1 Introduction

Every year, approximately 1.35million people died in road accident, so it is very important to reduce the frequency of car accident. In this project, a machine learning based system will be developed to predict the severity of an accident when given some variables, such as road condition, light condition, car speed, driver and so on. The objective of the model is that when the conditions are bad, the system will alert the driver to remind them to be more carefully.

2 Data

In this project, I use the dataset of Data-Collisions.csv. Firstly, I exploited the dataset to see all data. There are 38 columns and 194673 rows in total. I select ed the features of ROADCOND, LIGHTCOND, WEATHER, SPEEDING and SEVERIT YCODE . I am going to exploit how these features are related to the severity. There are lots of NaN in the features, I replaced NaN with 0 and Y with 1 in SPE EDING column. Then drop all the NaN in other features. After that there are 5 c olumns and 189337 rows in the new data frame. I encoded all the features wit h numbers and converted the data type to int. I extracted "WEATHER","LIGHTC OND","ROADCOND","SPEEDING" as X, "SEVERITYCODE" as y in machine learnin g feature sets. Please see the details in the Code.