**Introduction/Business Problem**

In order to reduce the frequency of car collisions in a location , an algorithm must be developed to predict the severity of an accident based on the current weather conditions, visibility conditions etc.

In most of the scenarios, Accidents occurs due to irresponsible drivers who does not pay attention while driving or drive above speed limit or drunk and drive etc. Apart from the above-mentioned scenarios, weather, visibility and road conditions related accidents are the uncontrollable factors. So these factors can be predicted bases on a particular model/ algorithm.

This model will alert drivers to remind them to be more careful based on the current weather conditions, visibility conditions.

**Data Section**

The data consists of 37 independent variables and 194,673 rows. The dependent variable, “SEVERITYCODE”, contains numbers that correspond to different levels of severity caused by an accident from 0 to 4.

Severity codes are as follows:

0: Little to no Probability (Clear Conditions)

1: Very Low Probability — Chance or Property Damage

2: Low Probability — Chance of Injury

3: Mild Probability — Chance of Serious Injury

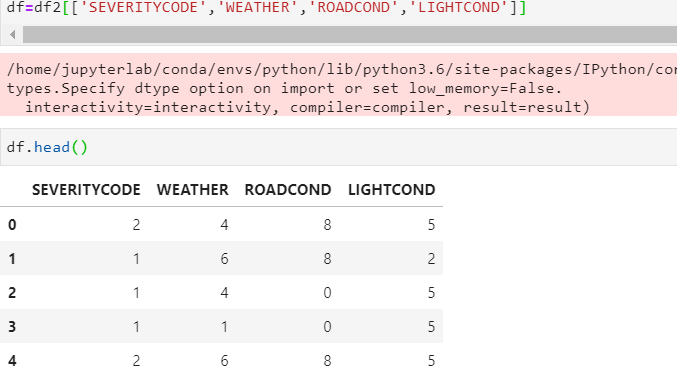
4: High Probability — Chance of Fatality

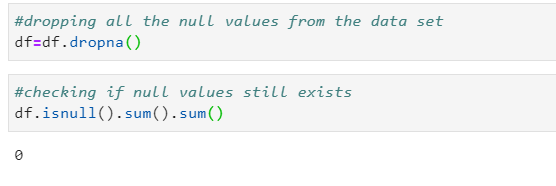
Furthermore, because of the existence of null values in some records, the data needs to be pre-processed before any further processing.

The dataset in the original form is not ready for data analysis. In order to prepare the data, first, we need to drop the non-relevant columns. In addition, most of the features are of object data types that need to be converted into numerical data types.

After analysing the data set, I have decided to focus on only four features, severity, weather conditions, road conditions, and light conditions, among others.

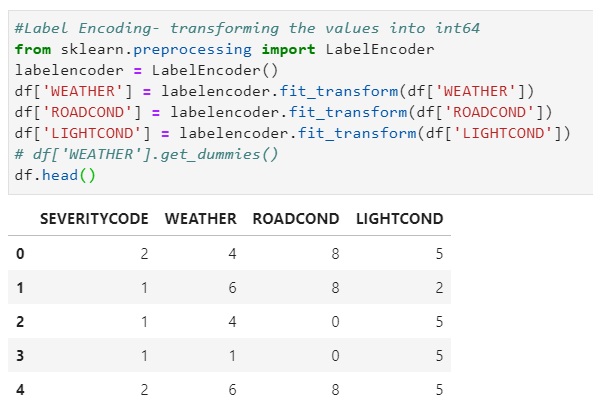
1) we need to remove non-relevant columns and Null fields:



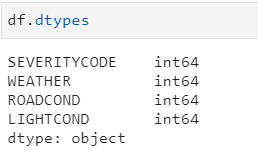


2)We must use label encoding to covert the features to our desired data type.

Label Encoding:



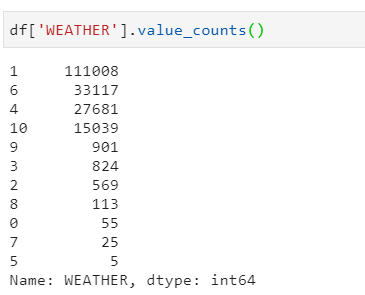
we do label encoding because some of the ML models does not supports object data type or string data data type. so we need to do label encoding to convert the features to similar data types so that we can use the data set in any of the machine learning models.

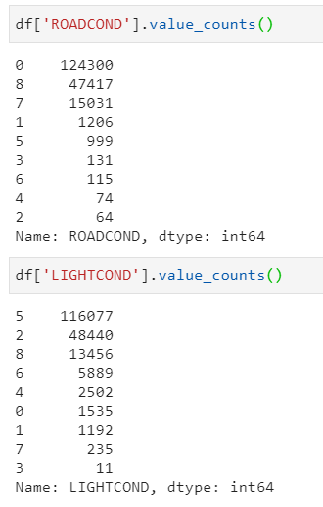


I have selected the following features to predict the severity of accidents. Among all the features, the following features have the most influence in the accuracy of the predictions:

WEATHER, ROADCOND, LIGHTCOND

I have used Value\_counts() for all the 3 to see the breakdowns of accidents occurring during the different conditions.





After balancing “SEVERITYCODE” feature, and standardizing the input , the data has been ready for building machine learning models.

# I have used 3 Machine learning models:

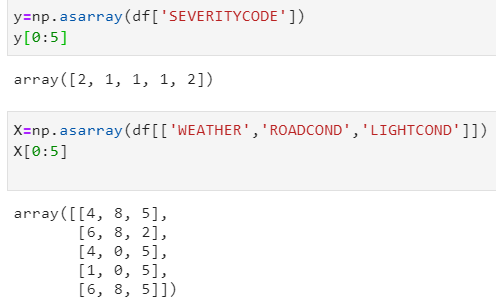
·        K Nearest Neighbour (KNN): KNN will help us predict the severity code of an outcome by finding the most similar to data point within k distance.

·        Decision Tree: A decision tree model gives us a layout of all possible outcomes so we can fully analyze the consequences of a decision. It context, the decision tree observes all possible outcomes of different weather conditions.

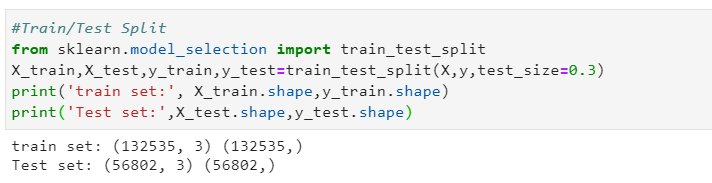
·        Linear Regression: our data set only provides us with two severity code outcomes, our model will only predict one of those two classes. This makes our data binary, which is perfect to use with logistic regression.

**#Initialization**

DEFINING X & Y:



# Train/Test Split



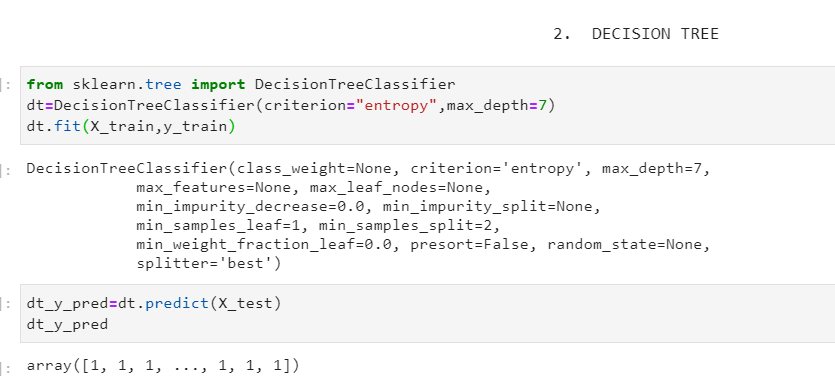
**Methodology:**

**APPLYING THE MACHINE LEARNING TECHNIQUES:**

**1. KNN (K-Nearest Neighbors)**



**2. Decision Tree**

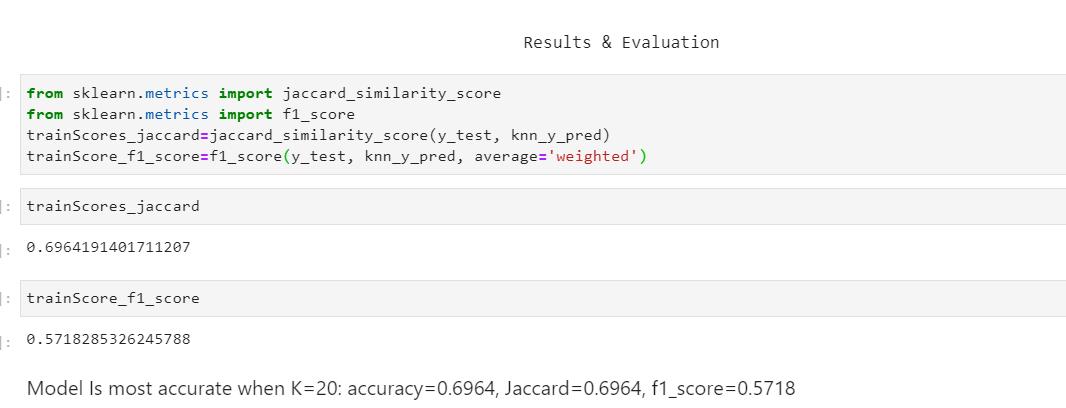


**3.Logistic regression**

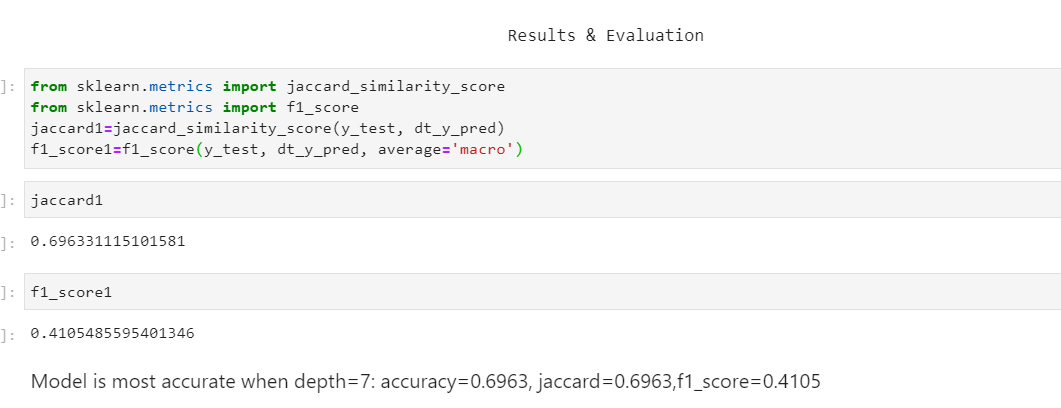


**Results & Evaluation:**

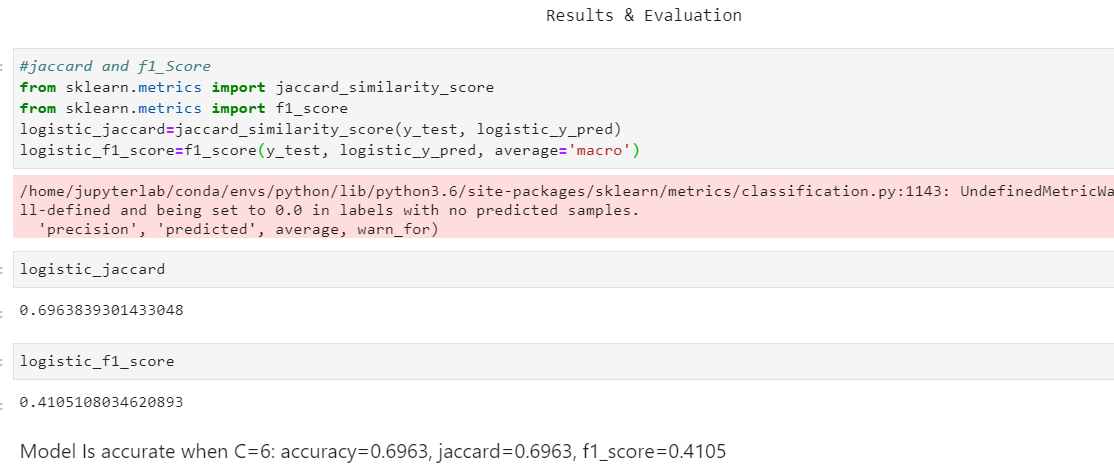
**1. KNN**

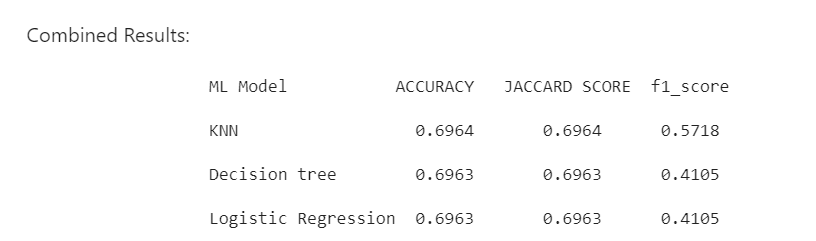


**2. Decision tree**



**3. Logistic regression**





So, From the results we can say that KNN is the best model to predict accident severity.

**Discussion**

At first, we had categorical data that was of type 'object'. So that kind of data cannot be used in all the ML models, so label encoding was used to created new classes that were of type int8; a numerical data type.

After we analyzed and cleaned the data and removed the null values, we used three ML models: KNN, Decision tree and Logistic Regression. In which logistic regression and decision trees gave us the ideal output but KNN has given us the most Accurate output as compared to other 2 ML models.

Evaluation metrics used to test the accuracy of our models were jaccard , f-1 score. Choosing different k, max depth and hyparameter C values helped to improve our accuracy output.

**Conclusion**

Based on data from weather conditions pointing to certain classes, we can conclude that particular weather conditions have different impact on whether or not travel could result in property damage or injury.