CMP417 – Engineering Resilient Systems

Machine Learning and Artificial Intelligence

Paul Michael Oates

2001642

25/03/2024

Contents

1. Introduction

Scottish Glen a small company in the energy sector have come under threat after comments made by their CEO. The IT department have been alerted by employees to threats made by a hacktivist group. Concerned about the data they have detected on their network; they require a Machine Learning (ML) classifier to categorise the network packet data according to their attack category. The aim of this report is to discuss and critically evaluate ML classifier algorithms, Random Forest and K-Means Clustering and to implement the most effective solution into their Intrusion Detection System (IDS).

An Intrusion Detection System or IDS is a way for the security team to detect malicious traffic on the monitored network. The IDS system that Scottish Glen wishes to implement consists of the following categories to help understand the traffic that goes across the network.

Malicious

- Reconnaissance
- Backdoor
- Denial of Service
- Exploits
- Analysis
- Fuzzers
- Worms
- Shellcode

Non-Malicious

- Normal
- Generic

These classifications will enable Scottish Glens IT team to have a better understanding of traffic on their network and mitigate attacks against it.

2. Machine learning algorithms

A Machine Learning (ML) algorithm is a set of rules, or a processes used by Artificial Intelligence with the aim to discover patterns, insights or potentially predict output values from a given input data. (IBM, ND)

Four distinct types of ML algorithms are: supervised, unsupervised, semi-supervised, and reinforcement. (IBM, ND) Each distinct type of ML algorithm has its own benefits and can be used for specialised tasks such as image identification or natural language processing. This enables engineers to develop tools for a wide range of purposes.

K-Means is an unsupervised ML algorithm which groups similar items together. Whilst Random Forest is a supervised ML algorithm that is a collaboration of decision trees that work together to provide a single output. (susmit_sekhar_bhakta, 2024)

2.1. K-Means Clustering

The K-Means algorithm aims to cluster data of equal variance. These clusters are identified by the programmer. The algorithm does this by dividing a set of N samples X into K disjoint clusters C, each are described by the mean μ (scikit-learn.org, ND). To explain if you were to throw

clothes on the floor then group related items near each other you have created clusters. This is a simple representation of how the algorithm works. The equation to calculate the K-Means algorithm is described below in Figure 1.

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

Figure 1- K-Means Algorithm (scikit-learn.org, ND)

The end goal of a K-Means algorithm is to divide a population set into clusters to compare and analyse. K-Means does this by grouping similar data together and repeating this process until the clusters stabilise. The following Figure 2 identifies an example of random numbers being classified in this manner.

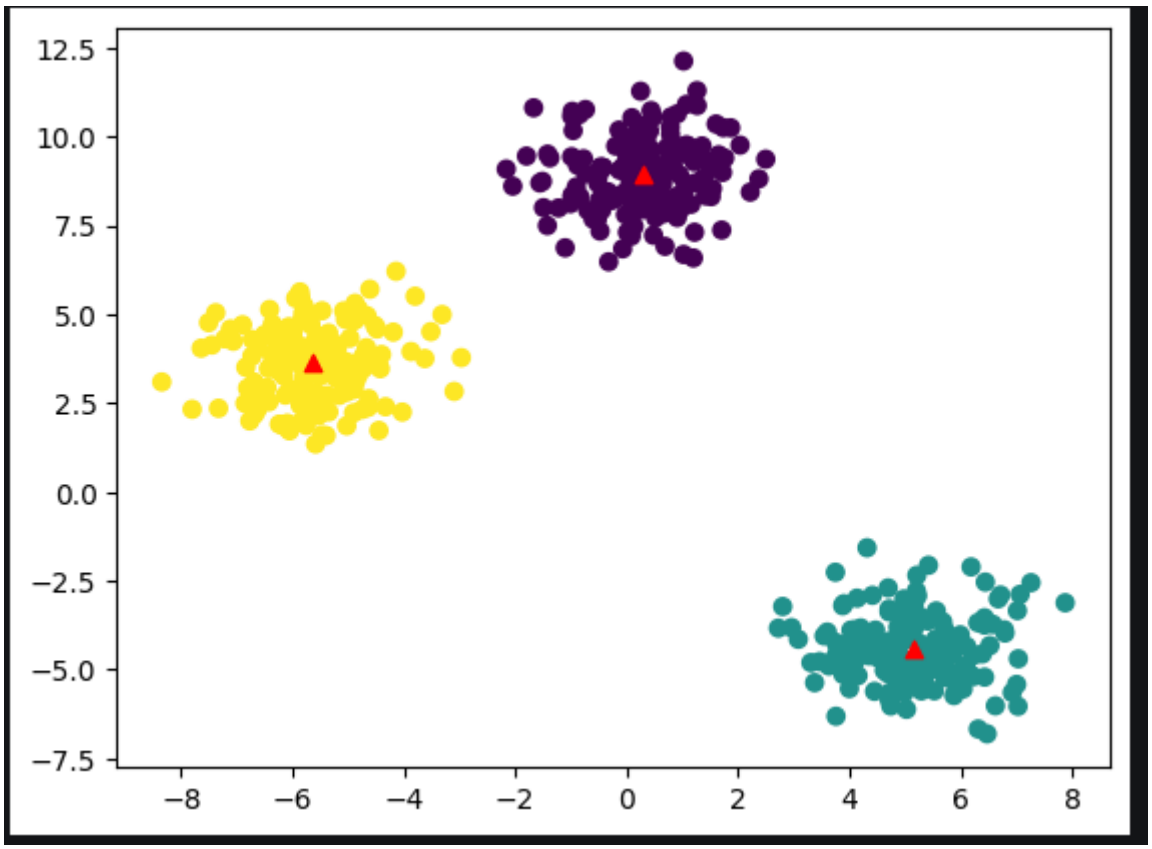

Figure 2 K-Means Clustering dataset (Geeks-For-Geeks, ND)

In relation to Scottish Glens network traffic this means that the algorithm would be able to group together the network data into the various predefined attack groups. Some advantages of using K-Means include its simple integration into the IDS system and its ability to scale to larger data sets meaning it would be ideal for monitoring large volumes of network traffic. K-Means ability to guarantee convergence around the predefined attacks would guarantee that similar exploits would be detected.

However, if the volume and variability of the training data is not sufficient it can impact the algorithm's ability to detect attacks. Another limitation of the K-Means algorithm is that outliers

in the training data such as rouge packets which could skew the cluster centres, making the IDS system less accurate. (Google Developer, 2022).

2.2. Random Forest

Random Forest is a ML learning algorithm that works by creating a number of decision trees during its training phase. When used to classify an item the algorithm aggregates the results of the decision trees enabling a collaborative decision-making process to occur leading to a predicted output. The collaborative output that occurs leads to more stable and accurate results than using a single decision tree. (Geeks-For-Geeks, 2024)

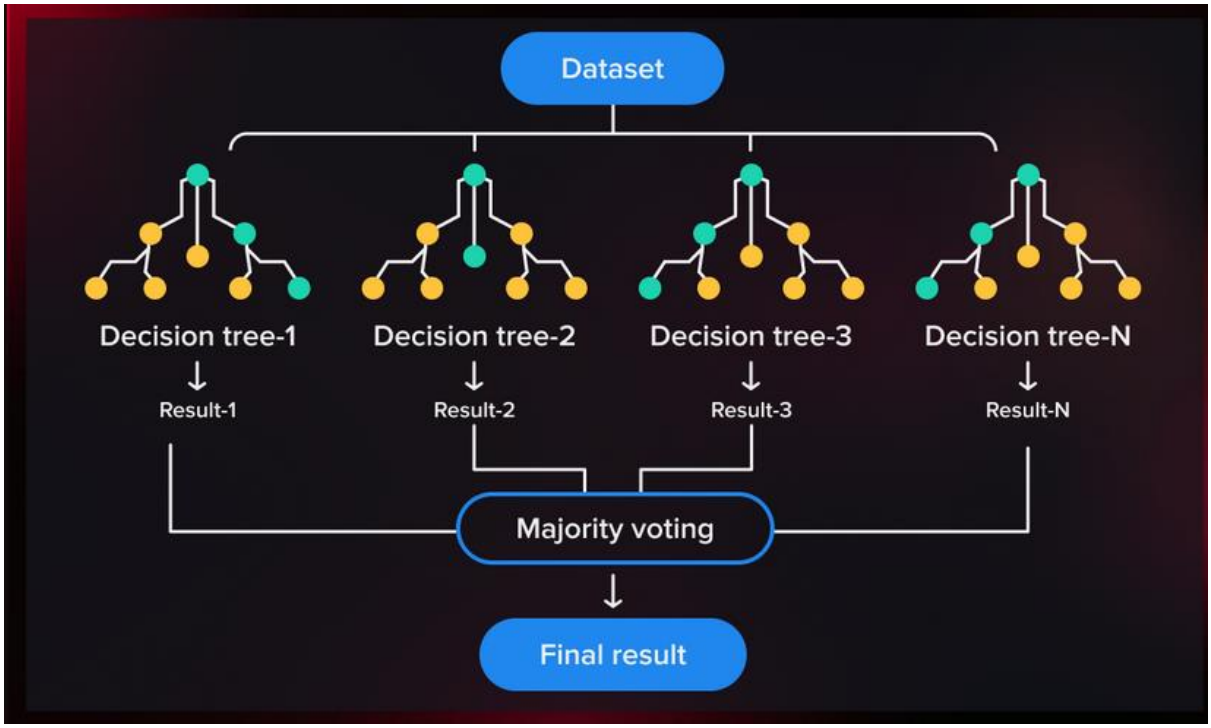

Figure 3- Random forest classification (Logunova, 2021)

Decision trees are a supervised method used for classification and regression. Their end goal is to predict the value by following simple decision rules similar to a flow chart. The tree is made up of decision nodes which help guide the algorithm to its decision as demonstrated in Figure 4:

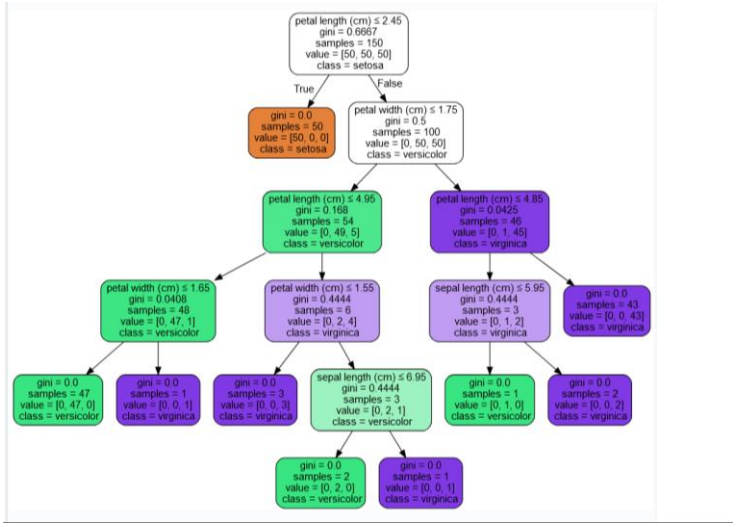

Figure 4- Decision tree which decides which kind of flower it is presented with (scikit-learn, ND))

When multiple decision trees are included within a Random Forest algorithm an extremely high level of accuracy is achieved. The Random Forest algorithm would be able to classify network traffic into several predefined categories. It is less sensitive to outliers and noisy data than K-Means and other algorithms because it averages the predictions of multiple decision trees (Sellahewa, 2023).

The Random Forest algorithm can also employ “bagging” or “bootstrap aggregation” to enhance model robustness. “Bagging” is the process of sampling the training data to creating multiple training sets with slight variations. These variations help to train the separate decisions trees which will employ majority voting to determine the outcome. This method enables a high level of accuracy within the Random Forrest classification algorithm. (IBM, ND) However, due to it containing multiple decision trees the algorithm would use substantial amounts of computer memory.

Another advantage of the Random Forest algorithm is it reduces the risk of “overfitting”, a common issue that occurs when the model is trained with so much data, it starts learning from the noise and inaccurate data entries in the training data. (Geeks-For-Geeks, 2024) A limitation of the Random Forest algorithm is that it is typically slower than K-Means at classifying data.

2.3. Conclusion

In conclusion, the Random Forest Algorithm will be selected over the K-Means algorithm due to its extremely high level of accuracy and since Scottish Glen is known to be under threat from threat actors, this is essential. K-Means can be too easily skewed by false data within training sets and therefore is likely to provide a larger number of false positives.

The K-Means algorithm could be used for a simpler IDS system. The Random Forest algorithm although more memory intensive, is robust to overfitting a significant challenge to machine learning algorithms. Random Forrest is the most suitable algorithm for Scottish Glen’s IDS system.

3. Design & Implementation

To correctly identify the various threats that could appear on the Scottish Glen network an IDS that runs the Random Forest classification algorithm must be developed to categorise the network packets according to their attack category. A data pipeline will be set up to allow the IDS system to build an appropriate Machine Learning model which will classify the network packets. This pipeline is known as the Machine Learning Workflow. (Datacamp, 2022) The stages of the data pipeline will include data ingestion/preprocessing, modelling, analysis, and communication of results identified.

Initially, at the data ingestion/preprocessing stage, a packet capturing tool such as Wireshark or Snort will be utilised to collect and store the data. This tool will be deployed at a network entry/exit point.

At the modelling stage, the data will be split into testing and training sets. The training data set is used to train the Random Forest algorithm to identify the predefined types of network traffic.

After initial modelling is complete the Random Forest algorithm will be analysed against the testing data. Evaluation metrics such as accuracy, precision, recall, F1-score, ROC curve, and AUC are calculated.

Following analysis, Hyperparameter Tunning should occur to optimise and improve the algorithm. Items such as the number of decisions and the depth of each decision tree will be modified to optimise and improve the model. Modelling and analysis will then be repeated several times until an optimal Machine Learning algorithm is developed.

The Random Forest algorithm will then be deployed on the Scottish Glen network. The results of the model will be communicated to the Security Team in real time. Therefore, any issues can be quickly identified and mitigated against.

4. Evaluation Metrics

For this model, standard evaluation metrics have been chosen to assess the effectiveness of the Random Forest classification algorithm's performance.

4.1. Confusion Matrix

Initially, a confusion matrix will be produced using the testing dataset and documenting the results into four categories. True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) (Geeks-For-Geeks, 2024). This evaluation matrix enables us to gain an understanding of how successful the model is at detecting malicious traffic as well as allowing for further analysis. Figure 5 identifies an example of a confusion Matrix.

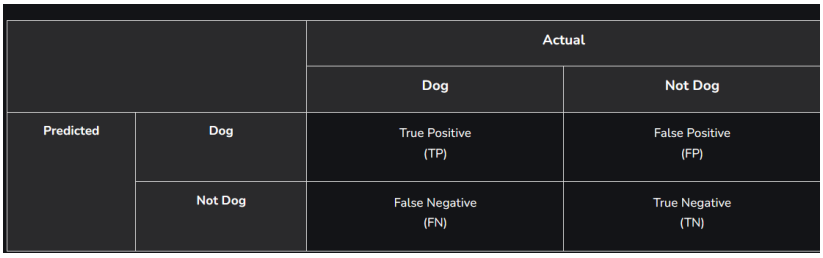

Figure 5 Confusion Matrix (Geeks-For-Geeks, 2024)

For Scottish Glen's Intrusion Detection System (IDS) a high TP is vital to ensure the security team do not miss attacks against the network.

4.2. Precision, Recall and F1-Score

By using the Confusion Matrix, it is possible to calculate several more detailed metrics including:

- Precision

This matrix calculates how accurate a model's positive predictions have been. (Geeks-For-Geeks, 2024) This is a key matrix for a successful IDS. The equation to calculate such a matrix is:

$$P = \frac{TP}{TP + FP}$$

In this equation P is the precision of the Random Forest algorithm when the total true positives (TP) and false positive (FP) are added up.

- Recall

Recall measures how effective the classification model is at identifying that all attacks are correctly identified. The equation to calculate such a matrix is as follows:

$$R = \frac{TP}{TP + FN}$$

In the recall equation R is the Recall of the Random Forest algorithm. "It is the ratio of the number of true positive (TP) instances to the sum of true positive and false negative (FN) instances." (Amsten, 2023)

- F1 -Score

This matrix is the calculation between precision and recall into a single score, it is the harmonic mean between these two values. This value is greater than 0 but less than 1 (0=<x =< 1). This matrix is used to calculate how robust and precise the algorithm is (Geeks For Geeks, 2023). The equation is as follows:

$$F1 = \frac{2 * P * R}{P + R}$$

In the equation above F1 is the F1-Score of the algorithm. P is the Precision and R is the Recall values.

4.3. ROC Curve

The ROC Curve or Receiver Operating Characteristic (ROC) Curve is the graphical representation of the TPR (True Positive Rate) and FPR (False Positive Rate) for the IDS model at various classification thresholds. (Google Developer, 2022)

- TPR

The True Positive Rate (TPR) is another name for Recall as discussed above. The equation is as follows:

$$TPR = \frac{TP}{TP + FN}$$

- FPR

The False Positive Rate (FPR) is the ratio of negative cases incorrectly predicted as positive. The equation is as follows:

$$FPR = \frac{FP}{FP + TN}$$

In this equation the False Positive Rate (FPR) of the algorithm is the proportion of false negatives with respect to all data points that are negative.

- ROC Curve

The ROC curve plots TPR against (x axis) FPR rate (y-axis) at different classification thresholds. Figure 6 identifies the TP vs FP rate at different classification thresholds:

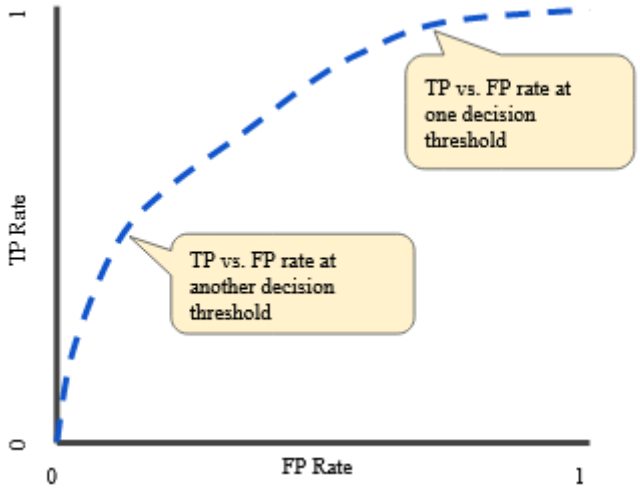

Figure 6- TP vs FP rate (Google Developer, 2022)

A successful IDS model ROC curve would hug the top-left corner, indicating high TPR with low FPR. (Science, 2022)

- AUC

The AUC or Area Under ROC curve is a metric that allows for comparison of different ROC Curves. The AUC is always a value between 0 and 1. The closer to one the better the performance of the IDS whilst a value equal to 0.5 is no better than tossing a coin. The higher the AUC the better the IDS algorithm would be at determining the IDS traffic. (Google Developer, 2022)

5. Conclusion

In conclusion, the Random Forest algorithm is the most desirable algorithm for the Scottish Glen IDS system as it is highly accurate in comparison to other algorithms.

The evaluation metrics employed should identify a high level of accuracy to ensure that the IDS system is detecting and classifying the correct network traffic, and that any threats detected are

being reported to the security team to allow them to take appropriate action to prevent or mitigate the attack.

The IDS system should be regularly tested and new models with higher accuracy and better detection rate should be trained. This will allow Scottish Glen a fuller and more detailed picture of traffic that goes across their network.

6. References

Amsten, 2023. *Evaluation Metrics in Machine Learning*. [Online]
Available at: https://www.geeksforgeeks.org/metrics-for-machine-learning-model/
[Accessed 21 03 2024].

Datacamp, 2022. *A Beginner's Guide to The Machine Learning Workflow*. [Online]
Available at: https://www.datacamp.com/blog/a-beginner-s-guide-to-the-machine-learning-workflow
[Accessed 21 03 2024].

Geeks For Geeks, 2023. *F1 Score in Machine Learning*. [Online]
Available at: https://www.geeksforgeeks.org/f1-score-in-machine-learning/
[Accessed 21 03 2024].

Geeks-For-Geeks, 2024. *Confusion Matrix in Machine Learning*. [Online]
Available at: https://www.geeksforgeeks.org/confusion-matrix-machine-learning/
[Accessed 21 03 2024].

Geeks-For-Geeks, 2024. *ML | Underfitting and Overfitting*. [Online]
Available at: https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/
[Accessed 24 03 2024].

Geeks-For-Geeks, 2024. *Random Forest Algorithm in Machine Learning*. [Online]
Available at: https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/
[Accessed 21 03 2024].

Geeks-For-Geeks, ND. *K means Clustering – Introduction*. [Online]
Available at: https://www.geeksforgeeks.org/k-means-clustering-introduction/
[Accessed 21 03 2024].

Google Developer, 2022. *Classification: ROC Curve and AUC*. [Online]
Available at: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
[Accessed 21 03 2024].

Google Developer, 2022. *k-Means Advantages and Disadvantages*. [Online]
Available at: https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages
[Accessed 21 03 2024].

IBM, ND. *What is a machine learning algorithm?*. [Online]
Available at: https://www.ibm.com/topics/machine-learning-algorithms
[Accessed 21 03 2024].

IBM, ND. *What is random forest?*. [Online]
Available at: https://www.ibm.com/topics/random-forest
[Accessed 25 03 2024].

Logunova, I., 2021. *Random Forest Classifier: Basic Principles and Applications*. [Online]
Available at: https://serokell.io/blog/random-forest-classification
[Accessed 21 03 2024].

Science, T. D., 2022. *Interpreting ROC Curve and ROC AUC for Classification Evaluation*. [Online]
Available at: https://towardsdatascience.com/interpreting-roc-curve-and-roc-auc-for-classification-evaluation-28ec3983f077
[Accessed 21 03 2024].

scikit-learn.org, ND. *2.3.2. K-means*. [Online]
Available at: https://scikit-learn.org/stable/modules/clustering.html#k-means
[Accessed 21 03 2024].

scikit-learn, ND. *1.10. Decision Trees*. [Online]
Available at: https://scikit-learn.org/stable/modules/tree.html
[Accessed 21 03 2024].

Sellahewa, K., 2023. *Random Forest Explained! (Regression and Classification Tasks)*. [Online]
Available at: https://www.linkedin.com/pulse/random-forest-explained-regression-classification-tasks-sellahewa/
[Accessed 21 03 2024].

susmit_sekhar_bhakta, 2024. *Random Forest Algorithm in Machine Learning*. [Online]
Available at: https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/
[Accessed 21 03 2024].