Prediction of severity of vehicular accidents

By Abraham Vergara

September 29, 2020

1. Introduction

The automobile is one of the most common methods of transportation worldwide. And according to the Centers for Disease Control and Prevention approximately 3,700 people die each in day in a car crash around the world, that amounts to nearly a million and a half people each year.

It is important then to look for methods and prevention strategies to be able to reduce the amount of accidents. This would not only result in saving human lives but could reduce drastically costs, to parties such as the government, health organizations or even the individuals themselves.

Some of the most contributing factors to driving accidents are: falling asleep at the wheel,loss of vehicle control (due to mechanical reasons or external e.g. the weather, other drivers), blind left turns, not staying in the proper lane.

Some of this factors can be attenuated, as is taking extra precautions during bad weather conditions or possibly even postponing the trip if possible.

The purpose of this work is to see the relationship between certain driving conditions and what type of accidents are drivers exposed to: severe, fatal or a slight driving accident. Also to develop a model that can predict the outcome of an accident taking into account the condtions of the road and some other factors.

This type of work could be of interest to the government: to be able to see the type of maintenance required on works, to insurance companies which can now have a better assessment of the probable type of accidents a driver can be involved in and finally to the drivers, it can help reduce the risks they take when driving.

2. Data

The dataset is under a OGL (Open Government License)

It has records of road accidents from the year 2005 to 2014 in United Kingdom. It is comprised of 4 files: the accident file which contains information about the accident severity, weather, lcocation, date, hour, day of the week, etc.
The vehicle file: which contains the details about the vehicles involved.
The casualty file which contains information about the severity, age, sex, casualty type, etc.
And finally a lookup file which has the description for the previously mentioned files.

For this particular study, the files that will be used are the accident file and the casualty file which contains the information deemed necessary to develop an adequate model to predict the severity of and accident taking into consideration the environmental variables.

The dataset can be found in [UK Accidents 10 years history with many variables](https://www.kaggle.com/benoit72/uk-accidents-10-years-history-with-many-variables/notebooks) at the Kaggle website.

Most of them have are numeric attributes which represent a particular state, however in the development notebook this numeric indices have been replaced by their corresponding 'word' representation to make the charts and information more easily understandable.

This notebook will be used for the Capstone Project of the Coursera Data Science Specialization. The project consists on a model to predict the severity of an accident taking into account the conditions of the day and the road.

Some of the more relevant attributes on the dataset are the following:

Table 1. Used attributes of dataset

Field	Description
Accident_Index	Unique identifier of the accident
Accident_Severity	Describes severity of accident (desired label)
Date	Date of the accident
Day_of_Week	Day of the week when the accident occurred
Road_Type	Type of road where the accident occurred
Speed_limit	Speed limit in the zone the accident occurred
Junction_Details	Description of type of junction where the accident happened (roundabout, 4-way,etc)
Junction_Control	Type of control at the junction (trafficlight, stop sign, etc)
Light_Conditions	Lighting conditions at time of accident
Weather_Conditions	Weather conditions when at time of accident
Road_Surface_Conditions	Conditons of the road surface at location of accident
Urban_or_Rural_Area	Area of acciden
Vehicle_Type	Description of Vehicle (car, van, truck, etc)
Age_Band_of_Driver	Age range for driver
Age_of_Vehicle	Vehicle's age

3.Methodology

The project consists on an analysis on the data discover the most relevant factors that contribute to vehicular accidents and that determine the severity of the accident. The analysis is limited to data from the years 2005 to 2010 in the UK.

The data is collected from Kaggle website and a first exploration is conducted. The most relevant attributes e.g. (weather,lighting,road conditions, etc.) that could contribute to determining the severity of the accident are left in a dataframe and the others are dropped. Next, data cleaning takes place, by dropping rows that have incomplete information or not useful to the analysis. These records are dropped with further impact on the project, due to the vast amount of records available.

Once the analysis is made, a model (k Nearest Neighbor) to determine the severity of a possible accident is developed. This will allow to establish which conditions lead to a more dangerous outcome in case of required travel.

4. Analysis

The first question to answer for the analysis is: what is the proportion of severity of accidents?

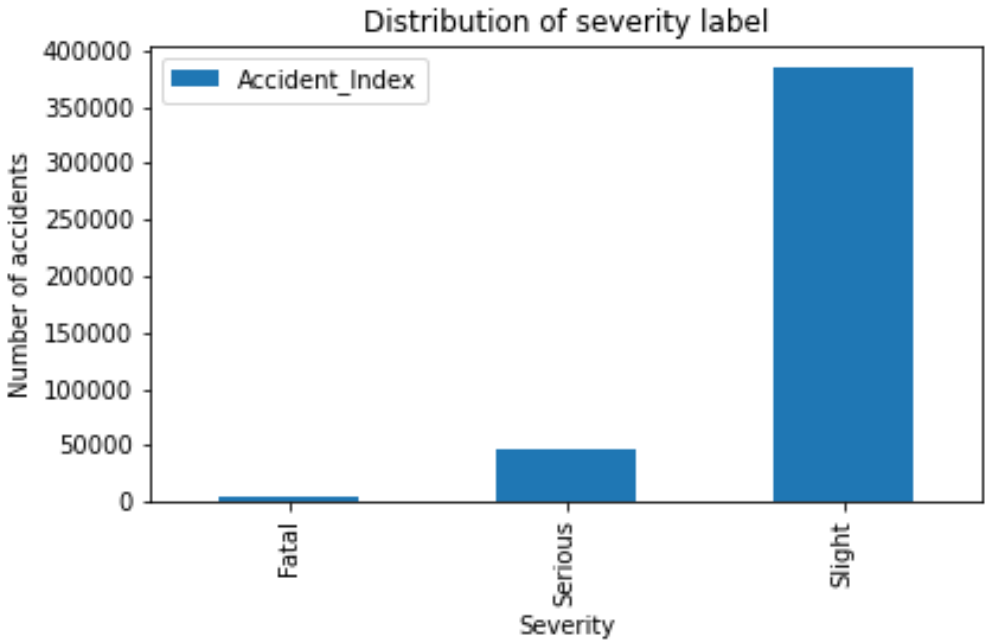

Figure 1. Initial proportion of label in dataset

As it can be seen the majority of accidents are not severe accidents. The proportions for the three classifications: fatal, serious and slight are as follows 1 percent, 11 percent and 88 percent. After this initial exploration, now is important to investigate the circumstances and attributes that contribute to the occurrence of the accident itself.

First, external factors such as weather, lighting and road conditions are explored. In the following image it can be seen the amount of accidents that occur and under which weather conditions. Some of the

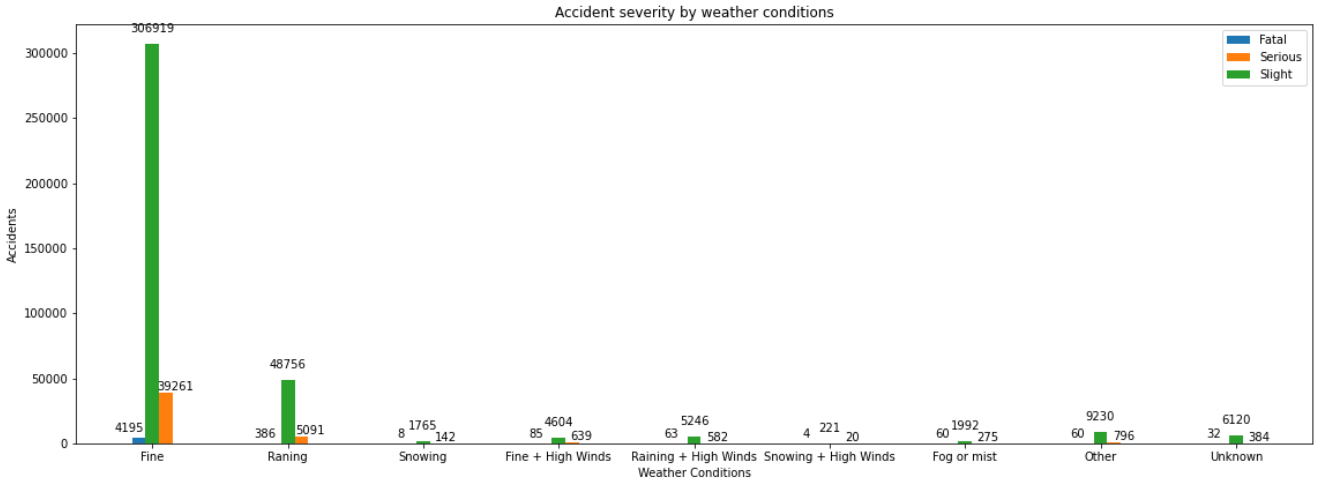

Figure 2. Figure of accident severity vs weather

conditions presented are: fine weather, raining, snowing, fine and high winds, raining and high winds, snowing and high winds, fog or mist, other and unknown. As it can be seen from the bar chart the majority of accidents occur under fine weather, for all three of the classifications. Still the most common severity is slight then serious and finally a severe accident.

The next external factor is lighting. Which is explored in the following figure.

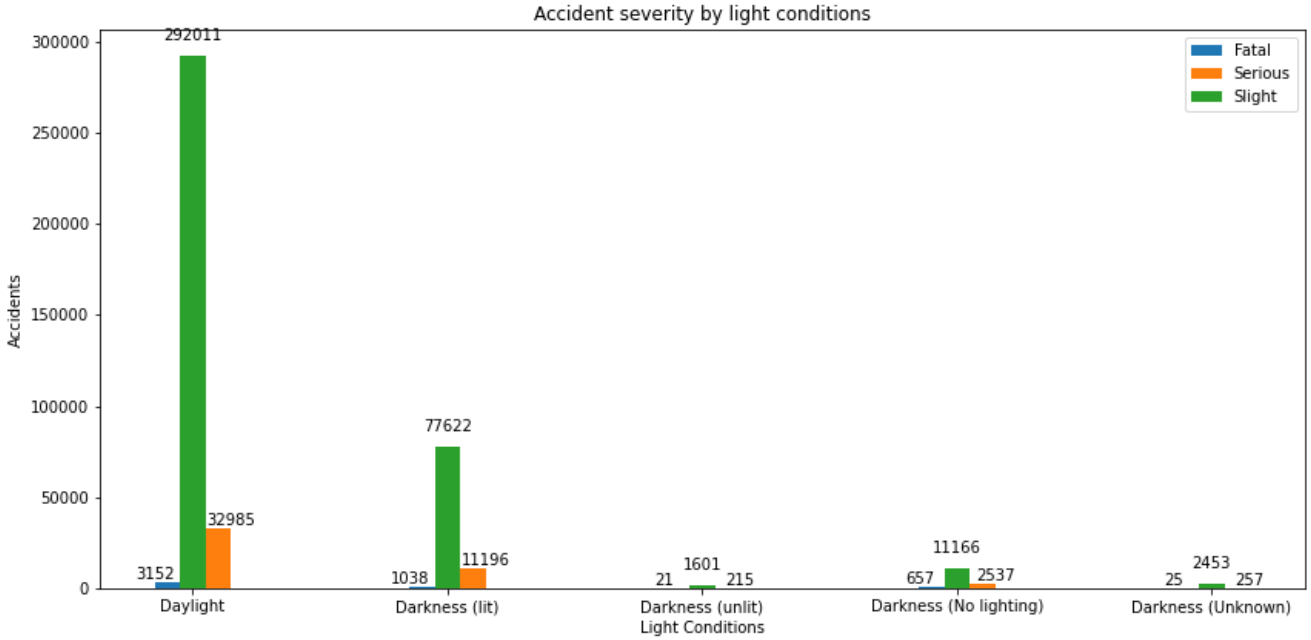

Figure 3. Accident Severity vs Lighting condition

In the above figure 3 it can be seen that the majority of the accidents occur during daylight, this trends applies for the tree severity classifications as shown. And some of the other lighting conditions are darkness (lit), darkness (unlit), darkness (without any type of lighting) and darkness unknown.

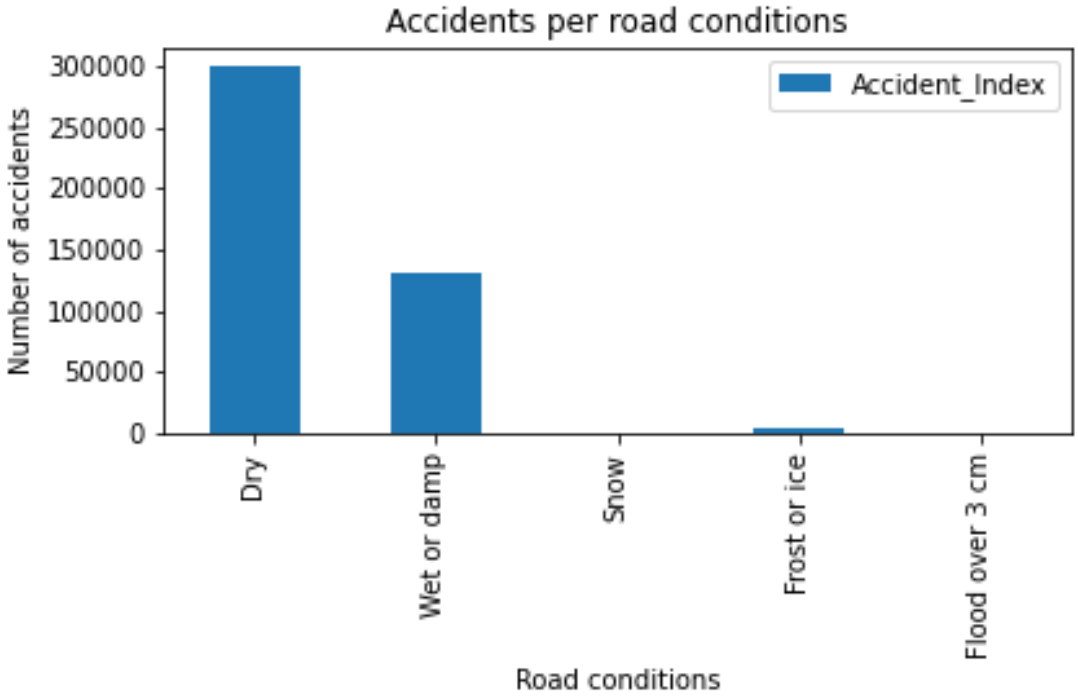

Figure 4. Accidents vs road condition

Another possible factor is the conditions of the road itself which is explored in the following figure.

The same trend observed in the previous figures continuous, the conditions which could be considered favorable for driving is the one that amounts the most accidents. The conditions present in the figure are: dry, wet or damp, snow, frost or ice and flood over 3 cm.

Once that factors directly related to the road are considered, the next would be circumstances that could affect the driver itself. Such as speed limit, day of the week, or internal variables such as age and sex of the driver.

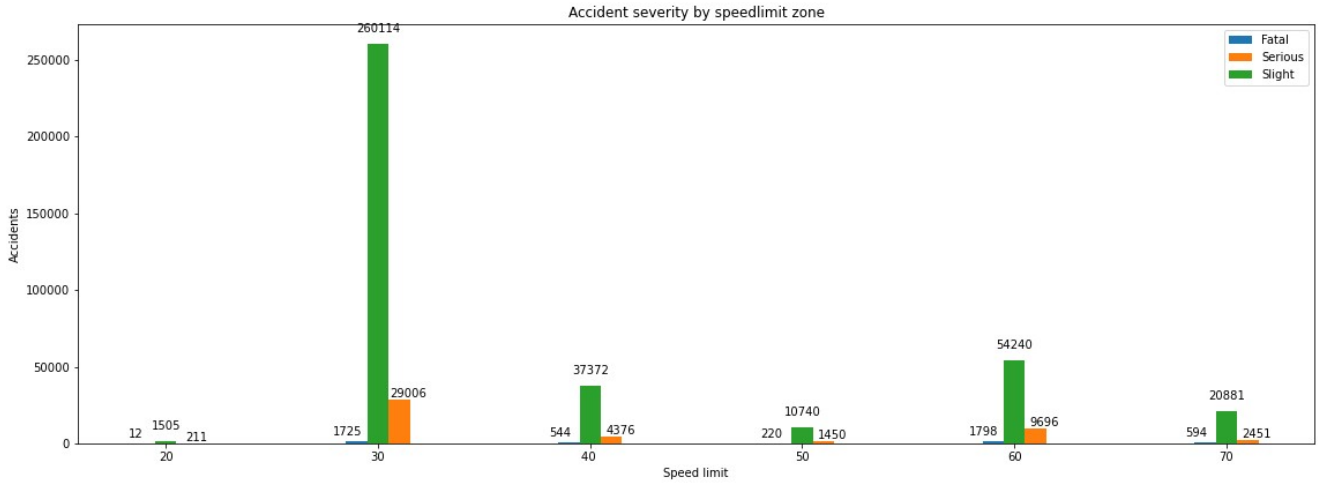

Figure 5. Accident severity by speed limit zone

The figure 5 presents the total of accidents grouped by severity in each of the speed limit zones. It can be seen that the zone with the most accidents is the 30 kilometers per hour. Once again, an attribute that could be considered 'favorable' for the driver is the one with the largest quantity of accidents.

One attribute that does have a considerable weight on the severity of the accident is the sex of the driver. In the following figure the amount of accidents by sex is presented.

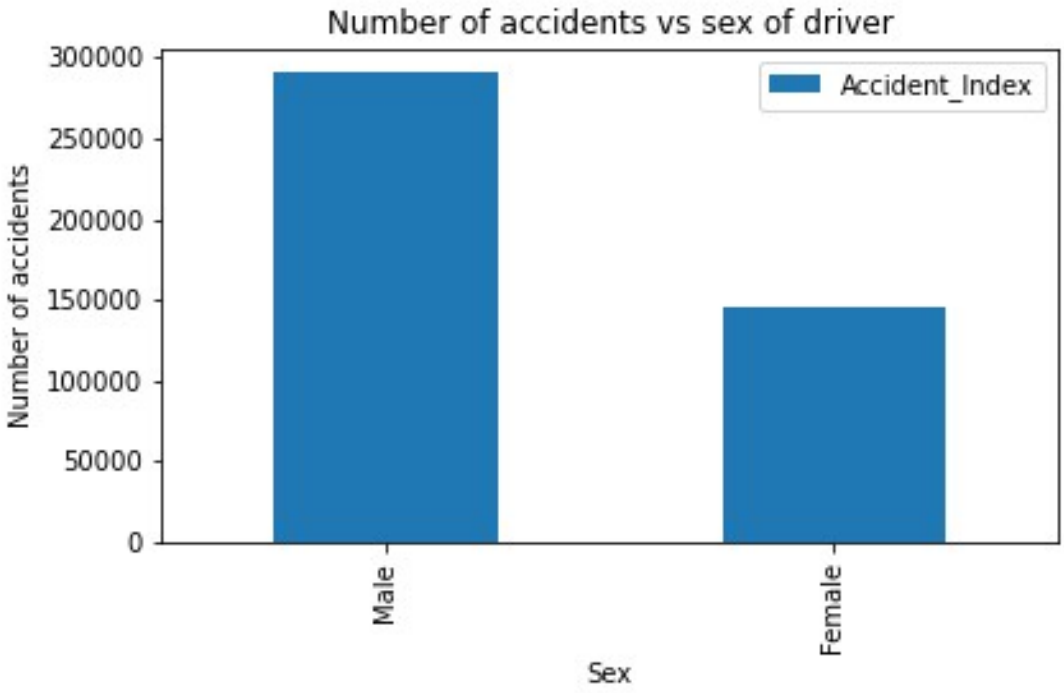

Figure 6. Accidents vs sex of driver

As it can be seen, the amount of accidents for the male drivers is considerably larger than the female one. And finally the age of the driver is presented in the following image.

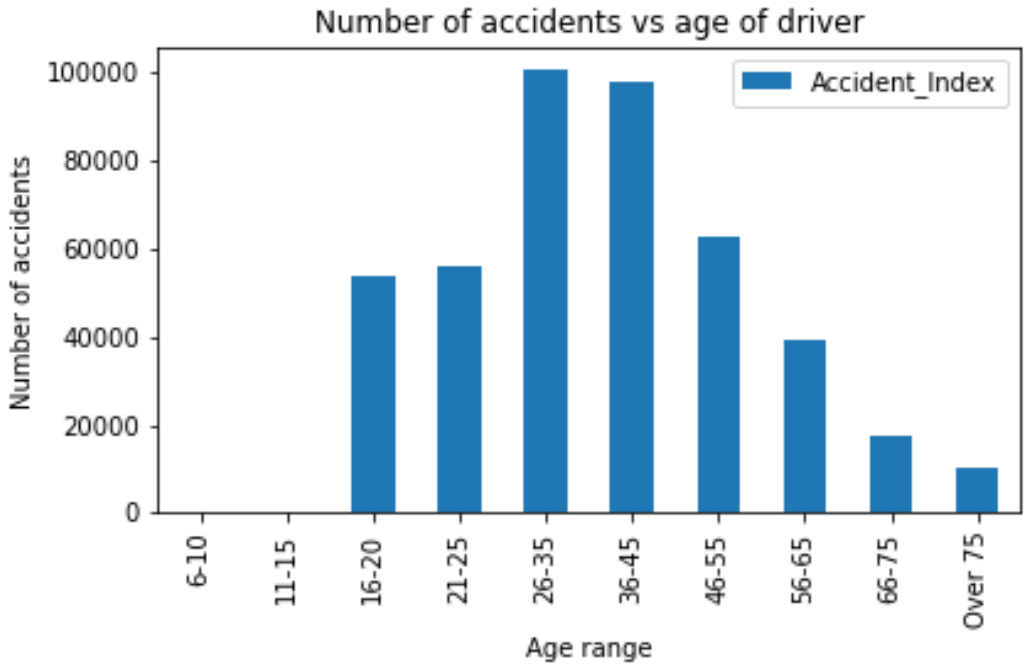

Figure 7. Accidents vs age range

The range with the most accidents is from 26 to 45 years old. This could be related to the fact that the majority of drivers are in this age range. This can not be determined directly from this data and would need to be explored in a subsequent project or data exploration.
Finally on last factor considered was the day of the week when the accident occurred. And as t can be seen from the following figure, it does not have a big impact on the amount of accidents that occur. Although the numbers raise as the week approaches Friday. Either way the reasons can not be determined at the moment being, with the data available.

For a model the K Nearest Neighbor approach was used. First data was split into training set and testing set using a 80/20 proportion. Dummy variables were used to make the data appropriate for the model. The training set was used to determine the best k.

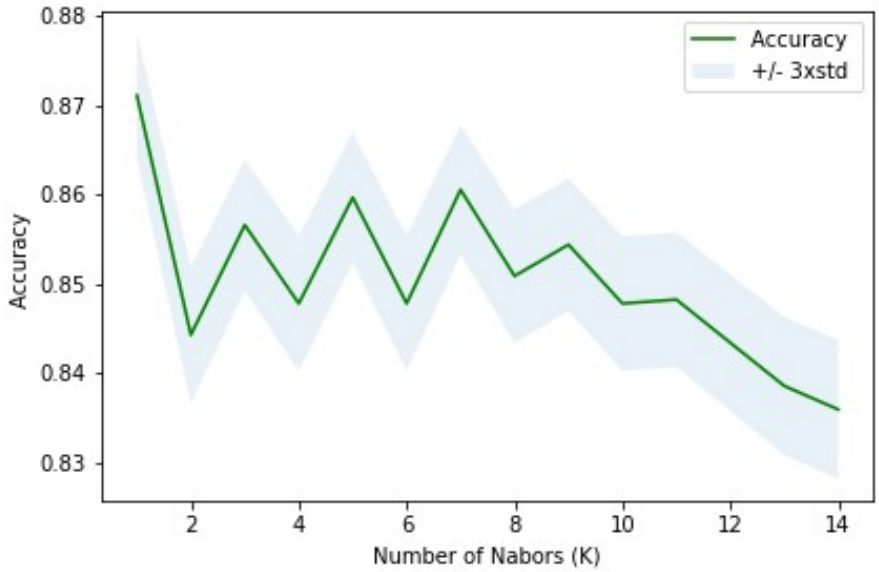

Figure 8. Finding of k for model

Then the model was trained and as metrics the Jaccard Index and F1-Score were used.

Table 2. Metrics of evaluation

Jaccard Similarity Score	F1-Score
0.77	0.87

For the purpose of this project and as an initial approach this could be considered satisfactory.

5. Results and discussion

Several things can be noted from the initial analysis. Contrary to a possible preconceived notion that it is probable that most accidents occur during poor external conditions (Light, Weather, Road, etc). The analysis reveals that it is not the case. Most accidents occur during good weather with no external conditions to impair the driver.

Most accidents occur in low-speed zones. And it is important to remark that there is a significant difference in the percentage of men and women who have accidents. One other factor is that there is no relation between the day of the week and the amount of accidents, except for Fridays which have a slightly higher number of accidents per day.

The biggest amount of accidents occur during daylight and the second most accidents occur during nigth time under well-lit conditions.

The model developed using K-NN has an Jaccard index of: 0.78 and F-1 Score of 0.88.

6. Conclusion

The results lead one to believe that most accidents occur due to internal, driver reasons and not external. Since good external conditions yield a much lower number of accidents. This could be to a number of

reasons: with good conditions drivers feel much more confident and do not exercise as much caution. Or there is more traffic on the roads. However, the initial hypothesis that bad weather conditions was the main cause of vehicular accidents can be discarded and some other possibilities need to be contemplated and explored, perhaps in other projects or data analysis. At the moment even though the model works, I would consider the possibility of other factors having a bigger impact on the severity of an accident, factors which are not available in this dataset and require a new study.

The model and determining the severity of an accident could see its main application with insurance companies, which could probably be the most interested party due to the reasons that their costs are relative to the probability of a driver having an accident.