# CASA0006 Assessment: Body Text

Student Number: 22186878

**Background**

This notebook presents the development and evaluation of a machine learning model applied to a vehicle crash dataset, to predict crash injury severity and determine the most influential contributing factors.

**Literature Review**

Vehicle crash injuries present a major public health and economic burden on a global scale. In response to this complex challenge, there is an extensive body of research, developed over decades, that has applied statistical and machine learning models to determine contributing factors in crash likelihood and improve prediction of crash severity (Savolainen *et al.*, 2011; Santos, Dias and Amado, 2022). Understanding the association of specific factors with crash outcomes allows for the design of evidence-based countermeasures, whether in the vehicle or in the built environment, whilst predictive capacity has potential implications for directing emergency response, and in the decision making of autonomous vehicles.

In recent years, there has been a broad spectrum of ML methods that have rapidly become the most popular choice for crash severity analysis due to their freedom from prior assumptions, robustness to outliers, and high performance in big data scenarios (Ziakopoulos and Yannis, 2020). Established methods include regression and decision trees (Jeong *et al.*, 2018), support vector machines (Effati, Thill and Shabani, 2015), neural networks (Delen *et al.*, 2017), and ensemble approaches such as Random Forest and Gradient Boosting (Iranitalab and Khattak, 2017; Mafi, AbdelRazig and Doczy, 2018; Wang and Kim, 2019). Of these various approaches, a systematic literature review by Santos et al. (2022) found that Random Forest (RF) models proved the highest performing algorithm in the most cases, achieving best performance, comparative to alternatives, in 70% of the studies it was applied.

Irrespective of the approach used, there are certain characteristics of vehicle crash data that need to be accounted for in designing a model. Firstly, these datasets predominantly collect an ordinal categorical outcome – such as crash severity categorised into fatal, serious, minor, or non-injury. These classes are generally heavily imbalanced, with a very high proportion of observations in the low-severity class, and a low proportion classed as severe/fatal. If this imbalance is not addressed, the model may produce overtly promising results with a high overall accuracy, but have very poor predictive performance for the high-severity classes, which are inherently the observations of greatest interest in crash analysis (Jeong *et al.*, 2018). In addition, crash events will have spatial and temporal dependence, where crashes occurring close in time and space are likely to share effects not otherwise captured in the data, which may introduce bias in model estimates (Savolainen *et al.*, 2011). Therefore, …

**Research Question**

How effectively can road and environmental factors predict vehicle crash severity?

**Methodology**

Data for this analysis are drawn from the New Zealand Transport Agency’s Crash Analysis System (CAS), which contains a record of all police-reported traffic crashes on New Zealand roadways from 1st January 2000 (Waka Kotahi NZ Transport Agency, 2023).

**Results**

[In notebook]

**Discussion**

**References**:

Delen, D. *et al.* (2017) ‘Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods’, *Journal of Transport & Health*, 4, pp. 118–131. Available at: https://doi.org/10.1016/j.jth.2017.01.009.

Effati, M., Thill, J.-C. and Shabani, S. (2015) ‘Geospatial and machine learning techniques for wicked social science problems: analysis of crash severity on a regional highway corridor’, *Journal of Geographical Systems*, 17(2), pp. 107–135. Available at: https://doi.org/10.1007/s10109-015-0210-x.

Iranitalab, A. and Khattak, A. (2017) ‘Comparison of four statistical and machine learning methods for crash severity prediction’, *Accident Analysis & Prevention*, 108, pp. 27–36. Available at: https://doi.org/10.1016/j.aap.2017.08.008.

Jeong, H. *et al.* (2018) ‘Classification of motor vehicle crash injury severity: A hybrid approach for imbalanced data’, *Accident Analysis & Prevention*, 120, pp. 250–261. Available at: https://doi.org/10.1016/j.aap.2018.08.025.

Mafi, S., AbdelRazig, Y. and Doczy, R. (2018) ‘Machine Learning Methods to Analyze Injury Severity of Drivers from Different Age and Gender Groups’, *Transportation Research Record*, 2672(38), pp. 171–183. Available at: https://doi.org/10.1177/0361198118794292.

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. Available at: https://doi.org/10.1016/j.jsr.2021.12.007.

Savolainen, P.T. *et al.* (2011) ‘The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives’, *Accident Analysis & Prevention*, 43(5), pp. 1666–1676. Available at: https://doi.org/10.1016/j.aap.2011.03.025.

Waka Kotahi NZ Transport Agency (2023) *Crash Analysis System (CAS) Open Data*. Available at: https://opendata-nzta.opendata.arcgis.com/search?tags=CAS (Accessed: 20 April 2023).

Wang, X. and Kim, S.H. (2019) ‘Prediction and Factor Identification for Crash Severity: Comparison of Discrete Choice and Tree-Based Models’, *Transportation Research Record*, 2673(9), pp. 640–653. Available at: https://doi.org/10.1177/0361198119844456.

Ziakopoulos, A. and Yannis, G. (2020) ‘A review of spatial approaches in road safety’, *Accident Analysis & Prevention*, 135, p. 105323. Available at: https://doi.org/10.1016/j.aap.2019.105323.