ENVIRONMENTAL QUALITY PREDICTION AND ITS DEPLOYMENT

Dr.S.Sridevi[1], Rakesh Jampala[2], B.Yaswanth[3]

[1]*Associate Professor, Department of ECE, Vel Tech Rangarajan Dr.Sagunthala R&D Institute Of Science And Technology, Chennai, India*
[2,3]*Student, Department of ECE, Vel Tech Rangarajan Dr.Sagunthala R&D Institute Of Science And Technology, Chennai, India*

---***---

Abstract - *The environment is the source of survival for the human. In the modern days, the degradation of the environment has been increased significantly, when we compared to the last few centuries. Examining and protecting environmental quality has become one of the essential aspects for the government in recent days. The meteorological and traffic factors, burning of fossil fuels, deforestation, industrial parameters, and mass development of civilization played a significant role in environmental quality. With a significant rise in environmental pollution, we need models which will record information about the environment and its pollutants. The deposition of harmful gases in the air, mass deforestation, and industrial factors are affecting the quality of people's lives around the world. Many researchers began to use the big data analytics approach as there environmental sensing networks and sensor data available. In this project, we implement ma�chine learning models to detect and predict environmental quality. Models in time series will be employed for the better prediction of environmental quality*

Key Words**: Environmental quality, Machine learning, Pollutants, Time series models.**

1. INTRODUCTION

With economic development and population rise in cities lead to environmental pollution problems involving air pollution, water pollution, noise and the shortage of land resources have attracted increasing attention. Among these, air pollution is significant problem, as it impact the human health. People exposure to pollutants has resulted in serious health problems. Both developing and developed countries are trying to figure out methods to ameliorate the present air pollution situations. Air pollution is usually caused by energy production from power plants, industries, residential heating, fuel burning vehicles, natural disasters etc. Human health concern is one among the important consequences of air pollution especially in urban areas. Artificial intelligence and machine learning in recent years has been one of the phenomenal advancement of technology Instead of just writing commands as standard works, the philosophy of artificial intelligence, in which the system makes its own choices gradually impacts all aspects of our society. Machine learning is an area where an artificial intelligence system gathers data from sensors and learns to behave in an environment. The ability of machine learning (ML) algorithms to adapt was one of the reasons we chose machine learning to predict the air quality index. The machine learning algorithms in such as naive Bayes classifier, logistic regression and decision tree classifier are implemented in this approach.

1.1 Aim of the project

The primary objective of this project is to analyze and predict the optimum quality of air using the machine learning algorithm. In addition, develop a machine learning model for higher efficacy and lower error rate for better prediction. And top of that to help the society with the optimum model of machine learning for a better tomorrow.

1.2 Project Domain

The project was designed to detect air quality, which is significant factor in contemporary society, as it impact individuals health. Therefore, We utilized Machine learning algorithms to predict the air quality. We designed the project using Supervised learning algorithms of Machine Learning algorithms such as Decision Tree, Support Vector Machine, and Naive Bayes.

1.3 Scope of the Project

As the quality of the air is degrading around the world there is a need for efficient machines for the monitoring air around the world. Inhaling the air with toxics will leads to several disease. With the help of the air quality we can impose the restrictions for saving the environment and can save humanity from many disasters.

1.4 Methodology

Classification is to determine the class to which each data sample of the methods belongs, which methods are used when the outputs of input data are qualitative. The purpose is to divide the whole problem space into a certain number of classes. A wide range of classification methods are present. There are diverse classification methods have been constructed for different data, since there is no particular method that works on every data set. the aim of classification is to assign the new samples to classes by using the pre-labeled samples. The most commonly used classification methods are described below

- Logistic Regression
- Naive Bayes
- Decision Tree
- Random Forest
- K-Nearest Neighbor
- Support Vector Machine

The evaluation of the air quality system is based on employing a confusion matrix. A confusion matrix may be a measure to work out the accuracy of a classifier or a classification model. The confusion matrix is plotted supported the particular class vs predicted class. The class indicates the expected results of the categories for pollutants and whereas the anticipated class indicates the anticipated results of the categories for pollutants and the prediction obtained by heuristic air quality system. With the assistance of confusion matrix, truth positives, false negatives, false positives and true negatives are calculated from the testing results obtained.

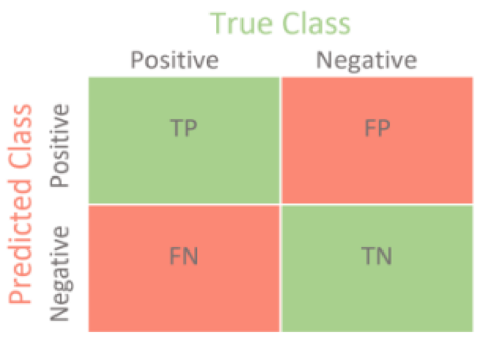


Fig-1: Confusion Matrix

The evaluation of the proposed system is measured using two parameters such as precision and recall respectively. Precision = (TP / TP + FP) Recall = (TP/ TP + FN) . The Precision achieved for Proposed System varies from 0.0 to 1.0 whereas Recall achieved for the proposed system varies from 0.0 to 1.0 based on the type of the heuristics selected by the user.

In the first measure, the info set is split into three parts as training, validation and test data by three-phase division in K-Fold method, and model selection and performance status are simultaneously performed. In the second measure, performance evaluation of classifier models generally uses a validation value. Validation value are often measured because the ratio of knowledge count detected or estimated correctly by the algorithm into all data within the data set.

2. DESCRIPTION OF WORK

Air quality is one among the significant problem that causes health problems, If left unrecognized . There are diverse range of pollutants that are continuously released to environment on daily basis by the different actions of humans and other processes. Many substantial systems are deployed to monitor the air quality continuous, as they pose threat to human survival. The purpose of this project is to show the efficacy of classification based models that are efficient in predicting the air quality.

2.1 Existing System

The existing models are based on image based classification technique. The one among such model works as follow. The concentration of PM2.5 would be calculated via a photographic process model. Observation discloses that the exposure chart displays slightly different indexes of large and small amounts of PM2.5 and is disposed to air quality. For contrast picture, the structures are destroyed and so the bulk of pixels are frequently 0 under a big concentration of PM2.5. The Weibull distribution is used to match the saturation map and to measure the value of the color feature. Finally, a mixture of the aforesaid features and a nonlinear mapping technique will evaluate the PM2.5 concentration of a photo. In contrast with the present techniques, the empirical and visualized properties of actual data collected confirm the consistency and dominance of the proposed process.

2.2 Proposed System

The proposed systems incorporates the machine learning domain in it. The system is based on classification based technique of supervised machine learning branch. It used the decision tree model to predict the air quality. For better accuracy, the decision tree model is compared with its supervised leaning counter parts such as, Logistic regression, Naive bayes, KNN, Random forest and Support Vector machine.

- Add a new heuristic characteristics with machine learning techniques to decrease the false positive in predicting the air quality.
- Made an effort to identify the finest model in machine learning of supervised method to predict the air quality with higher efficacy than the existing systems.
- Used different learning techniques such as Logistic regression, Naive bayes, KNN, Random forest, Decision Trees and Support Vector machine.

2.3 Feasibility Study

The feasibility of the project is analyzed in this part and puts forward a very general project plan and some commercial recommendations for price estimation. In the process of system analysis, the feasibility study of the proposed system was carried out. It is usually ensured that the planned system will not burden the company. For feasibility studies, it is important to understand the key requirements of the system.

3. MODULE DESCRIPTION

3.1 General Architecture

A system planning is the abstract model that defines the structure, conduct, and more interpretations of a system. An architecture picture is a formal description and illustration of a system, ordered in a way that supports cognitive about the structures and performances of the system. A system architecture can consist of system components and the sub-systems established , that will work collectively to device the overall system.

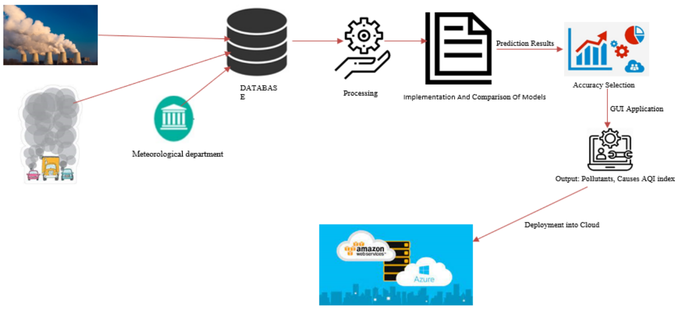


Fig -2: Architecture Diagram

3.2 ER Diagram

An entity–relationship model (or ER model) labels interconnected objects of concentration during a explicit field of data . A elementary ER model entails of entity categories (which organize the items of interest) and lay down relations which will exist among things (instances of these entity types). In software engineering, an ER model is usually designed to signify things a corporate must evoke so as to achieve business processes. Therefore, the ER model becomes an abstract data model, that describes a data or information assembly which can be employed in a database, typically a relational catalog.

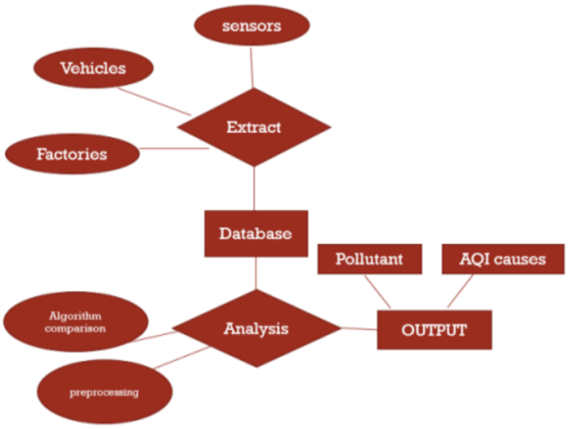
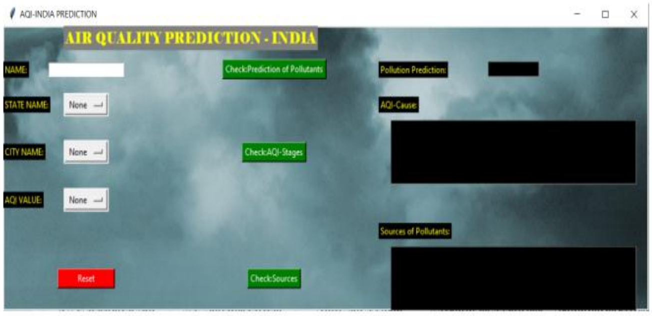


Fig -3: ER Diagram

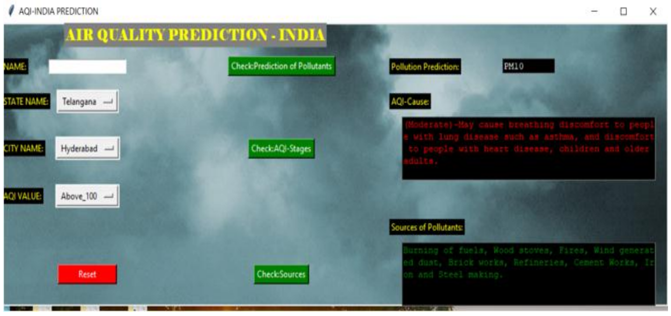
3.2 Input and Output Modules

The structure of input is designed in such a way that it emphases on controlling the quantity of input required, controlling the blunders, evading delay, avoiding extra steps and keeping the process simple. The input is premeditated in such a way so that it delivers security and ease of use with retaining the privacy. In this case, the project is designed in such a way that the models are trained using the dataset provided by the admin. In this implementation the project utilized diverse set of supervised learning algorithms and undergone the efficacy test with each other. Furthermore, the best model with higher efficiency is utilized to get output.


AQI-INDIA PREDICTION
AIR QUALITY PREDICTION - INDIA
Check:Prediction of Pollutants
Check:AQI-Stages
Reset
Check:Sources

NAME:
STATE NAME: None
CITY NAME: None
AQI VALUE: None
Pollution Prediction:
AQI-Cause:
Sources of Pollutants:

Fig -4: Input Design


AQI-INDIA PREDICTION
AIR QUALITY PREDICTION - INDIA
Check:Prediction of Pollutants
Check:AQI-Stages
Reset
Check:Sources

NAME:
STATE NAME: Telangana
CITY NAME: Hyderabad
AQI VALUE: Above_100
Pollution Prediction: PM10
AQI-Cause:
(Moderate)-May cause breathing discomfort to people with lung disease such as asthma, and discomfort to people with heart disease, children and older adults.
Sources of Pollutants:
Burning of fuels, Wood stoves, Fires, Wind generated dust, Brick works, Refineries, Cement Works, Iron and Steel making.

Fig -5: Output Design

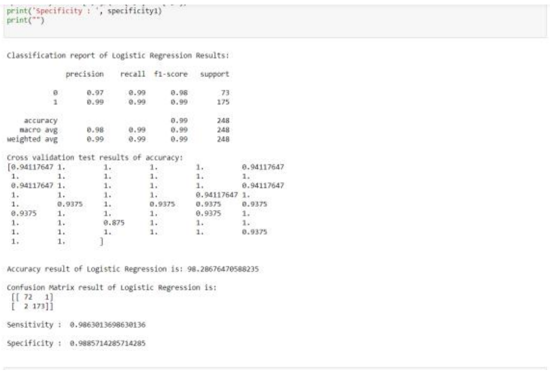
4. TESTING

The reason for the test part is to find bugs. Testing tries to find all problems and potential problems within the work item. Provides a way to visualize the useful solid ground of segments, sub assemblies, groupings and get started. The sample data is used for testing. However, it doesn't matter what quality the data used in the test questions are. The tests are designed to ensure that the system has been accurately associated with efficiency prior to live operating commands.

4.1 Types of Testing

- Unit Testing
- Functional Testing
- White box testing
- Black box Testing
- Regression Testing

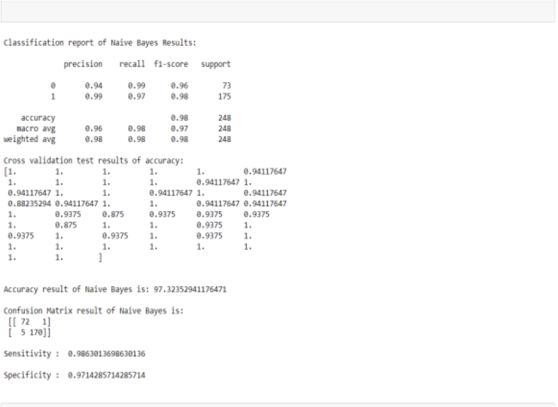
4.2 Test Results



```
print('Specificity : ', specificity1)
print("")
```

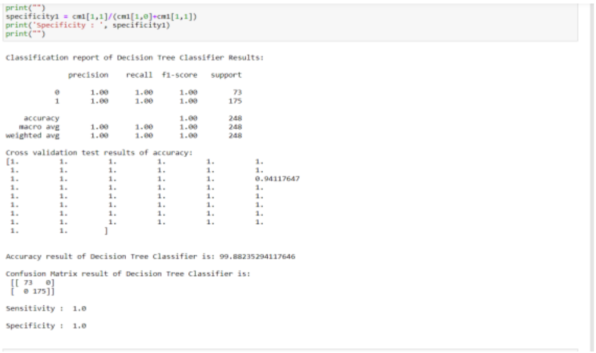
```
Classification report of Logistic Regression Results:

              precision    recall  f1-score   support

           0       0.97      0.99      0.98        73
           1       0.99      0.99      0.99       175

    accuracy                           0.99       248
   macro avg       0.98      0.99      0.99       248
weighted avg       0.99      0.99      0.99       248

Cross validation test results of accuracy:
[0.94117647 1.         1.         1.         1.         0.94117647
 1.         1.         1.         1.         1.         1.
 0.94117647 1.         1.         1.         1.         0.94117647
 1.         1.         1.         1.         0.94117647 1.
 1.         0.9375     1.         0.9375     0.9375     0.9375
 0.9375     1.         1.         1.         0.9375     1.
 1.         1.         0.875      1.         1.         1.
 1.         1.         1.         1.         1.         0.9375
 1.         1.        ]

Accuracy result of Logistic Regression is: 98.28676470588235

Confusion Matrix result of Logistic Regression is:
 [[ 72   1]
 [  2 173]]

Sensitivity :  0.9863013698630136

Specificity :  0.9885714285714285
```

Fig -6: Logistic Regression Result



```
Classification report of Naive Bayes Results:

              precision    recall  f1-score   support

           0       0.94      0.99      0.96        73
           1       0.99      0.97      0.98       175

    accuracy                           0.98       248
   macro avg       0.96      0.98      0.97       248
weighted avg       0.98      0.98      0.98       248

Cross validation test results of accuracy:
[1.         1.         1.         1.         1.         0.94117647
 1.         1.         1.         1.         0.94117647 1.
 0.94117647 1.         1.         0.94117647 1.         0.94117647
 0.88235294 0.94117647 1.         1.         0.94117647 0.94117647
 1.         0.9375     0.875      0.9375     0.9375     0.9375
 1.         0.875      1.         1.         0.9375     1.
 0.9375     1.         0.9375     1.         0.9375     1.
 1.         1.         1.         1.         1.         1.
 1.         1.        ]


Accuracy result of Naive Bayes is: 97.32352941176471

Confusion Matrix result of Naive Bayes is:
 [[ 72   1]
 [  5 170]]

Sensitivity :  0.9863013698630136

Specificity :  0.9714285714285714
```

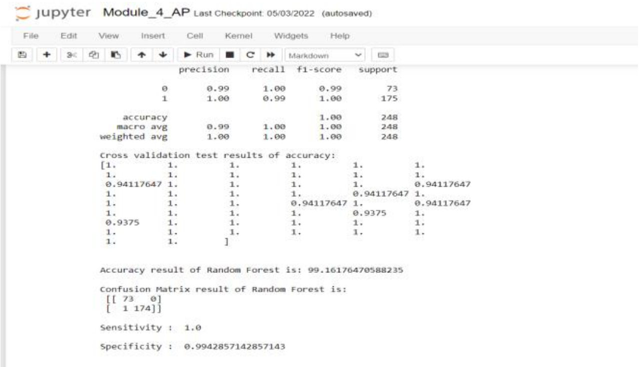
Fig -7: Naïve Bayes Result



```
print("")
specificity1 = cm1[1,1]/(cm1[1,0]+cm1[1,1])
print('Specificity : ', specificity1)
print("")
```

```
Classification report of Decision Tree Classifier Results:

              precision    recall  f1-score   support

           0       1.00      1.00      1.00        73
           1       1.00      1.00      1.00       175

    accuracy                           1.00       248
   macro avg       1.00      1.00      1.00       248
weighted avg       1.00      1.00      1.00       248

Cross validation test results of accuracy:
[1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         0.94117647
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.        ]


Accuracy result of Decision Tree Classifier is: 99.88235294117646

Confusion Matrix result of Decision Tree Classifier is:
 [[ 73   0]
 [  0 175]]

Sensitivity :  1.0

Specificity :  1.0
```

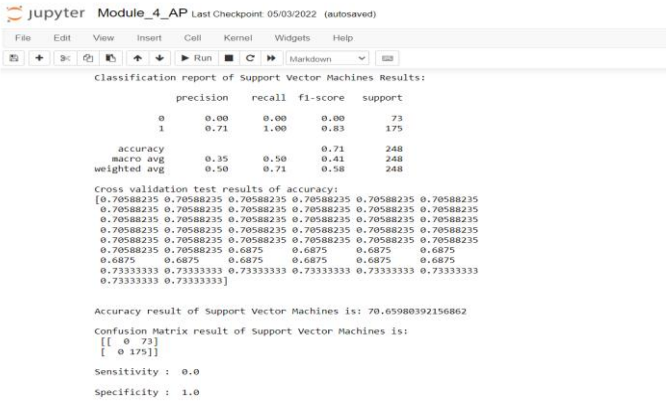
Fig -8: Decision Tree Result



jupyter Module_4_AP Last Checkpoint: 05/03/2022 (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Run
Markdown

```
              precision    recall  f1-score   support

           0       0.99      1.00      0.99        73
           1       1.00      0.99      1.00       175

    accuracy                           1.00       248
   macro avg       0.99      1.00      1.00       248
weighted avg       1.00      1.00      1.00       248

Cross validation test results of accuracy:
[1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 0.94117647 1.         1.         1.         1.         0.94117647
 1.         1.         1.         1.         0.94117647 1.
 1.         1.         1.         0.94117647 1.         0.94117647
 1.         1.         1.         1.         0.9375     1.
 0.9375     1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.        ]


Accuracy result of Random Forest is: 99.16176470588235

Confusion Matrix result of Random Forest is:
 [[ 73   0]
 [  1 174]]

Sensitivity :  1.0

Specificity :  0.9942857142857143
```

Fig -9: Random Forest Result



jupyter Module_4_AP Last Checkpoint: 05/03/2022 (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Run
Markdown

```
Classification report of Support Vector Machines Results:

              precision    recall  f1-score   support

           0       0.00      0.00      0.00        73
           1       0.71      1.00      0.83       175

    accuracy                           0.71       248
   macro avg       0.35      0.50      0.41       248
weighted avg       0.50      0.71      0.58       248

Cross validation test results of accuracy:
[0.70588235 0.70588235 0.70588235 0.70588235 0.70588235 0.70588235
 0.70588235 0.70588235 0.70588235 0.70588235 0.70588235 0.70588235
 0.70588235 0.70588235 0.70588235 0.70588235 0.70588235 0.70588235
 0.70588235 0.70588235 0.70588235 0.70588235 0.70588235 0.70588235
 0.70588235 0.70588235 0.70588235 0.70588235 0.70588235 0.70588235
 0.70588235 0.70588235 0.6875     0.6875     0.6875     0.6875
 0.6875     0.6875     0.6875     0.6875     0.6875     0.6875
 0.73333333 0.73333333 0.73333333 0.73333333 0.73333333 0.73333333
 0.73333333 0.73333333]

Accuracy result of Support Vector Machines is: 70.65980392156862

Confusion Matrix result of Support Vector Machines is:
 [[  0  73]
 [  0 175]]

Sensitivity :  0.0

Specificity :  1.0
```

Fig -10: Support Vector Machine Result

Methods	Test/True Accuracy
Logistic Regression	97.46
Naïve Bayes	97.38
Random Forest	99.16
Support Vector Machine	70.65
K Nearest Neighbor	97.61
Decision Tree	99.88

Table -1: Resulting Accuracies

5. RESULTS AND DISCUSSIONS

The development of this novel system is to provide the higher efficacy in detecting the air quality. By means of this new system, an individual can have an opportunity to assess the air quality to gain the knowledge of the air quality in the location. The designed new system uses the classification based algorithm for the prediction of air quality. For the efficacy terms, it is compared with five diverse sets of classification based algorithms of machine learning such as Logistic regression, Naive bayes, KNN, Random Forest and Support Vector Machine. The performance of the proposed system is at a pinnacle than the rest of the compared models. The developed model uses the decision tree to classify the data, which provide the output of the air quality data by taking input data. There are different classification techniques are evolved, since different algorithms have been developed to work with different sets of data, as there is no exact method that works on all data set. As stated in literature works, the goal of classification is to allocate the new samples to classes by using the pre-labeled samples.

6. CONCLUSION AND FUTURE ENHANCEMENTS

Prevention of air pollution is the need of the hour, so a influential machine learning system was established with the help of prediction model. Prediction of pollution events has become most important issue in major cities in India due to the increased expansion of the population and the associated impact of traffic capacities. Data from a variety of heterogeneous capitals were used and involved collection and cleansing for use in machine learning algorithms. The number of model parameters and optimized outputs were reduced with help of structure regularization which in turn, alleviated model complexity. The Decision Tree Algorithm gave the best results among all the algorithms, with an overall accuracy of 99.8.

Many efforts from both local and state administrations are done in order to understand and predict air quality index aiming to improve community health. With the progression of IoT substructures, big data knowledges, and machine learning techniques, real-time air quality monitor and evaluation is desirable for upcoming smart cities. In future a close working between authorities and also applying them with MLT background, which may provide boosting in prediction. This can be achieved by building operational models that adapt automatically to changes in environment. Also, more data can be included to increase data seasonality.

REFERENCES

[1] Acharjya, Debi Prasanna, and Kauser Ahmed (2019), ”A survey on big data analytics: challenges, open research issues and tools.” International Journal of Advanced Computer Science and Applications, vol.7,no.2, pp.511-518.

[2] A. Gnana Soundari, J. Gnana Jeslin, Akshaya A.C (2019),”Indian Air Quality Prediction And Analysis Using Machine Learning”, International Journal of Computer Applications Technology and Research ,Volume 8,Issue 09, 367-370.

[3] Abed Al Ahad M, Sullivan F, Demsar U, Melhem M, Kulu H(2020),” The Effect Of Air-pollution And Weather Exposure On Mortality And Hospital Admission And Implications For Further Research: A Systematic Scoping Review”. PLoS ONE 15(10): e0241415.

[4] D. Qin, J. Yu, G. Zou, R. Yong, Q. Zhao and B. Zhang (2019), ”A Novel Combined Prediction Scheme Based on CNN and LSTM for Urban PM2.5 Concentration,” in IEEE Access, vol.7, pp.20050-20059.

[5] G. Yue, K. Gu and J. Qiao (2019), ”Effective and Efficient Photo-Based PM2.5 Concentration Estimation,” in IEEE Transactions on Instrumentation and Measurement, vol.68, no.10, pp. 3962-3971.

[6] T. Zhang and R. P. Dick (2020), ”Estimation of Multiple Atmospheric Pollutants Through Image Analysis,” IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, pp. 2060-2064, doi:10.1109/ICIP.2019.8803130.

[7] S. Y. Muhammad, M. Makhtar, A. Rozaimee, A. Abdul, and A. A. Jamal (2019), “Classification model for air quality using machine learning techniques,” International Journal of Software Engineering and Its Applications, pp. 45-52.