Exploratory Data Analysis & Hyperparameter Tuning

1. Exploratory Data Analysis (EDA)

During the EDA phase, we performed the following steps and observations:

• Class Distribution: Plotted counts of each "Action Taken" category (Logged, Blocked, Ignored) to confirm balance and identify any skew.

• Feature Overview: Displayed dataset info to verify non-null counts and data types across all 25 original columns.

• Statistical Summary:

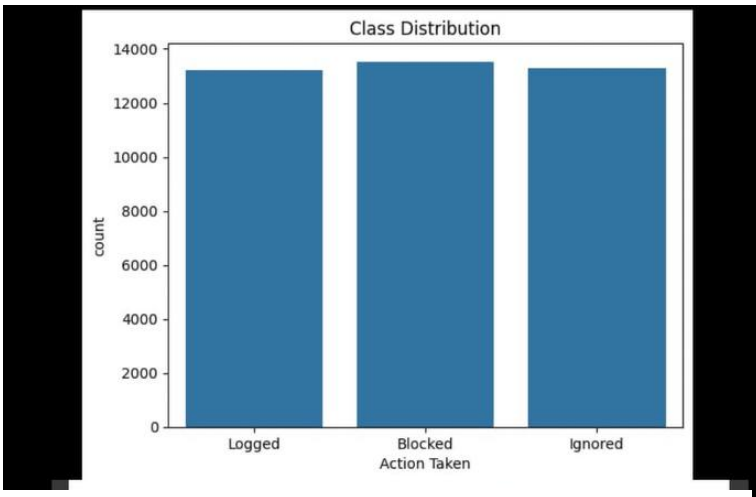
- Source/Destination Port: range 1 024–65 535, mean ≈33 000, high variability.

- Packet Length: mean ≈780 bytes, σ≈416, min 64, max 1 500.

- Anomaly Scores: uniform distribution on [0,100], mean ≈50, σ≈29.

• Missing Values: Identified five columns (~50% missing) and replaced nulls with explicit categories ("None", "No Data", "No Detection") to preserve all samples.

• Categorical Distributions: Visualized protocol usage (TCP/UDP/ICMP), traffic types (HTTP, DNS, FTP, etc.), and attack types (DDoS, Intrusion, Malware).



	Source Port	Destination Port	Packet Length	Anomaly Scores
count	40000.000000	40000.000000	40000.000000	40000.000000
mean	32970.356450	33150.868650	781.452725	50.113473
std	18560.425604	18574.668842	416.044192	28.853598
min	1027.000000	1024.000000	64.000000	0.000000
25%	16850.750000	17094.750000	420.000000	25.150000
50%	32856.000000	33004.500000	782.000000	50.345000
75%	48928.250000	49287.000000	1143.000000	75.030000
max	65530.000000	65535.000000	1500.000000	100.000000

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40000 entries, 0 to 39999
Data columns (total 25 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Timestamp             40000 non-null  object
 1   Source IP Address     40000 non-null  object
 2   Destination IP Address 40000 non-null  object
 3   Source Port           40000 non-null  int64
 4   Destination Port      40000 non-null  int64
 5   Protocol              40000 non-null  object
 6   Packet Length         40000 non-null  int64
 7   Packet Type           40000 non-null  object
 8   Traffic Type          40000 non-null  object
 9   Payload Data          40000 non-null  object
 10  Malware Indicators    20000 non-null  object
 11  Anomaly Scores        40000 non-null  float64
 12  Alerts/Warnings       19933 non-null  object
 13  Attack Type           40000 non-null  object
 14  Attack Signature      40000 non-null  object
 15  Action Taken          40000 non-null  object
 16  Severity Level        40000 non-null  object
 17  User Information      40000 non-null  object
 18  Device Information    40000 non-null  object
 19  Network Segment       40000 non-null  object
 20  Geo-location Data     40000 non-null  object
 21  Proxy Information     20149 non-null  object
 22  Firewall Logs         20039 non-null  object
 23  IDS/IPS Alerts        19950 non-null  object
 24  Log Source            40000 non-null  object
dtypes: float64(1), int64(3), object(21)
memory usage: 7.6+ MB
None
```

2. Hyperparameter Tuning (Random Forest)

To optimize the Random Forest classifier efficiently, we applied successive halving search:

• Subsampling: Extracted 10% of training data to accelerate tuning.

• Halving Strategy: Used HalvingRandomSearchCV with factor=2 and 2-fold CV, progressively narrowing from 529 to 2 candidates over 10 iterations.

• Parameter Space: n_estimators ∈ [50,200], max_depth ∈ [5,20].

• Results: Best params = {n_estimators:178, max_depth:12}, CV accuracy = 62.91%.

This method reduced computational cost compared to a full search while effectively finding strong hyperparameters.