Data Science Capstone Project

**Car Accident Severity Report**

***Introduction:***

The main objective of this project is to make drivers aware of the chances of a collision to reduce the number of collisions. A machine learning model which can predict the severity of the accident will be developed based on the considered dataset. This developed model will help the drivers to know the risk and be careful when the conditions are not favorable for driving.

***Data Understanding:***

In this project I am going to deal with a data set called ‘Data-Collisions’ which is take from

<https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv>

This data consists of ‘Car Collisions’ details for different conditions of weather, road, light conditions etc. in a community.

In this data our predictor is ‘SEVERITYCODE’ which is used to measure the severity of an accident from 0 to 4.

0 – Almost no Probability of accident

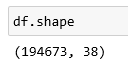
1 – Very low Probability of accident with a chance of property damage

2 – Low Probability of accident with a chance of minor injuries

3 – Mild Probability of accident with a chance of serious injuries

4 – High Probability of accident with a chance of fatality

The data we are going to use in this project is a huge dataset.



***Data Preperation:***

The data must be analysed properly and the missing values, misleading values must be filled. Feature Engineering must be carried out on the data to identify the proper variables which affect the SEVERITY.

As the Severity i.e. the target variable is a categorical variable the best way is to find the features which are highly correlated to target variable. As data is very large considerable data can be used as test data.

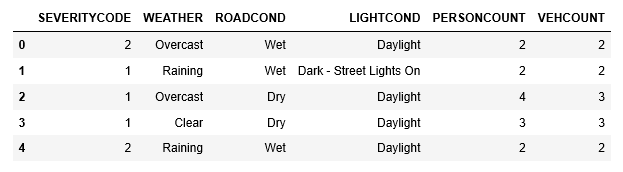
***Extracting Dataset and Balancing the Data***

We can clearly see from the below figure that there is a lot of data that is NaN and we can see some features does not have 95% of data. To make our data structured it is better to get rid of these kinds of features.

A picture containing building

Description automatically generated

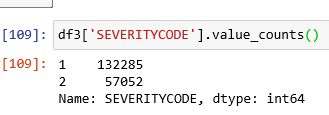
The correlation of the target variable with many independent variables in very low so removing the features which are not correlated with the target variable gives us the below dataset which is free of unwanted data.



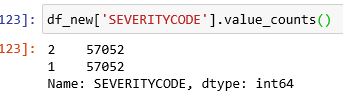
We also have some missing data in categorical features and we have eliminated these data as adding an approximate data here might cause a lot of variations in the data.

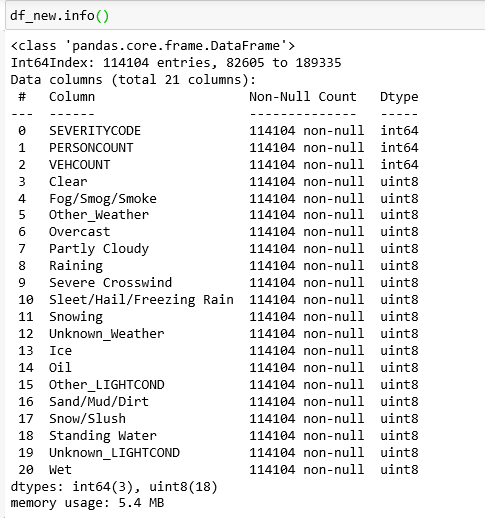
There are even categorical data present in our dataset, so we can use get\_dummies to convert them.

If we observe the target variable SEVERITYCODE it is not balanced properly.



Using downsampling the majority class we can deal this.





Now our data is balanced and is ready to be fed into Machine Learning model.