[11] D. N. S. Krishna Chaitanya Atmakuri, Dr. V Anandam, "A survey paper on spatial - temporal outliers influencing air quality," *International Journal Of Engineering Research And Development*, 2018.

[12] J. Brownleer, "11 classical time series forecasting methods in python (cheat sheet)," 2020.

[13] A. SINGH, "A gentle introduction to handling a non-stationary time series in python," *Analytics vidhya*, 2018.

[14] "5 statistical methods for forecasting quantitative time series," *Bistasolution.com*, 2068.

[15] A. Swalin, "How to handle missing data," *towards data science*, 2018.

[16] S. Faisal and G. Tutz, *Nearest Neighbor Imputation for Categorical Data by Weighting of Attributes.* PhD thesis, Department of Statistics, Ludwig-Maximilians-Universitat M unchen, Ludwigstrasse 33, D-80539, Germany. 2Department of Statistics, Ludwig-Maximilians-Universit at M unchen, Akademiestrasse 1, D-80799 Munich, Germany., 2017.

[17] J. Brownleer, "11 classical time series forecasting methods in python (cheat sheet)," 2020.

[18] A. SINGH, "A gentle introduction to handling a non-stationary time series in python," *Analytics vidhya*, 2018.

[19] Y. A. S. R. N. A. A. M. M. A. B. Mohamed Noor Noraziana, *, "Estimation of missing values in air pollution data using single imputation techniques," 2008.

[20] S. Vajapeyam, "Understanding shannon's entropy metric for information," 2014.

[21] J. D. Pleilr, "Qq-plots for assessing distributions of biomarker measurements and generating defensible summary statistics," *National Exposure Research Laboratory, Office of Research and Development, US Environmental Protection Agency, Research Triangle Park, NC 27711, USA*, 2016.

[22] K. R. Bo Qian, "Hurst exponent and financial market predictability," *Department of Computer Science University of Georgia Athens, GA 30601.*

[23] V. J. Biswanath Bishoi, Amit Prakash, "A comparative study of air quality index based on factor analysis and us-epa methods for an urban environment," 2009.

[24] P. G. Anikender Kumar, "Forecasting of air quality in delhi using principal component regression technique," *Centre for Atmospheric Sciences, Indian Institute of Technology Delh*, 2011.

# Acknowledgement

We are profoundly grateful to **Dr.Lata Ragha** for her expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion.

We would like to express deepest appreciation towards **Dr.S.M.Khot**, Principal,Fr.C.Rodrigues Institute of Technology, **Dr.Lata Ragha**, Head of Department of Computer Engineering and **Dr.Debrabata Datta**,Head of Radiology,Physics and Advisory Division,BARC,Trombay whose invaluable guidance supported us in completing this project. We would also like to appreciate the efforts taken by **Mrs.Rakhi Kalantri**, Project Coordinator in order to keep us on track.

At last we must express our sincere heartfelt gratitude to all the staff members of Computer Engineering Department who helped us directly or indirectly during this course of work.

**Project Group Members:**

1. Pragya Verma, 101661

_____

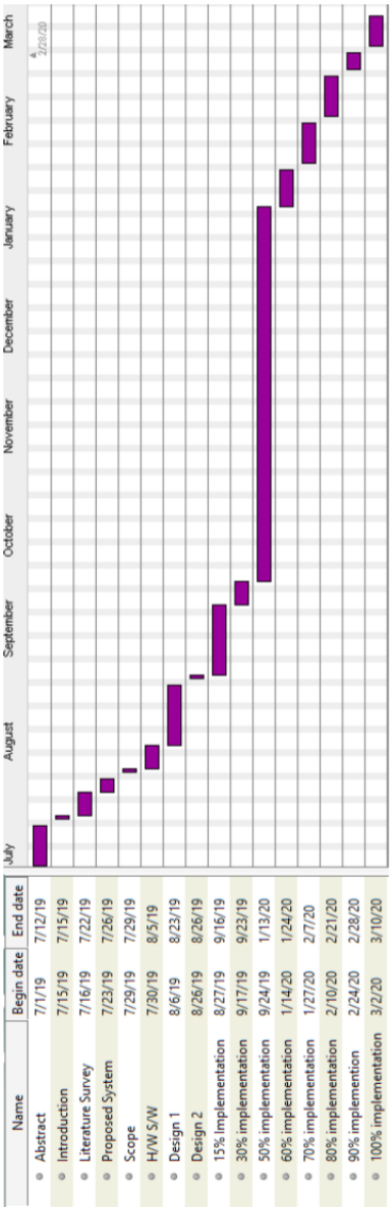
2. Sai Reddy, 101647

_____

3. Mithilesh Waghulade, 101663

_____

# Appendix A : Timeline Chart




| Name | Begin date | End date |
|---|---|---|
| Abstract | 7/1/19 | 7/12/19 |
| Introduction | 7/15/19 | 7/15/19 |
| Literature Survey | 7/16/19 | 7/22/19 |
| Proposed System | 7/23/19 | 7/26/19 |
| Scope | 7/29/19 | 7/29/19 |
| H/W S/W | 7/30/19 | 8/5/19 |
| Design 1 | 8/6/19 | 8/23/19 |
| Design 2 | 8/26/19 | 8/26/19 |
| 15% Implementation | 8/27/19 | 9/16/19 |
| 30% implementation | 9/17/19 | 9/23/19 |
| 50% implementation | 9/24/19 | 1/13/20 |
| 60% implementation | 1/14/20 | 1/24/20 |
| 70% implementation | 1/27/20 | 2/7/20 |
| 80% implementation | 2/10/20 | 2/21/20 |
| 90% implemention | 2/24/20 | 2/28/20 |
| 100% implementation | 3/2/20 | 3/10/20 |

Figure 6.1: Timeline Chart of the Entire Project

# Appendix B : Publication Details

This is appendix B