**Data science – Capstone project**

**Business problem**

Car crashes is the highest cause for accidental deaths and injuries. The Severity of a car accident might be influenced by a number of parameters. The ability to predict the severity could help to prevent future car crashes or reduce their severity.

**Data section**

I will use the “collision data” dataset for building the model. The label is the incident severity (1 or 2) and the rest of the variables are the independent variables. Since there are two categories of severity, I will consider this problem as a classification problem. The independent variables include both numerical and categorical variables. I will investigate the number of missing data for each variable and delete variables with more than 40% of missing data. I will also delete the column “severityCode.1” Since it is identical to the label (y) column. I will balance the data using under balancing method. I will standardize the data using the “standardScalar” method.