COMP20008 Assignment 2 Group Report

Group W17G3

Authors
Tan Lok Heng
Minh Triet Pham
Quynh Chi Dang
Francisco Palumbo

Executive Summary

This report investigates how external factors, specifically lighting, road geometry, and surface type, play a role in the severity of road accidents across Victoria, Australia. Using detailed crash data from Victoria Police, the research aims to explore whether these external factors, on their own or in combination, can help explain why some accidents are more serious or even fatal.

The analysis revealed that the greatest external contributor to severe car accidents was poor lighting. Similarly, roads that aren't at intersections, like dead-ends or open stretches, tend to see worse outcomes, likely due to higher speeds or less predictable traffic flow. Unsealed surfaces, like gravel or dirt, also showed slightly higher accident severity, though this varied depending on the road environment.

Machine learning models, including k-Nearest Neighbours and decision trees, were used to see if accident severity could be predicted based solely on these three factors. While the models performed above chance, their accuracy hovered around 63–64%, and they often misclassified serious crashes as minor. This suggests that although infrastructure plays a role, it is not the sole predictor; driver behaviour, vehicle condition, and traffic dynamics are also likely to be major factors.

Ultimately, the findings give clear direction for where to focus road safety improvements. Better lighting, safer road designs, and targeted upgrades in high-risk areas could help reduce the number of serious crashes. While no single factor explains every accident, understanding how these environmental risks interact is a step toward building safer roads for everyone.

Introduction

Context

Road traffic accidents are a significant and inevitable worldwide issue that harms public safety. In Australia alone, thousands of people are severely injured or killed every year, leading to immense socioeconomic damage. According to the Australian Government, there have been 1284 fatalities on Australian roads in the past 12 months (as of March 2025), representing a 1.2% increase from the previous year - equating to 4.7 deaths per 100,000 people. Furthermore, Victoria ranks second in road deaths in 2025 (Australian Government, 2025). Non-fatal but severe accidents are immense, often leading to long-term/permanent health consequences and heavy financial burdens. This never-ending issue presents a need - to deeply understand the external factors under which road accidents become more serious/fatal, ultimately allowing us to identify more effective improvements to Victorian roads.

Research Question & Purpose

Research Question: “How do external factors influence the severity of an accident?”

The “external factors” this report will be investigating refer to light condition, road geometry, and road surface condition. These external factors are distinct from behavioural and demographic variables, and are mainly objective conditions about the environment and features of Victorian roads. The main objective of this research is to investigate the influence of these factors in isolation and in combination with each other on accident severity. Moreover, it will allow us to identify the root problems - which combinations of factors often lead to serious/fatal accidents. By conducting this research, we can identify and highlight specific improvements to Victorian infrastructure or policies, such as improving lighting conditions in areas with high-risk external conditions. Ultimately, minimizing accident severity and decreasing the mortality rate of traffic accidents.

Data Sources

This research uses real-world data from the Victoria Police - allowing for reliable and comprehensive information about car accidents in Victoria (State of Victoria, 2025). The three datasets used are: accident.csv - environmental conditions and includes data on light condition, road geometry, road surface condition, and accident severity (outcome measure), vehicle.csv - context of accidents and attributes of vehicles involved, person.csv - data on individuals involved in accidents.

These datasets are significant because they consist of non-behavioural, non-vehicular external conditions, as well as our main outcome measure of interest - accident severity.

Methodological Framework

To investigate this research question, our research follows these 3 analysis layers:

1) Single-factor analysis

Investigating the effect of each individual external factor on accident severity. This involves identifying which conditions for each external factor (such as dark without lighting, or dead-end road geometries) are associated with serious/fatal accidents the most.

2) Combined-factor analysis

Researching interactions between pairwise and full combinations (2-3 external factors), such as light + geometry, or light + surface, and all three factors combined. Conducting this analysis allows us to understand how compound conditions affect accident severity, especially when there are multiple high-risk conditions.

3) Mutual information, clustering, and supervised model analysis

Mutual information correlation analysis: To investigate the quantitative strength of the relationship between external factors on accident severity.

Clustering analysis: To detect patterns of risk in the data space.

Supervised learning models (KNN and decision tree models): To predict accident severity based on the 3 external factors, and investigate its accuracy to strengthen our findings on how these factors affect severity. By conducting this analysis, we can also confirm whether these 3 factors alone can or cannot determine accident severity.

Methodology

Data Preprocessing 1

Categorical Feature Engineering

- **Light Condition**: consolidated the original seven-level light condition variable into five analytically meaningful categories:
    - Daylight (original: Day)
    - Limited light (original: Dusk/Dawn)
    - Dark with lighting (original: Dark street light on)
    - Dark without Lighting (originals: Dark street lights off, Dark no street lights)
    - Unknown (originals: Dark street lights unknown, Unknown)
- **Road Geometry**: Preserved all original categories as each represents a distinct roadway configuration.
- **Road Surface**: Dropped the single missing record and renamed "Not known" to "Unknown" for consistency.

This recategorization followed three guiding principles:

- **Functional Equivalence:** We merged categories with identical visibility conditions (e.g., "Dark street lights off" and "Dark no street lights") into a single "Dark without Lighting" category
- **Descriptive Clarity:** We refined original labels to more intuitively convey illumination conditions.
- **Data Integrity:** Rather than risk erroneous imputation, we consolidated the two infrequent categories ("Dark street lights unknown" and "Unknown"), which together comprised only ~0.01% of records, into a single "Unknown" category.

Correlation and Causal Analysis 2

Single Factor Analysis 2.1

We calculate the percentage of serious/fatal accidents for each external factor, as well as the mean and variance of accident severity for each category. This will be represented through bar charts. By conducting this analysis, we can isolate the risk that will lead to a severe crash.

Combination Analysis 2.2

We investigate the interaction between external factors and how they influence accident severity. Using the same approach as the single factor analysis, we present bar charts representing the top 10 combinations (of conditions across the 3 factors) with the highest percentage of serious/fatal accidents.

Mutual Information Correlation 2.3

We investigate the normalized mutual information correlation between each factor/combination and accident severity, allowing us to understand the associations without any ordinal relationships between categories.

Supervised Learning Models 3

We use four classification models to investigate whether light condition, road geometry, and road surface can independently and accurately predict the accident severity. These include two KNN classifiers and two Decision Tree classifiers. The predictive power of these models will then be evaluated.

k-Nearest Neighbors Models 3.1

We implemented two versions of KNN - an unweighted KNN model using the three external factors as equal predictive features, and a weighted KNN where features with higher variance in accident severity have a greater influence. Both of these models will produce continuous risk scores based on the percentage of serious/fatal accidents (derived from Single Factor Analysis 2.1). Accident severity is classified into "SIGNIFICANT" (serious/fatal) and "NORMAL" (minor/no-damage).

Decision Tree Models 3.2

We implemented two decision tree models - a standard model using the original categories for all external factors, and an optimized model with category ranges (similar risk profiles are grouped based on the percentage of serious / fatal accidents derived from Single Factor Analysis 2.1). The use of decision trees here is useful due to their ability to make rules based on how the external factors may influence accident severity in the real world.

Model Evaluation

All models were evaluated using stratified five-fold cross-validation (80%/20% train-test split). This allows us to assess performance while minimizing class imbalance effects. Moreover, we assessed predictive capacity and validated patterns in severity risk by comparing accuracy and confusion matrix components.

Data Clustering and Risk Profiling 4

We implemented k-means clustering to identify natural groupings in the 3D risk space. We transformed categorical variables into continuous numerical values using the same method as the KNN approach. We also used the elbow method to determine the most optimal k value.

For each cluster, we analysed:

- Combinations of infrastructure conditions by the percentage of serious/fatal accidents
- Commonly appearing categories within high-risk combinations
- Useful statistics (mean risk scores, percentage of serious/fatal accidents, variance)
- Stability of clusters based on the variance in accident severity

By conducting this, we can determine whether the natural groupings support our findings from the supervised learning models, and whether external factors alone can accurately and consistently predict accident severity.

Data Exploration & Analysis and Discussion & Interpretation

Correlation and Causal Analysis 2

Single Factor Analysis 2.1

This section examines how individual external factors - light condition, road geometry, and road surface - influence accident severity.

Light Condition vs Severity 2.1.1

The analysis of light conditions reveals that there is a significant impact on accident severity outcomes. Both the mean severity values and the percentage of serious/fatal accidents provides critical insights:

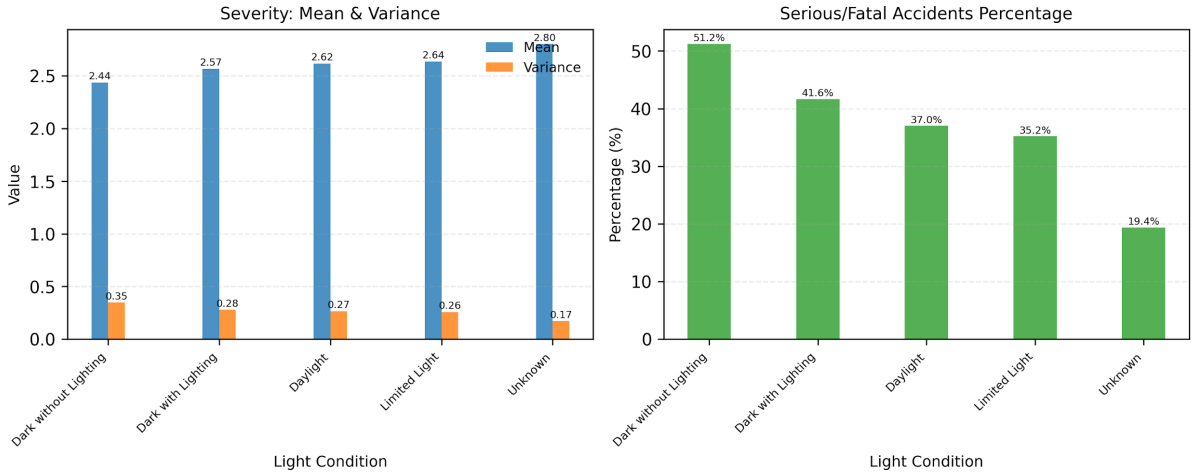

Figure 1: Light Conditions vs. Accident Severity Metrics

As shown in the left chart, “Dark without Lighting” has the lowest mean severity value (2.44 on a scale where lower values indicate higher severity). This shows that accidents occurring in dark environments without lights often lead to more severe outcomes.

This shows that accident severity in dark unlit conditions is more variable or less predictable, while daylight accidents show more consistent severity patterns.

Furthermore, the analysis of serious and fatal accident percentages reinforces these findings. The data shows that 51.2% of accidents occurring in “Dark without Lighting” conditions resulted in serious injury or fatality which is significantly higher than daylight conditions (37.0%). This clearly shows that poor lighting conditions increase the risk of severe accident outcomes.

Based on these findings, improving lighting conditions in poorly lit areas should be prioritized as an effective intervention to reduce accident severity.

Road Geometry vs Severity 2.1.2

Contrary to common expectations, our analysis reveals surprising patterns regarding road geometry and accident severity:

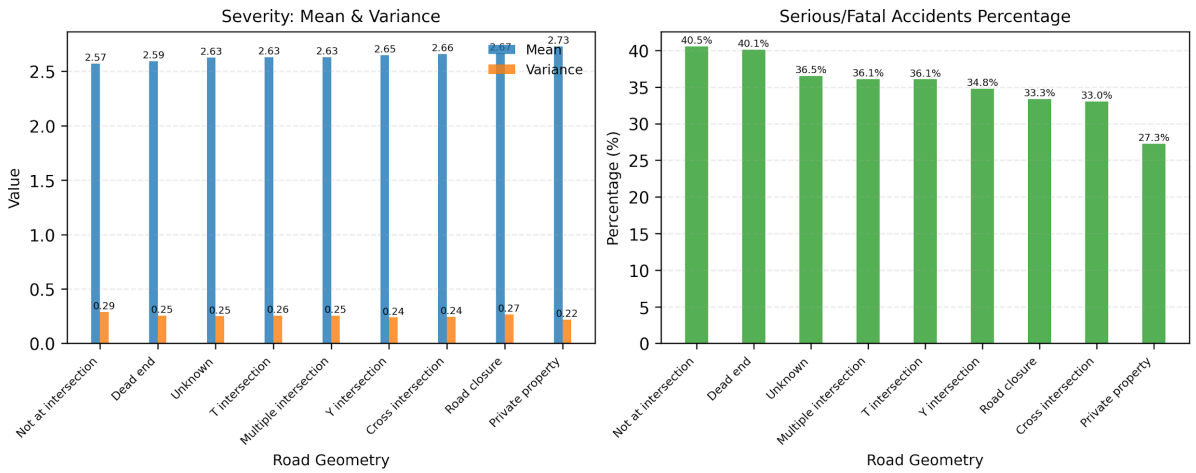

Figure 2: Road Geometry vs. Accident Severity Metrics

The severity mean analysis shows that "Not at intersection" (2.57) and "Dead end" locations (2.59) have the lowest mean severity values indicating more severe outcomes than various intersection types such as "Cross intersection" (2.66). The variance is also higher for non-intersection locations, suggesting more variable outcomes.

The percentage of serious/fatal accidents further confirms this unexpected finding. It is normally assumed that intersections present the greatest risk for severe accidents. However, from the chart above, "Not at intersection" and "Dead end" locations show the highest percentage of serious or fatal accidents (40.5% and 40.1% respectively). This indicates that not only does intersection location not affect accident outcome that much, but it also shows that accident outcome tends to be worse at non-intersection or dead end location.

These findings suggest that road safety initiatives should expand focus beyond intersections to include non-intersection road segments and dead-end locations, where accidents, though potentially less frequent, tend to result in more serious outcomes.

Road Surface vs Severity 2.1.3

Analysis of how road surface conditions affect accident severity shows less dramatic variations compared to other factors:

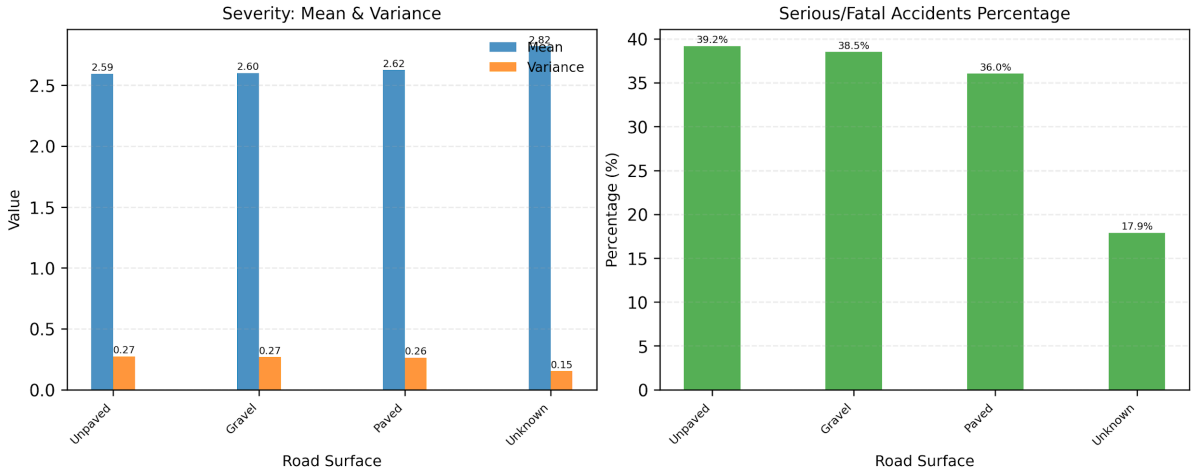

Figure 3: Road Surface vs. Accident Severity Metrics

The mean severity values show minimal differences across surface types, ranging from 2.59 for unpaved roads to 2.62 for paved surfaces - a difference of just 0.03 points. Variance values remain relatively consistent across surface types.

Looking at both severity metrics together reinforces this pattern. Unpaved and gravel surfaces show slightly higher percentages of serious or fatal accidents (39.2% and 38.5% respectively) compared to paved surfaces (36.0%), but the differences are not as pronounced as with other factors.

These findings indicate that while the road surface has some influence on accident severity, it appears less impactful than lighting conditions or road geometry. Thus, when prioritizing infrastructure improvements to reduce accident severity, road surface improvements should be considered after addressing more influential factors such as lack of lighting at key locations.

Combination Analysis 2.2

Investigating the effect of different combinations of light condition, road geometry, and road surface on the severity of accidents is significant. By conducting this analysis, we can identify what infrastructure to prioritize in order to minimize accident severity, ultimately leading to a lower mortality rate when travelling on Victorian roads.

Light Condition and Road Geometry 2.2.1

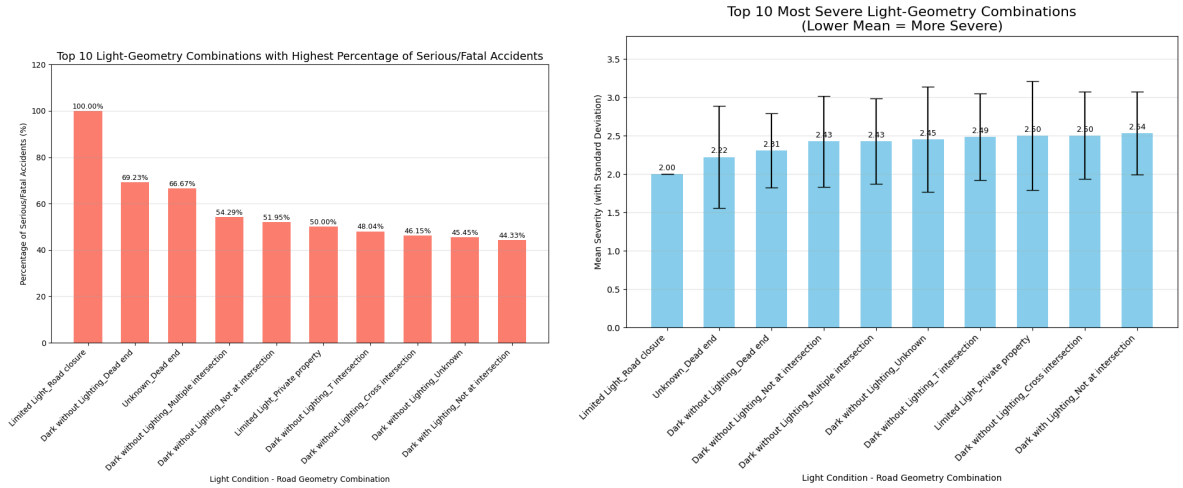

Figure 4: Top 10 Light Condition + Road Geometry Accidents

In both of these bar charts, the most severe accidents are dominated by “dark without lighting” combinations, with emphasis on combinations involving dead ends or road closures under dark conditions, and intersections with limited lighting. This suggests that poor lighting greatly amplifies the risks posed by specific road geometries. The analysis between light and road geometry furthers our core finding that lighting is the most critical external factor in determining accident severity, especially when compounded with high-risk road geometries.

Light Condition and Road Surface 2.2.2

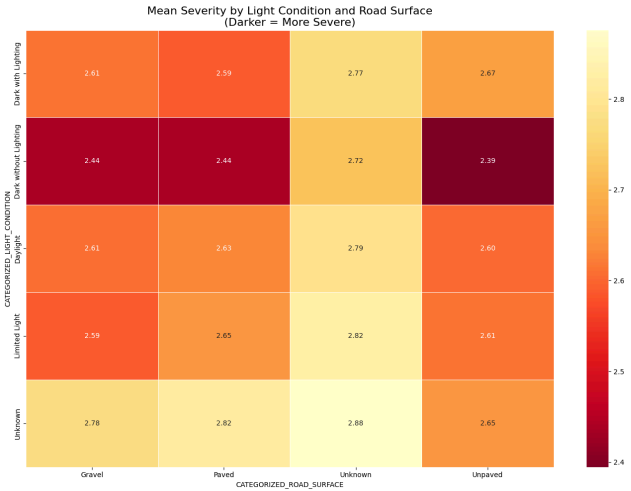

Figure 5: Light Condition + Road Surface Heatmap

The heatmap shows that combinations consisting of "dark without lighting" tend to lead to the worst accident outcomes. This confirms the dominance and significance of light conditions in connection with accident severity.

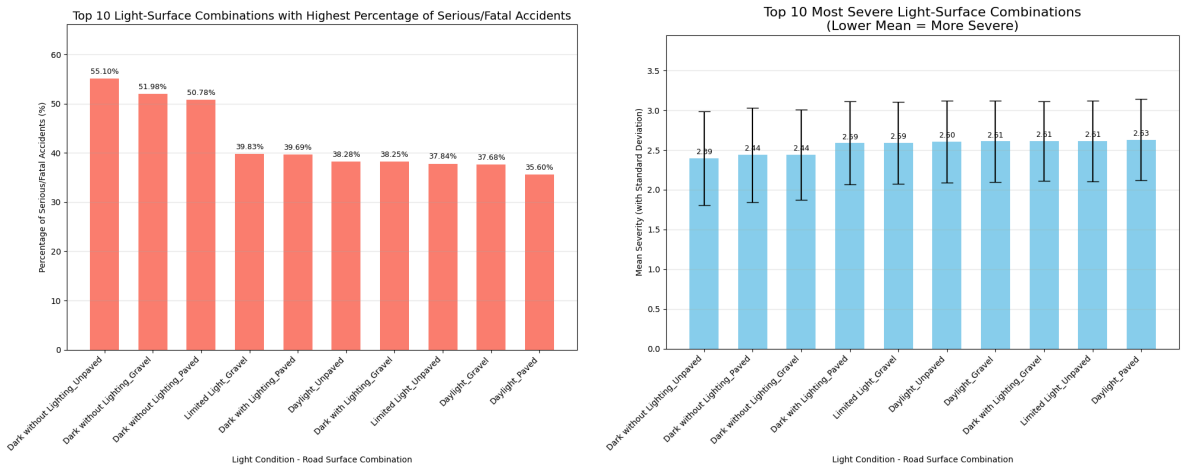

Figure 6: Top 10 Light Condition + Road Surface Accidents

In the 2 bar charts, the consistent appearance of "Dark without Lighting" in the top 10 most severe combinations show that lighting condition is once again the most influential external factor. On the other hand, road surface (gravel, paved, unpaved) show minimal variance in accident severity, further suggesting that under poor lighting conditions, road surface has a minimal impact on risk. This evidence supports our finding that improvement on light condition should be the main priority, with improving road surface being secondary.

Road Geometry and Road Surface 2.2.3

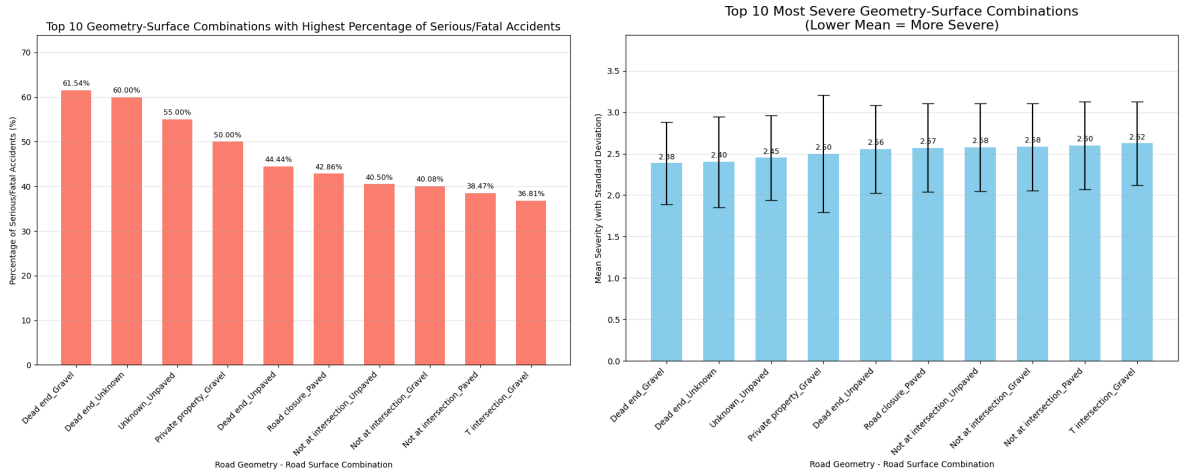

Figure 7: Top 10 Road Geometry + Road Surface Accidents

The two bar charts show that when road geometry is kept constant, there is minimal variation in accident severity between road surface types. However, regardless of road surface type, combinations consisting of dead ends consistently lead to higher accident severity. This is evident that out of these two external factors, road geometry is a more useful predictor of accident severity than road surface, especially when light condition is not taken into consideration.

Light Condition, Road Geometry and Road Surface 2.2.4

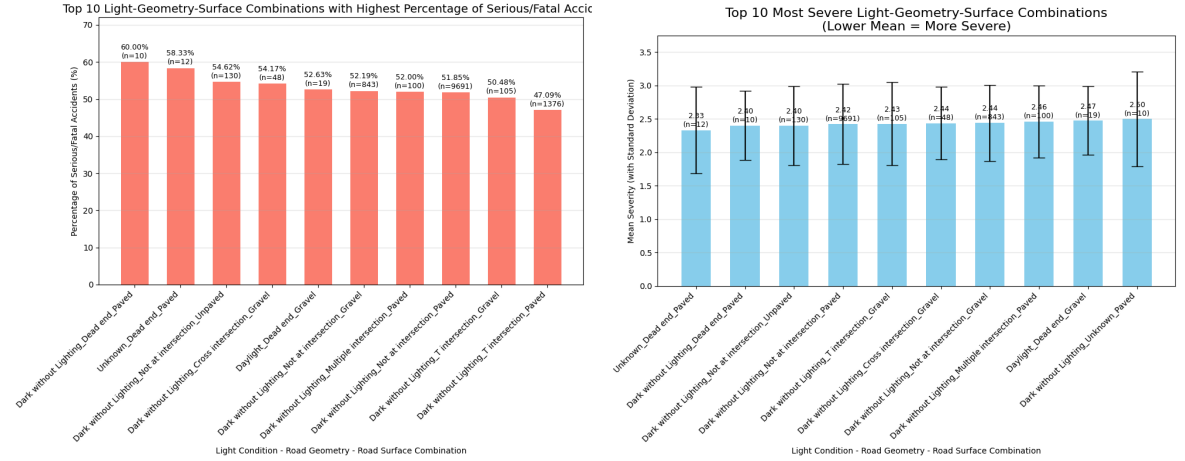

Figure 8: Top 10 Light Condition + Road Geometry + Road Surface Accidents

The involvement of “Dark without Lighting” is prevalent within the top 10 combinations with the highest percentage of serious/fatal accidents, with it being part of 80% of all the combinations here. This shows the significance of light condition on the accident severity, strongly justifying the improvement of light condition on Victorian roads. On the other hand, the involvement of road geometry here is insignificant because they are almost evenly distributed between the top 10 combinations - dead end (30%), not at intersection (40%), and intersection (40%). This carries over with road surface - despite most of the severe combinations being some combination of paved or gravel surfaces, road surface interestingly has negligible influence over accident severity. All in all, the bar chart strongly supports the finding that light condition is the most dominant factor in influencing accident severity.

From this tri-factor analysis, we can confidently suggest that improving light conditions on Victorian roads is critical in minimizing accident severity, with the first priority being roads with dead ends or intersections, and second priority being everywhere else with poor light condition, regardless of the road geometry and road surface.

Mutual Information Correlation Analysis 2.3

A mutual information correlation analysis allows us to quantitatively evaluate how much each external factor or a combination of factors contributes to accident severity.

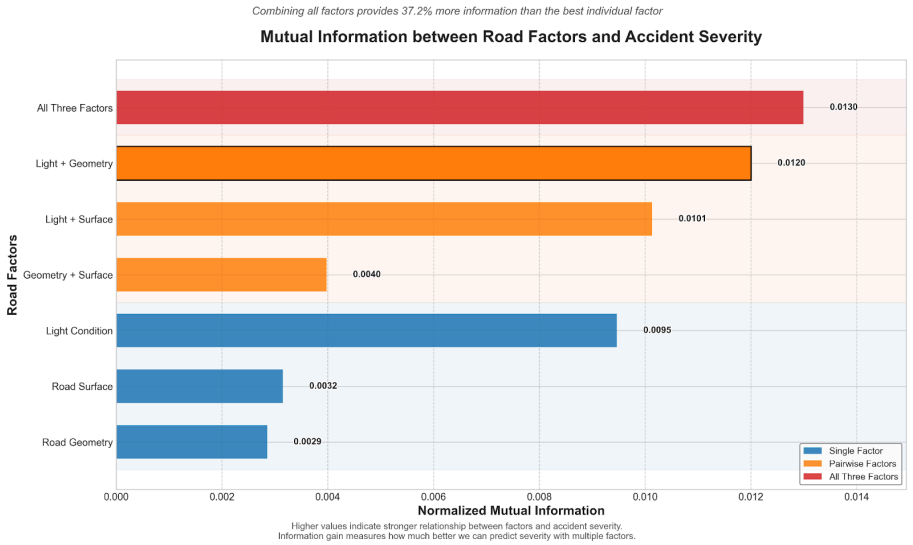

Figure 9: Mutual Information of Factors on Accident Severity

The analysis of single-variable mutual information shows that light condition has the highest individual mutual information with accident severity (0.0095) compared to road surface (0.0032) and road geometry (0.0029). Moreover, the combination of light condition and road geometry has the highest correlation (0.0120) out of all the two-factor combinations. Furthermore, road surface, when analysed both individually and in combinations, contributes the least to accident severity.

These findings suggest that not only is light condition the most dominant predictor for accident severity, it also represents the most influential compound combinations. This supports the prioritization of improvements on light conditions on Victorian roads, especially at areas with high-risk road geometries such as at dead ends or intersections. The mutual information correlation analysis concludes that for the most efficient use of resources, the risks of light condition and road geometry should be considered first, only then moving on to improving the road surface.

Supervised Learning Models and Evaluation Component 3

k-Nearest Neighbours Models 3.1

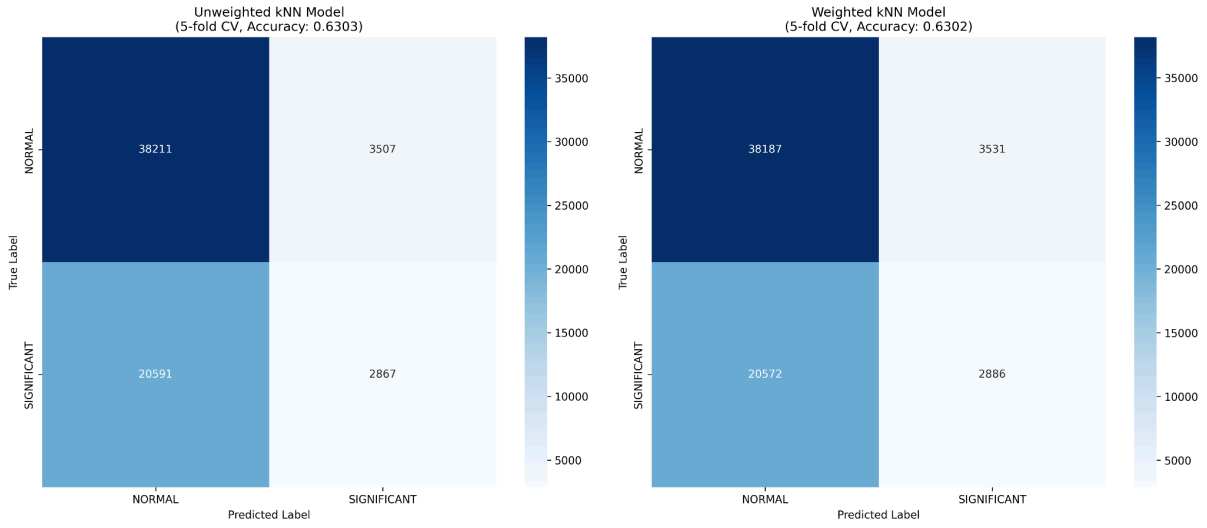

Figure 10: Unweighted vs. Weighted kNN Model Performance

Unweighted kNN Model 3.1.1

The unweighted kNN model, using equal weights for the three infrastructure risk factors, achieved an average accuracy of 0.63 across five-fold cross-validation. While this might initially appear as merely fair performance for a standard classification task (Santoso et al., 2021), contextual evaluation suggests otherwise. Research in accident severity prediction using only infrastructure factors typically reports accuracy ranges of 55-65% without incorporating behavioral or vehicle-specific features (Santos et al., 2022). The confusion matrix revealed a high rate of misclassifying "SIGNIFICANT" cases as "NORMAL," suggesting that either the equal weighting is ineffective or that infrastructure factors alone are insufficient for accurate prediction.

Variance-Weighted kNN Model 3.1.2

Adjusting weights based on feature variance did not improve accuracy, which remained around 0.63. This supports the conclusion that the primary limitation lies not in feature weighting but in the absence of other key predictors like driver behavior or vehicle conditions. Overall, while infrastructure factors contribute meaningfully, they cannot fully predict crash severity on their own.

Decision Tree Models 3.2

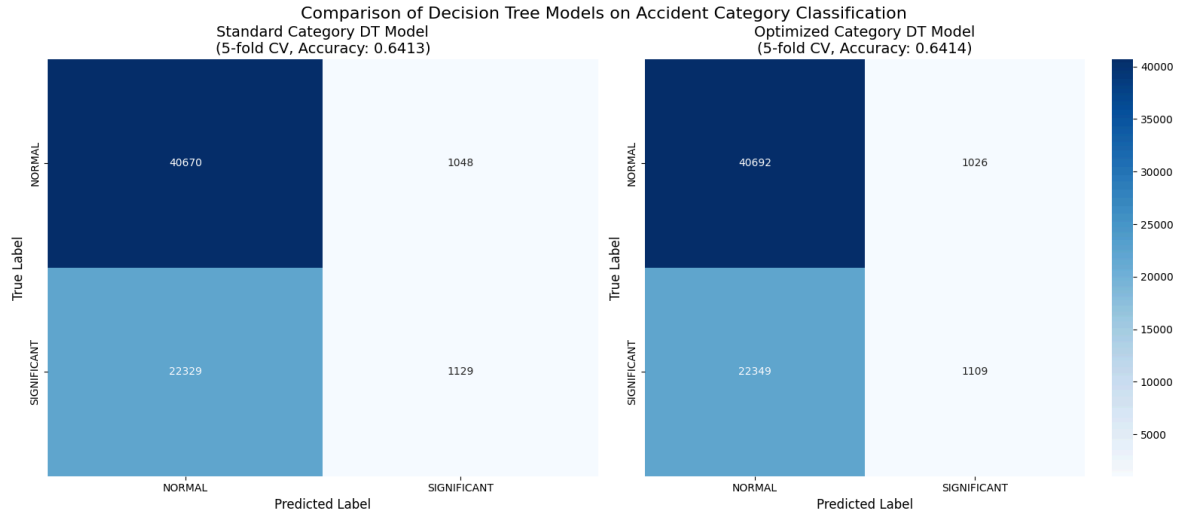

Figure 11: Standard vs. Optimized Decision Tree Model Performance.

Standard Decision Tree Model 3.2.1

The standard decision tree classifier, using original categorical values for the three infrastructure factors, achieved an average accuracy of 0.64 (64%); comparable to both kNN models. This consistency across distinct algorithms reinforces that the observed performance ceiling stems from the limited predictive power of infrastructure-only data, not the modeling technique. As with kNN, the model frequently misclassified “SIGNIFICANT” crashes as “NORMAL,” highlighting the need to incorporate additional factors for better accuracy.

Decision Tree with Optimized Categorical Groupings 3.2.2

A second decision tree model tested whether consolidating the original categories into broader, risk-aligned groupings would affect predictive performance. The grouping process is based on the percentage of serious / fatal accidents obtained from Single Factor Analysis’s result. The consolidation created the following category groups:

Light Condition:

- Poor Lighting: "Dark without Lighting"
- Adequate Lighting: "Dark with lighting," "Daylight," "Limited Light"
- Unknown Lighting: "Unknown"

Road Geometry:

- High Risk Geometry: "Not at intersection" and "Dead end"
- Moderate Risk Geometry: "Multiple intersection," "T intersection," "Y intersection," "Cross intersection," "Road closure"
- Low Risk Geometry: "Private property"

Road Surface:

- Known Road Surface: "Paved," "Gravel," and "Unpaved"
- Unknown Road Surface: "Unknown"

This optimized categorization model yielded similar accuracy of (~0.64), despite reducing category granularity. This supports the validity of consolidating infrastructure categories based on shared risk profiles and highlights the practical benefit of such groupings. For policy and planning, these broader categories provide a simplified yet effective framework for targeting high-risk infrastructure conditions (e.g., treating all "Poor Lighting" scenarios collectively).

Summary: The consistent performance across all four models: both kNN variants and both decision tree implementations, provides robust triangulation of our central finding: infrastructure factors meaningfully influence crash severity but cannot independently determine outcomes with high accuracy. This insight underscores the necessity of comprehensive approaches to traffic safety that address multiple risk dimensions simultaneously.

Data Clustering and Risk Profiling 4

3D Scatter Plot 4.1

Each dot in the plot represents a single data point, placed according to its values for the three features, (LIGHT_RISK, GEOMETRY_RISK and SURFACE_RISK).

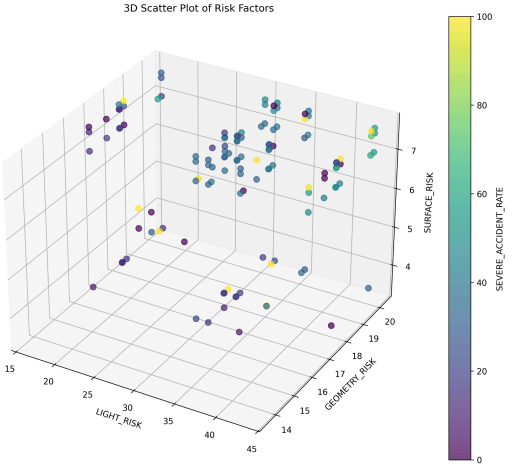

Figure 12: Overview of Risk to Severity.

The 3D scatter plot shows that accident severity is not solely determined by the three risk factors—LIGHT_RISK, GEOMETRY_RISK, and SURFACE_RISK. While some clustering is visible, the data is generally scattered. Severe accidents occur even under low-risk conditions, and high-risk areas sometimes lead to low-severity outcomes. This suggests that other factors that influence severity are at play.

3D Scatter Plot with k-Means Clustering 4.2

k-Means clustering is applied to the accident dataset to group similar data points based on three key risk factors: LIGHT_RISK, GEOMETRY_RISK and SURFACE_RISK.

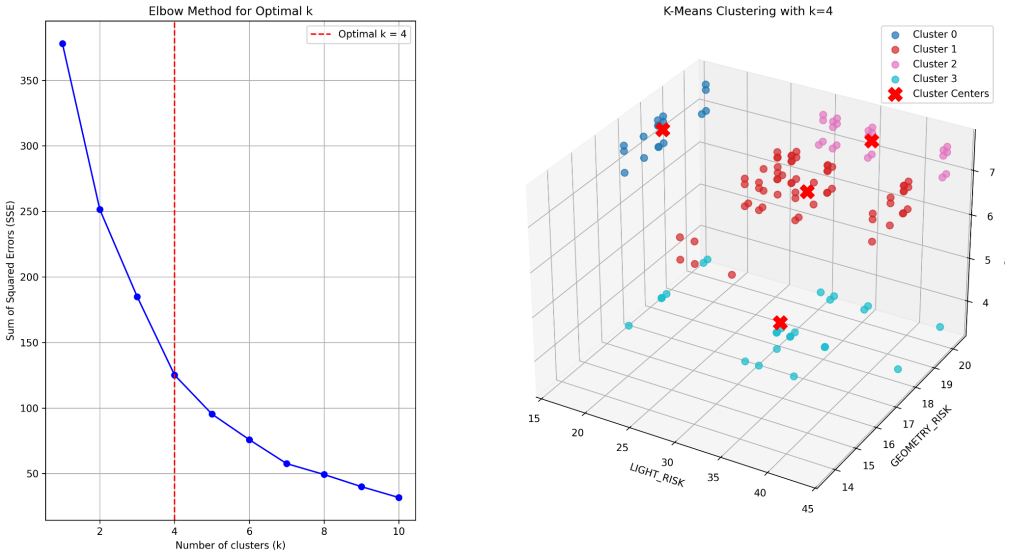

Figure 13: Elbow plot and 3D scatter plot with clustering.

Optimal k is 4.

There are clear boundaries between clusters, suggesting that the algorithm has successfully grouped similar data points together. Some clusters appear denser than others, which may indicate variation in the distribution of the data.

Risk Profiling 4.3

Cluster 0:

- Light Risk: Low light risk (~16.35), consisting of a lot of unknowns in the group.
- Geometry Risk: Moderate to high risk (16.48 - 20.23), predominantly at intersections.
- Surface Risk: Moderate risk (6.99 - 7.60), due to the variance in road types.

Key Traits: Cluster 0 primarily consists of intersections with an unknown light risk.

Cluster 1:

- Light Risk: Varies widely (~16.47 - 43.23), with a focus on poor lighting conditions.
- Geometry Risk: Moderate risk range (~16.47 - 18.24), mostly at different types of intersection structure.
- Surface Risk: Moderate risk (6.99 - 7.60), due to the variance in road types.

Key Traits: Cluster 1 primarily consists of dark, unlit road segments, often at intersections.

Cluster 2:

- Light Risk: Higher risk values (~16.35 - 35.14), influenced by dark lighting conditions.
- Geometry Risk: Consistently high risk (~20.02 - 20.22), with most not being at an intersection or dead end
- Surface Risk: Moderate risk (6.99 - 7.60), due to the variance in road types.

Key Traits: Cluster 2 consists of roads without intersections but with poor lighting.

Cluster 3:

- Light Risk: Most varied (~16.35 - 43.23), covering mostly daylight, but also a few limited lighting, and full darkness accidents.
- Geometry Risk: Broad range (~16.47 - 20.22), as it consists of varied road geometries.
- Surface Risk: Lowest among clusters (~3.47), due to unknown surface conditions.

Key Traits: Cluster 3 consists of roads with better lighting conditions, variable geometry, and minimal surface impact.

Accident Severity and Risk Profiles

- Cluster 0: Highest severe accident rate 58.33%.
- Cluster 1: Highest severe accident rate 54.17%.
- Cluster 2: Highest severe accident rate 60.0%.
- Cluster 3: Lower severe accident rates overall, highest at 31.82%.

Limitations and Improvement Opportunities

Dataset and Context Limitations 1

This study is limited to Victoria police records and focused on only three external factors, excluding many unrecorded accidents and important variables like vehicle characteristics and human factors.

Method Limitations 2

6.2.1 Proportional and Statistical Analysis Limitations 2.1

Our percentage-based analyses do not take into account differences in sample size. This can make some rare combinations appear more serious than they actually are. For example, some combinations are so small such as accidents on “Dead ends” in “Dark without lighting” make the statistical results might be unreliable even though they show high severity rates.

6.2.2 Mutual Information Analysis Limitations 2.2

Mutual information analysis shows correlation rather than causation, especially when working with incomplete or part of the datasets as in this research question. Therefore, it might give us an oversimplified picture of the complex relationships in accident data.

6.2.3 Supervised Machine Learning Limitations 2.3

Our models (around 64% accuracy) have several weaknesses. We only used 3 factors for prediction, while the accident severity depends on many other factors. The data was also imbalanced as there are too many minor accidents compared to serious ones. We also haven’t optimized our models yet. These models often missed serious accidents that occured in generally safer conditions such as daylight. This highlights the limitations of environmental factors alone for prediction.

6.2.4 Clustering Analysis Limitations 2.4

Our k-means clustering had trouble because most accidents (67%) happened in daylight, we used predetermined cluster count, potentially suboptimal distance metrics, and spherical cluster assumptions that may not match actual data distribution. This meant our analysis mostly just separated day and night accidents without finding more detailed patterns.

6.3 Improvement Opportunities 3

Dataset and Context Enhancements 3.1

We should expand our data from multiple states and add information about weather, visibility, traffic volume, driver behavior, vehicle characteristics.

Statistical Analysis Improvements 3.2

We need to set minimum sample counts, use bootstrapping for groups with little data, and apply Bayesian methods for better reliability assessment.

Advanced Correlation Analysis 3.3

We can use more powerful tools like partial correlation analysis, distance measurements, and graph models to see different types of relationships.

Enhanced Machine Learning Approaches 3.4

We should try models like XGBoost or LightGBM, balance data using SMOTE, create new features, and combine multiple models to reduce errors.

Advanced Clustering Techniques 3.5

Instead of k-means, we could use DBSCAN to handle irregularly shaped clusters, Gaussian mixture models for overlapping data, and hierarchical clustering to find more complex patterns.

These methodological improvements would enhance analytical rigor and provide more actionable insights for targeted road safety interventions.

Conclusion

This analysis has provided a structured breakdown of the risk factors, intersection types, and accident severity trends found within the dataset. By leveraging clustering techniques, we have identified key patterns in lighting conditions, road geometry, and surface characteristics, which directly impact the likelihood and severity of accidents.

The findings demonstrate that lighting conditions play a crucial role, with dark, unlit environments consistently linked to higher severe accident rates. Furthermore, complex road geometry, particularly in non-intersection locations, increases risks due to unclear traffic movement patterns, unpredictability and most likely speed. Surface risk varies across clusters, with gravel and unpaved roads exhibiting slightly higher accident severity, although cluster-specific trends reveal exceptions.

The segmentation of data into clusters allowed for the identification of distinct risk profiles, enabling a targeted approach to road safety improvements. By analysing these trends, relevant stakeholders can prioritise infrastructure changes, improve lighting conditions, and optimise clarity and safety at non-intersection roads to reduce severity of accidents.

References

1. Australian Government. (2025, April 14). *Monthly road deaths*. National Road Safety Data Hub. Retrieved from https://datahub.roadsafety.gov.au/progress-reporting/monthly-road-deaths
2. Santoso, J. T., Ginantra, N. L. W. S. R., Arifin, M., Riinawati, R., Sudrajat, D., & Rahim, R. (2021). Comparison of classification data mining C4.5 and Naïve Bayes algorithms of EDM dataset. *TEM Journal, 10*(4), 1738–1744. Retrieved from https://doi.org/10.18421/TEM104-34
3. Santos, K., Dias, J. P., & Amado, C. (2022). A literature review of machine learning algorithms for crash injury severity prediction. *Journal of Safety Research, 80*, 254–269. Retrieved from https://doi.org/10.1016/j.jsr.2021.12.008
4. State of Victoria. (2025). Victoria Road Crash Data. Retrieved from https://discover.data.vic.gov.au/dataset/victoria-road-crash-data.