1. **A description of the problem and a discussion of the background. (15 marks)**

**Intro**

The scope of the project is to create a model for road traffic accident classification by applying machine learning techniques. The data set contains information about the accidents across the Seattle area. Data includes a rich set of features such as an address, location, latitude, longitude, accident type, the severity of the accident, description of the accidents, weather condition, road condition, visibility condition, number of people and cars involved, and so on.

The chosen dataset contains data from crashes since 1/1/2004 until 9/9/2019. It contains information that might influence the accident occurrences such as weather conditions and post-accident information such as case numbers. The data set comes with a pdf file containing the description of the available features and values.

Such an analysis might present interest to people and organizations wanted to understand the impact of exogenous variables such as environmental ones or road conditions and more or less the endogenous ones such as the speeding or under influence factor. A better understanding of the impact of available variables might provide some guidance on efficient resource allocation or policy development.

In addition, describing the importance and the impact of the variables at hand on accident occurrence might help insurance companies to review their cost and premium allocations.

**Project Outline**

This is the initial version of the final report for the project. The whole analysis of the data set is done using Python language and its libraries. The report, notebooks, code, and data are available on my GitHub profile here.

The goal of the project is to predict the severity of the crash and which variables influence the accident occurrence. In the data set, the target variable is called SEVERITYDESC and takes two values: “Property Damage Only Collision” and “Injury Collision”.

To achieve this objective, the following steps will be taken:

* Feature Exploration (with data cleaning)
* Dimensionality Reduction
* Model Building
* Optimization and final model selection.

Even though the steps have been defined above, during the exploration of each step in part we might have to go back and forth to adjacent steps to make necessary adjustments into the data structure. The final report will reflect the main points of the process without describing all the intermediate steps and will exclude irrelevant explorations that sometimes are needed to better understand the data and/or the problem. Although, I will mention paths that were identified to solve other interesting problems while working on the data.

An exhaustive exploration of the features will be done to identify valuable insights about the features and values, and how they correlate with one another.

At this point in time, we can state with confidence that our problem is a binary classification one, and will be used an F1 score to evaluate the performance of different models that will be trained.