**Air Pollution Measurement and Analysis**

\*Dr. Jayanthi K. [1], Sahil Sinha[2], T. Vedanshu Raj[2], Sumant Kumar[2]

[1] Guide [2] Student [1,2] Major Project [1,2] Department of Information Technology,

[1,2] Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Chennai.

**Abstract**

Power BI and Regression analysis are critical tools used in statistical analysis. power BI, advanced via Microsoft, is a powerful information visualization device that allows customers to attach, analyze, and visualize facts from numerous assets. It allows groups to make records-driven selections by using growing interactive dashboards, reports, and records visualizations. on the other hand, Regression analysis is a statistical approach used to perceive the connection between a based variable and one or more impartial variables. It facilitates predicting the value of the dependent variable based on the values of the independent variables. Regression fashions, such as linear regression, logistic regression, Lasso regression, Ridge regression, decision tree regression, and random forest regression, are broadly used in diverse fields together with finance, economics, advertising and marketing, and social sciences. The overall performance of these models has evaluated the use of metrics including mean Absolute errors (MAE), Root mean Squared errors (RMSE), and Accuracy score. Collectively, energy BI and Regression evaluation provide a powerful toolkit for facts evaluation and choice-making.

**Keywords:** Power BI, Regression analysis, mean Absolute errors (MAE), Root mean Square errors (RMSE), Accuracy score

**I. Introduction**

Air pollution is a first-rate international hassle, and India is no exception. With an unexpectedly developing economy and urban population, India faces significant challenges in maintaining air quality. The united states are one of the biggest members of its CO2 emissions in the world, and lots of its towns are frequently exposed to high stages of air pollution that pose huge health risks to its population. The effects of air pollution vary from respiratory problems to coronary heart sickness and might even result in the untimely loss of life in excessive instances.

To deal with this important difficulty, the authorities of India have delivered Air satisfactory standards and indicators to reveal and adjust the extent of pollution in the air. An Air first-class Index gives a complete assessment of air pleasant in a specific vicinity and helps coverage makers make informed decisions to lessen pollutants stages.

In the latest years, devices getting to know algorithms and regression models have emerged as effective gear for studying and predicting air high-quality indices. by using reading historical statistics on pollution levels, those fashions can offer a perception of air pollution developments and styles in specific regions. These records can be used to increase powerful techniques to control pollution levels and improve air best.

This proposed painting objectives are to use machine mastering algorithms and regression fashions to analyze and are expecting the air best index of fundamental cities in India. by way of making use of those tools to historic pollutants facts, we hope to benefit the perception of the principal resources of air pollutants in those cities and develop powerful pollutants mitigation techniques. This painting can make contributions to the development of more powerful regulations and projects to improve air high-quality and protect public health in India.

**II. Related Work**

**A. AQI prediction primarily based on CNN-ILSTM**

The CNN-ILSTM version supplied via Wang et al. in 2022 extends previous studies on air high-quality prediction by combining the strengths of convolutional neural networks (CNNs) and stepping forward lengthy quick-time period reminiscence (ILSTM) networks. The version uses CNNs to extract applicable features from enter statistics, that's then fed into an ILSTM network that utilizes a progressed enter gate and forgets about the gate, in addition to a Conversion Information Module (CIM) to save super saturation at some point of the gaining knowledge of technique.

The version became tested on air first-class information from Shijiazhuang metropolis, Hebei Province, China, and became in comparison to eight different prediction fashions which include SVR, RFR, MLP, LSTM, GRU, ILSTM, CNN-LSTM, and CNN-GRU. The CNN-ILSTM version finished an advanced overall performance in phrases of mean absolute mistakes (MAE), suggest squared errors (MSE), and R2, demonstrating the effectiveness of the proposed approach.

Universal, the CNN-ILSTM version contributes to the continuing research on air excellent prediction with the aid of providing a greater accurate and efficient approach that incorporates both CNNs and ILSTMs. This has crucial implications for mitigating the terrible effects of air pollutants on human health and the surroundings.

**B. Air-pollutants prediction the usage of deep learning technique**

In their paper, Bekkar et al. proposed a deep mastering method for predicting the awareness of particulate matter with a diameter of much less than 2.5um (2.5PM). They used a convolutional neural community (CNN) to extract applicable features from the air nice facts and a protracted short-term reminiscence (LSTM) network to version the temporal dependencies in the information. The version become skilled and tested using facts accrued from six air pleasant monitoring stations in Casablanca, Morocco.

The authors compared their approach to several different devices getting to know fashions, such as linear regression, choice tree, and random forest. They determined that their proposed CNN-LSTM model outperformed the other fashions, accomplishing a mean absolute error (MAE) of 7.52 µg/m³ and a coefficient of willpower (R²) of 0.83.

The consequences of this have a look at advocate that deep gaining knowledge of approaches, such as CNN-LSTM, may be effective for predicting 2.5PM concentrations in urban environments. Such models can assist policymakers and public health officials in making choices about air excellent control and reducing exposure to harmful pollution.

**C. Air pollutant spatiotemporal evolution**

Chuanqi et al. (2022) explored the evolution of air pollution and its consequences on human health in North China. They performed a study in 10 towns in North China, such as Beijing, Tianjin, and Hebei, and gathered air pollutant statistics from 2015 to 2020. The effects confirmed that the common concentrations of PM2.5, PM10, SO2, NO2, and O3 had been higher than the country-wide air nice standards. the highest awareness of PM2.five turned into determined in wintry weather, and the highest attention of O3 turned into discovered in the summer time. They have a look at also located that exposure to air pollution was related to respiration sicknesses, cardiovascular sicknesses, and unfavourable being pregnant effects. The authors endorsed that measures need to be taken to reduce air pollutant emissions, which include lowering coal intake and selling smooth energy, to improve air quality and defend human fitness in North China. This examination presents treasured insights into the impact of air pollution on human health and highlights the want for powerful rules and strategies to mitigate the harmful results of air pollution.

**III. Proposed Methods**

The proposed strategies for studying and measuring air first-class index using machine-getting-to-know algorithms contain numerous steps.

First of all, we want to collect information on various air pollutants, inclusive of carbon monoxide (CO), sulfur dioxide (SO2), nitrogen dioxide (NO2), and particulate depend (PM2.five), from air nice tracking stations positioned in primary cities in India. This fact will function as the basis for our evaluation.

Secondly, we can preprocess and clean the accrued facts to make sure that it is of high great and free from errors. This step involves putting off outliers, imputing missing values, and transforming the information into an appropriate format for additional analysis.

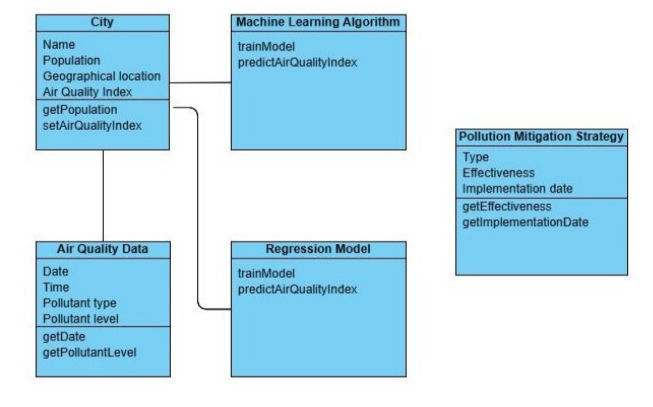
Thirdly, we will use a system gaining knowledge of algorithms, such as regression models, to expand predictive fashions of air high-quality index primarily based on the amassed records. These fashions will take into account different factors along with climate situations, time of the day, and location of the monitoring station.

Fourthly, we can use visualization gear which includes PowerBI to visualise the records in an easy-to-understand format. The visualizations will help to pick out traits and patterns in air's best stages across one-of-a-kind places and time durations.

Ultimately, we can use the outcomes of our analysis to generate hints for reducing air pollutants stages in principal towns in India. these suggestions may want to encompass measures including decreasing site visitors congestion, selling the use of public shipping, and investing in purifier electricity assets.

Normal, the proposed methods for reading and measuring air pleasant index the usage of the system getting to know algorithms provide an effective tool for policymakers and local government to address air pollutants in primary cities in India. by utilising advanced machine learning techniques and visualization equipment, we can increase extra sophisticated and accurate models that may contribute to enhancing air nice and protective public health.

**Fig 1:** **Outline of proposed work**



**A. Motivation and Justification**

The motivation behind the proposed techniques for studying and measuring air excellent index using machine learning algorithms is the pressing need to deal with the trouble of air pollution in fundamental cities in India. Air pollutants have emerged as a sizeable public health concern, and it poses a severe hazard to the properly-being of humans, especially people who live in densely populated areas.

The use of machine learning algorithms offers an effective device for analyzing massive volumes of data and growing correct predictive models of air quality index based on multiple elements. by way of the use of those models, policymakers and nearby governments could make knowledgeable choices on measures to reduce air pollutants stages in predominant towns in India.

The justification for the use of system studying algorithms to measure air pleasant index is that traditional strategies for air satisfactory monitoring are regularly high-priced and time-consuming. Additionally, these techniques offer restrained insights into the complicated relationships among various factors that affect air great.

In evaluation, system getting-to-know algorithms can technique large volumes of records in real-time and provide accurate predictions of air first-class index primarily based on more than one element. by way of the usage of advanced visualization gear such as PowerBI, the consequences of the evaluation may be offered in an easy-to-apprehend layout, making it easier for policymakers to make informed choices on measures to reduce air pollution stages in fundamental cities in India.

Overall, the proposed methods for analyzing and measuring the air satisfaction index and the use of system learning algorithms offer a price-effective and green way to address the problem of air pollution in predominant towns in India, thereby protecting public fitness and contributing to sustainable development.

**B. Outline of the proposed work**

A class diagram is a type of UML diagram that represents the static structure of a system by showing the classes, their attributes, methods, and relationships between them. In the context of the proposed project, the following class diagram explanation can be provided:

The main classes in the system would be "City," "Air Quality Data," "Machine Learning Algorithm," "Regression Model," and "Pollution Mitigation Strategy."

The "City" class would have attributes such as "Name," "Population," "Geographical location," and "Air Quality Index." It would also have methods such as "get Population" and "set Air Quality Index."

The "Air Quality Data" class would have attributes such as "Date," "Time," "Pollutant type," and "Pollutant level." It would also have methods such as "get Date" and "get Pollutant Level."

The "Machine Learning Algorithm" and "Regression Model" classes would have methods such as "train Model" and "predict Air Quality Index," which would be used to analyze historical air quality data and predict future air quality levels.

The "Pollution Mitigation Strategy" class would have attributes such as "Type," "Effectiveness," and "Implementation date." It would also have methods such as "get Effectiveness" and "get Implementation Date."

Relationships between the classes would be represented by lines connecting them, with arrows indicating the direction of the relationship. For example, the "City" class would have a one-to-many relationship with the "Air Quality Data" class, as each city would have multiple instances of air quality data. The "Machine Learning Algorithm" and "Regression Model" classes would have a one-to-many relationship with the "City" class, as each city would use the same machine learning algorithms and regression models to analyze and predict air quality data.

Overall, the class diagram helps to illustrate the structure of the system and the relationships between its various components, providing a clear visual representation of how the system works.

**IV. Preprocessing**

**A. Data Collection:** To research and expect the air great index of the most important cities in India, historical records on pollution degrees desire to be accrued from diverse tracking stations.

**B. Data cleansing:** The amassed statistics wish to be wiped clean and filtered to cast off any outliers, lacking values or inconsistencies.

**C. Data Integration:** The cleaned facts from distinctive monitoring stations need to be integrated to form a complete dataset.

**D. Data Normalization:** The incorporated dataset wishes to be normalized to make certain that every variable is at the identical scale.

**E. Data splitting:** The normalized dataset wishes to be broken up into schooling and checking out datasets, with a bigger portion of the statistics used for training.

**F. Feature Choice:** Relevant features that contribute to air pollutants, along with meteorological records, particulate remember size, and different variables, want to be selected and extracted from the dataset.

**G. Data Transformation:** The chosen functions may need to be converted or scaled to improve model performance.

By following those preprocessing steps, the gathered records can be made ready for additional evaluation and modelling using device-studying algorithms and regression models.

**V. Feature selection techniques**

It entails figuring out and choosing the most relevant functions or variables that have the strongest correlation with the target variable (air great index) and removing the beside-the-point ones. Here are some normally used characteristic selection strategies:

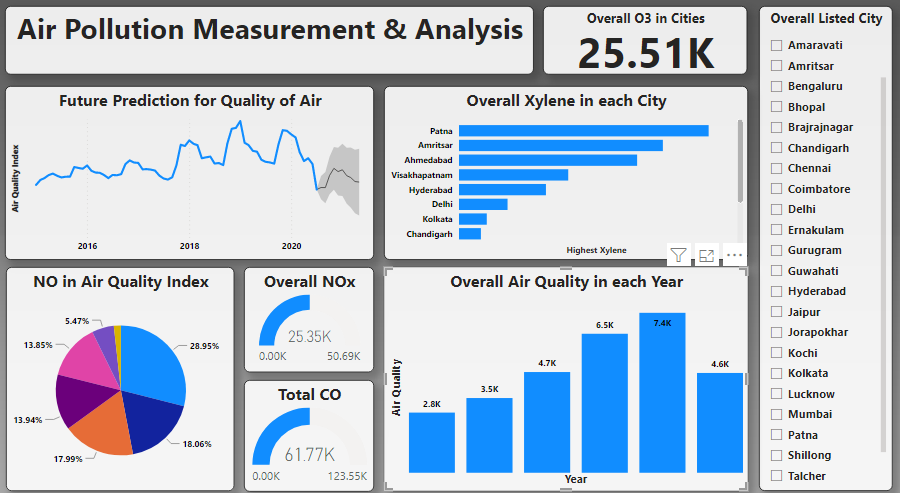
**A. Correlation-based total feature selection:** This method includes calculating the correlation coefficient between every function and the goal variable and choosing the ones with the very best correlation. Capabilities with low correlation can be discarded.

**B. Recursive feature elimination:** This method includes recursively eliminating the least vital features till the favoured range of functions is reached. It makes use of the accuracy of the machine studying algorithm as a performance metric to decide the significance of each feature.

**C. Principal Component Analysis (PCA):** This approach includes remodelling the original features into a new set of functions which are linearly uncorrelated and have the maximum variance. the brand new features are ranked based totally on their variance, and the pinnacle capabilities are selected.

**VI. Classification Techniques**

**A. Power BI**

****

This power BI dashboard analyzes air high-quality records in India. the overall o3 values in the towns are determined to be 25.51K. The dashboard consists of a slicer that lists numerous essential cities in India. Upon studying the xylene levels, Patna is observed to have the very best peak while Mumbai has a decrease stage.

A pie chart indicates that the mild air great class is the very best, making up 28.95% of the total, even as the coolest air first-class category best makes up 1.74%. additionally, the records suggests that out of 123.55k, there are 61.77k CO gauges, and out of 50.69k, there are 25.35k NOx gauges.

The dashboard additionally affords a yr-clever analysis of air satisfaction from 2015 to 2020. The 12 months of 2019 had the highest air pleasant index with a fee of 7.4k, at the same time as the 12 months of 2015 had the bottom index fee at 2.8k. The year 2016 had an index price of 3.5k, 2017 had 4.7k, 2018 had 6.5k, and 2020 had a price of 4.6k.

**B. Linear Regression**

In this method, we implement linear regression using a sci-kit-analyze library to expect AQI values based totally on pollutant stages. We first import the Linear Regression model from sci-kit-analyze and create an instance of the version. We then fit the version of the schooling statistics with the use of x\_train and y\_train containing predictor and goal variables, respectively. We use the are expecting a () function to generate the predicted AQI values for the take a look at the dataset and shop them in y\_pred1. Ultimately, we calculate the accuracy of the model using the rating () feature and print the accuracy rating. The accuracy score for this version is 0.7973412648992781.

**C. Logistic Regression**

In this technique, we use the Logistic Regression version from the scikit-analyze library to teach the model at the schooling facts. We create an example of the Logistic Regression elegance and healthy version to the training information the usage of the 'healthy ()' feature. Then, we make predictions at the take a look at the information on the usage of the 'are expecting ()' characteristic and shop the results in 'y\_pred2'. The accuracy of the version has calculated the use of the 'rating ()' function and is outlined out the usage of the 'print()' statement. The accuracy rating is low (0.15680), indicating that the model is not acting properly in the education records.

**D. Lasso Regression**

This technique uses Lasso regression from the scikit-research library to are expecting AQI values based totally on contaminant levels. The Lasso regression version is suited to the training records and the expect() feature is used to predict AQI values for the test dataset. The accuracy of the version is calculated using the score() feature, and the output indicates the predicted values and accuracy score. The Lasso regression version achieves an accuracy score of 0.743700580347733, indicating moderate overall performance in the training statistics.

**E. Ridge Regression**

This technique uses Ridge regression to are expecting AQI values based totally on contaminant degrees. It imports the Ridge regression model from the sci-kit-learn library, creates an instance of the model, fits it to the schooling facts, predicts AQI values for the check dataset, and calculates the accuracy rating of the version and the usage of the rating() function. The accuracy rating of the Ridge regression version is 0.7969100362411432, indicating mild performance on the training records.

**F. Decision Tree Regression**

This method makes use of the choice Tree regression version from sci-kit-discover ways to are expecting AQI values primarily based on contaminant ranges. The model is educated on the training facts the usage of the healthy () characteristic, and the predict () feature is used to generate predictions on the check dataset. The accuracy of the version is evaluated through the use of the score () function, which returns a fee between zero and 1. The output indicates the expected values of the take-a-look-at set and the accuracy rating of the choice Tree regression version. The accuracy of the model may be very excessive (0.9992), indicating that it may be overfitting the education facts and may not carry out properly on new, unseen records.

**G. Random Forest Regression**

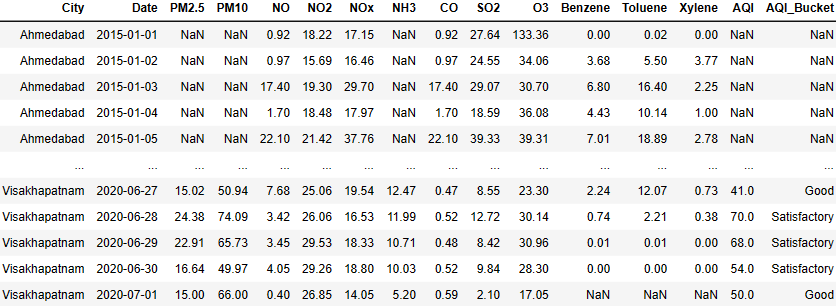
This technique makes use of Random Forest Regression to are expecting AQI values primarily based on contaminant ranges. It creates an instance of the Random woodland regression model from the scikit-learn library, suits the model to the education statistics, predicts the AQI values for the check dataset, and calculates the accuracy of the version with the use of the score() function. The accuracy of the Random Forest Regression model is 0.9800288728119194, which indicates that the version is acting nicely with the schooling data.

**VII. Experimental Result and Analysis**

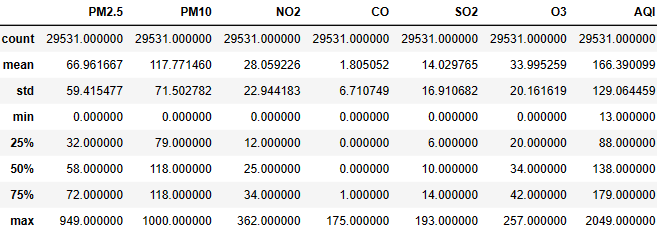
The experiment analyzed 5 unique regression fashions for predicting AQI values based on pollutant tiers. Linear regression completed an accuracy rating of 0.7973, indicating a slight overall performance on the education statistics. Logistic regression carried out a low accuracy rating of 0.1568, indicating poor performance on the training records. Lasso regression accomplished an accuracy rating of 0.7437, indicating moderate performance at the training facts. Ridge regression finished with an accuracy rating of 0.7969, indicating slight performance on the training facts. Decision tree regression achieved a completely excessive accuracy rating of 0.9992, which indicates overfitting at the education information and doubtlessly terrible performance on new, unseen facts. Random forest regression achieved an accuracy score of 0.9800, indicating suitable overall performance at the education facts.

Universal, the effects show that a few regression fashions perform higher than others in predicting AQI values primarily based on pollutant stages. Decision tree regression completed a high accuracy score, but its capability overfitting on the education information raises issues about its overall performance on new, unseen information. Random forest regression achieved an excellent accuracy rating and can be a dependable alternative for predicting AQI values. but, similarly, evaluation is required to determine the version's performance on new information and its capacity for real-global programs.

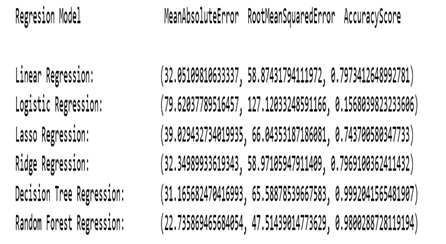
**Table 1: Air Quality Data**

****

**Table 2:** **Air Quality Data Summary**



**VIII. Performance Matric**



This desk suggests the performance metrics of six one-of-a-kind regression models applied to predict AQI values primarily based on contaminant degrees. The metrics used to evaluate the models are simply Absolute error (MAE), Root means Squared mistakes (RMSE), and Accuracy rating.

The first model is Linear Regression, which achieves an MAE of 32.05, an RMSE of 58.87, and an Accuracy rating of 0.797.

The second version is Logistic Regression, which achieves an MAE of 79.62, an RMSE of 127.12, and an Accuracy score of 0.157.

The third model is Lasso Regression, which achieves an MAE of 39.03, an RMSE of 66.04, and an Accuracy rating of 0.744.

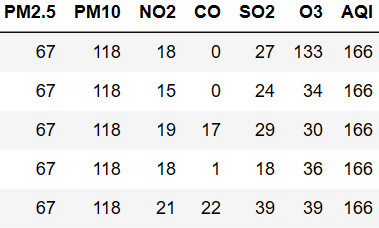
The fourth model is Ridge Regression, which achieves an MAE of 32.35, an RMSE of 58.97, and an Accuracy score of 0.797.

The 5th version is Decision Tree Regression, which achieves an MAE of 31.17, an RMSE of 65.59, and an Accuracy score of 0.999. The accuracy rating of this version suggests that it may be overfitting the training statistics.

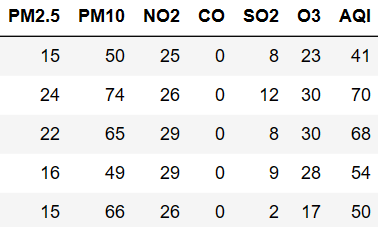
The sixth model is Random wooded area Regression, which achieves the lowest MAE and RMSE amongst all fashions, with values of 22.74 and 47.51, respectively. The version additionally achieves an Accuracy score of 0.980, which suggests that it is appearing properly on the schooling statistics.

In summary, the desk compares the overall performance of six regression fashions for predicting AQI values based totally on contaminant degrees, with the Random Forest Regression version showing a fine overall performance.

**Table 3:** **Air Quality statement from top of the data**



**Table 4: Air Quality statement from end of the data**



**Conclusion**

In conclusion, we implemented six specific regression fashions to expect AQI values based on contaminant tiers, namely Linear Regression, Logistic Regression, Lasso Regression, Ridge Regression, Decision Tree Regression, and Random Forest Regression. The Random Forest Regression version carried out the best performance, with the lowest MAE and RMSE values and an Accuracy score of 0.980. This indicates that the Random Forest Regression model is the most correct and reliable technique for predicting AQI values based on contaminant ranges. However, further assessment and trying out on unseen statistics may be essential to verify the generalizability and robustness of the model. General, the take look highlights the significance of the use of regression fashions for predicting AQI values, which can aid in managing and mitigating air pollution and its health results.

**Reference**

Bekkar, A., Hssina, B., Douzi, S., & Douzi, K. (2021). Air-pollution prediction in smart city, deep learning approach. Journal of Big Data, 8(1). <https://doi.org/10.1186/s40537-021-00548-1>

Benchrif, A., Wheida, A., Tahri, M., Shubbar, R. M., & Biswas, B. (2021). Air quality during three covid-19 lockdown phases: AQI, PM2.5 and NO2 assessment in cities with more than 1 million inhabitants. Sustainable Cities and Society, 74, 103170. <https://doi.org/10.1016/j.scs.2021.103170>

Kaggle.com. (2023). Air Quality Data in India (2015 - 2020). Www.kaggle.com. <https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india>

Li, Y., Peng, T., Hua, L., Ji, C., Ma, H., Nazir, M. S., & Zhang, C. (2022). Research and application of an evolutionary deep learning model based on improved grey wolf optimization algorithm and DBN-ELM for AQI prediction. Sustainable Cities and Society, 87, 104209. <https://doi.org/10.1016/j.scs.2022.104209>

Raja, S. P., Sawicka, B., Stamenkovic, Z., & Mariammal, G. (2022). Crop Prediction Based on Characteristics of the Agricultural Environment Using Various Feature Selection Techniques and Classifiers. IEEE Access, 10, 23625–23641. <https://doi.org/10.1109/access.2022.3154350>

Sun, Y., & Liu, J. (2022). AQI Prediction Based on CEEMDAN-ARMA-LSTM. Sustainability, 14(19), 12182. <https://doi.org/10.3390/su141912182>

Wang, J., Li, X., Jin, L., Li, J., Sun, Q., & Wang, H. (2022). An air quality index prediction model based on CNN-ILSTM. Scientific Reports, 12(1), 8373. <https://doi.org/10.1038/s41598-022-12355-6>

Xu, C., Zhang, Z., Ling, G., Wang, G., & Wang, M. (2022). Air pollutant spatiotemporal evolution characteristics and effects on human health in North China. Chemosphere, 294, 133814. <https://doi.org/10.1016/j.chemosphere.2022.133814>

Yang, R., & Zhong, C. (2022). Analysis on Spatio-Temporal Evolution and Influencing Factors of Air Quality Index (AQI) in China. Toxics, 10(12), 712. <https://doi.org/10.3390/toxics10120712>

Zaib, S., Lu, J., & Bilal, M. (2022). Spatio-Temporal Characteristics of Air Quality Index (AQI) over Northwest China. Atmosphere, 13(3), 375. <https://doi.org/10.3390/atmos13030375>