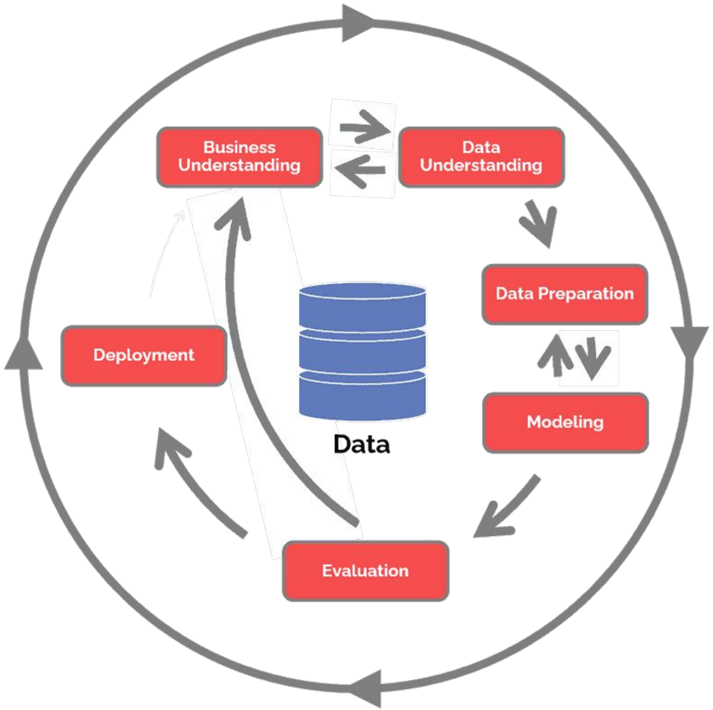
Project
Machine Learning

4th Year Data Science 9

Our Team

Plan

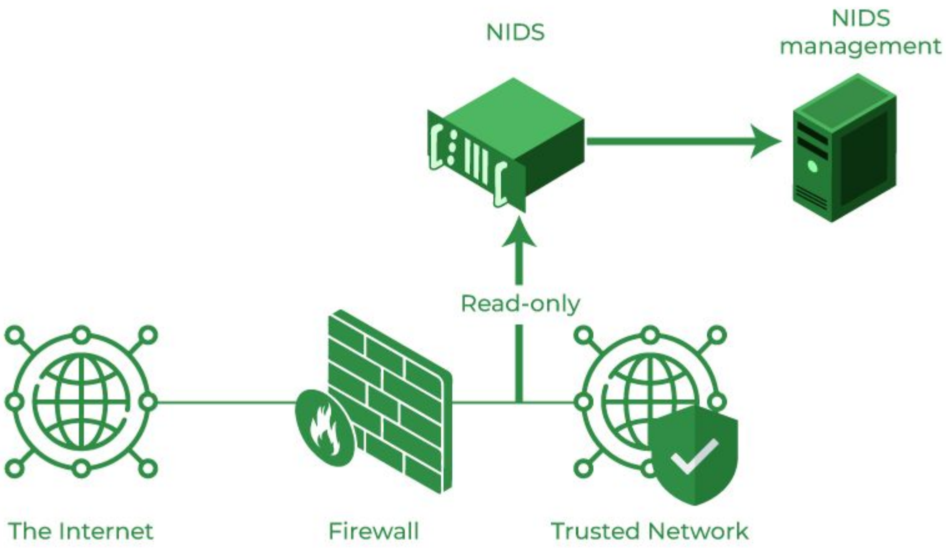
1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Deployment



1

Business Understanding

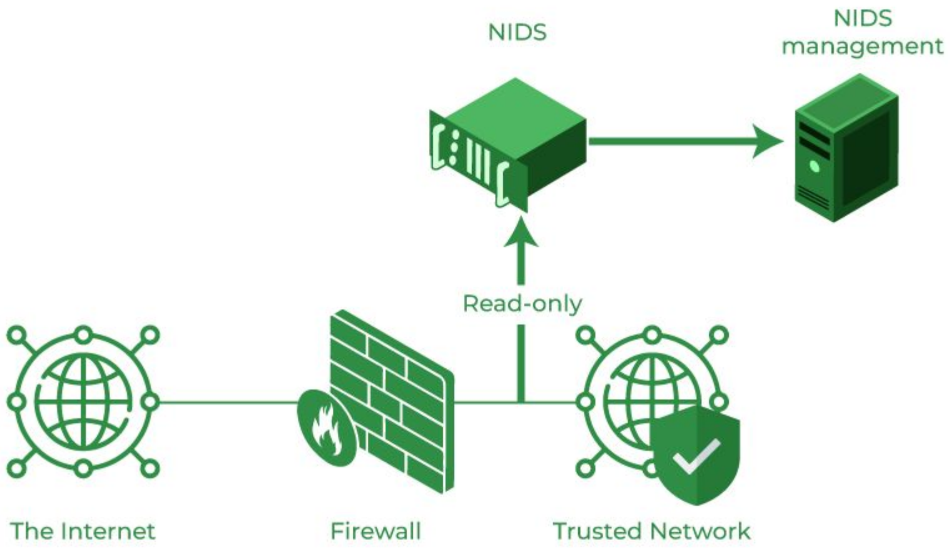
Intrusion Detection System: Introduction



The rapid evolution of technology has led to a significant increase in cybersecurity threats.

Intrusion attacks, whether external or internal, have become increasingly sophisticated, jeopardizing the confidentiality, integrity, and availability of data.

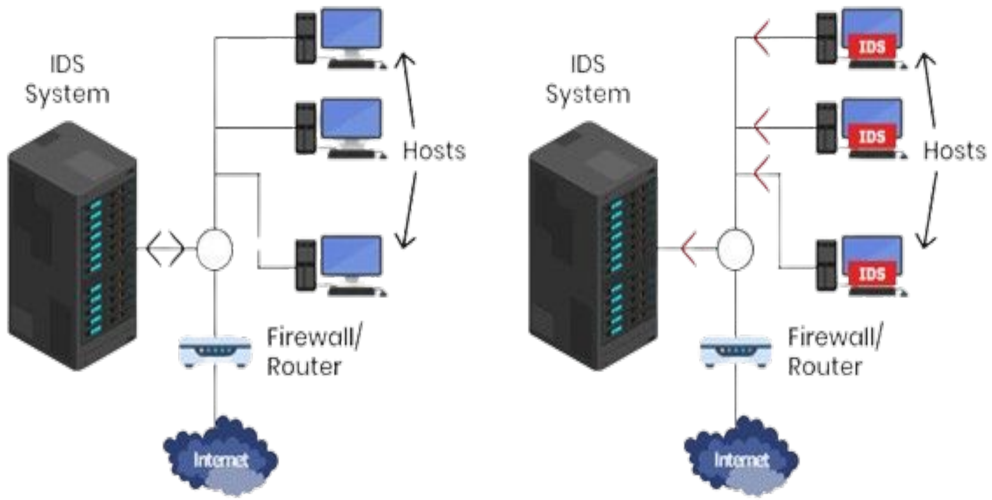
Intrusion Detection System: Introduction



In this context, implementing an Intrusion Detection System (IDS) using machine learning techniques becomes crucial to prevent, detect, and respond to potential malicious activities.

This project aims to apply various supervised and unsupervised machine learning algorithms to detect these malicious intrusions.

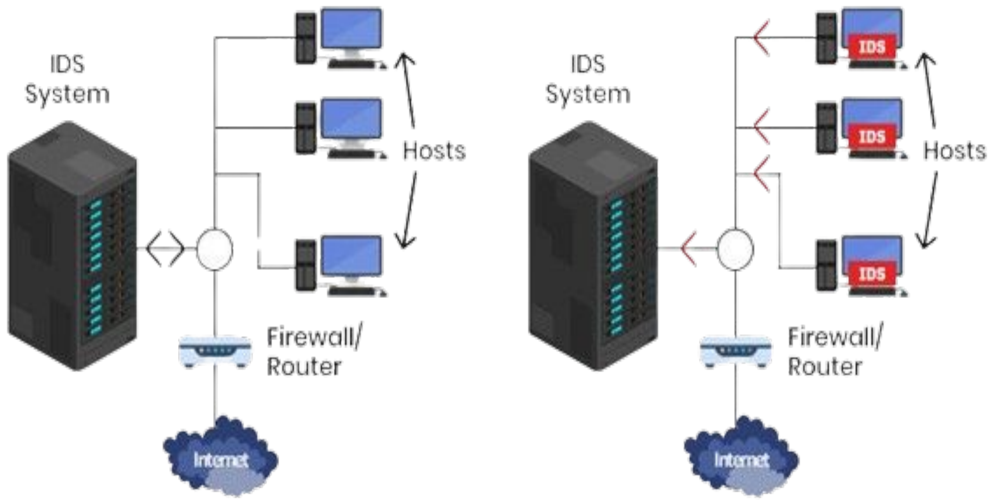
Intrusion Detection System: Families



1. **NETWORK Intrusion detection system**

- **The most widespread IDS**
- **Very Useful to the administrator to understand his network in real time.**
- **Can be placed in different points in the network**

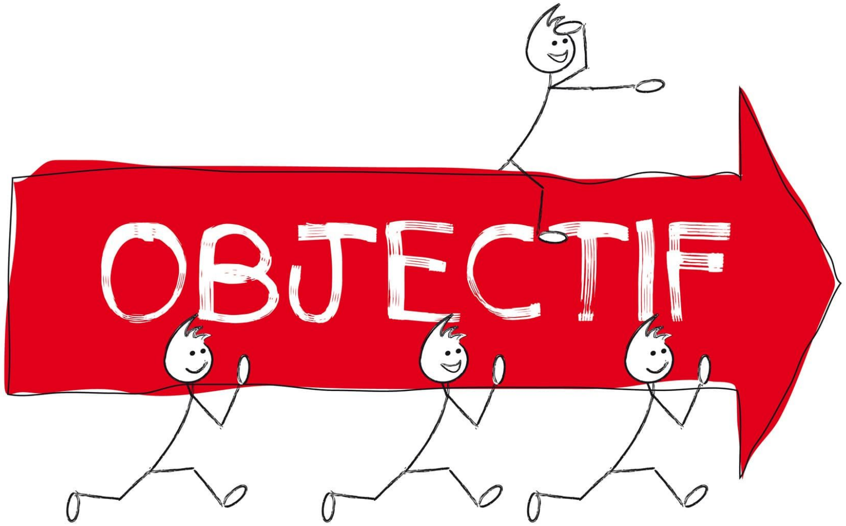
Intrusion Detection System: Families



1. **HOST Intrusion detection system**

    - **Deployed directly on the hosts being monitored**
    - **The analyses are strictly limited to the host on which the HIDS is installed**
    - **Act similarly to antivirus software.**

Intrusion Detection System: Objectifs



- Detection of suspicious activities
- Classification of Attacks

Intrusion Detection System: Statistics

01

By 2024, over 40% of new IDS solutions will utilize AI and ML techniques.

02

Approximately 80% of large enterprises use IDS to monitor their networks

03

Attacks targeting the UDP protocol can be more complex to detect because UDP is connectionless

2

Data Understanding

Data Understanding: Categories of Data

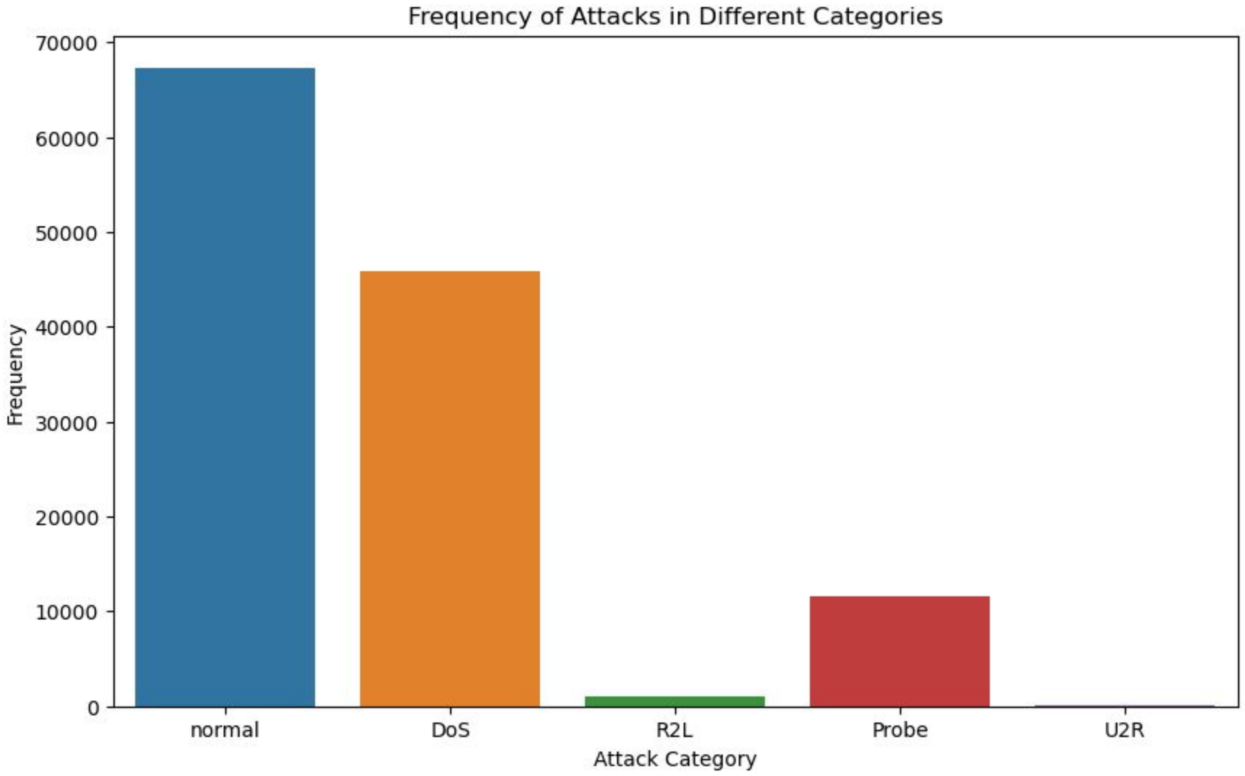
01 Intrinsic Features : Columns 1 to 9

02 Content Features : Columns 10 to 22

03 Time-Based Features : Columns 23 to 31

04 Host-Based Features : Columns 32 to 41

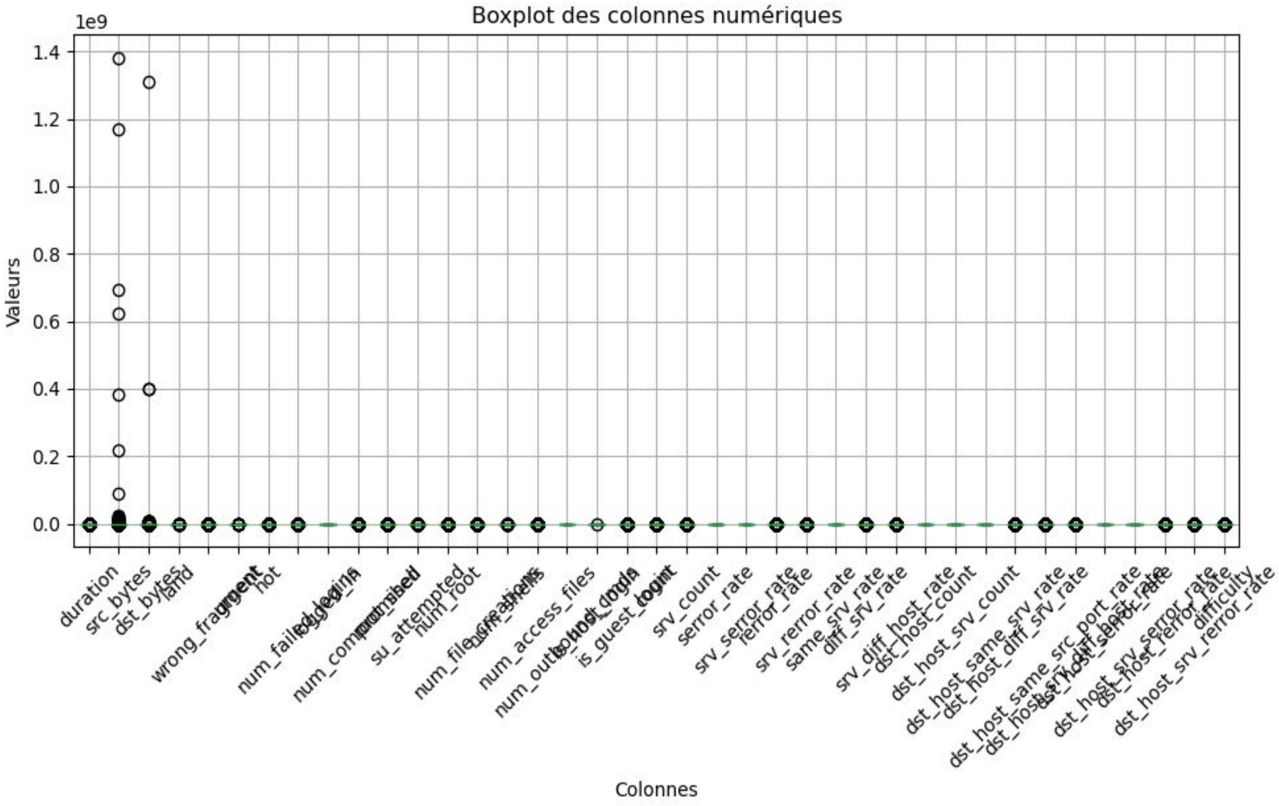
Data Understanding: Categories of Attacks



ATTACKS

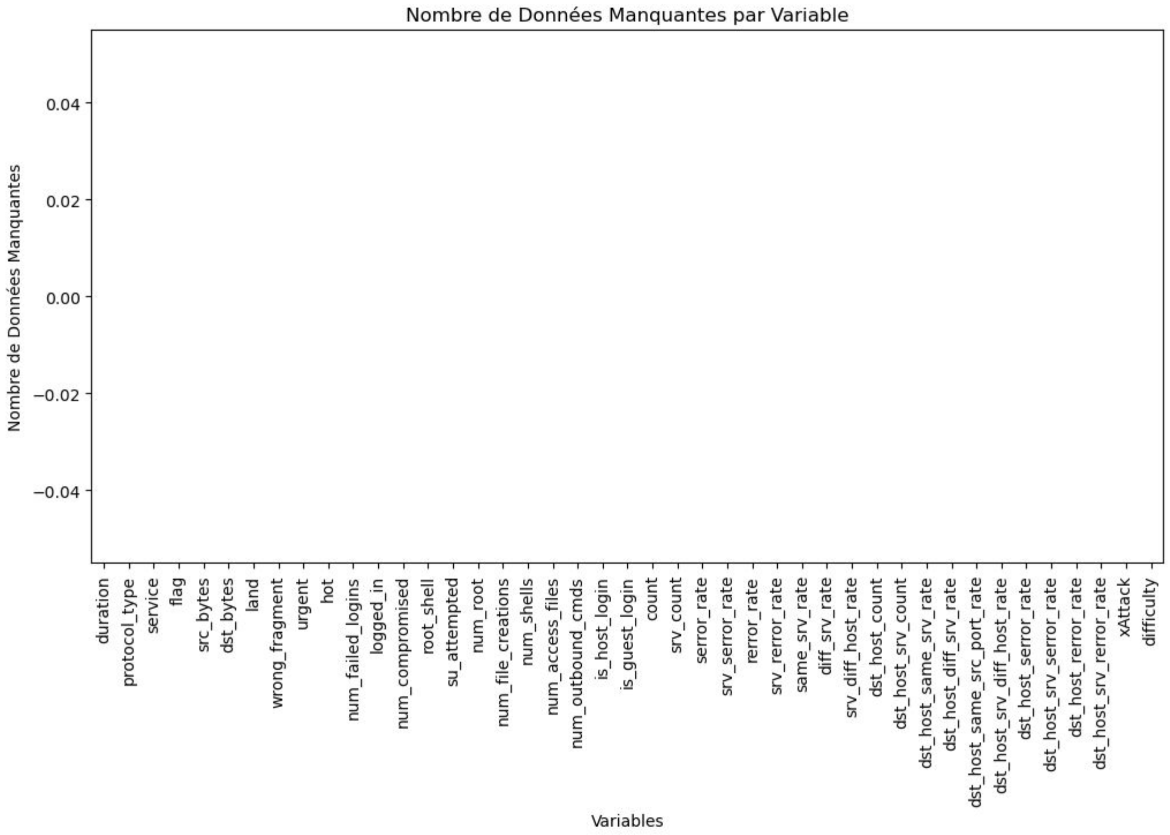
- **Normal**
- **Dos**
- **R2L**
- **Probe**
- **U2R**

Data Understanding: Outliers



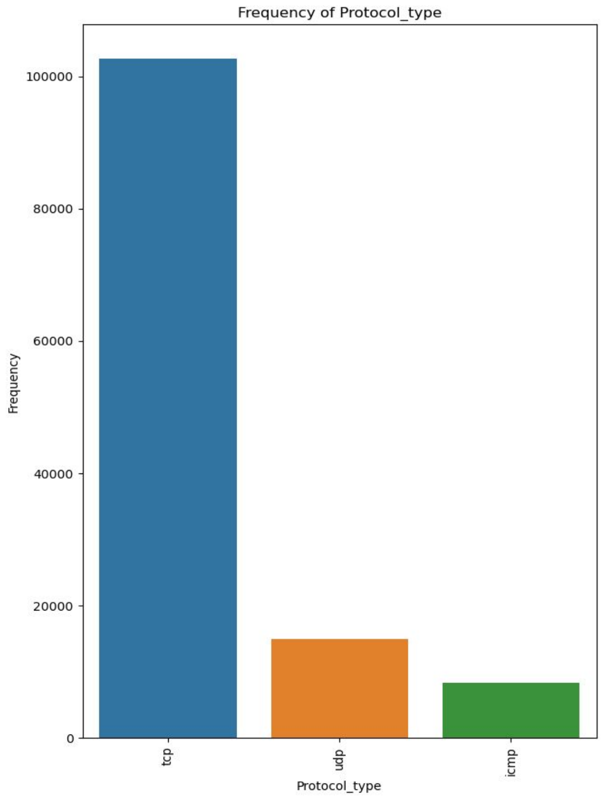
We notice the presence of some outliers in the 'duration' and 'src_bytes' variables.

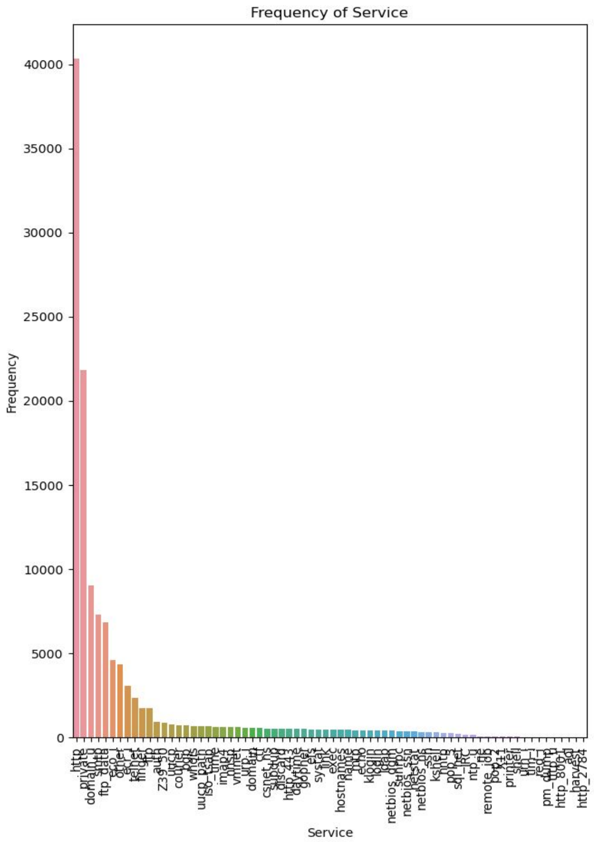
Data Understanding: Missing Values

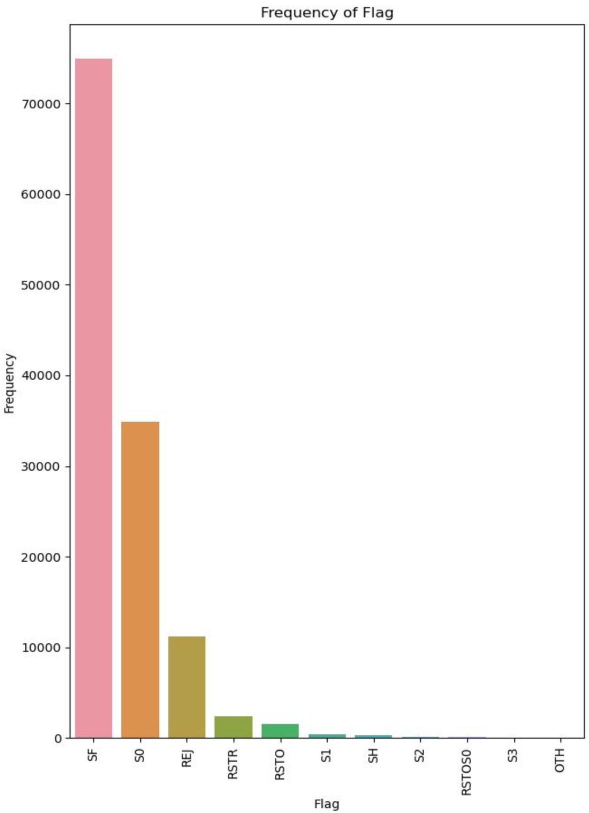


We notice that there are NO missing data in our dataset

Data Understanding: Categorical Variables

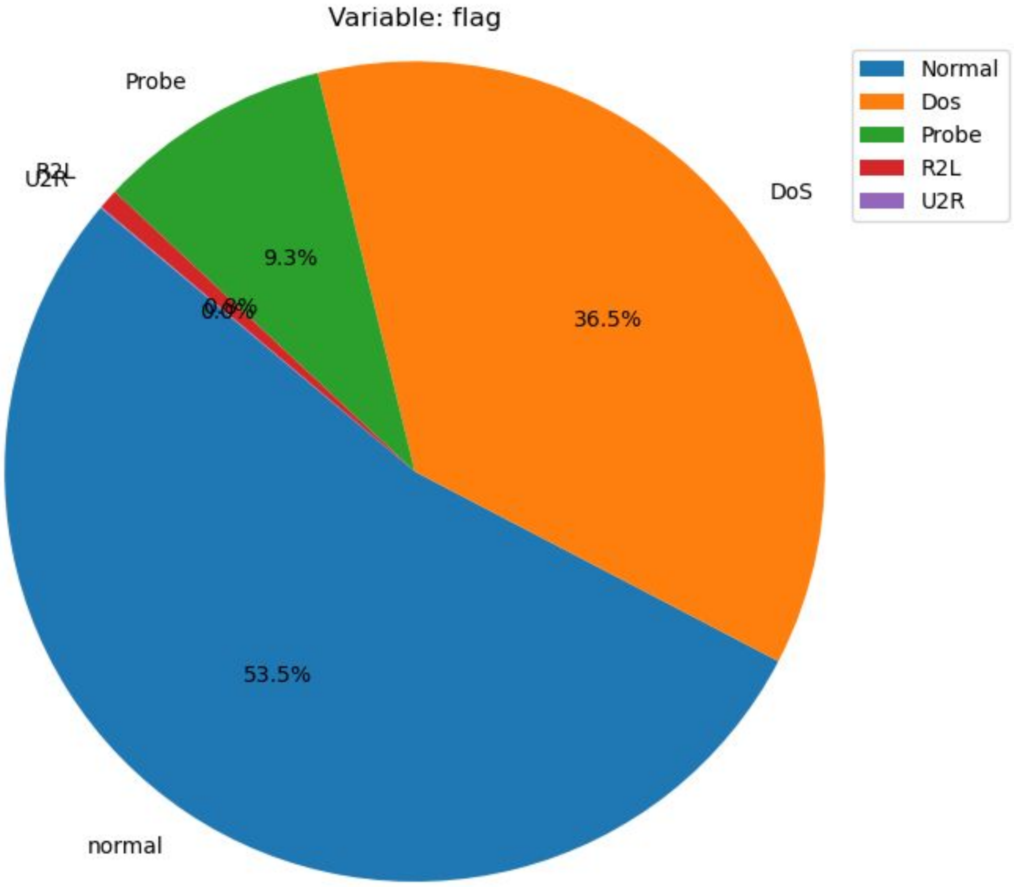






Our dataset contains only three categorical variables: Protocol, Service, and Flag.

Data Understanding: Dataset Balance



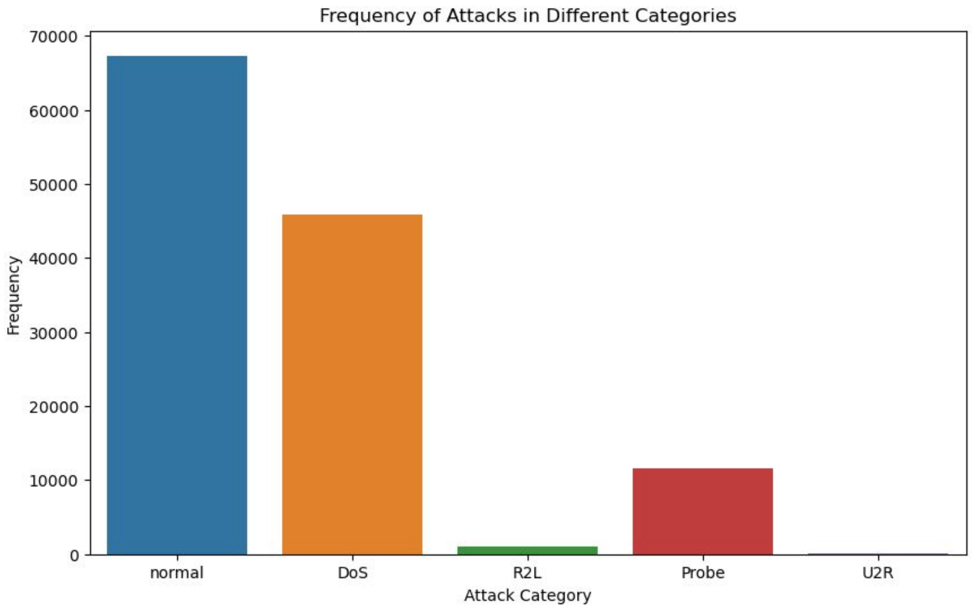
- **Dataset imbalance regarding attack categories.**
- **Very few data points for R2L and U2R attacks.**

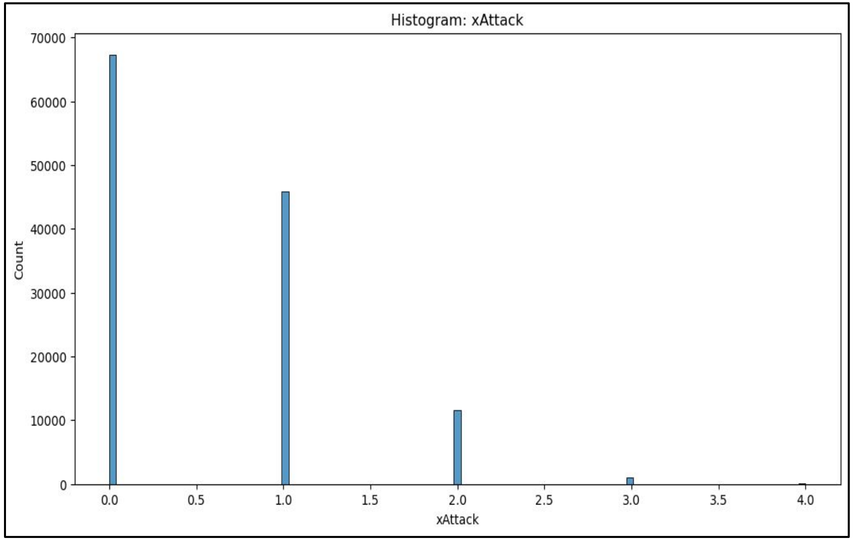
3

Data Preparation

Data preparation : Encoding

Encoding variable Attack



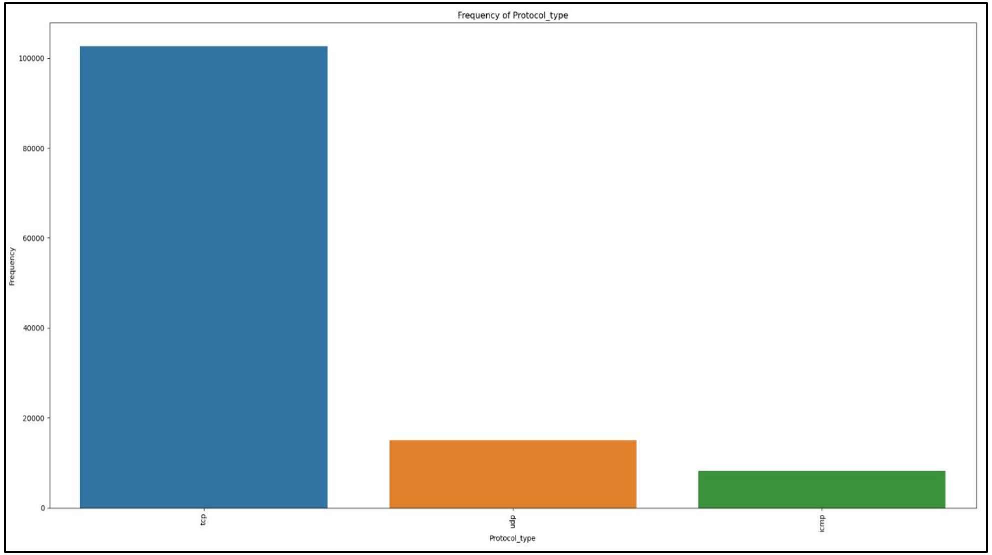


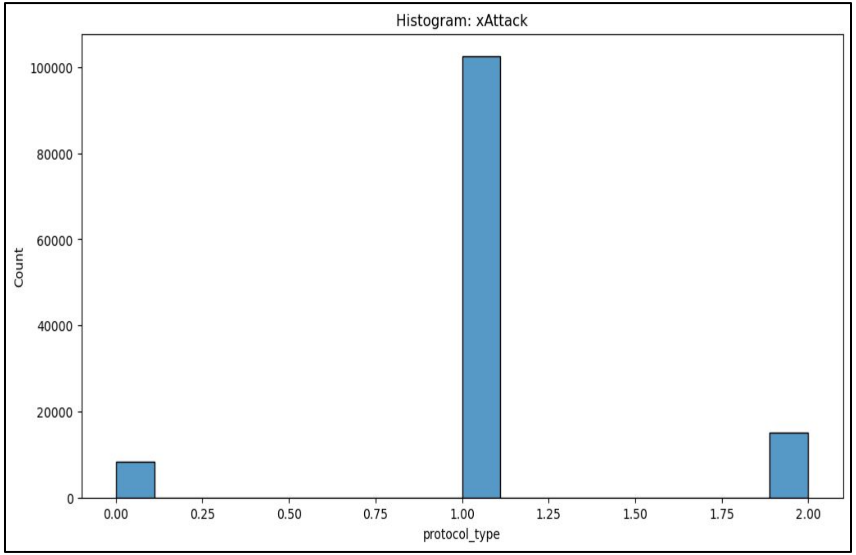
Before Encoding

After Encoding

Data preparation : Encodage

Encodage Protocol_type



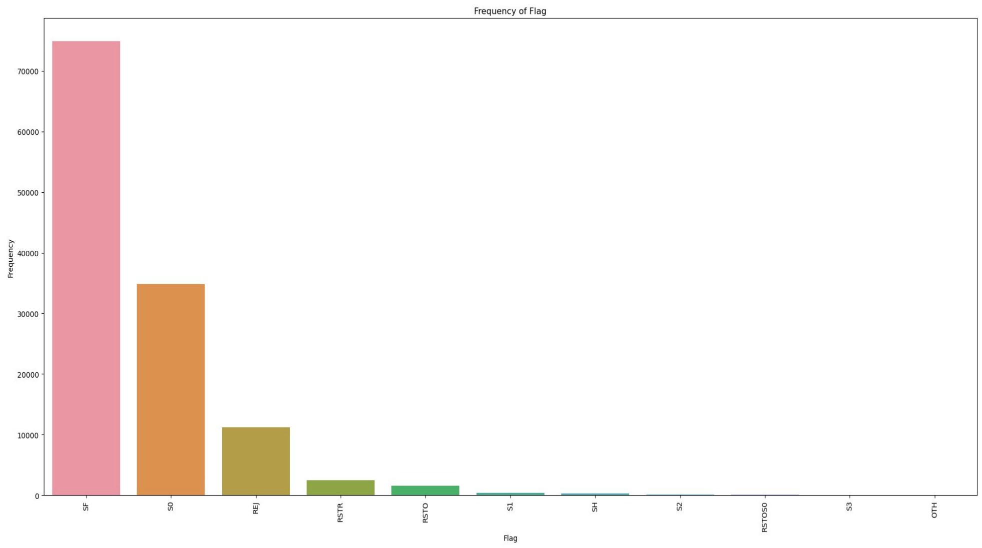


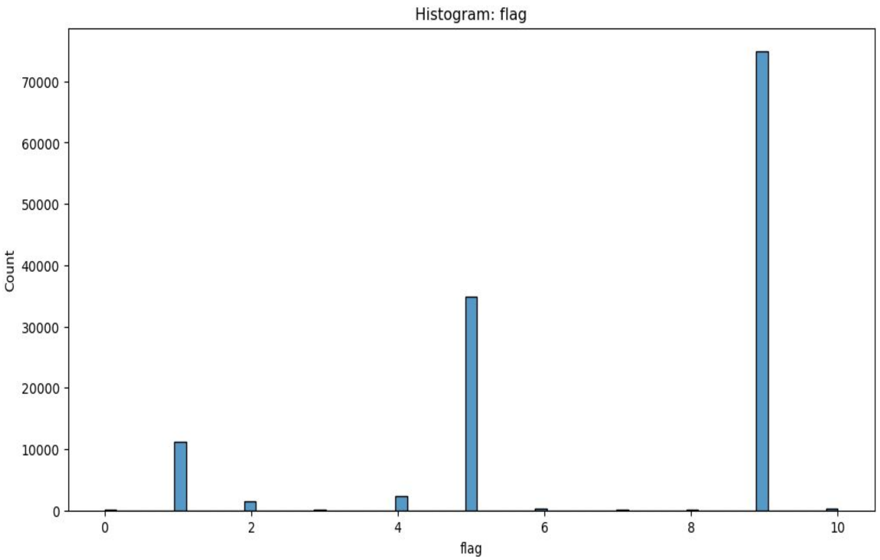
Before Encoding

After Encoding

Data preparation : Encodage

Encoding variable Flag



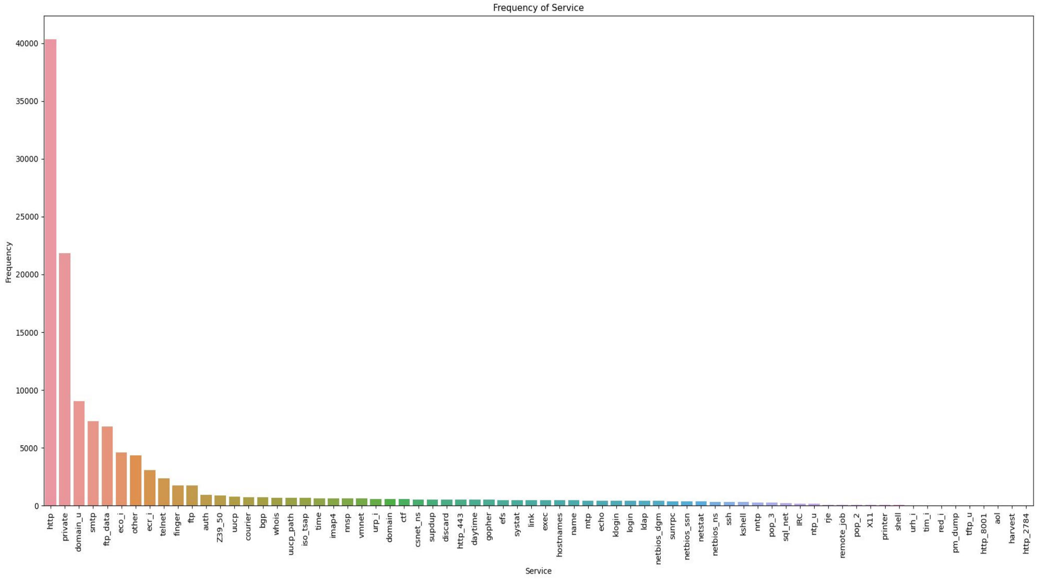


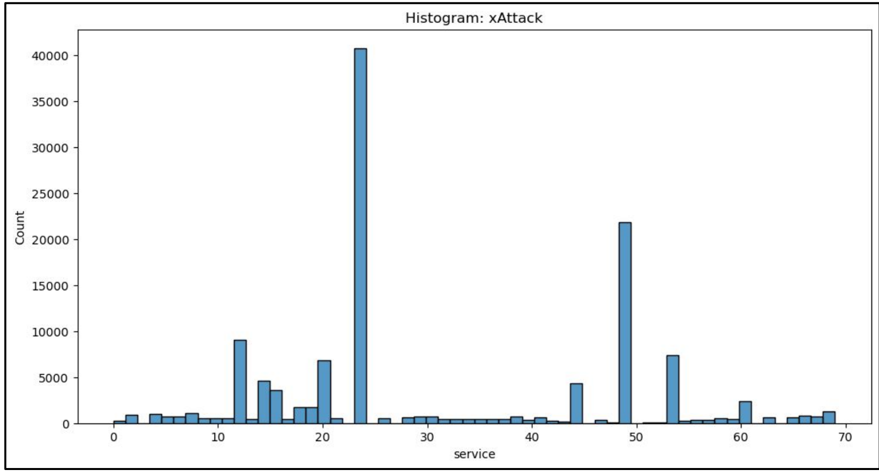
Before Encoding

After Encoding

Data preparation : Encodage

Encoding variable Service





Before Encoding

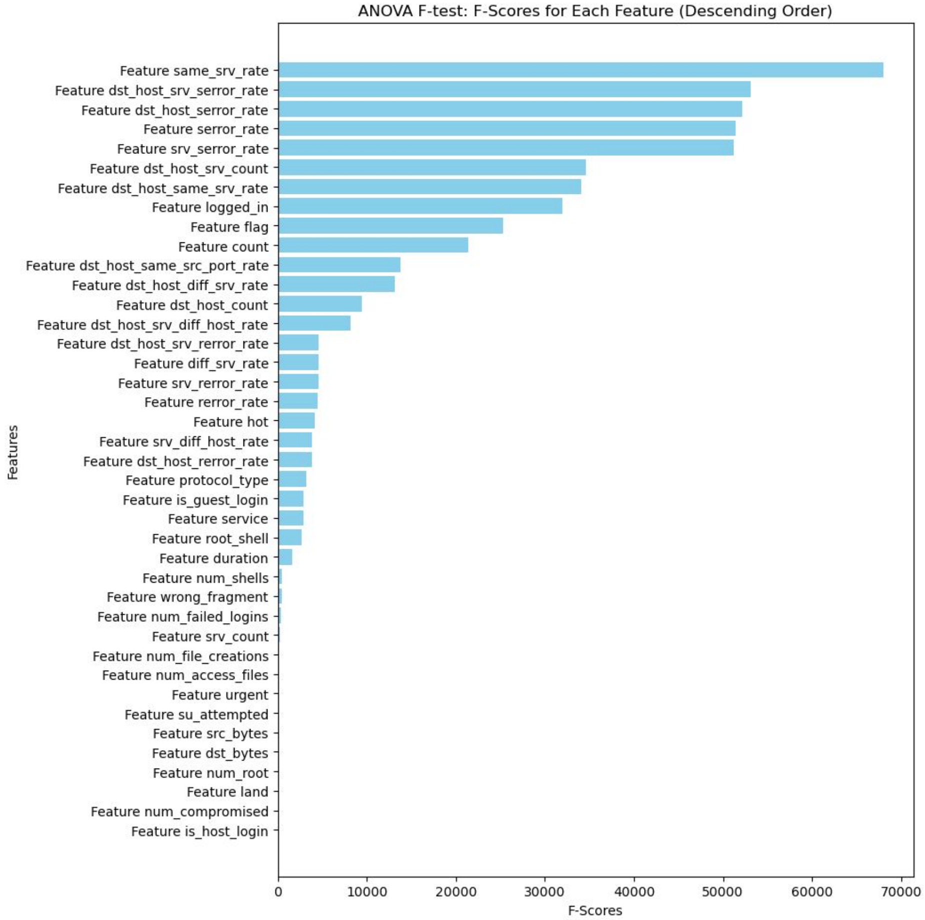
After Encoding

Data preparation : Standardization

	duration	protocol_type	service	flag	src_bytes	dst_bytes	land	wrong_fragment	urgent	hot	num_failed_logins	logged_in	num_compr
0	0.000000	0.5	0.289855	0.9	3.558064e-07	0.000000e+00	0.0	0.0	0.0	0.0	0.0	0.0	
1	0.000000	1.0	0.637681	0.9	1.057999e-07	0.000000e+00	0.0	0.0	0.0	0.0	0.0	0.0	
2	0.000000	0.5	0.710145	0.5	0.000000e+00	0.000000e+00	0.0	0.0	0.0	0.0	0.0	0.0	
3	0.000000	0.5	0.347826	0.9	1.681203e-07	6.223962e-06	0.0	0.0	0.0	0.0	0.0	1.0	
4	0.000000	0.5	0.347826	0.9	1.442067e-07	3.206260e-07	0.0	0.0	0.0	0.0	0.0	1.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
125968	0.000000	0.5	0.710145	0.5	0.000000e+00	0.000000e+00	0.0	0.0	0.0	0.0	0.0	0.0	
125969	0.000139	1.0	0.710145	0.9	7.608895e-08	1.106923e-07	0.0	0.0	0.0	0.0	0.0	0.0	
125970	0.000000	0.5	0.782609	0.9	1.616709e-06	2.931438e-07	0.0	0.0	0.0	0.0	0.0	1.0	
125971	0.000000	0.5	0.434783	0.5	0.000000e+00	0.000000e+00	0.0	0.0	0.0	0.0	0.0	0.0	
125972	0.000000	0.5	0.289855	0.9	1.094232e-07	0.000000e+00	0.0	0.0	0.0	0.0	0.0	1.0	

125973 rows × 43 columns

Data preparation : Feature selection



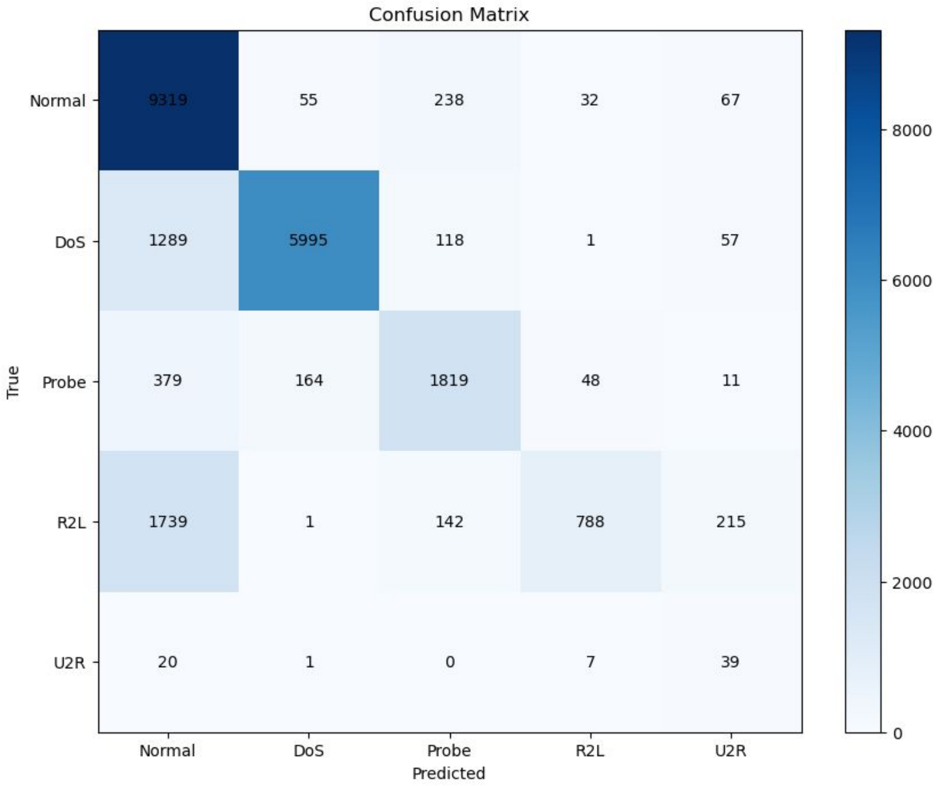
ANOVA F-TEST

According to the ANOVA test, we observed that 26 variables are significant, showing a high F-score

4

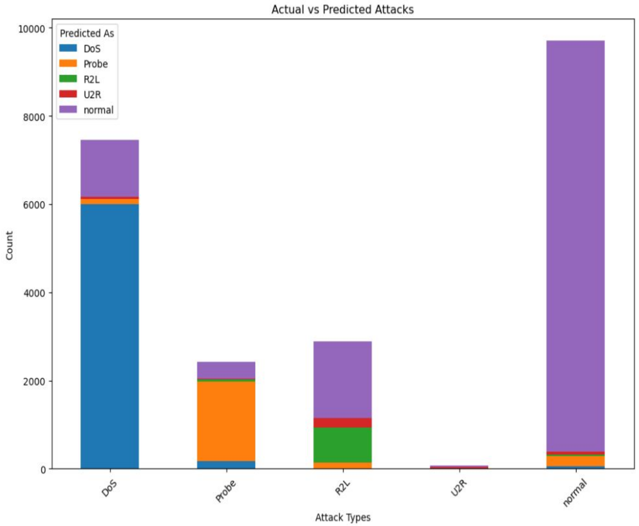
Modeling

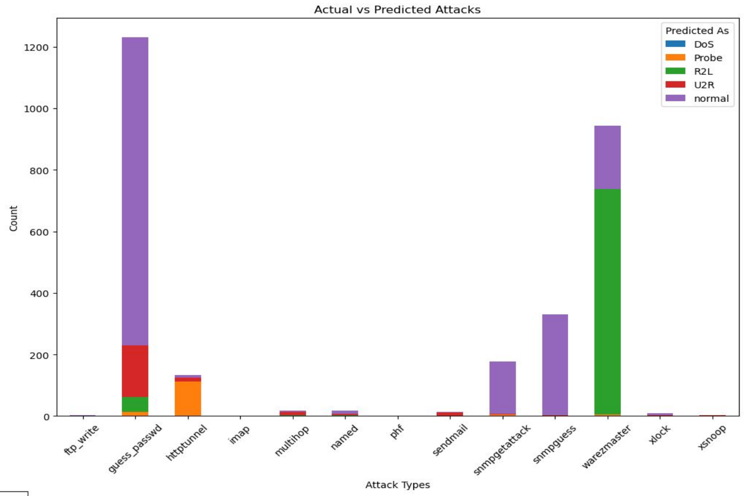
Modeling and Evaluation: KNN Before Split

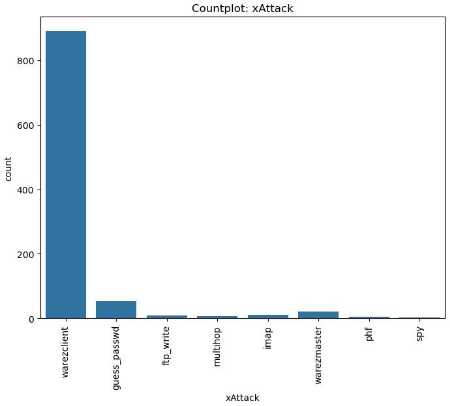


	precision	recall	f1-score	support
Normal	0.73	0.96	0.83	9711
DoS	0.96	0.80	0.88	7460
Probe	0.79	0.75	0.77	2421
R2L	0.90	0.27	0.42	2885
U2R	0.10	0.58	0.17	67
accuracy			0.80	22544
macro avg	0.70	0.67	0.61	22544
weighted avg	0.83	0.80	0.78	22544

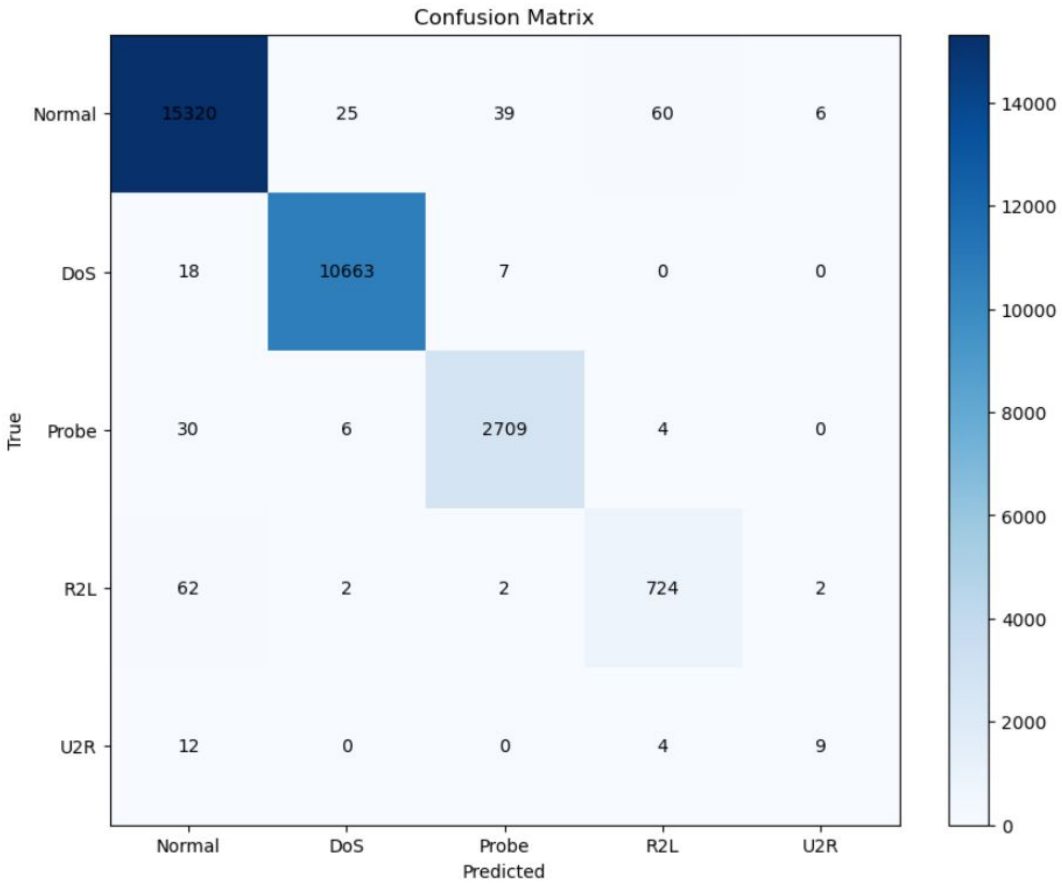
Modeling and Evaluation: KNN





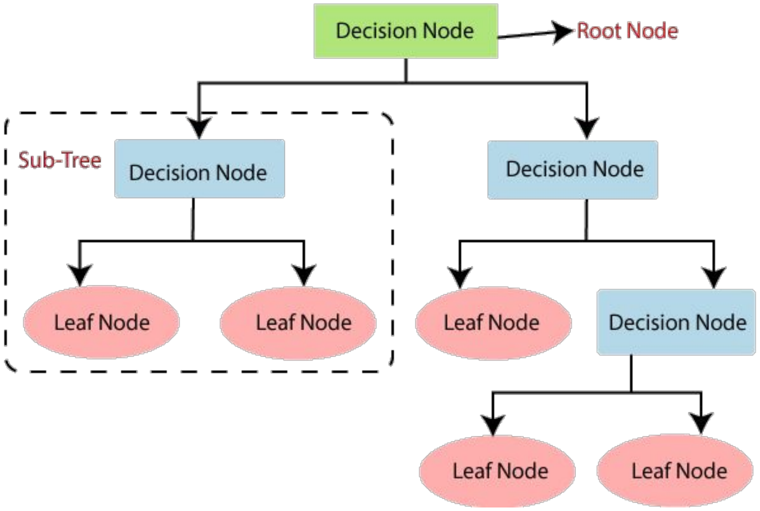


Modeling and Evaluation: KNN After split

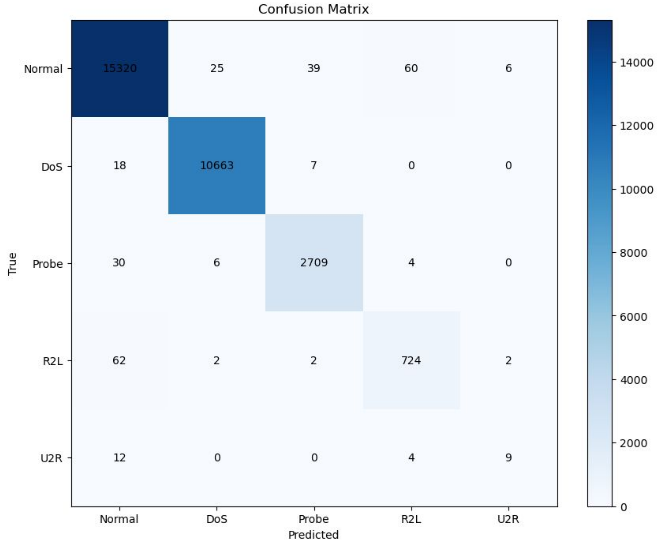


	precision	recall	f1-score	support
Normal	0.99	0.99	0.99	15450
DoS	1.00	1.00	1.00	10688
Probe	0.98	0.99	0.98	2749
R2L	0.91	0.91	0.91	792
U2R	0.53	0.36	0.43	25
accuracy			0.99	29704
macro avg	0.88	0.85	0.86	29704
weighted avg	0.99	0.99	0.99	29704

Modeling and Evaluation: DECISION TREE

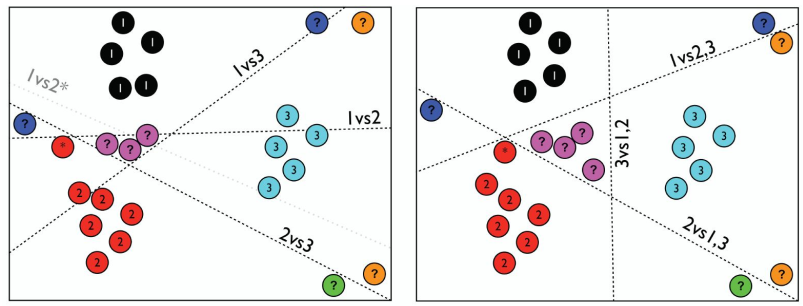


Criterion: entropy
max_depth: 19

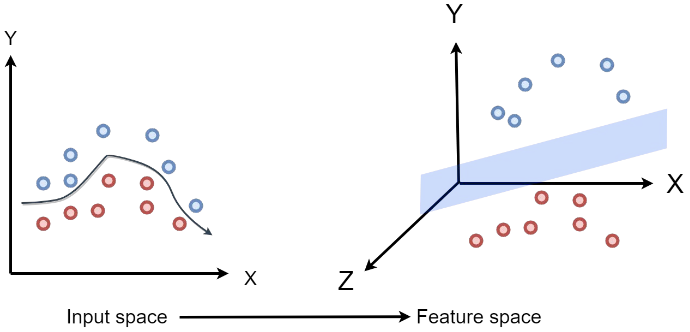


	precision	recall	f1-score	support
Normal	0.99	0.99	0.99	15450
DoS	1.00	1.00	1.00	10688
Probe	0.98	0.99	0.98	2749
R2L	0.91	0.91	0.91	792
U2R	0.53	0.36	0.43	25
accuracy			0.99	29704
macro avg	0.88	0.85	0.86	29704
weighted avg	0.99	0.99	0.99	29704

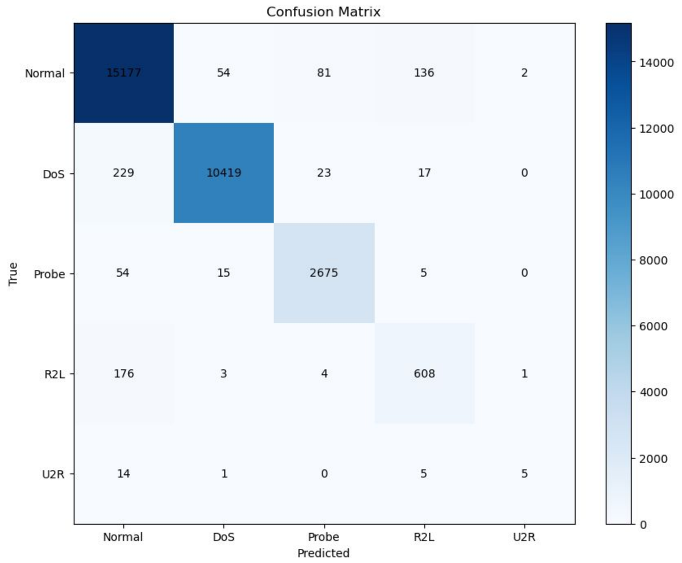
Modeling and Evaluation: SVM



(a) 1-vs-1 (b) 1-vs-All

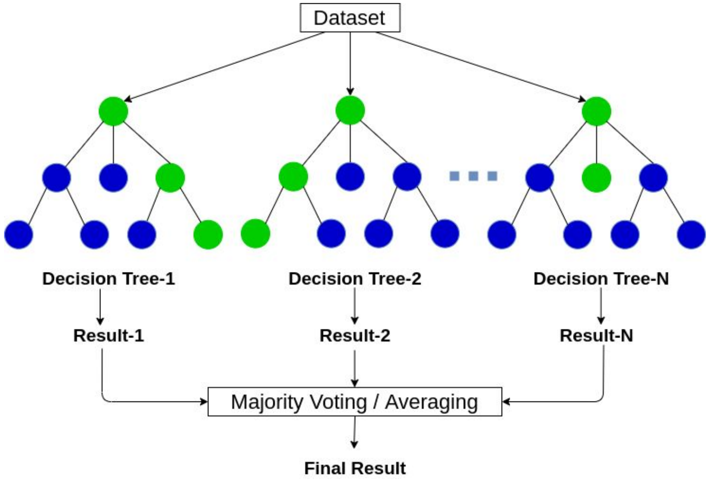


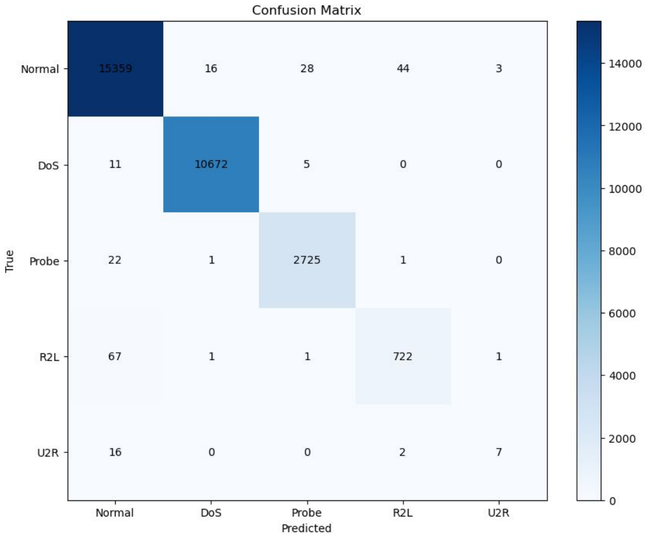
kernel Poly



	precision	recall	f1-score	support
Normal	0.97	0.98	0.98	15450
DoS	0.99	0.97	0.98	10688
Probe	0.96	0.97	0.97	2749
R2L	0.79	0.77	0.78	792
U2R	0.62	0.20	0.30	25
accuracy			0.97	29704
macro avg	0.87	0.78	0.80	29704
weighted avg	0.97	0.97	0.97	29704

Modeling and Evaluation: RANDOM FOREST

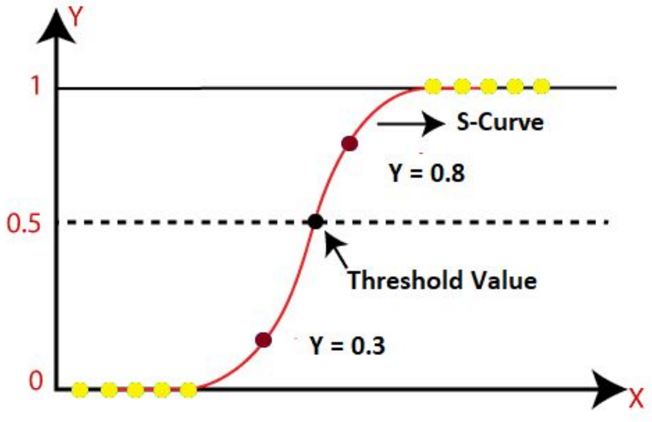




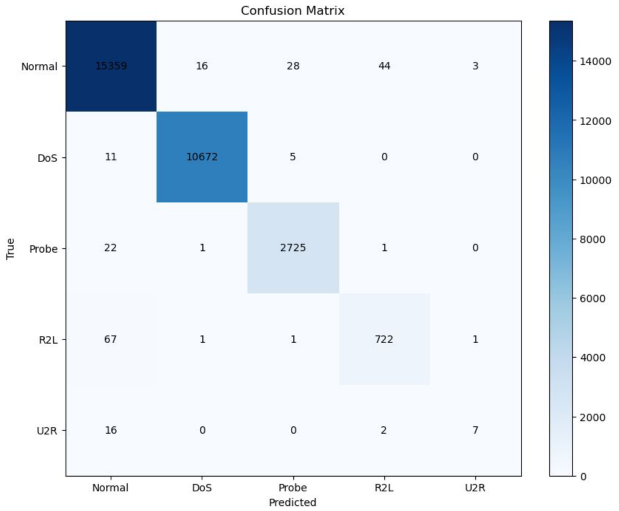
max_depth=30
min_samples_split=2
n_estimators = 1600

	precision	recall	f1-score	support
Normal	0.99	0.99	0.99	15450
DoS	1.00	1.00	1.00	10688
Probe	0.99	0.99	0.99	2749
R2L	0.94	0.91	0.93	792
U2R	0.64	0.28	0.39	25
accuracy			0.99	29704
macro avg	0.91	0.84	0.86	29704
weighted avg	0.99	0.99	0.99	29704

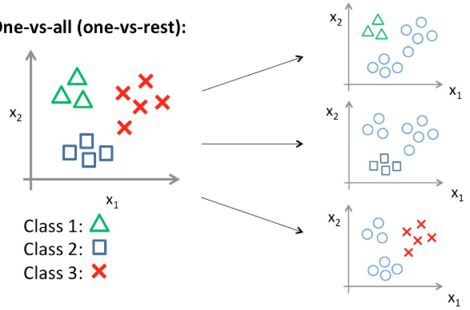
Modeling and Evaluation: LOGISTIC REGRESSION



Logistic regression



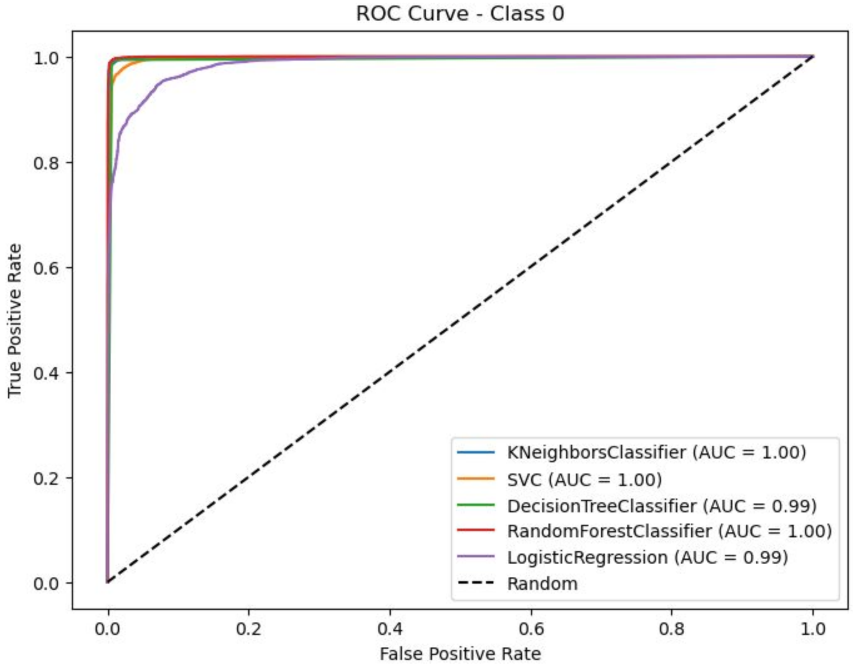
One-vs-all (one-vs-rest):

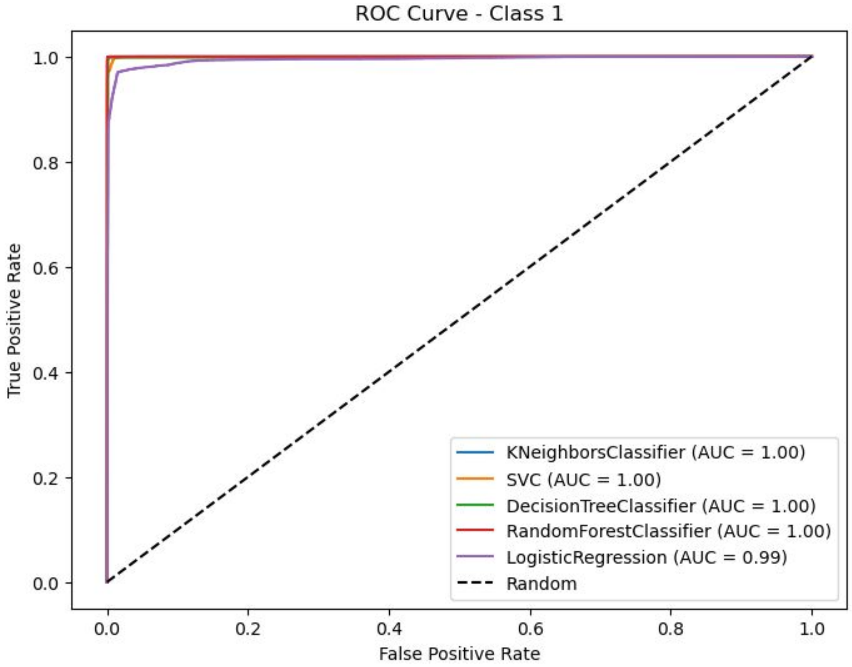


penalty="l2" ridge
max_iter=767

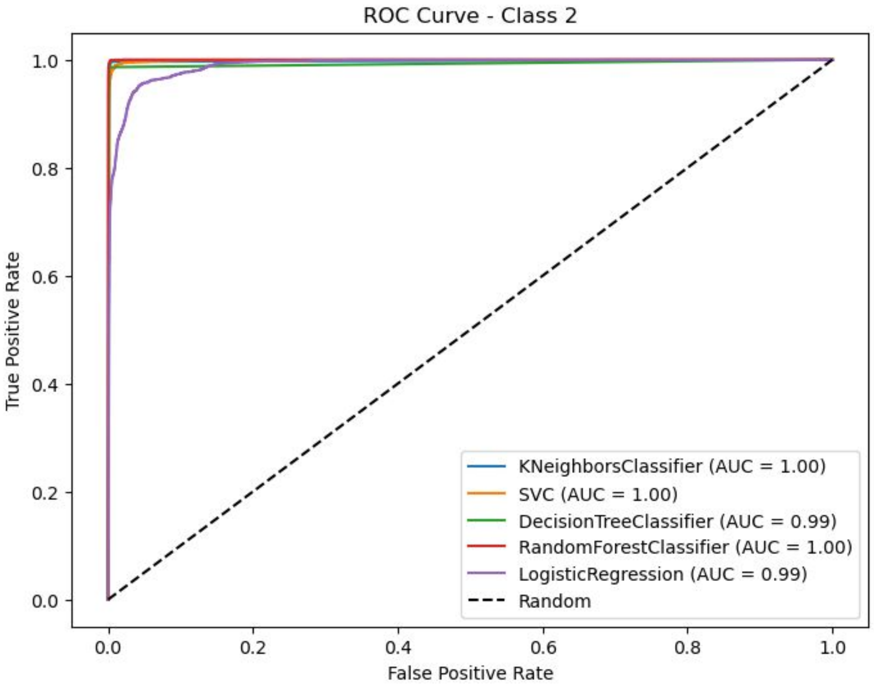
	precision	recall	f1-score	support
Normal	0.93	0.95	0.94	15450
DoS	0.97	0.97	0.97	10688
Probe	0.85	0.86	0.86	2749
R2L	0.55	0.25	0.35	792
U2R	0.77	0.40	0.53	25
accuracy			0.93	29704
macro avg	0.81	0.69	0.73	29704
weighted avg	0.93	0.93	0.93	29704

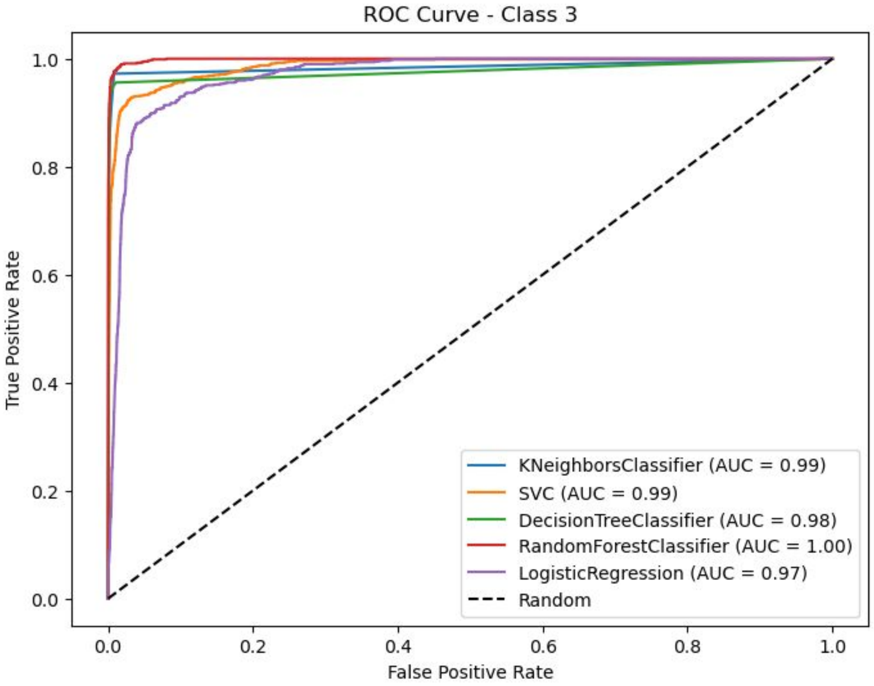
Modeling: Comparison



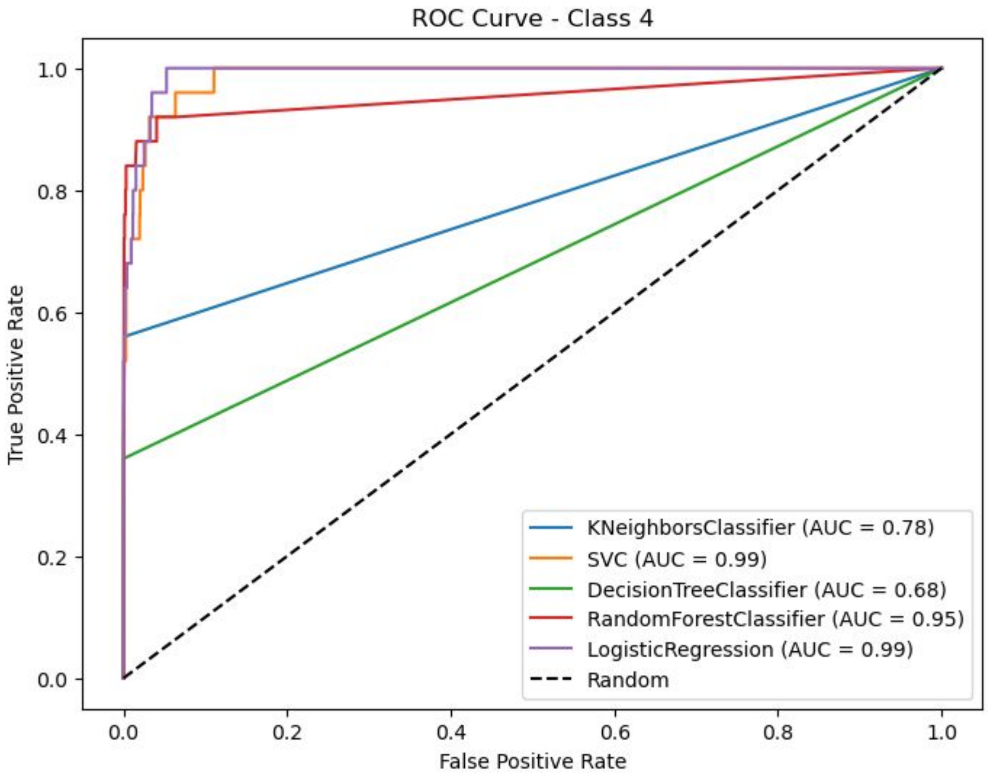


Modeling: Comparison



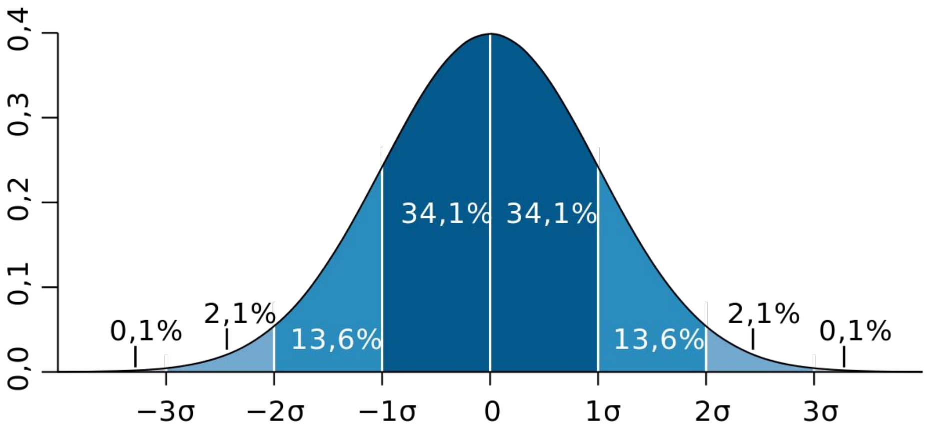


Modeling: Comparison



- **All models have very satisfactory performances**
- **Random Forest and Decision Tree yield the best results**
- **We choose Random Forest because it is more robust as it circumvents the issue of overfitting.**

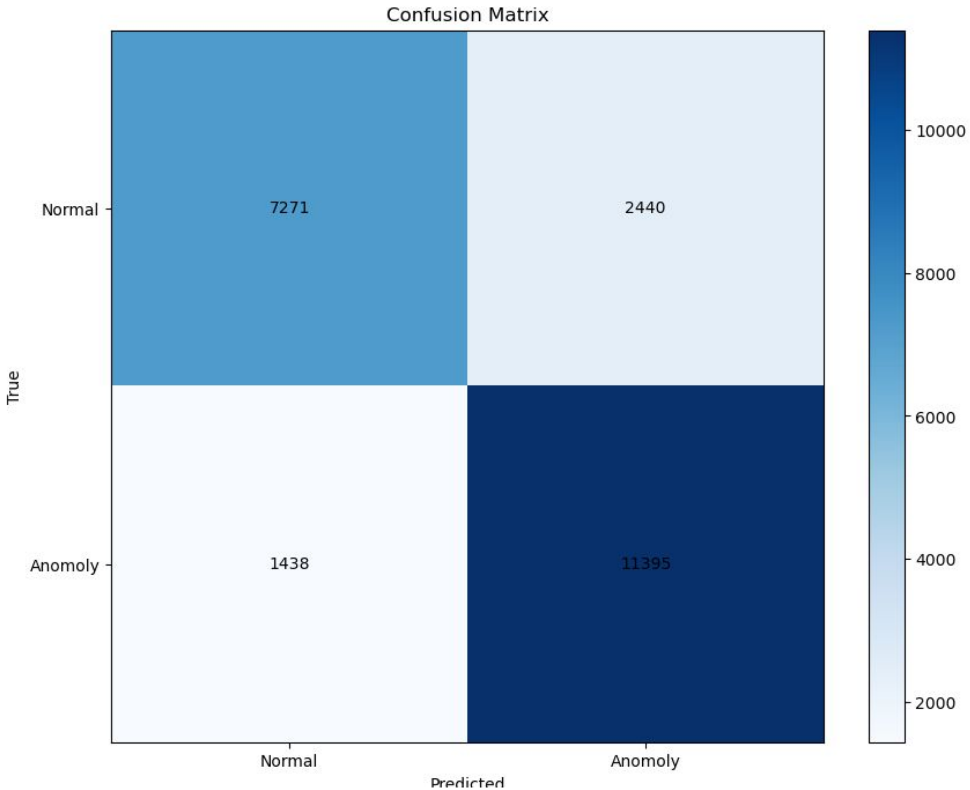
Modeling: Voting (Paper2)



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	
0	0.132780	0.440493	0.632142	0.577769	0.049004	0.831104	0.949828	0.258314	0.230383	0.648312	0.752327	0.704018	0.580687	0.468587	**Normal**
1	0.005808	0.124003	0.001635	0.969355	0.329880	0.833078	0.683989	0.021335	0.263994	0.005096	0.000364	0.584208	0.018393	0.004341	**Anomalie**

Seuil : 5% Consensus: 3

Modeling: Voting (Paper2)

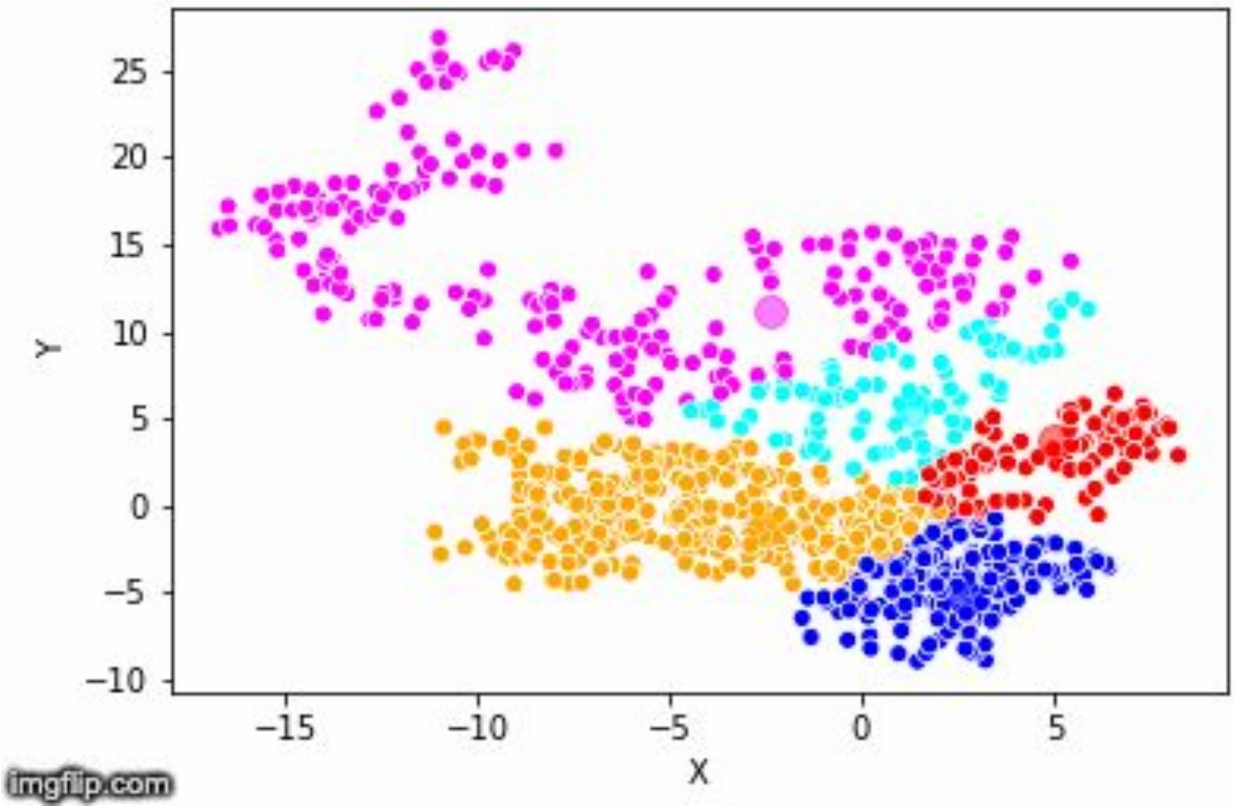


	precision	recall	f1-score	support
Normal	0.83	0.75	0.79	9711
Anomoly	0.82	0.89	0.85	12833
accuracy			0.83	22544
macro avg	0.83	0.82	0.82	22544
weighted avg	0.83	0.83	0.83	22544

F1_Score = 0.8545822708864557

d_norm_probs

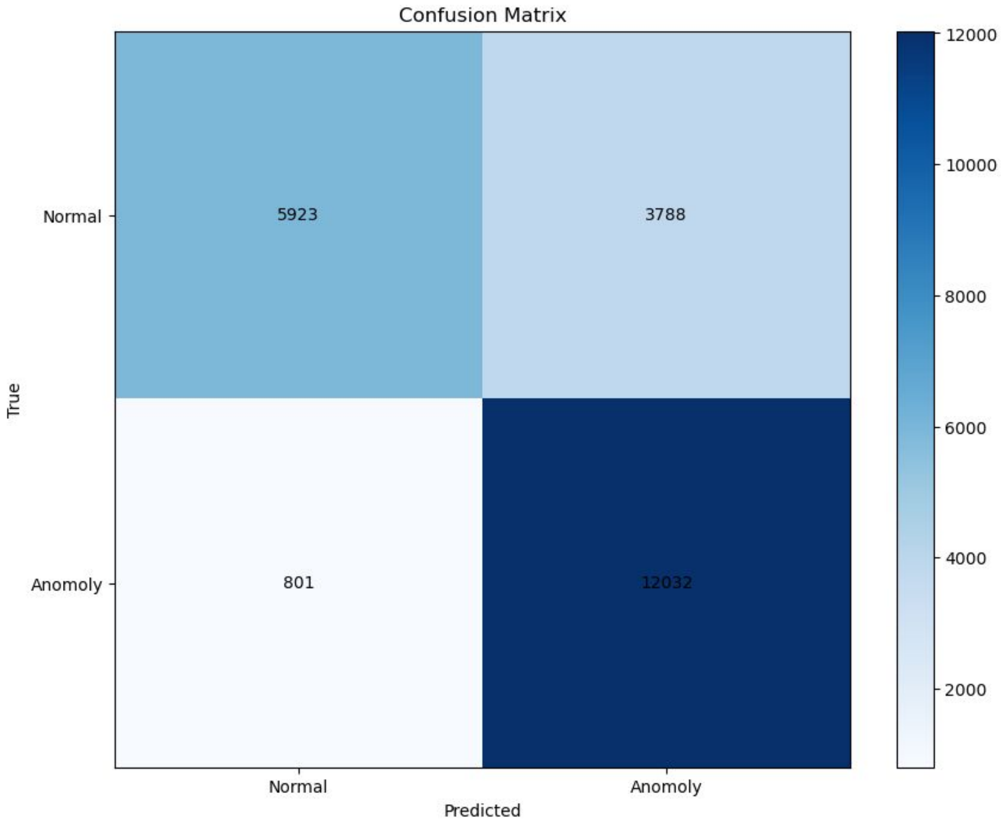
Modeling: K means anomaly detection (Paper2)



Clustering
Only with normal Observation

Prediction
We calculate the distance of the observation to the centroids

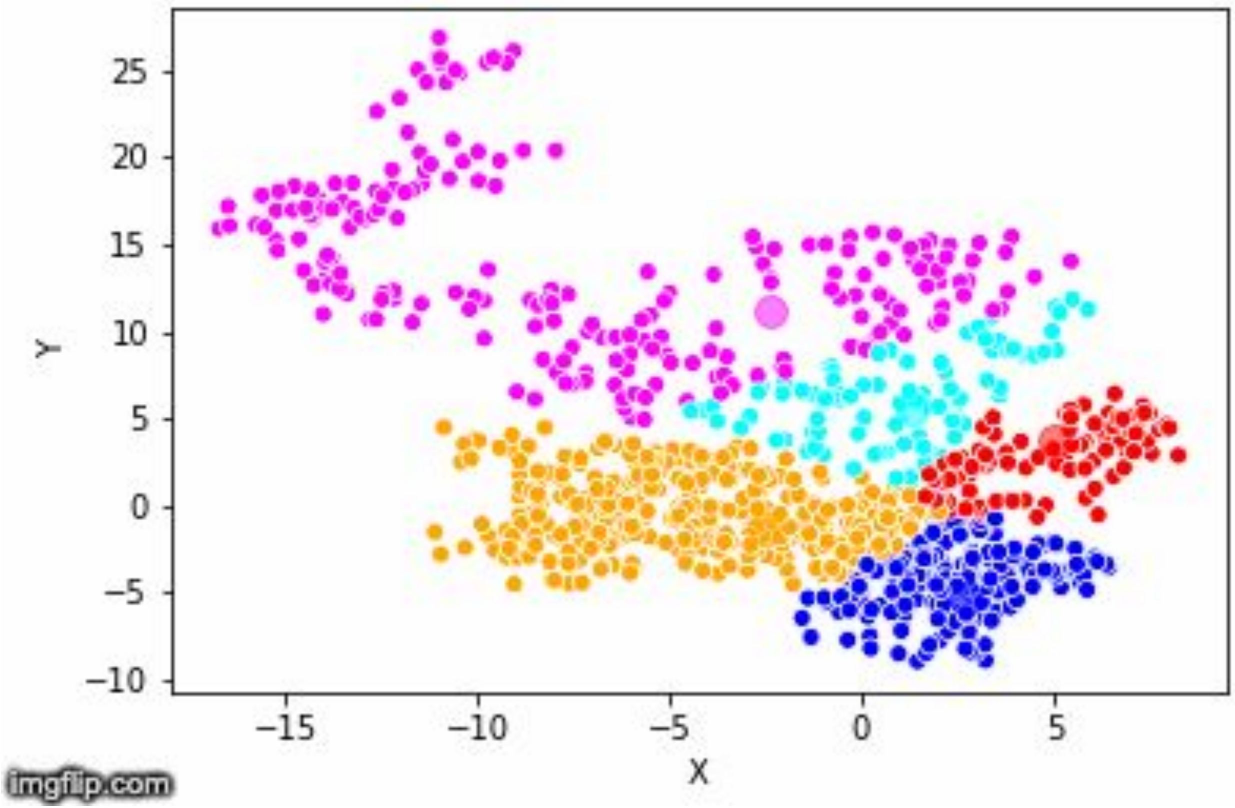
Modeling: K means anomaly detection (Paper2)



	precision	recall	f1-score	support
Normal	0.88	0.61	0.72	9711
Anomoly	0.76	0.94	0.84	12833
accuracy			0.80	22544
macro avg	0.82	0.77	0.78	22544
weighted avg	0.81	0.80	0.79	22544

F1_Score = 0.8398422503751789

Modeling: K means clusters (Paper2)

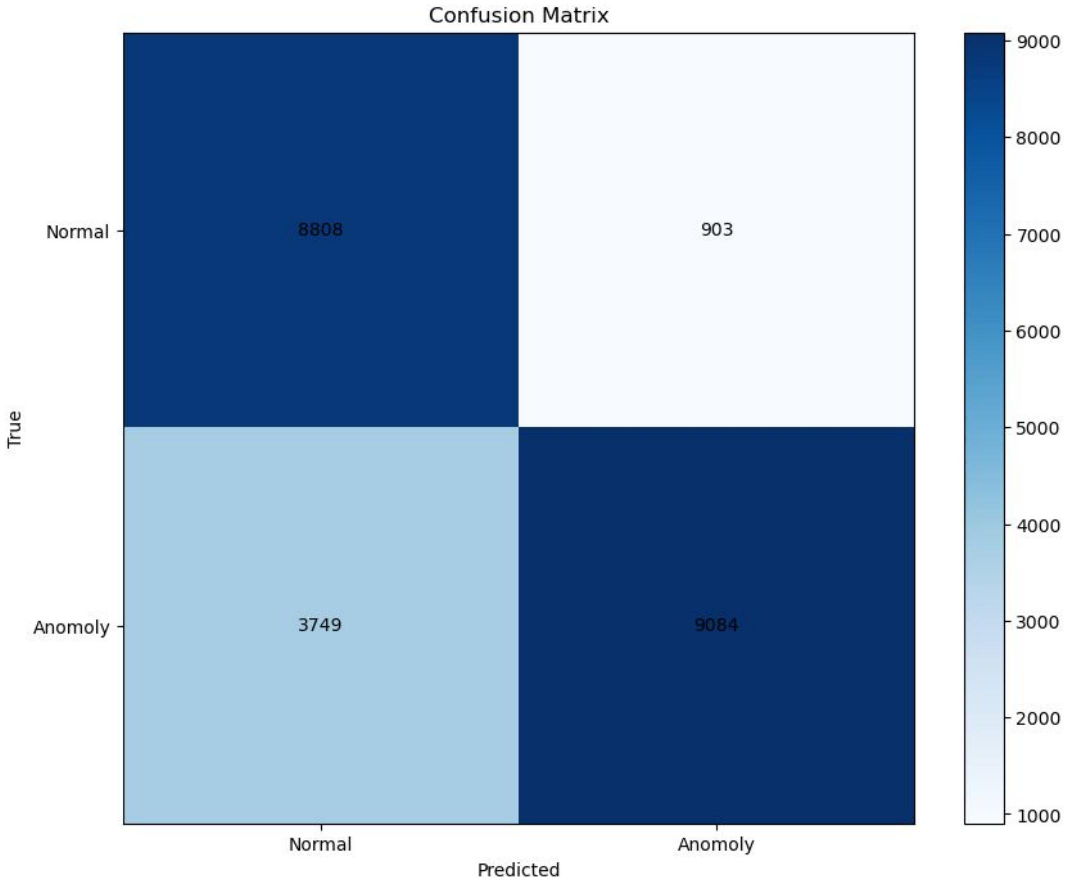


Clustering
With All Obeservations

Classification
We calculate the dominant class of each cluster:
0: Normal
1: Anomaly

Prediction
K Means determines the cluster to which the observation belongs, and afterward, we determine its class

Modeling: K means clusters (Paper2)

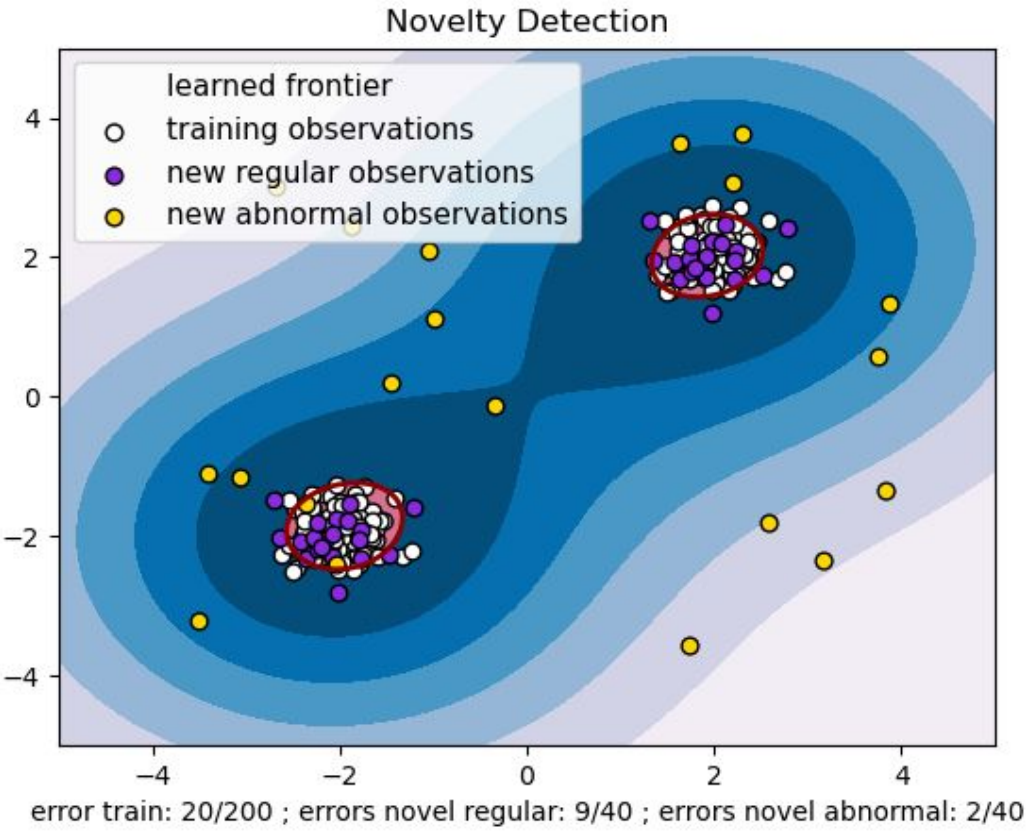


	precision	recall	f1-score	support
Normal	0.70	0.91	0.79	9711
Anomoly	0.91	0.71	0.80	12833
accuracy			0.79	22544
macro avg	0.81	0.81	0.79	22544
weighted avg	0.82	0.79	0.79	22544

F1_Score = 0.7961437335670465

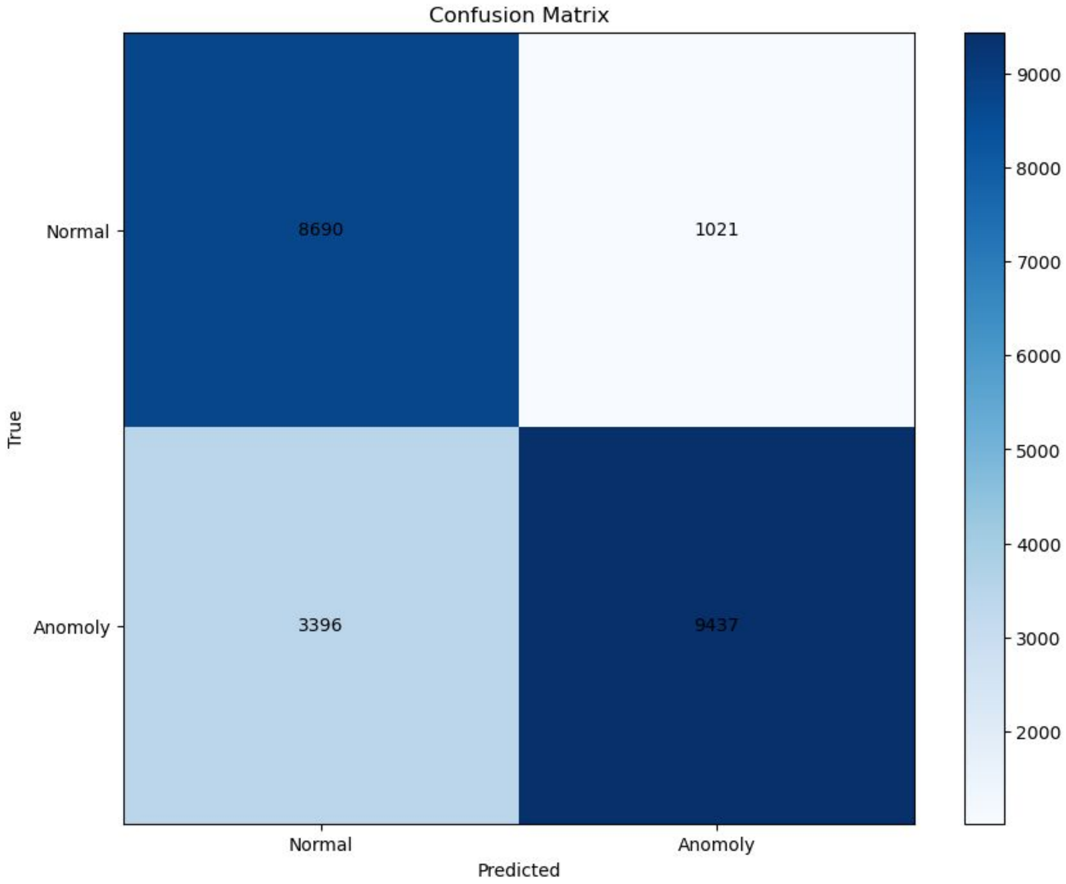
d_raw_probs

Modeling: SVM (Paper2)



OneClassSVM

Modeling: SVM (Paper2)



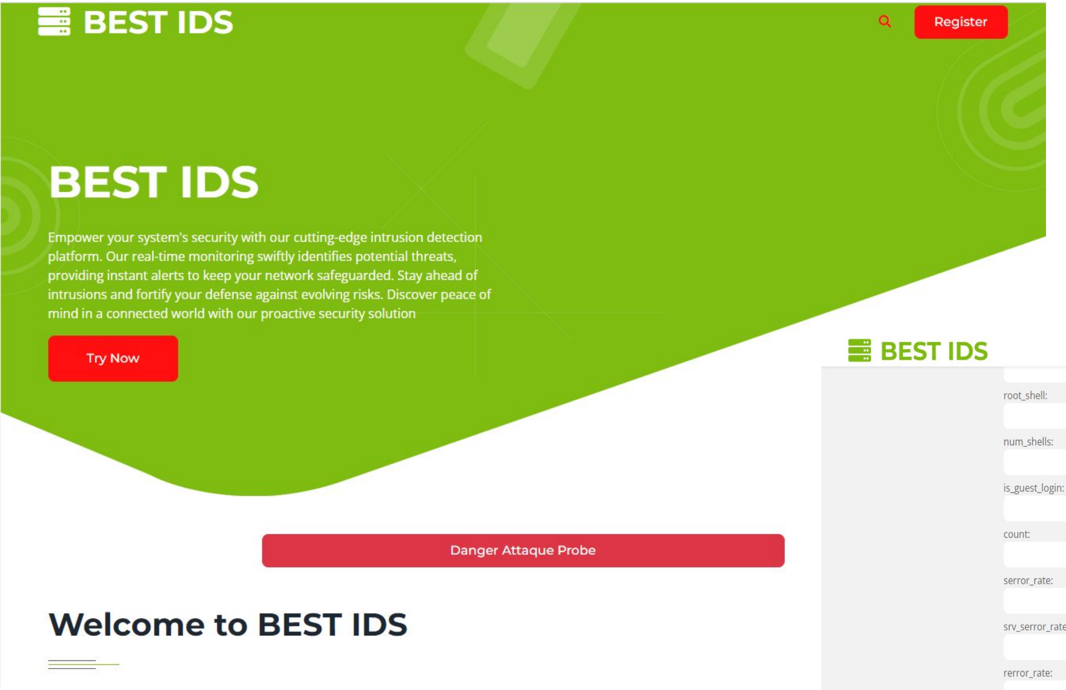
	precision	recall	f1-score	support
Normal	0.72	0.89	0.80	9711
Anomoly	0.90	0.74	0.81	12833
accuracy			0.80	22544
macro avg	0.81	0.82	0.80	22544
weighted avg	0.82	0.80	0.80	22544

F1_Score = 0.8103559314756774

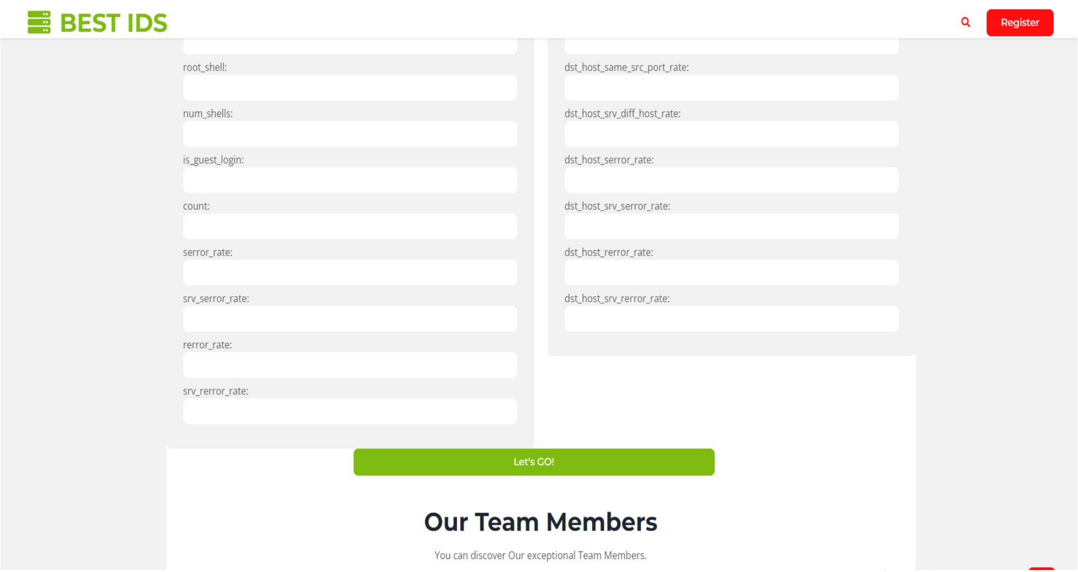
d_norm_pca

5 Deployment

Deployment : Django



The deployment using Random Forest that yielded the best modeling score.



THANK YOU