JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

AIR QUALITY PREDICTION USING RANDOM FOREST REGRESSION

Guntaka Sathvika[1] , Parchuri Poojitha[1] , Kandru Rakesh[1] , Lokesh Tadepalli[1] , Dr.Konda Chaitanya[2]

[1] b.tech student, Department of CSE, ANU college of engineering and technology, Acharya Nagarjuna University, Guntur, Andhra Pradesh.

[2] Assistant Professor, Department of CSE, ANU college of engineering and technology, Acharya Nagarjuna University, Guntur, Andhra Pradesh.

ABSTRACT :

Predicting air quality is essential to managing public health and the environment. In order to forecast air quality characteristics, this study suggests a supervised machine learning method that uses Random Forest Regression algorithm. By utilizing a dataset that includes a variety of pollutants, this model seeks to predict metrics related to air quality with a high degree of accuracy and consistency. The usefulness of Random Forest Regression in capturing intricate correlations between input factors and air quality indicators is determined through rigorous testing and assessment. The model's output displays encouraging performance indicators, such as strong generalization skills across various temporal and spatial contexts and high prediction accuracy. Through the development of useful tools that help the public, environmental organizations, and policymakers better understand and manage air pollution, this research advances the field of air quality monitoring systems.

Key words: Air Quality Prediction, Supervised Machine Learning, Random Forest Regression algorithm.

1. INTRODUCTION :

Quality of the air gained an international attention due to its profound effect on human health, ecosystems, and climate. Air pollution causes serious public health problems like respiratory and cardiovascular disorders[1] because of growing urbanization, industrial activity, and vehicle emissions. Thus, accurate air quality forecasting is essential for reducing its negative consequences and guiding policy choices.

Because machine learning approaches can handle complex and non-linear correlations in data, they are frequently utilized to forecast air Quality. The Random Forest Regression (RFR) technique has become well-known among them due to its resilience, accuracy, and interpretability [2]. In order to minimize overfitting and enhance generalization[3], Random Forest Regression is a technique which uses several Decision Trees while learning and provides the average forecast of individual Trees as output.

Recent research has shown that Random Forest Regression is a useful tool for predicting air quality. In order to estimate PM2.5 concentrations, for example, [4] used Random Forest Regression and discovered that it performed better than other ML Algorithms like SVM's and neural networks. In a similar vein, [5] used Random Forest Regression to estimate NO2 levels and was successful in obtaining resilience against missing data as well as excellent prediction accuracy.

Data collection, feature selection, model training, and validation are some of the crucial processes in the Random Forest Regression application for air quality prediction. Historical air quality measurements, meteorological information, and emission inventories[6] are common examples of data sources. In order to determine which factors—such as temperature, humidity, wind speed, and industrial emissions—have the greatest influence on air pollutant concentrations[7], feature selection is crucial.

In order to better understand how well Random Forest Regression can predict air quality, a complete model that takes into account a variety of anthropogenic and environmental factors is being developed. The goal of the project is to improve air quality forecast accuracy by utilizing the advantages of Random Forest Regression. This will help stakeholders and policymakers develop policies for managing air quality that are both effective and efficient.

Carbon Monoxide, Ammonia, Nitrogen Dioxide, ozone (O3), fine particles less than Ten micrometers in measurement (PM10), and fine particles less than 2.5 micrometers in measure (PM2.5) are the Six main pollutants that are present in India's record. So, the monitoring stations need to give others the grouping of particular toxin with its usual behaviour throughout a period. Regarding Carbon Monoxide and ozone (O3) the usual is thought to be greater than 8 hours, and the typical for other three is twenty-four hours.
The measuring unit per cubic meter is the milligram (microgram, depending upon the content of Carbon Monoxide). The associated indicators and air quality indices are shown in the following table.

AQI Category (Range)	PM_{10} 24-hr	$PM_{2.5}$ 24-hr	NO_2 24-hr	O_3 8-hr	CO 8-hr (mg/m^3)	SO_2 24-hr	NH_3 24-hr
Good (0-50)	0-50	0-30	0-40	0-50	0-1.0	0-40	0-200
Satisfactory (51-100)	51-100	31-60	41-80	51-100	1.1-2.0	41-80	201-400
Moderate (101-200)	101-250	61-90	81-180	101-168	2.1- 10	81-380	401-800
Poor (201-300)	251-350	91-120	181-280	169-208	10.1-17	381-800	801-1200
Very poor (301-400)	351-430	121-250	281-400	209-748*	17.1-34	801-1600	1201-1800
Severe (401-500)	430 +	250+	400+	748+*	34+	1600+	1800+

Figure: AQI Category, Pollutants and health breakpoints

As seen from the above picture each level is represented by the colour code that shows how people can grasp it easily.

2.RELATED WORK:

Several approaches to air quality prediction have been studied in the past, from conventional statistical models to more sophisticated machine learning methods. The use of supervised learning algorithms for air quality

forecasting tasks, such as SVM's [5,8], ANN's [9,10], Decision Trees [11], has been the subject of several studies.

A Support Vector Machine regression technique was used in a noteworthy work by Zhang et al. [4] to forecast PM2.5 concentrations using meteorological and air quality data. In comparison to conventional statistical models[18], their findings showed how well SVM captured complex correlations between input variables and PM2.5 levels, yielding high prediction accuracy.

When Pérez & Reyes [12] evaluated different machine learning methods for particulate matter forecasting, they found that ensemble methods performed better than individual models in most cases.

Using a large dataset of meteorological, traffic, and geographic data, Liang et al. [13] looked at ANN's to forecast condition of air. Their research demonstrated how neural networks can represent temporal dependencies and non-linear interactions in air pollution data, with promising outcomes for prediction accuracy and generalization.

Regression analysis was investigated by Mahanta et al. [17] to forecast urban air condition. They gave an example of how well regression algorithms can forecast urban air quality. To be able to examine relationship between air condition indicators as well as influencing factors, this study concentrated on a number of regression models, offering a thorough assessment of their prediction accuracy.

The application of large data programming models for air quality simulations was investigated by Ayyalasomayajula et al. [19, 20]. Their research concentrated on using big data technologies to manage the enormous volumes of data associated with air quality modelling. Their goal was to improve air quality simulation accuracy and efficiency by using scalable programming approaches.

In order to predict air quality, Alsaedi and Liyakathunisa [22] looked into the use of DL algorithms regarding the study about temporal and geographical data. In order to capture complex geographical and temporal patterns in air quality data, their study emphasized the use of deep learning models. They sought to increase the air quality models' capacity for prediction by utilizing deep neural networks.

For air quality forecasting problems, ensemble learning techniques have also been investigated in addition to individual machine learning algorithms. For example, Wang et al. [14] suggested a hybrid model to predict air pollution concentrations by integrating Genetic Algorithms and Decision Trees. Their hybrid strategy increased prediction performance and robustness by successfully addressing the shortcomings of separate models.

The use of ELM's to estimate the AQI was first described by Baran [16]. A study Symposium on artificial intelligence and data processing showed that ELMs could predict AQI quickly and accurately. ELM's main advantages are its quick learning, speed, and strong generalization capabilities, which make it ideal for jobs involving the monitoring and prediction of air quality in real time.

In 2020, Emam Hossain et al.[15] introduced a unique deep learning method to estimate the AQI. The primary aim of this work is forecasting AQI value for daily in Chattogram and Dhaka using the combined power of LSTM and GRU models. Despite the fact that numerous studies have attempted to forecast the AQI value, none of them have yet to make use of the combined strength of LSTM and GRU. In order to forecast daily AQI data, they consequently suggested a hybrid model that included the capabilities of two of the most potent time series analyzers (GRU and LSTM). Even while earlier studies have significantly advanced the subject of air quality prediction, there are still obstacles to overcome and room for growth. A prevalent constraint is the absence of uniformity in data gathering and preprocessing techniques, which may impact the repeatability and consistency of findings amongst research projects. Furthermore, because air pollution patterns are dynamic, prediction

models must be continuously improved and adjusted to guarantee accuracy and dependability in practical applications. By concentrating on Random Forest Regression, a potent ensemble learning technique, for air quality prediction, this study expands on the corpus of previous research. Through the utilization of a wide range of input factors and thorough testing, our goal is to provide new knowledge and approaches that will improve the efficiency of air quality monitoring and forecasting systems.

3.PROPOSED METHODOLY:

Because Random Forest Regression is good at capturing non-linear correlations, it is a good choice for modelling the complex dynamics of data on air quality. Through the combination of predictions from several decision trees trained on various subsets of the data, this technique reduces the possibility of overfitting. By enhancing generalization performance and robustness against noise, this ensemble technique produces more accurate predictions of air quality. This model has the ability to forecast the most accurate outcomes in a certain field. The model's accuracy falls between 80 and 90 percent.

3.1 DATASET DESCRIPTION:

The dataset used is Hourly Air Quality Data (2015-2020). This dataset covered various cities across India and nearly 30,000 areas in those cities. The dataset taken is of longer period (2015-2020) as well as standard and verified without any missing data. Longer period of data is helpful to predict the best accurate results for the future. This dataset consists of 16 columns and 29,530 rows(or)data. Columns are (Place, date, pm2.5, pm10, NitricOxide, NitrogenDioxide, NitrogenOxides, Ammonia, CarbonMonoxide, SulphurDioxide, ozone (O3), benzene, toluene, xylene, AQI_value, AQI_Bucket). Out of 16, 12 are pollutants. AQI_value represents the Air Quality Index value as well as AQI_Bucket represent air's purity (good, Acceptable, pathetic, ExtremelyPoor, Severely Polluted). The experimental data within the dataset has been collected from Kaggle.

3.2 DATA PRE-PROCESSING :

Pre-processing plays a crucial role in the field of ML. The data that has been taken for experimentation may contain irrelevant data(or) insufficient data. Sometimes, there is a chance of having a null values. So, it is important to handle this data and converting the data to a structure that the ML algorithm's could apply to the model after gathering. Preprocessing contains a lot of methods such as DataCleaning: This helps to eliminate the content that has been appended or categorized erroneously. DataImputation: A lot of machine learning frameworks come with scripts and ways to adjust or replace missing data. utilizing the data in the designated field's standard deviation, mean, and median to approximate missing values is a common technique. Oversampling: Increasing the number of insights utilizing methods such as bootstrapping or adding them to underrepresented groups could aid in addressing biases or mismatch within that data set. DataIntegration the lack of completeness of one data set may be addressed by By merging multiple datasets to form a large collections. DataNormalization: The amount of storage and computation required to train rounds will be determined by the volume of the data set. Normalization reduces the amount of data by lowering its order and magnitude.

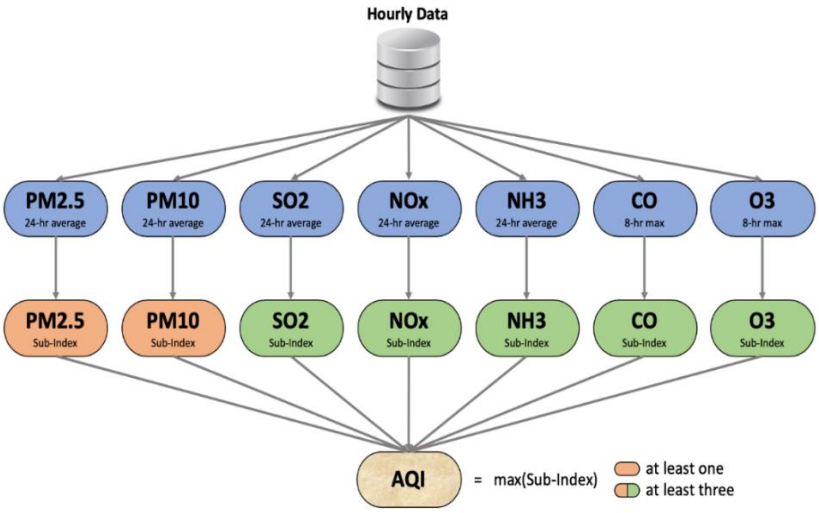
Fill the null values of the pollutants with their corresponding mean values.

Calculating Sub-Index for each Pollutant & Fill the AQI_value column's Null data by selecting the greatest readings from the Sub-Indexes.

3.2.1 CALCULATING SUB-INDEX AND AQI:

Seven metrics are used within AQI_value computation: pm2.5, pm10, NitrogenOxides, Ammonia, SulphurDioxide, CarbonMonoxide and ozone (O3). average for the last twenty-four hours is used for the first five particles, given that a minimum of sixteen values are present. The highest data from the last eight Hours

has been used for CO and O3. Every measure has been changed in the form of Sub_Index by using pre-defined groupings. Having the prerequisite requiring a minimum of one of PM2.5 and PM10 be accessible and at least 3 of 7, the Final AQI is the highest Sub-Index.



AQI levels and the sub-index of pollutants can be computed hourly using a method released by[21]. Related pollutant's Sub-Index is $IAQIp$, and the greatest data of $IAQIp$, that refers for a single air particle, is the AQI:

$$AQI = \text{Greatest of } \{IAQI1, IAQI2, IAQI3, \ldots, IAQIn\} \quad (1)$$

$$IAQIp=(IAQI_{Hi}-IAQI_{Lo})\ (Cp-BP_{Lo})\ /(BP_{Hi}-BP_{Lo})+IAQI_{Lo} \quad (2)$$

where C_p stands for mass concentration value for an air pollutant p, BP_{Hi} stands for largest value of the concentration limit from[21], BP_{Lo} stands for least value of the concentration limit from[21], $IAQI_{Hi}$ is the respective value of BP_{Hi} from[21], $IAQI_{Lo}$ is also the respective value of BP_{Lo} from [21].

3.3 ARCHITECHTURE OF A MODEL:

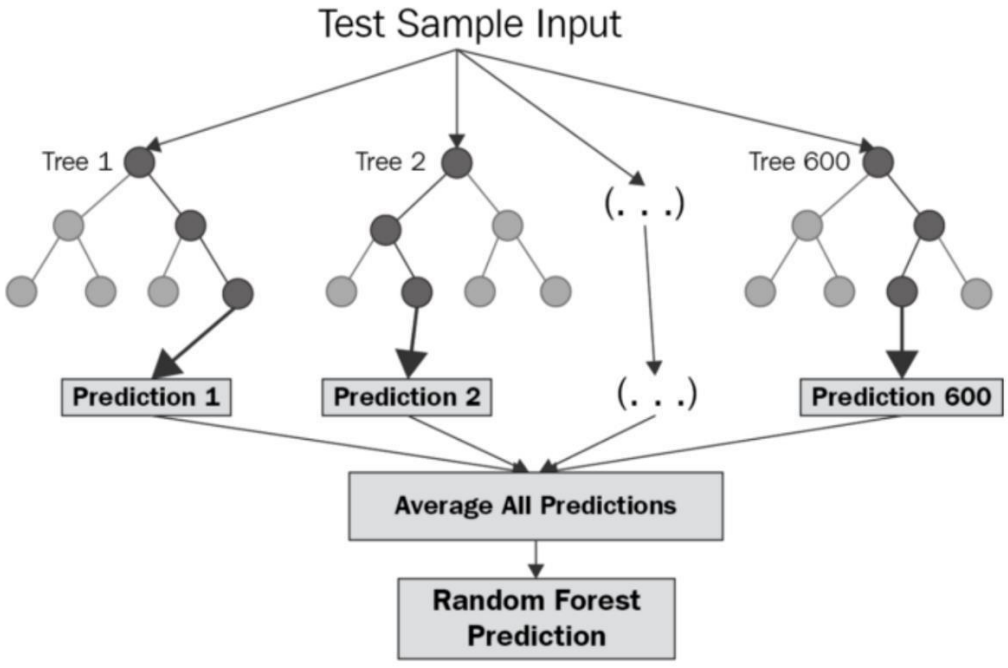
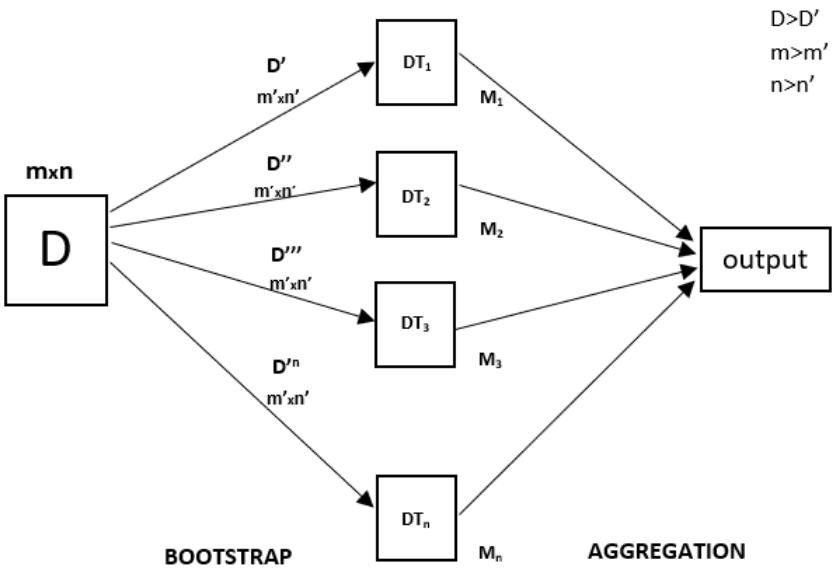


Figure: Random Forest Regression from[24]

The entire dataset can be splitted into train and test sets using some random state. The model is developed using Random Forest Regression. After that, it is trained and tested using the preprocessed data.

3.3.1 RANDOM FOREST REGRESSION:

By Chance In machine learning, Random Forest Regression is an ensemble methodology that may be used to accomplish regression tasks using more number of DT's and a technique called bagging.This helps in decreasing overfitting and increasing accuracy, the fundamental principle behind integrating more DT's is to get the end result more accurately rather than relying solely upon each DT alone. Although variance for each decision tree is considerable, the variance of the combined set is minimal because each decision tree is perfectly trained on the specific sample data when they are combined in parallel. The average of each decision tree's individual outputs constitutes the ultimate result. The below figure[25] shows how Random forest regression can be implemented.



Bootstrap: This technique uses randomized selections of the training data to train several models. **Aggregation:** To get a reliable and accurate overall result, it averages the predictions made by each individual tree.

3.4 RESULTS:

To estimate the quality of air, user's data is fed into the Random Forest Regression model. It computes the AQI after receiving the user's input values for each pollutant. The AQI bucket is then predicted using the AQI value; the table below illustrates this process. Lastly, based on user-provided new data from the front end, the anticipated Air Quality Index output is displayed on the front end along with the status of the air quality: good, moderate, unhealthy, unhealthy for strong individuals, and hazardous.

AQI	Air Pollution Level
Less than 50	good
51 to 100	Acceptable
101 to 200	Pathetic
201 to 300	Poor
301 to 400	Extremely Poor
Greater than 400	Severely Polluted

AQI Classification

In Addition to Random Forest Regression model various methods such as DT Classifier, Linear Regression, SVR, and AdaBoost Regressor are selected for the contrast testing. The table below lists and displays every accuracies of the matching models. Random Forest Regression demonstrated the highest accuracy in comparison.

Algorithm	Accuracy
Random Forest Regression	81.61%
Decision Tree Classifier	74.80%
Linear Regression	79.58%
SVR	56.99%
AdaBoost Regressor	47.55%

Accuracy table

4.CONCLUSION:

The development of the Random Forest Regression model to predict the quality of air using a variety of datasets, including pollution data, is the study's final project. High prediction accuracy and strong generalization across several locations and eras were displayed by this model. Our method has applications in public health management and environmental monitoring since it produces reliable and timely forecasts. This research shows that machine learning can be used to address air quality issues and promote sustainable ways for reducing the negative effects of pollution on society, even though more optimization is required. This model has an accuracy of 81.64%.

5.FUTURE WORK:

There are numerous ways to improve air quality forecast in the future. First off, forecast accuracy might be improved by incorporating other data sources like social media feeds and satellite photos. Second, for greater comprehension and confidence, models' interpretability—especially that of Random Forest Regression—must be improved. Opportunities for prompt reactions to shifting pollution dynamics are provided by adaptive modelling and real-time monitoring. Prediction utility and efficacy can be further increased by combining models with decision support systems and assessing ensemble learning techniques. Finally, thorough validation in real-world environments is essential for implementation in practice and stakeholder acceptance. Following these paths will help develop public health management plans and environmental monitoring techniques that are more successful.

6.REFERENCES:

[1] Brunekreef, B., & Holgate, S. T. (2002). Air pollution and health. *The Lancet*, 360(9341), 1233-1242.

[2] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

[3] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3), 18-22.

[4] Zhang, X., Tang, L., Luo, L., & Liang, J. (2017). A hybrid air quality early-warning system based on machine learning and multiple linear regression. *Atmospheric Environment*, 166, 98-105.

[5] Jiang, Y., Bai, Y., Yang, C., & Tang, T. (2020). A hybrid model for short-term air quality prediction. *Atmospheric Pollution Research*, 11(1), 218-229.

[6] Chen, T., Zhang, Y., Zhu, J., & Zhang, Z. (2018). Air quality prediction by deep learning model. *Neurocomputing*, 273, 155-162.

[7] Hu, X., Waller, L. A., & Wang, Y. (2017). Estimating ground-level PM2.5 concentrations in the southeastern U.S. using geographically weighted regression. *Environmental Research*, 156, 283-290.

[8] Li, G., Sun, G., Wang, Y., & Shen, Z. (2016). Application of an improved SVM-based prediction model to daily PM2.5 concentration in Beijing, China. *Atmospheric Pollution Research*, 7(4), 577-584.

[9] Elangasinghe, M. A., Singhal, N., Dirks, K. N., & Salmond, J. A. (2014). Development of an ANN-based air pollution forecasting system with explicit knowledge through sensitivity analysis. *Atmospheric Pollution Research*, 5(4), 696-708.

[10] Kurt, A., & Oktay, A. B. (2010). Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks. *Expert Systems with Applications*, 37(12), 7986-7992.

[11] Ganguly, S., Ray, A. K., & Kumar, R. (2015). Air quality prediction using decision tree-based approaches. *Procedia Computer Science*, 58, 199-204.

[12] Pérez, P., & Reyes, J. (2011). An integrated neural network model for PM10 forecasting. *Atmospheric Environment*, 45(10), 1713-1720.

[13] Liang, X., Zou, T., Guo, B., et al. (2019). Short-term air quality prediction using long short- term memory neural networks. *Atmospheric Environment*, 200, 218-229.

[14] Wang, J., Zhang, Z., Zheng, J., et al. (2017). A hybrid model for air pollutant concentration prediction based on decision trees and genetic algorithms. *Environmental Monitoring and Assessment*, 189(6), 279.

[15] Hossain, E., Farhana, S., & Islam, S. M. R. (2020). Air quality prediction in Dhaka and Chattogram using a hybrid deep learning model. *Journal of Air Quality, Atmosphere & Health*, 13(12), 1499-1512.

[16] B.Baran, "Prediction of Air Quality Index by Extreme Learning Machines," 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, Turkey, 2019, pp. 1-8, doi: 10.1109/IDAP.2019.8875910.

[17] Mahanta, S., Ramakrishnudu, T., Jha, R. R., & Tailor, N. (2019). Urban Air Quality Prediction Using Regression Analysis. In *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)* (pp. 1118-1123). IEEE. doi:10.1109/TENCON.2019.8929517.

[18] Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control*. John Wiley & Sons.

[19] H. Ayyalasomayajula, E. Gabriel, P. Lindner and D. Price, "Air Quality Simulations Using Big Data Programming Models," 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService), Oxford, 2016, pp. 182184, doi: 10.1109/BigDataService.2016.26.

[20] 2016, pp. 182184, doi: 10.1109/BigDataService.2016.26.

[21] Pai, P.; Karamchandani, P.; Seifneur, C. Simulation of the regional atmospheric transport and fate of mercury using a comprehensive eulerian model. Atmos. Environ. 1997, 31, 2717–2732. [CrossRef]

[22] China's Ministry of Environmental Protection. Available online: **http://kjs.mep.gov.cn/hjbhbz/bzwb/dqhjbh/jcgfffbz/201203/W020120410332725219541.pdf** (accessed on 27 November 2015). (In Chinese)

[23] Alsaedi, A. S., & Liyakathunisa, L. (2019). Spatial and temporal data analysis with deep learning for air quality prediction. In *2019 12th International Conference on Developments in eSystems Engineering (DeSE)* (pp. 581-587). IEEE. https://doi.org/10.1109/DeSE.2019.00111

[24] Keboola: The Ultimate Guide to Random Forest Regression (keboola.com)

[25] GeeksforGeeks: Random Forest Regression in Python - GeeksforGeeks