Predicting the Severity of an Accident

1. Introduction

1.1 Background

In traffic situations, passengers are prone to accidents on the roads. This can be due to different factors such as the weather conditions, the road conditions, the light conditions amongst other factors. These attributes are being stored by the traffic system in Seattle. It is highly recommended to be able to predict the severity of an accident based on the factors available to prepare for the casualty before the accident occurs.

1.2 Problem

The dataset provided for the Seattle city contains a total of 38 attributes (relating to the accidents that occur on the road) and the labelled data which describes the fatality of an incident. Given this dataset, the aim of this project is to select the necessary attributes that will be used to build a model that will help to predict the severity of an accident.

1.3 Interest

Residents of Seattle will find this helpful in predicting how severe an accident will be if they get into one based on the factors available. It will also be useful for traffic attendants and paramedic to prepare for accidents likely to happen. This will help reduce causality.

2. Data acquisition and cleaning

2.1 Data Sources

The dataset for Seattle road accidents was provided containing a total of 194673 observations and 37 attributes with many of them being categorical attributes.

2.2 Data Cleaning and Preprocessing

Out of the numerous attributes, only selected attributes were used because they relate to the severity code based on their description in the metadata. The attributes are: 'ADDRTYPE', 'COLLISIONTYPE', 'JUNCTIONTYPE', 'WEATHER','ROADCOND','LIGHTCOND', 'PERSONCOUNT','PEDCOUNT', 'VEHCOUNT', 'HITPARKEDCAR'. The other attributes were dropped either because they do no relate to the target variable or because they have a lot of missing values.

The following attributes are categorical values and needed to be changed to numerical values using the one hot encoding; 'ADDRTYPE', 'COLLISIONTYPE', 'JUNCTIONTYPE', 'WEATHER','ROADCOND','LIGHTCOND' while ‘HITPARKEDCAR’ was replace with 0 and 1 to represent its categorical values.