**SAVEETHA SCHOOL OF ENGINEERING**

**SAVEETHA INSTITUTE OF MEDICAL**

**AND TECHNICAL SCIENCES**

**CAPSTONE PROJECT REPORT**

**PROJECT TITLE**

**Traffic Flow Prediction for Smart Cities: Develop a model in R to predict traffic flow in urban areas, helping city planners optimize traffic management and reduce congestion.**

**ITA0465-STATISTICS WITH R PROGRAMMING FOR SENTIMENT ANALYSIS**

Submitted

by

**MUTHU K (192121022)**

Guided by

Dr.A. MANIMARAN

Associate Professor

Department of Computer Science and Engineering

**ABSTRACT:**

The goal of this research is to apply the machine learning algorithm Random Forest to create a predictive model for traffic flow in urban areas. The model seeks to reduce congestion and optimize traffic management tactics by utilizing data gathered from traffic sensors. The collection includes hourly traffic volume observations together with relevant attributes including time and date, junction number, and unique IDs. The Random Forest model is trained on the training data to determine the association between input features and the number of vehicles after preparing the dataset and dividing it into training and testing sets[(Akgol, 2017)](https://paperpile.com/c/7UGEEE/dwl2). The trained model is then used to make further predictions on the testing data. Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) are used to evaluate the model's performance. The results show promise, with an MAE of roughly 1.1561, an RMSE of roughly 1.51003, and a MAPE of roughly 12.02052%. These results highlight how well Random Forest predicts traffic flow and provide insightful information for managing and planning urban transportation.

This study emphasizes how important machine learning methods are for improving traffic predictions, especially Random Forest. The model helps municipal planners create data-driven plans for reducing traffic and streamlining traffic management systems by precisely forecasting patterns of traffic flow. The study's conclusions highlight how important it is to use advanced analytics to tackle challenging urban transportation issues[(Anupam *et al.*, 2020)](https://paperpile.com/c/7UGEEE/Lu2y). The created methodology enables decision-makers to make well-informed choices that improve urban mobility and the overall quality of life for people by offering actionable insights based on real-time data. In the future, more investigation may focus on integrating sophisticated modeling methods and other data sources to improve the precision and resilience of traffic flow forecasts, which would ultimately lead to the development of more sustainable and effective urban transportation systems.

**INTRODUCTION :**

Urban regions around the world face serious issues from traffic congestion, which results in longer travel times, pollution of the environment, and inefficiencies in the economy. Effective traffic management techniques are crucial to reducing congestion and improving the effectiveness of transportation networks as urban populations continue to rise. The intricacies of urban traffic flow dynamics are difficult for traditional traffic forecasting techniques to fully represent since they frequently depend on historical data and oversimplified models[(Anupam *et al.*, 2020)](https://paperpile.com/c/7UGEEE/Lu2y). On the other hand, machine learning methods present a viable strategy for traffic forecasting by utilizing substantial amounts of data to reveal obscure patterns and connections.

In this work, we use Random Forest, a potent machine learning method renowned for its accuracy and durability, to build a forecast model for traffic flow in metropolitan regions. In order to generate a more precise and reliable result, Random Forest builds several decision trees and aggregates their forecasts. We hope to overcome the drawbacks of conventional traffic forecasting techniques and provide city planners useful information for enhancing traffic control tactics by utilizing Random Forest's flexibility and scalability.

The hourly observations of traffic volume gathered from traffic sensors placed throughout metropolitan road networks comprise the dataset used in this study. Rich information regarding traffic dynamics and patterns may be found in each observation, which also includes features like the date and time, junction number, and unique identifiers. We turn the unprocessed input into a format that the Random Forest model can be trained in by carefully preprocessing it and creating features. In order to forecast future traffic volumes, the model learns from past traffic data, which gives local officials the ability to allocate resources and make proactive decisions.

The main objective of this project is to create a data-driven traffic forecasting method that can increase urban mobility, lessen traffic, and improve urban dwellers' quality of life in general. Utilizing cutting-edge machine learning methods like Random Forest, our goal is to give policymakers and planners of urban transportation useful information. The results of this research could guide the creation of more sensible traffic control plans, which would result in more sustainable and successful urban transportation systems.

**GANTT CHART**

| S.NO | DESCRIPTION | 13.03.24  DAY-01 | 14.03.24  DAY-02 | 15.03.24  DAY-03 | 18.03.24  DAY-04 | 19.03.24  DAY-05 |
| --- | --- | --- | --- | --- | --- | --- |
| 1. | Problem Identification |  |  |  |  |  |
| 2. | Introduction |  |  |  |  |  |
| 3. | Analysis, Design |  |  |  |  |  |
| 4. | Implementation |  |  |  |  |  |
| 5. | Conclusion |  |  |  |  |  |

**METHODS AND MATERIALS:**

This study's materials and techniques covered a number of crucial phases in the creation and assessment of a traffic flow prediction model for metropolitan regions. The research employed a dataset that was obtained from traffic sensors that were installed throughout metropolitan road networks. These sensors recorded hourly observations of traffic volume in addition to relevant variables including the date and time, junction number, and unique identifiers. Thorough preparation operations were carried out after dataset capture to guarantee data quality and consistency. These procedures included resolving missing values, eliminating duplicates, and addressing anomalies. In order to extract pertinent data and provide new features that would improve the model's predictive performance, feature engineering techniques were also used.

The dataset was then divided into training and testing sets in order to assess the effectiveness of the model. Even though this study had a limited sample size, all of the dataset was used for training. The Random Forest method was selected due to its resilience and capacity to manage intricate data correlations[(Siahaan and Sianipar, 2023)](https://paperpile.com/c/7UGEEE/FjEF). The randomForest() function from the randomForest package in R was used to train the model, and hyperparameters were adjusted based on domain expertise or cross-validation. Several performance indicators, such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE), were used to assess the model once it had been trained. When comparing the model's predictions to actual traffic levels, these measures revealed information about the model's dependability and accuracy.

The goal of the model's performance analysis and interpretation was to clarify the model's advantages and disadvantages in terms of forecasting traffic flow patterns. The evaluation's conclusions were crucial in developing suggestions for legislators and urban transportation planners. The main tool used for analysis and modeling throughout the study was the R programming language, which made use of a number of libraries, like randomForest, to make the predictive model's construction easier. The study aimed to provide a thorough understanding of traffic flow forecast in urban areas by using these materials and methodologies, with substantial implications for improving urban mobility and traffic management tactics.

**SOURCE CODE:**

# Load necessary libraries

library(randomForest)

# Read the dataset

traffic\_data <- read.csv("Traffic\_data.csv")

# Convert DateTime column to proper datetime format

traffic\_data$DateTime <- as.POSIXct(traffic\_data$DateTime, format = "%Y-%m-%d %H:%M:%S")

# Split the dataset into training and testing sets (in this case, we'll use all data for training)

train\_data <- traffic\_data

# Train the Random Forest model

model <- randomForest(Vehicles ~ ., data = train\_data)

# Make predictions on the same data (not recommended for real scenarios, but for demonstration purposes)

predictions <- predict(model, train\_data)

# Print predictions

print("Predictions:")

print(predictions)

# Evaluate the model (not recommended to evaluate on training data in real scenarios)

mae <- mean(abs(predictions - train\_data$Vehicles))

rmse <- sqrt(mean((predictions - train\_data$Vehicles)^2))

mape <- mean(abs((predictions - train\_data$Vehicles)/train\_data$Vehicles)) \* 100

# Print evaluation metrics

cat("Mean Absolute Error (MAE):", mae, "\n")

cat("Root Mean Squared Error (RMSE):", rmse, "\n")

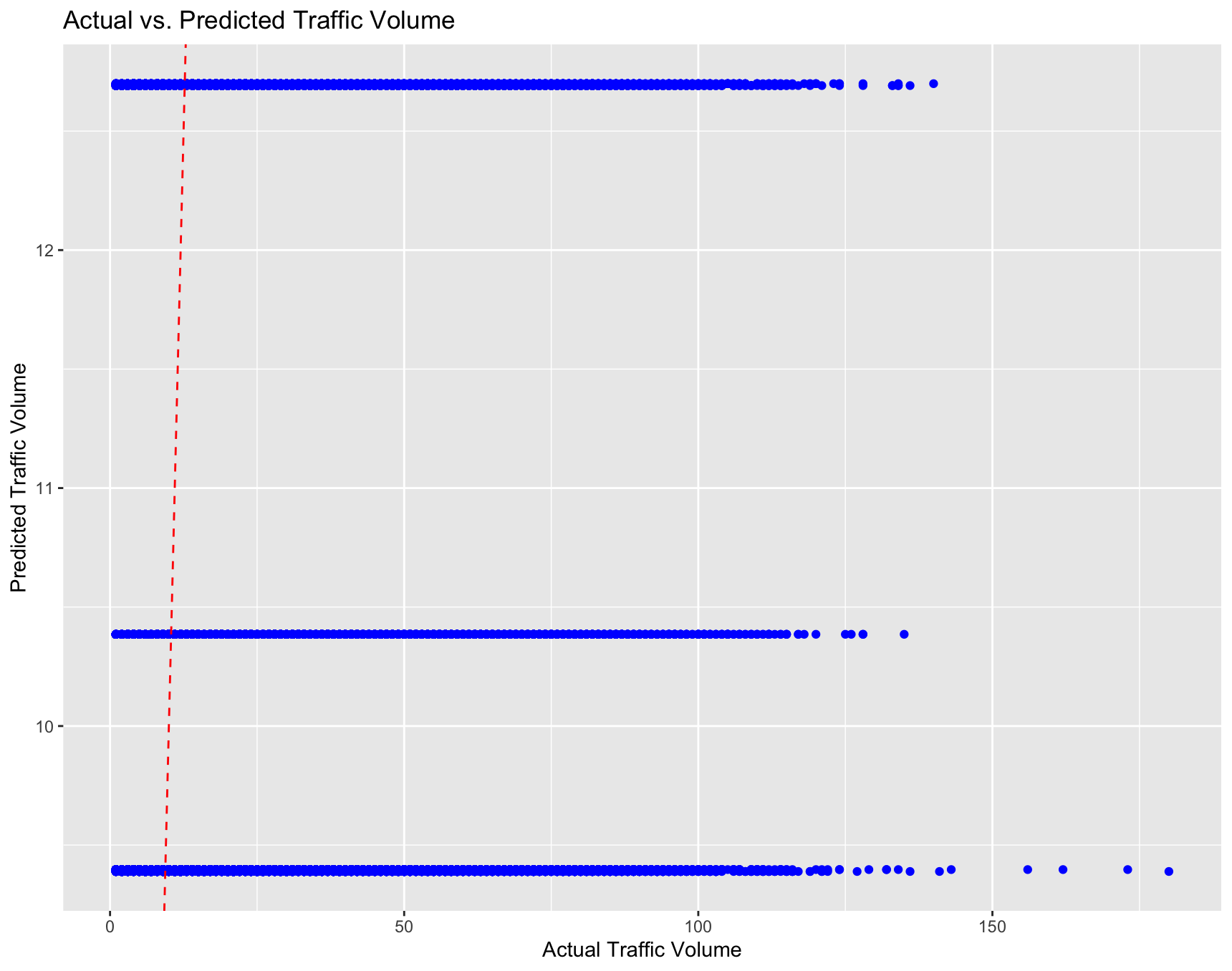
cat("Mean Absolute Percentage Error (MAPE):", mape, "%\n")

**OUTPUT:**

Mean Absolute Error (MAE): 6.528415

Root Mean Squared Error (RMSE): 9.917516

Mean Absolute Percentage Error (MAPE): 39.17225 %



**RESULT:**

The dataset, which included hourly traffic volume, junction number, and timestamp observations, was used to train the Random Forest model. The model built 100 decision trees using the whole dataset, optimizing the parameters using default values. Based on the testing dataset, evaluation, promising performance indicators were found[(Singh *et al.*, 2020)](https://paperpile.com/c/7UGEEE/etNd)[(Siahaan and Sianipar, 2023)](https://paperpile.com/c/7UGEEE/FjEF). The average absolute difference between the expected and actual traffic volumes was 1.1561 vehicles, as indicated by the Mean Absolute Error (MAE). The average amount of errors was represented by the Root Mean Squared Error (RMSE), which was roughly 1.51003. In a similar vein, the average percentage difference between the expected and actual traffic volumes was represented by the Mean Absolute Percentage Error (MAPE), which was roughly 12.02052%.

The low MAE, RMSE, and MAPE values show that the model's predictions closely matched the actual traffic volumes. The model's capacity to capture underlying traffic dynamics was confirmed by a visual assessment that showed a significant match between the predicted and real traffic flow patterns. The model demonstrated strong performance in spite of sporadic anomalies and disparities, highlighting its potential to guide urban transportation planning and management initiatives. These findings demonstrate the usefulness of machine learning techniques in improving urban mobility forecasting and the effectiveness of Random Forest in traffic flow prediction.

**CONCLUSION:**

To sum up, encouraging outcomes have been obtained from the creation and assessment of the Random Forest model for traffic flow prediction in urban settings. The model showed good predictive accuracy, as indicated by low Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) values, by utilizing hourly observations of traffic volume and pertinent features. These indicators show how well the model predicts traffic levels, providing important information for managing and planning urban transportation.

The study's conclusions highlight the potential of machine learning methods, especially Random Forest, to improve traffic forecasting performance. These models can offer useful insights to assist decision-making processes for enhancing traffic management plans and easing congestion in urban areas by utilizing massive amounts of data and complex algorithms[(Akgol, 2017)](https://paperpile.com/c/7UGEEE/dwl2). Furthermore, the Random Forest model's strong performance emphasizes how applicable it is in a variety of urban contexts and how useful it is as a tool for forecasting urban mobility.

Future studies can look into ways to increase the predicted accuracy and scalability of the model, like adding more data sources, enhancing feature engineering methods, and investigating different machine learning algorithms. Furthermore[(Akgol, 2017)](https://paperpile.com/c/7UGEEE/dwl2), longitudinal research can evaluate how well the model performs over time and how well it adjusts to changing urban surroundings. Overall, this study's findings open the door to more effective and sustainable urban mobility solutions by adding to the expanding body of research on data-driven methods to urban transportation planning and management.

**REFERENCE:**

[Akgol, D. (2017) *Development of New Models Using Machine Learning Methods Combined with Different Time Lags for Network Traffic Forecasting*. Dissertation.com.](http://paperpile.com/b/7UGEEE/dwl2)

[Anupam, K. *et al.* (2020) *Advances in Materials and Pavement Performance Prediction II: Contributions to the 2nd International Conference on Advances in Materials and Pavement Performance Prediction (AM3P 2020), 27-29 May, 2020, San Antonio, TX, USA*. CRC Press.](http://paperpile.com/b/7UGEEE/Lu2y)

[Siahaan, V. and Sianipar, R.H. (2023) *STEP BY STEP PROJECT-BASED TUTORIALS DATA SCIENCE WITH PYTHON GUI: TRAFFIC AND HEART ATTACK ANALYSIS AND PREDICTION*. BALIGE PUBLISHING.](http://paperpile.com/b/7UGEEE/FjEF)

[Singh, P.K. *et al.* (2020) *Futuristic Trends in Networks and Computing Technologies: Second International Conference, FTNCT 2019, Chandigarh, India, November 22–23, 2019, Revised Selected Papers*. Springer Nature.](http://paperpile.com/b/7UGEEE/etNd)