Research on traffic congestion prediction based on analyzable machine learning

Yixin Li [1, *], Wenqi Zhang [2]

[1] College of Mechanical and Automotive Engineering, Anhui Polytechnic University, Anhui, China, 241000

[2] College of Transportation Engineering, Shandong Jianzhu University, Yantai, China, 264000

* Corresponding Author Email: lyx210141144@163.com

Abstract. As motor vehicles proliferate around the world, traffic congestion increases, hindering travel, polluting the environment and hampering economic growth. Accurate congestion prediction is the key to optimizing traffic flow. In this study, the interaction between different traffic modes and congestion degree is deeply investigated, and a high-fidelity prediction model is established. The study analyses monthly, weekly, daily and hourly multi-modal traffic and congestion data and identifies congestion patterns. In this paper, Pearson's coefficient is applied to identify the key influencing factors, and congestion is predicted by KNN, logistic regression, decision tree and random forest models. The random forest model reaches 99.88% accuracy after 5-fold cross-validation and grid search optimisation. Further analysis using the Random Forest's feature importance scores revealed that traffic volume was the most important predictor of congestion, while time factors such as time and date played a smaller role. This study not only establishes a robust congestion prediction model, but also emphasizes the importance of traffic volume in congestion prediction. The findings provide valuable insights for traffic management agencies to take proactive measures to alleviate traffic congestion. By harnessing the predictive power of random forest models, these agencies can increase the efficiency of traffic management, improve the overall travel experience, and contribute to sustainable urban development. This study also lays a solid foundation for future research in the field of traffic congestion prediction and management.

Keywords: Pearson correlation coefficient, Decision Tree, Random Forest.

1. Introduction

With the rapid development of the global economy and the acceleration of urbanization, the number of motor vehicles has increased dramatically, leading to the worsening of traffic congestion. This problem not only significantly reduces people's travel efficiency and quality of life, but also poses a challenge to urban economic development and environmental protection [1]. In the face of an increasingly complex traffic environment, traditional traffic management tools are no longer able to cope with the growing traffic pressure, and therefore, there is an urgent need to adopt more intelligent and efficient solutions. In the era of artificial intelligence and big data, the method of predicting traffic congestion through machine learning algorithms has begun to receive more and more attention from scholars. Through the analysis and mining of massive traffic data, machine learning technology can predict potential traffic congestion in advance, helping the public to reasonably plan travel routes and reduce travel time and costs. At the same time, accurate traffic congestion prediction can also provide decision-making support for the traffic management department, optimize the road traffic strategy, effectively alleviate traffic pressure, and improve the overall efficiency and stability of the urban transportation system [2].

The construction of the traffic flow indicator system and prediction methods are two key components in the prediction of traffic congestion problems, the problem in this paper traffic flow prediction to study the short-time traffic. Zheng Shujian et al [3]. Listed the domestic and foreign traffic congestion evaluation index calculation methods, systematically sorted out the existing research results of traffic flow indicators, and compared and analyzed many indicators to build a more complete indicator system. Lu Erkan et al [4]. Used the three dimensions of time, speed and mileage

to construct a comprehensive index of urban traffic operation to evaluate urban traffic congestion. Tang Yi et al [5]. Proposed an improved time series model based on highway traffic flow data to overcome the shortcomings of the traditional time series model. Huang Helei et al [6]. Proposed the maximum deviation similarity criterion for overcoming the unsatisfactory effect of commonly used clustering algorithms. Although the above studies have achieved important results in the field of traffic congestion prediction, there are still some shortcomings. First, most of the existing studies focus on the analysis of single dimensions or limited indicators, and lack in-depth exploration of the comprehensive multi-dimensional characteristics. Second, traditional models show limitations in dealing with data with high nonlinearity and complexity, making it difficult to comprehensively capture the dynamic changes in traffic flow. In addition, current research needs to be improved in terms of model interpretability and visualization.

In summary, this paper aims to reveal the complex relationship between traffic congestion patterns and temporal features through the in-depth analysis of multi-dimensional temporal features. Specifically, this paper first analyzes the traffic data in different time dimensions, such as month, week, day, hour, etc., and uses Pearson's correlation coefficient to explore the correlation between temporal features, the number of each type of transportation and the traffic congestion situation. Subsequently, this paper establishes and compares various machine learning classification models by taking the number of each type of transportation as an input feature in order to predict traffic congestion. Through 5-fold cross-validation and grid search to optimize the model hyper parameters, it is finally found that the classifier based on the tree model performs the best, which provides empirical evidence for the selection of traffic congestion prediction models, and the technical roadmap of this paper is shown in Figure 1.

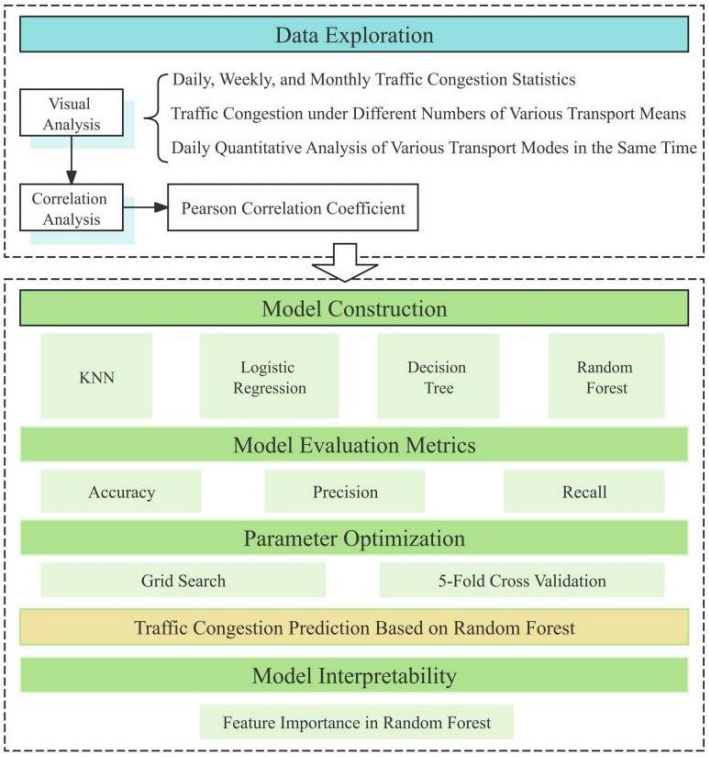

Figure 1. Technological route.

The innovation of this paper is that multiple time dimensions are analyzed and multiple machine learning models are compared for traffic congestion data, providing a new perspective on traffic congestion prediction. In addition, this paper provides an in-depth discussion on the interpretability of random forest and decision tree models in the classification task, revealing how these algorithms provide a decision basis for the classification results. The research results not only help to improve the accuracy of traffic congestion prediction, but also provide theoretical support and practical reference for the optimization of future intelligent transportation systems.

2. Data Exploration

2.1. Source of data

The traffic congestion data used in this paper was obtained from https://www.kaggle.com. The data records traffic congestion on roads from the 10th of one month to the 9th of the following month, with data sampling intervals of 15 minutes, and contains a total of 2976 observations.

The data contains information about Time, Car Count, Bike Count, Bus Count, Truck Count and Traffic Situation, some of which are shown in Table 1.

Table 1. Presentation of raw data.

Time	Date	Day of the Week	Car Count	Bike Count	Bus Count	Truck Count	Total	Traffic Situation
12:00:00 AM	10	Tuesday	31	0	4	4	39	low
12:15:00 AM	10	Tuesday	49	0	3	3	55	low
12:30:00 AM	10	Tuesday	46	0	3	6	55	low
...	...	...	...	...	...	...	...	...
11:15:00 PM	9	Thursday	15	4	1	25	45	normal
11:30:00 PM	9	Thursday	16	5	0	27	48	normal
11:45:00 PM	9	Thursday	14	3	1	15	33	normal

2.2. Data visualisation and analysis

Based on the above data, a time series graph of traffic congestion is plotted to observe the trend of traffic congestion conditions in different time periods.

By counting the traffic congestion condition in different time of the day and using the kernel density estimation curve to show the distribution pattern of the traffic condition, as shown in Figure 2.

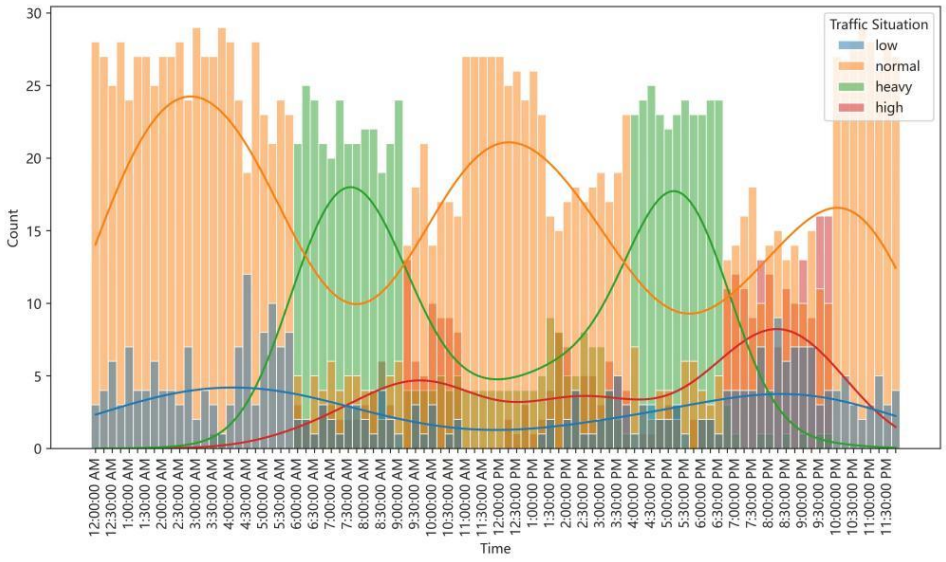

Figure 2. Distribution of traffic situation at different times of the day.

Looking at Figure 2 above, it can be seen that high traffic situation occurs during the morning rush hour (6am-9am) and evening rush hour (4pm-6.30pm), while traffic is smoother at night and in the early hours of the morning. This is in line with the daily situation, as residents generally follow a regular travelling pattern of 'nine-to-five' and are therefore prone to traffic congestion during this time, while the night time and early morning hours are smoother due to the reduced demand for work or study. The distribution of traffic situation as 'low', 'normal' and 'heavy' has a clear periodicity within a day, and 'high' is not cyclical.

Traffic congestion was analyzed by counting traffic congestion in weekly and respectively, as shown in Figure 3.

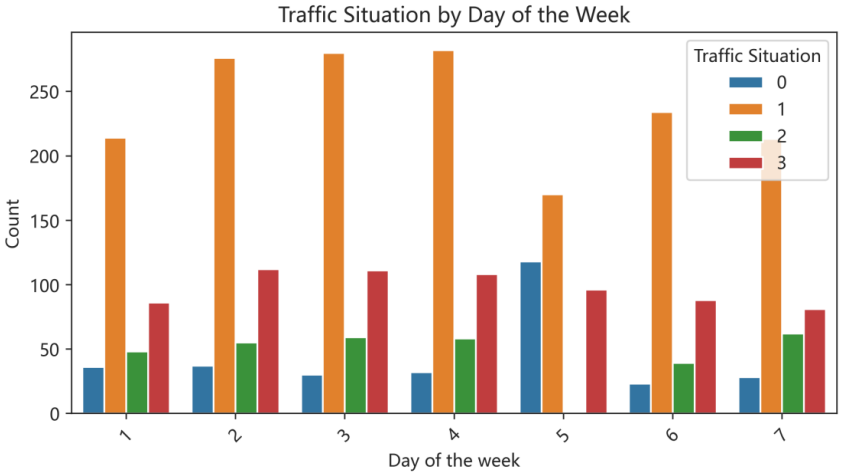

Figure 3. Weekly Traffic Situation Statistics.

As can be seen in Figure 3, from the weekly dimension, the traffic pressure on Friday is the smallest in a week, and there is not much difference in the traffic pressure from Tuesday to Thursday; from the monthly dimension, the traffic situation is 'normal' for the most time.

Next, the traffic conditions of different numbers of transport modes were counted and only partially presented due to space reasons, and the results are shown in Figure 4.

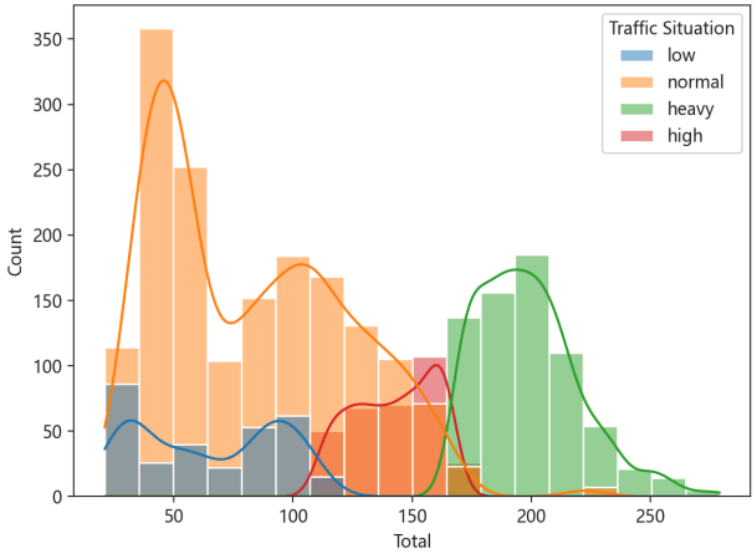

Figure 4. Traffic congestion with different numbers of transport modes.

As can be seen in Figure 4, the higher the number of cars, the more congested the traffic is when the number of trucks is lower. This phenomenon may be attributed to the fact that residents rely heavily on cars for their daily travels, leading to a rise in car ownership, which in turn raises the potential risk of traffic accidents and ultimately exacerbates traffic congestion. The demand for trucks is weak compared to the daily travel demand of residents, and in order to mitigate the impact on residents' daily travel, trucks are usually restricted to daytime hours and operate mainly at night and in the early hours of the morning when the demand for residents' travel is low. Also as a result, with fewer cars and more trucks, road conditions tend to be smoother and congestion is significantly reduced.

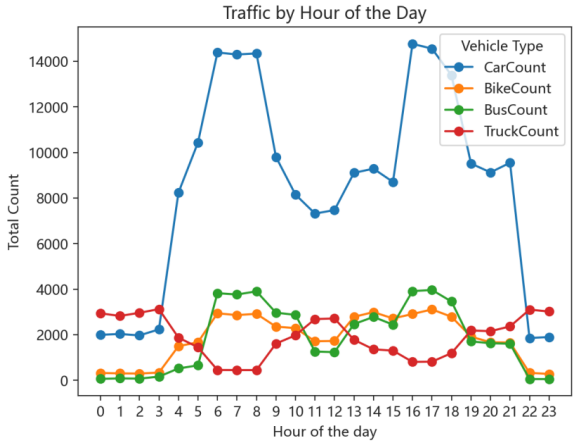

Figure 5. Statistics on the number of different vehicles at each time of the day.

By counting the number of different means of transport at different times of the day, the results of Figure 5 were obtained. It is observed that the number of cars peaks from 6am to 8am and 4pm to 6pm and the number of trucks peaks from 10pm to 3pm and 11pm to 12pm. The trend in the number of bikes and buses during the day follows the trend in the number of cars, but in smaller numbers.

Car traffic peaks in the morning and evening hours and coincides with people's commute times, indicating that a large number of people need to travel during these hours and use cars as a means of travelling. Bicycle and bus volume trends are similar to those of cars, indicating that these modes of transport also primarily serve commuting needs, but are relatively lightly used. Trucks, on the other hand, show peaks in traffic late at night and midday, which is related to the specific need of logistics transport to avoid congested hours, thus reaching peaks from 10 p.m. to 3 a.m. the next day, and from 11 to 12 noon.

2.3. Correlation coefficient analysis

By analyzed the statistics that have been produced, it is determined that traffic congestion conditions may be related to the time of day, day of week, and means of transport travel. To further investigate the relationship between these features and traffic situation, Pearson's correlation coefficient between the features and traffic situation is calculated and heat map is plotted.

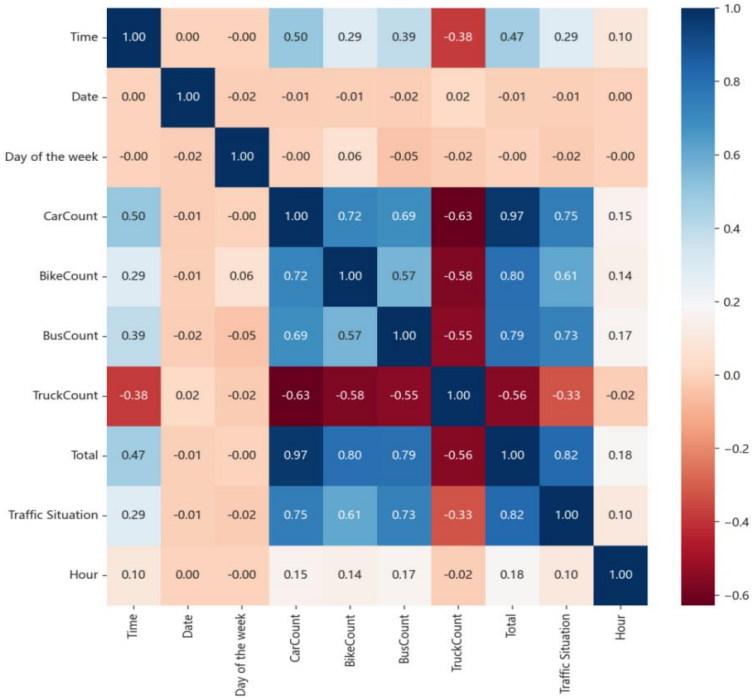

Figure 6. Heat map of correlation between different features and traffic conditions.

Looking at Figure 6, it is found that 'Car Count', 'Bus Count' and 'Total' have a positive correlation with a large correlation coefficient with the severity of the traffic situation; the number of trucks has a negative correlation with the severity of the traffic situation. The date and day of the week have little or no correlation with the traffic situation.

3. Model building

3.1. Introduction to modelling principles

In order to predict traffic congestion through the number of different transport types, this paper establishes four classification models, KNN [7], Logistic Regression, Decision Tree and Random Forest, and selects the optimal classification model based on the balanced accuracy.

Random forest taxonomy is an integrated learning method based on decision trees. Random forest taxonomy constructs a decision tree for each sub-datasets by extracting multiple sub-datasets from the original datasets in a playback manner [8]. After all the decision trees are constructed, Random Forest uses a voting method to make the final classification prediction by integrating the predictions of these trees. The advantage of random forest is that it is more resistant to noise, can effectively prevent overfitting and improve the generalization ability of the model. It also has some disadvantages, such as poorer model interpretability and longer training time. The principle of random forest classification is shown in Figure 7.

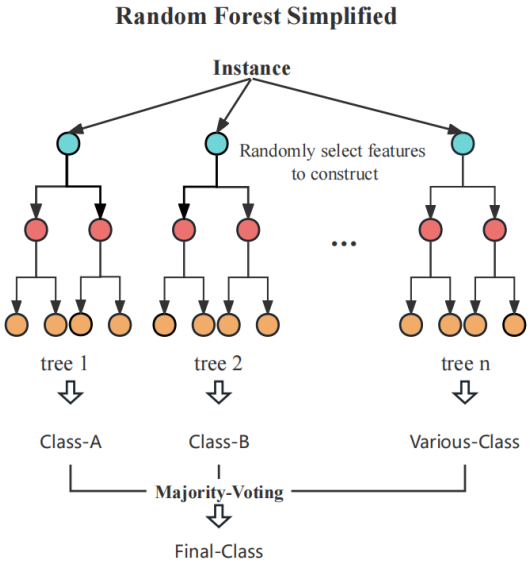

Figure 7. Random Forest Theory.

3.2. Indicators for model evaluation

Metrics were used to assess the performance of the classification model. Where TP denotes true Positive, TN denotes True Negative, FP denotes False Positive and FN denotes False Negative.

Accuracy: indicates the number of samples correctly classified as a proportion of the total number of samples.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision: indicates the proportion of samples predicted by the model to be positive cases that are actually positive cases.

$$\Pr ecision = \frac{TP}{TP + FP} \tag{2}$$

Recall: indicates the proportion of all samples that are actually positive cases that are correctly predicted to be positive by the model.

$$\text{Re}\,call = \frac{TP}{TP + FN} \tag{3}$$

3.3. Model hyperparameters

In this paper, we find the optimal combination of parameters through grid search and evaluate the model performance using 5-fold cross validation. The raw data is divided into 5 parts and each time one part of the data is used as a test set and the remaining part is used as a training set. Repeat the selection of 5 different test sets and finally calculate the average of the model evaluation metrics of all 5 tests as the estimation of model performance. In this case, grid search is a method that searches for all possible parameter combinations within a specified parameter range by exhaustive enumeration and uses cross-validation to evaluate the performance of each combination to find the optimal parameter configuration. The results of finding the best hyper parameters for KNN, Logistic Regression, Decision Tree, and Random Forest models are shown in Table 2 below:

Table 2. Model hyper parameters.

Model	Hyperparameters
KNN	k=1
Logistic Regression	C=1.5, penalty: L1
Decision Tree	criterion: entropy
Random Forest	max features: 4, n_estimators: 91

3.4. Results

The data was divided into 80% training set and 20% test set and the accuracy of the test results of the four models was obtained as shown in Table 3:

Table 3. Model accuracy.

Model	KNN	Logistic Regression	Decision Tree	Random Forest
Accuracy	0.8572	0.6904	0.9988	0.9959

The confusion matrix of the random forest is obtained as shown in Figure 8:

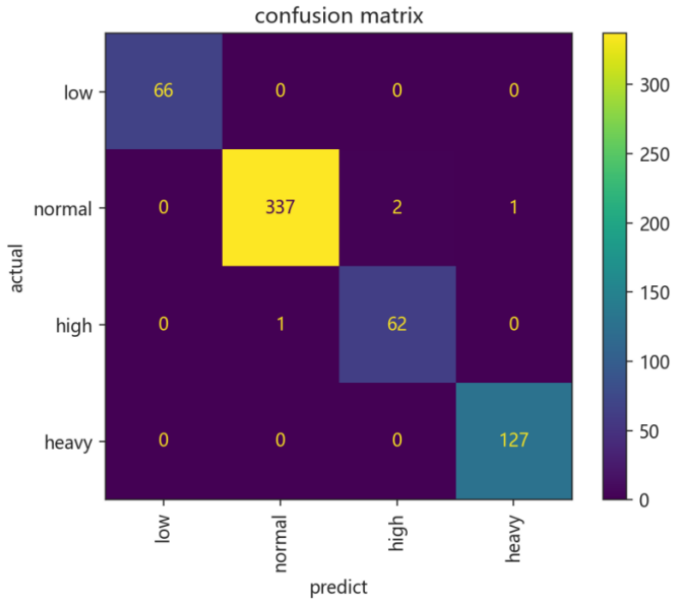

Figure 8. Random Forest Confusion Matrix.

From the figure, it can be seen that among the 596 test samples, only 4 prediction results do not match the real value, and the accuracy of the model is high, so Random Forest is chosen as the model for predicting traffic congestion.

3.5. Model interpretability

Random forests are collective decisions based on many decision trees, using each tree to analyses and predict the data independently, and ultimately arriving at the final prediction through a voting mechanism or averaging strategy. Random forests provide a way to quantify the importance of features, i.e., the Gini index is used as an evaluation metric to measure the extent to which each feature contributes to the prediction results when constructing all decision trees [9]. This helps to understand which features have the greatest impact on the model's decisions, resulting in strong interpretability.

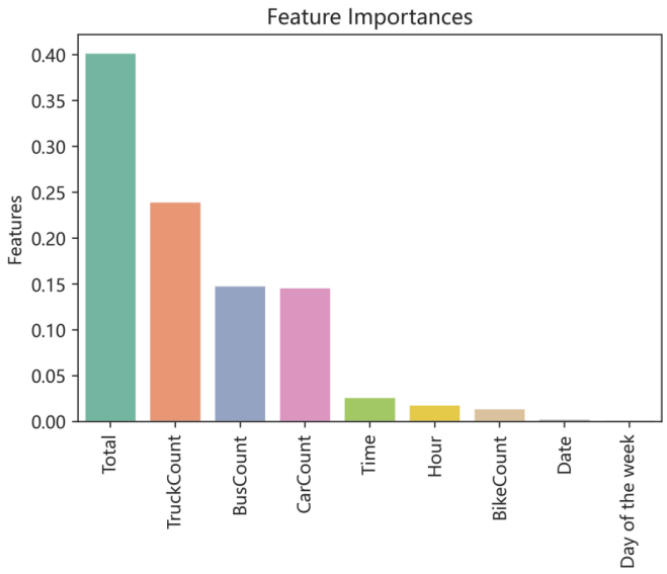

Figure 9. Feature importance score.

Observing Figure 9 above, it is found that ‘Total’ has the highest feature importance score, followed by ‘Trunk Count’, ‘Bus Count’, ‘Car Count’ in that order, while features such as ‘Time’, ‘Hour’, ‘Bike Count’, ‘Date’, ‘Day of the week’ have a lesser impact on the prediction results. The features of ‘Time’, ‘Hour’, ‘Bike Count’, ‘Date’, ‘Day of the week’ have less influence on the prediction results.

In the decision tree visualization graph, each node represents a feature or decision point, branches represent different feature takes or decision results, and leaf nodes represent the final prediction results [10]. The visualization diagrams allow to see how the decision tree splits and makes decisions based on the values of the features and how much each feature affects the prediction results. Decision tree visualization diagrams are very helpful in understanding and interpreting models, especially for complex decision tree models. The visualization diagrams allow to quickly find the key features that affect the prediction results, to understand the correlation and interactions between different features and how the model behaves in different situations.

4. Conclusions

This study reveals the close correlation between traffic congestion situation and the flow of different modes of transport through time series analysis of traffic congestion data of a road and Pearson correlation assessment. On this basis, the study constructs and compares KNN, Logistic Regression, Decision Tree and Random Forest classification models for predicting traffic congestion, and at the same time obtains the optimal hyper parameters of the above models through 5-fold cross-validation and grid search. The results show that the tree model can capture the complex nonlinear relationships in traffic data more effectively, and has a strong ability to predict traffic congestion.

This study not only reveals the regular relationship between traffic flow and congestion, but also makes a breakthrough in the accuracy of model prediction, which is especially important for practical applications in urban traffic management. By identifying high-impact factors, such as car, truck and bus flows, traffic management can take more targeted measures to alleviate congestion during peak periods. The data studied in this paper only used the number of various categories of transport as features to construct a model to predict traffic congestion. However, the causes of traffic congestion are complex and variable, involving weather conditions, traffic control measures and so on, and the influence of these variables not included in the analysis on traffic congestion cannot be ignored. Therefore, this study needs to further acquire more other influencing factors to enhance the robustness and generalization ability of the model. Through comparative analyses, this paper verifies that the tree model-based classifier works well in the task of traffic congestion prediction, which illustrates the nonlinear association between features and prediction targets, as well as the complexity of the traffic congestion phenomenon. This requires a strong nonlinear fitting capability of the prediction model to capture the complex patterns implied in the data for more accurate predictions. Given the performance of the tree model in this study, the next step will be to introduce the integrated model of XGBoost to further improve the prediction accuracy of the model.

References

[1] Li Gang. Analysis of traffic congestion characteristics and improvement countermeasures on urban roads in Hohhot [J]. Traffic and Transport, 2024, 37(S1): 130-133.

[2] Zhao Jinlan. The role and mitigation strategy of intelligent transport system in traffic congestion based on big data algorithm analysis [J]. Electronic Components and Information Technology, 2024, 8(05): 86-88.

[3] Zheng Shujian, Yang Jingfeng. Research on the calculation method of traffic congestion evaluation index at home and abroad [J]. Highway and Motor Transport, 2014(1): 57-61.

[4] Lu Erkan, Rao Yao, Chen Zian. Analysis of traffic congestion characteristics of large cities in Zhejiang Province based on entropy weight TOPSIS method [J]. Transportation Science and Management, 2023, 4(21): 25-28.

[5] Tang Yi,Liu Weining, Sun Dihua, et al. Application of improved time series model in short-time traffic flow forecasting on highways [J]. Computer Application Research, 2015, 32(1): 146-150.

[6] Huang Helie, Cai Yanguang, Cai Hao, et al. Traffic flow clustering algorithm based on maximum deviation similarity criterion [J]. Computer Application Research, 2018, 35(8): 154-164.

[7] Bao Zhikang, Chen Jixuan, Liu Yinxiao, et al. Research on DNA sequence classification based on machine learning [J]. Biochemistry, 2024, 10(03): 20-27.

[8] Huang Yifeng, Wang Zhanhai. Research on civil aviation accident class classification model based on improved raccoon algorithm optimised random forest [J]. Journal of Civil Aviation, 2024,8(04): 109-113

[9] Li Sanjiang, Jiang Zhilin, Lin Yiwei, et al. A dynamic assessment method of metro communication system health based on random forest algorithm [J]. Urban Rail Transit Research, 2024, 27(06): 265-269+275.

[10] Hu Tingrui, Gao Yifan, Chen Li, et al. Analysis of GIS spatial analysis course assessment based on decision tree [J]. Geospatial Information, 2024, 22(08): 129-132.