**PREDICTION OF AIR POLLUTION USING RANDOM FOREST CLASSIFIER**



**An Internship Project Report**

On

**Prediction of Air Pollution using Random Forest**

Submitted in partial fulfilment of the requirements for the award of the degree of

**BACHELOR OF TECHNOLOGY**

**In**

**COMPUTER SCIENCE AND ENGINEERING –**

**ARTIFICIAL INTELLIGENCE & MACHINE LEARNING**

**By**

|  |  |
| --- | --- |
| **G. Devanshi** | **22NN1A4216** |
| **V. Anusha** | **22NN1A4260** |

**N .Anitha 22NN1A4240**

**M. Lahya Priya 22NN1A4227**

**Under the Esteemed Guidance of**

**Ms. P. Silpa Chaitanya**

**Assistant Professor, CSE-AIML**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING –**

**ARTIFICIAL INTELLIGENCE & MACHINE LEARNING**

**VIGNAN’S NIRULA INSTITUTE OF TECHNOLOGY AND SCIENCE FOR**  **WOMEN**

**PEDAPALAKALURU, GUNTUR-522005**

**(Approved by AICTE, NEW DELHI and Affiliated to JNTUK, Kakinada.)**

**2020-2024**

**VIGNAN’S NIRULA INSTITUE OF TECHNOLOGY AND SCIENCE FOR**

**WOMEN**

**(Approved by AICTE, NEW DELHI and Affiliated to JNTUK, Kakinada)**

**PEDAPALAKALURU, GUNTUR-522005**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**



**CERTIFICATE**

This is to certify that the internship project report entitled **“****Prediction of Air Pollution using Random Forest”**, is a bonafide work of **G Devanshi(22NN1A4216), V Anusha(22NN4260), N Anitha(22NN1A4240) and M Lahya Priya(22NN1A4227)** submitted to the faculty of Computer Science And Engineering-Artificial Intelligence & Machine Learning, in the requirements for the award of degree of **BACHELOR OF TECHNOLOGY** in **COMPUTER SCIENCE AND ENGINEERING-ARTIFICIAL INTELLIGENCE & MACHINE LEARNING** from **VIGNAN’S NIRULA INSTITUTE OF TECHNOLOGY AND SCIENCE FOR WOMEN, GUNTUR.**

**Project Guide Head of the Department**

**Ms. P. Silpa Chaitanya Ms. P. Silpa Chaitanya(PhD)**

**Assistant Professor CSE-AIML**

**EXTERNAL EXAMINER**

**DECLARATION**

We hereby declare that the work described in this Internship project report, entitled **“Prediction of Air Pollution using Random Forest”** which is submitted by us for the award of **Bachelor of Technology** in the Department of **Computer Science and Engineering-Artificial Intelligence & Machine Learning** to the **Vignan’s Nirula Institute of Technology and Science for women**, affiliated to Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, is the result of work done by us under the guidance of **Ms. P . Silpa Chaitanya , Assistant Professor, CSE-AIML**.

The work is original and has not been submitted for any Degree/ Diploma of this or any other university.

**Place:**

**Date:**

**G Devanshi (22NN1A4216)**

**V Anusha (22NN1A4260)**

**N Anitha (22NN1A4240)**

**M Lahya Priya (22NN1A4227)**

**ACKNOWLEDGEMENT**

We express our heartfelt gratitude to our beloved principal **Dr. P. Radhika** for giving a chance to study in our esteemed institution and providing us all the required resources.

We would like to thank **Ms. P. Silpa Chaitanya, Assistant Professor, Head of the**

**Department of Computer science and Engineering-Artificial Intelligence & Machine Learning**, for his extended and continuous support, valuable guidance and timely advices in the completion of this project thesis.

We wish to express our profound sense of sincere gratitude to our Project Guide **Ms. P. Silpa Chaitanya Assistant Professor** without whose help, guidance and motivation this project thesis could not have been completed the project successfully.

We also thank all the faculty of the Department of Computer Science and Engineering Artificial Intelligence & Machine Learning for their help and guidance of numerous occasions, which has given us the cogency to build-up adamant aspiration over the completion of our project thesis.

Finally, we thank one and all who directly or indirectly helped us to complete our project thesis successfully.

**PROJECT ASSOCIATES**

|  |  |
| --- | --- |
| **G Devanshi** | **(22NN1A4216)** |
| **V Anusha** | **(22NN1A4260)** |
| **N Anitha** | **(22NN1A4240)** |
| **M Lahya Priya** | **(22NN1A4227)** |

**ABSTRACT**

Air pollution is a growing concern worldwide, and it has serious implications on human health, the environment, and the economy. We present the first comprehensive analysis government air quality observations for PM10, PM2.5, SO2, NO2 and O3 from the Central Pollution Control Board (CPCB). We address inconsistencies and data gaps in datasets using a rigorous procedure to ensure data representativeness. In this project, we explore the prediction of Air Quality Index (AQI) using the Random Forest algorithm. AQI is a measure of air pollution that is used to communicate the health risks associated with breathing polluted air. We use historical data collected from various air quality monitoring stations in a city and apply the Random Forest algorithm to predict AQI. This study aims to predict the AQI using machine learning algorithms. The AQI is a crucial indicator of air quality, and accurate forecasting can help mitigate the negative effects of air pollution on human health and the environment. The study utilizes data from air quality monitoring stations and meteorological sensors to train and evaluate various machine learning models, including Random Forest, Support Vector Regression, Linear Regression, Logistic Regression and Navie bayes. The accuracy of the algorithm is measured using the root mean square error. The mean square error and the mean absolute error). The results indicate that the Random Forest algorithm performs well in predicting AQI and has the potential to be used as a tool to monitor air quality and help in making decisions to reduce air pollution. The findings of this study can be used by policy makers, city planners, and environmental agencies to design effective strategies to combat air pollution.

**INDEX**

**CERTIFICATE PAGE**

**ACKNOWLEDGEMENTS**

**ABSTRACT----------------------------------------------------------------------------------5**

List of figures **---------------------------------------------------------------------------------7**

**Chapter 1: INTRODUCTION**

1.1 Introduction**-------------------------------------------------------------------------------8**

**Chapter 2:LITERATURE SURVEY**

2.1 Literature survey**-------------------------------------------------------------------------10**

**Chapter 3: AIR POLLUTION**

3.1 About Air pollution**----------------------------------------------------------------------27**

3.2 Types of air pollution**--------------------------------------------------------------------27**

3.3 Problems of air pollution**----------------------------------------------------------------29**

3.4 Prevention measures**---------------------------------------------------------------------31**

3.5 Future Scope**------------------------------------------------------------------------------32**

3.6 Proposed Methodology**------------------------------------------------------------------33**

**Chapter 4: Software Environment**

4.1 Introduction to Python **------------------------------------------------------------------37**

4.2 Why Choose Python **---------------------------------------------------------------------37**

4.3 Hardware Requirements**-----------------------------------------------------------------39**

4.4 Software Requirements**------------------------------------------------------------------39**

4.5 Steps to install Python**--------------------------------------------------------------------40**

4.6 Modules

4.6.1 Pandas**---------------------------------------------------------------------------42**

4.6.2 Numpy**--------------------------------------------------------------------------43**

4.6.3 scikit-learn**----------------------------------------------------------------------43**

**Chapter 5:FLASK**

5.1 Flask **---------------------------------------------------------------------------------------45**

5.2 Installation**---------------------------------------------------------------------------------45**

**Chapter 6:Working with Flask**

6.1 Flask work in python**--------------------------------------------------------------------47**

6.2 Template engine**--------------------------------------------------------------------------47**

**Chapter 7: PYTHON CODE**

7.1 View code **--------------------------------------------------------------------------------48**

7.2 Predict **------------------------------------------------------------------------------------50**

7.3 INIT **---------------------------------------------------------------------------------------51**

**Chapter 8: RESULTS AND DISCUSSION**

8.1 Final result **-------------------------------------------------------------------------------54**

**Chapter 9: Conclusion --------------------------------------------------------------------56**

**References**

**LIST OF FIGURES**

**FIGURE 1.1** FLOW CHART

**FIGURE 1.2** FLOWCHART OF TRAINING AND TESTING THE DATASET

**FIGURE 3.1** OZONE LAYER DEPLATION

**FIGURE 3.2** INCREASED ATMOSPHERE

**FIGURE 3.3** POLLUTANT EMISSION

**FIGURE 3.4** INDOOR AIR PROBLEM

**FIGURE 3.5** PERSISTENT ORGANIC POLLUTANTS

**FIGURE 3.6** CARDIOVASCULAR INVOLVEMENT

**FIGURE 3.7** THE COST OF CLIMATE CHANGE

**FIGURE 3.8** ENVIRONMENT DEGRADATION

**FIGURE 3.9** ACCURACY OF LOGISTIC REGRESSION

**FIGURE 3.10** ACCURACY OF NAVIE BAYES

**FIGURE 3.11** ACCURACY OF SVM

**FIGURE 3.12** ACCURACY OF RANDOM FOREST

**FIGURE 8.1** OPEN THE MAIN CODE AND COPY THE IP ADDRESS

**FIGURE 8.2** OUTPUT OF WEBPAGE

**FIGURE 8.3** OUTPUT OF THE MODEL CREATED

CHAPTER 1

INTRODUCTION

**1.1 INTRODUCTION**

In recent years, the issue of air pollution has garnered significant attention due to its detrimental effects on public health and the environment. According to the World Health Organization (WHO), air pollution is responsible for millions of premature deaths annually worldwide, making it a pressing global concern. Traditional methods of monitoring air quality rely on sparse networks of stationary sensors, which provide limited spatial coverage and temporal resolution. However, with the advent of machine learning (ML) techniques and the proliferation of IoT (Internet of Things) devices, there is a growing opportunity to improve the accuracy and efficiency of air pollution prediction.

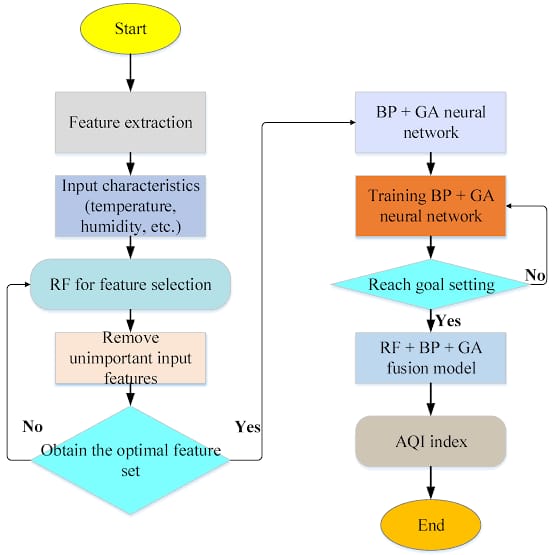
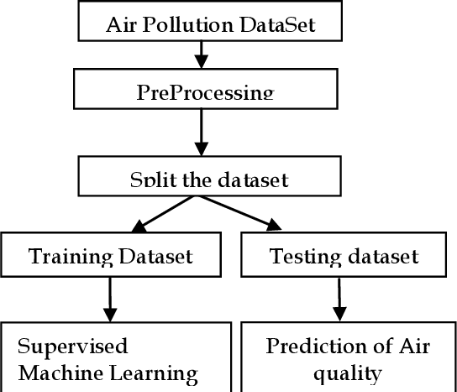
Machine learning offers a promising approach to address the complexities involved in modeling and predicting air quality. By leveraging vast amounts of data from various sources such as meteorological conditions, geographical features, and historical pollution levels, ML algorithms can learn intricate patterns and relationships that influence air quality dynamics. This capability enables more precise forecasting of pollution levels across different spatial and temporal scales, thereby supporting proactive measures for pollution control and public health management. 

Fig: 1.1 Flow chart

This paper explores the application of machine learning techniques in predicting air pollution levels. It begins by reviewing existing methods and challenges in air quality monitoring and prediction. Subsequently, it discusses the potential benefits of ML-based approaches, including improved accuracy, real-time monitoring capabilities, and scalability. Furthermore, the paper examines several ML models commonly used for air quality prediction, highlighting their strengths and limitations.

By integrating advanced data analytics with environmental science, ML-based air pollution prediction represents a pivotal advancement towards sustainable urban development and public health improvement. This research contributes to the growing body of knowledge aimed at harnessing technology to mitigate the adverse impacts of air pollution and promote healthier living environments.

****

**Fig:1.2 Flow chart of train and testing of dataset**

**CHAPTER 2**

**LITERATURE SURVEY**

[1] B.P. Pande came up with a very interesting study about the analysis air pollution and respiratory health. He analyzes the six years of air pollution data from Indian cities, focusing on twelve pollutants and the AQI. The dataset undergoes preprocessing and visualization to uncover patterns and data imbalance is addressed using resampling techniques. Five popular ML models are evaluated and compared using standard performance metrics. In this research XG Boost model achieved the highest accuracy among the Machine Learning models evaluated for predicting AQI values. The Support Vector Machine(SVM) model exhibited the lowest accuracy in comparison to other models in this research. While executing the dataset he knew that PM2.5 ad PM10 exhibit seasonal patterns with increased pollution with winter compared to summer. Additionally, SO2 levels have risen since 2018, while O3 levels remained variations in their concentration.

[2] K. Kumar carried out a detailed literal survey on the impact of air pollution on human health and environment leading to premature deaths and substantial economic losses globally. It emphasizes the importance of ground monitoring and remote sensing technologies in understanding and addressing air pollution dynamicas. In this research he used cluster analysis . Cluseter1 focuses on applying ml models to identify specific signals in software and VOCs. ML models have been utilized for monitoring various pollutants such as VOCs, PM2..5, CO and NO2 with Artificial Neural Network(ANN) being the most commonly used algorithm. The text also mentions the challenges of single sensor detection and the introduction of ML algorithms in data processing modules for calibration and enhanced detection efficiency, enabling rapid identification of gases in gas mixture bases on signal values.

[3] Aditya Kumar Agarwall and his team researched on the air pollutants in the laboratories of an Indian engineering institute. In their research they knew that the global concern of air pollution, particularly indoor air pollution in developing countries like India. It highlights the lack of guidelines for indoor Air Quality in India and emphasizes the need for continuous monitoring due to its negative impact on human health. This is focus on outdoor air pollution and the adverse effects of IAP on human health, economic growth and living habits. Here they used cluster analysis for indoor air pollutions include K-means, DBSCAN, Affinity Propagation and disadvantages with K-means being population due to its fast implementation. This research identify outlier concentration in PM and CO2 levels. It mentions the impact of surprise activities like renovation work and construction on air quality. Additionally, it discusses the correlation between different sized PM concentration in various laboratories, including distinct emission sources..

[4] S.C.Lee discusses the critical issue of air pollution in India, highlighting that over 660 million people are affected by poor air quality. It emphasizes the significant health impacts associated with air pollution, including reduced life expectancy. The introduction also outlines the challenges faced in implementing effective air quality policies and regulations. It sets the stage for the paper's focus on the need for reliable data, economically efficient regulatory frameworks, and the importance of piloting and evaluating new policies. **Linear Regression**: Often used for its simplicity and interpretability. **Random Forest**: A robust ensemble method that can handle non-linear relationships and interactions. **Support Vector** Machines **(SVM)**: Effective for classification tasks and can be used for regression as well. **Neural Networks**: Particularly deep learning models, which can capture complex patterns in large datasets. **Time Series Analysis**: Methods like ARIMA (Auto Regressive Integrated Moving Average) for forecasting air quality based on historical data.

[5] Depak Gaikar and his team members focused on predicting the Air Quality Index (AQI) using the Random Forest algorithm. It emphasizes that air pollution is a significant global concern with serious implications for human health, the environment, and the economy. In their research they used several machine learning algorithms that were applied for predicting the Air Quality Index (AQI). These algorithms include: **Random Forest (RF)** - Highlighted as a powerful algorithm for AQI prediction. **Support Vector Regression (SVR)** - Another machine learning technique used for forecasting. **Artificial Neural Networks (ANN)** - Mentioned as part of the comparative analysis for AQI prediction. **Linear Regression** - A standard statistical method used for prediction. **K-Nearest Neighbors (KNN)** - Included as one of the methods for comparison. **XG Boost** - Another algorithm referenced in the context of AQI prediction.  Random Forest algorithm performs well in predicting the Air Quality Index (AQI) and is noted for its accuracy compared to other machine learning methods. While it mentions that Random Forest is a powerful algorithm, it does not explicitly state which algorithm has the highest accuracy among all those tested, such as Linear Regression, Support Vector Regression, and Artificial Neural Networks.

[6] Air pollution is a significant health risk, being the leading cause of mortality from various diseases, including cardiovascular diseases, stroke, chronic obstructive pulmonary disease, and lung cancer. The concept of Disability Adjusted Life Years (DALYs) highlights the severe health burden of air pollution, combining years lost due to premature death and years lived with disability. It also adversely affects ecosystems, threatening flora and fauna and contributing to species extinction. Globally, approximately 9 out of 10 people are exposed to air pollution, which results in around 7 million deaths annually, according to the WHO. India is particularly affected, with average PM 2.5 levels reaching 40-50 μg per cubic meter—about four times the WHO's recommended limit. In 2019, air pollution was responsible for an estimated 1.7 million deaths in India, accounting for 18% of total deaths.

[7] Harish Kumar studied on the forecasting air pollution Particulate Matter(PM2.5) using machine learning Regression model. The research discusses the importance of air quality monitoring models using pollution sensors to predict pollution levels and assess air contamination. It highlights the use of machine learning methods, such as artificial neural networks, for forecasting pollution levels and the significance of combining machine learning with air pollution data for improved predictions. The abstract also emphasizes the need for effective air quality monitoring models due to increasing pollution levels and the importance of accurate pollution forecasts. The algorithms used in the research for air pollution prediction include linear regression, random forest regression, gradient boosting regression, K-neighbors regression, MLP regression, and Decision Tree regression CART. Additionally, the study compares the performance of these algorithms using cross-validation and performance criteria to measure accuracy. The results show that gradient boosting regression outperforms the other algorithms in terms of prediction accuracy. The research emphasizes the effectiveness of machine learning models, particularly the Gradient Boosting Regression algorithm, in predicting PM2.5 concentrations for air pollution monitoring. The study highlights the importance of accurate pollution forecasts for public health and environmental management. Overall, the research demonstrates the potential of machine learning techniques in improving air quality prediction models and addressing air pollution challenges in smart cities.

[8] In this research paper they investigate the performance and exhaust emissions of a diesel engine using artificial neural networks (ANNs). It specifically examines how variables such as injection pressure, engine speed and throttle position effect engine performance and emissions. In this they are study on some points. Those are **Injection Pressure** which is focuses on a design injection pressure of 150bar, with experiments conducted at four different pressures. **Throttle Positions** tested were 50%, 75% and 100%. **Performance Metrics** are used to measure various performance metrics including engine torque, power,brake mean, fuel flow and exhaust emission. **Neural Network Design**  is utilized a back-propagation learning algorithm with different configuration, including single and two hidden layers and employed Logistic sigmoid transfer function.

[9] A new neural network-based method was developed for short-term air pollution prediction specifically for SO2 pollution near the largest Slovenian thermal power plant at Sostanj due to high emissions. This method aims to predict and lower peak pollutant concentrations during critical meteorological conditions, especially in complex topography where traditional modeling methods may not be reliable. The results of this new method show great promise and can potentially be adapted for other key air pollutants with continuous measurement capabilities.

[10] Air pollution is a major source of worry for all living things. India has one of the world’s highest levels of air pollution. Rising population, unplanned growth, increased automotive traffic, stubble burning, industrial waste, fossil fuel combustion, powerplant emissions and a variety of other causes all contribute considerably to air pollution in developing countries. Particulate matter (PM) 2.5 is the most concerning of all air pollutants since it causes major health problems in individuals. Prediction and management of air quality have therefore become critical. Several machine learning algorithms were used in this work to examine dataset results. The results of our work suggest that for future predictions, logistic regression and autoregression can be efficaciously utilised for the analysis and forecasting of levels of PM2.5 in the future. Countries can lower the prevalence of strokes, and chronic and acute respiratory illnesses such as asthma, and lung cancer by reducing air pollution levels.

[11] David came up with analysis of air pollution data at a mixed source location using boosted regression trees. This paper investigates the use of boosted regression trees to analyze source characteristics in a complex location near a major international airport, focusing on hourly nitrogen oxides (NOX) concentrations. The study discusses model development, uncertainties quantification, and the importance of considering interactions between variables for improved model accuracy. It highlights the influence of different variables on predicted NOX concentrations, assesses aircraft take-off contributions to NOX levels, and compares model predictions with field campaign data, revealing faster dispersion of plumes from larger aircraft compared to smaller ones. The implications of this behavior are also explored in the context of air pollution near airports. Partial dependence plots were utilized to illustrate the relationships between nitrogen oxides (NOX) and each variable, showcasing the variable's impact on NOX concentrations while considering the average effects of all other variables in the model. These plots offer insights into how NOX concentrations are influenced by individual model variables. Confidence intervals around the variables were estimated using a bootstrap technique since boosted regression trees (BRT) do not inherently provide these estimates.

[12] Recent estimates using updated hazard ratio functions indicate a 30% increase in global premature mortality attributable to ambient PM2.5 exposure and a 17% increase due to O3 exposure from 2000 to 2015. The analysis reveals tradeoffs between changes in baseline mortality, population size, age distribution, and exposure, with global population growth alone contributing to a significant increase in mortality burden. The study emphasizes the importance of improving air quality and implementing air pollution mitigation strategies, alongside healthcare improvements, to address the growing burden of air pollution-related diseases worldwide.

[13] Dragomir, Elia Georgiana :One of the classical data mining techniques is k-nearest neighbor. This method uses the class of the k nearest neighbor to classify a new instance. The distance is calculated with one of the multiple mathematical distance metrics. In this paper, the technique is used in the air quality forecast domain in order to predict the value of the air quality index. This index is used to categorize the pollution level and to inform the population about some possible episodes of pollution.

[14] Carbajal-Hernández, José Juan, et al: In recent years, artificial intelligence methods have been used for the treatment of environmental problems. This work, presents two models for assessment and prediction of air quality. First, we develop a new computational model for air quality assessment in order to evaluate toxic compounds that can harm sensitive people in urban areas, affecting their normal activities. In this model we propose to use a Sigma operator to statistically asses air quality parameters using their historical data information and determining their negative impact in air quality based on toxicity limits, frequency average and deviations of toxicological tests. We also introduce a fuzzy inference system to perform parameter classification using a reasoning process and integrating them in an air quality index describing the pollution levels in five stages: excellent, good, regular, bad and danger, respectively. The second model proposed in this work predicts air quality concentrations using an autoregressive model, providing a predicted air quality index based on the fuzzy inference system previously developed. Using data from Mexico City Atmospheric Monitoring System, we perform a comparison among air quality indices developed for environmental agencies and similar models. Our results show that our models are an appropriate tool for assessing site pollution and for providing guidance to improve contingency actions in urban areas.

[15] Kumar, Anikender, and P. Goyal: As the impact of air pollutants on human health through ambient air address much attention in recent years, the air quality forecasting in terms of air pollution parameters becomes an important topic in environmental science. The Air Quality Index (AQI) can be estimated through a formula, based on comprehensive assessment of concentration of air pollutants, which can be used by government agencies to characterize the status of air quality at a given location. The present study aims to develop forecasting model for predicting daily AQI, which can be used as a basis of decision making processes. Firstly, the AQI has been estimated through a method used by US Environmental Protection Agency (USEPA) for different criteria pollutants as Respirable Suspended Particulate Matter (RSPM), Sulfur dioxide (SO2), Nitrogen dioxide (NO2) and Suspended Particulate Matter (SPM). However, the sub-index and breakpoint concentrations in the formula are made according to Indian National Ambient Air Quality Standard. Secondly, the daily AQI for each season is forecasted through three statistical models namely time series auto regressive integrated moving average (ARIMA) (model 1), principal component regression (PCR) (model 2) and combination of both (model 3) in Delhi. The performance of all three models are evaluated with the help of observed concentrations of pollutants, which reflects that model 3 agrees well with observed values, as compared to the values of model 1 and model 2. The same is supported by the statistical parameters also. The significance of meteorological parameters of model 3 has been assessed through principal component analysis (PCA), which indicates that daily rainfall, station level pressure, daily mean temperature, wind direction index are maximum explained in summer, monsoon, post-monsoon and winter respectively. Further, the variation of AQI during the weekends (holidays) and weekdays are found negligible. Therefore all the days of week are accounted same in the models.

[16] Singh, Kunwar P., et al: In this study, linear and nonlinear modeling was performed to predict the urban air quality of the Lucknow city (India). Partial least squares regression (PLSR), multivariate polynomial regression (MPR), and artificial neural network (ANN) approach-based models were constructed to predict the respirable suspended particulate matter (RSPM), SO2, and NO2 in the air using the meteorological (air temperature, relative humidity, wind speed) and air quality monitoring data (SPM, NO2, SO2) of five years (2005–2009). Three different ANN models, viz. multilayer perceptron network (MLPN), radial-basis function network (RBFN), and generalized regression neural network (GRNN) were developed. All the five different models were compared for their generalization and prediction abilities using statistical criteria parameters, viz. correlation coefficient (R), standard error of prediction (SEP), mean absolute error (MAE), root mean squared error (RMSE), bias, accuracy factor (Af), and Nash–Sutcliffe coefficient of efficiency (Ef). Nonlinear models (MPR, ANNs) performed relatively better than the linear PLSR models, whereas, performance of the ANN models was better than the low-order nonlinear MPR models. Although, performance of all the three ANN models were comparable, the GRNN over performed the other two variants. The optimal GRNN models for RSPM, NO2, and SO2 yielded high correlation (between measured and model predicted values) of 0.933, 0.893, and 0.885; 0.833, 0.602, and 0.596; and 0.932, 0.768 and 0.729, respectively for the training, validation and test sets. The sensitivity analysis performed to evaluate the importance of the input variables in optimal GRNN revealed that SO2 was the most influencing parameter in RSPM model, whereas, SPM was the most important input variable in other two models. The ANN models may be useful tools in the air quality predictions.

[17] Sivacoumar, R., et al: Jamshedpur, the steel city of India situated in the eastern part of India is affected by increasing air pollution levels as a result of concentrated industrial activities. The impact of NOX emissions resulting from various air pollution sources, viz. industries, vehicles and domestic, was estimated using Industrial Source Complex Short-Term gaussian dispersion model. The contribution of NOx concentration from industrial, vehicular and domestic sources was found to be 53, 40 and 7%. Further statistical analysis was carried out to evaluate the model performance by comparing measured and predicted NOx concentrations. The model performance was found good with an accuracy of about 68%.

[18] Gokhale, Sharad, and Namita Raokhande: Science of the total environment, 2008 - ElsevieR

There are several models that can be used to evaluate roadside air quality. The comparison  
of the operational performance of different models pertinent to local conditions is desirable  
so that the model that performs best can be identified. Three air quality models, namely the  
'modified General Finite Line Source Model'(M-GFLSM) of particulates, the 'California Line  
Source'(CALINE3) model, and the 'California Line Source for Queuing & Hot Spot  
Calculations'(CAL3QHC) model have been identified for evaluating the air quality.

[19] Bhanarkar, A. D., et al: Contribution of pollution from different types of sources in Jamshedpur, the steel city of India, has been estimated in winter 1993 using two approaches in order to delineate and prioritize air quality management strategies for the development of region in an environmental friendly manner. The first approach mainly aims at preparation of a comprehensive emission inventory and estimation of spatial distribution of pollution loads in terms of SO2 and NO2 from different types of industrial, domestic and vehicular sources in the region. The results indicate that industrial sources account for 77% and 68% of the total emissions of SO2 and NO2, respectively, in the region, whereas vehicular emissions contributed to about 28% of the total NO2 emissions. In the second approach, contribution of these sources to ambient air quality levels to which the people are exposed to, was assessed through air pollution dispersion modelling. Ambient concentration levels of SO2 and NO2 have been predicted in winter season using the ISCST3 model. The analysis indicates that emissions from industrial sources are responsible for more than 50% of the total SO2 and NO2 concentration levels. Vehicular activities contributed to about 40% of NO2 pollution and domestic fuel combustion contributed to about 38% of SO2 pollution. Predicted 24-h concentrations were compared with measured concentrations at 11 ambient air monitoring stations and good agreement was noted between the two values. In-depth zone-wise analysis of the above indicates that for effective air quality management, industrial source emissions should be given highest priority, followed by vehicular and domestic sources in Jamshedpur region.

[20] Singh, Kunwar P., Shikha Gupta, and Premanjali Rai: In this study, principal components analysis (PCA) was performed to identify air pollution sources and tree based ensemble learning models were constructed to predict the urban air quality of Lucknow (India) using the air quality and meteorological databases pertaining to a period of five years. PCA identified vehicular emissions and fuel combustion as major air pollution sources. The air quality indices revealed the air quality unhealthy during the summer and winter. Ensemble models were constructed to discriminate between the seasonal air qualities, factors responsible for discrimination, and to predict the air quality indices. Accordingly, single decision tree (SDT), decision tree forest (DTF), and decision tree boost (DTB) were constructed and their generalization and predictive performance was evaluated in terms of several statistical parameters and compared with conventional machine learning benchmark, support vector machines (SVM). The DT and SVM models discriminated the seasonal air quality rendering misclassification rate (MR) of 8.32% (SDT); 4.12% (DTF); 5.62% (DTB), and 6.18% (SVM), respectively in complete data. The AQI and CAQI regression models yielded a correlation between measured and predicted values and root mean squared error of 0.901, 6.67 and 0.825, 9.45 (SDT); 0.951, 4.85 and 0.922, 6.56 (DTF); 0.959, 4.38 and 0.929, 6.30 (DTB); 0.890, 7.00 and 0.836, 9.16 (SVR) in complete data. The DTF and DTB models outperformed the SVM both in classification and regression which could be attributed to the incorporation of the bagging and boosting algorithms in these models. The proposed ensemble models successfully predicted the urban ambient air quality and can be used as effective tools for its management.

[21] Wang, Jun, and Sundar A. Christopher: We explore the relationship between column aerosol optical thickness (AOT) derived from the Moderate Resolution Imaging Spectro Radiometer (MODIS) on the Terra/Aqua satellites and hourly fine particulate mass (PM2.5) measured at the surface at seven locations in Jefferson county, Alabama for 2002. Results indicate that there is a good correlation between the satellite‐derived AOT and PM2.5 (linear correlation coefficient, R = 0.7) indicating that most of the aerosols are in the well‐mixed lower boundary layer during the satellite overpass times. There is excellent agreement between the *monthly mean* PM2.5 and MODIS AOT (R > 0.9), with maximum values during the summer months due to enhanced photolysis. The PM2.5 has a distinct diurnal signature with maxima in the early morning (6:00 ∼ 8:00AM) due to increased traffic flow and restricted mixing depths during these hours. Using simple empirical linear relationships derived between the MODIS AOT and 24hr mean PM2.5 we show that the MODIS AOT can be used quantitatively to estimate air quality categories as defined by the U.S. Environmental Protection Agency (EPA) with an accuracy of more than 90% in cloud‐free conditions. We discuss the factors that affect the correlation between satellite‐derived AOT and PM2.5 mass, and emphasize that more research is needed before applying these methods and results over other areas.

[22] Sharma, M., E. A. McBean, and U. Ghosh: Soil-derived particles in suspended particulate matter (SPM) are reported to be responsible for acid rain neutralization in India. A study on the seasonal variability of SPM indicates that the winter season SPM levels are significantly less than the summer season SPM levels thereby suggesting the winter season to be critical for acid rain. Therefore, SO4 deposition is examined for the winter season at sensitive receptors in northern India using a long-range air transport model. The UW-LRT model is employed to estimate the SO4 deposition levels at six locations given the point source emission quantities from 12 major sources. The model results match favourably with literature-reported deposition levels. The model results argue for a well-designed monitoring program for measurement of sulphate deposition to better determine the extent of the problem and model calibration.

[23] Madan, Tanisha, Shrddha Sagar, and Deepali Virmani: Predicting air quality is necessary step to be taken by government as it is becoming the major concern among the health of human beings. Air quality Index measure the quality of air. Various air pollutants causing air pollution are Carbon dioxide, Nitrogen dioxide, carbon monoxide etc that are released from burning of natural gas, coal and wood, industries, vehicles etc. Air Pollution can cause severe disease like lungs cancer, brain disease and even lead to death. Machine learning algorithms helps in determining the air quality index. Various research is being done in this field but still results are still not accurate. Dataset are available from Kaggle, air quality monitoring sites and divided into two Training and Testing. Machine Learning algorithms employed for this are Linear Regression, Decision Tree, Random Forest, Artificial Neural Network, Support Vector Machine.

[24] Li, Chenchen, Yan Li, and Yubin Bao: In recent years, the prevention and control of environmental pollution attracted much attention, and the haze weather directly affects people's travel health. In order to effectively prevent and control air pollution, optimize the air quality evaluation system. In this paper, PM 2.5 , PM 10 , SO 2 , NO 2 , CO and O 3 \_8h are used as characteristic factors, and air quality index is used as a decision factor. A variety of regression algorithms are selected to establish a prediction model, and the accuracy and generalization ability of various algorithms are compared. The results show that the Random Forest Regression algorithm (RFR) and the Gradient Boosting Regression algorithm (GBR) can effectively predict the Air Quality Index (AQI) and the air quality level. This paper provides a reference for the establishment of the air quality model.

[25] The survival of humanity is heavily reliant on air quality, which has been adversely affected by ongoing industrial, transportation and domestic activities. Monitoring and predicting air quality is curcial. This research analyzes six years of air pollution data from 23 Indian cities, utilizing machine learning methods. Exploratory data analysis reveals hidden patterns, with a notable decrease in pollutant levels during the pandemic year 2020.The Guassian Naïve Bayes model achieves the highest accuracy, while the Support Vector Machine model shows the lowest accuracy. The study underscores the importance of machine learning in air quality analysis and prediction.

[26] Ibrahim Olanrewaju Alade in his study, the specific heat capacity of Alumina (Al2O3)/water nanofluid has been accurately evaluated using genetic algorithm/support vector regression (GA/SVR) model at volume fractions of 3.7–9.3%. The proposed (genetic algorithm/support vector regression) GA/SVR model was formulated using volume fractions and specific heat capacities of the alumina nanoparticles. The developed GA/SVR model is very accurate as determined from 99.998% correlation coefficient with experimentally obtained data and also has a root mean square error of 0.0014. Furthermore, the obtained results from the GA/SVR were compared with existing analytic models. Remarkably, the proposed model achieved an order of magnitude improvement over the model based on thermal equilibrium (Model II) and a two order of magnitude improvement over the model based on simple mixing rule for ideal gases (model I). Given the improvement in the accuracy, the proposed model would be useful for rapid and highly accurate estimation of the specific heat capacity of alumina/water nanofluids.

[27] In the populated and developing countries, governments consider the regulation of air as a major task. The meteorological and traffic factors, burning of fossil fuels, industrial parameters such as power plant semissions play significant roles in air pollution. Among all the particulate matter that determine the quality of the air, particulate matter(PM2.5) needs more attention. When it’s level is high in the air, it causes serious issues on people’s health. Hence, controlling it by constantly keeping a check on its level in the air is important. In this paper, Logistic Regression is employed to detect whether a data sample is either polluted or not pollute. Auto regression is employed to predict future values of PM2.5 based on the previous PM2.5 readings. Knowledge of level of PM2.5 in nearing years, month or week, enables us to reduce its level to lesser than the harmful range. This system attempts to predict PM2.5 level and detect air quality based on data set consisting of daily atmospheric conditions in a specific city.

Air pollution is a major environmental issue and

machine learning techniques play an important role in analyzing

and forecasting these data sets. Air quality is an outcome of

the complex interaction of several factors involving the chemical

reactions, meteorological parameters, and emissions from nat-

ural and anthropogenic sources. In this paper, we propose an

efﬁcient combined technique that takes the beneﬁts of statistical

techniques and machine learning techniques to predict/forecast

the Air Quality and Pollution in particular regions. This work

also indicates that prediction performance varies over different

regions/cities in India. We used time series analysis, regression

and Ada-boosting to anticipate PM 2.5 concentration levels in

several locations throughout Hyderabad on an annual basis,

depending on numerous atmospheric and surface parameters like

wind speed, air temperature, pressure, and so on. Dataset for

this investigation is taken from Kaggle and experimented with

proposed method and comparison results of our experiments are

then plotted.

Keywords—Air quality; forecasting; machine learning; statisti-

cal techniqu

Air pollution is a major environmental issue and

machine learning techniques play an important role in analyzing

and forecasting these data sets. Air quality is an outcome of

the complex interaction of several factors involving the chemical

reactions, meteorological parameters, and emissions from nat-

ural and anthropogenic sources. In this paper, we propose an

efﬁcient combined technique that takes the beneﬁts of statistical

techniques and machine learning techniques to predict/forecast

the Air Quality and Pollution in particular regions. This work

also indicates that prediction performance varies over different

regions/cities in India. We used time series analysis, regression

and Ada-boosting to anticipate PM 2.5 concentration levels in

several locations throughout Hyderabad on an annual basis,

depending on numerous atmospheric and surface parameters like

wind speed, air temperature, pressure, and so on. Dataset for

this investigation is taken from Kaggle and experimented with

proposed method and comparison results of our experiments are

then plotted.

Keywords—Air quality; forecasting; machine learning; statisti-

cal techniqu

Air pollution is a major environmental issue and

machine learning techniques play an important role in analyzing

and forecasting these data sets. Air quality is an outcome of

the complex interaction of several factors involving the chemical

reactions, meteorological parameters, and emissions from nat-

ural and anthropogenic sources. In this paper, we propose an

efﬁcient combined technique that takes the beneﬁts of statistical

techniques and machine learning techniques to predict/forecast

the Air Quality and Pollution in particular regions. This work

also indicates that prediction performance varies over different

regions/cities in India. We used time series analysis, regression

and Ada-boosting to anticipate PM 2.5 concentration levels in

several locations throughout Hyderabad on an annual basis,

depending on numerous atmospheric and surface parameters like

wind speed, air temperature, pressure, and so on. Dataset for

this investigation is taken from Kaggle and experimented with

proposed method and comparison results of our experiments are

then plotted.

Keywords—Air quality; forecasting; machine learning; statisti-

cal techniqu

Air pollution is a major environmental issue and

machine learning techniques play an important role in analyzing

and forecasting these data sets. Air quality is an outcome of

the complex interaction of several factors involving the chemical

reactions, meteorological parameters, and emissions from nat-

ural and anthropogenic sources. In this paper, we propose an

efﬁcient combined technique that takes the beneﬁts of statistical

techniques and machine learning techniques to predict/forecast

the Air Quality and Pollution in particular regions. This work

also indicates that prediction performance varies over different

regions/cities in India. We used time series analysis, regression

and Ada-boosting to anticipate PM 2.5 concentration levels in

several locations throughout Hyderabad on an annual basis,

depending on numerous atmospheric and surface parameters like

wind speed, air temperature, pressure, and so on. Dataset for

this investigation is taken from Kaggle and experimented with

proposed method and comparison results of our experiments are

then plotted.

Keywords—Air quality; forecasting; machine learning; statisti-

cal techniqu

Air pollution is a major environmental issue and

machine learning techniques play an important role in analyzing

and forecasting these data sets. Air quality is an outcome of

the complex interaction of several factors involving the chemical

reactions, meteorological parameters, and emissions from nat-

ural and anthropogenic sources. In this paper, we propose an

efﬁcient combined technique that takes the beneﬁts of statistical

techniques and machine learning techniques to predict/forecast

the Air Quality and Pollution in particular regions. This work

also indicates that prediction performance varies over different

regions/cities in India. We used time series analysis, regression

and Ada-boosting to anticipate PM 2.5 concentration levels in

several locations throughout Hyderabad on an annual basis,

depending on numerous atmospheric and surface parameters like

wind speed, air temperature, pressure, and so on. Dataset for

this investigation is taken from Kaggle and experimented with

proposed method and comparison results of our experiments are

then plotted.

Keywords—

Air pollution is a major environmental issue and

machine learning techniques play an important role in analyzing

and forecasting these data sets. Air quality is an outcome of

the complex interaction of several factors involving the chemical

reactions, meteorological parameters, and emissions from nat-

ural and anthropogenic sources. In this paper, we propose an

efﬁcient combined technique that takes the beneﬁts of statistical

techniques and machine learning techniques to predict/forecast

the Air Quality and Pollution in particular regions. This work

also indicates that prediction performance varies over different

regions/cities in India. We used time series analysis, regression

and Ada-boosting to anticipate PM 2.5 concentration levels in

several locations throughout Hyderabad on an annual basis,

depending on numerous atmospheric and surface parameters like

wind speed, air temperature, pressure, and so on. Dataset for

this investigation is taken from Kaggle and experimented with

proposed method and comparison results of our experiments are

then plotted.

Keywords—

[28] V.Devasekhar and P.Natarajan carried out air quality is an outcome of the complex interaction of several factors involving the chemical reactions, meteorological parameters and emission from natural and anthropogenic sources. In this paper, we propose an efficient combined technique that takes the benefits of statistical techniques and machine learning to predict/forecast the Air Quality and Pollution in particular regions. This work also indicates that prediction performance varies over different regions/cities in India. We used time series analysis, regression and Ada-boosting to anticipate PM2.5 concentration levels in several location throughout Hyderabad on an annual basis, depending on numerous atmospheric and surface parameters like wind speed, air temperature, pressure and so on. Dataset for this investigation is take from Kaggle and experimented with proposed method and comparison results of our experiments are then plotted.

[29] Air pollution continues to be a problem that affects us all worldwide, since it is estimated that 7 million people die each year due to repeated exposure to pollutants that cause chronic conditions such as severe respiratory diseases, cardiovascular problems and cancer. On the other hand, climatological effects lead to the deterioration of the planet’s ecosystems. Therefore, air quality monitoring systems are the main tools used by governments to control the emission of toxic gases into the atmosphere. This makes it possible to ensure the quality of life and the general well-being of the population, as well as to strengthen the agricultural and industrial sectors. The objective of this work is to present a review of the characteristics present in the mechanisms applied in the prediction of air pollution, with the main objective of synthesizing the knowledge found, identifying the models, approaches and variables that are most studied. The results show that the hybrid models based on CNN-LTSM are the most used. Other studies use GRU and ELM that have good results when making predictions. Suspended particles (PM10 and PM2.5) are the main object of study. Multivariate models are also more accurate and efficient when it comes to forecasting.

[30] Air pollution in the environment is growing daily as a result of urbanization and population growth, which causes numerous health issues. Information about air quality and environmental health risks provided by air pollutant data is crucial for environmental management. The use of artificial neural network (ANN) approaches for predicting air pollutants is reviewed in this research. These methods are based on several forecast intervals, including hourly, daily, and monthly ones. This study shows that ANN techniques forecast air contaminants more precisely than traditional methods. It has been discovered that the input parameters and architecture-type algorithms used affect the accuracy of air pollutant prediction models. ANN is therefore more accurate and reliable than other empirical models because they can handle a wide range of input meteorological parameters. Finally, research gap of neural networks for air pollutant prediction is identified. The review may inspire researchers and to a certain extent promote the development of artificial intelligence in air pollutant prediction.

[31] The impact of air pollution in Chennai metropolitan city, a southern Indian coastal city was examined to predict the Air Quality Index (AQI). Regular monitoring and prediction of the Air Quality Index (AQI) are critical for combating air pollution. The current study created machine learning models such as XG Boost, Random Forest, Bagging Regressor, and LGBM Regressor for the prediction of the AQI using the historical data available from 2017 to 2022. According to historical data, the AQI is highest in January, with a mean value of 104.6 g/gm, and the lowest in August, with a mean AQI value of 63.87 g/gm. Particulate matter, gaseous pollutants, and meteorological parameters were used to predict AQI, and the heat map generated showed that of all the parameters, PM2.5 has the greatest impact on AQI, with a value of 0.91. The log transformation method is used to normalize datasets and determine skewness and kurtosis. The XGBoost model demonstrated strong performance, achieving an R2 (correlation coefficient) of 0.9935, a mean absolute error (MAE) of 0.02, a mean square error (MSE) of 0.001, and a root mean square error (RMSE) of 0.04. In comparison, the Light GBM model's prediction was less effective, as it attained an R2 of 0.9748. According to the study, the AQI in Chennai has been increasing over the last two years, and if the same conditions persist, the city's air pollution will worsen in the future. Furthermore, accurate future air quality level predictions can be made using historical data and advanced machine learning algorithms.

[32] Vivek Singh air quality of India by using machine literacy to prognosticate the air quality indicator of a given area. Air quality indicator of India is a standard measure used to indicate the contaminant (so2, no2, rspm, spm etc.) situations over a period. We developed a model to prognosticate the air quality indicator grounded on literal data of former times and prognosticating over a particular forthcoming time as a Gradient decent boosted multivariable retrogression problem. we ameliorate the effectiveness of the model by applying cost Estimation for our prophetic Problem. Our model will be able for successfully prognosticating the air quality indicator of a total county or any state or any bounded region handed with the literal data of contaminant attention. In our model by enforcing the proposed parameter- reducing phrasings, we achieved better performance than the standard retrogression models. our model has 96 delicacy on prognosticating the current available dataset on prognosticating the air quality indicator of whole India, also we use AHP MCDM fashion to find of order of preference by similarity to ideal result.

[33] Air is one of the necessities of living things. Therefore, it is necessary to have good air quality. Air pollution can cause many negative impacts on life. Therefore, it is important to know the air quality in an area. Jakarta is one of the cities with poor air quality in Indonesia and the world. During the Covid-19 pandemic, the government implemented a large-scale social restriction policy, the impact of this policy was better air quality. But now it has started to back to normal, then it is important to control air quality. There are 5 locations to measure air quality in Jakarta. The results of the independence test between the location of air quality measurements and critical variables on air pollution indicate a relationship between the two variables. Moreover, there are differences in air quality before Covid-19 and during Covid-19 based on the results of the t-test. Air quality classification was carried out in this research using machine learning methods. Because there are several levels of air quality, the classification uses a multiclass classification. Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) Classifier are used in this research. Because based on the results of literature reviews, both methods produce high accuracy. The results of this research showed a comparison of the two methods. The comparison showed that the SVM method is better than the MLP Classifier.

[34] N.Srinivas Gupta, Yashvi Mohta, Khyati Henda, Raahil Armaan, B.Valrmathi, G.Arulkumaran researched on prediction of air quality index using machine learning techniques: a comprehensive analysis. An index for reporting air quality is called the air quality index (AQI). It measures the impact of air pollution on a person’s health over a short period of time. The purpose of the AQI is to educate the public on the negative health effects of local air pollution. The amount of air pollution in Indian cities has significantly increased. There are several ways to create a mathematical formula to determine the air quality index. Numerous studies have found a link between air pollution exposure and adverse health impacts in the population. Data mining techniques are one of the most interesting approaches to forecast AQI and analyze it. The aim of this paper is to find the most effective way for AQI prediction to assist in climate control. The most effective method can be improved upon to find the most optimal solution. Hence, the work in this paper involves intensive research and the addition of novel techniques such as SMOTE to make sure that the best possible solution to the air quality problem is obtained. Another important goal is to demonstrate and display the exact metrics involved in our work in such a way that it is educational and insightful and hence provides proper comparisons and assists future researchers. In the proposed work, three distinct methods—support vector regression (SVR), random forest regression (RFR), and CatBoost regression (CR)—have been utilized to determine the AQI of New Delhi, Bangalore, Kolkata, and Hyderabad. After comparing the results of imbalanced datasets, it was found that random forest regression provides the lowest root mean square error (RMSE) values in Bangalore (0.5674), Kolkata (0.1403), and Hyderabad (0.3826), as well as higher accuracy compared to SVR and CatBoost regression for Kolkata (90.9700%) and Hyderabad (78.3672%), while CatBoost regression provides the lowest RMSE value in New Delhi (0.2792) and the highest accuracy is obtained for New Delhi (79.8622%) and Bangalore (68.6860%). Regarding the dataset that was subjected to the synthetic minority oversampling technique (SMOTE) algorithm, it is noted that random forest regression provides the lowest RMSE values in Kolkata (0.0988) and Hyderabad (0.0628) and higher accuracies are obtained for Kolkata (93.7438%) and Hyderabad (97.6080%) in comparison to SVR and CatBoost regression, whereas CatBoost regression provides the highest accuracies for New Delhi (85.0847%) and Bangalore (90.3071%). This demonstrated definitely that datasets that had the SMOTE algorithm applied to them produced a higher accuracy. The novelty of this paper lies in the fact that the best regression models have been picked through thorough research by analyzing their accuracies. Moreover, unlike most related papers, dataset balancing is carried out through SMOTE. Moreover, all of the implementations have been documented via graphs and metrics, which clearly show the contrast in results and help show what actually caused the improvement in accuracy.

[35] Artificial intelligence (AI) is a technique in which computers are designed to do tasks just like humans, they are designed to think, walk, talk and do anything that a living thing can do. Machine Learning (ML) is a field of research devoted to understanding and ’learning’ building methods, that is, methods that improve data to improve the performance of a particular set of tasks. This study is concerned with combining data of pollutants, meteorological, and traffic data with statistical temporal-spatial feature engineering to provide multi-step-ahead air quality forecasts for 24 and 48 hours. It examines a multivariate time series approach to modeling and forecasting the pollution of PM2.5, PM10, and NO2 at three air quality stations in India. The data-driven approach is thus believed to be an excellent complement for the knowledge-driven model.

[36] Air pollution is the “world’s largest environmental health threat”[1], causing 7 million deaths[1] worldwide every year. Its major constituents are PM2.5, PM10 and the harmful green house gases S02, N02, C0 and other effluents from vehicles and factories affecting not only humans but also other living organisms both on land and sea. The only effective solution to this global issue is to implement machine learning algorithms to predict the AQI (Air Quality Index ) that can make the people aware of the condition of the air of a certain region such that certain actions could be issued by the government for the improvement of the air quality in the future. The prime objective behind this project is to predict the AQI based on the concentration of PM2.5, PM10,S02, N02, C0 as well as weather conditions like temperature, pressure and humidity[2].Hence the data set is combined from various web sources like cpcb.nic.in and uci repository in order to bring accuracy in the prediction and to justify whether the Quality of air is suitable or not. This prediction will be brought about with the help of some supervised machine learning algorithms and the observation and the result will state which algorithm is giving better accuracy in prediction of AQI and which one is giving less error.

[37] In recent years, people have been paying more and more attention to air quality because it directly affects people’s health and daily life. Effective air quality prediction has become one of the hot research issues. However, this paper is suffering many challenges, such as the instability of data sources and the variation of pollutant concentration along time series. Aiming at this problem, we propose an improved air quality prediction method based on the Light GBM model to predict the PM2.5 concentration at the 35 air quality monitoring stations in Beijing over the next 24 h. In this paper, we resolve the issue of processing the high-dimensional large-scale data by employing the Light GBM model and innovatively take the forecasting data as one of the data sources for predicting the air quality. With exploring the forecasting data feature, we could improve the prediction accuracy with making full use of the available spatial data. Given the lack of data, we employ the sliding window mechanism to deeply mine the high-dimensional temporal features for increasing the training dimensions to millions. We compare the predicted data with the actual data collected at the 35 air quality monitoring stations in Beijing. The experimental results show that the proposed method is superior to other schemes and prove the advantage of integrating the forecasting data and building up the high-dimensional statistical analysis.

[38] Pedro Mariano came up automated learning of air pollution predictive models that were trained using information gathered by a set of mobile low-cost sensors. Concretely, fast to compute machine learning methods (Decision Trees and Support Vector Machines) were used to build regression models that predict air pollution levels for a given location. The models were trained using the data collected by the Open Sense project, in particular, number of particulate matter, particle diameter, and lung deposited surface area (LDSA). We examined two different sets of attributes: one based on a geographical description of the location under analysis (e.g. distribution of households and roads), and another based on a time series of past air pollution observations in that location. Overall, we have found out that past measures lead to better pollution predictions. The best R2 score was 0.751 obtained with the model that predicts LDSA and was trained with the data set with time series attributes, while the worst R2 was 0.009 obtained with the geographical data set to predict number of particles. The performance of the best model is on par with similar air pollution systems

[39] In recent years, artificial intelligence methods have been used for the treatment of environmental problems. This work, presents two models for assessment and prediction of air quality. First, we develop a new [computational model](https://www.sciencedirect.com/topics/earth-and-planetary-sciences/computational-modeling) for air quality assessment in order to evaluate toxic compounds that can harm sensitive people in urban areas, affecting their normal activities. In this model we propose to use a Sigma operator to statistically asses air quality parameters using their historical data information and determining their negative impact in air quality based on toxicity limits, frequency average and deviations of toxicological tests. We also introduce a fuzzy inference system to perform parameter classification using a reasoning process and integrating them in an air quality index describing the pollution levels in five stages: *excellent*, *good*, *regular*, *bad* and *danger*, respectively. The second model proposed in this work predicts air quality concentrations using an autoregressive model, providing a predicted air quality index based on the fuzzy inference system previously developed. Using data from [Mexico](https://www.sciencedirect.com/topics/earth-and-planetary-sciences/mexico) City Atmospheric Monitoring System, we perform a comparison among air quality indices developed for environmental agencies and similar models. Our results show that our models are an appropriate tool for assessing site pollution and for providing guidance to improve contingency actions in urban areas.

[40] As the impact of air pollutants on human health through ambient air address much attention in recent years, the air quality forecasting in terms of air pollution parameters becomes an important topic in environmental science. The Air Quality Index (AQI) can be estimated through a formula, based on comprehensive assessment of concentration of air pollutants, which can be used by government agencies to characterize the status of air quality at a given location. The present study aims to develop forecasting model for predicting daily AQI, which can be used as a basis of decision making processes. Firstly, the AQI has been estimated through a method used by US Environmental Protection Agency (USEPA) for different criteria pollutants as Respirable Suspended Particulate Matter (RSPM), Sulfur dioxide (SO2), Nitrogen dioxide (NO2) and Suspended Particulate Matter (SPM). However, the sub-index and breakpoint concentrations in the formula are made according to Indian National Ambient Air Quality Standard. Secondly, the daily AQI for each season is forecasted through three statistical models namely time series auto regressive integrated moving average (ARIMA) (model 1), principal component regression (PCR) (model 2) and combination of both (model 3) in Delhi. The performance of all three models are evaluated with the help of observed concentrations of pollutants, which reflects that model 3 agrees well with observed values, as compared to the values of model 1 and model 2. The same is supported by the statistical parameters also. The significance of meteorological parameters of model 3 has been assessed through principal component analysis (PCA), which indicates that daily rainfall, station level pressure, daily mean temperature, wind direction index are maximum explained in summer, monsoon, post-monsoon and winter respectively. Further, the variation of AQI during the weekends (holidays) and weekdays are found negligible. Therefore all the days of week are accounted same in the models.

[41] With the development of industry, air pollution has become a serious problem. It is very important to create an air quality prediction model with high accuracy and good performance. Therefore, a new method of CT-LSTM is proposed in this paper, in which the prediction model is established by combining chi-square test (CT) and long short-term memory (LSTM) network model. CT is used to determine the influencing factors of air quality. The hourly air quality data and meteorological data from Jan. 1, 2017 to Dec. 31, 2018 are used to train the LSTM network model. The data from Jan. 1, 2019 to Dec. 31, 2019 are used to evaluate the LSTM network model. The AQI level of Shijiazhuang of Hebei Province of China from Jan. 1, 2019 to Dec. 31, 2019 is predicted with five methods (SVR, MLP, BP neural network, Simple RNN and this paper's new method). Then, a contrastive analysis of the five prediction results is made. The experimental results show that the accuracy of this new method reaches 93.7%, which is the highest in the five methods and the maximum error of this new method is 1. The correct number of days predicted by this new method is also the highest among the five methods, which is 342 days. The new method also shows good characteristics in MAE, MSE and RMSE, which makes it more accurate for people to predict the AQI level.

[42] Over the past few decades, due to human activities, industrialization, and urbanization, air pollution has become a life-threatening factor in many countries around the world. Among air pollutants, Particulate Matter with a diameter of less than 2.5𝜇𝑚 (𝑃𝑀2.5) is a serious health problem. It causes various illnesses such as respiratory tract and cardiovascular diseases. Hence, it is necessary to accurately predict the 𝑃𝑀2.5 concentrations in order to prevent the citizens from the dangerous impact of air pollution beforehand. The variation of 𝑃𝑀2.5 depends on a variety of factors, such as meteorology and the concentration of other pollutants in urban areas. In this paper, we implemented a deep learning solution to predict the hourly forecast of 𝑃𝑀2.5 concentration in Beijing, China, based on CNN-LSTM, with a spatial-temporal feature by combining historical data of pollutants, meteorological data, and 𝑃𝑀2.5 concentration in the adjacent stations. We examined the difference in performances among Deep learning algorithms such as LSTM, Bi-LSTM, GRU, Bi-GRU, CNN, and a hybrid CNN-LSTM model. Experimental results indicate that our method “hybrid CNN-LSTM multivariate” enables more accurate predictions than all the listed traditional models and performs better in predictive performance

[43] In the last decade, we have seen drastic changes in the air pollution level, which has become a critical environmental issue. It should be handled carefully towards making the solutions for proficient healthcare. Reducing the impact of air pollution on human health is possible only if the data is correctly classified. In numerous classification problems, we are facing the class imbalance issue. Learning from imbalanced data is always a challenging task for researchers, and from time to time, possible solutions have been developed by researchers. In this paper, we are focused on dealing with the imbalanced class distribution in a way that the classification algorithm will not compromise its performance. The proposed algorithm is based on the concept of the adjusting kernel scaling (AKS) method to deal with the multi-class imbalanced dataset. The kernel function's selection has been evaluated with the help of weighting criteria and the chi-square test. All the experimental evaluation has been performed on sensor-based Indian Central Pollution Control Board (CPCB) dataset. The proposed algorithm with the highest accuracy of 99.66% wins the race among all the classification algorithms i.e. Adaboost (59.72%), Multi-Layer Perceptron (95.71%), GaussianNB (80.87%), and SVM (96.92). The results of the proposed algorithm are also better than the existing literature methods. It is also clear from these results that our proposed algorithm is efficient for dealing with class imbalance problems along with enhanced performance. Thus, accurate classification of air quality through our proposed algorithm will be useful for improving the existing preventive policies and will also help in enhancing the capabilities of effective emergency response in the worst pollution situation.

[44] Air pollution is a growing environmental concern, especially in big cities. The effects of air pollution are harmful to both living beings and the environment. Air quality can be predicted using techniques like probability, statistics but these methods are complex to predict. Machine learning is a better approach to air quality prediction. Air quality forecasting is a crucial step to protect public health by providing an early warning against harmful air pollutants. Prediction of air quality will assist in initiating emergency measures to reduce the discharge of pollutants and mitigate the consequences. It was analysed that imbalance class distribution give inaccurate predictions. As a result, two well-known resampling techniques are used: Synthetic Minority Oversampling Technique (SMOTE) for oversampling and Neighbourhood Cleaning Rule (NCR) for undersampling. The effects of these approaches and their combination (SMOTE+NCR) on prominent machine learning classifiers K-Nearest Neighbours (KNN) and Naive Bayes are compared. The presented results demonstrate that KNN performed better on resampled data using SMOTE+NCR and Naive Bayes performed better on undersampled data using NCR.

[45] The economic and social impact of poor air quality in towns and cities is increasingly being recognised, together with the need for effective ways of creating awareness of real-time air quality levels and their impact on human health. With local authority maintained monitoring stations being geographically computational data-driven mechanisms are needed to address the data sparsity challenge. In this paper, we propose a machine learning based method accurately predict the Air Quality Index(AQI), using environmental monitoring data together with meteorological measurements. To do so , we develop an air quality estimation framework that implements a neural network that is enhanced with a noel Non-linear Autoregressive neural network with exogenous input(NARX), especially designed for time series prediction. The framework is applied to a case study featuring different monitoring sites in London, with comparisons against other standard machine-learning based predictive algorithms showing the feasibility and robust performance of the proposed method for different kinds for areas within an urban region.

[46] Puneeth Goswmi and his team members are researched on the air pollution prediction and hotpot detection using machine learning. Air pollution is a vital issue that affects day-to-day lives. It is observed that throughout the world, there is an instant need to overcome the monster of pollution. According to statistics, most of the polluted cities in the world are in India. This poses a serious need of the hour for the Indian scientists, engineers, and authorities as a whole to fight and reduce it as much as possible. The time has come when one needs to plan their outside activities on pollution levels and air quality status. Air Quality Index (AQI) varies daily; hence it is difficult to predict future trends for the same. The current study proposed a machine learning-based model that uses sensors, past/present pollutants concentration data, and satellite data to predict air pollution in the regions in India. We emphasize the fact that other than measurable pollutants (PM10, PM2.5, NO2, etc.); meteorological data like wind, temperature, and fire are also important factors in determining pollution. The model uses Long Short-Term Memory, which is the state-of-the-art technique used for time series prediction. The model could predict the concentration of the pollutants and calculate the AQI for the areas where data was available for the near future. The Root Mean Square Error on test data is 54. The results are quite promising and future model can be made, taking this as a base model. An inexpensive prediction technique can greatly help the administration in mitigating pollution.

[47] Modeling spatial and temporal dependencies is essential for achieving accurate air quality prediction. Current air quality prediction models often overlook the underlying causal relationships in the data and primarily focus on statistical correlations. As a result, these models lack sufficient predictive power for medium- and long-term forecasts. This paper introduces causality into air quality prediction and proposes a Causal Spatio-Temporal Convolutional Network (CSTCN). We utilize an attention mechanism to automatically assign attention weights to each air quality monitoring site, enabling the capture of causal relationships between sites in the spatial dimension. Conducting tests on the identified relationships ensures the causality of the data in space. Furthermore, convolution operations are applied to extract the spatio-temporal features of the monitoring stations, while also utilizing causal convolution to ensure the causality of the data over time. The experiments conducted on the Beijing air quality dataset demonstrate that CSTCN exhibits outstanding performance in medium- and long-term predictions.

[48] Haoran Wang and his team members cameup with the idea deciphering urban traffic impacts on air quality by deep learning and emission inventory. Air pollution is a major obstacle to future sustainability, and traffic pollution has become a large drag on the sustainable developments of future metropolises. Here, combined with the large volume of real-time monitoring data, we propose a deep learning model, iDeepAir, to predict surface-level PM2.5 concentration in Shanghai megacity and link with MEIC emission inventory creatively to decipher urban traffic impacts on air quality. Our model exhibits high-fidelity in reproducing pollutant concentrations and reduces the MAE from 25.355 µg/m3 to 12.283 µg/m3 compared with other models. And identifies the ranking of major factors, local meteorological conditions have become a nonnegligible factor. Layer-wise relevance propagation (LRP) is used here to enhance the interpretability of the model and we visualize and analyze the reasons for the different correlation between traffic density and PM2.5 concentration in various regions of Shanghai. Meanwhile, As the strict and effective industrial emission reduction measurements implementing in China, the contribution of urban traffic to PM2.5 formation calculated by combining MEIC emission inventory and LRP is gradually increasing from 18.03% in 2011 to 24.37% in 2017 in Shanghai, and the impact of traffic emissions would be ever-prominent in 2030 according to our prediction. We also infer that the promotion of vehicular electrification would achieve further alleviation of PM2.5 about 8.45% by 2030 gradually. These insights are of great significance to provide the decision-making basis for accurate and high-efficient traffic management and urban pollution control, and eventually benefit people's lives and high-quality sustainable developments of cities.

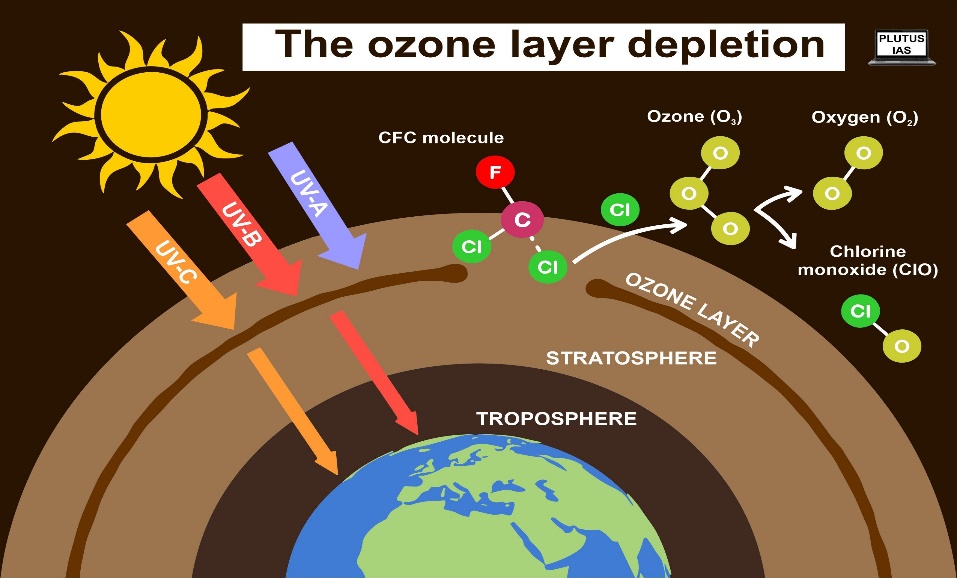
[49] With urbanization and increasing consumption, there is a growing need to prioritize sustainable development across various industries. Particularly, sustainable development is hindered by air pollution, which poses a threat to both living organisms and the environment. The emission of combustion gases containing particulate matter (PM 2.5) during human and social activities is a major cause of air pollution. To mitigate health risks, it is crucial to have accurate and reliable methods for forecasting PM 2.5 levels. In this study, we propose a novel approach that combines support vector machine (SVM) and long short-term memory (LSTM) with complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) to forecast PM 2.5 concentrations. The methodology involves extracting Intrinsic mode function (IMF) components through CEEMDAN and subsequently applying different regression models (SVM and LSTM) to forecast each component. The Naive Evolution algorithm is employed to determine the optimal parameters for combining CEEMDAN, SVM, and LSTM. Daily PM 2.5 concentrations in Kaohsiung, Taiwan from 2019 to 2021 were collected to train models and evaluate their performance. The performance of the proposed model is evaluated using metrics such as mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), and coefficient of determination (R2) for each district. Overall, our proposed model demonstrates superior performance in terms of MAE (1.858), MSE (7.2449), RMSE (2.6682), and (0.9169) values compared to other methods for 1-day ahead PM 2.5 forecasting.

**CHAPTER 3**

**AIR POLLUTION**

**Air pollution:-**

Air pollution refers to the presence of harmful substances in the air that can adversely affect human health, the environment, and the climate. These pollutants can include particulate matter (like dust and soot), gases (such as carbon monoxide, sulphur dioxide, and nitrogen oxides), ozone, and volatile organic compounds. Sources of air pollution vary widely and can include industrial emissions, vehicle exhaust, agricultural activities, and natural sources like wildfires and volcanic eruptions. Air pollution is a significant global issue with widespread implications for public health, ecosystems, and the economy.

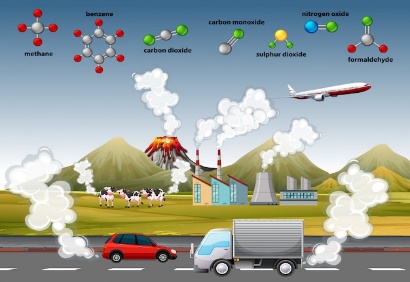


**Fig:3.1 Ozone layer depletion**

Air pollution occurs when harmful substances such as particulate matter, gases (like carbon monoxide and sulphur dioxide), ozone, and volatile organic compounds are released into the atmosphere. These pollutants can come from various sources such as vehicles, industrial activities, agriculture, and natural events like wildfires. They pose significant health risks to humans and animals, contribute to climate change, and can harm ecosystems and vegetation. Efforts to reduce air pollution involve regulatory measures, technological advancements, and changes in behaviour to mitigate emissions and improve air quality.

**Types of air pollution:-**

**1.Particulate Matter (PM)**: These are tiny particles suspended in the air, including dust, dirt, soot, and liquid droplets. PM can vary in size from coarse particles (PM10) to fine particles (PM2.5), which are small enough to penetrate deep into the lungs and cause respiratory issues.



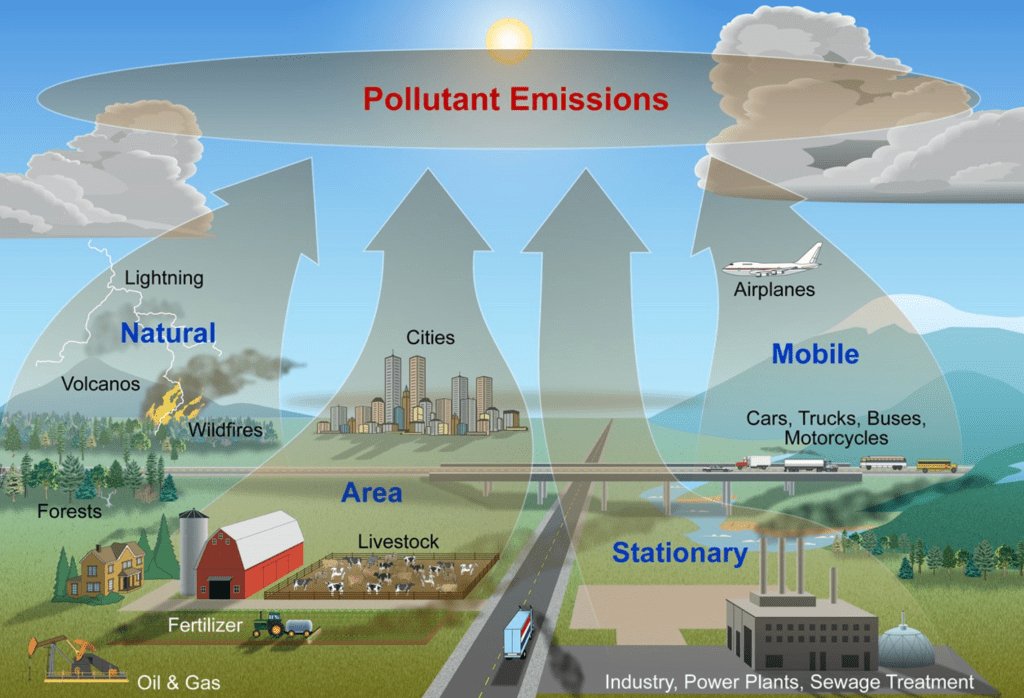
Top of Form

Bottom of Form

**Fig :3.2 Increased atmosphere CO2**

**2.Gaseous Pollutants**: This category includes pollutants such as carbon monoxide (CO), sulphur dioxide (SO2), nitrogen oxides (NOx), ozone (O3), and volatile organic compounds (VOCs). These gases can have direct health impacts and contribute to the formation of other pollutants like smog and secondary aerosols.

**3.Ozone (O3)**: Ground-level ozone is a secondary pollutant formed through complex chemical reactions involving NOx and VOCs in the presence of sunlight. It can irritate the respiratory system, exacerbate asthma, and damage crops.



**Fig:3.3 Pollutant Emission**

4. **Heavy Metals**: Pollutants like lead, mercury, and cadmium are toxic substances emitted from industrial processes and vehicle exhaust. They can accumulate in the environment and pose serious health risks, including neurological and developmental disorders.

5.**Persistent Organic Pollutants (POPs)**: These are organic compounds that resist environmental degradation and bioaccumulate in ecosystems and organisms. Examples include polychlorinated biphenyls (PCBs) and dioxins, which can have long-term health effects and ecological impacts.

6.**Radon**: A radioactive gas naturally occurring from the decay of uranium in soil and rocks.It

can seep into buildings and accumulate to levels hazardous to health.

**7.Indoor Air Pollutants**: Sources include tobacco smoke, household products (like paints and cleaners), building materials (like asbestos and formaldehyde), and indoor combustion appliances (like stoves and heaters). Poor indoor air quality can cause respiratory problems

and exacerbate allergies and asthma.

Each type of air pollutant has distinct sources, behaviour in the atmosphere, and impacts on human health, ecosystems, and the environment. Addressing air pollution requires understanding these sources and implementing strategies to mitigate emissions and improve air quality.Top of FormBottom of Form

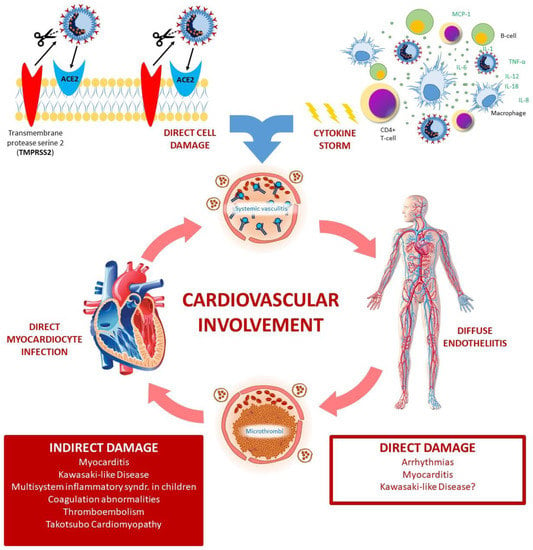
**Problems of Air pollution:-**Air pollution poses several significant problems for both human health and the environment:

**Fig :3.4 Indoor air problems**

**1.Respiratory and Cardiovascular Health Issues**: Exposure to air pollutants such as particulate matter (PM), ozone (O3), nitrogen dioxide (NO2), and sulphur dioxide (SO2) can lead to respiratory diseases such as asthma, bronchitis, and lung cancer. These pollutants can also aggravate cardiovascular conditions, leading to heart attacks and strokes.

**Fig:3.5 Persistent Organic Pollutants**

2.**Environmental Degradation**: Air pollution harms ecosystems by damaging plants, reducing crop yields, and disrupting natural habitats. Acid rain, caused by sulphur dioxide and nitrogen oxides reacting with water vapor in the atmosphere, further damages forests, lakes, and soils.

**3.Climate Change**: Certain air pollutants, like carbon dioxide (CO2), methane (CH4), and black carbon (soot), contribute to global warming and climate change. These pollutants trap heat in the atmosphere, leading to rising temperatures, melting glaciers, sea-level rise, and more frequent extreme weather events.

**Fig :3.6 Cardiovascular involvement**

**4.Economic Costs**: Air pollution results in significant economic losses due to healthcare expenditures for treating pollution-related illnesses, reduced worker productivity, damage to crops and forests, and increased energy consumption for air quality management and remediation.



**Fig :3.7 The cost of climate change**

5. **Social Impacts**: Air pollution disproportionately affects vulnerable populations such as children, the elderly, and low-income communities who may live near sources of pollution or lack access to healthcare. It also contributes to social inequalities by exacerbating health disparities.

**6.Visibility and Aesthetic Issues**: Particulate matter and other pollutants can reduce visibility, especially in urban areas and scenic landscapes, impacting tourism and quality of life.

Addressing air pollution requires comprehensive policies and strategies to reduce emissions from sources such as transportation, industry, agriculture, and energy production. Improving air quality not only protects human health but also preserves ecosystems and mitigates climate change impacts.

**Fig:3.8 Environment Degradation**

**Top of Form**

**Bottom of Form**

**Prevention measures:-**

Preventing air pollution involves a range of strategies aimed at reducing emissions of pollutants into the atmosphere. Here are some key prevention measures:

1.**Regulations and Standards**: Governments can establish and enforce air quality standards and regulations that limit emissions from industrial facilities, power plants, vehicles, and other sources. These regulations often include emission limits, technology requirements (such as catalytic converters for vehicles), and enforcement mechanisms to ensure compliance.

2.**Promotion of Clean and Renewable Energy**: Encouraging the use of clean energy sources such as wind, solar, hydroelectric, and nuclear power can reduce reliance on fossil fuels and decrease emissions of greenhouse gases and other pollutants.

3.**Improved Transportation Policies**: Implementing measures to reduce vehicle emissions, such as promoting public transportation, investing in electric and hybrid vehicles, improving fuel efficiency standards, and encouraging carpooling and cycling.

4.**Technological Advancements**: Developing and deploying technologies that reduce emissions from industrial processes, power generation, and transportation, such as scrubbers for smokestacks, low-emission vehicles, and cleaner production techniques.

**5.Waste Management**: Proper management of solid and hazardous waste to prevent open burning and reduce methane emissions from landfills.

6.**Education and Awareness**: Raising awareness about the health and environmental impacts of air pollution and promoting behaviours that reduce emissions, such as energy conservation and proper disposal of pollutants.

7.**International Cooperation**: Collaborating across borders to address transboundary air pollution issues, such as acid rain and cross-border transport of pollutants, through agreements and partnerships.

8.**Green Building Practices**: Designing and constructing buildings that are energy-efficient and incorporate materials with low emissions of volatile organic compounds (VOCs) and other pollutants.

9.**Forest and Land Management**: Protecting and restoring forests, wetlands, and other natural habitats that act as sinks for pollutants and help maintain air quality.

**10.Monitoring and Research**: Continuously monitoring air quality, conducting research on the sources and impacts of air pollution, and using data to inform policy-making and improve prevention strategies.

By implementing these prevention measures comprehensively and collaboratively, societies can work towards reducing air pollution and its associated impacts on human health, ecosystems, and the climate.

**FUTURE SCOPE**

The future scope of an air pollution prediction project is promising and multifaceted, with several areas of potential development and application:

1. **Enhanced Accuracy and Models**: Continuously improving predictive models using advanced machine learning algorithms (such as deep learning, ensemble methods) and incorporating more comprehensive datasets (including meteorological data, satellite imagery, and real-time sensor data) can enhance accuracy.
2. **Real-Time Prediction**: Developing capabilities for real-time or near-real-time prediction of air quality indices (AQI) can enable proactive measures to mitigate pollution impacts.
3. **Spatial Resolution**: Increasing the spatial resolution of predictions to provide localized forecasts can be beneficial for urban planning, public health interventions, and policy-making at a city or neighborhood level.
4. **Integration with IoT and Sensor Networks**: Integrating with Internet of Things (IoT) devices and sensor networks to gather continuous data and validate predictive models can improve reliability and responsiveness.
5. **Predictive Analytics for Health Impacts**: Linking air quality predictions with health data to forecast potential health impacts and provide early warnings to vulnerable populations.
6. **User-Friendly Interfaces**: Developing user-friendly interfaces (such as mobile apps or web dashboards) that deliver personalized air quality forecasts and recommendations to the public.
7. **Policy Support**: Supporting policymakers with predictive insights to design effective environmental policies and regulations aimed at reducing air pollution.
8. **Climate Change Adaptation**: Studying the interaction between air quality and climate change to anticipate future trends and adapt mitigation strategies accordingly.
9. **Cross-Domain Applications**: Exploring cross-domain applications, such as integrating air quality predictions with traffic management systems, urban planning, and renewable energy production, to optimize environmental sustainability.
10. **Global Collaboration**: Encouraging international collaboration and data sharing to address transboundary air pollution issues and develop globally applicable solutions.
11. **Ethical Considerations**: Addressing ethical considerations related to data privacy, equity in access to air quality information, and transparency in algorithmic decision-making.

Overall, the future of air pollution prediction projects lies in advancing technology, interdisciplinary collaboration, and leveraging data-driven insights to foster healthier and sustainable environments for communities worldwide.

**PROPOSED METHODOLOGIES**

The algorithms used to train the model are LinearRegression, Logistic Regression, Navie Bayes, SVM and Random Forest classifier

**LINEAR REGRESSION**

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data.

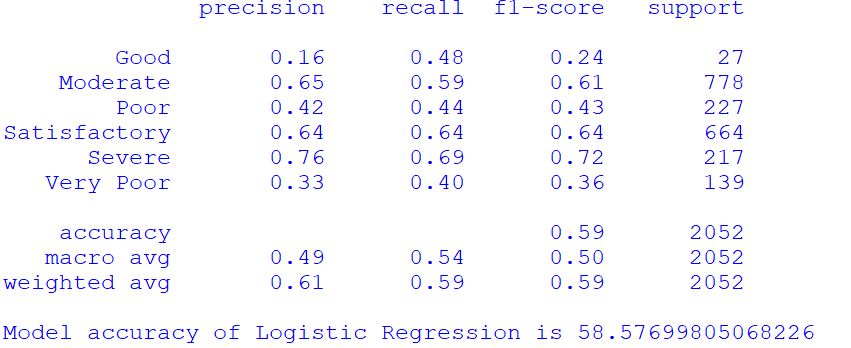
When there is only one independent feature, it is known as Simple Linear Regression, and when there are more than one feature, it is known as Multiple Linear Regression.

By applying this linear regression we got accuracy of **22.18916203256431**

**LOGISTIC REGRESSION**

Logistic regression is a data analysis technique that uses mathematics to find the relationships between two data factors. It then uses this relationship to predict the value of one of those factors based on the other. The prediction usually has a finite number of outcomes, like yes or no.

By applying this we get the accuracy of **58.57699805068226**

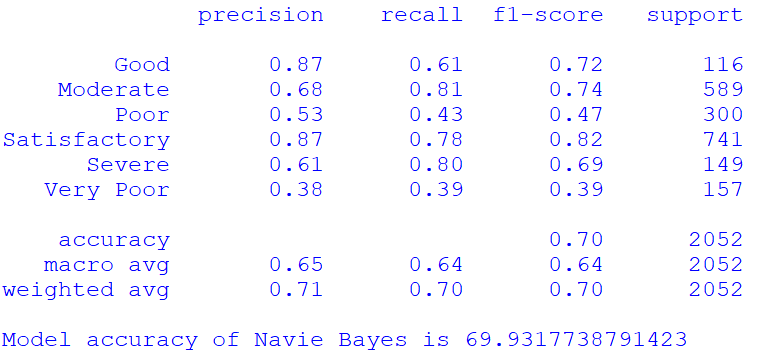


**Fig:3.9 Accuracy of logistic regression**

**NAVIE BAYES**

Naïve Bayes is part of a family of generative learning algorithms, meaning that it seeks to model the distribution of inputs of a given class or category. Unlike discriminative classifiers, like logistic regression, it does not learn which features are most important to differentiate between classes.

By applying Navie Bayes we get the accuracy of **69.9317738791423**

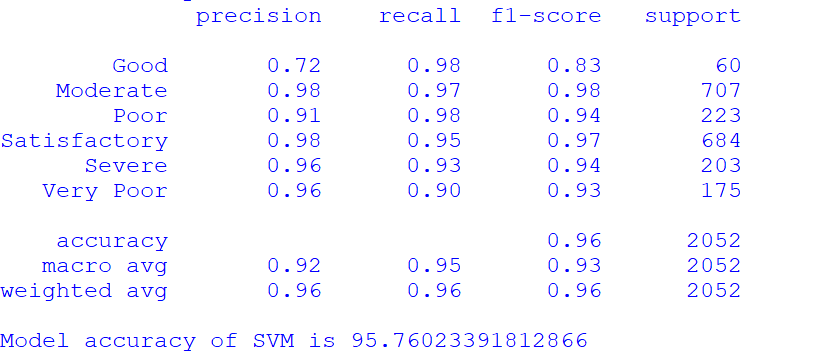


**Fig:3.10 Accuracy of Navie Bayes**

**SVM**

Support Vector Machine (SVM) is a powerful machine learning algorithm used for linear or nonlinear classification, regression, and even outlier detection tasks. SVMs can be used for a variety of tasks, such as text classification, image classification, spam detection, handwriting identification, gene expression analysis, face detection, and anomaly detection. SVMs are adaptable and efficient in a variety of applications because they can manage high-dimensional data and nonlinear relationships.

By applying SVM we get the accuracy of **95.76023391812866**



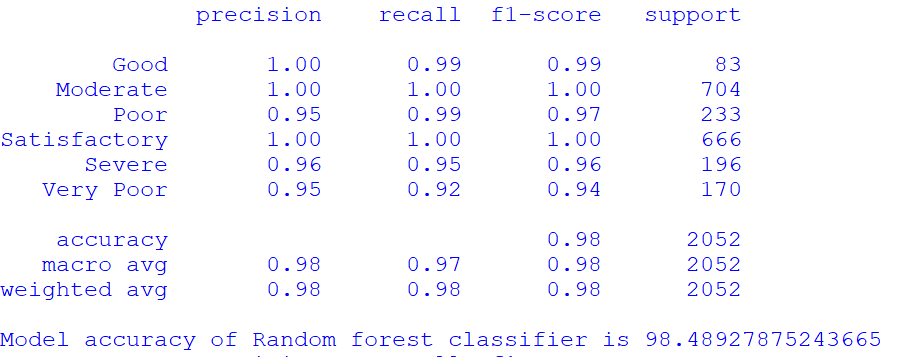
**Fig:3.11 Accuracy of SVM**

**RANDOM FOREST CLASSIFIER**

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning**,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

By applying this we get the accuracy of **98.48927875243665**



**Fig:3.12 Accuracy of Random Forest**

Since the accuracy of the Random forest is greater than all other algorithms, we select the Random forest for training and building the model.

**CHAPTER 4**

**SOFTWARE ENVIRONMENT**

**4.1 INTRODUCTION TO PYTHON**

Python is a high-level, interpreted scripting language developed in the late 1980s by Guido van Rossum at the National Research Institute for Mathematics and Computer Science in the Netherlands. The initial version was published at the alt. Sources newsgroup in 1991, and version 1.0 was released in 1994.

Python 2.0 was released in 2000, and the 2.x versions were the prevalent releases until December 2008. At that time, the development team made the decision to release version 3.0, which contained a few relatively small but significant changes that were not backward compatible with the 2.x versions. Python 2 and 3 are very similar, and some features of Python 3 have been back ported to Python 2. But in general, they remain not quite compatible.

Both Python 2 and 3 have continued to be maintained and developed, with periodic release updates for both. As of this writing, the most recent versions available are 2.7.15 and 3.6.5. However, an official End of Life date of January 1, 2020 has been established for Python 2, after which time it will no longer be maintained. If you are a newcomer to Python, it is recommended that you focus on Python 3, as this tutorial will do.

Python is still maintained by a core development team at the Institute, and Guido is still in charge, having been given the title of BDFL (Benevolent Dictator For Life) by the Python community. The name Python, by the way, derives not from the snake, but from the British comedy troupe Monty Python’s Flying Circus, of which Guido was, and presumably still is, a fan. It is common to find references to Monty Python sketches and movies scattered throughout the Python documentation

**WHY CHOOSE PYTHON**

For some applications that are particularly computationally intensive like graphics processing or intense number, this can be limiting. In practice,however,for most programs,the difference execution speed is measured in milliseconds, or seconds at most, and not appreciably noticeable to a human user.The expediency of coding in an interpreted language is typically worth it for most applications.

**Python is Free**

The Python interpreter is developed under an OSI-approved opensource license making it free to install use and distribute even for commercial purposes. A version of the interpreter is available for virtually any platform there is, including all flavors of Unix, Windows ,macOS, smart phones and tablets, and probably anything else you ever heard. A version even exists for the half dozen people remaining who use OS/2.

**Python is Portable**

Because Python code is interpreted and not compiled into native machine instructions, code written for one platform will work on any other platform that has the Python interpreter installed. (This is true of any interpreted language, not just Python.)

**Python is Simple**

Python-3 has 33 keywords, and Python-2 has 31.By contrast, C++ has 62, Java has 53, and Visual Basic has more than 120, though these latter examples probably vary Somewhat by implementation or dialect. Python code has a simple and clean structure that is easy to learn and easy to read.In fact,as you will see,the language definition enforces code structure that is easy to read. But It’s Not That Simple for all its syntactical simplicity, Python supports most constructs that would be expected in a very high-level language,including complex dynamic data types, structured and functional programming, and object-oriented programming. Additionally, a very extensive library of classes and functions is available that provides capability well beyond what is built into the language,such as database manipulation. Python accomplishes what many programming languages don’t: the language itself is simply designed, but it is very versatile in terms of what you can accomplish with it.

**Conclusion**

This section gave an overview of the Python programming language, including:

A brief history of the development of Python Some reasons why you might select Python as your language of choice.

Python is a great option, whether you are a beginning programmer looking to learn the basics, an experienced programmer designing a large application, or anywhere in between. The basics of Python are easily grasped, and yet its capabilities are vast. Proceed to the next section to learn how to acquire and install Python on your computer.

Python is an open-source programming language that was made to be easy-to-read and powerful. A Dutch programmer named Guido van Rossum made Python in 1991. He named it after the television show Monty Python's Flying Circus. Many Python.

Python is an interpreted language. Interpreted languages do not need to be compiled to run. A program called an interpreter runs Python code on almost any kind of computer. This means that a programmer can change the code and quickly see the results. This also means Python is slower than a compiled language like C, because it is not running machine code directly.

Python is a good programming language for beginners. It is a highlevel language, which means a programmer can focus on what to do instead of how to do it.

Writing programs in Python takes less time than in some other languages. Python drew inspiration from other programming languages like C, C++, Java, Perl, and Lisp. 32

Python has a very easy-to-read syntax. Some of Python's syntax comes from C, because that is the language that Python was written in. But Python uses whitespace to delimit code: spaces or tabs are used to organize code into groups. This is different from C. In C, there is a semicolon at the end of each line and curly braces ({}) are used to group code. Using whitespace to delimit code makes Python a very easy-to-read language.

**Python use [change / change source]**

Python is used by hundreds of thousands of programmers and is used in many places. Sometimes only Python code is used for a program, but most of the time it is used to do simple jobs while another programming language is used to do more complicated tasks. Its standard library is made up of many functions that come with Python when it is installed. On the Internet there are many other libraries available that make it possible for the Python language to do more things. These libraries make it a powerful language; it can do many different things.

Some things that Python is often used for are:

• Web development

• Scientific programming

• Desktop GUIs • Network programming

• Game programming

4.3 HARDWARE REQUIREMENTS

Windows 11 system Requirement

* 16GB of RAM
* 64GB of storage
* 64-bit processor
* 1GHz CPU clock speed
* An internet connection and a Microsoft account(for the initial setup)

4.4 SOFTWARE REQUIREMENTS

Latest OS:22631.3880

Edition: Windows 11 Home Single Language

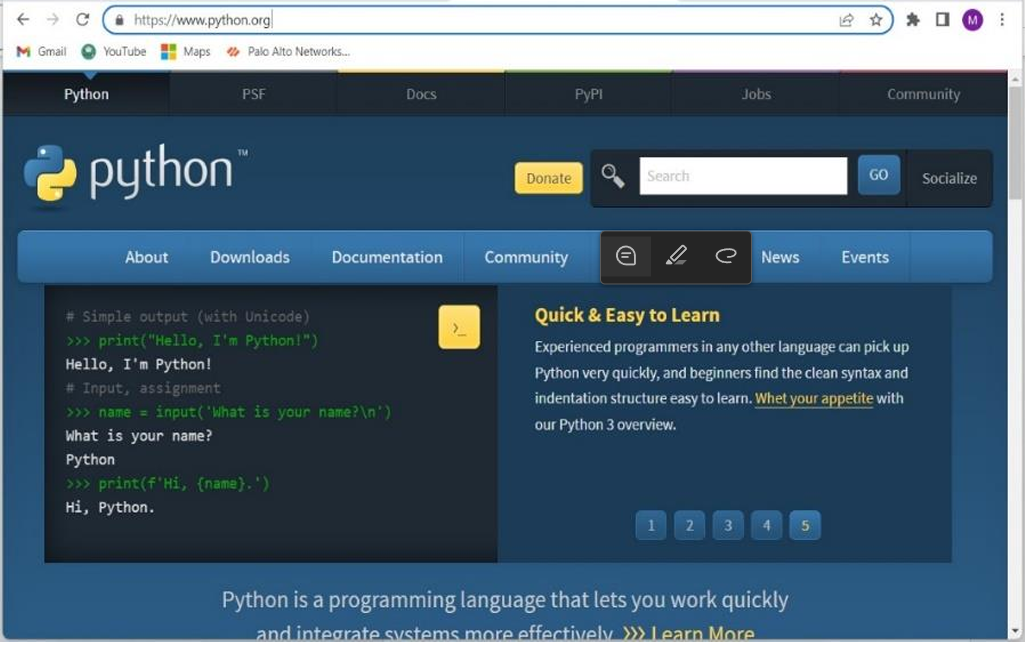
Version:23H2

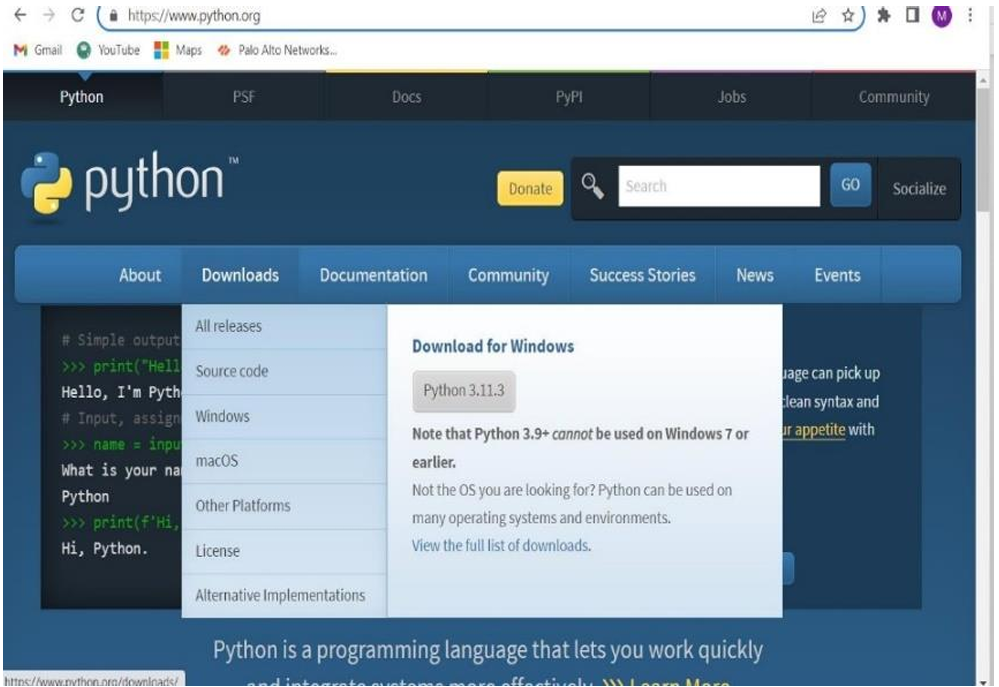
RAM: 16.0 GB (15.6 GB usable)

Processor: 13th Gen Intel(R) Core(TM) i7-1360P 2.20 GHz

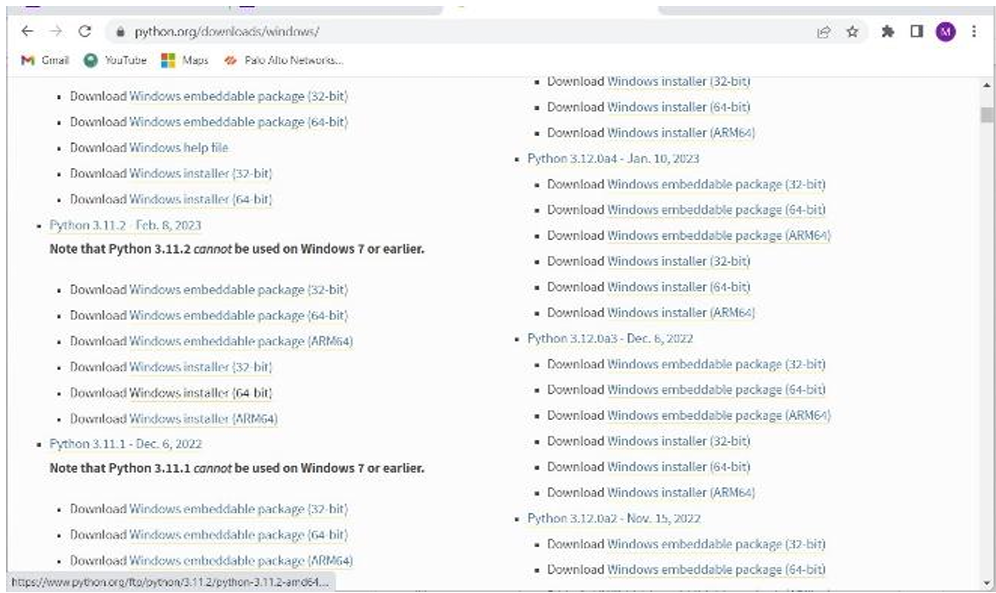
System type: 64-bit operating system, x64-based processor

**4.5 STEPS TO INSTALL PYTHONStep 1:** Search python.org

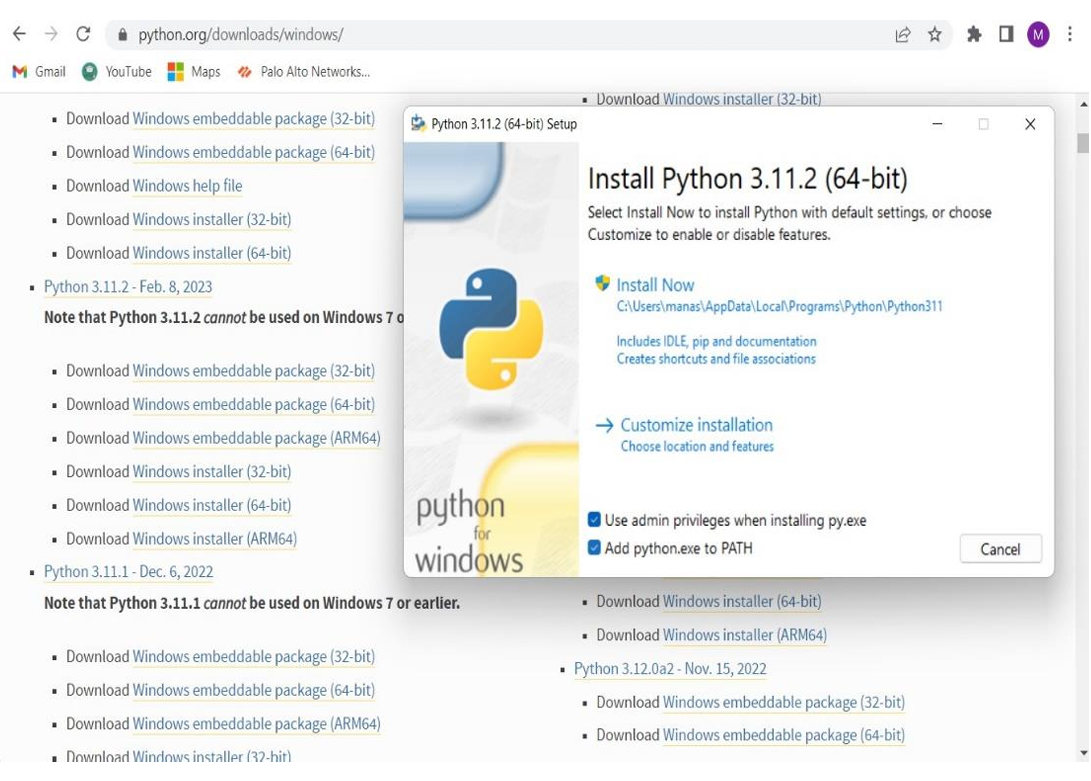


**Step 2:** Go to downloads and select windows

**Step 3:** Download Windows installer(64-bit)



**Ste 4:** Now select python.exe to path and install the IDLE



**Modules**

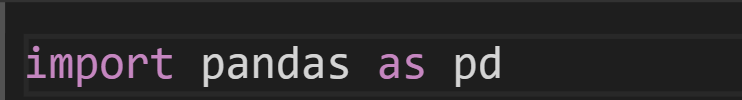
**Pandas :**

pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real-world data analysis in Python. Additionally, it has the broader goal of becoming the most powerful and flexible open source data analysis/manipulation tool available in any language. It is already well on its way toward this goal.

The two primary data structures of pandas, Series (1-dimensional) and DataFrame (2-dimensional), handle the vast majority of typical use cases in finance, statistics, social science, and many areas of engineering. For R users, DataFrame provides everything that R’s data.frame provides and much more. pandas is built on top of NumPy and is intended to integrate well within a scientific computing environment with many other 3rd party libraries.

Many of these principles are here to address the shortcomings frequently experienced using other languages / scientific research environments. For data scientists, working with data is typically divided into multiple stages: munging and cleaning data, analyzing / modeling it, then organizing the results of the analysis into a form suitable for plotting or tabular display. pandas is the ideal tool for all of these tasks

Here is that how we import pandas from the python Libraries



**Numpy :**

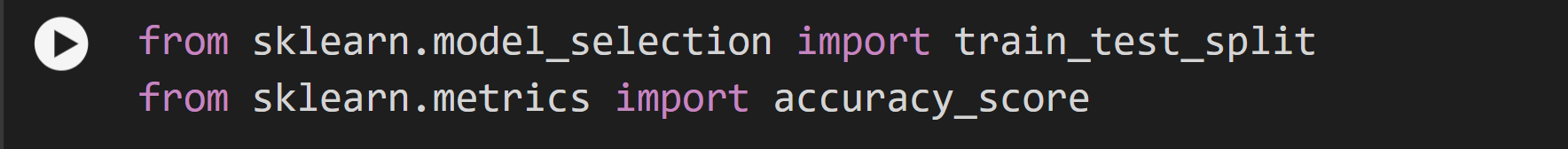
NumPy is a Python library used for working with arrays.It also has functions for working in domain of linear algebra, fourier transform, and matrices. NumPy stands for Numerical Python.

In Python we have lists that serve the purpose of arrays, but they are slow to process.NumPy aims to provide an array object that is up to 50x faster than traditional Python lists.The array object in NumPy is called ndarray, it provides a lot of supporting functions that make working with ndarray very easy.Arrays are very frequently used in data science, where speed and resources are very important.

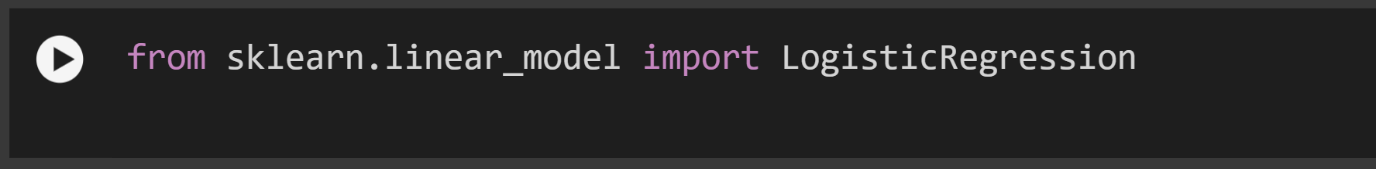
**Scikit-learn :**

scikit-learn (often abbreviated as sklearn) is a popular open-source machine learning library for Python. It provides simple and efficient tools for data mining and data analysis, and it is built on NumPy, SciPy, and Matplotlib. Scikit-learn is widely used for implementing and experimenting with a variety of machine learning models and algorithms.

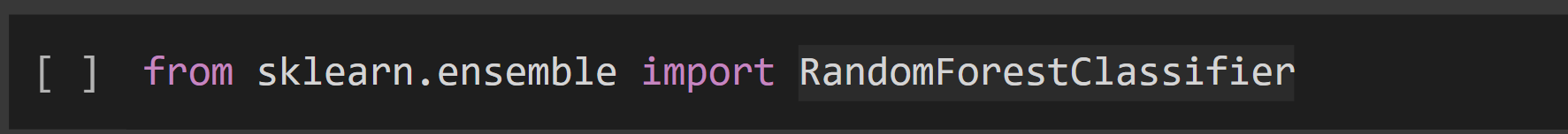
Here is that how we import sklearn from the python Libraries



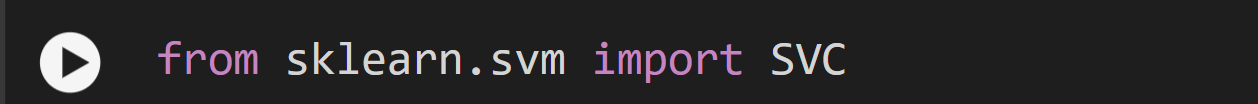
Here is that how we import sklearn for logistic regression



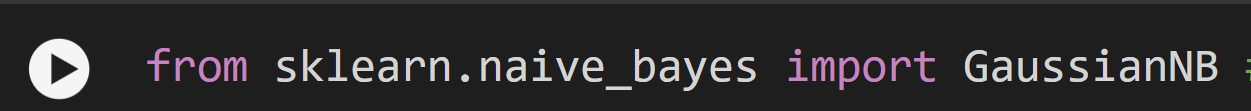
Here is that how we import sklearn for RandomForestClassifier



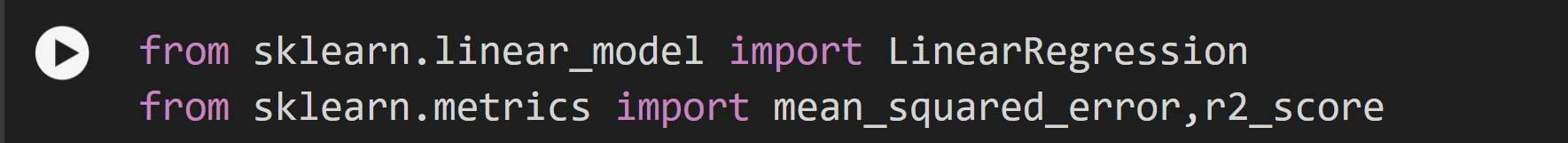
Here is that how we import sklearn for SVM(SupportVectorMachine)



Here is that how we import sklearn for naïve bayes



Here is that how we import sklearn for linear regression



**CHAPTER 5**

**ABOUT FLASK**

Flask is a web framework that allows developers to build lightweight web applications quickly and easily with Flask Libraries. It was developed by Armin Ronacher, leader of the International Group of Python Enthusiasts(POCCO). It is basically based on the WSGI toolkit and Jinja2 templating engine.

Flask is an API of Python that allows us to build web applications. It was developed by Armin Ronacher. Flask’s framework is more explicit than Django’s framework and is also easier to learn because it has less base code to implement a simple web application. Flask Python is based on the WSGI(Web Server Gateway Interface) toolkit and Jinja2 template engine.

**Advantages of Flask**

1. Flask is a **lightweight** backend framework with minimal dependencies.
2. Flask is **easy to learn** because its simple and intuitive API makes it easy to learn and use for beginners.
3. Flask is a **flexible** **Framework** because it allows you to customize and extend the framework to suit your needs easily.
4. Flask can be used with **any database** like:- SQL and NoSQL and with **any Frontend Technology** such as React or Angular.
5. Flask is **great for small to medium projects** that do not require the complexity of a large framework.
6. Flask Documentation

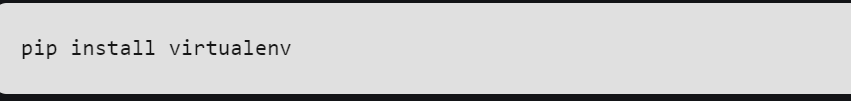
Flask is basically a Python module. It can work with **Python** only and it is a web-developing framework. It is a collection of libraries and modules. [Frameworks](https://www.geeksforgeeks.org/top-10-frameworks-for-web-applications/)are used for developing web platforms. Flask is such a type of **web application** framework. It is completely written in Python language. Unlike [Django](https://www.geeksforgeeks.org/django-tutorial/), it is only written in [Python](https://www.geeksforgeeks.org/python-programming-language/). As a new user, [Flask](https://www.geeksforgeeks.org/python-introduction-to-web-development-using-flask/) is to be used. As it is easier to handle. As it is only written in Python, before installing Flask on the machine, Python should be installed previously.

**Features of Python Flask:**

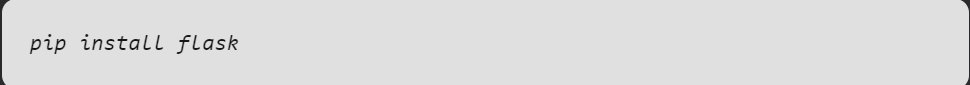
* Flask is easy to use and easily understandable for new users in Web Framework.
* It can also be used as any third-party plugin extension.
* It is also used for prototyping purposes.

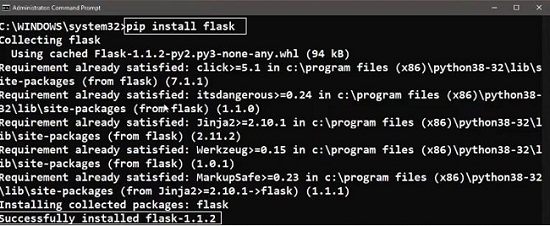
**Install Virtual Environment**

We use a module named virtualenv which is a tool to create isolated Python environments. [virtualenv](https://www.geeksforgeeks.org/creating-python-virtual-environment-windows-linux/) creates a folder that contains all the necessary executables to use the packages that a Python project would need.

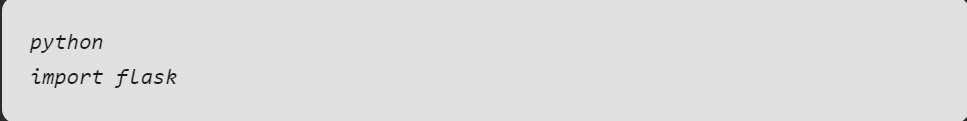


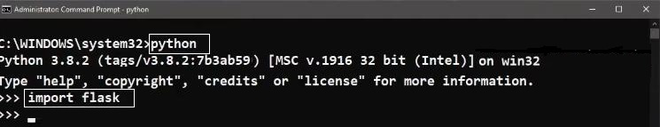
**Step 2:**At first, open the command prompt in administrator mode. Then the following command should be run. This command will help to install Flask using Pip in Python and will take very less time to install. According to the machine configuration, a proper Flask version should be installed. Wait for some time till the process is completed. After completionof the process, Flask is completed successfully, the message will be displayed. Hence Installation is successful*.*





**Step 3:**After that, also the following two commands should be run. These commands will start Flask in the command prompt. Hence, the process is completed successfully.

**

Top of Form

**Chapter 6**

**Working with Flask**

**6.1 Flask work in Python**

Flask uses a specific syntax to create links from a page to another . This is fact generates the link dynamically according to the decorator set to the function linked to. In addition it takes care of where the application is deployed

Flask is a web framework. This means flask provides you with tools, libraries and technologies that allow you to build a web application. This web application can be some web pages, a blog, a wiki or go as big as a web-based calendar application or a commercial website.

Flask is part of the categories of the micro-framework. Micro-framework are normally framework with little to no dependencies to external libraries. This has pros and cons. Pros would be that the framework is light, there are little dependency to update and watch for security bugs, cons is that some time you will have to do more work by yourself or increase yourself the list of dependencies by adding plugins. In the case of Flask, its dependencies a falsk.

**6.2 Template engine**

Have you ever built a website? Did you face the problem that to keep the style of the website consistent, you have had to write multiple times the same text? Did you ever tried to change the style of such website?

If your website contains only few pages, changing its style will take you some time but is doable. However, if you have a lot of pages (for example the list of items you sell in your store), this task become overwhelming.

Using templates you are able to set a basic layout for your pages and mention which element will change. This way you can define your header once and keep it consistent over all the pages of your website, and if you need to change your header, you will only have to update it in one place.

Using a template engine will save you a lot of time when creating your application but also when updating and maintaining it.

We are going to perform a very basic application with flask.

**CHAPTER 7**

**PYTHON CODE**

**VIEW CODE**

import pandas as pd

import joblib

data=pd. read\_ csv('Air\_Pollution\_Datasets.csv')

print(data)

data\_ cleaned=data. drop (columns= ['MID', 'Unnamed: 17', 'Unnamed: 18', 'Unnamed: 19', 'Unnamed: 20'])

print (data\_ cleaned)

columns\_ to\_ fill = ['PM2.5', 'PM10', 'NO', 'NO2', 'NOx','NH3', 'CO', 'SO2', 'O3', 'Benzene', 'Toluene', 'Xylene', 'AQI']

data\_ cleaned [columns\_ to\_ fill] = data\_ cleaned [columns\_ to\_ fill]. fillna (data\_ cleaned [columns\_ to\_ fill]. mean ())

print (data\_ cleaned. columns)

print (data\_ cleaned. isnull(). sum())

from sklearn. model\_ selection import train\_test\_split

from sklearn. metrics import accuracy\_score,classification\_report

features = data\_ cleaned [['PM2.5', 'PM10', 'NO', 'NO2', 'NOx', 'NH3','CO', 'SO2', 'O3', 'Benzene', 'Toluene', 'Xylene', 'AQI']]

target = data\_ cleaned ['AQI\_ Bucket']

x\_train, x\_test, y\_train,y\_test=train\_test\_split(features, target, test\_size=0.2,random\_state=42)

from sklearn. linear\_model import LogisticRegression

model=LogisticRegression()

model.fit(x\_train ,y\_train)

y\_pred=model.predict (x\_test)

dd=classification\_report (y\_pred, y \_test)

print(dd)

print ("Model accuracy of Logistic Regression is", accuracy\_ score(y\_test,y\_pred)\*100)

from sklearn.ensemble import RandomForestClassifier

model1=RandomForestClassifier()

model1.fit(x\_train,y\_train)

test\_pred1=model1.predict(x\_test)

aa=classification\_report(test\_pred1,y\_test)

print(aa)

print ("Model accuracy of Random forest classifier is", accuracy\_score(y\_test,test\_pred1)\*100)

from sklearn.svm import SVC

model2=SVC()

model2.fit(x\_train,y\_train)

pred=model2.predict(x\_test)

cc=classification\_report(pred ,y\_test)

print(cc)

print ("Model accuracy of SVM is",accuracy\_score(y\_test,pred)\*100)

from sklearn.naive\_bayes import GaussianNB # Fixed the typo in module name

model3 = GaussianNB()

model3.fit(x\_train, y\_train)

pred1=model3.predict(x\_test)

bb=classification\_report(pred1,y\_test)

print(bb)

print("Model accuracy of Navie Bayes is",accuracy\_score(y\_test,pred1)\*100)

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import mean\_squared\_error,r2\_score

from sklearn.preprocessing import LabelEncoder

label\_encoder = LabelEncoder()

y\_train\_encoded = label\_encoder.fit\_transform(y\_train)

y\_test\_encoded = label\_encoder.transform(y\_test)

model4=LinearRegression()

model4.fit(x\_train,y\_train\_encoded)

pr=model4.predict(x\_test)

r2=r2\_score(y\_test\_encoded, pr)

print(f"R-squared:{r2\*100}")

ee=classification\_report(pr,y\_test)

print(ee)

from sklearn.preprocessing import LabelEncoder

encoder=LabelEncoder()

data['AQI\_Bucket']=encoder.fit\_transform(data['AQI\_Bucket'])

data.head()

joblib.dump(model1,"my\_model.h5")

This code will train the model using the dataset and then it will predict the output of the data given.

**PREDICT**

import joblib

import pandas as pd

# Load the model

model = joblib.load("my\_model.h5")

# Input values for features

a = float(input("Enter the value of PM2.5: "))

b = float(input("Enter the value of PM10: "))

c = float(input("Enter the value of NO: "))

d = float(input("Enter the value of NO2: "))

e = float(input("Enter the value of NOx: "))

f = float(input("Enter the value of NH3: "))

g = float(input("Enter the value of CO: "))

h = float(input("Enter the value of SO2: "))

i = float(input("Enter the value of O3: "))

j = float(input("Enter the value of Benzene: "))

k = float(input("Enter the value of Toluene: "))

l = float(input("Enter the value of Xylene: "))

m = float(input("Enter the value of AQI: "))

# Create a dictionary with input values

d1 = {

"PM2.5": a, "PM10": b, "NO": c, "NO2": d, "NOx": e,

"NH3": f, "CO": g, "SO2": h, "O3": i, "Benzene": j,

"Toluene": k, "Xylene": l, "AQI": m}

# Create a DataFrame from the dictionary

d2 = pd.DataFrame([d1])

# Assuming d2 already has the correct column names, proceed to predict

p = model.predict(d2)

print(p)

**INIT**

from flask import Flask, render\_template, request, redirect, url\_for

import os

import sys

import pandas as pd # type: ignore

import joblib

app = Flask(\_\_name\_\_)

if sys.stderr is None:

sys.stderr=open(os.devnull,'w')

@app.route('/')

def index():

return render\_template('index.html')

@app.route('/submit', methods=['POST'])

def submit():

try:

model=joblib.load("my\_model.h5")

a=float(request.form['PM2.5'])

b=float(request.form['PM10'])

c=float(request.form['NO'])

d=float(request.form['NO2'])

e=float(request.form['NOx'])

f1=float(request.form['NH3'])

g=float(request.form['CO'])

h=float(request.form['SO2'])

i=float(request.form['O3'])

j=float(request.form['Benzene'])

k=float(request.form['Toluene'])

l=float(request.form['Xylene'])

m=int(request.form['AQI'])

d1 = {'PM2.5': a, 'PM10': b, 'NO': c, 'NO2': d, 'NOx': e,'NH3': f1, 'CO': g, 'SO2': h, 'O3': i, 'Benzene': j,'Toluene': k, 'Xylene': l, 'AQI': m}

# Create a DataFrame from the dictionary

d2 = pd.DataFrame([d1])

# Assuming d2 already has the correct column names, proceed to predict

result = model.predict(d2)

except ValueError:

result = "Invalid input! Please enter numbers only."

return redirect(url\_for('result', result=result))

@app.route('/result')

def result():

result = request.args.get('result')

return render\_template('result.html', result=result)

if \_\_name\_\_ == '\_\_main\_\_':

app.run(debug=True,port=8080)

**CHAPTER-8**

**FINAL RESULT**

**OUTPUT**

**ACTIVATING SERVER**



Fig:8.1 Open the main code and copy the IP address

**WEB PAGE**

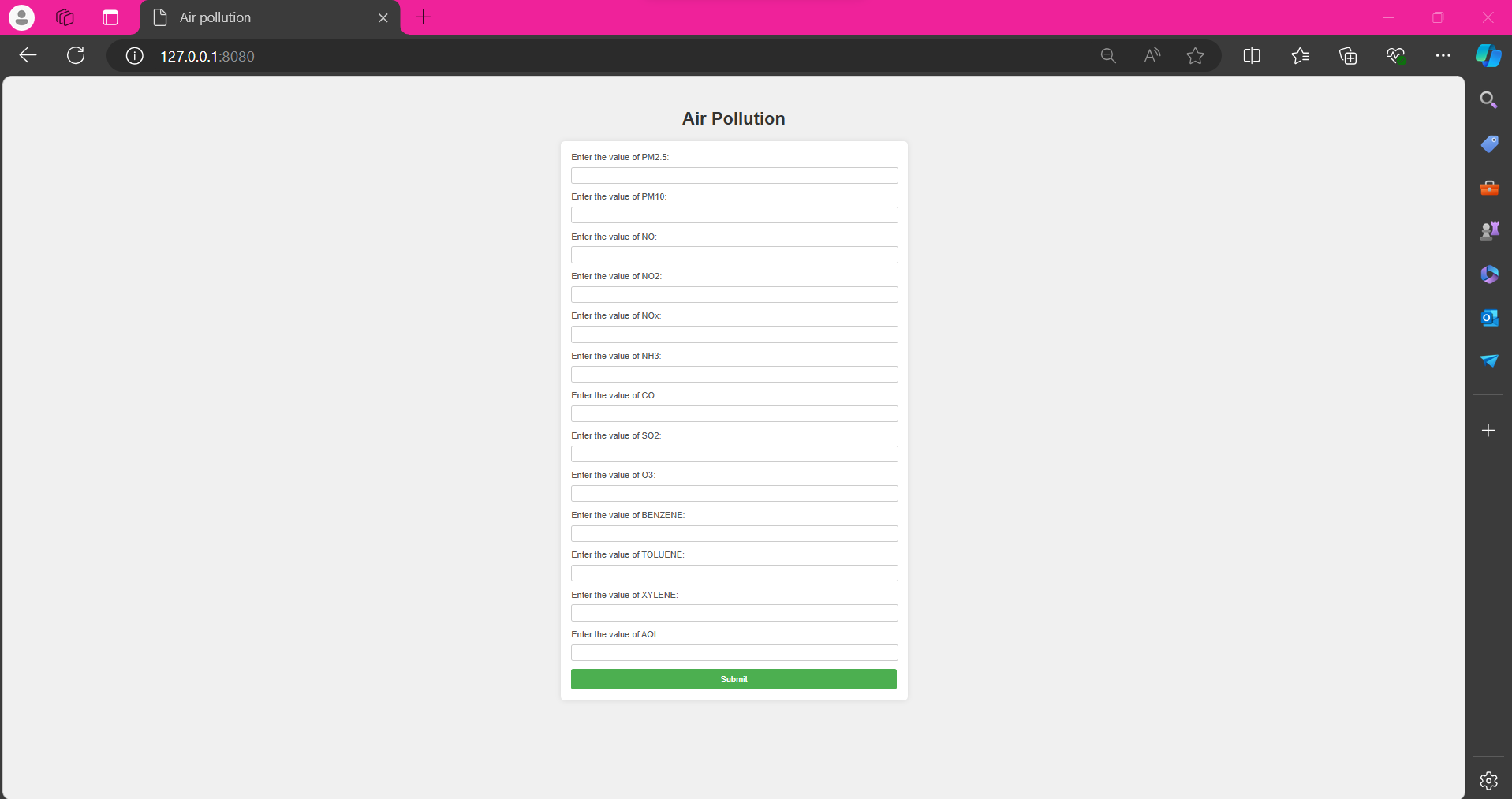


Fig:8.2 Output of webpage

After entering the values, the result that is the intensity of the air pollution will be predicted by the model. It will tell whether it is good, poor ,very poor,moderate or severe air pollution.

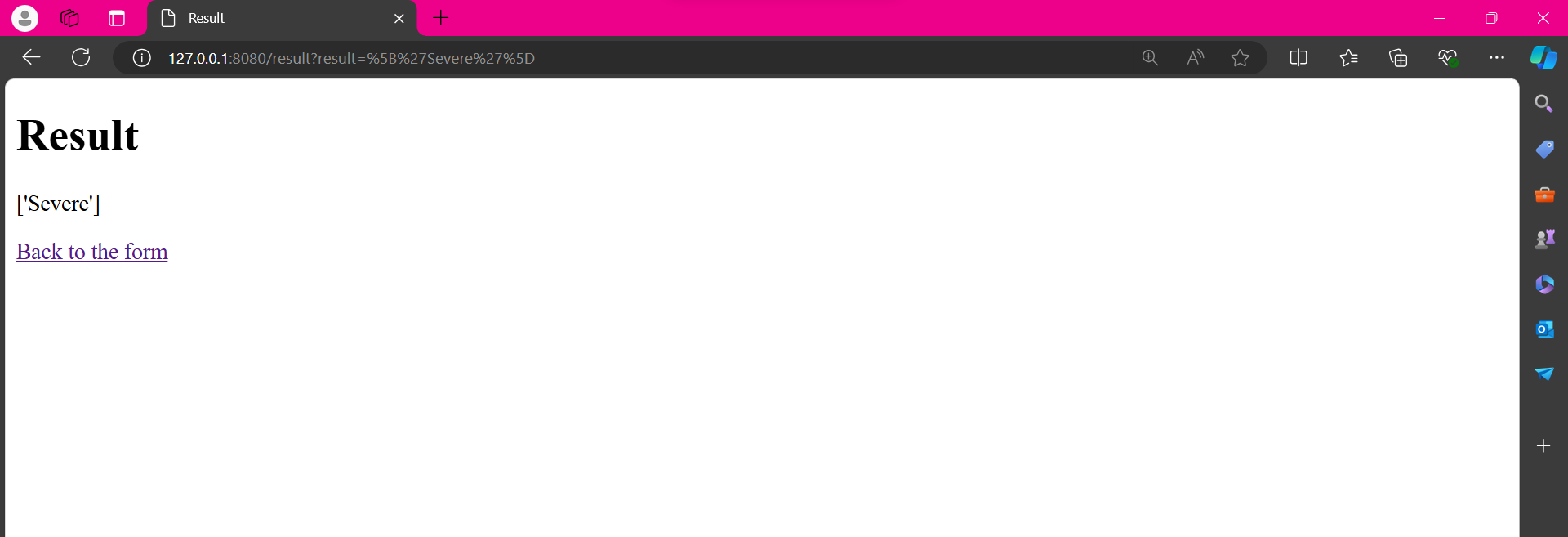


Fig:8.3 Output of the model created

**CHAPTER-9**

**CONCLUSION**

In conclusion, random forest is a powerful machine learning algorithm that can be used for air quality index prediction. It is a popular method for its ability to handle complex, high-dimensional datasets and to identify important features for prediction. By using random forest to analyze various air quality parameters, such as temperature, humidity, and particulate matter concentrations, it is possible to accurately predict the air quality index at a given location and time. However, it is important to note that prediction accuracy can be affected by the quality and quantity of data used to train the model, as well as other external factors such as weather conditions and human activity.

Despite significant advancements in air pollution control technology and management techniques, the desired outcome remains a long way off. Hence, there is a need to strengthen available policies and technologies on the identified research gaps in India’s air pollution related aspects. Further, the identified gaps can be divided into three areas re- search, policies and economic scale. First, a research constraint has been identified that necessitates a calibrated management system based on data generated by a broad countrywide network to take legislative steps to procure development and growth in the arena of air quality management while addressing prevalent lacunas. Second, this review also identifies a policy constraint in which a country’s national air quality standards must be revisited in light of higher background concentrations of pollutants. Third, the vehicle sector contributes significantly to air pollution which needs to be manage through strategic phasing of reduction of tail pipe emission. The review highlights that the sector has all three identified gaps, indicating that more study is needed to establish region-specific vehicle emission factors and assess real-time vehicular emissions to reinforce control activities.

REFERENCE

[1] Dragomir, Elia Georgiana. "Air quality index prediction using K-nearest neighbor technique no. 1 (2010): 103-108.

[2] Carbajal-Hernández, José Juan "Assessment and prediction of air quality using fuzzy logic and autoregressive models." Atmospheric Environment 60 (2012): 37-50.

[3] Kumar, Anikender and P. Goyal, “ Forcasting of daily air quality index in Delhi”, Science of th Total Environment 409, no. 24(2011): 5517- 5523.

[4] Singh Kunwar Petal. “Linear and nonlinear modelling approaches for urban air quality prediction, “ Science of the Total Environment 426(2012):244-255.

[5] Sivacoumar R, et al, “ Air pollution modelling for an industrial complex and model performance evaluation “, Environmental Pollution 111.3 (2001) : 471-477

[6] Gokhale sharad and Namita Raokhande, “Performance evaluation of air quality models for predicting PM10 and PM2.5 concentrations at urban traffic intersection during winter period”, Science of the total environment 394.1(2008): 9- 24.

[7] Bhanarkar, A. D., et al, “Assessment of contribution of SO2 and NO2 from different sources in Jamshedpur region, India, “Atmospheric Environment 39.40(2005):7745- India." Atmospheric Environment 39.40 (2005): 7745-7760.

[8] Singh Kunwar P., Shikha Gupta and Premanjali Rai, “ Identifying pollution sources and prediction urban air quality using ensemble learning methods”, Atmospheric environment80 (2013): 426-437.

[9] Wang Jun, and Sundar A. Christopher, “Intercomparison between satellite derived aerosol optical thickness and PM2. 5 Mass: Impliances for air quality studies”,Geophysical research letters30.21(2003).

[10] Sharma M E A McBean and U.Ghosh, “Prediction of atmospheric sulphate deposition at sensitive receptors in northern India”, Atmospheric Environment 29.16(1995): 2157- 2162

[11] T. Madan, S. Sagar, and D. Virmani, “Air quality prediction using machine learning algorithms –a review,” in 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020, pp. 140–145.

[12] C. Li, Y. Li, and Y. Bao, “Research on air quality prediction based on machine learning,” in 2021 2nd International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI), 2021, pp. 77–81.

[13]Jung,Chau-Ren,Bing-FangHwang,andWei-TingChen.”Incorporatinglong-termsatellite-basedaerosolopticaldepth,localizedlandusedata,andmeteorologicalvariablestoestimateground-levelPM2.5concentrationsinTaiwanfrom2005to2015.”*EnvironmentalPollution*237(1):1000-1010

.[14]Soh,Ping-Wei,Kai-HsiangChen,Jen-WeiHuang,andHone-JayChu.”Spatial-TemporalpatternanalysisandpredictionofairqualityinTaiwan.”*10thInternationalConferenceonUbi-mediaComputingandWorkshops(Ubi-Media)IEEE*(2),pp.1-6.

[15]Mehdipour,Vahid,DavidS.Stevenson,MahsaMemarianfard,andParveenSihag.”ComparingdifferentmethodsforstatisticalmodelingofparticulatematterinTehran,Iran.”*AirQuality,AtmosphereandHealth*11,no.(10)(3):1155-1165.

[16]Delavar,MahmoudReza,AminGholami,GholamRezaShiran,YousefRashidi,GholamRezaNakhaeizadeh,KurtFedra,andSmaeilHatefiAfshar.”ANovelMethodforImprovingAirPollutionPredictionBasedonMachineLearningApproaches:ACaseStudyAppliedtotheCapitalCityofTehran.”*ISPRSInternationalJournalofGeo-Information*8,no.2(4):99.

[17] Wang,Wei, Feiyue Mao, Lin Du, Zengxin Pan,Wei Gong, and Shenghui Fang. ”Deriving hourly PM2. 5 concentrations from himawari-8 aods over beijingtianjinhebei in China.” *Remote Sensing* 9, no. 8 (5): 858.

[18] Pan, Bingyue. ”Application of XGBoost algorithm in hourly PM2. 5 concentration prediction.” *In IOP Conference Series: Earth and Environ-mental Science,IOP Publishing* vol. 113, no. 1 (6), p. 012127.

[19] Sun, Xiaotong, Wei Xu, and Hongxun Jiang. ”Spatial-temporal prediction of air quality based on recurrent neural networks.” *In Proceedings of the 52nd Hawaii International Conference on System Sciences* 9 (7).

[20] Sharma, Nidhi, Shweta Taneja, Vaishali Sagar, and Arshita Bhatt. ”Forecasting air pollution load in Delhi using data analysis tools.” *Procedia computer science* 132 (8): 1077-1085.

[21] Martnez-Espaa, Raquel, Andrs Bueno-Crespo, Isabel Maria Timon-Perez, Jess Soto, Andrs Muoz, and Jos M. Cecilia. ”Air-Pollution Prediction in Smart Cities through Machine Learning Methods: A Case of Study in Murcia, Spain.” *J. UCS 24* 132, no. 3:(9) 261-276.

[22] Wang, Yikai, Xuefei Hu, Howard Chang, Lance Waller, Jessica Belle, and Yang Liu. ”A bayesian downscaler model to estimate daily PM2. 5 levels in the conterminous US.”*International journal of environmental research and public health* 15, no. 9 (10): 1999.

[23] Soh, Ping-Wei, Jia-Wei Chang, and Jen-Wei Huang. ”Adaptive deep learning-based air quality prediction model using the most relevant spatial-temporal relations.”*IEEE Access* 6 (11):38186-38199.

[24] Liu, Hui, Haiping Wu, Xinwei Lv, Zhiren Ren, Min Liu, Yanfei Li, and Huipeng Shi. ”An intelligent hybrid model for air pollutant concentra-tions forecasting: Case of Beijing in China.”*Sustainable Cities and Society 47* 1 (12): p.101471.

[25] KamiÅska, Joanna A. ”A random forest partition model for predicting NO2 concentrations from traffic flow and meteorological conditions.” *Science of The Total Environment 47* 651 (13): p. 475-483.

[26] Park, Seohui, Minso Shin, Jungho Im, Chang-Keun Song, Myungje Choi, Jhoon Kim, Seungun Lee et al. ”Estimation of ground-level par-ticulate matter concentrations through the synergistic use of satellite observations and process-based models over South Korea.”*Atmospheric Chemistry and Physics* 19, no. 2(14): p.1097-1113.

[27] Kang, Ganganjot Kaur, Jerry Zeyu Gao, Sen Chiao, Shengqiang Lu, and Gang Xie. ”Air quality prediction: Big data and machine learning approaches.” *International Journal of Environmental Science and Development* 9, no. 1 (15): p.8-16.

[28] Lyu, Baolei, Yuzhong Zhang, and Yongtao Hu. ”Improving PM2. 5 air quality model forecasts in China using a bias-correction frame-work.”*Atmosphere* 8, no. 8 (16): p.147.

[29] Rybarczyk, Yves, and Rasa Zalakeviciute. ”Machine learning approaches for outdoor air quality modelling: A systematic review.” *Applied Sciences* 8, no. 12 (17): p.2570.

[30] Environmental Protection Administration. AQI value and impact on health. Available from: <https://taqm.epa.gov.tw/taqm/en/>

[31] Karpatne, Anuj, Imme Ebert-Uphoff, Sai Ravela, Hassan Ali Babaie, and Vipin Kumar. ”Machine learning for the geosciences: Challenges and opportunities.” *IEEE Transactions on Knowledge and Data Engineering* (18).

[32] Grange, Stuart K., David C. Carslaw, Alastair C. Lewis, Eirini Boleti, and Christoph Hueglin. ”Random forest meteorological normalisation models for Swiss PM 10 trend analysis.” *Atmospheric Chemistry and Physics* 18, no. 9 (19): p.6223-6239.

[33] Zhao, Fang, andWeide Li. ”A Combined Model Based on Feature Selection andWOA for PM2. 5 Concentration Forecasting.”*Atmosphere* 10, no. 4 (20): p.223.

[34] Isakov, Vlad, Saravanan Arunachalam, Stuart Batterman, Sarah Bereznicki, Janet Burke, Kathie Dionisio, Val Garcia et al. ”Air quality model-ing in support of the near-road exposures and effects of urban air pollutants study (NEXUS).”*International journal of environmental research and public health* 11, no. 9 (21): p.8777-8793.

[35] Zhu, Dixian, Changjie Cai, Tianbao Yang, and Xun Zhou. ”A machine learning approach for air quality prediction: Model regularization and optimization.”*Big data and cognitive computing* 2, no. 1 (22): p.5.

[36] Suleiman, A., M. R. Tight, and A. D. Quinn. ”Applying machine learning methods in managing urban concentrations of traffic-related partic-ulate matter (PM10 and PM2. 5).” *Atmospheric Pollution Research* 10, no. 1 (23): p.134-144.

[37] Xie, Jingjing, XiaoxueWang, Yu Liu, and Yun Bai. ”Autoencoder-based deep belief regression network for air particulate matter concentration forecasting.” *Journal of Intelligent and Fuzzy Systems* 34, no. 6 (24): p.3475-3486.

[38] Yao, Fei, Jiansheng Wu, Weifeng Li, and Jian Peng. ”Estimating Daily PM2. 5 Concentrations in Beijing Using 750-M VIIRS IP AOD Retrievals and a Nested Spatiotemporal Statistical Model.”*Remote Sensing* 11, no. 7 (25): p.841.

[39] Balakrishnan, K., Dey, S., Gupta, T., Dhaliwal, R.S. (2018).The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: The Global Burden of Disease Study 2017. India State-Level Disease Burden Initiative Air Pollution Collaborators. Lancet Planet Health, 3, pp. e26–e39. DOI:10.1016/S2542-5196(18)30261-4.[45]

[40]. Siddique, S., Ray, M.R., Lahiri, T. (2011). Effects of air pollution on the respiratory health of children: a study in the capital city of India. Air Quality, Atmosphere & Health, 4(2),95–102, DOI: 10.1007/s11869-010-0079-2

[41] Gurjar, B. (2021). Air Pollution in India: Major Issues and Challenges. [online] The Energy and Resources Institute. April 05, 2021. Accessed October 14,2021. https://www.teriin.org/article/air-pollution-india-majorissues-and-challenges.

[42] Dandona, L. (2021). Health and economic impact of air pollution in the states of India: the Global Burden of Disease Study 2019. India State-Level Disease Burden Initiative Air Pollution Collaborators. Lancet Planet Health, 5(1), e25–e38, DOI:10.1016/S2542-5196(20)30298-9.

[43] Chang, T., Zivin, J.G., Gross, T., Neidell, M. (2016). The effect of pollution on worker productivity: evidence from callcenter workers in China. Institute for the Study of Labor (IZA) Discussion Paper #10027.June, 2016.DOI 10.3386/w22328. http://ftp.iza.org/dp10027.pdf

[44] Hansen-Lewis, J. (2018). Does air pollution lower productivity? Evidence from manufacturing in India. Brown University. Oct 11, 2018. <http://barrett.dyson.cornell> .edu/NEUDC/paper\_324.pdf (accessed Dec 11, 2020).

[45] World Bank, Institute for Health Metrics and Evaluation(2016). The cost of air pollution: strengthening the economic case for action. World Bank, Washington, DC . https://openknowledge.worldbank.org/handle/10986/25013

[46] Quaha E., Boon, T.L. (2003). The economic cost of particulate air pollution on health in Singapore. Journal of Asian Economics, 14(1), 73-90. DOI: 10.1016/S1049-0078(02) 00240-3.

[47] Guttikunda, S.K., Goel, R., Pant, P. (2014). Nature of air pollution, emission sources, and management in the Indian cities. Atmospheric Environment, 95, 501-510.

DOI: 10.1016/j.atmosenv.2014.07.006

[48] Central Pollution Control Board (CPCB), (2010). Air quality monitoring, emission inventory and source apportionment[46] study for Indian cities. National Summary Report 2011.Central Pollution Control Board, Govt. of India. <https://cpcb.nic.in/displaypdf.php?id=RmluYWxOYXRpb25hbFN1bW1hcnkucGRm>

[49] National Ambient Air Quality Status & Trend 2019. (CPCB 2020). Central Pollution Control Board, Ministry of Environment, Forest & Climate Change, Government of India.

Bottom of Form