**Logotipo

Descripción generada automáticamente**

**Detección de instalaciones fraudulentas utilizando datos históricos de consumo energéticos**

Susana María Álvarez Castillo

Luis Miguel Román Pereira

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor  
David Manuel Villanueva Valdes, Título académico más alto

Universidad de Antioquia  
Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2023

|  |  |
| --- | --- |
| **Cita** | (Álvarez Castillo & Román Pereira, 2023) |
| **Referencia**  **Estilo APA 7 (2020)** | Álvarez Castillo, S. M., & Román Pereira, L. M. (2023). Detección de instalaciones fraudulentas utilizando datos históricos de consumo energéticos, Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia. |

**** 

Especialización en Analítica y Ciencia de Datos, CohorteVI.

Centro de Investigación Ambientales y de Ingeniería (CIA).

|  |  |
| --- | --- |
|  | Diagrama  Descripción generada automáticamente con confianza media |

Centro de Documentación Ingeniería (CENDOI)

**Repositorio Institucional:** http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano: Julio Cesar Saldarriaga Molina

Jefe departamento: Diego José Luis Botia Valderrama

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

**Dedicatoria**

….

**Agradecimientos**

…

**Tabla de contenido**

[Resumen 8](#_heading=h.3rdcrjn)

[Abstract 9](#_heading=h.26in1rg)

[1. Descripción del problema 10](#_heading=h.35nkun2)

[1.1. Problema de negocio 10](#_heading=h.1ksv4uv)

[1.2. Aproximación desde la analítica de datos 11](#_heading=h.44sinio)

[1.3. Origen de los datos 12](#_heading=h.z337ya)

[1.4. Métricas de desempeño 12](#_heading=h.3j2qqm3)

[2. Objetivos 14](#_heading=h.4i7ojhp)

[2.1. Objetivo general 14](#_heading=h.2xcytpi)

[2.2. Objetivos específicos 14](#_heading=h.lcp9lpsiy89b)

[3. Datos 15](#_heading=h.2bn6wsx)

[3.1. Datos originales 15](#_heading=h.qsh70q)

[3.2. Datsets 16](#_heading=h.3as4poj)

[3.3. Analítica descriptiva 18](#_heading=h.1pxezwc)

[**Referencias 22**](#_heading=h.hs2u70ebmc7f)

**Lista de tablas**

[Tabla 1: Dataset 18](#_heading=h.vhxnxdaefrzp)

[Tabla 2: Descripción de indicadores. 19](#_heading=h.ty5tshnsa8ro)

[Tabla 3: Ejemplo variables numéricas 20](#_heading=h.x8dxb0icvdc1)

**Lista de figuras**

[Figura 1: Distribución por clases 19](#_heading=h.p06kq7wxz6m7)

[Figura 2: Distribución variables numéricas relevantes 20](#_heading=h.qwnwczijjknj)

[Figura 3: Box plot variables y consumos facturados 20](#_heading=h.489otwyoz5hl)

[Figura 4: Distribución por clases 22](#_heading=h.lod5t8yjdwu)

**Siglas, acrónimos y abreviaturas**

**CREG** American Psychological Association

**SQL** Lenguaje de consulta estructurada

**KV** Kilovoltio

**NaN** Not a Number

# Resumen

El fraude de electricidad provoca importantes pérdidas económicas para las compañías encargadas de suministrar energía eléctrica, disminuyendo sus ingresos por consumos no facturados. Estos aumentos en los gastos derivados de prácticas fraudulentas suelen repercutir en última instancia en los consumidores, lo que se traduce en tarifas eléctricas más elevadas. Mediante esta investigación, nos proponemos abordar el desafío de reducir las pérdidas no técnicas, identificando patrones que puedan señalar la probabilidad de que una instalación presente irregularidades o fraudes en el suministro de energía eléctrica.

# Abstract

The electricity fraud causes significant economic losses for the companies responsible for supplying electricity, reducing their revenue due to unreported consumption. These increases in expenses resulting from fraudulent practices often ultimately impact consumers, translating into higher electricity tariffs. Through this research, we aim to tackle the challenge of reducing non-technical losses by identifying patterns that could indicate the likelihood of irregularities or fraud in the supply of electrical energy at a given installation.

# Descripción del problema

## Problema de negocio

En la mayoría de países, las regulaciones y sanciones relacionadas con el fraude y el robo de energía eléctrica y servicios públicos pueden variar, pero generalmente se considera un delito y puede dar lugar a multas sustanciales y, en casos graves, penas de prisión.

Por ejemplo, en Colombia, estas acciones son consideradas delito y pueden resultar en la imposición de cuantiosas multas y la condena a varios años de prisión:

ARTÍCULO 256. Defraudación de fluidos. El que mediante cualquier mecanismo

clandestino o alterando los sistemas de control o aparatos contadores, se apropie de

energía eléctrica, agua, gas natural, o señal de telecomunicaciones, en perjuicio ajeno,

incurrirá en prisión de dieciséis (16) a setenta y dos (72) meses y en multa de uno

punto treinta y tres (1.33) a ciento cincuenta (150) salarios mínimos legales mensuales

vigentes.

Las pérdidas de energía a lo largo de la cadena de producción, transporte y consumo eléctrico son conocidas. Estas pérdidas impactan en la cantidad de energía comprada y en los ingresos de distribución. Así, la reducción de pérdidas beneficia a los consumidores, con efectos positivos en la accesibilidad tarifaria.

Existen dos tipos de pérdidas:

1. **Pérdidas Técnicas:** Representan la energía disipada entre el suministro de la distribuidora y los puntos de entrega. Estas pérdidas se deben a las leyes físicas y a las inherentes a los equipos de medición.
2. **Pérdidas No Técnicas:** Son la diferencia entre las pérdidas totales y las técnicas. Incluyen robos, errores de medición y facturación, entre otros.

Adicionalmente, en Colombia, la normativa que regula el umbral de pérdidas no técnicas se establece principalmente a través de la Comisión de Regulación de Energía y Gas (CREG) y se aplica a las empresas de distribución y comercialización de energía eléctrica en el país. Estas regulaciones establecen los límites permitidos de pérdidas no técnicas y definen las medidas que las empresas deben tomar para reducir y controlar estas pérdidas. Si las empresas no cumplen con los umbrales establecidos, están sujetas a la imposición de multas, sanciones y medidas correctivas, de acuerdo con la normativa vigente en Colombia.

## Aproximación desde la analítica de datos

A través de una aplicación para la reducción de pérdidas, se tiene acceso a la base de datos de clientes consumidores y sus respectivas características incluyendo la variable objetivo “was\_fraud”, dicha variable nos indica si un cliente cometió o no fraude. En este contexto se decide optar por desarrollar un modelo de aprendizaje supervisado, específicamente un modelo de clasificación. El modelo resultante, permitiría anticipar y evaluar el riesgo de fraude de un cliente en particular, resultando beneficioso en los siguientes aspectos:

1. **Identificación temprana de potenciales casos de fraude:** Los modelos pueden ayudar a las empresas a detectar clientes que podrían estar más inclinados a cometer fraude, permitiendo tomar medidas preventivas o de seguimiento.
2. **Optimización de recursos:** Al identificar con mayor precisión los casos con mayor probabilidad de fraude, se pueden asignar recursos y esfuerzos de manera más efectiva para la verificación y seguimiento de estos casos específicos.
3. **Mejora en la toma de decisiones:** Los modelos predictivos brindan información valiosa para la toma de decisiones más informadas y estratégicas en cuanto a políticas de prevención de fraude y estrategias de mitigación de riesgos.
4. **Mejora en la eficiencia operativa:** Al reducir el número de casos de fraude no detectados, se optimiza la eficiencia operativa y se minimizan las pérdidas económicas asociadas a estas actividades fraudulentas

## Origen de los datos

A través de código SQL, en una base de datos MySql, se obtienen los datos de instalaciones consumidoras de energía, extrayendo las principales características y comportamientos que durante un periodo de tiempo estas han presentado. El resultado de esta extracción, es almacenado en un archivo plano, para su posterior análisis.

En el contexto de este trabajo, se llevó a cabo la extracción de datos correspondientes a los 2 últimos años de historia de inspecciones de fraude y el último año de inspecciones normales, Esta estrategia se emplea para abordar el desbalanceo entre las clases de fraude y no fraude.

## Métricas de desempeño

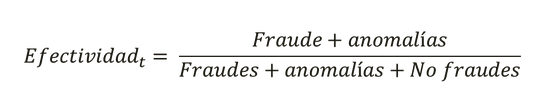
Métricas de machine learning: En la evaluación de los modelos se emplearán métricas como la exactitud, precisión, sensibilidad, y F1 score, para validar los resultados obtenidos:

* **Exactitud (Accuracy):** Proporción de todas las predicciones correctas, incluyendo tanto las instalaciones eléctricas identificadas correctamente como fraudulentas como aquellas identificadas correctamente como no fraudulentas.
* **Precisión (Precisión):** Porcentaje de las instalaciones eléctricas identificadas como fraudulentas que realmente son fraudes. Es decir, de todas las instalaciones eléctricas clasificadas como fraudulentas, ¿cuántas son realmente fraudes?
* **Recuperación (Recall o Sensibilidad):** Porcentaje de los casos reales de fraude que el modelo logra identificar correctamente. De todas las instalaciones eléctricas que realmente son fraudes, ¿cuántas fueron detectadas por el modelo?, esta métrica tiene más peso, respecto a las otras debido a que es importante capturar todos los casos fraudulentos.
* **Puntuación F1 (F1-Score):** Es una medida equilibrada que combina precisión y recall. Representa la precisión y completitud del modelo para detectar fraudes eléctricos. Ayuda a encontrar un equilibrio entre identificar correctamente fraudes y asegurarse de que las identificaciones positivas son realmente fraudes.

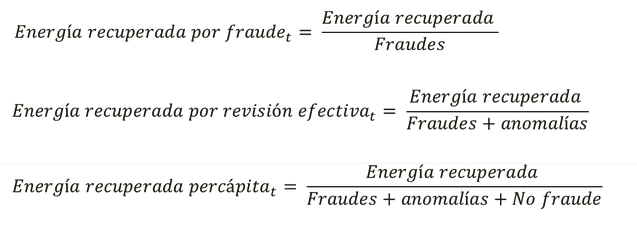
Métricas de negocio:

* Efectividad: El cálculo de efectividad contempla la relación de casos efectivos (positivos) en detección de fraudes e irregularidades, junto con el total de direccionamiento ejecutado. Se tiene como objetivo superar el 12% de efectividad actual de la empresa mediante la implementación del modelo propuesto en este proyecto.

A continuación la fórmula de cálculo:



* Índice de productividad: A través de los indicadores descritos a continuación, se espera superar los valores que actualmente presenta la compañía distribuidora de energía.



* Reducción de Pérdidas Económicas: Mide el impacto del modelo en la reducción de pérdidas económicas debidas al fraude. Compara las pérdidas antes y después de la implementación del modelo en términos económicos.

# Objetivos

## Objetivo general

Identificar pérdidas de energía no técnicas, mediante la reducción de casos de fraude, utilizando datos recopilados de inspecciones en un período de dos años.

## Objetivos específicos

* Identificar si existe relación significativa entre los patrones de consumo de energía o alguna correlación entre las diferentes variables, ubicación, tipo de cliente, tarifa, etc.
* Predecir instalaciones fraudulentas a través de un algoritmo de machine learning de clasificación.
* Evaluar la efectividad del modelo en la predicción de instalaciones fraudulentas
* Realizar un análisis descriptivo de los datos históricos de facturación y consumo de energía para alimentar el modelo y así indicar posibles casos de fraude.

# Datos

## Datos originales

La fuente principal de datos proviene de un software de reducción de pérdidas no técnicas, donde se almacenan las estructuras principales que describen las instalaciones consumidoras, su historial de consumo, revisión u otras órdenes de servicio, de una compañía de distribución de energía eléctrica en Chile. Esta información se guarda en un motor relacional de base de datos MySQL.

En el contexto de este trabajo, se llevará a cabo la extracción de datos, a través de código SQL, correspondientes a los 2 últimos años de historia de inspecciones de fraude y el último año de inspecciones normales, esta estrategia se usa para abordar el desbalanceo de clases desde la fuente, a través del submuestreo de los casos de 'no fraude'. Resultado de esta estrategia quedan disponibles 9277 registros, donde el 43% son datos etiquetados como fraude y 57% no fraude.

A continuación se describen las variables categóricas y numéricas que se extraen a través de una consulta sql desde la base de datos MySql:

Variables de Entrada:

Variables entrada categóricas: Presentan información acerca de la instalación eléctrica

|  |  |
| --- | --- |
| Categoría | Descripción |
| nivel\_de\_voltaje | Nivel de voltaje de la instalación. |
| actividad\_economica | Actividad económica asociada a la instalación. |
| provincia | Provincia de la instalación. |
| ID\_RISK\_AREA | Identificación de área de riesgo. |
| tipo\_cliente | Tipo de cliente. |
| tarifa | Tipo de tarifa del cliente. |

Variables de entrada numéricas: Son atributos que identifican el comportamiento de la instalación en el momento que se realizó la inspección, en el archivo existen 377 indicadores, ejemplos:

|  |  |
| --- | --- |
| Indicador | Descripción |
| N232 | Identifica el número de inspecciones en los 2M antes del último consumo. |
| N255 | Valor del consumo medido en la última facturación |
| N256 | Promedio del consumo medido en los últimos 3 meses |
| N266 | Cantidad de consumos medidos en los últimos 24 meses. |
| N285 | Cantidad de meses transcurridos entre la última revisión y la fecha del último consumo |
| N665 | Mínimo de consumo entre el primer consumo y cinco meses |

Variable de salida (Target):

was\_fraud: Indica si la instalación es fraudulenta o no. Variable binaria 1 indica fraude y 0 para no fraude.

## Datsets

Para la construcción del dataset, se procedió a una limpieza de datos. Se redujo el conjunto de datos original que constaba de 377 variables numéricas y 9105 registros a 68 variables y 9101 registros. Las estrategias empleadas se centraron en: Tratamiento de valores nulos. Eliminación de duplicados y variables numéricas con valores en cero (0), junto con la remoción de aquellas variables que mostraban una alta correlación.

Para los valores faltantes, se optó por imputar los valores numéricos utilizando la mediana, mientras que las variables categóricas se completaron con los valores más frecuentes. Además, se realizó un análisis de “outliers” con el fin de identificar posibles errores en la recopilación de datos, permitiendo validar la calidad de los datos y corregir posibles errores.

Se definió un diccionario llamado reemplazos que contiene pares clave-valor. Las claves son los valores originales que se encuentran en la columna 'tarifa', y los valores son los nuevos valores con los que se desea reemplazar. Esta transformación se usó para simplificar categorías reduciendo el número de agrupadores de 7 a 3 (Industria y comercio, Tarifa industrial y Hogar). Lo cual contribuyó a la optimización del análisis.

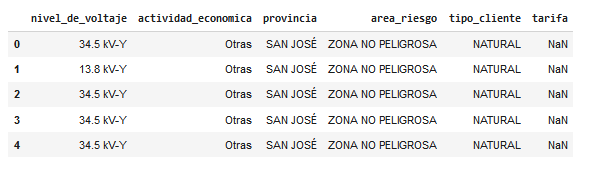
Para abordar el desbalance en la variable objetivo "was\_fraud" se utilizó la función de “Random Over Sampler” para aumentar el número de instancias de la clase minoritaria, mediante la creación de copias adicionales de esas instancias, hasta lograr un equilibrio deseado. Finalmente se crean variables dummy para representar de manera efectiva la información categórica en modelos que requieren datos numéricos como entrada.

## Analítica descriptiva

**Variables categóricas:**

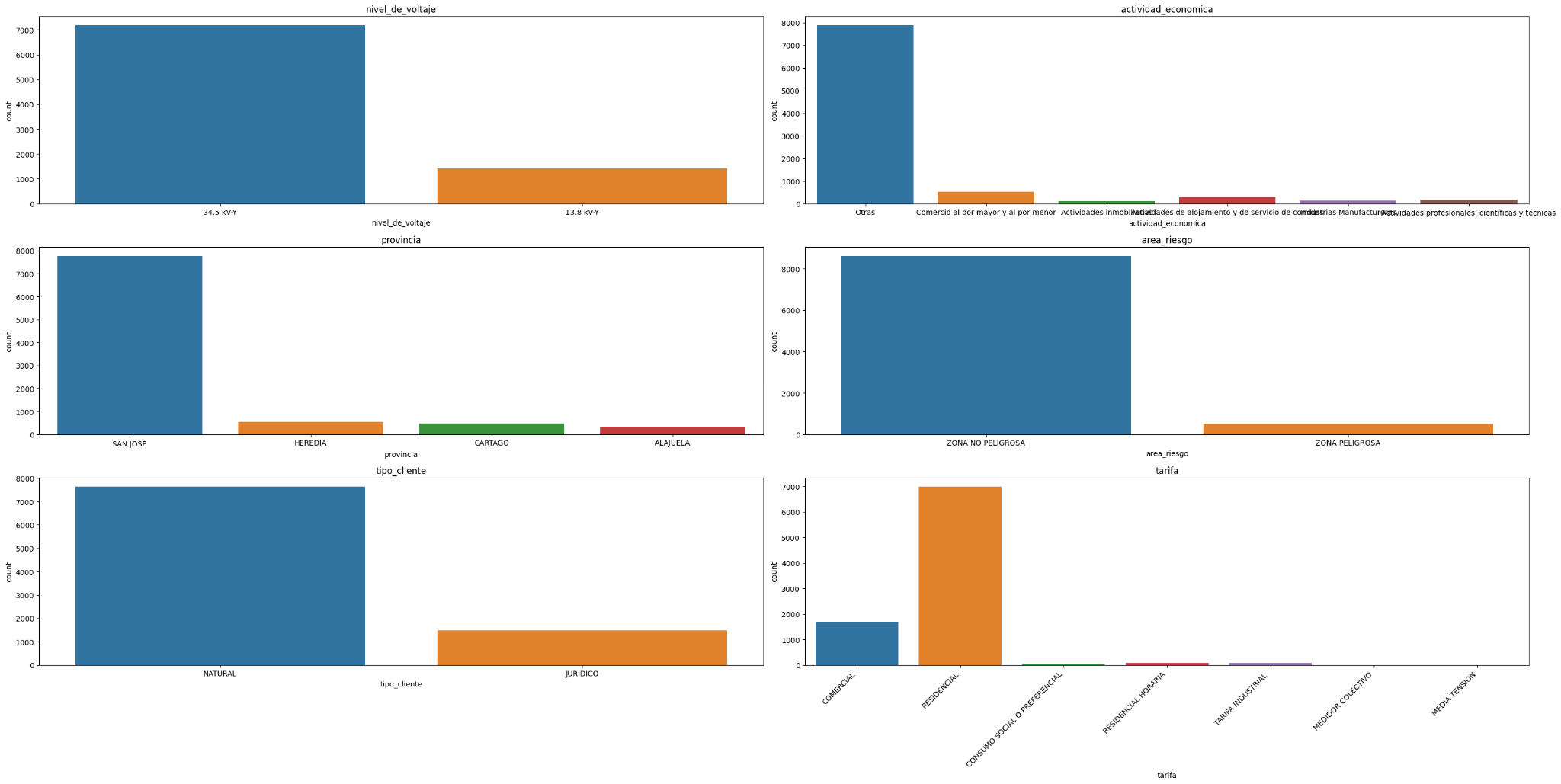
Entre las variables categóricas tenemos el nivel de voltaje, actividad económica, provincia, área de riesgo, tipo de cliente y tarifa, en la siguiente tabla se muestra algunos de los datos presentes en el dataset, se pueden observar algunos valores nulos.

##### [Tabla 1](#_heading=h.2u6wntf): Dataset



La siguiente Figura presenta la distribución de cada clase, en donde se observa cierto desbalance entre clases, por ejemplo actividad económica y tarifa presentan gran desequilibrio entre categorías:

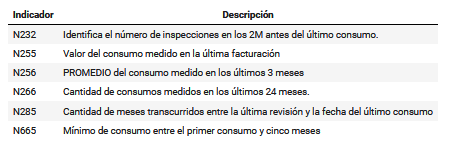
#### Figura 1: Distribución por clases



**Variables numéricas:**

Son atributos que identifican el comportamiento de la instalación en el momento que se realizó la inspección, en el archivo existen 377 indicadores, ejemplos:

##### [Tabla](#_heading=h.2u6wntf) 2: Descripción de indicadores.

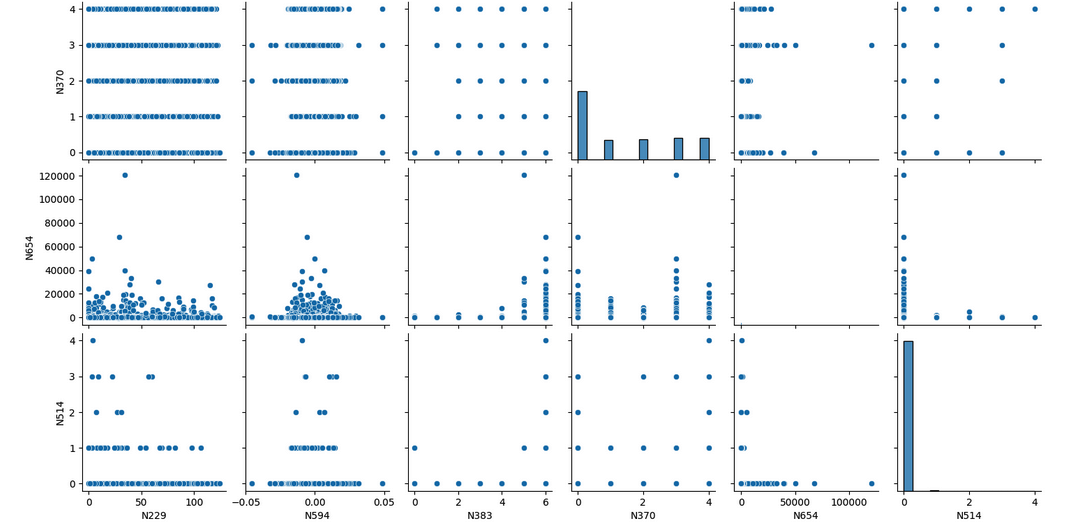


La siguiente tabla muestra alguno de los valores presentes en las variables numéricas:

##### [Tabla](#_heading=h.2u6wntf) 3: Ejemplo variables numéricas

A continuación, se presenta la distribución de algunas de las variables numéricas:

#### Figura 2: Distribución variables numéricas relevantes



#### Figura 3: Box plot variables y consumos facturados

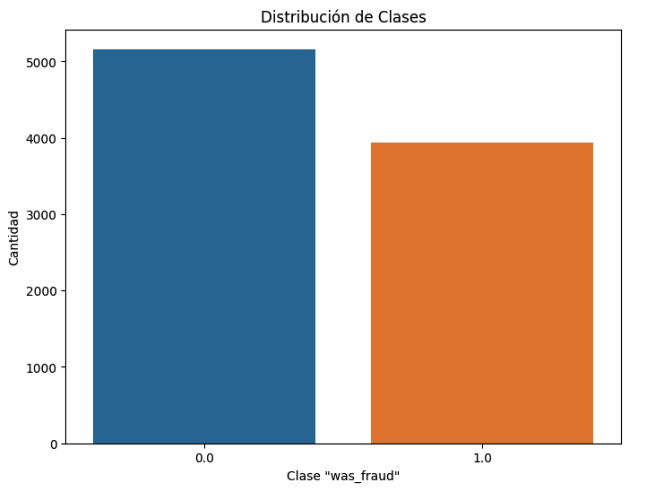
# 

Al tener una gran dimensionalidad, el reto durante la etapa de procesamiento se centrará en disminuir variables a través de tratamiento de valores nulos, eliminación de duplicados y variables numéricas con valores en 0 y remoción de aquellas variables que muestran una alta correlación.

**Variable de salida:**

was\_fraud: Indica si la instalación es fraudulenta o no. Variable binaria 1 indica fraude y 0 para no fraude, se encuentran 5167 casos de no fraude y 3938 con fraude:

#### Figura 4: Distribución por clases



# 

# 

# Referencias

Carrillo Rosales, A. J. (2019). Sistema inteligente basado en Machine Learning para la detección de fraude de facturación de agua potable.

Quintero Acuña, L. K. (2023). Aplicación de Machine Learning a un modelo tradicional de Prevención y detección de fraude en entidad financiera proyectado periodos trimestrales.

Bustos, T. J. P. P. (2021). Estrategia para la conceptualización de modelos de IA en el contexto de Gestión de la Energía.

Alvarez, F. (2020). Machine Learning en la detección de fraudes de comercio electrónico aplicado a los servicios bancarios. Ciencia y tecnología, 81-95.

Castro, C. B. M., & Alonso, J. A. V. (2022). Métodos y técnicas de Machine Learning e Inteligencia Artificial para el enfrentamiento al fraude en las Telecomunicaciones: Técnicas de minería de datos aplicadas a las gestión del fraude. Revista Cubana de Transformación Digital, 3(4), e182-e182.