

Dipartimento di Informatica
Corso di Laurea Magistrale in Sicurezza Informatica
Analisi dei Dati per la Sicurezza



UNIVERSITÀ
DEGLI STUDI DI BARI
ALDO MORO

Sintesi di una Pipeline KDD per la classificazione di dati relativi a connessioni DDoS

Supervisore:
Prof.ssa Appice Annalisa

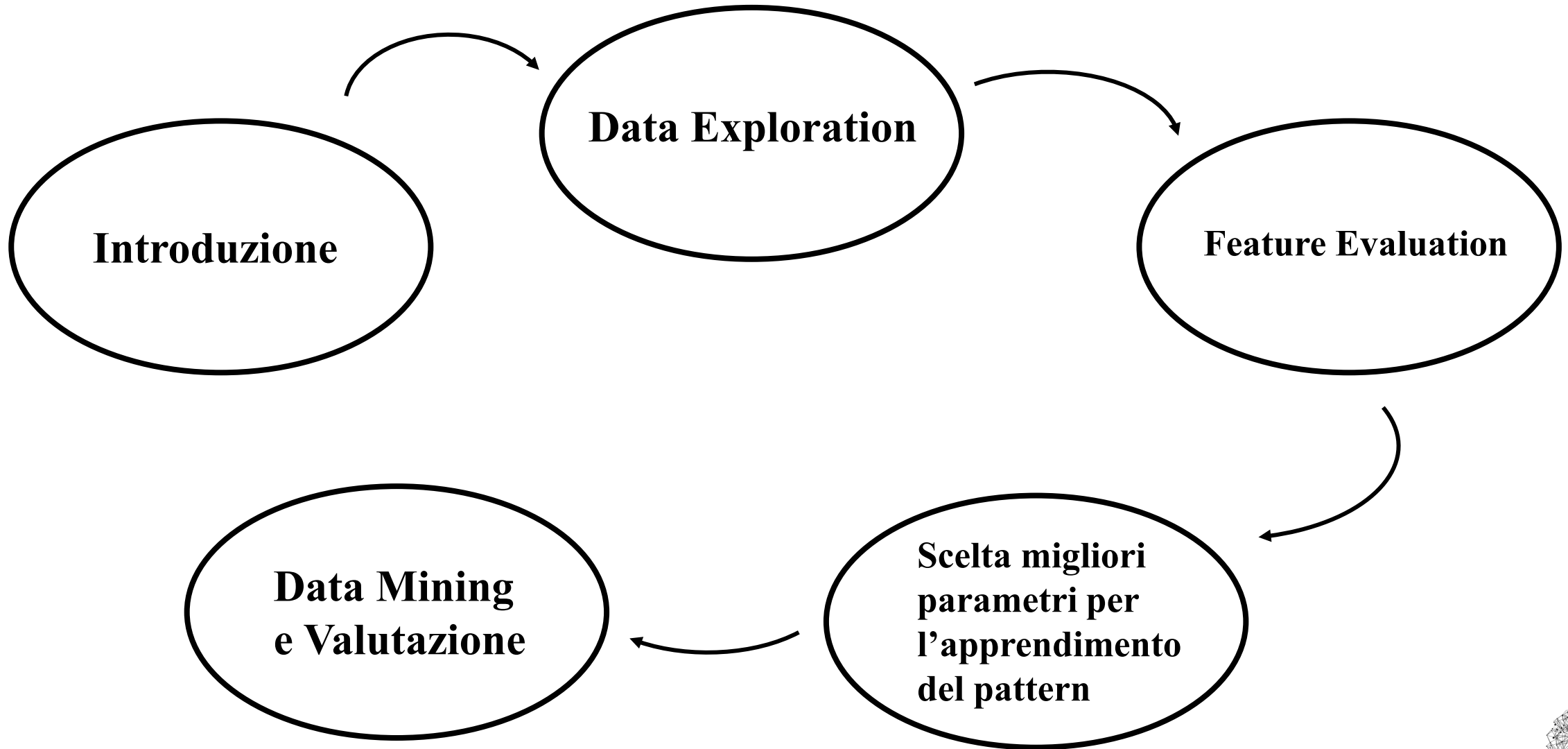
Studente:
del Vescovo Samuele
Matr.: 766196

Aprile 2023

A.A. 2021/2022



Outline



Introduzione

- L'attacco **DDoS** (Distributed Denial of Service) mira ad **esaurire** le **risorse computazionali** di un host target (in una rete di computers) in modo tale da rendere un servizio indisponibile [1].
- Come testimoniano Sharafaldin et al., diventa sempre più importante disporre di pattern in grado di individuare correttamente ed automaticamente connessioni potenzialmente veicolo di attacchi DDoS [2].
- L'**obiettivo** del lavoro proposto è sintetizzare una **pipeline KDD**, sfruttante algoritmi di **apprendimento automatico supervisionato**, al fine di **classificare connessioni** nelle diverse classi di attacchi DDoS (nel caso in esame sono “BENIGN”, “MSSQL”, “Syn”, “UDP”, “NetBIOS”).
- Il **dataset** scelto deriva da una **semplificazione** di quello proposto dal Canadian Institute for Cybersecurity, date le modeste risorse computazionali a disposizione [1] .



Dataset Iniziale

- Il **dataset di addestramento**, nella sua forma iniziale consiste in 10000 esempi descritti da 79 attributi (78 indipendenti e 1 dipendente) come testimonia l'uso della funzione “preElaborationClass”.
- Il **dataset di testing** è stato reperito dallo stesso benchmark e consiste in 2000 esempi descritti da 79 attributi (78 indipendenti e 1 dipendente)

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

Figura 1: Numero di esempi nel training set con numero di attributi

Label	
0	3000
1	2000
2	2000
3	2000
4	1000

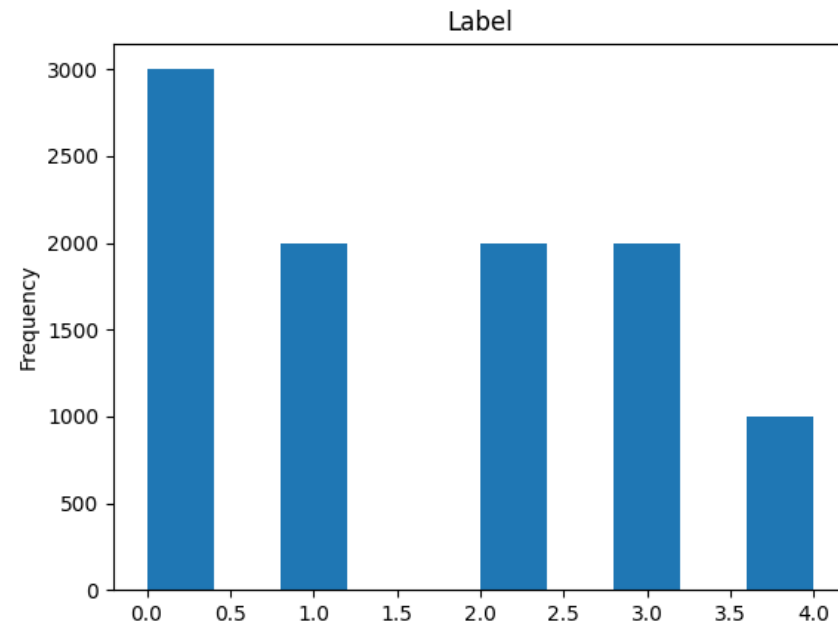


Figura 2: Distribuzione degli esempi di diverse classi

Data Exploration (1/6)

- Tramite apposite funzioni messe a disposizione dal framework “Pandas”, è possibile non solo **osservare** alcuni **parametri chiave** relativi alla **distribuzione** seguita da ogni **variabile indipendente** ma anche elaborare un **boxplot** che riassume tali informazioni.
- Ciò è utile al fine di **rilevare attributi più significativi** rispetto ad altri e **discriminanti** nei confronti della variabile dipendente (target).
- Di seguito verranno mostrati alcuni esempi di questi e confrontati con l’output delle tecniche di Features Evaluation.



Data Exploration (2/6)

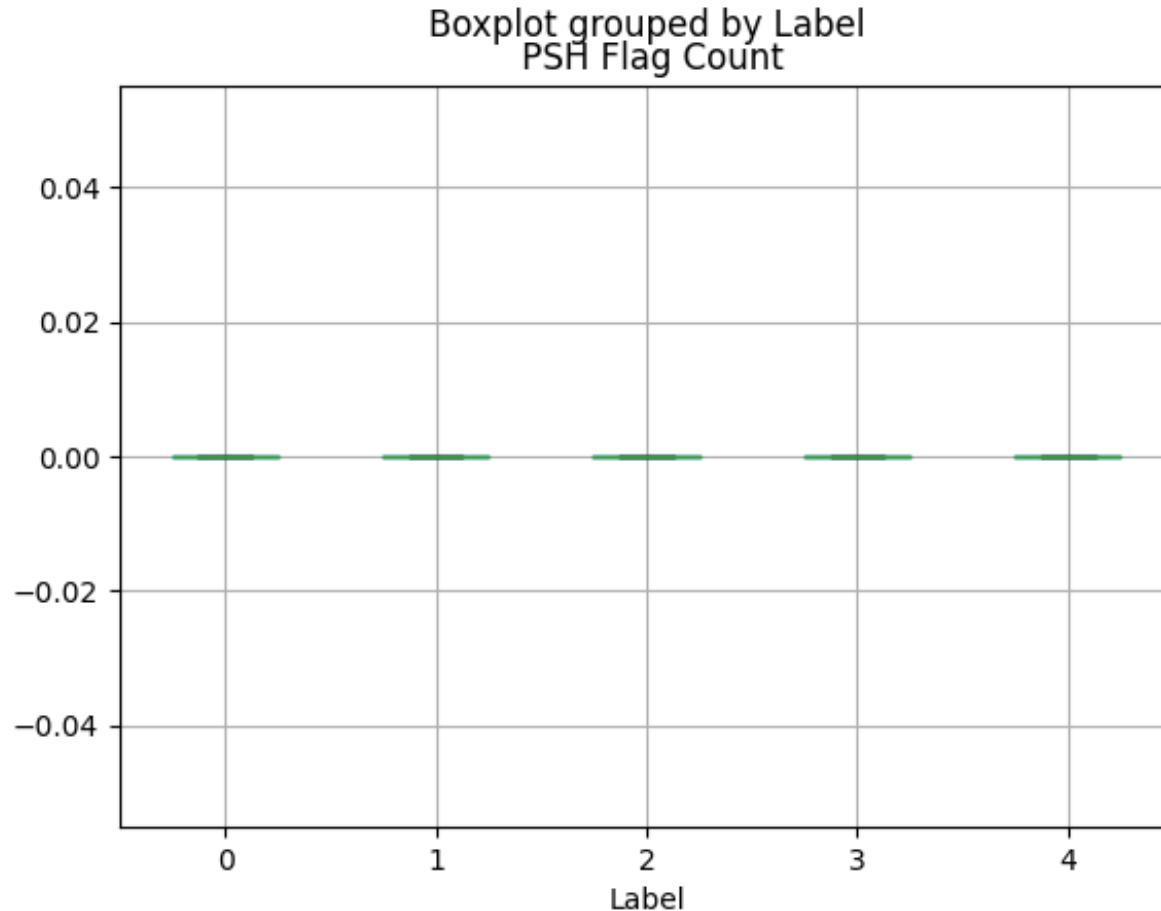


Figura 3: Esempio di attributo che presenta il valore massimo uguale al minimo

- L'attributo “PSH Flag Count” presenta il **valore massimo e minimo coincidenti** rispetto a tutte le classi.
- Per cui tale attributo è **inutile** ai fini del task di data mining e può essere **rimosso**.
- Tale caratteristica è osservata anche in altri 11 attributi.
- Inseguito a tale fase di scrematura, gli **attributi** in considerazione **saranno 66** (esclusa la label).



Data Exploration (3/6)

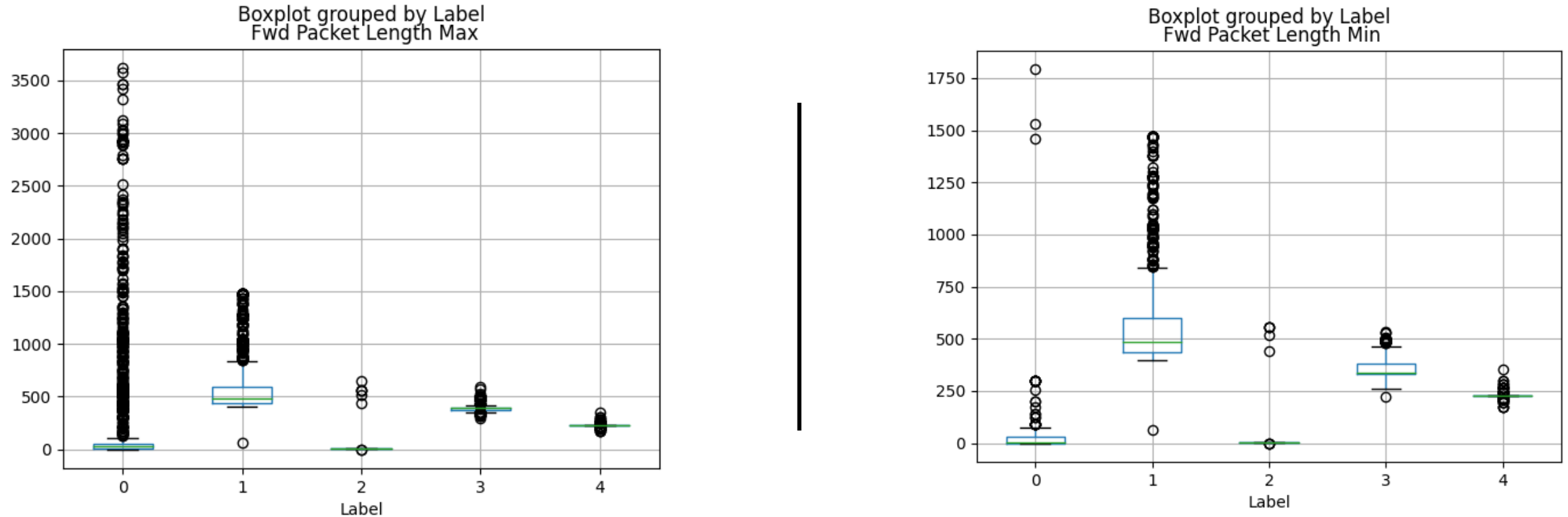


Figura 4: Esempio di attributi rappresentante di un trend

- Gli attributi “Fwd Packet Length Max” e “Fwd Packet Length Min” presentano un’ **ampiezza maggiore del terzo quartile** in riferimento alla **classe “1”** rispetto a tutte le altre classi (trend).
- Ci si aspetta che tali **attributi** siano nelle **posizioni intermedie** (tendenti alle prime) del **rank** prodotto dalle tecniche di **Features Evaluation**.



Data Exploration (4/6)

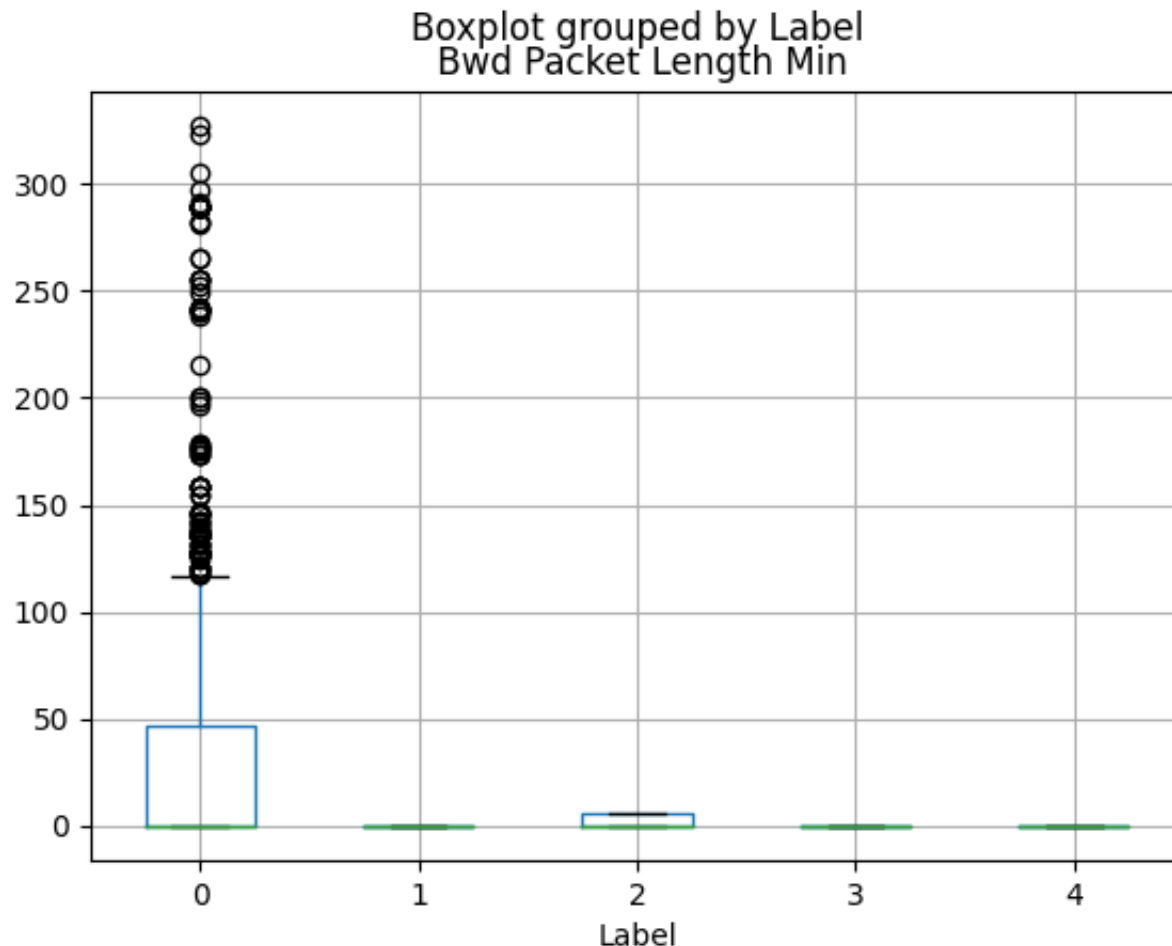


Figura 5: Esempio di attributo ritenuto poco discriminante

- L'attributo “**Bwd Packet Length Min**” presenta **box** molto **simili** tra le 4 **classi** di attacchi DDoS.
- Per cui ci si aspetta che tale **attributo** sia nelle **ultime posizioni** del rank prodotto da tecniche di Features Evaluation.



Data Exploration (5/6)

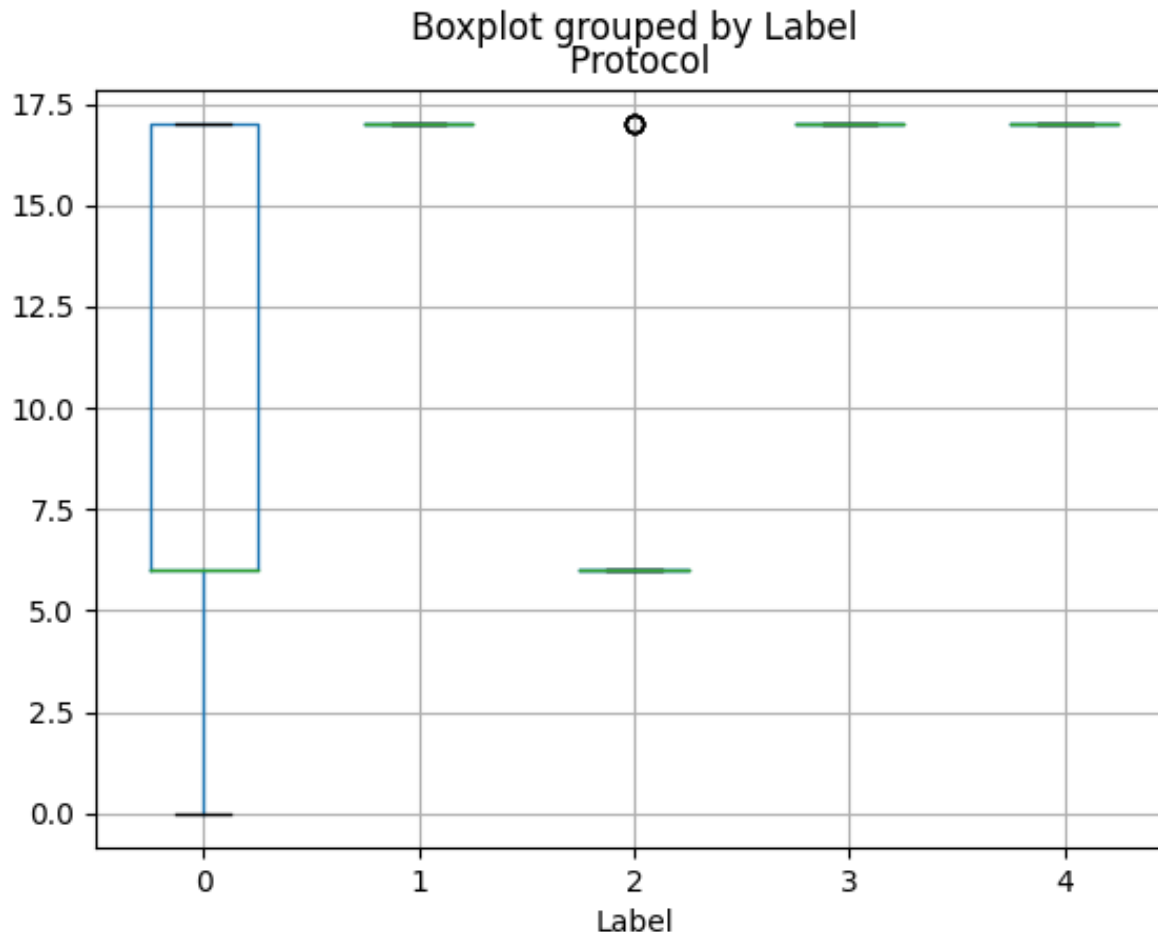
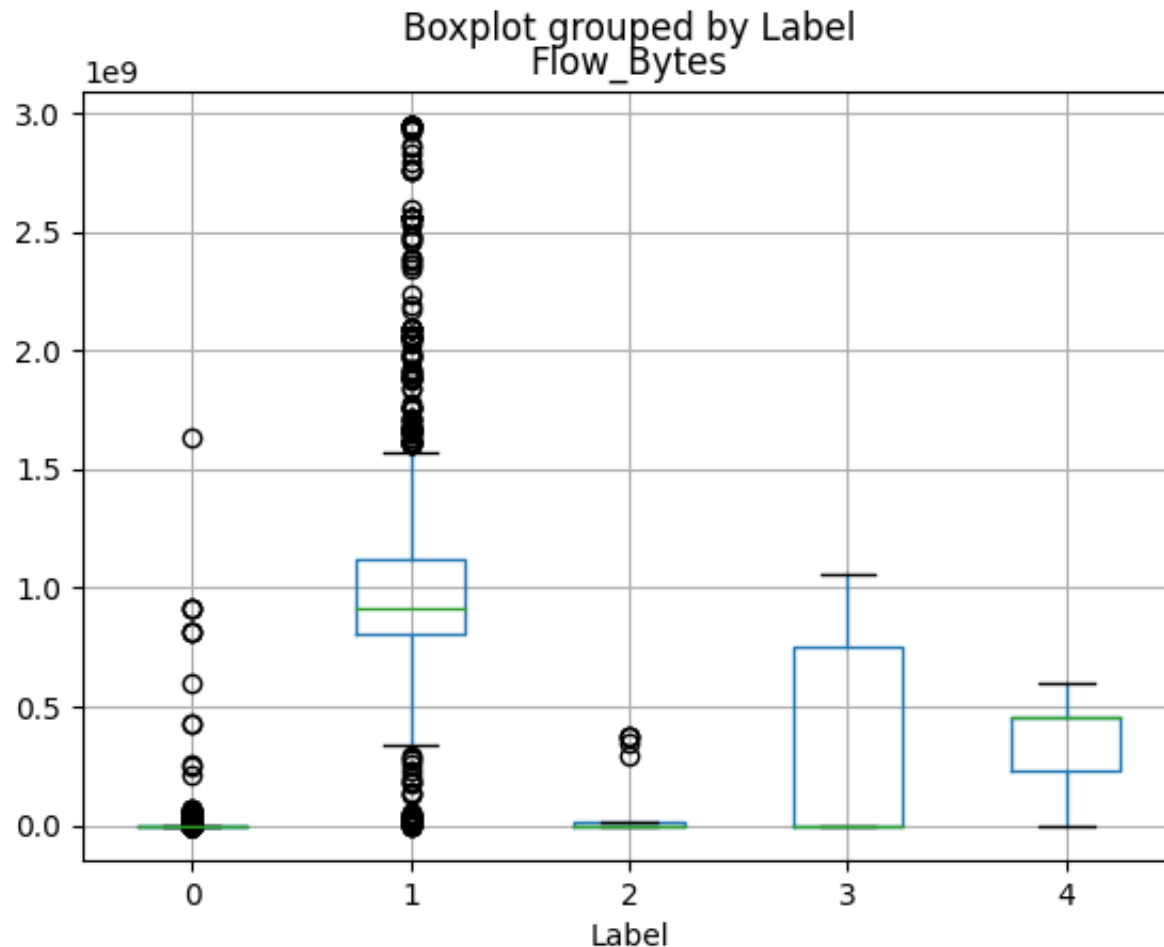


Figura 6: Esempio di attributo ritenuto mediamente discriminante

- L'attributo “**Protocol**” presenta **box simili** tra le **classi “1”, “3”, “4”** ma **quello** relativo alla **classe “2”** è **diverso** dalle altre.
- Inoltre, tale attributo presenta **box diversi** relativi alla **classe “0”** e **“2”** (in riferimento all'ampiezza del terzo quartile).
- Per cui ci si aspetta che tale **attributo** sia nelle **posizioni intermedie** del **rank** prodotto da tecniche di Features Evaluation.



Data Exploration (6/6)



- L'attributo “**Flow_Bytes**” presenta **box nettamente diversi** tra le **classi “1”, “3”, “4”** in riferimento ai valori di mediana e ampiezza dei quartili.
- Inoltre, i **box** relativi alle **classi “0” e “2”** sono **simili**.
- Per cui ci si aspetta che tale **attributo** sia nelle **prime posizioni** del **rank** prodotto da tecniche di Features Evaluation.

Figura 7: Esempio di attributo ritenuto altamente discriminante



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', 0.333040351732496), (' 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', 0.08713255224823957), (' 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)

random_state=42

Seed di numpy = 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', 0.18970172507684047), ('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', 0.04641642025544768), ('Active Std', 0.04576785797008798), ('Fwd PSH Flags', 0.04420173735039301), ('RST Flag Count', 0.04420173735039301), ('SYN Flag Count', 0.0007796186519307691)

MI Rank \neq IG Rank



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

- Si è provato ad applicare **un'algoritmo di scaling** (**MinMaxScaling** dove il nuovo minimo è “0” ed il nuovo massimo è ”1”) al dataset di addestramento ed a rivalutare le features tramite Mutual Info.
- Osservando la documentazione [3], è presente un parametro “**random_state**” che rappresenta il **seed** di un algoritmo **generatore di numeri pseudocasuali**; perciò (come accennato in precedenza) tale parametro è stato impostato ad un **valore costante** così come il **seed di numpy**.
- Detto ciò ci si aspetta che il ranking delle features in riferimento al dataset non scalato e quello in riferimento al dataset scalato **siano uguali** ma in realtà **sono diversi**, benchè tale operazione di scaling non modifica la distribuzione dei dati.
- Nella slide successiva, gli attributi etichettati con il colore **azzurro** si trovano in posizioni “invertite” nei due rank mentre quelli etichettati con il colore **arancione** si trovano in posizioni diverse nei due rank (ma in coppia)



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', 0.33184119086462927), (' 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]

Figura 8: Dataset proiettato lungo le componenti principali (l'ordine di queste è implicito)

- Il modello appreso per eseguire la pca verrà utilizzato anche nella fase di testing



Come scegliere la migliore configurazione?

```
1  Inizializza 'best_configuration_gini' ad una lista vuota di coppie <numero attributi, f1>
2  Inizializza 'best_configuration_entropy' ad una lista vuota di coppie <numero attributi, f1>
3
4  Per ogni criterio tra 'gini' ed 'entropy':
5
6      Inizializza 'list_number_feature_mean_f1' ad una lista vuota di coppie <numero attributi, f1>
7      Per ogni configurazione di feature F fino a 65 al passo di 5:
8          Esegui la Stratified 5-Fold Cross Validation ed aggiungi in 'list_number_feature_mean_f1'
9          la coppia <F, f1> dove f1 deriva dalla media delle f1_measure ottenute nelle 5 "trial"
10         della CV
11
12         Esegui la Stratified 5-Fold Cross Validation ed aggiungi in 'list_number_feature_mean_f1'
13         la coppia <F, f1> dove f1 deriva dalla media delle f1_measure ottenute nelle 5 "trial"
14         della CV ed F corrisponde alle 66 feature
15
16         Se il criterio è 'gini':
17             poni in 'best_configuration_gini' la coppia in 'list_number_feature_mean_f1' che presenta il valore di f1 massimo
18             altrimenti:
19             poni in 'best_configuration_entropy' la coppia in 'list_number_feature_mean_f1' che presenta il valore di f1 massimo
```

Figura 9: Pseudocodice relativo alla funzione utile a scegliere la migliore configurazione del decision tree per il train

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

Seed di numpy=42

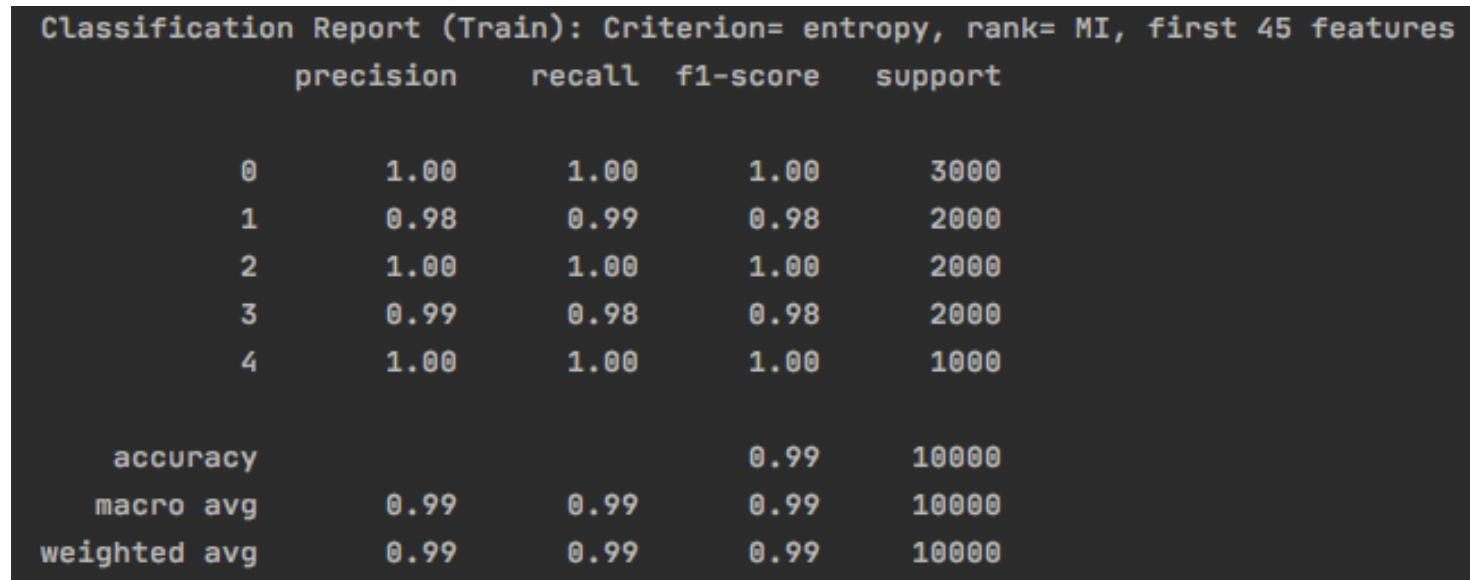
Parametri Decision Tree [4] → splitter="best", random_state=0, min_samples_split=500

Tipo f1-measure → weighted



Allenamento Miglior albero (MI Rank)

- **Migliore configurazione:**
 - Criterio di splitting: Entropy
 - Numero di attributi: 45
- **Sono presenti quegli attributi?**
 - Sono presenti tutti gli attributi esaminati tranne “Bwd Packet Length Min” (ritenuto pessimo)
- **Informazioni sull'albero:**
 - Numero di nodi: 47
 - Numero di foglie: 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

Figura 10: Classification report relativo al train del decision tree sul dataset ordinato secondo Mutual Info



Allenamento Miglior albero (IG Rank)

- **Migliore configurazione:**
 - Criterio di splitting: Entropy
 - Numero di attributi: 50
- **Sono presenti quegli attributi?**
 - Sono presenti tutti gli attributi
- **Informazioni sull'albero:**
 - Numero di nodi: 49
 - Numero di foglie: 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

Figura 11: Classification report relativo al train del decision tree sul dataset ordinato secondo Info Gain



Allenamento Miglior albero (PCA Rank)

- **Migliore configurazione:**
 - Criterio di splitting: Entropy
 - Numero di attributi: 40
- **Informazioni sull'albero:**
 - Numero di nodi: 61
 - Numero di foglie: 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

Figura 12: Classification report relativo al train del decision tree sul dataset ordinato secondo PCA



Testing Miglior albero (MI Rank)

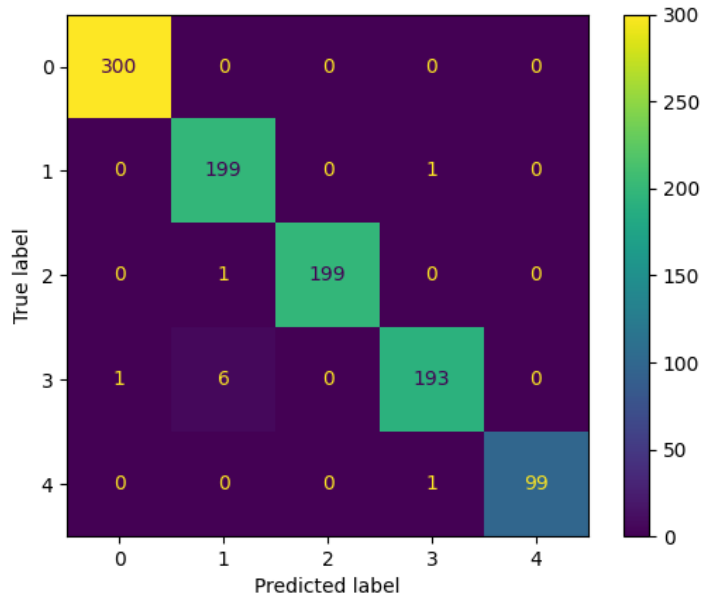


Figura 13: Matrice di confusione relativa al testing del decision tree sul dataset ordinato secondo MI

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

Figura 14: Classification report relativo al test del decision tree sul dataset ordinato secondo MI



Testing Miglior albero (IG Rank)

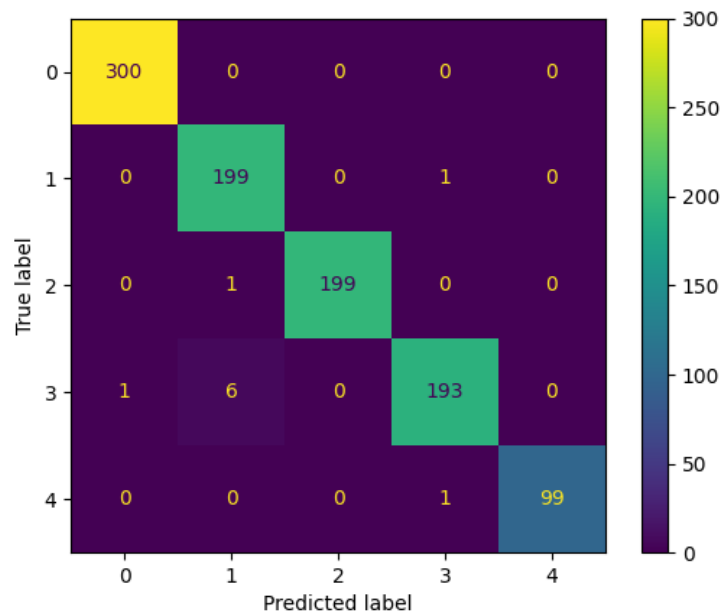


Figura 15: Matrice di confusione relativa al testing del decision tree sul dataset ordinato secondo IG

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

Figura 16: Classification report relativo al test del decision tree sul dataset ordinato secondo IG



Testing Miglior albero (PCA Rank)

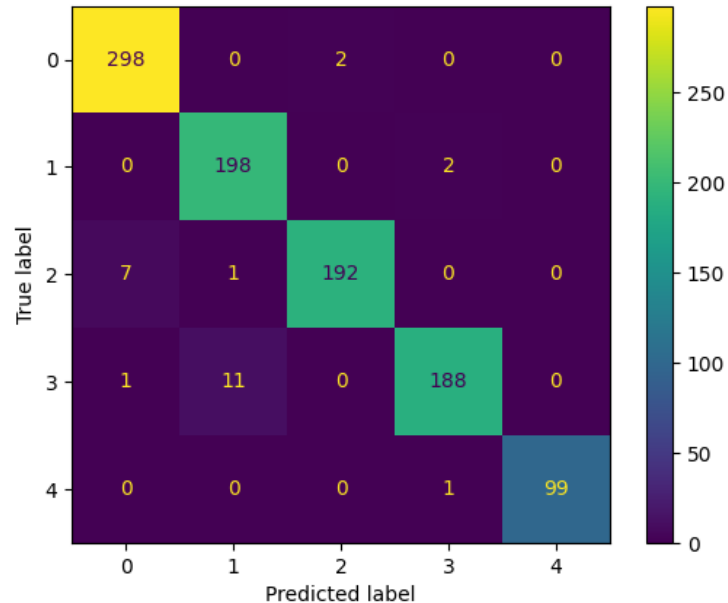


Figura 17: Matrice di confusione relativa al testing del decision tree sul dataset ordinato secondo PCA

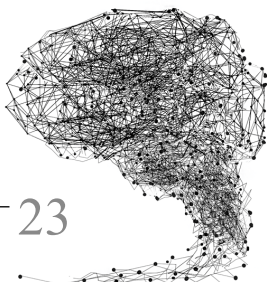
Classification Report (Test): Criterion= entropy, rank= PCA, first 40 features				
	precision	recall	f1-score	support
0	0.97	0.99	0.98	300
1	0.94	0.99	0.97	200
2	0.99	0.96	0.97	200
3	0.98	0.94	0.96	200
4	1.00	0.99	0.99	100
accuracy			0.97	1000
macro avg	0.98	0.97	0.98	1000
weighted avg	0.98	0.97	0.97	1000

Figura 18: Classification report relativo al test del decision tree sul dataset ordinato secondo PCA



UNIVERSITÀ
DEGLI STUDI DI BARI
ALDO MORO

GRAZIE!



Referenze

[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

