Comparative Analysis of Machine Learning for Predicting Air Quality in Smart Cities

KAMEL MAALOUL, LEJDEL BRAHIM
Informatique Department
El-Oued University
El Chott , El-Oued
ALGERIA

Abstract: - Ambient air pollution is the most harmful environmental risk to health. As urban air quality improves, health costs from air pollution-related diseases diminish. This is why air pollution is a major challenge for the public and government around the world. Deployment of the Internet of Things-based sensors has considerably changed the dynamics of predicting air quality. Air pollution can be predicted using machine learning algorithms Data-based sensors in the context of smart cities. In this paper, we performed pollution forecasting using machine learning techniques while presenting a comparative study to determine the best model to accurately predict air quality. Random Forest is an efficient algorithm capable of detecting air quality.

Key-Words: - Air Quality; Machine Learning; Random Forest; Smart Cities; Linear Regression; IOT.

Received: September 12, 2021. Revised: May 7, 2022. Accepted: June 3, 2022. Published: July 2, 2022.

1 Introduction

The smart city exploits information and communication technology (ICT) to optimize populations' lives in areas including such utilities, transportation, energy, and, most importantly, health. It also allows the government to make better use of the resources it has [1].The most pressing issues at the moment Smart cities have high levels of air pollution. The potential to affect the features of the atmosphere and increase health hazards to organisms, as well as cause their destruction, is characterized as air pollution [2]. Air pollution impacts people's health by exposing them to pollutants in the environment. Air pollution must be swiftly regulated in order to avoid a calamity. Environmental quality forecasting must be done quickly in order to take preventative measures in smart cities [3]. The capacity to retrieve data from sensors for smart cities is a significant benefit in addressing the issues and roadblocks that impede their development [4].

In the development of smart city concepts, air quality is a critical component. Air quality, on the other hand, is a big issue in many locations. The high population growth led to an increase in the use of transportation, energy, industry and vehicles. According to many studies, air quality in smart cities represents a major challenge for smart cities and that machine learning is an effective and better solution to predict air quality, although it is not used in many cities of the world. Different algorithms and techniques machine learning have been utilized in the past for various use cases, including Neural Networks, fuzzy a system, Support Vector Machine, Support Vector Machine for regression, fuzzy logic, Decision Trees, and K-Nearest Neighbor. Sophisticated and advanced machine learning techniques are increasingly being used, with 2.5 micrometers particulate matter (PM2.5) being the main target element; air quality data is best combined with other types of data. It is critical to guarantee that analysis of data is accurate and efficient. Data misinterpretation can lead to erroneous conclusions, which can be extremely dangerous. To guarantee the velocity and analysis of communication sources, artificial intelligence predictions and the usage of reliable data must be carefully considered [1].

We used multiple machine learning approaches to predict pollution in this work. In order to determine the best and most accurate predictive model for pollution estimate, we give a

comparative analysis of these methodologies based on acknowledged evaluation criteria. We investigated the processing time of various technologies using independent learning and Apache Spark, given the criteria for effective real-time analysis of smart city data. We proposed the best model in terms of both processing time and error rate in this research. The purpose of this research is to use a variety of machine learning models to provide an accurate and comparative examination of air quality. The results of data collection utilizing sensors at a certain site are reviewed and compared to identify the highest performing algorithm and accuracy.

The remainder of the paper is structured as follows: The relevant work is briefly discussed in Section 2 and how it differs from our submitted study. The third section discusses machine learning algorithms and how to use the model. The proposed model, empirical evaluation method, and data gathering process are all explained in Section 4. Section 5 contains the findings and their commentary. Finally, Section 6 summarizes the paper's findings as well as future study directions.

2 Related Works

In this section, we mention some previous studies related to our topic.

Martinez-Espana et al. [5] use of Machine Learning algorithms to forecast particulate matter concentration in the atmospheric air has been discussed. Machine learning technologies such as linear regression, ML Pregressor neighbors regressor, Decision Tree regressor, and gradient boosting regressor were used in the experiments. The authors convincingly demonstrated that Gradient Boosting Regression outperformed all other techniques.

Hamami et al. [6] proposes air quality category using class algorithms together with Logistic Regression, KKN, decision Tree, and Random forest set of rules. primarily based on experiment, decision tree model has the first-rate accuracy to categories air excellent degree as much as 100% with tuning numerous hyper parameters.

Fernando et al. [7] have a look at has been to find the maximum appropriate system studying method for predicting accurate air quality index in Colombo based upon PM2.five unique awareness.PM2.five concentration in Colombo were anticipated the usage of four correlated air pollutant concentrations including SO2, NO2, PM2.5, PM10. machine mastering techniques consisting of k-Nearest Neighboring, more than one Linear-Regression, Random forest, and support Vector Machines have been used to educate and examine the prediction models. Random forest changed into identified because the excellent appropriate prediction model after evaluating the models, with over 85% extra accuracy.

Abirami et al. [8] predict AQI accurately with utilizing the information set on one of a kind ML version with proper pre-processing method for locating nature regarding the air rests for the maximum component signified with the aid of its AQI (air high-quality index) value. It is tried to gauge the air situation inside the bounds of a definite region by using utilizing device learning strategies like aid vector regression (SVR), selection tree regression (DTR), a couple of linear regression (MLR) and random forest regression (RFR). RFR performed out the finest among all regression examples.

Murugan et al. [9] implement machine learning algorithms to discover the accuracy of the prediction of particulate be counted, PM2.five in air pollutants in smart towns . to test the implementation of device studying on this prediction, Multi-Layer Perceptron (MLP), and Random forest are selected and in comparison among those algorithms the usage of the Air pollution dataset. The outcome of this studies is that Random forest area gave the first-rate accuracy in prediction of Particulate count number.

On this paper, Sinnott et al. [10] explored using machine learning algorithms for prediction of pollution and specifically PM2.five. Prediction of pollutants activities is more and more essential in primary cities due to the increased urbanization of populations and associated effect on site visitors volumes. statistics from a selection of heterogeneous assets became used and concerned collection and cleaning for use in machine learning algorithms. Linear regression models, ANN models and LSTMs had been all explored. It turned into found that LSTM completed best and became able to are expecting high PM2.5 values with affordable accuracy. ANN and linear models have drawbacks in prediction of high PM2.five values but they provided reasonable standard performance.

In this paper, Pasupuleti et al. [11] are introducing a device that could continue that could take gift pollutants and with the help of beyond pollutants, we are walking an algorithm based totally at the machine learning to are expecting the future data of pollutants. The sensed information is stored within the Excel sheet for in addition evaluation. those sensors are used on the Arduino platform to accumulate the pollutant data. We are expecting the air high-quality index by means of the use of specific machine learning algorithms like linear regression, choice Tree and Random forest. From the consequences, we concluded that the Random forest set of rules gives better prediction of air satisfactory index.

Vu et al. [12] applied a machine learning-based random forest technique to determine the effectiveness of the Beijing Action Plan by separating the impact of meteorology on ambient air quality. The effects display that meteorological conditions have an important have an effect on at the annual adjustments in ambient air satisfactory. The movement plan has additionally been very effective in lowering primary pollutants emissions and enhancing Beijing's air pleasant. This work is a a hit instance of air excellent policy development in other areas of China and other developing international locations.

3 Proposed Architecture

Figure 1 describes the architecture of the proposed model. Data on air pollution is collected via sensors and saved as a dataset. This data set has already been preprocessed. The data set is then separated into two parts: a training data set and a test data set. The training dataset is subjected to more supervised machine learning methods. The collected results are then compared to the test data set and assessed.

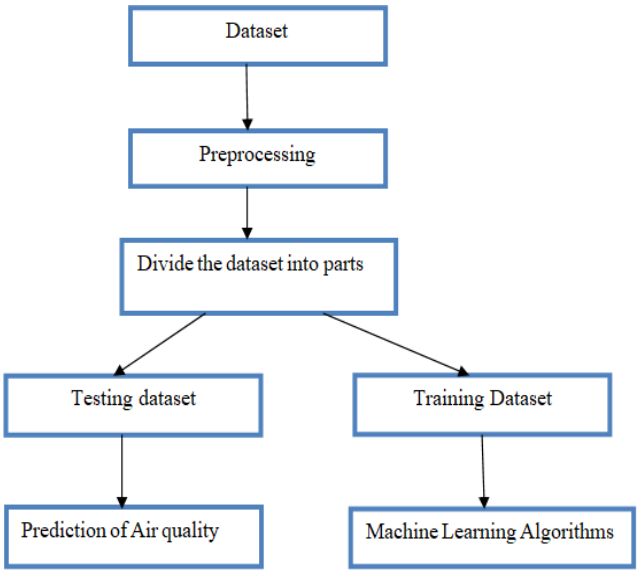


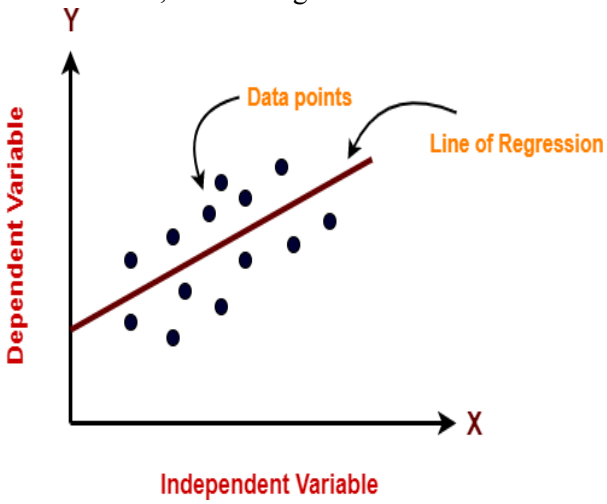
Figure 1: The architecture of the implemented System

4 Background Machine Learning Algorithms

The Supervised Machine Learning approach to air pollution prediction takes into account machine learning algorithms such as:

4.1 Linear Regression

When continuous variables are utilized to predict actual values, linear regression is used. Linear



Regression is a machine learning algorithm that performs a regression task and is based on supervised learning. Linear regression delivers a target prediction value based on independent variables, which is most frequently employed for discovering the connection between variables and forecasting [13]. In the Figure 2 above, on X-axis is the unbiased variable and on Y-axis is the output.

The regression line is the exceptional in shape line for a model. And our major objective on this algorithm is to locate this best in shape line.

Figure 2: Graph of Linear regression

Linear Regression may also lead to over-becoming but it may be averted the use of some dimensionality discount strategies, regularization strategies, and go-validation. It over-simplifies real-international issues via assuming a linear courting a few of the variables, therefore now not encouraged for realistic use-cases(Figure 2).

4.2 Decision Tree

Decision tree is a supervised learning algorithm that is used to describe a choice made in response to a circumstance. It can be used for classification as well as regression. The decision tree is always built from the top down. The Decision Tree Regression is a non linear and non-continuous entity. It's a decision-making function that accepts an attribute values vector as input and returns a decision. It can be used to address problems involving regression as well as classification. A decision tree makes a decision by performing a set of operations [14]. It can work with numerical and categorical features and easy to understand and interpret, visually intuitive.

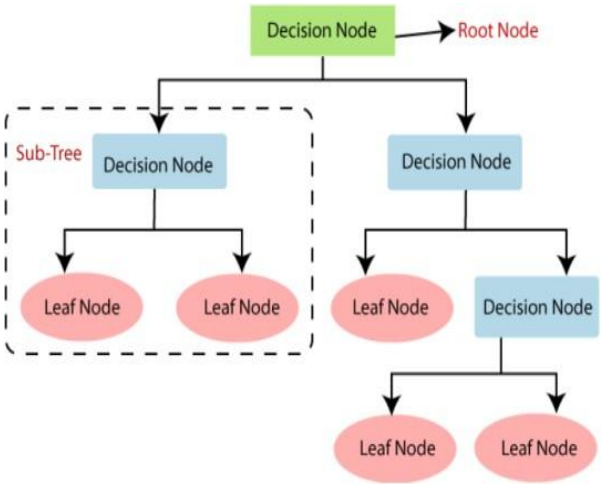


Figure 3: Decision Tree Classification Algorithm.

There are two nodes in every decision tree. One is decision node and the opposite is leaf node. selection nodes have many branches and leaf nodes are the resultant / final results of the decisions that were taken to attain to its point. these choices are made based on a given dataset(Figure 3).

4.3 Random forest

Random forest is a collection of decision trees used for regression and classification. To determine the majority vote, classification is used. The mean value is calculated using regression. This method is more accurate, resilient, and capable of handling a wide range of data types, including binary, category, and continuous data. Random Forest is nothing more than a collection of decision trees. A random forest is a meta-estimator (i.e., it integrates the results of numerous forecasts) that aggregates many decision trees and improves them. The number of functions that can be split on each node is limited to a percentage of the total (known as the hyper parameter) [15]. Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase(Figure 4).

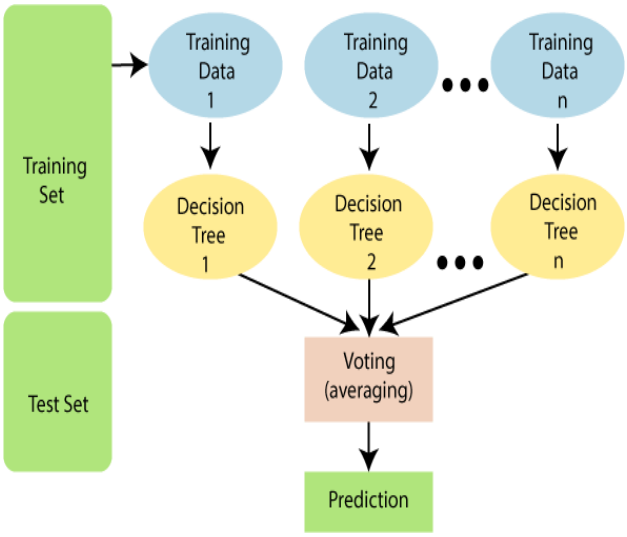
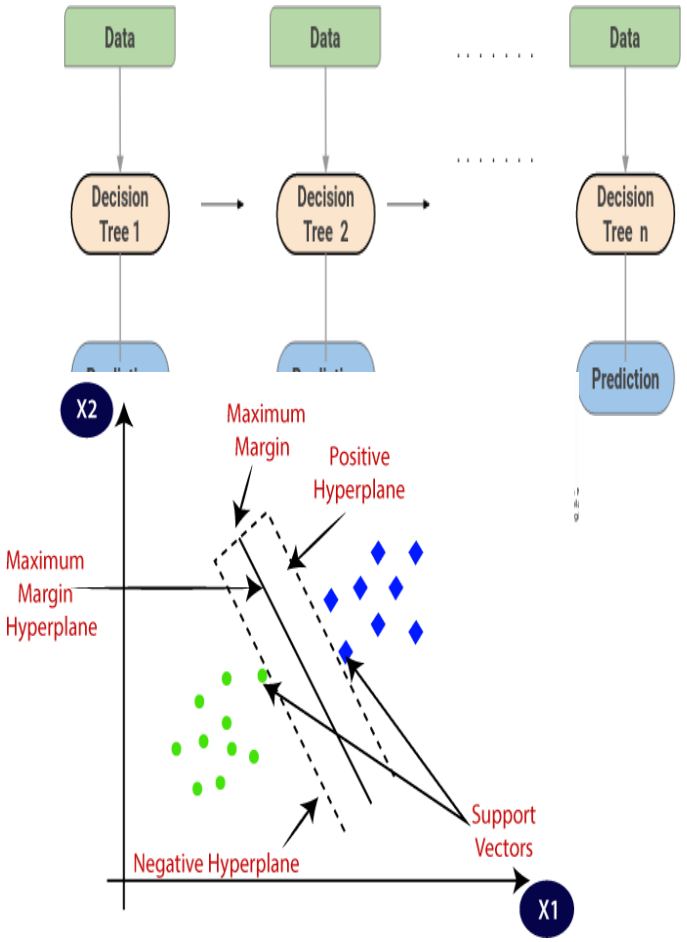


Figure 4: Random Forest Algorithm.

4.4 Support Vector Machines (SVM)

Support vector machines (SVMs, also known as support vector networks) are supervised learning

models that evaluate data and recognize patterns and are used for classification and regression analysis in machine learning. Support vector machines (SVMs) are supervised learning models with associated learning algorithms for classification and regression analysis in machine learning. A non-probabilistic



binary linear classifier, an SVM training technique builds a model that assigns new data measurements to one of two categories [16].

The working of the SVM algorithm can be understood by using an example. Suppose we have a dataset that has two tags (green and blue), and the dataset has two features x1 and x2. We want a classifier that can classify the pair(x1, x2) of coordinates in either green or blue(Figure 5).

Figure 5: Support Vector Machine Algorithm

4.5 Gradient boosting

Gradient boosting is a set of machine learning algorithms that merge a variety of weak learning models to generate a strong predictive model. When doing gradient boosting, decision trees are commonly employed. Gradient boosting models are getting popular as a measure of the ability to classify complex datasets [17]. It is a regression and classification machine learning technique that generates a prediction model in the form of an ensemble of weak prediction models, often decision trees.

A Gradient Boosting Machine or GBM combines the predictions from multiple decision trees to generate the final predictions. Keep in mind that all the weak learners in a gradient boosting machine are decision trees(Figure 6).

Figure 6: Gradient boosting Algorithm.

5 Methodology

Air pollution is a phenomenon, which is affected by different factors. In order for precise prediction, right identification of these parameters influencing the air pollution is necessary. We evaluate the high level aspects of the data more generally before starting a machine learning examination of the data. PM2.5 levels are also more concentrated on weekdays, as can be observed. As a result, it is clear that PM2.5 is influenced by human activity, and that it is especially linked to traffic levels.

5.1 Dataset

The pollutant data: This information was acquired in order to train the algorithm to detect air quality. CarbonMonoxide(CO), Sulphurdi oxide(SO2), and Ozone (O3) were to be included in the data set for 74 cities in China daily data from Oct. 2013 to April.2015 are used for air quality index prediction., are obtained from link http://106.37.208.233:20035/.

Meteorological data: Temperature, Wind Speed, Humidity, and Wind Direction are the meteorological data information set parameters that are utilized to train the system for 74 cities.

5.2 Data Preprocessing

Data preprocessing in machine learning refers to the technique of preparing (cleaning and organizing) raw data to make it fit to build and train machine learning models. Often, data preprocessing is the most important phase of a machine learning project. Data preprocessing is a step in the data mining and data analysis process that takes raw data and transforms it into a format that can be understood and analyzed by computers and machine learning.

6 Experiments

6.1 Results

The results obtained from the test were analyzed and as compared to decide the exceptional performing set of rules .The pc used for the test was an Intel i3-9100 with 16GB RAM and LGA 1151 card. We trained 5 models primarily based on decision Tree, Linear Regression, Random forest, SVM, and GBC. After education the 5 ML models, we provided them with trying out facts set for prediction.

We carried out the distinctive machine learning algorithms in Python the use of Jupyter notebook. the following plot indicates that all the capabilities that are considered for the prediction are correlated and as a consequence can be taken into consideration to teach the model.

- For the CO prediction, we get prediction accuracy as follows

Model Name	Prediction
Linear Regression	0.06%
Decision Tree	0.69%
Random forest	0.91%
Support Vector Machine	0.80%
Gradient Boosting	0.36%

Table 1: Prediction Probability of CO

From the previous table 1, it is clear that the Random forest algorithm is the best for others. To further clarify the results, we represented them with the following graph (Figure 7), which shows the extent of variation between the various applied machine learning algorithms:

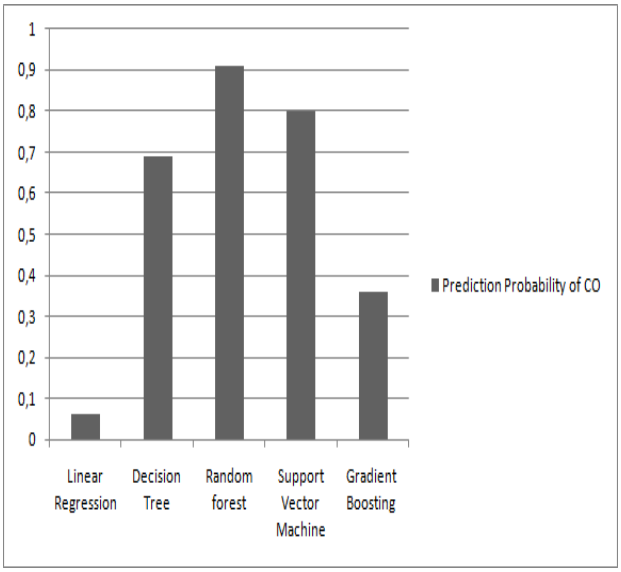


Figure 7: CO prediction probability.

- For the SO2 prediction, we get prediction accuracy as follows

Model Name	Prediction
Linear Regression	0.03%
Decision Tree	0.79%
Random forest	0.96%
Support Vector Machine	0.87%
Gradient Boosting	0.54%

Table 2: Prediction Probability of SO2

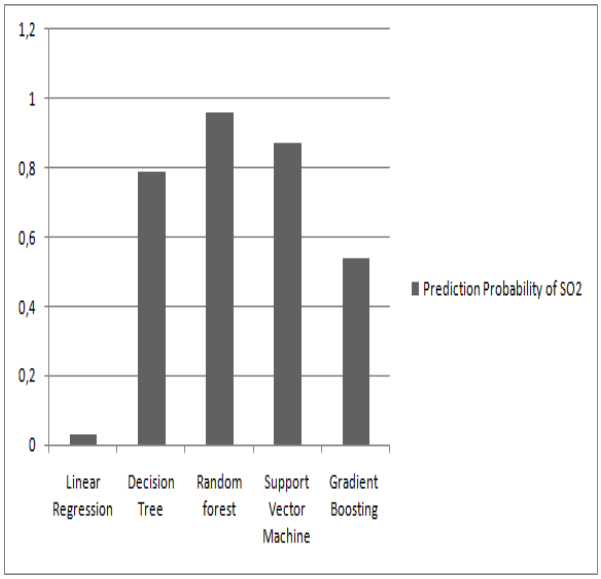


Figure 8: SO2 prediction probability.

From the previous table 2, it is clear that the Random forest algorithm is the best for others. To further clarify the results, we represented them with

the following graph(Figure 8), which shows the extent of variation between the various applied machine learning algorithms.

- For the O3 prediction, we get prediction accuracy as follow

Model Name	Prediction
Linear Regression	0.21%
Decision Tree	0.55%
Random forest	0.87%
Support Vector Machine	0.63%
Gradient Boosting	0.42%

Table 3: Prediction Probability of O3

From the previous table 3, it is clear that the Random forest algorithm is the best for others. To further clarify the results, we represented them with the following graph(Figure 9), which shows the extent of variation between the various applied machine learning algorithms:

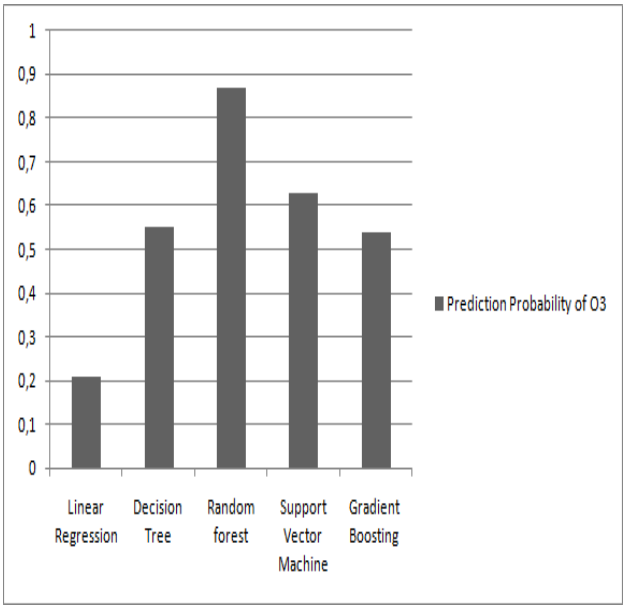


Figure 9: O3 prediction probability.

6.2 Discussion

From the previous tables, it can be seen that the:

- ✓ Linear Regression algorithm has much weaker indicators compared to the other models. This is due to the presence of several outliers and outliers from other data points.
- ✓ Decision Tree algorithm is a class capable of performing multi-class classification on a data set. They perform well even with the model from which the data is generated. But they are very complex and do not generalize well.
- ✓ The Support Vector Machine algorithm has good results. It is more productive in high dimensional spaces and when there is a margin of separation between categories as in our case.
- ✓ The Gradient Boosting algorithm has average results. The training process takes longer due to the fact that trees are built sequentially. It does not work well with sparse data.
- ✓ Random Forest has the advantage of being the easiest to understand and use of the five machine learning algorithms tested. Random Forest also outperforms other strategies in terms of processing time. However, it should be noted that the performance does not compare favorably to that of other models. Random Forest is a multi-tree ensemble approach. By integrating several trees, it lowers the over fitting of single trees. The peak values were identified using this model. Furthermore, the processing time was shorter than that of other models.

For the dataset, Random Forest algorithm performs the best among all five algorithms. We have validated our approach with field trials and have shown the performance comparison against different algorithms.

Through graphs, Random Forest is an efficient algorithm capable of detecting air quality. As shown in Figures 7,8 and 9, Random Forest algorithm made more accurate predictions than other algorithms.

7 Conclusion

We use machine learning algorithms such as linear regression, Decision Tree, Support Vector Machine ,Random Forest (RF) and Gradient Boosting to forecast the air quality index. We determined that the RF algorithm provides a better prediction of the air quality index based on the data. RF train each tree independently, using a random sample of the

data. This randomness helps to make the model more robust than a single decision tree, and less likely to adapt on the training data. There are typically two parameters in RF number of trees and number of features to be selected at each node.

We plan to examine the performance of these strategies in a multi-core context in the future. We also plan to look at the other elements that contribute to air pollution. The use of supervised learning is also being expanded, as is unsupervised learning. In the future, projected architecture can be integrated into a real-time environment once pollution measurement equipment is installed in smart cities.

References:

[1] Ameer, Saba and Shah, Munam Ali and Khan, Abid and Song, Houbing and Maple, Carsten and Islam, Saif Ul and Asghar, Muhammad Nabeel. Comparative analysis of machine learning techniques for predicting air quality in smart cities. *IEEE Access*, 2019, No. 7, pp.128325–128338.

[2] J. Sentian, F. Herman, C. Y. Yin and J. C. H. Wui. Long-term air pollution trend analysis in Malaysia. *International Journal of Environmental Impacts*, 2019, No.2, pp.309–324.

[3] Lu, Weizhen and Wang, Wenjian and Leung, Andrew YT and Lo, Siu-Ming and Yuen, Richard KK and Xu, Zongben and Fan, Huiyuan. Air pollutant parameter forecasting using support vector machines. In *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290)*;Publishing: IEEE, 2002; pp. 630–635.

[4] Iskandaryan, Ditsuhi and Ramos, Francisco and Trilles, Sergio.Air quality prediction in smart cities using machine learning technologies based on sensor data: a review. *Applied Sciences,* 2020, No.10, pp. 2401.

[5] Mart´ınez-Espan˜a, Raquel and Bueno-Crespo, Andres and Timon-Perez, Isabel Maria and Soto, Jesu´s A and Ortega, Andr´es Mun˜oz and Cecilia, Jose M.Air-Pollution Prediction in Smart Cities through Machine Learning Methods: A Case of Study in Murcia, Spain. *J. Univers. Comput. Sci.*, 2018, No.24, pp.261–276 .

[6] Hamami, Faqih and Dahlan, Iqbal Ahmad. Air Quality Classification in Urban Environment using Ma- chine Learning Approach. *In IOP Conference Series: Earth and Environmental Science*;Publishing: IOP Publishing, 2022; pp. 012004.

[7] Fernando, RM and Ilmini, WMKS and Vidanagama, DU.Prediction of Air Quality Index in Colombo.2022.

[8] Abirami, G and Girija, R and Das, Anindya and Sreenivasan, Navneeth. Predicting Air Quality Index with Machine Learning Models. In *Machine Learning and Deep Learning in Efficacy Improvement of Healthcare Systems;*Publishing: CRC Press, 2022; pp. 353–371.

[9] Murugan, Rishanti and Palanichamy, Naveen. Smart City Air Quality Prediction using Machine Learn- ing. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICI- CCS);*Publishing: IEEE, 2021; pp. 1048–1054.

[10] Sinnott, Richard O and Guan, Ziyue. Prediction of air pollution through machine learning approaches on the cloud. In *2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT);*Publishing: IEEE, 2018; pp. 51–60.

[11] Pasupuleti, Venkat Rao and Kalyan, Pavan and Reddy, Hari Kiran and others. Air quality prediction of data log by machine learning. In *2018 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS);*Publishing: IEEE, 2020; pp. 1395–1399.

[12] Freedman, David A.Statistical models: theory and practice. 2009*, cambridge university press* .

[13] Maimon, Oded Z and Rokach, Lior. Data mining with decision trees: theory and applications. Publishing: World scientific, 2014.

[14] Breiman, Leo. Bagging predictors. *Springer* 1996, 24, 123–140 .

[15] Zhao, Zhongliang and Carrera, Jose and Niklaus, Joel and Braun, Torsten. Machine Learning-Based Real-Time Indoor Landmark Localization. In *International Conference on Wired/Wireless Internet*

*Communication;*Publishing : Elsevier, 2018; pp. 95–106.

[16] Zhang, Yanru and Haghani, Ali. A gradient boosting method to improve travel time prediction. *Trans- portation Research Part C: Emerging Technologies,* 2015, 58, 308–324.

[17] Vu, Tuan V and Shi, Zongbo and Cheng, Jing and Zhang, Qiang and He, Kebin and Wang, Shuxiao and Harrison, Roy M. Assessing the impact of clean air action on air quality trends in Beijing using a machine learning technique. *Copernicus GmbH,* 2019, 19, 11303–11314 .

Contribution of individual authors to the creation of a scientific article (ghostwriting policy)

Kamel, Brahim carried out the simulation and the optimization.
Kamel has implemented the Algorithms .
Kamel has organized and executed the experiments of Section 6.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)