Android Detection Malware

Integrantes

Juan Pablo Manotas Cañas, CC: 1020487220, Bioingenieria

Carlos Daniel Quiros Carbajal, CC:1214743514, Ingenieria de sistema

Jharol Sebastian Agudelo Ramos, CC: 1214740141, Ingenieria mecanica

Progreso alcanzado

Luego de la selección del dataset , se creó un nuevo notebook que tiene como fin generar un nuevo dataset que cumple con los requerimientos del proyecto.

Para esto se importó el dataset original (355630 datos × 86 columnas) y se realizó una selección de 10000 datos aleatorios con el propósito de hacerlo más manejable con el recurso de procesamiento disponible generando así un nuevo dataset (10000 datos × 84 columnas) que posteriormente se subió a github.

Luego de esto, como el nuevo dataset no tenía valores nulos se llevó a cabo una eliminación de 500 filas de tres columnas con datos poco relevantes para nuestro fin. Por lo cual, se cumple el requisito de al menos tener un 5% de datos faltantes en al menos 3 columnas (5% en 4 columnas). Una vez realizado, las siguientes columnas fueron afectadas:

```
na_values = df_modified.isna().sum()
na_values[na_values != 0]
```

```
Fwd PSH Flags     500
 Bwd PSH Flags    500
 Fwd URG Flags    500
 Bwd URG Flags    500
```

Una vez cumplido este requerimiento, se procedió a analizar las variables categóricas, al observar que no contaba con la cantidad de variables categóricas necesarias para cumplir la exigencia de tener al menos el 10% de las columnas categóricas en el proyecto, fue necesario realizar un procedimiento para convertir variables numéricas a categóricas.

Para cumplir con la cantidad de variables categoricas se seleccionaron 5 columnas con datos fáciles de categorizar (pocas variables únicas). Se definió la categorización de cada una como se muestra a continuación:

Down/Up Ratio, PSH Flag count, ACK Flag Count, RST Flag Count, SYN Flag Count.

```
def categorizar(valor):
    if valor <= 3:
        return 'low'
    elif valor >= 5:
        return 'high'
    else:
        return 'medium'
```

```
def categorizar_psh(valor):
    if valor == 0.0:
        return 'send'
    else:
        return 'receive'
```

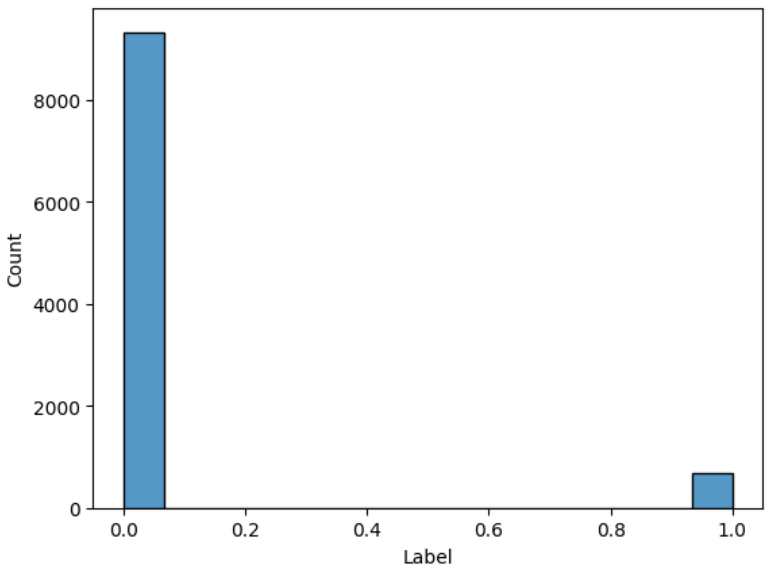
```
def categorizar_act(valor):
    if valor == 0.0:
        return 'not require'
    else:
        return 'require'
```

```
def categorizar_rst(valor):
    if valor == 0.0:
        return 'reset'
    else:
        return 'no reset'
```

```
def categorizar_syn(valor):
    if valor == 0.0:
        return 'no conection
    else:
        return 'conection'
```

La variable objetivo es ‘ Label’ y se decidió que todo lo relacionado con valores android (Android_Adware, Android_Scareware, Android_SMS_Malware) se nombraría con cero y los relacionados con benigno (Benign) se nombrarían con uno.

Luego se grafica variable de salida, se encuentra que el valor 0 es el más presente en el dataset.



Tras haber realizado la selección de datos y habiendo cumplido con todos los requerimientos del proyecto, se encontró lo siguiente:

- Tamaño del dataset: El tamaño del dataset es (1000, 84).
- Valores nulos: Las columnas Fwd PSH Flags, Bwd PSH Flags, Fwd URG Flags, Bwd URG Flags, tienen 500 valores nulos.
- Variables categóricas: Las columnas Flow ID, Source IP, Destination IP, Down/Up Ratio, PSH Flag count, ACK Flag Count, RST Flag Count, SYN Flag Count, son categóricas.
- Tipos de datos: Los tipos de datos que están presentes son float64 en su mayoría, algunos int64 y otros object en las variables que se convirtieron a categóricas.
- Inspección de datos numéricos: Se calcula el promedio, la desviación estándar, valor mínimo, máximo y percentiles de las columnas numéricas además se graficó la matriz de correlaciones.