Predicting the Severity of Traffic Accidents

Seshni Govender

October 2020

1 Introduction

1.1 Background

Traffic accidents are a significant cause of deaths, injuries, property damage and loss, and are major concerns for the public health and traffic safety. According to an article by the World Health Organization (WHO), approximately 1.35 million people die every year as a result of traffic accidents [1]. These accidents have negative effects on the direct participants of the incident physically and emotionally , and may incur considerable economic loss. Hence, given several factors, predicting the severity of such accidents can be useful since with this information a road user may be prompted to drive more carefully, use a different transportation route or possibly plan accordingly for such events .

1.2 Problem

Data that may determine the severity of a traffic accident may include road conditions, weather conditions, light conditions, alcohol or drug influences , whether the drivers were speeding or distracted, location and time, and number of participants and vehicles involved. This project aims to predict the severity (Injury or only property damage) of the traffic accident based on these data.

1.3 Interest

Road traffic safety institutions such as traffic police may be highly interested in the prediction of traffic accident severity to assist in improving overall road traffic safety. For example, if severe accidents are likely to occur at a certain time e.g. peak traffic hours, then more traffic police patrols can be dispatched to ensure that drivers adhere to the road rules. The general public may be highly interested in these predictions so that road users are likely to drive carefully, use different traffic routes or plan for a possible accident. Insurance companies may also be interested in these predictions to warn their clients of traffic accident possibilities.

2 Data

2.1 Data source

A shared data-source and metadata of the traffic accidents provided by the Seattle Police Department and recorded by Traffic Records for the Seattle city from 2004 to the present are used [2] [3]. The target or label is the accident "severity" in terms of human fatality or injury. There are 37 attributes in the data-set, some of which will not be useful and some contain empty data. The data contains only two types of severity categories:" injury" or "property damage". The data-set also has unbalanced labeled data with 136 485 accidents with a "property damage" severity and 58 188 accidents with a "injury" severity. This will need to be balanced so that the model will be unbiased. Supervised machine learning will be used to predict accident severity.

2.2 Data Cleaning

The data-set contained a lot of missing data. Rows which had more than half of the values in the columns as empty ("NaN") were dropped. For example, if the Feature set had 14 columns, then if a row had more than 7 of its column data as empty, it would be dropped.

To balance the data, a down sampling was performed on the majority class severity of type "property damage" or 1. This left a total sample of 114 320 accident records.

```
df_1 = df[df['SEVERITYCODE'] == 1]

df_2 = df[df['SEVERITYCODE'] == 2]

df_1.shape

(132630, 15)

df_2.shape

(57160, 15)

df_1_downsampled = resample(df_1,replace=False,n_samples=57160,random_state=123)

df_result = pd.concat([df_1_downsampled,df_2])

df_result.shape

(114320, 15)
```

Figure 1: Down sampling of collision data-set in Python

In the "ADDRTYPE" column which specifies the location of the accident, any unknown or missing row values were imputed using "Block" since the mass of the accidents occur at a "Block" address type.

The "SPEEDING"column had 9000 accidents where the driver(s) were speeding. The rest of the accidents did not have a value for this attribute. Either 1 of 2 options could be selected: (1) The empty values could be treated as a non-speeding accident (2) The "SPEEDING" column could be dropped altogether. Even though speeding may be a significant factor in predicting a car accident severity, it was decided that option (2) will be selected. It cannot be assumed to take the empty rows in the "SPEEDING" column as a non-speeding accident since there is insufficient data to make the assumption.

For the "UNDERINFL" attribute which is whether the driver was under the influence or drugs or alcohol, the data contained a mixture of "Y" or "1" for "Yes" and "N" or "0" for "No". To be consistent, all string values were changed to integers correspondingly. Any empty values were imputed using the most common class which was "1" representing the most accidents happening with the driver under the influence of alcohol or drugs.

The date and date-time columns were converted to Pandas date-times so that these values can be used to determine which day of the week the accident occur when selecting features.

For the weather conditions, road conditions and light conditions columns, any row with a column value of "Unknown" was also imputed using the most common class in that column.

A similar method was applied to all other categorical columns.

2.3 Feature Selection

After cleaning, the data-set contained 114 320 rows with 14 features. The first step was to determine which are the relevant attributes which can be used to predict the accident severity. There were several columns that have no meaning in predicting the accident severity. For example, the codes used to classify the intersections at which the accidents took place have no meaning in predicting accident severity. The "LOCATION" and "JUNCTIONTYPE" columns were removed because these columns contained similar information to the

"ADDRTYPE" column. The "STCOLDDESC" was also dropped because it contained similar information to the "COLLISIONTYPE" column. These columns were dropped which left the data-set with:

```
SEVERITYCODE      0
ADDRTYPE          0
COLLISIONTYPE     0
PERSONCOUNT       0
PEDCOUNT          0
PEDCYLCOUNT       0
VEHCOUNT          0
INCDATE           0
INCDTTM           0
UNDERINFL         0
WEATHER           0
ROADCOND          0
LIGHTCOND         0
HITPARKEDCAR      0
dayofweek         0
dtype: int64
```

Figure 2: Summary of column names and count of null values

It can be seen that no null values are included in the final dataframe. For the Feature data-set, a new column "dayofweek" was added which will be used to determine which days of the week an accident is most likely to be severe. The date and date-time columns will therefore not be used in the Feature set. The column "HITPARKEDCAR" was also removed because the data did not line up with the collision type. For example, many of the collision types were "Parked Car" but the "HITPARKEDCAR" column did not classify it as a "Y" for "Yes".

	SEVERITYCODE	ADDRTYPE	COLLISIONTYPE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	INCDATE	INCDTTM	UNDERINFL	WEATHER	ROADCOND	LIGHTCOND	HITPARKEDCAR	dayofweek
0	1	Block	Other	1	0	0	1	2013-03-27 00:00:00+00:00	2013-03-27 14:54:00	1	Clear	Dry	Dark - Street Lights On	N	2.0
1	1	Block	Parked Car	2	0	0	2	2006-12-20 00:00:00+00:00	2006-12-20 18:55:00	1	Clear	Dry	Daylight	N	2.0

Figure 3: Example of collision type correspondence with "Hit parked car" class

Finally, this will leave 11 features or attributes used to predict the accident severity. A sample of the final data-set is shown below.

	SEVERITYCODE	ADDRTYPE	COLLISIONTYPE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	INCDATE	INCDTTM	UNDERINFL	WEATHER	ROADCOND	LIGHTCOND	dayofweek
0	1	Block	Other	1	0	0	1	2013-03-27 00:00:00+00:00	2013-03-27 14:54:00	1	Clear	Dry	Dark - Street Lights On	2.0
1	1	Block	Parked Car	2	0	0	2	2006-12-20 00:00:00+00:00	2006-12-20 18:55:00	1	Clear	Dry	Daylight	2.0
2	1	Block	Other	1	0	0	1	2004-11-18 00:00:00+00:00	2004-11-18 10:20:00	1	Raining	Wet	Daylight	3.0
3	1	Block	Other	2	0	0	2	2013-03-29 00:00:00+00:00	2013-03-29 09:26:00	0	Clear	Dry	Daylight	4.0
4	1	Block	Parked Car	2	0	0	2	2004-01-28 00:00:00+00:00	2004-01-28 08:04:00	0	Overcast	Dry	Dark - Street Lights On	2.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
114315	2	Block	Angles	3	0	0	2	2012-09-22 00:00:00+00:00	2012-09-22 00:24:00	1	Raining	Wet	Daylight	5.0
114316	2	Block	Angles	2	0	0	2	2012-06-16 00:00:00+00:00	2012-06-16 18:01:00	1	Clear	Wet	Daylight	5.0
114317	2	Block	Head On	3	0	0	2	2012-05-28 00:00:00+00:00	2012-05-28 20:56:00	1	Clear	Dry	Daylight	0.0
114318	2	Intersection	Left Turn	3	0	0	2	2012-05-09 00:00:00+00:00	2012-05-09 10:09:00	1	Clear	Dry	Daylight	2.0
114319	2	Intersection	Cycles	2	0	1	1	2012-04-09 00:00:00+00:00	2012-04-09 18:37:00	1	Clear	Dry	Dusk	0.0

114320 rows × 14 columns

Figure 4: Final dataframe

3 Methodology

3.1 Exploratory Data Analysis

3.1.1 Relationship between Day of Week and Accident Severity

To verify that the day of the week has an impact on the accident severity, I plotted a histogram including this information as well as normalized the data to check the distributions between the two types of severity codes.

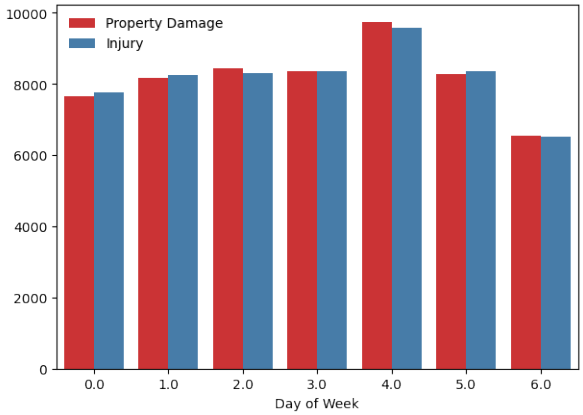

Figure 5: Count of day of week sorted by accident severity

From the histogram in Figure 5, it can be seen that the distributions between the severity types are balanced. This means that no particular day can determine whether an accident will more likely to be an "injury" type or "property damage" type severity. The plot only shows us on which day more accidents are likely to occur and this is not the target attribute that is needed. Therefore, the day of the week will not be a good attribute in determining the accident severity and will be dropped from the feature list.

3.1.2 Relationship between Address Type and Accident Severity

There are 3 types of addresses where accidents occur in the data-set: Alley, block and intersection. For the "Alley" and "Block" types, accidents are more likely to be of type "property damage". For the "Intersection" type, accidents are more likely to result in an "injury" type severity. This makes sense since intersections are more busy and traffic crosses which means accident impact is higher resulting in an injury.

```
ADDRTYPE      SEVERITYCODE
Alley         1              0.773743
              2              0.226257
Block         1              0.582155
              2              0.417845
Intersection  2              0.638052
              1              0.361948
Name: SEVERITYCODE, dtype: float64
```

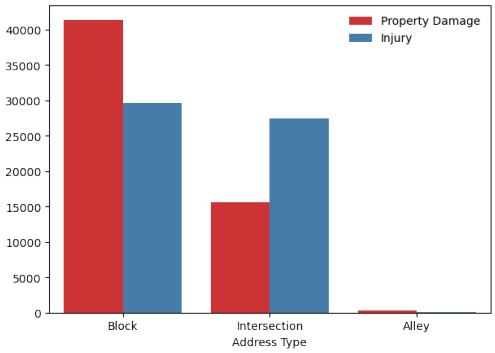

Figure 6: Histogram and normalized count of address type sorted by accident severity

3.1.3 Relationship between Collision Type and Accident Severity

There are 10 collision types included in the data-set and the distributions vary. 88% of the "Parked Car" collisions result in property damage while only 12% result in an injury. 63% of the "Rear Ended" collisions result in an "injury" type severity while 37% result in a "property damage" type severity. The collision types that are more likely to result in property damage include: Parked car, sideswipe, other and right turn. This means the majority of the collision types are likely to result in some form of an injury. The accident severity distributions between each collision type indicate that collision type is a good feature for predicting the accident severity. Collisions involving pedestrians and cycles should result more in an injury because people are less protected physically when riding a cycle or walking. Collisions types that are "Rear Ended" , "Angles", "Left Turn" and "Head On" are more likely to result in an "injury" since a moving cars that collide have more impact.

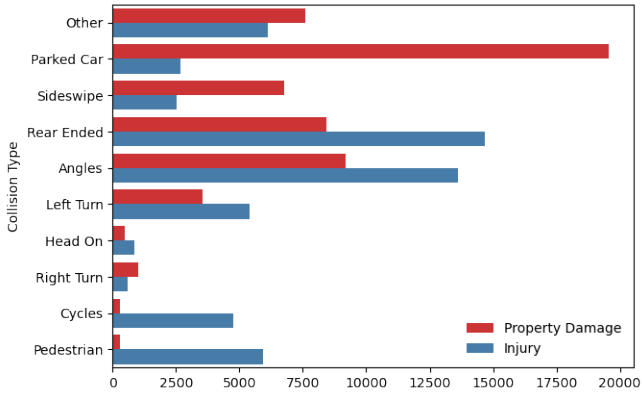

Figure 7: Count of collision type sorted by accident severity

3.1.4 Relationship between Alcohol or Drug Influence and Accident Severity

It is hypothesised that a driver who is under the influence of alcohol or drugs may cause a more severe "injury" type accident than a driver who is not. This is because an intoxicated driver is not in a mental state to make safe choices on the road in comparison to a sober driver who may be able to make choices to avoid accidents or at least reduce the impact. The histogram in Figure 8 shows this distribution.

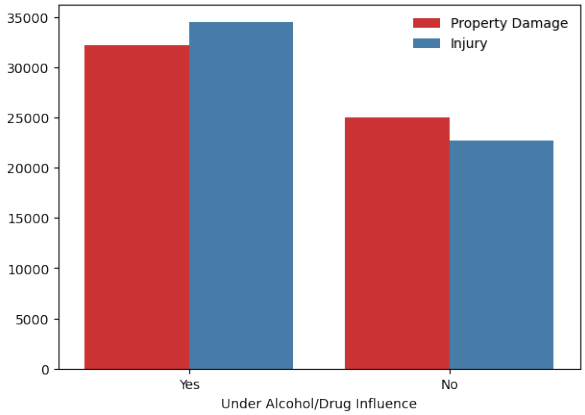

Figure 8: Count of alcohol or drug influence sorted by accident severity

3.1.5 Relationship between Weather and Accident Severity

There are 10 weather condition types included in the data-set and the distributions of accidents lean more toward "Clear", "Raining" and "Overcast" weather types. More "injury" type accidents are likely to occur

during rain and overcast weather types while for clear weather conditions, property damage is more likely to occur. This is valid since rainy and overcast conditions call for wetter roads and poor visibility. Some of the weather types such as "Sleet/Hail/Freezing Rain" and " Fog/Smog/Smoke" have a balanced distribution between the two severity types and will be dropped when encoding the feature set.

3.1.6 Relationship between Road Conditions and Accident Severity

The road conditions have a more varying distribution and even with condition similarities, the accident severity types may differ. For example, a "Wet" road condition results in more accidents being of type "injury" severity while "Standing Water" or "Ice" road condition result in more accidents being of type "property damage". The "Sand/Mud/Dirt" road condition will also be removed when encoding since the accident severity distribution is balanced.

3.1.7 Relationship between Light Conditions and Accident Severity

Generally darker lighting conditions result in an "injury" type accident severity. The "daylight" conditions, however, have a balanced distribution of accident severity types and will be dropped when encoding.

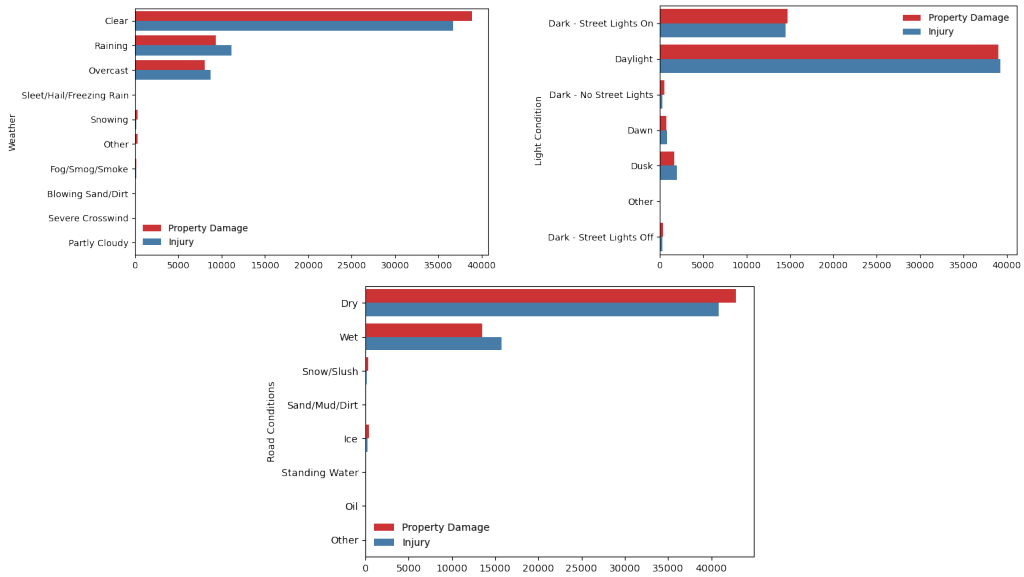

Figure 9: Counts of weather, light conditions and road conditions in accident

3.1.8 Relationship between Person,Vehicle,Pedestrian and Cyclist count and Accident Severity

It can be deduced that the more vehicles that are involved in an accident, the more likely an injury is to occur. Similarly if more than 5 people are involved in an accident, the likelihood of an "injury" type severity increases. The same hypothesis can be made for the number of pedestrians and cyclists involved in an accident.

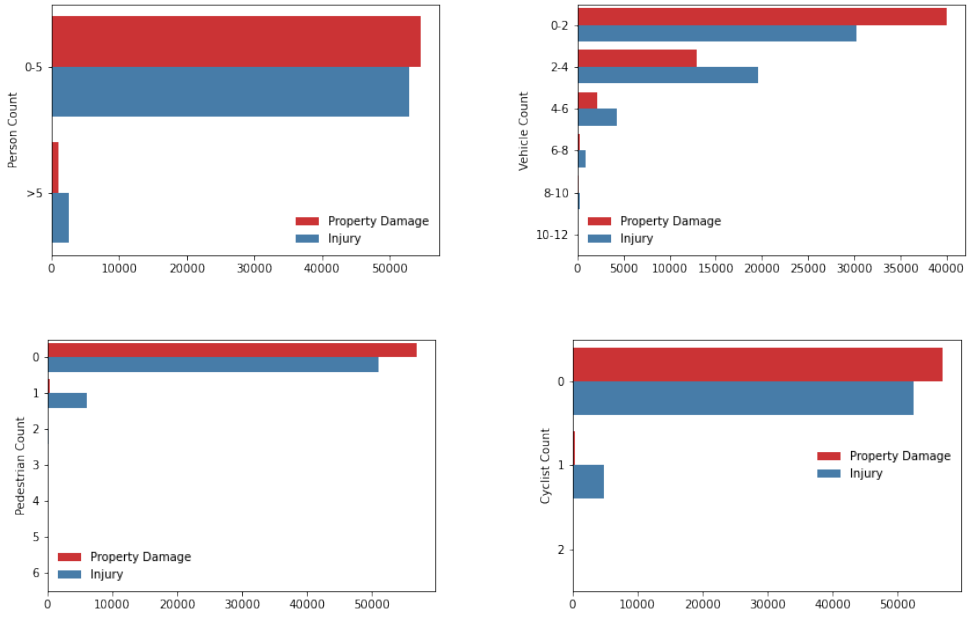

Figure 10: Counts of persons, vehicles, pedestrians and cyclists involved in accident

3.2 Predictive Modelling

The outcome of the modelling is to classify in which severity category (Property damage/Type 1 or Injury/Type 2) an accident may fall into based on labeled data. Hence, supervised machine learning using classification will be used. Four different methods will be used and compared against each other to determine which performs the best and can be deployed. The following algorithms will be used and are generally good for binary classification:

- K Nearest Neighbor(KNN)
- Decision Tree
- Support Vector Machine
- Logistic Regression

3.2.1 One Hot Encoding

Many of the attributes in the feature set need to be changed to numerical values for the machine learning models. Machine learning models better understand numerical data in comparison to categorical data and this is why encoding is necessary. The person count attribute was enhanced into two types: (1) Person Count within 0-5 people range (2) Person Count greater than 5 people. All attributes with the "Other" type were dropped since these do not provide much information about the accident. After encoding, 35 attributes were left for the feature set. Since only two severity codes were present in the data-set, this results in binary classification modelling. The label or target variable is the severity code which can be either 1 for "property damage" or 2 for "injury".

	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	UNDERINFL	PersonGroup 0-5	PersonGroup >5	Alley	Block	Intersection	Angles	...	Dry	Ice	Oil	Snow/Slush	Standing Water
0	0	0	1	1	1	0	0	1	0	0	...	1	0	0	0	0
1	0	0	2	1	1	0	0	1	0	0	...	1	0	0	0	0
2	0	0	1	1	1	0	0	1	0	0	...	0	0	0	0	0
3	0	0	2	0	1	0	0	1	0	0	...	1	0	0	0	0
4	0	0	2	0	1	0	0	1	0	0	...	1	0	0	0	0

Figure 11: Sample of feature set after One Hot Encoding

4 Results and Discussion

4.1 K Nearest Neighbor (KNN)

The KNeighborsClassifier from Sklearn was used for modelling. The data was split into 10% test data and 90% training data with a random state of 4. This left 11432 test samples and 102888 training samples. In order to maximise the accuracy, the accuracy score was plotted against varying numbers of neighbors. This resulted the best accuracy of 69% with k=13 neigbors.

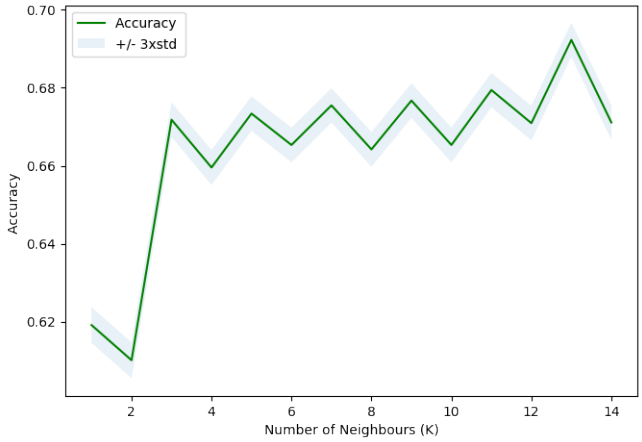

Figure 12: Plot of accuracy score against varying number of neighbors

After training the model and testing it on the test set, a confusion matrix was plotted to determine the number of true positives (number of accidents correctly predicted as "injury" or 2) and the number of true negatives (number of accidents correctly predicted as "property damage" or 1). 4217 samples were correctly predicted as type 2 which is 74% of the total number of type 2 accidents in the test set. 3697 samples were correctly predicted as type 1 which is 65% of the total number of type 1 accidents in the test set.

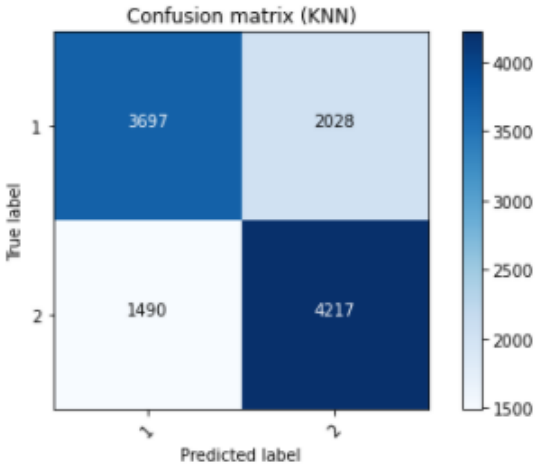

Figure 13: Confusion matrix for KNN model

4.2 Decision Tree

The DecisionTreeClassifier from Sklearn was used to model the decision tree. The data was split into 30% test data and 70% training data with a random state of 3. This left 34296 test samples and 80024 training samples. Different max depth parameters were used to find the best accuracy which resulted in an accuracy of 70% for a max depth parameter of 13. A sample of the decision tree can be seen below. The entire decision tree was too large to fit.

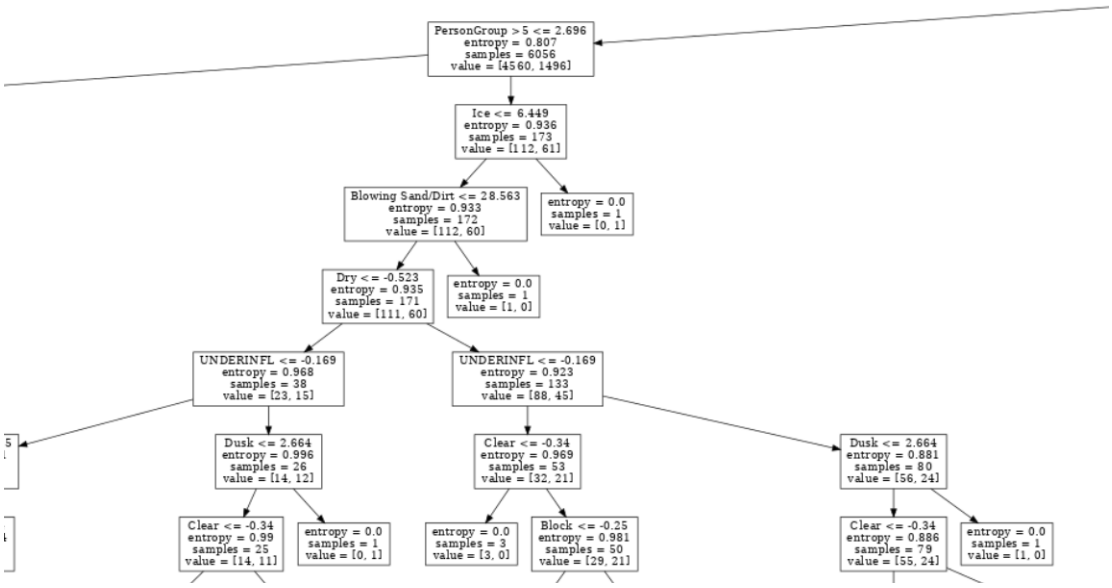

Figure 14: Decision Tree for predicting accident severity

A confusion matrix was plotted for the decision tree. 13345 samples were correctly predicted as type 2 which is 78% of the total number of type 2 accidents in the test set. 10769 samples were correctly predicted as type 1 which is 62% of the total number of type 1 accidents in the test set.

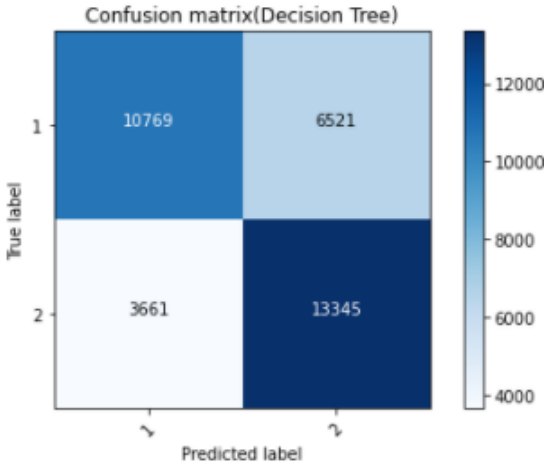

Figure 15: Confusion matrix for Decision Tree model

4.3 Support Vector Machine (SVM)

The svm classfifier from Sklearn was used to model the data. The data was split into 10% test data and 90% training data and trained using the 'rbf' kernel. This left 11432 test samples and 102888 training samples just as the KNN model used. The accuracy of the SVM model was 71%. A confusion matrix was plotted for the SVM model. 4451 samples were correctly predicted as type 2 which is 78% of the total number of type 2 accidents in the test set. 3653 samples were correctly predicted as type 1 which is 64% of the total number of type 1 accidents in the test set.

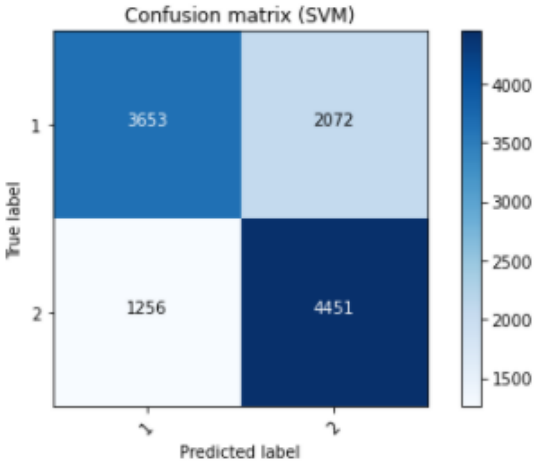

Figure 16: Confusion matrix for SVM model

4.4 Logistic Regression

The LogisticRegression classfifier from Sklearn was used to model the data. The data was split into 20% test data and 80% training data, and trained using the 'liblinear' solver and an inverse regularization strength of 0.01. This left 22864 test samples and 91456 training samples. The accuracy of the Logistic Regression model was 70%. A confusion matrix was plotted for the Logistic Regression model. 9299 samples were correctly predicted as type 2 which is 81% of the total number of type 2 accidents in the test set. 6664 samples were correctly predicted as type 1 which is 58% of the total number of type 1 accidents in the test set.

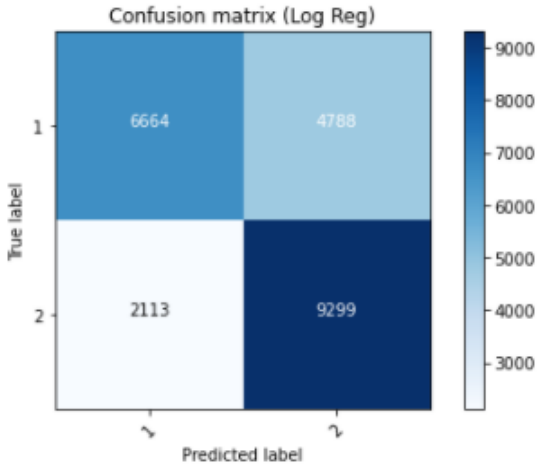

Figure 17: Confusion matrix for Logistic Regression model

4.5 Summary

The overall performance of the models were good. The classifer that performed the best was the SVM model with a 71% accuracy and the Decision Tree performed similarly with an accuracy of 70%. The worst performing model was the KNN model with a 69% accuracy. However,there were no large differences in performance with approximately 0.5% between the sequential models. Some models like the Decision Tree and Logistic Regression algorithms also used a higher test sample size which may have affected the performance.

Table 1: Performance of classification models

Algorithm	**KNN**	**Decision Tree**	**SVM**	**Logistic Regression**
Jaccard	0.6923	0.7031	0.7089	0.6981
F1	0.6916	0.7021	0.7074	0.6940
Log Loss	NA	NA	NA	0.5483
True Positives	4217	13345	4451	9299
True Negatives	3697	10769	3653	6664
False Positives	2028	6521	2072	4788
False Negatives	1490	3661	1256	2113

The Receiver Operating Characteristic (ROC) curves were plotted for each model as shown in Figure 18. The ROC curve represents the trade-off between the false positive rate (FPR) and true positive rate (TPR). The FPR and TPR represents the probability that the model will predict a class as its correct class. The curves that lie towards the top left have better performing models since this would indicate a probability closer to 1 or 100% (represents the ideal model). This confirms the results as before where the SVM model is the most accurate of all four models. Conversely, the closer the curve lies to the 45 degree "no prediction" dotted line, the less accurate the model is and this can be seen with the KNN model.

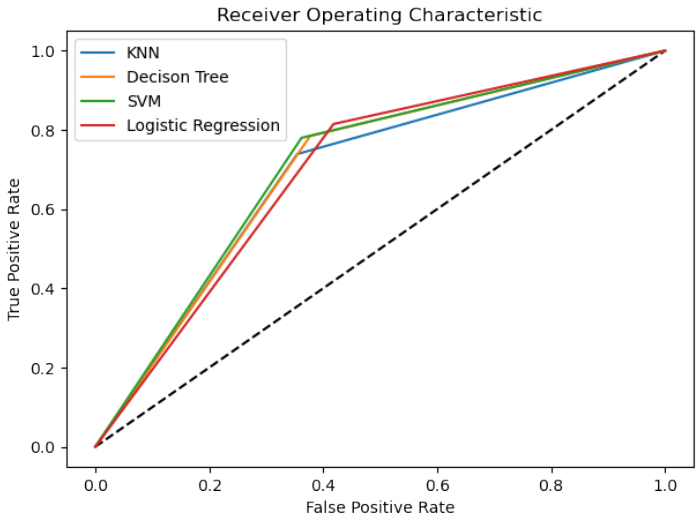

Figure 18: ROC curves for the classification model

5 Conclusions

In this project, the prediction of traffic accident severity using classification was investigated. Various attributes such as collision type, address type, weather, light and road conditions, alcohol or drug influence, and person, vehicle, pedestrian and cyclist count were used to predict the outcome of the accident in terms of severity (Type 1 for "property damage" and Type 2 for "injury"). Four different machine learning models were used to predict or classify the accident severity, namely: K Nearest Neighbor, Decision Tree, Support Vector Machine and Logistic Regression. The SVM model performed the best with an accuracy of 71%. These models can be useful in determining the outcomes of accidents for traffic safety institutions. For example, if a particular day has rainy weather, more safety measures and patrols can be put in place to ensure safer roads. Insurance companies can also warn drivers of severe weather storms that may prompt road users to drive more carefully. Additionally, road users can plan their traffic routes using the same information.

6 Future Recommendations

Although the models performed relatively well, averaging a 70% accuracy, improvements can be made. Firstly, more attributes can be included in the feature set such as "Speeding" which is a significant indicator for accident severity. The models generally tended to predict the outcome of "injury" better than "property damage" and more attributes can be added to the feature set to improve the predictions. Another improvement could be using data with different types of severity outcomes which may create more variance. Currently, there is significant overlap in predicting whether an accident severity is only "property damage" or "injury".

The models performance could also generally be improved by tuning the parameters within the models more and experimenting with different training and test sizes. Hence, if optimized, significant improvements can be made. To properly test the models, a different test set of data could be used as well.

7 References

1 https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

2 https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

3 https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf