**Introduction**

**1.1 Background**

An accident is when a vehicle collides with another vehicle, pedestrian, animal, obstacles, etc. Many factors contribute to collision including road environment, driving skills, speeding, impaired due to drug/alcohol, etc. The total number of persons and vehicles involved, pedestrians and injuries inflicted affect the severity of the accident.

**1.2 Problem**

The number of the accidents all around the world is on a continuous rise. A slight delay in the availability of the medical services may be the difference between life and death. This gap can sometimes be a result of the lack of knowledge of the severity of the accident. Having a clear picture of the accident can help authorities provide better emergency services. A model that can predict the severity by weighting in the different factors of weather, road, visibility, light conditions, etc. can be very effective to bridge the gap. Lot of the accidents are also the result of careless driving. How do the environmental factors impact, and their contributions? Taking all the factors into account is a necessity.

**Data**

The data required here is one which contains details about the environmental factors as well as the driver details like if the person was under some alcohol/ drug influence. Environmental factors can even lead to a pre-assessment of the probability of travelling unfriendly conditions and warning can be flagged in advance. The data will be used for classification using Logistic Regression and predict the probabilities of the severity of accidents. We will use a historical dataset that fulfils all the requirements. The dataset for this project is provided by Coursera.

The dataset has around 190k rows and 37 attributes including, speeding, road environments like visibility, light conditions and careless driving like inattention and driving under influence. Many attributes are unique identifiers for accident used by department of transportation and other organizations. These attributes are dropped as they will help predict the severity of the accident. Most of the attributes required close observation and data cleaning. For example, “speeding” attribute consisted of only ‘Y’ and ‘Nan’ values. Here, it can be inferred that only accident that involved speeding were documented as ‘Y’ and rest were left empty. The value “Nan” can be easily replaced as “N” for No or both “Y” and “Nan” can be replaced by “1” and “0” respectively.