

KLASIFIKASI TRAFFIC JARINGAN

Hustler – Institut Teknologi Bandung



Erwan
Poltak

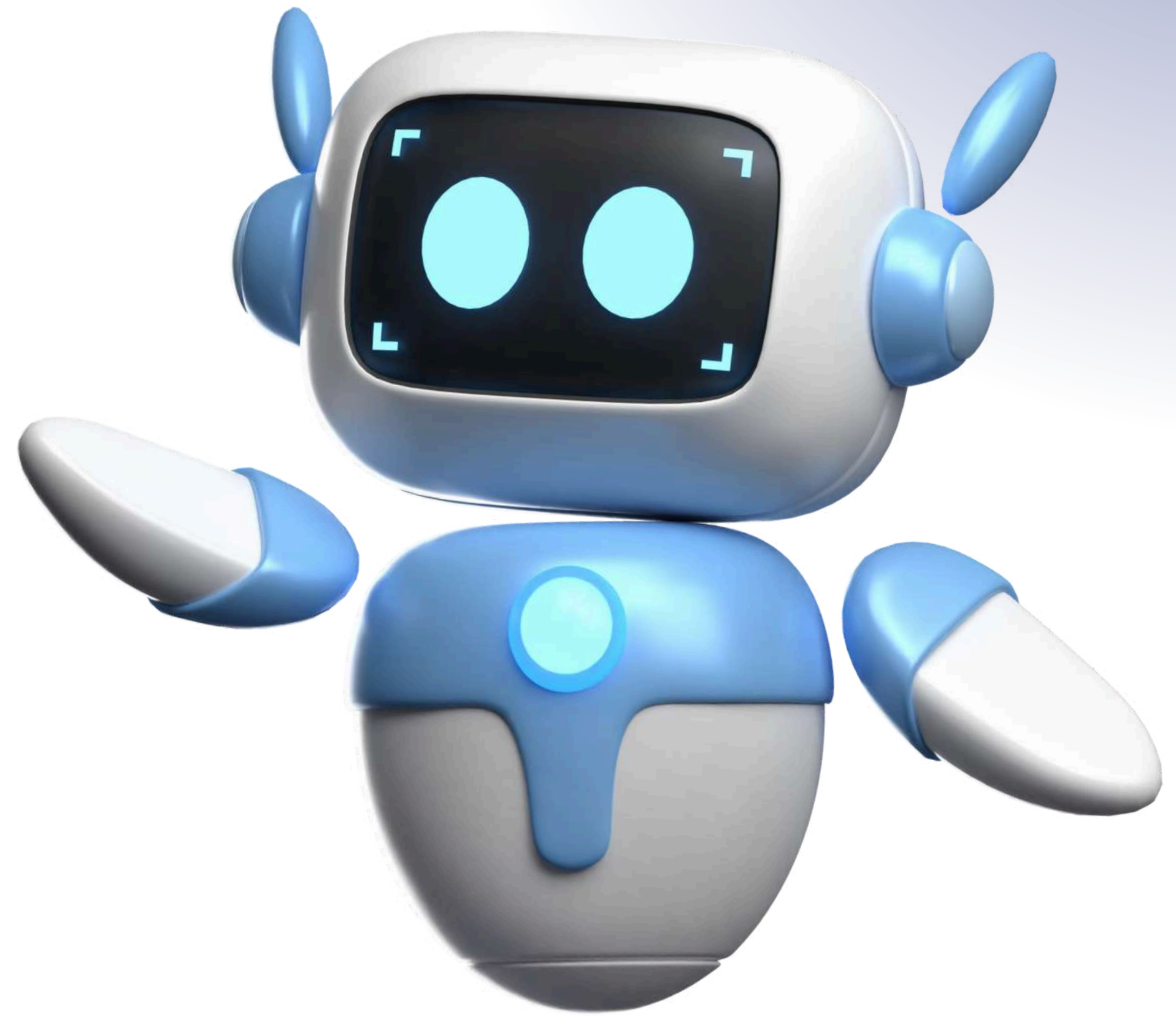


Viktor
Arsindiantoro



Mohamad
Maulana

Background



Langkah Awal Keamanan Jaringan di Era Digital

Ancaman Keamanan Volumen Data

Keamanan siber menjadi prioritas utama bagi perusahaan yang mengelola jaringan dengan lalu lintas tinggi, **volume data menjadi salah satu kunci** bisnis. Volume Data yang diproses seluruh dunia dengan pertumbuhan volume data yang terus meningkat **>100 Zettabytes** **400% setiap tahunnya**



Bruteforce
61 Juta
Serangan BruteForce
di Indonesia



Probing
29Juta
Serangan Probing
di Indonesia

Tantangan Utama Analisis Lalu Lintas Jaringan

Meninjau akan kebutuhan terkait keamanan volume data, perlu upaya menganalisis lalu lintas jaringan secara efektif guna mendeteksi potensi ancaman.

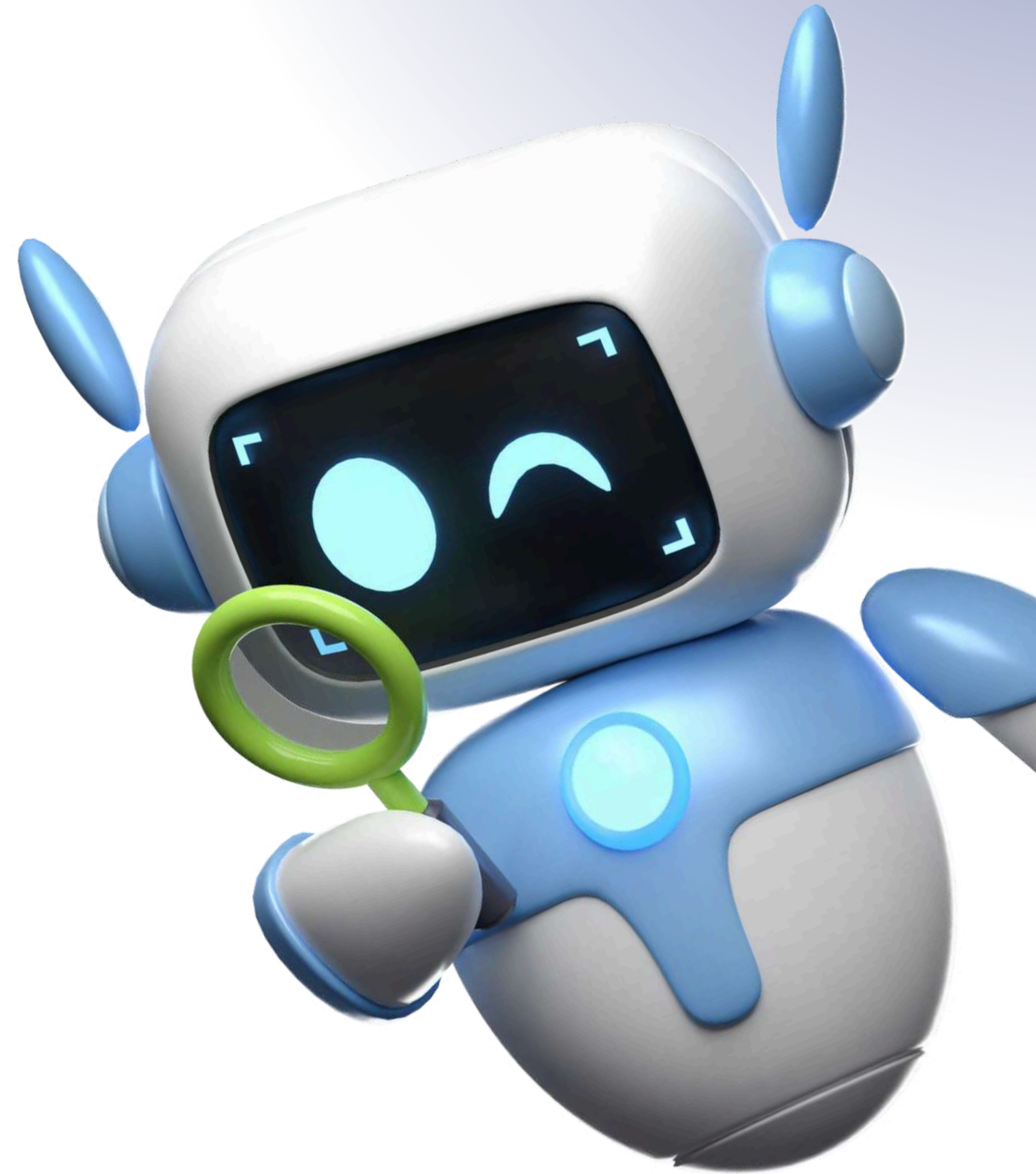
Tantangan Utama Lalu Lintas Jaringan

Keberagaman pola traffic
Kesulitan dalam persebaran distribusi

- Analisis Data
- Evaluasi Data
- Eksperimen Data

Analisis pada data dapat memberikan klasifikasi pola komunikasi dan potensi anomali di jaringan.

Exploratory Data Analytics (EDA)



Dataset 101

- id
- origin_host
- origin_port
- response_host
- response_port
- flow_duration
- forward_packets_per_sec
- backward_packets_per_sec
- flow_packets_per_sec
- down_up_ratio
- flow_FIN_flags
- flow_SYN_flags
- flow_RST_flags
- forward_PSH_flags
- backward_PSH_flags
- flow_ACK_flags
- forward_URG_flags
- backward_URG_flags
- flow_CWR_flags
- flow_ECE_flags
- forward_pkts_payload
- backward_pkts_payload
- flow_pkts_payload
- forward_iat
- backward_iat
- flow_iat
- payload_bytes_per_sec
- forward_subflow_packets
- backward_subflow_packets
- forward_subflow_bytes
- backward_subflow_bytes
- forward_bulk_bytes
- backward_bulk_bytes
- forward_bulk_packets
- backward_bulk_packets
- forward_bulk_rate

- backward_bulk_rate
- active
- idle
- forward_initial_window_size
- backward_initial_window_size
- forward_last_window_size
- traffic

Identifikasi Data

416.473 Rows & 43 Atribut

Integer

Float

Object

Background

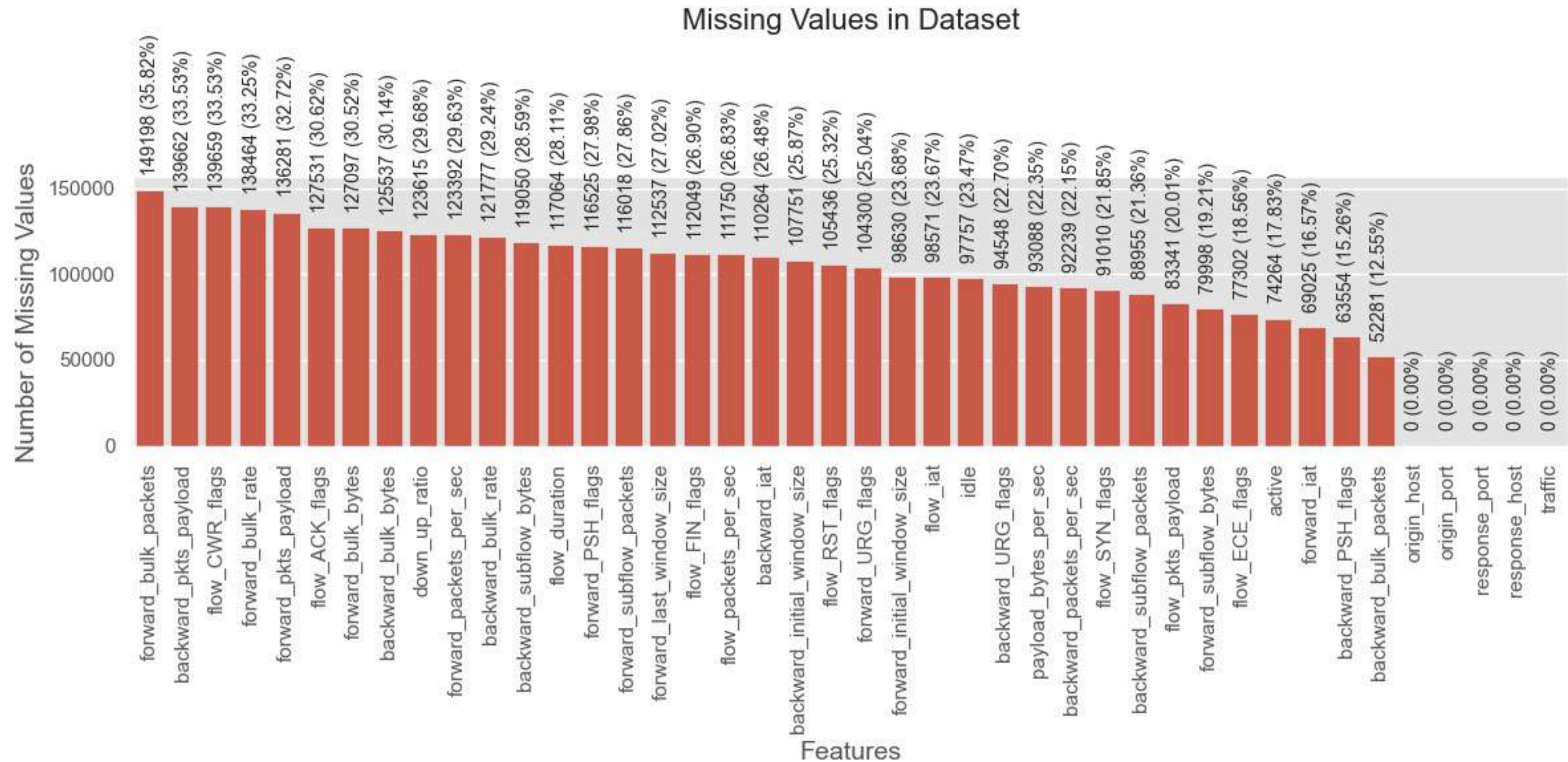
EDA

Data Preparation

Feature Engineering

Desain Eksperimen

Error Analysis



Dataset ini memiliki sejumlah fitur dengan missing values yang cukup signifikan

Background

EDA

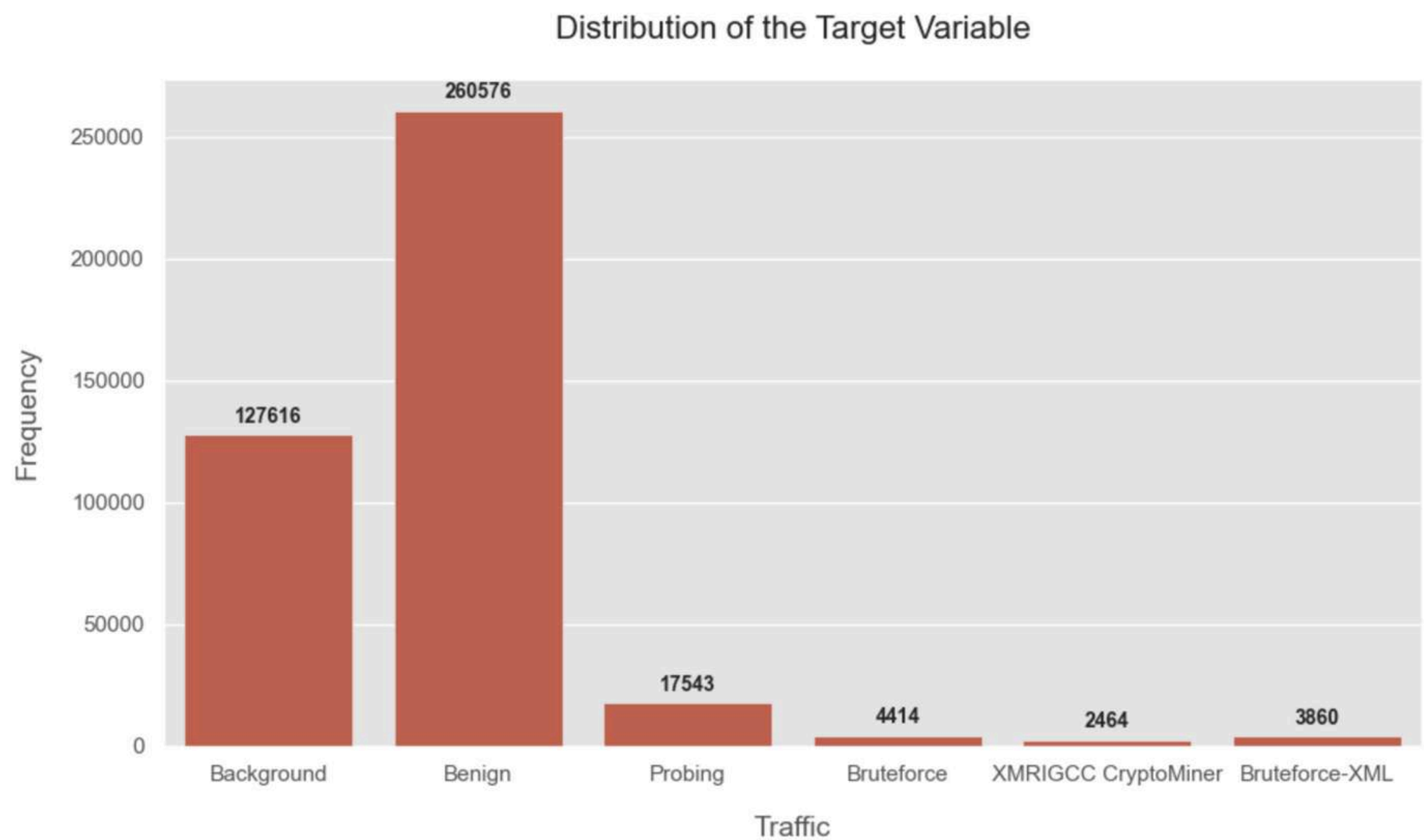
Data Preparation

Feature Engineering

Desain Eksperimen

Error Analysis

Faktor Yang Mempengaruhi Prediksi Model



Distribusi Target tidak merata

Distribusi kategori traffic pada dataset menunjukkan ketidakseimbangan yang mencolok, **Probing, Bruteforce, dan XMRIGCC CryptoMiner** memiliki jumlah instance yang **sangat sedikit** dibandingkan kategori **Benign dan Background**.

Ketidakseimbangan mempengaruhi **kemampuan model dalam mendeteksi traffic** berbahaya. Hal ini mempengaruhi **rendahnya sensitivitas model terhadap anomali**

Background

EDA

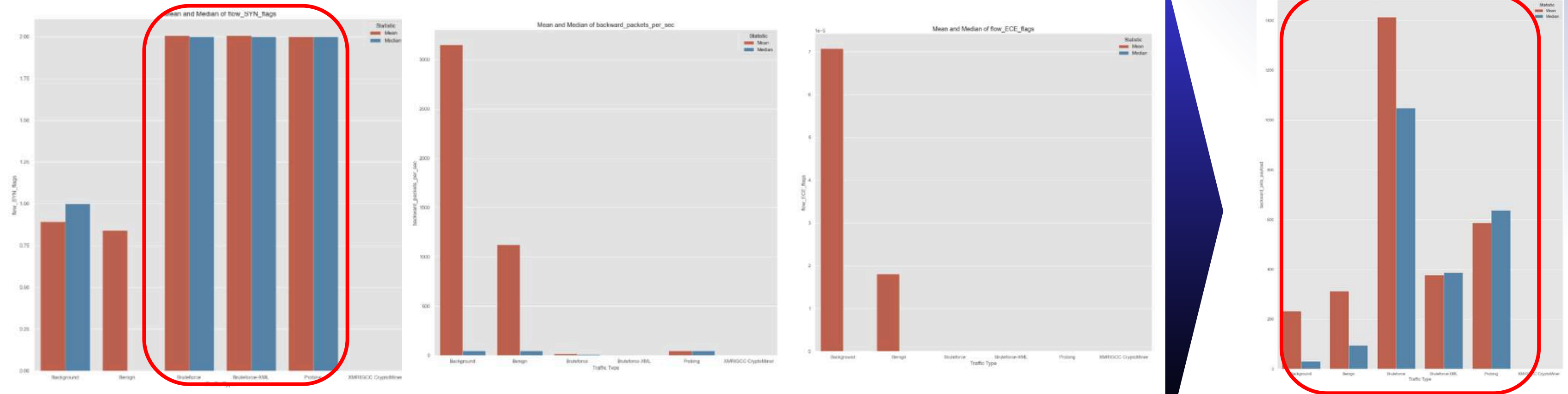
Data Preparation

Feature Engineering

Desain Eksperimen

Error Analysis

Faktor Yang Mempengaruhi Prediksi Model



Analisis distribusi mean dan median menunjukkan bahwa kesamaan nilai antar kategori traffic berisiko menyebabkan bias dalam prediksi. Selain itu, fitur dengan banyak nilai nol kurang efektif dalam membedakan traffic.

Background

EDA

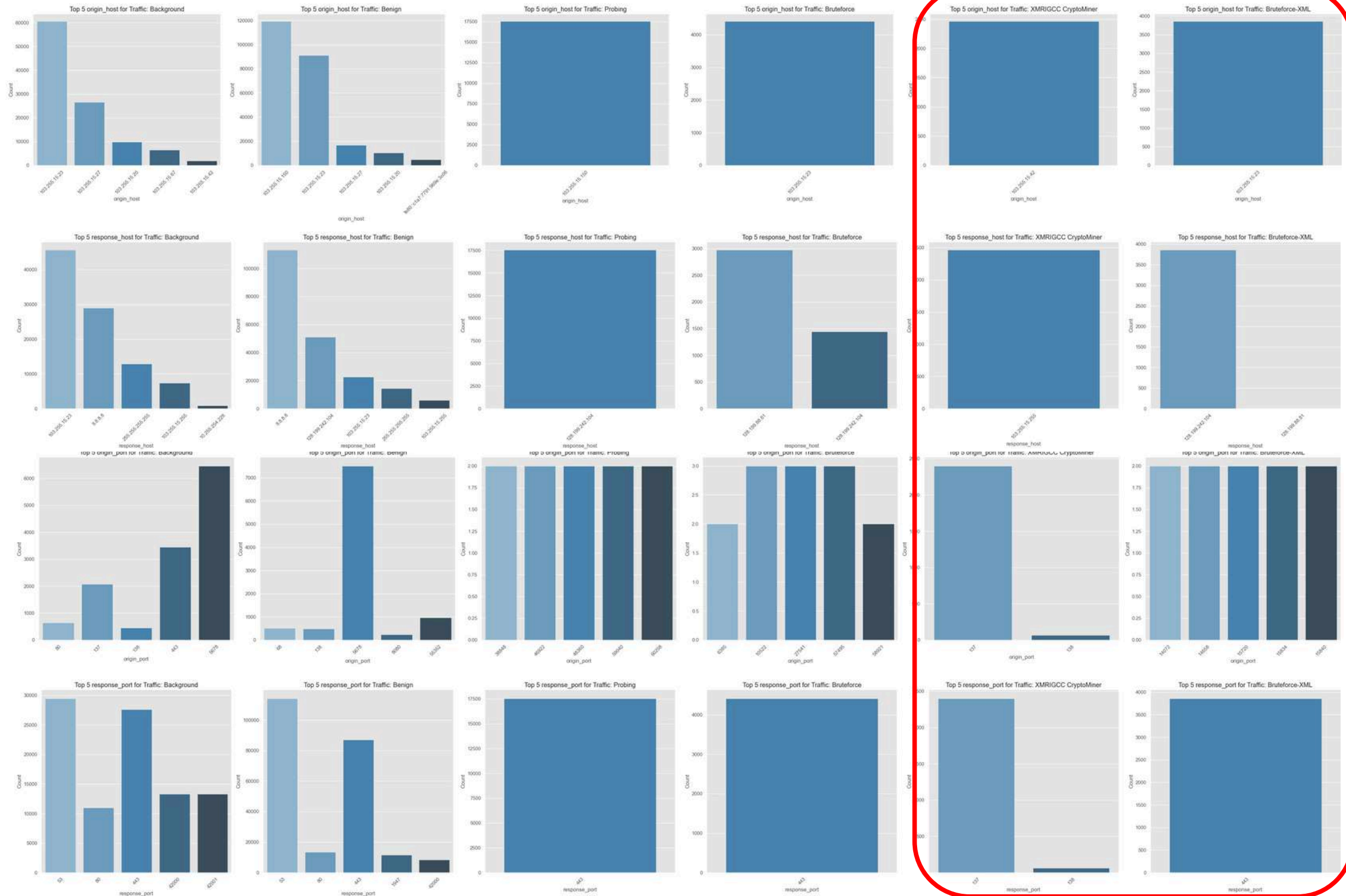
Data Preparation

Feature Engineering

Desain Eksperimen

Error Analysis

Faktor Yang Mempengaruhi Prediksi Model



Traffic attack seperti **Bruteforce dan XMRIGCC CryptoMiner** memperlihatkan pola serangan terfokus dengan penggunaan host dan port yang berulang, menargetkan IP tertentu. Variasi yang lebih rendah pada traffic serangan memudahkan model untuk mengidentifikasi pola serangan

Background

EDA

Data Preparation

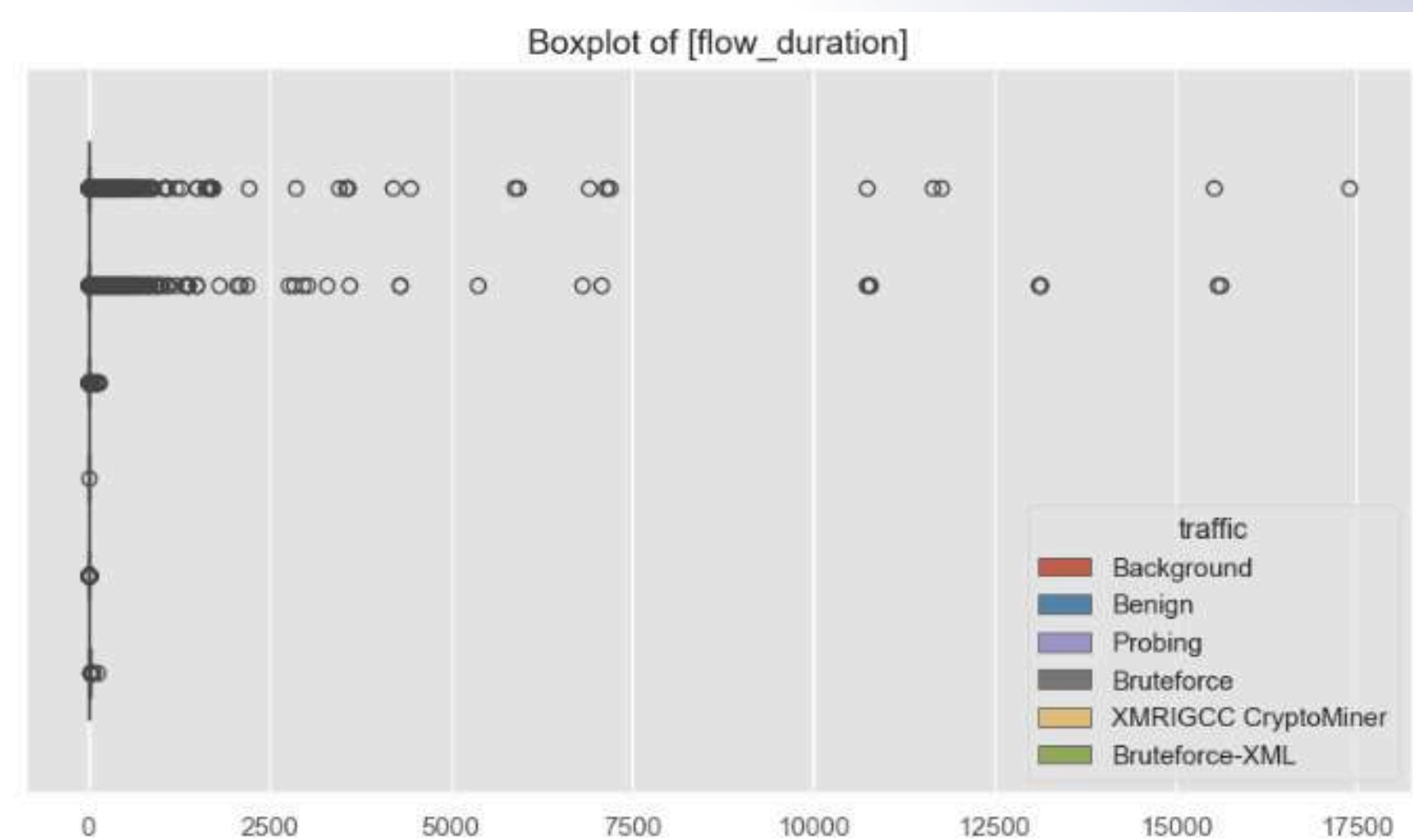
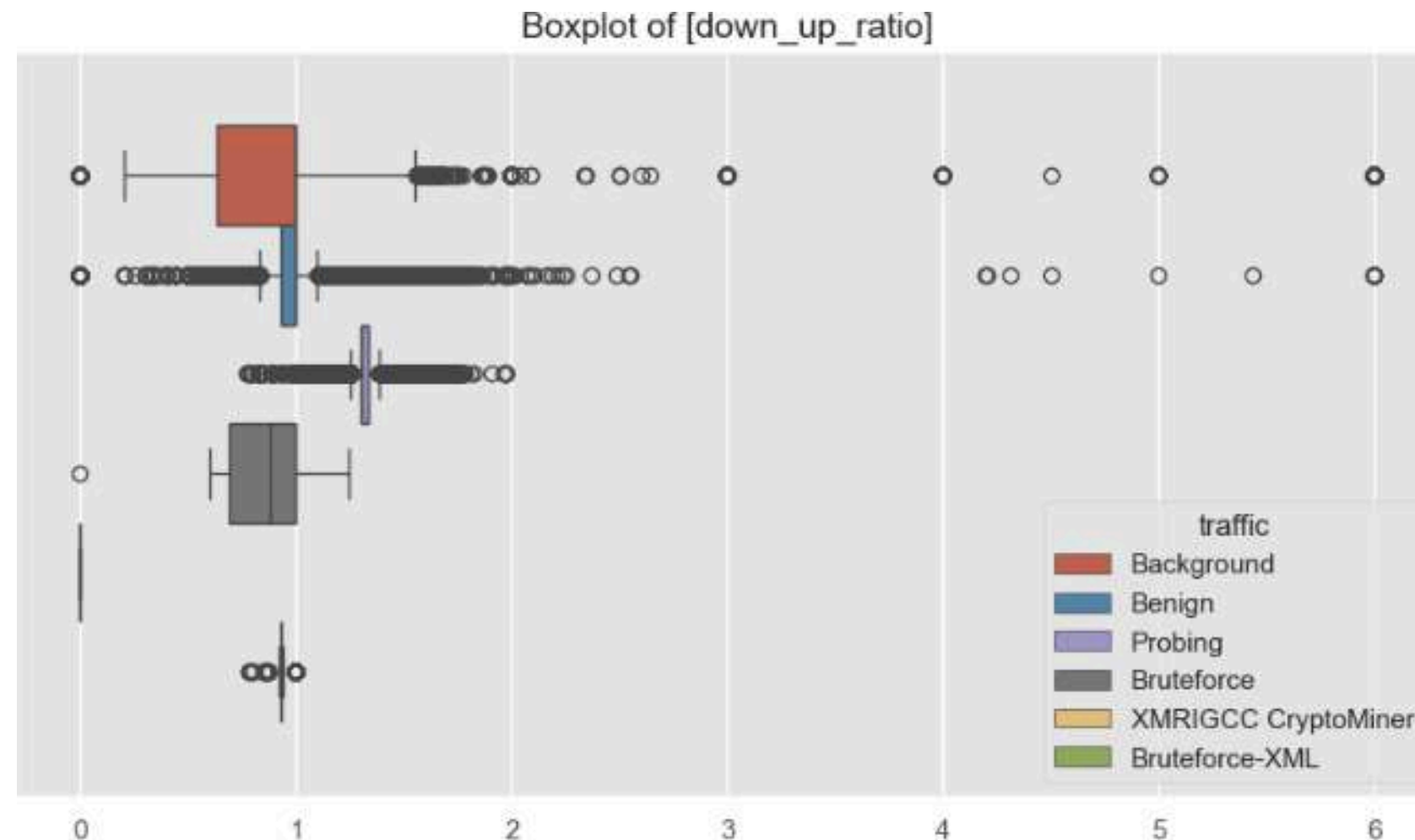
Feature Engineering

Desain Eksperimen

Error Analysis



Faktor Yang Mempengaruhi Prediksi Model



Boxplot menunjukkan outlier yang tidak dihapus karena penting untuk membedakan traffic normal dan serangan. Misalnya, fitur down_up_ratio dan flow_duration menunjukkan perbedaan signifikan pada serangan seperti Probing dan Bruteforce, membantu mengidentifikasi pola serangan melalui nilai ekstrem tersebut.

Background

EDA

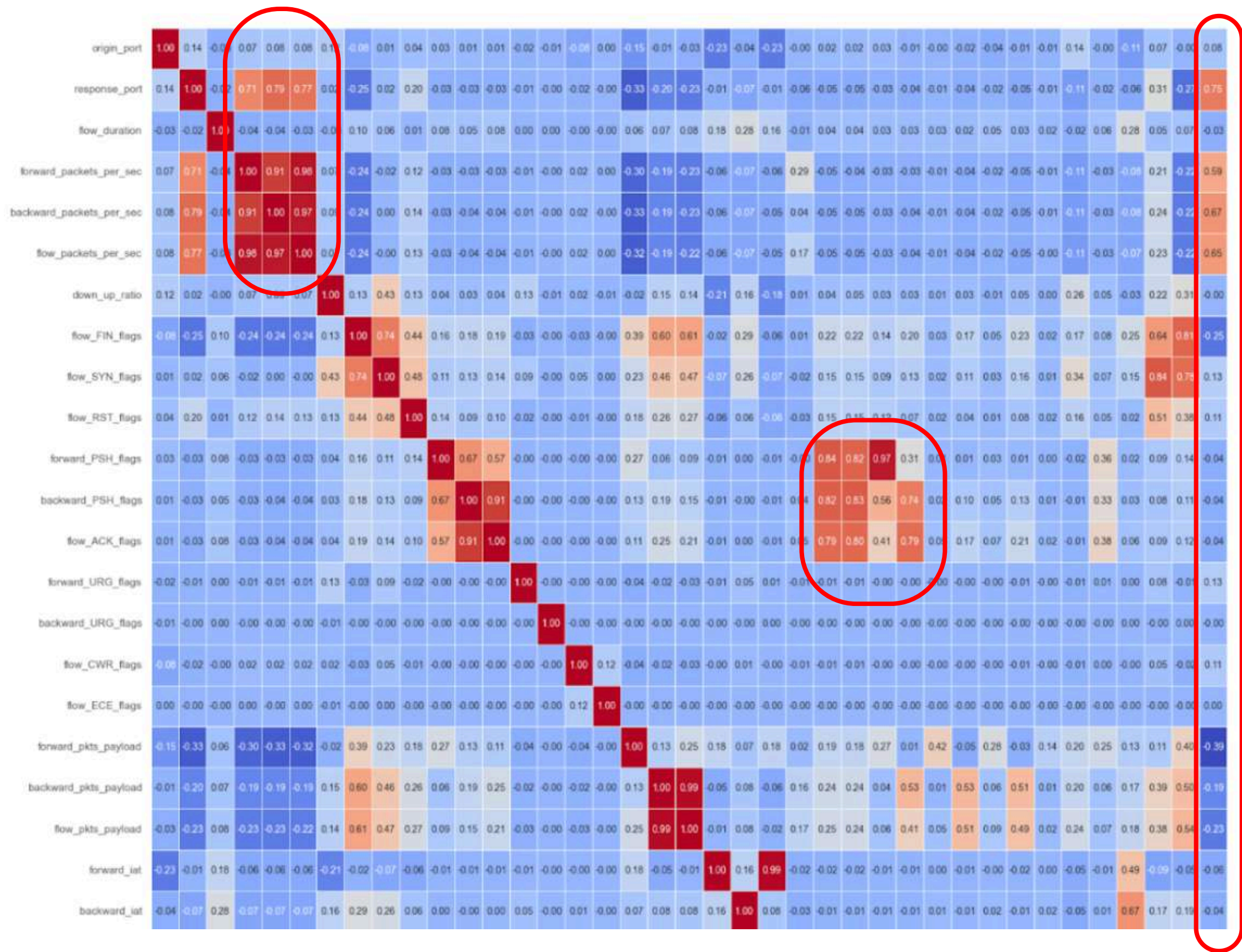
Data Preparation

Feature Engineering

Desain Eksperimen

Error Analysis

Correlation Matrix Antar Atribut dalam Dataset



Background

EDA

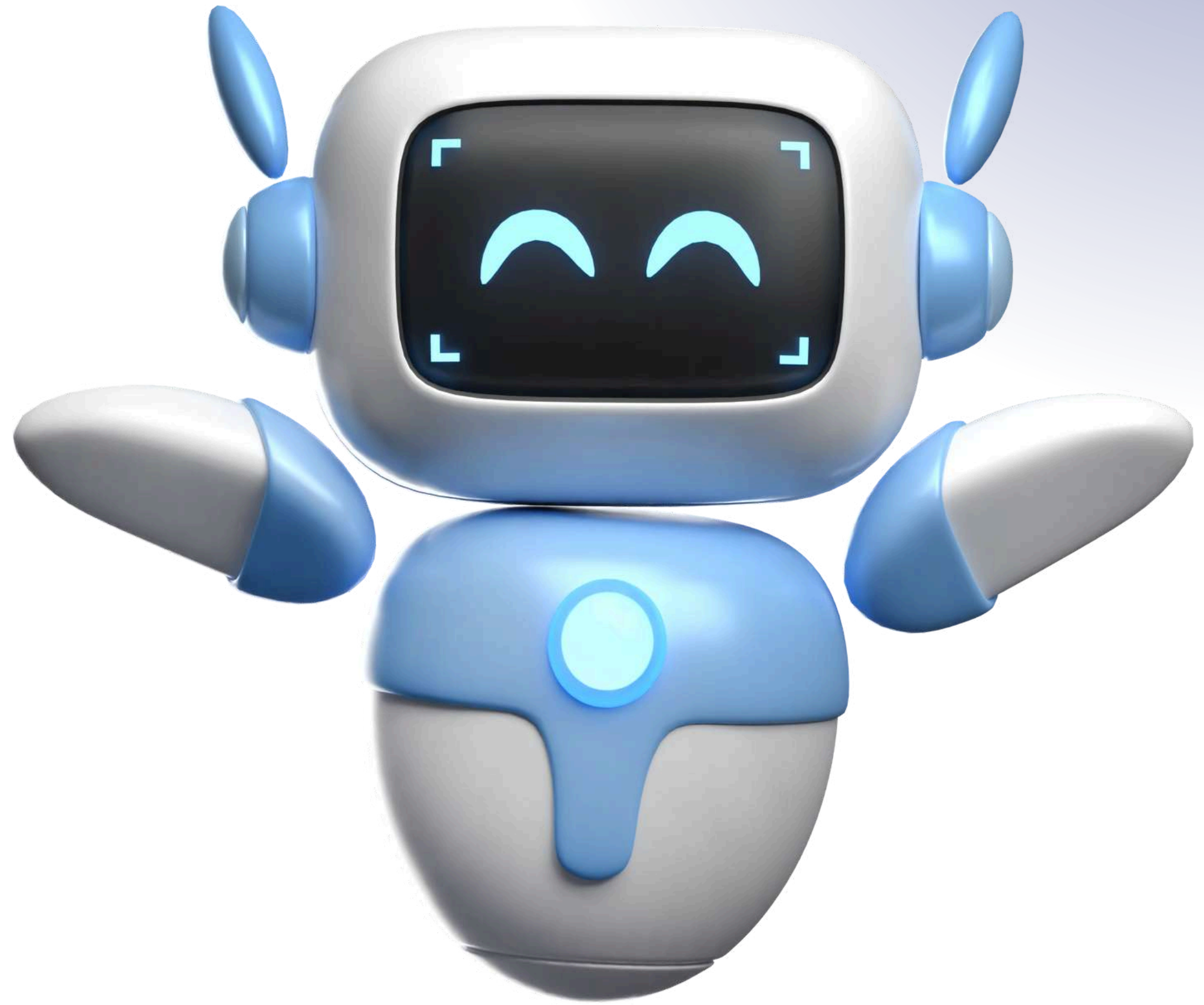
Data Preparation

Feature Engineering

Desain Eksperimen

Error Analysis

Data Preperation



5 Preprocessing Steps

Imputasi Forward, Backward, dan Flow Packets per Second

	forward_packets_per_sec	backward_packets_per_sec	flow_packets_per_sec
151776	0.000000	0.000000	0.000000
192474	13706.875817	13706.875817	27413.751634
244469	44.461235	57.799606	102.260842
372768	21290.883249	21290.883249	42581.766497



- Jika flow_packets_per_sec hilang:
$$\text{flow_packets_per_sec} = \begin{cases} 2 \times \text{forward_packets_per_sec} & \text{jika backward_packets_per_sec hilang} \\ 2 \times \text{backward_packets_per_sec} & \text{jika forward_packets_per_sec hilang} \\ \text{forward_packets_per_sec} + \text{backward_packets_per_sec} & \text{jika keduanya ada} \end{cases}$$
- Jika forward_packets_per_sec hilang:
$$\text{forward_packets_per_sec} = \begin{cases} \text{flow_packets_per_sec} - \text{backward_packets_per_sec} & \text{jika backward_packets_per_sec ada} \\ \text{backward_packets_per_sec} & \text{jika flow_packets_per_sec hilang} \end{cases}$$
- Jika backward_packets_per_sec hilang:
$$\text{backward_packets_per_sec} = \begin{cases} \text{flow_packets_per_sec} - \text{forward_packets_per_sec} & \text{jika forward_packets_per_sec ada} \\ \text{forward_packets_per_sec} & \text{jika flow_packets_per_sec hilang} \end{cases}$$

Imputasi Payloads Packets

	forward_pkts_payload	backward_pkts_payload	flow_pkts_payload
151776	201.0	0.0	201.000000
192474	0.0	0.0	0.000000
244469	83.9	637.0	396.521739
372768	0.0	0.0	0.000000



- Jika forward_pkts_payload atau backward_pkts_payload hilang:
$$\begin{aligned} \text{forward_pkts_payload} &= 2 \times \text{flow_pkts_payload} - \text{backward_pkts_payload} && \text{jika backward_pkts_payload ada} \\ \text{backward_pkts_payload} &= 2 \times \text{flow_pkts_payload} - \text{forward_pkts_payload} && \text{jika forward_pkts_payload ada} \\ \text{flow_pkts_payload} &= \frac{\text{forward_pkts_payload} + \text{backward_pkts_payload}}{2} && \text{jika keduanya ada} \end{aligned}$$

5 Preprocessing Steps

Imputasi Inter Arrival Time (IAT)

	forward_iat	backward_iat	flow_iat
151776	0.000000	0.000000	0.000000
192474	0.000000	0.000000	72.956085
244469	24990.558624	16342.163086	10223.410346
372768	0.000000	0.000000	46.968460



- Jika forward_iat atau backward_iat hilang:

$$\begin{aligned} \text{forward_iat} &= 4 \times \text{flow_iat} - \text{backward_iat} && \text{jika backward_iat ada} \\ \text{backward_iat} &= 4 \times \text{flow_iat} - \text{forward_iat} && \text{jika forward_iat ada} \\ \text{flow_iat} &= \frac{\text{forward_iat} + \text{backward_iat}}{4} && \text{jika keduanya ada} \end{aligned}$$

Imputasi Subflow Packets

	forward_subflow_packets	backward_subflow_packets
151776	1.0	0.0
192474	1.0	1.0
244469	10.0	13.0
372768	1.0	1.0



- Jika forward_subflow_packets hilang:

$$\text{forward_subflow_packets} = \begin{cases} \text{backward_subflow_packets} \times \text{down_up_ratio} & \text{jika backward_subflow_packets dan down_up_ratio ada} \\ 0 & \text{jika down_up_ratio} \leq 0 \text{ atau tidak ada data} \end{cases}$$

- Jika backward_subflow_packets hilang:

$$\text{backward_subflow_packets} = \begin{cases} \frac{\text{forward_subflow_packets}}{\text{down_up_ratio}} & \text{jika down_up_ratio} > 0 \text{ dan forward_subflow_packets ada} \\ 0 & \text{jika down_up_ratio} \leq 0 \text{ atau tidak ada data} \end{cases}$$

Imputasi Flow Duration

	forward_iat	backward_iat	flow_iat
151776	0.000000	0.000000	0.000000
192474	0.000000	0.000000	72.956085
244469	24990.558624	16342.163086	10223.410346
372768	0.000000	0.000000	46.968460



- Konversi nilai idle dan active:

$$\text{idle} = \frac{\text{idle}}{1,000,000}$$

$$\text{active} = \frac{\text{active}}{1,000,000}$$

- Jika flow_duration hilang:

$$\text{flow_duration} = \text{active} + \text{idle}$$

- Jika active hilang:

$$\text{active} = \text{flow_duration} - \text{idle}$$

- Jika idle hilang:

$$\text{idle} = \text{flow_duration} - \text{active}$$

Background

EDA

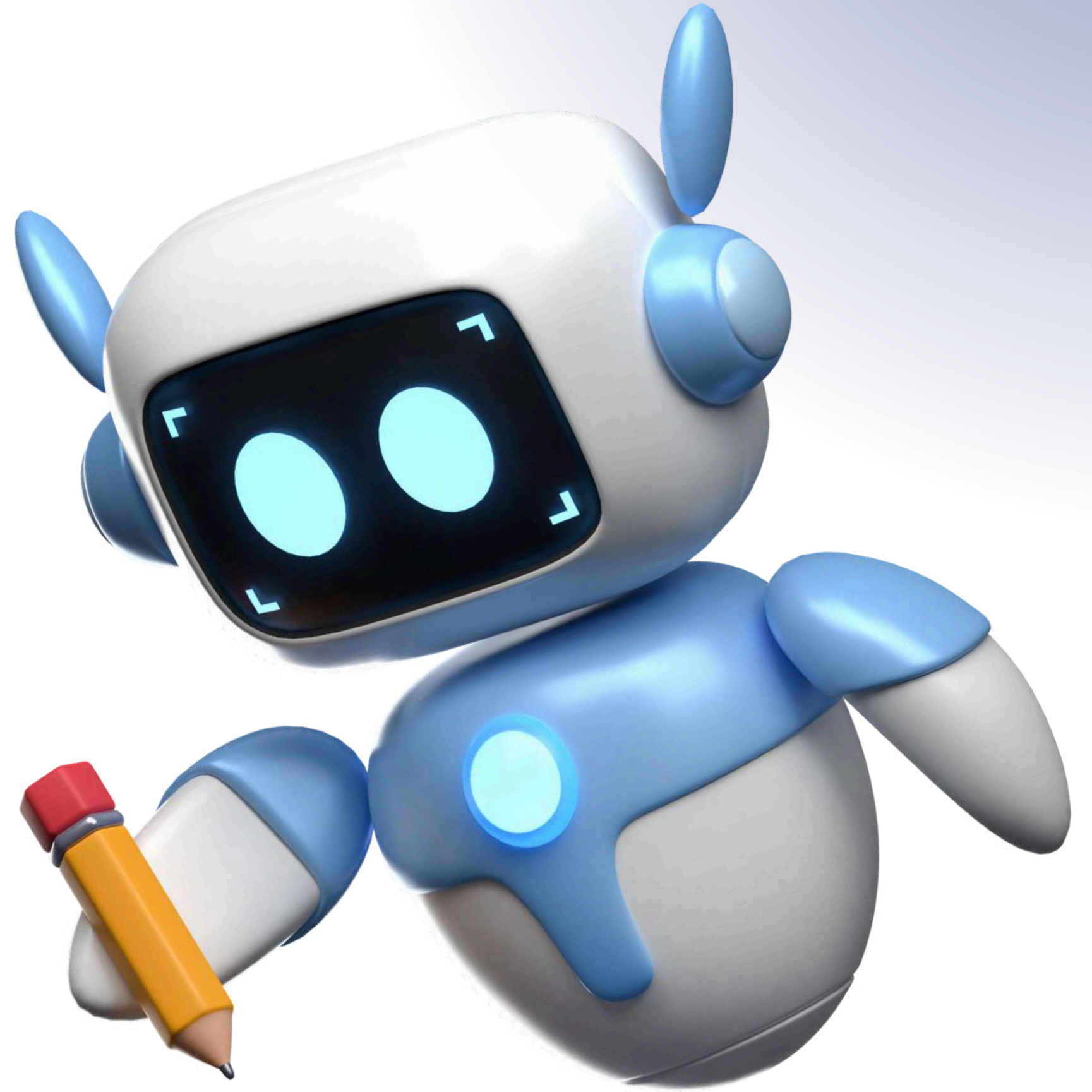
Data Preparation

Feature Engineering

Desain Eksperimen

Error Analysis

Feature Engineering



Feature Engineering

Konversi Fitur Tipe Object -> Kategori

- 1 Memungkinkan model memahami fitur sebagai kategori untuk efisiensi interpretasi

#	Column	Non-Null Count		Dtype
---	-----	-----		-----
0	origin_host	416473	non-null	object
1	origin_port	416473	non-null	int64
2	response_host	416473	non-null	object
3	response_port	416473	non-null	int64

Konversi origin_port dan response_port ke Kategori

- 1 Mengubah tipe data menjadi category agar model memperlakukan setiap port sebagai kelas yang berbeda, bukan sebagai angka kontinu.

Data columns (total 40 columns):				
#	Column	Non-Null Count		Dtype
---	-----	-----		-----
0	origin_host	416473	non-null	category
1	origin_port	416473	non-null	category
2	response_host	416473	non-null	category
3	response_port	416473	non-null	category

Feature Creation & Encoding

Binary Feature for PSH Flags (is_PSH)

- 1 Berfungsi sebagai indikator potensi serangan atau pola penggunaan jaringan yang tidak biasa.

Kombinasi Host dan Port

- 1 Memudahkan model mendeteksi pola tersembunyi yang bisa membedakan jenis traffic, termasuk serangan

Host-to-Port Ratio

- 1 Mendeteksi apakah sebuah host berinteraksi dengan banyak port (aktivitas normal) atau terbatas pada port tertentu (indikator serangan).

Frequency Encoding

- 1 Menerapkan frequency encoding pada origin_host dan response_host.
- 2 Menangkap popularitas atau frekuensi penggunaan host.

Port Service Type:

- 1 Mengklasifikasikan origin_port dan response_port berdasarkan layanan umum (HTTP, HTTPS, FTP, SSH).
- 2 Membantu model memahami jenis komunikasi jaringan dan memisahkan traffic normal dari traffic berbahaya.

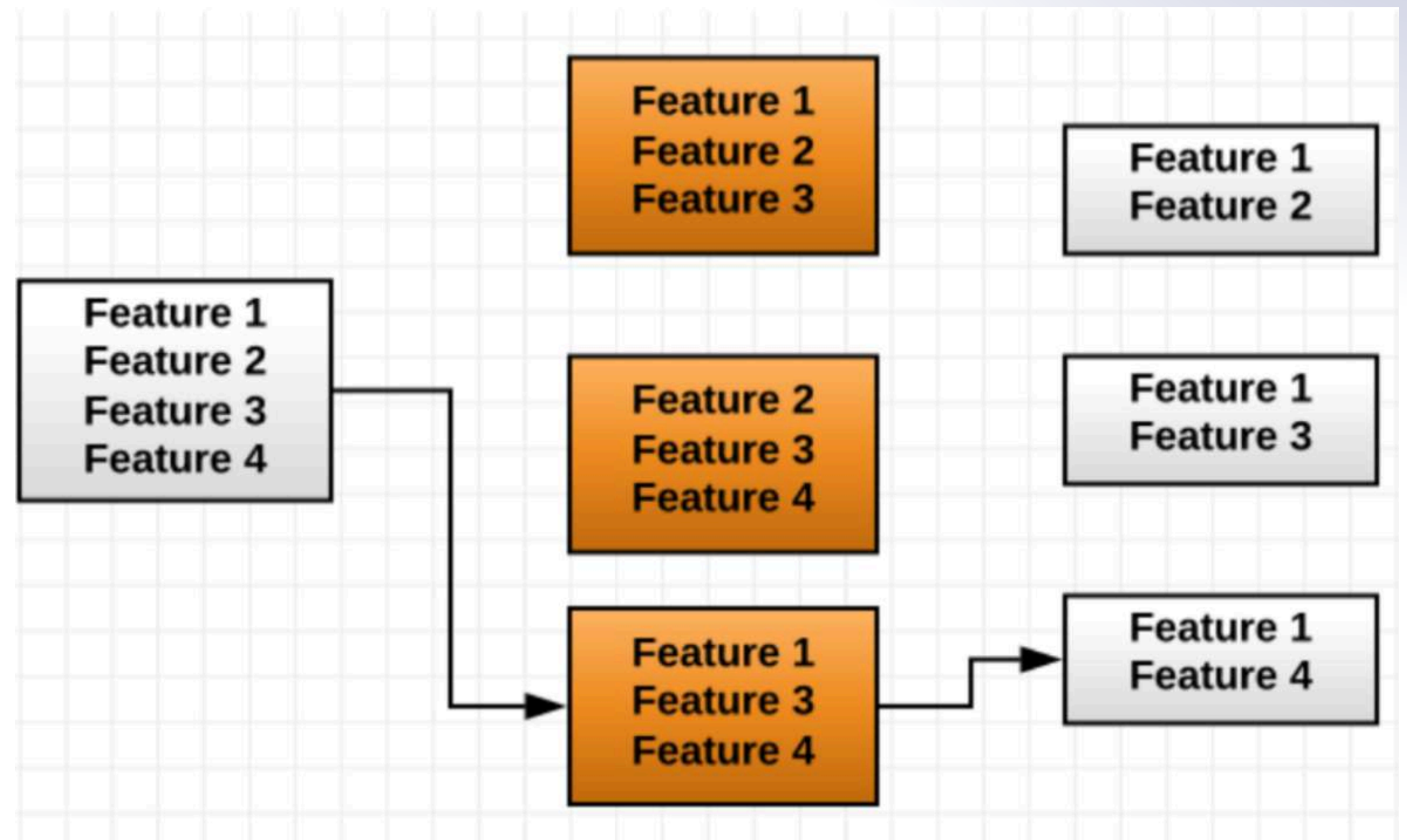
One Hot port service type

- 1 Dilakukan one-hot encoding pada fitur port service type (contoh: HTTP, HTTPS, FTP, SSH).
- 2 Setiap kategori pada origin_port_service_type dan response_port_service_type diubah menjadi kolom biner terpisah.

Feature Selection & Encoding

Backward Feature Elimination

- 1 Hanya mempertahankan fitur-fitur yang benar-benar penting untuk membuat dataset lebih ringan dan meningkatkan efisiensi prediksi.



Desain Eksperimen



Tujuan dan Skenario Eksperimen

Tujuan:

1. Mengetahui metode terbaik menangani missing value
2. Mengetahui teknik terbaik menangani data imbalance .
3. Mengetahui algoritma terbaik berdasarkan karakteristik dan kebutuhan data.

Skenario:

Penanganan missing value:

- 1 Baseline
- 2 Mean
- 3 Median

Penanganan data imbalance:

- 1 Baseline
- 2 Class Weight

Algoritma:

- 1 CatBoost
- 2 LightGBM

Validation Technique:

1. Stratified K-Fold 5 Fold
2. Split 80% Training Data, 20% Validation Data

Metrics Evaluasi:

$$\text{Accuracy} = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(y_{\text{true},i} = y_{\text{pred},i})$$

$$\text{Balanced Accuracy} = \frac{1}{C} \sum_{c=1}^C \frac{TP_c}{TP_c + FN_c}$$

$$\text{Score} = \frac{\text{Accuracy} + \text{Balanced Accuracy}}{2}$$

Background

EDA

Data Preparation

Feature Engineering

Desain Eksperimen

Error Analysis

Hasil Eksperimen

Mean Imputed							
Model	Fold					Mean	Std
	1	2	3	4	5		
CatBoost Baseline	0.8126	0.8147	0.8116	0.8133	0.8137	0.8132	0.001095
CatBoost ClassWeight	0.8640	0.8644	0.8629	0.8622	0.8622	0.8631	0.000839
LightGBM Baseline	0.7217	0.7154	0.7291	0.7255	0.7311	0.7245	0.005712
LightGBM ClassWeight	0.8267	0.8282	0.8235	0.8254	0.8254	0.8258	0.001775

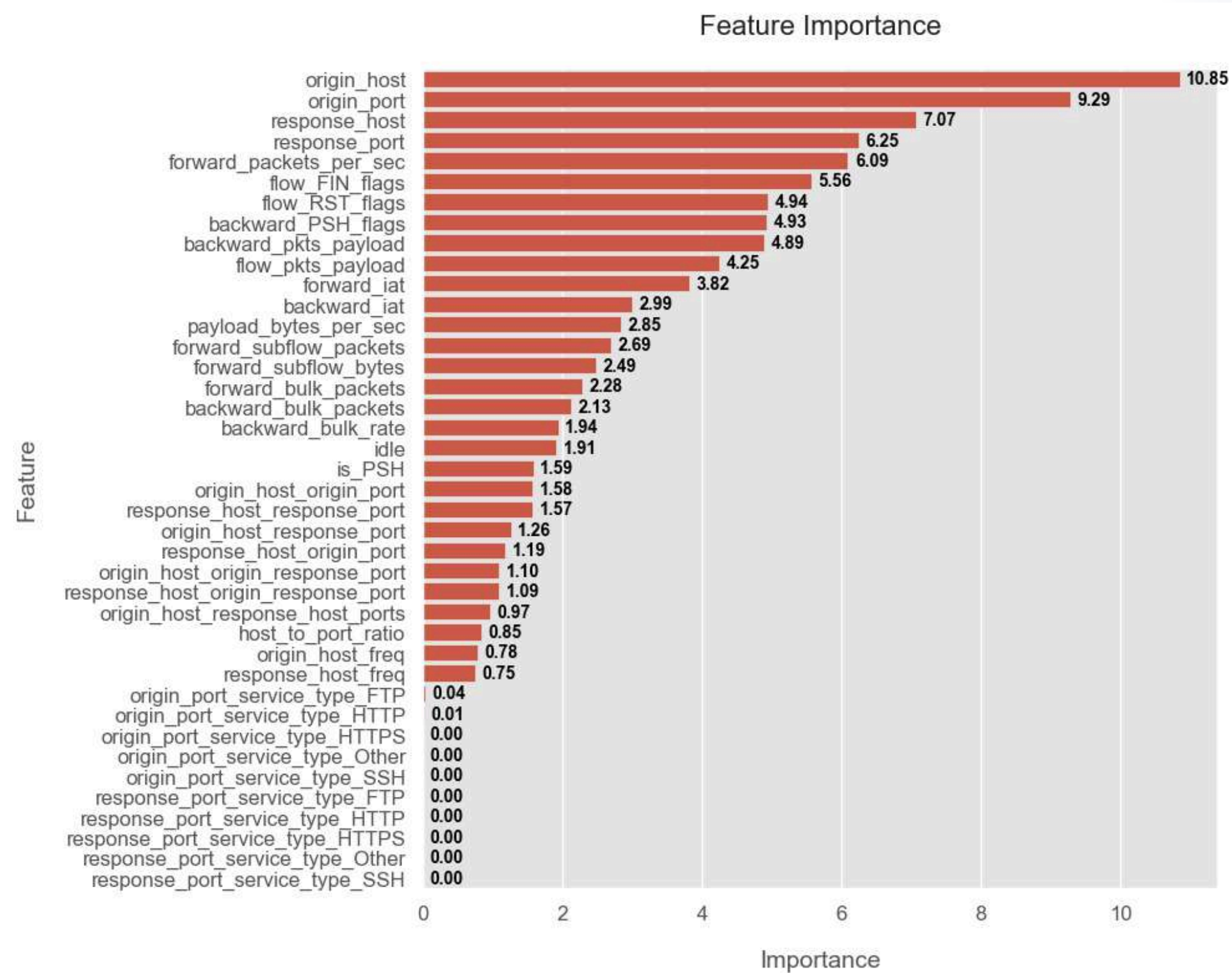
Median Imputed							
Model	Fold					Mean	Std
	1	2	3	4	5		
CatBoost Baseline	0.8035	0.8058	0.8017	0.8050	0.8040	0.8040	0.001542
CatBoost ClassWeight	0.8575	0.8578	0.8564	0.8562	0.8565	0.8569	0.000576
LightGBM Baseline	0.7160	0.7182	0.7195	0.7237	0.7148	0.7185	0.003167
LightGBM ClassWeight	0.8209	0.8214	0.8190	0.8196	0.8180	0.8198	0.001381

Hasil Eksperimen

Baseline							
Model	Fold					Mean	Std
	1	2	3	4	5		
CatBoost Baseline	0.8107	0.8140	0.8131	0.8123	0.8127	0.8126	0.001157
CatBoost ClassWeight	0.8645	0.8649	0.8627	0.8629	0.8626	0.8635	0.000853
LightGBM Baseline	0.7299	0.7331	0.7270	0.7380	0.7301	0.7316	0.004065
LightGBM ClassWeight	0.8274	0.8281	0.8240	0.8273	0.8260	0.8266	0.001558



Model Interpretability



Background

EDA

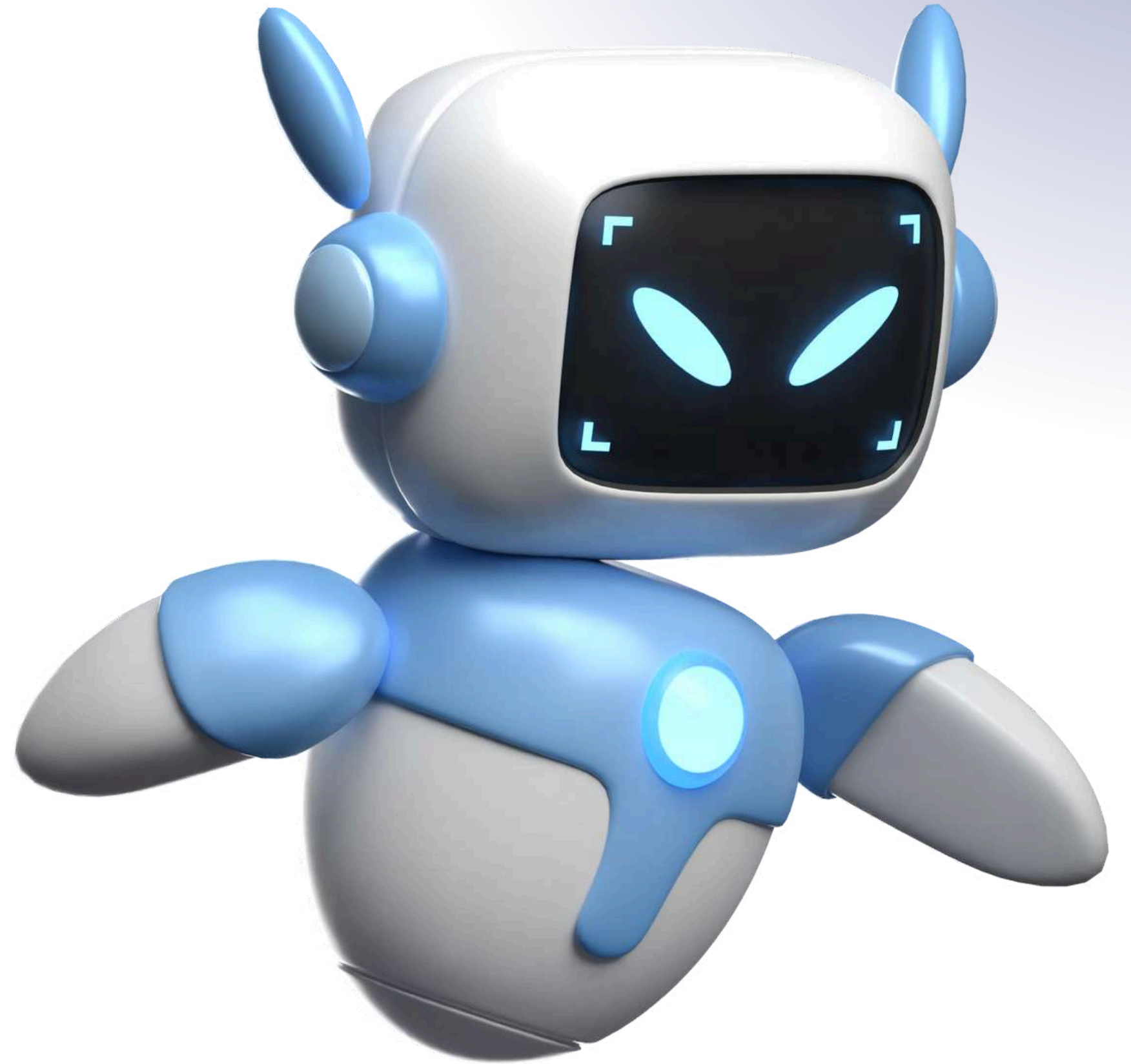
Data Preparation

Feature Engineering

Desain Eksperimen

Error Analysis

Error Analysis

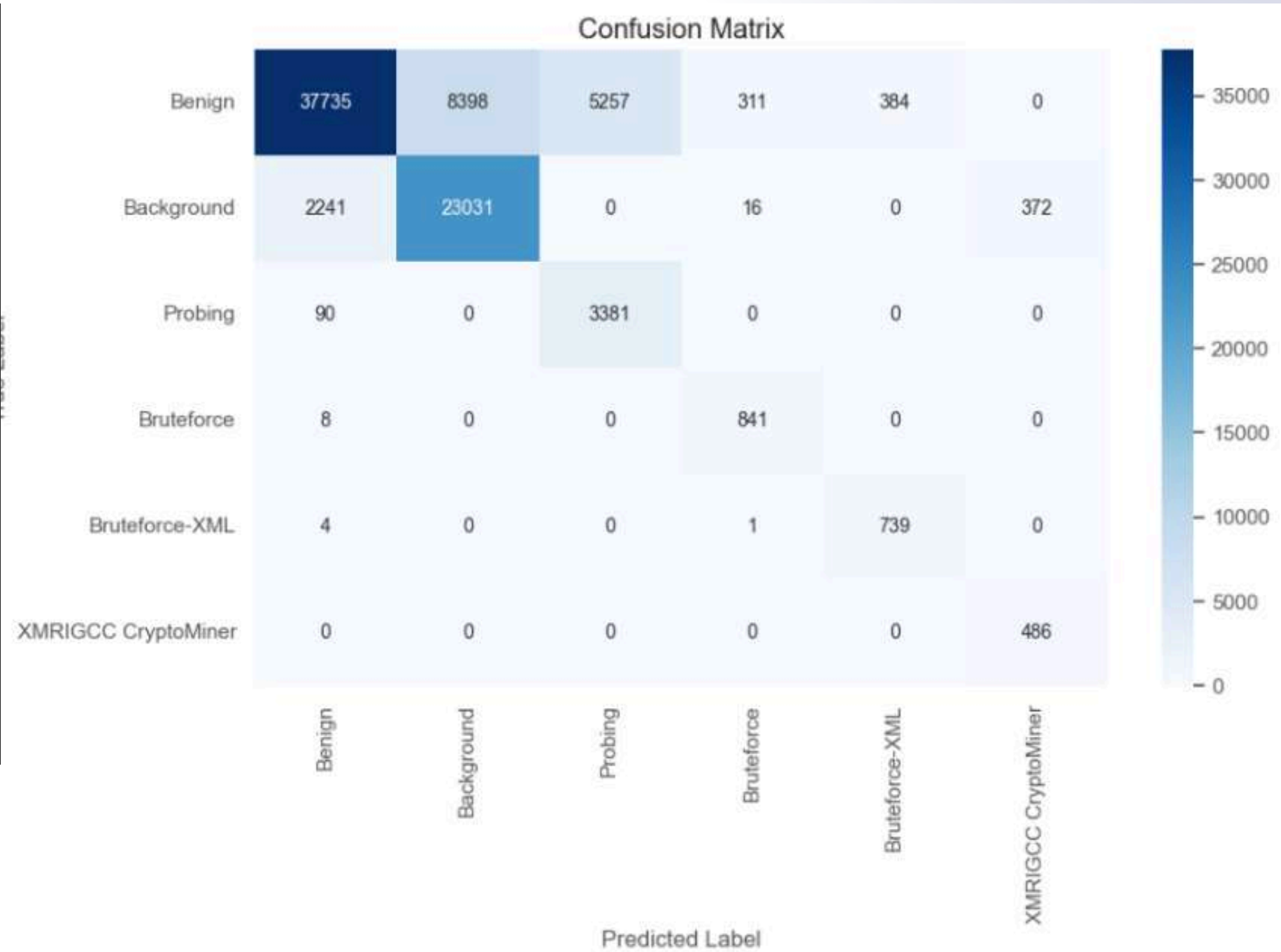


Hasil Analisis

Classification Report:

Score: 0.8624576007844553				
Balanced Accuracy: 0.9299935376034866				
Accuracy: 0.7949216639654241				
	precision	recall	f1-score	support
Background	0.73	0.90	0.81	25660
Benign	0.94	0.72	0.82	52085
Bruteforce	0.72	0.99	0.83	849
Bruteforce-XML	0.66	0.99	0.79	744
Probing	0.39	0.97	0.56	3471
XMRIGCC CryptoMiner	0.57	1.00	0.72	486
accuracy			0.79	83295
macro avg	0.67	0.93	0.76	83295
weighted avg	0.85	0.79	0.80	83295

Confussion Matrix:



Background

EDA

Data Preparation

Feature Engineering

Desain Eksperimen

Error Analysis



Threshold Tuning

```
Score: 0.8624576007844553
Balanced Accuracy: 0.9299935376034866
Accuracy: 0.7949216639654241
```

	precision	recall	f1-score	support
Background	0.73	0.90	0.81	25660
Benign	0.94	0.72	0.82	52085
Bruteforce	0.72	0.99	0.83	849
Bruteforce-XML	0.66	0.99	0.79	744
Probing	0.39	0.97	0.56	3471
XMRIGCC CryptoMiner	0.57	1.00	0.72	486
accuracy			0.79	83295
macro avg	0.67	0.93	0.76	83295
weighted avg	0.85	0.79	0.80	83295

```
Best Thresholds: [0.37, 0.18, 0.75, 0.66, 0.66, 0.01]
Score: 0.8721754687399257
Balanced Accuracy: 0.9182149149094689
Accuracy: 0.8261360225703823
```

	precision	recall	f1-score	support
Background	0.78	0.83	0.81	25660
Benign	0.91	0.81	0.86	52085
Bruteforce	0.75	0.99	0.86	849
Bruteforce-XML	0.69	0.99	0.81	744
Probing	0.49	0.88	0.63	3471
XMRIGCC CryptoMiner	0.57	1.00	0.72	486
accuracy			0.83	83295
macro avg	0.70	0.92	0.78	83295
weighted avg	0.85	0.83	0.83	83295

Background

EDA

Data Preparation

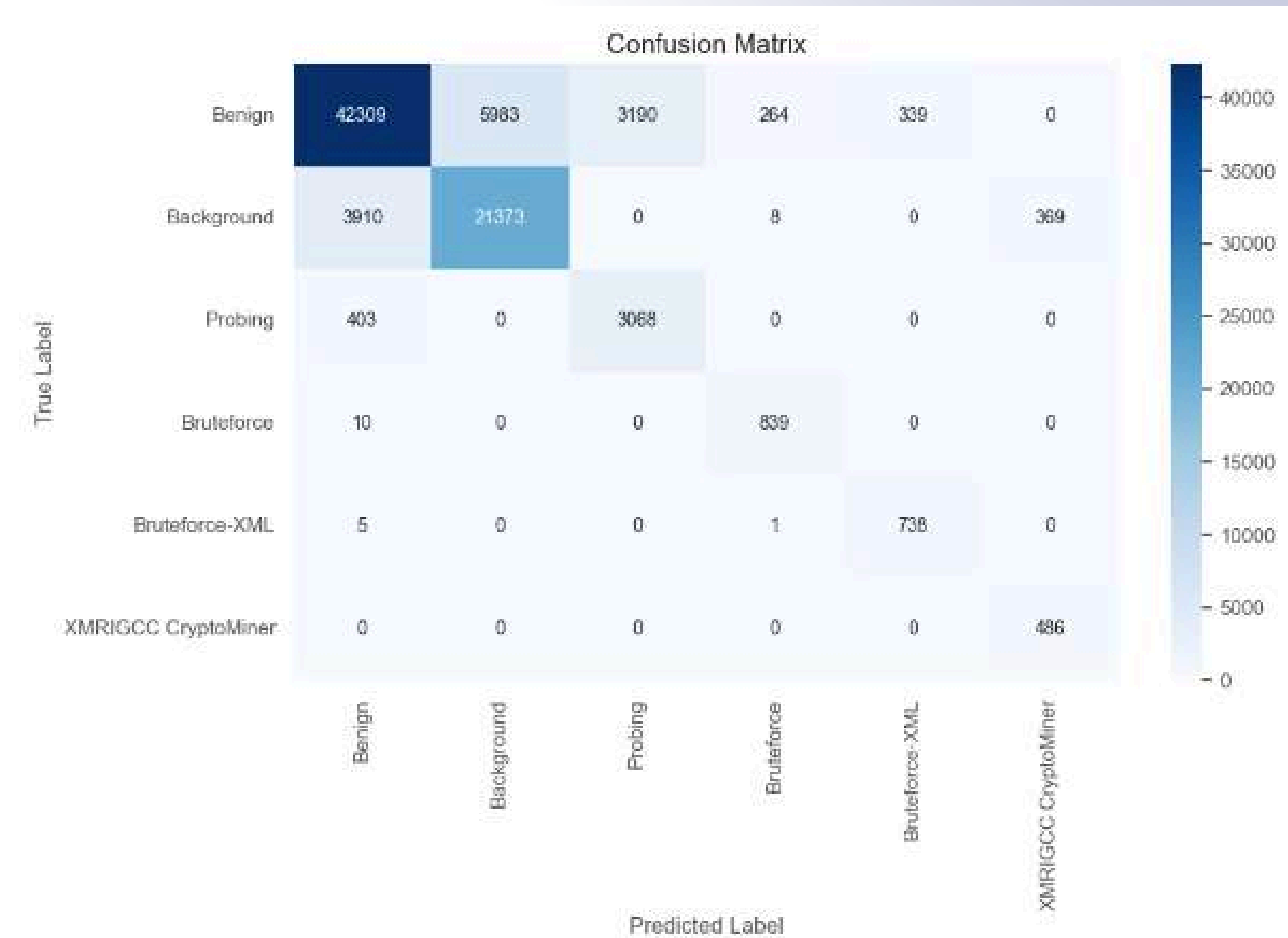
Feature Engineering

Desain Eksperimen

Error Analysis



Threshold Tuning



Background

EDA

Data Preparation

Feature Engineering

Desain Eksperimen

Error Analysis



Future Enhancement

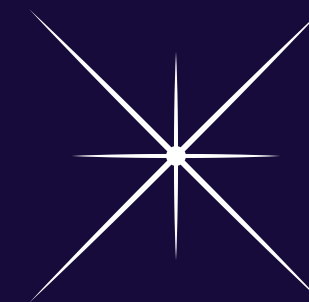
1. Model masih bergantung pada atribut statis seperti host dan port.
2. Kualitas data belum optimal, terdapat banyak nilai NaN.
3. Eksperimen untuk menangani ketidakseimbangan kelas terbatas karena kualitas data.
4. Sulit untuk menggeneralisasi prediksi jika terdapat kombinasi host dan port baru.

1. Meminimalkan nilai NaN untuk menghasilkan data yang lebih bersih.
2. Lakukan eksperimen tambahan seperti undersampling, SMOTE, dan Tomek Link.
3. Tambahkan fitur baru seperti timestamp atau informasi lain untuk memperbaiki pemisahan antar kelas.
4. Eksplorasi fitur yang lebih general, seperti pola trafik temporal atau indikator perilaku anomali, untuk mengurangi ketergantungan pada host dan port.



Kesimpulan

Penelitian ini mengeksplorasi **deteksi serangan** pada traffic jaringan menggunakan **feature engineering** dan **model klasifikasi** seperti **CatBoost** dan **LightGBM**. Hasilnya menunjukkan performa baik dalam **membedakan** traffic normal dan berbahaya, terutama dengan penggunaan **class weight** dan **imputasi baseline** yang meningkatkan **custom scoring**. **Threshold tuning** juga berhasil memaksimalkan prediksi, meskipun masih ada tantangan dalam membedakan kelas Probing dan CryptoMiner. Fitur **host, port, flag, IAT, dan packets** terbukti penting dalam **prediksi**, sementara fitur dengan importance rendah atau nol kurang berkontribusi dan perlu dievaluasi lebih lanjut.



TERIMA KASIH

“Mengubah Ketidaktahuan Menjadi Pengetahuan”