Network attack detection

Machine Learning course @ SUPSI

Luca Di Bello <luca.dibello@student.supsi.ch> - 2022/2023

Project goal:

Detect network attacks using machine learning techniques

Dataset

Data format: NetFlow Version 9 Flow-Record (by Cisco)

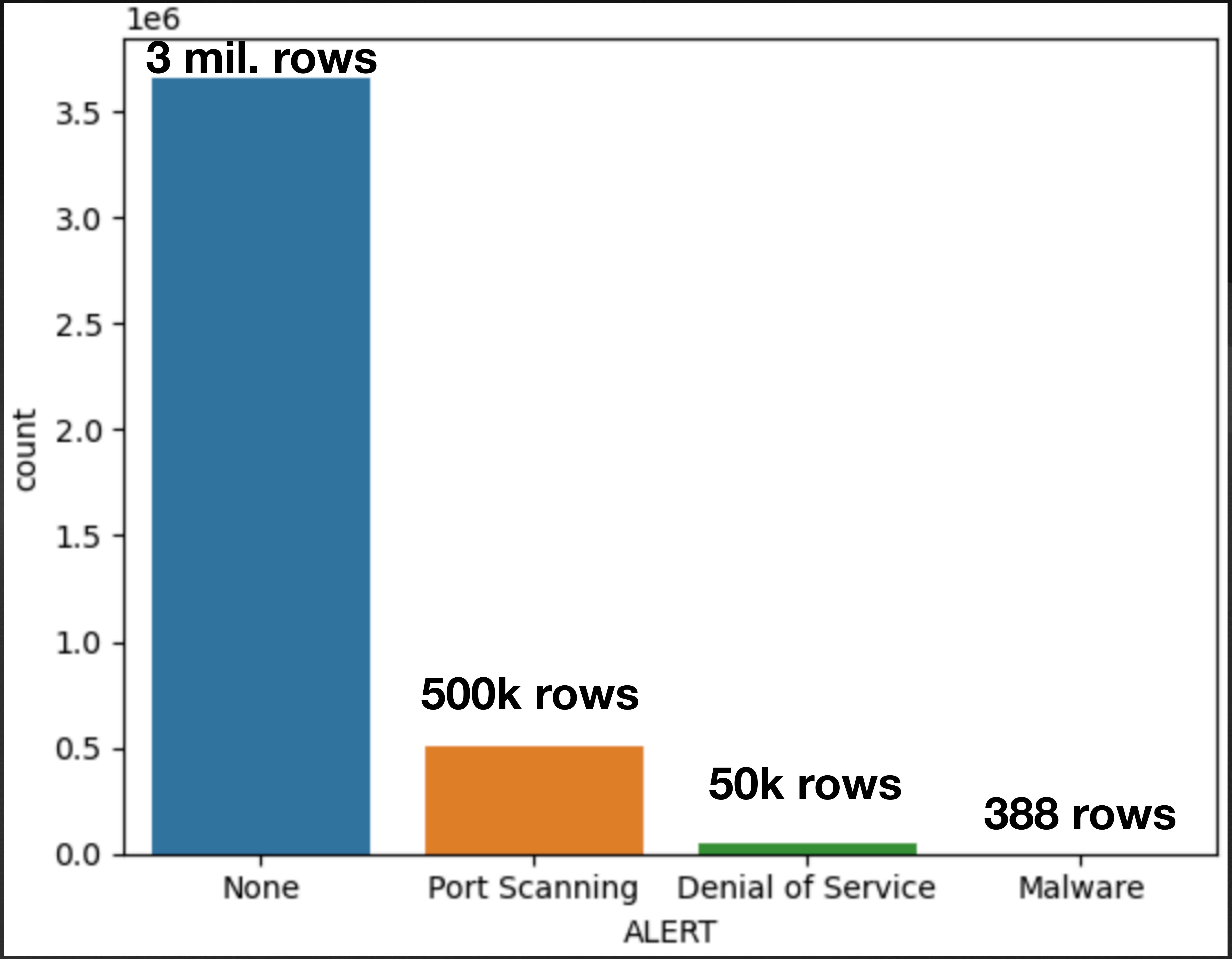
- **Two files** (total size = 1.01 GB)
  - **Train set:** ~4 mil. packets 14’066 unique hosts
  - **Test set:** ~2 mil. packets 6187 unique hosts

- **32 features**:
  - Protocol, Source/Dest. Ports, Packet size, etc. (View Cisco documentation)
  - **ALERT** (target) = DoS, Port Scanning, Malware, None

Dataset problems

- Known problems:
  - **Highly imbalanced** dataset
  - Dataset **too big**
  - Useless features
  - **Only classification**

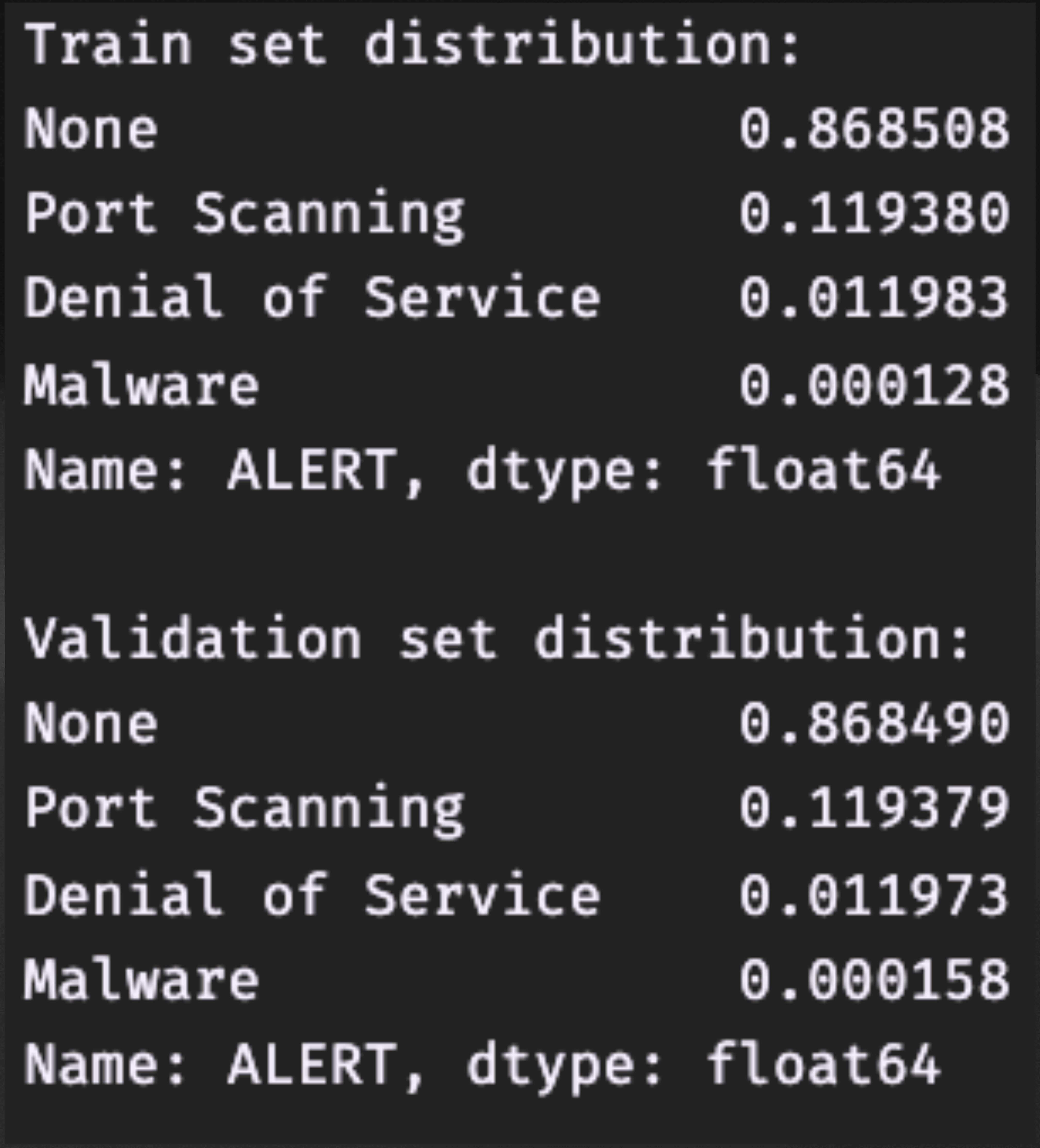
- Solutions:
  - Dataset sampling
  - `StatifiedShuffleSplit`



Train & Validation sets

(with `StratifiedShuffleSplit`)

- Dataset sampling
  - Train set = ~ 120k rows
  - Validation set = ~ 60k rows
    - (dev = 3% of the total size)
- Same target distribution



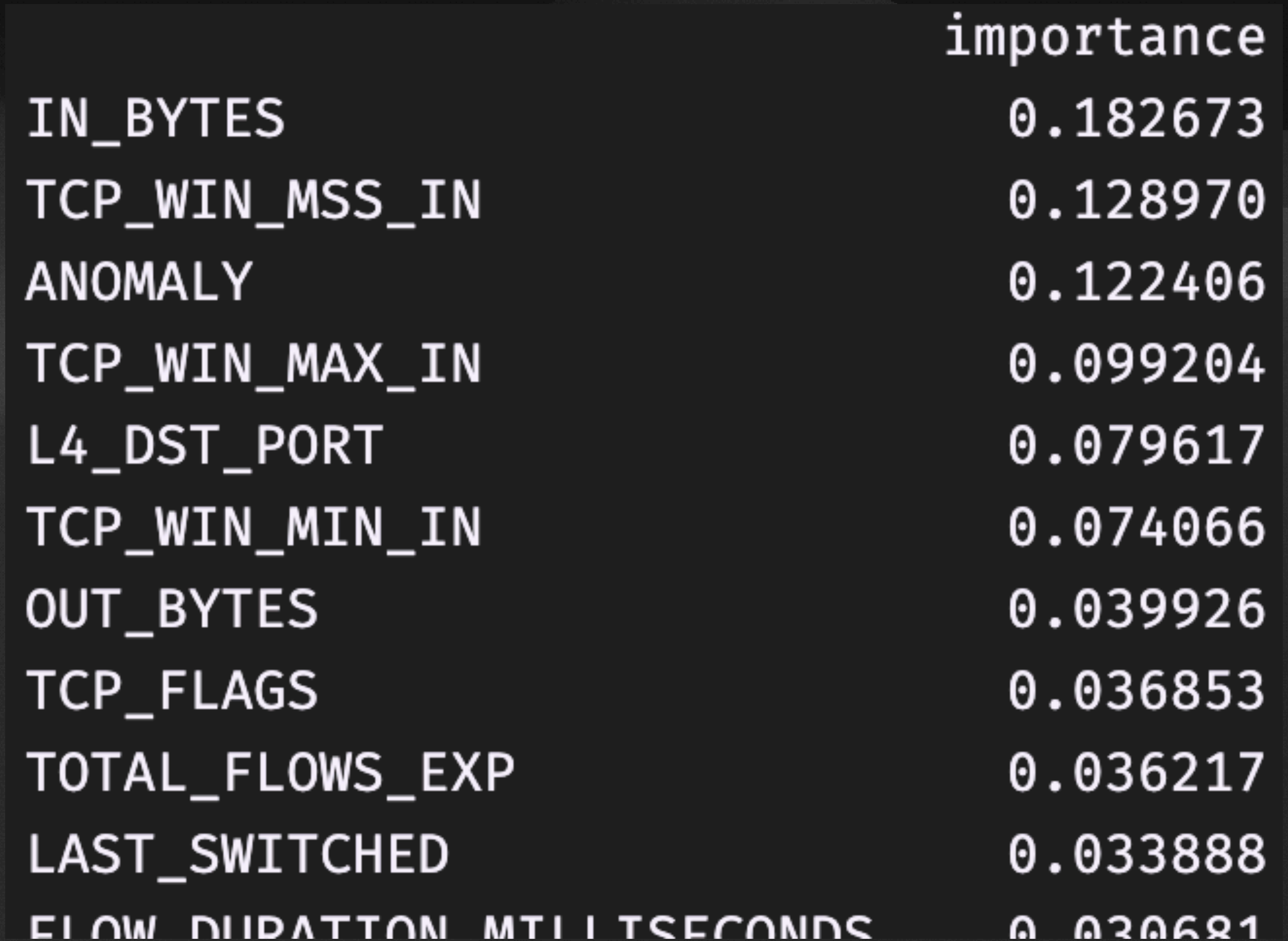
```
Train set distribution:
None                   0.868508
Port Scanning          0.119380
Denial of Service      0.011983
Malware                0.000128
Name: ALERT, dtype: float64

Validation set distribution:
None                   0.868490
Port Scanning          0.119379
Denial of Service      0.011973
Malware                0.000158
Name: ALERT, dtype: float64
```

Feature selection

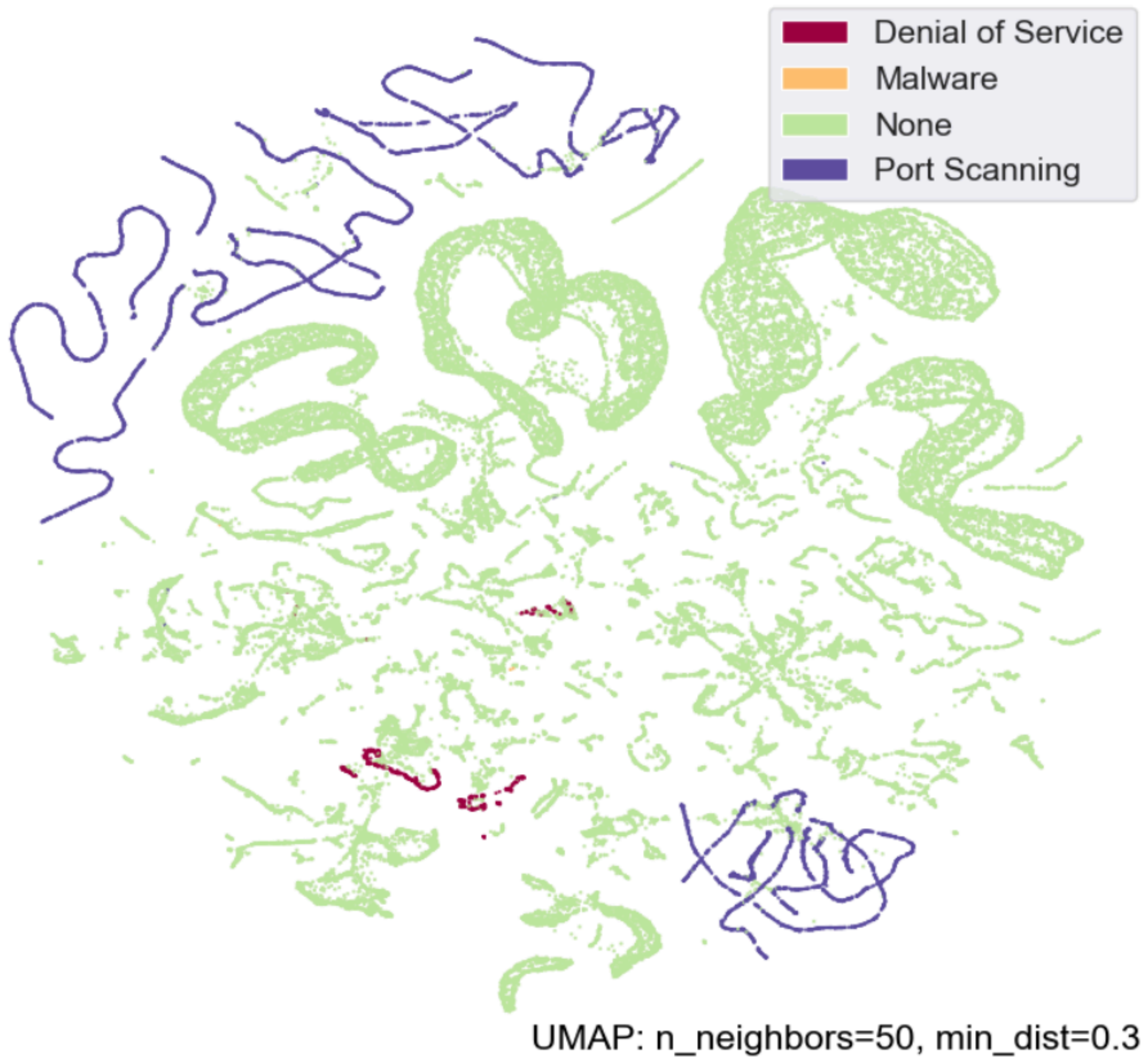
Many features, but which ones are important?

- Correlation matrix: a complete mess
  - By hand, visual method
  - Any correlation to malware
- Random Forest Regressor:
  - Better results!
  - Dynamic, with threshold (> 0.01)



```
                                importance
IN_BYTES                          0.182673
TCP_WIN_MSS_IN                    0.128970
ANOMALY                           0.122406
TCP_WIN_MAX_IN                    0.099204
L4_DST_PORT                       0.079617
TCP_WIN_MIN_IN                    0.074066
OUT_BYTES                         0.039926
TCP_FLAGS                         0.036853
TOTAL_FLOWS_EXP                   0.036217
LAST_SWITCHED                     0.033888
FLOW_DURATION_MILLISECONDS        0.030681
```

UMAP visualisation

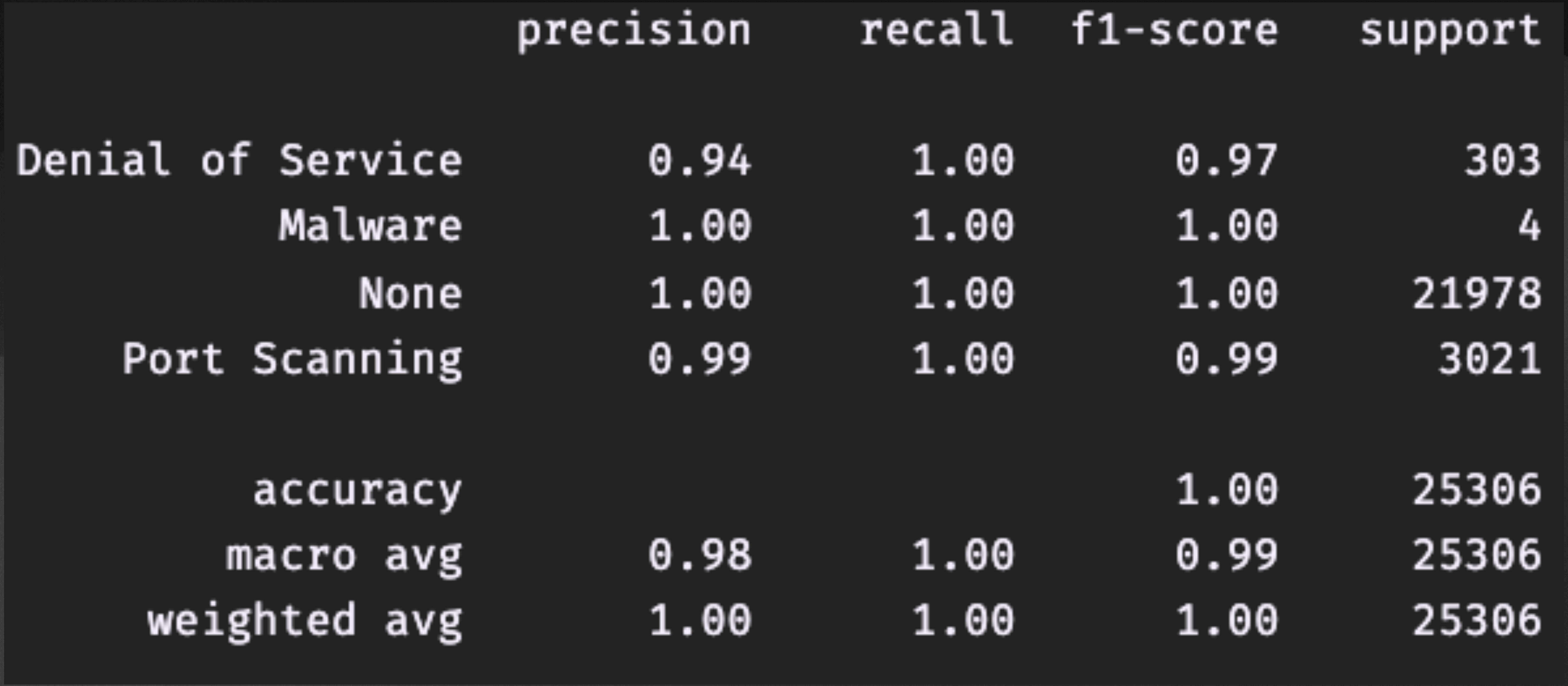


Classification: KNN + SVC + Bagging

(predictions and evaluation on validation set)

K-Nearest Neighbour

K = 3

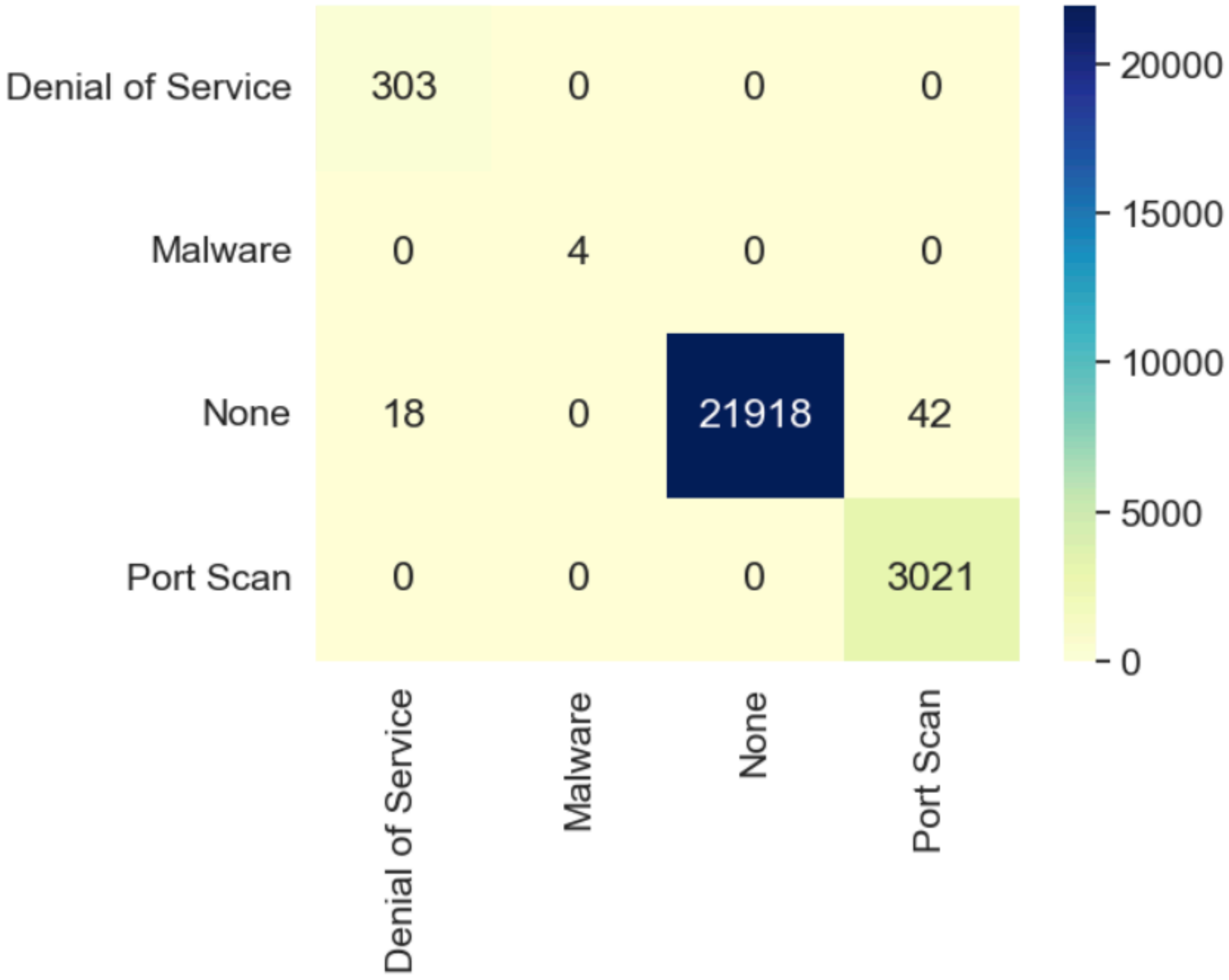


```
                   precision    recall  f1-score   support

Denial of Service       0.94      1.00      0.97       303
          Malware       1.00      1.00      1.00         4
             None       1.00      1.00      1.00     21978
    Port Scanning       0.99      1.00      0.99      3021

         accuracy                           1.00     25306
        macro avg       0.98      1.00      0.99     25306
     weighted avg       1.00      1.00      1.00     25306
```

Cross validation (5 folds) score: 0.997

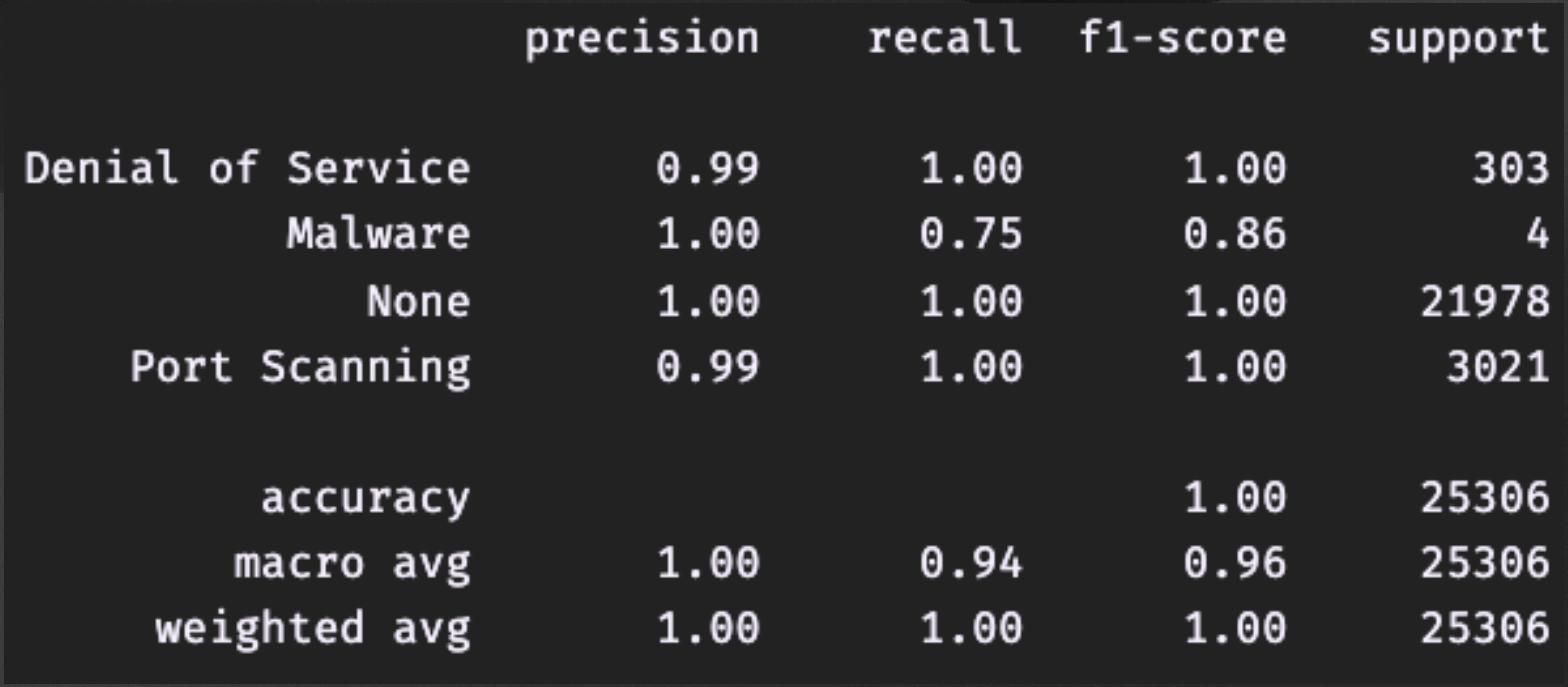


	Denial of Service	Malware	None	Port Scan
Denial of Service	303	0	0	0
Malware	0	4	0	0
None	18	0	21918	42
Port Scan	0	0	0	3021

Support Vector Classificator (SVC)

Pipeline: PCA + SVC

- PCA (10 components) - SVC (kernel=RBF, C=100 and Gamma=0.1)



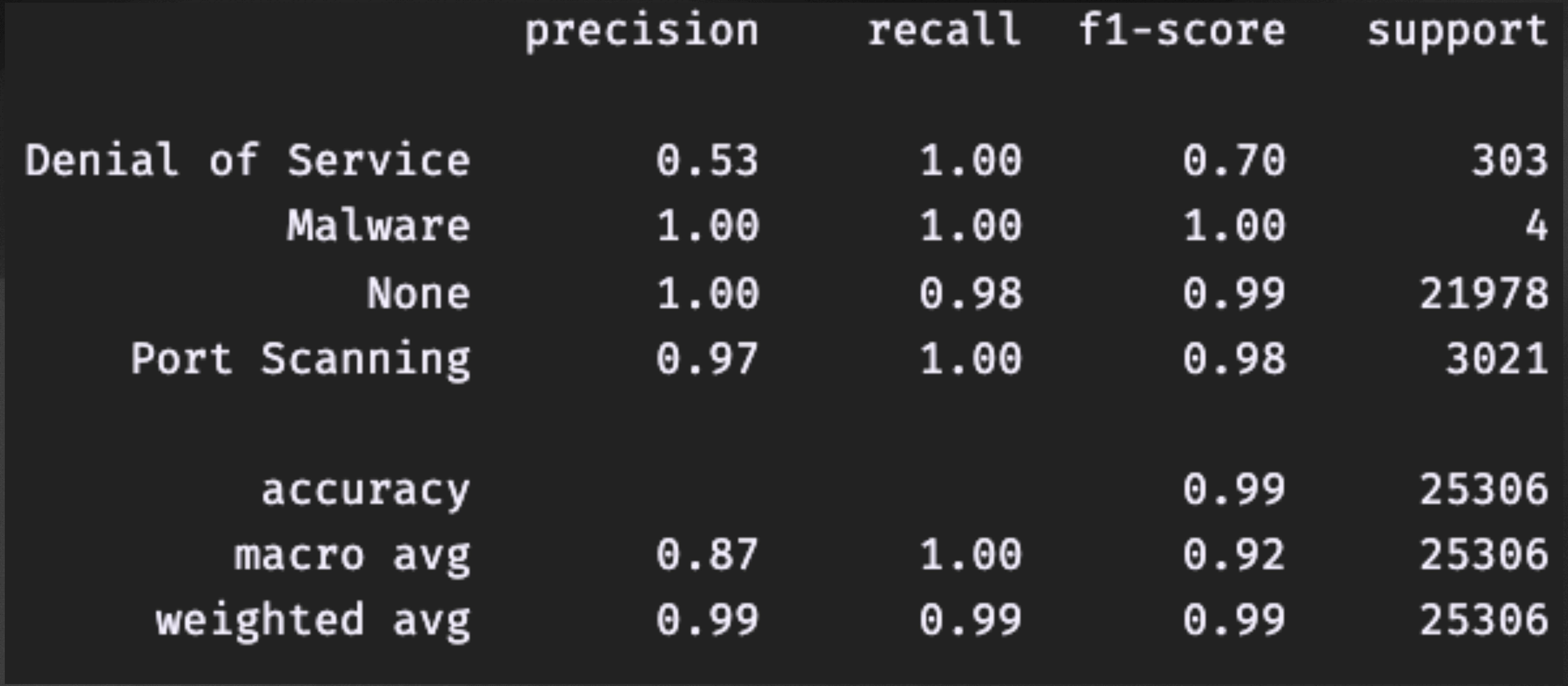
```
                   precision    recall  f1-score   support

Denial of Service       0.99      1.00      1.00       303
          Malware       1.00      0.75      0.86         4
             None       1.00      1.00      1.00     21978
    Port Scanning       0.99      1.00      1.00      3021

         accuracy                           1.00     25306
        macro avg       1.00      0.94      0.96     25306
     weighted avg       1.00      1.00      1.00     25306
```

Voting classifier

BaggingClassifier + 30x SVC estimators



```
                   precision    recall  f1-score   support

Denial of Service       0.53      1.00      0.70       303
          Malware       1.00      1.00      1.00         4
             None       1.00      0.98      0.99     21978
    Port Scanning       0.97      1.00      0.98      3021

         accuracy                           0.99     25306
        macro avg       0.87      1.00      0.92     25306
     weighted avg       0.99      0.99      0.99     25306
```

Why this precision?

Sources

- Cisco NetFlow v9 format
  - https://www.cisco.com/en/US/technologies/tk648/tk362/technologies_white_paper09186a00800a3db9.html
- Dataset source
  - https://www.kaggle.com/datasets/ashtcoder/network-data-schema-in-the-netflow-v9-format
- SciKit learn - StratifiedShuffleSplit
  - https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedShuffleSplit.html
- SciKit learn - BaggingClassifier
  - https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingClassifier.html