Predicting car accidents severity

Oumaima Saadi

30 October 2020

1.Introduction

1.1.Background

Every year the lives of approximately 1.35 million people are cut short as a result of a road traffic crash. Between 20 and 50 million more people suffer non-fatal injuries, with many incurring a disability as a result of their injury.

Road traffic injuries cause considerable economic losses to individuals, their families, and to nations as a whole. These losses arise from the cost of treatment as well as lost productivity for those killed or disabled by their injuries, and for family members who need to take time off work or school to care for the injured. Road traffic crashes cost most countries 3% of their gross domestic products.

Car accidents are usually caused by bad weather ,maybe a bad light , a bad road or maybe the driver is under influence ,and it can also be caused by high speeding , the severity of car accidents changes based on those features , and based on the type of collision , for example the severity of a Rear-end accident changes based on circumstances like speed.

1.2. Problem

Data that can be used to determine car accidents severity have to include severity code, road condition,light condition,speeding,if the driver is under influence, if the accident was caused by inattention, the collision type, and the adresse of the accident .The project aims to predict car accident severity based on some features.

1.3. Interest

The prediction of car accidents severity seems important , because it will help us to know conditions and features that cause accidents with high severity and injuries , so as we can avoid them by changing features that we have the ability to change like light or road condition for example.

2.Data acquisition and cleaning

2.1 Data Sources

Car accidents severity and circumstances dataset , was given by IBM on its data science specialization ,it includes all collisions provided by SPD and recorded by Traffic Records from 2004 to present.

2.2 Data Cleaning

The dataset had at first 38 columns and 194672 rows, once downloading it I realised that there are many problems.

the dataset had a lot of missing values , so i had to analyse each column to decide if those missing values had to be dropped or i should replace them by a specific value .

There was a major problem that i could not solve , it had relation with “speeding” and “inattentionInd” columns , the “speeding” column had 95% missing values, “inattentionInd” column, witch represente if the accident was caused by the driver inattention or not ,has 84% missing values , so i couldn’t drop all this rows because the dataset will become unsatisfying and short to make good decisions , also i couldn’t replace those values by another value , and those columns were important for predicting car severity.

The “underinfl” column which represent if the driver was under influence or not, had 4 types of values “N” and “0” which means the driver was not under influence,“Y” and ‘1’ which means the driver was under influence so i had to convert the two letters into numbers 0 and 1, or to do the opposite.

2.3 Feature Selection

After cleaning the data , there were 187504 rows and 38 columns , after examining the meaning of each feature , it was clear that there were features that have to be deleted .

The “speeding” and “inattentionind” feature were deleted ,even if they are important , because i couldn’t work on the data given after deleting 95% of it .

A lot of columns contain some information that doesn’t affect the severity column , so it isn't important in our project so it had to be deleted.

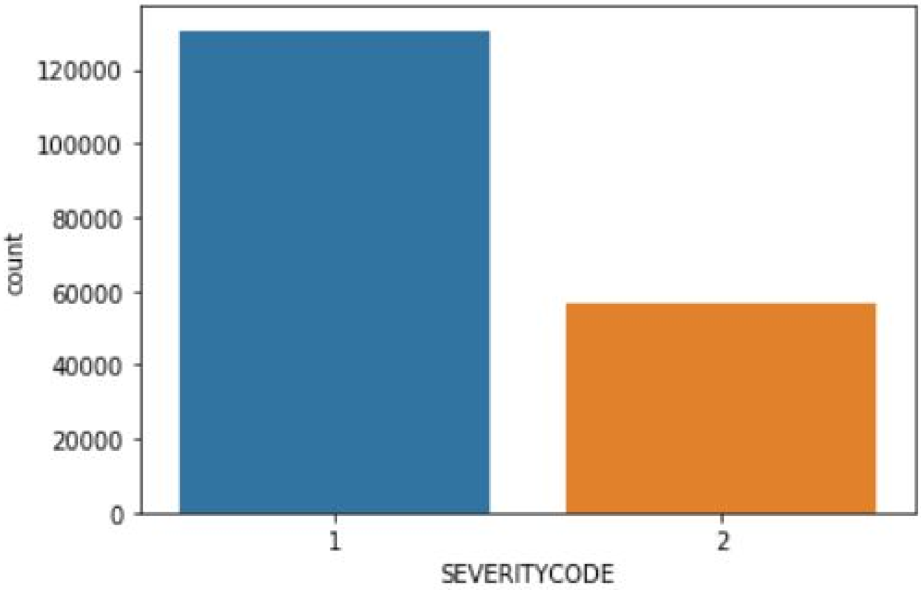
Finally we kept only 7 columns that seemed to be the most important.

Kept features	**Deleted features**	**Reason for dropping features**
SEVERITY COLLISIONTYPE ADDRTYPE	X,Y,INTKEY,LOCATION , EXCEPTRSNCODE, EXEPTRSNDESC,ST COLCODE,ST COLDDESC	the data is not important to predict the severity of car accidents
LIGHTCOND ROADCOND WEATHER UNDERINFL	INATTENTIONIND PEDROWNOTGRNT SPEEDING	more than 80% of the data were missing values.

3.Exploratory Data Analysis

3.1 Severity distribution

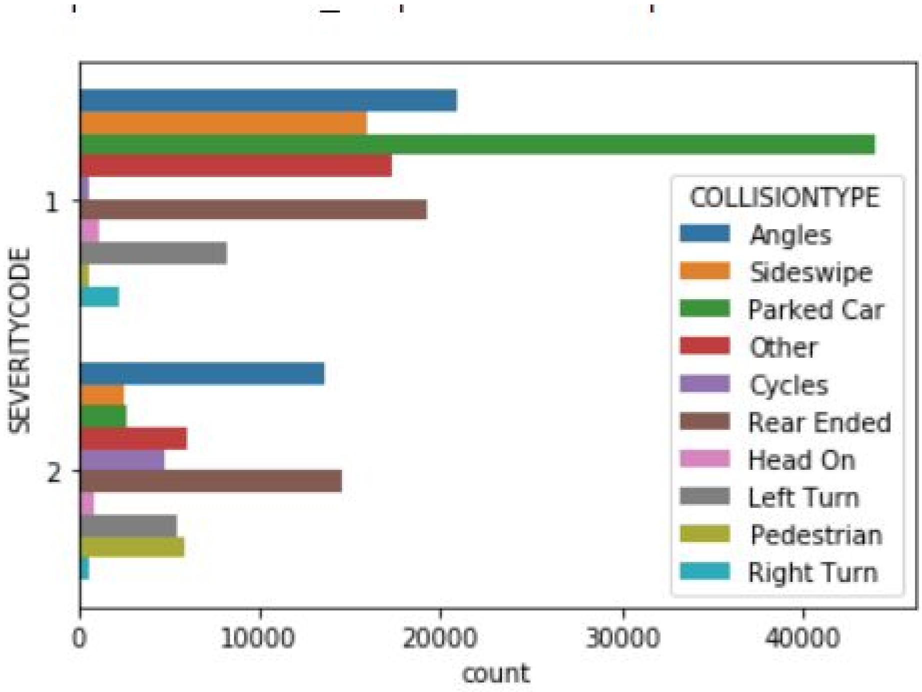
After plotting the number of each severity ,it’s clear that most accidents that belong to our data have property damage (severity code = 1) , so we can say that accidents with property damage are more frequent in the US than accidents having injuries.



So with the data given we can not compare for each type of feature (collision type for exemple) the number of accidents having property damage severity with the number of those having injuries , because for all collision types we will have the same results ⇒ accidents with Severity code 1 are more than those with severity code 2.

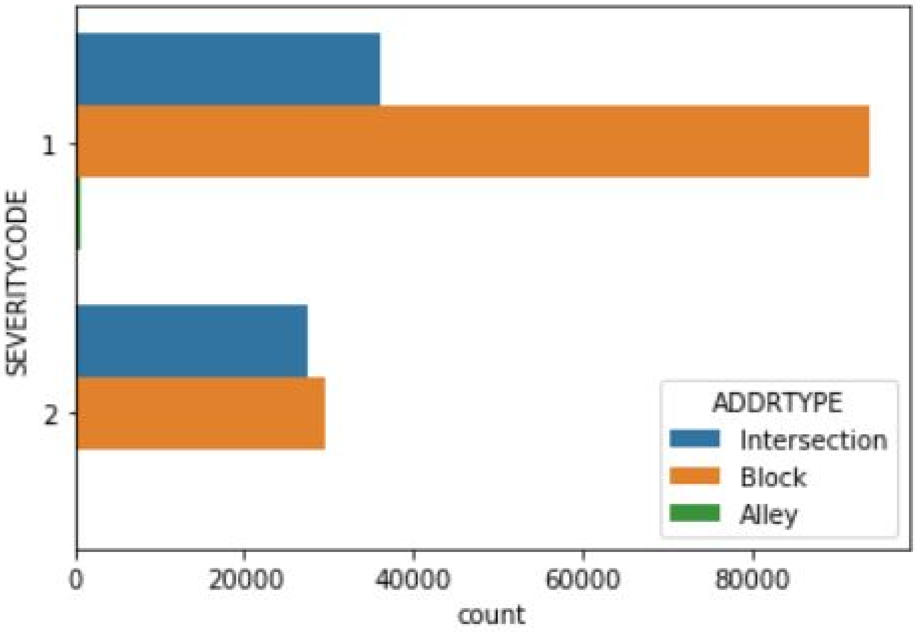
3.2 Relationship between severity and collision type

We can say that accidents that have only property damage are mostly parked car accidents or Angles accidents , where accidents which have injuries are mostly Rear-end accidents.



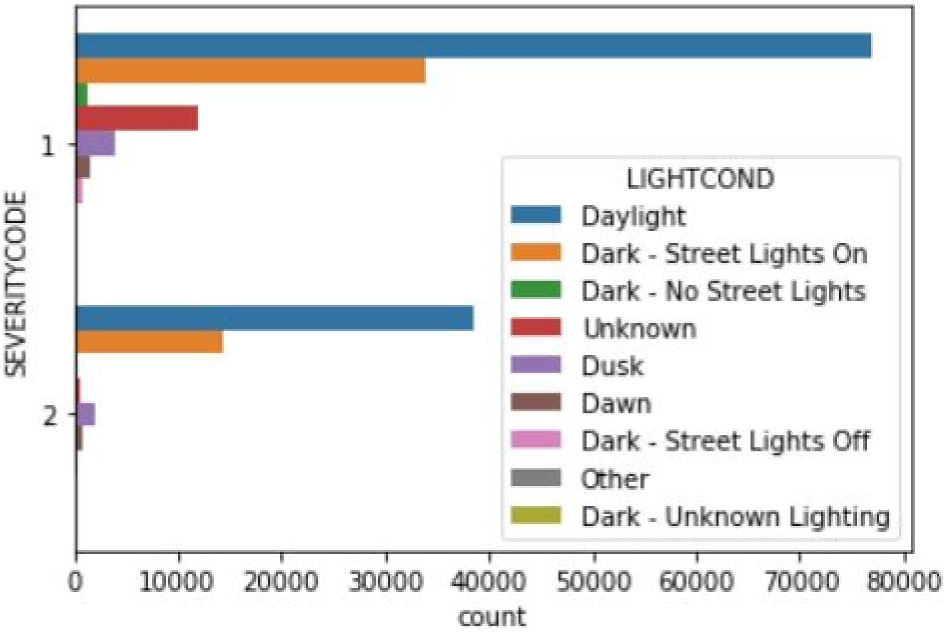
3.3 Relationship between severity and address type.

We can say that accidents with only property damage happen usually in blocks and sometimes in intersections whereas accidents that cause injuries happen usually in intersections and blocks, so we can conclude that accidents in intersections usually cause injuries.



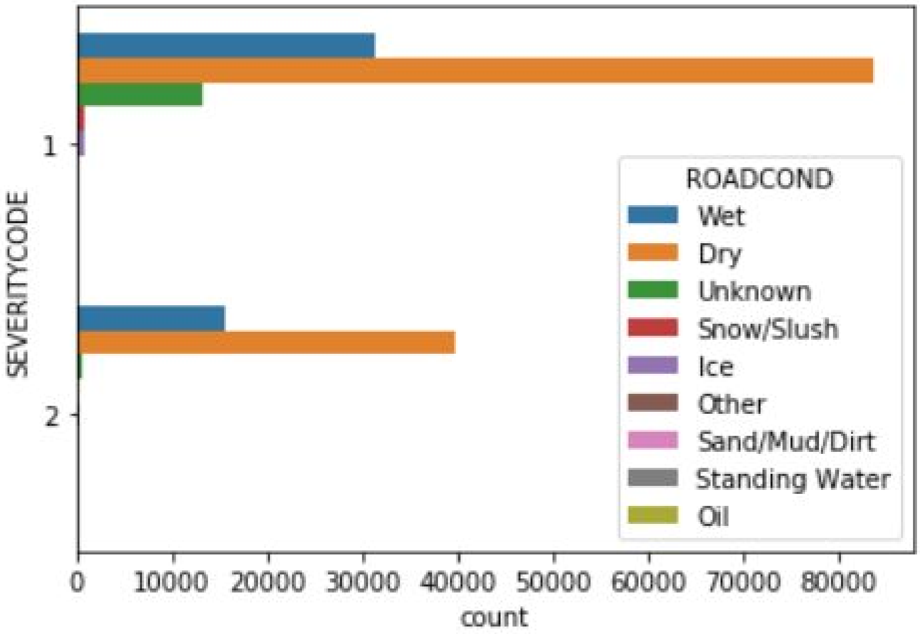
3.4 Relationship between severity and light condition .

Accidents are usually happening in the daylight condition and in the dark with street lights on and may sometimes have only property damage , or may have injuries.



3.5 Relationship between severity and road condition .

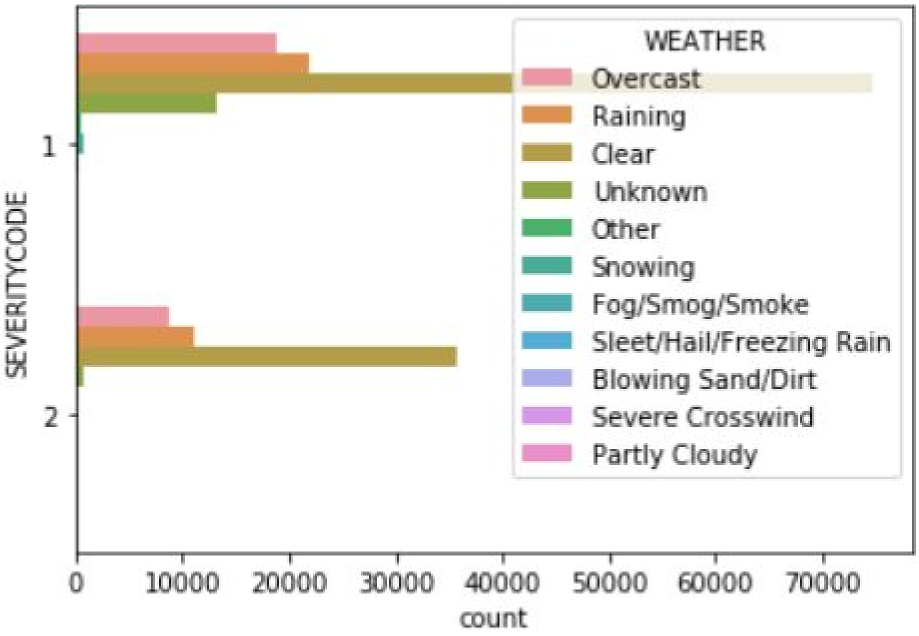
Accidents with property damage and those that cause injuries are happening usually in dry roads and sometimes in wet roads, and rarely when there's snow or ice in the road.



3.6 Relationship between severity and the weather .

For accidents that cause injuries , we can say that they are mostly happening when the weather is clear , and may happen when it’s raining or overcast , this can be explained by how drivers are acting : when it’s raining , drivers try to drive slowly and to pay attention , so accidents are less compared to when the weather is clear.

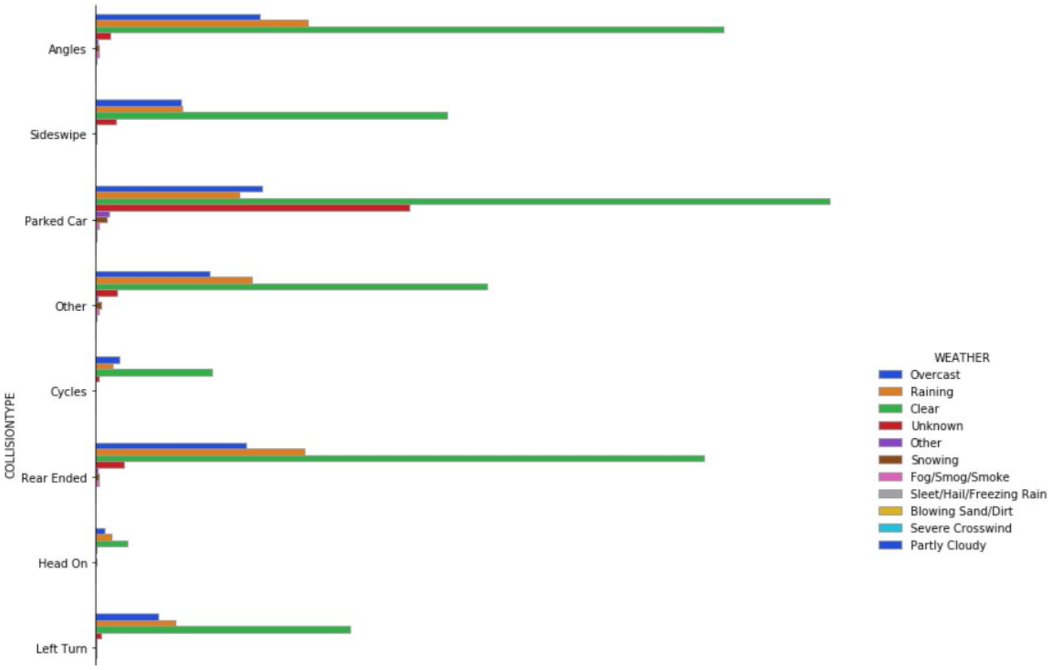
We can conclude the same thing for accidents that cause property damage.



3.7 Relationship between collision type and the weather .

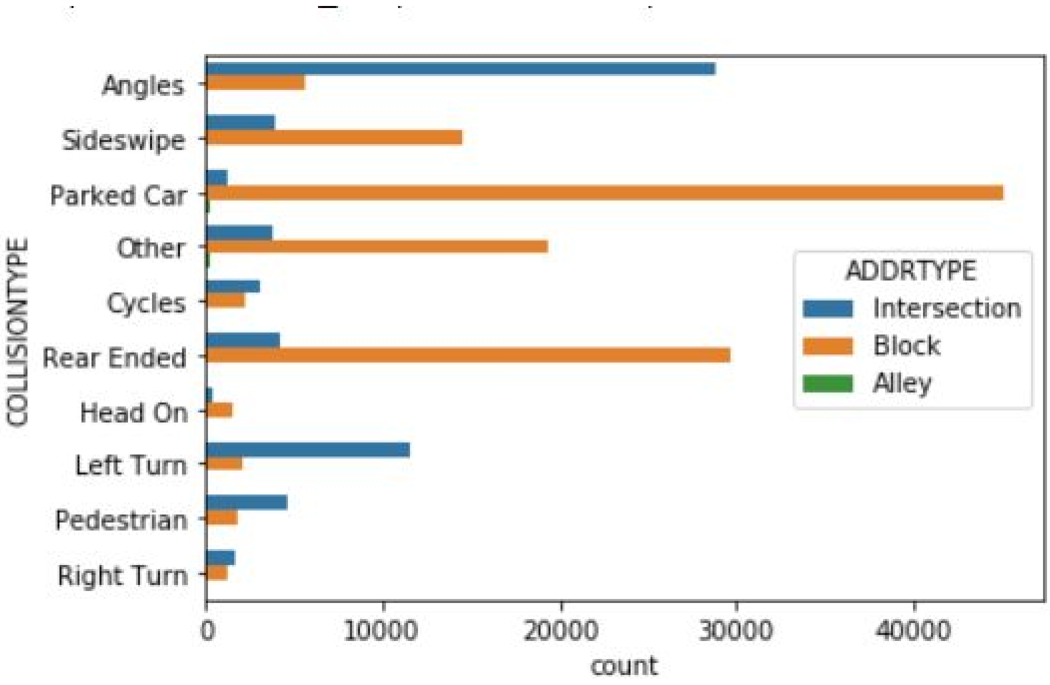
We can notice that accidents generally happen in clear weather .

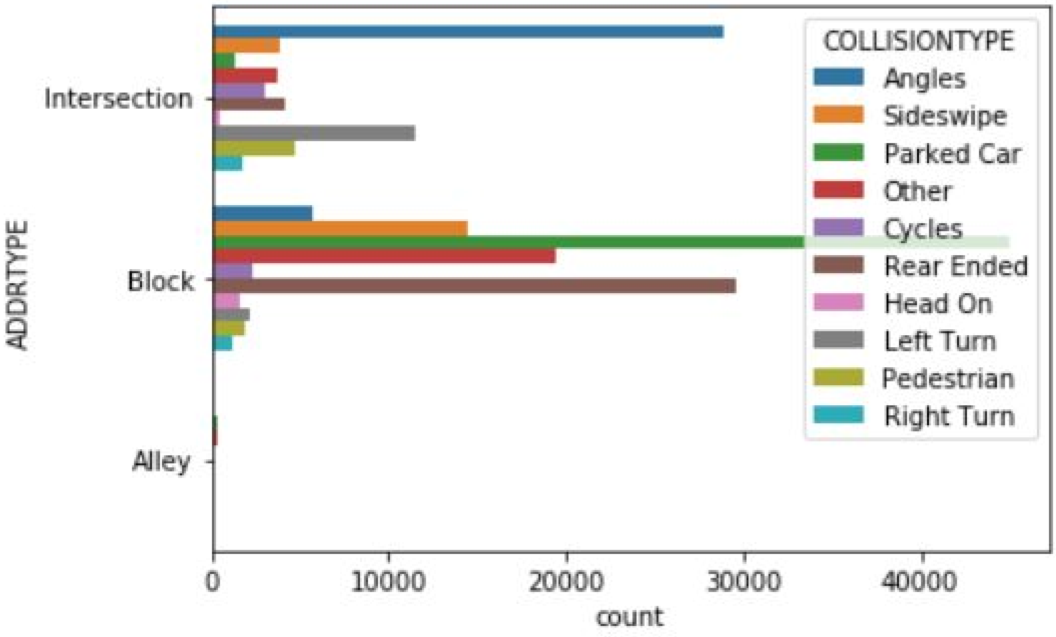
Parked car accidents happen when it’s overcast more than when it’s raining, but for all other types of accidents we can say that they happen when it’s raining more than when it’s overcast.



3.8 Relationship between address type and collision type.

We can notice that accidents that happen in intersections are mostly angles and left turn accidents ,in blocks parked car , rear ended, and sideswipe accidents happen the most.



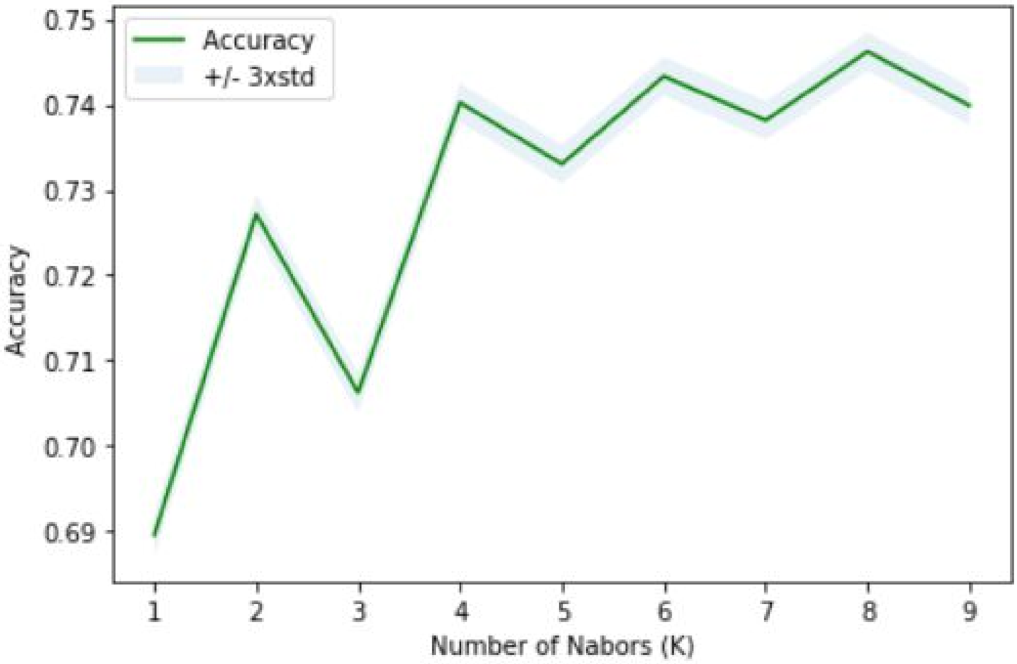


4.Predictive modeling

Despite we have only categorical data , i decided to use classification models : KNN ,tree decision, svm and logistic regression

4.1 KNN algorithm

Before deciding which number of neighbors i have to use, i decided to plot for all numbers from 1 to 9 the accuracy that matches with them and then decide the number of neighbors that give the high accuracy.



So we can conclude that k=8 is the number of neighbors which have high accuracy so we will use it as a parameter for building our model.

4.2 Other algorithms

After training our models , based on knn, svm, tree decision and logistic regression I calculate some different scores to evaluate those models ,the table shows the evaluation.

Algorithm	**Jaccard**	**F1-score**	**LogLoss**
KNN	0.75	0.69	NA
Decision Tree	0.75	0.69	NA
SVM	0.74	0.68	NA
LogisticRegression	0.7145	0.6735	0.56

5. Conclusion

In this study , I analyzed the relationship between features like light ,road ,weather,address, and collision type and the severity of an accident , i identified the impact of each value of a feature ,on the severity, and tried to predict if a car accident will have only property damage or will cause injuries ,based on different features using some classification algorithms like knn and tree decision , those models can be very useful on predicting the car accident severity in a specific road, light,weather and address.