PERFILAMIENTO Y SEGMENTACIÓN DE CLIENTES POR CONSUMOS

-Estudiantes

Daniela Útima Lasso

Paula Alejandra Lock Osorio

Jorge Leonardo Gonzalez Aguirre

Jhonathan Campuzano

Yeisson Torres Guarin

-Profesores:

Natalia Betancur Herrera

Frank Yesid Zapata Cataño

Andrés Felipe Sánchez Cano

Universidad De Caldas – Bootcamp Inteligencia Artificial

Mayo del 202

Tabla de Contenido

Introducción

En el ámbito financiero, las instituciones y empresas manejan grandes volúmenes de datos provenientes de diversas fuentes, como transacciones, movimientos bursátiles, comportamientos de clientes y variables económicas. Sin embargo, la correcta interpretación de estos datos resulta un desafío debido a su naturaleza compleja y la falta de etiquetas claras en muchos casos. A menudo, los métodos tradicionales de análisis no son capaces de detectar patrones o comportamientos ocultos en la información, lo que puede llevar a decisiones subóptimas o pérdidas significativas. Los modelos de machine learning no supervisados se presentan como una solución efectiva, ya que permite descubrir patrones, relaciones y anomalías en los datos sin necesidad de contar con etiquetas predefinidas. Este enfoque es especialmente relevante en el contexto financiero, donde los datos etiquetados pueden ser escasos o costosos de obtener.

Objetivo general:

Desarrollar e implementar un modelo de machine learning, no supervisado, enfocado en perfilamiento y la segmentación de clientes por consumo, con el fin de descubrir patrones y relaciones significativas que puedan contribuir a una toma de decisiones más informada y estratégica en el ámbito financiero.

Objetivo específicos:

- Preparar los datos para su uso en un modelo no supervisado, incluyendo la limpieza de datos, la normalización y la transformación de variables.
- Implementar diferentes técnicas de machine learning no supervisado, aplicando algoritmos como el clustering (por ejemplo, K-means,) o análisis de componentes principales (PCA) para identificar patrones o agrupaciones en los datos.
- Evaluar los resultados obtenidos y analizar la eficacia de los modelos

aplicados mediante métricas relevantes, como la calidad de las agrupaciones

- **Proponer aplicaciones prácticas en el sector financiero:** Identificar cómo los patrones que pueden utilizarse en áreas clave del sector financiero, como la segmentación de clientes, la identificación de riesgos financieros o la optimización de carteras de inversión.

Alcance del proyecto:

Este proyecto se centrará en el desarrollo de un modelo de machine learning no supervisado utilizando un conjunto de datos financieros de acceso público. Se abordarán los siguientes aspectos:

- **Selección de algoritmos:** Se implementarán diferentes técnicas de aprendizaje no supervisado, incluyendo clustering y reducción de dimensionalidad, para explorar los datos de manera eficaz, el proyecto se enfocará en identificar patrones en los datos que puedan ayudar en el perfilamiento, la segmentación de clientes en la identificación de oportunidades de inversión. Posteriormente se evaluarán los resultados obtenidos de los modelos, se valorarán en términos de su capacidad para extraer insights útiles y su aplicabilidad en el entorno financiero real.
- **Limitaciones:** Debido a la disponibilidad de datos y recursos, el alcance del proyecto estará limitado a un conjunto específico de datos financieros, lo que puede influir en la generalización de los resultados. Además, no se realizarán implementaciones en tiempo real ni se integrarán estos modelos en sistemas operativos de instituciones financieras. Este proyecto se desarrollará durante el bootcamp talento Tech en las fechas 01 de abril al 31 de mayo.

Metodología: CRISP-DM

Comprensión del Negocio

- **Descripción del problema desde el punto de vista del negocio:**Las entidades bancarias almacenan la información del proceso transaccional e histórico de movimientos de cada cuenta, pero los productos financieros no están alineados a patrones de uso regionales lo que condiciona su efectividad y esta desconexión limita la capacidad de tomar decisiones estratégicas basadas en el comportamiento real del cliente.

Esta BBDD contiene Información sobre el uso de productos financieros y cobertura, incluyendo la desagregación por género en la tenencia de los diferentes productos financieros.

- **Objetivos del negocio:** Usar los datos almacenados por cada entidad para identificar patrones de uso por región, brindando a la entidad un objetivo claro para ubicación de sucursales, corresponsales o métodos transaccionales efectivos, según el perfil del usuario por zona geográfica.

- **Criterios de éxito**

1. Identificar con al menos una precisión del 80% ,las regiones con mayor volumen de transacción y los métodos de pago más utilizados.
2. Generar recomendaciones accionables en la red de sucursales, cajeros o canales digitales.
3. Reducir costos operativos como reubicación de infraestructura subutilizada, incrementando canales más eficientes al menos en el 5% de las regiones.
4. Mejorar segmentación geográfica, resultando en un aumento de la conversión o uso de los servicios financieros por zonas objetivo.

Comprensión de los Datos

La fuente de datos utilizada en este proyecto proviene de los sistemas internos de las entidades financieras. la data se extrajo de www.superfinanciera.gov.co, específicamente se emplean datos históricos que contienen información sobre movimientos transaccionales, estas bases incluyen:

- Datos transaccionales: fecha, monto, tipo de operación (retiro, depósito, transferencia de pago), canal utilizado, número de retiros.
- Ubicación geográfica: Departamento, municipio.
- Identificadores de cliente: Cuenta de ahorros por salarios (hasta 1 smmlv, hasta 3 smmlv y 5 smmlv) permite análisis sin comprometer la privacidad.
- Datos de productos financieros vinculados: Tipo de cuenta, producto asociado (ahorro, consumo, pagos, giros y transferencias.

Estructura de los datos, se relacionan las columnas originales con las que se trabajó desde el inicio del proyecto:

Columnas	Descripción	Uso En Modelo?
TIPO_ENTIDAD	Clasificación institucional del intermediario financiero (banco, cooperativa, etc.), que puede influir en el acceso, cobertura y perfil transaccional.	**SI**
CODIGO_ENTIDAD	Identificador único numérico de la entidad. No aporta información para clustering.	**NO**
NOMBRE_ENTIDAD	Nombre completo de la entidad financiera. Información redundante para análisis numérico.	**SI**
FECHA_CORTE	Fecha del dato, útil para series de tiempo o segmentar por año/mes.	**NO**
UNICAP	Código geográfico del lugar donde opera la entidad (puede ser municipio, zona). Permite análisis espacial.	**SI**
DESCRIP_UC	Descripción del código geográfico. Es redundante con UNICAP.	**SI**
RENGLON	Clasificación económica o poblacional del cliente/segmento (puede indicar enfoque del servicio).	**NO**
DESC_RENGLON	Descripción larga de RENGLON. Redundante.	**SI**
TIPO	Puede indicar modalidad de operación o canal (ej. físico, digital). Útil si hay valores diversos.	**SI**
(1) NRO_CORRESPONSALES_FÍSICOS_PROPIOS	Cantidad de corresponsales propios de la entidad que operan físicamente. Indica infraestructura directa.	**SI**
(2) NRO_CORRESPONSALES_FÍSICOS_TERCERIZADOS	Cantidad de corresponsales físicos contratados por terceros. Refleja dependencia operativa externa.	**SI**

PERFILAMIENTO Y SEGMENTACIÓN DE CLIENTES POR CONSUMOS

(3) NRO_CORRESPONSALES_FÍSICOS_ACTIVOS	Total de corresponsales físicos que realizaron operaciones en el período. Mide uso real de la red.	**SI**
(4) NRO_CORRESPONSALES_FÍSICOS	Total general de corresponsales físicos. Redundante si ya se tienen propios y tercerizados.	**NO**
(5) NRO_DEPÓSITOS_CORRESPONSALES_FÍSICOS	Número total de depósitos realizados en corresponsales físicos. Refleja volumen de entrada de dinero por canal.	**SI**
(6) MONTO_DEPÓSITOS_CORRESPONSALES_FÍSICOS	Suma total de los depósitos. Ayuda a calcular ticket promedio y flujo de dinero.	**SI**
(7) NRO_GIROS_ENVIADOS_CORRESPONSALES_FÍSICOS	N° de transferencias enviadas. Ayuda a entender servicios de envío de dinero por canal físico.	**SI**
(8) MONTO_GIROS_ENVIADOS_CORRESPONSALES_FÍSICOS	Monto total girado. Relacionar con número de giros para ver comportamiento.	**SI**
(9) NRO_GIROS_RECIBIDOS_CORRESPONSALES_FÍSICOS	N° de giros recibidos en el canal.	**SI**
(10) MONTO_GIROS_RECIBIDOS_CORRESPONSALES_FÍSICOS	Valor total de giros recibidos. Complementa el campo anterior.	**SI**
(11) NRO_PAGOS_CORRESPONSALES_FÍSICOS	N° de pagos hechos en corresponsales (servicios, productos). Mide uso del canal físico para egresos.	**SI**
(12) MONTO_PAGOS_CORRESPONSALES_FÍSICOS	Monto total pagado. Puede reflejar consumo o gastos.	**SI**
(13) NRO_RETIROS_CORRESPONSALES_FÍSICOS	Número de retiros realizados. Muestra flujo de salida de efectivo.	**SI**
(14) MONTO_RETIROS_CORRESPONSALES_FÍSICOS	Monto total retirado. Importante para saber volumen de liquidez.	**SI**
(15) NRO_TRANSFERENCIAS_CORRESPONSALES_FÍSICOS	N° de transferencias (no giros). Puede indicar uso de cuentas digitales con canal físico.	**SI**
(16) MONTO_TRANSFERENCIAS_CORRESPONSALES_FÍSICOS	Monto total transferido. Complementa el anterior.	**SI**
(17) NRO_TRANSACCIONES_TRÁMITES_CORRESPONSALES_FÍSICOS	Cantidad de trámites distintos a pagos/retiros (actualizaciones, consultas).	**SI**
(18) MONTO_TRANSACCIONES_CORRESPONSALES_FÍSICOS	Suma total de los montos transaccionados en trámites diversos.	**SI**
(19) NRO_CTAS_AHORRO_HASTA_1SMMLV	Número de cuentas de ahorro con saldo hasta 1 salario mínimo mensual legal vigente (SMMLV).	**SI**
(20) SALDO_CTAS_AHORRO_HASTA_1SMMLV	Suma del saldo total en cuentas de ahorro con saldo hasta 1 SMMLV.	**SI**
(21) NRO_CTAS_AHORRO>1SMMLV_HASTA _3SMMLV	Número de cuentas de ahorro con saldo mayor a 1 SMMLV y hasta 3 SMMLV.	**SI**
(22) SALDO_CTAS_AHORRO>1SMMLV_HASTA _3SMMLV	Suma de saldo en cuentas de ahorro con saldo entre 1 y 3 SMMLV.	**SI**
(23) NRO_CTAS_AHORR>3SMMLV_HASTA_ 5SMMLV	Número de cuentas de ahorro con saldo entre 3 y 5 SMMLV.	**SI**
(24) SALDO_CTAS_AHORR>3SMMLV_HASTA_ 5SMMLV	Suma de saldo en cuentas con saldo entre 3 y 5 SMMLV.	**SI**
(25) NRO_CTAS_AHORRO_ACTIVAS	Total número de cuentas de ahorro activas del cliente.	**SI**
(26) SALDO_CTAS_AHORRO_ACTIVAS	Suma total del saldo en cuentas de ahorro activas.	**SI**
(27) NRO_CTAS_AHORRO_MUJERES	Número de cuentas de ahorro registradas a nombre de mujeres.	**SI**
(28) SALDO_CTAS_AHORRO_MUJERES	Suma del saldo en cuentas de ahorro a nombre de mujeres.	**SI**
(29) NRO_CTAS_AHORRO_HOMBRES	Número de cuentas de ahorro a nombre de hombres.	**SI**
(30) SALDO_CTAS_AHORRO_HOMBRES	Suma del saldo en cuentas de ahorro a nombre de hombres.	**SI**
(31) NRO_CTAS_AHORRO	Total número de cuentas de ahorro (hombres + mujeres).	**SI**
(32) SALDO_CTAS_AHORRO	Total saldo en cuentas de ahorro (hombres + mujeres).	**SI**
(41) NRO_CRÉDITO_CONSUMO_MUJERES	Número de créditos de consumo otorgados a mujeres.	**SI**

PERFILAMIENTO Y SEGMENTACIÓN DE CLIENTES POR CONSUMOS

(42) MONTO_CRÉDITO_CONSUMO_MUJERES	Monto total otorgado en créditos de consumo a mujeres.	**SI**
(43) NRO_CRÉDITO_CONSUMO_HOMBRES	Número de créditos de consumo otorgados a hombres.	**SI**
(44) MONTO_CRÉDITO_CONSUMO_HOMBRES	Monto total otorgado en créditos de consumo a hombres.	**SI**
(45) NRO_CRÉDITO_CONSUMO	Número total de créditos de consumo otorgados (hombres + mujeres).	**SI**
(46) MONTO_CRÉDITO_CONSUMO	Monto total otorgado en créditos de consumo.	**SI**
(47) NRO_CRED_CONS_BAJO_MONTO_MUJERES	Número de créditos de consumo de bajo monto otorgados a mujeres.	**SI**
(48) MONTO_CRED_CONS_BAJO_MONTO_MUJERES	Monto total en créditos de bajo monto otorgados a mujeres.	**SI**
(49) NRO_CRED_CONS_BAJO_MONTO_HOMBRES	Número de créditos de consumo de bajo monto otorgados a hombres.	**SI**
(50) MONTO_CRED_CONS_BAJO_MONTO_HOMBRES	Monto total en créditos de bajo monto otorgados a hombres.	**SI**
(51) NRO_CRED_CONS_BAJO_MONTO	Número total de créditos de consumo de bajo monto (hombres + mujeres).	**SI**
(52) MONTO_CRED_CONS_BAJO_MONTO	Monto total en créditos de bajo monto (hombres + mujeres).	**SI**
(53) NRO_CRÉDITO_VIVIENDA_MUJERES	Número de créditos de vivienda otorgados a mujeres.	**NO**
(54) MONTO_CRÉDITO_VIVIENDA_MUJERES	Monto total otorgado en créditos de vivienda a mujeres.	**NO**
(55) NRO_CRÉDITO_VIVIENDA_HOMBRES	Número de créditos de vivienda otorgados a hombres.	**NO**
(56) MONTO_CRÉDITO_VIVIENDA_HOMBRES	Monto total otorgado en créditos de vivienda a hombres.	**NO**
(57) NRO_CRÉDITO_VIVIENDA	Número total de créditos de vivienda otorgados (hombres + mujeres).	**NO**
(58) MONTO_CRÉDITO_VIVIENDA	Monto total otorgado en créditos de vivienda (hombres + mujeres).	**NO**
(59) NRO_MICROCRÉDITO_HASTA_ 1SMMLV	Número de microcréditos otorgados con monto hasta 1 SMMLV.	**NO**
(60) MONTO_MICROCRÉDITO_HASTA_1SMMLV	Monto total otorgado en microcréditos hasta 1 SMMLV.	**NO**
(61) NRO_MICROCRÉDITO_>1SMMLV_HASTA_ 2SMMLV	Número de microcréditos otorgados entre 1 y 2 SMMLV.	**NO**
(62) MONTO_MICROCRÉDITO_>1SMMLV HASTA_ 2SMMLV	Monto total otorgado en microcréditos entre 1 y 2 SMMLV.	**NO**
(63) NRO_MICROCRÉDITO_>2SMMLV_HASTA_3SMMLV	Número de microcréditos otorgados entre 2 y 3 SMMLV.	**NO**
(64) MONTO_MICROCRÉDITO_> 2SMMLV_HASTA_3SMMLV	Monto total otorgado en microcréditos entre 2 y 3 SMMLV.	**NO**
(65) NRO_MICROCRÉDITO_>3SMMLV_HASTA_ 4SMMLV	Número de microcréditos otorgados con monto entre 3 y 4 SMMLV.	**NO**
(66) MONTO_MICROCRÉDITO_>3SMMLV_HASTA_4SMMLV	Monto total otorgado en microcréditos entre 3 y 4 SMMLV.	**NO**
(67) NRO_MICROCRÉDITO_>4SMMLV_HASTA_10SMMLV	Número de microcréditos entre 4 y 10 SMMLV.	**NO**
(68) MONTO_MICROCRÉDITO_> 4SMMLV_HASTA_10SMMLV	Monto total otorgado en microcréditos entre 4 y 10 SMMLV.	**NO**
(69) NRO_MICROCRÉDITO_>10SMMLV_HASTA_25SMMLV	Número de microcréditos entre 10 y 25 SMMLV.	**NO**
(70) MONTO_MICROCRÉDITO_>10SMMLV_HASTA_ 25SMMLV	Monto total otorgado en microcréditos entre 10 y 25 SMMLV.	**NO**
(71) NRO_MICROCRÉDITO_MUJERES	Número total de microcréditos otorgados a mujeres.	**NO**
(72) MONTO_MICROCRÉDITO_MUJERES	Monto total otorgado en microcréditos a mujeres.	**NO**
(73) NRO_MICROCRÉDITO_HOMBRES	Número total de microcréditos otorgados a hombres.	**NO**
(74) MONTO_MICROCRÉDITO_HOMBRES	Monto total otorgado en microcréditos a hombres.	**NO**
(75) NRO_MICROCRÉDITO	Número total de microcréditos otorgados (hombres + mujeres).	**NO**
(76) MONTO_MICROCRÉDITO	Monto total otorgado en microcréditos (hombres + mujeres).	**NO**
(77) NRO_PRODUCTOS_A_NIVEL_NACIONAL	Número total de productos financieros activos a nivel nacional.	**NO**
(78) MONTO_SALDO_PRODUCTOS_A_NIVEL_NACIONAL	Monto total de saldo de productos financieros a nivel nacional.	**NO**

(79) NRO_CORRESPONSALES_FÍSICOS_PROPIOS_ACTIVOS	Número de corresponsales físicos propios activos (puntos de atención propios).	**NO**
(80) NRO_CORRESPONSALES_FÍSICOS_TERCERIZADOS_ACTIVOS	Número de corresponsales físicos tercerizados activos (terceros).	**NO**
(81) NRO_CORRESPONSALES_PROPIOS_MÓVILES	Número de corresponsales propios móviles (vehículos o unidades móviles de atención).	**NO**
(82) NRO_CORRESPONSALES_TERCERIZADOS_MÓVILES	Número de corresponsales móviles tercerizados (operados por terceros).	**NO**
(83) NRO_CORRESPONSALES_MÓVILES_ACTIVOS	Número total de corresponsales móviles activos (propios y tercerizados).	**NO**
(84) NRO_CORRESPONSALES_MÓVILES	Número total de corresponsales móviles (sin diferenciar activos o no).	**NO**
(85) NRO_DEPÓSITOS_CORRESPONSALES_MÓVILES	Número de depósitos realizados en corresponsales móviles.	**SI**
(86) MONTO_DEPÓSITOS_CORRESPONSALES_MÓVILES	Monto total de depósitos realizados en corresponsales móviles.	**SI**
(87) NRO_GIROS_ENVIADOS_CORRESPONSALES_MÓVILES	Número de giros enviados a través de corresponsales móviles.	**SI**
(88) MONTO_GIROS_ENVIADOS_CORRESPONSALES_MÓVILES	Monto total de giros enviados por corresponsales móviles.	**SI**
(89) NRO_GIROS_RECIBIDOS_CORRESPONSALES_MÓVILES	Número de giros recibidos en corresponsales móviles.	**SI**
(90) MONTO_GIROS_RECIBIDOS_CORRESPONSALES_MÓVILES	Monto total de giros recibidos en corresponsales móviles.	**SI**
(91) NRO_PAGOS_CORRESPONSALES_MÓVILES	Número de pagos realizados a través de corresponsales móviles.	**SI**
(92) MONTO_PAGOS_CORRESPONSALES_MÓVILES	Monto total de pagos realizados en corresponsales móviles.	**SI**
(93) NRO_RETIROS_CORRESPONSALES_MÓVILES	Número de retiros efectuados en corresponsales móviles.	**SI**
(94) MONTO_RETIROS_CORRESPONSALES_MÓVILES	Monto total de retiros efectuados en corresponsales móviles.	**SI**
(95) NRO_TRANSFERENCIAS_CORRESPONSALES_MÓVILES	Número de transferencias realizadas por corresponsales móviles.	**SI**
(96) MONTO_TRANSFERENCIAS_CORRESPONSALES_MÓVILES	Monto total de transferencias realizadas por corresponsales móviles.	**SI**
(97) NRO_TRANSACCIONES_TRÁMITES_CORRESPONSALES_MÓVILES	Número de transacciones o trámites diversos realizados en corresponsales móviles.	**SI**
(98) MONTO_TRANSACCIONES_CORRESPONSALES_MÓVILES	Monto total de transacciones o trámites diversos realizados en corresponsales móviles.	**SI**
year	Año en que se registraron los datos o la transacción. Permite análisis temporal y tendencias.	**SI**
month	Mes del año (1-12) del registro o transacción. Ayuda a capturar la estacionalidad.	**SI**
day	Día del mes (1-31) del registro o transacción. Puede reflejar patrones intra-mes.	**SI**

- Problemas detectados (datos faltantes, errores, duplicados)

Dentro de la detección de anomalía de la BBDD, se identifico estructuras de datos con caracteres especiales tales como tildes(´) y símbolos matemáticos tales como

(<,>) , por lo cual la normalización de los registros y encabezados fue relativamente sencilla, adicional no se tenía campos vacíos, ya que principalmente los datos son numéricos y por ende los resultados no concluyentes estaban en 0.

Como depuración de campos no útiles, se excluye campos de que representaban ID's, dando mayor importancia a identificación de la entidad financiera de la transacción, como otros ejemplos de exclusión se tiene campos de créditos de vivienda y microcréditos que no aportan información de valor, dado que el enfoque es consumos.

Preparación de los Datos

- **Limpieza de datos:**

  Esta fue una de las fases más críticas de nuestro proyecto de machine learning, dado que la base de datos contenía una amplia cantidad de información, por ejemplo, nos encontramos con algunos datos sucios o en un formato no adecuado para su análisis.

  Iniciamos con la identificación y corrección de problemas dentro del conjunto de datos, tales como tildes, duplicados, inconsistencias o errores en la entrada de los datos.

Detallamos la secuencia de limpieza utilizada:

- Eliminación de duplicados: Se verificaron las filas duplicadas en el conjunto de datos y se eliminaron aquellas que no aportan nueva información.
- Corrección de errores tipográficos y formatos inconsistentes: Se detectaron errores comunes en los valores de las variables (por ejemplo, caracteres extraños, fechas en diferentes formatos, etc.) y se corrigieron para asegurar la consistencia del conjunto de datos.

Algunas otras correcciones realizadas de este tipo fueron las siguientes:

- Pasar a minúsculas
- Reemplazar espacios por guión bajo

- Eliminar paréntesis y su contenido
- Eliminar caracteres especiales no deseados excepto guión bajo
- Reemplazar múltiples guiones bajos por uno solo
- Transformaciones realizadas (normalización, codificación, etc.)En esta fase se realizaron varias transformaciones a los datos con el fin de mejorar la calidad y la eficiencia del modelo. Las principales transformaciones aplicadas fueron las siguientes:

- **Normalización de datos:**
  - se identifico estructuras de datos con caracteres especiales tales como tildes(´)
  - símbolos matemáticos tales como (<,>)
  - No se tenían campos vacíos, ya que principalmente los datos son numéricos y por ende los resultados no concluyentes estaban en 0.
  - En la tabla anteriormente vista se realizó depuración de campos no útiles, se excluye campos de que representaban ID's, dando mayor importancia a identificación de la entidad financiera de la transacción, como otros ejemplos de exclusión se tiene campos de créditos de vivienda y microcréditos que no aportan información de valor, dado que el enfoque es consumos.

- **Codificación de variables categóricas:**
  - NOMBRE_ENTIDAD: Nombre textual de la entidad → categórica
  - UNICAP: Puede ser un código, si es texto o código discreto → categórica (aunque si fuera estrictamente numérico, sería cuantitativa)
  - DESCRIP_UC: Descripción textual de la unidad de captura → categórica
  - DESC_RENGLON: Descripción textual de renglón → categórica
  - TIPO: Tipo textual o categoría → categórica

- **Selección de variables:** La selección de variables se realizó con el objetivo de reducir la complejidad del modelo y mejorar su desempeño. Se llevaron a cabo las siguientes actividades:

- Análisis de correlación Se aplicó un análisis de correlación entre las variables numéricas para identificar y eliminar aquellas que estaban altamente correlacionadas, ya que esto podría introducir multicolinealidad en el modelo, lo cual afecta negativamente su capacidad de generalización.
- Eliminación de variables irrelevantes: Basado en el conocimiento del dominio y la exploración inicial de los datos, se eliminaron variables que no aportan información relevante al problema. Esto incluye variables con poca variabilidad o aquellas que no eran útiles para el análisis del comportamiento financiero de los clientes.

Se eliminaron las siguientes variables:

nro_credito_vivienda_mujeres, monto_credito_vivienda_mujeres, nro_credito_vivienda_hombres, monto_credito_vivienda_hombres, nro_credito_vivienda, monto_credito_vivienda, nro_microcredito_hasta_1smmlv, monto_microcredito_hasta_1smmlv, nro_microcredito_1smmlv_hasta_2smmlv, monto_microcredito_1smmlv_hasta_2smmlv, nro_microcredito_2smmlv_hasta_3smmlv, monto_microcredito_2smmlv_hasta_3smmlv, nro_microcredito_3smmlv_hasta_4smmlv, monto_microcredito_3smmlv_hasta_4smmlv, nro_microcredito_4smmlv_hasta_10smmlv, monto_microcredito_4smmlv_hasta_10smmlv, nro_microcredito_10smmlv_hasta_25smmlv, monto_microcredito_10smmlv_hasta_25smmlv, nro_microcredito_mujeres, monto_microcredito_mujeres, nro_microcredito_hombres, monto_microcredito_hombres, nro_microcredito, monto_microcredito

Modelado

- Algoritmos seleccionados y justificación

1. K-Means Clustering (por medio de H2O)

- Descripción:

K-Means es un algoritmo de aprendizaje no supervisado utilizado para agrupar datos en k grupos (clusters) basándose en la similitud de sus características.

- Justificación:
  - Es un método eficiente y ampliamente utilizado para segmentación de datos.
  - Permite detectar patrones, grupos homogéneos y posibles anomalías en los datos.
  - En el código, se utiliza K-Means para explorar la estructura interna del dataset y visualizar los clusters tras reducir la dimensionalidad con PCA.

2. PCA (Principal Component Analysis)

- Descripción:

PCA no es un algoritmo de clasificación, sino de reducción de dimensionalidad. Transforma las variables originales en un conjunto menor de variables (componentes principales) que explican la mayor varianza posible.

- Justificación:
  - Permite visualizar datos de alta dimensión en 2D o 3D, facilitando la interpretación visual de los clusters.
  - Ayuda a eliminar colinealidad y ruido, mejorando la eficacia de algoritmos como K-Means.
  - En el código, se usa para proyectar los datos en dos dimensiones antes de graficar y analizar los clusters.
  - Proceso de entrenamiento y validación

1. Selección y Preprocesamiento de Datos

- Carga de datos:Se importa el archivo CSV y se eliminan columnas irrelevantes o redundantes.
- Eliminación de valores nulos:

Se eliminan las filas con datos faltantes para asegurar la calidad del análisis.

- Selección de variables:

Se seleccionan variables numéricas con mayor varianza, ya que estas suelen aportar más información para diferenciar los grupos.

2. Transformación de Datos

- Codificación y escalado:

Se utiliza `HyperTransformer` de la librería `rdt` para transformar las variables categóricas y numéricas, asegurando que sean adecuadas para el modelo de clustering.

- Reducción de dimensionalidad (PCA):

Se aplica PCA para reducir los datos a dos componentes principales, lo que facilita la visualización y puede mejorar la calidad del clustering al eliminar ruido y redundancia.

3. Entrenamiento del Modelo de Clustering

- Definición de rango de k:

Se define un rango de valores para k (número de clusters) a probar, por ejemplo, de 2 a 9.

- Entrenamiento con H2O KMeans:

Para cada valor de k, se entrena un modelo KMeans utilizando H2O, agrupando los datos transformados en k clusters.

4. Validación y Selección del Número de Clusters

- Cálculo del SSE (Suma de Errores Cuadráticos):

Para cada modelo entrenado, se calcula el SSE, que mide la compacidad de los clusters.

- Método del codo:

Se grafica SSE vs. k. El punto donde la reducción de SSE empieza a disminuir ("codo") sugiere el número óptimo de clusters.

- *(Opcional)* Silhouette Score:

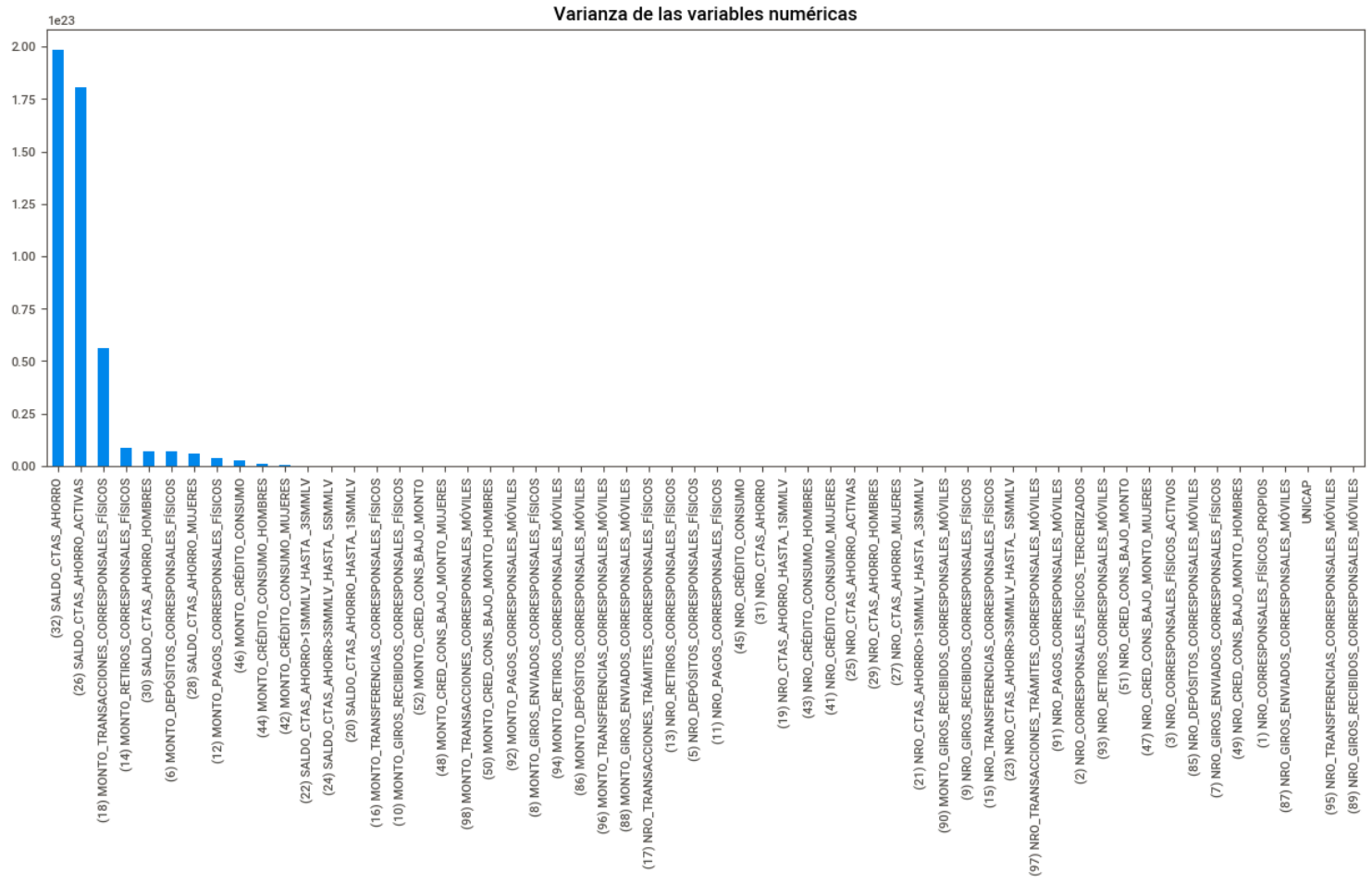
Puede calcularse la puntuación de silueta para evaluar la separación entre los clusters, aunque en tu código está comentado.

5. Visualización de Resultados

- Visualización de clusters:

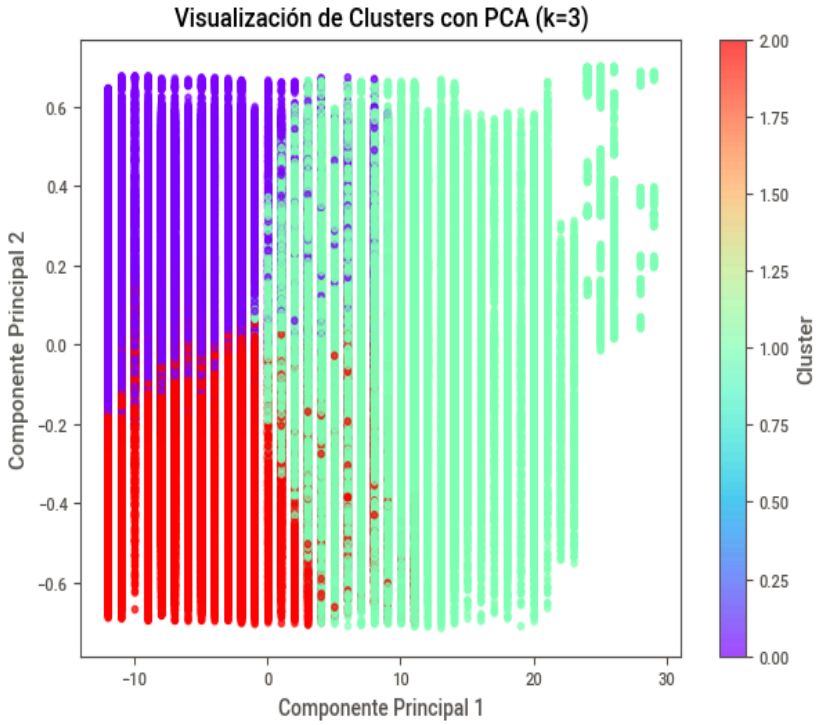
Se proyectan los datos reducidos por PCA en 2D y se colorean según su cluster asignado para verificar visualmente la separación y coherencia de los grupos.

- Métricas de evaluación

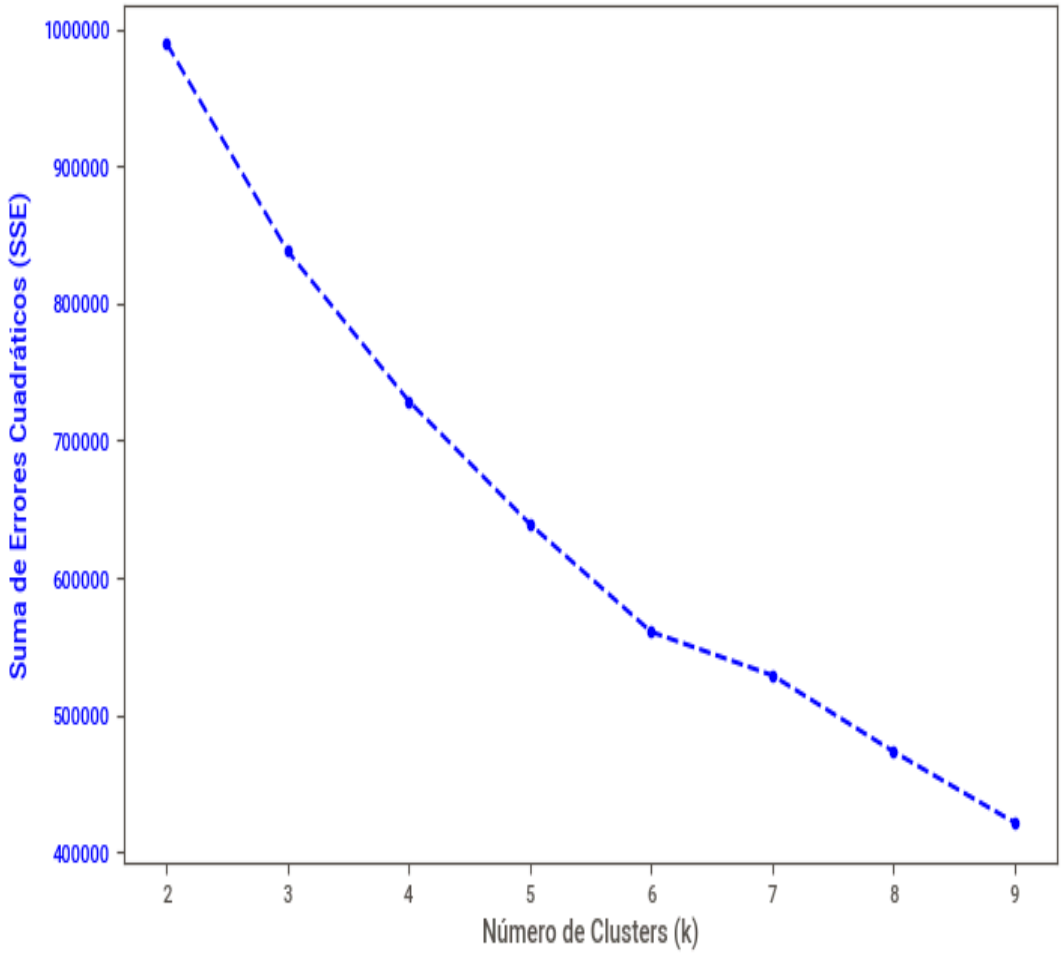


Se seleccionaron únicamente esas tres para evitar ruido y mejorar la eficiencia del modelo, ya que las demás tienen una contribución muy baja.

- Comparación entre modelos



La resultante del algoritmo de clustering con K=3 clusters, proyectados en dos componentes principales, usando análisis de PCA, esta reducción de dimensionalidad nos permite observar como los cluster se agrupan en 3 clusters distintos, representados con los colores.



Aca tenemos el método del codo,utilizado para determinar el número óptimo de cluster, indicando que K=3 es el número adecuado, dado que agregar más cluster a partir de ahí no mejora significativamente el modelo.

Evaluación

- Análisis de resultados

La precisión del proyecto da como resultado un % muy bajo al esperado, y por ende se requiere adicionar más información que pueda alimentar el modelo, con la información y el análisis de los cluster tenemos:

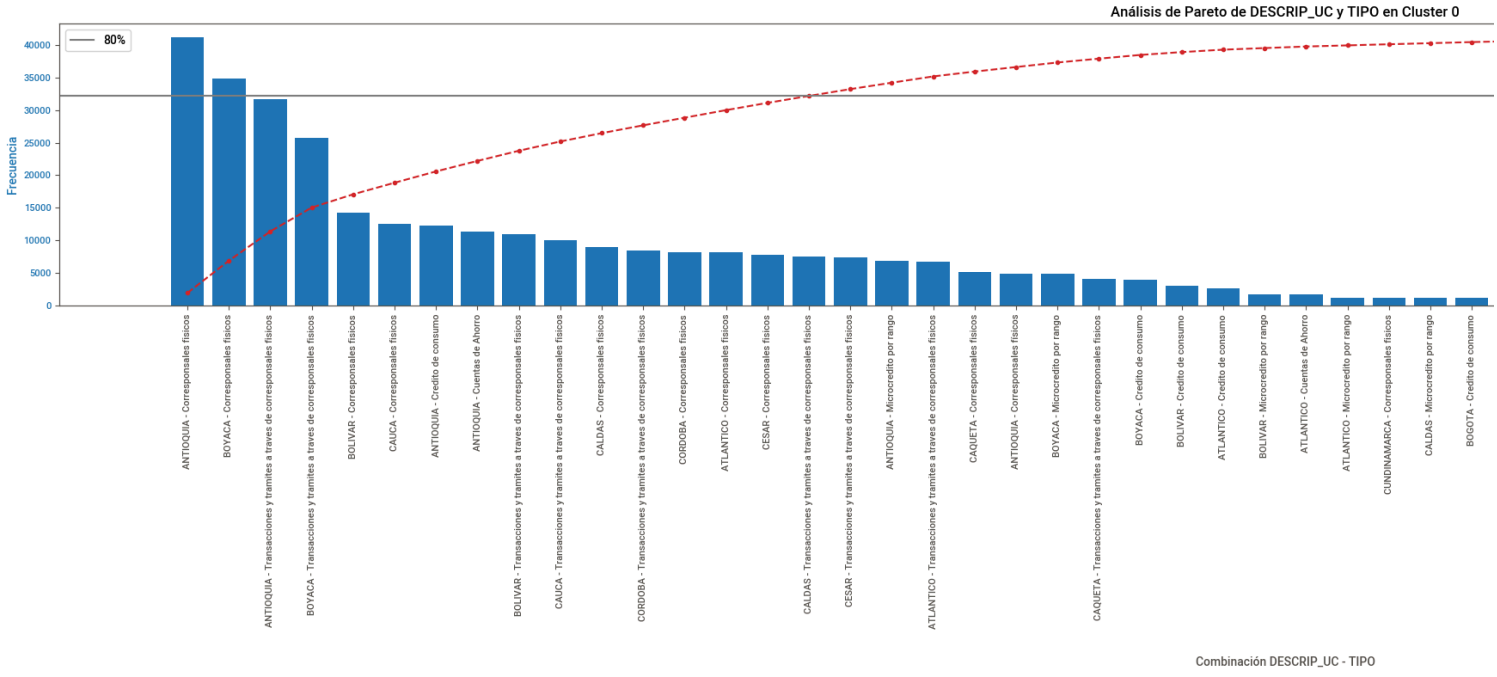
Clúster 0: Alta concentración regional y de tipo

- **Top combinación:** ANTIOQUIA – Compras físicas

- **Distribución:** Altamente concentrada; pocas combinaciones acumulan el 80% de los datos.

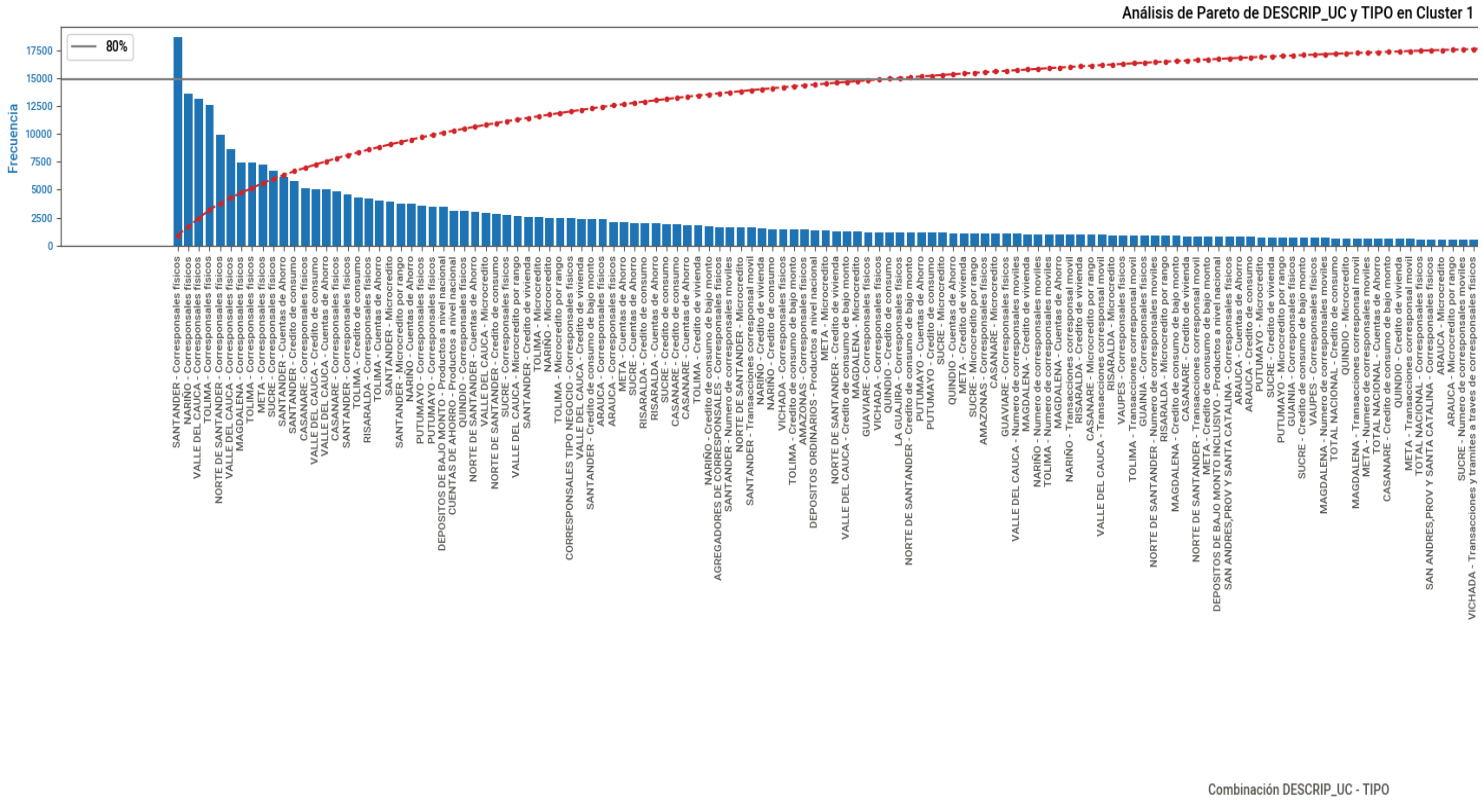
- **Interpretación:** Este grupo representa un segmento homogéneo, probablemente urbano, con comportamiento financiero centrado en compras físicas tradicionales.

- **Uso ideal:** Acciones de marketing masivas, canales físicos, alianzas con retailers.



Clúster 1: Diverso y descentralizado

- **Top combinación:** SANTANDER – Compras físicas (pero con menor dominio)
- **Distribución:** Muy dispersa, larga cola de combinaciones.
- **Interpretación:** Grupo heterogéneo en cuanto a ubicación y tipo de producto; combina regiones rurales, canales digitales y servicios no tradicionales.
- **Uso ideal:** Sub-segmentación adicional (por región o tipo de transacción), iniciativas de inclusión financiera y expansión de cobertura digital.



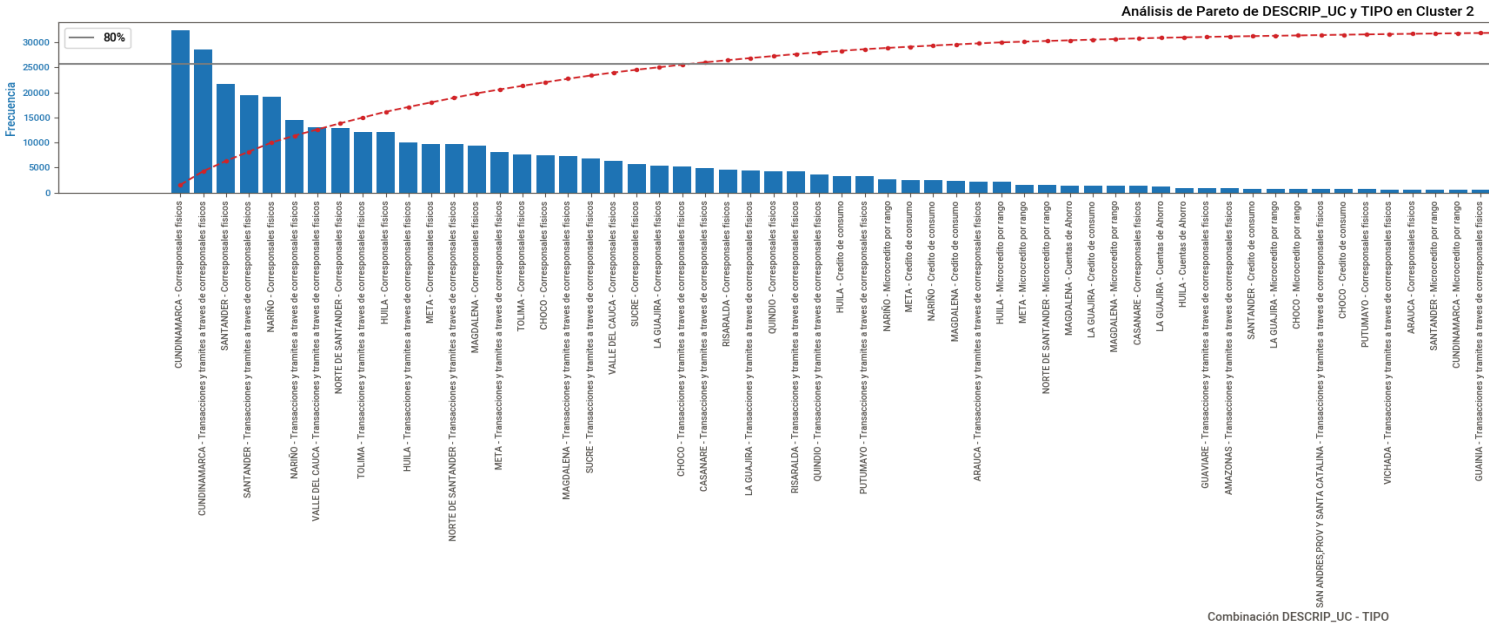
Clúster 2: Mezcla funcional y regional de corresponsales

- **Top combinación:** CUNDINAMARCA – Transacciones/trámites a través de corresponsales físicos

- **Distribución:** Moderadamente dispersa, con foco en servicios por corresponsales y presencia significativa de zonas apartadas.

- **Interpretación:** Usuarios que dependen de **corresponsales físicos** y servicios básicos, posiblemente en zonas semi-urbanas o rurales.

- **Uso ideal:** Fortalecer red de corresponsales, capacitar sobre productos digitales, alianzas con redes de pago locales.



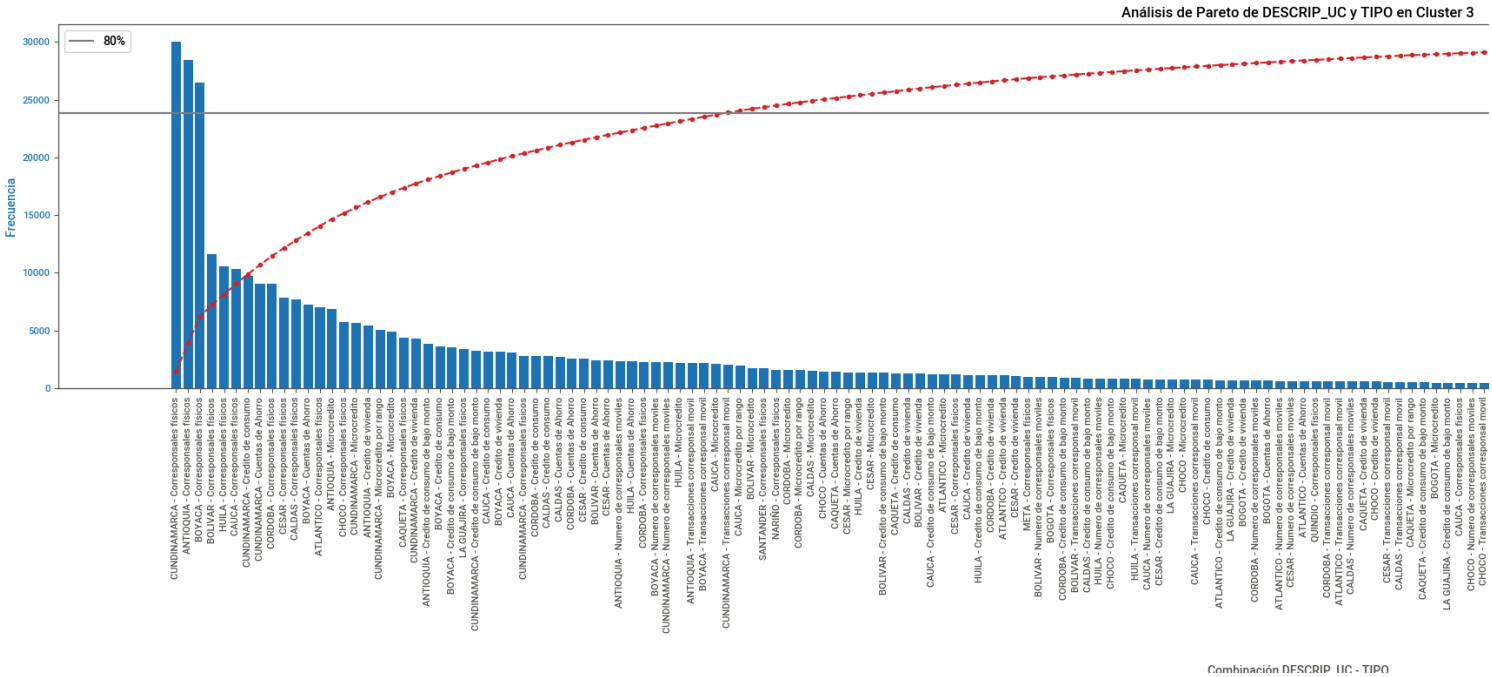
Clúster 3: Consumidores tradicionales y centralizados

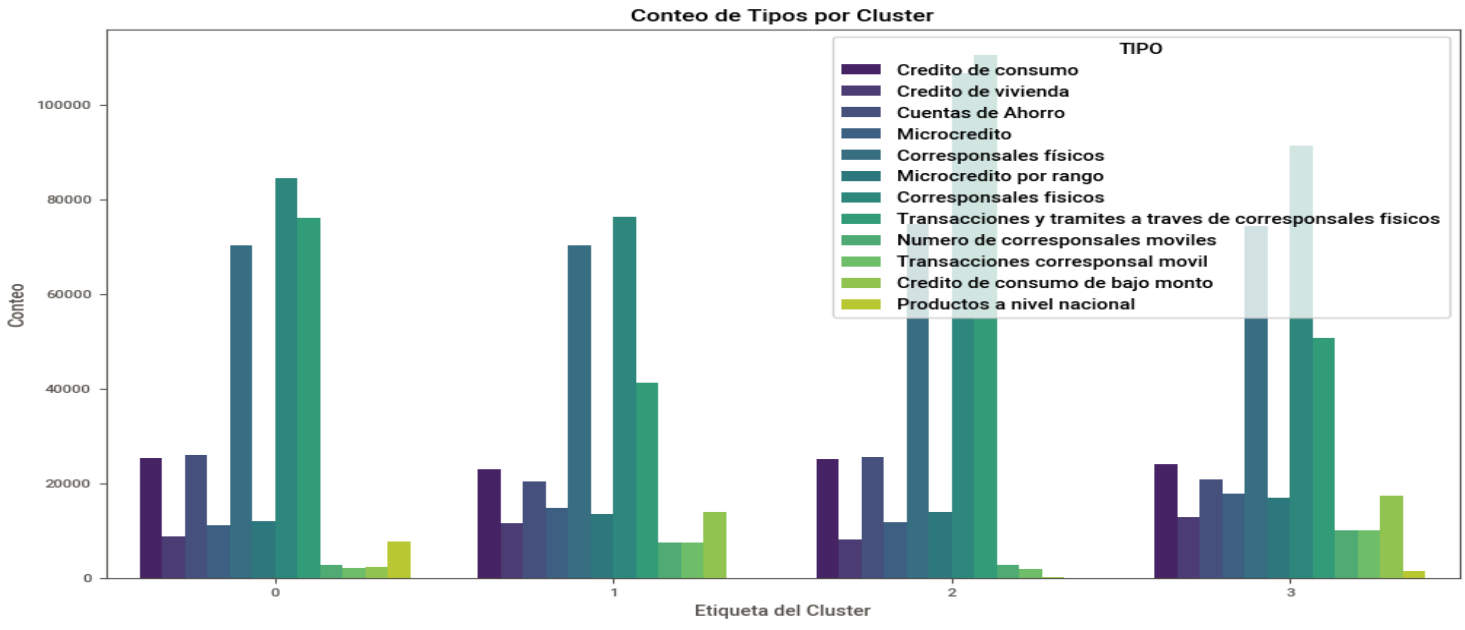
- **Top combinación:** CUNDINAMARCA – Compras físicas

- **Distribución:** Moderadamente concentrada, con más diversidad que el clúster 0.

- **Interpretación:** Usuarios principalmente físicos, pero con algo más de dispersión geográfica y de servicios. Predominan zonas como Cundinamarca y Santander.

- **Uso ideal:** Educación financiera para migración a canales digitales, campañas regionales semi-segmentadas.





- Validación con orientadores del Bootcamp.

Desde el inicio del programa cabe resaltar el buen acompañamiento, instrucciones y contenido otorgado por los orientadores, ahora bien, si nos enfocamos en los seguimientos que generaron al proyecto, encontramos que las sugerencias planteadas al código nos generaron una mayor tasa de éxito y eficacia en el modelo, donde inicialmente nos daba una precisión del modelo de 66.7%, después de aplicar y reestructurar las recomendaciones, aumentó a 99.5% lo que claramente fue una guía acertada.
Los puntos de control entre las clases y avances del proyecto nos dieron la posibilidad de presentar un modelo funcional y correctamente estructurado, donde cada participante del proyecto aportó partes fundamentales y que llevaron, en gran medida, al éxito del proyecto.

- Lecciones aprendidas

- Si bien existe un conocimiento amplio del tema, fue necesario realizar apoyos en sitios externos con respecto a validación de conceptos, investigativas de errores en el código, entre otros; los cuales fueron apuntados a través de herramientas como GPT y COPILOT.
- Tener la fuente de datos y el planteamiento del proyecto fue parte de la lección aprendida dado que fue uno de los principales retos, no solo por contenido útil de la BBDD si no por la estructura necesaria para desarrollar el proyecto, dentro de lo cual para el caso particular, se orientó de manera tal que como resultado, tenemos un modelo no supervisado.
- La importancia en la limpieza de datos, una buena limpieza de datos mejora significativamente los resultados del modelo, dado que este tipo de datas suelen venir con valores faltantes o mal estructurados.
- Selección de variables, si bien no todo los indicadores generan valor, identificar los más relevantes es clave para obtener los cluster o componentes útiles
- En el modelo se descubrió la relación entre bancos y clientes que no

hubiesen sido evidentes con un análisis tradicional.

- segmentación de clientes, logrando agrupar clientes con comportamientos similares, o dinámicas parecidas, volviéndolo muy útil al momento de personalizar estrategias.

- Limitaciones

  - Debido a la disponibilidad de datos y recursos, el alcance del proyecto estará limitado a un conjunto específico de datos financieros, se tenían 3 años inicialmente, con una cantidad de 1.700.000 registros aproximadamente, ende se toma la decisión de procesar la información para el modelo con 1 año, y realizar pruebas con el año continuo.
  - Este proyecto se desarrollará durante el bootcamp talento Tech en las fechas 01 de abril al 30 de mayo, lo que implicó una ejecución acelerada tomando todos los enfoques del proyecto, se debió de realizar una planificación y ejecución de funciones específicas de cada participante para lograr el éxito del mismo.

Implementación

- Plan de despliegue

Preparación del entorno de producción:

- Establecimiento de la base de datos actualizadas que contengan la información de consumo de los clientes en formatos adecuados.

Actualización del flujo de datos:

- Extracción, transformación y carga de datos para alimentar periódicamente el modelo con datos limpios y consistentes.

Implementación del modelo:

- Facilitación del modelo y entrega del mismo para despliegue y mantenimiento.

- Creación de API REST que permita recibir datos de cliente y devolver su segmentación correspondiente.

Monitoreo y mantenimiento:

- Configuración de métricas para evaluar estabilidad del modelo en producción, como distribución de segmentos y alertas ante desviaciones

- Planeación de un plan de retraining o actualización periódica para adaptarse a nuevos patrones de consumo.

Pruebas y validación Final:

- Realización de pruebas piloto con subconjunto de clientes para validar funcionalidad y el impacto en la estrategia comercial.

- Recopilación de feedback y ajustes antes del lanzamiento completo.

- Herramientas utilizadas

La principal herramienta utilizada para modelar, de forma colaborativa, fue google Colab, donde el equipo aportaba al proyecto de manera continua, así mismo se usa herramientas adicionales como google drive, herramientas de office y recursos de red.

- Consideraciones para mantenimiento

- Actualización de datos:

a. frecuencia de actualización
b. Procesamiento de nuevos datos pudiendo ser diaria mensual y semanal
c. Programar el reentrenamiento del modelo para mantener su precisión y relevancia
d. evaluación de desempeño, validar si hay datos nuevos para validar la

precisión de las segmentaciones

e. Documentación de cambios, lógica del modelo y variables usadas

Presupuesto

Los recursos financieros requeridos para el proyecto serían los siguientes:

Presupuesto para Desarrollo de Modelo (COP)	**Cantidad**
Concepto	**Valor (COP)**
Almacenamiento nube (200 GB x 12 meses)	116
Licencias Python	0
PC (Intel Core i7 10ª Gen, 32 GB RAM, 512 GB SSD. Ideal para tareas de programación estándar.	**5.000.000**
Total Aproximado	**5.116.000**

Manejo del Contenido Relacionado (Apéndices / Anexos)

- **Anexo A**: Diccionario de datos

La BBDD contiene información sobre el uso de productos financieros y cobertura, incluyendo la desagregación por género en la tenencia de los diferentes productos financieros, la cual contiene Filas 1,39M y Columnas 99.

Superintendencia Financiera de Colombia (2022-04-21). Inclusión Financiera. actualizacion 7 de marzo de 2025, de https://www.datos.gov.co/Econom-a-y-Finanzas/Inclusi-n-Financiera/kx2f-xjdq/about_data

- **Anexo B**: Código fuente.

Se tiene 3 libros de trabajo dado que en el libro principal se tiene 4 años de información (2021,2022,2023,2024) mientras que en los otros 2 libros, se segmento por año para revisar comportamientos anuales
https://drive.google.com/drive/folders/1aTm_3IKkPBEyh43y7Z3zHxH7L2X2LrDT

- **Anexo C**: Visualizaciones principales

https://drive.google.com/drive/folders/14ou-k3XQtpae85FJA9zDtJPqdnLaaJAQ

- EDA_bbdd_cluster_2_3.html
- EDA_bbdd_cluster_1_3.html
- EDA_bbdd_cluster_1_2.html
- EDA_bbdd_cluster_0_3.html
- EDA_bbdd_cluster_0_2.html
- EDA_bbdd_cluster_0_2.html
- EDA_bbdd_cluster_0_1.html
- EDA_ALL

- **Anexo D**: Detalles técnicos del modelo

Modelo no supervisado.
-RDT (plantilla de transformación numérica)
-k-means (Plantilla H2O)
-PCA
-Análisis exploratorio de datos EDA
-Diagrama del codo

- **Anexo E**: Plan de implementación técnica o de escalabilidad

Como plan de implementación, requerimos información adicional que complemente la BBDD y amplíe y focalice los cluster obtenidos, por ende si proveedor tiene más información el proyecto continúa con su curso de vida.

Conclusiones y Recomendaciones

- Conclusiones claves del análisis
  - Dentro del análisis de los datos, nos encontramos que de la amplia información inicial, pasando todos los pasos de transformación y PCA, la silueta no la obtenemos por la cantidad de datos, y al agrupar tenemos solo 4 clusters útiles (mejor agrupación).
  - Al segmentar y modelar los datos, a través de una discriminacion por año 2023 y 2024, no se presenta varianza o misma tendencia de comportamiento, por el contrario al tomar ambos años en conjuntos nos entrega un panorama distinto en la dimensionalidad de la información.
  - El principal cluster usable es región, lo que nos hace una segmentación posible por clientes_x_region, si bien significamente para la BBDD no se correlaciona más información de valor que la mencionada con su consumo, para entregar valor al modelo.
  - Dentro del cluster principal analizamos que hay un agrupamiento mayor o un top de datos en antioquia, si bien existen datos que se traslapan entre sí, de cierta manera la región(cluster 0) jala la importancia de la información.
  - Dentro de los casos prácticos y el objetivo del proyecto es viable segmentar por región y tipo de producto para analizar el segmento de clientes que requieren tipo de producto específico

- Recomendaciones para el negocio o para futuras investigaciones

- Como principal recomendación para futuras investigaciones,la BBDD debe de tener la estructura y datos suficientes para un buen análisis, si bien ésta debe de ser normalizada, si desde un inicio se establece campos criterios de forma homogénea, no será dispendioso el procesado de la misma.
- Considerando la optimización de tiempos y gestión de trabajo, es importante distribuir tareas dentro del equipo, y posteriormente unificar no solo conceptos sino ideas, de manera tal que siga

- Consideraciones éticas

Las étnica juega un papel fundamental en el desarrollo del proyecto ya que en

cuestión de temas financiero se debe de mantener la mayor transparencia posible, alguno de los aspectos fundamentales son:

- Segmentar al cliente sin entender qué variables conducen a cada grupo, puede generar sesgos existentes.
- El modelo podría agrupar personas o entidades de forma discriminatoria si las variables de entrada están correlacionadas con características sensibles (género, raza, localización geográfica o estrato). esto puede derivar en exclusión financiera o tratamiento desigual.
- los insights generados pueden usarse para prácticas éticamente cuestionables como la discriminacion de precios, perfilamiento financiero agresivo, o la segmentación para venta de productos no apropiados.
- Un enfoque responsable implica no solo competencias técnicas a nivel financiero, sino una visión crítica sobre el impacto social de los modelos.

1. Referencias bibliográficas

Superintendencia Financiera de Colombia (2022-04-21). Inclusión Financiera. actualizacion 7 de marzo de 2025, de https://www.datos.gov.co/Econom-a-y-Finanzas/Inclusi-n-Financiera/kx2f-xjdq/about_data