National College of Ireland

A Comprehensive Study to Forecast the Delhi and Bangalore Cities Air Pollution using Machine Learning Models

MSc Research Project
Data Analytics

Parth Darekar
Student ID: x19212739

School of Computing
National College of Ireland

Supervisor: Dr.Rashmi Gupta

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Parth Darekar
Student ID:	x19212739
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Dr.Rashmi Gupta
Submission Due Date:	16/08/2021
Project Title:	A Comprehensive Study to Forecast the Delhi and Bangalore Cities Air Pollution using Machine Learning Models
Word Count:	6723
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Parth Adesh Darekar
Date:	10th October 2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	Parth Adesh Darekar
Date:	16/08/2021
Penalty Applied (if applicable):	

A Comprehensive Study To Forecast The Delhi And Bangalore Cities Air Pollution Using Machine Learning Models.

Parth Darekar
x19212739@student.ncirl.ie

10th October 2021

Abstract

Air pollution is now a significant research area of topic in recent years, due to its detrimental implications. In today's environment, it is also recognized as one of the primary danger factors. In the existence of air pollution management systems, the first step is accurate air quality measurement, which contributes in the economic and social development of industrialized countries. For both systematic emissions management and public health and well-being, accurate air quality forecasting is critical. Delhi, India's capital, as well as Bangalore one of Information Technology hub of the country has been the world's most polluted city for the past two years.Different time series models, such as ARIMA (Autoregressive Integrated Moving Average) and SES (Simple Exponential Smoothing), have been effectively utilised in the past to forecast air pollution. Taking into consideration previous study as well as to anticipate on the future problems regarding air pollution, different time series models like SARIMA (Seasonal Autoregressive Integrated Moving Average) VAR (Vector Auto Regressive) VARMA (Vector Auto-Regressive Moving Average), ARFIMA(Auto Regressive Fractionally Integrated Moving Average), with the help of neural network model LSTM, have been used in this research study to forecast the Bangalore and Delhi cities air pollutants. This models can discover underlying trends, time series data analysis as well as can assist us in dealing with the problems regarding air pollution.

1 Introduction

Air pollution is a topic that many people are concerned about these days since it has a variety of harmful consequences on the environment and the economy around the world. It is described as the presence of one or more pollutants in outdoor or indoor air for an extended period of time that may harm the human, plant, or animal life, or unexpectedly interferes with normal life or property. Air pollution is a major issue in Asian countries. According to estimates, air pollution causes roughly 537,000 premature deaths in Asian countries. Even though the fact that air pollution is more prevalent in cities, the people who suffer the most are those who are poor or live in locations with poor air quality (Haq and Schwela; 2008). This research study has opted for Delhi and Bangalore city, from which Bangalore is one of India's recognized IT hub centers in India, as well as a

developed metropolitan city of the country. Whereas Delhi is the capital city of India which has other metropolitan cities in India, but due to a large number of industries and population, the city is being named as one of the air polluted cities in India since 2015.

1.1 Motivation and Background

For several years, air pollution has been a major worry in India. Air pollution has risen dramatically as a result of rapid development. Health-related problems such as stroke, heart disease, and lung cancer, to mention a few, have increased as a result of increased air pollution.(Haq and Schwela; 2008). Many of the pollutant measurement equipment has been deployed by the Indian government to monitor air pollution, particularly in heavily polluted cities such as Delhi, Bangalore, Gujarat, and Agra. As a result of increased air pollution in different areas of the country,health-related issues such as stroke, heart disease, and lung cancer diseases, have increased amongst people. Harmful air pollutants like particulate matter (PM2.5 and PM10), carbon monoxide (CO), Ozone (O3), nitrogen dioxides (NO2), and sulfur dioxide(S02) are the six most common air pollutants, which are usually found in India(Chaudhary et al.; 2018). For ensuring and making track of these pollutants the Indian government has placed pollutant monitoring sensors at several stations covering important pollution-prone locations to measure the rising pollution trend in the country. Also, the government has taken several initiatives to reduce pollution, including building a metro system, increasing public transportation, and enacting legislation such as the even-odd system for personal vehicles. The Air Quality Index for PM10 and Xylene was around 25 g/m3 in 2007, according to the Environmental Protection Agency, which utilizes sensory assessments of air pollutants in India, but it soared to 225 g/m3 in 2020.Hence this research work will deeply undergo the pollution level emission that will help in predicting the future air pollution concentration levels in Delhi and Bangalore city.

As shown in Figure 1, there are various key elements that contribute to air pollution. These variables can be divided into two categories namely primary factors and secondary ones. The key determinants are air pollutants such as solid particles, coal combustion, traffic volumes, and manufacturing emissions. Each of these sources has a unique spatial and temporal distribution. The secondary components, on the other hand, are mostly made up of meteorological data, topography, and time. People are becoming more interested in predicting future air quality because they can take greater precautions to avoid being ill if they know the air quality ahead of time. However, air quality prediction is a difficult undertaking, and improving prediction accuracy while lowering training time is a pressing and difficult topic in the field of air pollution.

In this project, we have chosen Delhi and Bangalore cities for predicting the air quality index, which is our research domain for this project. The dataset for this project is generated and described by the CPCB which was downloaded from the Kaggle website, the dataset consists of hour wise data for the pollutants like SO2, NO2, and PM2.5 from the year 2015-2020. The main focus of this study involves comparing the different machine learning models to forecast and predict the air quality index of the above-prescribed cities. Therefore the study is being showcased into different sections which will specify better understanding and delivery of the project.

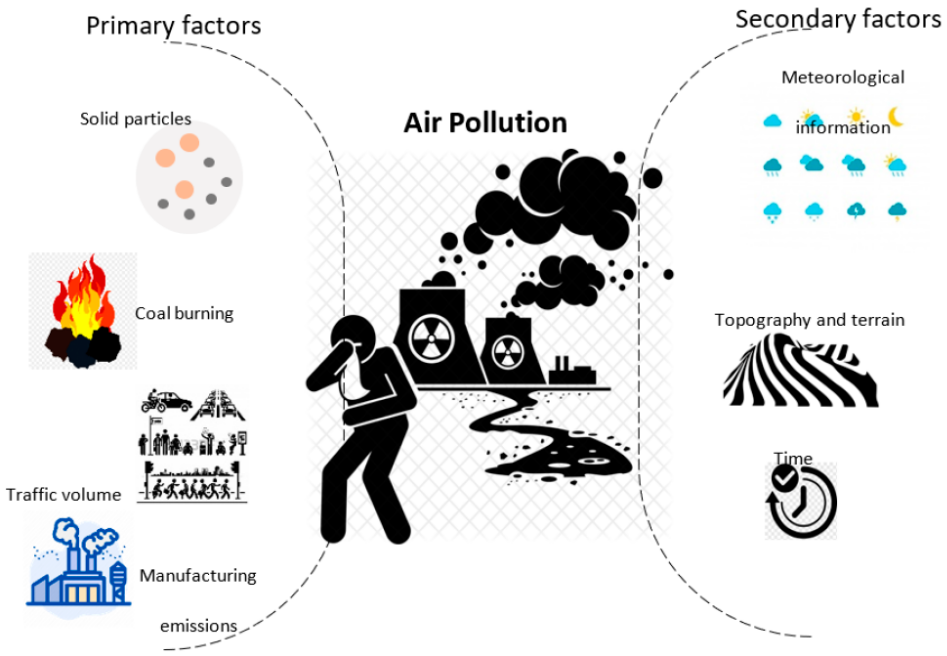


Figure 1: Air Pollution Factors

1.2 Objective and Contributions

Environmental health concerns and pollution have increased as a result of industrialization and urbanization, particularly in emerging nations like India. Due to its huge influence on human life, rapid increases in air pollution are a major source of concern. The study mainly focuses on the diseases which are caused due to air pollution like lung cancer, heart diseases, strokes, respiratory diseases. The study discussed in this paper focuses on forecasting air pollution concentrations and the Air Quality Index (AQI) in major Indian cities so that people are informed of pollution trends ahead of time. Previous research has revealed that air pollution is widespread, and the difficulty of estimating pollutant concentrations for the sensor or non-sensor locations has piqued the interest of researchers. The project's main contribution will be to develop a comprehensive and reliable emissions forecasting model, which will benefit both the environment and people's health. For this, we will be using two novel time series models namely ARFIMA and VARMAX model which will be compared with other models for predicting the Air quality index of Delhi and Bangalore city.

The initial part of the project is Section1 namely the introduction, which is further divided into two sections,i.e Motivation, and Background. This section explains the basic idea about the topic, why we choose the topic and the different causes of the topic. Then comes Section2 the Related work, which describes the previous work done by the authors related to this topic. The previous work prescribed by the authors describes different machine learning models used for predicting the air quality index. In Section3, the methodology used for the project is explained in detail.Section4, i.e implementation part which is the crucial part of the project.The implementation part consists of different machine learning models being performed on the dataset. The implementation results and evaluation is being described in Section5. Finally, the research project conclusion and future work is explained in Section6.

2 Related Work

Human existence depends on good air quality, which is deteriorating in many places throughout the world. Predicting the magnitude and rate of that depreciation has been and continues to be difficult. Planners and scientists will be able to enhance the outcomes for everyone if they have a better knowledge of the data and can make better forecasts.It is possible to employ a variety of predictive approaches and modeling to gain a better understanding of how air quality degrades over time and the potential impact of other man-made or natural factors.This section discusses a summary of the research work done with regards to air pollution to the current date,as well as the approaches utilized to enhance air quality prediction in three different sections which are showcased below:-

2.1 Review of studies on air pollution and its negative consequences on Humans

Multiple studies have been discussed by the authors in this area of study.Air pollution is one of the important aspect where as well a key factor in terms of long term diseases.As these air pollutants are so dangerous which can cause respiratory problems for the humans.Usually some of the diseases related to skin and eyes are caused due to air pollution,as well as some short term damages which are triggered through air pollution can cause severe death in human beings(Kunzli et al.; 2005).According to (Luo et al.; 2019) PM2.5 is the deadliest pollutant which hampers the lungs of the human beings,as well this particulate matter causes diseases like asthma and chest pain in the humans.The respiratory system is the primary route by which air contaminants enter the human body. Ozone, NO, and SO2, as well as fine particulate matter and dust, can irritate mucous membranes. These irritate the eyes, inflame the mouth and throat, impair lung function, and weaken the immune system, all of which contribute to the development of respiratory disorders.The environmental experts basically suggest that if the AQI is higher,the air pollution rate becomes higher,which eventually causes the diseases like headaches,pounding of heart and giddiness feelings (SIERRA-VARGAS and Teran; 2012).

Air pollution can be found at all levels,from the individual to the world.There are two types of consequences from atmospheric air pollution such as local effects and global effects.While local outcomes have an impact on human health,vegetation,raw materials,and cultural assets, whereas global outcomes which usually leads to greenhouse gas emissions,climate change,and ozone depletion in the atmosphere or stratosphere.Air pollution is a major worry for Indians, particularly in Delhi and Bangalore states as they are established with industries as well as the increase of vehicles day by day.According to a WHO (World Health Organization) report,the mortality report from air pollution is higher than the number of people who die from AIDS.Similarly in urban areas the air pollution rate is higher which usually affect the new born kid,kids below six and the old age people as they intake high amount of oxygen,as they have high breathing capacity.The people performing morning exercises in terms of Yoga are also at high risk if these harmful pollutant's gets penetrated into their body.According to one of the articles prescribed in the geographical researches natural disasters can act like agitator for spreading air pollution.For example one of the incident which burnt the Catastrophic forest,due to sudden emission of fire in Canada in the year 2002.This incident highly increased the rate of PM2.5 pollutants in the atmosphere.Hence natural disasters like dust storm,forest fire burning can contaminate the air and affect the environment.As a result,in the fight

against air pollution, research into pollutants like PM2.5,SO2,NO2 concentrations are taken into consideration for the future concentration predictions,as this will be beneficial for predicting future air quality and, ultimately, saving lives of the humans beings.

2.2 Examining the existing methods and techniques for Predicting Air Pollution using Machine Learning and Time series Approaches

In the field of intelligent computation, machine learning is a prominent sub-field. Its major goal is to extract information using computational approaches. Machine learning methods are widely utilized in environmental sciences for data processing, model emulation, climate prediction, AQI forecasting, oceanographic, and hydro logical forecasting(Peng; 2015).

Multiple authors have prescribed the study of predicting the air pollution in different ways.There are three types of AQI predictions the first one is,a simple empirical strategy that forecasts tomorrow's values based on today's data or relies solely on the dependency between projected pollutants and air pollution factors.Second one is Second one is the, physically-based techniques produce skewed projections because they are too complicated to be described by physically-based models.Finally, the third approach justifies the statistical techniques that are parametric or non-parametric,such as neural networks, outperform physically-based methods in terms of accuracy(Hajek et al.; 2015).Usually timeseries data for predicting the Air quality index have multiples 3 main important factors which are mentioned below:-

- Long-term increase or decreases in AQI levels are should be showcased in a trend or sequence manner.
- Seasonal AQI values should highlight elements such as the quarter of the year,the month, or the days of the week, as well as to compare AQI values.
- To accentuate the AQI's irregular rise and fall, make it cyclic. Smoothing is a technique for removing irregular roughness or noise from data so that AQI patterns may be seen more clearly..

Salcedo et al. (1999) invented a new time series method for studying air pollution in the Oporto area in 1999 ((PR, MT, and ML Sampling sites). A long-term trend was discovered, as well as cyclical and periodic elements. They examined daily amounts of strong acidity (SA) and black smoke using a stepwise approach known as the SATSA Model (BS). (Lin et al.; 2011) presented a study in which the author employed support vector machine regression models to predict China's air quality index to verify the air pollutants.For this the SVRLIA model with the help of logarithmic procedures where applied on the dataset.The SVRLIA model predicted accurate mixture of particulate matters with the help of logarithmic procedure.By determining the working of SVRLIA model,the model accurately predicts the concentrations such as particulate matters and nitrogen oxide/dioxide respectively.Hence from the future prospective this model can be used for real time data analysis for forecasting some scientific area of study.

According to (Freeman et al.; 2018) proposed a model Forecasting air quality time series using deep learning. Where they have used the deep learning techniques to predict air pollution time series. As we are familiar with the fact that air quality majorly relies

on time series data. And at the core of deep learning techniques we have neural networks which will find the underlying pattern of the data, and so to work with the time series data. In this paper they have used Long Short Term Memory (LSTM) to predict the 8hr averaged surface ozone (O3). They have reduced to 5 features from 25 features to train the LSTM model, which gave an improved accuracy which is been measured using Mean Absolute Error (MAE).For predictions out to 72 hours the MAE's were calculated less than 2

One of the studies prescribed by (Zhao et al.; 2018) used generalized models namely linear regression and logistic regression to predict and classify the air quality data of Taiwan.But the authors where unable to achieve the success which they wanted to achieve in forecasting the air quality data of Taiwan.As the models used by the authors where unable to handle the crucial data required for predicting the AQI,such as the overall patterns of the data which was in weakly format,outliers and the seasonality of the data.

According to the authors(Papacharalampous et al.; 2018) used time series data to predict the temperature and the air quality index of Greece country.For this study they proposed different methods which comprised of machine learning algorithms like Support vector machine and Neural network for predicting the air quality index.The further results from the SVM model were better as compared to other model in terms of its performance and accuracy.Finally, they came to the conclusion that, based on their score, there is no association between the time series parameters and forecast quality.Other factors used to infer AQI at non-sensor locations include meteorological data, traffic movement, human mobility, and point of interests (POIs).

One of the study prescribed by (Samal et al.; 2019)decribes and justifies their investigation into the role that fossil fuels have played in the depletion of our atmosphere in recent decades, these researchers also looked into how different types of vehicle pollution have contributed to the problem. Solution: time-series forecasting using SiRIMs and the Raphet Model Following a thorough and confidential interview procedure, these models are being analysed to see if they can be used to provide a rough prediction of pollution levels in the future, which is now underway. In order to deal with time-dependent data, linear regression techniques alone are insufficient; therefore, a different strategy has been developed. The researchers have taken this into consideration and used a time collection forecasting approach in a presumed secret interview in order to forecast future levels of poll respondents. In the Indian city of Bhubaneswar, an experimental study was carried out to test a proposed approach for predicting air pollution levels. Seasonality and holiday impacts are taken into account by time gathering forecasting, which is a systematised technique based on additive models that attempts to account for non-linear fluctuations such as seasonality and holiday influences.

(Kong et al.; 2021) have proposed an excellent model for predicting air pollution using a multivariate time series methods which consist of dynamic transfer models.Many studies and methods for time-series forecasting have been published to far, including deep-learning algorithms based on neural networks being utilized to forecast time series data. However, maximum model which we have seen and some which are mentioned here are where there are few investigations of real-time prediction of dynamic huge multivariate data, and the model is based on a stationary state. The researchers present a real-time prediction methodology for multivariate time-series data based on an ensemble method in this work.In terms of performance, the suggested approach can select multivariate time-series variables, which means it can take into account many features and forecast the forecast, and they also came with much-needed real-time adaptable auto regressive

models.As a result, the researchers tested the suggested model with simulated data and used it to forecast air quality measured by five sensors as well as server failures based on real-time performance log data.They discovered that their proposed strategy for predicting air pollution was effective and stable in both short and long-term prediction tests. Furthermore, they used conventional methods for abnormality detection and decided to focus on the current status of objects as either normal or abnormal based on provided data, protectively predict expected statuses of objects with provided real-time data, and implemented this in effective system management in cloud environments using the proposed method. In such particular system they worked with six different traditional time series algorithms namely ARIMA, ARIMAX, ANN (artificial Neural Network) ANNX (Artificial Neural Network with extra predictor) RNN (Recurrent Neural Network), VAR (Vector Auto Regressive) and also they have proposed two methods which are EDT- w and EDT-r. EDT is nothing but the Ensemble Dynamic Transfer Model. And based on the simulation dataset on which they have implemented the above mentioned methods and algorithms, we observe that the EDT-r w gave us pretty good RMSE score which is 0.4405 on target time t+12 which is pretty much better than the other algorithms. The above scores were obtained on the simulation 1 dataset which we extracted using VAR.sim function from package Multivar in R, which aims to generate VAR based model

2.3 Examining the existing methods and techniques for Predicting Air Pollution using Machine Learning and Deep Learning Approaches

Poor air quality is thought to be responsible for 6.5 million deaths worldwide each year. The presence of particulate materials, or minute particles in the air, affects air quality (PM). Particulate matter can occur in the air in a variety of particle sizes.One of the studies presented by the author (Zheng et al.; 2013) in Bejing to address the issue of air quality inference with the help of nearby sensory stations.Whereas in our research we are using the previous hour wise data to predict the air pollutants which are causing the environmental damage.In their research, (Zheng et al.; 2013) constructed a co-training based semi-supervised learning strategy that comprises of two distinct classifiers,a spatial classifier based on Artificial Neural Networks (ANN) and a temporal classifier based on Conditional Random Field (CRF).The author used five different sources of data to predict the the AQI. Similarly our research will be using stacked LSTM by taking previous time series data of sensory stations of Delhi and Bangalore cities as an input and analyzing and predicting the future output with the help of stacked LSTM and RNN(Recurrent neural Network) model.

Many researchers have employed pruned Neural Network and Lazy Learning algorithms to estimate and forecast the air quality level of various countries in recent years.Corani, one of the researchers, applied this technique to calculate Romania's AQI.The sudy comprised of 4years historical data,with this historical data the author developed a forecast model,based on the seasonality analysis of the data.While performing the seasonality analysis,it was observed that large amount of PM10 pollutants were observed in the city during the winters seasons rather than summers.The author investigated the accuracy of feed-forward neural networks against pruned neural networks as well as analyzed the slow learning results. Due to rise in the number of parameters t he FFNN model was over-fitted also the seasonality of the data started raising the number of parameters. According to (Freeman et al.; 2018) deep learning has been offered as a method for pre-

dicting air quality time series, according to the researchers. The use of neural network algorithms to make predictions about time series is common practise. It is well-known that air quality measurements are highly dependent on time series data in order to assess the quality of the air. Many of these techniques rely on neural networks, which may be used to discover the underlying pattern in data and then operate on that pattern. In this work, the LSTM model was employed to forecast the average surf ozone during an eight-hour period (3). A reduction in the number of features in the LSTM model resulted in better accuracy as measured by the Mean Absolute Error.

Below table describes the important literature review which will be beneficial as well as helpful for our research study.

Table 1: Summary details of the related work for air pollution detection using machine learning approach

Author(s)	Evaluation Paramter	Models Used	Factors Applied	Problems Addressed
Zheng et al. (2013)	MAE	LSTM(long short term memory) FFNN (Feed-Forward-Neural Network	O3 surface Ozone	By working on the O3 feature for a variety of average times determined for 03., the LSTM neural network model will assist in real time and present a clear image of how powerful the neural network is.
Boyadzhiev (2014)	K-S test,KMO,Bartlett's test	ARIMA,TBATS factor analysis, Box-Jenkins	NO2,NOx,PM10,SO2 and ground level O3	They highlighted the Box- Jenkins methodology in this post, which helps to illustrate the pollution concentration in the air, which is one of the distinctive aspects.
Hajek et al. (2015)	RMSE,MAE,Confusion Matrix	XGBoost Regressor, SVM, Decision Tree	PAHs Toxic air pollutants,Toxic metals,Urban aerosols	Obtaining more accurate one-day air quality index predictions and comparing the results in order to provide various recommendations to micro-regional government management
Zhao et al. (2018)	Predicting the seasonality and feasibility of the model	Time series Prophet Model for Seasonality of the data	Particulate Matter PM2.5	The study examines the seasonality of daily PM2.5 concentrations measured at 220 monitoring stations across the United States over a nine-year period (2007–2015) using Prophet, a freshly developed time-series analytic tool.
Freeman et al. (2018)	MAE,Parameter Sensitivity	Decision trees to evaluate input feature,LSTM and RNN	PM10, 03, Benzene	Air supervisors can forecast long-term air pollution concentrations using the methods given in this research by just monitoring essential parameters.
Samal et al. (2019)	KMO ,CI	SARIMA,	RSPM, SO2, NO2, SPM	The prophet model is used to predict air pollution. The advantage of the prophet model is that it can capture seasonal, cyclic, and holiday effects, just like in the stock market.
Kong et al. (2021)	RMSE	rim,rimx,nn(rtifiil neurl netwrk)nnx (rtifiil neurl netwrk with extra predictor)	CO(GT), PT08.S1(CO)	In this work, they present a Dynamic Transfer Model technique that outperforms classical time series analysis in terms of forecasting.

3 Methodology

This section would examine into the project's research framework in depth. It provides a clear intellectual justification for the models used. This segment also goes over each step of the process, starting with data gathering as performing different steps to transform the data for further implementation process.

For our research, specifically, we will be using CRISP-DM(Cross Industry Standard Product) methodology which provides a simple and straight forward approach for analyzing the data and implementing the models on the dataset. This approach follows five hierarchical steps, which will be further implemented on our dataset for predicting the

Air Quality Index of Delhi and Bangalore. The dataset for the study was taken from Kaggle, known as the Air Quality dataset. This dataset is then later pre-processed as per the CRSIP-DM approach.The CRISP-DM approach is showcased in figure(2).After the data is pre-processed the final dataset consists of 435735 rows and 13 columns and in these 13 columns,there consist 5 columns of various particulate which contribute to calculating the Air Quality Index. Those 5 particulate particles are:

- SO2:-Sulfur dioxide. It is a highly reactive gasses of the oxides of sulphur which is obtained from the the burning of sulphur or the items which contains sulfur
- NO2:-Nitrogen Dioxide is obtained in air from the burning of fuel. It is mainly emitted from vehicles like Cars, truck, buses etc.
- RSPM:-RSPM is a fraction of TSPM(Total suspended particulate matter) which is emitted wherever the combustion process takes place and in general this is considered as a particulate matter which has a diameter less than 2.5 micrometers, and is easily inhaled by humans unknowingly.
- SPM:-Suspended particulate matter this is also considered as a particulate matter as these particles are in the range of 2.5 micrometers to 10 micrometers in diameter.These particles are originated from a variety of stationary and mobile sources
- PM25:-It is a particulate matter which is of diameter 2.5 micrometers these particulate matters are originated from various sources like agricultural operation, and different industry processes, construction and also demolition activities, there are many other sources from which they get originated

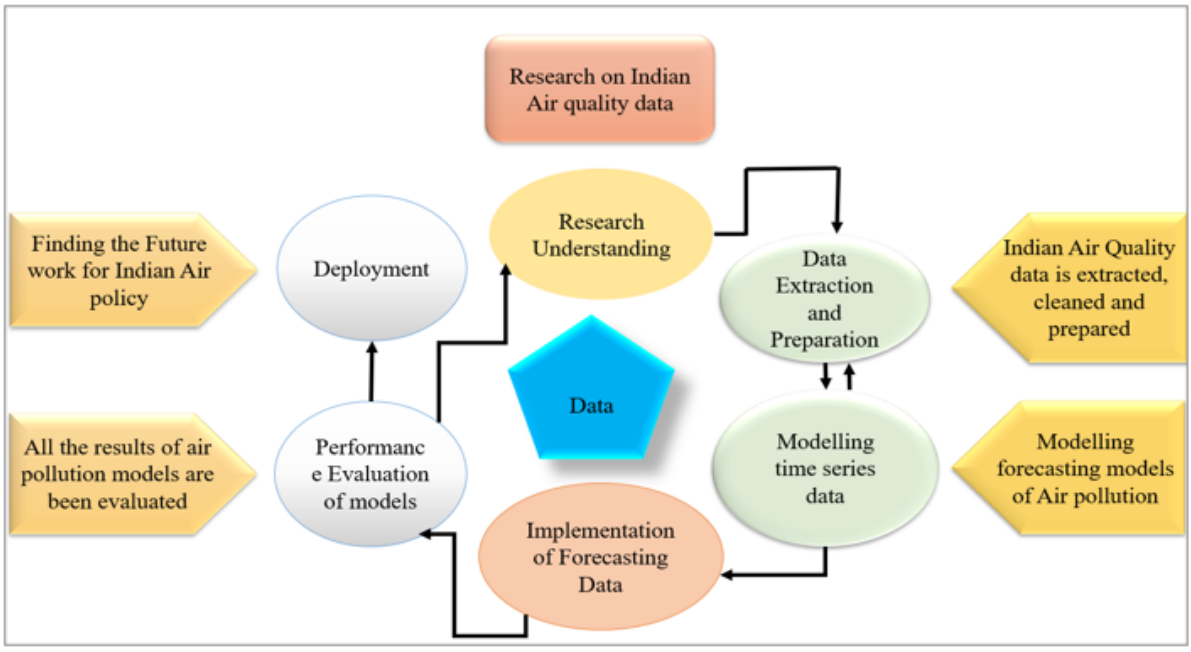


Figure 2: Proposed Methodology for Predicting Air Pollution

The five basic steps of CRISP-DM methodology are explained below:-

3.1 Business Understanding

This is the first and one of the important stages where the scope of the project is decided. The business understanding justifies the main objective of the project. As discussed in the earlier sections, the overall idea and understanding for predicting the air quality index have been taken into consideration. Finally, the methodology for this research work had been searched as well as researched thoroughly before proceeding for the implementation part.

3.2 Data Cleaning

As we will be forecasting the air pollution of only 2 locations which are Delhi and Bangalore. So From the original data, frame we will filter out only the details of Delhi and Bangalore and store them in a new data frame. Now we have a data frame that consists of 15210 rows and 13 columns and coming further we will be converting our date column into date-time format as our date column is an object data type.

3.3 Data Imputation

In the further step, we will be filling all the missing values with the mean of that particular column. As we are having a time series data. So to get a better forecast we cannot neglect the missing values as all the data of every date is important for us to do a better analysis. Feature Selection and Splitting the Data: Now comes the important part of our experiment as we will be using a univariate time-series model. So we need to select one such feature on which we will try to forecast using the time series models. So the one feature which we select is the so2 sulfur dioxide. Based on this feature we will be splitting our data. As the dataset we are using is a time series dataset so we just cannot do the traditional method of randomly splitting our dataset into train and test. We have to split it based on timestamp which is systematical as the earlier date data should be considered as the train set and the latest date of data should be considered as the test set For example,we have to dataset from the year 2015-2020. And this dataset is a time-series dataset similar to the one with which we are working. Suppose we want to split our data on a ratio scale of 80:20 as the 80 datasets should be considered or sent to the train set and the rest 20 as the test set. So the splitting will be done in such a manner like the data from the year 2015-2019 is considered as the train set and the latest timestamp data of the year 2019-2020 is considered as the test set. So this kind of splitting of data is known as Time Series data splitting. Similarly in our dataset, we will split our so2 column data into time series splitting format. Where the starting 80 of data is split into train set and the later 20 of data is considered as the test set. Now we have split our dataset into train and test according to the time series splitting format.

3.4 Feature Selection and Splitting the Data

Now comes the important part of our experiment as we will be using a univariate time-series model. So we need to select one such feature on which we will try to forecast using the time series models. So the one feature which we select is the so2 sulfur dioxide. Based on this feature we will be splitting our data. As the dataset we are using is a time series dataset so we just cannot do the traditional method of randomly splitting our dataset into train and test. We have to split it based on timestamp which is systematical as the

earlier date data should be considered as the train set and the latest date of data should be considered as the test set For example, we have to dataset from the year 2015-2020. And this dataset is a time-series dataset similar to the one with which we are working. Suppose we want to split our data on a ratio scale of 80:20 as the 80 percent dataset should be considered or sent to the train set and the rest 20 percent as the test set. So the splitting will be done in such a manner like the data from year 2015-2019 is considered as the train set and the latest timestamp data of the year 2019-2020 is considered as the test set. So this kind of splitting of data is known as Time Series data splitting. Similarly in our dataset,we will split our so2 column data into time series splitting format. Where the starting 80 percent of data is split into train set and the later 20percent of data is considered as the test set. Now we have split our dataset into train and test according to the time series splitting format.

3.5 Parameter Calibration

Parameter Calibration is a technique where we have to find the appropriate p,q,d values that need to be passed while implementing our univariate time series models like SARIMAX and ARFIMA which we will look further in detail. So the parameter calibration is done using the auto correlation plot and also the partial auto-correlation plot where the lag is set as 10. And also to check whether the time series data we are dealing with is stationary or not, to confirm that we will be implementing the ADF Test (Augmented Dicky- fuller).Let's first define the Dickey-Fuller test before moving on to the ADF test.

$$y_t = c + \beta t + \alpha y_{t-1} + \phi \Delta Y_{t-1} + e_t$$

From the above equation:

- y(t-1) = lg 1 f time series
- delta Y(t-1) = first differene f the series t time (t-1)

It seems to have a null hypothesis which is identical to the unit root test except that the coefficient is 1, stating that a unit root appears. If the sequence or the series is not rejected, it is believed to be non-stationary.If the series is not rejected, it is assumed to be non-stationary.

The ADF test is an enhanced version of the Dickey Fuller test, as the name implies. The ADF test adds to the Dikey-Fuller test statistic to account for slightly elevated regressive processes in the model.

We noticed that the p-value is less than 0.05 and the Dickey-Fuller statistics is smaller than the number of crucial values after running the Dickey-Fuller test, indicating that the time series data we're working with is static. The CRISP-DM approach has indeed been used to meet the project's needs and criteria.

For this research study we have used three-tier architecture to evaluate the project findings.

A three-tier architecture will be implemented due to the significant quantity of data processing that will be done on the data layer.A three-tier architectural methodology is proposed for conformation and review reasons since the research deals with the public

dataset and is being generated by the Central pollution control board. The architecture used for the study is prescribed below:-

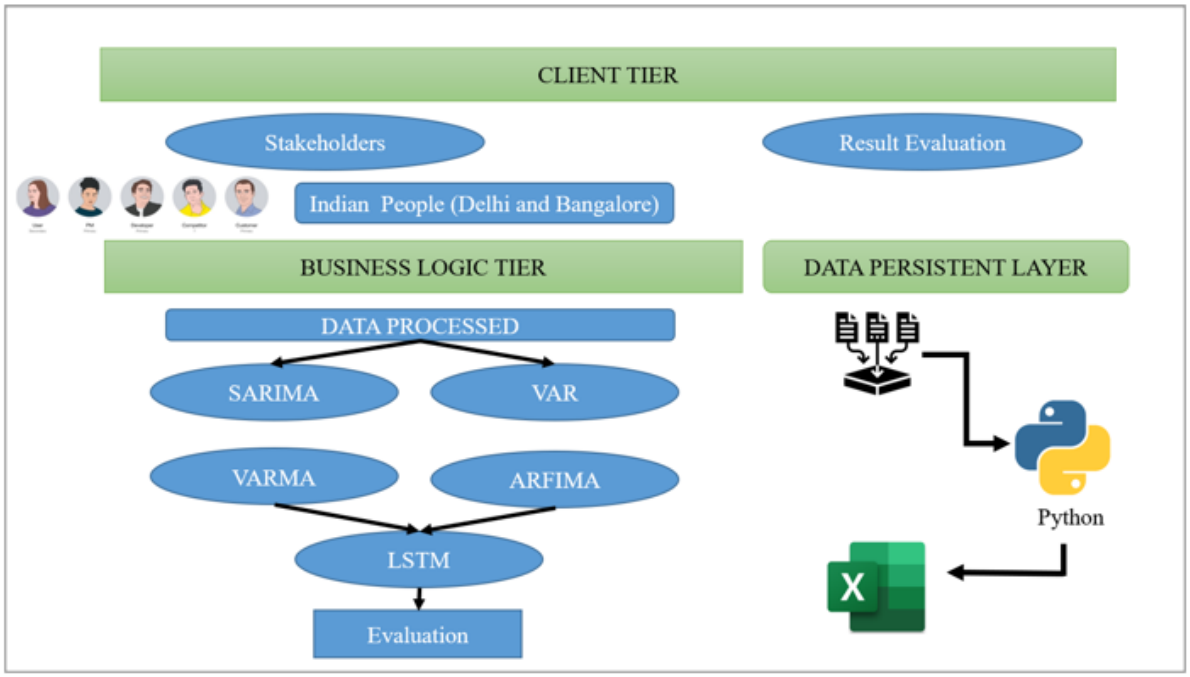


Figure 3: Design Specification

- Client Layer- The user interface layer, commonly referred as Client Layer (Tier 1), is transparent to the client and operates as the user interface.The findings are communicated to stakeholders via visuals in this layer.
- Logic Layer or Tier 2-This layer would incorporate relevant administrative functionality to the dataset, and also required deep learning and time series methods.It also provides the numerous assessment measures that were used to ensure that the findings were consistent.
- Data Layer or Tier3 -The topic-related objectives are represented by this layer. This layer will store both derived datasets. The techniques for cleaning and preprocessing the extracted data are also specified in this layer. It will also show you how to convert raw data into modeling data.

4 Implementation and Evaluation for Forecasting and Predicting the Air Quality Index

The implementation of this research work was done in python with the help of time series models and neural network algorithms like LSTM.After implementing all the models, the results were compared.

4.1 Implementing SARIMA Model

Starting with the SARIMA time series model, which is the equivalent to ARIMAX model. The SARIMAX is among the most intricate models we will have since it will incorporate

seasonality, integration variables. By setting the values of certain bound orders to zero, or by not providing bound data, the model can be simplified. It is a univariate time series model which we will be used to fit and forecast the so2 air pollutant for locations Delhi and Bangalore. Here we use statsmodels.api library to use the SARIMA function and fit the training data. Below is the output after implementing the SARIMA.

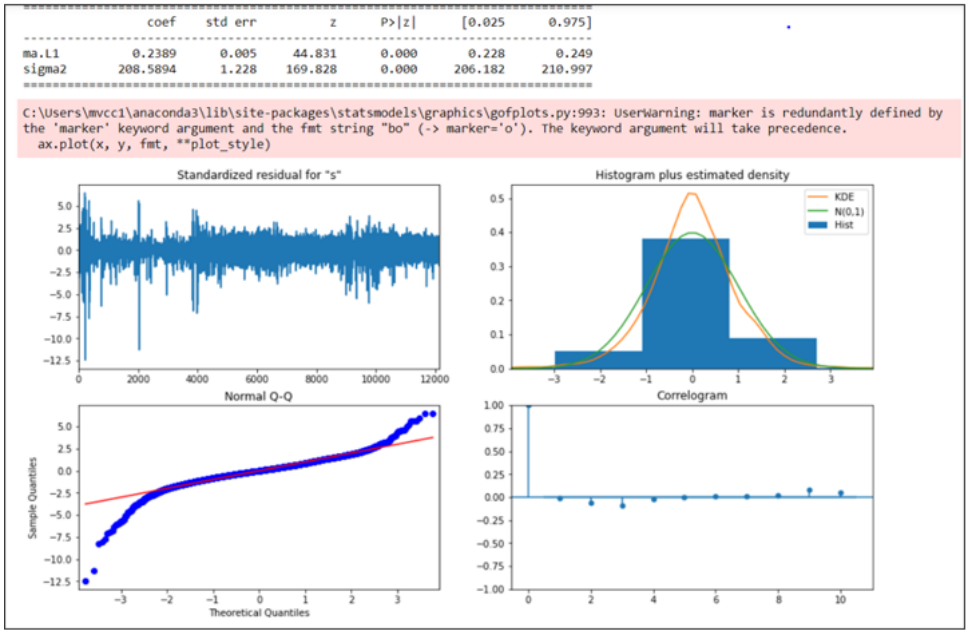


Figure 4: SARIMA Model

From the above output if we notice the Normal Quantile- plot we observe that there are Sample quantities and the theoretical quantities that are aligned properly with a minimal diversion between them. Visualizing the above prediction we can notice that the SARIMA has done a pretty decent job in forecasting the so2 air pollutant for locations Delhi and Bangalore.

4.2 Implementing Vector Auto Regressive

The vetr utregressive (VR) mdel n be nsiderd s the wrkhuse multivrite time collection mdel tht reltes urrent bservtins f vrible with st bservtins f itself nd st bservtins f specific vribles inside the system. VR mdels dissent frm univrite utregressive mdels s result f they ermit feedbk t ur among the vribles inside the mdel. Now we will try to fit the trained data for so2 air pollutants using the time series analysis model VAR (Vector Auto Regression). If we want to implement the vector Auto Regression in python we have to use the same package Statsmodel.We will try to display the result of the VAR output using a graph visualization of an actual and predicted model using the data which we have split. Below is the plot after forecasting the data using the multivariate time series algorithm VAR.

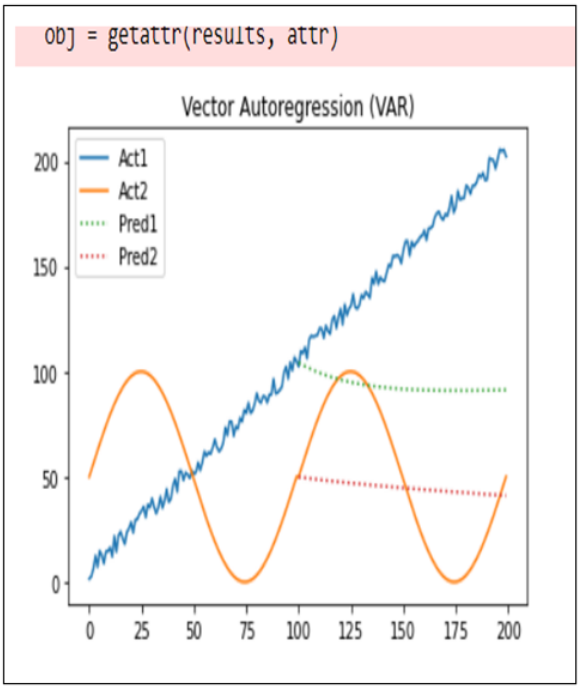


Figure 5: VAR Model

From the Above graph we can observe that the VAR was trying to capture the underlying time series pattern for the so2. We have also forecasted the model using another variant of VAR known as VARMA (Vector Auto regression Moving-Average) similar to the above time series models we have imported this function from the statsmodels package. Here is the final plotting for the forecast of so2 using VARMA

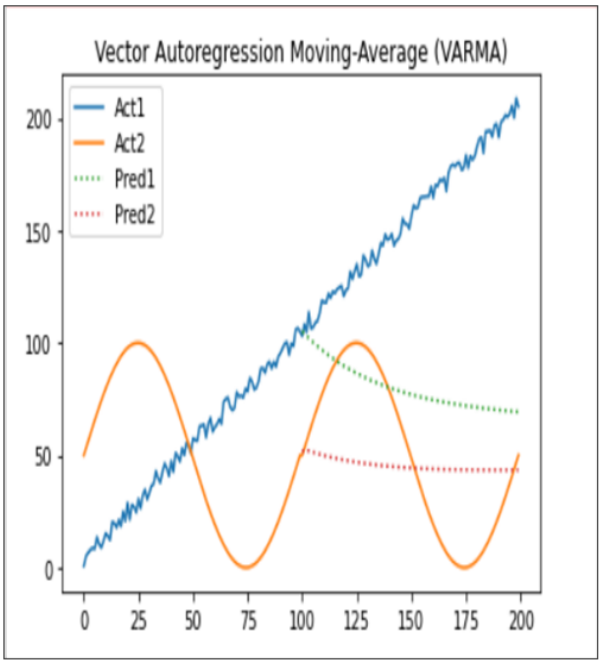


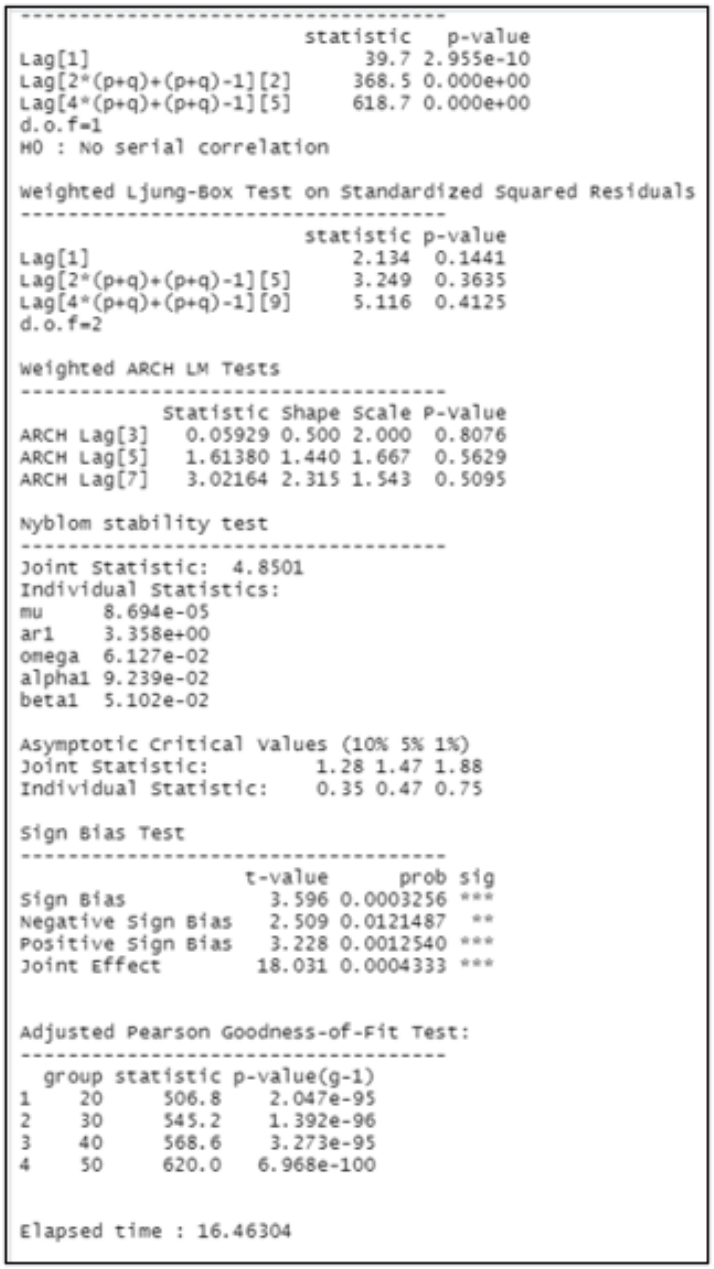
Figure 6: VARMA Model

4.3 Implementing ARFIMA(Auto Regressive Integrated Moving Average)

As,we have seen both the variant of multivariate time-series data. And if we observe according to obtained result SARIMA is pretty decent in forecasting the so2 air pollutant. Now we will be implementing our novel model ARFIMA for the Delhi and Bangalore dataset.Arfima is extended version of ARIMA.An autoregressive moving average process (with p, d, and q) that has been dynamically integrated is denoted by the operator's notation RFIM (p, d, q).Usually the model tries to capture the long-memory processes in the dataset which is provided.The ARFIMA model is present inside the Rugarch library which needs to be installed in python as well as R studio in order to leverage the power of ARFIMA model.After importing the packages we we will define the ARFIMA model using the function ugarchspec inside the ugarchspec we pass the parameters like arfima-T and fixed.pars-list(arfima-1),which in turn means that we are trying to forecast the data using the ARFIMA time series analysis method.And then we fit our data using ARFIMA and here is the result which we have obtained from fitting the so2 univariate variable in ARFIMA.

```
*---------------------------------*
*          GARCH Model Fit        *
*---------------------------------*

Conditional Variance Dynamics
-----------------------------------
GARCH Model     : sGARCH(1,1)
Mean Model      : ARFIMA(1,d,0)
Distribution    : norm

Optimal Parameters
------------------------------------
        Estimate  Std. Error   t value Pr(>|t|)
mu      6.297010    7.919500   0.79513  0.42654
ar1    -0.377230    0.013276 -28.41399  0.00000
arfima  1.000000          NA        NA       NA
omega   1.123942    0.195624   5.74541  0.00000
alpha1  0.083278    0.008190  10.16822  0.00000
beta1   0.903503    0.009595  94.16709  0.00000

Robust Standard Errors:
        Estimate  Std. Error   t value Pr(>|t|)
mu      6.297010   28.863370   0.21817 0.827300
ar1    -0.377230    0.013907 -27.12527 0.000000
arfima  1.000000          NA        NA       NA
omega   1.123942    0.510384   2.20215 0.027655
alpha1  0.083278    0.021390   3.89330 0.000099
beta1   0.903503    0.026455  34.15301 0.000000

LogLikelihood : -21166.87

Information Criteria
------------------------------------

Akaike       6.8385
Bayes        6.8439
Shibata      6.8385
Hannan-Quinn 6.8403

Weighted Ljung-Box Test on Standardized Residuals
------------------------------------
                        statistic   p-value
Lag[1]                       39.7 2.955e-10
Lag[2*(p+q)+(p+q)-1][2]     368.5 0.000e+00
Lag[4*(p+q)+(p+q)-1][5]     618.7 0.000e+00
d.o.f=1
H0 : No serial correlation

Weighted Ljung-Box Test on Standardized Squared Residuals
------------------------------------
                        statistic p-value
Lag[1]                      2.134  0.1441
Lag[2*(p+q)+(p+q)-1][5]     3.249  0.3635
Lag[4*(p+q)+(p+q)-1][9]     5.116  0.4125
```

Figure 7: ARFIMA Model



```
------------------------------------
                        statistic   p-value
Lag[1]                       39.7 2.955e-10
Lag[2*(p+q)+(p+q)-1][2]     368.5 0.000e+00
Lag[4*(p+q)+(p+q)-1][5]     618.7 0.000e+00
d.o.f=1
H0 : No serial correlation

Weighted Ljung-Box Test on Standardized Squared Residuals
------------------------------------
                        statistic p-value
Lag[1]                      2.134  0.1441
Lag[2*(p+q)+(p+q)-1][5]     3.249  0.3635
Lag[4*(p+q)+(p+q)-1][9]     5.116  0.4125
d.o.f=2

Weighted ARCH LM Tests
------------------------------------
            Statistic Shape Scale P-Value
ARCH Lag[3]   0.05929 0.500 2.000  0.8076
ARCH Lag[5]   1.61380 1.440 1.667  0.5629
ARCH Lag[7]   3.02164 2.315 1.543  0.5095

Nyblom stability test
------------------------------------
Joint Statistic:  4.8501
Individual Statistics:
mu     8.694e-05
ar1    3.358e+00
omega  6.127e-02
alpha1 9.239e-02
beta1  5.102e-02

Asymptotic Critical Values (10% 5% 1%)
Joint Statistic:         1.28 1.47 1.88
Individual Statistic:    0.35 0.47 0.75

Sign Bias Test
------------------------------------
                   t-value      prob sig
Sign Bias            3.596 0.0003256 ***
Negative Sign Bias   2.509 0.0121487  **
Positive Sign Bias   3.228 0.0012540 ***
Joint Effect        18.031 0.0004333 ***


Adjusted Pearson Goodness-of-Fit Test:
------------------------------------
  group statistic p-value(g-1)
1    20     506.8    2.047e-95
2    30     545.2    1.392e-96
3    40     568.6    3.273e-95
4    50     620.0   6.968e-100


Elapsed time : 16.46304
```

Figure 8: ARFIMA Model

From the above figure we can notice that in the optimal parameter section the t-value is pretty high than the statistic.This indicates that the model has performed well in certain circumstances.As our data is stationary type the ARFIMA model also has done a decent work.

4.4 Implementing LSTM Model(Long Short Term Memory)

LSTM is a sort of RNN model that contains memory and considers both current and prior data as input.As a result, the LSTM model's input at time t equals the model output at time t-1 plus additional input at time t.We are using this model to estimate future pollution concentrations,which is heavily dependent on prior pollutant concentrations,and air quality data.While performing the LSTM model we will take into consideration the

all the air pollutants that is SO2,NO2,SPM,RSPM,and PM2.5 as our feature and for the label part we will be sending the so2 data into label.In our dataset,we will split the data by considering the past 1 day data of the above 5 features and will try to predict the 2nd day so2 air pollutant volume.While the data was being divided into two sub parts train and test,we will be having 14449 data in the training set and 761 data points in the testing set.Based on the splitting the LSTM architecture will be constructed,which consist of 3 hidden layers,and in each layer there are 60 hidden neurons,whereas output layer consist of only 1 neuron,since our output is predicting only one outcome.We then compile the model using Adam optimizer and the metrics here used is Mean Absolute Error.And we train the model on 20 epochs with the batch size of 32.

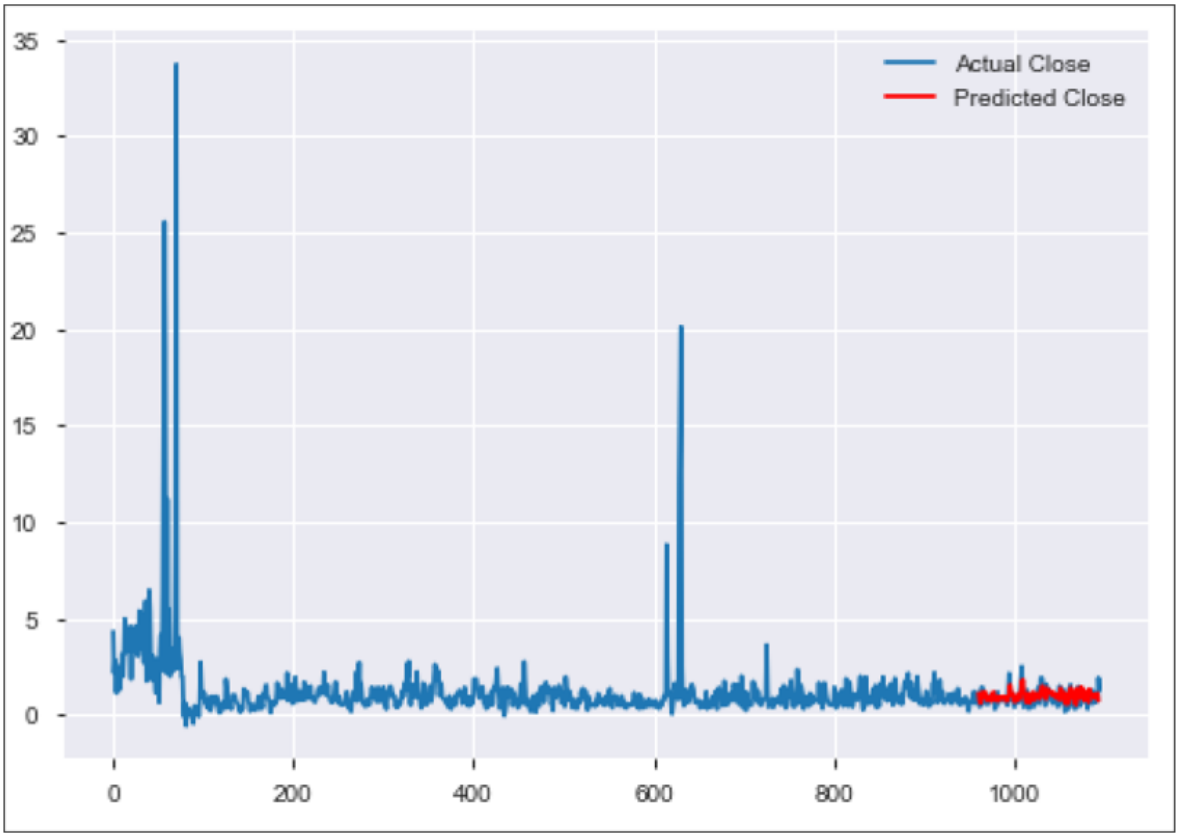


Figure 9: LSTM

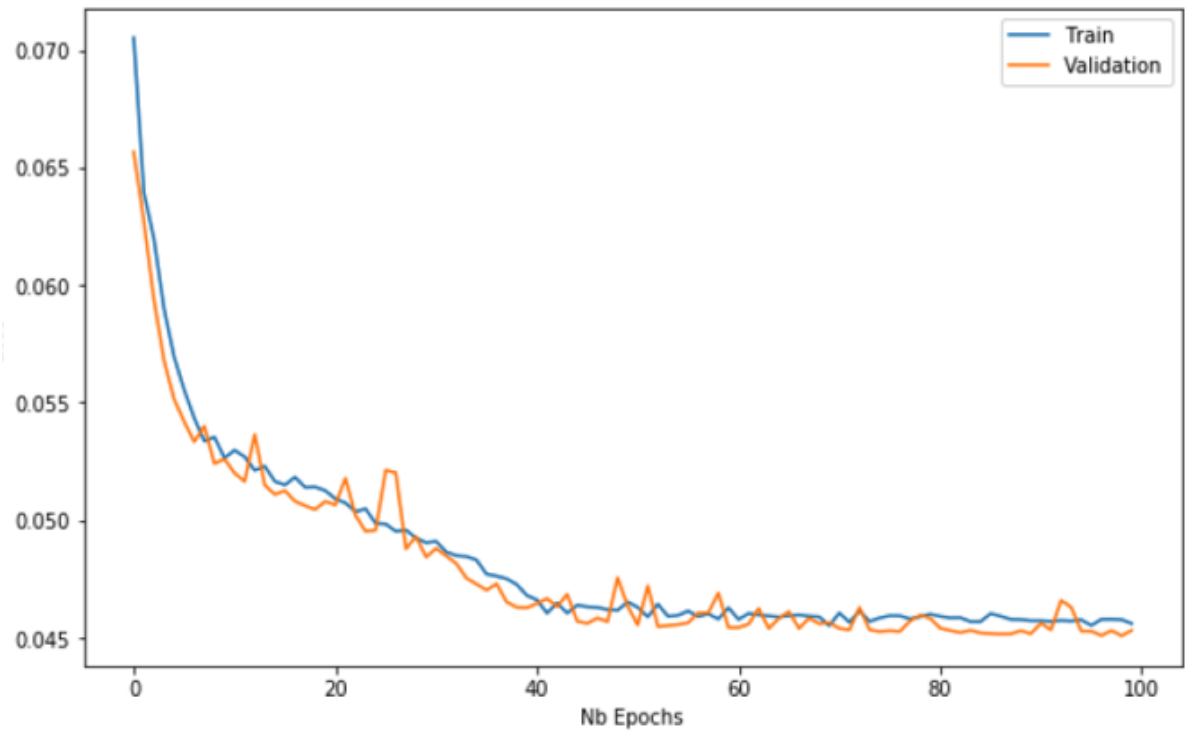


Figure 10: LSTM

```
Epoch 20/20
452/452 [==============================] - 2s 3ms/step - loss: 0.7671 - mae: 0.6664
Wall time: 33.2 s
```

Figure 11: LSTM

The predicted and observed values are overlapping, as seen in figure (9). SARIMA was able to achieve minimal forecasting errors when compared to other models employed, such as VAR,VARMA,ARFIMA,and LSTM, as it recalls past values in order to forecast future ones.The loss function value has decreased across the training cycles, showing that the model can predict general outcomes, as seen in figure(10). As a result, we may indicate that the model is resistant to over-fitting and will function well when dealing with various datasets. Whereas from figure(11) the MAE value for 20 epochs displayed 66percent,which specifies that model performed well as compared to other models.

4.5 Evaluation and Discussion

It's crucial to assess the model's performance using the right parameters. Based on,MAE (mean absolute error),and model execution time, we compared the performance of all the applicable time series models in this part, which was influenced by literature.Also we performed the Dickey-Fuller-Test to check whether our data is stationary or not.After performing the Dickey-Fuller-Test,the p-value was less than 0.05 which states that our dataset is stationary and hence we can proceeded with the further implementation process of models.

Table 2: Performance Comparison of Models

Model	MAE	Pollutant	Execution Time(secs)
SARIMA	62	SO2,NO2	13.6
VAR	56	SO2,NO2	7.97
ARFIMA	61	SO2,NO2,PM2.5	148
LSTM	66	SO2,NO2,PM2.5,PM10	78

In the above table we shown the comaprsion of the models.The table clearly specifies that LSTM model performed well as comapared with ther models.The novel models ARFIMA and VAR also produced some good results with the MAE value 61% and 56% respectivelly.As our data was stationary the overall results predicted by the models were accurate.As we used only one input layer with the help of one output layer the LSTM model performed well as compared to the other models,but the execution time taken was more as comapred to other models.But if we will execute more epochs,the accuracy,of the model will increase,as well as the execution time will increase.The overall implemented models performed well on the dataset,so these models in future can be applied on the live timeseries dataset which consist of long memories by taking into the consideration client requirements and approvals.Hence at the end we can conclude that based on the predicted output LSTM and SARIMA performed well,as well as our novel models showcased some good predictions.

5 Conclusion and Future Work

Hence from the above experiments we have come to conclusion that in our case for the univariate time series forecasting of the so2 feature for the location Delhi and Bangalore the SARIMA model has gave pretty good results compared to other time series algorithms.Whereas our novel models ARFIMA,VAR has also done a pretty decent job in terms of predicting the air pollution.Whereas the LSTM score of MAE is not as expected but if try to train the model for more number of epochs then there may be chances of giving us a pretty less loss and capture the necessary time series related information.

Future Work:-

The best and most consistent models could be studied and evaluated on a range of different datasets, to confirm their accuracy and performance evaluation on new data sets.To improve their performance, SARIMA,VAR,ARFIMA could be modified using different parameters.Further we can also work to en-capture the air pollution level for all the location of India as in this experiment we have worked with only 2 locations.

References

Boyadzhiev, T. (2014). Snezhana georgieva gocheva-ilieva, atanas valev ivanov, desislava stoyanova voynikova & doychin, *Stoch Environ Res Risk Assess* **28**: 1045–1060.

Chaudhary, V., Deshbhratar, A., Kumar, V. and Paul, D. (2018). Time series based lstm model to predict air pollutant's concentration for prominent cities in india, *Interna-*

tional Workshop on Utility-Driven Mining. ACM Conference on Knowledge Discovery and Data Mining, pp. 1–9.

Freeman, B. S., Taylor, G., Gharabaghi, B. and Thé, J. (2018). Forecasting air quality time series using deep learning, *Journal of the Air & Waste Management Association* **68**(8): 866–886.

Hajek, P., Olej, V. et al. (2015). Predicting common air quality index-the case of czech microregions, *Aerosol and Air Quality Research* **15**(2): 544–555.

Haq, G. and Schwela, D. (2008). Urban air pollution in asia, *Foundation Course on Air Quality Management in Asia, Stockholm Environment Institute* .

Kong, T., Choi, D., Lee, G. and Lee, K. (2021). Air pollution prediction using an ensemble of dynamic transfer models for multivariate time series, *Sustainability* **13**(3): 1367.

Kunzli, N., Gutschmidt, K., Pope, A., Romieu, I., Samet, J. and Smith, K. (2005). The global burden of disease due to outdoor air pollution, *J Toxicol Environ Health A* **68**(1314): 13011307.

Lin, K.-P., Pai, P.-F. and Yang, S.-L. (2011). Forecasting concentrations of air pollutants by logarithm support vector regression with immune algorithms, *Applied Mathematics and Computation* **217**(12): 5318–5327.

Luo, Z., Huang, J., Hu, K., Li, X. and Zhang, P. (2019). Accuair: Winning solution to air quality prediction for kdd cup 2018, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1842–1850.

Papacharalampous, G., Tyralis, H. and Koutsoyiannis, D. (2018). Univariate time series forecasting of temperature and precipitation with a focus on machine learning algorithms: A multiple-case study from greece, *Water resources management* **32**(15): 5207–5239.

Peng, H. (2015). *Air quality prediction by machine learning methods*, PhD thesis, University of British Columbia.

Salcedo, R., Ferraz, M. A., Alves, C. and Martins, F. (1999). Time-series analysis of air pollution data, *Atmospheric Environment* **33**(15): 2361–2372.

Samal, K. K. R., Babu, K. S., Das, S. K. and Acharaya, A. (2019). Time series based air pollution forecasting using sarima and prophet model, *proceedings of the 2019 international conference on information technology and computer communications*, pp. 80–85.

SIERRA-VARGAS, M. P. and Teran, L. M. (2012). Air pollution: impact and prevention, *Respirology* **17**(7): 1031–1038.

Zhao, N., Liu, Y., Vanos, J. K. and Cao, G. (2018). Day-of-week and seasonal patterns of pm2. 5 concentrations over the united states: Time-series analyses using the prophet procedure, *Atmospheric environment* **192**: 116–127.

Zheng, Y., Liu, F. and Hsieh, H.-P. (2013). U-air: When urban air quality inference meets big data, *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1436–1444.