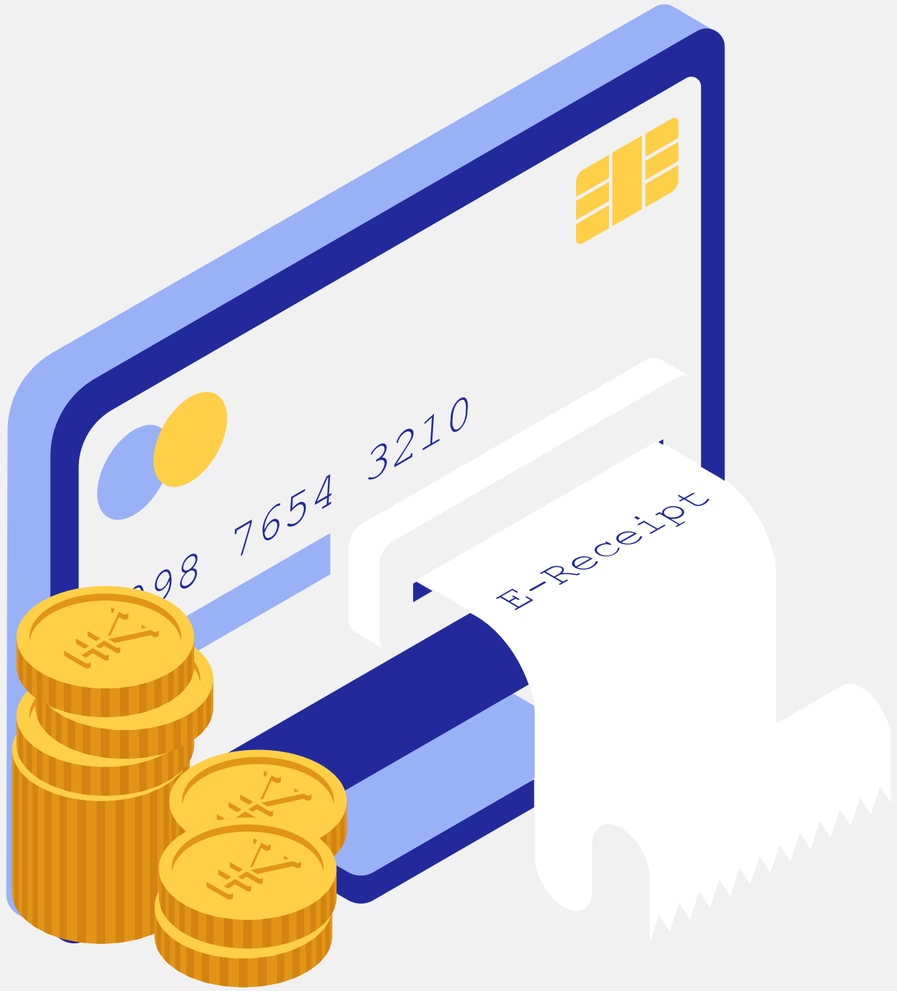


Christian Soriani Mat. n°764581

Progetto analisi dei dati per la sicurezza

A.A. 2021/2022

OBIETTIVI

Punto 1 Sviluppare una filiera KDD

Tramite l'uso di Python e SKLearn

Punto 2 Costruire un Albero Decisionale

Trovare la miglior configurazione

Punto 3 Valutare il pattern

Predizione del testing set

Punto 4 Generare le relative Matrici di confusioni

Verificare l'accuratteza del pattern scoperto

Punto 5 Stampare le metriche

Calcolo delle metriche per l'accuratezza

Data Set

Training Set ("trainDdosLabelNumeric.csv)

Data set contente valori collezionati dal Canadian Institute for CyberSecurity (2019). Attacchi che possono essere effettuati utilizzando protocolli basati su TCP/UDP.

Testing Set ("testDdosLabelNumeric.csv")

Data set utilizzato per la fase di valutazione.
(1000 x 79)

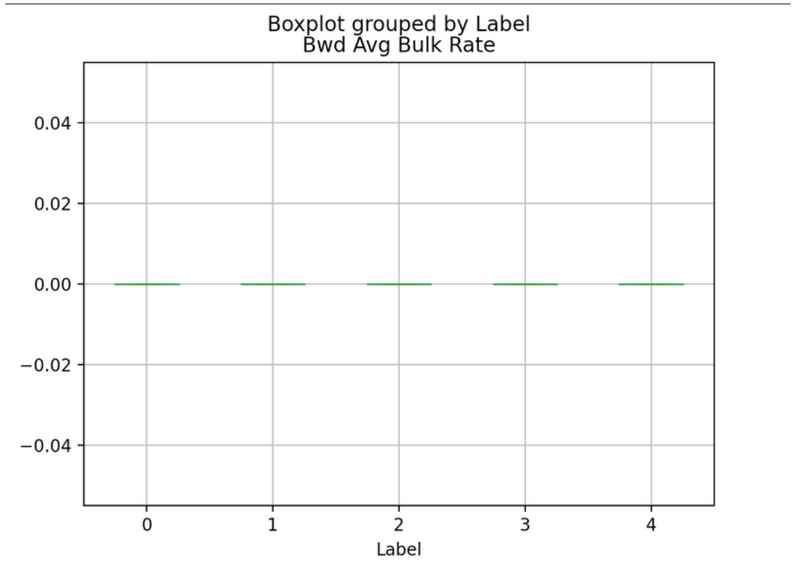
```
________SHAPE:________
(10000, 79)



________HEAD:________
    Protocol   Flow Duration   Total Fwd Packets  ...   Idle Max   Idle Min  Label
0          6              1                   2  ...        0.0        0.0      0
1          6          71271                   6  ...        0.0        0.0      0
2          6              2                   1  ...        0.0        0.0      0
3          6              1                   2  ...        0.0        0.0      0
4         17          20623                   2  ...        0.0        0.0      0

[5 rows x 79 columns]



________COLUMNS:________
Index([' Protocol', ' Flow Duration', ' Total Fwd Packets',
       ' Total Backward Packets', 'Total Length of Fwd Packets',
       ' Total Length of Bwd Packets', ' Fwd Packet Length Max',
       ' Fwd Packet Length Min', ' Fwd Packet Length Mean',
       ' Fwd Packet Length Std', 'Bwd Packet Length Max',
       ' Bwd Packet Length Min', ' Bwd Packet Length Mean',
       ' Bwd Packet Length Std', 'Flow_Bytes', 'Flow_Packets',
       ' Flow IAT Mean', ' Flow IAT Std', ' Flow IAT Max', ' Flow IAT Min',
       'Fwd IAT Total', ' Fwd IAT Mean', ' Fwd IAT Std', ' Fwd IAT Max',
       ' Fwd IAT Min', 'Bwd IAT Total', ' Bwd IAT Mean', ' Bwd IAT Std',
       ' Bwd IAT Max', ' Bwd IAT Min', 'Fwd PSH Flags', ' Bwd PSH Flags',
       ' Fwd URG Flags', ' Bwd URG Flags', ' Fwd Header Length',
       ' Bwd Header Length', 'Fwd_Packets', 'Bwd_Packets',
       ' Min Packet Length', ' Max Packet Length', ' Packet Length Mean',
       ' Packet Length Std', ' Packet Length Variance', 'FIN Flag Count',
       ' SYN Flag Count', ' RST Flag Count', ' PSH Flag Count',
       ' ACK Flag Count', ' URG Flag Count', ' CWE Flag Count',
       ' ECE Flag Count', ' Down/Up Ratio', ' Average Packet Size',
       ' Avg Fwd Segment Size', ' Avg Bwd Segment Size',
       ' Fwd Header Length.1', 'Fwd Avg Bytes/Bulk', ' Fwd Avg Packets/Bulk',
       ' Fwd Avg Bulk Rate', ' Bwd Avg Bytes/Bulk', ' Bwd Avg Packets/Bulk',
       'Bwd Avg Bulk Rate', 'Subflow Fwd Packets', ' Subflow Fwd Bytes',
       ' Subflow Bwd Packets', ' Subflow Bwd Bytes', 'Init_Win_bytes_forward',
       ' Init_Win_bytes_backward', ' act_data_pkt_fwd',
       ' min_seg_size_forward', 'Active Mean', ' Active Std', ' Active Max',
       ' Active Min', 'Idle Mean', ' Idle Std', ' Idle Max', ' Idle Min',
       'Label'],
      dtype='object')
```

Analisi del data set

- Individuazione degli attributi non utili alla classificazione
- Assenza di valori mancanti
- Valori numerici

```
count    10000.0
mean         0.0
std          0.0
min          0.0
25%          0.0
50%          0.0
75%          0.0
max          0.0
Name:  Fwd Avg Bulk Rate, dtype: float64
```

```
count    10000.0
mean         0.0
std          0.0
min          0.0
25%          0.0
50%          0.0
75%          0.0
max          0.0
Name:  Bwd Avg Bytes/Bulk, dtype: float64
```



Analisi del data set

- Utilizzo di tecniche quali:
  - Mutual Info
  - Info Gain
  - Pca

Abbiamo realizzato una classifica delle migliori feature in maniera decrescente.
Successivamente si è scelto di estrarre le prime N(10) feature, ovverro quelle più utili.

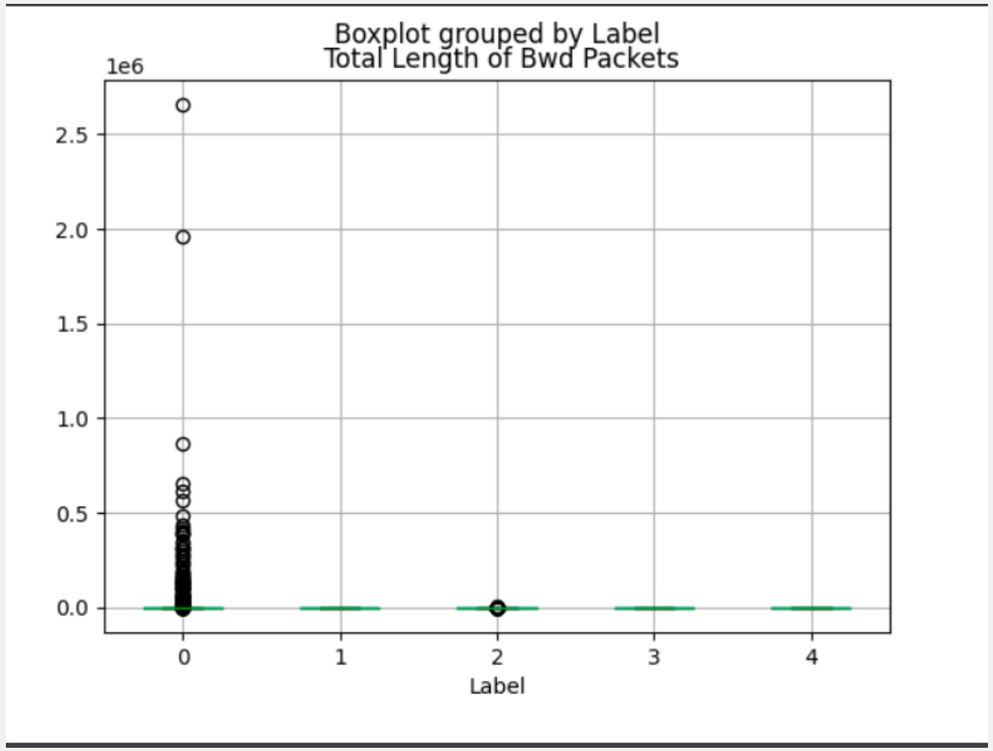
```
Index([' Average Packet Size', 'Total Length of Fwd Packets',
       ' Subflow Fwd Bytes', ' Avg Fwd Segment Size',
       ' Fwd Packet Length Mean', 'Flow_Bytes', ' Max Packet
       ' Min Packet Length', ' Packet Length Mean', ' Fwd Pac
       'Label'],
      dtype='object')
```

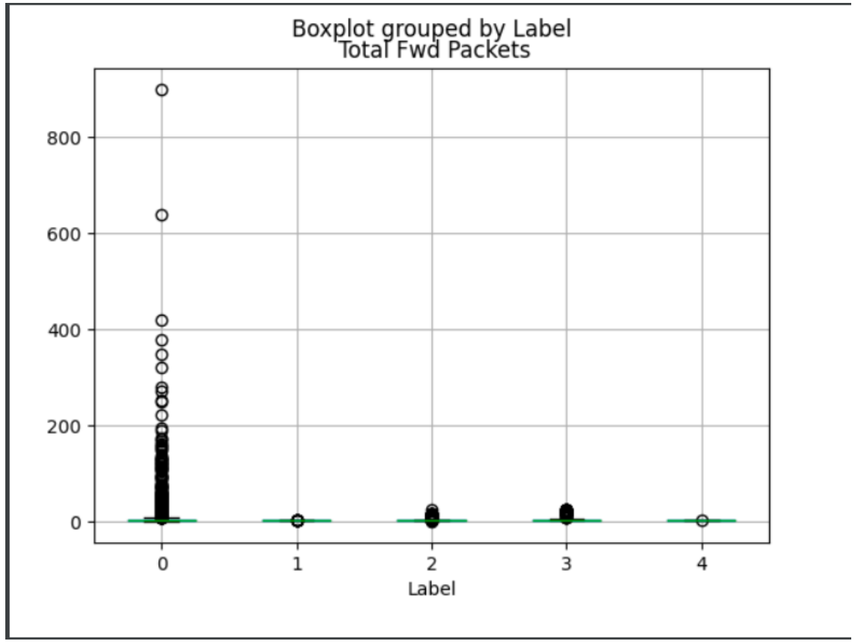
```
________PCA SELECTED________
              pc_1          pc_2  ...          pc_10  Label
0    -1.575375e+08  2.465755e+08  ... -180882.032955      0
1    -1.575230e+08  2.465192e+08  ... -147643.436348      0
2    -1.521068e+08  2.257800e+08  ... -165508.408246      0
3    -1.481905e+08  2.107832e+08  ... -174550.281200      0
4    -1.575363e+08  2.465711e+08  ... -150408.358560      0
            ...           ...  ...            ...    ...
9995 -4.183706e+07 -1.964745e+08  ... -102515.034721      4
9996 -9.968794e+07  2.505316e+07  ... -129899.422680      4
9997 -1.550770e+08  2.371541e+08  ... -156076.711768      4
9998 -1.551775e+08  2.375389e+08  ... -156122.377995      4
9999 -1.551283e+08  2.373505e+08  ... -156100.040043      4
```

```
rankingFeatureInfoGain = {list: 66} [('Flow_Bytes', 0.940305030995
 00 = {tuple: 2} ('Flow_Bytes', 0.9403050309957603)
 01 = {tuple: 2} (' Average Packet Size', 0.9102146306170038)
 02 = {tuple: 2} ('Total Length of Fwd Packets', 0.89816290063753
 03 = {tuple: 2} (' Subflow Fwd Bytes', 0.8981629006375397)
 04 = {tuple: 2} (' Packet Length Mean', 0.8796498907935585)
 05 = {tuple: 2} (' Fwd Packet Length Mean', 0.873590279058078
 06 = {tuple: 2} (' Avg Fwd Segment Size', 0.8735902790580787)
 07 = {tuple: 2} (' Max Packet Length', 0.8607302363583886)
 08 = {tuple: 2} (' Fwd Packet Length Max', 0.849027481844839)
 09 = {tuple: 2} (' Min Packet Length', 0.8443248214760471)
 10 = {tuple: 2} (' Fwd Packet Length Min', 0.8435961794299786)
```

Analisi del data set

- Nel data set sono stati individuati anche diversi outlier.






```
count    1.000000e+04
mean     2.269414e+03
std      3.956776e+04
min      0.000000e+00
25%      0.000000e+00
50%      0.000000e+00
75%      1.200000e+01
max      2.655090e+06
Name:  Total Length of Bwd Packets, dtype: float64
```

Configurazione Decision Tree

Utilizzo delle funzioni presenti in SKLearn per l'addestramento di un albero decisionale

```
________BEST CONFIGURATION EVER________
[('entropy', 35, 0.9928053074262679, 'mi')]
```

Pre-Elaborazione / Trasformazione

Utilizzo delle tecniche di Mutual Info, Information Gain e PCA per costruire un ranking delle feature più utili alla fase di addestramento

Best Configuration

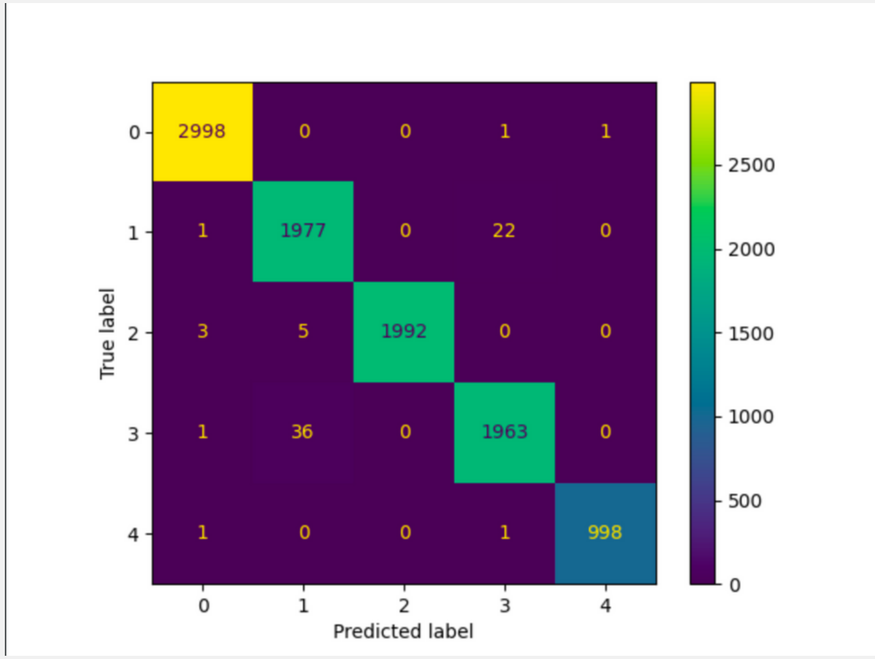
Per ogni configurazione ci si è focalizzati sulla ricerca del milgior criterio, del miglior numero di feature da impiegare nella fase di addestramento.
Per ogni configurazione è stata calcolato l'F1 Score, utilizzato successivamente per determinare il miglior approccio.

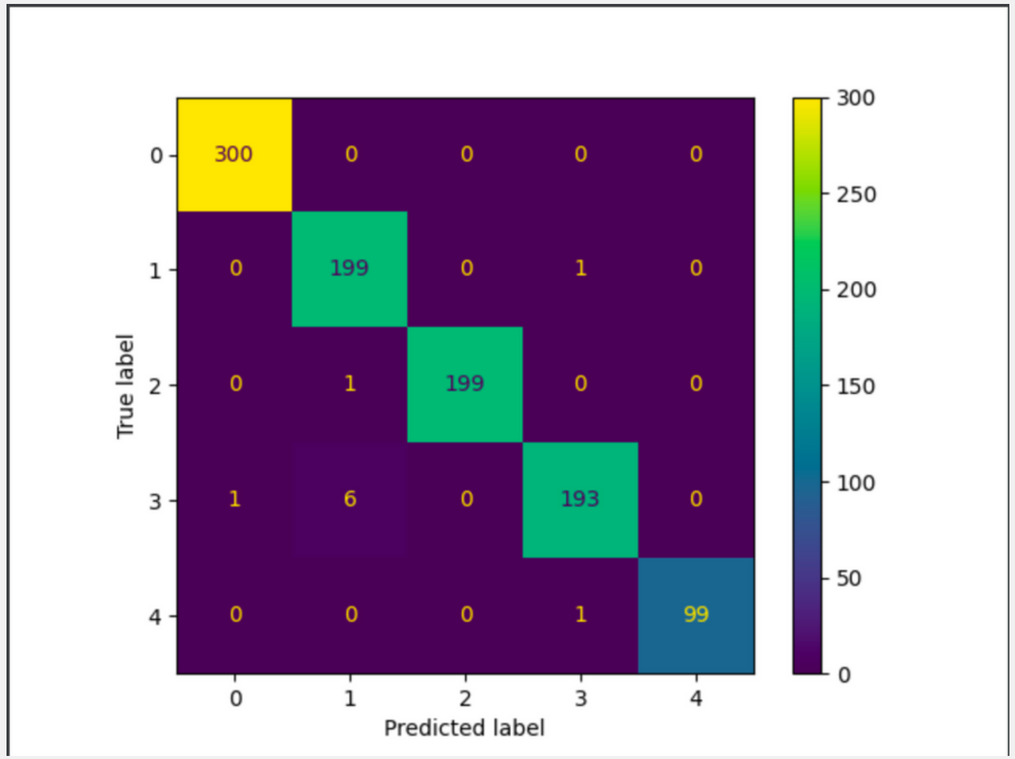
Scelta Del migliore

Prendiamo la configurazione con F1 Score maggiore. Se dovessero essere presenti più configurazioni con lo stesso score, tra queste si preferisce quella con numero di feature minore. Se anche in quest'ultimo caso dovessero figurare valori uguali estriamo casuale, sempre tra le stesse (minore)

Matrice di Confusione

Predisponiamo le matrici di confusione rispettivamente per il training set e per il testing set.
Notiamo che il nostro classificatore è affidabile.
Valori minimi nella parte superiore e inferiore.





Metriche di valutazione

SKLearn predispone funzioni per il calcolo delle apposite metriche di valutazione. Metriche calcolate sulla miglior configurazione

_____REPORT TRAINING_____

	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

_____REPORT TEST_____

	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

Grazie per l'attenzone