

cs-binome-preproc

May 20, 2023

1 Analyse et prétraitement des données

```
[1]: #pip install -U imbalanced-learn
```

```
Requirement already satisfied: imbalanced-learn in
c:\users\pappu\anaconda3\lib\site-packages (0.10.1)
Requirement already satisfied: joblib>=1.1.1 in
c:\users\pappu\anaconda3\lib\site-packages (from imbalanced-learn) (1.2.0)
Requirement already satisfied: scipy>=1.3.2 in
c:\users\pappu\anaconda3\lib\site-packages (from imbalanced-learn) (1.9.1)
Requirement already satisfied: numpy>=1.17.3 in
c:\users\pappu\anaconda3\lib\site-packages (from imbalanced-learn) (1.21.5)
Requirement already satisfied: scikit-learn>=1.0.2 in
c:\users\pappu\anaconda3\lib\site-packages (from imbalanced-learn) (1.0.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in
c:\users\pappu\anaconda3\lib\site-packages (from imbalanced-learn) (2.2.0)
Note: you may need to restart the kernel to use updated packages.
```

```
[2]: import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from imblearn.over_sampling import SMOTE
```

```
[4]: # Load the dataset
df = pd.read_csv("C:\LP-CIS\S6\Machine Learning &
↳CyberSec\Devoir\CICIDS2017_Thursday-WorkingHours-Morning-WebAttacks.
↳pcap_ISCX.csv")
```

```
[5]: df.head()
```

```
[5]:
```

	Destination	Port	Flow	Duration	Total Fwd Packets	Total Backward Packets	\
0		389		113095465	48	24	
1		389		113473706	68	40	
2		0		119945515	150	0	
3		443		60261928	9	7	
4		53		269	2	2	

	Total Length of Fwd Packets	Total Length of Bwd Packets	\
0	9668	10012	
1	11364	12718	
2	0	0	
3	2330	4221	
4	102	322	

	Fwd Packet Length Max	Fwd Packet Length Min	Fwd Packet Length Mean	\
0	403	0	201.416667	
1	403	0	167.117647	
2	0	0	0.000000	
3	1093	0	258.888889	
4	51	51	51.000000	

	Fwd Packet Length Std	...	min_seg_size_forward	Active Mean	\
0	203.548293	...	32	203985.500	
1	171.919413	...	32	178326.875	
2	0.000000	...	0	6909777.333	
3	409.702161	...	20	0.000	
4	0.000000	...	32	0.000	

	Active Std	Active Max	Active Min	Idle Mean	Idle Std	Idle Max	\
0	5.758373e+05	1629110	379	13800000.0	4.277541e+06	16500000	
1	5.034269e+05	1424245	325	13800000.0	4.229413e+06	16500000	
2	1.170000e+07	20400000	6	24400000.0	2.430000e+07	60100000	
3	0.000000e+00	0	0	0.0	0.000000e+00	0	
4	0.000000e+00	0	0	0.0	0.000000e+00	0	

	Idle Min	Label
0	6737603	BENIGN
1	6945512	BENIGN
2	5702188	BENIGN
3	0	BENIGN
4	0	BENIGN

[5 rows x 79 columns]

```
[7]: # optimising the dataset's size
import numpy as np
df = df.copy()

for column in df.columns:
    if df[column].dtype == np.int64:
        maxVal = df[column].max()
        if maxVal < 120:
            df[column] = df[column].astype(np.int8)
        elif maxVal < 32767:
```

```

        df[column] = df[column].astype(np.int16)
    else:
        df[column] = df[column].astype(np.int32)

    if df[column].dtype == np.float64:
        maxVal = df[column].max()
        minVal = df[df[column]>0][column]
        if maxVal < 120 and minVal>0.01 :
            df[column] = df[column].astype(np.float16)
        else:
            df[column] = df[column].astype(np.float32)

```

```

[8]: #after optimize of size
df.info()

```

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 170366 entries, 0 to 170365
Data columns (total 79 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Destination Port                      170366 non-null int32
1   Flow Duration                         170366 non-null int32
2   Total Fwd Packets                     170366 non-null int32
3   Total Backward Packets                170366 non-null int32
4   Total Length of Fwd Packets           170366 non-null int32
5   Total Length of Bwd Packets           170366 non-null int32
6   Fwd Packet Length Max                 170366 non-null int16
7   Fwd Packet Length Min                 170366 non-null int16
8   Fwd Packet Length Mean                 170366 non-null float32
9   Fwd Packet Length Std                 170366 non-null float32
10  Bwd Packet Length Max                 170366 non-null int16
11  Bwd Packet Length Min                 170366 non-null int16
12  Bwd Packet Length Mean                 170366 non-null float32
13  Bwd Packet Length Std                 170366 non-null float32
14  Flow Bytes/s                          170346 non-null float32
15  Flow Packets/s                        170366 non-null float32
16  Flow IAT Mean                         170366 non-null float32
17  Flow IAT Std                          170366 non-null float32
18  Flow IAT Max                          170366 non-null int32
19  Flow IAT Min                          170366 non-null int32
20  Fwd IAT Total                         170366 non-null int32
21  Fwd IAT Mean                          170366 non-null float32
22  Fwd IAT Std                           170366 non-null float32
23  Fwd IAT Max                           170366 non-null int32
24  Fwd IAT Min                           170366 non-null int32
25  Bwd IAT Total                         170366 non-null int32
26  Bwd IAT Mean                          170366 non-null float32

```

27	Bwd IAT Std	170366	non-null	float32
28	Bwd IAT Max	170366	non-null	int32
29	Bwd IAT Min	170366	non-null	int32
30	Fwd PSH Flags	170366	non-null	int8
31	Bwd PSH Flags	170366	non-null	int8
32	Fwd URG Flags	170366	non-null	int8
33	Bwd URG Flags	170366	non-null	int8
34	Fwd Header Length	170366	non-null	int32
35	Bwd Header Length	170366	non-null	int32
36	Fwd Packets/s	170366	non-null	float32
37	Bwd Packets/s	170366	non-null	float32
38	Min Packet Length	170366	non-null	int16
39	Max Packet Length	170366	non-null	int16
40	Packet Length Mean	170366	non-null	float32
41	Packet Length Std	170366	non-null	float32
42	Packet Length Variance	170366	non-null	float32
43	FIN Flag Count	170366	non-null	int8
44	SYN Flag Count	170366	non-null	int8
45	RST Flag Count	170366	non-null	int8
46	PSH Flag Count	170366	non-null	int8
47	ACK Flag Count	170366	non-null	int8
48	URG Flag Count	170366	non-null	int8
49	CWE Flag Count	170366	non-null	int8
50	ECE Flag Count	170366	non-null	int8
51	Down/Up Ratio	170366	non-null	int8
52	Average Packet Size	170366	non-null	float32
53	Avg Fwd Segment Size	170366	non-null	float32
54	Avg Bwd Segment Size	170366	non-null	float32
55	Fwd Header Length.1	170366	non-null	int32
56	Fwd Avg Bytes/Bulk	170366	non-null	int8
57	Fwd Avg Packets/Bulk	170366	non-null	int8
58	Fwd Avg Bulk Rate	170366	non-null	int8
59	Bwd Avg Bytes/Bulk	170366	non-null	int8
60	Bwd Avg Packets/Bulk	170366	non-null	int8
61	Bwd Avg Bulk Rate	170366	non-null	int8
62	Subflow Fwd Packets	170366	non-null	int32
63	Subflow Fwd Bytes	170366	non-null	int32
64	Subflow Bwd Packets	170366	non-null	int32
65	Subflow Bwd Bytes	170366	non-null	int32
66	Init_Win_bytes_forward	170366	non-null	int32
67	Init_Win_bytes_backward	170366	non-null	int32
68	act_data_pkt_fwd	170366	non-null	int32
69	min_seg_size_forward	170366	non-null	int8
70	Active Mean	170366	non-null	float32
71	Active Std	170366	non-null	float32
72	Active Max	170366	non-null	int32
73	Active Min	170366	non-null	int32
74	Idle Mean	170366	non-null	float32

```

75 Idle Std                      170366 non-null float32
76 Idle Max                      170366 non-null int32
77 Idle Min                      170366 non-null int32
78 Label                        170366 non-null object
dtypes: float32(24), int16(6), int32(28), int8(20), object(1)
memory usage: 40.3+ MB

```

```

[9]: # Feature selection
df = df[['Destination Port', '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/s', 'Flow Packets/s', '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', 'FIN Flag Count', 'SYN Flag Count',_
↳'RST Flag Count',
        'PSH Flag Count', 'ACK Flag Count', 'URG Flag Count', 'Label']]

```

```

[10]: #selecting the necessary columns : remaining 40
df.info()

```

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 170366 entries, 0 to 170365
Data columns (total 41 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Destination Port                      170366 non-null int32
1   Flow Duration                        170366 non-null int32
2   Total Fwd Packets                    170366 non-null int32
3   Total Backward Packets                170366 non-null int32
4   Total Length of Fwd Packets           170366 non-null int32
5   Total Length of Bwd Packets           170366 non-null int32
6   Fwd Packet Length Max                 170366 non-null int16
7   Fwd Packet Length Min                 170366 non-null int16
8   Fwd Packet Length Mean                170366 non-null float32
9   Fwd Packet Length Std                 170366 non-null float32
10  Bwd Packet Length Max                  170366 non-null int16

```

```

11 Bwd Packet Length Min      170366 non-null int16
12 Bwd Packet Length Mean    170366 non-null float32
13 Bwd Packet Length Std     170366 non-null float32
14 Flow Bytes/s              170346 non-null float32
15 Flow Packets/s            170366 non-null float32
16 Flow IAT Mean             170366 non-null float32
17 Flow IAT Std              170366 non-null float32
18 Flow IAT Max              170366 non-null int32
19 Flow IAT Min              170366 non-null int32
20 Fwd IAT Total             170366 non-null int32
21 Fwd IAT Mean              170366 non-null float32
22 Fwd IAT Std               170366 non-null float32
23 Fwd IAT Max               170366 non-null int32
24 Fwd IAT Min               170366 non-null int32
25 Bwd IAT Total             170366 non-null int32
26 Bwd IAT Mean              170366 non-null float32
27 Bwd IAT Std               170366 non-null float32
28 Bwd IAT Max               170366 non-null int32
29 Bwd IAT Min               170366 non-null int32
30 Fwd PSH Flags             170366 non-null int8
31 Bwd PSH Flags             170366 non-null int8
32 Fwd URG Flags             170366 non-null int8
33 Bwd URG Flags             170366 non-null int8
34 FIN Flag Count            170366 non-null int8
35 SYN Flag Count            170366 non-null int8
36 RST Flag Count            170366 non-null int8
37 PSH Flag Count            170366 non-null int8
38 ACK Flag Count            170366 non-null int8
39 URG Flag Count            170366 non-null int8
40 Label                     170366 non-null object
dtypes: float32(12), int16(4), int32(14), int8(10), object(1)
memory usage: 21.1+ MB

```

```
[18]: df['Label'].value_counts()
```

```

[18]: BENIGN                      168186
      Web Attack  Brute Force      1507
      Web Attack  XSS              652
      Web Attack  Sql Injection     21
      Name: Label, dtype: int64

```

```

[19]: from sklearn.preprocessing import LabelEncoder
      #from categorical data to numerical format
      label = LabelEncoder()
      df['Label'] = label.fit_transform(df['Label'])
      df['Label'].unique()

```

```
[19]: array([0, 1, 3, 2])
```

```
[20]: #show missing data  
df.isnull().sum()
```

```
[20]: Destination Port          0  
Flow Duration                0  
Total Fwd Packets            0  
Total Backward Packets       0  
Total Length of Fwd Packets  0  
Total Length of Bwd Packets  0  
Fwd Packet Length Max        0  
Fwd Packet Length Min        0  
Fwd Packet Length Mean       0  
Fwd Packet Length Std        0  
Bwd Packet Length Max        0  
Bwd Packet Length Min        0  
Bwd Packet Length Mean       0  
Bwd Packet Length Std        0  
Flow Bytes/s                 20  
Flow Packets/s               0  
Flow IAT Mean                0  
Flow IAT Std                 0  
Flow IAT Max                 0  
Flow IAT Min                 0  
Fwd IAT Total                0  
Fwd IAT Mean                 0  
Fwd IAT Std                  0  
Fwd IAT Max                  0  
Fwd IAT Min                  0  
Bwd IAT Total                0  
Bwd IAT Mean                 0  
Bwd IAT Std                  0  
Bwd IAT Max                  0  
Bwd IAT Min                  0  
Fwd PSH Flags                0  
Bwd PSH Flags                0  
Fwd URG Flags                0  
Bwd URG Flags                0  
FIN Flag Count               0  
SYN Flag Count               0  
RST Flag Count               0  
PSH Flag Count               0  
ACK Flag Count               0  
URG Flag Count               0  
Label                        0  
dtype: int64
```

```
[21]: # Handle missing data
df['Flow Bytes/s'].fillna(df['Flow Bytes/s'].mean(),inplace=True)
```

```
[22]: #handle infinite values and NaN values
df = df.replace([np.inf, -np.inf], np.nan)
df = df.dropna()
```

```
[23]: df.isnull().sum()
```

```
[23]: Destination Port          0
Flow Duration                 0
Total Fwd Packets             0
Total Backward Packets        0
Total Length of Fwd Packets    0
Total Length of Bwd Packets    0
Fwd Packet Length Max         0
Fwd Packet Length Min         0
Fwd Packet Length Mean        0
Fwd Packet Length Std         0
Bwd Packet Length Max         0
Bwd Packet Length Min         0
Bwd Packet Length Mean        0
Bwd Packet Length Std         0
Flow Bytes/s                  0
Flow Packets/s                0
Flow IAT Mean                 0
Flow IAT Std                  0
Flow IAT Max                  0
Flow IAT Min                  0
Fwd IAT Total                 0
Fwd IAT Mean                  0
Fwd IAT Std                   0
Fwd IAT Max                   0
Fwd IAT Min                   0
Bwd IAT Total                 0
Bwd IAT Mean                  0
Bwd IAT Std                   0
Bwd IAT Max                   0
Bwd IAT Min                   0
Fwd PSH Flags                 0
Bwd PSH Flags                 0
Fwd URG Flags                 0
Bwd URG Flags                 0
FIN Flag Count                0
SYN Flag Count                0
RST Flag Count                0
PSH Flag Count                0
```



```
ACK Flag Count      0
URG Flag Count      0
Label               0
dtype: int64
```

```
[24]: # find duplicated rows
dup_rows = df.duplicated()
df = df.drop_duplicates()
print(dup_rows)
```

```
0      False
1      False
2      False
3      False
4      False
...
170361  False
170362  False
170363  False
170364   True
170365  False
Length: 170231, dtype: bool
```

```
[25]: # check for duplicated columns
duplicated_cols = df.T.duplicated()

# get duplicated column names
duplicated_col_names = df.columns[duplicated_cols].tolist()

print("Duplicated column names:", duplicated_col_names)
```

```
Duplicated column names: ['Fwd URG Flags', 'Bwd URG Flags', 'SYN Flag Count']
```

```
[31]: # drop duplicated columns and keep only one column
df = df.drop(['Fwd URG Flags', 'Bwd URG Flags'], axis=1)
```

	Destination Port	Flow Duration	Total Fwd Packets \
0	389	113095465	48
1	389	113473706	68
2	0	119945515	150
3	443	60261928	9
4	53	269	2
...
170360	443	181	3
170361	55641	49	1
170362	45337	217	2
170363	22	1387547	41
170365	60146	50	1

	Total Backward Packets	Total Length of Fwd Packets \
0	24	9668
1	40	11364
2	0	0
3	7	2330
4	2	102
...
170360	1	18
170361	3	6
170362	1	31
170363	46	2728
170365	2	0

	Total Length of Bwd Packets	Fwd Packet Length Max \
0	10012	403
1	12718	403
2	0	0
3	4221	1093
4	322	51
...
170360	6	6
170361	18	6
170362	6	31
170363	6634	456
170365	0	0

	Fwd Packet Length Min	Fwd Packet Length Mean	Fwd Packet Length Std \
0	0	201.416672	203.548294
1	0	167.117645	171.919418
2	0	0.000000	0.000000
3	0	258.888885	409.702148
4	51	51.000000	0.000000
...
170360	6	6.000000	0.000000
170361	6	6.000000	0.000000
170362	0	15.500000	21.920311
170363	0	66.536583	110.129944
170365	0	0.000000	0.000000

	...	Bwd IAT Min	Fwd PSH Flags	Bwd PSH Flags	FIN Flag Count \
0	...	3	1	0	0
1	...	3	1	0	0
2	...	0	0	0	0
3	...	48	0	0	0
4	...	4	0	0	0
...
170360	...	0	0	0	1

170361	...	1	0	0	1
170362	...	0	1	0	0
170363	...	1	0	0	0
170365	...	1	0	0	0

	SYN Flag Count	RST Flag Count	PSH Flag Count	ACK Flag Count	\
0	1	0	0	1	
1	1	0	0	1	
2	0	0	0	0	
3	0	0	1	0	
4	0	0	0	0	
...	
170360	0	0	0	0	
170361	0	0	0	0	
170362	1	0	0	1	
170363	0	0	1	0	
170365	0	0	0	1	

	URG Flag Count	Label
0	0	0
1	0	0
2	0	0
3	0	0
4	0	0
...
170360	0	0
170361	0	0
170362	0	0
170363	0	0
170365	1	0

[159742 rows x 39 columns]

```
[33]: df.shape
```

```
[33]: (159742, 39)
```

```
[40]: #End of preprocessing step
# Save the preprocessed dataset
df.to_csv('preprocessed_dataset.csv', index=False)
```

```
[ ]:
```