International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 15 (2018) pp. 12202-12208

A Forecasting model to Predict the Health Impacts in Metropolitan Cities using Data Mining Techniques and Tools

Asha N[*], Dr. M P Indira Gandhi[**]

[*]*Research Scholar, Department of Computer Science, Mother Teresa University, Kodaikanal, India.*

[**]*Department of Computer Science, Mother Teresa University, Kodaikanal, India.*

Abstract

The colossal increase in continuous discharge of air pollutants by vehicles and factories in metropolitan areas are monitored by the monitoring stations in order to bring awareness among mankind to protect environment and human health. The Data collected at monitoring stations is enormously increasing on a day to day basis, so useful data should be mined to get the required information in order to take proper decisions like, which pollutant is mostly affecting the city and causing serious illness on health of mankind. Mining the entailed information from large amount of data requires data mining techniques to analyze the data patterns so as to predict the useful patterns. The nominal data is considered for prediction and Classification techniques such as Decision tree and Naïve Bayes Techniques, k Nearest Neighbor Classifier are used to predict the low, high and moderately affected areas in metropolitan city using Weka as a Data Mining Tool. A Forecasting Model is suggested, and predicts the air quality index for a given pollutant and observed air quality index and also predicts the health impacts caused by pollutants.

Keywords: Nominal data, Decision Tree, Naive bayes, k Nearest Neighbor Classifier, Weka

INTRODUCTION

Data Mining is extracting useful knowledge from a large database. A small effort is made in this paper to know how nominal data mining algorithms works on a pollution data set of a metropolitan city. As there is a massive increase in air pollution and every day the data that is being collected at the monitoring station is very huge. But the huge data is not properly analyzed to know the amount of damage causing to human health. So this paper emphasizes on predicting the health impacts caused due to Air pollution. The Techniques available in data mining to predict nominal data are Naive Bayes, k nearest Classifier and Decision tree Algorithms. These techniques classify the class label for a given training data set and also these models are very essential in defining the Predictive accuracy, Speed, Robustness, and Interpretability. Several Data Mining tools are used such as Weka, R Programming, Rapid Miner for large data bases.

CASE SUTY

Many Metropolitan cities are prone to air pollution. One such city is Bangalore, its not only known for its popularity in terms of IT, BT, garden city, etc..But also highly polluted with too much of vehicle population and emission caused by industries.

The KSPCB (Karnataka state pollution control board) is monitoring ambient air quality (AAQ) of Bangalore city at 15 locations using manual equipments under National Ambient Air Quality Monitoring Programme (NAMP) covering Industrial Area, Mixed Urban Area and Sensitive Area. The monitoring stations store air pollutants on a daily basis. The average annual archived data is collected from KSPCB website which is composed of PM10, SO2 and NO2 and air quality index of each pollutant from 2011-2015 are used for analysis and prediction [1].

PREVIOUS WORK

The Data mining techniques, used to predict the air quality index for any given pollutant by constructing a model. Also it is observed that maximum pollutant that is affecting the Bangalore city is Pm10 with a maximum air quality index which is a major pollutant released by vehicles by using the Regression Techniques such as Linear Regression, Multi Linear Regression, Multilayer Perceptron Model(MLP).. The air quality index (AQI) standards are classified as Moderate, Satisfactory, Good and Poor quality of air that is not permissible for a good breathing environment .The prediction provides decision making capability to the government to take proper action against the alarming rate of increase in vehicular population that has led to serious increase in PM10 concentration. The MLP is proved to be the best technique to predict Air quality Index. This work paves way for further analysis of predictions made by MLP that can be classified into Moderate, Satisfactory, Good and Poor which can be given as an input to Decision tree ,Naïve Bayes Algorithm and k nearest neighbor classifier to process nominal data and identify the poorly polluted area of Bangalore city using a data mining tool Weka [1].

The Flowchart 1 helps to understand the previous efforts made in predicting the annual AQI Index.

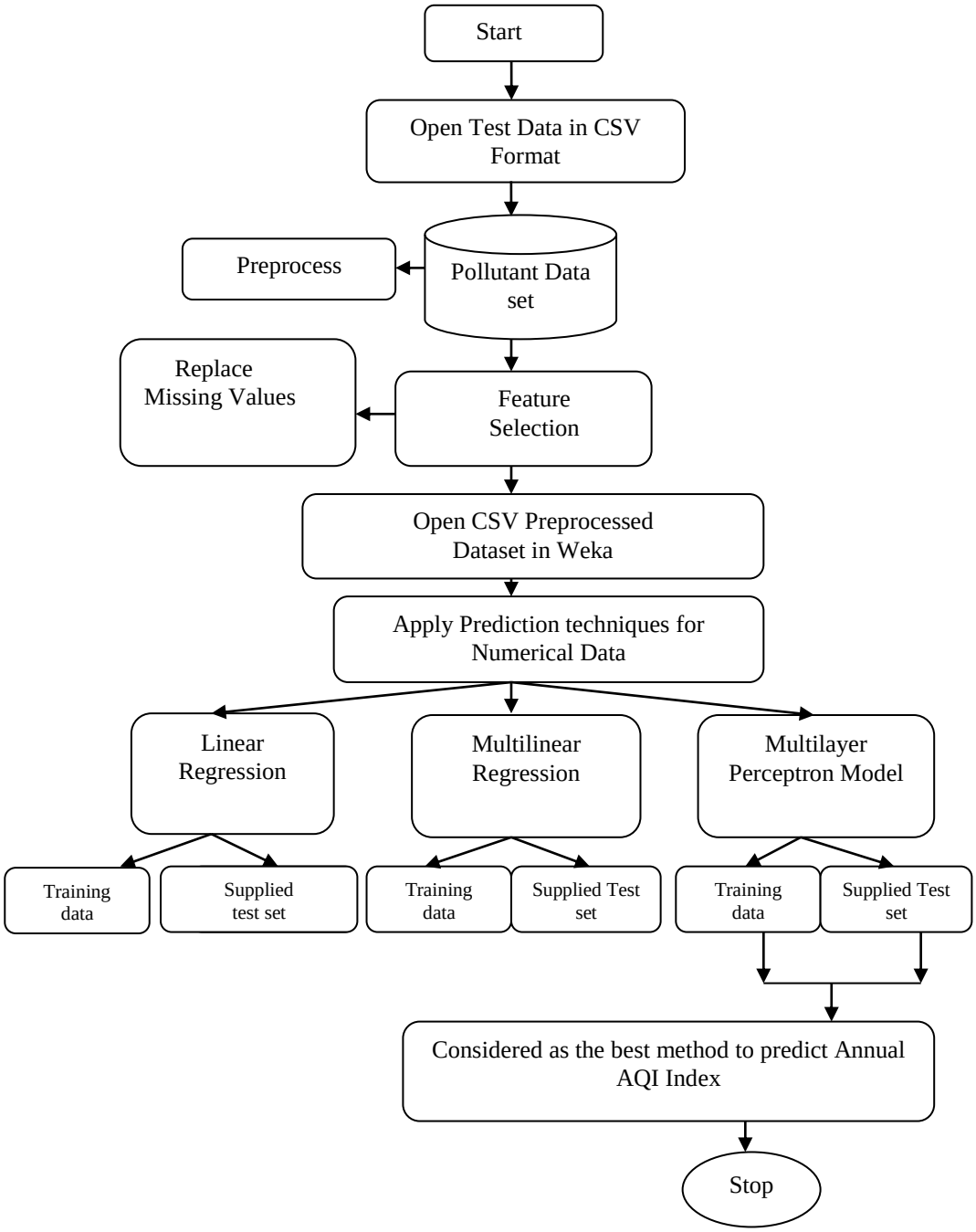

Figure 1: Prediction of AQI

PROPOSED FORECASTING MODEL

The Previous work has predicted air quality index using Numerical Predictive techniques of data mining. The techniques results with their predictions of air quality index and that technique with the less relative error or with more accurate predictions are considered. Since MLP was considered as the best predictive method to predict AQI with less relative error, this paper continued with the output obtained by MLP by classifying the health impacts based on Predicted Air Quality Index as per standards set by pollution control board which is as fallows in Table 1.

Table 1: Breathing Comforts set by KSPCB

Health Impacts	Range	Breathing Comforts
Good	0-50	Minimal Health Impact
Satisfactory	51-100	Minor Breathing Discomfort to Sensitive people
Moderate	101-200	Breathing discomfort to the people with lung, heart disease, children and older adults

Health Impacts	Range	Breathing Comforts
Poor	201-300	Breathing discomfort to people on prolonged exposure
Very Poor	301-401	Respiratory illness to the people on prolonged exposure
Severe	>401	Respiratory effects even on healthy people

The following code is used to classify the predicted air quality index into observed health impact as per the above standards.

IF(Predicted_air_QualityIndex<=50,"good",IF(Predicted_air_QualityIndex <=100,"satisfactory",IF(Predicted_air_QualityIndex>=101,"moderate",IF(Predicted_air_QualityIndex >=201,"poor")))).

So the output file is opened in Excel format and applied the above code to classify the predicted AQI into Good, Satisfactory, Moderate and Poor and named the column as Observed Health Impacts .An instance is as shown in Table 2

Table 2: Predicted AQI Index Classified into Observed Health Impacts

Observed AQI Index	Predicted AQI Index	Observed Health Impacts
181.5	182.7755	Moderate
211.9	213.0688	poor
50.8	49.49396	good
71.8	72.05328	satisfactory

The flowchart 2 discuses the present work. Since the data set was already preprocessed, now the work is enhanced in predicting health impacts using Decision tree, Naviee Bayes and K nearest classifier. The input file given to the flowchart 2 is an output of flowchart 1.

As discussed in Table 2, now the MLP output file as observed health impacts which was classified on Predicted AQI index. This file is opened in CSV format to predict health impacts which as shown in the following flowchart. Since the target variable is Observed Health Impacts and data is in Categorical form, we use above Classification Algorithms to predict the health impacts. A Data Mining tool called Weka is used for Prediction.

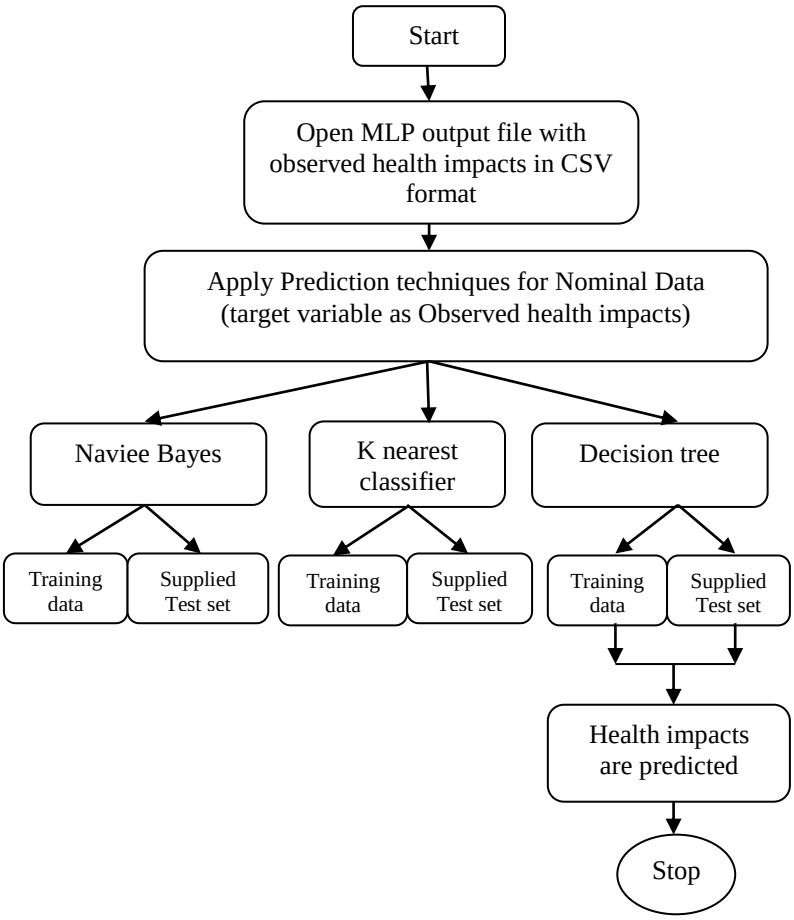

Figure 2: Prediction of Health Impacts

CLASSIFICATION TECHNIQUES

The following classification technique which works on Categorical data is considered for predicting the observed health impacts. The data classification processes are done in two steps:

1. Learning: Training data are analyzed by classification algorithm, here the class label Observed Health Impact is a target variable for which classification rules are defined.
2. Classification: Test data are used to estimate the accuracy of the classification rules, the classification is applied to new data set only if the accuracy is acceptable.

A. NAVIEE BAYES

A class Label called Observed health impacts are predicted using Bayesian classification. The data samples are described by the Year, Station, Pm10, Pm10 Index, So2, So2 Index,No2,No2 Index ,Observed AQI, Predicted AQI, Observed health Impacts .The unknown sample that is predicted is:

Predicted health impacts=(Predicted AQI<=50,"Predicted health impact=good", Predicted AQI<=100,"Predicted health impact=satisfactory", Predicted AQI>=101,"Predicted health impact=moderate", Predicted AQI>=201,"Predicted health impact=poor",)

So using Bayes Theorem,

$$P(C_i|X) = \frac{P(X|_{C_i})P(C_i)}{P(X)} \quad \text{----------→Equation 1}$$

Where, i=1, 2, 3….n

$P(X|C_i)P(C_i)$ The prior probability is computed based on training samples.

The algorithm is tested in two phases i.e., training data and test data.

B. K-nearest Neighbor Classifier

Nearest neighbor classifier are based on assumption. The training samples are defined by n-dimensional numeric attributes. Each sample describes a point in an n-dimensional numeric attributes. Hence training samples are stored in an n-dimensional space. In K-nearest neighbor classifier the given sample is searched in the pattern in space for the k training samples that are closest and nearest to the unknown sample. These K training samples are the k nearest neighbors of the unknown sample. These K training samples are the K nearest neighbors of the unknown samples. Closeness is defined in terms of Euclidian distance, where the Euclidean distance between two points.

x=(x1, x2, x3,….x_n)

y= (y1, y2, y3…..y_n) is

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(xi - yi)^2} \quad \text{----------→Equation 2}$$

The unknown sample is predicted health impacts are assigned the most common class among its k nearest neighbors. The unknown sample is assigned the class of the training sample that is closest to it in pattern space when k=1.

C. Decision Tree Induction

A decision tree is a tree like structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions. The top most nodes in a tree is the root node i.e. predicted air quality index. A typical decision tree in this problem represents the concept of predicting health impacts from the predicted AQI index as per the breathing standards set by pollution control board. To classify an Unknown sample, the attribute values of the sample are tested against the decision tree. A path is traced from the root to a leaf node that holds the class prediction for that sample. The tree for predicting health impacts is as fallows in Fig 1.

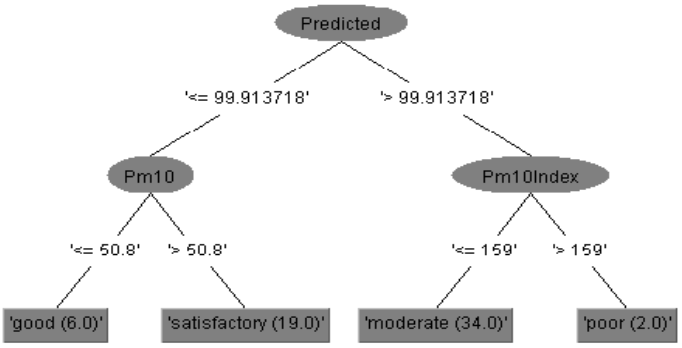

Figure 3: Tree structure of a Decision tree to predict Health Impacts

Picture 1 describes the attributes that are used to predict Health impacts.

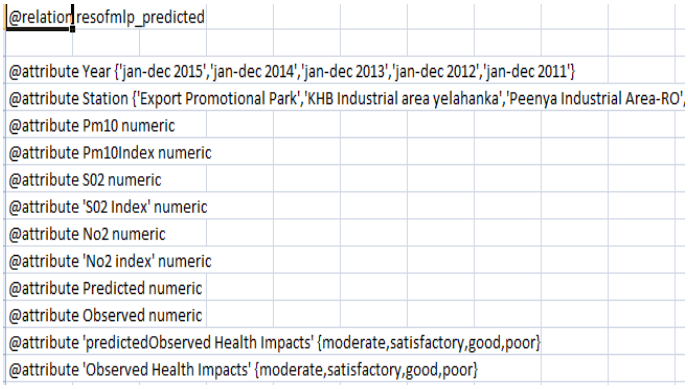

Figure 4: Attributes of MLP file

The following Picture 2 shows the output file in csv format opened in excel sheet shows that output that was obtained after executing the MLP file in Weka using Decision tree Algorithm and Obtained the Predicted Health Impacts which is exactly same as that of Observed Health Impacts.

```
@data
'jan-dec 2015','Export Promotional Park',181.5,154,4.9,6,21.6,27,182.775456,181.5,moderate,moderate
'jan-dec 2015','KHB Industrial area yelahanka',125.3,117,4.5,6,17,21,125.102423,125.3,moderate,moderate
'jan-dec 2015','Peenya Industrial Area-RO',134.7,123,5,6,20.6,26,133.488642,134.7,moderate,moderate
'jan-dec 2015','Peenya Gymkhana',127.3,118,11.7,15,26.9,34,125.861132,127.3,moderate,moderate
'jan-dec 2015','Yeshwanthpur police station',127.8,119,4.9,6,21.3,27,128.657283,127.8,moderate,moderate
'jan-dec 2015','Amco Batteries',136.6,124,5.1,6,19.9,25,134.207325,136.6,moderate,moderate
'jan-dec 2015','Banswadi Police Station',83.5,84,9.9,12,24.6,31,86.040928,83.5,satisfactory,satisfactory
'jan-dec 2015','central Silk Board',187.8,159,5.1,6,21.2,27,188.436121,187.8,moderate,moderate
'jan-dec 2015','DTDc House',149,133,4.7,6,18.8,24,150.181503,149,moderate,moderate
'jan-dec 2015','CAAQM City Railway Station',93.5,94,5.7,7,45.2,57,92.189681,93.5,satisfactory,satisfactory
'jan-dec 2015','CAAQM S G Halli',50.8,51,3.8,5,19.7,25,49.493957,50.8,good,good
'jan-dec 2015',Kajisonnenahalli,71.8,72,4.5,6,14.7,72,72.053281,71.8,satisfactory,satisfactory
'jan-dec 2015','Victoria Hospital',126.2,117,5,6,22.7,28,124.483526,126.2,moderate,moderate
'jan-dec 2015','Indira Gandhi Children Care',109.3,106,4.5,6,19.4,24,107.727331,109.3,moderate,moderate
'jan-dec 2015','Teri -Domlur',58.3,58,5.7,7,11.5,14,59.293749,58.3,satisfactory,satisfactory
'jan-dec 2014','Export Promotional Park',211.9,175,14.5,18,32.7,41,213.068833,211.9,poor,poor
```

Figure 5: Observations made after executing MLP file.

D. Comparitive Analysis of the Developed Models

The nominal values are predicted using the Navie bayes, k nearest classifier and Decision tree. The classification rules obtained are shown below in the table. The algorithms are analyzed based on predictive accuracy, speed, correctly classified instances and incorrectly classified instances as shown in Table 3.

PREDICTIVE ACCURACY OF THE ALGORITHMS IN TABLE FORMAT

The results of the algorithms with their respective predictive values are as shown. The tables below shows the attributes of air pollution of Bangalore city with their pollutant values and Average AQI index of observed and predictive values, and observed health impacts. The results show whether the predicted health impacts is matching the observed health impacts.

The analysis shows five tuples out of 61 instances. The results of the Navie bayes as discussed in table 4 and K nearest classifier as shown in table 5 shows certain incorrectly classified instances compared to Decision tree as shown in table 6. So compared to Navie bayes and k-nearest neighbor classifier, the Decision tree algorithm as given the best results, where the predicted health impacts is exactly same as that of observed health impacts,i.e the instances are correctly classified.

CONCLUSIONS

Therefore from the study of Comparative Analysis of tables, we could draw an inference that Decision tree algorithms are giving 100 percent accuracy and best results in terms of Predictive accuracy and time. The number of correctly classified instances is zero when compared to Naive Bayes and K-means Classifier. So we can conclude that Decision tree Algorithm gives best predictive accuracy for categorical data with respect to pollutant data set.

Also, there are two flaws in this work.

1. The data taken is very small as the data was collected from

KPCSB website. Since only the Annual Average data was provided from 2011-2015 in the website. So it is required to test the above model with huge data.

2. The second flaw is that air pollution not only depends on vehicle and industry emissions but also varies with temperature, humidity; wind. These environmental factors have to be taken into considerations to decide the breathing discomforts of Human health.

FUTURE WORK

The future work could be carried out in two steps

1. As numerical predictive algorithms like Linear Regression, Multi linear Regression, Multilayer perceptron Model was used to predict AQI and categorical predictive algorithms like Naive bayes, k means classifier and Decision tree to predict health impacts. This has led to the understanding of collection of huge amount of data with environmental factors causing air pollution and use clustering methods, so that any input to the cluster algorithms should correctly classify to the corresponding clusters.

2. Improvising the efficiency and accuracy of MLP to predict AQI.

Table 3: Comparative Analysis of the above Models

Models	Correctly Classified Instances	Incorrectly Classified Instances	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error	Total Number of Instances	Time	Predictive Accuracy
K nearest neighbor	53	8	0.7643	0.1218	0.2295	41.15	60.155	61	0	86.8852%
Navie Bayes	60	1	0.972	0.0082	0.0902	2.76	23.645	61	0.03	98.3607%
Decision tree	61	0	0	0	0	0	0	61	0	100%

Table 4: The results of Naïve Bayes are shown in the tabular column.

Year	Station name	PM10	Pm10 Index	So2	So2 Index	No2	No2 Index	Observed AQI index	Predicted AQI Index	Observed Health Impacts	Predicted Health Impacts
'jan-dec 2015'	'Export Promotional Park'	181.5	154	4.9	6	21.6	27	182.775	181.5	moderate	moderate
'jan-dec 2014'	'Amco Batteries'	219.3	180	13.8	17	32.8	41	218.162	219.3	poor	poor
'jan-dec 2013'	Kajisonnenahalli	73.8	74	12.5	16	28.7	36	73.683	73.8	satisfactory	satisfactory
'jan-dec 2012'	'CAAQM City Railway Station'	99	99	8.3	10	29.5	37	101.650	99	moderate	Satisfactory
'jan-dec 2011'	'CAAQM S G Halli'	20	20	10.5	13	24.7	31	19.2352	20	good	good

Table 5: The results of K nearest neighbor classifier are shown in the tabular column.

Year	Station name	PM10	Pm10 Index	So2	So2 Index	No2	No2 Index	Observed AQI index	Predicted AQI Index	Observed Health Impacts	Predicted Health Impacts
'jan-dec 2015'	'Export Promotional Park'	181.5	154	4.9	6	21.6	27	182.775	181.5	moderate	moderate
'jan-dec 2014'	'Amco Batteries'	219.3	180	13.8	17	32.8	41	218.162	219.3	poor	moderate
'jan-dec 2013'	Kajisonnenahalli	73.8	74	12.5	16	28.7	36	73.683	73.8	satisfactory	satisfactory
'jan-dec 2012'	'CAAQM City Railway Station'	99	99	8.3	10	29.5	37	101.650	99	moderate	satisfactory
'jan-dec 2011'	'CAAQM S G Halli'	20	20	10.5	13	24.7	31	19.2352	20	good	good

Table 6: The results of Decision Tree are shown in the tabular column.

Year	Station name	PM10	Pm10 Index	So2	So2 Index	No2	No2 Index	Observed AQI index	Predicted AQI Index	Observed Health Impacts	Predicted Health Impacts
'jan-dec 2015'	'Export Promotional Park'	181.5	154	4.9	6	21.6	27	182.775	181.5	moderate	moderate
'jan-dec 2014'	'Amco Batteries'	219.3	180	13.8	17	32.8	41	218.162	219.3	poor	poor
'jan-dec 2013'	Kajisonnenahalli	73.8	74	12.5	16	28.7	36	73.683	73.8	satisfactory	satisfactory
'jan-dec 2012'	'CAAQM City Railway Station'	99	99	8.3	10	29.5	37	101.650	99	satisfactory	satisfactory
'jan-dec 2011'	'CAAQM S G Halli'	20	20	10.5	13	24.7	31	19.2352	20	good	good

REFERENCES

[1] "Approaches in Predicting Urban Air quality in Bangalore City and Comparative Analysis of Predictive Models using Data Mining" in Global journal of Engineering Science and Research Jan 2017,Asha N ,Dr M P Indira Gandi, Kodaikanal.

[2] "Weka Approach for Comparitive Study of Classification Algorithm",in international journal of Advanced Research in Computer and Communication Engineering,vol 2,Issue 4,April 2013,Trilok Chand Sharma, Manoj Jain,Faridabad.

[3] "Comaprision of Different Classification Techniques UsingWeka for Hematological Data" in Americal Journal of Engineering Research,Vol 4,Issue 3 2015,Md.Nurul AMin,Md Absan Habib,Bangladesh.

[4] "Comapritive Analysis of Data Mining Tools and Classification Techniques using Weka in Medical Bioinformatics" in Computer Engineering and Intelligent systems,Vol 4, 2013, Satish Kumar David,Khalid Al Rubean,Suadi Arbia

[5] "Data Mining Methods for prediction of Air Pollution" internation Journal of Applied maths

,Computer Science, Vol 26 2016,Krzsztof Siwek,Stanislaw Osowski,Poland.

[6] kspcb.kar.nic.in / The Karnataka State Pollution Control Board for Prevention and Control of Water Pollution constituted by the Government of Karnataka.

[7] "Data Mining concepts and techniques" by Jiwaei han and Micheline Kamber.

[8] "Data mining methods and Models" by Daneil T Larose.