LITERATURE SURVEY

MANASA R, ANUSHA M, ANANYA M GUPTA, RITHIKA REDDY KARA

[1]Air pollution occurs when harmful or excessive quantities of substances including gases, particles, and biological molecules are introduced into the Earth's atmosphere. Air pollution in India is a serious issue, ranking higher than smoking, high blood pressure, child and maternal malnutrition, and risk factors for diabetes. The main focus of this project is learning about analysis of data by modelling using supervised algorithms i.e Linear Regression (regression), Logistic Regression, Random Forest Classifier, Decision Tree Classifier (classification). The main focus of this project is the Air Quality Index, and factors that affect AQI i.e so2, no2, spm, rspm. In this project a basic data analysis is done on India Air Quality data and the value of Air Quality Index based on given features of concentration of sulphur dioxide, nitrogen dioxide, respirable suspended particulate matter, suspended particulate matter were predicted and classification of the Air Quality and the analysis on whether it was good enough or poor was done.

[2]This paper mainly focuses on predicting AQI level based on the various pollutants NO2, SO2 and SPM observed in a region. This project was done to compare one algorithm with other algorithms and conclude on which is the better one. The Optimized Bayesian Network algorithm is compared with different algorithms to predict AQI level and analyze air quality based on a data set consisting of daily atmospheric conditions in the country and give an idea of which algorithm is best suited for predicting the future air quality. Experiment results prove that the Optimized

The Bayesian Network algorithm runs with best accuracy. This algorithm helps to predict the AQI level accurately with an accuracy of 99.63%, thus letting the government authorities and Pollution control board to make wise decisions about controlling the pollution in those affected regions.

[3]With this forecasting model, various knowledge about the data is extracted using

various techniques to obtain heavily affected regions in a particular region(cluster). This gives more information and knowledge about the cause and seniority of the pollutants.To calculate a location's Air Quality Index and determine the severity of air pollution in that region. This model is capable of predicting the current data with 97.91% accuracy; it will successfully predict the upcoming air quality index of any particular data within a given region. With this model the AQI can be forecast and it can alert the respected region of the country. Also it’s a progressive learning model capable of tracing back to the particular location that needed attention provided the time series data of every possible region needed attention. The air quality information utilized in this report originates from the Indian air quality checking and investigation stage, and incorporates the normal every day fine particulate issue (PM2.5), inhalable particulate issue (PM10), ozone (O3), CO, SO2, NO2 fixation and air quality record(AQI). In order to predict air quality, pm2\_5 is also an important attribute. The values of this must be recorded in future as these particles are responsible for various health effects including cardiovascular effects such as cardiac arrhythmias and heart attacks, and respiratory effects such as asthma attacks and bronchitis.

[4]The model is used to detect whether the sample is polluted or not. Among all the particulate matter that determine the quality of the air, Particulate matter (PM 2.5) needs more attention. When it’s level is high in the air, it causes serious issues on people’s health. Hence, controlling it by constantly keeping a check on its level in the air is important. Logistic regression is employed to detect whether the data sample is either polluted or not polluted.

Logistic regression is the appropriate regression model to conduct analysis when the dependent variable is dichotomous.

Autoregression is applied on the time series data set to predict the PM2.5 value 7 days prior to the current date.

When compared to other machine learning models applied on the data set, Logistic Regression suits the best for this system and gives higher accuracy.

Classify and distinguish the PM2.5 value generated based on the given sample atmospheric conditions to be polluted or not.

The model can only detect and predict the particulate matter of 2.5 and hence the model is restricted. Non-linear problems can't be solved with this model.

The analysis will help people establish a data source for small localities which are

usually left out in comparison to the large cities.

[5]The model forecasts the air quality of India by using machine learning to predict the air quality index of a given area. Air quality of a particular state or a country is a measure on the effect of pollutants on the respective regions, as per the Indian air quality standard pollutants are indexed in terms of their scale, these air quality indexes indicate the levels of major pollutants in the atmosphere.

By this data analysis we came to know that there are seasonal variations and

trend, in order to reduce these metrics, we resample the data month wise to

predict it month wise. By resampling the data, we can reduce the outlier more

more efficiently than raw data.

The model will successfully predict the upcoming air quality index of any particular data within a given region. Better performance than the standard regression models

Could have used different sources alongside the components that impact its fixation. Sensitive to outliers

Various knowledge about the data has been extracted to obtain heavily affected regions. Also plotted box plots which is very useful in EDA. The AQI formula is applied in order to calculate the AQI by using the linear regression algorithm for a particular year.

[6]The dataset includes a concentration of pollutants and meteorological factors.The dataset has been normalized.

These models provide methods to forecast the air pollution levels so that preventive measures can be taken by the people in order to minimize air pollution.

The accuracy score of the Artificial Neural Network, Support Vector Machine, and Random Forest-based model is 90.4%, 93.5%, and 99.4% respectively. The most efficient algorithm among these three is the Random forest algorithm and gives the highest accuracy.

Used an ensemble model which is better than a single decision tree. It decreases the over-fit.

Very less error since the random tree classifier model is used. The random forest classifier has great accuracy . It can perform both classification and regression.

Random tree classifier algorithm requires much time for training and also requires much computational power.

The goal of the ensemble method is to mix the predictions of several base estimators built with a given learning algorithm to enhance robustness over one estimator.

The accurate prediction helps people plan, decreasing the effects of harmful air pollutants on health and the cost associated and creating a cleaner and healthier environment.

[7]The regulation of environmental pollution has drawn public scrutiny. New Delhi, the capital city of India is one of the most polluted cities in the world. Various studies have been conducted to assess the air pollution trends and its hazardous consequences like one recent study has pointed out that the concentration of pollutants in Delhi is much higher than the permissible limits.Prediction of urban air pollution has become a must-have option for mitigating its negative effects. To forecast the air quality, a variety of machine learning algorithms have been used.

In this paper,they have implemented different classification and regression techniques like Linear Regression, SDG Regression, Random Forest Regression, Decision Tree Regression, Support Vector Regression, Artificial Neural Networks, Gradient Boosting Regression and Adaptive Boosting Regression to forecast the Air Quality Index of major pollutants like PM2.5, PM10, CO, NO2, SO2 and O3.

Considering overall performance, SVR and Neural Networks (MLP) best serve the purpose,but there is an unclear trend and wide fluctuations of air pollutants are also attributed to the emissions from pollution sources like transportation, industrial emissions etc. Those factors need to be considered as well.

The final conclusion is that we can forecast the air quality index using the above implementation of machine learning techniques. This information would assist the appropriate authorities in taking appropriate measures and in providing information to the general public as safety and precautionary measures.

[8]Air pollution has progressively increased in metropolitan areas during the previous few years.Cities like Gurugram, Faisalabad, Delhi, Beijing are few of the world's most polluted cities and have seen a dangerous rise in air pollution levels. Accurate forecasting will allow us to plan ahead of time, reducing the negative impacts on health and the costs connected with them. Local meteorological conditions have a significant impact on air pollution levels.

Generating deterministic models to study air pollutant behavior in environmental science research is often not very accurate because they are complex and need simulation at the molecular interaction level. Here comes machine learning to the rescue with high computing facilities to predict air pollution.The data were obtained from two sources. One contains Air quality data and the other contains Meteorological Data.

They tested the utility of current regression models in the sklearn library to predict air quality index values given prior weather data in this study.

This paper investigates how effective some available prediction models are in predicting the Air Quality Index(AQI) values given some input data, based on the pollution and meteorological information in New Delhi, India. They have performed regression analysis on the dataset;but the limitation of this paper is that accuracy can still be improved way better , their results show which meteorological factors affect the AQI values more and how useful the predictive models are to help in air quality forecasting.

[9]Air pollution is an important issue nowadays, being a factor which influences both human health and activities. There are many different chemical substances that contribute to it. These chemicals come from a variety of sources.SO we use ANN for air quality prediction.

Based on the results of the analysis, the model with neural network structure 7-20-4 outperforms the first model in terms of air quality prediction accuracy and R values. This model generates a much better R, indicating a strong correlation between the targets and predicted outputs.Regression analysis was performed to investigate the correlation between the actual and predicted results based on the value of correlation coefficient, R.This model still needs to be improved in order to provide a more accurate forecast of air quality. accuracy can be improved further.

Artificial neural networks have been used to solve a variety of environmental engineering problems with varying degrees of success. The study's goal is to create a neural network air quality prediction model. A prediction method based on a feed-forward neural network is developed in this study. This study takes into account several parameters such as sulphur dioxide (SO2), carbon monoxide (CO), nitrogen dioxide (NO2), nitric oxide (NO), temperature, relative humidity, and air velocity.

[10]This project analyses the data using 3 different models. Random forest, logistic regression, decision trees. This comparing and contrasting method gives the best possible way to analyze data and know you have the best result which can be obtained.

This method of comparing and contrasting was chosen to choose the best method.If only three methods are picked a clear picture of overall models is not taken into account.

The prediction of the air quality index is done using three different models. With the best model out of the selected 3 chosen, the impacts of the pollution were also considered based on the best model. With many types of regressions being used,a random forest was chosen as the best model.With that, the final air quality list was made.Predicting the air quality index.

[11]It uses the random forest to maintain the highest accuracy.The accuracy is not mentioned quantitatively.The report was based on the prediction of the Air Quality Index. The EDA was done to visualize the data very colorfully including a variety of graphs. Using the Random Forest method was the mode which gave the highest of accuracies.With the various scatter plots used to find the correlation of the inputs, as well as the feature importance was given primary importance. (PCA). The need to avoid the overfitting error is the primary reason to choose the random forest of the model and that makes it really essential and important in the analysis and choosing a right-fit model out of so many available. Along with the MSE, RMSE was also measured, the root mean squared error making it a very important aspect of paying attention to the errors as well.

It predicts the air quality index using the random forest model.

[12]They did not provide the code for how they had obtained the graphs. They just

mentioned the outputs. This report predicts the air quality index using machine learning models

and artificial neural networks for the next one hour. The time limit is really important here, as the checking is done only if required. It also provides the concentration of environmental factors such as pressure and dew drop concentration to gain more insight.This project is applicable enough to be used in various weather stations. The supervised models were the go to models picked here, so that the test data which can be used to test the model can give a better understanding of how the model really works. It predicts the air quality index,

using the artificial neural networks

References:

[1]Indian air quality data analysis , visualisation and prediction,Ashmita Dutta, NIT Meghalaya, 2019

[2]A comparative study on prediction of Indian air quality index using machine learning algorithms.Krishna Chaitanya Atmakuri,K V Prasad.

[3]Indian air quality data analysis,Ketaki patel, Anjali Nair, Shreyas Mavale, Rishikesh Mate, 2021

[4]<https://www.ijert.org/implementation-of-machine-learning-algorithms-for-analysis-and-prediction-of-air-quality> - Implementataion of ML algorithms for analysis and prediction of air quality, Dyuthi Sanjeev

[5]<https://www.ripublication.com/ijaerspl2019/ijaerv14n11spl_34.pdf> - Indian air quality prediction and analysis using ML, Mrs. A. Gnana Soundari (MTech, (PhD)), Mrs. J. Gnana Jeslin M.E

(PhD) and Akshaya A.C

[6]<https://www.researchgate.net/publication/327746379_Detection_and_Prediction_of_Air_Pollution_using_Machine_Learning_Models> - Detection and prediction of air pollution using ML models, Aditya C R, Chandana R Deshmukh, Nayana D K and Praveen Gandhi Vidyavastu

[7]Estimation of Air Pollution in Delhi Using Machine Learning Techniques,Chavi Srivastava;Shyamlin Singh,<https://www.researchgate.net/publication/332430367_Estimation_of_Air_Pollution_in_Delhi_Using_Machine_Learning_Techniques>

[8]Urban Air Quality Prediction Using Regression Analysis,Soubhik Mahanta;T. Ramakrishnudu;Rajat Raj Jha;Niraj Tailor

<https://ieeexplore.ieee.org/abstract/document/8929517/authors#authors>

[9]Air Quality Prediction Using Artificial Neural Network,Suraya Ghazali;Lokman Hakim Ismail;https://www.academia.edu/3062165/Air\_quality\_prediction\_using\_artificial\_neural\_network

[10]Air Quality Prediction by Machine Learning,:Ritik Sharma,Gaurav Shilikumar, Shivam Pissal

[11]Air Quality Index Prediction,Sachit Bandhu, Siddhant Saxena

[12]Air Quality Index Prediction using LSTM, Nithyashree K , S Bhumika,Sahana , Ranjitha V