**IBM Data Science Capstone: Car Accident Severity Report**

**Introduction/Business Problem**

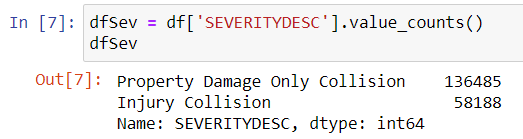
Are you a driver? Do you know how to drive? If yes, I ensure face this problem before such as you come across a terrible traffic jam on the other side of the highway. Long lines of cars barely moving. As you keep driving, police car start appearing from afar shutting down the highway. It is an accident and there's a helicopter transporting the ones involved in the crash to the nearest hospital. They must be in critical condition for all this to be happening.

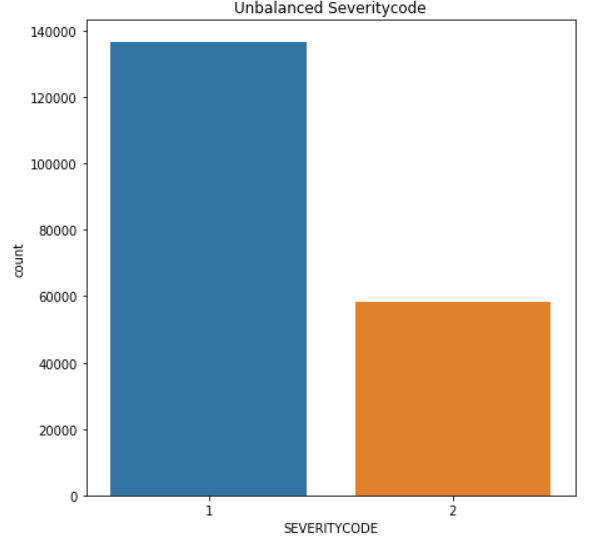
So I am going to predict the severity of an accident based on weather, road conditions and light conditions. It can warn you before by given the weather and the road conditions about the possibility of you getting into a car accident and how severe it would be, so that you would drive more carefully or even change your travel if you are able to.

**Data Section**

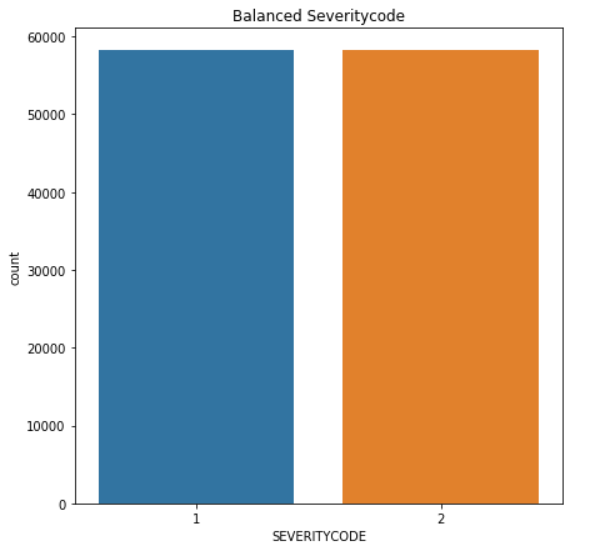
I will use the data that provided by the course to predict the severity of an accident based on weather, road conditions and light conditions. I will build a machine learning model to predict the possibility of you getting into a car accident and how severe it would be. For example, if the weather is overcast, the road conditions is wet and the light condition is daylight how many probability will you getting into a car accident and it is severity being either "Property damage only" or "Injury Collision"? It can be found by using the machine learning model to predict it.

In target variable SEVERITYCODE is unbalanced, so I need to balance the data to get a better machine learning model. As you can see in the picture below, more than twice the amount of 1 severity code than 2.

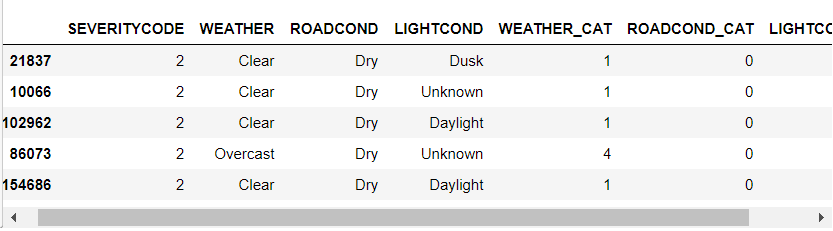




After balanced severitycode is shown as below:



For this model, there are many columns that I will not use for this model. I only use severitycode, whether, roadcon and lightcond. But, most of the features are type of object, when they should be numerical type. So I use label encoding to convert the features to our desired data type.

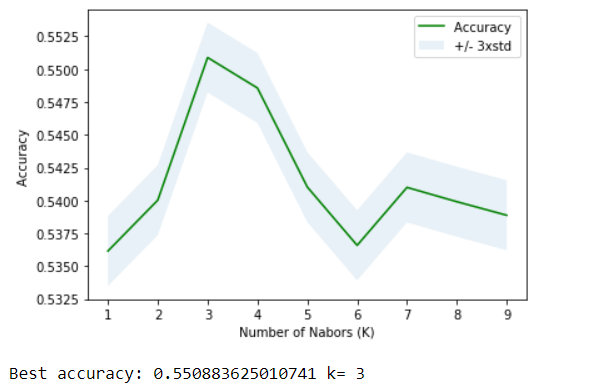


**Methodology**

Our data is now ready to be use for machine learning models. I will use the following models to predict:

**K-Nearest Neighbor (KNN)**

KNN classifying severity code by the output of their similarity which are near to each other. I will run K-Means into 3 cluster because when I analyse the K-Mean with accuracy classification score function it ensured me the 3 is the optimum k of K-Means.



**Decision Tree**

Decision tree will choose an attribute from the dataset and calculate the significance of attribute in splitting of data and split data based on the value of the best attribute. It will show which attribute is the purer node.

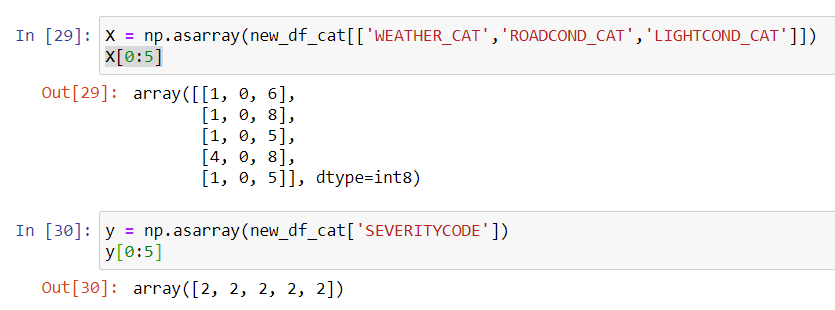
**Support Vector Machine (SVM)**

SVM mapping data to a high-dimensional feature space and finding a separator.

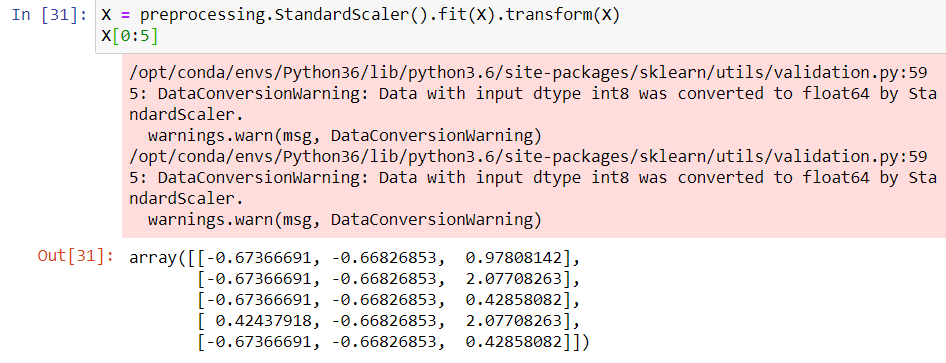
**Logistic Regression**

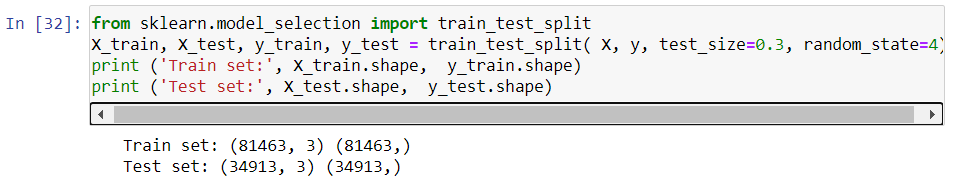
Logistic Regression use to predict the outcome of binary severity code based on whether, roadcon and lightcond.

First, I define X and Y.



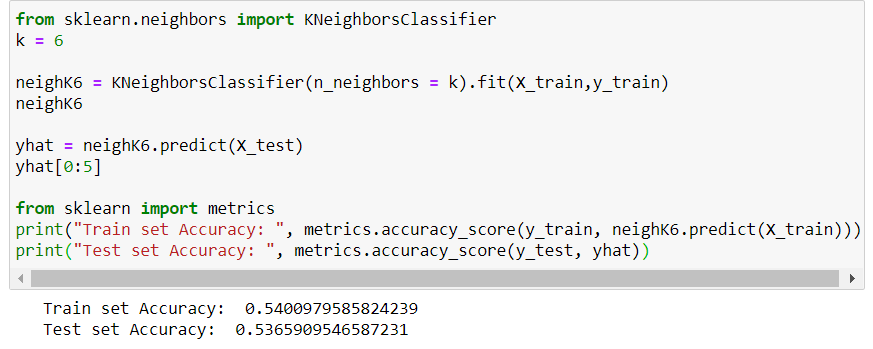
Second, I normalize the dataset.



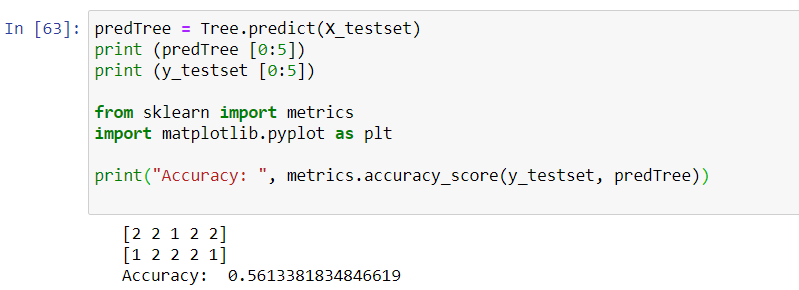
Then, I will use 30% of data for testing and 70% for training. 

Lastly will be all model and predictions:

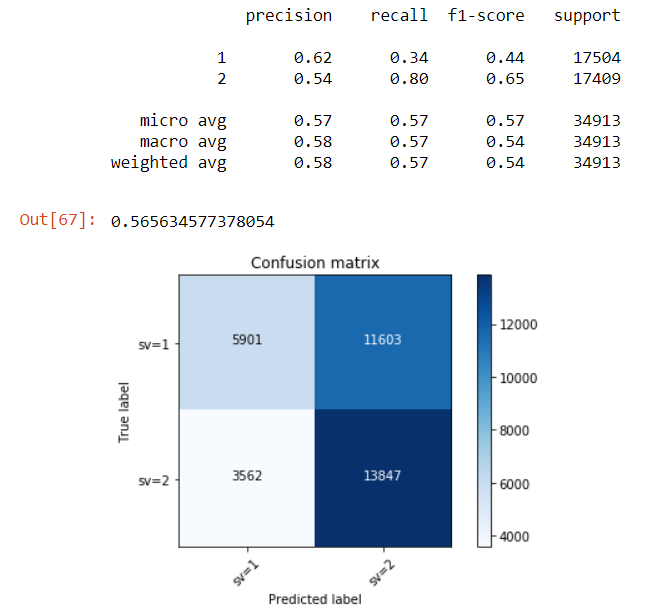
**K-Nearest Neighbor (KNN)**



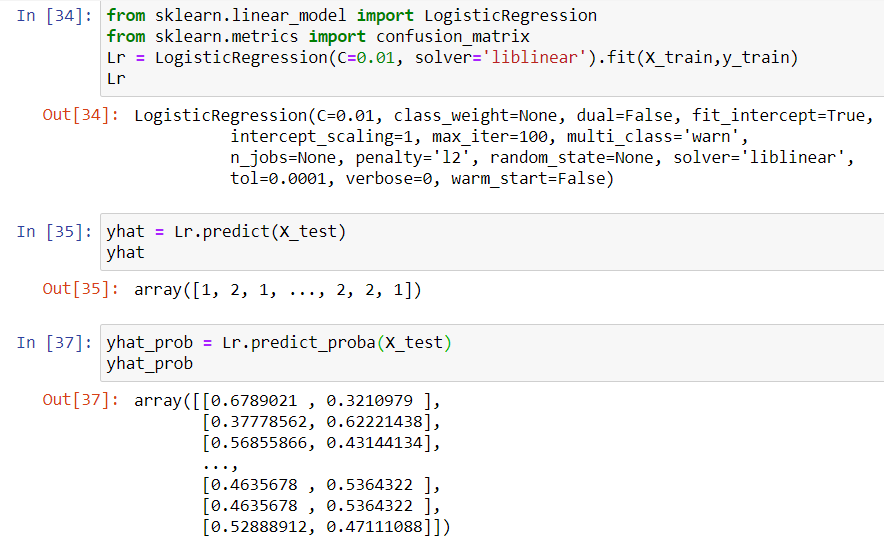
**Decision Tree**



**Support Vector Machine (SVM)**

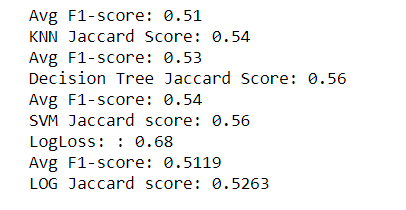


**Logistic Regression**



**Result**

After building all model, it is important to test and evaluate their performance by their accuracy of predict. I use F1-score, Jaccard Score and Log Loss to evaluate all model. F1-scare is the average of precision and recall for a model. From 0 to 1, 1 is the higher accuracy. Jaccard Score which looks at the difference between the predicted labels and predicted labels. From 0 to 1, 1 is the higher accuracy. Log loss is performance of a classifier where the predicted output is a probability value between 0 and 1. 0 is the higher accuracy.



**Discussion**

From the result section, it shown SVM is the most accuracy model to predict this case since it has the higher accuracy of Jaccard Score and F1-score compare to others.

Even though, SVM is the most accuracy model but the result of all model is almost the same. All of them are between 0.50 - 0.56 .

**Conclusion**

Since the result of all model is almost the same. Its means weather, road conditions and light conditions will cause the injury collision around 50% and property damage only collision 50%. In conclusion, please pay attention when you are driving because no matter how is the weather, road condition and light conditions you still have chance to get into an accident.