Which combination of driver, vehicle, and environmental factors most significantly influences road accident severity in Victoria, and how accurately can we predict whether an accident will result in a serious injury or fatality?

Group W{08}G{4}

Yuchao Bian
COMP20008
jonathanbian@student.unimelb.edu.au

Tianhao Tang
COMP20008
tianhaot1@student.unimelb.edu.au

Sandiv Malika Dathunarachchi
COMP20008
sdathunarach@student.unimelb.edu.au

Qinyi Zheng
COMP20008
qinyi.zheng.1@student.unimelb.edu.au

Executive Summary

In this report, we will be investigating how the combination of the speed zone and the road layout where the accident took place, the light condition, the day of a week, and the number of vehicles involved in the accident would affect the severity of the accident. Moreover, we will visualize the proportion of different severity of injuries either across different vehicle age groups, different speed zones where the accident occurred, different light conditions or for different types of road users. The following part of the report will focus on introducing how we preprocessed the data, and manifesting the result received from our model as well as analyzing the result and potential causes for the result. Additionally, clustering techniques were applied to driver demographic data to identify distinct risk profiles, further enriching our understanding of factors contributing to road accidents.

Methodology

This study used a multi-stage analytical approach to investigate the factors influencing road accident severity in Victoria and to assess the predictability of severe outcomes. The process involved data acquisition and preprocessing, visual analysis, and the development and evaluation of supervised machine learning models.

Data Preprocessing

The data frames selected for this topic were ‘person.csv’, ‘vehicle.csv’, and ‘accident.csv’ from Victoria Road Crash Data - Victorian Government Data Directory (2024). We firstly read the data files on Python and selected the attributes we were interested in to focus on.

Person Data Frame Processing

For the person file, we created a subset data frame by including columns 'ACCIDENT_NO', 'PERSON_ID', 'VEHICLE_ID', 'SEX', 'AGE_GROUP', 'INJ_LEVEL', 'SEATING_POSITION', 'HELMET_BELT_WORN', 'ROAD_USER_TYPE','ROAD_USER_TYPE_DESC', 'TAKEN_HOSPITAL', 'EJECTED_CODE'. To handle the missing values for each column, we filled in the missing ones with the mode value of that column, this is to include every data piece whilst minimizing the change of the distribution of each level of an attribute from the original data frame.

Vehicle Data Frame Processing

For the vehicle file, we created the subset by including columns 'ACCIDENT_NO', 'VEHICLE_ID', 'VEHICLE_YEAR_MANUF', 'ROAD_SURFACE_TYPE', 'ROAD_SURFACE_TYPE_DESC', 'VEHICLE_MAKE', 'VEHICLE_TYPE', 'VEHICLE_TYPE_DESC', 'LEVEL_OF_DAMAGE', 'INITIAL_IMPACT', 'DRIVER_INTENT', 'VEHICLE_MOVEMENT', 'CAUGHT_FIRE', 'TOTAL_NO_OCCUPANTS'. For the 'VEHICLE_YEAR_MANUF' column, we fixed the year formatting error by replacing zeroes with missing values to avoid having impossible vehicle ages in the following processing. Then except for the 'VEHICLE_YEAR_MANUF' column, we handled the missing values the same way as for the person file, to avoid having the age being negative values.

Accident Data Frame Processing

For the accident file, the subset included columns 'ACCIDENT_NO', 'ACCIDENT_DATE', 'ACCIDENT_TIME', 'ACCIDENT_TYPE', 'ACCIDENT_TYPE_DESC', 'DAY_OF_WEEK', 'LIGHT_CONDITION', 'NO_OF_VEHICLES', 'NO_PERSONS_KILLED', 'NO_PERSONS_INJ_2', 'NO_PERSONS_INJ_3', 'NO_PERSONS_NOT_INJ', 'NO_PERSONS', 'POLICE_ATTEND', 'ROAD_GEOMETRY', 'SEVERITY', 'SPEED_ZONE'. We used Regular Expression to extract the accident year from column 'ACCIDENT_DATE' and created a new column for the accident year. The file did not include any missing values.

Merging Data Frames

We merged each data frame output after our manipulation from above on the common primary key 'ACCIDENT_NO'. We firstly used inner join (based on the feedback received for Assignment 1 to ignore all the unmatching data pieces) to merge vehicle and accident data frames, and we calculated the vehicle age and created a new column for it from the difference between manufacture year and accident year. Moreover, we filled in the missing values for the car age column with the mean age. The intention of using mean was because having decimal places for age does not affect our following manipulations for grouping the vehicle ages. However, it will be problematic no matter what to use to fill the missing values, because it will just make one group having more values than the other. Then we finally merged the person data frame after the above mentioned manipulation to save memory space for our code.

Following the initial data merging and cleaning as described above, we created several derived features to facilitate analysis. These included SEVERITY_DESC (mapping numerical severity codes to descriptive labels), VEHICLE_AGE_GROUP (categorizing calculated vehicle age), SPEED_ZONE_GROUP (grouping numerical speed zones), LIGHT_CONDITION_DESC (mapping numerical light condition codes to descriptions), DAY_WEEK_DESC (mapping numerical day codes to day names), and ROAD_GEOMETRY_DESC (mapping numerical road geometry codes to descriptions). We then saved the processed data, including these derived features, into two distinct datasets: final_processed_data.csv for general visualization and exploration, and data_for_modeling.csv specifically prepared for the predictive modeling tasks.

Exploratory Data Analysis and Visualization

To visually explore the relationships between various factors and accident severity, we used the seaborn and matplotlib libraries in Python to generate a series of visualizations. The primary visualization technique employed was the normalized stacked bar chart. This method was chosen for its effectiveness in comparing the proportional distribution of different severity outcomes (Fatal, Serious Injury, Other Injury, Non-Injury) across categories of selected variables, such as speed zone groups, vehicle age groups, road user types, and light conditions. An initial bar chart was also generated to show the overall distribution of accident severity levels for contextual understanding.

The analysis focused on unique accident occurrences for factors like speed zone and light condition, unique vehicles for vehicle age, and individual persons for road user type.

Predictive Modelling of Accident Severity

To address the research question concerning the predictability of severe accidents, a binary classification task was defined. The SEVERITY attribute was transformed into a target variable, IS_SERIOUS_FATAL, where accidents resulting in a fatality (Severity code 1) or serious injury (Severity code 2) were labeled as '1', and all other outcomes were labeled as '0'. This binary target was created within the data_for_modeling.csv dataset, which was structured at the unique accident level.

The features selected for prediction from data_for_modeling.csv were SPEED_ZONE_GROUP, LIGHT_CONDITION_DESC, DAY_WEEK_DESC, ROAD_GEOMETRY_DESC, and NO_OF_VEHICLES. Prior to model training, any remaining missing values in these selected features were handled: numerical features were imputed with their median, and categorical features with their mode.

A ColumnTransformer from scikit-learn was employed for feature preprocessing. Numerical features (i.e., NO_OF_VEHICLES) were standardized using StandardScaler to have zero mean and unit variance. Categorical features (SPEED_ZONE_GROUP, LIGHT_CONDITION_DESC, DAY_WEEK_DESC, ROAD_GEOMETRY_DESC) were converted into a numerical format using OneHotEncoder, with the drop='first' parameter set to mitigate multicollinearity.

The preprocessed dataset was then split into a training set (70% of the data) and a testing set (30%) using train_test_split. Stratification based on the target variable IS_SERIOUS_FATAL was applied to ensure similar class proportions in both sets, and random_state=42 was used for reproducibility of the split.

Two supervised learning models were implemented and compared:

1. Decision Tree Classifier: Configured with max_depth=5 to control complexity and prevent overfitting, and class_weight='balanced' to address the class imbalance in accident severity.
2. K-Nearest Neighbors (KNN) Classifier: Implemented with n_neighbors=5 and utilizing n_jobs=-1 to leverage multiple CPU cores for potentially faster computation during the prediction phase.

Both models were integrated into scikit-learn Pipeline objects, which first applied the defined preprocessing steps to the data before fitting the respective classifier. Model performance was evaluated on the unseen test set using overall accuracy, precision, recall, F1-score (particularly for the 'Serious/Fatal' class), and the confusion matrix.

Clustering of Driver Profiles

An unsupervised clustering analysis was conducted to identify distinct driver demographic and risk profiles using K-Means. Features selected from the person dataset for drivers included 'AGE_GROUP', 'SEX', 'ROAD_USER_TYPE_DESC', and 'INJ_LEVEL'. These features were appropriately encoded *(briefly state the final encoding method, e.g., "with 'AGE_GROUP' one-hot encoded and others label encoded")* and then normalized *(state method, e.g., "using StandardScaler")*. The optimal number of clusters was determined using the Elbow Method. Due to initial challenges visualizing clusters with PCA, UMAP was employed for dimensionality reduction, enabling clearer visualization of cluster separation via scatter plots and aiding interpretation through a feature-UMAP component correlation heatmap

Data Exploration and Analysis

The Victorian road crash dataset was explored to identify patterns and factors associated with accident severity. To understand the fundamental distribution of accident outcomes, the overall severity levels were first examined. Figure 1 illustrates the distribution of unique accidents across the defined severity categories.

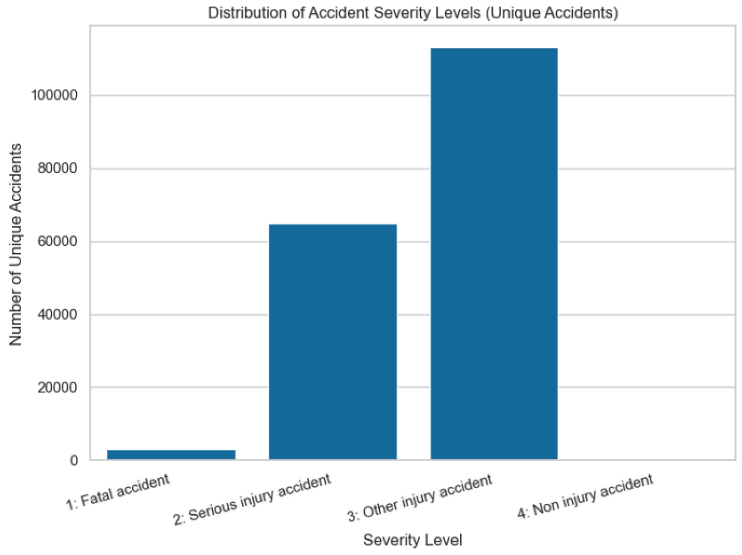

Figure 1.

In figure 1, we see that the vast majority of unique accidents are "Other injury accident" and "Non injury accident". "Serious injury accident" is the next most common, and "Fatal accident" is the least common. This is typical for road crash data. Fatalities are rare events compared to other injury types or non-injury accidents.

The influence of speed zones on accident severity was then investigated. Figure 2 presents the proportional distribution of severity levels for different speed zone groups, based on unique accident occurrences.

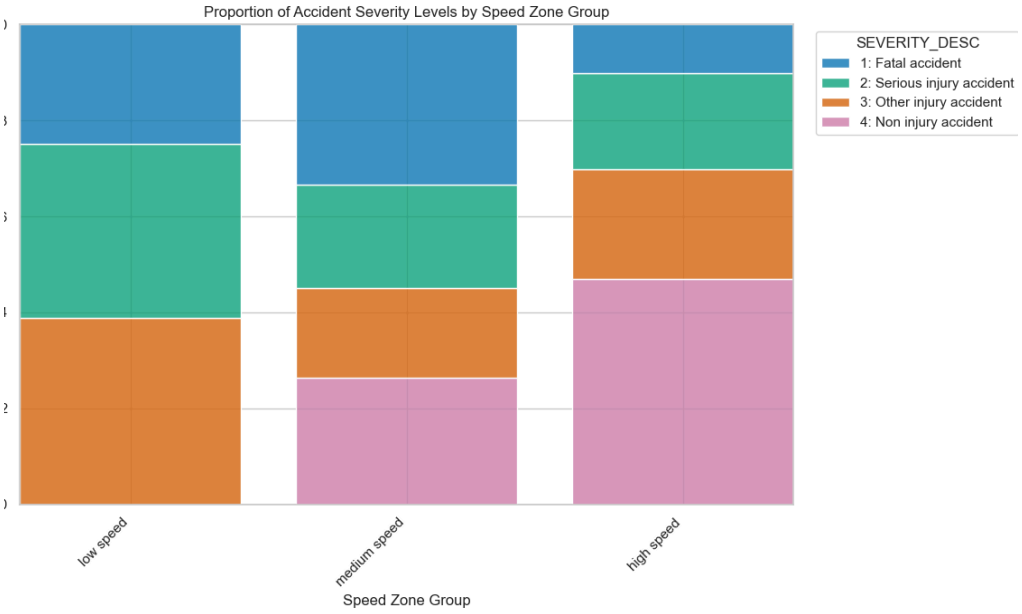

Figure 2.

In figure 2, we observe the following:

- Low Speed: Highest proportion of "Non injury accident" and "Other injury accident". Lowest proportion of "Fatal accident" and "Serious injury accident".
- Medium Speed: The proportion of "Non injury accident" decreases compared to low speed. The proportions of "Fatal" and "Serious injury" accidents start to increase.

- High Speed: Shows the lowest proportion of "Non injury" accidents and the highest proportions of "Fatal accident" and "Serious injury accident". The pink bar (Non-injury) is smallest here, and the combined height of the blue (Fatal) and green (Serious) bars is largest.

We can see that as the speed zone increases, the likelihood of an accident resulting in a fatality or serious injury also increases significantly. Conversely, lower speed zones are associated with a higher proportion of non-injury or other-injury accidents.

The relationship between the age of vehicles involved and accident severity was examined at the unique vehicle level, as shown in Figure 3.

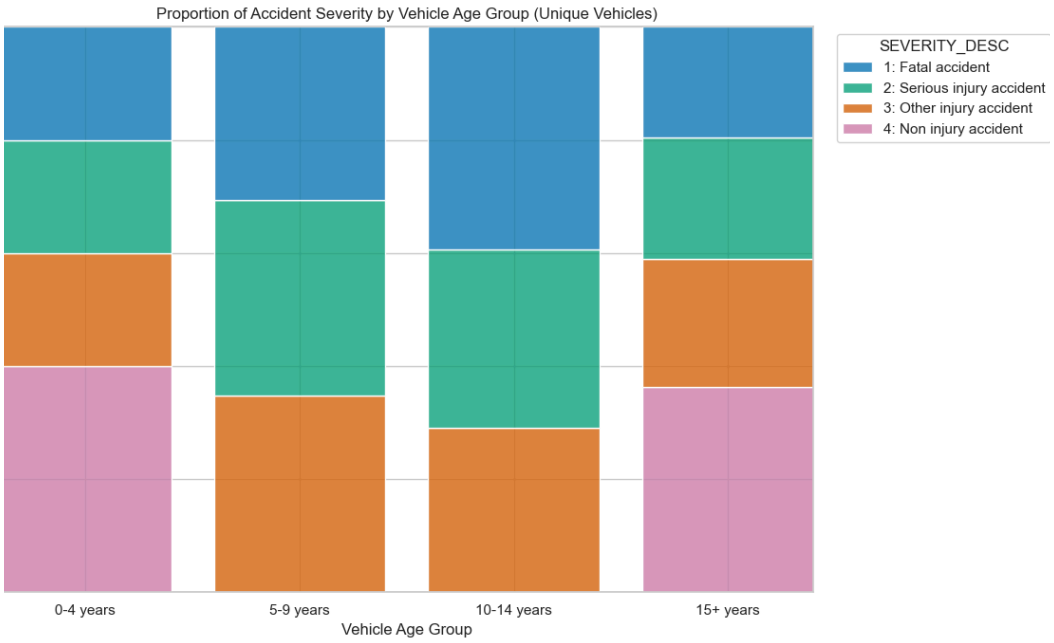

Figure 3.

In figure 3, we observe the following:

- 0-4 years (Newest): Appears to have a relatively high proportion of "Non injury" and "Other injury" accidents, and a lower proportion of "Fatal" and "Serious Injury" compared to the oldest category.
- 5-9 years & 10-14 years: The pattern seems somewhat similar to 0-4 years, though there might be subtle shifts. The "Fatal" and "Serious Injury" proportions might be slightly increasing.

- 15+ years (Oldest): This category seems to have the highest combined proportion of "Fatal" and "Serious Injury" accidents. The pink "Non injury" bar appears smallest here.

There's a potential trend suggesting that older vehicles (15+ years) are involved in proportionally more severe (fatal or serious injury) accidents. Newer vehicles, while still involved in accidents, seem to have a higher share of less severe outcomes. This could be due to better safety features in newer cars or other unknown factors related to drivers of older/newer vehicles.

To understand how severity outcomes differ among various road users, the proportion of injury severity by road user type was analyzed at the person level. Figure 4 displays these findings.

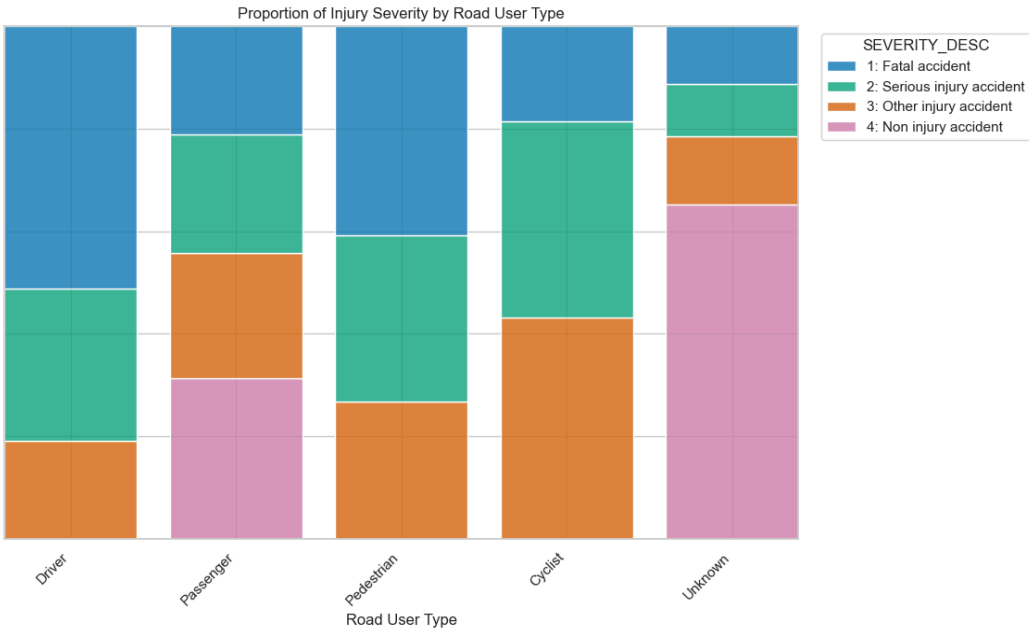

Figure 4.

In figure 4, we observe the following:

- Driver & Passenger: Have a substantial proportion of "Non injury" and "Other injury" accidents. The "Fatal" and "Serious Injury" proportions are present but smaller.
- Pedestrian: Shows a significantly higher proportion of "Fatal accident" and "Serious injury accident" compared to Drivers and Passengers. The "Non injury" bar is very small.
- Cyclist: Similar to Pedestrians, Cyclists also have a very high proportion of "Fatal" and "Serious Injury" accidents, and a very small proportion of "Non injury".

- Unknown: It’s hard for us to interpret what exactly this category represent, but it also shows a high proportion of severe outcomes.

From these observations, we can see that Pedestrians and Cyclists are much more vulnerable road users. When involved in an accident, they are far more likely to suffer fatal or serious injuries compared to vehicle occupants (Drivers/Passengers). This highlights their lack of physical protection.

Finally, the impact of ambient light conditions on accident severity was explored, with results presented in Figure 5 based on unique accident occurrences.

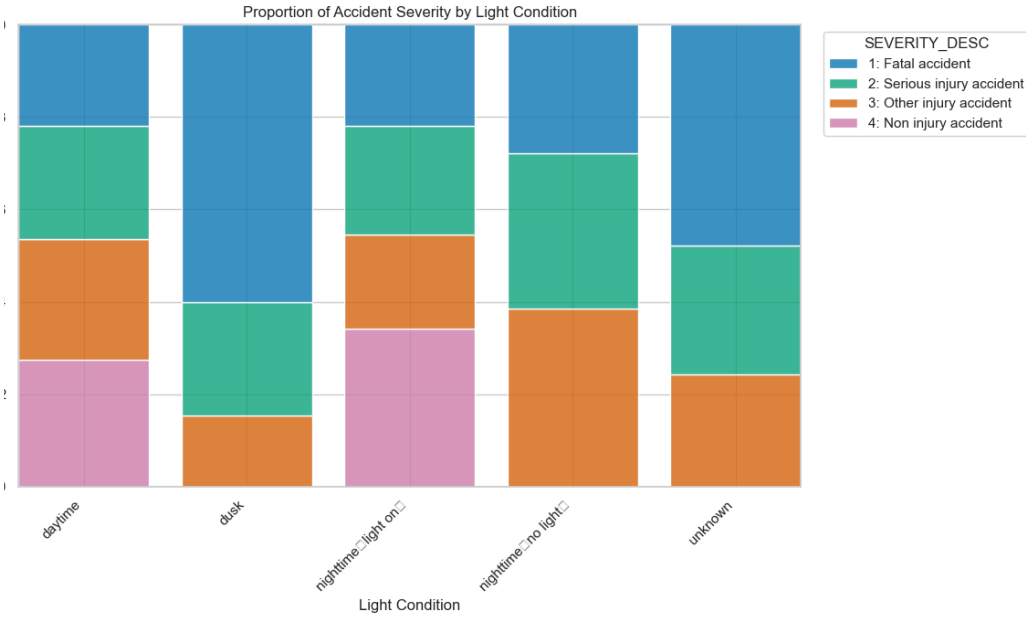

Figure 5.

In figure 5, we observe the following.

- Daytime: Has a large proportion of "Non injury" and "Other injury" accidents. Fatal and serious injuries are present but form a smaller proportion.
- Dusk: The proportion of "Fatal" and "Serious Injury" accidents appears to be slightly higher than in "Daytime".
- Nighttime (light on): Similar to "Dusk", possibly a bit higher proportion of severe outcomes than "Daytime".

- Nighttime (no light): This category shows a noticeably higher proportion of "Fatal" and "Serious Injury" accidents compared to "Daytime" and "Nighttime (light on)". The "Non injury" bar is smaller.
- Unknown: Again, it's hard to interpret what this is, but it also shows a relatively high proportion of severe outcomes.

From these observations we can interpret that accidents occurring in darker conditions, especially "Nighttime (no light)", tend to be more severe. While more accidents might happen in daylight (due to more traffic volume, not shown here), the proportion of those accidents that are fatal or serious increases when visibility is reduced. "Dusk" and "Nighttime (light on)" seem to be intermediate in terms of severity risk compared to full daylight or complete darkness.

Clustering analysis was performed to identify distinct demographic and risk profiles. The Elbow Method (Figure 6) was used to determine the optimal number of clusters.

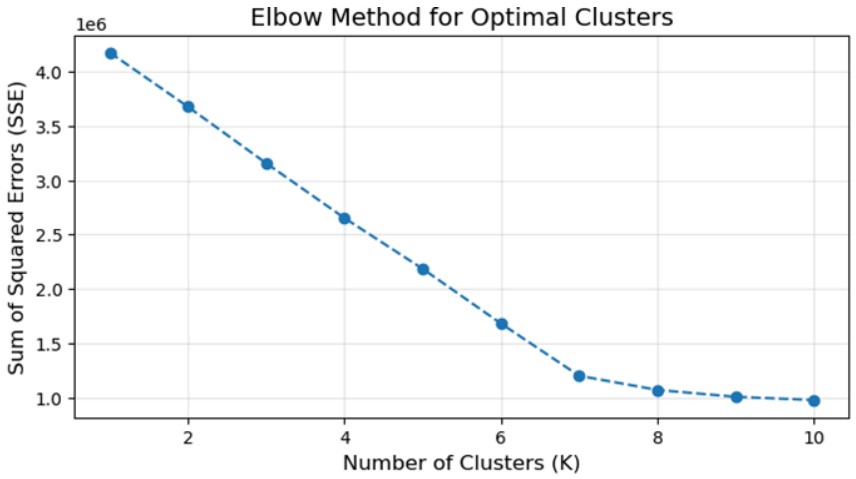

Figure 6.

As shown in Figure 6, the elbow point was identified at k=7... Figure 7 and 8 presents a scatter plot of these clusters.

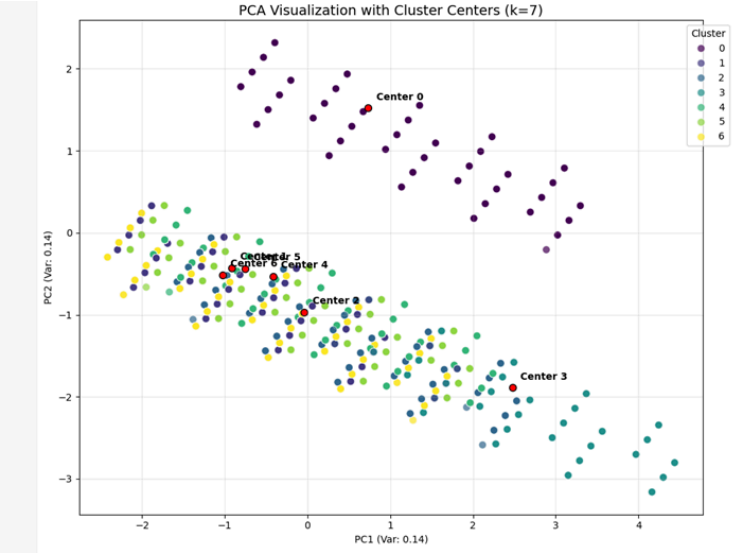

Figure 7.

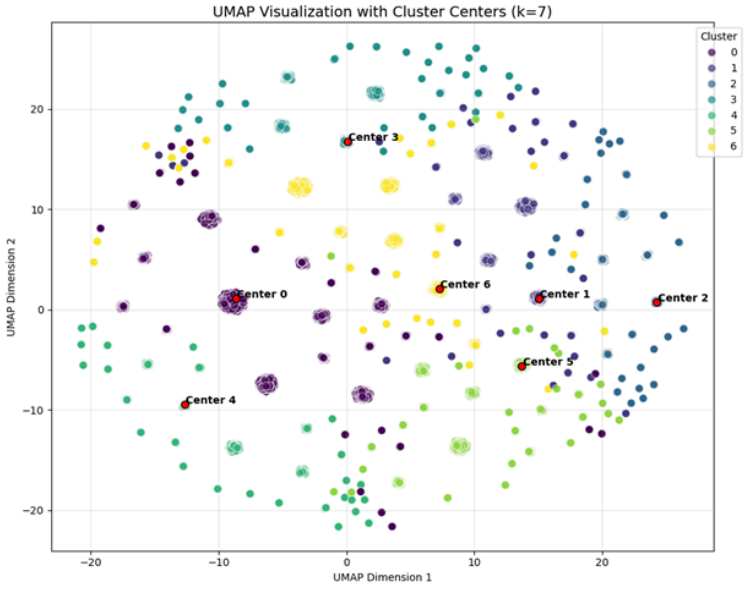

Figure 8.

These seven clusters were visualized using UMAP for dimensionality reduction, showing improved separation over PCA.

A correlation heatmap between original features and UMAP components (Figure 9) indicated that driver age was the primary factor differentiating these clusters. For example, age 18-29 strongly negatively correlated with UMAP1 (-0.55), while age 0-17 (0.53) showed notable positive correlations with UMAP2. Ages 40-49 (0.44) and over 70 (0.50) show positive correlation, ages 50-59

and 60-69 are negatively correlated on both dimensions. Gender and injury level had minimal influence on these cluster formations, while road user type showed a slight correlation.

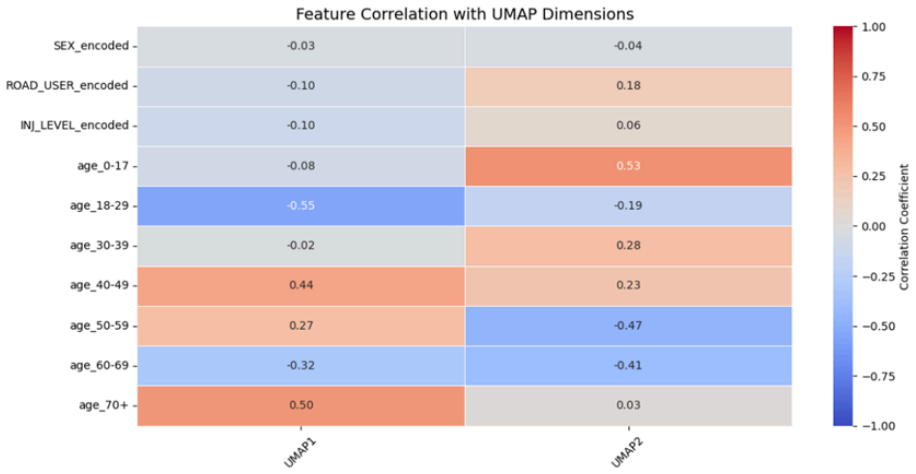

Figure 9.

Later, to this visual exploration, predictive modeling was done to assess the suitability of forecasting severe accident outcomes (Fatal or Serious Injury). A binary target variable, IS_SERIOUS_FATAL, was created, and an accident-level dataset was prepared using features identified as potentially influential, including SPEED_ZONE_GROUP, LIGHT_CONDITION_DESC, DAY_WEEK_DESC, ROAD_GEOMETRY_DESC, and NO_OF_VEHICLES. Two models, a Decision Tree Classifier and a K-Nearest Neighbors (KNN) Classifier, were trained and evaluated.

The Decision Tree Classifier achieved an overall accuracy of 60.97% on the test set. Its performance for the "Serious/Fatal" (Class 1) category was: Precision of 0.4804, Recall of 0.4422, and an F1-Score of 0.4605.

The confusion matrix for the Decision Tree was:

[[23756 9660]

[11263 8930]]

This indicates that the model correctly identified 8,930 serious/fatal accidents but missed 11,263 such incidents.

The K-Nearest Neighbors (KNN) Classifier, with k=5, yielded an overall accuracy of 58.55%. For the "Serious/Fatal" (Class 1) category, its performance was: Precision of 0.4182, Recall of 0.2566, and an F1-Score of 0.3181. The confusion matrix for the KNN model was:

[[26207 7209]

[15011 5182]]

The KNN model correctly identified 5,182 serious/fatal accidents and misclassified 15,011 as not serious/fatal.

Comparing the two models, the Decision Tree Classifier demonstrated better performance in predicting serious/fatal accidents, particularly evident in its higher F1-score and recall for the positive class.

Discussion and Interpretation

This research looked into elements affecting the seriousness of road accidents in Victoria and evaluated how well severe results could be predicted. Important understandings were gained by using visual examination and predictive modeling.

Visual examination (Figures 1-5) revealed that faster speed areas, old vehicles (over 15 years) and vulnerable road users like pedestrians or cyclists, along with low light conditions especially during "Nighttime (no light)", are linked to more fatal and serious injury accidents. These results highlight the obvious effect of greater impact forces, possibly decreased vehicle safety or driver issues in older cars. They also show lack of physical protection for pedestrians/cyclists and reduced visibility leading to severe consequences. Therefore, speed, age of the vehicle, type of road user and lighting appear as important factors influencing severity.

For forecasting severe accidents like serious or fatal ones versus others, we used predictive modeling. This involved a Decision Tree and KNN Classifier with features such as speed zone group and light condition. The results showed that the Decision Tree was more successful, having around 61% accuracy and an F1-score of 0.46 for the Serious/Fatal category compared to KNN which had approximately 58.5% accuracy and a lower F1-score at about 0.32.

Even though the Decision Tree is better, its F1-score shows only moderate prediction ability for serious cases with our current features. A large amount of major/fatal accidents were still wrongly classified (11,263 false negatives), highlighting how hard it is to predict these important but rare events. This implies that while chosen environmental and general accident factors are impactful, more detailed driver and vehicle-specific safety information would probably be necessary for big enhancements in predictive precision.

To conclude, factors such as the speed area, how old a vehicle is, type of road user and whether it's day or night clearly affect the seriousness of road accidents. Although our present prediction models that use only some of these elements show certain ability, they reach just moderate precision in predicting severe/deadly results. These conclusions emphasize the significance of determined risk factors for strategies concerning road safety and pointing out the difficulty involved in accurate severity forecasting.

Limitations and Improvement Opportunities

Though this research gave important understandings, several constraints must be recognized. The analysis depended on the accessible dataset that, even after preprocessing, might have contained natural biases or lacked detail in some variables (like wide age groups and generalized road surface types). Potentially critical elements such as specific driver behaviors (like fatigue and distraction), comprehensive vehicle safety ratings beyond just age, or accurate weather conditions at the moment of impact were not included in the dataset. These could greatly improve our grasp and prediction of severity.

In a methodological way, the predictive models used chose a limited number of features. If we perform more detailed feature engineering and add interaction terms, it may enhance their effectiveness. The models themselves (Decision Tree and KNN) were set up with basic settings; adjusting hyperparameters in detail was not considered at this time but is an obvious path for betterment. Also, although we applied class_weight='balanced', the modest F1-scores for "Serious/Fatal" category imply that dealing with class imbalance is still a problem. It means there is

room to consider using more sophisticated methods. In conclusion, the relationships and visual patterns we have found show connections but not direct causation.

Improvements include adding more comprehensive datasets, investigating advanced machine learning techniques such as Random Forest or Gradient Boosting. Also, strict hyperparameter optimization could be done to improve the predictive precision for severe accident results.

Conclusion

This project investigated the factors influencing road accident severity in Victoria and assessed the predictability of severe outcomes. The analysis revealed that higher speed zones, older vehicle age, vulnerable road user status (pedestrians and cyclists), and adverse light conditions are demonstrably associated with an increased proportion of fatal and serious injury accidents.

Predictive modeling using Decision Tree and K-Nearest Neighbors classifiers indicated that while it is possible to forecast severe accident outcomes to some extent, the accuracy with the current feature set is modest, with the Decision Tree showing comparatively better performance (F1-score of 0.46 for serious/fatal cases). The study underscores the significant impact of specific environmental, vehicle, and road user characteristics on accident severity. While direct prediction remains challenging, these identified factors are crucial for informing targeted road safety interventions and highlight areas where further data enrichment and advanced analytical approaches could yield more impactful results.

Reference

State Government of Victoria. (2024, December 11). *Victoria Road Crash Data - Victorian Government Data Directory*. Dataset - Victorian Government Data Directory. https://discover.data.vic.gov.au/dataset/victoria-road-crash-data