**CREACION DE UN MODELO PREENTRENADO QUE DETECTE TRANSACCIONES FRAUDULENTAS DE UN NEGOCIO TIPO ECOMMERCE**

**Metodología KDD (*Knowledge Discovery in Database*)**

Metodología desarrollada dentro un proceso iterativo que explora grandes cantidades de datos para determinar las relaciones entre estos, con la finalidad de encontrar información útil dentro de los repositorios de datos de una empresa.

EL proceso lo podremos describir en 6 pasos:

**1.- Selección de datos**

Consiste en determinar los datos a utilizar, como también las fuentes de información de donde se realiza la extracción de los datos.

Fuente de los datos: Utilizamos un repositorio de kaggle donde estos datos tienen más de 20 millones de transacciones generadas a partir de una simulación de mundo virtual multiagenterealizada por IBM.

*“Los datos aquí casi no tienen ofuscación y se proporcionan en un archivo CSV cuyo esquema se describe en la primera fila. Los datos cubren 2000 consumidores (sintéticos) residentes en los Estados Unidos, pero que viajan por el mundo. Los datos también cubren décadas de compras e incluyen múltiples tarjetas de muchos de los consumidores.”* (Kaggle.com)

**2.- Preprocesamiento.**

**Herramientas utilizadas:** Python, SQL Server, Power BI, Visual Studio 2022

Donde se realiza todos los procesos de limpieza de los datos ya extraídos. Generalmente es aquí donde algunos datos sufren transformaciones para que su posterior tratamiento sea más fácil de procesar, esta etapa es también llamada ETL.

**2.1.- Extracción:** La base de datos seleccionada contaba un número muy grande de transacciones por lo que se extrajo una muestra aleatoria de 2 millones de transacciones con un script de Python, de las cuales apenas 0.13% son transacciones fraudulentas, luego se seleccionan las variables críticas para un nuevo análisis.

https://www.kaggle.com/datasets/ealtman2019/credit-card-transactions/data

**Descripción de los datos Extraídos**

|  |  |  |
| --- | --- | --- |
| **Descripción de la tabla de datos MuestraTransacciones** | | |
| **Columna** | **Tipo de dato** | **Descripción** |
| User | Int | ID de los usuarios registrados |
| Card | Int | Número de tarjetas que posee un usuario |
| Year | Int | Año de Transacción (2006-2020) |
| Month | Int | Mes de la Transacción (1-12) |
| Day | Int | Día de la Transacción (1-31) |
| Time | Time | Hora de la transacción |
| Amount | Money | Monto de la transacción |
| Use\_Chip | VarChar(50) | Tipo de transacción |
| MerchantName | VarChar(50) | ID de tienda o comercio |
| Merchan\_City | VarChar(50) | Ciudad de ubicación del comercio |
| Merchant\_State | VarChar(50) | Siglas del Estado ubicación del comercio |
| Zip | VarChar(50) | Código ZIP de la ubicación del comercio |
| MCC | VarChar(50) | Código categoría de mercado |
| Errors | VarChar(50) | Tipo de error detectado en la transacción |
| Is\_Fraud | VarChar(50) | Si la transacción es un fraude o no |
| TransaccionID | Int | ID de las transacciones registradas (Primary key) |

Tabla 1 -descripción de la tabla MuestraTransacciones

|  |  |  |
| --- | --- | --- |
| **Descripción de la tabla Tarjetas** | | |
| **Columna** | **Tipo de dato** | **Descripción** |
| User | Int | ID de los usuarios registrados |
| Card\_index | int | Index de las tarjetas que posee un usuario |
| BancoCredito | VarChar(50) | Nombre de empresa crediticia |
| TipoTarjeta | VarChar(50) | Tipo de tarjeta |
| NumeroTarjeta | Bigint | Número de tarjeta (Primay Key) |
| Expires | VarChar(50) | Fecha de expiración de la tarjeta |
| CVV | Int | Código CVV del reverso de la tarjeta |
| Has\_Chip | VarChar(50) | Tipo de transacción |
| Cards\_Issued | Int | Número de tarjetas emitidas |
| LimiteCredito | Money | Límite de crédito en tarjeta |
| Acct\_Open\_Date | VarChar(50) | Fecha apertura de cuenta |
| Year\_PIN\_last\_Changued | Int | Fecha del último cambio del PIN |
| Card\_on\_Dark\_web | VarChar(50) | SI la tarjeta se encuentra reportada en la DarkWeb |

Tabla 2 - Descripción tabla Tarjetas

|  |  |  |
| --- | --- | --- |
| **Descripción de la tabla Usuarios** | | |
| **Columna** | **Tipo de dato** | **Descripción** |
| Person | VarChar(50) | Nombre del Cliente |
| Current\_Age | Int | Edad del cliente |
| Retirement\_Age | Int | Edad de retiro del cliente |
| Brith\_Year | Int | Año de nacimiento |
| Brith\_Month | Int | Mes de nacimiento |
| Brith\_Day | Int | Día de Nacimiento |
| Gender | VarChar(50) | Genero |
| Dirección | VarChar(50) | Dirección |
| Apartment | VarChar(50) | Numero de apartamento |
| Ciudad | VarChar(50) | Ciudad de residencia |
| State | VarChar(50) | Estado de residencia |
| Zipcode | Int | Código ZIP Dirección cliente |
| Latitude | Float | Latitud GPS |
| Longitude | Float | Longitud GPS |
| IngresoPerCapita | Money | Ingreso Per Capita |
| IngresoAnual | Money | Total Ingreso anual |
| DebitoTotal | Money | Total de débitos anuales |
| FICO\_Score | Int | Calificación FICO de score crediticio |
| NumeroTarjetasCredito | Int | Cantidad de tarjetas que posee el usuario |
| UserID | Int | ID de los usuarios registrados (Primay Key) |

Tabla 3 - Descripción tabal Usuarios

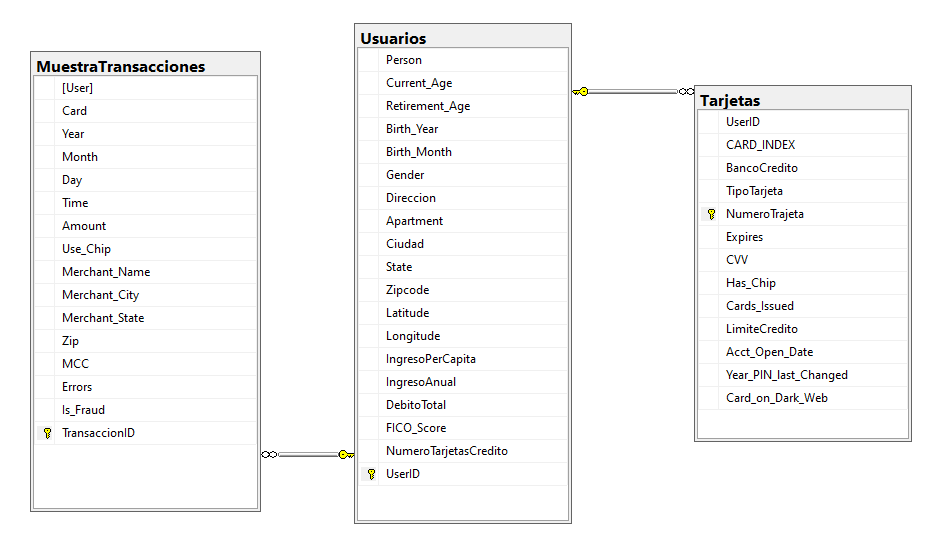


Ilustración 1-Diagrama de base datos original

**2.2.- Transformación:** Los datos son transformados en formato que se pueda cargar en la base de datos, la observación, sondeo y cambio del tipo de datos de cada columna y el procesamiento de datos nulos y erróneos usando SQL Server.

**2.3.- Carga de los datos:**

los datos son replicados en un nuevo almacén de datos FraudulentECoommerce agrupados en 4 tablas Dimensionales (DIM\_USUARIO\_DEST, DIM\_TARJETA\_DEST, DIM\_TIEMPO\_DEST, DIM\_COMERCIO\_DEST) y una tabla de Hechos (H\_TRANSACCION\_DEST), ya se puede hacer diversas gráficas para visualizar tendencias y relaciones entre variables, el modelo ETL permite hacer consultas más complejas y con más detalle, como los que se presentan más adelante.

**Descripción de los datos cargados**

|  |  |  |
| --- | --- | --- |
| **Descripción de la tabla DIM\_USUARIO\_DEST** | | |
| **Columna** | **Tipo de dato** | **Descripción** |
| UserID | Int | ID de los usuarios registrados (Primay Key) |
| Person | VarChar(50) | Nombre del Cliente |
| Gender | VarChar(50) | Genero |
| Direccion | VarChar(50) | Dirección |
| Ciudad | VarChar(50) | Ciudad de residencia |
| IngresoPerCapita | Money | Ingreso Per Capita |
| IngresoAnual | Money | Total Ingreso anual |
| DebitoTotal | Money | Total de débitos anuales |
| FICO\_Score | Int | Calificación FICO de score crediticio |

Tabla 4 - Descripción de la tabla Dimensional Usuario

|  |  |  |
| --- | --- | --- |
| **Descripción de la tabla DIM\_TARJETA\_DEST** | | |
| **Columna** | **Tipo de dato** | **Descripción** |
| Número de tarjeta | Bigint | Número de tarjeta (Primay Key) |
| BancoCredito | VarChar(50) | Nombre de empresa crediticia |
| TipoTarjeta | VarChar(50) | Tipo de tarjeta |
| Has\_Chip | VarChar(50) | Tipo de transacción |
| LimiteCredito | Money | Límite de crédito en tarjeta |

Tabla 5 - Descripción de la tabla dimensional Tarjeta

|  |  |  |
| --- | --- | --- |
| **Descripción de la tabla DIM\_TIEMPO\_DEST** | | |
| **Columna** | **Tipo de dato** | **Descripción** |
| TiempoID | Int | Índice de los registros de tiempo (Primay Key) |
| Year | Int | Año de Transacción (2006-2020) |
| Month | Int | Mes de la Transacción (1-12) |
| Day | Int | Día de la Transacción (1-31) |
| Time | Time | Hora de la transacción |
| Dia\_Semana | VarChar(20) | Día de la semana registrada en la transacción |

Tabla 6 - Descripción de la tabla dimensional Tiempo

|  |  |  |
| --- | --- | --- |
| **Descripción de la tabla DIM\_COMERCIO\_DEST** | | |
| **Columna** | **Tipo de dato** | **Descripción** |
| ComercioID | Int | Índice de los Comercio registrados (Primary Key) |
| MerchantName | VarChar(50) | ID de tienda o comercio |
| Merchan\_City | VarChar(50) | Ciudad de ubicación del comercio |
| Merchant\_State | VarChar(50) | Siglas del Estado ubicación del comercio |
| Zip | VarChar(50) | Código ZIP de la ubicación del comercio |

Tabla 7 - Descripción de la tabla dimensiona Comercio

|  |  |  |  |
| --- | --- | --- | --- |
| **Descripción de la tabla H\_TRANSACCION\_DEST** | | | |
| **Columna** | **Tipo de dato** | | **Descripción** |
| UserID | Int | ID de los usuarios registrados (Primay Key) | |
| Número de tarjeta | Bigint | Número de tarjeta (Primay Key) | |
| TiempoID | Int | Índice de los registros de tiempo (Primay Key) | |
| ComercioID | Int | | Índice de los Comercio registrados (Primary Key) |
| Amount | Money | | Monto de la transacción |
| Errors | VarChar(50) | | Tipo de error detectado en la transacción |
| Is\_Fraud | VarChar(50) | | Si la transacción es un fraude o no |

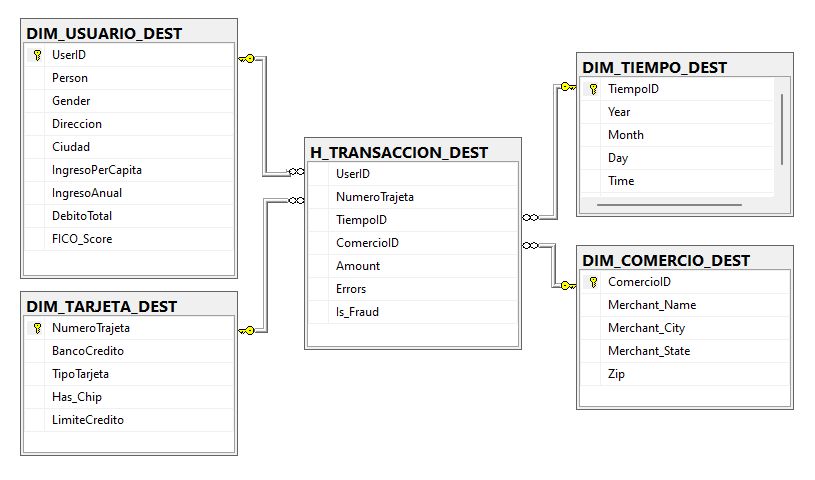


Ilustración 2- Modelo Estrella de la base de datos FraudulentECommece

Adicionalmente en este paso se hizo un chequeo exploratorio de algunas variables y sus incidencias en relación a la variable objetivo

Según el muestreo se puede observar relaciones muy evidentes en relación a los meses, días y picos de actividad en ciertas horas del día, datos que pueden servir para optimizar recursos en el monitoreo de transacciones en los tiempos determinados.

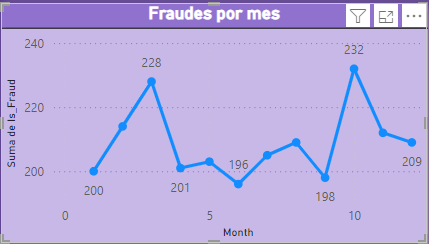


Ilustración 3 - Tabla de relación Mes - Transacciones Fraudulentas

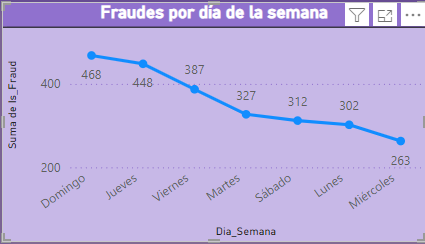


Ilustración 4 - Tabla de relación Día de la semana - Transacciones Fraudulentas

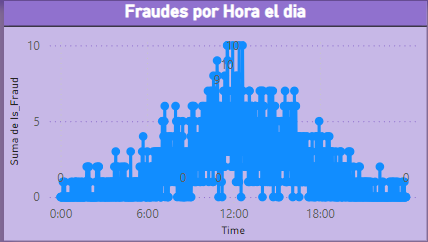


Ilustración 5 - tabla de relación Hora del día - Transacciones Fraudulentas

**3.- Transformación**

**Herramientas utilizadas:** Python, Google Colab

todos los cambios que no se realizaron en el paso anterior son llevados a cabo en esta etapa, donde el dato se prepara finalmente para poder ser procesado por algún algoritmo de minería de datos. Las operaciones más comunes realizadas son de agregación de datos o normalización.

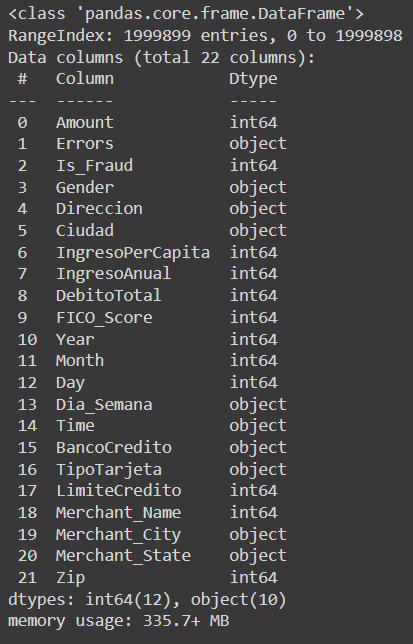


Ilustración 6 - Información de las variables utilizadas en la tabla de datos

**3.1.- Eliminación de columnas innecesarias:**

Las columnas que no tomaremos en cuenta para el entrenamiento del modelo de predicción serán en primera instancia los datos geográficos y Gender, ya que, aunque en el análisis exploratorio se detectó lugares donde existe mayor incidencia de actividades fraudulentas, no creemos que será relevante para identificar su legitimidad, además de la columna día por ser un numero ordinal entre 1 y 31 que podría causar confusión en el entrenamiento

|  |  |
| --- | --- |
| Columnas eliminadas | Tipo de dato |
| Dirección | Object |
| Day | Int |
| Merchant\_Name | Object |
| Merchant\_city | Object |
| Merchant\_State | Object |
| Zip | Int |
| Ciudad | Object |
| Gender | Object |

Tabla 8 - Columnas eliminadas del Data Frame

**3.2.- transformación de los datos**

Procedemos con la transformación de datos de las variables categóricas de tipo object que no van a ser reconocidas en el entrenamiento como son: Errors, Dia\_Semana, Time, BancoCredito, TipoTarjeta. Las transformaciones serán de tipo Encoding OHE (One Hot Encodig), lo que hace es crear nuevas columnas iguales a cada categoría de las variables a transformar dándoles valores binarios, esto además de darle valores numéricos a las variables también aumenta la dimensión de la Data Frame.

|  |  |  |  |
| --- | --- | --- | --- |
| Columnas Transformadas | Tipo de dato | Tipo Transformación | Columnas creadas |
| Errors | Object | OHE | SIN ERROR, Insufficient Balance, Technical Glitch, Bad PIN,  Bad Card Number, Bad CVV, Bad Expiration, Bad Zipcode |
| BancoCredito | Object | OHE | Amex, Mastercard, Visa, Discover |
| TipoTarjeta | Object | OHE | Credit, Debit, Debit (Prepaid) |
| Dia\_Semana | Object | OHE | Miércoles, Domingo, Martes, Lunes, Sábado, Jueves, Viernes |
| Time | Object | Label\_encoder | Time |

Tabla 9 - Columnas transformadas

**4.- Minería de datos**

corresponde al modelo propiamente tal, en donde los algoritmos son aplicados con el objetivo de encontrar las relaciones que puedan existir

**5.- Interpretación**

finalmente, se identifican los patrones encontrados en la etapa anterior que puedan ser de utilidad. Todo esto bajo métricas de evaluación que deben ser consistente al objetivo de búsqueda.

**6.- Evaluación**

Evaluación del modelo