Prediction of severity of road accidents using multiclass classification

Lakshika Girihagama

1. Introduction

How to predict accurately severity of road collisions ahead remains one of the key questions for many. People often see road accidents with variable severity including property damages to injuries and ultimately unfortunate fatalities. Severity of road collisions can be influenced by various factors including personal behaviors (e.g. drugs under influence-DUI, speeding), environmental conditions (e.g. rain), and technological factors (e.g. car breakdown). Accurate prediction of severity of road accidents and their influencing factors can help people and authorities to adapt precautionary safety measures so that avoiding the risk may be possible. On the other hand, based on severity, it also helps authorities to efficiently allocate resources during an accident and eventually minimize the after costs associated with road collisions. Therefore, to potentially mitigate the road collisions and their severity, it is critical to accurately predict severity of road collisions and the contributing factors. Hence, in this analysis, I address the question of how to accurately predict severity of road accident using factors such as whether or not a driver involved was under the influence of drugs or alcohol, weather conditions during the time of the collision, the condition of the road during the collision, whether or not speeding was a factor in the collision, and the light conditions during the collision.

2. Data acquisition and preprocessing

2.1. Data source

The data set contains the all types of road collisions recorded in the city of Seattle. The data were provided by the Seattle Police Department (SPD) that was recorded by Traffic Records (Data Source: gisdata.seattle.gov). This data set includes 221266 records of all types of collisions, their severity, various sub-categories of conditions (37) involving the accident. The time span for this data set ranges from 2004 to Present. The data record is updated weekly.

2.2 Data preprocessing

The data set was recorded as Comma Separated Values (CSV) format and it was loaded to Jupyter Notebook to convert into a pandas dataframe for easy analysis. The definition of severity of accidents given by the are as follows (Table 1).

Table 1: The code that corresponds to the severity of the collision (Data source: gisdata.seattle.gov).

Code	Description
0	Unknown
1	Property damage only
2	Injuries

2b	Serious injuries
3	Fatalities

2.2.1. Redundant Data

There were subcategories with redundant data for prediction of severity of road accidents. For instance, descriptions of the accident, technical codes used by the SPD were to name a few. Therefore, columns with redundant data were removed. As a result, elements in the data frame were reduced for inexpensive computational capability.

2.2.2 Missing values

There were missing values corresponds to weather, road conditions, light conditions which cannot be replaced by any statistical means. Therefore, entries correspond to missing values in those columns were removed. As an example, weather conditions during the time of the collision record contains 26420 missing values (about 12% of the total) compared to 194969 to available weather condition records.

2.2.3 Redefining the accident severity code

The code that corresponds to the severity of the collision is categorized into five classes, namely, 0: unknown, 1: property damage, 2: injuries, 2b: serious injuries, 3: fatalities. The value count of severity code revealed that there were less than 2% serious injuries and less than 1% of fatalities. Further, the unknown class didn't provide predictive information such that, removal of the entries corresponds to unknown severities were performed. As a result, the severity code was redefined into four major classes where 1 represents the property damage and 2 represents the personal injury, 3 repesents the serious injury, and 4 represents the fatality (Table 2).

New Severity Code	Description
1	Property damage only
2	Injuries
3	Serious injuries
4	Fatalities

2.2.3 Convert datetime object

Next, the date time object is converted so, the day of week is extracted as a subcategory which may be crucial to understand whether there is a pattern associated with road collisions with what day of the week.

2.2.3 Convert mixed data into 0 or 1

Further, the columns that contains mixed data such as “Y”, “N”, “0”, “1” entries were converted to 0 or 1. For instance, entries in the UNDERINFL column have “Y” and others have “1” and vice versa. Therefore, all the entries were converted to 0 or 1. Also, it is assumed that blank entries in SPEEDING, PEDROWNOTGRNT, and HITPARKEDCAR are as “0” because it was only recorded when there was a positive effect on the accident.

2.3. Convert categorical data into numerical data

One-hot encoding approach is performed to convert categorical data so the machine learning models can be applied for accurate prediction of severity of accidents. For instance, weather data contained 11 subcategories such as whether it was clear, rainy, partly cloudy etc. Here, I used pandas pd.get_dummies function to create columns of weather conditions with 0s and 1s. However, subcategories named as ‘unknown’ and ‘other’ were removed. Similarly, one-hot encoding was performed for ROADCOND, LIGHTCOND, ADDRTYPE, and SDOT_COLCODE. similarly, redundant information was removed from those subcategories.

2.3 Balancing data

From the distribution of number of accidents-based severity code reveals that there were less than 0.1% of fatalities, less than 2% of serious injuries, ~30% of injuries and more than 60% of property damage. Because of this imbalance in distribution, the machine learning models outcomes will be biased. To overcome this issue, down sampling of majority class was performed. In that case, number of samples similar to SEVERITYCODE=4 were randomly picked for SEVERITYCODE=1,2,3 such that, dataframe size was significantly reduced. Also, column data contain variable units. Thus, standardization of columns was performed to obtain zero mean and unit variance.

2.4. Feature selection

One of the important steps to obtain an accurate predictive model is select appropriate features to avoid overfitting or underfitting of the model. By looking at each subcategory, it is evident that some features are redundant. Table 3 summarizes the final feature set that was used for the predictive models in this analysis. The target variable is the SEVERITYCODE.

Table 3. The selected feature categories for machine learning model for multiclass classification.

Feature	Description
X, Y	Latitude and Longitude of the collision
ADDRTYPE	Collision address type: • Alley • Block • Intersection
PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT	The total number of people involved in the collision, the number of pedestrians involved in the collision, the number of bicycles

	involved in the collision, and the number of vehicles involved in the collision.
INATTENTIONIND, UNDERINFL, PEDROWNOTGRNT	Whether or not collision was due to inattention, whether or not a driver involved was under the influence of drugs or alcohol, and whether or not the pedestrian right of way was not granted.
WEATHER	A description of the weather conditions during the time of the collision.
ROADCOND	The condition of the road during the collision.
SDOT_COLCODE	
LIGHTCOND	The light conditions during the collision.
SPEEDING	Whether or not speeding was a factor in the collision
dayofweek	The day of the week

The final feature set contains 76 subcategories for the machine learning models.

Methodology

The analysis will be using various multiclass classification algorithms such as KNN, decision tree, SVM, and Logistic regression analysis to provide better prediction. The best method will be chosen based on the model's accuracy evaluated using Jaccard index, F1-score, and Log Loss. As the first step before modelling, the data set was divided into training and the test set. Only 30% of randomly selected data points were used as the test set. The final step in this process is standardizing the feature set so that it gives zero mean and unit variance.

3. Results and Discussion

3.1. Data visualization.

Before the application of machine learning models, it is useful to understand some of data distribution. In such, I plotted the number of accidents associated with each subcategory that will be used in the machine learning models. The majority of the severity resulting from accidents can be categorized as property damage which is accounted for 68% of total accidents (**Figure 1**). Injuries are accounted for ~30%, serious injuries are accounted for ~1.6%, and fatalities are accounted for ~0.2% of the total accidents.

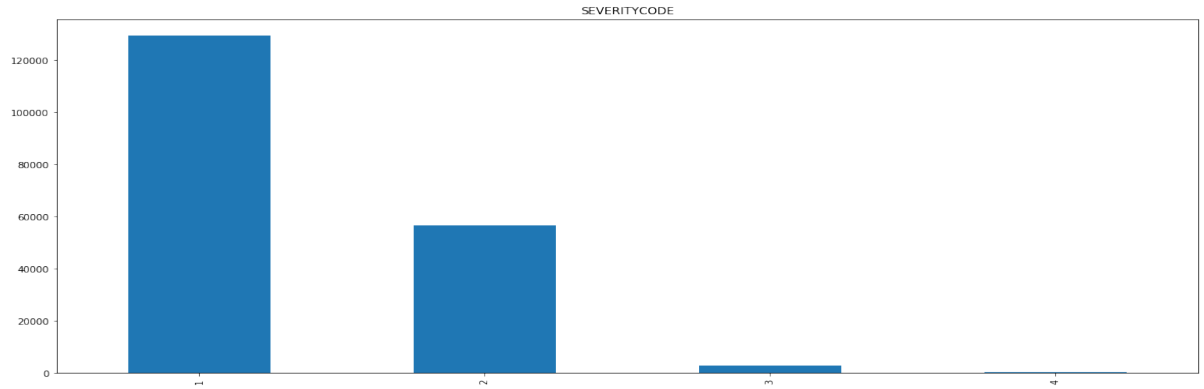


Figure 1. Number of accidents are plotted against the severity code. While code 1 represents the property damage, the code 2, 3, and 4 represent the injuries, serious injuries, and fatalities, respectively.

It is also evident that the majority of accidents occur on Fridays and minimum occurs Mondays (**Figure 2**).

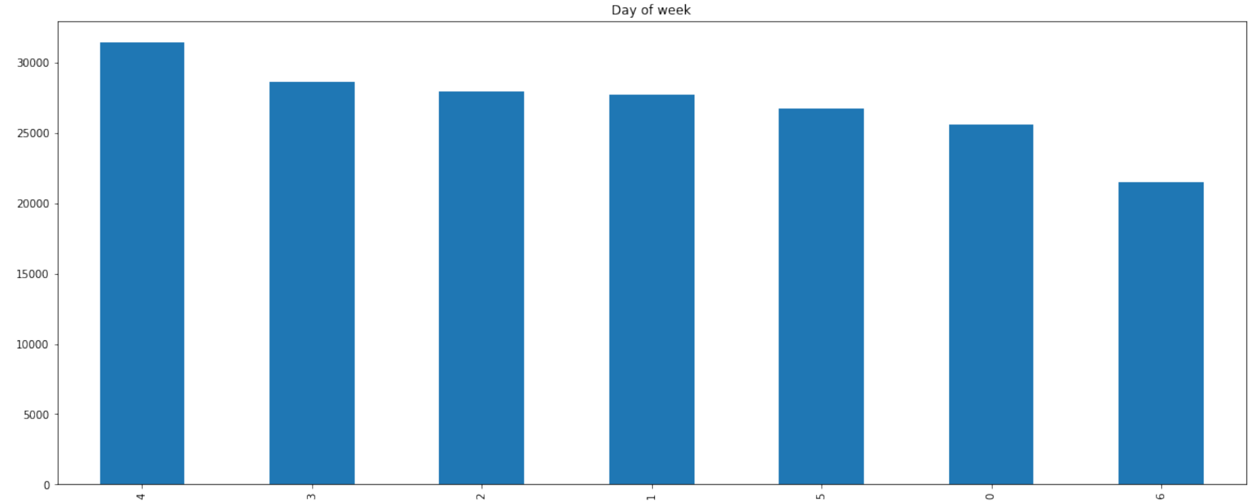


Figure 2. Distribution of number of accidents with respect to the day of week. 0: Monday, 1: Tuesday, 2: Wednesday, 3: Thursday, 4: Friday, 5: Saturday, and 6: Sunday.

Distribution of accidents based on the type of location reveals that the majority of accidents occur at blocks compared to that of intersections (**Figure 3**).

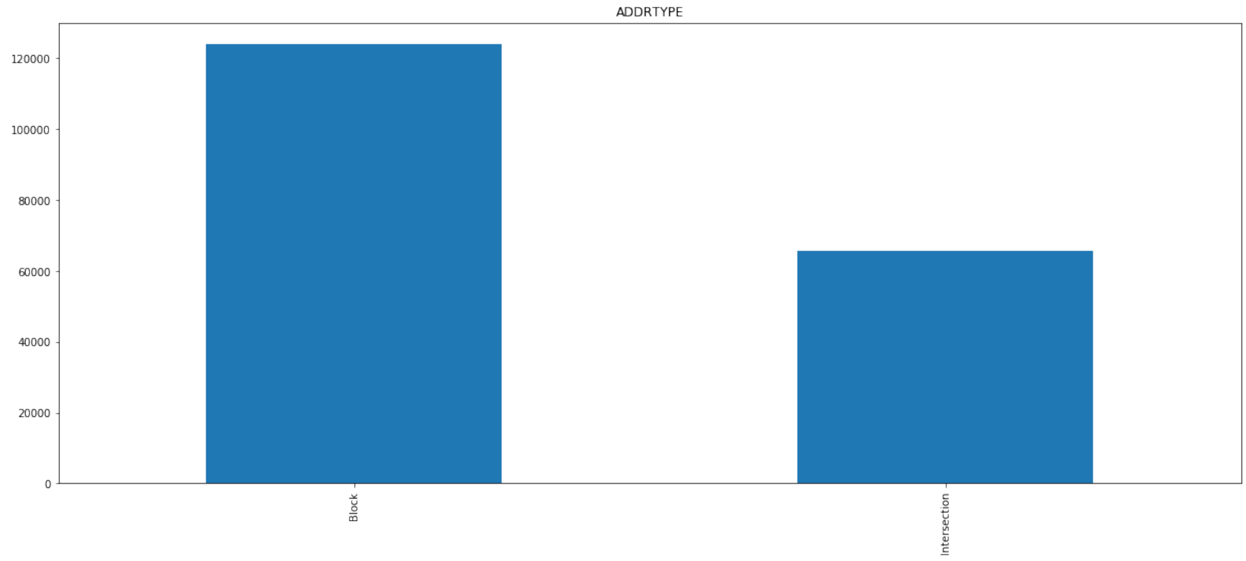


Figure 3. Distribution of accidents plotted against the address type.

The effect of environmental conditions such as weather conditions, light conditions, and road conditions on the road accidents during last 16 years are shown in **Figure 4**. The effect of weather conditions show that majority of accidents occur when there was clear weather, but significant amount of accidents occurs when it was raining or overcast. While wet road conditions affect for significant amount of accidents, the majority of accidents record dry road conditions. Although limited lighting (although with streetlights on) account for number of accidents, the majority of accidents were happened during daylight.

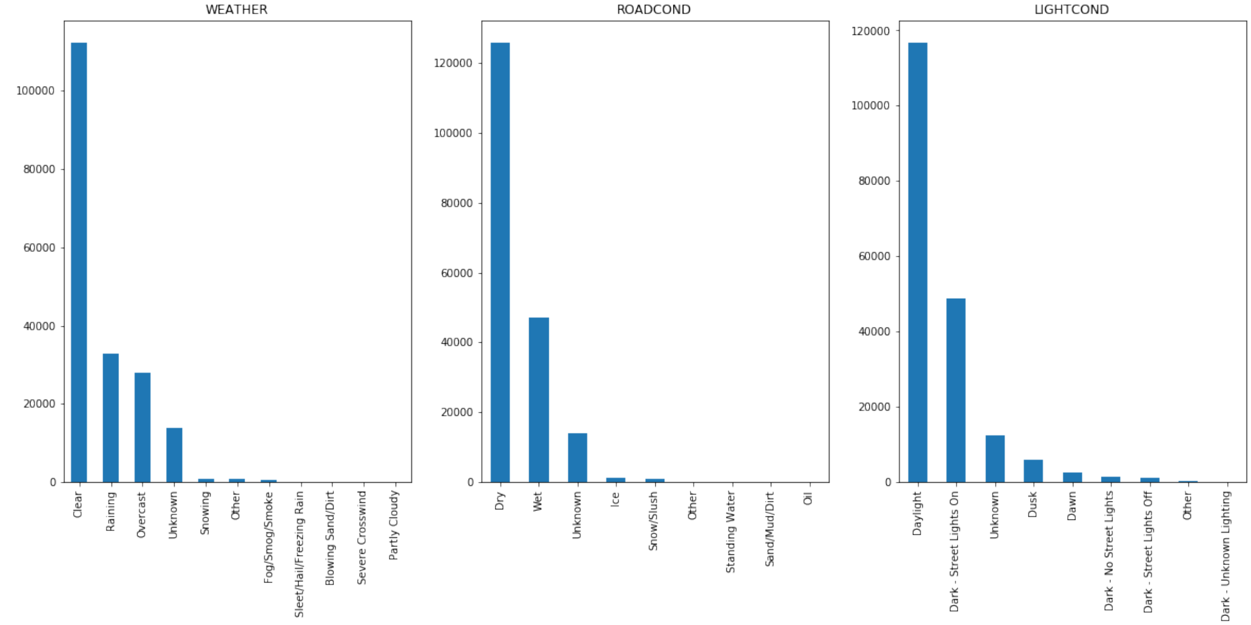


Figure 4. The effect of environmental conditions on road accidents.

The personal behaviors such as whether or not collision was due to inattention, whether or not a driver involved was under the influence of drugs or alcohol, and whether or not the pedestrian right of way was not granted or speeding show that majority of accidents have negative response from bad personal behaviors (**Figure 5**).

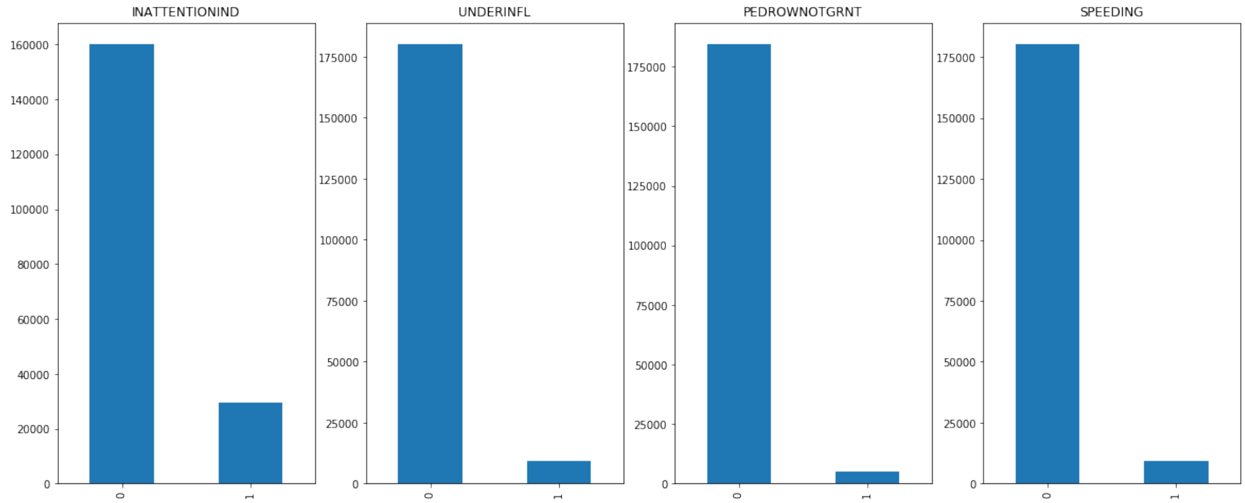


Figure 5. Number of accidents based on personal behaviors.

3.2 Machine Learning Models

Now we examine the outcomes from the machine learning algorithms that was used in this analysis for multiclass classification of severity of road accidents. Jaccard index, F1-score, and LogLoss were used to compare the algorithms that provides the best prediction.

3.2.1. KNN model

The K-Nearest Neighbor model was built using scikitlearn KNeighborsClassifier model. The best accuracy is acquired when the K value reaches to 28 (**Figure 6)** with an accuracy of 0.51. The Jaccard score and F1-score for KNN model are 0.513 and 0.507, respectively. The confusion matrix for KNN model is shown below (Figure 7).

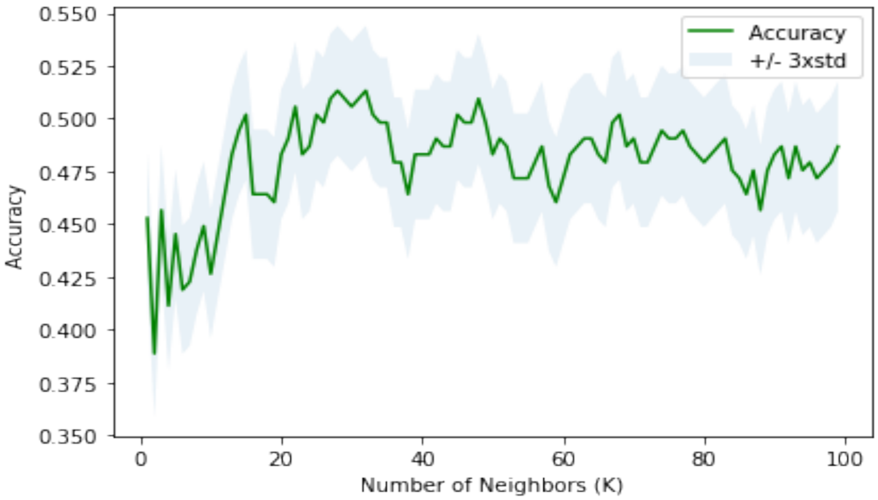


Figure 6. KNN model accuracy with respect to variable K neighbors. The highest accuracy of 0.51 was achieved when K=28.

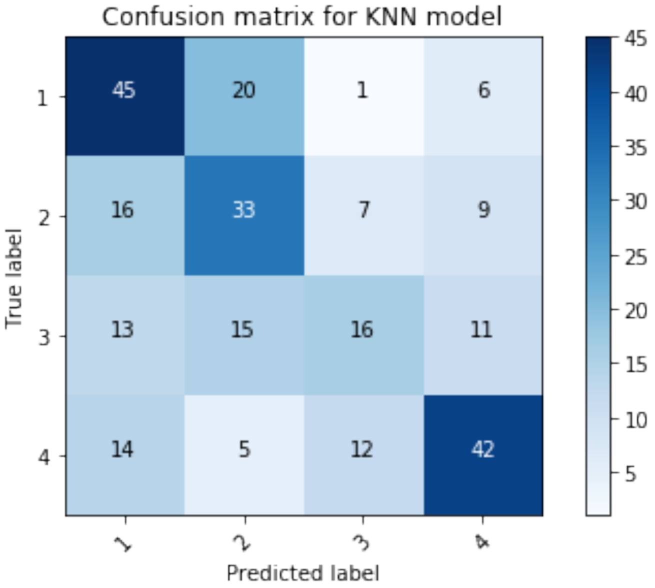


Figure 7. The confusion matrix for severity index acquired from KNN model. the severity code is defined as four major classes where 1 represents the property damage and 2 represents the personal injury, 3 represents the serious injury, and 4 represents the fatality.

3.2.2. Decision Tree model

The goal to create severity code (target variable) by learning simple decisions aquaired from similar feature data. For this analysis, scikitlearn DecisionTreeClassifier was used with entropy criterion. The maximum depth of the tree was set to 6. The Jaccard accuracy and F1 score are 0.44 and 0.43, respectively. The confusion matrix for decision tree is shown below (**Figure 8**).

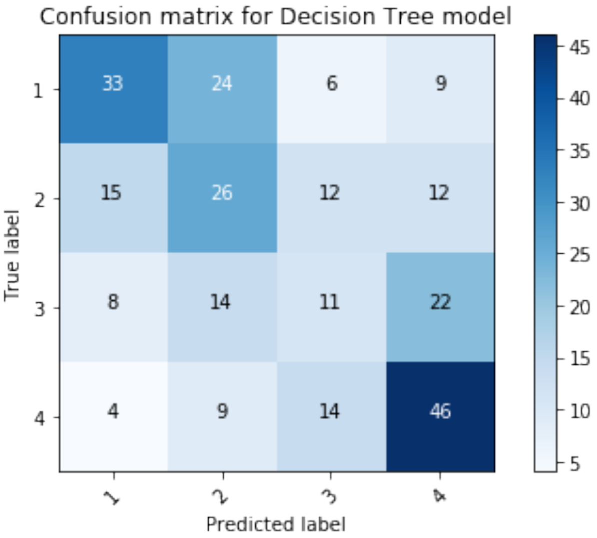


Figure 8. Confusion matrix for severity code for road accidents obtained from decision tree classification model.

3.2.3. Support Vector Machine (SVM) model

The objective of the SVM method is to find a hyperplane that distinctively classifies the severity code for road accidents. In this method, scikitlearm svm model is used. However, SVM model is preliminary desined for binary classification. Therefore, for multiclass classification, one-vs-one method was used where decision_function_shape='ovo'. The model predicted Jaccard accuracy and F1 score are 0.52 and 0.519, respectively. The confusion matrix for SVM model is shown below (**Figure 9**).

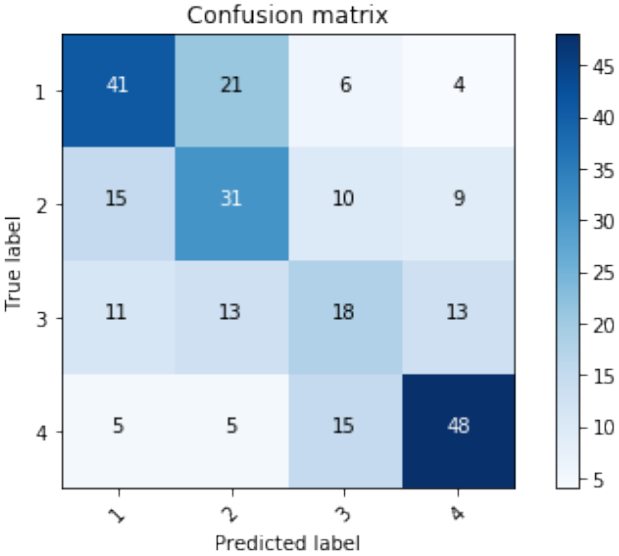


Figure 8. Confusion matrix for severity code for road accidents obtained from SVM multiclass classification model.

3.2.4. Logistic Regression model

Similar to SVM, Logistic regression model is primarily designed for binary classification. However, for multiclass classification with Logistic Regression, the training algorithm used the one-vs-rest (OvR) scheme with 'lbfgs' solver. The predicted Jaccard accuracy, F1 score, and logloss were 0.52, 0.51 and 1.17, respectively. The confusion matrix for logistic regression is given below (**Figure 9**).

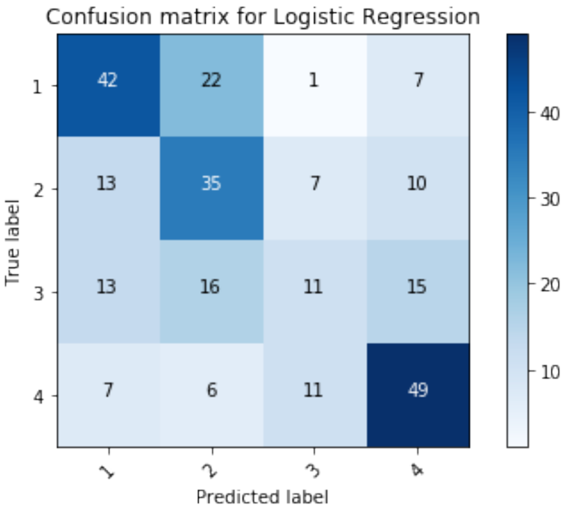
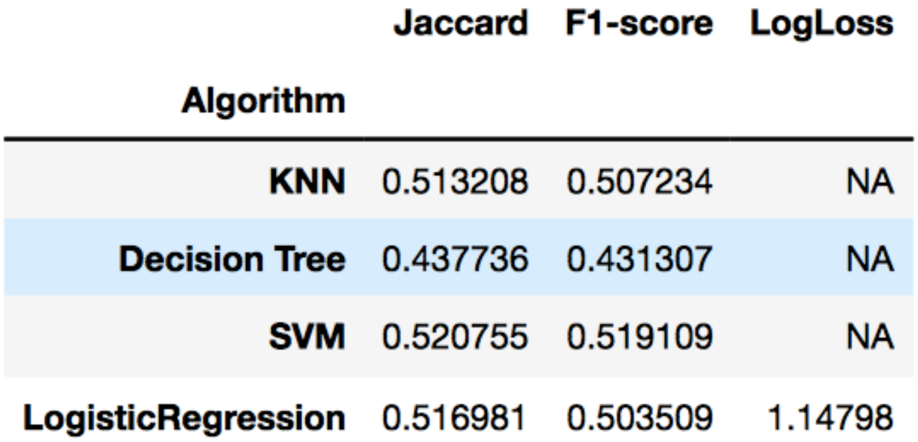


Figure 9. Confusion matrix for severity code for road accidents obtained from Logistic Regression multiclass classification model.

3.3 Evaluation of the predictive models

The summary of accuracy measured from various schemes such as Jaccard accuracy, F1 score, and LogLoss are shown below (Table 4). From the analysis it is evident that the Decision Tree model poorly predicts (~43%) the severity of road collision compared to that of KNN (~51%), SVM (~52%), and Logistic regression models (~51%).

Table 4. Accuracies obtained from KNN, SVM, Decision Tree, and Logistic Regression multi classification models.



	Jaccard	F1-score	LogLoss
Algorithm			
KNN	0.513208	0.507234	NA
Decision Tree	0.437736	0.431307	NA
SVM	0.520755	0.519109	NA
LogisticRegression	0.516981	0.503509	1.14798

4. Conclusion

In this study, accidents recorded by the Seattle Police Department were used to predict severity of accident when happened. To predict the road accident severity, I used multiclass classification using few machine learning algorithms such as K-nearest neighbor, decision tree, Support Vector

Machine and Logistic Regression. The model accuracies were evaluated using Jaccard accuracy, F1-score and LogLoss. This is a useful tool for authorities during an accident such that they can use this model to allocate resources depending on the severity of accidents and if possible, take precautionary safety measures before it happens. Based on model predictions acquired in this study, it showed that Decision Tree model predicts poorly compared to the other models. However, overall models were able to obtain approximately 52% of accuracy. In other words, probability of prediction can be true or false will be half. This may be due to multiclass classification. Hence, for future direction I would like combine severity data into two classes such binary classification may help to improve the models' accuracy.