International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

TRAFFIC FLOW PREDICTION USING MACHINE LEARNING

K. SAI MANASWINI
Student
Computer Science and Engineering
Tirumala Engineering College, Narasaraopeta
Email: saimanaswinikancharla2002@gmail.com

K. DEEPTHI
Student
Computer Science and Engineering
Tirumala Engineering College, Narasaraopeta
Email: deepu.kattera25@gmail.com

P.V.SAI ESWAR REDDY
Student
Computer Science and Engineering
Tirumala Engineering College, Narasaraopeta
Email: eswarnani3655@gmail.com

M. SOMA SEKHAR
Student
Computer Science and Engineering
Tirumala Engineering College, Narasaraopeta
Email: somusekhar949@gmail.com

Dr. Lalu Naik. Ph.D
ASST PROFESSOR, DEPARTMENT OF CSE
Tirumala Engineering College, Narasaraopeta
Email: lalunaik12321@gmail.com

Abstract **– As a result of the ongoing increase in the number of vehicles, there are increasingly severe traffic bottlenecks. The amount of time and money that users of transportation spend traveling will immediately change as a result. To some extent, this issue can be mitigated by projecting future traffic patterns. Three machine learning algorithms—Linear Regression, Decision Tree, and Support Vector Machine—are used in this work to predict the traffic flow after the data has been preprocessed using Selenium, OSS, and Message Queue. Next, we examine real-time Beijing traffic data as it shows up on the Baidu map. According to this study, Random Forest is the most accurate of all three techniques, with an accuracy of 0.719. Second are logistic regression and SVM.**

Keywords - ***Traffic FlowPrediction; Decision Tree; Random Forest; SVM; Baysian Ridge; CNN; LSTM***

I. INTRODUCTION

Due to its potential to improve transportation management and lessen urban traffic congestion, the application of deep learning and machine learning to traffic prediction has attracted a lot of attention lately. To close the gaps and lay the groundwork for our project, which aims to close those gaps, we identify key points that each study article may have missed by taking insights from a large number of them.

Weather-based traffic flow prediction is one area of research for intelligent transportation systems. The effectiveness of standard prediction methodologies in estimating short-term traffic flow is limited due to the complexity of the affecting components. Lu Wang et al.'s short-term traffic flow prediction model, which makes use of variational modal decomposition and short- and long-term memory networks, has shown encouraging results in producing accurate forecasts [1].
Jingyi Lu and others Furthermore, it has been discovered that adding meteorological data to traffic flow forecasting models increases forecast accuracy. Research has demonstrated that the accuracy of power price estimates in day-ahead markets can be greatly enhanced by utilizing next-day weather forecasts, indicating that weather forecasts may also be useful in anticipating traffic patterns [2]. Mohammed Bashir et al. Furthermore, it has been discovered that the use of weather variables in soybean volatility forecasting models—such as clear sky index, cloud cover, relative humidity, atmospheric pressure, precipitation, temperature, and wind speed—outperforms models without weather data, demonstrating the weather's predictive power in this area as well [3].

As per the authors' discourse, Wang et al. devised an innovative approach that enhances the precision of traffic flow forecasts through the employment of deep learning methodologies, specifically Long Short-Term Memory (LSTM) neural networks, AdaBoost, and gradient descent. As per the authors' discourse, Wang et al. devised an innovative approach that enhances the precision of traffic flow forecasts through the employment of deep learning methodologies, specifically Long Short-Term Memory (LSTM) neural networks, AdaBoost, and gradient descent.[4]

Aditya with every one of them The Support Vector Regression approach is a useful tool for traffic flow

prediction, according to a number of studies. Traffic flow forecasts on an hourly, daily, monthly, or even annual basis can be made using it.[5]

Our research aims to advance traffic prediction by providing a thorough methodology that considers several machine learning and deep learning techniques in addition to meteorological factors. In order to accomplish this, we will examine these study publications to identify any gaps in knowledge. We want to fill in the following details in order to develop a trustworthy traffic forecast model that might significantly enhance urban transportation management, reduce traffic, boost road safety, and support a more sustainable urban environment, these gaps in the literature.

II. METHOD

To ascertain the traffic state, we employ datamining as the prediction model. Data mining is a technique that uses specific algorithms to sift through massive amounts of data in search of hidden information. Combining databases, machine learning, artificial intelligence, statistics, and many other topics is known as data mining.

Three steps are involved in data mining:

1) Rata preparing herself. We first define the question and the main problem in this part. After that, the database is set up and the necessary data is collected. We then preprocess the data after that.
2) acquiring information.We acquire pertinent information from the internet and other sources and accurately evaluate the data.
3) interpreting and analyzing data. Using certain techniques, we first select and construct the appropriate model in this section. After that, evaluate the model and report the results.

Four ways are usually used in data mining.

1) Neural network methodology. It draws inspiration from the human brain's neural network, which is trained to operate as a nonlinear prediction model.By mimicking the action of human brain neurons, neural network algorithms may carry out numerous data mining tasks, such as grouping, classification, feature mining, and so forth. Associative memory, non-linear learning, and anti-interference are a few of its many benefits. In complicated circumstances, it also produces precise forecast results. Still, it cannot obtain learning stages in the process and is not suitable for processing high dimensional data. Second, interpreting the results is difficult. Thirdly, making predictions takes a long time[6].

2) The algorithm used by Bayes. This type of classification method is grounded in mathematical statistics and probability. There are two types of Bayesian algorithms. Naive Bayes is the first. Categorization accuracy is higher when each attribute functions independently of the others.If not, it might be lower. The second is the Bayes network with tree augmentation. The ability to recognize the dependencies among the data's intrinsic features and utilize those connections to categorize the information. Large databases can be used with Bayesian. It is easy to use, quick, and accurate.[7].

3) the algorithm for linear regression. The least square function is a statistical technique known as "linear regression" that is used to depict the relationship between one or more independent variables and dependent variables. A linear combination of one or more model parameters makes up this 31 function. When there are multiple independent variables, the condition is referred to as numerous regression. When each attribute in the data is a number and the prediction is a number as well, It is going to be regarded as the finest algorithm[8].

4) Decision Tree Approach. It is a technique for categorizing data according to a collection of guidelines; it resembles the tree structure's flowchart. It is not required to have a protracted building procedure. It is easy to categorize, define, and comprehend. One drawback is that it could be challenging to identify rules based on the fusion of several factors.The approach is especially useful for processing large volumes of data and performs well when handling non-numerical data. A means of illustrating regulations, such as the values that will be earned in what situations, is with a decision tree[9].

In this study, we use the SVM, Decision Tree, and Linear Regression techniques to make the prediction. Next, these three models are compared to decide which prediction model is the best.

A. Web retrieving

1) Selenium

Browser control may be achieved with Python APIs thanks to an application called Selenium.Using Selenium, our team launches a browser, goes to an online map page, and snaps a screenshot [10].

2) WebStorage.

The decision to employ an Object Storage System (Object Storage System) to hold the picture data so that the screenshots can be saved for a later photo and data analysis. When it comes to handling large volumes of data, the Ass system is more scalable than the File Storage System since it is flattened[11].

3) MessageQueue.

We stored the data regarding the examined and unexplored using Message Buffer. photographs

temporarily.Focusing on a tiny area of the map means we can only develop one picture fetching module and one image analyzing module.However, when more road analysis is needed, this strategy is unsatisfactory.Message Queue increases the scalability and resilience of our system.[12]

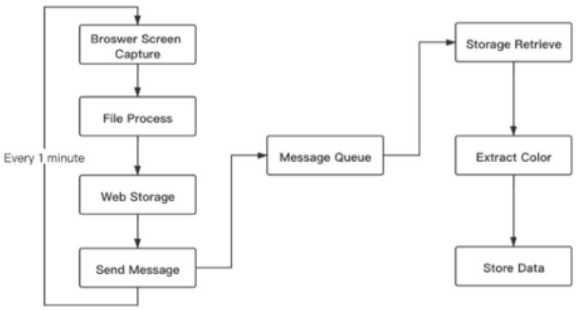

Figure.1 Process of original data retrieving and analysis

B. Machine Learning

1) Support Vector Machine(SVM):

One supervised machine, technique that is well-known for its efficiency in classification and regression applications is the Support Vector Machine (SVM). Situations with high-dimensional feature spaces and complex data distributions are especially well-suited for it. Finding the ideal hyperplane to divide data points into distinct groups or forecast continuous values in regression tasks is the fundamental idea of support vector machines (SVM).

The Support Vector Machine (SVM) algorithm has become a potent tool for predicting traffic congestion because it is accurate and adaptable when modeling intricate traffic patterns. It is a suitable choice for our research because of its capacity to manage noisy inputs, non-linear interactions, and high-dimensional data.

2) Decision Tree:

Widely utilized and adaptable machine learning algorithms, decision trees are renowned for their efficacy and interpretability in both classification and regression problems. They have found use in a number of industries, including banking, healthcare, and more.

It's crucial to remember that Decision Trees do have certain drawbacks, such as their tendency to overfit, which can be lessened by employing ensemble approaches or pruning, among other suitable strategies. They are a useful option in prediction due to their interpretability and simplicity of use.

3) Random Forest:

Over the years, there have been notable developments in machine learning, with the Random Forest algorithm being one of the most notable. Applications for Random Forest's robust and adaptable ensemble learning technique can be found in many different fields. We will explore the foundations of Random Forest, its construction, benefits, and real-world applications in this thorough review.

An ensemble learning method called Random Forest is applied to both regression and classification problems. Each decision tree in the ensemble was trained using a distinct subset of the data. The concept of adding randomization to the building process, which reduces overfitting and enhances model performance, is reflected in the term "Random Forest".

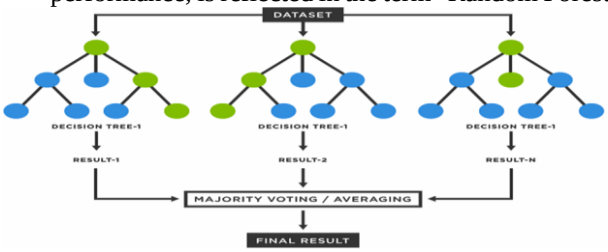

Figure.2 Structure of Random Forest

4) Bayesian Ridge Regression:

A statistical method used in regression analysis and machine learning is the Bayesian Ridge regression algorithm. This particular kind of linear regression was created to address a few issues with more conventional linear regression models.

C. DEEP LEARNING MODELS:

Convolutional neural networks (CNNs) and long short-term memory networks (LSTMs) are two effective deep learning models in the field of traffic congestion prediction. Whereas LSTMs are adept at capturing temporal correlations in traffic time series data, CNNs are particularly good at identifying spatial patterns from visual data, such as traffic camera photos. Through the integration of CNNs' proficiency in extracting spatial features and LSTMs' aptitude in modeling sequences and traffic conditions that change over time, these models provide a comprehensive understanding and forecasting of traffic congestion, thereby enabling more efficient traffic management and congestion mitigation tactics. Let's talk about each algorithm in detail.

1) Convolutional Neural Networks (CNNs):

Convolutional neural networks, or CNNs, are useful for analyzing and forecasting traffic congestion. A particular kind of deep learning model called CNN is mostly intended for handling and analyzing structured grid data, such pictures and movies. CNNs have revolutionized domains such as image classification, object recognition, and image segmentation, and have become an essential tool in computer vision applications.

Although CNNs are typically used to process image data, they may also be modified to handle non-image data by applying a technique known as 1D or 2D convolution, depending on the structure of the data. Examples of this type of data include tabular traffic statistics and sensor data.

2) Long Short-Term Memory (LSTM):

Recurrent neural networks (RNNs) with Long Short-Term Memory (LSTM) architecture are specifically made to handle and model sequential input. When it comes to identifying and comprehending patterns, dependencies, and connections among data sequences, LSTMs are especially good. They are extensively employed in many different applications, such as speech recognition, time series analysis, and natural language processing. LSTMs are different from conventional RNNs in that they can preserve long-term dependencies and address the issue of the vanishing gradient. Specialized memory cells and gating mechanisms enable long-term sequences of selectively storing, updating, and retrieving information, which is how LSTMs accomplish this. Because of this, LSTMs are a good fit for tasks involving time series data where comprehending context and distant relationships is crucial for accurate predictions and modelling.

A Long Short-Term Memory network (LSTM) is a customized deep learning model used to analyze and anticipate traffic conditions over time in the context of traffic congestion prediction. Because they can collect and model the intricate temporal correlations and patterns found in traffic data, LSTMs are a good fit for this purpose.

III. PERFORMANCE EVALUATION

We accomplish this by assessing their performance using the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), two widely used metrics. These metrics enable us to assess the degree to which our forecasts match the actual facts. When it comes to forecasting traffic congestion, we can objectively determine which prediction techniques are effective and which ones would require improvement by utilizing RMSE and MSE.

Every model must be examined once it has been completed. For activities involving deep learning and machine learning, evaluation measures are crucial.

A. Evaluation Parameters:

Metrics for evaluation are useful in determining how well the prediction model works. Two significant prediction model metrics are covered in this section.

Mean Squared ERROR (MSE):

The statistical model's error level is measured by mean squared error, or MSE. The average squared difference between the observed and anticipated values is evaluated. The MSE is equal to 0 in a model that has no errors. The value of the model increases with the error. Another name for the mean squared error is the mean squared deviation (MSD).

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n}$$

Rooted Mean Squared Error (RMSE):

A key indicator in predictive modeling is the root mean square error (RMSE), whose value represents a model's performance. The average difference between the anticipated and actual values of a statistical model is measured by RMSE. It is the standard 33 deviation of the residuals in mathematics. The distance between the data points and the regression line is represented by residuals.

The RMSE measures the degree to which these residuals have been dispersed, providing insight into how well the observed data adheres to the expected values. The RMSE decreases as the data points approach the regression line because the model has less error. Predictions made by a model with lower error are more accurate.

RMSE values are expressed in the same units as the dependent (outcome) variable and can span from zero to positive infinity. When the predicted and actual numbers are exactly the same, the result is 0. Low RMSE values show that the model has more accurate predictions and matches the data well. Higher levels, on the other hand, indicate greater mistake and less accurate predictions.
The RSME formula for a sample is the following:

$$MSE = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{N - P}}$$

IV. COMPARISION AND DISCUSSION

A. COMPARISION

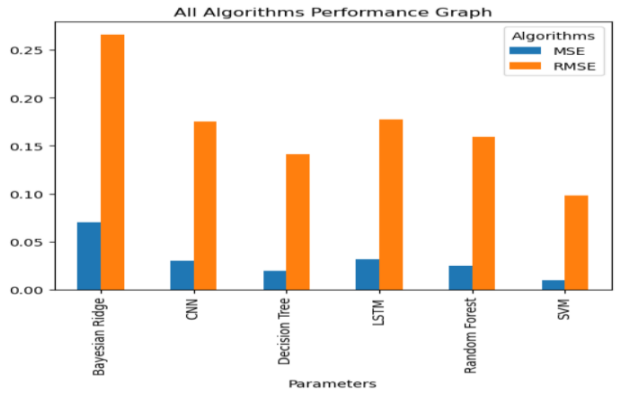

Figure.3 All Algorithms Performance Graph

The Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) values for each method are displayed in the bar chart (Figure.3) that was previously mentioned. The y-axis on this chart shows the associated error numbers, and the x-axis reflects the various algorithms or parameters.

Algorithm Name	**Mean Squared Error (MSE)**
Random Forest	0.00965160562044798
LSTM	0.030584202891979474
CNN	0.018035024129894092
Decision Tree	0.019936270731283737
SVM	0.025308814693551094

Baysian Algorithm	0.07062080416605596

Table1. All Algorithms Mean Squared Error (MSE) Performance

Algorithm Name	**Rooted Mean Squared Error (RMSE)**
Random Forest	0.09824258557493273
LSTM	0.17488339798842964
CNN	0.13429454244269978
Decision Tree	0.14119585946933336
SVM	0.15908744354458365
Baysian Algorithm	0.26574575098401093

Table2. All Algorithms Rooted Mean Squared Error (RMSE) Performance

A thorough comparison of all the models is given in Tables.1 and 2. These tables display the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) values for each model. A succinct and understandable evaluation of the model's prediction accuracy performance is made possible by the tabular data.

V. CONCLUSION

During our thorough analysis of several traffic prediction algorithms, we found that their efficacy and accuracy varied. With the lowest Mean Squared Error (MSE) of 0.00965 and Root Mean Squared Error (RMSE) of 0.098242, the Random Forest algorithm stood out as the most accurate in predicting traffic patterns. The LSTM (Long Short-Term Memory) model, which showed excellent accuracy with an MSE of 0.03058 and an RMSE of 0.17488, came in close second.

Closely lagging the LSTM, the CNN demonstrated robust performance as well, with an MSE of 0.01803 and an RMSE of 0.13429. With an MSE of 0.01993 and an RMSE of 0.14119, the Decision Tree model demonstrated its performance in properly predicting traffic.

Despite its proficiency, the Bayesian Ridge algorithm exhibited slightly more mistakes, with an RMSE of 0.26574 and an MSE of 0.07062. It nevertheless continued to offer insightful forecasts. Lastly, with an MSE of 0.025308 and an RMSE of 0.15908, the Support Vector Machine (SVM) demonstrated respectable performances.

In conclusion, our analysis showed that the Random Forest algorithm was the best option for accurate traffic predictions, with LSTM and CNN coming in close second. These models outperformed others in terms of accuracy and prediction error minimization.

VI. REFERENCES

[1] Jingyi, Lu. (2023). An efficient and intelligent traffic flow prediction method based on LSTM and variational modal decomposition. Measurement: Sensors, doi: 10.1016/j.measen.2023.100843

[2] Bashir, Mohammed., Nandini, Krishnaswamy., Mariam, Kiran. (2019). Multivariate Time-Series Prediction 10.1109/ANCS.2019.8901870 for Traffic in Large WAN Topology. doi:

[3] Lu, Wang., Rui, Wu., Weichun, Ma., Weiju, Xu. (2023). Examining the volatility of soybean market in the MIDAS framework: The importance of bagging-based weather information. International Review of Financial Analysis, doi: 10.1016/j.irfa.2023.102720

[4] Aditya, Srivastava., Aryan. (2023). A Survey Paper on Traffic Prediction Using Machine Learning. International Journal For Science Technology And Engineering, doi: 10.22214/ijraset.2023.51390

[5] Gobezie, Ayele, and Marta Sintayehu Fufa. "Machine learning and deep learning models for traffic flow prediction: A survey." (2020).

[6] Agafonov A. A., Short-Term Traffic Data Fore casting: A Deep Learning Approach, [J]Optical Memory and Neural Networks Volume 30, Issue1. 2021.PP1-10

[7] Banchhor Chitrakant; Srinivasu N., Analysis of Bayesian optimization algorithms for big data classification based on Map Reduce framework, [J]Journal of Big Data Volume 8, Issue1.2021.

[8] Peng Jianan; Liu Wei; Bretz Frank; Hayter A J, Simultaneous confidence tubes for comparing several multivariate linear regression models., [J]Biometrical Journal 2021.

[9] Camilo Gutierrez-Osorio, Cesar Pedraza.Modern data sources and techniques for analysis and forecast of road accidents: A review[J]. Journal of Traffic and Transportation Engineering (English Edition), 2020, 7(04):432-446.

[10] R. S.Chaulagain, S. Pandey, S.R.Basnet and S.Shakya, "CloudBased Web Scraping for Big Data Applications," 2017 IEEE International Conference on Smart Cloud (SmartCloud), 2017, pp. 138-143, doi: 10.1109Cloud.2017.28.

[11] M. Factor, K. Meth, D.Naor, O. Rodehand1. Satran, "Object storage: the future building block for storage systems,"2005 IEEE International Symposium on Mass Storage Systems and Technology, 2005, pp. 119 123,doi: 10.1109/LGDI.2005.1612479.

[12] Johansson, L. and Dossot, D, "RabbitMQ Essentials: Build distributed and scalable applications with message queuing using RabbitMQ, 2nd Edition, "Packt Publishing, 2020.