

Predicting the Severity of an Accident

Lyndall Fourie

September 2020

1. Introduction

1.1 Background

Every year millions of people across the world are involved in road accidents. The severity of these accidents differ and an understanding of which factors impact the severity of an accident may assist emergency response units, hospitals and road side assistance companies to plan accordingly. The insight may also assist motorists to plan their journeys.

1.2 Problem

Many factors may influence the severity of a road accident. The selection of factors and the way each factor contributes to the severity of an accident are often difficult to assess. Machine Learning (ML) will assist in understanding the patterns and make predictions that will assist the target audience to predict the severity of an accident in order to plan accordingly.

1.3 Scope

The focus of this study is to develop a ML model to assist Seattle City in predicting the severity of accidents within its border. Various factors may influence the severity of the accidents and will be assessed for use in the prediction model, but only factors that may influence the severity of the accident before it happens will be included in the scope. For instance, both weather and the angle of impact may influence the severity of the accident, but the target audience can only act according to the weather.

The goal of the study is: *Predict the severity of an accident, based on preselected features.*

2. Data acquisition and cleaning

2.1 Data source

The data source is the IBM Capstone Project dataset. The following is important to note:

- The data was provided by the Seattle Traffic Department and contains information on accidents during the period of 1 January 2004 to 20 May 2020;
- The data set contains data of accident classes based on the severity of accidents. The label for the data set is therefore severity, which describes the fatality of an accident;

- The columns contains various factors (features) that may influence the severity of an accident.
- The data set has a total of 194674 cases; and
- The original data set contained 5 classes namely: 3 – *fatality*, 2b – *serious injury*, 2 – *injury*, 1 – *property damage* and 0 – *unknown*. The data set provided for the Capstone project contains only two classes namely: 2 – *injury* and 1 – *property damage*.

2.2 Data limitations

The following are potential data limitations:

- The data has unbalanced labels which will lead to a biased ML model. It should therefore be balanced during the course of the project;
- The data set contains incomplete records and fields with incorrect formats. Data cleaning will be required to ensure the data set contains information that will be useable by the ML algorithm;
- Features engineering will be required to increase the predictability of the model; and
- The exclusion of 3 of the 5 severity classes from the data set may have an impact on certain statistical features of the data i.e. statistical significance between the features of severity classes 3 and 1 would have been high whereas the statistical significance between the features of severity classes 2 and 1 may be statistically insignificant. This may have an impact on the usability of the ML model.

2.3 Feature engineering

The sourced data set contained 38 features (columns) that were both relevant and irrelevant to the scope of the project namely, *to predict the severity of accidents within the Seattle City border*. The features that are relevant to the scope of the project are listed in column 1 of Table 1 and were included in a new data set. The features that are not relevant to the scope of the project are listed in column 2 of Table 1, with column 3 describing the reason why it is not relevant to the scope of the project.

It may be useful to understand the impact of the ‘Day of the Week’, ‘Month of the Year’ and ‘Time of Day’ features on the severity of an accident. These features were derived from the INCDATE and INCDTTM columns and listed in column 1.

Table 1: Features included and excluded from the project data set

Columns kept in the data set	Columns dropped (excluded) from the data set	Reason for exclusion from the data set
SEVERITY CODE	X	More than 10 000 unique entries
OBJECTID	Y	More than 10 000 unique entries
SEVERITYDESC	INCKEY	More than 10 000 unique entries
INCDATE	ADDRTYPE	Similar to JUNTIONTYPE
INCDTTM	SEVERITYCODE	Duplicate column
JUNTIONTYPE	STATUS	No value for the analysis
WEATHER	COLDTKEY	More than 10 000 unique entries
ROADCOND	REPORTNO	More than 10 000 unique entries
LIGHTCOND	INTKEY	Too many empty fields (>50%)
Day of the Week	LOCATION	More than 10 000 unique entries

Columns kept in the data set	Columns dropped (excluded) from the data set	Reason for exclusion from the data set
Month of the Year	EXCEPTRSNCODE	Too many empty fields (>50%)
	EXCEPTRSNDESC	Too many empty fields (>50%)
	SDOT_COLCODE	Will not predict the severity of an accident before it happens.
	SDOT_COLDESC	Will not predict the severity of an accident before it happens.
	INATTENTIONID	Too many empty fields (>50%)
	PEDROWNOTGRNT	Too many empty fields (>50%)
	SDOTCOLNUM	More than 10 000 unique entries
	SPEEDING	Too many empty fields (>50%)
	COLLISSIONTYPE	Will not predict the severity of an accident before it happens.
	UNDERINFL	No indication of what 0 and 1 refer to.
	SEGLANEKEY	Too many fields with '0' (>50%)
	CROSSWALKKEY	Too many fields with '0' (>50%)
	HITPARKEDCAR	Too many fields with 'N' (>50%)
	INCDATE	Replaced by 'Day of the Week' and 'Month of the Year'.
	INCDTTM	Replaced by 'Time of Day'.
	PERSONCOUNT	Will not predict the severity of an accident before it happens.
	PEDCOUNT	Will not predict the severity of an accident before it happens.
	PEDCYLCOUNT	Will not predict the severity of an accident before it happens.
	VEHCOUNT	Will not predict the severity of an accident before it happens.
	Time of Day	Implied by LIGHTCOND

2.4 Data Cleaning

The following data cleaning activities were performed:

- Created a new data frame with the selected features;
- Identified the columns with missing values:
 - CCOLLISIONTYPE (4904) - 2.5% of dataset
 - JUNCTIONTYPE (6329) - 3.25% of dataset
 - WEATHER (5081) - 2.6% of dataset
 - ROADCOND (5012) - 2.6% of dataset
 - LIGHTCOND (5170) - 2.7% of dataset
- Further analysis on the csv dataset indicated that in 99.57% of the cases, where the COLLISIONTYPE field is empty, the WEATHER, ROADCOND and LIGHTCOND fields are also empty. The empty COLLISIONTYPE fields will therefore not contribute to the scope of the analysis and were deleted. Further analysis indicated similar scenarios for WEATHER, ROADCOND and LIGHTCOND. All rows with empty fields in these columns were therefore deleted.
- The empty values in the JUNCTIONTYPE column were replaced with 'Unknown'.

- In order to derive the 'Day of the Week' and 'Month of the Year', the INCDATE and INCDTTM column formats were changed to 'datetime' format.

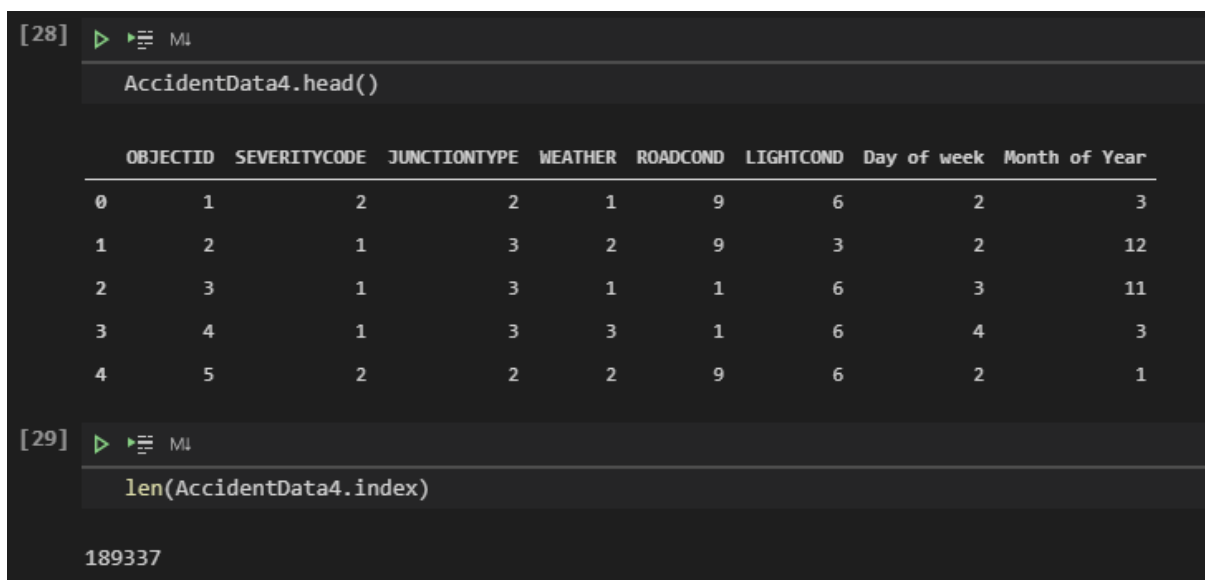
Number of cases were reduced from 194 673 to 189 337.

The formats of the columns, which will be used for further analysis, are in the format of 'int64', 'object' and 'datetime64'. This will be sufficient for exploratory analysis. To enable machine learning, all columns will be transformed to the appropriate formats at the specific stage in the project.

To further prepare the data for the Machine Learning algorithms, the following data preparation activities were performed:

- Converted data types to 'int64'; and
- Dropped the column 'SEVERITYDESC' of type 'object' as it conveyed the same information as column 'SEVERITYCODE', which was in the correct format.

Figure 1 illustrates the final columns that were used in the analysis.



```
[28] In [28]: AccidentData4.head()
```

	OBJECTID	SEVERITYCODE	JUNCTIONTYPE	WEATHER	ROADCOND	LIGHTCOND	Day of week	Month of Year
0	1	2	2	1	9	6	2	3
1	2	1	3	2	9	3	2	12
2	3	1	3	1	1	6	3	11
3	4	1	3	3	1	6	4	3
4	5	2	2	2	9	6	2	1

```
[29] In [29]: len(AccidentData4.index)
```

189337

Figure 1: Columns used in the ML algorithms

3. Exploratory Data Analysis

Exploratory data analysis is used to learn more about the data before applying machine learning models. This enables the practitioner to summarise data characteristics, discover trends and relationships and provide the practitioner with a mental model of the data in order to assess its relevance.

As part of this analysis, the following were assessed:

- Number of accidents per type of feature i.e. weather condition; and
- Relationship between various features (X) and the severity class (Y labels).

A total of 189 337 records (cases) were assessed.

3.1 Severity and Weather

The number of accidents per weather condition is illustrated in Figure 2. The codes are describes as follows:

- 1 = Overcast
- 2 = Raining
- 3 = Clear
- 4 = Blowing sand / dirt
- 5 = Fog / smog / smoke
- 6 = Other
- 7 = Partly cloudy
- 8 = Severe crosswind
- 9 = Sleet / hail / freezing rain
- 10 = Snowing
- 11 = Unknown

Most accidents occur during clear weather, followed by raining and overcast weather.

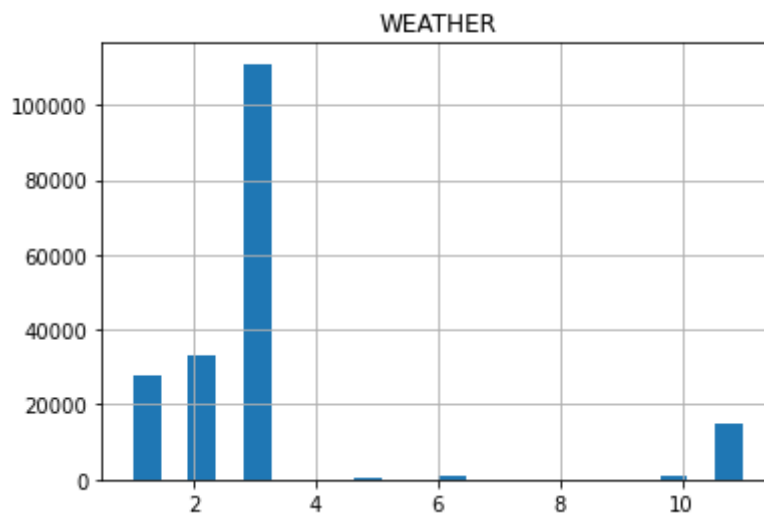


Figure 2: Number of accidents due to type of weather condition

Although this analysis provides some insight, it doesn't take into account the number of days in a year certain weather conditions are experienced and this may skew the results. For example, there are probably more days in the year with clear conditions than with severe crosswind. The same logic would apply to the other features.

For this reason the practitioner decided to analyse the data based on type of weather (or other feature / X) as a percentage of the class (Y). For instance, when the weather is clear, what percentage of accidents only result in damage to property (class 1) and what percentage of accidents result in injury (class 2). This is in line with the scope of this study namely: *Predict the severity of an accident, based on preselected features.*

The percentage of accidents only resulting in damage to property ('SEVERITYCODE' 1 -) and percentage of accidents result in injury ('SEVERITYCODE – 2') due to weather conditions are shown in Figure 3 (percentage is shown to two decimal places).

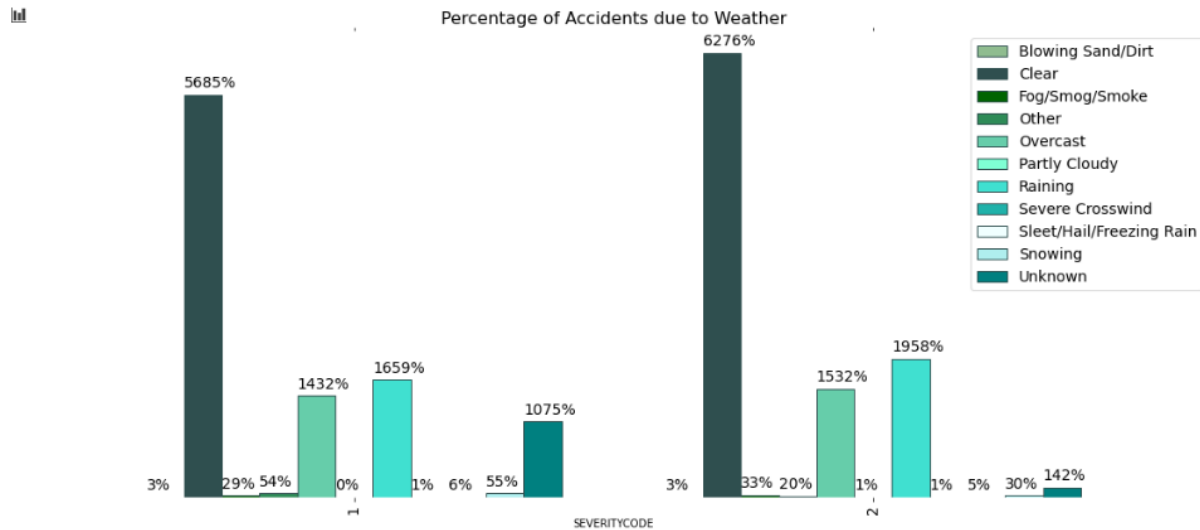


Figure 3: Percentage of accidents per weather condition based on class

In both severity classes the percentage frequency follows the same pattern, except where the weather condition 'Unknown' is significantly higher in severity class 1 (damage to property). This is a limitation of the data set, but were not deleted during data cleaning as it relates to a significant number of cases. It is therefore identified as a factor that may have an impact on the usability of the model.

The following weather conditions are slightly higher in severity class 2 (injury):

- Clear – 5.9% difference
- Overcast – 1% difference
- Raining – 2.99% difference

3.2 Severity and Road Condition

The percentage of accidents only resulting in damage to property ('SEVERITYCODE' 1 -) and percentage of accidents result in injury ('SEVERITYCODE – 2') due to road conditions are shown in Figure 4 (percentage is shown to two decimal places).

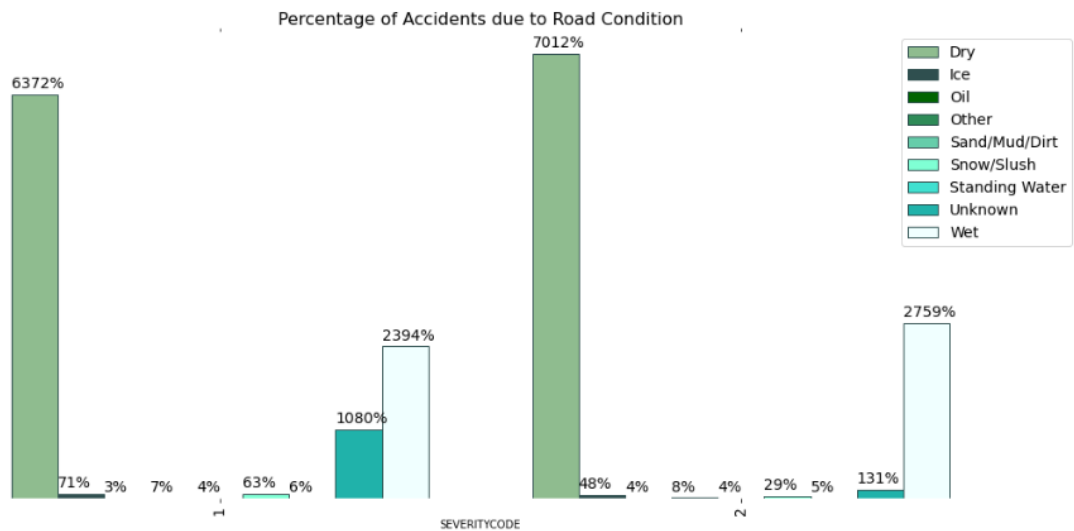


Figure 4: Percentage of accidents per road condition based on class

In both severity classes the percentage frequency follows the same pattern, except where the road condition 'Unknown' is significantly higher in severity class 1 (damage to property). As with the weather feature, this is a limitation of the data set, but were not deleted during data cleaning as it relates to a significant number of cases. It is therefore identified as a factor that may have an impact on the usability of the model.

The following road conditions are slightly higher in severity class 2 (injury):

- Dry – 6.4% difference
- Wet – 3.65 difference

The results correspond with the weather feature analysis where accidents in severity class 2 (injury) were more likely during clear and raining conditions.

3.3 Severity and Light Condition

The percentage of accidents only resulting in damage to property ('SEVERITYCODE' 1 -) and percentage of accidents result in injury ('SEVERITYCODE – 2') due to light conditions are shown in Figure 5 (percentage is shown to two decimal places).

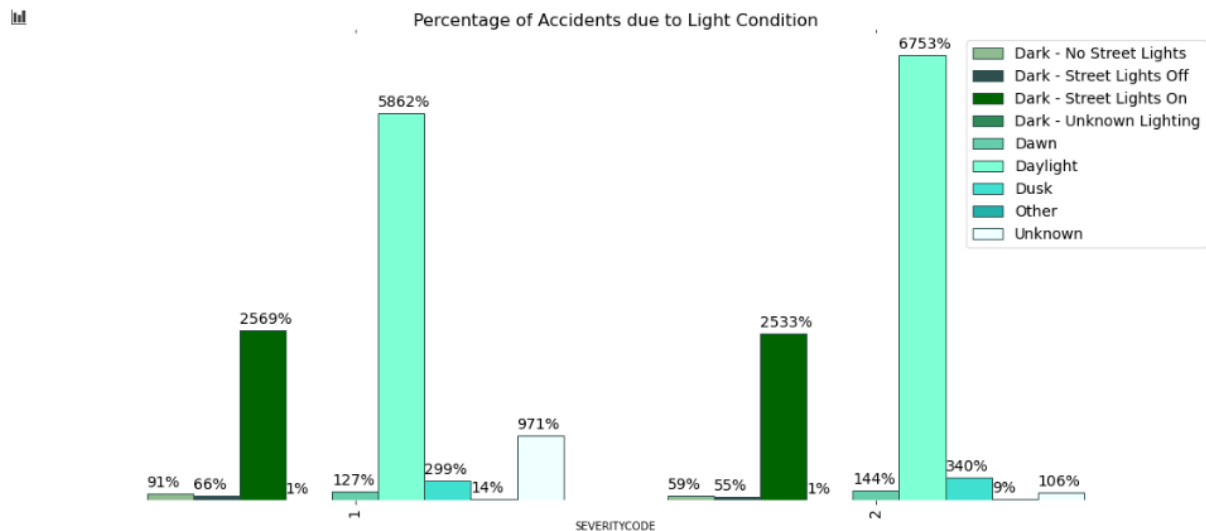


Figure 5: Percentage of accidents per light condition based on class

In both severity classes the percentage frequency follows the same pattern, except where the road condition 'Unknown' is significantly higher in severity class 1 (damage to property). As with the weather and road condition features, this is a limitation of the data set, but were not deleted during data cleaning as it relates to a significant number of cases. It is therefore identified as a factor that may have an impact on the usability of the model.

The following light conditions are slightly higher in severity class 2 (injury):

- Daylight – 8.91% difference

3.4 Severity and Day of Week

The percentage of accidents only resulting in damage to property ('SEVERITYCODE' 1 -) and percentage of accidents result in injury ('SEVERITYCODE – 2') due to the day of week are shown in Figure 6 (percentage is shown to two decimal places).

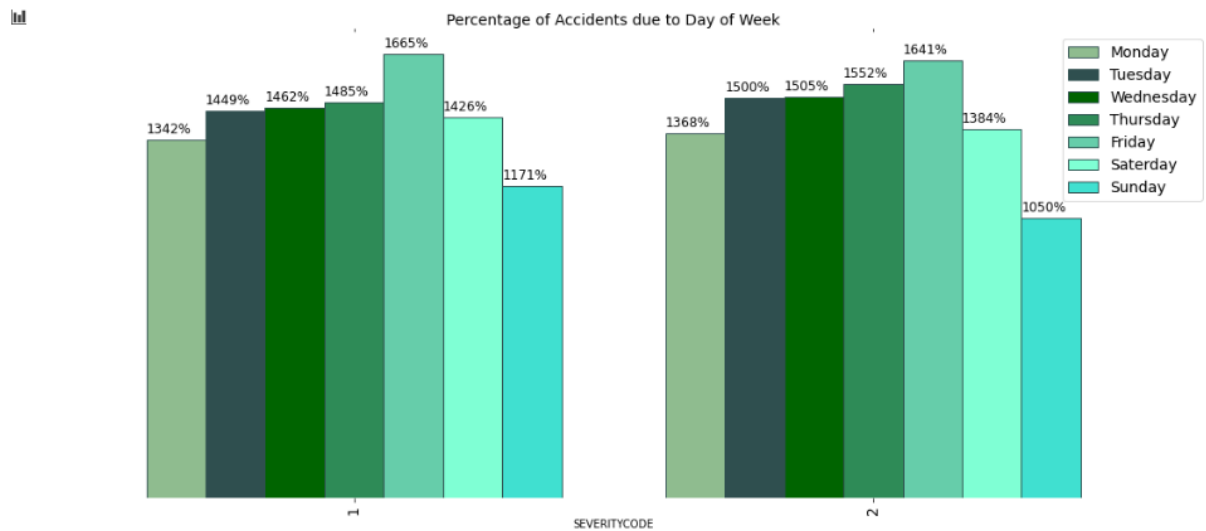


Figure 6: Percentage of accidents per day of week based on class

In both severity classes the percentage frequency follows the same pattern, i.e. most accidents occur on a Friday. There is no statistically significant difference between the severity of the accident and day of week. Regardless of this observation, the practitioner has decided to use this feature in the Machine Learning models as the Machine Learning algorithms may be sensitive to small differences in data.

3.5 Severity and Month of Year

The percentage of accidents only resulting in damage to property ('SEVERITYCODE' 1 -) and percentage of accidents resulting in injury ('SEVERITYCODE – 2') due to the month of year are shown in Figure 7 (percentage is shown to two decimal places).

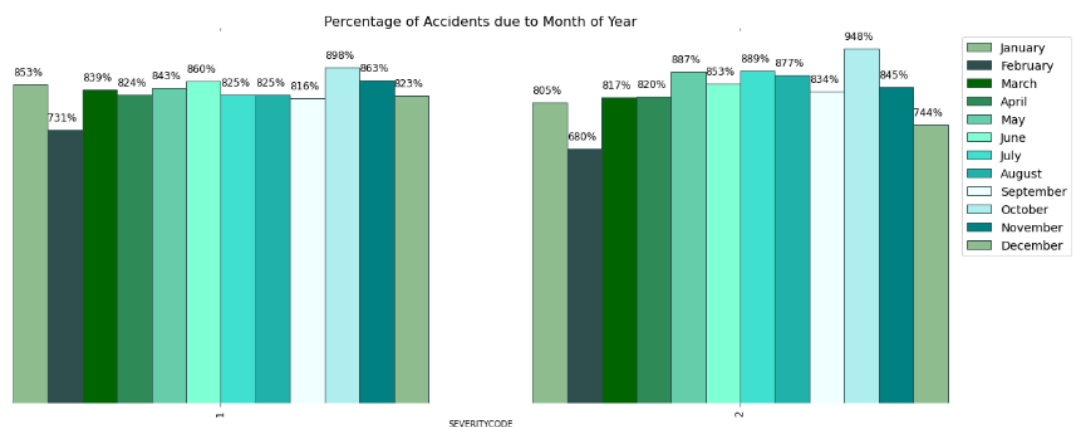


Figure 7: Percentage of accidents per month of year based on class

In both severity classes the percentage frequency follows the same pattern. There is no statistically significant difference between the severity of the accident and month of year. Regardless of this observation, the practitioner has decided to use this feature in the Machine Learning models as the Machine Learning algorithms may be sensitive to small differences in data.

3.6 Severity and Junction Type

The percentage of accidents only resulting in damage to property ('SEVERITYCODE' 1 -) and percentage of accidents resulting in injury ('SEVERITYCODE' – 2') due to the junction type are shown in Figure 8 (percentage is shown to two decimal places).

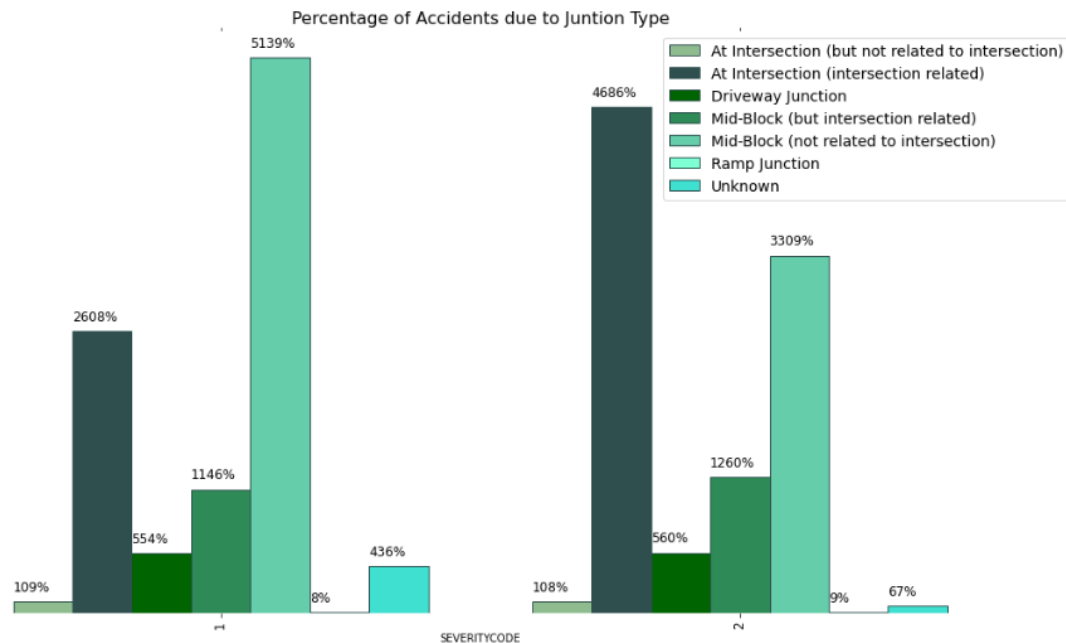


Figure 8: Percentage of accidents per junction type based on class

The percentage frequency between the two severity classes do not follow the same pattern. As with weather condition and road condition, the unknown factor contributes significantly. The following types of junction result in a higher probability for class 1 (property damage only):

- Midblock – 18.3%

The following types of junction result in a higher probability for class 2 (injury):

- At intersection (intersection related) – 20.78%

4. Predictive Modelling

Classification type of Machine Learning models will be used to *predict the severity of an accident, based on preselected features*. The features were analysed in more detail in Section 3.

A shortcoming of the data is that it has unbalanced labels which will lead to biased Machine Learning models. Approaches to deal with unbalanced data are as follows:

- Use different Machine Learning algorithms
- Use accuracy validation techniques such as F1 – score, Jaccard index and Precision / Recall metrics.

4.1 Model Development

The following Machine Learning classification algorithms were used:

- K-Nearest Neighbours (KNN);
- Decision Trees;
- Support Vector Machines (SVM); and
- Logistic Regression (LogLoss)

The following approach was used to develop the models:

- a) Define features (X)
- b) Convert the Pandas data frame to a numpy array
- c) Define the labels (Y)
- d) **Normalise the data** (this was not done for Decision Trees as it is not a requirement for the algorithm)
- e) Split the data set into Train and Test sets (Train = 0.8 (151 469), Test = 0.2 (370868))
- f) Train the model using the training data set
With the KNN algorithm the accuracy was calculated for different k values. Selected the k value (k = 8) with the highest accuracy
- g) Use the model to predict, using the test data set

4.2 Model Evaluation

Table 2 lists the performance of the test sets for each classification model. Best performances are labelled in red.

Table 2: Evaluation performance of the models

	KNN	Decision Tree	SVM	LogLoss
Accuracy	0.67986	0.695204	0.69616	0.69579
F1 – score	0.63159	0.570208	0.57145	0.57799
Jaccard index	0.64755	0.69520	0.69616	0.69299
Precision / Recall score	0.68	0.7	0.7	0.58328
LogLoss score	-	-	-	0.58323

The SVM model scored the highest in three of the four evaluation techniques.

5. Conclusion

The goal of the study was to *predict the severity of an accident, based on preselected features*. The dataset used contained a sufficient number of cases, but certain data cleaning activities had to be performed to ensure the outcome of the analysis was useful and could be used by the Machine Learning algorithms.

Notable data shortcomings that could not be corrected by the data cleaning and normalisation activities were:

- The feature called 'Unknown':
This may influence the usability of the Machine Learning model.
- Only two classes from the original data set were provided:
The exclusion of 3 of the 5 severity classes from the data set may have an impact on certain statistical features of the data i.e. statistical significance between the features of severity classes 3 and 1 would have been high whereas the statistical significance between the features of severity classes 2 and 1 may be statistically insignificant. This may have an impact on the predictability of the ML model.
- The data had unbalanced labels which would have lead to a biased Machine Learning model:
Different machine learning algorithms and model validation techniques were used to minimise the risk.

Four different classification algorithms were used namely KNN, Decision Tree, SVM and LogLoss. The models were validated using different validation techniques namely Accuracy, F1 – score, Jaccard index and Precision / Recall score. The SVM model scored the highest in three of the four model validation techniques.

6. Future Work

The inclusion of the other severity classes may enhance the usability of the model. The practitioner would also recommend the guideline that 'Unknown' should not be a selection option when capturing the data, or that fields should not be kept blank, as this further influences the usability of the model.