## Title: Car Accident Severity Prediction Report

Name: Rajashree.

# Table of Content

Contents

[**Title**: Car Accident Severity Prediction Report 1](#_Toc51955300)

[Table of Content 1](#_Toc51955301)

[**1.** **Introduction | Business Understanding** 2](#_Toc51955302)

[**2.** **Data Understanding** 2](#_Toc51955303)

[**2.1 Balancing the dataset** 3](#_Toc51955304)

[**2.2 Down-sampling** 3](#_Toc51955305)

[**3.** **Methodology** 4](#_Toc51955306)

[**3.1 K Nearest Neighbour (KNN) algorithm** 4](#_Toc51955307)

[**3.2 Decision Tree algorithm** 4](#_Toc51955308)

[**3.3 Initialization** 4](#_Toc51955309)

[**3.4 Pre - processing & Normalizing Data Set** 5](#_Toc51955310)

[**3.5 Train/Test Split** 5](#_Toc51955311)

[**3.6 Training Data Set** 6](#_Toc51955312)

[**4.** **Result and Analysis** 7](#_Toc51955313)

[**4.1 KNN** 7](#_Toc51955314)

[**4.2 Decision Tree** 7](#_Toc51955315)

[**5.** **Discussion** 8](#_Toc51955316)

[**6.** **Conclusion** 8](#_Toc51955317)

## **Introduction | Business Understanding**

These days car accident is matter of great concern. As an objective to lessen the frequency of car collision, an algorithm must be developed to predict the severity of an accident. The severity of the accident depends upon various factors like weather condition, road condition, speed of vehicle, light intensity factor, different atmospheric conditions like fog, intense wind, rain etc.

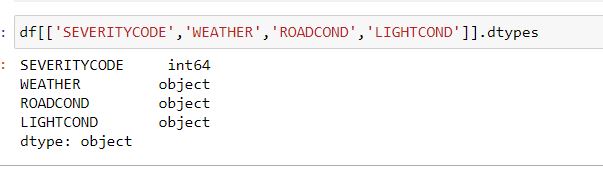
If some of the collision causing factors are been priorly alerted, depending upon the data received precautionary measures can be taken against the factors which has high severity, resulting in lowering the risk of collision. In order to meet this condition we can create ML model which can predict the outcome i.e. severity of the accident If we are successful in prediction of the factors responsible for the collision, it will have direct impact on decreasing the car collision cases.

## **Data Understanding**

There are different factors responsible for collision of the car such as, weather, road condition, light condition, etc. For manipulation consider this factors as attributes. There are total 38 attributes which includes numeric as well as categorical data. Shape of the data set is 38\*194674.

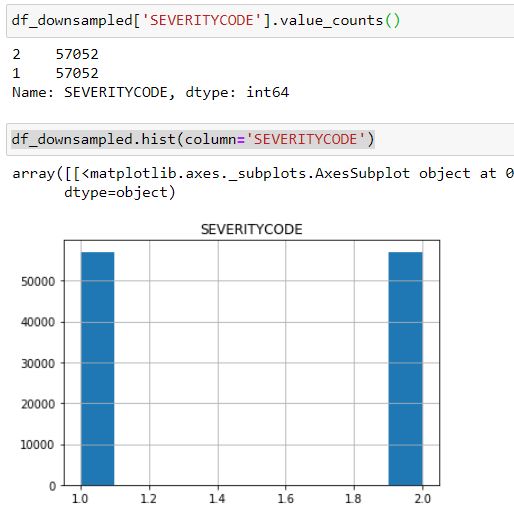
Here most important and targeted entity is the ‘SEVERITYCODE’, it can be predicted by training our model. Hence can be called as dependent variable. As it is said to be dependent, need to relay on some independent variable i.e. attributes as mentioned above.

In this data set, ‘SEVERITYCODE 1-> Property Damage Only Collision’ & ‘SEVERITYCODE 2 -> Injury Collision’. In this project we are extracting three attributes namely Weather Condition, Road Condition & Light Condition which are independent and can be used to predict the outcome. Weather Condition, Road Condition & Light Condition labelled as WEATHERCOND, ROADCOND & LIGHTCOND respectively.



### **2.1 Balancing the dataset**

Imbalanced classes put ‘accuracy’ out of business. This is a surprisingly common problem in machine learning (specifically in classification),



1. Balanced Data
2. Imbalanced Data

From the above information we can state that:

Output for class 2 is 57052 and that for class 1 is 132285. Here almost 70% of the data is biased towards class 1. This will affect the model. Hence randomly we have to down sample the model which will contain the output as class 1. Thus the data set will be balanced as shown above.

**2.2 Down-sampling**

Down-sampling involves randomly removing observations from the majority class to prevent its signals from dominating the learning algorithm.

## **Methodology**

As the data set is well labelled we can use Supervised Machine Learning technique to train data and predict outcome. Our predictor value contains binary data i.e. 1 & 2 ( Two outcomes only ) so can use K - Nearest Neighbour Algorithm, Decision Tree algorithm, Logistic model, etc.

For this project we are using K nearest Neighbour algorithm & Decision Tree algorithm.

### **3.1 K Nearest Neighbour (KNN) algorithm**

K-Nearest Neighbour is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection. The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.

### **3.2 Decision Tree algorithm**

Decision Tree is one of the most powerful and popular algorithm. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables.

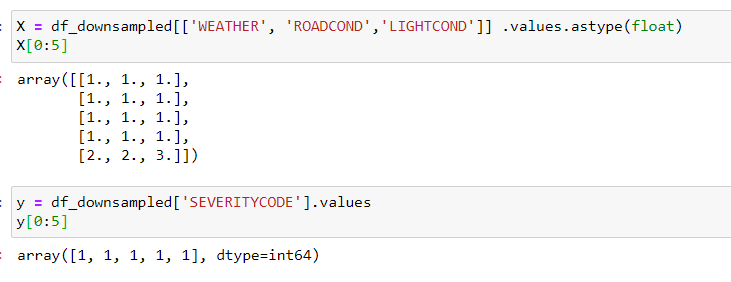
Decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non-parametric supervised learning method for classification problems.

### **3.3 Initialization**

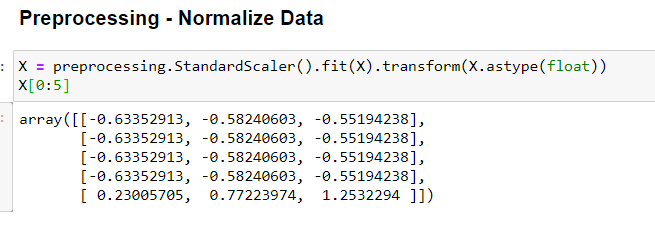
Defining dependent and independent variables –

X – Dependent

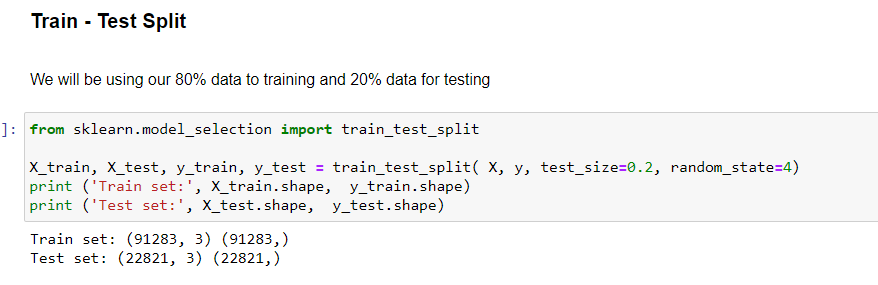
Y- Independent



### **3.4 Pre - processing & Normalizing Data Set**

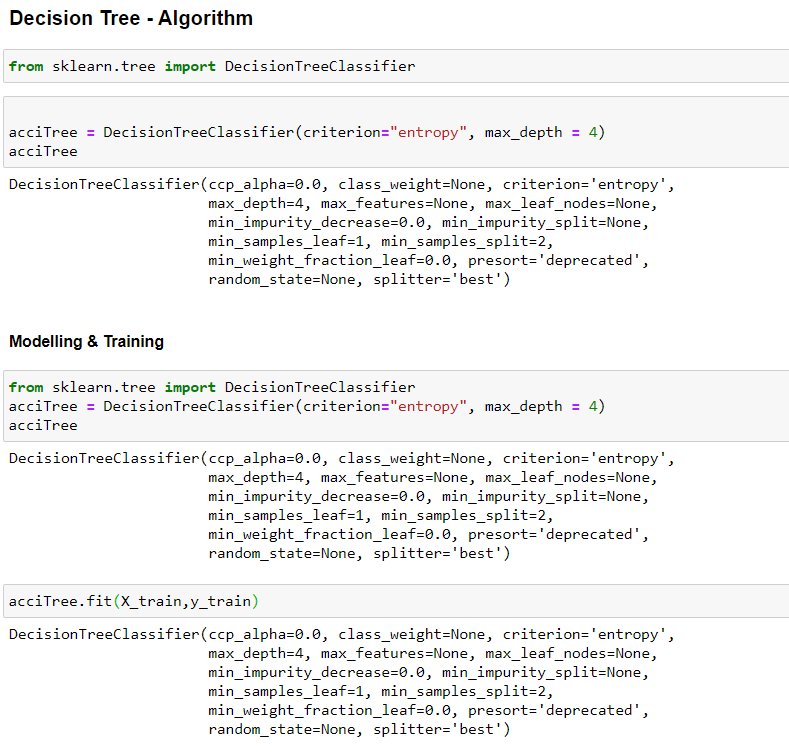


### **3.5 Train/Test Split**



### **3.6 Training Data Set**

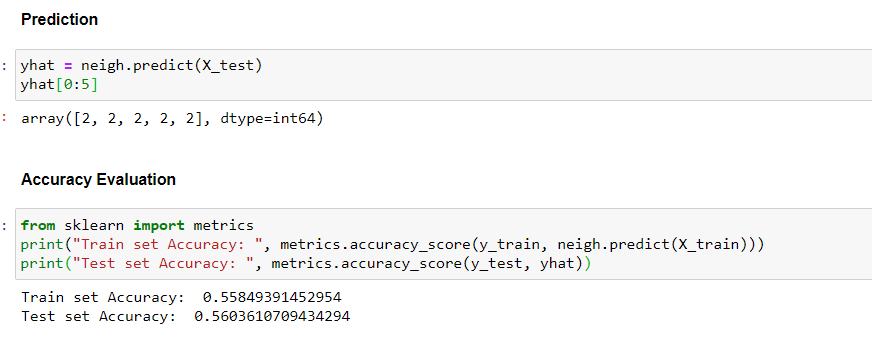




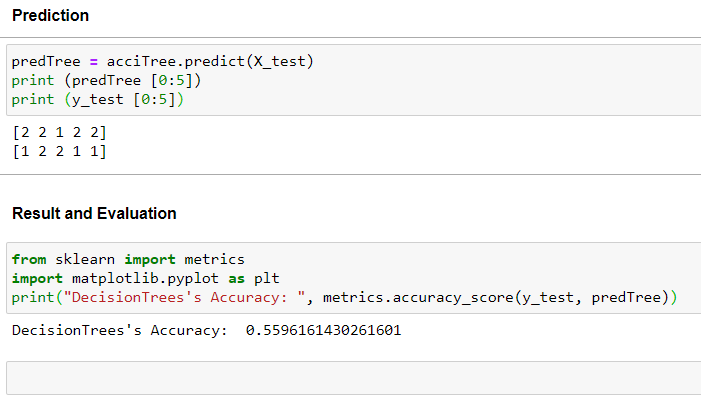
## **Result and Analysis**

Now we will check the accuracy of our models:

### **4.1 KNN**



### **4.2 Decision Tree**



## **Discussion**

At the start of notebook, we had categorical data that was of type 'object'. This type of data cannot be given through an algorithm, so label encoding was used for creation of new classes.

After tackling with this problem we came across another problem i.e. imbalanced data.As mentioned above class 1 was larger than class 2. Thus, down sampling was the option for getting the balanced data. Majority class was down-sampled. As a result we got the balanced data. Once we analysed and cleaned the data, it was then fed through two ML models; K-Nearest Neighbour and Decision Tree.

We came to conclusion that Decision tree has more significant values over KNN. Output can be predicted in better way using Decision tree than that of the KNN algorithm.

## **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). Decision Tree Algorithm has more significant output than that of KNN Algorithm.