Intrusion Detection
KDD CUP 99 Data Set

[1]

Name: Wilmar QUIROGA

Project: Intrusion detection

Class: Machine Learning for networks
Professor: Andrea ARALLDO

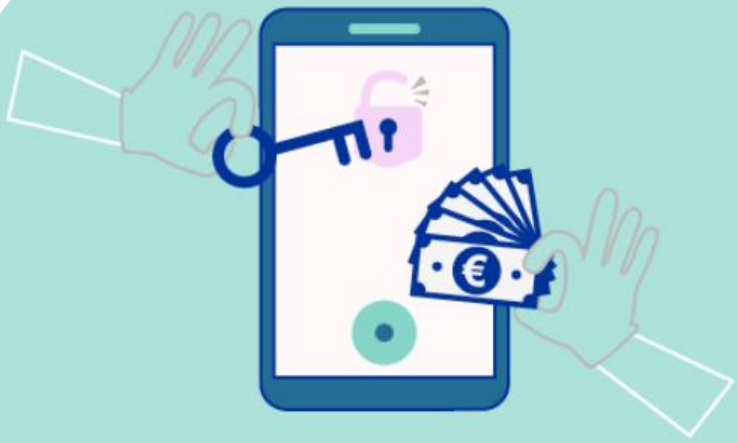
Introduction

The importance of cyber security is given because according to the European Commission:

- In 2020, the amount of data stolen on a monthly basis in the EU exceeded 10 terabytes.
- In the EU, ransomware stands out as a major cyber threat.
- Distributed Denial of Service (DDoS) attacks also rank among the highest threats.
- The estimated annual cost of cybercrime to the global economy was €5.5 trillion at the end of 2020 [3]

[2]

Top Cyber Threats in Europe



Ransomware attacks

Attacks where cybercriminals take control of a target's asset and demand a ransom to restore its availability.

60% of affected organisations may have paid ransom demands.

Distributed denial-of-service (DDoS) threats

Attacks preventing users of a network or a system from accessing relevant information, services and other resources.

July 2022 saw the largest ever recorded attack against a European customer.

[3]

Social engineering threats

Threats that attempt to exploit a human error or human behaviour to gain access to information or services.

82% of data breaches involved a human element.

Supply-chain attacks

An attack strategy targeting an organisation through vulnerabilities in its supply chain with the potential to induce cascading effects.

Supply chain incidents accounted for 17% of intrusions in 2021 compared to less than 1% in 2020.

[3]

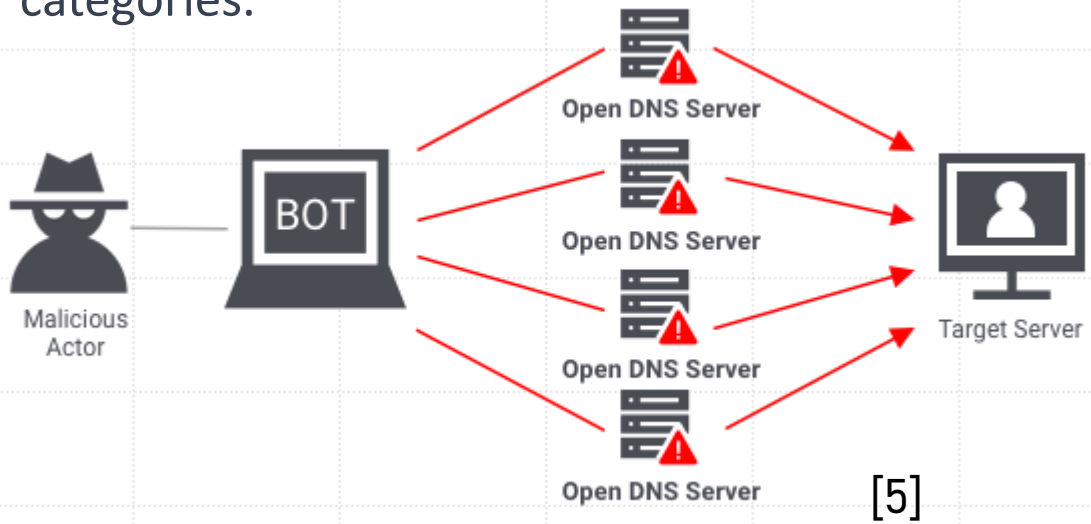
KDD CUP 99 Data Set

- The DARPA'98 IDS evaluation program collected around 4 gigabytes of compressed raw (binary) tcpdump and System events data over 7 weeks of network traffic, so The KDD training dataset has roughly 4.9 million single connection vectors with 41 features, labeled as either normal or an attack

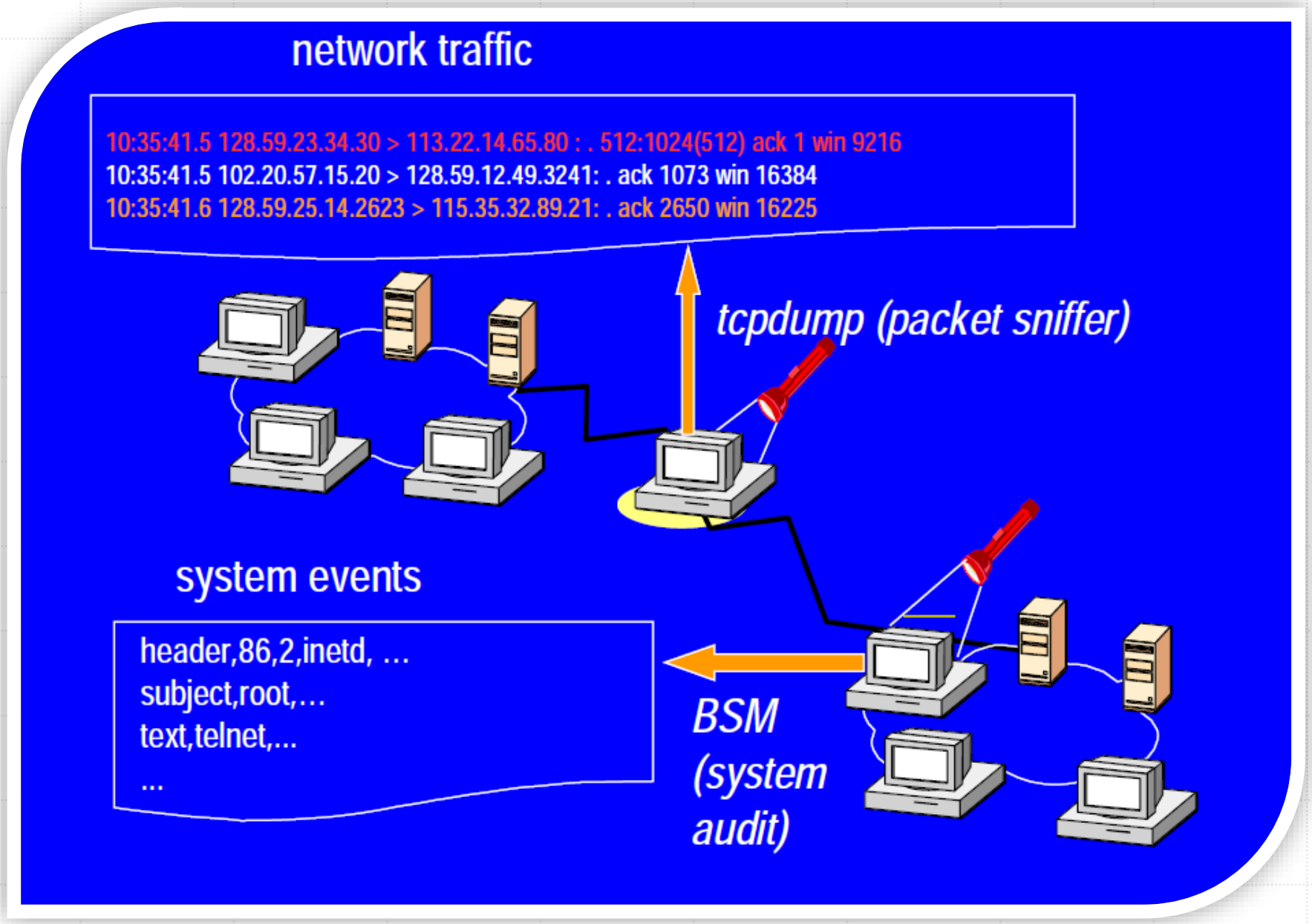
[4]

- The simulated attacks belong to one of four categories.

1) Denial of Service Attack (DoS):
2) User to Root Attack (U2R):
3) Remote to Local Attack (R2L):
4) Probing Attack:



[5]


network traffic
10:35:41.5 128.59.23.34.30 > 113.22.14.65.80 : . 512:1024(512) ack 1 win 9216
10:35:41.5 102.20.57.15.20 > 128.59.12.49.3241: . ack 1073 win 16384
10:35:41.6 128.59.25.14.2623 > 115.35.32.89.21: . ack 2650 win 16225
tcpdump (packet sniffer)
system events
header,86,2,inetd, …
subject,root,…
text,telnet,...
...
BSM
(system
audit)

[6]

Features in the Dataset

Basic features: This group includes all the characteristics that can be obtained from a TCP/IP connection, many of which result in a delay in detection.

Traffic features: This category is composed of features that are calculated within a window interval and is divided into two groups.

"same host" features: This group of features analyzes only connections from the past 2 seconds that have the same destination host as the current connection.

"same service" features: This category involves analyzing only the connections that have the same service as the current connection within the past 2 seconds.

Features:

Basic Features	Type
Duration	Int64
Protocol type	Object(Categorical)
Service	Object(Categorical)
Flag	Object(Categorical)
Source bytes	int64
Destination Bytes	Int64
Land	Int64
Wrong Fragment	int64
Urgent	int64
Hot	int64

Content Features	Type
num_failed_logins	Int64
logged_in	int64
num_compromised	int64
root_shell	int64
su_attempted	int64
num_root	Int64
num_file_creations	Int64
num_shells	int64
num_access_files	int64
num_outbound_cms	int64
is_host_login	int64
is_guest_login	int64
count	int64

Traffic Features	Type
srv_count	Int64
serror_rate	Float64
srv_serrot_rate	Float64
rerror_rate	Float64
srv_serror_rate	Float64
same_srv_rate	Float64
diff_srv_rate	Float64
srv_diff_host_rate	Float64
dst_host_count	int64
dst_host_srv_count	int64
dst_host_same_srv_rate	int64
dst_host_diff_srv_rate	int64
dst_host_same_src_port_rate	int64
dst_host_srv_diff_host_rate	int64
dst_host_serror_rate	int64
dst_host_srv_serror_rate	int64
dst_host_rerror_rate	int64
dst_host_srv_rerror_rate	int64

Categorize the types of attacks

Denial of Service Attack (DoS)

- Back
- Land
- Neptune
- Pod
- Smurf
- Teardrop

User to Root Attack (U2R)

- Buffer overflow
- Loadmodule
- perl
- rootkit

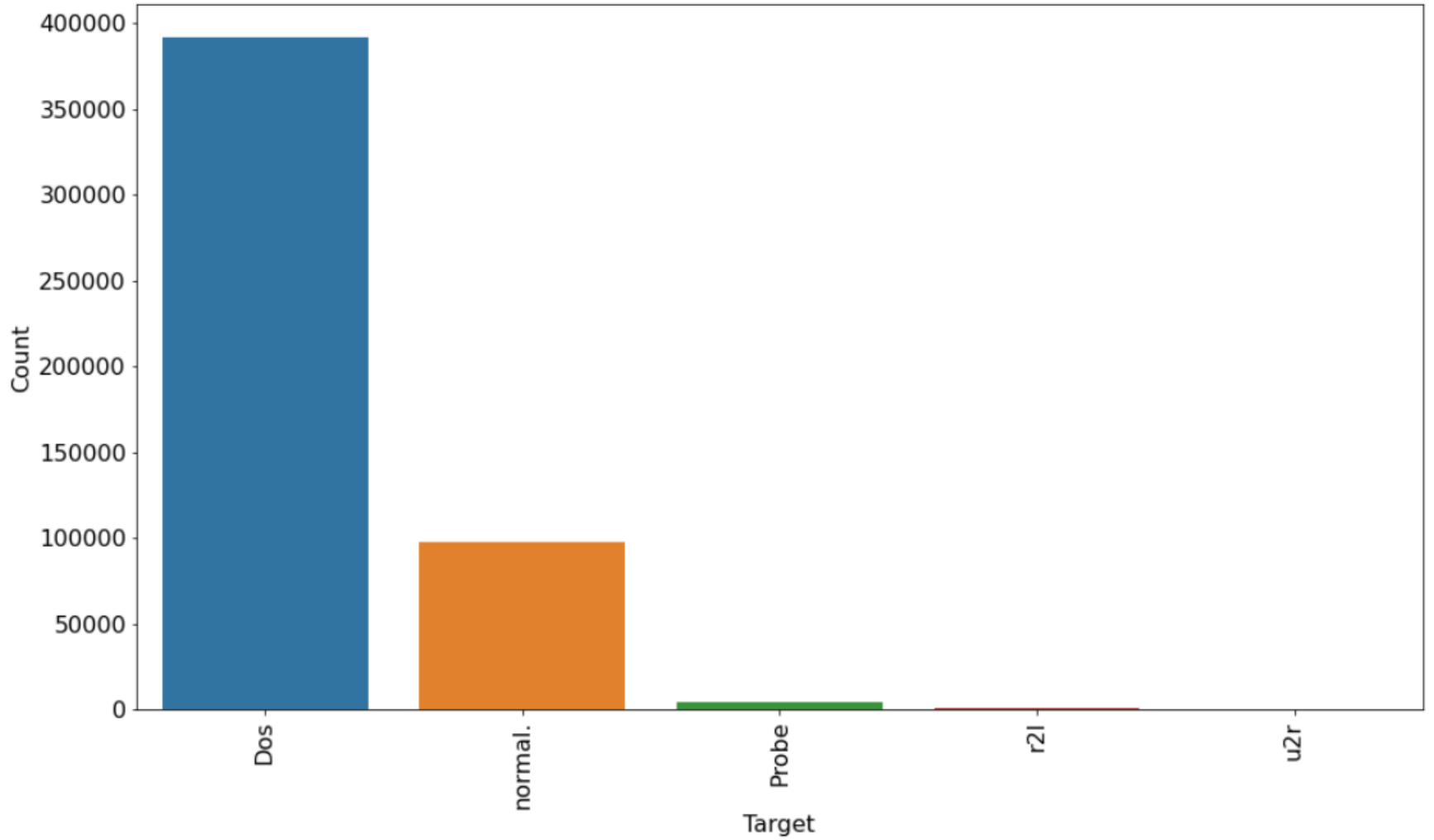
Probing Attack:

- Ip sweep
- Nmap
- Port sweep
- Satan

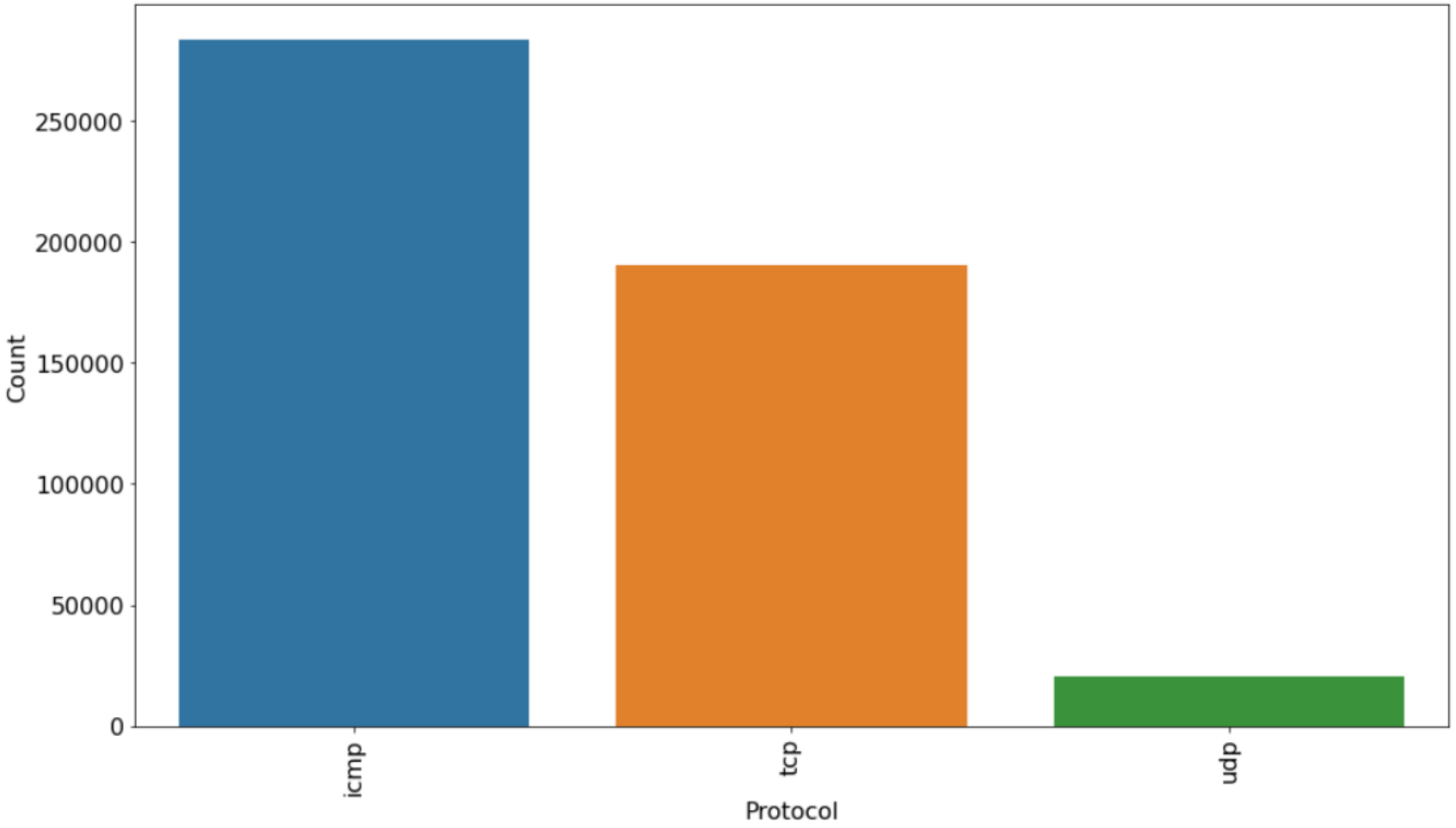
Remote to Local Attack (R2L)

- ftp write
- Guess password
- imap
- Multihop
- Phf
- Spy
- Warez client
- Warez master

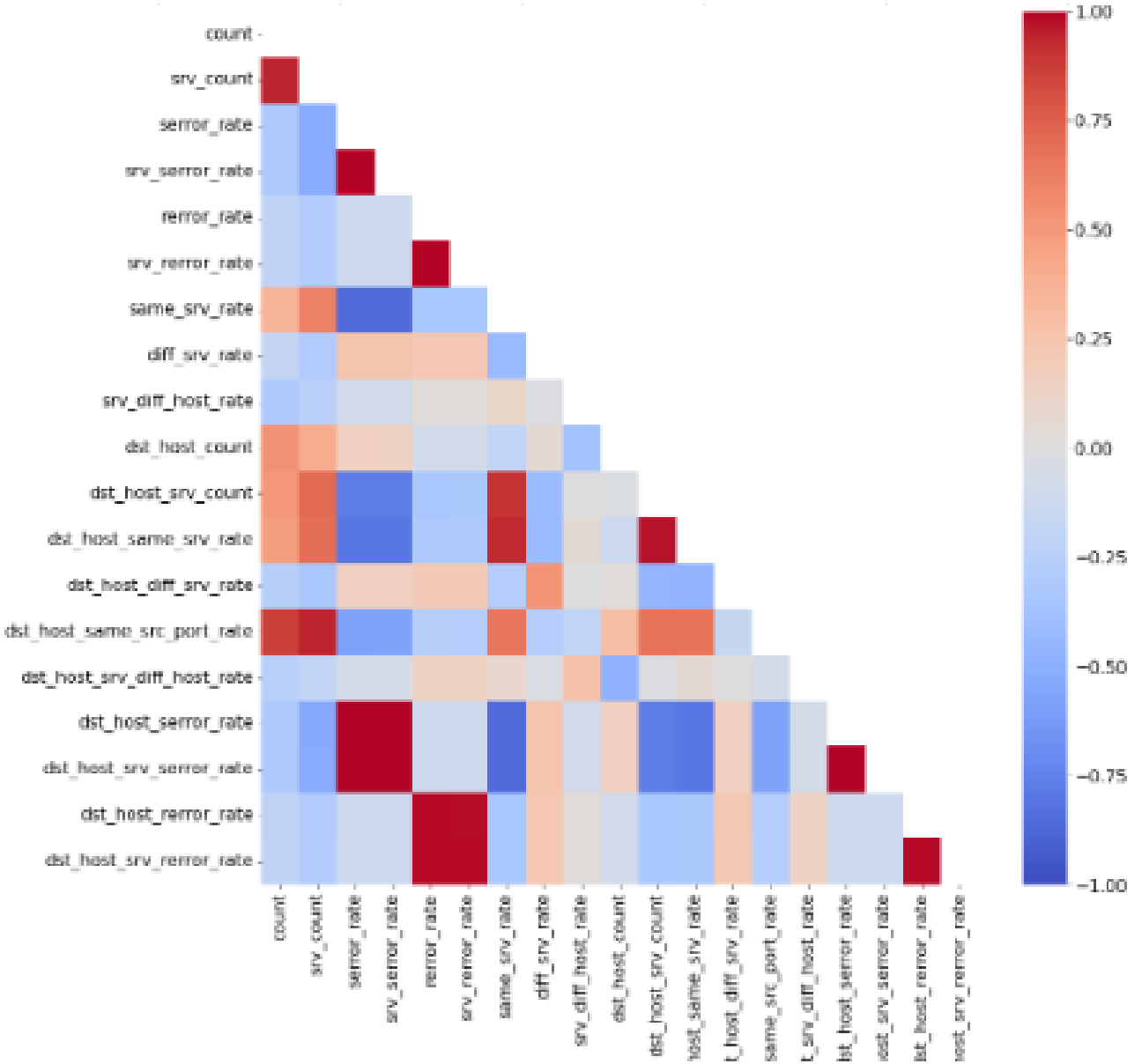
Distribution of number of attacks by category



Protocol distribution



Finding correlated variables



Most correlated variables

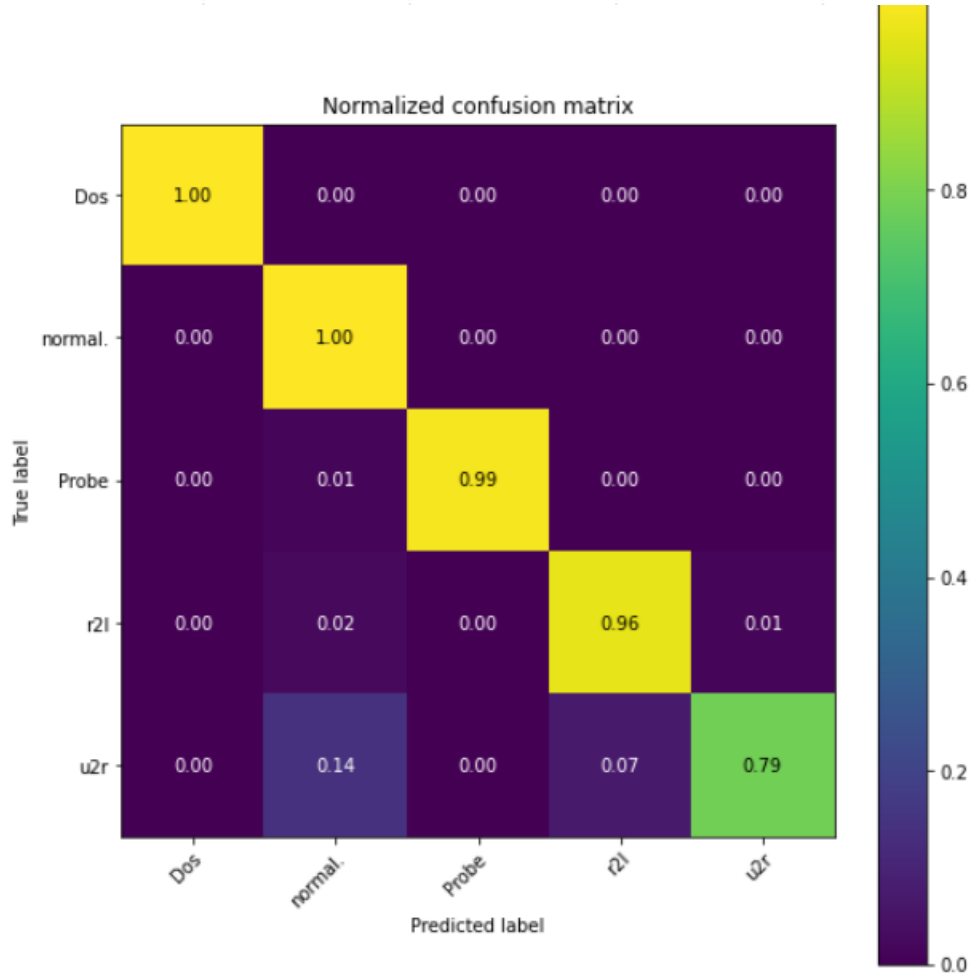
Variable 1	Variable 2	Correlation
serror-rate	dst_host_srv_serror_rate	0.997
serror-rate	dst_host_serror_rate	0.998
srv_count	dst_host_same_src_port_rate	0.994
count	dst_host_same_src_port_rate	0.86
rerror_rate	dst_host_srv_serror_rate	0.985
rerror_rate	dst_host_rerror_rate	0.986
srv_rerror_rate	dst_host_srv_rerror_rate	0.986
srv_rerror_rate	dst_host_rerror_rate	0.986
same_srv_rate	dst_host_same_srv_rate	0.927
same_srv_rate	dst_host_srv_count	0.898
hot	is_guest_login	0.843
num_root	num_compromised	0.993
num_compromised	su_attempted	0.701

Eliminate the most correlated characteristics and those that have a unique value such as num_outbound_cmds and is_host_login

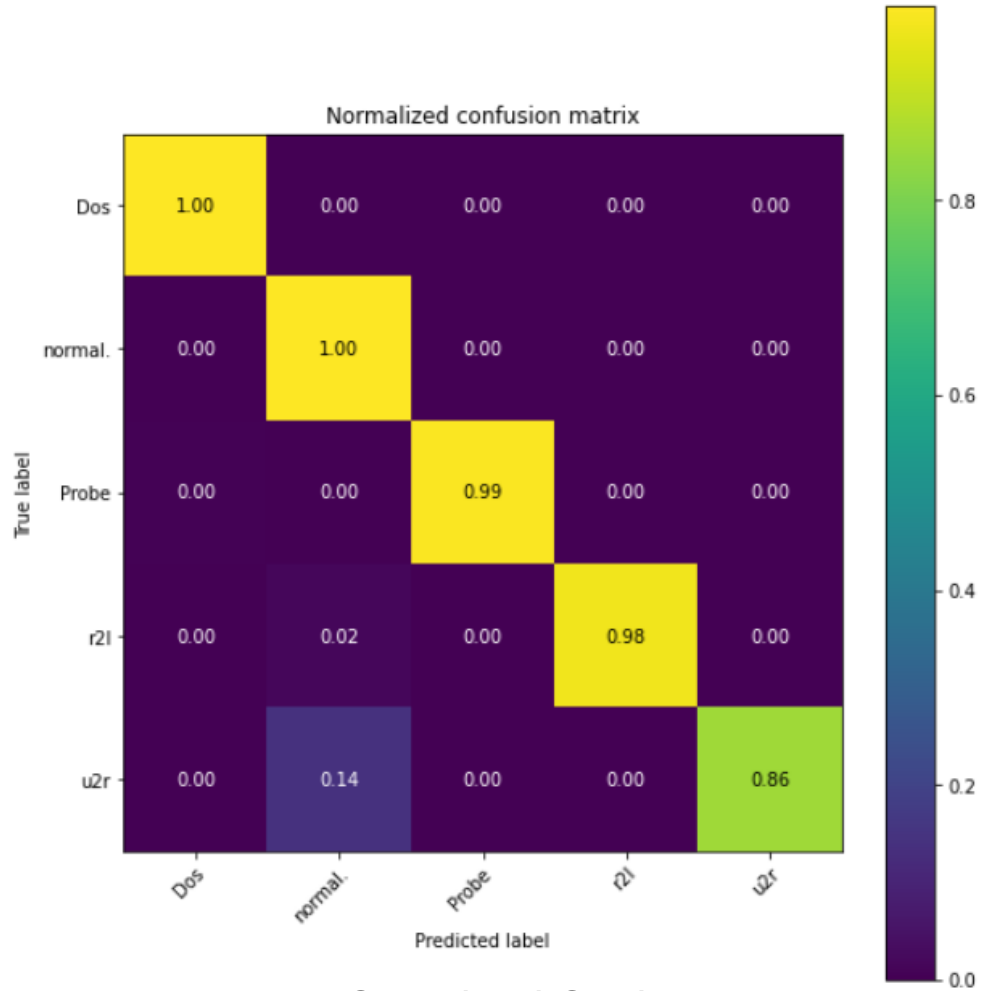
Models Performance

Model	MME Train	MME Test	Train Accuracy	Test Accuracy
Decision Tree Classifier	0.003	0.034	0.999	0.999
Logistic Regression	0.111	0.119	0.992	0.991
Neural Networks	0.179	0.178	0.975	0.975
Support vector machine	0.114	0.118	0.993	0.992

Confusion Matrix Decision Tree Classifier

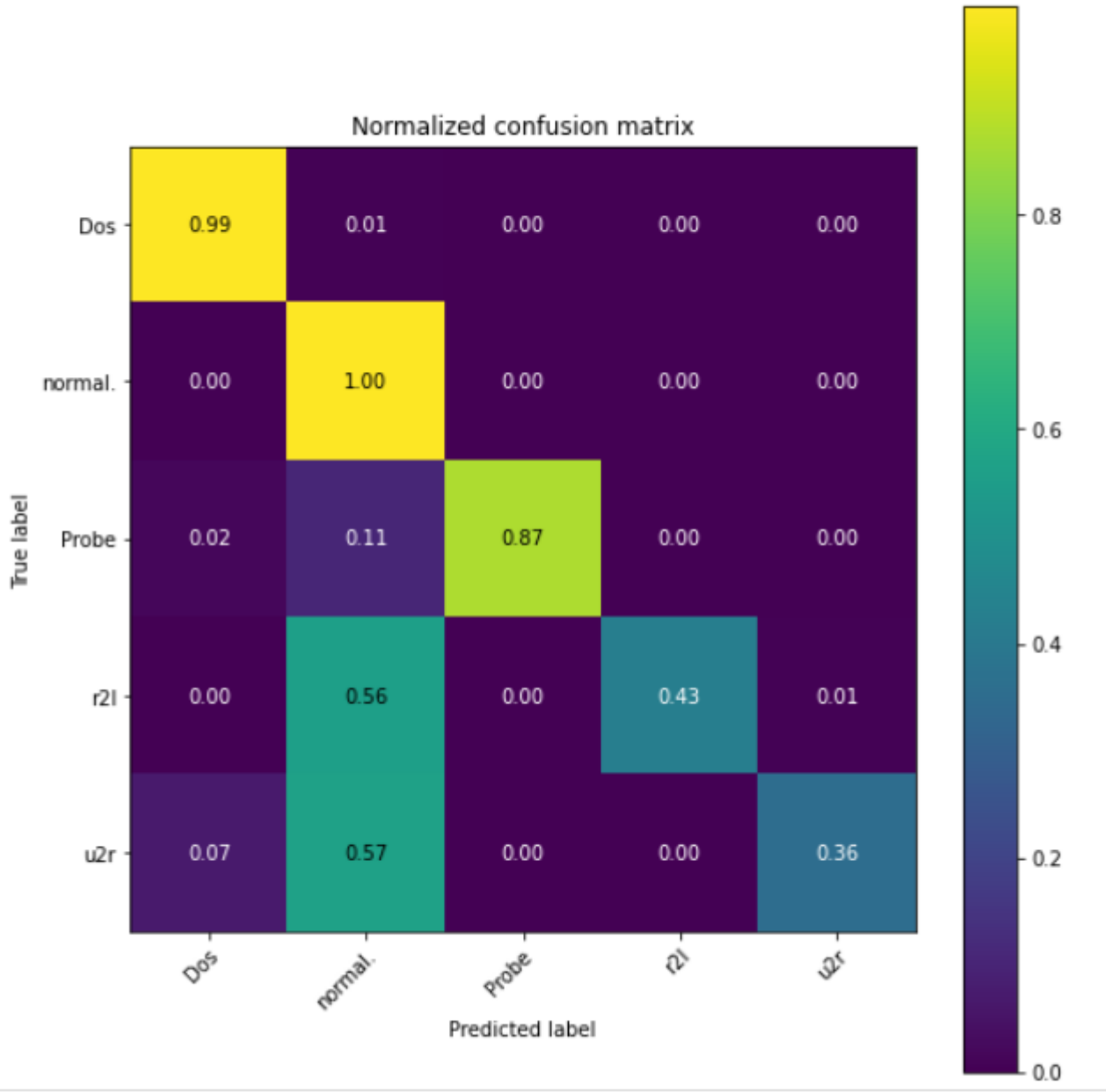


MinMax Scaler

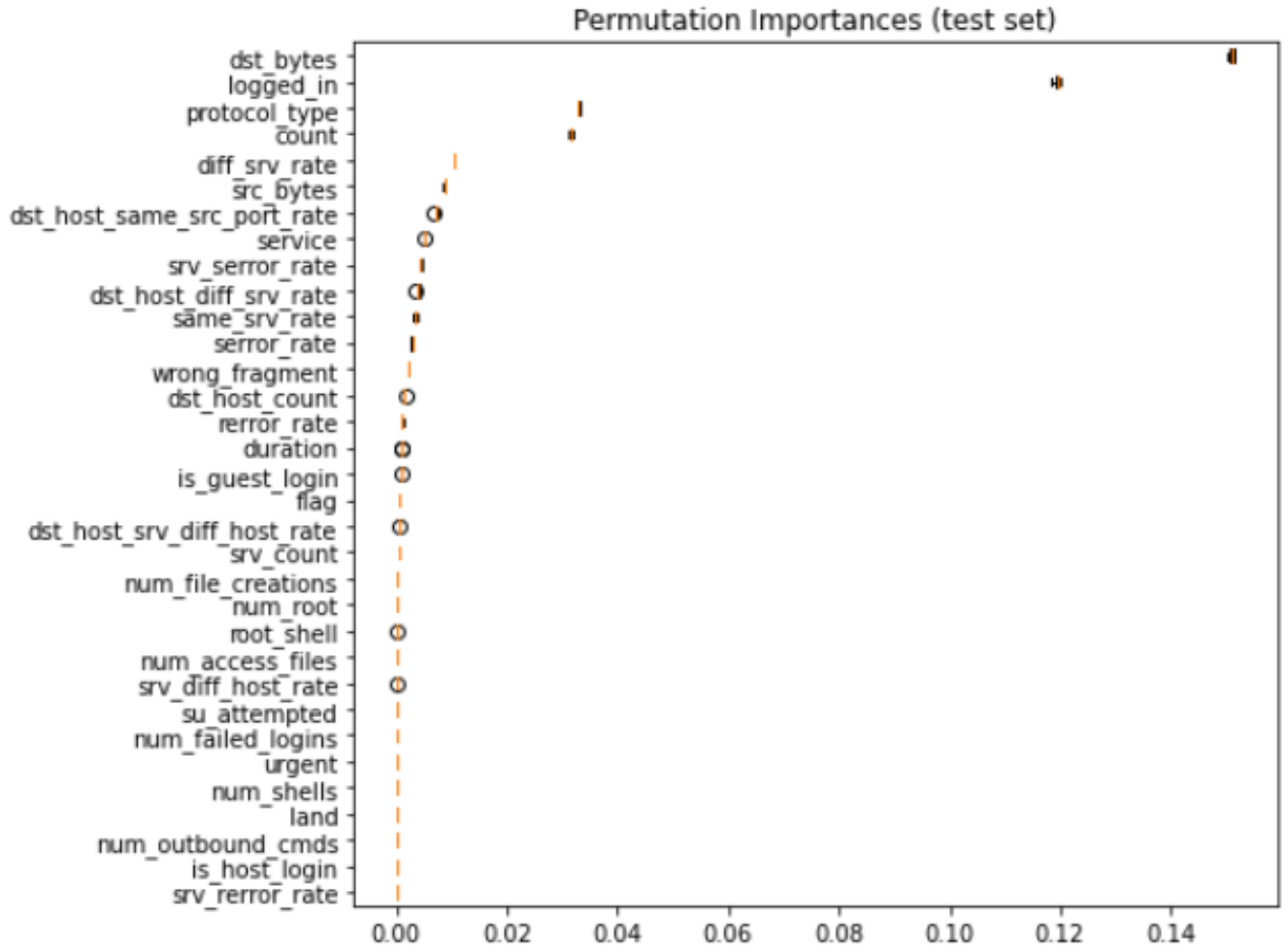


Standard Scaler

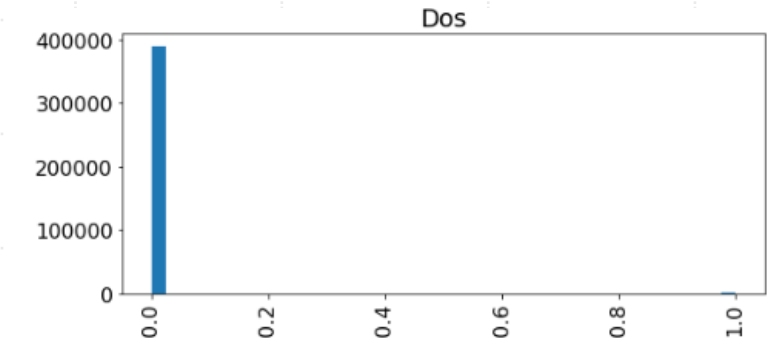
Confusion Matrix Logistic Regression

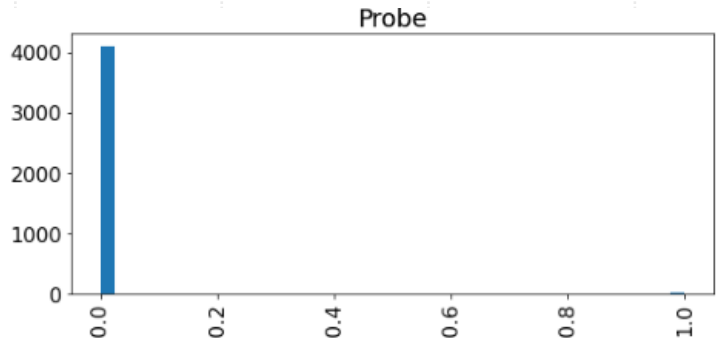


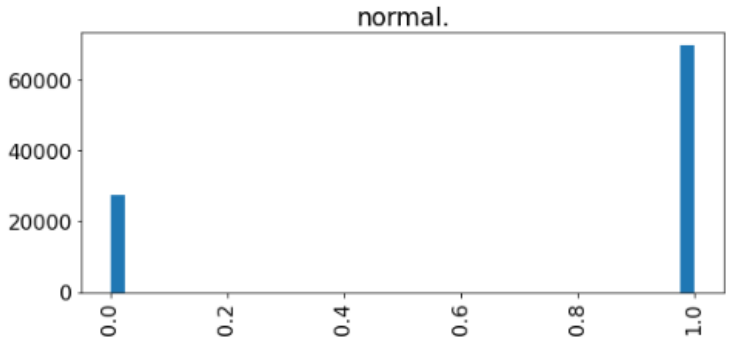
Feature importance Decision Tree Classifier

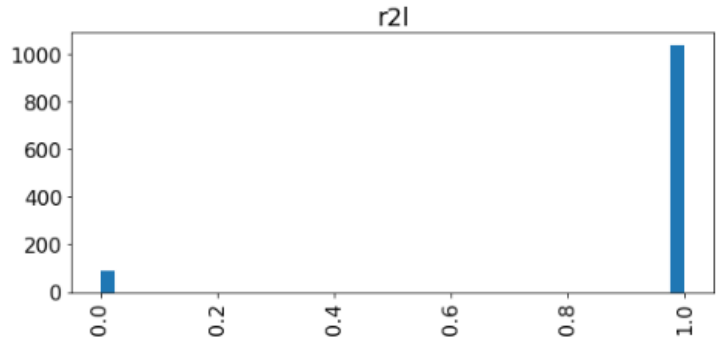


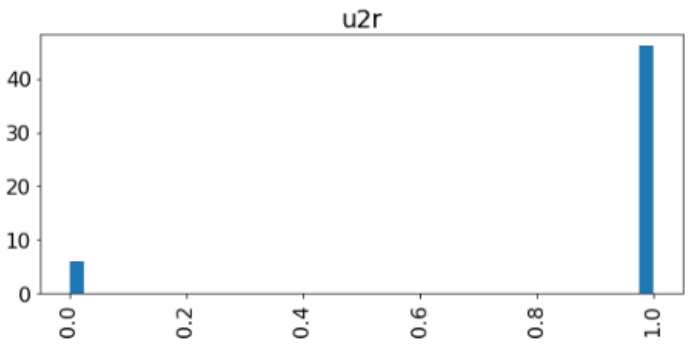
Histogram of loggin_in feature Group by Target

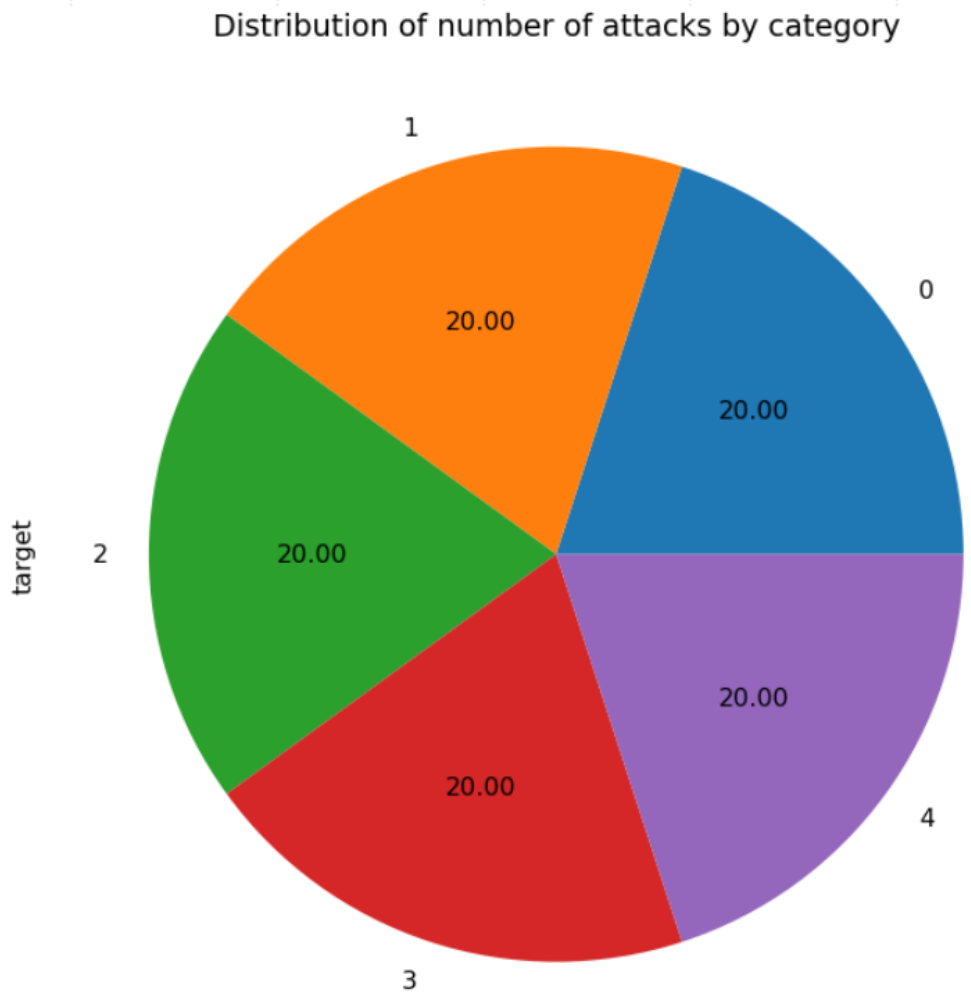












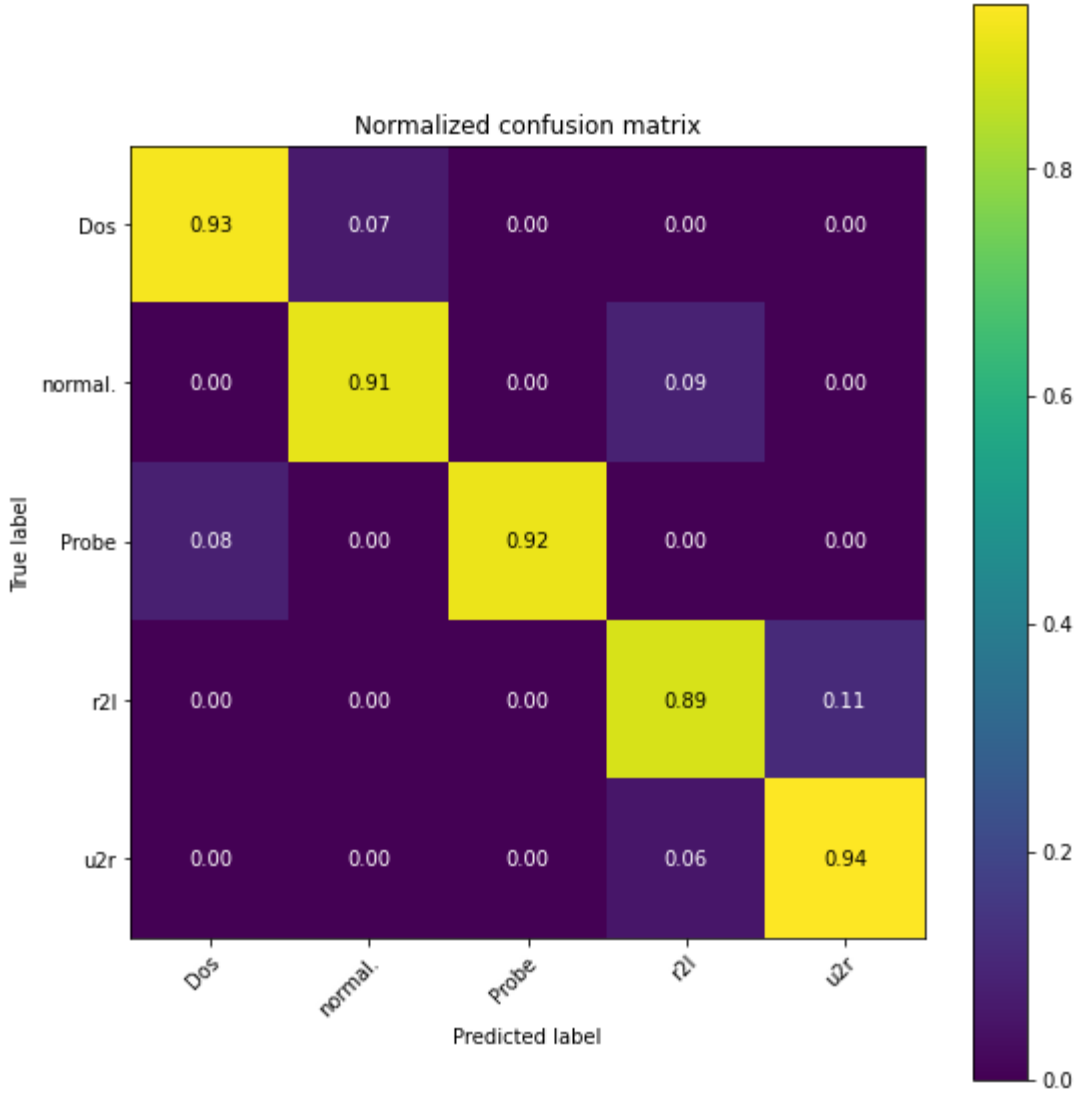
Balanced Dataset

- Dos: 0
- Normal: 1
- Probe: 2
- r2l: 3
- u2r: 4

Decision Tree Classifier

Model	MME Train	MME Test	Train Accuracy	Test Accuracy
Decision Tree Classifier	0.0	0.411	1.0	0.923

Confusion Matrix Balanced Dataset



Conclusions

- The dataset is unbalanced, that is, although we have a very good accuracy, the less frequent classes such as U2R have a lower accuracy.
- The dataset has many repeated samples meaning that some samples may appear in the test which may increase the accuracy.

References:

- [1] https://towardsdatascience.com/building-an-intrusion-detection-system-using-deep-learning-b9488332b321
- [2] https://kratikal.com/blog/watch-out-for-these-5-major-network-security-attacks/
- [3] https://www.consilium.europa.eu/en/infographics/cyber-threats-eu/
- [4] https://www.ecb.torontomu.ca/~bagheri/papers/cisda.pdf
- [5] https://www.ecb.torontomu.ca/~bagheri/papers/cisda.pdf
- [6] http://wenke.gtisc.gatech.edu/ids-readings.html