**DATA 1202 - DATA ANALYSIS TOOLS ANALYTICS**

DATA ANALYTICS USING MACHINE LEARNING

Submitted TO : Dr. Sk. Md. Mizanur Rahman

Submitted By : Promise joshy (100873024)

Vimal vijayraj (100892118)

Aswin philip (100906431)

Submitted On : 18/04/2024

**Introduction**

This report's goal is to assess the effectiveness of three machine learning classifiers using a dataset made up of features taken out of system processes: Random Forest, Gradient Boosting, and Logistic Regression. The study includes an overview of the dataset, details the process of developing, testing, and refining the classifiers, and concludes with a discussion and comparison of the outcomes.

**Information about the dataset**

The evaluation's dataset is distinguished by its thorough depiction of system processes, which includes a range of attributes that shed light on the features and behaviour of these processes. The dataset's distinctive traits are broken down as follows:

* Number of instances:100,000 instances make up the dataset; each instance is a snapshot of a system process at a particular moment in time. Machine learning models may be robustly trained and evaluated thanks to this large dataset.
* Number of Features: There are 35 features in all, comprising both category and numerical variables in the dataset. These features record a variety of system process characteristics, such as timestamp data (e.g., milliseconds), memory management properties (e.g., page offset, page table entries), and performance metrics (e.g., CPU time).
* Distribution of Classes: There are an equal number of instances of each class in the dataset, indicating a balanced distribution of classes. There are specifically 50,000 cases classified as "malware" and 50,000 instances classified as "benign." By maintaining this balance, bias towards any one class is avoided and the machine learning models are trained on representative data from both classes.
* Rich Details Regarding System Operations: The dataset's thorough coverage of system process attributes is one of its most notable qualities. The dataset contains comprehensive metrics pertaining to process priority, memory management, CPU utilisation, context switches, file system activity, and more, in addition to standard data like timestamp and process status. This abundance of data makes it possible to perform complex analysis and classification tasks by offering insightful knowledge about the traits and behaviour of system operations.

**Proper split of dataset**

*Training Set*: The majority of the dataset is included in the training set, which is used to train the machine learning model. During this stage, the model discovers relationships and patterns in the data.

*Testing Set*: Consisting of data not utilised in training, the testing set is set aside for assessing the model's performance. To evaluate the model's accuracy and capacity for generalisation, the predictions it made on this unobserved data are compared to the true labels.   
Relevance:   
  
By preventing memory of the training set, overfitting is prevented and the model is ensured to generalise well to fresh data.   
Evaluation with objectivity: Offers a neutral standard by which to measure the model's effectiveness.   
Analysing bias and variance helps determine if the model is overfitting (high variance) or underfitting (high bias).   
For classification jobs where the class distributions are unbalanced, stratified splitting is essential as it preserves the class proportions in both sets.

**Building the three classifiers**

*Random Forest*:

*Adaptability:* Suitable for both classification and regression applications, Random Forest is a flexible ensemble learning technique. It adapts well to overfitting and performs admirably with high-dimensional data.

*Effectiveness:* During training, Random Forest creates several decision trees and votes to merge their predictions, creating a more reliable and accurate model. It successfully manages feature interactions and non-linear relationships.

Random Forest offers a metric for feature importance, which is helpful in determining which characteristics in the dataset are the most pertinent.

*Gradient Boosting:*

*High Accuracy*: This technique is renowned for its capacity to identify intricate correlations in data and for having a high predicted accuracy.

*Sequential Learning:* Gradient Boosting develops trees in a sequential fashion, with each tree fixing the mistakes of its predecessor, in contrast to Random Forest. Improved performance is frequently the result of this sequential learning approach.

*Regularisation:* Gradient Boosting is resistant to noisy data because it incorporates regularisation parameters to manage model complexity and avoid overfitting.

*Logistic regression:*   
Interpretability: The Logistic Regression model is a straightforward and comprehensible linear model that is extensively employed in binary classification jobs.

*Efficiency*: Large-scale datasets are a good fit for the computationally efficient method of logistic regression. When there is a linear or almost linear relationship between the characteristics and the target variable, it works well.

*Probabilistic Interpretation*: Because logistic regression offers probabilistic explanations for forecasts, it is a good choice for jobs requiring comprehension of the degree of certainty or uncertainty in predictions.

These classifiers were selected due to their ability to handle various data kinds and levels of complexity, as well as their complimentary strengths. While Logistic Regression provides simplicity, interpretability, and efficiency, making it a useful baseline model for comparison, Random Forest and Gradient Boosting are strong ensemble methods that can capture intricate patterns in the data. All things considered, the amalgamation of these classifiers offers a sturdy structure for proficiently assessing the classification task.

**Training the classifier**

Fitting each chosen model to the training data allows the classifiers to learn the underlying relationships and patterns seen in the dataset. This is a thorough description of the training procedure:   
  
*Data Preparation:*

The dataset is preprocessed to make sure it is in a format that the classifiers can use before training. In order to extract pertinent information, this may entail addressing missing values, encoding categorical variables, scaling numerical features, and doing feature engineering.

*Classifier Selection:*

The Random Forest, Gradient Boosting, and Logistic Regression classifiers are selected and instantiated with pre-configured or default parameters. The selection of these algorithms is predicated upon their adaptability and efficacy in addressing categorization assignments of diverse complexity.

*Training Set:*

The classifiers are trained using the training set, which normally consists of 80% of the original dataset. It is made up of labelled instances, each of which has characteristics taken from system processes together with the labels (such "malware" or "benign") that correlate to those qualities.

*Model Fitting*:

The fit() method is used to optimise the model parameters in order to minimise the discrepancy between the actual and predicted labels on the training set. This process is repeated for each classifier. The classifiers pick up pattern recognition and prediction skills depending on the input features during training.

*Learning Process:*

To minimise a given loss function, the model's parameters are iteratively adjusted throughout the training phase. Multiple decision trees are trained sequentially or concurrently for ensemble methods like Random Forest and Gradient Boosting. Each tree learns from the mistakes of its predecessors.

*Evaluation and Validation:*

During training, the classifiers' performance may be tracked using methods like cross-validation to evaluate how well they generalise and spot possible problems like under- or overfitting.

*Model Optimisation:*

Hyperparameter tuning or model selection approaches may be used to further optimise the classifiers, depending on the performance metrics and validation findings. The goal of this iterative approach is to guarantee robustness on unknown data and enhance model performance.

**Testing the three classifiers**

Testing the classifiers involves evaluating their performance on a separate set of data that was not used during training – the reserved testing set. Here's an explanation of the testing process:

*Unseen Data:*

The testing set contains instances with known features but withheld labels. This ensures that the classifiers are evaluated on data they haven't seen before, simulating real-world scenarios where they encounter new, unseen instances.

*Prediction:*

Each trained classifier is used to make predictions on the features of the instances in the testing set. The classifiers apply the learned patterns and relationships from the training phase to classify or predict the labels of the testing instances.

*Evaluation Metrics:*

The predictions made by the classifiers are compared to the true labels of the testing instances. Various evaluation metrics are computed to assess the performance of the classifiers, including accuracy, precision, recall, F1-score, and area under the ROC curve (ROC AUC).

*Assessing Generalization:*

By testing the classifiers on unseen data, we evaluate their generalization ability – their capacity to make accurate predictions on instances they haven't been trained on. This step is crucial for assessing how well the models perform in real-world scenarios and whether they suffer from issues like overfitting or underfitting.

*Comparison and Analysis:*

The performance of the classifiers is analyzed and compared using the evaluation metrics. This helps in understanding their strengths and weaknesses, identifying which model performs best for the given task, and gaining insights into areas for improvement or further optimization.

Overall, testing the classifiers on unseen data is a critical step in the machine learning pipeline, providing valuable insights into their performance and ensuring their reliability and effectiveness in real-world applications. It validates the models' ability to generalize beyond the training data and guides decision-making regarding their deployment and use.

**Explaining and comparing the results**

**Random Forest**

Confusion Matrix:

[[10000 0]

[ 0 10000]]

Evaluation Metrics:

Accuracy: 1.0

Precision: 1.0

Recall: 1.0

F1-score: 1.0

ROC AUC: 1.0

**Discussion: Random Forest achieved perfect accuracy, precision, recall, and F1-score, indicating excellent performance in classifying both classes (malware and benign).**

**Gradient Boosting**

Confusion Matrix:

[[9991 9]

[ 0 10000]]

Evaluation Metrics:

Accuracy: 0.99955

Precision: 0.99955

Recall: 0.99955

F1-score: 0.99955

ROC AUC: 1.0

**Discussion: Gradient Boosting also performed exceptionally well with high accuracy and minimal misclassifications.**

**Logistic Régression**

Confusion Matrix:

[[6445 3555]

[1446 8554]]

Évaluation Metrics:

Accuracy: 0.74995

Precision: 0.76158

Recall: 0.74995

F1-score: 0.74714

ROC AUC: 0.81619

**Discussion: Logistic Regression achieved lower accuracy and F1-score compared to Random Forest and Gradient Boosting, indicating that it struggled with classification, especially in distinguishing between malware and benign instances.**