School of Engineering and Applied Science (SEAS), Ahmedabad University

B.Tech (ICT) Semester VI : Machine Learning (CSE 523)

- Group No : S_ECC6
- Group Members :
  1) Kalagee Anjaria (AU1741052) (B.Tech)
  2) Nidhi Golakiya (AU1741062) (B.Tech)
  3) Rajvee Kadchha (AU1741064) (B.Tech)
- Project Title: AQI prediction using ML techniques
- Project Area: Environment and climate change

1 Introduction

1.1 Background

- Air pollution is rising day by day which is a serious issue. In cities like Delhi , Ghaziabad, Noida it is a prime issue. AQI (Air Quality Index) is basically an index for measuring air pollution and various pollutants like PM2.5, NO_2 , O_3 , CO etc. PM2.5 is a very harmful pollutant, it is for all particles having radius less than 2.5. These particle do not settle down and stay longer in air which if inhaled is harmful.There are six levels for AQI. Low value of AQI means the air is safe to be inhaled. Pollution is increasing due to urbanisation, more people are living in urban areas which leads to more number of vehicles, more traffic ad more harmful gasses released into the environment.
- AQI prediction has been carried out using somewhat old approaches past the years.Such methods include manual collection and correction of data. They also have some disadvantages like :
  1) Less accuracy
  2) There may involve some difficult mathematical calculations
  3) Use disorganized methods for prediction

  But, as the technology has advanced considerably, new approaches using ML and big data algorithms have replaced the older ones. [1]

  For predicting AQI, support vector regression(SVR) and linear models like multiple linear regression which consists of gradient descent, stochastic gradient descent, mini-batch gradient descent can be used. From all these models, SVR shows better performance in terms of quality. [2]

  Comparing ANN-based models with multiple linear regression models exhibit that regression models

perform better for predicting CO and PM10 values, and mixed results for NO_2 and O_3. [3]
Ensemble Learning is a type of ML techniques that creates multiple predictors to address the same problem. It performs one single prediction by the combination of results. It's advantage is to provide a better accuracy, compared to the performance of each predictor taken individually. [4]
One more approach that one article focuses upon is predicting air quality using Long Short Term Memory(LSTM). The results show that LSTM can predict the AQI well. [5]
PM2.5 concentration is predicted using a hybrid model along with some other regression models like LR, SVR, RF, DT. The hybrid model is combined of decision tree and light gradient boosting model. [6]
PCA dimensionality reduction method is used to extract the principle components to reduce the effect of redundant features. [7]

- Multiple linear regression is a method in ML which includes one dependent variable and more than one independent variable. In general, it can be expressed as : $Y = b_1 + b_2X_2 + ... + b_kX_k + c$
Principle components can be computed by computing covariance of input data matrix. PCA can be applied with MLR to predict AQ index. We can get a rough idea on how this is done by the following figure : [8]

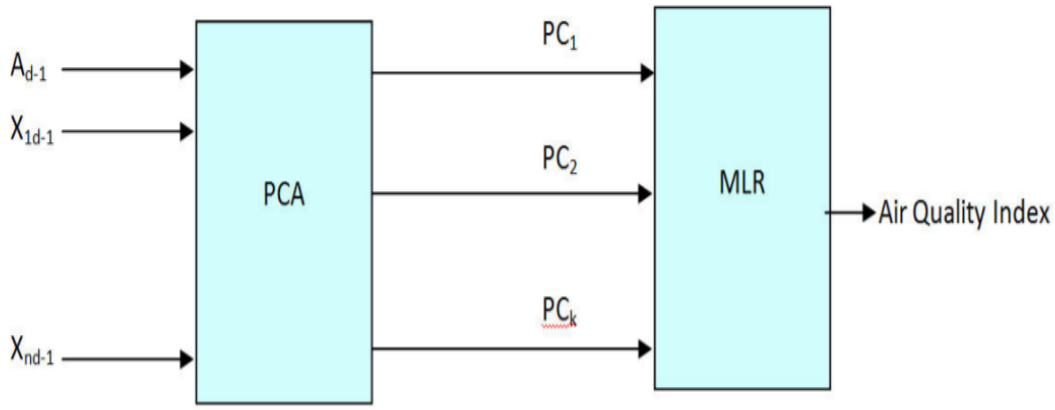


Figure 1: Architecture of PCR model

Using Naive Forecast approach, the dataset is split into two parts for training and testing. Then the outliers are removed. Linear regression is applied. After that, gradient boost algorithm is used to optimize the model. [9]
In our base article, four regression techniques have been used for predicting AQI, which include:- Decision Tree regression, Random Forest regression, Multi-Layer Perceptron regression and Gradient Boosting regression. All these four techniques are compared to determine which gives lesser error. And for comparing the errors between actual and predicted data, the evaluation parameters used are:- Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). [10]

1.2 Motivation

Predicting AQI has become almost a necessity in the highly polluted cities, so that people can take appropriate precautions regarding their heaths. For that, accurate measurements of AQI is needed. There are many existing techniques for predicting AQI (Air Quality Index). But using an appropriate technique which gives us accurate results in less time is a must.

Our base article, as mentioned in the previous section, predicts the AQI using 4 different regression techniques. Unlike most other literatures which performs the same task using only one or two techniques. Using just 1-2 techniques restricts us to accept whatever results are exhibited, and in however manner the prediction model fits the dataset. Also, other works have considered limited stations/cities data and so the dataset is limited for comparison. So, keeping all this in mind, the authors of our base article used four different very efficient techniques and then compared the results for different cities to see which model fits best for which city. Hence by comparing, they could conclude which algorithm gave the most accurate prediction results with lesser error rates and much lower processing time.

1.3 Problem Statement

- In our project, we've predicted the AQI for R.K Puram station of Delhi. We have acquired the dataset from an official government website for datasets. Then, we have cleaned the dataset, because it is not necessary that all the values are availabe for all the parameters given. For calculating the AQI, first we have to calculate the AQI for all the parameters taken, and the maximum AQI out of all these paramters is considered to be the AQI for that city. Then for predicting the future AQI, we have used the technique of linear regression. Also, we have applied the techniques used in the base article on our dataset.

  The performance metrics for our project is RMSE, MSE and accuracy. By calculating RMSE and MSE, we can identify the performance for the regression model. The lesser the RMSE and MSE, the better is the performance. For the classification, accuracy is used as a performance metric. The more the accuracy the better has the classification model performed. We have used confusion matrices to measure the accuracy.

2 Data Acquisition / Explanation of Data set

- We have acquired our dataset from the site of Central Control Room for Air Quality Management - All India. [11] In this site, we need to specify the state name, city name and station name for which we want the data. Then we need to select the parameters which will become the independent variables of our dataset. Then the format(tabular/graph), criteria(15 min/30 min/hourly/4 hours/8 hours/24 hours/annual average), from date and to date(with time). As we submit this information, we can view the dataset according to the information/parameters selected. There are 3 options as to in which format you want to download : pdf, excel and word.

Our dataset is for the R.K Puram station of Delhi. The parameters of our dataset are :
$CO_2, SO_2, O_3, PM10, PM2.5$ and NO_2.
The data is for every 24 hours of the time 18:00 starting from 4th March 2105 to 8th April 2020(after cleaning the null values). We selected Delhi because as we know, Delhi is among the most polluted cities of India so predicting AQI for such a city is more beneficial. Then this particular area and all these criterias were selected because more data was available for it with less null values.

3 Machine Learning Concept Used

- Linear Regression

  AQI which is dependent on all of those 6 dependent variables. For predicting the AQI, we have applied linear regression because in our dataset the independent variables are linearly related to the dependent variable. Another reason for applying regression to our dataset is that our output is continuous and not discrete. Regression analysis is important as it describes the relationship between the dependent and independent variables. Through regression, we can determine which independent variable has more impact on the AQI predicted.We have removed all the null values. The final AQI which is considered the maximum value obtained from all the variables' values. So, we obtained the maximum value from each row of our data and added one column for their corresponding AQI. We also added a column for dummy variable with all 1's for performing MLE on the data. Thereafter, we split the data into train data(80%) and test data(remaining 20%). Then we implemented MLE using the following equation: ML = $(X^T \cdot X)^{-1} \cdot X^T \cdot Y$ By using the formula Y = $X^T \cdot \theta_{ML}$ we have predicted the AQI value for the training data as well as testing data. Then, we have also obtained the RMSE for both the training and testing data.

- PCA(Principal Component Analysis)

  To form X,matrix of all six independent variables is constructed. The dimension of X is (1506 6). Then, as all of our parameters are not in the same range, we will normalise it first, and then all the operations are performed on this normalised X. Then, Covariance matrix is formed byX^T·X.The dimension of covariance matrix will be (p p) i.e. (6 6) in our case Eigen vectors and Eigen values are found by inbuilt function present in numpy library. And we will get eigen value and eigen vectors related to our number of parameters. In our case we will get 6 values. Projection matrix is formed by B^T·B, where B is matrix formed by eigen vectors of all components upto m, where m is dimension upto which we want to reduce it. Shape of projection matrix is (6 6).

- SVM(Support Vector Machine) Our problem is prediction problem. But we can also classify it. As mentioned AQI can be mesured in levels. tags are Healthy, Hazardous, good,moderate,etc. It is classified into six different levels based on Aqi values. Then we are predicting class based on concentration of SO_2, NO_2, etc SVM is the technique for classification which best seperates the features into different domains. We have used python inbuilt classifier for SVM. We have applied all the kernels

for SVM, which are linear, polynomial and rbf. In polynomial and rbf kernels, we have to provide values of C and gamma. For polynomial, we have taken C=300 and gamma=0.0001 and for rbf C=50 and gamma=0.0001 .C is used to decide the boundry and gamma is used to adjust the curvature of boundry.C and gamma values are decided by using SVC fit.

4 Analysis/ Algorithm

- ## Linear Regression

We have 6 parameters in our dataset.There are 1231 examples in our training set and 309 examples in testing set.

Here, our data point are IID.Conditional probability distribution is ,

P(Y/X,θ) = P($y_1, y_2, ..., y_N/x_1, x_2, ..., x_N, \theta$)

P(Y/X,θ) = $\prod_{n=1}^{N} P(y_n/x_n, \theta)$

By taking negative log likelihood ,

$$L(\theta) = -logP(Y/X, \theta)$$

$$\therefore L(\theta) = -log \prod_{n=1}^{N} P(y_n/x_n\theta)$$

$$\therefore L(\theta) = -\sum_{n=1}^{N} logP(y_n/x_n\theta)$$

Let Y = $X^T \cdot \theta + \epsilon$, $\epsilon \sim N(0,\sigma^2)$

$$L(\theta) = -\sum_{n=1}^{N} logP(y_n/x_n^T\theta, \sigma^2)$$

$$\therefore L(\theta) = -\sum_{n=1}^{N} log\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\frac{-(y_n - x_n^T\theta)^2}{2\sigma^2}$$

$$\therefore L(\theta) = \sum_{n=1}^{N} log\frac{1}{\sqrt{2\pi\sigma^2}} + \frac{1}{2\sigma^2}(y_n - x_n^T\theta)^2$$

$$\therefore L(\theta) = \frac{1}{2\sigma^2}(y_n - x_n^T\theta)^2 + Constant$$

$$\therefore L(\theta) = \frac{1}{2\sigma^2}(Y - X\theta)^T \cdot (Y - X\theta)$$

$$\therefore L(\theta) = \frac{1}{2\sigma^2}(||Y - X\theta||)^2$$

$$\longrightarrow \frac{dL(\theta)}{d\theta} = \frac{1}{2\sigma^2}\frac{d}{d\theta}[(Y^T - \theta^T X^T)(Y - X\theta)]$$

$$\therefore 0^T = \frac{1}{2\sigma^2}[2Y^T X + 2\theta^T X^T X]$$

$$Y^T X = \theta^T X^T X$$

$$\therefore \theta_{ML}^T = (Y^T X)(X^T X)^{-1}$$

$$\therefore \boxed{\theta_{ML} = (X^T X)^{-1}(Y^T X)}$$

- PCA(Principal Component Analysis)

The algorithm for PCA is as follows :

Step-1 : Input is High dimensional(6-D) dataset X (1540 x 6).Normalize X.

Step-2 : Find Covariance matrix S = $X^T \cdot X$(6 x 6)

Step-3 : Find Eigen values and Eigen Vector of S. Where Eigen values are $\lambda_1 > \lambda_2 > ... > \lambda_6$. And corresponding eigen vectors are $v_1, v_2, .., v_6$.

Step-4 : Form New Basis Vector B = $[v_1, v_2, ...v_M]$ where M is the dimension we want to reduce.

Step-5 : Form Projection matrix P = $B \cdot B^T$

Step-6 : The original X can be reconstructed as , $X_{Reconstruct} = X \cdot P^T$

Step-7 : Compute the RMSE on X and $X_{Reconstruct}$

- SVM(Support Vector Machine)

We have 6 features(6-D) and classifies into 6 different classes.

So equation of hyerplane is,

$$Y = w_0 + w_1 \cdot x_1 + ... + w_{66}$$

$$\therefore Y = w_0 + \sum_{i=1}^{6} w_i \cdot x_i$$

$$\therefore Y = W^T \cdot X$$

Y = b + $W^T \cdot X$, where b = biased item , X = input feature vector,W = weights

if $Y_i(W^T \cdot X) \geq 1$ which means X_i is correctly classified. Let ζ_i be slack variables.

$$\therefore Y_i(W^T \cdot X) \geq 1 - \zeta_i$$

if $\zeta > 0$ then points are not correctly classified and $\zeta_i = 0$ then points are correctly classified.

Now, $\underset{w,b}{minimize}\ \frac{1}{2}||w||_2 + \sum_{i=1}^{n} \zeta_i$

$$L(w, b, \zeta, \alpha, \beta) = \frac{1}{2} w \cdot w + C \sum_{i=1}^{n} \zeta_i + \sum_{i=1}^{n} \alpha_i [y_i(x \cdot w + b) - 1 + \zeta_i] - \sum_{i=1}^{n} \beta_i \zeta_i$$

To find the vector w and the scalar b such that the hyperplane represented by w and b maximizes the margin distance minimizes the loss term subjected to the condition that all points are correctly

classified.

If the data points are not linearly separable ,

$\longrightarrow \underset{\alpha}{maximize} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j X_i^T X_j$ subject to $\alpha_i > 0$ for all i=1,2,...,n and $\sum_{i=1}^{n} \alpha_i y_i = 0$

Kernel is a way of computing the dot product of two vector X and Y in high dimesional feature space.

$\longrightarrow \underset{\alpha}{maximize} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathrm{K}(X_i^T X_j)$ subject to $\alpha_i > 0$ for all i=1,2,...,n and $\sum_{i=1}^{n} \alpha_i y_i = 0$

To apply kernel means to replace dot product of two vector by kernel function.

Linear kernel : $\mathrm{K}(X_1, X_2) = X_1^T \cdot X_2$

Polynomial kernel : $\mathrm{K}(X_1, X_2) = (a \cdot X_1^T \cdot X_2)^p$

Gaussian kernel : $\mathrm{K}(X_1, X_2) = \exp^{-\gamma ||X_1 - X_2||^2}$

5 Coding and Simulation

5.1 Simulation Framework

Num Component is a controlling parameter in PCA. Num component is a number on which we want to reduce our dimension. For example if Num component is 3 then in our case we are reducing our 6-D data to 3-D data. As we go on increasing Num component that means we are covering more and more spread of data.

Gamma and C are also controlling parameters in SVM. While using inbuilt SVC in python we are supposed to provide value of Gamma and C with name of the kernel. Polynomial and RBF kernel requires these parameters. Value of C determines how many data samples are placed in different classes. If C is low then we are covering outliers and general decision boundary is found. If C is high then more careful boundary is found.While Gamma parameter adjust the curvature of the decision boundary. It means it determines how far the influence of the data point reaches. If Gamma is low then it reaches far and if high then reaches close.

5.2 Results

- Used Tool-Jupyter Notebook-Python

5.2.1 Results from Base Article Reproduction

- Prediction using the models : Decision Tree , Random Forest , Gradient Boosting Regression on our dataset

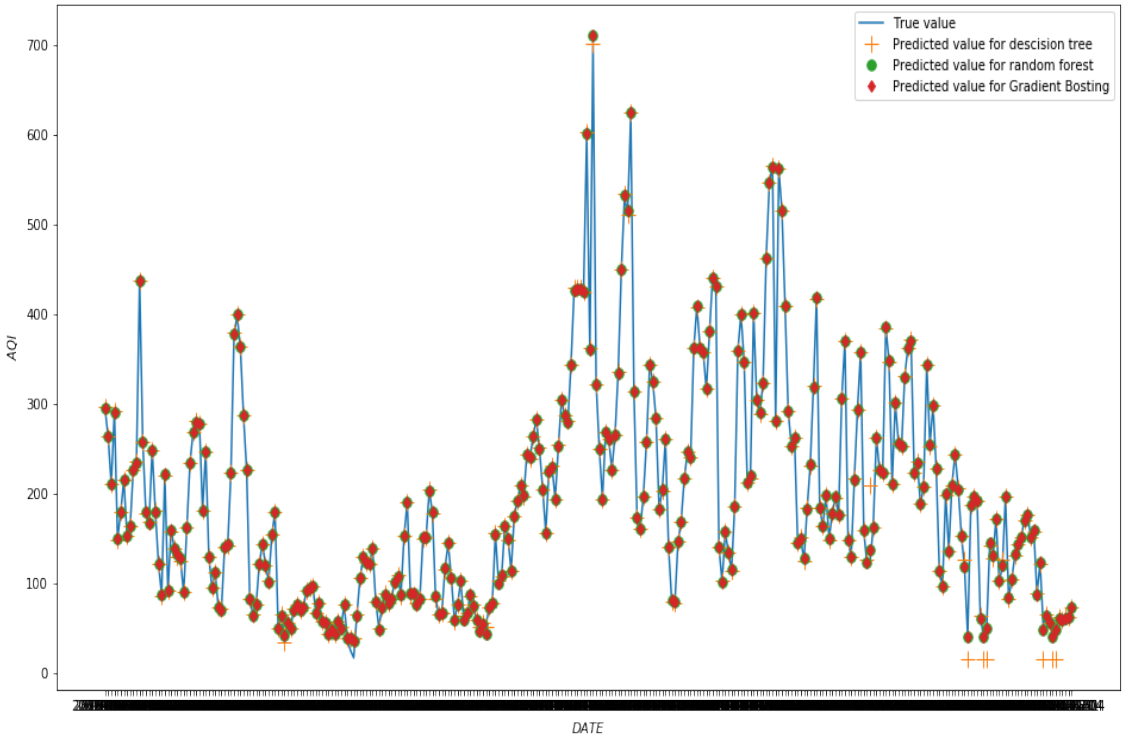


Figure 2: Prediction of 1)Decision Tree 2)Random Forest 3)Gradient Boosting

As we can see from the figure-2 that all the three regression models performs well for prediction. And that all the predicted data points are close to actual data points.So, from the figure it is difficult to say, which model performs better.

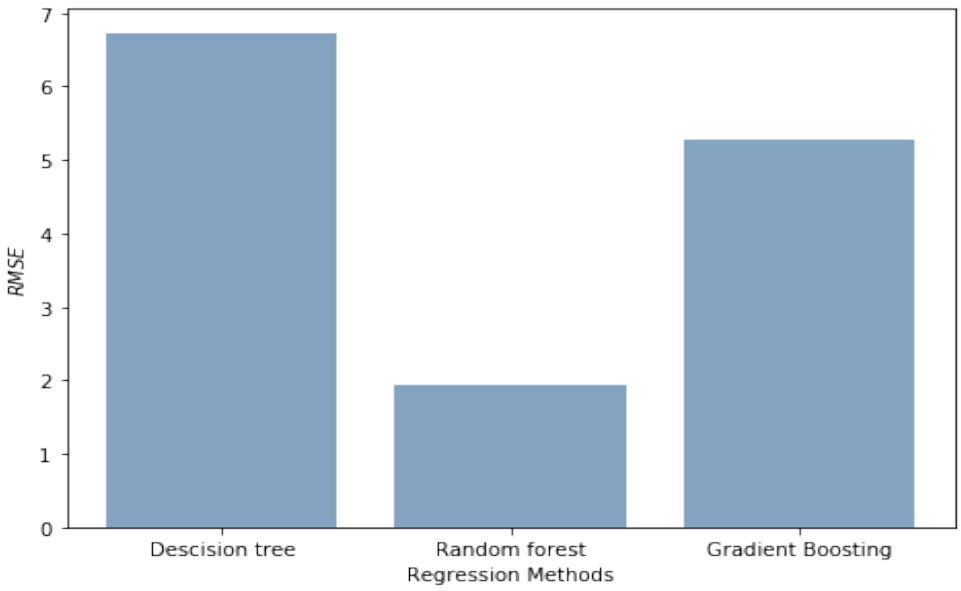


Figure 3: Rmse of 1)Decision Tree 2)Random Forest 3)Gradient Boosting

From the bar graphs in Figure-3 it can be said for our prediction model Random forest regression performs better. As rmse of random forest regression is the lowest. Then Gradient boosting regression is with the 2nd lowest rmse and Decision Tree performs the worst of all three.

5.2.2 Linear Regression

- Prediction of LR model

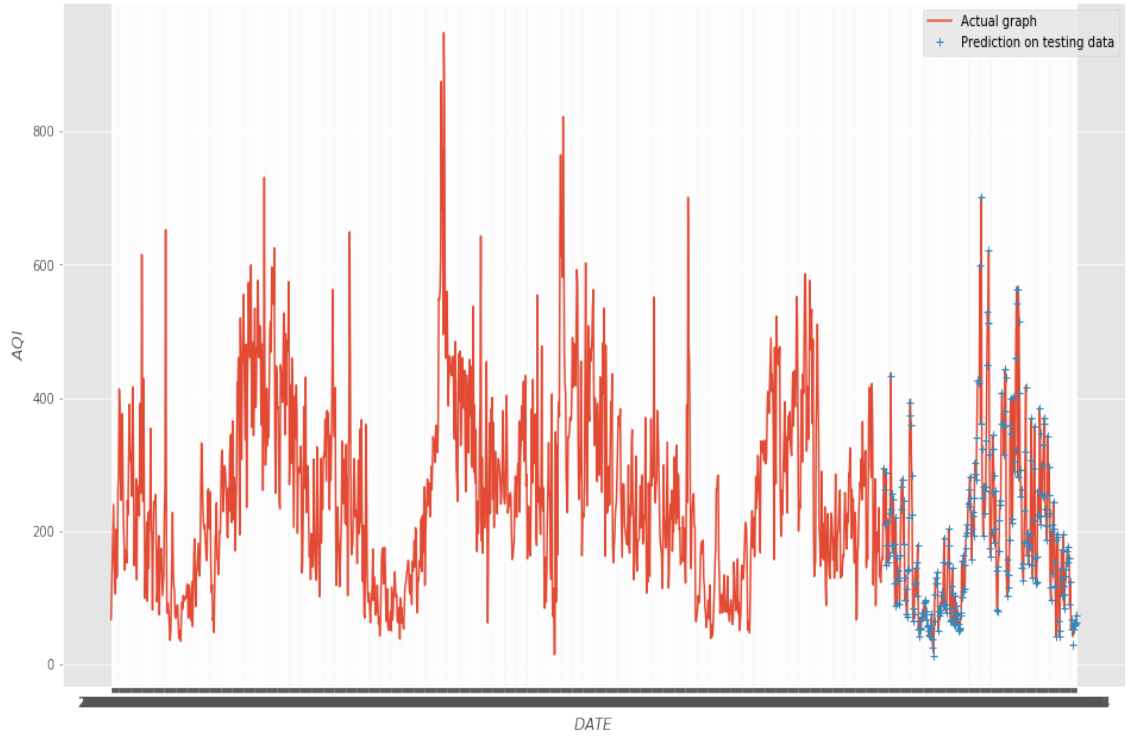


Figure 4: Output of Linear Regression Model

We have used Linear Regression as our predicted value Y(i.e. AQI) depends linearly on independent parameters X(i.e. concentration of different gases). So, LR gives very good performance. Blue + as markers for predicted value of AQI for test dataset .As we can see, the predicted values fit almost nearly with the true values of AQI corresponding to the dates. So, the RMSE for testing data is 1.3434. The outliers on the training set is more, that's why the RMSE for the training set it is 5.416.

5.2.3 PCA

- MSE for different dimensions

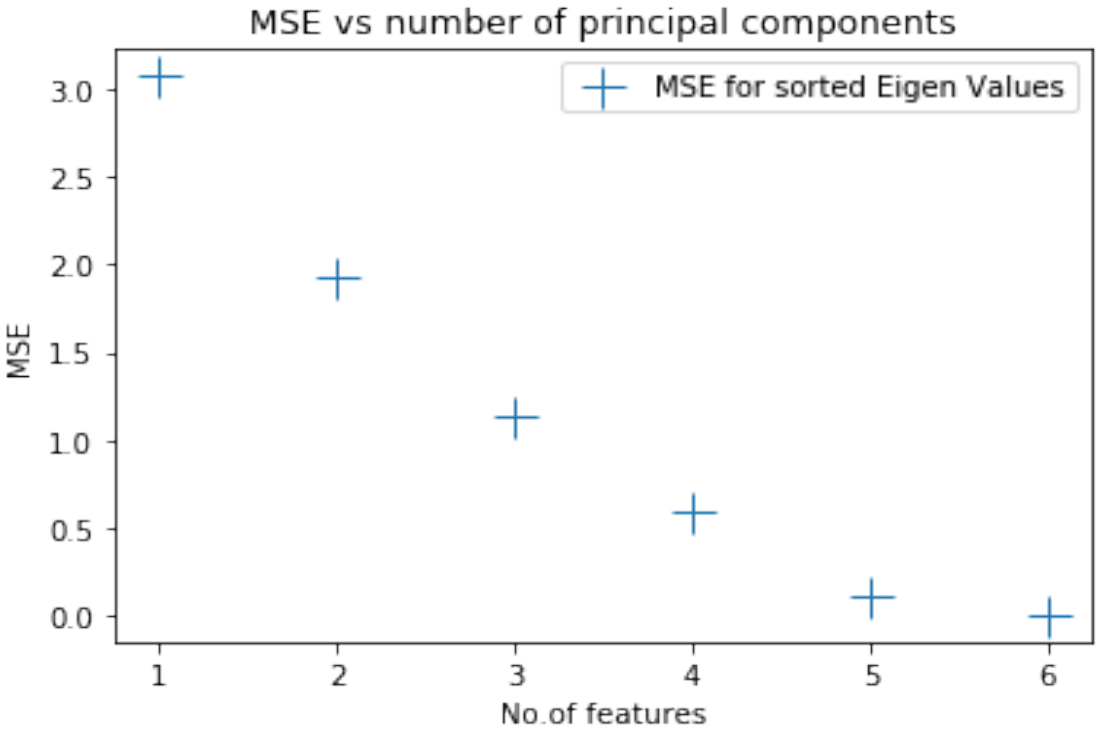


Figure 5: MSE for taking different no. of features into consideration

From figure, we can see that as number of feature increases, mse is decreasing that is because prediction is better when more number of features is provided.That is because as number of feature increases we are considering more spread.Also information loss decreases.

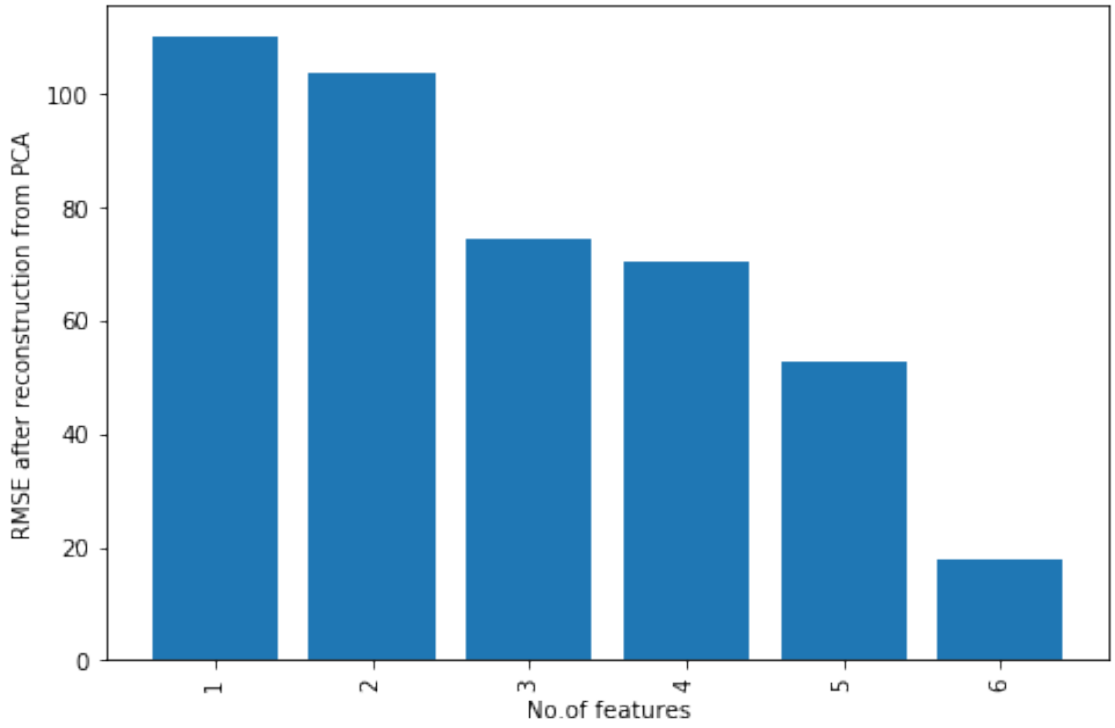
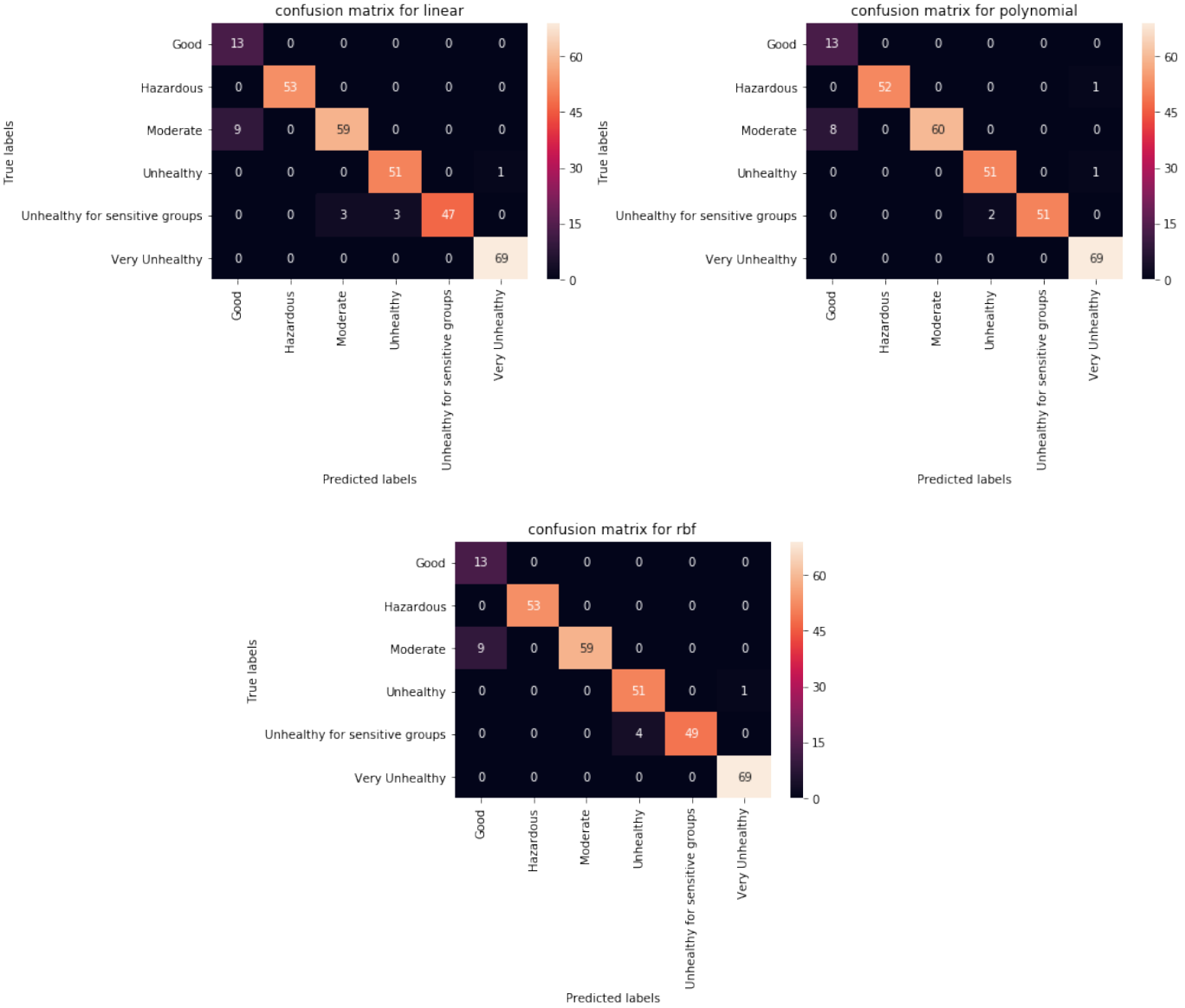


Figure 6: RMSE after doing linear regression on reconstructed X from PCA

Here for 1 feature only we have got very high value of rmse.That means information loss is higher. It is because as already mentioned, reducing features for already less features and records does not help for better performance. So, as number of features increases RMSE decreases. And also one point to note that this error is very much as compared to the linear regression on actual X.Because as it is calculated by taking into consideration reconstructed matrix. We have error in values of X, and so we get more error in prediction.So, rmse is so high.

5.2.4 SVM



We have classified AQI into different levels as shown in figure.Confusion matrix is performance evaluation for classification problem, where X axis contains predicted tags and Y axis contains true labels.So, in confusion matrix, diagonal entries should have higher values.We can see that almost all three kernels perform nearly same. It is difficult to say from the confusion matrix that which kernel performs better .Few miss classifications can be seen,which are off-diagonal entries.So by looking at Confusion matrix alone we cannot derive any conclusion. So for that we have calculated accuracy. For linear kernel, we are getting 0.94805 accuracy.For polynomial with C=300 and gamma=0.0001 , accuracy is 0.96103.And for rbf(Radial Basis Function or Gaussian) with C=50 and gamma=0.0001 , accuracy is 0.95454. So, from accuracy we can say that polynomial with C=300 and gamma=0.0001 performs better of all.

6 Conclusion

- In the results section, we have compared the techniques used in our base article. From that, we inferred that Random forest performs the best. Also, we have implemented linear regression as well on the same dataset. The RMSE of LR technique is 1.3424 and the RMSE of Random Forest technique is 1.9462. Thus, from this, we can conclude that the linear regression model better fits this dataset than the Random forest regression. Linear regression fits better on our dataset as the dependent variable(AQI) is linearly related with the independent variables(concentration of gases). And LR is the best model for data that are linearly related.

- We can apply both regression and classification on our problem statement. When we need to predict the measure or the numerical value of AQI, we apply the appropriate regression technique to the dataset. Through regression, we predict the AQI values(numerical). And when we want to know whether the air quality is good/hazardous/moderate/unhealthy/unhealthy for sensitive groups/very unhealthy i.e the effect of AQI, then classification is applied.

7 Contribution of team members

7.1 Technical contribution of all team members

Tasks	Kalagee Anjaria	Nidhi Golakiya	Rajvee Kadchha
Reproduction of base article	✓		✓
Mathematical analysis and algorithm		✓	✓
LR, PCA, SVM	✓	✓	✓

7.2 Non-Technical contribution of all team members

Tasks	Kalagee Anjaria	Nidhi Golakiya	Rajvee Kadchha
Finding dataset	✓	✓	✓
Literature survey	✓	✓	
Report writing	✓	✓	✓

References

[1] G. Kaur, J. Gao, S. Chiao, S. Lu, and G. Xie, "Air quality prediction: Big data and machine learning approaches," *International Journal of Environmental Science and Development*, vol. 9, pp. 8–16, 01 2018.

[2] S. S. Ganesh, S. H. Modali, S. R. Palreddy, and P. Arulmozhivarman, "Forecasting air quality index using regression models: A case study on delhi and houston," in *2017 International Conference on Trends in Electronics and Informatics (ICEI)*, 2017, pp. 248–254.

[3] Y. Zhou, S. De, G. Ewa, C. Perera, and K. Moessner, "Data-driven air quality characterisation for urban environments: a case study," *IEEE Access*, vol. 6, pp. 77 996 – 78 006, 12 2018.

[4] Y. Rybarczyk and R. Zalakeviciute, "Machine learning approaches for outdoor air quality modelling: A systematic review," *Applied Sciences*, vol. 8, p. 2570, 12 2018.

[5] Y. Jiao, Z. Wang, and Y. Zhang, "Prediction of air quality index based on lstm," in *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, 2019, pp. 17–20.

[6] A. S. K. J. Angelin Jebamalar, "Pm2.5 prediction using machine learning hybrid model for smart health," *International Journal of Engineering and Advanced Technology*, vol. 9, pp. 6500–6504, 10 2019.

[7] Y. Zhang, Y. Wang, M. Gao, Q. Ma, J. Zhao, R. Zhang, Q. Wang, and L. Huang, "A predictive data feature exploration-based air quality prediction approach," *IEEE Access*, vol. PP, pp. 1–1, 02 2019.

[8] A. Kumar and P. Goyal, "Forecasting of air quality in delhi using principal component regression technique," *Atmospheric Pollution Research*, vol. 2, pp. 436–444, 10 2011.

[9] A. A. Mrs. A. Gnana Soundari, Mrs. J. Gnana Jeslin M.E, "Indian air quality prediction and analysis using machine learning," *International Journal of Applied Engineering Research*, vol. 14, pp. 181–186, 11 2019.

[10] S. Ameer, M. Shah, A. Khan, H. Song, C. Maple, S. Islam, and M. Asghar, "Comparative analysis of machine learning techniques for predicting air quality in smart cities," *IEEE Access*, vol. PP, pp. 1–1, 06 2019.

[11] "Central control room for air quality management - all india." [Online]. Available: https://app.cpcbccr.com/ccr//caaqm-dashboard-all/caaqm-landing/data