**Capstone Project - Car Accident Severity – Report**

**INTRODUCTION**

**Car Accidents have become a part of our everyday life, and also causes 1.35 million deaths in a year, and 3700 deaths in a single day, with millions more sus-training serious injuries and living with long-term adverse health consequences. Globally, road crashes are a leading cause of death among young people, and the main cause of death among those aged 15-29 years. Road injuries are currently estimated to be the seventh leading cause of death across all age groups globally, and are predicted to become the seventh leading cause of death by 2030**

**Leveraging the tools and all the information nowadays available, an extensive analysis to predict tragic accidents and its severity would make a difference to the death toll. Analysing a sign cant range of factors, including weather conditions, locality, type of road and lighting among others, an accurate prediction of the severity of the accidents can be performed. Thus, trends that commonly lead to severe incidents can help identifying the highly severe accidents. This kind of information could be used by emergency services, to send the exact required staff and equipment to the place of the accident, leaving more resources available for accidents occurring simultaneously. Moreover, this severe accident situation can be warned to nearby hospitals which can have all the equipment ready for a severe intervention in advance, and thus saving countless lives**

**DATA**

#### **We will use France’s accident data from 2005 - 2016, which can be obtained from**[**https://www.kaggle.com/ahmedlahlou/accidents-in-france-from-2005-to-2016**](https://www.kaggle.com/ahmedlahlou/accidents-in-france-from-2005-to-2016)

#### **it consists of 5 data which are as followed**

**1.caracteristics.csv**

**2.holidays.csv**

**3.places.csv**

**4.users.csv**

**5.vehicles.csv**

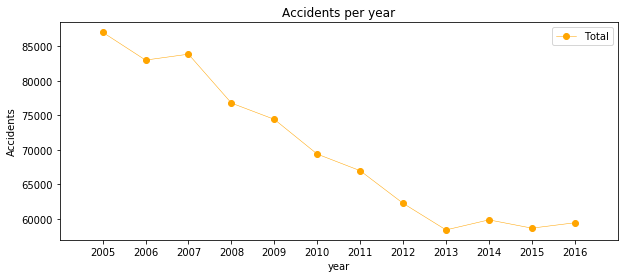
**Then we sorted the data, and found out it has 50% filled with NaN or 0, so** we  **keept with replacing the missing values, then analysis was divided in two groups of features. The rest group had in all features a label which described other cases, for instance the feature describing the atmospheric conditions had a value of 9 for any other atmospheric condition not labelled with the other 8 values. Therefore, the missing values and outliers were replaced with the other cases label for the features of atmospheric conditions, type of collision, road category and the surface conditions. For the second group of features instead, the distribution of their values was analysed. Then two features were dropped, the infrastructures and reserved lanes, as the outliers represented more than 75% of its data. Finally, with the rest of the features with missing values, the number of lanes, the road pro le and shape and the situation at the time of the accident, the NaN and outliers were replaced with the feature’s most popular value. Last format changes were performed to the school and department values. The school feature had all samples divided either in the 0 or the 100 values, thus all the 100 values were replaced with a 1. Similarly, the department feature had an extra 0 added at the units position, so all values were divided by 10. Regarding the type of the data, all features had a coherent data type except for the date feature which was de ned with the string type. I used the to data function of pandas to de ne the date feature with the datetime type. After all, 24 features remained.**

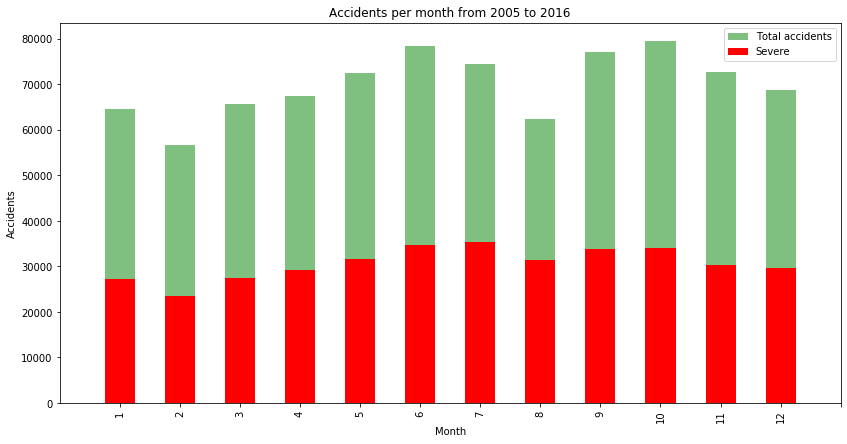
**INTERESTS**

**Stakeholders like the Government and Medic would be highly interested in the prediction as it will help the staff to bring required equipment and thus saving the life and the resources, thus making it more efficient.**

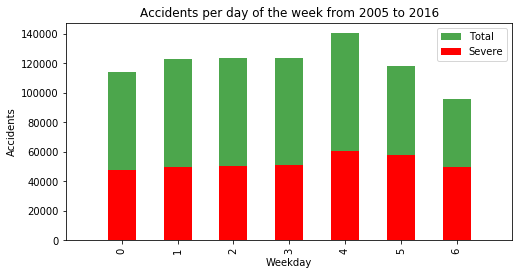
**DATA ANALYSING**

**First the data analysed simply by plotting anccidents by the year and we got this results**

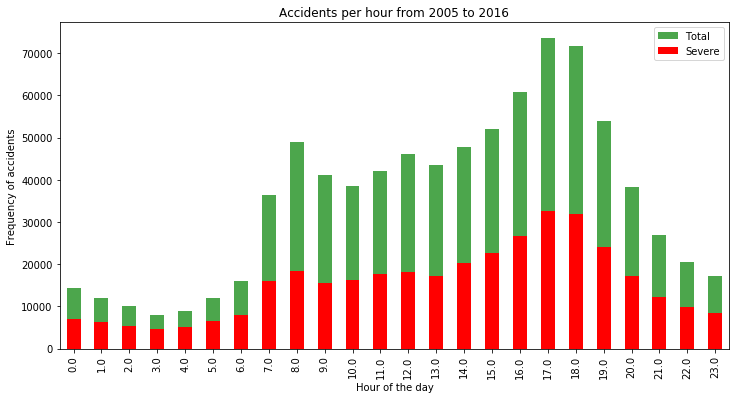
****

**As u can see there is a tread which is declining, now lets see the graph based on the month of the year**

**As you can see June, July, September, October have the most accidents, now let’s go more deeper into the graph and look at the day of these accidents**

****

**Now that its seen, let’s go deeper and look at the time of these accidents**

****

**As you see 5 pm - 6 pm has the most accidents recorded.**

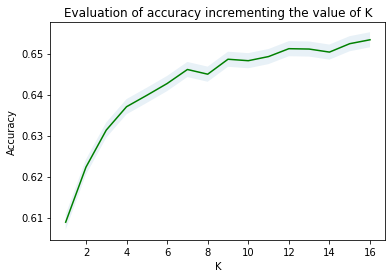
**DATA MODELING (PREDICTIVE MODELING)**

**Now let’s jump into predicting the severity of the accident, in this project we will use 3 different approaches namely:**

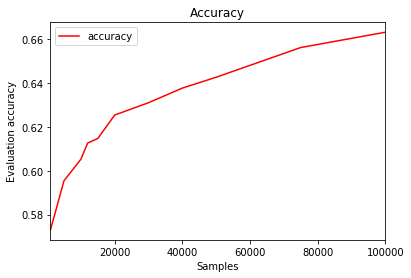
1. **Decision Tree, Random**
2. **Forest Logistic**
3. **Regression K-Nearest Neighbour**
4. **Supervised Vector Machine**

**But before that we have to set the train/test data, so we will split the data by 80/20 Then the data was standardized giving zero mean and unit variance to all features. The decision tree model was upgraded to the random forest. With the default random forest, the features were sorted by impurity-based importance in the prediction of the severity. Thus, the 10 least important features were dropped to decrease the computation complexity for the KNN and SVM models. Keeping with 13 features the accuracy stayed the same and the computational time decreased significant. After evaluating the parameters for each algorithm these were the models. Random Forest: 10 decision trees, maximum depth of 12 features and maximum of 8 features compared for the split. Logistic Regression: c=0.001. KNN: k=16 SVM: size of the training set= 75,000 samples**

**Now lets see the accuracy of KNN values incrementing the value of K:**

****

**Now lets see KNN model’s accuracy while incrementing the value of K:**

****

**RESULTS**

Time taken : 403.9158980846405

Jaccard : 0.6617499122008131

precision recall f1-score support

0 0.66 0.82 0.73 94297

1 0.67 0.46 0.54 73700

accuracy 0.66 167997

macro avg 0.66 0.64 0.64 167997

weighted avg 0.66 0.66 0.65 167997

**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 or being a holiday would be the most relevant ones, yet I identified the department, the day and time of the accident, the road category and type of collision among 11 the most important features that the gravity of the accident. I built and compared 4 different classification models to predict whether an accident would have a high or low severity. These models can have multiple application in real life. For instance, imagine that emergency services have an 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.**

**OBSERVASTION:**

**One problem I think these features had is that the target of this classification problem was simplified to 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.**