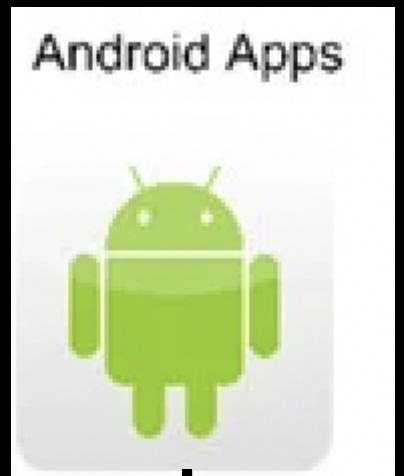
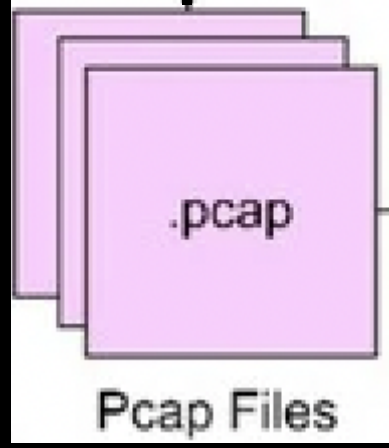
Shielding Android Devices

Using Machine Learning to Detect Malware

Network Flow Diagram



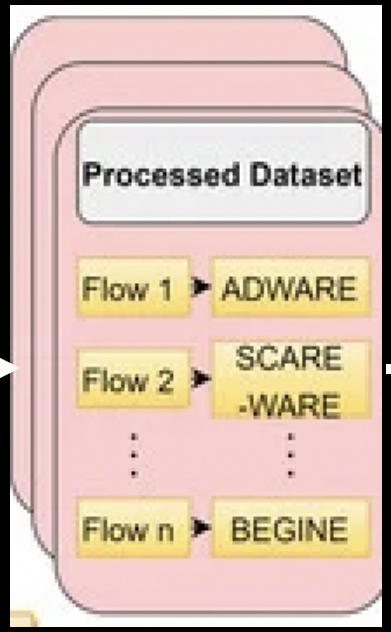


CICFlowMeter-V3

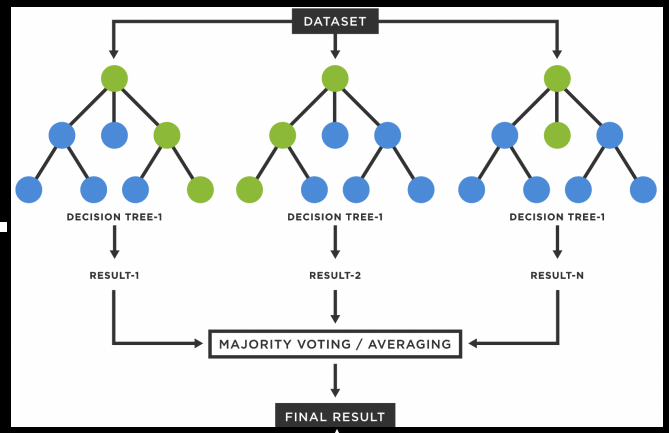
Network Traffic Information

- Flow ID
- Source IP
- Destination
- .....

My Dataset



RandomForestClassifier



The Data

- **ORIGINAL DATASET**
  - 86 columns and 355,630 rows
- **BINARY CLASSIFICATION**
  - 1 = Malware
  - 0 = Benign
  - Class imbalance: 93% Malware, 7% Benign
    - SMOTE
- **FEATURE SELECTION**
  - Mann-Whitney U Test for numerical columns
  - Chi2 test for categorical columns
  - p-value < 0.05
- **F2 SCORING METRIC**
  - Harmonic mean between with emphasis on recall

Destination Port	Protocol	Flow Duration	Total Fwd Packets	Total Backward Packets	Total Length of Fwd Packets	Total Length of Bwd Packets	Fwd Packet Length Max
443.000000	6	37027	1	1	0.000000	0.000000	0.000000
443.000000	6	36653	1	1	0.000000	0.000000	0.000000
443.000000	6	534099	8	12	1011.000000	11924.000000	581.000000
443.000000	6	9309	3	0	0.000000	0.000000	0.000000
443.000000	6	19890496	8	6	430.000000	5679.000000	218.000000
443.000000	6	19196263	5	2	0.000000	0.000000	0.000000
443.000000	6	4304325	1	2	0.000000	0.000000	0.000000
443.000000	6	37926	1	2	23.000000	0.000000	23.000000
443.000000	6	36631	1	2	23.000000	0.000000	23.000000
443.000000	6	36535	1	2	0.000000	31.000000	0.000000

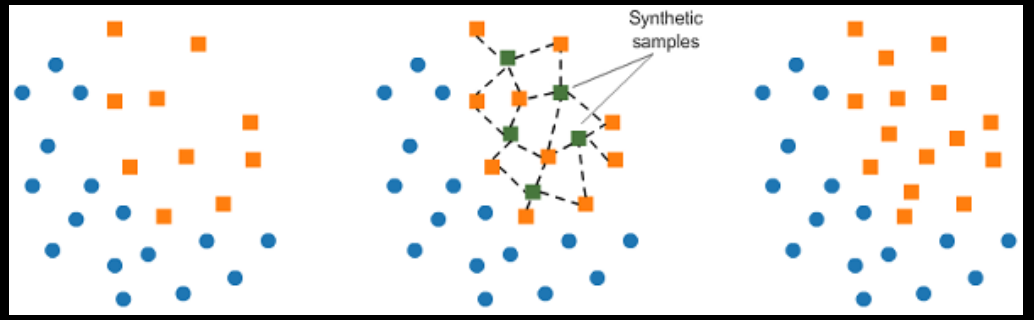
$$F2 = \frac{(1 + beta^2)(precision * recall)}{((beta^2 * precision) + recall)}$$

Model Selection

- **RANDOMFORESTCLASSIFIER:**
  - Reduces overfitting by averaging the predictions of multiple trees
  - Feature selection for tree construction
  - Handles: numerical, categorical, outliers
- **XGBOOSTCLASSIFIER:**
  - Gradient boosting for speed and performance.
  - Handles: missing values, numerical, categorical
- **LGBMCLASSIFIER:**
  - Gradient boosting for speed and efficiency.
  - Designed for large datasets and high-dimensional data.
  - "**Gradient-based One-Side Sampling**" (**GOSS**) reduces the number of data points to consider during tree construction, making it faster than other boosting methods.

Classifier	Preprocessor	Sampler	CV f2 Score	CV F1	CV Recall	CV Precision
RandomForestClassifier	RobustScaler	SMOTE	0.888826	0.890909	0.887476	0.894520
LGBMClassifier	MinMaxScaler	SMOTE	0.875125	0.881374	0.871266	0.892758
LGBMClassifier	StandardScaler	SMOTE	0.872793	0.879819	0.868495	0.892720
XGBClassifier	StandardScaler	SMOTE	0.866005	0.876265	0.860042	0.895837
RandomForestClassifier	StandardScaler	SMOTE	0.865381	0.875577	0.859423	0.894954
XGBClassifier	RobustScaler	SMOTE	0.863801	0.873708	0.857919	0.892309
LGBMClassifier	RobustScaler	SMOTE	0.852035	0.865657	0.844230	0.891837
RandomForestClassifier	MinMaxScaler	SMOTE	0.851265	0.865473	0.843227	0.892996
XGBClassifier	MinMaxScaler	SMOTE	0.849302	0.864674	0.840901	0.895032

SMOTE (Synthetic Minority Over-sampling Technique)



randomly picks a point from the minority class and computes the k-nearest neighbors for this point. The **synthetic points are added** between the chosen point and its neighbors

Comparing Baseline model and Optimized model after Tuning

Hyperparameter tuning: Optuna

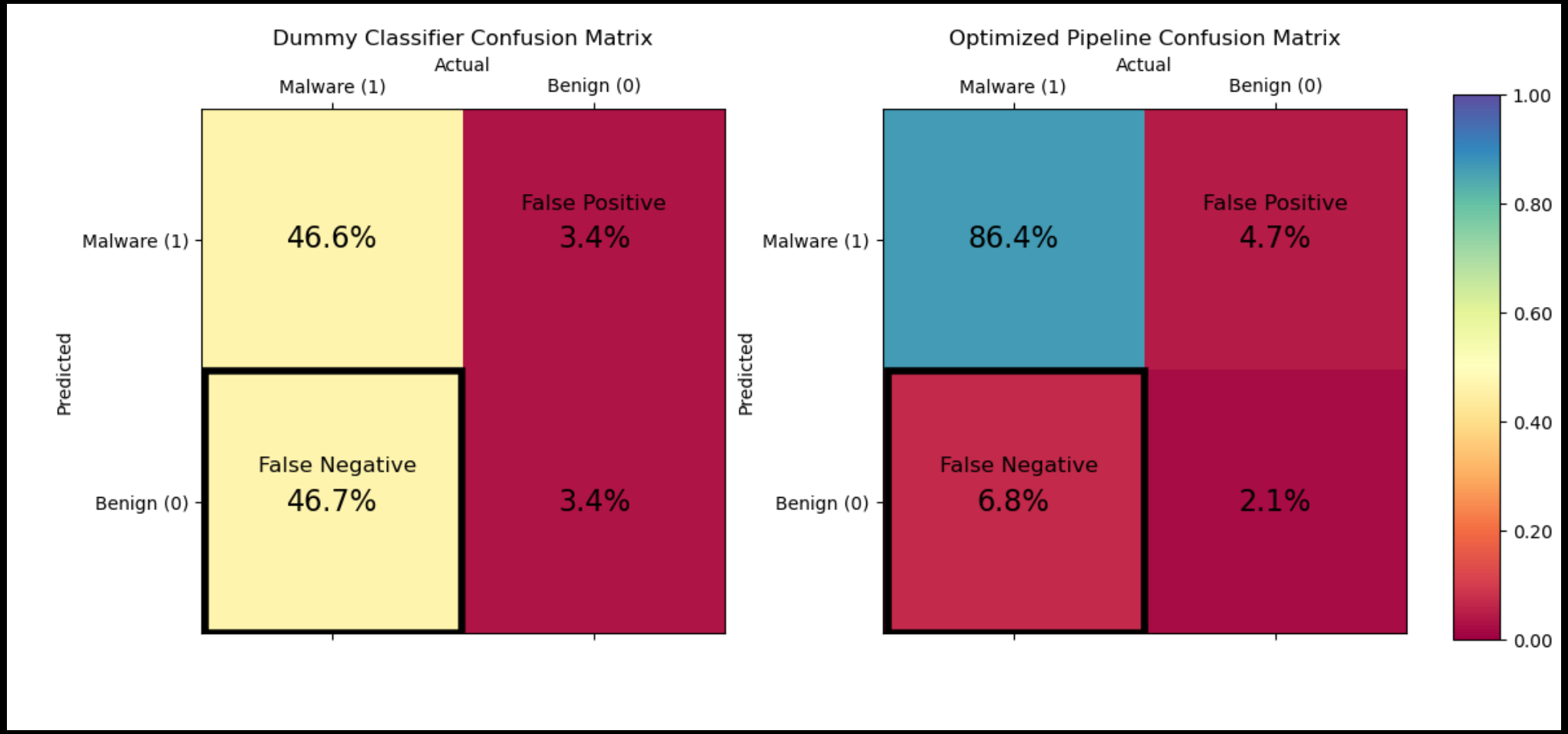
	f2_score_training	max_depth	n_estimators	min_samples_split	min_samples_leaf	max_features
rfc_base	0.888826	67	100	2	1	auto
rfc_optimized	0.888280	34	190	2	1	sqrt

* same in both models

- **MODEL COMPLEXITY:**
  - Lower complexity often reduces the risk of overfitting, as it prevents the model from fitting the noise in the training data.
- **NUMBER OF TREES:**
  - **Positive effect**:
    - generalization is average of all trees. More trees = less affected by randomness
  - **Negative effect**:
    - Overfit if trees are deep and complex.
    - Computationally more expensive

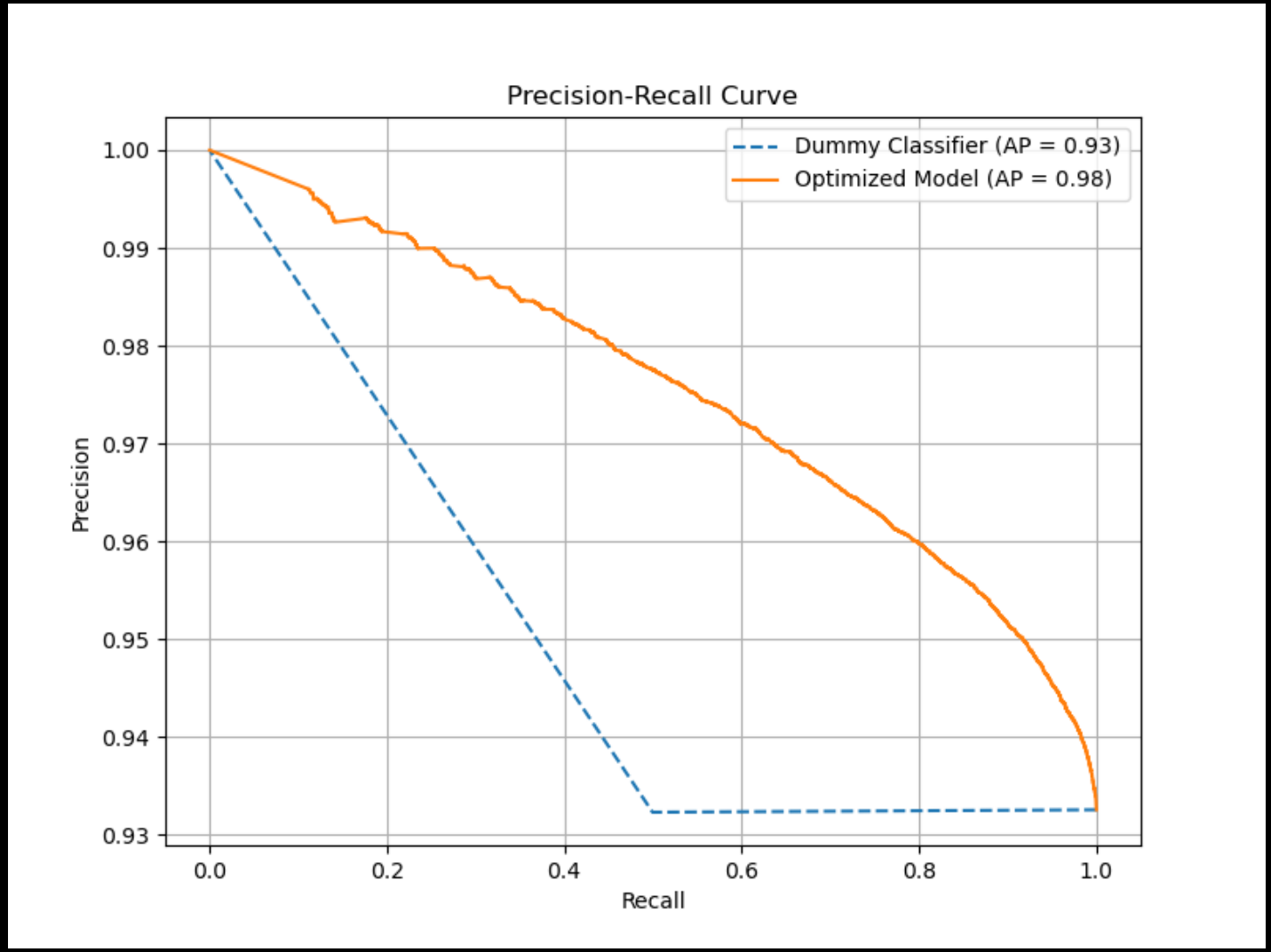
Confusion Matrices: Test Evaluation

Classifier	F2 Score
Dummy Classifier	0.528231
rfc_optimized	0.887812



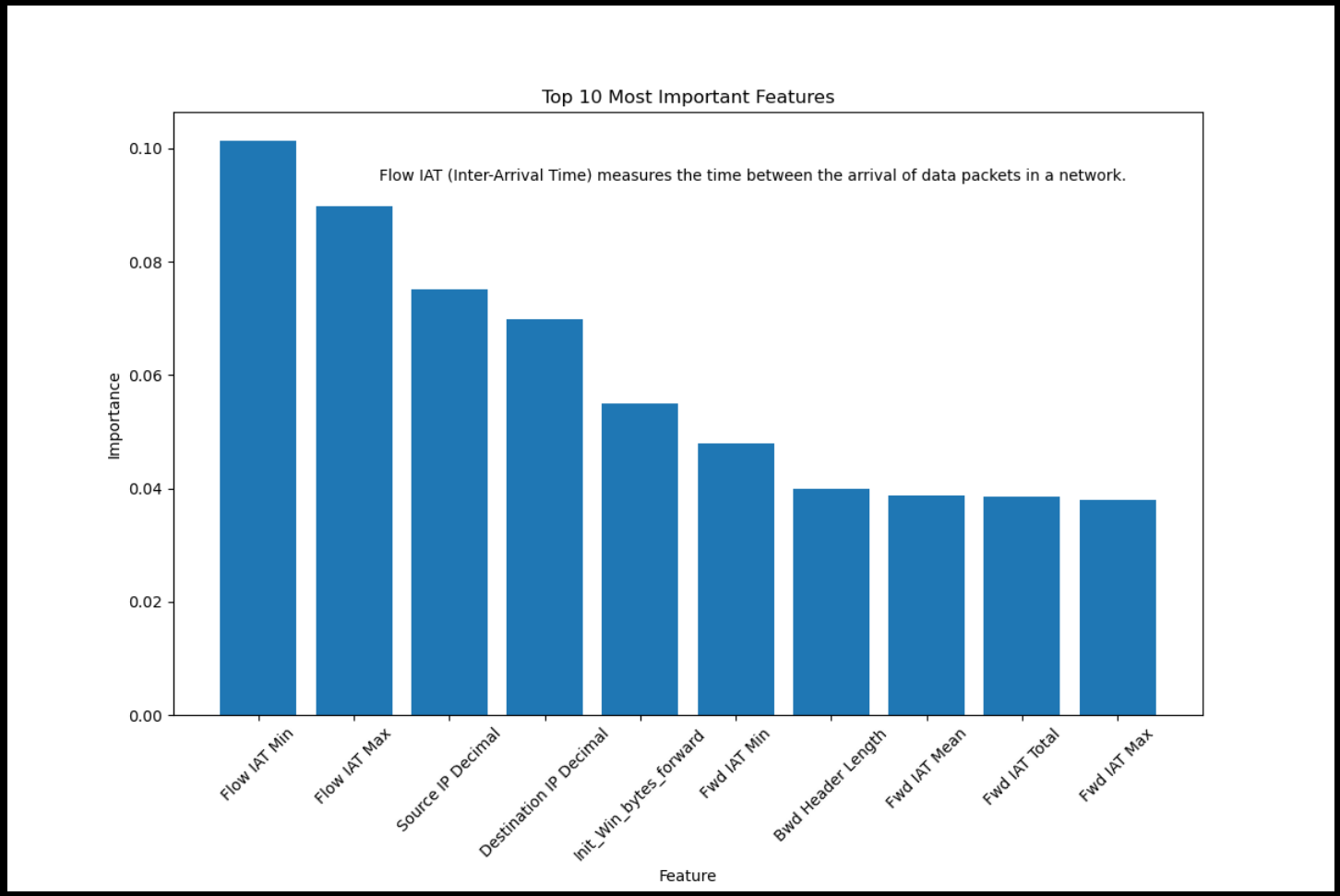
Precision-Recall Curve

- **VISUALIZATION OF PERFORMANCE**
  - Shows each models precision and recall values at various thresholds
- **THRESHOLDS**
  - Decision threshold to determine class label
  - Each point is assigned a probability of the label and if it is above the threshold it is assigned that label
- **SMOOTH CURVE**
  - Indicates high precision and recall values
- **CLOSER TO TOP RIGHT**
  - Indicates better performance
- **DUMMY GRAPH**
  - Slope Change means TP = FN



Most Important Features Optimized Model

- **FLOW IAT (INTER-ARRIVAL TIME)**
  - is a measure of the time between the arrival of data packets in a network.
- **FLOW IAT MIN:**
  - This feature represents the minimum time interval between two consecutive packets in a network flow.
- **FLOW IAT MAX:**
  - This feature represents the maximum time interval between two consecutive packets in a network flow.



Most Important Features Original Data

- **FLOW IAT MIN:**
  - low IAT time = transmit data quickly
    - DoS (Denial of Service) attack
    - Fast data exfiltration.
- **FLOW IAT MAX:**
  - high IAT = bursty traffic
    - Malware hiding communication by mimicking normal traffic patterns
    - Exfiltrating data in small bursts.