Sistema de detección de Fraude - Proyecto final para Coderhouse

**Integrantes**

· Sanchez Federico Matías

· Vaisman Ricardo

· Rodriguez Miguel

# Tabla de contenido

[**Tabla de contenido**](#_lziuyttk2gs2) **2**

[**Introducción**](#_klvtaa8b1ti2) **3**

[Problema a resolver](#_xkczs2mwfjwc) 3

[Contexto](#_gjzfv9ps0c2p) 3

[Preguntas y objetivos de la investigación](#_t1ic6cvlr3y9) 3

[**Data Acquisition**](#_90g74wdqijny) **3**

[Fuente del dataset y criterios de selección](#_cy1xpj12ahtt) 3

[Información sobre el conjunto de datos](#_1e572nm0754o) 3

[**Análisis Univariado**](#_v8gu4bdvrace) **5**

[Analisis general](#_7n7y5lmhg1b1) 5

[Variable target](#_5pi04kzfol9p) 5

[Nulos y valores en cero](#_nrhjhieq4xrv) 6

[Limpieza variables categóricas](#_m3dkbp2c4dcm) 6

[Bar charts](#_81707f6jppbt) 7

[Box plots](#_aie5hkm0g82i) 7

[Análisis de frecuencias Categóricas](#_92s7b1ons5uv) 7

[Conclusiones](#_prwehwfcekuq) 7

[**Análisis Bivariado**](#_cs94y3l46oge) **7**

[Box plots](#_2iwlj48mgx74) 7

[**Análisis Multivariado**](#_ef604jywrwih) **7**

[Tabla de correlación](#_4le6qrzdvxnx) 7

[**Análisis de componentes principales**](#_1c9w5h3amyxn) **7**

[**Indicadores**](#_kuvpy4h9bjjc) **7**

# Introducción

## Problema a resolver

Se necesita implementar un sistema de detección de fraudes para transacciones con la criptomoneda Ethereum (ETH).

## Contexto

Este conjunto de datos contiene filas de fraudes conocidos y transacciones válidas realizadas con Ethereum, un tipo de criptomoneda.

## Preguntas y objetivos de la investigación

Entender que variables de las transacciones de ETH dan indicios para creer que una transacción es fraudulenta o no.

# Data Acquisition

## Fuente del dataset y criterios de selección

Ethereum Fraud Detection Dataset

<https://www.kaggle.com/vagifa/ethereum-frauddetection-dataset>

Seleccionamos el siguiente set datos ya que contaba con una variable TARGET ya definida. Contiene un número considerable de columnas y representa de una manera muy entendible una transacción de ETH.

Posee 9841 registros y 51 columnas lo consideramos es una buena cantidad de información para analizar y crear un modelo de ML.

## Información sobre el conjunto de datos

Aquí hay una descripción de las filas del conjunto de datos:

* Índex: El número de índice de una fila.
* Address: la dirección de la cuenta Ethereum.
* FLAG: si la transacción es fraude o no.
* Avg min between sent tnx: tiempo promedio entre transacciones enviadas para la cuenta en minutos.
* Avg min between received tnx: tiempo promedio entre transacciones recibidas para la cuenta en minutos.
* Time Diff between first and last (Mins): diferencia de tiempo entre la primera y la última transacción.
* Sent Tnx: número total de transacciones normales enviadas
* Received Tnx: número total de transacciones normales recibidas
* NumberofCreated\_Contracts: Número total de transacciones de contrato creadas.
* UniqueReceivedFrom\_Addresses: Total de direcciones únicas desde las cuales la cuenta recibió transacciones
* UniqueSentTo\_Addresses20: Total de direcciones únicas desde las cuales la cuenta envió transacciones
* MinValueReceived: valor mínimo en Ether recibido alguna vez
* MaxValueReceived: valor máximo en Ether recibido alguna vez
* AvgValueReceived5Average: Valor promedio en Ether recibido alguna vez
* MinValSent: Valor mínimo de Ether jamás enviado
* MaxValSent: Valor máximo de Ether jamás enviado
* AvgValSent: valor promedio de Ether enviado alguna vez
* MinValueSentToContract: valor mínimo de Ether enviado a un contrato
* MaxValueSentToContract: valor máximo de Ether enviado a un contrato
* AvgValueSentToContract: valor promedio de Ether enviado a contratos
* TotalTransactions(IncludingTnxtoCreate\_Contract): número total de transacciones
* TotalEtherSent:Ether total enviado para la dirección de la cuenta
* TotalEtherReceived: Ether total recibido para la dirección de la cuenta
* TotalEtherSent\_Contracts: Total Ether enviado a direcciones de contrato
* TotalEtherBalance: Saldo total de Ether después de las transacciones promulgadas
* TotalERC20Tnxs: número total de transacciones de transferencia de token ERC20
* ERC20TotalEther\_Received: Total de transacciones recibidas del token ERC20 en Ether
* ERC20TotalEther\_Sent: Total de transacciones ERC20token enviadas en Ether
* ERC20TotalEtherSentContract: transferencia de token ERC20 total a otros contratos en Ether
* ERC20UniqSent\_Addr: número de transacciones de token ERC20 enviadas a direcciones de cuenta únicas
* ERC20UniqRec\_Addr: número de transacciones de token ERC20 recibidas de direcciones únicas
* ERC20UniqRecContractAddr: número de transacciones de token ERC20 recibidas de direcciones de contrato únicas
* ERC20AvgTimeBetweenSent\_Tnx: tiempo promedio entre transacciones enviadas con token ERC20 en minutos
* ERC20AvgTimeBetweenRec\_Tnx: tiempo promedio entre transacciones recibidas del token ERC20 en minutos
* ERC20AvgTimeBetweenContract\_Tnx: tiempo promedio del token ERC20 entre transacciones de token enviadas
* ERC20MinVal\_Rec: valor mínimo en Ether recibido de las transacciones de token ERC20 para la cuenta
* ERC20MaxVal\_Rec: valor máximo en Ether recibido de las transacciones de token ERC20 para la cuenta
* ERC20AvgVal\_Rec: valor promedio en Ether recibido de las transacciones de token ERC20 para la cuenta
* ERC20MinVal\_Sent: valor mínimo en Ether enviado desde transacciones de token ERC20 para la cuenta
* ERC20MaxVal\_Sent: valor máximo en Ether enviado desde transacciones de token ERC20 para cuenta
* ERC20AvgVal\_Sent: valor promedio en Ether enviado desde transacciones de token ERC20 para la cuenta
* ERC20UniqSentTokenName: número de tokens ERC20 únicos transferidos
* RC20UniqRecTokenName: número de tokens ERC20 únicos recibidos
* ERC20MostSentTokenType:token más enviado para la cuenta a través de la transacción ERC20
* ERC20MostRecTokenType: token más recibido para la cuenta a través de transacciones ERC20

# Análisis Univariado

## Analisis general

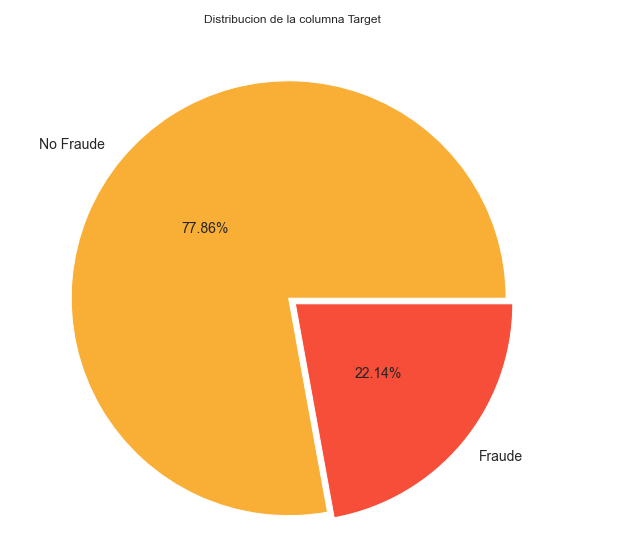
En el análisis general vemos que tenemos 51 variables, de estas solo **FLAG**, **ERC20 most sent token type** y **ERC20\_most\_rec\_token\_type** son variables categóricas.

Las variables **Unnamed:0**, **Index** y **Address** no se van a utilizar ya que son variables que o son índices o el address de la cuenta que no tienen ninguna relación con nuestra variable objetivo.

Después se hizo un df.describe() sobre las variables continuas, y encontramos algo muy curioso en algunas variables y es que desde el percentil 25 al 75 el valor es cero, esto nos está diciendo que para algunas variables el 75% o más de los datos son un cero. Más adelante se hará el análisis a fondo de esta variables.

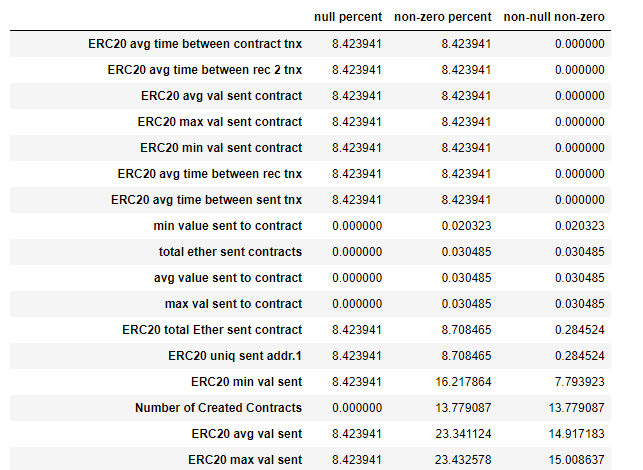
## Variable target

En nuestra variable target tenemos 2179 ejemplos de fraude y 9841 ejemplos totales, lo que nos da un 22% de ejemplos de fraude en el data set. Esto quiere decir que tenemos un dataset desbalanceado y que antes de hacer el modelo tendremos que emplear alguna técnica para balancear o si no podemos entrenar el modelo a pensar que todos los casos no son fraude.



## Nulos y valores en cero

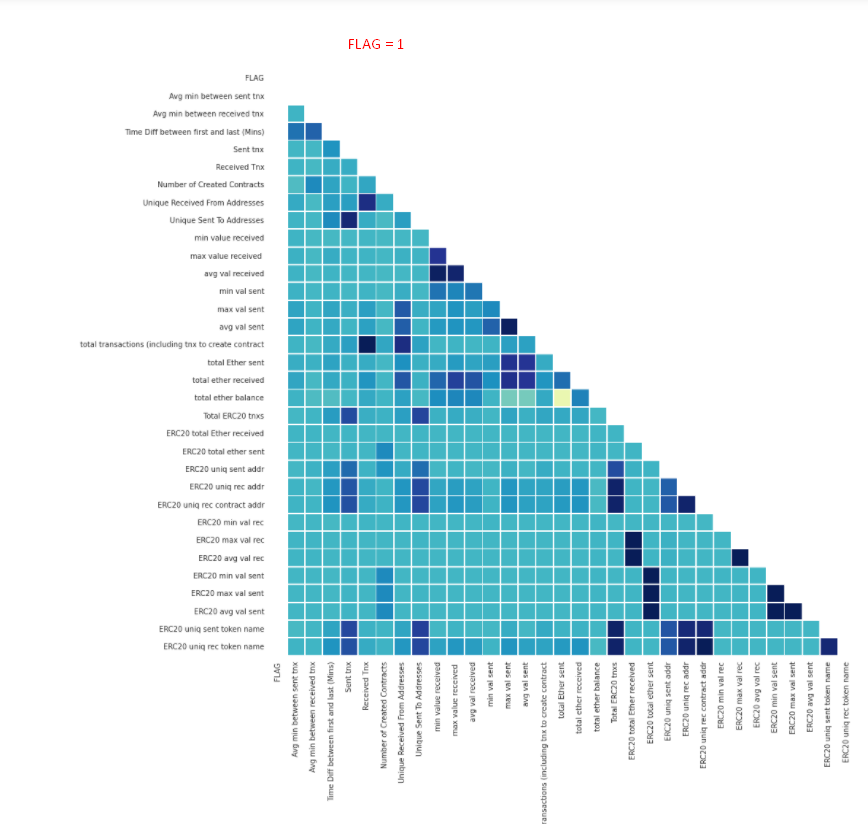
Se hizo un conteo de nulos y valores diferente a cero para cada una de las variables, esto ya que como vimos en el cuadro de describe habían unas variables que el el percentil 75 tenían cero, lo que nos da la idea de que esas variables pueden solo tenemos valores en cero y no aportar información al modelo.



Como se puede observar las primeras 13 variables de esta lista tienen menos del 0.3% de valores distintos a cero o nulo, es decir menos de 30 datos en un dataset de 9841 registros. Consideramos que estas variables no deberían entrar en el modelo porque no aportan información al problema ya que más del 99,7% de los datos van a ser nulos o cero.

Además observamos que las variables que tienen valores nulos son las de ERC20 que es un protocolo de ETH. Lo que parece indicarnos que esas cuentas no hicieron ninguna transacción por ese protocolo. Con esto pensamos agregar otra variables que indique si tiene información sobre las transacciones ERC20 o no.

# Correlación de la variable Target



Podemos observar las distintas correlaciones que tienen las variables entre sí para un Fraude o una transacción sin Fraude. Tenemos una fuerte correlación en variables que representa el protocolo de ETH.

## Limpieza variables categóricas

Para los siguientes variables categóricas reemplazamos el valor ‘0’ por Nan

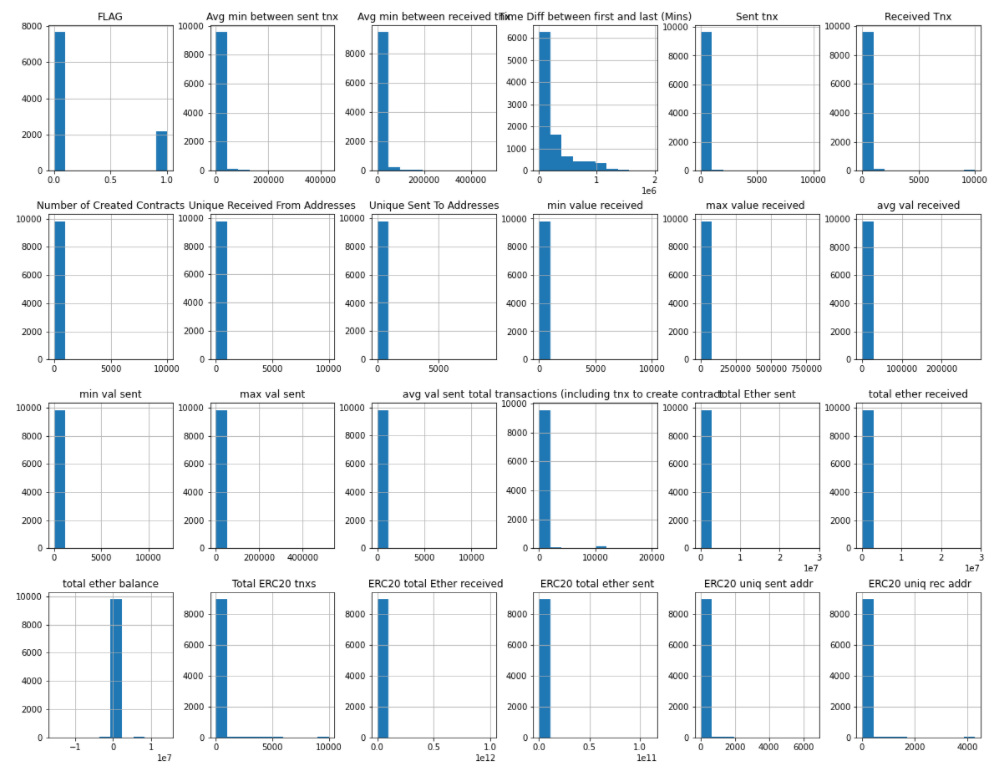
ERC20\_most\_rec\_token\_type

ERC20 most sent token type

Consideramos que tiene más sentido para una variable categórica que tenga un valor null a un valor numérico para representar la nada.

## Bar charts

En los Bar charts podemos observar que la mayoría de los valores tienen un mismo valor.

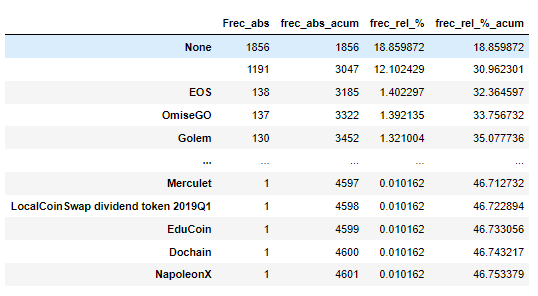


Como varias variables tienen una gran densidad de datos cerca de los valores a cero, estamos pensando en hacer dos tipos de modelos. uno que incluya esos valores extremos y otro sin esos valores extremos para ver cómo se comportan los modelos.

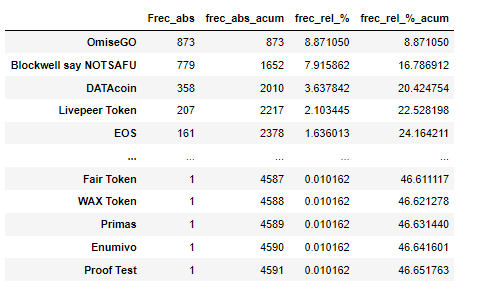
## 

## Análisis de frecuencias Categóricas

Variable : ERC20 most sent token typ



Variable : ERC20\_most\_rec\_token\_type

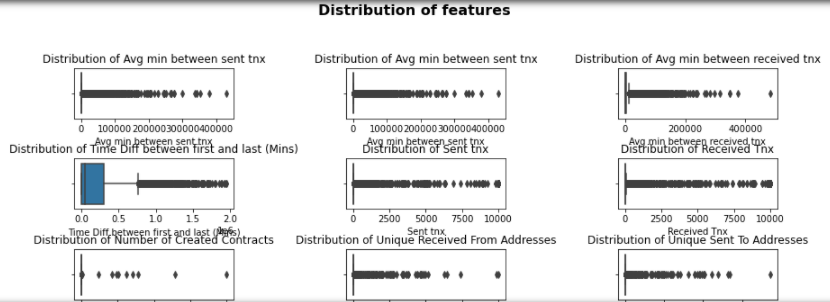
Para las 2 variables categóricas podemos observar que la frecuencia es en promedio baja.

## 

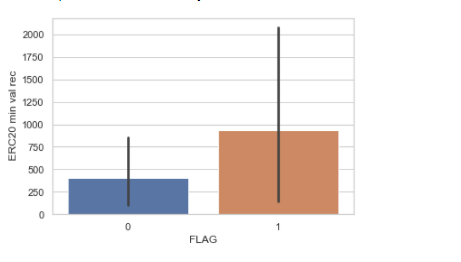
## 

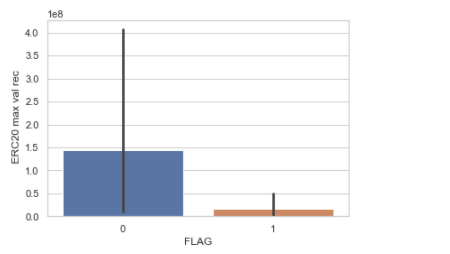
# Análisis Bivariado

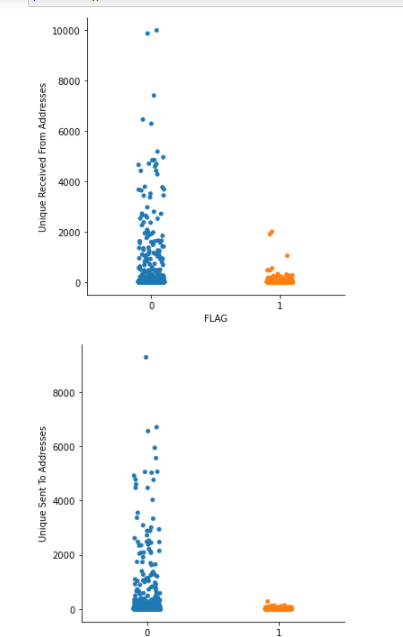
## Box plots

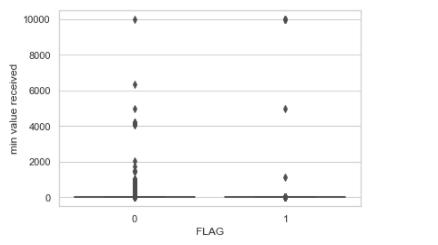


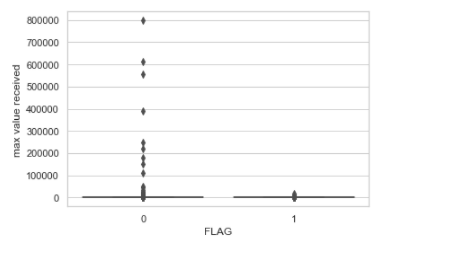
La gran mayoría de la dispersión tiene más al valor cero. Es una particularidad que se da en la mayoría de las variables









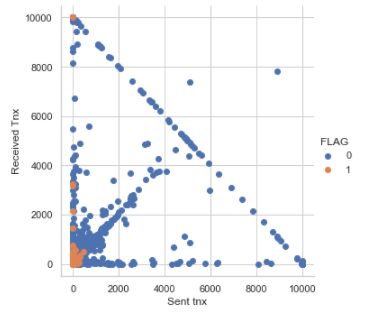


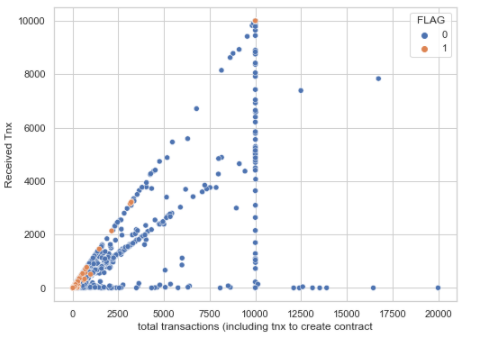
En la mayoría de las variables tenemos pares. Es decir , un valor máximo o mínimo. Podemos observar este patrón en la mayoría,para los valores máximos tenemos muy pocos fraudes. En cambio , para los valores mínimos tenemos muchas más Fraudes.

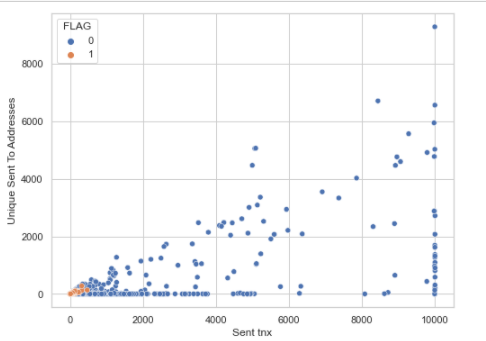
Con esto podemos concluir que la mayoría de los fraudes representan transacciones de menor costo.

# Análisis Multivariado

## 







Con todo el análisis multivariado que realizamos, podemos concluir que los fraudes se van en la mayoría de los valores mínimos , es decir, que más tienden a cero.

Podemos marcar más esta tendencia generan indicadores de corte

# 

# Indicadores

En base al análisis multivariado se generaron indicadores de corte

* **FLAG\_RECEIVED\_VS\_RECEIVED\_ADRESS** : Indicador de corte para las variables : Received Tnx , Unique Received From Addresses , FLAG
* **FLAG\_SENT\_ADRR\_VS\_SENT\_TRX** : Indicador de corte para las variables : Unique Sent To Addresses , Sent tnx ,FLAG
* **FLAG\_ERC20UNIQ\_VS\_ERC20\_TNXS** : Indicador de corte para las variables : ERC20 uniq sent addr , Total ERC20 tnxs , FLAG
* **FLAG\_TOTAL\_TRX\_VS\_RECEIVED\_TNX** : Indicador de corte para las variables : total transactions (including tnx to create contract, Received Tnx, FLAG