**Traffic Prediction Based on Air Quality Using Regression Model Analysis in IoT Based Smart City**

Manjaiah.D.H1, Praveena Kumari M.K2, Harishkumar K S3

1 Department of Computer Science Mangalore University Mangalore, India Email:drmdhmu@gmail.com

2 Department of Computer Science Mangalore University Mangalore, India

Email:narayan.praveena@gmail.com

3Department of Computer Science and Engineering Presidency University Bangalore, India Email:harishkumar@presidencyuniversity.in

**Abstract.** The forecasting of urban mobility and analysis of vehicle traffic patterns are vital elements of the "smart city" paradigm. Good traffic forecasting can aid in route planning and traffic jam alleviation. In addition to traffic-specific data such as speed and time, other aspects related to road traffic are air and noise pollution. Air contamination emissions are frequently linked with traffic quantity. In this paper, we present an air pollution-based traffic forecasting method. We contend that air contaminants data can improve traffic predictions. Our technique employs harmful airborne gases such as carbon monoxide, nitrogen dioxide, sulfur dioxide, particulate matter, and ozone. These gases were chosen due to their being related to transportation. The data used for the study is real-time traffic flow and air pollution data from Aarhus City of Denmark for August and September months in the year 2014. We have undertaken an evaluation of 9 regression models, K-Nearest Neighbor Regression, Support Vector Machine Regression, CART Regression, Random Forest Regression, Gradient Boosting Machine Regression, EXtreme Gradient Boosting Regression, Light Gradient Boosting Machine Regression, Catboost Regression, and Multilayer Layer Perceptron Regression to find out which model gives better accuracy. The performance of these regression models is tested using statistical measures such as Root Mean Squared Error, Mean Absolute Error, Mean Squared Error, and Coefficient of Determination. The best results were obtained with the Gradient Boosting Machine Regression model with 99.92% accuracy.

**Keywords:** Traffic, Regression, Smart City, ITS, Forecast, Air Quality.

1. Introduction

Automatic travel increased due to a combination of the global economy’s swift growth and the globe’s steadily growing human intensity. This has led to higher levels of traffic jams. Traffic has a major influence on many elements of everyday life, including the amount of time spent in traffic jams, the amount of polluted air generated, the amount of resources and petrol consumed, and the cost of building and maintaining transportation and road networks [1]. In certain large cities, traffic-related pollutants now account for a significant fraction of all air pollution and are the primary cause of it [2]. Of course, living in a metropolis has a lot to do with the quality of the air. Significant health issues brought on by air pollution include cancer of the lungs, pulmonary obstructive disease, cardiovascular disease, stroke, and infections of the lungs [3]. Families belonging to the general public and inner-city motorists have suffered from weakened psychological health and lowered standard of living as a result of the traffic jam [4]. Current research has shown an increased risk of illness and death for motorists, travelers, and those who live close to major highways as a result of traffic congestion, which also worsens the quality of the air and raises vehicle emissions [5]. Air quality and traffic congestion, both concepts are connected and numerous towns are tackling this problem by placing devices that gauge contamination in the atmosphere and traffic volume. Traffic fumes have been the primary contributor to air pollution in several areas. Particulate Matter(PM), Carbon Monoxide(CO), Ozone(O3), Nitrogen Dioxide(NO2) and Sulphur Dioxide(SO2) are the contaminants having the most compelling proof for safety for humans concern. Both short and prolonged exposure to these different contaminants can result in issues with health. There are no levels below which certain contaminants have negative consequences. The latest figures show that fine particulate matter or PM2.5, is to blame for almost four million mortality worldwide caused by cardiopulmonary conditions like persistent lung infections, preterm births and other disorders [6]. According to the survey of Online Master of Science in Civil Engineering, the typical resident of one of the seventy five major cities in U.S. experienced a seven-hour travel delay in the year 1982. By 2001, that number had increased to 26 hours of delays annually, and the most blocked times of the day, formerly known as “rush hour”, had expanded to approximately six hours per day. Additionally, trips during “rush hour” typically took nearly 40% longer than trips during other periods of the day[7]. The ability to choose the least congested and simplest route to their location or to modify the length of the journey to account for the projected time of arrival resulting from traffic will assist motorists to avoid jams. Multiple investigations have demonstrated how road congestion data can be utilized to forecast the pollutants in the air.

### Kumar K et al [8], in their study, have examined six years' worth of air pollution data from twenty-three Indian cities to analyze and forecast the health of the air. Different deep learning models were put up by Bekkar A et al [9] for the simulation of PM2.5 concentration, utilizing the air quality data containing the concentration of air contaminants and the overall conditions at twelve places, provided by the Beijing Municipal Environmental Monitoring Centre. They did not, however, consider traffic size in their studies. Travel is made simpler and more pleasant for travelers who are aware of the best path to take to get where they are going. Traffic administration is the key element of smart cities. One of the most significant amenities offered by the smart city platform is smart transportation. Traffic congestion increases pollution in the air, which harms stability in many municipalities. Motorists can avoid gridlocked roadways with the use of intelligent congestion plans, which lowers the level of toxins. It can be hard to foresee congestion transmission with any degree of accuracy because of the constantly changing irregular behavior of road infrastructures. Cognitive transportation is the key element in urban environments and is a vital topic in this field. Our study indicates that pollutants in the atmosphere have a significant impact on traffic predictions. Forecasting traffic will be more accurate when the degree of pollution is taken into account. Road traffic has consistently been used to predict air quality in previous investigations. The proportions of pollutants generated can be used to measure or extrapolate the number of automobiles on the road. The aim of our study is to build a model to forecast traffic considering air pollution data to give accurate results. If the outcome of our implementation is satisfactory then the maintenance costs can be lowered by reducing the number of traffic sensors. This means that instead of using traffic sensors to predict the traffic flow the model can do that using air quality. We have performed a comparative analysis of 9 different regression models, “K-Nearest Neighbor Regression, Support Vector Machine Regression, CART Regression, Random Forest Regression, Gradient Boosting Machine Regression, EXtreme Gradient Boosting Regression, Light Gradient Boosting Machine Regression, CatBoost Regression, and Multilayer Layer Perceptron Regression” to find out which model gives better accuracy. The methodology of the proposed technique is mentioned in Section 3. The performance of these regression models is tested using statistical measures such as Root Mean Square Error, Mean Square Error, Mean Absolute Error, and Coefficient of Determination for estimating traffic intensity.

The remaining portion of the paper is structured as follows. An analysis of related research is presented in Section 2. The architecture, methodology and regression models are described in Section 4. Section 5 has a discussion of the findings and results. The conclusion and future work are found in Section 6.

### 

1. Literature Review

### Almeida et al. [10] have investigated both statistical and deep learning methods for comprehending and forecasting the city transport pattern. Their study and experiment in this regard using statistical algorithms such as SARIMA, and neural network algorithms such as Feed Forward Neural Networks, Long Short-Term Memory, Convolution Neural Networks, and Hybrid Long Short-Term Memory-Convolution Neural Networks have shown that statistical models are significantly better than neural network algorithms at predicting traffic counters data in the short-term, even when unusual traffic situations are noticed. Convolution Neural Networks have been proven to be accurate and stable for forecasts over the long term.

### Menguc K et al [11] gave a viable, economical methodology to aid decision-makers by estimating changes in traffic flow caused by the construction of additional roads to existing systems. The study employs the Extreme Gradient Boosting (XGboost) method, a tree-based technique that delivers 85% accuracy in estimating traffic flow patterns in Istanbul. The model suggested by them is a static model that enables town administrators to conduct comprehensive studies between projects involving modifications to the town's transportation system.

### The HighD, a high-quality trajectory dataset is used by Yuan C et al. [12] to investigate the connection between clashes and transportation characteristics while accounting for heterogeneity while building forecasting algorithms for recognizing conflict-prone situations in real-time. Their comparison study included XGB (Boosting), Random Forest (Bagging), SVM (Single-classifier), and MLP (Deep neural network). The results show that (1) traffic pattern features have an important effect on the probability of conflict recurrence; and (2) XG Boosing trained on an under-sampled dataset is the most effective model, with an AUC of 0.871 and accuracy of 0.867.

### Time series traffic flow data are utilized by Lu J et al. [13] to propose a VMD-LSTM framework based on the VMD method. According to their findings, VMD can break the original unsteady series into more stable modal components, and the LSTM model may remove long-term information reliance from past data. The empirical comparison done by them shows that, in relation to other standard models, the suggested model for prediction performs superior in several metrics, and the correctness and consistency of the forecasting outcomes are better.

### Ahmed et al.[14] provided a full understanding of the impact of road accident injuries and the elements that contribute to them. The New Zealand road accident dataset is utilized for research purposes. Because there is no possibility of overfitting, the precision of the RF model outperforms Decision Jungle, AdaBoost, XGBoost, LGBM and CATBoost by 3% to 15%. In this study, the Shapley value is also employed as an understandable ML method to analyze the accuracy of models. They succeeded to identify the link between the relevant factors on overall model efficacy and individual data points by performing global and local SHAP analysis.

### Khajavi H et al [15] used a combination of RF, SV regression, and response surface approach to forecasting CO2 emissions in 30 important Chinese cities. In addition, seven optimizers are used to tune the Random Forest, and two optimizers are used to adjust the Support Vector Regression methods' hyper-parameters. The exactness of the given approaches is contrasted using the statistical indexes Standard Error (SE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Relative Absolute Error (RAE), and coefficient of determination (R2). The collected findings show that the Support Vector Regression with Harris Hawk optimizer has the best training accuracy with an R2 value of 0.9999, followed by the Random Forest with the Slime Mould Algorithm with an R2 value of 0.9641.

### Aljuaydi F et al [16] predicted motorway traffic under uncommon events, multivariate prediction models based on the MLP, One-dimensional Neural Network (1-D CNN), the Long Short-term Memory network, 1D-CNN LSTM, and autoencoder LSTM networks were created using a source of information with many instances and five attributes. The proposed multivariate forecasting approaches capture traffic patterns under uncommon occurrences. The 1D-CNN LSTM forecasting model gives more precise estimates.

### The data utilized by EIGhanam et al [17] in their work was gathered from the TomTom Move O/D Analysis portal for the towns of Dubai and Sharjah in the United Arab Emirates. To build the EV prediction of demand, multiple machine learning (ML) algorithms are trained on the dataset, including Random Forest(RF), Extreme Gradient Boosting (XGBoost), Multilayer Perceptron and Linear Regression Models. The MLP outperforms every other model tested, with a symmetric mean absolute percentage error of 20% on both the training and testing data subsets and a much shorter training time than RF and XGBoost.

### Ramachandra N R et al [18] used DAM (Deep Autoencoder), DBN (Deep Belief Network), RF (Random Forest), and LSTM (Long Short Term Memory) four machine learning techniques in their suggested model. The accuracy, precision, recall, and error value metrics of machine learning algorithms are used to assess the success of the proposed approach. LSTN achieved 95.2% performance among the four techniques.

### Zeinalnezhad M et al [19] implemented Nonlinear Multivariate Logistic Regression and Adaptive Neuro-Fuzzy Inference System models which were designed and tested to achieve the lowest feasible loss in forecasting contaminants CO, SO2, O3 and NO2. The study data was gathered from an isolated surveillance station in Tehran. In their study, the comparison of the accuracy of both models resulted in a good performance of the Adaptive Neuro-Fuzzy Inference System than regression techniques when estimating time-series data.

### Tang et al [20] suggested a hybrid model for roadway movement prediction that comprises noise mitigation methods and support vector machines. They simply employed 3 characteristics in their experiment: volume, speed, and occupancy. But because they did not include air quality data, the rate of errors remains substantial.

### In their research on air pollution forecasting, Le V D et al [21] used datasets on traffic density and mean driving pace. In their research, they suggested using the Convolutional Long Short-Term Memory which combines Long Short-Term Memory with Convolutional Neural Networks, to simulate air quality for the whole town at once.

### Zhap J et al [22], in their study, used Geographically Weighted Regression models to connect CO, NO2, and PM10 concentrations at Traffic Analysis Zone with a variety of influencing variables, such as congestion, the road system, social demographics, and industrial factors.

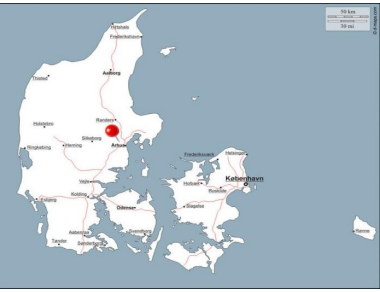
1. Study Area and Methodology
   1. *Data set Description*

### Our study is based on real-world traffic statistics from Aarhus, Denmark. We are using the Vehicle Traffic Dataset and Pollution Dataset collected from the city of Aarhus, Denmark. It is a large-scale publicly available IoT data. This IoT data contains traffic, air pollution, weather, cultural event, social event, library event, and parking datasets. For the study, we have used two datasets: Traffic and Air pollution data. The city administration has installed 449 sensor pairs along the city's key thoroughfares. The sensors capture the number of vehicles passing every five minutes. The air pollution dataset includes parameters for pollutants like carbon monoxide, nitrogen dioxide, sulfur dioxide, particulate matter, and ozone that are discharged into the atmosphere by moving cars.

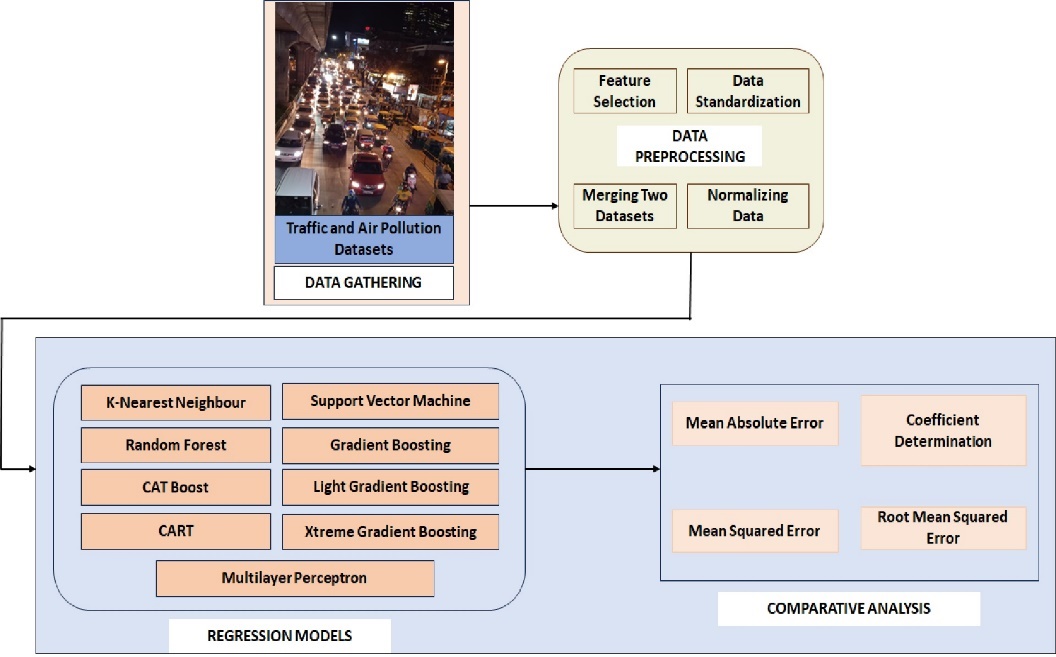
* 1. Models

For forecasting the traffic condition, we have utilized machine learning strategies, such as K-Nearest Neighbor Regression, Support Vector Machine Regression, CART Regression, Random Forest Regression, Gradient Boosting Machine Regression, Xtreme Gradient Boosting Regression, Light Gradient Boosting Machine Regression, CatBoost Regression, and Multilayer Layer Perceptron Regression. Air Pollution and Traffic datasets contain information on the same locations at the same timestamp. We have joined the two data sets based on the timestamp. After joining each row contains traffic data and pollution data for the same location at the same timestamp. For the analysis, we have retained only vehicle intensity from the traffic data and have used all the features from pollution data. Different techniques were applied for the model optimization; in this regard, feature engineering, feature Transformation, standardization, and hyperparameter tuning were used to improve the model performance.

Figure 1 shows the data source collection location. Figure 2 is the proposed architecture of the prediction models for forecasting the condition of traffic.



**Figure 1** Data source location



**Figure 2** Proposed architecture of the prediction models for forecasting traffic using air quality data

### **K-Nearest Neighbor Regression**

KNN regression is a nonparametric method that matches the relationship amongst independent variables and continuous outcomes by aggregating data in the same area. The size of the neighborhood must be determined by the researcher or can be determined via cross-validation to fix the size that reduces the mean squared error. While the strategy appears to be enticing it rapidly becomes unworkable as the dimension grows, i.e. when there are a lot of variables that are independent.

### Support Vector Machine Regression

### SVM regression, often known as Support Vector Regression (SVR), is a regression-based machine learning algorithm. It differs from typical linear regression techniques in that rather than fitting a straight line to the data points, it finds something called a hyperplane that best suits the data points in a space that is continuous. It seeks the function that best forecasts the continuous output value given a specific input value. SVR can use both nonlinear as well as linear kernels. A linear kernel is a basic dot product of 2 input vectors, whereas a non-linear kernel is a more complicated function capable of capturing complex data patterns. The kernel used is determined by the qualities of the data and the degree of difficulty of the task at hand.

* + 1. **CART Regression**

CART is a Machine Learning prediction technique that shows how the values of the variable being studied can be anticipated by considering other factors. It is a decision tree in which every branch is divided into variables that are predictors and each node contains an end prediction for the outcome variable. Nodes in the decision tree are divided into sub-nodes based on an attribute’s threshold value. The root node is used as the basis for training, and it is divided into two parts based on the most effective attribute and the threshold value. Furthermore, the subsets are divided utilizing the same approach. This process is repeated until the tree has the last pure subset or has the most number of leaves conceivable in that expanding tree.

* + 1. **Random Forest Regression**

Random Forest is an ensemble technique that can handle both regression and classification problems by combining many decision trees and an approach called Bootstrap and Aggregation or Bagging. The core idea is to use numerous decision trees to determine the result instead of depending on specific decision trees.

* + 1. **Gradient Boosting Machine Regression**

Gradient Boosting is a robust boosting technique that combines numerous poor learners into powerful learners by training every fresh model to reduce the loss function of the preceding model, such as mean squared error or cross-entropy, using gradient descent. The approach determines the gradient of the loss function with regard to the present ensemble’s estimates in every round and then trains a new poor model to minimize this gradient. The new model’s projections are subsequently added to the ensemble, and the procedure continues until an endpoint is reached.

* + 1. **EXtreme Gradient Boosting Regression**

Extreme Gradient Boosting is a supervised machine learning approach that uses decision trees as its base estimators. Gradient Boosting methods create powerful models for forecasting through the combination of poor models. The model’s decision trees are constructed progressively to allow the following trees to minimize the shortcomings of earlier trees.

* + 1. **Light Gradient Boosting Machine Regression**

Light Gradient Boosting Regression is a gradient-boosting approach based on decision trees that enhances model performance while using a smaller amount of memory. To solve the limitations of the histogram-based strategy, which is widely used in all Light Gradient Boosting Regression systems, it employs two novel techniques: Gradient-based One Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). They work together to make the model perform effectively and offer it a distinct benefit over approaches that are equivalent to Light Gradient Boosting Regression.

* + 1. **Catboost Regression**

CatBoost can incorporate a variety of data sources, including continuous and discrete values. This approach is effective at predicting medium to long-term load. CatBoost was created with the theoretical idea of Gradient Boosting and decision trees in mind. Boosting works by merging numerous poor models and running them through an aggressive search algorithm to increase the accuracy of a prediction model. CatBoost can manage sparse, diverse, and categorical data.

* + 1. **Multilayer Layer Perceptron Regression**

MLP consists of densely connected layers that translate any input dimension to the required dimension. A neural network with numerous layers is referred to as a multi-layer perception. To build a neural network, we connect neurons so that their outputs become the inputs of other neurons. Multi-layer perceptions are neural networks that can be used to solve binary/multiple-class regression and classification issues.

1. Performance Criteria

Some of the statistical evaluations are used to evaluate the model performance such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Square Error (MSE), and coefficient of determination (*R*2). The criteria formulas are shown in below:

Some of the statistical evaluations are used to evaluate the model performance such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Square Error (MSE), and coefficient of determination (*R*2). The criteria formulas are shown in below:

Some of the statistical evaluations are used to evaluate the model performance such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Square Error (MSE), and coefficient of determination (*R*2). The criteria formulas are shown in below:

Mean Square Error (MSE), Mean Absolute Error(MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R2) are some of the statistical evaluations used to evaluate model performance. The formulas for the criterion are listed below.

(1)

(2)

in (1) and (2) w is the number of observations, is the predicted value and is the actual value.

(3)

where, N shows the number of observations, shows the standard deviation of the observation X, shows the standard deviation of Y, shows the observed values, is the mean of the observed values shows the calculated values, and is the mean of the calculated values.

(4)

where n is the number of observations, is the predicted value and is the actual value.

1. Results and Discussion

Dataset is divided into train and test datasets in 8:2 ratio. Hyperparameters of all the algorithms are tuned to get high accuracy. We performed the experiment on Jupyter Notebook. For viewing the results matplot and seabornlibrariess are used. Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Square Error (RAE), R-Square (R2) are used for assessing the performance of the models.

The goal of this comparison analysis is to find and choose the most effective traffic forecasting approach among a set of regression approaches. In this section, the assessment of each model and its total performance is shown, demonstrating its usefulness in traffic prediction. Table 1 summarizes all test metrics for the nine regression models. Fig. 4(a) is shows the Mean Absolute Error values for the nine regression techniques.MAE computes the absolute average difference between the observed and anticipated data and Gradient Boosting is giving the lowest value i.e. 0.112727. Fig. 4(b) and Fig 4(c) shows the Mean Square Error and Root Mean Square Error values respectively for the nine regression techniques. The RMSE is a typical method for calculating a model’s error in anticipating quantitative findings. The lower the RMSE, the better the fit to the data. In our study Gradient Boosting regression is giving the lowest values i.e. 0.020468 and 0.143065 respectively.

4(d) is showing the R-Squared value of nine regression techniques. R-square is a rating metric that indicates the regression model’s performance. The ideal R-square value is one. The closer the R-square factor to one, the more accurately the model fits.In our study Gradient Boosting regression is giving the best value i.e.0.999242. Fig. 3 summarizes the above-discussed results. Gradient Boosting Regression is better in every aspect when compared with other regression approaches.

**Table 1** Test metrics for the nine regression models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **MAE** | **MSE** | **RMSE** | **R2 Square** |
| ***K-Nearest Neighbor*** | 1.120071 | 3.149554 | 1.774811 | 0.883301 |
| ***Support Vector Machine*** | 2.146205 | 12.362537 | 3.516040 | 0.541996 |
| ***CART*** | 1.542576 | 5.948167 | 2.438886 | 0.779634 |
| ***Random Forest*** | 0.721655 | 1.137878 | 1.066714 | 0.957844 |
| ***Gradient Boosting Machine*** | **0.112727** | **0.020468** | **0.143065** | **0.999242** |
| ***Extreme Gradient Boosting*** | 0.601935 | 0.756439 | 0.869735 | 0.971976 |
| ***Light Gradient Boosting Machine*** | 1.109146 | 2.477313 | 1.573948 | 0.908221 |
| ***CatBoost*** | 1.173650 | 2.694791 | 1.641582 | 0.900164 |
| ***Multi Layer Perceptron*** | 1.889581 | 7.683268 | 2.771871 | 0.715352 |

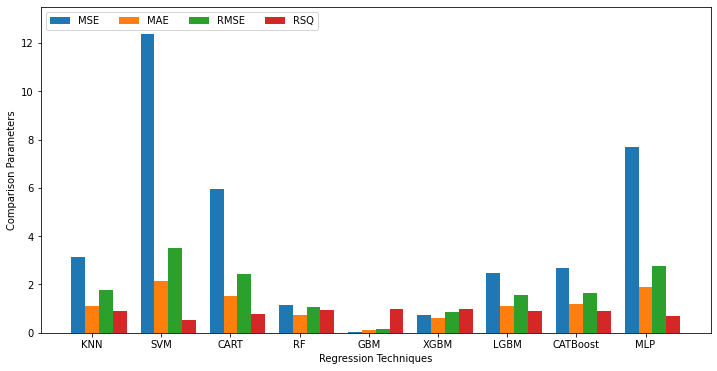


Figure 3 Summary of comparisons between nine regression techniques

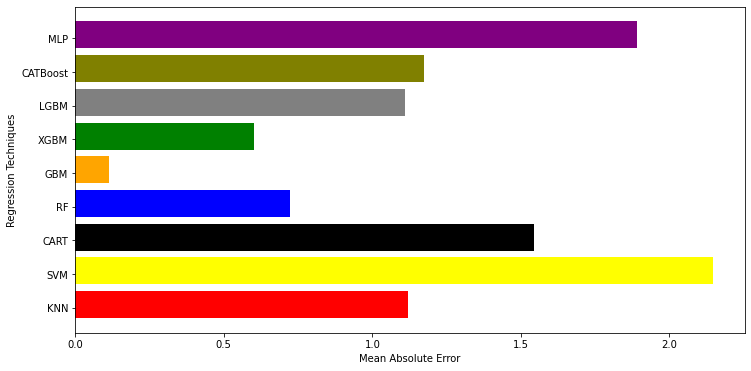


Figure 4(a): MAE for different regression techniques

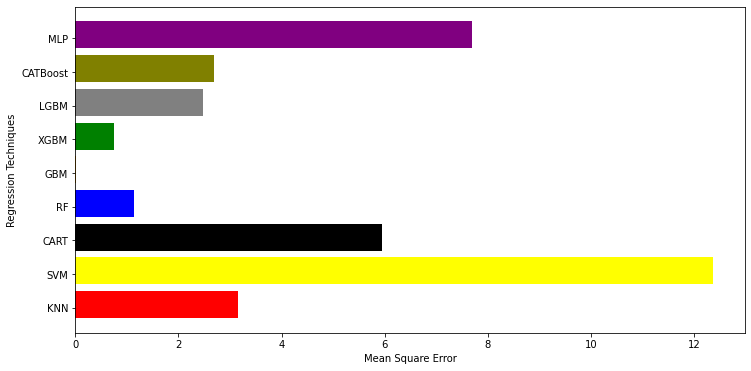


Figure 4(b): MSE for nine regression techniques

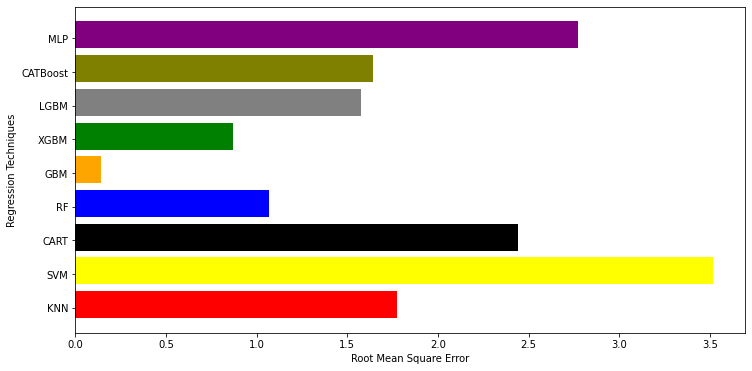


Figure 4(c): RMSE for nine regression techniques

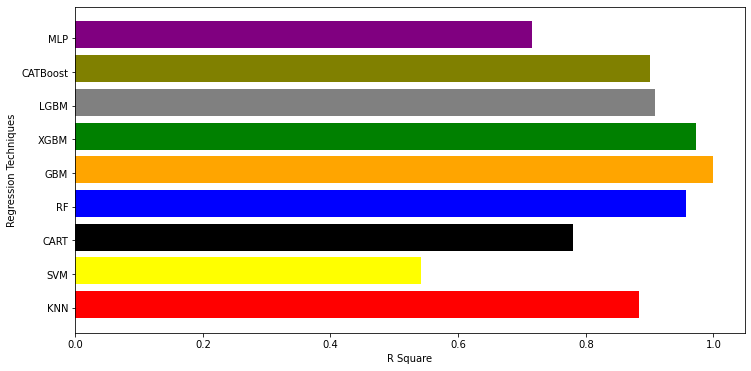


Figure 4(d): R-Squared for nine regression techniques

1. Conclusion and Future Work

In this work, we compared nine regression models K-Nearest Neighbor Regression, Support Vector Machine Regression, CART Regression, Random Forest Regression, Gradient Boosting Machine Regression, Xtreme Gradient Boosting Regression, Light Gradient Boosting Machine Regression, CatBoost Regression, and Multilayer Layer Perceptron Regression to find out which model gives better accuracy. The performance of these regression models is tested using statistical measures such as Root Mean Squared Error, Mean Absolute Error, Mean Squared Error, and Coefficient of Determination. The best results were obtained with the Gradient Boosting Machine Regression model with 99.92% accuracy followed by EXtreme Gradient Boosting Regression model with 97.20% accuracy. The findings from the experiment support the general success rate of the comprehensive strategy that we presented. We intend to implement Ensemble methods (stacking, bagging, and boosting) as well as neural networks to predict traffic considering more instances than used in this study.

References

1. Kuang, Y., Yen, B. T., Suprun, E., & Sahin, O. (2019). A soft traffic management approach for achieving environmentally sustainable and economically viable outcomes: An Australian case study. Journal of environmental management, 237, 379-386.
2. Lu, J., Li, B., Li, H., & Al-Barakani, A. (2021). Expansion of city scale, traffic modes, traffic congestion, and air pollution. *Cities*, *108*, 102974.
3. Borck, R., & Schrauth, P. (2021). Population density and urban air quality. *Regional Science and Urban Economics*, *86*, 103596.
4. Iamtrakul, P., & Chayphong, S. (2021). The perception of Pathumthani residents toward its environmental quality, suburban area of Thailand. *Geographica Pannonica*, *25*(2), 136-148.
5. Zhang, K., & Batterman, S. (2013). Air pollution and health risks due to vehicle traffic. *Science of the total Environment*, *450*, 307-316.
6. Thangavel, P., Park, D., & Lee, Y. C. (2022). Recent insights into particulate matter (PM2. 5)-mediated toxicity in humans: an overview. *International journal of environmental research and public health*, *19*(12), 7511.
7. Available: https://onlinemasters.ohio.edu/blog/traffic-congestion-problems-and-solutions/ dated 26 July 2022.
8. Kumar, K., & Pande, B. P. (2023). Air pollution prediction with machine learning: a case study of Indian cities. International Journal of Environmental Science and Technology, 20(5), 5333-5348.
9. Bekkar, A., Hssina, B., Douzi, S., & Douzi, K. (2021). Air-pollution prediction in smart city, deep learning approach. Journal of big Data, 8(1), 1-21.
10. Almeida, A., Brás, S., Oliveira, I., & Sargento, S. (2022). Vehicular traffic flow prediction using deployed traffic counters in a city. *Future Generation Computer Systems*, *128*, 429-442.
11. Menguc, K., Aydin, N., & Yilmaz, A. (2023). A Data-Driven Approach to Forecasting Traffic Speed Classes Using Extreme Gradient Boosting Algorithm and Graph Theory. *Physica A: Statistical Mechanics and its Applications*, *620*, 128738.
12. Yuan, C., Li, Y., Huang, H., Wang, S., Sun, Z., & Li, Y. (2022). Using traffic flow characteristics to predict real-time conflict risk: A novel method for trajectory data analysis. *Analytic methods in accident research*, *35*, 100217.
13. Lu, J. (2023). An efficient and intelligent traffic flow prediction method based on LSTM and variational modal decomposition. *Measurement: Sensors*, 100843.
14. Ahmed, S., Hossain, M. A., Ray, S. K., Bhuiyan, M. M. I., & Sabuj, S. R. (2023). A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance. *Transportation research interdisciplinary perspectives*, *19*, 100814.
15. Khajavi, H., & Rastgoo, A. (2023). Predicting the carbon dioxide emission caused by road transport using a Random Forest (RF) model combined by Meta-Heuristic Algorithms. *Sustainable Cities and Society*, *93*, 104503.
16. Aljuaydi, F., Wiwatanapataphee, B., & Wu, Y. H. (2023). Multivariate machine learning-based prediction models of freeway traffic flow under non-recurrent events. *Alexandria engineering journal*, *65*, 151-162.
17. ElGhanam, E., Hassan, M., & Osman, A. (2022, December). Machine Learning-Based Electric Vehicle Charging Demand Prediction Using Origin-Destination Data: A UAE Case Study. In *2022 5th International Conference on Communications, Signal Processing, and their Applications (ICCSPA)* (pp. 1-6). IEEE.
18. Ramchandra, N. R., & Rajabhushanam, C. (2022). Machine learning algorithms performance evaluation in traffic flow prediction. *Materials Today: Proceedings*, *51*, 1046-1050.
19. Zeinalnezhad, M., Chofreh, A. G., Goni, F. A., & Klemeš, J. J. (2020). Air pollution prediction using semi-experimental regression model and Adaptive Neuro-Fuzzy Inference System. *Journal of Cleaner Production*, *261*, 121218.
20. Tang, J., Chen, X., Hu, Z., Zong, F., Han, C., & Li, L. (2019). Traffic flow prediction based on combination of support vector machine and data denoising schemes. *Physica A: Statistical Mechanics and its Applications*, *534*, 120642.
21. Le, V. D., Bui, T. C., & Cha, S. K. (2020, February). Spatiotemporal deep learning model for citywide air pollution interpolation and prediction. In 2020 IEEE international conference on big data and smart computing (BigComp) (pp. 55-62). IEEE.
22. Xu, C., Zhao, J., & Liu, P. (2019). A geographically weighted regression approach to investigate the effects of traffic conditions and road characteristics on air pollutant emissions. Journal of Cleaner Production, 239, 118084.