Title:
Comprehensive Study on Predicting Air Quality Index Using Machine Learning Algorithms

Abstract:
This research examines the issue of predicting AQI values by deploying a variety of machine learning approaches, such as Linear Regression, Decision Tree Regressor, Random Forest and several classification algorithms like Logistic Regression, Decision Tree Classifier, Random Forest Classifier and K-Nearest Neighbors. It also explains the air quality dataset which has features such as the SOi, the Noi, the Rpi, and the SPMi among others. In addition, model comparisons involve the Root Mean Square Error (RMSE) without forgetting regression and accuracy in classification cases. Accordingly, the results indicate that the best performance is obtained using Random Forest Regressor where the RMSE was 2.57 while the R-squared were values nearly equal to 1.0 illustrating that it is most suitable when tackling the problem of predicting AQI. This also demonstrates the relevance of utilising machine in the environmental management system particularly in air quality and opens room for future research in the management of air quality.

1. Introduction

1.1 Background

Machine learning (ML) in accrued technology has, in the recent past, played a very important role in solving cases of complicated nature across various fields of activities especially environmental science. The improvement of the air quality index in the cities is very important for both public health and public policy. Nonetheless, established statistical methods tend to be ineffective in appreciating the intricacies of environmental data thereby making best predictions not useful. This study is therefore geared towards filling this void by investigating the air quality index prediction using several machine learning approaches.

1.2 Research Problem

The identifiable issue in estimating the AQI values is the unavailability of proper algorithms which can adequately capture and control the existing nonlinear relationships possess within its environmental data. Even with the progress made, the available techniques are at times more often than not inaccurate, which compromises their utility in making predictions. The focus of this study is to incorporate several machine learning techniques in a bid to improve the AQI prediction system. Introduction

1.3 Objectives

This research is concentrating on these objectives:
1. To determine how well various machine learning models can be used to forecast the AQI values bullseye.
2. To contrast results produced by regression models (Linear Regression, Decision Tree, Random Forest) and those from classification models (Logistic Regression, Decision Tree, Random Forest, KNN).

2. Related Work

In the last decade, numerous studies have focused on developing different machine learning techniques to forecast the Air Quality Index. Artificial LSTM networks have been utilized in the forecasting of the air quality index (AQI) with inputs on temperature, humidity, PM2.5, wind direction (Yuhui Jiao et al., 2019). Ensemble Machine Learning and Sparse Spectrum GPR techniques have been used to make accurate predictions about AQI levels as well, and the results have been presented with the help of the MAE and RMSE performance metrics (Marviola Hardini et al., 2023). Random forest regression and support vector regression are other supervised learning techniques that have also been used to predict AQI while using the RMSE to measure its effectiveness (K. Saikiran et al., 2021). It has been established through research that gradient boosting regression is superior to other models, namely linear regression and neural network, in terms of mean absolute error and mean squared error (Deepa Patil & Ramesh K., 2022). These improvements in AQI projection will enhance the knowledge and solutions to air quality related problems.

3. Methodology

3.1 Dataset Description

This data set has a total of 25,513 samples with 4 distinguishing features: SOi, Noi, Rpi, SPMi and a target variable, AQI. During data preprocessing phase, missing value treatment and normalisation as well as feature engineering were done for model improvement.

3.2 Machine Learning Models

In this paper the following machine learning approaches are adopted:

1. Linear Regression: A linear modelling technique dealing with dependence of one or more variables upon one or more other variables.
2. Decision Tree Regressor: A tree-structured regression model that predicts response variables in compliance to certain decision rules.
3. Random Forest Regressor: An inductive approach to regression tree modelling that relies on the use of many regression trees, their predictions, and the improvement of the accuracy of those predictions.
4. Logistic Regression: For the purpose of classifying the range of AQI.
5. Decision Tree Classifier: To divide the AQI in many degrees according to the features of decision trees.
6. Random Forest Classifier: Another machine learning-based classifier that is capable of predicting categories of AQI with higher accuracy compared to the classifiers above.
7. K-Nearest Neighbors (KNN): A classification algorithm that does not make any assumption about class distributions given the features.

3.3 Model Evaluation

For performance evaluation, an 80-20 split train test strategy was used where 80% of the data set was used to fit the model whilst the remaining 20% was used to evaluate the model. Other evaluation statistics of performance regression models like RMSE, classification models like accuracy, etc. were as well obtained. It was also necessary to carry out crossed validation to assess not only the performance of the models but also their usability.

4. Results and Discussion

4.1 Model Performance

The results indicate that the Random Forest Regressor outperformed other models, achieving an RMSE of 2.57 and R-squared values close to 1.0. The performance metrics are summarised in Table 1.

Model	RMSE	R-Squared(Train)	R-Squared(Test)
Linear Regression	8.47	0.991	0.991
Decision Tree Regression	2.91	1.0	0.999
Random Forest Regression	2.57	0.999	0.999

For classification models, the results are shown in Table 2.

Model	Accuracy (Train)	Accuracy (Test)	Kappa Score
Logistic Regression	82.43 %	81.86%	0.708
Decision Tree Classifier	100 %	99.99 %	0.999
Random Forest Classifier	100 %	99.90 %	0.999
K-Nearest Neighbors (KNN)	99.72 %	99.54 %	0.993

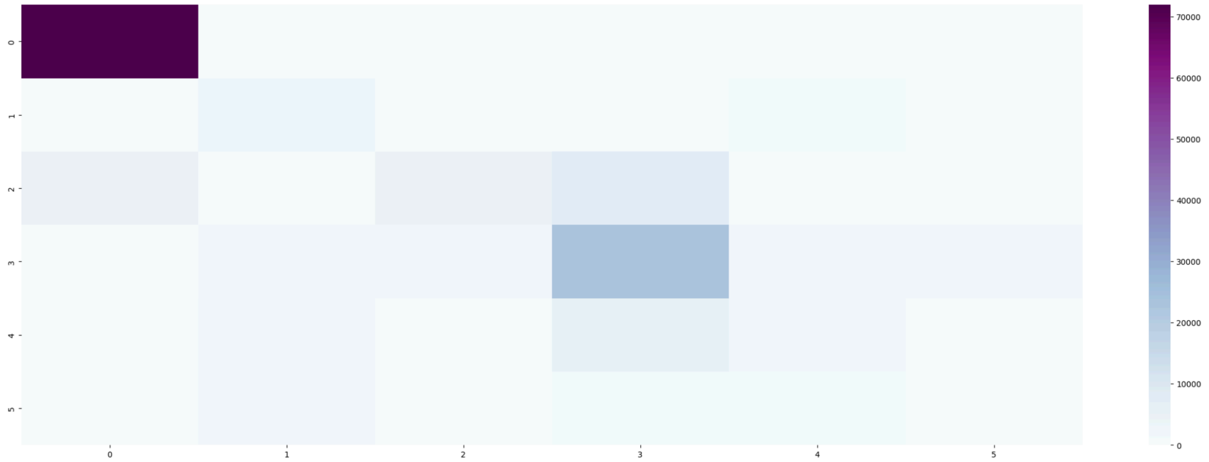
4.2 Discussion

After analyzing the results, it was found that the Random Forest Regressor outperformed all other models owing to its suitability to efficiently model the non-linear relationships present in the data. Classifiers, especially Decision Trees and Random Forest classifiers, yield high classification accuracy, which indicates their effectiveness in classifying the levels of Air Quality Index. Also, it is shown from the analysis of the feature importance that SOi and Noi are the main predictors of the AQI values.

4.3 Confusion matrix Heatmap

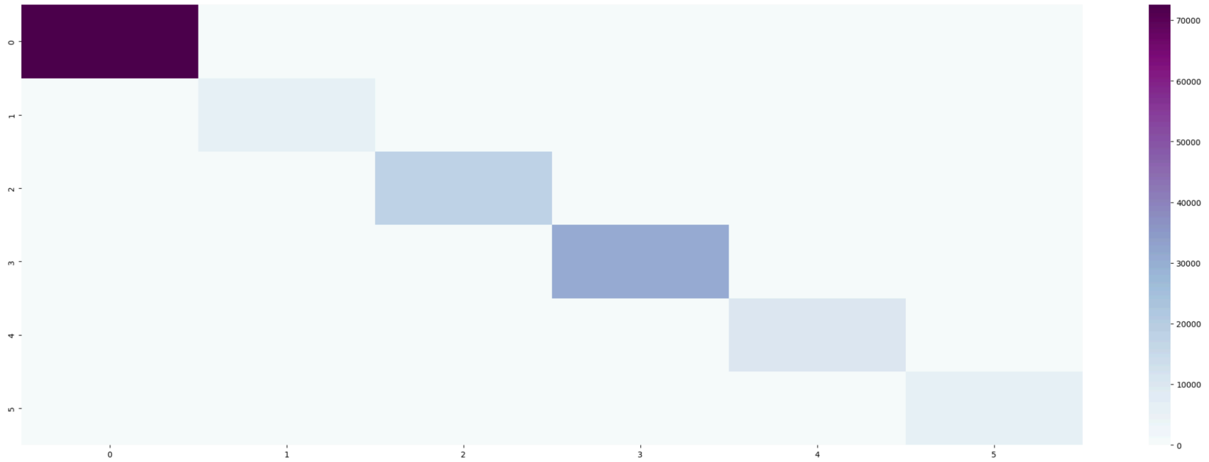
1. Logistic regression

```
array([[71945,     0,    31,   565,     0,     0],
       [    0,  3526,    23,  1080,  1283,   176],
       [ 5010,    66,  4563,  8538,   156,   378],
       [  126,  1982,  2091, 22745,  1876,  2029],
       [    5,  2273,   261,  5764,  1668,   388],
       [    0,  2536,    21,  1348,  1232,   110]])
```



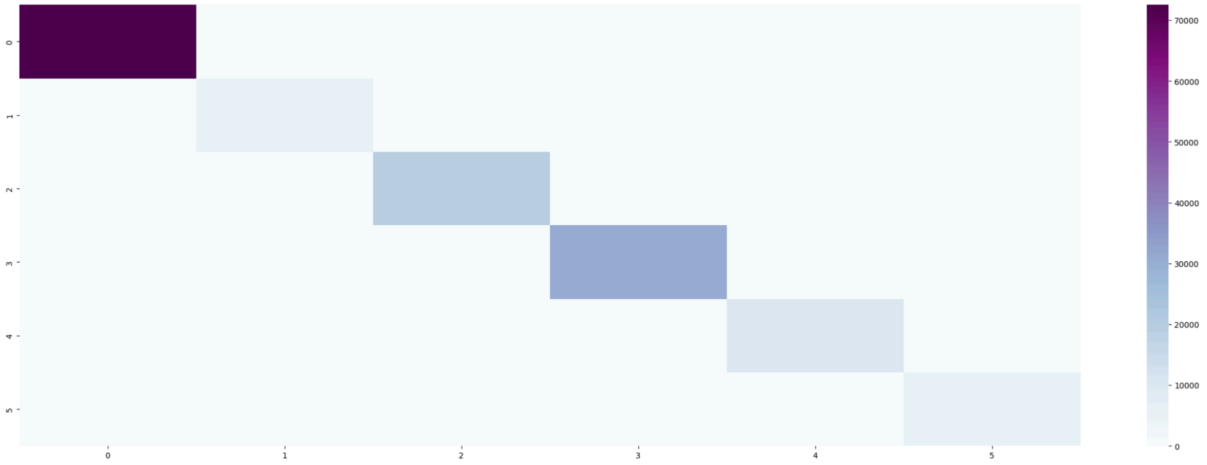
2. Decision Tree Classifier

```
array([[72531,     0,     9,     1,     0,     0],
       [    0,  6087,     0,     0,     0,     1],
       [   12,     0, 18699,     0,     0,     0],
       [    0,     0,     3, 30846,     0,     0],
       [    0,     0,     0,     0, 10359,     0],
       [    0,     0,     0,     0,     1,  5246]])
```



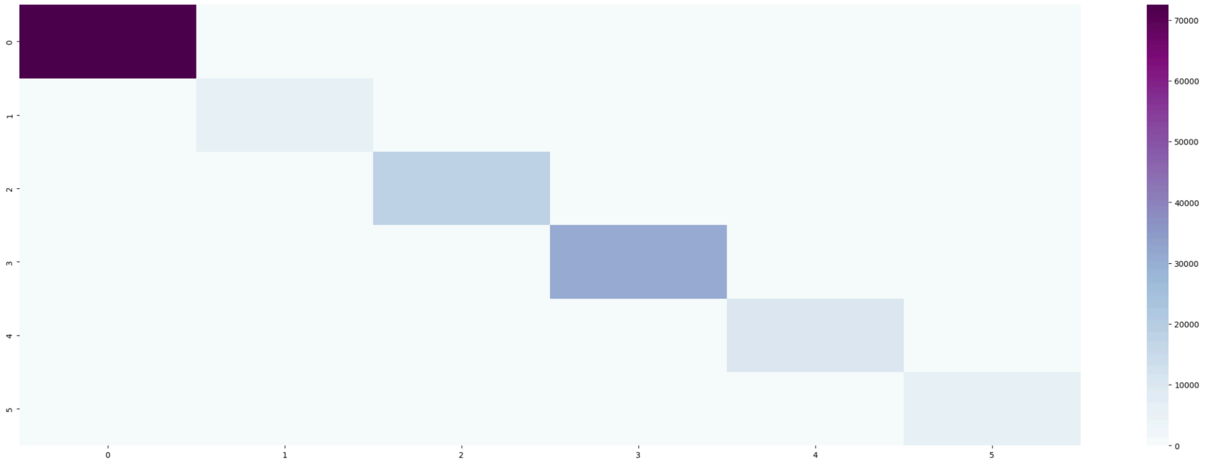
3. Random Forest Classifier

```
array([[72531,     0,     9,     1,     0,     0],
       [    0,  6086,     0,     0,     0,     2],
       [    6,     0, 18705,     0,     0,     0],
       [    0,     0,     1, 30848,     0,     0],
       [    0,     0,     0,     3, 10356,     0],
       [    0,     0,     0,     0,     1,  5246]])
```



4.K-Nearest Neigbours

```
array([[72494,     0,    46,     1,     0,     0],
       [    0,  6076,     0,     0,     0,    12],
       [   60,     0, 18537,   114,     0,     0],
       [    0,     0,    43, 30783,    23,     0],
       [    0,     0,     0,    52, 10283,    24],
       [    0,    69,     0,     0,    29,  5149]])
```



5. Conclusion

The present work purposes that machine learning techniques particularly Random Forest can efficiently predict AQI values and perform better than the conventional methods. The research shows the need to resort to ensemble techniques to enhance the predictions of the model. On the other hand, the drawbacks such as tree-based models are prone to overfitting imply that the next step would be to consider more advanced strategies such as hyperparameter optimization and hybrid models that could also use deep learning techniques.

6. References

1.Yuhui Jiao, Zhifeng Wang, Yang Zhang,"Prediction of Air Quality Index Based on LSTM",IEEE Joint International Information Technology and Artificial Intelligence Conference,2019
2.Marviola Hardini, Richard Andre Sunarjo, Marsani Asfi, Mochamad Heru Riza Chakim, Yulia Putri Ayu Sanjaya,"Predicting Air Quality Index using Ensemble Machine Learning",ADI Journal on Recent Innovation (AJRI),2023
3.K. Saikiran, G. Lithesh, Birru Srinivas, S. Ashok,"Prediction of Air Quality Index Using Supervised Machine Learning Algorithms",Access,2021
4.Deepa Patil, Ramesh K.,"Prediction of Air Quality Index through machine learning",IEEE North Karnataka Subsection Flagship International Conference (NKCon),2022