Predicting the severity of vehicle collisions based on environmental factors – a case study

This case study is the Applied Data Science Capstone Project in the IBM Data Science Professional Program. This study does not only serve as a demonstration of the skills that I have acquired, hopefully it can also demonstrate the power and versatility of common data science methods. The audience of this article are not necessarily just scientist and researchers but the general public and especially those with an interest in the applications of data science.

All the code as well as the used data set can be found on GitHub <https://github.com/dcadosch/Coursera_Capstone>

# Introduction

At least since we use wheeled motor vehicles for transportation, traffic collisions are an unfortunate but inherent aspect of it. The severity of accidents varies for a number of reasons. Roads may not be designed in a safe way or crucial parts of vehicles may fail while driving due to insufficient maintenance or due to engineering faults of the vehicle manufacturer. The driver may be distracted or violate safety rules and laws by speeding or driving under the influence of alcohol or drugs. There are also environmental factors that may influence the severity of a collision. A law-abiding driver who takes appropriate care of her or his vehicle can only influence a few of these factors that govern the outcome of a collision. However, it is difficult to judge these factors without a thorough analysis of the data of past collision events.

In this case study I am creating a machine learning model that estimates the severity of collision events based on a number of environmental factors. The severity of collisions is divided into multiple classes, which means that we need to employ a classification algorithm. In order to be able to choose an optimal classification algorithm, we will deploy multiple methods and compare their performance by measuring several evaluation metrics.

This case study will focus on traffic accidents in an urban environment in the United States of America. Hence, it might not be applicable to situations that differ considerably from such a scenario. The general approach of this study might nonetheless serve as a guide to create a similar model with data from another source.

# Data understanding

The data to train and evaluate the model was obtained from the 'City of Seattle Open Data Portal' (<https://data.seattle.gov/Land-Base/Collisions/9kas-rb8d>). The data set contains all reported vehicle collisions in Seattle from January 1, 2004 to October 9, 2020. The table contains 40 columns and 221,525 incidents. Not all of the columns contain relevant or usable information for the task at hand. Furthermore, not all incidents have a complete set of information. These incidents may need to be removed from the data set.

The most relevant feature is the severity of an accident. In the original data set the severity is divided into four classes: property damage, injury, serious injury, and fatality. The model will be trained to predict the severity class based on other available information. Independent features that a driver may observe before a collision occurs are for example the road and light conditions, the weather and the kind of road (s)he is in. All these features are divided into about ten classes each. Other features such as the month of the year or the time of the day may be extracted from the timestamp in the data set. Collision data that can only be obtained after the collision occurred such as the type of collision or whether any person involved was driving under influence will not be used in the training of the classification algorithms.

To get an initial overview of the nature of the data that will be used it might be helpful to have a geographic visualization. The following map shows all recorded collision incidents in 2020 up to October 9 for which the coordinates were available. From this map we can see the boundaries of the geographic area from which the data stems. We can further see that the incidents are not homogeneously distributed. There are certain areas where incidents are clustered. That clustering might be due to higher amounts of traffic, particularly dangerous roads or simply due to stochastic effects. Since we do not know the reason for the clustering, we will not use the location data any further in our study.

