Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 5 (2023)

Application of Different Machine Learning Methods to Predict Traffic Flow

K.B.V. Satya Prakash[1], Dr. Kiran Kumar Billa[2*], Dr. Siva Kumar Perumal[3], Dr. M. Satya Srinivas[4], Dr. MVSS Nagendranath[5]

[1]Student, M. Tech, Department of CSE, Sasi Institute of Technology & Engineering, Tadepalligudem
[2*]Assoc. Prof & Faculty by Research, Department of ME, Sasi Institute of Technology & Engineering, Tadepalligudem
[3]Assoc. Prof, Department of CSE, Sasi Institute of Technology & Engineering, Tadepalligudem
[4]Asst. Prof, Department of CSE, Sasi Institute of Technology & Engineering, Tadepalligudem
[5]Prof. & HOD, Department of CSE, Sasi Institute of Technology & Engineering, Tadepalligudem

Abstract*:* Nowadays, with the proliferation of automobiles, it's become more difficult to precisely anticipate traffic volumes. In recent years, traffic bottlenecks have become more widespread. A dataset containing information about traffic volumes is used. Find the accuracy, mean absolute error, mean squared error, and Root Mean Squared Error (RMSE) for the anticipated traffic volume using Logistic Regression, Support vector Machines, random forest method, and Naive-Bayes machine learning algorithms. Results from using Logistic Regression outperformed those from using Support Vector Machines, Random Forest and Naïve Bayes classifiers are verified using different metrics like MSE, MAE and Regression values. Moreover, out of all the algorithms used, Logistic Regression obtained the lowest MSE of 1.1775e-30, lowest MAE of 7.9028e-16 and higher Regression Coefficient of 0.9999997. Future plans for road construction and widening might be informed by the anticipated amount of traffic.

Keywords: Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), machine learning, Root Mean Squared Error (RMSE).

About 1.5 lakh people die in road accidents in India each year, with an average of 1,130 accidents and 422 fatalities every day, or 47 accidents and 18 fatalities per hour, according to official statistics [1]. For those aged 15 to 49, injuries sustained in car accidents rank first among all causes of mortality worldwide. Around 1.3 million individuals lose their lives in traffic-related incidents every year. In order to reduce the number of accidents and fatalities, it is crucial to identify the traffic volume early on in order to enlarge the existing highways. The project's overarching goal is to use logistic regression to forecast traffic volumes using existing traffic data sets.

1. Introduction

Although Road accidents are common, they are still the most unfortunate thing that can happen to a motorist. The worst part is that we drive around making the same errors over and over again. It is the carelessness of drivers, rather than a lack of awareness, of traffic laws and safety protocols, that is the primary cause of collisions and accidents[2]. Human mistake is the leading cause of accidents and crashes. Some of the most typical human actions that lead to accidents are like Distracted Driving, Drunk Driving, Speeding, Jumping Red Lights, and not using safety equipment like helmets and seat belts 6. Disregarding lane markings and improperly passing. These are the most typical driving behaviors that cause accidents, according to studies from throughout the world.

Going too fast for conditions causes the majority of fatal incidents. Achieving greatness is inherent to the human spirit. Man will surely reach limitless speed if given the opportunity. When we share the road with other drivers,

however, we can't help but follow closely behind other vehicles. Accidents are more likely and more severe while traveling at higher speeds. Automobiles traveling at higher speeds are more likely to be involved in accidents, and those accidents that do occur are more likely to be catastrophic. Increasing the speed increases the danger. The braking distance increases as the vehicle's speed increases [3]. According to the law of concept, a slower car will stop instantly, while a quicker one will take a long time to stop and will skid over a long distance as well. When a car is going fast, it will have more of an impact when it crashes, which means more people will be hurt. Accelerating reduces one's capacity to foresee what's coming next, which in turn increases the likelihood of making a mistake in judgment and, ultimately, a collision.

As a general rule, people drink to mark special occasions. When combined with driving, however, it transforms joy into tragedy. Alcohol impairs focus. It makes people's bodies respond faster. When the brain sends signals to the limbs, the reaction time increases. It causes vertigo, which impairs eyesight. People are less afraid and more willing to take chances when they're drunk. All these things when driving create accidents and many a times it proves deadly. The risk of an accident increases twofold for every 0.05-point rise in blood alcohol concentration. In addition to alcohol, several substances and medications may impair the focus and coordination needed for safe driving. To start, it's not a good idea to drink alcohol [4]. But please do not drive after consuming alcohol, even if you believe that your celebration would be incomplete without it. Get a buddy who doesn't drink alcohol to drive you home.

Even a little distraction while driving may lead to serious collisions. Anything inside or outside the car might be a potential distraction. These days, chatting on a cell phone while driving is the biggest distraction. The larger part of the brain is responsible for speaking on the phone, while the smaller part is responsible for driving. Reaction time and judgment are both negatively impacted by this brain division. One of the causes of crashes is this. It is dangerous to talk on the phone and drive [5]. You should pull over to the side of the road and answer an emergency call. One example of a potential roadside distraction are like adjusting vehicle mirrors while driving, Stereo/Radio in vehicle, Street animals, Banners and billboards. To be safe during diversions and other types of outside distractions, the motorist should not be distracted by these items and should lower speed.

At many junctions, drivers often disregard the traffic signal and go through the intersection nonetheless. Time savings is the driving force behind red light jumping. Most people think it's a waste of time and gas to stop at a red light. Commuters save time and arrive at their destinations safely and on schedule when traffic signals are obeyed correctly by all cars, according to studies. When a driver disregards a red light, he puts himself and everyone else on the road in danger. The crossing becomes a complete shambles when one car does this, which encourages other drivers to do the same. The major reason for traffic bottlenecks is the disorder at intersections. Everyone is late for something at some point. Another observation is that red light jumpers tend to cross the junction faster in an effort to avoid wrecks and challans. However, this strategy frequently backfires since it impairs their ability to gauge the flow of traffic, leading to accidents.

Not only is it now illegal to operate a motor vehicle without a seat belt, but two-wheeler riders are now subject to penalties for failing to wear protective headgear. Research showing that using safety equipment like seat belts and helmets may significantly lessen the impact of an accident led to its legalization [6]. The odds of surviving a catastrophic accident are double for those who wear helmets and seat belts. You may avoid serious injury or even death with the help of safety equipment. The use of helmets has been mandated, leading to a significant decrease in two-wheeler fatalities. For maximum protection, always use safety equipment that meet the standards and always knot them correctly. A subfield of computer science known as "machine learning" allows computers to pick up new skills without human intervention. One of the most fascinating machine learning systems ever seen. It makes the computer more like a person by giving it the capacity to learn, as the name implies. There may be more places than you think using machine learning right now. There are primarily four categories of machine learning, and they are: Supervised, Unsupervised, Semi-Supervised and, Reinforcement Machine Learning. A key distinction between supervised and unsupervised learning is the absence of human oversight, as the term implies. What this implies is that in unsupervised machine learning, the computer learns from an unlabeled dataset and makes predictions about its own output with no human intervention. Without human oversight, models trained in unsupervised learning use unlabeled or unclassified data to make predictions.

The goal of data science is to draw actionable insights from the vast and growing amounts of data collected by modern businesses. As a whole, data science entails cleaning and organizing data, doing complex analyses on that data, and then presenting the findings to stakeholders so that they may see patterns and make informed choices.

In order to get data ready for certain types of processing, it is necessary to clean, aggregate, and change the data.[7] Algorithms, analytics, and artificial intelligence models must be developed and used for analysis. Software runs the show, finding patterns in the data and turning them into predictions that companies may use to their advantage. Thoroughly planned tests and studies are necessary to validate these predictions, and the results should be shared using data visualization tools that everyone can use to see trends and patterns.

One of the difficulties in predicting traffic flows is dealing with datasets that are uneven. On datasets with unbalanced data, conventional binary classification algorithms sometimes provide less-than-ideal results, leaving users dissatisfied. Secondly, using an unbalanced dataset to determine the most important parameters for predicting traffic flow. Thirdly, accuracy, mean square error, and root mean square error may be determined using a variety of machine learning techniques, including Logistic Regression, Support Vector Machines, Random Forest classifier, and Naïve Bayes[8].

One of the main challenges is establishing accurate predictions of traffic flows and then developing and constructing new road infrastructure appropriately. Numerous individuals lose their lives or become disabled every day in vehicle accidents caused by poorly maintained roads around the nation, which has a knock-on effect on the financial stability of the families of those who have passed away. Better road infrastructure and public education on the importance of road safety (by government agencies, nonprofits, and individuals) may lower the number of road fatalities [9]. In order to better understand how to use machine learning techniques to forecast traffic flow, this scholarly article provides a comprehensive overview of the existing literature. With an emphasis on regression fit, MAE, MSRE, and scalability, the article discusses a range of ML algorithms and how they are applied to the problem of traffic flow prediction. It also indicates possible future study fields and addresses the difficulties and limits of present methodologies. The article covers all the bases, giving readers an idea of what the profession is like and how machine learning may change things for the better in areas like urban planning and traffic management. In order to tackle this issue, traffic data pertaining to the number of bikes, cars, buses, and trucks was collected at fifteen-minute intervals throughout the day for a month. Logistic Regression, Support Vector Machines, Random Forest, and Naive-Bayes algorithms were among the many machine learning techniques used to forecast traffic flows[10]. When the class unbalanced distribution is ignored, conventional binary classification algorithms often underperform the dataset, resulting in the less-than-ideal outcome of accurately identifying the majority class while mislabeled the minority. The unequal distribution of previous knowledge has a significant effect on final discrimination in many cases, which is one explanation for this. As a result, the planned effort employs.

Machine learning is becoming more important and promising for transportation engineers, especially in traffic prediction. To reduce traffic, regression models and libraries like pandas, os, numpy, matplotlib, and pyplot are utilized for traffic prediction. This must be done to reduce traffic and make it simpler to access. Traffic flow data and congestion flow data from start to finish may be seen using hourly data. Users' probable road conditions may be displayed. Comparing the mean square errors of data from the prior year with the current year might indicate traffic accuracy. User may also discover the average number of automobiles on the route with traffic prediction. S. Shahriari et al. studied several non-parametric and parametric traffic volume prediction. The fundamental drawback of parametric approaches is poor prediction accuracy. Non-parametric approaches forecast better but lack theoretical backing, hence they are critiqued. This research applies bootstrap to the parametric ARIMA model [11] to improve forecast accuracy while retaining theory adherence. Each ARIMA model in E-ARIMA is created using a random subsample of data. Comparing E-ARIMA with parametric and non-parametric approaches like ARIMA and LSTM tests the model's validity. A year of traffic count data on four Sydney arterial routes is utilized for calibration and validation. The findings imply that ensemble modeling enhances prediction accuracy. This study presents E-ARIMA, a short-term traffic volume forecast ensemble of ARIMA models. A controlled experiment tests E-ARIMA's potential. Next, traffic count data from four important roadways in Sydney, Australia, is utilized to compare E-ARIMA against ARIMA and LSTM, two popular traffic volume

prediction algorithms. E-ARIMA greatly increases the forecast accuracy of the standard ARIMA model, according to the findings. Our results confirm past research that LSTM has greater prediction accuracy than ARIMA, however applying the ensemble approach to ARIMA increases the model's prediction accuracy, thus E-ARIMA's prediction accuracy measures are higher than LSTM's. This supports the study's claim that "the ensemble method improves prediction accuracy". Each ARIMA model represents data differently, however E-ARIMA represents data more accurately. In addition, E-ARIMA surpasses LSTM in spatial validity across the four highways with reduced root mean square error and mean absolute percentage error. E-ARIMA is more transferable than LSTM since it uses parametric ARIMA models. This study suggests ensemble approaches may improve traffic volume forecast. Hourly data was obtained for this investigation. Thus, this research only predicts 1-hour horizons. Model application on 5 minute and 15 minute prediction horizons would be interesting to study. This research used the ensemble approach on ARIMA models, however it can also be used on LSTM models. Future research may examine more spatially transferable machine-learning methodologies. This research did not attempt to further machine learning algorithms. To increase spatial transferability, the ARIMA model should include road physical design factors like lanes and shoulders. Another research topic the authors are considering is adding GARCH models to the ensemble. Priyaet al. [12] discussed Traffic congestion is society's biggest issue. This article focused on road traffic forecast. This machine learning and Hidden Markov model project will assist users choose the best path to their destination. Traffic volume prediction is the important technology in intelligent transportation systems, according to Mao et al. Traffic volume prediction uses BP neural network worldwide. This project sought to improve BP neural network psilas traffic flow prediction. Traffic volume prediction using subsection learning of double-layer BP neural networks[13]. Jingshi road traffic volume was predicted using the enhanced approach in Jinan city, then compared to subsection-learning and conventional methods. Using subsection-learning, average relative tolerance dropped 2.52%. Improved BP neural network predicts traffic volume. Nazirkar Reshma Ramchandra et al. [14] discussed work zones, weather conditions, and special occasions like festivals and festivities caused traffic congestion because drivers behaved differently. Road traffic may be predicted using machine learning. Multiple domains impact communication technology. Since technology has advanced, traffic forecasting uses machine learning. This model incorporates DAN, DBN, RF, and LSTM. The suggested model may be evaluated using machine learning measures including accuracy, precision, recall, and error value. The aforementioned four methods provide 95.2% accuracy for LSTM. Ge Zheng et.al [15] proposed Traffic prediction in Intelligent Transportation Systems (ITSs) enables sophisticated transportation management and services to alleviate rising traffic congestion. Traffic prediction has progressed from basic statistical models to complicated deep learning model integration in recent decades. We evaluate contemporary hybrid deep learning models for traffic prediction in this research. We examined and taxonomized models via feature extraction approaches to achieve this. We examine their components and architecture. Ten models from our taxonomy representing various architectural choices were compared for performance. We reconstitute the chosen models and run a series of similar comparison tests using three well-known real-world datasets from large-scale road networks. Y. Gao et al. [16] Accurate traffic speed forecasts may assist traffic management departments make better decisions and enhance road monitoring, as well as help drivers plan their routes and arrive safely. Considering the unavailability of traffic speed data, this research provides a technique for predicting traffic speed using the multi temporal traffic flow volume of previous and later moment states. Traffic flow volume data was used to derive traffic patterns from prior and subsequent moment states. Second, five forecasting models—LSTM, BP, classification and regression trees, k-nearest neighbor, and support vector regression—were compared. Finally, sensitivity analysis tests for various traffic patterns were performed using the best prediction model. A real-data case study showed that the LSTM model predicts time and space better than other models. This traffic pattern "previous = 3 and later = 3" may predict traffic speed more correctly and robustly across circumstances.

N. K. ChikkaKrishna, P. Rachakonda, T. Tallam [17], suggested "Short-Term Traffic Prediction Using Fb-PROPHET and Neural-PROPHET. Developing a STTP model to anticipate traffic patterns is one of the biggest challenges today. Since car and traffic numbers are rising rapidly, congestions occur. A project to anticipate short-term road traffic is gaining steam because precise traffic flow forecasts help reduce road congestion. Facebook-developed Fb-Prophet models predicted time series data patterns. This work built STTP models to estimate traffic volume using Fb-PROPHET and Neural-PROPHET. Classified traffic volume count for seven days-24 hours on

National Highway 744 in Tamil Nadu was pneumatically collected. MAE, RMSE, and MAPE tests were used to evaluate the model's fit. The proposed study helps traffic management organizations plan and allocate routes to prevent congestion. Wanqiu Lou, et.al [18] studied traditional prediction approaches fail for short-term traffic flow time series because they are nonlinear, non-stationery, complicated, and stochastic. This research provides a glowworm swarm optimization (GSO)-optimized least square support vector regression (LSSVR) short-term traffic flow forecasting model. Two key learning parameters that affect model prediction are determined by the GSO algorithm. The LSSVR-GSO model is tested using traffic flow data from one highway stretch in Chengdu, China. The experimental findings reveal that the LSSVR-GSO model predicts better than the genetic algorithm and back-propagation neural network models. Aljuaydi et al. [19] provide multivariate machine learning-based highway traffic flow prediction models for non-recurrent incidents. MLP, CNN, LSTM, CNN-LSTM, and Auto encoder LSTM networks have been created to anticipate traffic flow after a road collision and rain. Model performance is assessed using an input dataset with five characteristics (flow rate, speed, density, road incident, and rainfall) and two standard metrics (Root Mean Square error and Mean Absolute error). Zihao Wang et al. [20] First, we analyze traffic features and compare state-of-the-art traffic feature creation methods. Then, we propose a novel encrypted traffic feature concept designed for encrypted malicious traffic analysis. We also present an encrypted malicious traffic detection system. Deep learning and machine learning algorithms make up the two-layer detection framework. Comparative trials show it outperforms conventional deep learning and ResNet and Random Forest. Additionally, we curate a dataset of only public datasets to train the deep learning model. The combined dataset is more complete than any available dataset. Many nations have implemented different reaction measures to slow the spread of the COVID-19 pandemic since 2020. These measures have affected everyday life, the economy, education, social and recreational activities, and domestic and foreign travel. Various policies and strategies indirectly affect urban traffic mobility. Such methods and strategies have altered urban traffic flows. However, their effects are unquantified. Estimating the influence of these coupled but diverse policies and actions on urban traffic flows is difficult. This research presents an artificial neural networks (ANN) model that links urban traffic flows with the enforced response measure and other parameters. The findings demonstrate that the selected ANN model can accurately and efficiently map the complicated link between traffic flows and response measures. Close to observed values are expected. They cluster around the regression line with R2 0.9761. Additionally, the model may be used to predict demand from any post-pandemic response measure. Mega-event traffic may be managed using this approach. In catastrophe or emergency circumstances, traffic flow estimations are crucial for operations and planning [21]. Yan Zheng Chunjiao Dong et al. [22] present a fusion deep learning model with spatial–temporal correlation to forecast urban road traffic flow. First, this study argues that the traffic flow of a section in the urban road network relies on its own time series and on the traffic flow of other sections in the vicinity. Thus, wavelet decomposition and dynamic temporal warping are used to filter parts with traffic flow similarity to the target region. Second, differential approach reconstruction of unstable time series into stationary time series improves prediction accuracy. Finally, we feed the CNN-LSTM fusion deep learning model for traffic flow prediction the extracted traffic flow data of a comparable part as an independent variable and the target section as a dependent variable. We found that the suggested model is more accurate and stable than the benchmark models. The MAPE can achieve 92.68%, 93.39%, 85.14%, and 76.14% at 5 min, 15 min, 30 min, and 60 min, and the other evaluation indices are better than the benchmark models. Boris Medina-Salgado et al. [23] also proposed “Urban traffic flow prediction techniques: A review” Transport infrastructure has improved, yet traffic issues continue to spread owing to population growth in metropolitan areas that utilize these transports. Congestion control issues affect people via air pollution, fuel consumption, traffic violations, noise pollution, accidents, and time loss. clever Transport Systems use the internet of things and clever algorithms to gather data from numerous sources and analyze it to optimize traffic flow. Traffic data processing and modeling are difficult owing to road network complexity, space–time relationships, and varied traffic patterns. This review study groups smart techniques used for mobility data analysis in urban traffic flow prediction, shows their results, and describes and analyzes their benefits and drawbacks. Given the above, (iv) the data sets used in the literature and available for use are shown, (v) the quantifiable results of precision of the various techniques were compared, highlighting advantages and limitations, which allows us to (vi) identify the related challenges and (vii) propose a general taxonomy in which the knowledge acquired in this traffic flow review converges from a computational approach. Nigam et al. [24] Fog, rain, and snow impair driver sight, vehicle movement, and road capacity. In an Intelligent

Transportation System, traffic operation and management need accurate prediction of macroscopic traffic stream factors like speed and flow. Because the traffic stream is non-linear and complicated, predicting these factors is difficult. Because layer-wise design removes buried abstract representation, deep learning models predict traffic stream factors better than shallow learning models. Traffic is affected by weather conditions due to hidden factors. Terrain limits prevent the weather station-road distance from immediately affecting traffic during rain. Waterlogging results from persistent rainfall weakening the drainage system and soil absorption. To capture the geographical and extended influence of weather, we presented a soft spatial and temporal threshold mechanism. To fill meteorological data gaps, spatial interpolation is performed. We created CNN-LSTM and LSTM-LSTM hybrid deep learning models. Former model in hybrid model extracts spatio-temporal characteristics, which later model employs as memory. The latter model predicts traffic stream characteristics using features and temporal input. Multiple trials verify the deep learning model's performance. The trials reveal that a deep learning model trained with traffic and rainfall data predicts better than one without. The LSTM-LSTM model extracts long-term traffic-weather dependence better than previous models. Vladimir Shepeleva et al. [25]. Published a hybrid model for continuous road traffic spatio-temporal dependency monitoring and prediction. The model uses a YOLOv4 CNN and two-level long-short-term memory. Considering transport infrastructure and climatic elements, the hybrid model's first module collects and fills road traffic parameter data. Recurrent neural networks (RNNs) provide better predictions using interpreted and aggregated data. Experimental findings reveal that the suggested model yields 82–97% prediction accuracy over a 20-minute time span with few observations. Paul et al. [26] published a model that might assist drivers make smarter judgments, minimize traffic congestion, carbon emissions, and traffic operation efficiency. Traffic flow prediction aims to give such information. With the fast development and implementation of Intelligent Transportation Systems, traffic flow prediction is gaining interest. It is essential for the effective deployment of ITS subsystems including advanced traveler information systems, traffic management systems, public transit systems, and commercial vehicle operations. Historical and real-time traffic data from inductive loops, radars, cameras, mobile GPS, crowd sourcing, social media, and others is crucial to traffic flow prediction. Traffic data are expanding due to old and new traffic sensor technologies, ushering in Big Data transportation. Transportation management and control are becoming data-driven.

2. Methodology

In order to manage vehicle movement, reduce congestion, and generate the ideal (least time- or energy-consuming) route, traffic prediction involves anticipating the density and volume of traffic flow. In order to produce reliable traffic forecasts, it is necessary to take into account all of the elements that affect traffic. Consequently, you will need to collect data from a number of primary sources.

A comprehensive map with road networks and other features is required initially. An excellent approach to get comprehensive and current information is to connect to such worldwide mapping data suppliers as OSM, HERE, Google Maps, or TomTom. To gather traffic data, including how many cars were there at a specific time, how fast they were going, and what kinds of vehicles were passing through (trucks, light vehicles, etc.). Loop detectors, cameras, weigh-in motion sensors, radars, and other sensor technologies are used to get this data. Thankfully, it is not necessary to individually put these devices everywhere. Those service providers make it easy to access this data by compiling it from a network of sensors, a variety of third-party sources, or by using GPS probe data. Current, historical, and anticipated weather data is also required because of the influence that weather has on road conditions and driving speed.

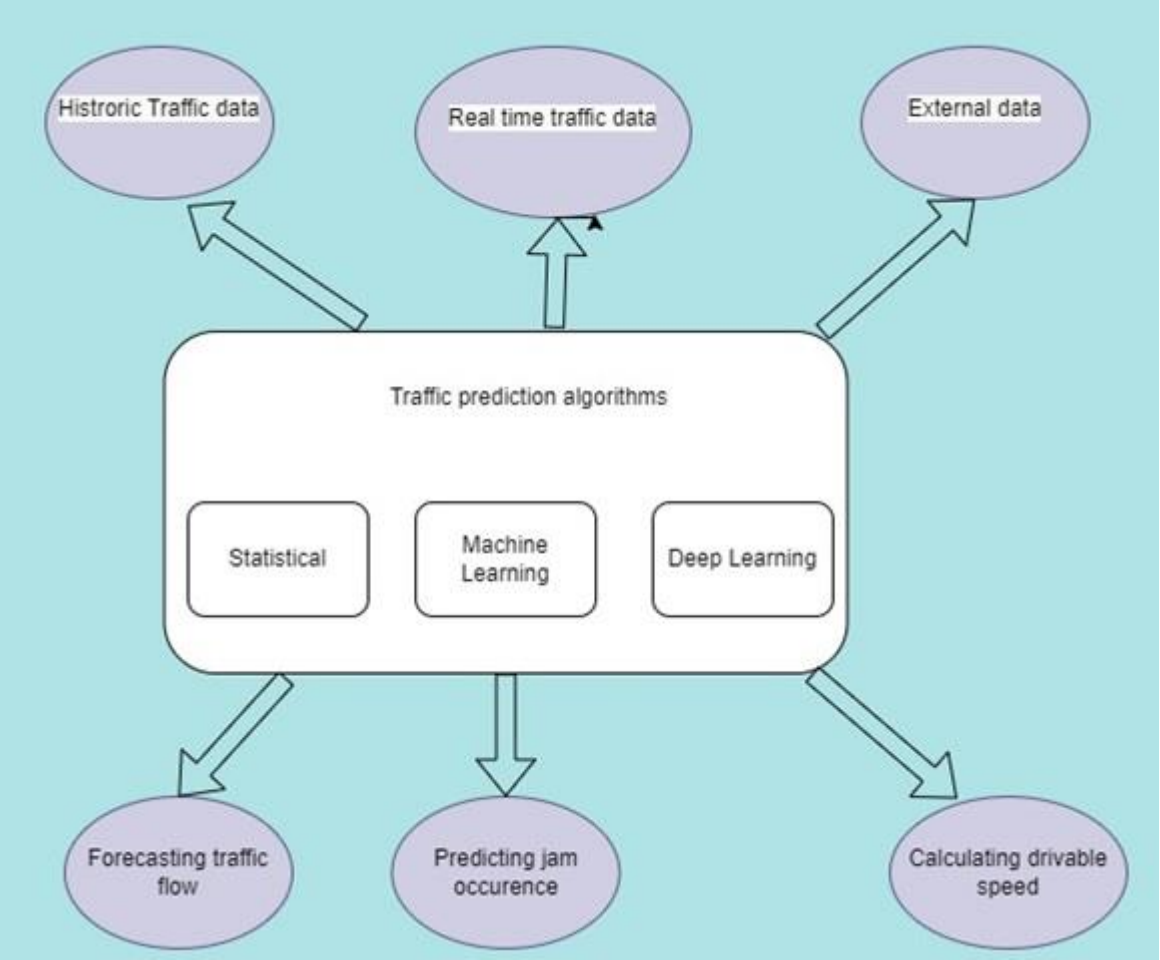

Fig1: strategies for traffic forecasting

In order to get reliable predictions, the Random Forest algorithm builds many decision trees and then blends their data. Assuming there is enough training data, it is quick and accurate. When making predictions, the K-Nearest Neighbors (KNN) algorithm looks for features that are similar to one another. In terms of short-term traffic flow prediction, experiments using the KNN model showed an accuracy of above 90%.

Machine learning methods such as logistic regression, support vector machines, and random forest approaches were used in this study to analyze normalized traffic flow volume. The following metrics were determined by using the aforementioned methods: accuracy, mean absolute error, mean squared error, and R-squared error. Logistic regression outperformed support vector machines and random forest in terms of accuracy and error rate.

The dataset on traffic flow volume has nine different characteristics. The attributes include date, time, traffic volume, holiday, temperature, rain, snow, clouds overall, weather primary, and weather description. Traffic volume is one of the nine factors that may be used to forecast traffic flow in various scenarios. There is a training set and a testing set inside the dataset. The data used for training is 80% while the data used for testing is 20%. The collected data set underwent data pre-processing prior to being fed into the classification algorithms. To make the model work better, we standardized the data to a range of 0 to 1.

2.1 Logistic Regression:

Among the many Machine Learning algorithms that fall under the umbrella of Supervised Learning, logistic regression stands out. With this method, one may utilize a group of independent factors to make predictions about a categorical dependent variable. The logistic regression model is used to forecast the value of a dependent variable that is categorical. It provides probabilistic values that fall between 0 and 1, rather than the precise values of yes or no, true or false, etc. The main difference between Logistic Regression and Linear Regression is their

application. while dealing with regression issues, one should use logistic regression, and while dealing with classification difficulties, one should use linear regression. We fit a "S" shaped logistic function in logistic regression instead of a regression line. This function predicts two maximum values, 0 and 1, instead of one. Logistic Regression can categorize new data using both continuous and discrete datasets and offer probabilities, making it a key machine learning approach. Using various kinds of data, Logistic Regression can sort observations into categories and quickly find the best variables to utilize for classification. You can see the logistic function in the picture below:

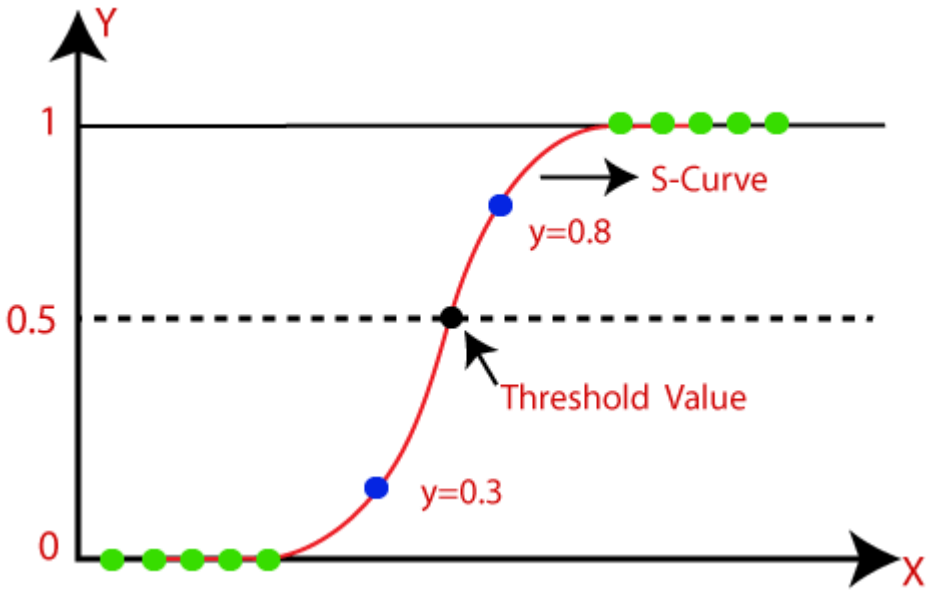

Fig2: Representation of Logistic Regression

To convert the anticipated values into probabilities, we employ the sigmoid function. It converts any number between 0 and 1 into another real number. Since the logistic regression result can't be greater than 1, it takes the shape of a "S" curve. One name for the S-shaped curve is the logistic or Sigmoid function. The idea of a threshold value, which specifies the likelihood of a value of 0 or 1, is used in logistic regression. For example, numbers north of the threshold tend to be 1, whereas values south of the threshold likely to be 0. If you know the equation for linear regression, you may use it to get the logistic regression equation. Here are the mathematical procedures to acquire the equations for Logistic Regression:

General equation of the straight line is

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$$\frac{y}{1-y}\text{; 0 for y= 0, and infinity for y=1}$$

But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

Prior to using the classification model, the dataset is subjected to data pre-processing. In order to improve a model's performance, the data was first normalized between 0 and 1. This was one of the primary data pre-processing procedures. There are a few different approaches to dealing with missing values: either completely disregarding the value, replacing it with a numeric value, using the mode value, or using the mean value of the

property. Analyzing the Training Data Set using Logistic Regression: A logistic regression model is fitted to the data set used for training purposes. Sklearn model selection includes a train_test_split function that divides data arrays into training and testing subsets, respectively. The two subsets will be randomly divided by Sklearn train test split by default. Training uses 90% of the data in the proposed study, while testing uses the remaining 10%.

```
import pandas as pd
df = pd.read_csv("C:\\Users\\kbvsatyaprakash\\Downloads\\Traffic.csv",
usecols=['Date', 'CarCount', 'BikeCount', 'BusCount', 'TruckCount', 'Total']);
df2 = scaler.transform(df)
x = df2[:, :-1]
y = df2[:, :-1]
```

Fig 3. Data Preprocessing stage

The first collection of data used to educate machine learning models is known as training data or a training dataset. In order to train machine learning algorithms to do a certain job or generate predictions, they are given training datasets. The testing data, on the other hand, is a subset of the training dataset that is used to evaluate the final model fit in an objective manner. The purpose of the testing data is to provide an objective assessment of how well the final model fits the training dataset. Analyzing the Training Data Set using Logistic Regression: A logistic regression model is fitted to the data set used for training purposes. Sklearn model selection includes a train_test_split function that divides data arrays into training and testing subsets, respectively. The two subsets will be randomly divided by Sklearn train test split by default.

```
from sklearn.model_selection import train_test_split
from sklearn import linear_model
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaler.fit(df)
df2 = scaler.transform(df)
x = df2[:, :-1]
y = df2[:, :-1]
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.1,
random_state=1)
reg = linear_model.LinearRegression()
reg.fit(X_train, y_train)
y_predict = reg.predict(X_test)
```

Fig 4. Predicting Results

Training uses 90% of the data in the proposed study, while testing uses the remaining 10%. The first collection of data used to educate machine learning models is known as training data or a training dataset. In order to train machine learning algorithms to do a certain job or generate predictions, they are given training datasets. The testing data, on the other hand, is a subset of the training dataset that is used to evaluate the final model fit in an objective manner. The purpose of the testing data is to provide an objective assessment of how well the final model fits the training dataset. Fog, rain, and snow impair driver sight, vehicle movement, and road capacity. In an Intelligent Transportation System, traffic operation and management need accurate prediction of macroscopic traffic stream factors like speed and flow. Because the traffic stream is non-linear and complicated, predicting these factors is difficult. Because layer-wise design removes buried abstract representation, deep learning models predict traffic stream factors better than shallow learning models. Traffic is affected by weather conditions due to hidden factors. Terrain limits prevent the weather station-road distance from immediately affecting traffic during rain. Waterlogging results from persistent rainfall weakening the drainage system and soil absorption. To capture the geographical and extended influence of weather, we presented a soft spatial and temporal threshold

mechanism. To fill meteorological data gaps, spatial interpolation is performed. We created CNN-LSTM and LSTM-LSTM hybrid deep learning models. Former model in hybrid model extracts spatio-temporal characteristics, which later model employs as memory. The latter model predicts traffic stream characteristics using features and temporal input. Multiple trials verify the deep learning model's performance. The trials reveal that a deep learning model trained with traffic and rainfall data predicts better than one without. The LSTM-LSTM model extracts long-term traffic-weather dependence better than previous models.

2.2 Support Vector Machine (SVM):

For both classification and regression tasks, one of the most common supervised learning techniques is support vector machine, or SVM. Nevertheless, its main use is in Machine Learning classification tasks. In order to make it easier to assign fresh data points to the proper category in the future, the SVM method seeks to find the optimal line or decision boundary that may partition n-dimensional space into classes. The term for this optimal decision boundary is a hyperplane. In order to construct the hyperplane, SVM selects the outlying points and vectors. A Support Vector Machine method is named after these out-of-the-ordinary instances, which are known as support vectors. Take a look at the graphic below; a decision boundary or hyperplane divides the image into two distinct groups. To divide data arrays into training and testing subsets, Sklearn's train_test_split function is used. Sklearn train test split will automatically divide the data into two sections at random.Training uses 90% of the data in the proposed study, while testing uses the remaining 10%. The first collection of data used to educate machine learning models is known as training data or a training dataset. In order to train machine learning algorithms to do a certain job or generate predictions, they are given training datasets

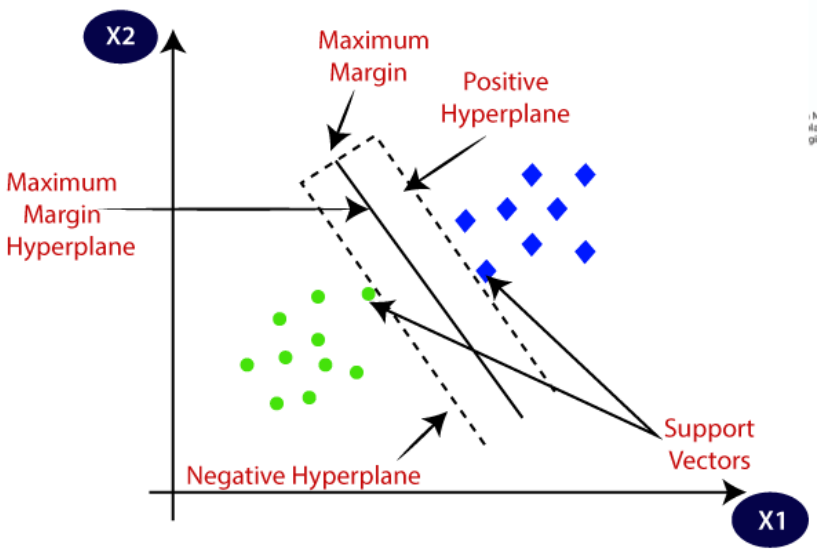

Fig 5: Support Vector Machines

The testing data, on the other hand, is a subset of the training dataset that is used to evaluate the final model fit in an objective manner.

```
import pandas as pd
df = pd.read_csv("C:\\Users\\kbvsatyaprakash\\Downloads\\Traffic.csv",
usecols=['Date', 'CarCount', 'BikeCount', 'BusCount', 'TruckCount', 'Total']);
df2 = scaler.transform(df)
x = df2[:, :-1]
y = df2[:, :-1]
```

Fig 6: Data Preprocessing

```
from sklearn.metrics import mean_absolute_error, mean_squared_erro
import matplotlib.pyplot as plt
mae = mean_absolute_error(y_test, y_predict)
mse = mean_squared_error(y_test, y_predict)
r2 = r2_score(y_test, y_predict)
print('SVR Mean Absolute Error: {}'.format(mae))
print('SVR Mean Squared Error: {}'.format(mse))
print('SVR R-squared: {}'.format(r2))
plt.scatter(y_test, y_predict)
plt.xlabel('True Values')
plt.ylabel('Predictions')
plt.show()
```

Fig 7: Predicting results

2.3 Random Forest Algorithm:

An effective and flexible supervised machine learning algorithm, Random Forest grows and combines multiple decision trees to form a "forest." It excels at both classification and regression problems. "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Rather than depending on a single decision tree, Random Forest takes the predictions from each tree and predicts the final output based on the majority votes of predictions. With more trees in the forest, accuracy increases and the problem of overfitting is prevented. For a more precise forecast, Random Forest develops a network of decision trees and then merges them.

Multiple uncorrelated models (the separate decision trees) perform much better when combined, according to the reasoning underlying the Random Forest model. A "vote" or classification is cast by each tree in a Random Forest classification. Based on the number of "votes," the forest selects the categorization that has the most support. Random Forest takes an average of all the trees' outputs while doing regression. The most important thing to note is that the decision trees that compose the Random Forest model have very little association with one another. Although some decision trees may make mistakes, the aggregate will be more accurate, leading to a positive result. Random forest's inherent capacity to rectify decision trees' propensity for overfitting to their training set is its most appealing feature. Overfitting causes erroneous results; fortunately, this algorithm's execution using the bagging approach and random feature selection almost entirely eliminates this issue. In addition, Random Forest often maintains its accuracy even when some data is absent. The "bagging" approach is often used to train decision trees in ensembles, similar to how Random Forest trees are taught. Machine learning ensemble algorithms like Bootstrap Aggregation are what the "bagging" process is all about. The goal of an ensemble technique is to improve prediction accuracy above that of a single machine learning algorithm by combining the results of many such algorithms. A second example of an ensemble approach is Random Forest.

```
import pandas as pd
df = pd.read_csv("C:\\Users\\kbvsatyaprakash\\Downloads\\Traffic.csv",
usecols=['Date', 'CarCount', 'BikeCount', 'BusCount', 'TruckCount', 'Total']);
df2 = scaler.transform(df)
x = df2[:, :-1]
y = df2[:, :-1]
```

Fig 8: Data Preprocessing

In order to generate sample datasets for each model, Bootstrap randomly executes feature sampling and row sampling from the dataset. By combining them, aggregation lowers these sample datasets to observation-based summary statistics. Using Bootstrap Aggregation, algorithms with a large variance, such decision trees, may have their variance reduced. When training on a dataset that is very sensitive to even little changes, the ensuing mistake is known as variance. When the variance is high, the algorithm will mistakenly model the dataset's noise for the desired signal. Overfitting is the name given to this issue. A model that has been overfitted will do well during training but will fail to differentiate between signal and noise when tested. Bagging is a high variance machine learning algorithm's implementation of the bootstrap approach.

```
from sklearn.metrics import mean_absolute_error, mean_squared_error,
import matplotlib.pyplot as plt
mae = mean_absolute_error(y_test, y_predict)
mse = mean_squared_error(y_test, y_predict)
r2 = r2_score(y_test, y_predict)
print('Random Forest Mean Absolute Error: {}'.format(mae))
print('Random Forest Mean Squared Error: {}'.format(mse))
print('Random Forest R-squared: {}'.format(r2))
plt.scatter(y_test, y_predict)
plt.xlabel('True Values')
plt.ylabel('Predictions')
plt.show()
```

Fig 9: Predicting Results

2.4 Naïve Bayes's Classifier

A technique for tackling classification issues, the Naïve Bayes algorithm is a kind of supervised learning algorithm that relies on Bayes theorem. Its primary use is in text categorization using a high-dimensional training set. When it comes to constructing rapid machine learning models that can generate quick predictions, one of the most successful and easy classification algorithms is the Naïve Bayes Classifier. It makes predictions based on the likelihood of an item, since it is a probabilistic classifier. The Naïve Bayes Algorithm is widely used for tasks such as spam filtering, sentiment analysis, and article classification. Simply put, the Naïve Bayes algorithm is made up of the terms "Naïve" and "Bayes." Plainly innocent: The reason it is referred to as Naïve is because it presumes that a feature's presence is unrelated to the occurrence of other characteristics. As an example, if we were to identify fruits based on their color, shape, and flavor, we would know that a red, round, and delicious fruit is an apple. So, everything about it tells you that it's an apple, and none of the features rely on each other.

```
import pandas as pd
df = pd.read_csv("C:\\Users\\kbvsatyaprakash\\Downloads\\Traffic.csv",
usecols=['Date', 'CarCount', 'BikeCount', 'BusCount', 'TruckCount', 'Total']);
df2 = scaler.transform(df)
x = df2[:, :-1]
y = df2[:, :-1]
```

Fig 10: Data Preprocessing

```
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
import matplotlib.pyplot as plt
r2 = r2_score(y_test, y_predict)
mae = mean_absolute_error(y_test, y_predict)
mse = mean_squared_error(y_test, y_predict)
accuracy = nb_classifier.score(X_test, y_test)
print('Naive Bayes Mean Absolute Error: {}'.format(mae))
print('Naive Bayes Mean Squared Error: {}'.format(mse))
print('Naive Bayes R-squared: {}'.format(r2))
plt.scatter(y_test, y_predict)
plt.xlabel('True Values')
plt.ylabel('Predictions')
plt.show()
```

Fig 11: predicting results

3. Results and Discussion

The results are obtained by applying Logistic Regression, Support Vector Machines, Random Forest and Naïve Bayes classification algorithms are applied on the traffic data set and the following plots are obtained from visual studio code.

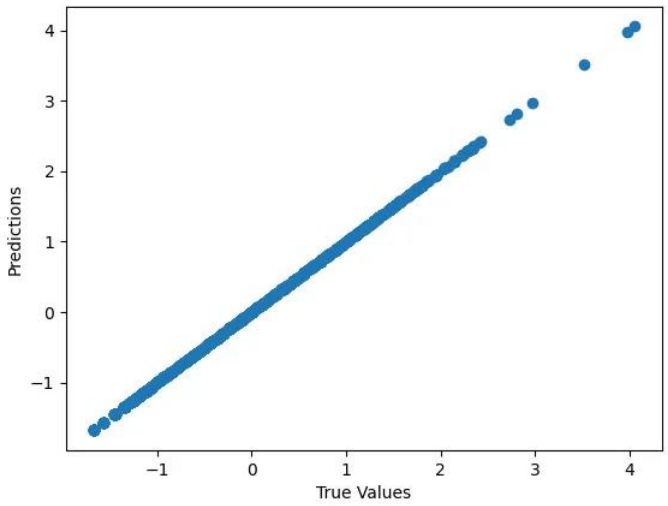

Fig 12: Predicted vs Actual

Mean Absolute Error (MAE) is a measure of how close the predictions from a model are to the actual values. It is calculated by averaging the absolute differences between the predicted values and the actual values. MAE is particularly useful when the dataset contains outliers that can heavily influence the accuracy of a model. The lower the MAE, the better the model's predictions. In figure, it can be seen that the MAE is very very low value (7.9028e-16) that shows the model is very significant as there is very low margin for error.

The Mean Squared Error (MSE) is a statistical metric for gauging the degree of inaccuracy in a model. Determine the average squared deviation from the expected value to the actual value. When there are no mistakes in the model, the MSE equals 0. Its worth grows in proportion to the magnitude of the model mistake. The values of MSE for the Logistic regression is 1.1775e-30 which is very value also signifies that the error values are very low and the model is significant. And finally, the value of Regression is 0.9999997 that clearly shows the predicted data is aggrable with the experimented data and the model is significant.

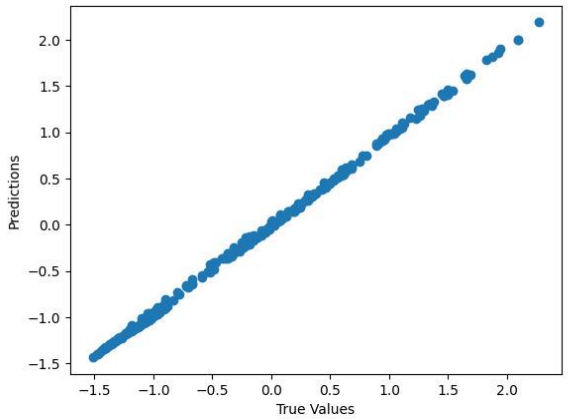

Fig 13: Predicted vs Actual

Support Vector Machines has a mean squared error (MSE) of 0.00205, and MAE of 0.03834 which is extremely low and indicates a statistically significant model. Lastly, the regression value of 0.99772 indicates that the model is statistically significant and that the projected data agrees with the experimental data. But the values are little higher than Logistic regression values.

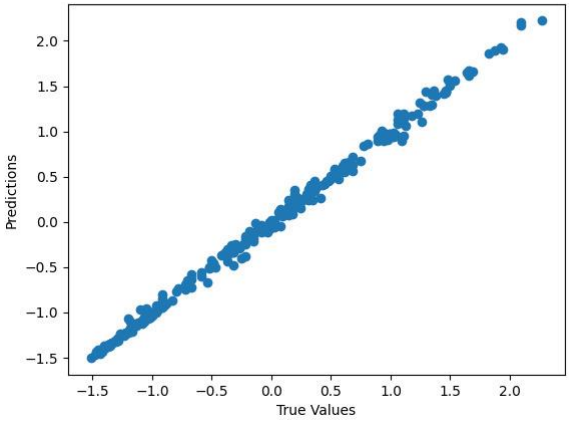

Fig 14: Predicted vs Actual

0.00271 is the mean Squared error (MSE) value and 0.03674 is the Mean Absolute Error for Random Forests, which is very low and indicates a significant model. In conclusion, the model is statistically significant, since the Regression value of 0.99 indicates that the predicted data agrees with the experimental data.

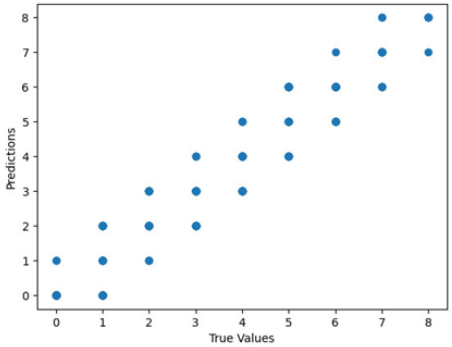

Fig 15: Predicted vs Actual

Naïve Bayes Classifier has an MSE of 0.402, and MAE of 0.582 which is extremely lower and indicates a significant model with very small error values but it is higher when compared to other machine learning methods. At last, the regression value of 0.9201 demonstrates that the model is statistically significant and that the projected data agrees with the experimental data.

4. Conclusions

National Highways Authorities will be able to better accommodate the increasing traffic population with the use of a 98.85% accurate traffic flow forecast model, which will allow them to improve and enlarge existing highways as well as offer other routes. Predictions of traffic flows will show where delays are most likely to happen, allowing for the creation of alternate routes or other solutions tailored to specific areas. Concerns such as property acquisition, compensation, and the construction of bypass routes or widened roadways are involved. The model outperformed the approaches described in the literature, according to the experimental data. People feel safer and more trusting after using this strategy. The proposed models are significant as the MSE, MAE and Regression values are above the acceptable threshold values. Though all (Logistic Regression, Support vector machines, Random Forests and Naïve-Bayes Classifiers) machine learning methods are significant, the particular data set shows better predictions with the Logistic Regression method with its least MSE, MAE and higher R value, which is a significant finding in the project.

References

1. K. K. Billa, M. Deb, G. R. K. Sastry, and S. Dey, “Experimental investigation on dispersing graphene-oxide in biodiesel/diesel/ higher alcohol blends on diesel engine using response surface methodology,” Environ. Technol. (United Kingdom), vol. 0, no. 0, pp. 1–18, 2021, doi: 10.1080/09593330.2021.1916091.
2. C. G. Rajulu, A. G. Krishna, and T. Babu Rao, “An integrated evolutionary approach for simultaneous optimization of laser weld bead characteristics,” Proc. Inst. Mech. Eng. Part B J. Eng. Manuf., vol. 232, no. 8, pp. 1407–1422, 2018, doi: 10.1177/0954405416667431.
3. N. Yilmaz, E. Ileri, A. Atmanli, A. D. Karaoglan, U. Okkan, and M. S. Kocak, “Predicting the Engine Performance and Exhaust Emissions of a Diesel Engine Fueled with Hazelnut Oil Methyl Ester: The Performance Comparison of Response Surface Methodology and LSSVM,” J. Energy Resour. Technol. Trans. ASME, vol. 138, no. 5, pp. 1–7, 2016, doi: 10.1115/1.4032941.
4. B. Ashok, K. Nanthagopal, V. Anand, K. M. Aravind, A. K. Jeevanantham, and S. Balusamy, “Effects of n-octanol as a fuel blend with biodiesel on diesel engine characteristics,” Fuel, vol. 235, no. July 2018, pp. 363–373, 2019, doi: 10.1016/j.fuel.2018.07.126.
5. Y. Jiotode and A. K. Agarwal, “Endoscopic combustion characterization of Jatropha biodiesel in a compression ignition engine,” Energy, vol. 119, pp. 845–851, 2017, doi: 10.1016/j.energy.2016.11.056.
6. K. K. Billa, G. R. K. Sastry, and M. Deb, “Optimization of next-generation alcohols and fishoil methyl ester blends in a single cylinder DI-CI engine using response surface methodology,” Int. J. Eng. Adv. Technol., vol. 9, no. 1, pp. 631–645, 2019, doi: 10.35940/ijeat.F8746.109119.
7. I. E. Agency, “Global EV Outlook 2021,” 2021. [Online]. Available: www.iea.org/t&c/.
8. K. K. Billa, G. R. K. Sastry, and M. Deb, “ANFIS Model for Prediction of Performance-Emission Paradigm of a DICI Engine Fueled with the Blends of Fish Oil Methyl Ester , n-Pentanol and Diesel,” vol. 17, no. 1, pp. 115–133, 2020.
9. S. Javed, Y. V. V Satyanarayana Murthy, R. U. Baig, and D. Prasada Rao, “Development of ANN model for prediction of performance and emission characteristics of hydrogen dual fueled diesel engine with Jatropha Methyl Ester biodiesel blends,” J. Nat. Gas Sci. Eng., vol. 26, pp. 549–557, 2015, doi: 10.1016/j.jngse.2015.06.041.
10. G. Ministry of Petrolium & Natural Gas, “Energiging India’s progress,” Ministry of Petrolium & Natural Gas.
11. C. N. Huang and H. T. Shen, “Maximum hydrogen production by using a gasifier based on the adaptive control design,” Int. J. Hydrogen Energy, vol. 44, no. 48, pp. 26248–26260, 2019, doi: 10.1016/j.ijhydene.2019.08.087.

12. Siroos Shahriari, Milad Ghasri, S. A. Sisson & Taha Rashidi (2020) Ensemble of ARIMA: combining parametric and bootstrapping technique for traffic flow prediction.
13. S. Priya, D. Singh, H. Sharma and J. Yadav, "Prediction of Traffic Flow Propagation Using Machine Learning Algorithms," 2022
14. Mao and S. Shi, "Research on Method of the Subsection Learning of Double-Layers BP Neural Network in Prediction of Traffic Volume," 2009.
15. Nazirkar Reshma Ramchandra, C. Rajabhushanam "Machine learning algorithms performance evaluation in traffic flow prediction" 2021.
16. Ge Zheng, Wei Koong Chai, Vasilis Katos "Hybrid deep learning models for traffic prediction in large-scale road networks".2023.
17. Y. Gao, C. Zhou, J. Rong, Y. Wang and S. Liu, "Short-Term Traffic Speed Forecasting Using a Deep Learning Method Based on Multi temporal Traffic Flow Volume," in IEEE Access, vol. 10, pp. 82384-82395, 2022.
18. N. K. Chikka Krishna, P. Racha konda and T. Tallam, "Short - Term Traffic Prediction Using Fb-PROPHET and Neural-PROPHET," 2022.
19. Wanqiu Lou, Yingjie Zhou, Peng Sheng and Junfeng Wang, "An improved least square support vector regression algorithm for traffic flow forecasting" 2014.
20. Fahad Aljuaydi, Benchawan Wiwatanapataphee, Yong Hong Wu, "Multivariate machine learning-based prediction models of freeway traffic flow under non-recurrent events" Alexandria Engineering Journal, Volume 65, 2023.
21. Zihao Wang, Vrizlynn L. L. Thing "Feature mining for encrypted malicious traffic detection with deep learning and other machine learning algorithms", 2023.
22. Mohammad Shareef Ghanim,[a,*] Deepti Muley,[b] and Mohamed Kharbeche[c] "ANN-Based traffic volume prediction models in response to COVID-19 imposed measures", 2022.
23. Yan Zheng Chunjiao Dong,* Daiyue Dong and Shengyou Wang "Traffic Volume Prediction: A Fusion Deep Learning Model Considering Spatial–Temporal Correlation".2021
24. Boris Medina-Salgado[a,b] , Eddy Sánchez-Dela Cruz [a]∗ , Pilar Pozos-Parra[c] , Javier E. Sierra[b] "Urban traffic flow prediction techniques: A review".2023
25. Archana Nigam, Sanjay Srivastava "Hybrid deep learning models for traffic stream variables prediction during rainfall".2023
26. Vladimir Shepeleva *, Ivan Slobodina , Zlata Almetovaa , Dmitry Nevolinb , Andrei Shvecova "A Hybrid Traffic Forecasting Model for Urban Environments Based on Convolutional and Recurrent Neural Networks", 2023.
27. Anand Paul, Naveen Chilamkurti, Alfred Daniel, Seungmin Rho Intelligent Vehicular Networks and Communications, Fundamentals, Architectures and Solutions, 2017, pages 177-184 Book chapter-8, Big Data Collision analysis framework.