Australasian Transport Research Forum 2023 Proceedings
29 November – 1 December, Perth, Australia
Publication website: http://www.atrf.info

Predicting injury risk using Big Data: The case of Metropolitan Melbourne

Ali Soltani [1 2] , Betsabeh Tanoori[3] , Christopher J. Pettit [4]

[1] Injury Studies, Flinders University, Bedford Park, SA, 5042, Australia

[2] UniSA Business, University of South Australia, SA, 5001, Australia

[3] Department of Computer Engineering, Shiraz University, Shiraz, Iran

[4] City Futures Research Centre, University of New South Wales, Sydney, 2052, Australia

Email for correspondence (presenting author): ali.soltani@flinders.edu.au

Abstract

This research paper investigates the use of big data to analyse road accidents in the Melbourne metropolitan area. Using the Crash dataset and the Random Forest model, the study sought to determine the relationship between various factors and the proportion of accident victims who were maimed or slain. Results demonstrated that factors such as weekdays, months, and weather conditions can influence the severity of accidents. The study provides policymakers and transportation authorities with valuable insights for devising strategies to enhance road safety and reduce accident risk. Additional research can investigate other potential accident severity factors.
Keywords: Big data; Road crashes; Machine Learning; Correlation analysis; Victoria.

1. Introduction

Big data plays a key role in identifying road collision trends, which informs policies and measures to enhance road safety. Researchers may find patterns and trends in road crashes by analysing massive volumes of data from crash records, weather reports, and traffic volume data (Pika et al., 2021). This may involve identifying high-risk junctions, road segments, and characteristics that cause crashes, such as speeding, distracted driving, and bad weather (Jiang et al., 2017). Policymakers and transportation agencies may use this information to plan targeted actions to minimise the frequency and severity of road crashes, such as better road infrastructure, greater law enforcement, or public awareness campaigns.

Technology and data availability are expanding the role of big data in collision and traffic accidents. Connected devices and the Internet of Things (IoT) provide a variety of transportation data, including traffic volume, weather, and vehicle speed (Kitchin, 2014). Data analytics technologies and methods have made processing and analysing large volumes of data faster and simpler. This helps academics and policymakers to uncover patterns and trends in road crashes. Transportation organisations and policymakers may use big data to enhance road safety and lower the frequency and severity of crashes. Thus, big data's role in road safety will likely increase in the future.

Using big data for road safety research has several benefits: By combining and analysing massive volumes of data from many sources, big data may provide a completer and more accurate picture of road safety trends than traditional methods of data collection and analysis (Wang et al., 2018). Big data may be used to discover patterns and trends in road crashes, such as high-risk zones or particular factors that cause crashes. By keeping tabs on traffic and road

conditions in real time, big data may improve emergency response times (Papadimitriou et al., 2019). Policymakers and transportation organisations may utilise big data to enhance road safety by identifying particular risk factors and implementing focused interventions to address the core causes of crashes. Through the use of big data analytics, we can determine which road safety measures should be prioritised for funding and where those funds would have the greatest impact.

Australia's road safety issues include lengthy distances between major population centres, a significant percentage of rural roads, and a high number of heavy vehicles (Naweed et al., 2014; Faulks, 2012). Compared to countries with more densely populated regions and shorter travel distances, these characteristics may lead to greater incidence of crashes and deaths.

Despite the above, Australia has made significant progress in road safety over the past few decades, but more can be done (World Health Organisation, 2015). In terms of fatality rate per 100,000 population in 2020, Australia ranked 20th out of 36 countries, with a rate of 4.26. Australia's fatality rate decreased by 25.4% between 2011 and 2020. During the same time period, the OECD median rate decreased by 34.6% (BITRE, 2022).

Road crashes are a significant source of injury and death in Australia (Scott-Parker & Oviedo-Trespalacios, 2017), hence road crash studies are crucial. The Australian Institute of Health and Welfare reports that road crashes are the leading cause of death for 15–24-year-olds and the second leading cause for 25–34-year-olds. Road crashes have a significant economic effect in addition to the human cost (Blincoe et al., 2002; Clements & Kockelman, 017).

These data demonstrate Australia's need to improve road safety (Malik et al., 2020). Road collision studies help experts understand crash causes and find ways to lessen their frequency and severity. This information may help save lives, decrease injuries, and lower the economic consequences of road crashes. Thus, road crash studies improve road safety and protect Australians' health and well-being.

The purpose of this paper is to investigate road safety elements in Greater Melbourne, Victoria, by analysing 15 years (2006-2020) of car crash data to identify patterns and trends. The research topic may therefore be formulated as follows: What are the most notable patterns and trends in road safety elements in Greater Melbourne, Victoria, over a 15-year period based on car crash data, and how can this information be used to enhance road safety?

1. Background

Due to the numerous factors that contribute to road accidents, the use of big data crash analysis is becoming more prevalent. To predict the factors contributing to road accidents, the application of large-scale machine learning techniques with the Apache Spark framework and Decision Tree algorithms has been explored. According to Ait-Mlouk et al. (2017), the proposed model can assist decision-makers with road safety analysis and improvement.

In addition to data integration, support vector machines, correlation machines, and multinomial goodness have been utilised to predict traffic data. Lokala et al. (2017) explored the use of SVM's methodology for text classifiers based on traffic data. Through data preprocessing and data selection, the R programming language has been used to analyse traffic data and visualisations and to select high-frequency accident locations. Chen (2017) also modelled data using decision tree, linear regression, and random forest algorithms, and validated the model's accuracy.

Xie et al. (2017) have highlighted the benefit of big data analysis in providing more accurate estimations of the effects of risk factors and facilitating the identification of large-scale hotspots. The Hadoop framework is recommended by Park et al. (2016) for the development of a problem-solving, problem-predicting model. In addition, they used a sampling technique to balance the data, corrected the data by classifying it into groups, and applied classification analysis to enhance the accuracy of their predicting.

Shao et al. (2020) investigated the effect of updating rail equipment using the maximal information coefficient (MIC) to examine the big data analysis of rail equipment accidents. The application of big data analysis has produced promising results in predicting and preventing road accidents. Utilising various algorithms, frameworks, and techniques in data analysis has allowed for more precise predictions, identification of hotspots, and resolution of data imbalance. Further studies on big data analysis decision and its application in road safety can provide decision-makers in the transportation industry with valuable insights.
There have been numerous studies on the use of big data in crash analysis, but there are still certain research gaps that need to be filled (Marjani et al., 2017; fan et al., 2014; Wamba et al., 2015; Chen et al., 2015). The gathering and analysis of crash data are not standardised. This might cause data errors and make it hard to compare studies. This hinders detailed studies and forecast model accuracy. While crash data is essential for understanding crash patterns and identifying risk factors, integrating other data sources such as weather, traffic, and driver behaviour can provide more comprehensive insights into crash causes. Most crash data is acquired retrospectively, limiting real-time judgements and interventions. Real-time data collection and analysis may speed up interventions and minimise crashes. While most crash studies focus on motor vehicle crashes, pedestrians and bicyclists need to be included in crash analysis. This may detect distinct risk variables and enhance road safety interventions. The advantages of big data analysis must be delivered equally among populations. Including diverse road users in data gathering and analysis is important.

2. Methodology

The methodology employed in this paper involves using big data to analyse road crashes. Specifically, we utilized the Victoria Crash dataset, which contains 65535 samples from greater Melbourne Area. Our objective was to evaluate the importance of features and develop two models. Random Forest is selected as the suitable type for modelling.
Random Forest is a common ensemble learning approach for classification and regression. Multiple decision trees are grown independently on a random subset of training data and features (Katuwal et al., 2018). The decision trees are trained via bootstrapped aggregation, or bagging, in which each decision tree is trained on a separate random subset of the training data. The Random Forest technique separates each decision tree node depending on the feature with the largest information gain during training (Breiman, 2001). Information gain measures how much a feature reduces prediction uncertainty. The technique does this recursively until a stopping requirement is satisfied, such as a maximum tree depth or a minimum number of samples per leaf node. After all decision trees are trained, the algorithm predicts the class (classification) or value (regression) by aggregating their predictions. The most frequent prediction among all trees is used in classification. The average of tree predictions is the final prediction in regression. Random Forest outperforms several other machine learning models in accuracy, scalability, and overfitting resistance. It can handle high-dimensional datasets with many features (Hegelich, 2016).

3. Results

4.1 Correlation studies

In this investigation, the target feature was PerNum1, which represents the proportion of people injured or killed in an accident. To investigate the relationship between this target feature and

various factors, scatter plots were constructed and Pearson correlation coefficients were calculated for the Weekdays, Months, and Air Condition features.
Figure 1 depicts the correlation between PerNum1 and the various values of the weekday feature through a scatter plot. The results show that Saturdays (0.029) and Sundays (0.043) have the highest correlation values with the number of people injured or killed in accidents, whereas Mondays (0.002) and Thursdays (0.017) have the lowest correlation values with crash severity, everything else being equal. Notably, the negative correlation between Fridays, Thursdays, and Tuesdays and severity of crashes suggests that accidents on these days are less severe.

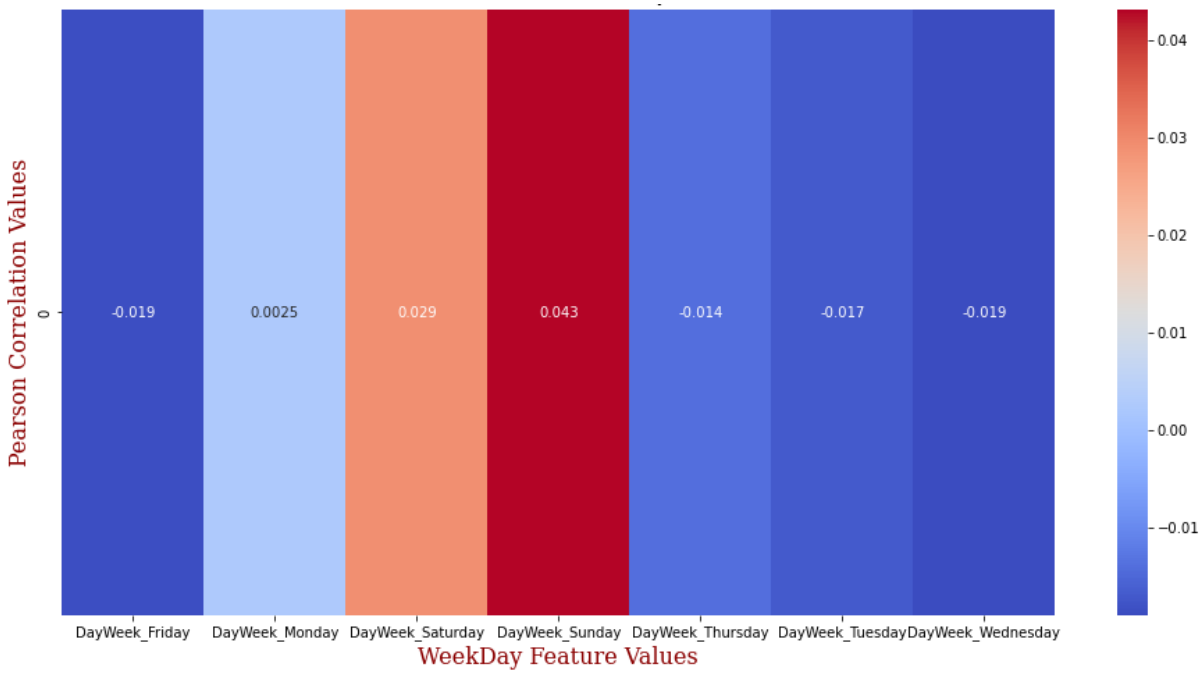

Figure 1: Correlation between crash severity and days

Figure 2 illustrates the correlation between PerNum1 and the various months. The results indicate that January has the highest correlation (0.018) with the number of people injured or killed in accidents, followed by December (0.014). May (0.001), October (-0.003), and April (0.001) are the months with the lowest correlation between month and crash severity. Intriguingly, the negative correlations between the months of August, February, July, and June, October, and September imply that accidents in these seasons occur with less severity, resulting in lower costs and damages.

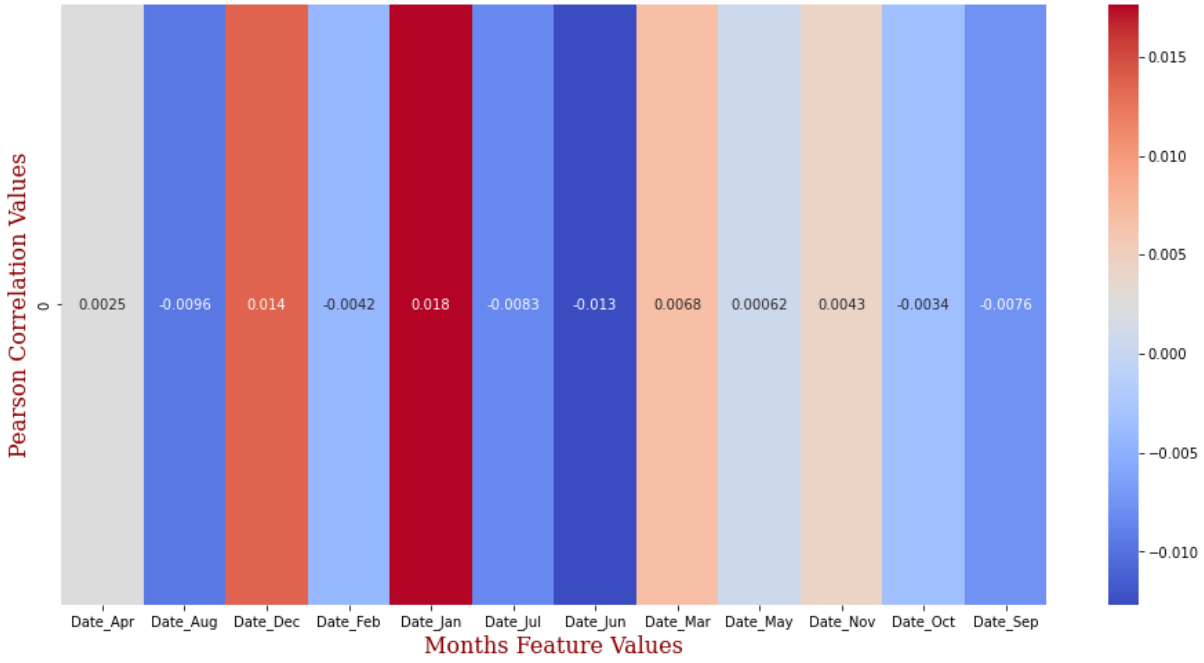

Figure 2: Correlation between crash severity and months

Figure 3 depicts the correlation between PerNum1 and the various values of the Air Condition feature. The results indicate that Fog (0.027), Rainy (0.039), and Strong Wind (0.02) weather conditions have the highest correlation with the number of people injured or killed in accidents, making them the worst weather conditions in terms of road safety. Intriguingly, the negative correlation between clear air conditions and crash severity suggests that in clear weather, the severity of incidents is low. In addition, snowing has a weak correlation with crash severity, which may be a result of the small proportion of snow-covered days.

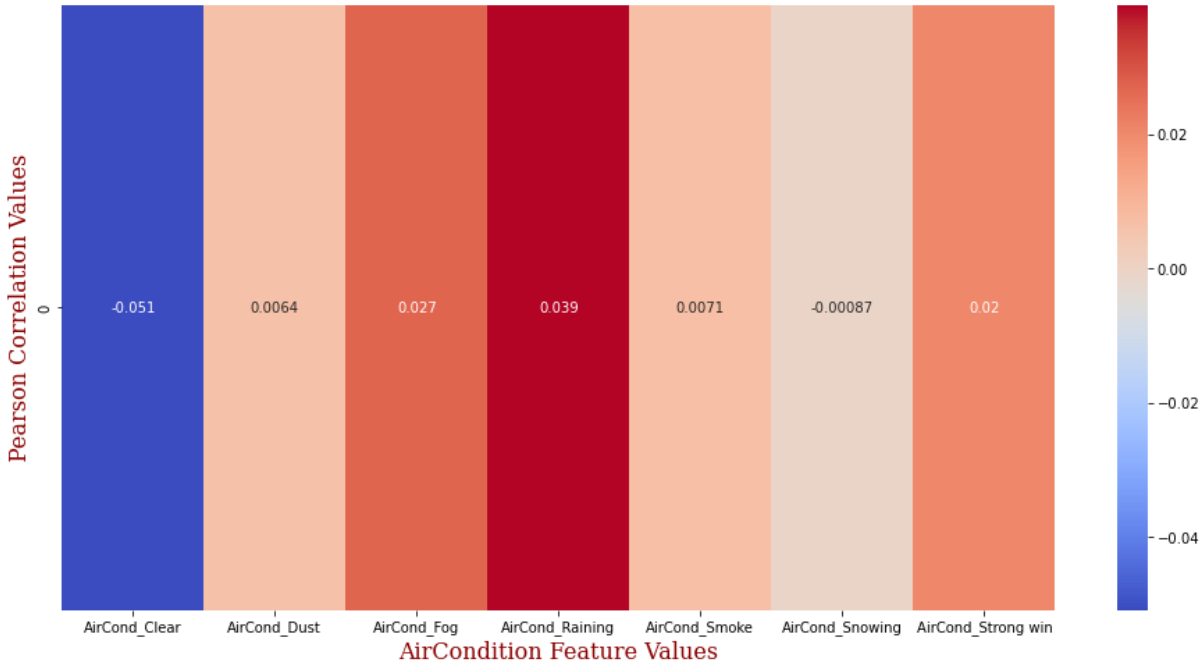

Figure 3: Correlation between crash severity and weather condition

The RF model attempts to predict the percentage of individuals injured or slain in the accident (as the severity index), which is a regression task employing the PerNum1 target column. The

range of target values for this regression task is [0,1]. To complete these tasks, we followed a three-step procedure consisting of preprocessing, cross-validation, and training and evaluation. In the subsequent sections of the paper, each step's specifics are described.

4.2 Pre-processing

Our dataset with mixed numeric and categorical features, duplicated samples, and missing values required management during the preprocessing phase. First, we eliminated duplicate samples based on accident numbers, resulting in 58,982 unique samples. Subsequently, we addressed missing values in features such as LGA_NAME, Light, TypeInters, SpeedZone, and AirCond, where undetermined values were considered missing. Filling in missing values is crucial for attaining an accurate model, so we utilised the most frequent approach by replacing missing values with the most frequent value in the entire feature column. In addition, we converted the time attribute into two attributes, AM_Peak and PM_Peak, each with a value of one between 6:30 and 9:30 AM and 15:30 and 18:30 PM. After completing the missing values stage, categorical features were converted to numeric values using one-hot encoding. This method generates a new dummy feature for each unique value in the nominal feature column, resulting in a binary value that specifies the location of an example. Finally, one-hot encoding features and numeric features were concatenated to generate a feature vector of length 67. The following table lists all features and their respective categories.
We utilised one-hot encoding to convert categorical features into numeric values in order to represent our features numerically. This method required the creation of a new artificial feature for every unique value in the nominal feature column. We converted the "Location" feature into the following five new features: TOWNS, SMALL_TOWNS, MELB_URBAN, SMALL_CITIES, and RURAL_VICTORIA. Then, binary values were used to specify the precise location of an example. For example, if a sample's "Location" feature value was RURAL_VICTORIA, it would be encoded as RURAL_VICTORIA=1, SMALL_TOWNS=0, MELB_URBAN=0, SMALL_CITIES=0, TOWNS=0.

Table 1: The features used in training models

features	Type	Size of numeric feature vector
LGA_NAME	categorical	19
Location	categorical	5
Date	categorical	12
DayWeek	categorical	7
Light	categorical	6
TypeInters	categorical	8
AirCond	categorical	7
SpeedZone	numeric	1
AM_Peak	numeric	1
PM_Peak	numeric	1
Total number of features		67

4.3 Cross-validation

For fitting and evaluating the models, the K-fold cross-validation method was utilised. The dataset was randomly divided into k-folds without replacement, with (k - 1) folds used as the training set and one fold used to evaluate model performance. This procedure was repeated k times in order to acquire k models and performance estimates, with the final model performance being the mean of the k folds. Both the classification and regression problems were solved using a k value of 10.

4.4 Classification task

The objective of the regression assignment was to predict the proportion of individuals who were injured or slain in an accident; the Random Forest model was utilised for this purpose. The squared error criterion was utilised to train the model with a maximal tree depth of 10 and 100 trees. To evaluate the performance of the model, 10-fold cross-validation was applied, resulting in 53084 training samples and 5898 testing samples. The average importance of the forty features from the 10-folds cross-validation is depicted in Table 2, and Figure 4 depicts the ten most significant features. The 10-fold cross-validation yielded an average root mean square error (RMSE) of 0.2021, and $\pm$ 0.01 standard deviation (SD), which was used to evaluate the model's performance.

Table 2: The average importance of 40 features

FEATURE NAME	IMPORTANCE
1) TYPE_COLLISION WITH A FIXED OBJECT	0.4940
2) TYPE_VEHICLE OVERTURNED (NO COLLISION)	0.1157
3) TYPE_NO COLLISION AND NO OBJECT STRUCK	0.1157
4) DCA_VEHICLE COLLIDES WITH VEHICLE PARKED ON LEFT OF ROAD	0.0948
5) DCA_REAR END(VEHICLES IN SAME LANE)	0.0337
6) TYPE_COLLISION WITH SOME OTHER OBJECT	0.0191
7) DCA_OTHER ON PATH	0.0138
8) SPEEDZONE	0.0099
9) TYPE_STRUCK ANIMAL	0.0097
10) DCA_HEAD ON (NOT OVERTAKING)	0.0090
11) DCA_OUT OF CONTROL ON CARRIAGEWAY (ON STRAIGHT)	0.0035
12) DCA_STRUCK ANIMAL	0.0025
13) DCA_LOAD OR MISSILE STRUCK VEHICLE	0.0025
14) TYPE_FALL FROM OR IN MOVING VEHICLE	0.0024
15) TYPEINTERS_NOT AT INTERSECTION	0.0019
16) DAYWEEK_TUESDAY	0.0019
17) DAYWEEK_SATURDAY	0.0019

18) LIGHT_DARK STREET LIGHTS ON	0.0018
19) LIGHT_DAY	0.0018
20) DAYWEEK_SUNDAY	0.0017
21) TYPEINTERS_CROSS INTERSECTION	0.0016
22) LGA_NAME_CASEY	0.0016
23) AIRCOND_RAINING	0.0015
24) DAYWEEK_FRIDAY	0.0015
25) DATE_JUL	0.0015
26) LGA_NAME_MONASH	0.0015
27) DCA_LEFT OFF CARRIAGEWAY INTO OBJECT/PARKED VEHICLE	0.0015
28) DAYWEEK_WEDNESDAY	0.0014
29) DATE_JAN	0.0014
30) LGA_NAME_MORNINGTON PENINSULA	0.0014
31) DAYWEEK_THURSDAY	0.0014
32) LGA_NAME_FRANKSTON	0.0014
33) DATE_JUN	0.0013
34) DATE_OCT	0.0013
35) LOCATION_RURAL_VICTORIA	0.0013
36) LOCATION_MELB_URBAN	0.0013
37) AIRCOND_CLEAR	0.0012
38) DATE_SEP	0.0012
39) DATE_AUG	0.0012
40) DATE_APR	0.0012

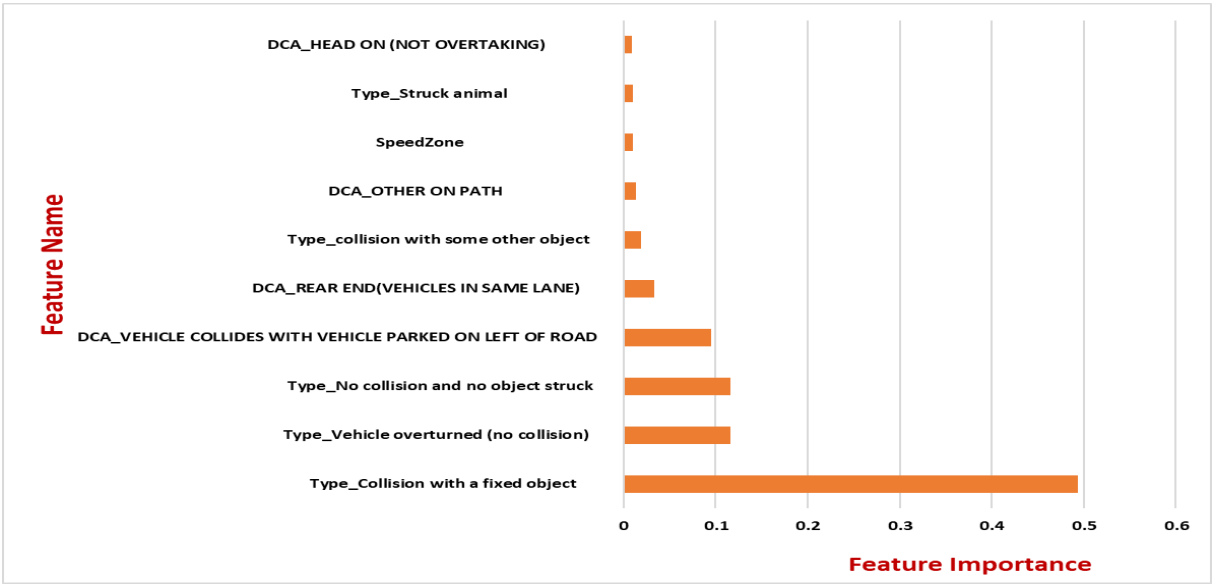

Figure 4: The important top 10 features

5. Conclusion

The contribution of this research is the application of big data analytics to the analysis of road collisions and the evaluation of the significance of various factors in predicting the severity of accidents. Specifically, the research demonstrates the use of the Victorian Crash dataset to examine the relationship between the percentage of persons injured or killed in an accident and factors such as weekdays, months, and weather. In addition, the study constructs and evaluates a Random Forest model for predicting the percentage of accident victims who were injured or killed.

The study's findings can aid policymakers and transportation authorities in devising strategies to enhance road safety and reduce the likelihood of accidents. In addition, the study emphasises the potential of big data analytics for analysing road accident data and identifying significant factors that contribute to accident severity. The study also identifies several directions for future research, such as investigating additional factors that may contribute to the severity of accidents, such as road conditions, driver behaviour, and vehicle type.

References

Ait-Mlouk, A., Gharnati, F., & Agouti, T. (2017). Application of big data analysis with decision tree for road accident. Indian Journal of Science and Technology, 10(29), 1-10.

Australian Institute of Health. (2012). A picture of Australia's children 2012. AIHW.

Bureau of Infrastructure and Transport Research Economics (BITRE) (2022). International road safety comparisons 2020 BITRE, Canberra ACT.

Blincoe, L. J., Seay, A. G., Zaloshnja, E., Miller, T. R., Romano, E. O., Luchter, S., & Spicer, R. S. (2002). The economic impact of motor vehicle crashes, 2000 (No. DOT-HS-809-446). United States. National Highway Traffic Safety Administration.

Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

Chen, C. (2017). Analysis and forecast of traffic accident big data. In ITM Web of Conferences (Vol. 12, p. 04029). EDP Sciences.

Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A. V., & Rong, X. (2015). Data mining for the internet of things: literature review and challenges. International Journal of Distributed Sensor Networks, 11(8), 431047.
Clements, L. M., & Kockelman, K. M. (2017). Economic effects of automated vehicles. Transportation Research Record, 2606(1), 106-114.
Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. National science review, 1(2), 293-314.
Faulks, I. J., Lane, M., & Irwin, J. D. (2012, August). Traffic policing and road safety for individuals and for populations. In ACRS National Conference–'A Safe System: Expanding the reach.
Hegelich, S. (2016). Decision trees and random forests: machine learning techniques to classify rare events. European Policy Analysis, 2(1), 98-120.
Jiang, Z., Shekhar, S., Jiang, Z., & Shekhar, S. (2017). Spatial Big Data. Spatial Big Data Science: Classification Techniques for Earth Observation Imagery, 3-13.
Katuwal, R., Suganthan, P. N., & Zhang, L. (2018). An ensemble of decision trees with random vector functional link networks for multi-class classification. Applied Soft Computing, 70, 1146-1153.
Kitchin, R. (2014). The data revolution: Big data, open data, data infrastructures and their consequences. Sage.
Lokala, U., Nowduri, S., & Sharma, P. K. (2017). Road accidents bigdata mining and visualization using support vector machines. World Academy of Science, Engineering and Technology-International Journal of Computer and Systems Engineering, 10(8).
Malik, S., Swapan, M. S. H., & Khan, S. (2020). Sustainable mobility through safer roads: Translating road safety strategy into local context in western australia. Sustainability, 12(21), 8929.
Marjani, M., Nasaruddin, F., Gani, A., Karim, A., Hashem, I. A. T., Siddiqa, A., & Yaqoob, I. (2017). Big IoT data analytics: architecture, opportunities, and open research challenges. IEEE access, 5, 5247-5261.
Naweed, A., Balakrishnan, G., & Thomas, M. J. (2014). Challenges facing simulator use in transportation research: lessons from a road safety case study. Road & Transport Research: A Journal of Australian and New Zealand Research and Practice, 23(2), 60-70.
Papadimitriou, E., Filtness, A., Theofilatos, A., Ziakopoulos, A., Quigley, C., & Yannis, G. (2019). Review and ranking of crash risk factors related to the road infrastructure. Accident Analysis & Prevention, 125, 85-97.
Park, S. H., Kim, S. M., & Ha, Y. G. (2016). Highway traffic accident prediction using VDS big data analysis. The Journal of Supercomputing, 72, 2815-2831.
Pika, A., ter Hofstede, A. H., Perrons, R. K., Grossmann, G., Stumptner, M., & Cooley, J. (2021). Using big data to improve safety performance: an application of process mining to enhance data visualisation. Big Data Research, 25, 100210.
Scott-Parker, B., & Oviedo-Trespalacios, O. (2017). Young driver risky behaviour and predictors of crash risk in Australia, New Zealand and Colombia: Same but different?. Accident Analysis & Prevention, 99, 30-38.
Shao, F., Yang, S., Sun, B., Jia, L., Dong, Y., & Wang, D. (2020). The big data analysis of rail equipment accidents based on the maximal information coefficient. Journal of Transportation Safety & Security, 12(7), 959-976.
Soltani, A., & Askari, S. (2017). Exploring spatial autocorrelation of traffic crashes based on severity. Injury, 48(3), 637-647.
Soltani, A., Azmoodeh, M., & Qadikolaei, M. R. (2023). Road crashes in Adelaide metropolitan region, the consequences of COVID-19. Journal of Transport & Health, 30, 101581.

Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How ‘big data’can make big impact: Findings from a systematic review and a longitudinal case study. International journal of production economics, 165, 234-246.
Wang, Y., Kung, L., Wang, W. Y. C., & Cegielski, C. G. (2018). An integrated big data analytics-enabled
World Health Organization. (2015). Global status report on road safety 2015. World Health Organization.
Xie, K., Ozbay, K., Kurkcu, A., & Yang, H. (2017). Analysis of traffic crashes involving pedestrians using big data: Investigation of contributing factors and identification of hotspots. Risk analysis, 37(8), 1459-1476.