For this week, you will be required to submit the following:

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

Clearly define a problem or an idea of your choice. Remember that data science problems always target an audience and are meant to help a group of stakeholders solve a problem, so make sure that you explicitly describe your audience and why they would care about your problem.

1. A description of the data and how it will be used to solve the problem. (**15 marks)**

Describe the data that you will be using to solve the problem or execute your idea. So, make sure that you provide adequate explanation and discussion, with examples, of the data that you will be using.

**Business Understanding:** We want to be able to predict the severity of a road accidents under certain conditions like current weather, road and visibility. It makes it possible to alert the drivers for the current risks and also to better allocate resources, like police and medical emergency teams, according to the conditions that will make severe accidents more likely to happen in a certain area.

**Data understanding:** The example dataset to be used has 33 columns. The target variable will be 'SEVERITYCODE', as it is a measure of the severity of the accident and varies between 1 and 2.

The attributes 'WEATHER', 'ROADCOND' and 'LIGHTCOND' will be used to weigh the severity of an accident.

Values for SEVERITYCODE represent the following consequences of the accident:

* 1. Property Damage Only Collision
  2. Injury Collision

**Data Preparation:** The original dataset needs some preparation to be fit for analysis.

From the 33 columns of this dataset, many are not useful for this model. Many of the features need to be converted to numerical type, using label encoding, which will create new numerical columns.

4947 rows have all the three attributes in study in clear, so they will be removed and not considered for analysis.

10209 rows will be discarded from the analysis because all the three attributes are 'Unknown' and would be useless to train the model.

In the remaining 179517 rows we notice a great unbalance in the target variable:

SEVERITYCODE=1 122604 (68%)

SEVERITYCODE=2 56818 (32%)

This can be fixed by dropping the majority class in order to achieve a 50% representation of each class.

**Result section**: Describing your results and also any output supporting the results.

**Discussion section**: This is an important section as it shows all the inferences and analysis that you did on your project.

**Conclusion section:** Concluding statement of the project and also its future scope.

LOCATION – WEATHER – ROADCOND – LIGHTCOND – ST\_COLCODE (describes the collision) – SPEEDING - INCDTTM (date and time of the incident)

Try to get a most accidents area and or date

needs more attention from authorities? Road needs to be worked on?

Business model:

Is the number of accidents only limited because of the type of weather occurring or is it somehow occurring in a specific location? The idea