ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

Mantenimiento preventivo de cajeros automáticos basado en el análisis de detección de anomalías en los registros históricos de inactividad por error de los módulos críticos aplicando técnicas de ciencia de datos y aprendizaje no supervisado

PROYECTO DE TITULACIÓN

Previo la obtención del Título de:

Magister en Ciencias de Datos

Presentado por:

Christian Alejandro Jaramillo Espinoza

GUAYAQUIL - ECUADOR

Año: 2023

DEDICATORIA

A las personas más importantes en mi vida, por su comprensión y apoyo constante e incondicional a lo largo de este proyecto.

Ana y Flor

AGRADECIMIENTOS

Mi agradecimiento a todas las personas que hicieron posible la realización de este trabajo:

A mis profesores que impartieron sus conocimientos sobre ciencia de datos para lograr este objetivo.

A mi tutora por su guía y consejos durante todo este proceso.

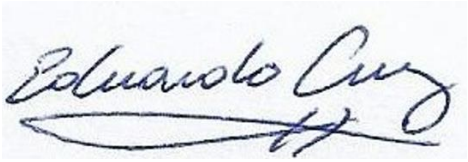
A la empresa y compañeros donde laboro por su apertura en la resolución de dudas y procesos del negocio.

DECLARACIÓN EXPRESA

“Los derechos de titularidad y explotación, me corresponde conforme al reglamento de propiedad intelectual de la institución; Yo, Christian Alejandro Jaramillo Espinoza doy mi consentimiento para que la ESPOL realice la comunicación pública de la obra por cualquier medio con el fin de promover la consulta, difusión y uso público de la producción intelectual"

Christian Alejandro Jaramillo Espinoza

COMITÉ EVALUADOR



Msc. Karen Calva Yaguana

PROFESOR TUTOR

Mg. Eduardo Cruz Ramírez

PROFESOR EVALUADOR

RESUMEN

Los tiempos de inactividad de los cajeros automáticos pueden tener consecuencias significativas en la experiencia del usuario y en la operatividad general de los servicios financieros. Al ser un medio fundamental para la realización de transacciones bancarias dentro o fuera del horario de atención bancaria tradicional, se han convertido en una parte vital de la infraestructura financiera. Los tiempos de inactividad por fallas técnicas pueden impactar de diversas formas como la satisfacción del usuario, pérdidas financieras, costes operativos.

En el presente proyecto se propone el uso de aprendizaje de máquina no supervisado mediante el algoritmo Isolation Forest para lograr detectar a tiempo aquellos cajeros automáticos que puedan convertirse en un problema mayor y, por ende, realizar mantenimiento preventivo a dichos cajeros anómalos para que los tiempos de inactividad y las llamadas de soporte al centro de atención al cliente de la empresa proveedora de los cajeros disminuyan.

Además, se provee al cliente un módulo adicional sobre detección de anomalías dentro de la aplicación de monitoreo que ya poseen, en el cuál, pueden visualizar de manera tabular y gráfica aquellos cajeros previamente detectados por el modelo. Con esta información proporcionada, el cliente deberá analizarla y tomar la decisión pertinente de realizar mantenimiento preventivo.

Como resultado de la evaluación del modelo y el mantenimiento preventivo realizado, el cliente analizado mediante la solución propuesta se ahorraría $42.840 anual a mediano plazo, un 62.63% menos de su gasto anual presupuestado en mantenimientos correctivos en todos sus cajeros. En cuanto a las llamadas de soporte se reduciría en 272 un 15.2% menos para este cliente que en las condiciones actuales.

Palabras Clave: cajero automático, tiempos de inactividad, isolation forest, detección de anomalías

ABSTRACT

ATM downtime can have a significant impact on the user experience and the overall operability of financial services. As a fundamental means of conducting banking transactions within or outside traditional banking hours, they have become a vital part of the financial infrastructure. Downtime due to technical failures can impact in many ways such as user satisfaction, financial losses, operational costs.

This project proposes the use of unsupervised machine learning using the Isolation Forest algorithm to achieve early detection of those ATMs that may become a major problem and, therefore, perform preventive maintenance to these anomalous ATMs so that downtime and support calls to the customer service center of the ATM provider company decrease.

The customer is also provided with an additional module on anomaly detection within the monitoring application they already have, in which they can visualize in tabular and graphical form those ATMs previously detected by the model. With this information provided, the client must analyze it and make the pertinent decision to perform preventive maintenance.

As a result of the evaluation of the model and the preventive maintenance performed, the client analyzed through the proposed solution would save $42,840 annually in the medium term, 62.63% less than its budgeted annual expenditure on corrective maintenance in all its ATMs. Support calls would be reduced by 272, 15.2% less for this client than under current conditions.

***Keywords**: ATM, downtime, isolation forest, anomaly detection*

ÍNDICE GENERAL

ABREVIATURAS

CAC	Centro de atención al cliente
PIN	Número de identificación personal
EMV	Europay, Mastercard y Visa
WCSS	Suma de los cuadrados dentro del clúster
CH	Índice Calinski-Harabasz
DBI	Índice Davies-Bouldin
IDE	Entorno de desarrollo integrado
CSV	Valores Separados por Comas
t-SNE	Incrustación estocástica de vecinos en distribución t
uMAP	Aproximación y proyección de colectores uniformes

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO 1

1. PLANTEAMIENTO DE LA PROBLEMÁTICA

Un cajero automático es un dispositivo computacional que está diseñado para realizar las funciones más importantes de un banco o institución financiera, mediante el uso de tarjetas que contienen los datos personales del cliente e ingreso de su contraseña hacen uso de los servicios que presta [1]. Ofrece la posibilidad de llevar a cabo una variedad de transacciones financieras en las que incluye retiros, depósitos, transferencias de fondos y la consulta de información general sobre su saldo. [2]

Los cajeros automáticos pueden considerarse uno de los servicios más importantes del sector bancario. La inversión en cajeros y su impacto en el sector bancario crece constantemente en todo el mundo. Los propietarios se enfrentan cada vez más al reto de mejorar la calidad del servicio al cliente y, al mismo tiempo, aumentar la fiabilidad y seguridad del equipo, mientras que los fabricantes y proveedores están comprometidos a garantizar la calidad e integridad de los equipos para mantener la confianza de sus clientes y compradores.

1.1 Descripción del problema

Los cajeros automáticos son altamente susceptibles a errores e indisponibilidad al ser un dispositivo demasiado complejo con múltiples módulos interconectados, para restaurar el servicio de estos cajeros la institución financiera realiza una llamada al CAC para realizar soporte de mantenimiento correctivo de manera presencial.

El Centro de Atención al Cliente (CAC) de la empresa “ABC” brinda a sus clientes atención oportuna ante una llamada o reporte de falla por medio de la presencia de un ingeniero de campo en el sitio de ubicación del cajero para brindar el soporte necesario, hoy en día se ha desencadenado una situación crítica debido a la agobiante cantidad de llamadas de atención de clientes sobre cajeros fuera de servicio debido a errores y problemas técnicos (125 llamadas promedio por día de acuerdo a un experto de la empresa). Este hecho ha sobrepasado significativamente la capacidad de respuesta del personal técnico existente (55 en todo el país) y ha generado una necesidad urgente de atención adicional, dado que a un técnico de

campo se le asigna no más de dos llamadas por día dada la cantidad de horas que conlleva realizar mantenimiento completo a un cajero (3 horas promedio por llamada).

La indisponibilidad del cajero automático por falta de mantenimiento es un problema significativo que afecta a propietarios y usuarios. Los cajeros son equipos altamente complejos y sofisticados que operan continuamente durante todo el día para proporcionar servicios financieros, entregar y recibir dinero, lo que lo hace muy crucial en términos de beneficios y satisfacción hacia el cliente, es por esto por lo que los bancos han empezado a prestarle más atención a este problema. [3]

Debido a la frecuencia y el volumen de transacciones que realizan, los cajeros están expuestos a una gran cantidad de desgaste y problemas técnicos, por tanto, requieren mantenimiento regular para garantizar su correcto funcionamiento y evitar malos ratos a los usuarios y penalidades por instituciones reguladoras.

1.2 Justificación del problema

Los cajeros son dispositivos críticos, y su mantenimiento regular es esencial para garantizar un funcionamiento confiable y seguro. Sin embargo, el mantenimiento basado en el tiempo, donde el mantenimiento se realiza en intervalos fijos y preestablecidos o peor aun cuando el cajero ya se encuentra inutilizable, puede ser costoso y puede no ser eficiente en términos de tiempo y recursos.

De acuerdo con BANRED (La red interamericana de cajeros automáticos) en Ecuador se distribuyen alrededor de cinco mil ochocientos cajeros entre bancos, cooperativas de ahorro y crédito, mutualistas y otras instituciones financieras [4]. Si un cajero no recibe el mantenimiento adecuado, es posible que experimente una variedad de problemas, como atascos de billetes o tarjetas, fallas lógicas o físicas de dispositivos, problemas de software, y otros problemas técnicos que pueden causar una interrupción en el servicio. Esto puede llevar a que no esté disponible por un tiempo incalculable, lo que puede ser extremadamente desfavorable para los usuarios y causar una pérdida de ingresos para la institución financiera dueña del equipo.

Cada operación de los dispositivos de un cajero como la lectora de tarjetas, dispensador de dinero, impresora y demás es almacenada en registros. Estos

registros contienen información descriptiva del equipo y tiempos de actividad e inactividad que pueden ayudar a construir un modelo preventivo adecuado mediante el análisis de las series de tiempos que provee.

El mantenimiento preventivo debido a la detección temprana de anomalías es una alternativa prometedora por su capacidad de identificar automáticamente cuando debe realizarse [5]. Por lo que la finalidad es programar el soporte en sitio en el momento en el que el mantenimiento sea más rentable y antes de que el equipo pierda rendimiento y funcionalidad.

El principal objetivo del mantenimiento preventivo es aumentar la disponibilidad de las máquinas reduciendo al mínimo el mantenimiento no planificado causado por averías. Saber cuándo y por qué va a fallar una máquina ofrece muchas ventajas, como una mejor planificación del mantenimiento, la gestión de piezas en stock y otras optimizaciones de costes relacionadas con el proceso de mantenimiento.

El mantenimiento preventivo por detección de anomalías utiliza técnicas avanzadas de ciencia de datos y aprendizaje de máquina para identificar cuáles serían los cajeros problemáticos y permitiría a los coordinadores técnicos de mantenimiento tomar medidas y estrategias necesarias antes de que la falla sea de mayores proporciones.

Ante lo mencionado, el Centro de atención al Cliente (CAC) puede recomendar a sus clientes por la aplicación de monitoreo realizar mantenimiento a sus cajeros en el momento oportuno con ayuda de una solución de detección de anomalías realizada mediante el análisis de los registros históricos diarios de fallos en los módulos críticos del cajero.

Con esto se intenta disminuir progresivamente a largo plazo la tasa de reportes o llamadas de fallos al CAC y aliviar la carga humana y operacional, disminuir los fallos de los cajeros, aumentar la confianza, reputación, fidelización y satisfacción del cliente, mejorar la imagen de la empresa a la interna ante el uso de nuevas tecnologías de inteligencia artificial y a la externa brindando un impacto positivo en el mercado y diferenciándola de sus competidores.

1.3 Objetivos

1.3.1 Objetivo General

Desarrollar el prototipo de un módulo de la aplicación de monitoreo de cajeros automáticos usando técnicas de ciencia de datos y aprendizaje no supervisado para incrementar la eficiencia y fiabilidad del cajero automático, incrementar la satisfacción del usuario y dueño del equipo, reducir gastos y recursos del proveedor del servicio y mejorar la imagen de la empresa.

1.3.2 Objetivos Específicos

Analizar los registros históricos diarios de inactividad por error de los dispositivos críticos de los cajeros automáticos.

Implementar técnicas de ciencias de datos y aprendizaje no supervisado que identifique patrones o valores anómalos que detecten los cajeros automáticos potencialmente problemáticos.

Desarrollar el prototipo del módulo de detección de anomalías que indique los cajeros automáticos identificados para realizar mantenimiento preventivo.

1.4 Metodología

Para cumplir con los objetivos deseados se aplicará el análisis de anomalías de los tiempos de inactividad de los módulos críticos del cajero automático con ayuda de las técnicas más usadas en este campo entre los que se encuentran:

- K-means clustering (Agrupamiento k-medias).
- Isolation Forest (Bosques de aislamiento).

Para lograr el objetivo se realiza el siguiente procedimiento:

1. **Recopilación y limpieza de datos**

   Limpiar e imputar los datos para tener un conjunto de datos acorde que pueda ser usado en los modelos de aprendizaje de máquina para obtener una detección coherente.

2. **Preprocesamiento de datos**

   Convertir los datos a otros formatos requeridos como, por ejemplo: la fecha o el tiempo para poder ser procesados por los modelos. Enmascarar la información sensible por seguridad.

3. **Normalización de datos**

   Normalizar los datos para que funcionen de manera más efectiva y produzcan resultados más precisos.

4. **Evaluar el algoritmo K-means**

   Evaluar para agrupar los datos en clústeres basados en sus características, determinar el número óptimo de clústeres utilizando métodos como el codo e identificar los clústeres que contienen los cajeros automáticos normales y aquellos que podrían estar propensos a errores.

5. **Evaluar el algoritmo Isolation Forest**

   Evaluar para detectar anomalías dentro de los clústeres identificados en el algoritmo K-means, ajustar los hiperparámetros según sea necesario para lograr un equilibrio entre la detección de anomalías y la minimización de falsos positivos.

6. **Analizar y evaluar resultados**

   Con los resultados de K-means e Isolation Forest analizar y evaluar los grupos detectados y la puntuación de anomalías para presentar los resultados.

1.5 Resultados esperados

Detectar los cajeros automáticos anómalos mediante uso de modelos de aprendizaje no supervisado que indique que un cajero ha aumentado sus tiempos de inactividad por error y esté propenso a salir de servicio.

Que el prototipo del módulo de detección de anomalías presente de manera adecuada e intuitiva al cliente los cajeros detectados para que tome las decisiones que crea pertinentes para realizar mantenimiento preventivo.

Mediante los mantenimientos preventivos realizados a aquellos cajeros detectados como anómalos lograr que la tasa de llamadas al CAC descienda a mediano o largo plazo.

Entregar un valor agregado al cliente mediante el módulo de detección que incremente la confianza y satisfacción del cliente.

Generar un impacto positivo y consolidación del aprendizaje de máquina e inteligencia artificial en los procesos de la empresa.

1.6 Conjunto de datos

El conjunto de datos comprende la historia de tiempos de inactividad por error de 2.526 cajeros automáticos de 13 instituciones financieras de 5 países durante 12 semanas entre junio y agosto del año 2023.

En cada observación del conjunto de datos se define la cantidad en segundos de inactividad por día de seis módulos críticos de un cajero automático e incluye varias descripciones categóricas de cada uno de ellos.

1.6.1 Variables

En total el conjunto de datos contiene 212.184 registros donde se encuentra las siguientes 14 características:

- **CUSTOMER**
  Cliente dueño del cajero automático (Tipo categórica).

- **ID**
  Identificación del cajero automático (Tipo categórica).

- **FAMILY**
  Familia o generación del cajero automático (Tipo categórica).

- **MODEL**
  Modelo dentro de la familia del cajero automático (Tipo categórica).

- **FUNCTION**

Funcionalidad principal del cajero automático (Tipo categórica).

- **SITE**
  Tipo de cajero automático de acuerdo con su ubicación (Tipo categórica).

- **COUNTRY**
  País donde el cajero automático está ubicado (Tipo categórica).

- **DATETIME**
  Indica el día en que la observación fue registrada (Tipo fecha).

- **CARD_DOWNTIME**
  Cantidad de segundos de inactividad por error reportados de la lectora de tarjetas en el día (Tipo numérico).

- **CASH_DOWNTIME**
  Cantidad de segundos de inactividad por error reportados del dispensador de efectivo en el día (Tipo numérico).

- **ACCEPTOR_DOWNTIME**
  Cantidad de segundos de inactividad por error reportados del aceptador de efectivo en el día (Tipo numérico).

- **DEPOSITOR_DOWNTIME**
  Cantidad de segundos de inactividad por error reportados del aceptador de cheques en el día (Tipo numérico).

- **EPP_DOWNTIME**
  Cantidad de segundos de inactividad por error reportados del teclado electrónico en el día (Tipo numérico).

- **PRINTER_DOWNTIME**
  Cantidad de segundos de inactividad por error reportados de la impresora en el día (Tipo numérico).

	CUSTOMER	ID	MODEL	FUNCTION	SITE	FAMILY	COUNTRY	DATETIME
1443	Customer USA 1	1119HF	ADA 047	Recycler	Branch	Gen 2000	United States	2023-08-31
1444	Customer USA 3	581009IH	SND 045 A	Recycler	Branch	Gen 2020	United States	2023-07-08
1445	Customer USA 4	932929MTA	XGN 0877	Recycler	Drive Up	Gen 2010	United States	2023-08-03
1446	Customer USA 5	323000CS	XGN 0977	Recycler	Branch	Gen 2010	United States	2023-06-20
1447	Customer ECU 3	3209091C	ADA 225	Cash Dispenser	Branch	Gen 2000	Colombia	2023-08-11
1448	Customer USA 1	3090HF	ADA 057	Recycler	Branch	Gen 2000	United States	2023-08-07
1449	Customer USA 7	42004801BSVM	SND 094	Recycler	Branch	Gen 2020	United States	2023-07-10
1450	Customer USA 6	3270AJ	SND 074	Recycler	Branch	Gen 2020	United States	2023-08-04
1451	Customer USA 1	9529HF	ADA 067	Recycler	Branch	Gen 2000	United States	2023-07-14
1452	Customer ECU 1	8382404EYGTA	SND 045 A	Recycler	Lobby	Gen 2020	Ecuador	2023-06-28
1453	Customer USA 1	8450HF	ADA 047	Recycler	Branch	Gen 2000	United States	2023-07-21
1454	Customer ECU 3	4640091C	ADA 225	Cash Dispenser	Lobby	Gen 2000	Colombia	2023-07-05
1455	Customer USA 1	3050HF	ADA 027 A	Recycler	Branch	Gen 2000	United States	2023-07-21
1456	Customer USA 1	1219HF	SND 094	Recycler	Branch	Gen 2020	United States	2023-08-30
1457	Customer USA 3	960009IH	SND 074	Recycler	Branch	Gen 2020	United States	2023-07-15
1458	Customer USA 4	423929MTA	XGN 0877	Recycler	Drive Up	Gen 2010	United States	2023-08-17
1459	Customer COL 1	1150MTA	ADA 225	Cash Dispenser	Branch	Gen 2000	Colombia	2023-06-17
1460	Customer USA 1	7119HF	SND 094	Recycler	Branch	Gen 2020	United States	2023-07-28
1461	Customer ECU 1	1682404EYGTA	SND 045 A	Recycler	Lobby	Gen 2020	Ecuador	2023-07-27
1462	Customer USA 2	014000CN	ADA 047	Recycler	Branch	Gen 2000	United States	2023-07-18
1463	Customer ECU 3	7700091C	ADA 225	Cash Dispenser	Lobby	Gen 2000	Colombia	2023-08-29
1464	Customer ECU 3	6019091C	ADA 225	Cash Dispenser	Branch	Gen 2000	Colombia	2023-09-01
1465	Customer ECU 1	4922104OIUTA	XGN 0555	Cash Dispenser	Lobby	Gen 2010	Ecuador	2023-08-05
1466	Customer USA 1	3050HF	ADA 027 A	Recycler	Branch	Gen 2000	United States	2023-07-22
1467	Customer USA 1	8540HF	ADA 027 A	Recycler	Branch	Gen 2000	United States	2023-06-13

CARD_DOWNTIME	CASH_DOWNTIME	ACCEPTOR_DOWNTIME	DEPOSITOR_DOWNTIME	EPP_DOWNTIME	PRINTER_DOWNTIME
35125	0	0	0	0	0
35113	0	0	0	0	0
35104	0	86400	0	3524	0
35020	0	0	0	0	0
35014	0	0	0	0	0
35011	0	0	0	0	0
34983	52867	0	0	0	0
34949	0	0	0	0	0
34931	0	0	0	0	0
34924	0	0	0	0	0
34916	35214	34910	34916	0	34915
34893	0	0	0	0	0
34874	17537	17537	17537	79	17536
34839	34847	35537	34847	0	34845
34832	1116	958	1680	0	0
34801	1907	3113	3114	0	2114
34695	0	0	0	0	0
34665	43501	43536	86400	0	0
34659	0	0	0	0	0
34633	0	0	0	0	0
34585	0	0	0	0	0
34553	41114	0	0	86400	0
34508	0	0	0	0	0
34479	0	0	0	0	0
34441	34441	34441	34441	0	34441

Figura 1.1 Vistazo del conjunto de datos

Fuente: Elaboración propia

1.6.2 Análisis exploratorio de datos

- Secuencia temporal de inactividad del módulo lector de tarjetas durante un período de dos años, para este proyecto solo tomará en cuenta las últimas doce semanas.

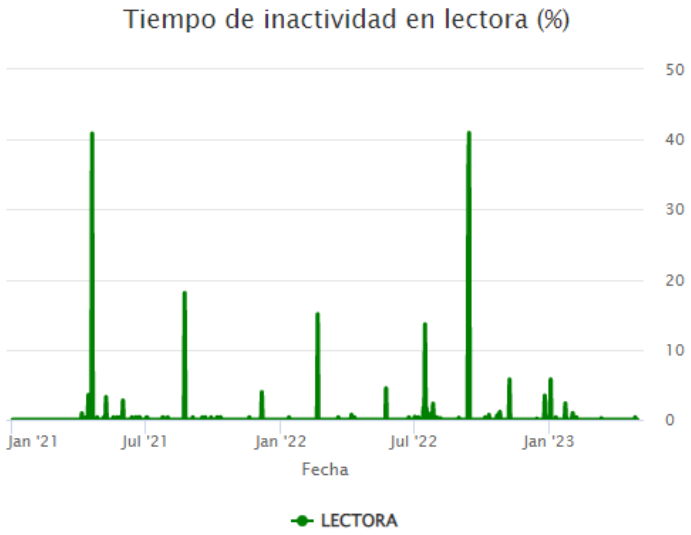


Figura 1.2 Secuencia temporal de inactividad diaria de la lectora de tarjetas

Fuente: Elaboración propia

- Matriz de correlación de las variables numéricas y variables categóricas significativas.

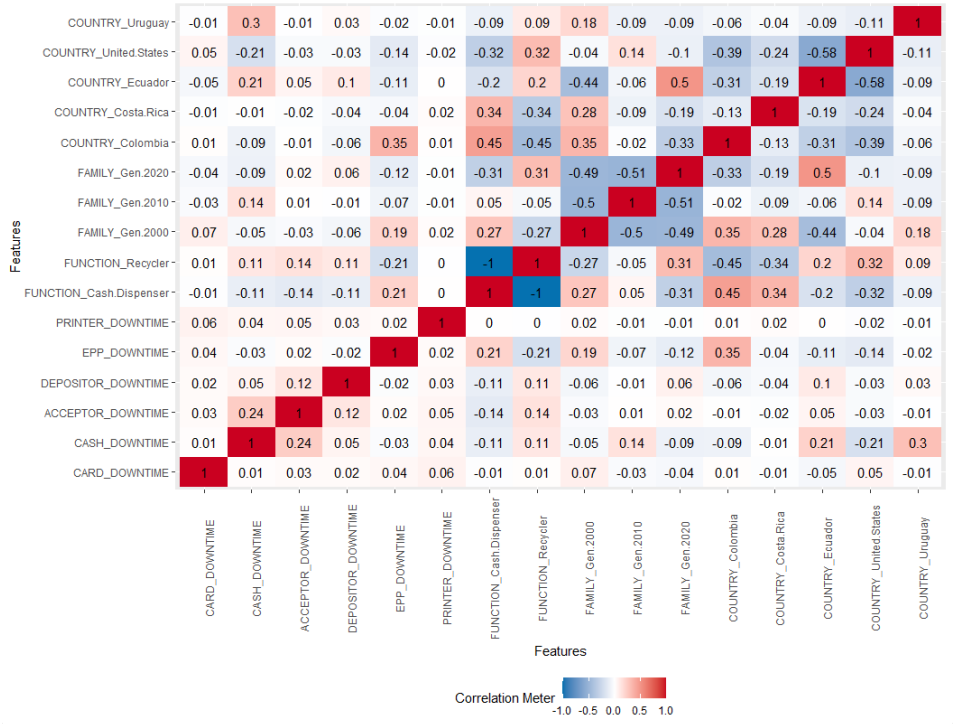


Figura 1.4 Matriz de correlación del conjunto de datos

Fuente: Elaboración propia

- Diagrama de violín de los tiempos de inactividad de los seis módulos críticos

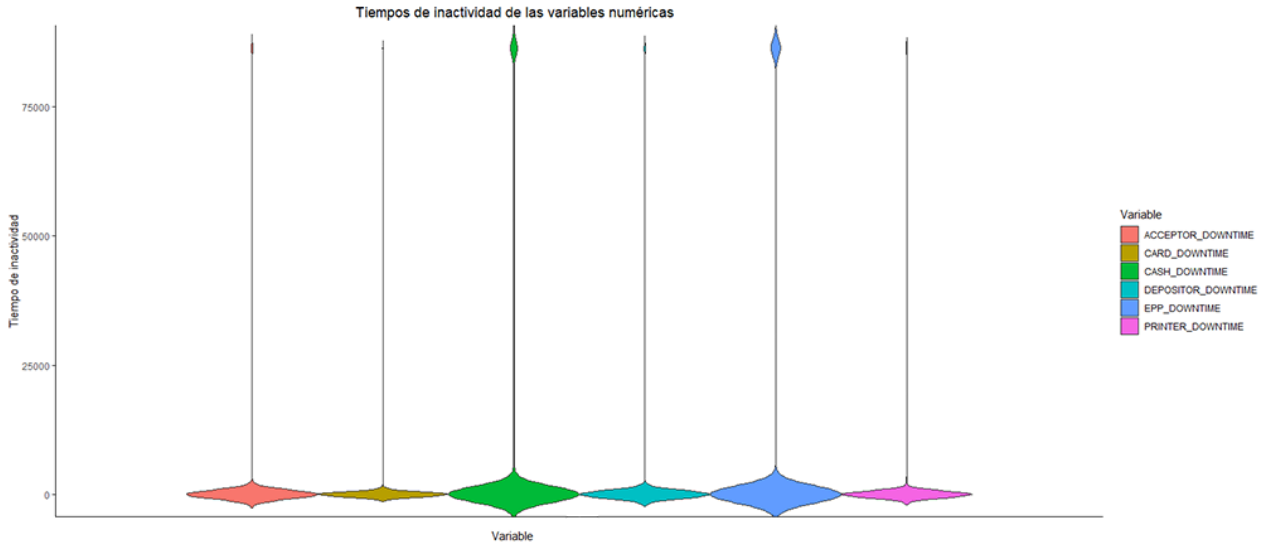


Figura 1.5 Diagrama de violín de las variables numéricas

Fuente: Elaboración propia

- Gráfico Quantile-Quantile (QQ) de los tiempos de inactividad de los seis módulos críticos del cajero automático.

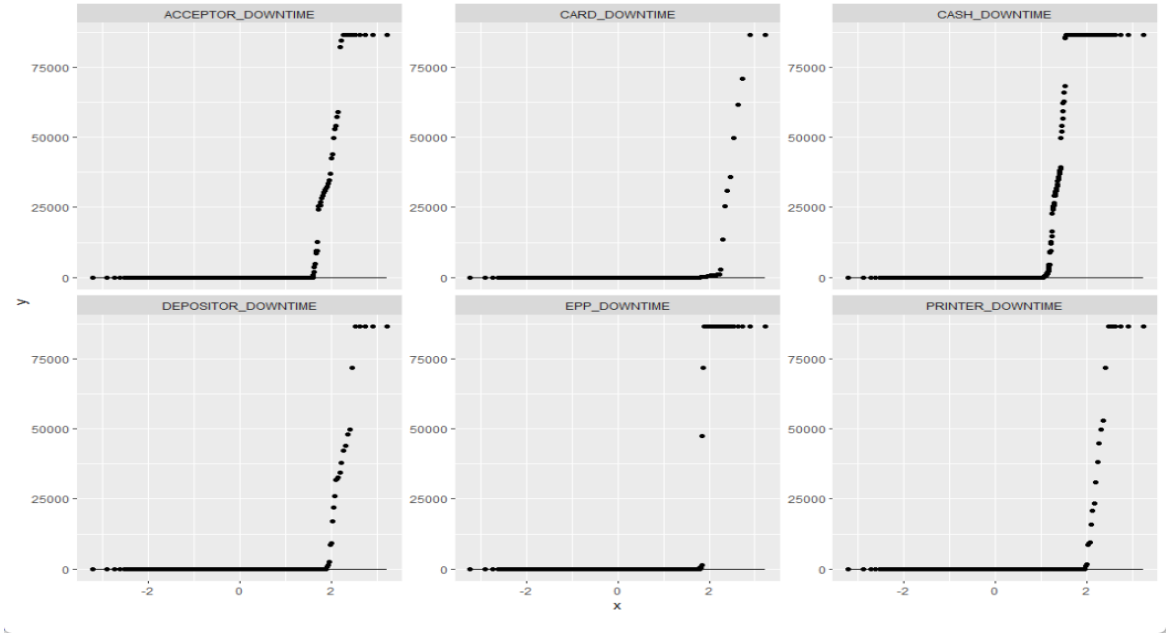


Figura 1.6 Gráficos QQ del conjunto de datos

Fuente: Elaboración propia

- Métricas estadíticas descriptivas de las características del conjunto de datos.

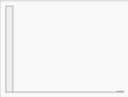
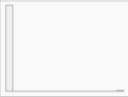
No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	CUSTOMER [character]	1. Customer ECU 1 2. Customer USA 1 3. Customer COL 1 4. Customer ECU 3 5. Customer CRC 1 6. Customer USA 2 7. Customer USA 3 8. Customer USA 4 9. Customer USA 6 10. Customer USA 5 [3 others]	63252 (29.8%) 48888 (23.0%) 19320 (9.1%) 17724 (8.4%) 15036 (7.1%) 10164 (4.8%) 9744 (4.6%) 6888 (3.2%) 6888 (3.2%) 5880 (2.8%) 8400 (4.0%)		212184 (100.0%)	0 (0.0%)
2	ID [character]	1. 0000MTA 2. 0009HF 3. 001000RA 4. 001009IH 5. 0014MTA 6. 00172RC 7. 0019HF 8. 002009IH 9. 0022562EYGTA 10. 0029HF [2516 others]	84 (0.0%) 84 (0.0%) 84 (0.0%) 84 (0.0%) 84 (0.0%) 84 (0.0%) 84 (0.0%) 84 (0.0%) 84 (0.0%) 84 (0.0%) 211344 (99.6%)		212184 (100.0%)	0 (0.0%)
3	MODEL [character]	1. SND 045 A 2. ADA 225 3. XGN 0055 4. XGN 0977 5. ADA 047 6. SND 094 7. SND 002 C 8. XGN 0573 9. XGN 0877 10. XGN 0555 [24 others]	35616 (16.8%) 33852 (16.0%) 20748 (9.8%) 17136 (8.1%) 12264 (5.8%) 10752 (5.1%) 10668 (5.0%) 9912 (4.7%) 9072 (4.3%) 7224 (3.4%) 44940 (21.2%)		212184 (100.0%)	0 (0.0%)
4	FUNCTION [character]	1. Cash Dispenser 2. Recycler	75180 (35.4%) 137004 (64.6%)		212184 (100.0%)	0 (0.0%)
5	SITE [character]	1. Branch 2. Drive Up 3. Island 4. Lobby 5. Office 6. Walk Up	118692 (55.9%) 5124 (2.4%) 1428 (0.7%) 44604 (21.0%) 41916 (19.8%) 420 (0.2%)		212184 (100.0%)	0 (0.0%)
6	FAMILY [character]	1. Gen 2000 2. Gen 2010 3. Gen 2020	68796 (32.4%) 72660 (34.2%) 70728 (33.3%)		212184 (100.0%)	0 (0.0%)
7	COUNTRY [character]	1. Colombia 2. Costa Rica 3. Ecuador 4. United States 5. Uruguay	37044 (17.5%) 15036 (7.1%) 66948 (31.6%) 89796 (42.3%) 3360 (1.6%)		212184 (100.0%)	0 (0.0%)
8	DATETIME [Date]	min : 2023-06-11 med : 2023-07-22 max : 2023-09-02 range : 2m 22d	84 distinct values		212184 (100.0%)	0 (0.0%)
9	CARD_DOWNTIME [numeric]	Mean (sd) : 634.7 (6277.7) min ≤ med ≤ max: 0 ≤ 0 ≤ 86400 IQR (CV) : 0 (9.9)	4200 distinct values		212184 (100.0%)	0 (0.0%)
10	CASH_DOWNTIME [numeric]	Mean (sd) : 6115.3 (20067.6) min ≤ med ≤ max: 0 ≤ 0 ≤ 86400 IQR (CV) : 0 (3.3)	13964 distinct values		212184 (100.0%)	0 (0.0%)
11	ACCEPTOR_DOWNTIME [numeric]	Mean (sd) : 2054.6 (11087.2) min ≤ med ≤ max: 0 ≤ 0 ≤ 86400 IQR (CV) : 0 (5.4)	8535 distinct values		212184 (100.0%)	0 (0.0%)
12	DEPOSITOR_DOWNTIME [numeric]	Mean (sd) : 1500.8 (10054.3) min ≤ med ≤ max: 0 ≤ 0 ≤ 86400 IQR (CV) : 0 (6.7)	4761 distinct values		212184 (100.0%)	0 (0.0%)
13	EPP_DOWNTIME [numeric]	Mean (sd) : 2348.5 (13976) min ≤ med ≤ max: 0 ≤ 0 ≤ 86400 IQR (CV) : 0 (6)	900 distinct values		212184 (100.0%)	0 (0.0%)
14	PRINTER_DOWNTIME [numeric]	Mean (sd) : 1095.8 (8484.1) min ≤ med ≤ max: 0 ≤ 0 ≤ 86400 IQR (CV) : 0 (7.7)	3880 distinct values		212184 (100.0%)	0 (0.0%)

Figura 1.3 Estadística descriptiva de las variables

Fuente: Elaboración propia

CAPÍTULO 2

2. MARCO TEÓRICO

2.1 Conceptos generales

2.1.1 Sobre el cajero automático y sus módulos

2.1.1.1 Definición de cajero automático

Un cajero automático es una computadora que te permite realizar operaciones financieras, entre las que destacan el ingreso o retiro de dinero en efectivo, sin la necesidad de la presencia de un empleado de un banco [6].

Dependiendo del cajero automático y la entidad bancaria, un cajero permite hacer una serie de operaciones u otras. No obstante, teniendo en cuenta el avance de la digitalización y la continua implantación de nuevas funciones en los cajeros automáticos, estos pueden realizar un sinfín de operaciones.

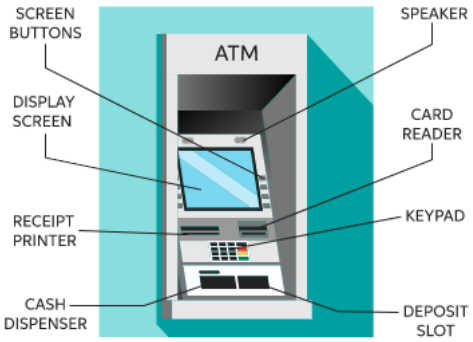


Figura 2.1 Ilustración de un cajero automático y sus módulos

Fuente: https://www.gstsuvidhakendra.org/introduction-about-atm

Entre las funciones más frecuentes que se pueden realizar a través de un cajero automático cabe destacar:

- Retirado de efectivo.
- Depósito de efectivo.
- Transferencias de dinero.
- Obtención y cambio de contraseñas de la cuenta bancaria.
- Gestión de control de cuentas bancarias.
- Recargas telefónicas

- Pago de servicios públicos y privados.
- Contratar otros productos financieros.

2.1.1.2 Módulos del cajero automático

Los módulos de un cajero automático son los componentes físicos y electrónicos que permiten que el cajero realice diversas funciones financieras y transacciones en beneficio de los clientes.

Debido al análisis de los tiempos de inactividad por error de ciertos módulos críticos e indispensables a realizar en este proyecto, a continuación, se definen brevemente cada uno de ellos:

Lector de tarjetas

Es un dispositivo que puede decodificar la información contenida en la banda magnética o el microchip de una tarjeta de crédito o débito, como el número de cuenta, la información del titular y el código de autorización contenidos.

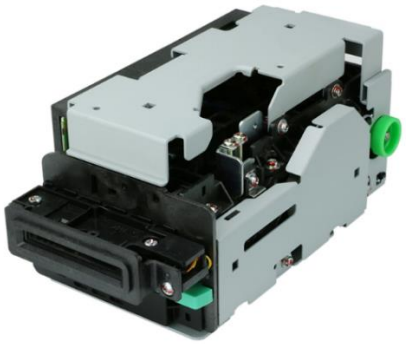


Figura 2.2 Lector de tarjetas de un cajero Wincor PC280
Fuente: https://happyatms.mx/wincor-nixdorf-01750173205

Dispensador de efectivo

El módulo dispensador de dinero de un cajero automático es un componente esencial responsable de dispensar efectivo a los clientes durante las operaciones de retiro. Es la parte del cajero que físicamente almacena y dispensa billetes a los clientes.

Figura 2.3 Dispensador de efectivo de un cajero GRG CDM8240
Fuente: https://www.alibaba.com/product-detail/GRG-Banking-ATM-parts-GRG-Bill_62105203828.html

Reciclador de efectivo

Un módulo reciclador de efectivo en un cajero automático es un componente sofisticado que desempeña un papel clave en la gestión de las transacciones en efectivo. A diferencia de un simple cajero automático que sólo dispensa dinero a los clientes, un reciclador de efectivo tiene capacidades adicionales tanto para dispensar como para aceptar depósitos en efectivo**.**

Figura 2.4 Reciclador de efectivo de un cajero NCR 6687
Fuente: https://www.tigeratmparts.com/atm-machine-parts-diebold-nixdorf-3700-dispenser-00149280000l_p289.html

Depositario de cheques

Es un componente que permite a los clientes depositar cheques en papel para su procesamiento. Está diseñado para agilizar el proceso de depósito de cheques,

haciéndolo más cómodo para los clientes y reduciendo la necesidad de manipulación manual de los cheques por parte del personal del banco.

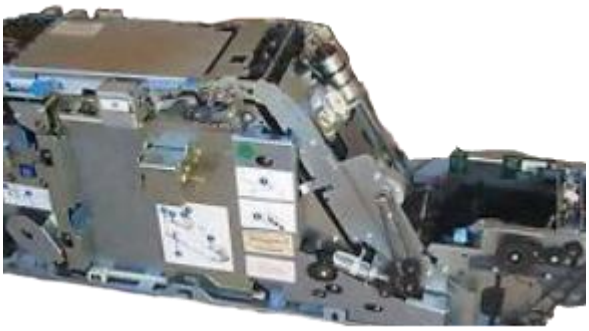


Figura 2.5 Depositario de cheques de un cajero Diebold Nextgen 3700
Fuente: https://www.tigeratmparts.com/ncr-selfserv-87-recycler-ncr-6687-bank-atm-machine-exterior-through-the-wall-cash-recycler_p388.html

Teclado electrónico

Es un dispositivo de entrada seguro que los usuarios utilizan para introducir su número de identificación personal (PIN) o valores de montos durante las transacciones. La función principal es garantizar la confidencialidad y seguridad del dato introducido.

Figura 2.6 Teclado electrónico de un cajero Diebold Opteva 520
Fuente: https://www.tigeratmparts.com/49249440721b-diebold-epp7-keyboard-atm-machine-parts_p18.html

Impresora

Es un componente responsable de imprimir recibos para los clientes. Cuando los clientes realizan varias transacciones como retiradas de efectivo, consultas de

saldo, depósitos o transferencias de fondos, el cajero automático suele proporcionar un recibo impreso como registro de la transacción.

Figura 2.7 Impresora de recibos de un cajero NCR 66XX

Fuente: https://www.alibaba.com/product-detail/ATM-Machine-NCR-66XX-Series-Thermal_1680095020.html

2.1.2 Definición de anomalía

Una anomalía se refiere a cualquier cosa que se desvíe del patrón, comportamiento o condición esperados o típicos dentro de un contexto dado. Son datos que se diferencian del resto por algún motivo, lo que hace pensar que se han generado por una causa distinta [7] y pueden darse en diversos campos.

Figura 2.8 Representación de anomalías

Fuente: https://www.cuemath.com/data/outlier

Las anomalías puntuales

Conocidas como anomalías individuales se refieren a instancias u observaciones específicas de datos que se desvían significativamente del comportamiento esperado o normal dentro de un conjunto de datos. Estas anomalías se caracterizan por su aislamiento y suelen ser distintas e inusuales en comparación

con la mayoría de los puntos de datos de un conjunto de datos determinado. Las anomalías puntuales son las más fáciles de detectar, ya que implican la identificación de puntos de datos individuales que sobresalen significativamente de la norma.

Las anomalías contextuales

Conocidas como anomalías condicionales, se refieren a puntos de datos o sucesos que se consideran anómalos no de forma aislada, sino dentro de un contexto específico o en determinadas condiciones. En otras palabras, estas anomalías no son intrínsecamente anormales, sino que pasan a serlo cuando se comparan con un conjunto particular de circunstancias o un contexto predefinido.

Las anomalías colectivas, también conocidas como anomalías de grupo o anomalías colectivas, se refieren a situaciones en las que un grupo o una colección de puntos de datos muestran un comportamiento anómalo cuando se consideran en conjunto, aunque los puntos de datos individuales puedan parecer normales cuando se examinan de forma aislada. En otras palabras, las anomalías colectivas implican un patrón de anomalías que surgen cuando se analizan múltiples puntos de datos en conjunto.

En definitiva, para la realización de este proyecto nos enfocaremos en las denominadas anomalías puntuales.

2.1.3 Aplicaciones con detección de anomalías

La detección de anomalías es una técnica versátil y valiosa con una amplia gama de aplicaciones en diversos ámbitos [8].

En ciberseguridad, detección de actividades o comportamientos inusuales en la red que puedan indicar un ciberataque o un acceso no autorizado. Identificación de código maligno o comportamiento del software que podría ser señal de infección o virus. Identificación de actividad inusual de inicio de sesión o de cuenta que puede indicar una cuenta de usuario comprometida.

En la gestión de la energía, supervisión de los datos de la red eléctrica en busca de anomalías que puedan indicar fallos en los equipos. Detección de patrones

inusuales de consumo de energía en edificios o instalaciones industriales para mejorar la eficiencia energética.

En la gestión de la cadena de suministro, detección de anomalías en los niveles de inventario, los pedidos o la logística de la cadena de suministro para optimizar las operaciones y reducir las interrupciones. Identificación de desviaciones en el rendimiento o el comportamiento de los proveedores que puedan plantear riesgos para la cadena de suministro.

En el análisis del comportamiento del cliente, detección de patrones de compra o transacciones inusuales para evitar compras fraudulentas. Identificación de comportamientos inusuales de los usuarios en sitios web o aplicaciones, que pueden requerir investigación o intervención personalizada.

En el monitoreo ambiental, identificación de patrones climáticos inusuales o puntos de datos medioambientales que puedan sugerir impactos del cambio climático. Detección de actividad sísmica o meteorológica inusual para emitir alertas tempranas de terremotos, huracanes o inundaciones.

En el análisis financiero, identificación de movimientos de precios o volúmenes de negociación inusuales que puedan indicar manipulación del mercado o anomalías en la negociación. Detección de comportamientos financieros inusuales o incoherencias en los informes crediticios a la hora de evaluar la solvencia.

En la supervisión de redes y sistemas, identificación de patrones de tráfico o comportamientos de red inusuales que podrían ser indicativos de ataques o fallos en la red. Supervisión de los registros y métricas del sistema para detectar comportamientos anómalos que puedan indicar fallos del sistema o violaciones de la seguridad.

En la salud, identificación de valores inusuales en la incidencia de enfermedades o en los síntomas de los pacientes para detectar a tiempo brotes de enfermedades. Detección de registros sanitarios de pacientes o resultados de pruebas médicas inusuales que puedan indicar problemas de salud.

En la fabricación y control de calidad, identificación de productos o componentes defectuosos en los procesos de fabricación mediante la detección de desviaciones de las normas de calidad. Predicción de fallos de máquinas o equipos mediante la supervisión de los datos de los sensores para detectar anomalías en el rendimiento.

En la detección de fraudes, detección de patrones de gasto o transacciones inusuales en tarjetas de crédito que puedan ser señal de actividad fraudulenta. Identificación de reclamaciones anómalas o comportamientos en el seguro que sugieran reclamaciones fraudulentas.

En este proyecto se utilizará el análisis de anomalías sobre los tiempos de inactividad por error de cajeros automáticos para detectar aquellos cajeros o grupos de cajeros más propensos a sufrir averías o indisponibilidad a corto, mediano o largo plazo.

2.2 Breve historia de la detección de anomalías en cajeros automáticos

Figura 2.9 Cajero automático en los años 60
Fuente: https://www.wired.com/2010/09/0902first-us-atm/

En la década de 1960 se introdujeron los primeros cajeros automáticos, la detección de anomalías durante esta época era de nula a rudimentaria y se basaba principalmente en reglas sencillas y límites de transacciones para señalar actividades sospechosas. [9]

En la década de 1970 empezaron a surgir redes de cajeros automáticos que conectaban cajeros de distintos bancos y regiones. Los métodos de detección de

anomalías mejoraron con el uso de sistemas basados en reglas y patrones de transacciones para detectar posibles fraudes.

En la década de 1980 la tecnología de banda magnética se convirtió en la norma para las tarjetas de los cajeros automáticos, lo que permitió una autenticación más segura. Los sistemas de control de transacciones se hicieron más sofisticados, incorporando algoritmos basados en reglas para identificar patrones sospechosos y limitar el acceso no autorizado.

En la década de 1990, se introdujeron las tarjetas con chip EMV (Europay, Mastercard y Visa), que ofrecían una mayor seguridad mediante la autenticación basada en chip. Los sistemas de detección de anomalías evolucionaron para incluir algoritmos capaces de analizar los datos de las transacciones en busca de patrones inusuales y detectar actividades fraudulentas con mayor eficacia. [10]

En la década de 2000 fue testigo de la expansión de las redes de cajeros automáticos en todo el mundo, lo que hizo necesario mejorar las medidas de seguridad. Los sistemas de supervisión de transacciones en tiempo real, junto con el análisis de macrodatos, permitieron a los bancos detectar anomalías rápidamente y reducir los falsos positivos.

En la década de 2010 y 2020 se integraron análisis avanzados de datos e inteligencia artificial en sistemas de detección de anomalías, lo que permitió una detección más precisa de los patrones de fraude. Se implantaron tecnologías para proteger contra el robo de tarjetas, una forma habitual de fraude en cajeros automáticos.

2.3 Soluciones de analítica relacionadas para detección de anomalías

Existen ciertos casos en donde se utilizó la detección de anomalías en cajeros automáticos.

En el artículo “Detección de fraudes usando un modelo de comportamiento” se centra en el desarrollo de detección de actividades fraudulentas en cajeros automáticos [11]. La idea central gira en torno a la creación de un modelo de comportamiento que capte y analice los patrones únicos de uso de los cajeros por parte de cada usuario mediante un algoritmo de detección de anomalías que

compara las transacciones de los cajeros automáticos en tiempo real con el modelo de comportamiento. Si una transacción se desvía significativamente del comportamiento esperado basado en el modelo, se marca como potencialmente fraudulenta.

Los autores recopilan una gran cantidad de datos de varios cajeros automáticos, como marcas de tiempo de las transacciones, tipos de transacciones, ubicaciones e identificadores de usuario. Proponen un método para construir un modelo de comportamiento específico del usuario mediante el análisis de los datos históricos de transacciones de cada usuario de cajero automático.
El modelo de comportamiento captura los patrones típicos de transacción, incluyendo la frecuencia de las transacciones, la consistencia de la ubicación y los tipos de transacción.

Los autores presentan los resultados de sus experimentos, que demuestran la eficacia de su sistema de detección del fraude basado en el comportamiento.
Comparan el rendimiento de su enfoque con el de los métodos tradicionales basados en reglas y destacan la mayor precisión y los menores índices de falsos positivos conseguidos con el modelo basado en el comportamiento.

Por otra parte, el artículo titulado "Detección de anomalías de estado de las transacciones en cajeros automáticos basado en aprendizaje no supervisado" recolecta información para detectar anomalías de cuatro indicadores: el desbalance de volumen de transacciones por días, el índice de fallo en las transacciones, la lentitud en el procesamiento, los tiempos de respuesta muy extensas de las transacciones bancarias donde construye un esquema de detección de anomalías mediante las técnicas de agrupación de K-means y red neuronal BP. [12]

Las transacciones son clasificadas con diferentes grados de advertencia para diferentes condiciones anormales.
Alerta roja: Cuando los datos detectados se desvían de los previstos en más de un 50%, indica que la condición anormal es especialmente grave.
Alerta naranja: Cuando la desviación entre los datos detectados y los previstos se sitúa entre el 35% y el 50%, indica que la condición anormal es grave.

Alerta amarilla: Cuando la desviación entre los datos detectados y los previstos se sitúa entre el 20% y el 35%, indica que la anomalía es relativamente grave.
Alerta azul: cuando la desviación entre los datos detectados y los previstos se sitúa entre el 10% y el 20%, indica que la anomalía no es grave.
Alerta verde: cuando la desviación entre los datos detectados y los previstos se sitúa entre el 0% y el 10%, indica que el estado de las operaciones está aproximadamente dentro de la normalidad.

Los autores llegan a la conclusión de que clasificación de las anomalías no es lo suficientemente precisa, debido a que el rango de datos es demasiado pequeño.

2.4 Técnicas de detección de anomalías

El aprendizaje no supervisado, también conocido como aprendizaje automático no supervisado, utiliza algoritmos de aprendizaje automático para analizar y agrupar conjuntos de datos no etiquetados. Estos algoritmos descubren patrones ocultos o agrupaciones de datos, su capacidad para descubrir similitudes y diferencias en la información lo convierten en la solución ideal para el análisis exploratorio de datos, las estrategias de venta cruzada, la segmentación de clientes y el reconocimiento de imágenes. [13]

Los modelos de aprendizaje no supervisado se utilizan para tres tareas principales: agrupación, asociación y reducción de la dimensionalidad. A continuación, definimos dos métodos para el desarrollo de este proyecto.

2.4.1 K-means clustering (Agrupación K-medias)

Es un tipo de aprendizaje no supervisado, que se utiliza cuando se tienen datos sin etiquetar (es decir, datos sin categorías o grupos definidos). El objetivo de este algoritmo es encontrar grupos en los datos, con el número de grupos representado por la variable K. El algoritmo trabaja de forma iterativa para asignar cada punto de datos a uno de los K grupos basándose en las características que se proporcionan. Los puntos de datos se agrupan en función de la similitud de las características. Los resultados del algoritmo de agrupación de K-medias son:

a). Los centroides de los K clústeres, que pueden utilizarse para etiquetar nuevos datos.

b). Etiquetas para los datos de entrenamiento (cada punto de datos se asigna a un único clúster).

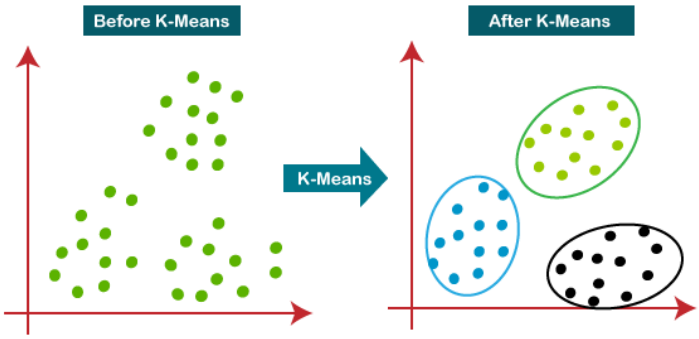


Figura 2.10 Representación de Agrupación K-means
Fuente: https://www.analyticsvidhya.com/blog/2021/04/k-means-clustering-simplified-in-python/

En lugar de definir los grupos antes de observar los datos, el clustering permite encontrar y analizar los grupos que se han formado.

Cada centroide de un clúster es una colección de valores de características que definen los grupos resultantes. El análisis de los pesos de las características del centroide puede utilizarse para interpretar cualitativamente qué tipo de grupo representa cada conglomerado. [14]

Implementación de Agrupación K-Means:

Paso 1: Elegir el número K de clústeres, es decir, K=2 (aunque existe el método del codo para determinar el K óptimo), para segregar el conjunto de datos y colocarlos en diferentes clústeres respectivos. Elegiremos 2 puntos al azar que actuarán como centroides para formar el clúster.

Paso 2: A continuación, asignaremos cada punto de datos a un gráfico de dispersión en función de su distancia al punto K o centroide más cercano. Para ello, trazaremos una mediana entre ambos centroides.

Paso 3: Los puntos a la izquierda de la línea están cerca del centroide azul, y los puntos a la derecha de la línea están cerca del centroide amarillo. Los de la

izquierda forman un conglomerado con el centroide azul, y los de la derecha con el centroide amarillo.

Paso 4: Repita el proceso eligiendo un nuevo centroide. Para elegir los nuevos centroides, encontraremos el nuevo centro de gravedad de estos centroides, como se muestra a continuación.

Paso 5: A continuación, reasignaremos cada punto de datos al nuevo centroide. Repetiremos el mismo proceso anterior (utilizando una línea mediana). El punto de datos amarillo en el lado azul de la línea mediana se incluirá en el conglomerado azul.

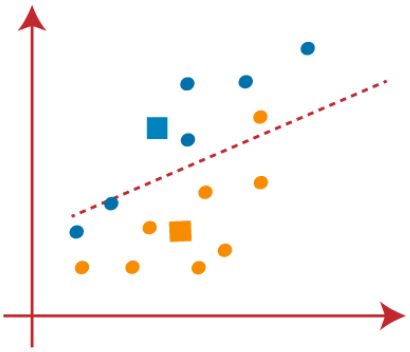


Figura 2.11 Implementación K-medias (Paso 5)
Fuente: https://www.analyticsvidhya.com/blog/2021/04/k-means-clustering-simplified-in-python

Paso 6: Como se ha producido la reasignación, repetiremos el paso anterior de encontrar nuevos K centroides.

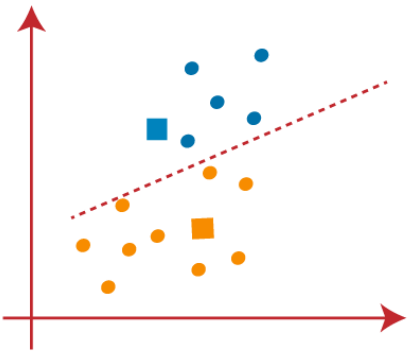


Figura 2.12 Implementación K-medias (Paso 6)
Fuente: https://www.analyticsvidhya.com/blog/2021/04/k-means-clustering-simplified-in-python

Paso 7: Repetiremos el proceso anterior para encontrar el centro de gravedad de k centroides, como se muestra a continuación.

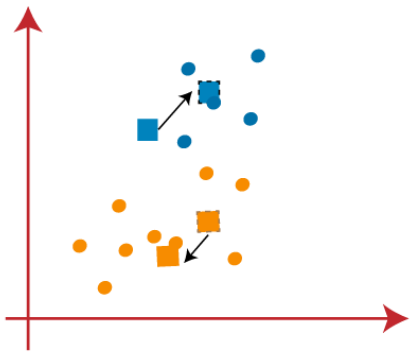


Figura 2.13 Implementación K-medias (Paso 7)
Fuente: https://www.analyticsvidhya.com/blog/2021/04/k-means-clustering-simplified-in-python

Paso 8: Después de encontrar los nuevos k centroides, trazaremos de nuevo la línea mediana y reasignaremos los puntos de datos, como en los pasos anteriores.

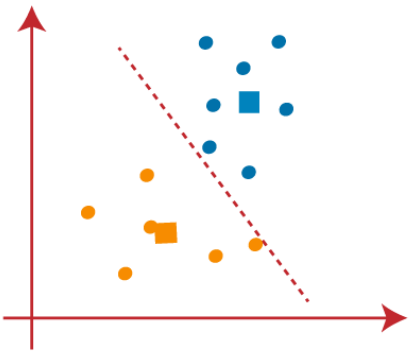


Figura 2.14 Implementación K-medias (Paso 8)
Fuente: https://www.analyticsvidhya.com/blog/2021/04/k-means-clustering-simplified-in-python

Paso 9: Finalmente segregaremos los puntos según la línea mediana, de modo que se formen dos grupos y no se incluya ningún punto diferente en un solo grupo.

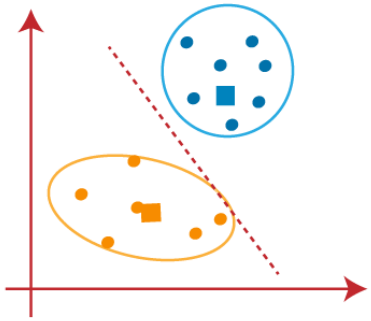


Figura 2.15 Implementación K-medias (Paso 9)
Fuente: https://www.analyticsvidhya.com/blog/2021/04/k-means-clustering-simplified-in-python

El clúster final formado quedaría así:

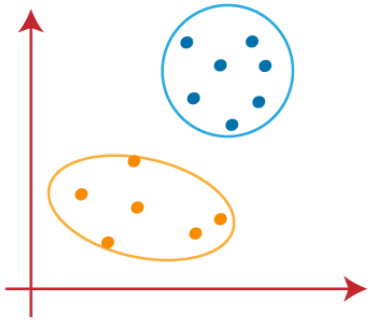


Figura 2.16 Implementación K-medias (Resultado final)
Fuente: https://www.analyticsvidhya.com/blog/2021/04/k-means-clustering-simplified-in-python

La selección del número óptimo de clústeres para el algoritmo no debe ser aleatorio. Todos y cada uno de los clústeres se forman calculando y comparando las distancias medias de cada punto de datos dentro de un clúster con respecto a su centroide.

Podemos elegir el número correcto de clústeres con la ayuda del método “Within Cluster Sum of Squares” (WCSS). WCSS es la suma de los cuadrados de las distancias entre los puntos de datos de cada clúster y su centroide.

La idea principal es minimizar la distancia (por ejemplo, la distancia euclidiana) entre los puntos de datos y el centroide de los clústeres. El proceso se itera hasta alcanzar un valor mínimo para la suma de distancias. [15]

Método del codo

Estos son los pasos para seguir para encontrar el número óptimo de clúster utilizando el método del codo:

Paso 1: Ejecutar la agrupación de K-medias en un conjunto de datos dado para diferentes valores de K (que van de 1 a 10).

Paso 2: Para cada valor de K, calcule el valor WCSS.

Paso 3: Trazar un gráfico o curva entre los valores WCSS y el número respectivo de clústeres K.

Paso 4: El punto agudo de la curva o un punto (parecido a la articulación de un codo) del gráfico, como un brazo, se considerará como el mejor u óptimo valor de K.

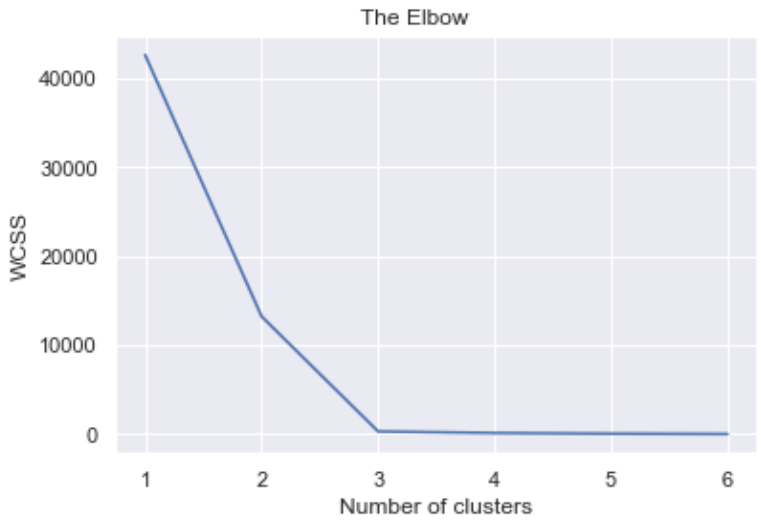


Figura 2.17 Gráfica del método del codo
Fuente: https://www.analyticsvidhya.com/blog/2021/04/k-means-clustering-simplified-in-python

Este método muestra que 3 es el número ideal de clústeres.

2.4.2 Isolation Forest (Bosques de aislamiento)

Es una técnica para identificar valores atípicos en los datos, el método emplea árboles binarios para detectar anomalías, lo que se traduce en una complejidad de tiempo lineal y un bajo uso de memoria que resulta muy adecuado para procesar grandes conjuntos de datos.

Desde su introducción ha ganado popularidad como algoritmo rápido y fiable para la detección de anomalías en diversos campos [16]. Es similar a los Bosques Aleatorios, se construyen a partir de árboles de decisión, como aquí no hay etiquetas predefinidas, se trata de un modelo no supervisado.

En un Isolation Forest, los datos submuestreados aleatoriamente se procesan en una estructura de árbol basada en características seleccionadas al azar. Las muestras que se adentran más en el árbol tienen menos probabilidades de ser anomalías, ya que necesitaron más cortes para aislarlas. Del mismo modo, las muestras que acaban en ramas más cortas indican anomalías, ya que al árbol le resultó más fácil separarlas de otras observaciones.

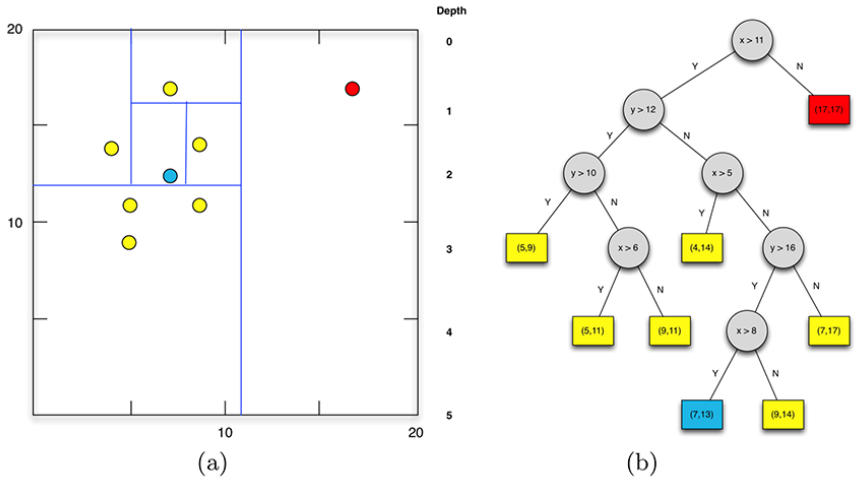


Figura 2.18 Representación de árbol del método Isolation Forest

Tomado de https://zhuanlan.zhihu.com/p/32841893

Implementación de Isolation Forest:

Paso 1: Cuando se da un conjunto de datos, se selecciona una submuestra aleatoria de los datos y se asigna a un árbol binario.

Paso 2: La ramificación del árbol comienza seleccionando primero una característica aleatoria (del conjunto de todas las características N). A continuación, la ramificación se realiza en un umbral aleatorio (cualquier valor en el intervalo de valores mínimo y máximo de la característica seleccionada).

Paso 3: Si el valor de un punto de datos es menor que el umbral seleccionado, pasa a la rama izquierda o a la derecha. De este modo, un nodo se divide en las ramas izquierda y derecha.

Paso 4: Este proceso desde el paso 2 se continúa recursivamente hasta que cada punto de datos está completamente aislado o hasta que se alcanza la profundidad máxima (si está definida).

Paso 5: Los pasos anteriores se repiten para construir árboles binarios aleatorios.

Una vez creado un conjunto de árboles del Isolation Forest, se completa el entrenamiento del modelo. Durante la puntuación, un punto de los datos se recorre a través de todos los árboles que se entrenaron anteriormente. Ahora, se asigna una "puntuación de anomalía" a cada uno de los puntos de datos en función de la profundidad del árbol necesaria para llegar a ese punto. Esta puntuación es una agregación de la profundidad obtenida de cada uno de los árboles. Se asigna una puntuación de anomalía de -1 a las anomalías y de 1 a los puntos normales en función del parámetro de contaminación (porcentaje de anomalías presentes en los datos) proporcionado [17].

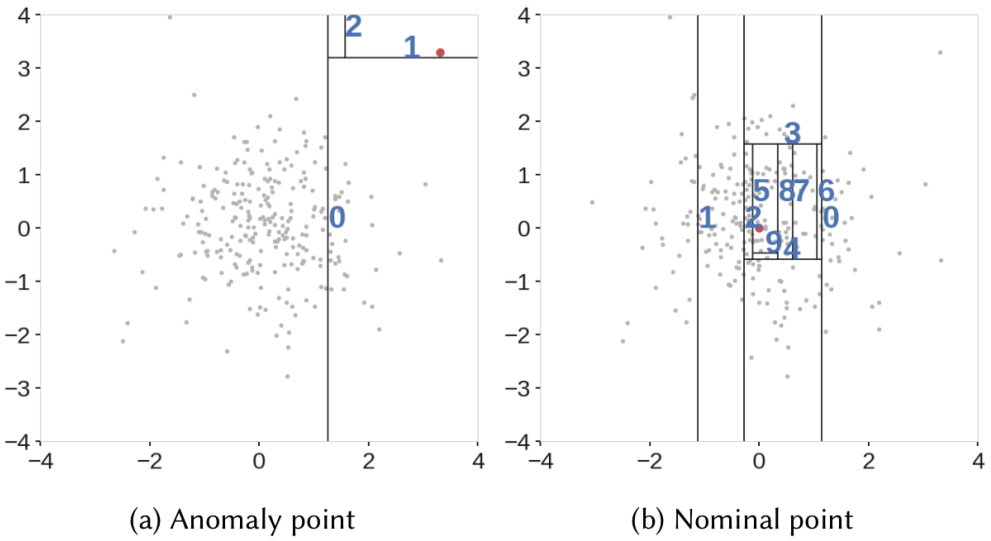


Figura 2.19 Aislamiento de un valor anómalo usando Isolation Forest

Fuente: S. Hariri, M. C. Kind and R. J. Brunner, "Extended Isolation Forest," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 4, pp. 1479-1489, 1 April 2021

2.5 Métricas de evaluación

Evaluar el rendimiento de un sistema de detección de anomalías es esencial para garantizar su eficacia y fiabilidad. Para evaluar el rendimiento de un modelo o algoritmo se utilizan diversos parámetros y técnicas de evaluación. La elección de la métrica depende de las características específicas de los datos y de los objetivos de la tarea de detección de anomalías.

A continuación, se presentan algunas métricas de evaluación comunes a utilizar durante el desarrollo de este proyecto:

2.5.1 Puntuación de silueta

Esta métrica cuantifica cuán similar es cada punto de datos a su propio clúster en comparación con otros clústeres cercanos. Los valores de silueta oscilan entre -1 y 1, donde valores más altos indican que los puntos están mejor asignados a sus clústeres.

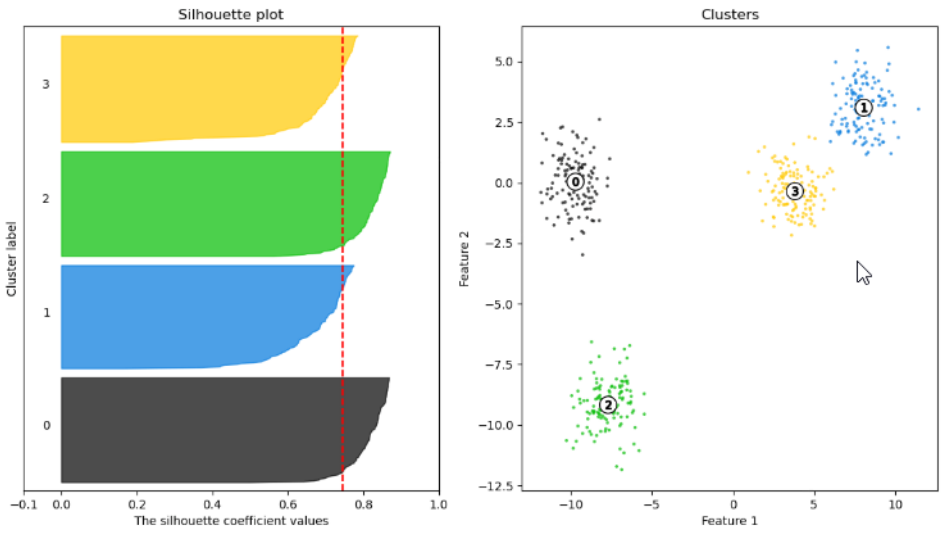


Figura 2.20 Ejemplo del método puntuación de silueta

Fuente: https://www.baeldung.com/cs/silhouette-values-clustering

2.5.2 Índice Calinski-Harabasz Index (CH)

Es una de las medidas de evaluación de los algoritmos de clustering. Se suele utilizar para evaluar la división de un algoritmo de agrupación K-Medias para un número determinado de clústeres [18].

El índice de Calinski-Harabasz (también conocido como Criterio de la Relación de Varianza) se calcula como una relación entre la suma de la dispersión interclúster y la suma de la dispersión intraclúster para todos los clústeres (donde la dispersión es la suma de las distancias al cuadrado).

Un índice alto significa una mejor agrupación, ya que las observaciones de cada clúster están más próximas entre sí, mientras que los propios clústeres están más alejados unos de otros.

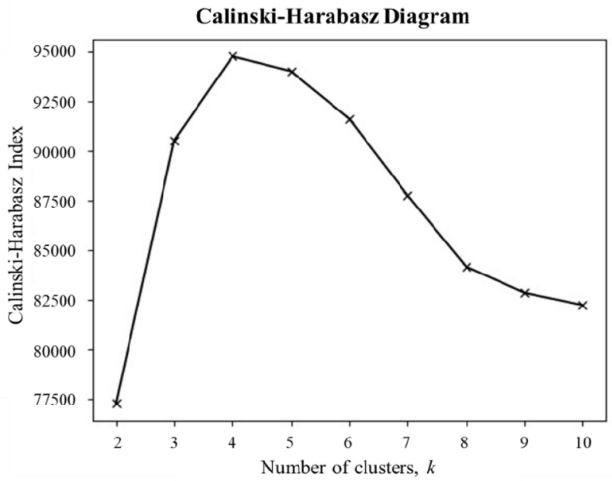


Figura 2.21 Representación del Índice Calinski-Harabasz

Fuente: Xu, Jieyan et al. (2020). "Clustering-based probability distribution model for monthly residential building electricity consumption analysis"

2.5.3 Índice Davies-Bouldin (DBI)

Considera como la mejor métrica de evaluación para técnicas de agrupación [19]. El índice de Davies-Bouldin es una de las medidas de evaluación de los algoritmos de clustering. Se suele utilizar para evaluar que tan bueno es la división de un algoritmo de agrupación K-Means para un número determinado de clústeres. Un valor más bajo de DBI indica una mejor calidad de agrupamiento, donde los clústeres están más separados y son más homogéneos.

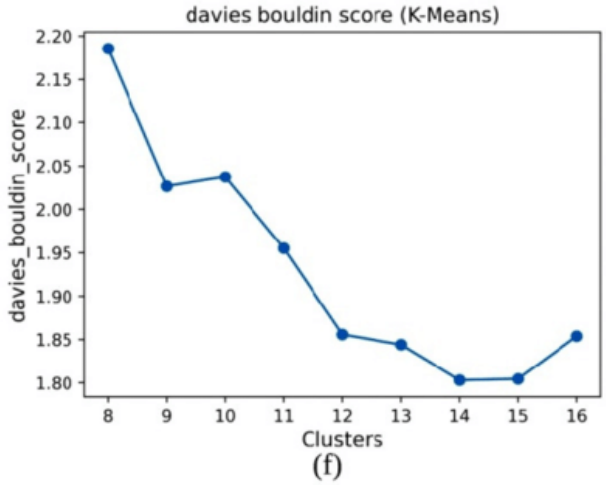


Figura 2.22 Representación del Índice Davies Bouldin

Fuente: https://www.researchgate.net/figure/Davies-Bouldin-index-for-different-number-of-clusters-when-the-SOM-of-the-reduced-madelon_fig6_228988686

2.5.4 Puntuación de anomalía

La puntuación de anomalía en Isolation Forest se refiere a la medida utilizada para determinar el grado de anomalía o anormalidad de un punto de datos dentro de un conjunto de datos. En el algoritmo Isolation Forest, cada punto de datos recibe una puntuación de anomalía, que indica la facilidad con la que fue aislado por el bosque de árboles de aislamiento.

Esta puntuación de anomalía se calcula en función de la profundidad media de los árboles que aíslan un punto de datos específico. Los puntos anómalos tienden a aislarse más rápidamente que los puntos normales.

Normalmente, los puntos con puntuaciones de anomalía más altas se consideran más anómalos o atípicos porque necesitaron menos divisiones o pasos para ser aislados dentro de los árboles de aislamiento. Por el contrario, los puntos con puntuaciones de anomalía más bajas se consideran más normales o comunes dentro del conjunto de datos.

2.6 Software y herramientas de analítica a utilizar

2.6.1 Lenguaje Python

El proyecto será desarrollado en el lenguaje de programación de alto nivel Python que se destaca por su legibilidad y sintaxis clara y concisa, se ha vuelto

extremadamente popular en la comunidad de programadores debido a su facilidad de uso y versatilidad. [20]

2.6.2 R y RStudio

R es un entorno y lenguaje de programación el cual proporciona una amplia variedad de técnicas estadísticas como manipulación de datos, análisis estadístico, aprendizaje automático, modelado y predicciones, entre otras.
R está disponible como Software Libre bajo los términos GNU de la Free Software Foundation en forma de código fuente. Se compila y se ejecuta en plataformas UNIX (Linux y Mac OS) y Windows.

RStudio es un entorno de desarrollo integrado (IDE) para R. Incluye una consola, edición y ejecución directa de código, así como herramientas para gráficos, historial, depuración y gestión del espacio de trabajo [21].

2.6.3 PyCaret

Es una biblioteca de Python diseñada para facilitar y acelerar el proceso de desarrollo y evaluación de modelos de aprendizaje automático. Se enfoca en simplificar las tareas comunes de preparación de datos, entrenamiento de modelos, evaluación de modelos y selección de los modelos más adecuados. [22].

2.6.4 StreamLit

Es una librería de desarrollo de código abierto en Python que simplifica la creación y distribución de aplicaciones web personalizadas y atractivas para tareas relacionadas con el aprendizaje automático y la ciencia de datos. En cuestión de minutos se puede desarrollar y publicar aplicaciones de datos interactivas poderosas [23].

CAPÍTULO 3

3. DISEÑO E IMPLEMENTACION

3.1 Exploración y validación de datos y fuentes

3.1.1 Fuente de datos

Todo cajero automático de todos los clientes que hayan adquirido la aplicación de monitoreo recolecta todo tipo de información al detalle del equipo, desde datos de software y hardware instalados hasta información de sensores de los diferentes módulos del cajero.

Toda esta información es enviada a un servidor centralizado donde el cliente dispone de acceso para monitorear en línea el estado actual de sus equipos.

3.1.2 Extracción de datos

Debido a la ausencia de un API u otro método para obtener los datos fácilmente, la única forma de descargarlos es manera manual. La descarga se realiza por cliente y rango de fechas donde se obtiene un archivo .csv con la información descriptiva e inactividad por error de los módulos críticos de cada cajero automático detallado en el capítulo 1.

3.1.3 Exploración de datos

La exploración de datos en un conjunto de datos es un proceso crucial para comprender y aprovechar al máximo la información que contiene.

A continuación, se detalla de manera descriptiva las características del conjunto de datos:

- Con respecto a la distribución de los clientes.

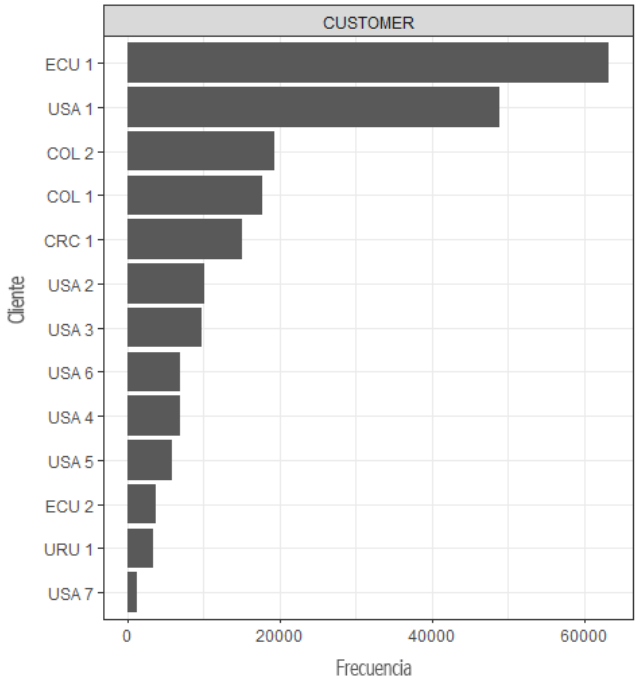


Figura 3.1 Histograma de frecuencias de la variable Customer

Fuente: Elaboración propia

Aproximadamente alrededor del 30% corresponde al cliente “ECU 1” seguido del cliente “USA 1” con 23% y el cliente “COL 2” con un 9%.

- Con respecto a la familia del cajero o generación.

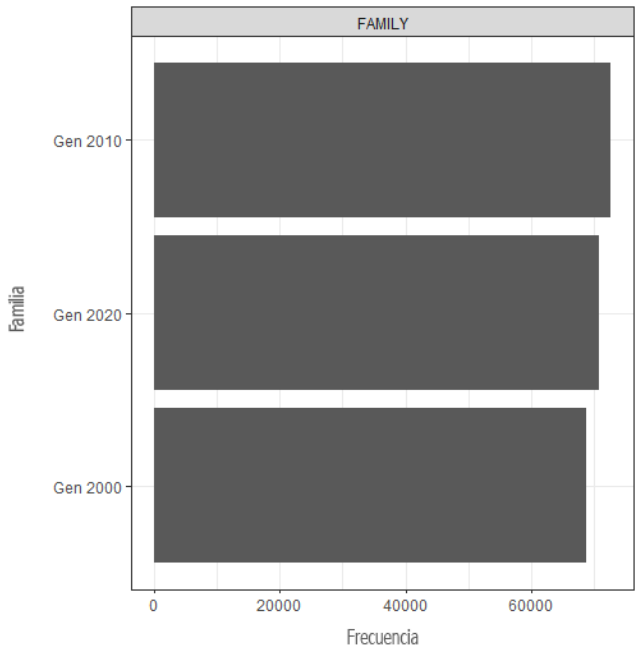


Figura 3.2 Histograma de frecuencias de la variable Family

Fuente: Elaboración propia

La generación del 2010 corresponde al 34%, la generación del 2020 al 33% y la generación del 2000 al 32%.

- Con respecto a la distribución de acuerdo con el modelo del cajero.

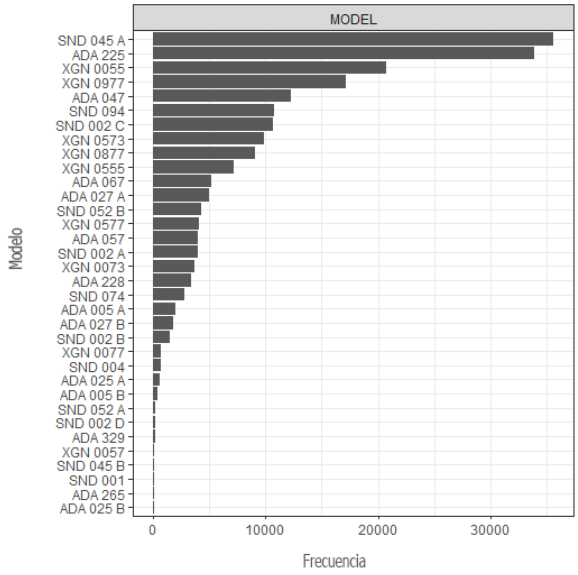


Figura 3.3 Histograma de frecuencias de la variable Model

Fuente: Elaboración propia

El modelo “SND 045 A” comparte en primera posición con el 17% de los datos seguido del modelo “ADA 225” con un 16% y el modelo “XGN 0055” con un 10% aproximadamente.

- Con respecto a la distribución con la función del cajero.

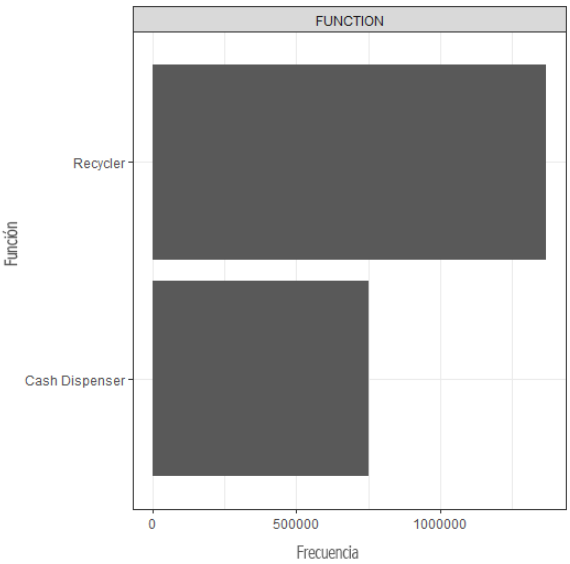


Figura 3.4 Histograma de frecuencias de la variable Function

Fuente: Elaboración propia

Según su función la mayoría de los datos con 65% los cajeros son “Recycler” mientras que el 35% son “Cash Dispenser”.

- Con respecto a la distribución del tipo o ubicación del cajero.

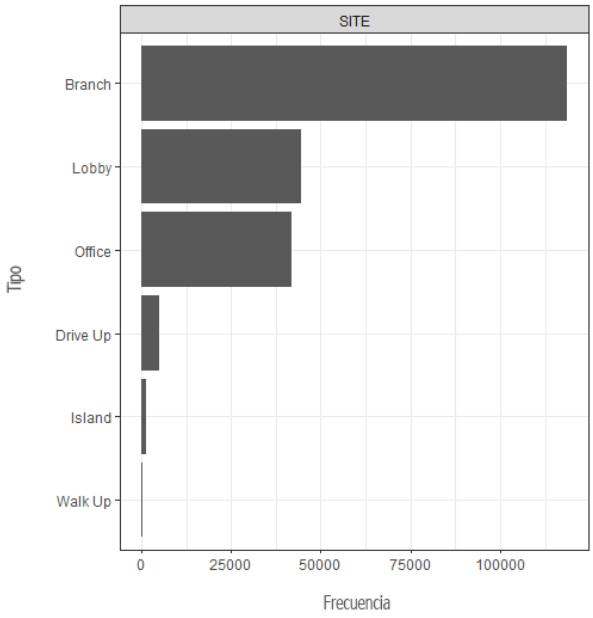


Figura 3.5 Histograma de frecuencias de la variable Site

Fuente: Elaboración propia

Según su tipo o ubicación el 56% corresponde a tipo “Branch” aproximadamente mientras que el 21% son de tipo “Office” y el 20% de tipo “Lobby”.

- Con respecto al país del conjunto de datos:

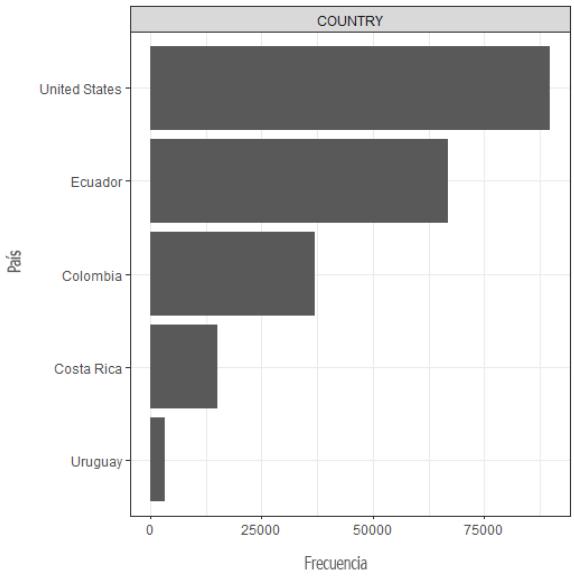


Figura 3.6 Histograma de frecuencias de la variable Country

Fuente: Elaboración propia

En primer lugar, con el 42% pertenece a Estados Unidos, en segundo lugar 32% a Ecuador y el tercer puesto Colombia con el 17%.

3.1.4 Limpieza y verificación de datos

Una vez obtenido el conjunto de datos inicial se verifica la integridad de la información para asegurarte de que se haya recopilado de manera precisa y completa y no se hayan generados errores durante la extracción.

Varias características categóricas como nombre del cliente, modelo del cajero, familia del cajero fueron cambiados con valores genéricos para conservar la privacidad de la información de los clientes.

En la categoría “Function” hubo que imputar los valores “Unknown” a su correspondiente función de forma manual ya sea “Recycler” o “Cash Dispenser”.

Algunos cajeros no tenían las secuencias temporales completas para ciertos días seguidos o incluso semanas completas debido a que eran cajeros nuevos y disponían de aquellos datos por lo que se eliminó toda la información para ese cajero del conjunto de datos.

3.2 Prototipos de modelos

Para el análisis de este proyecto se definió que los modelos a utilizar serían K-means e Isolation Forest, para el entrenamiento de dichos modelos se utilizan los datos de 12 semanas entre junio y agosto del 2023 con los valores promediados semanalmente de inactividad por error de todos los módulos.

3.2.1 K-means

En primera instancia ser probó el método K-means donde el objetivo principal es agrupar un conjunto de datos en grupos o clústeres basados en similitudes entre las muestras. Cada grupo contiene muestras que son más similares entre sí que con las muestras de otros grupos.

Pipeline del modelo

Figura 3.7 Diagrama de flujo de creación del modelo K-means

Fuente: Librería Pycaret

1. El conjunto de datos es sometido a imputación numérica, si no se encontrase algún valor numérico se imputa con la media de los datos existentes.
2. El conjunto de datos es sometido a imputación categórica, si no se encontrase algún valor categórico se imputa con la moda (valor de mayor frecuencia) de los datos existentes.
3. El conjunto de datos es sometido a la codificación One-hot, una técnica en donde los datos categóricos son convertidos a datos binarios, es decir de ceros y unos, paso fundamental para entrenar el modelo con este tipo de características.

Tabla 3.1 Conjunto de datos luego de la codificación One-hot

ID	CUSTOMER_USA 1	CUSTOMER_ECU 1	CUSTOMER_USA 3	CUSTOMER_USA 4	CUSTOMER_CRC 1	CUSTOMER_USA 6	...
002D7833	1	0	0	0	0	0	...
00385E3F	1	0	0	0	0	0	...
0041F161	0	0	1	0	0	0	...
0057BF92	0	0	0	0	0	1	...
007ED0CC	0	0	0	0	1	0	...
...	...	...	...	...	...	...	...

4. El conjunto de datos es sometido a la agrupación de categorías, es decir, todas aquellas columnas dependientes o que estén relacionadas como son las columnas de tiempos de inactividad de las semanas para cada módulo del cajero generan columnas adicionales al conjunto de datos con medidas estadísticas descriptivas como el máximo, mínimo, media, mediana, moda y desviación estándar de cada fila de las columnas relacionadas.

Tabla 3.2 Conjunto de datos luego de la agrupación de categorías

ID	min(group_1)	max(group_1)	mean(group_1)	std(group_1)	median(group_1)	mode(group_1)	...
002D7833	-0.044593	-0.286349	-0.122405	-0.280957	-0.058522	-0.044626	...
00385E3F	-0.044593	-0.374393	-0.154037	-0.369832	-0.058522	-0.044626	...
0041F161	-0.044593	-0.374393	-0.154037	-0.369832	-0.058522	-0.044626	...
0057BF92	-0.044593	-0.368222	-0.152795	-0.363864	-0.058522	-0.044626	...
007ED0CC	-0.044593	0.22141	0.118229	0.272986	-0.045284	-0.044626	...
...	...	...	...	...	...	...	...

5. El conjunto de datos es sometido a normalización bajo el método Z-score que transforma los datos en una distribución normal estándar con una media de 0 y una desviación estándar de 1.

Tabla 3.3 Conjunto de datos luego de la normalización

ID	W0	W1	W2	W3	W4	W5	...
002D7833	-0.044593	-0.286349	-0.12241	-0.280957	-0.058522	-0.044626	...
00385E3F	-0.044593	-0.374393	-0.15404	-0.369832	-0.058522	-0.044626	...
0041F161	-0.044593	-0.374393	-0.15404	-0.369832	-0.058522	-0.044626	...
0057BF92	-0.044593	-0.368222	-0.1528	-0.363864	-0.058522	-0.044626	...
007ED0CC	-0.044593	0.22141	0.118229	0.272986	-0.045284	-0.044626	...
...	...	...	...	...	...	...	...

El conjunto de datos de entrenamiento antes de la ejecución del pipeline del modelo comprendía (2526 filas, 78 columnas), al finalizar el pipeline quedó conformado por (2526 filas, 170 columnas) de las cuales:

- 72 columnas resultantes corresponden a las 12 semanas de cada uno de los 6 módulos del cajero.
- 62 columnas resultantes corresponden a las 7 columnas categóricas luego de la codificación One-hot.
- 36 columnas resultantes corresponden a la agrupación de características relacionadas.

Para finalizar se definió tres agrupaciones en K-means (K = 3) para categorizar los cajeros automáticos de acuerdo con su condición sobre tiempos de servicio.

Por ejemplo:

- **Grupo 0**: Cajeros normales
- **Grupo 1**: Cajeros problemáticos
- **Grupo 2**: Cajeros regulares

Luego del entrenamiento de modelo se obtiene las siguientes métricas más utilizadas en K-means con los siguientes resultados:

Tabla 3.4 Métricas de entrenamiento del modelo K-means

Silhouette	Calinski-Harabasz	Davies-Bouldin
0.36	261.3399	1.3142

De acuerdo con el gráfico de siluetas la calidad de un agrupamiento es “apropiado” como lo indica Rousseeuw, Peter (1987) [24] debido al valor mayor a 0 con una puntuación promedio de 0.36

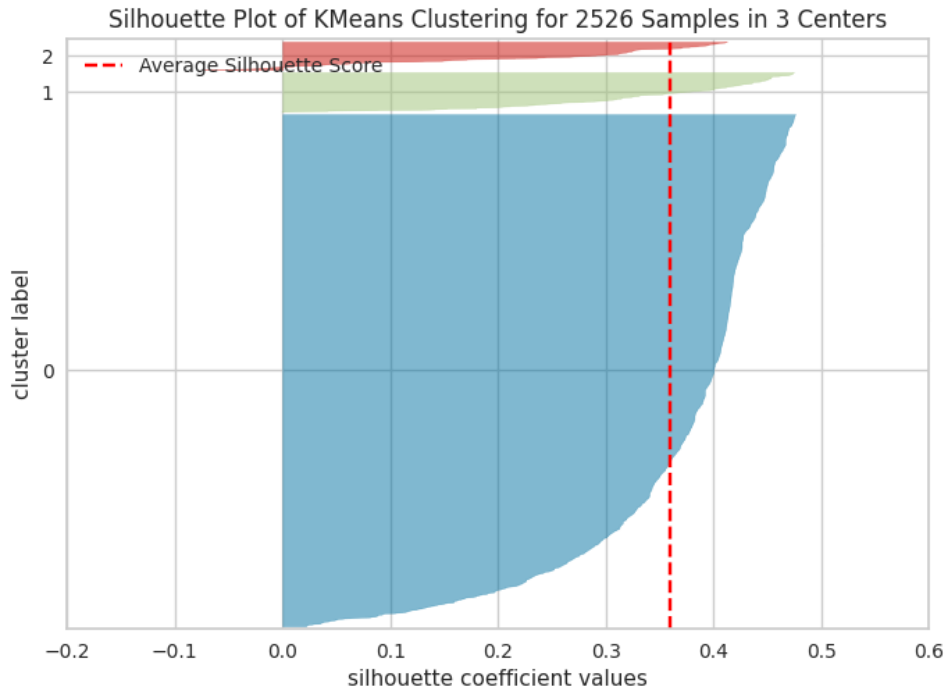


Figura 3.8 Gráfico de siluetas del modelo K-means

Fuente: Librería Pycaret

El mapa de distancias generado de K-means nos indica lo bien distanciados que están los centroides uno de otro.

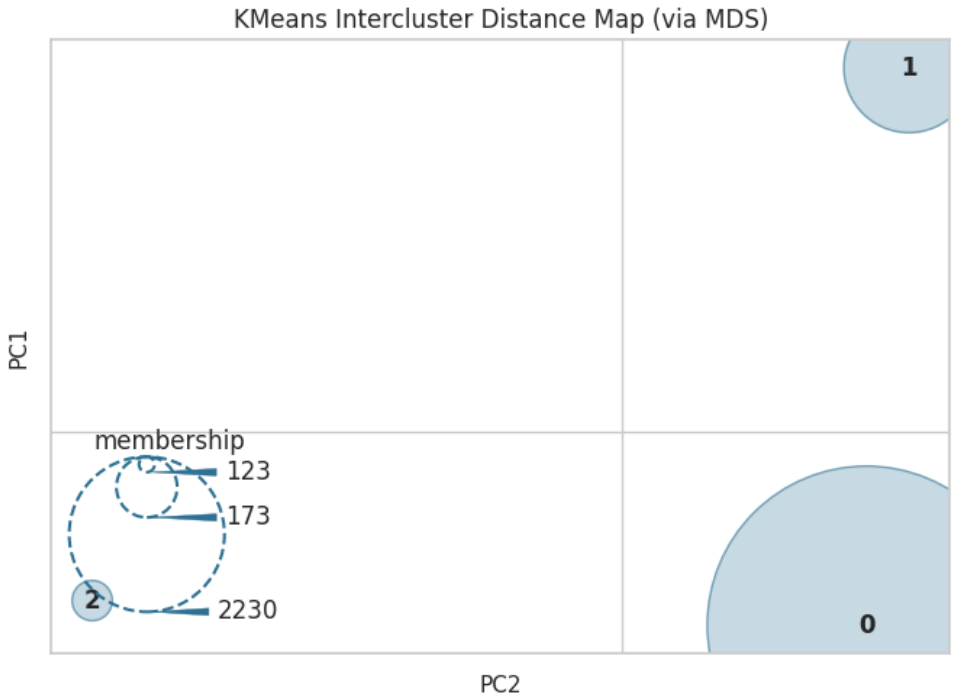


Figura 3.9 Mapa de distancias del modelo K-means

Fuente: Librería Pycaret

El histograma de frecuencia de los grupos nos proporciona una visualización de la cantidad de cajeros asignado a cada grupo.

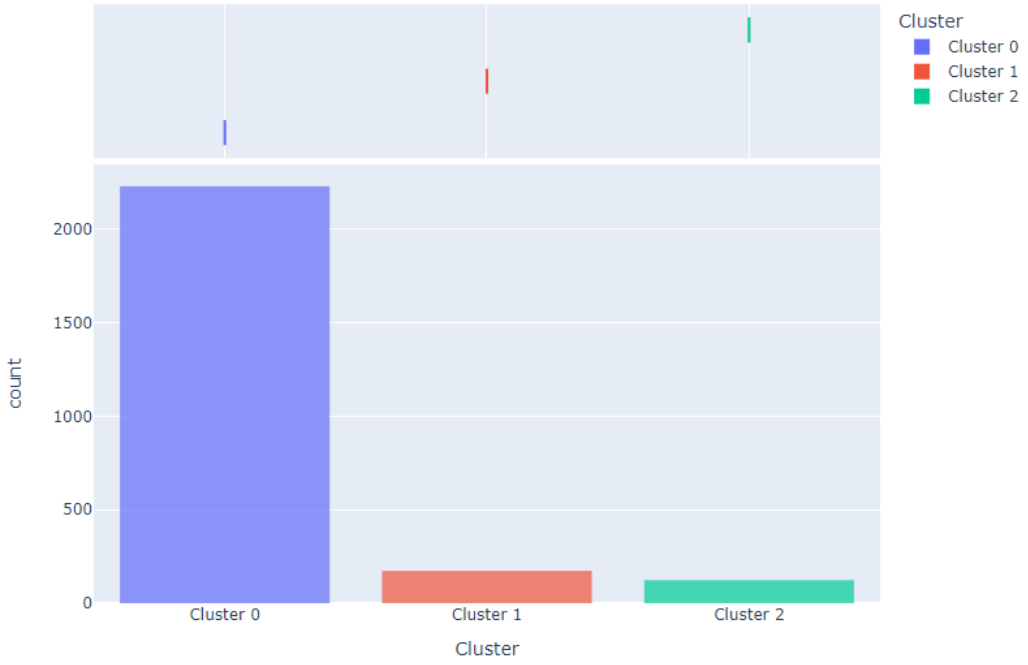


Figura 3.10 Histograma de frecuencias de agrupación del modelo K-means

Fuente: Librería Pycaret

Visualización 2D uMAP de las componentes principales de los tres grupos identificados del K-means.

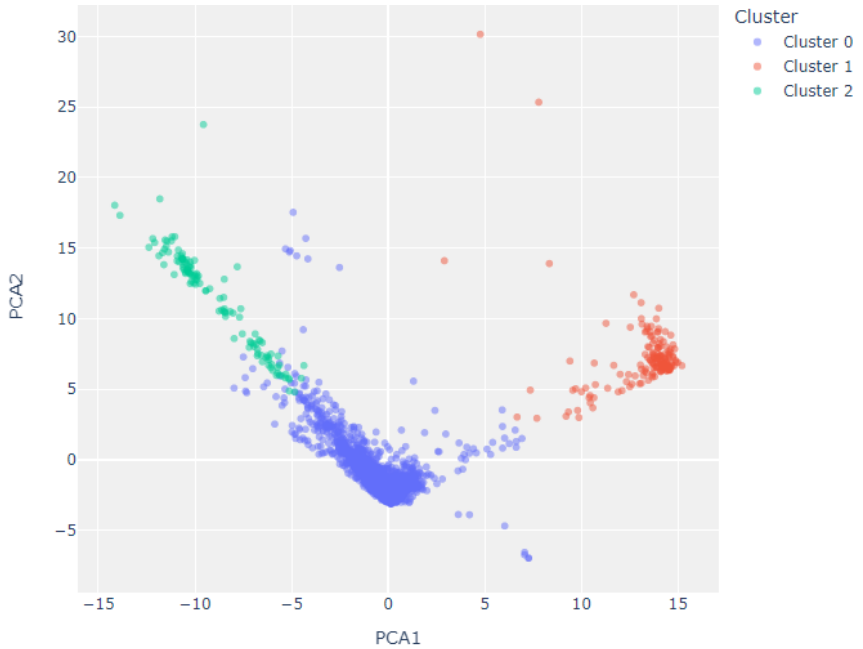


Figura 3.11 Visualización 2D de los grupos identificados del modelo K-means

Fuente: Librería Pycaret

Visualización 3D t-SNE de las componentes principales de los tres grupos identificados del K-means.

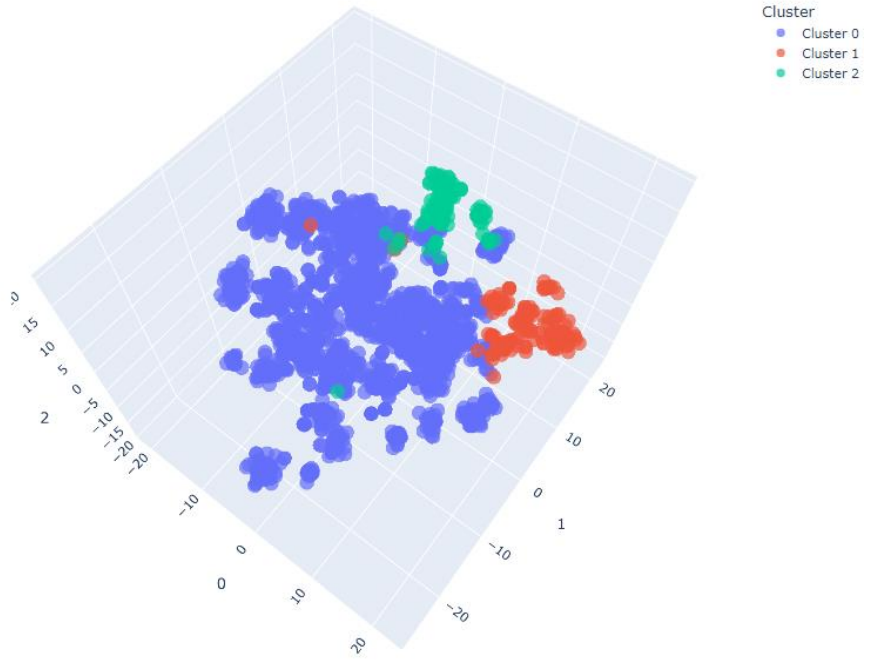


Figura 3.12 Visualización 3D de los grupos identificados del modelo K-means

Fuente: Librería Pycaret

El modelo K-means definió los siguientes grupos:

- **Grupo 0** (2.330 cajeros normales): Este grupo está formado por la mayoría de los cajeros automáticos que se consideran "normales" debido a la cercanía con el origen y a características específicas, es decir, baja tasa de fallos y tiempos de inactividad.

- **Grupo 1** (173 cajeros problemáticos): Este grupo está formado por cajeros automáticos considerados como anómalos o problemáticos debido a que está muy alejado del origen y de los demás grupos. Estos cajeros tienen características específicas que los hacen diferentes de los otros grupos, como la alta tasa de fallos y tiempos de inactividad.

- **Grupo 2** (123 cajeros regulares): Este grupo está formado por cajeros automáticos que no están tan cerca del origen como los del Grupo 0 pero tampoco tan alejados para ser considerados peores que los del Grupo 1.

Tienen características que los hacen distintos de los otros dos grupos, como la tasa de fallos y tiempos de inactividad intermedios.

A continuación, se presenta el valor con mayor frecuencia de las variables categóricas, promedio y desviación estándar de las variables numéricas de los grupos encontrados en el modelo K-means.

Tabla 3.5 Resultados del entrenamiento del modelo K-means

Grupo	Categoría	Cantidad de cajeros
0	Normales	2230
1	Problemáticos	173
2	Regulares	123

Grupo	Moda de Cliente	Moda de Modelo	Moda de Función	Moda de Familia	Moda de Tipo	Moda de País
0	ECU 1 (661)	SND 045 A (423)	Recycler (1465)	Gen 2020 (839)	Branch (1296)	United States (1068)
1	COL 2 (106)	ADA 225 (99)	Cash Dispenser (129)	Gen 2000 (140)	Branch (116)	Colombia (173)
2	ECU 1 (92)	XGN 0573 (83)	Recycler (122)	Gen 2010 (93)	Office (87)	Ecuador (93)

Grupo	Lector de tarjetas Promedio Inactividad (s)	Dispensador de efectivo Promedio Inactividad (s)	Depósito de efectivo Promedio Inactividad (s)	Cheque Promedio Inactividad (s)	Teclado Electrónico Promedio Inactividad (s)	Impresora Promedio Inactividad (s)
0	1791.83	626.75	3194.54	1480.32	343.83	1070.33
1	2920.69	1085.98	2406.91	463.64	81076.7	1788.37
2	5600.19	144.91	64284.09	3330.5	128.77	584.41

Grupo	Lector de tarjetas Desv. Est Inactividad (s)	Dispensador de efectivo Desv. Est Inactividad (s)	Depósito de efectivo Desv. Est Inactividad (s)	Cheque Desv. Est Inactividad (s)	Teclado Electrónico Desv. Est Inactividad (s)	Impresora Desv. Est Inactividad (s)
0	7648.83	5238.44	9292.35	7755.15	4754.44	6377.9
1	12150.64	5194.25	6505.32	5843.51	20098.9	7893.26
2	12430.46	1318.12	26032.15	10110.54	1705.15	2612.63

3.2.2 Isolation Forest

Para hacer uso del método Isolation Forest no es necesario normalizar los datos, lo que es una ventaja ya que no se ve afectado negativamente por la presencia de valores atípicos que es lo que este proyecto desea analizar. En cuanto a los hiperparámetros del modelo se configuró “fraction” en 0.05 y “n_estimators” en 200 sus valores por defecto.

- **fraction**: Especifica la fracción de valores atípicos en el conjunto de datos. Es decir, representa la proporción esperada de valores anómalos en el conjunto de datos. Un valor típico de "fraction" puede estar en el rango de 0 a 1, donde 0 indica que no se esperan valores atípicos y 1 significa que se espera que todos los valores sean atípicos.

- **n_estimators**: Especifica el número de árboles (estimadores) que se construirán en el modelo. Cuantos más árboles se construyan, mayor será la precisión del modelo, pero también se incrementará el tiempo de procesamiento.

Pipeline del modelo

Figura 3.13 Diagrama de flujo de creación del modelo Isolation Forest

Fuente: Librería Pycaret

El pipeline del Isolation Forest es similar al del K-means solo que no se procesa agrupación de categorías ni normalización z-score.

1. El conjunto de datos es sometido a imputación numérica, si no se encontrase algún valor numérico se imputa con la media de los datos existentes.
2. El conjunto de datos es sometido a imputación categórica, si no se encontrase algún valor categórico se imputa con la moda (valor de mayor frecuencia) de los datos existentes.
3. El conjunto de datos es sometido a la codificación One-hot, una técnica en donde los datos categóricos son convertidos a datos binarios, es decir de ceros y unos, paso fundamental para entrenar el modelo con este tipo de características.

El conjunto de datos de entrenamiento antes de la ejecución del pipeline del modelo comprendía (2526 filas, 78 columnas), al finalizar el pipeline quedó conformado por (2526 filas, 134 columnas) de las cuales:

- 72 columnas resultantes corresponden a las 12 semanas de cada uno de los 6 módulos del cajero.
- 62 columnas resultantes corresponden a las 7 columnas categóricas luego de la codificación One-hot.

Luego del entrenamiento del modelo se obtiene los siguientes resultados:

Visualización 3D t-SNE de las anomalías detectadas por el modelo con el conjunto de entrenamiento.

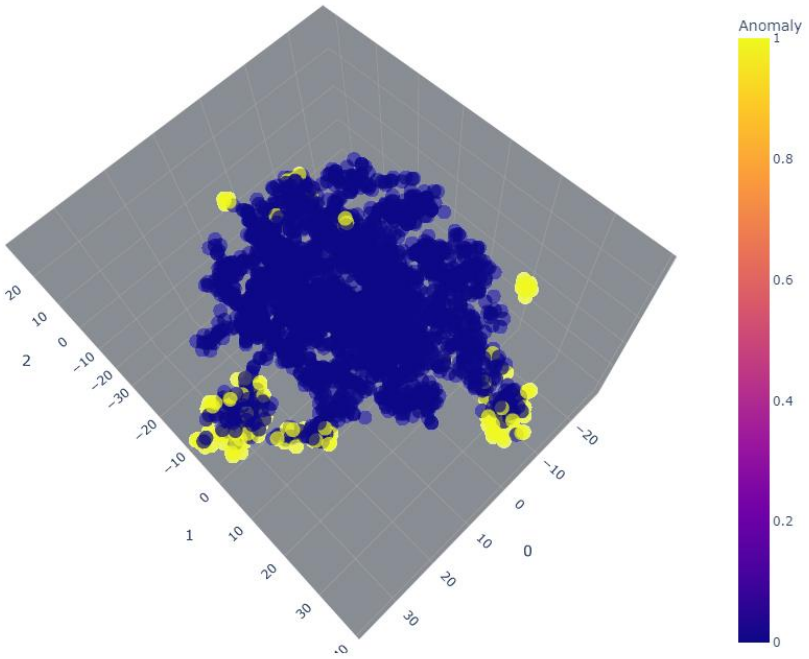


Figura 3.14 Visualización 3D de las anomalías identificadas en el entrenamiento (Isolation Forest)

Fuente: Librería Pycaret

Visualización 2D uMAP de las anomalías detectadas por el modelo con el conjunto de entrenamiento.

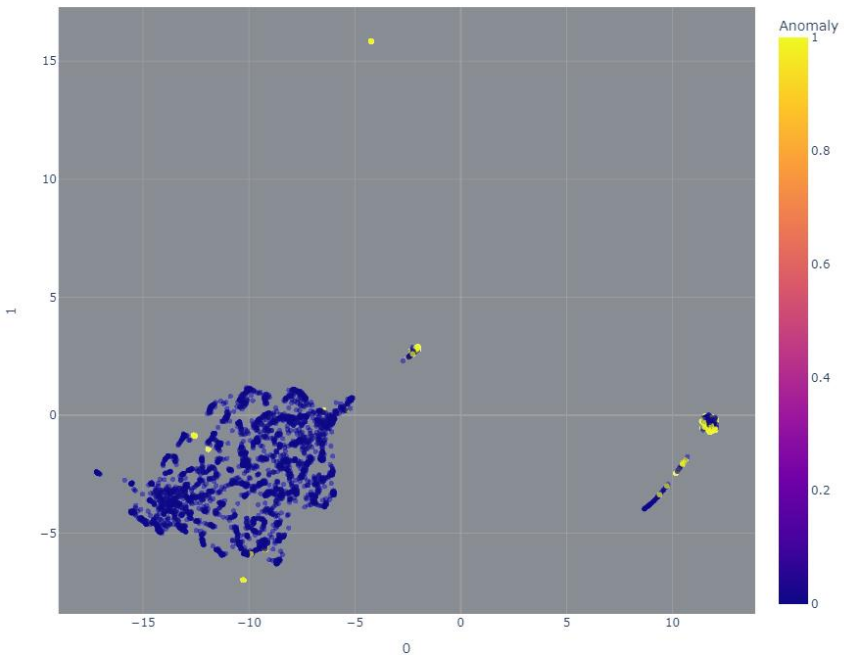


Figura 3.15 Visualización 2D de las anomalías identificadas en el entrenamiento (Isolation Forest)

Fuente: Librería Pycaret

A continuación, se presenta una tabla con información del valor con mayor frecuencia de las variables categóricas (moda), promedio y desviación estándar de las variables numéricas de los grupos encontrados como anómalos y no anómalos en el modelo Isolation Forest.

Tabla 3.6 Resultados del entrenamiento del modelo Isolation Forest

Grupo	Categoría	Cantidad de cajeros	Puntuación Promedio
No anomalía	Normales	2399	-0.119545
Anomalía	Problemáticos	127	0.034801

Grupo	Moda de Cliente	Moda de Modelo	Moda de Función	Moda de Familia	Moda de Tipo	Moda de País
No anomalía	USA 1 (569)	SND 045 A (416)	Recycler (1524)	Gen 2020 (828)	Branch (1362)	United States (1052)
Anomalía	ECU 1 (36)	ADA 228 (21)	Recycler (107)	Gen 2000 (74)	Branch (51)	Colombia (49)

Grupo	Lector de tarjetas Promedio Inactividad (s)	Dispensador de efectivo Promedio Inactividad (s)	Depósito de efectivo Promedio Inactividad (s)	Cheque Promedio Inactividad (s)	Teclado Electrónico Promedio Inactividad (s)	Impresora Promedio Inactividad (s)
No anomalía	1498.66	429.43	4873.99	1146.77	4596.9	829.28
Anomalía	12555.82	4512.96	29562.57	8187.99	29770.77	6131.18

Grupo	Lector de tarjetas Desv. Est Inactividad (s)	Dispensador de efectivo Desv. Est Inactividad (s)	Depósito de efectivo Desv. Est Inactividad (s)	Cheque Desv. Est Inactividad (s)	Teclado Electrónico Desv. Est Inactividad (s)	Impresora Desv. Est Inactividad (s)
No anomalía	5778.36	3344.88	14141.89	6494.92	19189.13	4745.69
Anomalía	25413.25	17139.74	36324.8	19038.94	40821.74	18833.75

3.2.3 Análisis comparativo de los resultados

A continuación, una comparación de los grupos del K-means contra los grupos detectados del Isolation Forest.

Tabla 3.7 Comparación de los resultados del modelo K-means e Isolation Forest

Grupo (K-means)	Anomalía (Isolation Forest)	Cantidad	Puntuación de anomalía promedio
Grupo 0	No	2194	-0.127471
	Si	**36**	**0.039764**
Grupo 1	No	127	-0.03492
	Si	**46**	**0.03202**
Grupo 2	No	78	-0.034382
	Si	**45**	**0.033672**

Conocidos los resultados del modelo K-means donde los cajeros se agrupan por características y patrones similares, el grupo 1 denota una tasa alta de tiempo de inactividad por error en el módulo teclado electrónico, mientras que el grupo 2 denota una tasa alta de tiempos de inactividad por error en el módulo de dispensador de efectivo, desde la perspectiva del experto del Centro de Atención al Cliente, ambos grupos son problemáticos debido a los altos tiempos de inactividad por error indistinto del módulo que lo genere, por lo que, ambos deberían ser considerados para mantenimiento preventivo. Por tanto, el modelo K-means no es efectivo para detección de anomalías en cajeros automáticos por tiempos de inactividad de acuerdo con lo antes mencionado

De acuerdo con la tabla 3.5 del total de 127 anomalías detectadas por el modelo Isolation Forest solo 46 corresponden a lo que el modelo K-means detectó como cajeros problemáticos del grupo 1, del grupo 0 de cajeros normales detectó 36, mientras que del grupo 2 de cajeros regulares detectó 45.

Mientras que con los resultados del Isolation Forest, el grupo no anómalos congrega aquellos cajeros automáticos con baja tasa de inactividad sin distinguir de los módulos, mientras que los del grupo anómalos fueron aislados debido a su alta tasa de inactividad en cualquiera de sus módulos lo que lo hace un modelo de óptimo de detección de anomalías, por lo que se decide usar el modelo Isolation Forest por sobre el modelo K-means para el desarrollo de este proyecto.

3.2.4 Evaluación del modelo

Para la evaluación del modelo Isolation Forest seleccionado se utilizará la última semana de agosto del 2023 del conjunto de datos de entrenamiento para evaluar la detección de cajeros anómalos obteniendo los siguientes resultados.

Visualización 3D t-SNE de las anomalías detectadas por el modelo con el conjunto de pruebas.

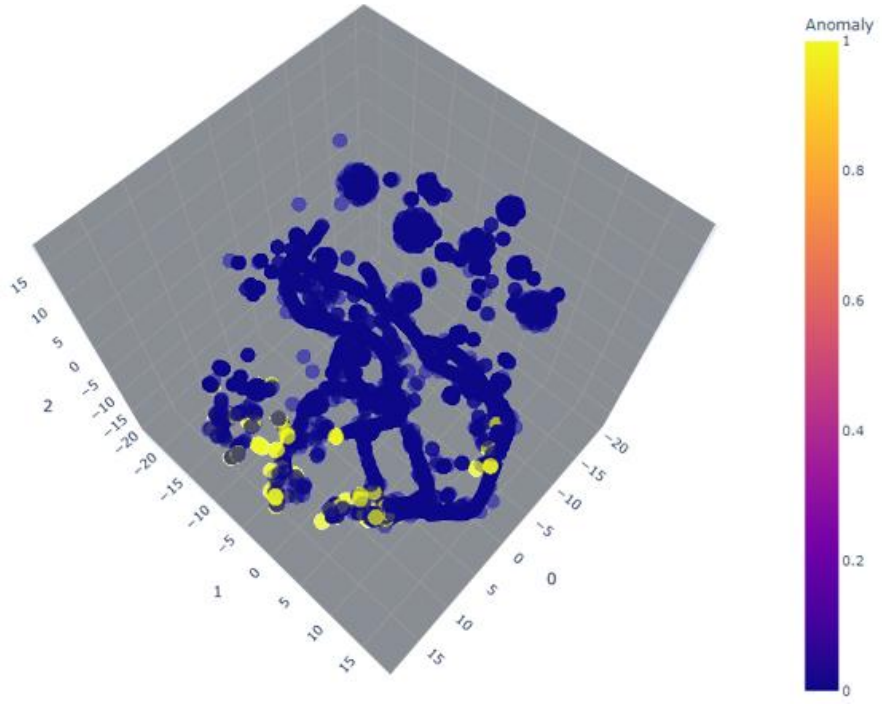


Figura 3.16 Visualización 3D de las anomalías identificadas en la evaluación (Isolation Forest)

Fuente: Librería Pycaret

Visualización 2D uMAP de las anomalías detectadas por el modelo con el conjunto de pruebas.

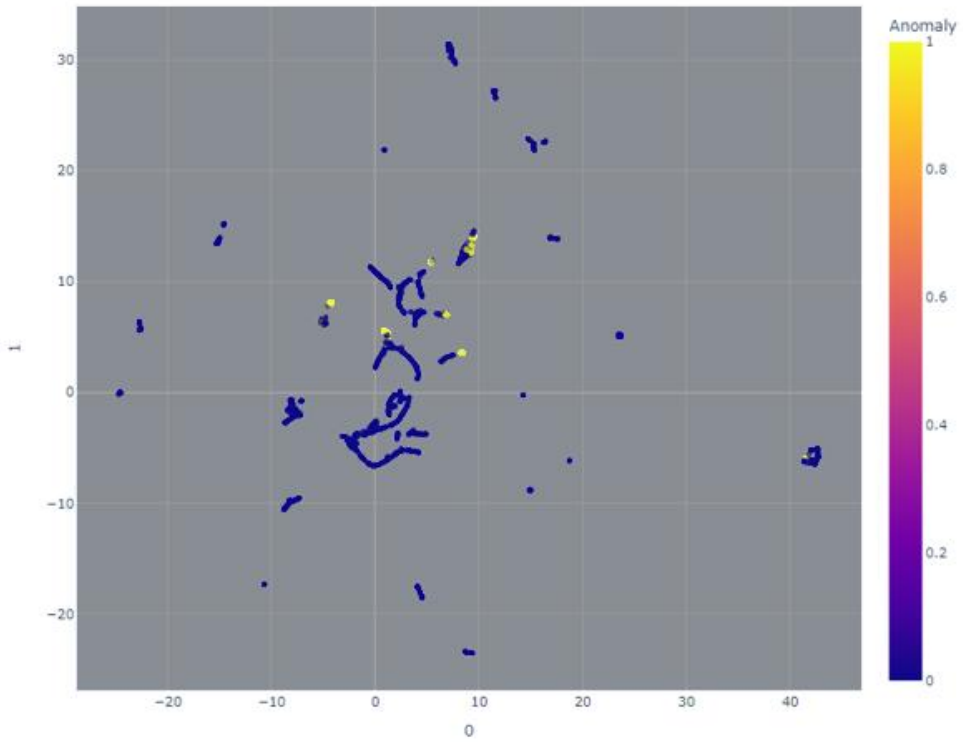


Figura 3.17 Visualización 2D de las anomalías identificadas en la evaluación (Isolation Forest)

Fuente: Librería Pycaret

Resultados estadísticos de la evaluación del modelo donde se incluye moda, promedio y desviación estándar:

Tabla 3.8 Métricas descriptivas de la evaluación del modelo Isolation Forest

Grupo	Categoría	Cantidad de cajeros	Puntuación Promedio
No anomalía	Normales	2326	-0.111053
Anomalía	Problemáticos	215	0.050181

Grupo	Moda de Cliente	Moda de Modelo	Moda de Función	Moda de Familia	Moda de Tipo	Moda de País
No anomalía	ECU 1 (688)	SND 045 A (403)	Recycler (1480)	Gen 2020 (808)	Branch (1332)	United States (1039)
Anomalía	ECU 1 (71)	ADA 225 (38)	Recycler (160)	Gen 2000 (104)	Branch (90)	Colombia (76)

Grupo	Lector de tarjetas Promedio Inactividad (s)	Dispensador de efectivo Promedio Inactividad (s)	Depósito de efectivo Promedio Inactividad (s)	Cheque Promedio Inactividad (s)	Teclado Electrónico Promedio Inactividad (s)	Impresora Promedio Inactividad (s)
No anomalía	222.7	4488.65	1297.04	852.95	4899.16	444.26
Anomalía	6220.96	29533.98	15969.39	9309.52	27174.54	8421.82

Grupo	Lector de tarjetas Desv. Est Inactividad (s)	Dispensador de efectivo Desv. Est Inactividad (s)	Depósito de efectivo Desv. Est Inactividad (s)	Cheque Desv. Est Inactividad (s)	Teclado Electrónico Desv. Est Inactividad (s)	Impresora Desv. Est Inactividad (s)
No anomalía	1480.22	13936.82	4496.67	5224.98	19847.1	2657.46
Anomalía	17168.33	33041.87	24950.2	21184.79	39885.14	18697.44

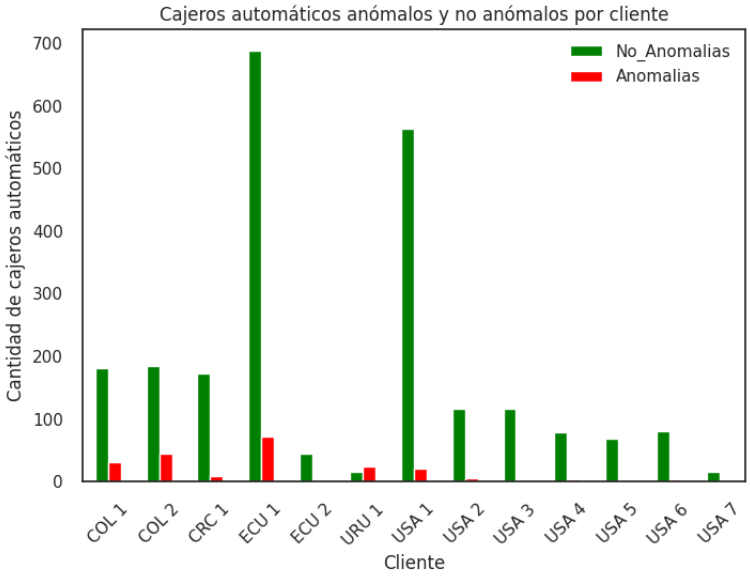


Figura 3.18 Gráfico de barras de la cantidad de cajeros anómalos y no anómalos por cliente

Fuente: Librería Pycaret

Tabla 3.9 Detalle de la cantidad de cajeros anómalos y no anómalos por cliente

Cliente	Cajeros No Anómalos	Cajeros Anómalos	Cajeros Anómalos (%)
COL 1	181	31	14.62
COL 2	185	45	19.57
CRC 1	173	8	4.42
ECU 1	688	71	9.35
ECU 2	44	0	0
URU 1	16	24	60
USA 1	563	21	3.6
USA 2	116	6	4.92
USA 3	116	1	0.85
USA 4	79	3	3.66
USA 5	69	1	1.43
USA 6	81	3	3.57
USA 7	15	1	6.25

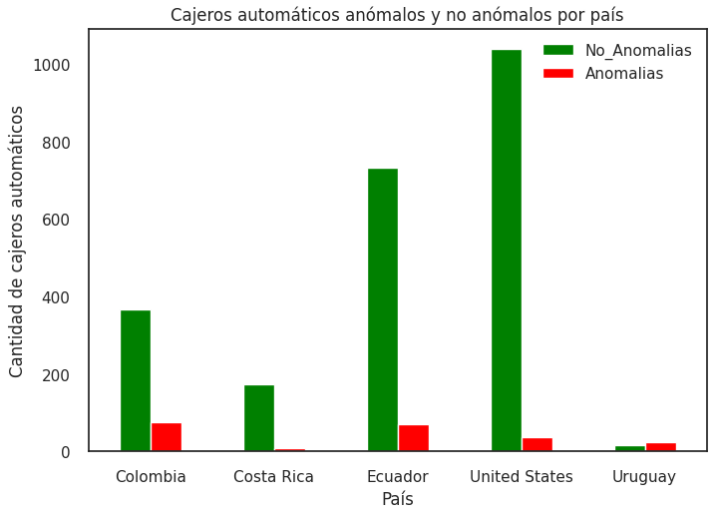


Figura 3.19 Gráfico de barras de la cantidad de cajeros anómalos y no anómalos por país

Fuente: Librería Pycaret

Tabla 3.10 Detalle de la cantidad de cajeros anómalos y no anómalos por país

País	Cajeros No Anómalos	Cajeros Anómalos	Cajeros Anómalos (%)
Colombia	366	76	17.19
Costa Rica	173	8	4.42
Ecuador	732	71	8.84
United States	1039	36	3.35
Uruguay	16	24	60

3.3 Infraestructura para proceso y almacenamiento

3.3.1 Almacenamiento de datos en la nube

El conjunto de datos de entrenamiento y pruebas será almacenado en Google Cloud Storage, un servicio que permite a las empresas y desarrolladores almacenar y gestionar datos no estructurados en la nube de forma que sea escalable, segura y siempre accesible. Google replica automáticamente los datos en múltiples ubicaciones geográficas, lo que garantiza su durabilidad y disponibilidad, además de ofrecer opciones de seguridad avanzadas, como el control de acceso basado en políticas y la encriptación de datos.

3.3.2 Procesamiento de datos

Se hará uso de Google Colaboratory, o Colab, una versión como servicio de Jupyter Notebook que permite escribir y ejecutar código Python a través del navegador. Jupyter Notebook es una creación gratuita y de código abierto del Proyecto Jupyter.

Es un cuaderno de laboratorio interactivo que incluye no sólo notas y datos, sino también código que puede manipularlos. El código puede ejecutarse dentro del cuaderno, que, a su vez, puede capturar la salida del código.

3.4 Plataformas y prototipos de visualización

Para visualizar los resultados se hará uso de la herramienta StreamLit, una biblioteca de Python que facilita la creación de aplicaciones web interactivas para mostrar los resultados del modelo.

La detección de anomalías es un proceso que se realiza bajo demanda en el momento que el usuario accede al módulo de detección en la aplicación de monitoreo, la evaluación se realiza con el conjunto de datos registrados de la última semana para cada cajero automático.

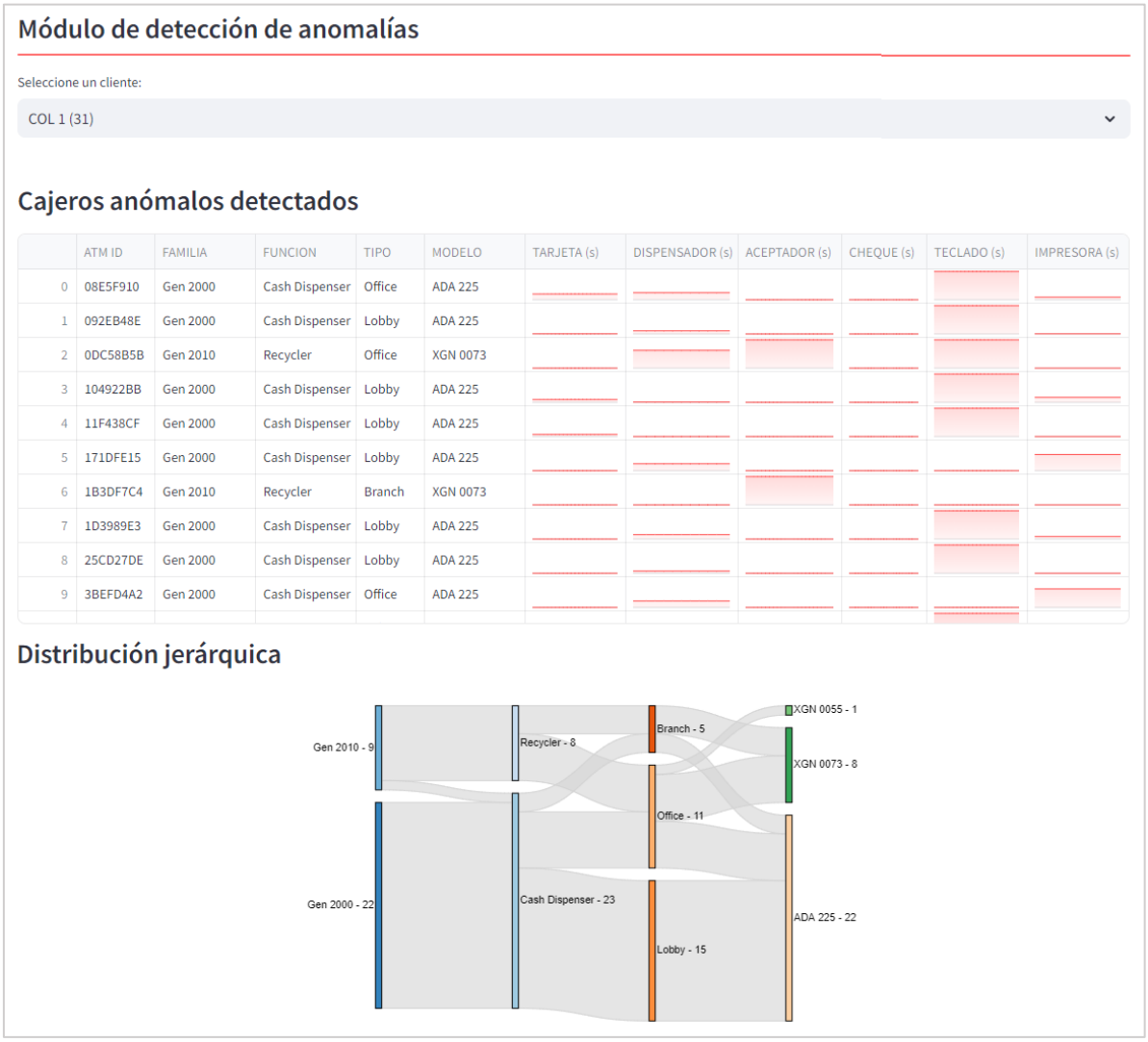


Figura 3.20 Prototipo de visualización de resultados usando StreamLit

Fuente: Elaboración Propia

CAPÍTULO 4

4. ANÁLISIS DE RESULTADOS

4.1 Estrategias para validación de proyecto

4.1.1 Análisis de las puntuaciones de anomalías

Dado el conjunto de datos de entrenamiento se verificó las puntuaciones de anomalía asignadas a cada punto del conjunto de datos. Se comparó las puntuaciones de los datos que fueron consideradas como no anomalías con las de los datos que fueron consideras como anomalías.

Para todo el conjunto de datos de entrenamiento los resultados promedio de puntuación de anomalías es:

Tabla 4.1 Detalle de la cantidad de cajeros anómalos y no anómalos

Anomalía	Cantidad de cajeros	Puntuación de Anomalía
No	2326	-0.111053
Si	215	0.050182

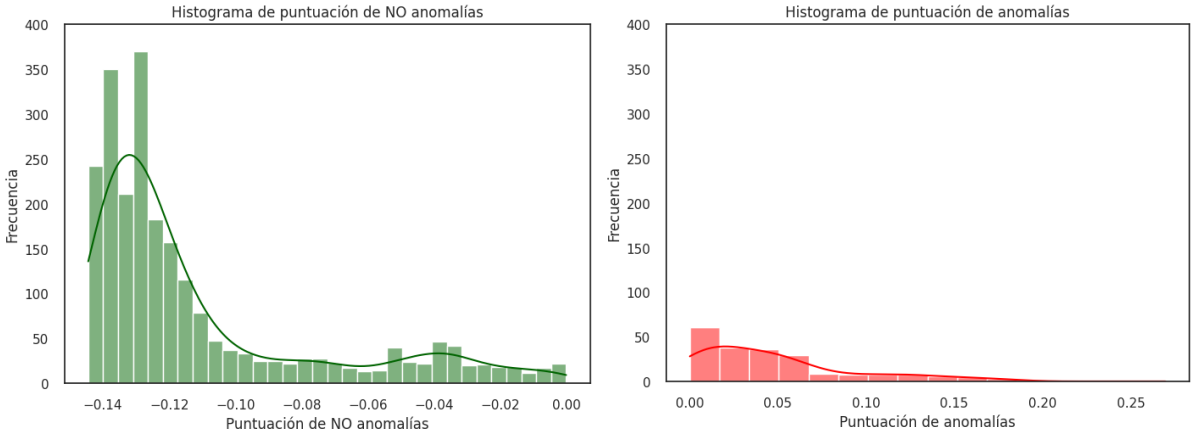


Figura 4.1 Histograma de frecuencias de la puntuación de anomalías y no anomalías

Fuente: Elaboración Propia

Todas las no anomalías promedian un valor menor a cero, mientras que las anomalías promedian un valor mayor a cero. Si se desglosa el promedio de puntuaciones por clientes se obtiene:

Tabla 4.2 Detalle de la puntuación de cajeros anómalos y no anómalos por cliente

Cliente	Cantidad de cajeros	Puntuación de anomalía promedio	
		No	Si
COL 1	31	-0.094821	0.0714
COL 2	45	-0.086259	0.059677
CRC 1	8	-0.110963	0.049766
ECU 1	71	-0.103542	0.04418
ECU 2	0	-0.116282	-
URU 1	24	-0.059062	0.021755
USA 1	21	-0.128547	0.055717
USA 2	6	-0.125417	0.05565
USA 3	1	-0.123064	0.006746
USA 4	3	-0.102682	0.051951
USA 5	1	-0.117536	0.026155
USA 6	3	-0.118217	0.039503
USA 7	1	-0.113451	0.02191

El promedio por cliente indica que todas las anomalías detectadas poseen una puntuación mayor a cero mientras que las no anomalías tienen una puntuación menor a cero, con lo que se validan los resultados obtenidos.

El resultado de anomalías de cajeros automáticos sería siempre visible desde el módulo de anomalías de la aplicación de monitoreo a la que el cliente tiene acceso y diseñado mediante el prototipo realizado en este proyecto.

Es decisión del cliente realizar mantener mantenimiento preventivo a los cajeros que el prototipo y modelo ha detectado como anómalos para que los tiempos de inactividad por error disminuyan a mediano o largo plazo y, por ende, las llamadas de atención al CAC de estos equipos decrezcan.

4.1.2 Visualización de resultados

En el prototipo del módulo de detección de anomalías se muestra los resultados del modelo, el usuario observa un listado de los cajeros detectados con todas las variables cualitativas y cuantitativas mediante gráficos de línea en miniatura que representa los tiempos de inactividad promediados por la semana del análisis para cada dispositivo.

Adicional, se muestra un diagrama de sankey para que el usuario tenga una visión más amplia de la distribución jerárquica basados en las variables categóricas de los cajeros detectados como anómalos.

Mediante estos resultados el cliente tendrá la noción de los cajeros anómalos en su flota y deberá tomar las acciones necesarias mediante mantenimiento preventivo para contrarrestar los tiempos de inactividad del cajero y disminuir las llamadas de atención.

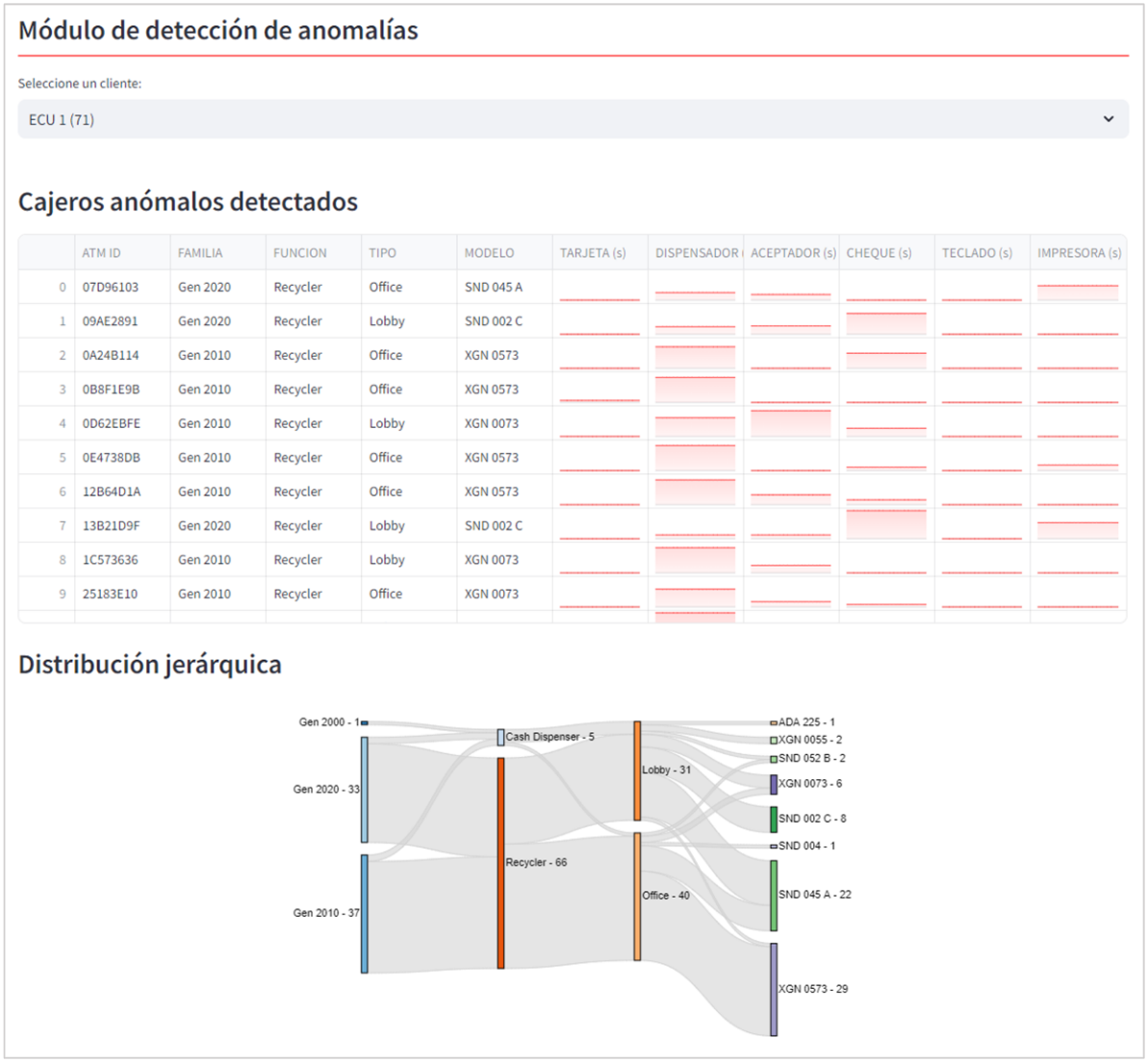


Figura 4.2 Prototipo del módulo de detección de anomalías

Fuente: Elaboración Propia

4.2 Puesta en marcha y funcionamiento

El desarrollo del proyecto se realiza en un cuaderno Jupyter con ayuda de Google Colab, donde se realiza todo el preprocesamiento de los datos. La creación y entrenamiento de los modelos se realizan con ayuda de la librería PyCaret, la muestra de resultados y visualizaciones se realiza mediante la librería StreamLit, los conjuntos de datos generados y modelos creados son almacenados en la plataforma Google Cloud Storage para luego ser leída por el prototipo.

Toda la implementación anterior es convertida a un script Python que es almacenada en GitHub, una plataforma en línea que se utiliza para alojar, gestionar y colaborar en proyectos de desarrollo de software para finalmente ser desplegada en Heroku, una plataforma para administrar aplicaciones web y servicios en línea lista para ser desplegada y usada por el usuario final sin necesidad preocuparse por la gestión de la infraestructura.

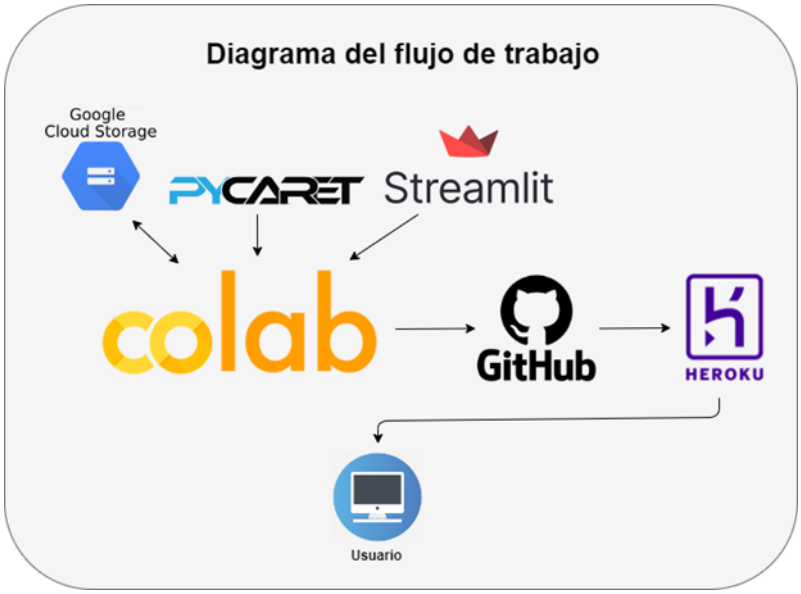


Figura 4.3 Flujo de trabajo del proyecto

Fuente: Elaboración Propia

4.3 Pruebas de funcionalidad

4.3.1 Pruebas de detección de anomalías

Se verifica la capacidad del modelo para identificar anomalías con un conjunto de datos alterno que contengan anomalías conocidas y si el modelo de los detecta adecuadamente.

Antes de realizar la prueba evaluamos el modelo nuevamente solo con el resultado de los cajeros no anómalos (2.326) probado anteriormente, del cual se obtuvo el siguiente resultado:

Tabla 4.3 Resultado de la reevaluación del modelo sin anomalías

Anomalía	Cantidad de cajeros	Puntuación de Anomalía
No	2326	-0.111053
Si	0	-

Acto seguido, de todos los cajeros no anómalos (2.526) se seleccionó explícitamente los 5 últimos para elevar arbitrariamente sus tiempos de inactividad en todos sus módulos y verificar si son detectados como anomalías.

Tabla 4.4 Detalle del conjunto de datos sin anomalías modificados

No.	ATM ID	LECTORA (S35)	DISPENSADOR (S35)	ACEPTADOR (S35)	CHEQUE (S35)	TECLADO (S35)	IMPRESORA (S35)
1	00385E3F	0	0	0	0	0	0
...	...	...	...	...	...	...	...
2517	FCBAD4B5	8.714286	0	0	0	0	0
2518	39530DD1	0	781.28	0	0	0	0
2519	FE8174E8	0	0	0	0	0	0
2520	020F00FF	0	0	0	0	0	18678.71
2521	FF2764A7	0	0	0	0	0	0
2522	FF27A3DF	50000	50000	50000	50000	50000	50000
2523	FF3F07B6	50000	50000	50000	50000	50000	50000
2524	FF45AEEC	50000	50000	50000	50000	50000	50000
2525	FFB9EF19	50000	50000	50000	50000	50000	50000
2526	FFF10B72	50000	50000	50000	50000	50000	50000

CLIENTE	PAIS	FUNCION	FAMILIA	TIPO	MODELO
USA 1	United States	Cash Dispenser	Gen 2000	Branch	ADA 005 A
...	...	...	...	...	...
USA 1	United States	Recycler	Gen 2010	Branch	XGN 0877
CRC 1	Costa Rica	Cash Dispenser	Gen 2000	Branch	ADA 225
ECU 1	Ecuador	Recycler	Gen 2020	Office	SND 045 A
USA 2	United States	Recycler	Gen 2000	Branch	ADA 047
ECU 1	Ecuador	Recycler	Gen 2020	Office	SND 045 A
CRC 1	Costa Rica	Cash Dispenser	Gen 2000	Branch	ADA 225
USA 4	United States	Cash Dispenser	Gen 2000	Branch	ADA 225
ECU 1	Ecuador	Recycler	Gen 2020	Office	SND 045 A
ECU 1	Ecuador	Recycler	Gen 2010	Office	XGN 0573
COL 1	Colombia	Cash Dispenser	Gen 2000	Lobby	ADA 225

Luego de la evaluación del modelo con este conjunto de datos modificado se obtiene el siguiente resultado:

Tabla 4.5 Resultado de la reevaluación del modelo con datos modificados

Anomalía	Cantidad de cajeros	Puntuación de Anomalía
No	2321	-0.111091
Si	5	0.317716

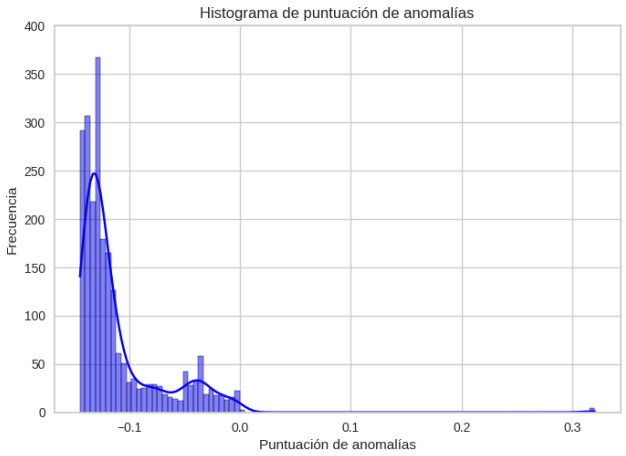


Figura 4.4 Histograma de frecuencias de puntuación del conjunto de datos modificado

Fuente: Elaboración Propia

Los cinco cajeros anómalos detectados se grafican en el histograma de puntuación con un valor aproximado a 0.31.

4.3.2 Pruebas de visualización

Se verifica que los resultados gráficos del modelo mantengan coherencia con los valores del conjunto de datos y del modelo resultante utilizado.

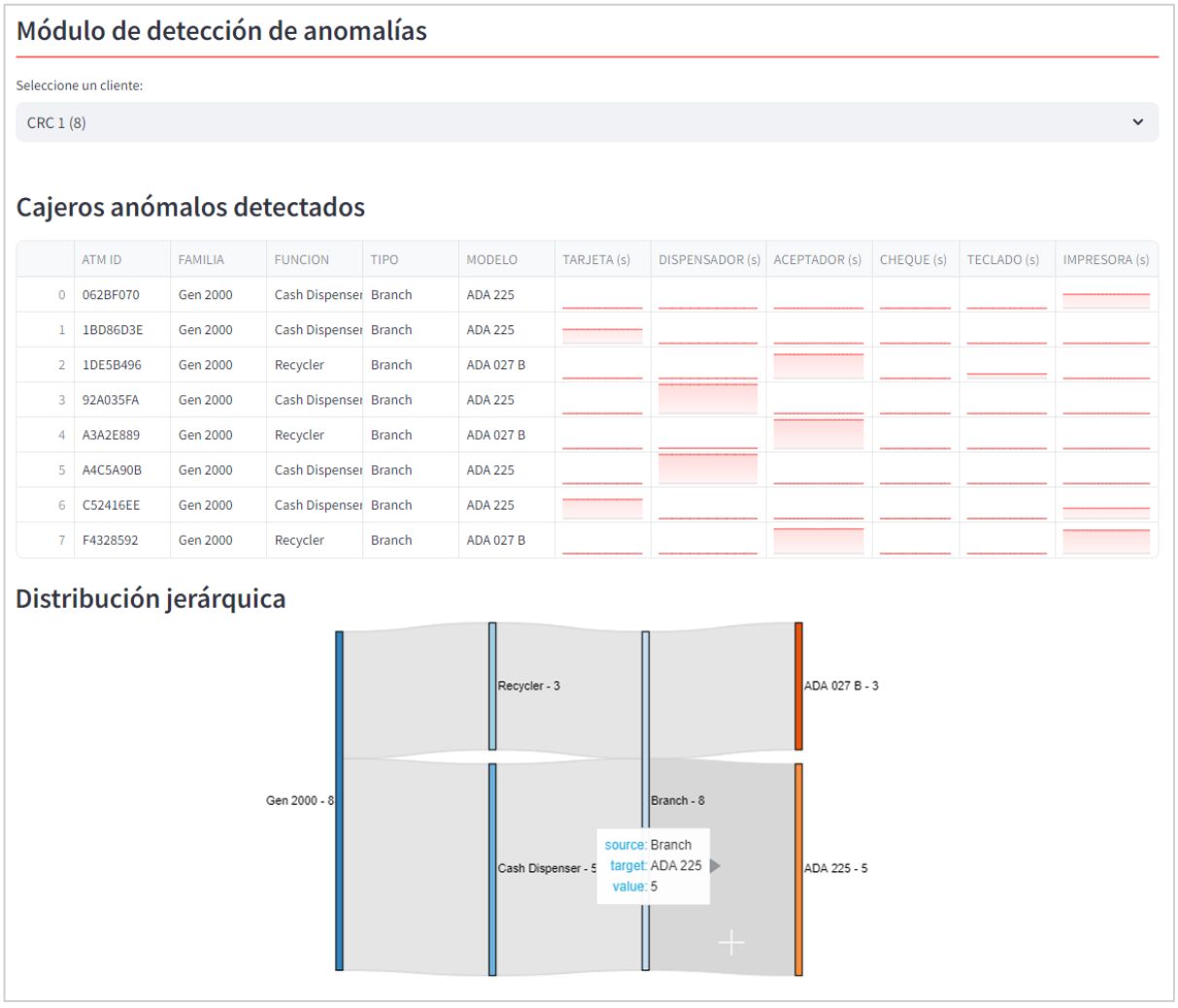


Figura 4.5 Visualización de resultados en el prototipo de anomalías

Fuente: Elaboración Propia

De las 8 anomalías detectadas para el cliente “CRC 1” por el modelo Isolation Forest y graficadas en el diagrama sankey tenemos:

- De acuerdo con su familia: todos son cajeros de “Gen 2000”.
- De acuerdo con su función: cinco son “Cash Dispenser” y tres son “Recycler”.
- De acuerdo con su tipo: todos son “Branch”.

- De acuerdo con su modelo: cinco son “ADA 225” y tres son “ADA 027 B”.

En definitiva, la visualización concuerda con la tabla resultado del modelo mostrado anteriormente.

4.3.3 Pruebas de privacidad

Se verifica la anonimización de los datos para no afectar la privacidad e integridad del cliente y sus equipos.

Para el nombre del cliente simplemente se lo reemplaza por un valor genérico, mientras que para los ID de los cajeros se hace uso del algoritmo hash no criptográfico “MurmurHash” conocido por su velocidad y calidad de distribución, lo que lo hace popular para tareas como la indexación de tablas hash [25].

Cajeros anómalos detectados

	ATM ID	FAMILIA	FUNCION	TIPO
0	11280BFE	Gen 2000	Recycler	Branch
1	2ED0BB7A	Gen 2010	Recycler	Branch
2	33795B99	Gen 2020	Recycler	Branch
3	42902879	Gen 2000	Recycler	Branch
4	4ADE2B44	Gen 2000	Recycler	Branch
5	5FBEAF1D	Gen 2000	Recycler	Branch
6	63343DFB	Gen 2000	Recycler	Branch
7	7042F1F9	Gen 2010	Recycler	Branch
8	7296D54A	Gen 2010	Recycler	Branch
9	80BE8762	Gen 2020	Recycler	Branch

Figura 4.6 Resultado de la función MurmurHash

Fuente: Elaboración Propia

El proceso se realizó en R durante el preprocesamiento del conjunto de datos mediante la siguiente función.

```
hash_string <- function(text) {
  hash <- digest(text, algo = "murmur32", serialize = FALSE)
  return(hash)
}
data10$ID <- sapply(data10$ID, hash_string)
```

Figura 4.7 Código en R de la función MurmurHash

Fuente: Elaboración Propia

4.4 Análisis costo - beneficio

El análisis de costo - beneficio de un modelo de detección de anomalías debido a tiempos de inactividad en cajeros automáticos implica evaluar los costos asociados con la implementación del modelo en comparación con los beneficios. Para este proyecto se utilizan herramientas de terceros para realizar el prototipo de lo que sería el módulo de detección de anomalías del software propio de monitoreo.

Si la compañía decide realizar este proyecto lo haría con sus propios recursos humanos, por lo que no tendría un costo asociado a este proyecto. Lo mismo sucede con la infraestructura necesaria como servidores, hosting de aplicaciones, por lo que tampoco un costo asociado y lo mismo para la obtención de datos ya que posee herramientas propias implementadas para conseguirlos.

Ante lo antes mencionado, el enfoque del beneficio a la empresa se verá reflejado en la disminución de carga operativa y humana del centro de atención al cliente (CAC) dependiendo de si los clientes realizan las recomendaciones de mantenimiento preventivo en el momento que son detectadas, si así el cliente lo considera.

A continuación, un análisis costo - beneficio desde la perspectiva del cliente y su impacto sobre las llamadas de atención o soporte al CAC:

Costo del mantenimiento correctivo en las condiciones actuales.

La siguiente información fue directamente consultada a expertos del área de servicio de la empresa “ABC”:

- Costo del soporte a un cajero
  $60 promedio por hora.

- Horas promedio de soporte a un cajero por un especialista
  3 horas.

- Porcentaje promedio de cajeros con problemas técnicos constantes por error de cualquier tipo de un cliente
  5%.

- Cantidad de veces promedio que un cajero problemático (del 5%) es atendido al año
  10 veces.

- Cantidad de veces promedio que un cajero no problemático (del 95%) es atendido al año
  2 veces.

A continuación, se define un ejemplo del gasto actual en cajeros problemáticos de un cliente “A” suponiendo que posee 100 cajeros en su flota y se encuentran en servicio a disposición de los usuarios.

- Cajeros problemáticos
  5 (100 cajeros * 0.05)

- Cantidad de soportes a los cajeros problemáticos al año
  50 veces (5 cajeros * 10 veces)

- Costo del soporte a los cajeros problemáticos por año
  $9.000 ($60 * 50 veces* 3 horas)

En conclusión, el cliente “A” gasta $9.000 en mantenimiento correctivo a sus 5 cajeros problemáticos al año.

Costo del mantenimiento preventivo con la solución propuesta.

De acuerdo con la Tabla 3.9, el cliente “ECU 1” posee 759 cajeros automáticos en su flota, la cantidad de cajeros problemáticos de acuerdo con los expertos consultados de la empresa “ABC” es aproximadamente 38 (5% del total).

Mediante la solución propuesta el modelo detectó 71 cajeros como anómalos, tomando como base este resultado y considerando que el cliente ha tomado en cuenta la expresa recomendación de realizar mantenimiento preventivo a sus cajeros anómalos en un tiempo prudencial se obtiene los siguientes costos:

Tabla 4.6 Comparación de costos de mantenimiento

Descripción	Solución actual	Solución propuesta
Costo del mantenimiento por hora	$60	$60
Horas promedio de mantenimiento	3	3
Cajeros problemáticos atendidos	38	71
Cantidad de soportes por cajero al año	10	2
Costo total anual del mantenimiento	$68.400	**$25.560**

En definitiva, el ahorro del cliente “ECU 1” sería de $42.840 (62.63%) siempre y cuando el cliente cumpla con las recomendaciones del modelo.

Llamadas de atención con las condiciones actuales.

De acuerdo con la Tabla 3.9, el cliente “ECU 1” posee 759 cajeros automáticos, lo que de acuerdo con el experto de servicios de la empresa “ABC” genera al año:

Tabla 4.7 Detalle del total de llamadas en la condición actual

Descripción	Cajeros	Llamadas promedio anual por cajero	Llamadas totales
Cajeros no anómalos	725	2	1450
Cajeros anómalos	34	10	340
			1.790 llamadas al año

Llamadas de atención con solución propuesta.

Mientras que con la detección apropiada de anomalías y los preventivos mantenimientos realizados a aquellos cajeros se obtiene el siguiente resultado:

Tabla 4.8 Detalle del total de llamadas con la solución propuesta

Descripción	Cajeros	Llamadas promedio anual por cajero	Llamadas totales
Cajeros no anómalos	688	2	1376
Cajeros anómalos	71	2	142
			1.518 llamadas al año

El cliente “ECU 1” en las condiciones actuales genera aproximadamente 1.790 llamadas de atención al año mientras que con la solución propuesta generaría 1.518 llamadas al año, una rebaja de 272 (15.2%) de llamadas al año.

Si de acuerdo con el inciso 1.1, al año todos los clientes del país generan 45.625 llamadas (125 llamadas promedio al día * 365 días), solo con el análisis de este cliente “ECU 1” el total de llamadas disminuiría a 45.353, que en términos porcentuales se reduciría un 0.60% anual (sin haber tomado en cuenta a los demás clientes).

CONCLUSIONES Y RECOMENDACIONES

Entre las conclusiones más importantes de este proyecto se tiene:

- Con ayuda de técnicas de ciencias de datos, aprendizaje de máquina no supervisado y varias herramientas especializadas en este contexto se ha logrado concretar un prototipo de detección de anomalías que se convierte en una gran utilidad para que los distintos clientes puedan consultar si existen cajeros automáticos con altos tiempos de inactividad por error que puedan incurrir a que salgan de servicio.

- Una vez que se consolida el prototipo de detección y el cliente se acostumbra a seguir sus recomendaciones la tasa de llamadas de atención disminuirá a ritmo moderado al igual que la carga operativa del centro de atención al cliente.

- El prototipo de detección logró catalogar adecuadamente los cajeros de los distintos clientes usando métricas como puntuación de anomalía para tomar la decisión de etiquetarlo como anómalos y presentarlos en una tabla con su información descriptiva y distribución jerárquica.

Entre las recomendaciones más importantes de este proyecto se tiene:

- Es decisión exclusiva del cliente realizar mantenimiento preventivo de aquellos cajeros detectados como anómalos, si sigue las recomendaciones adecuadas generará ahorros financieros para la institución a mediano y largo plazo.

- El cliente debe realizar revisiones regulares de detección, no más allá de una vez por semana por medio del módulo de detección en la aplicación de monitoreo, para estar al tanto de posibles cajeros problemáticos que antes no fueron identificados y que ahora pudieran presentar esta situación de anomalía.

- Definir algún medio o canal de comunicación en el que se pueda informar al cliente de los hallazgos de la detección de anomalías de forma automática.

BIBLIOGRAFÍA

[1] U. N. A. Razimi, & F. Y Ahmed & N. R. Mustapa. (2019). ATM Reporting System, IEEE 9th International Conference on System Engineering and Technology (ICSET), 166-171. Accedido el 10 de septiembre, 2023, desde https://ieeexplore.ieee.org/document/8906494

[2] Yogesh Kumawat & Manish Dubey. (2017). A Review Paper on ATM Transaction. International Journal of Advance Research in Computer Science and Management Studies, Volume 5, Issue 4, 23-26. Accedido el 10 de septiembre, 2023, desde http://ijarcsms.com/docs/paper/volume5/issue4/V5I4-0020.pdf

[3] Antoine Guillaume & Christel Vrain & Wael Elloumi. (2021). Predictive maintenance on event logs: Application on an ATM Fleet. Accedido el 10 de septiembre, 2023, desde https://arxiv.org/abs/2011.10996

[4] BANRED La red interbancaria más grande del Ecuador. (2023). https://www.banred.fin.ec

[5] Jiang, Y., Wang, W., & Zhao, C. (2019). A Machine Vision-based Realtime Anomaly Detection Method for Industrial Products Using Deep Learning. Chinese Automation Congress (CAC). https://doi.org/10.1109/cac48633.2019.8997079

[6] Economipedia. Cajero automático (2020). https://economipedia.com/definiciones/cajero-automatico.html

[7] Domina los datos (2023). https://dominalosdatos.com/deteccion-anomalias

[8] López, L., Acosta, N., & Gago, A. (2019). Detección de anomalías basada en aprendizaje profundo: Revisión. Revista Cubana de Ciencias Informáticas, 13(3), 107-123. Recuperado el 23 de septiembre de 2023, de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2227-18992019000300107

[9] The Atlantic. A Brief History of the ATM (2015). https://www.theatlantic.com/technology/archive/2015/03/a-brief-history-of-the-atm/388547

[10] Bank Info Security. The history of EMV (2011). https://www.bankinfosecurity.com/interviews/history-emv-i-933

[11] Laimek, R., & Kaothanthong, N. (2018). ATM Fraud Detection using Behavior Model. 5th Asian Conference on Defense Technology. https://doi.org/10.1109/acdt.2018.8593092

[12] Zeng, Y., Zhang, Z., Guo, H., Chen, Y., Shen, S., Zhu, L. & Yu, B. ATM Transaction Status Feature Analysis and Anomaly Detection. Studies in Engineering and Technology. Volume 6. Number 1. https://doi.org/10.11114/set.v6i1.3829

[13] IBM. What is unsupervised learning? (2023). https://www.ibm.com/topics/unsupervised-learning

[14] Oracle AI & Data Science Blog. Introduction to K-means Clustering (2016) https://blogs.oracle.com/ai-and-datascience/post/introduction-to-k-means-clustering

[15] Analytics Vidhya. K Means Clustering (2023). https://www.analyticsvidhya.com/blog/2021/04/k-means-clustering-simplified-in-python

[16] Liu, F., Ting, K. M., & Zhou, Z. (2008). Isolation Forest. Eighth IEEE International Conference on Data Mining. https://doi.org/10.1109/icdm.2008.17

[17] Analytics Vidhya. Anomaly detection using Isolation Forest (2023). https://www.analyticsvidhya.com/blog/2021/07/anomaly-detection-using-isolation-forest-a-complete-guide

[18] PyShark. Calinski-Harabasz Index for K-Means Clustering Evaluation using Python (2023). https://pyshark.com/calinski-harabasz-index-for-k-means-clustering-evaluation-using-python

[19] Wijaya, Y. A., Kurniady, D. A., Setyanto, E., Tarihoran, W. S., Rusmana, D., & Rahim, R. (2021). Davies Bouldin Index Algorithm for optimizing clustering case studies mapping school facilities. TEM Journal, 1099–1103. https://doi.org/10.18421/tem103-13

[20] Python (2023). https://www.python.org/doc/

[21] BookDown. Introducción a R y SIG (2019). https://bookdown.org/chescosalgado/intro_r/intro.html

[22] PyCaret 3.0. An open-source, low-code machine learning library in Python (2023). https://pycaret.gitbook.io/docs/

[23] Streamlit. Streamlit documentation (2023). https://docs.streamlit.io

[24] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

[25] MurmurHash – a FAST hashing algorithm (2014). https://www.saguiitay.com/murmurhash-a-fast-hashing-algorithm/

GLOSARIO

Cajero automático. - Es un dispositivo electrónico que permite a los usuarios realizar diversas transacciones bancarias sin necesidad de interacción humana ni de visitar una sucursal bancaria física.

Mantenimiento correctivo. - Es un tipo de mantenimiento que se realiza después de detectar un problema o fallo y que requiere una acción correctiva para devolver el equipo, la maquinaria o el sistema a su estado operativo.

Mantenimiento preventivo. - Es un tipo de mantenimiento con un enfoque proactivo de inspeccionar, revisar y mantener periódicamente equipos, maquinaria o sistemas para evitar averías, fallos u otros problemas antes de que se produzcan.

Anomalía. - Es algo que se desvía de lo que se considera normal, esperado o estándar dentro de un contexto específico.

Pipeline. - Es una serie de pasos o etapas de procesamiento que recopilan, procesan, transforman y mueven datos de un sistema o ubicación a otro.

Tiempo de inactividad. - Es el periodo o duración de tiempo en la que el dispositivo no puede realizar su función prevista debido a mantenimiento, reparación, mal funcionamiento u otras razones.

K-means. - Es un popular algoritmo de aprendizaje automático no supervisado que se utiliza para agrupar datos en distintos grupos o clústeres en función de su similitud.

Isolation forest. - Es un algoritmo de detección de anomalías utilizado en el aprendizaje automático para identificar anomalías o valores atípicos en un conjunto de datos.

Clúster. - Es una agrupación de puntos de datos que comparten características o propiedades similares dentro de un conjunto de datos.

EMV. - Se refiere a la tecnología y las normas utilizadas en las tarjetas de pago (como las tarjetas con chip) y los sistemas de procesamiento de pagos para mejorar la seguridad y reducir el fraude durante las transacciones en persona.

Cash dispenser. - Es un tipo de cajero automático que funciona principalmente para dispensar efectivo a los usuarios.

Recycler. - Es un tipo de cajeros automático avanzado que combina las funciones de retiro y depósito de efectivo en una sola máquina.

Branch. - Es un tipo de cajero automático que está situado dentro o muy cerca de una sucursal bancaria física.

Lobby. - Es un tipo de cajero automático situado en el interior o en el vestíbulo de un edificio, como un banco, un centro comercial o cualquier otro establecimiento accesible al público.

One-hot. - Es una técnica utilizada en el aprendizaje automático y el procesamiento de datos para representar variables categóricas como vectores binarios.

Z-Score. - Es una medida estadística que indica cuántas desviaciones estándar hay entre un punto de datos y la media de un conjunto de datos.

MurmurHash. - Es una familia de funciones hash no criptográficas diseñadas para realizar hashing de datos rápidamente con bajas tasas de colisión.

APÉNDICES

```python
#Módulo de detección de anomalías sobre los tiempos de
#inactividad de cajeros automáticos usando Isolation Forest.
#Christian Jaramillo Espinoza - MCD 2023

import streamlit as st
import pandas as pd
import numpy as np
import holoviews as hv
from google.cloud import storage
from google.oauth2 import service_account
from pycaret.clustering import *
from pycaret.anomaly import AnomalyExperiment
from datetime import datetime

project_id = 'mcd-proyecto'
bucket_name = "mcdproyectobucket"
file_name = "dataset-v6-testweek35-ofuscated.csv"
iforest_model_name = "iforest_model_downtime"

#dataframes
data = []
data_g = []
categories = []
data_pivot = []
data_pivot_no_geo = []
cluster_anomaly = []
anomalies = []
merged = []

@st.cache_data
def load():
    #Descarga del conjunto de datos
    credentials = service_account.Credentials.from_service_account_file("google-credentials.json")
    storage_client = storage.Client(project=project_id, credentials=credentials)

    bucket = storage_client.get_bucket(bucket_name)

    blob = bucket.blob(file_name)
    dataset_filename = "dataset.csv"
    blob.download_to_filename(dataset_filename)

def evaluate():
    global data
    global data_g
    global categories
    global data_pivot
    global data_pivot_no_geo
    global cluster_anomaly
    global anomalies
    global merged

    #Preprocesamiento de los datos
    data = pd.read_csv("dataset.csv", sep=";", encoding="UTF-8")
    data['DATETIME'] = pd.to_datetime(data['DATETIME'])

    categories = data[["CUSTOMER", "ID", "MODEL", "FUNCTION", "FAMILY", "SITE", "STATE", "CITY", "COUNTRY"]]
    categories = categories.drop_duplicates(keep='first').reset_index()
    categories.set_index("ID", inplace=True)
    categories.drop(columns=["index"], inplace=True)
```

Implementación del módulo de anomalías I

```
60
61          downtime = data[["ID", "DATETIME", "WEEK", "CARD_DOWNTIME", "CASH_DOWNTIME", "ACCEPTOR_DOWNTIME",
62                           "DEPOSITOR_DOWNTIME", "EPP_DOWNTIME", "PRINTER_DOWNTIME"]]
63
64          data_g = downtime.groupby(by=["ID", "WEEK"]).agg(
65                      CARD_DOWNTIME = ("CARD_DOWNTIME", "mean"),
66                      CASH_DOWNTIME = ("CASH_DOWNTIME", "mean"),
67                      ACCEPTOR_DOWNTIME = ("ACCEPTOR_DOWNTIME", "mean"),
68                      DEPOSITOR_DOWNTIME = ("DEPOSITOR_DOWNTIME", "mean"),
69                      EPP_DOWNTIME = ("EPP_DOWNTIME", "mean"),
70                      PRINTER_DOWNTIME = ("PRINTER_DOWNTIME", "mean")
71                    ).reset_index()
72
73          week=35
74          data_g_c = pd.DataFrame(data_g, copy=True)
75
76          for i in range(week+1, week+12):
77            data_g_b = pd.DataFrame(data_g, copy=True)
78            data_g_b["WEEK"] = i
79            data_g_c = data_g_c.append(data_g_b).reset_index(drop=True)
80
81          data_g = pd.DataFrame(data_g_c, copy=True)
82
83          data_pivot = data_g.pivot_table(index='ID',
84                                          columns='WEEK',
85                                          values=["CARD_DOWNTIME", "CASH_DOWNTIME",
86                                                  "ACCEPTOR_DOWNTIME", "DEPOSITOR_DOWNTIME",
87                                                  "EPP_DOWNTIME", "PRINTER_DOWNTIME"],
88                                          aggfunc='mean', sort=False)
89          data_pivot.columns = [f'W{i}' for i in range(data_pivot.columns.size)]
90          data_pivot_week_columns = data_pivot.columns
91
92          data_pivot_no_geo = pd.DataFrame(data_pivot, copy=True)
93          data_pivot_no_geo["CUSTOMER"] = categories["CUSTOMER"]
94          data_pivot_no_geo["MODEL"] = categories["MODEL"]
95          data_pivot_no_geo["FUNCTION"] = categories["FUNCTION"]
96          data_pivot_no_geo["FAMILY"] = categories["FAMILY"]
97          data_pivot_no_geo["SITE"] = categories["SITE"]
98          data_pivot_no_geo["COUNTRY"] = categories["COUNTRY"]
99
100         step = 12
101         columnsByDevice = {}
102         columnsByDeviceEvents = {}
103
104         for i in range(0, len(data_pivot_week_columns.values), step):
105             chunk = data_pivot_week_columns[i:i+step].values
106             key = f'group_{i//step + 1}'
107             columnsByDevice[key] = chunk
108
109         auth = {"project": project_id, "bucket": bucket_name}
110         iforest_model = load_model(model_name=iforest_model_name, platform="gcp", authentication=auth)
111
112         iforest_setup = AnomalyExperiment()
113         result_iforest = iforest_setup.predict_model(iforest_model, data=data_pivot_no_geo)
114
115         no_anomalies = result_iforest[result_iforest["Anomaly"] == 0]
116         anomalies = result_iforest[result_iforest["Anomaly"] == 1]
117
118         merged = anomalies.reset_index()
```

Implementación del módulo de anomalías II

```
if __name__ == '__main__':
    st.set_page_config(layout="wide")

    load()

    evaluate()

    #Obtención de los clientes
    st.subheader('Módulo de detección de anomalías', divider='red')

    customer_count = anomalies.groupby("CUSTOMER").agg(Cantidad = ("CUSTOMER","count")).reset_index()
    customer_count["CUSTOMER_B"] = customer_count["CUSTOMER"].astype(str) + "   (" + customer_count["Cantidad"].astype(str) + ")"
    n_anomalies = 0

    def custom_format(option):
        n anomalies = customer count.loc[customer count["CUSTOMER"] == option]["Cantidad"]
    data_filtered = pd.DataFrame(merged, copy=True)
    data_filtered = data_filtered.loc[data_filtered["CUSTOMER"] == customerSelected]

    data_filtered =  pd.DataFrame({"ID": data_filtered["ID"],
                                   "FAMILY": data_filtered["FAMILY"],
                                   "FUNCTION": data_filtered["FUNCTION"],
                                   "SITE": data_filtered["SITE"],
                                   "MODEL": data_filtered["MODEL"],
                                   "CARD_DOWTIME": [np.array(data_filtered.loc[data_filtered["ID"] == id].melt()[1:13]["value"]) for id in data_filtered["ID"]],
                                   "CASH_DOWTIME": [np.array(data_filtered.loc[data_filtered["ID"] == id].melt()[13:25]["value"]) for id in data_filtered["ID"]],
                                   "ACCEPTOR_DOWTIME": [np.array(data_filtered.loc[data_filtered["ID"] == id].melt()[25:37]["value"]) for id in data_filtered["ID"]],
                                   "DEPOSITOR_DOWTIME": [np.array(data_filtered.loc[data_filtered["ID"] == id].melt()[37:49]["value"]) for id in data_filtered["ID"]],
                                   "EPP_DOWTIME": [np.array(data_filtered.loc[data_filtered["ID"] == id].melt()[49:61]["value"]) for id in data_filtered["ID"]],
                                   "PRINTER_DOWTIME": [np.array(data_filtered.loc[data_filtered["ID"] == id].melt()[61:73]["value"]) for id in data_filtered["ID"]]
                                   })

    anomalies_by_customer = anomalies.loc[anomalies["CUSTOMER"] == customerSelected]

                                   "EPP_DOWTIME": [np.array(data_filtered.loc[data_filtered["ID"] == id].melt()[49:61]["value"]) for id in data_filtered["ID"]],
                                   "PRINTER_DOWTIME": [np.array(data_filtered.loc[data_filtered["ID"] == id].melt()[61:73]["value"]) for id in data_filtered["ID"]]
                                   })
    #Datos filtrados por cliente
    anomalies_by_customer = anomalies.loc[anomalies["CUSTOMER"] == customerSelected]

    df_fam_fun = anomalies_by_customer.groupby(['FAMILY', 'FUNCTION'])['W0'].count().reset_index()
    df_fam_fun.columns = ['source', 'target', 'value']

    df_fun_sit = anomalies_by_customer.groupby(['FUNCTION', 'SITE'])['W0'].count().reset_index()
    df_fun_sit.columns = ['source', 'target', 'value']

    df_sit_mod = anomalies_by_customer.groupby(['SITE', 'MODEL'])['W0'].count().reset_index()
    df_sit_mod.columns = ['source', 'target', 'value']

    links = pd.concat([df_fam_fun, df_fun_sit, df_sit_mod], axis=0)

    hv.extension('bokeh')
    links_filtered = links.loc[links["value"] > 0]
    nlinks = len(links_filtered)

    def hide_hook(plot, element):
        plot.handles["xaxis"].visible = False
        plot.handles["yaxis"].visible = False
        plot.handles["plot"].border_fill_color = None
        plot.handles["plot"].outline_line_color = None
```

Implementación del módulo de anomalías III

```
181         if nlinks > 0:
182             #Creación de gráfica sankey
183             sankey = hv.Sankey(links_filtered, label='')
184             sankey.opts(width=650, height=375, hooks=[hide_hook], toolbar=None, default_tools = [],
185                         label_position='outer', edge_color='lightgray', node_color='index', cmap='tab20c', node_padding=20)
186 
187         col1, col2 = st.columns([2, 1])
188 
189         with col1:
190             #Visualización del dataframe
191             st.subheader("Cajeros anómalos detectados")
192             st.dataframe(data_filtered.reset_index(drop=True),
193                          hide_index=False,
194                          use_container_width=True,
195                          column_config={
196                             "ID": st.column_config.TextColumn(label="ATM ID", width="small"),
197                             "FAMILY": st.column_config.TextColumn(label="FAMILIA", width="small"),
198                             "FUNCTION": st.column_config.TextColumn(label="FUNCION", width="small"),
199                             "SITE": st.column_config.TextColumn(label="TIPO", width="small"),
200                             "MODEL": st.column_config.TextColumn(label="MODELO", width="small"),
201                             "CARD_DOWTIME": st.column_config.LineChartColumn("TARJETA (s)", y_min=0, y_max=86400, width="small",
202                                                                              help="Promedio semanal del tiempo de inactividad de la lectora de tarjetas"),
203                             "CASH_DOWTIME": st.column_config.LineChartColumn("DISPENSADOR (s)", y_min=0, y_max=86400, width="small",
204                                                                              help="Promedio semanal del tiempo de inactividad del dispensador de efectivo"),
205                             "ACCEPTOR_DOWTIME": st.column_config.LineChartColumn("ACEPTADOR (s)", y_min=0, y_max=86400, width="small",
206                                                                                  help="Promedio semanal del tiempo de inactividad del aceptador de efectivo"),
207                             "DEPOSITOR_DOWTIME": st.column_config.LineChartColumn("CHEQUE (s)", y_min=0, y_max=86400, width="small",
208                                                                                   help="Promedio semanal del tiempo de inactividad del depósito de cheques"),
209                             "EPP_DOWTIME": st.column_config.LineChartColumn("TECLADO (s)", y_min=0, y_max=86400, width="small",
210                                                                             help="Promedio semanal del tiempo de inactividad del teclado electrónico"),
211                             "PRINTER_DOWTIME": st.column_config.LineChartColumn("IMPRESORA (s)", y_min=0, y_max=86400, width="small",
212                                                                                 help="Promedio semanal del tiempo de inactividad de la impresora de recibos"),
213                          })
214         with col2:
215             #Visualización del gráfico Sanky
216             st.subheader("Distribución jerárquica")
217 
218             if nlinks > 0:
219                 st.bokeh_chart(hv.render(sankey, backend='bokeh'))
220 
221         if st.session_state.selectbox_customers != customerSelected:
222             st.session_state.selectbox_customers = customerSelected
```

Implementación del módulo de anomalías IV