# Detecting Host Based intrusion using Machine Learning

Edwin Lin

# AGENDA

01 Intro to IDE

02 Problem Definition

03 Host-Based dataset

04 Machine Learning Models

05 Application and results

06 Conclusion

# intrusion detection system (IDS)

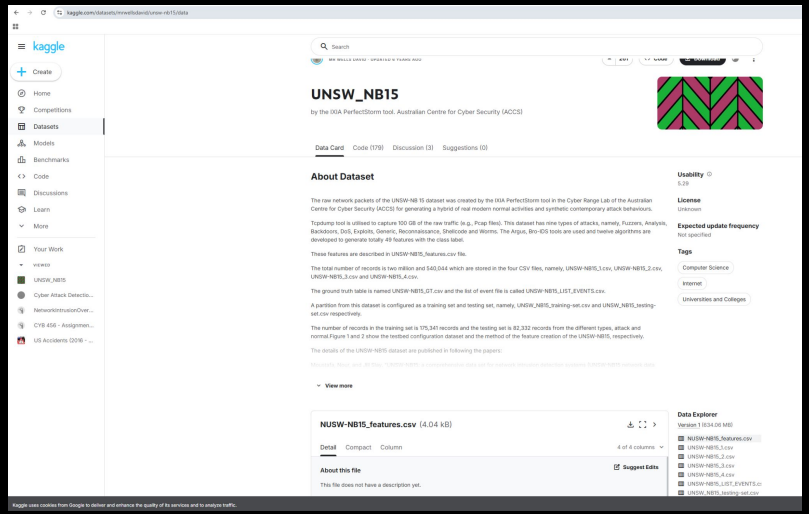
Misuse-based methods: detect known attacks by using signatures of those attacks

Anomaly-based techniques: model the normal network and system behavior and
identify anomalies as deviations from normal behavior.

Hybrid techniques: combine misuse and anomaly detection

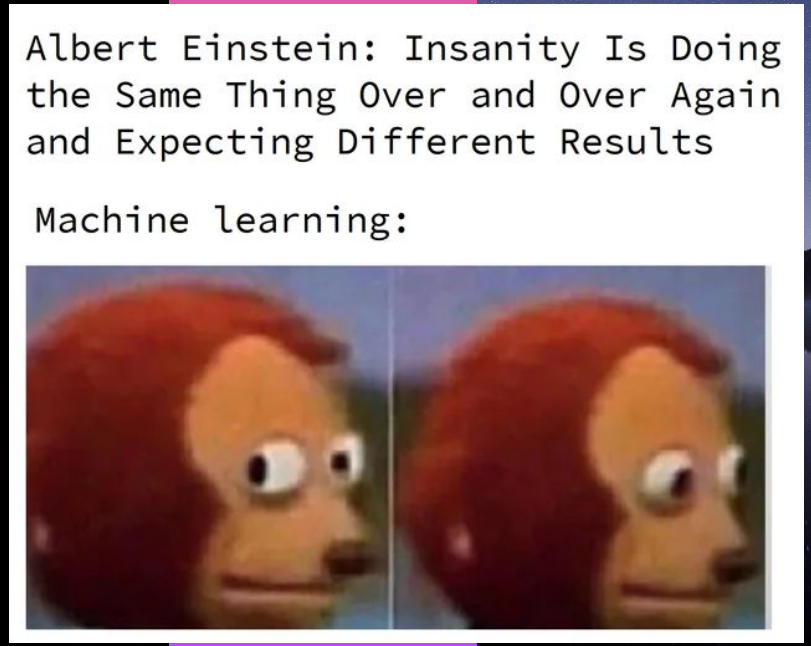
# Problem Statement

Using a hybrid detection method on a UNSW_NB15 dataset for host based intrusion detection using machine learning



# What is Machine Learning?

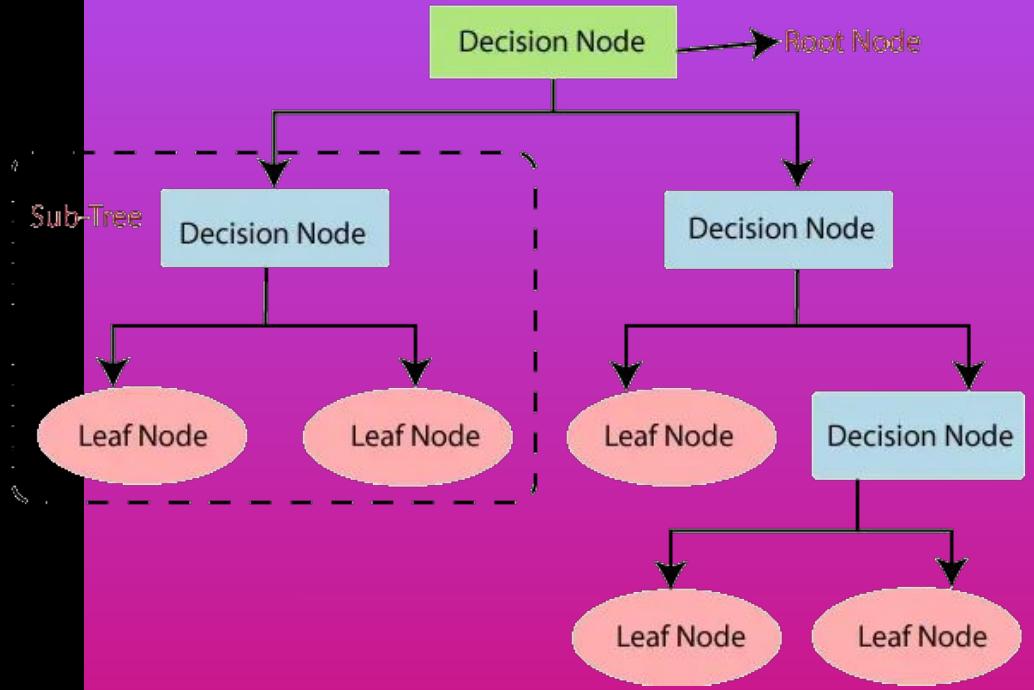
Machine Learning is like if you combine statistics, coding, and telling it to predict



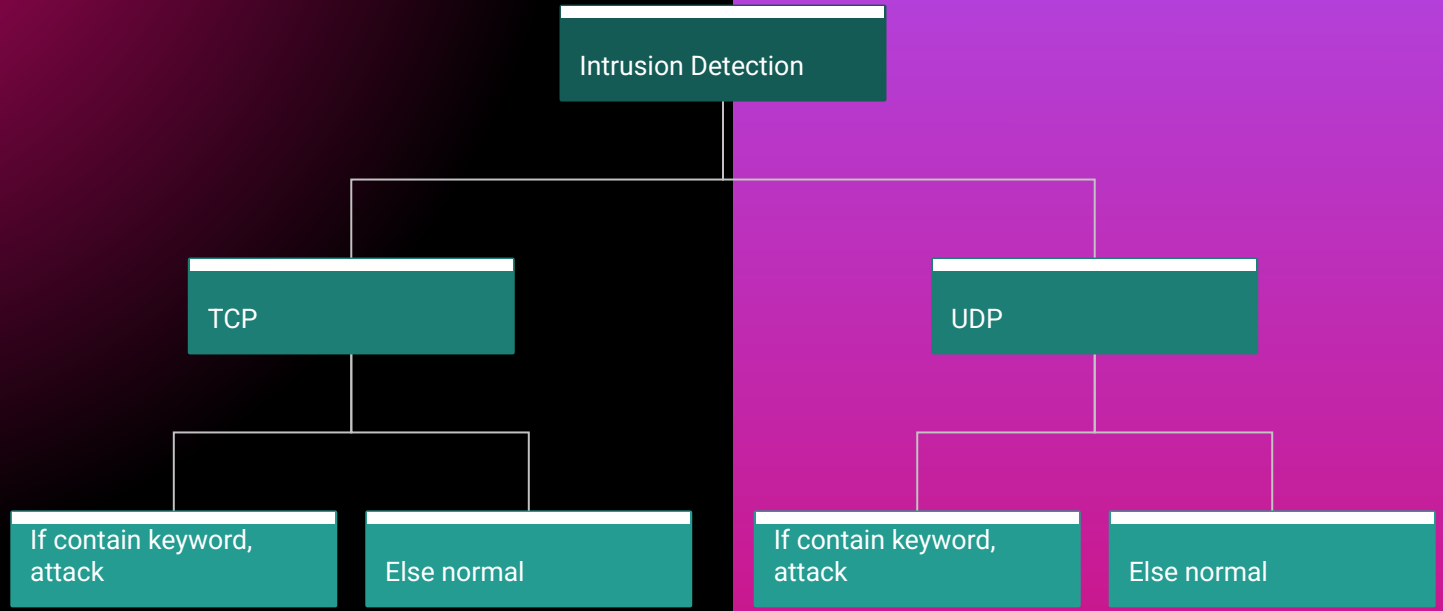
# Decision Tree

- Our Misused based method
- Supervised Learning
- 2 Labeling options:
  Normal = 0
  Attack = 1

Each node level represents a feature, based on that feature, makes a decision



# Example:


Intrusion Detection
TCP
UDP
If contain keyword, attack
Else normal
If contain keyword, attack
Else normal

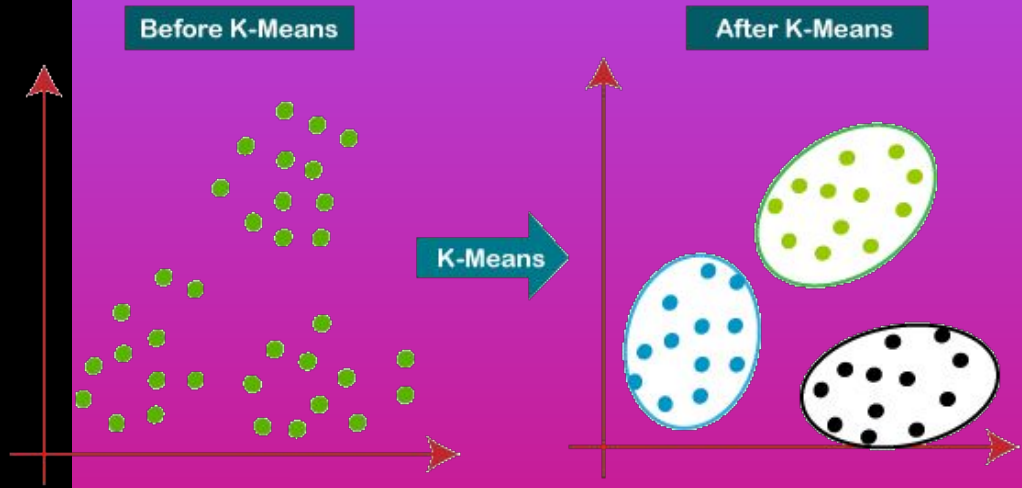
# K-Means clustering

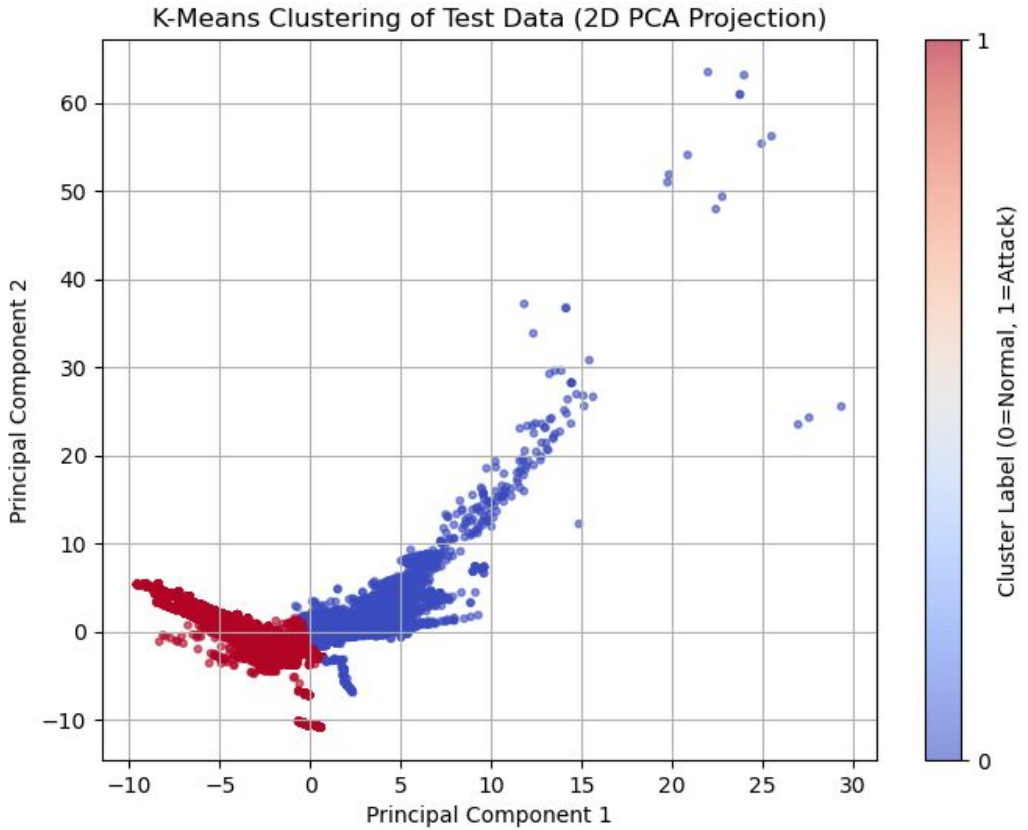
Anomaly-based approach

Unsupervised Learning

Find interesting patterns for Anomalies

Centroid clusters



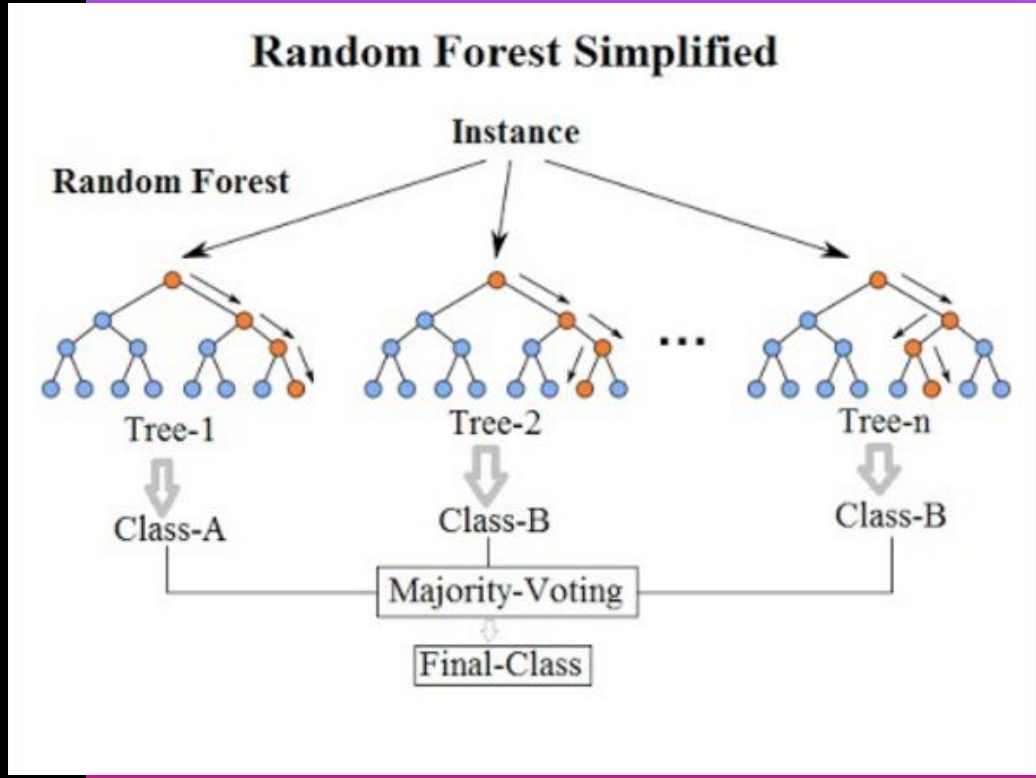

K-Means Clustering of Test Data (2D PCA Projection)
Principal Component 2
Principal Component 1
Cluster Label (0=Normal, 1=Attack)
60
50
40
30
20
10
0
−10
−10
−5
0
5
10
15
20
25
30
1
0

# Random Forest

Hybrid approach

Multiple decision trees

Ensembled ML model



# Performance Analysis

K means performance the worst as expected

High False Positives as expected for normal

99% accuracy in finding attacks
Great performance.

Decision Tree Report:

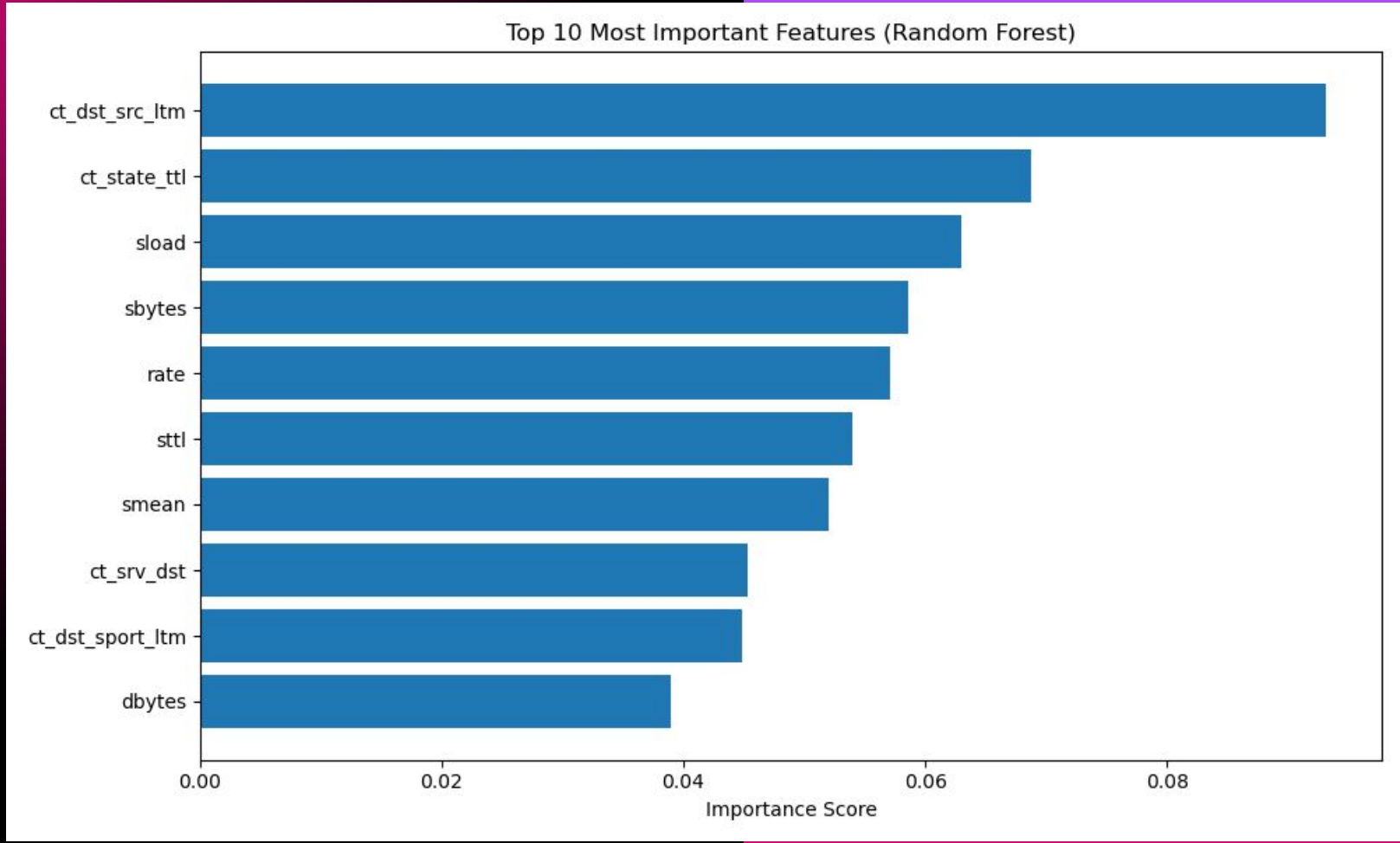
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.77 | 0.97 | 0.86 | 56000 |
| 1 | 0.98 | 0.86 | 0.92 | 119341 |
| accuracy | | | 0.90 | 175341 |
| macro avg | 0.88 | 0.91 | 0.89 | 175341 |
| weighted avg | 0.91 | 0.90 | 0.90 | 175341 |

K-Means Report (Anomaly Detection):

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.54 | 0.88 | 0.67 | 56000 |
| 1 | 0.92 | 0.65 | 0.76 | 119341 |
| accuracy | | | 0.72 | 175341 |
| macro avg | 0.73 | 0.77 | 0.72 | 175341 |
| weighted avg | 0.80 | 0.72 | 0.73 | 175341 |

Random Forest Report (Hybrid Detection):

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.77 | 0.98 | 0.86 | 56000 |
| 1 | 0.99 | 0.86 | 0.92 | 119341 |
| accuracy | | | 0.90 | 175341 |
| macro avg | 0.88 | 0.92 | 0.89 | 175341 |
| weighted avg | 0.92 | 0.90 | 0.90 | 175341 |


Top 10 Most Important Features (Random Forest)
ct_dst_src_ltm
ct_state_ttl
sload
sbytes
rate
sttl
smean
ct_srv_dst
ct_dst_sport_ltm
dbytes
0.00
0.02
0.04
0.06
0.08
Importance Score

# THANK YOU

Any questions?