**Coursera Capstone Report**

**Car Accident Severity- Seattle, USA**

**By: George Saliba**

**Date: 09/09/2020**

**Introduction**

According to online sources, the USA experiences an average of approximately 102 vehicle crashes per day. From approximately 6 million crashes a year, over 35,000 crashes are fatal. To improve the country's road safety, it is important to understand the causes of accidents as well as the conditions leading to the incident.

Moreover, understanding the factors prior to the incidents through graphical and statistical analysis, including the level of severity, likelihood of collisions, and probability of injury; can further improve the education sector within road safety as well as give governmental bodies the data needed to build safer roads and raise awareness of the main factors and conditions leading to road incidents.

The aim of this project is to use publicly available data- published by the Seattle Deport of Transportation (SDOT) in the form of a dataset, to identify the relevant information that allow for the comprehension on the number, injuries, and severity of collision occur in Seattle.

The data will be analyzed mathematically and graphically through data science methodologies such as data visualization tools and machine learning models in order to provide insight on traffic collisions in Seattle.

**Data**

The data used throughout this project is extracted from the SDOT's website entitled "Collisions-All Years". This dataset includes all types of collisions, whether be vehicle, bicycle or pedestrian, filed by the Seattle Police Department.

The dataset provided includes 65,000 collisions from 2004 to present day, updated on a weekly basis. The severity is also an important factor that is incorporated within the collision's dataset with 5 codes, explained:

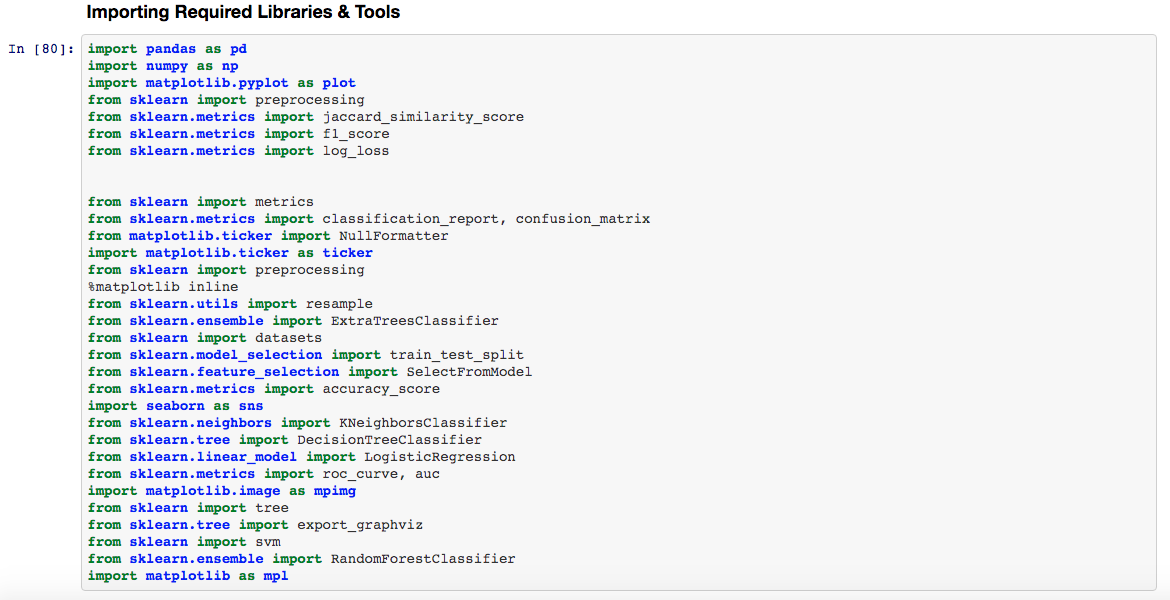
* 0: unknown
* 1: property damage
* 2: injury
* 2b: serious injury
* 3: fatality

The actual dataset only includes codes 1 and 2 without the inclusion of any other alternatives; but gives 37 attributes related to the collisions reported such as day, time, month, weather etc. All of which will be used to build our model and explain trends related to traffic collisions.

The full description of the data can be found at: <https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf>

**Methodology**

The first important step taken was to import the needed libraries into my Jupyter Notebook. The screenshot below represents the above:



Next, the dataset, as a .csv file was imported into the Notebook and the codes to generate the following were implemented:

1)

*# List of Columns* df\_data\_1.columns

2)

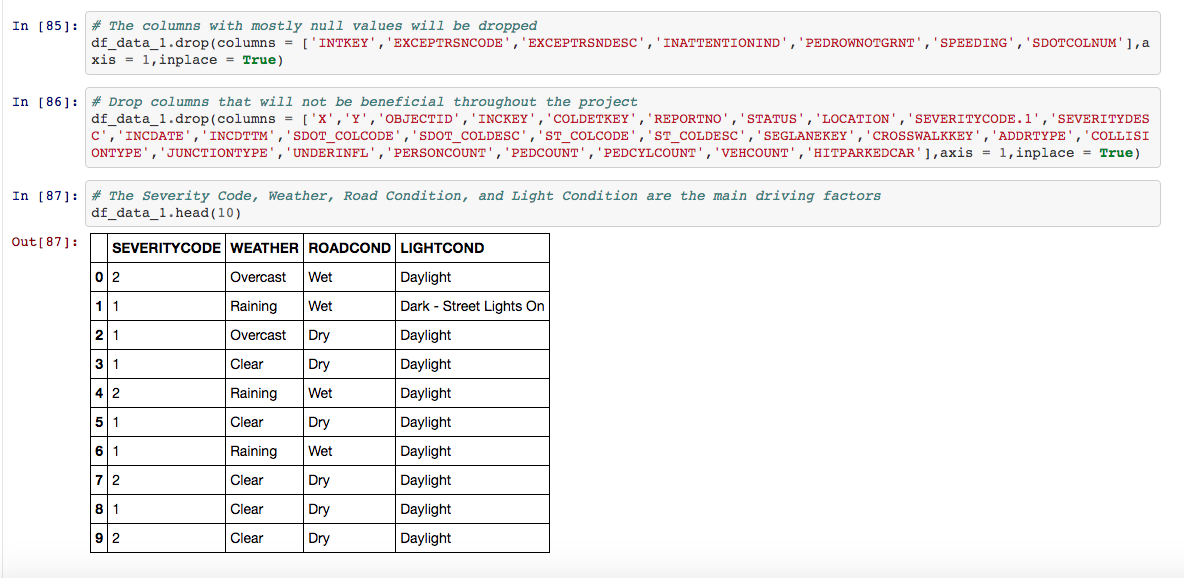
*# Check Data that has missing values*df\_data\_1.info()

3)

*# Check size of Data* df\_data\_1.isnull().sum()

Then, all the columns with null values where dropped as well as the columns that will not be beneficial throughout the report.

This makes the remaining factors: Severity Code, Weather Conditions, Road Conditions, and Light Conditions.



The NaN values were also adjusted within the Severity Code, Weather Conditions, Road Conditions, and Light Conditions columns using the following code:

*#We need to adjust each of the 4 factors by replacing NaN values with unknown values#Weather*df\_data\_1['WEATHER'].replace(np.NaN, "Unknown", inplace=**True**)

*#Road Condition* df\_data\_1['ROADCOND'].replace(np.NaN, "Unknown", inplace=**True**)

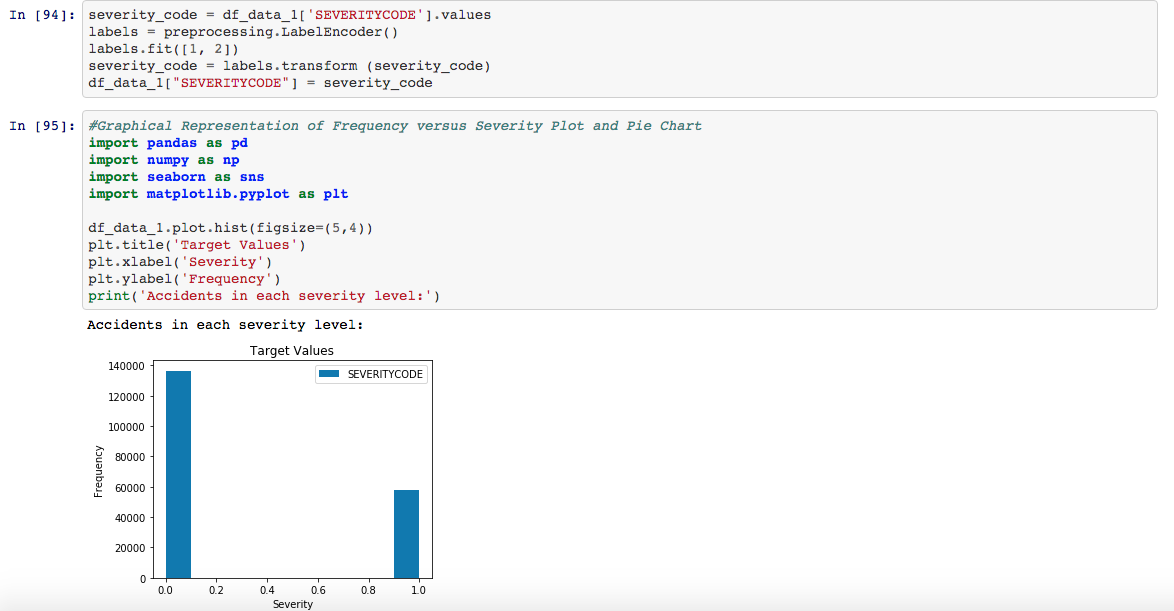
*#Light Condition*df\_data\_1['LIGHTCOND'].replace(np.NaN, "Unknown", inplace=**True**)

After the data was checked and successfully implemented, the ***Exploratory Data Analysis*** phase began.

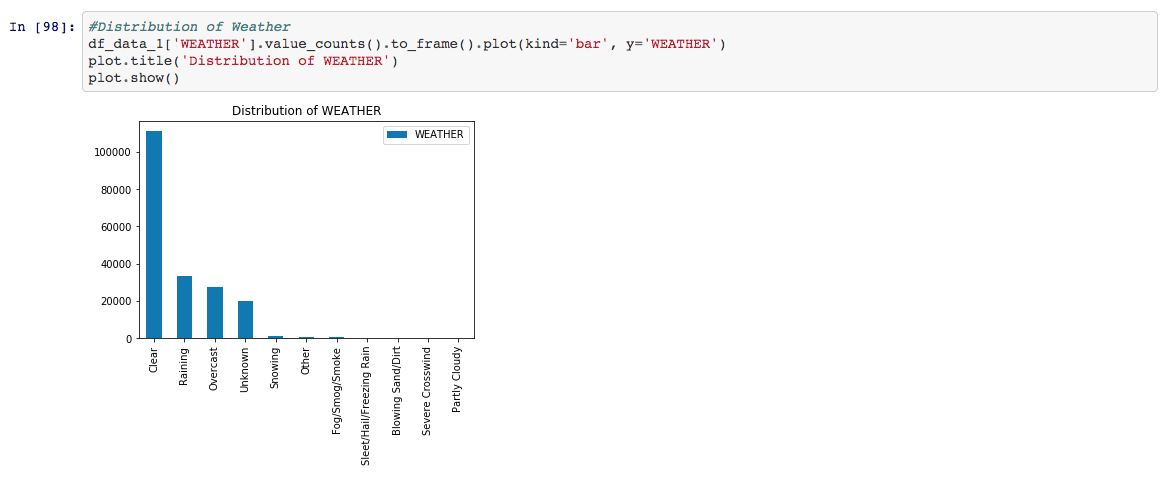
This section of the Notebook was dedicated towards understanding the relationship between the Severity accidents and the other factors (Weather, Road and Light Conditions).

The following screenshots represent the codes used to execute the needed graphs.

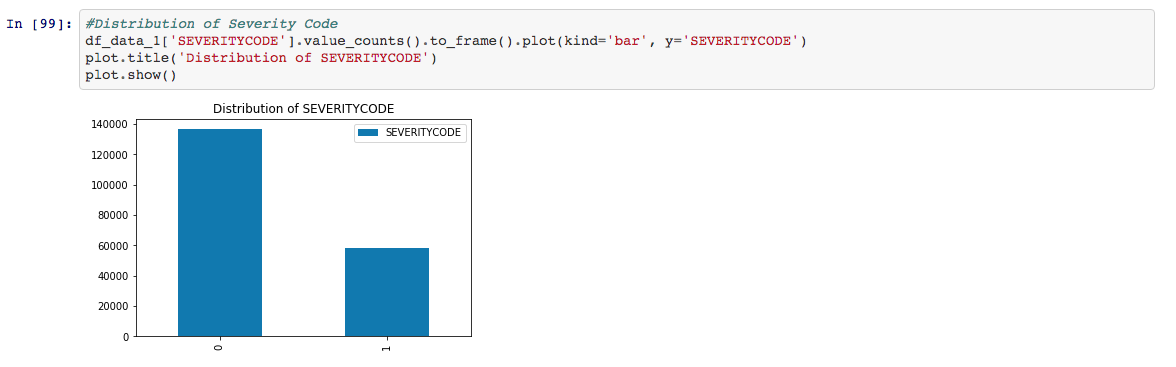
**FREQUENCY vs. SEVERITY PLOT**



**DISTRIBUTION OF WEATHER**



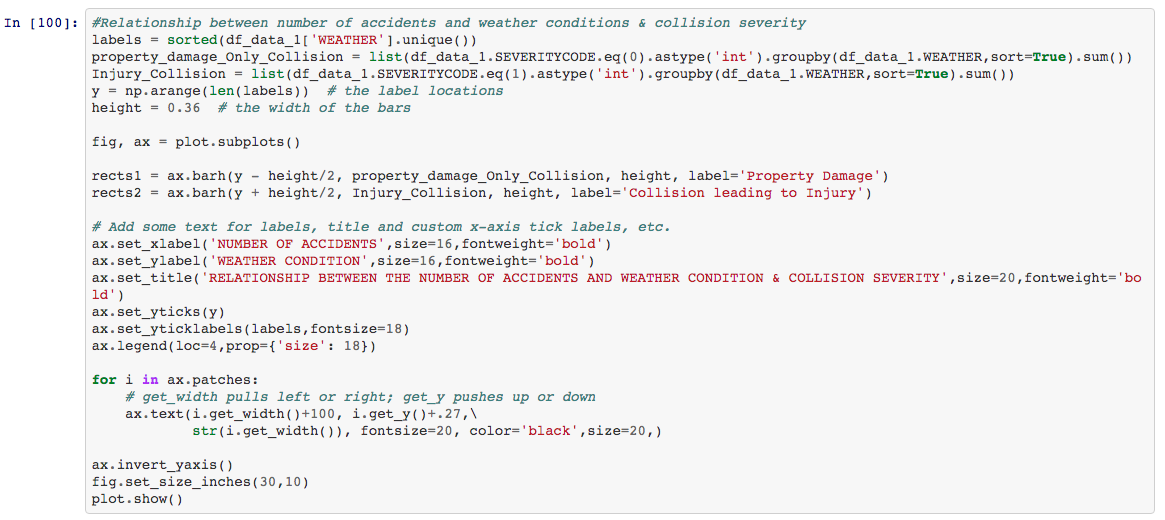
**DISTRIBUTION OF SEVERITY CODE**



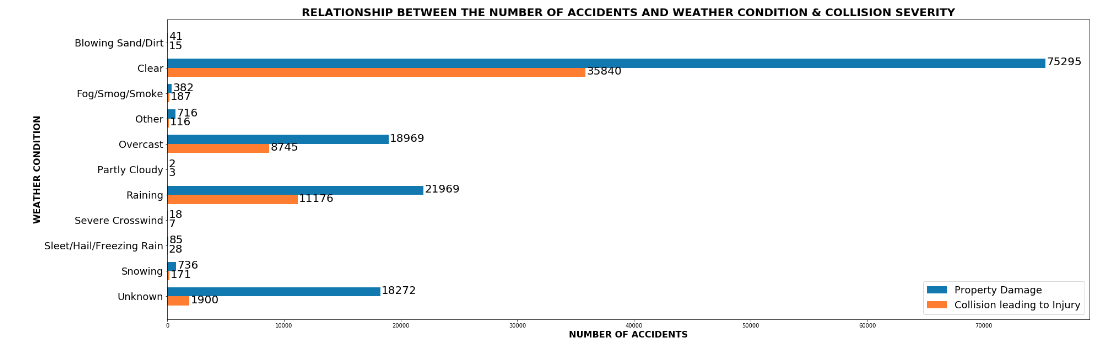
After generating these graphs, 2 main relationships have also been determined:

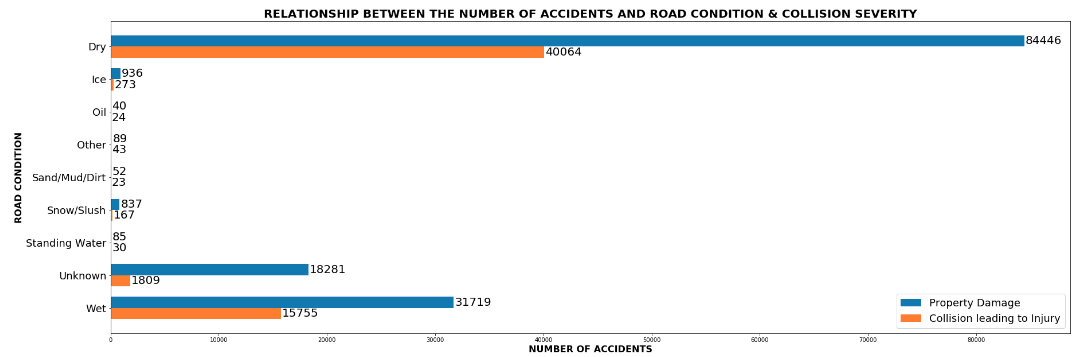
* The relationship between number of accidents and weather conditions & collision severity
* The relationship between number of accidents and road conditions & collision severity

The same code was used for both, but the terms ‘WEATHER’ and ‘WEATHERCOND’ have been replaced with ‘ROAD’ and ‘ROADCOND’.



The relationship between number of accidents and weather conditions & collision severity



The relationship between number of accidents and road conditions & collision severity

After generating these relational graphs, the ***Data Preparation*** phase was started and the following was achieved:

* Encoding and replacing each of the elements’ subsections into values for grouping purposes
* The data frame conditions have been set (X)
* The SEVERITYCODE has been set on the axis (Y)
* Preprocessing and splitting was completed using the following codes
  + Preprocessing
    - **from** **sklearn** **import** preprocessing
    - X= preprocessing.StandardScaler().fit(X).transform(X)
    - X[0:5]
  + Splitting
    - **from** **sklearn.model\_selection** **import** train\_test\_split
    - X\_train, X\_test, y\_train, y\_test = train\_test\_split( X, Y, test\_size=0.2, random\_state=4)

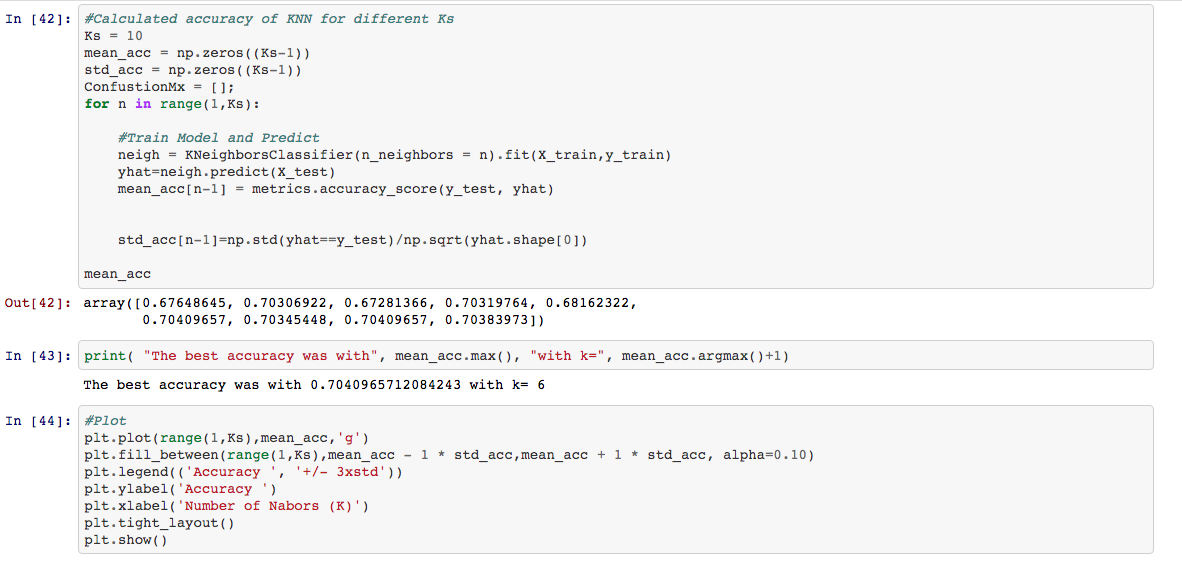
**Results**

The results section is catered towards the models formed through Machine Learning techniques; K-Nearest Neighbor, Logistical Regression, and Decision Tree models.

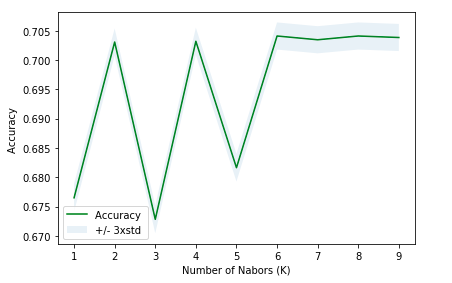
K-Nearest Neighbor

*The following lines of code were used:*

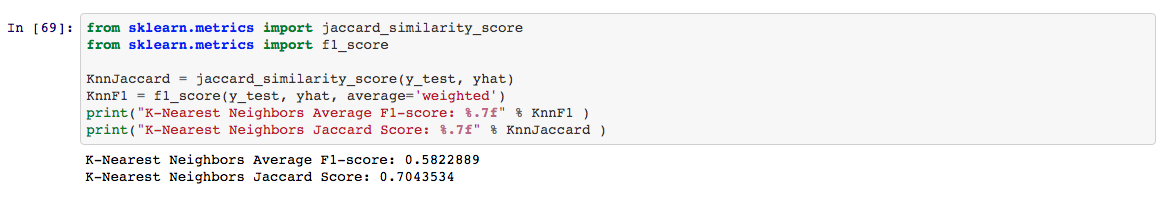




*The Model*



*Average F-1 Score and Jaccard Score*



Logistical Regression

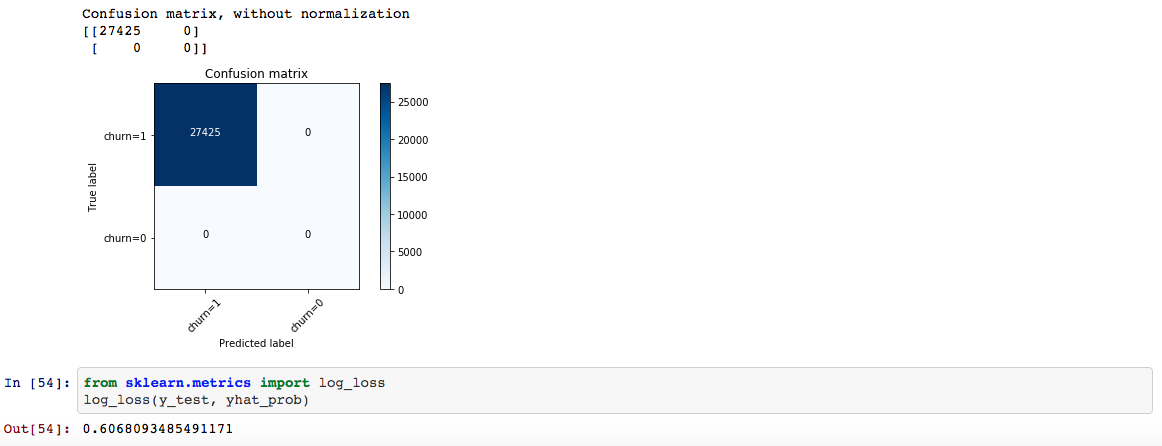
*The following lines of code were used:*



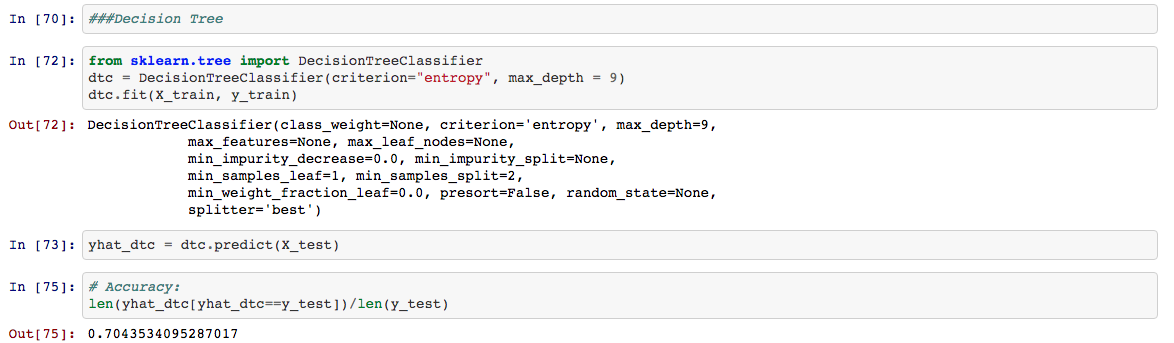




*The Model & Log-Loss*



Decision Tree Model & Accuracy



**Discussion & Conclusion**

Based on all the results generated, the Decision tree model displayed an accuracy of 0.70435, while the KNN gave off an Average F-1 score of 0.58 and a Jaccard score of 0.70435 and logistical regression also gave the same Jaccard score, similar to the Decision tree’s accuracy and that of the KNN.

The graphs showed an important relationship between the number of accidents and road conditions and severity of collision. On ‘DRY’ road conditions, there was an estimate of 84,446 collisions that lead to property damage and 40,064 that lead to injuries; the ‘DRY’ road served the highest number of accidents. Taking a look at the weather conditions, ‘CLEAR’ weather resulted in the highest number of accidents with 75,295 leading to property damage and 35,840 leading to injuries.

Based on all the collisions that took place, 70.1% of them lead to property damage while 29.9% lead to injuries.

To conclude, the results give beneficial insights on the severity of vehicle collisions with focus on the different factors such as weather, road and light conditions. In terms of modeled results and values, they can be improved with more accurate and precise input data.