Predicting AQI Using Meteorological and Particulate Matter Data: A Comparative Analysis

Sharika Anuj

Received November 14, 2024
Accepted February 11, 2025
Electronic access March 31, 2025

AQI (air quality index) is an important metric for understanding air quality and its effect on our health. Although there is significant research on predicting AQI values using deep learning methods, several models have neither been tested nor compared for their efficacy in predicting AQI values. In this paper, different machine learning models, including Naive Bayes, Perceptron, Multilayer Perceptron, Random Forest, K-nearest Neighbor, and Gradient Boosting, are compared to determine the best models for predicting AQI values. The methodology involves feeding in a dataset containing meteorological and particulate matter data. Many of the models needed hyperparameter tuning in order to deliver the optimal accuracy when predicting the AQI values. Based on testing, the Gradient Boosting method and Random Forest Algorithm, both belonging to the Ensemble methods, performed the best, with the smallest mean absolute error of all six: 0.07 and 0.02, respectively. These findings suggest that the Ensemble methods are the best in performing such predictions. This experiment aimed to find the most accurate model in determining AQI values. These models can be implemented in special tools in order to give policymakers and government authorities the ability to accurately assess the state of the air quality in a certain environment and take action to reduce its effects on their citizens.

Keywords: AQI (Air quality index), machine learning, mean absolute error (MAE), regressive machine learning model

Introduction

Air quality is a major global concern today (Ahmadi et al., 2015)[1]. Almost seven million people die every year due to air-pollution related diseases (Ghada Sahbeni et al., 2019)[2]. In light of this fact, avoiding such health risks is very important. To do so, we must be capable of determining accurate AQI (air quality index) values and share this information with the public. As deep learning is gaining traction in the scientific world and is surpassing human intelligence (Nowak et al., 2018)[3], investigation into its uses in AQI value prediction has the potential to address common health and respiratory issues.

This paper sought to determine the best performing regressive machine learning models for predicting AQI levels in California. In addition to my findings, one can infer from my research which specific models are best for performing other tasks similar to predicting AQI levels. Referring to my research, the best performing models can also be applied to projects that require efficiency in recognition of trends and patterns in datasets.

Main Objectives

- This research implements machine learning models AQI values prediction using meteorological and particulate matter data
- The performance evaluation of the machine learning models are based on mean absolute error (MAE)

Machine learning methods are commonly used for prediction of many different outcomes such as cognitive disease (Bratić, Kurbalija & Ivanović, 2018)[4], greenhouse gas emissions (Abderrachid et al., 2020)[5], and much more. Many research papers have already sought to find solutions using deep learning and machine learning techniques to accurately calculate AQI levels. However, prior research does not cover specific regressive machine learning and algorithms such as Naive Bayes, Random ForesT, Perceptron, MLP (Multilayer Perceptron), KNN (K-nearest neighbor), and Gradient Boosting. My research was necessary to test these models and how they performed with respect to predicting AQI levels. Existing literature on prediction of AQI values used data from satellite imagery instead of ground surfacing to train their model (Rowley & Karakus, 2023)[6]. Using this data, they sought to create a machine learning model which predicted AQI values with high accuracy. Their results showed that the model had overestimated the AQI values of testing data input by 20%. Another research paper attempted to predict AQI values using different data sources: real-time sensors detecting levels of polluting gases in the nearby environment. Since this data was being tracked in real-time, their methods of prediction involved time-series analysis and linear regression (Mani et al., 2021)[7]. Another paper used data sources from particulate matter concentration such as carbon monoxide to determine AQI levels (Kataria & Puri, 2021)[8]. They use different machine learn-

ing algorithms, including Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), and Artificial Neural Networks (ANN). Some researchers have already conducted tests with meteorological data in addition to particulate matter levels, albeit with different algorithms. For instance, one paper discussed the use of artificial neural networks (ANN algorithms) to determine AQI values using meteorological data (Maleki et al., 2019)[9]. Another paper conducted research using a Long Short-Term memory (LSTM) based deep learning model with a Support vector regression model also using meteorological data (Janarthanan et al., 2021)[10]. One paper argues that including meteorological data as part of the training dataset is important, as this could play a role in AQI levels in certain areas (Ji et al., 2020)[11]. It has been proven that sunshine duration can affect the AQI value in an environment because it is an important thermal factor which can influence "temperature inversion" (the existence of warmer air over cooler air), and impact the presence of pollutants (Zhang et al., 2019)[12]. Another meteorological factor which affects the AQI is humidity, since it can prevent the pollutants in the air from spreading and reducing their concentration. Much of the literature mentioned at the beginning does not include meteorological data. On the other hand, the literature which did include this data, only implemented and assessed a limited number of algorithms. Many more models needed to be tested with different data set combinations to be able to determine not just an accurate method of predicting AQI levels, but also be able to conclude which model performed the best. Such a model may also perform well in tasks that involve the analysis and prediction of trends and patterns in large datasets such as the one that is used in this paper.

The following study is organized into different sections: Section II consists of the dataset description, information on the models, and implementation, Section III includes the results from the experiment, Section IV includes the discussion and in-depth analysis of the results, and Section V includes the conclusion of the paper.

Methodology

A. *Dataset Description*

In this project two datasets were used: the 2020 California Environmental Conditions Kaggle dataset (as shown by Figure 1) [11] and the 2020 California Air Quality Kaggle dataset (As shown by Figure 2) [12]. The first one is an environment conditions dataset that was gathered by CIMIS weather stations. This dataset comprises 14 numerical values for meteorological data (such as maximum/ minimum air temp, solar radiation, and average wind speed). These values were listed daily from 2018 to 2020. The second dataset comprises information on PM2.5 concentration (fine particulate matter) and AQI values for each day of 2020. Since the first dataset had more information than the other, both had to be sorted according to the recorded date and location of the weather station. This was done by converting them into csv files and importing the pandas library.

	Stn Id	Stn Name	CIMIS Region	Date	ETo (in)	\
7	109	Carneros	San Francisco Bay	2020-02-01	0.07	
16	140	Twitchell Island	San Francisco Bay	2020-02-01	0.05	
25	157	Point San Pedro	San Francisco Bay	2020-02-01	0.07	
47	47	Brentwood	San Francisco Bay	2020-02-01	0.07	
62	170	Concord	San Francisco Bay	2020-02-01	0.05	
...	...	...	...	...	...	
30472	253	Pescadero	San Francisco Bay	2020-09-18	0.15	
30488	170	Concord	San Francisco Bay	2020-09-18	0.15	
30489	171	Union City	San Francisco Bay	2020-09-18	0.15	
30497	191	Pleasanton	San Francisco Bay	2020-09-18	0.17	
30508	211	Gilroy	San Francisco Bay	2020-09-18	0.20	

Fig. 1 Section of Dataset 1

# Daily Mean PM2....	# DAILY_AQI_VALUE	Site Name
Mean measurement for the day.	Air quality index	
-3 379	0 419	Fresno - Garland 2% Stockton-Hazelton 2% Other (29412) 96%
8.6	36	Livermore
4.5	19	Livermore
14.2	55	Livermore
10.9	45	Livermore
7.8	33	Livermore
6.2	26	Livermore

Fig. 2 Section of Dataset 2

B. *Models*

This study uses algorithms chosen from a wide variety of commonly used machine learning models to test their performance.

1. ***Naive Bayes***

   Based on the Bayes Theorem, the Naïve Bayes model is used in a wide variety of classification tasks. This model works on the assumption that every feature makes an independent and equal contribution to the outcome. Though this may not usually be true in the real world, this algorithm may prove to be helpful when analyzing AQI values based on different meteorological values and particulate matter data. The Bayes Theorem allows this model to make connections between input and output values using conditional probability and likelihood. An advantage of

the Naive Bayes model is that this algorithm doesn't require a lot of training data. However, this model is generally used more for classification tasks and less for pattern analysis (Raj, Goswami, Mhatre & Agrawal, 2024)[13]. In a real-world scenario, Naive Bayes can be used for making real-time predictions due to its efficiency.

2. ***Random Forest***

   Part of the ensemble methods, the Random Forest machine learning model makes predictions based on multiple decision trees by combining their outputs to reach a final result. When analyzing complex datasets with many inputs, a random forest algorithm can produce high accuracy levels. As the combined datasets in my research contain more than fifteen parameters, the Random Forest Algorithm was chosen for its efficiency. The number of estimators, which is a parameter for the Random Forest algorithm, was increased and improved the accuracy level of the program. Random Forest also has the ability to determine which are the most important features in the dataset that have the most pronounced effect on the final output (Rigatti et al., 2017)[14].

3. ***Perceptron***

   The Perceptron model is a single-layered neural network used for supervised learning of various binary classifiers. This algorithm is known as a Feedforward Neural Network. These are known to be able to learn complex patterns within the dataset. Due to its efficiency with using binary classifiers, Perceptron is recommended for use in binary sorting and classification tasks. Additionally, with its single-layered feature, this algorithm is used for linear regression tasks and performs well on problems with logical operations. However, there is risk of overfitting with the data and could be more difficult to train. Since my dataset contains several meteorological data parameters as well as multiple particulate matter values, this model may not be the most capable for determining the AQI values. Our dataset may require a model or neural network that is capable of solving nonlinear problems due to the existence of multiple parameters in the data set. While working with this model, the parameters were tweaked to deliver the highest performance level. Even so, the performance plateaued and could not be improved.

4. ***Multilayer Perceptron (MLP)***

   he Multilayer Perceptron model (MLP), similar to Perceptron, is also a feedforward neural network. However, MLPs are multi-layered (Figure 3) and are thus capable of solving nonlinear problems. This may prove to be especially helpful when dealing with multiple input values, similar to the dataset used in this paper. Being multi-layered, this model can analyze complex trends and patterns between the different input values (in this case, meteorological data and particulate matter values) (Singh & Bannerjee, 2019)[15]. Due to its capability of learning nonlinear relationships in the data, this is a powerful model for classification, regression, and pattern recognition. In particular, pattern recognition is essential for accurate AQI value prediction. However, MLPs are harder to train due to the amount of data that needs to be fed and the hyperparameter tuning needed. This also results in MLPs taking longer to converge.

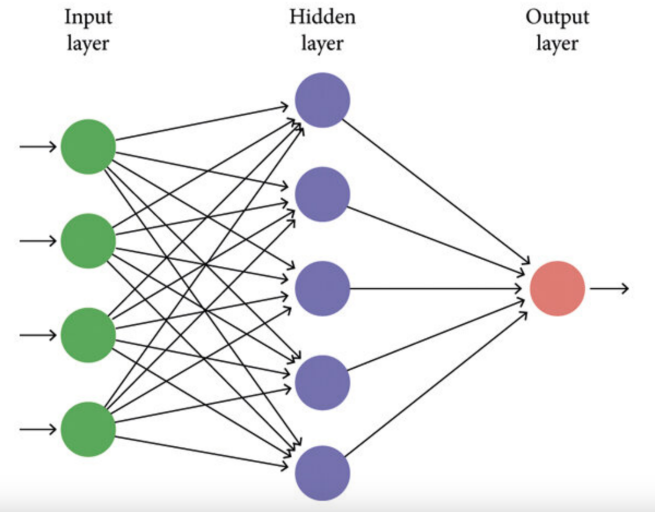

Fig. 3 Structure of MLP Model

5. ***K-nearest Neighbor***

   The K-nearest neighbor model, often referred to as "KNN", is an algorithm with a simpler structure when compared to MLP. This algorithm addresses queries by analyzing classes or values of the neighboring nodes. Because of the algorithm's structure, KNN is mainly used for classification tasks instead of predicting trends and patterns or conducting time-series analysis. This approach allows the model to make predictions based on the local structure of the data. KNN makes these analyses by taking the average of the neighbors or using the value attached to the majority of the neighbors (Zhang et al., 2016)[16]. This trait makes this algorithm more forgiving towards outliers within the dataset. KNN is used in applications such as fraud detection or image recognition systems.

6. ***Gradient Boosting***

Similar to the Random Forest model, the Gradient Boosting Model is also part of the ensemble methods. Gradient Boosting trains models in a step-by-step fashion: this algorithm improves the model in each step, trying to improve its accuracy each time and bringing down the error (Figure 4). In the end, the method involves the calculation of the average of all the outputs from each model that was used. In other words, Gradient boosting combines multiple weak learning models to create one that is much stronger (Bentéjac, Csörgő, & Martínez-Muñoz, 2020)[5]. The model is able to analyze complex relationships between the values in the data and is usually not prone to overfitting. This approach is best suited for classification tasks or complex regression algorithms.

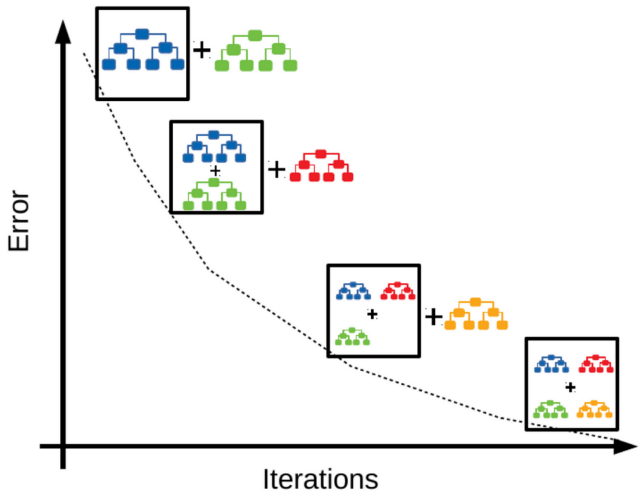

Fig. 4 Structure of Gradient Boosting

C. *Procedure*

Overview of Methodology

1. Data collection
2. Data pre-processing
3. Data split
4. Model training
5. Model testing
6. Evaluation metrics
7. Graphical analysis of data

The first step towards training the models is data collection. Data is gathered from reputed weather sources, and platforms (such as Kaggle). In this project, the data used included weather information (dataset 1) and particulate matter values (dataset 2). The next step is to make sure the entire data is combined into one concise set that can be loaded into a model. With the help of the pandas library, this can be done by sorting the datasets by the same rows and columns. Missing values in the dataset were handled using imputation techniques. For numerical columns, missing values were replaced with the mean of the respective column. This ensures the dataset remains statistically consistent. Python's pandas library was used for this operation, employing the *fillna()* method. The dataset was normalized using the Min-Max scaling technique to ensure all features contributed equally to the model. This method scales the data to a range of [0, 1], which improves model convergence during training. Normalization was performed using *sklearn.preprocessing.MinMaxScaler*.

The analysis in this study uses models imported from the *sklearn* library. Next, the data is separated into training and testing datasets in an 80-20 ratio. The training datasets must be loaded into the model, while the testing datasets are used to determine the mean absolute error (the average measure of the difference between predicted and actual values). Each model has been trained with specific and unique parameters in order to ensure the model's optimal performance. Though the Naive Bayes does not require any parameters, the other five models need them.

The Random Forest Regressor model needs the *number of estimators* to be set to 20 and the random state to be set to 42.

The Perceptron needed the following parameters to be set: max iterations as 1000, float stopping criterion (tol) as 0.003, eta0 (learning parameter) as 0.01, and random state as 42.

The MLP regressor needed the following parameters: hidden layer sizes as 100, 50, and 25, maximum iterations as 1000, alpha as 0.01, solver as "adam", learning rate as "adaptive", activation as "logistic", and random state as 42.

The only parameter needed for the K-Nearest neighbors regressor is the number of neighbors to be set to 20.

For the last model, Gradient Boosting, the only parameter required is the number of estimators which is set to 100.

The above parameters were used to optimize the performance of the different models. Once trained using the dataset, the model is evaluated for generalization using the testing dataset. The criteria used for this evaluation include the mean absolute error (MAE) and root mean squared error (RMSE). MAE is defined as the average of the size of error of the values that are predicted in comparison to the actual values, as shown by Eq. (1).

Other options to evaluate each model's accuracy include mean squared error (MSE) and root mean squared error

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \quad (1)$$

(RMSE). MSE is the average of the squares of the error as shown by Eq. (2). RMSE, on the other hand, is simply the square root of MSE as shown by Eq. (3).

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \quad (2)$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \quad (3)$$

The squaring of the measure error in MSE gives too much weight to the outliers within the data which is deemed unnecessary for this comparative analysis (Wilmot & Matsuura, 2005)[17]. Though RMSE also includes the squaring of the error in its formula, the square root of the entire sum reduces the weightage given to the outliers. Thus, MAE and RMSE were chosen as the evaluation criteria for this study due to its ability to assess the prediction accuracy without giving too much weight to the outliers within the results.

The results are provided graphically for the analysis. This can be done by importing the *matplotlib* library to produce a bar graph as shown below (Fig 5 & 6). The actual values from the data and the predicted values are shown for comparison (Fig 7). The overall structure of the study is experimental with some cross-sectional elements.

This study analyzed the performance of the multiple methods to evaluate the model with highest accuracy for predicting AQI. This experimental study involves controlled evaluation metrics, manipulation of algorithmic approaches, and observation of outcomes. Feature selection was performed to identify the most impactful predictors for AQI prediction. The Random Forest model's built-in feature importance metric was utilized to rank the features. Features contributing less than 0.01 importance score were excluded. This ensures the model focuses on variables with higher predictive power, enhancing its performance and reducing overfitting. The study contains cross-sectional elements because of the analysis of data from a single period and the capturing of the performance of each algorithm at that moment.

Results

Using the code that exercised these models, followed by tests that were run to determine the accuracy of each of the six models (Naive Bayes, Random Forest algorithm, Perceptron, MLP, KNN, and Gradient Boosting), the ensemble methods outperformed the others, delivering a mean absolute error of less than 0.1 and a root mean squared error of less than 1.0. The Random Forest algorithm was the best performing considering the mean absolute error of 0.02 and performed second best considering the root mean squared error of 0.54. Gradient Boosting performed the second best considering the mean-absolute error of 0.07 and performed the best considering the RMSE of 0.38. Naive Bayes produced a mean absolute error of 2.01 and a RMSE of 3.51, while the Multilayer Perceptron model produced a mean absolute error of 2.84 and a RMSE of 3.88. The KNN and Perceptron models had the lowest performance, delivering an mean absolute error of 14.10 and 22.70, respectively, and RMSE of 20.04 and 30.32. In Figure 7, the comparative values for actual versus predicted with Random Forest (RF) Model are shown. The line graphs overlap closely and are indistinguishable.

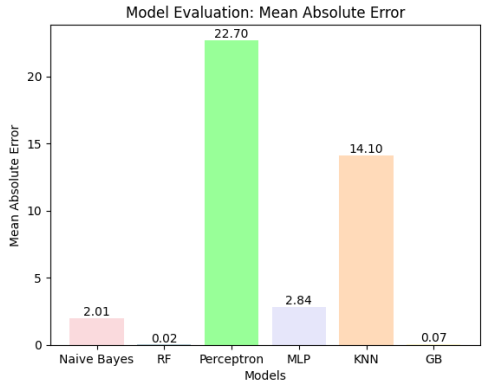

Fig. 5 Model Evaluation (MAE)

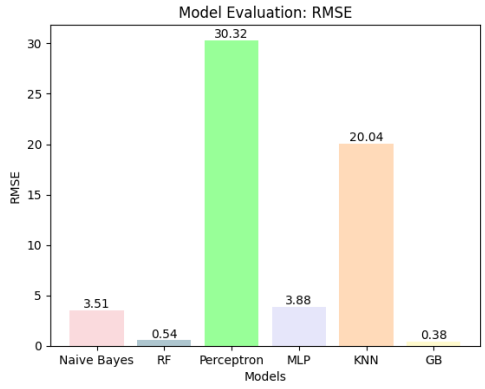

Fig. 6 Model Evaluation (RMSE)

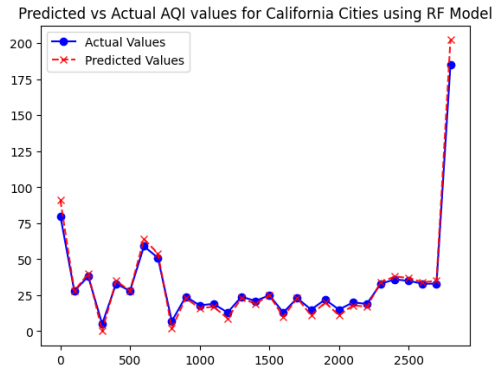

Fig. 7 Predicted vs Actual AQI values for California cities using RF Model

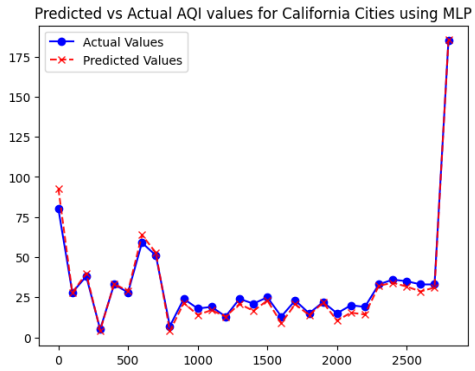

Fig. 10 Predicted vs Actual AQI values for California Cities using MLP

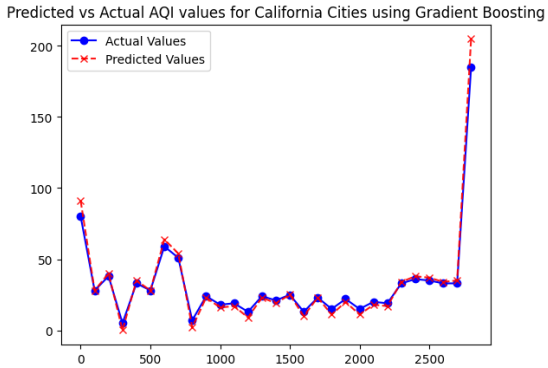

Fig. 8 Predicted vs Actual AQI values for California Cities using Gradient Boosting

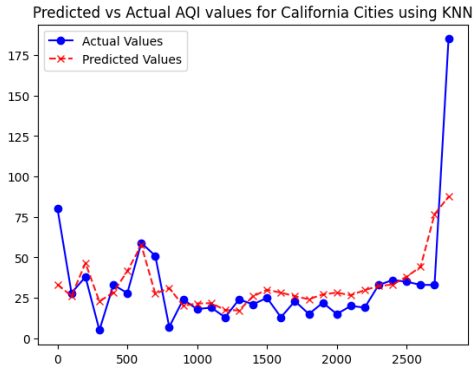

Fig. 11 Predicted vs Actual AQI values for California Cities using KNN

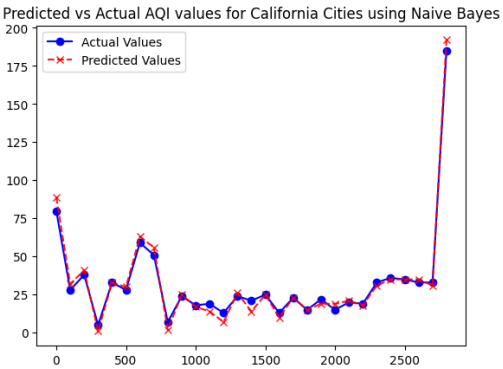

Fig. 9 Predicted vs Actual AQI values for California Cities using Naive Bayes

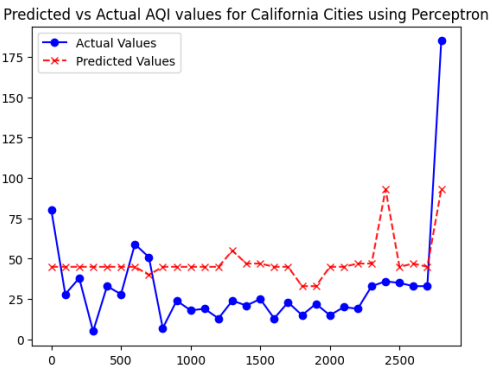

Fig. 12 Predicted vs Actual AQI values for California Cities using Perceptron

Discussion

The Random Forest and Gradient Boosting models, both being part of the ensemble methods, performed the best out of the six models that were tested as shown in Figure 5 and 6, delivering the highest mean absolute error. When the gradient boosting and random forest's predictions were compared to the actual values, as shown in Figures 5 and 6, the difference was smaller when contrasted with the other models' performance. The ensemble methods are methods that combine multiple models to produce a final model that is much better than the rest. These methods aggregate two or more learners (such as regressive learning and neural networks) to produce a better model. This approach of improving simple models to create more complex and more

accurate ones can be used to prove how the Random Forest and Gradient Boosting algorithms performed the best overall. Though the Random Forest is usually preferred for classification tasks, by using the Regression function, the model was still able to produce high accuracy levels. The Gradient Boosting function is used for regression tasks in addition to classification tasks. This model performed well because it was able to handle datasets with non-linear relationships and complex trends and patterns. The Gradient Boosting showed better results than Random Forest when evaluated using RMSE. This piece of data illustrates that Gradient Boosting had fewer outliers when compared to Random Forest.

The reason the Naive Bayes model did not perform as efficiently as the Random Forest and Gradient Boosting models, as shown by Figure 5, 6, and 9, is because the approach used is not modeled on real-life situations, and is therefore inaccurate. The Naive Bayes' approach treats each and every variable as independent of each other. This assumption cannot always be applied to problems involving the complex relationships between variables in real-life situations.

The K-nearest neighbor model also did not perform well, seen in Figure 5, 6, and 11, since the algorithm is mainly used for classification tasks instead of regression problems. The approach involves the analysis of the nodes nearest to the current cell. Through this analysis, certain conclusions can be deduced to determine under what classes / groups the current node falls into. Because of this classification-based approach, this model was unable to produce the best results out of all the models tested.

The Perceptron and MLP (multilayer perceptron) models were least suited for the topic of this paper. The perceptron model is generally used for binary classification tasks and thus was not able to produce the best results, seen in Figure 5, 6, and 12. This project did not require classification skills, but rather regression skills. On the other hand, the MLP model is known for being able to analyze complex trends and patterns within the dataset. These characteristics for this approach made them suitable for analyzing the complex AQI value dataset. However, the reason MLP did not produce the best results, as shown in Figure 5, 6, and 10, is because it is very hard to train. Even after a lot of hyperparameter tuning, the accuracy could not be improved. For this reason, the MLP model would be hard to implement in real-life situations and solutions.

When the Random Forest model was used to predict the AQI values of stations in California across many days in 2020, there was barely any difference when compared to the actual values as shown in Figure 7. This highlights the level of accuracy with the Random Forest Regression, and is therefore the best model to be used when predicting AQI levels. Being an ensemble method, the Gradient Boosting model can also be used due its similarity in accuracy levels.

Conclusion

This paper sought to determine the best regressive machine learning algorithms for determining AQI values accurately using weather data. The original hypothesis was that the Multilayer Perceptron model would outperform the other six models. However, the findings in this paper show that Random Forest and Gradient Boosting models performed the best due to their smaller computational resources and ability to deal with complex data, showing better accuracy. Additionally, the MLP model proved to be much harder to train as compared to the others since arriving at a high accuracy using hyperparameter tuning was challenging. A limitation to this paper is that this research only tested two ensemble methods, so future work can be implemented to see whether there are other deep learning models that outperform Random Forest Regressor and Gradient Boosting. By conducting enough tests and determining the most accurate model for AQI value prediction, more opportunities will open up to improve our deteriorating environment and make lasting changes in our daily lives to prevent more damage. To expand on this research, one can seek to try other deep learning models that could outperform the ones listed in this paper. Additionally, using the data presented in this paper, one can seek to perform an in-depth analysis of the effect of certain weather data and meteorological factors on AQI values in an environment. This can be done using the Random forest model to produce a plot showing the impact of each feature within the data. This can also be tested by using the grid search technique, where the model is trained on various key hyperparameters. The results of the performance of the models in this paper can be further studied to compare their performance during different seasons or geographical regions. Another possible improvement is to utilize these high-performing models to determine other possible factors that affect AQI values in a specific area and conduct in-depth ablation studies.

Acknowledgment

Thank you for the guidance of Maria Stamatopoulou from University College London in the development of this research paper.

References

1 A. Ahmadi, M. Abbaspour, R. Arjmandi and Z. Abedi, *Air Quality Risk Index (AQRI) and its application for a megacity*.

2 G. Sahbeni, G. Nagy and T. Gurgenidze, *Examination of Air Quality Indexes (AQI) role in urban air quality assessment*.

3 A. Nowak, P. Lukowicz and P. Horodecki, *Assessing Artificial Intelligence for Humanity: Will AI be the Our Biggest Ever Advance ? or the Biggest Threat [Opinion*.

4 B. Bratić, V. Kurbalija and M. Ivanović, *Machine Learning for Predicting Cognitive Diseases: Methods, Data Sources and Risk Factors*, https://doi.org/10.1007/s10916-018-1071-x.

5 C. Bentéjac, A. Csörgő and G. Martínez-Muñoz, *A comparative analysis of gradient boosting algorithms*.

6 O. Karakuş and A. Rowley, *Predicting air quality via multimodal AI and satellite imagery*.

7 G. Mani, J. K. Viswanadhapalli and A. A. Stonie, *Prediction and Forecasting of Air Quality Index in Chennai using Regression and ARIMA*.

8 A. Kataria and V. Puri, *AI- and IoT-based hybrid model for air quality prediction in a smart city with network assistance*.

9 H. Maleki, A. Sorooshian, G. Goudarzi, Z. Baboli, Y. Tahmasebi Birgani and M. Rahmati, *Air pollution prediction by using an artificial neural network model*.

10 R. Janarthanan, P. Partheeban, K. Somasundaram and P. Navin Elamparithi, *A deep learning approach for prediction of air quality index in a metropolitan city*.

11 M. Ji, Y. Jiang, X. Han, L. Liu, X. Xu, Z. Qiao and W. Sun, *Spatiotemporal Relationships between Air Quality and Multiple Meteorological Parameters in 221 Chinese Cities*.

12 Y. Zhang, *Dynamic effect analysis of meteorological conditions on air pollution: A case study from Beijing*, https://doi.org/10.1016/j.scitotenv.2019.05.360.

13 R. Kumar, B. K. Goswami, S. M. Mhatre, S. Agrawal and M. Issue.5, *Naive Bayes in Focus: A Thorough Examination of its Algorithmic Foundations and Use Cases*, https://doi.org/10.38124/ijisrt/IJISRT24MAY1438, PP :-2078-2081:-.

14 S. Rigatti.

15 J. Singh and R. Banerjee, 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC, Erode, India, pp. 35–40,.

16 Z. Zhang, *Introduction to machine learning: K-Nearest Neighbors*.

17 C. Willmott and K. Matsuura, *Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance*.