Capstone Project

Weigang Wei

1 Introduction

1.1 Business problem

It is aimed to build a model to predict the severity of an accident based on a database which contains a set of accident severity data. This model is prepared for the local transportation department and this model may be used to set up diverted routes, and temporary speed limits etc.

2 Data

2.1 General description of the dataset

The dataset contains The general information of the dataset is shown below.

```
#   Column        Non-Null Count   Dtype
---  ------          --------------   -----
 0   SEVERITYCODE    194673 non-null  int64
 1   X               189339 non-null  float64
 2   Y               189339 non-null  float64
 3   OBJECTID        194673 non-null  int64
 4   INCKEY          194673 non-null  int64
 5   COLDETKEY       194673 non-null  int64
 6   REPORTNO        194673 non-null  object
 7   STATUS          194673 non-null  object
 8   ADDRTYPE        192747 non-null  object
 9   INTKEY          65070 non-null   float64
 10  LOCATION        191996 non-null  object
 11  EXCEPTRSNCODE   84811 non-null   object
 12  EXCEPTRSNDESC   5638 non-null    object
 13  SEVERITYCODE.1  194673 non-null  int64
 14  SEVERITYDESC    194673 non-null  object
 15  COLLISIONTYPE   189769 non-null  object
 16  PERSONCOUNT     194673 non-null  int64
 17  PEDCOUNT        194673 non-null  int64
 18  PEDCYLCOUNT     194673 non-null  int64
 19  VEHCOUNT        194673 non-null  int64
 20  INCDATE         194673 non-null  object
 21  INCDTTM         194673 non-null  object
 22  JUNCTIONTYPE    188344 non-null  object
 23  SDOT_COLCODE    194673 non-null  int64
 24  SDOT_COLDESC    194673 non-null  object
```

```
25  INATTENTIONIND  29805 non-null   object
26  UNDERINFL       189789 non-null  object
27  WEATHER         189592 non-null  object
28  ROADCOND        189661 non-null  object
29  LIGHTCOND       189503 non-null  object
30  PEDROWNOTGRNT   4667 non-null    object
31  SDOTCOLNUM      114936 non-null  float64
32  SPEEDING        9333 non-null    object
33  ST_COLCODE      194655 non-null  object
34  ST_COLDESC      189769 non-null  object
35  SEGLANEKEY      194673 non-null  int64
36  CROSSWALKKEY    194673 non-null  int64
37  HITPARKEDCAR    194673 non-null  object
```

The aim of this project is to predict the severity which is described by the severity code. The description of the codes are:

- Code 1: Property Damage Only Collision, and
- Code 2: Injury Collision.

The number of different severity is shown in Figure 1.

```
Column is:  SEVERITYCODE
1    136485
2     58188
Name: SEVERITYCODE, dtype: int64
```

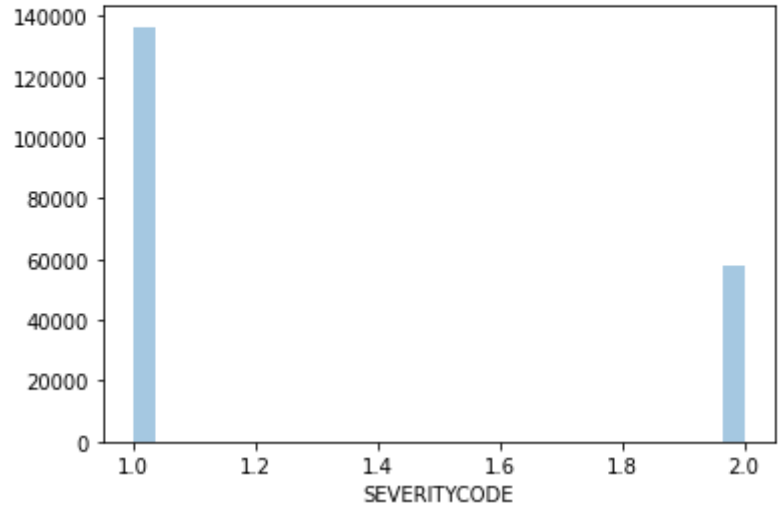


Figure 1: Number of different severity in the data set

19 number columns contains NaN value, and the percentage of NaN in 7 columns exceeds 30%, as shown in Figure 2.

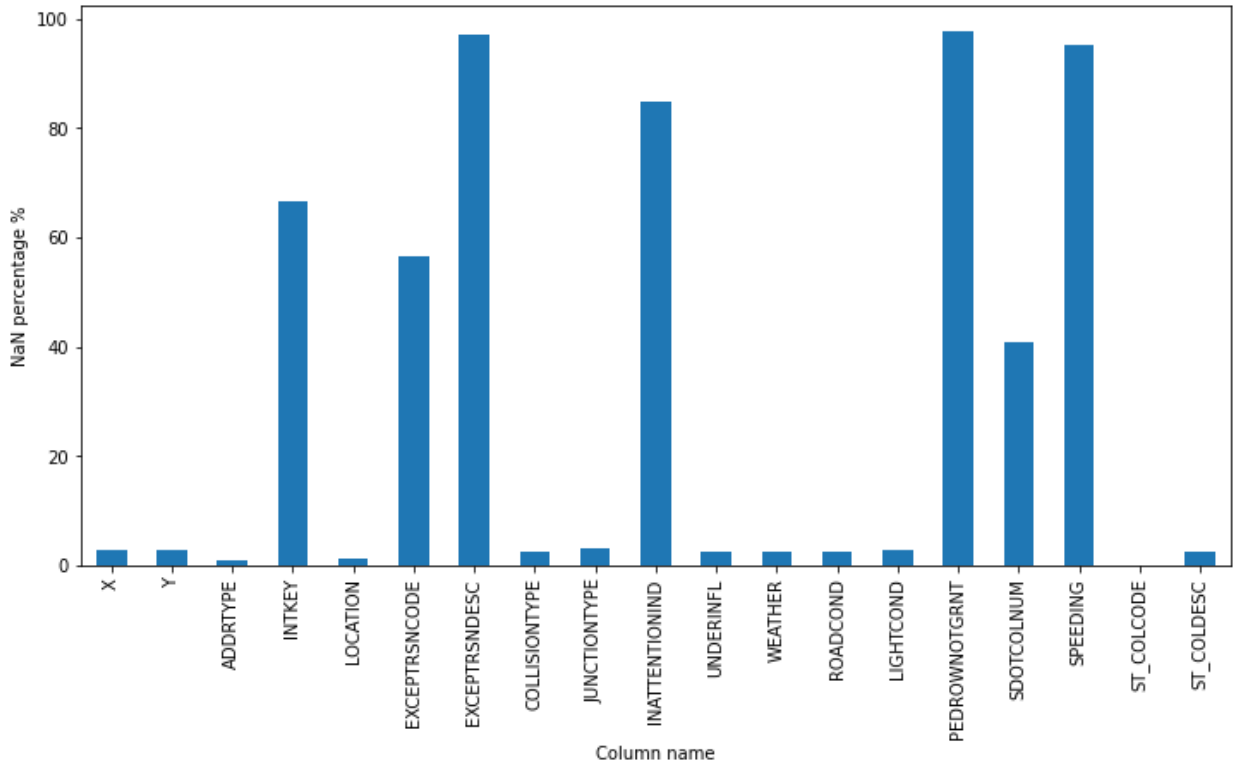


Figure 2: Columns having NaN values and the percentage of NaN

3 Methodology

3.1 Data cleaning

Columns containing coordinates, id and unexplainable information

Columns which are clearly not related to the severity of an accident are X, Y, OBJECTID, INCKEY, INTKEY, STATUS, LOCATION, EXCEPTRSNCODE, EXCEPTRSNDESC, UNDERINFL, ST_COLDESC, SEGLANEKEY, UNDERINFLL.

Column PERSONCOUNT, PEDCOUNT, and PEDCYLCOUNT

The involved persons column “PERSONCOUNT” includes the persons within the vehicle. This is not strongly related to an accident severity. This column will be removed from any further analysis.

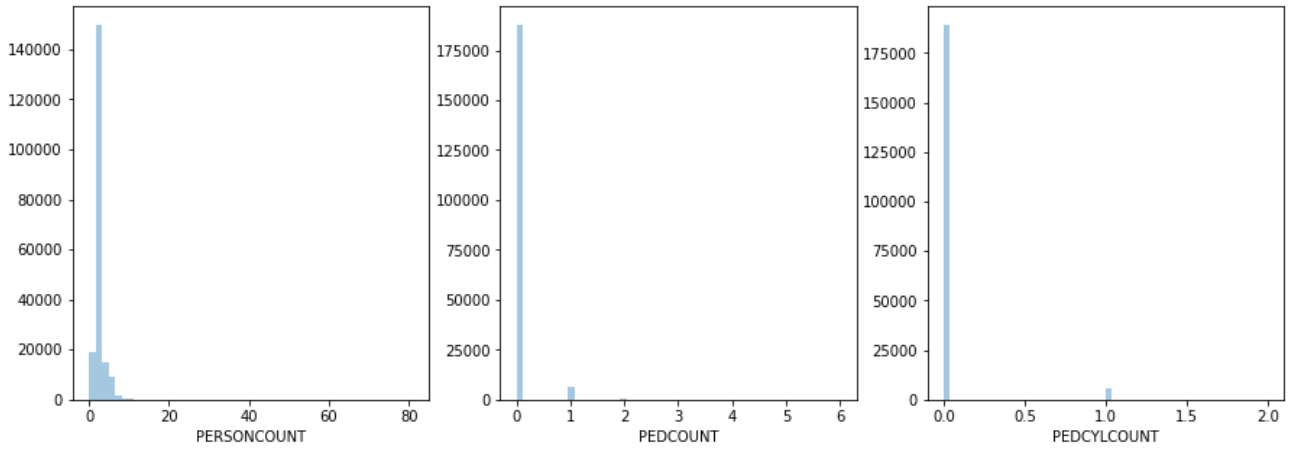


Figure 3: Attributes and distributions of PERSONCOUNT, PEDCOUNT AND PEDCLYCOUNT

The involved pedestrians and cyclists “PEDCYLCOUNT” includes the pedestrian count “PEDCOUNT”. Therefore the “PEDCOUNT” column is redundant data, and it will be removed from any further analysis.

Column WEATHER and ROADCOND

The attributes in the column WEATHER are “Raining”, “Clear” , “Unknown” etc . The attributes in column ROADCOND are “Wet”, “Dry”, “Unknown” etc. These two columns have redundant information. Therefore, only the ROADCOND column is considered for further analysis.

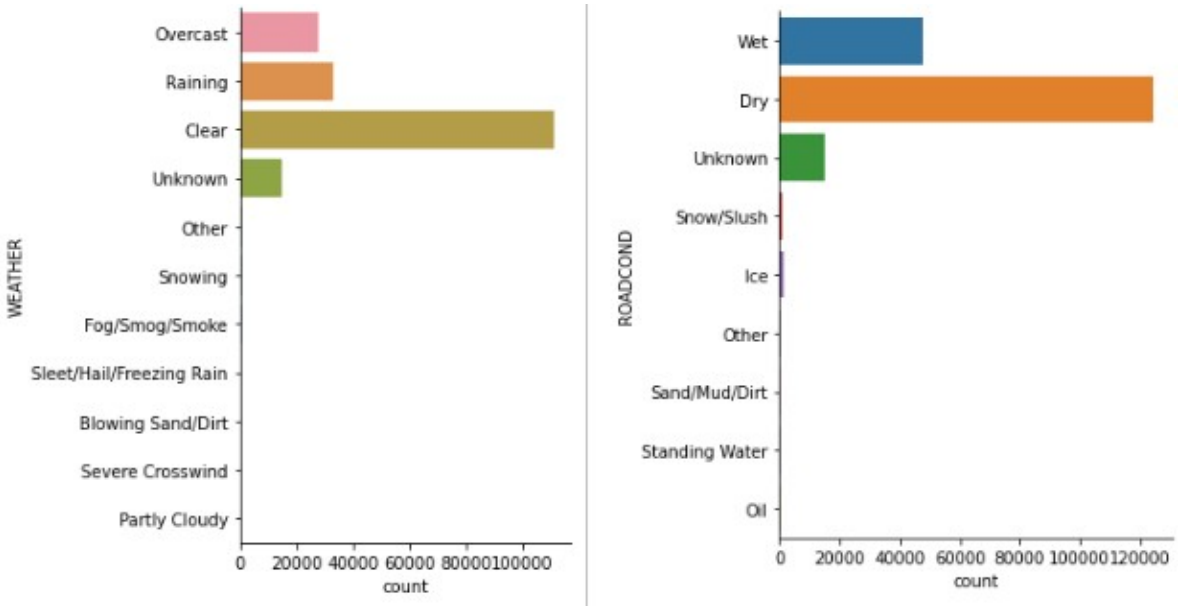


Figure 4: Attributes and distribution of column WEATHER and ROADCOND

Column SDOT_COLDESC, and SDOT_COLCODE

Columen SDOT_COLDESC is the description of the column SDOT_COLCODE, which is the indication of the collision type. This is likely related to the accident severity.

Therefore, column SDOT_COLCODE will be used for further analysis. Column SDOT_COLDESC will be removed as redundant information.

Summary of the columns to be removed

As discussed above, some columns will not be considered in further analysis and modelling. They are summarised in Table 1.

Table 1: Columns not considered in further analysis and modelling to predict accident severity

Column name	Reasons not included in further analysis
X	They are the coordinates which are not related to the accident severity
Y	
SEVERITYDESC and SEVERITYCODE.1	SEVERITYCODE.1 is the same as SEVERITYCODE. SEVERITYDESC is the description of SEVERITYCODE. Therefore, these two columns are redundant information.
OBJECTID	An identification number which is not related to the accident severity
INCKEY, INTKEY	
INATTENTIONIND, INATTENTIONIND	Meaningless column, removed
STATUS	This contains binary information “Match” and “Unmatch” however what status are they are not clear. Therefore, this column is not used for further analysis
LOCATION	Accurate address of an accident, which is not related to the severity
EXCEPTRSNCODE	More than 50% data is NaN
EXCEPTRSNDESC	More than 90% data is NaN
UNDERINFL	This contains binary data “N”, “Y”, 0 and 1. However the meaning of these data is unknown. Therefore this column is not considered in further analysis
ST_COLDESC	This is the description of ST_COLCODE.
SEGLANEKEY	An identification number and not related to the accident sereviry
CROSSWALKKEY	
PERSONCOUNT	Redundant information of PEDCOUNT
PEDCYLCOUNT	
WEATHER	Redundant information of ROADCOND
SDOT_COLDESC	Redundant information of SDOT_COLCODE
JUNCTIONTYPE	Redundant information of ADDTYPE

3.2 Data organisation for modelling

Collision type

The collision type does not have a clear patten related to the accident severity as shown in Figure 5.

For example, “Angles” is recorded more than 35000, however, this can be severe and not severe. Therefore, this column is removed for any further analysis.

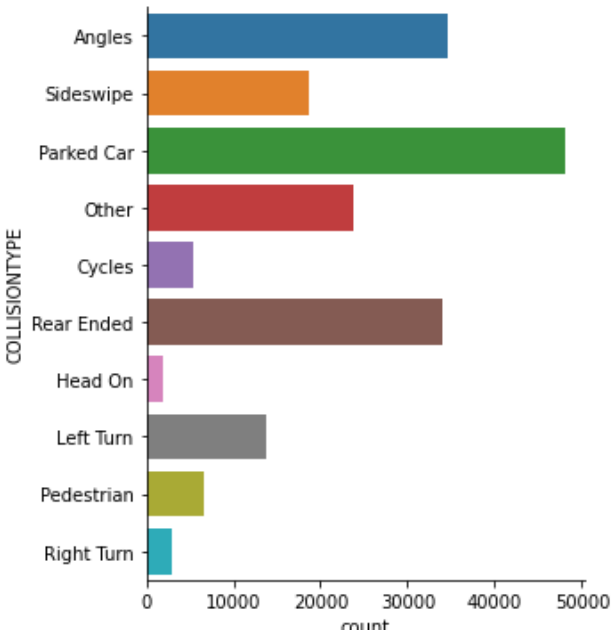


Figure 5: Collision type

Speeding

In the original data only 9333 over 194673 cases are recorded as speeding. The majority is remaining NaN. It is assumed that the remaining data is not on the condition of over speed.

[“Y”, “N”] are replaced by [1, 0]. The distribution plot is shown in Figure 6.

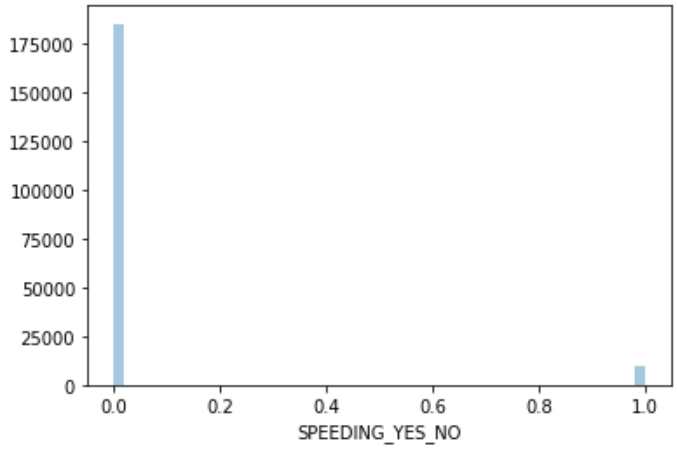


Figure 6: Speeding distribution plot

Address type

The address type in the dataset includes “Intersection”, “Block” and “Alley”, as shown in Figure 7. As the attribute “Alley” only have 751 samples, the attribute of these rows are replaced by “Block”.

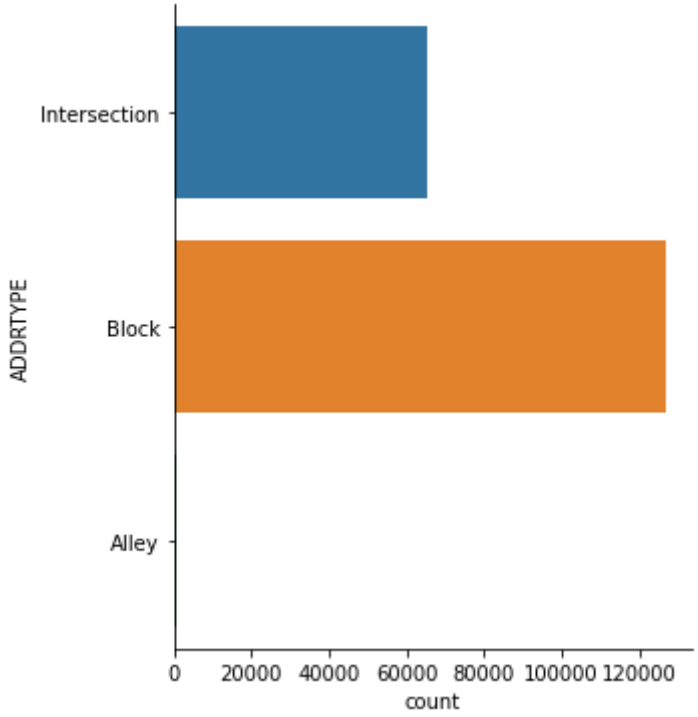
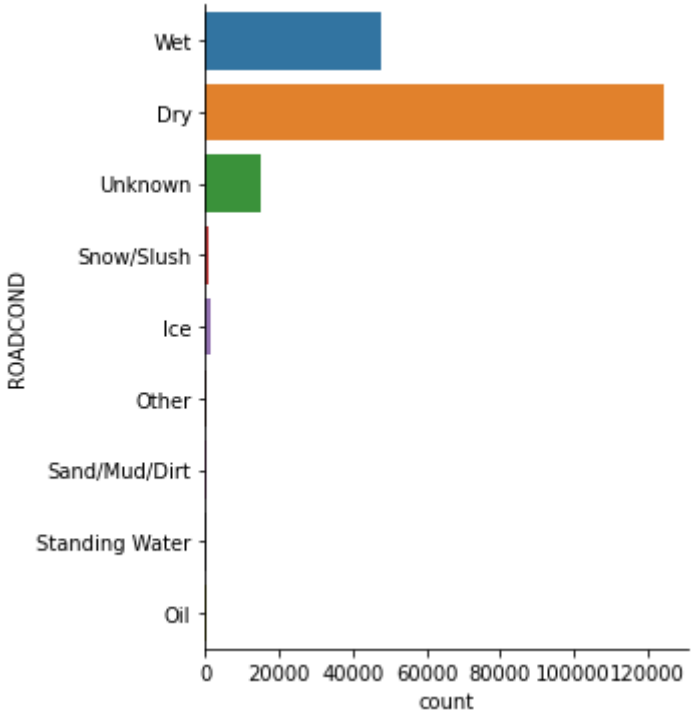


Figure 7: Address type

Road condition

There are 9 different road conditions in this column. These road conditions can be grouped to wet and dry surface road. The “Unknown” samples will be grouped into dry surface road group.



Light condition

Light condition will be included in the predicting model. The values of the light condition are also grouped to “Bright” and “Dark”. Where "Daylight", "Unknown", "Dusk", "Dawn", "Other" are categorised as “Bright” and "Dark - Street Lights On", "Dark - No Street Lights", "Dark - Street Lights Off", "Dark - Unknown Lighting" are categorised as “Dark”.

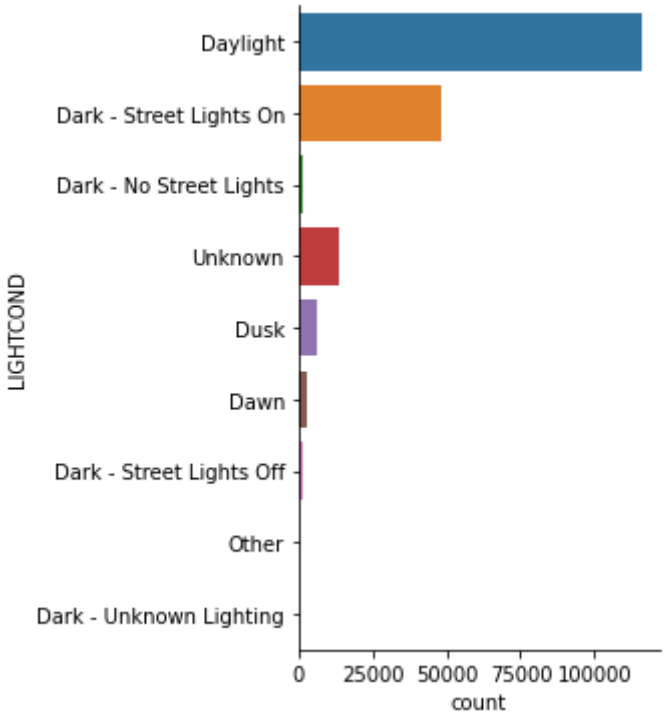


Figure 8: Light condition

Overall view of the data for modelling

The overall view of the data after data cleaning is shown below.

```
#   Column           Non-Null Count   Dtype
---  ------           --------------   -----
0   SEVERITYCODE     194673 non-null  int64
1   PEDCOUNT         194673 non-null  int64
2   VEHCOUNT         194673 non-null  int64
3   SDOT_COLCODE     194673 non-null  int64
4   SPEEDING_YES_NO  194673 non-null  int64
5   Block            194673 non-null  uint8
6   Intersection     194673 non-null  uint8
7   Dry              194673 non-null  uint8
8   Wet              194673 non-null  uint8
9   Bright           194673 non-null  uint8
10  Dark             194673 non-null  uint8
dtypes: int64(5), uint8(6)
```

As the address type “Block and Intersection”, road condition “Dry and Wet”, and light condition “Bright ad Dark” are binary value pairs, the features “Intersection”, “Dry” and “Bright” are also dropped.

The features used in modelling are shown below.

```
#   Column         Non-Null Count   Dtype
---   ------            --------------    -----
 0    SEVERITYCODE      194673 non-null   int64
 1    PEDCOUNT          194673 non-null   int64
 2    VEHCOUNT          194673 non-null   int64
 3    SDOT_COLCODE      194673 non-null   int64
 4    SPEEDING_YES_NO   194673 non-null   int64
 5    Block             194673 non-null   uint8
 6    Wet               194673 non-null   uint8
 7    Dark              194673 non-null   uint8
dtypes: int64(5), uint8(3)
```

Relation between the features and the severity

From data visualisation, it is found that the most significant features could affect the severity is the involved pedestrians and speeding as shown in Figure 9 and Figure 10. Relation between severity and other features are not possible to directly identified therefore, they are not illustrated in this report.

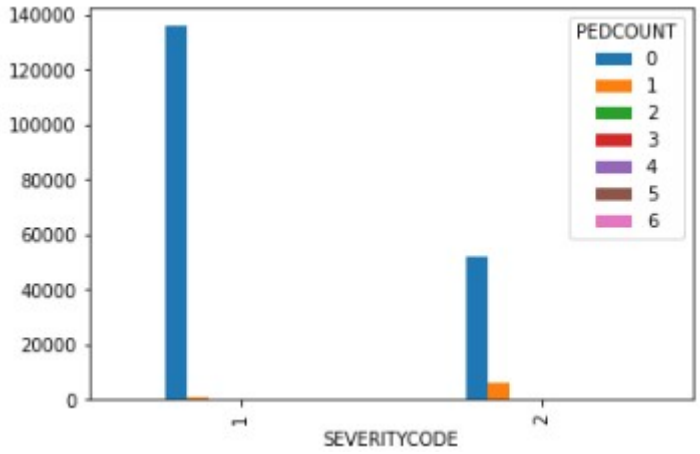


Figure 9: Relation between involved number of pedestrians

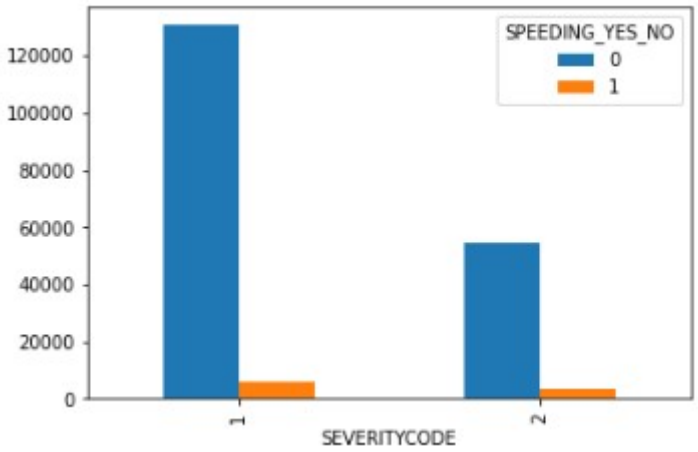


Figure 10: Relation between the severity and the speeding

3.3 Split data to training set and testing set

80% of the dataset will be used for training and 20% of the data will be used for testing.
All the data will be standardised before splitting, training and testing.

3.4 Machine learning models

To predict the accident severity either 1 (not severe) or 2 (severe) is a classification problem. Therefore the following machine learning models are used.

K-nearest neighbour

This algorithm will classify the input information to belong to a certain group based on the similarity of this input information and its neighbours.

K is the number of neighbours used in the similarity comparison. The sklearn from "sklearn.neighbors" module "KneighborsClassifier" will be used.

K will affect the predicting accuracy. Therefore, the best K is determined based on the predicting error. To avoid over fit, the greatest K in the testing is 10.

Based on the training data, the best K is 8 as shown in Figure 11.

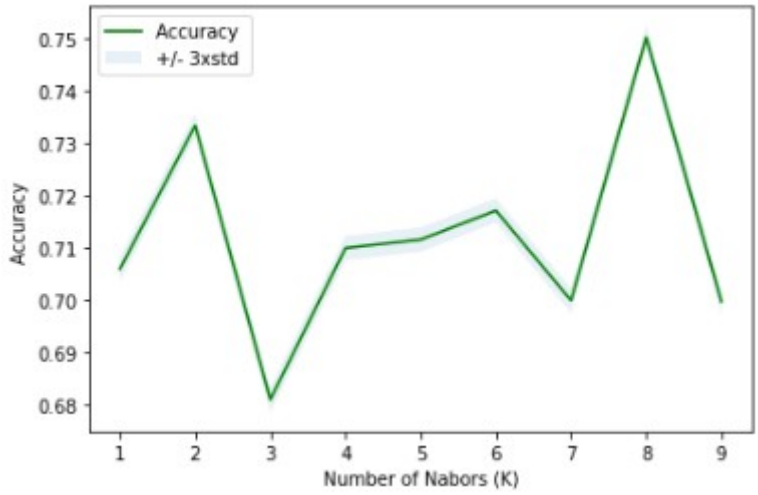


Figure 11: Find the best K

Decision tree

Decision trees are built by splitting the training set into distinct nodes, where one node contains all of or most of one category of the data. This is also used in data classifying.

The sklearn.tree module DecisionTreeClassifier is used in this report.

Logistic regression

The goal of logistic regression is to build a model to predict the class of each accident and also the probability of each sample belonging to a class.

The sklearn.linear_model module LogisticRegression is used in this report.

Support vector machine

A Support Vector Machine is a supervised algorithm that can classify cases by finding a separator. SVM works by first mapping data to a high dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. Then, a separator is estimated for the data. This can also be used for this project.

The sklearn module svm is used in this report.

Evaluation methods

f1_score, jaccard_score and log_loss are used to evaluate the predicting machine learning models.

4 Results

The predicting results of an accident severity with different machine learning algorithms are shown in Table 2.

Table 2: Predicting accuracy

Algorithm	f1_score	jaccard_score	log_loss
KNN	0.70	0.73	--
Decision tree	0.68	0.73	--
SVM	0.70	0.74	--
Logistic regression	0.72	0.67	0.55

The above table indicates that the SVM method gives the best predicting output. The accuracy of other machine learning methods is slightly lower than the SVM however not significant. The confusion matrix is shown in Figure 12.

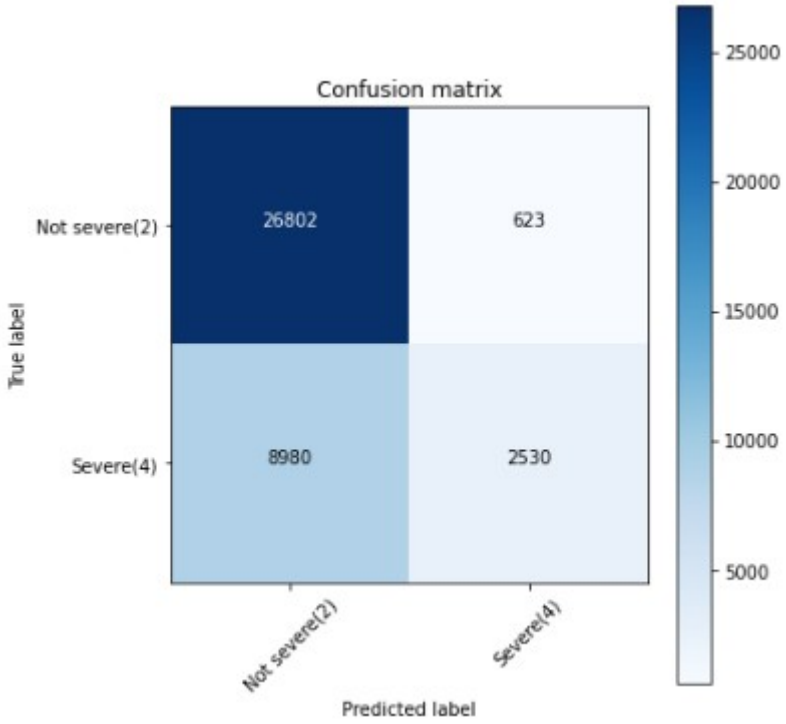


Figure 12: Confusion matrix of the SVM predicting model

5 Discussion

The data samples with severity code 1 (not severe) is 2.3 times more than that of severity code 2 (severe). Therefore the machine learning algorithm is not evenly fitted. It could tend to predict more code 1 cases.

The road conditions, and light conditions are simplified to binary values to make the model simple and possible to fit on a low performance laptop. The predicting precision may be increased if all these parameters are included.

6 Conclusion

The accident severity dataset was used and the project is aiming to find a model to predict the accident severity for local transportation department.

The dataset was analysed and the unrelated or redundant features are dropped. The directly related features are modified so that make it suitable for machine learning.

Classify machine learning models, such as K-Nearest Neighbout, Decision Tree, Support Vector Machine and Logistic Regression have been used in this project. The prediction accuracy was evaluated with F1 score, Jaccard score and log loss.

It is found that the Support Vector Machine is the best algorithm to predict the accident severity with F1 score of 0.70 and Jaccard score of 0.74.

The accident severity may be predicted with relative high confidence.