# Capstone Project - Car Severity Analysis model

Jesús A Molero Cano

October 22, 2020

## 1 Introduction. Problem definition

**In this project we create a model able to define the severity of an accident thanks to a big set of data from the city of Seattle**. However, the conclusions and the model can be applied to other several cities with similar characteristics in America, Europe, Asia, Africa or any other part of the world.

The results of this project are very useful in car accident prevention, and they may save the lives of many citizens. Just to be aware of the magnitude of mortal accidentes, car accidents were responsible of 33,654 deaths just in 2018 in the US and of an average of 1.35 million people deaths worldwide each year. Those traumatic statistics are a just a sign of the importance of this issue globally.

Moreover, these results may be useful to Law Enforcement in their activities of traffic control. The results may indicate the conditions that are likely to cause an accident, what may help in the prevention activities.

## 2 Data

Based on definition of our problem, we will create a model able to predict severity of an accident giving certain conditions:

External conditions: Weather, road condition, light at the time of the accident, road state. Internal conditions: number of vehicles in the accident, speeding, drivers/people in the accident, pedrastians involved, alcohol influence and type of crash intercourse.

The CSV file given by the IBM Coursera course is very detailed an it offers a good source for developing this project. The raw data reachs almost 20000 rows, this quantity is enough in order to accomplish the task. Not all the attributes will be used in the project, this is the reason why some of them are being deleted. The Data given by IBM will be prepared and filtered in the following process. **The empty data will be filled with the most frequent values in the dataframe**.

Following data sources or a given data the model will generate the required information:

Severity in case of an accident with certain conditions. This severity will have a rate of certainty given by the models developed in this project. The models developed in this project are:

- Decision tree model.
- Logistic Regression.

The accuracy of both models is checked by the following Data Science classification accuracy tools:

- F1 Score.
- Jaccard score.
- Precision score.

**Also, the results of these models will give a good estimation of the quality of the data used in this project**. Sometimes, the attributes do not offer enough detail in order to produce a good model, this conclusion is important for a further deeper analysis of the issue with an improved data set with more data attributes.