Chapter 17

Analyzing AQI before Covid '19:

Experimental Study of 3 Years for Intelligent Environment Conducted at North Indian Zone to Extract Knowledge

Rohit Rastogi
https://orcid.org/0000-0002-6402-7638
Department of CSE, ABES Engineering College, Ghaziabad, India

Sheelu Sagar
https://orcid.org/0000-0001-6393-9793
Amity University, Noida, India

Neeti Tandon
Vikram University, Ujjain, India

ABSTRACT

In the populated and developing countries, governments consider the regulation and protection of environment as a major task and should take into consideration the concept of smart environment monitoring. The main motive of these systems is to enhance the environment with various technology including sensors, processors, data sets, and other devices connected across the globe through a network. This system can help in monitoring air quality. Also, these factors contribute a lot to air pollution. So, forecasting air quality index using an intelligent environment system includes a_machine learning model to predict air quality index for NCR (National Capital Region). The values of major pollutants like SO_2, PM2.5, CO, PM10, NO_2, and O_3. The authors have implemented different machine learning algorithms of classification and regression techniques. To make their prediction more accurate, mean square error, mean absolute error, and R square errors have been considered. The chapter helps to frame a structured view of air quality prediction methods in the reader's mind and also gives suggestions for other prediction methods as well. The real challenge is to decide which method will be applied in predicting air quality. Hence, it is important to test and use all these methods.

DOI: 10.4018/979-8-3693-2109-6.ch017

INTRODUCTION

This research paper discusses the different parameters of air quality and environment using various machine learning algorithms. This paper also introduces us with the reason behind the contamination of air. Pollution can cause harm to not only air but to water bodies as well. It can take the form of noise, heat and even light. The substances which are responsible for pollution can be either foreign substance or they may have been naturally occurred in the nature. Pollution is mostly classified as to be caused by any one single source rather than many.

Intelligent Environment Systems

In this era of globalization, every country in the world is facing problems related to environment. In order to control these problems, it has become a primary concern of thinking for the various organizations and governments. This emerging problem produces a need of monitoring of environment and finding more environment nourishing solutions. And, this need brings smart monitoring techniques into the picture (Dipanda, A. et al., 2016).

Intelligent Environment Systems plays an important role in approximately all sectors. This field is becoming a must have for cities with increased industrialization, high population and massive transportation, as these sectors are the main reason behind increasing pollution (Wilson, T., et al., 2018).

Types of Pollution

Pollution is mainly of four types: water pollution, soil pollution, noise pollution and air pollution. We here elaborate detail about air pollution. A factor which induces air pollution is stubble farming, motor vehicle emission, topological factor, and open construction work. The ordinance of environmental contamination has drawn public examination. NCR(National Capital Region) one of the most contaminated territories in the world (Tripathi, C.B. et al., 2019).

Components in Pollution Particles

Different researches have carried out several experiments and have come to a conclusion that the concentration of pollutants in NCR is alarmingly higher as compared to any other region (Srivastava, C. et al., 2018).This has made the lives of all the residents less for up to 6 years. While some researchers have (Aggarwal, P. et al., 2015) concluded that pollution has affected human fitness. Hence, we enhancing the air quality forecasting is one of the best objectives for civilization. Sulphur dioxide, PM2.5 and NO are major pollutants found in the air. Sulphur Dioxide is a gas, present in air(Gallardo, M. et al., 2017).

This combines easily with different substances to form harmful substances like Sulphur acid, sulfurous acid etc. Sulfur dioxide affects social fitness when it is inhaled in. It causes a burning feeling in the nose, throat, and airways to result into coughing, wheezing, the brevity of breath, or a tense feeling around the chest. The concentration of Sulphur dioxide in the environment affects the places we can live in (Bhalgat, P. et al., 2019).PM2.5 is also known as fine particulate matter (2.5 micrometers is one 400th of a millimeter). Fine particulate matter (PM2.5) is important among the pollutant index because it is a big concern to people's health when its level in the air becomes high (Pandey, G. et al., 2013). So, it has been categorized according to Air quality index table.

Research Problem Introduction and Motivation

- With an average of 98.6 Particulate Matter (PM 2.5) concentrations, Delhi was the most polluted city in the world. 21 cities out of 30, which were the most polluted, were from India.
- India leads the charts for the most polluted cities in the world. All those cities in India, who were to be monitored as per WHO, didn't report for the annual pollution exposure 2019.

RELATED PREVIOUS WORK

Numerous other models exist to check the concentration of pollutants in cities like Delhi. Traditionally analytical models and statistical models include synthetic variation models and atmospheric dispersal models, which were applied for prognostication. Recently it was seen that machine learning methods give a more accurate result in cases of prognostication models.

Machine Learning Models

We know that machine learning has made our lives easier because of their accurate predictions. But when these algorithms are combined with that of AI's, the results become clearer. A machine learning approach takes various different factors into account unlike the statistical approach. Artificial Neural Networks (ANNs) have emerged to be one of the several broadly accepted methods for the prognostication of air quality (M. Baawain et al., 2014). Many researchers have build their models based on regression. Artificial intelligence algorithms such as fuzzy logic, generative algorithm, Principal component analysis (PCA) along with ANNs have been applied to create such models like Adaptive Neuro Fuzzy Interface System (ANFIS) model (Sharma, A.K., et al., 2018) etc. Another machine learning models that have been recognized add Support Vector Machine (SVM) situated model, PCA-SVMand several also. A modified Lasso and Ridge regression technique mode0l (Siris, V.A. et al., 2019) where K-nearest neighbour algorithm has also been implemented to determine concentrations of PM2.5, SO2 and PM10. Another study conducted in Quito, Ecuador (Singh, S. et al., 2018), worked on six meteorological constituents for predicting AQI concentrations. K. Hu et al. (M. Baawain et al., 2014), planned a machine training model HazeEst for prognosticating the air index. Here, first, the method was evaluated using seven distinctive regression techniques and finally, SVR (Support vector regressors) was chosen as the ultimate prognostication model. The main goal is to prognosticate an air contamination level in an urban area with the ground data set (Siris, V.A. et al., 2019).

Regression Techniques Applications

This method has used the Linear regression and Support vector regression for the forecast of the contamination of the next month, the next day and any date of future. The method improves to prognosticate any date contamination details within one period based on independent parameters and examining pollution parts and determine future pollution. Time Series Analysis was also used for the identification of future data points which have seasonality and trends in air pollution prediction (Bhalgat, P. et al., 2019).

This designed method performs two significant tasks (i). Identifies the levels of pollutants (S02, PM2.5, CO, benzene) based upon provided meteorological values. (ii) Prognosticates the level of pollutants for a special date (Kumar, S. et al., 2009). Logistic regression is used to identify a data sample is either contaminated or not contaminated. Auto regression is used to prognosticate projected values of pollutants based upon the early pollutants' interpretations. The prime aim is to prognosticate the air pollution level within a particular area with the raw data set (Pandey, G. et al., 2013).

- Air quality prediction using machine learning model by Huiping Peng (2013).
- Nandigala Venkat Anurag, Yagnavalk Burra, S. Sharanya they carried out case study on Air Quality Index Prediction with Meteorological Data Using Feature Based Weighted Xgboost by analyse trend in air quality with year wise the explore key factor that responsible for Air Quality Index predicition.
- Chavi Srivastava, Shyamli Singh, Amit Prakash Singh they also carried out case study on Estimation of Air Pollution in Delhi Using Machine Learning Techniques by estimate the value of AQI by train machine learning model on various regression model like linear regression and etc. (S.A. Rijwan et al., 2007).
- Shivangi Nigam, B.P.S. Rao, N. Kumar, V. A. Mhaisalkar evaluate the Air Quality Index – A Comparative Study for Assessing the Status of Air Quality they evaluate dataset to find specific trends in season related to previous years data with present year and find some pattern in dataset which is use full to find different factor which is best fit for estimate value of AQI and trends according season and lot of factor having correlation between them.

METHODOLOGY ADOPTED IN RESEARCH

Data Source

To prognosticate the air quality of The NCR area, authors' team wanted the pollutant concentration of all the elements available in the air which will be available in the **cpcb.nic.in** the website, which holds all the data that contaminates the area every year. Research Team used data from several stations which measures many elements present in the atmosphere. Data is taken from 10 different stations in NCR. These data are stored in the form of a table which consists of a total of 3469 rows and having 8 columns in each row. The AQI formulae will be applied in order to calculate the AQI by using the various regression algorithms for a particular year.

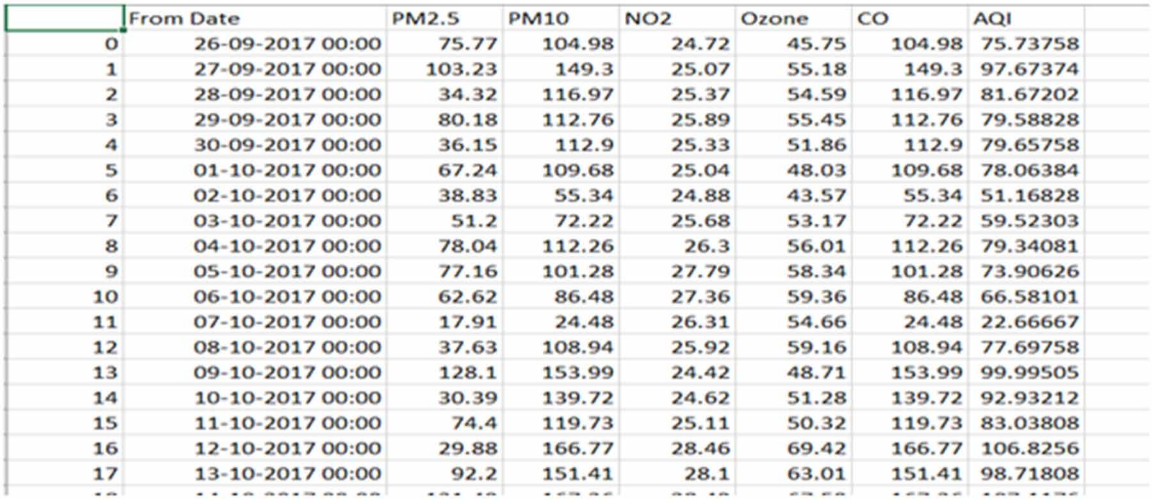
The first step, to build such a model is to collect the raw dataset. The air pollutant dataset was collected from government website cpcb.com which place all record of air pollutant of every year in day wise format, of various formats. The snapshot for the same is shown in Table 1.

Duration: 1stJanuray, 2017 – 1st January, 2020
Various Station: Anand Vihar, Delhi – DPCC,Indirapuram, Ghaziabad - UPPCB,AshokVihar, Delhi – DPCC,Bawana, Delhi – DPCC, and etc
Daily Time Span: 0:00 to 23:59hr (24 hrs)

1. Number of tuples recorded per day: 1
2. Total number of tuples in wanted duration: 3469 after cleaning

NOTE: The data is collected from different station

Table 1. Sample Dataset of Air Pollutant



	From Date	PM2.5	PM10	NO2	Ozone	CO	AQI
0	26-09-2017 00:00	75.77	104.98	24.72	45.75	104.98	75.73758
1	27-09-2017 00:00	103.23	149.3	25.07	55.18	149.3	97.67374
2	28-09-2017 00:00	34.32	116.97	25.37	54.59	116.97	81.67202
3	29-09-2017 00:00	80.18	112.76	25.89	55.45	112.76	79.58828
4	30-09-2017 00:00	36.15	112.9	25.33	51.86	112.9	79.65758
5	01-10-2017 00:00	67.24	109.68	25.04	48.03	109.68	78.06384
6	02-10-2017 00:00	38.83	55.34	24.88	43.57	55.34	51.16828
7	03-10-2017 00:00	51.2	72.22	25.68	53.17	72.22	59.52303
8	04-10-2017 00:00	78.04	112.26	26.3	56.01	112.26	79.34081
9	05-10-2017 00:00	77.16	101.28	27.79	58.34	101.28	73.90626
10	06-10-2017 00:00	62.62	86.48	27.36	59.36	86.48	66.58101
11	07-10-2017 00:00	17.91	24.48	26.31	54.66	24.48	22.66667
12	08-10-2017 00:00	37.63	108.94	25.92	59.16	108.94	77.69758
13	09-10-2017 00:00	128.1	153.99	24.42	48.71	153.99	99.99505
14	10-10-2017 00:00	30.39	139.72	24.62	51.28	139.72	92.93212
15	11-10-2017 00:00	74.4	119.73	25.11	50.32	119.73	83.03808
16	12-10-2017 00:00	29.88	166.77	28.46	69.42	166.77	106.8256
17	13-10-2017 00:00	92.2	151.41	28.1	63.01	151.41	98.71808

Data Preprocessing

Removing Rows With Missing Values

The easiest way to take care of such missing values is to remove the whole row. But, these information should be contained in different rows so that there is no loss of data. Deleting the whole row has not been proved to be so beneficial.

Fixing Errors in the Structure

Typographical or grammatical errors should be always avoided. But, if any how these error are present in dataset, we need to remove it so that they do not cause a problem with the machine learning model. Such errors create confusion in the code.

Calculating AQI

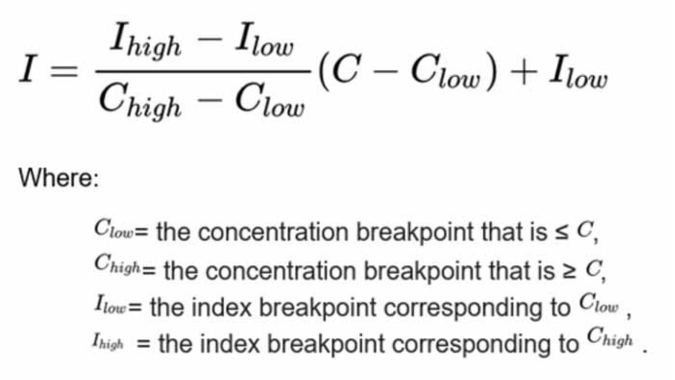
An air quality index (AQI) is used by the government to tell the people how much the cities are polluted. Public health risks increase with the increase in AQI. Different countries have their own AQI. Here, we had focused on the AQI used in India.

Computing AQI

The air quality index is a linear function of the pollutant concentration. At the boundary between AQI categories, there is a discontinuous jump of one AQI unit. To convert from concentration to AQI this equation from Figure 1 is used.

Table 2 is showing AQI Category, pollutants and Health Breakpoints.

Figure 1.The equation to calculate AQI

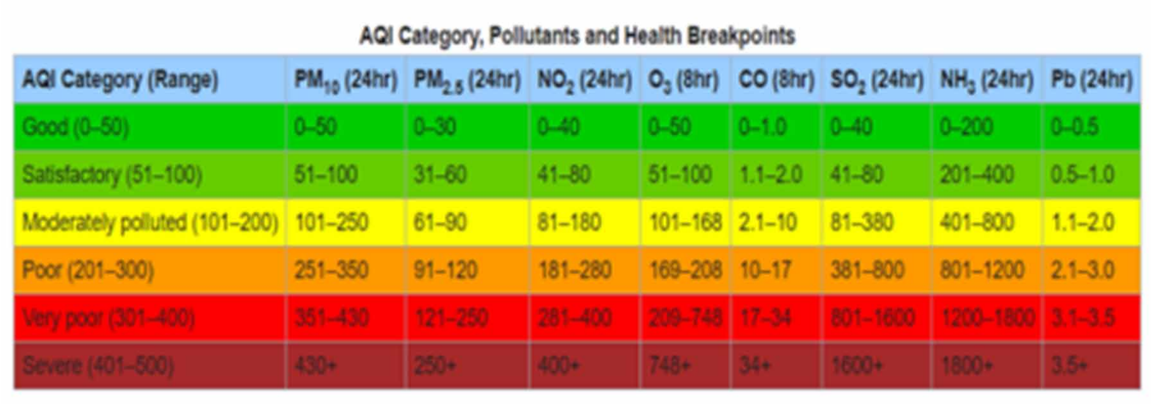


$$I = \frac{I_{high} - I_{low}}{C_{high} - C_{low}}(C - C_{low}) + I_{low}$$

Where:

C_{low}= the concentration breakpoint that is ≤ C,
C_{high}= the concentration breakpoint that is ≥ C,
I_{low}= the index breakpoint corresponding to C_{low} ,
I_{high} = the index breakpoint corresponding to C_{high} .

Table 2. EPA table



AQI Category, Pollutants and Health Breakpoints

AQI Category (Range)	PM_{10} (24hr)	$PM_{2.5}$ (24hr)	NO_2 (24hr)	O_3 (8hr)	CO (8hr)	SO_2 (24hr)	NH_3 (24hr)	Pb (24hr)
Good (0–50)	0–50	0–30	0–40	0–50	0–1.0	0–40	0–200	0–0.5
Satisfactory (51–100)	51–100	31–60	41–80	51–100	1.1–2.0	41–80	201–400	0.5–1.0
Moderately polluted (101–200)	101–250	61–90	81–180	101–168	2.1–10	81–380	401–800	1.1–2.0
Poor (201–300)	251–350	91–120	181–280	169–208	10–17	381–800	801–1200	2.1–3.0
Very poor (301–400)	351–430	121–250	281–400	209–748	17–34	801–1600	1200–1800	3.1–3.5
Severe (401–500)	430+	250+	400+	748+	34+	1600+	1800+	3.5+

Data Pre-Processing

Following Steps were performed for data preprocessing and collection.

Data Refinement

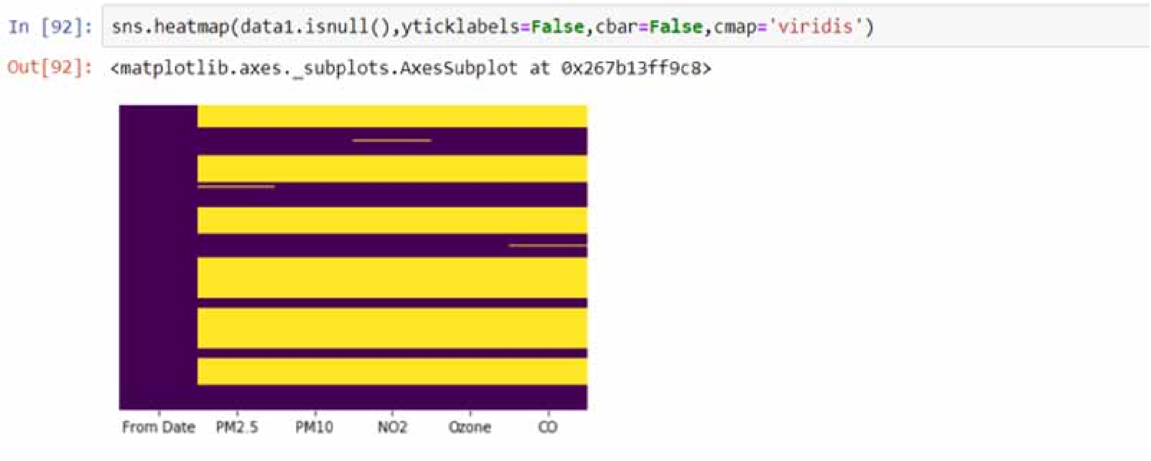
The data to be analyzed by first cleaned, by removing all the unfavorable values. The missing values in case of the target object, i.e., the pollutants are estimated using an imputer function to perform the interpolation. The strategy used here for estimation is the mean value. Data pre-processing projection of the null values by heat map as shown in Fig. 2 and Fig. 3.

Removing null rows having maximum number of null values as show in Fig 4.6

Fill Null Values

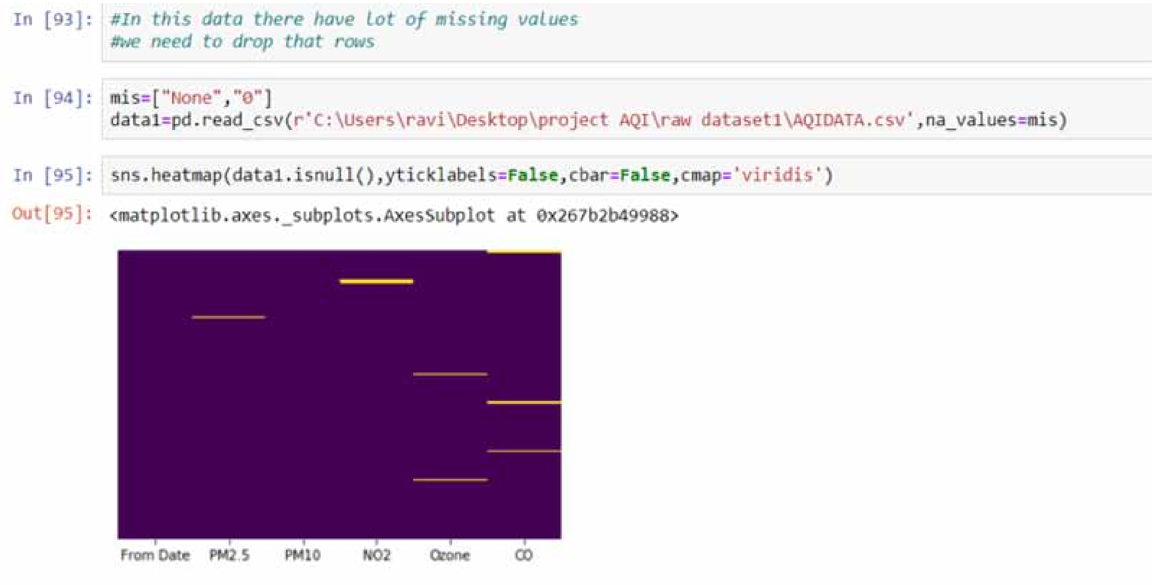
Shown in Fig. 4 and Correlation of data is shown in Fig. 5.

Figure 2. Heat map of our dataset



```
In [92]: sns.heatmap(data1.isnull(),yticklabels=False,cbar=False,cmap='viridis')
Out[92]: <matplotlib.axes._subplots.AxesSubplot at 0x267b13ff9c8>
```

Figure 3. Heat map after removing null rows



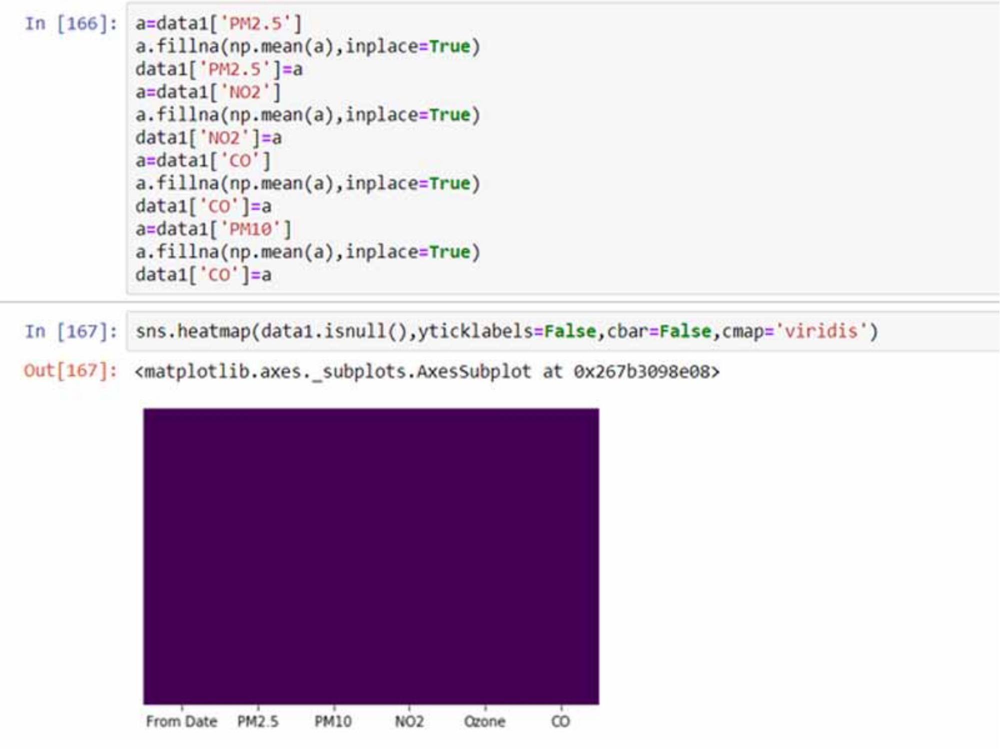
```
In [93]: #In this data there have lot of missing values
         #we need to drop that rows

In [94]: mis=["None","0"]
         data1=pd.read_csv(r'C:\Users\ravi\Desktop\project AQI\raw dataset1\AQIDATA.csv',na_values=mis)

In [95]: sns.heatmap(data1.isnull(),yticklabels=False,cbar=False,cmap='viridis')
Out[95]: <matplotlib.axes._subplots.AxesSubplot at 0x267b2b49988>
```

Now, visualizing the dataset. We'd try to find the correlation between attributes using 'heat map' as shown in Fig. 5.

Feature Selection

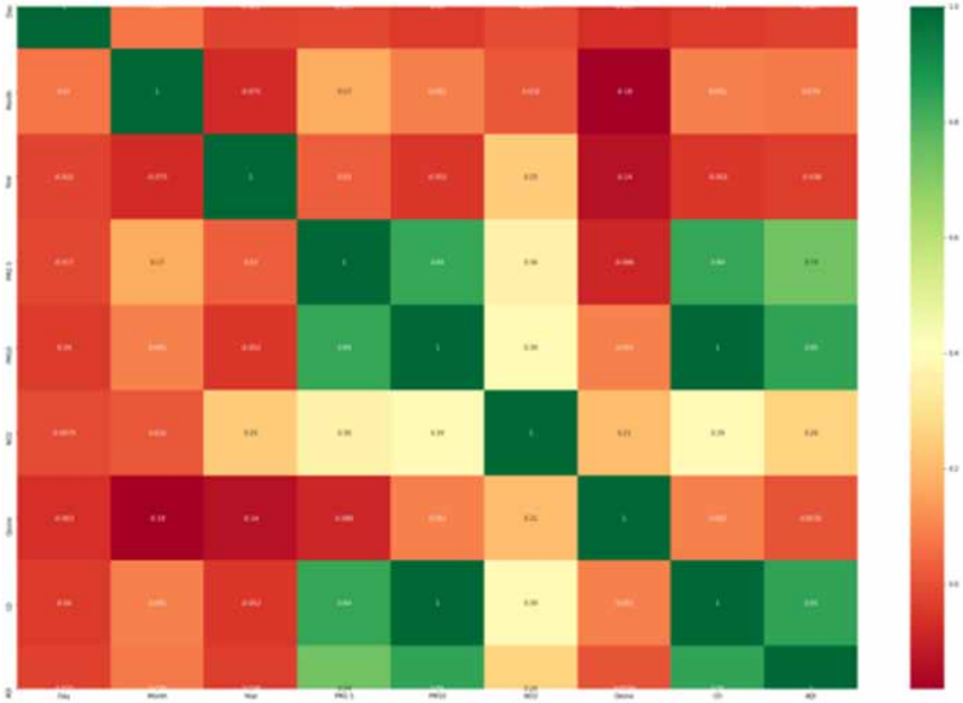
Feature selection is the method of choosing a subset from the features that contain important data. In the case of unnecessary data, feature extraction implies used. Feature extraction includes the choice of best input parameters of the chosen input dataset. The unified dataset which was gathered is used for

Figure 4. Filling missing values by mean



```
In [166]: a=data1['PM2.5']
          a.fillna(np.mean(a),inplace=True)
          data1['PM2.5']=a
          a=data1['NO2']
          a.fillna(np.mean(a),inplace=True)
          data1['NO2']=a
          a=data1['CO']
          a.fillna(np.mean(a),inplace=True)
          data1['CO']=a
          a=data1['PM10']
          a.fillna(np.mean(a),inplace=True)
          data1['CO']=a

In [167]: sns.heatmap(data1.isnull(),yticklabels=False,cbar=False,cmap='viridis')

Out[167]: <matplotlib.axes._subplots.AxesSubplot at 0x267b3098e08>
```

Figure 5. Correlation in dataset



further analysis. The maximum amount of inputs available for review is seven; hence all the inputs are selected for the computations (As per Fig. 6).

Training the Model

Data Splitting was done as 80% for training and 20% for testing

RESULTS AND DISCUSSIONS

Collective Analysis

The machine learning model to predict air pollution was one the most important objective of the research paper. We here predict air pollution on particular data i.e. 1 January 2020.We take data of previous years the data from www.cpcb.nic.in .By using linear regression, Lasso-Ridge regression, KNN and support vector machine. We actually compare actual value and predicted values. Self-explanatory Data Visualization as shown in Fig. 7 to Fig. 9 as per order of Year, Month and their averages.

After evaluation of different type of regression model, it was found that the best fit model for predict AQI is K-nearest neighbor model which having accuracy of 97.5 percentage. This model is best fit.

Here, we show that the almost all values is equal some particular values show anomalous behavior. If we see average of AQI in month wise Fig.8 and 9, we easily say that AQI not only depends on concentration of particle it also show that AQI also depends on temperature, humidity, etc. It conclude that we need more data and Add more column in dataset of other factor like temperature and etc.

We clearly see that, if we particularly show average July and august month in 2020 it is far more differ than other years same month.

Applying Various Repressors'

Here we used various regression algorithm for predicting dependent variable such that

- Multiple linear regression
- Lasso and Ridge regression

Figure 6. Model for training the dataset

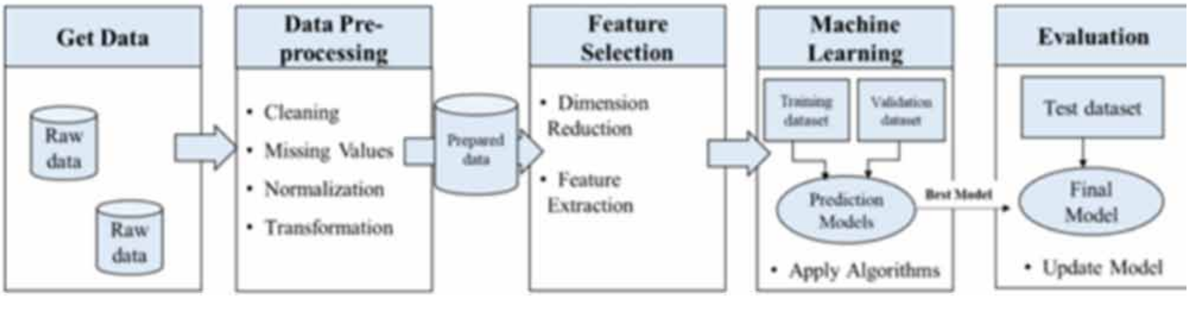


Figure 7. Average of AQI year wise

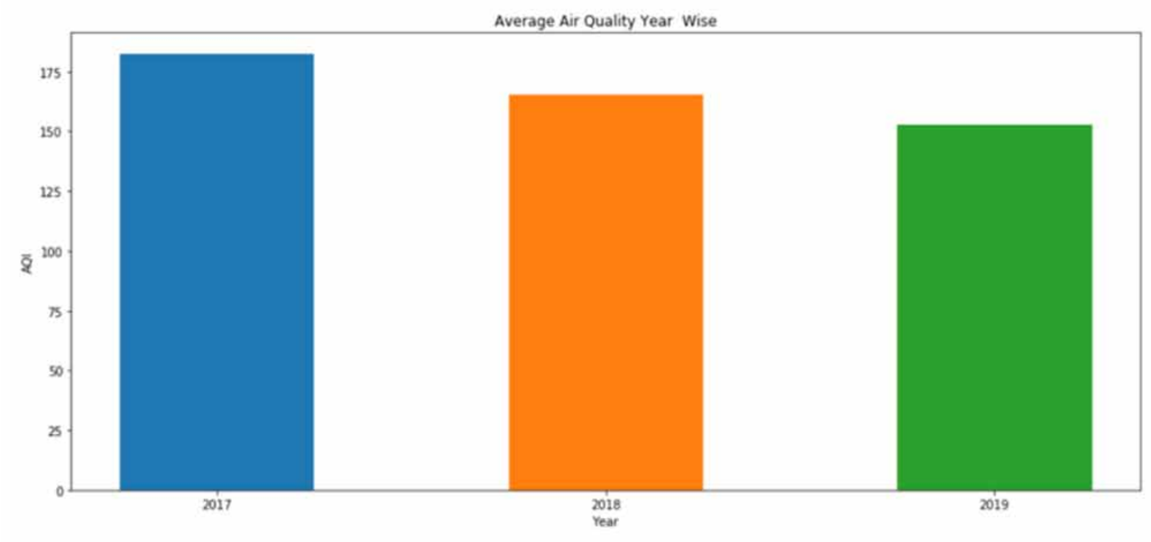
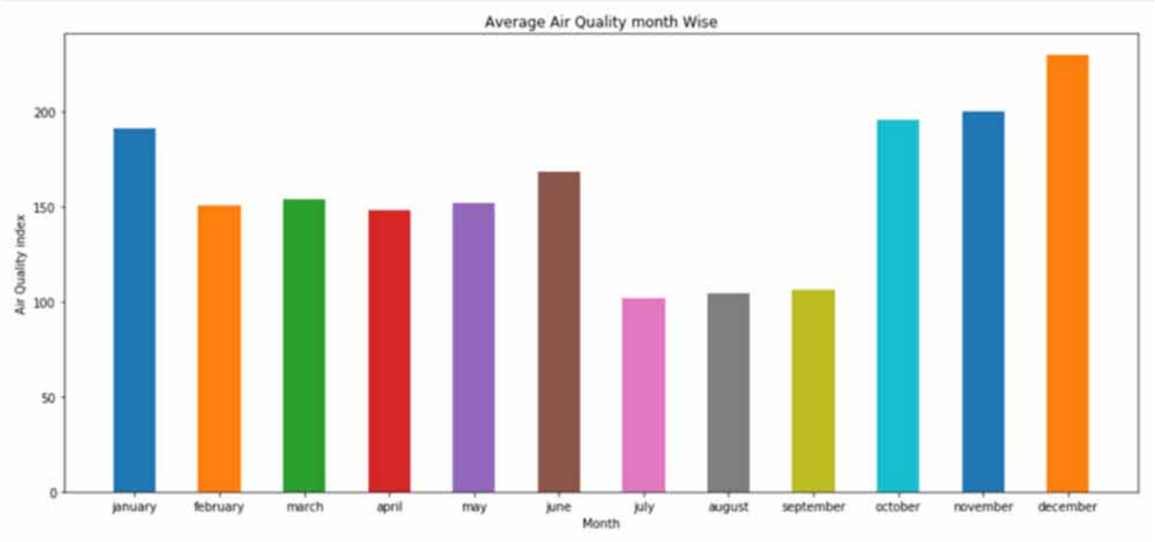


Figure 8. Average of AQI MonthWise



- Support vector regressors
- k-nearest neighbors regression

K nearest neighbor is best fit model which having accuracy of 97.5% Its is best model of predicting air quality index prediction on independent variable (PM2.5, PM10 and etc) as shown in Fig. 10 and Fig. 11.

As shown Fig. 11 comparing actual values with predicted values.

Fig.11. shows comparison between actual values and predicted values here we show that actual values is almost equal to predicted values except some values.

Figure9. Average of AQI according to month

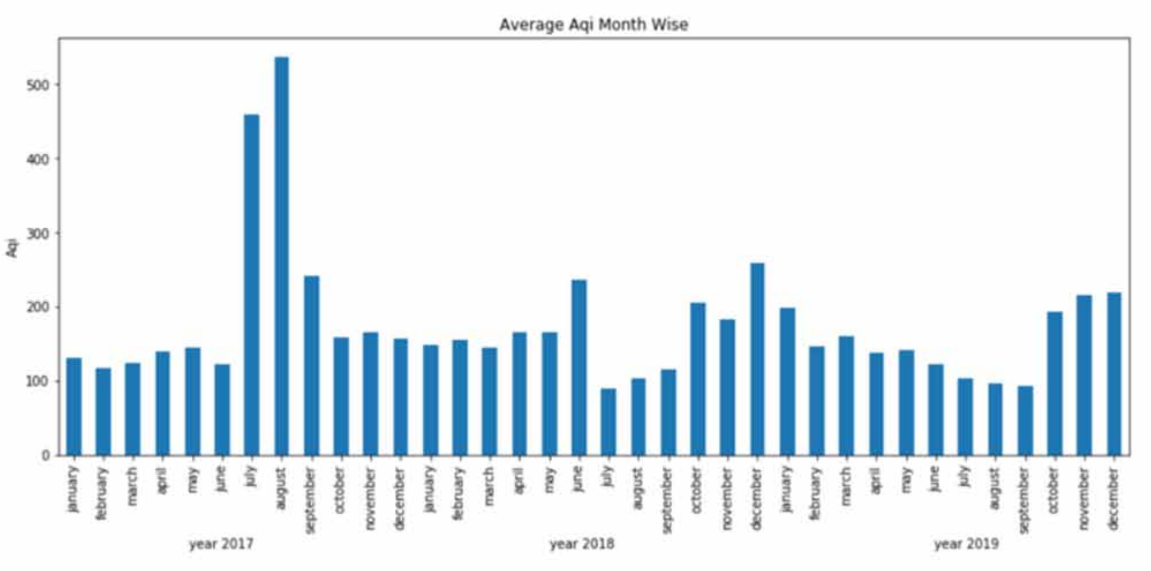
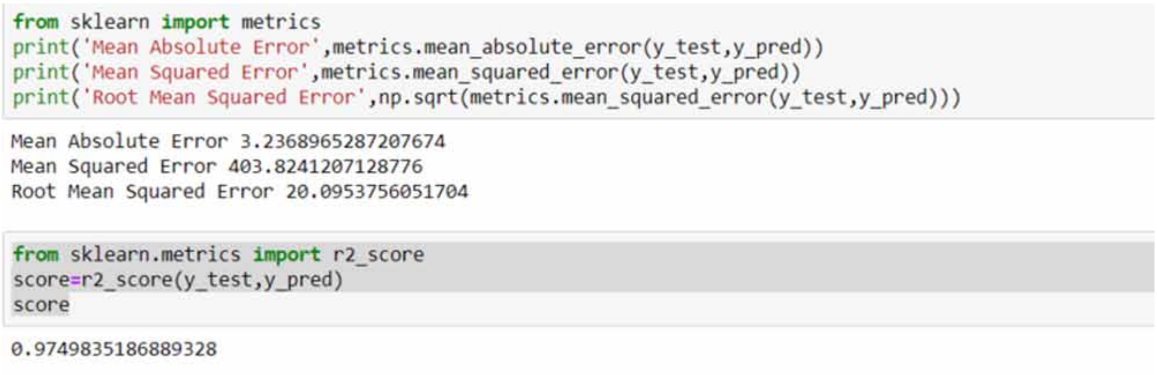


Figure 10. Accuracy of k-nearest neighbour



```
from sklearn import metrics
print('Mean Absolute Error',metrics.mean_absolute_error(y_test,y_pred))
print('Mean Squared Error',metrics.mean_squared_error(y_test,y_pred))
print('Root Mean Squared Error',np.sqrt(metrics.mean_squared_error(y_test,y_pred)))
```

```
Mean Absolute Error 3.2368965287207674
Mean Squared Error 403.8241207128776
Root Mean Squared Error 20.0953756051704
```

```
from sklearn.metrics import r2_score
score=r2_score(y_test,y_pred)
score
```

```
0.9749835186889328
```

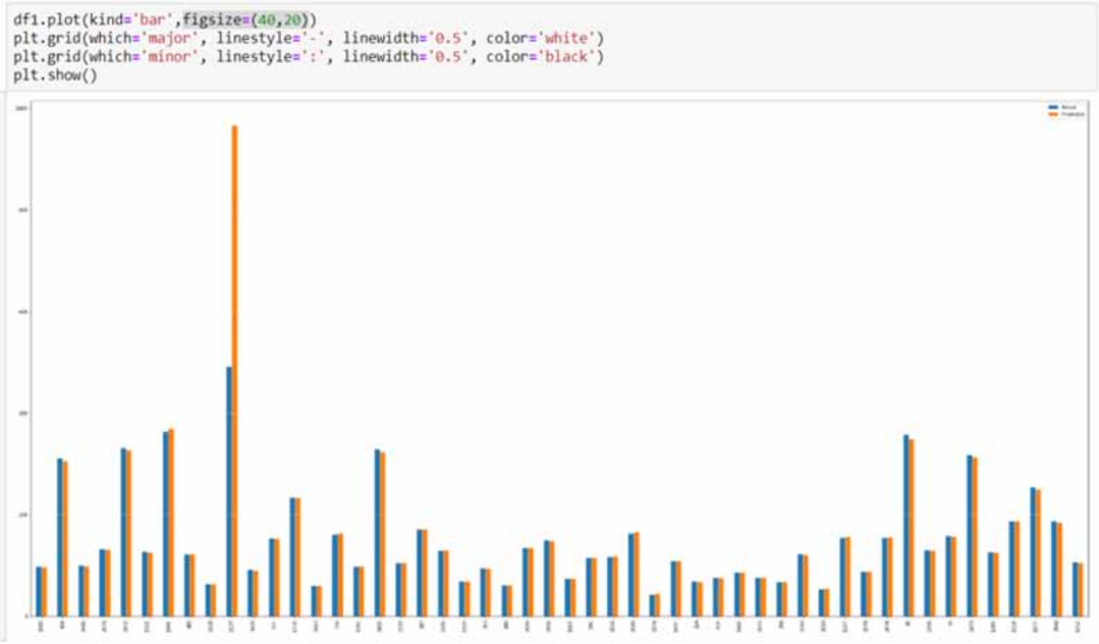
Comparison With Existing State-of-the-Art Technologies

This project has used several factors like temperature, occasion and weather conditions in order to improve the results.

Performance Evaluation

Based on the comparison of out-of-sample RMSE among the models, along with taking into account the interpretability of the model, we have decided that k nearest neighbor is the best model. For reference, following is the performance of the models and comparison of their in-sample and out-of-sample RMSE values (as per Table 3).

Figure 11. Comparison between actual values and predicted values



```
df1.plot(kind='bar',figsize=(40,20))
plt.grid(which='major', linestyle='-', linewidth='0.5', color='white')
plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')
plt.show()
```

Evaluation

After implementing different types of algorithm, we need to evaluate which algorithm is best for predict AQI of NCR. Table no 2 show root mean and r squared values of algorithm which mention in above paragraph.

NOVELTIES IN THE WORK

This research used various algorithms of regression and classification like LR, SDGR, RFR, DTR, SVM, ANN. To make the predictions more accurate the reasearch team have used MSE and MAE with R square. For predicting air quality index of NCR(national capital region) in different aspects of like stubble farming, Motor vehicle emission, and open construction practice which impact of air quality of NCR.

Table 3. Model evaluation

Models	Mean Absolute Error	Root Mean Squared Error
Multiple Linear Regression	38.386647	65.075391
Lasso Regression	38.344869	65.059393
Ridge Regression	38.385283	65.075104
Support Vector Regressor	0.099856	0.387394
K-Nearest Neighbors	3.236896	20.095375

FUTURE RESEARCH DIRECTIONS

There is a lot of scope in future regarding this topic as various aspects can be covered regarding this field. Also, air pollution and environment analysis can also be done with the help of Advanced Machine Learning Algorithms or Deep Neural Networks. In order to build a good predicting model, we can find the air quality index by taking various parameters into consideration like temperature, humidity, etc. Also, models implemented in this paper can be applied to a greater number of stations forincreaseing the training input (Baawain, M. 2014). It is also observed that the approach can also help to identify the predictor(s) for which the variance is not properly captured (reason for heteroscedasticity). This will also help in solving the problem for normality as well. We can still search for other avenues in order to look for quality controlled data. The next research seekers can also train their model for next year data and also, solve the problem of auto-correlations (Hornos, M.J. et al., 2018).

LIMITATIONS

Various limitations were faced in writing this challenging paper and carrying out such a difficult task. The experiment needs further more tiny constraints for future predicting AQI. In this paper, instruments applied for finding readings were limited to some range and generated the limited results (Jalan, I. et al., 2019). So, to get more refined results, one needs more refined instruments. Due to the limitation of the data, some aspects are still left to be covered which can be interesting research area. Also, this experiment was confined to a small place, so in order to view the environmental situation from different perspectives; one may needs to extend future research to different parts of the country or continent (Hornos, M.J., 2017).

CONCLUSION

In the populated and developing countries, governments consider the regulation of air as a major task. Monitoring air qualityusing Intelligent and Smart Environment Solutions is a necessity due to various pollution causing activities including stubble burning and open construction practices contributing a lot in the air pollution. So, we can forecast air quality index using machine learningalgorithms working as a part of Smart Environment systems in order to predict air quality index for NCR zone at India.

For making a advanced training and predicting model, any researcher seeker and interested can take various parameters into examination like temperature, concentration of each gas in the atmosphere, pressure, etc. Place of experiment is also a major parameter to take into consideration as each and every place has its own environment related problems.

Air quality indexes of major pollutants like PM2.5, PM10, CO, NO2, SO2 and O3.in recent years machine learning in most emerging technology for predicting on historical data with 99.99% of accuracy. Advanced fields including Artificial Intelligence, Data Analytics, Deep Learning Algorithms, etc. can be very helpful in saving the environment from pollution.

REFERENCES

Aggarwal, P., & Jain, S. (2015). Impact of air pollutants from surface transport sources on human health: A modelling and epidemiological approach. *Environment International*, *83*, 146–157. doi:10.1016/j.envint.2015.06.010 PMID:26142107

Baawain, M. (2014). *Systematic Approach for the Prediction of Ground Level Air Pollution (around an Industrial Port) Using an Artificial Neural Network*. Aerosol Air Qual. doi:10.4209/aaqr.2013.06.0191

Bhalgat, P. (2019). Air Quality Prediction using Machine Learning Algorithms. *International Journal of Computer Applications Technology and Research, 8*(9).

Bhalla, N. (2018). Who is Responsible for Delhi Air Pollution? Indian Newspapers’ Framing of Causes and Solutions. *International Journal of Communication*, *12*, 41–64.

Dipanda, A., Damiani, E., & Yetongnon, K. (2016). Special issue on intelligent systems and applications in vision, image and information computing. *Journal of Reliable Intelligent Environments*, *2*(3), 117–118. doi:10.1007/s40860-016-0032-8

Gallardo, M., Lavado, L., Panizo, L., & Titolo, L. (2017). A constraint-based language for modelling intelligent environments. *Journal of Reliable Intelligent Environments*, *3*(1), 55–79. doi:10.1007/s40860-017-0040-3

Hornos, M. J. (2017). Application of Software Engineering techniques to improve the reliability of Intelligent Environments. *Journal of Reliable Intelligent Environments*, *3*(1), 1–3. doi:10.1007/s40860-017-0043-0

Hornos, M. J., & Rodríguez-Domínguez, C. (2018). Increasing user confidence in intelligent environments. *Journal of Reliable Intelligent Environments*, *4*(2), 71–73. doi:10.1007/s40860-018-0063-4

Jalan, I., & Hem, H. (2019). What is Polluting Delhi’s Air? Understanding Uncertainties. *Emissions Inventories.* www.ceew.in

Kumar, S. (2009). *Air Pollution and Climate Change: Case Study National Capital Territory of Delhi.*

Pndey, G., Zhang, B., & Jian, L. (2013). Predicting sub-micron air pollution indicators: a machine learning approach. *Environmental Science: Processes & amp. Impacts*, *15*(5), 996–1005.

Rizwan, S. A. (2007). Center for Community Medicine. All India Institute of Medical Sciences, New Delhi, India

Sharma A.K., Baliyan P.,(Mar. 2018). Air pollution and public health: The challenges for Delhi, India. *Reviews on Environmental Health, 01, 33*(1), 77-86.

Singh, S., & Singh, A. P. (2018). Estimation of Air Pollution in Delhi Using Machine Learning Techniques. *International Conference on Computing, Power and Communication Technologies.* Research Gate. https://www.researchgate.net/publication/332430367

Siris, V. A., Fotiou, N., Mertzianis, A., & Polyzos, G. C. (2019). Smart Application-aware IoT data Collection. *Journal of Reliable Intelligent Environments*, *5*(1), 17–28. doi:10.1007/s40860-019-00077-y

Srivastava, C., Singh, A. P., & Singh, S. (2018). *Estimation of Air Pollution in Delhi Using Machine Learning Techniques.* Research Gate. https://www.researchgate.net/publication/332430367

Tripathi, C. B., Baredar, P., & Tripathi, L. (2019, October 10). Air pollution in Delhi: Biomass energy and suitable environment policies are sustainable pathways for health safety. *Current Science*, *117*(7), 1153. doi:10.18520/cs/v117/i7/1153-1160

Wilson, T., Legay, A., & Sedwards, S. (2018). Group abstraction for assisted navigation of social activities in intelligent environments. *Journal of Reliable Intelligent Environments*, *4*(2), 107–120. doi:10.1007/s40860-018-0058-1

KEY TERMS AND DEFINITIONS

Data Visualization: Data visualization is used to represent dataset and understand it through variousplots and graphs. In air quality, we have used heatmap, correlation matrix and boxplot to visualize the data. Through, Data visualization, we get to know the relationbetween various attributes of the dataset. Python offers multiple libraries for datavisualization and analysis.

k-Nearest Neighbour Regression: K nearest neighbors is a simple algorithm that stores all available cases and predict the numerical target based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970’s as a non-parametric technique.

Lasso and Ridge Regression: Ridge regression and Lasso regression are very similar in working to Linear Regression. The only difference is the addition of *the l1* penalty in Lasso Regression and *the l2* penalty in Ridge Regression. The primary reason why these penalty terms are added is two ensure there is regularization, shrinking the weights of the model to zero or close to zero to ensure that the model does not overfit the data.

Multiple Linear Regression: Multiple Linear Regression (MLR) is a statistical technique for finding the linear relation between the independent variables (predictors) and the dependent or response variable. The general MLR model is built from N observations of the multiple predictor variables xk (k = 1, 2,.., m) and the observed target data y.

Support Vector Regressor: 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.