**CAPSTONE PROJECT - CAR ACCIDENT REPORT**

**MARIA RENEE CARRASCO**

September 2020

1. **Introduction**
   1. **Background**

A car accident is an unplanned event that sometimes has inconvenient or undesirable consequences. There are too many situations that can cause an accident, some of them are our responsibility but others not.

* 1. **Problem**

The Seattle government is interested to develop an algorithm that could help to avoid car accidents, considering many variables that could affect the prediction like Weather, Road conditions and light conditions.

* 1. **Objective**

The principal objective of the project is try to define if there are a possibility and what is the level of probability to be involved in a car accident, taking in account some attributes that could add or rest reasons to change the circumstances.

1. **Data acquisition**
   1. **Data sources**

Data is obtained from the Seattle police department and accident traffic record department from 2004.

First, it is necessary to understand the problem and the data which we are going to work. For that, all the attributes will be analyzed and prepared to construct a data set, considering missing data, type of data, correlations between data, etc. It is necessary to construct a supervised model and perform an evaluation to confirm the accuracy of it.

1. **Exploratory Data Analysis**

The data contains 37 attributes, where the independent variables are “WEATHER”, “ROADCOND” and “LIGHTCOND”, and the dependent variable is the “SEVERITY CODE”.

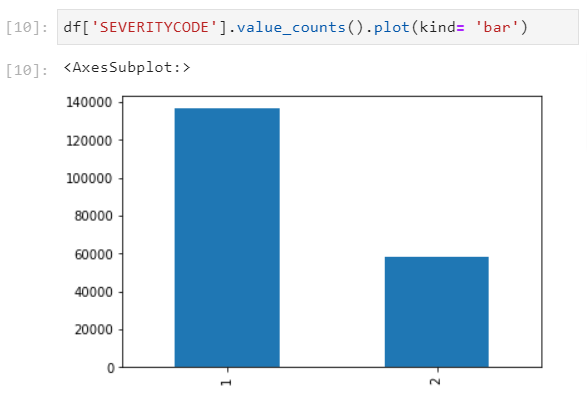


Fig.1

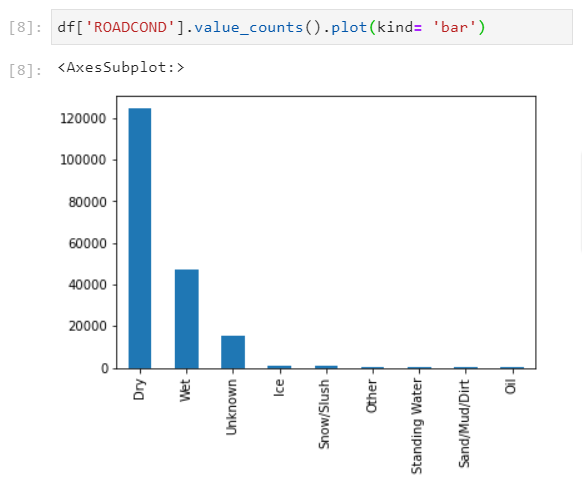


Fig.2

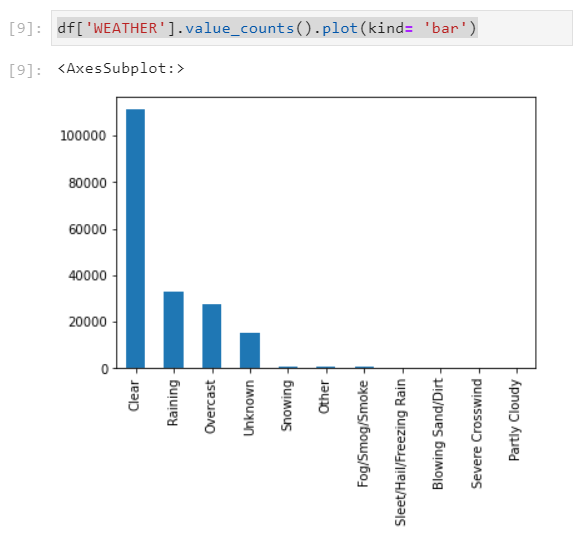


Fig.3

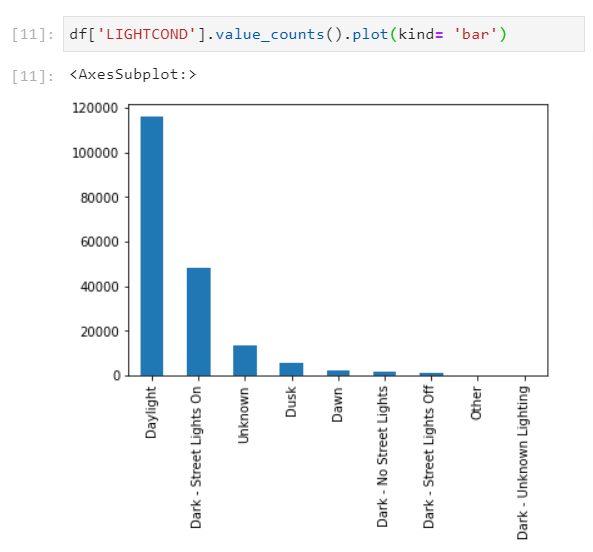
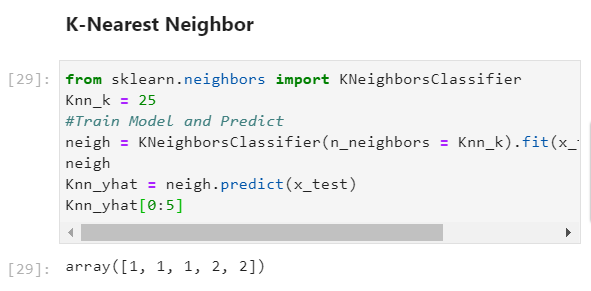
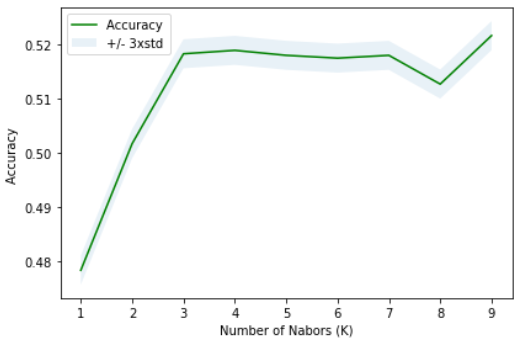


Fig.4

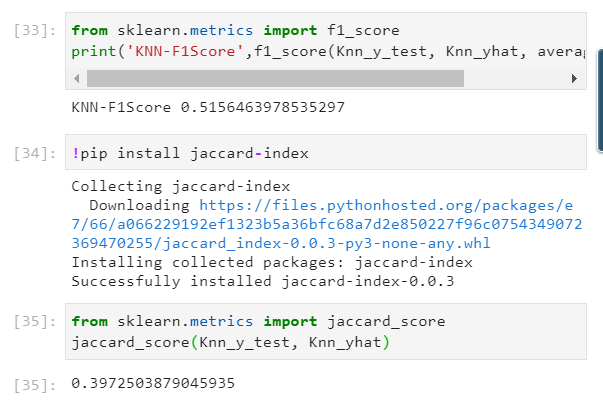
1. **Methodology**
   1. **K-Nearest Neighbor (KNN)**

Algorithm for supervised learning to solve both regression and classification problems. The value K (Knn\_k) is the number of nearest neighbors to examine. It is necessary to increase the value of k, and see which k is the best for your model.





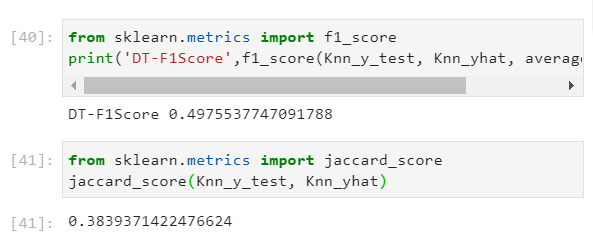
* + 1. **KNN Evaluation**



* 1. **Decision tree**

Decision trees are built are built using recursive partitioning to classify the data. For that, it is necessary to splitting the training set into distinct nodes, where one node contains all of most of one category of the data.

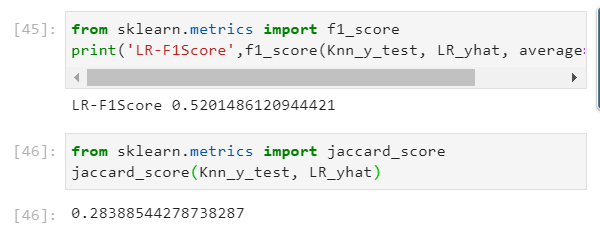
* + 1. **Decision Tree Evaluation**



* 1. **Logistic Regression**

The dependent variable (“SEVERITYCODE”) is finite or categorical. Logistic regression has been applied to understand the relationship between dependent variable (“SEVERITYCODE”) and other attributes (“WEATHER”, “ROADCOND”, “LIGHTCOND”). In this way, it is possible to predict the car accident severity according to the variable selected to be analyzed.

* + 1. **Logistic Regression Evaluation**



1. **Conclusion**

Based on the data, it is possible to conclude that particular conditions have impact at a different scale that could result in property damage or injury. The following table shows that the model classification KNN is the best model to predict car accident.

|  |  |  |  |
| --- | --- | --- | --- |
| **MODEL** | **F1 SCORE** | **JACCARD SCORE** | **ACCURACY** |
| **KNN** | 0.52 | 0.40 | 0.51 |
| **Decision Tree** | 0.50 | 0.38 | 0.56 |
| **Logistic regression** | 0.52 | 0.28 | 0.53 |