Data Science Capstone Project

**Car Accident Severity Report**

**Abhilash**

**September 20, 2020**

1. ***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.

1. ***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

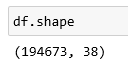
1 – Very low Probability of accident with a chance of property damage

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.



1. ***Data Preparation:***

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.

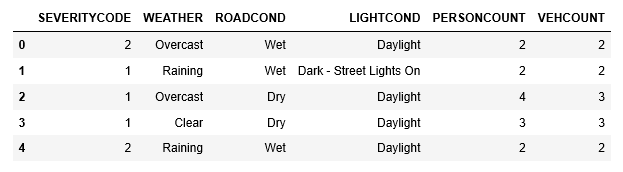
*Extracting Dataset and Balancing the Data*

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

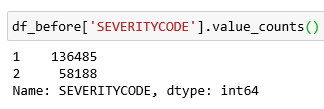
The correlation of the target variable with many independent variables in very low so removing the features which are not correlated with the target variable gives us the below dataset which is free of unwanted data.



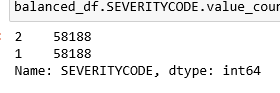
We also have some missing data in categorical features and we have eliminated these data as adding an approximate data here might cause a lot of variations in the data.

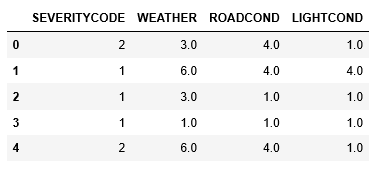
There are even categorical data present in our dataset, so we can use get\_dummies to convert them.

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



Using downsampling the majority class we can deal this.



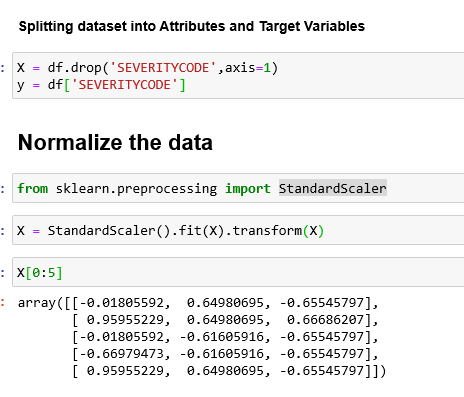


Now our data is balanced and is ready to be fed into Machine Learning model.

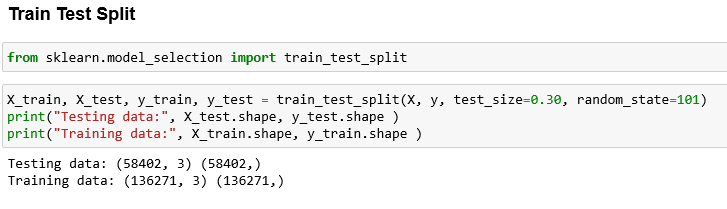
1. ***Methodology:***

As our data is has a categorical target value It comes under a classification problem and we will choose K-Nearest Neighbour(k-NN), Decision Tree and Logistic Regression Model.

Let us split our dataset into features and target variables and normalize the features value.

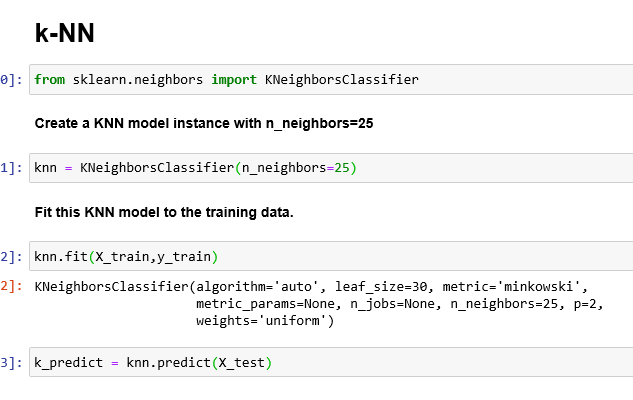


Now split the data set into testing data and training data.



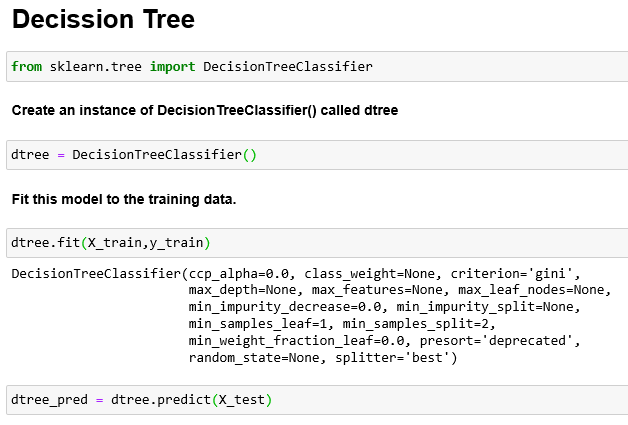
*4.1 k-NN*

The k-nearest neighbours algorithm (k-NN) is a non-parametric method proposed by Thomas Cover used for classification and regression. the input consists of the k closest training examples in the feature space and output is a class membership. An object is classified by a plurality vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours.



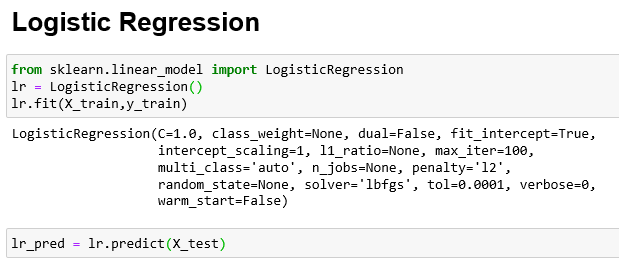
*4.2 Decision Tree:*

Decision tree learning is one of the predictive modelling approaches used in statistics, data mining and machine learning. It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. Decision trees are among the most popular machine learning algorithms given their intelligibility and simplicity.



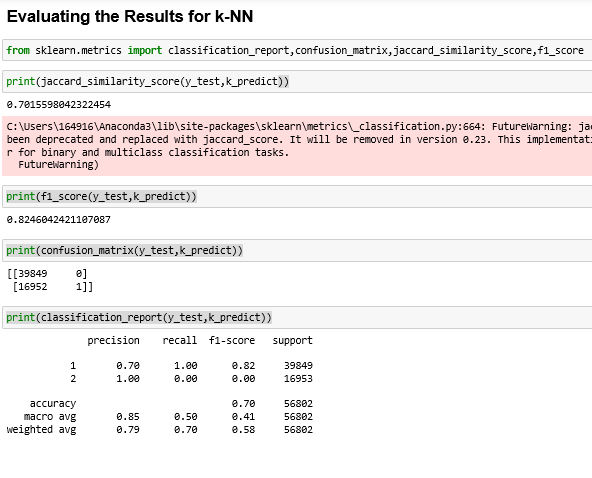
*4.3 Logistic Regression:*

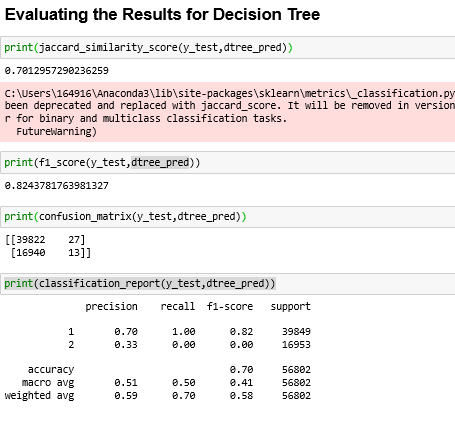
the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1, with a sum of one.

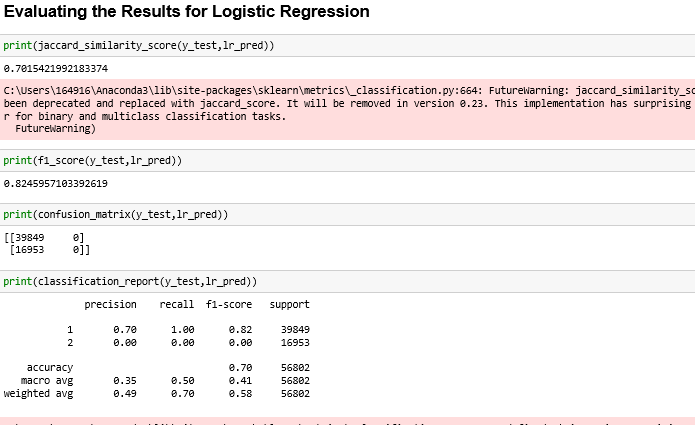


1. ***Results and Evaluation:***

Using all the 3 methods we have got almost same level of prediction. Below images show the f1score, jaccard\_similarity\_score, confusion\_matrix and classification\_report of the 3 models.







1. ***Discussion:***

In the beginning of this notebook, we had categorical data that was of type 'object'. This is not a data type that we could have fed through an algorithm, so label encoding was used to created new classes that were of type int8; a numerical data type.

After solving that issue we were presented with another - imbalanced data. As mentioned earlier, class 1 was nearly three times larger than class 2. The solution to this was downsampling the majority class with sklearn's resample tool. We downsampled to match the minority class exactly with 58188 values each.

Once we analyzed and cleaned the data, it was then fed through three ML models; K-Nearest Neighbor, Decision Tree and Logistic Regression. Although the first two are ideal for this project, logistic regression made most sense because of its binary nature.

Evaluation metrics used to test the accuracy of our models were jaccard index, f-1 score and logloss for logistic regression. Choosing different k, max depth and hyparameter C values helped to improve our accuracy to be the best possible.

1. ***Conclusion:***

Based on historical data from weather conditions pointing to certain classes, we can conclude that particular weather conditions have a somewhat impact on whether or not travel could result in property damage (class 1) or injury (class 2)