Grupo Bimbo – Dirección Global de Auditoría Interna.

Gerencia de Auditoría de TI & Analítico

Caso práctico

Presenta:

Jessica Olivares Lopez

Junio 2024

Índice General

1 Introducción 3

1.1 Objetivos 3

1.1.1 Objetivo general 3

1.1.2 Objetivos específicos 3

2 Exploración de datos 3

2.1 Exploración de datos **¡Error! Marcador no definido.**

2.1.1 Aperturas 3

2.1.2 Movtos_Ctas_2015 5

2.1.3 Relación de los dos conjuntos de datos 12

3 Implementación de estrategias para la identificación de operaciones atípicas 13

3.1.1 Conjunto de datos auxiliar 13

3.1.2 Análisis Univariado 13

3.1.3 Análisis Multivariado 17

4 Conclusiones 19

5 Bibliografía **¡Error! Marcador no definido.**

1 Introducción

En el presente flujo de trabajo se obtienen diferentes observaciones respecto a la identificación de datos que pueden provocar un daño reputacional a la empresa o un incumplimiento de obligaciones legales o fiscales, a partir de la identificación de observaciones (transacciones) y cuentas atípicas de acuerdo con el comportamiento de los datos.

1.1 Objetivos

1.1.1 Objetivo general

Implementar y evaluar estrategias de análisis que permitan identificar transacciones fraudulentas o desviaciones en procesos de negocio.

1.1.2 Objetivos específicos

1. Realizar cinco pruebas de análisis de datos, con objetivo y descripción.
2. Realizar una valoración de resultados y conclusiones.

2 Exploración de datos

En esta sección se muestra el proceso de exploración y limpieza de datos, con el objetivo de entender la estructura y características de la información disponible, considerando dos fuentes de datos que se describen a continuación.

1. Aperturas: Contiene registros de los movimientos de aperturas realizadas desde el año 2013 a 2015.
2. Movtos_Ctas_2015: Registros de las transacciones realizadas en el año 2015 con los tipos de transacción:
    a. Depósitos en cheque (DCHQ)
    b. Depósitos en efectivo (DEFE)
    c. Retiro efectivo cajero (RECJ)
    d. Retiro efectivo ventanilla (REVN)

2.1.1 Aperturas

Este conjunto de datos está integrado por 5 variables, como se puede observar en la Figura 1 y descrita en la tabla 1.

Tabla 1. Descripción de conjunto de datos Aperturas

Variable	Núm. de valores únicos	Observación
Numero_de_cuenta	6980	Identificador de cuenta (variable discreta)
Monto_de_apertura	6889	Monto inicial (variable continua)
Fecha_Apertura	578	Fecha de apertura
Cliente	5809	Identificador de Cliente (variable discreta)
Tipo_de_Instrumento	2	Tipo de instrumento (cuenta) (variable categórica)

	Numero_de_cuenta	Monto_de_apertura	Fecha_Apertura	Cliente	Tipo_de_Instrumento
0	00MH985399	16046.40	3/6/2014	100540012	Cuenta corriente
1	00CU982321	379072.47	11/24/2014	100540013	Cuenta corriente
2	00CU982323	1603.86	2/12/2013	100540014	Cuenta corriente
3	00CU982324	3207.72	2/19/2013	100540015	Cuenta corriente
4	00CU982509	880377.46	9/9/2013	100540016	Cuenta corriente

Figura 1. Primeras 5 observaciones del conjunto de datos que contiene el registro de Aperturas.

	count	mean	std	min	25%	50%	75%	max
Monto_de_apertura	6980.0	1.678120e+06	5.189515e+07	-944104707.1	1.525060e+03	13071.5	9.400463e+04	2.926603e+09
Cliente	6980.0	1.005433e+08	2.196100e+03	100540010.0	1.005411e+08	100543510.5	1.005453e+08	1.005470e+08

Figura 2. Descripción del conjunto de datos Aperturas en base a las variables numéricas.

Observaciones:

1.- Se tiene un mayor número de cuentas en comparación del conteo de clientes en el banco.

Se deduce que se tienen casos en los que un cliente tiene más de una cuenta asociada, tomando en cuenta que hay dos tipos de instrumentos o cuentas disponibles; Cuenta corriente y Cuenta de Ahorro.

Mediante el uso de pandas, se realizaron los respectivos filtros para agrupar el conjunto de datos por 'Cliente', y obtuvo que, se tienen clientes con 1, 2 y 3 cuestas asociadas. Por lo que, los clientes con 3 cuentas, se consideran cómo situaciones anormales, teniendo un total de **38** Clientes en esta situación.

```
clientes_con_3_cuentas.describe().T
```

	count	mean	std	min	25%	50%	75%	max
Cliente	38.0	1.005401e+08	50.489447	100540018.0	1.005400e+08	100540127.5	1.005401e+08	100540148.0
num_cuentas	38.0	3.000000e+00	0.000000	3.0	3.000000e+00	3.0	3.000000e+00	3.0

Figura 3. Descripción del subconjunto de clientes con 3 cuentas aperturadas en el banco.

2.- El valor mínimo para el Monto de apertura es negativo.

Al abrir una cuenta bancaria, generalmente se requiere un depósito inicial mínimo, que varía según el banco y el tipo de cuenta. Las cuentas deben comenzar con un saldo positivo o al menos cero. Una cifra negativa no es permitida para la apertura de cuentas, ya que implica una deuda o un descubierto, algo que no se puede establecer como condición inicial para una nueva cuenta. Sin embargo, se pone a consideración del banco Dräxlmaier, para validar si esta observación se ajusta a los requerimientos del proceso de apertura de cuenta.

Se obtiene el subconjunto de cuentas en esta situación en (df_MontoNegativo) obteniendo un total de **854** cuentas aperturadas en esta situación.

2.1.2 Movtos_Ctas_2015

Este conjunto de datos está integrado por 8 variables, como se puede observar en la Figura 1 y descrita en la tabla 2.

Tabla 2. Descripción de conjunto de datos de registros de las transacciones realizadas en el año 2015.

Variable	Núm. de valores únicos	Observación
Numero_de_cuenta	6990	Identificador de cuenta (variable discreta)
M_transaccion	31891	Monto de la transacción (variable continua)
Sucursal_nombre	7	Sucursal donde se efectuó la transacción (Variable categórica)
Fecha_Transaccion	331	Fecha de la transacción
Hora_transaccion	3785	Hora de la transacción
T_transaccion	4	Indica el tipo de transacción (variable categórica)
Operador	152	Identificador del operador (variable discreta)
Referencia	65532	Identificador de la operación (variable discreta)

	Numero_de_cuenta	Monto_transacción	Sucursal_nombre	Fecha_Transaccion	Hora_transaccion	Tipo_de_Transacción	Operador	Referencia
0	00CU962253	0.0	CUAUHTEMOC	08/01/2015	05:13:49 a. m.	DEFE	CU5027	3065181
1	00FO960212	0.0	FORÁNEAS	06/01/2015	05:13:46 a. m.	DEFE	FO7506	3065264
2	00MH961270	0.0	MIGUEL HIDALGO	02/05/2015	05:13:51 a. m.	DEFE	MH5507	3065482
3	00FO989961	0.0	FORÁNEAS	4/29/2015	01:18:20 a. m.	DEFE	FO7507	3064174
4	00MH960322	0.0	MIGUEL HIDALGO	9/23/2015	01:46:25 a. m.	DEFE	MH5506	3064514

Figura 4. Primeras 5 observaciones del conjunto de datos de registros de las transacciones realizadas en el año 2015.

	count	mean	std	min	25%	50%	75%	max
M_transaccion	65534.0	3.193806e+05	1.673933e+07	0.0	633.60	2875.89	13154.55	2.924383e+09
Referencia	65534.0	3.032767e+06	1.891805e+04	3000001.0	3016383.25	3032766.50	3049149.75	3.065532e+06

Figura 5. Descripción del conjunto de datos de registros de las transacciones realizadas en el año 2015 para variables numéricas.

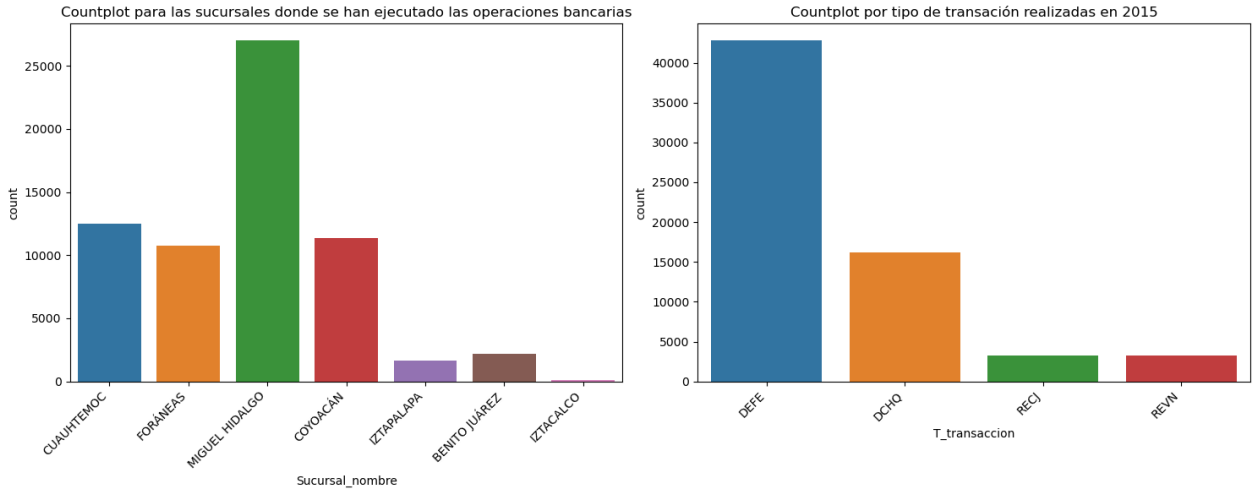

Figura 6. Countplot de las variables categóricas Sucursal_nombre y T_transaccion.

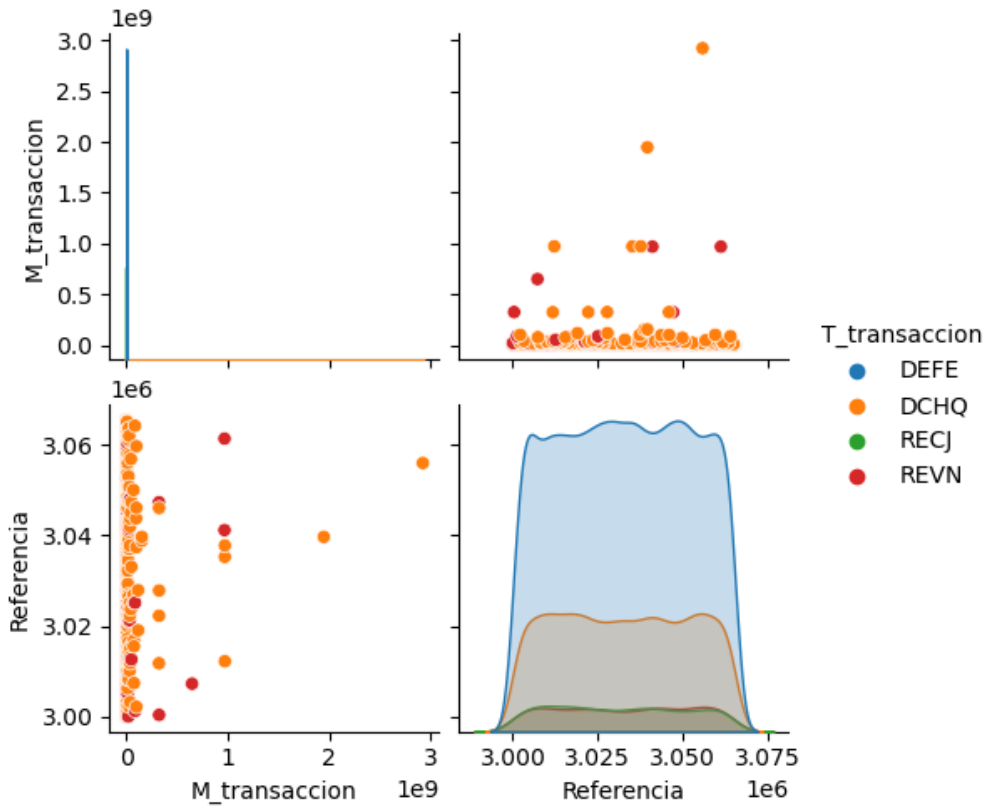

Figura 7. Pairplot de las variables categóricas Sucursal_nombre y T_transaccion.

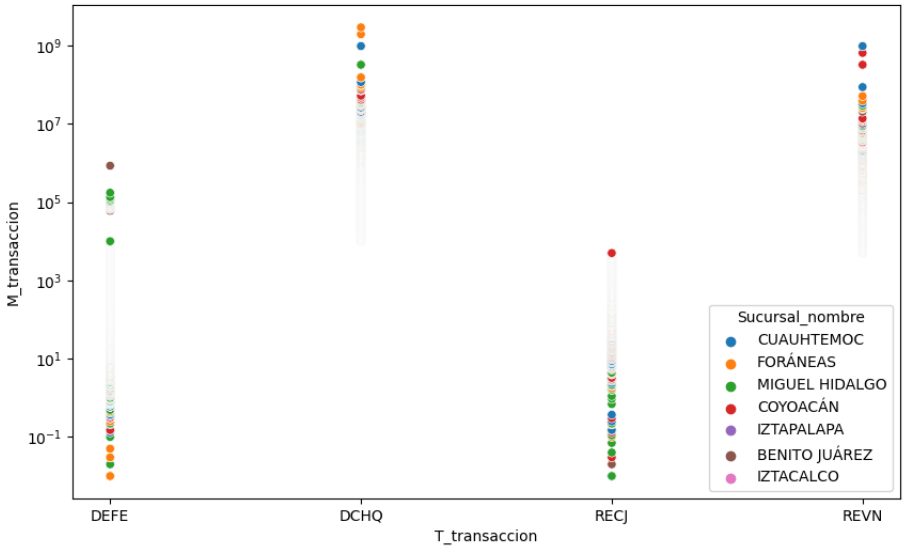

Figura 8. Grafica de distribución de las variables Sucursal_nombre y T_transaccion a partir de M_transaccion.

Observaciones:

1.- En base a la tabla 2. Se tiene menor número de valores únicos para las referencias de transferencia (65532) en comparación del número de registros en el conjunto de datos (65534).

Se realizado un filtrado de los valores duplicados con padas, para revisar los registros y determinar si se trata de un registro totalmente duplicado, o si se tienen dos operaciones asignadas a una misma referencia.

```
duplicados_referencia = df[df['Referencia'].duplicated()]
print(duplicados_referencia)
```

```
      Numero_de_cuenta  M_transaccion Sucursal_nombre Fecha_Transaccion  \
29874       00MH988235        2257.86  MIGUEL HIDALGO         8/24/2015
47269       00BJ962502       10533.54   BENITO JUÁREZ        05/05/2015

      Hora_transaccion T_transaccion Operador  Referencia
29874   04:08:42 a. m.          DEFE   MH5506     3004637
47269   01:49:04 a. m.          DCHQ   BJ0109     3064984
```

Figura 9. Referencias identificadas como duplicadas en las transacciones realizadas en el año 2015.

Al realizar una buscada en específico de estas dos referencias con ayuda de Excel, se tienen operaciones del mismo tipo, monto, operador, sucursal, en diferente fecha asociadas a una misma referencia. Por la anomalía de esta situación se consideran como transferencias atípicas, ***se recomienda revisar una investigación mayor considerando esta información y determinar si se trata de un error de sistema o algo mal intencionado***.

00MH988235	2257.86	MIGUEL HIDALGO	4/30/2015	02:56:08 a. m.	DEFE	MH5506	3004637
00MH988235	2257.86	MIGUEL HIDALGO	8/24/2015	04:08:42 a. m.	DEFE	MH5506	3004637

00BJ962502	10533.54	BENITO JUÁREZ	05/05/2015	02:56:08 a. m.	DCHQ	BJ0109	3064984
00BJ962502	10533.54	BENITO JUÁREZ	05/05/2015	01:49:04 a. m.	DCHQ	BJ0109	3064984

2.- Se tienen valores nulos

Mediante el uso de pandas de identificaron valores nulos en los campos de Fecha_Transaccion y Hora_transaccion. Se válido y se trata de los mismos registros (6).

```
Numero_de_cuenta     0
M_transaccion        0
Sucursal_nombre      0
Fecha_Transaccion    6
Hora_transaccion     6
T_transaccion        0
Operador             0
Referencia           0
dtype: int64
```

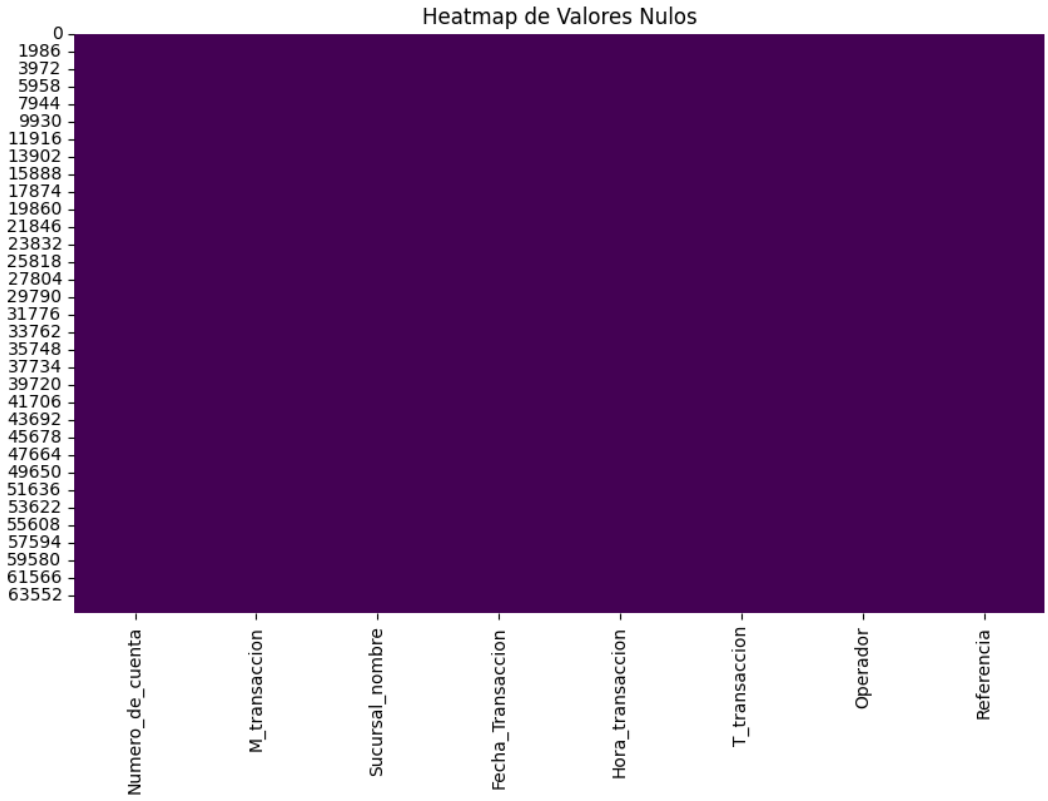

Figura 10. Heatmap de Valores Nulos.

	Numero_de_cuenta	M_transaccion	Sucursal_nombre	Fecha_Transaccion	Hora_transaccion	T_transaccion	Operador	Referencia
14786	00IP989289	526.24	IZTAPALAPA	NaN	NaN	DEFE	IP1047	3002889
22533	00CU980770	1200.00	CUAUHTEMOC	NaN	NaN	DEFE	CU5056	3001903
27452	00CU980883	1861.04	CUAUHTEMOC	NaN	NaN	DEFE	CU5032	3001905
32341	00CU984227	2757.48	CUAUHTEMOC	NaN	NaN	DEFE	CU5056	3001871
40858	00CU980883	5582.80	CUAUHTEMOC	NaN	NaN	DEFE	CU5029	3001904
49404	00CU962219	13583.70	CUAUHTEMOC	NaN	NaN	DCHQ	CU5052	3001902

Figura 11. Subconjunto de valores nulos del conjunto de datos de registros de las transacciones realizadas en el año 2015.

Nota: Con motivos de revisar si estos valores nulos tienen mayor contexto en el conjunto de datos, se validó si los clientes asociados a estas transacciones están dentro del subconjunto de clientes con 3 cuentas aperturadas. Obteniendo como resultados que, en este caso, los nulos no están asicados dichos clientes.

3.- Desbalance de observaciones considerando las variables categóricas.

De acuerdo con la gráfica de la figura 8 con la distribución de las observaciones, se puede visualizar que se tiene mayor cantidad de transacciones desde la sucursal del Miguel Hidalgo.

```
#Cálculo de porcentaje de observaciones por sucursal
porcentajes = df['Sucursal_nombre'].value_counts(normalize=True) * 100
print("Cálculo de porcentaje de observaciones por sucursal")
print(porcentajes)
```

```
Cálculo de porcentaje de observaciones por sucursal
Sucursal_nombre
MIGUEL HIDALGO    41.167943
CUAUHTEMOC        19.029817
COYOACÁN          17.340617
FORÁNEAS          16.440321
BENITO JUÁREZ      3.318888
IZTAPALAPA         2.507096
IZTACALCO          0.195318
```

Además, también considerando el conteo de transacciones realizadas en 2015 por tipo de transacción, se tiene un desbalance de observaciones, dónde la clase mayoritaria está asociada con las transacciones mediante depósitos en efectivo.

```python
#Cálculo de porcentaje de observaciones por Tipo de transacción
porcentajes = df['T_transaccion'].value_counts(normalize=True) * 100
print("Cálculo de porcentaje de observaciones por Tipo de transacción")
print(porcentajes)
```

```
Cálculo de porcentaje de observaciones por Tipo de transacción
T_transaccion
DEFE    65.248573
DCHQ    24.712363
RECJ     5.044710
REVN     4.994354
Name: proportion, dtype: float64
```

2.1.3 Relación de los dos conjuntos de datos

En este apartado se pretende tomar ventajas de la disponibilidad de la información proporcionada por el conjunto de datos de Aperturas, en base a las transacciones realizadas identificar si se tiene algún movimiento realizado desde una cuenta que no se encuentra en los movimientos de aperturas realizadas desde el año 2013 a 2015.

Mediante el uso de un outer join con pandas, considerando como llave "Numero_de_cuenta", se identificaron 10 transacciones, en su mayoría mediante Deposito en efectivo (DEFE), asociadas a cuentas no registradas con base a la información disponible. Considerando que se tiene conocimiento que la institución financiera tiene 5 años de presencia en el sector, es importante validar antes de confirmar que estos registros son anómalos, si las cuentas involucradas fueron aperturadas desde otro periodo de tiempo válido. Sin embargo, por el hecho de que durante 1 año sólo se tiene un movimiento para cada cuenta, aumenta la incertidumbre.

	Numero_de_cuenta	M_transaccion	Sucursal_nombre	Fecha_Transaccion	Hora_transaccion	T_transaccion	Operador	Referencia	Cuenta
31276	00MH960449	7.00	MIGUEL HIDALGO	01/01/2015	10:55:04 a. m.	DEFE	MH5555	3025643	left_only
42131	00BJ968504	47.26	BENITO JUÁREZ	5/28/2015	09:23:19 a. m.	DEFE	BJ0109	3036916	left_only
51965	00BJ968421	228.36	BENITO JUÁREZ	11/09/2015	01:59:28 a. m.	DEFE	BJ0107	3026347	left_only
54571	00MH960448	365.97	MIGUEL HIDALGO	07/02/2015	09:22:55 a. m.	DEFE	MH5554	3028415	left_only
55781	00BJ968458	481.19	BENITO JUÁREZ	11/11/2015	09:22:09 a. m.	DEFE	BJ0116	3026342	left_only
58627	00MH960452	1106.64	MIGUEL HIDALGO	8/24/2015	09:24:54 a. m.	DEFE	MH5558	3011308	left_only
60996	00MH960447	2618.50	MIGUEL HIDALGO	01/09/2015	05:52:09 a. m.	DEFE	MH5552	3042337	left_only
61345	00FO964179	2902.84	FORÁNEAS	4/18/2015	09:23:19 a. m.	DEFE	FO7558	3045816	left_only
61522	00BJ968506	3081.75	BENITO JUÁREZ	03/02/2015	01:53:02 a. m.	DEFE	BJ0116	3007083	left_only
65107	00BJ968505	88466.28	BENITO JUÁREZ	05/01/2015	03:22:32 a. m.	REVN	BJ0116	3044901	left_only

Figura 12. Transacciones realizadas desde cuentas no registradas en movimientos de aperturas realizadas desde el año 2013 a 2015.

3 Implementación de estrategias para la identificación de operaciones atípicas

Una vez realizado el análisis exploratorio de ambas fuentes de datos disponibles, se propone una serie de estrategias de análisis y técnicas de machine learning para la identificación de datos atípicos considerando las tracciones realizas en el 2015. Partiendo de lo anterior, las técnicas valoradas han sido seleccionadas a partir de la necesidad de trabajar de manera no supervisada por la naturalidad de los datos.

3.1.1 Conjunto de datos auxiliar

Mediante el uso de un left join tomando como llave el número de cuenta, se creó un conjunto de datos auxiliar para la implementación de las técnicas univariables y multivariables que se muestran más adelante. En este conjunto de datos, se decidió conservar la variable de tipo de transacción, debido a que ayuda a identificar outliers. Además, también se creó la variable meanMontoTransferecia la cual concentra el promedio de todas las transacciones asociadas a cada cuenta.

	Numero_de_cuenta	M_transaccion	T_transaccion	Monto_de_apertura	meanMontoTransferencia
0	00CU962253	0.000000e+00	1	1.449880e+05	1.812351e+04
1	00FO960212	0.000000e+00	1	6.125068e+05	9.007453e+03
2	00MH961270	0.000000e+00	1	2.767345e+05	7.756204e+03
3	00FO989961	0.000000e+00	1	-7.558069e+05	2.148114e+04
4	00MH960322	0.000000e+00	1	2.419651e+05	8.819989e+05
...	...	...	...	...	...
65529	00FO988163	9.717436e+08	0	9.828636e+08	1.966066e+08
65530	00FO960033	9.723555e+08	0	9.882445e+08	6.588820e+07
65531	00CU964678	9.723555e+08	0	1.004064e+09	2.010504e+06

Figura 13. Conjunto de datos auxiliar para el apoyo de la implementación de los métodos univariados y multivariados.

3.1.2 Análisis Univariado

3.1.2.1 Diagrama de cajas

En esta sección se pretende a partir de los gráficos de Boxplot identificar que variables toman mayor relevancia para localizar los outliers o datos atípicos.

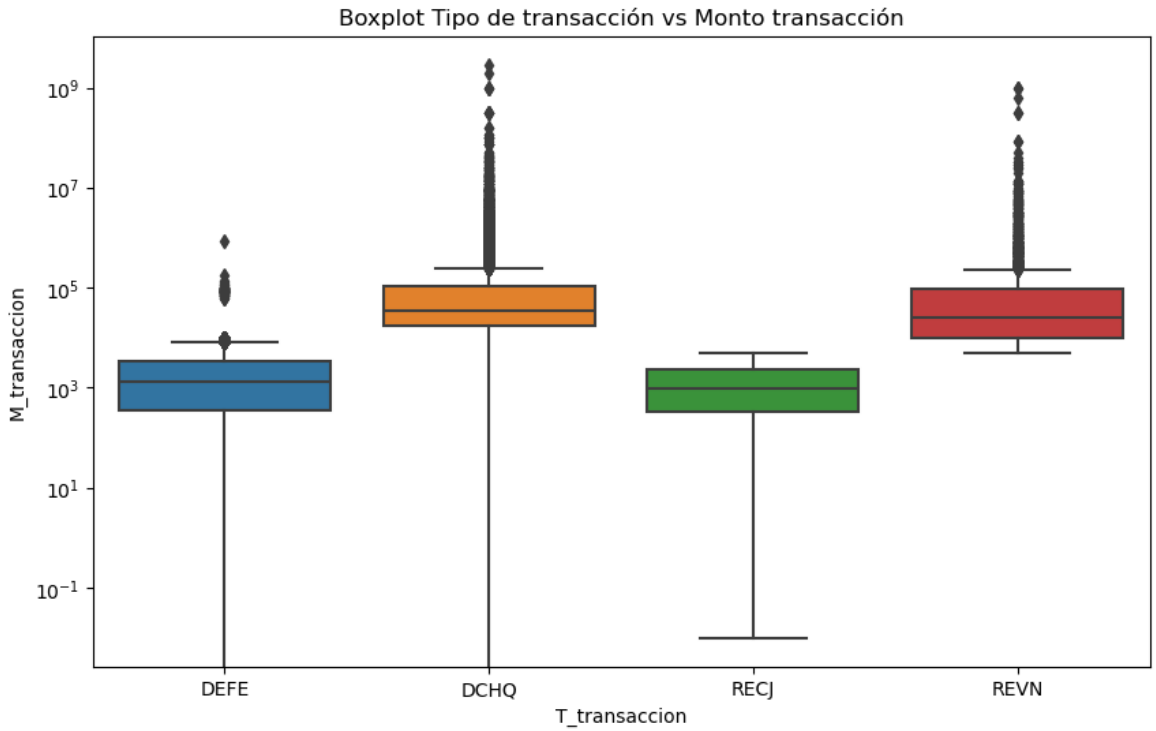

Figura 14. Boxplot Tipo de transacción vs Monto transacción

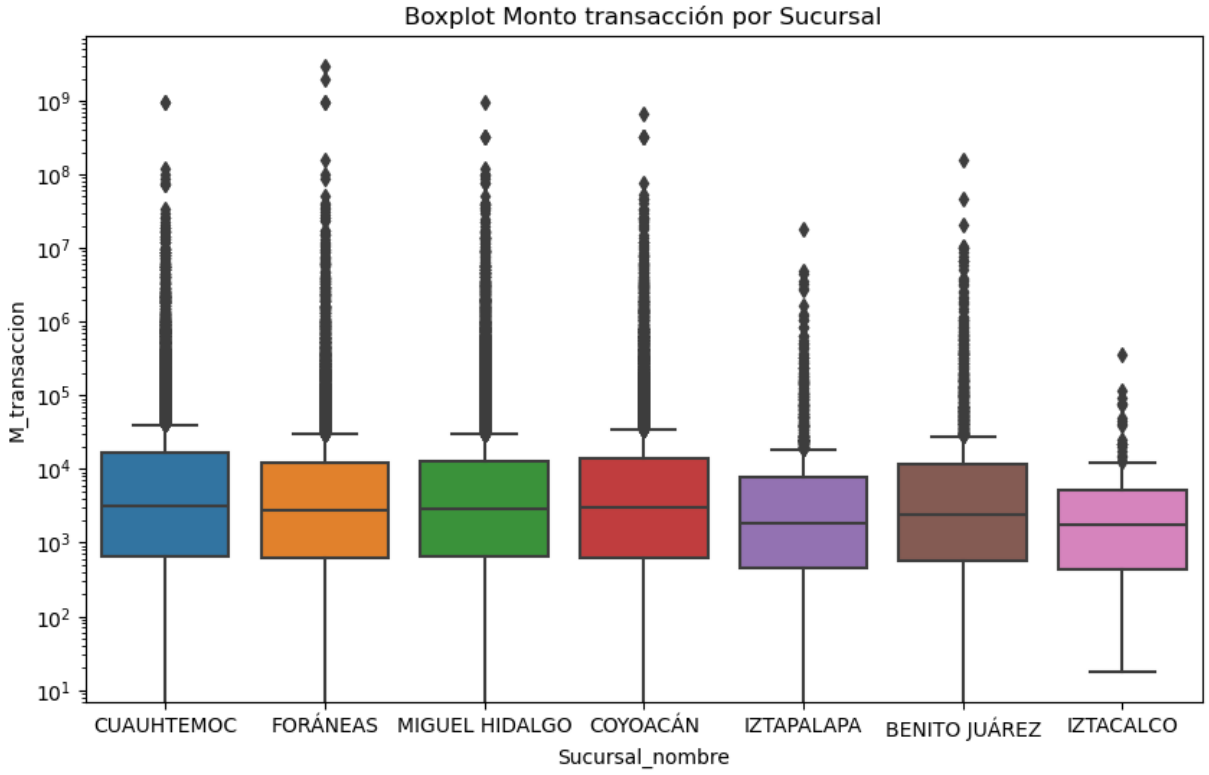

Figura 15. Boxplot Monto transacción por Sucursal

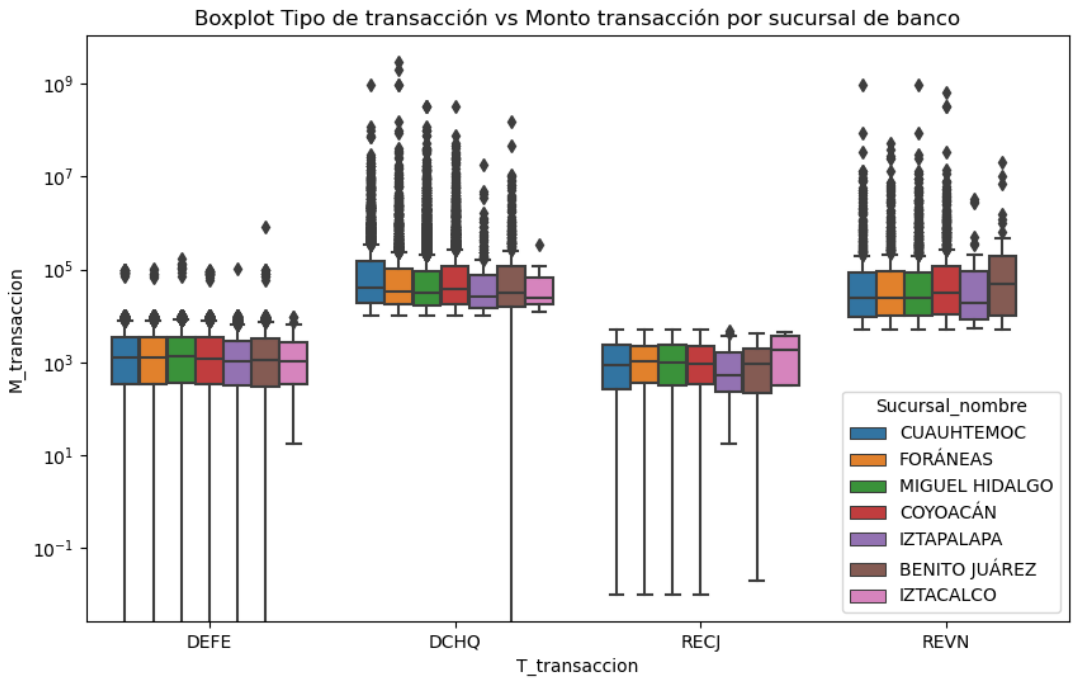

Figura 16. Boxplot Tipo de transacción vs Monto transacción por sucursal de banco

Observaciones:

1.- La distribución de los datos de tipo de transacción con respecto del monto de transacción, ayuda visualmente a identificar los ourliers mediante el diagrama de cajas. Por lo que, se considera una variable relevante que se puede utilizar para determinar si una transacción es o no fraudulenta.

2.- En la figura 14 se puede observar que las transacciones realizadas mediante retiro de efectivo en cajero automático no presentan caso de outliers. Seguido de los depósitos en efectivo. Esto tiene sentido, debido a que generalmente los cajeros automáticos controlar la cantidad máxima de dinero que se puede retirar.

3.- Las transacciones de Retiro de efectivo en ventanillas y depósitos en cheques, son de acuerdo con la figura 14, son las transacciones con más casos de outliers.

3.1.2.2 IQR

El IQR es una medida estadística que se puede emplear para identificar outliers. Los valores que están más allá de 1.5 veces el IQR por debajo del primer cuartil (Q1) o por encima del tercer cuartil (Q3) se consideran outliers.

Mediante el cálculo del IQR considerando el Monto de transacción del conjunto de datos auxiliar y el cálculo de los umbrales inferior y superiores, se determinó que observaciones (transacciones) según IQR eran ourliers. Obtenidos la relación de observaciones con ourliers de la figura 17.

Proceso para el cálculo del IQR y los umbrales inferior y superior

```
# Calcular Q1 (primer cuartil) y Q3 (tercer cuartil)
df_iqr = df_final
Q1 = df_iqr['M_transaccion'].quantile(0.25)
Q3 = df_iqr['M_transaccion'].quantile(0.75)
IQR = Q3 - Q1

# Definir los límites para outliers
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Identificar outliers
df_iqr['is_outlier'] = (df_iqr['M_transaccion'] < lower_bound) | (df_iqr['M_transaccion'] > upper_bound)
df_iqr
```

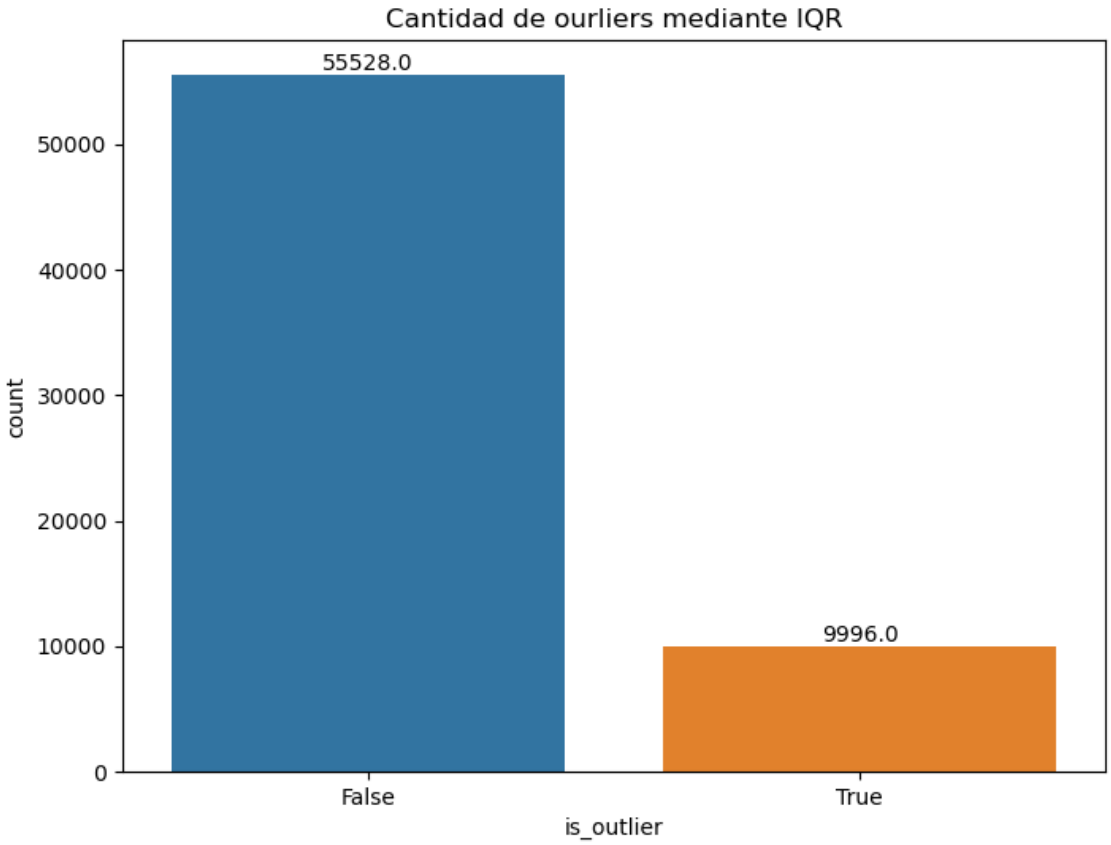

Figura 16. Conteo de outliers determinados a partir de IQR para todas las observaciones del conjunto de datos auxiliar.

3.1.3 Análisis Multivariado

3.1.3.1 IsolationForest

El objetivo de esta sección es poder determinar en base a más de una variable que transacciones pueden ser fraudulentas. Para eso se hizo uso del algoritmo de IsolationForest mediante el apoyo de la librería de scikit-learn.

Para poder hacer uso de este algoritmo para la creación de un modelo y predecir los datos atípicos se utilizaron las siguientes variables.

T_transaccion	Monto_de_apertura	meanMontoTransferencia	is_outlier	IsolationForest
1	0.00	0.00	0	1
1	0.00	0.00	0	1
1	0.00	0.00	0	1
1	0.00	0.00	0	1
1	0.00	0.00	0	1
...	...	...	...	...
0	5160689.49	5160689.49	1	-1
0	5638748.96	5638748.96	1	-1
0	8041205.96	8041205.96	1	-1
3	-10247956.58	10247956.58	1	-1
0	18581218.85	18581218.85	1	-1

Figura 17. Dataframe empleado en el entrenamiento del modelo con IsolationForest.

Como se puede observar en la figura anterior, las variables categóricas han sido decodificadas mediante LabelEncoder() de scikit-learn. Además, es importante señalar que para esta técnica se optó por tomar el promedio de las trasferencias realizadas en el 2015 por cuenta. De esta manera de reduce el número de observaciones, lo cual permitió en este ejercicio explorar técnicas más complejas computacionalmente.

```python
#Tratamiento de l variable categoríca T_transaccion
le = LabelEncoder()
df_final['T_transaccion'] = le.fit_transform(df_final['T_transaccion'])
# Mostrar la asignación de categorías a valores numéricos
print("Asignación de categorías a valores numéricos:")
for categoria, codigo in zip(le.classes_, range(len(le.classes_))):
    print(f"{categoria}: {codigo}")
```

```
Asignación de categorías a valores numéricos:
DCHQ: 0
DEFE: 1
RECJ: 2
REVN: 3
```

Finamente, se obtuvieron las siguientes cantidades de observaciones asociados a outliers y no ourliers (inliers).

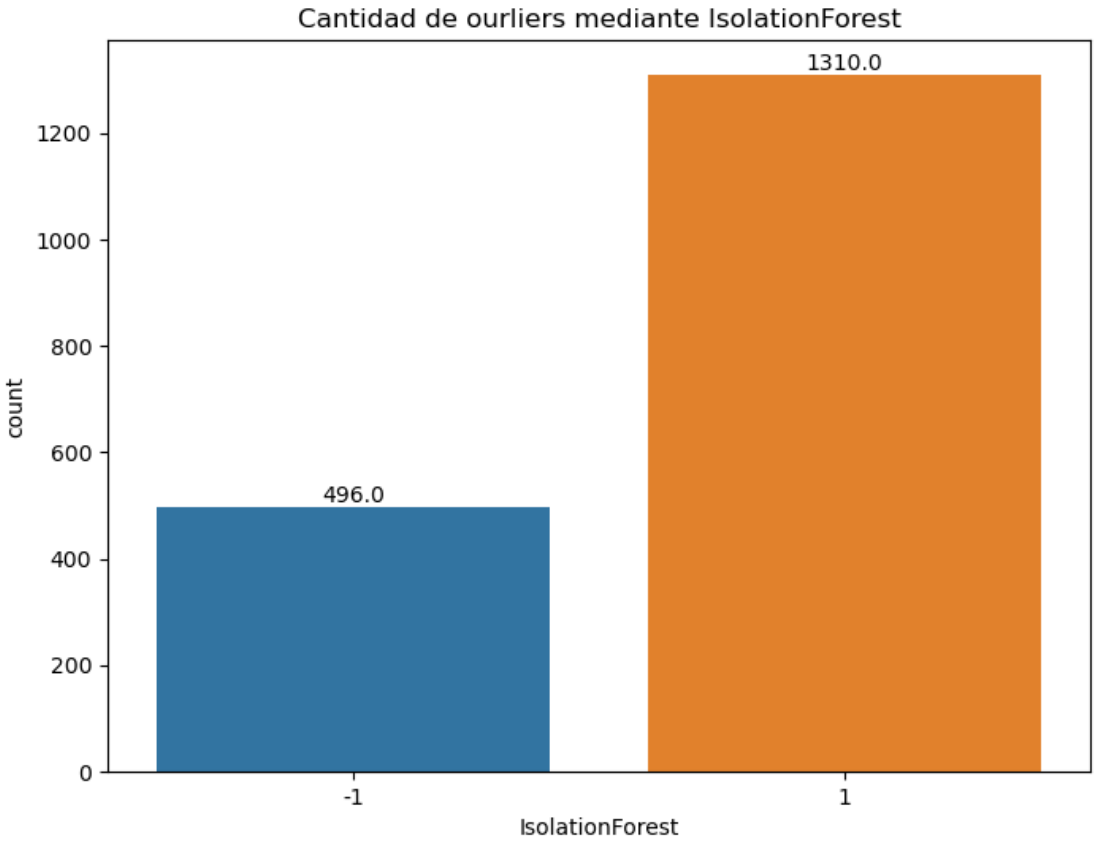

Figura 18. Conteo de cuentas asociadas a comportamientos no comunes (outliers) mediante IsolationForest.

4 Conclusiones

La identificación de casos o situaciones atípicas puede ser empleado en distintos casos de uso. En específico para identificación de fraudes o desviaciones en procesos de negocio, es de gran utilidad encontrar la estrategia que permita en la mayor cantidad identificar estos procesos. El análisis exploratorio permite comprender la estructura y composición de los datos, identificar comportamientos, que variables agregan mayor valor para localizar los casos atípicos, mediante estos supuestos se puede tomar ventaja e implementar métodos basados en Machine Learning que automaticen estos procesos, cómo lo es el uso de IsolationForest. Además, mediante las observaciones presentadas en la exploración de datos, como el uso de cuentas no registradas para efectuar transacciones, se pueden generar observaciones de áreas de oportunidad que robustezcan la seguridad de los procesos dentro de la organización,

Para este problema, nos enfrentamos a una situación de un conjunto de datos muy segmentado, es decir, de los 5 años de presencia de la financiera en el sector, sólo se está disponiendo de 1 años de transacciones efectuadas. Sin embargo, aumentar también el número de características introduciría mayores consumos de recursos y aumentar la dificultad del análisis y visualización de datos, sin embargo podría enriquecer más los procesos automáticos basados en ML.

Las estrategias valoradas en este trabajo se pueden mejorar considerando otros factores, por ejemplo, para el análisis de datos, hacer uso de las fechas de las transacciones e identificar en que época del año se presentan más las operaciones fraudulentas o casos atípicos. También para el proceso de automatización mediante machine learning, se pueden explorar más algoritmos para realizar puntos de comparación y generar una propuesta más robusta de acuerdo con la ergonomía de los datos.