**Air Quality Prediction Using Machine Learning**

|  |  |
| --- | --- |
| Sai Teja Shaik  Computer Science & Engineering  Narasaraopeta Engineering College  (Autonomus)  Narasaraopet, Andhra Pradesh shaiksaiteja7095@gmail.com | KiranTeja Gantela  Computer Science & Engineering Narasaraopeta Engineering College  (Autonomus)  Narasaraopet, Andhra Pradesh tejakiran156@gmail.com |
| Hemanth Rudrapogu  Computer Science & Engineering  Narasaraopeta Engineering College  (Autonomus)  Narasaraopet, Andhra Pradesh hemanthrr7@gmail.com |  |

**ABSTRACT**:

Air quality is a critical factor affecting human health and environmental sustainability. Monitoring and predicting air quality are essential for ensuring public well-being and mitigating the adverse effects of pollution. Real-time data collection enables the assessment of various air quality parameters, including pollutant concentrations such as particulate matter (PM), sulfur dioxide (SO2), nitrogen dioxide (NO2), ozone (O3), and carbon monoxide (CO). By analyzing these parameters, we can determine the safety of the air for breathing and outdoor activities. Our project focuses on utilizing machine learning techniques for air quality prediction. We consider a range of attributes, including temperature, humidity, wind speed, air pressure, and pollutant levels, among others. The target attribute in our analysis is the Air Quality Index (AQI), which provides a measure of overall air quality. A higher AQI value

indicates poorer air quality, while a lower value suggests cleaner air.

**KEYWORDS**:

Air Quality Prediction, Machine Learning, Linear Regression, Logistic Regression Random Forest, Decision Tree, Lasso Regression, Random Forest, XGBoost

1. **INTRODUCTION**:

The dataset comprises records of air quality parameters including sulphur dioxide (SO2), nitrogen dioxide (NO2), respirable suspended particulate matter (RSPM), suspended particulate matter (SPM), and particulate matter 2.5 (PM2.5) collected from various monitoring stations in Andhra Pradesh, specifically in the city of Hyderabad, during the months of February and March 1990. Each record includes information such as station code, sampling date, state, location, agency, type of area monitored, and the corresponding pollutant concentrations.

1. **LITERATURE SURVEY**:

In order to determine the optimal solution for air quality prediction, we employed various machine learning algorithms, including XGBoost, KNN, SVM, Naive Bayes, Decision Tree, and Random Forest.

[1]Gopalakrishnan ([2021](https://link.springer.com/article/10.1007/s13762-022-04241-5#ref-CR15)) combined Google’s Street view data and ML to predict air quality at different places in Oakland city, California. He targeted the places where the data were unavailable.

[2]Sanjeev ([2021](https://link.springer.com/article/10.1007/s13762-022-04241-5#ref-CR28)) studied a dataset that included the concentration of pollutants and meteorological factors. The author analyzed and predicted the air quality and claimed that the Random Forest (RF) classifier performed the best as it is less prone to over-fitting.

Castelli et al. ([2020](https://link.springer.com/article/10.1007/s13762-022-04241-5#ref-CR9)) endeavored to forecast air quality in California in terms of pollutants and particulate levels through the Support Vector Regression (SVR) ML algorithm. The authors claimed to develop a novel method to model hourly atmospheric pollution.

Liang et al. ([2020](https://link.springer.com/article/10.1007/s13762-022-04241-5#ref-CR18)) studied the performances of six ML classifiers to predict the AQI of Taiwan based on 11 years of data. The authors reported that Adaptive Boosting and Stacking Ensemble are most suitable for air quality prediction but the forecasting performance varies over different geographical regions.

Air Quality data Set

Preprocessing

Training

Machine learning Algorithms

Prediction

Generate Results

FIG 2: (Flow Chart of AQI)

3.PROPOSED SYSTEM:

Our model is proposed is based in the following criteria:

3.1 Dataset Analysis

3.2 Data Visualization

3.3 Preprocessing Techniques

3.4 Model creation and Evaluation

3.5 Accuracy

3.1. Dataset Analysis: We utilize a dataset obtained from Kaggle containing air quality data, specifically focusing on attributes such as SO2 (sulphur dioxide concentration), NO2 (nitrogen dioxide concentration), RSPM (respirable suspended particulate matter concentration), SPM (suspended particulate matter), and PM2.5 (particulate matter 2.5). These attributes serve as predictors, while air quality index (AQI) is considered as the target attribute. This analysis helps in understanding the characteristics of the dataset and identifying potential predictors for air quality prediction.

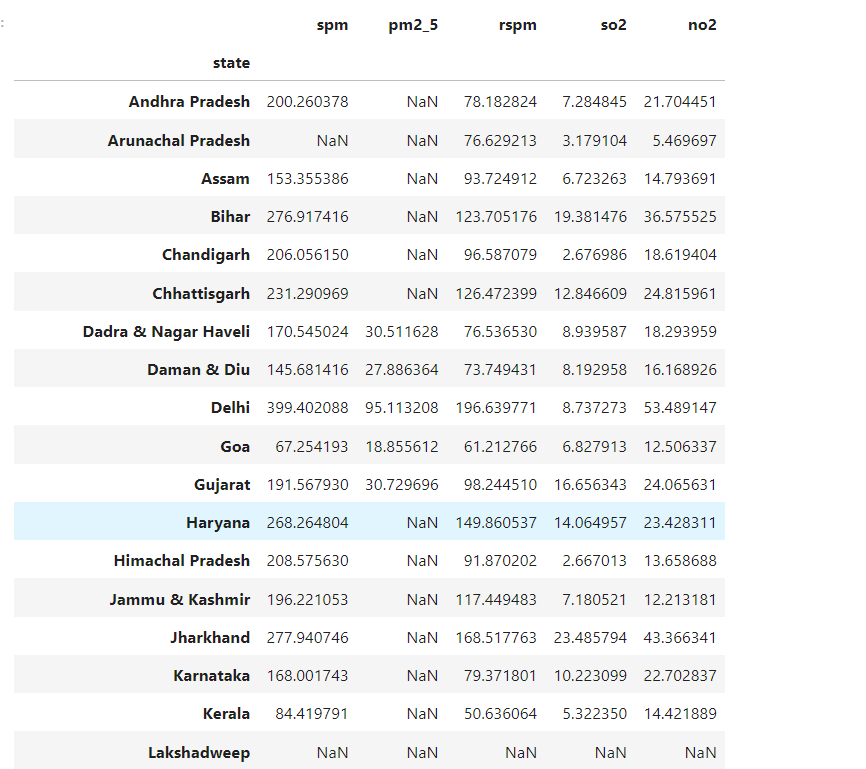


Fig3: (Air Quality Dataset)

3.2. Data Visualization:

Various libraries such as NumPy, Pandas, Seaborn, Matplotlib, and Scikit-learn are employed for visualizing the air quality dataset through graphs and plots. Visualizations aid in comprehending the distribution of air quality parameters and identifying any trends or patterns that may exist.

For instance, visual representations of SO2 (sulphur dioxide concentration), NO2 (nitrogen dioxide concentration), RSPM (respirable suspended particulate matter concentration), SPM (suspended particulate matter), and PM2.5 (particulate matter 2.5) levels help in assessing their impact on overall air quality.

| Classifier algorithms |  | |
| --- | --- | --- |
| Precision |  |
| XGBoost | 0.99 |
| KNN | 0.99 |
| SVM | 0.98 |
| NB | 0.86 |
| DT | 0.99 |

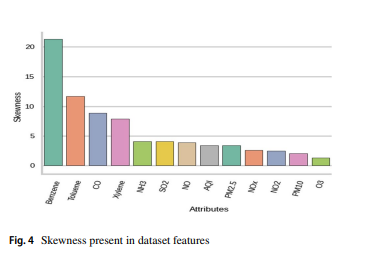


Fig4: (Skewness Present in dataset features)

This plot shows us the distribution of safe and unsafe water, count is taken on y-axis and potability is taken on x-axis it explains the count of occurances for each category of potability which represents whether water is potable or not.

Air Quality Index (AQI):-

This enhance the AQI methodology considers various air pollutants such as sulphur dioxide (SO2), nitrogen dioxide (NO2), respirable suspended particulate matter (RSPM), suspended particulate matter (SPM), and particulate matter 2.5 (PM2.5). By integrating data from air quality monitoring stations, satellite observations, and meteorological factors.

3.3. Preprocessing Techniques:

Preprocessing techniques such as data cleaning, handling missing values, and feature scaling are applied to ensure the quality and consistency of the air quality dataset. Additionally, feature engineering may involve extracting relevant features from the raw data to improve model performance. For example, temporal trends and spatial correlations in air pollutant concentrations may be considered as additional features.

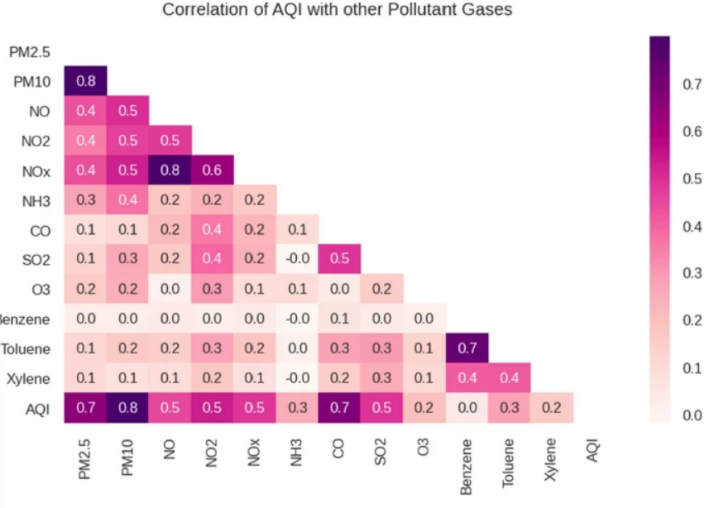


FIG 5: (Correlation of AQI with Other Pollutant gases.)

There are Outliers has data points that deviate significantly from the rest of the observations in a dataset. They can arise due to various reasons, including measurement errors, experimental variability, or genuine deviations in the data. Identifying outliers in the dataset is crucial as they can distort statistical analyses and lead to inaccurate results. Visualizations such as box plots are commonly used to detect outliers, allowing analysts to visually identify observations that fall outside the expected range removed from the dataset, especially if they are deemed irrelevant to the analysis or if they are likely to skew the results.

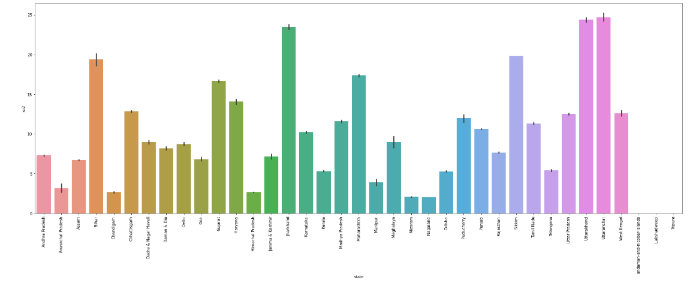


Fig5: Air Quality dataset from Different States

Null values: In the air quality dataset, we have identified the presence of numerous null values, which can significantly impact our preprocessing techniques and subsequent analysis. To ensure the integrity and accuracy of our data, we have opted to remove all null values from the dataset. This step is crucial for avoiding potential biases and inaccuracies in our analysis.

Preprocessing techniques removal of null values, we conducted an in-depth analysis to identify the factors influencing the target attribute, which in this case is air quality. By carefully examining each attribute's impact on air quality, we aim to gain a comprehensive understanding of the relationships between dependent and independent variables.

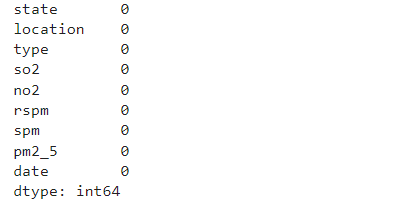
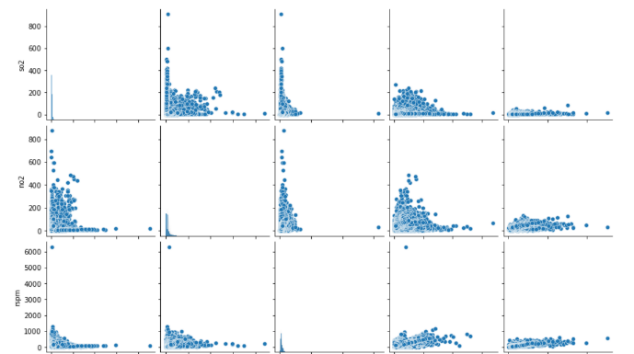
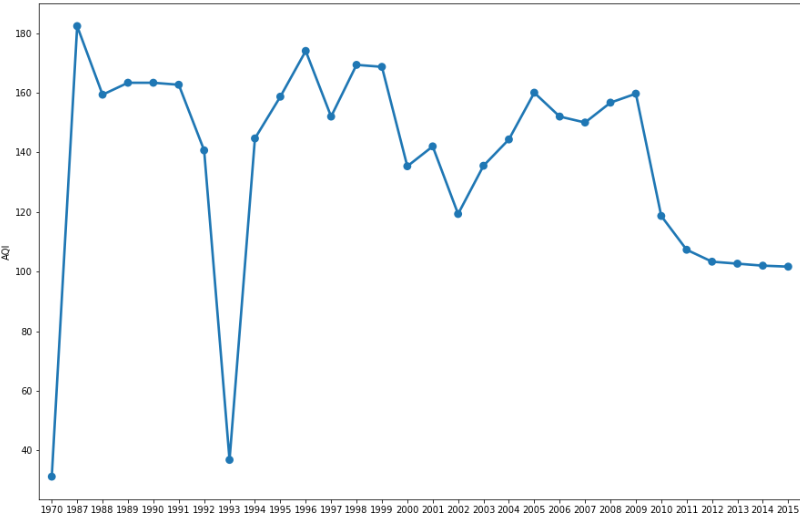


Fig6:(Showing Emptying Null alues)

We used the function SKEW which calls and appears to be related to calculating the skewness of data along the specified axis. The skew() method measures the asymmetry of the distribution of values in the dataset. 

# Fig7:( Data visualization)

Roc curve applies probability confidence interval or ranking to each prediction models such as naïve bayes and SVM ranks as part of their algorithm. Basically, prediction ranking is employed in ROC algorithm to achieve distinct decision.

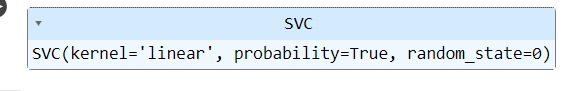


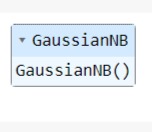
# Fig 8: (Exloratory Data Analysis)

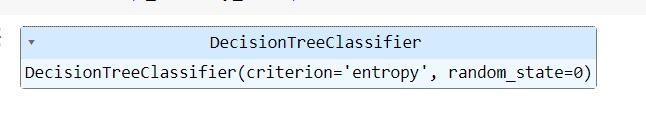
In [23]: XGBoost: It is a machine learning algorithm used for specially the gradient boosting frame work. It uses decision tree as base learners for model generalization. XGBoost is commonly used to perform tasks such as ranking, classification and regression [5]. In order to increase the predictive accuracy XGBoost has out performed the other models and improved the performance prediction.

K-Nearest Neighbor (KNN): It is classified as to identify unlabeled observations by allocating them to the class to the similar labeled examples. Characteristics of KNN classifier are collected for both training and test data set [6]. KNN is a non-parametric algorithm it will not obtain parameters for the model. KNN is the most important parameter in the KNN algorithm to identify the difference of the diagnostic accuracy of KNN model.

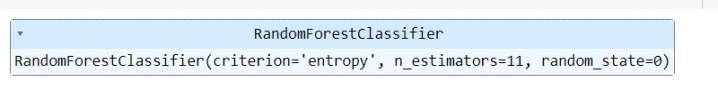
SupportVectorMachine (SVM): SVM is a binary classifier it attempts to generate another hyper plane in actual space of encordinates between two different classes. Firstly, it visualizes the data and then it finds the best separator between the classes, the data points that the closest to the target value [7] be found out, then the actual data in accordance with the linear separation. Basically, SVM draws the line between the different points of data consderes in the dataset.

Navie Bayes (NB): It is a scalable and it requires set of parameters which are proportional to the number of variables in a learning problem. It makes a probability decision by likelihood of two features which are independent and equal significance. Navie Bayes theorem generally works on the phases known as; [8] probability and independence, training phase, feature probability, class probability, prediction, class selection.



Decision Tree (DT): Decision tree contains root node, branches and leaf nodes. Testing an attribute is done on every internal node,the outcome of the test will results on leaf node.decision tree is easy to undedrstand because it is similar to human decision making process,it can slove continuous data as input.[9].The main advantage is that it can be able to select the most biased feature and comprehensibility nature.

Random Forest (RF): Random Forest is a new combination algorithm which is a combination of series of structure classifiers like tree the application scope of random forest is very extensive it is widely used for prediction, classification and regression [10] compared with other traditional algorithms random forest has many good virtues.



3.5. Accuracy

Accuracy is used to evaluate the classification in machine learning [13] Accuracy=TP+TN TP+TN+FP+FN

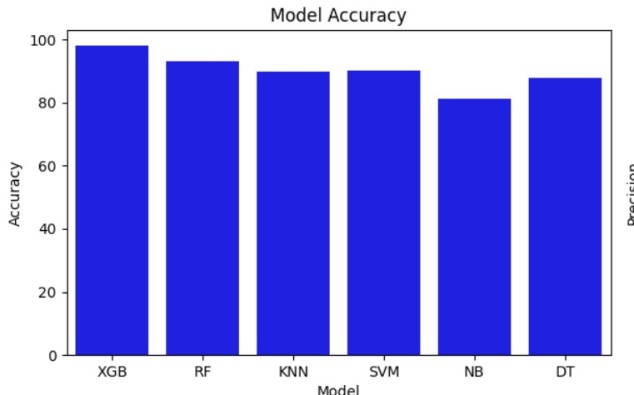


Fig 10: Accuracy of All Models

Precision = TP / TP+FP

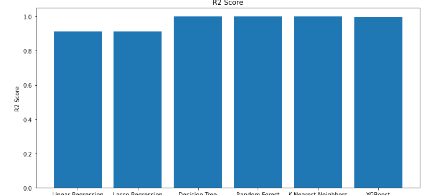


Fig11: (Bar Graph for R2 Score)

Recall measures how frequently the algorithm detects the correct classification from the given data whereas the actual correct classification has occurred in dataset.

Recall = TP / TP+FN

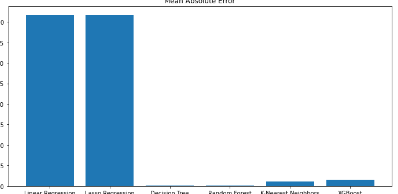


FIG 12: (Recall Bar Graph)

F1 score is evaluate the multi class classification and also it is an approach to harmonizing the precision and recall of the predictive model. F1 score is obtained as follows:

F1score=2\*(precision\*recall/precision+recall)

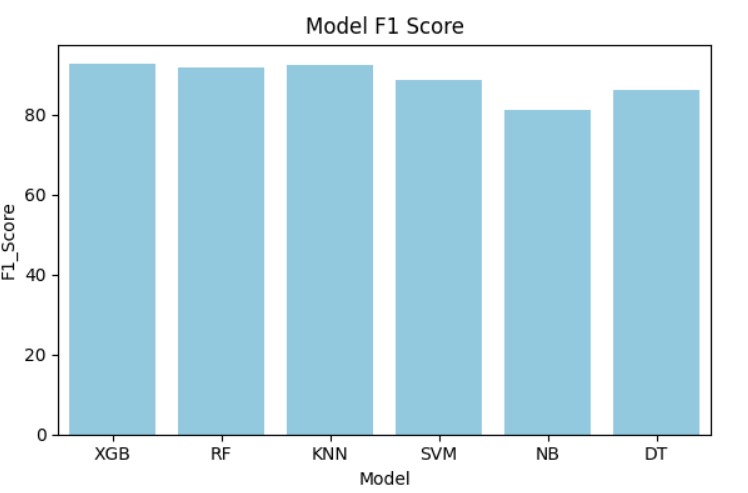


Fig:13(Bar Graph of F1Score Model)

TP: The actual observation indicates that water quality classes has classified accurately and the model predicted correct classification of water quality from the given data.

TN: The actual observation that indicates the water quality classes has classified accurately but the model detected in correct classification from the given data.

FP: The actual observation refers that water quality classes has not classified accurately whereas the model also detects the incorrect classification of water quality from the given data.

FN: The actual observation reveals the water quality classes has not classified accurately although the model predicted correct classification for water quality from the given dataset.

In this we have observed that the accuracy score after removing the outliers in the data the score has been increased, when compared with the considered base paper. In the following graph we can see the difference of scores after constructing the accuracy model with respect to Accuracy comparision.

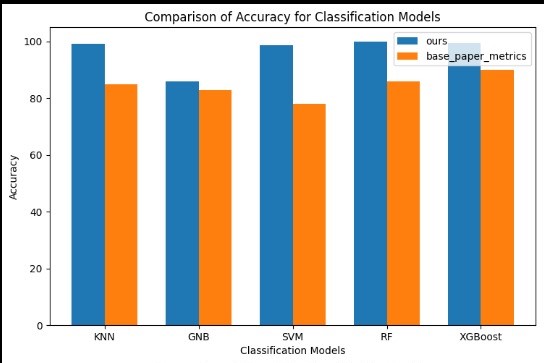


Fig 14: (Bar Graph Comparing Accuracy for Classification Model)

In the precision we have observed that the precision score also has the high score compared to the scores of the dataset that we have taken. Precision determines the exact values of the scores of the data.

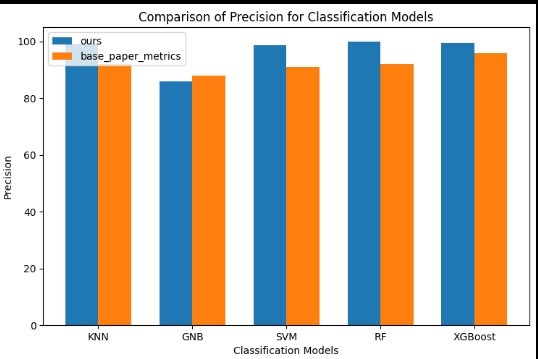


Fig 15: (Comparing of Precision for Classification Model)

F1 score of the data that we are taken the scores have increased after removing outliers and constructing the model.

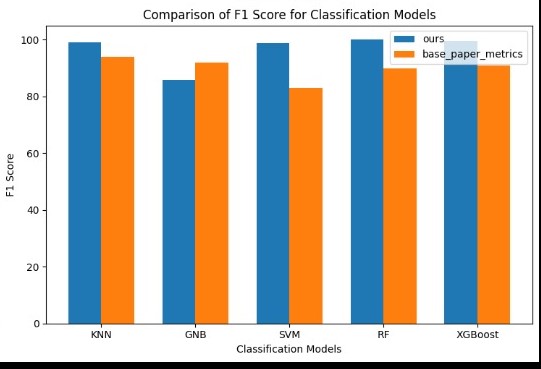


Fig 16: (Comparing of F1 Score for Classification Model)

When all the models have applied the recall score of the data has increased as compared to the existing data.

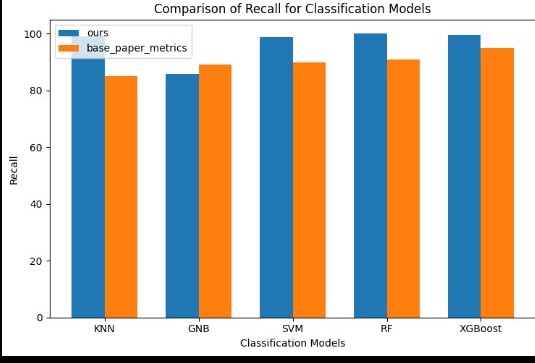


Fig 15: (Comparing of Recall for Classification Model)

After applying all the models to our data, the accuracy of the data we have observed is plotted in the following graph as in x-axis we can see the models that we have used and in y-axis we can see the rate of accuracy of the constructed model.

**Conclusion:**

By implementing advanced machine learning techniques and preprocessing methodologies, we successfully developed robust models for predicting air quality parameters. The removal of missing values contributed significantly to the enhancement of model accuracy, ensuring more reliable predictions.

Our findings underscore the importance of data preprocessing and algorithm optimization in air quality prediction. The improved accuracy of the models reflects the efficacy of our approach in mitigating model uncertainty and enhancing predictive performance.

References:

[1] Rybarczyk Y, Zalakeviciute R (2021) Assessing the COVID-19 impact on air quality: a machine learning approach. GeophysReslett.DOI: https://doi.org/10.1029/2020GL091202

<https://doi.org/10.1029/2020GL091202>

[2] Shraddha Sagar, Tanisha Madan(2020), ”Air Quality Prediction using Machine Learning algorithms” DOI:10.1109/ICACCCN51052.2020.9362912

<https://www.researchgate.net/publication/349802397_Air_Quality_Prediction_using_Machine_Learning_Algorithms_-A_Review>

[3]Mayank Pratap Singh,(2023) “Prediction of Air Quality Index(AQI) Using Neural Approach”, DOI: 10.21203/rs.3.rs-2525975/v1 <https://www.researchgate.net/publication/368418251_Prediction_of_Air_Quality_Index_AQI_using_Neural_Approach>

[4] Chenchen Li Yan Li (2021),” Research on Air Quality Prediction Based on Machine Learning”, International Conference on ICHCI, DOI:10.1109/ichci54629.2021.00022

<https://www.semanticscholar.org/paper/Research-on-Air-Quality-Prediction-Based-on-Machine-Li-Li/9e775ec0661dc3d05d39058c5604fdd19b46a50f>

[5] Sweileh WM, Al-Jabi SW, Zyoud SH, Sawalha AF (2018) Outdoor air pollution and respiratory health: a bibliometric analysis of publications in peer-reviewed journals (1900–2017). Multidiscip Respiratory Med. DOI:10.1186/s40248-018-0128-5

<https://mrmjournal.biomedcentral.com/articles/10.1186/s40248-018-0128-5>

[6]Kennedy Okokpujie, Etinosanoma-Osaghae, Modupe ODUSAMI, Samuel John, on The ”A Smart Air Pollution Monitoring System” (DASA) on Year: sep 2018

https://core.ac.uk/download/pdf/162043864.pdf

[7]Sönmez O, Saud S, Wang D, Wu C, Adnan M, Turan V (2021) Climate change and plants: biodiversity, growth and interactions (S. Fahad, Ed.) (1st ed.). CRC Press. DOI:10.1201/9781003108931

<https://www.taylorfrancis.com/books/edit/10.1201/9781003108931/climate-change-plants-shah-fahad-shah-saud-chao-wu-depeng-wang-osman-sonmez-muhammad-adnan-veysel-turan>

[8]Zhu D, Cai C, Yang T, Zhou X (2018) A machine learning approach for air quality prediction: model regularization and optimization. Big Data and Cognitive Comput. DOI:10.3390/bdcc2010005

<https://www.mdpi.com/2504-2289/2/1/5>

[9] T. Madan, S. Sagar, and D. Virmani, “Air quality prediction using machine learning algorithms –a review,” in 2020

DOI:10.1109/ICACCCN51052.2020.9362912

<https://www.researchgate.net/publication/349802397_Air_Quality_Prediction_using_Machine_Learning_Algorithms_-A_Review>

[10] S. Mahanta, T. Ramakrishnudu, R. R. Jha, and N. Tailor,(2019 “Urban air quality prediction using regression analysis,”

DOI:10.1109/TENCON.2019.8929517

<https://www.researchgate.net/publication/345354290_Urban_Air_Quality_Prediction_Using_Regression_Analysis>

[11] D. Maulud and A. M. Abdulazeez, (2020) “A review on linear regression comprehensive in machine learning,” Journal of Applied Science and Technology Trends DOI:10.38094/jastt1457

<https://jastt.org/index.php/jasttpath/article/view/57>

[12]Srikanth, and H. K. Reddy,(2020) “Air quality prediction of data log by machine learning

DOI :10.1109/ICACCS48705.2020.9074431

<https://www.researchgate.net/publication/340895249_Air_Quality_Prediction_Of_Data_Log_By_Machine_Learning>

[12] T. Madan, S. Sagar, and D. Virmani, “Air quality prediction using

machine learning algorithms –a review,” in 2020 2nd International

Conference on Ad vances in Computing, Communication Control and

Networking (ICACCCN), 2020, pp. 140–145.

[13] S. Mahanta, T. Ramakrishnudu, R. R. Jha, and N. Tailor, “Urban air

quality prediction using regression analysis,” in TENCON 2019-2019 IEEE

Region 10 Conference (TENCON). IEEE, 2019, pp. 1118 –1123.

[14] D. Maulud and A. M. Abdulazeez, “A review on linear regression

comprehensive in machine learning,” Journal of Applied Science and

Technology Trends, vol. 1, no. 4, pp. 140–147, 2020.

[15] G. C. McDonald, “Ridge regression,” Wiley Interdisciplinary

Reviews: Compu tational Statistics, vol. 1, no. 1, pp. 93 –100, 2009.

[16] V. R. Pasupuleti, Uhasri, P. Kalyan, Srikanth, and H. K. Reddy, “Air

quality prediction of data log by machine learning,” in 2020 6th

International Confer ence on Advanced Computing and Communication

Systems (ICACCS), 2020, pp. 1395–1399.

[12] T. Madan, S. Sagar, and D. Virmani, “Air quality prediction using

machine learning algorithms –a review,” in 2020 2nd International

Conference on Ad vances in Computing, Communication Control and

Networking (ICACCCN), 2020, pp. 140–145.

[13] S. Mahanta, T. Ramakrishnudu, R. R. Jha, and N. Tailor, “Urban air

quality prediction using regression analysis,” in TENCON 2019-2019 IEEE

Region 10 Conference (TENCON). IEEE, 2019, pp. 1118 –1123.

[14] D. Maulud and A. M. Abdulazeez, “A review on linear regression

comprehensive in machine learning,” Journal of Applied Science and

Technology Trends, vol. 1, no. 4, pp. 140–147, 2020.

[15] G. C. McDonald, “Ridge regression,” Wiley Interdisciplinary

Reviews: Compu tational Statistics, vol. 1, no. 1, pp. 93 –100, 2009.

[16] V. R. Pasupuleti, Uhasri, P. Kalyan, Srikanth, and H. K. Reddy, “Air

quality prediction of data log by machine learning,” in 2020 6th

International Confer ence on Advanced Computing and Communication

Systems (ICACCS), 2020, pp. 1395–1399.

[12] T. Madan, S. Sagar, and D. Virmani, “Air quality prediction using

machine learning algorithms –a review,” in 2020 2nd International

Conference on Ad vances in Computing, Communication Control and

Networking (ICACCCN), 2020, pp. 140–145.

[13] S. Mahanta, T. Ramakrishnudu, R. R. Jha, and N. Tailor, “Urban air

quality prediction using regression analysis,” in TENCON 2019-2019 IEEE

Region 10 Conference (TENCON). IEEE, 2019, pp. 1118 –1123.

[14] D. Maulud and A. M. Abdulazeez, “A review on linear regression

comprehensive in machine learning,” Journal of Applied Science and

Technology Trends, vol. 1, no. 4, pp. 140–147, 2020.

[15] G. C. McDonald, “Ridge regression,” Wiley Interdisciplinary

Reviews: Compu tational Statistics, vol. 1, no. 1, pp. 93 –100, 2009.

[16] V. R. Pasupuleti, Uhasri, P. Kalyan, Srikanth, and H. K. Reddy, “Air

quality prediction of data log by machine learning,” in 2020 6th

International Confer ence on Advanced Computing and Communication

Systems (ICACCS), 2020, pp. 1395–1399.

[1] M. Castelli, F. M. Clemente, A. Popovič, S. Silva, and L. Vanneschi,

“A machine learning approach to predict air quality in california,”

Complexity, vol. 2020, 2020.

[2] M. Du n, Z. Xu, Y. Chen, and L. Wu, “Short-term air quality prediction

based on fractional grey linear regression and support vector machine,”

Mathematical problems in engineering, vol. 2020, 2020.