Caracterización de Perfiles de Clientes en el ROS (Reporte de Operación Sospechosa) y Diseño de Reglas para Actualizar la Evaluación de Riesgo LAFT

Cesar Augusto Saenz Jimenez

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

Asesor
Ronald Akerman Ortiz García, Magíster (MSc)

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2024

Cita	(Saenz Jimenez, 2024)
Referencia **Estilo APA 7 (2020)**	Saenz Jimenez, C.A., (2024). *Caracterización de Perfiles de Clientes en el ROS (Reporte de Operación Sospechosa) y Diseño de Reglas para Actualizar la Evaluación de Riesgo LAFT* Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.

Especialización en Analítica y Ciencia de Datos, Cohorte IV.

Centro de Investigación Ambientales y de Ingeniería (CIA).

UNIVERSIDAD DE ANTIOQUIA
Vicerrectoría de Docencia

Sistema de Bibliotecas

Centro de Documentación Ingeniería (CENDOI)

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

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.
Decano: Julio Cesar Saldarriaga Molina
Jefe departamento: Diego José Luis Botia Valderrama

Caracterización de Perfiles de Clientes en el ROS (Reporte de Operación Sospechosa) y Diseño de Reglas para Actualizar la Evaluación de Riesgo LAFT

Dedicatoria

Queridos amigos, familiares y seres queridos,
Con gran emoción y agradecimiento, quiero dedicar este trabajo de grado a todos ustedes ya que han sido una fuente constante de apoyo y motivación en mi vida.

A mis amigos y compañeros, gracias por los momentos, por las conversaciones profundas y por ayudarnos a mantener el equilibrio en momentos de estrés. Ustedes son parte de mi familia y estoy muy agradecido de tenerlos en mi vida.

A mis Asesores de Monografía, gracias por su guía, apoyo y paciencia en cada una de las etapas de este trabajo. Sus conocimientos y experiencia han sido fundamentales para alcanzar este logro.

Finalmente, quiero dedicar este trabajo de grado a mi propio esfuerzo y dedicación. Este logro es una muestra del compromiso y perseverancia. Espero que este trabajo pueda servir para contribuir al conocimiento en mi campo de estudio y para mejorar la calidad de vida de nuestra sociedad.

Con gratitud y cariño.

Agradecimientos

Dedico este pequeño espacio a mi familia por su amor incondicional, su paciencia y su constante aliento. Gracias por creer en mí y por apoyarme en cada uno de mis sueños y metas.

Agradezco a la empresa Bancolombia. Agradezco por su confianza en mí, al depositar la responsabilidad de poder manejar sus datos, Su compromiso con la seguridad y la transparencia financiera me inspira a trabajar con la mayor dedicación y profesionalismo.

Reconozco con profunda gratitud el apoyo y la colaboración de mis asesores David Manuel Villanueva Valdés y Ronald Ortiz quienes han sido pilares fundamentales en mi desarrollo académico y profesional Agradezco su valioso tiempo y esfuerzo dedicados a la revisión de este trabajo de grado. Su compromiso y dedicación en la evaluación y retroalimentación del proyecto han sido fundamentales para mi desarrollo académico y profesional.

Me siento muy afortunado de haber contado con sus orientaciones y asesoramientos, el cual ha sido esencial para alcanzar los objetivos planteados en mi trabajo de grado. Gracias a sus comentarios y sugerencias, he podido mejorar y perfeccionar el trabajo final, lo que me permitirá llevarlo a la práctica para la empresa que estoy asesorando.

Una vez más, agradezco su labor y el tiempo dedicado a la revisión de este trabajo. Sus valiosas contribuciones son una gran ayuda en mi formación y desarrollo como profesional en la Especialización en Analítica y Ciencia de Datos.

Tabla de contenido

Lista de tablas

Lista de figuras

Siglas, acrónimos y abreviaturas

APA	American Psychological Association
Esp.	Especialista
MSc	Magister Scientiae
Párr.	Párrafo
UdeA	Universidad de Antioquia
ROS	Reporte de Operación Sospechosa
LAFT	Lavado de Activos y Financiación del Terrorismo
UIAF	Unidad de Información y Análisis Financiero
SARLAFT	Sistema de Administración del Riesgo de Lavado de Activos y Financiación al Terrorismo

Resumen

El presente proyecto tiene como objetivo desarrollar una herramienta adicional que identifique patrones de relación entre las características de los clientes con calificación en los Reportes de Operaciones Sospechosas (ROS). Y estos patrones convertirlos en reglas que sirván como parámetros para evaluar el riesgo (LAFT) (Superintendencia Financiera de Colombia, 2003), Logrando Así reclasificar aquellos clientes cuya calificación inicial haya estado incorrecta, en cumplimiento de la Circular Externa 027 de la Superintendencia Financiera de Colombia. Este marco regulatorio busca mitigar los riesgos asociados al lavado de activos y financiamiento del terrorismo, un problema que afecta gravemente la integridad del sistema financiero en Colombia (Superintendencia Financiera de Colombia, 2019).

La segmentación de riesgos implica un análisis profundo para dividir elementos en grupos homogéneos y heterogéneos, utilizando distintas variables relacionadas con los clientes, productos y jurisdicciones. Este proceso es para nosotros esencial ya que identifica patrones sospechosos y mejora las estrategias de prevención y detección de las actividades ilícitas.

El reto principal radica en manejar grandes volúmenes de datos y variables diversas, asegurando que los modelos de análisis que se apliquen sean precisos y eficientes. Se requiere una metodología robusta que permita comprender y clasificar adecuadamente los datos para cumplir con las exigencias regulatorias y fortalecer las medidas contra el lavado de activos y la financiación del terrorismo.

Palabras clave: ROS, LAFT, SARLAFT, Clúster. Reglas de Asociación.

Abstract

The present project aims to develop an additional tool to identify relationship patterns between customer characteristics and their ratings in the Suspicious Activity Reports (SARs). These patterns will be converted into rules that will serve as parameters for evaluating risk (LAFT) (Superintendencia Financiera de Colombia, 2003), thus adjusting the rating of those customers whose initial assessment was incorrect, in compliance with External Circular 027 of the Superintendencia Financiera de Colombia. This regulatory framework seeks to mitigate the risks associated with money laundering and terrorist financing, a problem that seriously affects the integrity of the financial system in Colombia (Superintendencia Financiera de Colombia, 2019).

Risk segmentation involves a deep analysis to divide elements into homogeneous and heterogeneous groups, using different variables related to customers, products, and jurisdictions. This process is essential for us as it identifies suspicious patterns and improves prevention and detection strategies for illicit activities.

The main challenge lies in handling large volumes of data and diverse variables, ensuring that the analysis models applied are accurate and efficient. A robust methodology is required to properly understand and classify data in order to meet regulatory requirements and strengthen anti-money laundering and terrorist financing measures.

*Keywords***:** Clustering, Association Rules, Prediction, ROS, LAFT, SARLAFT.

1. Descripción del problema

Este proyecto aborda la necesidad que tiene el Grupo Bancolombia en mejorar la gestión de riesgos LA/FT siendo una iniciativa fundamental para fortalecer la posición de la organización en la lucha contra el lavado de activos y la financiación del terrorismo. La implementación exitosa de este proyecto permitirá cumplir con las regulaciones vigentes, prevenir actividades ilícitas y optimizar la eficiencia de sus procesos, lo cual es fundamental para evitar sanciones regulatorias, pérdidas financieras y daños reputacionales (Compliance, 2024).

Para resolver este problema, se propone utilizar la metodología CRISP-DM, que incluye la implementación de algoritmos no supervisados. Dado que en este caso no se dispone de una variable objetivo o datos etiquetados, esta metodología permite aplicar diversas técnicas según los objetivos y la naturaleza del proyecto. Entre estas técnicas se encuentran los modelos de clusterización, siendo el algoritmo K-Prototypes especialmente adecuado para interpretar datos mixtos y segmentar a los clientes según sus características únicas (Reddy, 2023). Este enfoque permitirá identificar patrones de comportamiento y, a partir de ellos, generar reglas de asociación utilizando el algoritmo FP-Growth. Esto facilitará la detección temprana de actividades ilícitas y mejorará la eficacia en la gestión de riesgos financieros (Ali, 2019).

1.1. Problema de negocio

Bancolombia busca implementar técnicas para optimizar la calificación de los clientes reportados en los Reportes de Operaciones Sospechosas (ROS), ya que se han detectado calificaciones incorrectas durante el seguimiento. Este problema requiere reprocesos y ajustes de parámetros para recalificar a los usuarios, por lo que constantemente se trabaja en la implementación de modelos que evalúan el nivel de riesgo de cada cliente. Estos modelos consideran diversas variables, incluyendo productos, canales y jurisdicciones, para mejorar la precisión en la calificación del riesgo.

El objetivo es evaluar el nivel de riesgo LAFT de cada cliente, considerando variables cualitativas y cuantitativas, como aspectos demográficos, actividad económica y presencia en listas de PEP-desmovilizado-cliente. También se tienen en cuenta variables transaccionales del cliente, como productos, canales y jurisdicciones de transacciones. Esta calificación determina la exposición al riesgo LAFT de cada cliente para el Banco y se utiliza para la toma de decisiones y gestiones en el área de Cumplimiento (UIAF, 2014).

1.2. Aproximación desde la analítica de datos

El Grupo Bancolombia, comprometido con la mejora continua, sugiere ajustar los parámetros de los modelos y técnicas utilizados para la gestión del riesgo de Lavado de Activos y Financiación del Terrorismo (LA/FT). Es esencial combinar técnicas cualitativas y cuantitativas, aprovechando diversas fuentes de información para medir el riesgo de manera más objetiva.

Para ello, se propone segmentar a los clientes mediante un modelo de clusterización. Este enfoque permitirá aplicar las técnicas y ajustes necesarios para mejorar la efectividad de los resultados. El objetivo es analizar los riesgos para cada factor de riesgo, determinando en cada segmento la probabilidad de ocurrencia y la magnitud del impacto en caso de materializarse.

La segmentación resultante se presentará en forma de clústeres utilizando el algoritmo *K-Prototypes*. El motivo de Clústerizar a los clientes es lograr una agrupación coherente basada en

características compartidas, permitiendo una comprensión más detallada y específica de cada grupo generado. Esta agrupación facilita la aplicación de técnicas avanzadas de análisis y permite identificar patrones que no serían evidentes en un análisis global de los datos.

Para cada grupo, se generarán reglas de asociación aplicando el modelo FP-Growth, con el fin de analizar los riesgos y causas identificados previamente. Las reglas de asociación ayudarán a entender cómo diferentes factores se relacionan entre sí dentro de cada clúster, permitiendo identificar combinaciones de características que aumentan el riesgo. Este análisis permitirá determinar la probabilidad de ocurrencia y el impacto de las consecuencias en cada segmento de manera precisa. Así, se pueden diseñar estrategias de mitigación de riesgos más efectivas y dirigidas específicamente a las características y comportamientos de cada clúster de clientes

1.3. Origen de los datos

Los datos utilizados para este proyecto son el resultado de calificación obtenida según el modelo que actualmente tienen implementado para la Evaluación del Riesgo LAF correspondiente al periodo con corte a febrero del 2024 y entregados por el funcionario del área en un archivo en Excel.

1.4. Métricas de desempeño

1.4.1 Métricas de desempeño Modelo Clusterización K-Prototype

Para lograr evaluar la precisión del modelo de clusterización al usar el algoritmo *K-Prototypes*, se pretende entender la calidad de los clústers formados, puedo citar:

- **Coeficiente de Silhouette:** Esta Métrica mide qué tan similar es un punto a su propio clúster (cohesión) en comparación con otros clústeres (separación). Lo que quiere decir que un valor cercano a 1 indica clústers bien definidos, mientras que valores cercanos a -1 indican lo contrario (Rodriguez, 2023).

- **Suma de Distancias Cuadradas (Inertia):** La Inertia mide la suma de las distancias cuadradas de cada punto a su clúster más cercano. Un valor menor de Inertia indica clústers más compactos (Chauhan, 2023).
- **Davies-Bouldin Score:** Esta Métrica mide la relación entre la dispersión dentro de los clústers y la separación entre los clústers. Lo que indica que un valor menor es que los clústers más compactos y están bien separados (Rodriguez D. , 2023).
- **Calinski-Harabasz Score**: Por último esta métrica mide la ratio de la suma de la dispersión entre clústers y la suma de la dispersión dentro de los clústers. Lo que indica que un valor mayor es porque los clústers están mejor definidos (Rodriguez D. , www.analyticslane.com, 2023).

1.4.2 Métricas de desempeño Modelo Reglas Asociación Modelo FP-Growth

Estas Métrica asegura la calidad, relevancia y utilidad de las reglas descubiertas en el proceso de análisis de datos. Estas métricas permiten evaluar la efectividad del modelo en identificar patrones significativos que puedan ser utilizados para la gestión del riesgo (LA/FT).

- **Soporte**: Se emplearon valores dentro del rango de 0 a 1 para determinar la frecuencia con la que aparece un conjunto de elementos en la base de datos. Se consideraron las frecuencias más altas como indicativas de mayor confiabilidad.
- **Confianza**: Se utilizaron valores entre 0 y 1 como parámetros para evaluar la fiabilidad de las reglas de asociación.
- **Lift**: Se emplearon valores entre 0 y 1 para el análisis del Lift. Un valor superior a 1 sugiere que la ocurrencia del consecuente es más probable dado el antecedente, lo que indica una asociación positiva.
- **Convicción**: Los valores utilizados estuvieron en el rango de 0 a 1 para evaluar la fuerza de implicación de una regla. Un alto valor de convicción indica que el consecuente rara vez aparece sin el antecedente, lo que sugiere una relación fuerte entre ambos.

La evaluación del rendimiento de las reglas de asociación se basa en diversos conceptos. Si se busca identificar reglas comunes en el conjunto de datos, la métrica de soporte resulta adecuada. En cambio, para encontrar reglas sólidas, la métrica de confianza es la mejor en este caso.

2. Objetivos

2.1.Objetivo general

Generar reglas de asociación para mejorar el análisis de riesgos financieros. Este modelo utilizará técnicas de aprendizaje automático para examinar grandes conjuntos de datos e identificar patrones de comportamiento de los clientes de manera eficiente y precisa. Las reglas de asociación generadas nutrirán el modelo de calificación para los clientes del Reporte de Operación Sospechosa (ROS).

2.2.Objetivos específicos

- Desarrollar un modelo de clasificación utilizando el algoritmo K-Prototype para segmentar a los clientes del GRUPO BANCOLOMBIA según sus características individuales. Esto implica recopilar y limpiar datos, definir variables relevantes e implementar el algoritmo para una segmentación eficiente.
- Utilizar el modelo de clasificación desarrollado para realizar un análisis detallado junto con expertos del área, con el fin de identificar las principales características de cada segmento de clientes. Esto incluye interpretar los resultados del modelo y colaborar con funcionarios del área para comprender los patrones y comportamientos observados en cada segmento
- Proporcionar al GRUPO BANCOLOMBIA herramientas de análisis financiero para detectar, gestionar y reducir los riesgos relacionados con actividades ilícitas. Esto se logrará mediante la creación de reglas de asociación basadas en el algoritmo FP-Growth, permitiendo la detección temprana de operaciones sospechosas y fortaleciendo la confianza de reguladores y clientes en los servicios bancarios

3. Datos

3.1. Datos originales

La base de Datos Proporcionada por la vicepresidencia de riesgos financieros del Grupo Bancolombia, con un total de 25849 registros. Utilizando el método MissingIndicator se calcula el porcentaje de los datos faltantes de la base de datos entregada. Identificando un promedio del 29.963% de datos ausentes, lo cual representa un desafío significativo para abordar durante el análisis. La base de datos incluye información detallada sobre los clientes, distribuida en 28 columnas categóricas y 18 columnas numéricas, abarcando datos de identificación que son proporcionados al momento de la vinculación, información sociodemográfica, origen de los recursos, descripción del reporte ROS, y reporte de las listas de control e historial de transacciones. Con un tamaño total de 6052 KB, por su tamaño esta base de datos permite un análisis eficiente, siendo suministrada por el analista de riesgos encargado de apoyarme en un archivo de formato Excel. La diversidad de la información contenida en las columnas plantea desafíos y oportunidades para aplicar técnicas avanzadas de preprocesamiento y modelado, fundamentales para la metodología a aplicar CRISP-DM.

Tabla 1
Características Dataset Entregado

Dataset statistics	
Number of variables	47
Number of observations	25489
Missing cells	221339
Missing cells (%)	18.5%
Duplicate rows	18
Duplicate rows (%)	0.1%
Total size in memory	6.052 MiB
Average record size in memory	6052000.0 B

Variable types	
Numeric	19
Categorical	28

Nota. Datos extraídos del Reporte Dataset Entregado (Librería y_data_profile)

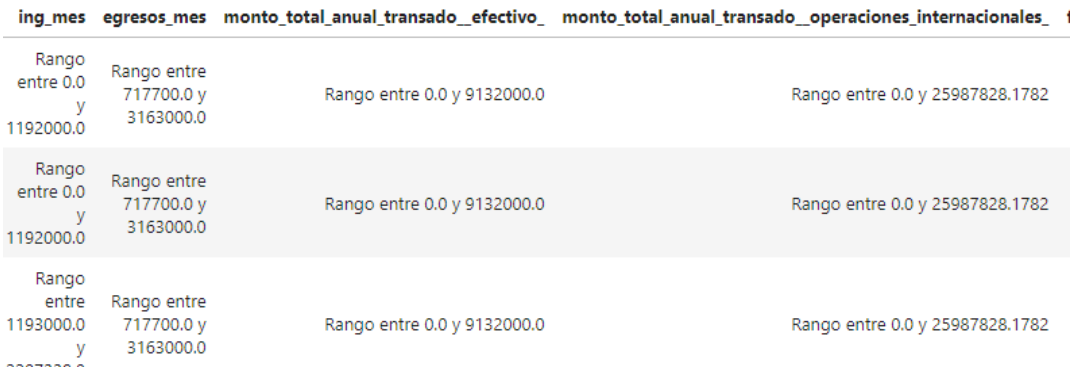
3.2. Datasets

Debido a restricciones de confidencialidad en la información suministrada, se llevó a cabo un proceso de transformación y limpieza del conjunto de datos. Se realizaron varios pasos, que incluyeron:

- Eliminar columnas que contenían codificaciones internas del Banco, reemplazándolas por una columna continua con su descripción.
- Se identificaron y eliminaron columnas con datos personales del cliente que no eran relevantes para el modelo.
- Se aplicaron técnicas para reducir la dimensionalidad, como fue utilizar la matriz de correlación de Pearson para identificar columnas irrelevantes y eliminarlas sin afectar el rendimiento del modelo.
- Se dividió el conjunto de datos en variables categóricas y numéricas para aplicar técnicas de imputación adecuadas a cada tipo de dato.
- Para las variables numéricas, se utilizó la biblioteca Scikit-Learn y el módulo SimpleImputer y se asignó el valor '0' a los valores faltantes, según las sugerencias del analista que me asignaron para el proyecto.
- Para las variables categóricas, también se utilizó Scikit-Learn y el módulo SimpleImputer para asignar la categoría 'Sin Información' a los valores faltantes, por recomendación del analista de riesgos.
- Después de la transformación, se unificó nuevamente el conjunto de datos.
- Se identificaron varias columnas con valores numéricos dispersos, lo que dificultaba su interpretación. Por lo tanto, se segmentaron los valores de estas columnas en 10 bloques y se asignaron a su categoría correspondiente para una mejor comprensión como se logra ver a continuación:

Figura 1

División de Rangos Para Valores Dispersos



ing_mes	egresos_mes	monto_total_anual_transado__efectivo_	monto_total_anual_transado__operaciones_internacionales_	f
Rango entre 0.0 y 1192000.0	Rango entre 717700.0 y 3163000.0	Rango entre 0.0 y 9132000.0	Rango entre 0.0 y 25987828.1782	
Rango entre 0.0 y 1192000.0	Rango entre 717700.0 y 3163000.0	Rango entre 0.0 y 9132000.0	Rango entre 0.0 y 25987828.1782	
Rango entre 1193000.0 y	Rango entre 717700.0 y 3163000.0	Rango entre 0.0 y 9132000.0	Rango entre 0.0 y 25987828.1782	

Nota: El grafico muestra el resultado final luego de aplicar la división de los rangos para las columnas que tenían datos muy dispersos

Repositorio: https://github.com/Cesar012782/Caracterizacion-Perfiles-Clientes.

3.3. Analítica descriptiva

A continuación logra evidenciar un breve resumen de cómo se encuentra el dataset entregado

Tabla 2

Resumen Analítica Descriptiva Dataset

Name Columns	Value
motivo_ros	(73.1%) is highly imbalanced
tipologia	(61.7%) is highly imbalanced
delito_fuente	(55.7%) is highly imbalanced
tipo_cli	(75.3%) is highly imbalanced
sociedad_ccial_civ	(69.2%) is highly imbalanced
pais_nacim	(90.8%) is highly imbalanced
cv	(76.4%) is highly imbalanced
pais_origen_recursos1	(92.2%) is highly imbalanced
riesgo_pais_origen_de_recursos	(96.9%) is highly imbalanced
pais_residencia1	(98.6%) is highly imbalanced
riesgo_pais_residencia	(99.5%) is highly imbalanced
ctrl_terc	(52.3%) is highly imbalanced
frecuencia_total_anual_transada_operaciones_internacionales_	(66.9%) is highly imbalanced
segm_comercial	(6.6%) missing values, has 1685
tipo_cli	(2.9%) missing values, has 748
sociedad_ccial_civ	(96.3%) missing values, has 24553

ciiu	(21.4%) missing values, has 5457
riesgo_actividad_economica	(21.4%) missing values, has 5457
riesgo_ocupacion	(12.8%) missing values, has 3255
riesgo_ciudad_de_residencia	(3.4%) missing values, has 870
pais_nacim	(9.6%) missing values, has 2441
cv	(9.6%) missing values, has 2441
pais_origen_recursos1	(18.0%) missing values, has 4597
riesgo_pais_origen_de_recursos	(18.0%) missing values, has 4597
pais_residencia1	(3.6%) missing values, has 922
riesgo_pais_residencia	(3.6%) missing values, has 922
f_vinc	(2.9%) missing values, has 748
estado_cli	(3.2%) missing values, has 818
ctrl_terc	(2.9%) missing values, has 748
f_ingreso_lc	(73.5%) missing values, has 18730
desc_subcateg	(73.5%) missing values, has 18730
ing_mes	(2.9%) missing values, has 749
egresos_mes	(2.9%) missing values, has 749
Dataset has 1604	**(6.3%) duplicate rows**

Nota: Características iniciales del Dataset entregado

El analista de datos **Asignado** recomendó una herramienta **que utiliza el banco llamado** y_data_profiling, la cual es utilizada para generar informes detallados y completos sobre conjuntos de datos. Estos informes son altamente personalizables e incluyen una amplia gama de características, como estadísticas del conjunto de datos, distribución de valores, datos faltantes, uso de memoria, entre otros. Esta herramienta resulta muy útil para explorar y analizar eficazmente los datos **(pypi, 2024).**

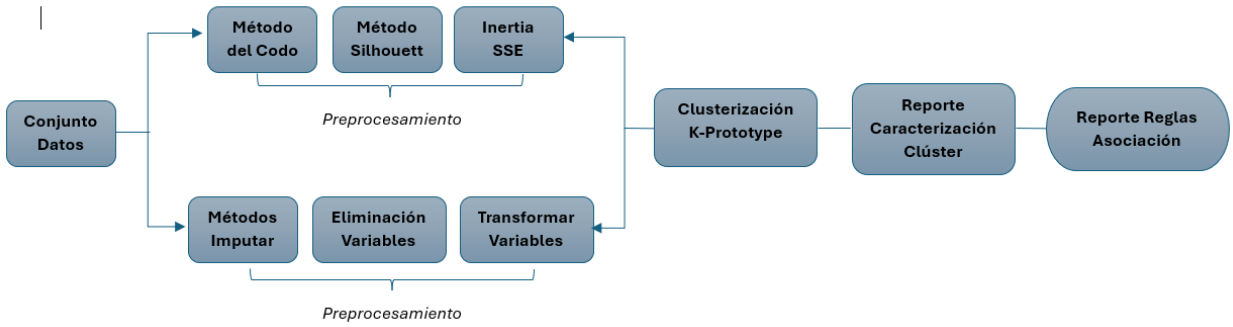
4. Proceso de analítica

4.1. Pipeline principal

A continuación, se relaciona la manera como se implementó el pipeline principal Utilizando la metodología CRISP-DM.

- Business Understanding, Se basa en una necesidad que actualmente presenta el Banco
- Data Understanding, Identificando como está compuesta identificando la relación de los Datos
- Data Preparation, Se realiza la limpieza y transformación de los Datos
- Modeling, se decide implementar el algoritmo de clusterización K-Prototype con el fin de segmentar de acuerdo con las características identificadas
- Evaluation, Se genera el reporte con las características representativas por cada clúster
- Deployment, Se comprueba con la generación de las reglas de asociación por cada Clúster.

Tabla 3
Flujo de Trabajo



Nota: Proceso a realizar durante el proyecto

4.2. Preprocesamiento

Se definieron diferentes métodos de preparación para el Dataset entre lo que podemos resaltar métodos para reducir la dimensionalidad

Tabla 4

Relación Variables que se Retiraron

Variable	**Columna Similar-Motivo**
nombre	Var Personal
num_caso	Id Caso
cod_ciiu	Igual ciiu
cod_ocup	Igual ocup
cod_ciudad_dirp	Igual nombre_ciudad_dirp
cod_pais_nacim	Igual pais_nacim
pais_origen_recursos	Igual pais_origen_recursos1
pais_residencia	Igual pais_residencia1
cod_categ_lc	Igual desc_categ
riesgo_cliente__ric_	Calificación Banco
cod_subcateg_lc	Igual desc_categ
cod_nivel_cat	Igual desc_subcateg
act_econom	Igual ciiu
desc_categ	Igual desc_subcateg
motivo_ingreso_a_listas_de_control	Igual desc_subcateg
ros	Igual motivo_ros
cod_tipo_doc	Relación Cercana a 0
f_vinc	Relación Cercana a 0
f_ingreso_lc	Relación Cercana a 0

Nota: Relación de variables que se retiraron debido a que se encontraron variables con mismos significados o baja Relación identificada con matriz de Pearson y variables codificadas.

Dado que un gran porcentaje de los datos presentaba valores faltantes y su tratamiento debía realizarse de manera independiente según el tipo de dato, se procedió imputar de la siguiente manera:

- Datos numéricos, Se utilizó la biblioteca Scikit-Learn y el módulo SimpleImputer para asignar el valor '0' a los valores faltantes. Esta decisión se tomó según las sugerencias del analista asignado al proyecto.
- Datos categóricos, También se utilizó Scikit-Learn y el módulo SimpleImputer para asignar la categoría 'Sin Información' a los valores faltantes, siguiendo la recomendación del analista de riesgos.

Además Se detectó que el personal encargado de registrar los datos había ingresado cadenas de texto en campos que debían contener valores numéricos. Este problema fue expuesto y corregido adecuadamente.

Debido a que se encontraron variables con datos numéricos dispersos, para este caso se identificaron seis columnas, lo que dificultaba su interpretación y para mejorar esto, se procedió a segmentar estos valores y se reclasificaron según su rango.

Se propuso aplicar técnicas para reducir la dimensionalidad del dataset que aplican para datos categóricos como fueron estos 2 modelos:

- **Matriz de Cramér's V**, Se implementó una matriz para comparar la asociación entre variables categóricas, El objetivo era evaluar la fuerza de la asociación entre las variables pero el resultado no fue óptimo, ya que no se obtuvieron sugerencias para eliminar (Analisis y Decisión Transformando Datos en decisiones, 2019).
- **Matriz de Theil's U**. Esta matriz utiliza el índice de incertidumbre de Theil, esta medida sirve para medir la asociación entre variables categóricas, esta medida distingue la asociación entre la primera y segunda variable. Al igual que la matriz anteriormente utilizada, esta tampoco identificó alguna columna que pudiera eliminarse (torchmetrics, 2023).

Al no obtener resultados relevantes con las técnicas mencionadas anteriormente para la eliminación de variables, se optó por transformar el conjunto de datos utilizando LabelEncoder de la librería scikit-learn. Permitiendo convertir los datos categóricos en valores numéricos, habilitando el uso de la Matriz de Correlación de Pearson. Finalmente se estableció un umbral de correlación de 0.5 para considerar una correlación como insignificante. Mediante este proceso, se identificaron 4 columnas que no aportaban valor significativo al modelo. Esta selección fue

validada por el personal de soporte, quienes confirmaron que dichas columnas eran redundantes, Sin embargo, la variable "doc" no fue eliminada, ya que se utilizará para cruzarlos con los valores reales, dado que los datos personales fueron entregados enmascarados

Tabla *5*
Matríx de Correlación de Pearson

4.3.Modelos

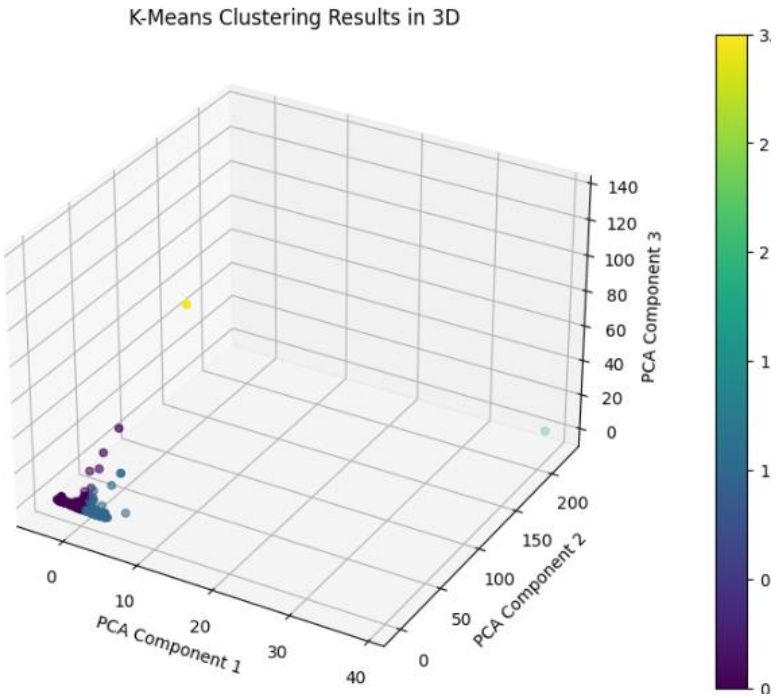
Al principio se realizaron varias pruebas en búsqueda del algoritmo que mejor se ajuste al tipo de datos entregado entre los que aplique se encontraban algoritmos como:

El primer modelo que se utilizó fue usando el algoritmo de clusterización ***Kmeans***, Por su gran popularidad debido a su simplicidad, eficiencia y facilidad de implementación, siendo particularmente útil en aplicaciones de segmentación, compresión de datos y análisis de patrones. En este caso, se observó un desbalanceo significativo en la asignación de los clústers, donde la

mayoría de los puntos se concentraron en solo dos grupos. Debido a esta situación, se optó por descartar el uso de este algoritmo (Anthony Barrios, 2023).

Figura 2

K-Means Clustering Resultados



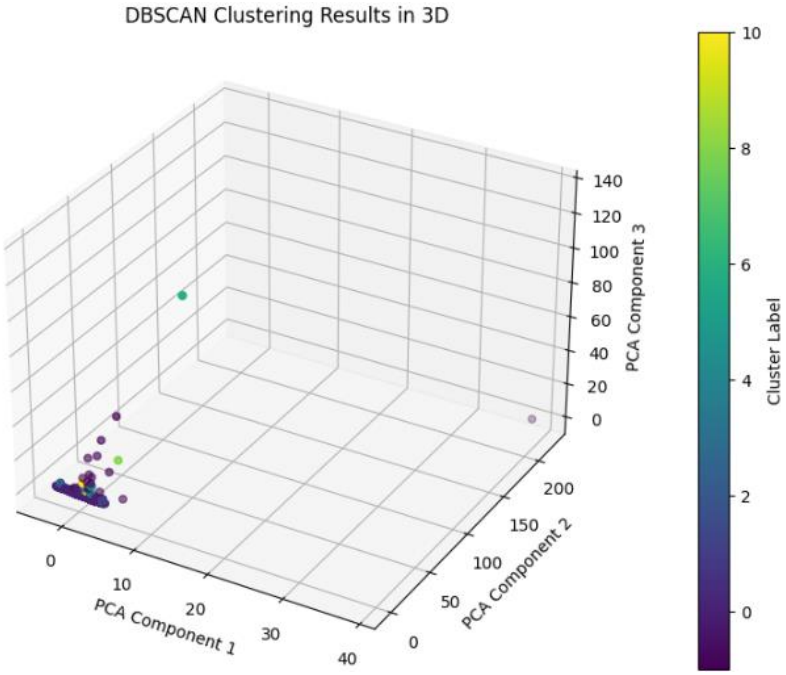
Nota: Resultado y gráfico en 3D utilizando el algoritmo Kmeans

En la búsqueda de un algoritmo más eficiente y que se ajustara a los datos opte por *Dbscan*, este algoritmo es adecuado para datos que forman clústers de formas arbitrarias. A diferencia de K-means este puede identificar clústers de cualquier forma, siempre que haya suficiente densidad de puntos, pero se obtenia el mismo comportamiento que kmeans al momento de Clústerizar asignando gran cantidad solo a 2 clusters y al graficarlos daba como resultado (Balaji C, 2023).

s

Figura 3

Resultado DBSCAN Clustering



Nota: Resultado y gráfico en 3D utilizando el algoritmo DBSCAN

El siguiente algoritmo que utilice fue ***Hdbscan***, debido a que supera diversas limitaciones de técnicas previas y aporta funcionalidades adicionales para el manejo de datos con densidades variables. Destaca su capacidad para identificar clústers más estables y robustos en la detección de anomalías. Sin embargo, al igual que DBSCAN, presenta un alto consumo de memoria bloqueando por completo el rendimiento de la computadora, motivo por el cual solo evidencio los parámetros que utilice, (Udit, 2022)

Figura 4

Parámetros Algoritmo HDBSCAN

```python
import hdbscan
import gower

# Calcular la matriz de distancias de Gower
gower_dist_matrix = gower.gower_matrix(DataFinal_)

# Aplicar HDBSCAN usando la matriz de distancias de Gower
hdbscan_clusterer = hdbscan.HDBSCAN(metric='precomputed')
labels = hdbscan_clusterer.fit_predict(gower_dist_matrix)

# Añadir los labels al dataframe original
DataFinal_['cluster'] = labels
DataFinal_.groupby(['Cluster']).size().sort_values(ascending=False)
```

Nota: parámetros Utilizados para el algoritmo HDBSCAN antes que se consumiera los recursos computacionales

Entre varias búsquedas de artículos y referenciados por otros científicos intente con el algoritmo ***Aglomerative Clustering***, Esta técnica de agrupamiento de datos brindó la capacidad de explorar la estructura jerárquica de los datos y detectar clústers de forma arbitraria, sin la necesidad de predefinir un número específico de clústers. Sin embargo, es importante destacar que su implementación presenta un alto costo computacional, al igual que el modelo anterior solo puedo evidenciar los parámetros que utilice debido a que bloqueaba mi computadora y no finalizaba la tarea (Harshit Dawar, 2020).

Figura 5

Parámetros Aglomerative Clustering

```
from sklearn.cluster import AgglomerativeClustering
import gower

# Calcular la matriz de distancias de Gower
gower_dist_matrix = gower.gower_matrix(DataFinal_)

# Aplicar Agglomerative Clustering usando la matriz de distancias de Gower
agg_clustering = AgglomerativeClustering(n_clusters=5, affinity='precomputed', linkage='average')
labels = agg_clustering.fit_predict(gower_dist_matrix)

# Añadir los labels al dataframe original
DataFinal_['cluster'] = labels
DataFinal_.groupby(['Cluster']).size().sort_values(ascending=False)
```

Nota: parámetros Utilizados para el algoritmo Aglomerative Clustering antes que se consumiera los recursos computacionales

Finalmente decidí Utilizar el algoritmo ***K-Prototype*** Este algoritmo es una extensión de k-means creada para trabajar datos mixtos además de que utiliza validación cruzada para asegurarse de que los resultados del clustering sean consistentes (Yashwanth Reddy, 2023).

4.4.Métricas

Las Métricas utilizadas son de vital importancia ya que lo que se pretende es evaluar la calidad de los clústers, que por finalidad proporcionan información valiosa sobre la cohesión interna y la separación entre los clústers, lo que permite determinar si los clústers identificados son realmente representativos de los datos y si existen patrones distintivos que pueden ser explotados para generar reglas de asociación útiles.

El análisis de estas métricas proporcionó una visión más completa de la calidad de los clústers generados por el modelo *K-Prototypes*, permitiendo evaluar su precisión y efectividad. Para validar estos resultados, se realizó una verificación basada en estos parámetros:

- **Coeficiente de Silhouette:** el silhouette_score más alto es 0.060598476 para la configuración de 6 clústers, init='Huang', n_init=10, max_iter=150, verbose=1. Este valor sugiere que los clústeres están medianamente definidos. (**Concepto)**
- **Inertia:** la configuración con la menor Inertia es 25897.55343 para 6 clústers, init='Huang', n_init=15, max_iter=100, verbose=1, indicando que los clústeres formados son los más compactos en esta configuración. (**Concepto)**
- **Davies-Bouldin Score:** el davies_bouldin_score más bajo es 3.324735479 para la configuración de 6 clústers, init='Huang', n_init=10, max_iter=150, verbose=1, indicando que esta configuración tiene los clústeres más compactos y mejor separados (**Concepto)**
- **Calinski-Harabasz Score:** el calinski_harabasz_score más alto es 1435.224414 para la configuración de 6 clústers, init='Huang', n_init=15, max_iter=100, verbose=1, concluyendo que esta configuración tiene los clústeres más bien definidos y separados (**Concepto)**

Para una mejor comprensión de las Métricas se representó en esta gráfica:

Tabla 6
Métricas de Desempeño

Parámetros					Métricas Desempeño			
clústers	init	n_init	max_iter	verbose	silhouette_score	Inertia	davies_bouldin_score	calinski_harabasz_score
6	**Huang**	**10**	**150**	**1**	**0,060598476**	**26089,12239**	**3,324735479**	**1432,295611**
6	**Huang**	**15**	**100**	**1**	**0,051381537**	**25897,55343**	**3,438233816**	**1435,224414**
6	Cao	10	150	1	0,055133775	26278,31516	3,457279769	1413,046944
6	Cao	15	100	1	0,052874286	26196,41436	3,559521397	1416,04021

Nota: Parametrización y comparación de la métrica de desempeño obtenida en cada iteración

Las configuraciones (n_clusters=num_clusters, init='Huang', n_init=10, max_iter=150, verbose=1) y (n_clusters=num_clusters, init='Huang', n_init=15, max_iter=100, verbose=1) se destacan cada una en diferentes métricas:

La primera parametrización se logra observar que tiene el mejor “silhouette_score” y “davies_bouldin_score”, y también tiene un buen “calinski_harabasz_score”.

La segunda configuración tiene la menor Inertia y el mejor “calinski_harabasz_score”.

Dado que la primera configuración tiene un mejor “silhouette_score” y “davies_bouldin_score”, que son métricas clave para evaluar la separación y cohesión de los clústers y es lo que busco para mi modelo, la configuración (n_clusters=num_clusters, init='Huang', n_init=10, max_iter=150, verbose=1) es considerada la mejor opción en este caso.

Para la generación de las reglas de Asociación entre las iteraciones que se realizaron para la búsqueda de los mejores parámetros se realizaron tomando en cuenta estos valores que me definían lo siguiente para el:

Segmento 1: Lift: 1.5, Leverage: 0.02, Conviction: 1.3 donde

- **Lift** 1.5: Este segmento tienen una fuerte asociación, ocurriendo 1.5 veces más a menudo que si fueran independientes.
- **Leverage** 0.02: Hay una pequeña pero positiva diferencia en la co-ocurrencia observada de los elementos en las reglas comparado con lo esperado si fueran independientes.
- **Conviction** 1.3: Las reglas sugieren que el antecedente predice el consecuente con una probabilidad razonable, aunque no es extremadamente fuerte.

Segmento 2: Lift: 1.2, Leverage: 0.01, Conviction: 1.1 donde,

- **Lift** 1.2: Aquí hay una moderada asociación entre los elementos de las reglas.
- **Leverage** 0.01: Las reglas tienen una co-ocurrencia observada ligeramente superior a lo esperado por independencia.
- **Conviction** 1.1: El antecedente predice el consecuente con una ligera confianza.

Estos valores tomados como ejemplo para evaluar la calidad de las reglas de asociación generadas para cada segmento, lo que me permitió entender mejor la precisión y efectividad del modelo FP-Growth en el análisis de riesgo y segmentación de los clientes.

En la búsqueda de los mejores parámetros tuve en cuenta los siguiente:

- Identificar los parámetros clave a ajustar, como **min_support** y **min_threshold**.
- Utilizar técnicas como la búsqueda en cuadrícula (*Grid Search*) para probar combinaciones de parámetros.
- Evaluar cada combinación utilizando las métricas de rendimiento (**Lift**, **Leverage**, **Conviction**)

Por ello se creó una función para identificar los mejores parámetros dando como resultado lo siguiente:

Resultado de la función, me está demorando más de lo normal

5. Metodología

5.1. Baseline

Luego de realizar varias pruebas con diferentes modelos y observando que se obtuvo una mejor distribución usando el algoritmo K-Prototype, Se decidió usar este modelo por el comportamiento que tuvo para manejar grandes volúmenes de datos además de la variedad de ítems por variable que tiene el dataset entregado, luego de definir el algoritmo que mejor comportamiento obtenido, se procedió a buscar los mejores parámetros que consistían en:

- n_clusters, correspondía al número de clúster basado en las métricas mencionadas en el capítulo **1.3 Métricas de desempeño** y basándome en el método del codo y el método de Silhouette el valor óptimo de k es 6 Clústers
- init, utilizado para inicializar los centroides de los clústers se usaron 'Huang' y 'Cao', luego de varias iteraciones se identifico que mi modelo se comportaba mejor con 'Huang'
- max_iter, este parámetro que define el número máximo de iteraciones del algoritmo de clustering use entre 100 y 150 iteraciones, para mi caso el valor optimo fue 150
- n_init, este parámetro define el número de veces que el algoritmo se ejecuta con diferentes inicializaciones de centroides se usó valores entre 10 y 15, el mejor valor se obtuvo con 10.
- verbose, este parámetro define para establecer la semilla del generador de números aleatorios siempre se usó el número 1.

En conclusión, se logró un mejor rendimiento del algoritmo de clusterización usando K-Prototype al iterar con estos parámetros, lo que también dio como resultado una mejor optimización del modelo.

5.2. Validación

Luego de haber encontrado los mejores parámetros donde mi modelo era eficiente y eficaz procedió a validar el desempeño del resultado de mi modelo, para ello existen métricas que evalúan el rendimiento del modelo realizo como lo es:

- **Inertia (Costo Total o Suma de los Cuadrados de las Distancias):** Es la suma de las distancias cuadradas entre cada punto y el centroide del clúster al que pertenece.
- **Davies-Bouldin Index:** Mide la dispersión dentro de los clústers y la separación entre los clústers. Es el promedio de la razón entre la dispersión dentro del clúster y la distancia entre los clústers
- **Calinski-Harabasz Index (Variance Ratio Criterion):** Ratio de la suma de la dispersión entre clústers a la suma de la dispersión dentro de los clústers,

Figura 6

Métricas Rendimiento Algoritmo K-Prototype

Métrica	Resultado
Inertia	257084.16347774977
Davies-Bouldin	2.6463596112159715
Calinski-Harabasz	1681.3904643192116

Nota: Resultado Obtenido para medir el rendimiento del algoritmo K-Prototype

5.3.Iteraciones y evolución

Durante la ejecución del algoritmo K-Prototype, encontrar opciones de preprocesamiento adecuadas resultó ser un desafío debido a su complejidad. Aunque el comportamiento del algoritmo era adecuado, el proceso de generación de reglas de asociación se observó un alto consumo de recursos computacional. Para abordar esto, se implementaron varias transformaciones, que incluyeron: 1) Optimización del tamaño de los segmentos, 2) Mejora del paralelismo para aprovechar todos los núcleos de la CPU disponibles, y 3) Aplicación de filtros de reglas mediante la parametrización de min_support y min_threshold para identificar reglas más sólidas.

5.4. Herramientas

Para facilitar el desarrollo del proyecto, se emplearon diversas herramientas que optimizaron el proceso y permitieron un trabajo colaborativo eficiente. Entre las principales herramientas se encuentran:

5.4.1. Jupyter Notebook: Esta aplicación web popular en la comunidad de ciencia de datos y programación permite crear documentos que integran código en tiempo real, visualizaciones y texto narrativo. Nos facilita la comprensión y el análisis de los datos.

5.4.2. Google Collab: Es una herramienta valiosa para el desarrollo de proyectos de Machine Learning, especialmente aquellos que requieren un alto poder de procesamiento y colaboración entre usuarios

5.4.3. Scikit-learn: Esta biblioteca de machine learning en Python brinda herramientas simples y eficientes para el análisis de datos y la modelización predictiva. Incluye una amplia gama de algoritmos para tareas como clasificación, regresión, clustering, reducción de dimensionalidad y muchas más.

5.4.4. Git: Este sistema de control de versiones permite realizar un seguimiento de los cambios en el código y gestionar diferentes versiones del mismo. Permite que cada desarrollador puede trabajar en su propia copia y posteriormente sincronizar los cambios con el repositorio central.

5.4.5. GitHub: Esta plataforma de alojamiento de repositorios en la nube facilita la colaboración y el control de versiones en proyectos. Posee una interfaz intuitiva y sus funcionalidades colaborativas la convierten en una herramienta esencial para el desarrollo en equipo.

Figura 7

Características de hardware y Software Utilizado en Collab

```
Platform processor: x86_64
Platform architecture: ('64bit', 'ELF')
Machine type: x86_64
'System's network name: 7766920ddebe
Platform information: Linux-6.1.85+-x86_64-with-glibc2.35
Operating system: Linux
System info: Linux
Python build no. and date: ('main', 'Nov 20 2023 15:14:05')
Python compiler: GCC 11.4.0
Python SCM: GCC 11.4.0
Python implementation: CPython
Python version: 3.10.12
```

Nota, Se dio por inicio en entorno Collab buscando entornos de ejecución con mejores características

Figura 8

Características de hardware y Software Utilizado en Jupyter

```
Platform processor: Intel64 Family 6 Model 142 Stepping 9, GenuineIntel
Platform architecture: ('64bit', 'WindowsPE')
Machine type: AMD64
'System's network name:' CesarS
Platform information: Windows-10-10.0.19045-SP0
Operating system: Windows
System info: Windows
Python build no. and date: ('tags/v3.12.2:6abddd9', 'Feb  6 2024 21:26:36')
Python compiler: MSC v.1937 64 bit (AMD64)
Python SCM: MSC v.1937 64 bit (AMD64)
Python implementation: CPython
Python version: 3.12.2
```

Nota: Por políticas del banco y buscando uniformidad en la entrega del proyecto se cambió el Ide a Jupyter Notebook

6. Resultados y discusión

Los resultados iniciales mostraron clústers muy desbalanceados debido a la complejidad de los datos. Para abordar esto, se realizaron transformaciones de los datos a valores numéricos utilizando técnicas de procesamiento de texto: se tokenizó, se aplicaron técnicas de lematización y se eliminaron stopwords. Posteriormente, se utilizó K-means, lo que resultó nuevamente en un desbalanceo de los datos, como se observa en la **Figura 7**. El objetivo era encontrar la técnica más óptima para realizar la clusterización.

También se exploraron técnicas para reducir la dimensionalidad, como PCA. Sin embargo, esto resultó en una asignación de clústers muy difícil de interpretar, como se muestra en la Figura 8. Se aplicaron técnicas de Hierarchical Clustering y se representaron con dendrogramas, pero estos resultaron ser muy confusos **Figura 9.**

Los resultados iniciales mostraron un desbalance en los clústers. Esto se debió, en parte, a las técnicas de imputación utilizadas. Para las variables numéricas, se empleó KNNImputer con los parámetros n_neighbors=5 y weights="uniform", seguido de una escalación con RobustScaler. Para las variables categóricas, se imputaron los valores faltantes utilizando la estrategia 'most_frequent', la cual no resultó muy satisfactoria en este caso.

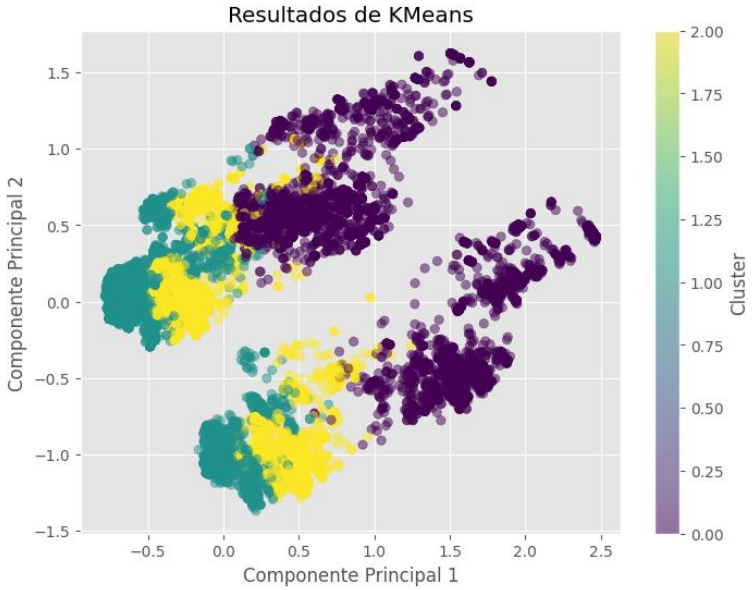
Luego de probar los modelos con los parámetros mencionados en el apartado **4.3 Modelos** y tras varias iteraciones, como se ilustró en el capítulo **4.4 Métricas**, se definió el modelo de clusterización K-Prototype. Después de aplicar todos los ajustes descritos, se obtuvo el resultado final mostrado en la **Figura 10.**

A solicitud del analista de riesgo encargado, se generó un reporte con los indicadores característicos de cada variable en cada clúster para evaluar sus características. Este reporte fue exportado a un archivo de texto para su análisis. **Figura 11.**

Se procedió a generar las reglas de asociación basadas en la clusterización. Inicialmente, se encontraron problemas de rendimiento debido a la alta demanda de recursos del algoritmo. Sin embargo, ajustando los parámetros, como se menciona en el apartado **5.3 Iteraciones y evolución**, se logró una generación de las reglas de asociación de manera más eficiente y distribuida **Figura 12**.

Figura 9

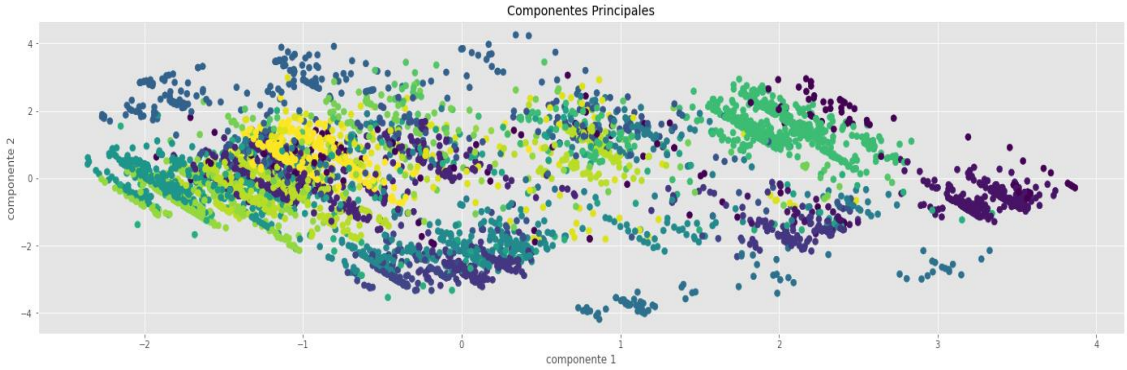
Algoritmo Kmeans



Nota: Esta técnica usando Embedings fue muy compleja y no lograba obtener resultados óptimos

Figura 10

Reducción de la Dimensionalidad PCA



Nota: Esta técnica daba resultados muy complejos para interpretar, por ende se descartó.

Figura 11

Distancia euclídea, Link Average

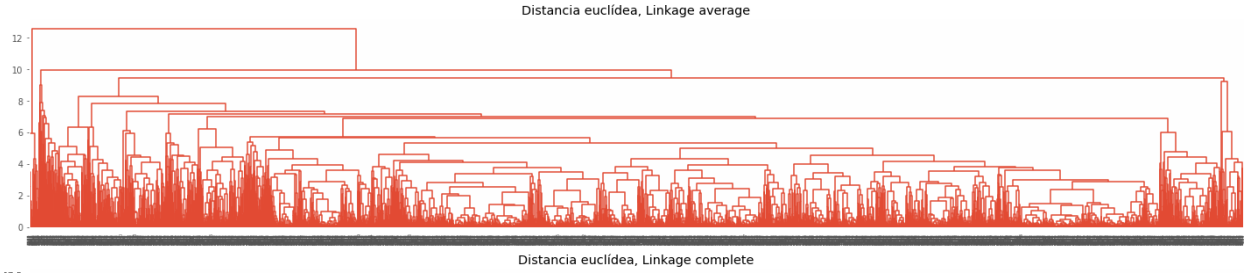


Figura 12

Resultado Final Clústerizar K-Prototypes

```
#Agrupo por cluster y cuento cuantos registros asigno a cada cluster
DatOrig.groupby('cluster').cluster.count().sort_values(ascending=False)
```

```
cluster
1    5204
0    5183
3    5059
2    5022
4    5021
Name: cluster, dtype: int64
```

Nota: Resultado Obtenido después de Clusterizar con k-Prototypes

Figura 13

Reporte Caracterización por Clúster

```
Resultados del Cluster 'Cluster_02':
De la columna 'motivo_ros': El 67.03% de los Registros es 'Posible Operacion LA' Con un total de 3366 Items
De la columna 'tipologia': El 51.75% de los Registros es 'Posible Operacion LA' Con un total de 2599 Items
De la columna 'delito_fuente': El 46.30% de los Registros es 'Posible Operacion LA' Con un total de 2325 Items
De la columna 'segm_comercial': El 44.11% de los Registros es 'PERSONAL' Con un total de 2215 Items
De la columna 'tipo_cli': El 92.91% de los Registros es 'PERSONA NATURAL' Con un total de 4666 Items
De la columna 'sociedad_ccial_civ': El 96.08% de los Registros es 'Sin Informacion' Con un total de 4825 Items
De la columna 'ciiu': El 46.59% de los Registros es 'ASALARIADOS' Con un total de 2340 Items
De la columna 'riesgo_actividad_economica': El 62.58% de los Registros es 'BAJO' Con un total de 3143 Items
De la columna 'ocup': El 34.79% de los Registros es 'EMPLEADO' Con un total de 1747 Items
De la columna 'riesgo_ocupacion': El 53.50% de los Registros es 'Medio' Con un total de 2687 Items
De la columna 'nombre_ciudad_dirp': El 10.97% de los Registros es '7bb20f00' Con un total de 551 Items
De la columna 'riesgo_ciudad_de_residencia': El 65.17% de los Registros es 'ALTO' Con un total de 3273 Items
De la columna 'pais_nacim': El 83.11% de los Registros es 'COLOMBIA' Con un total de 4174 Items
De la columna 'cv': El 83.41% de los Registros es 'MEDIO' Con un total de 4189 Items
De la columna 'pais_origen_recursos1': El 79.29% de los Registros es 'COLOMBIA' Con un total de 3982 Items
De la columna 'riesgo_pais_origen_de_recursos': El 82.20% de los Registros es 'MEDIO' Con un total de 4128 Items
De la columna 'pais_residencia1': El 96.22% de los Registros es 'COLOMBIA' Con un total de 4832 Items
De la columna 'riesgo_pais_residencia': El 96.71% de los Registros es 'MEDIO' Con un total de 4857 Items
De la columna 'estado_cli': El 54.00% de los Registros es 'ACTIVO' Con un total de 2712 Items
De la columna 'ctrl_terc': El 57.79% de los Registros es 'CLIENTE' Con un total de 2902 Items
De la columna 'desc_subcateg': El 72.86% de los Registros es 'Sin Informacion' Con un total de 3659 Items
De la columna 'ing_mes': El 36.36% de los Registros es 'Rango entre 0.0 y 1192000.0' Con un total de 1826 Items
De la columna 'egresos_mes': El 61.11% de los Registros es 'Rango entre 0.0 y 713411.0' Con un total de 3069 Items
De la columna 'monto_total_anual_transado__efectivo_': El 59.92% de los Registros es 'Rango entre 0.0 y 9132000.0' Con un total de 3009
Items
De la columna 'monto_total_anual_transado__operaciones_internacionales_': El 75.45% de los Registros es 'Rango entre 0.0 y 25987828.1782'
Con un total de 3789 Items
De la columna 'frecuencia_total_anual_transada__operaciones_internacionales_': El 88.21% de los Registros es 'Rango entre 0.0 y 80.0' Con un
total de 4430 Items
De la columna 'frecuencia_total_anual_transada__efectivo_': El 86.94% de los Registros es 'Rango entre 0.0 y 94.0' Con un total de 4366
Items
--------------------------------------------------
```

Nota: Resultado obtenido al analizar el clúster 02

6.1. Métricas

Luego de varias iteraciones y buscando los mejores parámetros relacionados en la **Tabla 5** se concluyó que los mejores resultado se obtenían basados en estos valores (n_clusters=5, init='Cao', n_init=10, max_iter=150, verbose=1) ya que ofrecía un mejor rendimiento, Para el algoritmo de reglas de asociación los mejores parámetros identificados fueron (min_support= 0.1 y min_threshold= 1.0) este con el fin de establecer la asociación perfecta al generar las reglas de asociación dando como resultado lo demostrado en la **Figura 12**.

6.2. Evaluación cualitativa

Para evaluar la precisión del modelo de clusterización, luego de exportar las características más representativas por cada variable dentro de cada clúster, como se muestra en la **Figura 11** tomando como ejemplo el clúster número 02. Este análisis permite al analista Identificar las tendencias presentes por cada grupo de clúster y determinar la tendencia, confiabilidad y relevancia de las reglas dentro de cada segmento representativo en cada clúster.

Es esencial asegurarse de que los clústers sean adecuados para los datos utilizados, evitando que sean demasiado específicos, lo cual podría indicar un problema de sobreajuste (overfitting). Este análisis implica revisar detalladamente los resultados obtenidos para cada clúster.

Por otro lado, si los clústers no logran capturar de manera efectiva la estructura subyacente de los datos, podría haber un problema de sobreajuste (underfitting). Este análisis lo realiza el analista de riesgos al examinar el resultado de la extracción de características de cada clúster.

Es importante destacar que, para demostrar este análisis, el analista de riesgos sugiere hacer referencia a los pasos realizados en lugar de mostrar directamente los resultados obtenidos. Esto se debe a que el banco ha establecido que los resultados de los clientes no pueden ser mostrados directamente.

6.3. Consideraciones de producción

Es importante destacar que este modelo tiene como objetivo realizar una validación adicional para generar calificaciones más precisas y con un propósito específico. Por lo tanto, es crucial que un analista de riesgos revise los resultados antes de integrarlos al modelo de calificación en el ROS. Esta revisión permitirá realizar ajustes al modelo o agregar técnicas que mejoren su efectividad.

Para asegurar que el algoritmo pueda ser implementado en producción, se cumplieron todos los requisitos mínimos establecidos por el Grupo Bancolombia. Esto incluyó el uso de las librerías recomendadas por el banco y el IDE de desarrollo permitido, que en este caso es Jupyter Notebook. Además, se aplicaron todas las recomendaciones proporcionadas por el analista de datos durante la etapa final del proyecto. Por último, el analista de datos sugirió varias modificaciones en el script para adaptar el algoritmo a las condiciones específicas requeridas.

7. Conclusiones

El proyecto logró cumplir con los objetivos establecidos y abordar el problema identificado de manera efectiva. La limpieza y transformación de los datos, junto con la implementación de algoritmos avanzados de clasificación y asociación, contribuyeron significativamente a mejorar la calidad y utilidad del dataset. Los resultados obtenidos demostraron la viabilidad y eficacia de la segmentación precisa y la identificación de patrones de riesgo, lo que facilita una gestión más eficiente del riesgo LA/FT.

Sin embargo, el proyecto enfrentó algunos desafíos, como la complejidad en la imputación de datos y la necesidad de equilibrar la reducción de dimensionalidad con la preservación de información relevante. Estos desafíos subrayan la importancia continua de desarrollar y aplicar técnicas avanzadas de análisis de datos para mejorar la precisión y eficacia de los modelos.

En resumen, el proyecto proporciona una base sólida para futuras aplicaciones en la gestión del riesgo LA/FT, destacando la importancia crítica de la calidad de los datos y la robustez de los modelos analíticos en la toma de decisiones informadas.

8. Recomendaciones

Para mejorar la calidad y precisión en la gestión del riesgo LA/FT, es recomendable emplear métodos automatizados para la detección y corrección de valores anómalos al ingresar los datos. Además, se sugiere implementar diversas técnicas avanzadas de imputación y limpieza de datos. Explorar métodos adicionales de reducción de dimensionalidad, como la Selección de Características Basada en Importancia, puede ayudar a conservar características relevantes sin perder información crucial.

Además, proponer nuevos algoritmos de clusterización y comparar su desempeño con K-Prototypes podría optimizar la segmentación de clientes. También sería beneficioso investigar algoritmos híbridos para la generación de reglas de asociación, lo que mejoraría la extracción de patrones frecuentes y proporcionaría insights más valiosos. La integración de modelos predictivos basados en Deep Learning permitiría anticipar comportamientos y riesgos de los clientes con mayor precisión.

Evaluar la viabilidad de implementar sistemas de análisis y detección de patrones en tiempo real utilizando tecnologías de Big Data podría mejorar la respuesta y vigilancia ante comportamientos sospechosos. Por último, realizar estudios que integren múltiples fuentes de datos, como análisis de redes geoespaciales, ofrecerá una visión más completa del comportamiento del cliente y los riesgos asociados.

Estas recomendaciones están dirigidas a mejorar la precisión y eficacia en la gestión del riesgo LA/FT mediante el desarrollo y aplicación de técnicas avanzadas de análisis de datos.

Referencias

Ali, A. (2019, 02 03). *Medium*. Retrieved from Medium: https://medium.com/machine-learning-researcher/association-rule-apriori-and-eclat-algorithm-4e963fa972a4

Analisis y Decisión Transformando Datos en decisiones. (2019, 07 16). *Gráfico de correlaciones entre factores. Gráfico de la V de Cramer*. Retrieved 06 02, 2024, from https://analisisydecision.es/grafico-de-correlaciones-entre-factores-grafico-de-la-v-de-cramer/

Anthony Barrios. (2023, 08 08). *Tutorial del Algoritmo K-Means en Python*. Retrieved from Medium: https://medium.com/latinxinai/tutorial-del-algoritmo-k-means-en-python-d8055751e2f3

Balaji C. (2023, 11 11). *DBSCAN Clustering*. Retrieved from Medium.com: https://medium.com/@balajicena1995/dbscan-clustering-2a577d384e61#:~:text=Density%2Dbased%20spatial%20clustering%20of%20applications%20with%20noise(DBSCAN),with%20minimum%20size%20and%20density.

Chauhan, N. S. (2023, 09 25). *www.datasource.ai*. Retrieved from Métricas De Evaluación De Modelos En El Aprendizaje Automático: https://www.datasource.ai/es/data-science-articles/metricas-de-evaluacion-de-modelos-en-el-aprendizaje-automatico

Compliance. (2024). *Compliance Sistema de Información*. Recuperado el 20 de 05 de 2024, de Compliance.com.co: https://www.compliance.com.co/el-impacto-economico-y-reputacional-del-lavado-de-activos/#:~:text=Impacto%20en%20la%20sociedad%3A%20Puede,la%20democracia%20engendrando%20violencia%20interna.

Harshit Dawar. (2020, 05 30). *Deep dive Agglomerative Clustering!* Retrieved from Medium: https://medium.com/analytics-vidhya/deep-dive-agglomerative-clustering-e9af2bfd8daf

Koehrsen, W. (2018, 10). *towardsdatascience.com*. Retrieved from Medium: https://towardsdatascience.com/neural-network-embeddings-explained-4d028e6f0526

Mahmood, M. S. (2021, 07 11). *towardsdatascience.com*. Retrieved 06 01, 2024, from Medium: https://towardsdatascience.com/factor-analysis-of-mixed-data-5ad5ce98663c

pypi. (2024, 05 07). *ydata-profiling 4.8.3*. Retrieved from pypi.org: https://pypi.org/project/ydata-profiling/

Reddy, Y. (2023, 04 10). *Medium*. Retrieved from Medium: https://medium.com/@reddyyashu20/k-means-kmodes-and-k-prototype-76537d84a669

Rodriguez, D. (2023, 06 23). *www.analyticslane.com*. Retrieved from Analitics Lane: https://www.analyticslane.com/2023/06/23/numero-optimo-de-clusteres-con-silhouette-e-implementacion-en-python/

Rodriguez, D. (2023, 06 30). *www.analyticslane.com*. Retrieved from El índice de Davies-Bouldinen para estimar los clústeres en k-means e implementación en Python: https://www.analyticslane.com/2023/06/30/el-indice-de-davies-bouldinen-para-estimar-los-clusteres-en-k-means-e-implementacion-en-python/

Rodriguez, D. (2023, 06 16). *www.analyticslane.com*. Retrieved from Identificar el número de clústeres con Calinski-Harabasz en k-means e implementación en Python: https://www.analyticslane.com/2023/06/16/identificar-el-numero-de-clusteres-con-calinski-harabasz-en-k-means-e-implementacion-en-python/

RStudio, R. b., & Ayala, J. (2020, 10 02). *Minería de datos, Reducción de dimensionalidad*. Retrieved from https://rpubs.com/JairoAyala/574796

Superintendencia Financiera de Colombia. (2003, 09 15). *LAVADO DE ACTIVOS*. Retrieved from www.superfinanciera.gov.co: https://www.superfinanciera.gov.co/publicaciones/18899/normativaconceptos-y-jurisprudencia-conceptoshistorico-doctrina-y-conceptos-anteriores-superintendencias-bancaria-y-de-valores-doctrinas-y-conceptos-financieros-indice-generallavado-de-activos-18899/

Superintendencia Financiera de Colombia. (2019, 11 27). *Circular Externa 027 - Superintendencia Financiera de Colombia*. Retrieved from www.superfinanciera.gov.co: https://www.google.com/url?client=internal-element-cse&cx=012067830255839863893:xs8mx2mrbzy&q=https://www.superfinanciera.gov.co/loader.php%3FlServicio%3DTools2%26lTipo%3Ddescargas%26lFuncion%3Ddescargar%26idFile%3D1041184&sa=U&ved=2ahUKEwiNzKiY_s-GAxUJTD

torchmetrics. (2023). Retrieved from https://torchmetrics.readthedocs.io/en/v0.11.3/nominal/theils_u.html

Udit. (2022, 12 30). *Discovering the Power of HDBSCAN Clustering for Unsupervised Learning*. Retrieved from Medium.com: https://itsudit.medium.com/discovering-the-power-of-hdbscan-clustering-for-unsupervised-learning-d67273e28c5b

UIAF. (2014). *Ministerio Hacienda*. Retrieved 06 01, 2024, from https://uiaf.gov.co/sites/default/files/2022-06/documentos/archivos-anexos/Lo%20que%20debe%20saber%20sobre%20LAFT-1.pdf

Yashwanth Reddy. (2023, 04 10). *K-means, kmodes, and k-prototype*. Retrieved from Medium.com: https://medium.com/@reddyyashu20/k-means-kmodes-and-k-prototype-76537d84a669

Anexos

Ali, A. (03 de 02 de 2019). *Medium*. Obtido de Medium: https://medium.com/machine-learning-researcher/association-rule-apriori-and-eclat-algorithm-4e963fa972a4

Analisis y Decisión Transformando Datos en decisiones. (16 de 07 de 2019). *Gráfico de correlaciones entre factores. Gráfico de la V de Cramer*. Obtido em 02 de 06 de 2024, de https://analisisydecision.es/grafico-de-correlaciones-entre-factores-grafico-de-la-v-de-cramer/

Anthony Barrios. (08 de 08 de 2023). *Tutorial del Algoritmo K-Means en Python*. Obtido de Medium: https://medium.com/latinxinai/tutorial-del-algoritmo-k-means-en-python-d8055751e2f3

Balaji C. (11 de 11 de 2023). *DBSCAN Clustering*. Obtido de Medium.com: https://medium.com/@balajicena1995/dbscan-clustering-2a577d384e61#:~:text=Density%2Dbased%20spatial%20clustering%20of%20applications%20with%20noise(DBSCAN),with%20minimum%20size%20and%20density.

Chauhan, N. S. (25 de 09 de 2023). *www.datasource.ai*. Obtido de Métricas De Evaluación De Modelos En El Aprendizaje Automático: https://www.datasource.ai/es/data-science-articles/metricas-de-evaluacion-de-modelos-en-el-aprendizaje-automatico

Compliance. (2024). *Compliance Sistema de Información*. Recuperado el 20 de 05 de 2024, de Compliance.com.co: https://www.compliance.com.co/el-impacto-economico-y-reputacional-del-lavado-de-activos/#:~:text=Impacto%20en%20la%20sociedad%3A%20Puede,la%20democracia%20engendrando%20violencia%20interna.

Harshit Dawar. (30 de 05 de 2020). *Deep dive Agglomerative Clustering!* Obtido de Medium: https://medium.com/analytics-vidhya/deep-dive-agglomerative-clustering-e9af2bfd8daf

Koehrsen, W. (10 de 2018). *towardsdatascience.com*. Obtido de Medium: https://towardsdatascience.com/neural-network-embeddings-explained-4d028e6f0526

Mahmood, M. S. (11 de 07 de 2021). *towardsdatascience.com*. Obtido em 01 de 06 de 2024, de Medium: https://towardsdatascience.com/factor-analysis-of-mixed-data-5ad5ce98663c

pypi. (07 de 05 de 2024). *ydata-profiling 4.8.3*. Obtido de pypi.org: https://pypi.org/project/ydata-profiling/

Reddy, Y. (10 de 04 de 2023). *Medium*. Obtido de Medium: https://medium.com/@reddyyashu20/k-means-kmodes-and-k-prototype-76537d84a669

Rodriguez, D. (23 de 06 de 2023). *www.analyticslane.com*. Obtido de Analitics Lane: https://www.analyticslane.com/2023/06/23/numero-optimo-de-clusteres-con-silhouette-e-implementacion-en-python/

Rodriguez, D. (30 de 06 de 2023). *www.analyticslane.com*. Obtido de El índice de Davies-Bouldinen para estimar los clústeres en k-means e implementación en Python: https://www.analyticslane.com/2023/06/30/el-indice-de-davies-bouldinen-para-estimar-los-clusteres-en-k-means-e-implementacion-en-python/

Rodriguez, D. (16 de 06 de 2023). *www.analyticslane.com*. Obtido de Identificar el número de clústeres con Calinski-Harabasz en k-means e implementación en Python: https://www.analyticslane.com/2023/06/16/identificar-el-numero-de-clusteres-con-calinski-harabasz-en-k-means-e-implementacion-en-python/

RStudio, R. b., & Ayala, J. (02 de 10 de 2020). *Minería de datos, Reducción de dimensionalidad*. Obtido de https://rpubs.com/JairoAyala/574796

Superintendencia Financiera de Colombia. (15 de 09 de 2003). *LAVADO DE ACTIVOS*. Obtido de www.superfinanciera.gov.co: https://www.superfinanciera.gov.co/publicaciones/18899/normativaconceptos-y-jurisprudencia-conceptoshistorico-doctrina-y-conceptos-anteriores-superintendencias-bancaria-y-de-valores-doctrinas-y-conceptos-financieros-indice-generallavado-de-activos-18899/

Superintendencia Financiera de Colombia. (27 de 11 de 2019). *Circular Externa 027 - Superintendencia Financiera de Colombia*. Obtido de www.superfinanciera.gov.co: https://www.google.com/url?client=internal-element-cse&cx=012067830255839863893:xs8mx2mrbzy&q=https://www.superfinanciera.gov.co/loader.php%3FlServicio%3DTools2%26lTipo%3Ddescargas%26lFuncion%3Ddescargar%26idFile%3D1041184&sa=U&ved=2ahUKEwiNzKiY_s-GAxUJTD

torchmetrics. (2023). Obtido de https://torchmetrics.readthedocs.io/en/v0.11.3/nominal/theils_u.html

Udit. (30 de 12 de 2022). *Discovering the Power of HDBSCAN Clustering for Unsupervised Learning*. Obtido de Medium.com: https://itsudit.medium.com/discovering-the-power-of-hdbscan-clustering-for-unsupervised-learning-d67273e28c5b

UIAF. (2014). *Ministerio Hacienda*. Obtido em 01 de 06 de 2024, de https://uiaf.gov.co/sites/default/files/2022-06/documentos/archivos-anexos/Lo%20que%20debe%20saber%20sobre%20LAFT-1.pdf

Yashwanth Reddy. (10 de 04 de 2023). *K-means, kmodes, and k-prototype*. Obtido de Medium.com: https://medium.com/@reddyyashu20/k-means-kmodes-and-k-prototype-76537d84a669

.