

Università degli Studi di Bari Aldo Moro
Laurea Magistrale in Sicurezza Informatica



Analisi dei Dati per la Sicurezza
Report Caso di Studio A.A. 2024/25

Prof.ssa Annalisa Appice

Dott. Matteo Esposito

Sommario

1. DATASET	5
2. PRESENZA DI ATTRIBUTI CON VALORI MANCANTI	6
3. ATTRIBUTI INUTILI.....	7
4. PROPRIETA' STATISTICHE DEGLI ATTRIBUTI UTILI	8
4.1 Protocol	8
4.2 Flow Duration	9
4.3 Total Fwd Packets	10
4.4 Total Backward Packets	11
4.5 Total Length of Fwd Packets	12
4.6 Total Length of Bwd Packets.....	13
4.7 Fwd Packet Length Max.....	14
4.8 Fwd Packet Length Min	15
4.9 Fwd Packet Length Mean	16
4.10 Fwd Packet Length Std	17
4.11 Bwd Packet Length Max	18
4.12 Bwd Packet Length Min	19
4.13 Bwd Packet Length Mean	20
4.14 Bwd Packet Length Std	21
4.15 Flow_Bytes	22
4.16 Flow_Packets.....	23
4.17 Flow IAT Mean.....	24
4.18 Flow IAT Std	25
4.19 Flow IAT Max	26
4.20 Flow IAT Min.....	27
4.21 Fwd IAT Total	28
4.22 Fwd IAT Mean.....	29
4.23 Fwd IAT Std	30
4.24 Fwd IAT Max	31
4.25 Fwd IAT Min.....	32
4.26 Bwd IAT Total	33
4.27 Bwd IAT Mean	34

4.28 Bwd IAT Std	35
4.29 Bwd IAT Max	36
4.30 Bwd IAT Min	37
4.31 Fwd PSH Flags.....	38
4.32 Fwd Header Length	32
4.33 Bwd Header Length.....	33
4.34 Fwd_Packets.....	34
4.35 Bwd_Packets	35
4.36 Min Packet Length.....	36
4.37 Max Packet Length	37
4.38 Packet Length Mean.....	38
4.39 Packet Length Std	39
4.40 Packet Length Variance	40
4.41 SYN Flag Count	41
4.42 RST Flag Count.....	42
4.43 ACK Flag Count	43
4.44 URG Flag Count	44
4.45 CWE Flag Count	45
4.46 Down/Up Ratio.....	46
4.47 Average Packet Size.....	47
4.48 Avg Fwd Segment Size.....	48
4.49 Avg Bwd Segment Size	49
4.50 Fwd Header Length.1	50
4.51 Subflow Fwd Packets.....	51
4.52 Subflow Fwd Bytes	52
4.53 Subflow Bwd Packets	53
4.54 Subflow Bwd Bytes.....	54
4.55 Init_Win_bytes_forward	55
4.56 Init_Win_bytes_backward	56
4.57 act_data_pkt_fwd	57
4.58 min_seg_size_forward	58
4.59 Active Mean	59

4.60 Active Std	60
4.61 Active Max.....	61
4.62 Active Min	62
4.63 Idle Mean	63
4.64 Idle Std	64
4.65 Idle Max.....	65
4.66 Idle Min	66
5. DISTRIBUZIONE DELLE CLASSI.....	67
6. FEATURE SELECTION.....	68
7. ADDESTRAMENTO DEI MODELLI	69
7.1 ALBERI DECISIONALI.....	69
7.2 ENSEMBLE SVM.....	69
7.3 Z-SCORE STANDARDIZATION	70
8. RISULTATI.....	71
8.1 STRUTTURA DEGLI ALBERI	71
8.2 CONFIGURAZIONE ENSEMBLE	74
8.3 MATRICI DI CONFUSIONE E CLASSIFICATION REPORT	75
9. CONCLUSIONI	79

1. DATASET

Un sottoinsieme di dati raccolti dal Canadian Institute for Cybersecurity nel 2019. Il dataset contiene attacchi che possono essere eseguiti utilizzando protocolli basati su TCP/UDP.

- Training set: 10.000 campioni
- Test Set: 1.000 campioni
- 78 attributi + 1 classe

	Mapping	#samples
BENIGN	0	3000
MSSQL	1	2000
Syn	2	2000
UDP	3	2000
NetBIOS	4	1000
TOT		10000

2. PRESENZA DI ATTRIBUTI CON VALORI MANCANTI

Utilizzo il seguente codice per verificare la presenza di valori mancanti all'interno del dataset.

```
#Controllo missing values
for col in data.columns:
    missing_count = data[col].isnull().sum()
    print(f"Colonna '{col}' n. {missing_count} missing values")
```

L'analisi evidenzia che non sono presenti valori nulli.

3. ATTRIBUTI INUTILI

Utilizzo il seguente codice per verificare la presenza di attributi inutili.

```
def removeColumns(df, columns):
    """
        Rimuove dal DataFrame le colonne il cui valore minimo è
        uguale al valore massimo (colonne costanti).

        Args:
            df (pd.DataFrame): Il DataFrame da cui rimuovere le
                colonne.
            columns (list): Lista di nomi di colonne da
                controllare.

        Returns:
            tuple:
                - pd.DataFrame: DataFrame con le colonne rimosse.
                - list: Lista delle colonne rimosse.
    """
    removedColumns = []

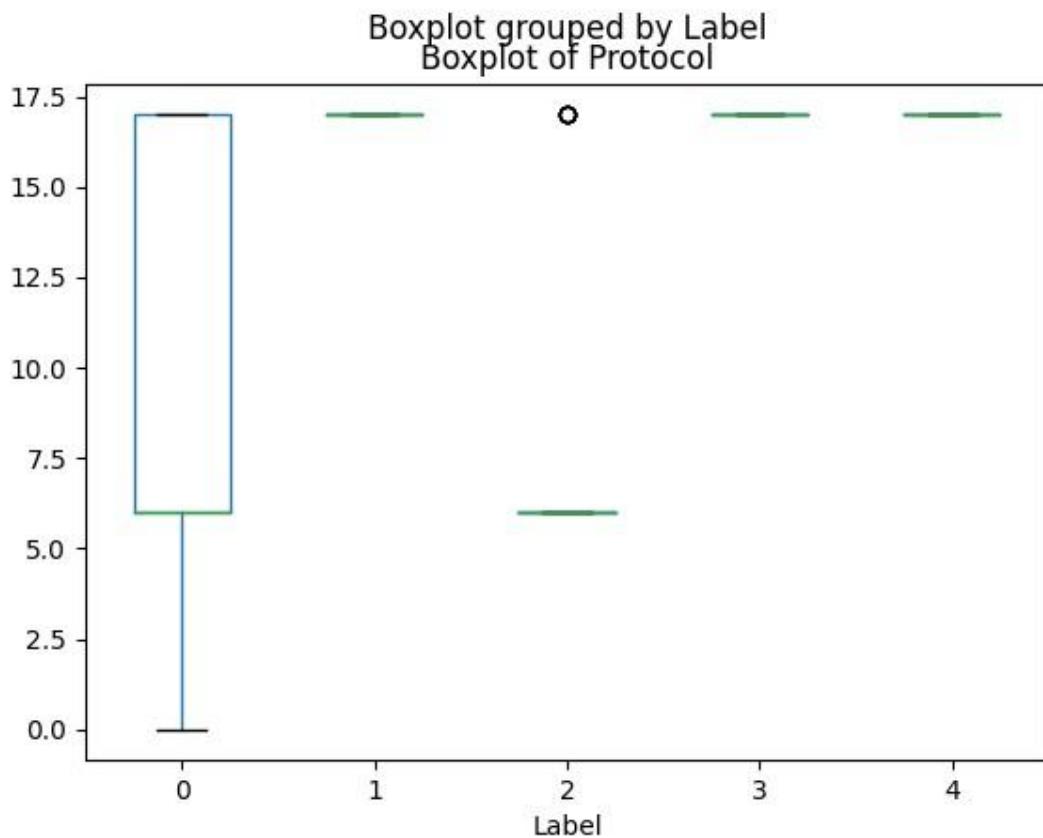
    for c in columns:
        if df[c].min() == df[c].max():
            removedColumns.append(c)

    df.drop(columns=removedColumns, inplace=True)
    return df, removedColumns
```

Colonne dove il valore minimo è uguale al valore massimo: ' Bwd PSH Flags', ' Fwd URG Flags', ' Bwd URG Flags', 'FIN Flag Count', ' PSH Flag Count', ' ECE Flag Count', 'Fwd Avg Bytes/Bulk', ' Fwd Avg Packets/Bulk', ' Fwd Avg Bulk Rate', ' Bwd Avg Bytes/Bulk', ' Bwd Avg Packets/Bulk', 'Bwd Avg Bulk Rate']

4. PROPRIETA' STATISTICHE DEGLI ATTRIBUTI UTILI

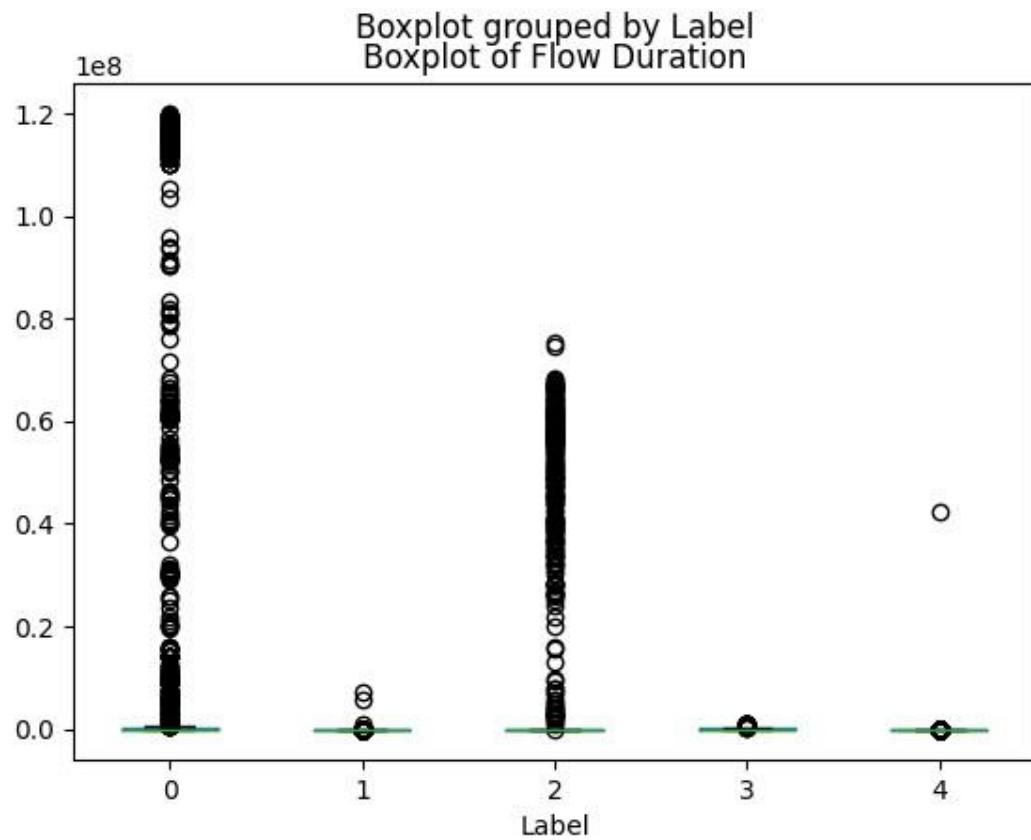
4.1 Protocol



- count 10000.000000
- mean 12.372700
- std 5.495017
- min 0.000000
- 25% 6.000000
- 50% 17.000000
- 75% 17.000000
- max 17.000000

Il boxplot mostra la presenza di outlier nella classe 2, con un valore di 17. La classe 0 presenta valori compresi tra 0 e 17, mentre classe 1, 3 e 4 hanno valori pari a 17. La classe 2 registra valori pari a 6.

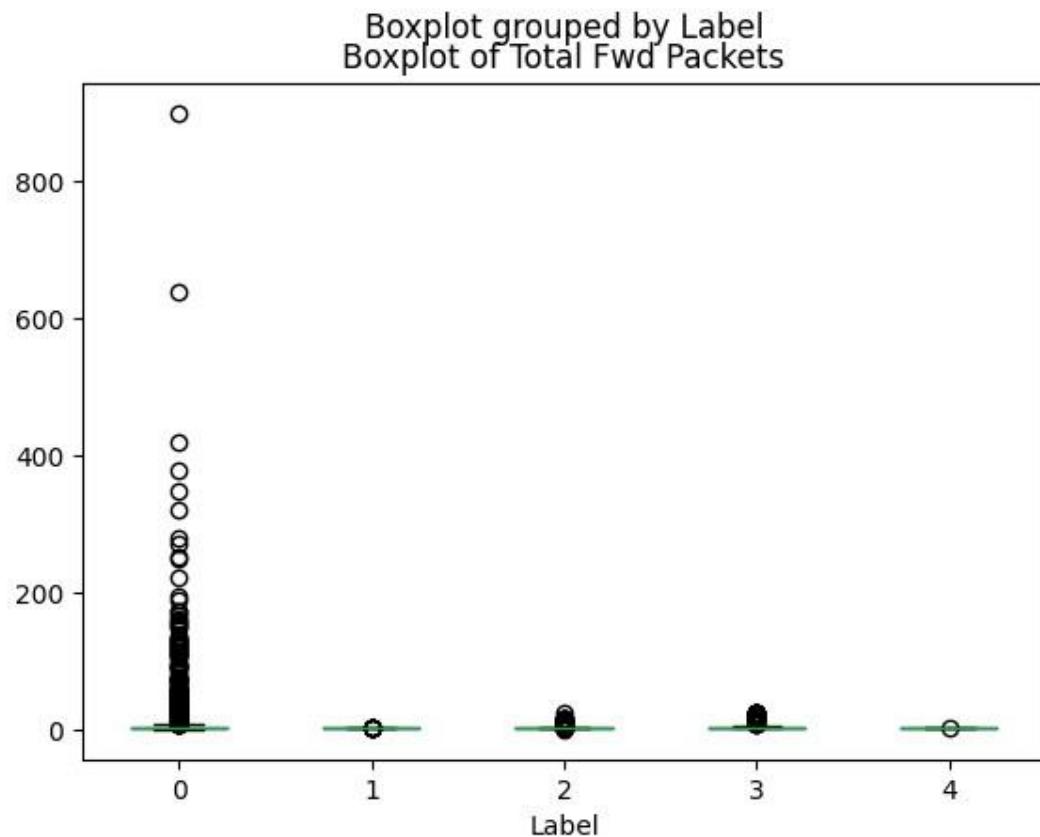
4.2 Flow Duration



- count 1.000000e+04
- mean 5.052362e+06
- std 2.086003e+07
- min 1.000000e+00
- 25% 1.000000e+00
- 50% 4.700000e+01
- 75% 3.113775e+04
- max 1.199499e+08

Tutte e cinque le classi presentano outlier, nello specifico Classe 0 e 2 presentano outlier che superano enormemente i limiti superiori.

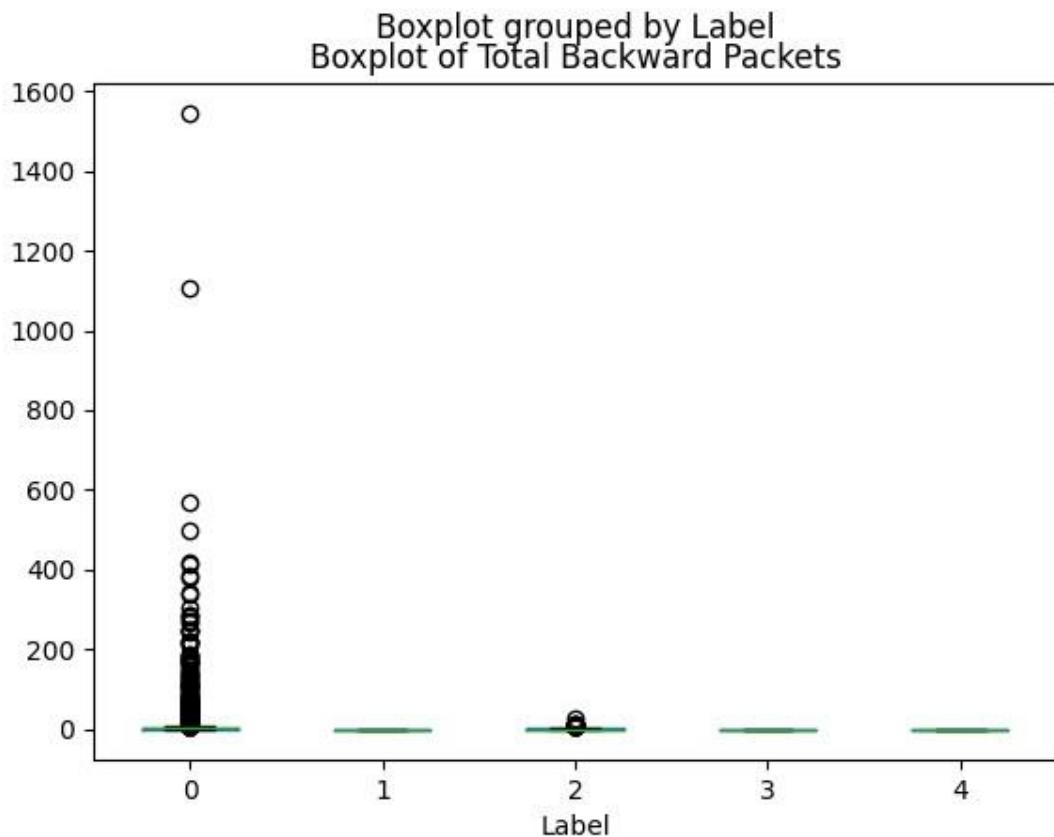
4.3 Total Fwd Packets



- count 10000.000000
- mean 4.974100
- std 17.905778
- min 1.000000
- 25% 2.000000
- 50% 2.000000
- 75% 2.000000
- max 899.000000

Tutte e cinque le classi presentano outlier, nello specifico la Classe 0 ha outlier che superano di gran lunga il valore più frequente (2).

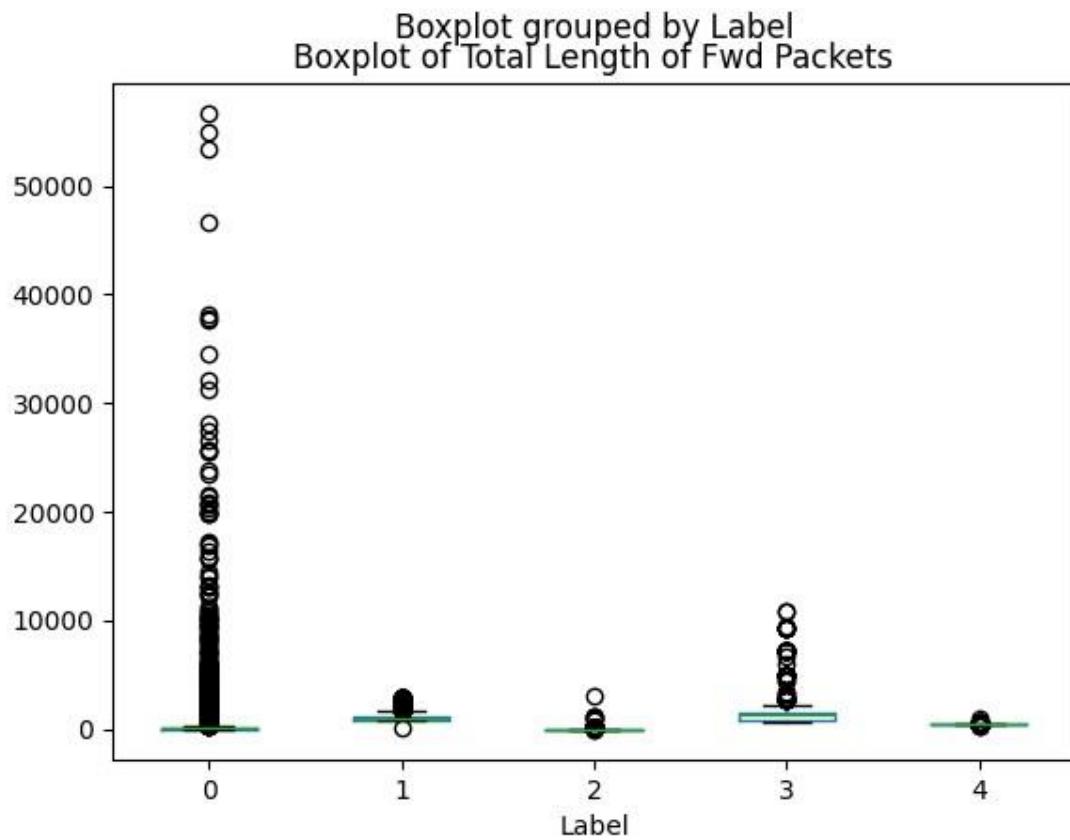
4.4 Total Backward Packets



- count 10000.000000
- mean 3.349400
- std 26.611075
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 2.000000
- max 1543.000000

Classe 0 e 2 presentano outlier, nello specifico la Classe 0 ha outlier (max 1543) che superano di gran lunga il valore più frequente (compreso tra 0 e 2).

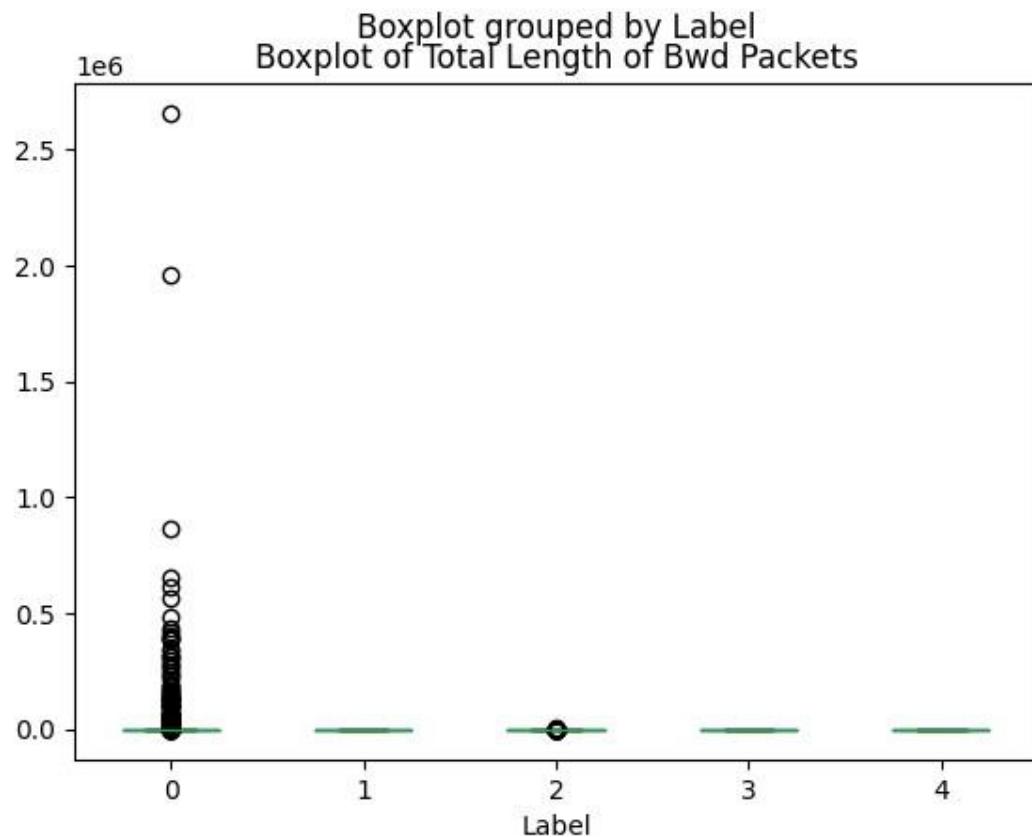
4.5 Total Length of Fwd Packets



- count 10000.000000
- mean 832.031800
- std 2049.949291
- min 0.000000
- 25% 12.000000
- 50% 458.000000
- 75% 1000.000000
- max 56622.000000

Tutte e cinque le classi presentano outlier, nello specifico la Classe 0 ha outlier che superano di gran lunga il valore più frequente (compreso tra 12 e 1000).

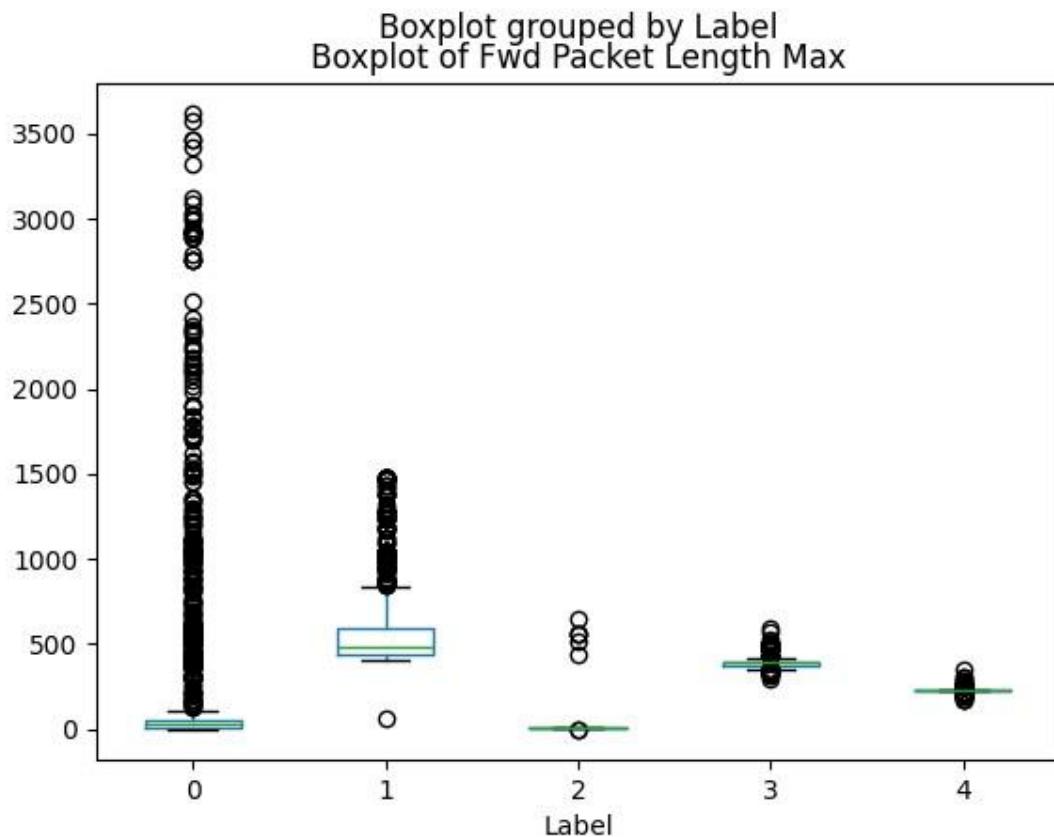
4.6 Total Length of Bwd Packets



- 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 $2.655090e+06$

Classe 0 e 2 presentano outlier, nello specifico la Classe 0 ha outlier (max 2.655.090) che superano di gran lunga il valore più frequente (compreso tra 0 e 12).

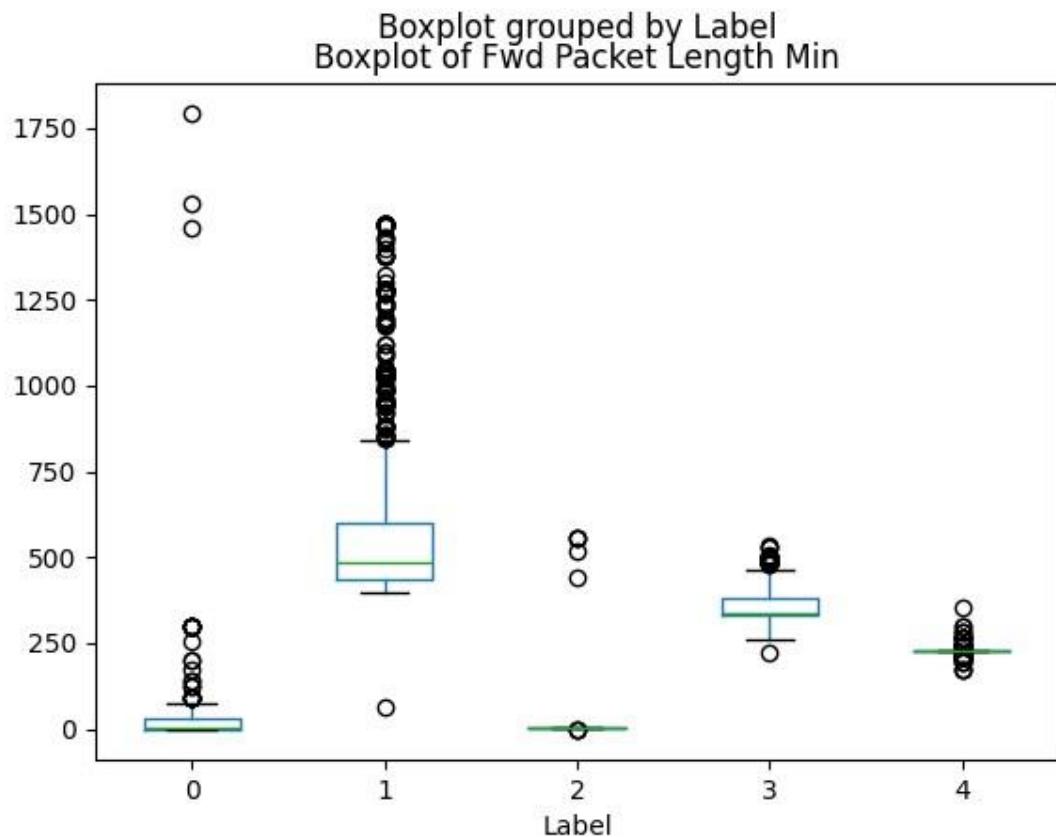
4.7 Fwd Packet Length Max



- count 10000.000000
- mean 266.889100
- std 322.176744
- min 0.000000
- 25% 6.000000
- 50% 229.000000
- 75% 403.000000
- max 3617.000000

Tutte e cinque le classi presentano outlier, classe 1, 2 e 3 presentano anche outlier inferiori.

4.8 Fwd Packet Length Min

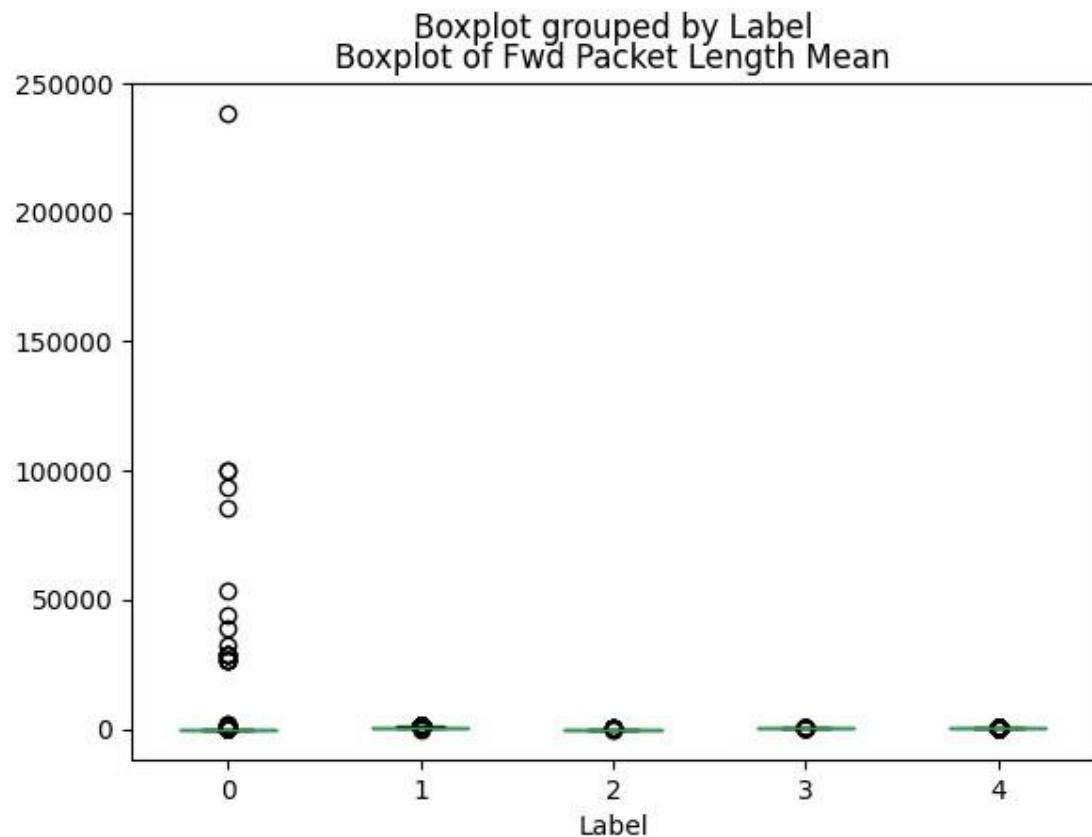


- count 10000.000000
- mean 213.788400
- std 245.620054
- min 0.000000
- 25% 6.000000
- 50% 211.000000
- 75% 383.000000
- max 1791.000000

Tutte e cinque le classi presentano outlier, nello specifico la Classe 0 ha outlier (max 1791) che superano di gran lunga i valori più frequenti.

Classe 1, 2, 3, 4 presentano anche outlier inferiori.

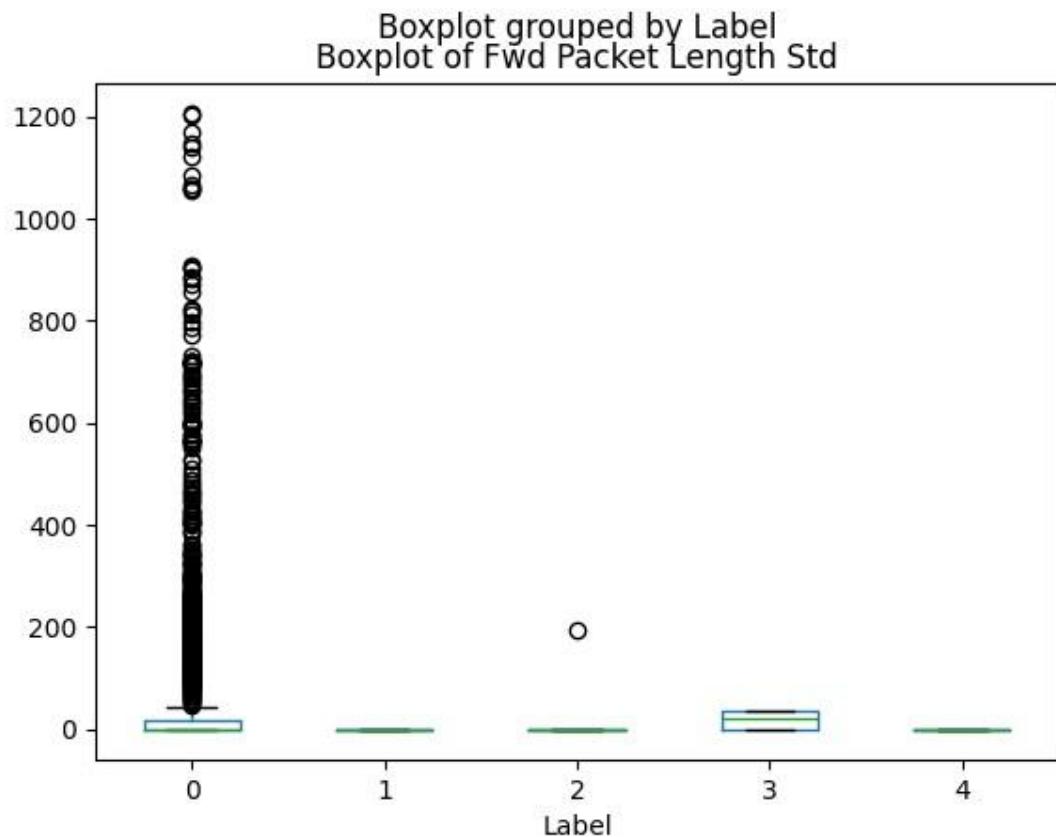
4.9 Fwd Packet Length Mean



- count 10000.000000
- mean 335.006267
- std 3299.525185
- min 0.000000
- 25% 6.000000
- 50% 229.000000
- 75% 383.000000
- max 238125.000000

Tutte e cinque le classi presentano outlier, nello specifico la Classe 0 ha outlier (max 238.125) che superano di gran lunga i valori più frequenti.

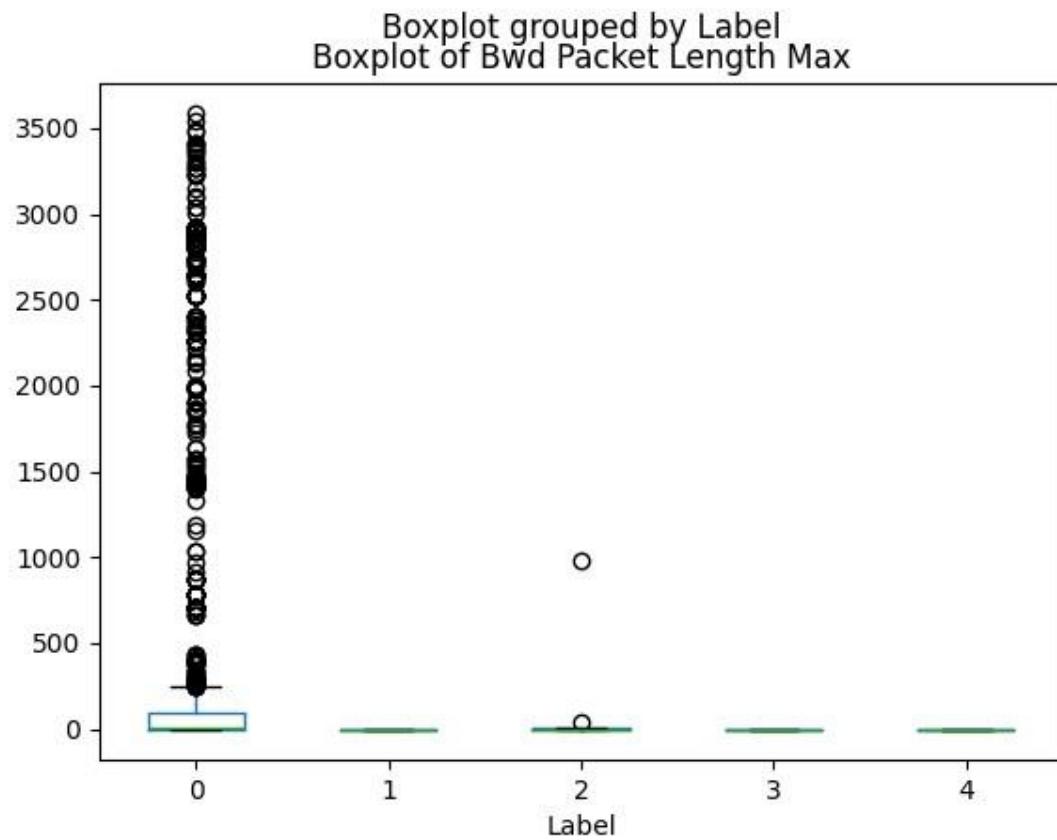
4.10 Fwd Packet Length Std



- count 10000.000000
- mean 18.225182
- std 77.235121
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 0.000000
- max 1205.477522

Classe 0 e 2 presentano outlier, in particolare la Classe 0 ha outlier (valore max 1.205) che superano di gran lunga i valori più frequenti.

4.11 Bwd Packet Length Max

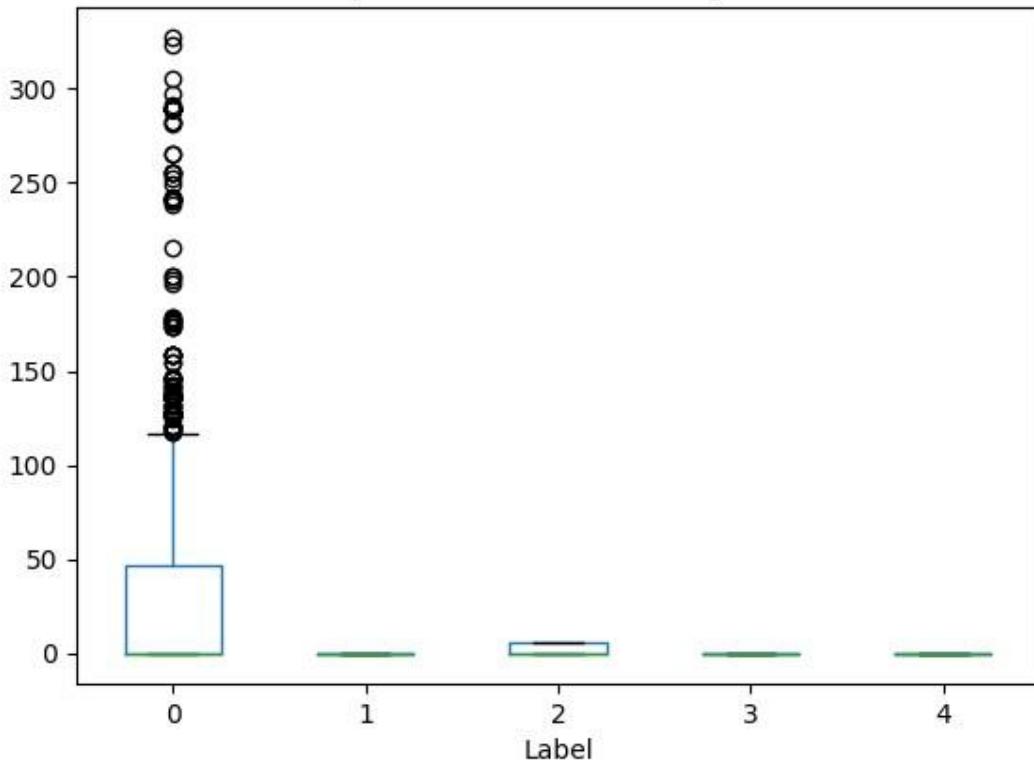


- count 10000.000000
- mean 98.885900
- std 446.663368
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 6.000000
- max 3583.000000

Classe 0 e 2 presentano outlier, in particolare la Classe 0 ha outlier (max 3.583) che superano di gran lunga i valori più frequenti.

4.12 Bwd Packet Length Min

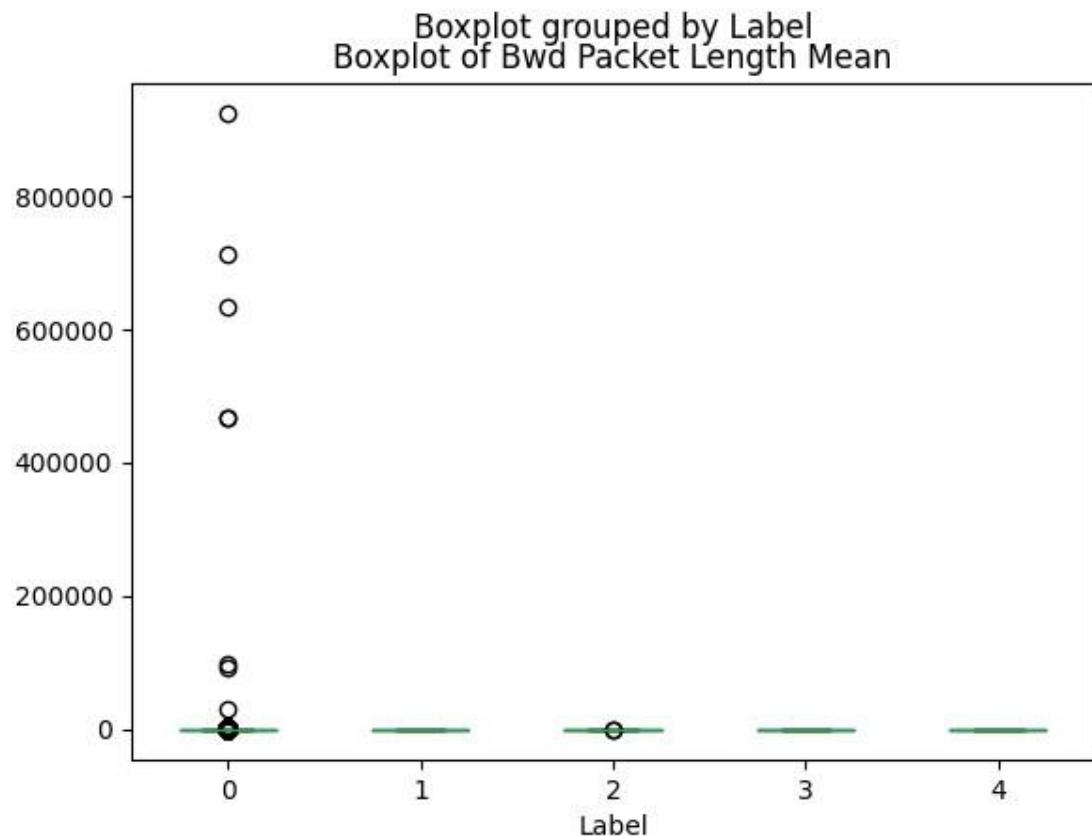
Boxplot grouped by Label
Boxplot of Bwd Packet Length Min



- count 10000.000000
- mean 8.426500
- std 28.499972
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 0.000000
- max 327.000000

La classe 0 è l'unica che presenta outlier che superano di gran lunga i valori più frequenti (valore max 327).

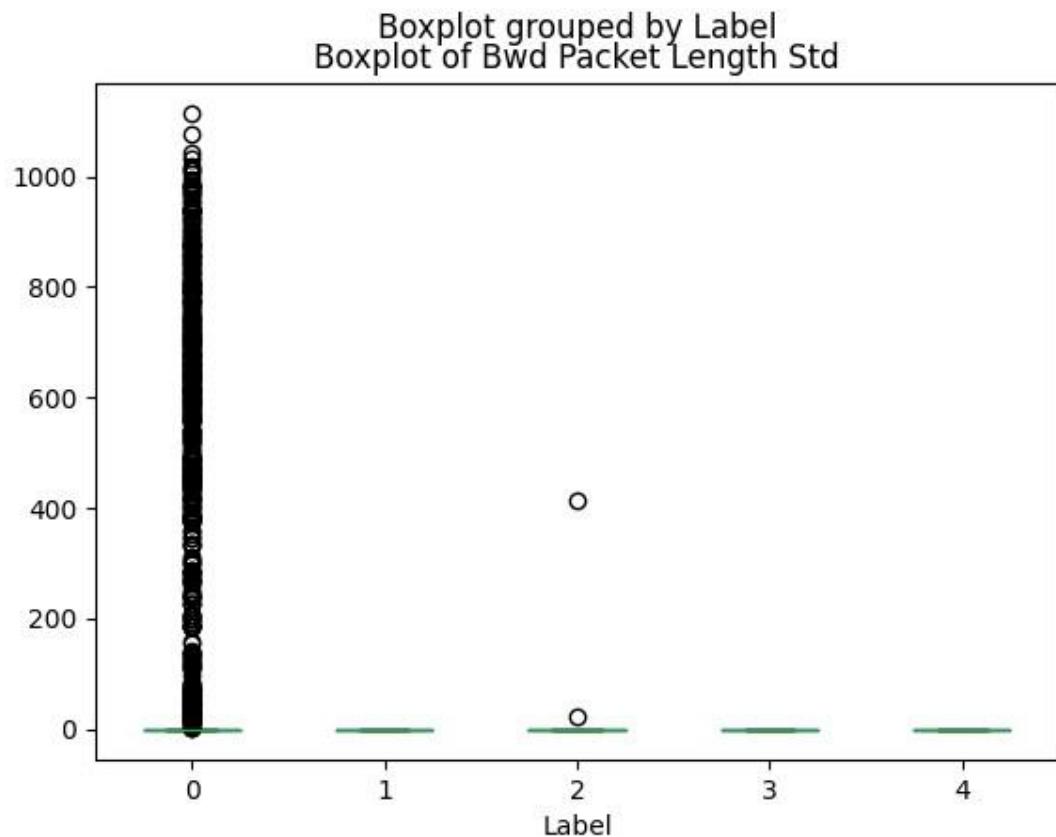
4.13 Bwd Packet Length Mean



- count 10000.000000
- mean 375.263850
- std 14894.726524
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 6.000000
- max 924125.000000

Classe 0 e 2 presentano outlier, in particolare la Classe 0 ha outlier (max 924125) che superano di gran lunga i valori più frequenti.

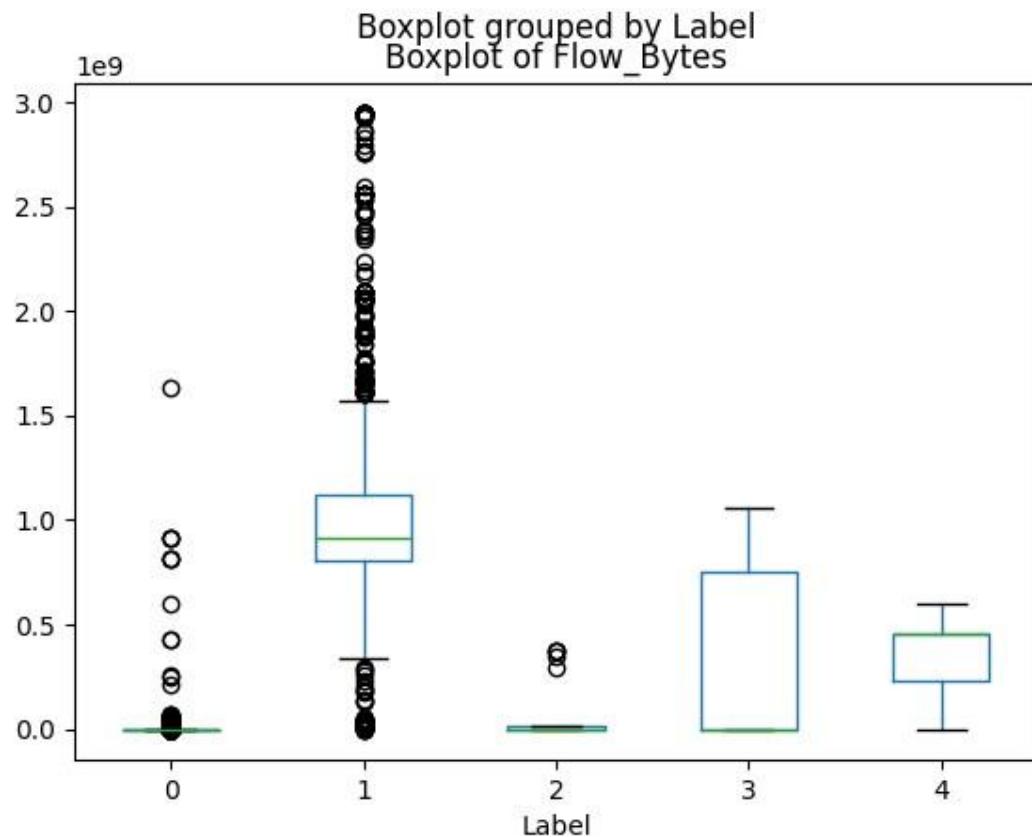
4.14 Bwd Packet Length Std



- count 10000.000000
- mean 27.253425
- std 132.108142
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 0.000000
- max 1112.854940

Classe 0 e 2 presentano outlier, in particolare la Classe 0 ha outlier (max 1.112,85) che superano di gran lunga i valori più frequenti.

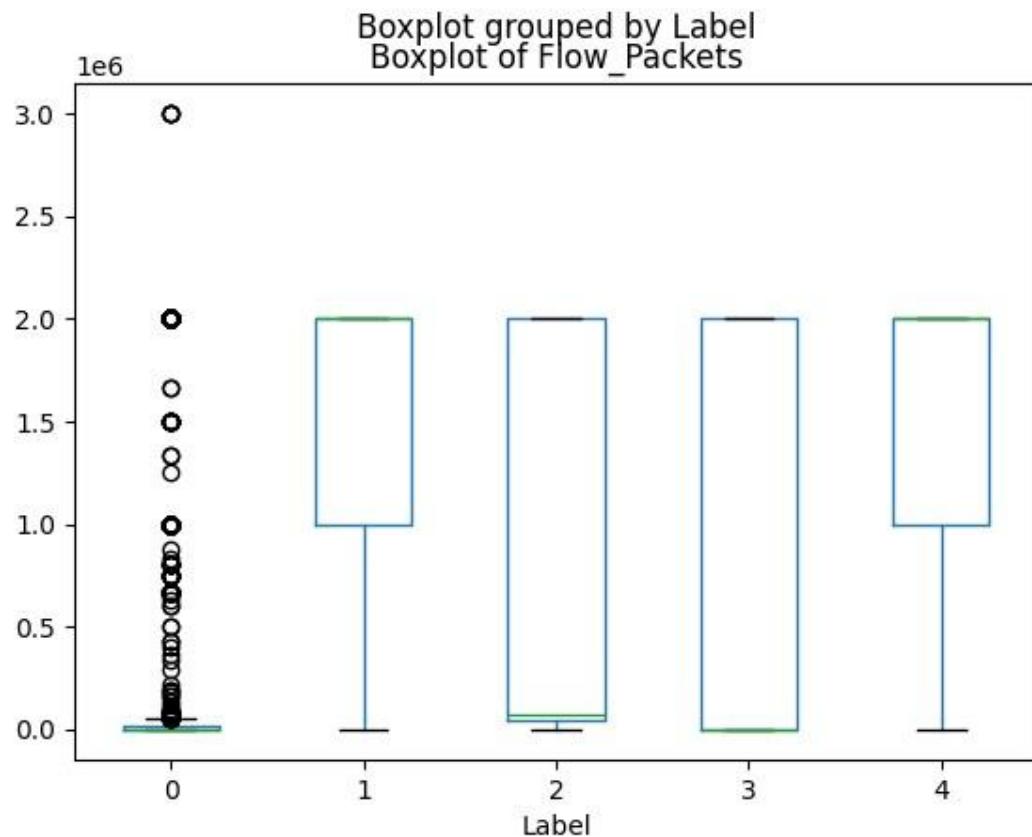
4.15 Flow_Bytes



- count 1.000000e+04
- mean 2.784807e+08
- std 4.588245e+08
- min 0.000000e+00
- 25% 1.288851e+04
- 50% 9.541667e+06
- 75% 4.580000e+08
- max 2.944000e+09

Classe 0, 1 e 2 presentano outlier, in particolare Classe 1 ha sia outlier superiori che inferiori.

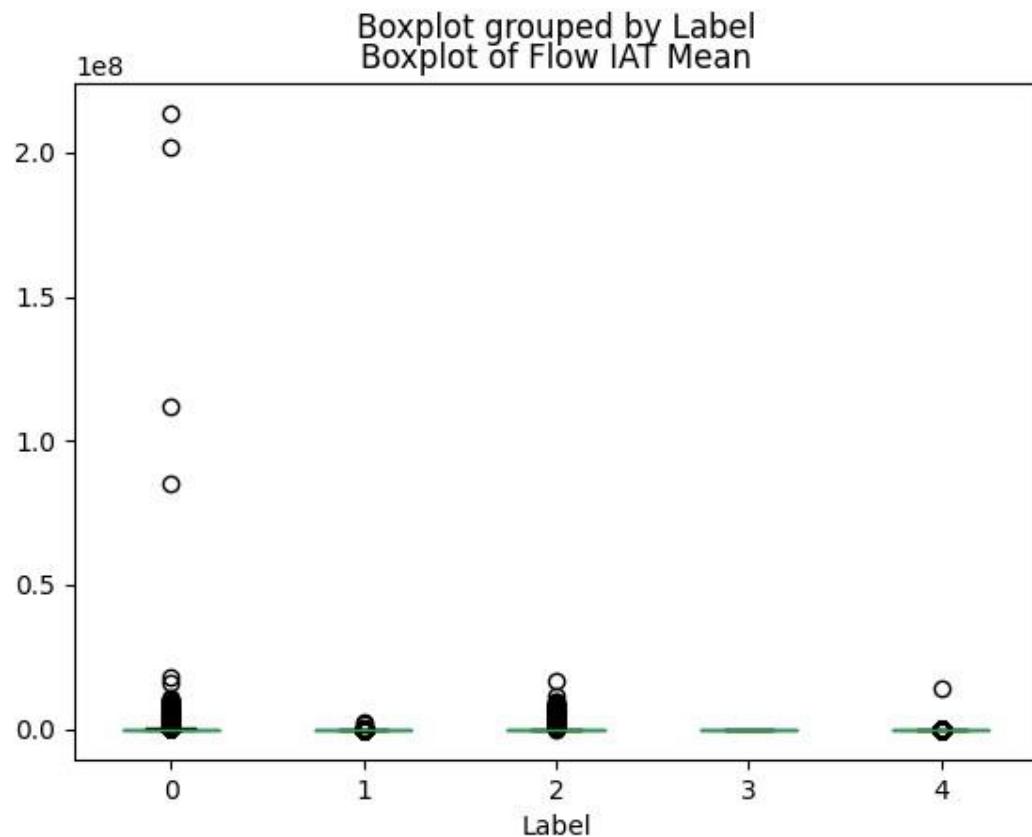
4.16 Flow_Packets



- count $1.000000e+04$
- mean $8.762053e+05$
- std $9.464913e+05$
- min $7.475528e-02$
- 25% $1.532832e+02$
- 50% $4.687500e+04$
- 75% $2.000000e+06$
- max $3.000000e+06$

La classe 0 è l'unica che presenta outlier, alcuni di questi superano di gran lunga i valori più frequenti (valore max 3.000.000).

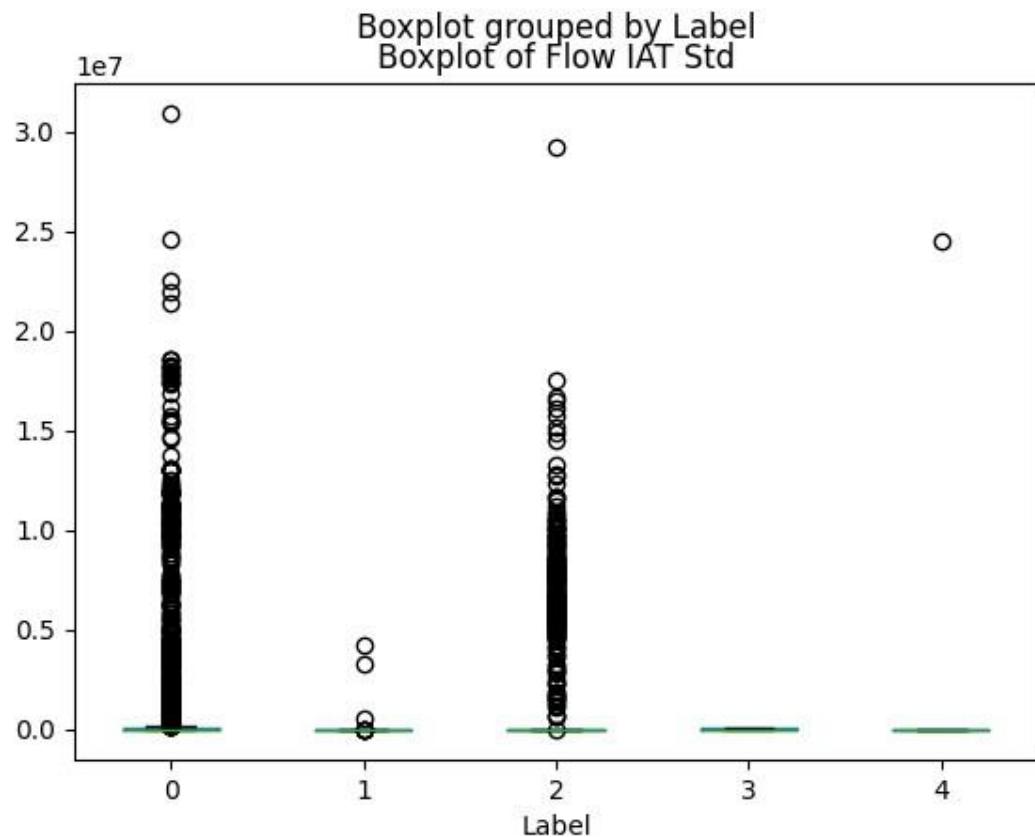
4.17 Flow IAT Mean



- count $1.000000\text{e+}04$
- mean $2.734815\text{e+}05$
- std $3.385072\text{e+}06$
- min $5.000000\text{e-}01$
- 25% $1.000000\text{e+}00$
- 50% $3.133333\text{e+}01$
- 75% $8.393483\text{e+}03$
- max $2.133164\text{e+}08$

Classe 0,1,2 e 4 presentano outlier, in particolare la Classe 0 ha outlier (max 213.316.400) che superano di gran lunga i valori più frequenti.

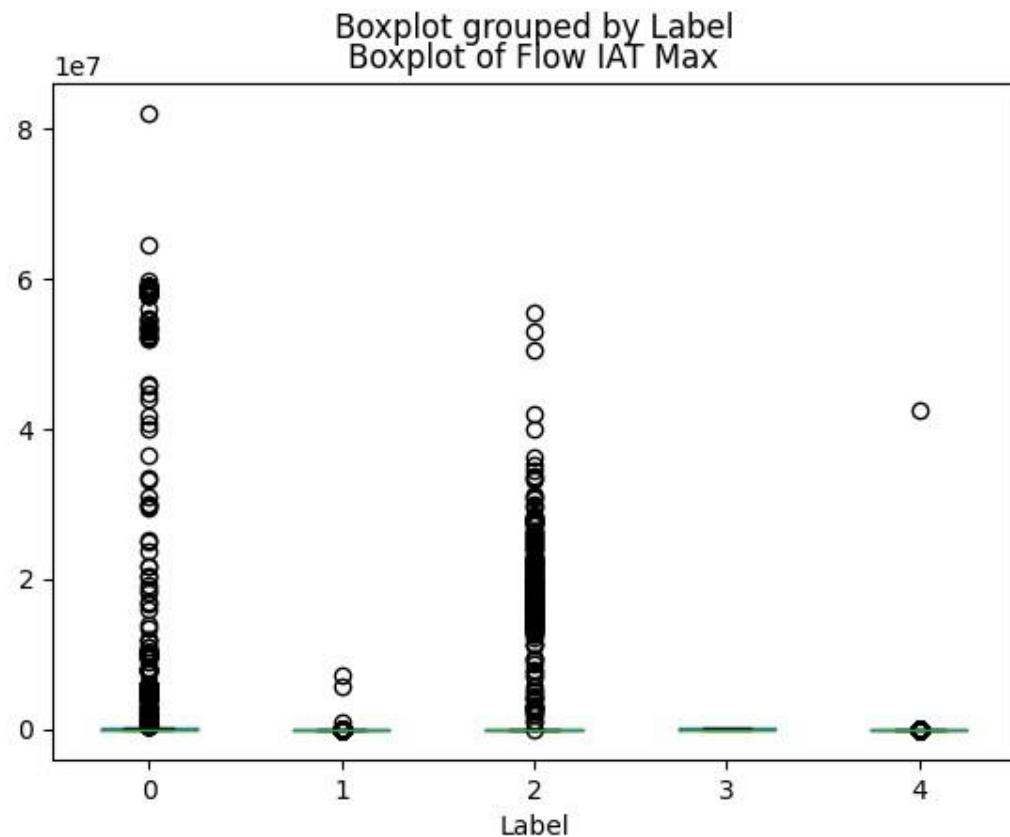
4.18 Flow IAT Std



- count $1.000000e+04$
- mean $4.955549e+05$
- std $2.040065e+06$
- min $0.000000e+00$
- 25% $0.000000e+00$
- 50% $0.000000e+00$
- 75% $1.256485e+04$
- max $3.089281e+07$

Classe 0,1,2 e 4 presentano outlier che superano il baffo superiore.

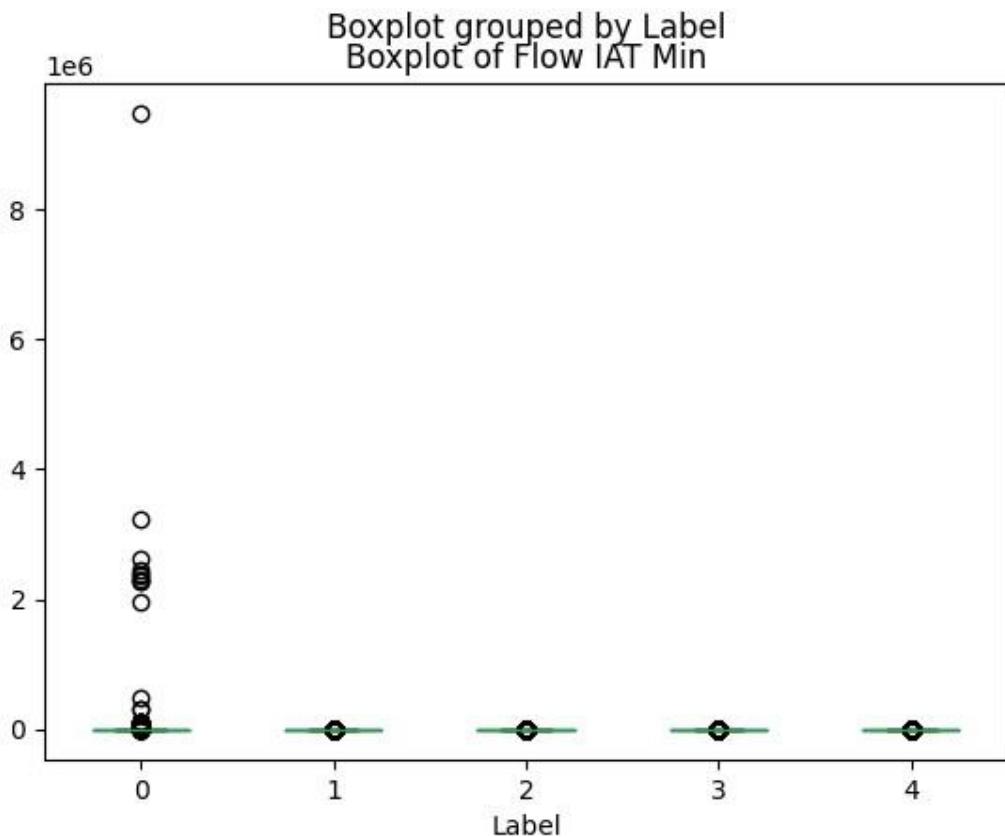
4.19 Flow IAT Max



- count $1.000000e+04$
- mean $1.967495e+06$
- std $8.750638e+06$
- min $1.000000e+00$
- 25% $1.000000e+00$
- 50% $4.700000e+01$
- 75% $2.820800e+04$
- max $8.201220e+07$

Classe 0,1,2 e 4 presentano outlier che superano il baffo superiore.

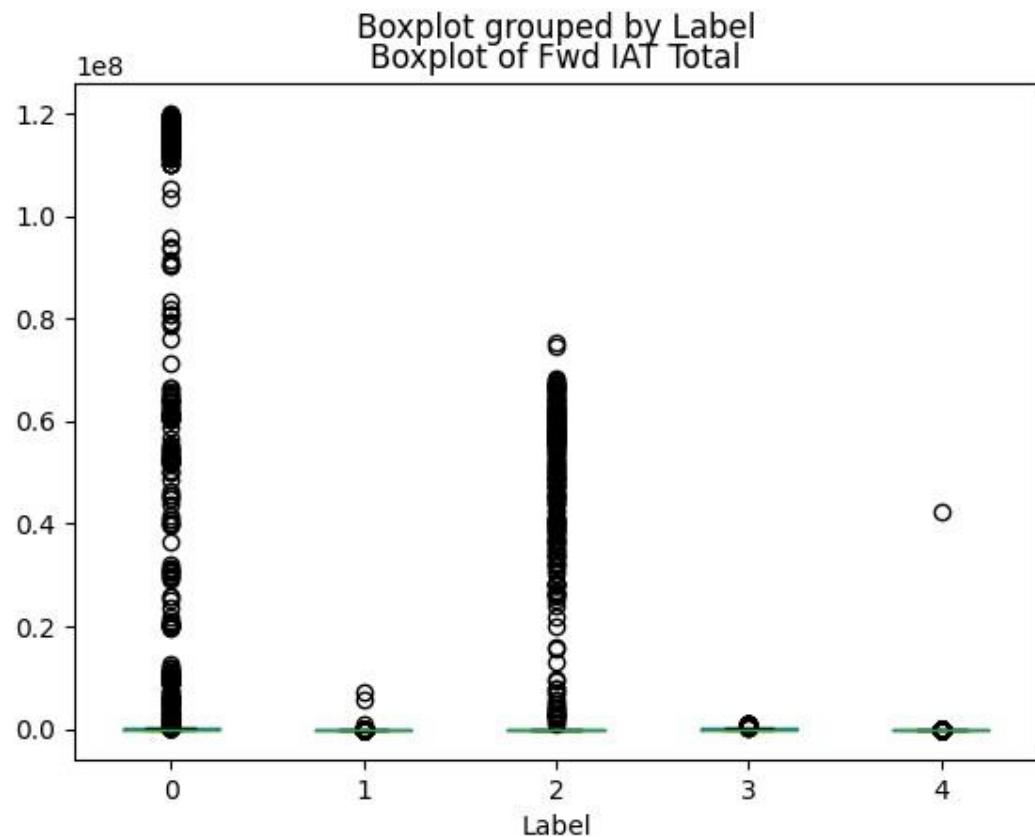
4.20 Flow IAT Min



- count 1.000000e+04
- mean 3.766345e+03
- std 1.201319e+05
- min 0.000000e+00
- 25% 1.000000e+00
- 50% 1.000000e+00
- 75% 2.000000e+00
- max 9.472020e+06

Tutte e cinque le classi presentano outlier, nello specifico la Classe 0 ha outlier (max 9.472.020) che superano enormemente i valori più frequenti.

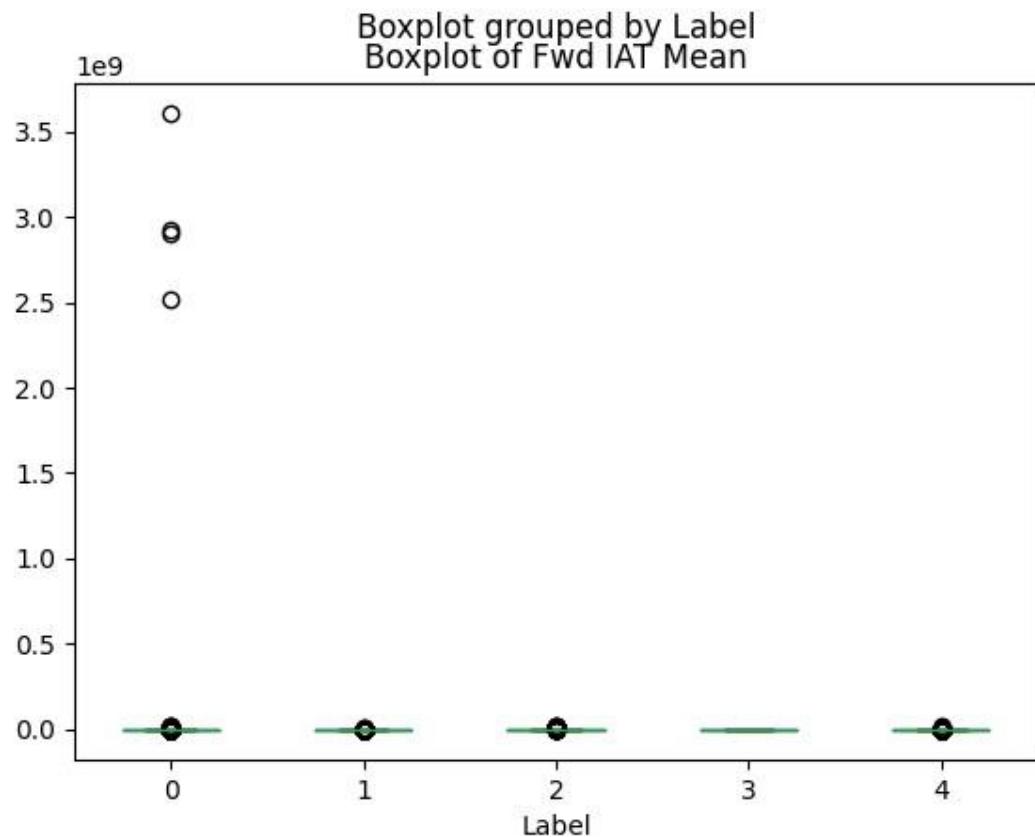
4.21 Fwd IAT Total



- count $1.000000e+04$
- mean $4.956191e+06$
- std $2.083816e+07$
- min $0.000000e+00$
- 25% $1.000000e+00$
- 50% $1.000000e+00$
- 75% $4.900000e+01$
- max $1.199499e+08$

Tutte e cinque le classi presentano outlier, nello specifico la Classe 0 ha outlier (max 9.472.020) che superano enormemente i valori più frequenti.

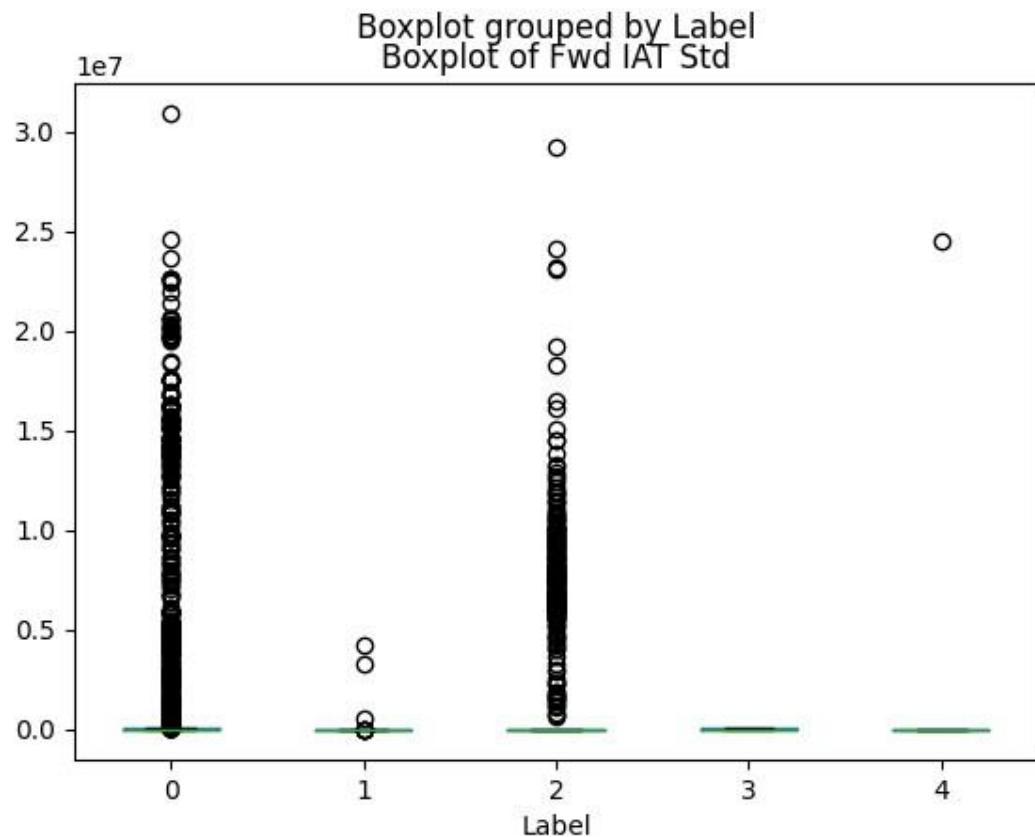
4.22 Fwd IAT Mean



- count $1.000000\text{e+}04$
- mean $1.488125\text{e+}06$
- std $6.023732\text{e+}07$
- min $0.000000\text{e+}00$
- 25% $1.000000\text{e+}00$
- 50% $1.000000\text{e+}00$
- 75% $4.900000\text{e+}01$
- max $3.604371\text{e+}09$

Classe 0, 1, 2 e 4 presentano outlier, nello specifico la Classe 0 ha outlier (max 3.604.371.000) che superano enormemente i valori più frequenti.

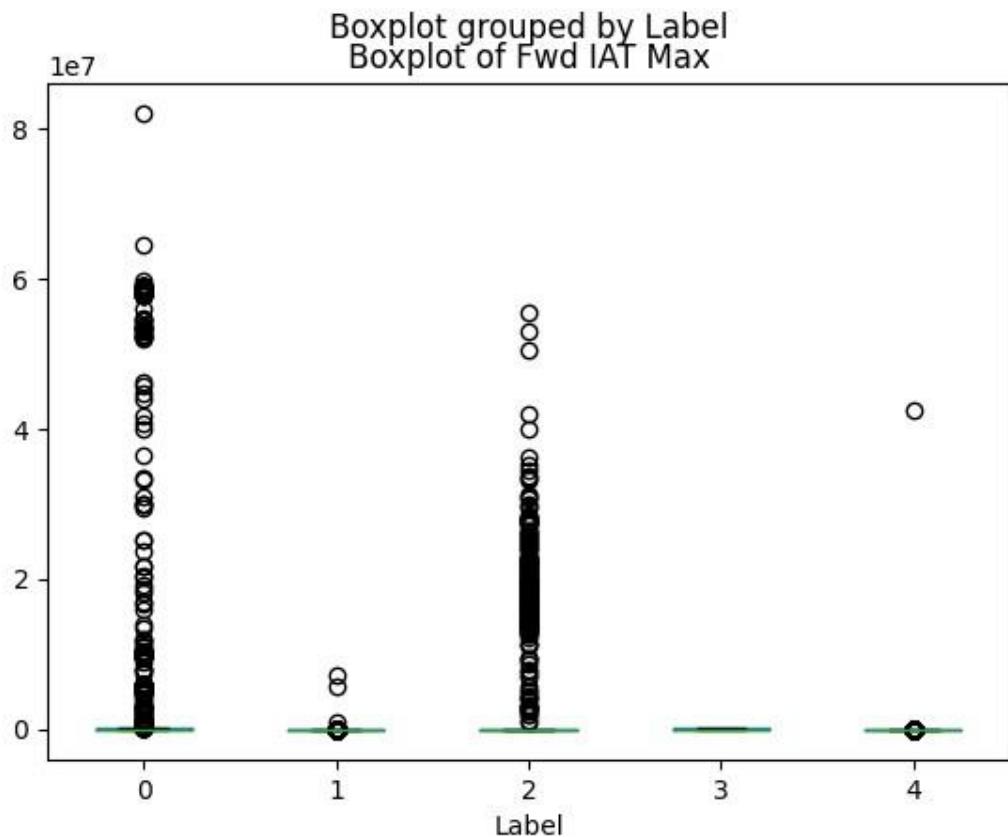
4.23 Fwd IAT Std



- count $1.000000e+04$
- mean $5.920651e+05$
- std $2.506531e+06$
- min $0.000000e+00$
- 25% $0.000000e+00$
- 50% $0.000000e+00$
- 75% $0.000000e+00$
- max $3.089281e+07$

Classe 0, 1, 2 e 4 presentano outlier, nello specifico la Classe 0,2 e 4 hanno outlier che superano enormemente il baffo superiore.

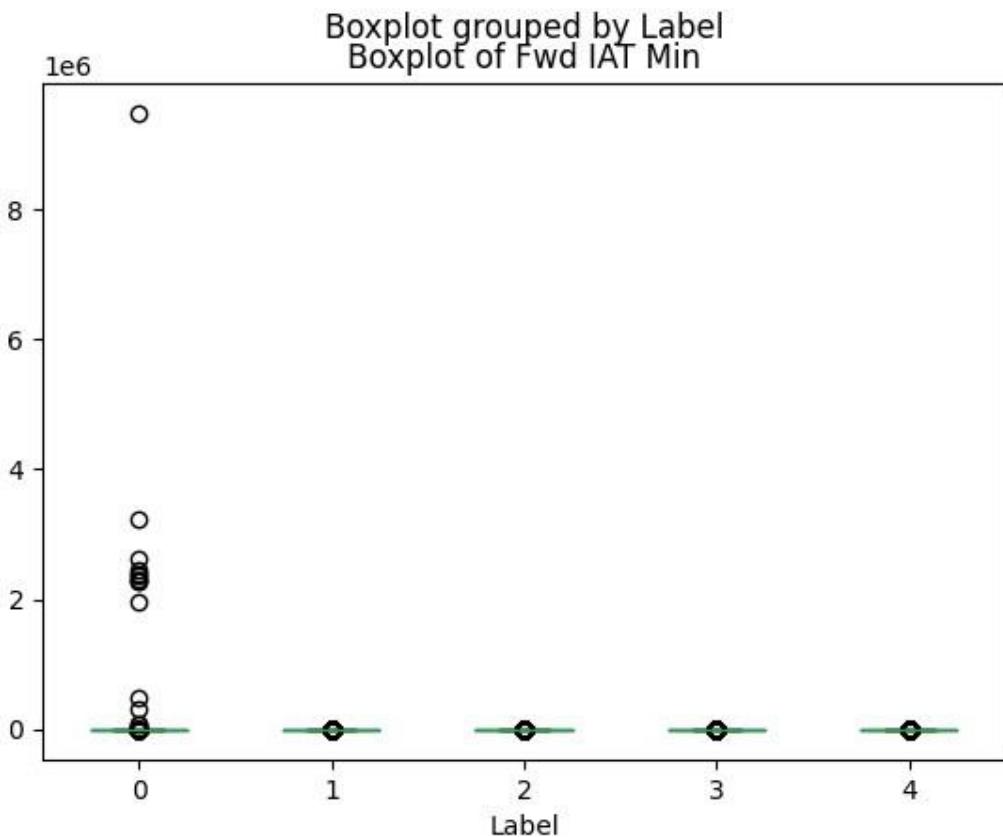
4.24 Fwd IAT Max



- count $1.000000\text{e+}04$
- mean $1.918364\text{e+}06$
- std $8.750342\text{e+}06$
- min $0.000000\text{e+}00$
- 25% $1.000000\text{e+}00$
- 50% $1.000000\text{e+}00$
- 75% $4.900000\text{e+}01$
- max $8.201220\text{e+}07$

Classe 0, 1, 2 e 4 presentano outlier, nello specifico la Classe 0,2 e 4 hanno outlier che superano enormemente il baffo superiore.

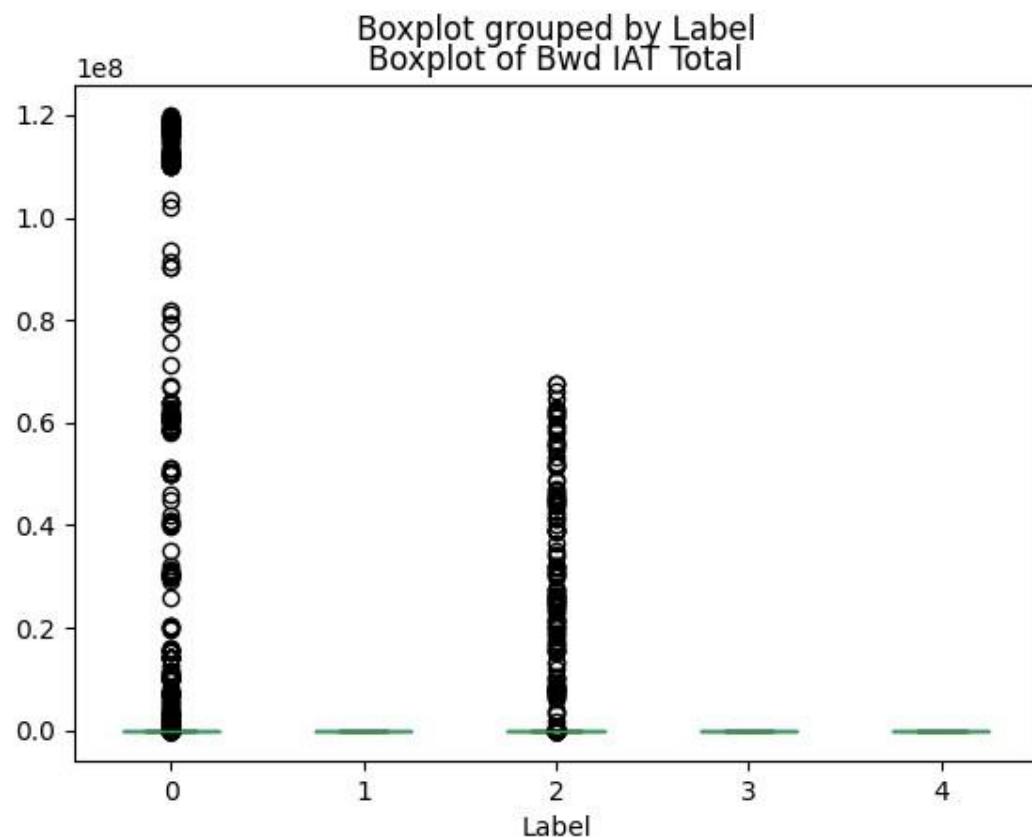
4.25 Fwd IAT Min



- count $1.000000e+04$
- mean $3.233667e+03$
- std $1.200038e+05$
- min $0.000000e+00$
- 25% $1.000000e+00$
- 50% $1.000000e+00$
- 75% $2.000000e+00$
- max $9.472020e+06$

Tutte e cinque le classi presentano outlier, nello specifico la Classe 0 ha outlier (max 9.472.020) che superano enormemente i valori più frequenti.

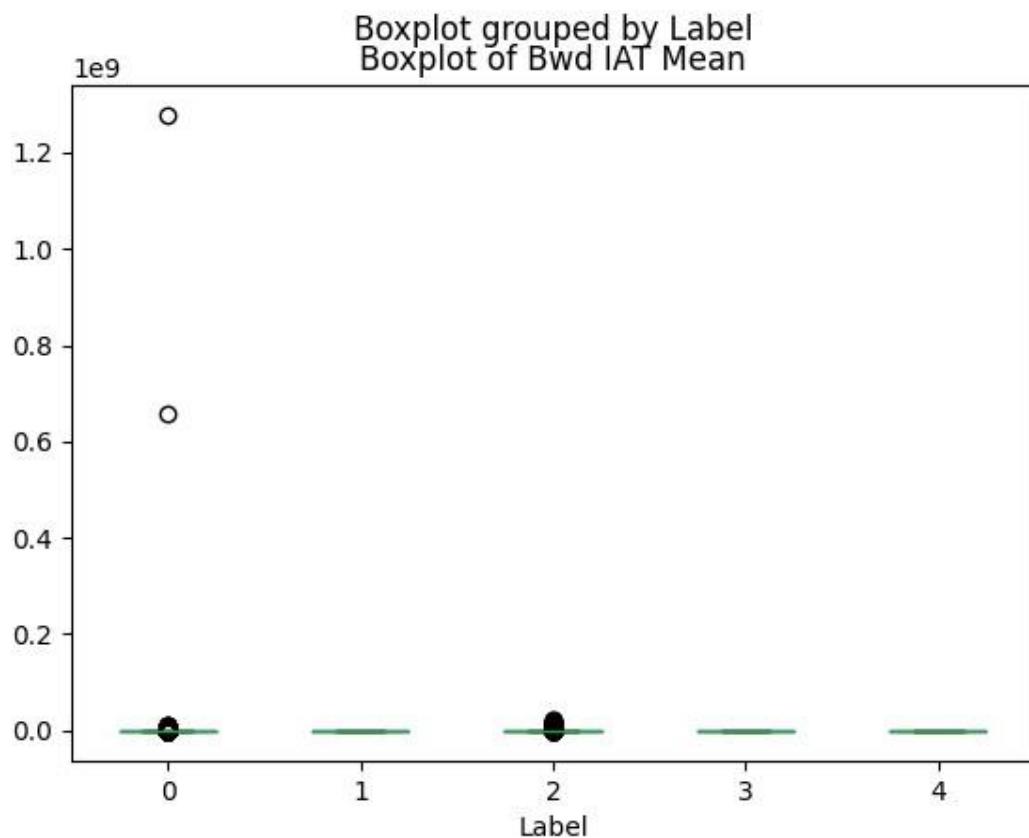
4.26 Bwd IAT Total



- count $1.000000\text{e+}04$
- mean $3.551726\text{e+}06$
- std $1.813216\text{e+}07$
- min $0.000000\text{e+}00$
- 25% $0.000000\text{e+}00$
- 50% $0.000000\text{e+}00$
- 75% $1.000000\text{e+}00$
- max $1.198525\text{e+}08$

Classi 0 e 2 presentano outlier che superano di gran lunga il baffo superiore.

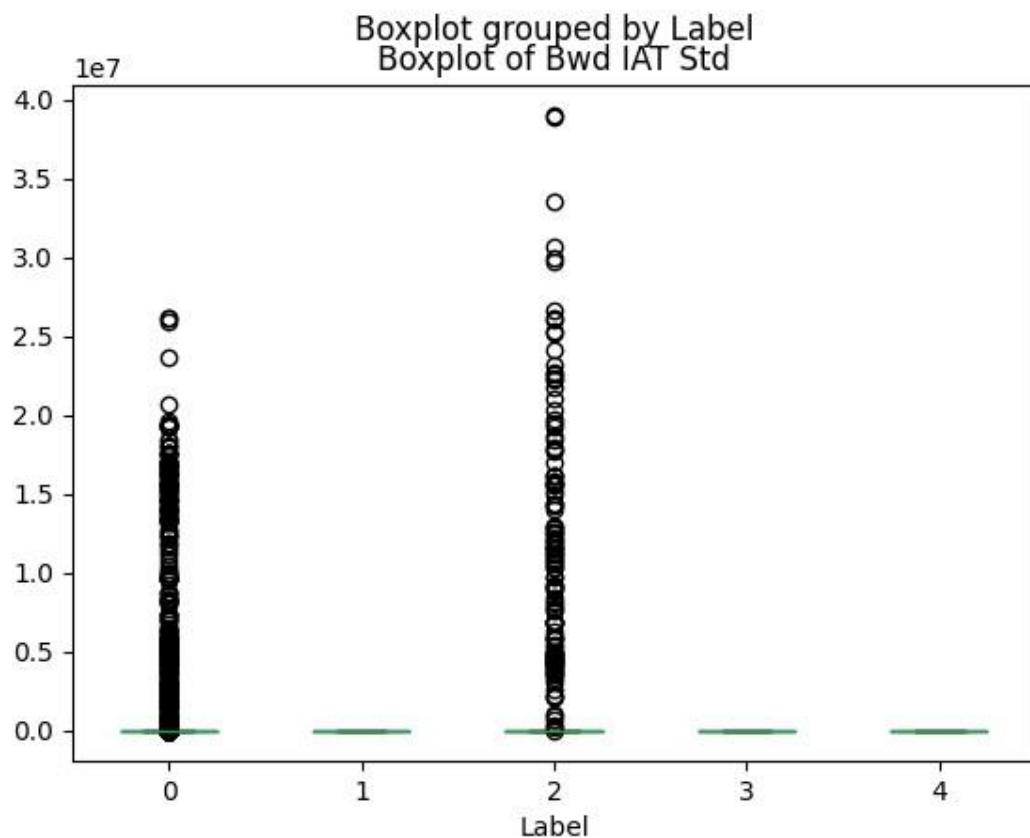
4.27 Bwd IAT Mean



- count 1.000000e+04
- mean 3.981695e+05
- std 1.440115e+07
- min 0.000000e+00
- 25% 0.000000e+00
- 50% 0.000000e+00
- 75% 1.000000e+00
- max 1.276447e+09

Classi 0 e 2 presentano outlier, il valore massimo di 1.276.447.000 supera
enormemente il baffo superiore.

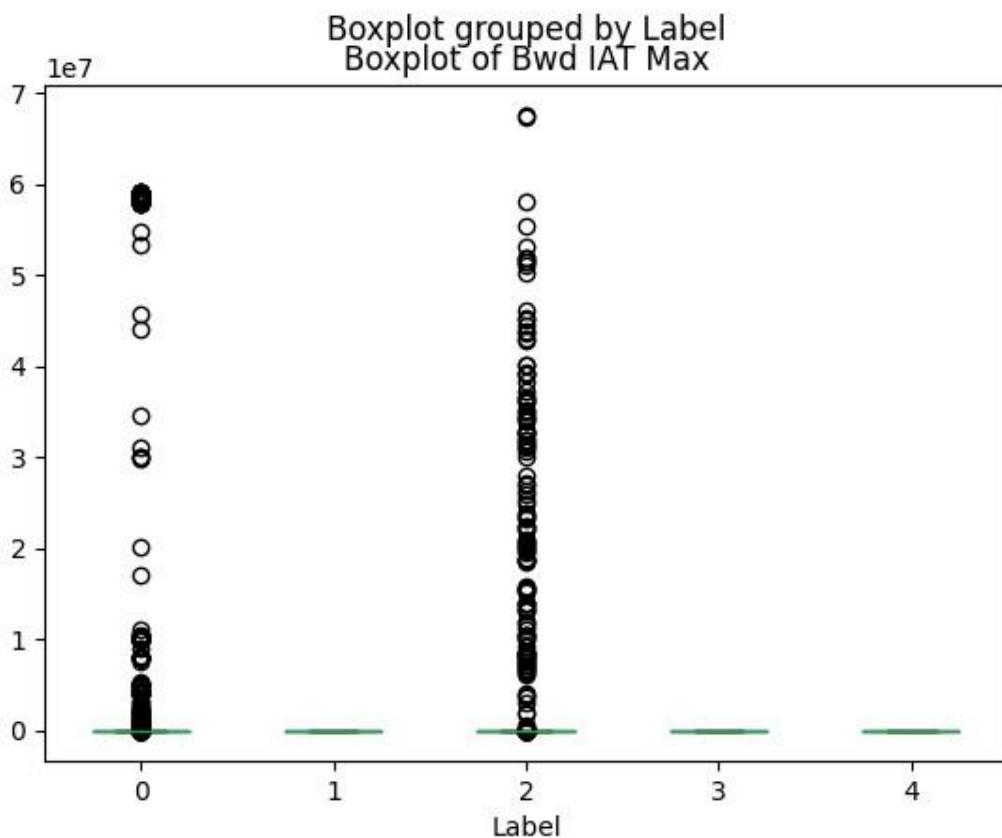
4.28 Bwd IAT Std



- count $1.000000\text{e+}04$
- mean $4.678660\text{e+}05$
- std $2.482373\text{e+}06$
- min $0.000000\text{e+}00$
- 25% $0.000000\text{e+}00$
- 50% $0.000000\text{e+}00$
- 75% $0.000000\text{e+}00$
- max $3.897542\text{e+}07$

Classi 0 e 2 presentano outlier, il valore massimo di 38.975.420 supera enormemente il baffo superiore per la classe 2.

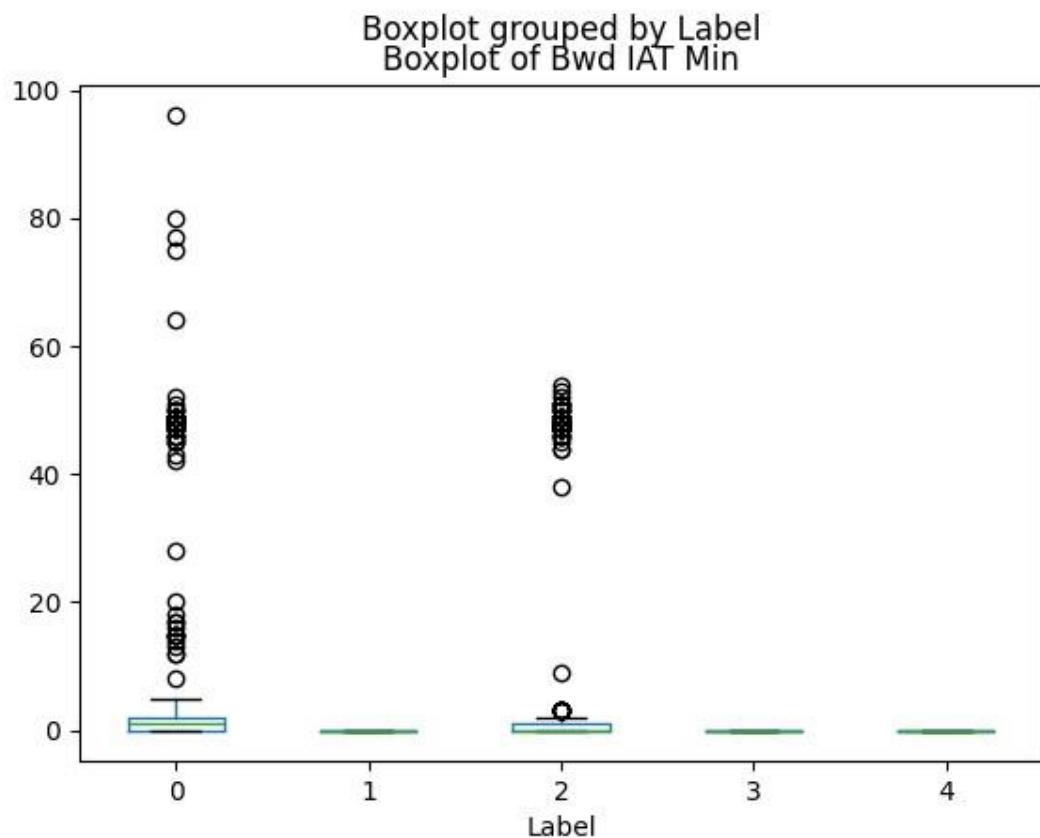
4.29 Bwd IAT Max



- count 1.000000e+04
- mean 1.560932e+06
- std 8.364169e+06
- min 0.000000e+00
- 25% 0.000000e+00
- 50% 0.000000e+00
- 75% 1.000000e+00
- max 6.750742e+07

Classi 0 e 2 presentano outlier, il valore massimo di 67.507.420 supera enormemente il baffo superiore per la classe 2.

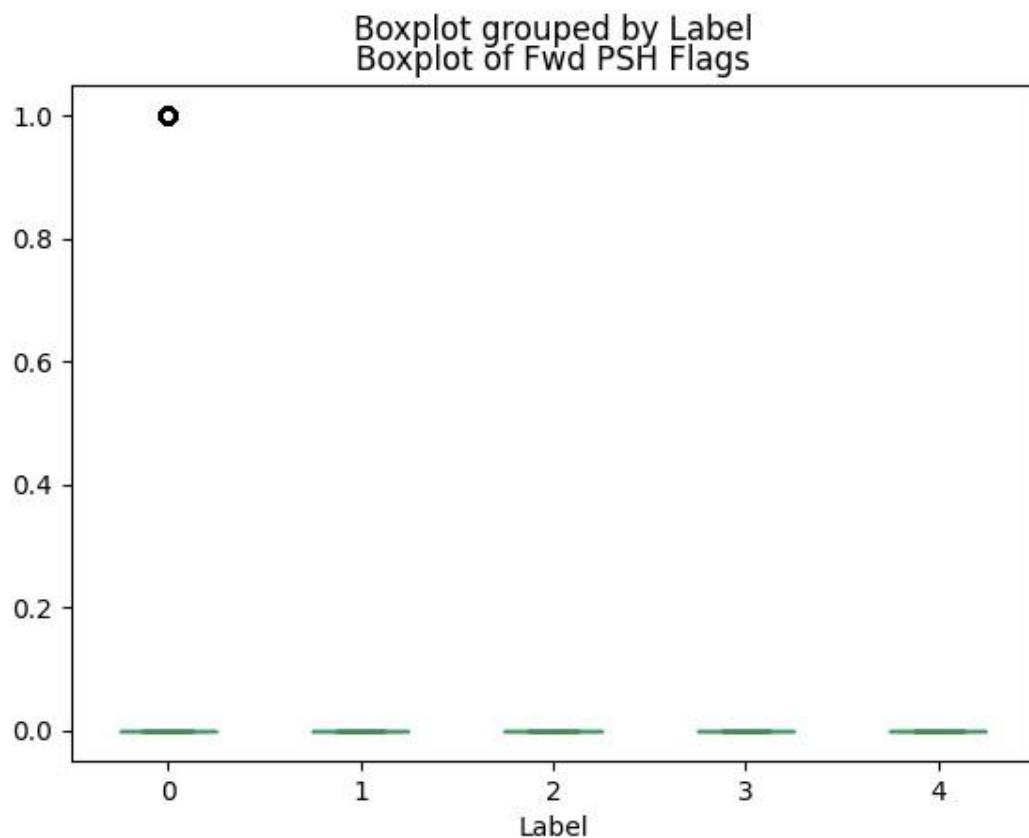
4.30 Bwd IAT Min



- count 10000.000000
- mean 0.988600
- std 5.089033
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 1.000000
- max 96.000000

Classi 0 e 2 presentano outlier, il valore massimo di 96 supera enormemente il baffo superiore per la classe 2.

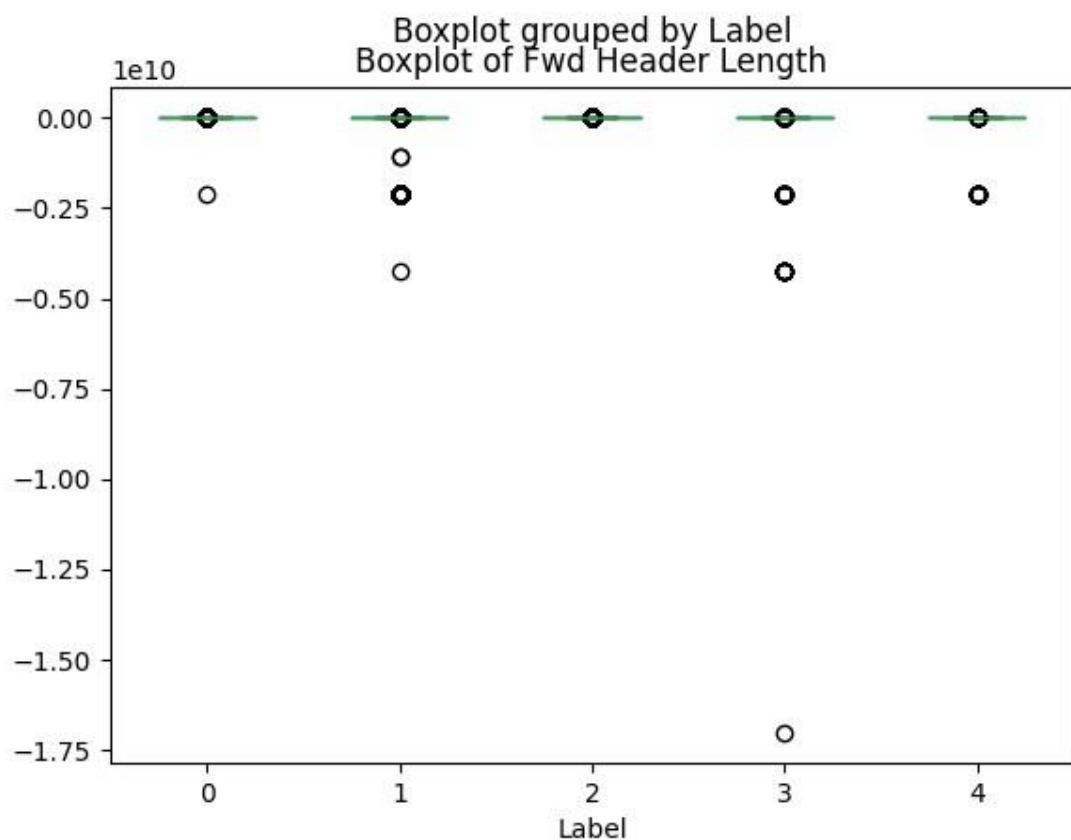
4.31 Fwd PSH Flags



- count 10000.000000
- mean 0.055800
- std 0.229547
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 0.000000
- max 1.000000

Dalle analisi emerge che la feature contiene valori booleani (0 e 1), assume 1 solo per Classe 0.

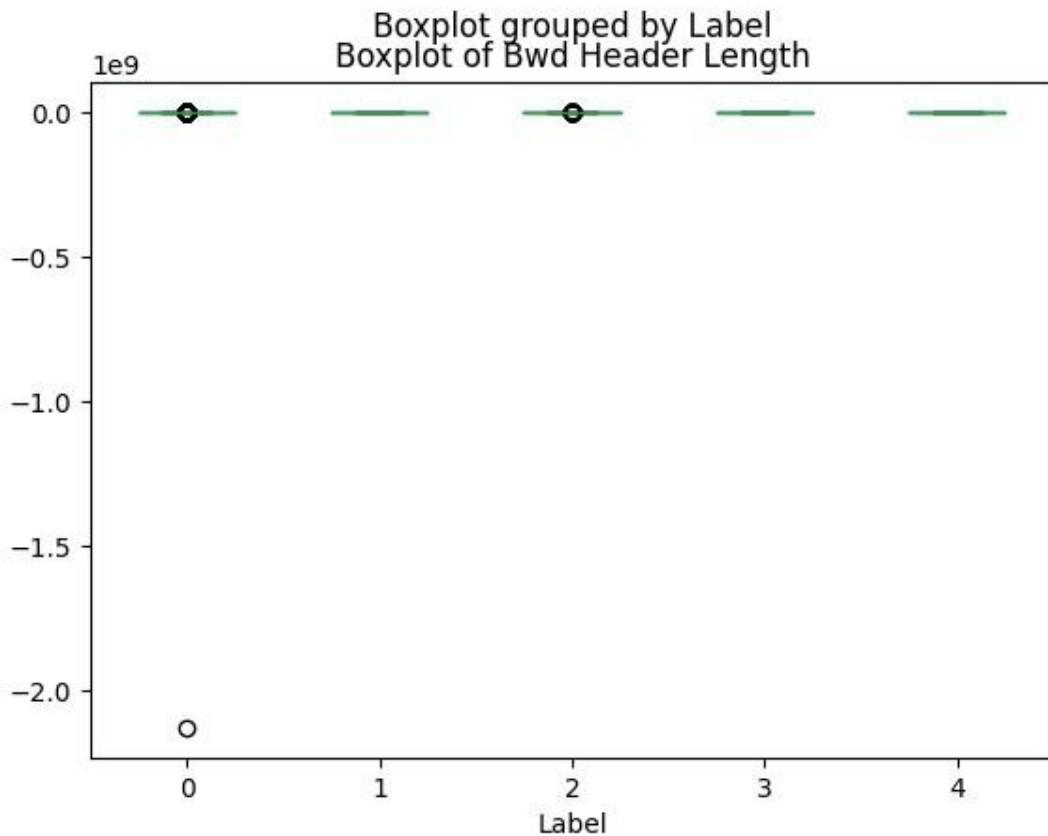
4.32 Fwd Header Length



- count $1.000000e+04$
- mean $-6.057485e+07$
- std $4.012595e+08$
- min $-1.700350e+10$
- 25% $2.800000e+01$
- 50% $4.000000e+01$
- 75% $6.400000e+01$
- max $2.878400e+04$

Tutte e 5 le classi presentano degli outlier inferiori, in particolare Classe 3 ha un outlier che supera enormemente il baffo inferiore –17mld

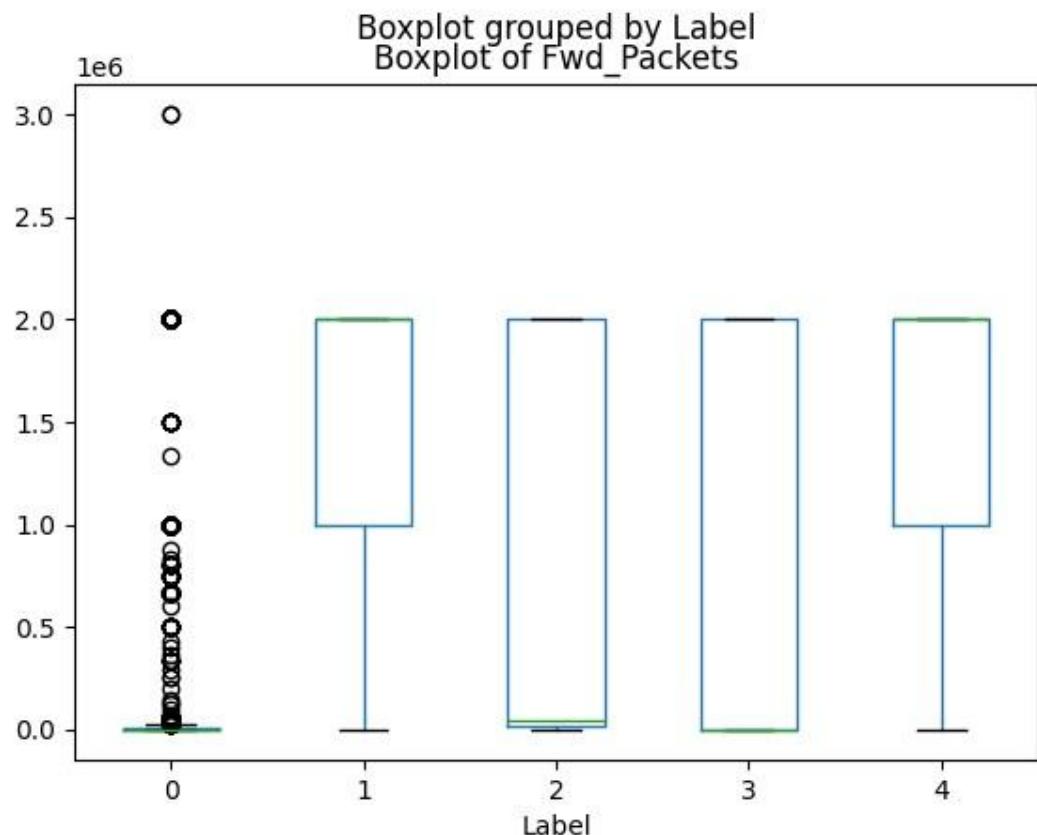
4.33 Bwd Header Length



- count 1.000000e+04
- mean -2.124557e+05
- std 2.125430e+07
- min -2.125430e+09
- 25% 0.000000e+00
- 50% 0.000000e+00
- 75% 4.000000e+01
- max 4.939200e+04

Classe 0 e 2 presentano outlier, in particolare il valore minimo 2.125.430.000 supera di gran lunga il baffo inferiore.

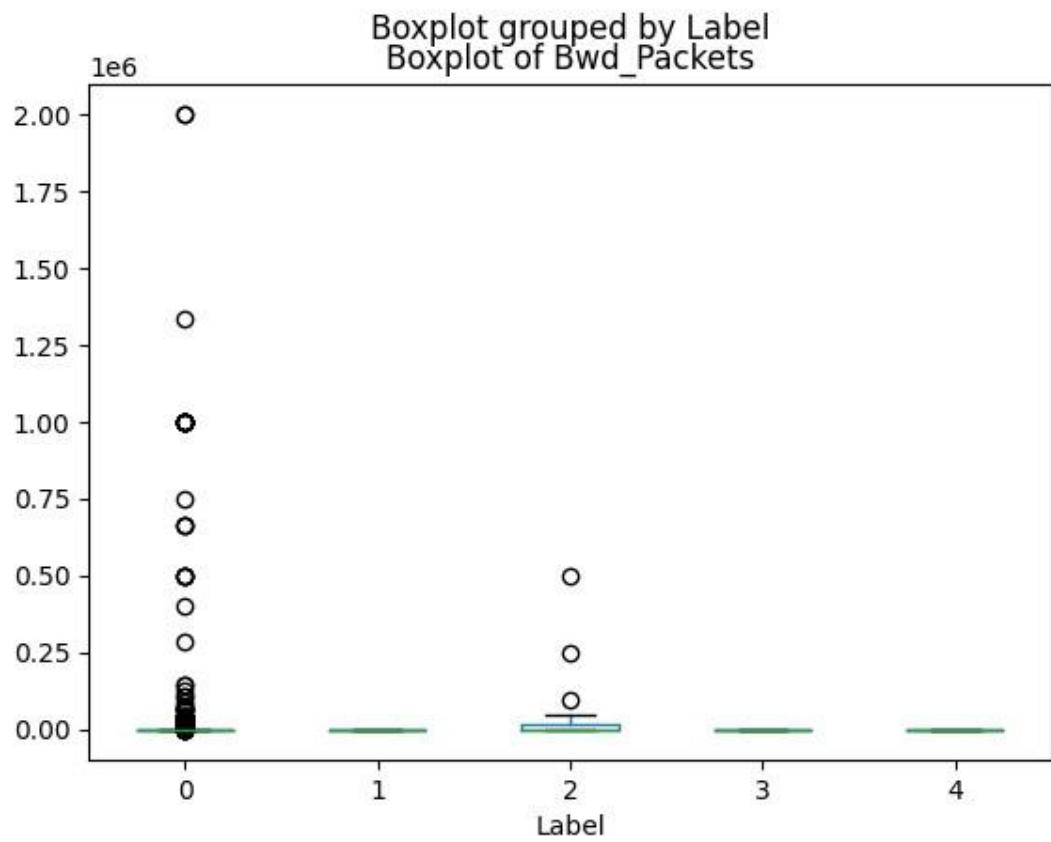
4.34 Fwd_Packets



- count 1.000000e+04
- mean 8.716025e+05
- std 9.478612e+05
- min 7.475528e-02
- 25% 7.739876e+01
- 50% 4.255319e+04
- 75% 2.000000e+06
- max 3.000000e+06

L'unica classe a presentare outlier è la classe 0 con valore max 3.000.000

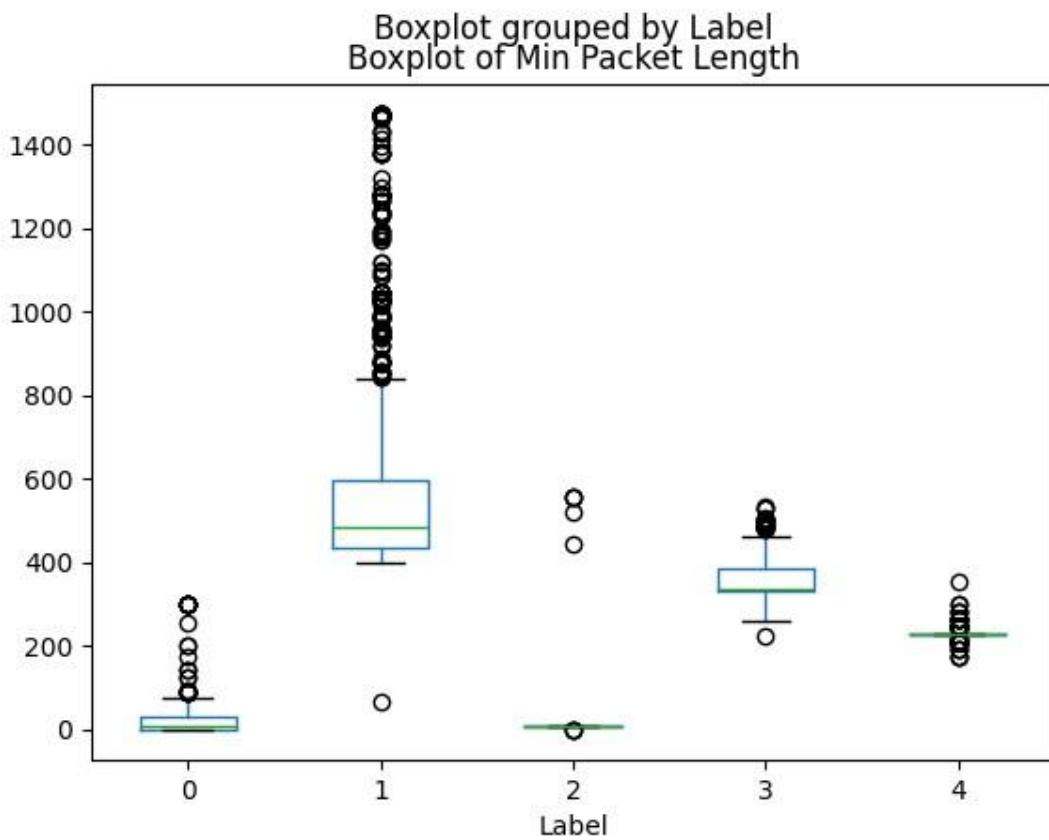
4.35 Bwd_Packets



- count 1.000000e+04
- mean 4.602816e+03
- std 5.160721e+04
- min 0.000000e+00
- 25% 0.000000e+00
- 50% 0.000000e+00
- 75% 1.103628e+00
- max 2.000000e+06

Classe 0 e 2 presentano outlier superiori.

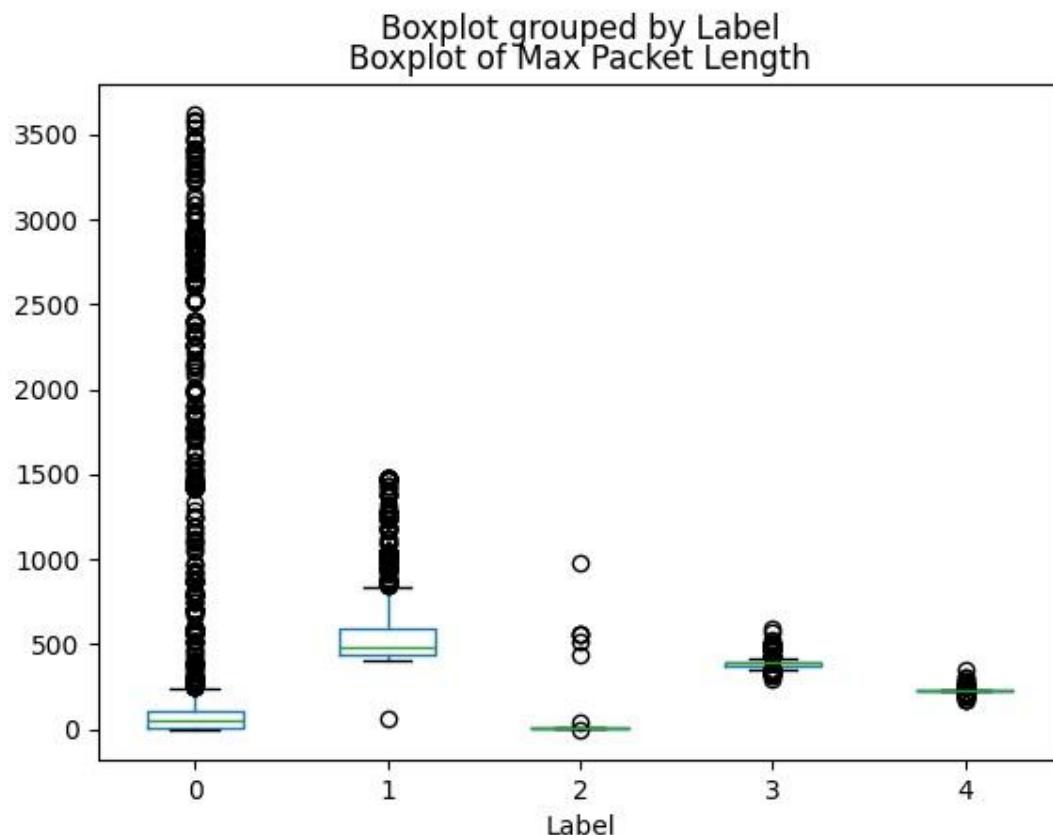
4.36 Min Packet Length



- count 10000.000000
- mean 213.236100
- std 244.527403
- min 0.000000
- 25% 6.000000
- 50% 211.000000
- 75% 383.000000
- max 1472.000000

Tutte e 5 le classi presentano outlier, in particolare Classe 1,3 e 4 presentano anche outlier inferiori.

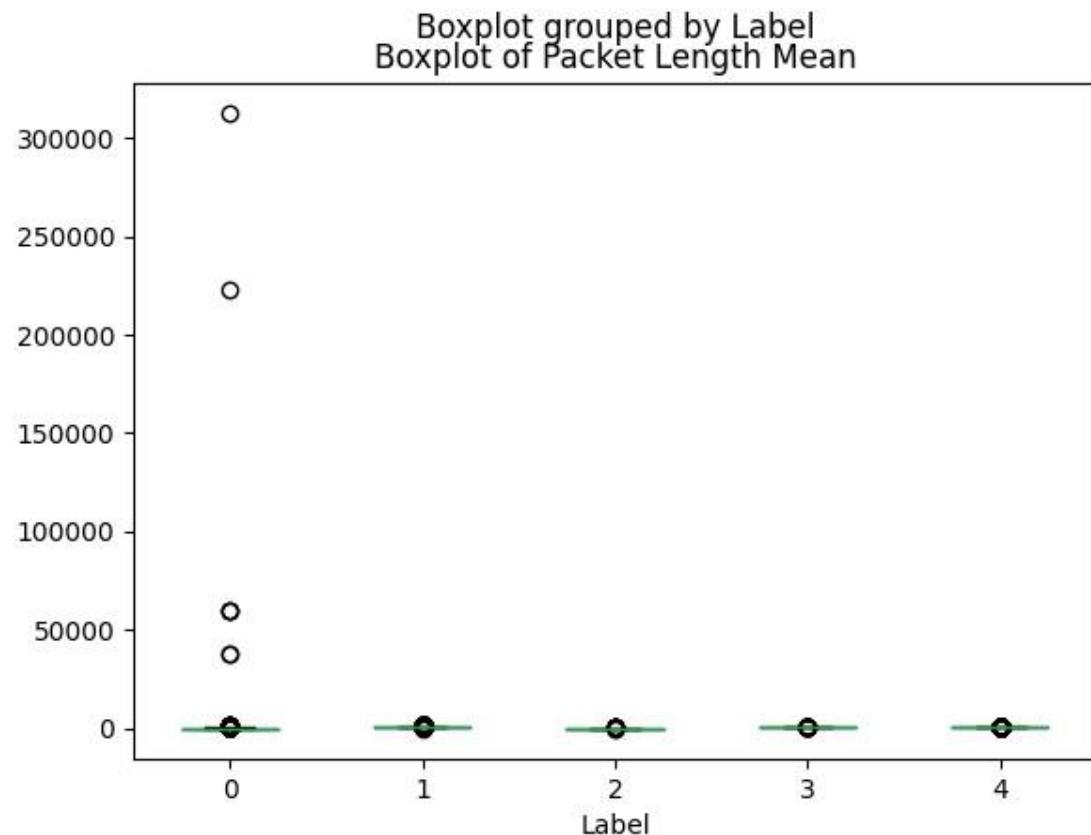
4.37 Max Packet Length



- count 10000.000000
- mean 325.721700
- std 484.730972
- min 0.000000
- 25% 6.000000
- 50% 229.000000
- 75% 405.000000
- max 3617.000000

Tutte e 5 le classi presentano outlier, in particolare Classe 1, 2, 3 e 4 presentano anche outlier inferiori.

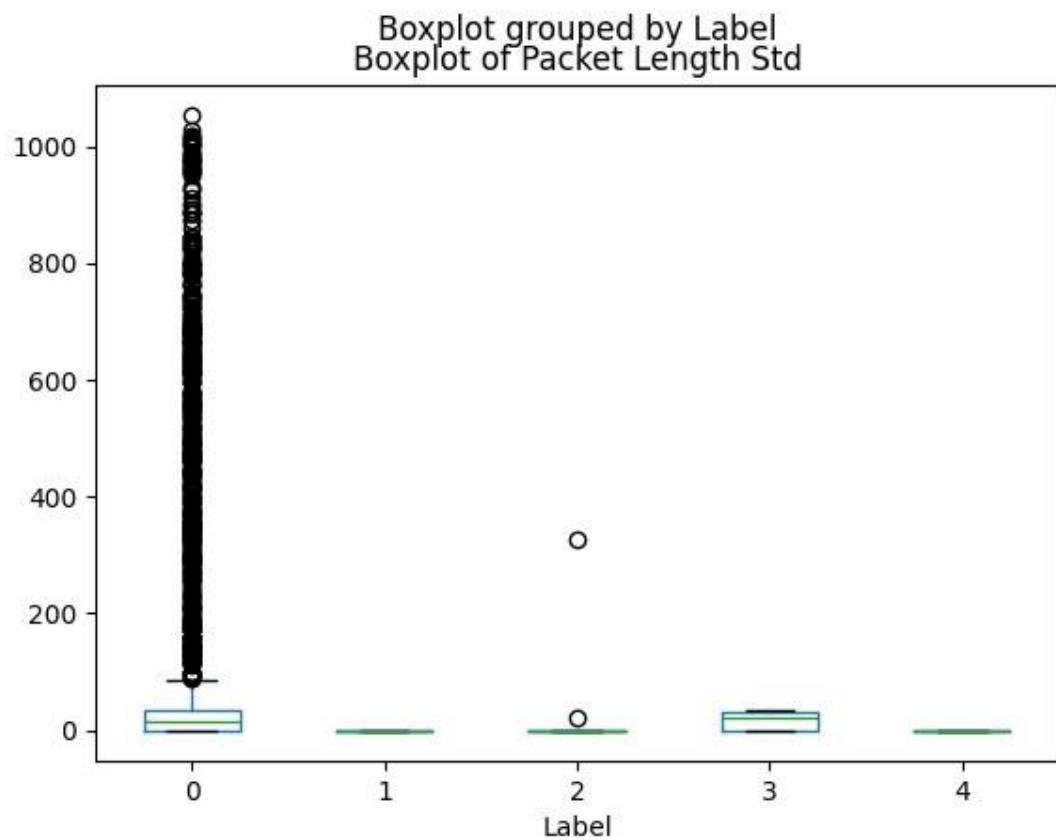
4.38 Packet Length Mean



- count 10000.000000
- mean 314.810701
- std 4009.218666
- min 0.000000
- 25% 6.000000
- 50% 229.000000
- 75% 401.000000
- max 312375.000000

Tutte e cinque le classi presentano outlier, il valore massimo di 312.375 supera
enormemente il baffo superiore per la classe 0.

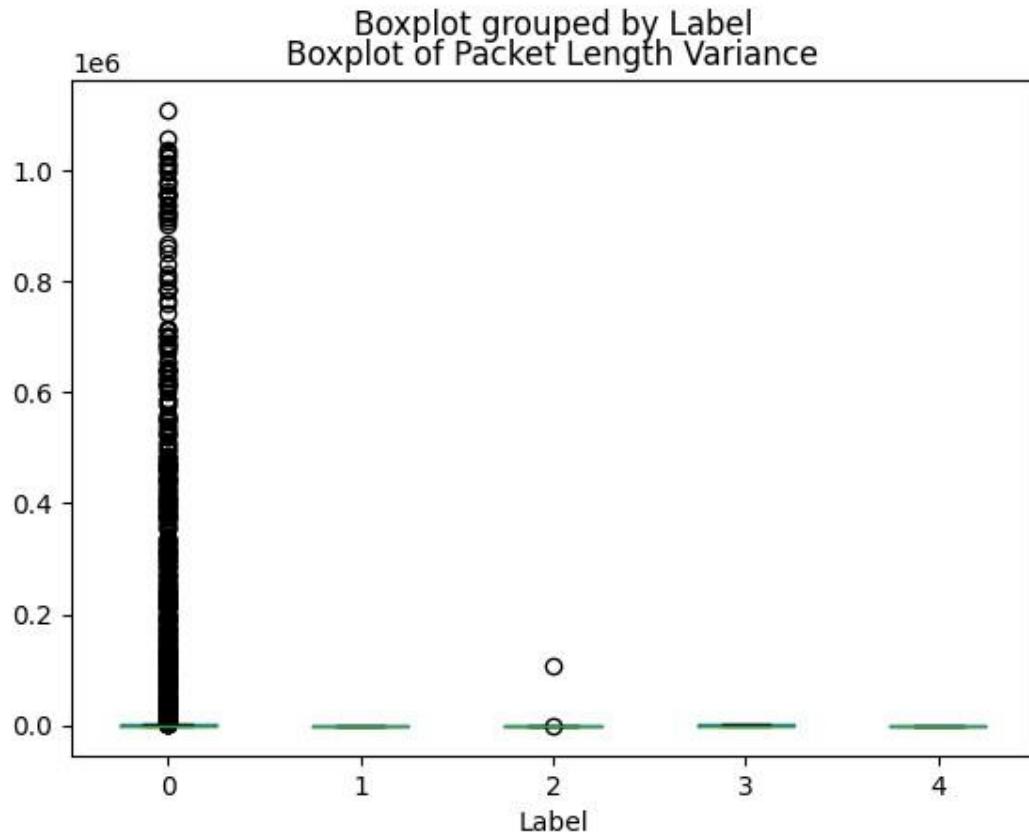
4.39 Packet Length Std



- count 10000.000000
- mean 32.782734
- std 123.568815
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 14.433757
- max 1052.315670

Classe 0 e 2 presentano outlier, il valore massimo di 1.052 supera enormemente il baffo superiore per la classe 0.

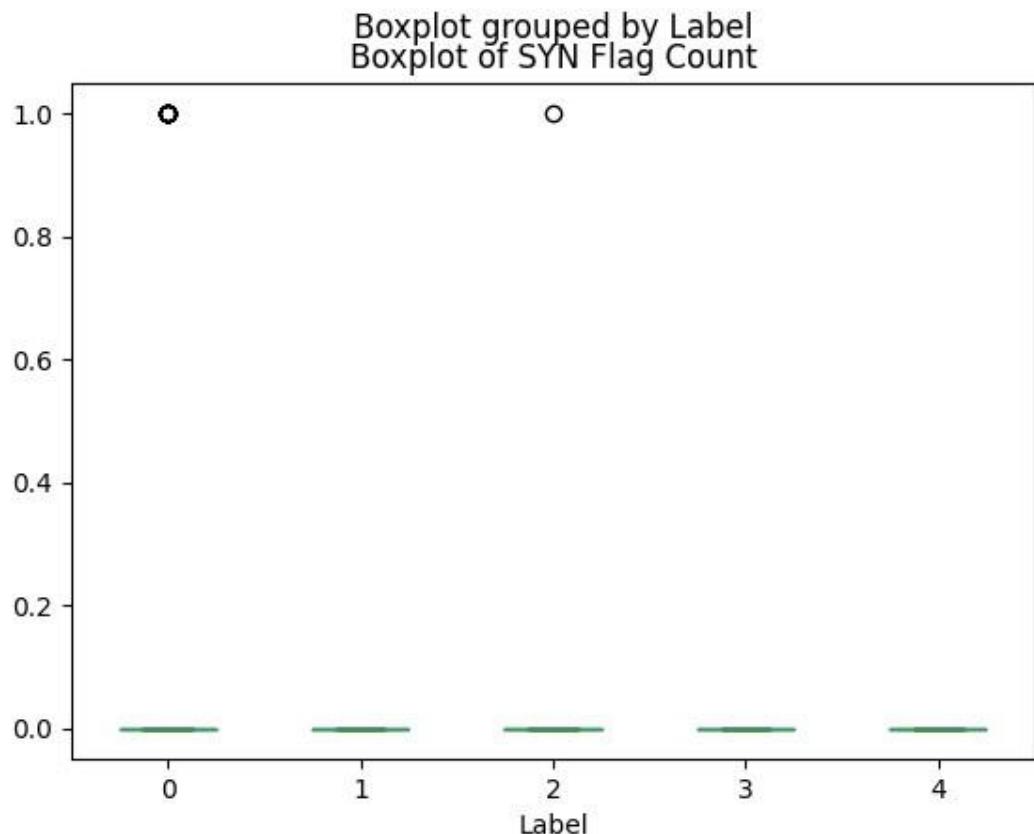
4.40 Packet Length Variance



- count 1.000000e+04
- mean 1.634243e+04
- std 9.071925e+04
- min 0.000000e+00
- 25% 0.000000e+00
- 50% 0.000000e+00
- 75% 2.083333e+02
- max 1.107368e+06

Classe 0 e 2 presentano outlier, il valore massimo di 1.107.368 supera enormemente il baffo superiore per la classe 0.

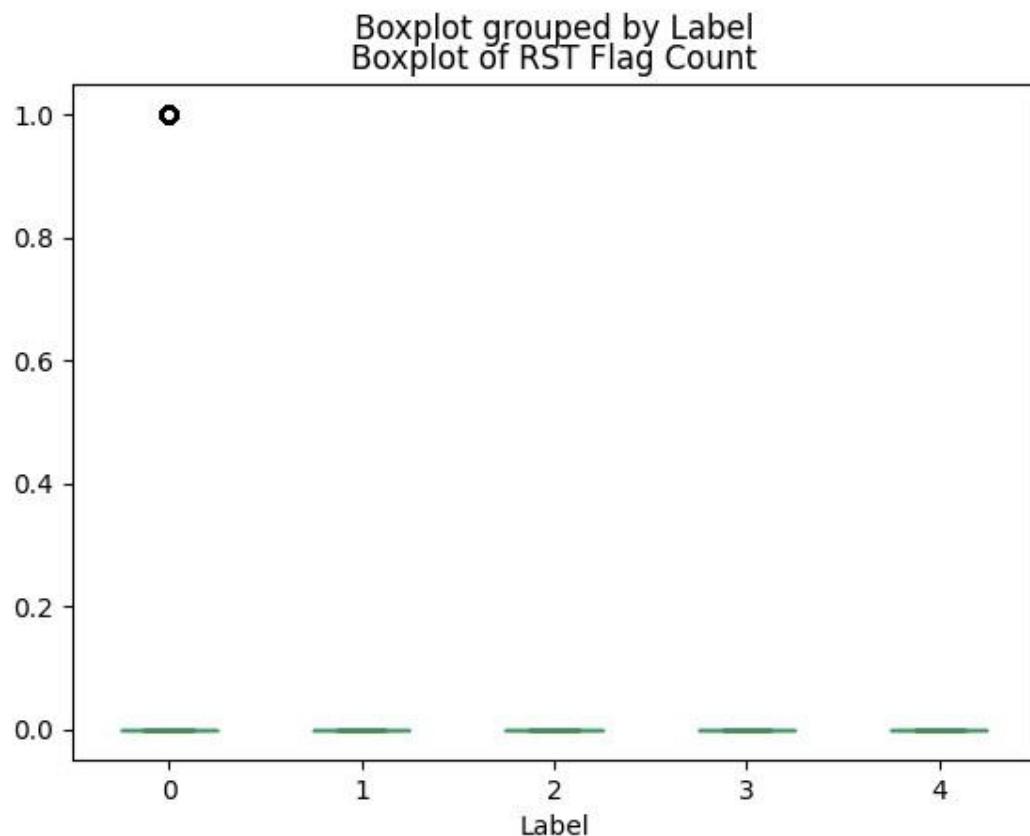
4.41 SYN Flag Count



- count 10000.000000
- mean 0.001300
- std 0.036034
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 0.000000
- max 1.000000

Dalle analisi emerge che la feature contiene valori booleani (0 e 1), assume 1 solo per Classe 0 e 2.

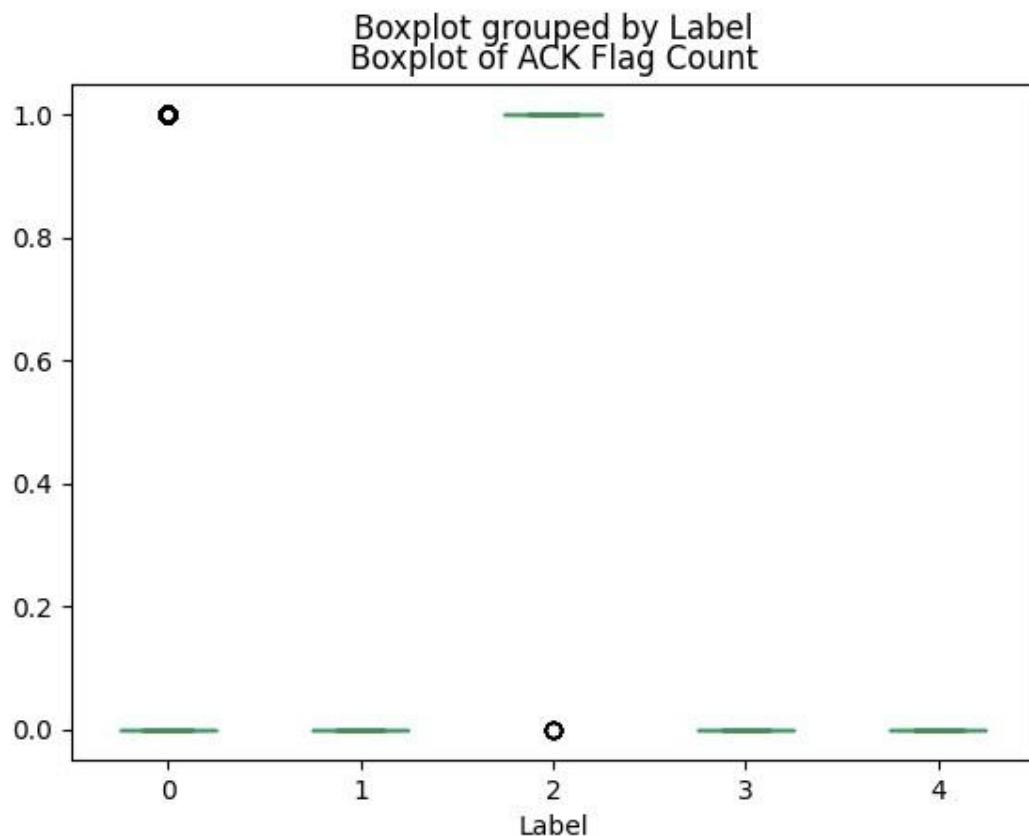
4.42 RST Flag Count



- count 10000.000000
- mean 0.055800
- std 0.229547
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 0.000000
- max 1.000000

Dalle analisi emerge che la feature contiene valori booleani (0 e 1), assume 1 solo per Classe 0.

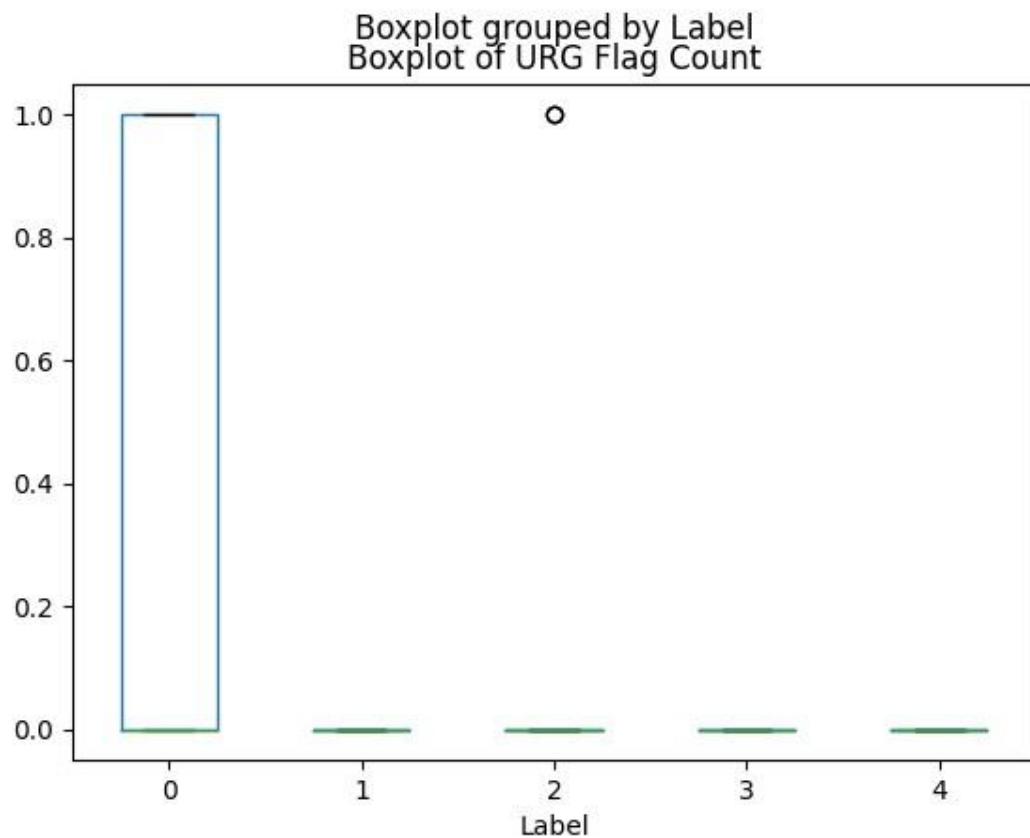
4.43 ACK Flag Count



- count 10000.000000
- mean 0.261200
- std 0.439311
- min 0.000000
- 25% 0.000000 • 50% 0.000000
- 75% 1.000000
- max 1.000000

Dalle analisi emerge che la feature contiene valori booleani (0 e 1), assume 1 di frequente per la classe 2 e di rado per la classe 0.

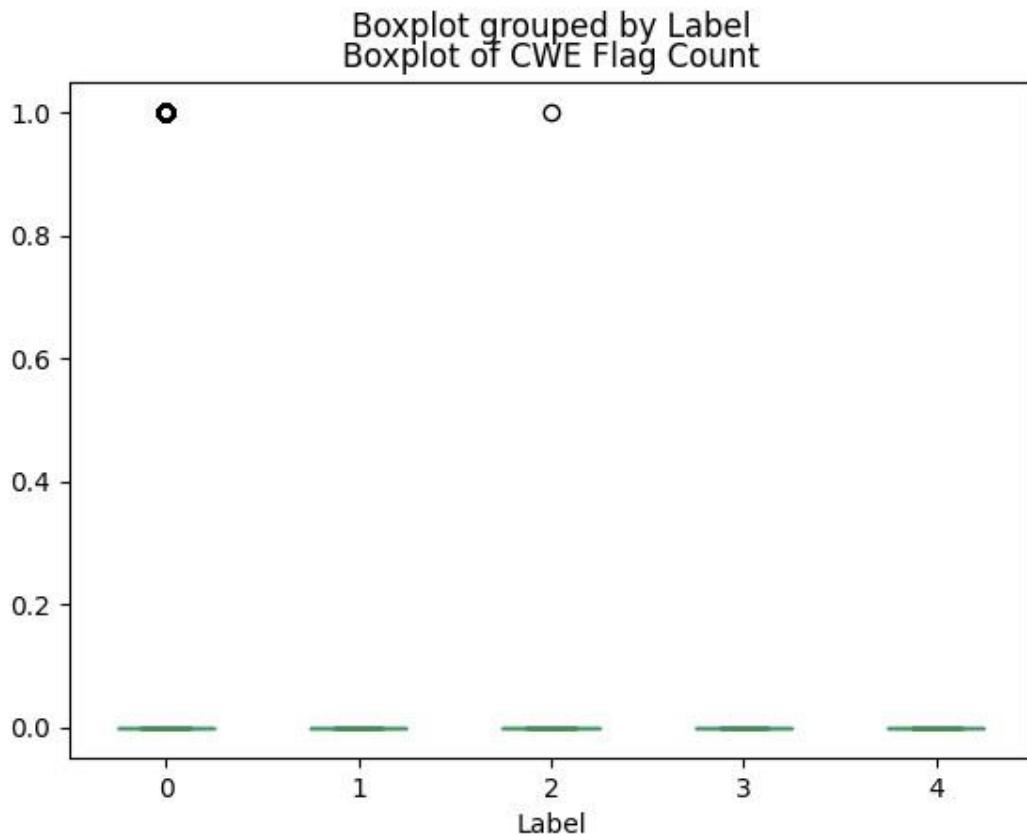
4.44 URG Flag Count



- count 10000.000000
- mean 0.148600
- std 0.355712
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 0.000000
- max 1.000000

Dalle analisi emerge che la feature contiene valori booleani (0 e 1), assume 1 solo per Classe 0 e 2.

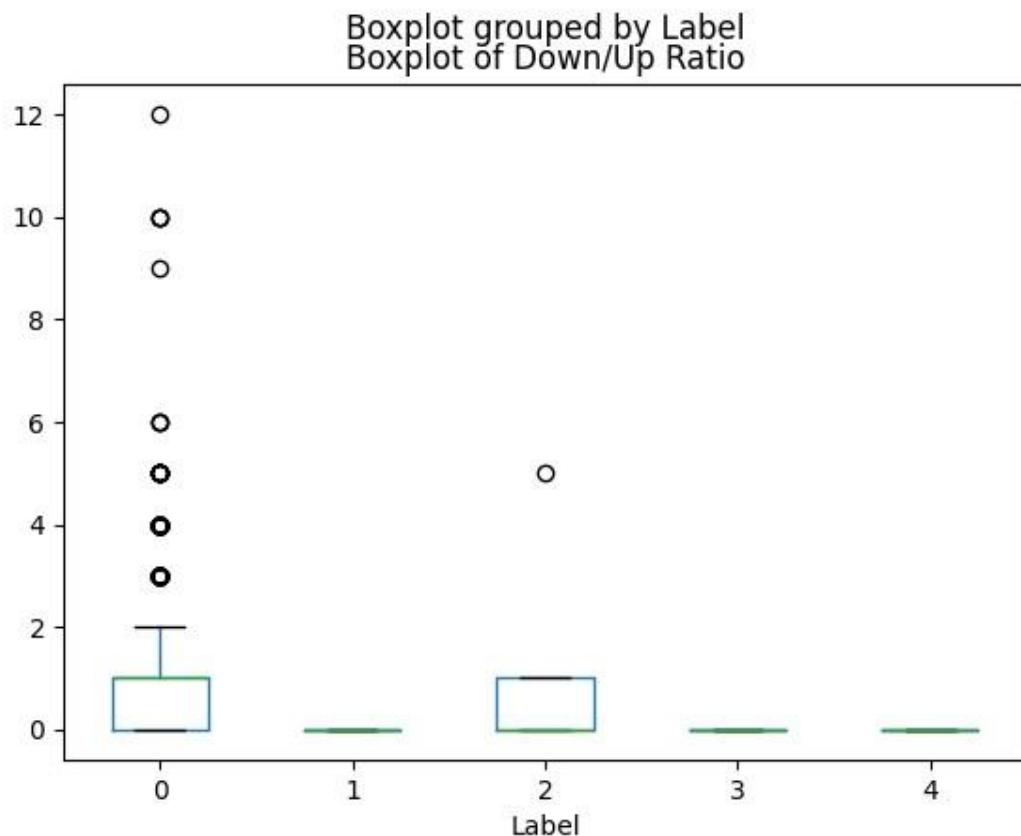
4.45 CWE Flag Count



- count 10000.00000
- mean 0.06370
- std 0.24423
- min 0.00000
- 25% 0.00000
- 50% 0.00000
- 75% 0.00000
- max 1.00000

Dalle analisi emerge che la feature contiene valori booleani (0 e 1), assume 1 solo per Classe 0 e 2.

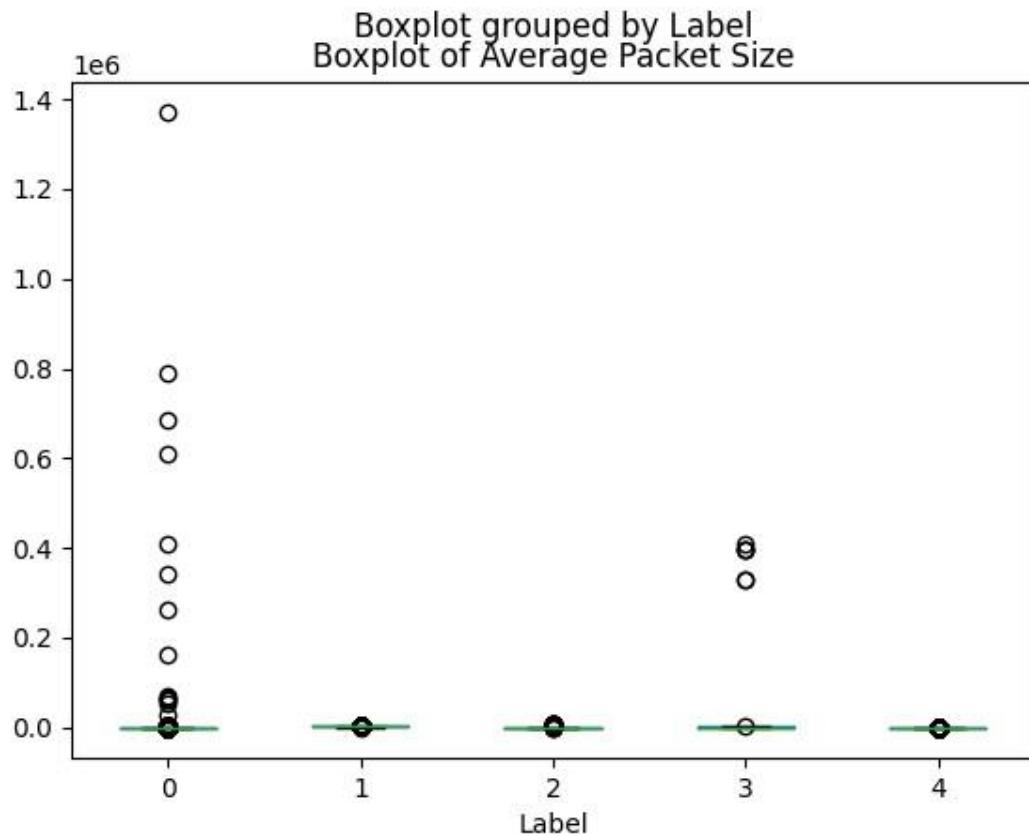
4.46 Down/Up Ratio



- count 10000.000000
- mean 0.318100
- std 0.702255
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 0.000000
- max 12.000000

Classe 0 e 2 presentano outlier, il valore massimo di 12 supera enormemente il baffo superiore per la classe 0.

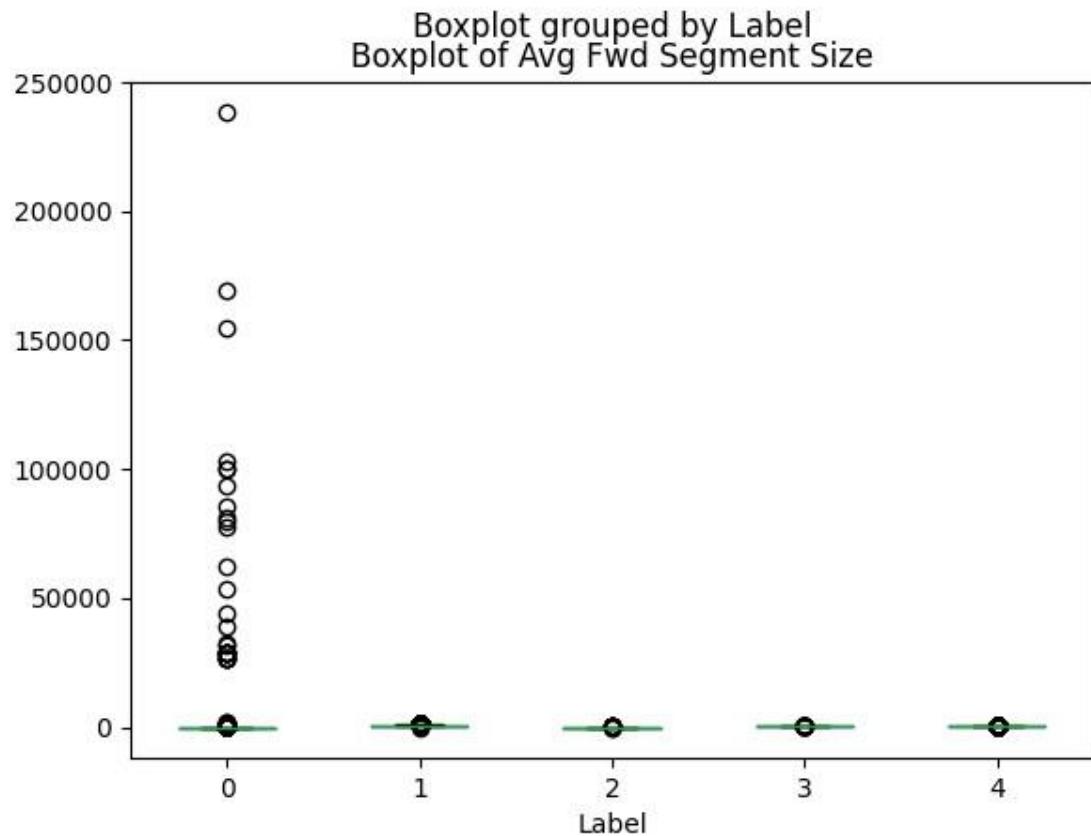
4.47 Average Packet Size



- count $1.000000e+04$
- mean $1.071255e+03$
- std $2.139569e+04$
- min $0.000000e+00$
- 25% $9.000000e+00$
- 50% $3.435000e+02$
- 75% $5.970000e+02$
- max $1.369875e+06$

Tutte e cinque le classi presentano outlier, nello specifico la Classe 0 ha outlier (valore max 1.369.875) che superano di gran lunga i valori più frequenti.

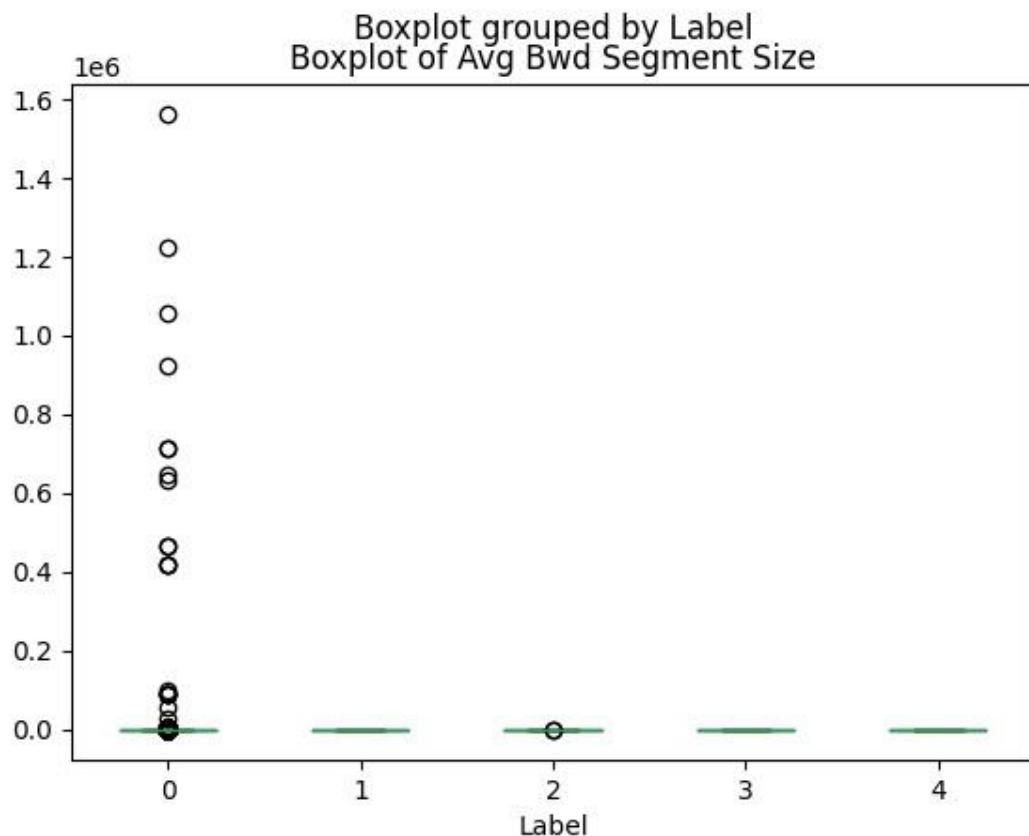
4.48 Avg Fwd Segment Size



- count 10000.000000
- mean 413.805489
- std 4431.180017
- min 0.000000
- 25% 6.000000
- 50% 229.000000
- 75% 383.000000
- max 238125.000000

Tutte e cinque le classi presentano outlier, nello specifico la Classe 0 ha outlier (valore max 238.125) che superano di gran lunga i valori più frequenti.

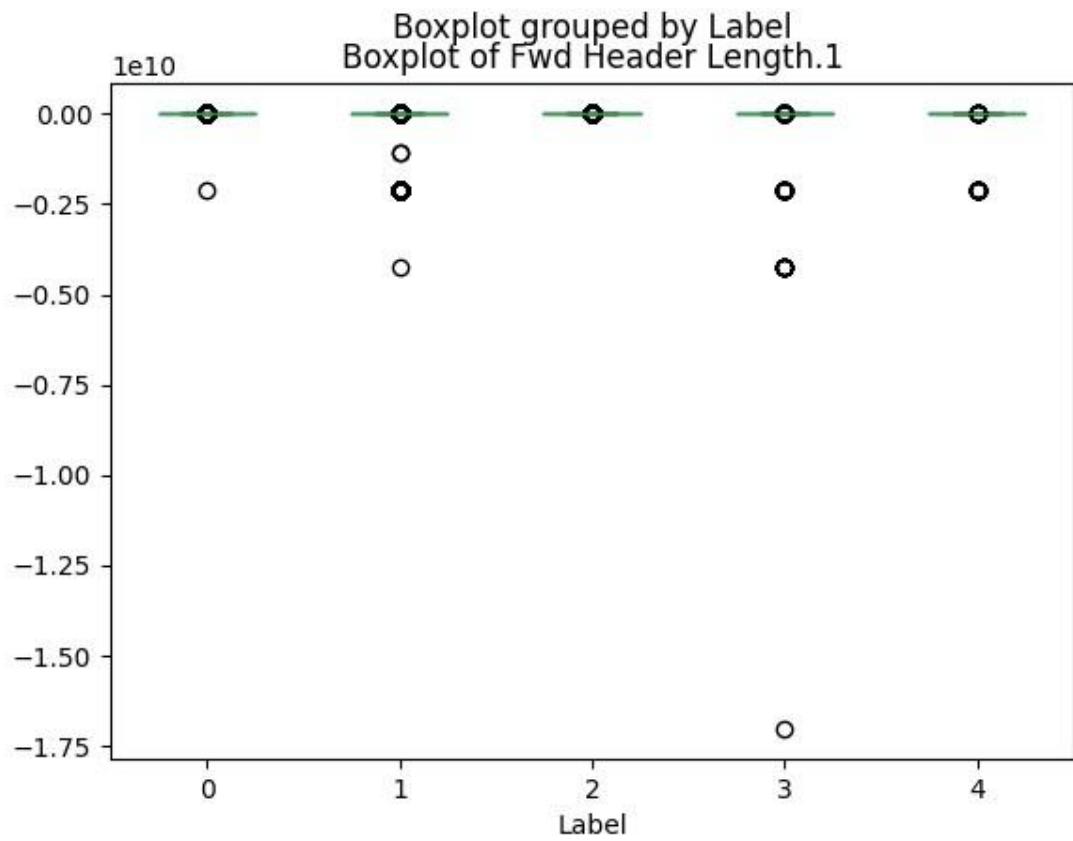
4.49 Avg Bwd Segment Size



- count 1.000000e+04
- mean 1.052065e+03
- std 2.956934e+04
- min 0.000000e+00
- 25% 0.000000e+00
- 50% 0.000000e+00
- 75% 6.000000e+00
- max 1.560784e+06

Classe 0 e 2 presentano outlier, nello specifico la Classe 0 ha outlier (valore max 1.560.784) che superano di gran lunga i valori più frequenti.

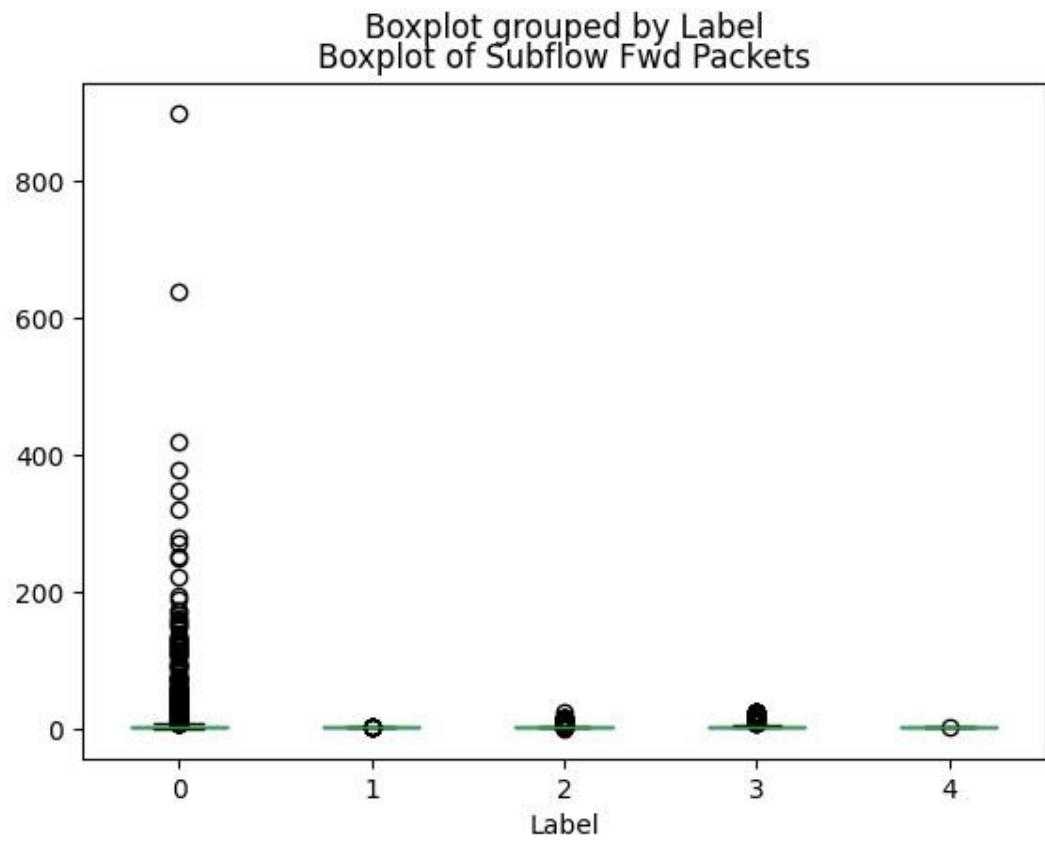
4.50 Fwd Header Length.1



- count 1.000000e+04
- mean -6.057485e+07
- std 4.012595e+08
- min -1.700350e+10
- 25% 2.800000e+01
- 50% 4.000000e+01
- 75% 6.400000e+01
- max 2.878400e+04

Tutte e cinque le classi presentano outlier, il valore minimo di -17mld supera
enormemente il baffo superiore per la classe 3.

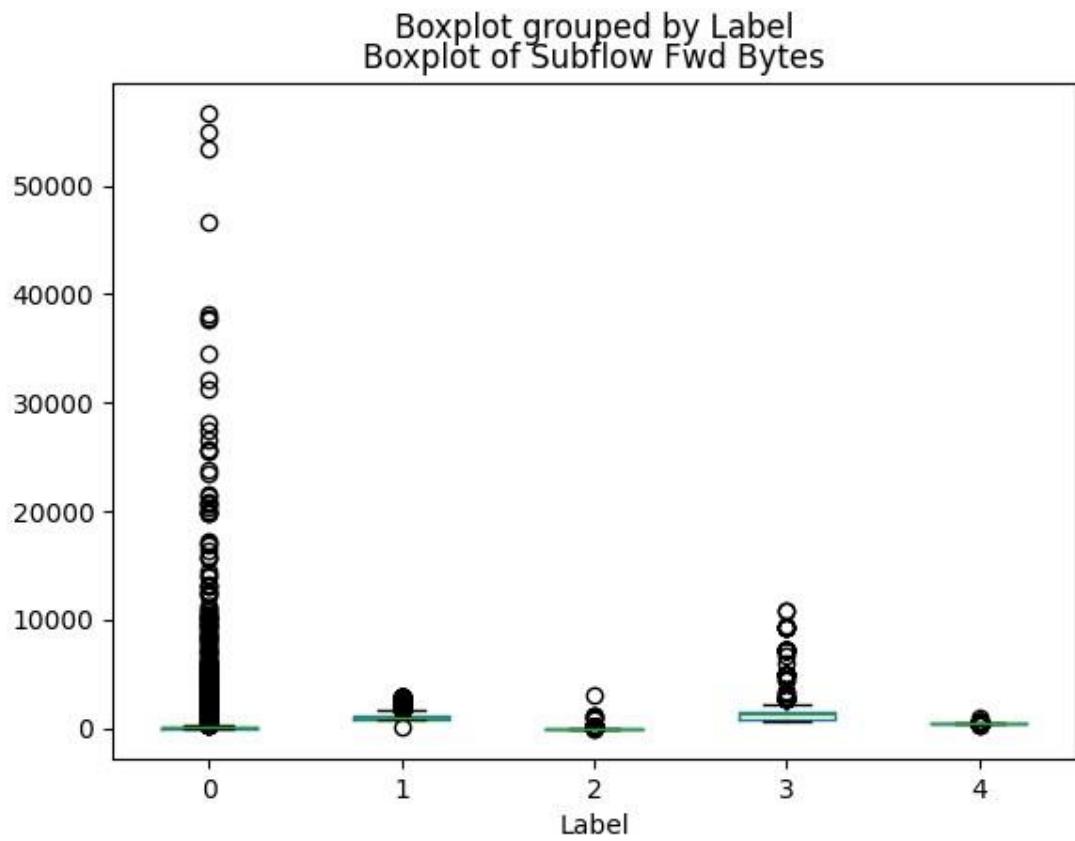
4.51 Subflow Fwd Packets



- count 10000.000000
- mean 4.974100
- std 17.905778
- min 1.000000
- 25% 2.000000
- 50% 2.000000
- 75% 2.000000
- max 899.000000

Tutte e cinque le classi presentano outlier, il valore massimo di 899 supera
enormemente il baffo superiore per la classe 0.

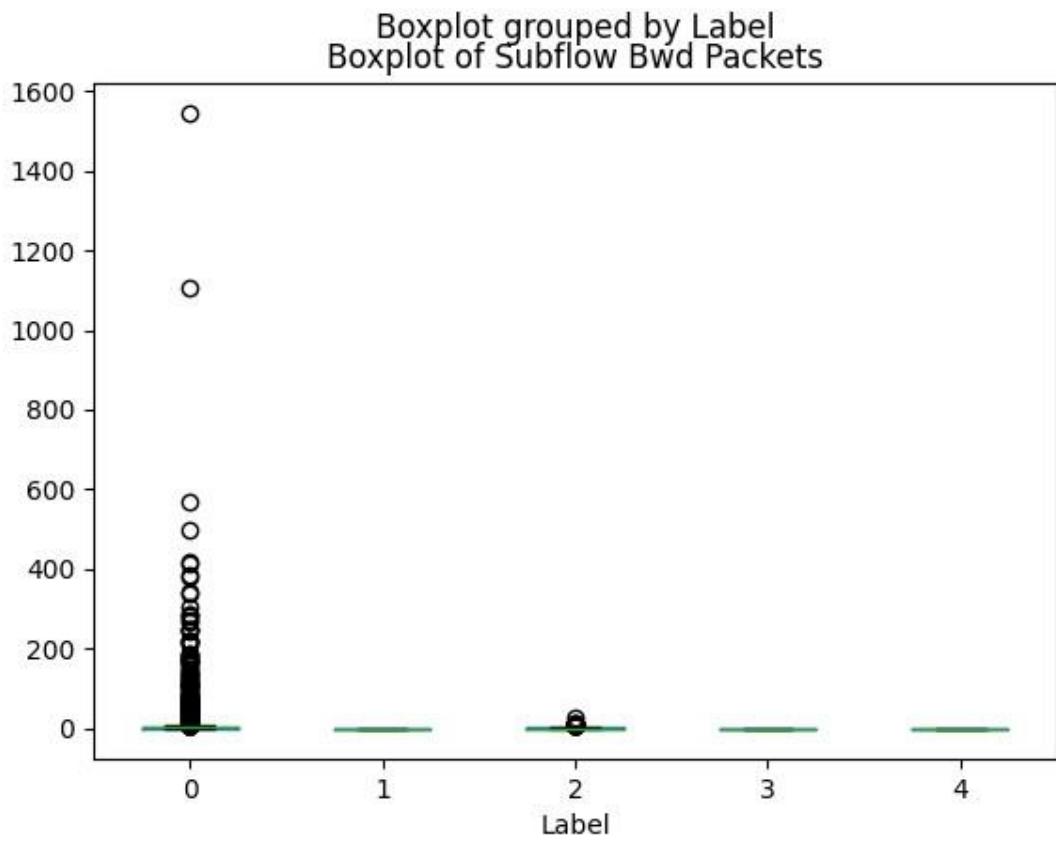
4.52 Subflow Fwd Bytes



- count 10000.000000
- mean 832.031800
- std 2049.949291
- min 0.000000
- 25% 12.000000
- 50% 458.000000
- 75% 1000.000000
- max 56622.000000

Tutte e cinque le classi presentano outlier, il valore massimo di 56.622 supera
enormemente il baffo superiore per la classe 0.

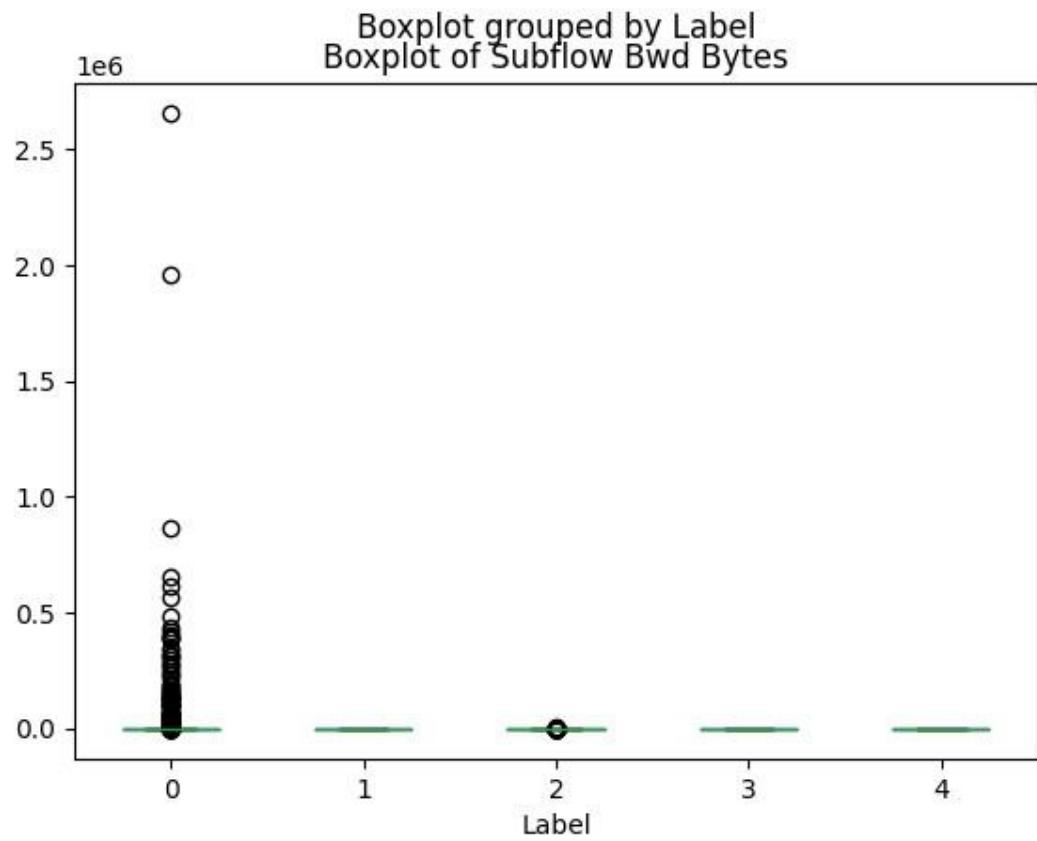
4.53 Subflow Bwd Packets



- count 10000.000000
- mean 3.349400
- std 26.611075
- min 0.000000
- 25% 0.000000
- 50% 0.000000
- 75% 2.000000
- max 1543.000000

Classe 0 e 2 presentano outlier, nello specifico la Classe 0 ha outlier (valore max 1.543) che superano di gran lunga il valore più frequente (compreso tra 0 e 2).

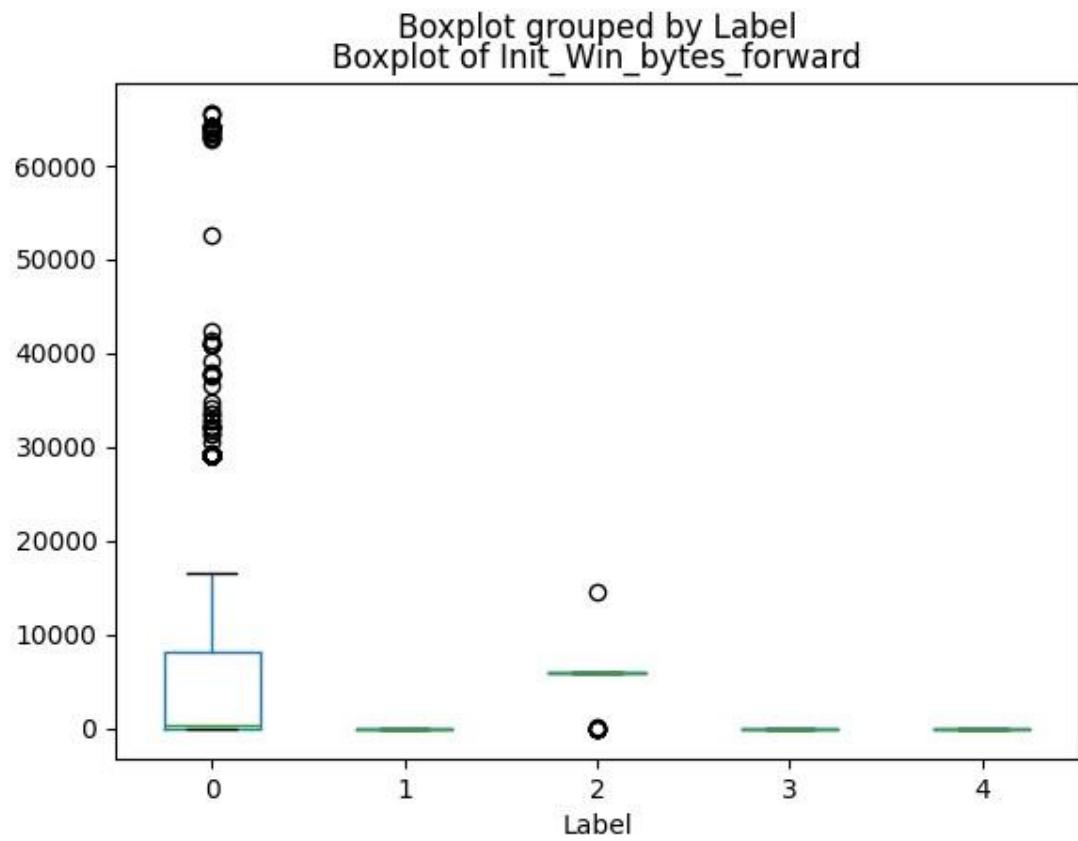
4.54 Subflow Bwd Bytes



- 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 $2.655090e+06$

Classe 0 e 2 presentano outlier, nello specifico la Classe 0 ha outlier (max 2.655.090) che superano di gran lunga il valore più frequente (compreso tra 0 e 12).

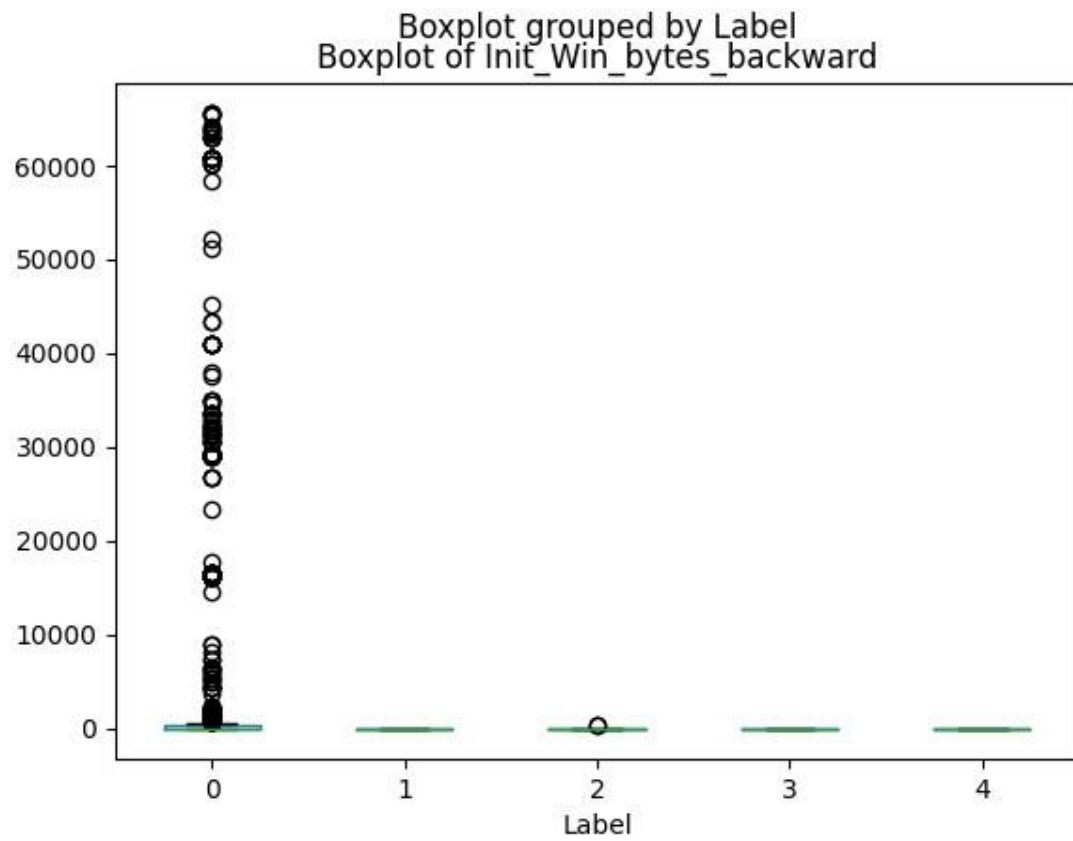
4.55 Init_Win_bytes_forward



- count 10000.000000
- mean 2908.308800
- std 6525.600525
- min -1.000000
- 25% -1.000000
- 50% -1.000000
- 75% 5840.000000
- max 65535.000000

Classe 0 presenta outlier che superano enormemente il baffo superiore, Classe 2 invece presenta outlier sia nella parte alta che nella parte bassa.

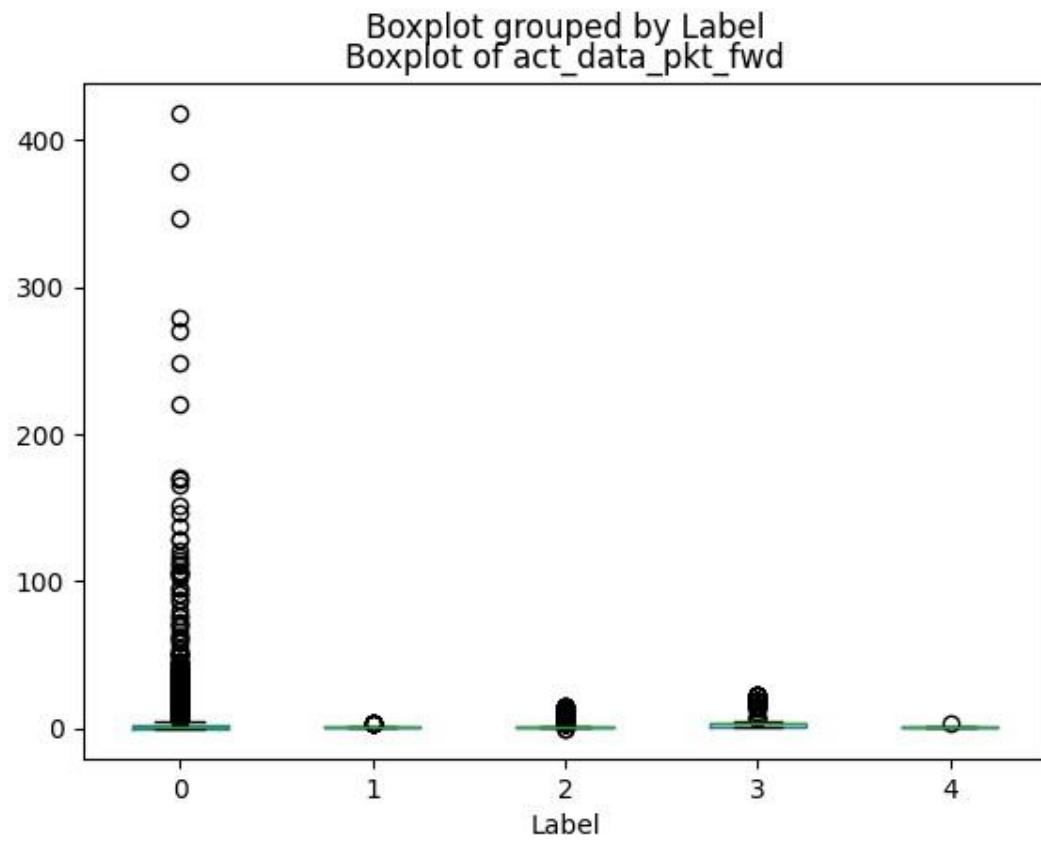
4.56 Init_Win_bytes_backward



- count 10000.000000
- mean 830.581600
- std 5503.753359
- min -1.000000
- 25% -1.000000
- 50% -1.000000
- 75% -1.000000
- max 65535.000000

Classe 0 e 2 presentano outlier, nello specifico la Classe 0 ha outlier (max 65.535) che superano di gran lunga il valore più frequente (-1).

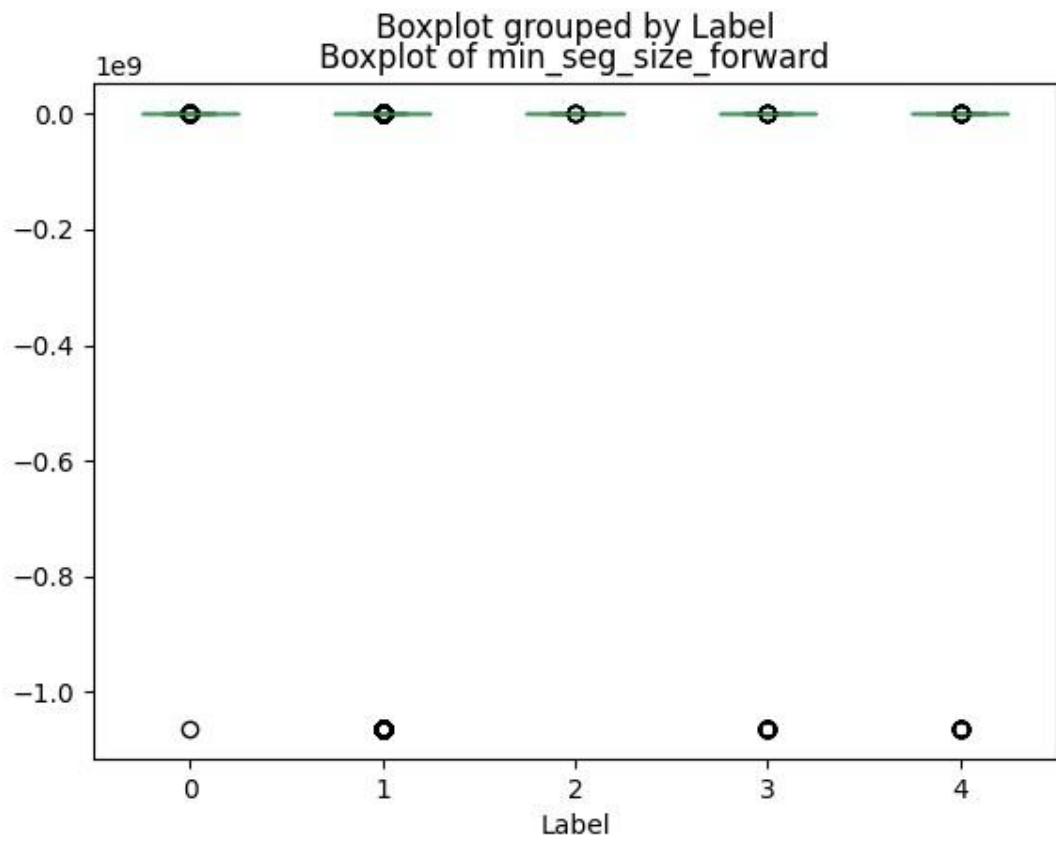
4.57 act_data_pkt_fwd



- count 10000.000000
- mean 2.924500
- std 11.476444
- min 0.000000
- 25% 1.000000
- 50% 1.000000
- 75% 1.000000
- max 418.000000

Tutte e cinque le classi presentano outlier, nello specifico la Classe 0 ha outlier che superano di gran lunga il valore più frequente (1).

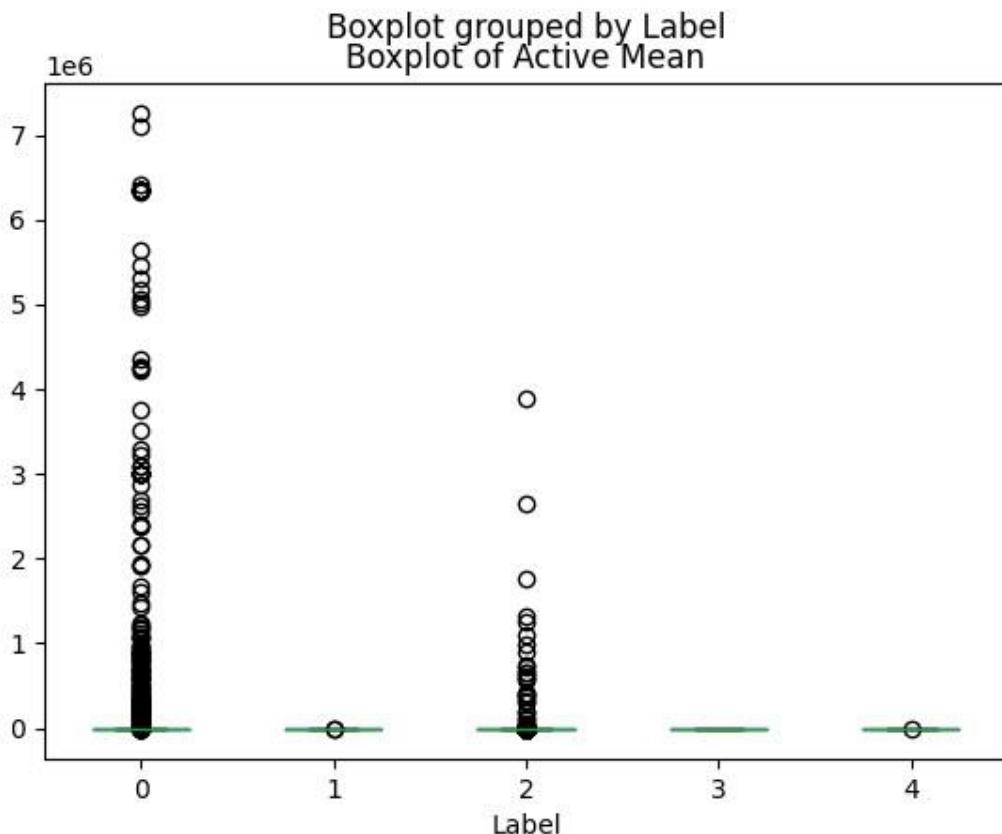
4.58 min_seg_size_forward



- count 1.000000e+04
- mean -2.837457e+07
- std 1.713244e+08
- min -1.062719e+09
- 25% 8.000000e+00
- 50% 2.000000e+01
- 75% 2.000000e+01
- max 1.480000e+03

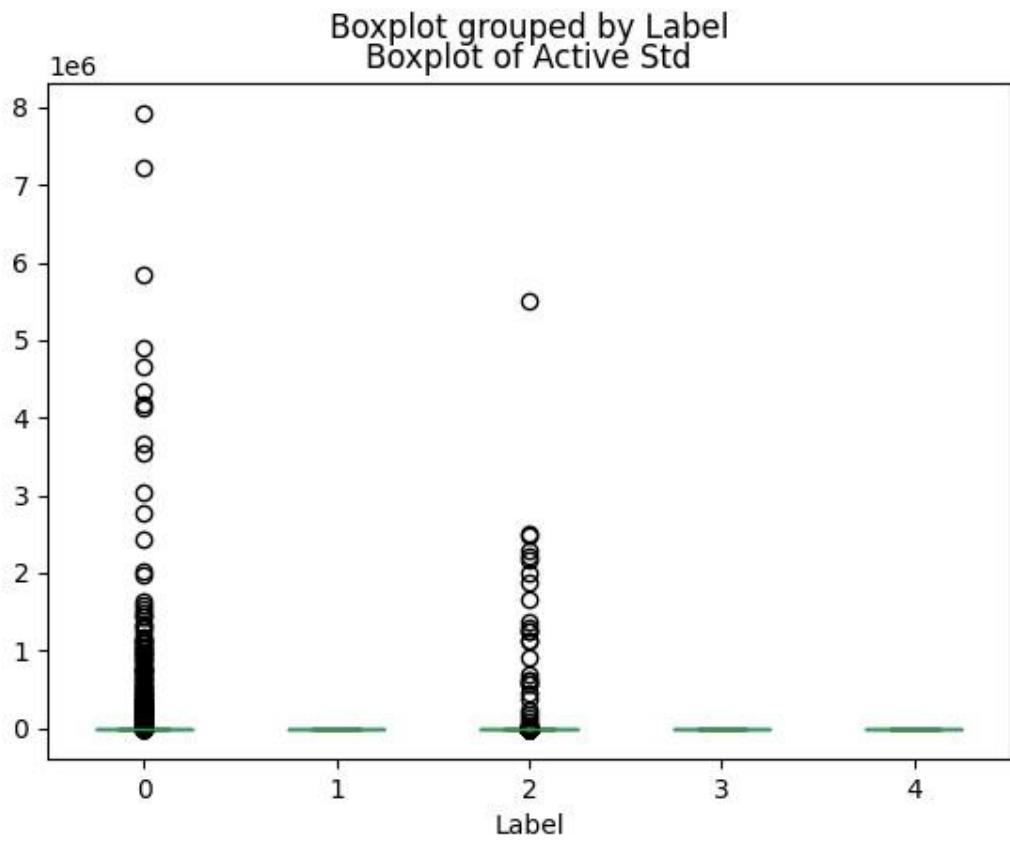
Tutte le classi presentano degli outlier, in particolare Classe 0, 1, 3 e 4 presentano outlier che superano enormemente il baffo inferiore (<-1mld).

4.59 Active Mean



Classe 0,1,2 e 4 presentano outlier, in particolare Classe 0 e 2 hanno Outlier che superano di gran lunga il baffo superiore.

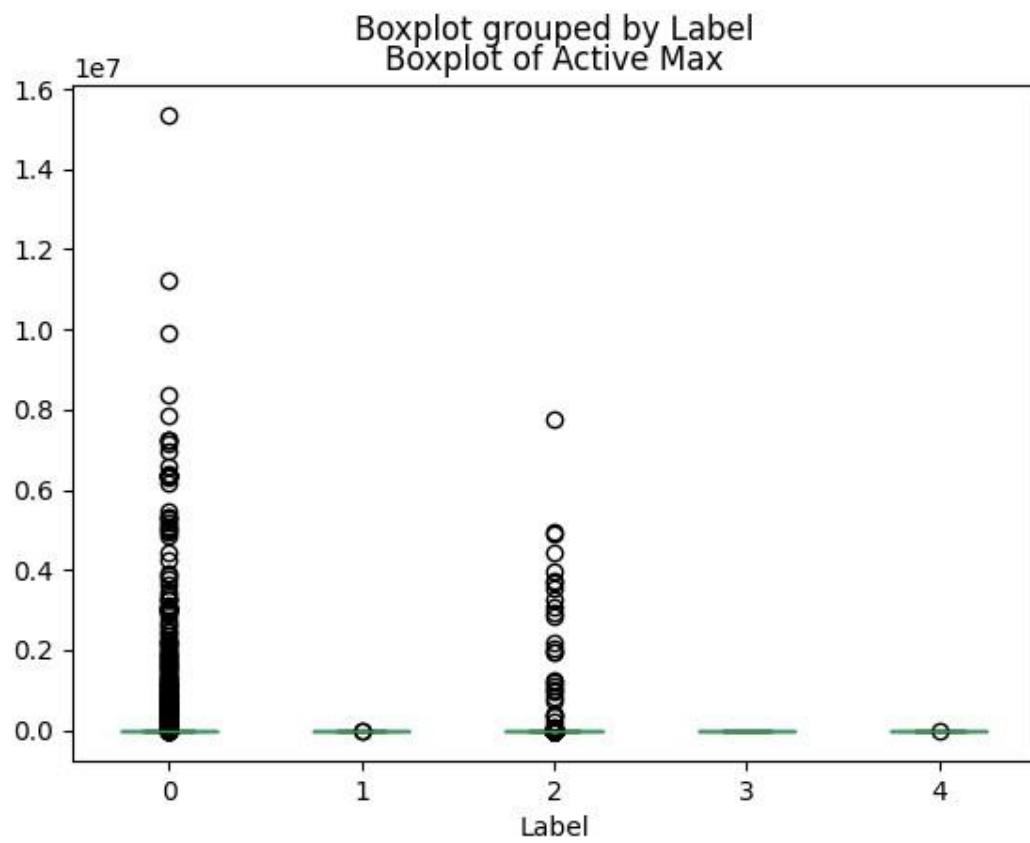
4.60 Active Std



- count 1.000000e+04
- mean 1.752317e+04
- std 2.095494e+05
- min 0.000000e+00
- 25% 0.000000e+00
- 50% 0.000000e+00
- 75% 0.000000e+00
- max 7.914871e+06

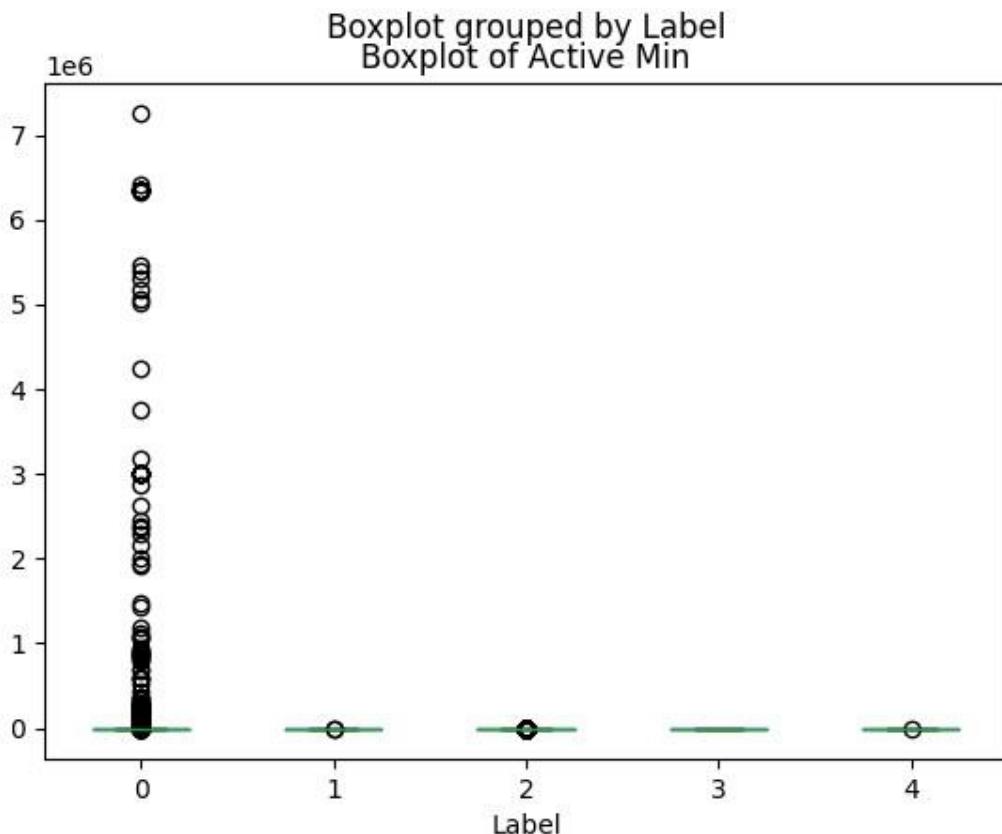
Classe 0 e 2 presentano outlier che superano di gran lunga il baffo superiore.

4.61 Active Max



Classe 0,1,2 e 4 presentano outlier, in particolare Classe 0 e 2 hanno outlier che superano di gran lunga il baffo superiore.

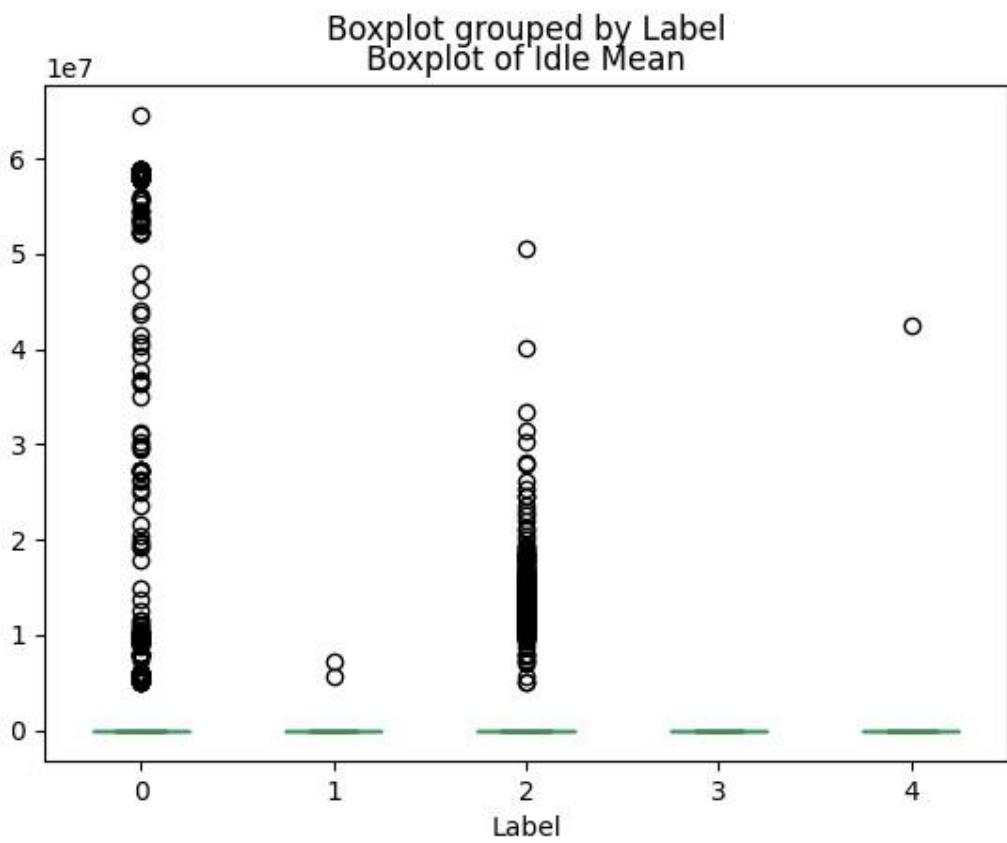
4.62 Active Min



- count 1.000000e+04
- mean 2.347295e+04
- std 2.964444e+05
- min 0.000000e+00
- 25% 0.000000e+00
- 50% 0.000000e+00
- 75% 0.000000e+00
- max 7.247923e+06

Classe 0,1,2 e 4 presentano outlier, in particolare Classe 0 ha outlier che superano di gran lunga il baffo superiore.

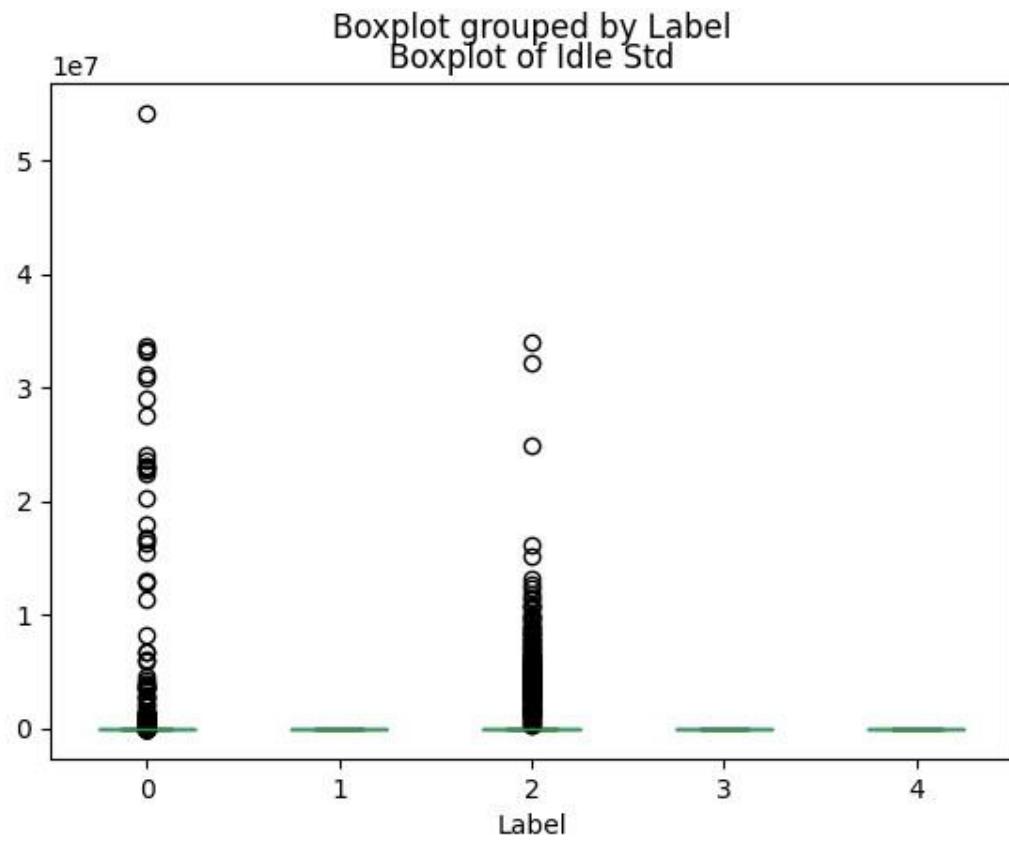
4.63 Idle Mean



- count 1.000000e+04
- mean 1.700478e+06
- std 8.205343e+06
- min 0.000000e+00
- 25% 0.000000e+00
- 50% 0.000000e+00
- 75% 0.000000e+00
- max 6.451177e+07

Classe 0,1,2 e 4 presentano outlier, in particolare Classe 0, 2 e 4 hanno outlier che superano di gran lunga il baffo superiore.

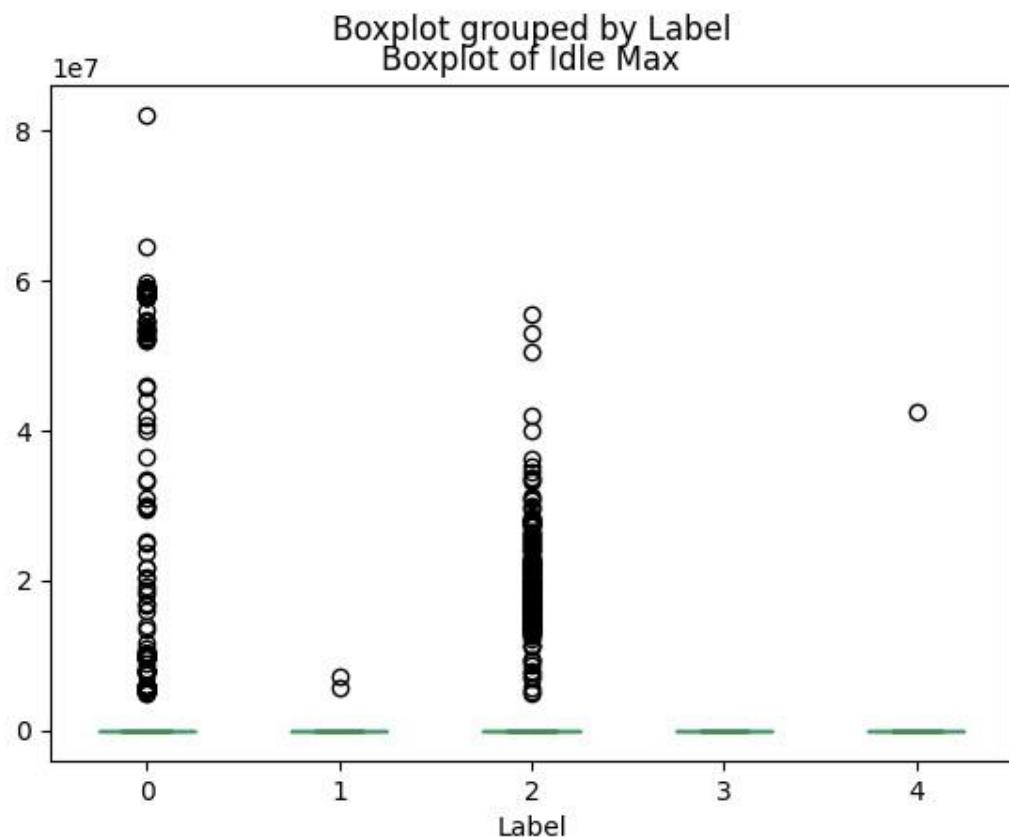
4.64 Idle Std



- count 1.000000e+04
- mean 1.801478e+05
- std 1.583956e+06
- min 0.000000e+00
- 25% 0.000000e+00
- 50% 0.000000e+00
- 75% 0.000000e+00
- max 5.407594e+07

Classe 0 e 2 presentano outlier che superano di gran lunga il baffo superiore.

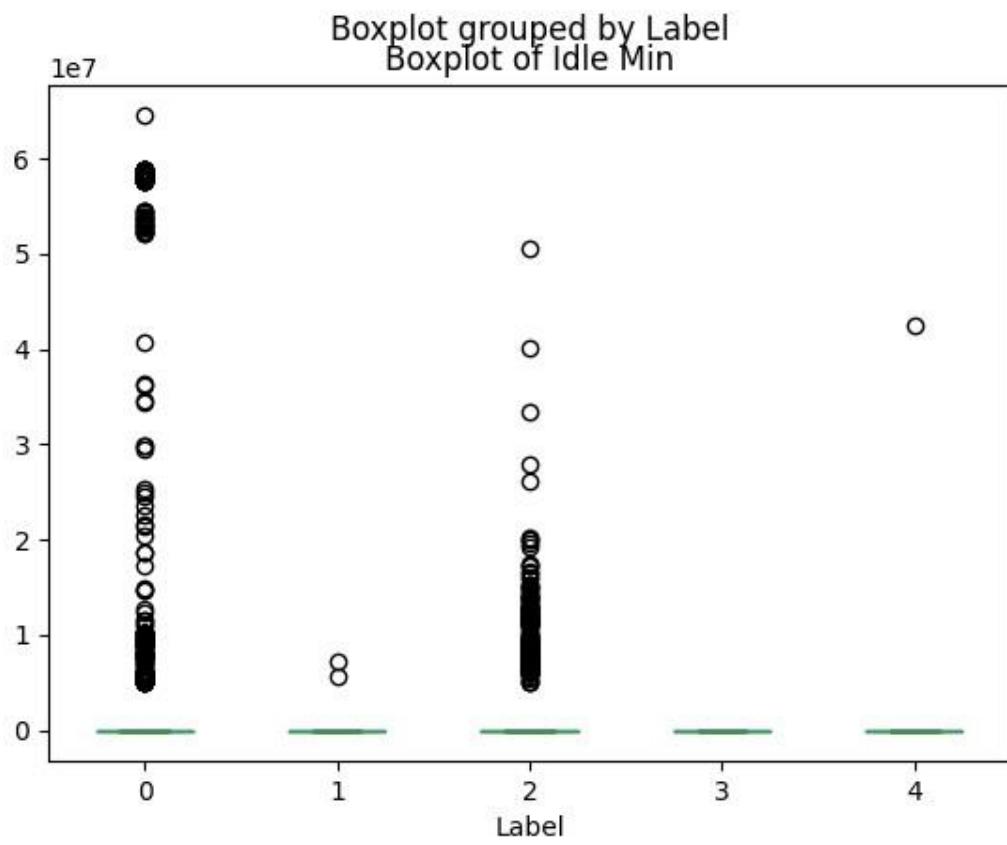
4.65 Idle Max



- count $1.000000\text{e+}04$
- mean $1.876333\text{e+}06$
- std $8.740673\text{e+}06$
- min $0.000000\text{e+}00$
- 25% $0.000000\text{e+}00$
- 50% $0.000000\text{e+}00$
- 75% $0.000000\text{e+}00$
- max $8.201220\text{e+}07$

Classe 0,1,2 e 4 presentano outlier, in particolare Classe 0, 2 e 4 hanno Outlier che superano di gran lunga il baffo superiore.

4.66 Idle Min

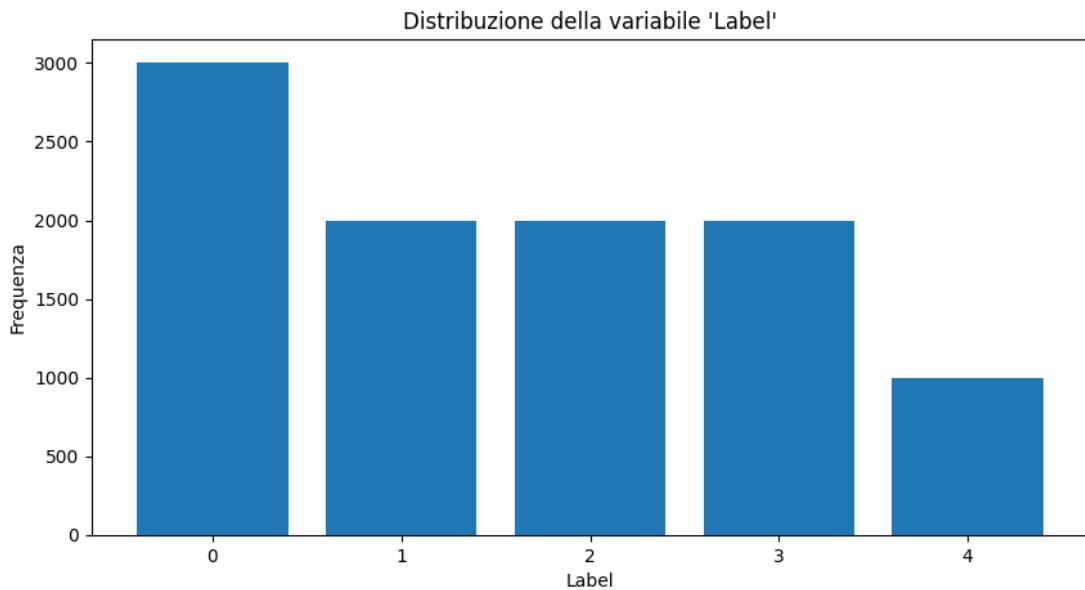


- count 1.000000e+04
- mean 1.541433e+06
- std 7.909073e+06
- min 0.000000e+00
- 25% 0.000000e+00
- 50% 0.000000e+00
- 75% 0.000000e+00
- max 6.451177e+07

Classe 0,1,2 e 4 presentano outlier, in particolare Classe 0, 2 e 4 hanno Outlier che superano di gran lunga il baffo superiore.

5. DISTRIBUZIONE DELLE CLASSI

Il seguente istogramma mostra la distribuzione della variabile target Label.



Le classi risultano sbilanciate: la classe 0 è nettamente più rappresentata (con 3000 istanze), mentre la classe 4 è significativamente meno frequente (con 1000). Mentre classi 1, 2 e 3 presentano le stesse frequenze (con 2000 istanze). Questo sbilanciamento potrebbe influenzare le performance dei modelli, penalizzando la precisione sulle classi meno rappresentate.

6. FEATURE SELECTION

Sono stati utilizzati due noti algoritmi di Feature Selection: Mutual Info Ranking(MI) e Information Gain (GI), con l'obiettivo di individuare le variabili indipendenti più rilevanti rispetto alla variabile target del nostro dataset. Entrambi gli approcci restituiscono un ordinamento delle feature sulla base di un criterio informativo, ma si basano su fondamenti differenti.

A partire dai ranking prodotti dai due metodi, estraiamo le prime dieci feature secondo ciascun criterio, confrontandole per individuare eventuali sovrapposizioni o differenze rilevanti.

Rank	Mutual Information	MI Score	Information Gain	GI Score
1	Average Packet Size	1.3934	Flow_Bytes	0.9403
2	Total Length of Fwd Packets	1.3903	Average Packet Size	0.9102
3	Subflow Fwd Bytes	1.3887	Total Length of Fwd Packets	0.8982
4	Avg Fwd Segment Size	1.3663	Subflow Fwd Bytes	0.8982
5	Fwd Packet Length Mean	1.3656	Packet Length Mean	0.8796
6	Flow_Bytes	1.3614	Fwd Packet Length Mean	0.8736
7	Max Packet Length	1.3535	Avg Fwd Segment Size	0.8736
8	Min Packet Length	1.3486	Max Packet Length	0.8607
9	Packet Length Mean	1.3446	Fwd Packet Length Max	0.8490
10	Fwd Packet Length Min	1.3433	Min Packet Length	0.8443

I due metodi presentano una notevole convergenza: sette feature su dieci sono comuni ad entrambi gli insiemi. Tale convergenza suggerisce che queste variabili presentano una forte correlazione informativa con la variabile target, indipendentemente dal metodo impiegato per la valutazione.

7. ADDESTRAMENTO DEI MODELLI

Sono stati utilizzati e confrontati diversi modelli di classificazione, con l'obiettivo di identificare l'approccio più efficace per la predizione della variabile target. Le tecniche utilizzate includono alberi decisionali e un ensemble di Support Vector Machine.

7.1 ALBERI DECISIONALI

La configurazione degli alberi decisionali è stata oggetto di una procedura di ottimizzazione, eseguita tramite 5-Fold Cross-Validation, al fine di valutare in modo robusto le performance medie (F1-Score) associate a diverse combinazioni di parametri. Il processo ha previsto la variazione congiunta di due dimensioni:

Criterio di suddivisione:

- gini
- entropy

Numero di feature selezionate, secondo tre diverse strategie di ordinamento:

- Ranking Mutual Info (MI)
- Ranking Information Gain (GI)
- Principal Component Analysis (PCA)

Per ciascuna combinazione di criteri e numero di feature (aggiunte progressivamente una alla volta), è stato allenato un albero su ciascuna delle 5 fold e calcolata la media degli F1-Score. La configurazione che ha massimizzato questa media è stata considerata ottimale e impiegata per il training finale del modello sull'intero training set.

Successivamente, il modello è stato testato sul dataset di test esterno per la valutazione definitiva delle prestazioni.

7.2 ENSEMBLE SVM

Infine, è stato utilizzato un ensemble di Support Vector Machine (SVM), come segue:

1. Estrazione dei dati tramite campionamento stratificato senza rimessa: sono stati generati 10 sottoinsiemi casuali dell'80% del training set, mantenendo la distribuzione delle classi.
2. Randomizzazione delle feature: per ciascun sottoinsieme, sono state selezionate 20 variabili indipendenti in modo casuale tra tutte quelle disponibili. Questo ha portato alla formazione di dieci dataset differenti, ciascuno con una configurazione unica di campioni e variabili.
3. Addestramento di 10 modelli SVM distinti, uno per ciascun dataset, ottimizzando i relativi iperparametri tramite 5-Fold Cross-Validation, utilizzando come metrica il Weighted F1-Score per garantire un bilanciamento adeguato in presenza di eventuali classi sbilanciate.

4. L'ensemble di 10 SVM è stato ottimizzato tramite Grid Search con 5-Fold CV, focalizzato sul weighted F1-score per bilanciare le classi. La griglia esplora:
 - C (0.1, 1, 10, 100, 1000);
 - gamma (1, 0.1, 0.01, 0.001, 0.0001);
 - Kernel 'rbf'.
5. Costruzione dell'ensemble: i 10 modelli SVM così ottenuti sono stati combinati tramite una regola di maggioranza. Per ogni osservazione del test set, la classe predetta è stata determinata come quella più frequentemente predetta dai 10 modelli.

7.3 Z-SCORE STANDARDIZATION

Nel processo di pre-processing dei dati, è stato utilizzato lo Z-Score standardization (o StandardScaler) per normalizzare le feature prima dell'applicazione della PCA e dell'addestramento dei modelli SVM. Questa tecnica consiste nella trasformazione di ciascuna variabile affinché abbia media 0 e deviazione standard 1.

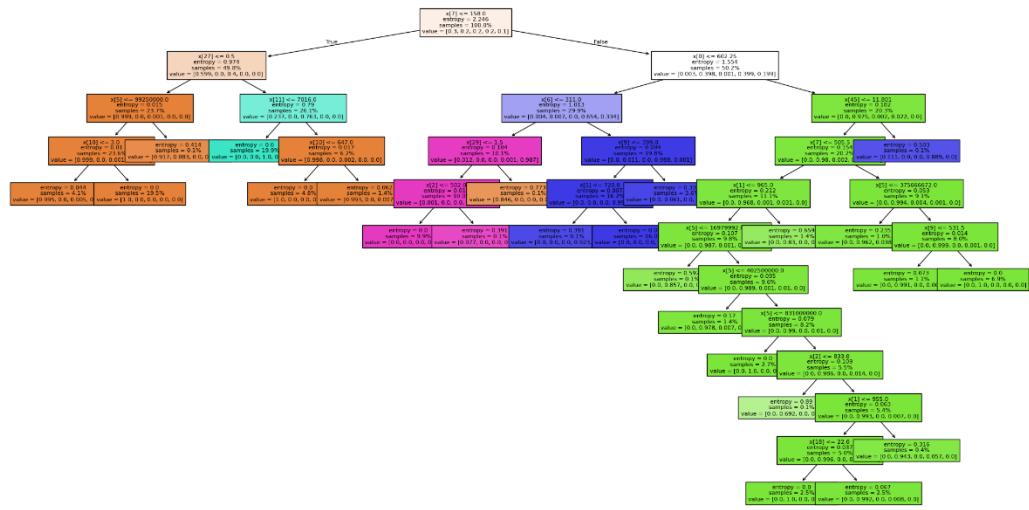
Questa scelta è stata dettata da precise necessità:

- Per la PCA, la standardizzazione previene il rischio che le feature con scale numeriche più ampie dominino artificialmente i componenti principali, anche quando non sono realmente più informative. Trasformando tutte le variabili in una scala comune (media zero e deviazione standard uno), ci assicuriamo che ogni feature contribuisca equamente all'analisi.
- Nel caso delle SVM la standardizzazione è altrettanto cruciale. Questo tipo di modello basa le sue decisioni sul calcolo delle distanze tra i punti, e senza un'adeguata normalizzazione, le feature su scale diverse distorcerebbero completamente queste misurazioni. Lo Z-Score mantiene l'equilibrio, permettendo al modello di valutare correttamente l'importanza relativa di ciascuna caratteristica.

8. RISULTATI

8.1 STRUTTURA DEGLI ALBERI

Per ciascuno dei tre alberi addestrati (basati su MI, IG, PCA), si riporta la rappresentazione grafica della struttura decisionale risultante.



Albero Decisionale con Feature selezionate via Mutual Information

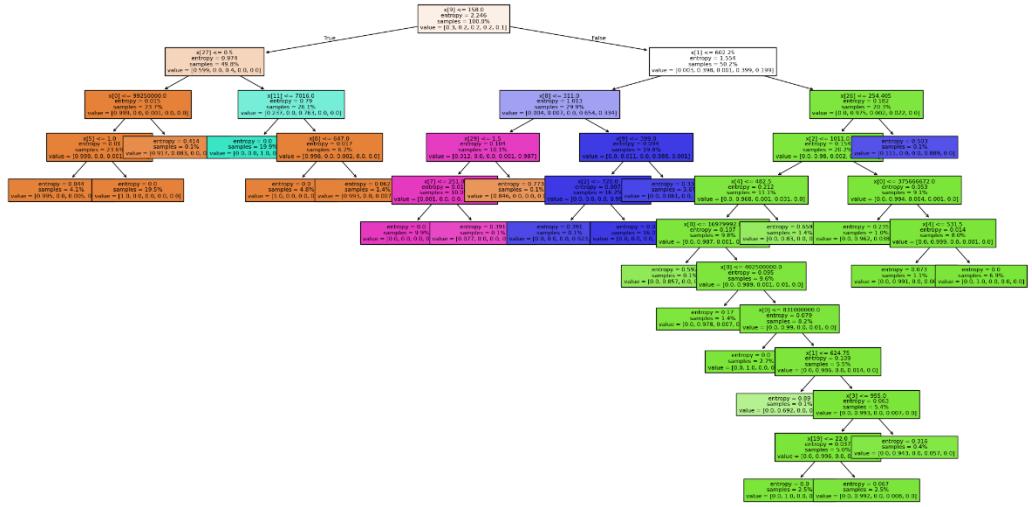
Migliore configurazione trovata:

Criterio: entropy

Numero di feature selezionate: 47

Miglior F1 Score: 0.9929036564851096

L'albero ha 47 nodi e 24 foglie.



Albero Decisionale con Feature selezionate via Information Gain

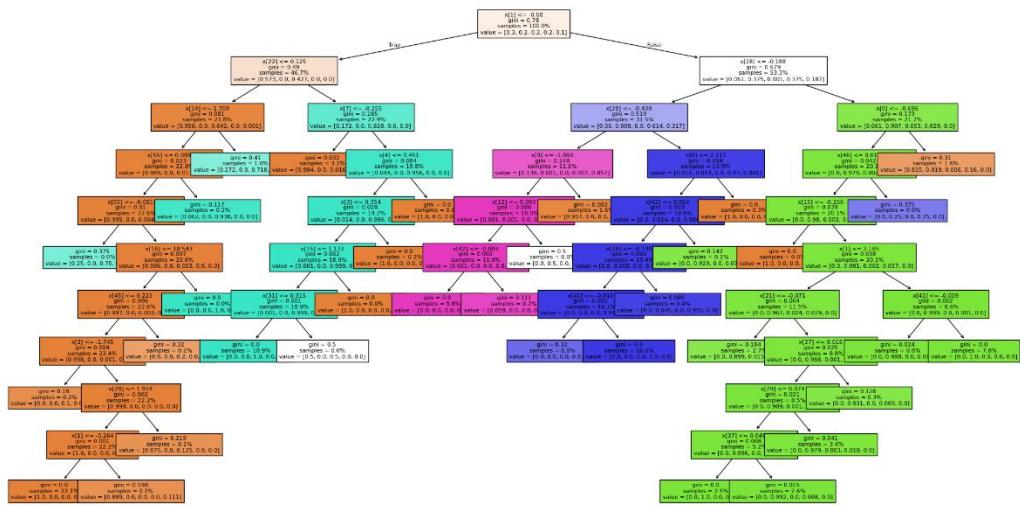
Migliore configurazione trovata:

Criterio: entropy

Numero di feature selezionate: 35

Miglior F1 Score: 0.9929036564851096

L'albero ha 47 nodi e 24 foglie.



Albero Decisionale con Feature selezionate via PCA

Migliore configurazione trovata:

Criterio: gini

Numero di feature selezionate: 56

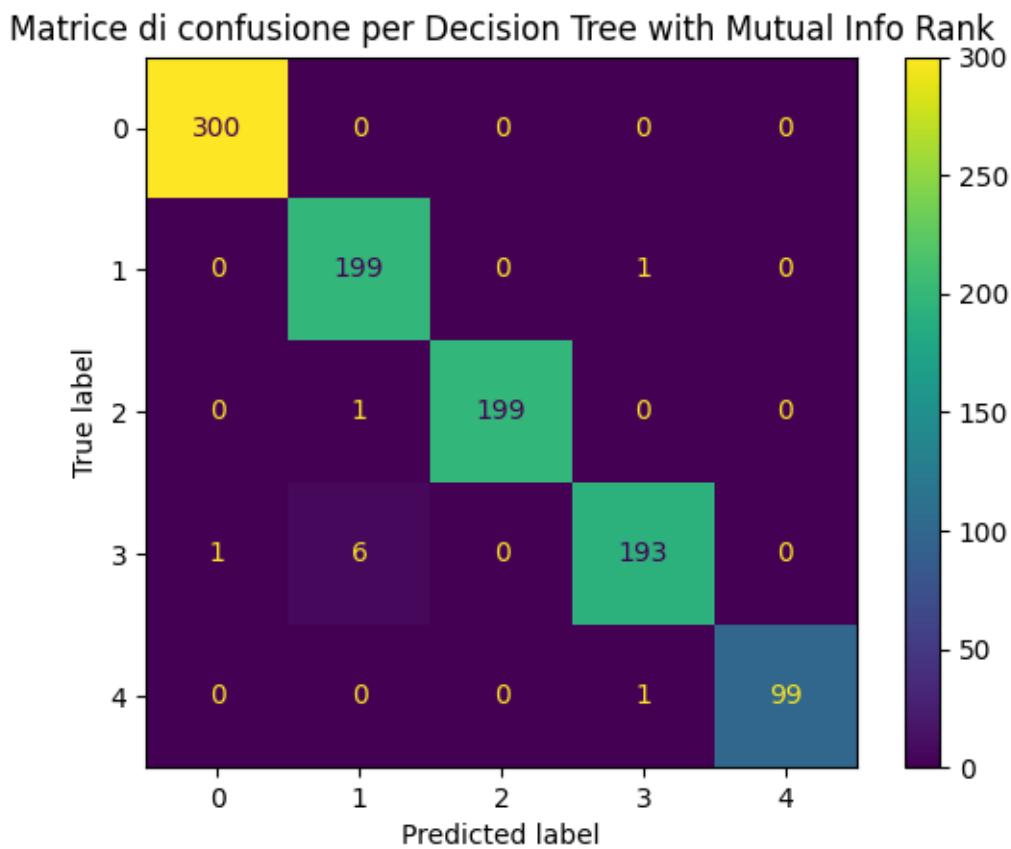
Miglior F1 Score: 0.9843874042879672

L'albero ha 67 nodi e 34 foglie.

8.2 CONFIGURAZIONE ENSEMBLE

Modello	Migliori Parametri	F-score Pesato
SVM 1	{'C': 1000, 'gamma': 1, 'kernel': 'rbf'}	0.9834
SVM 2	{'C': 1000, 'gamma': 1, 'kernel': 'rbf'}	0.9803
SVM 3	{'C': 1000, 'gamma': 1, 'kernel': 'rbf'}	0.9539
SVM 4	{'C': 1000, 'gamma': 1, 'kernel': 'rbf'}	0.9805
SVM 5	{'C': 1000, 'gamma': 1, 'kernel': 'rbf'}	0.9609
SVM 6	{'C': 1000, 'gamma': 1, 'kernel': 'rbf'}	0.9541
SVM 7	{'C': 1000, 'gamma': 1, 'kernel': 'rbf'}	0.9573
SVM 8	{'C': 1000, 'gamma': 1, 'kernel': 'rbf'}	0.9823
SVM 9	{'C': 1000, 'gamma': 0.01, 'kernel': 'rbf'}	0.9831
SVM 10	{'C': 1000, 'gamma': 1, 'kernel': 'rbf'}	0.9647

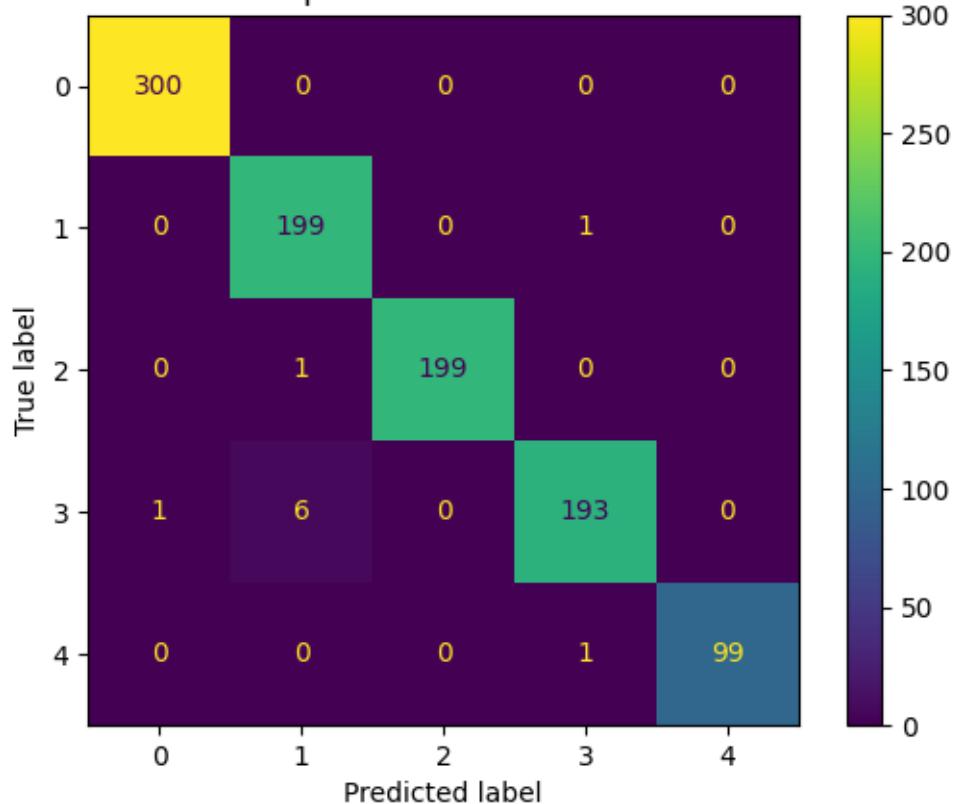
8.3 MATRICI DI CONFUSIONE E CLASSIFICATION REPORT



	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

Matrice di confusione per Decision Tree with Information Gain



precision recall f1-score support

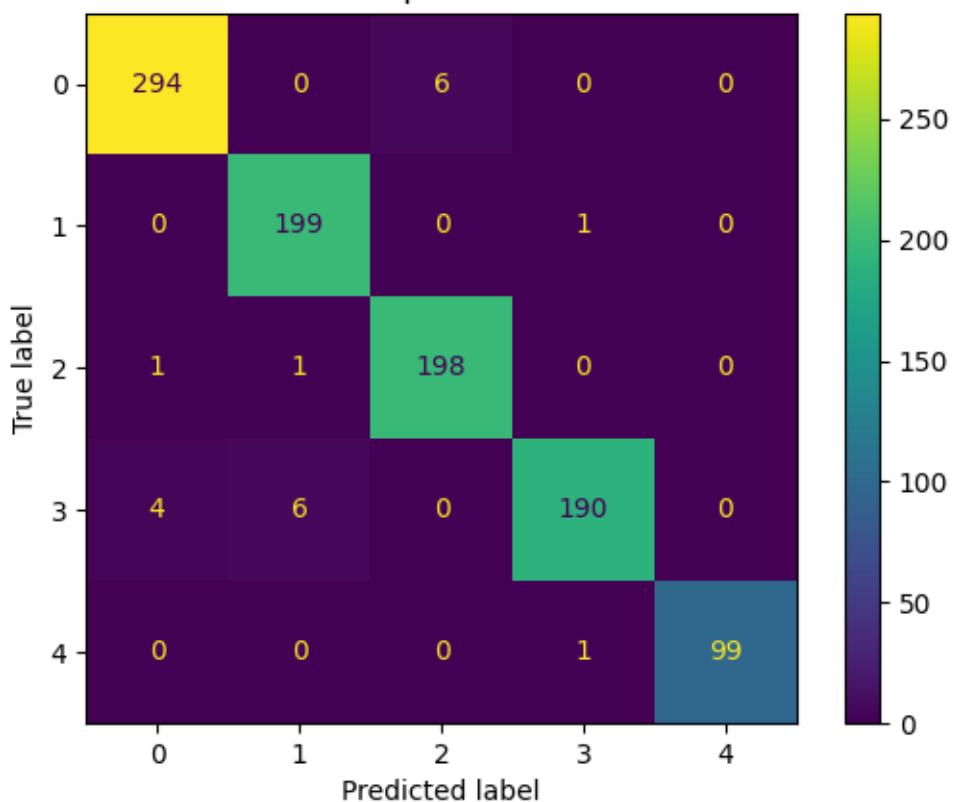
	precision	recall	f1-score	support
0	1.00	1.00	1.00	300
1	0.97	0.99	0.98	200
2	1.00	0.99	1.00	200
3	0.99	0.96	0.98	200
4	1.00	0.99	0.99	100

accuracy 0.99 1000

macro avg 0.99 0.99 0.99 1000

weighted avg 0.99 0.99 0.99 1000

Matrice di confusione per Decision Tree with PCA



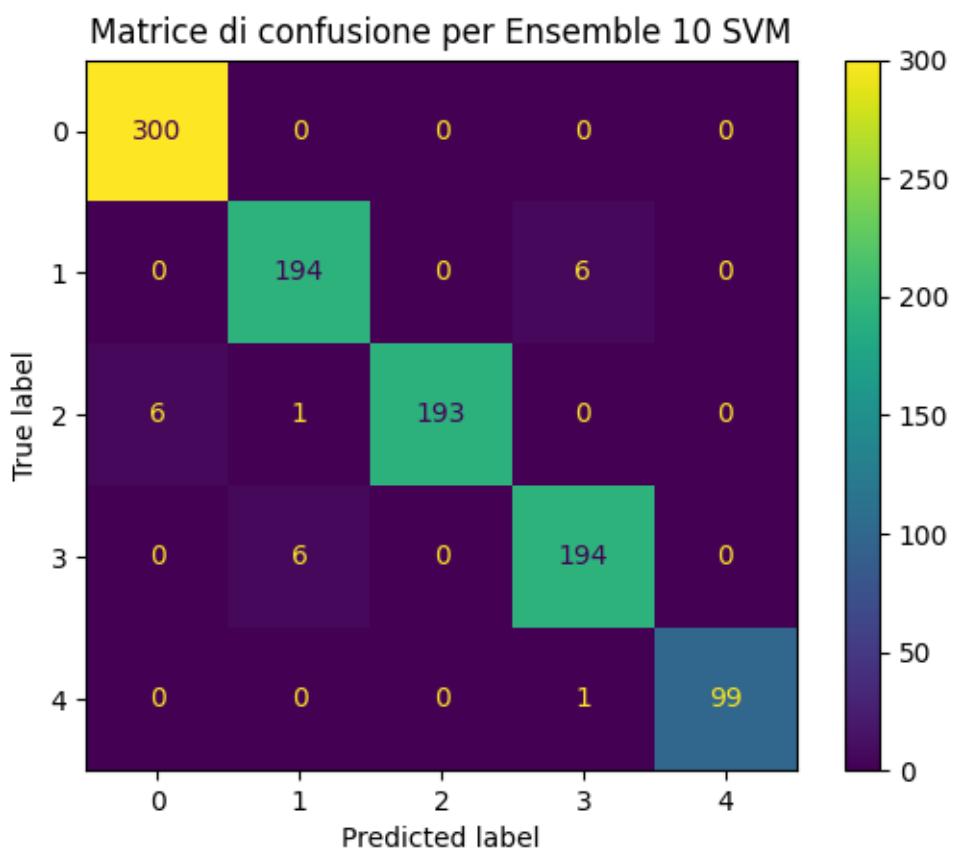
precision recall f1-score support

	0	1	2	3	4
0	0.98	0.98	0.98	300	
1	0.97	0.99	0.98	200	
2	0.97	0.99	0.98	200	
3	0.99	0.95	0.97	200	
4	1.00	0.99	0.99	100	

accuracy 0.98 1000

macro avg 0.98 0.98 0.98 1000

weighted avg 0.98 0.98 0.98 1000



precision recall f1-score support

	precision	recall	f1-score	support
0	0.98	1.00	0.99	300
1	0.97	0.97	0.97	200
2	1.00	0.96	0.98	200
3	0.97	0.97	0.97	200
4	1.00	0.99	0.99	100

accuracy 0.98 1000

macro avg 0.98 0.98 0.98 1000

weighted avg 0.98 0.98 0.98 1000

9. CONCLUSIONI

Il Decision Tree con Mutual Information raggiunge il 99% di accuratezza. Il gemello con Information Gain ottiene risultati identici (sempre 99%).

La versione con PCA scende al 98% di accuratezza e mostrando qualche difficoltà con gli attacchi SYN (classe 2).

Infine, l'ensemble di SVM si attesta sul 98%, risultando meno preciso dei primi due Decision Tree.