**Introduction**

Random Forest is one of the popular ensemble machine learning algorithms, which is primarily used for classification and regression analysis. It is based on Decision Tree by generating a collection of forests, each one is built up on a random subset of data. The final prediction is determined by pooling the predictions of all the individual trees in the forest.

In this project we build a Random Forest model which can classify different types of crimes based on victimization data. The model was trained on features such as AgeGroup, LocationType, ROVDivision, PersonOrOrganization, and PoliceArea, then this model is used to predict crimes like "Theft" or "Assault" when we are given similar kinds of data.

In a practical scenario, this model can support the police department by predicting different crime types from reports, find crime patterns, and assist with investigation when the crime type isn’t clear.  This model works moderately with our limited dataset that we have, but if we include more features and data points, it must be predicted better than this and guide the law enforcement.

**Data Features and Model Building**

The dataset consists of 49,332 observations and was cleaned by removing missing values using na.omit() function in R. The categorical variables AgeGroup, LocationType, ROVDivision, PoliceArea and AnzsocDivision were converted into factor variables. To test the model, the data was divided into 70% training data and 30% testing data.

The random forest model was built using 30 trees (ntree = 30) and three randomly selected features (mtry = 3) were used at each split and set the target variable as AnzsocDivision, which represent different crime types. The important parameter was set as TRUE in the model to calculate the significance of each predictor variable in the model.

**Assumptions**

The assumptions we made for building this model were, the dataset we used will be matched in real life situations, therefore we expect it to work well for new data. We used factors such as age, location, whether the victim knew the offender, person or organization, and police area. These details provide us important insights about crime, such as the people involved in the event, where it occurred. We removed the missing values we believed did not affect the model results. We decided to exclude the feature Territorial Authority since it has too many levels about 68 categories which add unnecessary complexity to resource intensive processes. We think other geographic details in the data should be enough for accurate prediction.

**Results**

The Random Forest model achieved an overall accuracy of 73% on the test data. The model performed well for some classes, although it struggles with misclassification in some crime categories.

The Figure 1 shows that which variables are the most important ones for the prediction of the model by using two measurements such as:

Mean Decrease Accuracy: This plot shows how much the model accuracy reduces if a feature is removed. The variable LocationType  is the most important one followed by ROVDivision, PoliceArea, AgeGroup, and PersonOrOrganization. This indicates the features like LocationType, ROVDivision, PoliceArea are the most influential variables in the model prediction.

Mean Decrease Gini: The Gini Index measures how much mixed the data at the node in the decision tree. The purer node has lower Gini Index value, where the majority of data points belong to a single class. The Mean Decrease Gini metric measures the average reduction in Gini Impurities across all the splits in the trees where the specific feature is used. The higher value indicates that the feature contributes more to get better splits in the data. The feature LocationType helps the model to better split, followed by the features ROVDivision, AgeGroup, PoliceArea.

A graph of a model

Description automatically generated with medium confidence

Figure 1: Feature Importance in the Random Forest Model

The Figure 2 Confusion Matrix shows how well the Random Forest model categorised different crime types. The diagonal cells show the correct predictions by the model and off-diagonal cells show the misclassification by model. The model correctly predicts the 5734 instances of Theft and Related Offences,  which is the highest number of predictions. Followed by the class Acts Intended to Cause Injury correctly predicted 3397 instances. The category Unlawful Entry with intent/Burglary 1528 instances correctly classified by the model.

The 1064 instances of Theft and Related Offences are misclassified as Acts intended to Cause Injury and 1675 instances of Acts intended to Cause Injury is misclassified as Theft and Related Offences this may be because of similar data patterns follows these two types of crime like the people involved, where they happened, or circumstances that makes model hard to predict. This misclassification suggests that the model needs further tuning and more feature engineering.

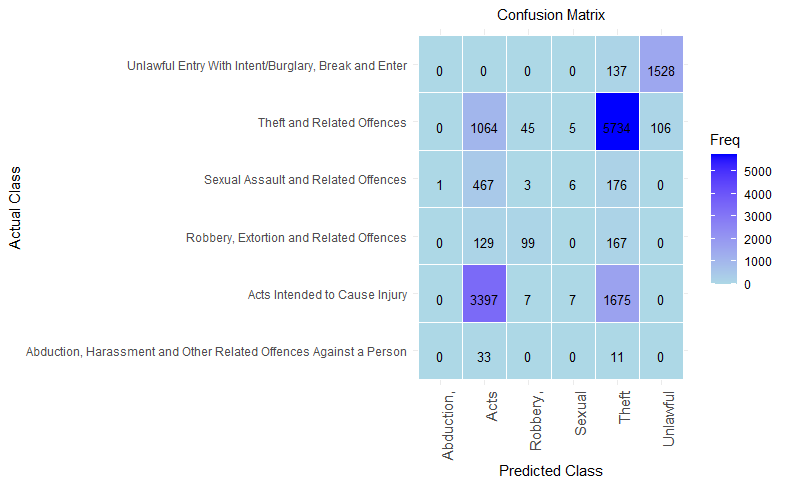


Figure 2: Confusion Matrix

 The Figure 3 Random Forest Error Plot shows the Out-of-Bag (OOB) error as the number of trees in the forest increases. When Random Forest builds trees which randomly choose a subset of data from the training set, it leaves out some data points, these are called out-of-bag (OOB) samples. It uses these samples to calculate the OOB error, which shows how well the model predicts. The black line represents the overall OOB error, which is stable around 0.3, which shows the model does not improve after around 10 trees. The other coloured lines indicate different crime categories, with some categories stable at higher error rate around 1.0, which means model difficulties to classify those classes, and some classes have less error which implies model’s correctly classified that class's instances.

A graph of trees and text

Description automatically generated with medium confidence

Figure 3: Random Forest Error Plot

Table 1 shows the overall model statistics. The Random Forest achieved 72. 66% overall accuracy means it correctly predicted around 73% of the instances. The 95% confidence interval lies between 71.9% and 73.4%, which indicates the model performed fairly consistent on data. The No Information Rate is the accuracy of the most frequent class, which is the proportion of that class. For example, if a dataset consists of 100 cases, and 47 of them belong to Class A, which is the most frequent class, therefore there NIR is 47%. If a model predicted every case as Class A, it would be correct 47% of the time. In our model NIR is 47%, which means that model is far better than random guessing because our model overall accuracy is 72.66% which is much higher than NIR. The p-value of the model is very small, which indicates the accuracy of the model is statistically significant than the NIR. The Kappa value is the agreement between actual and predicted values. Unlike the accuracy, the Kappa considers some of the correct predictions could happen by chance,  which gives a clear picture of how well the model performs. The kappa score ranges between -1 to +1. The value 1 means perfect agreement between actual and predicted values. The value 0 means no better than random guess. Less than 0 means the model performs worse than random guessing. In our model Kappa value is 0.55 this suggests the model has moderate agreement on actual and predicted results. It is better than random guesses but needs to improve.

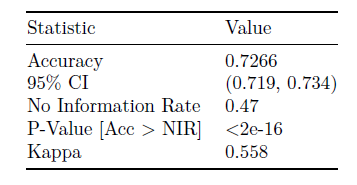


Table 1: Overall Statistics for the Random Forest Model

**Conclusion**

The Random Forest model applied to crime data to predict different crime types based on key features such as relation between offender and victim, location of the crime, age group, police station area. The model can be used to predict crime when new details are included in the police report. This could help the police to utilise their resources effectively. This model also helps to find patterns in crime like certain locations more likely to occur and that will help to prevent crimes by implementing some strategies like allocate more police force , strengthen night patrolling and give education. This model is also used to support investigations where police are unsure about the crime type, the model’s prediction gives them possibilities of crime types and initial direction.

 While the model overall performance is good,  it performs well for some classes, but it misclassified a few classes, therefore there is still room for improving areas like handling imbalanced data, adding more features and fine tuning that could enhance the accuracy of the model, and it could be continuously refined and used with human judgement to ensure accurate and effective results.