Predicting Traffic Accident Severity

BLESSED MUTENGWA

SEPTEMBER 2020

Table of Contents

1. Introduction

1.1 Background

Road traffic accident statistics are recorded everyday due to different reasons, which some will be reviewed in the following passages. According to a research conducted by the World Health Organization (WHO) there were 1.35 million road traffic deaths globally in 2016, with millions more sustaining serious injuries and living with long-term adverse health consequences. Globally, road traffic crashes are a leading cause of death among young people, and the main cause of death among those aged 15 to 29.

Analysis of some of the above mentioned contributions to road severities to make a model for predicting the chances of these accidents can be designed and implemented for future use. Such insights, could allow law enforcement bodies to allocate their resources more effectively in advance of potential accidents, preventing when and where a severe accident is bound to occur. By so doing, it will result in saving resources, life and unnecessary expenses unaccounted for.

If the system is well adopted, and cities trust the predictions from the model, nations could be forewarned before the execution of accidents by announcing on radios, and broadcasting on televisions which routes to use on a particular day to avoid traffic jams or to suffer the effects of bad weather.

Governments should be highly interested in accurate predictions of the severity of an accident, in order to reduce the time of arrival and thus save a significant amount of people each year. Others interested could be private companies investing in technologies aiming to improve road safeness. Why run costs for rushing the injured to the hospital, fuelling fire trucks to accident scenes to rescue the injured when it can be avoided before it happens. As with most Governments that have its people at heart, they should be more interested in adopting the severity predicting systems for the safety of its people.

1.2 Problem

Data that might contribute to determining the likeliness of a potential accident occurring might include information on previous accidents such as road conditions, weather conditions, exact time and place of the accident, type of vehicles involved in the accident, information on the users involved in the accident and of course the severity of the accident as well as the road condition. This project aims to forecast the severity of accidents based on past data.

1.3 Interest

Governments should be highly interested in accurate predictions of the severity of an accident, in order to reduce the time of arrival and to make a more efficient use of the resources both in usage and distribution, and thus a significant number of people are saved each year. Others interested could be private companies investing in technologies aiming to improve road safety and network coverage.

2. Data

2.1 Data source

The original dataset for this project was downloaded from the Applied Data Science Capstone Project from the following link : https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv along with its metadata pdf. I selected the most relevant attributes for use as features for the prediction of accident severity.

2.2 Feature Selection

For a better prediction on the target variable, the selection of the best correlated features was very useful and important. Feature selection included the following columns and the description of how they contributed when they were recorded SEVERITYCODE, COLLISIONTYPE, OBJECTID, ADDRTYPE, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, INJURIES, SERIOUSINJURIES, FATALITIES, JUNCTIONTYPE, WEATHER, ROADCOND, LIGHTCOND, HITPARKEDCAR. The SEVERITYCODE has a code that corresponds to the severity of the collision: that is a 3 if it was a fatality, or a 2b if it was a serious injury, or a 2 if it was an injury, 1 if it was a prop damage and lastly a 0 if it was an unknown. COLLISIONTYPE recorded the type of collision which occurred. OBJECTID has the ESRI unique identifier. ADDRTYPE has the collision address type: • Alley • Block • Intersection. PERSONCOUNT has the total number of people involved in the collision. PEDCOUNT has the number of pedestrians involved in the collision, this is entered by the state. PEDCYLCOUNT is the number of bicycles involved in the collision, this is entered by the state as well. VEHCOUNT has the number of vehicles involved in the collision, this too is entered by the state. INJURIES is the number of total injuries in the collision. This is entered by the state. SERIOUSINJURIES records the number of serious injuries in the collision. This is entered by the state. FATALITIES has the number of fatalities in the collision. This is entered by the state. JUNCTIONTYPE is a category of junction at which collision took place. WEATHER is a description of the weather conditions during the time of the collision. ROADCOND, has the condition of the road during the collision. LIGHTCOND describes the light conditions during the collision. HITPARKEDCAR tells whether or not the collision involved hitting a parked car. (Y/N)

2.3 Description

The dataset that resulted from the feature selection consisted in 182,895 samples, each one d escribing an accident and 15 different features. These features where the following: recorded SEVERITYCODE, COLLISIONTYPE, OBJECTID, ADDRTYPE, PERSONCOUNT, PEDCOUNT, PEDCYLCO UNT, VEHCOUNT, INJURIES, SERIOUSINJURIES, FATALITIES, JUNCTIONTYPE, WEATHER, ROADCO ND, LIGHTCOND, HITPARKEDCAR.

2.4 Data Cleaning

The main aim of Data Cleaning is to identify and remove errors and duplicate data, in order to create a reliable dataset. This improves the quality of the training data for analytics and enables accurate decision-making. During this stage we filled out missing values, removed rows with missing values so that our model could train with a dataset that had all the conditions filled

with no missing information. I also fixed errors in the structure by encoding my data with numbers which the machine learning algorithms understood in the pre-processing. Lastly I reduced the data for proper data handling which could not conflict with the specs of my working environment

3. Exploratory Data Analysis

First, the distribution of the target's values was visualized. The plot confirmed that it is a balanced labelled dataset as the samples are divided 55-45 with more cases of lower severity.

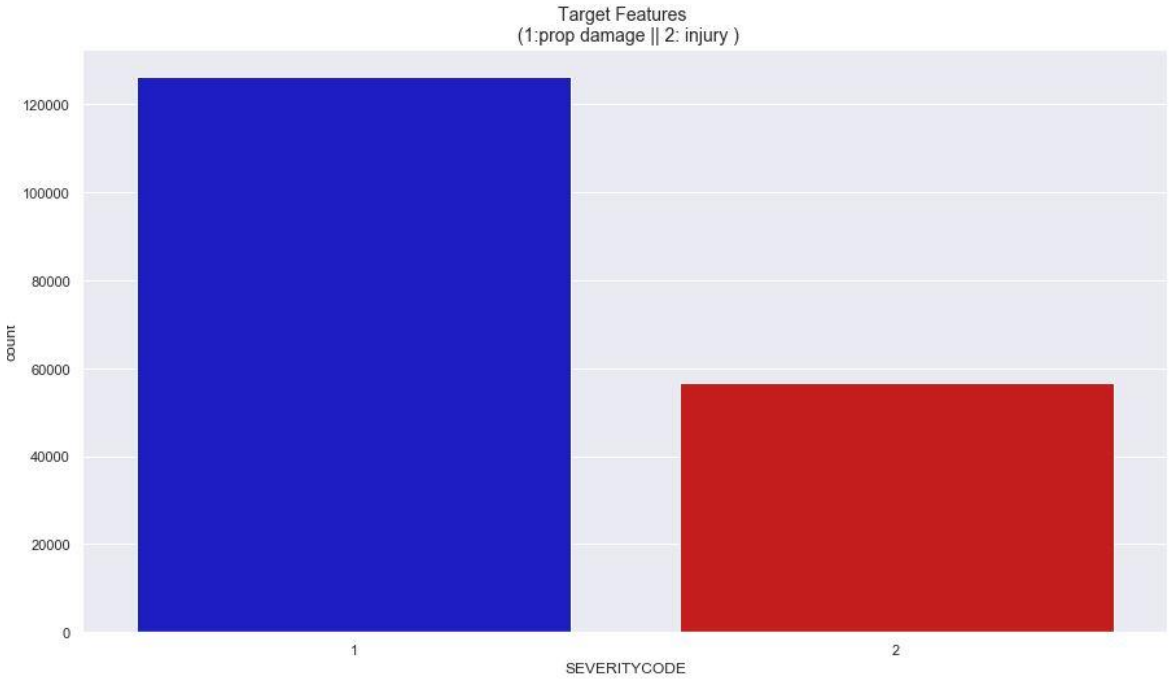

Fig 1 Distribution of severity types

Then a seasonality analysis was performed, visualizing the global trend of daily accidents as well as the amount of accidents grouped by years and the month of the year.

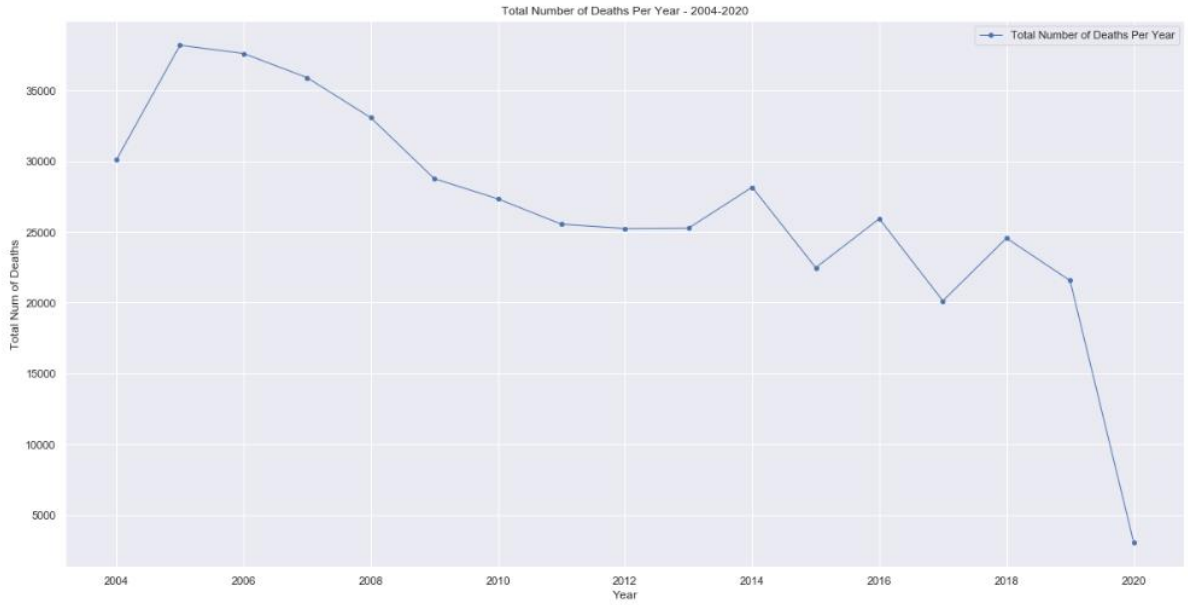

Fig 2 Line plot of total amount of accidents per year.

Total number of vehicles that have been involved in accidents was plotted on a horizontal bar graph which clearly showed that most vehicles were involved in accidents during the 10th month (October).

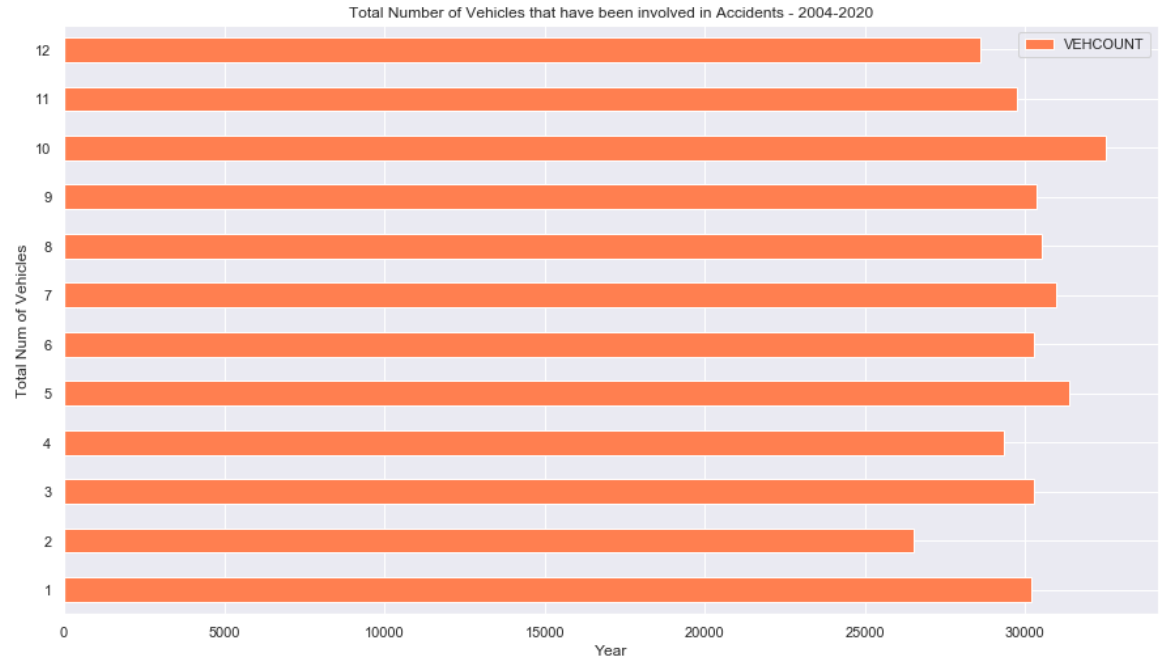

Fig 3 Total Number of Vehicles per Month from 2004 to date

In Fig 4. It is clear that most accidents occurred on the 4th day which if Friday and very low on the 6th day (Sunday).

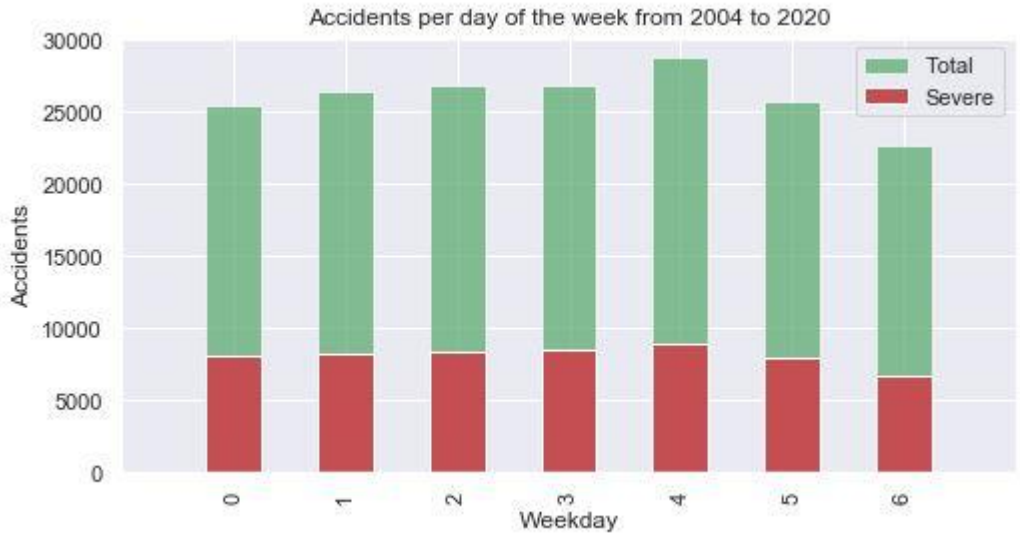

Fig 4 Total number of accidents per day

4. Predictive Modelling

Different classification algorithms have been tuned and built for the prediction of the level of accident severity. These algorithms provided a supervised learning 7 approach predicting with certain accuracy and computational time. These two properties have been compared in order to determine the best suited algorithm for his specific problem. Firstly, the 182,895 rows whe re split 80/20 between the training and test sets, afterwards an additional 80/20 split was per formed among the training samples creating the validation set for the development of the mo dels. Then the data was standardized giving zero mean and unit variance to all features. Three different approaches were used: Logistic Regression, K-Nearest Neighbour, Supervised Vector Machine The same modus operandi was performed with each algorithm. With the train and v alidation sets the best hyper parameters were selected and using the test set the accuracy an d computational time for the development of the models were calculated. The decision tree m odel was upgraded to the random forest. With the default random forest, the features were s orted by impurity based importance in the prediction of the severity. Thus, the 10 least impor tant features were dropped to decrease the computation complexity for the KNN and SVM m odels.

After evaluating the parameters for each algorithm these were the results:
Logistic Regression: c=0.01.
KNN: k=14
SVM: size of the training set= 75,000 samples.

The following visualizations show how the parameters for KNN and SVM models were selecte d. The SVM model is computationally inefficient with huge sample sets. Therefore, an equilibr ium between accuracy and computational time was a found evaluating different training sizes . The training set was reduced from 117,052 to 75,000 rows. On Figure:7, the accuracy is incr easing as the training size does, however Figure:9 shows how this comes with an important in creasing of the computational time.

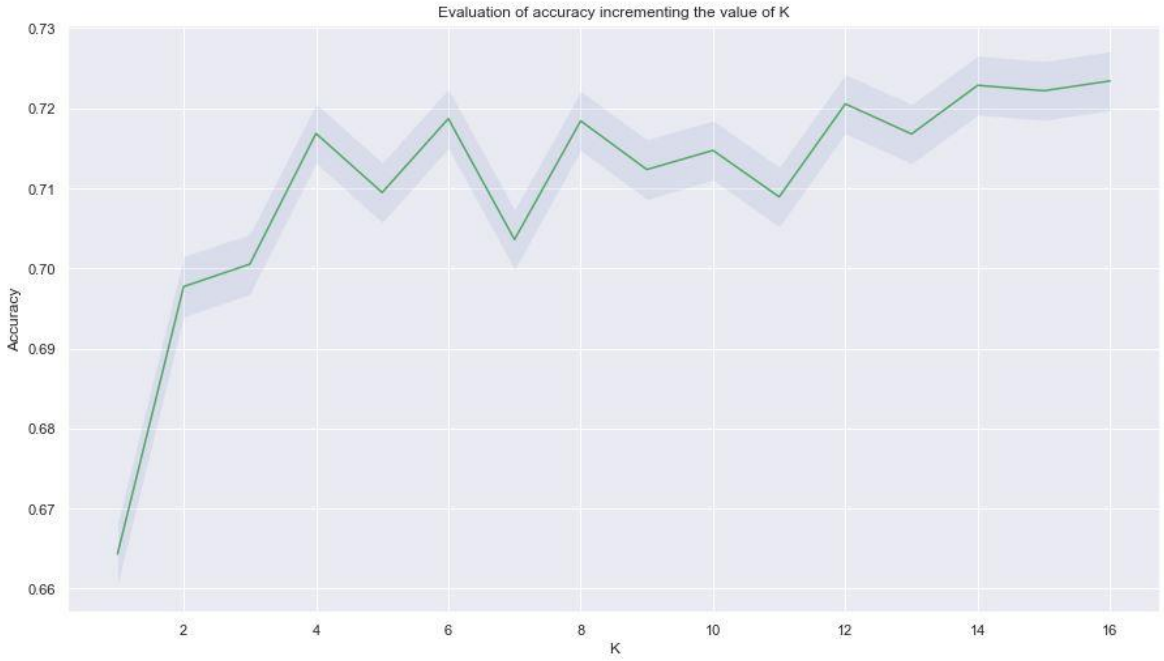

Fig 5 Evaluation of the KNN

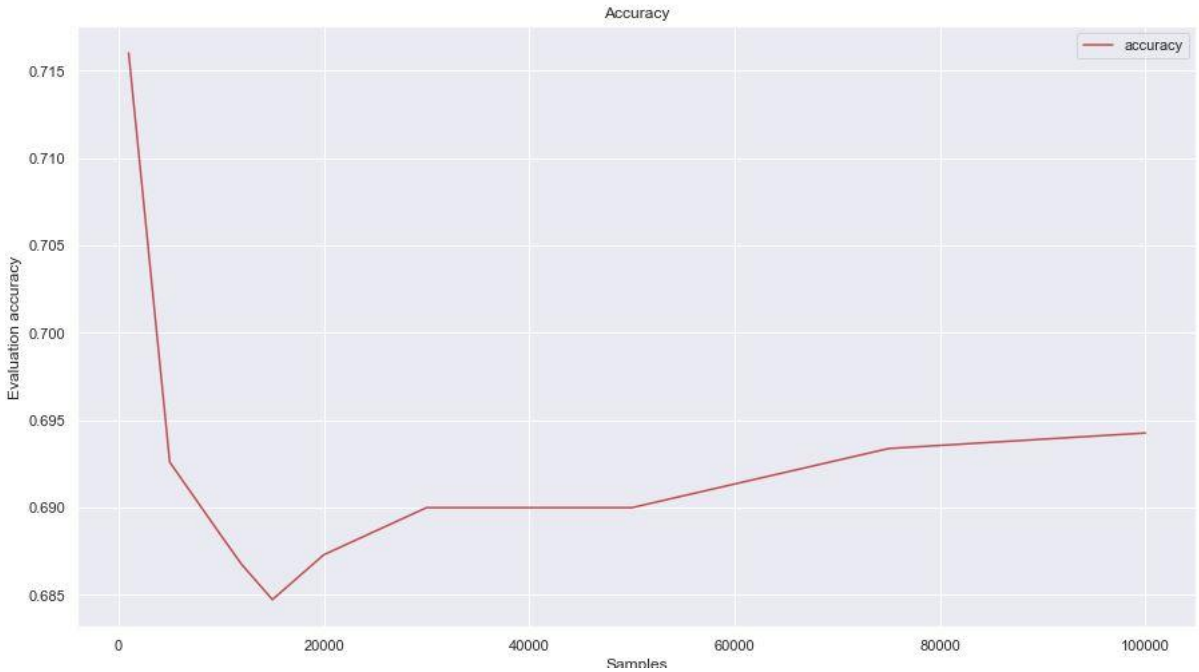

Fig 6 Accuracy of KNN models increasing the value of K

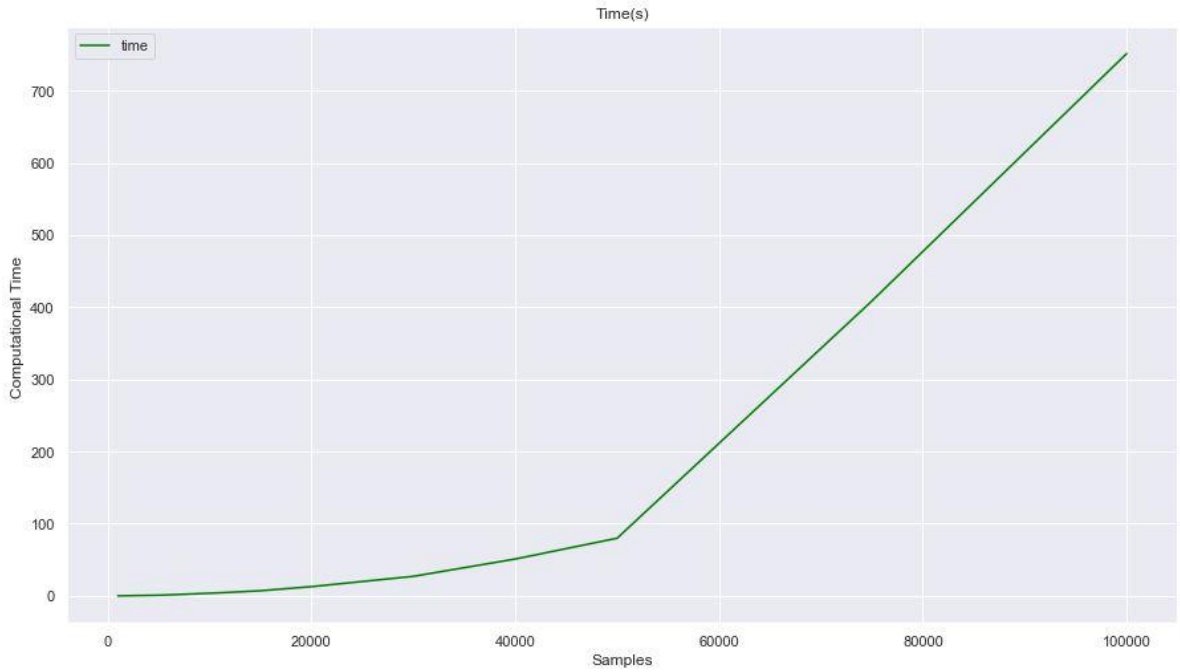

Fig 7 Accuracy of SVM increasing the training sample's size.

5. Results

Algorithm	Jaccard	f1-score	Precision	Recall	Time(s)
Logistic Regression	0.69	0.82	0.69	0.99	0.53
KNN	0.72	0.83	0.74	0.89	19.49
SVM	0.69	0.82	0.69	0.99	393.62

Fig 8 Model Results Summary

The metrics used to compare the accuracy of the models are the Jaccard Score, f1-score, Precision1 and Recall2. This table reports the results of the evaluation of each model. Algorithm Jaccard f1-score Precision Recall Time(s) Logistic 0.69 0.82 0.69 0.99 0.53 KNN 0.72 0.83 0.74 0.89 19.49 SVM 0.69 0.82 0.69 0.99 393.62. In this case, the recall is more important than the precision as a high recall will favour that all required resources will be equipped up to the severity of the accident. The logistic regression and SVM models have similar accuracy, however the computational time from the regression is far better than the other two models. With no doubt the logistic regression is the best model. It has the accuracy of 0.69 and a total execution time of 0.53 seconds.

6. Conclusion

In this study, I analysed the relationship between severity of an accident and some characteristics which describe the situation that involved the accident. Initially I thought that features such as atmospheric conditions, the lighting, would be the most relevant ones, yet I identified the location, the time of the year and time of the accident, the road category and type of collision among 11 the most important features that affect to the gravity of the accident. I built and compared 3 different classification models to predict whether an accident

would have a high or low severity. These models can have multiple applications in real life. For instance, imagine that emergency services have this application with some default features such as date, time and department/municipality and then with the information given by the witness calling to inform on the accident they could predict the severity of the accident before getting there and so alert nearby hospitals and prepare with the necessary equipment and staff. Also by identifying the features that favour the most the gravity of an accident, these could be tackled by improving road conditions or increasing the awareness of the population.

7. Observation

I was able to achieve 74% accuracy in the training of my classification algorithms. However, there was still significant variance that could not be predicted by the models in this study. I think other features like speed or drug influence of the driver could be used to predict a more accurate classification. These are characteristics that may be impossible of knowing right now, but at the incredible peace that technology is evolving nowadays, soon cars will be able to track them so that the emergency services could adopt them. One problem I think these features portrayed was that the target of this classification problem was simplified into two different classes, low and high severity. Labelling severity with a range of punctuation from 0 to 100, for instance, could allow the possibility of developing regression model. The next step on this problem could be to add an accident prediction model able to not just predict the accuracy but also the critical time and spots where potential accidents can occur in advance.