Naveen Kumar,2021, 9:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

Air Quality Predictive Analysis Using Machine Learning

Naveen Kumar, Prathap, Aditya Madeva, Manoj Naik, Asst. Prof. Shilpashree
Department of Information Science & Engineering,
Dayananda Sagar College of Engineering,
Bangalore,India.
naveekumar345@gmail.com, poojariprathap@gmail.com, aditya.moger@gmail.com, manunaik004@gmail.com,
shilpashree-ise@dayanandasagar.edu

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 amount so harmful substances such as sulfur dioxides, nitrogen oxides, carbon on oxides, and hydrocarbons are released into the atmosphere, lasting a long time and in concentration sex ceding tolerable environmental limits. As are sultof this,people's respiratory and cardio vascular system will get affected. There fore,we need to develop the models that will record the information about the concentrations of air pollutants (SO2,NO2,COetc). 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 implementation of these algorithms could be very efficient in predicting air pollutants.

Keywords:- Air Quality, Machine Learning, Linear Regression, Decision Tree.

I. INTRODUCTION

In the developing countries like India, the speedy increase in population and economic upswing in cities have led to environmental problems such as air pollution,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, 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 controlling is thus becoming more and more significant.Real-time air quality information, such as the concentration of PM2.5, PM10, NO2, is an important Aspect for pollution management and protecting human beings from damages caused by air pollutants.

II. PROPOSED SYSTEM

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

1. Steps Involved:

1.1 Gathering Data:

- Downloading data from web sites
- Downloading already tested data

1.2 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
...	...	...	...	...	...	...	...	...

International Journal of Science, Engineering and Technology

An Open Access Journal

- Pair plot

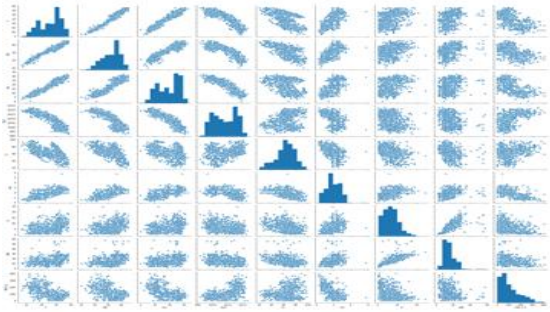

Fig 1.Pairplot

1.3 Correlation Matrix with Heatmap:

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

1.4 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 each feature of our data, the higher the score more important or relevant is the feature towards our output variable.

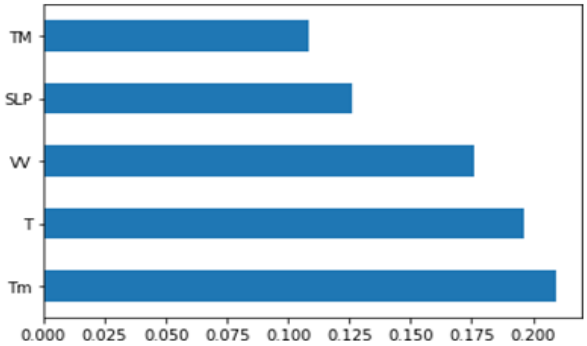

Fig 2.Graph of feature importance for better visualization.

1.5 Decision Tree Regressor:

- Decision tree builds or develpos a regression or classification models which are 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.

1.6 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 a predictions on data not used to train the model.
- Even though simple to use and intercept, there are times when the procedure should not be used, such as when we have small dataset and situations where additional configuration is required, such as when it is used for classification and the dataset is not balanced.

1.7 Tree Visualization:

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

1.8 Hyperpara meter Tuning Decision Tree Regressor:

- Hyper para meter tuning is searching the hyperparameter space for a set of values that will optimize 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.

{'max_depth': 5,
'max_features': None,
'max_leaf_nodes': 50,
'min_samples_leaf': 1,
'min_weight_fraction_leaf': 0.1,
'splitter': 'best'}

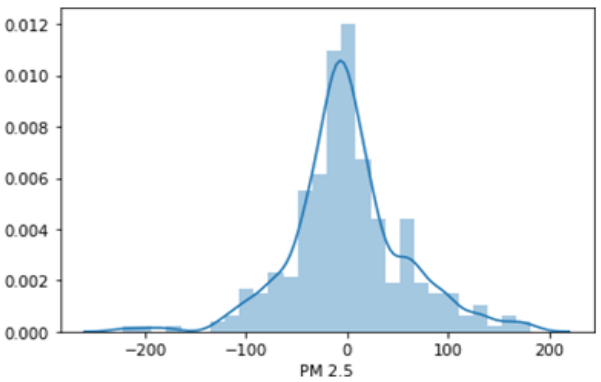

Fig 3.Model Evaluation

1.9 Regression Evaluation Metrics:

International Journal of Science, Engineering and Technology

An Open Access Journal

Here are three common evaluation metrics for regression problems:

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

1.9.1 Comparing these metrics:

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

1.9.2Decision Tree Regressor:

```
print('MAE:', metrics.mean_absolute_error(y_test, prediction))
print('MSE:', metrics.mean_squared_error(y_test, prediction))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, prediction)))
```

```
MAE: 22.473640752032527
MSE: 1880.607923494665
RMSE: 43.365976565674906
```

1.9.3Linear Regression

```
print('MAE:', metrics.mean_absolute_error(y_test, prediction))
print('MSE:', metrics.mean_squared_error(y_test, prediction))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, prediction)))
```

```
MAE: 44.83624126628639
MSE: 3687.5430309324192
RMSE: 60.725143317512384
```

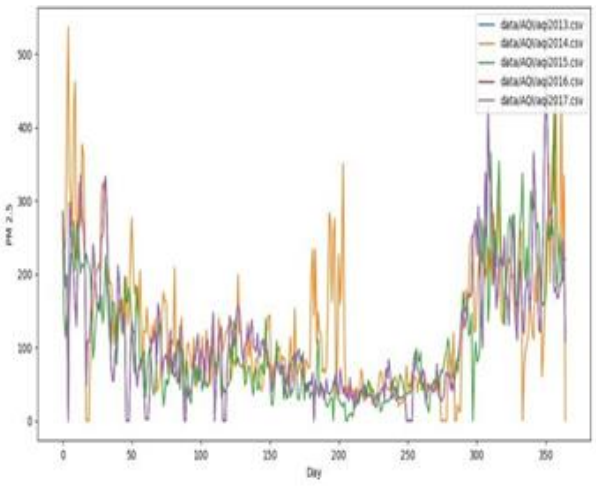

Fig 4.Evaluation Metrics result

2. System Architecture:

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

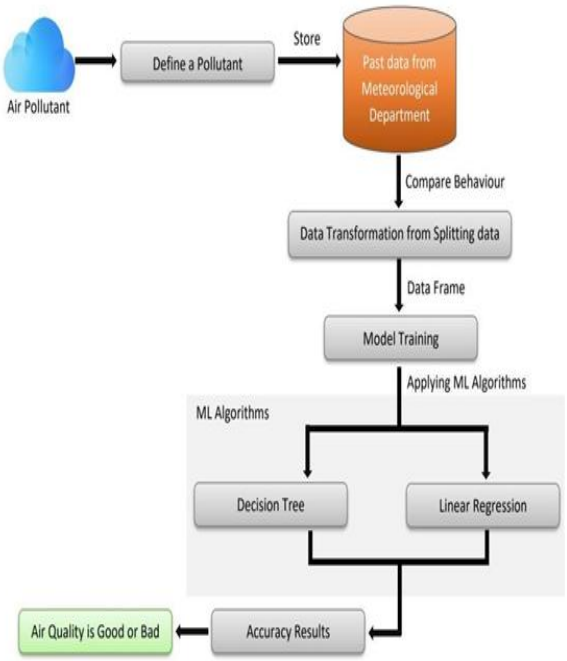

Fig 5.System Architecture

3. Data Flow Diagram:

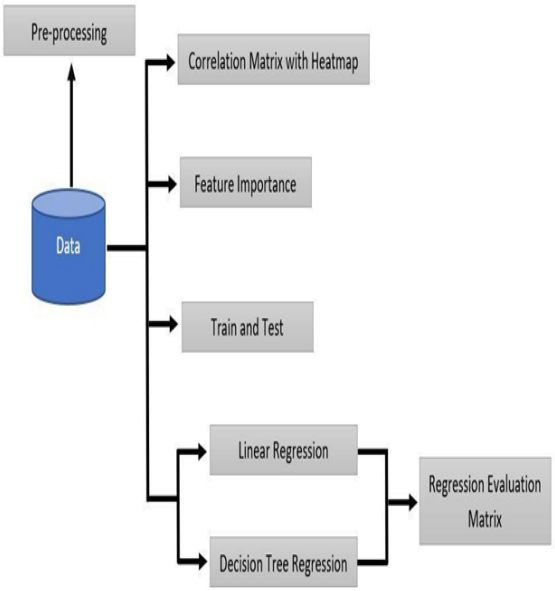

Fig 6.Data Flow Diagram.

International Journal of Science, Engineering and Technology

An Open Access Journal

A data-flow diagram (DFD) is a graphical representation of the "stream" of information through a data framework. DFDs can likewise 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 the information stream among the capacities.

III. CONCLUSIONS

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 efficient for 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.

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 Air Quality 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 with Parameter 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." 10.1007/978-3-540-70706-6_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://www.ijstr.org/