Department of Computer Science
Master's Degree in Cybersecurity

Data Analysis for Security

UNIVERSITÀ
DEGLI STUDI DI BARI
ALDO MORO

Synthesis of a KDD Pipeline for DDoS Connections Data Classification

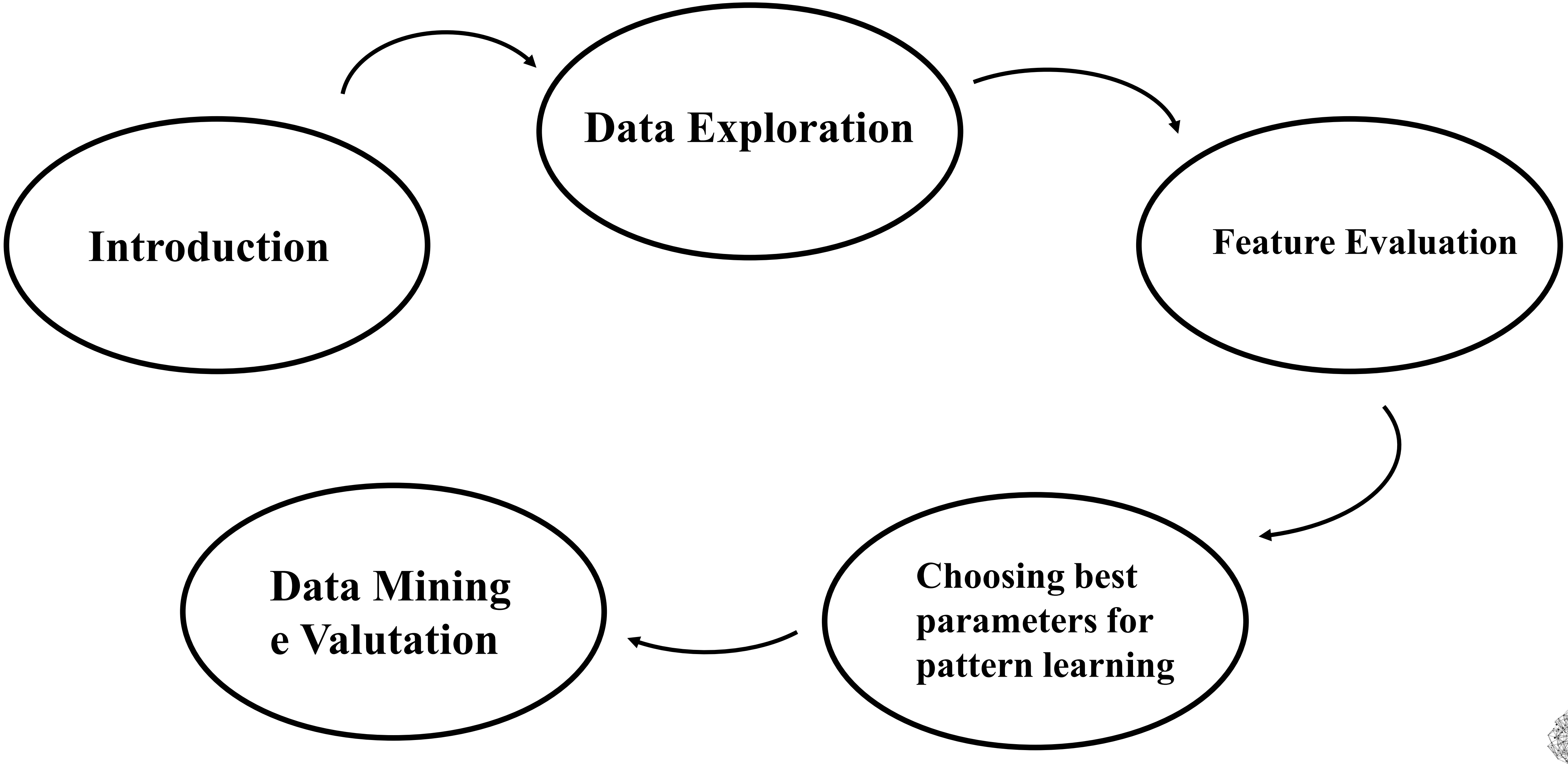
Supervisor:
Prof. Appice Annalisa

Student:
del Vescovo Samuele
Matr.: 766196

April 2023

A.A. 2021/2022

Outline



Introduction

Data Exploration

Feature Evaluation

Choosing best
parameters for
pattern learning

Data Mining
e Valutation

Introduction

- The Distributed Denial of Service (**DDoS**) attack aims to **exhaust** the **computational resources** of a target host (in a computer network) in order to make a service unavailable [1].

- As mentioned by Sharafaldin et al., it is becoming increasingly important to learn "patterns" capable of correctly and automatically identifying connections potentially related to DDoS attacks [2].

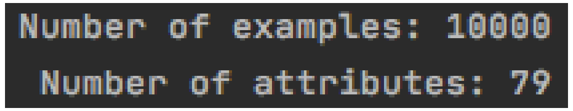
- The **goal** of this work is to synthesize a **KDD pipeline**, based on **supervised machine learning algorithm** in order to **classify connections** in the different classes of DDoS attacks ("BENIGN", "MSSQL", "Syn", "UDP" , “NetBIOS”).

- The chosen **dataset** derives from a **simplification** of the one proposed by the Canadian Institute for Cybersecurity [1].

Initial Dataset

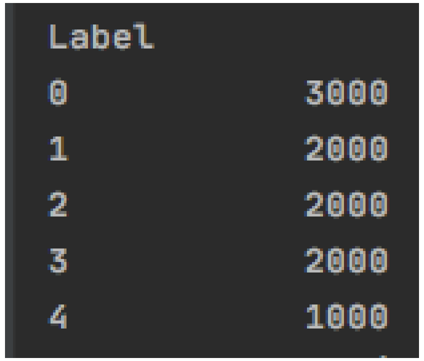
- The **training dataset**, in its initial form consists of 10000 examples described by 79 features (78 independent and 1 dependent) as evidenced by the "preElaborationClass" function.

- The **testing dataset** was retrieved from the same benchmark and consists of 2000 examples described by 79 features (78 independent and 1 dependent).



```
Number of examples: 10000
 Number of attributes: 79
```

Figure 1: The number of examples in the training set with number of features



```
Label
0            3000
1            2000
2            2000
3            2000
4            1000
```

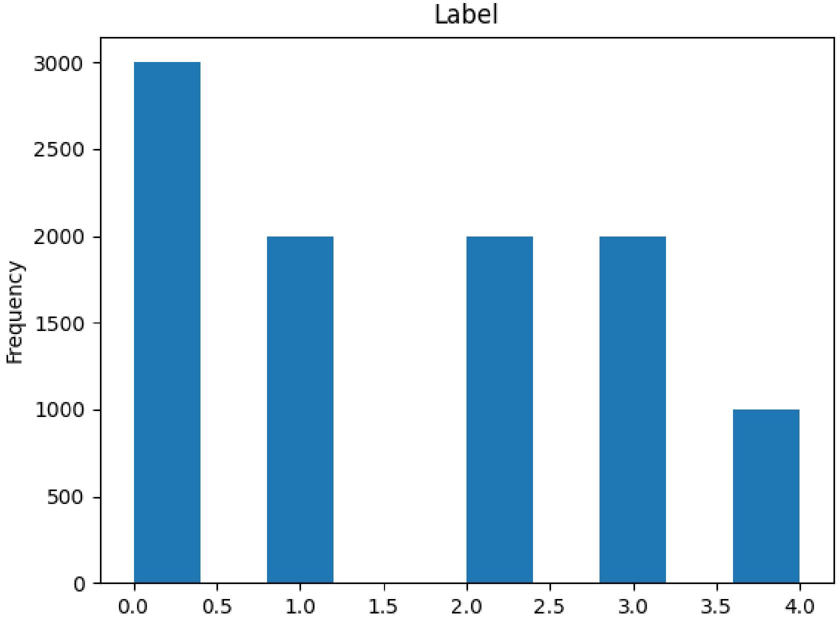


Figure 2: Distribution of examples of different classes

Data Exploration (1/6)

- Through specific functions coded in the "Pandas" framework, it is possible not only to **observe** some **key parameters** relating to each **independent variable distribution** but also to develop a **boxplot** that summarizes this information.

- This is useful in order to detect features that are **more significant** than others and **discriminating** with respect to dependent variable (target).

- Some examples of these will be shown below and compared with the output of the Features Evaluation techniques.

Data Exploration (2/6)

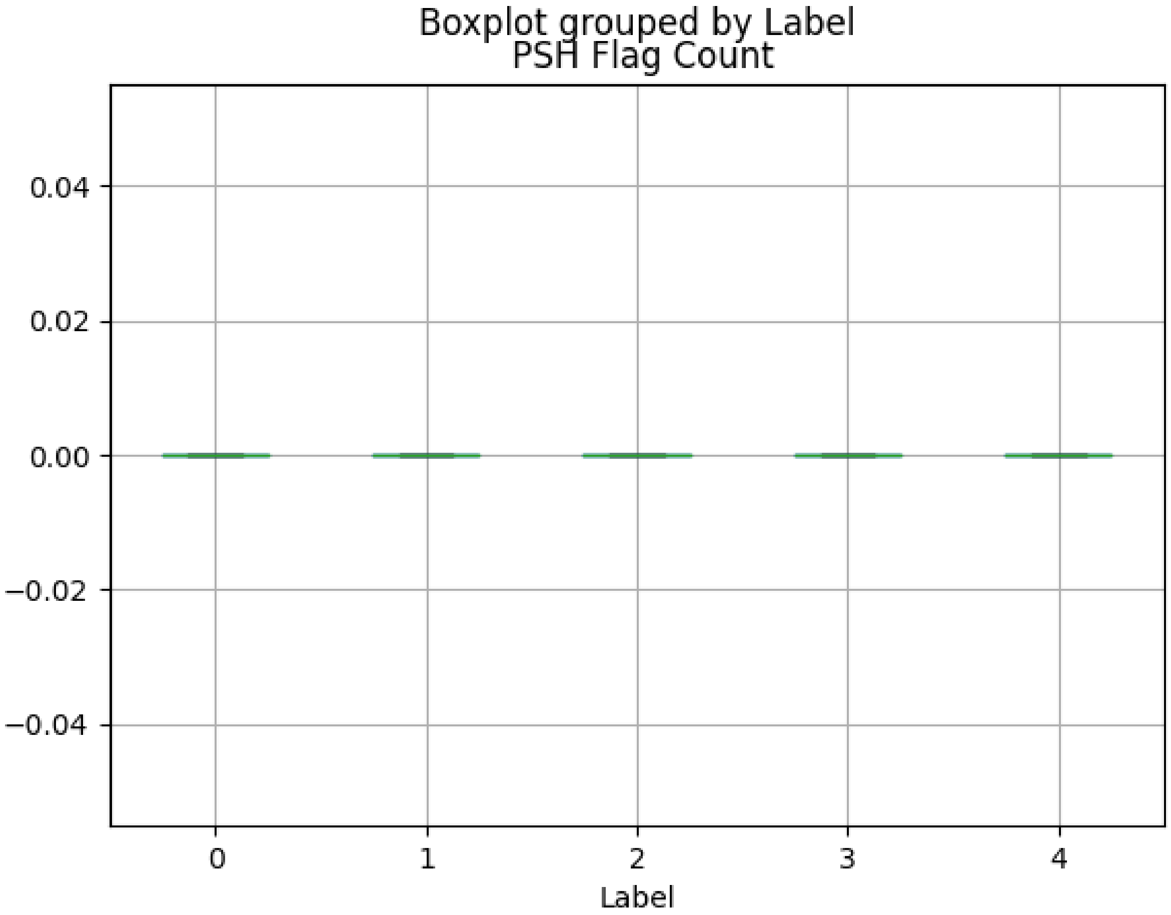
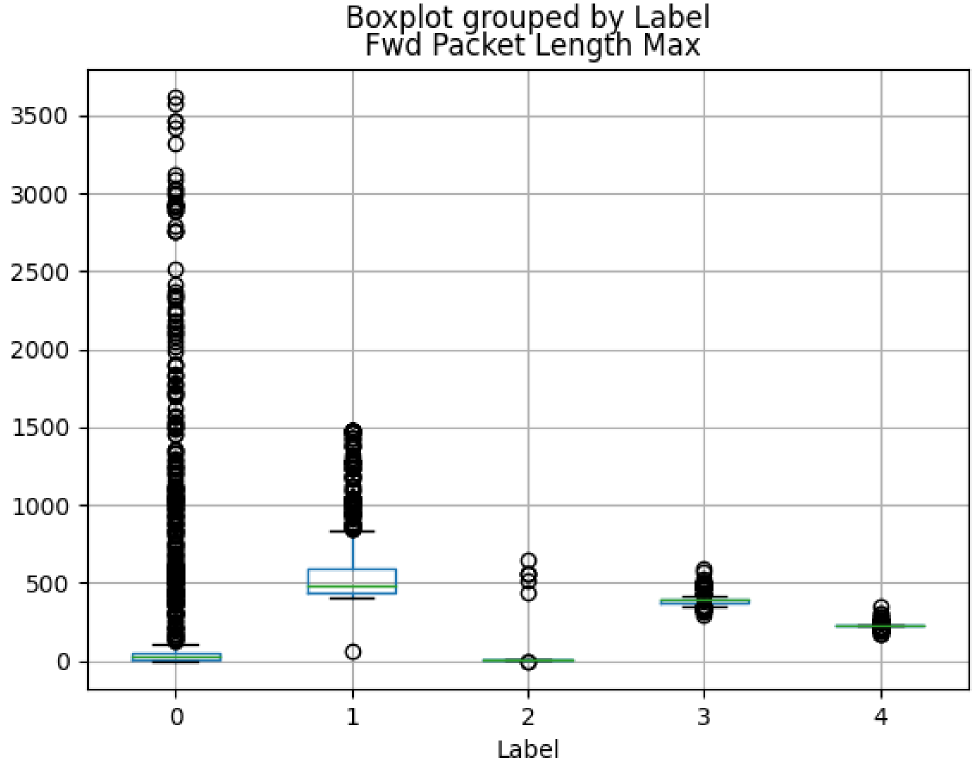


Figure 3: Example of a feature that has the maximum value equal to the minimum

- The “PSH Flag Count” feature has the same **maximum** and **minimum value** for all classes.
- This feature is **useless** to the data mining task and can be removed.
- This characteristic is also observed in 11 other features.
- Following this screening phase, the features taken into consideration will be **66** (excluding the label).

Data Exploration (3/6)



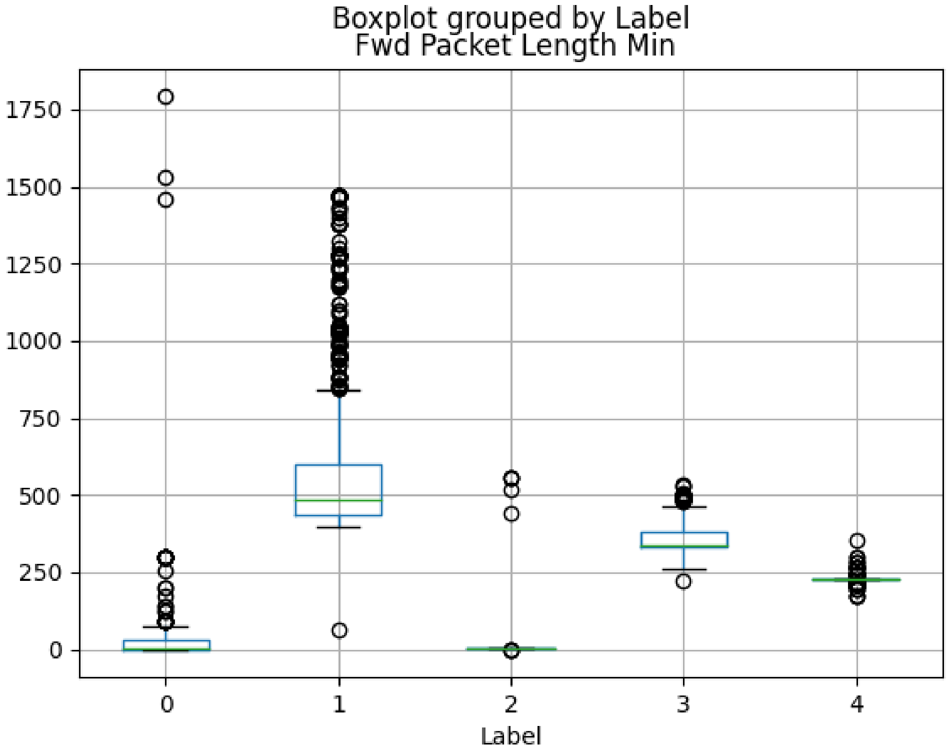


Figure 4: Example of features representing a trend

- The features “Fwd Packet Length Max” and “Fwd Packet Length Min” present a **greater amplitude** of the **third quartile** for the class “1” compared to all other classes (trend).

- It is expected that these **features** are in the **intermediate positions** (tending to the top) of the **rank** obtained by the **Features Evaluation** techniques.

Data Exploration (4/6)

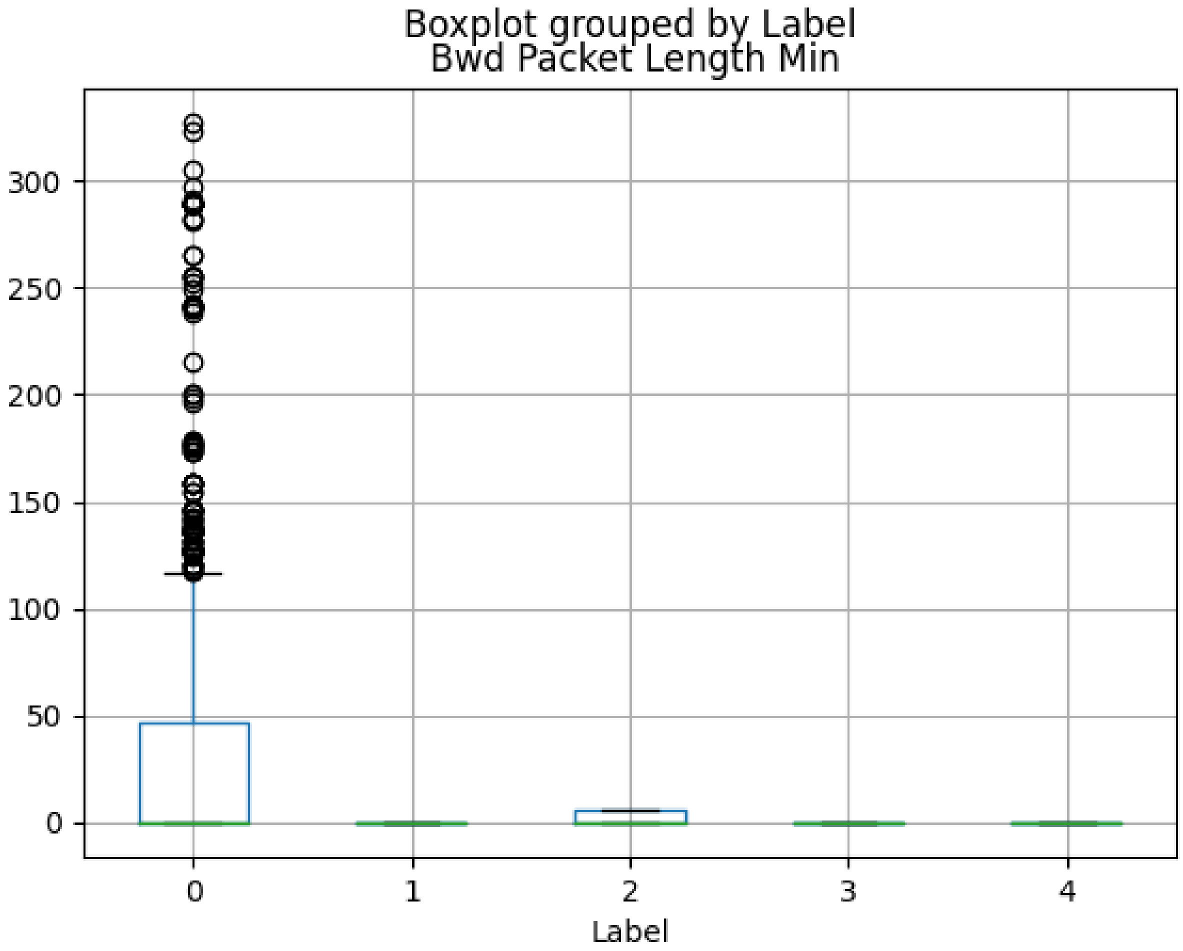


Figure 5: Example of a feature considered not very discriminating

- The "Bwd Packet Length Min" feature has very **similar boxes** between the 4 **classes** of DDoS attacks.

- It is expected that this **feature** is in the **last positions** of the rank obtained by Features Evaluation techniques.

Data Exploration (5/6)

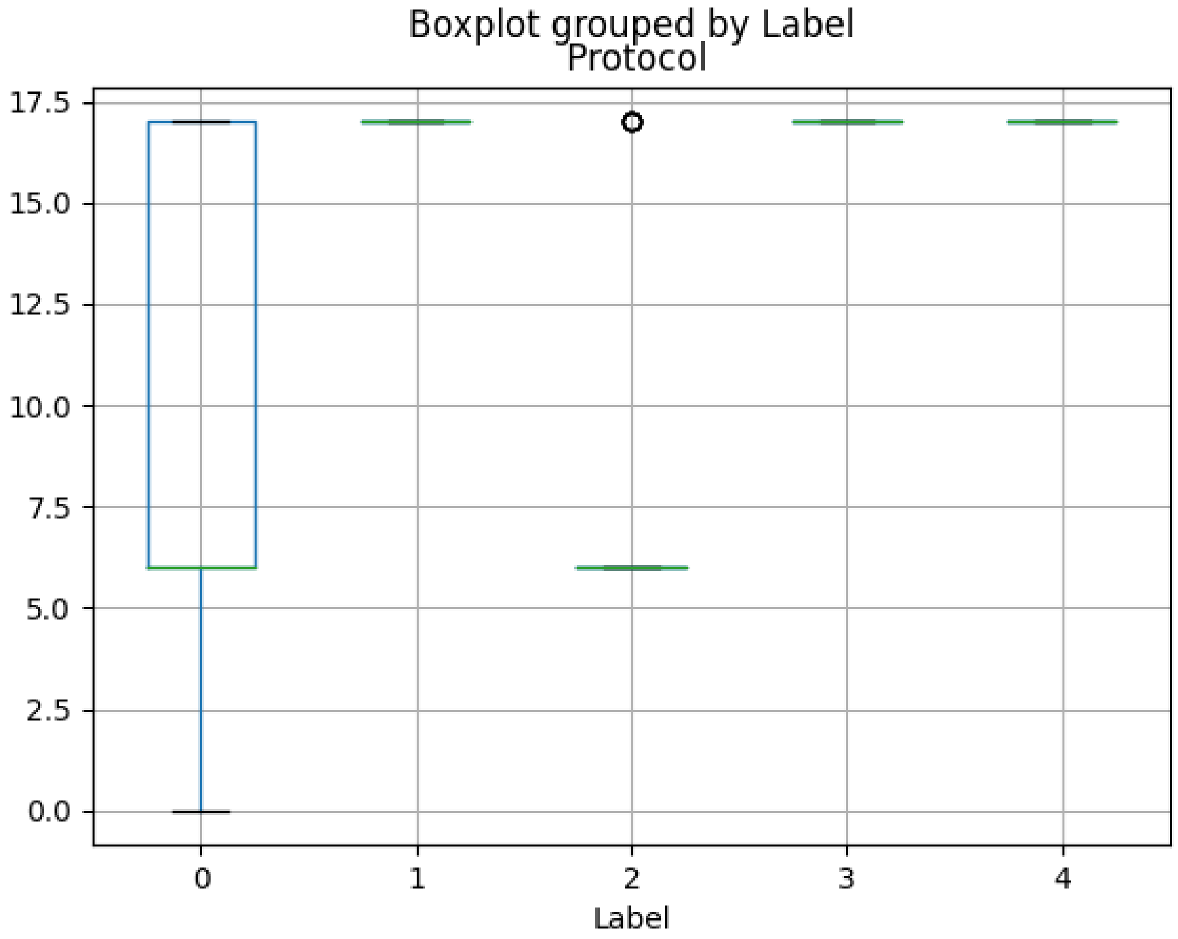


Figure 6: Example of an feature considered to be on average discriminating

- The "Protocol" feature has **similar boxes** between **classes "1", "3", "4"** but the one relating to **class "2"** is different from the others.

- Furthermore, this feature has **different boxes** relating to **class "0" and "2"** (size of the third quartile).

- It is expected that this **feature** is in the **intermediate positions** of the rank obtained by Features Evaluation techniques.

Data Exploration (6/6)

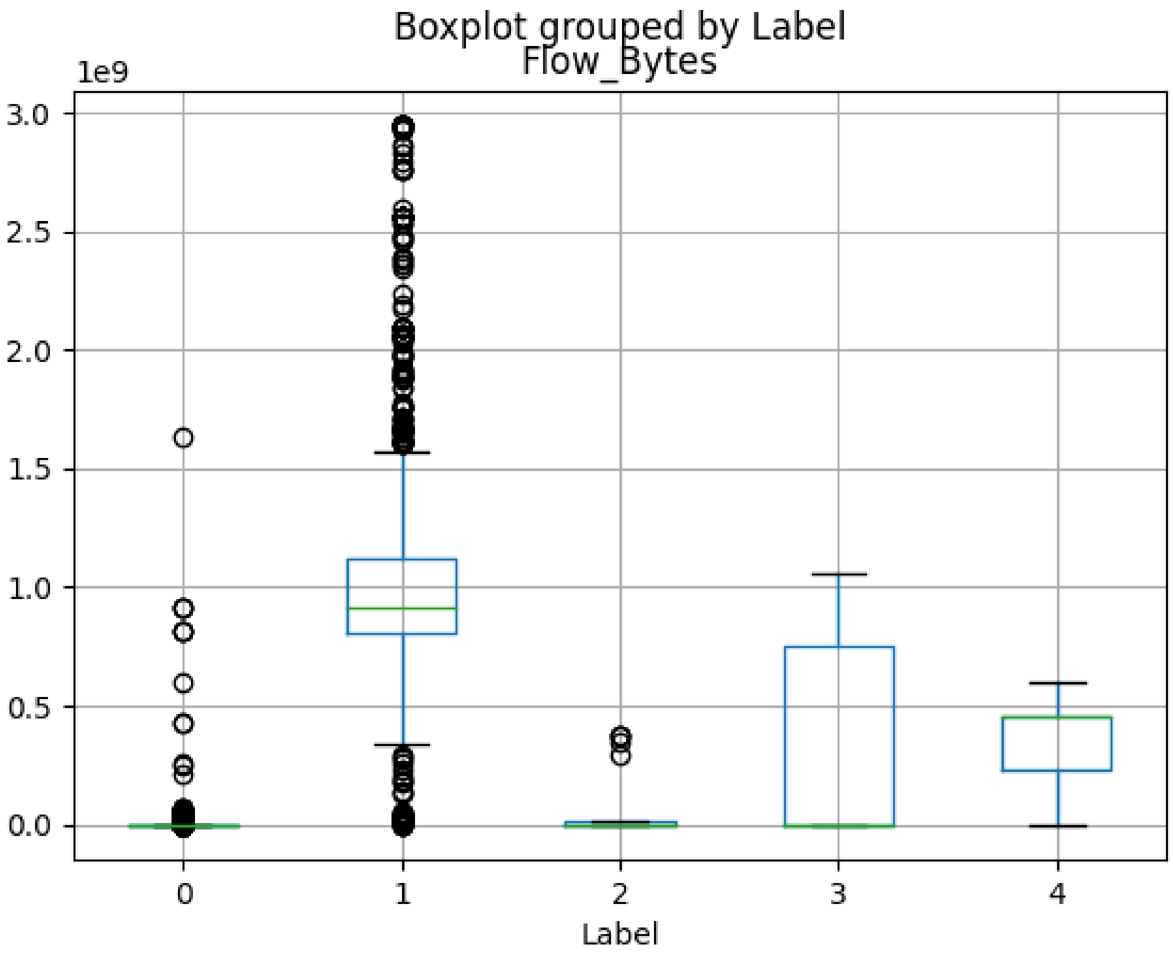


Figure 7: Example of a feature considered highly discriminating

- The "Flow_Bytes" feature presents **different boxes** between the **classes "1", "3", "4"** observing the median and amplitude values of the quartiles.

- The **boxes** related to **classes "0" and "2"** are **similar**.

- It is expected that this feature is in the **top positions** of the rank obtained by Features Evaluation techniques.

Feature Evaluation: MI Rank

- (' Average Packet Size', 1.3934006378924466), ('Total Length of Fwd Packets', 1.3902809876637139), (' Subflow Fwd Bytes', 1.3887076352069208), (' Avg Fwd Segment Size', 1.366270300291472), (' Fwd Packet Length Mean', 1.3656162206600995), ('Flow_Bytes', 1.3613799868669945), (' Max Packet Length', 1.3535445478071175), (' Min Packet Length', 1.348555021000605), (' Packet Length Mean', 1.3445635380055723), (' Fwd Packet Length Min', 1.3433241110286873), (' Fwd Packet Length Max', 1.3259154236594968), ('Init_Win_bytes_forward', 0.7747075413798603), (' Flow Duration', 0.6485522543748727), (' Flow IAT Mean', 0.6473879126265007), ('Flow_Packets', 0.6471338253158803), (' Flow IAT Max', 0.6385838184357386), ('Fwd_Packets', 0.6382552831296682), (' Flow IAT Std', 0.6055254478045582), (' Fwd Header Length', 0.5517727816731264), (' Fwd Header Length.1', 0.5445200186011923), ('Fwd IAT Total', 0.5168617660311243), (' Fwd IAT Max', 0.5093666320876884), (' Protocol', 0.50800505960873), (' Fwd IAT Mean', 0.5039607303207718), (' Packet Length Variance', 0.45755401119364736), ('Bwd_Packets', 0.45005936643085276), (' Packet Length Std', 0.4455814679573735), (' ACK Flag Count', 0.40605934045642345), (' min_seg_size_forward', 0.3950240226379522), (' act_data_pkt_fwd', 0.36639303517640665), (' Subflow Bwd Bytes', 0.3508136862678737), (' Bwd Header Length', 0.3507016807529699), ('Bwd IAT Total', 0.3457414040151936), (' Total Length of Bwd Packets', 0.3419224587160472), (' Total Backward Packets', 0.3387019588257769), (' Subflow Bwd Packets', 0.33732492262053526), ('Bwd Packet Length Max', 0.33696285147330585), (' Bwd IAT Max', 0.3358151077790148), (' Bwd IAT Mean', 0.33304035173249 6), (' Bwd Packet Length Mean', 0.33099420923253975), (' Avg Bwd Segment Size', 0.32884398779062396), (' Total Fwd Packets', 0.31910854222048535), (' Init_Win_bytes_backward', 0.31705814263355325), ('Subflow Fwd Packets', 0.3131257770806486), (' Fwd IAT Std', 0.3089135684037041), (' Fwd Packet Length Std', 0.3032180831990887), (' Bwd IAT Min', 0.28355429882768224), (' Bwd Packet Length Min', 0.2524770516631678), (' Down/Up Ratio', 0.2440625247673167), (' URG Flag Count', 0.21454509954632117), (' Flow IAT Min', 0.17878693042252403), (' Fwd IAT Min', 0.17425798081424304), (' Bwd IAT Std', 0.09513154098718157), (' Idle Max', 0.0871325522482395 7), (' CWE Flag Count', 0.08596463816711042), ('Active Mean', 0.08563421244151259), (' Active Min', 0.08457533867703981), (' Active Max', 0.08188426448286945), ('Idle Mean', 0.07898003353106553), (' Bwd Packet Length Std', 0.07421372291602113), (' Idle Min', 0.06968683541258391), (' RST Flag Count', 0.06929297610568641), (' Idle Std', 0.06694676194420612), ('Fwd PSH Flags', 0.06691395274750889), (' Active Std', 0.05821756380607468), (' SYN Flag Count', 0.005935646173135023)

random_state=42

Numpy Seed = 42

Feature Evaluation: IG Rank

- ('Flow_Bytes', 0.9403050309957603), (' Average Packet Size', 0.9102146306170038), ('Total Length of Fwd Packets', 0.8981629006375397), (' Subflow Fwd Bytes', 0.8981629006375397), (' Packet Length Mean', 0.8796498907935585), (' Fwd Packet Length Mean', 0.8735902790580787), (' Avg Fwd Segment Size', 0.8735902790580787), (' Max Packet Length', 0.8607302363583886), (' Fwd Packet Length Max', 0.849027481844839), (' Min Packet Length', 0.8443248214760471), (' Fwd Packet Length Min', 0.8435961794299786), ('Init_Win_bytes_forward', 0.4810499353288393), ('Fwd_Packets', 0.455467963941771), ('Flow_Packets', 0.45433243499510056), (' Flow IAT Mean', 0.4518795162634386), (' Flow Duration', 0.4427004867276081), (' Flow IAT Max', 0.4344445842848571), (' Flow IAT Std', 0.42782931994560125), (' Fwd Header Length', 0.3614212377531054), (' Fwd Header Length.1', 0.3614212377531054), ('Fwd IAT Total', 0.3514424121618669), (' Fwd IAT Mean', 0.35023716499135826), (' Fwd IAT Max', 0.34859183722367026), (' Protocol', 0.3130129861623332), ('Bwd_Packets', 0.2898192619325598), (' Packet Length Std', 0.28802419887441466), (' Packet Length Variance', 0.28802419887441466), (' ACK Flag Count', 0.2590554718786756), (' min_seg_size_forward', 0.25194972077027644), (' act_data_pkt_fwd', 0.23022495268022625), (' Bwd Header Length', 0.22659002074353207), (' Fwd IAT Std', 0.22639482864838267), ('Bwd IAT Total', 0.2214034174783881), (' Bwd IAT Mean', 0.22077285834151106), (' Bwd IAT Max', 0.22046863550455975), (' Total Length of Bwd Packets', 0.2194047029624847), (' Subflow Bwd Bytes', 0.2194047029624847), (' Bwd Packet Length Mean', 0.20993940390453836), (' Avg Bwd Segment Size', 0.20993940390453836), ('Bwd Packet Length Max', 0.2095458963002551), (' Total Backward Packets', 0.20875262472326706), (' Subflow Bwd Packets', 0.20875262472326706), (' Fwd Packet Length Std', 0.2038114650423003), (' Total Fwd Packets', 0.1986536930379722), ('Subflow Fwd Packets', 0.1986536930379722), (' Init_Win_bytes_backward', 0.1897017250768404 7), (' Bwd IAT Min', 0.17465661627792217), (' Bwd Packet Length Min', 0.1539795040076889), (' Down/Up Ratio', 0.14613569841843566), (' URG Flag Count', 0.1309552859645371), (' Fwd IAT Min', 0.12172723049732936), (' Flow IAT Min', 0.112521223433803), (' Bwd IAT Std', 0.07377566544621961), ('Idle Mean', 0.059897424110549546), (' Idle Max', 0.059897424110549546), (' Idle Min', 0.059897424110549546), ('Active Mean', 0.05814504456452341), (' Active Max', 0.057915843221322594), (' Active Min', 0.05660012881939358), (' CWE Flag Count', 0.05044661219672453), (' Idle Std', 0.0478455236023001), (' Bwd Packet Length Std', 0.0464164202554476 8), (' Active Std', 0.045767857970087 98), ('Fwd PSH Flags', 0.04420173735039301), (' RST Flag Count', 0.04420173735039301), (' SYN Flag Count', 0.0007796186519307691)

MI Rank ≠ IG Rank

Feature Evaluation: MI Rank (Scaled) (1/2)

- We tried to apply a **scaling algorithm** (**MinMaxScaling** where the new minimum is "0" and the new maximum is "1") to the training dataset and to re-evaluate the features via Mutual Info Rank.

- Looking at the documentation [3], there is a “**random_state**” parameter which represents the seed of a **pseudorandom number generator algorithm**; therefore (as mentioned previously) this parameter was set to a **constant value** as was the **numpy seed**.

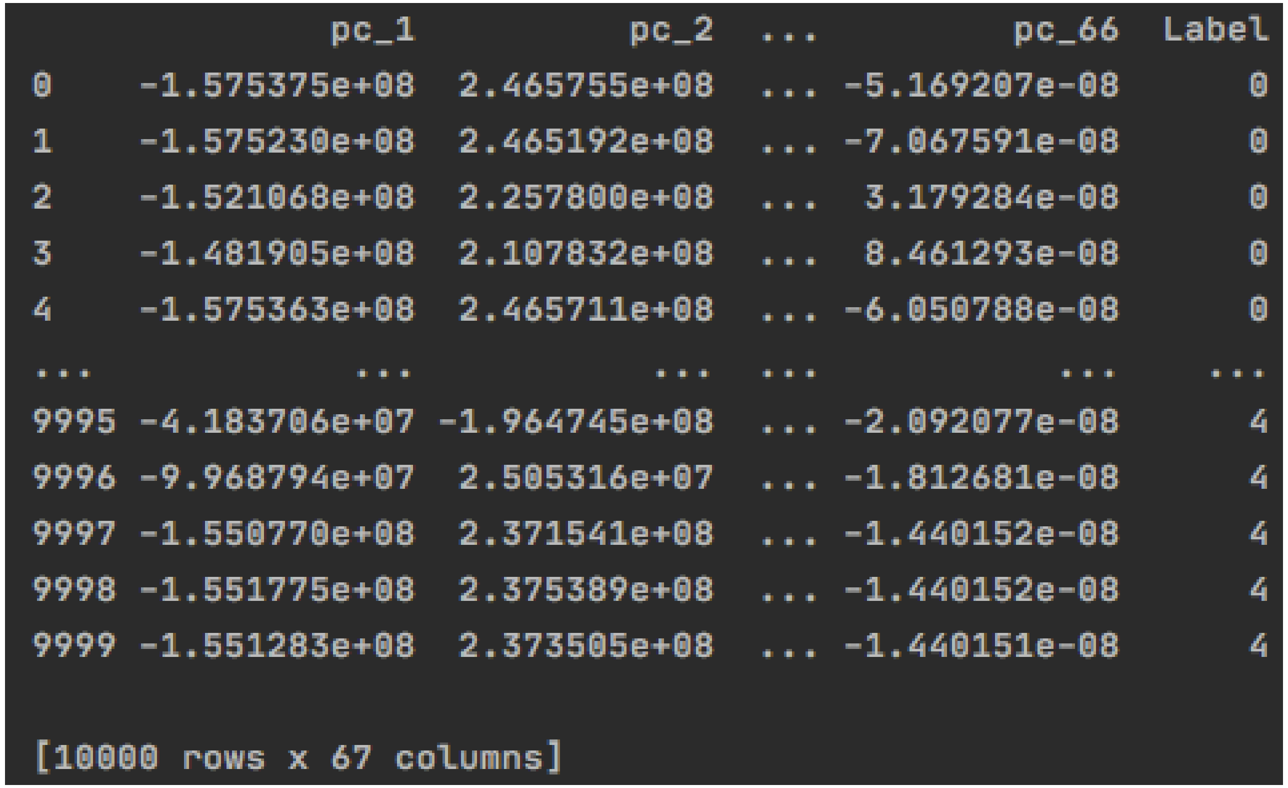
- It is expected that the features ranking related to the unscaled dataset and that related to the scaled dataset should be **the same** but **they are different**. That is strange because the scaling operation does not change the distribution of the data.

- In the next slide, the features labeled with the blue color are oberved in "reversed" positions in the two ranks while those labeled with the orange color are observed in different positions in the two ranks (but in pairs).

Feature Evaluation: MI Rank (Scaled) (2/2)

- (' Average Packet Size', 1.3934006378924466), ('Total Length of Fwd Packets', 1.3902809876637139), (' Subflow Fwd Bytes', 1.3887076352069208), (' Avg Fwd Segment Size', 1.366270300291472), (' Fwd Packet Length Mean', 1.3655662206600994), ('Flow_Bytes', 1.355899852097116), (' Max Packet Length', 1.3535445478071175), (' Min Packet Length', 1.348555021000605), (' Packet Length Mean', 1.3445635380055723), (' Fwd Packet Length Min', 1.3431241110286873), (' Fwd Packet Length Max', 1.3259154236594968), ('Init_Win_bytes_forward', 0.7745575413798602), (' Flow Duration', 0.6485522543748727), (' Flow IAT Mean', 0.6471515244124537), ('Flow_Packets', 0.6470870014762247), (' Flow IAT Max', 0.6385838184357386), ('Fwd_Packets', 0.638026574020921), (' Flow IAT Std', 0.6052801572555881), ('Fwd IAT Total', 0.5168617660311243), (' Fwd IAT Max', 0.5093666320876884), (' Protocol', 0.5078412044230662), (' Fwd Header Length', 0.4860303716078469), (' Fwd Header Length.1', 0.47980919082213047), (' Packet Length Variance', 0.45755401119364736), (' Fwd IAT Mean', 0.45236881035018195), ('Bwd_Packets', 0.44968529665460943), (' Packet Length Std', 0.4455814679573735), (' ACK Flag Count', 0.40605934045642345), (' min_seg_size_forward', 0.3951684273998568), (' act_data_pkt_fwd', 0.3662714539798255), (' Bwd Header Length', 0.3509012421126627), (' Subflow Bwd Bytes', 0.3508136862678737), ('Bwd IAT Total', 0.3457414040151936), (' Total Length of Bwd Packets', 0.3419224587160472), (' Total Backward Packets', 0.3387019588257769), (' Subflow Bwd Packets', 0.33732492262053526), ('Bwd Packet Length Max', 0.33696285147330585), (' Bwd IAT Max', 0.3358151077790148), (' Bwd IAT Mean', 0.3318411908646292 7), (' Bwd Packet Length Mean', 0.33099420923253975), (' Avg Bwd Segment Size', 0.32884398779062396), (' Total Fwd Packets', 0.3190613763790804), (' Init_Win_bytes_backward', 0.31705814263355325), ('Subflow Fwd Packets', 0.3130477649150891), (' Fwd IAT Std', 0.3089135684037041), (' Fwd Packet Length Std', 0.3032180831990887), (' Bwd IAT Min', 0.2835507273991107), (' Bwd Packet Length Min', 0.2524570516631681), (' Down/Up Ratio', 0.24393085810064985), (' URG Flag Count', 0.21454509954632117), (' Flow IAT Min', 0.17878693042252403), (' Fwd IAT Min', 0.17425798081424304), (' Bwd IAT Std', 0.09513154098718157), (' Idle Max', 0.08713255224823957), (' CWE Flag Count', 0.08596463816711042), ('Active Mean', 0.08563421244151259), (' Active Min', 0.08457533867703981), (' Active Max', 0.08188426448286945), ('Idle Mean', 0.07898003353106553), (' Bwd Packet Length Std', 0.07421372291602113), (' Idle Min', 0.06968683541258391), (' RST Flag Count', 0.06929297610568641), (' Idle Std', 0.06694676194420612), ('Fwd PSH Flags', 0.06691395274750889), (' Active Std', 0.05821756380607468), (' SYN Flag Count', 0.005935646173135023)

Feature Evaluation: PCA Rank

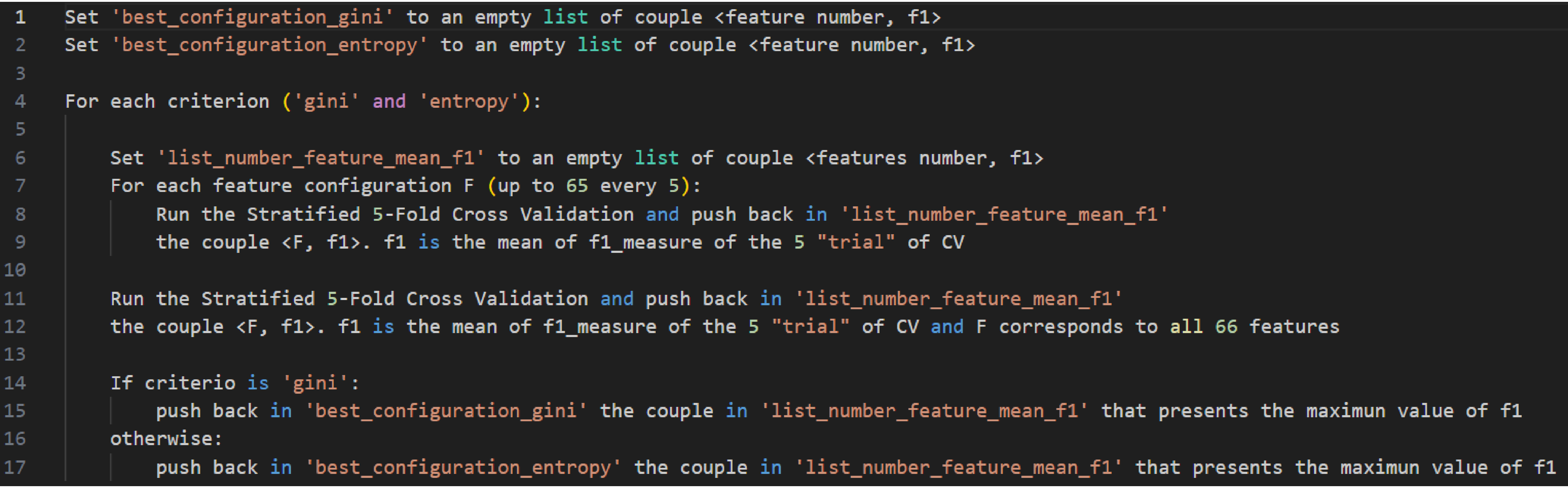


```
                pc_1          pc_2  ...         pc_66  Label
0    -1.575375e+08  2.465755e+08  ... -5.169207e-08      0
1    -1.575230e+08  2.465192e+08  ... -7.067591e-08      0
2    -1.521068e+08  2.257800e+08  ...  3.179284e-08      0
3    -1.481905e+08  2.107832e+08  ...  8.461293e-08      0
4    -1.575363e+08  2.465711e+08  ... -6.050788e-08      0
...            ...           ...  ...           ...    ...
9995 -4.183706e+07 -1.964745e+08  ... -2.092077e-08      4
9996 -9.968794e+07  2.505316e+07  ... -1.812681e-08      4
9997 -1.550770e+08  2.371541e+08  ... -1.440152e-08      4
9998 -1.551775e+08  2.375389e+08  ... -1.440152e-08      4
9999 -1.551283e+08  2.373505e+08  ... -1.440151e-08      4

[10000 rows x 67 columns]
```

Figure 8: Dataset described by the main components (the order of these is implicit)

- The model learned to perform the PCA will also be used in the testing phase.

How to choose the best configuration?



```
Set 'best_configuration_gini' to an empty list of couple <feature number, f1>
Set 'best_configuration_entropy' to an empty list of couple <feature number, f1>

For each criterion ('gini' and 'entropy'):

    Set 'list_number_feature_mean_f1' to an empty list of couple <features number, f1>
    For each feature configuration F (up to 65 every 5):
        Run the Stratified 5-Fold Cross Validation and push back in 'list_number_feature_mean_f1'
        the couple <F, f1>. f1 is the mean of f1_measure of the 5 "trial" of CV

    Run the Stratified 5-Fold Cross Validation and push back in 'list_number_feature_mean_f1'
    the couple <F, f1>. f1 is the mean of f1_measure of the 5 "trial" of CV and F corresponds to all 66 features

    If criterio is 'gini':
        push back in 'best_configuration_gini' the couple in 'list_number_feature_mean_f1' that presents the maximun value of f1
    otherwise:
        push back in 'best_configuration_entropy' the couple in 'list_number_feature_mean_f1' that presents the maximun value of f1
```

Figure 9: Pseudocode relating to the function useful for choosing the best decision tree configuration for the train

K-Fold CV Parameters [5]→ K=5, seed=42 e shuffle = true

Numpy seed =42

Decision Tree Parameter [4] → splitter="best", random_state=0, min_samples_split=500

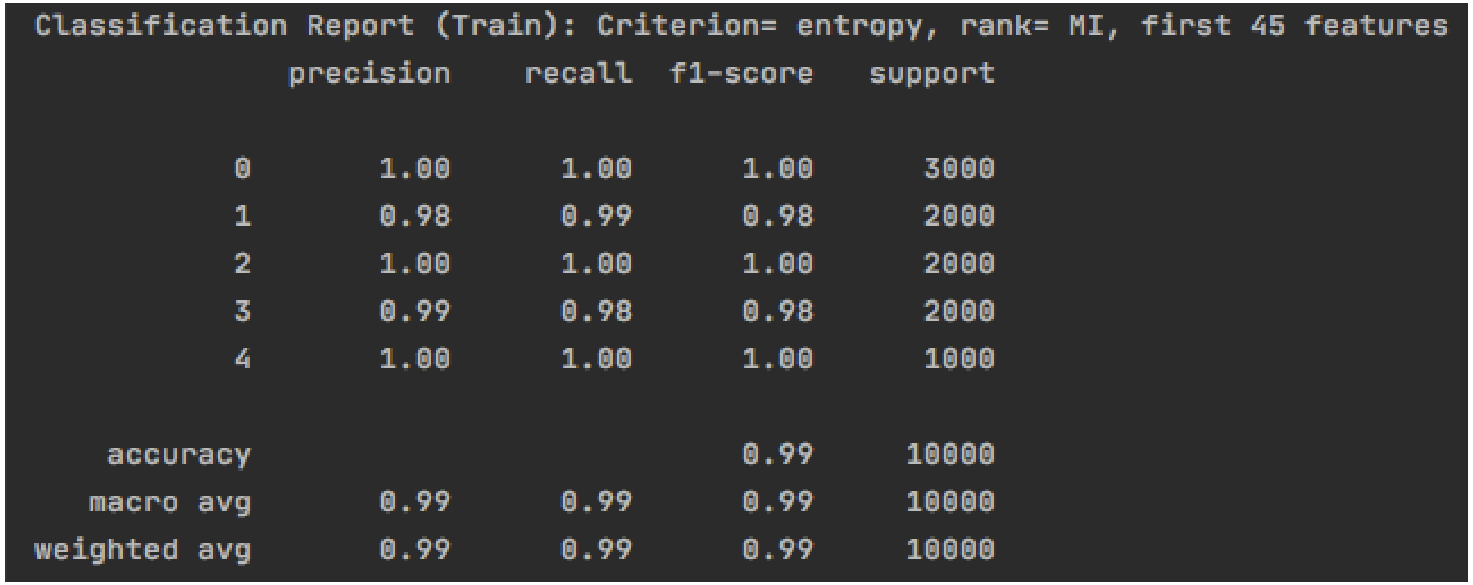
F1-Score Type → weighted

Best Tree Training (MI Rank)

- **Best Configuration**:
  - Splitting Criterion: Entropy
  - Features Number: 45

- **Are the best expected feature present in this configuration?**
  - All the features examined are present except “Bwd Packet Length Min” (considered terrible)

- **Details about the learned tree**:
  - Nodes number: 47
  - Leaf number: 24



```
Classification Report (Train): Criterion= entropy, rank= MI, first 45 features
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      3000
           1       0.98      0.99      0.98      2000
           2       1.00      1.00      1.00      2000
           3       0.99      0.98      0.98      2000
           4       1.00      1.00      1.00      1000

    accuracy                           0.99     10000
   macro avg       0.99      0.99      0.99     10000
weighted avg       0.99      0.99      0.99     10000
```

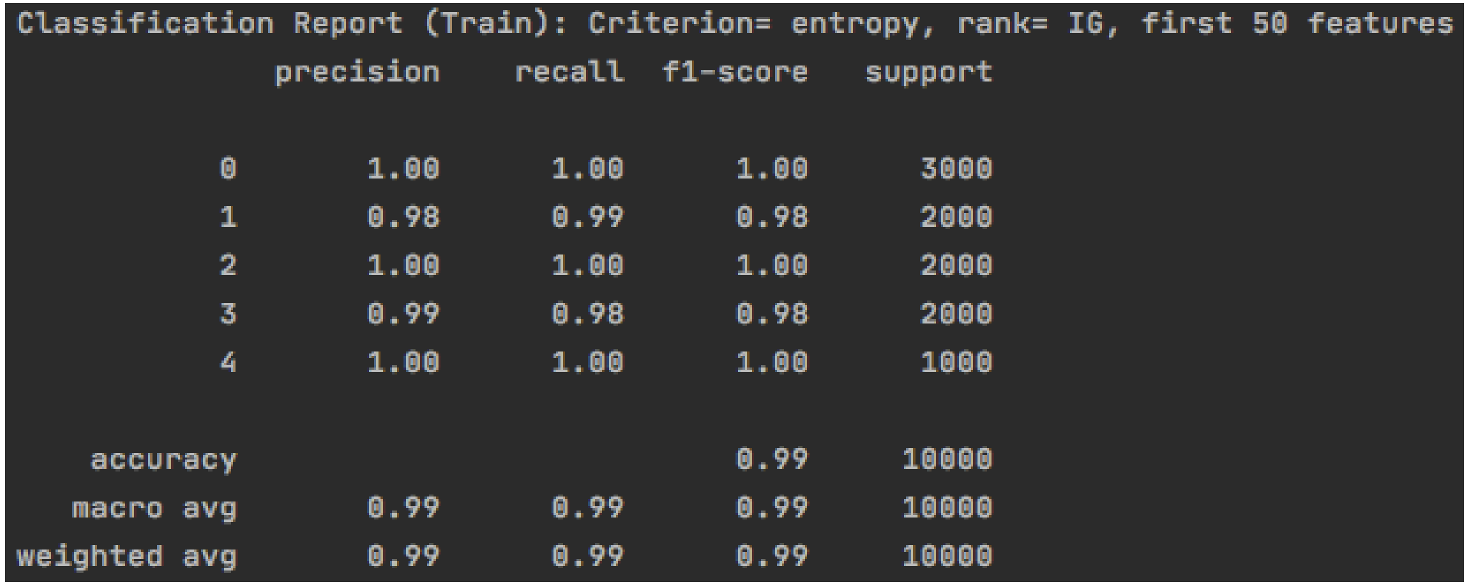
Figure 10: Classification report relating to the decision tree trained on the dataset sorted according to Mutual Info

Best Tree Training (IG Rank)

- **Best Configuration**:
  - Splitting Criterion: Entropy
  - Features Number: 50

- **Are the best expected feature present in this configuration?**
  - All the features examined are present

- **Details about the learned tree**:
  - Nodes number: 49
  - Leaf number: 25



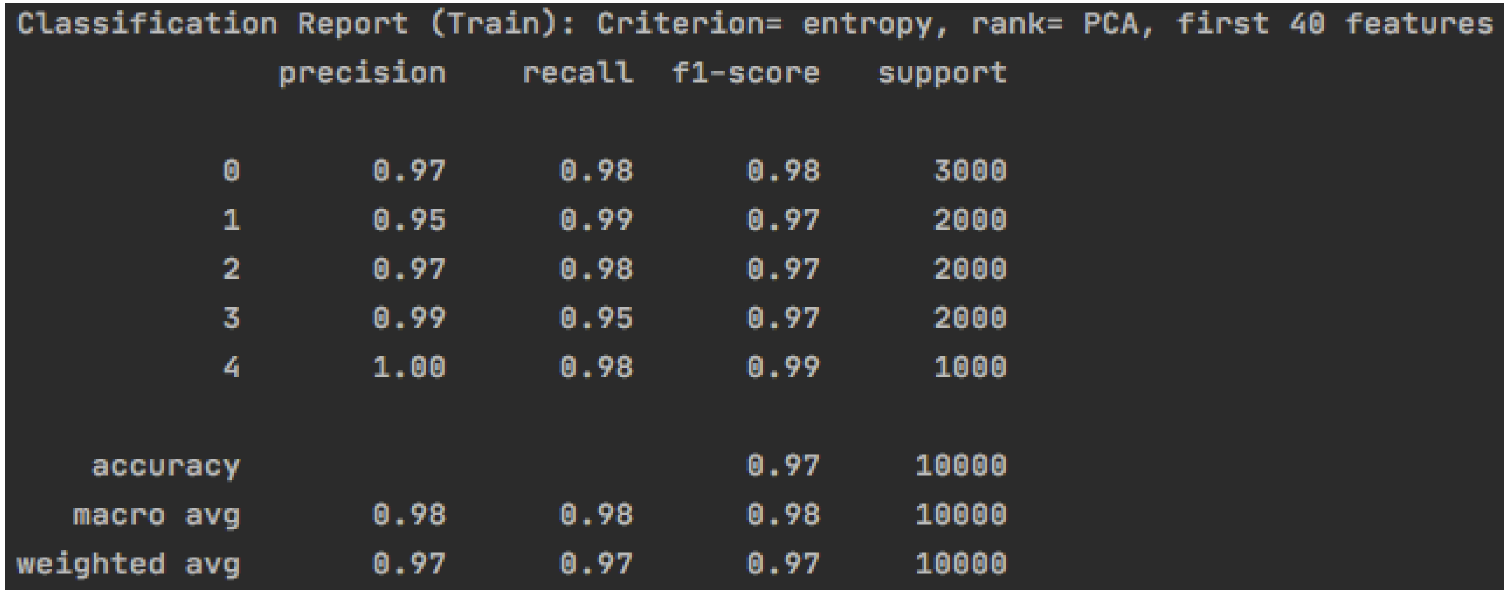
```
Classification Report (Train): Criterion= entropy, rank= IG, first 50 features
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      3000
           1       0.98      0.99      0.98      2000
           2       1.00      1.00      1.00      2000
           3       0.99      0.98      0.98      2000
           4       1.00      1.00      1.00      1000

    accuracy                           0.99     10000
   macro avg       0.99      0.99      0.99     10000
weighted avg       0.99      0.99      0.99     10000
```

Figure 11: Classification report relating to the decision tree trained on the dataset sorted according to Info Gain

Best Tree Training (PCA Rank)

- **Best Configuration** :
    - Splitting Criterion: Entropy
    - Features Number: 40

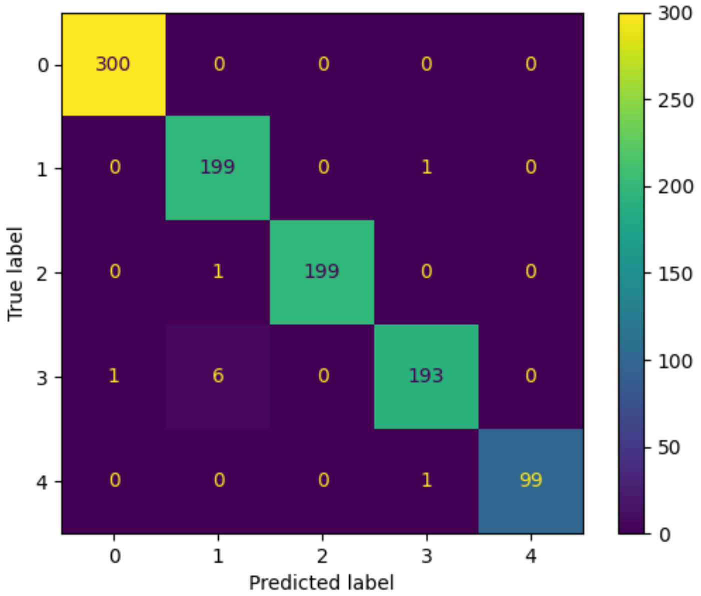
- **Details about the learned tree**:
    - Nodes number: 61
    - Leaf number: 31



```
Classification Report (Train): Criterion= entropy, rank= PCA, first 40 features
              precision    recall  f1-score   support

           0       0.97      0.98      0.98      3000
           1       0.95      0.99      0.97      2000
           2       0.97      0.98      0.97      2000
           3       0.99      0.95      0.97      2000
           4       1.00      0.98      0.99      1000

    accuracy                           0.97     10000
   macro avg       0.98      0.98      0.98     10000
weighted avg       0.97      0.97      0.97     10000
```

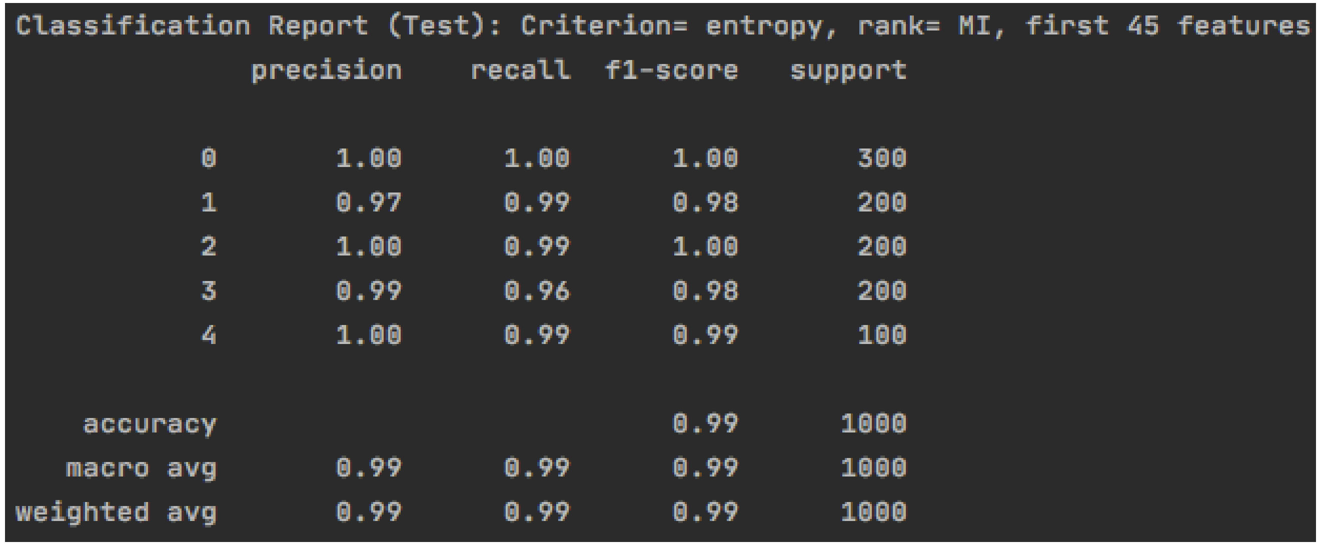
Figure 12: Classification report relating to the decision tree trained on the dataset sorted according to PCA

Best Tree Testing (MI Rank)



	0	1	2	3	4
0	300	0	0	0	0
1	0	199	0	1	0
2	0	1	199	0	0
3	1	6	0	193	0
4	0	0	0	1	99

Figure 13: Confusion matrix relating to the testing of the decision tree on the dataset sorted according to MI



```
Classification Report (Test): Criterion= entropy, rank= MI, first 45 features
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       300
           1       0.97      0.99      0.98       200
           2       1.00      0.99      1.00       200
           3       0.99      0.96      0.98       200
           4       1.00      0.99      0.99       100

    accuracy                           0.99      1000
   macro avg       0.99      0.99      0.99      1000
weighted avg       0.99      0.99      0.99      1000
```

Figure 14: Classification report relating to the decision tree tested on the dataset sorted according to MI

Best Tree Testing (IG Rank)

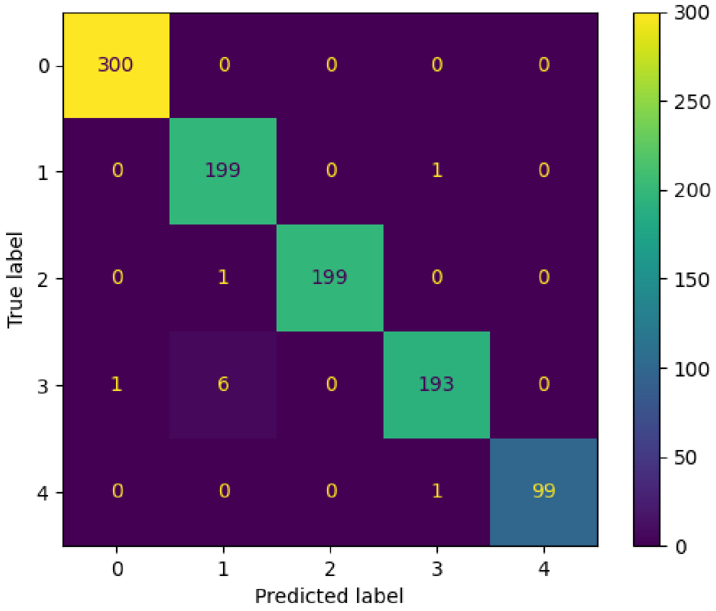
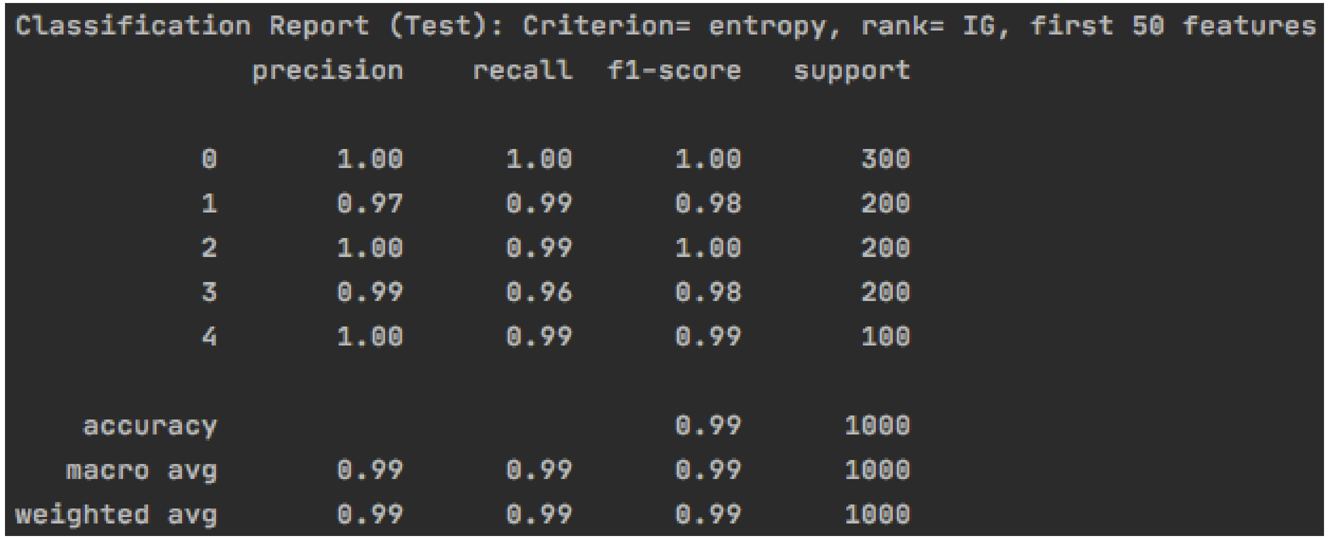


Figure 15: Confusion matrix relating to decision tree tested on the dataset sorted according to IG



```
Classification Report (Test): Criterion= entropy, rank= IG, first 50 features
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       300
           1       0.97      0.99      0.98       200
           2       1.00      0.99      1.00       200
           3       0.99      0.96      0.98       200
           4       1.00      0.99      0.99       100

    accuracy                           0.99      1000
   macro avg       0.99      0.99      0.99      1000
weighted avg       0.99      0.99      0.99      1000
```

Figure 16: Classification report relating to the decision tree tested on the dataset sorted according to IG

Best Tree Testing (PCA Rank)

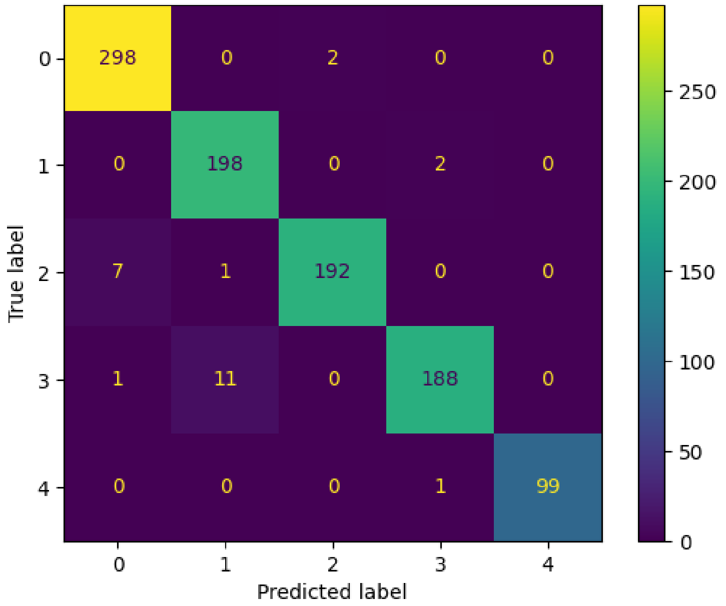
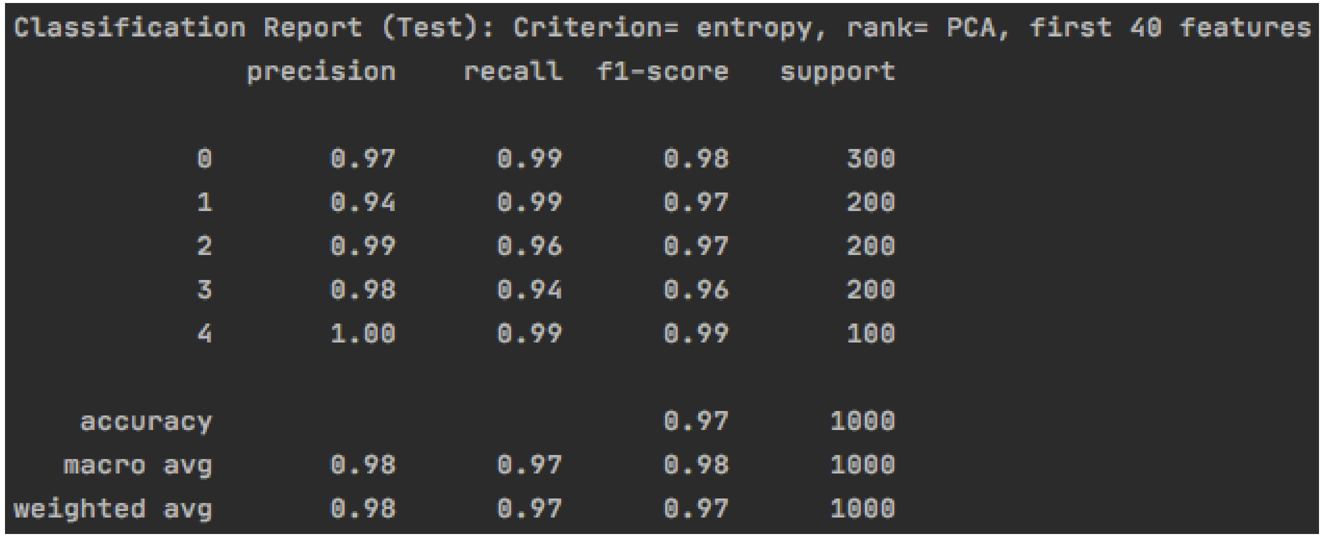


Figure 17: Confusion matrix relating to decision tree testing on the dataset sorted according to PCA

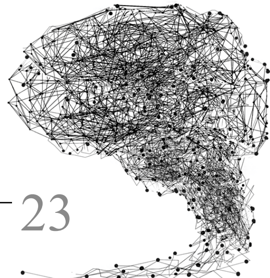


```
Classification Report (Test): Criterion= entropy, rank= PCA, first 40 features
              precision    recall  f1-score   support

           0       0.97      0.99      0.98       300
           1       0.94      0.99      0.97       200
           2       0.99      0.96      0.97       200
           3       0.98      0.94      0.96       200
           4       1.00      0.99      0.99       100

    accuracy                           0.97      1000
   macro avg       0.98      0.97      0.98      1000
weighted avg       0.98      0.97      0.97      1000
```

Figure 18: Classification report relating to the decision tree tested on the dataset sorted according to PCA

UNIVERSITÀ
DEGLI STUDI DI BARI
ALDO MORO

THANKS!



References

[1] https://www.unb.ca/cic/datasets/ddos-2019.html

[2] I. Sharafaldin, A. H. Lashkari, S. Hakak and A. A. Ghorbani, "Developing Realistic Distributed Denial of Service (DDoS) Attack Dataset and Taxonomy," 2019 International Carnahan Conference on Security Technology (ICCST), Chennai, India, 2019, pp. 1-8, doi: 10.1109/CCST.2019.8888419.

[3] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_classif.html

[4] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier

[5] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html