Analysis of Road Environment Factors Affecting Accident Severity

Group W20G7

Yurim CHO	**Ruohan Zhao**	**Leming Lin**	**Jasmine Tao**
COMP20008	COMP20008	COMP20008	COMP20008
yurimc@student.unimelb.edu.au	ruohzhao@student.unimelb.edu.au	scott.lin@student.unimelb.edu.au	jitao1@student.unimelb.edu.au

Executive Summary

This report analyses the effect on accident severity by road environment factors such as road type, intersection type and speed limit. Using data pre-processing, we corrected invalid and missing data and encoding data into categorical data made analysis more effective. Correlation analysis shows a weak linear relationship between individual features and accident severity. This gives the need for multifactor approaches. Supervised learning models –logistic power and decision trees– were used to evaluate predictive power. In the analysis, speed zones and road geometry significantly influence severity but there were imbalance data which give challenge in model performance. Clustering using k-means showed higher accident severity in intersections and high-speed intercity roads. While, residential areas had more frequent but less severe incidents. These findings indicate that combinations of road environments do influence the severity of accidents rather than a single factor in the road environment. This suggests that to prevent accidents, it is crucial to consider various factors of road characteristics. Further study should explore deeply on how these factors are connected to each other and how these connections affect the outcomes.

Introduction

Accidents are a major concern since they often lead to serious injuries, fatalities, and significant disruptions to operations. Understanding of the elements affecting accident severity is required for developing successful safety strategies and preventive measures.

This study investigates the relationship between severity of accidents and important factors of road environment–road type, intersection road type, and speed limit. These variables are thought to be crucial in determining the location and details of accidents as well as the severity of their outcomes.

This analysis involves various stages to analyse the dataset. We start with data pre-processing to ensure quality and consistency of data. Then, using correlation analysis to identify significant relationships between variables. We build and evaluate models that predict accident severity using supervised learning methods. Finally, we use clustering and risk profiling to identify hidden patterns and categorise similar accidents according to risk levels.

Using these predictive and descriptive research techniques, this study intends to deliver practical insights that will support efforts at safety planning and accident prevention.

Methodology

Data Pre-processing

The data used for this research are the information of accident(accidents.csv) and the information of accident location (accident_locaiton.csv). The data was pre-processed for analysis and modelling using methods of data integration, validation, transformation , and encoding.

The data was first merged with an accident dataset and accident location dataset with pivot of accident number. Then required data for research was selected from the merged dataset. The variables of data selected were accident number, road geometry , description of road geometry, severity of accident, speed limit of accident location, road name, road type, name of intersection , and type of intersection. Thus, the dataset could be efficiently used for further processing data and analysis.

Validating data quality of speed zone variable(SPEED_ZONE), showing speed limit of accident location, was processed next. In the dataset, there were invalid data such as 777, 888, 999 which did not fit in the range of 30-110. Handling this data, first replaced invalid data to empty data and then calculated the mode of each road type(ROAD_TYPE) group's speed limit to replace empty data. Invalid data with no mode within the group was imputed using the global mode across all data. This process of validating data quality ensured consistency and reliability of speed limit data(SPEED_ZONE).

Next, transformation of the speed zone variable(SPEED_ZONE) was processed with discretise those continuous values. To enhance analysis efficiency as our research focuses on severity of accident and speed zone, Speed zone variable is categorised with range of 1-4 with 1 to be 30-40 km/h, 2 to be 50-60 km/h, 3 to be 70-80 km/h, 4 to be 90-110 km/h. This category was manually binned since the data of speed limit was fixed with 30, 40, 50, 60, 70, 75, 80, 90, 100, 110 and to analyse the data, it is more efficient to separate manually. Normally, speed zone 30-40 km/h is often at school zone or residential area, speed zone of 50-60 km/h is on the main road of city, 70-80 km/h is applied on the outside of city or road with a few intersections, and the speed zone of 90-110 km/h is normally at highways. After categorising data, it was stored in SPEED_ZONE_CAT. These speed zone types are showing the location of road and types of road, which may help deeply interpreting the relationship between speed zone and severity of accident.

Lastly, the encoding of road type and intersection road type was processed for analysis of relationship between road type, intersection road type and severity of accident. The road type was separated based on the location and characteristics of road type. Rural road types include chase, drive, park, driveway, alley, circuit, way. These road types indicate roads that do not have a high population or connection with residential and commercial areas. Intercity road types include bypass, parkway, freeway, highway, tunnel, throughway. The characteristic of these road types is that these roads have a high speed limit and are for long distance travel. Residential road types include circle, square, cross, row, crescent, bay, place, plaza, court, boulevard, avenue, street, road, and bend. These road types have many pedestrians and bicycle riders since they are in the residential area or in the city. The categorised data was stored in ROAD_TYPE_CAT for road type and ROAD_TYPE_INT_CAT for intersection road type with 0 to be rural road type, 1 to be intercity road type, 2 to be residential road type and 3 to be unknown or special road type.

Correlation Analysis

In this research, Pearson correlation analysis and visualisation tools were applied to examine the correlation between road-related features and accident severity. To support the numerical correlation analysis, three visualisation methods, including boxplots, count plots, and heatmaps, were utilised to illustrate the relationship between important variables for their ability to reveal different patterns.

Under various speed zone categories, the boxplot of accident severity(SEVERITY) can vividly contrast the differences and display the distribution range and concentration of accident severity. Boxplots are effective in revealing the median, spread, and outliers, but they do not reflect frequency or sample size. The countplot was used to compare the frequency of different severity levels across categories. While useful for showing frequency distributions, countplots cannot show statistical dispersion or variability. Heat maps can efficiently demonstrate correlations using colour gradients. The Pearson correlation coefficient intuitively reflects the linear relationship between two variables through the numerical interval from -1 to 1. The correlation can be expeditiously and precisely demonstrated as positive or negative, accompanied by a heat map of the correlation matrix, which strongly improves the readability and interpretability of the results.

Supervised Methods and Evaluation

We used two supervised classification models—logistic regression and decision trees—to train on the dataset to investigate how road types, intersections, and speed zones influence accident severity. We choose logistic regression because the speed seems to have a robust positive correlation with the accident severity. We can easily find if there exists a significant linear impact on accident severity from these features by reviewing the coefficients after training. We chose the decision tree model because it does not require the assumption of a linear relationship between features and labels. The decision tree will determine the impact of features and the nonlinear relation among features by itself. Otherwise, it is easy to visualise and interpret, which makes it very comfortable to analyse the relation between features and labels.

The logistic regression model was trained first. We extracted five features—ROAD_GEOMETRY, ROAD_TYPE_CAT, ROAD_TYPE_INT_CAT, and SPEED_ZONE_CAT—from the pre-processed dataset, and set SEVERITY as the label. Because this experiment aims to investigate the impact of features, we delete the data that has unknown or not normal road type in ROAD_TYPE_CAT and ROAD_TYPE_INT_CAT to avoid these data messing up the relationship. And all these features except SPEED_ZONE are discrete, so we use one-hot to encode them. We then used the sklearn method to split the training and testing datasets with a ratio of 4:1 and keep them at the same stratification. As for model parameters, the default 100 maximum iteration is unsatisfactory in this experiment, so we extend it to 1000. After observing the distribution of the raw dataset, there is a heavy bias between the amounts of these four levels of severity. To balance the amount of data, we used the LogisticRegression parameters to adjust the weight of each label. Finally, we were going to determine the impact by checking the coefficient of each feature.

Next, we trained the decision tree model. We chose our familiar one, ID3, with information entropy, and used the same method to split the dataset and balance it. We added a limitation of a maximum depth of 3 to avoid unlimited tree increases or overfitting. For better interpretation, we drew a picture of the tree and determined the impact of features on accident severity based on the choice of branches.

We decided to use the Classification Report to evaluate these two models. Accuracy cannot reflect the specific conditions of different labels when the dataset is imbalanced. The Classification Report contains precision, recall, and F1-score. Precision is the correct ratio in prediction, and recall is the ratio of objects labelled correctly. The F1-score combines the precision and recall scores to balance these two values well.

Clustering and risk profiling

In order to analyse the relationship between road features and accident risks, we applied k-mean clustering techniques to group the accident data with similar features including road geometry, road type, and speed zones. The goal was to identify if there are any hidden structural patterns that could help to inform road safety strategies.

In our previous data preprocessing, the three key features including road geometry, road type, and speed zones has been categorised into numerical datas, which has made it a lot easier to apply the clustering techniques. The dataset was concentrated into 4 columns including 'ACCIDENT_NO','ROAD_GEOMETRY', 'SPEED_ZONE_CAT', ‘ROAD_TYPE_CAT’, where accident number is used as a unique identifier for each accident.

Before the clustering, one hot encoding was applied to make our data suitable to cluster later. This method transfers categorical features into a binary vector format, which avoids incorrect interpretations of category labels. Then, we evaluated the optimal number of clusters using the elbow method, which examined the sum of squared errors (SSE) for k value from 1 to 10. Once we produce an elbow graph, we are able to identify the elbow which is the optimal k from the graph. After that we have done some visualisation in order to analyse the pattern, and clustering was performed using kmean.

During the clustering implementation, we once considered applying Principal Component Analysis (PCA) to reduce dimensionality to 2D. However, consider that PCA might reduce the clarity of meaning of the clusters, which might not help us to understand the patterns and relationship between the features and accident risk, hence we eventually decided not to use it.

For visualisation, we produced two plot to support the interpretation of the clustering results. The first is a bar chart counting the number of accidents associated with each cluster. This plot gives us an overview of the distribution of the number of accidents across the clusters. Through this plot, we can determine if certain combinations of road types, geometries, and speed zones are more frequently involved in accidents.

Moreover, The second graph is a pie chart presenting the distribution of the 4 severity within each cluster. This visualisation helps to reveal the amount of different severity of accidents in each cluster. By transferring the number of each severity case into percentage and having them in one pie chart, we can easily compare the seriousness and risk profile of each cluster.

Lastly, we exported the frequency counts of each feature (road type, geometry, and speed zone) for every cluster into a CSV file to analyse the characteristics of each cluster. This made it easier to interpret what each cluster represented and helped to identify patterns of the clusters. Additionally, we have also counted the number of each severity within the clusters, which makes it easier to determine the seriousness of accidents involved in each cluster.

Data Exploration and Analysis

4.3 Correlation Analysis

4.3.1 Relationship between speed zone category and accident severity

The analysis of speed zones revealed an unexpected result: accident severity did not vary significantly across different speed categories. One would expect that higher speed zones would result in more severe accidents, but the boxplot visualisation (Figure 4.3.1) did not show a significant difference.

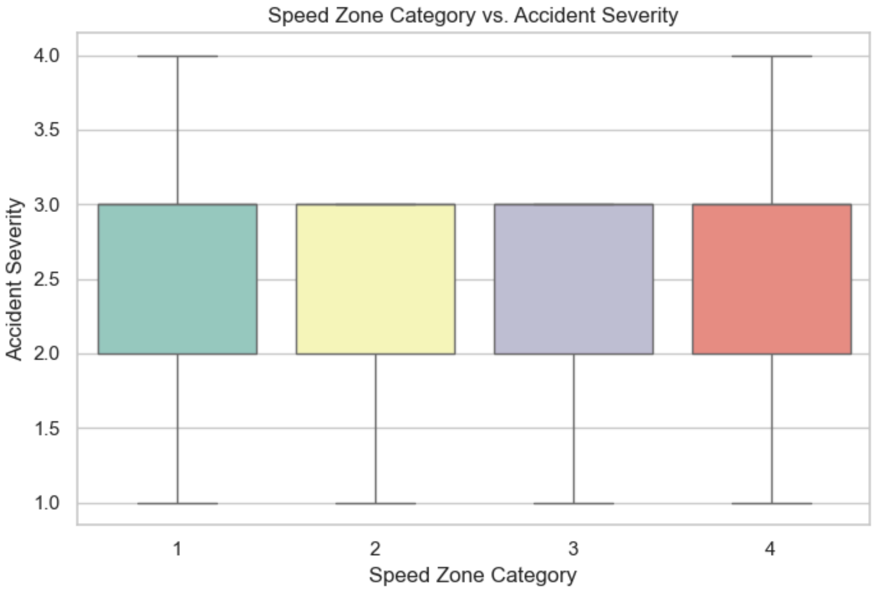


Figure 4.3.1: Speed Zone Category vs. Accident Severity

This may indicate:

- Data Filling Problems: Potential mislabeling or incorrect filling of speed zone values.
- Outlier Influence: Certain outliers might bias the distribution.
- Resolution Limitation: The categorisation might be too broad to capture subtle differences.

Further investigation is required to verify the speed zone data to ensure that it accurately reflects the actual speed limits during accidents.

4.3.2 Distribution of accident severity by road type category

A clear relationship is observed between road types and accident severity by the countplot (Figure 4.3.2). The analysis shows that Category 2 roads have the highest concentration of severe accidents. This category probably refers to significant roads like highways or main city arteries where more serious accidents are caused by higher speeds and traffic density..

On the other hand, there were very few serious accidents in categories 0 and 1, which may indicate a safer low-speed environment. Category 3 has a higher level of severity, which means that in some cases even less frequent roads may be dangerous.

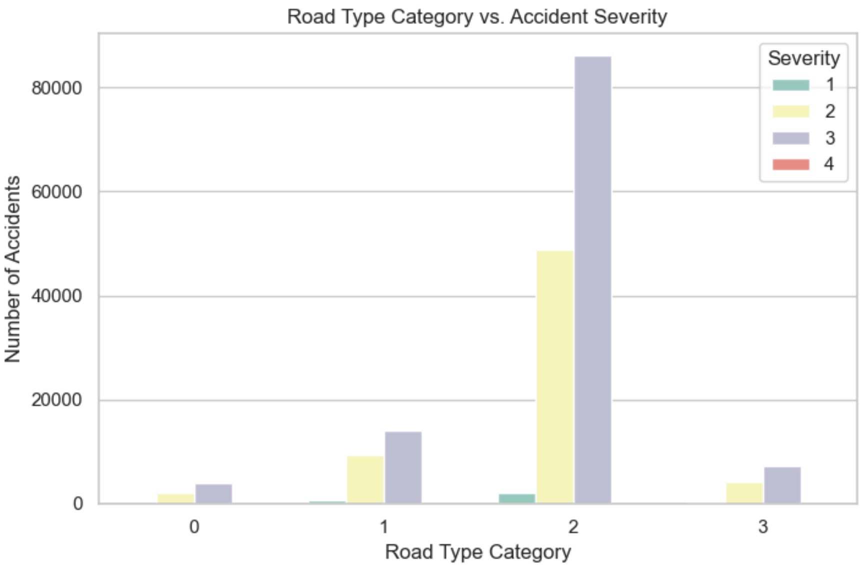


Figure 4.3.2: Road Type Category vs. Accident Severity

4.3.3 Heatmap of road geometry vs accident severity

The heatmap (Figure 4.3.3) analysis of road geometry types against accident severity shows that accidents occurring at intersections (labeled as "At intersection") are significantly more frequent and more severe. This is particularly true for severity levels 2 and 3, which indicate that intersections are critical areas for improving road safety.

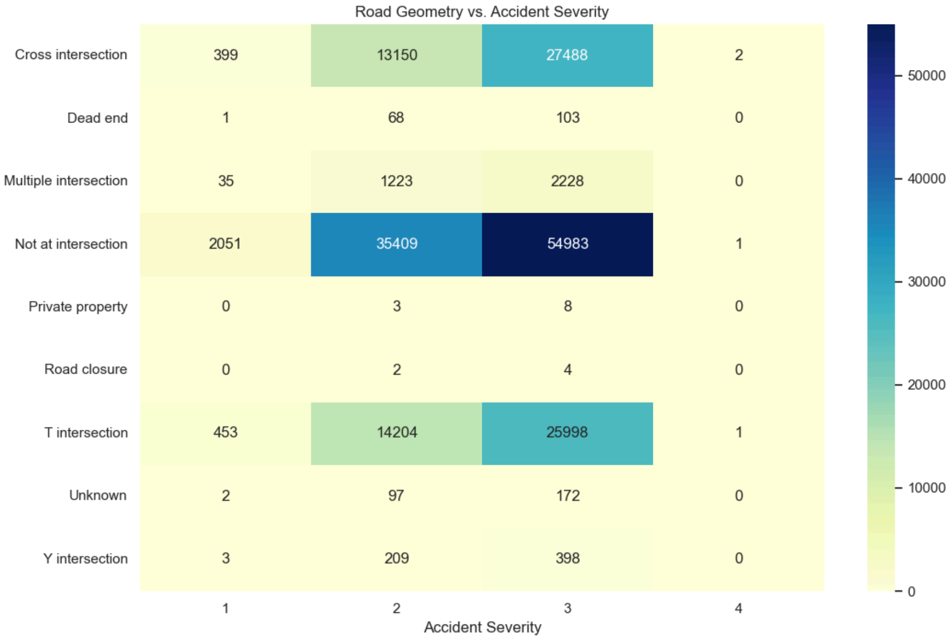


Figure 4.3.3: Road Geometry vs. Accident Severity

The analysis also revealed some unexpected results:

Private Property and Road Closure entries showed minimal, yet existent, accident data. These should be verified for accuracy, as it is uncommon for accidents to be recorded in these scenarios.

Dead End roads also reported some accidents, although very few, which is plausible but rare.

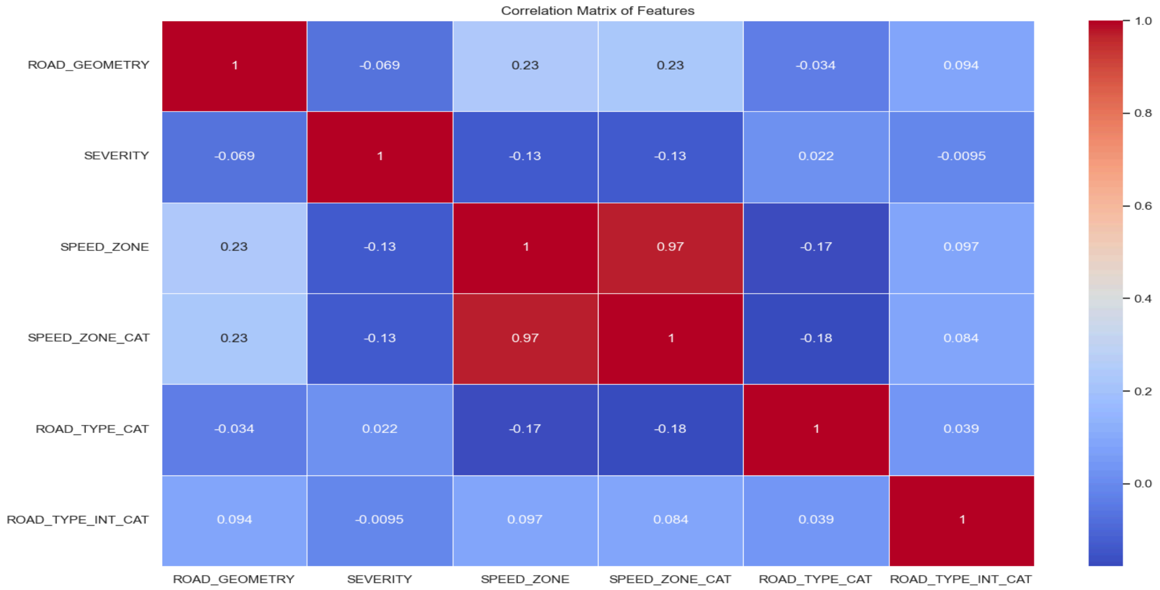
The data suggests that T-intersections and Cross intersections are also high-risk geometries. Safety measures such as improved signage, enhanced visibility, and speed reduction strategies could reduce the severity of accidents in these zones.

4.3.4 Heatmap of correlation matrix of features

The correlation matrix (Figure 4.3.4) illustrates the relationship between the features, and finds that SPEED_ZONE and SPEED_ZONE_CAT are highly correlated.

ROAD_TYPE_CAT and ROAD_TYPE_INT_CAT show low correlation, suggesting these two categories are fairly independent.

Surprisingly, SEVERITY has low correlation with all other features, indicating that single-factor analysis may be insufficient for predicting accident severity.



	ROAD_GEOMETRY	SEVERITY	SPEED_ZONE	SPEED_ZONE_CAT	ROAD_TYPE_CAT	ROAD_TYPE_INT_CAT
ROAD_GEOMETRY	1	-0.069	0.23	0.23	-0.034	0.094
SEVERITY	-0.069	1	-0.13	-0.13	0.022	-0.0095
SPEED_ZONE	0.23	-0.13	1	0.97	-0.17	0.097
SPEED_ZONE_CAT	0.23	-0.13	0.97	1	-0.18	0.084
ROAD_TYPE_CAT	-0.034	0.022	-0.17	-0.18	1	0.039
ROAD_TYPE_INT_CAT	0.094	-0.0095	0.097	0.084	0.039	1

Figure 4.3.4: Correlation Matrix of Features

This low correlation suggests the need for multi-feature analysis: Combining features (e.g., SPEED_ZONE and ROAD_GEOMETRY) may improve predictive power.

4.3.5 Relationship between intersection road type and accident severity

The relationship between intersection road types and accident severity is visualised to understand which types of intersections are more prone to severe accidents. The analysis (Figure 4.3.5) reveals that Category 2 intersections have the highest number of accidents, with a significant portion classified as high severity (Level 3). This suggests that certain types of intersections are high-risk zones that require more safety measures.

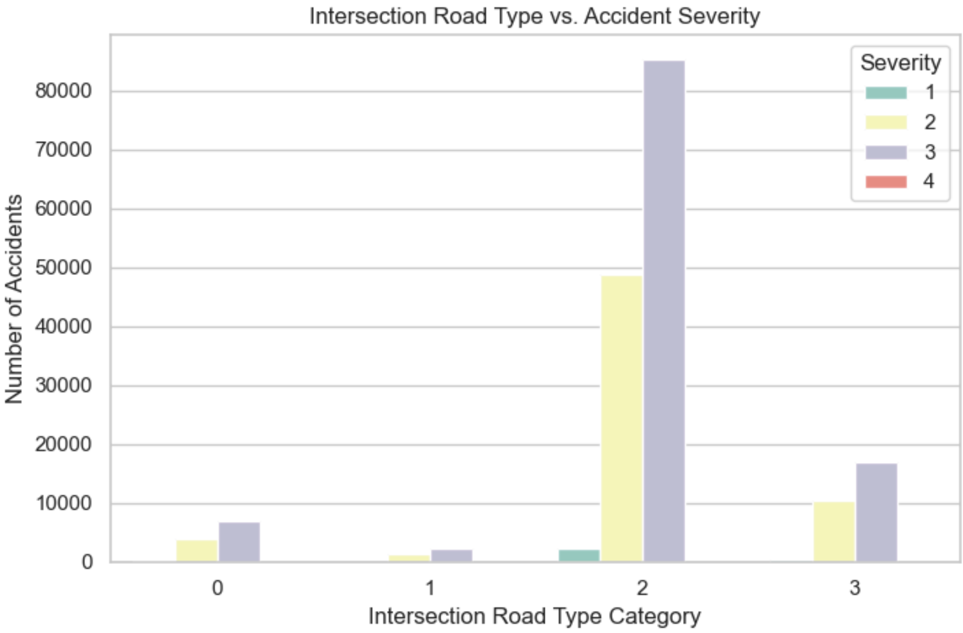
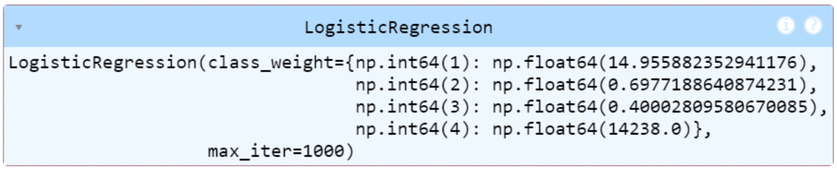


Figure 4.3.5: Intersection Road Type vs. Accident Severity

The findings suggest that Category 2, which likely represents major highways or heavily trafficked urban intersections, is disproportionately represented in severe accidents. This demonstrates the need for enhanced safety features, such as better traffic signals and more controlled intersections in these areas.

4.4 Supervised Methods and Evaluation

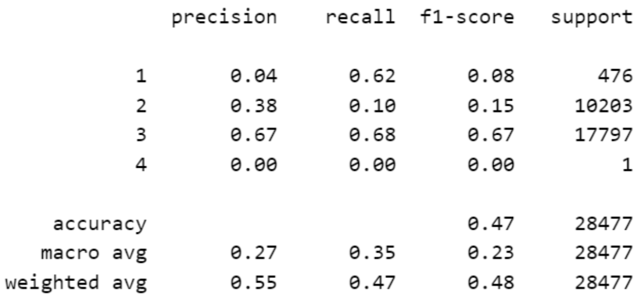
Logistic Regression



```
LogisticRegression(class_weight={np.int64(1): np.float64(14.955882352941176),
                                 np.int64(2): np.float64(0.6977188640874231),
                                 np.int64(3): np.float64(0.40002809580670085),
                                 np.int64(4): np.float64(14238.0)},
                   max_iter=1000)
```

Figure 4.4.1 Logistic regression model

The first thing we noticed was the weight of different labels in Figure 4.4.1. Due to the bias, levels 2 and 3's weights are below 1, while level 1's is about 15, and the most abnormal one is level 4 with 14238. Then, to the evaluation part of the model.



```
              precision    recall  f1-score   support

           1       0.04      0.62      0.08       476
           2       0.38      0.10      0.15     10203
           3       0.67      0.68      0.67     17797
           4       0.00      0.00      0.00         1

    accuracy                           0.47     28477
   macro avg       0.27      0.35      0.23     28477
weighted avg       0.55      0.47      0.48     28477
```

Figure 4.4.2 Logistic regression classification report

From the classification in Figure 4.4.2, it is easy to see that the model performs poorly in level 4, with an F1-score of 0. This is because of the extreme bias in the dataset; even a weight balance cannot save this label. So we put the main focus on levels 1, 2 and 3. Level 3 has the most data and the best performance here. The F1-score is 0.67, and both precision and recall reach a high value of around 0.68. Level 2 got 0.38 on precision while only got 0.10 on recall. That means the model is hard to determine level 2. In contrast, the level 1 has a very low precision with 0.04 and a high recall with 0.62, which means the model misclassifies a large number of other classes of samples as level 1. The reason for the result we think there may be some nonlinear relationships between features and labels.

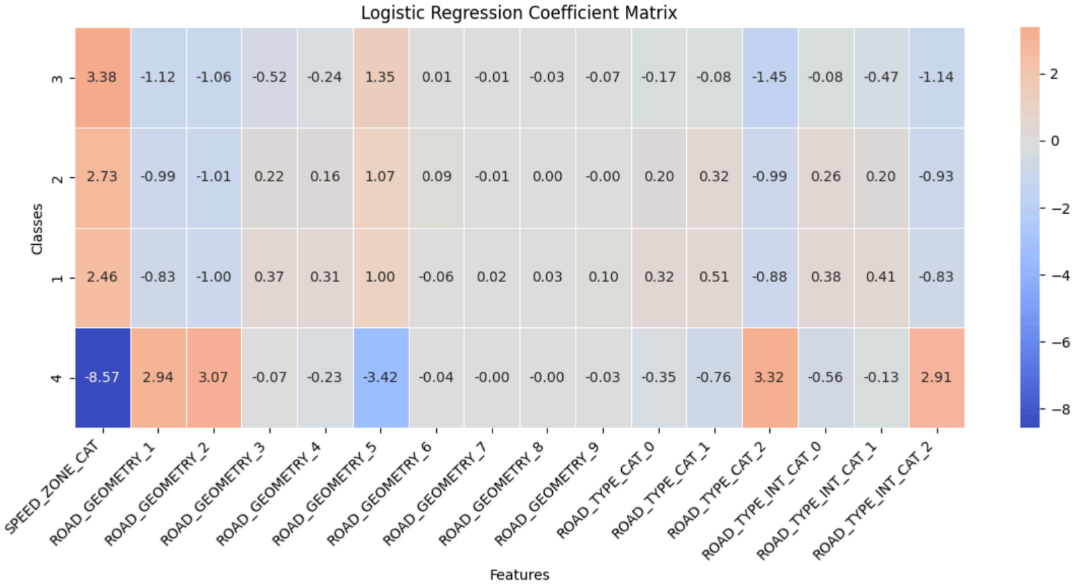
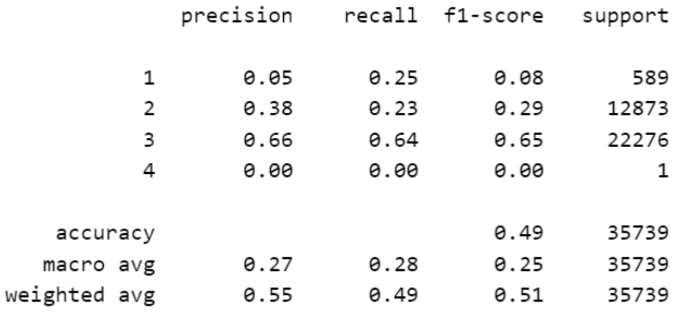


Figure 4.4.3: Logistic regression coefficient matrix

Next, in Figure 4.4.3, we turn to the coefficient matrix of the model. As above, we only look at levels 1, 2, and 3. Speed plays an important role in all three levels, with the highest coefficient values. The second one is geometry 5, which corresponds to the not-at-intersection. From there, the result shows that the speed zone has a great impact on accident severity.



```
              precision    recall  f1-score   support

           1       0.05      0.25      0.08       589
           2       0.38      0.23      0.29     12873
           3       0.66      0.64      0.65     22276
           4       0.00      0.00      0.00         1

    accuracy                           0.49     35739
   macro avg       0.27      0.28      0.25     35739
weighted avg       0.55      0.49      0.51     35739
```

Figure 4.4.4 decision tree classification report

In the decision tree shown in Figure 4.4.4,   the result is better, with higher accuracy and less training time. It has a similar result to level 3 with LR one. The main difference between them is the level 2. In DT, the F1-score and recall value have a significant improvement, which means the decision tree is more suitable for the project with the nonlinear relationship.

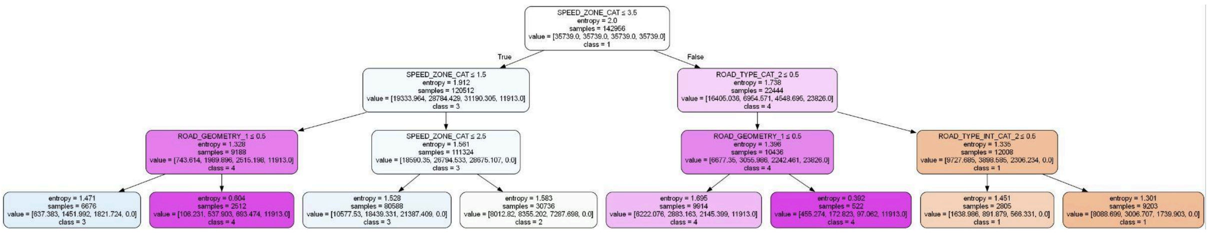


Figure 4.4.5 decision tree structure

This is the decision tree structure shown in Figure 4.4.5, we can find that the most important thing is the speed zones 3 and 4. However, through the analysis, we find that the amount of data affects the decision tree. Most of the final label is level 3. So maybe we need to balance the amount or use another value to replace the information gain.

Limitations and future improvement

Through the experiment, we find some limitations for these two models:

- Logistic regression can only learn linear decision boundaries, for nonlinear problem, maybe use some method transfer nonlinear to linear.
- Decision tree choose split point based on local optimum.
- Decision tree may influenced by amount of data in information gain.
- Decision tree can have more branches than two of each node for a better split with multiclass features.
- Both tend to prefer features with more classes or larger ranges.

4.5 Clustering and risk profile analysis

By applying the Elbow method, we found the optimal k to be 5, and we clustered the data into 5 clusters. To create a risk profile for the clusters, we have produced a CSV file which summarises the accident counts for each feature within each cluster. This allows us to determine what each cluster might represent and if there is a trend or pattern on the high risk road types.

By applying the Elbow method, we found the optimal k to be 5, and we clustered the data into 5 clusters. To create a risk profile for the clusters, we have produced a CSV file which summarises the accident counts for each feature within each cluster. This allows us to determine what each cluster might represent and if there is a trend or pattern on the high risk road types.

Table 4.5.1: table shows the data summary and accident count for each feature in cluster 0

Cluster	**Category**	**data**	**Count**
0	SPEED_ZONE	90-110	25182
0	SPEED_ZONE	30-40	272
0	ROAD_TYPE	residential_road_type	12340
0	ROAD_TYPE	intercity_road_type	11000
0	ROAD_TYPE	Other	1964
0	ROAD_TYPE	rural_road_type	150
0	ROAD_GEOMETRY	not at intersection	23605
0	ROAD_GEOMETRY	T-intersection	822
0	ROAD_GEOMETRY	Cross intersection	655
0	ROAD_GEOMETRY	multiple intersection	185
0	ROAD_GEOMETRY	Y-intersection	117
0	ROAD_GEOMETRY	Unknown	53
0	ROAD_GEOMETRY	Dead end	11

In Cluster 0, it focuses on a higher speed zone in residential and intercity road types with various intersection types but mainly focuses on the roads that is not at an intersection. The speed zones in cluster 0 are mostly ranging from 90-110, with a small amount of 30-40 speed zone accidents included. This cluster might represent high-speed roads in intercity areas and main traffic road in the city where usually have a higher speed limit and a less complicated road geometry (straight road).

Table 4.5.2: table shows the data summary and accident count for each feature in cluster 1

Cluster	Category	data	Count
1	SPEED_ZONE	50-60	26642
1	SPEED_ZONE	70-80	9119
1	SPEED_ZONE	30-40	2432
1	SPEED_ZONE	90-110	1641
1	ROAD_TYPE	residential_road_type	32838
1	ROAD_TYPE	intercity_road_type	2948
1	ROAD_TYPE	Other	2121
1	ROAD_TYPE	rural_road_type	1927
1	ROAD_GEOMETRY	T-intersection	39834

Cluster 1 concentrates on the roads that has “T-intersection”, with the speed zones and road types vary. However, it is shown that cluster 1 has speed zones mainly range from 50-80, which is considered to be some common, moderate speed zones. This might suggest that Cluster 1 represents the roads with T intersection across all areas.

Table 4.5.3: table shows the data summary and accident count for each feature in cluster 2

Cluster	Category	data	Count
2	SPEED_ZONE	50-60	29668
2	SPEED_ZONE	30-40	3122
2	SPEED_ZONE	90-110	1310
2	ROAD_TYPE	residential_road_type	30482
2	ROAD_TYPE	Other	1598
2	ROAD_TYPE	intercity_road_type	1155
2	ROAD_TYPE	rural_road_type	865
2	ROAD_GEOMETRY	Cross intersection	31543
2	ROAD_GEOMETRY	multiple intersection	2018
2	ROAD_GEOMETRY	Y-intersection	296

2	ROAD_GEOMETRY	Unknown	125
2	ROAD_GEOMETRY	Dead end	109
2	ROAD_GEOMETRY	Private property	7
2	ROAD_GEOMETRY	Road closure	2

Cluster 2 represents complicated intersections including cross intersection, Y intersection and multiple intersection, with speed limits mostly below 70 km/h and in residential areas. This indicates that this cluster might represent the complex traffic areas which have intersections.

Table 4.5.4: table shows the data summary and accident count for each feature in cluster 3

Cluster	Category	data	Count
3	SPEED_ZONE	70-80	29453
3	SPEED_ZONE	30-40	5589
3	ROAD_TYPE	residential_road_type	24667
3	ROAD_TYPE	intercity_road_type	7748
3	ROAD_TYPE	Other	1968
3	ROAD_TYPE	rural_road_type	659
3	ROAD_GEOMETRY	not at intersection	24682
3	ROAD_GEOMETRY	Cross intersection	8841
3	ROAD_GEOMETRY	multiple intersection	1283
3	ROAD_GEOMETRY	Y-intersection	149
3	ROAD_GEOMETRY	Unknown	66
3	ROAD_GEOMETRY	Dead end	20
3	ROAD_GEOMETRY	Road closure	1

Cluster 3 indicates a number of accidents from mainly residential roads, which have a moderate high speed limit and various road geometry. This could mean that cluster 3 represents residential zones or connecter roads of intercity and residential with a moderately high speed limit.

Table 4.5.5: table shows the data summary and accident count for each feature in cluster 4

Cluster	Category	data	Count
4	SPEED_ZONE	50-60	44265
4	ROAD_TYPE	residential_road_type	36505
4	ROAD_TYPE	Other	4036
4	ROAD_TYPE	rural_road_type	2516
4	ROAD_TYPE	intercity_road_type	1208
4	ROAD_GEOMETRY	not at intersection	44157
4	ROAD_GEOMETRY	Y-intersection	48

4	ROAD_GEOMETRY	Dead end	32
4	ROAD_GEOMETRY	Unknown	27
4	ROAD_GEOMETRY	Private property	1

Cluster 4 mainly includes low-speed roads especially 50-60 km/h speed zones and a mix of various road types, especially residential. The intersection types are mostly not at an intersection. It is considered that this cluster represent the common residential areas, where traffic is at moderate speed and have a less complicated traffic condition and intersections.

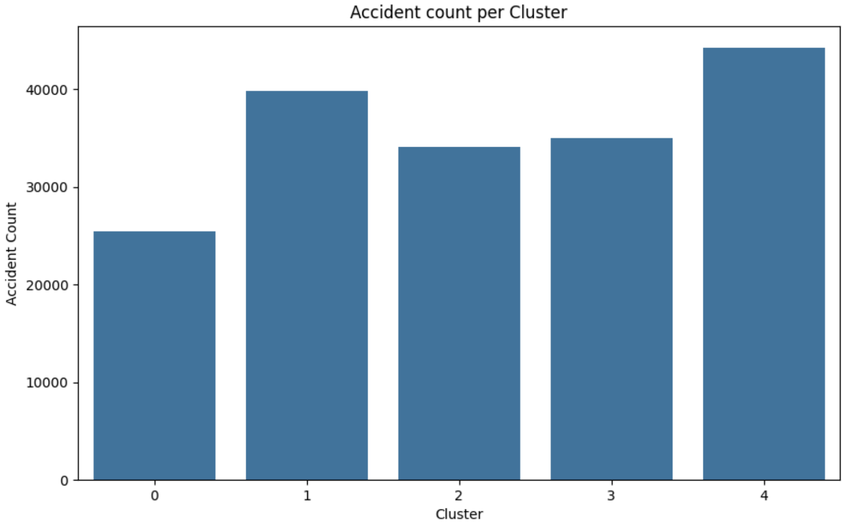


Figure 4.5.1 Accident count per cluster bar chart

As shown in Figure 4.5.1, the bar chart shows the accident count for each cluster. With cluster 4 having the highest accident number and Cluster 0 with a slightly low number of accident compare to the others. Cluster 0 who have the lowest number of accidents represents a more simple road geometry compare to the others, while cluster 1,2,3 all represents a more complex road geometry. Cluster 4 also represent simple road geometry with moderate speed limits, however, the high accident count might be due to the higher dataset size for cluster 4, as cluster 4 represent roads that are more commonly used.

Pie chart showing Severity distribution per cluster

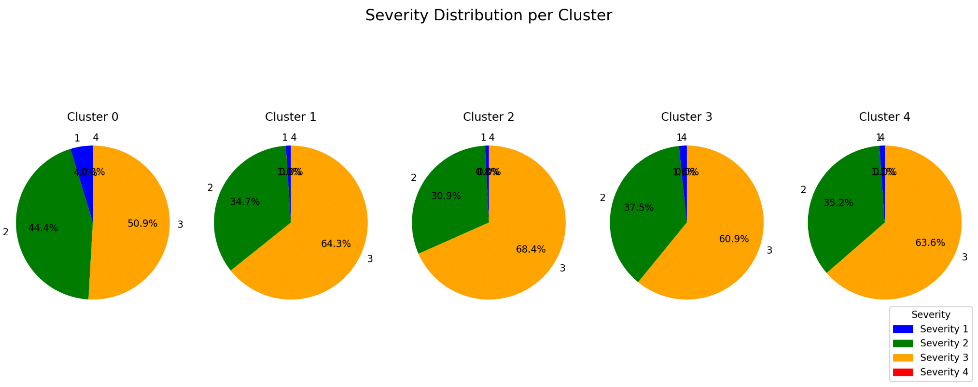


Figure 4.5.2 Severity distribution per cluster pie chart

In the severity distribution pie chart (figure 4.5.2), Cluster 0 shows a higher proportion of low severity accidents and fewer high-severity cases. This suggests that accidents in Cluster 0 tend to be less severe than the other clusters

Clusters 1, 2, and 4 all have a large proportion of severity level 3 accidents, while the number of level 1 cases is low. This may indicate that the road types in these clusters are more likely to be involved in serious accidents.

Cluster 3 also has a similar number of severity level 3 cases with cluster 1, 2 and 4, but its distribution appears slightly more balanced. It contains a relatively higher proportion of low-severity cases, which might suggest that serious accidents are less dominant in this cluster.

Overall, Cluster 4, which represents moderate speed limit, not at intersection roads, has the highest number of accidents and a large proportion of severe cases. Less complicated road geometry are less involved in accidents and have a lower risk of having severe accidents.

Limitations and future improvement

This clustering and rick profile task still have several limitations, and future improvement could be done, which could include:

- Some tests could be done to see if the k value is optimal and the clusters are separated.
- Including more features to the clustering process to ensure that each cluster can be more independent to each other
- The pie chart(figure 4.2) can be improved by separating the percentages for the low proportion, this could be writing a function to move the cluster 1 or cluster 4 percentage closer to the centre, which helps to separate the percentage from overlapping.

Discussion and Interpretation

In this project we focused on researching the question : What is the impact of road types, intersections, and speed zones on accident risk? In order to analyse the dataset, the tasks have been splitted into 4 parts: data preprocessing, correlation, supervised methods and clustering and risk profiling.

The analysis reveals that the road types and road geometry are more related to accident number and severity than speed zones. In both supervised model and clustering tasks, it is shown that although “not at intersection” roads are involved frequently in accidents, the severity is generally low, suggesting that simple road geometries have a lower risk of having severe accidents, and at the same time, complicated intersections and residential areas leads to more severe and higher amount of accidents. Furthermore, in correlation and clustering, it is suggested that the relationship between speed zones and accident risk is weak. Even though in the supervised model, it indicates that higher speed zones are related to more severity level 4 cases, since the severity level 4 has a limited number of cases, it may lead to biases and cause extreme results.

The combined findings indicate that road geometry and road types are highly related to accident risks. Roads that have complicated intersections are more likely to be involved in severe accidents. This highlights the need to enhance awareness and increase the safety features such as traffic signals and better controlled intersections in these areas. To enhance safety, some improvements can be:

1. Improving safety measures at T-intersections and major highways.
2. Refining speed zone categorisation and validating data accuracy.
3. Applying advanced machine learning techniques for multi-feature analysis to predict accident severity more accurately.

Conclusion

In conclusion, this research explored the impact of road types, intersections, and speed zones on accident risk and accident severity. Through data cleaning, data preprocessing, correlation analysis, supervised learning model and clustering, we found that accident risk and severity is related to both road geometry and road type, but less related to speed zones. Our findings highlight that complex intersections are associated with a higher risk of severe cases, while moderate speed limit residential roads are more frequently involved in accidents but less severe ones.