Predicting the Severity of a Car Accident Using Seattle Collision Data

Sankhadeep Bhowmick

October 1, 2020

1. Introduction

1.1 Background

Car accidents are unfortunate but very common an incident around the world. The city of Seattle in the Washington state of United States of America is no exception with a good amount of car accidents seen every year over more than the past decade. As the number of cars is increasing accidents are becoming more commonplace these days. And whenever there is an accident, it involves injuries, destruction of properties, fatalities etc. and is followed by road blocks. The situation immediately demands the intervention of the paramedic team, the police department and the nearest hospital authorities. So it becomes absolutely necessary for these authorities to be totally prepared for an accident and respond immediately. Therefore, it would be very advantageous for these authorities to predict such a situation in advance on a given day. For example, the hospital authorities could increase their attending staff strength on a day likely to lead to more accidents.

1.2 Problem

Data that contribute to determining the severity of a car accident might include road condition, weather condition, lighting condition, inattention, speeding, influence of drugs, number of persons involved, number of pedestrians involved, number of cyclists involved, number of vehicles involved, collision type, junction type, address type, location etc. that describe the overall situation the accident might take place in. This project aims to predict the severity of car accidents that might happen in the future based on data related to these fields.

1.3 Interest

Firstly, the hospitals would be very much interested in accurate prediction of the severity of an accident so as to maintain sufficient amount of staffs and doctors on a given day to handle the situation efficiently. Secondly, the police department would be very interested in accurately predicting the same to increase alertness on a given day if the probability of a severe accident turns out to be high enough. Third, the Department of Transportation would be very much interested in accurately predicting the same as to warn the people of the chances of getting into a severe accident on a given day through online portals and/or by other means.

2. Data acquisition and cleaning:

2.1 Data sources

The collisions data is recorded by Seattle Department of Transportation, Traffic Management Division, Traffic Records Group and maintained by ArcGIS and can be found in the Kaggle dataset here and the metadata can be found from here. The collisions dataset includes all types of collisions from year 2004 to present. The collisions dataset is what I'm going to be working with in this project.

2.2 Data Cleaning

The data downloaded from the above mentioned source was saved in a table. There were many entries in the dataset which said that no person, pedestrian, cyclists or vehicles were involved in those collisions. Analysing such entries a bit further they were found to be irrelevant and hence deleted from the dataset.

There were several columns (features) in the dataset that were added by various authorities to mainly identify the incident instance and were irrelevant to our final goal of developing a predictive model. Those columns were dropped. **Our aim is to build a predictive model which can foretell or predict accurately the severity of any future accident given the conditions (features) that can be pre specified**. There were several columns in our dataset which only made sense only when a collision had already taken place; so those columns were also dropped. Also we want our model to be independent of the exact locations of the collisions, therefore columns indicating the exact locations of the collisions were dropped and rather the columns describing the properties of such locations were kept. Below is a table of the dropped columns and the reason for their dropping:

Table 1: Dropping of irrelevant columns (features) during data cleaning

Dropped Column	Reason for dropping
OBJECTID, INCKEY, COLDETKEY, REPORTNO, STATUS, INTKEY, EXCEPTRSNCODE, EXCEPTRSNDESC, SDOTCOLNUM	Irrelevant. Added by authorities only for identification purpose.
ST_COLCODE, ST_COLDESC, PEDROWNOTGRNT, COLLISIONTYPE, INJURIES, SERIOUSINJURIES, FATALITIES, SDOT_COLCODE, SDOT_COLDESC, HITPARKEDCAR	Can be known only after a collision.

SEVERITYDESC	Irrelevant. SEVERITYCODE kept instead.
INCDATE	Irrelevant. Prediction should work for any given day.
INCDTTM	Model to be independent of time. LIGHTCOND kept instead.
X, Y, LOCATION, SEGLANEKEY, CROSSWALKKEY	Model to be independent of exact location. ADDRTYPE, JUNCTIONTYPE kept instead.

There were a lot of missing values as well as data inconsistencies in many of the kept features fields. For example, the 'ADDRTYPE' feature which is a categorical variable, describes the types of the addresses where the collisions took place had 2011 missing values. I did not drop this column rather replaced the missing values with the label "others", as it seems a potential candidate in determining the severity of an accident and hence could be a valuable input to our model. In the exact similar way I also handled the missing values for the categorical features 'JUNCTIONTYPE', 'WEATHER', 'ROADCOND', 'LIGHTCOND' by replacing the missing values with the label "others" for the same reason.

The 'INATTENTIONID' feature which is a categorical variable and supposed to have a 'Y' for 'yes' and 'N' for 'no' values, but instead it had data inconsistency where the 'yes' values were labelled by 'Y' but the 'no' values were kept as missing values. I replaced all the missing values with 0's and replaced the 'Y' labels with 1's.

The 'UNDERINFL' feature had data inconsistency too. Some of the 'yes' values were labelled as 'Y' whereas the others as 1's and in a similar fashion some 'no' values were labelled as 'N' and the others as 0's. I handled the situation by converting all the 1's to 'Y's and all the 0's to 'N's. Moreover this feature also had 26479 missing values which I labelled as 'UN' which stands for 'unknown'.

The 'SPEEDING' feature had data consistency where the 'yes' values were correctly labelled as 'Y's but the 'no' values were indicated using missing values. I replaced the missing values with 0's to indicate a 'no' and the 'Y's to 1' to indicate a 'yes'.

At last I checked whether the data-types of the features were consistent with the values they contained or not. I found out the 'INATTENTIONID' feature had a data-type inconsistency where the data-type of this feature was 'object' but the values it contained were 0's for a 'no' and 1's for a 'yes'. Therefore I converted the data-type of this feature from 'object' to 'int64' to make it consistent with the values it contained.

2.3 Feature Selection

After data cleaning there were 201786 samples and 13 features in the cleaned data. To better select the features the Pearson correlation coefficient was calculated for the dataset to get an

idea about the extent to which the features were correlated to the target variable 'SEVERITYCODE'. As Pearson correlation function only works on numeric features I had to encode the categorical features into numeric values and then calculate the Pearson correlation.

Examining the Pearson correlation of different features with the target variable I found out that none of the features displayed a high positive or negative correlation, with 'PEDCOUNT' being the most positively correlated with a Pearson correlation coefficient of 0.28. This is understandable because none of the features contained continuous values and most of them were categorical and they were encoded into numerical values which were discrete in nature.

Depending on the above observation none of the features were dropped and all of them were kept to be better understood during the Exploratory Data Analysis phase.

3. Exploratory Data Analysis

3.1 Relationship between number of persons involved in an accident and number of pedestrians, cyclists among them

While trying to get an idea about how many of the persons involved in an accident were pedestrians and how many were cyclist, it was seen that neither the number of pedestrians nor the number of cyclist were correlated with the number of persons involved in a straight positive manner, i.e. more number of pedestrians or cyclists did not always individually pertained to more number of persons involved. This is probably because the number of persons involved may consist of passengers in a vehicle that was involved in the accident and/or obviously a combination of the pedestrians and cyclists.

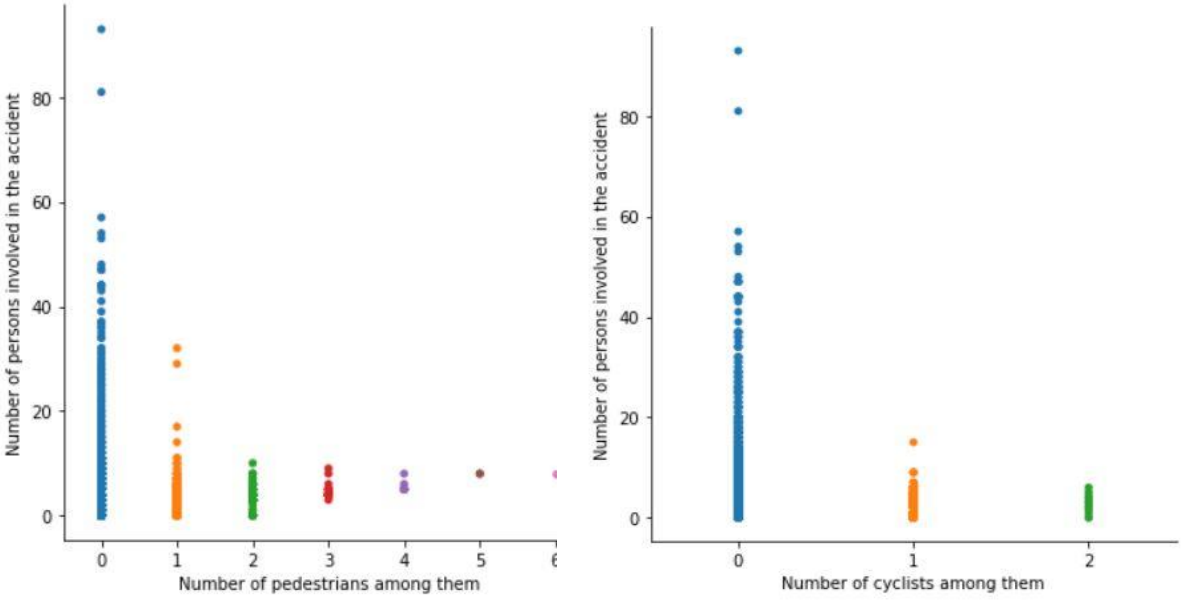


Figure 1. Number of persons involved and number of pedestrians and cyclists among them

3.2 Relationship between the number of person involved in an accident and the number of vehicles involved

While trying to get an idea about how the number of persons involved in an accident and the number of vehicles involved, a somewhat positive linear relationship is seen but not very strong. This may well be for the same earlier stated fact that the persons involved in an accident come from different categories such as pedestrians, cyclists etc. not only from the passengers in the vehicles involved. But there were two distinct outliers detected which showed a high person involved count while the vehicle count was low. This fact can be explained by the scenario where one or more of the vehicles involved are buses or such vehicles carrying many passengers.

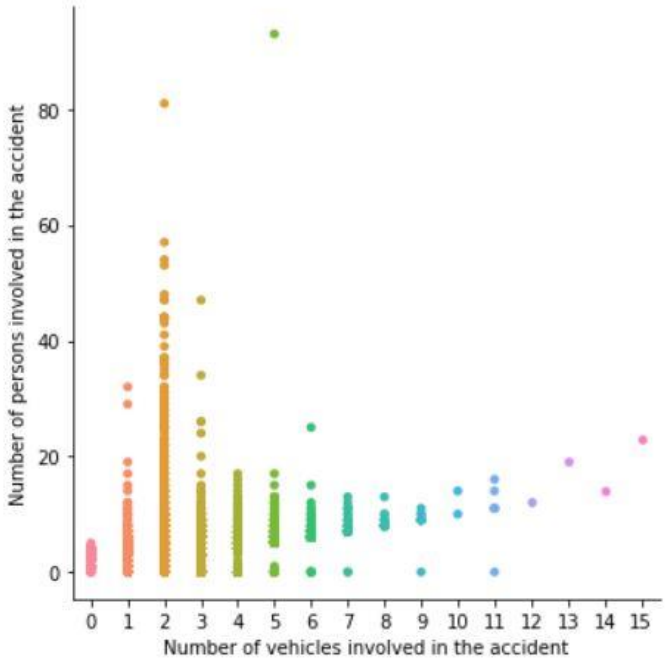


Figure 2. Number of persons involved and number of vehicles involved

From the above observation and the observations in 3.1 it was clear that the two data points which had the value for 'PERSONCOUNT' greater than 80 were outliers and hence all the entries containing a value greater than or equal to 80 for 'PERSONCOUNT' were deleted from our dataset.

3.3 Relationship between severity of an accident and address type

While trying to get an idea about which address types contributed to which category of accidents it was found out that "Other" address type contributed to the most severe of the accidents, i.e. of the SEVERITYCODE category '3', i.e. "fatality". While blocks and alleys contributed to accidents of SEVERITYCODE category '2' the most, intersections contributed to the SEVERITYCODE category '1' the most.

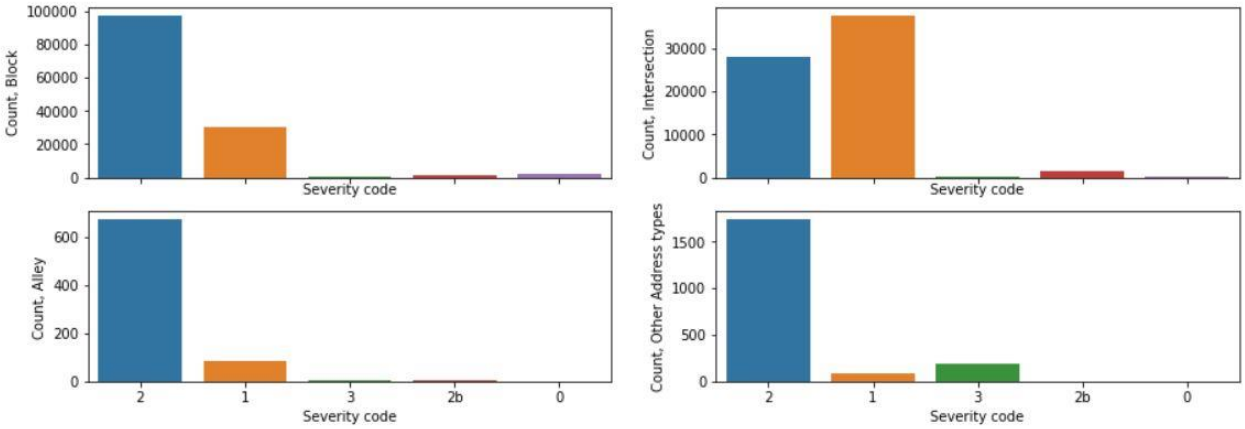
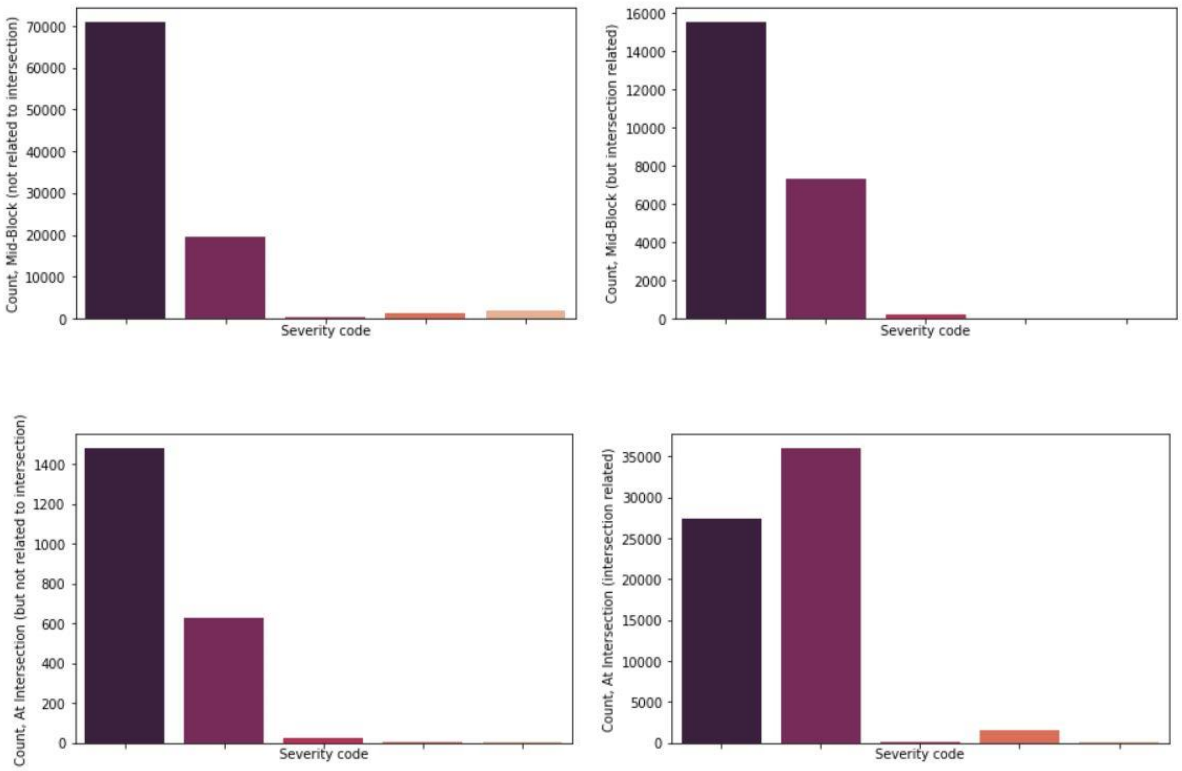


Figure 3. Distribution of the severity code categories over different address types

3.4 Relationship between severity of an accident and junction type

It was tempting to keep only one independent feature between 'ADDRTYPE' and 'JUNCTIONTYPE' for modelling and drop the other, but on a close observation of the labels they contain, it become evident that they both are needed be kept. While trying to get an idea about which type of junction contributed to which severity code category of accidents the most, it was found that the most severe accidents, i.e. most of the accidents with a SEVERITYCODE of ‘3’ took place mid-block (but not related to intersection)., while some unknown locations contributed the most to the SEVERITYCODE category ‘2’ i.e. injury.



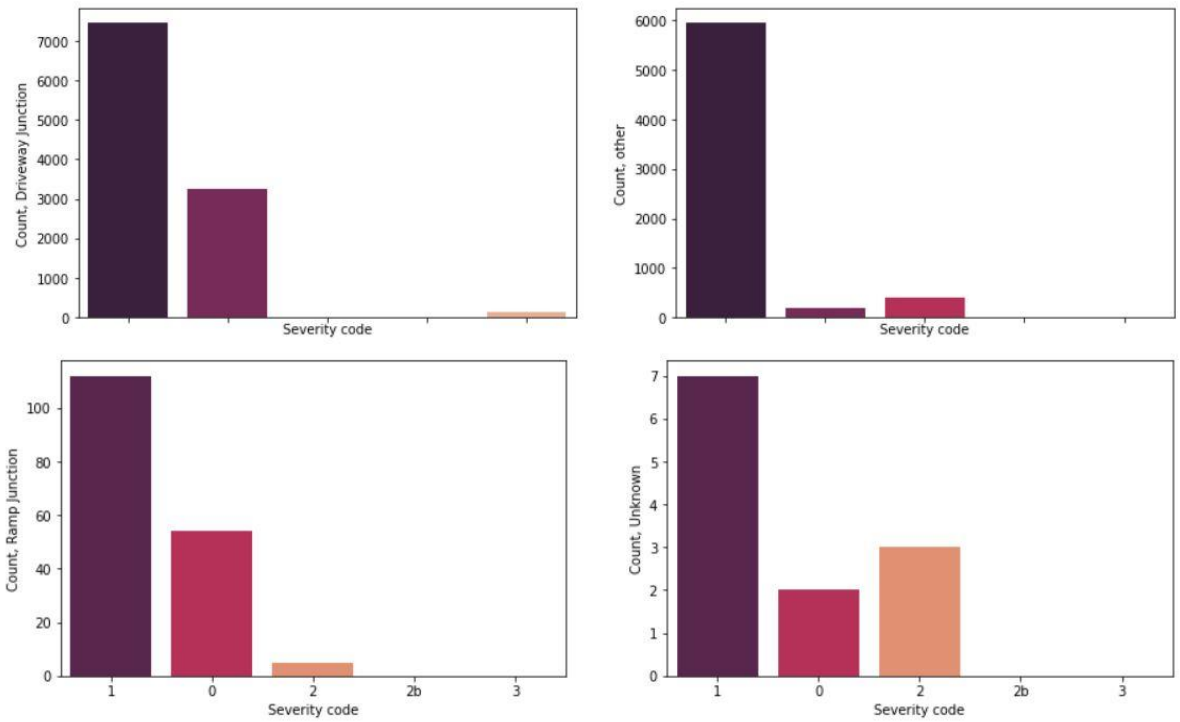


Figure 4. Distribution of the severity categories over different junction types

3.5 Relationship between severity of an accident and number of persons involved

While analysing the relationship between the severity of the accidents and the number of persons involved in an accident higher number of persons involved did not necessarily implied higher degree of severity. This is understandable from the fact that even only one person is involved in an accident but that person died in that accident, the accident would be marked with the SEVERITYCODE '3' i.e. 'fatality'. The highest person count was observed in the severity code category '1' i.e. property damage.

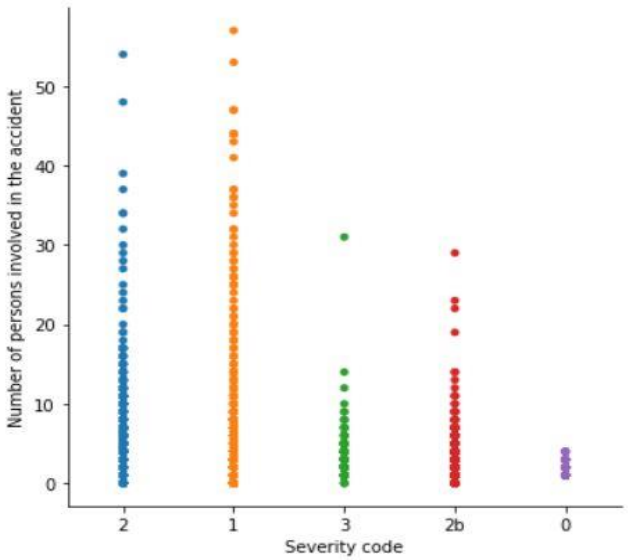


Figure 5. Distribution of number of persons involved over the severity categories

3.6 Relationship between severity of an accident and number of pedestrians involved

During this analysis it was observer that every severity code category involved both low and high number of pedestrians. This observation can also be understood by the same fact stated in point number 3.5. Severity code category '2' had seen the most number of pedestrians involved for a single accident instance though.

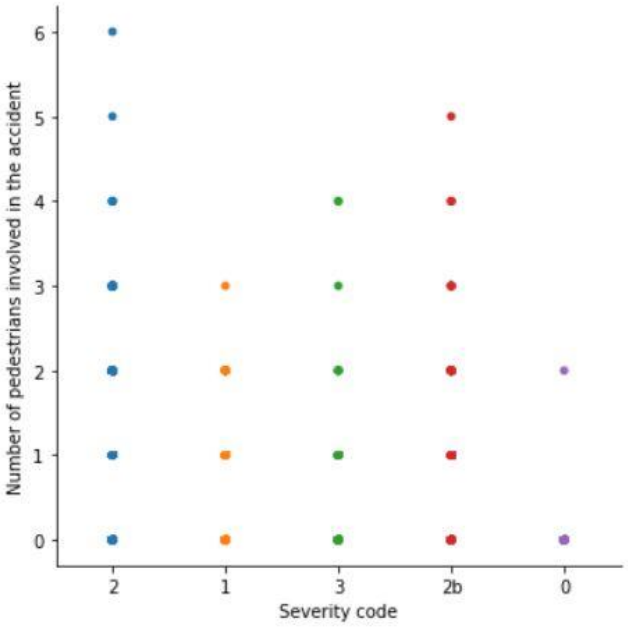


Figure 6. Distribution of number of pedestrians involved over severity categories

3.7 Relationship between severity of an accident and number of cyclists involved

While trying to get an idea about cyclists involved in an accident contributed to which category of severity code the most, it was seen that all the categories except the SEVERITYCODE '0' had seen low, moderate and also max number of cyclists involved.

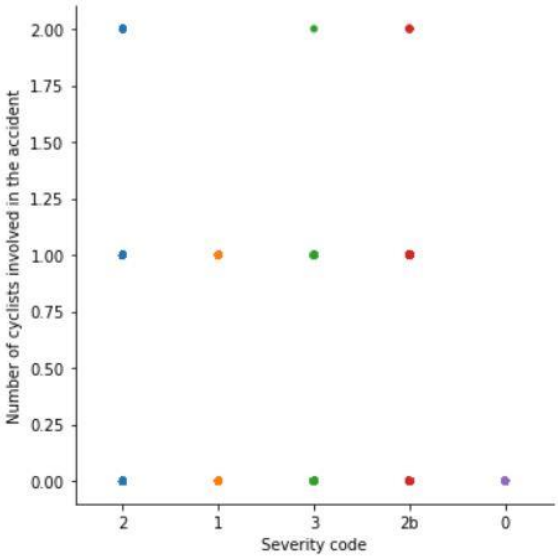


Figure 7. Distribution of number of cyclists involved over severity categories

3.8 Relationship between severity of an accident and number of vehicles involved

It was observed that vehicles led to more severe categories of accidents. Although all the severity categories had seen low, moderate and high number of vehicle counts for different instance of collisions but the severity code category '2b' i.e. serious injury had seen the highest second highest and third highest counts of vehicles for three different instances of collisions/accidents.

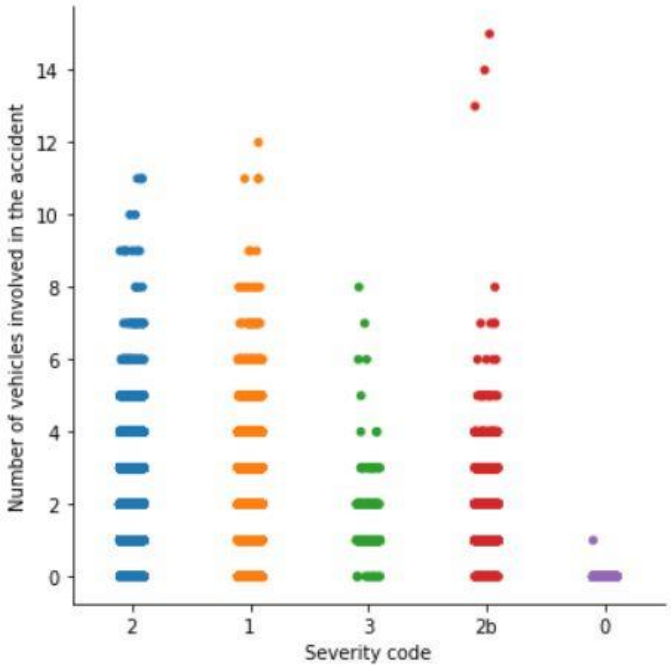


Figure 8. catplot (jitter = True) of distribution of number of vehicles involved over severity categories

3.9 Relationship between severity of an accident and inattention of the person/s involved

In the accidents where inattention of one or more of the persons involved in the accident was a causing factor to the accidents, most such accidents led to the severity code category '1' i.e. property damage.

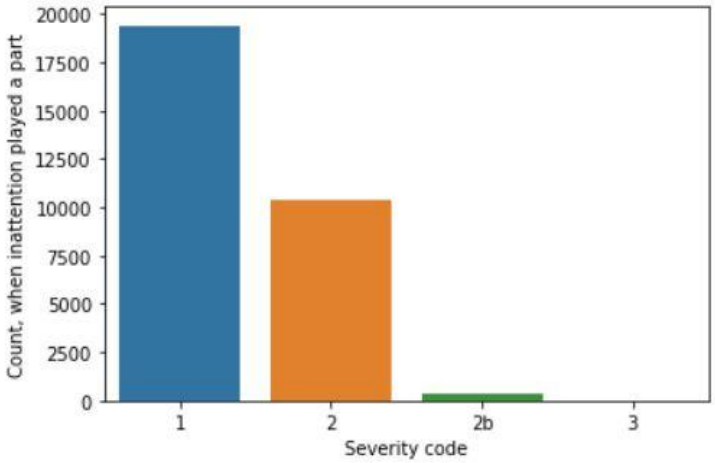


Figure 9. countplot of INATTENTIONIND counts over different severity categories

3.10 Relationship between severity of an accident and when person/s involved were under the influence of drug/alcohol

While trying to analyse the above said, it was found that most accidents due to such condition had fallen into severity code category ‘1’ yet a sizable amount of accidents also had fallen into more severe categories like ‘2’ and ‘2b’. There also were 93 such accidents which saw fatalities too i.e. were of severity code category ‘3’.

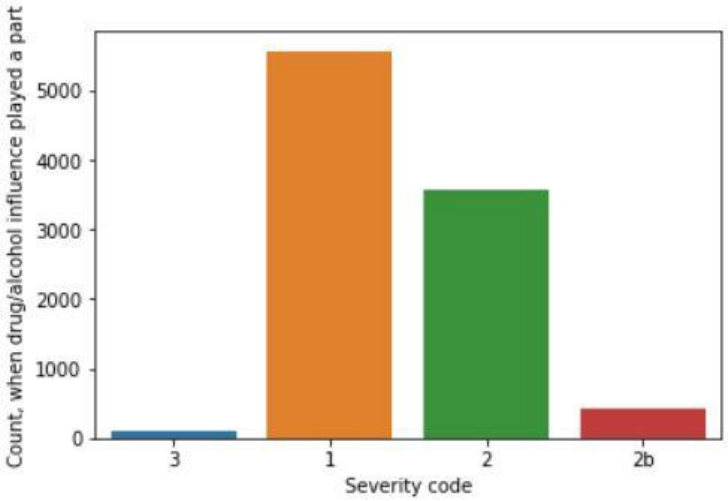


Figure 10. countplot of UNDERINFL counts over different severity codes

3.11 Relationship between severity of an accident and speeding of vehicle/s

While trying to analyse which severity categories speeding of vehicle/s involved in the accidents contributed to the most, to my surprise I found out that speeding of vehicles only led to accidents with SEVERITYCODE ‘1’ i.e. property damage, none of the other severity levels were seen when speeding was a factor in the accidents. And also over 86% of the rows in the original dataset were missing the values for the SPEEDING field. Therefore, I removed the SPEEDING variable from our data in order to avoid bias in our model towards this severity code category.

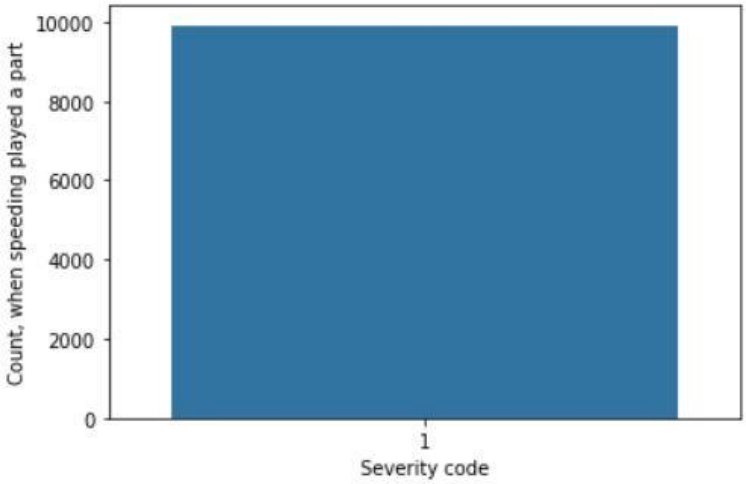


Figure 11. countplot of SPEEDING counts over different severity codes

3.12 Relationship between severity of an accident and weather condition

While trying to get an idea about what type of weather conditions led to the most severe of accidents, I found out to my surprise that “Clear” weather conditions contributed to most of the severity category ‘3’ accidents, i.e. fatality. “Clear” weather also contributed to the next most severity category ‘2b’ the most.

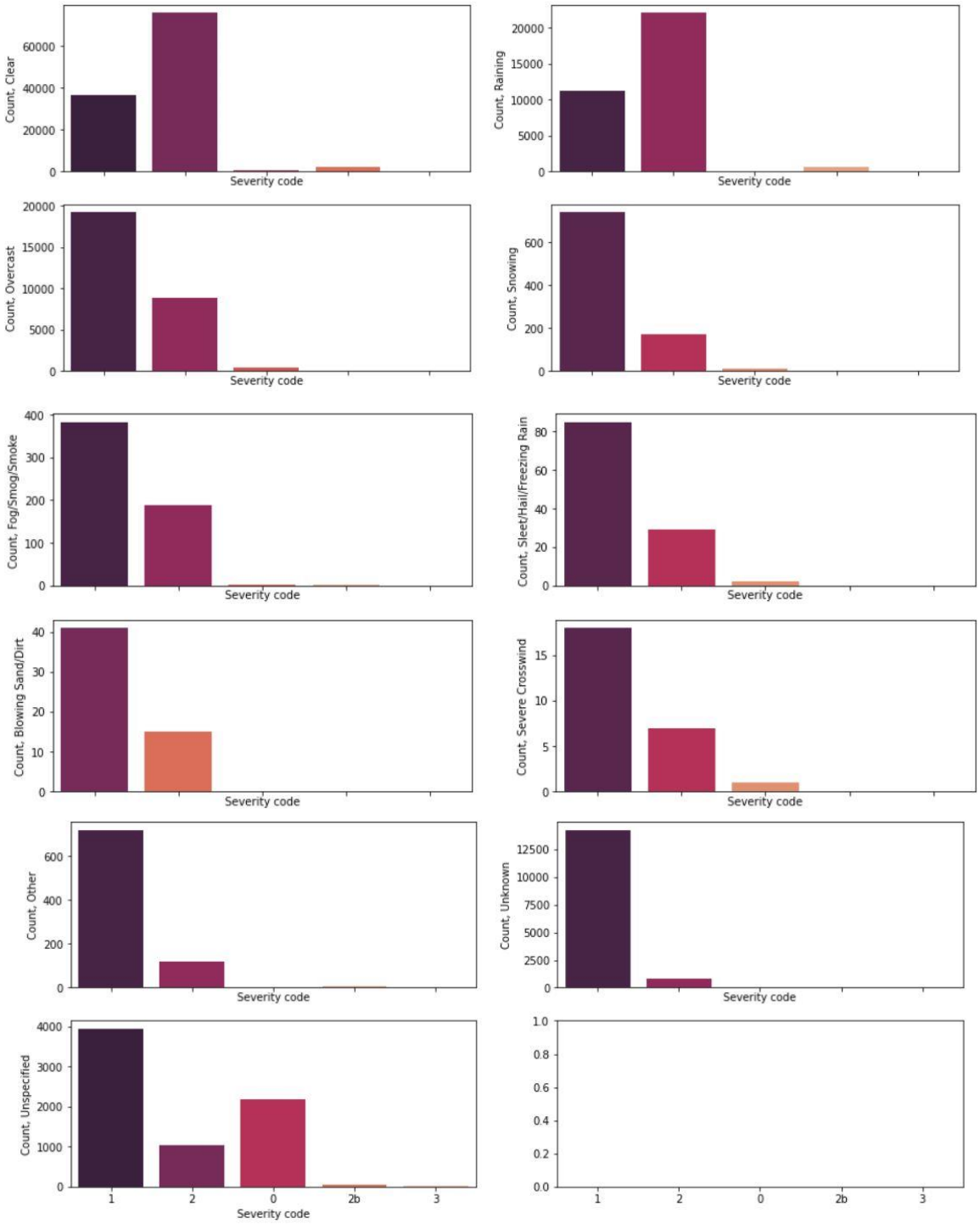


Figure 12. Distribution of the severity categories over different weather conditions

3.13 Relationship between severity of an accident and road condition

While trying to get an idea about which type of road condition contributed to accidents of which severity category, it was found that “Dry” road conditions contributed to most of the severity level ‘3’ accidents.

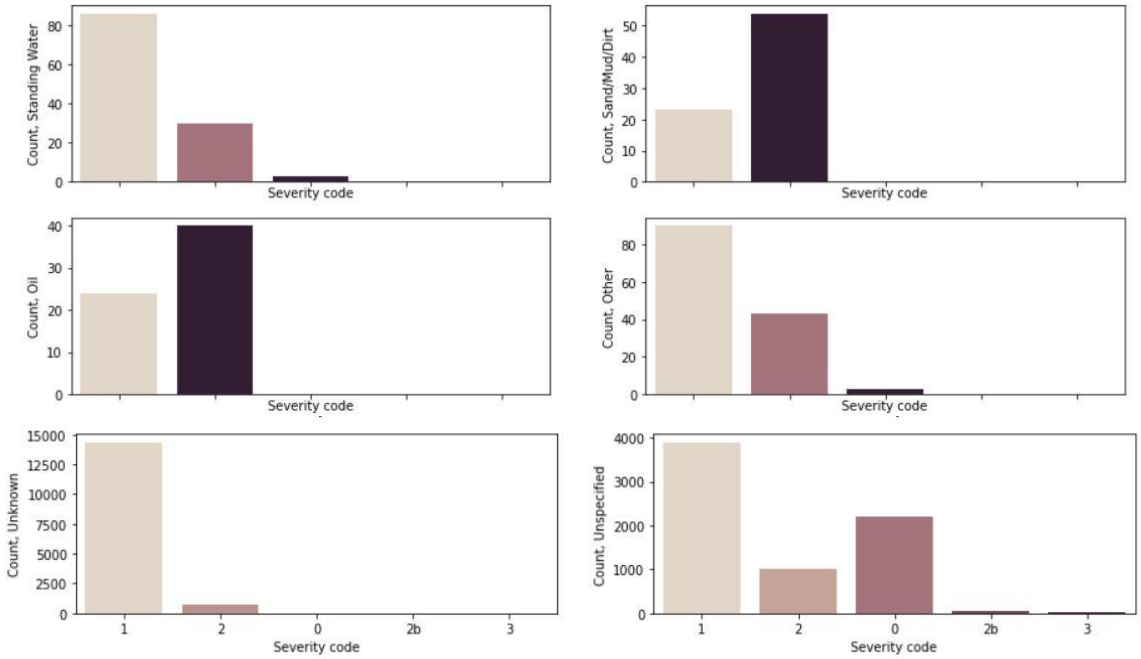
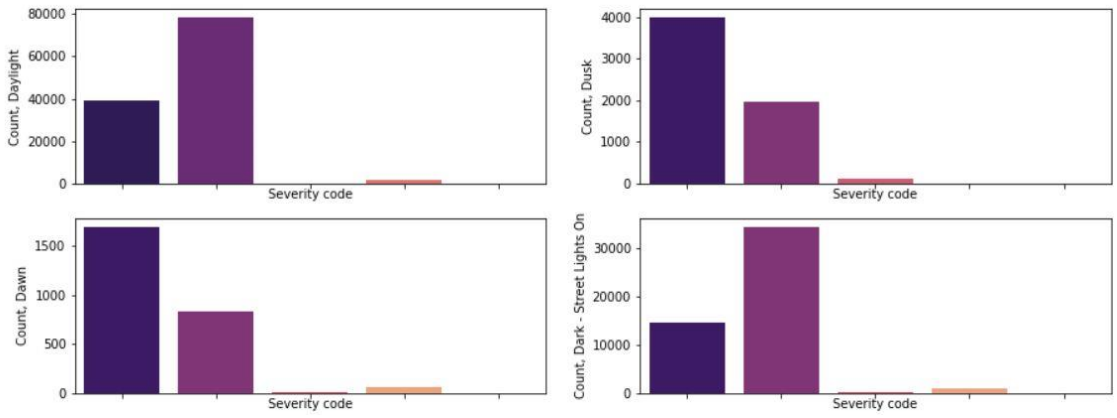


Figure 13. Distribution of the severity categories over different road conditions

3.14 Relationship between severity of an accident and light condition

While trying to get an idea about the number of accidents of various severity categories in the different light conditions, it was found that most of the accidents of severity level ‘3’ happened in “Daylight” condition.



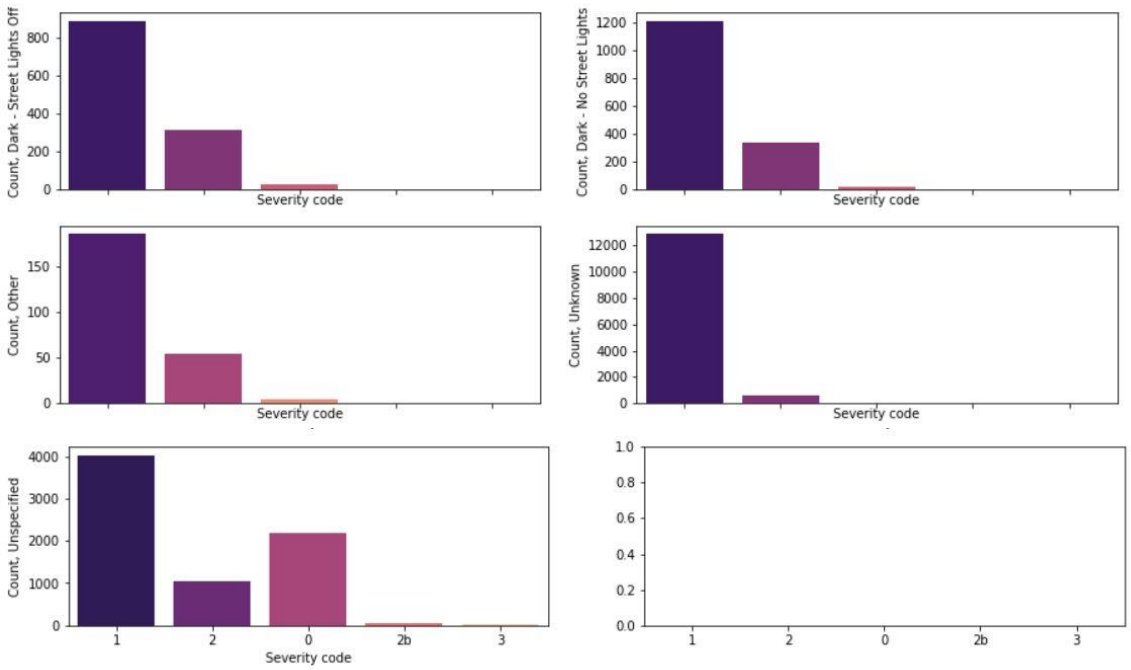


Figure 14. Distribution of the severity categories over different light conditions

4. Predictive Modeling

Our objective in this project was to build a machine learning model that could predict/foretell the severity level of a future accident given a feature set instance. The severity of an accident was given by five different labels/classes namely, '3', '2', '2b', '1' and '0'. So there were multiple labels/classes there in our target variable 'SEVERITYCODE" in our dataset. So **this problem fell into the category of "Multi-class Classification"**. Therefore, I built several classification models using different classification algorithms and the model with the highest accuracy can be selected for implementation.

4.1 K Nearest Neighbors Classification Model

I built the first model using the "K Nearest Neighbors" classification algorithm. While doing that I first wanted to obtain an optimal value for the hyper parameter "k". But I encountered a problem there. The cleaned dataset was too large. Even after splitting the dataset into training and testing datasets, the training dataset contained over 1.61 lakh rows and probing this huge data multiple times with different values of "k" in order to find the optimal "k" was hugely time consuming. As a solution to this problem I selected a relatively small portion of the dataset for finding the optimal "k" value and split that into training and testing data. Finally after finding out the optimal value for "k", I fitted the model with the whole original training data and that "k" value.

4.2 Decision Tree Classification Model

Before I built and trained a decision tree model with the training data I first wanted to find an optimal value for the hyper parameter "max_depth" for the decision tree. So what I did was I further split the training data into training and testing data and found out the mean accuracy for a few values of "max_depth" by repeatedly training and testing the decision tree model with varied "max_depth" values. This step was quick to produce the results and was not slow as in our KNN model. The "max_depth" value which gave out the maximum mean accuracy was selected to feed the final model with. Finally I fitted a decision tree model with the whole original training data and the best "max_depth" value.

4.3 Logistic Regression Classification Model

Before I built and trained a logistic regression model with the training data I first wanted to find the best "solver" to use for our model. But the same problem as in the "K Nearest Neighbors" model was here: the training dataset was huge and probing the dataset multiple times with different solvers was hugely time consuming. As a solution to this problem I did the same as I did in the "K Nearest Neighbors" model and found out the best solver. Finally I fitted a logistic regression model with the whole original training dataset.

4. Model Evaluation

The fitted models were used to predict the severity code labels using the test features set and the accuracy of the predictions for the models were evaluated using various evaluation metrics. The confusion matrices for each of the models were also generated.

4.1 Accuracies of the different models

Table 2: Accuracy score of different models

	Jaccard Score	**F1 Score**	**Subset Accuracy Score**	**Log Loss**
K Nearest Neighbor Model	0.561288	0.686157	0.731645	NA
Decision Tree Model	0.566244	0.689493	0.731645	NA
Logistic Regression Model	0.552580	0.671888	0.732512	0.611181

4.2 Confusion Matrices of different models

Confusion matrices are a great way to visualize how many of the labels got predicted correctly and which labels were mistaken as which labels and also how many. Below are the confusion matrices for our models.

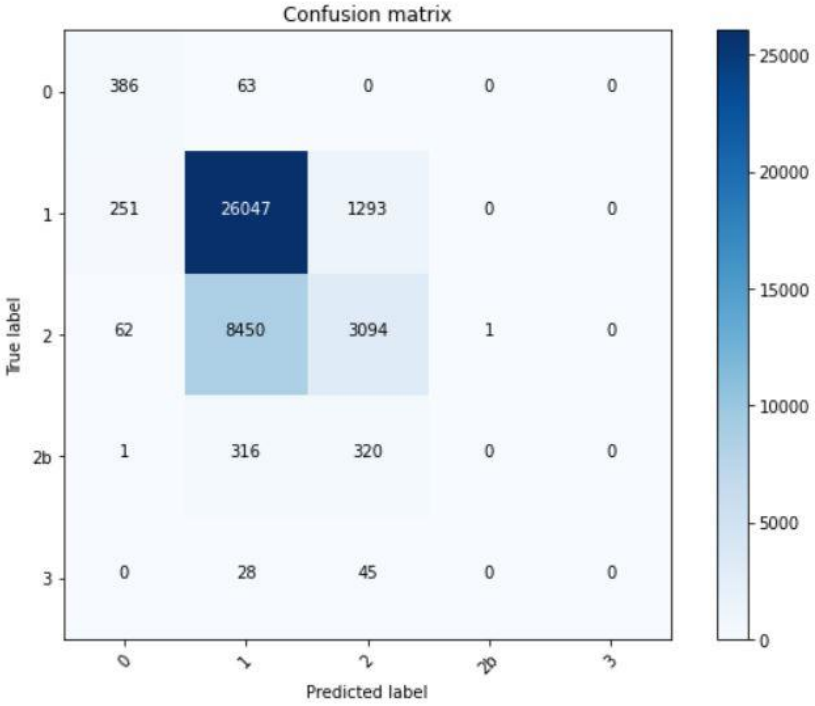


Figure 15. Confusion Matrix for the K Nearest Neighbors model

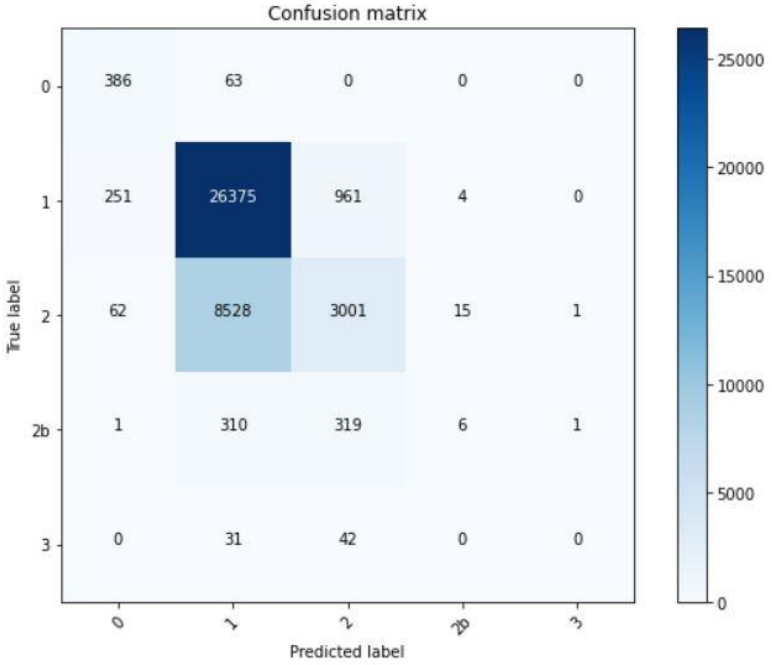


Figure 16. Confusion Matrix for the Decision Tree model

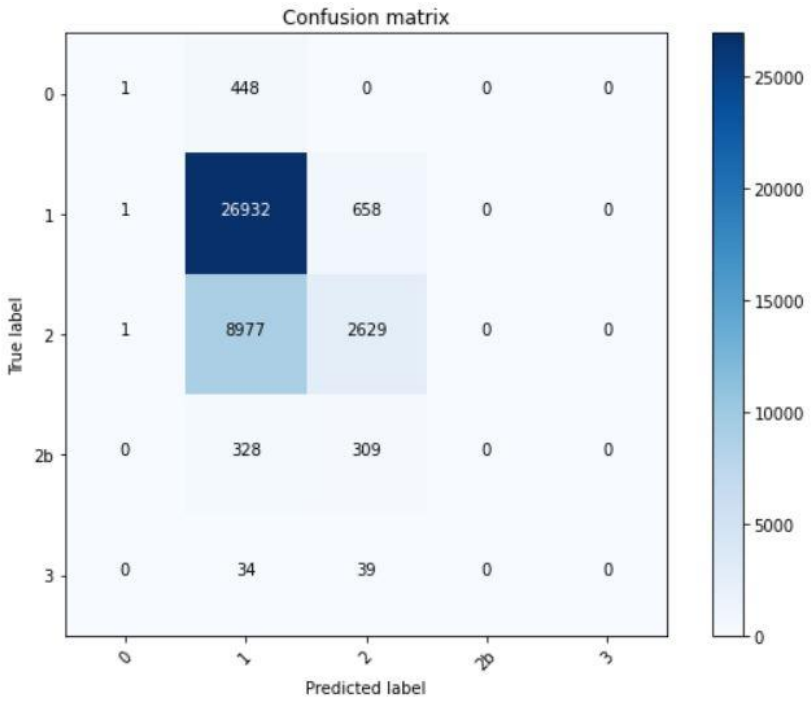


Figure 17. Confusion Matrix for the Logistic Regression model

5. Conclusions

In this study, I analysed the relationship between the severity of car accidents and the data from SDOT on the factors that describe the overall situation the accidents took place in. I identified vehicle count, person count, pedestrian count, weather condition, road condition, light condition among the most important features that determines the severity of a car accident. I built three classification models to predict the severity of future car accidents and suggested to take the model with highest accuracy for implementation. These models can be very useful helping the hospitals and police department to be better prepared for accidents in advance. For example, it can help the hospitals to maintain sufficient amount of staffs and doctors on a given day to handle such situations efficiently, etc.

5. Future Directions

I was able to achieve ~69%, ~68% and ~67% accuracy in the Decision Tree model, K Nearest Neighbors model and Logistic Regression model respectively. However, there was still significant variance that could not be predicted by the models in this study. Many of the relevant features had a lot of missing data as well as data marked as “Unknown”. I think better record keeping could lead to more accuracy to the models.

Our aim was to build the models using only those features which can be pre specified at any given instant of time. But there were features in the original SDOT dataset that only become relevant after a collision has taken place. Some of those features which are relevant can be included to build the models to see whether it results in a better accuracy or not.