What combinations of time, road, vehicle condition and driver profiles are most strongly associated with high-risk accidents, and how can we use these to predict accident severity?

Group W10G01

Debao Li debaol@student.unimelb.edu.au	Jingqi Xu jingqixu@student.unimelb.edu.au	Junhao Cui junhaocui@student.unimelb.edu.au	Ruoyu Li ruoyuli@student.unimelb.edu.au

Executive Summary

This report explores the effectiveness of machine learning models in predicting vehicle accident severity whilst also identifying what combinations of time, road, vehicle conditions, and driver profiles are most susceptible to high-risk accidents. By understanding these patterns and implementing these models, it can help enforce safer road policies to reduce the likelihood of severe road accidents.

To address this research question, real-world accident datasets obtained from the Victoria Road Crash Data provided by the State of Victoria were used and were applied to two supervised classification models: K-Nearest Neighbor (KNN) and Decision Trees. We also examined the impacts of different data preprocessing strategies, leveraging encoding systems such as one-hot encoding and label encoding to evaluate their effectiveness on model performance.

In addition to the predictive models, clustering analysis using K-Means was also conducted to detect hidden patterns within accident profiles based on time of day, and speed zones. The analysis revealed three main clusters, lighter everyday vehicles with higher time and speed variability (cluster 0), heavy commercial vehicles with more early morning accidents most likely due to fatigue (cluster 1), and public transport vehicles which showed lower frequency of accidents which may reflect the impact of more regulated training (cluster 2).

Our findings showed that the Decision Tree classifier was more effective at predicting crash severity (66%) compared to KNN (60%). While this suggests a clear improvement, in the broader context of predicting accident severity and making decisions on life-or-death scenarios, a 6% increase in accuracy remains marginal and suggests that there is room for improvement in future research. Additionally, we found that different encoding methods can significantly influence model performance, both in accuracy and computational efficiency, highlighting the importance of choosing the correct preprocessing pipelines.

Future work should explore more advanced classifiers such as random forests or gradient boosting as well as more comprehensive feature engineering and preprocessing techniques such as target encoding. These enhancements could help achieve more robust and insightful predictions for accident severity.

Introduction

Road accidents are a constant public safety concern in Victoria, often resulting in serious injuries or fatalities. This project aims to investigate the research question: What combinations of time, road, vehicle conditions, and driver profiles are most strongly associated with high-risk accidents, and how can we use these to predict accident severity? To explore this question, we analyzed comprehensive datasets collected from Victoria Police Records which document detailed attributes of accidents, vehicles, and individuals involved. Through examining these interconnected variables, we aim to identify key patterns and risk factors that contribute to severe accidents.

Using a combination of data cleaning, correlation analysis, supervised learning, and clustering, we explore how environmental and vehicle-related variables interact to influence accident severity. The goal is not only to uncover high-risk scenarios but also to build predictive models that support proactive safety measures. Ultimately, the findings aim to contribute to safer road environments through better understanding of when and under what conditions the likelihood of serious accidents is elevated.

Methodology

Our methodology is structured into three phases: data preparation and feature selection, supervised machine learning and validation, and data clustering and risk profiling. Each phase builds upon the prior phases, ensuring a coherent flow from extracting raw datasets to fine-tuning machine learning models.

- Data Preparation and Feature Processing
  We begin by merging accidents.csv, vehicle.csv, and person.csv using accident_id as the common key joining all the tables. Once the datasets are combined, they are then preprocessed using various methods such as label encoding, and one-hot encoding. Once the data has been filtered and cleaned, the dataset will partake in feature selection. We employed a hybrid feature selection strategy that combines research paper analysis with statistical analysis where initially, we rank the association between each feature and injury level and choose the top 10-20 features from the merged dataset that has the highest mutual information scores. Once the features are identified, we can utilize research papers to help justify the inclusion of certain features.

- Supervised Machine Learning and Validation
  Once the datasets were preprocessed and the features were selected, we implemented two different machine learning algorithms, K nearest neighbor (KNN), and decision trees. These algorithms will help predict the severity of accidents based on the 10-20 features obtained through the mutual information grid. The prediction then gets validated through various cross validation processes like stratified K-fold analysis.

- Data Clustering and Risk Profiling
  Additionally, clustering techniques such as K-Means are also applied to identify hidden patterns present in the accident data. The clustering analysis will focus on identifying the correlations between accidents by time of day, and road surfaces, effectively assisting in the research question to identify which patterns show the highest accident severity.

Data Preparation and Feature Processing

First, the four datasets were obtained from the CSV files and reorganized to prepare the data for machine learning. To clean the data, techniques such as merging, feature selection, one-hot encoding and feature engineering were implemented. The first section will focus on the use of merging, one-hot encoding and feature engineering. The second section will focus on feature selection and mutual information.

Data Exploration and Analysis

Data Preparation

The first step taken to prepare the datasets was merging the data, which created a single dataset which included all data from the four given CSV files. By observing the combined dataset, there were many unnecessary or redundant columns which contributed little to the data, so feature selection was implemented in this stage. The columns that were kept were deemed as necessary for the mutual information and machine learning sections, and included information such as gender, age groups, accident time and location. One-Hot Encoding was applied for age and converted into binary data, which showed that there were a significant number of 'unknown' genders. For age groups, Feature Engineering was applied, which reorganized the given age groups into 10-year age groups. Unknown age groups were replaced with the most common value. In the end, a table of cleaned data was built, which can be used for machine learning and mutual information sections.

The original AGE_GROUP variable was mapped into broader categories and encoded as integers ranging from 0 to 5, representing: child, teen, young adult, adult, middle age, and senior. These mapped age groups were then combined with SEX to construct a heatmap illustrating the average injury level across different demographic combinations. After removing outliers and unknown values, a heatmap was constructed as shown in figure 1 which revealed that the groups with the highest injury levels are male children, followed by young adult, adult, and middle-aged males. This suggests that males within these age groups are more likely to be involved in high-risk accidents. In particular, the high injury level observed among children may be attributed to a lack of physical protection and awareness. In contrast, females generally show significantly lower injury levels across all age groups, indicating a potential gender-based difference in either accident severity or behavior.

A second heatmap was then created to show which combination of vehicle body styles and road geometry has accident occurrences which result in the highest injury level. As shown in figure 1, three vehicle styles that are involved in the highest number of accidents are small wagons, sedans and wagons, and the three road geometries for the highest number of accidents are T-intersections, cross intersections and not at intersections. The 'others' category includes every other type of vehicle style or road geometry. From the heatmap, the injury level from accidents for all vehicle body types is lower when compared to other road geometries in the visual. In addition, accidents involving wagons of all road geometry types are higher than the other vehicle types. Injury level for accidents involving small wagons is lower than wagons but higher than sedans.

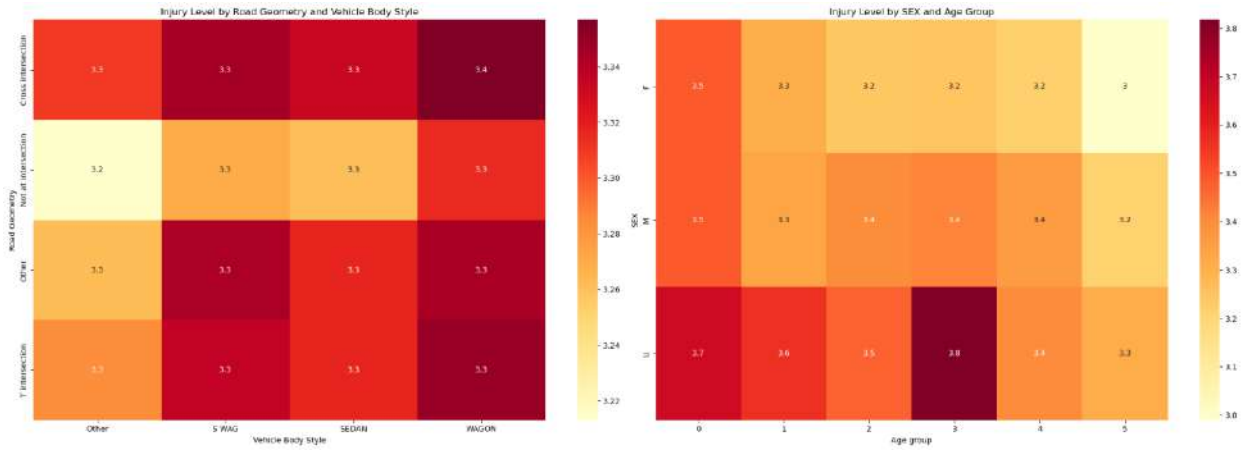

Figure 1: (Left) Heatmap showing the injury levels for the vehicle body styles and road geometries that are involved in the highest number of accidents. (Right) Heatmap showing how injury level is related to different age groups and gender of accident victims. (0: children; 1: teenagers; 2: young adults; 3: adults; 4: middle-aged adults; 5: seniors)

Feature Processing (MI, selection)

Selecting a correct set of features is essential for data generation and training purposes. To achieve this, two methods are used during the data construction phase: the use of mutual information (MI), and support from reliable research literature (Rolison et al., 2018; Wang et al., 2013). Mutual Information (MI) measures the amount of information that two random variables provide about each other (Orlitsky, 2003). When the mutual information value between two variables approaches 1, it indicates a strong dependency between them. This helps the analyst identify which variables are most relevant for data modeling. In Python, a table was constructed to display the mutual information values for most of the columns across the different CSV files. Specifically, the columns were evaluated based on their relationship with the injury level of individuals involved in accidents. This helped determine which features should be retained for final data construction.

Feature	Mutual Information
SEVERITY	0.213508
ROAD_USER_TYPE_DESC	0.037687
DCA_DESC	0.035039
HELMET_BELT_WORN	0.030648
LEVEL_OF_DAMAGE	0.018813
SEX	0.018205
ACCIDENT_TYPE_DESC	0.018078
VEHICLE_MODEL	0.017659
AGE_GROUP	0.015753
ACCIDENT_TIME	0.012358
SPEED_ZONE	0.006149
VEHICLE_BODY_STYLE	0.001403
VEHICLE_MAKE	0.001348
VEHICLE_TYPE_DESC	0.001313
ROAD_GEOMETRY_DESC	0.001012
SEATING_CAPACITY	0.001003

Table 1: Mutual Information table; Chosen column vs Injury Level

Discussion and Interpretation

Data Preparation

After data preparation and cleaning, and creating heatmaps related to the research question, the following interpretation can be made:

- The three vehicle body styles that are involved in the highest number of accidents (wagon, small wagon, sedan) are common vehicle body styles for cars, which shows a potential bias in the data as there are a greater number of data collected which is related to accidents involving cars. The 'others' category includes motorcycles, bikes and scooters which have been grouped together, as there is less data for each unique vehicle body style under the 'others' category.
- The bias of the data towards car types such as wagons and sedans could mean machine learning models might not be able to accurately predict for vehicle types such as motorcycles and scooters.
- For the four road geometries involved in the highest number of accidents, accidents 'not at intersection' have significantly lower injury levels than the other categories. This road geometry type is broad, which could explain why the data is skewed towards a lower injury level than the other categories, as it includes a wide range of injury levels. Interestingly, the 'others' category for road geometry has similar injury level values when compared to accidents at cross intersections and T-intersections.
- People with sex 'U' (unknown) have higher injury levels across all age groups when compared to injury levels for male and females. This could mean data collection which could mean a higher G/K ratio which allowed for greater protection of privacy for the people involved in the accidents, however this means there is poorer data preservation, as gender data might be more often grouped under the unknown category.
- A higher injury level is evident across all genders for category 0 in age groups (children), which shows that children are at higher risk to greater injury in car accidents, due to varying factors such as seat belt usage, and which seat they were in during the accident. Data privacy could also mean data for children that were not severely injured might have not been recorded for privacy protection. Injury level for seniors of all genders is consistently lower than the other age groups, which be explained by factors such as greater experience in driving due to older age, or the wide age group that was categorized as seniors may have skewed the data.

Feature Selection

As shown in table 1, it is clear which features hold stronger mutual information values. Columns with values containing more than three zeros after the decimal point were not considered significant, as they do not exhibit a strong correlation with injury level (and are not shown in the figure). To further support the column selection process, scientific research papers on related topics were consulted. These papers indicate that age group and sex of drivers significantly influence the likelihood of a collision. Additionally, road geometry was also identified as a key factor in causing accidents. However, its MI value in this dataset was relatively low, possibly due to the variability in the data sources and how different columns relate to injury levels.

Limitations and Improvement Opportunities

Data Preparation

Limitations for the merging section include the selection of using an ‘outer’ merge technique, which combines every column in the dataset together. Poor matching of datasets could add many ‘unknown’ values into the data, which would have to be cleaned or replaced with additional code, since large number of unknown data can create noise and reduce effectiveness of machine learning models. As an improvement, columns containing unknown values could be feature engineered, and unknown data can be replaced with the most common occurring data. Additionally, applying One-Hot Encoding to columns such as ‘ROAD_USER_TYPE_DESC’ can add many new columns for each unique data in that column. Feature engineering could also be applied to the other columns, such as accident time, which can be categorized into parts of the day (morning, afternoon, evening, night).

Another limitation in the data preparation process lies in the SEX category, where a significant number of entries are labeled as "Unknown." This introduces outliers and uncertainty when constructing visualizations such as heatmaps, potentially skewing the interpretation of injury levels across demographic groups. The presence of these unknown values can distort the true distribution and hinder accurate insights. To address this issue, one possible solution is to impute the missing values by replacing "Unknown" with either "Male" or "Female" based on contextual patterns or distributional assumptions. This can enhance the accuracy and readability of the heatmap or any other visual analysis, leading to more reliable conclusions.

Feature Selection

A key limitation in the feature selection process is the presence of unknown or missing values in critical columns such as AGE_GROUP, SEX, and HELMET_BELT_WORN. These missing entries reduce the effectiveness of statistical techniques like mutual information, as incomplete data can distort the true relationship between features and the target variable. Although some unknowns were handled through imputation or removal, these decisions may introduce bias or lead to the exclusion of potentially valuable data points. A more robust strategy—such as probabilistic imputation or leveraging auxiliary datasets—could improve the integrity of the analysis (Nguyen, 2025).

Another limitation is the inconsistency between mutual information values and existing research literature. For instance, scientific studies consistently emphasize the importance of factors like road geometry and driver demographics (e.g., age and sex) in determining accident severity. However, the mutual information scores for some of these features were relatively low. This discrepancy may be due to limitations in the dataset, such as uneven distribution, lack of granularity, or the influence of confounding variables. It highlights the need to balance quantitative analysis with domain knowledge—statistical scores alone should not be the sole basis for feature selection, especially in fields as complex as road safety.

Supervised Machine Learning and Validation

Following the methodology, supervised machine learning models were constructed based off the preprocessed datasets. To ensure robust and accurate feature prediction models, KNN and Decision Tree classifiers were implemented. The first section will focus on KNN classifiers while the second section will focus on Decision Tree classifiers.

KNN Analysis

Data Exploration and Analysis

The first classifier used was the K Nearest Neighbor or KNN classifier. Due to its reliance on distance metrics, the preprocessed datasets had to be processed again, which led to an interesting sub comparison between the implementation of label encoding and one hot encoding. Although the key comparison evaluates the effectiveness of one classifier to the other (i.e KNN and Decision Trees), this sub comparison provides more insightful information on the choices made in the preprocessing steps and how these elements could consequently affect the overall quality and accuracy of models. Once the dataset was normalized and processed for KNN, a stratified K-Fold cross validation was applied to the top 15 features for both encodings. The results for both approaches are displayed below where each approach has a confusion matrix as well as a classification report on model accuracy.

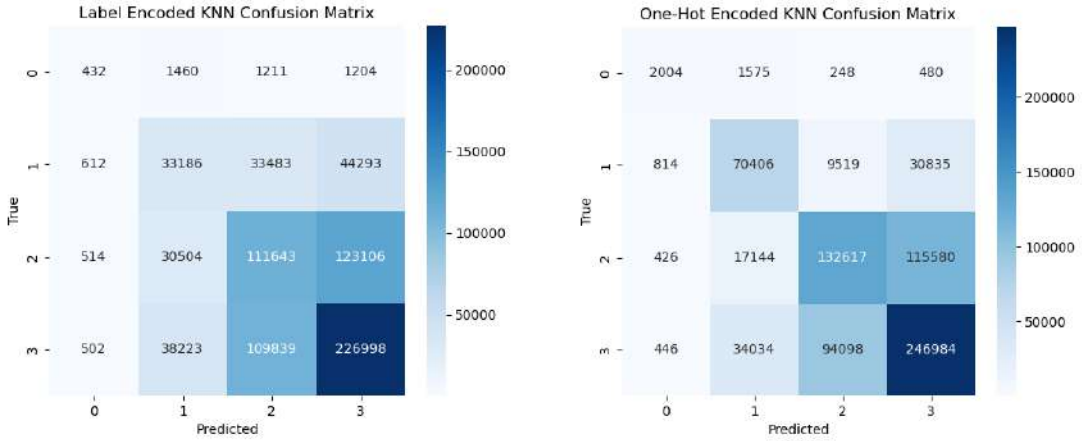

Figure 2: Confusion matrix for both label and one hot encoding

	precision	recall	f1-score	support
1	0.21	0.1	0.14	4307.0
2	0.32	0.3	0.31	111574.0
3	0.44	0.42	0.43	265767.0
4	0.57	0.6	0.59	375562.0
accuracy	0.49	0.49	0.49	0.49
macro avg	0.39	0.36	0.37	757210.0
weighted avg	0.49	0.49	0.49	757210.0

	precision	recall	f1-score	support
1	0.54	0.47	0.5	4307.0
2	0.57	0.63	0.6	111574.0
3	0.56	0.5	0.53	265767.0
4	0.63	0.66	0.64	375562.0
accuracy	0.6	0.6	0.6	0.6
macro avg	0.58	0.56	0.57	757210.0
weighted avg	0.6	0.6	0.6	757210.0

Table 2: Accuracy scores for label (left) and one-hot encoding (right)

Discussion and Interpretation

Through the visualization provided by the confusion matrix as well as the accuracy scores, various interpretations of the results can be identified. The key interpretations are displayed below:

- Model performance between label encoding and one hot encoding
  Throughout the implementation process of both KNN classifiers, the different encoding methods led to significant deviations in accuracy even though all other variables such as simulation number and feature number remained fixed. This can be shown in table 2 where label encoding achieved an average accuracy of 49% while one hot encoding had an average accuracy of 60%. This indicates that the way categorical features are interpreted can be a major factor when establishing the model's ability to learn. The implications of the lower accuracy for label encoding can be traced to its artificial assignment of ordinal relationships between different descriptions within a category. Since KNN is a distance-based model, the misleading distances assigned through label encoding can alter and reduce the accuracy of the KNN classifier. Meanwhile, one-hot encoding treats all categories equitably, preserving the non-ordinal nature of the features, which could be a key factor to its higher accuracy. By understanding the factors that influence model accuracy, in this case, the encoding methodologies, it allows a better understanding of what tools are more effective when predicting accident severity.

- Confusion Matrix between label encoding and one hot encoding
  As shown in figure 2, the differences in confusion matrix patterns between label and one-hot encoding indicated that one-hot encoding had a stronger diagonal alignment while the label encoded model frequently confused fatal injuries with other injury categories. This suggests that label encoding introduced noise to the KNN classifier which made it harder for the model to establish clear boundaries between severity levels. These factors could be traced back to the idea of assigning artificial relationships between descriptions within a feature as mentioned earlier, which leads to misleading distances. Since our research seeks to predict high-risk accident conditions, the misclassification of injury categories have significant implications in real-world applications like emergency prioritization and risk communication which can lead to more overestimations and under-preparations to life-or-death scenarios.

- Tradeoff between accuracy and computational performance
  While accuracy is the key determining factor in evaluating model effectiveness, it must also be balanced against computational efficiency, especially when working with larger datasets. During the training and testing phase, there was a clear trade-off between label and one hot encoding. While label encoding degraded model accuracy by around 11%, its execution time was comparatively faster than one-hot encoding with execution times for model prediction being around 15 seconds compared to one-hot encoding which took around 5 minutes. This contrast in runtime highlights the computational cost of one-hot encoding, especially when there are multiple columns as this increases the dimensionality of the dataset, effectively slowing down the distance calculations in KNN. This raises an interesting question: To what extent can speed be more important than accuracy? The general interpretation will depend on the context and conditions but for the case of predicting accident severity, accuracy should be prioritized as a misclassified injury can lead to life threatening consequences. Thus, in this context, the additional processing time required by one-hot encoding is a reasonable and justifiable trade-off for its higher accuracy and reliability of results.

Limitations and Improvement Opportunities

When the preprocessed dataset was first obtained, the features contained both integer and string based categorical data. While this is generally fine for most prediction models, as mentioned earlier in the discussion, due to the nature of KNN having a distance metric, it can only take in integer data. This resulted in additional preprocessing through one-hot encoding and label encoding, resulting in an inefficient allocation of time. To improve future analysis, depending on the supervised machine learning model used, two separate datasets should be created during the preprocessing phase, one for numerical, and the other one being a combination.

Another limitation that is targeted especially for KNN classifiers is its scalability when it comes to multiple features. Since KNN requires numerical data, various features from the preprocessed dataset had to be one-hot encoded and label encoded. While 10-20 additional columns will not affect the processing time of KNN classification too much, features such as "VEHICLE_MODEL", which had over 12,000 distinct descriptions will make it impossible to effectively process and predict outputs in a reasonable time. Luckily, as shown in figure y in the preprocessing section, "VEHICLE_MODEL" did not have the highest mutual information score which meant removing this feature from the KNN classifier will not impact the accuracy of the predictions too much. However, this becomes problematic when features with very high cardinality have a high mutual information score as removing the feature will lead to significant deviations towards overall model accuracy. One possible approach would be to use alternate encoding methods such as target encoding which replaces each category with the mean target value for that category.

Furthermore, while it is evident from the sub-comparison that one-hot encoding is more effective than label encoding for KNN predictions with an 11% improved performance, looking at the greater scope, especially in the context of the research question, 60% accuracy is still not feasible when it comes to predicting life-threatening situations. Hence, another analysis will be conducted comparing the effectiveness of KNN with another prediction classifier, decision trees.

Decision Tree Analysis

Data Exploration and Analysis

The Decision Tree model was evaluated using Stratified K-Fold Cross-Validation to ensure balanced representation of severity classes across folds. The accuracy table and confusion matrix are shown below.

	precision	recall	f1-score	support
1	0.69	0.69	0.69	4307.0
2	0.65	0.68	0.67	111574.0
3	0.62	0.65	0.64	265767.0
4	0.71	0.67	0.69	375562.0
accuracy	0.67	0.67	0.67	0.67
macro avg	0.67	0.67	0.67	757210.0
weighted avg	0.67	0.67	0.67	757210.0

Table 3: Accuracy scores for decision tree classifier

The model showed very stable accuracy across all folds, indicating consistent performance. While the overall accuracy may appear moderate (~66.6%).

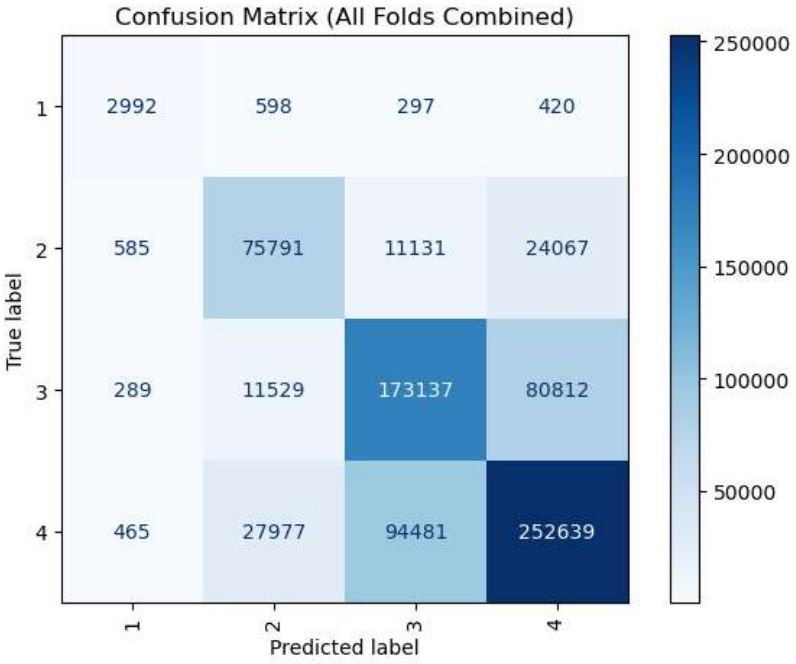

Figure 3: Confusion matrix for decision tree classifier

From the confusion matrix and the accuracy table, we can see that:

- P(Predicted Fatal | Actual Fatal) = 2,992 / 4,307 ≈ 0.695 - this is the recall for fatal injuries.
- P(Actual Fatal | Predicted Fatal) = 2,992 / (2,992 + 585 + 289 + 465) ≈ 0.686 - this is the precision for fatal injuries.
- These conditional probabilities suggest that when the model predicts a fatal injury, it is correct nearly 69% of the time — which is strong given the rarity of fatal events.
- The model shows reasonably good performance in predicting injury levels, especially at the extremes — Fatal injuries (1) and No injury (4). The F1-score for both of these classes was 0.69, indicating that the model is fairly reliable in identifying the most and least severe cases.
- The biggest confusion occurred between Serious (2) and Other Injury (3) levels. These were often misclassified as each other, likely because they share similar accident conditions — such as moderate speed, common road types, or common accident mechanisms. This overlap in context reduces the model's ability to clearly separate them.
- Our model handled Fatal Injuries reasonably well (F1 = 0.69), despite this class having far fewer examples. This suggests that certain patterns — such as high speed, vulnerable road users (e.g., pedestrians), or specific accident types — are distinctive enough to allow the model to correctly classify many fatal cases. However, some fatal cases were still misclassified as Serious or Other injuries, which can have serious consequences if used in real-world triage or prioritization systems.
- The Not Injured class had the most data, and the model performed best here in raw count, but precision and recall are balanced, indicating that the model is not just guessing this label, but actually learning from patterns in the data.
- Overall, the model captures injury level trends fairly well but could be improved to better separate moderate injury levels.

Discussion and interpretation

- The Decision Tree model was developed to predict injury levels in traffic accidents using features such as accident type, user role, road geometry, and demographic data. The model achieved an average accuracy of 66.6% across five stratified folds and showed balanced performance across all injury classes. This indicates that the model was able to generalize reasonably well, despite challenges such as class imbalance and overlapping feature conditions between moderate injury categories.

- Injury level 1 (Fatal) and 4 (Not Injured) had the highest precision and F1-scores (both around 0.69), which suggests that these extreme cases were easier for the model to classify due to distinct patterns, such as high-speed impact or pedestrian involvement in the case of fatalities, and low-impact accidents in the case of no injuries. But in contrast, Serious (2) and Other Injury (3) classes were often misclassified as each other. These mid-level categories had more feature overlap and shared characteristics such as vehicle collisions at moderate speeds, resulting in more confusion for the model.

- The use of one-hot encoding worked effectively with the decision tree, as the tree structure can make splits on individual categories without being misled by artificial ordinal relationships. So this contrasts with models like KNN, where label encoding led to lower performance due to its reliance on numerical distance metrics that falsely imply order between categories.

- When comparing the Decision Tree to the KNN classifier (as reported in a similar task), the difference in performance was clear. While KNN achieved only ~60% accuracy with one-hot encoding and ~49% with label encoding, the Decision Tree consistently reached 66.6% accuracy with one-hot encoding. This suggests that the decision tree is more suitable for this type of categorical and structured data, especially because it does not rely on distance calculations, which are heavily influenced by encoding choices in KNN.

- Our Decision Tree model trained and predicted faster than the KNN model, which required significantly more computation time with one-hot encoding. This shows that the Decision Tree not only outperforms KNN in accuracy but also scales better with large feature sets and high-dimensional data.

- These findings highlight the importance of choosing the right model and encoding method for the task. While one-hot encoding improved performance in both cases, the Decision Tree was more robust and interpretable. It was better able to capture non-linear relationships between features and injury levels and provided stable, explainable predictions across multiple folds.

- Understanding these strengths can be crucial for real-world applications, such as injury risk assessment and automated accident triage where both prediction accuracy and speed are essential. Based on our results, the decision tree model presents a more effective approach for injury level prediction than KNN, particularly in large datasets with complex categorical variables.

Limitations and Improvement Opportunities

- One major limitation of this analysis is that not all columns in the dataset were used. The selection of features was based on assumptions about which attributes were most relevant to predicting accident severity, such as vehicle severity, accident type, road user type, and road geometry. However, by leaving out other potentially useful variables like VEHICLE_MODEL, VEHICLE_TYPE_DESC, or ACCIDENT_DATE, the model may have missed patterns that could improve performance. For example, certain vehicle types may be more commonly involved in severe accidents, or accidents occurring during specific months or seasons may show different severity patterns. Excluding these features reduces noise but also limits the scope of insight and the model's ability to generalize.

- The model struggled to clearly distinguish between Serious Injury (2) and Other Injury (3), with frequent misclassifications between these two. This may be due to overlapping feature conditions in the dataset, but it also suggests that the input features do not provide enough granularity to capture subtle differences between moderate injury levels.

- The encoding of some variables, such as sex, was done using separate one-hot columns (SEX_F, SEX_M, SEX_U) rather than a single label or category. This increases the dimensionality of the data and may not have helped the model, especially since the sex variable appeared to have limited influence on the severity prediction. Similarly, ACCIDENT_TIME was used as a raw value and not processed into more meaningful categories like time of day, which could have helped identify high-risk periods (e.g., night-time accidents being more severe).

- While the accuracy of decision tree classifiers is higher than KNN classifiers with decision trees being 6.66% more accurate, when observing the broader context in terms of the research question, this increase remains suboptimal, especially when mortality depends on these predictions. To improve, a wider range of models and classifiers should be explored as well as alternative encoding methodologies to enhance predictive performance.

Data Clustering and Risk Profiling

Data Exploration and Analysis

We conducted a clustering analysis to identify hidden patterns between accidents by time of day, and speed zones using K-Means clustering. The clustering was conducted on the aggregated mean values of speed, time, seating capacity and weight and were grouped by 'VEHICLE_TYPE_DESC', 'TRAFFIC_CONTROL_DESC', and 'ROAD_GEOMETRY_DESC'. The numerical features were then normalized to ensure consistency when running distance-based metrics such as K-Means. The K-Means elbow plot indicated that 3 clusters was the most appropriate. Understanding this, a scatter plot was constructed, illustrating clusters by hour of accident on the x axis and the speed zone on the y axis.

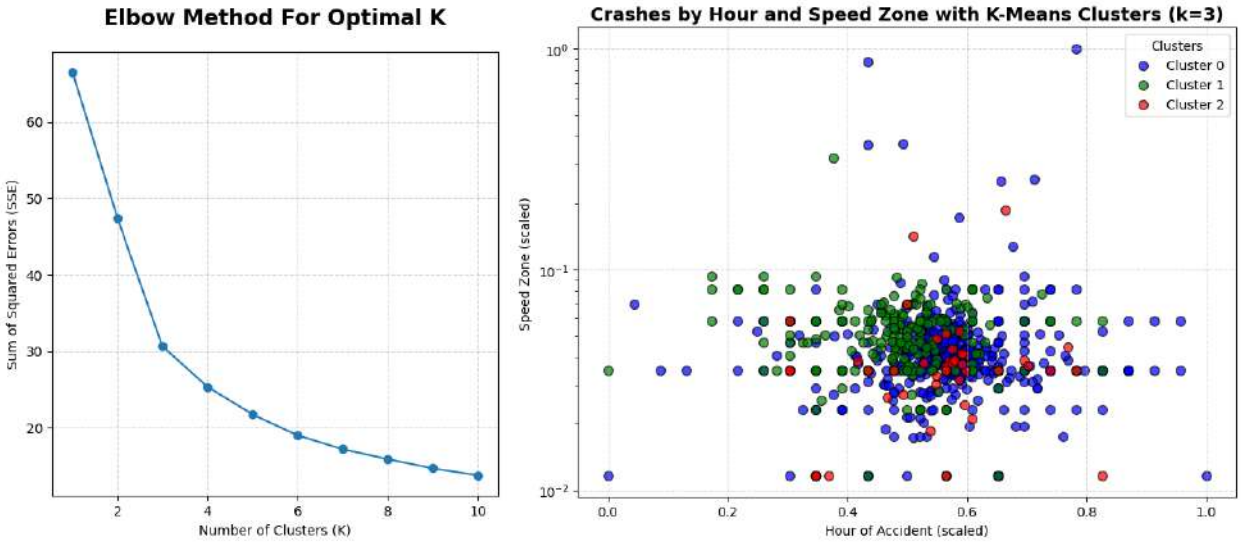

Figure 4: Identifying the most optimal k value through elbow method and then applying the kth value (k=3) into a cluster

Discussion and Interpretation

Here are the key insights from the clustering results, the cluster csv's are obtained through the clustering code located in the "AClustering.py" section.

- Cluster 0: Light vehicles
  As denoted by the cluster 0 csv output from the clustering code segment, vehicles in this cluster are typically everyday people's vehicles such as SUV's and sedans. Accidents in this cluster occurred most frequently between 1-4pm, potentially aligning with after school or after work congestion and general daytime traffic densities. There was also greater speed zone variability which includes higher speed accidents. This suggests that accidents may involve reckless or inexperienced drivers who are new to driving. These insights highlight how inexperience and high-density traffic periods contribute to more unpredictable accident profiles, suggesting the need for targeted safety policies and stricter enforcement during peak traffic hours.
- Cluster 1: Heavy Commercial vehicles
  This cluster consists of vehicles that weigh more than 4.5 tons with moderate speed variance and accidents primarily occurring between 3am-12pm as shown in figure 4. The early hours suggest that accidents are likely a result of fatigue related driving and long-haul schedules (Motor et al., 2016). This cluster emphasizes the risk profiles for commercial drivers and the impact of fatigue on accident timing suggesting that even adequate employment training can be counteracted by fatigue. These findings support greater enforcement of policies promoting mandatory rest periods and better logistic schedules to reduce accident frequency.
- Cluster 2: Public transportation and buses
  This was the least frequent cluster, consisting primarily of buses with low accident speed and time variance and a concentrated accident period between 12-3pm. Compared to other vehicle driver profiles, public transport drivers travel on familiar routes and typically require specialized certification and training which can be reflected by its lower frequency and tighter variance which creates a tangible impact on safety outcomes. This cluster reinforces the value of driver certifications and suggests that a broader adaptation of mandatory training is effective at reducing the probability for accidents.

Limitations and Improvement Opportunities

One limitation of the cluster analysis was the lack of clustering variety for feature selection as only four features: speed zone, hour of accident, seating capacity, and tare weight were used for the clustering. This omits other potentially more influential features such as the day of the week, or driver age, which could reveal more significant patterns. To improve the clustering analysis, more features should be incorporated to allow the cluster to capture more comprehensive risk profiles for more precise observations.

Another limitation was the lack of comprehensive outlier handling. While "SPEED_ZONE != 999" was effective at handling extreme outliers, other potential outliers in speed or weight may have distorted the overall cluster centroids. To improve, a more robust preprocessing and outlier handling method such as the Z-score should be considered (Ektamaini, 2020).

There were also limitations in the presentation of the clustering. While a log scale on the y axis allowed a better visualization of the clusters, the meaning of the speed zone was lost, making it harder to interpret the results. To improve, the actual speed zones should be kept after the normalization in a list which could then be used in the cluster graph.

Conclusion

Road accidents are affected by factors such as drivers, road and vehicle conditions. Using Victoria Police data for vehicles, accidents and drivers, the starting datasets were cleaned and prepared for machine learning. Feature engineering, selection and Mutual Information tables were important in selecting data that was important to the research question. From the cleaned data, it is evident that many accidents in the datasets involve car types such as sedans and wagons. Additionally, cross and T-intersections were clearly shown as the road type that more dangerous accidents occur at.

To predict injury levels from these conditions, both supervised and unsupervised machine learning methods were applied. K-Nearest Neighbours (KNN) was tested but showed limitations due to its reliance on distance metrics and encoding sensitivity. In contrast, the Decision Tree model performed more reliably, achieving ~66.6% accuracy and handling categorical features more effectively. It identified key patterns linking fatal and serious injury outcomes to factors like speed zones, road geometry, and vulnerable road users. However, in the context of accident severity where mortality is being evaluated, further research should be conducted to identify more accurate prediction models

K-Means clustering was also used to uncover common accident profiles based on time of day and vehicle characteristics. It revealed clear groupings such as light vehicle accidents during peak traffic, fatigue-related incidents involving heavy commercial vehicles, and low-risk public transport accidents during the day. These insights help explain how accident risk varies with time, road layout, and vehicle type. Ultimately, our findings exhibits that machine learning can effectively uncover patterns in accident data and predict injury outcomes based on contextual conditions, which supports the research aim by highlighting how specific combinations of time, road, and vehicle factors contribute to high-risk accidents and how these patterns can inform prevention and response strategies.

References

Alon Orlitsky. (2003). Information Theory. *Elsevier EBooks*, 751–769. https://doi.org/10.1016/b0-12-227410-5/00337-9

Ektamaini. (2020, May 10). *Z score for Outlier Detection - Python*. GeeksforGeeks. https://www.geeksforgeeks.org/z-score-for-outlier-detection-python/

Motor, in, Committee on National Statistics, Board on Human-Systems Integration, Division of Behavioral and Social Sciences and Education, Transportation Research Board, & National Academies of Sciences, Engineering, and Medicine. (2016, August 12). *Fatigue, Hours of Service, and Highway Safety*. Nih.gov; National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK384974/

Nguyen, M. (2025). Chapter 11 Imputation (Missing Data) | A Guide on Data Analysis. In *bookdown.org*. https://bookdown.org/mike/data_analysis/imputation-missing-data.html

Rolison, J. J., Regev, S., Moutari, S., & Feeney, A. (2018). What are the factors that contribute to road accidents? An assessment of law enforcement views, ordinary drivers' opinions, and road accident records. *Accident Analysis & Prevention*, *115*(0001-4575), 11–24. https://doi.org/10.1016/j.aap.2018.02.025

State of Victoria. (2025). Victoria Road Crash Data. Retrieved from https://discover.data.vic.gov.au/dataset/victoria-road-crash-data.

Wang, C., Quddus, M. A., & Ison, S. G. (2013). The effect of traffic and road characteristics on road safety: A review and future research direction. *Safety Science*, *57*, 264–275. https://doi.org/10.1016/j.ssci.2013.02.012