**Car Accident Severity**

1. **Introduction**
   1. **Background:** There are many means of transport like buses, trains, planes and by personal vehicles like cars. We often travel short/long distances by cars. We see many accidents and road blocks while travelling and we never know what can happen to us. It’s not always human’s fault in an accident, sometimes it depends on the weather conditions and road conditions also.
   2. **Problem:** Due to different type of light conditions and road conditions there are chances for the accidents to happen in a particular region more. An accident is so sudden that for an ambulance to reach there takes time so to avoid this we can make a model that predicts the accident severity.
   3. **Interest:** We can help travelers to get to know whether to travel on a specific road at a specific time by telling them the severity of the car accident we can use the prior knowledge to predict the same. The other utilization of this analysis is to make ready and aware police department or ambulances about the vulnerable places/roads and weather conditions for road/ car accidents in advance.
   4. **Process:** The prediction of car accident severity can be done by studying the previous car accident data of that route and by making prediction model to get exciting results. We can use machine learning algorithms to predict the car accident severity.
2. **Data:** I have used the data provided by Coursera. It contains 38 columns. We will import the csv file into the panda dataframe.

**2.1) Data Cleaning:**

1. We will delete the unwanted columns from the dataframe.
2. Second we will eliminate the rows with unknown or null values.
3. Then we will select the features/columns to be used in the classification model. We will assign values to the columns like weather, road condition, light conditions.
4. There are about 50k fewer records for severity 2, so to balance the no. of records in the dataset we will take around 56k records for severity 1.
5. We will split the dataset into test and train datasets. Finally, we will fit our training dataset in different classification algorithms like SVM, decision trees, nearest neighbor etc. Then we will use the test data for prediction and evaluation.

**2.2) Feature Selection:**

We will use the columns on which we will predict the severity of the road accident. The columns are: ADDRTYPE, JUNCTIONTYPE, WEATHER, ROADCOND, LIGHTCOND. The column names mentioned can be used as the features which will decide the severity of the accident.

**2.3) Data Labeling:** To use the classification algorithms we will assign numeric values to each column’s values.

**2.4) Data Splitting:** We will split the final data into 80% of train dataset and 20% of test dataset.

1. **Predictive modeling:**

There are two types of models, regression and classification, that can be used. We will use the classification model to predict the severity as we have to classify the values between 1 and 2 severities.

**3.1) Classification Models:**

The application of classification is much more straightforward. I divided the samples into two classes (severity code =1 and severity code =2). The number of records for both the classes were uneven so to balance the number of records for both the classes, I sampled out the records from the class having more records. We can use the binary classification algorithms like logistic regression, k nearest neighbor, support vector, Naïve Bayes or decision tress.

**3.2) Performance of classification models**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **SVM** | **Decision Tree** | **KNN** | **Logistic Regression** | **Gradient Boosting** | **XG Boost** | **Random forest** | **Gaussian NB** |
| **F-1 Score** | 0.5879 | 0.5809 | 0.56625 | 0.5871 | 0.58796 | 0.33703 | **0.59285** | 0.58481 |
| **Accuracy** | 0.590393 | 0.5894 | 0.56783 | 0.5888 | 0.59084 | 0.501726 | **0.592949** | 0.58949 |

**F-1 score:** In [statistical](https://en.wikipedia.org/wiki/Statistics) analysis of [binary classification](https://en.wikipedia.org/wiki/Binary_classification), the **F1 score** (also **F-score** or **F-measure**) is a measure of a test's accuracy. The highest possible value of **F**1 is 1, indicating perfect precision and recall, and the lowest possible value is 0, if either the precision or the recall is zero.

**Accuracy:** It is the number of correctly predicted data points out of all the data points. More formally, it is defined as the number of true positives and true negatives divided by the number of true positives, true negatives, false positives, and false negatives.

Clearly, we can see that the highest value of the F-1 score and accuracy is gained by the model Random Forest classifier. We can use the severity of the accident by using the machine learning model: Random Forest

1. **Conclusion:**

In this study, I analyzed the data and extract the features that can predict the severity of the accident according the external factors the type of road, road conditions, weather conditions etc. The Random Forest classifier turned out to me most accurate model from all. These models can be very useful in helping police department, hospital management and travelers to take extra care in the regions where the accident severity is more than the other locations according to the present weather and road conditions.

1. **Future directions:**

I was able to acquire the 59% accuracy from the provided data. It only provides the severity between 1 and 2. However, if the data has more values for severity codes of higher level (severity=4), then we can train model accordingly so that model can predict more precisely between the severity 1,2,3 and 4(where 4 being the most severe as fetal). More data would help improve model performance significantly.

We did not consider the vehicle type and condition of the vehicle as the feature. Two other features we can take in consideration are: day of the week and the time of the day. As traffic is more heavy during the office hours and on the weekends.

So to prepare a more accurate model for prediction of severity of road accidents we have to take the above points into consideration.