**Introduction/Business Problem**

For years, road accidents have ravaged lives causing physical, financial and mental effects on individuals involved and families. These accidents can be individually driven (Driving Under Influence (DUI), poor vehicle maintenance, exceeding speed limits etc.) or physically driven (bad roads, unfavorable weather conditions, inadequate traffic signs and symbols). Road accidents are a danger to every society and as such national governments recognize this and put in place necessary infrastructures that can help alleviate this problem and keep the public safe. However, the efficacy of these infrastructures are often times poor and driven by little or limited considerations. Therefore, the objective of this project is to build a model that takes into consideration both physical and non-physical parameters in order to predict accidents severity and help mitigate this menace. Seeing as the demographic of road users is conventionally limited to age, the recipients of this project is aimed at the general public. In addition, national governments can also benefit from this project as the model can be used to determine areas that are at high risk of severe accidents and allocate the necessary resources to these areas to help combat this issue.

**Data**

The data to be utilized for this project is the Seattle accident data which has been sourced from Kaggle (an opensource platform for datasets) and made available in csv format. The dataset contains 38 attributes including both physical and non-physical parameters and 194,673 cases of road accidents. In the initial stage, data analysis and visualization techniques would be employed to derive initial insights of the data and also determine the level of correlation that exist between the label and predictors. Since this project aims to be predict accidents severity, the SEVERITYCODE attribute would be used as the label, while other attributes e.g. VEHCOUNT (number of vehicles involved in collision), INJURIES (number of total injuries in the collision), SERIOUSINJURIES (number of serious injuries in the collision), PERSONCOUNT (number of people involved in the collision), PEDCOUNT (number of pedestrians involved in the collision), INCDATE (incident date), WEATHER (weather conditions during collision), ROADCOND (road condition during collision), SPEEDING (whether or not speeding was a factor) among others would acts as independent variables (predictors) of the model. The data contains some missing values and as such some feature engineering would be performed on some columns (attributes) to fill in missing data. However, columns that have too much missing values will be dropped from the dataset.