Menghadapi Tantangan Keamanan Era Digital: Pendekatan Pembelajaran Mesin untuk Analisis Lalu Lintas Jaringan

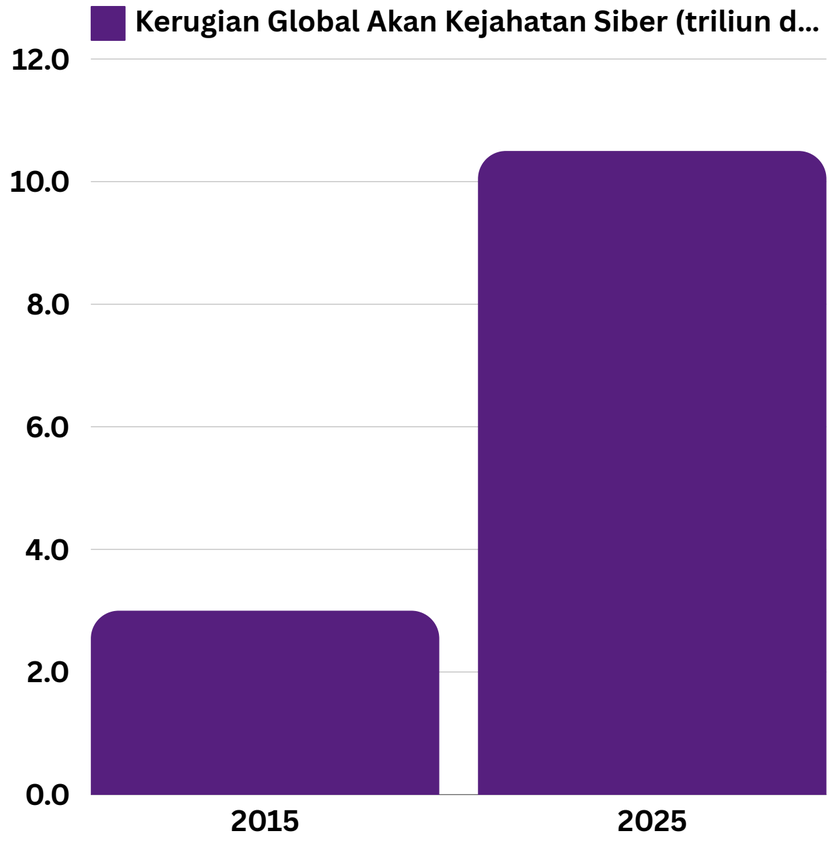
Timur Tengah

Anthony Edbert Feriyanto
Universitas Indonesia

Filbert Aurelian Tjiaranata
Universitas Indonesia

M. Arvin Wijayanto
Universitas Indonesia

Latar Belakang



"Lebih dari 8,5 miliar catatan data terekspos pada tahun 2019 saja, dengan 71% serangan yang berhasil dimotivasi secara finansial"
– *IBM X-Force Threat Intelligence Index*

"Rata-rata waktu yang dibutuhkan untuk mengidentifikasi dan mengatasi pelanggaran data adalah 280 hari"
– *Ponemon Institute*

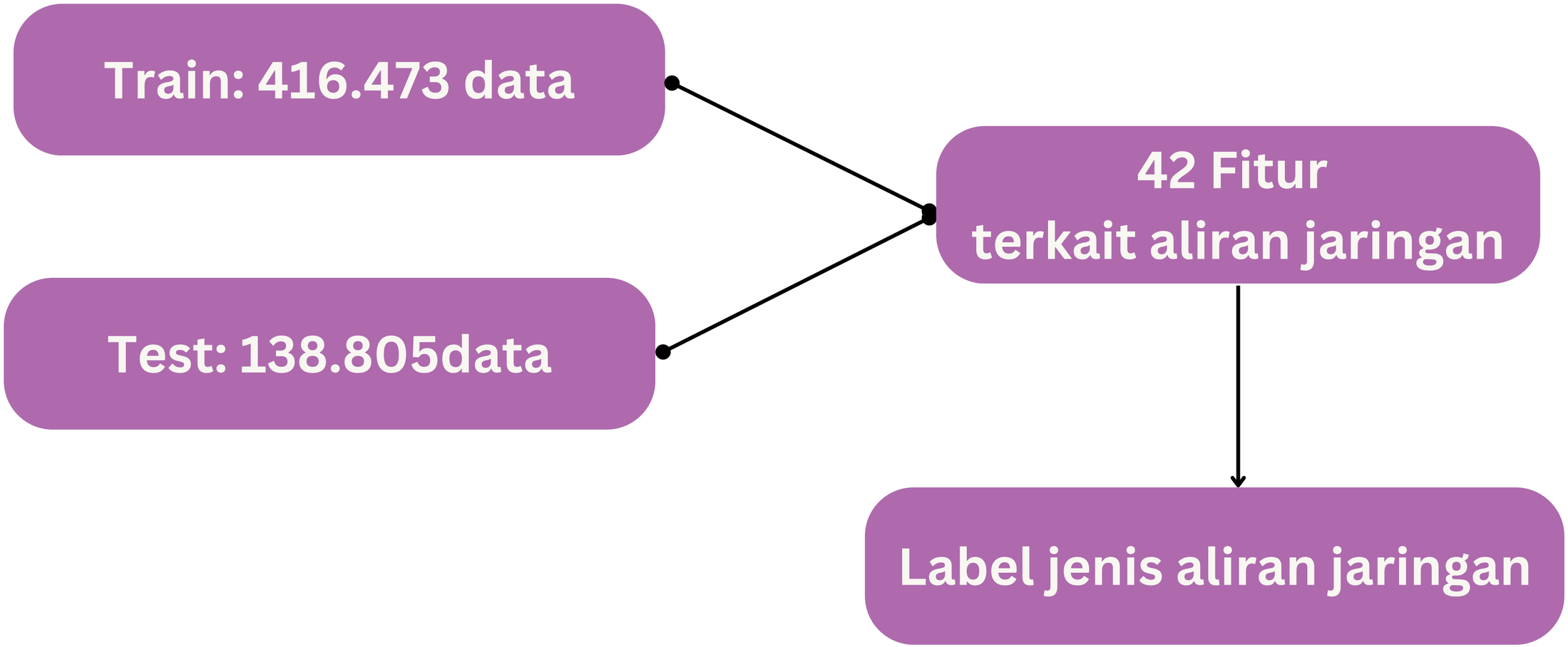
Tujuan

DATA QUEST
PART OF AIRNOLOGY

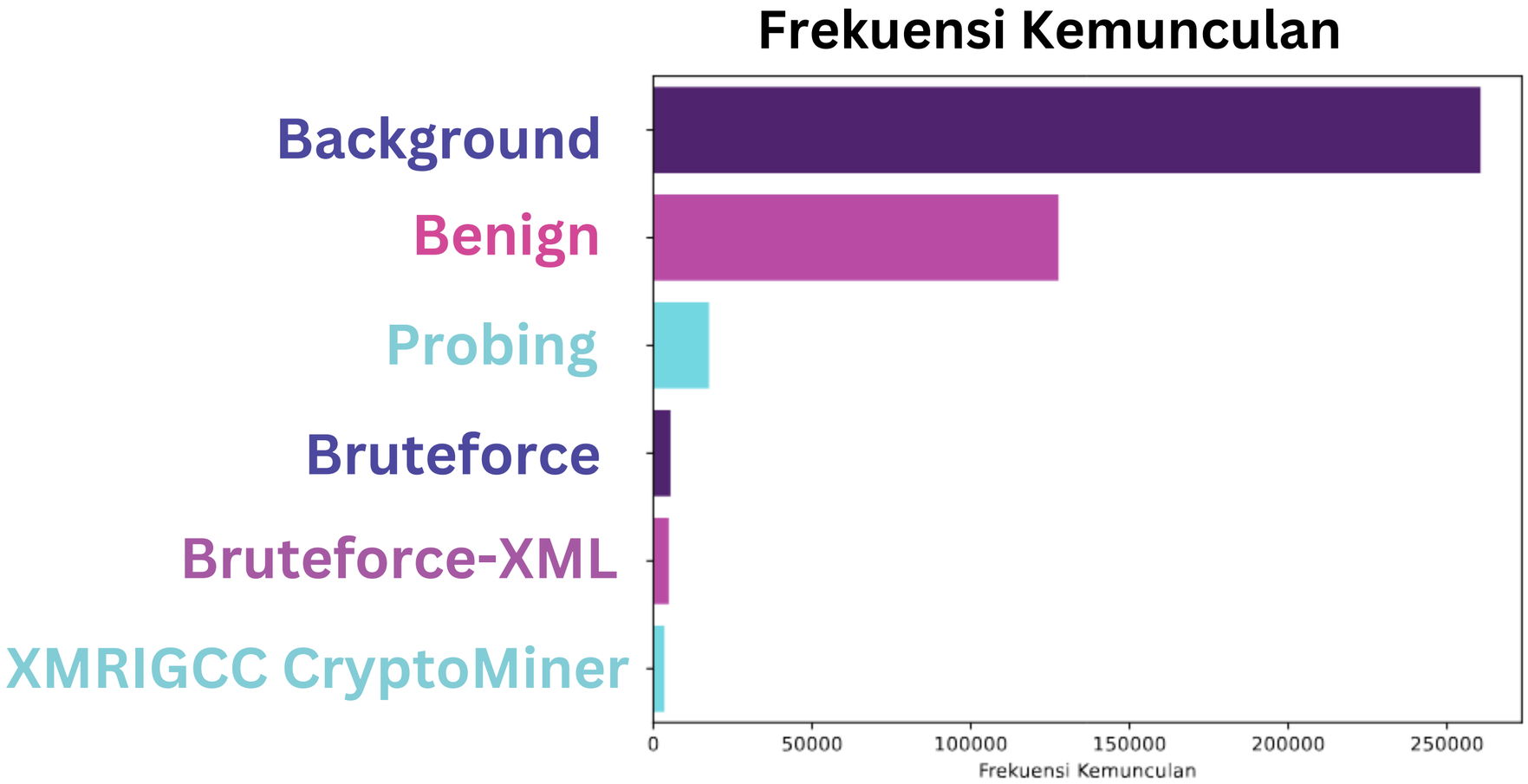
Menciptakan **model prediktif** yang dapat menganalisis volume data besar secara **real-time**, dengan kemampuan untuk mendeteksi anomali dan mengklasifikasikan berbagai jenis lalu lintas dengan **akurasi tinggi**

Analisis dan Pembahasan

Deskripsi Dataset



Deskripsi Label Jenis Aliran Jaringan



DATA QUEST
PART OF AIRNOLOGY

Exploratory Data Analysis

Tipe Data

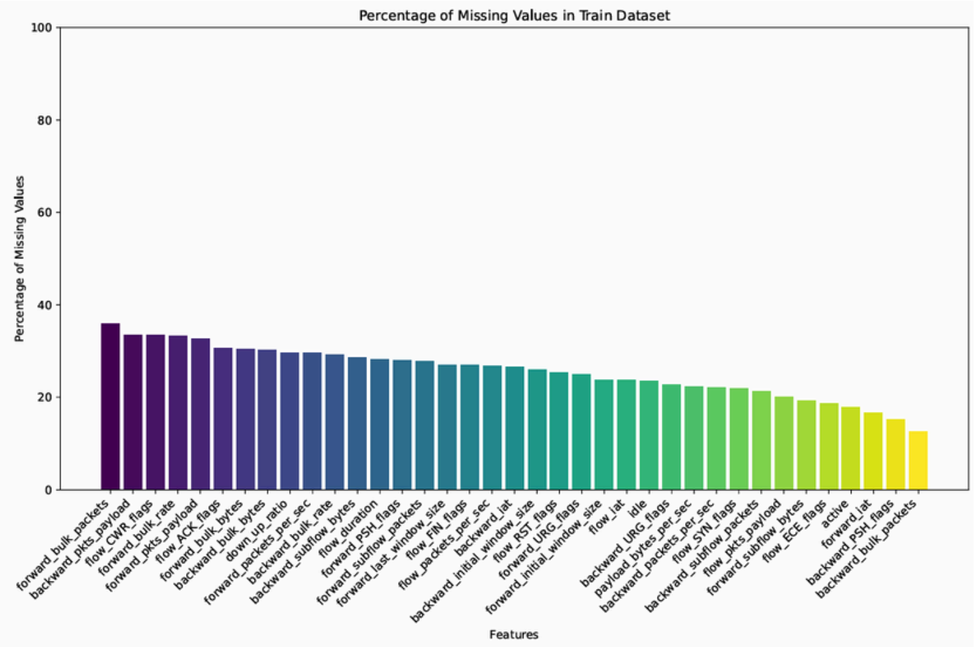
id	object
origin_host	object
origin_port	int64
response_host	object
response_port	int64
flow_duration	float64
forward_packets_per_sec	float64
backward_packets_per_sec	float64
flow_packets_per_sec	float64
down_up_ratio	float64
flow_FIN_flags	float64
flow_SYN_flags	float64
flow_RST_flags	float64
forward_PSH_flags	float64
backward_PSH_flags	float64
flow_ACK_flags	float64
forward_URG_flags	float64
backward_URG_flags	float64
flow_CWR_flags	float64
flow_ECE_flags	float64
forward_pkts_payload	float64
backward_pkts_payload	float64
flow_pkts_payload	float64
forward_iat	float64
backward_iat	float64
flow_iat	float64
payload_bytes_per_sec	float64
forward_subflow_packets	float64
backward_subflow_packets	float64
forward_subflow_bytes	float64
backward_subflow_bytes	float64
forward_bulk_bytes	float64
backward_bulk_bytes	float64
forward_bulk_packets	float64
backward_bulk_packets	float64
forward_bulk_rate	float64
backward_bulk_rate	float64
active	float64
idle	float64
forward_initial_window_size	float64
backward_initial_window_size	float64
forward_last_window_size	float64
traffic	object

Terdapat 4 Kolom yang bertipe data String sehingga harus dilakukan pre-processing lebih lanjut

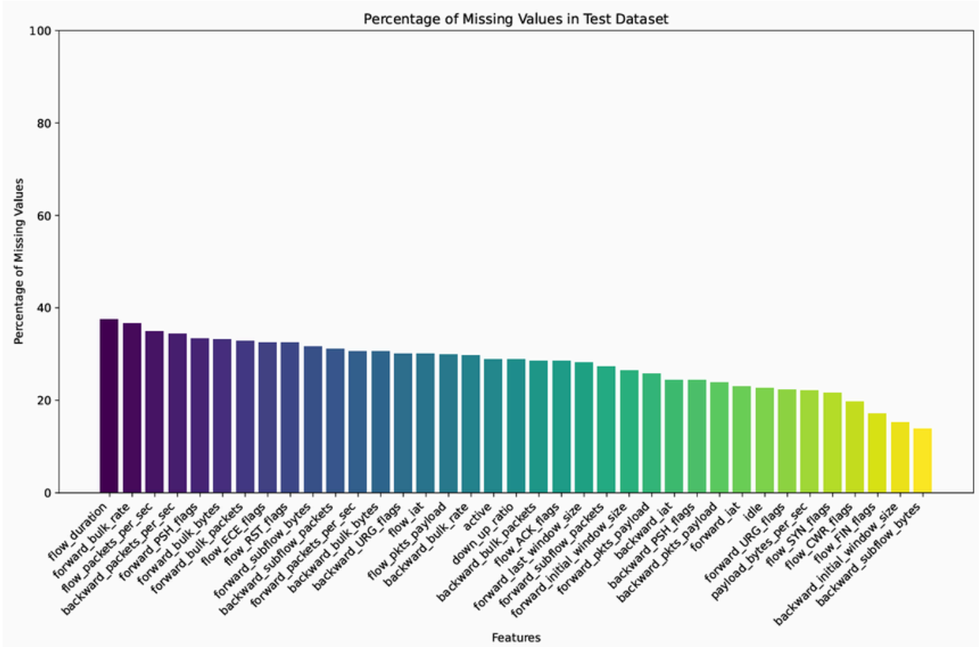
Missing Values

DATA QUEST
PART OF AIRNOLOGY

Train data

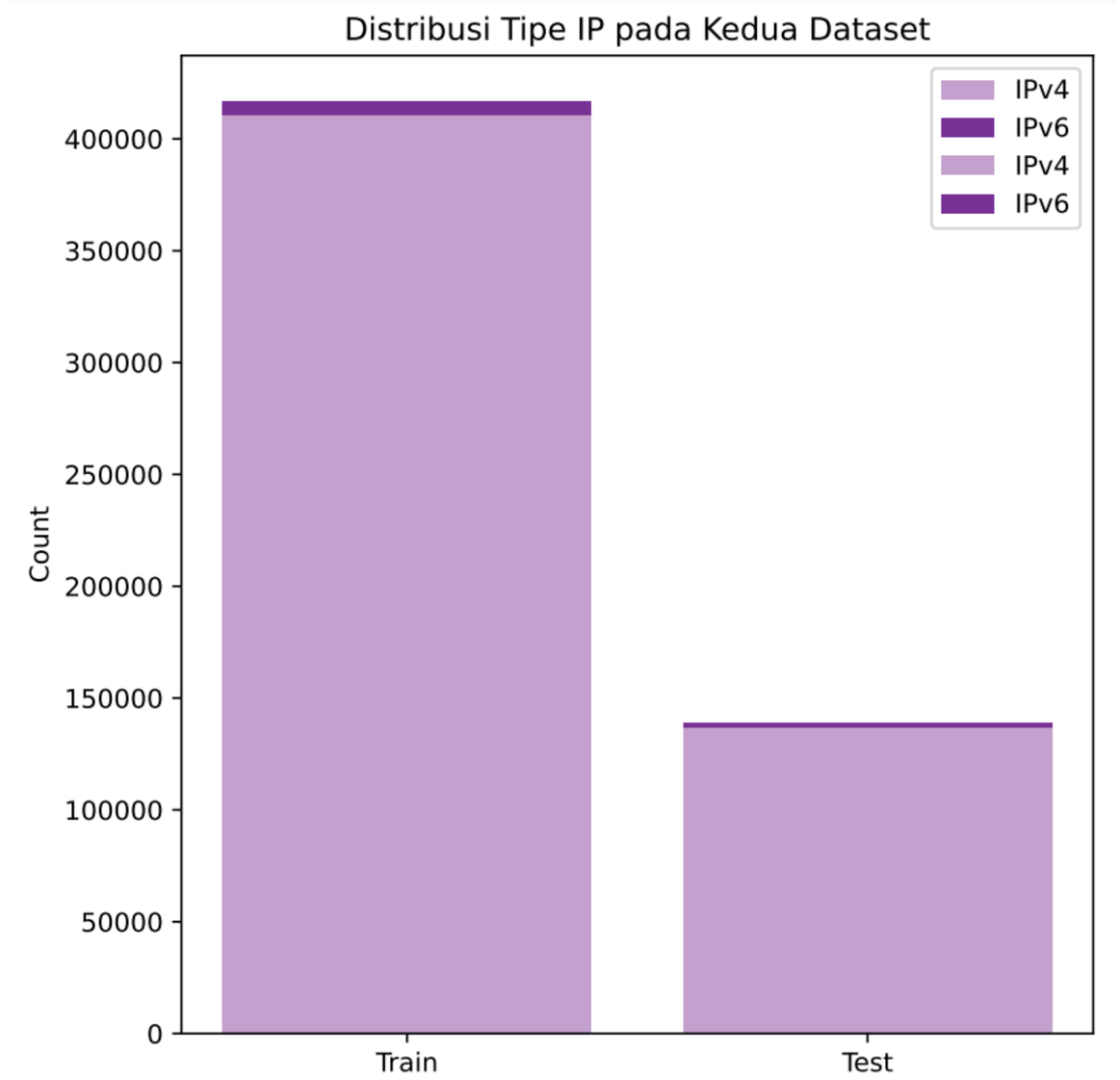


Test data



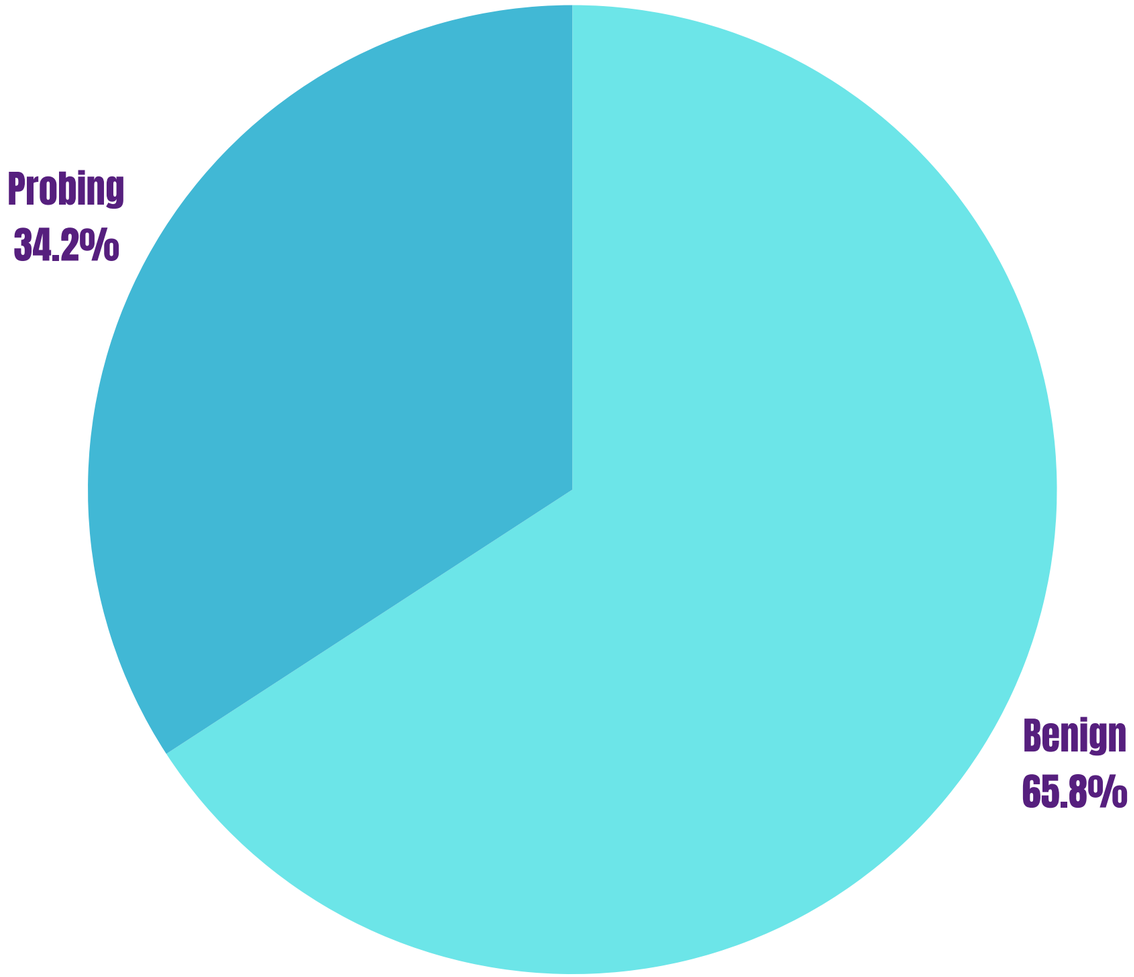
Persentase Missing Values pada dataset mencapai 35%

Alamat IP



Distribusi Tipe IP yang terdiri dari IPV6 dan dimayoritasi IPV4

Keunikan Dataset - ‘Sang Pelaku’



Ada anomali pada origin host ’103.255.15.150’ dengan response host ’128.199.242.104’ pada response port 443

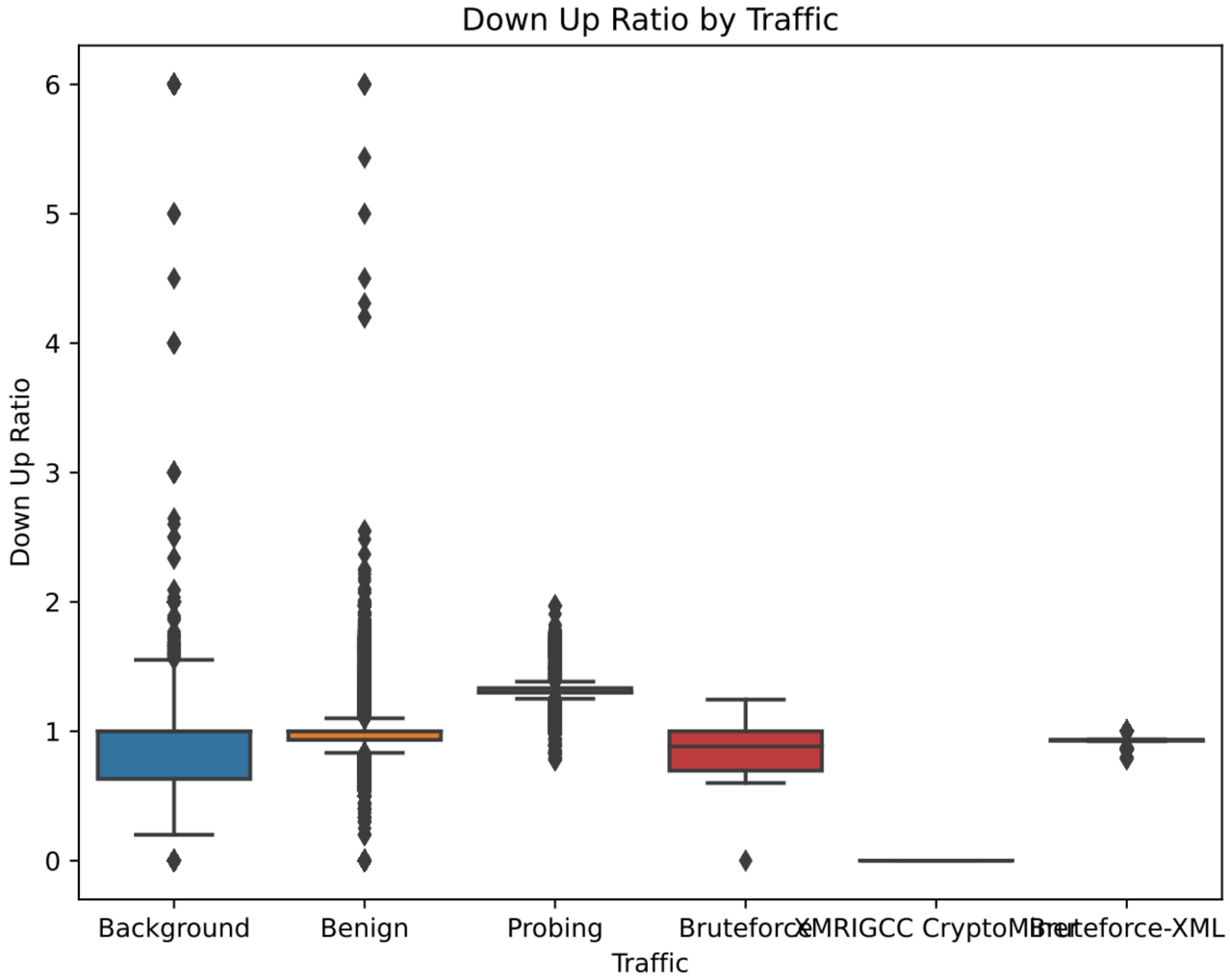
Keunikan Dataset - Ciri Khas Fitur pada Setiap Kelas

DATA QUEST
PART OF AIRNOLOGY

Terdapat beberapa fitur pada data yang memiliki persebaran value yang berbeda sehingga dapat dianalisis untuk mencari karakteristik berbeda antar suatu label

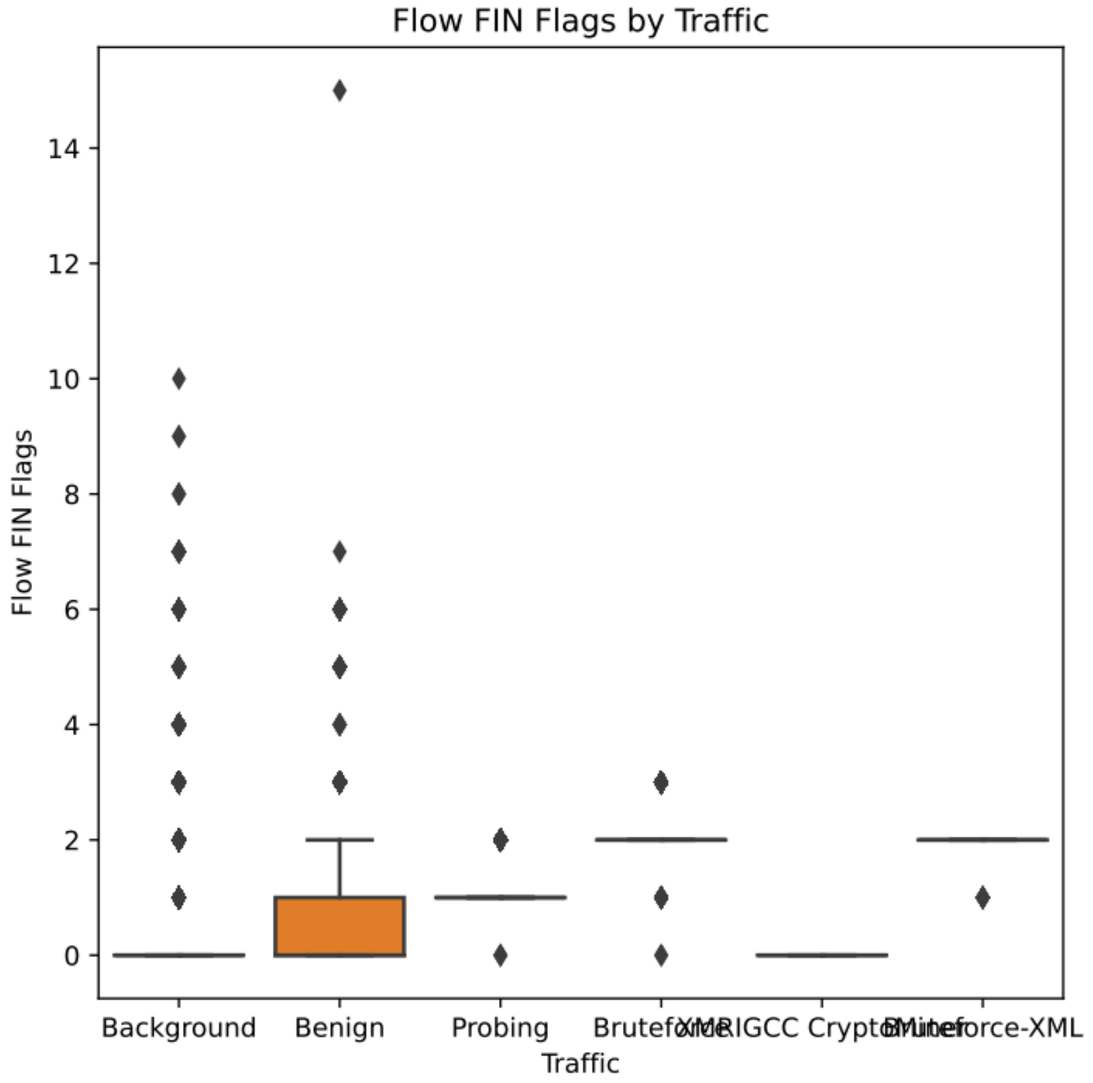
Keunikan Dataset - Ciri Khas Down Up Ratio pada Setiap Kelas

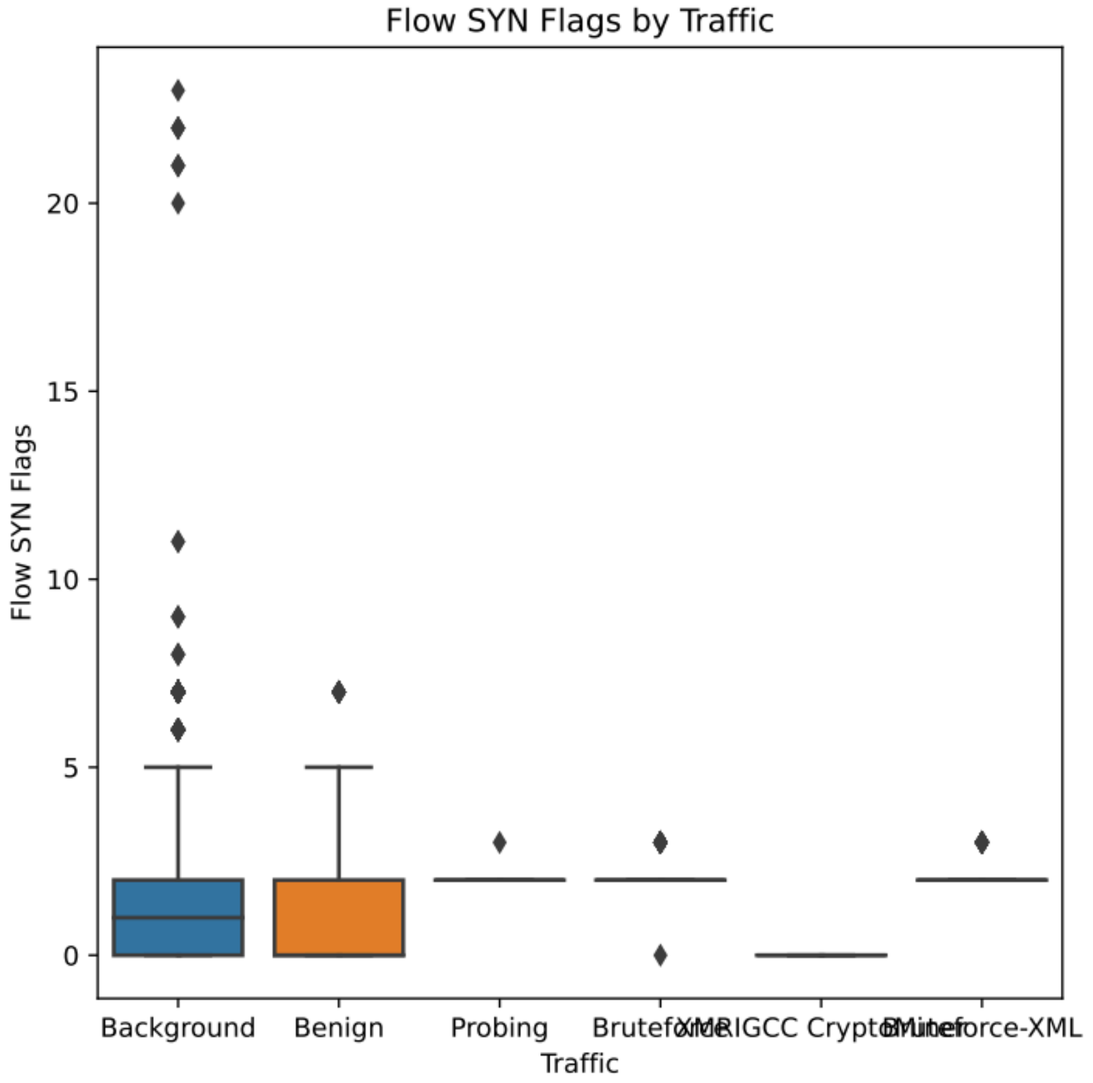
DATA QUEST
PART OF AIRNOLOGY



Keunikan Dataset - Ciri Khas Flow FIN dan SYN Flags Feature pada Setiap Kelas

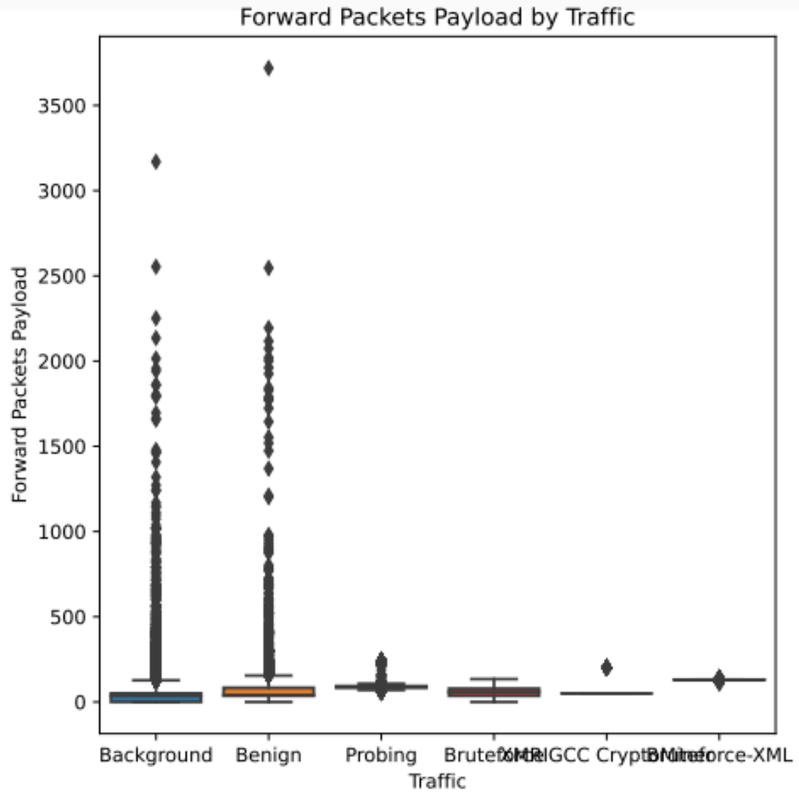
DATA QUEST
PART OF AIRNOLOGY

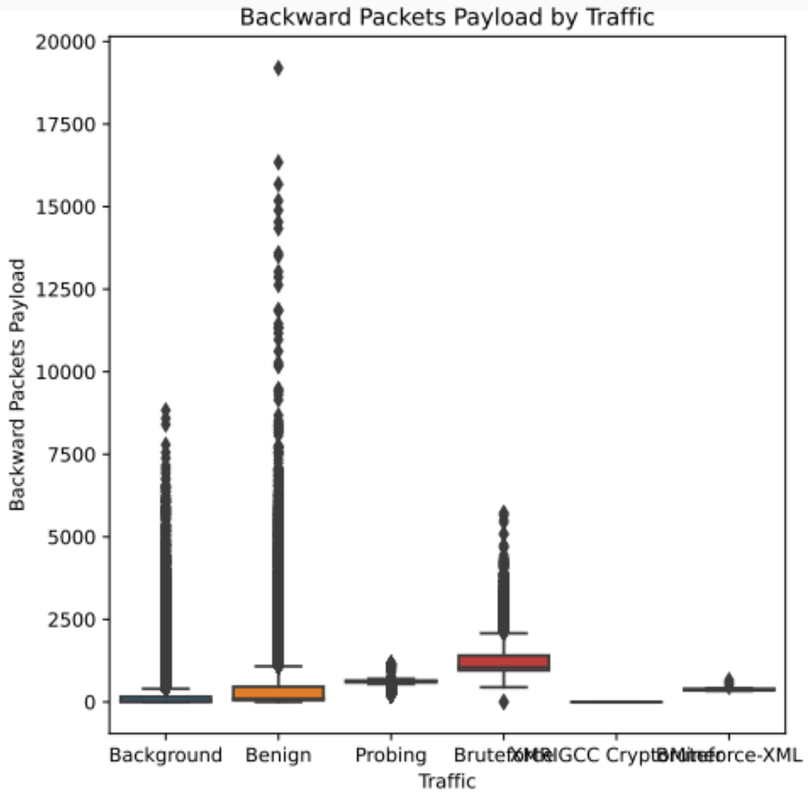


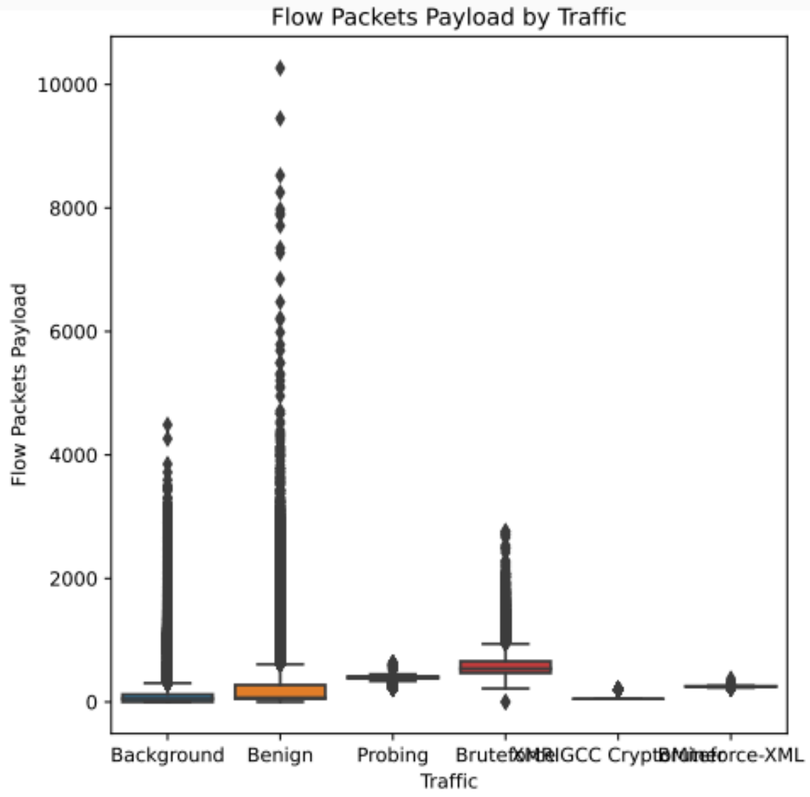


Keunikan Dataset - Ciri Khas Forward Packets Payload, Backward Packets Payload, Flow Packets Payload Feature pada Setiap Kelas

DATA QUEST
PART OF AIRNOLOGY







Keunikan Dataset - Ciri Khas Fitur pada Setiap Kelas

DATA QUEST
PART OF AIRNOLOGY

1. **Outlier yang sangat mendominasi box plot pada Benign dan Background**
2. **XMRIGCC Cryptominer yang selalu memiliki value 0 di keenam fitur tersebut**
3. **Perbedaan distribusi data untuk fitur backward packets payload dan flow packets payload**

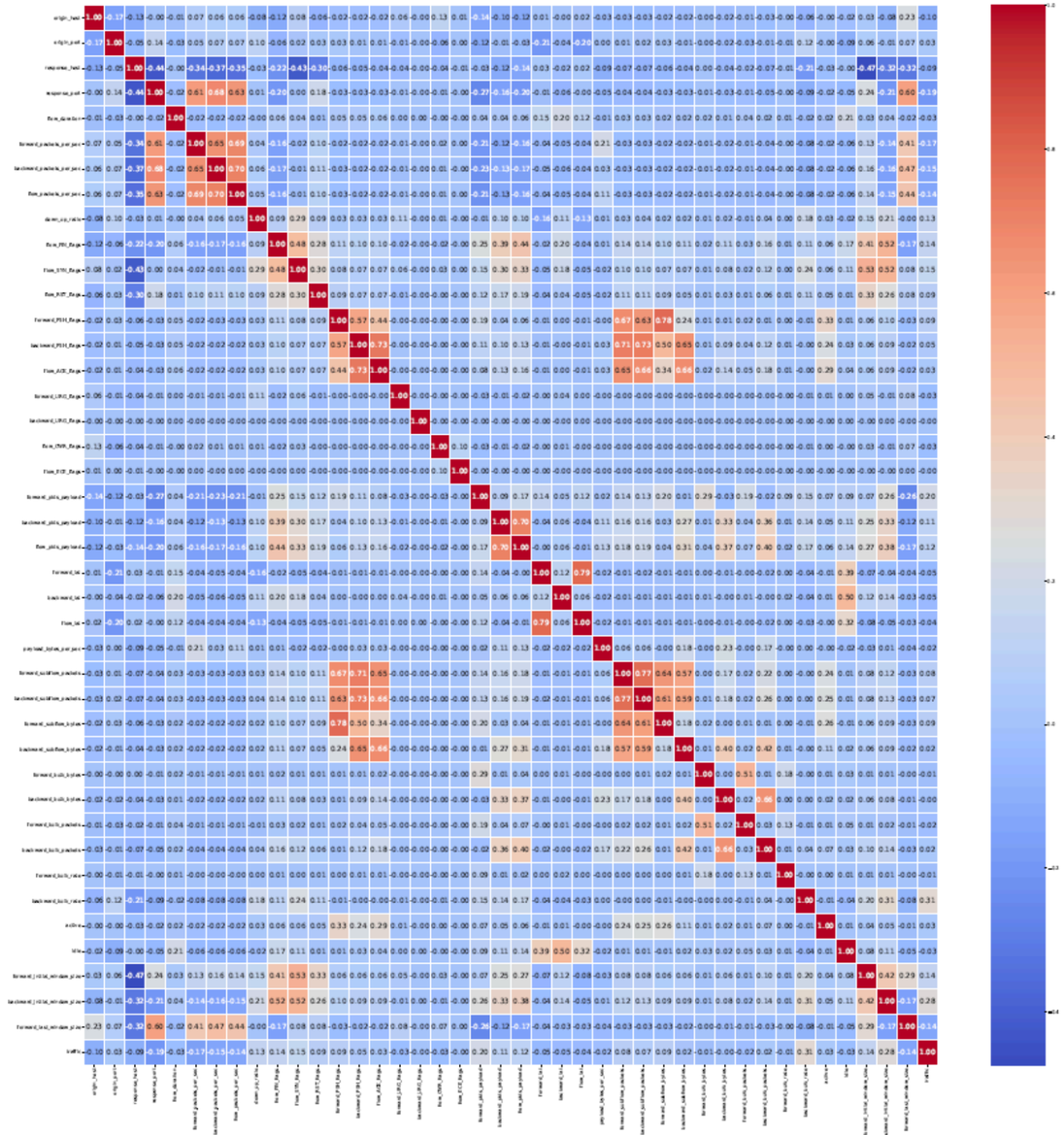
Correlation Analysis

DATA QUEST
PART OF AIRNOLOGY

Size of Correlation	Interpretation
0.90 to 1.00 / -0.90 to -1.00	Very high positive / negative correlation
0.70 to 0.90 / -0.70 to -0.90	High positive / negative correlation
0.50 to 0.70 / -0.50 to -0.70	Moderate positive / negative correlation
0.30 to 0.50 / -0.30 to -0.50	Low positive / negative correlation
0.00 to 0.30 / -0.00 to -0.30	Very low positive / negative correlation

Tabel 3: Interpretation of Correlation Coefficient

Correlation Analysis



1. Tidak terdapat very high correlation
2. Terdapat beberapa korelasi fitur yang bersifat high correlation, seperti :
   - flow_packets_per_sec dan backward_packets_per_sec,
   - forward_subflow_bytes dan forward_PSH_flags,
   - flow_ACK_flags dan backward_PSH_flags,
   - forward_subflow_packets dan backward_PSH_flags
   - backward_subflow_bytes dan backward_PSH_flags
3. Mayoritas korelasi antar fitur bersifat very low correlation

DATA QUEST
PART OF AIRNOLOGY

Preprocessing

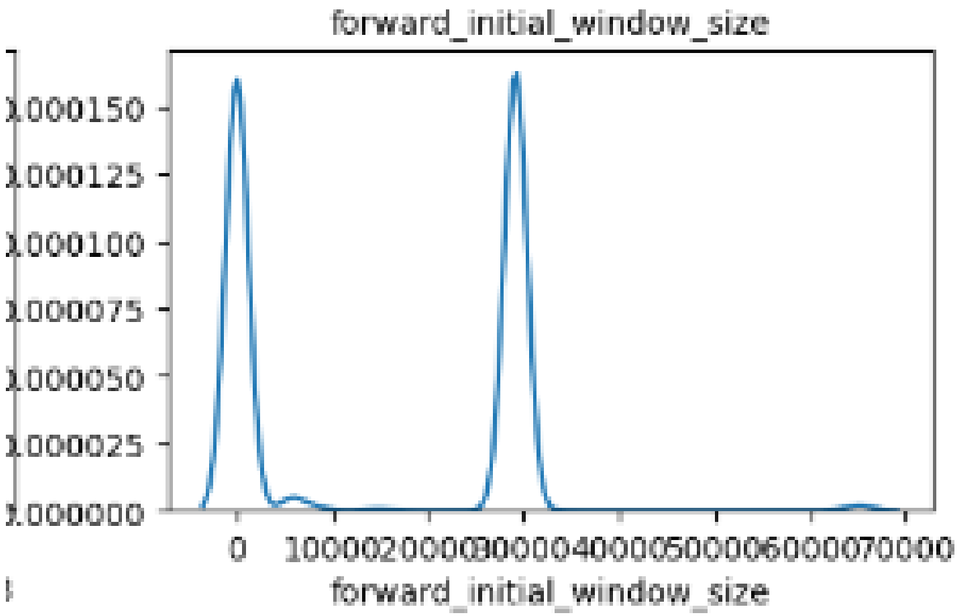
Preprocessing - Special Imputation

DATA QUEST
PART OF AIRNOLOGY

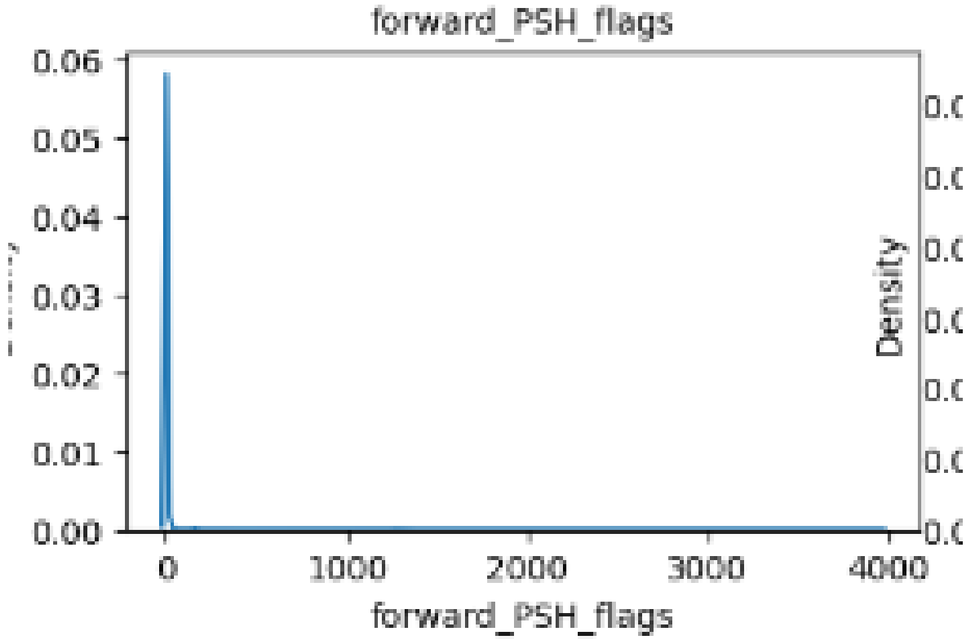
id	forward_packets_per_sec	backward_packets_per_sec	flow_packets_per_sec
CkwI1TlUCRApPfcJl	11125.474801	NaN	22250.949602
CBlrcc3dvtaHzyV4zj	30174.848921	30174.848921	60349.697842
CdpSX33u29yjDvnVzi	0.322699	0.242025	0.564724
CT23VJ1KsoKeCdWpx2	NaN	82.478178	164.956355
C6OJU51P50bwNKvnY6	NaN	NaN	72.5162256
CWhujWex7PeGY4Q78	36.078793	46.902431	NaN
CA4mfd5WpmkjY07Qk	76.190116	76.190116	NaN
CE0eax4wzKi6EeAWc5	34.350759	38.167510	72.518269
C4EbK82xAdSARKweec	NaN	56.906226	NaN
CJk4Yq3AHNRTsiz1rf	NaN	39.660199	79.320398

$$\texttt{flow\_packets\_per\_sec} = \texttt{forward\_packets\_per\_sec} + \texttt{backward\_packets\_per\_sec}$$

General Feature Imputation
Modus dan Median sebagai Metode Imputasi Terbaik



Median



Mode

General Feature Imputation
Ketidakefektifan Iterative dan KNN Imputer

DATA QUEST
PART OF AIRNOLOGY

Iterative imputer tidak efektif dalam melakukan imputasi terhadap data yang tidak terdistribusi secara normal

KNN Imputer tidak efektif karena waktu komputasi yang lama dalam menghadapi ratusan ribu NaN values

DATA QUEST
PART OF AIRNOLOGY

Feature Engineering

Pemisahan Segmen Alamat IP

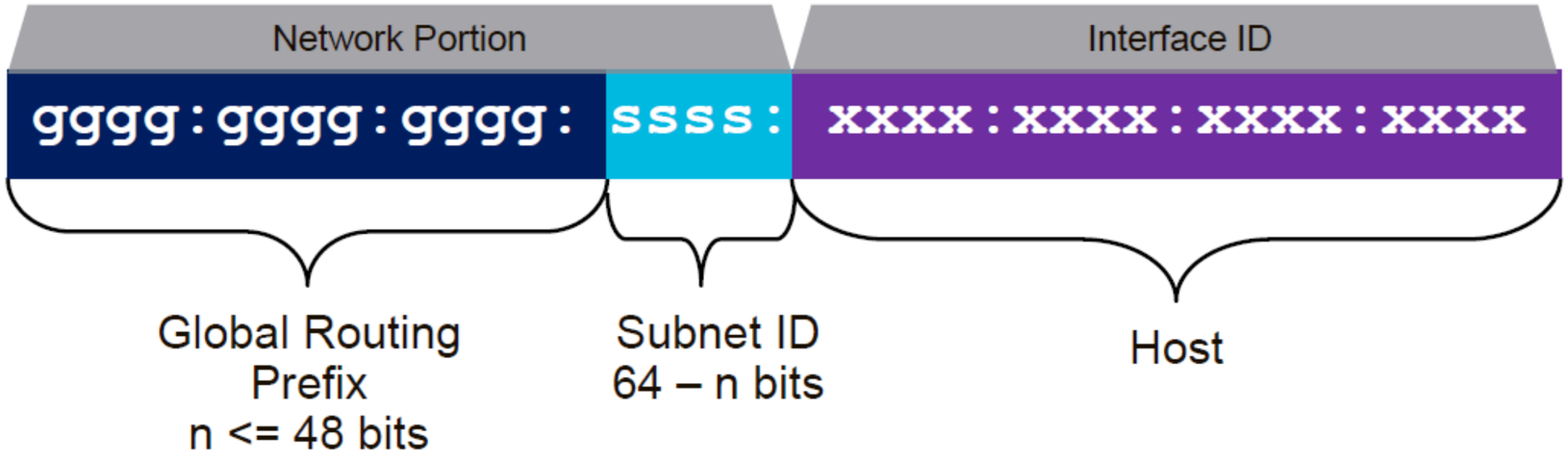
DATA QUEST
PART OF AIRNOLOGY

IPv4

17.172.224.47

8 bits (1 byte)	8 bits (1 byte)	8 bits (1 byte)	8 bits (1 byte)

IPv6



Kesamaan Subnet pada Alamat IP Server dan Klien

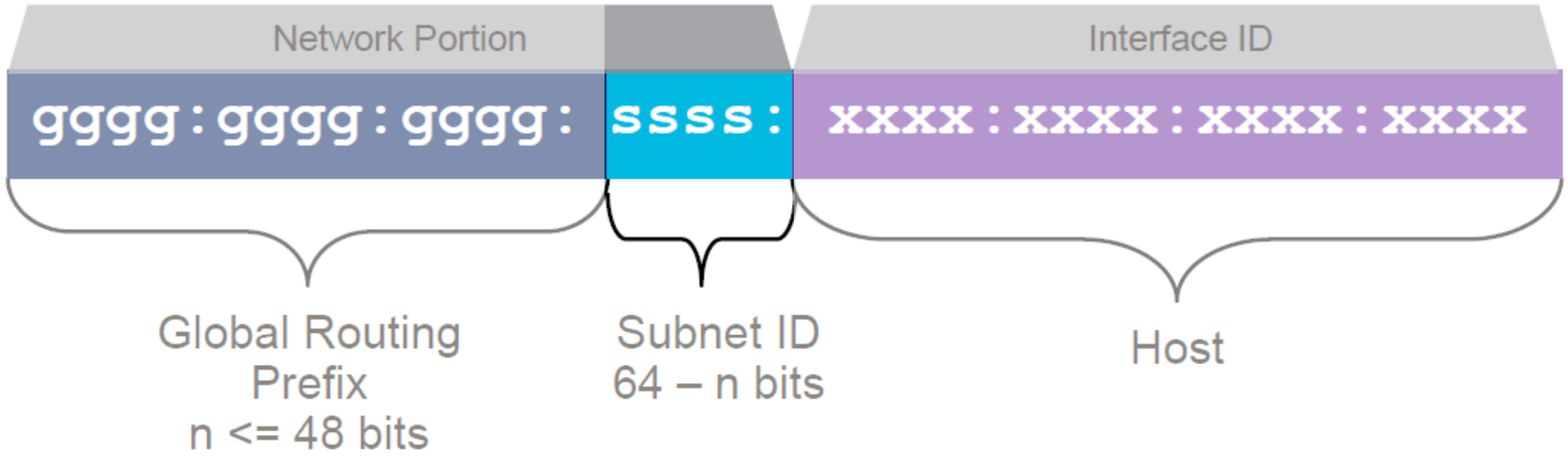
DATA QUEST
PART OF AIRNOLOGY

IPv4

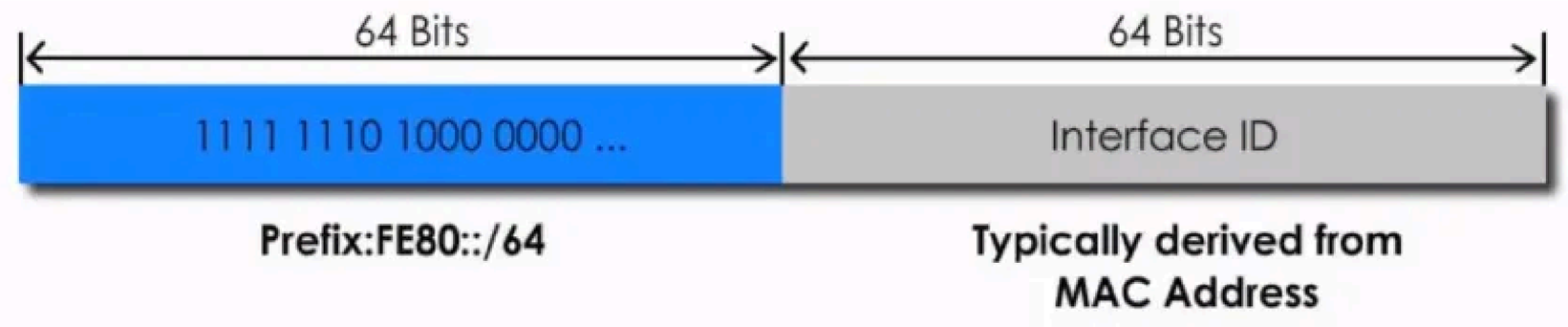
17.172.224.47

8 bits (1 byte)	8 bits (1 byte)	8 bits (1 byte)	8 bits (1 byte)

IPv6



Klasifikasi Alamat IP



0xFE80

0xFEC0

0xFF00

blocks[0] == 10 or

(blocks[0] == 192 and blocks[1] == 168) or (blocks[0] == 172 and 16 <= blocks[1] <= 31)

IPv4

Identifikasi Origin dan Response Port Umum

DATA QUEST
PART OF AIRNOLOGY

All traffic data with:

origin_host **'103.255.15.150'**
response_host **'128.199.242.104'**
response_port **443**

→ **has similar traffic?**

→ **if yes, how many?**

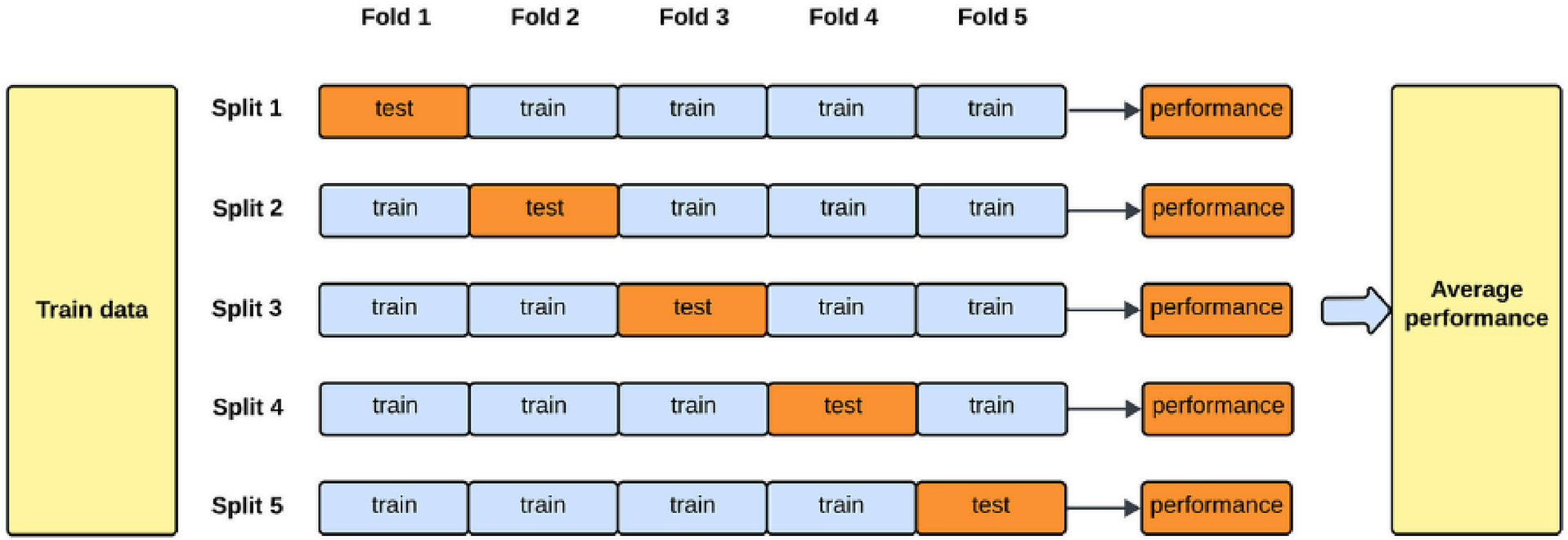
Similar traffic is considered based on the similarity of:

'**flow_duration**', '**forward_packets_per_sec**', '**active**'
'**backward_packets_per_sec**', '**flow_packets_per_sec**'

DATA QUEST
PART OF AIRNOLOGY

Modelling

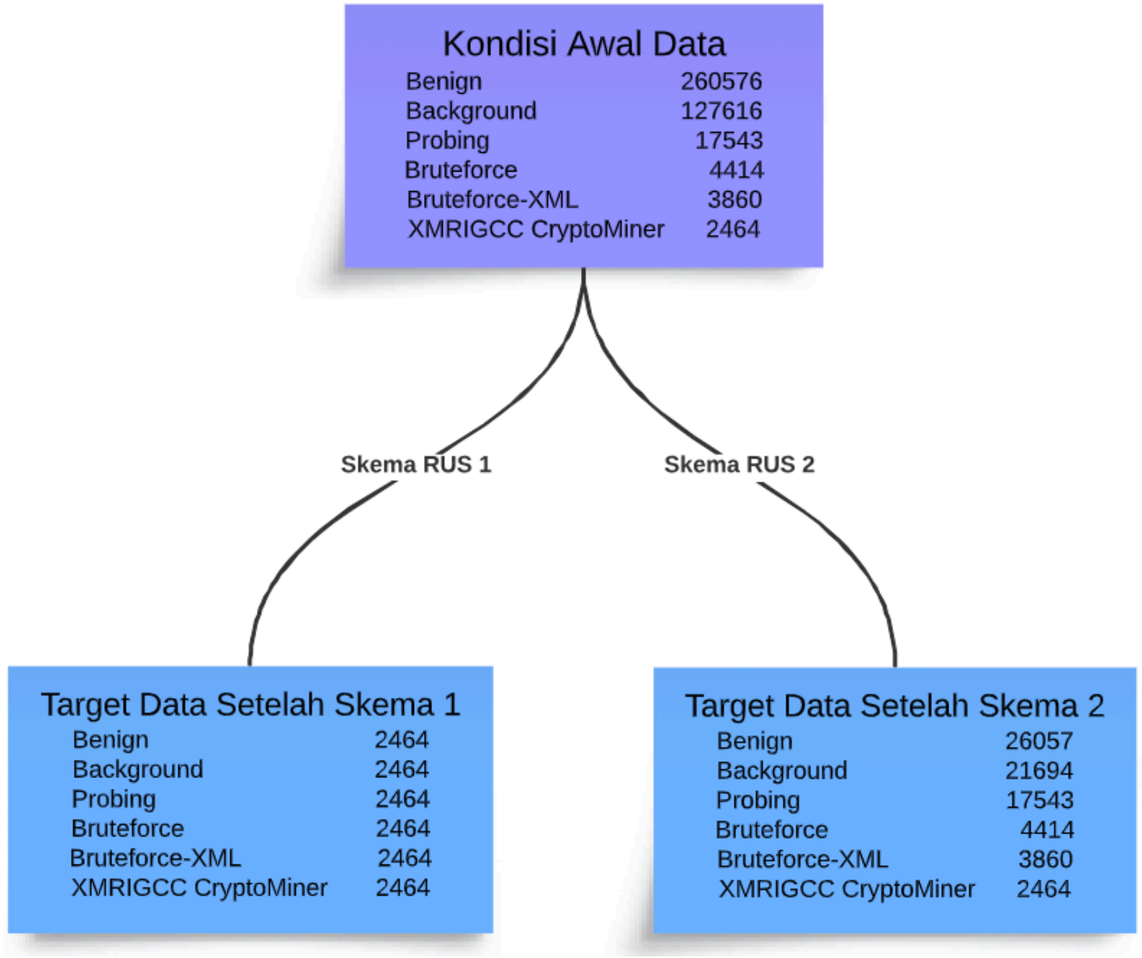
Stratified K-Fold sebagai Validator



Mengatasi Ketidakseimbangan Kelas pada Dataset

DATA QUEST
PART OF AIRNOLOGY

2 Skema RUS



Class Weighting

$$CW_k = \frac{max_{c=1}^{K}(\sum_{t_i=c} w_i)}{\sum_{t_i=k} w_i}$$

```
n_samples / (n_classes * np.bincount(y))
```

Training time CPU vs GPU

DATA QUEST
PART OF AIRNOLOGY

Setelah enable GPU pada CatBoost model, proses validasi stratified 5-fold dilaksanakan hanya selama 30 menit. Hal ini meningkatkan waktu pelatihan, dimana stratified 5-fold untuk model CatBoost CPU mencapai 5 jam!

Hasil Analisis

Hasil Analisis

DATA QUEST
PART OF AIRNOLOGY

Skema Eksperimen	*Private Leaderboard*	Public Leaderboard	Stratified K-Fold
LGBM + RUS Skema 1	0, 80839	0, 80769	0, 8206
LGBM + RUS Skema 2	0, 79874	0, 79851	0, 8198
LGBM Auto Balance Weighting	0, 82060	0, 82240	0, 82170
CatBoost Auto Balance Weighting dengan CPU	0, 87270	0, 87257	0, 87201
CatBoost Auto Balance Weighting dengan GPU	0, 84221	0, 84337	0, 84566
CatBoost + RUS Skema 2 dengan CPU	0, 85140	0, 85504	0, 85387
CatBoost + RUS Skema 2 dengan GPU	0, 82230	0, 82224	0.81240

Model terbaik: CatBoost Auto Balance Weighting CPU

Mengapa model sama dengan CPU dan GPU memiliki hasil berbeda?

Hasil Analisis: CPU vs GPU

DATA QUEST
PART OF AIRNOLOGY

Perbedaan Default Value

The default value depends on the processing unit type and other parameters:

- CPU: 254
- GPU in PairLogitPairwise and YetiRankPairwise modes: 32
- GPU in all other modes: 128

Pembatasan Range Value

Note

The maximum depth of the trees is limited to 8 for pairwise modes (YetiRank, PairLogitPairwise and QueryCrossEntropy) when the training is performed on GPU.

sumber: catboost documentation

Hasil Analisis: RUS vs Weighting

DATA QUEST
PART OF AIRNOLOGY

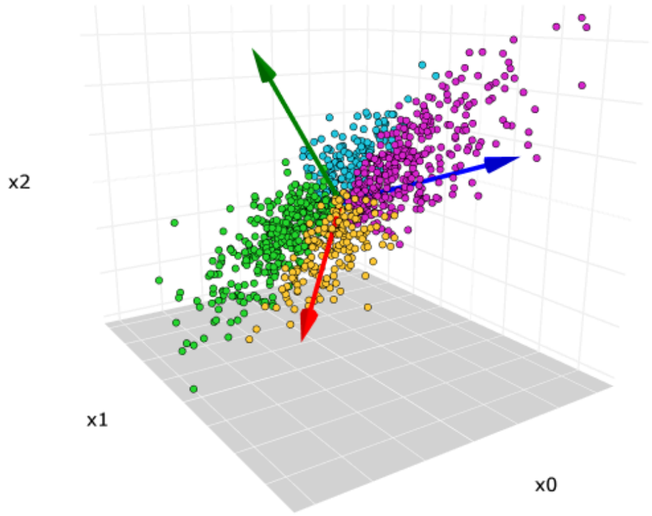
Skema Eksperimen	*Private Leaderboard*	Public Leaderboard	Stratified K-Fold
LGBM + RUS Skema 1	0, 80839	0, 80769	0, 8206
LGBM + RUS Skema 2	0, 79874	0, 79851	0, 8198
LGBM Auto Balance Weighting	0, 82060	0, 82240	0, 82170
CatBoost Auto Balance Weighting dengan CPU	0, 87270	0, 87257	0, 87201
CatBoost Auto Balance Weighting dengan GPU	0, 84221	0, 84337	0, 84566
CatBoost + RUS Skema 2 dengan CPU	0, 85140	0, 85504	0, 85387
CatBoost + RUS Skema 2 dengan GPU	0, 82230	0, 82224	0.81240

Keunggulan weighting terhadap RUS dapat terlihat pada performa metriks. Hal ini disebabkan hilangnya informasi data mayoritas ketika dilakukan RUS

Hasil Analisis: Mengatasi waktu latih yang lama

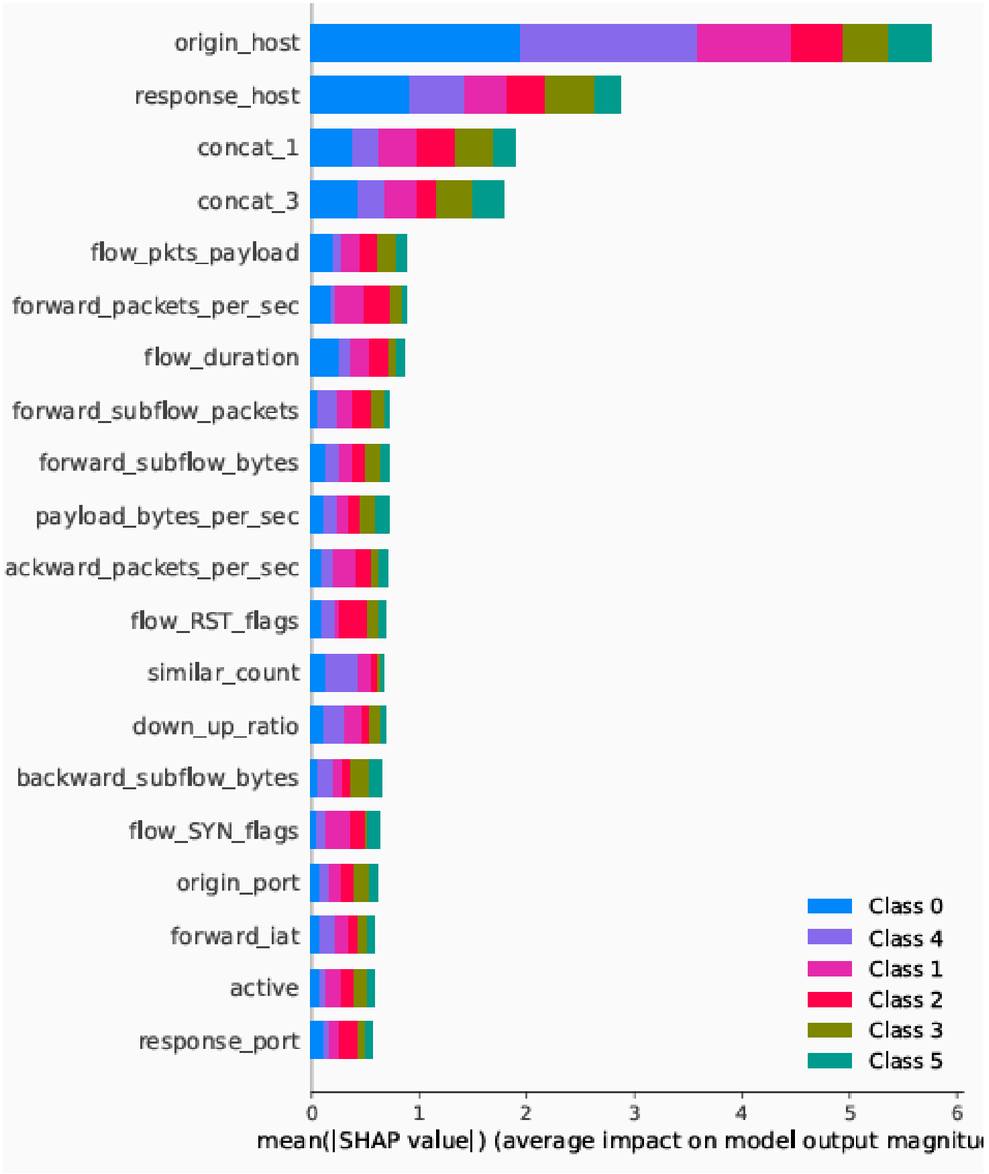
DATA QUEST
PART OF AIRNOLOGY

Model berbasis CPU memberikan hasil yang paling baik pada performa metriks. Sayangnya, proses pelatihannya lama.



PCA Analysis

Hasil Analisis: SHAP



Kesimpulan

Hasil eksperimen menunjukkan bahwa metode Auto Balance Weighting dengan algoritma CatBoost memberikan performa terbaik, terutama saat pengujian dengan CPU, dan efektif mengatasi ketidakseimbangan data antar kelas dalam deteksi intrusi jaringan.

DATA QUEST
PART OF AIRNOLOGY

Saran

Saran

1. Pengembangan Model Berbasis GPU
2. Pengujian di Lingkungan Jaringan Nyata

Thank You!