ISSN No:-2456-2165

An Effective System to Forecast Environmental Pollution using Air Quality Index

Mohammad Ishaque Ali, K. Yasudha

Abstract:- One of the major issues we are facing all around the world is "GLOBAL WARMING" and its effects on our daily life which is only due to poor Air quality. It is something when the earth's surface temperature increases slowly and makes the air quality poor. The poor air quality affects human health in many ways. The Machine Learning Algorithm (ML), is used to predict Air quality Index. The main focus of the paper is to establish the relationship between dependent data and independent data. Here, we go through the various feature engineering processes to analyze that which machine learning algorithm must be used that can give the best result and less overfitting.

Keywords:- *Air Pollution, Particulate matter, Decision Tree Regression, Feature Engineering, Human Freedom Index, Linear Regression, Machine Learning, Mutual Information, Random Forest Regression.*

I. INTRODUCTION

In today's era, mitigating the Pollution is one of the major tasks for the government body as well as in our society. India is second most populous countries in the world. There is an increased level of pollution due to very high population exist in India.. In India, Bangalore is one of the main metropolitan cites and here we can calculate the Air Quality Index. This model is capable to predict air quality index using one of the main pollutants known as Particulate Matter(PM).Particulate Matter is key component of air pollution, which enters into the human body when they breathe and defect their lungs, which causes serious heart diseases. PM 10 is the dust particles which have a diameter of 10 micrometer and PM 10 is also called as core dust particles. PM 2.5 is generally referred to as an atmospheric particulate matter which has diameter of around 2.5 micrometer and it is about 2.8 percent of diameter of human hair. The increasing of PM 2.5 level causes invisibility in our environment. This is the mixture of indoor activities as well as outdoor activities done by human being. The outdoor activity refers to the pollution caused by fossil fuels, burning of coal in industries and forest fires. The main source of generation of PM 2.5 is thermal power plant ,because here the coal is continuously burnt to produce electricity. This is also generated by indoor activity like traditional cooking, smoking etc.

II. DATA SET

Data set represents information of Air Quality Index with PM2.5 value in the City of Bangalore. It contains 10 columns and 2,187 rows. The dataset has the following attributes such as: Average total annual, Annual average wind speed, Number of days with rain, Number of days with snow, Number of days with storm, Number of foggy days, Number of days with tornado, Number of days with hail. AQI is an index for reporting daily air quality. It tells how clean or polluted air is, and what associated health effects might be a concern . The AQI focuses on health effects which may be experienced within a few hours or days after breathing polluted air. EPA (Environment Protection Act) calculates the AQI for five major air pollutants regulated by the Clean Air Act: ground-level ozone, particle pollution (also known as particulate matter), carbon monoxide, sulfur dioxide, and nitrogen dioxide.

III. PRE-PROCESSING THE DATA SET

Data is available with these three things : incompleteness, noise, inconsistency. Incompleteness leads the lacking of attributes. Noise will deviates from actual result. and Incompleteness is a kind of discrepancies.

Data cleaning : data cleaning is the process to fill the missing values and smooth the noisy data. This process ensures the consistency in data set. This task can be done using the package called sklearn .preprocessing. This package will use the methods given below :

- Using mean.
- Using median
- Using Most Frequent value

Data Reduction : The data reduction involves attributes subset selection, Dimensionality reduction. This can also be referred as feature engineering

IV. MULTIVARIATE ANALYSIS

This is first and basic step to be performed in explore the Data analysis. This helps to understand the data and analyze the data .and it also help us to go to right path whenever the machine learning model is to be implemented. Here Multivariate analysis is considered ,so there is a library which will be used called 'seaborn' is used, and in this library there is function called 'pairplot' is used. With the help of this function it can be measured that how the values are plotted and also the diagrams which is depicted will lead to the idea of correlation.

ISSN No:-2456-2165

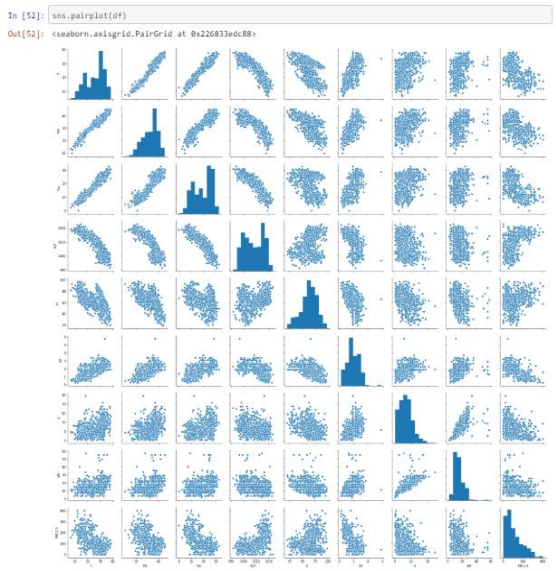

Fig 1

Here my Dependent feature is PM2.5 and remaining is my Independent features. Here the multivariate analysis is done .It helps us to compare various features with respect to each other. When I compare PM with T (maximum temperature) we can see that when T is increasing then PM is also increasing , so there is a correlation. If I have many features and I want to see how other feature is behaving with respect to each other, so we can use sns.pairplot () function. Here some of the diagrams which have some pattern and some of the diagrams have no pattern. This will help us judge which machine learning algorithm I must use.

V. FEATURE IMPORTANCE:

Feature importance gives us a score for each feature of data, the higher the score more important or relevant is the feature towards output variable. Extratreeregressor is also important because it will see the outliers

```
from sklearn.ensemble import ExtraTreesRegressor
import matplotlib.pyplot as plt
model = ExtraTreesRegressor()
model.fit(X,y)
```

```
C:\Users\krish.naik\AppData\Local\Continuum\anaconda3\envs\myenv\lib\site-pa
t.py:246: FutureWarning: The default value of n_estimators will change from
0.22.
  "10 in version 0.20 to 100 in 0.22.", FutureWarning)
```

```
ExtraTreesRegressor(bootstrap=False, criterion='mse', max_depth=None,
          max_features='auto', max_leaf_nodes=None,
          min_impurity_decrease=0.0, min_impurity_split=None,
          min_samples_leaf=1, min_samples_split=2,
          min_weight_fraction_leaf=0.0, n_estimators=10, n_jobs=None,
          oob_score=False, random_state=None, verbose=0, warm_start=False)
```

```
In [105]: print(model.feature_importances_)
```

```
[0.14624008 0.07974868 0.30984738 0.11468752 0.07902165 0.16893677
 0.05704029 0.04447722]
```

Fig 2

VI. ALGORITHM DESCRIPTION

A. *Linear Regression*

Linear regression establishes the relationship between dependent variable and independent variable by fitting the linear equation to given data .The scattered plot is used to connect these features. If there is one independent variable and one dependent variable ,in this case we call It as simple linear regression. To represents the simple linear regression in the form of equation we use "y= mx+c", where m is slope, c is intercept , x is independent variable and y is dependent variable. If There are more than one independent variable and one dependent variable ,in this case we call it as Multiple linear regression.

B. *Ridge and Lasso Regression :*

Regularization is a technique to handle the expensive and complex record in a smooth way. It reduce the complexity of model and shrink magnitude of the coefficient of regression. There are two types of Regularization :

- Ridge Regression
- Lasso Regression

- Ridge Regression: Ridge Regression is the technique which adds the difference between actual value and predicted value and penalty of the difference.
- Lasso Regression: This is almost same like ridge regression , but only difference is ,in the ridge regression magnitude of co-efficient can not be zero(0), but almost equal to zero. but in the lasso regression the magnitude of co-efficient can be zero. This can also be act as a feature selection.

C. *Decision Tree Algorithm-*

Decision Tree is decision-making tool which uses a tree structure representation, where two results will be considered either yes or no. Decision tree implementation can be done in two different ways, first way is classification and second way is regression. It breaks the data set into subset concurrently the tree will also be developed. Decision tree make regression model in the arrangement of a tree structure. The final structure of a tree will be with decision node and leaf node.

D. *Random Forest Regression:*

A Random Forest is a machine learning algorithm which can be implemented with both regression and classification technique. Random forest algorithm follow bagging technique using multiple decision trees. The main concept to combine various decision trees is to examine result with less overfitting. It splits the decision tree into many trees .after this step the outcome is measured by merging various decision tree.

For implementing this concept the equation the equation is used

$$g(x)=f0(x)+f1(x)+f2(x)+...$$

where , g is sum of models, fi is models.

VII. RESULT

When all the features are selected as linear regression algorithm it gives accuracy of trained is 0.48 and accuracy of test data is 0.49. Accuracies of each algorithms are analyzed with training data sample and testing data sample. At last Random forest regression is selected , which gives highest accuracy and less overfitting as compare to other.

Algorithm	Test data accuracy
Linear Regression	0.49
Ridge Regression	0.50
Lasso Regression	0.50
Decision tree	0.69
Random forest	0.76

Table 1

VIII. CONCLUSION

A air quality index system plays an main role nowadays in monitoring air quality. A keen attention towards PM 2.5 is needed. Random forest algorithm produce highest accuracy and less overfitting than remaining algorithm .This model will help us predict the particulate matter2.5. Multivariate analysis can be used to better analyze that which machine learning algorithm is used to improve the performance and reduce the complexity. Feature selection technique can also be used for better prediction.

REFERENCES

[1]. Shweta Taneja, Dr. Nidhi Sharma, Kettun Oberoi, Yash Navoria ,"Predicting Trends in Air Pollution in Delhi using Data Mining", IEEE(2016).
[2]. https://www.researchgate.net/publication/335911816_Air_Quality_Prediction_using_Machine_Learning_Algorithms.
[3]. https://www.analyticsvidhya.com/blog/2016/10/complete-study-of-factors-contributing-to-air-pollution/.
[4]. http://ccs1.hnue.edu.vn/hungtd/DM2012/DataMining_BOOK.pdf