CSCE 5380.001 Data Mining

Project Report Group– 04

Regression-Based Traffic Flow Prediction: A Multi-Model analysis.

Teammates:

Sushma Chowdary Aari –11713230

Vijaythanav Allena –11715799

Lekhitanand Bandi –11727419

Naga Lakshmi kedareswari Parchuri –11662335

Bharatwaz Lkkly Srinivasulu –11699441

Abstract:

The Mian objective of our project is to predict the traffic flow patterns using machine learning algorithms, which plays a crucial role in the effective traffic management system and a smart city planning. Traffic flow is influenced by a various factor such as time of days, weather conditions, road status, and urban events. Our focus is to analyses the intricate nature of traffic congestion, uncover its root causes, and suggest practical solutions. To achieve this type, we utilize several regression-based models in machine learning models like Random Forests, Gradient Boosting Machines, and Neural Networks, as they are well-suited for capturing and explaining the non-linear relationships to define the complex dynamics of traffic systems.

The practical outcomes of our project are very highly impactful, leading to noticeable developments in the urban traffic flow and reduce the strain on infrastructure caused by congestion. Ultimately, this research(study) also supports the advancement of more efficient and sustainable transportation systems, that contributing to broader objectives such as minimizing urban carbon emissions and increasing the overall quality of life in the cities. By combining robustness methodologies with in the real-world applications, our project not only tackles the present days situations urban issues but also it lays the foundation for the future innovations in traffic and transportation management.

Keywords: Traffic flow prediction, Machine learning algorithms, Urban mobility, Traffic congestion, Regression models, Urban planning, Traffic management, Non-linear relationships, Sustainable transportation systems, Data analysis.

Introduction:

With the more people moving to more cities and more infrastructure being built in those cities, traffic control has become much difficult now-a-days. At its core, traffic in cities is a complicated issue that involves planning spaces, people's actions, and the uncertainty of the environmental factors. Traffic control systems that also works well keep cities running, make sure people to get back into the work on time, and lower environmental impact of road travellers.

In the previous days, unchanged models based on historic trends and simple generalise were used to track and control the traffic flow patterns. But these ways don't work for all the problems that comes up in the cities today, where things changes quickly and differentiate. Vital models that can adapt to real-time data and make the correct projections are needed for today's traffic systems. These all models are the premise of intelligent transportation systems (ITS).

At this point, traffic prediction is both a diagnostic tool and a guideline for how to control the traffic and plan cities. Cities can stay away from problems before they happen by planning ahead

for traffic stages and using strategies like traffic signal control, minimal pricing on toll roads, and best routes for emergency services. There are many options, such as shorter travel times and less gas usage, good air quality, and less stress on infrastructure.

The area whereas our project can also be used is where data science and urban planning meet. We use the huge amounts of data that current traffic systems produce and the most up to the date machine learning algorithms to find trends and make best predictions. The ideas behind this project come from time series analysis, pattern recognition, and predictive modelling, which are all the concepts of machine learning. We look at the other work that which is connected to this one as well, like using deep learning to get features and using group of methods to make forecast more accurate.

The aim of this project is to help the smart cities become smarter by giving us a good thought of how traffic works. Our work is in line with the aim of maintainable urban growth and the search for the good quality of life for the city tenant. We want to make sure all that the roads and bridges in the cities to be well constructed and smooth as the need for transportation grows.

Background:

Now a days Traffic control has become more tough as more people move to cities and infrastructure in the cities grows to settle them. At its point, traffic in the cities is a difficulted issue that involves planning spaces, people's actions, and the unbalanced of the environmental factors. Traffic control systems that work well when we keep cities running, make sure the people get to work on time, and lower the environmental risk of road travellers.

Previously, static models based on historic trends and simple extrapolation were used to track and control traffic flow patterns. But these ways don't work for the issues that come up in cities today, where things change quickly and changed. Dynamic models that can adapt to real-time data and make correct forecasts are needed for today's traffic controls. These models are the premise of intelligent transportation systems (ITS).

At this point, traffic predictions is both a diagnostic tool and a guideline for how to control traffic and plan cities. Cities can avoid problems before they could happen by planning ahead for traffic stages and using strategies like adaptive traffic signal control, dynamic pricing on toll roads, and smart routes for emergency services. There are many perks, such as shorter travel times and less fuel use, better air quality, and less stress on infrastructure.

The area where our project can be used is where data science and urban planning meet. We use the huge amounts of data that modern traffic systems produce and the most up-to-date machine learning algorithms to find trends and make smart predictions. The ideas behind this project come from time series analysis, pattern recognition, and predictive modelling, which are all areas of

machine learning. We look at other work that is connected to this one as well, like using deep learning to get features and using ensemble methods to make predictions more accurate.

The aim of this project is to help the smart cities become smarter by giving us a good thought of how traffic works. Our work is in line with the aim of maintainable urban growth and the search for the good quality of life for the city tenant. We want to make sure all that the roads and bridges in the cities to be well constructed and smooth as the need for transportation grows.

Experiment Methodology:

Dataset:

For this project, we are going to use a dataset which contains the 48,120 number of records. Each entry corresponds to traffic data at a given time and junction.

- **Date Time**: The Date-Time attribute displays the date and time of data entry ,when the vehicle count was taken.
- **Junction**: The junction attribute describes about the traffic intersection where the vehicle count was considered.
- **Vehicles**: The vehicle attribute keeps track of the how many automobiles that exist and it provides a count measure of traffic volume at the certain junction and time, as well as info on traffic density and flow patterns.
- **ID**: The ID attribute can be described as a unique identifier number for each data point, and it is most likely to be made up of date and an extra integer to make sure that each timestamp is unique across the collection.

Firstly, we have taken all the useful modules which are required to train the dataset using different regressor models and then evaluate them. We have loaded our dataset on Google Colab and displayed the different columns of our dataset. We have then dropped the 'ID' column which is not required for the first analysis and extracted date features from the DateTime column which are used in our analysis. Now, we proceed to the Data Analysis. First, we have displayed bar plots to show the no.of vehicles at each junction for each particular year as shown below:

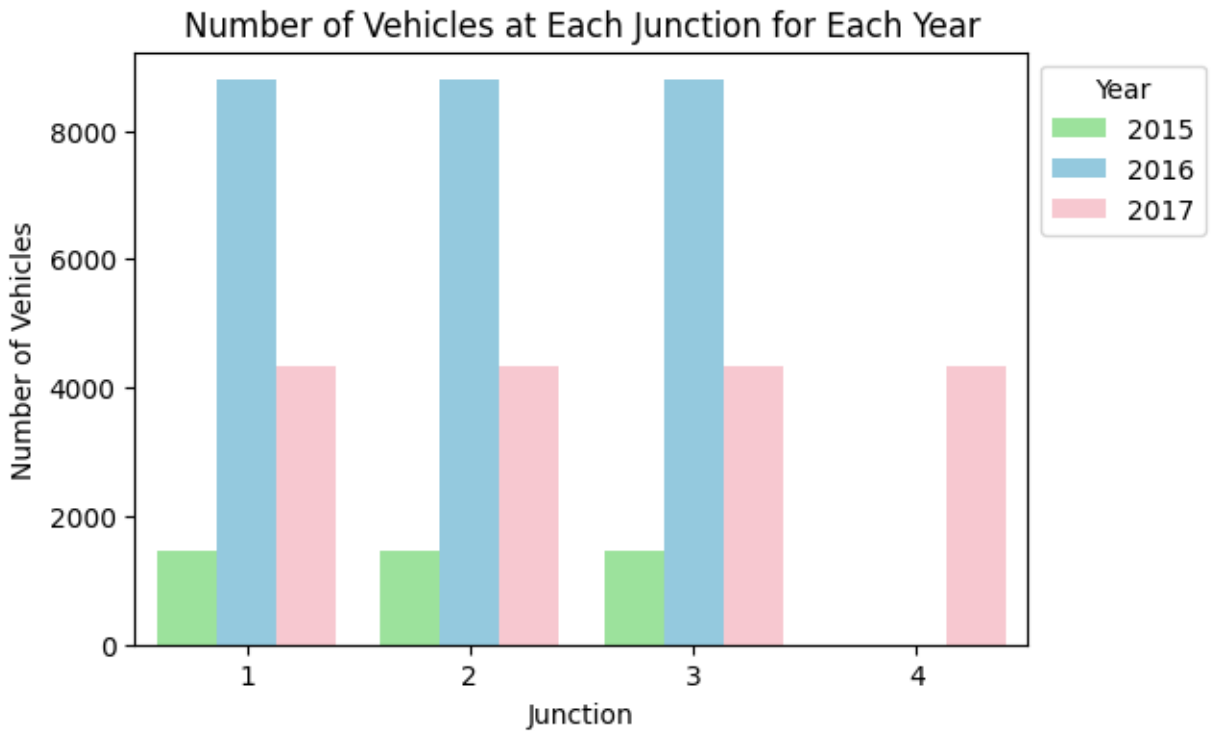

Fig 1 : Number of vehicles at each junction for each year

Next, we have shown line plots to show the relationship between the no.of vehicles at different times of a particular day and on different days at each of the junctions.

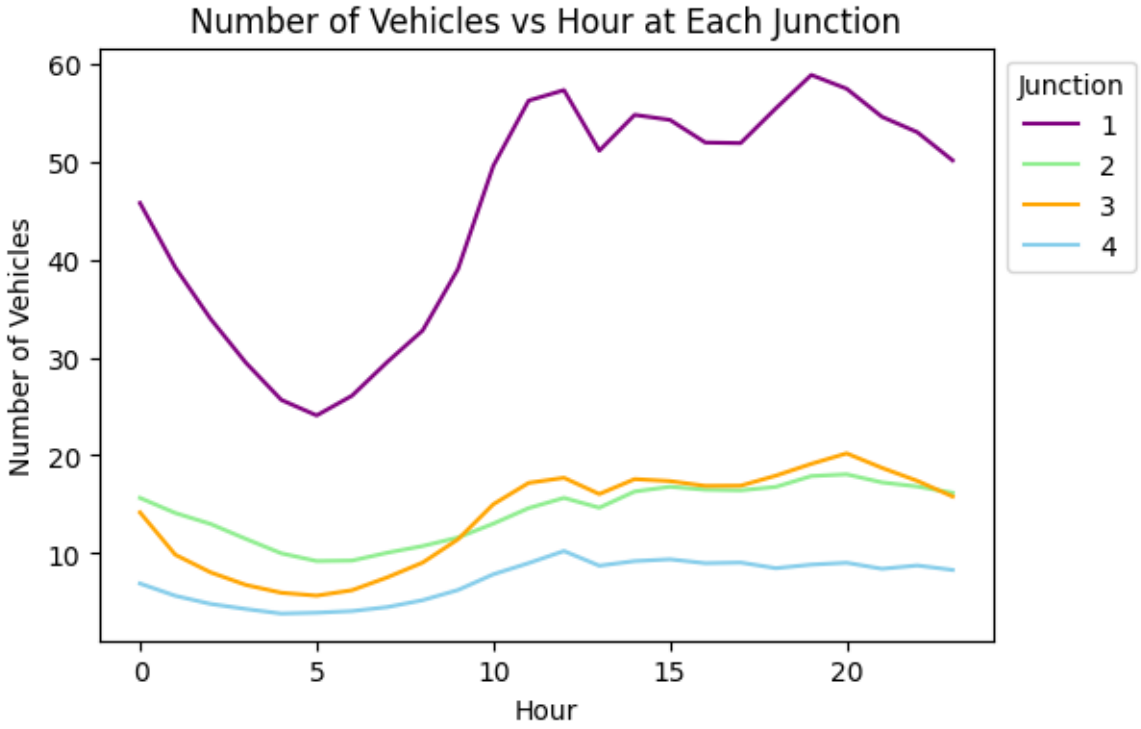

Fig 2 : Number of Vehicles vs Hour at each junction

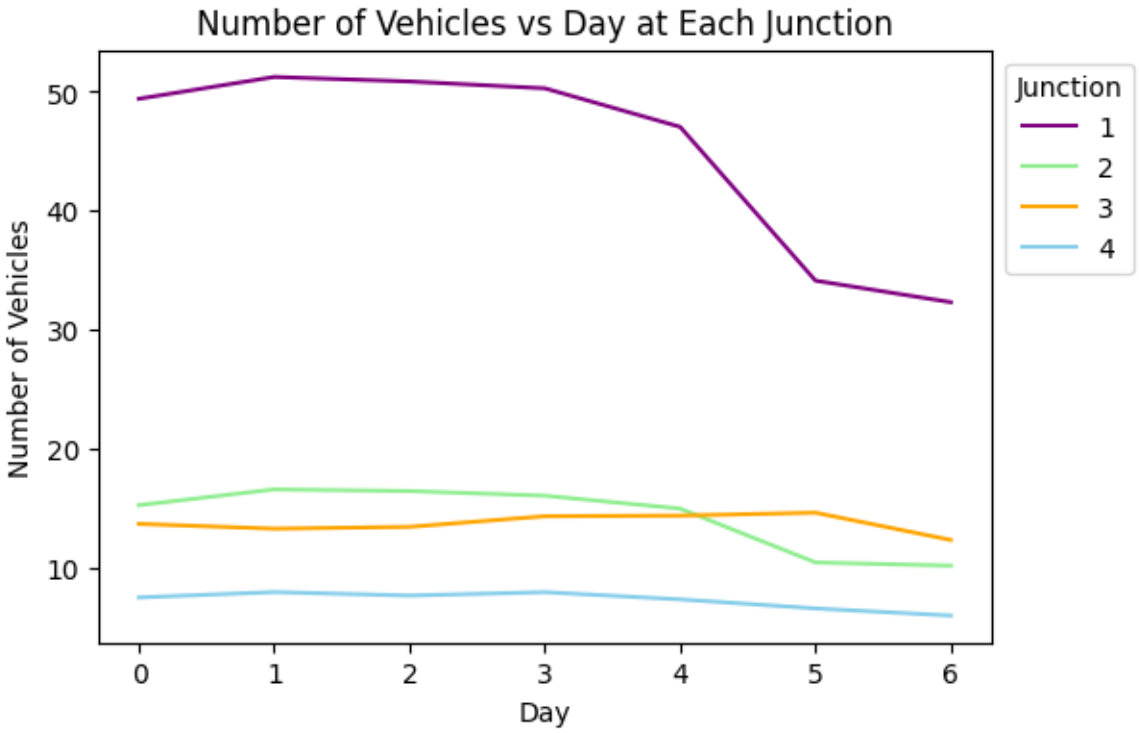

Fig 3 : Number of Vehicles vs Day at each junction

After that, we have displayed the distribution of traffic over time at each of the junctions using the line plots as below:

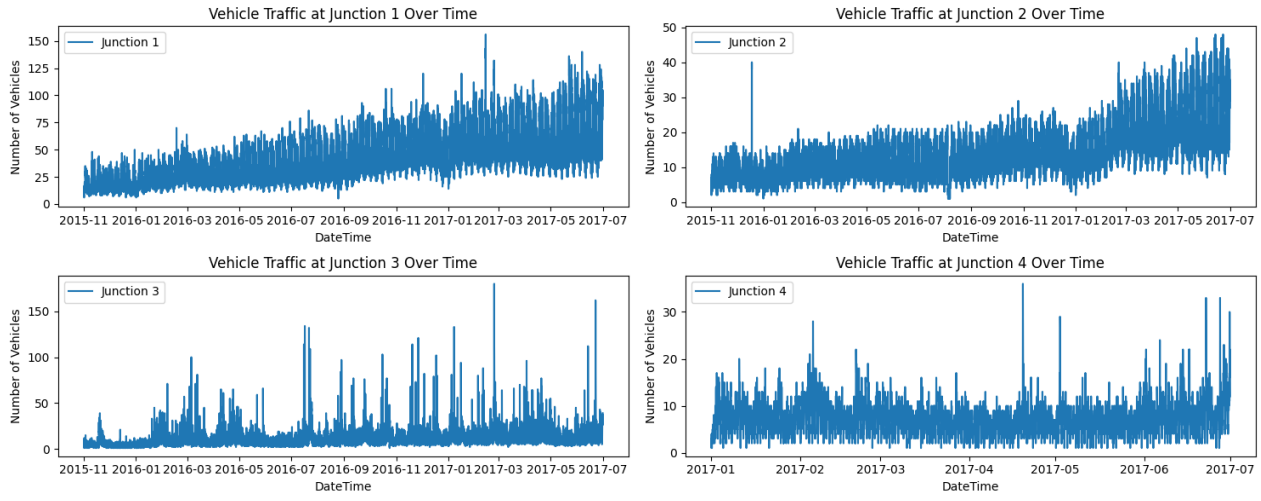

Fig 4 : Vehicle Traffic over time for each junction

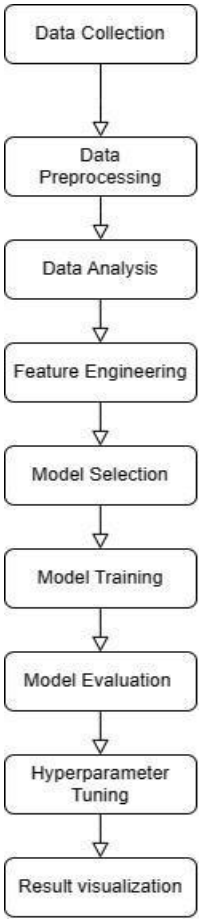

Fig 5 : Flow Chart of the model

Data Mining Algorithms

Various regression models are deployed to analyse and forecast the traffic patterns.

- **Decision Tree Regression**: This model is very simple and capable of collecting complex interactions that are non-linear by merging data into various leaves and branches which are depends on the decision nodes. It is very helpful for first explorations into data structure.
- **Random Forest Regression Model**: A group of strategies which develops the basic characteristics of decision trees via using multiple branches to boost forecast stability and accuracy. This model has a lower probability likely to undergo overfit and produces best rate.
- **XGBoost Regressor Model**: This regression is also known as a gradient boosting model because of its ability and performance in dealing with a vast variety of structured data attributes. It also improves the capability of the model to focus on challenging patterns in data, giving rise to more accurate forecast.

After the data analysis, we are now into proceeding to train our dataset on three different regressor models. We have taken the Vehicles as the target first and the rest of the columns as the feature

vectors and then divided our dataset into training and testing sets with a test size of 20%. Then, we have trained this data on three models. We have also performed hyperparameter tuning process using GridSearchCV to find the better parameters for the XG Boost Regression model. Once the models are trained, next we have to visualized the actual vs predicted vehicles on scatter plots for all three models as shown below :

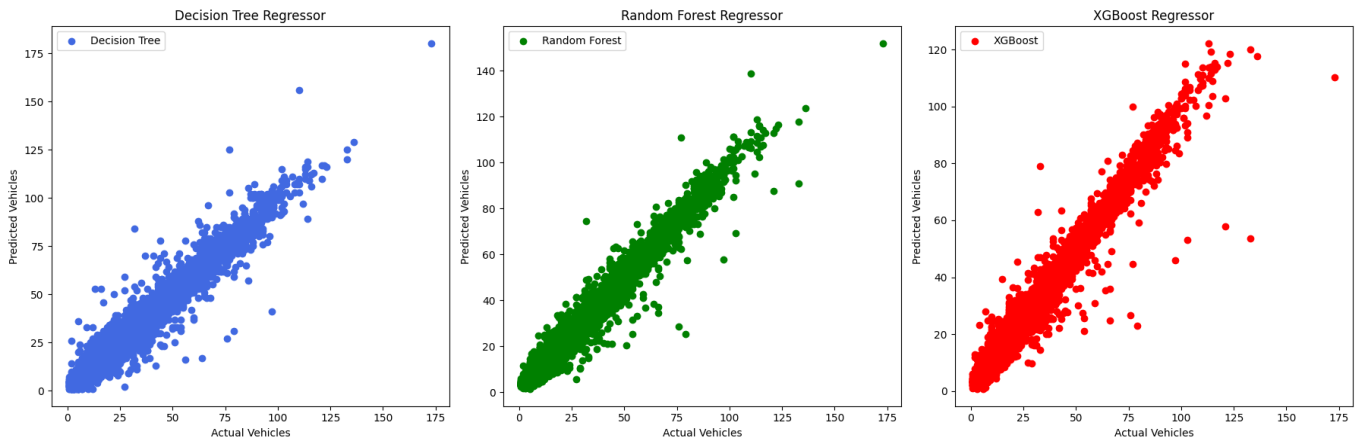

Fig 6 : Visualization of Actual vs Predicted Vehicles for Regression Models

For Now, we have used seaborn's distplot function to create the kernel density estimate plots with histograms of all the errors to show the error distribution among the different models in the estimation.

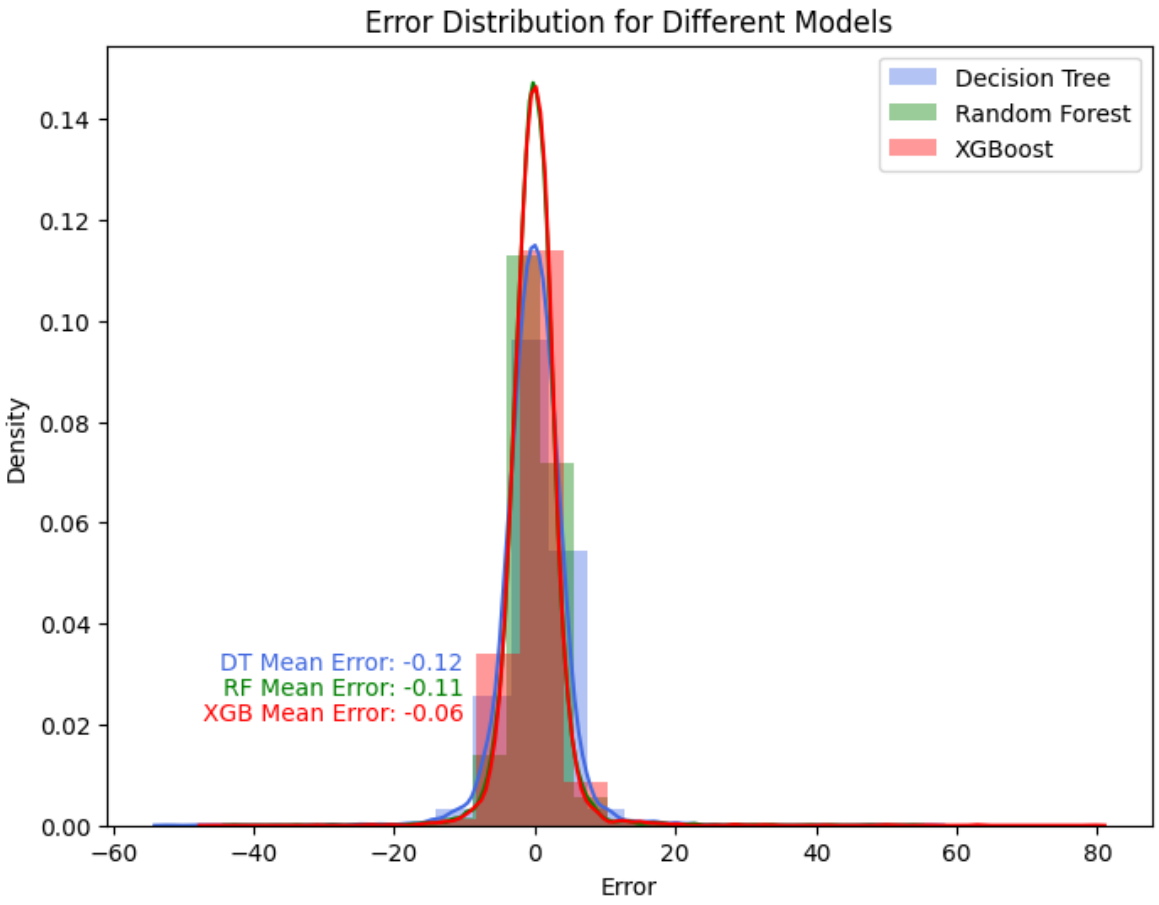

Fig 7: Error Distribution for Regression Models

Evaluation Metrics

To equitably evaluate the performance of the predictive models, the following metrics are used here:

1. **Mean Squared Error (MSE)**: This metric used to measure the average squared difference between the actual and predicted values, giving a clear indication of model accuracy. A lower MSE value shows a model with higher precision in forecasting traffic volumes.
2. **R2 Score**: This also known as the coefficient of determination, by this metric assesses the proportion of variance in dependent variable that is where predictable from the independent variables. It is a key for indicator of model fit, with a higher R2 score denoting a model that can better explain the variability in traffic data collected.

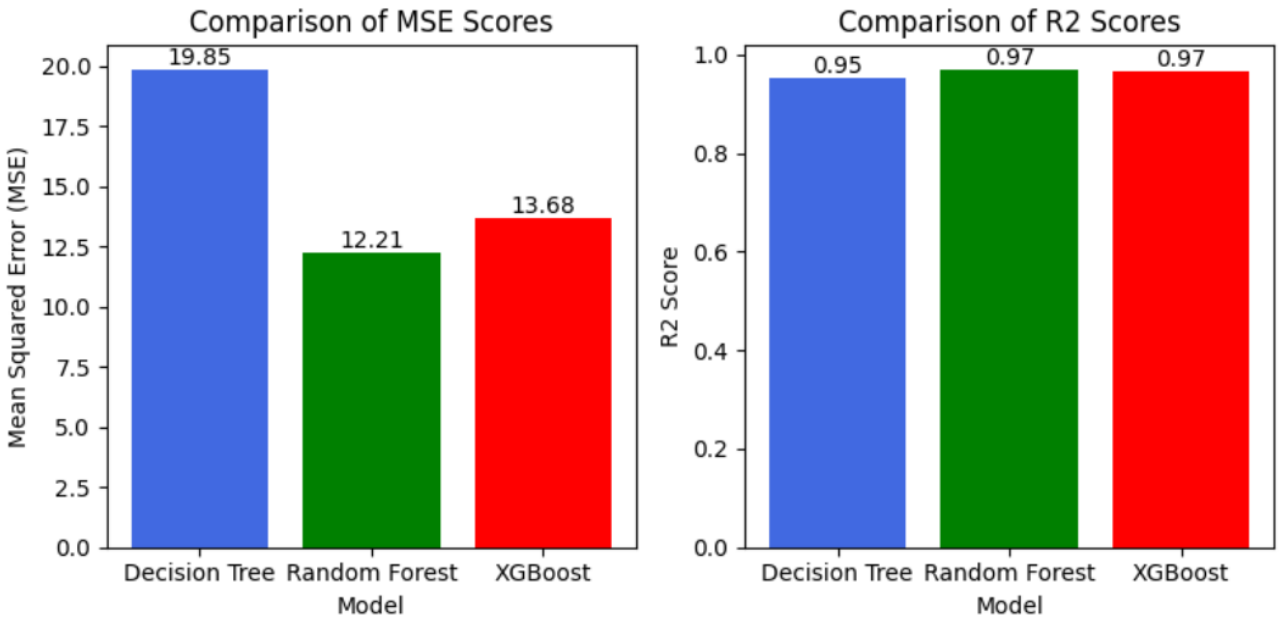

Fig 8: Comparison of MSE and R2 Scores

Results :

The complete tabulation and visualization of results are performed to concise present the effectuality of each model. This also includes the comparing MSE and R2 scores across models, as well as detailing the training and testing scores to illustrate each model's learning and abstraction capabilities. And this detailed approach provides a robust framework for analysing and predicting traffic flow patterns, provide that all aspects of the experimental setup, data handling, and result analysis are accurate addressed. Finally, we have tabulated the train and test scores of various models to evaluate their performance.

Model	Training Score	Test Score
Decision Tree	1	0.951287
Random Forest	0.995275	0.97004
XGBoost	0.982204	0.966432

Fig 9: Model Performance Scores on Train and Test Data

Related Work:

Predicting the traffic flow and obstruction with high-tech computer models is an important part of smart transportation systems for making traffic control better than that and lowering obstruction. From the recent research, like those that using Regressor models, shows that these tools are very good at accurately forecasting how traffic will move. For instance, using Random Forest Regressor as a post-processing tool has led to big performance gains in real-world situations like Tamil Nadu's IT corridor. In the same way, other studies have used to environmental data, like weather, along with machine learning algorithms to guess it when traffic will be bad in places like Delhi. This method considers not only previous and present traffic data, but also outside factors such as temperature and humidity. This creates a complete model change that can accurately predict future traffic violations.

Comparing the different predictive models helps identifying the most effective methods for traffic forecasting. For instance, a study in Beijing found nonparametric regression models to be more accurate and adaptable than traditional approaches. This highlights the importance of the selecting models based on traffic conditions and data characteristics. As forecasting techniques evolve, especially with innovative models and diverse datasets, traffic management systems can become more accurate, efficient, and safer.

Conclusion:

This study focused on analyzing traffic flow using machine learning , offering significant advantages for urban traffic planning and management. Traffic patterns at different times and intersections were examined using multiple regression models applied to a well-sourced and detailed dataset.

The models used to predict traffic patterns varied in their effectiveness. Among them, random Forest and XGBoost demonstrated the strongest performance, achieving lower Mean Squared values and the higher R2 score compared to the others. These models were also effective in capturing the non-linear relationships present in the data. Additionally, visualizations such as bar charts, line graphs, scatter plots, and distribution graphs provided deeper insights into both traffic patterns and model performance. These visual outputs highlighted temporal trends and showcased the model's prediction accuracy across various times and conditions in the datset.

Limitations:

There are few limitations despite the successes of the project. Which are shown below,

- **Scope of the data**: The dataset included only specific time points and locations, which may not fully capture the variability of traffic conditions or represents diverse urban environments accurately.
- **Narrow Feature Selection**: The focus was primarily on basic variables like time and junctions, potentially overlooking other critical factors such as accidents, road closures, or unexpected real-time events that significantly impact traffic flow.
- **Model Generalization**: While the models performed well on the test data, their effectiveness may vary across different cities or traffic scenarios, limiting their generalizability.

Future Work:

These are some works that can make our results stronger, and we can also consider the points mentioned below.

- **Expanding Data Collection**: Gathering data from different places and situations can help build models that work well in all kinds of areas.
- **Advanced Feature Engineering**: Adding more useful and detailed information like live traffic updates or data from social media can help make the model give more accurate results.
- **Deep Learning Approaches**: Using neural networks can help find hidden patterns in the data that simple models might miss, especially when the traffic is hard to predict.
- **Real-time Prediction Integration**: Building a live prediction system that works with traffic control can give quick suggestions an help improve traffic flow and road usage.

References:

[1]. Deshpande, Minal and Preeti R. Bajaj. “Performance Improvement of Traffic Flow Prediction Model using Combination of Support Vector Machine and Rough Set.” *International Journal of Computer Applications* 163 (2017): 31-35.

[2]. Asad, Moumita, Rafed Muhammad Yasir, Dr. Naushin Nower and Dr. Mohammad Shoyaib. “Traffic Congestion Prediction Using Machine Learning Techniques.” *ArXiv* abs/2206.10983 (2022): n. pag.

[3]. Lee, Jiwan, Bonghee Hong, Kyungmin Lee and Yang-Ja Jang. “A Prediction Model of Traffic Congestion Using Weather Data.” *2015 IEEE International Conference on Data Science and Data Intensive Systems* (2015): 81-88.

[4]. Priambodo, Bagus and Azlina Ahmad. “Predicting Traffic Flow Based on Average Speed of Neighbouring Road Using Multiple Regression.” *International Visual Informatics Conference* (2017).

[5]. Priambodo, Bagus and Yuwan Jumaryadi. “Time Series Traffic Speed Prediction Using k-Nearest Neighbour Based on Similar Traffic Data.” (2018).

[6]. Das, Anuradha, Swadhin Kumar Barisal and Pratik Dutta. “A Comparative Analysis on Traffic Flow Prediction.” *2022 International Conference on Machine Learning, Computer Systems and Security (MLCSS)* (2022): 318-324.

[7]. Wang, Jingyuan, Fei Hu, Xiaofei Xu, Dengbao Wang and Li Li. “A Deep Prediction Model of Traffic Flow Considering Precipitation Impact.” *2018 International Joint Conference on Neural Networks (IJCNN)* (2018): 1-7.

[8]. Huang, Wenhao, Guojie Song, Haikun Hong and Kunqing Xie. “Deep Architecture for Traffic Flow Prediction: Deep Belief Networks With Multitask Learning.” *IEEE Transactions on Intelligent Transportation Systems* 15 (2014): 2191-2201.

[9]. Bao, Xuexin, Dan Jiang, Xuefeng Yang and Hong Wang. “An improved deep belief network for traffic prediction considering weather factors.” *alexandria engineering journal* (2020): n. pag.

[10]. Liu, Zhiquan, Yao Hu and Xiangying Ding. “Urban Road Traffic Flow Prediction with AttentionBased Convolutional Bidirectional Long Short-Term Memory Networks.” *Transportation Research Record* 2677 (2023): 449 - 458.