Air Quality Strategies, Exploratory Data Analysis using Machine Learning

Meka Naga Ramyasri
PG scholar, Department of MCA, DNR College, Bhimavaram, Andhra Pradesh.
A.Naga Raju
(Assistant Professor), Master of Computer Applications, DNR college, Bhimavaram, Andhra Pradesh.

Abstract: Now-a-days world is making tremendous growth with the help of various technologies such as advance computers, self-driving vehicles, and robot based surgeries and many more. This technologies putting adverse effect on environment which is leading to polluting of natural resources such as AIR and WATER. Already environmentalist warn world about global warming which can trigger natural calamities such as Earth quake or tsunamis. Polluted Air or water can further degrade health condition or respiratory disease patients such as asthma and lung infections and this can affect normal people's health. To reduce air pollution government are taking extreme steps of reducing all industrial equipment's which are emitting carbon. Carbon is the main enemy of harming ozone layer which is playing important role in protecting world from dangerous sunlight.The most challenging and mysterious stage of qualitative research is data analysis. As a result, understanding these processes is critical for performing qualitative research as well as reading, comprehending, and interpreting it. Using exploratory data analysis (EDA), data visualization, and data research, this project on the EPA's Air Quality System (AQS) seeks to get comprehensive insights into the air quality measurements and trends across diverse locations and contaminants. In order to better understand the dynamics of air quality, develop policies, and make informed decisions, the research analyzes the public AQS data to seek out patterns, anomalies, and potential links in the data. Using rigorous EDA approaches and compelling visualizations, the project seeks to identify significant findings about pollutant concentrations, temporal fluctuations, spatial distributions, and potential implications for public health and the environment. In this application different ML algorithms are used for air quality analysis. Machine learning techniques like decision tree, random forest and support vector machine are used for air quality analysis.

I. INTRODUCTION

The impact of poor air quality on people is multifaceted, affecting both physical health and overall quality of life. The general problem of poor air quality stems from the presence of pollutants in the atmosphere, often emitted from various sources such as industrial activities, transportation, and natural processes. Air quality is a critical environmental factor that directly influences the well-being of individuals and communities. The negative effects of air pollution extend to climate change, with certain pollutants acting as greenhouse gases that contribute to global warming.

Analysing data for air quality is crucial for several reasons. Firstly, it provides insights into the extent and nature of pollution, helping authorities identify pollution sources and patterns. Data analysis allows for the monitoring of air quality trends over time, enabling early detection of emerging issues and the assessment of the effectiveness of implemented measures. The importance of data analysis in the context of air quality lies in its ability to inform policies, empower communities, and contribute to the overall well-being of both individuals and the planet.

Everything that surrounds us makes up the environment. Due to natural disasters and human activity, the atmosphere is becoming increasingly contaminated; air pollution is one of the significant serious issues. The climate variables include temperature control, relative humidity, wind direction, atmospheric wind speed, and the concentration of air contaminants in ambient air. We experience significantly greater heat because sweat doesn't evaporate into the air when the humidity is higher.

Urbanization is one of the primary causes of air pollution because it increases the amount of transportation infrastructure, which releases more pollutants into the atmosphere. Industrialization is another primary driver of air pollution. Sulphur dioxide (SO2), particulate matter (PM), carbon monoxide (CO), Nitrogen oxide (NO), and others are the primary pollutants. Carbon Monoxide is created due to inadequate oxidation of propellants such as oil, gas and other fuel types. Nitrogen oxide is produced when thermal fuel is ignited; it causes nausea and dizziness; carbon monoxide causes vomiting and headaches; nitrous oxide is formed when thermal fuel is ignited; and smoking produces benzene, which irritates the lungs.

Particulate matter 2.5 micrometres in diameter or smaller have a more significant impact on human health. Action must be taken to reduce air pollution in the environment. The Air Quality Index (AQI) is a tool used to gauge air quality. Air quality was formerly predicted using traditional techniques like statistics and probability. However, those techniques are exceedingly challenging to utilize nowadays. Thanks to technological advancements, gathering information using sensors on air pollutants is now quite simple. A robust analysis is required to evaluate the raw data and find the contaminants.

Machine learning, deep learning, recursive neural networks and convolutional neural network guarantee the accuracy of AQI prediction so that the proper course of action can be taken. Reinforcement learning, unsupervised learning, supervised learning, and the three types of learning algorithms used in machine learning, a branch of artificial intelligence. We used a supervised learning strategy in the research that is being proposed. Numerous algorithms fall under the category of supervised learning algorithms, including Random Forest, Linear Regression, SVM, Nearest Neighbor, kernel SVM and Naïve Bayes. Our method chooses Random Forest because it produces superior results than any other algorithms in accurately predicting air pollution.

II. LITERATURE SURVEY

Monitoring air quality is an essential component of an environmental impact assessment program. It can be done with either direct air samplers or bio monitoring plants. Monitoring air pollution with live creatures provides low-cost information on the nature and quantity of contaminants. Lichens can be used as bio monitoring organisms because of their sluggish growth rate, ability to survive longer than vascular plants, and reliance on atmospheric nutrients. The fact that lichens do not lose their parts and thus remain a repository of knowledge makes them an extremely important category of plants compared to other vascular plants. Lichen bio monitoring in a diversified and large geographic area of India can be a cost-effective way to monitor the air quality of such a large area. [1]

In developing countries, transitioning to green energy transportation systems—most notably electric vehicles—is essential to reducing global warming and improving the quality of urban air. Given the detrimental effects of urban air quality on health, it is imperative to accurately estimate pollutant levels and implement emission reduction techniques in order to maintain public health. This report notes that the adoption of electric vehicles is expected to increase from less than 1% to 10% in three years, and it projects the impact of green energy transport systems on the quality of the air in Lahore and Islamabad, Pakistan. This work is novel in two ways. First, air quality index (AQI) changes were predicted using remote sensing data from the Sentinel-5P satellite prior to, during, and following COVID-19. Second, the amounts of NO2, SO2, and CO in the atmosphere were predicted using deep learning models, such as long short-term memory (LSTM) and bidirectional LSTM, and machine learning models, such as decision trees and random forest regression. findings show that installing green energy transport systems can improve air quality in developing countries by over 98%. [3]

One of the main environmental issues in a smart city environment is how to deal with air pollution. The ability to monitor pollution data in real-time allows urban areas to assess the traffic conditions in their area and make appropriate judgments. Different machine learning tools have been employed for pollution prediction in previous research; however, to have a better understanding of these techniques' processing times for many datasets, a comparative comparison of these strategies is frequently necessary. This research presents a comparative analysis of the best model for accurately forecasting air quality concerning data quantity and processing time. They have completed the pollution prediction using four advanced regression approaches. Using a variety of available data sets, we have conducted tests and estimated pollution using Apache Spark. Regression model comparisons have been made using root mean square error (RMSE) and mean absolute error (MAE) as evaluation metrics. [4]

The analysis of air quality and continuous monitoring of air pollution levels are essential topics of study for environmental science. There is an actual effect this issue has on people's health and quality of life. Determining the circumstances that lead to high concentrations of pollutants and, more importantly, promptly predicting such occurrences are extremely important since they make it easier for civil protection to impose targeted preventive and protective measures. The innovative threefold intelligent hybrid system of coupled machine learning algorithms, known as HISYCOL (henceforth), is the subject of this research study. It first addresses the relationship between the circumstances that lead to the emergence of excessive concentrations of contaminants. Conversely, it suggests and displays an ensemble system that forecasts air pollution levels through the combination of machine learning algorithms. The utilization of clustered datasets in this modeling endeavor is crucial and contributes to its hybrid nature. Additionally, by clustering the data vectors and tracing hidden knowledge using unsupervised machine learning, this method increases the accuracy of existing forecasting models. To forecast the concentrations of each air pollutant even more accurately, the item finally uses the Mamdani fuzzy inference method.[5]

A method based on the Internet of Things (IoT) has been developed to monitor the level of noise and air quality inside a designated area. Accepting the required steps is necessary to manage and appropriately screen this situation. The Internet of Things enables data to be transferred as a variable or parameter between a device and the Internet. This approach works well for environmental monitoring because of its increased technological influence. Wi-Fi is used to help define a framework to structure IoT sound and air pollution observation. Cloud-based data access is then set up for periodic storage and access. [6]

To enhance air quality, which can enhance a city's quality of life, it is necessary to utilize the environmental data that has been gathered for the creation of smart cities. Because air pollutants are so dangerous, air pollution is becoming an important concern for society. Pollutants hurt health and can lead to heart and respiratory issues. It could be harmful if air pollution levels rise above acceptable levels. It is possible to design software or a tool whose output can be used to monitor and control air pollution, forecast levels of air pollution, and anticipate health issues associated with air pollution. This work focuses on air analysis using data on different air contaminants, including NO2, SO2, CO, and O3. The dataset, which includes air pollutants and their accompanying AQI values, can be downloaded from the Kaggle website. To anticipate the health issue, this paper uses the Decision Tree J48 method and Naive Bayes. [7]

A database is used to store the monitored data, along with a server and wireless sensor nodes, to enable smart air pollution monitoring. In an air pollution monitoring system, gas sensors provide huge amounts of data. The amount of data to be processed and analyzed is too large for traditional methods to handle. Using data mining techniques, the heterogeneous data are transformed into information that may be used to make decisions. One of the most popular clustering techniques in data mining for organizing large amounts of data is the K-Means algorithm. This work proposes an improved K-means clustering algorithm for the analysis of air pollution data. The real-time monitored pollutant datasets are used to compute the correlation coefficient. The correlation coefficient is used to generate the Air Quality Index (AQI) value, which indicates the degree of air pollution in a given area. In terms of accuracy and execution time, the suggested improved K-Means clustering algorithm is contrasted with the Possibilistic Fuzzy C-Means (PFCM) clustering algorithm. [8]

III. PROPOSED METHOD

Provide an overview of the importance of air quality analysis and its impact on public health, the environment, and quality of life. Clean air is a fundamental component of a healthy and sustainable environment. To address air quality concerns and develop effective strategies for its improvement, it is essential to employ a systematic and data-driven approach. This methodology outlines the framework for conducting Exploratory Data Analysis (EDA) and Visualization for the development of Effective Air Quality Strategies (AQS EDA and Visualization).

3.2 Methodology

We discuss data selection, pre-processing, analysis, modelling, and evaluation in the proposed methodology. Specific machine learning algorithms are utilized here to predict the air quality and to select the process with the highest accuracy. The algorithms are SVM, random forests, and decision trees. For the implementation Python code is used.

3.2.1 **Data Selection**

The data refers to the Environmental Protection Agency (EPA) constructs trends related to air quality by using data gathered from monitoring stations situated throughout the country. This data is generated in Kaggle. These measurements are sourced from the EPA's Air Quality System (AQS). Various entities responsible for collecting environmental data submit their information to the EPA through this system.

For predicting, air quality the selection is a necessary step. In the dataset, we have the county code, parameter name, sample duration, pollutant standard, year, datum, method name, units of measure, event type, metric use, observation count, observation percent, completeness indicator, valid day count, required day count, exceptional data count, null data count and valid day count and many other factors of people taken for pre-processing. The foundation of this methodology is the collection of comprehensive air quality data from various sources. It is crucial to ensure data accuracy, reliability, and relevance. With that values and using the above dataset, we are training SVM, Random forest and DT algorithms and then comparing their performance in terms of accuracy.

3.2.2 Data Pre-processing

We split dataset into train and test part where application used for training 80% dataset and for testing 20%. Before any analysis can take place, the collected data must undergo thorough pre-processing. This includes data cleaning to handle missing values, outliers, and data inconsistencies.

3.2.3. Data transformation

After pre-processing the data, we have to perform the data transformation. Data transformation is a crucial step in the analysis of air quality data. Data transformation is essential to ensure that the dataset is well-prepared for subsequent exploratory data analysis and visualization. A clean, organized, and appropriately transformed dataset facilitates more meaningful insights into air quality trends, patterns, and potential strategies for improvement.

It involves manipulating and modifying the raw air quality data to make it more suitable for analysis and visualization. It involves converting, cleaning, and structuring raw data into a format that is suitable for analysis and visualization. Data transformation is a critical component of preparing air quality data for effective analysis and decision-making. The specific steps and techniques used will depend on the nature of the data and the objectives of the analysis.

Identify and handle missing or incomplete data points. This may involve imputing missing values using statistical methods or removing rows with missing data to ensure the dataset is complete. Cleanse the data by addressing outliers, errors, or inconsistencies. This could include correcting data entry mistakes, validating sensor readings, and addressing extreme values that may skew the analysis.

Normalize or standardize numerical variables to bring them to a common scale. This is important when dealing with features that have different units or magnitudes, ensuring that each variable contributes proportionately to the analysis. Create new features or variables that may enhance the analysis. Convert data types to ensure compatibility with analysis tools and algorithms.

3.2.4. Modelling

Modelling air quality is a complex process that involves the use of mathematical, statistical, and computational techniques to simulate, predict, and analyze the concentration and distribution of air pollutants in the atmosphere. Air quality models are essential tools for understanding pollution sources, assessing environmental impacts, and developing effective strategies for air quality management.

3.2.5 Evaluation

Evaluating air quality using Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) can be a valuable approach, especially when assessing the performance of air quality models or predictions. MSE and RMSE are valuable metrics for evaluating air quality models, they should be considered alongside other relevant metrics, domain-specific knowledge, and environmental regulations to make comprehensive assessments of air quality and to guide policy decisions and public health initiatives.

3.2.6 **Logical Flow of System**

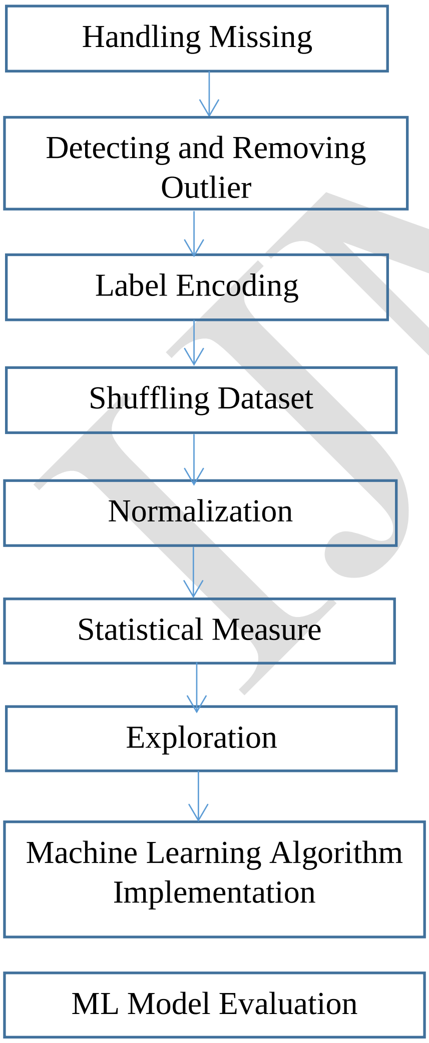

Fig.1 Flowchart

IV. RESULT

We have implemented this work using JUPYTER notebook and below are the code and output screens with blue color comments,

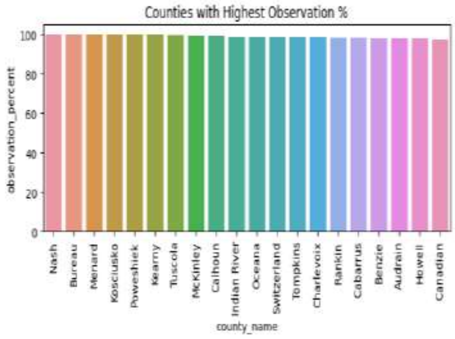

Fig.2 Countries with highest observation percentage

In above screen we are finding and plotting graph of to 20 counties with highest number of Air Quality Observation %. In graph x-axis represents county name and y-axis represents air quality observation %

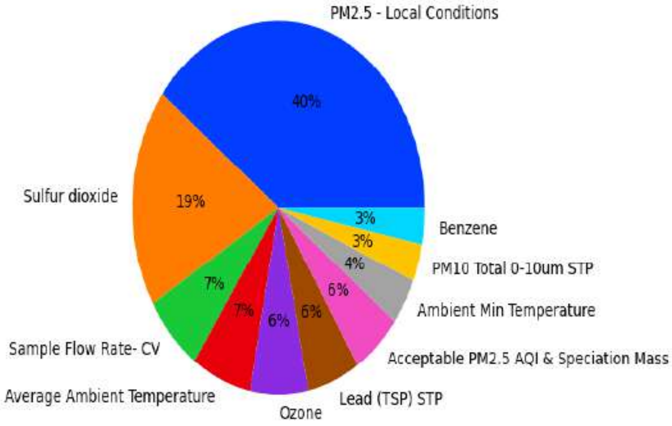

Fig.3 PM 2.5 – Local conditions

In above screen finding and plotting graph of top 10 pollutants found in air quality

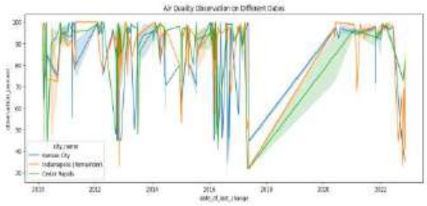

Fig.4 Air quality observation on different dates

In above screen finding and plotting graph of different cities of air quality observation where x-axis represents Date and y-axis represents air quality and different line represents different cities and by using this graph we can see which year has highest or lowest observation for each city

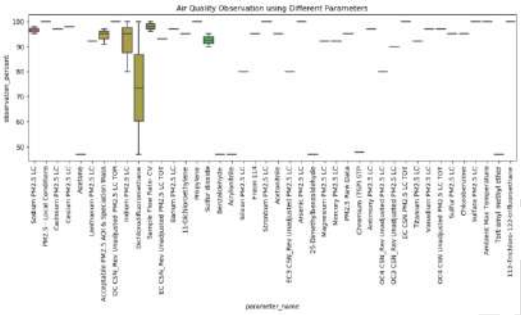

Fig.5 Air quality observation with different parameters

In above box plot graph finding different air pollutants found in dataset where x-axis represents pollutant name and y-axis represents observation

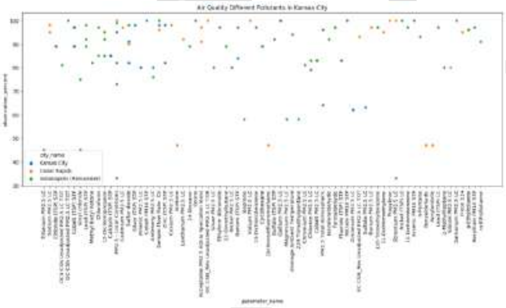

Fig.6 Air quality different pollutants in Kansas city

In above scatter plot also we can see air quality observation % in different cities where different dots represents different cities. X-axis contains pollutant name and y-axis contains air quality observation values

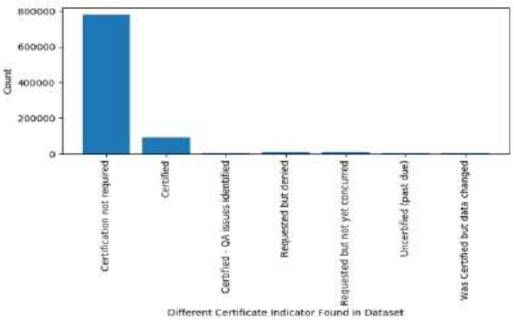

Fig.7Different certificate indicator found in dataset

In above screen finding and plotting graph of different certification available in Hawaii where x-axis represents certificate names and y-axis represents count of applied certificates

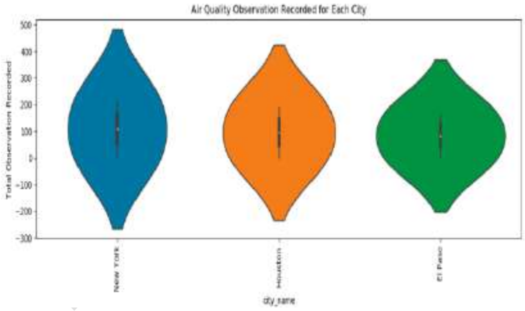

Fig.8 Air Quality observation recorded found in dataset

In above graph plotting cities with highest number of air quality pollutant observation where x-axis represents city name and y-axis represents observation

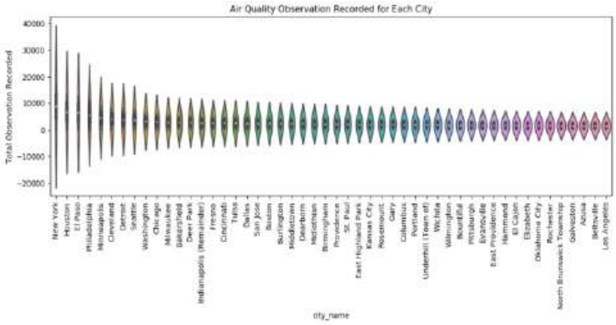

Fig.9 Air quality analysis recorded for each city

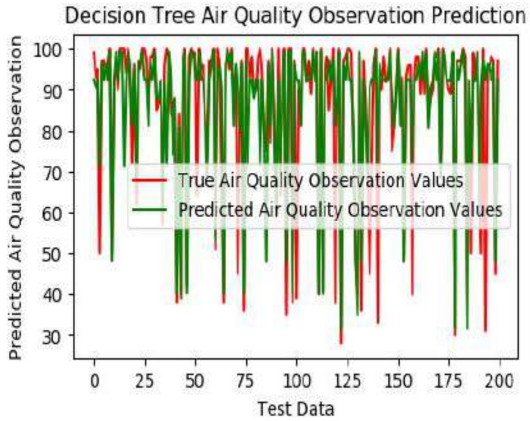

Fig.10 DT Air quality observation prediction

In above graph x-axis represents test number of records and y-axis represents observed Air Quality where RED line represents TEST data observed air quality and green line represents Predicted air quality and in above graph we can see nearly 95% both lines are overlapping so we can see predicted and original observed values are too close

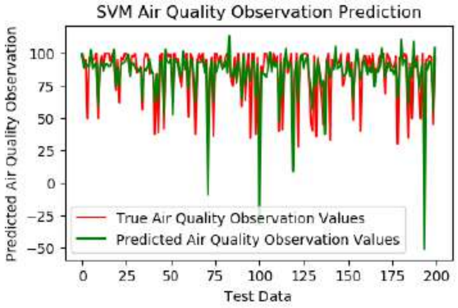

Fig.11 SVM Air quality observation prediction

In above graph both lines are not fully overlapping so SVM predictions are not accurate or closed with the test values

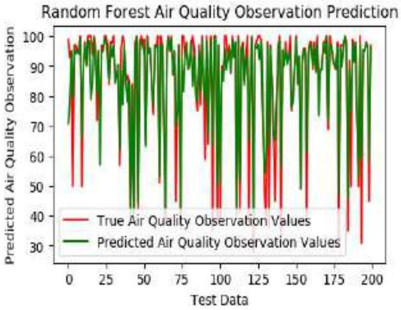

Fig.12 RF Air quality observation prediction

Above is the random Forest predicted graph

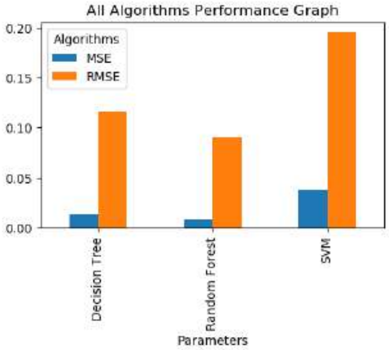

Fig.13 All algorithms performance graph

In above screen displaying MSE and RMSE graph of each algorithm where x-axis represents algorithm names and y-axis represents MSE and RMSE values in different color bars and in all algorithms Random Forest and decision tree got less MSE and RMSE values so they are good in performance

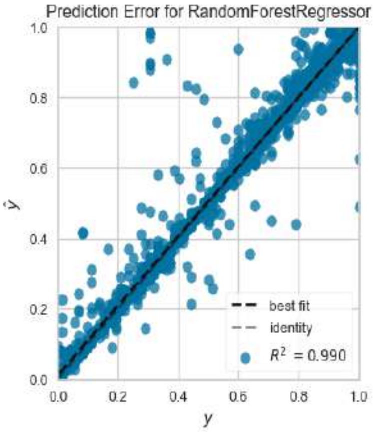

Fig.14 Prediction error for Random forest regressor

In above screen using YELLOW BRICK we are finding Prediction ERROR for Random Forest where Best FIT and Identify lines should be fully overlap if algorithm predictions are closed to test data and in above graph RSQUARE value is 0.98 for Random Forest

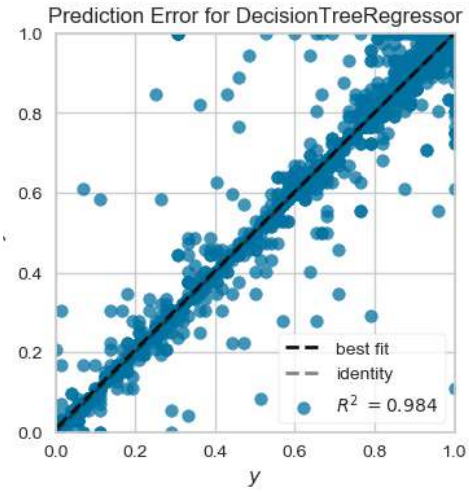

Fig.15 Prediction error for Decision Tree Regressor

In above is the decision tree Prediction Error graph and here also both lines are fully overlapping with 0.97% as the RSQUARED

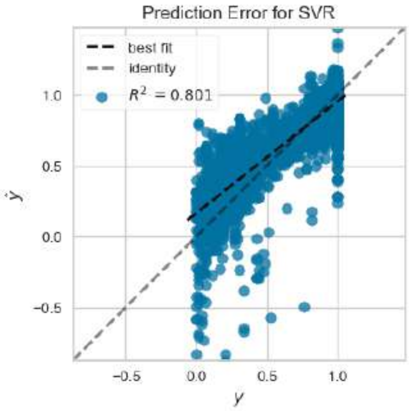

Fig.16 Prediction error for SVR

In above SVM graph both lines are not fully overlapping so we can say SVM predicted values and Test values are not close to each other and its RSQUAARE is 0.69% which is not good and consider to be worst performer.

V. CONCLUSION AND FUTURE SCOPE

As we conclude this research, it is evident that the transformation of AQS data into actionable intelligence is more than an academic exercise; it is a commitment to a healthier, cleaner world. We are employing YELLOWBRICKS visualizer which will visualize how best the algorithm fit with the train and test prediction values. In propose work we are employing Yellow Brick to visualize best fit with error prediction with RSQUARE metrics. RSQUARE will comes in range between 0 and 1 where any value closer to 0 will consider model performance as worst and the value closer to1 will be consider as BEST. In propose application we have tried to work with SVM, Random Forest and Decision Tree where SVM giving RSQUARE as 0.69, Random Forest as 0.98 and Decision Tree as 0.97. The impact extends to public awareness and engagement, fostering a sense of shared responsibility in the fight against air pollution. The future scope for air quality involves addressing a range of challenges and opportunities related to air pollution and environmental sustainability. The future of air quality management is closely tied to sustainable development and improving the quality of life for people around the world. As environmental awareness grows, addressing air pollution and promoting clean air will remain a global priority.

.

REFERENCES

[1] [US ENVIRONMENTAL PROTECTION AGENCY]. [Air Quality Annual Summary], Retrieved from[https://www.kaggle.com/datasets/epa/air-quality].

[2] Chakraborty, P.; Tandon, N.; Bajpai, R. Monitoring of air quality in an urban area of India using lichens. Environ. Monit. Assess. 2000, 64, 513–525.

[3] Moretti, M.; Becagli, S.; Cappelletti, F. A multi-disciplinary study of air quality in Florence, Italy. Atmos. Environ. 2010, 44, 2701–2711.

[4] S. Ameer, M. A. Shah, A. Khan, H. Song, C. Maple, S. U. Islam, and M. N. Asghar, ''Comparative analysis of machine learning techniques for predicting air quality in smart cities,'' IEEE Access, vol. 7, pp. 128325–128338, 2019.

[5] I. Bougoudis, K. Demertzis, and L. Iliadis, ''HISYCOL a hybrid computational intelligence system for combined machine learning: The case of air pollution modeling in Athens,'' Neural Compute. Appl., vol. 27, no. 5, pp. 1191–1206, Jul. 2016.

[6] D. Ganeshkumar, ''Air and sound pollution monitoring system using cloud computing,'' Int. J. Eng. Res., vol. V9, no. 6, Jun. 2020.

[7] R. W. Gore and D. S. Deshpande, ''An approach for classification of health risks based on air quality levels,'' in Proc. 1st Int. Conf. Intell. Syst. Inf. Manage. (ICISIM), Oct. 2017, pp. 58–61.

[8] G. R. Kingsy, R. Manimegalai, D. M. S. Geetha, S. Rajathi, K. Usha, and B. N. Raabiathul, ''Air pollution analysis using enhanced K-means clustering algorithm for real time sensor

data,'' in Proc. IEEE Region 10 Conf. (TENCON), Nov. 2016, pp. 1945–1949.

[9] Jakkula, Vikramaditya. "Tutorial on support vector machine (svm)." *School of EECS, Washington State University* 37.2.5 (2006): 3.

[10] De Ville, Barry. "Decision trees." Wiley Interdisciplinary Reviews: Computational Statistics 5.6 (2013): 448-455.

[11] Chai, Tianfeng, and Roland R. Draxler. "Root mean square error (RMSE) or mean absolute error (MAE)." Geoscientific model development discussions 7.1 (2014): 1525,1534

[12] Bao, R., Zhang, A., 2020. Does lockdown reduce air pollution? Evidence from 44 cities in northern China. Sci. Total Environ. 731, 139052

[13] Bekkar, A., Hssina, B., Douzi, S., Douzi, K., 2021. Air-pollution prediction in smart city, deep learning approach. J. Big Data 8, 1–21. https://doi.org/10.1186/S40537-021- 00548-1/FIGURES/17.