Universidad Nacional Autónoma de México

Facultad de Estudios Superiores Acatlán

Actuaría

Proyecto Final

Profesor: Juan Carlos Gonzalez Granado

Materia:Analisis Multivariado

Semestre 2024-II

Elaborado por:

Torres Ferrer Luis Noe
Hernadez Estrada Yahir Emiliano
Hernandez Vargas Hector
Sanchez Barajas Ramon
Rojas Duran Claudio

Índice

1 Introducción

1.1 Análisis Multivariado en la Toma de Decisiones

En el ámbito de la investigación y la toma de decisiones basadas en datos, el análisis multivariado juega un papel fundamental al permitir la comprensión de múltiples variables simultáneamente. Este tipo de análisis permite identificar patrones y relaciones, transformando grandes cantidades de información en conocimiento útil para la toma de decisiones informadas; es decir, genera tendencias en grandes conjuntos de datos que serían difíciles de detectar mediante análisis univariados o bivariados tradicionales. Además, el análisis multivariado permite:

- Las herramientas proporcionadas por la estadística permiten a los analistas, empresas o investigadores tomar decisiones fundamentadas en datos. El análisis estadístico permite que las decisiones estén respaldadas por datos, lo que aumenta la probabilidad de obtener resultados confiables.
- A través de técnicas estadísticas, es posible descubrir patrones y relaciones ocultas entre las variables.
- No todos los datos tienen patrones lineales o fáciles de identificar. La estadística es una herramienta útil para medir la variabilidad y comprender las diferencias entre individuos, grupos o circunstancias. Esto es fundamental porque ayuda a anticipar varios escenarios y administrar el riesgo.

En este contexto, en este proyecto, se llevará a cabo un análisis detallado de un conjunto de datos que contiene información financiera y demográfica de clientes de tarjetas de crédito. Los datos incluyen variables como ingresos anuales, límites de crédito, calificaciones crediticias, número de tarjetas de crédito, edad, educación, género, condición de estudiante, estado civil, origen étnico y saldos actuales en tarjetas de crédito. Estas variables ofrecen una rica fuente de información que puede ser utilizada para realizar análisis financieros. Para explorar sus propiedades estadísticas, se visualizará la distribución de las variables a través de histogramas y se analizará la relación entre ellas mediante matrices de correlaciones y varianza.

A lo largo del desarrollo del proyecto, se procederá primero a la obtención de estadísticas descriptivas básicas, tales como la media, mediana, moda, desviación estándar y percentiles, con el fin de comprender la distribución y la dispersión de los datos. A continuación, se emplearán histogramas para visualizar la distribución de cada una de las variables involucradas, lo que permitirá identificar posibles sesgos, tendencias o anormalidades en los datos.

Posteriormente, se llevará a cabo un análisis de correlaciones entre las variables. Este análisis es fundamental para detectar dependencias o relaciones lineales entre las variables, lo que podría tener importantes implicaciones en la toma de decisiones o en la modelización de datos. Para ello, se generará una matriz de correlaciones que permitirá visualizar de forma estructurada la relación entre cada par de variables. Además, se analizará la matriz de varianzas y covarianzas para examinar la magnitud de la variabilidad conjunta entre las variables, lo que es clave en el análisis multivariado para identificar la homogeneidad o heterogeneidad dentro del conjunto de datos.

El análisis de datos multivariado proporcionará una visión integral y detallada del conjunto de datos, facilitando la extracción de conclusiones valiosas y la identificación de patrones o relaciones que de otro modo pasarían desapercibidos. Asimismo, permitirá construir una base sólida para posibles análisis más avanzados o modelos predictivos, fundamentales en la toma de decisiones empresariales, actuariales o de investigación.

1.2 Planteamiento

En un entorno financiero altamente competitivo, las instituciones que ofrecen servicios de crédito enfrentan el reto de gestionar adecuadamente la información de sus clientes para optimizar sus estrategias comerciales, minimizar riesgos y maximizar la rentabilidad. Los datos financieros y demográficos que las empresas recaban de sus clientes, como ingresos anuales, calificaciones crediticias, límites de crédito y datos personales como la edad, educación, género y estado civil, representan una fuente valiosa de información. Sin embargo, el volumen y la complejidad de estos datos dificultan la extracción de información clara y accionable sin un análisis riguroso. El principal problema radica en cómo aprovechar de manera efectiva estos grandes volúmenes de datos para identificar patrones significativos, correlaciones y relaciones que permitan mejorar la toma de decisiones. En concreto, es necesario entender cómo diferentes características financieras y demográficas influyen en el comportamiento crediticio de los clientes, en particular, en la forma en que utilizan sus tarjetas de crédito, y qué riesgos representan para la organización. Sin un análisis exhaustivo y sistemático, las instituciones financieras podrían estar tomando decisiones basadas en información incompleta o errónea, lo que conlleva Además, en la gestión del riesgo crediticio, es fundamental identificar a los clientes que representan un alto riesgo de impago, así como aquellos con potencial de crecimiento en el uso de productos crediticios. Para ello, es necesario implementar un análisis multivariado que permita analizar no solo cada variable individualmente, sino también las interacciones entre ellas. En este sentido, las instituciones deben ser capaces de identificar cuáles son las variables más relevantes que predicen el comportamiento crediticio, tales como los ingresos, el número de tarjetas de crédito, o los límites crediticios, y cómo estos factores varían entre diferentes grupos.

1.3 Objetivo General:

Este conjunto de datos proporciona una visión integral de varios factores que pueden influir en el comportamiento financiero y la solvencia crediticia de un individuo, lo que lo convierte en un recurso valioso para el análisis multivariado y el modelado para un área totalmente financiera.

- **Ingresos y Límite de Crédito:** Examinar la relación entre los niveles de ingreso y los límites de crédito asignados.
- **Calificación Crediticia:** Comprender cómo diversos factores, como el ingreso, el número de tarjetas y la educación, influyen en las calificaciones crediticias.
- **Análisis Demográfico:** Analizar las diferencias en el comportamiento financiero y las características crediticias según el género, el estado civil y el origen étnico.

- **Edad y Comportamiento Financiero:** Investigar cómo el comportamiento financiero, como el saldo y el uso de tarjetas de crédito, cambia con la edad.
- **Modelado de riesgo crediticio:** Desarrollar modelos para predecir el riesgo crediticio en función de datos demográficos y financieros.
- **Segmentación de clientes:** Identificar diferentes segmentos de clientes para adaptar los productos financieros y las estrategias de marketing.
- **Programas de educación financiera:** Diseñar programas de educación financiera específicos para grupos demográficos con el fin de mejorar la alfabetización y el comportamiento financieros.

1.4 Definiciones auxiliares

Como se mencionó al principio del documento, el análisis multivariado nos permite estudiar la relación que existe entre las variables para poder extraer e interpretar la información del conjunto de datos. En éste sentido será bueno definir algunos conceptos un tanto básicos pero de suma importancia para poder entender el desarrollo de dicho proyecto.

Población: Conjunto de personas, eventos u objetos de los cuales nos interesa estudiar alguna característica.

Muestra: Subconjunto de una población de tamaño n

Variable: Característica o atributo de una población. Existen dos tipos de variables: cualitativas y cuantitativas.

Variable cualitativa: Aquella que se usa para representar cualidades o características no contabilizables de la población, a su vez existen dos tipos: ordinales y nominales.

Variable ordinal: Es útil para representar categorizaciones, calificaciones, escalas, etc; las cuales siguen un orden o jerarquía inherente para guardar sentido.

Variable nominal: Es útil para representar múltiples valores, nombres, incluso clasificaciones, en las que no existe un orden y jerarquía para su comprensión.

Variable cuantitativa: Es aquella que se usa para representar cualidades o características contabilizables de la población, a su vez existen dos tipos: discretas y continuas.

Variable discreta: Variable qué, solo toma valores enteros "n.^{en} un conjunto finito de valores separados de R.

Variable continua: Es aquella que toma valores en un subconjunto infinito de R.

Variable aleatoria: Función X : Ω → R que asigna a cada resultado en el espacio muestral Ω un número real. Es decir, X es una función medible que mapea los resultados de un experimento aleatorio a valores num´ericos reales

1.5 Variables redundantes

La detección de variables redundantes que puedan existir en nuestro conjunto de datos, haciendo referencia a una variable que proporciona información redundante que ya está contenida en otras variables del conjunto de datos. En otras palabras, es una variable que no agrega información nueva o útil al análisis y que podría ser eliminada sin afectar significativamente los resultados.

Algunos de los métodos para poder detectar dichas variables redundantes son:

- Matriz de correlaciones.
- Análisis de componentes principales.
- Grupos (Clustering)
- Análisis de Varianza Inflada por Tolerancia.

En el caso del método de la matriz de correlaciones, discriminaremos como variables redundantes a aquellas que estén indicadas con alta correlación.

1.6 Valores atípicos

Los valores atípicos son aquellas observaciones cuyos valores son muy diferentes a otras observaciones del mismo grupo de datos. Estos pueden ser ocasionados por errores de procedimiento, acontecimientos extraordinarios, valores extremos o por causas que desconocemos.

Estos datos atípicos son un problema, ya que tienen un efecto de distorsión en el resultado del análisis de los datos. Por esta razón, es importante identificarlos para tratarlos de manera adecuada dichos metodos son los siguientes:

1.6 Método del rango intercuartílico (RIC)

El rango intercuartílico es una medida de variabilidad basada en la división de un conjunto de datos en cuartiles. Comprende el rango entre el primer y el tercer cuartil y se denota por:

$$RIC = (Q_3 - Q_1)$$

Bajo este método, se considerarán como valores atípicos aquellos datos que queden por debajo de:

$$Q_1 - 1{,}5 \times RIC$$

o por encima de:

$$Q_3 + 1{,}5 \times RIC$$

1.6 Método del z-score

Los puntajes z o puntajes estándar miden la distancia entre un punto de datos y la media del conjunto en términos de desviación estándar.

Para usar este puntaje en la detección de valores atípicos, establecemos un umbral en función del nivel de importancia o de los requisitos específicos del conjunto de datos. En general, una puntuación z mayor a 3 se considera un caso atípico. Es decir, si cumplen la condición:

$$\frac{X - \mu_x}{\sigma_x} > 3$$

se consideran valores atípicos.

1.6 Distancia de Mahalanobis

La distancia de Mahalanobis es un criterio que depende de los parámetros estimados de la distribución multivariada. Describe la distancia entre cada punto de datos y el centro de masa. La Distancia de Mahalanobis se define como:

$$MSD_i = \sqrt{(x_i - \bar{x})^T S_n^{-1} (x_i - \bar{x})}$$

Donde T denota la matriz transpuesta, $\bar{x}$ expresa la media del vector muestral y S_n la matriz de covarianza muestral. Se lleva a cabo también un proceso de normalización, ya que para datos multivariantes normales los valores de la distancia de Mahalanobis tienen aproximadamente una distribución chi-cuadrado con p grados de libertad. Bajo este criterio, las observaciones que se encuentren lejos del centro de masa, es decir, con distancia de Mahalanobis grande se señalan como valores atípicos. Recordemos que estos metodos son para datos multivariados, para los datos univariados se usan otros metodos.

1.7 Componentes principales

El análisis de componentes principales busca algunas combinaciones lineales que puedan ser utilizadas para resumir los datos, perdiendo en el proceso la menor cantidad de información posible. Este intento de reducir la dimensión puede ser descrito como un "Resumen eficiente" de los datos.

Definición: Si x es un vector aleatorio con media μ y covarianza σ, entonces la transformación de componentes principales es:

$$x \to y = \Gamma'(x - \mu)$$

Donde Γ es ortogonal, $\Gamma'\Sigma\Gamma = A$ es diagonal y $\lambda_1 > \lambda_2 > \ldots > \lambda_p > 0$. La estricta positividad de los eigenvalores λ_i está garantizada si Σ es definida positiva. Esta representación de Σ sigue del teorema de descomposición espectral. El i-ésimo componente principal de x puede definirse como el i-ésimo elemento del vector y, es decir:

$$y_i = \gamma_i'(x - \mu)$$

La combinación lineal con la varianza más grande será la primera componente principal, definida por analogía sobre la expresión anterior como:

$$y^{(1)} = (X - 1\bar{x}')g^{(1)}$$

En este caso, $g^{(1)}$ es el eigenvector estandarizado correspondiente al eigenvalor más grande de S (es decir, $S = GLG'$).

De forma similar, el i-ésimo componente principal se define como:

$$y^{(i)} = (X - 1\bar{x}')g^{(i)}$$

La correlación entre x y el vector de componentes principales y se calcula. Por simplicidad, se asume que la media de x y por lo tanto de y es 0. La covarianza de x y y es:

$$E(xy) = E(xx'\Gamma) = \Sigma\Gamma = \Gamma\Lambda$$

Por lo tanto, la covarianza entre x_i y y_i es $\gamma_{ij}\lambda_j$. Ahora x_i y y_i tienen varianzas σ_{ii} y λ_i respectivamente, así su correlación ρ_{ij} es entonces:

$$\rho_{ij} = \frac{\gamma_{ij}\lambda_j}{(\sigma_{ii}\lambda_j)^{1/2}} = \gamma_{ij}\left(\frac{\lambda_j}{\sigma_{ii}}\right)^{1/2}$$

Cuando Σ es una matriz de correlación, $\sigma_{ii} = 1$, entonces $\rho_{ij} = \gamma_{ij}(\lambda_j)^{1/2}$. Podemos decir que la proporción de variación de x_i "explicada" por y_i es ρ_{ij}^2. Entonces, dado que los elementos de y no están correlacionados, cualquier conjunto G de componentes explica una proporción:

$$\rho_{iG}^2 = \sum_{j\in G}\rho_{ij}^2 = \frac{1}{\sigma_{ii}}\sum_{j\in G}\lambda_j\gamma_{ij}^2$$

de la variación en x_i. El denominador de esta expresión representa la variación en x_i que se explicará, y el numerador da la cantidad explicada por el conjunto G. Cuando G incluye todos los componentes, el numerador es el elemento (i,i)-ésimo de $\Gamma\Lambda\Gamma'$, que por supuesto es solo σ_{ii}, de modo que la relación es uno.

Notemos que la parte de la variación total contabilizada por los componentes en G se puede expresar como la suma de todas las variables p de la proporción de variación en cada variable explicada por los componentes en G, donde cada variable se pondera por su varianza; es decir:

$$\sum_{j\in G}\lambda_j = \sum_{i=1}^{p}\sigma_{ii}\rho_{iG}^2$$

1.8 Escalamiento multidimensional

1.9 Análisis de factores

1.10 Análisis de regresión

El objetivo principal de la regresión es predecir el valor de la variable dependiente basado en los valores de las variables independientes y entender la naturaleza de sus relaciones.

Regresión lineal

La regresión lineal es el modelo de regresión más simple y común. Se asume que la relación entre la variable dependiente y y las variables independientes X1,X2,...,Xp es lineal. El modelo se expresa como:

$$\text{y} = \beta_0 X + \varepsilon$$

donde:

- β_0 es la intersección o término constante.
- β_1 es el vector de coeficientes de regresiónn que representa la magnitud y la dirección de la relación entre cada variable independiente y la variable dependiente.
- ε es el término de error que captura la variabilidad en y no explicada por las variables independientes.

Regresión logística

La **regresión logística** se utiliza cuando la variable dependiente es categórica, especialmente binaria (con dos categorías). El modelo estima la probabilidad de que la variable dependiente tome un valor específico (e.g., éxito o fracaso). La relación se modela utilizando la función logística:

$$logit(p) = log(\tfrac{p}{1-p}) = \beta_0 + \beta_1 X$$

Donde p es la probabilidad de que la variable dependiente sea 1.

Criterios de evaluación

Los modelos de regresión se evalúan utilizando diversos criterios y estadísticas:

- Coeficiente de Determinación (R^2): Indica la proporción de la varianza en la variable dependiente explicada por las variables independientes.
- Error cuadrático medio (MSE): Mide la magnitud promedio de los errores de predicción.
- Estad´ıstico F: Prueba la significancia global del modelo.
- Pruebas t: Evalúan la significancia individual de cada coeficiente de regresión.

1.11 Árboles de decisión

Los árboles de decisión son modelos predictivos utilizados tanto para clasifica ción como para regresión. Estos modelos utilizan un conjunto de reglas de decisión derivadas de los datos para predecir el valor de una variable objetivo.

Los árboles de decisión son priorizados debido a su capacidad para manejar tanto variables categóricas como continuas.

La construcción de un árbol de decisión implica los siguientes pasos:

1. Selección de la mejor división: En cada nodo, seleccionar la variable y el punto de división que mejor separen los datos según una métrica específica.
2. División del nodo: Crear ramas basadas en la mejor división.
3. Repetición: Repetir los pasos anteriores para cada nodo resultante hasta que se cumpla un criterio de detención (profundidad máxima del árbol, número mínimo de muestras en un nodo, etc.).
4. Asignación de valores: Asignar valores a los nodos de hoja basándose en la mayoría de clases (para clasificación) o en la media de los valores (para regresión).

Para seleccionar la mejor división en cada nodo se mide la impureza de un nodo, la incertidumbre o la reducción en la variabilidad de los valores de la variable objetivo.

Para evitar el sobreajuste, se puede aplicar una técnica llamada **poda**. Puede ser:

1. Previa: Donde se detiene la construcción del árbol en base a la profundidad máxima o mínimo de muestras en un nodo
2. Posterior: Donde luego de construir el árbol completo se eliminan nodos que no proporcionan poder predictivo significativo.

1.12 Análisis de discriminante lineal

El Análisis de Discriminante Lineal **(LDA)** es una técnica estadística utilizada para encontrar una combinación lineal de características que separe o caracterice dos o más clases de objetos o eventos.

El objetivo principal del **LDA** es proyectar los datos en un espacio de menor dimensión con la máxima separabilidad de clases.

El **LDA** se basa en varias suposiciones clave:

- Normalidad: Las características siguen una distribución normal multivariante.
- Covarianza igual: Las clases tienen la misma matriz de covarianza.
- Linealidad: Las relaciones entre las características son lineales.

El proceso de **LDA** implica los aiguientes pasos:

1. **Calcular las medias de las clases y la media general**

$$\mu_i = \frac{1}{N_i} \sum_{x \in C_i} X$$

$$\mu = \frac{1}{N} \sum_{i=1}^{k} N_i \mu_i$$

Donde:

- N_i es el número de muestras en la clase i.
- μ_i es la media de la clase i.
- N es el número total de muestras.
- μ es la media general.

2. **Calcular la matriz de varianza dentro de la clase** S_W

$$S_w = \sum_{i=1}^{k} \sum_{x \in C_i} (X - \mu_i)(X - \mu_i)$$

3. **Calcular la matriz de varianza entre clases** S_B

$$S_B = \sum_{i=1}^{k} N_i (\mu_i - \mu)(\mu_i - \mu)^T$$

4. **Resolver la ecuación generalizada de valores propios**

$$W = S_W^{-1} SB$$

Donde W es la matriz de transformación de datos. Los k vectores propios correspondientes a los mayores valores propios forman la base del espacio de características reducido.

5. **Proyección de las muestras**

Proyectar las muestras originales (X de dimensión M), en el nuevo espacio ($k < M$) de características utilizando:

$$Y = XV_k$$

Donde V_k es la matriz de vectores propios de W.

1.13 Distancia de Mahalanobis

La distancia de Mahalanobis es una medida de distancia que considera la correlación y dispersión de los datos en un espacio multidimensional. A diferencia de la distancia euclidiana, que solo mide la distancia en línea recta, la distancia de Mahalanobis ajusta la distancia en función de la distribución de los datos, permitiendo identificar valores atípicos y realizar el análisis discriminante.

Además, se lleva a cabo un proceso de normalización, ya que para datos multivariantes normales los valores de la distancia de Mahalanobis tienen una distribución chi-cuadrada con p grados de libertad.

Bajo el criterio anterior, aquellas observaciones (datos) que se encuentren lejos del centro de masa, es decir, con una distancia de Mahalanobis grande se considerarán como datos atípicos (outliers).

1.13 Cálculo de la Distancia de Mahalanobis

La distancia de Mahalanobis $D_M(X_i)$ entre un vector de observaciones $\mathbf{X}$ y el vector de medias $\bar{\mathbf{X}}$ se define como:

$$D_M(X) = \sqrt{(X_i - \overline{X})^T \ S^{-1}(X_i - \overline{X})}$$

Donde:

- $\overline{X}$ es la media muestral de la respectiva variable
- S^{-1} es la matriz de varianzas y covarianza de los datos

Este cálculo permite considerar tanto la dispersión de los datos como las correlaciones entre variables.

2 Desarrollo

2.1 Nuestros datos

Obtuvimos nuestros datos de el sitio web https://www.kaggle.com/datasets/rassiem/credit-data/data, el cual es *un conjunto de datos completo sobre los atributos financieros de los clientes de tarjetas de crédito*

2.2 Analisis descriptivo y exploratorio de los datos

Este conjunto de datos contiene información relacionada con las características financieras y demográficas de las personas, que puede ser útil para diversos análisis, como la calificación crediticia, la segmentación de clientes o los estudios de comportamiento financiero. El conjunto de datos incluye las siguientes columnas:

2.2 Naturaleza de nuestras Variables

Para esta parte procederemos a clasificar nuestras variables de acuerdo a nuestro set de datos.
Variables Numéricas:

- *Limit:* El límite de crédito asignado a la cuenta de tarjeta de crédito del individuo.
- *Rating:* Una puntuación de calificación crediticia para el individuo.
- *Cards:* El número de tarjetas de crédito que posee el individuo.
- *Age:* La edad del individuo.
- *Education:* El número de años de educación completados por el individuo.
- *Balance:* El saldo actual en la cuenta de tarjeta de crédito del individuo.
- *Income:* El ingreso anual del individuo (en miles de dólares).

Y **variables categóricas**

- *Gender:* El género del individuo (masculino/femenino).
- *Student:* Si el individuo es estudiante (Sí/No).
- *Married:* El estado civil del individuo (si esta casado Sí/No).
- *Ethnicity:* El origen étnico del individuo.

Estas ultimas variables son perceptibles a que para una mejor lectura podamos reorganizarlas con valores 0 y 1 como las variables Dummy.

2.2 Resumen estadístico

Con la librería de pandas podemos generar un resumen estadístico de las columnas numéricas o categóricas de nuestro set de datos. Es muy útil para obtener una visión general de los datos y verificar su distribución, dispersión y posibles anomalías.

Por defecto, este método proporciona estadísticas descriptivas para columnas numéricas (como las de tipo float64 o int64). Sin embargo, también se puede aplicar a columnas categóricas. A partir de los resultados de `df.describe()`, se pueden hacer varias observaciones

	dtype	#missing	#duplicates	#unique	min	max	avg	std dev
Income	float64	0	0	399	10.354000	186.634000	45.218885	35.244273
Limit	int64	0	0	387	855.000000	13913.000000	4735.600000	2308.198848
Rating	int64	0	0	283	93.000000	982.000000	354.940000	154.724143
Cards	int64	0	0	9	1.000000	9.000000	2.957500	1.371275
Age	int64	0	0	68	23.000000	98.000000	55.667500	17.249807
Education	int64	0	0	16	5.000000	20.000000	13.450000	3.125207
Gender	object	0	0	2	nan	nan	nan	nan
Student	object	0	0	2	nan	nan	nan	nan
Married	object	0	0	2	nan	nan	nan	nan
Ethnicity	object	0	0	3	nan	nan	nan	nan
Balance	int64	0	0	284	0.000000	1999.000000	520.015000	459.758877

Figura 1: Estadísticas Básicas

```
1 df.describe()
```

	Income	Limit	Rating	Cards	Age	Education	Balance
count	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000
mean	45.218885	4735.600000	354.940000	2.957500	55.667500	13.450000	520.015000
std	35.244273	2308.198848	154.724143	1.371275	17.249807	3.125207	459.758877
min	10.354000	855.000000	93.000000	1.000000	23.000000	5.000000	0.000000
25%	21.007250	3088.000000	247.250000	2.000000	41.750000	11.000000	68.750000
50%	33.115500	4622.500000	344.000000	3.000000	56.000000	14.000000	459.500000
75%	57.470750	5872.750000	437.250000	4.000000	70.000000	16.000000	863.000000
max	186.634000	13913.000000	982.000000	9.000000	98.000000	20.000000	1999.000000

Figura 2: Estadísticas Básicas

sobre las características estadísticas de los datos:

- **Ingresos (Income):**
  - Media: 45.22, con una desviación estándar de 35.24, lo cual sugiere una alta variabilidad en los ingresos de los individuos.
  - Rango: Desde un mínimo de 10.35 hasta un máximo de 186.63, lo que indica una amplia dispersión en los ingresos.
- **Límite de Crédito (Limit):**
  - Media: 4735.6, con una desviación estándar de 2308.2.
  - Rango amplio, de 855 a 13913, lo que podría significar una gran variación en los límites de crédito asignados.
- **Calificación Crediticia (Rating):**
  - Media: 354.94, con desviación estándar de 154.72.
  - Rango: Desde 93 a 982, indicando grandes diferencias en las calificaciones crediticias.
- **Número de Tarjetas (Cards):**
  - Media: 2.96 tarjetas, con una desviación estándar baja de 1.37.
  - Rango limitado: de 1 a 9 tarjetas. La mayoría de los individuos parecen tener entre 2 y 4 tarjetas (percentiles del 25 % y 75 %).
- **Edad (Age):**

  - Media: 55.67 años, con desviación estándar de 17.25.
  - Rango: Desde 23 a 98 años, lo que muestra una variabilidad moderada.

- **Educación (Education):**
  - Media: 13.45 años de educación (aproximadamente secundaria o preparatoria), con una desviación estándar de 3.12.
  - Rango: Desde 5 a 20 años, lo que indica diferencias en los niveles educativos.

- **Saldo (Balance):**
  - Media: 520.02, con una desviación estándar de 459.76, lo que indica una dispersión significativa en los saldos.
  - Rango: Desde 0 hasta 1999, sugiriendo que algunos individuos no tienen saldo de deuda, mientras que otros tienen saldos elevados.

- **Resumen:** Los datos muestran variabilidad considerable en variables financieras (ingresos, límite de crédito, saldo) y factores demográficos (edad, educación), lo cual sugiere posibles diferencias en comportamientos financieros entre subgrupos. Estos datos pueden servir para analizar relaciones, como cómo los ingresos afectan el límite de crédito o cómo la educación puede relacionarse con el saldo y la calificación crediticia.

Por la naturaleza de nuestros datos sospechamos que estos no han sido limpiados, es decir, no están estandarizados. Lo cual haremos mas adelante para un mejor manejo de esta información.

2.2 Apartado Gráfico

Procedemos a visualizar nuestras variables mediante histogramas.

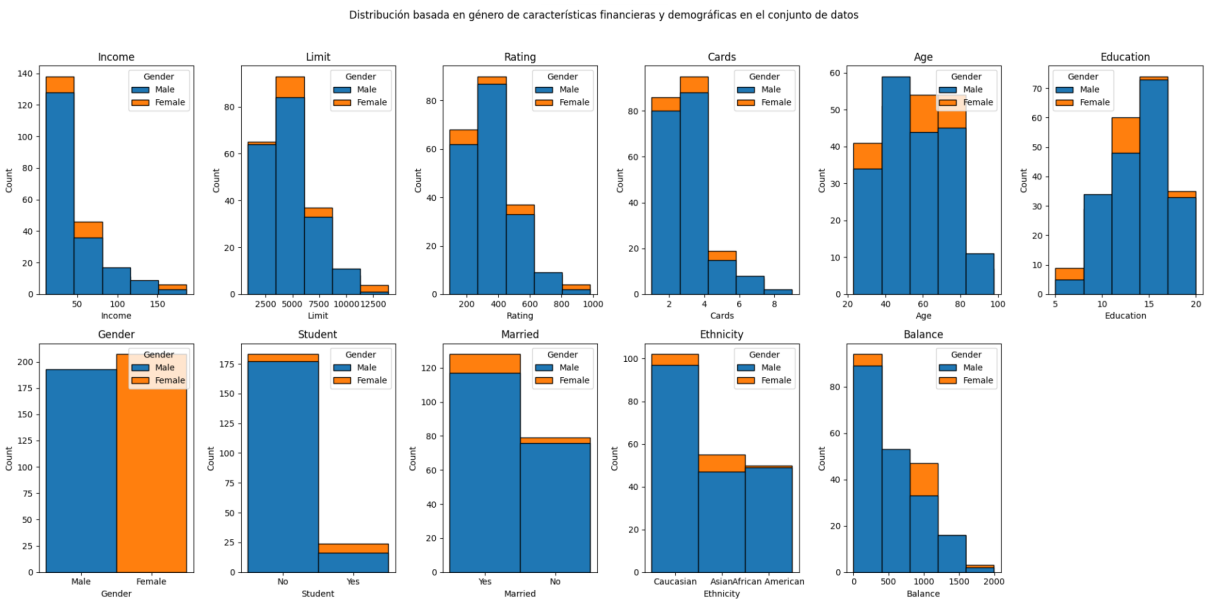


Figura 3: Histogramas

Con base en los histogramas, podemos deducir varias características y patrones en los datos financieros y demográficos, diferenciados por género:

- **Ingresos (Income)**: La mayoría de los individuos tienen ingresos bajos, concentrados en menos de 50 unidades. Los hombres parecen tener una representación ligeramente mayor en los ingresos altos, aunque tanto hombres como mujeres tienen ingresos similares en los niveles más bajos.
- **Límite de Crédito (Limit)**: La mayoría de los límites de crédito están por debajo de 7500 unidades. Los hombres tienen límites más altos en comparación con las mujeres, especialmente en los niveles superiores.
- **Calificación Crediticia (Rating)**: La distribución de la calificación crediticia es amplia, pero la mayoría se concentra alrededor de los valores bajos y medios (menos de 500). Las diferencias de género no son tan pronunciadas en este caso.
- **Número de Tarjetas (Cards)**: La mayoría de las personas tienen entre 2 y 4 tarjetas. No se observan diferencias significativas por género en la distribución de la cantidad de tarjetas.
- **Edad (Age)**: La mayor parte de la población en el conjunto de datos se concentra entre los 40 y 70 años. Hay una distribución similar entre géneros en esta variable.
- **Educación (Education)**: La mayoría de los individuos tienen entre 10 y 15 años de educación. Aquí también se observa una distribución relativamente pareja entre hombres y mujeres.

- **Género (Gender)**: La población está dividida casi en partes iguales entre hombres y mujeres.
- **Estudiante (Student)**: La gran mayoría no son estudiantes, y no se observan grandes diferencias por género en esta categoría.
- **Estado Civil (Married)**: Hay más personas casadas que no casadas en el conjunto de datos, y esta tendencia es similar para ambos géneros, aunque los hombres parecen estar ligeramente más representados en la categoría de çasado".
- **Etnicidad (Ethnicity)**: La mayoría de las personas se identifican como Çaucasian", con representación menor de .Asianz .African American". Hay una ligera diferencia en la proporción de géneros dentro de estas etnias, especialmente en la categoría Çaucasian".
- **Saldo (Balance)**: La mayoría de los saldos están por debajo de las 1000 unidades, y los hombres parecen tener saldos ligeramente más altos que las mujeres en las categorías superiores.
- **Conclusiones Generales**:
  - **Distribución Similar en Muchas Variables**: En la mayoría de las variables (edad, educación, número de tarjetas), no hay grandes diferencias en la distribución entre hombres y mujeres.
  - **Diferencias en Variables Financieras**: En variables como límite de crédito e ingresos, los hombres parecen estar sobrerrepresentados en los niveles altos, lo cual podría indicar ciertas tendencias de asignación de crédito y capacidad económica.
  - **Segmentación Demográfica y Etnográfica**: La variable étnica muestra una representación mayoritaria de personas caucásicas, lo cual podría tener implicaciones en el análisis de comportamiento financiero y acceso a crédito en función de la etnicidad.

Notamos que la variable student esta muy desiquilibrada asi como Balance contiene bastantes ceros.

2.2 Identificación de valores atípicos

Por otro lado observamos los boxplots.

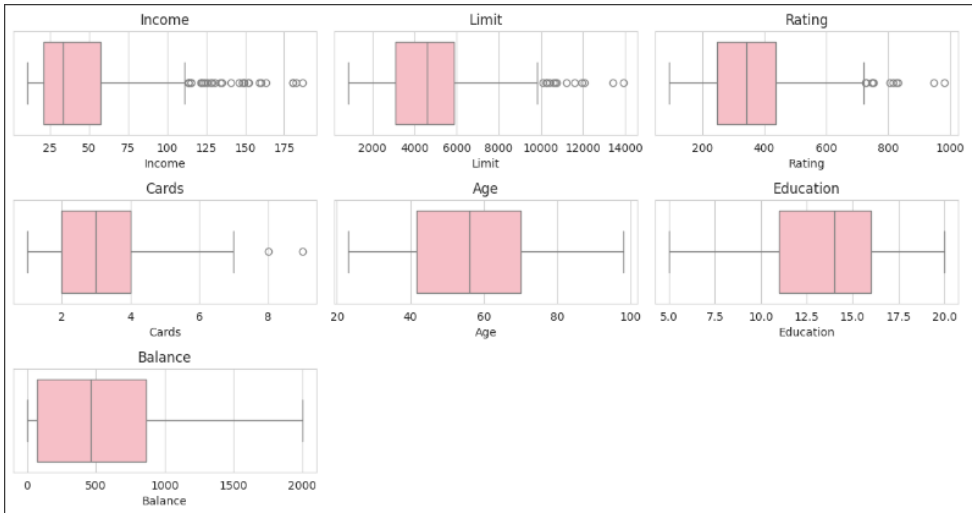


Figura 4: Histogramas

Como los diagramas de caja ayudan a identificar la mediana, el rango intercuartílico y los valores atípicos en cada variable, lo que permite una rápida evaluación de la dispersión y centralización de los datos. Podemos observar con respecto a cada columna un mejor detalle.

- **Income:**
  - La mediana del ingreso está cerca de los 60.
  - Existen varios valores atípicos hacia el extremo superior, que representan ingresos significativamente más altos que el rango intercuartílico.
- **Limit:**
  - La mediana del límite de crédito está alrededor de 5,000.
  - Hay valores atípicos a partir de los 10,000, indicando que algunos individuos tienen límites de crédito mucho mayores que el promedio.
- **Rating:**
  - La mediana está cerca de los 400, lo cual podría representar una puntuación de crédito promedio.
  - Existen varios valores atípicos altos, lo que indica que algunos individuos tienen una calificación de crédito excepcionalmente alta.
- **Cards:**
  - La mayoría de las personas tiene entre 2 y 4 tarjetas de crédito.
  - Hay unos pocos valores atípicos donde las personas tienen más de 6 tarjetas de crédito.

- **Age:**
  - La edad está centrada alrededor de los 50-60 años.
  - No hay valores atípicos visibles en este gráfico, lo que sugiere que las edades están relativamente distribuidas de manera uniforme.
- **Education:**
  - La mediana de la educación está alrededor de 12-13 años (aproximadamente la educación secundaria completa).
  - No hay valores atípicos evidentes en este gráfico, lo que indica que la educación está distribuida de manera bastante homogénea.
- **Balance:**
  - La mediana del balance (posiblemente saldo de deuda) está alrededor de 500.
  - No se observan valores atípicos en este gráfico, y los valores están bastante distribuidos sin excesos significativos.

En general, los diagramas de caja ayudan a identificar la mediana, el rango intercuartílico y los valores atípicos en cada variable, lo que permite una rápida evaluación de la dispersión y centralización de los datos.

2.2 Matriz de varianzas

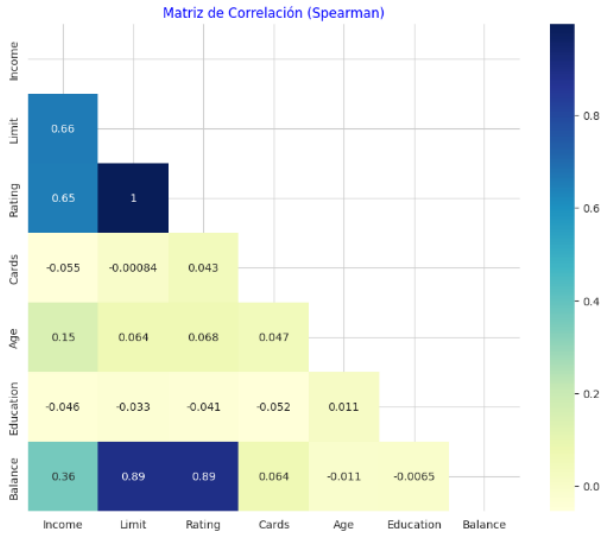


Figura 5: Descripción de la imagen.

Análisis de algunas de las correlaciones más destacadas:

- **Limit y Balance:** Existe una fuerte correlación positiva entre *Limit* y *Balance* (0.89). Esto sugiere que a medida que aumenta el límite de crédito, también tiende a aumentar el balance o deuda de los individuos.
- **Rating y Balance:** También hay una alta correlación positiva entre *Rating* y *Balance* (0.89), lo que indica que las personas con una calificación de crédito más alta tienden a tener balances más altos.
- **Income y Limit:** La correlación entre *Income* y *Limit* es 0.66, lo que muestra una relación moderada positiva. Esto sugiere que los ingresos están relacionados con el límite de crédito, aunque no de manera tan fuerte como con el balance.
- **Income y Rating:** La correlación entre *Income* y *Rating* es de 0.65, lo que también indica una relación positiva moderada entre el ingreso y la calificación de crédito.
- **Otras correlaciones:** La mayoría de las otras relaciones, como las que involucran *Cards*, *Age*, y *Education*, muestran valores de correlación cercanos a cero. Esto sugiere que no hay una relación lineal significativa entre estas variables y las demás en el contexto de esta matriz de correlación.

En resumen, los datos sugieren que el límite de crédito (*Limit*), el balance (*Balance*) y la calificación de crédito (*Rating*) están relacionados entre sí y también tienen una relación moderada con el ingreso (*Income*). Las demás variables (*Cards*, *Age*, y *Education*) no muestran correlaciones significativas con las demás.

2.2 Graficos Biplot

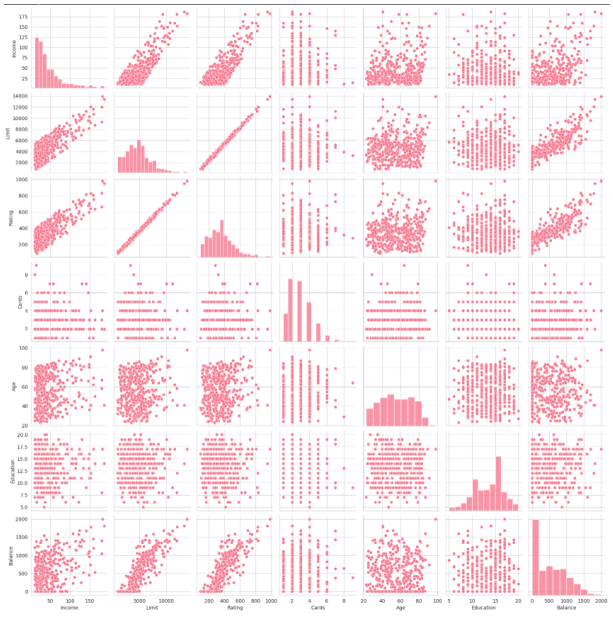


Figura 6: Descripción de la imagen.

Relaciones Lineales Notables:

- **Income vs. Limit** y **Rating vs. Balance**: Se observan relaciones lineales positivas claras, lo que confirma las correlaciones positivas encontradas anteriormente en la matriz de correlación.
- **Limit vs. Rating**: También hay una relación lineal positiva destacada, lo que sugiere que estas variables están relacionadas en el conjunto de datos.

Distribuciones Univariadas:

- **Income**: Tiene una distribución sesgada a la derecha, lo que significa que la mayoría de los ingresos están en el rango inferior, pero existen algunos valores significativamente altos.
- **Limit y Rating**: Ambas variables parecen tener distribuciones más concentradas en torno a un valor central, con algunos valores extremos hacia el extremo superior.
- **Age y Education**: Estas variables muestran distribuciones relativamente más uniformes y sin sesgos significativos.

Relaciones entre Variables Discretas y Continuas:

- **Cards**: Se trata de una variable discreta, y la mayoría de los puntos están alineados en niveles específicos en sus relaciones con otras variables. Esto muestra que el número de tarjetas (*Cards*) no varía de manera continua y tiene un número limitado de valores posibles.

- **Education**: Aunque no es tan discreta como *Cards*, también muestra agrupamientos específicos, especialmente en su relación con otras variables como *Age* y *Balance*.

Variables con Poca o Ninguna Correlación Visible:

- **Age** y **Education** con otras variables no presentan patrones claros en los gráficos de dispersión, lo que concuerda con las correlaciones bajas observadas en la matriz de correlación. Estas variables parecen estar distribuidas de manera independiente respecto a otras variables como *Income* y *Limit*.

Resumen de las Relaciones Observadas

Las relaciones más fuertes parecen estar entre las variables financieras (*Income*, *Limit*, *Rating*, *Balance*). Estas relaciones sugieren que personas con mayores ingresos o límites de crédito tienden a tener balances y calificaciones de crédito más altas.
Variables como *Age*, *Education* y *Cards* tienen una influencia menor o no muestran patrones claros con las otras variables.
Este tipo de gráfico es útil para identificar relaciones y patrones visuales en los datos, así como para detectar posibles relaciones lineales o no lineales entre variables. En general, el gráfico muestra que las variables financieras están fuertemente relacionadas, mientras que las variables demográficas y de educación tienen una menor correlación con estas.

2.2 Interacciones entre Balance y las demás columnas

Figura 7: Descripción de la imagen.

- **Ingresos y Balance:** Hombres y mujeres pueden tener diferentes valores de Balance en niveles similares de ingresos. A medida que los ingresos aumentan, el Balance generalmente también aumenta. Sin embargo, esta relación no varía mucho según el género.
- **Límite de Crédito y Balance:** A medida que aumenta el Límite de Crédito, el Balance también aumenta tanto para hombres como para mujeres. Esta fuerte relación positiva es similar sin importar el género.
- **Calificación Crediticia y Balance:** A medida que aumenta la Calificación Crediticia, los valores de Balance también aumentan tanto para hombres como para mujeres. Esta es otra relación positiva en la que la diferencia de género no es significativa.
- **Tarjetas y Balance:** El número de tarjetas no muestra una tendencia clara de aumento o disminución en los valores de Balance. Sin embargo, tener más tarjetas podría estar relacionado con un Balance más alto para ambos géneros.
- **Edad y Balance:** No hay una tendencia clara de aumento o disminución en los valores de Balance con la edad. Tanto hombres como mujeres tienen valores de Balance similares en varios grupos de edad.
- **Educación y Balance:** No se observa una tendencia clara de aumento o disminución en los valores de Balance según el nivel de educación. El nivel de educación no parece tener un efecto significativo en el Balance.
- **Género y Balance:** Este gráfico muestra directamente la distribución de Balance por categoría de género. No hay una diferencia significativa en las distribuciones de Balance entre hombres y mujeres.

- **Estudiante y Balance:** Ser estudiante o no, no afecta significativamente los valores de Balance. Tanto los estudiantes hombres como mujeres tienen valores de Balance similares.
- **Estado Civil y Balance:** El estado civil no afecta significativamente los valores de Balance. Tanto hombres como mujeres tienen valores de Balance similares independientemente de su estado civil.
- **Etnicidad y Balance:** No hay una diferencia significativa en los valores de Balance entre las diferentes etnias. Tanto hombres como mujeres tienen valores de Balance similares en varios grupos étnicos.

  Las variables financieras como el Límite de Crédito y la Calificación Crediticia tienen un fuerte efecto positivo en el Balance, y este efecto es similar sin importar el género. Otras variables (Ingresos, Edad, Educación, Tarjetas, Estudiante, Estado Civil, Etnicidad) tienen efectos menos notorios en el Balance, y estos efectos no varían mucho según el género. El género en sí mismo no parece tener un efecto significativo en el Balance, lo que significa que no hay una diferencia significativa en la distribución de Balance entre hombres y mujeres.

2.3 Análisis de Componentes Principales (PCA)

El Análisis de Componentes Principales (PCA) es una técnica estadística que busca encontrar combinaciones lineales de variables que puedan resumir el conjunto de datos, minimizando la pérdida de información. Su objetivo es reducir la complejidad de los datos mientras se conserva la mayor cantidad posible de variabilidad, optimizando así el análisis y la visualización de estos.

En nuestro estudio de PCA, comenzamos con la estandarización de los datos, un paso clave para asegurar que cada variable contribuya de manera equitativa al análisis, independientemente de sus escalas originales. Posteriormente, realizamos la optimización para calcular los *loadings*, o cargas, que maximizan la varianza en cada componente principal.

Para determinar el número óptimo de componentes principales, fijamos un umbral de varianza explicada (normalmente esta entre 75 %-95 %). Este umbral permite retener la mayor parte de la información con el menor número de componentes. Decidimos realizar el análisis con 8 componentes principales a partir de las 15 variables originales, obteniendo los siguientes resultados:

```
[ ] data3=pd.DataFrame({'PCA1':pca_3[:,0],'PCA2':pca_3[:,1],'PCA3':pca_3[:,2],'PCA4':pca_3[:,3],'PCA5':pca_3[:,4],'PCA6':pca_3[:,5],'PCA7':pca_3[:,6],'PCA8':pca_3[:,7]})
    data3.head()
```

	PCA1	PCA2	PCA3	PCA4	PCA5	PCA6	PCA7	PCA8
0	-1.237876	1.053816	-0.593969	0.435355	0.504847	-2.076574	-0.498902	-1.049863
1	2.229418	-1.923547	-1.685528	1.582598	0.148771	1.365878	0.880569	-1.330745
2	1.889139	-1.492836	-0.715599	-0.606449	1.970124	0.927666	-0.254352	-1.821625
3	3.801434	-1.514449	-1.331251	-0.044570	1.462369	-0.580188	-1.246269	-1.509645
4	-0.040214	0.991661	-0.208069	0.036471	-0.549653	-0.570772	0.913251	-2.086051

```
pca3.explained_variance_ratio_
```

```
array([0.2642799 , 0.13057167, 0.10858693, 0.08834163, 0.08319058,
       0.07901335, 0.07466389, 0.0712236 ])
```

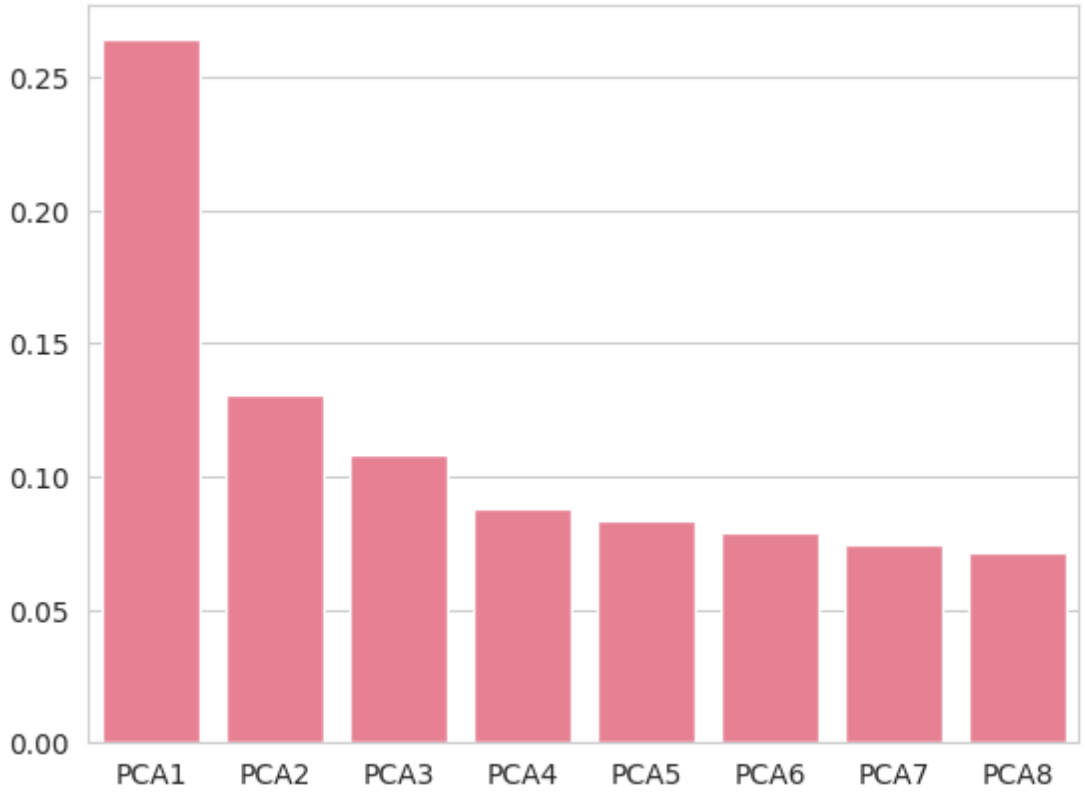
```
[ ] pca3.explained_variance_ratio_.sum()
```

```
0.8998715344160287
```

Es decir

- PCA1 = 0.264
- PCA2 = 0.130
- PCA3 = 0.108
- PCA4 = 0.088
- PCA5 = 0.083
- PCA6 = 0.079
- PCA7 = 0.074
- PCA8 = 0.071

Cada componente principal explica un porcentaje de la variabilidad total en el conjunto de datos. Al sumar las varianzas explicadas de los 8 componentes principales, se alcanza un total de 0.89, es decir, un 89 % de la varianza total. Esto significa que el ACP logra capturar la mayoría de la variabilidad presente en los datos, sacrificando solo un 11 % de la información original al reducir el número de variables de 15 a 8. Gráficamente, se ve de la siguiente manera.



2.4 Codificación de Variables Categóricas

Las columnas incluyen variables relacionadas con ingresos (`Income`), límites de crédito (`Limit`), calificaciones crediticias (`Rating`), y otras características como edad (`Age`), nivel educativo (`Education`), género (`Gender`), estado civil (`Married`), y saldos de tarjeta de crédito (`Balance`), las cuales no tendrán gran repercusión a la hora de analizar los datos estadísticamente.

La situación cambia en el caso de las "Variables categóricas", las cuales incluyen variables como `Gender`, `Student`, `Married`, y grupos étnicos (`Caucasian`, `Asian`, `African Amer`). Estas variables han sido transformadas a valores binarios (0 y 1), un procedimiento conocido como **dummificación**, lo cual es útil para modelos de *machine learning* y análisis estadísticos.

2.4 Finalidad del Conjunto de Datos

Este conjunto es ideal para:

- *Modelos de riesgo crediticio*: Usar variables como ingresos, límite de crédito, saldos actuales y calificación crediticia para predecir la probabilidad de incumplimiento.
- *Segmentación de clientes*: Identificar grupos de clientes con características comunes para estrategias de marketing.
- *Estudios demográficos y financieros*: Analizar cómo las características personales (edad, género, estado civil) influyen en el comportamiento financiero.

2.4 Codificación

A continuación, se presentan las variables categóricas transformadas a formato binario, con el significado de cada valor:

- *Gender*: se refiere al género de los clientes:
  - 0: Representa a las personas del género femenino.
  - 1: Representa a las personas del género masculino.
- *Student*: indica si el cliente es un estudiante o no:
  - 0: El cliente no es estudiante.
  - 1: El cliente es estudiante.
- *Married*: indica el estado civil del cliente:
  - 0: El cliente no está casado (soltero, divorciado, viudo, etc.).
  - 1: El cliente está casado.
- *Caucasian*: indica si el cliente se identifica como caucásico (de origen europeo o blanco).
  - 0: El cliente no se identifica como caucásico.

- 1: El cliente se identifica como caucásico.

- *Asian* y *African*: indican si el cliente se identifica como asiático o afrodescendiente respectivamente:
  - 0: El cliente no se identifica.
  - 1: El cliente se identifica.

Estas variables codificadas nos ayudarán a realizar la correlación entre variables como `Income`, `Limit`, `Rating`, y `Balance` para identificar patrones y el uso de modelos de regresión para predecir `Balance` basado en las demás variables.

2.5 Cálculo de la Distancia de Mahalanobis

El siguiente código en Python calcula la distancia de Mahalanobis para cada observación en un conjunto de datos y detecta valores atípicos usando un umbral basado en la distribución chi-cuadrada.

```
1
2  # Cargar los datos desde un archivo CSV
3  import pandas as pd
4  import numpy as np
5  from scipy.spatial.distance import mahalanobis
6  from scipy.stats import chi2
7
8  # Cargar el dataset
9  data = pd.read_csv('Credit_alterado.csv')
10
11 # Distancia de Mahalanobis
12 def calculate_mahalanobis(df2):
13     # Calcular la media y la matriz de covarianza del DataFrame
14     mean = np.mean(df2, axis=0)
15     cov_matrix = np.cov(df2.values.T)
16
17     # Usar pseudo-inversa en lugar de la inversa regular
18     inv_cov_matrix = np.linalg.pinv(cov_matrix)  # Cambiado a pinv
19
20     # Calcular la distancia de Mahalanobis para cada observacion
21     mahalanobis_distances = df2.apply(lambda row: mahalanobis(row, mean,
       inv_cov_matrix), axis=1)
22
23     return mahalanobis_distances
24
25 # Distancia de Mahalanobis robusta
26 from sklearn.covariance import MinCovDet
27
28 def mahalanobis_robusta(df2):
29     # Calcular la media y la matriz de covarianza del DataFrame
30     array_df2 = df2.select_dtypes(include=np.number).to_numpy()
31     mcd = MinCovDet().fit(array_df2)
```

```
    mean = mcd.location_
    cov_robusta = mcd.covariance_

    # Usar pseudo-inversa en lugar de la inversa regular
    inv_cov_matrix = np.linalg.pinv(cov_robusta)

    # Calcular la distancia de Mahalanobis robusta para cada observacion
    mahalanobis_distances_robusta = df2.select_dtypes(include=np.number)
    .apply(lambda row: mahalanobis(row, mean, inv_cov_matrix), axis=1)

    return mahalanobis_distances_robusta

# Deteccion de outliers
def detect_outliers_mahalanobis(df2, threshold=0.95):
    # Calcular las distancias de Mahalanobis
    distances = calculate_mahalanobis(df2)

    # Calcular el valor critico de la chi-cuadrada para detectar
    outliers
    chi2_threshold = chi2.ppf(threshold, df2.shape[1])

    # Identificar los outliers (distancias mayores al valor critico)
    outliers = distances > np.sqrt(chi2_threshold)

    return outliers
```

Listing 1: Cálculo de la distancia de Mahalanobis y detección de outliers

La función **calculate_mahalanobis** calcula la distancia de Mahalanobis para cada observación, utilizando la media y la matriz de covarianza del conjunto de datos.

La función **mahalanobis_robusta** calcula una versión robusta de la distancia de Mahalanobis utilizando el estimador de mínima covarianza, lo que mejora la precisión en presencia de valores atípicos.

Finalmente, la función **detect_outliers_mahalanobis** utiliza la distribución chi-cuadrada para identificar valores atípicos en el conjunto de datos, comparando las distancias de Mahalanobis con un umbral.

2.5 Conclusión

La distancia de Mahalanobis es una herramienta valiosa en el análisis multivariado, ya que no solo mide la distancia entre las observaciones y las medias, sino que también tiene en cuenta las correlaciones entre las variables. Este enfoque permite identificar valores atípicos en los datos y realizar análisis discriminante de manera más precisa.

3 Conclusiones

En resumen, este proyecto no solo validó la utilidad del análisis multivariado en entornos financieros, sino que también proporcionó un marco metodológico robusto para la toma de decisiones basada en datos, destacando oportunidades para optimizar productos crediticios y políticas de riesgo, como por ejemplo:.

3.1 Importancia del Análisis Multivariado en Decisiones Financieras

El estudio demostró que el análisis multivariado es una herramienta fundamental para comprender las relaciones complejas entre múltiples variables en contextos financieros. A través de técnicas como PCA, regresión y la distancia de Mahalanobis, se logró identificar patrones clave, como la fuerte correlación entre el límite de crédito, el saldo y la calificación crediticia, lo que permite a las instituciones financieras optimizar estrategias de riesgo y marketing.

3.2 Reducción de Dimensionalidad y Eficiencia

El Análisis de Componentes Principales (PCA) permitió reducir la dimensionalidad de los datos conservando el 89 % de la variabilidad original. Esto facilitó la visualización y el manejo de la información sin perder insights críticos, destacando la utilidad de PCA en conjuntos de datos con múltiples variables correlacionadas.

3.3 Identificación de Valores Atípicos

Mediante métodos como el rango intercuartílico, z-score y la distancia de Mahalanobis, se detectaron observaciones atípicas que podrían distorsionar los análisis. Estos hallazgos son cruciales para mejorar la calidad de los modelos predictivos y garantizar decisiones basadas en datos representativos.

3.4 Segmentación y Comportamiento del Cliente

El análisis reveló que variables demográficas como género y etnicidad tienen menor impacto en el saldo de tarjetas de crédito en comparación con factores financieros (ingresos, límite de crédito). Esto sugiere que las estrategias de segmentación deberían enfocarse más en características económicas que en demográficas para predecir comportamientos crediticios.

3.5 Limitaciones y Recomendaciones

El desbalance en variables como Student y la presencia de ceros en Balance sugieren la necesidad de técnicas de muestreo o ponderación para mejorar los modelos.

Futuros estudios podrían incorporar análisis no lineales o técnicas de machine learning avanzado para capturar relaciones más complejas.

4 Fuentes

Referencias

[1] Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis* (6th ed.). Pearson Prentice Hall.

[2] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.

[3] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.