

Research on the Rules and Prediction of British Traffic Accidents Based on Data Mining

by

Chen Tao

Supervisor: Nik Whitehead

Project submitted as part of the requirements for the degree of master of computing

Faculty of Architecture, Computing and Engineering

University of Wales Trinity Saint David



November 2021

Declaration

I, Chen Tao declare that I am the sole author of this project; that all references cited have been consulted; that I have conducted all work of which this is a record, and that the finished work lies within the prescribed word limits.

This work has not previously been accepted as part of any other degree submission.

Signed Chen Tao

Date 2021.11.7

Form of Consent

I, Chen Tao hereby consent that my project, submitted in candidature for the BEng Computer Systems and Electronics degree, if successful, may be made available for inter-library loan or photocopying (subject to the law of copyright), and that the title and abstract may be made available to outside organisations.

Signed Chen Tao

Date 2021.11.7

Acknowledgements

Time flies, and the life of graduate students in the UK is coming to an end. Thinking back to the online class in September last year, it seemed to be right in front of my eyes. During this year, I have not only learned professional knowledge, but more importantly, I have been in contact with more outstanding teachers and classmates. What I learned from them has benefited me a lot. I am especially interested in these lives. Thank you to the distinguished person.

First of all, I would like to sincerely thank my mentor nik. Nik's profound professional knowledge, rigorous scientific research attitude, and positive outlook on life affect me all the time and push me to the next level. In life, Nik is also like a loving mother, always caring about the various problems that I encounter in life. For example, during the epidemic, Nik always cares about my health.

Secondly, I would like to thank my classmates for their help. Exchanging professional knowledge with them allows me to understand the content of the class more quickly, and also makes the ordinary study life more interesting.

Finally, thank my parents for caring and taking care of me all the time. Thank my girlfriend, thank you for your understanding and support, which gave me the motivation to move forward.

Abstract

Based on British traffic accident data, this paper conducts research from three aspects: traffic accident visualization analysis, traffic accident cause rule mining and severity prediction. By mining the potential rules of the cause of the accident and predicting the severity of the accident, the corresponding early warning strategies are formulated for the driver. Provide accident warning information and suggestions with traffic management departments.

First, carry out statistical analysis on the temporal and spatial characteristics and influencing factors of accidents at home and abroad. From a statistical perspective, analyze the time characteristics and spatial layout of traffic accidents in the UK, and proceed from people, vehicles, roads and the environment to explain the theoretical mechanism of the factors that affect the accident, and analyze the impact of multiple factors on the accident. It provides a basis for the selection and division of attribute factors in the early warning of accident cause rules and the early warning of accident severity.

Secondly, the Apriori algorithm is used to mine the accident cause rules and formulate an early warning strategy for the accident rules. The results show that the association rules obtained by the Apriori algorithm are similar to the results of data visualization, and the conclusions that have not been obtained in the data visualization process are mined. According to the rules obtained, an early warning strategy of

accident rules is formulated to provide drivers with early warning information of potential rules leading to accidents.

Next, build three traffic accident severity prediction models and compare their performance. Through data balancing processing, the prediction accuracy of the model is improved. Finally, it is concluded that the performance of the random forest model after data balancing is the best.

Keywords: Data mining, Association rules, J48, Bayes, Random forest

Contents

Declaration	i
Acknowledgements	ii
Abstract	iii
1 Introduction	1
2 Background research / Literature review	3
2.1 ANALYSIS OF ACCIDENT FACTORS.....	3
2.2 DISTRIBUTION OF TRAFFIC ACCIDENTS	9
2.3 ASSOCIATION RULES	13
2.3 FORECAST OF THE SEVERITY OF TRAFFIC ACCIDENTS.....	18
2.3.1 <i>Research status</i>	18
2.3.2 <i>Bayesian analysis</i>	22
2.4 SUMMARIZE	24
2.4.1 <i>Analysis of accident factors</i>	24
2.4.2 <i>Association rules</i>	24
2.4.3 <i>Forecast of the severity of traffic accidents</i>	25
3 Research Methodology	25
3.1 RESEARCH METHODOLOGY	25
3.1.1 <i>Philosophy</i>	25
3.1.2 <i>Strategy</i>	26

3.1.3	<i>Approaches</i>	26
3.1.4	<i>Data Collection & Evaluation</i>	28
3.1.5	<i>Ethical Issues, Risks & Limitations</i>	29
3.1.6	<i>Reliability, Validity & Generalisability</i>	29
3.1.7	<i>Development Methodology</i>	30
3.2	RESOURCES	31
3.3	PROJECT PLANNING	31
4	Experiment / Testing	32
4.1	DATA PRE-PROCESSING	32
4.2	VISUALIZATION FOR THE DISTRIBUTION CHARACTERISTICS AND INFLUENCING FACTORS OF TRAFFIC ACCIDENTS ..	33
4.2.1	ACCIDENT TEMPORAL AND SPATIAL DISTRIBUTION	34
	CHARACTERISTICS	34
4.2.2	<i>Analysis of accident factors</i>	38
4.3	ANALYSIS OF ACCIDENT CAUSE BASED ON APRIORI ALGORITHM	47
4.3.1	<i>secondary pre-processing</i>	47
4.3.2	<i>Association rules</i>	50
4.3.3	<i>Classification models</i>	56
5	Results and discussion	66
5.1	DATA VISUALIZATION	66
5.1.1	<i>Accident temporal and spatial distribution characteristics</i>	66
5.1.2	<i>Analysis of accident factors</i>	68
5.1.3	<i>Recommendations for preventing traffic accidents</i>	72
5.2	ASSOCIATION RULES	73
5.2.1	<i>Slight traffic accident</i>	74

5.2.2	<i>Serious traffic accident</i>	75
5.2.3	<i>Fatal traffic accident</i>	75
5.2.4	<i>Advice on preventing traffic accidents</i>	76
5.3	COMPARISON OF CLASSIFIERS.....	77
5.3.1	<i>J48 algorithm</i>	77
5.3.2	<i>Random forest algorithm</i>	78
5.3.3	<i>Bayes algorithm</i>	80
6	Conclusions	81
7.1	RESEARCH SUMMARY	81
7.2	SUGGESTIONS FOR FURTHER WORK	83
7	Reflection.....	83
7.1	DIFFICULTY.....	83
7.2	SURPRISE	84
7.3	REWARDS	85
8	References	86

1 Introduction

With the development of smart cities, smart, efficient, safe and green transportation has become the primary goal of urban residents. Travel safety, as one of the most important parts, has also received a lot of attention. Road traffic safety and safe travel are key issues in social development. According to the latest research, safety is an indispensable part of the development of smart cities, and the safety of road users is also the most important consideration in the development of smart cities [1].

The World Health Organization has released the following important fact: Approximately 1.35 million people die in road traffic accidents every year. In other words, 1 person will die in a traffic accident every 25 seconds. The damage caused by road traffic accidents caused most countries to lose 3% of their GDP. Road traffic accidents are also the main cause of deaths of children and teenagers [2].

It can be seen that road traffic injuries have brought huge economic losses to individuals, families and the entire country. There are huge differences in road traffic death rates in various regions. The risk of death from road traffic injuries is highest in Africa (26.6 deaths per 100,000 people), and the lowest in Europe (9.3 deaths per 100,000 people) [2].

Although motorization has improved the lives of many individuals and societies, these benefits have also brought more costs. Although the number of deaths due to road traffic accidents in high-income countries has declined in recent decades, the burden of road traffic injuries for most of the world's population is rising sharply in terms

of social and economic costs [3]. In a developing country, injuries and deaths caused by road traffic accidents are a major public health problem. It is predicted that by 2021, the number of road deaths is expected to increase to nearly 2 million per year. In addition, among the leading causes of death and disability worldwide, road traffic accidents are expected to rise to the third place (following ischemic heart disease and unipolar major depression) [4].

At present, the early warning methods for vehicle accidents mostly use sensors, cameras and other equipment to obtain surrounding vehicle information and road information, and make warning prompts for the vehicle before the collision and lane change. The potential rules of accident occurrence can be obtained in advance from traffic accident data, and early warning information of accidents can be provided to the driver from a macro perspective, which makes up for the lack of time in early warning using cameras and other equipment [5]. Through the above analysis, we propose a vehicle early warning method that includes the rules of the cause of the accident and the severity of the accident. By digging out the potential factors that lead to accidents, improving existing algorithms, we can obtain more effective association rules for the cause of accidents, and predict the severity of the accident after it occurs, and formulate corresponding early warning strategies before the occurrence of vehicle accidents. Provide drivers with early warning information, improve their driving vigilance, and then play a role in enhancing driving safety, which has good theoretical significance and high application value.

2 Background research / Literature review

The research on the early warning of vehicle accidents has always been a hotspot and focus in the field of road traffic and vehicle safety. Early warning of vehicle accidents on urban roads refers to the use of statistical reasoning and data mining methods to study the rules governing the impact of various factors such as people, vehicles, roads, and the environment on vehicle traffic accidents, and then to grasp the rules of the occurrence and development of traffic accidents, through identification The safety status of the current operating environment of the vehicle, and the prediction of whether the vehicle will have an accident in the future, so as to provide the driver with early warning information through a certain early warning method before the accident, and minimize the risk of the accident, or the shortest after the accident Reduce accident damage within time[6].

2.1 Analysis of accident factors

In the early studies, scholars from all over the world generally believed that traffic accidents were mainly caused by human causes. Guang nan Zhang [7] conducted in-depth research on driver behavior and found that the driver's violation of traffic laws would increase the probability of traffic accidents. An immediate implication is: if the traffic violation rate could be reduced or controlled successfully, then the rate of serious injuries and fatalities would be reduced accordingly. Shinar [8] conducted in-depth

research on the factors that affect whether drivers use seat belts and found that the education level and age of the motor vehicle driver will affect the use of seat belts.

With the deepening of the research on traffic accidents, scholars have found that traffic accidents are not only caused by human factors, but also factors such as car conditions, road conditions, and changes in the traffic environment can also affect the occurrence of accidents. Therefore, the research on the influencing factors of traffic accidents has gradually expanded from the earliest human factors to factors such as people, vehicles, roads, and the environment.

Li-Yen Chang et al [9]. established a classification and regression tree (CART) model to analyze the traffic accident data that occurred in Taipei in 2001 and found that driver factors, vehicle parameters, and road environment have a significant impact on the severity of traffic accidents. The results indicate that the most important variable associated with crash severity is the vehicle type. Pedestrians, motorcycle and bicycle riders are identified to have higher risks of being injured than other types of vehicle drivers in traffic accidents.

In order to study the characteristics of factors affecting the severity of highway traffic accidents, Rongjie Yu et al [10]. established a Logit model and a support vector machine model, and selected factors such as weather conditions, temperature, road slope values, and driving speed for analysis. The analysis results showed bad weather and traffic conditions will affect the severity of highway traffic accidents.

At the same time, many countries have invested a lot of manpower and material resources, and established accident databases and specialized research institutions to compare and analyze the influencing factors that lead to traffic accidents from different perspectives, trying to find the general rules and main factors of accidents.

In order to study the differences in the characteristics of the influencing factors of traffic accidents on urban roads and suburban roads, Theofilatos [11] analyzed the road traffic accident data published by Greece and found that factors such as the driver's driving age, urban road intersections and non-motorized vehicles are urban roads. The main influencing factors of traffic accidents, weather conditions are the main factors of traffic accidents on suburban roads.

In order to study the impact of different road conditions and traffic conditions on the frequency of accidents, Aidoo [12] conducted in-depth analysis of the collected traffic accident data and found that the road alignment, weather conditions and road lighting conditions would affect the frequency of traffic accidents.

Jianfeng Xi et al [13]. built an analysis model of the influencing factors of highway rear-end accidents based on data on highway accidents in China from 2010 to 2015, and analyzed the effects of driving age, weather, accident time, visibility, fatigue, and driving speed on highways. The specific impact of a rear-end collision.

Zhiyuan Sun et al [14]. used binary logistic regression analysis and regression tree analysis methods to analyze the factors affecting the severity of road traffic accidents in Beijing. The results of the study found that there are differences between different

urban functional areas. The type of accident, cross-sectional location, lighting conditions, visibility, signal control, road physical isolation facilities, time of occurrence, and road type have a significant impact on the severity of the accident.

Yuhuan Zhang et al. [15], based on data from road traffic accidents in Shenzhen, applied geographic detection methods to analyze various factors affecting accidents and casualties in the area, and found that the cause of the accident and the type of the main responsible party have the greatest impact on death and injury accidents, followed by regional and time interval. However, road factors, lighting, terrain, etc. only have an impact on the death toll.

In general, the analysis of factors affecting traffic accidents has changed from single-factor analysis to multi-factor analysis. However, the systematic analysis of traffic accidents is not comprehensive enough, ignoring the impact of the interaction between traffic accident factors on the accident. Therefore, the correlation between the influencing factors of the accident still needs further study.

In recent years, Chinese scholars have conducted in-depth analysis of traffic accident data over the years by learning from foreign research methods, focusing on the impact of various factors on the occurrence of traffic accidents.

In order to analyze the impact of various factors on the characteristics and development trends of highway traffic accidents, Zhao Jinbao[16], conducted in-depth analysis of the collected traffic accident data, taking into account factors such as people,

vehicles, roads, and the environment, and finally found that the highway The main factors of traffic accidents are location, weather, time, and vehicle type.

Xu Dejie et al [17]. studied the impact of human, vehicle, road, and environmental factors on major traffic accidents. Through analysis of the 2005-2010 major accident data released by Gansu Province, they finally found that the main cause of such accidents was human factors.

In order to study the impact of various factors on the severity of traffic accidents, Yuan Quan [18] selected some factors from the factors such as people, vehicles, roads and the environment to analyze. The research showed that bad weather and road environment have a significant impact on the severity of traffic accidents.

Zhang Yuan [19] selected the binomial Logit model to analyze the factors that affect the severity of traffic accidents of drivers of different genders. The model analysis results show that whether the driver wears a seat belt, whether the vehicle is insured, the attributes of road safety, the type of intersection, and the lighting conditions have significant effects on male and female drivers.

At the same time, Chinese scholars have also used Bayesian models, multiple regression models, and correlation analysis methods to conduct in-depth analysis of accident influencing factors and have achieved remarkable results.

Xu Hongguo[20]constructed a Bayesian network to analyze the causes of traffic accidents. The analysis results show that the addition of traffic facilities such as signal

lights and traffic signs has a significant impact on reducing the severity of traffic accidents.

Zhu Jiazheng [21] conducted research on the influencing factors of road traffic accidents in the construction area, constructed a number of logistic regression models, and used classification regression trees and other methods for analysis. Finally, it was found that the main reasons leading to road traffic accidents in the construction area were illegal driving, bad weather conditions and road conditions without physical separation.

In order to explore the influence of various factors on the severity of highway tunnel traffic accidents, Ma Zhuanglin [22] built a neural network model for analysis. The research results showed that the main factors affecting the severity of highway tunnel traffic accidents were the proportion of large vehicles, road conditions and weather conditions. The influence of other factors is relatively small.

Chen Yanyan and Li Xiangnan [23] selected a binary Logistic regression model to analyze the factors affecting the severity of road traffic accidents in megacities and their temporal and spatial distribution characteristics, and found that visibility, lighting conditions and traffic signal methods have a significant impact on the severity of traffic accidents.

Based on historical data, Fan Xiantao [24] used the Bayesian model to analyze the impact of three typical unfavorable weather of rain, snow, and fog on highway traffic accidents in Shanxi Province, and obtained the main causes of highway traffic accidents

in mountainous areas. The mountainous expressway traffic management department provided valuable advice on how to reduce the impact of adverse weather on traffic accidents.

Hu Ji and Yan Zhangcun[25] constructed an ordered Logit and Probit model, and applied the model to analyze the factors affecting the severity of traffic accidents. The results show that factors such as whether seat belts are worn, road alignment, and road surface type have a significant impact on the severity of the rollover accident.

Zeng Qiang and Wang Xuesong[26] selected a time-space interaction model to analyze the influence of road design parameters, traffic characteristics, and meteorological factors on the number of seasonal accidents in the fitted road section. They found that the proportion of vehicle types, road section curvature, slope, wind speed, precipitation and other factors have significant influence on the accident frequency.

From the above analysis, it can be seen that there are various causes of traffic accidents, and changes in the road environment and road users will all have an impact on traffic accidents. Therefore, when conducting an in-depth analysis of traffic accidents, specific problems must be analyzed in detail, and the impact of various conditions must be fully considered.

2.2 Distribution of traffic accidents

In order to study the relationship between the occurrence of traffic accidents and road design parameters, Ruediger et al [27]. analyzed two-lane highways. The study

showed that different road changes have a significant impact on the law and trend of traffic accidents.

In order to study the law of the incidence of rural road traffic accidents, Ponte G et al [28]. analyzed the collected rural road traffic accident data in Adelaide, Australia, and found that the increase in vehicle speed would have a significant impact on the occurrence of traffic accidents.

Murat Selim Çepni[29] used GIS tools to analyze the relationship between urban road accidents and road infrastructure, environment and traffic related variables, and obtained the distribution of accidents in Kocaeli city.

Hayidso Temesgen Haile [30] used GIS to analyze accident data in Hosanna, Ethiopia, and obtained nine areas where road traffic accidents occur frequently.

El-Said MM Zahran [31] used ArcGIS software to analyze the network kernel density estimation method, the GetisOrd GI method, and the risk-based method, and analyzed the application of three different hot spot analysis methods in identifying and sorting traffic accident hot spots. The relative advantages and disadvantages of this method in hot spot analysis.

In order to study the distribution of road traffic accidents in my country, based on the data of road traffic accidents in my country from 2007 to 2011, Gao Tianzhu [32] deeply analyzed the distribution of road traffic accidents in my country in time, space, types of deaths, traffic accident patterns, and different climates. The management of the accident provides a theoretical basis.

Based on principal component analysis and fault tree analysis, Ding Yulei [33] compares the differences in the distribution of major and general traffic accidents in my country, and provides a reference for the prevention of major and general accidents based on the analysis results.

In order to study the influencing factors of urban traffic accident rate, Cai Jun [34] selected road grade, road section form, road network density, road intersection form, road speed, etc. for analysis. The research results show that the wide and sparse urban road network, lack of branch roads, and unclear street functions have a significant impact on the high traffic accident rate in China.

In order to test the distribution of hotspot sections and temporal-spatial interaction patterns of road traffic accidents in Qingshan District, Wuhan, Wu Ruilong [35] selected the network space linear nearest neighbor index Knox spatiotemporal test method to analyze traffic accident data in this region over the years. The analysis results show that there are obvious aggregation and distribution patterns in time and space for road traffic accidents in Qingshan District, Wuhan.

In order to analyze the distribution of traffic accidents on suburban roads, Guo Xiaokui[36] selected the kernel density analysis method to calculate the density of traffic accidents in different time periods, seasons and locations, and obtained the distribution characteristics of road traffic accidents in the suburbs of the study area.

Based on the traffic accident data collected by the mobile phone App method, Guo Lin [37] used the improved K-means clustering algorithm to analyze the temporal and

spatial distribution characteristics of traffic accidents in the study area, and found black spots of road traffic accidents in the study area.

Based on three spatial analysis methods of improved buffer analysis, overlay analysis and nuclear density clustering, Wang Hai [38] analyzes the areas and causes of traffic accidents. The results show that the improved method can more effectively identify accident-prone locations and cause analysis than traditional methods.

Lu Huapu [39] analyzed the road traffic accident data in Shenzhen from 2014 to 2016, and calculated whether to consider the nuclear density value under the two conditions of road network density, and obtained the spatial distribution of road traffic accidents in Shenzhen under the two conditions.

From the above research, we can see that the research on the distribution law of traffic accidents is mainly based on the statistical data of accidents to study the law in time and space.

It can be seen that starting from factors such as people, vehicles, roads, and the environment, studying the causes of accidents is the mainstream research direction. However, the focus of the above-mentioned research is on single factor. Although it can be compared vertically within the same factor, it is still too one-sided and lacks the horizontal correlation between different influencing factors. However, the causes of traffic accidents are often combined by a variety of factors and influence each other. It is difficult to find the cause of accidents from many complicated traffic characteristics. With the advancement of data mining and machine learning technology, analysis

methods represented by association rules have increasingly entered the field of traffic safety research.

2.3 Association rules

Association rule mining is an important data mining method, mainly to find the implicit association rules in the data set. It is also called shopping basket analysis. It was first proposed by American expert Agrawal [40] in 1993. Because this method is named after the sales case of beer and diapers in Wal-Mart supermarkets at that time, it is called shopping basket analysis, which is also a classic application scenario of association rules. At that time, Wal-Mart merchants used data mining methods to analyze a large amount of detailed data in order to find customers' purchasing habits. Unexpectedly, beer and diapers were purchased together the most often. After investigation, it was found that young fathers were mainly buying diapers at that time. They always buy some beer after buying diapers. This conclusion also provides a basis for merchants to display their products.

Now we are in an era of big data development. In recent years, more and more experts and scholars have begun to conduct in-depth research on data mining and apply it in practice. After years of research, various fields have verified the association rule algorithm, and produced good social and economic value. In the field of traffic safety, it is mainly used to analyze the causes of traffic accidents, and to find hidden causes of accidents and combined risk factors of accidents through analysis.

Commonly used association rule algorithms mainly include: Apriori algorithm [41], FP growth algorithm [42], and Grey Relational Analysis [43].

Yu et al. [44] used the Apriori algorithm to dig out the important association rules between the authenticity of the accident and the influencing factors. Compared with previous studies, the accident variables are included more comprehensively, including environment, management level, and driver and vehicle status.

Faisal et al. [45] used the Apriori algorithm and the EM clustering algorithm to mine the potential causes of accidents in KSA's Alghat-Provence traffic accident data. Since the EM clustering algorithm requires an unquantified number of iterations to get an optimal solution, the Apriori algorithm can find All frequent itemsets outperform the EM algorithm in terms of time and results.

Fatima et al. [46] used the Apriori algorithm to analyze the French traffic accident data, and sorted the obtained association rules through the ELECTRE method, revealing the conclusion that there is a narrow correlation between the driver and the pedestrian, the normal road surface and the normal atmospheric conditions, and the road surface and the pedestrian. And choose the higher-ranking accident rule as the criterion for accident prevention.

Hu Zhenbo [47] combined the idea of weighting with association rules and applied it to traffic accident data. The information entropy theory calculates the weights of the factors that may affect the accident, and proposes a dynamic FP-Growth algorithm, adding weighted support and confidence vectors. Describe the relationship between

association rules and time changes, find the most applicable association rules for each time period, and analyze the impact of various factors on traffic accidents.

In view of the characteristics of tunnel traffic accident data, Liu Yunxiang[48] also improved the FP-Growth algorithm, assigning different weights to each attribute, and obtaining potential association rules more efficiently, making the results meaningful, and using the improved WFP-Growth algorithm to establish The mining model is used to mine the tunnel traffic accident data to find out the combination of frequent factors that cause the tunnel traffic accident, and find the potential relationship between the factors that cause the tunnel traffic accident.

Chen Jingxi[49] started with single-factor, multi-factor, and multi-category accidents, selected PCA feature extraction algorithm, Apriori algorithm, and K-Means clustering algorithm for different analysis levels to establish a comprehensive analysis model, and statically sampled the data to improve mining efficiency.

Deng Xiaoheng et al. [50] used the analytic hierarchy process to reduce accident attributes, combined association rules and genetic algorithms, optimized the efficiency of association rule mining, and obtained more effective accident rules.

Aiming at highway traffic accidents under non-extreme severe weather conditions, Jing Liantian[51] uses association rules and decision trees to explore the relationship between the factors affecting the accident.

Yang Jianhong[52] is based on the principle of the apriori algorithm. By adding filtering conditions, the time complexity of the algorithm is controlled to $O(n)$ to solve

the problem of low efficiency of the original algorithm. Finally, the star model is constructed and the data mining experiment is performed using the weka tool. The correlation and causality of each feature are deeply explored, and relevant conclusions can be produced that can effectively reduce the occurrence of accidents and reduce the severity of accidents. For example, male drivers who drive non-motorized vehicles are most likely to run red lights between 6 and 11 o'clock.

Pande and Abdel-Aty[53] conducted a systematic analysis of safety accident cases that occurred on a straight road in a state in the United States. Through the analysis of the links between different influencing factors and traffic accidents, they found that the main factors leading to road traffic accidents is that road lighting conditions are insufficient.

Liling Li et al. [54] used the heuristic association rules of the Apriori algorithm to establish a classification model with a naive Bayes classifier, and used a simple k-means clustering algorithm to generate clusters. On the basis of statistics, association rules, classification models and clustering, some methods and suggestions that can effectively prevent road traffic accidents are put forward.

In the process of researching traffic accident data in an urban area for the past three years, Dong Liyan[55] effectively integrated association analysis and rough sets, and proposed a simplified algorithm for decision rules based on preference information, which greatly improved the data mining experiment. operation efficiency.

Shangwei[56] from Fudan University effectively integrated the original traffic accident case records of a highway, classified and integrated various factors of the traffic accident, and transformed the data to correspond to different fields, and finally formed a new database. Then analyze the recorded related data according to the multi-dimensional association rules, and obtain the complex cross-relevance between the multi-dimensional elements.

Zhang Xin [57] from Tongji University used the optimization algorithm of cluster analysis method to obtain a new feasibility study plan, which was applied to the cluster study of road traffic accidents, and effectively analyzed the weight values of different attribute factors.

In the process of studying the black spots of traffic accidents on a certain section of highway, Wu Hao [58] classified and defined the differences in the attributes of road traffic accidents. Through modeling and combined with the improved Apriori algorithm, the historical data of traffic accidents were analyzed. Through the analysis, it provides decision-making assistance and technical support for relevant units and users to find road black spots (accident-prone points).

Weng et al. [59] used association rules to analyze marine transportation accidents. Based on marine transportation data collected from 2004 to 2014, they found out the relationship between natural conditions, external environment and other factors that caused the accident.

Based on multi-dimensional and multi-layer association rules, Hong Yuanyuan [60] summarized the regular characteristics of low-grade highway traffic accidents and the combination of attributes that easily lead to accidents.

Zhou Jin[61] proposed the MC_Apriori algorithm for fast mining of frequent itemsets, analyzed the 203 accidents published by RAIB, obtained the accident cause data set, and found out the association rules of the railway transportation accident cause.

It can be seen from the above research that searching for potential rules hidden among the influencing factors of traffic accidents through association rules is one of the commonly used methods to solve the cause analysis of accidents. One of the most used is the apriori algorithm.

2.3 Forecast of the severity of traffic accidents

For the related research on traffic accident prediction, it is not only necessary to predict whether the accident will occur, but also to predict the severity of the accident after it occurs. Therefore, the predictive modeling and analysis of the severity of the accident has received great attention from scholars.

2.3.1 Research status

Early research methods for the severity of traffic accidents were mainly based on binary regression, including Logit and Probit models, focusing on the analysis of the impact of accident factors on the fatality rate. Kim et al. [62] used the Logit regression model to analyze the traffic accident data in North Carolina, USA, and studied the related

factors of four different degrees of traffic accidents through control variables. The results showed that under severe weather and poor light conditions, alcohol Reasons such as rear driving, speeding and frontal collision can increase the probability of death after an accident by more than two times. Speeding is the most serious. If the vehicle speed exceeds 80.5km/h, the probability of death increases by more than 16 times.

Liu Haizhu[63] divided the traffic accidents in Changchun City into three severity levels, introduced direct economic property loss factors, and established a road traffic severity prediction model through the cumulative Logistic regression model. The results showed that the model has good predictive capabilities.

With the in-depth study of traffic safety theory, the analysis of accident fatality rate alone cannot meet the safety requirements. In order to avoid the limitations of the accident severity classification model, the multiple regression model has gradually become a research tool favored by scholars, including the MNL model, OL model and ordered corresponding model, etc.

Zhao et al. [64] compared the MNL model with other regression models, and found that the MNL model has a better degree of fit for traffic accidents involving vehicle collisions, but requires that each category of accident severity be independent of each other, based on the degree of property damage and injury. The accident classification is difficult to do.

Celika and Oktay[65] used the MNL model to analyze traffic accidents in Turkey and found that the driver's age, education level, road conditions and weather conditions are

the main factors affecting the severity of the accident, and based on this conclusion, they made a more accurate prediction of the severity.

Kockelman et al. [66] analyzed the factors influencing the severity of traffic accidents of various types of collisions and found that compared to cars, pickup trucks or SUVs, they are more likely to cause serious traffic accidents, and the driver's injury degree is generally lower than that of the passengers. In addition, factors such as the driver's gender and age also have a significant impact on the severity of the accident.

In addition, with the development of data mining technology, machine learning has become one of the effective methods for analyzing the severity of traffic accidents by virtue of its powerful learning and predictive capabilities. It is increasingly used to study traffic accidents such as people, vehicles, roads, and the environment. A complex pattern of influence between factors and the severity of the accident.

Delen et al. [67] used artificial neural networks to simulate the potential nonlinear relationship between the severity of the accident and collision-related factors, and conducted a sensitivity analysis of the established neural network model. The results showed the driver's gender, age, vehicle use, and whether or not to drink and drive. Significantly related to the severity of the accident.

Sohn et al. [68] used the decision tree model, the Logit model and the neural network model to construct the prediction model of the severity of the accident, and found that the prediction accuracy of different models is not much different, and

wearing safety belts and other protective devices can greatly reduce the severity of the accident.

W Budiawan et al. [69] studied 501 traffic accidents on toll roads in Indonesia. According to factors such as accident date, road type, weather, road conditions, accident time, driver gender, vehicle type, etc., the naive Bayes method was used to establish the serious accidents. However, due to the small number of data samples and the large gap in the number of different types of accidents, the accuracy of the model for predicting the severity of accidents and the probability of accidents is only 39.49%.

Sun Yixuan[70] analyzes events using mining methods such as classification, regression, cluster analysis, and association rule mining and he constructed a two-class and multi-class TPMSVM classification prediction model to classify and analyze the severity of the accident, and applied feature selection to the analysis of the factors affecting the severity of the accident to filter out the core feature variables.

Wang Lei [71] used Bayesian network, random forest model, BP neural network model and SVM model to predict the severity of road tunnel accidents. The results show that the Bayesian model is the most reliable predictor of accident severity, with an accuracy rate of 78%.

Shi Xuehuai et al. [72] proposed a traffic accident severity prediction method that combines CNN neural network and Xgboost ensemble algorithm, and considers the approaching road conditions of the accident. Experimental results show that this model

has a prediction accuracy of 91.51% on the test set, which is better than a single model. Good forecasting effect.

It can be seen that machine learning algorithms such as support vector machines, random forests, neural networks, and Bayesian theory are often used to predict the severity of traffic accidents. Among them, the support vector machine has the characteristics of simple structure and good adaptability, but different kernel functions will have different effects on the accuracy of the results; integrated algorithms such as random forest have high accuracy in predicting the severity of accidents, but they are prone to overfitting phenomenon; neural network has certain limitations in the prediction of severity, local search makes it more accurate in predicting the severity of some traffic accidents, but sometimes the optimal solution cannot be obtained; Bayesian network can obtain each type of accident The distribution probability of severity, and the ability of reverse reasoning, can accurately grasp the relationship between different influencing factors, but the accuracy of prediction is often not ideal due to the large difference in accident types. It can be solved by data balancing.

2.3.2 Bayesian analysis

Bayesian network is currently a popular method to solve uncertainty problems, and it can perform probabilistic reasoning on uncertainty problems. In 1986, Professor Pearl [73] first proposed Bayesian networks. After a lot of research by many scholars, Bayesian networks gradually transformed from the original basic theoretical research into the application of reasoning to uncertain problems. In the 21st century, Bayesian networks

have been widely used in data mining [74], fault diagnosis [75], and condition monitoring [76].

The current research on bass networks is mainly divided into theoretical research and application research, both of which have achieved certain results. In theory, there are strategy optimization algorithms [77] and genetic algorithms [78] applicable to Bayesian network structure learning and parameter learning; in software applications, more mature Bayesian network software such as Hugin, Netica, and Genie have also been developed [79].

Xiaoli Ma et al. [80] used DS evidence theory and Genie software to build a Bayesian network prediction model for dangerous goods road transportation accidents, and applied the model to analyze dangerous goods road traffic accidents. Through the analysis, the complexities of dangerous goods road transportation accidents can be obtained. The logical relationship and the uncertain relationship between factors provide theoretical support for transportation companies and government departments to take effective measures to reduce the risk of road transportation of dangerous goods.

China's research on Bayesian networks and its applications is relatively late, and its application in traffic accident prediction is also in the active exploration stage.

In order to study the main causes of traffic accidents in mountainous sections of expressways, Lu Yao [81] conducted an in-depth analysis of 2,635 traffic accidents on Jingkun Expressway by constructing a Bayesian network model, and obtained the main

causes of accidents on mountainous expressways in the region. Provide a theoretical basis for traffic management departments to prevent traffic accidents.

Liu Zhiqiang[82]analyzed the causes of frequent highway accidents under haze weather through Bayesian network models, and the results showed that changes in vehicle speed and haze levels would affect the occurrence of accidents.

In order to study the relationship between the various influencing factors and events in the road traffic accident chain, Xiong Xiaoxia[83]selected the Bayesian network method to analyze the US 100-Car natural driving research data, and the research results showed that the influencing factors of different types of accidents exist difference.

2.4 Summarize

2.4.1 Analysis of accident factors

The distribution of traffic accidents in time and space has a certain degree of clustering. Through the analysis of the characteristics of accidents in time and space, it is helpful to understand the potential laws of accidents.

The influencing factors of traffic accidents are mainly divided into human factors, vehicle factors, road factors and environmental factors, and statistical analysis can be performed on the data items recorded in the data set.

2.4.2 Association rules

When choosing the association rule algorithm, considering that the FP-growth algorithm performs poorly when there are many itemsets, when faced with long

transaction data, the depth of the FP tree is very large, and the subsequent sub-problems that need to be solved have become particularly numerous. , Although it only needs to traverse the data set twice, the efficiency will drop rapidly. Although the Apriori algorithm needs to traverse the database many times, it is more common because of its concise thinking and high scalability.

2.4.3 Forecast of the severity of traffic accidents

Because of its unique advantages in reasoning analysis, Bayesian networks have been widely cited in failure analysis and risk prediction. However, the application of Bayesian networks to traffic problem analysis is relatively rare. Traffic accidents are characterized by randomness and uncertainty, so Bayesian networks are also suitable for cause analysis and risk prediction of traffic accidents.

3 Research Methodology

3.1 Research Methodology

3.1.1 Philosophy

Choosing Positivism as research philosophy can be better.

Positivism is that Knowledge is based on the observation or experience of real phenomena in an objective and real world, and objective experiments are performed to observe these phenomena.

Techniques of data mining definitely take some experiment in order to find which algorithm is suitable for our aim.

3.1.2 Strategy

1) Statistical analysis of traffic accidents in the UK. Discussed from the four factors of human, vehicle, road and environment, expounded the impact of each factor on the accident in detail, and provided a basis for the selection and classification of accident attribute factors in the early warning of accident rules and the early warning of accident severity.

2) Use the improved Apriori algorithm to formulate the early warning strategy of accident rules.

3) By building three classifiers and comparing their performance, we can verify the conclusions of the literature review that Bayes is applicable to the prediction of the severity of traffic accidents.

4) Combining the results of data visualization and association rule algorithms, a traffic rule warning strategy is proposed to reduce the number of road traffic accidents.

3.1.3 Approaches

the literature review can be used to justify chosen research approaches

3.1.3.1 Analysis of Causes of Traffic Accidents

From the literature review, it can be seen that searching for potential rules hidden among the influencing factors of traffic accidents through association rules is one of the common methods to solve the cause analysis of accidents. However, the results obtained by classic association rule algorithms often contain many redundancies. Information and a large amount of repetitive content lead to the ineffectiveness of association rules, and it is difficult to use them in actual accident prevention. Therefore, the application of association rules in the cause analysis of traffic accidents is still being explored.

This project will use the traditional association rule algorithm (apriori) to verify that the classic association rule algorithm can also be applied to the data mining project of traffic accidents.

3.1.3.2 Forecast of the severity of traffic accidents

Machine learning algorithms such as support vector machines, random forests, neural networks, and Bayesian theory are often used to predict the severity of traffic accidents. Among them, the support vector machine has the characteristics of simple structure and good adaptability, but different kernel functions will have different effects on the accuracy of the results; integrated algorithms such as random forest have high accuracy in predicting the severity of accidents, but they are prone to over-simulation. Synergistic phenomenon; neural network has certain limitations in the prediction of severity, local search makes it more accurate in predicting the severity of some traffic accidents, but sometimes the optimal solution cannot be obtained; Bayesian network

can obtain each type of accident The distribution probability of severity, and the ability of reverse reasoning, can accurately grasp the relationship between different influencing factors, but the accuracy of prediction is often not ideal due to the large difference in accident types. It can be solved by data balancing.

By comparing the performance of J48, random forest and bayes three classifiers, it is verified that bayes is the best model. However, the authenticity of this proposition is still unknown and needs to be verified in this project.

3.1.4 Data Collection & Evaluation

3.1.4.1 Data collection

What data will you collect? How will you collect it? How will you analyse it once you've got it?

The data come from the Open Data website of the UK government, where they have been published by the Department of Transport.

The dataset comprises of two csv files:

- AccidentInformation.csv: every line in the file represents a unique traffic accident (identified by the AccidentIndex column), featuring various properties related to the accident as columns. Date range: 2018-2020
- Vehicle_Information.csv: every line in the file represents the involvement of a unique vehicle in a unique traffic accident, featuring various vehicle and passenger properties as columns. Date range: 2018-2020

The two above-mentioned files/datasets can be linked through the unique traffic accident identifier (Accident_Index column).

Once we have got this dataset, we can statistically process this data and do some predictions by data mining algorithm after preprocessing there data.

3.1.4.2 Evaluation

This project mainly aims at the accuracy and recall rate of the classifier to evaluate the quality of the classifier.

3.1.5 Ethical Issues, Risks & Limitations

This project does not raise any ethical issues, because this is an open source dataset offered by UK government. There are no any risks that might cause this project to fail at present.

What are the limitations/scope of your project?

If there are traffic data sets from all over the world, then this model has more generalization ability, that is, it can be applied in the transportation field of other countries

3.1.6 Reliability, Validity & Generalisability

- Reliability

The research gives the same result if run a second time or by a different researcher

- Validity

This is an appropriate way to research this topic, which will use common algorithms for data mining, such as apriori and Bayes.

- Generalisability

Since the UK data set is used to build the model, it may be better applied to the traffic in the UK.

But, we can believe that it can have super generalization ability with enough data.

3.1.7 Development Methodology

CRISP-DM breaks the process of data mining into six major phases: [84]

1) Business Understanding

By digging out the underlying rules of the cause of the accident and predicting the severity of the accident, the corresponding early warning strategies are formulated to provide drivers and traffic management departments with accident early warning information and suggestions

2) Data Understanding

The data come from the Open Data website of the UK government, where they have been published by the Department of Transport. Therefore, we do not need to collect data by ourselves.

3) Data Preparation

We can preprocess our data before use.

4) Modeling

Algorithm : Apriori ; Bayes;J48;Random forest

Tools: Weka; Tableau prep builder; Tableau

5) Evaluation

We need to evaluate the results of models

6) Deployment

Apply the model to traffic problems

3.2 Resources

Resources Required:

Tableau desktop

Tableau prep builder

Notepad++

Weka

3.3 Project Planning

Table 1 Project Planning

Component of study	Time span
Literature review	2021.7-2021.8
Study algorithms	2021.8-2021.9
Data preprocessing	2021.8-2021.9
Data visualization	2021.9-2021.10
Build classifiers	2021.9-2021.10

Write the report	2021.10-2021.11
------------------	-----------------

4 Experiment / Testing

This section should include a description of any formal experiments that have been carried out or of any tests that have been performed to get the desired result.

4.1 Data pre-processing

This project uses a tool, tableau prep builder, to complete data preprocessing.

Step 1:

The original data set contains the UK traffic accident data for the past five years. We only need the data for the past three years (2018-2020). Therefore, the UK traffic accident data for the past three years is filtered through data processing tools.

Step 2:

Delete useless fields. The data set contains many fields that are not used in the project, such as vehicle_reference. Direct deletion is the most direct way.

Step 3:

Remove outliers. There are some unreasonable values in the data that need to be deleted. For example, the driver's age is 5 years old, which is obviously unreasonable.

Step 4:

Numerical transformation. The data in the data set is recorded in encoded form, such as the field `road_type`. When its value is 1, it represents roundabout. We need to convert the encoded value of the field into a meaningful string value. Such as 1->roundabout.

Step 5:

Combine certain values into one value. For example, minibuses and buses can be merged into bus. This operation can reduce the complexity of the classifier and facilitate our data analysis.

Step 6:

Missing value handling. Due to the large amount of data, more than 500,000 records, and not much missing value, deleting empty data is the simplest and most direct way.

After finishing the data preprocessing, import the data set into Tableau Desktop to realize the visualization of the data.

4.2 Visualization for the distribution characteristics and influencing factors of traffic accidents

The composition of the urban road traffic system is complex, and factors such as personnel, vehicles, roads and the environment interact and interact, resulting in different temporal and spatial characteristics of the accident distribution. The causes of accidents are very different, and the severity of the accidents are also different from each other. It is difficult to realize accidents through simulation. Prevention and management of Therefore, mining and analyzing the historical data of urban road traffic

accidents to find the potential information is an effective method to study traffic accident-related problems, and it can lay the foundation for the mining of accident cause rules and the prediction of serious accidents

4.2.1 Accident temporal and spatial distribution characteristics

The distribution of traffic accidents in time and space has a certain degree of clustering. Through the analysis of the characteristics of accidents in time and space, it is helpful to understand the potential laws of accidents.

4.2.1.1 Analysis of accident time characteristics

(1) Year

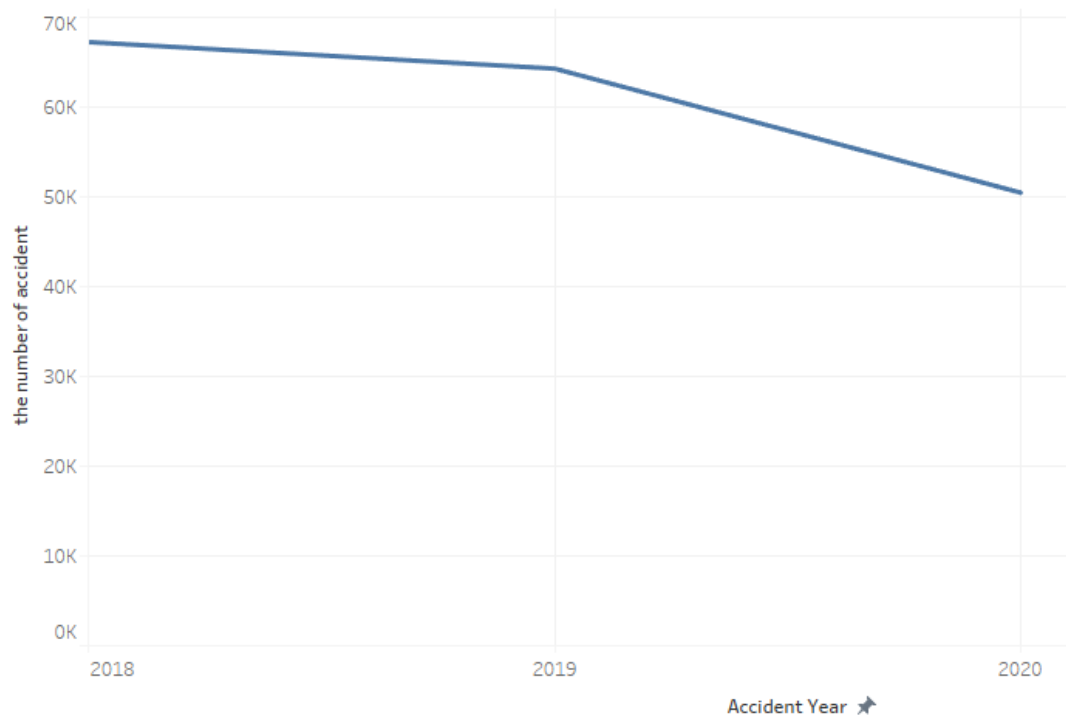


Figure 1 The relationship between the number of accidents and year

(2) Month

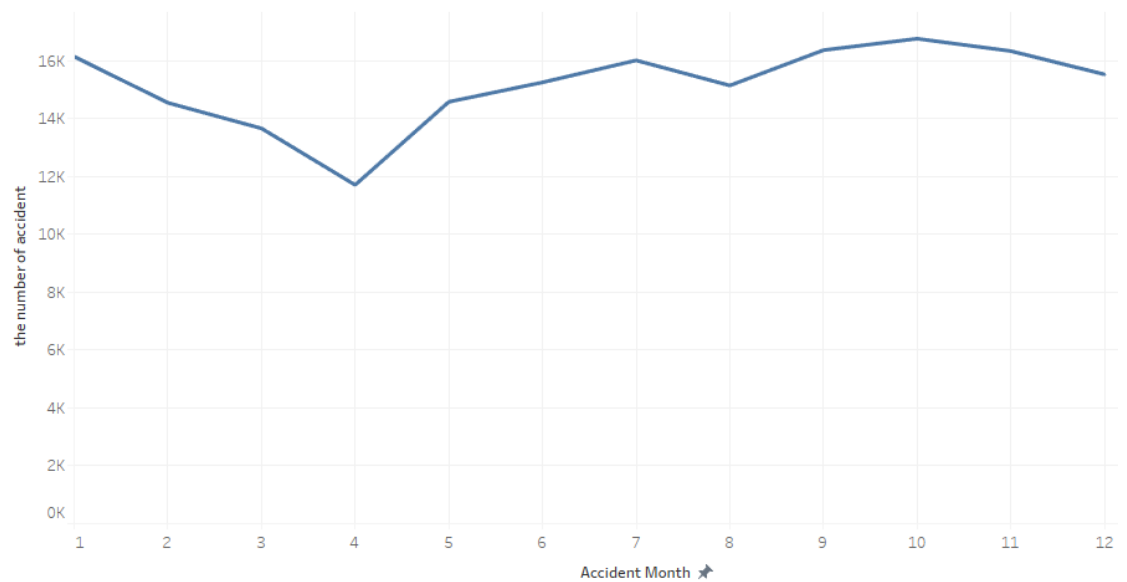


Figure 2 The relationship between the number of accidents and month

(3) Week

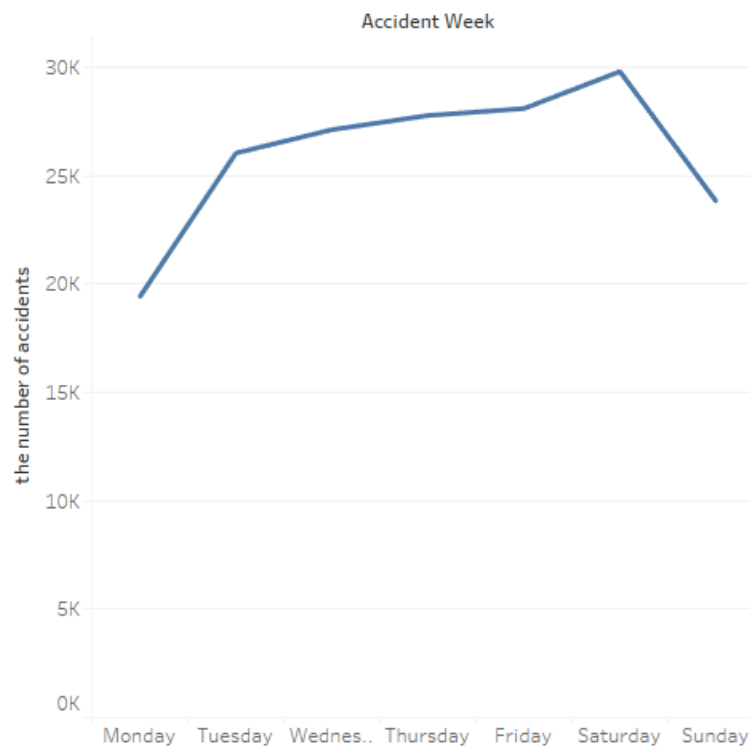


Figure 3 The relationship between the number of accidents and weeks

(4) Time

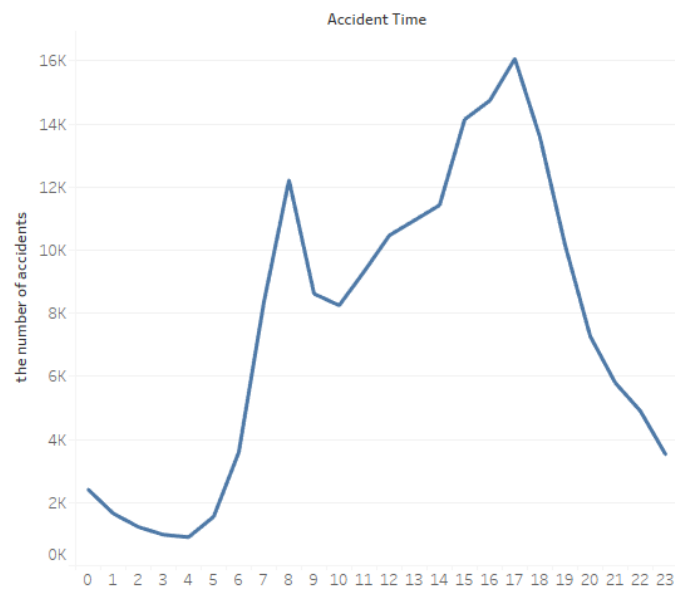


Figure 4 The relationship between the number of accidents and time

4.2.1.2 Accident spatial distribution analysis

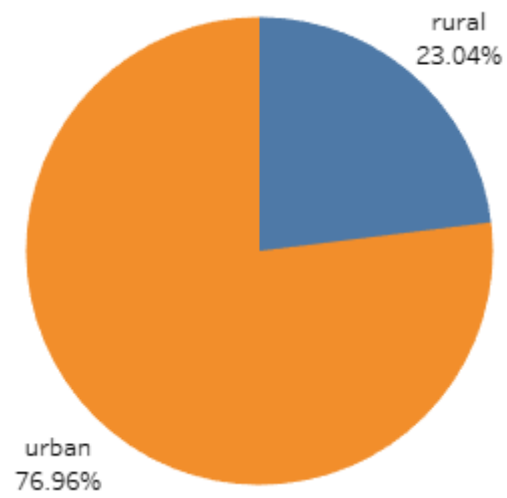


Figure 5 Frequency distribution of urban and rural traffic accidents

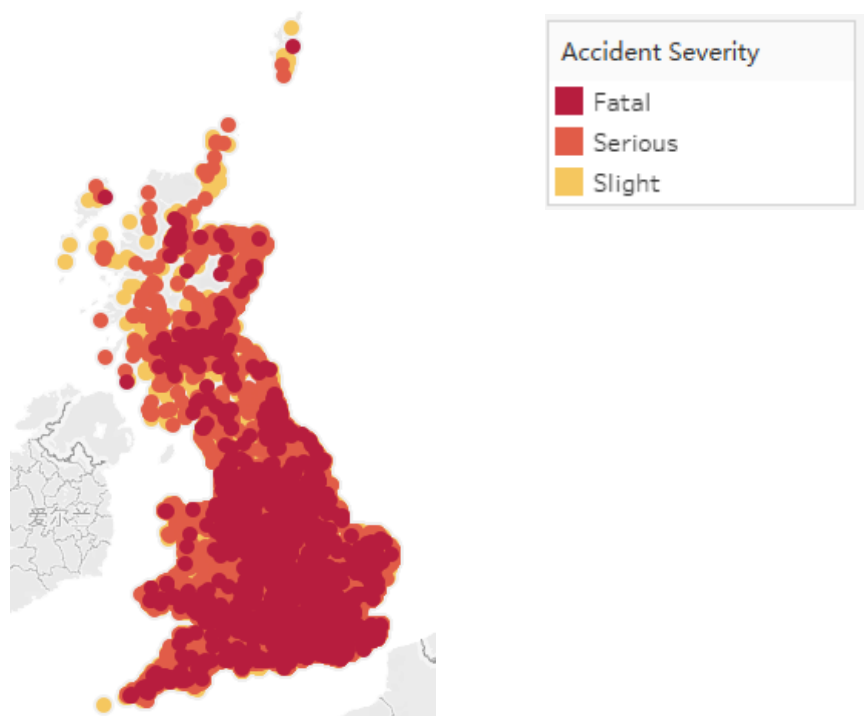


Figure 6 Spatial distribution of traffic accidents

4.2.2 Analysis of accident factors

The influencing factors of traffic accidents are mainly divided into human factors, vehicle factors, road factors and environmental factors. The following is a statistical analysis of the data items recorded in the data set.

4.2.2.1 Human factors

People are the main body of traffic and are also participants in traffic accidents. Human factors include drivers, passengers and pedestrians, among which driver factors are the main cause of traffic accidents. According to incomplete statistics, traffic accidents caused by drivers account for more than 87.5% of the total number, which requires special attention.

When a driver is driving a vehicle, it is mainly affected by two aspects: physiological factors and psychological factors. Physiological factors include objective conditions such as gender, age, and driving experience. Psychological factors include internal aspects such as the driver's mental state, personality characteristics and emotions while driving, which are often difficult to quantify.

(1) Gender

The driving performance brought about by gender differences of drivers is mainly reflected in the handling habits and emergency response. Female drivers often have minor accidents due to unskilled handling, such as scratches caused by forgetting to

close the door; male drivers have relatively long driving experience, so they have relatively rich driving experience, but they often experience aggressive driving. For example, speeding, drunk driving and other undesirable or even illegal phenomena often occur among male drivers. The probability of car accidents does not decrease but increases, and they are more likely to cause more serious traffic accidents.

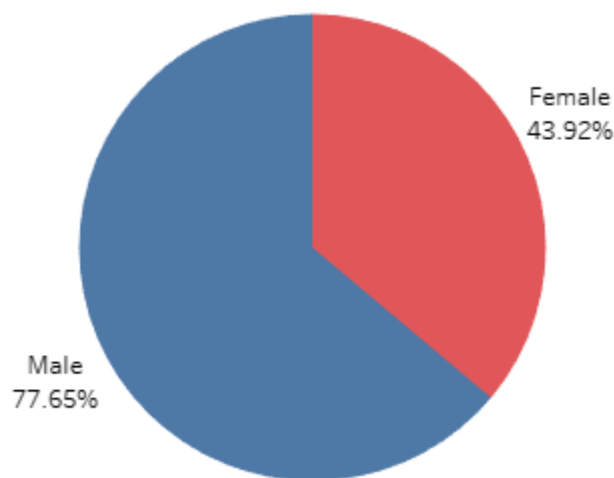


Figure 7 Sexual distribution of traffic accidents

(2) Age

The influence of age on the driver's driving vehicle is mainly reflected in the aspects of vision, reaction ability and experience accumulation. Starting from adulthood at the age of 18, with the increase of age, the body's ability to respond to emergencies reaches its peak between 20 and 30 years of age. It has a good ability to deal with emergencies that require a short reaction time, but in the age group, some people are radical and reckless and they have a tend to drive dangerously. After that, the body response of the

human body begins to decline, and the driver is more inclined to avoid danger, but the accumulation of driving experience makes it easier for vehicle manipulation under extreme conditions; after the age of 50, the human body's His eyesight and responsiveness declined rapidly, and his driving style became more cautious. The age distribution of drivers is shown in Figure 8.

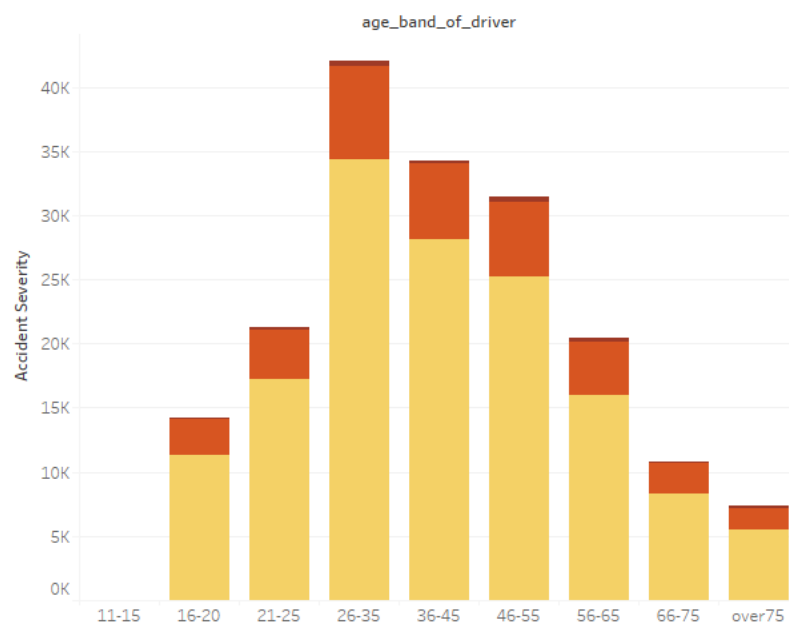


Figure 8 The age distribution of drivers

4.2.2.2 Vehicle factors

Vehicles are the carriers of drivers and passengers. They not only protect personnel in traffic accidents, but may also cause casualties. Liu Bohai's research [85] on traffic accidents in Shenzhen found that among all the accidents, the proportion of inter-vehicle accidents was as high as 83.78%, and the proportions of electric vehicles and

bicycles were 4.29% and 4.73%, respectively. Therefore, the accidents on urban roads are still vehicle-to-vehicle accidents. Inter-collision is dominant.

The factors that affect the vehicle are divided into the safety performance of the vehicle itself and the interaction between the vehicles. There are fewer accidents caused by problems with the safety performance of the vehicle itself, but they are often more serious. When the vehicle is running, the steering system, braking system, electrical system, etc. are coordinated and coordinated. Any problem in any link may cause the driver to make operation errors or even lose control of the vehicle.

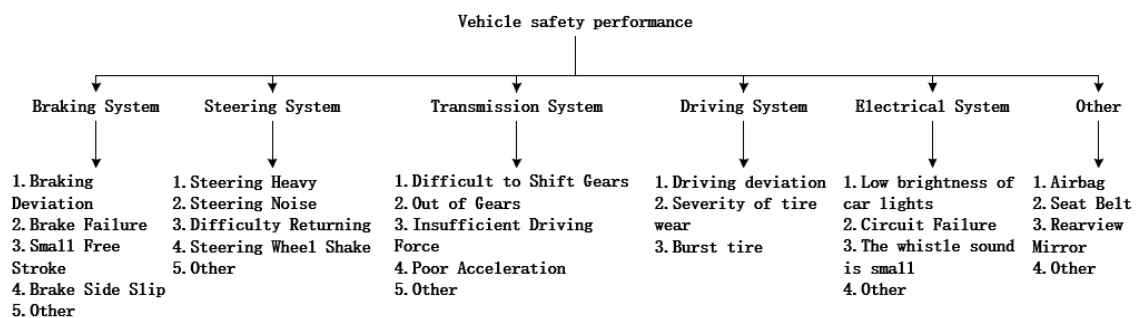


Figure 9 Common faults affecting vehicle safety performance

There is also a phenomenon of mutual influence between vehicles on the road. Common types of vehicle are cars, bus, heavy trucks, etc.

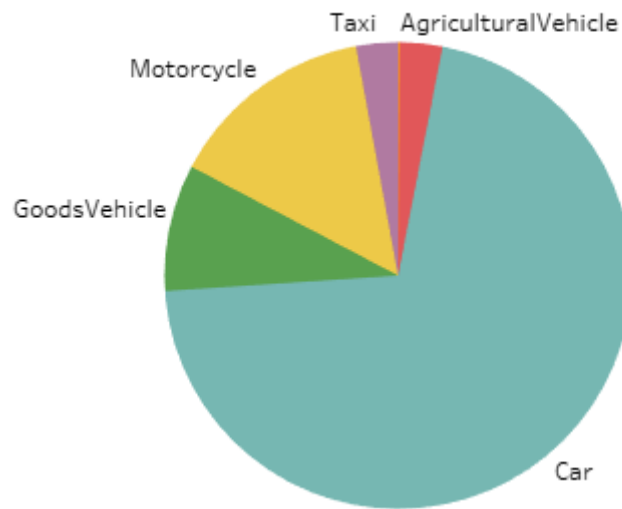


Figure 10 Proportion of various vehicles in traffic accidents

4.2.2.3 Road factors

Different roads have different physical characteristics, which will have different effects on vehicles. When a driver is driving a vehicle, he is affected by road conditions, road geometric characteristics, different road type combinations, and road physical facilities, and is prone to accidents such as slipping and rear-end collision and scratching against roadblocks. Among them, road conditions are mainly determined by road attachments. Drivers have different driving cognitions for different road conditions such as dry, wet, icy, etc.; road geometric characteristics and road combinations are mainly determined by local road conditions, especially overpass ramps and intersections. Other places are often high-frequency accident areas. In addition, the shoulders of road, obstacles and traffic signs on the road may also cause vehicle bumps and cause accidents. The composition of common road factors is shown in Figure 11.

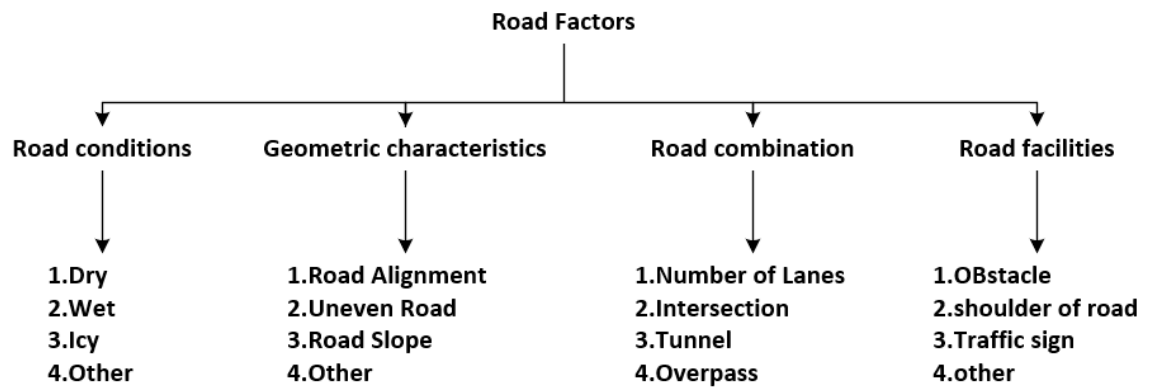


Figure 11 Road factors affecting traffic accidents

(1) Road Surface Condition

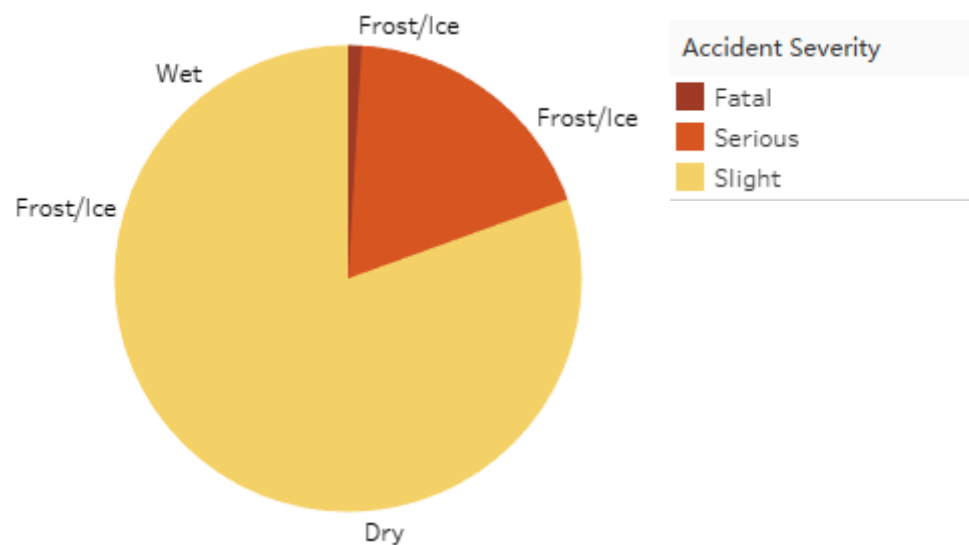


Figure 12 The impact of road surface conditions on the traffic accidents

(2) Junction control

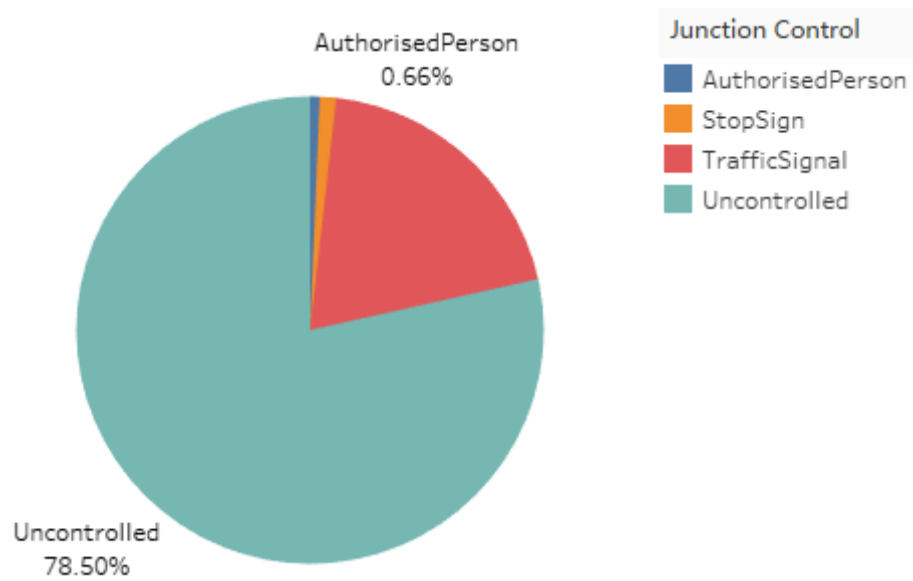


Figure 13 Pie chart of junction control

(3) Junction details

Table 2 Proportion of traffic accidents at various junction

Junction Detail	Accident Severity		
	Fatal	Serious	Slight
Crossroads	0.15%	2.96%	13.49%
MiniRoundabout	0.01%	0.36%	2.03%
MultipleJunction	0.01%	0.43%	1.87%
NotAtJunction	0.00%	0.00%	0.00%
Other	0.09%	1.59%	6.13%
PrivateDriveOrEntrance	0.04%	0.75%	3.07%
Roundabout	0.08%	2.05%	12.01%
SlipRoad	0.03%	0.37%	1.95%
Tjunction	0.52%	10.03%	39.98%

(4) Road type

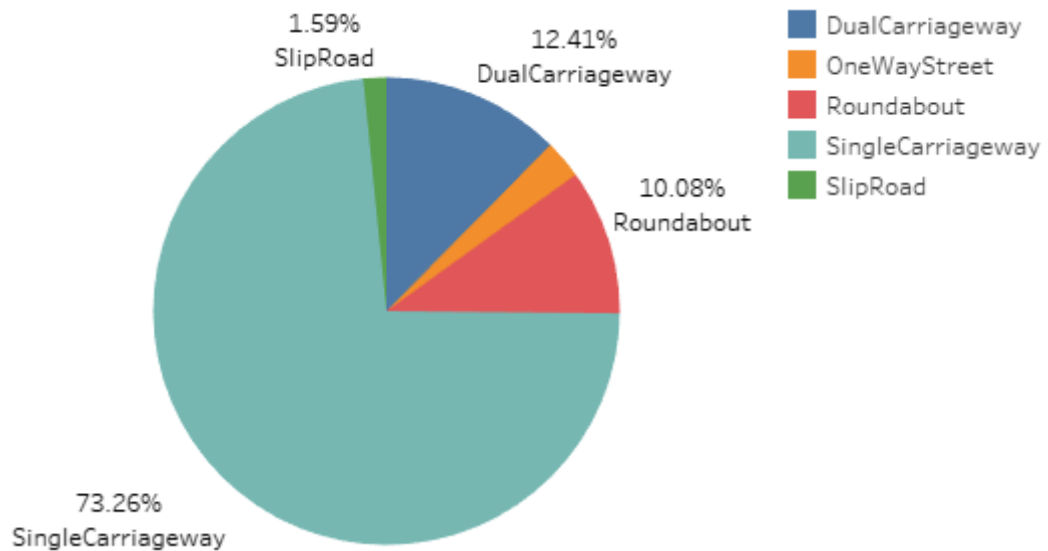


Figure 14 Proportion of the frequency of traffic accidents on different types of roads

4.2.2.4 Environmental factors

The impact of environmental factors on traffic accidents is mainly divided into natural environment and artificial environment, as shown in Figure 15. Among them, the most researches on weather conditions and light conditions in the natural environment, because rain, snow, and fog will cause the ground to become wet and icy, reduce visibility, and affect the driver's vision and vehicle driving stability. Studies have found that traffic accidents caused by weather such as rain and snow are 3 to 4 times higher than usual.

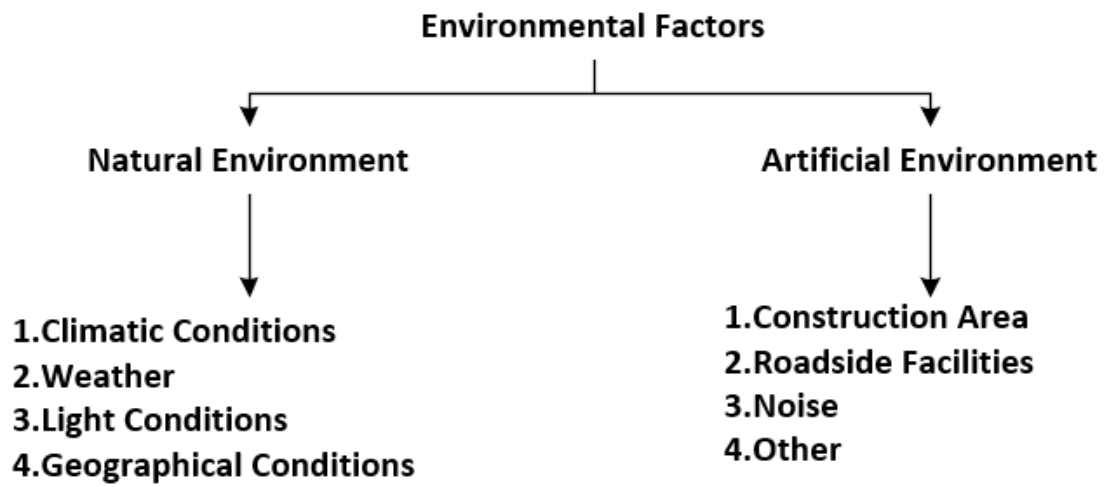


Figure 15 The impact of environmental conditions on traffic accidents

(1) Weather conditions

Table 3 Proportion of the severity of traffic accidents under various weather conditions

Accident Se..	Weather Conditions				
	Fine	Fog	Raining	Snowing	other
Fatal	0.80%	0.00%	0.11%	0.00%	0.01%
Serious	15.77%	0.06%	2.39%	0.06%	0.36%
Slight	66.65%	0.26%	11.02%	0.42%	2.08%

(2) Light conditions

Table 4 Proportion of the severity of traffic accidents under various light conditions

Light Conditions	conditions		
	Fatal	Accident Severity Serious	Slight
DarknessLight	0.26%	4.77%	18.98%
DarknessNoLighting	0.08%	0.69%	2.07%
daylight	0.59%	13.18%	59.38%

4.3 Analysis of accident cause based on Apriori algorithm

4.3.1 secondary pre-processing

4.3.1.1 Filter data

Since there are nearly 180,000 traffic data records in the United Kingdom in the past three years, it is considered to filter out the data of the most recent year (2020) as the data for correlation analysis. This is a preferable method, because at this stage, we only need to find the potential interrelationships between the various attributes of the accident. Too little data volume is conducive to simplifying the complexity of the model, preventing the occurrence of over-fitting, and improving the effect of the model. It can be seen from Figure 16 that after the data filtered by tableau prep builder, the total data records are only about 30,000, which is a good result.

24 fields 33,221 rows						
Filter Values... Rename Fields... Create Calculated Field... 1 Recommendation Search						
#	#	#	#	#	#	#
accident_index	accident_year	longitude	latitude	number_of_vehicles	number_of_casualti...	accident_month
2,020,010,287,195	2,020	0.11498	51.420052	2	1	12
2,020,070,649,306	2,020	-2.894066	53.195801	2	1	10
2,020,010,286,403	2,020	-0.104612	51.498764	2	1	12
2,020,040,933,014	2,020	-2.247207	53.791687	2	1	2
2,020,010,285,727	2,020	-0.150988	51.515898	2	1	12
2,020,010,286,834	2,020	-0.130367	51.507712	2	1	12
2,020,070,003,961	2,020	-2.514621	53.260303	1	1	1
2,020,070,069,416	2,020	-2.884892	53.191501	1	1	2
2,020,052,000,933	2,020	-3.075293	53.421623	2	1	5
2,020,052,000,933	2,020	-3.075293	53.421623	2	1	5
2,020,052,001,810	2,020	-3.033567	53.379455	2	2	9
2,020,052,001,114	2,020	-2.946424	53.500259	2	1	5
2,020,052,001,114	2,020	-2.946424	53.500259	2	1	5
2,020,052,001,137	2,020	-2.94106	53.43978	2	1	6
2,020,100,935,380	2,020	-1.461938	54.970579	2	2	3
2,020,010,286,624	2,020	-0.212966	51.358767	2	1	12

Figure 16 Filter out UK traffic data for 2020

4.3.1.2 Formatting data into acceptable form

Once the data has been cleaned, it can be prepared for loading into WEKA that will read data from Excel files but it gives the data columns meaningless names. It is much better if the data is preformatted into ARFF format before trying to load it into WEKA.

ARFF format file is split into 3 sections:

- A relation identifier
- A header showing the attributes of the data
- The data

ARFF header

@RELATION ACCIDENT2020

@ATTRIBUTE accident_month {1,2,3,4,5,6,7,8,9,10,11,12}

@ATTRIBUTE accident_time NUMERIC

@ATTRIBUTE accident_week {Monday,Tuesday,Wednesday,Thursday,Friday,Saturday,
Sunday}

@ATTRIBUTE road_type {DualCarriageway,OneWayStreet,Roundabout,SingleCarriage
way,SlipRoad}

@ATTRIBUTE junction_detail {Crossroads,MiniRoundabout,MultipleJunction,NotAtJunc
tion,Other,PrivateDriveOrEntrance,Roundabout,SlipRoad,Tjunction}

@ATTRIBUTE junction_control {AuthorisedPerson,StopSign,TrafficSignal,Uncontrolled}

@ATTRIBUTE light_conditions {DarknessLight,DarknessNoLighting,DarknessUnknown,
daylight}

@ATTRIBUTE weather_conditions {Fine,Fog,Raining,Snowing}

@ATTRIBUTE road_surface_conditions {Dry,Flood,Frost/Ice,Snow,Wet}

@ATTRIBUTE area {rural,urban}

@ATTRIBUTE vehicle_type {AgriculturalVehicle,Bus,Car,GoodsVehicle,Motorcycle,Taxi}

@ATTRIBUTE vehicle_manoeuvre {ChangingLane,GoingAhead,MovingOff,Overtaking,
Parked,Reversing,Slowing,Turning,Waiting}

@ATTRIBUTE sex_of_driver {Male,Female}

@ATTRIBUTE age_band_of_driver {11-15,16-20,21-25,26-35,36-45,46-55,56-65,66-75,
over75}

@ATTRIBUTE accident_severity {Slight,Serious,Fatal}

Put the header and data into notepad++, and save the file as accident2020.arff.

If there are problems, WEKA will tell you which line the problem is at. The line number is the line in the data file, not the line of the data section, which is a good reason to keep the data file open in Notepad++ while you load the data so you can find any errors easily.

After correcting the error, you will get a perfect ARFF file, which can be imported into WEKA, and then do the association rule analysis of traffic accidents.

4.3.2 Association rules

4.3.2.1 First experiment

In this part, the apriori algorithm needs to be used to obtain the association rules of the class, that is, which conditions may determine the severity of the accident. In practical applications, it is necessary to use the minimum support and the minimum confidence to find the associated results. When the minimum support (Minsupport) is set to 0.1 and the minimum confidence (Minconfidence) is set to 0.9, the best rule cannot be found. The result is shown in Figure 17. Therefore, you need to change the value of the parameter to get the best rule.

```
Associator output

accident_severity
=== Associator model (full training set) ===

Apriori
=====

Minimum support: 0.1 (5070 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 25

Size of set of large itemsets L(2): 84

Size of set of large itemsets L(3): 107

Size of set of large itemsets L(4): 81

Size of set of large itemsets L(5): 33

Size of set of large itemsets L(6): 3

Best rules found:
```

Figure 17 Association rule result 1

Set the minimum support (Minsupport) to 0.02; set the minimum confidence (Minconfidence) to 0.25, the result is shown in Figure 19. The association rule result of the experiment adopts the expression form of "item num.2=>conclusion num.2", where the number after the item is the record frequency under this condition, and the number after the conclusion is the record frequency that meets this conclusion.

```
Associator output

Apriori
=====

Minimum support: 0.5 (25350 instances)
Minimum metric <confidence>: 0.25
Number of cycles performed: 10

Generated sets of large itemsets:

Size of set of large itemsets L(1): 8

Size of set of large itemsets L(2): 4

Best rules found:

1. vehicle_type=Car 38890 ==> accident_severity=Slight 31808    conf:(0.82)
2. area=urban 37121 ==> accident_severity=Slight 30268    conf:(0.82)
3. weather_conditions=Fine vehicle_type=Car 32465 ==> accident_severity=Slight 26440    conf:(0.81)
4. light_conditions=daylight 36643 ==> accident_severity=Slight 29335    conf:(0.8)
5. light_conditions=daylight weather_conditions=Fine 32164 ==> accident_severity=Slight 25577    conf:(0.8)
6. weather_conditions=Fine 42495 ==> accident_severity=Slight 33725    conf:(0.79)
7. road_surface_conditions=Dry 35489 ==> accident_severity=Slight 28157    conf:(0.79)
8. weather_conditions=Fine road_surface_conditions=Dry 35280 ==> accident_severity=Slight 27986    conf:(0.79)
9. junction_control=Uncontrolled 40246 ==> accident_severity=Slight 31775    conf:(0.79)
10. road_type=SingleCarriageway 37480 ==> accident_severity=Slight 29514    conf:(0.79)
```

Figure 18 Association rule result 2

This result has an obvious flaw, that is, a certain type of association rule is obtained. The most likely reason is that this is an unbalanced data set, so the number of instances of slight accidents is far greater than the number of serious and fatal accidents. In the second experiment, we can separate the data of each type of accident, and then analyze the association rules of each type of traffic accident separately.

4.3.2.2 Second experiment

WEKA filters can help us separate three types of traffic accidents (minor, serious, and fatal). The specific operations are as follows:

Filter-> unsupervised-> instance-> RemoveWithValues-> set the values of
attributeindex and nominalindices

Here is to separate each type of traffic accident by deleting.

The results of each type of traffic accident are shown below:

(1) Slight accident

Selected attribute			
Name: accident_severity		Type: Nominal	
Missing: 0 (0%)		Unique: 0 (0%)	
		Distinct: 1	
No.	Label	Count	Weight
1	Slight	40458	40458.0
2	Serious	0	0.0
3	Fatal	0	0.0

Class: accident_severity (Nom) Visualize All



Figure 19 Data separation result 1

Associator output

```

Apriori
=====

Minimum support: 0.65 (26298 instances)
Minimum metric <confidence>: 0.25
Number of cycles performed: 7

Generated sets of large itemsets:

Size of set of large itemsets L(1): 8

Size of set of large itemsets L(2): 3

Best rules found:

1. weather_conditions=Fine 33725 ==> accident_severity=Slight 33725    conf:(1)
2. vehicle_type=Car 31808 ==> accident_severity=Slight 31808    conf:(1)
3. junction_control=Uncontrolled 31775 ==> accident_severity=Slight 31775    conf:(1)
4. area=urban 30268 ==> accident_severity=Slight 30268    conf:(1)
5. road_type=SingleCarriageway 29514 ==> accident_severity=Slight 29514    conf:(1)
6. light_conditions=daylight 29335 ==> accident_severity=Slight 29335    conf:(1)
7. road_surface_conditions=Dry 28157 ==> accident_severity=Slight 28157    conf:(1)
8. weather_conditions=Fine road_surface_conditions=Dry 27986 ==> accident_severity=Slight 27986    conf:(1)
9. sex_of_driver=Male 26994 ==> accident_severity=Slight 26994    conf:(1)
10. junction_control=Uncontrolled weather_conditions=Fine 26550 ==> accident_severity=Slight 26550    conf:(1)

```

Figure 20 Association rules for slight traffic accidents

(2) Serious accident

Selected attribute

Name: accident_severity	Distinct: 1	Type: Nominal
Missing: 0 (0%)		Unique: 0 (0%)

No.	Label	Count	Weight
1	Slight	0	0.0
2	Serious	9725	9725.0
3	Fatal	0	0.0

Class: accident_severity (Nom) Visualize All

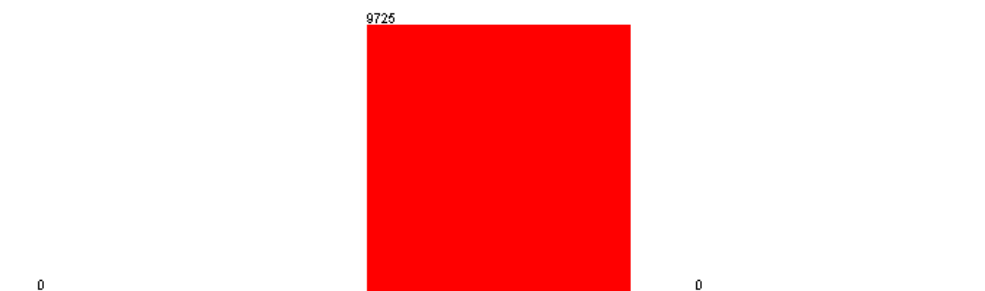


Figure 21 Data separation result 2

```

Associator output

Apriori
=====

Minimum support: 0.65 (6321 instances)
Minimum metric <confidence>: 0.25
Number of cycles performed: 7

Generated sets of large itemsets:

Size of set of large itemsets L(1): 8

Size of set of large itemsets L(2): 4

Best rules found:

1. weather_conditions=Fine 8325 ==> accident_severity=Serious 8325    conf:(1)
2. junction_control=Uncontrolled 8022 ==> accident_severity=Serious 8022    conf:(1)
3. road_type=SingleCarriageway 7551 ==> accident_severity=Serious 7551    conf:(1)
4. sex_of_driver=Male 7101 ==> accident_severity=Serious 7101    conf:(1)
5. road_surface_conditions=Dry 6963 ==> accident_severity=Serious 6963    conf:(1)
6. light_conditions=daylight 6961 ==> accident_severity=Serious 6961    conf:(1)
7. weather_conditions=Fine road_surface_conditions=Dry 6930 ==> accident_severity=Serious 6930    conf:(1)
8. junction_control=Uncontrolled weather_conditions=Fine 6895 ==> accident_severity=Serious 6895    conf:(1)
9. vehicle_type=Car 6760 ==> accident_severity=Serious 6760    conf:(1)
10. area=urban 6581 ==> accident_severity=Serious 6581    conf:(1)

```

Figure 22 Association rules for serious traffic accidents

(3) Fatal accident

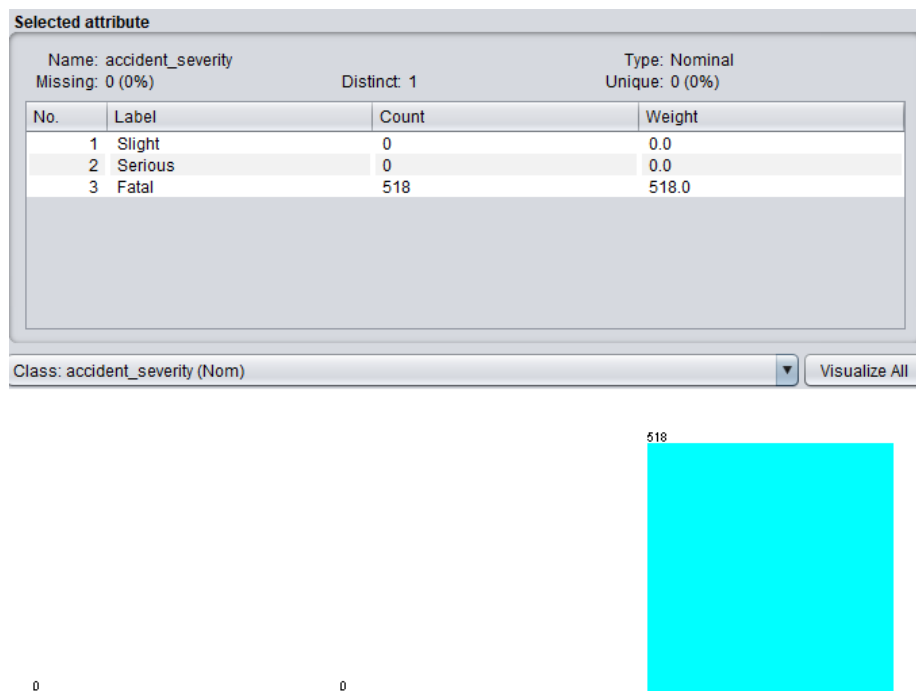


Figure 23 Data separation result 3

```
Associator output

Apriori
=====

Minimum support: 0.65 (337 instances)
Minimum metric <confidence>: 0.25
Number of cycles performed: 7

Generated sets of large itemsets:

Size of set of large itemsets L(1): 7

Size of set of large itemsets L(2): 6

Best rules found:

1. junction_control=Uncontrolled 449 ==> accident_severity=Fatal 449    conf:(1)
2. weather_conditions=Fine 445 ==> accident_severity=Fatal 445    conf:(1)
3. sex_of_driver=Male 418 ==> accident_severity=Fatal 418    conf:(1)
4. road_type=SingleCarriageway 415 ==> accident_severity=Fatal 415    conf:(1)
5. junction_control=Uncontrolled weather_conditions=Fine 387 ==> accident_severity=Fatal 387    conf:(1)
6. road_type=SingleCarriageway junction_control=Uncontrolled 372 ==> accident_severity=Fatal 372    conf:(1)
7. road_surface_conditions=Dry 369 ==> accident_severity=Fatal 369    conf:(1)
8. weather_conditions=Fine road_surface_conditions=Dry 364 ==> accident_severity=Fatal 364    conf:(1)
9. road_type=SingleCarriageway weather_conditions=Fine 363 ==> accident_severity=Fatal 363    conf:(1)
10. junction_control=Uncontrolled sex_of_driver=Male 358 ==> accident_severity=Fatal 358    conf:(1)
```

Figure 24 Association rules for fatal traffic accidents

4.3.3 Classification models

In order to obtain the severity of traffic accidents under the combination of various attribute values, we use a classification model to achieve this task. In addition, we will analyze and compare the performance of several classification models.

The test option chooses cross-validation or percentage split to create several classification models.

10-fold cross-validation:

- 1) Divide dataset into 10 parts(folds)
- 2) Hold out each part in turn

3) Average the results

4) Each data point used once for testing, 9 time for training

Stratified corss-validation: ensure that each fold has the right proportion of each class value. Weka's default is stratified cross-validation

Percentage split: if set 90%, 90% of dataset for training, 10% of dataset for testing.

How to choose test option?

when the amount of data is large, you can choose percentage split. Since the total amount of data in the data set exceeds 30,000 and there are only three classes, it is better to use percentage split.

Generally, when the amount of data is relatively small, in order to allow the classifier to be fully trained, we will use cross-validation to train and test the model. But in the case of a large amount of data, training the model in this way will increase the training time of the model. Therefore, in this part, the percentage split method is used to train the model.

It can be seen from Figure25 and Figure 26 that when percentage split is selected, the prediction effect of the model is better. Therefore, Percentage split will be used in the process of building models.

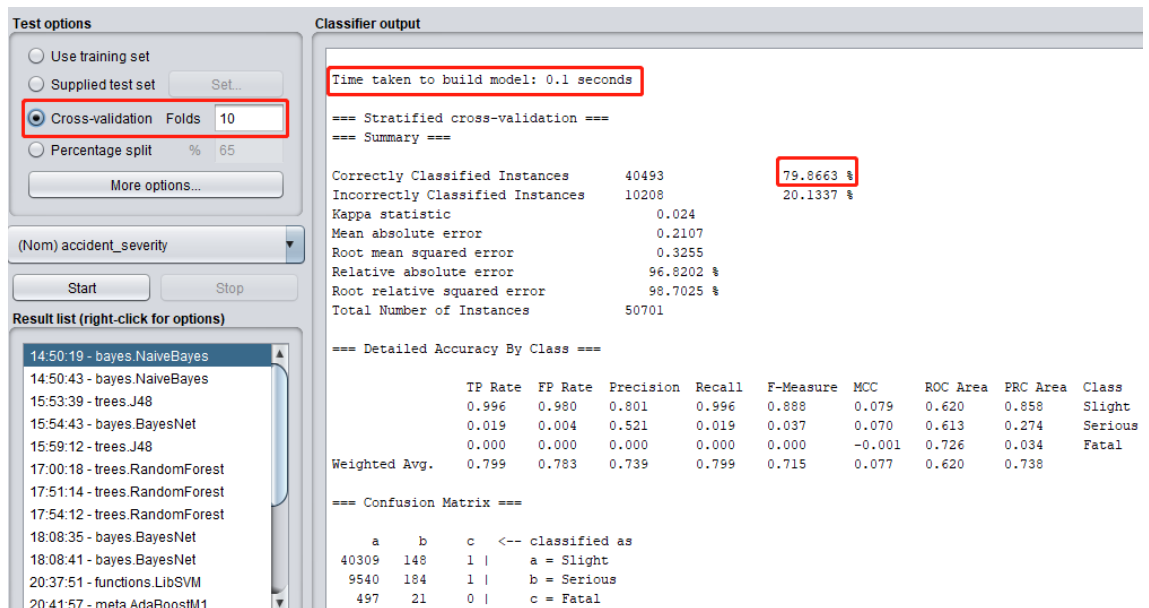


Figure 25 Results of NaiveBayes using cross-validation

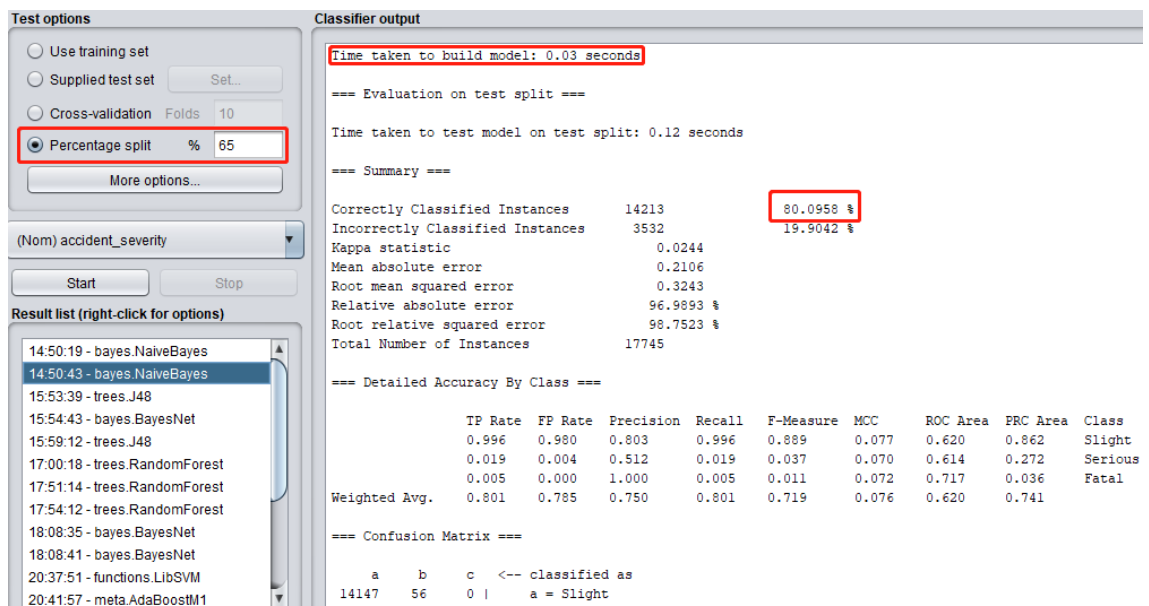


Figure 26 Results of NaiveBayes using percentage split (65%)

4.3.3.1 First experiment

In this part, we initially use J48, Bayes, and Random Forest algorithms to classify the traffic data set, and analyze the results.

The test option chooses percentage split.

(1) J48 algorithm

Classifier output

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances

14203

80.0394 %

Incorrectly Classified Instances

3542

19.9606 %

Kappa statistic

0

Mean absolute error

0.2171

Root mean squared error

0.3284

Relative absolute error

99.9905 %

Root relative squared error

99.9999 %

Total Number of Instances

17745

=== Detailed Accuracy By Class ===

TP Rate

FP Rate

Precision

Recall

F-Measure

MCC

ROC Area

PRC Area

Class

1.000

1.000

0.800

1.000

0.889

?

0.500

0.800

Slight

0.000

0.000

?

0.000

?

?

0.500

0.189

Serious

0.000

0.000

?

0.000

?

?

0.500

0.011

Fatal

Weighted Avg.

0.800

0.800

?

0.800

?

?

0.500

0.676

=== Confusion Matrix ===

a

b

c

<-- classified as

14203

0

0

a = Slight

3353

0

0

b = Serious

189

0

0

c = Fatal

Figure 27 Results of JRIP using percentage split (65%)

(2) Random forest algorithm

Classifier output

Time taken to test model on test split: 1.49 seconds

=== Summary ===

Correctly Classified Instances	13935	78.5292 %
Incorrectly Classified Instances	3810	21.4708 %
Kappa statistic	0.0501	
Mean absolute error	0.208	
Root mean squared error	0.3325	
Relative absolute error	95.762 %	
Root relative squared error	101.2407 %	
Total Number of Instances	17745	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.965	0.926	0.807	0.965	0.879	0.076	0.604	0.849	Slight
	0.067	0.035	0.307	0.067	0.111	0.063	0.596	0.248	Serious
	0.026	0.001	0.200	0.026	0.047	0.069	0.694	0.054	Fatal
Weighted Avg.	0.785	0.748	0.706	0.785	0.725	0.074	0.604	0.727	

=== Confusion Matrix ===

a	b	c	<-- classified as
13704	490	9	a = Slight
3116	226	11	b = Serious
165	19	5	c = Fatal

Figure 28 Results of random forest using percentage split (65%)

(3) Bayes algorithm

```

Classifier output

=== Summary ===

Correctly Classified Instances      14212      80.0902 %
Incorrectly Classified Instances    3533      19.9098 %
Kappa statistic                    0.0249
Mean absolute error                 0.2105
Root mean squared error             0.3243
Relative absolute error             96.9432 %
Root relative squared error         98.7322 %
Total Number of Instances          17745

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0.996    0.979    0.803     0.996    0.889     0.080    0.621    0.862    Slight
      0.019    0.004    0.504     0.019    0.037     0.069    0.615    0.272    Serious
      0.005    0.000    0.333     0.005    0.010     0.041    0.717    0.037    Fatal
Weighted Avg.   0.801    0.784    0.742     0.801    0.719     0.077    0.621    0.742

=== Confusion Matrix ===

      a      b      c  <-- classified as
14146    57      0 |  a = Slight
 3286   65      2 |  b = Serious
  181     7      1 |  c = Fatal
  
```

Figure 29 Results of random forest using percentage split (65%)

As can be seen from Figures 27, 28, and 29, the number of examples correctly classified by these three algorithms is quite substantial (about 80%). However, after confidently observing the results of the three algorithms, we can find that the three algorithms are only good at classifying slight traffic accidents, but are at a loss for serious and fatal traffic accidents. The reason for this is probably because the traffic data set used in the experiment is one unbalanced data set.

Data imbalance is also called data skew, which refers to the large difference in the number of samples in each category of the data set. Taking traffic accident data as an example, as shown in Figure 30, fatal accidents account for only 1.02% of all traffic accidents. The imbalance of data distribution may cause the majority class to dominate, thereby ignoring the impact of the minority class. Therefore, it is necessary to balance the data first.

In the second test, the data will be preprocessed for the third time, and the above three classification models will be established through WEKA again.

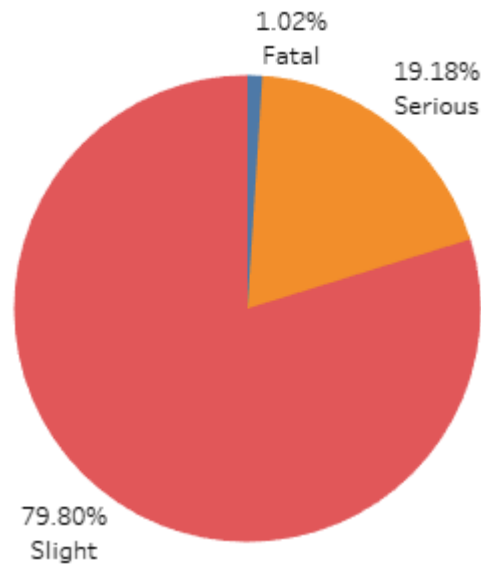


Figure 30 Proportion of traffic accidents of different severity

4.3.3.2 Second experiment

This part will preprocess the data for the third time, and use the processed data to build a classifier.

4.3.3.2.1 Third pre-processing

In terms of data sampling, the main sampling methods are divided into two types: under-sampling (down-sampling) and over-sampling (up-sampling). As shown in Figure 31, under-sampling refers to randomly selecting a small number of samples from most types of data, and ultimately ensuring the same data for each type of sample. Over-sampling, which is the opposite of under-sampling, re-sampling from a minority of data to add new samples, also achieves the goal of equal amount of data for each type of sample. In addition, under the condition that the total amount of data remains

unchanged, under-sampling and over-sampling can be combined to achieve the effect of data balance by mixing sampling.

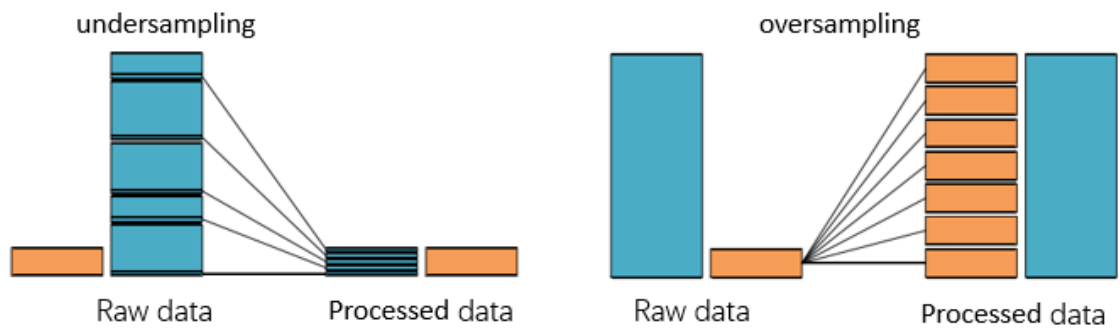


Figure 31 Undersampling and oversampling

we use the filter (resample) in WEKA to address the imbalanced dataset which can produce a random subsample of a dataset using either sampling with replacement or without replacement. The specific steps are as follows:

filter->supervised->instance->Resample-> biasToUniformClass = 1 -> apply

biasToUniformClass -- Whether to use bias towards a uniform class. A value of 0 leaves the class distribution as-is, a value of 1 ensures the class distribution is uniform in the output data.

The result is shown in the figure 24 below:

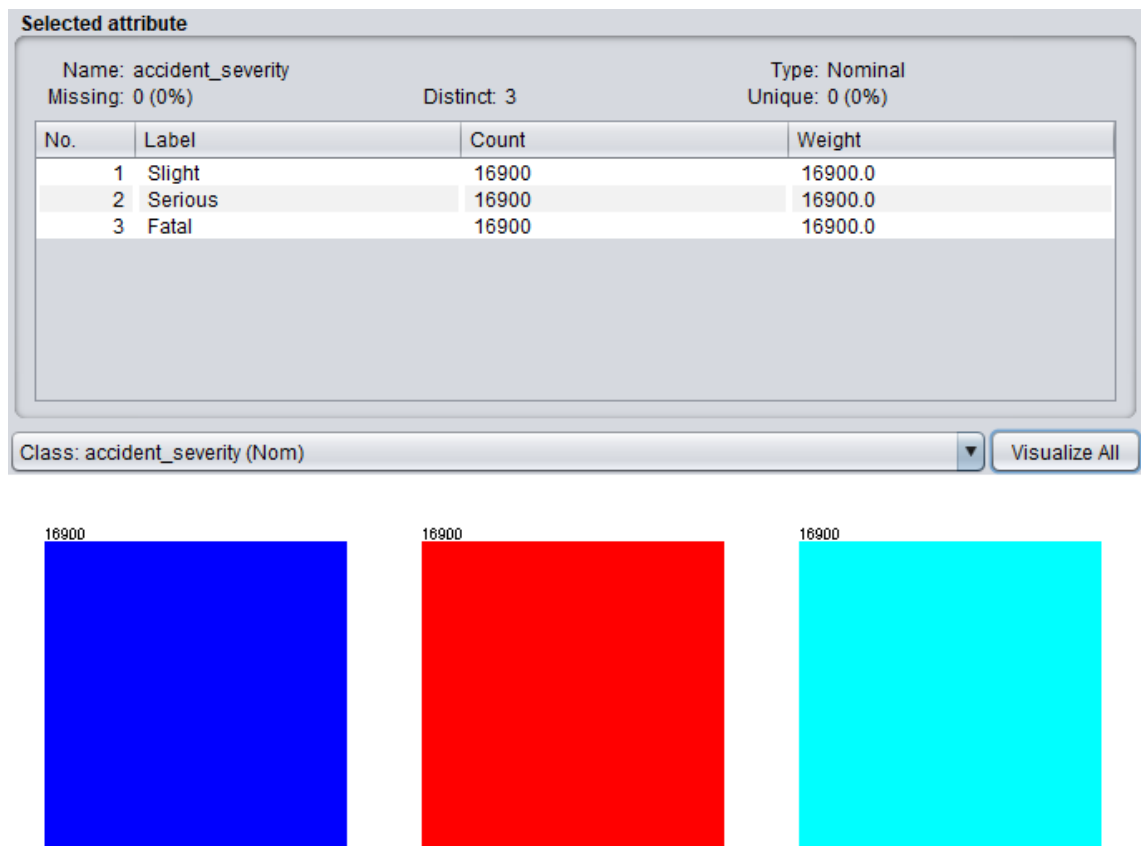


Figure 32 The result of data processing

4.3.3.2.2 Building the classifiers

After many experiments, it is found that when the percentage split is set to 65%, the constructed classifier works best. Therefore, only the result graph when percentage split is equal to 65 is pasted here.

(1) J48 algorithm

```

Classifier output

Time taken to test model on test split: 0.02 seconds

=== Summary ===

Correctly Classified Instances      13610      76.6977 %
Incorrectly Classified Instances    4135      23.3023 %
Kappa statistic                    0.6505
Mean absolute error                0.1729
Root mean squared error            0.3388
Relative absolute error             38.9113 %
Root relative squared error         71.8671 %
Total Number of Instances          17745

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                0.627    0.147    0.681     0.627    0.653      0.490    0.826     0.692    Slight
                0.675    0.160    0.678     0.675    0.677      0.516    0.845     0.685    Serious
                1.000    0.042    0.921     1.000    0.959      0.939    0.990     0.963    Fatal
Weighted Avg.   0.767    0.117    0.760     0.767    0.763      0.648    0.887     0.780

=== Confusion Matrix ===

  a    b    c  <-- classified as
3721 1892  321 |  a = Slight
1740 3991  182 |  b = Serious
  0    0 5898 |  c = Fatal

```

Figure 33 The second result of J48 classifier

(2) Random forest algorithm

```

Classifier output

Time taken to test model on test split: 1.36 seconds

=== Summary ===

Correctly Classified Instances      15245      85.9115 %
Incorrectly Classified Instances    2500      14.0885 %
Kappa statistic                    0.7887
Mean absolute error                0.1449
Root mean squared error            0.2514
Relative absolute error             32.6015 %
Root relative squared error         53.3293 %
Total Number of Instances          17745

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                0.759    0.088    0.813     0.759    0.785      0.684    0.945     0.871    Slight
                0.819    0.116    0.778     0.819    0.798      0.694    0.947     0.913    Serious
                1.000    0.007    0.985     1.000    0.993      0.989    1.000     1.000    Fatal
Weighted Avg.   0.859    0.071    0.859     0.859    0.858      0.789    0.964     0.928

=== Confusion Matrix ===

  a    b    c  <-- Classified as
4504 1378  52 |  a = Slight
1035 4843  35 |  b = Serious
  0    0 5898 |  c = Fatal

```

Figure 34 The second result of random forest classifier

(3) Bayes algorithm

Classifier output

=== Summary ===

Correctly Classified Instances	8540	48.1262 %
Incorrectly Classified Instances	9205	51.8738 %
Kappa statistic	0.2219	
Mean absolute error	0.3925	
Root mean squared error	0.4491	
Relative absolute error	88.3207 %	
Root relative squared error	95.2638 %	
Total Number of Instances	17745	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.535	0.307	0.466	0.535	0.498	0.221	0.683	0.506	Slight
	0.295	0.198	0.427	0.295	0.349	0.109	0.613	0.428	Serious
	0.614	0.273	0.528	0.614	0.568	0.330	0.744	0.585	Fatal
Weighted Avg.	0.481	0.259	0.474	0.481	0.472	0.220	0.680	0.506	

=== Confusion Matrix ===

a	b	c	<-- classified as
3173	1381	1380	a = Slight
2312	1747	1854	b = Serious
1318	960	3620	c = Fatal

Figure 35 The second result of bayes classifier

5 Results and discussion

This chapter includes a summary of the results of the experiments done above and some valuable discussions.

5.1 Data visualization

5.1.1 Accident temporal and spatial distribution characteristics

5.1.1.1 Analysis of accident time characteristics

(1) Year

From the above figure1, it can be seen that the number of traffic accidents in the UK in the past three years has been declining year by year. This may be because the relevant departments in the UK have formulated better traffic policies and people's awareness of traffic accident prevention has increased.

(2) Month

As can be seen from the above figure2, April is the month with the least accidents. From a seasonal perspective, autumn and winter are the seasons with the highest frequency of accidents. The reason may be that the season will affect the road conditions (ice, rain), so pay attention to the prevention of traffic accidents in autumn and winter.

(3) Week

From the above figure3, we can see that the number of accidents during working days is significantly higher than that on non-working days. The number of accidents steadily rises from Monday to Saturday, and reaches a peak on Saturday. The arrival of the weekend caused a rapid decline in the number of accidents. Since Friday is the last working day of the week, the working people feel relaxed after working for the week, reducing their driving vigilance, and the increase in traffic flow has caused more accidents. Therefore, especially for working days on Fridays Pay special attention to accident prevention.

(4) Time

The result of the analysis is that the number of accidents in different hours is quite different. Among them, 8 o'clock and 17:00 are the high-frequency time of accidents every day. The peak period of commuting brings traffic problems such as traffic jams, and the driver's mental state of eagerness to work, Leading to an increase in the number of accidents.

5.1.1.2 Accident spatial distribution analysis

Accidents are mainly concentrated in the central and southern parts of the United Kingdom. The characteristics of the accident cluster are that the regional economy is developed, the permanent population is large, and the road environment is complex. Therefore, the number of accidents is positively correlated with these factors.

5.1.2 Analysis of accident factors

5.1.2.1 Human factors

(1) Gender

It can be seen from the figure7 that the gap in the proportion of traffic accidents caused by male and female drivers is relatively small. The ratio of the number of men and women holding a driver's license is the main reason for this phenomenon.

(2) Age

It can be seen from the figure8 above that accidents occur at most in people aged 26-35, which may be due to lack of driving experience. With age, the number of accidents gradually decreases. This is because older people may have more driving experience and do not like to drive fast.

The accidents are mainly slight accidents. Serious accidents mainly happen to people aged 26-35. It may be because people in this age group prefer to drive fast, so the severity of the accident has been increased.

5.1.2.2 Vehicle factors

Among all types of vehicles, cars accounted for 85.08%, which is related to the fact that the car base is large; as the main vehicles for transporting goods in cities, there are also many trucks, accounting for 10.54%. Due to the large size of the truck, the driver is prone to blind spots in the field of vision, which can lead to vehicle accidents. Especially the drivers of heavy trucks often experience fatigue driving due to long-term transportation of materials across regions, which will affect the severity of the accident. There are further upgrades. From the books, motorcycles are also the most commonly used means of transportation, accounting for the second largest proportion. Therefore, we must pay special attention to safety when riding motorcycles, because the probability of traffic accidents is also very high.

5.1.2.3 Road factors

(1) Road surface conditions

As can be seen from the above figure 12, serious and fatal traffic accidents usually occur on icy or frost roads. This tells us that under such conditions, we must drive at a speed limit, so as to ensure safer driving.

(2) Junction control

Obviously, when the junction is uncontrolled, the more frequent traffic accidents occur, accounting for 78.5% of the total traffic accident frequency. In the case of human supervision, the frequency of accidents is the least, only 0.66%, which shows that human control may be the best way to prevent traffic accidents. Therefore, when human resources are sufficient, as many people as possible should be arranged to control traffic at intersections. If human resources are insufficient, human supervision should also be set up on high-frequency road sections where accidents occur.

(3) Junction details

The visualization results show that when the vehicle is not at the junction, the frequency of traffic accidents is the least, almost 0%. When a vehicle is at a T-shaped junction, the frequency of accidents is the highest, and serious and fatal traffic accidents are more likely to occur. Therefore, it is necessary to increase traffic control at T-junctions and increase people's safety awareness at T-junctions.

(4) Road type

From the figure14, it can be concluded that the accident frequency of single carriageway and dual carriageway is 73% and 10%, respectively. The possible reason is that the driver is driving too fast on this road section. It is a good idea to add traffic speed limit signs on these road sections.

5.1.2.4 Environmental factors

(1) Weather conditions

The weather statistics at the time of traffic accidents are shown in table 3. It can be found that traffic accidents mostly occur on sunny and rainy days, accounting for 83.22% and 13.52% respectively. Only 0.48% and 0.33% of accidents occur on snowy and foggy days. This phenomenon is closely related to the number of different weather occurrences. During the three years, sunny and rainy days accounted for the majority. In addition, the small number of traffic accidents in snowy and foggy days may also be due to people's reduction in travel under these conditions and more caution when driving.

(2) Light conditions

It can be seen from the table 4 that accidents occur more frequently during the day. This is a normal phenomenon, because people's daily activities are during the day and people prefer to be at home at night. However, the data shows that the number of accidents at night when there are lights is greater than when there are no lights. This is

not in line with the normal logic. The reason may be that most of the roadsides in the UK have built street lights, which leads to unreasonable statistics.

5.1.3 Recommendations for preventing traffic accidents

Through the analysis of this part, suggestions for preventing traffic accidents can be put forward from the following aspects:

1. Time

From the results of data visualization, in winter, we should pay extra attention to traffic safety, because snowy weather may cause traffic accidents. The solution is to drive at a low speed and install anti-skid devices on the wheels.

On working days, we must pay special attention to traffic safety during commuting time, because during this time period, the traffic flow is too large and it is easy to cause traffic accidents. The solution is to let the relevant departments restrict the traffic flow during this time period, and arrange traffic police at various traffic arteries to manage the traffic order on the spot.

2. Human

Relevant departments and the media need to strengthen traffic safety education for citizens, especially young people. In the final analysis, people are the main body of traffic accidents. Only when people's safety awareness increases, there will naturally be a safe driving environment.

3. Vehicle

From the perspective of the vehicles involved in the accident, cars accounted for the majority, and cars are generally private cars. Therefore, the regular maintenance of vehicles is very important, and a car with good performance can help reduce the incidence of traffic accidents.

4. Road

The results of data visualization show that serious and even fatal traffic accidents are prone to occur on icy/frost roads. Therefore, in extreme weather, relevant departments should organize relevant personnel to clean up the snow and ice on the road in time. In addition, we must also call on people to slow down on this kind of road.

On road sections where accidents often occur, such as T-junction, single carriageway, set up speed signs and other reminders.

5. Infrastructure

Some road infrastructure, such as street lights, should be added to reduce the accident rate of driving at night.

5.2 Association rules

Since there are many rules obtained through the apriori algorithm, here are some valuable rules for discussion.

5.2.1 Slight traffic accident

(1) Experimental results:

vehicle_type=Car ==> accident_severity=Slight

junction_control=Uncontrolled ==> accident_severity=Slight

area=urban==> accident_severity=Slight

road_type=SingleCarriageway==> accident_severity=Slight

light_conditions=daylight==> accident_severity=Slight

sex_of_driver=Male==> accident_severity=Slight

(2) Conclusion:

1. More slight traffic accidents in cars
2. When junction are not uncontrolled, slight traffic accidents are more likely to occur.
3. Most slight traffic accidents happen during the day.
4. More slight traffic accidents in the urban
5. Man are more likely to cause slight traffic accidents
6. Slight traffic accidents are more likely to happen on single carriageway.
7. Frequent slight traffic accidents during the day.

It can be seen from the conclusion that the association rules obtained by the data mining algorithm are roughly similar to the results of data visualization, which shows the feasibility of the algorithm.

5.2.2 *Serious traffic accident*

(1) Experimental results:

vehicle_type=Car ==> accident_severity=Serious

junction_control=Uncontrolled ==> accident_severity= Serious

area=urban==> accident_severity= Serious

road_type=SingleCarriageway==> accident_severity= Serious

light_conditions=daylight==> accident_severity= Serious

sex_of_driver=Male==> accident_severity= Serious

(2) Conclusion:

1. More serious traffic accidents in cars
2. When junction are not uncontrolled, serious traffic accidents are more likely to occur.
3. Most serious traffic accidents happen during the day.
4. More serious traffic accidents in the urban
5. Man are more likely to cause serious traffic accidents
6. Serious traffic accidents are more likely to happen on single carriageway.
7. Frequent serious traffic accidents during the day.

The conclusion here is still similar to the result of data visualization.

5.2.3 *Fatal traffic accident*

(1) Experimental results:

road_type=SingleCarriageway,junction_control=Uncontrolled==>accident_severity
=Fatal

junction_control=Uncontrolled,sex_of_driver=Male==> accident_severity=Fatal

(2) Concussion:

1. Uncontrolled single carriageways are more prone to serious traffic accidents
2. Men driving on uncontrolled roads are more likely to cause serious traffic accidents.

The above two conclusions are not obtained from the data visualization results, which shows that the association rule algorithm can indeed dig out the potential information of the data.

The classic association rule algorithm performs well on this data set, and this conclusion is also different from the conclusion of the literature review, which shows that for different data sets, using the same algorithm, the results may also be different.

5.2.4 Advice on preventing traffic accidents

From the results of the above analysis, we can make some suggestions on reducing the number of traffic accidents. For example, most traffic accidents happen to men. Therefore, we should strengthen the supervision of male drivers. Measures include improving safety awareness and preventing drunk driving. As much as possible, arrange personnel to monitor the traffic on site.

5.3 Comparison of classifiers

Two evaluation measures are mainly analyzed here: recall and accuracy.

5.3.1 J48 algorithm

5.3.1.1 Experimental results:

Table 5 J48 experimental results

Measures	Imbalanced dataset	Balanced dataset
Accuracy	80.0394 %	76.6977 %
Recall	0.8	0.767
Time1	1.94s	1.43s
Time2	0.1s	0.02s

Time1: Time taken to build model

Time2: Time taken to test model on test split

=== Confusion Matrix ===

Imbalanced dataset:

```
a  b  c <-- classified as
14203  0  0 |  a = Slight
3353  0  0 |  b = Serious
189  0  0 |  c = Fatal
```

Balanced dataset:

a b c <-- classified as

3721 1892 321 | a = Slight

1740 3991 182 | b = Serious

0 0 5898 | c = Fatal

5.3.1.2 Conclusion:

It can be seen from the experimental results that the imbalance of the data has a very bad effect on the performance of the model. E.g. increase in time cost.

For the J48 classifier, the accuracy and recall of the classifier in the balanced data set seem to decrease compared to the unbalanced data set. But in fact, it can be seen from the confusion matrix that the former can perfectly classify serious traffic accidents and fatal traffic accidents, while the latter is difficult to do. This means that the classifier performance obtained by balancing the data set is better.

5.3.2 Random forest algorithm

5.3.2.1 Experimental results:

Table 6 the experimental results of random forest

Measures	Imbalanced dataset	Balanced dataset
Accuracy	78.5292 %	85.9115 %
Recall	0.785	0.859
Time1	11.27s	6.91s

Time2	1.14	1.36s
-------	------	-------

Time1: Time taken to build model

Time2: Time taken to test model on test split

=== Confusion Matrix ===

Imbalanced dataset:

a b c <-- classified as

13704 490 9 | a = Slight

3116 226 11 | b = Serious

165 19 5 | c = Fatal

Balanced dataset:

a b c <-- classified as

4504 1378 52 | a = Slight

1035 4843 35 | b = Serious

0 0 5898 | c = Fatal

5.3.2.2 Conclusion:

The classifier trained on the balanced data set is almost better in performance than the classifier trained on the unbalanced data set, and the accuracy of the random forest classifier exceeds 85%. It can also show good performance when classifying serious traffic accident data and fatal traffic accident data, which can be observed from the confusion matrix. The only drawback is that it has a higher time cost compared to the J48 classifier.

Therefore, the random forest classifier performs well on this data set, and can be used as a traffic accident prediction model for use in reality to predict the severity of traffic accidents.

5.3.3 Bayes algorithm

5.3.3.1 Experimental results:

Table 7 Bayes experimental results

Measures	Imbalanced dataset	Balanced dataset
Accuracy	80.0902 %	48.1262 %
Recall	0.801	0.481
Time1	0.07s	0.07s
Time2	0.03s	0.04s

Time1: Time taken to build model

Time2: Time taken to test model on test split

=== Confusion Matrix ===

Imbalanced dataset:

```

a   b   c  <-- classified as
14146  57   0 |   a = Slight
3286   65   2 |   b = Serious
181    7   1 |   c = Fatal

```

Balanced dataset:

a b c <-- classified as

3173 1381 1380 | a = Slight

2312 1747 1854 | b = Serious

1318 960 3620 | c = Fatal

5.3.3.2 Conclusion:

The Bayesian classifier is completely different from the first two classifiers. The Bayesian classifier obtained by balancing the data set is far inferior in performance to the case of training on the unbalanced data set. Surprisingly, on the balanced data set, the accuracy of the Bayesian classifier is only about 48%. The possible reason is that the model is overfitting. If we want to improve the performance of the model, we may need a data set that is more suitable for the Bayesian algorithm.

Therefore, the Bayes classifier is not suitable for this data set, that is, the experimental results are inconsistent with the literature review.

6 Conclusions

7.1 Research summary

Vehicle accident research is the hotspot and focus of road traffic safety research. This article is based on traffic accident data on British roads, starting from three aspects: accident data visualization, vehicle accident cause rule mining, and vehicle accident

severity prediction. Based on the results, the prevention of traffic is proposed. Accident recommendations and classification models of the severity of the accident. The report mainly carried out the following research work:

- (1) Using the traffic accident data of the UK from 2018 to 2020, compare and analyze the differences in the temporal and spatial characteristics of the two, and discuss the relevant factors of traffic accidents from the four levels of personnel, vehicles, roads and the environment, and analyze from a statistical perspective. The influences of various factors on traffic accidents are discussed, which lays the foundation for the selection of influencing factors in the formulation of accident early warning strategies.
- (2) Based on UK traffic accident data, the apriori association rule algorithm is used to obtain the attributes that are most closely related to the severity of traffic accidents by adjusting parameters and processing data sets. Based on this, an early warning strategy of accident rules was formulated to provide drivers with early warning information about the rules of the cause of accidents.
- (3) Based on the UK accident data, the original data is balanced for the large gap in the number of accidents of different severity, and then three classifiers are constructed and their performance is compared to obtain a traffic accident severity prediction model. The prediction from the model is accurate. Rate and recall rate verify the effectiveness of the model. The balanced data also improves

the accuracy of the prediction of the severity of the accident and provides the driver with more accurate warning information.

7.2 Suggestions for further work

At present, the research on traffic accidents is still in the development stage.

While obtaining the above research results, due to the limitation of research time and own ability, the research content still has limitations, including the following two aspects:

- (1) The British traffic accident data used in this article has fewer attributes, especially the information about vehicle factors only involves the types of vehicles. If more comprehensive accident information can be obtained, more detailed warning content can be provided.
- (2) Under the balanced data set, the reason for the performance drop of the Bayesian model is temporarily unknown. It may be due to the low quality of the data set. Subsequent work can verify this conjecture by obtaining a high-quality data set.

7 Reflection

7.1 Difficulty

- (1) Data collection

There are many database sets on the Internet, with a wide variety of data sets, and the quality of the data sets is uneven, and it is necessary to understand whether the selected data set is meaningful to address practical problems.

(2) Data preprocessing

After downloading the data set from the Internet, it cannot be used for data mining tasks immediately, because the original data set is dirty and needs to be preprocessed before it can be used. The preprocessing technology is also a bit complicated, such as data type conversion, grouping, and aggregation, etc. Finding the right data preprocessing tool is a good start. I mainly use tableau, weka and MS excel to complete data preprocessing.

(3) Model selection

There are many classification models, so you have to compare and analyze multiple classification models before you can make a choice. This process is a long one.

7.2 Surprise

With the development of big data analysis technology, the birth of new technology has brought convenience to our daily lives. From this project, I learned that we can reduce the incidence of traffic accidents through in-depth data mining. This is a very meaningful project, because the number of people who die in traffic accidents across

the country each year far exceeds your imagination. This project is a guarantee for People's travel safety plays an important role.

7.3 Rewards

After completing this data mining task, I learned a lot of knowledge, such as data preprocessing techniques, commonly used statistical techniques, commonly used data mining algorithms, and so on. I believe this knowledge is very helpful to my personal growth and employment in the future.

8 References

- [1] TRANSPORT I. Safety is integral to development of smart cities according to new research[EB/OL].[1stOctober]. <https://www.intelligenttransport.com/transport-news/81699/safety-integral-smartcities/>
- [2] ORGANIZATION W H. Road traffic injuries[EB/OL]. [1st October]. <https://www.who.int/news-room/fact-sheets/detail/road-trafficinjuries>
- [3] SHANTHI A, MARTHA H, ROBYN N. Road-traffic injuries: confronting disparities to address a global-health problem[J]. The Lancet, 2006,367(9521).
- [4] BAGULEY C. THE IMPORTANCE OF A ROAD ACCIDENT DATA SYSTEM AND ITS UTILISATION, 2002.
- [5] Wan Yu. Research on Early Warning Methods of Vehicle Accidents Based on Urban Traffic Accident Data[D]. Wuhan University of Technology, 2020.
- [6] Yang Yanfang. Research on Analysis and Forecast of Urban Road Network Traffic Flow and Accident Early Warning Methods[D]. Beijing Jiaotong University, 2017.
- [7] GUANGNAN Z, KELVIN K W Y, GUANGHAN C. Risk factors associated with traffic violations and accident severity in China[J]. Accident Analysis and Prevention, 2013,59.
- [8] DAVID S, EDNA S, RICHARD C. Self-reports of safe driving behaviors in relationship to sex, age, education and income in the US adult driving population[J]. Accident Analysis and Prevention, 2001,33(1).
- [9] CHANG L Y, WANG H W. Analysis of traffic injury severity: an application of non-parametric classification tree techniques[J]. Accid Anal Prev, 2006,38(5): 1019-1027.
- [10] YU R, ABDEL-ATY M. Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data[J]. Safety science, 2014,63: 50-56.
- [11] ATHANASIOS T, DANIEL G, GEORGE Y. Factors Affecting Accident Severity Inside and Outside Urban Areas in Greece[J]. Traffic Injury Prevention, 2012,13(5).
- [12] AIDOO E N, AMOH-GYIMAH R, ACKAAH W. The effect of road and environmental characteristics on pedestrian hit-and-run accidents in Ghana[J]. Accid Anal Prev, 2013,53: 23-27.
- [13] JIANFENG X, HONGYU G, JIAN T, et al. Analysis of influencing factors for rear-end collision on the freeway[J]. Advances in Mechanical Engineering, 2019,11(7).
- [14] SUN Z, WANG J, CHEN Y, et al. Influence Factors on Injury Severity of Traffic Accidents and Differences in Urban Functional Zones: The Empirical Analysis of Beijing.[J]. International journal of environmental research and public health, 2018,15(12).

- [15]YUHUAN Z, HUAPU L, WENCONG Q. Geographical Detection of Traffic Accidents Spatial Stratified Heterogeneity and Influence Factors[J]. International Journal of Environmental Research and Public Health, 2020,17(2).
- [16]JINBAO Z, WEI D. Traffic Accidents on Expressways: New Threat to China[J]. Traffic Injury Prevention, 2012,13(3).
- [17]Xu Dejie, Qian Yongsheng, Wang Min, et al. Analysis of characteristics of major road traffic accidents in Gansu Province[J]. China Public Safety (Academic Edition), 2011(04): 100-103.
- [18]Yuan Quan, Li Yibing, Chen Kang. Analysis of the Characteristics of the Significant Factors Leading to Major Traffic Accidents and Safety Countermeasures[J]. China Forensic Expertise, 2015(05): 34-40.
- [19]Zhang Yuan, Deng Yuanchang, Lin Qingfeng, et al. Analysis of influencing factors of car accidents with drivers of different genders[J]. Safety and Environmental Engineering, 2019,26(03): 166-170.
- [20]Xu Hongguo, Zhang Huiyong, Zong Fang. Bayesian Network Modeling for the Cause Analysis of Traffic Accidents[J]. Journal of Jilin University(Engineering and Technology Edition), 2011,41(S1): 89-94.
- [21]Zhu Jiazheng. Research on Analysis Model of Factors Affecting Injury and Death Degree of Traffic Accidents in Road Work Area[D]. Beijing Jiaotong University, 2016.
- [22]Ma Zhuanglin, Shao Chunfu, Li Xia. Influencing factors of the severity of road tunnel traffic accidents based on Logistic model[J]. Journal of Jilin University (Engineering and Technology Edition), 2010,40(02): 423-426.
- [23]Chen Yanyan, Li Xiangnan, Sun Zhiyuan, et al. Identification method of influencing factors of road traffic accidents based on Logistics regression[J]. Transportation Technology and Economy, 2018,20(5): 1-5, 11.
- [24]Fan Xiantao. Research on Traffic Safety Risk Analysis of Mountain Expressway under Adverse Weather[D]. Chinese People's Public Security University, 2019.
- [25]Hu Ji, Yan Zhangcun, Lu Xiaozhao, et al. Analysis of influencing factors of traffic accident severity based on ordered Logit and Probit models[J]. Journal of Safety and Environment, 2018,18(03): 836-843.
- [26]Zeng Qiang, Wang Xuesong, Zhang Xuan, et al. Analysis of influencing factors of expressway seasonal accident frequency based on time-space interaction model[J]. China Journal of Highway and Transport, 2020,33(11): 255-263.
- [27]LAMM R, PSARIANOS B, MAILAENDER T. Highway Design and Traffic Safety Engineering Handbook[M]. Donnelley & Sons Company, 1999: 23-56.
- [28]G P, J M A. Travelling Speed and the Risk of Crash Involvement on Rural Roads[J]. Report, 2001.
- [29]MURAT S Ç, OZAN A. A GIS Approach to Evaluate Infrastructure Variables Influencing the Occurrence of Traffic Accidents in Urban Roads[J]. International Journal of Environment and Geoinformatics, 2017,4(1).

- [30] TEMESGEN H H, DESSALEGN O G, ASHENIF M A. Identifying Road Traffic Accidents Hotspots Areas Using GIS in Ethiopia: A Case Study of Hosanna Town[J]. *Transport and Telecommunication Journal*, 2019,20(2).
- [31] M. M Z, TAN S J, NURUL A A B M, et al. Evaluation of various GIS-based methods for the analysis of road traffic accident hotspot[J]. *MATEC Web of Conferences*, 2019,258.
- [32] Gao Tianzhu. Research on the characteristics and prevention of road traffic accidents in my country[D]. Chang'an University, 2014.
- [33] Ding Yulei. Analysis of Characteristics and Influencing Factors of Major Traffic Accidents[D]. Southeast University, 2016.
- [34] Cai Jun. Research on the Occurrence Rules of Traffic Accidents and Urban Road System Planning and Design[J]. *Chinese Safety Science Journal*, 2005(04): 16-20.
- [35] Wu Ruilong, Zhu Xinyan, Su Wei, et al. Analysis on the temporal and spatial distribution pattern of urban road traffic accidents[J]. *Surveying and Spatial Geographic Information*, 2018,41(07): 103-106.
- [36] Guo Xiaokui, Li Rui, Chen Luogang, et al. Research on the temporal and spatial distribution of traffic accidents on suburban roads based on GIS[J]. *Journal of Hebei University of Technology*, 2017,46(05): 62-67.
- [37] Guo Lin, Zhou Jibiao, Dong Sheng, et al. Analysis of urban road traffic accidents based on improved K-means algorithm[J]. *China Journal of Highway and Transport*, 2018,31(04): 270-279.
- [38] Wang Hai. Identification and cause analysis of frequent traffic accidents based on spatial analysis technology[D]. Tsinghua University, 2014.
- [39] Lu Huapu, Luo Shengxi, Li Ruimin. Research on the Spatial Distribution Characteristics of Road Traffic Accidents in Shenzhen Based on GIS Analysis[J]. *China Journal of Highway and Transport*, 2019,32(08): 156-164.
- [40] RAKESH A, TOMASZ I, ARUN S. Mining association rules between sets of items in large databases[J]. *ACM SIGMOD Record*, 1993,22(2).
- [41] Liu Yian, Yang Bin. Research on an improvement of Apriori algorithm in association rule mining[J]. *Journal of Computer Applications*, 2007(02): 418-420.
- [42] Liu Bohai. Research on Influencing Factors of Road Traffic Accidents Based on Association Rules and Spatial Autocorrelation[D]. Beijing Jiaotong University, 2019.
- [43] Zhao Huaixin, Sun Xingxing, Xu Qianqian, et al. Analysis of correlation factors between highway freight volume and cargo turnover based on grey entropy method[J]. *Journal of Traffic and Transportation Engineering*, 2018,18(04): 160-170.
- [44] SHUAI Y, YUANHUA J, DONGYE S. Identifying Factors that Influence the Patterns of Road Crashes Using Association Rules: A case Study from Wisconsin, United States[J]. *Sustainability*, 2019,11(7).

- [45]FAISAL M N A, ABDELMONEIM A M H. Usage Apriori and clustering algorithms in WEKA tools to mining dataset of traffic accidents[J]. Journal of Information and Telecommunication, 2018,2(3).
- [46]FATIMA Z E M, MOHAMMED C A, KHALID Z. Data mining combined to the multicriteria decision analysis for the improvement of road safety: case of France[J]. Journal of Big Data, 2019,6(1).
- [47]Hu Zhenbo. Data mining of traffic accidents based on association rules[D]. Tianjin University, 2017.
- [48]Liu Yunxiang, Han Bei. Correlation analysis of tunnel traffic accidents based on improved FP algorithm[J]. Modern Electronic Technology, 2018,41(17): 141-144.
- [49]Chen Jingxi. Analysis of traffic accidents based on K-Means[D]. Chang'an University, 2018.
- [50]Deng Xiaoheng, Zeng Detian. Cause analysis model of traffic accidents based on AHP and hybrid Apriori-Genetic algorithm[J]. Application Research of Computers, 2019,36(06): 1633-1637.
- [51]Jing Liantian. Research on the relationship between non-extreme severe weather conditions and highway traffic accidents[J]. Highway, 2015,60(01): 114-119.
- [52]Yang Jianhong. Association analysis of data mining and its application in road traffic accidents[D]. Guangzhou University, 2018.
- [53]ANURAG P, MOHAMED A. Market basket analysis of crash data from large jurisdictions and its potential as a decision support tool[J]. Safety Science, 2007,47(1).
- [54]L. L, S. S, G. H. Analysis of road traffic fatal accidents using data mining techniques: 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA)[C], 2017.
- [55]Dong Liyan, Liu Guangyuan, Yuan Senmiao, et al. Application of data mining technology in traffic accident analysis[J]. Journal of Jilin University (Science Edition), 2006(06): 951-955.
- [56]Shang Wei, Shang Ning, Qin Minggui, et al. Multidimensional association rules analysis of traffic accidents[J]. Computer Applications and Software, 2006(02): 40-42.
- [57]Zhang Xin, Guan Zhichao, Yang Dongyuan. Research on Urban Traffic Simulation Algorithm Based on Multi-Object Data Mining[J]. Journal of Sun Yat-sen University (Natural Science Edition), 2007(S2): 210-214.
- [58]Wu Hao, Li Junguo. Road traffic accident data mining model based on association rule theory[J]. Application of Electronic Technology, 2009,35(02): 135-138.
- [59]WENG J, ZHU J, YAN X, et al. Investigation of work zone crash casualty patterns using association rules.[J]. Accident; analysis and prevention, 2016,92.

- [60]Hong Yuanyuan. Cause analysis of low-grade highway traffic accidents based on multi-dimensional and multi-layer association rules mining[D]. Southeast University, 2017.
- [61]Zhou Jin. Research on the Causes and Safety Risk Analysis Methods of Railway Transportation Accidents[D]. Beijing Jiaotong University, 2018.
- [62]KIM J, KIM S, ULFARSSON G F, et al. Bicyclist injury severities in bicycle-motor vehicle accidents.[J]. Accident; analysis and prevention, 2007,39(2).
- [63]Liu Haizhu. Analysis of factors affecting the severity of road traffic accidents and establishment of prediction models[D]. Jilin University, 2014.
- [64]ZHAO S, KHATTAK A. Motor vehicle drivers' injuries in train-motor vehicle crashes[J]. Accident Analysis and Prevention, 2015,74.
- [65]ÇELİK A K, OKTAY E. A multinomial logit analysis of risk factors influencing road traffic injury severities in the Erzurum and Kars Provinces of Turkey.[J]. Accident; analysis and prevention, 2014,72.
- [66]KARA M K, YOUNG-JUN K. Driver injury severity: an application of ordered probit models[J]. Accident Analysis and Prevention, 2002,34(3).
- [67]DELEN D, SHARDA R, BESSONOV M. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks.[J]. Accident; analysis and prevention, 2006,38(3).
- [68]SO Y S, HYUNGWON S. Pattern recognition for road traffic accident severity in Korea[J]. Ergonomics, 2001,44(1).
- [69]BUDIAWAN W, SAPTADI S, SRIYANTO, et al. Traffic Accident Severity Prediction Using Naive Bayes Algorithm - A Case Study of Semarang Toll Road[J]. IOP conference series. Materials Science and Engineering, 2019,598(1): 12089.
- [70]Sun Yixuan. Analysis of road traffic accidents based on data mining[D]. Beijing Jiaotong University, 2014.
- [71]Wang Lei, Qiu Feng, Xia Yongxu, et al. Road tunnel traffic accident prediction based on road environmental factors[J]. Tunnel Construction (Chinese and English), 2019,39(08): 1301-1307.
- [72]Shi Xuehuai, Qi Yong, Zhang Weibin, et al. Traffic accident severity prediction method based on combined model[J]. Application Research of Computers, 2019,36(08): 2395-2399.
- [73]PEARL J. Fusion, Propagation, and Structuring in Belief Networks[J]. Artificial Intelligence, 1986,29(3): 241-288.
- [74]Li Hongxi, Zhang Lianfeng, Zheng Zhongyi. Research on the Cause Chain of Ship Man-made Collision Based on Data Mining[J]. Journal of Dalian Maritime University, 2014,40(02): 10-12.
- [75]Si Dongsen, Zhang Yingjun, Lang Kun. Cause analysis of container ship collision accident based on improved BN[J]. Chinese Safety Science Journal, 2019,29(10): 31-37.
- [76]Qian Chao, Dai Liang, Lin Shan, et al. Traffic state prediction based on random repeated hill climbing method[J]. Transportation Systems Engineering and Information, 2016,16(01): 162-168.

- [77]Liu Haoran, Zhang Liyue, Fan Ruixing, et al. Bayesian network structure learning algorithm based on improved whale optimization strategy[J]. Journal of Electronics & Information Technology, 2019,41(06): 1434-1441.
- [78]Wang Chunfeng, Jiang Yan. Bayesian network structure hybrid learning method based on bee colony algorithm and genetic algorithm[J]. Journal of Henan Normal University (Natural Science Edition), 2015,43(04): 16-20.
- [79]HADDAWY P. An Overview of Some Recent Developments in Bayesian Problem-Solving Techniques[J]. AI Magazine, 1999,20(2): 11-19.
- [80]XIAOLI M, YINGYING X, JIAN L, et al. Causation Analysis of Hazardous Material Road Transportation Accidents by Bayesian Network Using Genie[J]. Journal of Advanced Transportation, 2018,2018.
- [81]Lu Yao. Research on Bayesian Network Model Diagnosis and Reasoning of Mountain Expressway Traffic Accidents[D]. Huazhong University of Science and Technology, 2019.
- [82]Liu Zhiqiang, Wang Ling, Zhang Aihong, et al. Research on the Occurrence Mechanism of Highway Traffic Accidents in Smoggy Days Based on Bayesian Model[J]. Journal of Chongqing University of Technology (Natural Science), 2018,32(01): 43-49.
- [83]Xiong Xiaoxia, Chen Long, Liang Jun, et al. Research on Generation and Evolution of Road Traffic Accident Chain Based on Bayesian Network Model[J]. Highway and Transportation Science and Technology, 2018,35(05): 99-107.
- [84] Savolainen P T,M annering F L,Lord D,et al.The statis- tical analysis of highway crash injury severities: are- view and assessment of methodological alternatives [J] ,Accident Analysis & Prevention,2011,43 (5): 1665 -1667.
- [85] Liu Bohai. Research on influencing factors of road traffic accidents based on association rules and spatial autocorrelation[D]. Beijing: Beijing Jiaotong University, 2019.