**A picture containing text, font, logo, graphics

Description automatically generated**  
**Machine Learning and Data Visualization – MAT 3120.3**

**Assignment 2: Machine Learning Modelling**

**Name: \*\*\*\*\*\*\*\*\*\*\*\*\***

**Student ID: \*\*\*\*\*\*\*\*\*\*\*\***

Contents

[Introduction 3](#_Toc162963433)

[Data Preparation 3](#_Toc162963434)

[Machine Learning Models 3](#_Toc162963435)

[Result & Discussion 4](#_Toc162963436)

[Conclusion and Recommendations 6](#_Toc162963437)

## Introduction

This project aims to develop machine learning models for accurate incident detection, focusing on classifying malicious events using R code. Two machine learning models that are trained are logistic lasso regression and classification tree models. The hyper parameters of the models are fine-tuned, and their performances are assessed on the unseen test set. Evaluation of the performance includes specificity, sensitivity, recall, precision, F1 score, and accuracy are considered. The document is structured and organized into sections covering data preparation, machine learning models, results and discussions, and conclusion with recommendations.

## Data Preparation

Several data preparation steps are implemented to clean and preprocess the data for analysis:

* Filtering the data to retain observations with class values of either 0 or 1. Instance where the class value was null or negative are \*\*\*\*\*\*\*\*\*\* to ensure accurate class labels for \*\*\*\*\*\*\*\*\*\*  
  \*\*\*\*\*\*\*. Having accurate class labels is a necessary step to \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
  \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*.
* Cleaning the \*\*\*\*\*\*\*\*\* column by merging \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* into one master category. Invalid values \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* are replaced with \*\*\*\*\*\*\*\*. Similar cleaning is also applied to the \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* column where the categories like \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*.
* Further filtering to remove the observations where \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

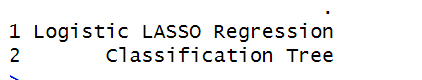
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*.

* Removing missing values from the dataset to retain only complete cases, resulting \*\*\*\*\*\*\*\*\*\*\*  
  \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*.
* Resolving any data balancing issues and splitting the dataset into training and testing sets. Additionally, two training sets are generated which includes the training sets with balanced training samples and a training set with unbalanced training samples.

## Machine Learning Models

In this assignment, two machine learning models are trained are the logistic lasso regression model and a classification tree model. Initially, both the models are trained with default parameters using both balanced and unbalanced training datasets. Subsequently, 5 folds cross validations is applied to fine tune the hyperparameters of both models.

For the logistic lasso regression model, the hyperparameters tuned is the \*\*\*\*\* value, as it is the sole parameter available for tuning in this context. Meanwhile, for the classification tree model, the \*\*\*\*\*\*\* value is tuned. The range of the cp values explored spans from \*\*\*\*\*\*\*\*\* to \*\*\*\*\*\*\*\*\*, with a step size of \*\*\*\*\*\*\*\*\*\*\*. Similarly, the \*\*\*\*\*\*\*\*\*\* grid ranges from \*\*\*\*\*\*\*\* to \*\*\*\*\*\*\*\* with a step size of \*\*\*\*\*\*\*\*.



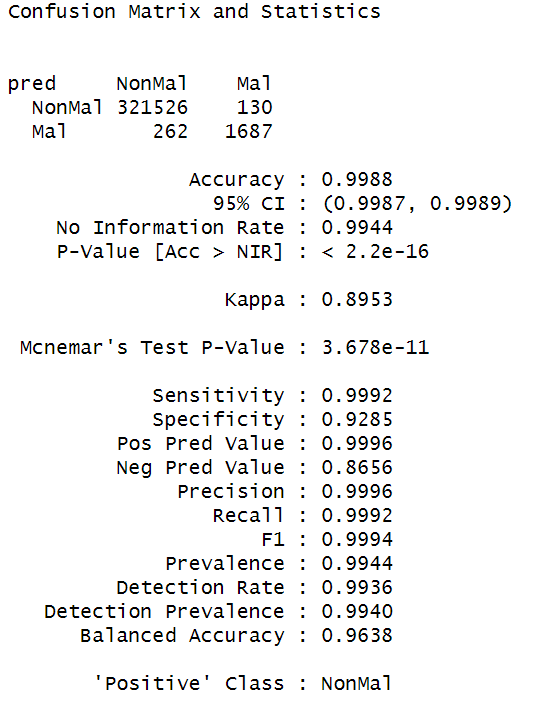
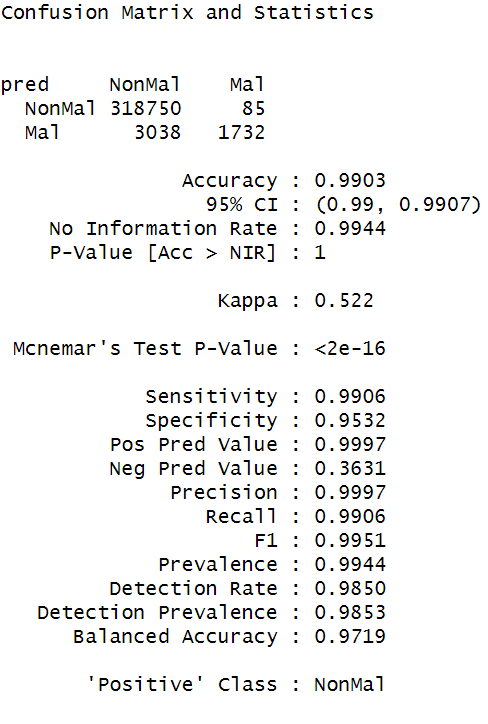
**Figure 1: Machine Learning Models Trained**

## Result & Discussion

The evaluation of tuned decision tree models, trained on two distinct training datasets, is conducted on the test set, yielding the following results.

For the model trained on the training dataset with balanced samples, it achieves an accuracy of \*\*\*\*\*\*\*\* on the test set. Although the model misclassifies a few observations, the 95% confidence interval (CI) indicates that the accuracy lies between \*\*\*\*\*\*\*\* & and \*\*\*\*\*\*\*\*. The kappa statistic for this model stands at \*\*\*\*\*\*\*\* which is good indicating a commendable performance. While sensitivity, precision, recall, and F1 score all exceed \*\*\*\*\*\*\*\*, its notable that the specificity \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*.

On the other hand, the model that is trained on the training dataset with unbalanced samples exhibits an accuracy of \*\*\*\*\*\*\*\* on the test set. Similarly, a small number of observations are misclassified. The 95% CI suggests, and accuracy of the model lies between \*\*\*\*\*\*\*\* and \*\*\*\*\*\*\*\* . The kappa statistic for this model reaches \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*, indicating \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*. Despite sensitivity, precision, recall, and F1 score all \*\*\*\*\*\*\*\*, the model’s \*\*\*\*\*\*\*\* remains lower compared to the \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*.

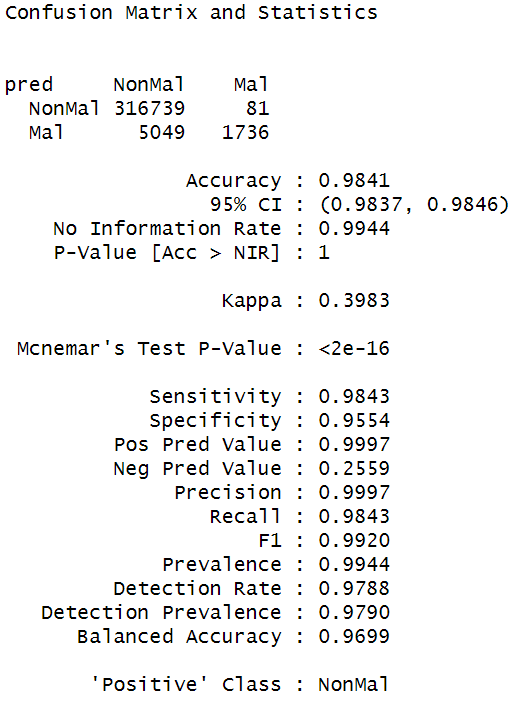


**Figure 2: Performance of classification tree model Figure 3: Performance of classification tree model**

**trained on balanced training set trained on unbalanced training set**

Moving onto the logistic regression model, the model is trained on the balanced training dataset achieving an accuracy of \*\*\*\*\*\*\*\* on the test set. The 95% CI indicates that the accuracy lies between \*\*\*\*\*\*\*\* and \*\*\*\*\*\*\*\*. The kappa statistic for this model is \*\*\*\*\*\*\*\*, reflecting a satisfactory performance. The sensitivity, precision, recall, and F1 score are \*\*\*\*\*\*\*\*, however the \*\*\*\*\*\*\*\*.

In contrast, the model trained on the unbalanced dataset achieves an accuracy of \*\*\*\*\*\*\*\* on the test set. The 95% of CI suggests an accuracy range of the model between \*\*\*\*\*\*\*\* and \*\*\*\*\*\*\*\*. The kappa statistic for this model reaches \*\*\*\*\*\*\*\*, indicating that it is good. Similarly, sensitivity, precision, recall, and F1 score are \*\*\*\*\*\*\*\*. However, the specificity of this model \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*. Specifically, the specificity obtained using the model trained on balanced samples is \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*.

 A screenshot of a computer

Description automatically generated

**Figure 4: Performance of Logistic Lasso Figure 5: Performance of Logistic Lasso**

**Regression model trained on balanced training set Regression model trained on unbalanced training set**

## Conclusion and Recommendations

Based on the results obtained from training the models, fine tuning their parameters, and evaluating their performances on the test set, the decision tree model emerges as the superior classifier compared to the logistic lasso regression model. It exhibits higher \*\*\*\*\*\*\*\*, \*\*\*\*\*\*\*\*, \*\*\*\*\*\*\*\*, \*\*\*\*\*\*\*\*, and \*\*\*\*\*\*\*\* across \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*.

Among the classification tree models trained on \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*, the unbalanced dataset model \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*. However, it is noted that the \*\*\*\*\*\*\*\* of the unbalanced dataset model is \*\*\*\*\*\*\*\* \*\*\*\*\*\*\*\* \*\*\*\*\*\*\*\* \*\*\*\*\*\*\*\* compared to that of the balanced dataset model.

Therefore, despite the \*\*\*\*\*\*\*\* overall performance of the unbalanced dataset model, the recommended choice is \*\*\*\*\*\*\*\* \*\*\*\*\*\*\*\* model trained on the \*\*\*\*\*\*\*\* dataset, as it provides a better balance between performance metrics, including specificity.