# Introduction/Business Problem

Traffic accidents are an actual problem and responsible of millions of casualties every year around the world. In 2019, the number of traffic fatalities in Canada was 1,922; which implied an increase of 3.6% compared to 2017

Our target is to create a model that can warn of the possibility of being involved in a car accident and the potential severity based on the weather and road conditions and could be use by the traffic authorities to prevent\reduce the motor vehicle fatalities.

Reducing the traffic mortality will benefit the society in general and specially to the drivers that will save their lives or will avoid severe accidents, also their families will skip a traumatic and economic drama. Insurance companies will be able to provide more competitive products. Safer roads and transportation infrastructures could be optimized by government and granted to citizens

# Data

The example dataset from the Seattle city does only contain collisions with severity 1 or 2, which could be insufficient for the building a proper model. Therefore we have extracted from [here](https://data.seattle.gov/) a fresh dataset with all type of severity collisions

The worksheet has a total of 40 attributes, with information from collisions in Seattle from 2004 up to now. Some of the more relevant attributes are:

* SEVERITYCODE: this is the target variable, the rest of attributes will be used to predict the severity of a motor accident. This is a categorical variable, but since a classification algorithm will be used, it is not required to convert it to a numerical variable
* WEATHER: categorical variable that describes the weather conditions. It must be converted to numerical variable or use dummy variables
* ROADCOND: another categorical variable that must be processed (using the one hot encoding) before creating the model.

The rest of the attributes are not relevant for our model and will be discarded

Duplicated data or rows with empty values will be also discarded. After that, there is only one row with target value “0” (unknown)

Data is unbalanced, therefore we should proceed to balance it before training & validating, otherwise the model will be biased

# Methodology

For solving our business problem, we need to use a supervised machine learning model. Since the target variable is categorical, we should use a classification algorithm, the Decision Tree would be an optimal solution

As described before, the target of our project is predict the severity of an accident based on the weather and road conditions, therefore our independent variable “X” will consist of WEATHER and ROADCON attributes, in total 20 dummy variables.

The sklearn library we will use to split the our dataset in train and test parts, as result of some tests it was the 80% training – 20% testing share provides the highest accuracy in the test

# Results

As result, our model is able to predict with a 82% of accuracy the test values after being fit with the training data

# Discussion

The target data is unbalanced, this could potentially led to biased model. Our model is behaving quite well over the test data and satisfactory level. However, in further steps it will interesting to investigate how balance the data, unluckily this topic was out of the scope of this certification, and was not explained in the material

dfSeattle['SEVERITYCODE'].value\_counts()

1 133449

2 57611

2b 3039

3 338

0 1

# Conclusion

 The decision tree supervised model has shown that the weather and road conditions have a real and high correlation with the severity of a collision