Analysis of Road Accident Patterns and Risk Factors Using Data Science and Machine Learning Techniques

Field: Technology and Computer

Thanakrit Ponimthai
Darunsikkhalai Science School
Samut Songkhram, Thailand

Werawish Chotijurangkul
Darunsikkhalai Science School
Bangkok, Thailand

Polrapat Roemraksachaikul
Bangpakokwittayakom School
Bangkok, Thailand

Shayne Chawengchot
Roong Aroon School
Bangkok, Thailand

Sukree Sinthupinyo
Faculty of Engineering,
Chulalongkorn University
Bangkok, Thailand

Sivakorn Malakul
The Institute for the Promotion of
Teaching Science and Technology
Bangkok, Thailand

Abstract: This study investigates the factors contributing to road accidents in Thailand, with a focus on weather conditions, vehicle types, date, location, and time. The primary objective is to develop predictive models that identify key factors influencing accidents and provide actionable insights to improve road safety. The research methodology involved data collection, preprocessing, and exploratory data analysis, including the use of heatmaps to visualize correlations. Subsequently, various predictive models were trained and evaluated for their performance and accuracy. The findings highlight critical factors that significantly impact accident occurrences, offering valuable implications for policy-makers and stakeholders in designing targeted interventions to enhance road safety.

Keywords: Road Accident, Data Science, K-mean Clustering, Data Visualization

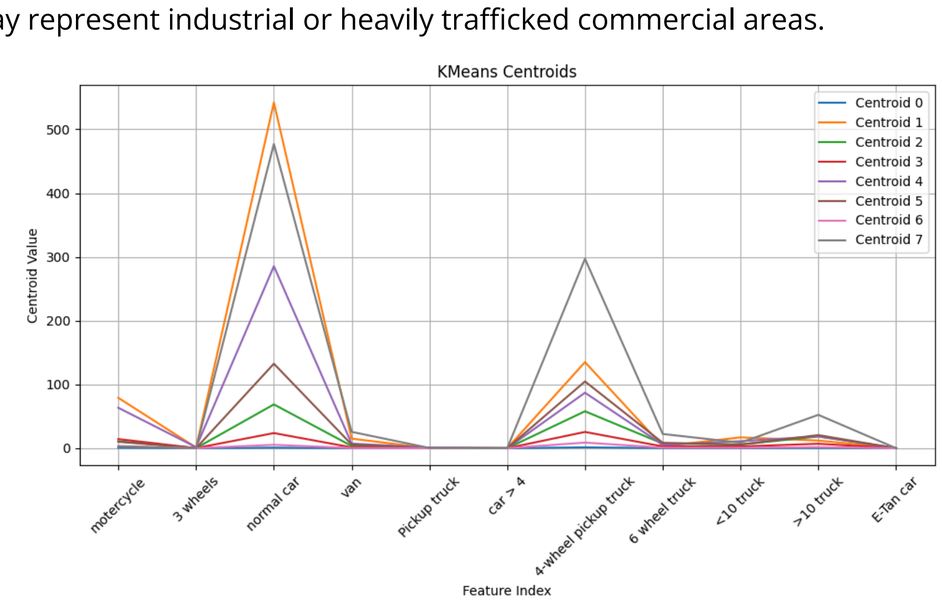
Introduction:

Road accidents in Thailand pose significant challenges to public safety and economic stability. This research uses k-means clustering to analyze factors like date, location, time and vehicle characteristics, aiming to enhance road safety and support evidence-based policies. Insights will guide targeted interventions and public awareness campaigns to reduce accidents and their impact.

Results, Discussion and Conclusion:

The clustering results reveal distinct patterns in vehicle distributions across the dataset. Cluster 0 is the largest group, encompassing 13,118 grids with notably sparse vehicle density, indicating areas with minimal traffic activity. Cluster 1, in contrast, is the smallest cluster, containing only 1 grid that is predominantly characterized by private cars, though it has fewer 4-wheel pickup trucks compared to other clusters. Cluster 2 includes 63 grids, which show a low presence of both private cars and pickup trucks, suggesting regions with light vehicle activity. Cluster 3 comprises 217 grids and exhibits no significant dominance of any specific vehicle type, reflecting a more uniform vehicle distribution.

Meanwhile, Cluster 4 represents 5 grids where private cars are highly prevalent, accompanied by a moderate presence of motorcycles and pickup trucks. This cluster may represent areas where personal vehicles are the primary mode of transportation. Similarly, Cluster 5, which includes 30 grids, shows a balanced distribution of private cars and 4-wheel pickup trucks, possibly indicating suburban regions with mixed vehicle usage. Cluster 6, covering 1,996 grids, has sparse vehicle activity similar to Cluster 0, which might correspond to rural or less populated areas. Finally, Cluster 7, with only 3 grids, stands out due to its high density of private cars and a moderate presence of 4-wheel pickup trucks, as well as larger vehicles like 6-wheel and >10-wheel trucks. This combination suggests that Cluster 7 may represent industrial or heavily trafficked commercial areas.



The graph displays the centroids of each cluster from K-Means analysis. The X-axis represents vehicle types (e.g., motorcycle, 3 wheels, normal car, van, etc.), while the Y-axis shows the centroid values for each feature.

Acknowledgements:

This research has received funding support from the NSRF via the Program Management Unit for Human Resources & Institutional Development, Research and Innovation [grant number B46G670083]

Methodology:

Data Collection

This study investigates road accident data from Thailand collected between 2019 and 2023 [1]. The dataset, obtained from the government website, includes variables such as year, date, time of occurrence, weather conditions, number and types of vehicles involved, minor injuries, severe injuries, and fatalities.

Data Preprocessing

To ensure data quality and reliability, researchers implemented a systematic cleaning process. Records containing null values, unknown entries, or inconsistencies across any variables were excluded. This process was crucial for minimizing potential biases and ensuring the dataset's integrity, forming a robust foundation for analysis. By removing incomplete or unreliable entries, the study aimed to produce credible and accurate insights.

Data Analysis and Visualization

Researchers used a heatmap to examine correlations among variables and determine feature importance. A geographic grid was developed to map accident severity levels across regions, aiding in the analysis of spatial patterns. Severity levels were calculated using a Weighted Severity Index (WSI)-based formula:

$$Severity\ Level = 5D + 2S + M + V,$$

where D represents deaths, S represents severe injuries, M represents minor injuries, and V represents vehicles involved. This formula emphasizes severe outcomes, aligning with established WSI methodologies. [2,3] The analysis provided insights into regional disparities in accident severity, helping prioritize critical areas for intervention.

Data Clustering

The Elbow Method was applied to determine the optimal number of clusters (K) for k-means clustering. This approach involved calculating the within-cluster sum of squares (WCSS) for varying K values and identifying the "elbow" point where WCSS reductions became marginal. Once the optimal K was determined, k-means clustering segmented the dataset into meaningful clusters, grouping data points with similar characteristics. This process revealed underlying patterns and relationships in the data, offering deeper insights into accident trends and characteristics.

Insight Evaluation

Key regions with high accident severity or significant deficincies were identified using heatmaps and k-means clustering. These findings highlight areas with pronounced safety challenges, offering opportunities for targeted interventions. The insights can guide evidence-based policies and strategic measures aimed at improving road safety and addressing regional disparities for long-term development and enhancement.

References

[1] National Statistical Office, "Road Accident Dataset," *Data.go.th.* [Online]. Available: httpsxs//data.go.th/dataset/gdpublish-roadaccident. [Accessed: November 2024].

[2] D. M. Reddy and K. N. Chaitanya, "Road Accident Black Spot Analysis Using Weighted Severity Index Method at L B Nagar Zone Hyderabad," *Civil Engineering and Architecture*, vol. 11, no. 1, pp. 237–247, 2023, doi: 10.13189/cea.2023.110120.

[3] N. N. Kumar, T. Ilango, K. Anvesh, M. Sowbhagya, S. Shashidhar, and T. S. R. Chand, "Identification of Black Spots & Accident Analysis on SH–4," *Journal of Emerging Technologies and Innovative Research,* vol. 6, no. 4, pp. 479–485, Apr. 2019.