DATA SCIENCE

**IBM Data Science Capstone:**

Seattle Road Accident Report.

horizontal line



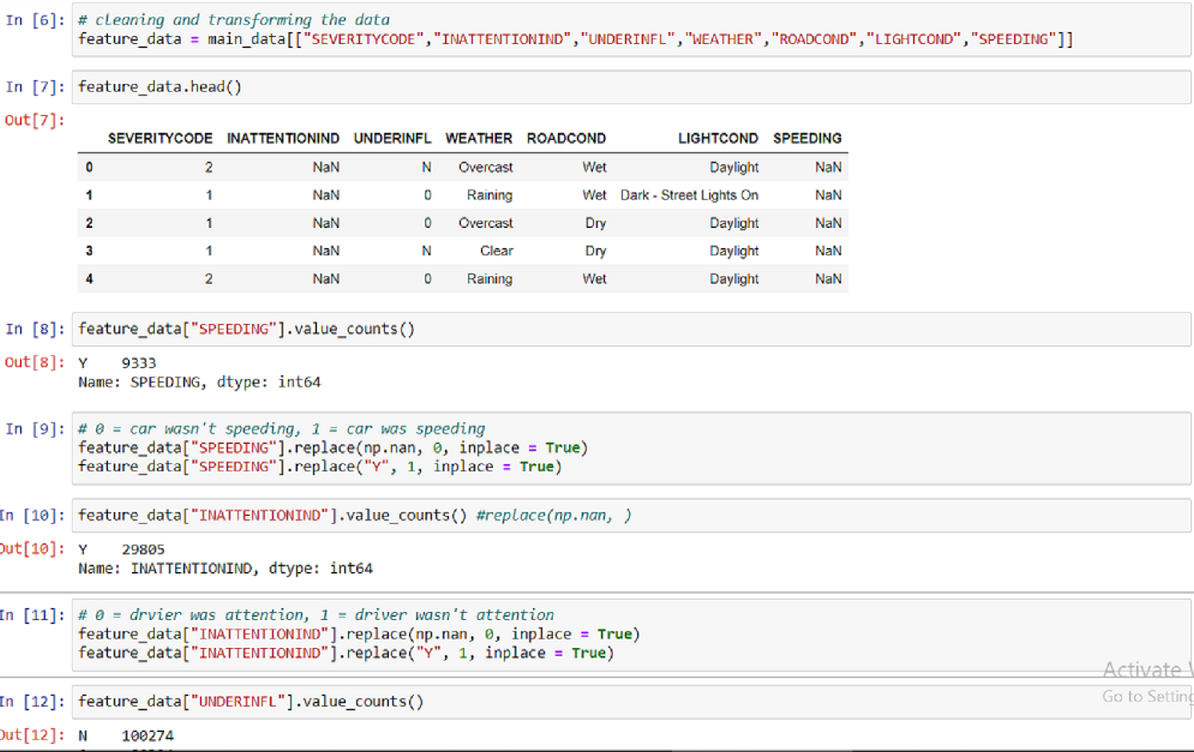
## Introduction | Business Understanding

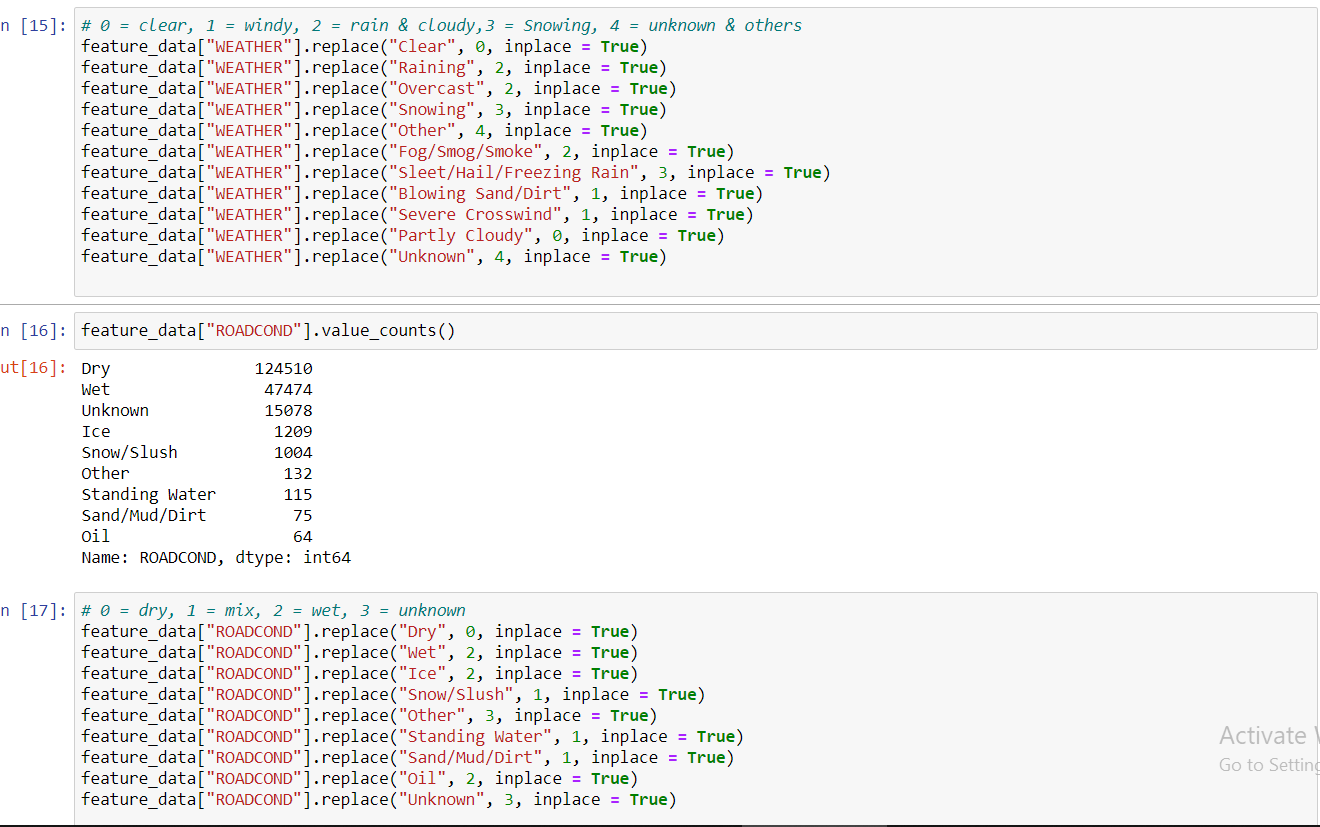
Seattle is a seaport city on the West Coast of the United States. Seattle is the largest city in both the state of Washington and the Pacific Northwest region of North America. According to U.S. Census data released in 2019, the Seattle metropolitan area's population stands at 3.98 million, making it the 15th-largest in the United States. Seattle residents get around by car, trolley, streetcar, public bus, bicycle, on foot, and by rail. With such bustling streets, it’s no surprise that Seattle sees car accidents every day. Climate of Seattle is one more factor which led to the road accidents. Climate of Seattle changes frequently from extremely hot to extremely cold. To reduce the car accident in Seattle city a machine learning model must be developed to predict the severity of an accident and alert the driver showing the chance of accident given weather conditions , road and visibility conditions. This will remind the driver to be more careful.

## Methodology

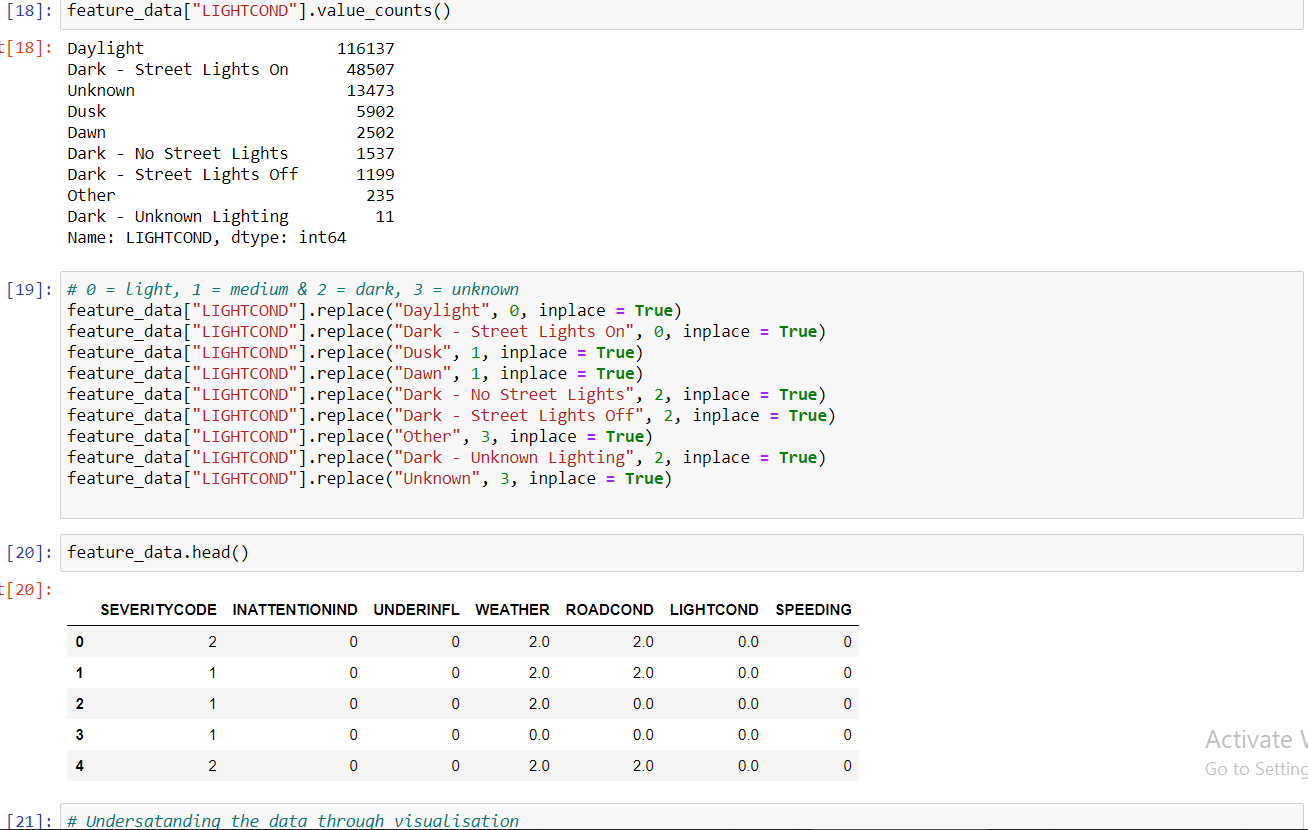
* **Cleaning and Transforming the Data.**

The data which i extracted from main data have categorical values that i converted them to numerical values, and also the categorical values belong to the same group so i put different categorical values to the group. the rows containing the nan values are dropped.





The final cleaned and transformed data looks like this:

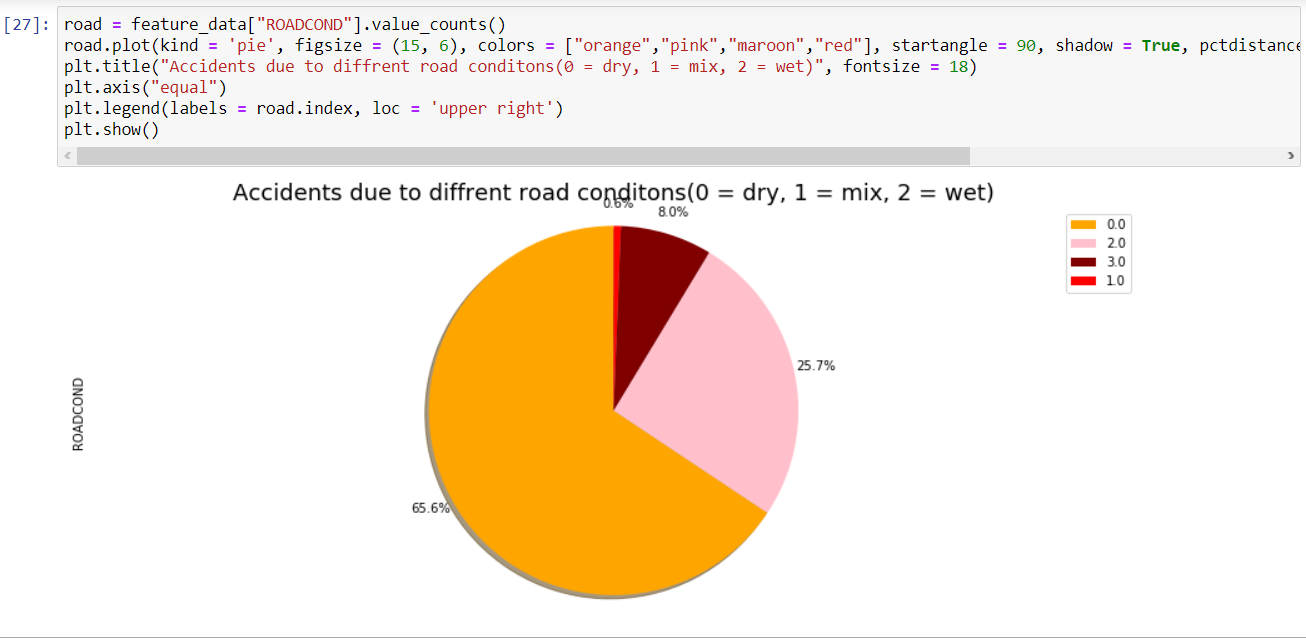


* **Exploratory Data Analysis:**

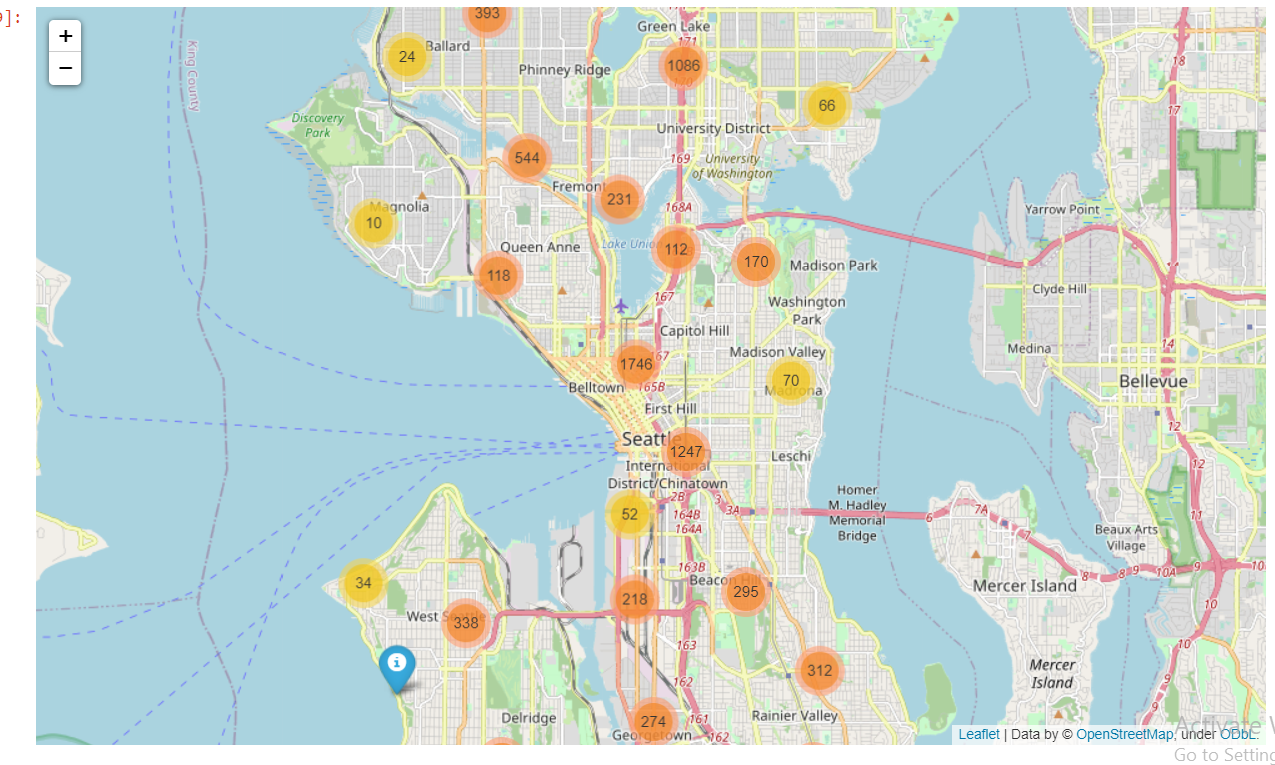
1. There is a high chance of Accidents in bad weather when I did analysis to see the number of accidents in seattle. The more number of accidents happened in clear weather.

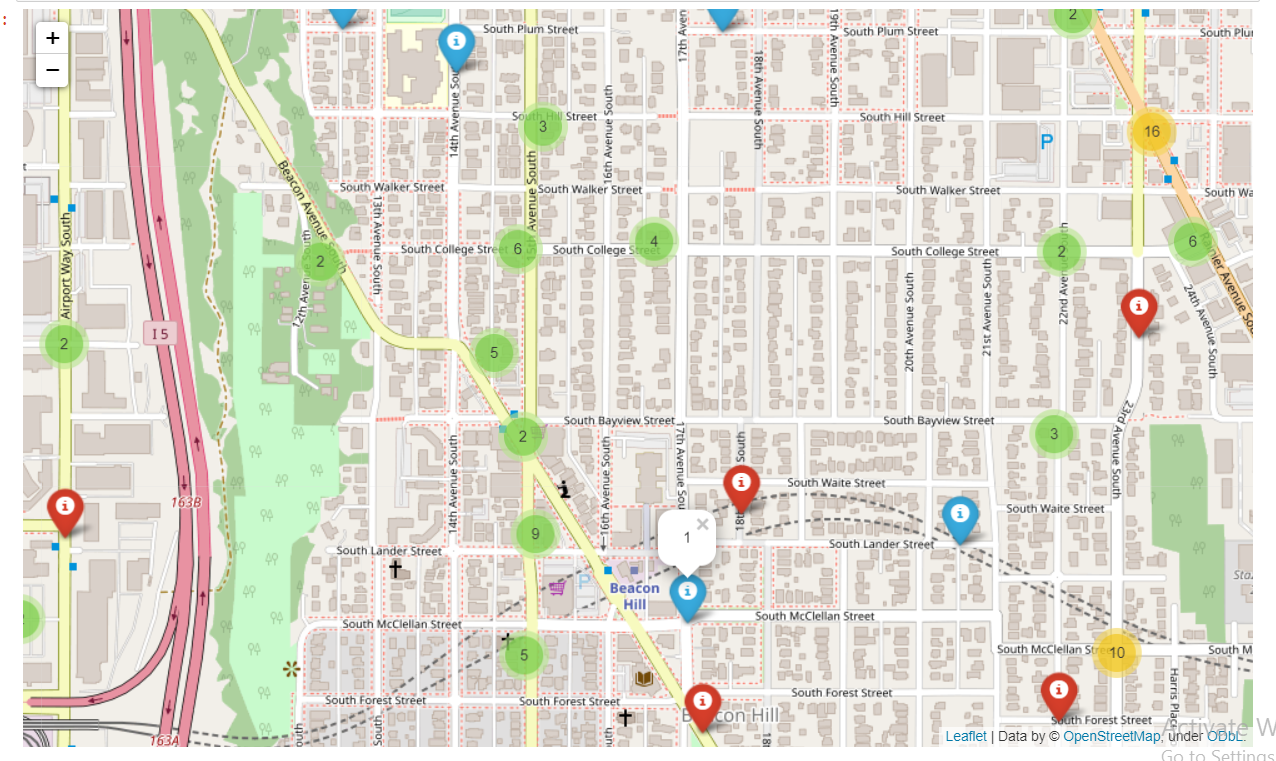


1. The conditions of Roads are one of the factors which lead to the road accidents. But in Seattle most of the accidents happened on dry roads.



1. The below map shows the places where accidents happened some of the accidents happened in front of educational centres and the workplace.





* **Machine Learning Model:**

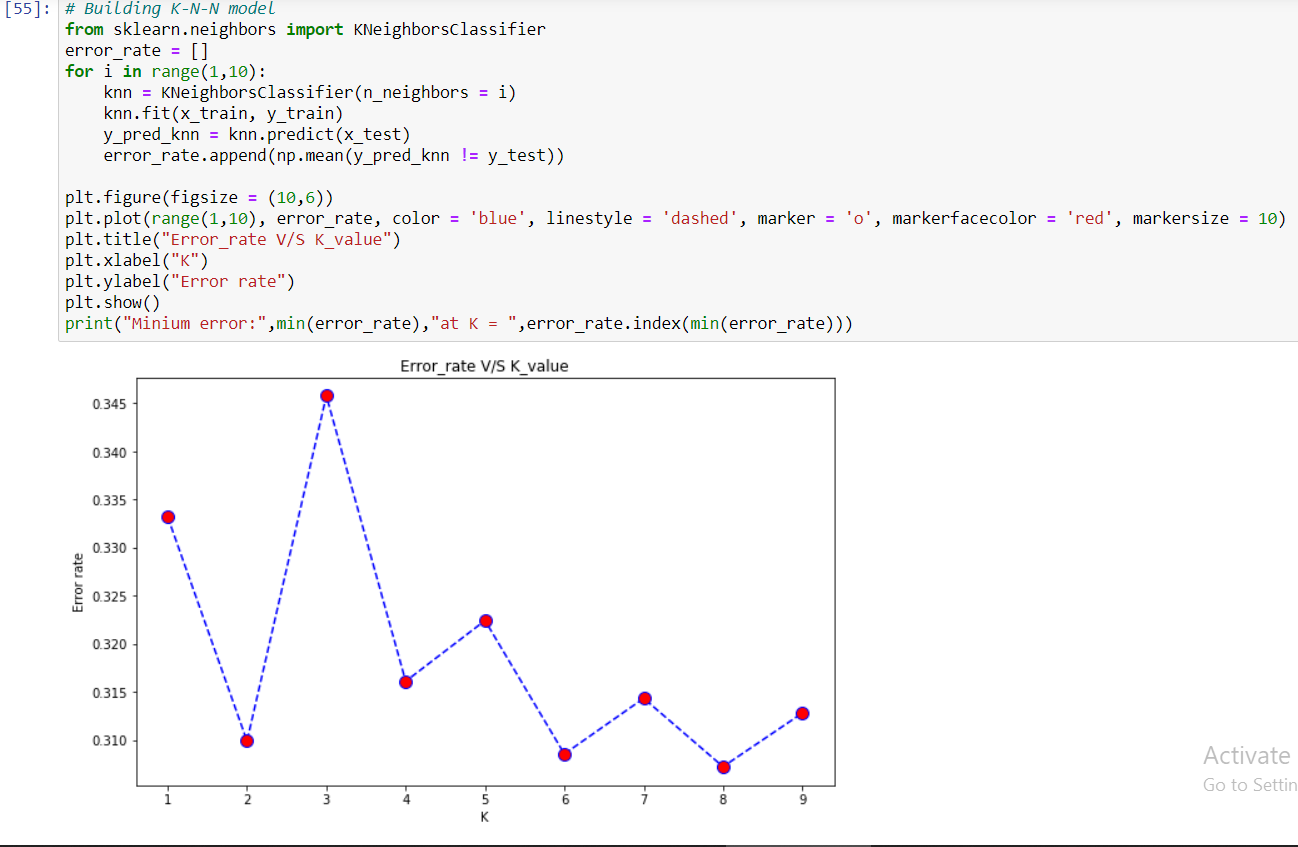
Since the data have both dependent and independent variables the supervised machine learning model will be used. For my model I choose SEVERITYCODE as a target variable which has binary type data. 1 for property damage and 2 for property and injury. So the classification model will be used. The data set is splitted into train test data sets and the 20 percent of the data is used for testing the model.

1. Logistic Regression:



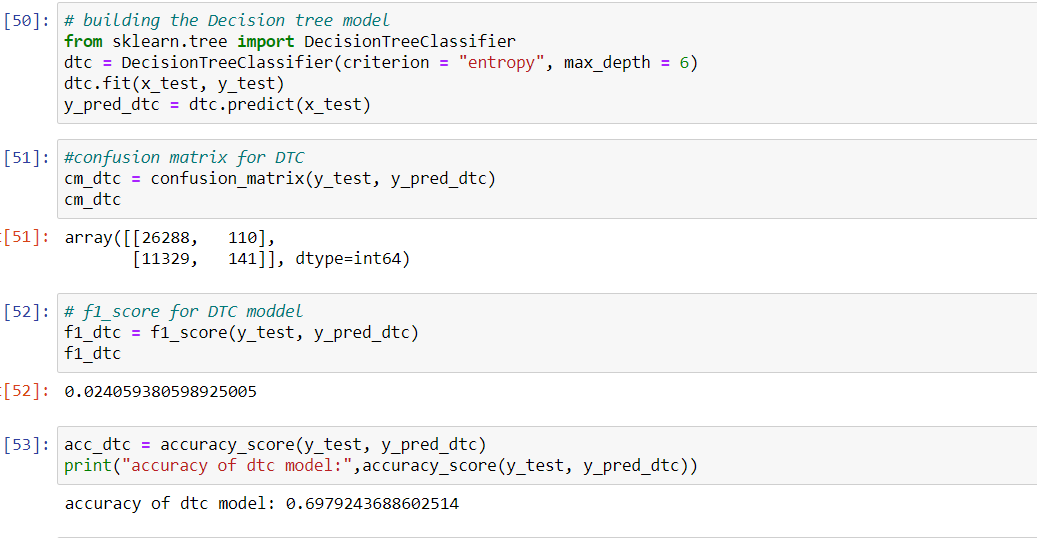
1. KNeighborsClassifier model:

The model give best performance at n\_neighbors = 8, the minimum error rate is at k = 8



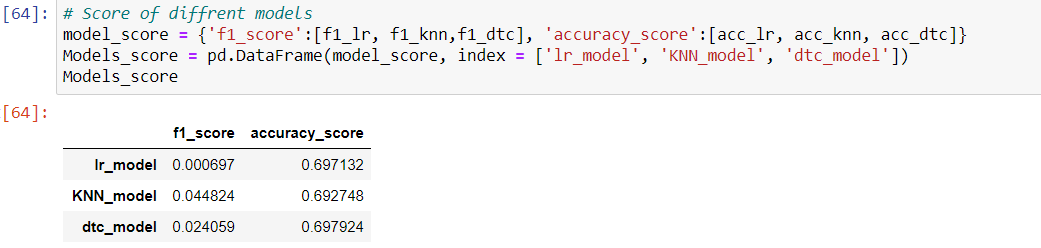


1. DecisionTreeClassifier:



## Result

Now we will check the f1\_score and accuracy score of our model:



The Decision Tree Classifier model has highest accuracy. So the Decision Tree Classifier is most suitable for this data set.

## 

## 

## Discussion

In the beginning the data is a categorical type of data ‘object’. So I transform them into int type of data by using the label encoder of sklearn. After this the nan values were dropped or replaced by the minority data values. Then after analyzing the data the transformed data is fed through the ML model .

Evaluation matrix is used to evaluate the predicted result of the model accuracy score and the f1\_score is used to calculate the accuracy of the model. Choosing different K and max\_depth helped to improve the accuracy of the model to the best possible.

## Conclusion

Based on analysis performed on data, weather conditions have a somewhat impact on accidents. The more accidents happened in clear weather. So we can conclude that the driver has to be more careful while driving on a clear day otherwise it can result in property damage and injury damage.