e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

AIR QUALITY PREDICTIVE ANALYSIS USING MACHINE LEARNING

Prathapa[*1], Dishanth S[*2], Anujith Shetty[*3], Mrs. Navya S Rai[*4]

[*1,2,3,4]Dept, I.S.E, A.J. Institute Of Engineering And Technology, Mangalore, Karnataka, India

ABSTRACT

Examining and protecting the air quality has become one of the most essential activities for the government in many industrial and urban areas today. With the rapid development of various industries and motorized transportation, large amounts of harmful substances such as sulfur dioxides, nitrogen oxides, carbon monoxides, and hydrocarbons are released into the atmosphere, lasting a long time and in concentrations exceeding tolerable environmental limits. As a result of this, people's respiratory and cardiovascular systems will get affected. Therefore, we need to develop models that will record the information about the concentrations of air pollutants (SO2, NO2, CO etc). In this paper, we are using two machine learning algorithms (Linear Regression and Decision Tree) are used to predict the concentration of air pollutants in the environment. The results are promising and the implementation of these algorithms could be very efficient in predicting air pollutants.

I. INTRODUCTION

In developing countries like India, the speedy increase in population and the economic upswing in cities have led to environmental problems such as airpollution, water pollution, noise pollution, and many more. Urban air pollution is a major problem in both developed and developing countries, as atmospheric pollutants have a huge effect on human health.

Numerous illnesses such as lung cancer and asthma may be caused by various atmospheric pollutants. In addition, some other serious environmental problems can also result from air pollution, such as acid rain, ozone depletion, and the greenhouse gas effect. For example, SO2 and NO2 are the main causes of acid rain, while CO2 and NO2 are the main reasons for the greenhouse gas effect.

Air pollution monitoring and control is thus becoming more and more significant. Real-time air quality information, such as the concentration of PM2.5, PM10, and NO2, is an important aspect of pollution management and protecting human beings from damages caused by air pollutants. In the atmosphere, obtaining, and maintaining the high quality of the air, it is one of the biggest challenges facing the cities of"megacities" with a large population, businesses, and industries. As the population grows, so does transportation, and the consumption of electricity and fuel, and increases. Also, there is a huge amount of waste that is thrown out on the wasteland, we know that as well. The atmospheric air is highly polluted, it is a serious threat to all species of living organisms on the planet.

II. METHODOLOGY

System Design

The system architecture gives the description of the various internal components that have been integrated to become the final system. It also gives a clear idea about the working of all the process that happens within the system and the communication processes amongst the various parts organized. In the following figure, we can see how ML algorithms are implemented toget accurate results i.e. whether the air quality is good or bad. The system architecture is as follows:

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

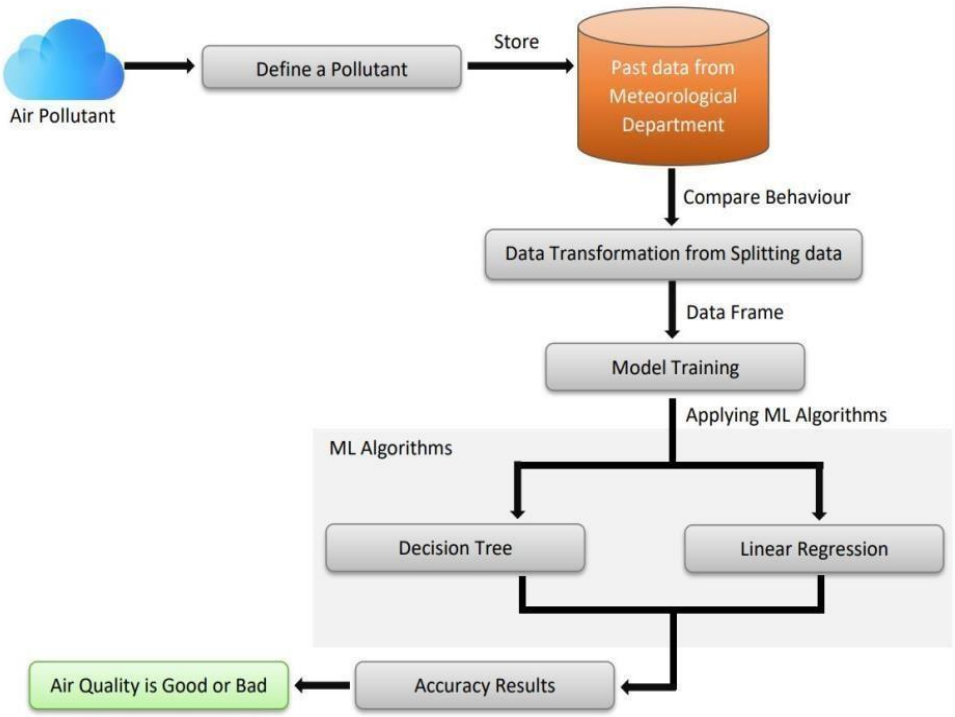

Figure 2.1 System Architecture

Data-flow diagram (DFD) is the graphical representation of the "stream" of information through a data framework, which capture, manipulate, store, and distribute databetween system and environment. It can be utilized for the perception of information handling. A DFD model uses an exceptionally predetermined number of primitive images to speak to the capacities performed by a framework and logical information flow of the system.

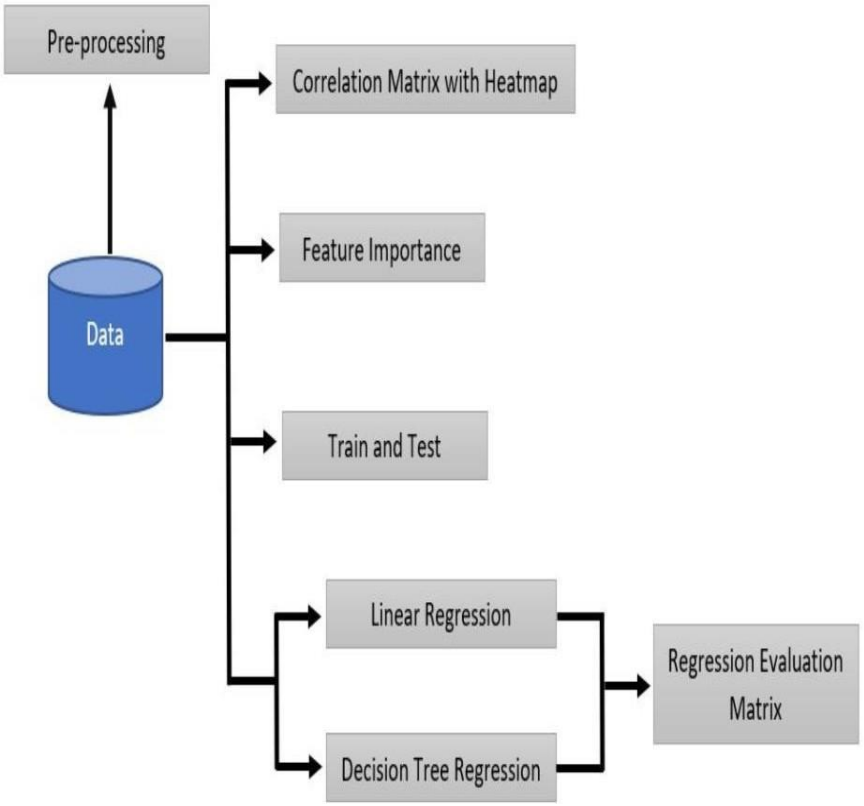

Figure 2.2 Data flow Diagram

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

A flowchart is a type of diagram that depicts a workflow, process or system. A flowchart can also be defined as a diagrammatic representation or the blueprint of an algorithm, a step-by-step procedure to solving a task. The flowchart shows the steps as boxes of various kinds, and their order by connecting the boxes with arrows. This diagrammatic representation gives a solution model to the given problem. Flow charts are used in analyzing, designing, documenting or managing a process or program in various fields.

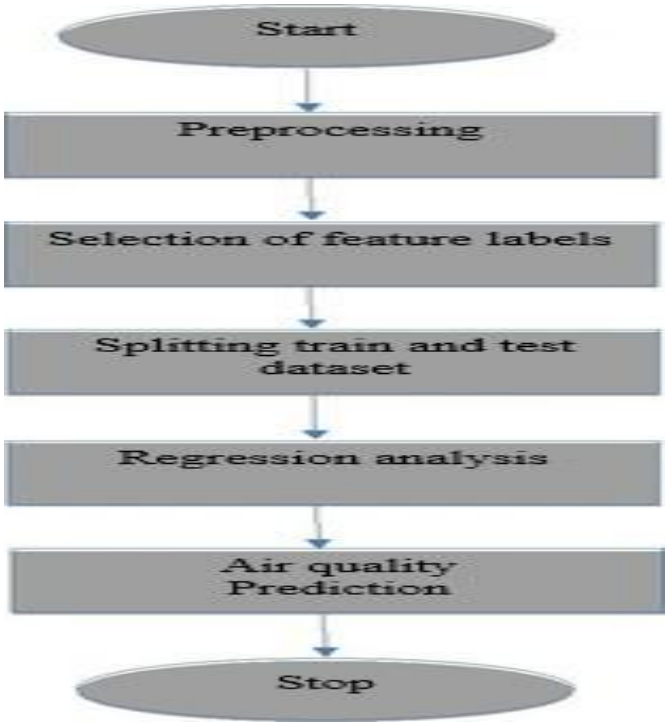

Figure 2.3 Flow Chart

III. MODELING AND ANALYSIS

The proposed model is capable of predicting concentration of air pollutants for the upcoming days.

Steps Involved:

Gathering Data:

- Downloading data from web sites
- Downloading already tested data

Preprocessing Data:

- **Checking Null value**

	T	TM	Tm	SLP	H	VV	V	VM
0	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

Correlation Matrix with Heatmap:

- Correlation Matrix states how the featuresare related to each other or the target variable.
- Correlation can be positive
- Heatmap makes it easy to identify which featuresare most related to the target variable, we will plot the heatmap of correlated features using the seaborn library.

Feature Importance:

- We can get the feature importance of each feature of our dataset by using the feature importance property of the model.
- Feature importance gives us a score for eachfeature of our data, the higher the score more important or relevant is the feature to our output variable.

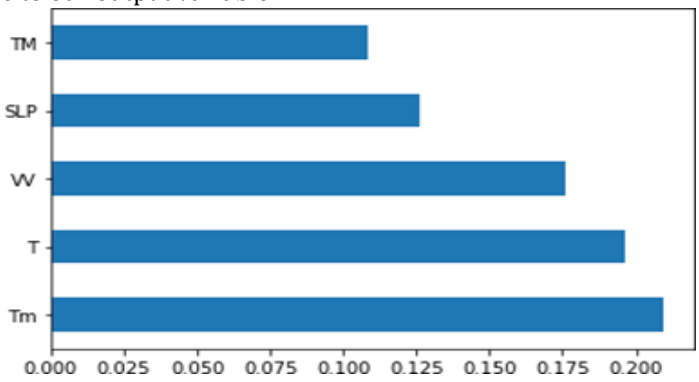

Fig 2. Graph of feature importance for bettervisualization

Decision Tree Regressor:

- Decision tree builds or develops a regression or classification model which is in the form of a tree structure.
- It breaks down the dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.
- The final result is in the form of a tree with decision nodes and leaf nodes.

Train Test split:

- Train Test split is a procedure that is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.
- Even though simple to use and intercept, there are times when the procedure should not beused, such as when we have a small dataset and situations where additional configuration is required, such as when it is used for classification and the dataset is not balanced.

Tree Visualization:

- Decision trees are a popular tool in decisionanalysis method.
- They can support decisions that helps to thevisual representation of each decision.

Hyperparameter Tuning Decision Tree Regressor:

- Hyperparameter tuning is searching the hyperparameter space for a set of values that willoptimize our model architecture.
- There are two specific hyperparameters. They are:

Max depth : This is the maximum number of children nodes that can grow out from the decision tree until the tree is cut off.

Min samples leaf : This is the minimum number of data points, that are required to be present in the leaf node.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

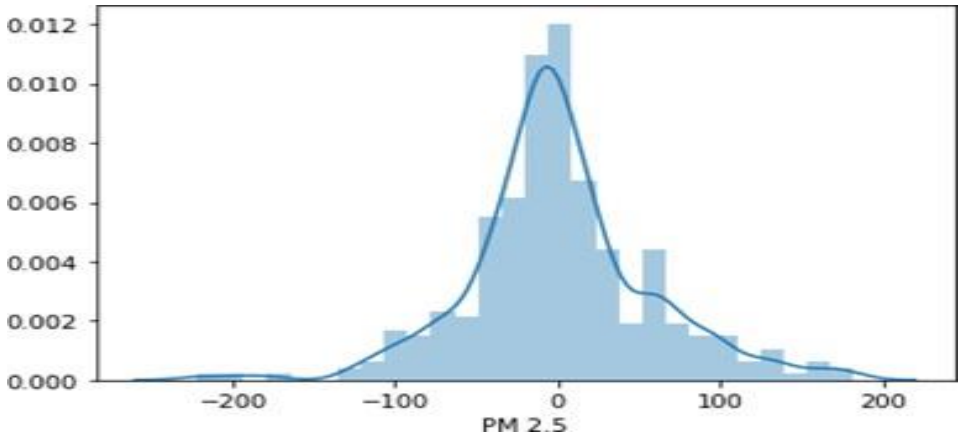

Fig 3.2 Graph of feature importance for better visualization

Regression Evaluation Metrics:

Here are three common evaluation metrics forregression problems:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)

Comparing these metrics:

- MAE is the easiest to understand because it'sthe average error.
- MSE is more popular than MAE because MSE "punishes" larger errors, which tends to be usefulin the real world.
- RMSE is even more popular than MSE because RMSE is interpretable in the "y" units.

IV. CONCLUSION

This paper has proposed air quality analysis and a forecast based on an intelligent algorithm with parameter optimization and decision rules. SA and DT were used to achieve the best classification accuracy and classify air quality by the obtained decision rules, and they were shown to be efficientfor generating decision rules.

Regression analysis techniques are used to predict the concentration of Carbon monoxide C.O. in the environment. This research provided a prediction model for improving air quality and this model could effectively improve people's living environment and protect people's health.

V. REFERENCES

[1] Haotian Jing, Yingchun Wang, "Research on Urban Air Quality Prediction Based on Ensemble Learning of XGBoost", E3S Web of Conferences 165, 02014 (2020) CAES2020. https://doi.org/10.1051/e3sconf/202016502014.

[2] Zepeng Qin, Chen Cen, Xu Guo, "Prediction of AirQuality Based on KNN-LSTM", ICSP 2019 IOP Conf. Series: Journal of Physics: Conf. Series 1237 (2019) 042030 IOP Publishing doi:10.1088/1742-6596/1237/4/042030.

[3] Chou-Yuan Lee 1* , Zne-Jung Lee 1 , Jian-Qiong Huang 1 , Fu-Lan Ye 1 , Zheng-Yuan Ning 1 and Cheng-Fu Yang 2*, "Urban Air Quality Analysis and Forecast Based on Intelligent Algorithm withParameter Optimization and Decision Rules", 1 School of Technology, Fuzhou University of International Studies and Trade, Fuzhou 350202, China;

[4] Barai, Sudhirkumar & Dikshit, A. & Sharma, Sameer. (2007). "Neural Network Models for Air Quality Prediction: A Comparative Study." 10.10 07/978-3-540-70706-6_27., https://www.researchgate.net/publication/225620689.

[5] A.Aarthi, P.Gayathri, N.R.Gomathi , S.Kalaiselvi , Dr.V.Gomathi, "Air Quality Prediction Through Regression Model", International Journal Of Scientific & Technology Research Volume 9, Issue 03, March 2020 ISSN 2277-8616, http://w ww.ijstr.org/