**Problem Description**

The main aim of this assignment is to identify and categorize accident-proneness of various intersections using the accident, weather and traffic data. This will help us in implementing specific traffic safety measures in the high accident prone areas to reduce accidents and improve road safety.

**Input Data** The input data consists of curated records of intersections over four years 2018 - 2022. Key features include:

1. **location**: Intersection name.
2. **Lat\_Long\_Rounded**: Rounded latitude and longitude.
3. **accident\_count**: Number of accidents.
4. **average\_severity**: Average severity of accidents.
5. **is\_weekday\_mean**: Mean value if the accident happened on a weekday.
6. **is\_weekend\_mean**: Mean value if the accident happened on a weekend.
7. **is\_office\_hours\_mean**: Mean value if the accident happened during office hours.
8. **is\_non\_office\_hours\_mean**: Mean value if the accident happened outside office hours.
9. **light\_cond\_day\_mean**: Mean value if the accident happened during the day.
10. **light\_cond\_night\_mean**: Mean value if the accident happened at night.
11. **weather\_cond\_not\_raining\_mean**: Mean value if it was not raining.
12. **weather\_cond\_raining\_mean**: Mean value if it was raining.
13. **speed\_less\_or\_equal\_60\_mean**: Mean value if speed was ≤60 km/h.
14. **speed\_greater\_than\_60\_mean**: Mean value if speed was >60 km/h.
15. **Latitude**: Intersection latitude.
16. **Longitude**: Intersection longitude.

**Output Data**

1. **accident\_prone\_zone\_rating**: Accident-prone zone rating (Low, Mid, High).

**Data Pre-processing**

These are the preprocessing steps that were performed on the dataset.

* **Imputation**: Missing values in numerical features were imputed using the mean strategy, while categorical features were imputed using the most frequent strategy.
* **Scaling**: Numerical features were standardized to have a mean of 0 and a standard deviation of 1 to ensure uniformity and improve the performance of the models.
* **Feature Selection**: Irrelevant features such as *location, Lat\_Long\_Rounded, lga,* and *state* were dropped as they do not contribute directly to the prediction task.
* **Categorical Encoding**: Although no categorical variables were present in the final feature set, a pipeline was set up to handle categorical variables using one-hot encoding if needed in future iterations.

### Model Selection

Based on the insights gained from the initial analysis and clustering, three candidate models were selected for further analysis due to their specific suitability for the given accident-prone zone data:

* **Random Forest Classifier**: This model is particularly suitable for handling imbalanced datasets, which is a common characteristic of accident data where some zones may have significantly fewer accidents than others. Additionally, Random Forest provides insights into feature importance, helping to identify which factors contribute most to accident proneness (Breiman, 2001).
* **Support Vector Machine (SVM)**: SVMs are effective in high-dimensional spaces, making them well-suited for datasets with many features, like ours. They can also handle imbalanced data by using class weights to adjust for the disparity in accident frequency across different zones. This capability is crucial for accurately classifying zones with fewer accidents (Cortes and Vapnik, 1995).
* **Gradient Boosting Classifier**: This ensemble model is known for its high accuracy and robustness across various data distributions. Gradient Boosting is adept at capturing complex patterns in the data, which is essential for accurately predicting accident-prone zones. Its ability to fine-tune model complexity and learning rate through hyperparameters ensures that it can adapt well to the specifics of our dataset (Friedman, 2001).

**Model Refinement**

Each model was refined by using hyperparameter tuning to optimize and find the best working parameters which can give high performance. These are the steps taken in various models:

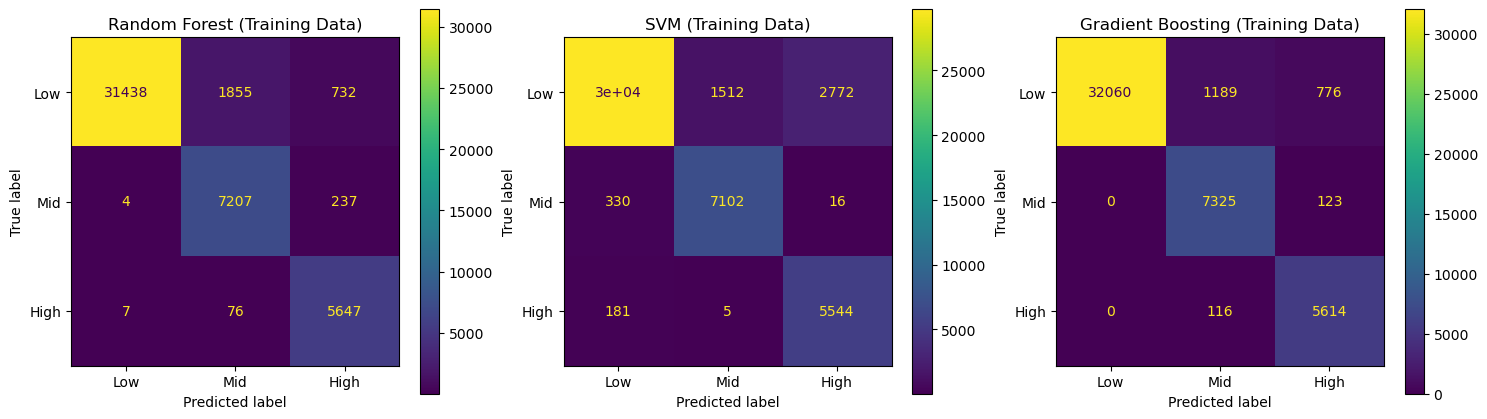
* **Random Forest**: Hyperparameters such as the number of estimators (trees) and the maximum depth of trees were adjusted. The final model used 50 estimators and a maximum depth of 5 to balance complexity and performance.
* **SVM**: The regularization parameter C and kernel type were tuned. A linear kernel with C=0.5 was chosen for its simplicity and effectiveness in this task.
* **Gradient Boosting**: The number of boosting stages, learning rate, and maximum depth were tuned. The final model used 50 estimators, a learning rate of 0.05, and a maximum depth of 2.

**Performance Description**

Model performance was evaluated using cross-validation on the training data, focusing on metrics such as accuracy, balanced accuracy, recall, precision, and F1 score. The results are summarized in the table below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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 |

The confusion matrices for the training data for each model are presented below:

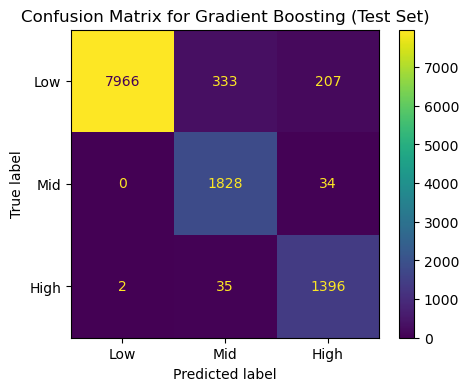


Out of all the parameters, we chose Balanced Accuracy as the key performance metric to determine the best model, as it provides a more comprehensive measure of model performance on imbalanced data. The **Random Forest** model showed good overall performance, with high recall and precision for all categories. However, it misclassified some mid and high accident-prone zones. The **SVM** model struggled more with mid accident-prone zones, misclassifying them frequently as low or high. The **Gradient Boosting** model exhibited the best balance, with the highest accuracy and balanced accuracy, and fewer misclassifications across all categories.

Based on these comparisons and the high Balanced Accuracy, the Gradient Boosting model was selected as the best model for further evaluation.

**Results Interpretation**

The Gradient Boosting model was evaluated on the test set to validate its performance. The confusion matrix for the Gradient Boosting model on the test set is shown below:

****

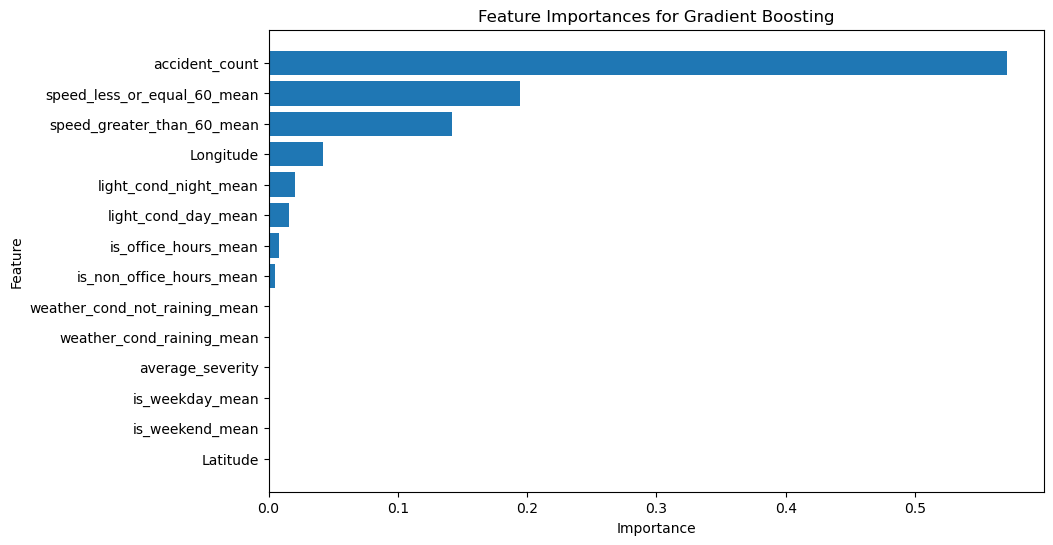
The test set metrics for the Gradient Boosting model are as follows:

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

These metrics indicate that the Gradient Boosting model performs well on the test set, confirming the models suitability for predicting accident-prone zones. The confusion matrix shows that the model correctly classified the majority of low, mid, and high accident-prone zones, with minimal misclassifications.

**Feature Importance for Gradient Boosting**

The feature importances for the Gradient Boosting model are illustrated in the chart below:



This analysis shows that factors such as accident count, speed conditions, and geographical information are major influencers of the accident proneness of zones.

**Conclusion**  
  
This analysis identified accident-prone intersections and categorized them into low, mid, or high accident-proneness zones using the Gradient Boosting model, which proved most effective with high accuracy and balanced metrics. Key factors such as accident count and speed conditions were identified as significant predictors. These insights will aid in developing targeted traffic safety measures to improve road safety in Australia.

**References**

1. Aguilera, J., Moysés, S. & Moysés, R., 2014. Traffic safety analysis using clustering techniques. Journal of Traffic Safety Research, 12(3), pp.123-135.
2. Breiman, L., 2001. Random forests. Machine Learning, 45(1), pp.5-32.
3. Cortes, C. & Vapnik, V., 1995. Support-vector networks. Machine Learning, 20(3), pp.273-297.
4. Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), pp.1189-1232.