Vyaan- Air Quality Analysis And Prediction

Submitted in partial fulfillment of the requirements

For the degree of

Bachelor of Engineering in Information Technology

by

Bhakti Zaware 17IT1042
Saniya Upadhyaya 17IT2005
Yashsingh Manral 17IT2002

Supervisor

Mr. Murali Parameswaran

D Y PATIL
DEEMED TO BE
UNIVERSITY
RAMRAO ADIK
INSTITUTE OF TECHNOLOGY
NAVI MUMBAI

Department of Information Technology

Dr. D. Y. Patil Group’s

Ramrao Adik Institute of Technology

Nerul, Navi Mumbai 400706.

(Affiliated to University of Mumbai)

(2021)

D Y PATIL
DEEMED TO BE
UNIVERSITY
RAMRAO ADIK
INSTITUTE OF TECHNOLOGY
NAVI MUMBAI

Ramrao Adik Institute of Technology

(Affiliated to the University of Mumbai)
Dr. D. Y. Patil Vidyanagar, Sector 7, Nerul, Navi Mumbai 400 706.

CERTIFICATE

This is to certify that, the Project-I titled

“Geospatial Analysis, Visualization and Prediction of Earth’s Surface Temperature”

is a bonafide work done by

Bhakti Zaware
Saniya Upadhyaya
Yashsingh Manral

and is submitted in the partial fulfillment of the requirement for the degree of

Bachelor of Engineering
in
Information Technology
to the
University of Mumbai

Supervisor

Mr. Murali Parameswaran

Project Co-ordinator	Head of Department	Principal
(Mrs. Reshma Gulwani)	(Dr. Ashish Jadhav)	(Dr. Mukesh D. Patil)

Project Report Approval for B.E.

This is to certify that the project entitled *"Air Quality, Analysis and Visualization"* is a bonafide work done by *Bhakti Zaware, Saniya Upadyaya and Yashsingh Manral* under the supervision of *Mr. Murali Parameswaran*. This project has been approved for the award of *Bachelor's Degree in Information Technology, University of Mumbai.*

Examiners:

1. ....

2. ....

Supervisors:

1. ....

2. ....

Principal:

....

Date: 9th December 2020

Place: Mumbai

Declaration

I declare that this written submission represents my ideas in my own words and where other's ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Bhakti Zaware	17IT1042	(B.Z.)
Saniya Upadhyaya	17IT2005	(S.U.)
Yashsingh Manral	17IT2002	(Y.M.)

Date: 18th May 2021

Contents

Abstract

Today we live in a world where things around us are experiencing an escalation in emerging technologies like artificial intelligence, machine learning, cloud, internet of things and so forth to provide high level of comfort to humankind with minimum human intervention. A major challenge faced by many industries, corporations, or industries is to control and regulate air quality. Almost every generation of humankind is equally affected by air which is contaminated by harmful pollutants; looking forward to all of this air pollution prediction is an increasingly important necessity. It can impact individuals and their health, example the individuals which undergo major respiratory problems are the ones who are greatly affected by air pollution. With the advent of modern air quality monitoring and pollution control systems, a prediction framework aids the process of finding effective solutions to complex problems.

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Introduction of your project concept and assign some suitable heading

Data analytics is the science of analyzing raw data to make conclusions about that information. Data is extracted from various sources and is cleaned and categorized to analyze different patterns.

This project aims in building a prediction model and providing the results based on the pollution data. This project focuses to investigate machine learning based techniques for air quality prediction. In this project, a particulate matter of less than 2.5 micro-meters (PM2.5) was collected from various sources. Firstly data analysis was conducted on pollution, furthermore data cleansing was analyzed with different machine learning models including decision tree, random forest regression and support vector machine. To overcome the drawbacks of these algorithms we implemented LSTM model for more accurate results. In future, this work can also be determined by employing different artificial intelligence techniques.

With the economical and technological advances of cities, there are many problems related to environmental pollutions which are escalating. In particular, air pollution has an adverse effect on humankind through the exposure of harmful pollutants, increasing its interest in air and causing a huge impact on the scientific community. Air pollution comes out from many sources for example factories, industries, homes, activities related to business, etc. Air pollution has a direct impact on humankind through the exposure of harmful pollutants which has increased the rate of air pollution. World Health Organization gives the following definition for air pollution “Air pollution is a combination of the indoors and outdoors environment by chemical, physical or biological agent that modifies the natural characteristics of the atmosphere”. With excessive rise of population and lack of measures to control pollution there is a high probability that the measures used to prevent air pollution may worsen in the coming future. India experiences the highest number of emissions from coal-fired power plants in order to be precise almost 76,000 of the 90,980 deaths are linked to power plants.

1.2 Problem Definition

Almost every city across India comprises and utilizes information and advanced communication technologies to provide better health, flexible transport and energy related facilities and also enables the government to make use of its valuable resources. Better pollution control energy conservation, waste management, traffic control policies, and betterment in public safety and security are among the objectives of developing a smart city.

In the recent years there has been a hike in urban population due to industrialization and migration of people from rural to urban areas. According to a UN report, considering an approximate amount around 54 to 66 percent of the world's population shall migrate to urban areas by 2050. With the rise in population, demand of transportation and energy have also been increased, also increasing the sources of pollution emissions which has indeed termed to be a major concern for local and national governments as well as the leaders on global stage. Local and national government ensures to provide a better lifestyle for its inhabitants by controlling pollution-related diseases.

1.3 Scope of Project

Air Quality is one of the major issues in India and people tend to be ignorant to this problem. The top 10 polluted cities in the world consists of 9 Indian cities and all of them are from the Northern Province of India. As claimed by the UN, the PM2.5 shouldn't top 10 mcg/m3 averages throughout the year or 25 mcg/m3 in any 24-hour time. A recent study shows that WHO has claimed that 4.2 million fatalities from outside air pollution however, has undervalued the effect on heart disease.

Hence, our project has the following objectives:

- Predicting Air Quality on the basis of Air Quality Historical Data Platform with six variables by creating a model.
- One of the main objectives of this project is to spread awareness about the effects of Pollution and its effect on every individual.
- How every individual is responsible for pollution and what can one do make air clean for every individual to breathe.

The scope of our project is to maximize the accuracy of our model by adding large amount of data from different sites as well. We have analyzed various papers and algorithms that were previously used to spread awareness about air pollution.

We found a few drawbacks with the other published paper and sites:

- Many of the previous projects have been limited in their scope due to the usage of static data.
- Available solutions don’t predict the future Air Quality Index (AQI).
- The intention of the data has not been clearly explained to the people by the previous studies.

This project solves the following problems:

- We have made use of dynamic data which will be useful for increasing the potential of our model.
- It will predict the future Air Quality Index (AQI).
- The intention of the data will be clearly explained to the citizens through our model.

1.4 Relevance and Motivation of Project

We developed an inclination for Machine Learning and Analytics through our curriculum, due to its popularity and wide industrial applications. These fields tend to be very popular because they are useful for gaining insights from large amount of data and they do this by experienced leveraging algorithms and discovering patterns.

Within the scope of analytics and machine learning, we chose a topic that would help us to bring about a larger impact on our environment. We chose a topic that would be beneficial for everyone and therefore will be relevant for curbing the impact of air pollution.

Even though the citizens are aware of the global warming caused by air pollution, they are not motivated on a daily basis to take measures for preventing the effects of air pollution.

The seriousness of the issue can be very well understood through visualizations and interactive platforms. A progress can only be made if a country as a whole will unite and take steps towards a pollution free environment.

1.5 Organization of Report

The Chapter 1 contains problem definition, scope of the project, objective and motivation of the entire project. It gives the general overview of our project. The Chapter 2 includes the various research papers in the field of analytics and machine learning. The literature survey tells us about the existing system and wide range of implementation of various algorithms.

Chapter 3 includes the planning and formulation of the project. The Chapter 4 contains the proposed work in terms of its framework and methodologies used, hardware and software requirements. Chapter 5 gives the detailed flow diagrams which help to understand the flow of the system working. Chapter 6 shows the Results and required validation.

Chapter 2
Review of Literature

Mrs A. Gnana Soundari et al. [1] have used Neural Network Model and Box Plot analysis is used to calculate the air quality index based on historical data of previous years and predicting over a specific upcoming year as a Gradient decent boosted multivariable regression problem. Along with this, AHP MCDM technique is used to find of order of preference by similarity to ideal solution. In conclusion, this model has an accuracy of 96% and can also predict the upcoming air quality index of any particular data within a given region. This paper originates from checking china air quality, investigation stage and then incorporates the everyday fine particulate issue (PM2.5), inhalable particulate issue (PM10), Ozone (O3), CO, SO2, NO2 fixation and air quality record (AQI).

Saba Ameer et al. [2] made use of Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, Ann Multi-Layer Perception Regression that is used to present a comparative study to determine the best model for predicting air quality index with respect to data size and processing time. Furthermore, Apache Spark is used for implementing experiments and performing pollution estimation using multiple datasets. In conclusion, of all the algorithms used, Random Forest regression was the best technique, performing well for pollution prediction for data sets of varying size and location and having different characteristics. The worst performing algorithm was Gradient boosting regression out of all the algorithms, since it achieved highest processing time in almost every data set and had given a very high error rate in most of the cases.

Aditya C R et al. [3] have used Logistic regression which is employed to detect whether a data sample is either polluted or not polluted. Auto-regression is employed to predict future values of PM2.5 based on the previous PM2.5 readings. Logistic Regression has termed to be the best for this system, having mean accuracy and standard deviation of 0.998859 and 0.000612 respectively. Auto-regression was applied on time series data tend to predict the PM2.5 value 7 days prior to the current date, generated the Mean Squared Error (MSE) of 27.0. The difference between current date and the date on which the value of PM2.5 is to be predicted should be reduced to lower the MSE**.** Both of the above models can be efficiently used to detect the quality of air and predict the level of PM2.5 in the future.

Heni Patel et al. [4] started with Multivariate Multistep Time Series Prediction which is done Using Random Forest for predicting the air pollution. A number of trees were created and each of them was trained on a subset of time series data. Along with this, a feature selection technique is used using Genetic Algorithm which is used for finding the most relevant inputs for predictive model. In conclusion, Multivariate Multistep Time Series Prediction Using Random Forest technique not only improves the performance but also reduce the complexity of the air pollution prediction model. Also here there is a feature selection technique used which makes the prediction even better.

Zhiying Meng [5], published in Journal of Software Engineering and Applications have used five different machine learning models which are used in the prediction of ground ozone level and their final accuracy scores are compared. Among Logistic Regression, Decision Tree, Random Forest, AdaBoost, and Support Vector Machine (SVM), SVM has the highest test score of 0.949. This research paper exploits simple methods of forecasting and calculates the accuracy in predicting ground ozone level which can thus be beneficial for the environmentalists for their research. Further, the train and test scores are compared and it reached a conclusion that the most accurate method is SVM which predicts the binary outcomes having a test score of 0.949.

K. Mahesh Babu et al. [6], published in International Journal of Innovative Technology and Exploring Engineering (IJITEE) have used Logistic Regression, Random Forest, K-Nearest Neighbors, decision tree and support vector machines to predict the quality of air. Here, the air quality dataset is pre-processed with respect to uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments, data validation, data cleaning/preparing. Data pre-processing is next, independent features were examined and correlation between in dependant and dependent features were calculated. This research paper concluded that Decision Tree algorithm works best for predicting air quality. The first-rate accuracy on public test set is good parameter values of decision tree method process by way of accuracy with regard to classification.

Gaganjot Kaur et al. [7], published in International Journal of Environmental Science and development, Vol 9, No 1 January 2018, they have implemented Artificial Neural Network Model, Genetic Algorithm-ANN model, Random forest Model, Decision tree model, Least Squares Support Vector Machine Models and Deep Belief Network. The research paper aims to evaluate various big-data and machine learning based techniques for air quality prediction. This paper used various methods like artificial intelligence, decision trees, deep learning etc to review the published research results relating to air quality evaluation. This paper also highlights the issues, challenges and needs of this topic.

Sr. No	Problem Statement	Technique	Strengths	Limitations
1	Forecast the air quality of India by using machine learning to predict the air quality index of a given area	Neural Network Model, Box Plot analysis and AHP MCDM	Boosting Algorithm is a victor among the most prevalent learning insights	There are seasonal variations and trend, in order to reduce these metrics, we resample the data month wise to predict it month wise
2	The pollution prediction using four advanced regression techniques and have presented a comparative study to analyze the best model for accurately predicting the air quality with reference to data size and processing time.	Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, Ann Multi-Layer Perception Regression	Random Forest Regression achieve the lowest mean absolute error and RMSE.	Gradient boosting regression has performed worst as it has achieved highest processing time in almost all data sets and has given a very high error rate in most cases.
3	This system attempts to predict PM2.5 level and detect air quality based on a data set consisting of daily atmospheric conditions in a specific city.	Logistic Regression and Autoregression	The results show that machine learning models (logistic regression and autoregression) can be efficiently used to detect the	Data Size is Small

			quality of air and predict the level of PM2.5 in the future.	
4	Predicting air pollution for our smart city using data mining technique	Time Series, Random Forest Algorithm	Multivariate Multistep Time Series Prediction Using Random Forest technique improve the performance and reduce the complexity of the air pollution 5prediction model.	Feature selection technique must be improved which make our prediction even better.
5	Classifying ground ozone level based on big data and machine learning models.	Principal Component Analysis and Logistic Regression, Adaboost, Decision Tree, Random Forest, Support Vector Machine	This research reached a conclusion that using SVM to predict binary outcomes is the most accurate method, which has the highest test score of 0.949	PCA is proven to be unnecessary. Only Ozone Factor.
6	Concentrates on performing an effective analysis on all the major works done in this aspect using machine learning algorithms	Linear Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors.	The first-rate accuracy on public test set is good parameter values of decision tree method.	Implementation is not discussed.
7	Big-data and machine learning based techniques for air quality forecasting.	Artificial Neural Network Model, Genetic Algorithm-ANN Model, Random Forest, Decision Tree, Least Squares Support Vector Machine Models and Deep Belief Network.	They find decision tree to be the best estimator.	Data quality and validation issue. Real-time air quality monitor and supervision for air resources.

Chapter 3

Planning and Formulation

Initially we decided the system of our project as a whole but to begin with the actual implementation we had to carry out survey of the existing systems. At the start, we were focused on implementing the project without considering factors that play an equally important role in such prediction-based projects. We instantiated the project work by extensively studying the machine learning algorithm and related research paper and publications after two months of rigorous of literature survey we shortlisted few papers which were most relevant to our project. From our research we came to know that the most important factors of our project will be various existing dataset, sources of datasets, tools and previous studies related to Air Quality Index. After the Literature survey was conducted, we began with the first stage of our project that was to gather data from various AQI sites which is more suitable to our project and in a format that is easy to work with. After finding the relevant data source, our next phase was to preprocess the data which consisted of identifying and eliminating outliers. Hence we made use of ample amount of data which was required to carry out the implementation of literature survey algorithms.

After the actual implementation with the available algorithms that were more admissible to our problem statement, we shortlisted a few regression models and studied their results. After analyzing the results, we experienced many flaws and defects that were comprising the accuracy as well as results of our model. To overcome these limitations, we finalized the LSTM model as it gives promising results to make predictions in time-series forecasting models rather than giving pre-specified and fixed results. The final piece of the project is forecasting the AQI of a particular city on an interactive platform to make it user friendly for the citizens to analyze the severity of the air pollution problem.

Chapter 4
Methodologies

4.1 Proposed System

We collected the data from heterogenous real time dataset platforms. After collecting the data, we performed data wrangling which is responsible for transforming one raw data into another format having the intention to make it more appropriate for a variety of purposes. Since the data is collected from different platforms, it contains inconsistent as well as recurring data and missing values. In order to get accurate prediction result, the dataset must be cleaned, missing values should either be deleted or filled with mean values to eliminate redundancy. First we divided the data into training and testing data, and passed it into the model accordingly. We implemented various algorithms like decision tree, random forest and support vector machine to analyze the behaviour of the dataset and to find their maximum accuracy. After testing the model, the accuracy of the model is estimated by using parameters like MSE, RMSE, and overall accuracy. Based on the results, we can determine the best algorithm for our classification model. Now since the model has been built, it is also important to check its robustness and stability. This can effectively be done scoring and predicting using the testing data set which is available to us. This will be in turn be beneficial to check as to how the predicted values are compared to the actual data. The proposed system serves as a means to educate the common public about global warming and spread awareness about the rising surface temperature of the Earth and its ill effects.

4.2 Proposed Methodology

The algorithms which we have implemented using python programming Language and pre-processing and time series evaluation was conducted using Panda. Machine learning algorithms have been employed by using scikit learn library which is an open-source machine learning library. To plot the graphs effectively the plotly library has been used. The evaluation of performance has been conducted using sklearn metrics. We implemented three algorithms in our project that are Decision Tree, Random Forest, Support Vector Machine.

After experiencing such limitations from the above-mentioned algorithms we have decided to implement LSTM (Long Short Term Memory Network) as it intends to improve and investigate the usage of LSTM to forecast air quality through dataset and variable selection—considering a large dataset with more parameters and measurements, which can support more accurate predictive models for air pollutants and particulates, in particular, PM2.5.

4.3 System Requirements

Data is a very crucial aspect in any Data science projects and the same applies for our project. Data is collected from multiple AQI sources which is further cleaned, processed and transformed into a homogenous format for smooth execution.

Python is a language which is highly interpreted, object-oriented programming language with dynamic semantics. It consists of data structures which are combined with dynamic typing and dynamic findings, making it attractive for Rapid Application Development and scripting language which serves the purpose of connecting the existing components together.

It supports modules and packages which enhances program modularity and code reusability.

We have used Juypter Notebook which is an open source application responsible for creating and sharing documents but contains live code, visualizations and equations. Python is proficient for cleaning and transformation of data, numerical simulation, statistical modeling, data visualization, machine learning, analytics etc.

Python consists of various libraries and dependencies which are helpful in carrying out operations like pre-processing, visualization, training and testing of the model. The libraries which we have used in this project are pandas, NumPy, matplotlib, seaborn, sci-kit learn, keras and TensorFlow.

Chapter 5

Design of System

5.1 System Design

The system model describes how the AQI model is constructed and accuracy is obtained for prediction. Preprocessing is carried out on the raw data obtained from AQICN platform. We split the data into training and testing tests and then we trained our prediction model. The training data will be used for training the algorithm and later it will be used for building the predictive model. On the other hand, the test data will be useful in validating the model and also see how the comparisons are made with the actual data.

The algorithm used for our prediction model is Long Short Term Memory Network (LSTM).

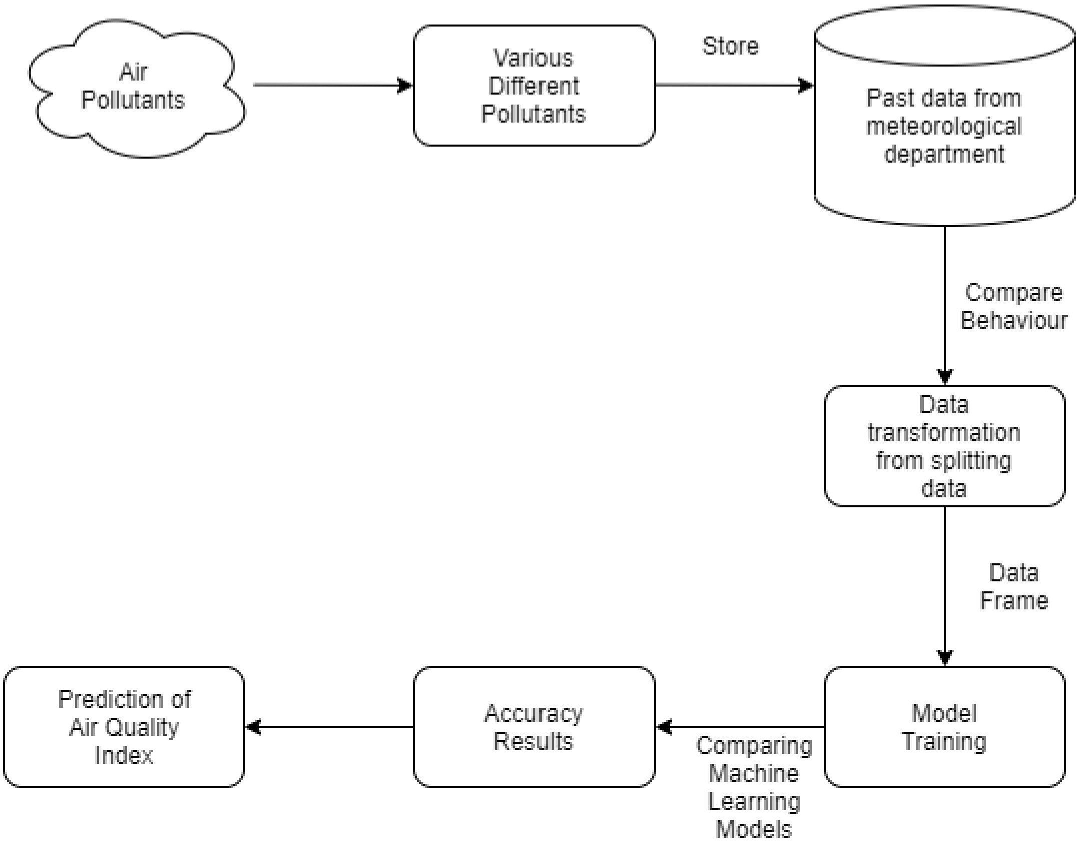

Air Pollutants
Various Different Pollutants
Store
Past data from meteorological department
Compare Behaviour
Data transformation from splitting data
Data Frame
Model Training
Comparing Machine Learning Models
Accuracy Results
Prediction of Air Quality Index

Fig.1 System design

LSTM promises to learn the context which is beneficial for making predictions in time series forecasting problems, rather having its context pre-specified and fixed. LSTM networks are used on time-series data for classification, processing, and making predictions.

We have called our dataset .csv file and renamed the columns. We have converted the values into numerical forms for easy processing. We have beautiful soup which will go to a particular link and will pass the html (aqicn.org) which means it will fetch the value on that basis enclosed by a particular tag. We are fetching the current values for all the parameters and storing it in a variable called live_data. We have manage the missing values by filling them up with mean of the columns. We are sequencing the data based on dates. We have applied min-max scalar that will convert our dataset in the range of 0 to 1. We have used min-max scalar here because our LSTM works better for shorter ranges of values. In LSTM, we train our model to particular epoch range where the weight gets updated, so initially our epoch value is close to 0 and whenever the weight gets updated there is a minute difference between the initial value and the updated value; since the difference is very small, the loss is also minimal and our model predicts more accuracy. The predicted value are then transformed into actual range by using inverse transform method. Finally using the prediction model, our AQI is predicted.

5.2 Data Flow Diagrams

In simplified terms, the process consists of 5 steps:

1. Data Source and Data Collection
2. Data Cleaning
3. Data Preprocessing
4. Build, Train, Test and Deploy Model
5. Predict Future AQI

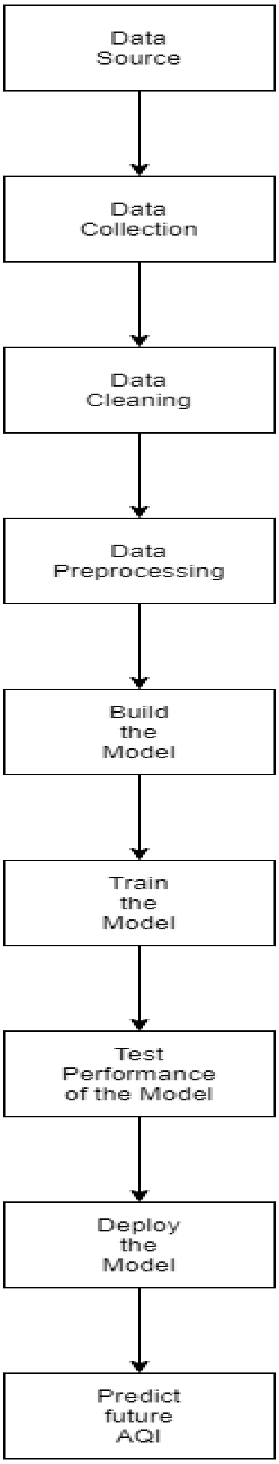


Fig 2. Data Flow Diagram

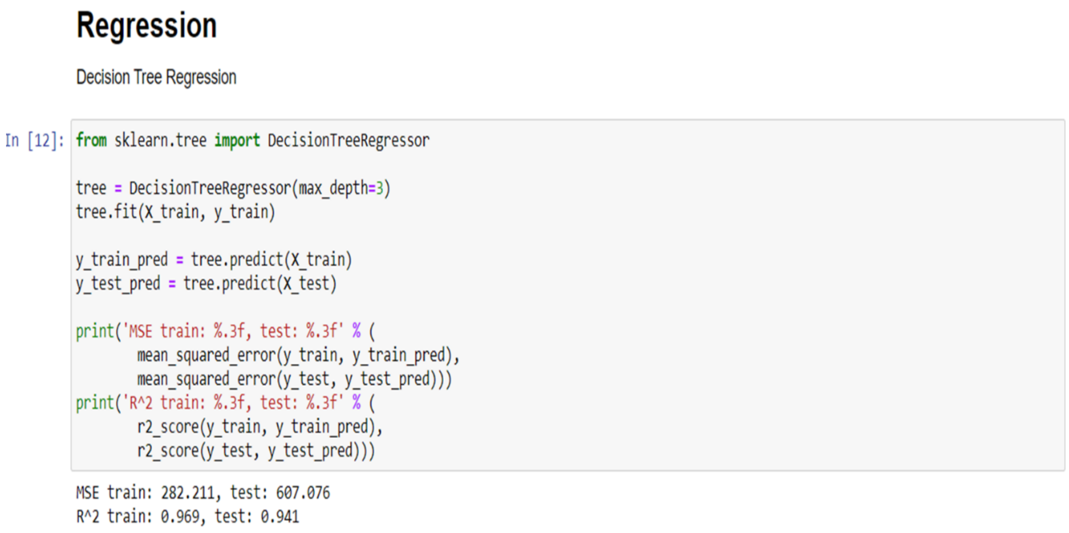
First, we collected the data from an appropriate AQI source. We cleaned the data and managed the missing values. Then, we build a model and trained it by the training dataset. Later, we tested the model’s performance using the testing dataset. Once the model was built successfully, we deployed it. Then this model predicts future AQI.

Chapter 6

Experimental Results

We have analyzed and processed our dataset using decision tree, random forest and SVM. We have taken six parameters in to consideration that are PM2.5, CO, SO2, NO2, O3 and PM10; amongst these parameters we have considered PM2.5 to be a dependent variable and the others are considered to be independent variables. We have divided the data set into training and testing sets and used the training set to train the model and testing set to test the model. We then evaluated the model performance based on an error metric to determine the accuracy of the model. These algorithms produced various results based on data with some percentage of errors. The dataset we have used for our project is from Anand Vihar-Delhi from 2014-17.

The first algorithm which we have used is Decision tree algorithm, which acts like a decision support tool and uses a tree like model of decisions consisting of their possible consequences, chance event outcomes, research costs and utility. It is used for classification and regression problems across the globe. When we implemented decision tree algorithm on our training dataset, it produced mean square error (MSE) of 282.211 and root square error of 607.067. Simultaneously, on execution of decision tree algorithm on testing data, it produced mean square error (MSE) of 0.969 and root square error of 0.941. The conclusion obtained from the above scenario is that the model is getting over fitted. Overfitting is a concept which occurs in regression algorithm where the model learns the details of the training data to an extend where it negatively impacts the performance of the model on new data.



Regression

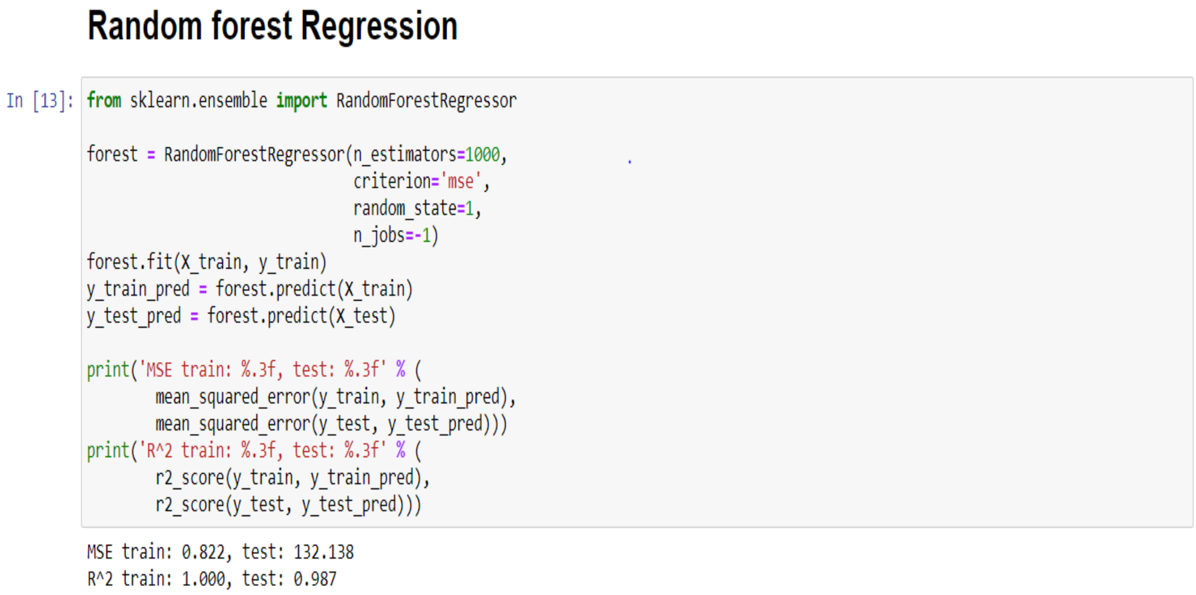
Decision Tree Regression

```
In [12]: from sklearn.tree import DecisionTreeRegressor

         tree = DecisionTreeRegressor(max_depth=3)
         tree.fit(X_train, y_train)

         y_train_pred = tree.predict(X_train)
         y_test_pred = tree.predict(X_test)

         print('MSE train: %.3f, test: %.3f' % (
                 mean_squared_error(y_train, y_train_pred),
                 mean_squared_error(y_test, y_test_pred)))
         print('R^2 train: %.3f, test: %.3f' % (
                 r2_score(y_train, y_train_pred),
                 r2_score(y_test, y_test_pred)))
```

```
MSE train: 282.211, test: 607.076
R^2 train: 0.969, test: 0.941
```

Fig 3. Implementation of Decision tree regression

The second algorithm which we implemented is Random forest regression used for both, classification and regression problems as well. It operates by constructing a multitude of decision trees at training time and outputting the class and outputting the class that is a mode of the classes or mean or average prediction of the individual trees. When we implemented random forest regression on our training dataset, it produced mean square error (MSE) of 0.822 and root square error of 1.000. Simultaneously, on execution of random forest on testing data, it produced mean square error (MSE) of 132.138 and root square error of 0.987. Therefore, we can deduce the conclusion that random forest regression also exhibits overfitting.



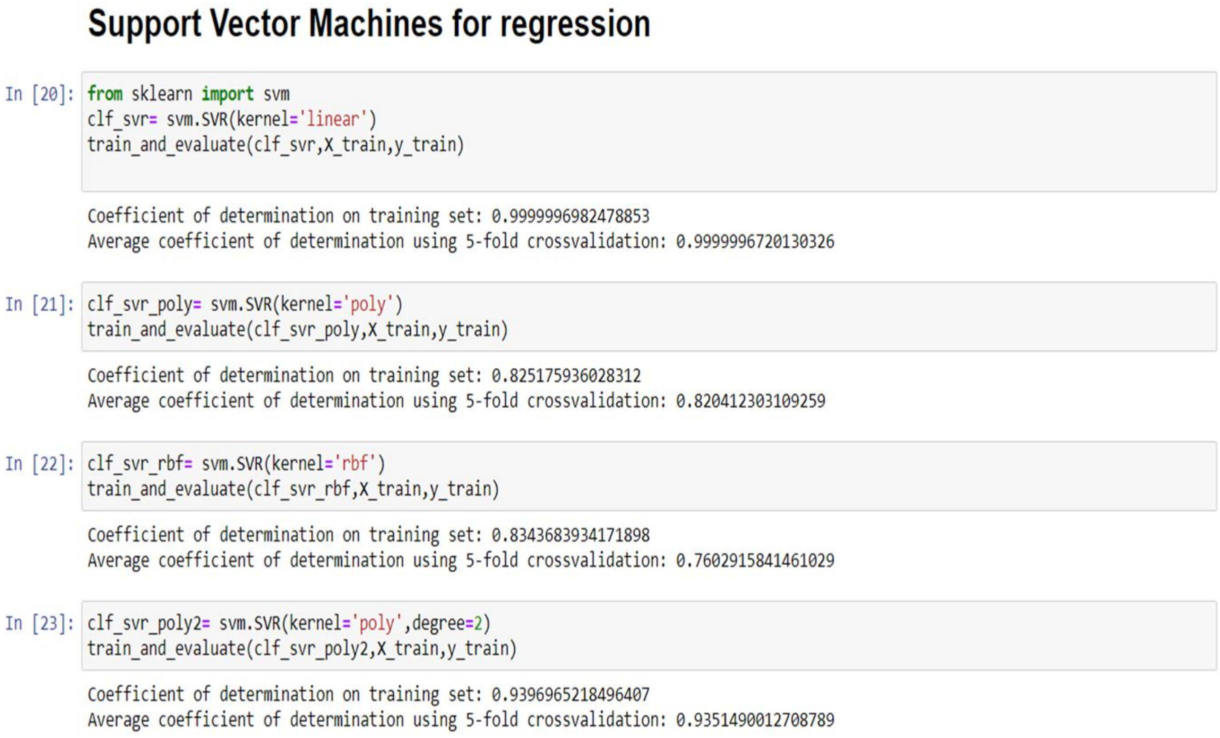
Random forest Regression

```
In [13]: from sklearn.ensemble import RandomForestRegressor

         forest = RandomForestRegressor(n_estimators=1000,
                                        criterion='mse',
                                        random_state=1,
                                        n_jobs=-1)
         forest.fit(X_train, y_train)
         y_train_pred = forest.predict(X_train)
         y_test_pred = forest.predict(X_test)

         print('MSE train: %.3f, test: %.3f' % (
                 mean_squared_error(y_train, y_train_pred),
                 mean_squared_error(y_test, y_test_pred)))
         print('R^2 train: %.3f, test: %.3f' % (
                 r2_score(y_train, y_train_pred),
                 r2_score(y_test, y_test_pred)))
```

```
MSE train: 0.822, test: 132.138
R^2 train: 1.000, test: 0.987
```

Fig 4. Implementation of Random Forest Regression

Lastly, we implemented Support Vector Machine Algorithm which is used for solving linear and non-linear problems. During its implementation, we have made use of three kernels which are linear kernel, poly kernel and rbf kernel. A linear kernel is used when the data is linearly separable i.e. it can be separated using a single line. A polynomial kernel represents a similarity of vectors in a feature space over polynomials of the original variables, allowing learning of non-linear models. RBF kernel is a function whose value depends on the distance from the origin or from some point. We have obtained coefficient of determination from different kernels which we have mentioned and according to that we can infer that linear kernel has exhibited the most promising results.



Support Vector Machines for regression

```
In [20]: from sklearn import svm
         clf_svr= svm.SVR(kernel='linear')
         train_and_evaluate(clf_svr,X_train,y_train)

         Coefficient of determination on training set: 0.9999996982478853
         Average coefficient of determination using 5-fold crossvalidation: 0.9999996720130326

In [21]: clf_svr_poly= svm.SVR(kernel='poly')
         train_and_evaluate(clf_svr_poly,X_train,y_train)

         Coefficient of determination on training set: 0.825175936028312
         Average coefficient of determination using 5-fold crossvalidation: 0.820412303109259

In [22]: clf_svr_rbf= svm.SVR(kernel='rbf')
         train_and_evaluate(clf_svr_rbf,X_train,y_train)

         Coefficient of determination on training set: 0.8343683934171898
         Average coefficient of determination using 5-fold crossvalidation: 0.7602915841461029

In [23]: clf_svr_poly2= svm.SVR(kernel='poly',degree=2)
         train_and_evaluate(clf_svr_poly2,X_train,y_train)

         Coefficient of determination on training set: 0.9396965218496407
         Average coefficient of determination using 5-fold crossvalidation: 0.9351490012708789
```

Fig 5. Implementation of SVM

In this work, we have analyzed and compared four existing schemes for solving the air pollution prediction issue. e. The techniques were Decision Tree algorithm, Random Forest regression, support-vector machine. We have compared the techniques with respect to rmse and mse. The results demonstrated decision tree and random forest algorithm does not completely validate each other because they are giving 100 percent results which signifies that the model is getting over-fitted. On the other hand, SVM is not suitable for large datasets and does not perform well when the data set has more i.e. target classes are overlapping. These drawbacks have been tackled by LSTM. We constructed a LSTM model and we executed the model on our dataset.

Our model is giving an rmse value of 0.9716 which indicates that accuracy of our model is pretty good, this shows that air quality prediction of our model will be almost perfectly accurate. We will try to import this model on a flask framework for it to be more interactive as well as user friendly.

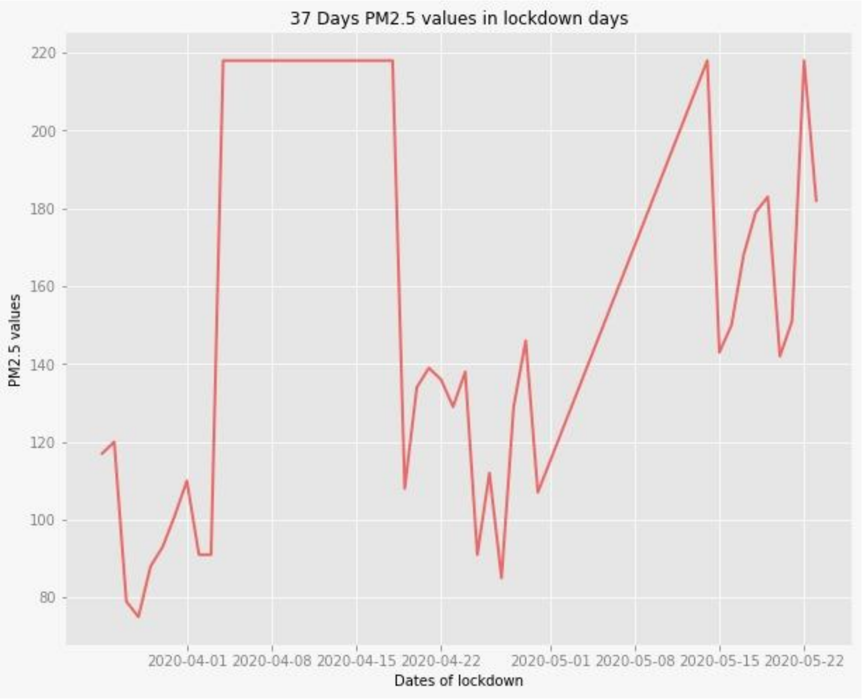


Fig 6: 37 days PM 2.5 values in lockdown days

Through this graph, we have mapped the PM2.5 values for a particular period during lockdown. This graph indicates that the AQI value was relatively low initially but increased gradually due to some reasons.

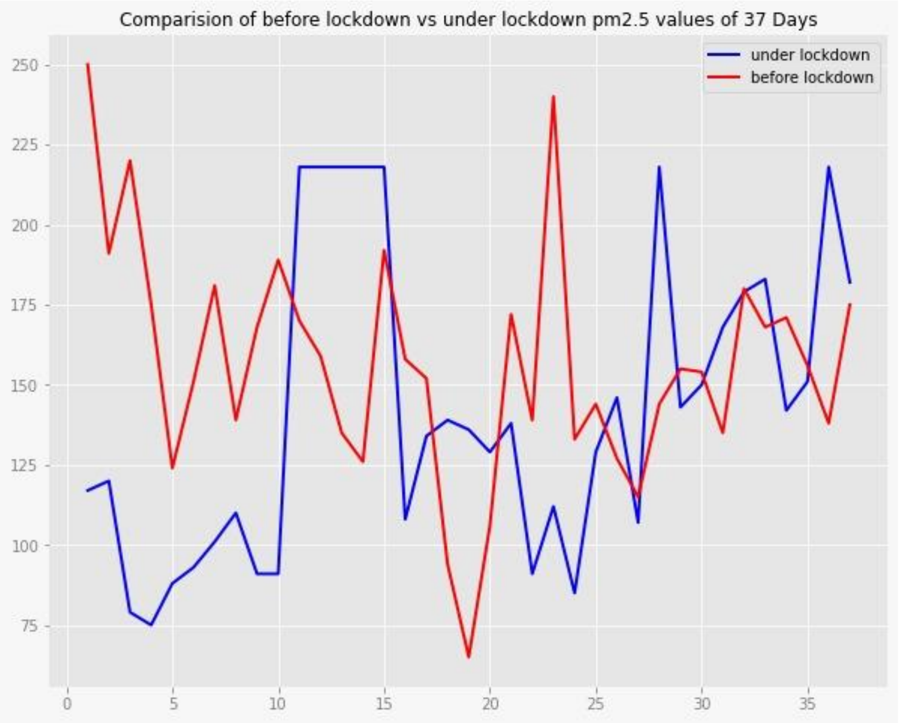


Fig 7: Comparison of before lockdown vs under lockdown of 37 days

This graph indicates that though the AQI values increased during the lockdown period but still the amount of pollution was lesser than it would have been before the lockdown or normal days. The red line indicates the AQI values that would have been true before the lockdown and the blue line indicates the actual AQI values that were mapped during the lockdown period.

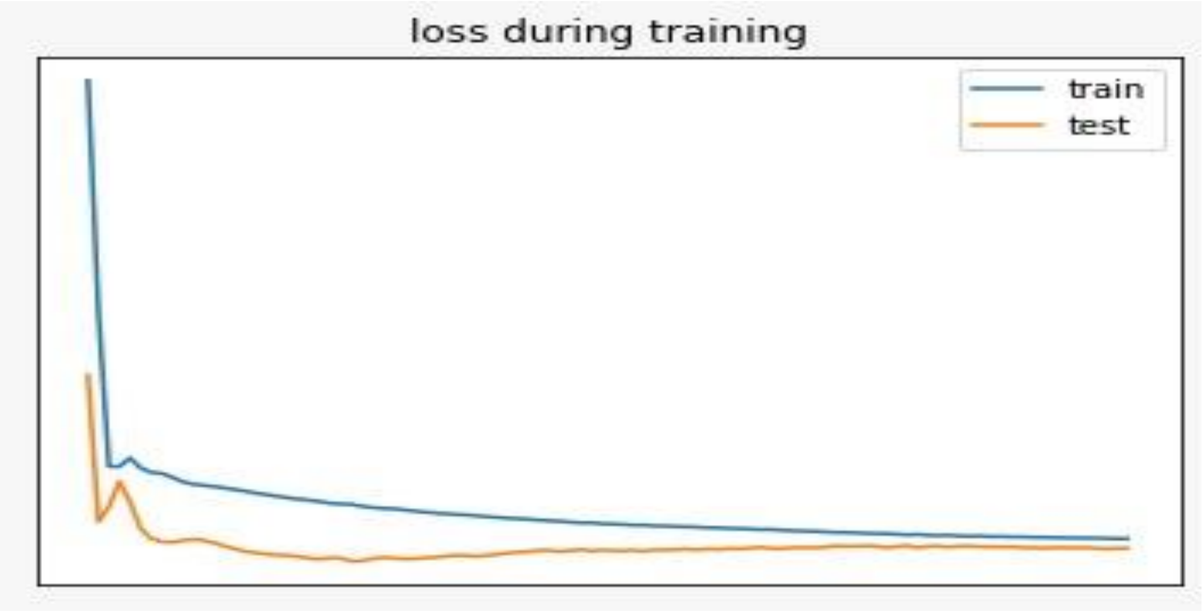
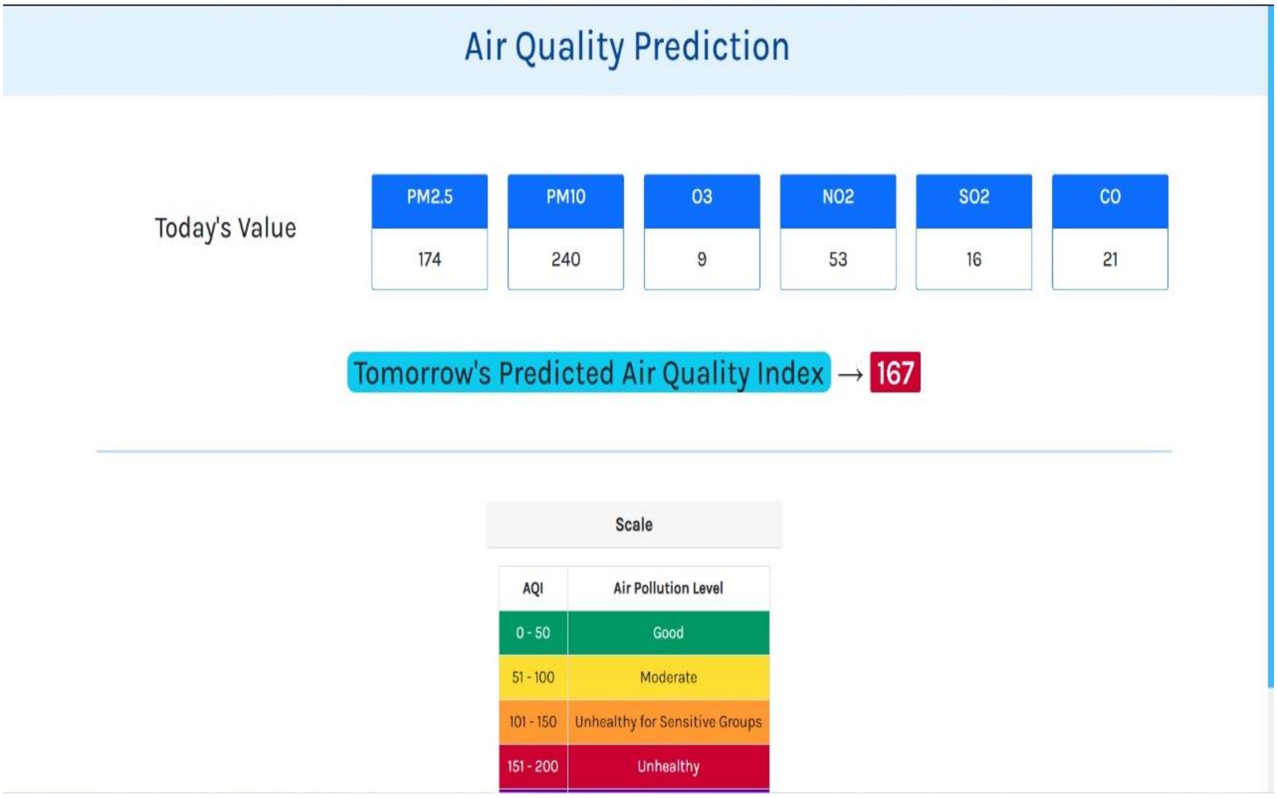


Fig 8: Loss during Training

Here is the graph indicating the loss occurred during the training and testing of the data. From the graph we can infer that, initially the loss occurred from training dataset is comparatively more than testing dataset. After consecutive iterations, as the weights were updated, we observed that the loss occurred from both the datasets reduced to minimal extent.



Air Quality Prediction

	PM2.5	PM10	O3	NO2	SO2	CO
Today's Value	174	240	9	53	16	21

Tomorrow's Predicted Air Quality Index → 167

Scale

AQI	Air Pollution Level
0 - 50	Good
51 - 100	Moderate
101 - 150	Unhealthy for Sensitive Groups
151 - 200	Unhealthy

Fig 9: Interface through which the predicted values are displayed.

We have integrated predicted values with the flask which will be displayed on this html interface. From the above image, we get our final AQI value, and it is also supported with particular ranges which indicates the severity of the air pollution.

Chapter 7

Conclusion

Thus, we have implemented a fairly accurate model with machine learning and analytics with fairly simple, user-friendly and fundamental implementation. Through this project, we have made a collective effort to demonstrate a way through which we can implement technology to convey the effects of air pollution on our day to day lives. This project will benefit our community and will evoke a sense of awareness to take preventive measures against air pollution so that our earth will be a better place to live in.

Literature Cited

[1] Mrs. A. Gnana Soundari, Mrs. J. Gnana Jeslin, Akshaya A.C , Indian Air Quality Prediction and Analysis using Machine Learning

[2] Saba Ameer, Munam Ali Shah, Abid Khan, Houbing Song, Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities

[3] Aditya C R, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu, Detection and Prediction of Air Pollution using Machine Learning Models

[4] Heni Patel, Swarndeep Saket, Air Pollution Prediction System for Smart city using Data Mining Technique

[5] Zhiying Meng, Ground Level Prediction Using Machine Learning

[6] K. Mahesh Babu, J. Rene Beulah, Air Pollution Prediction Using Machine Learning Supervised Learning Approach.

[7] *Gaganjot Kaur Kang, Jerry Zeyu Gao, Sen Chiao, Shengqiang Lu, and Gang Xie*, Air Quality Prediction: Big Data and Machine Learning Approaches.

[8] Madhuri VM, Samyama Gunjal GH, Savitha Kamalapurkar, Air Pollution Prediction Using Machine Learning Supervised Learning Approach.

[9] Fabiana Martins Clemente, Aleš Popovič, Sara Silva and Leonardo Vanneschi, A Machine Learning Approach to Predict Air Quality in California.

[10] Khushi Maheshwari, Sampada Lamba, Air Quality Prediction using Supervised Regression Model