

Cisco – Ariel University API Security Detection Challenge 2023

Moriya Bitton | | Victor Kushnir

[GitHub Link](#)



Original Features

We have 6 datasets, each with the same original features.

#	Column	Non-Null	Count	Dtype
0	request.headers.Host	4282	non-null	object
1	request.headers.User-Agent	4282	non-null	object
2	request.headers.Accept-Encoding	4282	non-null	object
3	request.headers.Accept	4282	non-null	object
4	request.headers.Connection	4282	non-null	object
5	request.headers.Accept-Language	4282	non-null	object
6	request.headers.Sec-Fetch-Site	4282	non-null	object
7	request.headers.Sec-Fetch-Mode	4282	non-null	object
8	request.headers.Sec-Fetch-User	4282	non-null	object
9	request.headers.Sec-Fetch-Dest	4282	non-null	object
10	request.headers.Set-Cookie	4282	non-null	object
11	request.headers.Date	4282	non-null	object
12	request.method	4282	non-null	object
13	request.url	4282	non-null	object
14	request.body	4282	non-null	object
15	response.status	4282	non-null	object
16	response.headers.Content-Type	4282	non-null	object
17	response.headers.Content-Length	4282	non-null	object
18	response.status_code	4282	non-null	int64
19	response.body	4282	non-null	object
20	request.headers.Cookie	566	non-null	object
21	response.headers.Location	401	non-null	object
22	request.headers.Content-Length	299	non-null	object
23	response.headers.Set-Cookie	299	non-null	object
24	attack_type	4282	non-null	object
25	label	4282	non-null	object

Preprocessing Data

In each dataset, we repeat the same preprocessing for our specific dataset:

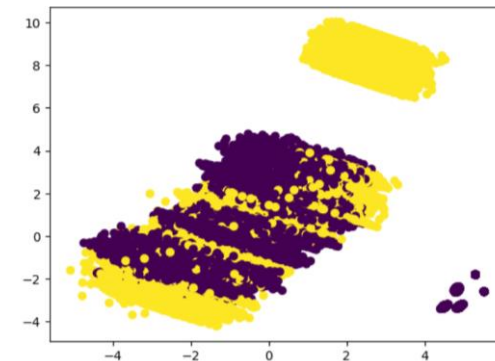
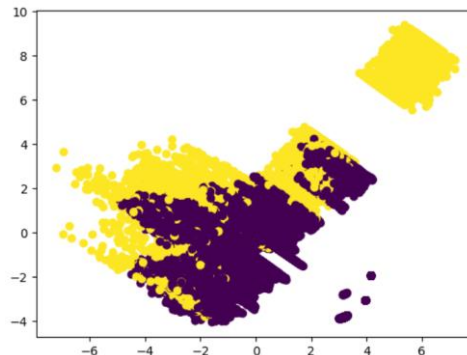
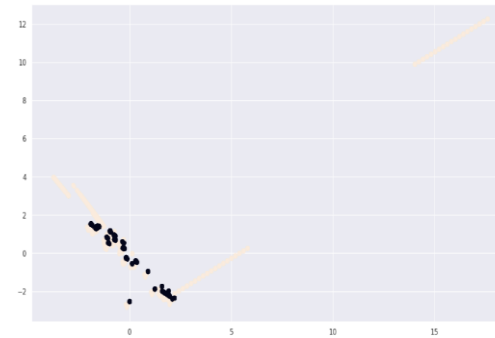
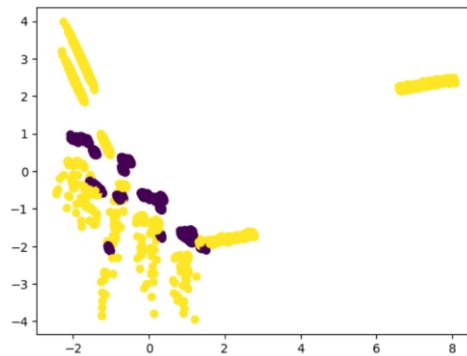
- Replace all Nan values with the string 'Null'
- Remove columns that have the same values for all rows
- Create new features from the request.url and request.headers
- Preprocess using only LabelEncoder()

```
COLUMNS_TO_REMOVE = [
    'request.body',
    'response.headers.Content-Length',
    'request.headers.Date',
    'request.headers.Accept',
    'request.headers.Connection',
    'request.headers.Sec-Fetch-User',
]
```



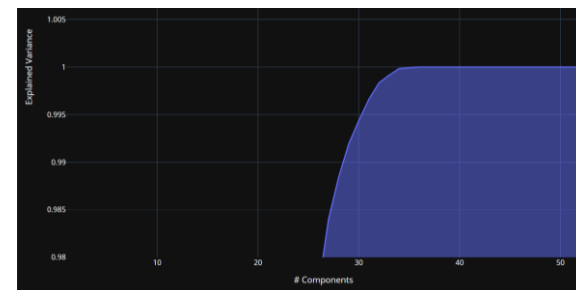
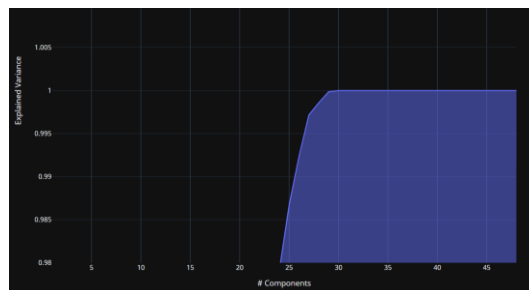
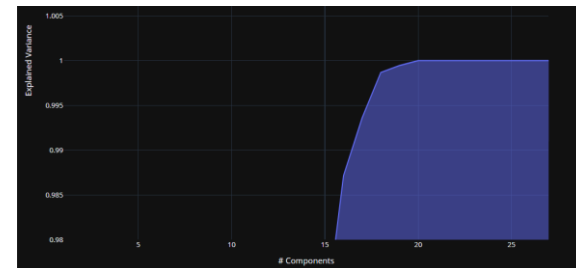
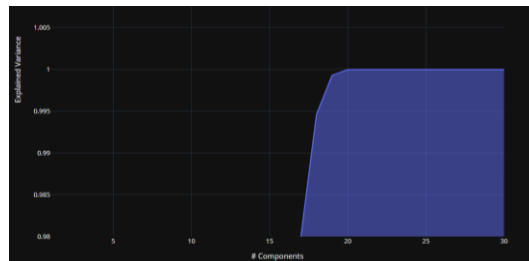
PCA in 2D

Compressing the data into two components allows us to analyze its distribution.



PCA - Ratio

Using this ratio, we can find out how much information we lose compared to how many features we have.



Standard Scaler

Standard Scalers are pre-processing techniques in machine learning and data analysis that standardize constants in datasets to have a mean of 0 and a standard deviation of 1.

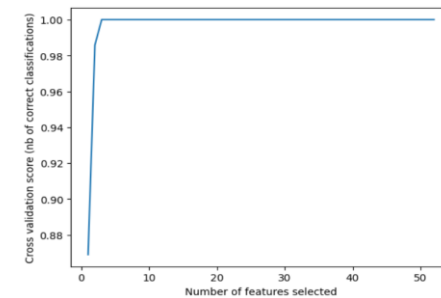
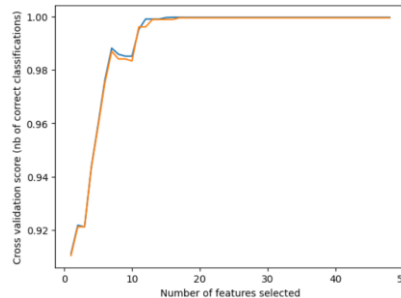
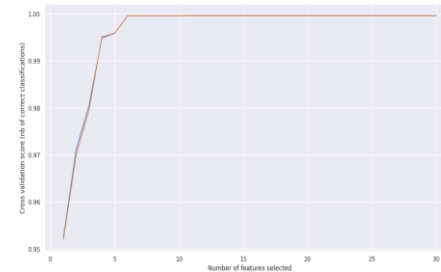
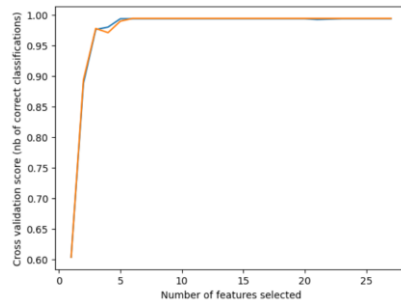
Outliers in the data can also be reduced through standardization.



Feature Selection

Using RFECV we found the optimal number of features.

Now, a grid search object finds the best hyperparameters for the model.



Random Forest Classifier

The RandomForestClassifier is an ensemble learning algorithm that uses multiple decision trees to predict the future.

Except for Task_4_Attack, which was 97% accurate, our model was 100% accurate across all datasets.

We then repeat our preprocessing steps for the test data, just as we did for the training data.



The END

Moriya Bitton | | Victor Kushnir

[GitHub Link](#)

