STEP 1: Starting data acquisition from Kaggle...

Start downloading dataset...

Download successfully.

Path to dataset: /Users/tranvankhoi/.cache/kagglehub/datasets/chethuhn/network-intrusion-dataset/versions/1

🚀 Starting DDoS Detection System Pipeline...

================================================================================

List of csv files: ['Thursday-WorkingHours-Afternoon-Infilteration.pcap\_ISCX.csv', 'Monday-WorkingHours.pcap\_ISCX.csv', 'Friday-WorkingHours-Morning.pcap\_ISCX.csv', 'Friday-WorkingHours-Afternoon-PortScan.pcap\_ISCX.csv', 'Friday-WorkingHours-Afternoon-DDos.pcap\_ISCX.csv', 'Tuesday-WorkingHours.pcap\_ISCX.csv', 'Wednesday-workingHours.pcap\_ISCX.csv', 'Thursday-WorkingHours-Morning-WebAttacks.pcap\_ISCX.csv']

Start reading file: Thursday-WorkingHours-Afternoon-Infilteration.pcap\_ISCX.csv

Shape: (288602, 79)

Columns: 79

Start reading file: Monday-WorkingHours.pcap\_ISCX.csv

Shape: (529918, 79)

Columns: 79

Start reading file: Friday-WorkingHours-Morning.pcap\_ISCX.csv

Shape: (191033, 79)

Columns: 79

Start reading file: Friday-WorkingHours-Afternoon-PortScan.pcap\_ISCX.csv

Shape: (286467, 79)

Columns: 79

Start reading file: Friday-WorkingHours-Afternoon-DDos.pcap\_ISCX.csv

Shape: (225745, 79)

Columns: 79

Start reading file: Tuesday-WorkingHours.pcap\_ISCX.csv

Shape: (445909, 79)

Columns: 79

Start reading file: Wednesday-workingHours.pcap\_ISCX.csv

Shape: (692703, 79)

Columns: 79

Start reading file: Thursday-WorkingHours-Morning-WebAttacks.pcap\_ISCX.csv

Shape: (170366, 79)

Columns: 79

Shape: (2830743, 79)

Columns: Index([' Destination Port', ' Flow Duration', ' Total Fwd Packets',

' Total Backward Packets', 'Total Length of Fwd Packets',

' Total Length of Bwd Packets', ' Fwd Packet Length Max',

' Fwd Packet Length Min', ' Fwd Packet Length Mean',

' Fwd Packet Length Std', 'Bwd Packet Length Max',

' Bwd Packet Length Min', ' Bwd Packet Length Mean',

' Bwd Packet Length Std', 'Flow Bytes/s', ' Flow Packets/s',

' Flow IAT Mean', ' Flow IAT Std', ' Flow IAT Max', ' Flow IAT Min',

'Fwd IAT Total', ' Fwd IAT Mean', ' Fwd IAT Std', ' Fwd IAT Max',

' Fwd IAT Min', 'Bwd IAT Total', ' Bwd IAT Mean', ' Bwd IAT Std',

' Bwd IAT Max', ' Bwd IAT Min', 'Fwd PSH Flags', ' Bwd PSH Flags',

' Fwd URG Flags', ' Bwd URG Flags', ' Fwd Header Length',

' Bwd Header Length', 'Fwd Packets/s', ' Bwd Packets/s',

' Min Packet Length', ' Max Packet Length', ' Packet Length Mean',

' Packet Length Std', ' Packet Length Variance', 'FIN Flag Count',

' SYN Flag Count', ' RST Flag Count', ' PSH Flag Count',

' ACK Flag Count', ' URG Flag Count', ' CWE Flag Count',

' ECE Flag Count', ' Down/Up Ratio', ' Average Packet Size',

' Avg Fwd Segment Size', ' Avg Bwd Segment Size',

' Fwd Header Length.1', 'Fwd Avg Bytes/Bulk', ' Fwd Avg Packets/Bulk',' Fwd Avg Bulk Rate', ' Bwd Avg Bytes/Bulk', ' Bwd Avg Packets/Bulk',

'Bwd Avg Bulk Rate', 'Subflow Fwd Packets', ' Subflow Fwd Bytes',

' Subflow Bwd Packets', ' Subflow Bwd Bytes', 'Init\_Win\_bytes\_forward',

' Init\_Win\_bytes\_backward', ' act\_data\_pkt\_fwd',

' min\_seg\_size\_forward', 'Active Mean', ' Active Std', ' Active Max',

' Active Min', 'Idle Mean', ' Idle Std', ' Idle Max', ' Idle Min',

' Label'],

dtype='object')

✅ STEP 1 COMPLETED: Data acquisition successful

================================================================================

🔄 STEP 5: Starting comprehensive data preprocessing...

🔍 Starting outlier detection...

📊 Destination Port: Capped 627245 outliers (22.16%)

📊 Flow Duration: Capped 532586 outliers (18.81%)

📊 Total Fwd Packets: Capped 332693 outliers (11.75%)

📊 Total Backward Packets: Capped 308291 outliers (10.89%)

📊 Total Length of Fwd Packets: Capped 446972 outliers (15.79%)

📊 Total Length of Bwd Packets: Capped 619046 outliers (21.87%)

📊 Fwd Packet Length Max: Capped 664214 outliers (23.46%)

📊 Fwd Packet Length Min: Capped 14958 outliers (0.53%)

📊 Fwd Packet Length Mean: Capped 188020 outliers (6.64%)

📊 Fwd Packet Length Std: Capped 663959 outliers (23.46%)

📊 Bwd Packet Length Max: Capped 637112 outliers (22.51%)

📊 Bwd Packet Length Min: Capped 85557 outliers (3.02%)

📊 Bwd Packet Length Mean: Capped 472283 outliers (16.68%)

📊 Bwd Packet Length Std: Capped 654269 outliers (23.11%)

📊 Flow Bytes/s: Capped 532458 outliers (18.81%)

📊 Flow Packets/s: Capped 306567 outliers (10.83%)

📊 Flow IAT Mean: Capped 573536 outliers (20.26%)

📊 Flow IAT Std: Capped 599510 outliers (21.18%)

📊 Flow IAT Max: Capped 531354 outliers (18.77%)

📊 Flow IAT Min: Capped 512300 outliers (18.10%)

📊 Fwd IAT Total: Capped 665766 outliers (23.52%)

📊 Fwd IAT Mean: Capped 671833 outliers (23.73%)

📊 Fwd IAT Std: Capped 660519 outliers (23.33%)

📊 Fwd IAT Max: Capped 666292 outliers (23.54%)

📊 Fwd IAT Min: Capped 535064 outliers (18.90%)

📊 Bwd IAT Total: Capped 565783 outliers (19.99%)

📊 Bwd IAT Mean: Capped 563909 outliers (19.92%)

📊 Bwd IAT Std: Capped 599454 outliers (21.18%)

📊 Bwd IAT Max: Capped 568604 outliers (20.09%)

📊 Bwd IAT Min: Capped 284946 outliers (10.07%)

📊 Fwd PSH Flags: Capped 131478 outliers (4.64%)

📊 Fwd URG Flags: Capped 315 outliers (0.01%)

📊 Fwd Header Length: Capped 383532 outliers (13.55%)

📊 Bwd Header Length: Capped 353252 outliers (12.48%)

📊 Fwd Packets/s: Capped 337419 outliers (11.92%)

📊 Bwd Packets/s: Capped 254052 outliers (8.97%)

📊 Min Packet Length: Capped 9994 outliers (0.35%)

📊 Max Packet Length: Capped 623901 outliers (22.04%)

📊 Packet Length Mean: Capped 462792 outliers (16.35%)

📊 Packet Length Std: Capped 559721 outliers (19.77%)

📊 Packet Length Variance: Capped 654542 outliers (23.12%)

📊 FIN Flag Count: Capped 100151 outliers (3.54%)

📊 SYN Flag Count: Capped 131478 outliers (4.64%)

📊 RST Flag Count: Capped 686 outliers (0.02%)

📊 URG Flag Count: Capped 268420 outliers (9.48%)

📊 CWE Flag Count: Capped 315 outliers (0.01%)

📊 ECE Flag Count: Capped 689 outliers (0.02%)

📊 Down/Up Ratio: Capped 27294 outliers (0.96%)

📊 Average Packet Size: Capped 409714 outliers (14.47%)

📊 Avg Fwd Segment Size: Capped 188020 outliers (6.64%)

📊 Avg Bwd Segment Size: Capped 472283 outliers (16.68%)

📊 Fwd Header Length.1: Capped 383532 outliers (13.55%)

📊 Subflow Fwd Packets: Capped 332693 outliers (11.75%)

📊 Subflow Fwd Bytes: Capped 446972 outliers (15.79%)

📊 Subflow Bwd Packets: Capped 308291 outliers (10.89%)

📊 Subflow Bwd Bytes: Capped 619046 outliers (21.87%)

📊 Init\_Win\_bytes\_forward: Capped 439029 outliers (15.51%)

📊 Init\_Win\_bytes\_backward: Capped 285710 outliers (10.09%)

📊 act\_data\_pkt\_fwd: Capped 310279 outliers (10.96%)

📊 min\_seg\_size\_forward: Capped 2149 outliers (0.08%)

📊 Active Mean: Capped 558829 outliers (19.74%)

📊 Active Std: Capped 205534 outliers (7.26%)

📊 Active Max: Capped 558829 outliers (19.74%)

📊 Active Min: Capped 558829 outliers (19.74%)

📊 Idle Mean: Capped 567283 outliers (20.04%)

📊 Idle Std: Capped 229243 outliers (8.10%)

📊 Idle Max: Capped 567283 outliers (20.04%)

📊 Idle Min: Capped 567283 outliers (20.04%)

✅ Outlier detection complete. Shape maintained: 2830743 rows

🚀 Creating network-specific features...

✅ Created 17 network-specific features

After preprocessing:

Shape: (2830743, 97)

Binary Label:

Binary\_Label

0 2273097

1 557646

Name: count, dtype: int64

Attack Rate: 0.19699633629757277

✅ STEP 2-5 COMPLETED: Data preprocessing successful

================================================================================

🎯 STEP 6: Starting ensemble feature selection...

Initial number of features: 95

🔗 Analyzing feature correlations...

📊 Found 67 highly correlated pairs (threshold: 0.95)

📉 Removed 32 redundant features

📈 Features reduced from 95 to 63

Top 5 most correlated pairs:

• forward\_mean\_activity ↔ forward\_total\_activity: 1.000

• Destination Port\_is\_registered ↔ Destination Port\_is\_wellknown: 1.000

• Avg Fwd Segment Size ↔ Fwd Packet Length Mean: 1.000

• Avg Bwd Segment Size ↔ Bwd Packet Length Mean: 1.000

• Fwd Header Length.1 ↔ Fwd Header Length: 1.000

After removing zero variance: 38

🧠 Applying ensemble feature selection...

Final chosen features: 10

🏆 Top 10 features by ensemble scoring:

1. Flow Bytes/s\_to\_Flow Packets/s\_ratio: 0.982

2. Average Packet Size: 0.440

3. Destination Port: 0.433

4. forward\_mean\_activity: 0.419

5. Packet Length Variance: 0.368

6. Init\_Win\_bytes\_forward: 0.358

7. Subflow Fwd Bytes\_to\_Subflow Bwd Bytes\_ratio: 0.351

8. Min Packet Length: 0.342

9. Bwd Packet Length Min: 0.315

10. Init\_Win\_bytes\_backward: 0.312

✅ STEP 6 COMPLETED: Feature selection successful

================================================================================

📏 STEP 7: Starting data standardization and balancing...

⚖️ Start balancing data ...

Class 0 (BENIGN): 2,273,097

Class 1 (ATTACK): 557,646

After balancing: 1 100000

0 100000

Name: count, dtype: int64

✅ STEP 7 COMPLETED: Data standardization and balancing successful

================================================================================

🎉 PIPELINE COMPLETED: Enhanced preprocessing and feature selection completed!

📊 Final dataset shape: (200000, 10)

📋 Selected features: 10

🚀 Dataset is now ready for model training!

================================================================================

================ Random Forest Results ================

Accuracy: 0.9973

Precision: 0.9987

Recall: 0.9958

F1 Score: 0.9973

Confusion Matrix:

[[20071 25]

[ 83 19821]]

Classification Report:

precision recall f1-score support

0 1.00 1.00 1.00 20096

1 1.00 1.00 1.00 19904

accuracy 1.00 40000

macro avg 1.00 1.00 1.00 40000

weighted avg 1.00 1.00 1.00 40000

======================================================

================ Decision Tree Results ================

Accuracy: 0.9969

Precision: 0.9980

Recall: 0.9957

F1 Score: 0.9969

Confusion Matrix:

[[20057 39]

[ 85 19819]]

Classification Report:

precision recall f1-score support

0 1.00 1.00 1.00 20096

1 1.00 1.00 1.00 19904

accuracy 1.00 40000

macro avg 1.00 1.00 1.00 40000

weighted avg 1.00 1.00 1.00 40000

======================================================

================ KNN Results ================

Accuracy: 0.9953

Precision: 0.9940

Recall: 0.9966

F1 Score: 0.9953

Confusion Matrix:

[[19977 119]

[ 67 19837]]

Classification Report:

precision recall f1-score support

0 1.00 0.99 1.00 20096

1 0.99 1.00 1.00 19904

accuracy 1.00 40000

macro avg 1.00 1.00 1.00 40000

weighted avg 1.00 1.00 1.00 40000

======================================================

================ SVM Results ================

Accuracy: 0.9727

Precision: 0.9560

Recall: 0.9907

F1 Score: 0.9730

Confusion Matrix:

[[19188 908]

[ 185 19719]]

Classification Report:

precision recall f1-score support

0 0.99 0.95 0.97 20096

1 0.96 0.99 0.97 19904

accuracy 0.97 40000

macro avg 0.97 0.97 0.97 40000

weighted avg 0.97 0.97 0.97 40000

======================================================

🏆 Best model by F1 score: Random Forest (0.9973)

✅ Saved best model as ddos\_best\_model.joblib

✅ Saved scaler as ddos\_scaler.joblib