Traffic Congestion Prediction Using Supervised Machine Learning Algorithm

Snir Levi, Stav Harpaz

Table of Contents

Abstract

In this research we analyzed the efficiency of supervised machine learning algorithms on traffic prediction. As our primary dataset we used the 'Traffic Prediction Dataset'. In addition, we used the same dataset from the article [*] to see if we can get close to their outstanding results (~99% accuracy).

For our research we mainly used the 'Traffic Prediction Dataset' dataset from Kaggle (https://www.kaggle.com/datasets/fedesoriano/traffic-prediction-dataset/data).
The dataset consists of 4 attributes and 48120 instances which were further divided into 70% for training and 30% for testing. We also used the relevant temperature from the 'Weather_Data' dataset (https://www.kaggle.com/datasets/alioraji/weather-data-nov-2015), and combined it with our data to see its effects on traffic congestion.
The dataset was used to formulate predictive models for traffic congestion using three supervised machine learning algorithms: Decision Tree Classifier, Support Vector Machine and Random Forest Classifier.
The formulation and simulation of the predictive models were carried out using Jupyter.
For the original dataset (without the temperature) the results show Classification Accuracy (%) of 82.98, 82.14, 82.94 for Random Forest Classifier, Support Vector Machine and Decision Tree Classifier algorithms respectively.

For the combined dataset (with the temperature) the results show Classification Accuracy (%) of 83.24, 82.47, 83.24 for Random Forest Classifier, Support Vector Machine and Decision Tree Classifier algorithms respectively. We can see a slight improvement with the weather data.
In an attempt to improve the accuracy, we added to the combined data two more features that were derived from the date – 'month' and 'day_in_month'. The results show Classification Accuracy (%) of 86.24, 84.87, 86.12 for Random Forest Classifier, Support Vector Machine and Decision Tree Classifier algorithms respectively. We can see a noticeable improvement with the added data.

As for the article dataset, we managed to achieve an accuracy of 91.73% with a Gradient Boost Tree. An Impressive result, but still quite far from the article results.

Introduction

Traffic congestion poses significant challenges, impacting economic efficiency, environmental sustainability, public health, and individual well-being. It leads to increased travel time, fuel consumption, and vehicle wear and tear, translating into higher economic costs for both individuals and businesses. Congested roads also result in elevated emissions, contributing to air pollution and climate change.
The quality of life for commuters is adversely affected, with longer travel times reducing personal time and contributing to stress and frustration. Health risks escalate due to pollution-related respiratory problems, heart disease, and psychological stress. Furthermore, congestion heightens safety risks by increasing the likelihood of accidents and delaying emergency response vehicles. It also signals inefficiencies in

transportation infrastructure usage, underscoring the need for improved traffic management and public transportation options.

Supervised Machine Learning (ML) is a powerful approach for predicting traffic conditions, leveraging historical data to predict future traffic patterns. Here are several reasons why it's particularly suited for traffic prediction:

1. Leveraging Historical Data: Traffic patterns often follow predictable trends based on time of day, day of the week, and other factors. Supervised ML models can learn from historical data where the traffic conditions are known, allowing these models to uncover and utilize these patterns to make accurate predictions.

2. Handling Complex Relationships: Traffic flow can be influenced by a wide range of factors, including weather conditions, accidents, road closures, and events. Supervised ML can handle these complex, non-linear relationships between various predictors and traffic conditions, which traditional statistical methods might struggle with.

3. Real-time Decision Making: By using supervised ML for traffic prediction, transportation authorities can make informed, real-time decisions to manage traffic flow, such as adjusting signal timings, issuing travel advisories, and rerouting traffic in response to predicted congestion.

4. Improving Safety and Efficiency: Accurate traffic predictions can enhance road safety by anticipating and mitigating congestion-related accidents. It also improves efficiency by helping drivers avoid traffic jams, saving time and reducing fuel consumption.

5. Adaptability: Supervised ML models can be continuously updated with new data, making them adaptable to changing traffic patterns and urban development. This ensures that the models remain accurate and relevant over time.

6. Variety of Models: There's a wide range of supervised ML models available, including regression, decision trees, and neural networks, each with its strengths. This variety allows for the selection and optimization of the best model for specific traffic prediction tasks.

In summary, supervised ML provides a flexible, powerful, and adaptive approach to predicting traffic conditions, leveraging historical data to make accurate predictions and support intelligent transportation systems.

The Algorithms Used in the Research

Primary Research

SVM (Support Vector Machine)
SVM is a supervised learning algorithm that finds the best boundary to separate different classes by maximizing the margin between data points of the classes.

Decision Tree Classifier
A decision tree classifier makes decisions by splitting data through a series of binary decisions, forming a tree structure from the root to the leaves, which represent the outcomes.

Random Forest Classifier
A random forest is an ensemble of decision trees, which improves prediction accuracy and reduces the risk of overfitting by averaging the results of individual trees.

Secondary Research

Stochastic Gradient Descent (SGD)
An optimization method used to minimize the error of a model (e.g., in logistic regression, neural networks). It updates the model's parameters using only a single sample or a small batch of samples at each iteration, making it efficient for large datasets.
We used it with the 'hinge' loss. When you use the `SGDClassifier` from `sklearn.linear_model` with `loss="hinge"`, it is equivalent to using a linear Support Vector Machine (SVM). The SGDClassifier with hinge loss optimizes the same objective as a standard linear SVM model, but it does so using stochastic gradient descent for optimization, making it more suitable for large datasets.

K-Nearest Neighbors (KNN)
A simple, instance-based learning algorithm where the class of a sample is determined by the majority class among its k-nearest neighbors. It's used for classification and regression tasks.

Gradient-Boosted Tree
An ensemble learning technique that builds models in a stage-wise fashion. It constructs new models that predict the residuals or errors of prior models and then combines these models through a weighted sum to make the final prediction more accurate.

Literature Review

The paper [*] presents a review of supervised machine learning algorithms for traffic congestion prediction. The authors discuss the advantages and disadvantages of each algorithm and provide examples of how they have been used in real-world applications.

The authors begin by defining traffic congestion and discussing the importance of traffic congestion prediction. Traffic congestion is a major problem in many cities, and can lead to increased travel times, pollution, and accidents. Traffic congestion prediction can be used to help improve traffic management and reduce the negative impacts of congestion.

The paper then reviews the results of several studies that have used supervised machine learning algorithms for traffic congestion prediction. These studies have shown that supervised machine learning algorithms can be effective in predicting traffic congestion. However, the authors note that the performance of these algorithms can vary depending on the specific data set and the algorithm that is used.

The authors present three supervised machine learning algorithms: Classification Tree, Support Vector Machine and Ensemble (RUSBoosted), to build a model that can be used to predict traffic congestion.

- **Classification Tree:** A decision tree algorithm that can be used for both classification and regression tasks. It works by recursively splitting the data into smaller and smaller subsets until each subset contains only one type of data point. It is a simple and easy-to-understand algorithm, but it can be less accurate than other algorithms.

- **Support Vector Machine (SVM):** A binary classification algorithm that can be used for both linear and non-linear problems. It works by finding the hyperplane that maximizes the margin between the two classes of data points. It is a more accurate algorithm than classification trees, but it can be more computationally expensive.

- **Ensemble (RUSBoosted):** An ensemble algorithm that combines multiple decision trees to improve overall accuracy. It is used for both classification and regression tasks. It is more robust than individual decision trees. It can be computationally expensive.

The dataset was obtained from Kaggle dataset repository (https://www.kaggle.com/datasets/bobaaayoung/trafficvolumedatacsv).
The dataset consists of 15 attributes, e.g.: date_time, is_holiday, air_pollution_index, temperature, traffic_volume, etc.
The data consists of information collected from 33750 instances and was split into two, namely: 70% training and 30% testing.

Results

The accuracy is 0.998, 0.999, 0.565 for Classification Tree algorithm, Support Vector Machine and Ensemble (RUSBoosted) respectively. The execution time and prediction speed of the Classification Tree algorithm is better compared to the other two algorithms. The area under curve is 0.66, 0.69, 0.69 for Classification Tree algorithm, Support Vector Machine and Ensemble (RUSBoosted) respectively.

Conclusion

The authors conclude their research by discussing the challenges of traffic congestion prediction. These challenges include the need for large and accurate data sets, the need to account for the dynamic nature of traffic, and the need to consider the cost of implementing traffic congestion prediction systems.

Supervised machine learning algorithms can be a valuable tool for traffic congestion prediction. However, it is important to carefully consider the advantages and disadvantages of each algorithm before selecting one for a particular application.

[*] 'Traffic Congestion Prediction using Supervised Machine Learning Algorithms' by Taiwo, E. O., Ogunsanwo, G. O., Alaba, O. B, Ogunbanwo, A. S

Materials and Methods

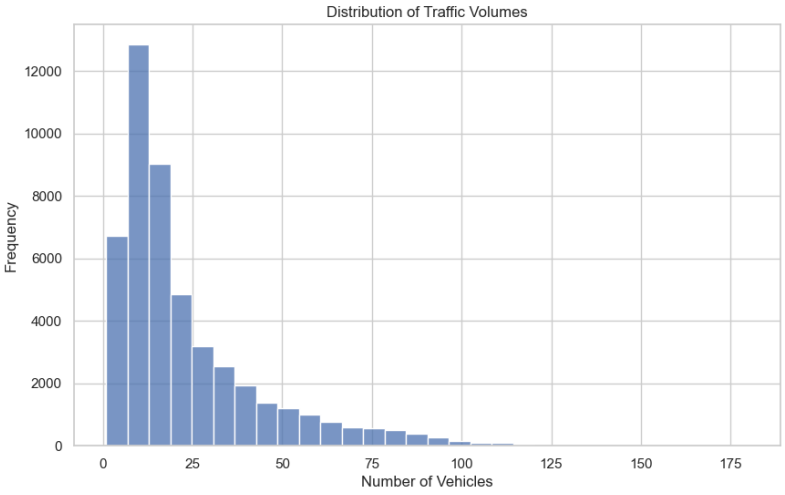
Datasets and Attributes

The datasets were obtained from Kaggle dataset repositories, we used the 'Traffic Prediction Dataset' (https://www.kaggle.com/datasets/fedesoriano/traffic-prediction-dataset/data), which consists of 4 attributes namely: DateTime, Junction, Vehicles, ID.
The data did not have any information about weather, so we later combined it with the relevant temperature from the 'Weather_Data' dataset (https://www.kaggle.com/datasets/alioraji/weather-data-nov-2015), and compared their results.
The article dataset (https://www.kaggle.com/datasets/bobaaayoung/trafficvolumedatacsv).
To predict the state of the traffic we created a threshold variable and divided the target to high/low traffic accordingly.

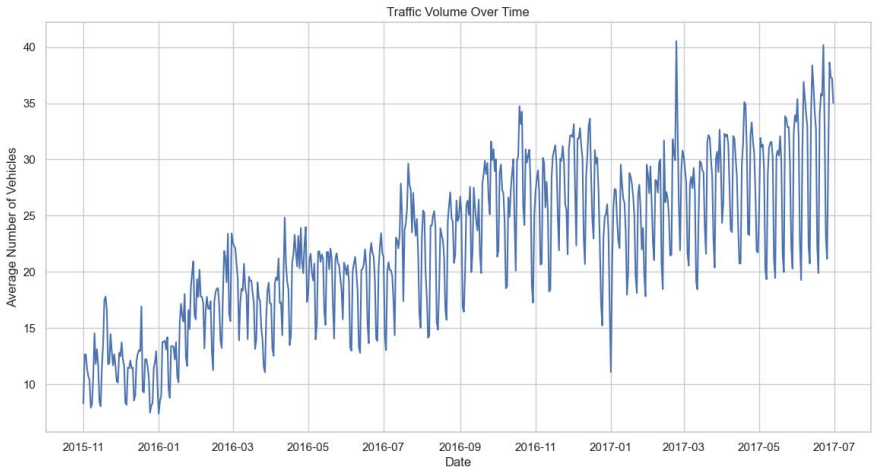
Data Analysis

We created various graphs and diagrams to see how the features behave and impact the number of vehicles on the road (target).

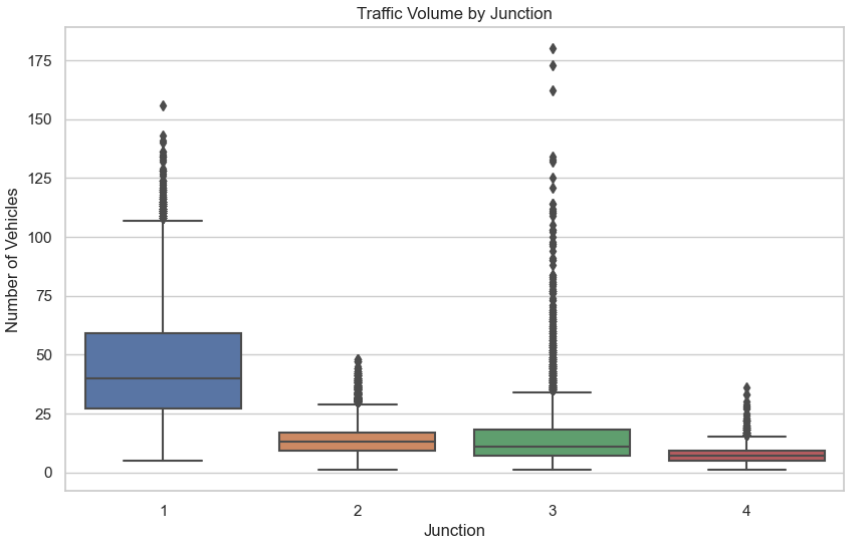
Primary Research



Frequency of vehicle number

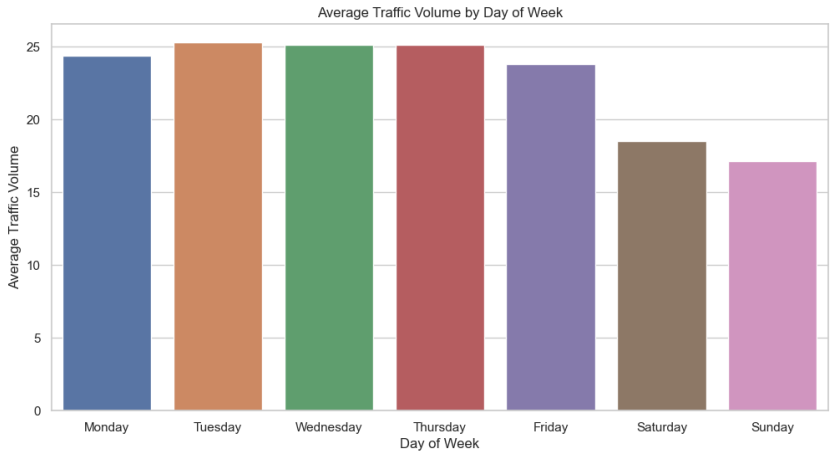


Average number of vehicles per day through out all the dates in the data

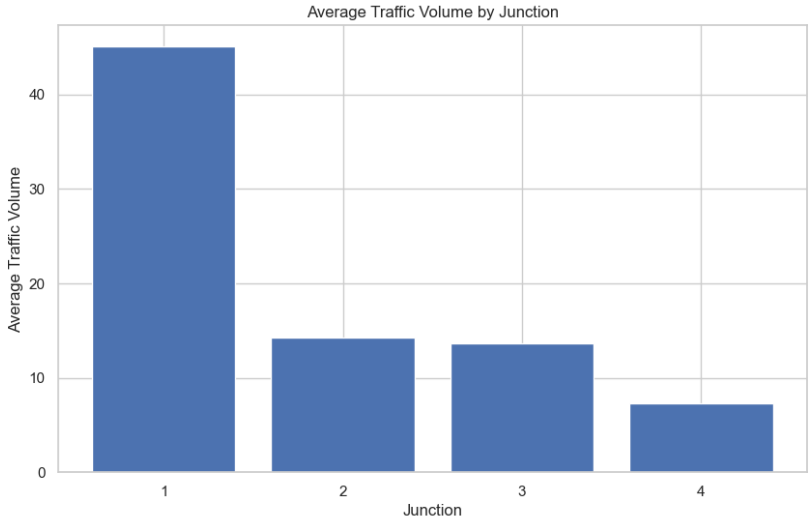


Boxplot* of the number of vehicles in each junction.

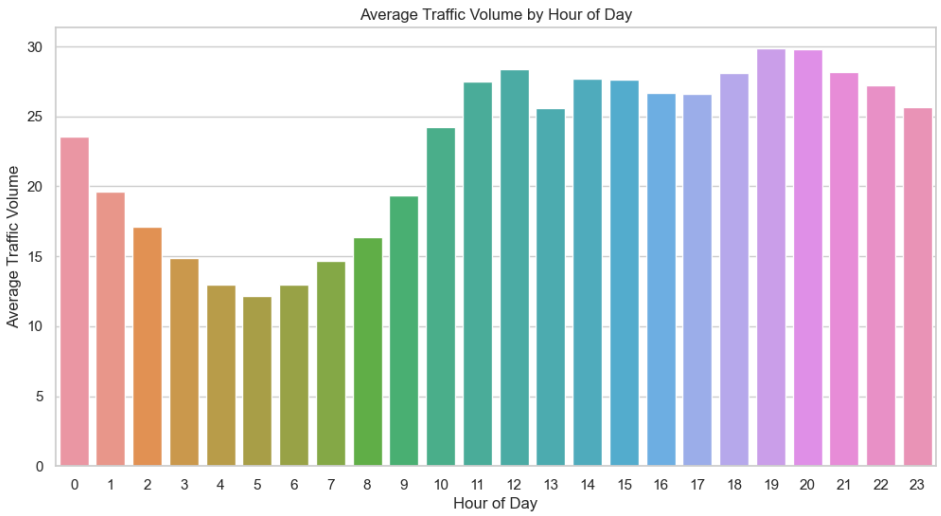
* A boxplot in matplotlib is a standardized way of displaying the distribution of data based on a five-number summary: minimum, first quartile, median, third quartile, and maximum, visually highlighting the central tendency, dispersion, and skewness of the data.



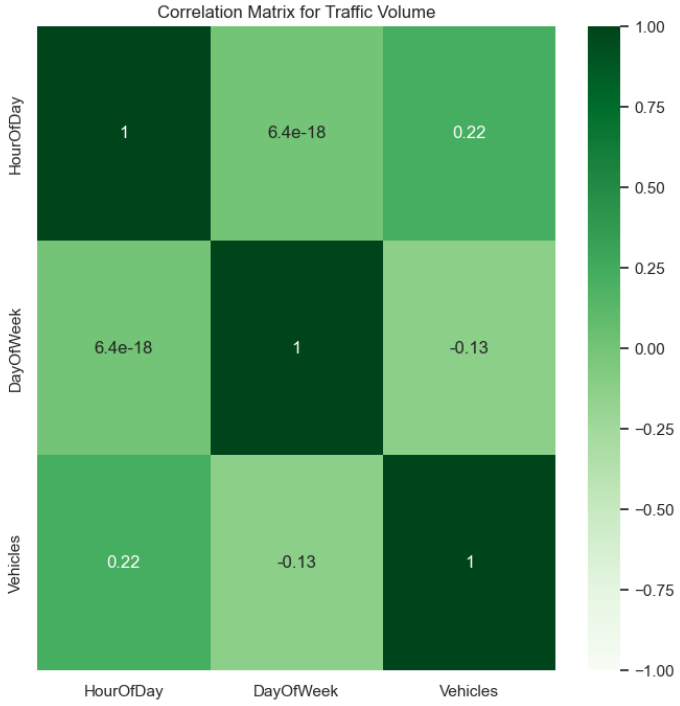
Average traffic volume by day of the week



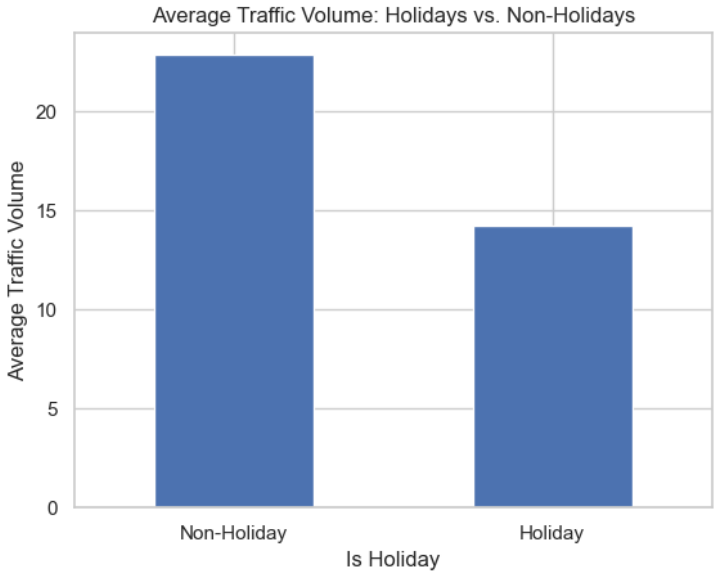
Average traffic volume by junction



Average traffic volume by hour of day

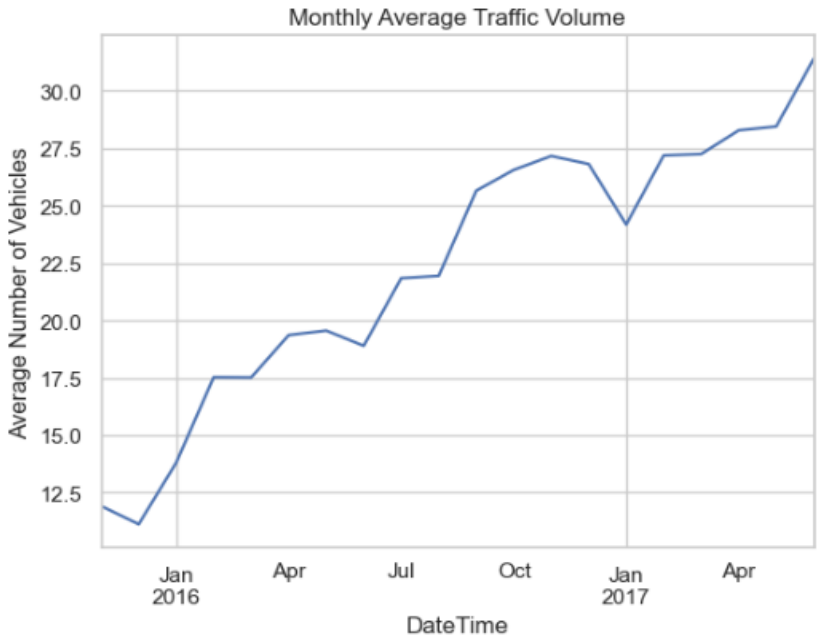


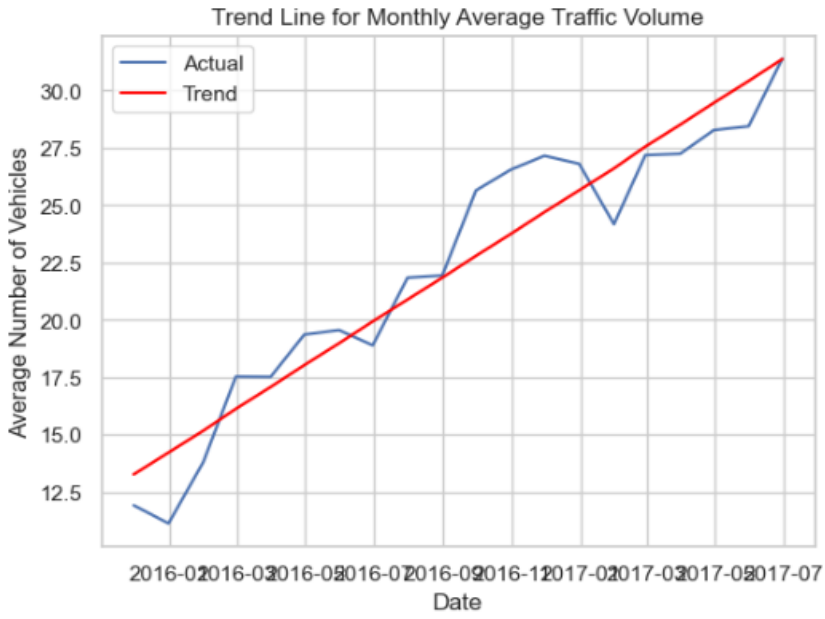
A correlation matrix



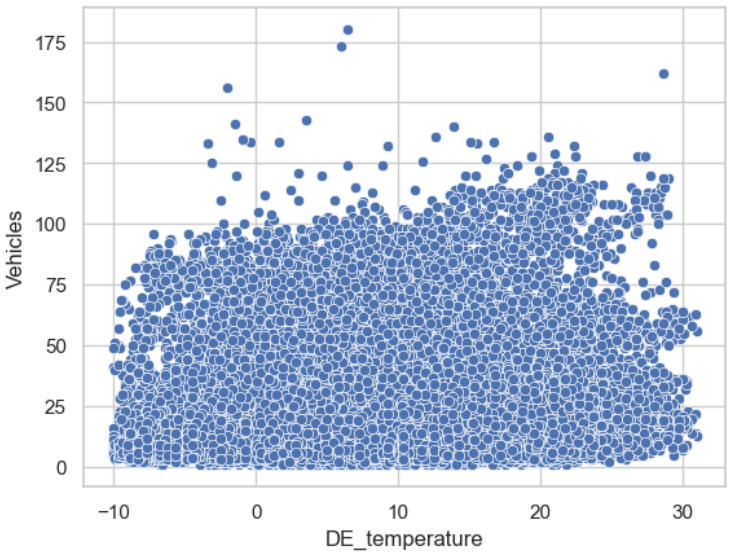
Average traffic volume by holiday.

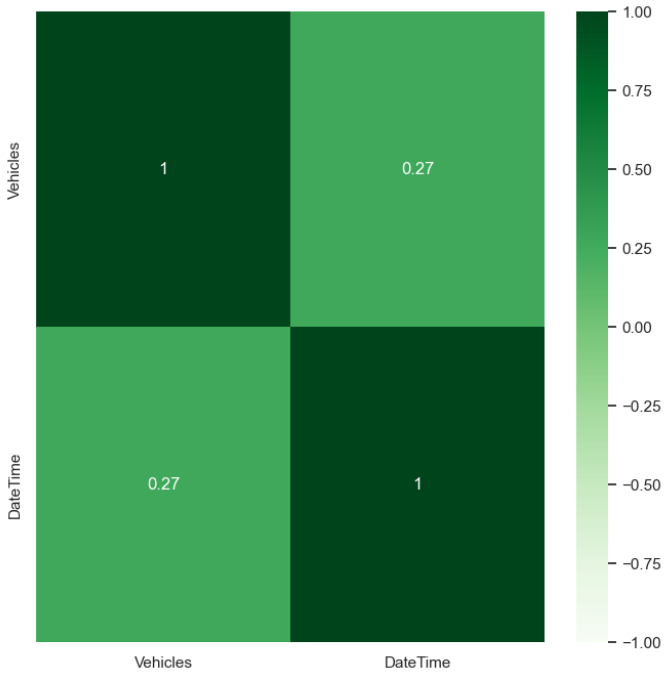
We added a new column 'is_holiday', and injected well known holidays to see if they impact the traffic.





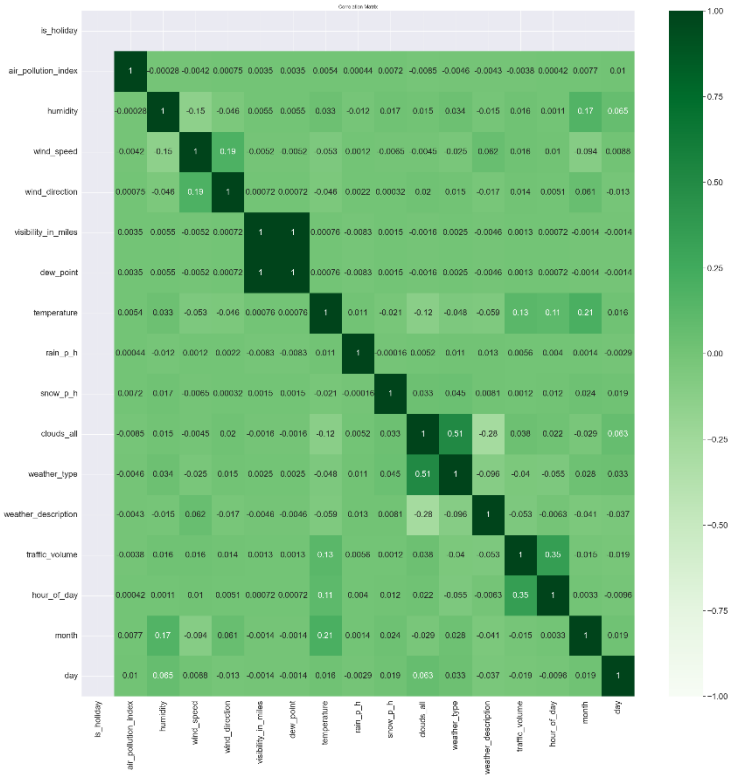
The first chart displays a rising trend in monthly traffic volumes with some fluctuations, including a notable dip in October 2016 before recovering.
The second chart, with a trend line in red against actual traffic in blue, confirms a consistent increase over time, hinting at urban development, population growth, or altered traffic patterns.
This upward trend implies the need for revised traffic management and infrastructure planning to address congestion.



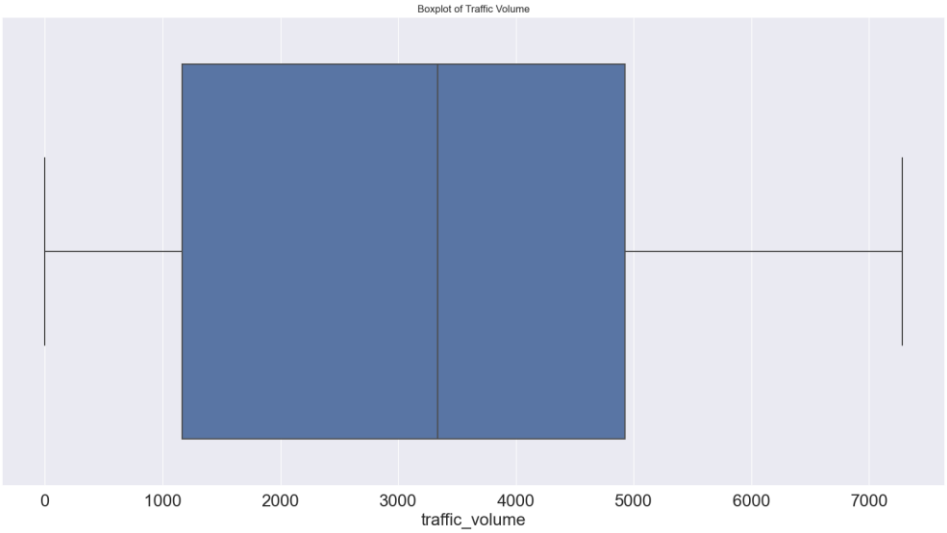


The correlation matrix shows a slight positive correlation (0.27) between time and vehicle count, indicating a minor increase in vehicles over time, though the relationship is weak.
Scatter Plot reveals no apparent link between temperature and vehicle numbers, suggesting temperature isn't a strong traffic predictor.

Secondary Research



A correlation matrix for all features.
We can see that some features are very weakly correlated to the target, such as: air_pollution_index, rain_p_h, snow_p_h, visibility_in_miles.
And that visibility_in_miles and dew_point have a 1 to 1 correlation, so we can keep just one of them.



boxplot of traffic_volume (target).

The median traffic volume, indicated by the line inside the box, appears to be just above 3000 vehicles. There do not appear to be any outliers, as there are no points beyond the 'whiskers' of the boxplot, which means there are no traffic volume counts that are unusually high or low.
This boxplot suggests that while there is variability in traffic volumes, there are no extreme anomalies, and most of the traffic volumes are clustered around the median value.

Results

Primary Research

Model	Train Accuracy	Test Accuracy
SVM (Support Vector Machine)	84.11%	84.87%
Decision Tree Classifier	86.64%	85.26%
Random Forest Classifier	87.31%	86.24%

Secondary Research

Model	Train Accuracy	Test Accuracy
SGD with 'hinge' loss	66.91%	66.67%
KNN	87.04%	84.88%
Gradient-Boosted Tree	92.87%	91.73%

Comparing to the Article Research Results

Primary Research

Model	Article Accuracy	Our Accuracy
SVM (Support Vector Machine)	99.9%	84.87%
Decision Tree Classifier	99.8%	85.26%
Ensemble	56.5%	86.24%

Differences in accuracy between the "Article Accuracy" and "Our Accuracy" as listed in the table can be attributed to several factors even if the data differs:

1. **Data Quality and Quantity**: The dataset used by the article may be more balanced, or cleaner compared to the one we have used, which can significantly influence model performance.
2. **Feature Selection**: The article's models might use a different set of features or may have engineered features more effectively to capture the underlying patterns in the data.
3. **Model Parameters and Tuning**: The models cited in the article might be better tuned with more optimal hyperparameters, leading to higher accuracy.
4. **Preprocessing Techniques**: Differences in how the data was preprocessed (normalization, handling missing values, encoding categorical variables) can cause variations in model performance.
5. **Model Version and Implementation**: There might be differences in the version of the machine learning algorithms or the specific implementations used which could impact accuracy.
6. **Evaluation Methodology**: The discrepancy could also be due to different evaluation methodologies, such as the split of training and testing data, or the use of different cross-validation strategies.

Secondary Research

We tried to reach their results by using different algorithms on the same dataset. The best accuracy score we managed to get is 91.73% with the Gradient-Boosted Tree algorithm.
Achieving a maximum accuracy of 91.73% with a Gradient-Boosted Tree model is indicative of a robust and well-performing model, albeit still short of the article's reported accuracies.
The discrepancies highlight the importance of understanding all aspects of a machine learning pipeline, from preprocessing and feature selection to model selection and hyperparameter tuning, to replicate or improve upon previously reported results.

Conclusion

In this seminar paper, we explored the efficacy of various supervised machine learning algorithms on traffic prediction using a primary dataset for model training and testing. We attempted to match the high accuracy rates reported in a related article but achieved a maximum accuracy of 91.73% with the Gradient-Boosted Tree algorithm. We discuss the factors that might contribute to the differences in accuracy, such as data quality, feature selection, and model tuning, and suggests that further optimization and understanding of the article's methodology could close the performance gap.

For Further Research

The SVM and Decision Tree Classifier show a significant drop in accuracy from the article to our results, which might indicate a need for improved data preprocessing, feature engineering, or model tuning in your approach.
Interestingly, the Ensemble method, which typically combines predictions from multiple models to improve accuracy, shows a significantly better performance in our results compared to the article. This could suggest that our ensemble method is well-tailored to our specific dataset, or it could indicate that the article's ensemble approach was suboptimal or not well-suited to their data.

Further investigation into the specifics of the article's methodology and a more thorough optimization process for our models could help in closing the gap between our results and those reported in the article.