ASEAN
Journal of Scientific and Technological Reports
Online ISSN:2773-8752

Research article

Profiling Festival-Period Traffic Accidents in Thailand: Clustering and Risk Factors

Witchaya Rattanametawee[1], Sriamporn Rebankoh[2], Khwansiri Sirimangkhala[3], and Naowarat Manitcharoen[4]*

[1] Faculty of Science, Mahasarakham University, Mahasarakham, 44150, Thailand
[2] College of Industrial Technology, King Mongkut's University of Technology North Bangkok, Bangkok, 10800, Thailand
[3] Big Data Institute (Public Organization), Ministry of Digital Economy and Society, Bangkok, 10900, Thailand
[4] College of Industrial Technology, King Mongkut's University of Technology North Bangkok, Bangkok, 10800, Thailand
* Correspondence: naowarat.n@cit.kmutnb.ac.th

Citation:
Rattanametawee, W.; Rebankoh, S; Sirimangkhala, K.; Manitcharoen, N. Profiling festival-period traffic accidents in Thailand: clustering and risk factors. *ASEAN J. Sci. Tech. Report.* **2025**, *28*(5), e259830. https://doi.org/10.55164/ajstr.v28i5.259830.

Article history:
Received: June 17, 2025
Revised: August 30, 2025
Accepted: September 6, 2025
Available online: August 14, 2025

Abstract: The analysis of road accident data currently focuses on collecting and examining information from multiple perspectives, including factors such as the date and time of occurrence, details of the injured parties, alcohol consumption, and safety measures such as seatbelt and helmet usage. This study employs clustering techniques to group similar accident events. In addition to clustering, this study integrates logistic regression and decision tree techniques to enhance predictive capabilities. Logistic regression is used to estimate the probability of accident severity based on contributing factors, while decision tree modeling helps identify key decision rules that influence accident outcomes. To analyze the severity of road accidents, logistic regression can be employed to model the probability of severe outcomes based on contributing factors. A case study in Thailand is conducted to explore accident trends, which helps develop effective safety measures and policies. The findings emphasize the need to refine procedures for transferring emergency patients and implement stricter safety protocols to enhance care efficiency and mitigate adverse consequences.

Keywords: Clustering; graffic accident; logistic regression.

1. Introduction

Road accidents during the Thai New Year and Songkran festivals are a significant public safety concern, with Thailand experiencing the highest road traffic injury (RTI) mortality rate globally. The festive periods, particularly from April 11 to 17, see a marked increase in accidents, attributed to factors such as increased travel, alcohol consumption, and inadequate road safety awareness. Thailand ranks first in road accident fatalities, with a rate of 36.2 per 100,000 people [1]. During Songkran, the number of accidents significantly exceeds normal days, with a notable increase in severe injuries and fatalities [2, 3]. Key risk factors include alcohol consumption, age, and driving behavior [4]. EMS utilization during these holidays is linked to higher mortality rates among severe RTI patients, indicating potential inefficiencies in emergency response [3]. Despite efforts to improve road safety during these festive periods, the persistent high rates of accidents suggest a need for enhanced public awareness and preventive measures. Addressing the underlying causes of these accidents remains crucial for reducing fatalities and injuries. These pose a significant concern for public health and safety. The high frequency of accidents during

these periods necessitates an in-depth analysis to identify underlying patterns and risk factors.

This study aims to address this gap by conducting a clustering analysis of traffic accident data, with a specific focus on the New Year and Songkran festivals in Thailand. This research holds substantial importance for several reasons: Identification of Risk Factors. By employing clustering techniques, the study will identify distinct groups of accidents based on shared characteristics. The results will enable the pinpointing of specific risk factors. Evaluation of Existing Measures: The findings can be used to evaluate the effectiveness of existing safety measures and identify areas where improvements are needed. Policy Recommendations: The results of this research can be used to inform the development of evidence-based policies and programs aimed at reducing road traffic accidents during festive periods. The specific objectives of this study are to analyze the patterns and characteristics of road traffic accidents that occur during the New Year and Songkran festivals in Thailand, identify the key risk factors associated with these accidents, and develop clusters of accidents based on shared characteristics. The study will cover a specific timeframe, from 2008 to 2014, encompassing both the New Year and Songkran festivals. The analysis will focus on national-level data collected by the National Institute for Emergency Medicine.

Logistic regression is a widely used statistical method for classification tasks, particularly in predicting binary outcomes. It estimates the probability of a dependent variable belonging to a specific category (e.g., success/failure) based on one or more independent variables. By applying the logistic function, the model ensures that predictions remain within the range of 0 to 1, making it well-suited for binary classification problems [5]. Logistic regression parameters are estimated using the maximum likelihood method, which iteratively determines the model that best fits the data. The model's coefficients represent the effect of independent variables on the log-odds of the predicted outcome, making their interpretation essential for understanding the influence of different factors [6]. Decision trees are a versatile and interpretable machine learning tool used across various fields for classification and regression tasks. They mimic human decision-making by creating a tree-like model that splits data based on the most informative features. This method is particularly valued for its ability to handle both categorical and numerical data, as well as its ease of understanding for human users [7]. While decision trees offer significant advantages in terms of interpretability and ease of use, they may not always perform as well as more complex models in certain scenarios, particularly when dealing with high-dimensional data or intricate relationships [8]. Clustering methodologies are of paramount importance in the realms of data mining and machine learning, facilitating the aggregation of data points based on their inherent similarities without the need for pre-established labels. A multitude of approaches are available, each characterized by distinctive algorithms and practical applications, rendering clustering a highly adaptable instrument across a diverse array of disciplines. K-means is the most widely used method, where data points are assigned to clusters based on their proximity to centroids [9]. It excels in small to medium datasets but struggles with larger datasets [10]. A comparative analysis reveals that, although K-means enjoys widespread recognition, Nonnegative Matrix Factorization (NMF) frequently achieves superior accuracy, particularly within lower-dimensional datasets [11]. Clustering methodologies are integral to big data analytics, facilitating the categorization of extensive datasets to derive enhanced insights. While clustering methodologies possess substantial analytical power, they are not immune to challenges, including susceptibility to noise and the need for parameter calibration, which can complicate their implementation in practical scenarios.

The application of tourism data in Thailand to provide suitable management recommendations. It utilizes clustering techniques in conjunction with logistic regression to categorize provinces. The analysis, which utilizes 14 variables, found that the optimal number of groups is three: primary provinces (18 provinces), secondary provinces (29 provinces), and tertiary provinces (30 provinces) [12]. It also highlights that the key variables influencing these provincial groupings are the number of overnight stays by foreign tourists, the number of domestic tourists, and their average expenditure. This analysis aims to develop a system for classifying the severity of road accidents in Thailand. It uses several machine learning algorithms, including Logistic Regression (LR), Random Forest (RF), and K-Nearest Neighbor (KNN). Additionally, different feature selection techniques were applied to create a total of nine predictive models. The findings support targeted road safety policies and highlight the importance of integrating data-driven approaches to

reduce accident severity [13]. This study focuses on combining clustering with logistic regression and decision trees, analyzing beyond simple grouping to create a robust predictive model. First, clustering organizes a large dataset into meaningful groups based on similarities. Then, logistic regression can be applied to these clusters to predict the probability of a new data point belonging to a specific group. Simultaneously, a decision tree can be used to visually explain the key rules or criteria that define each cluster, providing clear and interpretable insights into what makes each group unique. By integrating these three techniques, the study gains the ability not only to group data but also to predict and explain future outcomes accurately.

2. Materials and Methods

2.1 Data

This article analyzes traffic accident data for the period 2008-2014, specifically during the New Year and Songkran festivals in Thailand. These data were obtained from the National Institute for Emergency Medicine in Thailand, a government agency responsible for overseeing and improving emergency medical services (EMS) nationwide [14]. It was established to enhance the quality, accessibility, and efficiency of emergency medical care, ensuring that people in need receive timely and adequate medical attention during emergencies. Descriptions of the dataset are as follows.

Table 1. Data description.

Column name	Description	Possible values
Date	Date of incident	13 Apr 2010, …
Time	Time of incident	20:01-21:00, …
Festival	The festival where the incident occurred	New year, Songkran
Gender	Victim's gender	Male, Female
Age	Victim's age	1, … , 99
RoadType	Road classification	Highway, Rural road, Urban road, Unknown
Status	Status of the victim	Driver, Passenger, Pedestrian, Unknown
Vehicle	Vehicle involved in the accident	Bicycle, Bus, Car, Car/Taxi, Motorcycle, Motorized tricycle, Pickup truck, Tricycle, Truck, Van, Other, Unknown
AnotherVehicle	The other driver's vehicle	Bicycle, Bus, Car, Car/Taxi, Motorcycle, Motorized tricycle, Pickup truck, Tricycle, Truck, Van, Other, Unknown
Measure	Preventive measures	Helmet, Seat belt, Unknown, No measure
Drinking	Drinking alcohol	Yes, No, Unknown
Transfer	The person who transported the patient	ALS, BLS, Died instantly, Foundation/Volunteer, FR, ILS, Not Transfer, Police officer, Victim/Relatives
ReferAdmit	Has he/she been admitted to the hospital?	Yes, No
Solution	Treatment outcome	Recovery, Died instantly, Died subsequently
Duration	Treatment duration	6 (Days), 15 (Days), …
Province	Province Name	Bangkok, Kalasin, …

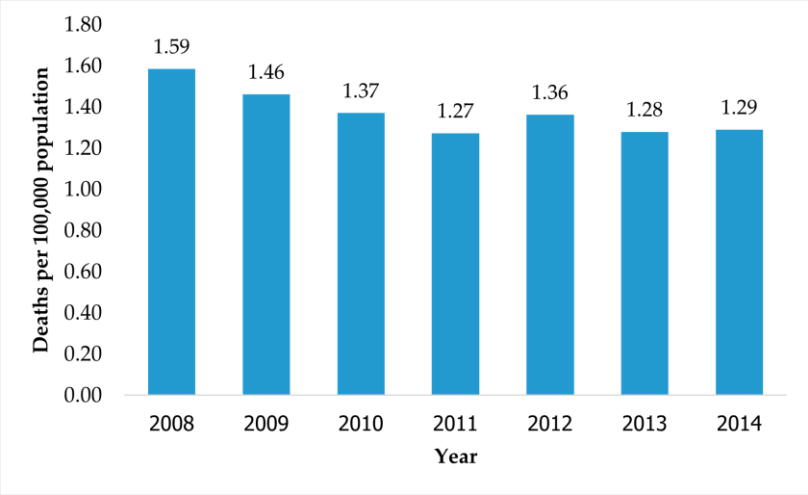

Figure 1. Deaths per 100,000 population by road accident during the New Year and Songkran festival.

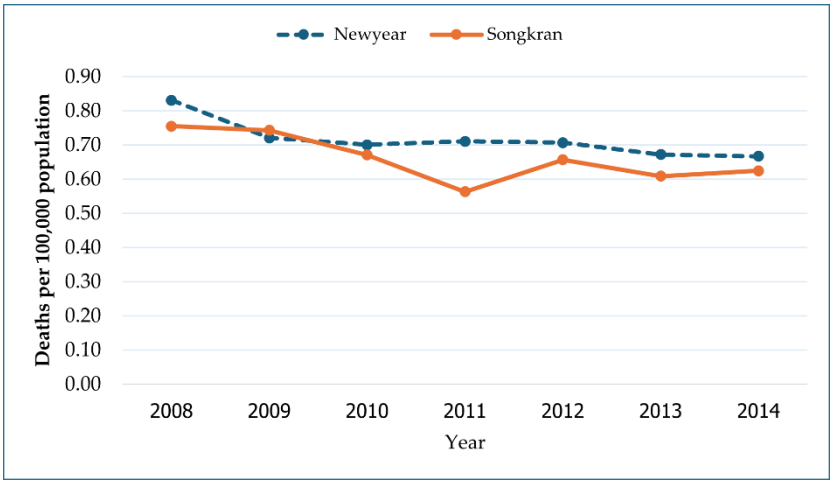

Figure 2. Comparison of deaths per 100,000 population during the New Year and Songkran festivals.

Figure 1 illustrates the annual deaths per 100,000 population from 2008 to 2014. Deaths peaked in 2008, then generally declined with some fluctuation through 2014, though rates remained relatively high throughout the period. Figure 2 compares deaths during the New Year and Songkran festivals from 2008 to 2014. The New Year consistently had higher fatalities than Songkran across all years. Both festivals showed declining death tolls over time, with a notable dip during the New Year 2011. Despite safety improvements, the New Year continues to pose a greater risk than Songkran

2.2 Methodology

Analyzing road accident data is crucial for understanding the factors contributing to these incidents and developing effective strategies for prevention. This process involves several key steps, each contributing to a comprehensive understanding of the data and its implications. Here is a detailed breakdown of the data analysis process for road accidents.

2.2.1. Importing Libraries

The first step involves importing the necessary libraries into the Python environment. Libraries such as Pandas are essential for data manipulation and cleaning, while NumPy is used for numerical operations. Matplotlib and Seaborn are essential for data visualization, enabling the creation of informative charts and graphs. Scikit-learn (sklearn) is indispensable for machine learning tasks, including feature selection and

model building. Other libraries, such as imblearn and statsmodels, may be necessary for upsampling and statistical analysis.

2.2.2 Importing Data

The subsequent step entails importing road accident data from government databases. This data is typically loaded into a Pandas DataFrame, offering a structured and efficient format for handling tabular data.

2.2.3 Removing Duplicated Rows

Duplicate entries can distort analytical outcomes and lead to erroneous conclusions. Therefore, it is essential to identify and eliminate duplicate rows from the dataset, ensuring that each record is unique and contributes meaningfully to the analysis.

2.2.4 Detecting and Removing Outliers

Outliers, defined as data points that deviate significantly from the norm, can have a substantial impact on statistical analyses and the performance of machine learning models. Techniques such as box plots and z-scores are employed to detect these anomalies, which are then either removed or appropriately transformed based on the study's context.

2.2.5 Data Preparation

Data preparation encompasses the cleaning and transformation of raw data to render it suitable for analysis. This process includes handling missing values, standardizing or normalizing numerical variables, and encoding categorical variables. The integrity of the subsequent analysis is heavily dependent on the quality of data preparation.

2.2.6 Assessing Multicollinearity via VIF

Multicollinearity, the phenomenon where independent variables exhibit high correlations, can result in unstable and unreliable regression coefficients. The Variance Inflation Factor (VIF) is computed to quantify multicollinearity, enabling the identification and remediation of these issues within the dataset.

2.2.7 Upsampling

Road accident datasets often suffer from class imbalance; for example, severe accidents may be underrepresented compared to minor accidents. Upsampling techniques, such as SMOTE, are implemented to balance the class distribution, thereby enhancing the performance and robustness of machine learning models.

2.2.8 Encoding Categorical Variables

Since machine learning models predominantly require numerical input, categorical variables must be transformed into numerical formats. One-hot encoding and label encoding are standard methods used for this purpose, where one-hot encoding creates binary variables for each category, and label encoding assigns a unique integer to each category.

2.2.9 Dataset Partitioning

For evaluating model performance, the dataset is partitioned into three subsets: training, validation, and testing sets. The training set is used to build the model, the validation set to fine-tune hyperparameters, and the testing set to evaluate the model's final performance. In this setup, the dataset is divided into training and testing subsets using an 80:20 ratio. Specifically, 80% of the data is allocated for model training, while the remaining 20% is reserved for testing.

2.2.10 Feature Selection Using Logistic Regression and Decision Trees

Feature selection involves identifying a subset of relevant variables that most effectively predict the target outcome. Techniques such as logistic regression and decision trees are employed to evaluate the importance of each feature, thereby enhancing model performance, mitigating overfitting, and improving interpretability.

2.2.11 Visualization of Feature Coefficients

The visual representation of the coefficients derived from the logistic regression model provides insight into the relationships between the predictors and the outcome variable. Such visualizations facilitate the identification of key factors contributing to road accidents.

2.2.12 Clustering

Clustering is applied to the accident data to identify distinct clusters based on shared characteristics. This clustering approach is particularly advantageous for targeting interventions and developing prevention strategies.

2.2.13 Conclusions

The final step involves synthesizing the findings from the data analysis. This includes identifying the principal factors that contribute to road accidents, evaluating the performance of the machine learning models employed, and formulating evidence-based recommendations for prevention strategies. Additionally, the analysis of clusters derived from clustering provides nuanced insights into the heterogeneous characteristics of different accident groups, thereby informing more targeted interventions and safety measures.

3. Results and Discussion

3.1 Results

This section presents the primary findings from logistic regression and clustering analyses. We first discuss the most influential factors contributing to road accidents, as identified through logistic regression coefficients. We then examine the clustering analysis results, which categorize the data into distinct groups based on road type, vehicle type, and transfer method.

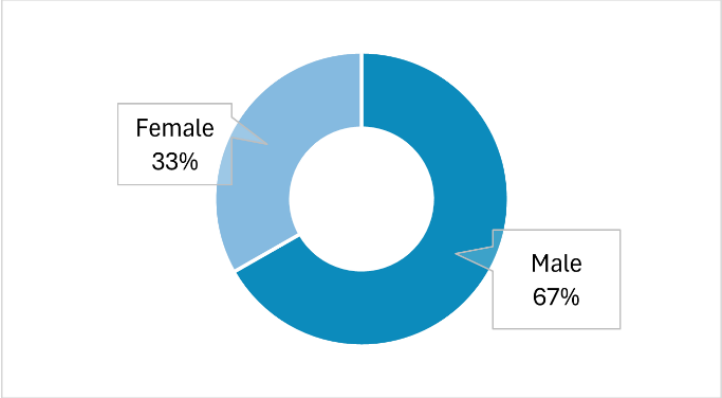

Figure 3. Accident Victim Statistics by Gender.

Figure 3 shows men account for 262,273 individuals (67%) compared to 130,858 women (33%), representing a 2:1 ratio. Similarly, Figure 4 illustrates a significantly larger gender disparity, with men comprising 4,876 individuals (79.36%) and women comprising 1,286 (20.64%).

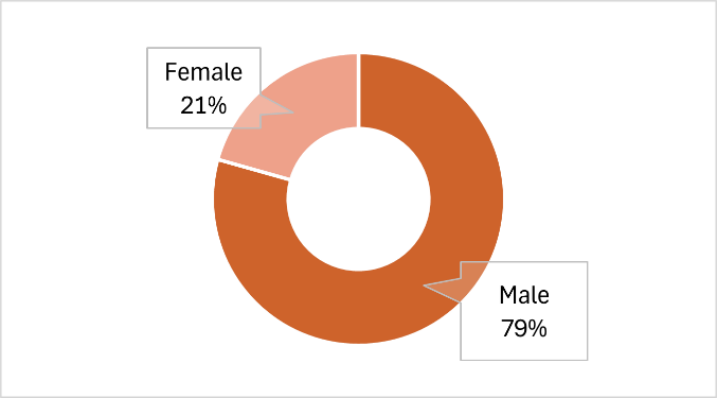

Figure 4. Mortality statistics by gender.

According to Figure 4, men comprise 4,876 individuals (79.36%), while women comprise 1,286 individuals (20.64%). This notable difference indicates a higher proportion of male cases compared to female cases, suggesting the potential influence of gender-related factors on the observed outcomes.

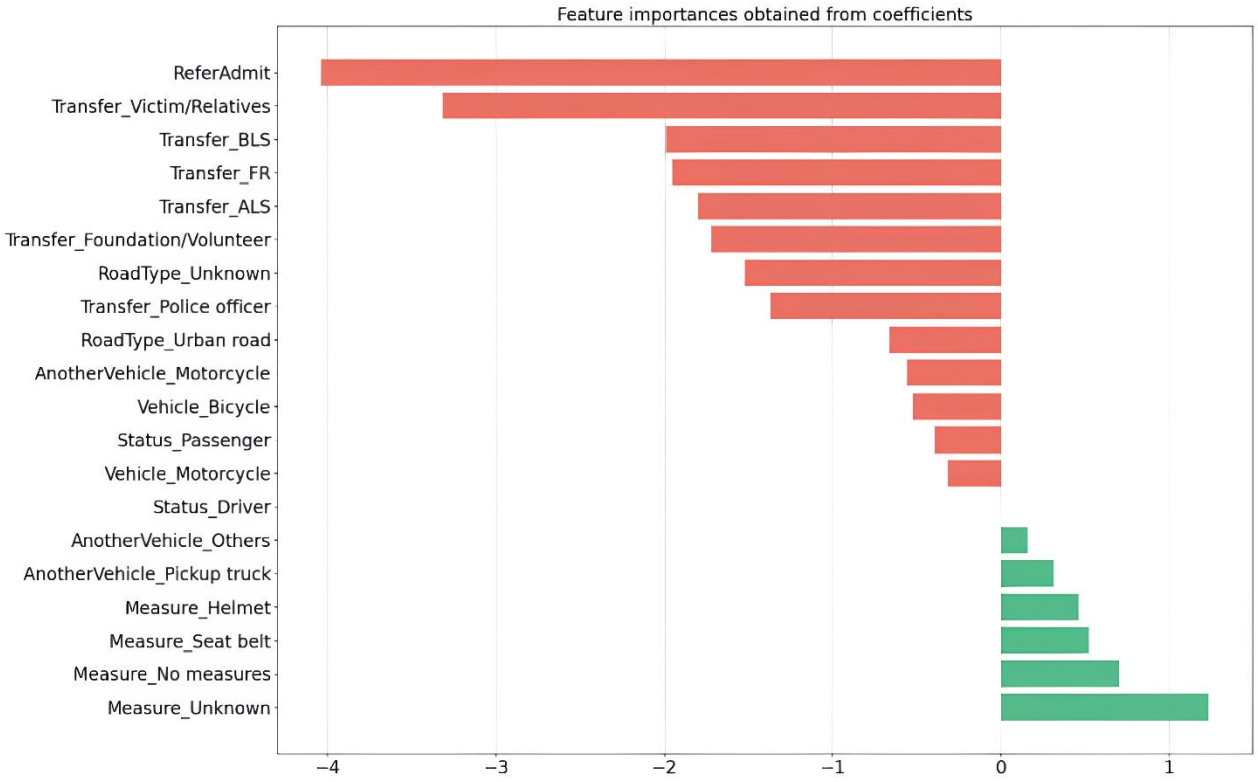

Figure 5. The features of importance derived from logistic regression coefficients.

Figure 5 displays the feature importances derived from the logistic regression coefficients. Red bars indicate features that reduce accident probability (ReferAdmit, Transfer_Victim/Relatives, Transfer_BLS), while green bars show features that increase it (Measure_Unknown, Measure_No measures, Measure_Seat belt). The analysis reveals that transfer methods and safety measures significantly impact outcomes, suggesting that improved transfer systems and safety protocols could enhance care efficiency."

The analysis utilized logistic regression to predict the likelihood of road accidents. Prior to modeling, the data underwent a thorough cleaning process to ensure accuracy and reliability. Multicollinearity was assessed using the Variance Inflation Factor (VIF) to identify and address any potential issues among predictor variables. The final model achieved a high accuracy rate of 98.67 %, demonstrating its strong predictive capability.

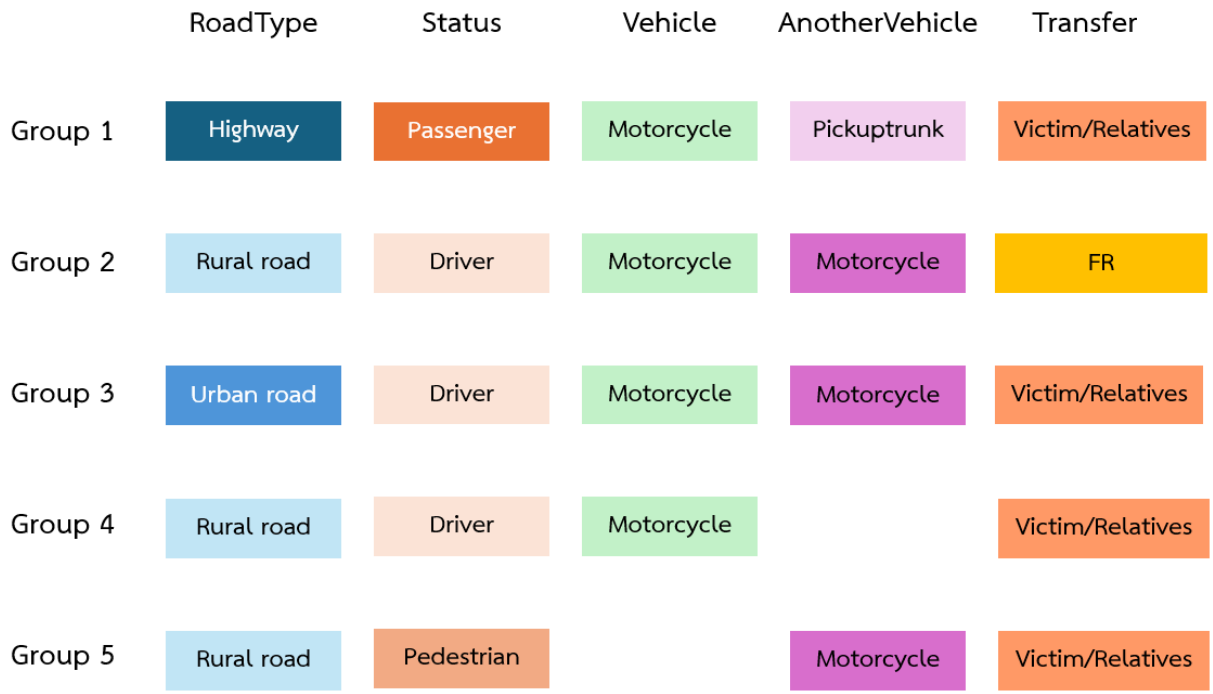

Figure 6. Scenarios concerning road incidents and traffic patterns.

This visual, Figure 6, represents the results of a clustering analysis, categorizing road accident data into five distinct groups based on key features: Road Type, Status, Vehicle, Another Vehicle, and Transfer. Each group is defined by specific combinations of these features, offering insights into common accident scenarios. Detailed Analysis is described as follows.

Group 1: Highway accidents involving motorcycle passengers colliding with pickup trucks, with victim/relative transfers

Group 2: Rural road accidents between motorcycle drivers and other motorcycles, using FR transfer method

Group 3: Urban motorcycle-to-motorcycle collisions involving drivers, with victim/relative transfers

Group 4: Single-vehicle motorcycle accidents on rural roads with victim/relative transfers

Group 5: Pedestrian-motorcycle accidents on rural roads with victim/relative transfers

3.2 Discussion

The clustering analysis reveals five distinct accident patterns characterized by road type, user role, vehicle involvement, and transfer methods. Motorcycles emerge as the dominant factor, appearing in every cluster either as the primary or secondary vehicle involved. The analysis identifies critical accident scenarios: motorcycle-pickup truck collisions on highways, motorcycle-to-motorcycle crashes in both rural and urban settings, single-vehicle motorcycle accidents, and motorcycle-pedestrian incidents in rural areas. These patterns highlight environment-specific risks and reveal important gaps in emergency response systems. Notably, rural accidents often rely on transfers of victims or their relatives rather than professional emergency services, which can potentially impact care outcomes. The findings underscore the need for targeted interventions: enhanced motorcycle safety protocols for highways, improved collision prevention strategies for urban and rural motorcycle interactions, and strengthened emergency transfer systems, particularly in rural areas where professional emergency response may be limited."

4. Conclusions

This study reveals significant patterns in road accident data during festive periods. A notable gender disparity emerged, with males being twice as likely to experience accidents and four times more likely to suffer fatal outcomes compared to females. The analysis demonstrates that patient transfer methods have a critical influence on injury severity, highlighting the urgent need to strengthen emergency response systems. The clustering analysis identified motorcycles as the dominant factor across all accident scenarios, spanning highways, urban roads, and rural settings, and involving drivers, passengers, and pedestrians. Importantly, the study confirms the protective value of safety measures, with the use of helmets and seatbelts showing positive impacts on outcomes, reinforcing the importance of sustained prevention campaigns. These findings underscore the need for targeted interventions, including gender-specific safety education, motorcycle-focused safety protocols tailored to different road environments, enhanced emergency transfer systems (particularly in rural areas), and continued promotion of protective equipment use. Such targeted policies and resource allocation could significantly improve road safety and reduce accident burden during high-risk festive periods. Future analyses should incorporate environmental and infrastructure variables to enhance predictive accuracy. Key additions include road surface conditions (dry/wet), road geometry (straight/curve/intersection), weather conditions (clear/fog/rain), and lighting conditions (daylight/nighttime with/without lighting). Integrating these factors would provide deeper insights into contextual influences on accident severity and improve the model's ability to capture complex environmental relationships.

5. Acknowledgements

The authors gratefully acknowledge the Open Government Data of Thailand for providing access to valuable datasets that facilitated this research.

Funding: This research was funded by the College of Industrial Technology, King Mongkut's University of Technology North Bangkok (Grant No. Res-CIT0629/2023).

Author Contributions: All the authors have contributed by reading and revising the current study, including numerical simulation, writing, and review

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Simmachan, T.; Wongsai, N.; Wongsai, S.; Lerdsuwansri, R. Modeling road accident fatalities with underdispersion and zero-inflated counts. *PLoS ONE* **2022**, *17*(11), e0269022. https://doi.org/10.1371/journal.pone.0269022

[2] Palwisut, P. A Risk Prediction Model of Road Accidents During Long Holiday in Thailand Using Ensemble Learning with Decision Tree Approach. *Suan Sunandha Sci. Technol. J.* **2023**, *10*(2), 213–221. https://doi.org/10.53848/ssstj.v10i2.499

[3] Riyapan, S.; Thitichai, P.; Chaisirin, W.; Nakornchai, T.; Chakorn, T. Outcomes of Emergency Medical Service Usage in Severe Road Traffic Injury during Thai Holidays. *West J. Emerg. Med.* **2018**, *2*, 266–275. https://doi.org/10.5811/westjem.2017.11.35169

[4] Boonserm, E.; Wiwatwattana, N. Using Machine Learning to Predict Injury Severity of Road Traffic Accidents During New Year Festivals From Thailand's Open Government Data. In *Proceedings of the 2021 9th International Electrical Engineering Congress (iEECON)*, Pattaya, Thailand, **2021**; pp 464–467. https://doi.org/10.1109/iEECON51072.2021.9440287

[5] Dong, Q.; Chen, X.; Huang, B. *Logistic regression*; Elsevier BV, **2024**; pp 141–152.

[6] Dey, D.; Haque, M. S.; Islam, M. M. *et al.* The proper application of logistic regression model in complex survey data: a systematic review. *BMC Med. Res. Methodol.* **2025**, *25*, 15. https://doi.org/10.1186/s12874-024-02454-5

[7] Saranceva, S. Applying the Decision Tree Method in the Field of Management Activities. *Ergodizajn* **2024**, 241–246. https://doi.org/10.30987/2658-4026-2024-2-241-246

[8] Fleissner, M.; Zarvandi, M.; Ghoshdastidar, D. Decision Trees for Interpretable clusters in mixture modelsand deep representations. **2024**.

[9] Sonia, Y. Study of Existing Methods & Techniques Of K-Means Clustering. **2024**. https://doi.org/10.53555/kuey.v30i4.1755

[10] Velunachiyar, S.; Sivakumar, K. Some Clustering Methods, Algorithms and their Applications. *Int. J. Recent Innovation Trends Comput. Commun.* **2023**. https://doi.org/10.17762/ijritcc.v11i6s.6946

[11] Basiri, F.; Amer, A.; Ranjbar Naserabadi, M. J.; Moghimi, M. M. A Comparative Study of K-means and NMF Clustering Algorithms. In *2023 2nd International Engineering Conference on Electrical, Energy, and Artificial Intelligence (EICEEAI)*, Zarqa, Jordan, **2023**; pp 1–4. https://doi.org/10.1109/EICEEAI60672.2023.10590510

[12] Boonkrong, P.; Simmachan, T.; Sittimongkol, R.; Lerdsuwansri, R. Data-Driven Approach in Provincial Clustering for Sustainable Tourism Management in Thailand. *Thail. Stat.* **2025**, *23*, 481–500.

[13] Simmachan, T.; Wongsai, S.; Lerdsuwansri, R.; Boonkrong, P. Impact of COVID-19 Pandemic on Road Traffic Accident Severity in Thailand: An Application of K-Nearest Neighbor Algorithm with Feature Selection Techniques. *Thail. Stat.* **2024**, *23*, 129–143. https://doi.org/10.1371/journal.pone.0309234