What is the impact of road surface type, road surface condition, intersection type, and speed zones on accident severity?

By Vanesa Reategui Gutierrez

Executive summary

This report investigates the impact of road types, road conditions, intersection types, and speed zones on accident severity. Relevant data was used from the following datasets provided by **VicRoads**: Accidents, Vehicle and Road Surface Condition. The data was prepared by managing anomalies in speed zone data, handling missing values, removing duplicates and one-hot encoding categorical variables. The relationship between the pre-crash variables and crash severity were investigated using correlation analysis and two different supervised learning models. The correlation analysis primarily consisted of using mutual information to explore non-linear associations between road features and severity levels. The analysis revealed that most individual features had weak associations, however there were some notable patterns identified. Visualizations such as stacked bar charts and distributions graphs displayed the trends. Supervised learning models were developed to predict whether a vehicle's damage would be severe given road surface, road surface condition, intersection type, and speed. The models also revealed the extent to which each variable contributed to predicting accident severity. Overall, the report demonstrates that speed zones have the most important impact on severity, while road surface condition and certain intersection types also have notable impact.

Introduction

This report explores the relationship between pre-crash metrics and accident severity, through the analysis of data provided by **VicRoads**. The intention of our report is to identify whether road surface type, road surface condition, intersection type, and speed zones have a tangible impact on the severity of an accident. Hence, the investigation aims to highlight the most pertinent pre-crash data, with outcomes that could have potential use in increasing road safety.

To investigate the factors contributing to road accident severity, we explored the correlation between our pre-crash conditions and crash outcomes. Our analysis focused on investigating how road surface type, surface condition, intersection geometry, and speed zones relate to specifically to the crash severity index provided in the Accident dataset.

We assessed these relationships by using correlation techniques such as Mutual Information, measuring the dependency between categorical and numerical features. In addition to individual feature analysis, we introduced combined features, such as road surface and intersection type, to reveal higher risk conditions that may not be evident in isolation. These insights are intended to enhance understanding of how different road factors interact, which could be used to increase overall road safety.

We also used two different supervised learning models to determine whether selected pre-crash conditions could assist in predicting of accident severity, as measured by vehicle damage. Our objective was to determine whether a meaningful relationship exists between these specific conditions and the level of vehicle damage in a crash. A *k*-Nearest-Neighbours (KNN) algorithm and a Random Forest algorithm were used to model these relationships. By applying two models, we increased the likelihood of building an accurate model and introduced the opportunity to compare their results to further ascertain any connection.

Both models used speed zone, intersection type, road surface type and condition as pre-crash indicators per our overall investigation. We chose to use vehicle damage, represented by the most damaged vehicle involved in the accident, as an indicator of accident severity ahead of injuries or fatalities, due to the limitations in the dataset. Fatalities were too severe and rare to be of use in training, while injuries were too common and not classified specifically enough. Although vehicle damage is not a reliable metric in classifying accident severity to its occupants, it is still important in overall severity and could have potential applications in road safety or for insurance purposes.

Methodology

Data-preprocessing

The Vehicle, Accident, and Road Surface Condition datasets (sourced from the VicRoads website) were used throughout the investigation, as they contained the features most relevant pre-crash features and severity metrics. These datasets were then merged using each accident's unique number as a key identifier. The descriptive variables such as road surface condition, road surface type and road geometry were handled using one hot encoding to facilitate statistical analysis.

Each task filtered the new merged dataset according to their specific objectives and requirements, retaining only the relevant features. Because most accidents involve more than one vehicle, the vehicle dataset possesses a many-to-one relationship with the accident dataset. This did not affect the road surface condition and therefore our correlation analysis tasks did not need to explicitly handle this. For the supervised learning models, only the most damaged vehicle was included in the dataset as we considered this appropriately represented the severity of the crash.

Some of our correlation analysis involved more specific data preprocessing, specifically we:

- Created new data columns representing each road type under wet and dry conditions to more accurately calculate severity rates and mutual information (MI).
- Generated combination columns for road surface type and intersection geometry; combinations with fewer than 30 occurrences were filtered out to ensure statistical reliability.
- Extracted TRAFFIC_CONTROL_DESC from filtered_vehicle.csv and created a new DataFrame to examine the relationship between traffic control type, speed zone, and severity.

For the supervised learning task, the vehicle damage index was further consolidated into two categories: Severe (rating 4 and 5) and Not Severe (rating 6 (no damage), 1, 2 and 3), excluding unknown values. Reconsidering the rating as a binary classification allowed for clearer separation between high-risk crashes and others. It also improved model interpretability and consistency across evaluation metrics. Additionally, the data was filtered to only include valid Victorian speed limits (30km/h-110km/h), removing entries containing an unknown invalid speed limit.

As the KNN algorithm uses distance calculations, we normalised the data to ensure equal weighting. Normalisation was not required for the random forest model as it inherently insensitive to the scale of the input features (Scikit-learn developers, 2024).

To ensure accuracy and consistency in analysis, all data processing and calculations were conducted using Python libraries. *Pandas* was used for data preparation and manipulation, whilst *scikit-learn* was employed for computing mutual information and implementing supervised learning models.

Correlation Analysis

This report employed Mutual Information (MI) and Normalized Mutual Information (NMI) to investigate the relationship between categorical variables. These metrics allowed us to quantify the correlation between road features and severity of road accidents. MI and NMI were chosen as they allow for the identification of linear and non-linear dependencies, which is ideal for the data used. MI and NMI identify which road categories contribute most to severe road accidents, providing insight that supports the report's analysis.

Supervised Learning Model Implementation

The KNN and Random Forest models were chosen to serve different purposes. KNN was implemented a more basic model to act as a baseline for how distance-based supervised learning algorithms would perform. The algorithm classifies inputs based on the most common label among their *k* nearest neighbours, using Euclidean distance, as such we normalised the features to ensure equal weighting. We used *scikit-learn*'s implementation of KNN and selected an 85/15 train-test split to maximise model training, which due to the magnitude of the dataset it still left us a sufficiently large test set. A K value of 5 was found to be optimal during training.

Random Forest was implemented to be the higher performing model and to be compared with KNN. Random Forest works by building an ensemble of decision trees, training them with a random subset of the data then aggregating their predictions to decrease variance and reduce overfitting which is typically found with standard decision trees (Scikit-learn developers, 2024). The model was implemented using *scikit-learn*'s *RandomForestClassifier*, with the same train/test split as the KNN model. Random Forest was also chosen for its ability to provide insight into feature performance. By accessing the *feature_importances_* array, we were able to isolate the most influential features in the model training which provided further insight to the overarching report question.

To evaluate the models, confusion matrices were used to visually compare performance. In addition to this, we produced classification reports which allowed for direct comparison of F1-score, accuracy, and precision.

Data Exploration and Analysis

Primary Correlation Analysis

This analysis draws on mutual information scores and visualizations to reveal patterns across speed zones and intersection types. The severity variable consists within a range from 1 to 4, where 1 represents the most severe and 4 the least severe.
The relationships explored in our primary correlation analysis were:

- Intersection Type (road geometry) and severity
- Speed Zone and severity

The bar graph (Figure 1) presents the Normalized Mutual Information (NMI) values for Road Geometry and Speed Zone in relation to accident severity. Both features show weak associations, with NMI scores of 0.003 and 0.009, respectively. However, Speed Zone demonstrates a stronger relationship with severity compared to Road Geometry. This suggests that speed limits may play a greater role in influencing accident severity than the type of intersection or road layout. Nevertheless, the low NMI values indicate that neither feature independently explains much of the variation in severity, highlighting the complexity of factors contributing to crash outcomes.

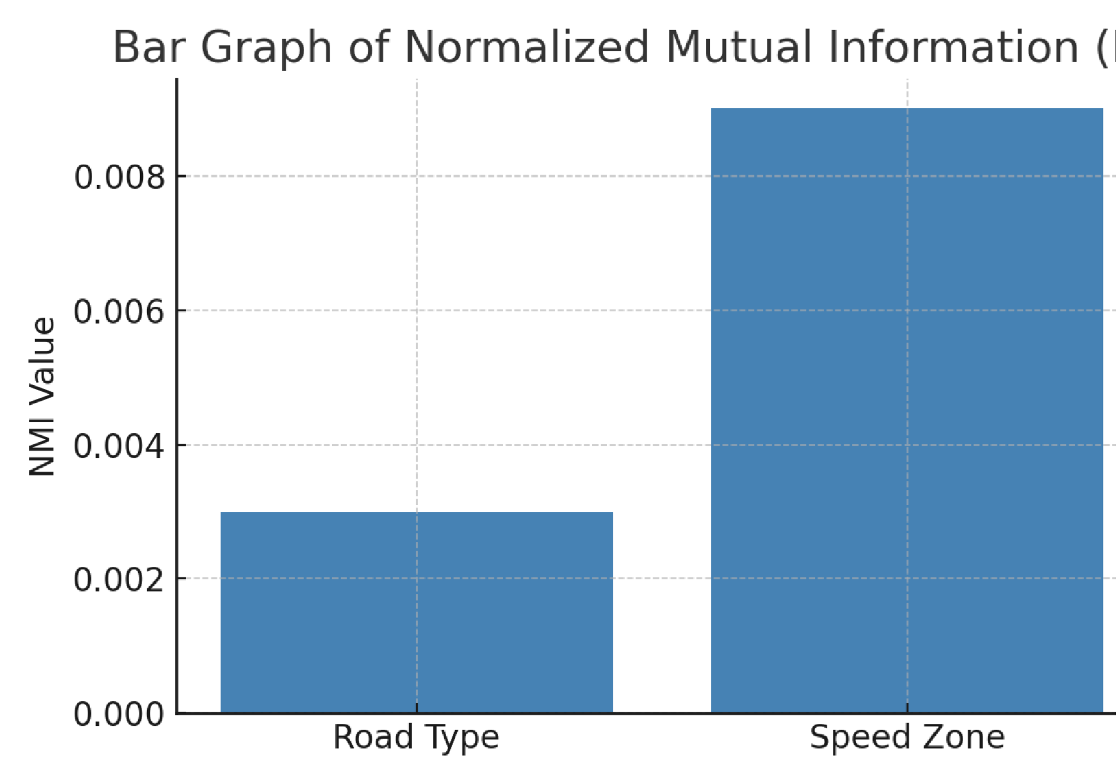


Figure 1: Bar graph displaying NMI values for intersection type and speed zone

The MI analysis (Table 1) of accident severity and intersection type indicates that accidents which occur not at intersection have the strongest association with severity, with an MI score of 0.014. This is noticeably higher than any other intersection. Cross intersection shows the next highest association (MI

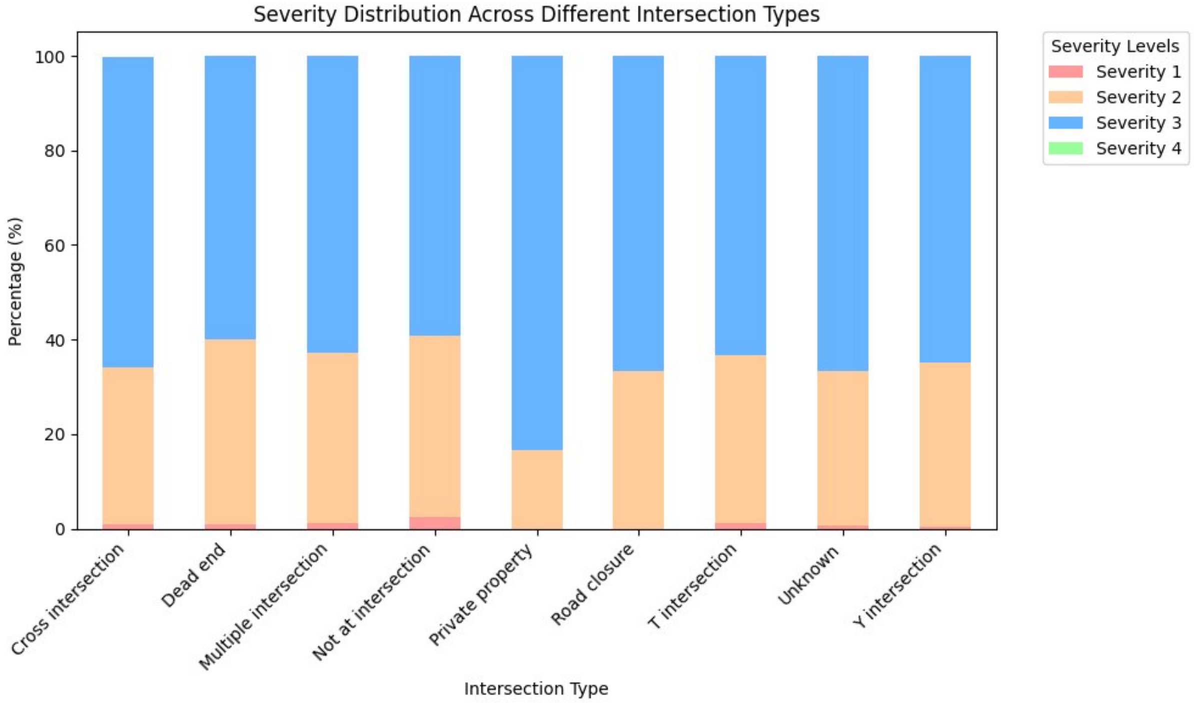
0.006), while private properties and multiple intersections have the lowest MI, suggesting minimal correlation with severity of accident. These results indicate that particular intersection types are more strongly correlated with severity, but that overall intersection type has a weak association with crash outcome.

Table 1: Mutual Information values for each intersection type

Intersection Type	**Mutual Information**
Not at Intersection	0.014
Cross Intersection	0.006
Road Closure	0.003
T-Intersection	0.002
Dead-end	0.002
Y-Intersection	0.002
Unknown	0.002
Private Property	0.0002
Multiple Intersection	0.000

The bar graph (Figure 2), displaying the percentage distribution of severity levels across different intersection types, shows that trends are consistent despite variations in road geometry. The graph indicates that severity level 3 is the most common across all intersection types, with level 1 being much rarer and level 4 almost negligible. Accidents that occurred not at an intersection had the highest proportion of severity level 1 crashes compared to other intersection types. Although the percentage difference is small, it aligns with the mutual information results, where “not at an intersection” had the highest MI score (0.014). The overall uniformity in severity distribution across road geometries may explain the low mutual information scores observed in Figure 1, suggesting that road geometry has a weak correlation with accident severity. While MI and NMI show that road geometry alone does not strongly influence crash severity, the slightly elevated fatal crash rate for non-intersection roads may indicate a higher risk for severe outcomes in these areas.

Figure 2: Bar graph displaying severity for each intersection type



The box plot (figure 3) of severity vs speed zone displays the distribution of speed limits for each severity level (with 1 being the most severe and 4 being the least). Severity level 1 has a wide range of speed zones (30 km/h to 110 km/h), with a median of 80 km/h, suggesting that the most severe accidents tend to occur on high-speed roads. Severity level 2 also spans the full speed range but has a lower median of 60 km/h, indicating that these crashes more commonly occur at moderate speeds within a narrower spread. Severity level 3 has a similar range (30 km/h to 110 km/h) and a median of 60 km/h, with accidents more evenly distributed across speed zones. Severity level 4 has the smallest range (40 km/h to 100 km/h) and the most compact box, suggesting that lower severity accidents are associated with slower speeds and occur in more consistent conditions and showing the lack of data for severity 4.

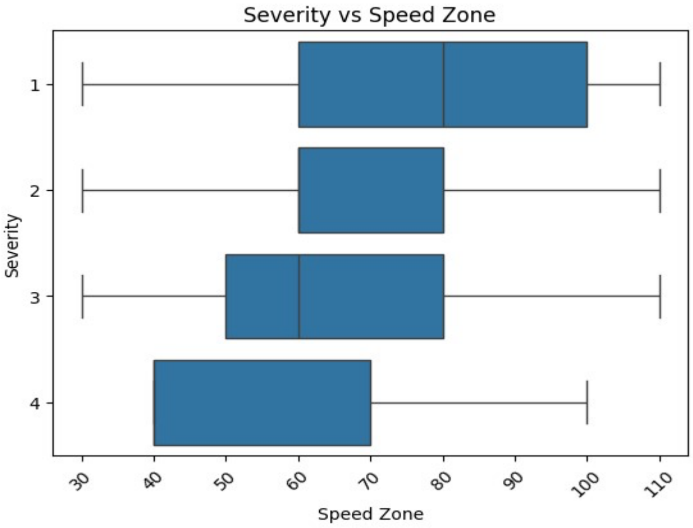


Figure 3: Box plots displaying the distribution of speed zones for each severity level

The MI analysis (Table 2) between accident severity and speed zone show that 60km\h has the highest MI value (0.008640), indicating the strongest relationship with severity level, followed by 100km/h (0.006442). Speed zones of 30km/h, 70km/h, 90km/h have a lack of relationship with severity level, likely due to the rarity of these speed zones and thus rarity of serious accidents in these speed zones on Victorian roads.

Table 2: Mutual Information values for each speed zone

Speed Zone	**Mutual Information**

60	0.009
100	0.006
80	0.003
50	0.002
40	0.001
110	0.0002
30	0.000
70	0.000
90	0.000

Further Correlation Analysis

For further correlation analysis we explored the following relationships:

- Road surface type vs accident severity (for both dry and wet conditions)
- Road surface and geometry combinations vs fatality rate
- Traffic control and speed zone combinations vs accident severity

Comparison of road surface and surface condition (Figure 4) vs accident severity level suggests that there is no consistent difference between wet and dry conditions for severity level distribution.

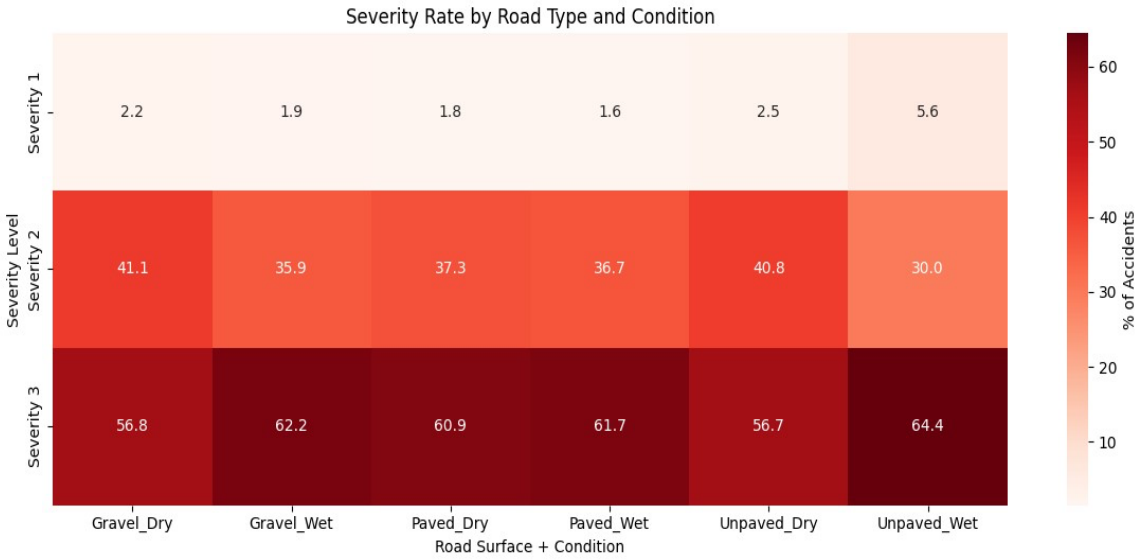


Figure 4: Frequency heatmap for road surface and condition vs accident severity level

Figure 5 reinforces the distribution of accident severities on wet and dry gravel roads. The differences between dry and wet conditions are largely negligible across all severity levels, suggesting that surface condition alone may not significantly influence crash severity on gravel roads.

Severity level 3 accounts for the highest proportion of crashes in both dry and wet conditions, with a slightly higher rate observed in wet conditions, aligning with expectations of the dataset. Severity level 2 accidents are slightly more common in dry conditions, while severity level 1 (the most severe) remains rare under both conditions.

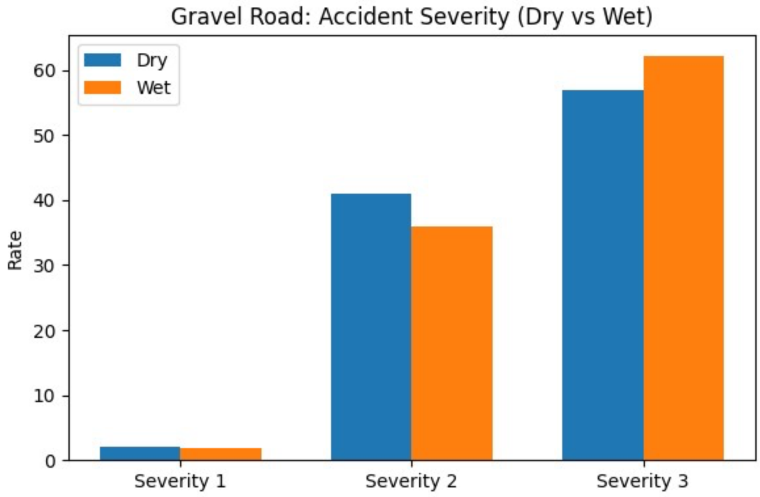


Figure 5: Bar charts comparing the severity makeup for wet and dry gravel roads

The distribution for unpaved roads (figure 6) is largely similar to the gravel roads. However, a noticeable difference is that level 1 makes up a larger proportion of accidents for the wet unpaved roads, suggesting that an accident in these conditions could lead to a more severe outcome.

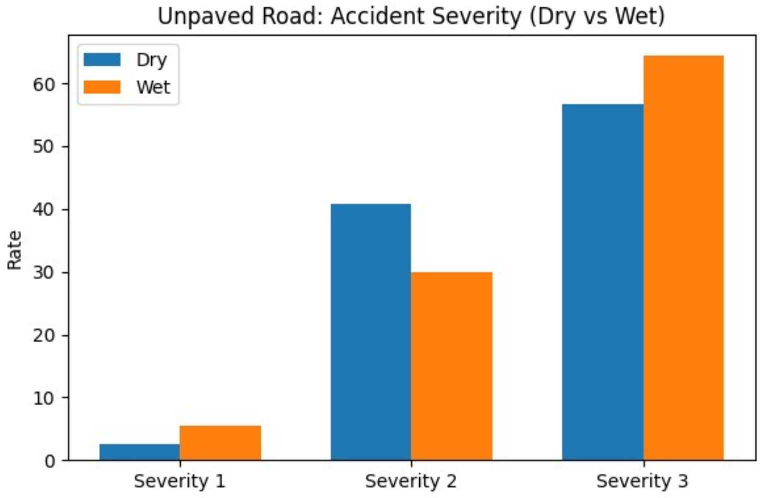


Figure 6: Bar charts comparing the severity makeup for wet and dry unpaved roads

Table 3 has generally low mutual information (MI) values across road surface types; however, unpaved roads exhibit a notably higher MI score compared to both paved and gravel roads.

This coincides with the findings from the severity rate calculations, which indicate that unpaved roads have the highest proportion of severity level 1 accidents. This suggests that unpaved surfaces, particularly when wet, are associated with a greater risk of severe crashes, making them the most dangerous road type in the dataset based on MI. The elevated MI value for unpaved roads reflects their stronger relationship with accident severity relative to other surfaces.

Table 3: Mutual Information values for road surfaces

Surface	**Mutual Information**
Gravel	0.0007
Paved	0.00002

Unpaved	0.002

Fatality rates across combinations of road surface and geometry are generally low (Figure 7), likely due to the small number of fatal accidents recorded in the dataset. Despite this, some patterns exist.

Unpaved cross intersections exhibit the highest fatality rate, even though they have one of the lowest overall severity rates. This suggests that whilst accidents at unpaved cross intersections are less frequent, they are more likely to be fatal. Other combinations such as paved or gravel intersections show lower fatality rates, suggesting lower risk of fatal accidents.

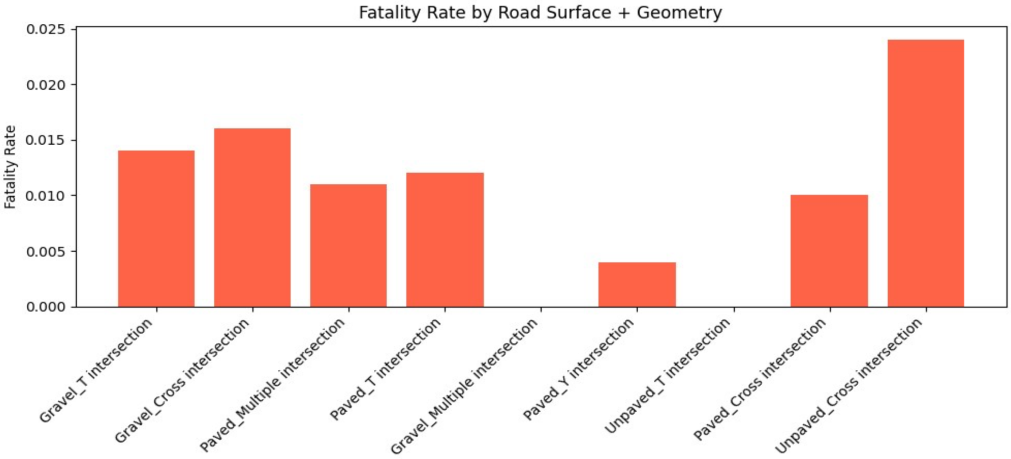


Figure 7: Mutual information values for road surface and intersection type with risk

The accident severity of levels 1 and 2 was then combined with accidents involving a fatality to create a new "risk" target variable. Combining target values into one variable was useful as fatalities are quite rare in the dataset; therefore, combining fatalities with accident severity yielded a stronger signal when computing Mutual Information (Table 4). It also made the data easier to interpret and draw hypotheses from.

However, computing MI values for this new variable against road surface and intersection type resulted in very low values, suggesting weak correlation. This suggests the lack of a concrete conclusion from the fatality rate was valid.

Table 4: Mutual information values for road surface and intersection type with risk

Short Combo Name	Mutual Information
Paved_DESC Cross intersection	0.001046
Paved_DESC T intersection	0.000115
Gravel_DESC T intersection	0.00000377
Paved_DESC Y intersection	0.00000371
Paved_DESC Multiple intersection	0.00000331
Unpaved_DESC Cross intersection	0.00000312
Unpaved_DESC T intersection	0.00000071
Gravel_DESC Multiple intersection	0.00000016
Gravel_DESC Cross intersection	0.00000013

Comparing traffic control and speed zone combinations with severity rates highlights a few key relationships (Figure 8). The “No control” traffic description has three separate speed zone entries, indicating a relatively high correlation with accident severity. Overall, the higher speed zones such as “90–110” have a greater proportion of severe accidents (levels 1 and 2), compared to lower speed zones which have more level 3 accidents. The “Stop sign, 90–110” combo has the highest proportion of level 1 accidents, and is one of only two entries where level 2 accidents outnumber level 3, further highlighting the dangers at high speeds. Additionally, “No control, 90–110” also has a considerable level 1 severity rate, reinforcing the risks associated with high-speed, low-control environments.

The higher speed zones exhibit a greater spread of severity levels, whereas lower speed zones are heavily skewed toward level 3. This spread can be seen clearly in the bar graph, with green bars (level 3) dominating lower-speed combinations.

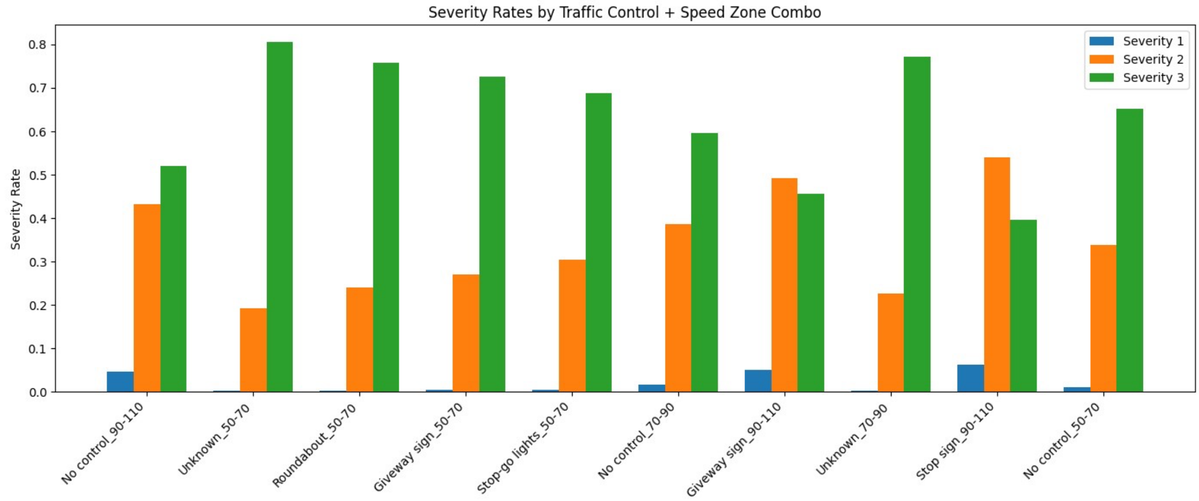


Figure 8: Bar charts comparing the severity makeup for wet and dry unpaved roads

The “No control, 90–110” combination also has a considerably higher MI value (Table 5) compared to all other traffic control/speed zone combination This supports the hypothesis that it is the most dangerous traffic control-speed zone combination. Overall, while all combos have relatively low MI values, this is expected due to the rarity of severe crashes. We expand upon this in the limitations section.

Table 5: Mutual information values for each speed zone x traffic description combination

Combo	Total Accidents	MI with Binary Severity
No control_90-110	32,471	0.00447
Unknown_50-70	3,389	0.00092
Roundabout_50-70	6,599	0.00088
Giveway sign_50-70	10,293	0.00075
Stop-go lights_50-70	28,158	0.00073
No control_70-90	36,701	0.00063
Giveway sign_90-110	1,759	0.00049
Unknown_70-90	1,681	0.00028
Stop sign_90-110	490	0.00024

No control_50-70	75,303	0.00018

Supervised Learning Model

The KNN model achieved an overall accuracy of 57%, calculated by dividing the 6638 true positives and 5178 true negatives by the 20729 total test cases as shown by the confusion matrix (figure 9). The model had a better accuracy in both the prediction of true positives and true negatives, suggesting that the features used to train the model have some level of impact on predicting the heaviest vehicle damage.

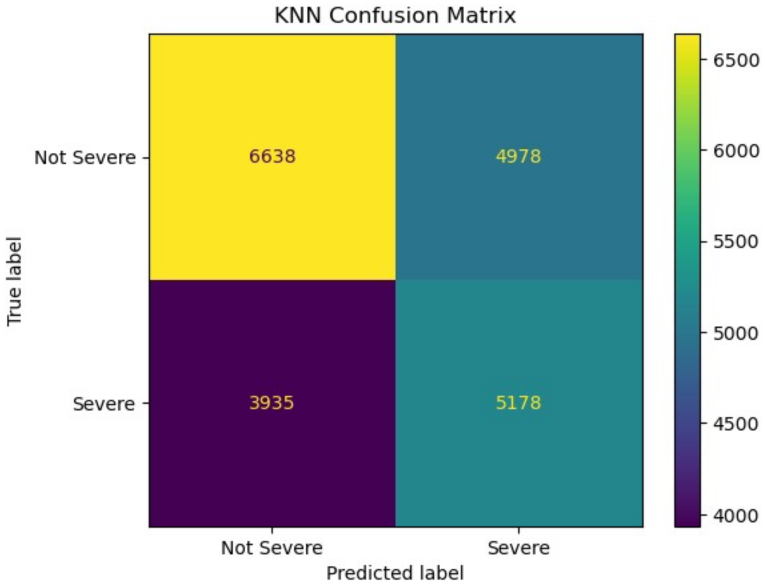


Figure 9: KNN Confusion Matrix, produced by *scikit-learn*'s *ConfusionMatrixDisplay*

The classification report (Table 6) for KNN shows the precision, recall, and F1 score for the model. Notably, the model had a 0.598 F1-score for the Not Severe category, and a 0.537 F1 score for the severe category. Both the Confusion Matrix and the classification report indicate that the model did a better job in the prediction of accidents in the category of Not Severe.

Table 6: KNN tabulated classification report, produced by *scikit-learn*'s *classification_report*

Outcome	Precision	Recall	F1-score	Support
Not Severe	0.628	0.571	0.598	11616
Severe	0.510	0.568	0.537	9113

Comparatively, the Random Forest model achieved an overall accuracy of 63.7%, with an F1 score of 0.699 for the Note Severe category and 0.542 for the Severe category (Table 8). The recall for the Not Severe category was 0.752, indicating that the model did a good job of identifying most Not Severe crashes, but the performance was far weaker for the Severe category (0.490) highlighting a potential issue with the identification of severe crashes.

The detailed evaluation of the two models is in the discussion section.

Table 7: Random Forest tabulated classification report, produced by *scikit-learn*'s *classification_report*

Outcome	Precision	Recall	F1-score	Support
Not Severe	0.654	0.752	0.699	11616
Severe	0.606	0.490	0.542	9113

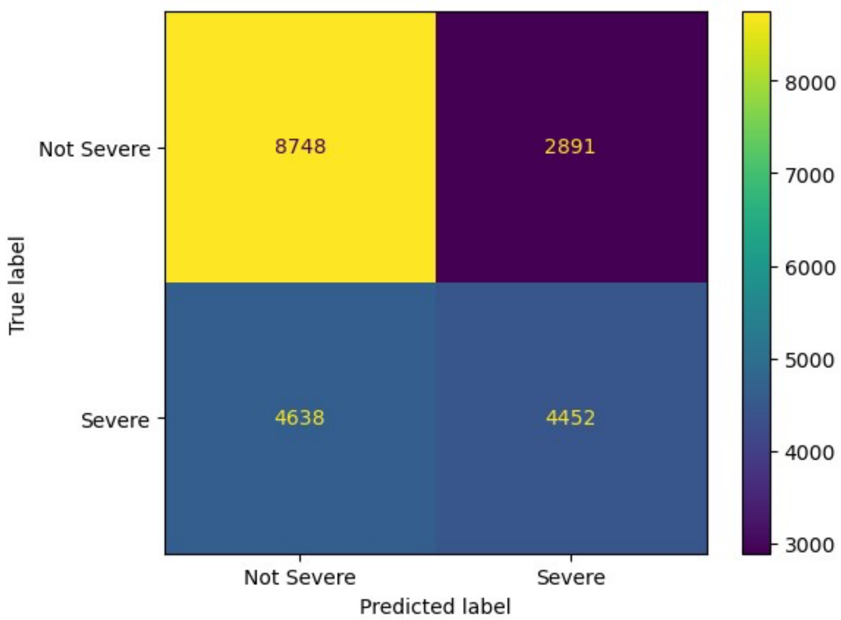


Figure 10: Random Forest Confusion Matrix, produced by *scikit-learn*'s *ConfusionMatrixDisplay*

Displaying the most important features in the Random Forest model (Figure 11) reveals that speed zone had by far the highest impact on the training of the model, suggesting that a clear relationship exists with severity. This supports the findings in the correlation analysis. Additionally, the accident not occurring at an intersection was of decent feature importance supporting the conclusion made in our primary analysis that not being at an intersection has an impact on overall severity. Other important features included the surface condition. Road surface type was found to not be that useful, and thus no meaningful relationship can be drawn

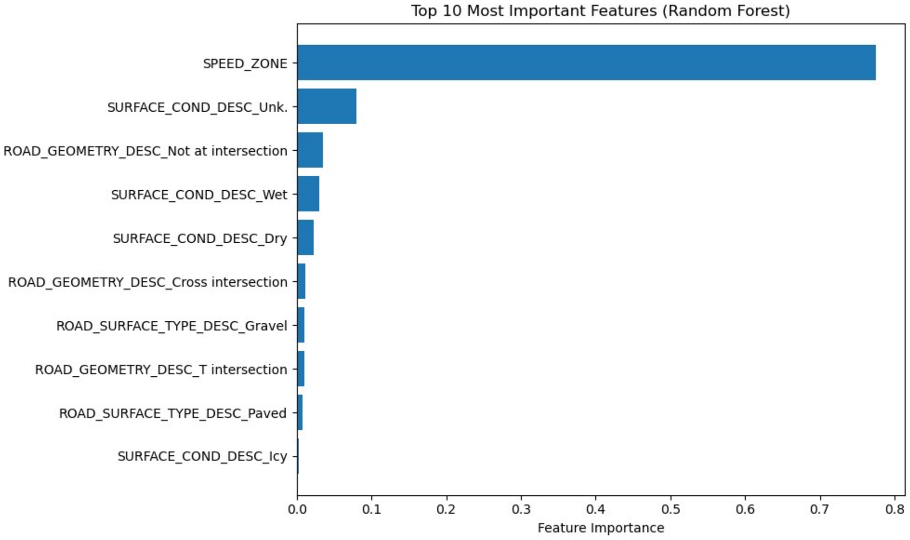


Figure 11: Bar chart of the most important features used for training by Random Forest

Discussion and Interpretation

Correlation Analysis

Overall, all correlation analyses produced low MI values, which may suggest weak correlation. However, given the size of the dataset, several dozen factors will influence accident severity/fatality. Therefore, a single feature or a combination of a couple is unlikely to produce a high MI value. The low MI values indicate that road geometry or speed zone don't independently explain the variation in accident severity. Other variables or variables in combinations may be able to explain the severity of an accident.

However, a key relationship identified through MI value was accidents that did not occur at intersections and accident severity. A closer look at the proportion of severe accidents (severity 1) that occur not at intersection suggests that accidents that do not occur at intersections are more likely to be severe. This could be due to a number of factors, but most likely due to the fact that collisions will occur at higher speed if not at an intersection, reducing severity. When coming to an intersection, cars will slow down and thus the severity of accidents will on average decrease. Alternatively, when an accident occurs not at an intersection, it is likely on a road where cars are travelling at full speed leading to worse outcomes.

Another trend identified was that road accidents with speed limits of 60km/h had a higher MI with severity. This could be due to a higher speed combined with pedestrian use (pedestrians are not found near 100km/h or 80km/h roads) leading to more fatal outcomes.

Speed zone had a higher overall MI with severity than road geometry, suggesting that it is more important in road safety and crash outcomes. This is expected, as the speed one is travelling at when they crash is more important to their outcome than the way they crash.

Higher speed zones also correlated with severity; this could be due to more reckless driving in higher speed areas. Drivers may be more aware of their surrounding when driving at lower speeds, thus can identify threats quickly and avoid them. **Supervised Learning Models**

The results from our supervised learning models suggest that the combination of our pre-crash features have a tangible impact in the predicting of accident severity. Thus, further highlighting that crash severity is dependent on these variables.

With the Random Forest important features array we identified that speed zone was the by far most influential feature in the training of the model. This suggests that the strongest relationship between our metrics and accident severity, as determined by vehicle damage, was the speed zone the accident occurred in. This result supports the conclusion drawn that speed zone is a more important factor than road geometry. Again, speed zone influencing the training of the model so heavily is not a surprising result given how much a car is damaged during the crash is going to drastically affected by the speed the car was travelling during the event of the crash.

For similar reasoning, Not at Intersection was a lesser, but still important factor in the training of Random Forest. As mentioned in the correlation analysis, cars that are not at an intersection are likely travelling the speed limit, so in the event of a crash the outcome is going to be a lot more severe than a crash involving a car that is slowing down for an intersection.

Another insight that is provided from the feature importance is the slight impact of road surface condition. Although not as influential as speed zone, road surface condition still played a notable role, particularly when it was not known, or the surface was wet. This is likely because wet conditions can reduce traction and increase the damage sustained to a vehicle during a crash. **Supervised Learning Model Evaluation**

Overall, the Random Forest model performed better in the prediction of vehicle damage severity, given both models were trained on the same data. By implementing Random Forest, we increased the accuracy by 6.7% (63.7% vs 57%) from KNN. Additionally, Random Forest had a better F1 score for both the Severe and Not Severe classes, indicating a better balance between recall and precision. The only metric KNN outperformed Random Forest was Severe recall (0.568 vs 0.490). There is an imbalance in the recall for Severe and Not Severe categories for Random Forest, which was probably caused by an imbalance in the training data which did not affect KNN as much. As a result, Random Forest far outperforms KNN for Not Severe recall (0.752 vs 0.571). If we had directly handled the class imbalance, then Random Forest likely would have outperformed KNN in both categories, but by a more event amount.

Random Forest was the more precise model in both categories and therefore had a better overall F1 score for both categories. This clearly indicates that Random Forest was the higher performing model and is of much more use in the prediction of crash severity. This superiority was expected due to the ability of Random Forest to handle diverse scaling and reduce overfitting (through the aggregation of results of multiple decision trees.

Ultimately, the performance of the models is probably not sufficient for any real-world applications, 63.7% accuracy is not good enough for ongoing practical applications. However the results helped confirm links between the pre-crash data and severity, and the feature importance from Random Forest allowed us to highlight the strongest of these links.

Limitations

Correlation Analysis

The dataset is as follows:

- Level 1 severity accidents: 2526 cases
- Level 2 severity accidents: 53117 cases
- Level 3 severity accidents: 90754 cases

This imbalance can lead to faulty correlation metrics. Mutual information, for example, measures how much a feature reduces uncertainty in the target. In this case, many features would only slightly shift the target from the dominant class and therefore may appear irrelevant.

Binary encoding of the severity levels was used to try to mitigate this limitation; however, it was only partially successful.

Similarly, the distribution of road surfaces is extremely imbalanced:

- Gravel roads: 4959 cases
- Paved roads: 140506 cases
- Unpaved roads: 517 cases

This imbalance will produce slightly inaccurate metrics, especially in the correlation analysis of road surface and road condition (wet/dry) with severity. Possible workarounds, such as undersampling/oversampling, will lead to further inaccuracies and therefore are better avoided.

The dataset contains a total number of 146,400 accident entries, however only 2526 accidents reported fatalities of 1 or more. Accidents resulting in fatalities account for only 1.7% of all accidents. This imbalance may explain the unreliability of fatality rate values in the above correlation analysis. Merging the fatality rate with accident severity may address this however it also correlates different types of outcomes.

Furthermore, in many correlation analysis cases we filtered out data with less than 30/50 entries to reduce statistical noise, whilst necessary this may have got rid of important entries that could change our correlation values and ultimately conclusions.

In most correlation analysis tasks above the only correlation metrics applied are a form of accident severity rate and Mutual Information. Whilst applicable, Mutual information quantifies how much knowing one variable reduces uncertainty in the other, MI does not indicate directionality or causality. For example, a high MI value between high-speed zones and severity does not necessarily indicate high speeds caused accidents of higher severity, therefore relying on mutual information alone may lead to faulty conclusions. Albeit limited correlation metrics are viable for the given structure of analysis.

Supervised Learning Model

Both supervised learning models struggled to identify cases of severe vehicle damage, which was shown in both the confusion matrices and recall scores. These errors likely suggest that the data used to train the models did not contain enough variety. The commonality of the conditions (e.g. how often it is dry, 60km/h speed zone) likely meant that the models did not receive enough variety to correctly identify new or rare crash cases. Additionally, the imbalance between Severe and Not Severe cases in the training data probably lead to the higher accuracy when identifying Not Severe cases, particularly for the Random Forest Model. A more concerted effort to address the imbalance may have led to a more even split of accuracy across the two classifications.

Another limitation was the definition of severity. Severity not a concrete value and rather an abstract concept. As such we had to define our own ways of measuring severity, but given the many possible metrics we could have used (fatality, injuries, vehicle damage, etc) this proved harder than initially imagine. For the supervised learning models, we chose to use vehicle damage for reasons outlined in the method. Initially, we divided damage severity into three classes (Minor, Moderate, Severe), but the moderate group proved difficult to identify due to its overlap with other classes. As a result, both models produced very inconsistent results when attempting to isolate the moderate cases. To address this, we turned severity into a binary classification — grouping levels 1–3 and 6 into "Not Severe" and 4–5 into "Severe." This improved the accuracy of the models, whilst still producing interesting results. However, it is possible that the binary classification was not nuanced enough to train the models to produce more profound results.

Finally, the dataset only included recorded crashes, thus unreported crashes or near-miss incidents were not considered. This introduces bias, as the model is trained on a limited view of crash outcomes conditions. Very minor collisions and crashes may not be reported and hence not included in the data. This could cause problems for the models when identifying very low risk outcomes.

Conclusion

The findings from both our machine learning models and correlation analysis indicate that speed zone is the most important factor in determining accident severity. This is in-line with expectations, as one's speed when they crash is going to have the most impact on the crash severity. Accidents occurring not at intersections were also highlighted in both analysis and supervised learning, likely due to cars travelling at higher speeds when they crash on a road vs when they crash at an intersection. Road surface condition had a noticeable impact on severity, reinforcing the notion that wet roads can potentially lead to more dangerous outcomes. However, road surface type and the alternative intersection types had minimal impact, indicating they are less useful in predicting severity. Ultimately, our results suggest that features related to speed, and to a lesser extent conditions, have the most impact on the severity of a crash when it occurs.

References

Scikit-learn developers. (2024). *1.11.2 documentation: Random forests and other randomized tree ensembles*. https://scikit-learn.org/stable/modules/ensemble.html#random-forests-and-other-randomized-treeensembles

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