Introduction/Business Problem

I have been hired by the city council of Seattle to build a model that will enable them to predict the severity of road accidents in order to protect lives, propreties and infrastrucrures damages. In fact, the number of road traffic accidents increases every year and the damages whether on proprieties such as cars, infrastures or on humans injury, death are quite alarmant. The pressure is now on the council as the communities accuse the council not doing much to alleviate the problem, the budget allocates every year for road infrastrustures repairs increases and always almost runs on decifit. Not to talk about the cost of other damages and the loss of a loved one. Therefore, I have the task to build a reliable and appropriate model to predict accidents severity and allow people to take informed decision on their driving.

The question is to build a model that will predict the severity of road accidents Therefore what are the conditions that affect the severity of road accidents? Are the data already available within the council or am I going to fetch them somewhere else? These are some of the issues at hands before talking about any modeling.

While discussing the issues with the council as I tried to understand the problem, I was given access to their database where is stored years of the Seattle Collision dataset in a csv format.
Next, I have to define measurable conditions that could affect the severity of the road accidents.
In the literature, it has been reported that many studies aim at predicting the severity of an accident using various information from the accident in order to understand what causes an accident to be fatal. Specifically, Chong et al. [1] identified that the seat belt usage, the light conditions and the alcohol usage of the driver are the most important features whereas Abellan´ et al. [2] found that the type and cause of the accident, the light condition, the sex of the driver and the weather were the most important features. Using BoxPlots, a way of statistically representing the distribution of the data through five main dimensions in addition to the aforementioned information, I identified the weather conditions, road conditions and light conditions as important measurable attributes to weigh the road accidents severity. Interestingly, less than 3% data of these attributes were missing.
Furthermore, it has also been reported that road accidents prediction are seen whether as (1) a regression problem, predicting the risk of accidents translated into different ways, or (2) a binary classification problem, predicting whether an accident will occur. I choose to approach it as a classification problem because the dataset only contains 2 severity codes implying a classification issue. Nevertheless I will be using different algorithm models in order to deliver a more accurate solution.