Data Description and Approach

The data being used has been gathered by the UK Department of Transport and it includes all traffic accidents during 2018 in the Greater Manchester area. Greater Manchester is a large metropolitan area in the north of England with an approximate population of 2.8 million and includes 10 boroughs: Bolton, Bury, Oldham, Rochdale, Stockport, Tameside, Trafford, Wigan and the cities of Manchester and Salford.

The dataset is publicly available on the following site: https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

The data includes more than 122,000 data points and more than 30 attributes such as:

- Accident data: location, number of vehicles involved, date
- Environmental data: light conditions, weather conditions
- Others: local authority, special conditions, police attendance

The dataset uses 3 different attributes to identify the severity of the accidents – Fatal (1), Serious (2) and Slight (3).

To solve the problem, firstly the plan is to understand the available data and the available attributes. Next, look for trends on which attributes are the most relevant to use. Afterwards, ensure the data is ready for modelling including data balancing, filling missing data and generally cleaning the dataset. Since the aim of the analysis is to predict and categorise accident severity, I will use a model using supervised machine learning techniques. Lastly, evaluate the model to ensure the business objectives are achieved using metrics such as recall, precision, and F1-score.