**PROJECT REPORT**

**PROJECT TITLE: Microsoft Classifying Cyber security Incidents with Machine Learning**

**Project Overview**

This project aims to enhance SOC efficiency by developing a machine learning model to predict the **triage grade** of cybersecurity incidents, categorizing them as:

* **True Positive (TP)**: Genuine security threats.
* **Benign Positive (BP)**: Alerts that appear as threats but are benign.
* **False Positive (FP)**: False alarms with no security impact.

This model will support SOC analysts by filtering alerts, reducing false positives, and prioritizing genuine threats, improving response time and overall security posture.

### 1.Data Overview

#### **Dataset**

* **Source**: GUIDE dataset, consisting of cyber security incident data and response feedback.
* **Train Data**: train.csv – used to train the model.
* **Test Data**: test.csv – used to evaluate the model's performance.

#### **Key Features**

The model uses features such as:

* **MitreTechniques, Category, EntityType**: Indicators of attack technique, type, and target.
* **DetectorId, AlertTitle, DeviceId, IpAddress**: Related to the incident’s source and specifics.
* **Target Variable (IncidentGrade)**: Categorizes each incident as TP, BP, or FP.

**2. Data Preprocessing and Feature Engineering**

The following steps were undertaken to prepare the data for modeling:

* **Label Encoding**: All categorical features were label-encoded for consistency.
* **Handling Missing Values**: Missing data was addressed by imputation or by dropping features with excessive null values.
* **Oversampling**: The minority classes were oversampled to address class imbalance, particularly to boost FP and BP instances.

**3. Model Training and Selection**

Several classification algorithms were evaluated for this task, including Decision Tree, Random Forest, and Gradient Boosting, with **macro-F1 score, precision, and recall** as key performance metrics.

* **Chosen Model**: Based on evaluation, the model selected provides high macro-F1, precision, and recall scores, ensuring a balanced and robust performance across classes (TP, BP, FP).

**4. Evaluation Metrics**

The model was evaluated on the test.csv dataset to ensure it generalizes well to unseen data.

* **Macro-F1 Score**: Measures overall model performance by averaging F1 scores across all classes.
* **Precision**: Focuses on reducing false positives, helping SOC analysts avoid unnecessary actions.
* **Recall**: Ensures that genuine threats (TPs) are captured effectively, minimizing the risk of overlooking critical incidents.

**5. Key Insights and Findings**

1. **Feature Importance**: The top contributing features were MitreTechniques, Category, and EntityType, which had the highest correlation with the incident grade. These insights can guide SOC teams on which factors to prioritize in initial triage.
2. **Class Distribution and Imbalance**:
   * The dataset contained a majority of TP instances, with fewer FP and BP labels.
   * Oversampling was effective in improving model accuracy for BP and FP cases, leading to a balanced performance across all classes.
3. **Model Performance on Test Data**:

### ****Classification Metrics****

| **Class** | **Precision** | | **Recall** | | **F1-Score** |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP (0)** | | 0.59 | | 0.80 | 0.68 |  |
| **BP (1)** | | 0.74 | | 0.49 | 0.59 |  |
| **FP (2)** | | 0.66 | | 0.65 | 0.66 |  |
|  | |  | |  |  |  |
|  | |  | |  |  |  |

## **Insights from Metrics**

1. **Precision vs. Recall**:
   * Precision for **Class 1 (BP)** is high (0.74), meaning the model effectively reduces false positives in this class.
   * Recall for **Class 0 (TP)** is high (0.80), ensuring genuine threats are captured with minimal misses.
   * Recall for **Class 1 (BP)** is lower (0.49), indicating some benign incidents are incorrectly flagged as threats.
2. **Class Distribution Impact**:
   * Balanced performance across all three classes with F1-scores around 0.59–0.68, demonstrating the model's ability to generalize.
3. **ROC-AUC**:
   * The score of 0.83 is encouraging and indicates good overall performance for multi-class classification.

**6. Recommendations for SOC Teams**

1. **Guided Response**:
   * **True Positives (TP)**: Alerts flagged as TP should be prioritized and investigated immediately. The SOC should have a predefined, streamlined response plan in place.
   * **Benign Positives (BP)**: BP alerts can be deprioritized or tagged for periodic review. An automated notification could inform analysts about these, reducing manual intervention.
   * **False Positives (FP)**: Incidents flagged as FP should be logged but require minimal attention. SOC could implement automated flagging to filter these from the primary alert dashboard, reducing noise.
2. **Model Integration and Feedback Loop**:
   * **Continuous Monitoring**: The model’s performance should be monitored regularly, especially as new data becomes available. Re-train and fine-tune the model periodically to maintain high precision and recall.
   * **Feedback Incorporation**: Analysts should be able to provide feedback on the model’s classification for continual improvement. Integrate this feedback into periodic re-training.
3. **Adopting an Alert Triage Strategy**:
   * By leveraging the model, SOC can establish clear triage protocols based on model-predicted grades, allowing for a more structured incident response flow.
   * Reducing FP and BP responses through automation will free up analysts to focus on high-risk incidents