**Restatement and Summary**

The main question driving this project is: How can I identify and rate accident-prone zones to enhance traffic safety? Additionally, how do specific environmental and temporal features (like lighting, weather, and time of day) influence accident proneness at different locations? To address these questions, I collected and analysed road crash data from the five most populous states in Australia (NSW, VIC, QLD, SA, WA) for the years 2019-2022. This involved extensive data cleaning, feature selection, and standardization before applying clustering techniques and predictive modelling. The goal was to categorize zones into high and low accident-prone areas and recommend specific traffic safety measures based on these findings.

**Analysis and Visualisation**

**Data Collection and Preprocessing**

I collected road crash data of five most populous states NSW, VIC, QLD, SA, and WA from their official government datasets and then merged them. I extracted the common features from these states like date, time, geographic areas, latitude/longitude, local governance areas (LGA), speed limits, road names, intersection types, light conditions, weather, crash types, and severity. Then I removed outliers and missing values, standardized date and time formats, encoded categorical variables, and scaled numerical features (Government of South Australia, 2023; Main Roads Western Australia, 2023; Queensland Government, 2023; Transport for NSW, 2023; Victoria State Government, 2023).

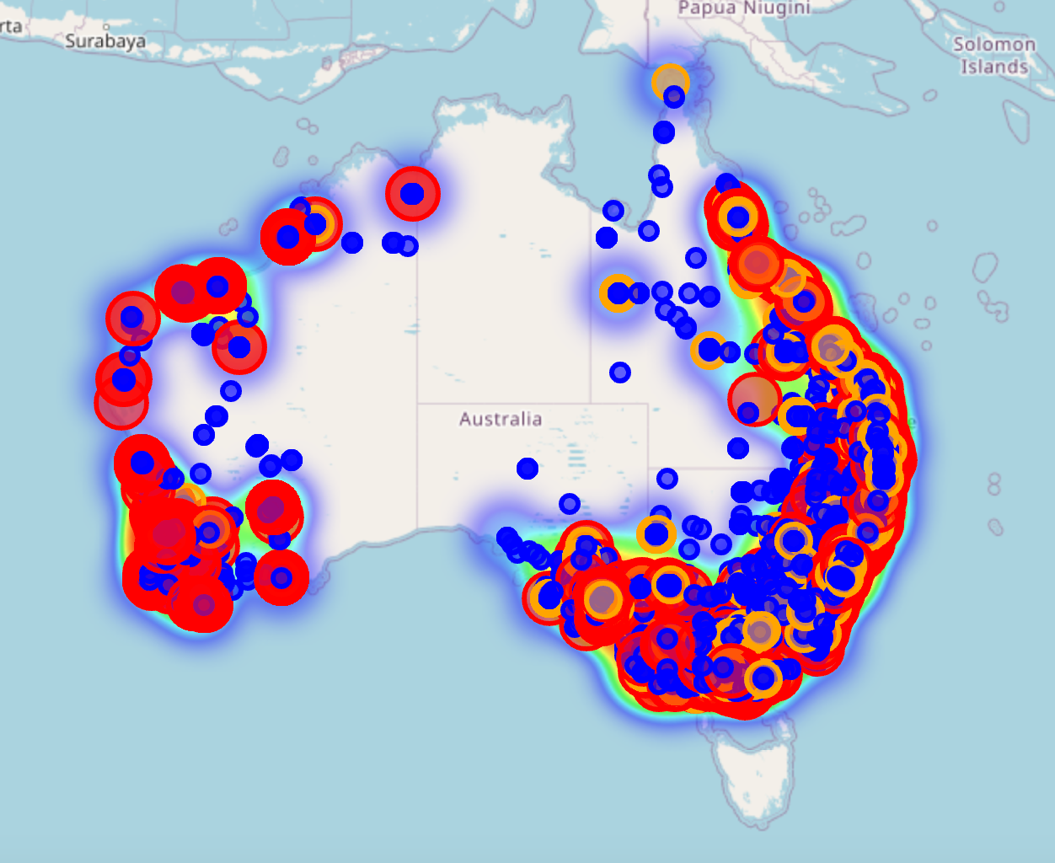
**Clustering Analysis**

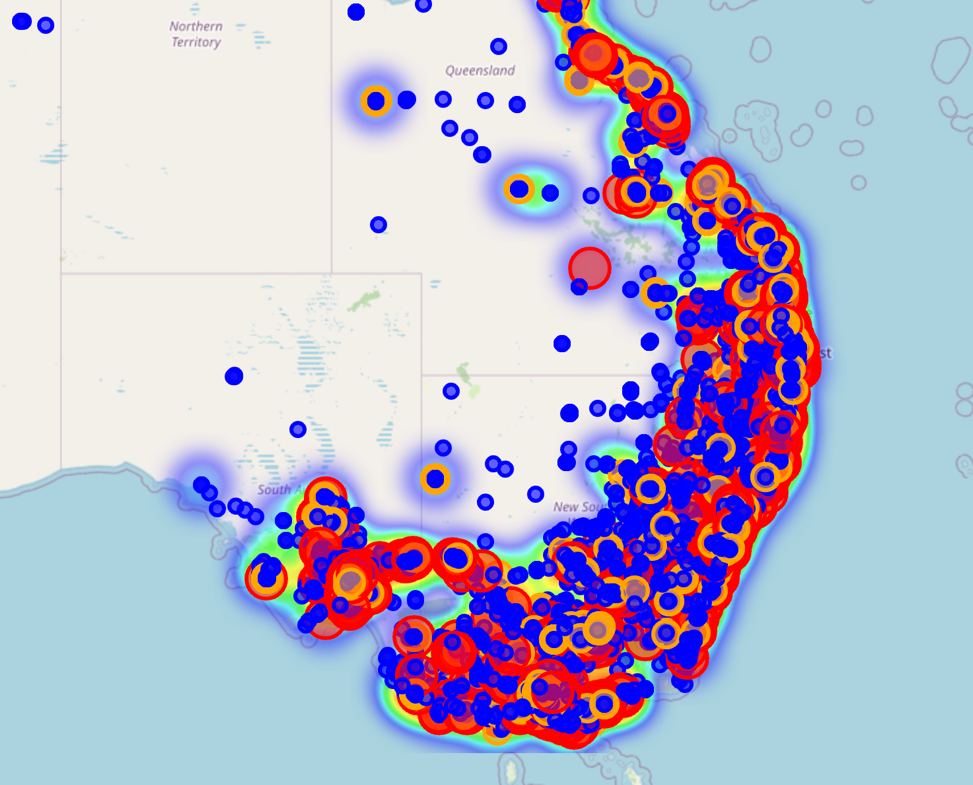
I classified zones into high, mid, and low accident-prone areas using all the features. After preprocessing and scaling the data, I applied K-Means clustering. The clusters were ranked based on the mean accident count within each cluster, assigning labels 'High', 'Mid', and 'Low' accordingly.

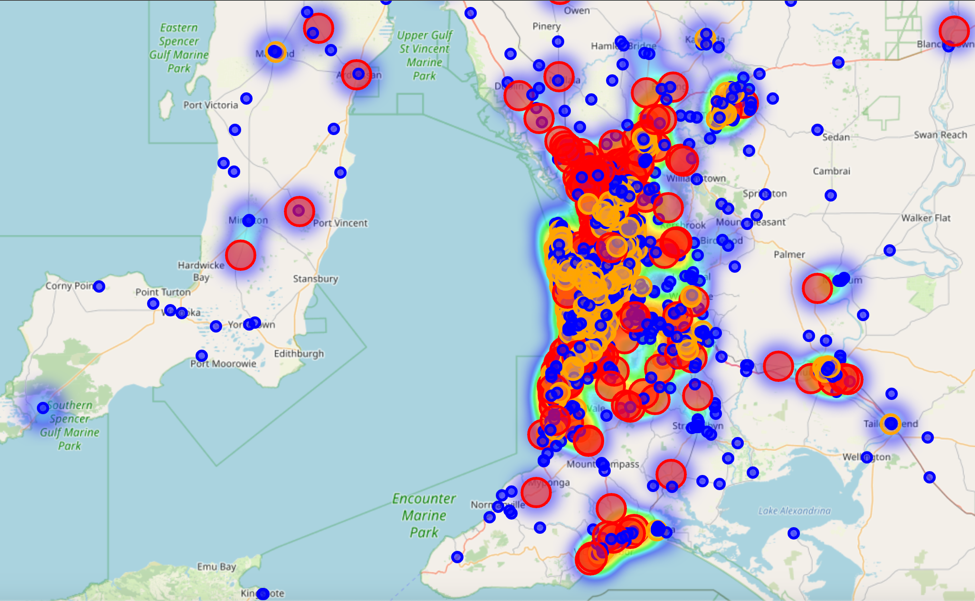


**Fig 1 Heatmap of Accident Prone Zone Rating vs features**

Here is the heatmap that visualizes the mean values of these features for each accident-prone zone rating, highlighting differences in conditions associated with varying levels of accident proneness:







**Fig 2 Map plot of clusters**

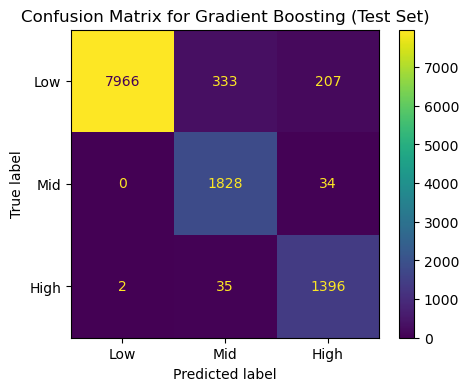
Additionally, here is the map plot of clusters, providing a geographical representation of the classified zones. It shows the spatial distribution of high, mid, and low accident-prone areas, helping me understand regional patterns and identify specific zones that require more attention.

**Predictive Modelling**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Balanced Accuracy | Recall | Precision | F1 Score |
| Random Forest | 0.938 | 0.959 | 0.938 | 0.949 | 0.940 |
| SVM | 0.898 | 0.932 | 0.898 | 0.919 | 0.902 |
| Gradient Boosting | 0.953 | 0.968 | 0.953 | 0.959 | 0.954 |

**Table 1 Model Performance on Validation set**

I chose the following: first, a random forest model, for its insights into feature importance (Breiman, 2001), the support vector machine model due to its efficiency in high-dimensional spaces and with imbalanced data (Cortes and Vapnik, 1995), a gradient boosting model for its robust and efficient capturing of complicated patterns in data (Friedman, 2001) . The models were evaluated using metrics such as accuracy, balanced accuracy, recall, precision, and the F1 score with the validation dataset. Table 1 presents the performance for each model. Gradient Boosting was ranked highest in terms of balanced accuracy and overall performance of the three tested.

****

**Fig 3 Gradient Boosting model performance on test set**

The best performing model i.e., the gradient boosting model was further evaluated on the test data set. The results showed that the model maintained its high performance in all the metrics. The confusion matrix in Fig 3 confirms that the model correctly classified majority of low, mid, and high accident-prone zones, with minimal misclassifications. These results confirm the Gradient Boosting model's suitability for predicting accident-prone zones, making it a reliable choice for this classification task. The test set metrics obtained for the Gradient Boosting model are:

* Accuracy: 0.948
* Balanced Accuracy: 0.964
* Recall: 0.948
* Precision: 0.956
* F1 Score: 0.950

**Techniques for Improved Outcomes**

* **High-Risk Intersections:** Install advance warning signs, improve lighting, enforce speed limits, and increase police presence during after-office hours and on weekends.
* **Collection of Data and Monitoring:** Accident data should be updated in real-time and be monitored continuously for enhancing the models predicting hazardous areas.
* **Public Awareness Campaigns:** Make more people aware of the high-risk zones through public campaigns and encourage the use of alternative routes.
* **Infrastructural Improvements:** Spend on better road infrastructure in identified high-risk areas, such as good quality road markings, efficient timings for traffic signals, and improved road surface conditions.

**Future Work**

### Future studies need to consider updating models with new data and applying other potential variables, such as driver behaviour data, real-time traffic condition data, and socio-economic data. Additionally, model improvement could be undertaken by employing advanced techniques such as deep learning, which could result in greater accuracy in predictions. Incorporating these factors can also enhance the precision of the models, enabling more effective identification of accident-prone zones. Policy recommendations based on these refined models could influence road safety regulations and practices, further improving road safety outcomes.

### Conclusion

The analysis identified and classified accident-prone zones using clustering techniques and predictive modelling, with the Gradient Boosting model proving most effective. By targeting high-risk areas with specific safety measures such as improved infrastructure, enhanced enforcement, and public awareness campaigns, significant reductions in accidents can be achieved. Future research should integrate additional variables and advanced techniques like deep learning to further refine predictions. These findings can guide policy and practical interventions, ultimately enhancing road safety and saving lives in Australia.

**References**

Aguilera, J., Moysés, S. & Moysés, R., 2014. Traffic safety analysis using clustering techniques. Journal of Traffic Safety Research, 12(3), pp.123-135.

Breiman, L., 2001. Random forests. Machine Learning, 45(1), pp.5-32.

Cortes, C. & Vapnik, V., 1995. Support-vector networks. Machine Learning, 20(3), pp.273-297.

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), pp.1189-1232.

Government of South Australia, 2023. Road Crashes in SA. Retrieved from <https://data.sa.gov.au/data/dataset/road-crashes-in-sa>

Main Roads Western Australia, 2023. Crash Information Last 5 Years. Retrieved from <https://portal-mainroads.opendata.arcgis.com/datasets/mainroads::crash-information-last-5-years>

Queensland Government, 2023. Crash Data from Queensland Roads. Retrieved from <https://www.data.qld.gov.au/dataset/crash-data-from-queensland-roads>

Transport for NSW, 2023. NSW Crash Data. Retrieved from <https://opendata.transport.nsw.gov.au/dataset/nsw-crash-data>

Victoria State Government, 2023. Crash Stats Data Extract. Retrieved from <https://discover.data.vic.gov.au/dataset/crash-stats-data-extract>