

Predicting air quality in Delhi due to pollution from bordering states

A report submitted in partial fulfillment of the requirements

Of

Mini-Project (ISL64)

In

Sixth Semester

By

1MS17IS105 Sarthak Jain

1MS17IS124 Tilak Singh

1MS17IS133 Vinay Biradar

1MS18IS416 Umer Faruk

Under the guidance of

S.R Mani Sekhar

Asst. Professor

Dept. of ISE, RIT



RAMAIAH

Institute of Technology

DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

RAMAIAH INSTITUTE OF TECHNOLOGY

(AUTONOMOUS INSTITUTE AFFILIATED TO VTU)

M. S. RAMAIAH NAGAR, M. S. R. I. T. POST, BANGALORE - 560054

2019-2020

RAMAIAH INSTITUTE OF TECHNOLOGY

(Autonomous Institute Affiliated to VTU)

M. S. Ramaiah Nagar, M. S. R. I. T. Post, Bangalore – 560054

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING



RAMAIAH
Institute of Technology

CERTIFICATE

This is to certify that the project work entitled “**Predicting air quality in Delhi due to pollution from bordering states**” is a Bonafede work carried out by **Sarthak Jain** bearing **USN: 1MS17IS105**, **Tilak Singh** bearing **USN: 1MS17IS124**, **Vinay Biradar** bearing **USN:1MS17IS133**, **Umer Faruk** bearing **USN:1MS18IS416** in partial fulfillment of requirements of Mini-Project (ISL64) of Sixth Semester B.E. It is certified that all corrections/suggestions indicated for internal assessment has been incorporated in the report. The project has been approved as it satisfies the academic requirements in respect of project work prescribed by the above said course.

Signature of the Guide

Mr. S R Mani Sekhar

Asst. Professor
Dept. of ISE, RIT,
Bangalore-54

Signature of the HOD

Dr. Vijaya KumarBP

Professor and Head,
Dept. of ISE, RIT
Bangalore-54

Other Examiners

Name of the Examiners:

- 1.
- 2.

Signature

Acknowledgement

We have taken efforts in successfully completing this project. However, it would not have been possible without the kind support and help of many individuals and the organizations. We would extend our sincere gratitude to everyone who helped us in completing this project.

We are also grateful to our Institution, Ramaiah Institute of Technology with its ideals and inspiration for having provide us with the facilities, which has made this project a success.

We are pleased to acknowledge **Mr. S.R. Mani Sekhar, Assistant Professor**, Department of ISE for his invaluable guidance during the course of this project work, without his guidance, this project would have been an uphill task.

We take this opportunity to extend our gratitude to our Head of Department, **Dr. Vijaya Kumar B P**, who co-operated with us regarding some issues and for his invaluable information, exemplary guidance, monitoring and constant encouragement through the course of this project. We would like to extend our hearty gratitude to **Dr. N V Naidu**, the principal, for the kind support and permission to use the facilities available in the Institute.

Lastly, we would also like to thank all the faculties of ISE for their co-operation for the smooth development of this project.

May 2020

Sarthak Jain (1MS17IS105)

Tilak Singh (1MS17IS124)

Vinay Biradar (1MS17IS133)

Umer Faruk (1MS18IS416)

Abstract

Urban air pollution prediction becomes a compelling alternative to restrain its detrimental consequences. A thick layer of smog blankets can be seen in Delhi in the month of November which is the main season of stubble burning. Out of all the parameters in analyzing air quality, we considered the change in $PM_{2.5}$ data to analyze due to the burning. Particulate Matter (PM) which is generally measured in terms of the mass concentration of particles within certain size classes: PM_{10} or coarse (with an aerodynamic diameter of less than 10 micron) and $PM_{2.5}$ or fine (with an aerodynamic diameter of less than 2.5 micron). $PM_{2.5}$ has a greater residing time in air when compared to PM_{10} because of the balance between the downward acting force of gravity and aerodynamic drag force. Therefore, Particulates are the main component of air pollution. Numerous machine learning techniques have been adopted to forecast the air quality. But none on them focuses primarily on the issues of stubble burning and its effect on Delhi's air. In this paper, we present an attempt to estimate the value of $PM_{2.5}$ in the National capital mainly due to stubble burning in neighboring states. For this the $PM_{2.5}$ data is taken into consideration for 9 ground-based continuous air quality monitoring stations in the neighboring states of Delhi for the duration of 6 months & 15 days in the calendar year of 2019 to analyze the agricultural reason behind the slumping of air quality in Delhi day by day. The model used here is based on Stacking regression which is an ensemble learning technique to combine the prediction capabilities of multiple regression models via a meta-regressor. The individual regression models or base models are trained using the out-of-folds predictions based on the complete training set then, the meta-regressor is trained based on the outputs (meta-features) of the individual regression models in the ensemble. The technique is then evaluated by RMSE (Root Mean Square Error), MSE (Mean absolute Error), R^2 (R-squared) and MAE (Mean Average Error). Finally, the proposed model is validated with the real-time data of Delhi for the stubble burning period. The proposed model will illustrate the effect of stubble burning and give us an estimate of $PM_{2.5}$ based on the $PM_{2.5}$ values of stubble burning hotspots in Haryana and Punjab which will help us to predict the extent of deterioration in air quality in case of increase in stubble burning.

Contents

1 Introduction	1
1.1 Motivation	1
1.2 Scope	2
1.3 Objectives	2
1.4 Proposed Model	3
1.5 Organisation of Report	3
2 Literature Review	4
3 System Analysis and Design	
3.1 Workflow	7
3.1.1 Data Description	8
3.1.2 Data Preprocessing	8
3.2 Performance Check Measures	9
3.3 Technical Stack	11
4 Modelling and Implementation	12
4.1 Algorithms	
4.1.1 Stacked Regression	13
4.1.2 XGBoost	26
4.1.3 Boosting Gradient	27
5 Testing, Results and Discussion	23
5.1 Testing	28
5.2 Results	28
5.3 Numerical Values	34
5.4 Interface	34
6 Conclusion and Future Work	35
References	36

Chapter 1

Introduction

1.1 Motivation

A country is said to be developed when there is prosperity i.e. people have food to eat, a hygienic water to drink and a good air to breath. The latter two are the biggest challenges for any country. Due to urbanization and construction of buildings across the world there is a degradation in the quality of both water and air and hence causing water and air pollution respectively. In our project we mainly focused on air pollution, and found out that if the degradation continues like this, then people will start experiencing serious health problems in a matter of 10 years or so. So, it is our duty to start contributing towards the reduction of air pollution else we ourselves will be affected by it. In our quest for reducing the air pollution, we came across data of air quality in Delhi and observed that, for the past few years the air quality index of Delhi is deteriorating day by day. The people are suffering from diseases such as lung cancer, asthma, heart diseases, etc. The Delhi government proposed a report in which it was very clear that the quality of air in Delhi is at its lowest in the history. This motivated us to analyze the reason behind it. In our examination we found out that besides all the traditional fuel-based pollution, pollution due mining, etc., one of the noteworthy reasons which fascinated us the most was the effect of Stubble burning due to bordering states like Punjab and Haryana. In our indagation we found out that the agricultural burning that takes place in states like Haryana and Punjab causes the flow of harmful substances from their region to Delhi. Furthermore, according to the information provided by Indian government it is evident that the stubble burning has 35% [\[1\]](#) effect on pollution in Delhi. So, the very fact that we can examine this effect, its consequence on atmosphere and its repercussions on humans motivated us to implement this project so that we could aware the people and the government to take necessary steps in controlling the air pollution in Delhi and making it a better place to live.

1.2 Scope

The main user base for this project are central and state government to estimate the $PM_{2.5}$ value with the input from neighboring states $PM_{2.5}$ data. With the help of our model they can maintain the regulations on stubble burning and reduce the pollution. The primary goal of this project is to predict the effect of burning on humans, since there are many factors in examining like values of NO_2 , SO_2 , CO , $PM_{2.5}$ etc. we choose to predict the change in values of $PM_{2.5}$. In future implementation we might take other parameters in-consideration like change in the values of NO_2 , SO_2 , CO_2 , CO etc.

1.3 Objective

The final goal of the project was as follows.

1. Collecting the $PM_{2.5}$ data of region causing stubble burning and estimating $PM_{2.5}$ value in target region.
2. After estimating $PM_{2.5}$ value, the objective of this project is to examine the repercussion of stubble burning in target area on human beings such as what fatal diseases can be spread because of it.

1.4 Proposed Model

The user will be provided a web-based application platform in which they will be able to input the data of $PM_{2.5}$ for the cities like Panipat, Sirsa, Bhiwani, Rohtak, Patiala, Ludhiana, Kaithal, Karnal, Jind. Based on the input provided we will be estimating the value of $PM_{2.5}$ for the places in Delhi and quality of air for the human health.

1.5 Organization of Report

In order to explain the developed model, the following sections are covered:

- 1) **Literature Review** describing the study of the existing systems and techniques taken into account prior to development of the proposed system.
- 2) **System Analysis and Design** provides a detailed walk through of the software engineering methodology adopted to implement the model, an overview of the system and the modules incorporated into system.
- 3) **Modelling and Implementation** provides a deeper insight into the working model. The various modules and their interactions are depicted using descriptive diagrams.
- 4) **Testing** the model to ensure bug/error free model along with the **Results** obtained. **Discussions** then provides detailed analysis on quality assurance measures.
- 5) **Conclusion** about the results obtained after successfully running the model and **Future Scope** of the model is highlighted.

Chapter 2

Literature Review

Stubble burning^[6] is intentionally setting fire to the straw stubble that remains after grains, like paddy, wheat, etc., have been harvested. The practice was widespread until the 1990s, when governments increasingly restricted its use.

Stubble burning in Punjab and Haryana in northwest India has been cited as a major cause of air pollution in Delhi. Consequently, the government is considering implementation of the 1,600 km long and 5 km wide Great Green Wall of Aravalli. In late September and October each year, farmers mainly in Punjab and Haryana burn an estimated 35 million tons of crop waste from their paddy fields after harvesting as a low-cost straw-disposal practice to reduce the turnaround time between harvesting and sowing for the second (winter) crop. Smoke from this burning produces a cloud of particulates visible from space and has produced a what has been described as a "toxic cloud" in New Delhi, resulting in declarations of an air-pollution emergency. For this, the NGT (National Green Tribunal) instituted a fine of Rs. 2lac on the Delhi Government for failing to file an action plan providing incentives and infrastructural assistance to farmers to stop them from burning crop residue to prevent air pollution.

Although harvesters such as the Indian-manufactured "Happy Seder" that shred the crop residues into small pieces and uniformly spread them across the field are available as an alternative to burning the crops, some farmers complain that the cost of these machines is prohibitive compared to burning the fields.

Helpful Effects

- Kills slugs and others pests.
- Can reduce nitrogen.

Harmful Effects

- Loss of nutrients.
- Pollution from smoke.
- Damage to electrical and electronic equipment from floating threads of conducting waste.
- Risk of fire spreading out of control.

In the paper, “Assessment of contribution of agricultural residue burning on air quality of Delhi using remote sensing and modelling tools^[8]”, the authors proposed the model in which they observe the higher baseline value in Delhi. Local meteorological conditions of the regions like Punjab and Haryana were also observed. The observation used remote sensing for analysis in which it was found that the particulates transmission from the region of Punjab and Haryana is not likely to happen, but when the study was led by the usage of Aerosol Robotics Network (AERONET) it showed that some carbonaceous aerosol subtype was identified over Delhi region.

In the Hybrid Single Particle Lagrangian Integrated Trajectory Model (HYSLIT) proposed, it was showed that northwesterly winds intersect the agricultural residue burning regions, which might have transported the burnt stubble particulates towards Delhi during the months of October or November.

Research was performed by Dipti Grover and Smita Chaudhary on “Ambient air quality changes after stubble burning in rice-wheat system in an agricultural state of India^[9]” in which it was concluded that it is clearly evident that there is a significant increase in concentration levels of $PM_{2.5}$, SO_x , and NO_x during crop residue burning period as compared to non-burning period. In their study, they under took three regions of Haryana and analysed their $PM_{2.5}$, SO_x , and NO_x values and its effect on Delhi. In research, the AQI and concentration of primary pollutants were determined in rice and wheat crop season. During the research it was founded that the concentration of the three parameters taken into consideration exceeded the NAAQS values by 78%, 71% and 53% respectively. There was a striking increase in $PM_{2.5}$ value by 3.5 times.

In the paper “Ambient air Quality during wheat and rice crop stubble burning episodes in Patiala [\[10\]](#)” the researchers studied the stubble burning practices in five different regions of Patiala to find the level of change in aerosol, SO₂, NO₂. Aerosols were collected on GMF/A and QMF/A sheets for 24 hrs along with separate collection of SO₂ and NO₂ throughout the year and the results obtained during stubble burning periods were compared with non-stubble burning period. The results clearly pointed out that there were an increase in the levels of concentration of aerosol, SO₂, NO₂ during stubble burning period.

Chapter 3

System Analysis and Design

3.1 Workflow

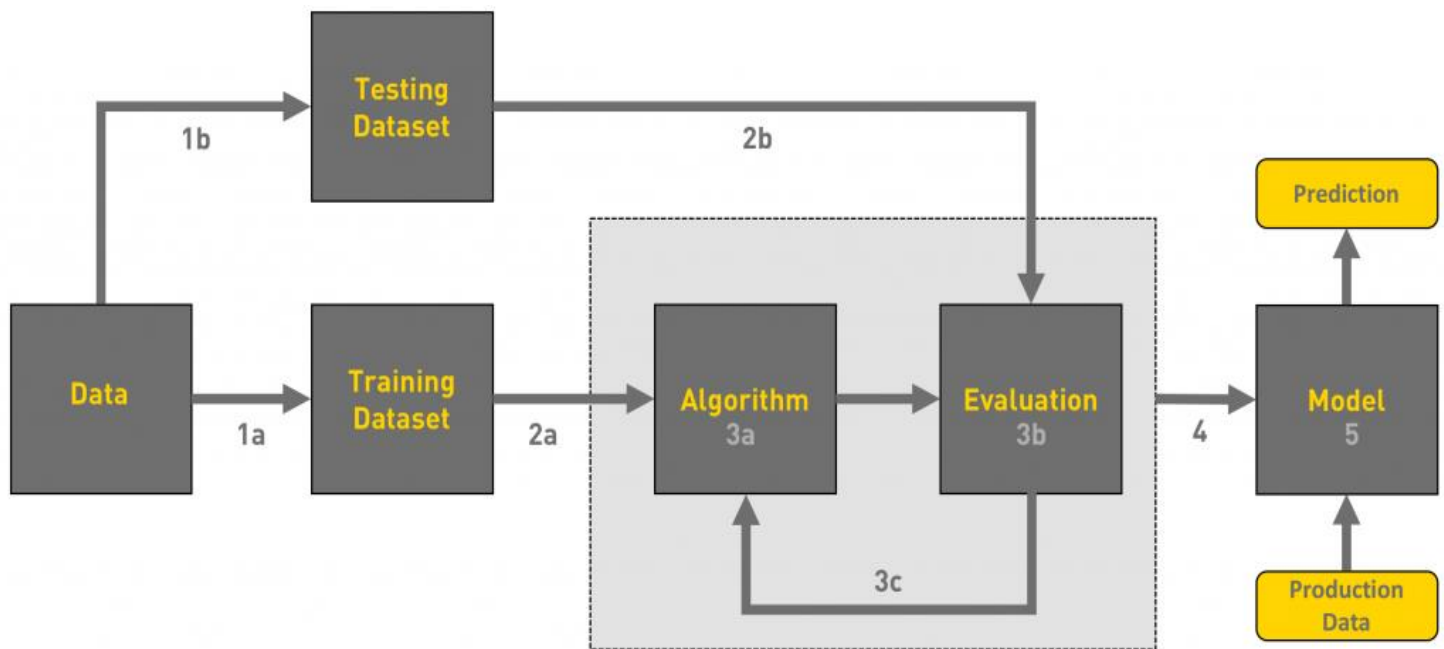


Fig: 3.1 Stages in Implementation of algorithm^[14]

We can define the machine learning workflow in 3 stages.

1. Gathering data.
2. Data pre-processing
3. Researching the model that will be best for the type of data
4. Training and testing the model
5. Evaluation

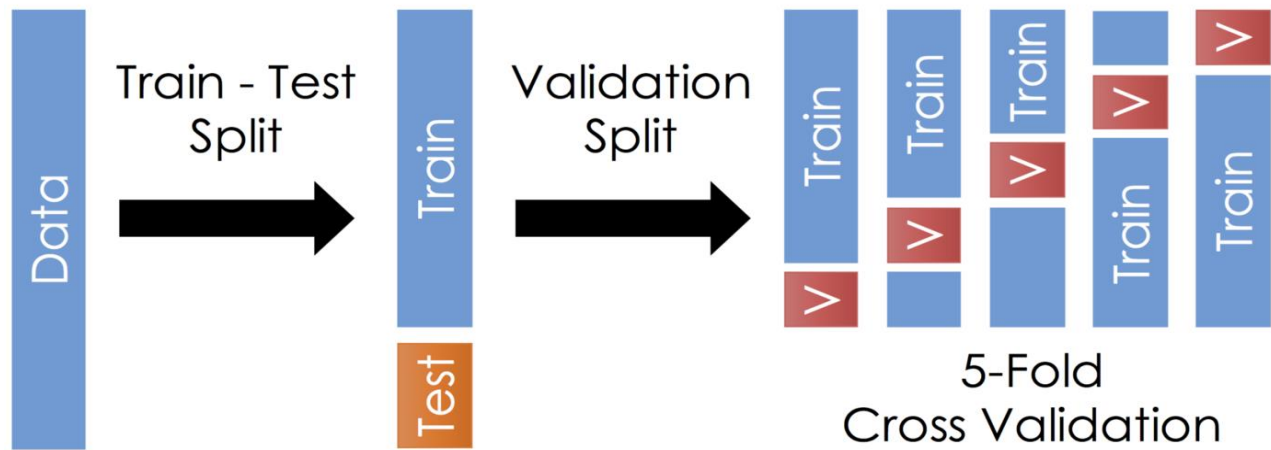


Fig: 3.2 Data split and implementation of 5-Fold Cross Validation

3.1.1 Data Description

The data source used in this project is taken from acqni.org ^[13] which is collecting air quality and air pollution data such as PM2.5 (fine particulate matter), PM10 (respirable particulate matter), NO2 (nitrogen dioxide), SO2 (sulphur dioxide), CO (carbon monoxide) and O3 (ozone) from Delhi Pollution Control Committee (Government of NCT of Delhi) and CPCB- India Central Pollution Control Board. The data is collected for Delhi and its surrounding states Haryana and Punjab. The data is stored in a file named “**revisedDataset.csv**”.

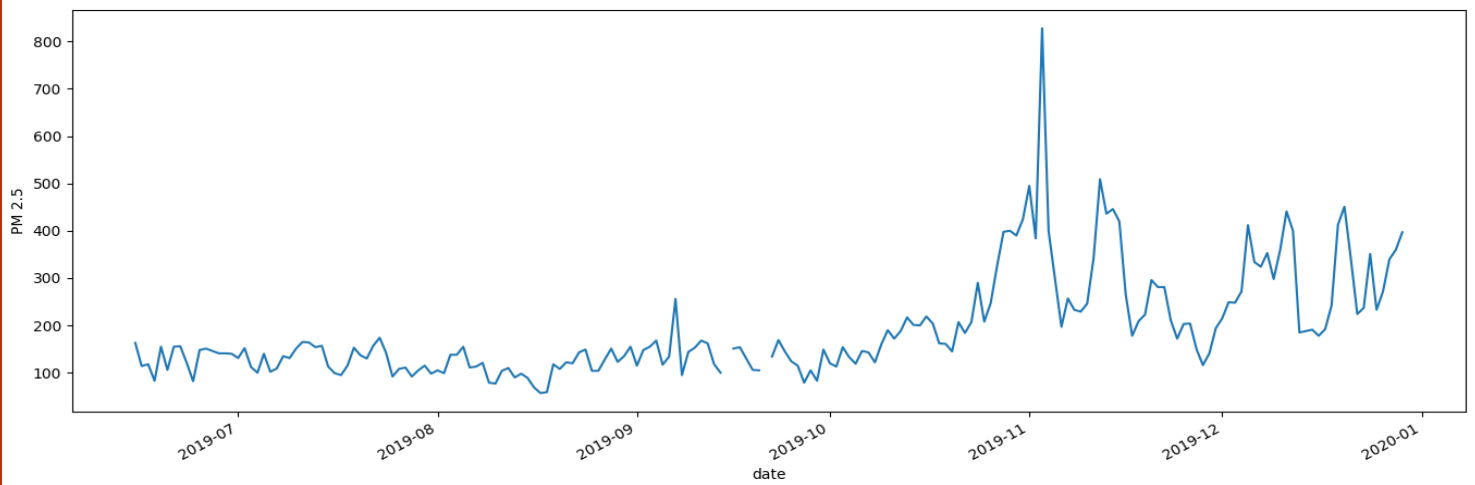


Fig 3.3- PM2.5 values of Delhi

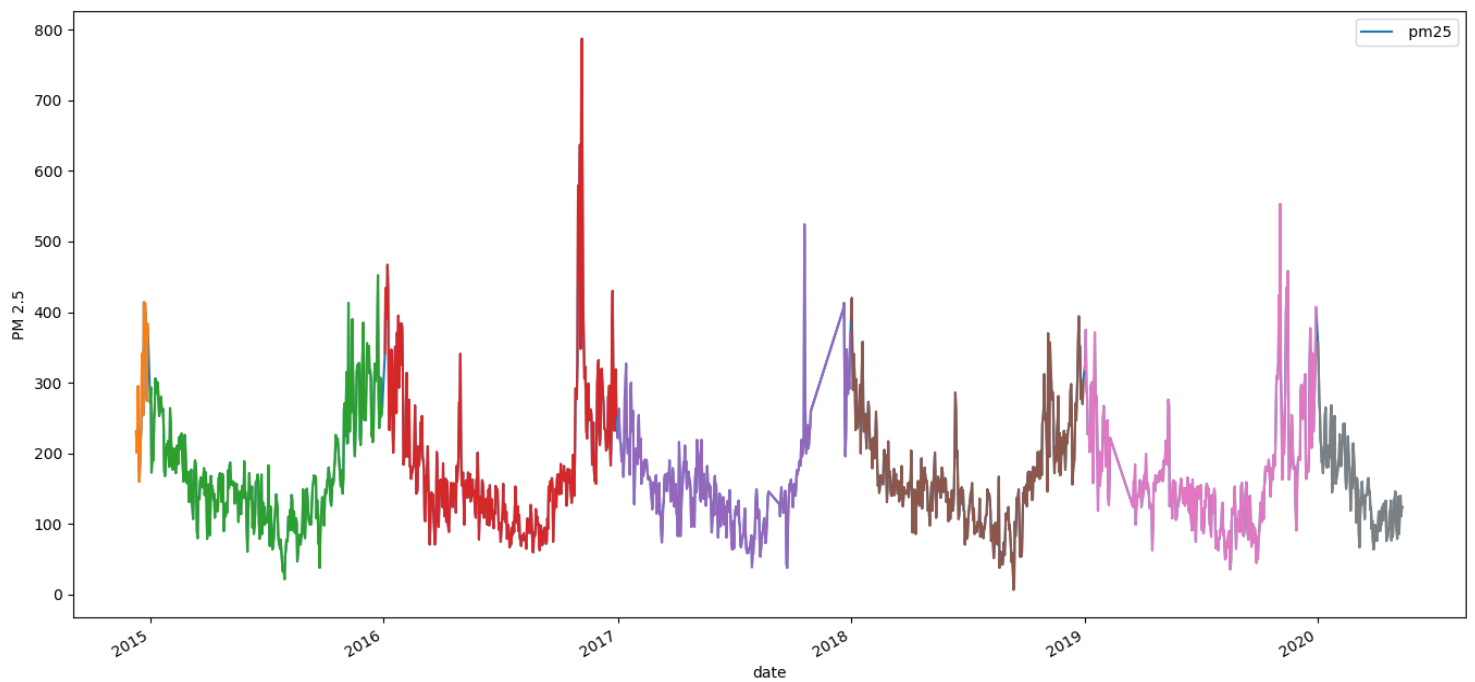


Fig: 3.4- PM2.5 value of Delhi highest during stubble burning period

3.1.2 Pre-processing of Data

Initially the columns contain the values contains the data of PM_{2.5} (fine particulate matter), PM₁₀ (respirable particulate matter), NO₂ (nitrogen dioxide), SO₂ (sulphur dioxide), CO (carbon monoxide) and O₃ (ozone). we have eliminated the fields of PM₁₀ (respirable particulate matter), NO₂ (nitrogen dioxide), SO₂ (sulphur dioxide), CO (carbon monoxide) and O₃ (ozone) because this attribute has weak correlation with the target label and the impact of stubble burning on ambient air quality significantly increases the PM₁₀ and PM_{2.5} concentrations during crop residue burning periods which in turn increases the PM₁₀ and PM_{2.5} concentrations which deteriorates the air quality in Delhi. The PM_{2.5} (fine particulate matter) data of Delhi and its surrounding states Haryana and Punjab is taken under observation. The cities of Haryana and Punjab taken under consideration are Hisar, Panipat, Bhiwani, Jind, Karnal, Kaithal, Rohtak, Ludhiana and Patiala. The dataset was consisting of NULL values, which can lead to a bias in the Results therefore, the dataset was cleaned by dropping unnecessary attributes and also, by dropping the tuples with null value of one or more attribute.

3.2 Performance Check Measures

In this project we measured the performance using MSE, RMSE, R squared and MAE. All the matrices are explained as follows.

3.2.1 RMSE

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). It is used to measure of the differences between values predicted by a model or an estimator and the values observed.

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}}$$

3.2.2 MAE

Mean absolute error (MAE) is a measure of errors between paired observations expressing the same phenomenon.

$$MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n} = \frac{\sum_{i=1}^n |e_i|}{n}.$$

3.2.3 R-squared

R-squared (R^2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

$$R\text{-Squared} = 1 - \frac{SS_{\text{regression}}}{SS_{\text{total}}}$$

3.2.4 MSE

MSE (Mean Squared Average) is the average of the squared error that is used as the loss function for least squares regression: It is the sum, over all the data points, of the square of the difference between the predicted and actual target variables, divided by the number of data points.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \tilde{y}_i)^2$$

3.3 Tech Stack

The technology stack used in the development of this project are as follows –







	Python: Interpreted high-level, general-purpose programming language.
	Sci-kit Learn: Python well documented Open Source ML Library
	Pandas: Software library written in Python for data manipulation and analysis.
	Flask: Micro-web framework written in python.
	CSS: Styling sheet used to design interface.
	HTML5: HTML5 is used for structuring and presenting content.

Fig 3.5 Technical Stack

Chapter 4

Modelling and Implementation

Implementation is the process of converting the designed system architecture into working modules where it is made sure that all the function and non-functional requirements are met.

4.1 Algorithms

Machine learning algorithms are programs that can learn from data and improve from experience, without human intervention. Learning tasks may include learning the function that maps the input to the output, learning the hidden structure in unlabeled data; or ‘instance-based learning’, where a class label is produced for a new instance by comparing the new instance (row) to instances from the training data, which were stored in memory. ‘Instance-based learning’ does not create an abstraction from specific instances.

4.1.1 Stacked Regression

Stacked regression is a method for forming combinations of different predictors to give improved predicted accuracy. The idea is to use cross-validation data and least square under non-negativity constraints to determine the coefficients in the combination. There are basically two models in Stack Regression, Base Model and Meta Model. The different predictors we can be classified into Base Model and Meta Model.

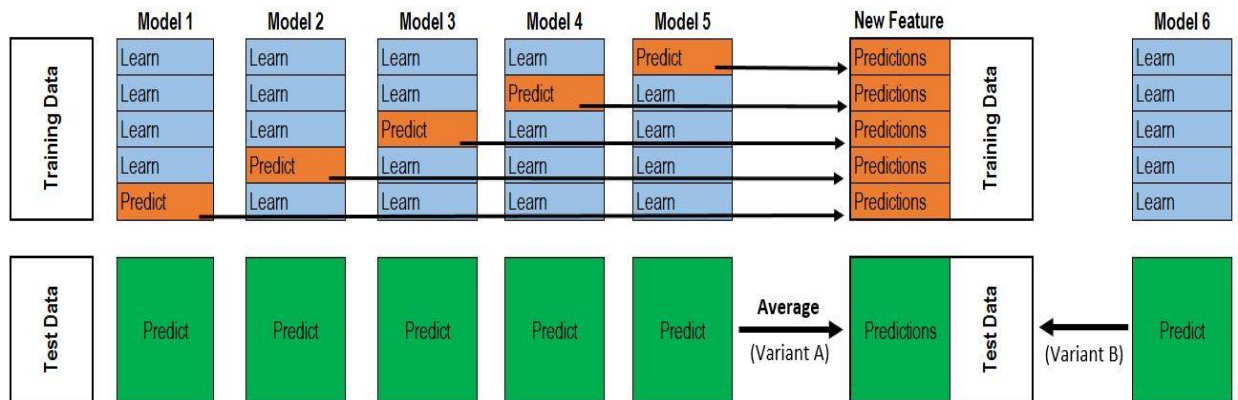


Fig 4.1

4.1.1(a) Base Model Stacked Regression

While analyzing the $PM_{2.5}$ values from Hisar, Panipat, Bhiwani, Jind, Karnal, Kaithal, Rohtak, Ludhiana and Patiala and comparing it with $PM_{2.5}$ values of Delhi the following graphs were obtained.

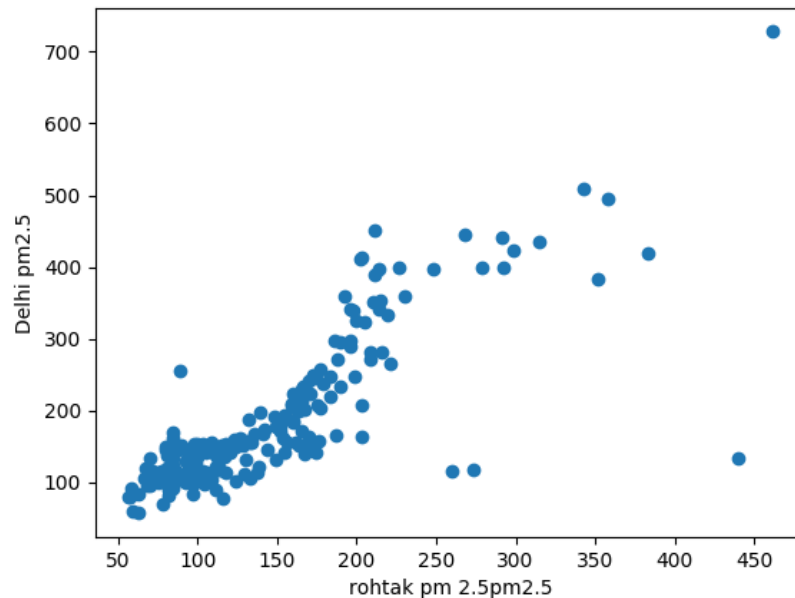


Fig: 4.2 Scattered graph of Delhi v/s Rohtak

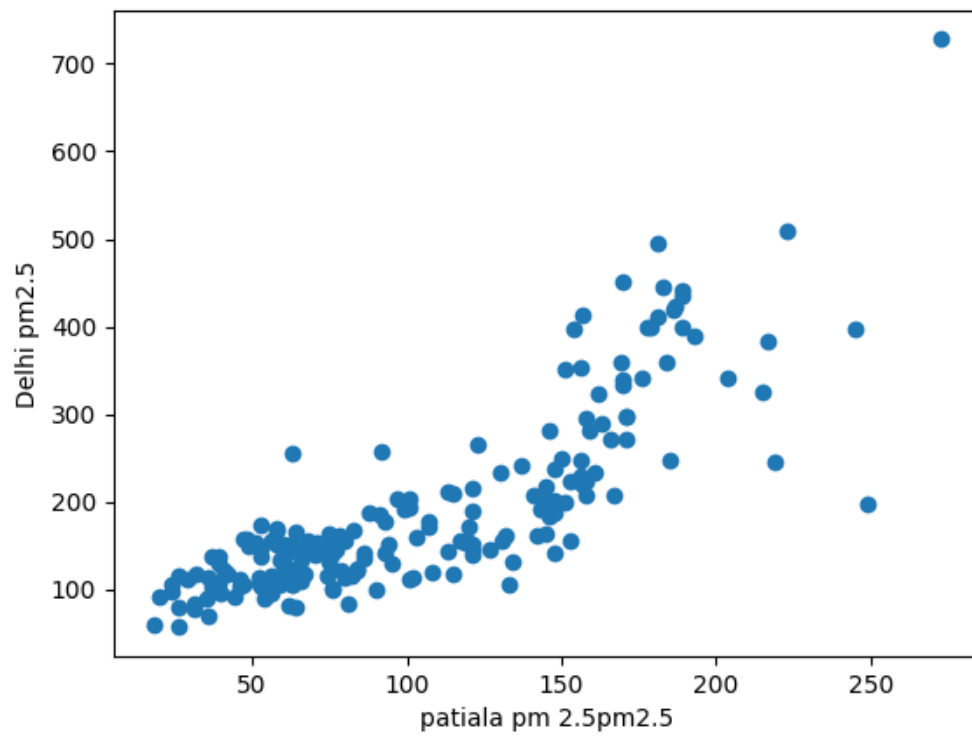


Fig 4.3 Scattered graph of Delhi v/s Patiala

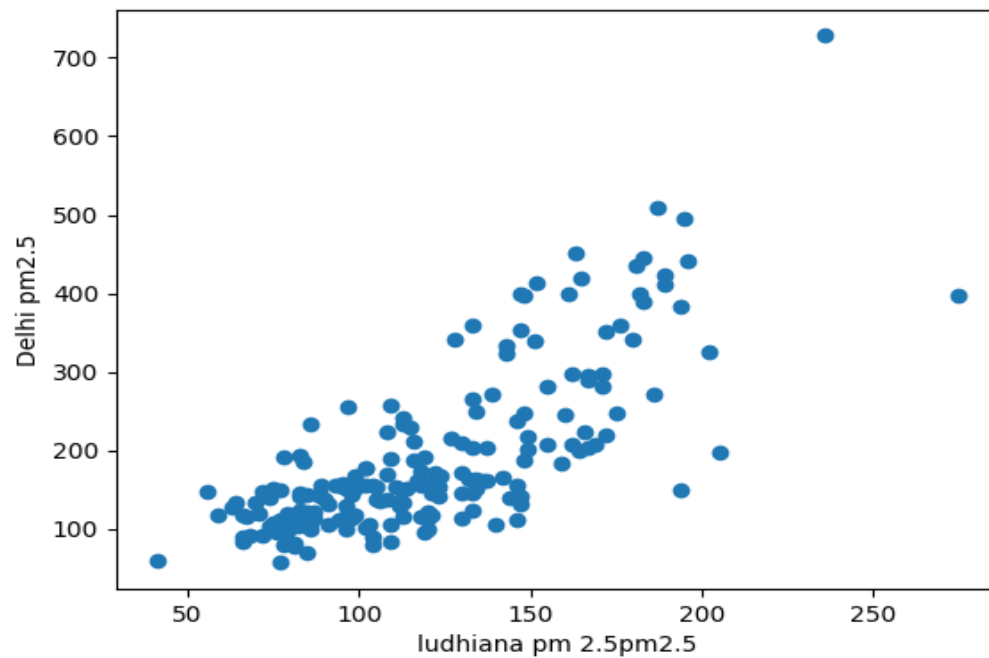


Fig: 4.4 Scattered graph of Delhi v/s Ludhiana

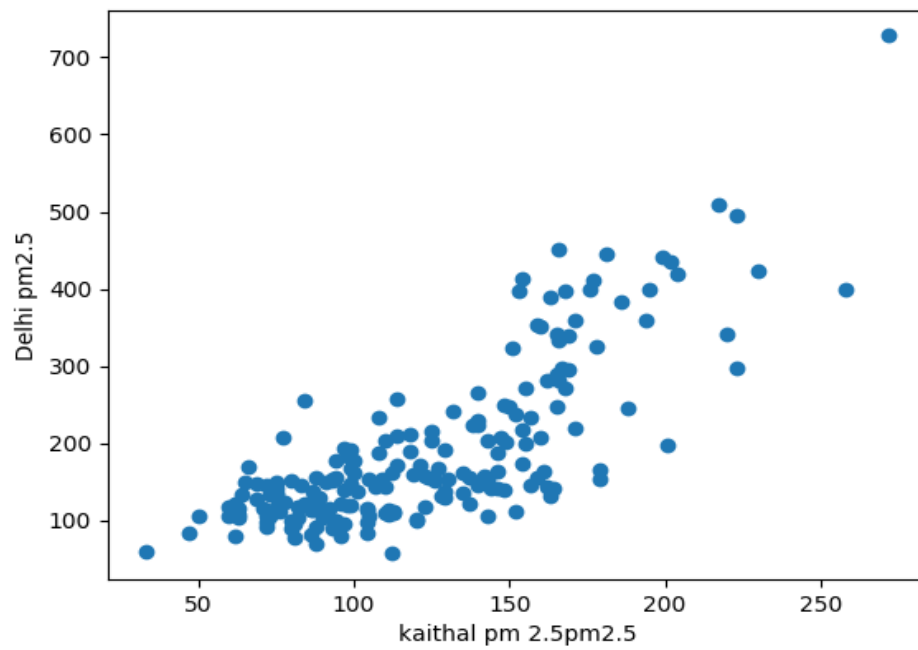


Fig: 4.5 Scattered graph of Delhi v/s Kaithal

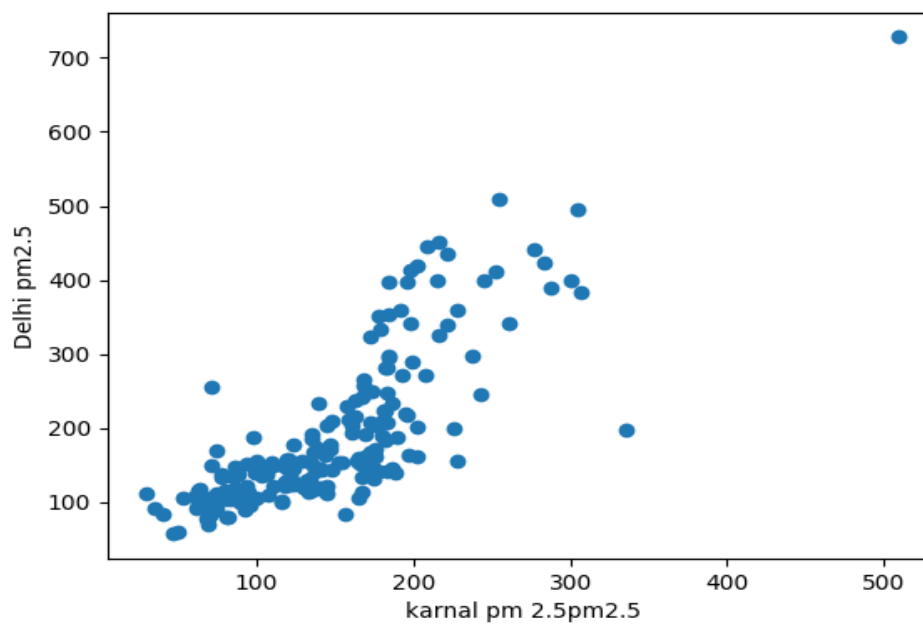


Fig: 4.6 Scattered graph of Delhi v/s Karnal

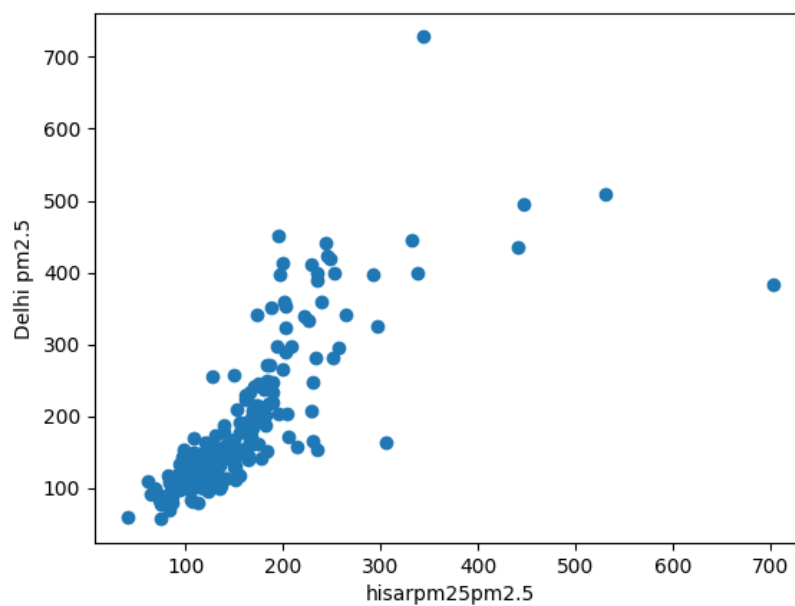


Fig: 4.7 Scattered graph of Delhi v/s Hisar

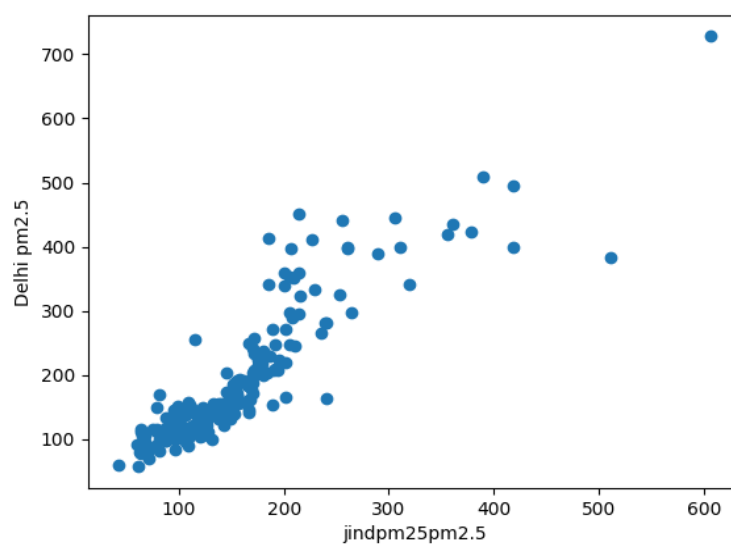


Fig: 4.8 Scattered graph of Delhi v/s Jind

It can be observed from the graph of comparison between PM_{2.5} of Delhi and other neighboring cities that the graph consists of both Linear and Non-Linear features. Hence an ensembled base model should be used to predict the values such that we can inculcate both Linear and Non-Linear features. Hence the following were implemented under Base Stacked Regression.

1) Linear Regression

Linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variable or independent variable. The case of one explanatory variable is called Simple Linear Regression. For more than one explanatory variable, the process is called Multiple Linear Regression.

To estimate the PM_{2.5} value of Delhi by taking all the nearest area's PM_{2.5} data Multi Linear Regression is taken into consideration since when the model is given all the value then the line obtained was not covering all the points hence Simple Linear Regression was not taken in calculating accuracy.

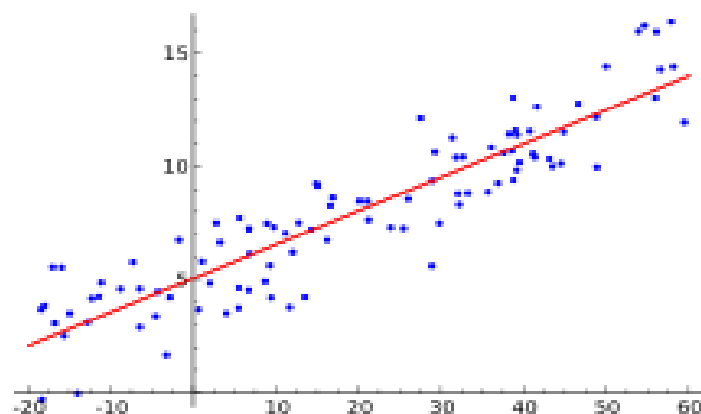


Fig: 4.9 Simple Linear Graph

2) Random Forest

Random forest is also known as random decision forests, are a popular ensembled method that can be used to build predictive models for both classification and regression problems. Ensemble methods use multi learning models to gain better predictive results – In the case of random forest, the model creates an entire forest of random uncorrelated decision tree to arrive at the best possible answer.

Pros:

- 1) Random forest has less variance than a single decision tree. It means that it can work correctly for a large range of data items than single decision trees.
- 2) Random forest is extremely flexible and have very high accuracy.
- 3) They also do not require preparation of input data. You do not have to scale the data.
- 4) Maintains accuracy even when a large proportion of the data are missing. S

Cons:

- 1) The main disadvantage of Random forest is their complexity. It is much harder and time-consuming to construct than decision trees.
- 2) It also requires more computational resources and are also less intuitive.
- 3) Time consuming than any other algorithm.

The pseudo code for the random forest algorithm can split into two stages.

- Random forest creation pseudo code.
- Pseudo code to perform prediction from the created random forest classifier.

First, let's try to begin with random forest creation pseudo code:

A. Random Forest Pseudo Code:

- a. Randomly select “k” features from total “m” features, where $k \ll m$.
- b. Among the “k” features, calculate the node “d” using the best split point.
- c. Split the node into daughter nodes using the best split.
- d. Repeat 1 to 3 steps until “l” number of nodes has been reached.
- e. Build forest by repeating steps 1 to 4 for “n” number times to create “n” trees.

B. Random Forest Prediction Pseudo Code:

To perform the prediction using the trained random forest algorithm uses the below pseudocode.

- a) Takes the test features and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target)
- b) Calculate the votes for each predicted target.
- c) Consider the high voted predicted target as the final prediction from the random forest algorithm.

3) K-nearest Neighbors

The K-nearest neighbors (KNN) algorithm is a type of supervised machine learning algorithms. It is a lazy learning algorithm since it doesn't have a specialized training phase. Rather, it uses all of the data for training while classifying a new data point or instance. KNN is a non-parametric learning algorithm, which means that it doesn't assume anything about the underlying data. This is an extremely useful feature since most of the real-world data doesn't really follow any theoretical assumption e.g. linear-separability, uniform distribution, etc.

Working:

The intuition behind the KNN algorithm is one of the simplest of all the supervised machine learning algorithms. It simply calculates the distance of a new data point to all other training data points. The distance can be of any type e.g. Euclidean or Manhattan etc. It then selects the K-nearest data points, where K can be any integer. Finally, it assigns the data point to the class to which the majority of the K data points belong.

Suppose you have a dataset with two variables, which when plotted, looks like the one as follows.

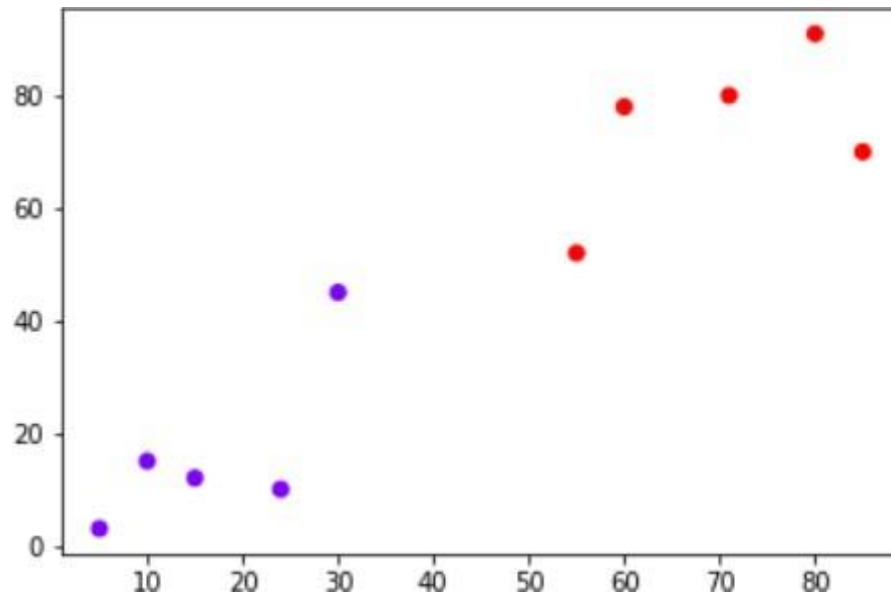


Fig 4.10

The task is to classify a new data point with 'X' into "Blue" class or "Red" class. The coordinate values of the data point are $x=45$ and $y=50$. Suppose the value of K is 3. The KNN algorithm starts by calculating the distance of point X from all the points. It then finds the 3 nearest points with least distance to point X. This is shown in the figure below. The three nearest points have been encircled.

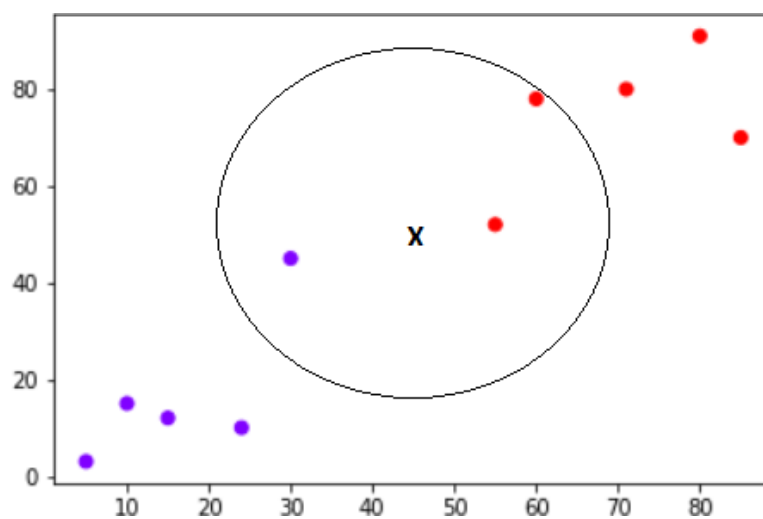


Fig 4.11

The final step of the KNN algorithm is to assign new point to the class to which majority of the three nearest points belong. From the Figure 4.4 we can see that the two of the three nearest points belong to the class "Red" while one belongs to the class "Blue". Therefore, the new data point will be classified as "Red".

4.1.1(b) Meta Model Stacked Regression

In this approach, we add a meta-model on averaged base models and use the out-of-folds predictions of these base models to train our meta-model.

The procedure, for the training part, may be described as follows:

1. Split the total training set into two disjoint sets (here train and holdout)
2. Train several base models on the first part (train)
3. Test these base models on the second part (holdout)
4. Use the predictions from 3) (called out-of-folds predictions) as the inputs, and the correct responses (target variable) as the outputs to train a higher-level learner called **meta-model**.

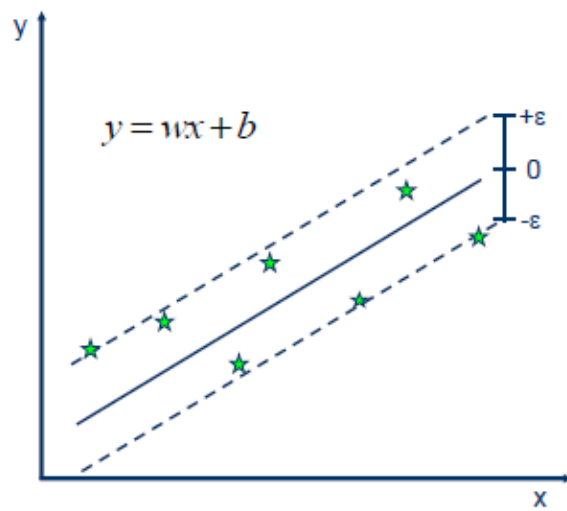
The first three steps are done iteratively. If we take for example a 5-fold stacking, we first split the training data into 5 folds. Then we will do 5 iterations. In each iteration, we train every base model on 4 folds and predict on the remaining fold (holdout fold).

So, we will be sure, after 5 iterations, that the entire data is used to get out-of-folds predictions that we will then use as new feature to train our meta-model in the step 4.

For the prediction part, we average the predictions of all base models on the test data and used them as **meta-features** on which, the final prediction is done with the meta-model.

1) Support Vector Machine

Support Vector Machine - Regression (**SVR**) Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm (maximal margin). The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences. First of all, because output is a real number it becomes very difficult to predict the information at hand, which has infinite possibilities. In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM which would have already requested from the problem. But besides this fact, there is also a more complicated reason, the algorithm is more complicated therefore to be taken in consideration. However, the main idea is always the same: to minimize error, individualizing the hyperplane which maximizes the margin, keeping in mind that part of the error is tolerated.



• Solution:

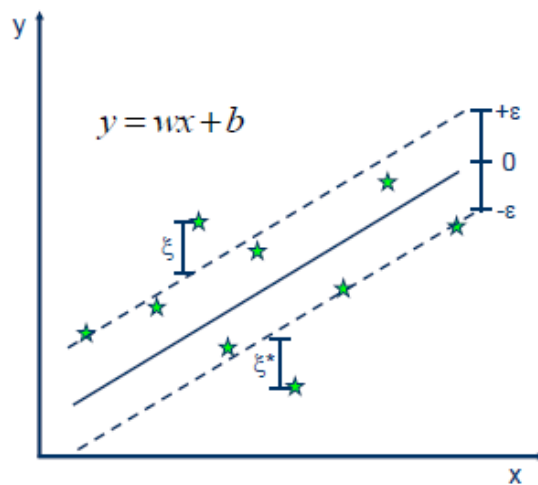
$$\min \frac{1}{2} \|w\|^2$$

• Constraints:

$$y_i - wx_i - b \leq \varepsilon$$

$$wx_i + b - y_i \leq \varepsilon$$

Fig: 4.12



• Minimize:

$$\frac{1}{2} \|w\|^2 + C \sum_{i=1}^N (\xi_i + \xi_i^*)$$

• Constraints:

$$y_i - wx_i - b \leq \varepsilon + \xi_i$$

$$wx_i + b - y_i \leq \varepsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \geq 0$$

Fig: 4.13

Linear SVR:

$$y = \sum_{i=1}^N (\alpha_i - \alpha_i^*) \cdot \langle x_i, x \rangle + b$$

Non- Linear SVR:

$$y = \sum_{i=1}^N (\alpha_i - \alpha_i^*) \cdot \langle \varphi(x_i), \varphi(x) \rangle + b$$

$$y = \sum_{i=1}^N (\alpha_i - \alpha_i^*) \cdot K(x_i, x) + b$$

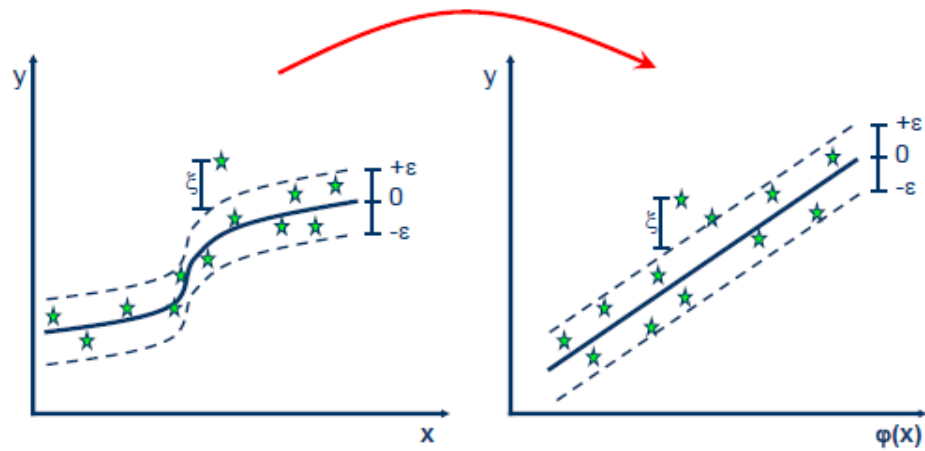


Fig: 4.14

Kernel Function:

$$y = \sum_{i=1}^N (\alpha_i - \alpha_i^*) \cdot \langle \varphi(x_i), \varphi(x) \rangle + b$$

$$y = \sum_{i=1}^N (\alpha_i - \alpha_i^*) \cdot K(x_i, x) + b$$

4.1.2 XGBoost

XGBoost, short for “Extreme Gradient Boosting”, was introduced by Chen in 2014. Since its introduction, XGBoost has become one of the most popular machine learning algorithms. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. GBM divides the optimization problem into two parts by first determining the direction of the step and then optimizing the step length. Different from GBM, XGBoost tries to determine the step directly by solving.

The two reasons to use XGBoost are also the two goals of the project:

1. Execution Speed - Its execution is really fast when compared to other implementations of gradient boosting.
2. Model Performance.

How it works:

In XG Boost, model is fit on the gradient of loss generated from the previous step.

In XG Boost, the gradient boosting algorithm is modified so that it works with any differentiable loss function.

4.1.3 Boosting Gradient

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision tree. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

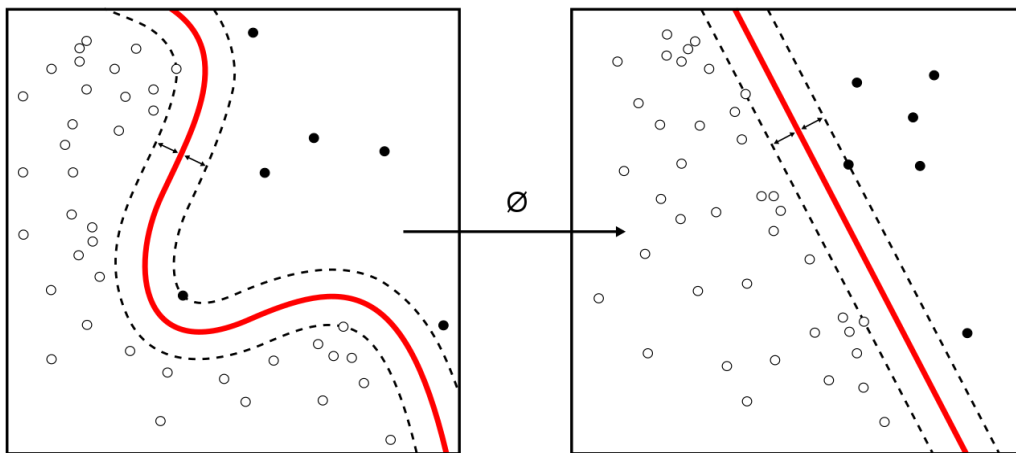


Fig: 4.15

Chapter 5

Testing, Results and Discussions

5.1 Testing

Testing is one of the ways of assessing the system which helps to detect the quality of the model we trained by giving the specific input and evaluating the expected output.

The project involved applications of several different algorithms such as Linear Regression, Random Forest, K-Nearest regression, Gradient boosting, XG Boost, SVR, Stacked Regression. For testing purposes, standard `sklearn.train_test_split` is used. The dataset was split into 25% testing data and 75% training data.

5.2 Results and Discussions

The results of the various applied algorithms are presented here with the help of various performance metrics such as RMSE, MSE, MSA and R-Squared in the form of graphs. Different outcomes have been compared as well.

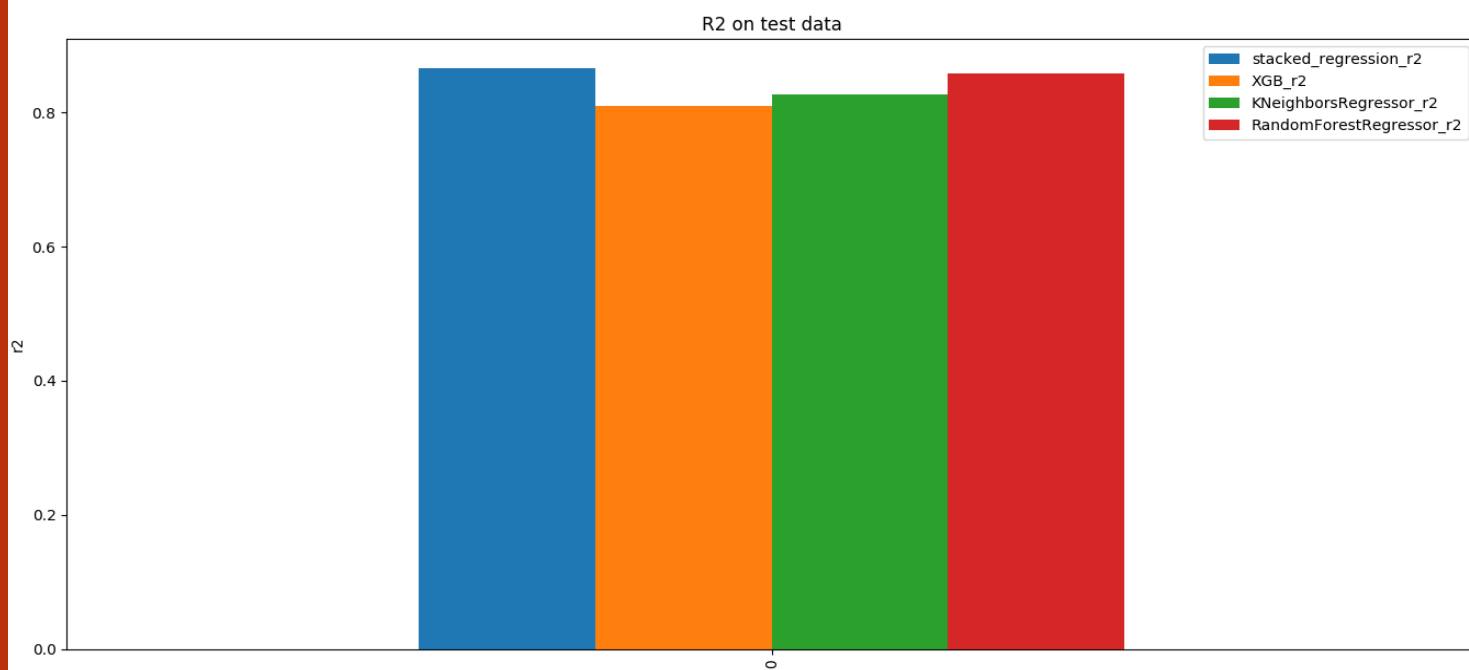


Fig: 5.1- R² on testing dataset

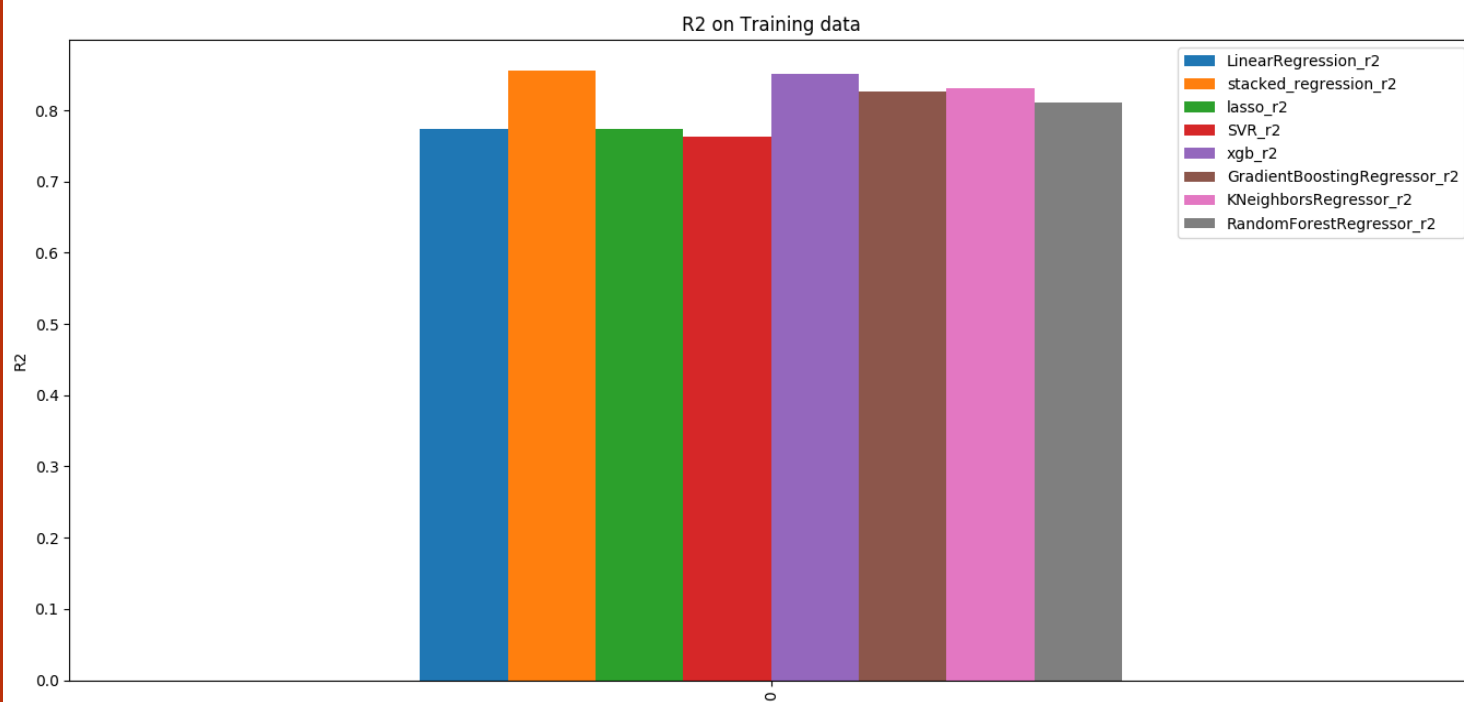


Fig: 5.2- R² on training dataset

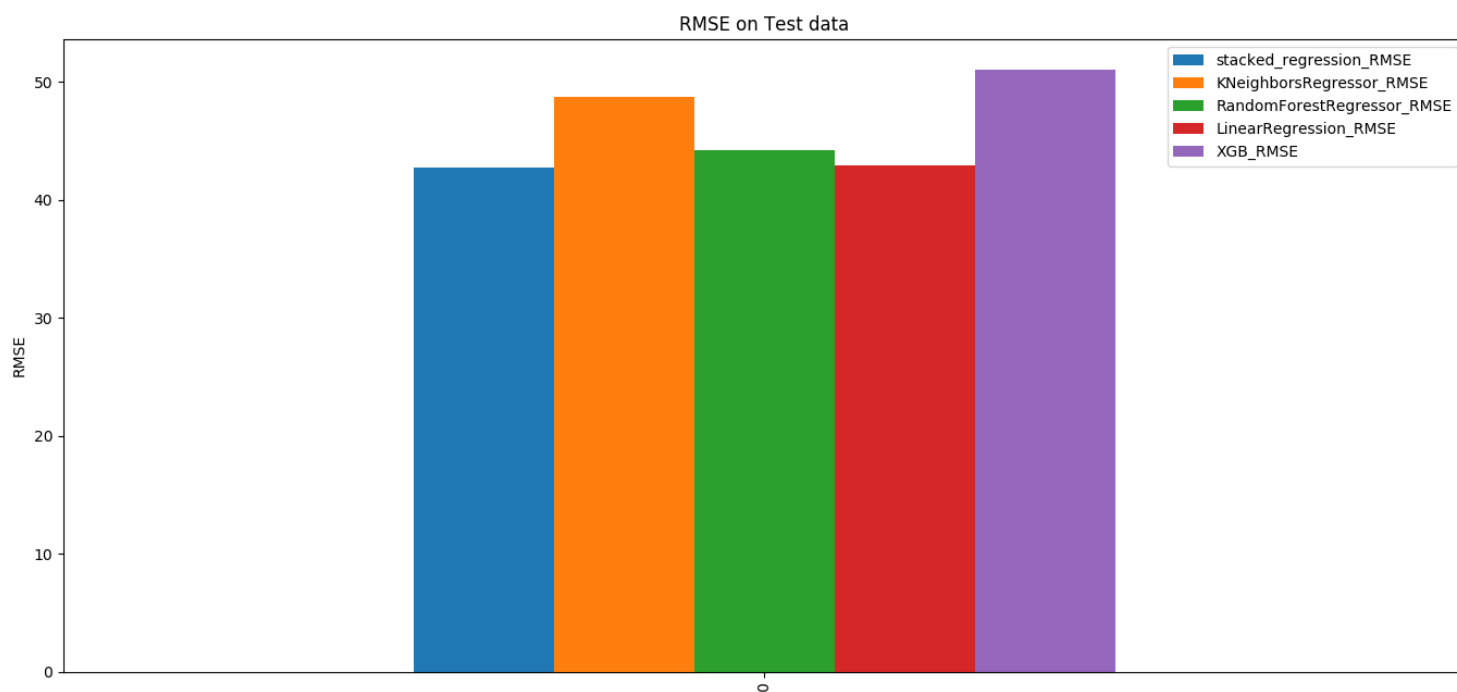


Fig: 5.3- RMSE on testing dataset

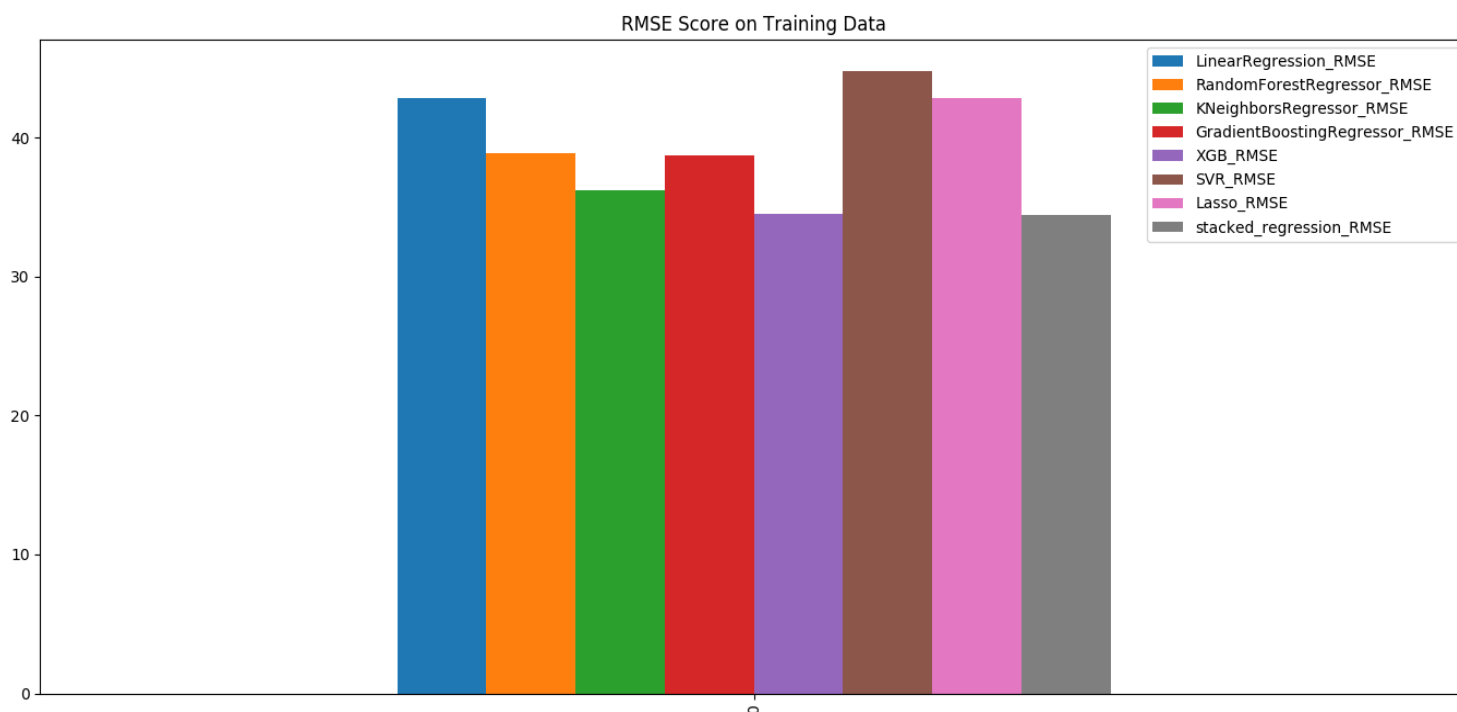


Fig: 5.4- RMSE on training dataset

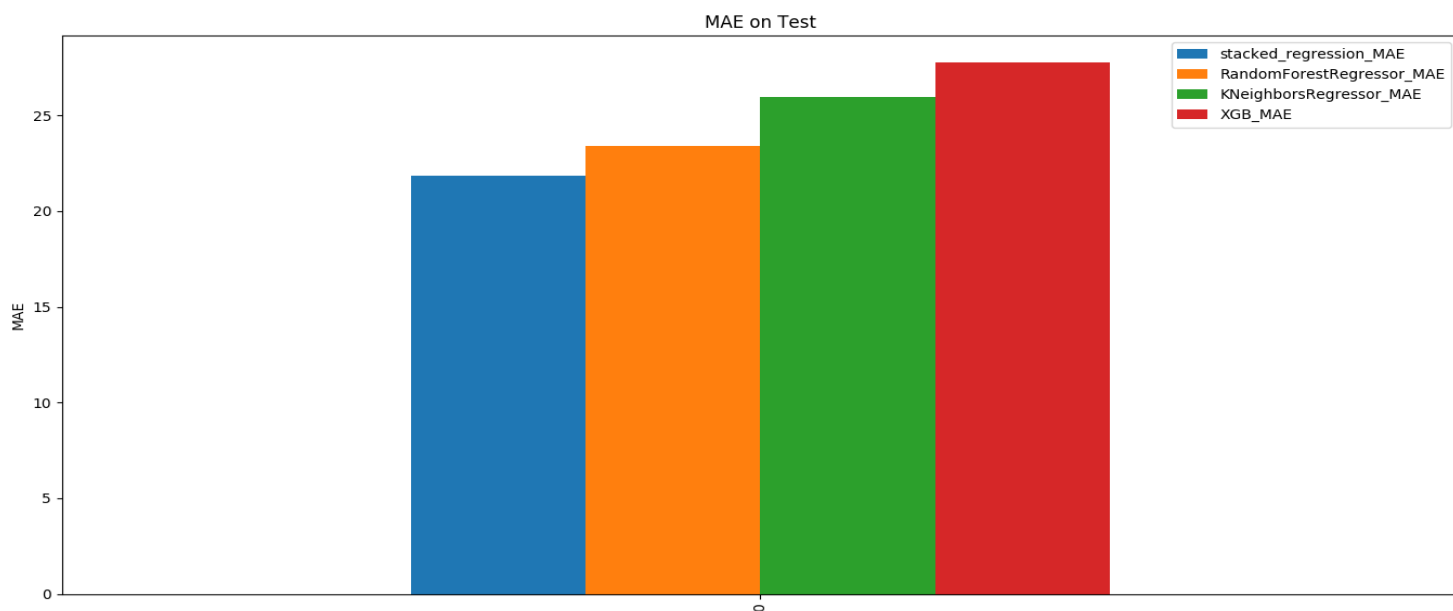


Fig: 5.5- MAE on testing dataset

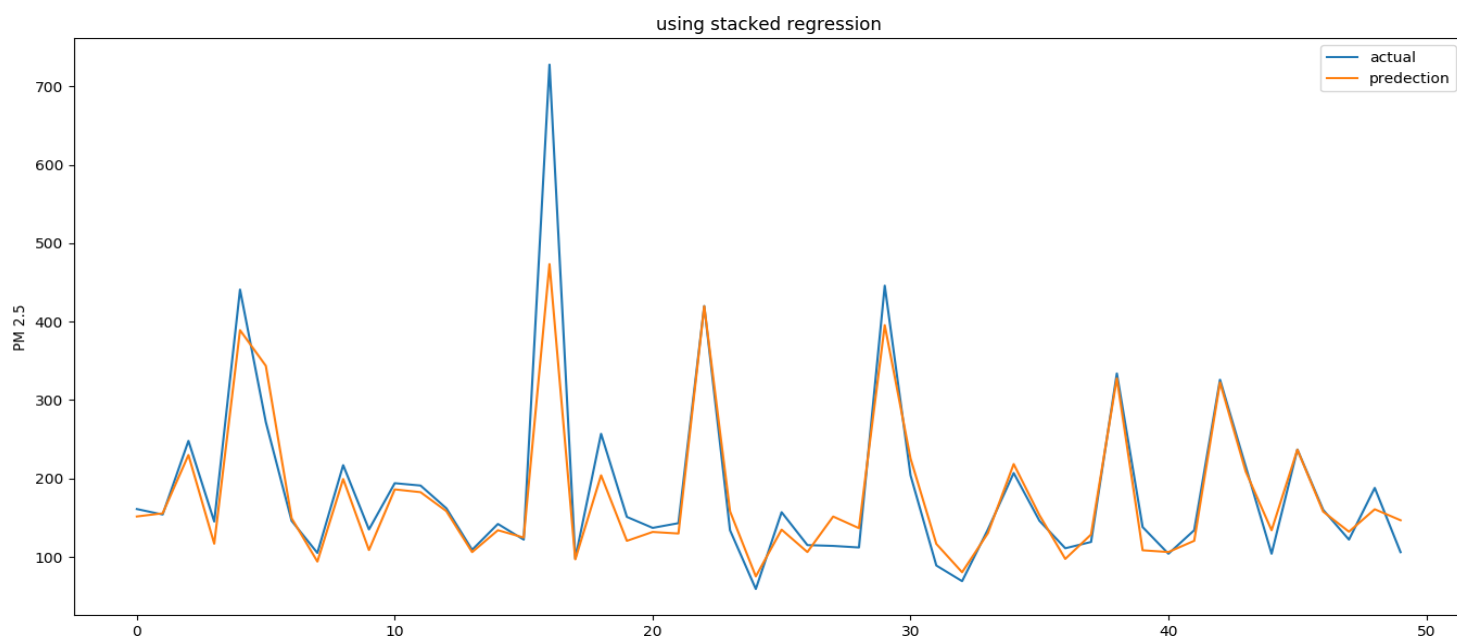


Fig: 5.6- Stack Regression. Actual v/s Predicted

	Actual	Predicted
0	161151.52615511856	
1	154155.53719235192	
2	248229.98026478679	
3	145 116.7484474288	
4	441389.22391644662	
5	272 343.7093359629	
6	146149.25885975033	
7	10593.947633654033	
8	217199.30719450363	
9	135108.68393756055	
10	194186.08132659269	
11	191182.52548311412	
12	162158.36392281735	
13	109106.04676157493	
14	142134.00417256863	
15	122124.69812056929	
16	728473.48595940753	
17	9896.821812196498	
18	257204.04468662505	
19	151120.43748157365	
20	137131.78383660722	
21	143129.79577950988	
22	420419.76176856794	
23	134158.04577673826	
24	5975.060222990605	
25	157134.76458644732	
26	115106.15482589135	
27	114151.46645546681	
28	112136.64103994824	
29	446395.60414243065	
30	204225.45590216852	
31	89116.53493803358	
32	6980.349769970513	
33	135130.23137633365	
34	207218.41083355067	
35	146 153.1375209873	
36	11197.402364730851	
37	119128.54481348835	
38	334 327.7638457105	
39	138 108.3827214448	
40	104106.18855258525	
41	134 120.5035877499	
42	326322.58145817399	
43	215 208.7201614521	
44	104133.91844428427	
45	237 236.687135047	
46	160157.78520174891	
47	122132.11454249662	
48	188160.73230198197	
49	106146.74903436161	

Fig: 5.7- Stack Regression CSV file

	Stacking				
	Actual	regression	XGB	KNN	RANDOM_FOREST
0	161151.52615511856	165.43413	149.5	175.27	
1	154155.53719235192	156.10117	152.5	157.76	
2	248229.98026478678	239.51228	241	238.82	
3	145 116.74844742881	118.765945	114.5	120.79	
4	441389.22391644662	401.66815	406	414.47	
5	272 343.7093359629	358.925	383	355.68	
6	146149.25885975033	142.64435	149.5	161.72	
7	10593.947633654033	79.59978	101.5	98.14	
8	217199.30719450363	202.76414	194.5	204.72	
9	135108.68393756055	111.18531	105	112.99	
10	194186.08132659269	179.29128	209	173.33	
11	191182.52548311411	172.95506	181.5	176.62	
12	162158.36392281735	159.68466	160.5	158.29	
13	109106.04676157493	97.26761	102	108.43	
14	142134.00417256863	117.74139	127.5	128.57	
15	122124.69812056929	132.37407	132	134.63	
16	728473.48595940753	430.0112	447.5	462.72	
17	9896.821812196498	94.799225	110.5	98.26	
18	257204.04468662505	186.59105	181	218.29	
19	151120.43748157365	118.21612	110.5	134.02	
20	137131.78383660722	145.54816	128.5	139.01	
21	143129.79577950988	135.45442	137.5	130.94	
22	420419.76176856794	425.633	412	440.45	
23	134158.04577673826	167.84363	116	153.93	
24	5975.060222990605	72.72869	87.5	78.06	
25	157134.76458644733	134.7493	130.5	135.44	
26	115106.15482589135	89.9222	108	106.47	
27	114151.46645546681	154.51637	154.5	142.2	
28	112136.64103994824	131.98512	138.5	135.99	
29	446395.60414243065	404.14462	400	406.37	
30	204225.45590216852	221.336	228.5	227.72	
31	89116.53493803358	158.13777	113	124.74	
32	6980.349769970513	80.213554	84.5	90.7	
33	135130.23137633365	126.177505	116.5	140.27	
34	207218.41083355067	222.30974	210	230.62	
35	146 153.1375209873	159.55292	146.5	146.64	
36	11197.402364730851	127.18172	101.5	102.24	
37	119128.54481348835	112.60485	137.5	124.18	
38	334 327.7638457105	325.85645	302.5	380.8	
39	138 108.3827214448	118.35148	106.5	111.58	
40	104106.18855258525	131.97078	117	113.96	
41	134 120.5035877499	96.34629	142.5	111.49	
42	326322.58145817399	318.74893	320	335.68	
43	215 208.7201614521	209.64314	226.5	203.55	
44	104133.91844428427	144.51144	144	133.95	
45	237 236.687135047	244.19418	208.5	225.92	
46	160157.78520174891	142.96667	161	153.96	
47	122132.11454249663	142.36604	137	127.04	
48	188160.73230198197	132.727	166.5	155.35	
49	106146.74903436161	151.4052	143.5	140.38	

Fig: 5.8-Final Results

5.3 Numerical Values

After the long analysis and examination of the different models, following observations were made by the use of different performance measure matrices.

R² (R-Squared) values:

1. Stacked Regression: 0.866
2. Random Forest Regressor: 0.8579
3. K-Neighbors Regressor: 0.8271
4. XG-Boost: 0.8103

RMSE values:

1. Stacked Regression: 42.7714
2. Random Forest Regressor: 44.1652
3. K-Neighbors Regressor: 48.7101
4. XG-Boost: 51.0294

MAE values:

1. Stacked Regression: 21.858
2. Random Forest Regressor: 23.422
3. K-Neighbors Regressor: 25.95
4. XG-Boost: 27.777

5.4 Interface



Predict Delhi PM 2.5 Analysis

panipat pm 2.5

Bhowani pm 2.5

rohtak pm 2.5

patiala pm 2.5

ludhiana pm 2.5

kathal pm 2.5

karnal pm 2.5

hisar pm25

jind pm25

Predict

Chapter 6

Conclusion and Future Work

In present times, if the deteriorating rates of quality in air continues, then there will be a havoc all around the world. Average age span will reduce and our future generations will be born in an environment full of particulates that might cause them diseases like asthma, cancer etc. from the early age itself. There needs to be a set regulation on fuel-based pollution, mining pollution and pollution due to agriculture. In our project we targeted the pollution due to agricultural burning called Stubble Burning. We created an awareness that due to the burning of agricultural waste there will be an air pollution not only in the cities nearby but also in different states as well. We specifically confined our study for the state like Delhi but in future works, different states can also be taken into consideration. Since there are many factors to find the quality of air such as concentration of NO_2 , SO_2 , SO , $\text{PM}_{2.5}$ etc. we undertook the examination of $\text{PM}_{2.5}$ and found its change in Delhi due to the neighboring states. Also, the data set we worked upon was quite limited. So, in future we can accumulate more data and find more accuracy in our model that we applied here in this project. Furthermore, beside working on $\text{PM}_{2.5}$ this project can be further implemented by examining other parameters like NO_2 , SO_2 , SO etc.

In nutshell, we can agree that the air quality of not only Delhi but throughout the globe is worsening day by day. We need to take action against the increase in air pollution by proposing different models or algorithms and working on it.

References

- 1) A Machine Learning Approach for air quality prediction: Model Regularization and Optimization [2018]. Author- Dixian Zhu, Changije Cai, Tianbao Yang, Xun Zhou.
- 2) IEEE Paper - Predicting trends in Air Pollution in Delhi using Data Mining [2016]. Author – Shweta Taneja, Nidhi Sharma, Kettun Oberoi, Yash Navoria.
- 3) IEEE Paper – Estimation of air pollution in Delhi Using Machine Learning Techniques [2018]. Author – Chavi Srivastava, Shyamli Singh, Amit Prakash Singh.
- 4) IEEE Paper – Research of Air pollution impact of Straw burning based on modis [2010]. Author – Qing Li, Qiao Wang, Zhongting Wang, Jinglei Ding, Xiang Zhao, Lijuan Zhang, Chunyan Zhau, Xin Yang.
- 5) Effects of agriculture crop residue burning on children and young on PFTs in North West India [2010]. Author – Amit Awasthi, Nirankar Singh, Susheel Mittal, Prabhat K. Gupta, Ravinder Agarwal.
- 6) https://en.wikipedia.org/wiki/Stubble_burning
- 7) Analysis of aerosol and carbon monoxide characteristics over Arabian Sea during crop residue burning period in the Indo-Gangetic Plains using multi-satellite remote sensing datasets [2009]. Author – K.V.S Badarinath, Sailesh Kumar Kharol, Anu Rani Sharma, V. Krishna Prasad.
- 8) Assessment of contribution of agricultural residue burning on air quality of Delhi using remote sensing and modelling tools. Author – Moorthy Nair, Hemant Bherwani, Suman Kumar, Sunil Gulia, Sanjeev Goyal, Rakesh Kumar.
- 9) Ambient air Quality change after stubble burning in rice-wheat system in an agricultural state of India. Author – Dipti Grover, Smita Chaudhary.
- 10) Ambient air Quality during wheat and rice crop stubble burning episodes in Patiala. Author – Susheel K Mittal, Nirankar Singh, Ravinder Agarwal, Amit Awasthi, Prabhat K. Gupta.
- 11) SAFAR information about Stubble burning printed in “The Print” journal in the link [here](#).
- 12) The Weighted Multiple Meta-Models Stacking Method for Regression Problem. Author -Dong Wang, Xishun Yue.
- 13) aqicn.org
- 14) <https://towardsdatascience.com/workflow-of-a-machine-learning-project-ec1dba419b94>