# Microsoft: Classifying Cybersecurity Incidents with Machine Learning

## 1. Objective

The objective of this project is to classify cybersecurity incidents by predicting their triage grade using machine learning techniques. This classification aims to assist Security Operation Centers (SOCs) in managing and prioritizing incidents more efficiently, enabling quicker responses to critical threats.

## 2. Methodology

1. Data Exploration & Preprocessing:

- Data Source: The training and test datasets contained millions of rows with features like 'AlertTitle', 'Category', 'EntityType', 'EvidenceRole', and others. The target variable is 'IncidentGrade'.
- Missing Values: Columns like 'MitreTechniques' and 'IncidentGrade' had missing values handled by imputation or dropping.
- Feature Extraction: Extracted features like Year, Month, Day, and Hour from timestamp data.
- Feature Encoding: Label encoding used for categorical features.

2. Data Balancing:

- SMOTE was used to balance the skewed target class 'IncidentGrade'.

3. Model Selection & Training:

- Tested: Logistic Regression, Random Forest, Decision Tree, XGBoost, LightGBM, Gradient Boosting.

4. Model Evaluation:

- Metrics: Accuracy, Precision, Recall, F1 Score, Macro-F1 Score.

# Microsoft: Classifying Cybersecurity Incidents with Machine Learning

## 3. Model Performance Summary

| Model | Accuracy | Macro-F1 | Precision | Recall |
|---|---|---|---|---|
| Logistic Regression | 0.63 | 0.54 | 0.64 | 0.55 |
| Decision Tree | 0.70 | 0.67 | 0.70 | 0.66 |
| Random Forest | 0.70 | 0.67 | 0.70 | 0.66 |
| XGBoost | 0.68 | 0.62 | 0.71 | 0.61 |
| LightGBM | 0.68 | 0.61 | 0.72 | 0.61 |
| Gradient Boosting | 0.64 | 0.55 | 0.68 | 0.56 |

# Microsoft: Classifying Cybersecurity Incidents with Machine Learning

## 4. Findings

1. Best Performing Model:

- Random Forest showed the best Macro-F1 Score (0.67), indicating a strong balance of precision and recall.

2. Accuracy vs Macro-F1:

- Though several models had similar accuracy (~0.70), Random Forest performed better across all classes.

3. Error Analysis:

- Logistic Regression struggled with minority class recall, affecting classification reliability.

## 5. Rationale for Model Selection

- Random Forest was chosen for its ensemble stability and strong all-around metrics.
- XGBoost/LightGBM had higher precision but lower recall, making them less ideal for imbalanced class scenarios.

# Microsoft: Classifying Cybersecurity Incidents with Machine Learning

## 6. Model Improvement

1. Hyperparameter Tuning: Optimize n_estimators, max_depth, etc. using GridSearch.

2. Cross-Validation: Use k-fold CV for robustness.

3. Feature Engineering: Enhance understanding by deriving new features from 'EntityType', 'Category', etc.

## 7. Conclusion

This project demonstrates the application of machine learning to cybersecurity incident triage. Random Forest emerged as the best model with a balanced performance. With further tuning and feature engineering, the model can be deployed to help SOCs rapidly and accurately classify security incidents.