COMP3217 – CPS Coursework

# Part A

## Problem

The problem for this task is to develop a machine learning model that can accurately classify system traces into two categories: normal events (labelled as ‘0’) and abnormal events (labelled as ‘1’). The training data consists of 6000 labelled system traces, with each trace containing 128 features representing measurements from 4 phasor measurements units (PMUs) and additions columns for the control panel logs, snort alerts, and relay logs. The 129th column is the label of each of the events. Figure 1 shows the configuration of the power system framework that is considered for this task.

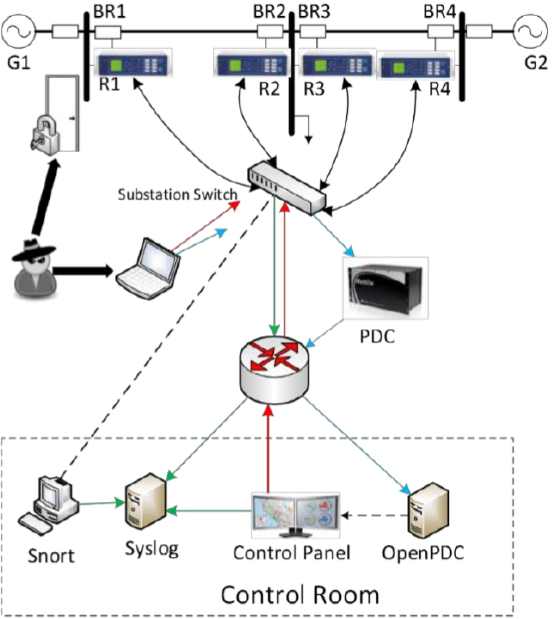


Figure : The configuration of the power system framework in question

The various components in Figure 1 are as follows:

* G1 and G2 are power generators
* R1 through R4 are Intelligent Electronic Devices (IEDs) that can switch the breakers on or off
* BR1 through BR4 are the breakers

There are also two lines:

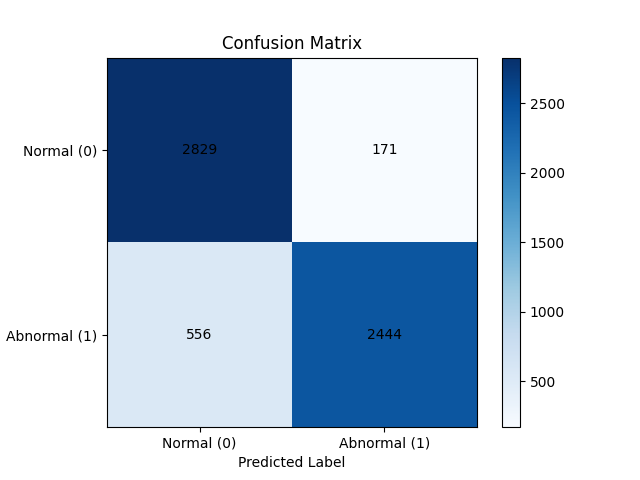
* Line One spans from breaker one (BR1) to breaker two (BR2)
* Line Two spans from breaker three (BR3) to breaker four (BR4)

The goal is to use this training data to train a machine learning model that can predict the labels for 100 unlabelled system traces in the testing data. The predicted labels for the testing data need to be outputted in a file named "TestingResultsBinary.csv" with the same format as the training and testing data. The final machine learning model that has been chosen is the Decision Tree Classifier.

## Results

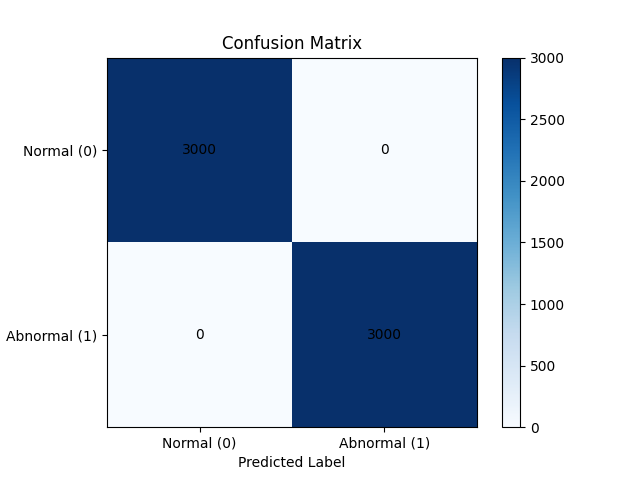
My initial implementation was a logistic regression model using a 5-fold cross-validation method. I chose this value of cross-validation because it gave the ideal combination of speed and accuracy. This implementation provided the following results:

* Cross-Validation Scores: [0.88242153, 0.84580924, 0.82083022, 0.86144498, 0.74269521]
* Average Cross-Validation Score: 0.830640236
* F1 Score: 0.8705253784505788 = 0.871 (3 significant figures)
* Confusion Matrix:



I was unsatisfied with the accuracy of my implementation, so I decided to change the machine learning model to a Decision Tree Classifier model. This implementation gave the following results:

* Cross-Validation Scores: [0.90748143, 0.8024107, 0.83874174, 0.9082209, 0.83455836]
* Average Cross-Validation Score: 0.858282626
* F1 Score: 1.0
* Confusion Matrix:



## Analysis

Possible training errors in my chosen implementation could include overfitting whereby the model becomes too specialised to the training data and then fails to generalise to unseen data, resulting in poor performance on the testing data and therefore lower accuracy. This can be solved by implementing a data cleaning feature. Another plausible training error can be the lack of Hyperparameter tuning which was solved by implementing cross-validation methods. The cross-validation technique provides a more reliable estimate of the model’s performance, reduces the risk of overfitting, and improves generalisation to any unseen data by selecting the relevant hyperparameters.

To determine the accuracy of my mode, I had added an F1 score and a confusion matrix. The F1 score provides a single value that represents the model's accuracy by considering both false positives and false negatives. By combining these features, the F1 score is a useful measure for evaluating a model’s accuracy, especially when dealing with possible unbalanced datasets. A confusion matrix is a table that provides a detailed breakdown of the model's predictions compared to the actual classes or labels of the data. It displays the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The confusion matrix helps in determining the accuracy of a model by providing insights into different types of errors made by the model. In summary, the confusion matrix provides a detailed breakdown of predictions and actual classes, allowing for a deeper analysis of the model's performance.

# Part B

## Problem

This section is like that in Part A in terms of the core concept of the problem. The problem for this task is to develop a machine learning model that can accurately classify system traces into three categories: normal events (labelled as ‘0’), data injection attack events (labelled as ‘1’), and command injection attack events (labelled as ‘2’). The training data consists of 6000 labelled system traces, with each trace containing 128 features and the 129th column is the label of each of the events. The predicted labels for the testing data need to be outputted in a file named "TestingResultsMulti.csv" with the same format as the training and testing data. The final machine learning model that has been chosen is the Decision Tree Classifier.

## Results

The initial implementation was a logistic regression model. The results were as follows:

* Cross-Validation Scores: [0.62453492, 0.48173773, 0.55829501, 0.58854752, 0.52553884]
* Average Cross-Validation Score: 0.555730804
* F1 Score: 0.6430273118300941
* Confusion Matrix:

A picture containing screenshot, diagram, text, rectangle

Description automatically generated

The finalised implementation was a Decision Tree model. The results were as follows:

* Cross-Validation Scores: [0.55304954, 0.50665427, 0.49712888, 0.5135223, 0.5359259]
* Average Cross-Validation Score: 0.521256178
* F1 Score: 1.0
* Confusion Matrix:

A blue and white squares

Description automatically generated with low confidence

## Analysis