TRAFFIC PREDICTION BASED ON AIR QUALITY USING REGRESSION MODEL ANALYSIS IN IOT BASED SMART CITY

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 approach utilizes noxious airborne gases, including CO, NO2, SO2, PM, and O3.

Thse gases were chosen due to their bein g related to transportation. The study utilized real-time traffic flow and air pollution information obtained from Aarhus City in Denmark during the months of August and September 2014.

We have undertaken an evaluation of 9 r egression models, K-Nearest Neighbor, Support Vector Machine, CART, Random Forest, Gradient Boosting, Extreme Gradient Boosting, Light Gradient Boosting, Catboost, and Multilayer Layer Perceptron to find out which model gives better accuracy. We assessed the effectiveness of these regression models through statistical metrics, including Root Mean Squared Error, Mean Absolute Error, Mean Squared Error, and Coefficient of Determination. The Gradient Boosting Machine Regression model delivered the most favorable outcomes, achieving an accuracy rate of 99.92%.

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

Authors

Manjaiah. D.H
Department of Computer Science
Mangalore University
Mangalore, India
drmdhmu@gmail.com

Praveena Kumari M.K
Department of Computer Science
Mangalore University
Mangalore, India
narayan.praveena@gmail.com

Harishkumar K S
Department of Computer Science and Engineering
Presidency University
Bangalore, India
harishkumar@presidencyuniversity.in

I. 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 significantly impacts various aspects of daily life, including the duration spent in traffic congestion, the generation of air pollution, the consumption of resources and gasoline, and the expenses associated with constructing and upkeeping transportation and road systems [1]. In some major cities, pollutants stemming from traffic constitute a substantial portion of overall air pollution and serve as its primary source [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]. Recent studies have indicated a heightened health and mortality risk for drivers, commuters, and individuals residing in proximity to major highways due to traffic congestion, which additionally deteriorates air quality and escalates 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]. Based on the Online Master of Science in Civil Engineering survey, the average inhabitant of one of the seventy-five major U.S. cities faced a travel delay of seven hours in 1982. By 2001, this figure had surged to 26 hours of yearly delays, and the previously defined "rush hour" had extended to about six hours each day. Furthermore, journeys during these congested times, once known as "rush hour," typically took nearly 40% more time than trips made during other times 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 exam-ined six years' worth of air pollution data from twenty-three Indian cities to analyze and fore-cast 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 con-centration 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 conges-tion 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. Our study's objective is to construct a predictive model for traffic by incorporating air pollution data to enhance accuracy. If our implementation yields satisfactory results, it could potentially reduce maintenance costs by decreasing the reliance on traffic sensors. In other words, the model could predict traffic flow using air quality data instead of traditional traffic sensors. We conducted a comparative analysis involving nine distinct regression models as part of our research: “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. We evaluate the effectiveness of these regression models by employing statistical metrics like Root Mean Square Error, Mean Square Error, Mean Absolute Error, and Coefficient of Determination to estimate traffic intensity.

The remainder of the paper is organized as follows: Section 2 provides an examination of relevant prior research. Section 4 outlines the architecture, methodology, and regression models. Section 5 offers a discussion of the findings and results. Finally, Section 6 encompasses the conclusion and outlines future research directions.

II. 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. In the research, the Extreme Gradient Boosting (XGboost) approach is utilized, which is a tree-based method, achieving an 85% accuracy rate in predicting traffic flow patterns in Istanbul. Their proposed model is a static one, enabling city officials to perform in-depth analyses when considering projects related to changes in the city's transportation system.

Yuan C et al. [12] employed the HighD dataset, which is a high-quality trajectory dataset, to explore the relationship between traffic incidents and various transportation characteristics. They aimed to develop predictive algorithms capable of identifying conflict-prone situations in real-time while accounting for heterogeneity. Their comparative analysis involved the use of different algorithms, including XG Boosting (Boosting), Random Forest (Bagging), SVM (Single-classifier), and MLP (Deep neural network). The findings indicate two key points: (1) traffic pattern features significantly influence the likelihood of recurring conflicts, and (2) XG Boosting, when trained on an under-sampled dataset, emerged as the most effective model, achieving an AUC of 0.871 and an accuracy of 0.867.

Lu J et al. [13] make use of time series traffic flow data to introduce a VMD-LSTM framework, which is built upon the VMD method. Their research reveals that VMD effectively decomposes the initial volatile series into more stable modal components, while the LSTM model helps eliminate the dependence on long-term information from past data. Their empirical comparison indicates that, when compared to other conventional models, the proposed prediction model outperforms them across various metrics, leading to more accurate and consistent forecasting results.

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 managed to establish a connection between the pertinent factors affecting both the overall model performance and individual data points by conducting global and local SHAP analyses.

Khajavi H et al. [15] employed a combination of Random Forest (RF), Support Vector Regression (SVR), and a response surface approach to predict CO2 emissions in 30 significant Chinese cities. They utilized seven optimizers to fine-tune the Random Forest model and two optimizers to adjust the hyperparameters of the Support Vector Regression methods. The precision of these approaches was compared using statistical metrics such as 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 results revealed that the Support Vector Regression with the Harris Hawk optimizer achieved the highest training accuracy, boasting an R2 value of 0.9999. Following closely was the Random Forest model with the Slime Mould Algorithm, which achieved an R2 value of 0.9641

Aljuaydi F et al. [16] forecasted motorway traffic during unusual events by developing multivariate prediction models. These models were based on various neural network architectures, including the MLP, One-dimensional Neural Network (1-D CNN), Long Short-term Memory network, 1D-CNN LSTM, and autoencoder LSTM networks. They employed a dataset with a substantial number of instances and five attributes as their data source. The suggested multivariate prediction methods excel at capturing traffic patterns during unusual events, with the 1D-CNN LSTM forecasting model providing the most accurate predictions.

EIGhanam et al. [17] collected their data from the TomTom Move O/D Analysis portal, focusing on the cities of Dubai and Sharjah in the United Arab Emirates. To create a predictive model for electric vehicle (EV) demand, they trained various machine learning (ML) algorithms on this dataset. These algorithms included Random Forest (RF), Extreme Gradient Boosting (XGBoost), Multilayer Perceptron (MLP), and Linear Regression Models. Among these models, the Multilayer Perceptron (MLP) outperformed all others. It achieved a symmetric mean absolute percentage error of 20% on both the training and testing data subsets and required considerably less training time compared to RF and XGBoost.

Ramachandra N R et al. [18] employed four machine learning techniques, namely DAM (Deep Autoencoder), DBN (Deep Belief Network), RF (Random Forest), and LSTM (Long Short Term Memory), in their proposed model. They evaluated the effectiveness of their approach using metrics such as accuracy, precision, recall, and error values for these machine learning algorithms. Among the four techniques, LSTM achieved the highest performance at 95.2%.

Zeinalnezhad M et al. [19] applied Nonlinear Multivariate Logistic Regression and the Adaptive Neuro-Fuzzy Inference System models. These models were specifically designed and tested with the goal of minimizing prediction errors when forecasting contaminants such as 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, velocity, and occupancy. However 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 Regresion 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.

III. AREA OF RESEARCH AND RESEARCH APPROACH

1. **Description of the Dataset:** Our research relies on authentic traffic data sourced fromAarhus, 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 events, social events, library events, and parking datasets. For the study, we have used two datasets: Traffic and Air pollution data. The municipal authorities have positioned 449 pairs of sensors along the primary routes within the city. The sensors record the count of vehicles passing by every five minutes. Meanwhile, the air pollution dataset contains measurements for pollutants such as carbon monoxide, nitrogen dioxide, sulfur dioxide, particulate matter, and ozone emitted into the air by vehicles in motion.

2. **Models:** To predict traffic conditions, we have employed regression techniques, such as KNN, SVM, CART, RF, GB, XGB, LGB, CatBoost, and Multilayer Layer Perceptron. 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: Illustrates the location where data was collected, while Figure 2 depicts the suggested architecture for the traffic condition prediction models.

Figure 1: Location of data source

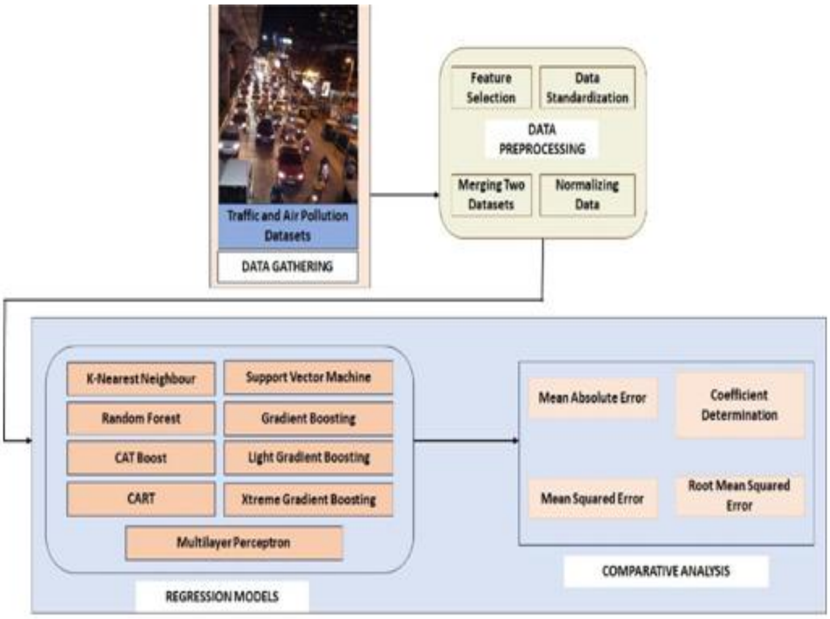

Figure 2: Proposed Design of Traffic Forecasting Models Utilizing Air Quality Data

- **K-Nearest Neighbor Regression:** KNN regression is a non-parametric technique that establishes a connection between independent variables and continuous outcomes by grouping data points in the same vicinity. The researcher needs to specify the neighborhood's size or determine it through cross-validation to find the size that minimizes the mean squared error. However, this approach becomes impractical as the dimensionality increases, particularly when there are numerous independent variables.

- **Support Vector Machine Regression:** SVM regression, also recognized as Support Vector Regression (SVR), is a machine learning algorithm primarily used for regression tasks. 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 choice of kernel is based on the characteristics of the data and the complexity of the task.

- **CART Regression:** CART, which stands for Classification and Regression Trees, is a machine learning technique used for predicting how the values of a studied variable can be forecasted by considering various influencing factors. It functions as a decision tree, where each branch represents predictive variables, and each node contains a final prediction for the target variable. Nodes in the decision tree are split into sub-nodes based on a particular attribute's threshold value. The initial root node is divided into two parts using the most informative attribute and its corresponding threshold value. This process continues iteratively, creating subsets based on the same methodology, until the tree either reaches a fully homogeneous subset or attains the maximum possible number of leaves for that expanding tree. ng, 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.

- **Random Forest Regression:** Random Forest is an ensemble method capable of addressing both regression and classification tasks by amalgamating multiple decision trees through a technique known as Bootstrap and Aggregation, or Bagging. The fundamental concept is to utilize a multitude of decision trees to make predictions, rather than relying on individual decision trees.

- **Gradient Boosting Machine Regression:** Gradient Boosting is a robust boosting method that enhances the performance of multiple weak learners by training each new model to minimize the loss function of the previous model, which could be metrics like mean squared error or cross-entropy, using gradient descent. This approach calculates the gradient of the loss function concerning the current ensemble's predictions in each iteration and then trains a new weak model to reduce this gradient. The predictions of the new model are then incorporated into the ensemble, and this process iterates until a stopping point is reached.

- **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.

- **Light Gradient Boosting Machine Regression:** Light Gradient Boosting Regression is a gradient-boosting technique that relies on decision trees to improve model performance while conserving memory resources. To address the shortcomings of the widely adopted histogram-based strategy in all Light Gradient Boosting Regression systems, it incorporates two innovative techniques: Gradient-based One Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). These techniques collaborate to enhance the model's efficiency and provide it with a distinct advantage over comparable approaches in the Light Gradient Boosting Regression domain.

- **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.

- **Multilayer Layer Perceptron Regression:** An MLP (Multi-Layer Perceptron) comprises densely connected layers that transform input dimensions into the desired dimensions. A neural network with multiple layers is commonly referred to as a multilayer perceptron. In constructing a neural network, neurons are interconnected in a way that their outputs serve as inputs to other neurons. The multi-layer perceptron is a type of neural network that can effectively handle tasks involving binary or multiclass regression as well as classification problems.

IV. 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:

Several statistical assessments are employed to assess model performance. They are Mean Square Error (MSE), Mean Absolute Error(MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R2). The formulas for these criteria are provided below.

$$RMSE = \sqrt{\frac{\sum_{r=1}^{w}(y_r - \hat{y}_r)^2}{w}} \tag{1}$$

$$MSE = \frac{\sum_{r=1}^{w}(y_r - \hat{y}_r)^2}{w} \tag{2}$$

in (1) and (2) w is the number of observations, $\hat{y}_r$ is the predicted value and y_r is the actual value.

$$R^2 = \left[\frac{1}{N}\,\frac{\sum_{i=1}^{N}[(Y_i - \bar{Y})(X_i - \bar{X})]}{\sigma_y \sigma_x}\right]^2 \tag{3}$$

where, N represents the total number of observations, $\sigma_x$represents the standard deviation of the observation X, σ_y represents the standard deviation of Y, X_i represents the observed values, $\bar{X}$ represents the mean of the observed values Y_i represents the calculated values, and $\bar{Y}$ represents the mean of the calculated values.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{4}$$

where n represents the number of observations, y_i represents the predicted value and x_i represents the actual value.

V. FINDINGS AND CONVERSATION

The dataset is split into training and testing sets with an 80:20 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 seaborn libraries 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 aim of this comparative analysis is to identify and select the most effective traffic forecasting approach from a set of regression methods. In this section, we present an evaluation of each model along with their overall performance, showcasing their utility in predicting traffic. Table 1 provides a summary of all the test metrics for the nine regression models. Figure 4(a) displays the Mean Absolute Error (MAE) values for these nine regression techniques. MAE quantifies the average absolute difference between observed and predicted data, with Gradient Boosting yielding the lowest value at 0.112727. Figures 4(b) and 4(c) depict the Mean Square Error (MSE) and Root Mean Square Error (RMSE) values, respectively, for the nine regression techniques. RMSE is a common measure for assessing a model's accuracy in predicting quantitative outcomes, where a lower RMSE indicates a better fit to the data. In our study, Gradient Boosting regression achieves the lowest values for RMSE, specifically 0.020468 and 0.143065, respectively.

Figure 4(d) displays the R-Squared values for nine regression techniques. R-Squared is a performance metric that assesses how well a regression model fits the data. An ideal R-Squared value is one, and the closer the R-Squared value is to one, the better the model's fit. In our study, Gradient Boosting regression stands out with the highest R-Squared value, which is 0.999242. Figure 3 provides a concise overview of the results discussed above, highlighting that Gradient Boosting Regression outperforms other regression approaches in every aspect.

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

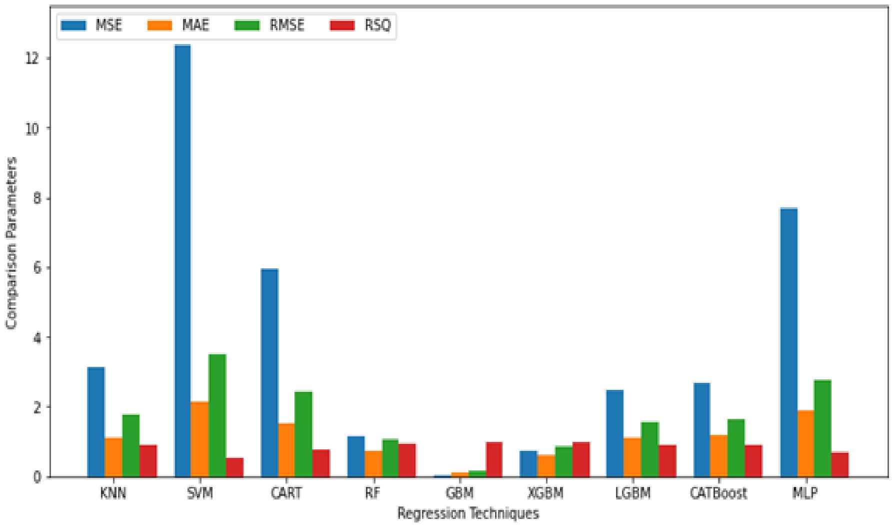

Figure 3: Summary of comparisons between nine regression techniques

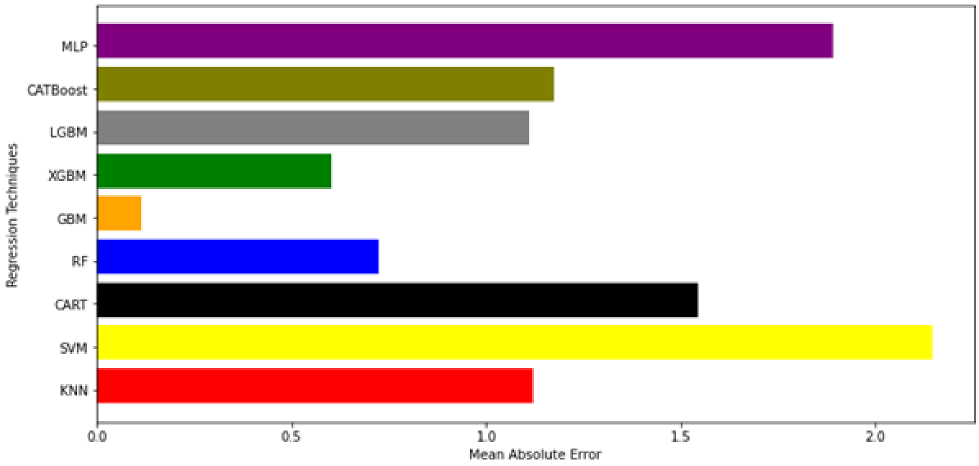

Figure 4(a): MAE for different regression techniques

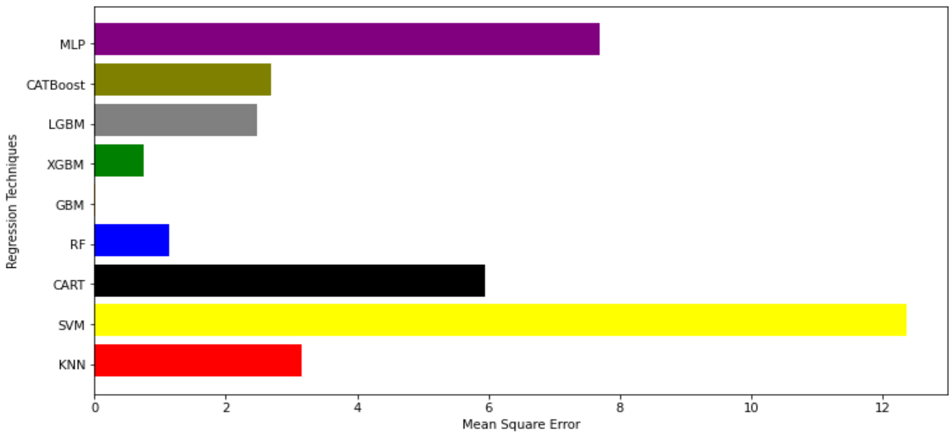

Figure 4(b): MSE for nine regression techniques

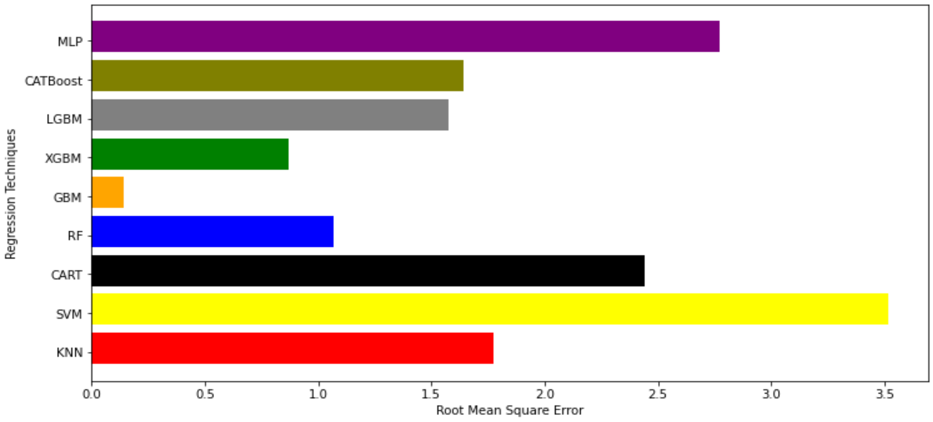

Figure 4(c): RMSE for nine regression techniques

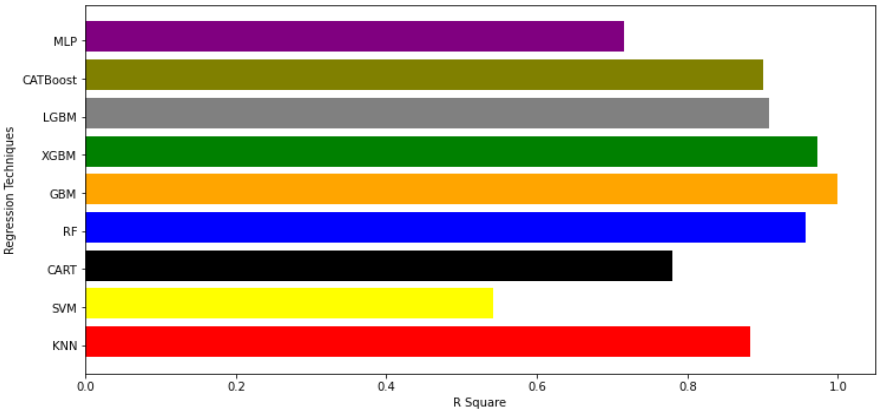

Figure 4(d): R-Squared for nine regression techniques

VI. SUMMARY AND FUTURE PROSPECTS

In this study, we conducted a comparison of nine regression models, including 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 Perceptron Regression. Our objective was to identify the model that offers the highest level of accuracy. To assess the performance of these regression models, we employed statistical metrics such as Root Mean Squared Error, Mean Absolute Error, Mean Squared Error, and Coefficient of Determination. The most favorable results were achieved by the Gradient Boosting Machine Regression model, which achieved an impressive accuracy of 99.92%, followed by the Xtreme Gradient Boosting Regression model with an accuracy of 97.20%. 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.