**Business Understanding:**

**Background of the study**

The usage of car has increased, especially in past few decades since the 60’s, as the price of the cars are more affordable to the open public this gives us a correlation between the increase cars on the roads and the increase of accidents. According to the 2017 WSDOT data, a car accident occurs every 4 minutes and a person dies due to a car crash every 20 hours. Fatal crashes went from 508 in 2016 to 525 in 2017, resulting in the death of 555 people. This number has stayed relatively steady for the past decade. Conditions of road, light, weather and a multiple of human factors, are some of the many factors that can influence a car accident

**Problem and Interested parties**

To help reduce the severity and frequency of car collision, this project wants to use the Seattle car collision data to generate insights on how modeling can help reduce accidents. The objective of this project is to be able to predict the severity of a road accidents under certain conditions like weather, road and visibility. It can be use for specific resources, like municipal police or emergency teams (fire fighters or hospitals), as a way to predict the flux of accidents that they can expect in a specific day. But it can be used by the public, since the common car drivers or a pedestrian. In the future it can be useful to predict the location of the most common places of an accident to eld place and help the government officials to improve roads or be more watchful to certain roads. By looking into road condition factors and address types such as intersection and block, we could see if there is any possible improvement in road condition and city planning.

**Data understanding:**

**Data Source**

The Data used in the following project is the Car Collision dataset of the city of Seattle, from IBM Cloud which contains among others the collision data, address, and a few different conditions such as Weather, road and light. The dataset includes 38 columns and 194673 observations.

**Data cleaning**

The first objective of the Dataset is going to be cleaning the “noise” or remove all the columns that we are not going to require for the base of this specific study. It will be need to have in regard the missing data, which there are about 10527 records in this category. The target variable will be 'SEVERITYCODE', as it is a measure of the severity of the accident and varies between 1 and 2. The attributes 'WEATHER', 'ROADCOND' and 'LIGHTCOND' are going to be converted from Categorical to Numerical values, as part of the normalization process. Values for SEVERITYCODE represent the following consequences of the accident: Property Damage Only Collision and Injury Collision Lastly, we need to have in consideration that this is unbalanced labeled dataset, so we are going to need to normalize the data first of all, before they being used.

**Feature selection**

It will be considered as well the factor 'SPEEDING', to use as a dependent regarding the Weather and Road conditions. The attributes 'WEATHER', 'ROADCOND' and 'LIGHTCOND' will be used to weigh the severity of an accident. But it is going to be considered the “ST\_COLCODE” (condition of accident) and “INCDTTM” (date and time of the accident).

**Data Preparation**

The original dataset needs some preparation to be fit for analysis.

From the 33 columns of this dataset, many are not useful for this model. Many of the features need to be converted to numerical type, using label encoding, which will create new numerical columns.

4947 rows have all the three attributes in study in clear, so they will be removed and not considered for analysis.

10209 rows will be discarded from the analysis because all the three attributes are 'Unknown' and would be useless to train the model.

In the remaining 179517 rows we notice a great unbalance in the target variable:

SEVERITYCODE=1 122604 (68%)

SEVERITYCODE=2 56818 (32%)

This can be fixed by dropping the majority class in order to achieve a 50% representation of each class.

LOCATION – WEATHER – ROADCOND – LIGHTCOND – ST\_COLCODE (describes the collision) – SPEEDING - INCDTTM (date and time of the incident)

**Methodology:**

All the information regarding the data that was utilized in the project was provided by IBM. It was used the Github repository to store the information, while Jupiter Lab was the program selected to process the database in the course of the project.

**Result section:** Describing your results and also any output supporting the results.

**Discussion section:** This is an important section as it shows all the inferences and analysis that you did on your project.

**Conclusion section:** Concluding statement of the project and also its future scope.