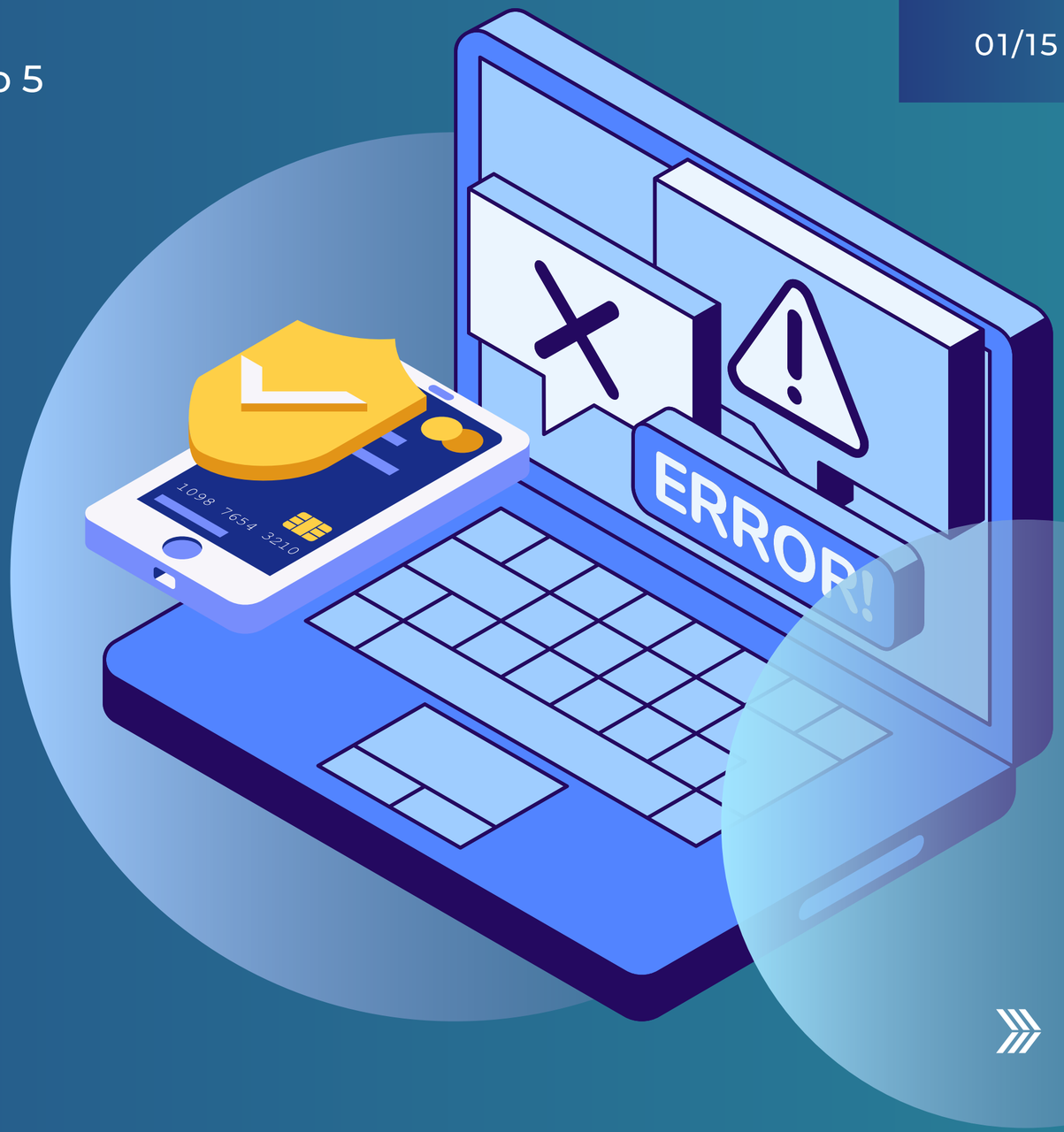
A Machine Learning Approach By group 5

DDOS DETECTION SYSTEM



OUTLINE

INTRODUCTION & OBJECTIVES

- DDoS attacks disrupt services by flooding networks
- Aim: develop ML pipeline to distinguish benign vs. attack traffic
- Requirements: high accuracy, low false alarms, real-time capability

DATA OVERVIEW

- Source: Kaggle – chethuhn/network-intrusion-dataset - CIC-IDS- 2017
- Records: 2,830,743 flows; Features (initial): 79
- Class distribution: 80.3% Benign (2,273,097), 19.7% Attack (557,646)

DATA PREPROCESSING

- Exploratory Data Analysis (EDA)
- Outlier Handling: IQR-based capping to preserve data integrity
- Missing & Infinite Values: median imputation
- Normalization: StandardScaler (zero mean, unit variance)
- Class Balancing: Undersampling to 100k samples per class

FEATURE ENGINEERING & SELECTION

Engineered 17 domain-specific features: flow ratios, traffic asymmetry, port classification

Removed redundant features (corr > 0.95) & zero-variance

Applied ensemble selection: F-test (0.4), MI (0.3), RF importance (0.3)

Top 10 features selected for modeling

Top 10 Đặc Trưng theo F-Score filecite turn7file0 L26-L33:

Hạng	Đặc Trưng	F-Score	Giải thích an ninh
1	Flow Bytes/s_to_Flow Packets/s_ratio	665,469.22	Tỷ lệ byte/gói bất thường trong lưu lượng bùng phát
2	Min Packet Length	412,077.32	Kích thước gói nhỏ nhất bất thường
3	Bwd Packet Length Min	327,263.43	Độ biến thiên kích thước gói ngược cho lưu lượng phản hồi
4	Bwd Packet Length Max_to_Min_ratio	191,499.72	Biến động kích thước gói ngược cho thấy nhiễu tấn công
5	Packet Length Variance	140,937.65	Độ nhất quán kích thước gói trong các cuộc tấn công DDoS
6	Fwd IAT Std	122,818.07	Độ lệch chuẩn thời gian giữa các gói tới luật tấn công
7	forward_mean_activity	120,198.36	Hoạt động trung bình chiều tới phản ánh các đợt bùng nổ
8	Avg Bwd Segment Size	117,990.12	Kích thước trung bình phân đoạn ngược cho lưu lượng hoàn thiện
9	Average Packet Size	116,279.41	Phân phối kích thước gói chung khác biệt giữa benign và attack
10	Avg Fwd Segment Size	114,051.74	Sự nhất quán phân đoạn chiều tới trong chuỗi tấn công

MODELING METHODOLOGY

Model	Key Hyperparameters
Decision Tree	criterion='gini', max_depth=None
Random Forest	n_estimators=50, random_state=42
KNN	n_neighbors=5, metric='minkowski'
SVM	kernel='rbf', C=1.0, gamma='scale'

EVALUATION RESULTS

Model	Precision	Recall	F1 Score	Accuracy
Decision Tree	0.9980	0.9957	0.9969	0.9969
Random Forest	0.9987	0.9958	0.9973	0.9973
KNN	0.9940	0.9966	0.9953	0.9953
SVM	0.9560	0.9907	0.9730	0.9727

CONCLUSION

FUTURE WORK

- Random Forest achieved the best balance (F1 = 0.9973)
- Pipeline ready for real-time deployment (FastAPI, Docker)
- Future: integrate deep learning, extend to other attack types, stream processing

THANK YOU!