Assessment Proposal

Proposed Research Question:
How can a unified Random Forest model be employed to predict the severity of traffic accidents in central London, and what is the impact of accident timing and vehicle type on accident severity?

Background:
The study is to rely on data sets which are accessible to the public and are related to traffic incidents, the vehicles in those incidents, and the recorded casualties, these being obtained from official sources such as the Department for Transport. In terms of academic motivation, the research has been influenced by a study that Santos et al. (2022) presented, in which an extensive overview of machine learning methodologies used for determining how severe crash injuries might be was provided. Moreover, O'Toole et al. (2018) elaborates on deeper and more intricate aspects of data-related work such as preprocessing, feature modification, and adjustments in model parameters, has also been taken into consideration.

Purpose:
Traffic accidents, which are quite impactful, come with huge human and economic costs. These impacts are particularly noticeable in urban centers that are densely populated, like central London. The existing literature, which has looked into these matters, points out that Random Forest is one of those models that tend to do better than other models. Since we aim to predict the severity level (e.g., slight, serious, fatal) of traffic accidents, this research is framed as a multi-class classification problem. The Random Forest classifier will be employed to handle these discrete severity categories. In this study, the focus will be on central London because of the specific and complex nature of traffic there. The goal is to provide insights, targeted ones, into how the timing of accidents and the type of vehicle involved could influence the severity of the accidents. The results, which will come from this, are expected to help with urban traffic management decisions, how resources are allocated, and the development of safety measures that are more effective.

Methodologies:
In the methods section, there will be steps taken for integrating and preprocessing the data. This means merging the accident data from the DfT, along with casualty and vehicle data. The data will also be cleaned up, with missing values being dealt with. Geospatial association will be applied, especially focusing on the area of central London, as this is where the analysis will be concentrated.

1. Data Integration and Preprocessing: Merge and clean the DfT accident, casualty, and vehicle datasets, applying missing value treatment and geospatial association, specifically for central London.
2. Model Development: Employ a unified Random Forest model for multi-class classification, using grid search for hyperparameter tuning.
3. Model Evaluation and Interpretation: Evaluate model performance via confusion matrices, accuracy, and ROC curves, while using feature importance analysis to interpret the influence of accident timing and vehicle type on accident severity.

[1] Santos, K., Dias, J.P. and Amado, C., 2022. A literature review of machine learning algorithms for crash injury severity prediction. Journal of safety research, 80, pp.254-269.
[2] O'Toole, S.E. and Christie, N., 2018. Deprivation and road traffic injury comparisons for 4–10 and 11–15 year-olds. Journal of Transport & Health, 11, pp.221-229.