Fase #2: Limpieza de datos y selección.

Grupo 6

Recordando el dataset...

Variables Iniciales

IPV4_SRC_ADDR

Es la direccion origen

L4_SRC_PORT:

La direccion IPV4 del destino

IPV4_DST_ADDR

Numero del puerto de origen

L4_DST_PORT

Numero del puerto de origen

PROTOCOL

Identificador del protocolo

L7_PROTO

Numero del protocolo de la capa 7

IN_BYTES

Numero entrante de bytes

IN_PKTS

Numero entrante de paquetes

OUT_BYTES

Numero saliente de bytes

OUT_PKTS

Numero saliente de bytes

TCP_FLAGS

Acumulativo de todas las banderas de TC

CLIENT_TCP_FLAGS

Acumulativo de todos los indicadores TCP del cliente

Variables Iniciales

SERVER_TCP_FLAGS

Acumulativo de todos los indicadores TCP del servidor

FLOW_DURATION_MILLISECONDS

Duracion del flujo en segundos

DURATION_IN

transmisión del cliente al servidor (mseg)

DURATION_OUT

Duración de la transmisión del cliente al servidor (mseg)

MIN_TTL

Flujo mínimo TTL

MAX_TTL

TTL de flujo máximo

LONGEST_FLOW_PKT

Paquete más largo (bytes) del flujo

SHORTEST_FLOW_PKT

Paquete más corto (bytes) del flujo

MIN_IP_PKT_LEN

Longitud del paquete IP de mayor flujo observado

MAX_IP_PKT_LEN

Longitud del paquete IP de mayor flujo observado

SRC_TO_DST_SECOND_BYTES

Src a dst Bytes/seg

DST_TO_SRC_SECOND_BYTES

dst to src bytes/seg

Variables Iniciales

RETRANSMITTED_IN_BYTES

Número de bytes de flujo TCP retransmitidos (src->dst)

RETRANSMITTED_IN_BYTES

Número de paquetes de flujo TCP retransmitidos (src->dst)

RETRANSMITTED_OUT_BYTES

Número de bytes de flujo TCP retransmitidos (dst->src)

RETRANSMITTED_OUT_PKTS

Número de paquetes de flujo TCP retransmitidos (dst->src)

SRC_TO_DST_AVG_THROUGHPUT

Src a thpt promedio dst (bps)

DST_TO_SRC_AVG_THROUGHPUT

DST a thpt promedio de origen (bps)

NUM_PKTS_UP_TO_128_BYTES

Paquetes cuyo tamaño de IP <= 128

NUM_PKTS_128_TO_256_BYTES

Paquetes cuyo tamaño de IP > 128 y <= 256

NUM_PKTS_256_TO_512_BYTES

Paquetes cuyo tamaño de IP > 256 y <= 512

NUM_PKTS_512_TO_1024_BYTES

Paquetes cuyo tamaño de IP > 512 y <= 1024

NUM_PKTS_1024_TO_1514_BYTES

Paquetes cuyo tamaño de IP > 1024 y <= 1514

TCP_WIN_MAX_IN

Ventana máxima de TCP (src->dst)

Variables Iniciales

TCP_WIN_MAX_OUT

Ventana máxima de TCP (dst->src)

DNS_QUERY_TYPE

Tipo de consulta DNS (p. ej., 1=A, 2=NS…)

Dataset

Base de datos origen

ICMP_TYPE

Tipo ICMP * 256 + código ICMP

DNS_TTL_ANSWER: FTP_COMMAND_RET CODE

TTL del primer registro A (si lo hay)

MAX_IP_PKT_LEN

Longitud del paquete IP de mayor flujo observado

ICMP_IPV4_TYPE

Tipo ICMP

Label

Label

SRC_TO_DST_SECO ND_BYTES

Src a dst Bytes/seg

DNS_QUERY_ID

ID de transacción de consulta de DNS

Attack

Tipo de etiqueta

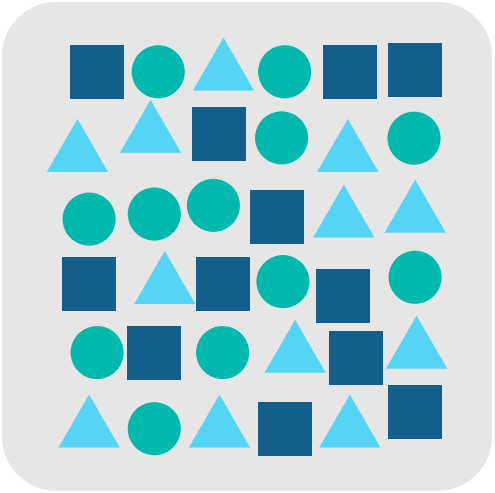
DST_TO_SRC_SECO ND_BYTES

dst to src bytes/seg

Entonces...

PROCESO

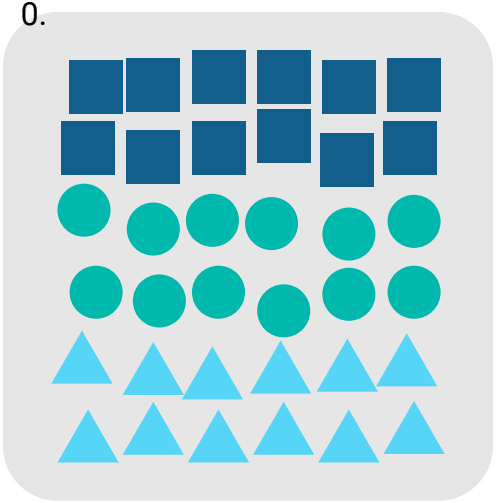
data



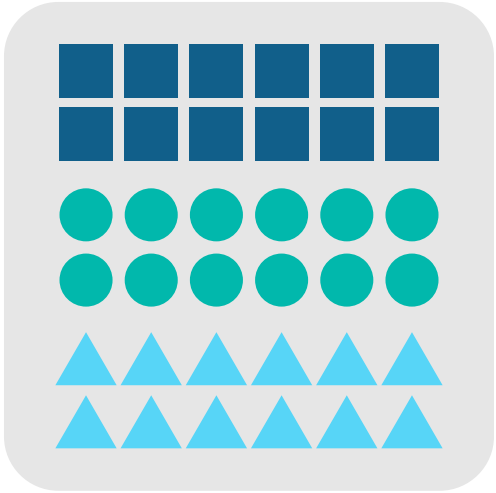
Tomamos el dataset inicial, el cual contaba con 46 columnas y 13 GB de información.

limpieza

Para la limpieza se eliminan columnas repetitivas y columnas con valores iguales a 0.



selección



La selección se realizó en base a los datos restantes, los cuales consideramos relevantes.

La limpieza se basó en:

Luego de hacer un repaso en la base de datos, nos percatamos que existen algunas columnas que eran de cierta forma repetidas, por lo que estás decidimos eliminarlas y solo quedarnos con una de esas variables, así como también, eliminamos dos o tres variables que no representan un valor para el modelo.

DNS_QUERY_ID

NUM_PKTS_UP_TO_128_BYTES

RETRANSMITTED_IN_BYTES

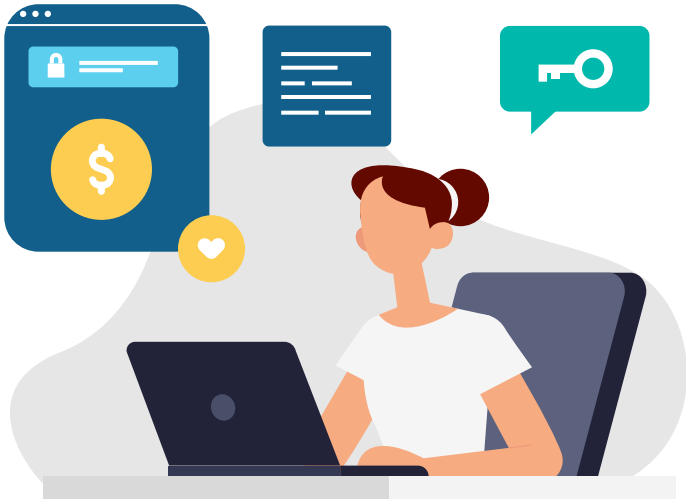
Label

ICMP_TYPE

DNS_TTL_ANSWER: FTP_COMMAND_RET CODE

Selección de datos

La selección de datos consistió en la utilización de los datos restantes, si bien es cierto que sabemos que la cantidad de variables no contribuye a la calidad del modelo, consideramos que el resto de variables pueden ser de utilidad para el modelo.



Gracias