Data Mining and Machine Learning approach for Air Quality Index Prediction.

Mayuresh Mohan Londhe
M.Sc in Data Analytics
National College of Ireland
Dublin, Ireland
x20137265@student.ncirl.ie

***Abstract*— In recent years, Air Pollution has increased drastically and having worse effect of that on all the living beings. Majority of Countries in the world battling with increasing Air Pollution Levels. So, it has become a necessity to control and predict the Air Quality Index. In this research project, we will be implementing Data Mining and Machine Learning models to predict the AQI and Classify the AQI into buckets. For AQI prediction we have implemented five regression models Principal Component, Partial Least Square, Principal Component with Leave One Out CV, Partial Least Square with Leave One Out CV, Multiple regression AQI Data of Multiple Indian Cities. AQI Index further gets classified into 6 Different Categories called Buckets "Good, Satisfactory, Moderate, Poor, Very Poor and Severe" based on the value of the AQI. To predict the AQI bucket we have developed three classification models which are Multinomial Logistic Regression and K Nearest Neighbor and K Nearest Neighbors with repeat CV Classification algorithm. From the Air Quality Dataset of Different Indian Cities PLS model with Leave One Out Cross Validation was best at dimension reduction considering only the 5th component from all the models. In terms of accuracy PLS model was best with Lowest RMSE. From Station Wise Data of Indian Cities KNN Model with Repeated CV and Tune Length 10 performed best in terms of accuracy and AUC.**

Keywords—AQI, KNN, MLR, PLS, PCR, AUC, RMSE, MAE, MAPE, ROC

I. INTRODUCTION

As the globalization and industrialization era started the world witnessed radical lifestyle changes but as we progressed further, we encountered demerits of globalization and industrialization in terms of different types of pollutions. Out of various pollutions, Air pollution is the most hazardous and widely affected all over the globe. Researchers and Innovators are profoundly working towards finding different solutions to mitigate the hazard caused by poor air quality using prediction and monitoring methods. In 2018 European Parliament declared Breeze Technologies as the Most Promising Start-Up of the Year which focuses on monitoring Air Quality.

Air pollution occurs when a combination of different hazardous gases and solid particles reaches a detrimental level. Major factors for this pollution are Industrial and Vehicle emissions. Global Warming, Acid Rains, Extinction of animal species and health problems are effects of Air Pollution. Particles that are as small as 0.01mm can cause illness and cardiovascular disease. Air Quality index is computed by measuring the level of different air pollutants such as PM2.5, PM10, CO, NO, NO2, SO2, O3 in the air.

Air Quality in the majority of Indian states and cities is proven to be unsafe. As a result, the count of cardiovascular diseases and chronic illnesses increased. Effectively the number of deaths due to Air pollution is 1.67 Million in the year 2019 out of 6.7 million deaths globally. In the list of most polluted cities in the world, 21 out of 30 were from India.

By predicting the Air Quality Index (AQI) using historical data could help to prevent deaths due to Air Pollution. This would assist government and environmental organizations to implement preventive measures.

This research paper is presented using the following template. First Section of the template provides an introduction of the research topic which focuses on highlighting the motivation behind the selection of the topic, present situation and research question. Section 2 elucidates related work, which mainly provides information regarding Literature presented by other researchers, Key results achieved by them, positive and Negative points of the research, Selection of models and methodologies their limitations. Section 3 provides a description of Data Mining Methodologies implemented in research. The fourth section describes models implemented and the evaluation of their results. Conclusion and future scope is the last section which is a summary of the research and further work that can be carried out to improvise and extend further.

II. RELATED WORK

In the paper [1], Multiple regression methods are applied to a dataset to predict the AQI for Delhi and Houston cities. Two different AQI scales were used to classify AQI values into 6 different groups. Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RSME) performance evaluation methods used to evaluate the performance and accuracy of the model. Out of implemented models, multiple linear regression model with batch mini-batch gradient outperforms in terms of High Performance. The number of observations used for prediction is relatively low. Comparative studies can be done with respect to results obtained Multiple Linear regression model from paper [1] and the model implemented in this research paper.

In [2], Multinominal Logistics Regression (MLR), Decision Tree and K mean algorithm are examined for the prediction. Only a few important Pollutants were taken into consideration when implementing the model. AQI labels were divided into 3 categories Moderate, Good, and Unhealthy. Accuracy and Error of the model are predicted based on correct labeling of the data according to AQI levels. Out of all implemented models, MLR performed better than the Decision Tree model with an error rate of 0.442 and 0.666 respectively. Data Labeling, Data Splitting, Machine

Learning, Process, Prediction Process were involved in model building. Here there is relatively less scope to compare the performance of Multinominal Logistic Regression with the model implemented in this paper as a number of AQI labels are not identical.

The work in paper [3], includes the requirement to build an Air Quality prediction model for the Northern Thailand region. Linear regression, neural networks, and genetic programming. algorithms were compared and analyzed. In the first model Data gathered from 12 different stations was averaged month-wise for the first five months of the Year and AQI was predicted for an entire month. In the second model, the data was split between Unhealthy Air and Unpolluted Air and forecasted station wise. Apart from Air Pollutants and AQI other environmental factors were involved in the analysis. Linear Regression performed well in comparison to the other two models with an accuracy of 97.78. Root Mean Square Error (RMSE) is adopted to calculate the effectiveness of the model.

[4] Here, an Accuracy of 98.67%, 97.78%, 98.67%, 94.22% and 99.56% was recorded for models K nearest neighbors (KNN), Support Vector Machine (SVM), Naïve Bayesian classifier, random forest and neural network respectively. Models were built using Preliminary Knowledge, Hardware Development, Data Gathering, Data Pre-Processing, Development and training of the model and Testing and Evaluation steps. Data was collected by privately owned multiple sensors after that data was preprocessed and labeled. Only 750 observations were taken into consideration. Here the source of the data is not a government or organization which could be a demerit as sensors used by a government agency could be more sophisticated in data gathering and sensing.

[5], In this different Auto Regression Models were implemented to predict the Air Quality Index (AQI) seven days prior. Hourly data of 11 days was further divided into 8 days of Training and 3 Days of Testing. Values used for P, D, Q while implementing the Arima model are not shared. Visualization of the time series, Stationaries the series, Plot Parameters, Build Model, make predictions were the steps involved. It was predicted with a Mean Square Error of 27.00. Only four pollutants and other atmospheric quantities were considered to predict Air Quality.

The author in [6], have used a different approach to Air Quality Prediction. The data of Pollutants concentrations and meteorological attributes were used separately. First Meteorological data was clustered using K means. Results from the clustering were then used to build a hybrid model. Results from the Hybrid model were compared with K Means, Support Vector Machine, Neural Network, Deep Leaning. The hybrid model proved to be better than all other models with respect to the error rate.

In the paper [7], Air Quality Indexes are computed and clustered using pollutants concentration amount. For each pollutant, there is a different scale used to decide the AQI and Respective Label. K means and Fuzzy C means algorithms were implemented for clustering the computed AQI into labels. In which Fuzzy C Means exhibited 100% accuracy which is better than K Means. Data was not split between training and test data due to which accuracies obtained casts doubt. We will be interested in analyzing the method used to compute the AQI from Pollutants concentration.

[8] AQI and Weather data of identical timeframe was merged to compare the relationship between the AQI values with Weather and to predict the AQI multiple regression models were implemented on merged data. Only RMSE was taken into consideration for evaluating the models. All the models were having RMSE in the close range between 40 to 44 except KNN with an RMSE of 58.

[9] Here, the source of the dataset is the same as one of the datasets used in the implementation of this research paper. There is a significant difference in approaches, in [9], Data of Delhi city for 31 Days from 37 different stations was collected. We will be interested in a comparison of the results of K nearest neighbor and Multinomial Logistic Regression with Support Vector Machine. Observations used in the approach are relatively less as the mean of the AQI was taken for 37 stations for 31 Days.

The dataset used in [10], is relatively large with fifty thousand observations also it was divided into a 70:30 ratio for training and testing. Light Gradient Boosting model performs better than Xtreme Gradient Boosting which is based on MSE Value.

III. DATA MINING METHODOLOGY

Knowledge Discovery in Databases methodology is adopted for analyzing the databases. This is an iterative process defined by a sequence of steps. The primary objective of KDD methodology is to extract the hidden knowledge from the databases. The below figure represents the steps involved in KDD methodology.

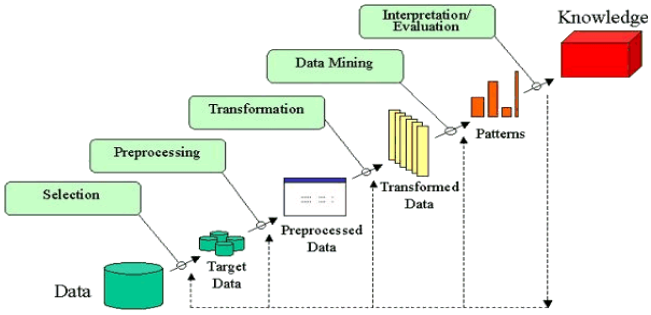

Fig:1 KDD Methodology Steps

We need to select the number of related datasets, three large datasets with more than 10000 rows and more than 10 columns related to AQI of Indian cities are selected. Air Quality Dataset of Different Indian Cities contains 16 columns which is a day-wise AQI of cities. Station Wise Data of Indian Cities contains 16 columns which is station wise data of each day for different states. Hourly Data of Indian Cities contains 11 columns, which is a station wise, Three days data of different cities.

The first step is selection. In which we have selected the Target Data from datasets. From the Air Quality Dataset of Different Indian Cities, we selected AQI Data of 13 Cities

Ahmedabad, Amritsar, Bengaluru, Chandigarh, Chennai, Delhi, Gurugram, Hyderabad, Jaipur, Kolkata, Lucknow, Mumbai, Vishakhapatnam. These are prominent cities in India where the majority of Urban and industrial activities take place. For Dataset two, we selected data of 2 stations from Andhra Pradesh State,6 stations from Bihar state, 5 Stations from Delhi state, 6 Stations from Maharashtra state. For the Hourly Data of Indian Cities, we have considered data of all the cities available.

Data pre-processing is the second step in KDD methodology. Data often contains noise, outliers and null values which could degrade the quality of analysis, due to which we need to handle noise, outliers and null values appropriately. AQI's are computed based on values obtained from sensor devices which provides us the amount of pollutant present in the atmosphere. In rare situations, these sensor devices fail or they need to be serviced to obtain accurate results from the device. This failure of the device to sense the pollutant value results in null value computation of that pollutant for that period.

Below Fig 2. represents the number of null values in each column of the Air Quality Dataset of Different Indian Cities.

City	Date	PM2.5	PM10	NO	NO2
0	0	380	2496	308	273
NOx	NH3	CO	SO2	O3	Benzene
145	2443	120	505	419	262
Toluene	Xylene	AQI	AQI_Bucket		
1214	5801	382	382		

Fig 2. Null Values Column wise Air Quality Dataset of Different Indian Cities.

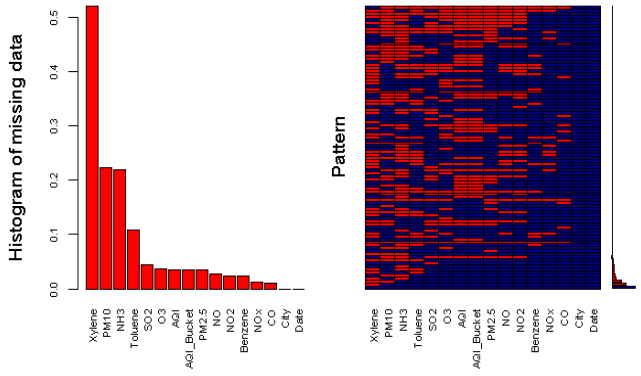

Fig 3. Graphical Representation of Null Values

At First, we can see that AQI and AQI Bucket columns have 382 rows which indicate that all row doesn't contain any record of pollutants due to which we will remove those 382 Rows from Air Quality Dataset of Different Indian Cities. Now we are remaining with pollutants columns, we will be using two different methods to impute the null values based on the number of missing values in the column. For columns that contain Null Values less than equal to 550, we will be using the fill function. As this is daily data, we are assuming that there won't be a significant change in AQI of the Pollutant from Previous Day or One Day Ahead. The values will be imputed in both the direction which means first it will fill values in the down direction which will take the value of previous day to fill the missing value but if the previous day value is also null it will get imputed when there will be upward imputation which will take the value of one day after. Now we are left with Column PM10, NH3, Toluene, xylene. It won't be appropriate to impute null values with a previous method when there are a greater number of missing values. We will be using the impute_lm function which will allow us to use a regression model to impute the missing values. Multiple linear regression is used for all the remaining columns but independent variables are used depending on the type of pollutant.

PM10 ~ Benzene + O3+ SO2 + CO + NO + PM2.5+ NO2

NH3 ~ NO+NO2+NOx

Toluene ~ Benzene + O3+ SO2 + CO + NO + PM2.5

Xylene ~ Benzene + O3+ SO2 + CO + NO + PM2.5+ NO2

For Station Wise Data of Indian Cities, we first created 19 subsets of the dataset which contains data of respective stations and then they were merged using R bind function. Here we have used multiple methods to impute missing values i.e., removing insignificant columns, Mean value Imputation, mice imputation. Firstly, we removed Columns Xylene, Toluene and NH3 as missing counts are 11118,5189,4390 respectively. After this step for column PM2.5, we imputed the mean value of the column values in place of null values. Next, we applied the predictive mean matching method is used which helps reduce bias introduced during the imputation process by imputing real values from sampled data. We have kept the value m as 5 which is the default, maxit value to 3, as the number columns are more and it does require significant computation time. This will generate set of values for 8 variables. We selected Station Wise Data of Indian Cities computed by PMM to impute the Null values. Now we removed Station ID and Date Column to perform feature scaling before building a model. Below fig 4. display the location of Missing Values in the dataset.

Fig 4. Missing Value Plot

Hourly Data of Indian Cities contains concentrations of Pollutants which are in Âµg/mÂ³. First of all, we need to convert these concentration values into AQI and then compute the AQI. This step is done in Microsoft Excel. AQI Calculator Excel was downloaded, using the formulas given in the document AQI values were computed for pollutants and then the Overall AQI value was computed by taking maximum value among all parameters for the particular instance of time. After this step Dataset was again imported in R software. The initial dataset was containing null values which were imputed using the fill function but the Dataset of AQI was having many instances where Pollutant AQI was Zero and Negative values were present. In ideal

situations AQI value cannot be Zero or Negative so using abs() function in r we converted all negative values of numeric columns to positive values. After this, we converted all Zero values to Null Values and Imputed them using the fill function as this dataset contains data of Three days taken at unspecified time intervals.

Descriptives

	co	no2	o3	pm10	pm25	so2
N	10974	11319	10665	10356	11809	10413
Missing	1338	993	1647	1956	503	1899
Mean	1198.113	42.38287	30.71347	215.8405	110.1448	15.22448
Median	960.0000	27.70000	14.61000	183.7850	81.60000	9.210000
Minimum	-4930.000	-162.6000	-5.490000	-59.07000	-999.0000	-2.600000
Maximum	9700.000	854.4000	414.6000	3457.300	1000.000	1338.900

Fig 5. Null Values from Initial Dataset

Descriptives

	coaqi	no2aqi	o3aqi	pm10aqi	pm25aqi	so2aqi
N	12040	12115	12120	11757	12130	12153
Missing	272	197	192	555	182	159
Mean	209.6916	51.44670	36.63146	193.4465	208.5384	16.99056
Median	195.0000	34.68750	17.96000	154.8400	177.2167	11.50000
Minimum	-2465.000	-203.2500	-5.490000	-59.07000	-1665.000	-3.250000
Maximum	770.5882	778.6667	338.3302	4184.125	976.9231	367.3625

Fig 6. Null Values after Converting Zero to Null

The third and important step of the KDD process is Data Transformation. As all three datasets were having the majority of columns with numerical values there was little scope for data transformation. In all three Datasets, the Date column was in factors format. Using as. Date () function converted them in Date format. Hourly Data of Indian Cities contains row-wise information of pollutants unlike column-wise data of the other two datasets. It would be difficult to perform analysis on such a dataset, to convert the row-wise data of pollutants into column-wise we used the Spread function which takes the Key-Value pair as an argument. We passed Parameter and Value columns to the spread function which provided us five different columns of individual pollutants. Later this dataset was exported for AQI Conversion.

All the three Datasets were then randomly sampled into Training and Testing Datasets in a 75:25 ratio.

	location	city	local	parameter	value
1	GIDC, Ankleshwar - GPCB	Ankleshwar	2020-10-28T17:00:00+05:30	o3	39.35
2	GIDC, Ankleshwar - GPCB	Ankleshwar	2020-10-28T17:00:00+05:30	co	910.00
3	GIDC, Ankleshwar - GPCB	Ankleshwar	2020-10-28T17:00:00+05:30	so2	20.31
4	GIDC, Ankleshwar - GPCB	Ankleshwar	2020-10-28T17:00:00+05:30	no2	17.15
5	GIDC, Ankleshwar - GPCB	Ankleshwar	2020-10-28T17:00:00+05:30	pm10	88.39
6	GIDC, Ankleshwar - GPCB	Ankleshwar	2020-10-28T17:00:00+05:30	pm25	0.00
7	NISE Gwal Pahari, Gurugram - IMD	Gurugram	2020-10-28T17:00:00+05:30	co	1490.00
8	NISE Gwal Pahari, Gurugram - IMD	Gurugram	2020-10-28T17:00:00+05:30	no2	18.43
9	NISE Gwal Pahari, Gurugram - IMD	Gurugram	2020-10-28T17:00:00+05:30	pm10	271.96
10	NISE Gwal Pahari, Gurugram - IMD	Gurugram	2020-10-28T17:00:00+05:30	pm25	138.27

Fig 7. Initial View of Hourly Data of Indian Cities

	city	local	co	no2	o3	pm10	pm25	so2
1	Jaipur	2020-10-26T05:15:00+05:30	1140	178.80	6.36	134.05	148.60	8.90
2	Jaipur	2020-10-26T06:00:00+05:30	1110	193.75	6.88	147.00	147.60	8.72
3	Jaipur	2020-10-26T06:30:00+05:30	980	178.33	5.16	147.60	146.77	10.10
4	Jaipur	2020-10-26T07:15:00+05:30	1200	166.94	9.96	155.48	105.38	11.25
5	Jaipur	2020-10-26T07:45:00+05:30	950	128.73	17.35	155.57	105.38	12.27
6	Jaipur	2020-10-26T09:00:00+05:30	1190	83.22	37.07	218.41	82.74	9.17
7	Jaipur	2020-10-26T09:30:00+05:30	1330	93.04	50.09	220.96	82.05	14.59
8	Jaipur	2020-10-26T12:15:00+05:30	9150	3.92	82.38	0.00	0.00	84.76
9	Jaipur	2020-10-26T19:00:00+05:30	2530	109.89	16.46	142.34	30.64	8.89
10	Jaipur	2020-10-26T19:30:00+05:30	2140	87.37	28.59	142.34	29.80	14.49

Fig 8. After Column wise conversion of Pollutants

Model Building:

In this section, we will be discussing supervised learning algorithms used to build the model on a selected dataset. We will be looking at 3 Regression and 2 classification models.

1. Air Quality Dataset of Different Indian Cities:

In this Dataset, the dependent variable is Continuous type and all of the independent variables are also Continuous type that's why we choose to implement Partial Least Square Regression Principal Component Regression. Both have significant similarities and differences. Both the techniques focus on dimension reduction while forecasting but Partial Least Square Regression is Supervised methodology while Principal Component Regression is Unsupervised methodology.

*A. **Principal Component Regression:***

One of the Popular dimension reduction techniques Principal Component Analysis in which we extract a low dimensional set of features from a large number of attributes. This approach is used in Principal Component Regression as well. In this methodology, PCR focuses on deriving the first M Principal Components and after this process, it uses these components as the Independent Variable in linear regression model which is fit using least squares. The important point of discussion is the Optimal Value of M to consider during the Regression process, as the number of Components increases the Bias in the model decreases at the same time variance increases. Each obtained component in the process is a linear combination of all the original variables. One of the requirements of the PCR is to standardize each variable. Which brings all the variables on the same scale. It is essential because if the Variables are not standardizing the variables with Large Variance will have significant involvement in derived components. If all the variables are having the same unit then it is not necessary to standardize the variables [11].

*B. **Partial Least Squares Regression:***

In the previous methodology, we discussed the unsupervised approach to dimension reduction using PCR as the response variable Y is not considered while obtaining components. This becomes the Drawback for PCR as directions that is a good fit for predictor variables not necessarily be a good fit for predicting response. Partial Least Squares Regression uses a similar approach to dimension reduction but it takes Predictor as well response variable in extracting components. Now we will look at how PLS computes the direction of the first component [11].

Here Z_1 is the first direction

P is Standardize Predictors

φ_{j1} is the coefficient from simple linear regression of Y and X_j

X_j is Original Predictor

Y is the Dependent /Response Variable.

Here the Highest weight is assigned to the predictor which is strongly related to the response variable.

To Find out the second direction of the PLS we first compute the residual between Z_1. These residuals are information which the first direction couldn't able to explain. Then we compute Z_2 in the same way as Z_1. This procedure is repeated M times and in the end this Z_1, Z_2,…..Z_m, are then regressed to predict Y using Least Squares[11].

Tune Length: It is a number of levels used by each tuning parameter. In order to Tune the model automatically, we specify tune length. Which will generate the number of values of the response variable and out of that optimal value would be selected.

Leave one Out Cross-Validation: For resampling purpose, we have selected Leave one Out Cross-Validation. The advantage of using such a method is it considers all the training data for model building.

2. Station Wise Data of Indian Cities:

The dependent variable in Station Wise Data of Indian Cities is of categorical type. So, we have implemented two classification algorithms to predict the class of AQI. It is essential to classify the AQI indexes to correct categories as class names will be the easiest way for the layman to understand the Quality of Air.

C. K Nearest Neighbor:

This is a classification algorithm that is supervised in nature. It classifies unlabeled data by using similar examples of data and their labels. This is implemented using a simple concept of Nearest Neighbor where for every record in the test data set , KNN identifies k records in nearly similar training data. The class of Majority k neighbors is then assigned to an unlabeled test record.

Algorithm:

KNN(D,d,k)

here D is the Training Data,

d is the Testing Data,

k is the number of records from D to select.

Step 1: Calculate the distance between d and Every record in D.

Step 2: Select k records from D which are nearest to d as per the calculated distance. This will create a subset of D; we call it P.

Step 3: Assign class of most frequently occurred label or Majority class label od P to d.

Here distances can be calculated in multiple ways such as Euclidian Distance, Manhattan Distance, Minkowski Distance and Hamming Distance.

There are multiple rules to select the appropriate value of k, ideally, it lies between 3-10. One of the rules is to use k as the square root of the total number of observations.

Results obtained can be well evaluated using Confusion Matrix, ROC Curve, AUC Value.

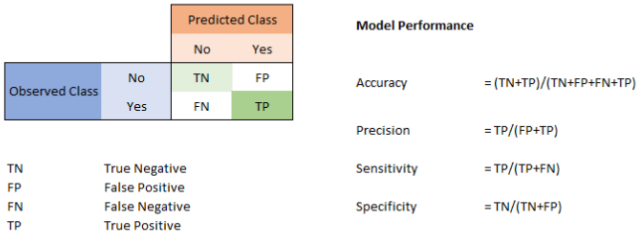

Fig 9. Confusion Matrix

Confusion Matrix: This is a tabular representation of how the model has classified the class labels. This helps us evaluate the model on various aspects, such as Accuracy, Precision, Sensitivity, Specificity.

Roc Curve: This is used to depict the performance of the classification model at every level of classification. This plots the graph of Specificity vs Sensitivity.

It not easy to interpret and compare the graphs that's why we use AUC Value.

AUC Value: This is Area Under the Curve value, which ranges from 0 to 1. Here, we look for a value close to 1. For two different looking graphs the AUC Value could be similar.

D. Multinomial Logistic Regression:

This is an extension to the Binomial Logistic Regression where labels are Binary Categorical values such as 0 and 1, True and False. Multinomial Logistic Regression is used when the Dependent Variable is having multiple classes.

$$E(y) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}$$

Fig 10. Logistic Regression Equation

Here E(y) is the Probability of Logistic Regression, in which y could be coded as 0 or 1 when implementing Binary Logistic Regression, in the case of multiple regression it could take multiple values. The value of E(y) indicates the Probability that y=1 given a particular set of values for X_1, X_2, ……, X_p.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

Fig 11. Logit Transformation

Here Left-Hand side is called Log-Odds, if there is a one-unit change in X_1 there will be a change in log odds by β_1 [12].

The results of the Logistic Regression will be evaluated by the Confusion Matrix.

3. Hourly Data of Indian Cities:

E. *Multiple Linear Regression:*

This uses a similar approach of regression as Linear Regression only difference is the number of predictors used to build the model.

Below is the Regression equation of Multiple Linear Regression.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

Here, Y is a dependent Variable,

X1,X2,…..Xp are independent variables,

β_0 is Y – Intercept (Constant Term)

$\beta_1, \ldots\ldots, \beta_p$ are regression (Slope) coefficients for each independent variable.

ε is the Error Term (Residuals).

IV. Evaluation

In the above section, we looked at the 5 different models, now we will be looking at evaluation metrics used to evaluate and compare these model performances. There are numerous methods by which we can accomplish this. It is dependent on what type of model such as (Regression, Classification, Dimension Reduction) we have implemented. Some metrics are model specific as well.

A. *Principal Component and Partial Least Squares Regression*

As these Models are Regression type models along with dimension reduction feature, we have selected R Square, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), RMSEP (Root Mean Square Error of Prediction), % of Variability in predictor and Outcome Variable described by Number of Components.

```
> summary(modelplsfit$finalModel)
Data:   X dimension: 8086 12
        Y dimension: 8086 1
Fit method: oscorespls
Number of components considered: 11
TRAINING: % variance explained
         1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
X          37.17    47.48    55.45    67.31    73.31    78.45    83.59    86.74
.outcome   71.11    84.72    88.50    88.98    89.18    89.19    89.20    89.20
         9 comps  10 comps  11 comps
X          90.62     96.01     98.86
.outcome   89.21     89.21     89.21
```

Fig 12 Model Summary PLS

Here we can see the percentage of variance explained by each component. X denotes the percentage of variability from predictors captured by each component and outcome denotes the percentage of variability from outcome variables captured by components. The decision to select the number of components is dependent on the goal of the analysis. Ideally, the component which is explaining more variability from both types of variables and Less number of components used to explain the variability. From Fig 12. we can interpret that model with 11 Components is an ideal choice but the number of dimensions are not reduced, if we look at the 6th Component the which is explaining 78.45% and 89.19% variability in both Predictors and Outcome variables is a better selection as it explains optimal variability as well as significantly reduced the number of variables.

```
Number of components considered: 11
TRAINING: % variance explained
         1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
X          37.67    53.15    64.65    72.86    79.78    85.33    90.46    93.74
.outcome   62.14    63.11    71.33    71.46    71.51    79.15    80.10    87.82
         9 comps  10 comps  11 comps
X          96.73     98.93     99.56
.outcome   88.33     89.17     89.18
```

Fig 13. Model Summary PCR

```
TRAINING: % variance explained
     1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps
X      76.37    85.75    89.57     91.3    93.00    94.12    95.67    96.67    98.12
AQI    43.16    65.83    75.08     83.1    87.25    88.51    88.85    89.16    89.20
     10 comps  11 comps  12 comps
X       99.76     99.89    100.00
AQI     89.20     89.21     89.21
```

Fig 14. Model Summary PLS Leave One Out CV

```
TRAINING: % variance explained
     1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 comps
X      76.89    86.38    90.11    92.61    94.92    96.88    98.14    99.16    99.57     99.77
AQI    37.64    54.97    63.59    67.16    68.07    68.23    68.95    70.67    82.90     89.07
     11 comps  12 comps
X       99.93    100.00
AQI     89.19     89.21
```

Fig 15. Model Summary PCR Leave One Out CV

From Fig 12,13,14,15. we can say that the PLS Leave One Out Model is better at explaining the variability of both types of variables. If We compare the 5th component of all the models, PLS Leave One Out Model is explaining more variance than other models.

Now we will look at their prediction accuracy and Error Metrics.

RMSE: It is the root mean square between Actual Values and Predictions. It should be as low as possible.

MAE: It is the mean absolute error between True and Actual Values. This also needs to be as low as possible.

MAPE: This indicates to us that how accurately the model has done the prediction. Minimum MAPE value is desirable.

R Square: The amount of Variance of the dependent variable explained by independent variables is called as R Square. This should be close to 1, the higher the better [14].

From Fig 16. we can interpret that the PLS model has improved its accuracy in testing than training. The RMSE is reduced but MAE and MAPE value increased also there is an improvement in R Square value.

PLS Model Evaluation								
Method	PLS Training Evaluation Results				PLS Testing Evaluation Results			
	RMSE	R Sqare	MAE	MAPE	RMSE	R Sqare	MAE	MAPE
Caret Package	52.28	0.89	24.63	NA	50.74	0.91	25.63	NA
Mlmetrics Package	52.28	NA	24.63	0.16	50.74	NA	25.63	0.16
Manual Computation	NA	NA	NA	15.59	NA	NA	NA	15.82

Fig 16. PLS Evaluation Results

PCR Model Evaluation								
Method	PLS Training Evaluation Results				PLS Testing Evaluation Results			
	RMSE	R Sqare	MAE	MAPE	RMSE	R Sqare	MAE	MAPE
Caret Package	52.34	0.89	24.79	NA	50.91	0.91	25.91	NA
Mlmetrics Package	52.34	NA	24.79	0.16	50.91	NA	25.91	0.16
Manual Computation	NA	NA	NA	15.75	NA	NA	NA	16.08

Fig 17. PCR Evaluation Results

PCR model from fig has also improved in testing which means that both PLS and PCR not overfitting the data. When compare thembased on metrics mentioned above both the model seemto similar accuracy and error rate.

PLS Leave One Out Evaluation Results					
PLS Training Evaluation Results			PLS Testing Evaluation Results		
RMSE	MAE	MAPE	RMSE	MAE	MAPE
68.48	32.11	0.20	70.59	33.49	0.20

Fig 18. PLS Leave One Out Evaluation Results

From fig 18. we can see that RMSE and MAE values have increased in the testing phase which indicates that data is overfitted. In comparison to PLS with Leave One Out Cross-validation PLS and PCR models in fig are performing better.

PCR Leave One Out Evaluation Results					
PLS Training Evaluation Results			PLS Testing Evaluation Results		
RMSE	MAE	MAPE	RMSE	MAE	MAPE
85.98	41.40	0.25	91.26	43.01	0.25

Fig 19. PCR Leave One Out Evaluation Results

PLS model has performed best when considering Accuracy as a measure.

B. *K Nearest Neighbor and Multinomial Logistic Regression:*

These two algorithms are classification algorithms therefore evaluation methods used here will be different. We have used Confusion Matrix, ROC Curve and AUC Value as evaluation methods.

	KNN with Cross Validation & Tune Length 12				
k	5	10	15	20	25
Accuracy	83.93	82.96	83.11	81.69	81.54

Fig 20. KNN Test Data

	KNN with Repeated Cv Cross Validation and Tune length 10									
k	5	7	9	11	13	15	17	19	21	23
Accuracy	83.02	82.28	82.34	82.14	82.27	82.43	82.52	82.49	82.29	82.27

Fig 21. KNN with Repeated CV & TL 10 Training Data

fig 20,21. represents the accuracy of KNN and KNN Repeated CV and Tune Length 10 with respect to two Different K Values. Here, as the value of K is increasing accuracy is decreasing. Both models give the best accuracy at K=5. But we can not compare both the models based on Fig as one of them is for Training Data and another one is for Test Data.

```
> summary(test_set$AQI_Bucket)
      Good   Moderate       Poor Satisfactory     Severe  Very Poor
       188       1340        568        742        461        585
```

Fig 22. Count of Each Label in Test Data

test_set$AQI_Bucket \ y_pred	Good	Moderate	Poor	Satisfactory	Severe	Very Poor	Row Total
Good	135 0.718 0.865 0.035	2 0.011 0.001 0.001	0 0.000 0.000 0.000	51 0.271 0.067 0.013	0 0.000 0.000 0.000	0 0.000 0.000 0.000	188 0.048
Moderate	0 0.000 0.000 0.000	1200 0.896 0.863 0.309	49 0.037 0.087 0.013	91 0.068 0.119 0.023	0 0.000 0.000 0.000	0 0.000 0.000 0.000	1340 0.345
Poor	0 0.000 0.000 0.000	88 0.155 0.063 0.023	434 0.764 0.772 0.112	0 0.000 0.000 0.000	0 0.000 0.000 0.000	46 0.081 0.079 0.012	568 0.146
Satisfactory	21 0.028 0.135 0.005	98 0.132 0.071 0.025	0 0.000 0.000 0.000	623 0.840 0.814 0.160	0 0.000 0.000 0.000	0 0.000 0.000 0.000	742 0.191
Severe	0 0.000 0.000 0.000	0 0.000 0.000 0.000	2 0.004 0.004 0.001	0 0.000 0.000 0.000	395 0.857 0.923 0.102	64 0.139 0.110 0.016	461 0.119
Very Poor	0 0.000 0.000 0.000	2 0.003 0.001 0.001	77 0.132 0.137 0.020	0 0.000 0.000 0.000	33 0.056 0.077 0.008	473 0.809 0.811 0.122	585 0.151
Column Total	156 0.040	1390 0.358	562 0.145	765 0.197	428 0.110	583 0.150	3884

Fig 23. Confusion Matrix KNN K=5 Test Data

test_set$AQI_Bucket \ knnPredictAQI	Good	Moderate	Poor	Satisfactory	Severe	Very Poor	Row Total
Good	135 0.718 0.865 0.035	2 0.011 0.001 0.001	0 0.000 0.000 0.000	51 0.271 0.067 0.013	0 0.000 0.000 0.000	0 0.000 0.000 0.000	188 0.048
Moderate	0 0.000 0.000 0.000	1200 0.896 0.862 0.309	49 0.037 0.087 0.013	91 0.068 0.119 0.023	0 0.000 0.000 0.000	0 0.000 0.000 0.000	1340 0.345
Poor	0 0.000 0.000 0.000	88 0.155 0.063 0.023	439 0.773 0.777 0.113	0 0.000 0.000 0.000	0 0.000 0.000 0.000	41 0.072 0.070 0.011	568 0.146
Satisfactory	21 0.028 0.135 0.005	100 0.135 0.072 0.026	0 0.000 0.000 0.000	621 0.837 0.814 0.160	0 0.000 0.000 0.000	0 0.000 0.000 0.000	742 0.191
Severe	0 0.000 0.000 0.000	0 0.000 0.000 0.000	2 0.004 0.004 0.001	0 0.000 0.000 0.000	395 0.857 0.927 0.102	64 0.139 0.110 0.016	461 0.119
Very Poor	0 0.000 0.000 0.000	2 0.003 0.001 0.001	75 0.128 0.133 0.019	0 0.000 0.000 0.000	31 0.053 0.073 0.008	477 0.815 0.820 0.123	585 0.151
Column Total	156 0.040	1392 0.358	565 0.145	763 0.196	426 0.110	582 0.150	3884

Fig 24, Confusion Matrix KNN, K=5 Repeated CV & TL 10Test Data

Fig 22,23,24. Provides us the details of how Models have classified data into each class. This is important in evaluation as we can find out number of instances on which model has incorrectly predicted the class.

```
Overall Statistics

               Accuracy : 0.8393
                 95% CI : (0.8274, 0.8508)
    No Information Rate : 0.345
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7941

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Good Class: Moderate Class: Poor
Sensitivity              0.71809          0.8955       0.7641
Specificity              0.99432          0.9253       0.9614
Pos Pred Value           0.86538          0.8633       0.7722
Neg Pred Value           0.98578          0.9439       0.9597
Prevalence               0.04840          0.3450       0.1462
Detection Rate           0.03476          0.3090       0.1117
Detection Prevalence     0.04016          0.3579       0.1447
Balanced Accuracy        0.85620          0.9104       0.8627
                     Class: Satisfactory Class: Severe Class: Very Poor
Sensitivity                       0.8396        0.8568            0.8085
Specificity                       0.9548        0.9904            0.9667
Pos Pred Value                    0.8144        0.9229            0.8113
Neg Pred Value                    0.9618        0.9809            0.9661
Prevalence                        0.1910        0.1187            0.1506
Detection Rate                    0.1604        0.1017            0.1218
Detection Prevalence              0.1970        0.1102            0.1501
Balanced Accuracy                 0.8972        0.9236            0.8876
```

Fig 25. Overall Statistics KNN, K=5 Test Data

```
Overall Statistics

               Accuracy : 0.8411
                 95% CI : (0.8293, 0.8525)
    No Information Rate : 0.345
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7963

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Good Class: Moderate Class: Poor Class: Satisfactory
Sensitivity              0.71809          0.8955      0.7729               0.8369
Specificity              0.99432          0.9245      0.9620               0.9548
Pos Pred Value           0.86538          0.8621      0.7770               0.8139
Neg Pred Value           0.98578          0.9438      0.9611               0.9612
Prevalence               0.04840          0.3450      0.1462               0.1910
Detection Rate           0.03476          0.3090      0.1130               0.1599
Detection Prevalence     0.04016          0.3584      0.1455               0.1964
Balanced Accuracy        0.85620          0.9100      0.8674               0.8959
                     Class: Severe Class: Very Poor
Sensitivity                 0.8568          0.8154
Specificity                 0.9909          0.9682
Pos Pred Value              0.9272          0.8196
Neg Pred Value              0.9809          0.9673
Prevalence                  0.1187          0.1506
Detection Rate              0.1017          0.1228
Detection Prevalence        0.1097          0.1498
Balanced Accuracy           0.9239          0.8918
```

Fig 26. Overall Statistics KNN, K=5 with Repeated CV & TL 10

From Fig 21,26. we can find out that the model with repeated CV has improved from Training Data. After comparing two models the repeated CV model provided slightly better accuracy.

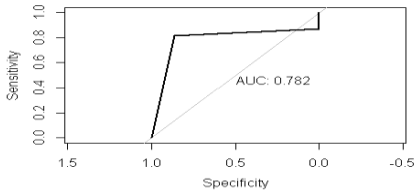

Fig 27. ROC curve KNN K= 5 Test Data

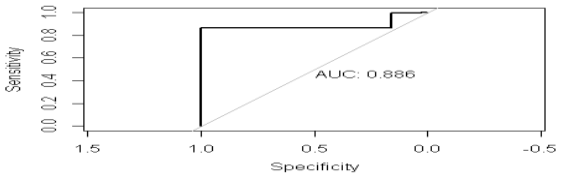

Fig 28. ROC Curve KNN, K=5 Repeated CV & TL 10Test Data

From Fig we can make out that the Specificity of the models remains constant at 1.0 while Sensitivity Starts to increase when it reaches 0.8 level, it remains constant and Specificity starts to decrease. Here AUC value of KNN with repeated CV is better in comparison. We are looking for AUC as high as possible which ranges from Zero to One.

test_set$AQI_Bucket	AQI_TEST_PREDICTED_LOGISTICS Good	Moderate	Poor	Satisfactory	Severe	Very Poor	Row Total
Good	129 0.686 0.838 0.033	1 0.005 0.001 0.000	0 0.000 0.000 0.000	58 0.309 0.079 0.015	0 0.000 0.000 0.000	0 0.000 0.000 0.000	188 0.048
Moderate	0 0.000 0.000 0.000	1122 0.837 0.780 0.289	90 0.067 0.165 0.023	119 0.089 0.162 0.031	1 0.001 0.002 0.000	8 0.006 0.014 0.002	1340 0.345
Poor	0 0.000 0.000 0.000	143 0.252 0.099 0.037	347 0.611 0.638 0.089	3 0.005 0.004 0.001	3 0.005 0.007 0.001	72 0.127 0.122 0.019	568 0.146
Satisfactory	25 0.034 0.162 0.006	161 0.217 0.112 0.041	0 0.000 0.000 0.000	556 0.749 0.755 0.143	0 0.000 0.000 0.000	0 0.000 0.000 0.000	742 0.191
Severe	0 0.000 0.000 0.000	0 0.000 0.000 0.000	2 0.004 0.004 0.001	0 0.000 0.000 0.000	378 0.820 0.894 0.097	81 0.176 0.138 0.021	461 0.119
Very Poor	0 0.000 0.000 0.000	11 0.019 0.008 0.003	105 0.179 0.193 0.027	0 0.000 0.000 0.000	41 0.070 0.097 0.011	428 0.732 0.727 0.110	585 0.151
Column Total	154 0.040	1438 0.370	544 0.140	736 0.189	423 0.109	589 0.152	3884

Fig 29. Confusion Matrix Logistic Regression Model

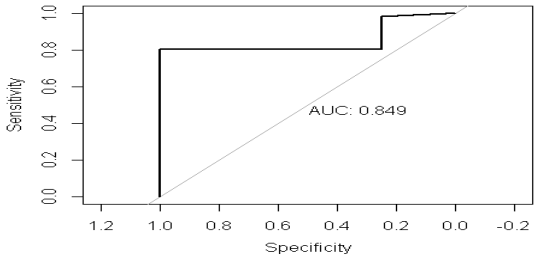

Fig 30. ROC Curve Logistic Regression Model

Training	Test
75.59	76.21

Table 1. Multinomial Logistic Regression Accuracy

CoxSnell	Nagelkerke	McFadden
0.8795418	0.913273	0.6416193

Table 2. R Square Values for Multinomial Logistic Regression

There could be an argument over which R^2 to consider as specified in Table 2. Looking at the R Square values from all three methods, we can say that the Logistic Regression Model is doing good in terms of class prediction

After comparing and evaluating all the classification models we can say that KNN with Repeated CV and Tune Length 10 has provided better accuracy and AUC.

C. Multiple Linear Regression

Multiple Linear Regression							
Training Evaluation Results				Test Evaluation Results			
RMSE	R Sqare	MAE	MAPE	RMSE	R Sqare	MAE	MAPE
67.88384	0.788	42.39624	16.1	69.42095	0.776	43.44164	16.3

Fig 31. MLR Evaluation Metrics

Multiple Linear Model here seems to be overfitting as the errors in the test evaluation has increased. Model is predicting with good R^2 value which could be improved further. Initially 6 predictors were used for model building but after looking at Model Summary we found that two of the predictors are insignificant. We removed those variables and re run the model but even after removing those variables model's performance hasn't increased significantly.

```
> summary(mlr_aqi)

Call:
lm(formula = AQI ~ pm25aqi + coaqi + no2aqi + pm10aqi + so2aqi +
    o3aqi, data = aqi_train)

Residuals:
    Min      1Q  Median      3Q     Max
-521.70  -36.19   -7.92   23.14 1926.30

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.594179   1.847985  12.226   <2e-16 ***
pm25aqi      0.419577   0.006170  68.000   <2e-16 ***
coaqi        0.419824   0.007520  55.825   <2e-16 ***
no2aqi      -0.167021   0.017184  -9.720   <2e-16 ***
pm10aqi      0.478621   0.005909  81.002   <2e-16 ***
so2aqi      -0.069326   0.036366  -1.906   0.0566 .
o3aqi        0.006097   0.014694   0.415   0.6782
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 67.9 on 9228 degrees of freedom
Multiple R-squared:  0.7881,    Adjusted R-squared:  0.788
F-statistic:  5721 on 6 and 9228 DF,  p-value: < 2.2e-16
```

Fig 32. MLR with 6 Predictors

From Fig 32. we can see that P values for SO2AQI and 03AQI are greater than 0.05. Value of R Square and Adjusted R Square are 0.7881 and 0.788 respectively.

```
> summary(mlr_aqi_1)

Call:
lm(formula = AQI ~ pm25aqi + coaqi + pm10aqi + no2aqi, data = aqi_train)

Residuals:
    Min      1Q  Median      3Q     Max
-749.15  -36.36   -7.55   22.92 1987.48

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 21.986338   1.727033   12.73   <2e-16 ***
pm25aqi      0.432688   0.006062   71.38   <2e-16 ***
coaqi        0.415656   0.007545   55.09   <2e-16 ***
pm10aqi      0.463533   0.005942   78.01   <2e-16 ***
no2aqi      -0.154251   0.016970   -9.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 68.24 on 9230 degrees of freedom
Multiple R-squared:  0.7899,    Adjusted R-squared:  0.7898
F-statistic:  8674 on 4 and 9230 DF,  p-value: < 2.2e-16
```

Fig 33. MLR with 6 Predictors

Here we can see that after removing the insignificant variables ideally the values of R Square and Adjusted R Square should have increased but this not the case here.

V. CONCLUSIONS AND FUTURE WORK

In this research, we analyzed and evaluated Five different models on 3 different datasets. All the three datasets were related but they were using different nature, scale and approach for data representation. Due to which models applied on the particular dataset are only compared for evaluation. From the Air Quality Dataset of Different Indian Cities, we were looking to predict the quantitative AQI values along with dimension reduction techniques, in which the PLS model with Leave One Out Cross Validation was best at dimension reduction considering only 5th component from all the models. In terms of accuracy PLS model was best with Lowest RMSE. The objective of implementing models on Station Wise Data of Indian Cities was to correctly classify and predict the AQI Labels, in which the KNN Model with Repeated CV and Tune Length 10 performed best in terms of accuracy and AUC. The aim for Hourly Data of Indian Cities was to find significant factors that could lead to accurate AQI prediction from a range of Pollutants. Out of Six Pollutants, only Four were contributing significantly in model prediction.

The analysis and results obtained from this research encourage the possibility of further work, in order to improve the model’s performances and research goal.

The data analyzed in this research doesn’t include weather statistics. AQI values are also dependent on surrounding weather due to which inclusion of such data could lead to more precise AQI prediction.

The data of AQI can be viewed as a Time Series, which leads to consideration of Time Series Algorithms for analysis.

REFERENCES

[1] 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," 2017 International Conference on Trends in Electronics and Informatics (ICEI), Tirunelveli, 2017, pp. 248-254, doi: 10.1109/ICOEI.2017.8300926.

[2] K. Nandini and G. Fathima, "Urban Air Quality Analysis and Prediction Using Machine Learning," 2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE), Bangalore, India, 2019, pp. 98-102, doi: 10.1109/ICATIECE45860.2019.9063845.

[3] S. Srikamdee and J. Onpans, "Forecasting Daily Air Quality in Northern Thailand Using Machine Learning Techniques," 2019 4th International Conference on Information Technology (InCIT), Bangkok, Thailand, 2019, pp. 259-263, doi: 10.1109/INCIT.2019.8912072.

[4] T. M. Amado and J. C. Dela Cruz, "Development of Machine Learning-based Predictive Models for Air Quality Monitoring and Characterization," TENCON 2018 - 2018 IEEE Region 10 Conference, Jeju, Korea (South), 2018, pp. 0668-0672, doi: 10.1109/TENCON.2018.8650518.

[5] N. Tomar, D. Patel and A. Jain, "Air Quality Index Forecasting using Auto-regression Models," 2020 IEEE International Students' Conference on Electrical,Electronics and Computer Science (SCEECS), Bhopal, India, 2020, pp. 1-5, doi: 10.1109/SCEECS48394.2020.216.

[6] D. Ao, Z. Cui and D. Gu, "Hybrid model of Air Quality Prediction Using K-Means Clustering and Deep Neural Network," 2019 Chinese Control Conference (CCC), Guangzhou, China, 2019, pp. 8416-8421, doi: 10.23919/ChiCC.2019.8865861.

[7] R. K. Grace, K. Aishvarya S., B. Monisha and A. Kaarthik, "Analysis and Visualization of Air Quality Using Real Time Pollutant Data," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 34-38, doi: 10.1109/ICACCS48705.2020.9074283.

[8] S. Mahanta, T. Ramakrishnudu, R. R. Jha and N. Tailor, "Urban Air Quality Prediction Using Regression Analysis," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, 2019, pp. 1118-1123, doi: 10.1109/TENCON.2019.8929517.

[9] U. Mahalingam, K. Elangovan, H. Dobhal, C. Valliappa, S. Shrestha and G. Kedam, "A Machine Learning Model for Air Quality Prediction for Smart Cities," 2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET), Chennai, India, 2019, pp. 452-457, doi: 10.1109/WiSPNET45539.2019.9032734.

[10] Y. Su, "Prediction of air quality based on Gradient Boosting Machine Method," 2020 International Conference on Big Data and Informatization Education (ICBDIE), Zhangjiajie, China, 2020, pp. 395-397, doi: 10.1109/ICBDIE50010.2020.00099.

[11] James G., Witten D., Hastie T., Tibshirani R. (2013). An Introduction to Statistical Learning. Springer.

[12] Pallant, Julie. Ebook: SPSS Survival Manual: a Step by Step Guide to Data Analysis Using IBM SPSS, McGraw-Hill Education, 2020

[13] Olson, D. L., & Delen, D. (2008). *Advanced data mining techniques*. Chap 9. Springer Science & Business Media.