Tabla de contenido

[Contexto 2](#_Toc85303387)

[Objetivo 2](#_Toc85303388)

[Fuente de datos 2](#_Toc85303389)

[Preparación del entorno de ejecución 2](#_Toc85303390)

[1) Importar las bibliotecas necesarias 2](#_Toc85303391)

[2) Importar los datos 2](#_Toc85303392)

[Selección de características 3](#_Toc85303393)

[Evaluación visual 3](#_Toc85303394)

[Matriz de correlaciones 4](#_Toc85303395)

[Selección de variables 5](#_Toc85303396)

[Aplicación del algoritmo 5](#_Toc85303397)

[Conclusiones de los clústeres obtenidos 8](#_Toc85303398)

[Clúster 0: 8](#_Toc85303399)

[Clúster 1: 9](#_Toc85303400)

[Clúster 2: 10](#_Toc85303401)

[Clúster 3: 10](#_Toc85303402)

[Clúster 4: 11](#_Toc85303403)

[Clúster 5: 12](#_Toc85303404)

[Clúster 6: 13](#_Toc85303405)

[Conclusiones 14](#_Toc85303406)

[Link de Google Colab 14](#_Toc85303407)

# Contexto

Objetivo: Obtener clústeres de casos de usuarios, con características similares, evaluados para la adquisición de una casa a través de un crédito hipotecario con tasa fija a 30 años.

## Fuente de datos

* ingresos: son ingresos mensuales de 1 o 2 personas, si están casados.
* gastos\_comunes: son gastos mensuales de 1 o 2 personas, si están casados.
* pago\_coche
* gastos\_otros
* ahorros
* vivienda: valor de la vivienda.
* estado\_civil: 0-soltero, 1-casado, 2-divorciado
* hijos: cantidad de hijos menores (no trabajan).
* trabajo: 0-sin trabajo, 1-autonomo, 2-asalariado, 3-empresario, 4-autonomos, 5-asalariados, 6-autonomo y asalariado, 7-empresario y autónomo, 8-empresarios o empresario y autónomo
* comprar: 0-alquilar, 1-comprar casa a través de crédito hipotecario con tasa fija a 30 años.

# Preparación del entorno de ejecución

## Importar las bibliotecas necesarias

import pandas as pd               # Para la manipulación y análisis de datos

import numpy as np                # Para crear vectores y matrices n dimensionales

import matplotlib.pyplot as plt   # Para la generación de gráficas a partir de los datos

import seaborn as sns             # Para la visualización de datos basado en matplotlib

%matplotlib inline

## Importar los datos

Fuente de datos: Hipoteca.csv

from google.colab import files

files.upload()

# Para importar los datos desde Drive

#from google.colab import drive

#drive.mount('/content/drive')

Hipoteca = pd.read\_csv("Hipoteca.csv")

Hipoteca

Pantalla de un video juego

Descripción generada automáticamente con confianza media

Hipoteca.info()

Un texto con letras negras

Descripción generada automáticamente

print(Hipoteca.groupby('comprar').size())

Imagen de la pantalla de un celular de un mensaje en letras blancas

Descripción generada automáticamente con confianza baja

Se puede observar que **135 usuarios desean alquilar**, mientras que **67 usuarios desean comprar** una casa a través de un crédito hipotecario con tasa fija a 30 años.

# Selección de características

## Evaluación visual

sns.pairplot(Hipoteca, hue='comprar')

plt.show()

Texto, Pizarra

Descripción generada automáticamente

sns.scatterplot(x='ahorros', y ='ingresos', data=Hipoteca, hue='comprar')

plt.title('Gráfico de dispersión')

plt.xlabel('Ahorros')

plt.ylabel('Ingresos')

plt.show()

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

## Matriz de correlaciones

Una matriz de correlaciones es útil para analizar la relación entre las variables numéricas. Se emplea la función corr().

CorrHipoteca = Hipoteca.corr(method='pearson')

CorrHipoteca

Imagen que contiene ventana, reloj, tabla, computadora

Descripción generada automáticamente

print(CorrHipoteca['ingresos'].sort\_values(ascending=False)[:10], '\n')   #Top 10 valores

Imagen que contiene Texto

Descripción generada automáticamente

Se muestra la correlación que tiene la variable **ingresos** con las demás variables.

# Mapa de calor de la relación que existe entre variables

plt.figure(figsize=(14,7))

MatrizInf = np.triu(CorrHipoteca)

sns.heatmap(CorrHipoteca, cmap='RdBu\_r', annot=True, mask=MatrizInf)

plt.show()

Gráfico, Histograma

Descripción generada automáticamente

## Selección de variables

1. A pesar de existir 2 correlaciones altas, entre 'ingresos' y 'ahorros' (0.71) y 'trabajo' e 'hijos' (0.69); éstas se tomarán en cuenta para obtener una segmentación que combine las variables mediante la similitud de los elementos.
2. Se suprimirá la variable 'comprar' debido a que representa inherentemente un agrupamiento, y fue un campo calculado con base a un análisis hipotecario preliminar.

MatrizHipoteca = np.array(Hipoteca[['ingresos', 'gastos\_comunes', 'pago\_coche', 'gastos\_otros', 'ahorros', 'vivienda', 'estado\_civil', 'hijos', 'trabajo']])

pd.DataFrame(MatrizHipoteca)

#MatrizHipoteca = Hipoteca.iloc[: , 0:9].values     #iloc para seleccionar filas y columnas según su posición

Interfaz de usuario gráfica, Texto, Calendario

Descripción generada automáticamente

# Aplicación del algoritmo

**Algoritmo: Jerárquico Ascendente**

Cuando se trabaja con clustering, dado que son algoritmos basados en distancias, es fundamental escalar los datos para que cada una de las variables contribuyan por igual en el análisis.

from sklearn.preprocessing import StandardScaler, MinMaxScaler

estandarizar = StandardScaler()                               # Se instancia el objeto StandardScaler o MinMaxScaler

MEstandarizada = estandarizar.fit\_transform(MatrizHipoteca)   # Se calculan la media y desviación y se escalan los datos

pd.DataFrame(MEstandarizada)

MinMaxScaler se usa cuando no hay datos fuera de rango. Estos datos atípicos se pueden remover de nuestro análisis.

Cuando hay datos atípicos se usa la función StandardScaler.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

#Se importan las bibliotecas de clustering jerárquico para crear el árbol

import scipy.cluster.hierarchy as shc

from sklearn.cluster import AgglomerativeClustering

plt.figure(figsize=(10, 7))

plt.title("Casos de hipoteca")

plt.xlabel('Hipoteca')

plt.ylabel('Distancia')

Arbol = shc.dendrogram(shc.linkage(MEstandarizada, method='complete', metric='euclidean')) #Utilizamos la matriz estandarizada

#plt.axhline(y=6, color='orange', linestyle='--') # Hace un corte en las ramas

#Probar con otras mediciones de distancia (chebyshev, cityblock, etc.)

Gráfico

Descripción generada automáticamente

Se puede observar que hay 7 ramas en nuestro árbol, cada una con su respectivo color.

#Se crean las etiquetas de los elementos en los clústeres

MJerarquico = AgglomerativeClustering(n\_clusters=7, linkage='complete', affinity='euclidean')

MJerarquico.fit\_predict(MEstandarizada)

MJerarquico.labels\_

Imagen en blanco y negro

Descripción generada automáticamente con confianza media

Hipoteca = Hipoteca.drop(columns=['comprar'])

Hipoteca['clusterH'] = MJerarquico.labels\_

Hipoteca

Pantalla de juego de computadora

Descripción generada automáticamente con confianza baja

#Cantidad de elementos en los clusters

Hipoteca.groupby(['clusterH'])['clusterH'].count()

Imagen que contiene Tabla

Descripción generada automáticamente

Hipoteca[Hipoteca.clusterH == 6] # Filtramos los clusters etiquetados como 6, que son 6 elementos

Pantalla de un video juego

Descripción generada automáticamente con confianza media

Los datos etiquetados con el 6, son los que están en color rojo:

Gráfico

Descripción generada automáticamente

CentroidesH = Hipoteca.groupby('clusterH').mean()

CentroidesH

Los centroides es el promedio de los valores contenidos en los clústeres.

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza baja

## Conclusiones de los clústeres obtenidos

### Clúster 0:

1. Este clúster está conformado por **30 casos de una evaluación hipotecaria.**
2. Con un **ingreso promedio mensual de 3421.13 USD**
3. Con **gastos comunes promedios de 846.47 USD**
4. Tienen un pago **promedio mensual de su coche de 309.93 USD**
5. **Otros gastos en promedio de 527.23 USD**.

Estos gastos en promedio representan el 49.21 % del ingreso promedio mensual, es decir, casi la mitad del salario mensual.

1. Por otro lado, este grupo de usuarios tienen un **ahorro promedio de 24289.63 USD**
2. y un **valor promedio de vivienda (a comprar o hipotecar) de 295590.7 USD**
3. Además, su **estado civil en promedio es: 0.23** [ 0-soltero, 1-casado, 2-divorciado ]

En este punto se puede observar que en su mayoría son solteros.

Interfaz de usuario gráfica

Descripción generada automáticamente

1. Tienen en promedio, 0.0 hijos menores, es decir, **no tienen hijos menores.**
2. y tienen en promedio un tipo de trabajo 2.0 [ 0-sin trabajo, 1-autónomo, 2-asalariado, 3-empresario, 4-autónomos, 5-asalariados, 6-autónomo y asalariado, 7-empresario y autónomo, 8-empresarios o empresario y autónomo ]

Se puede observar que la **mayoría son Autónomos:**

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Basándome en mi propio criterio, yo consideraría no darle el crédito a este grupo de usuarios, ya que sus gastos son muy elevados (gastan la mitad de lo que ganan) y además en su mayoría son trabajadores autónomos, por lo que se corre el riesgo de que no paguen a tiempo porque sus ingresos son algo volátiles, es decir, no tienen un ingreso fijo. Por otra parte, su ahorro promedio no es muy considerable.

### Clúster 1:

1. Este clúster está conformado por **51 casos de una evaluación hipotecaria.**
2. Con un **ingreso promedio mensual de 6394.02 USD**
3. Con **gastos comunes promedios de 1021.63 USD**
4. Tienen un **pago promedio mensual de su coche de 192.27 USD**
5. **Otros gastos en promedio de 533.04 USD**

Estos gastos en promedio representan el 27.32 % del ingreso promedio mensual.

1. Por otro lado, este grupo de usuarios tienen un **ahorro promedio de 54382.53 USD**
2. y un **valor promedio de vivienda (a comprar o hipotecar) de 421178.76 USD**
3. Además, su **estado civil en promedio es: 1.49** [ 0-soltero, 1-casado, 2-divorciado ]

Se puede observar que en su mayoría son casados.

Imagen que contiene Texto

Descripción generada automáticamente

1. Tienen en **promedio, 2.25 hijos menores**

La mayoría tiene un hijo:

Un conjunto de letras blancas en un fondo blanco

Descripción generada automáticamente con confianza media

1. y tienen en **promedio un tipo de trabajo 6.31** [ 0-sin trabajo, 1-autónomo, 2-asalariado, 3-empresario, 4-autónomos, 5-asalariados, 6-autónomo y asalariado, 7-empresario y autónomo, 8-empresarios o empresario y autónomo ]

Se puede observar que hay más asalariados:

Un conjunto de letras blancas en un fondo blanco

Descripción generada automáticamente con confianza media

Basándome en mi propio criterio, yo consideraría sí darle el crédito a este grupo de usuarios, ya que sus gastos son muy bajos en comparación a sus ingresos, la mayoría son casados, tienen hijos y son asalariados y empresarios, por lo que tienen un ingreso constante y el riesgo de que incumplan con los pagos es mínimo. En sumo a lo anterior, su ahorro promedio es una cantidad considerable.

### Clúster 2:

1. Este clúster está conformado por **35 casos de una evaluación hipotecaria.**
2. Con un **ingreso promedio mensual de 6599.54 USD**
3. Con **gastos comunes promedios de 1087.43 USD**
4. Tienen un **pago promedio mensual de su coche de 204.77 USD**
5. **Otros gastos en promedio de 362.6 USD**

Estos gastos en promedio representan el 25.07 % del ingreso promedio mensual.

1. Por otro lado, este grupo de usuarios tienen un **ahorro promedio de 51863.03 USD**
2. y un **valor promedio de vivienda (a comprar o hipotecar) de 515494.26 USD**
3. Además, su **estado civil en promedio es: 0.69** [ 0-soltero, 1-casado, 2-divorciado ]

En su mayoría son solteros:

Un conjunto de letras blancas en un fondo blanco

Descripción generada automáticamente con confianza media

1. Tienen en **promedio, 0.23 hijos menores**

La mayoría no tiene hijos:

Un conjunto de letras blancas en un fondo blanco

Descripción generada automáticamente con confianza media

1. y tienen en **promedio un tipo de trabajo 2.89** [ 0-sin trabajo, 1-autónomo, 2-asalariado, 3-empresario, 4-autónomos, 5-asalariados, 6-autónomo y asalariado, 7-empresario y autónomo, 8-empresarios o empresario y autónomo ]

Es decir, en su mayoría son empresarios:

Texto

Descripción generada automáticamente con confianza media

Basándome en mi propio criterio, yo consideraría sí darle el crédito a este grupo de usuarios, ya que sus gastos son bajos en comparación a sus ingresos, la mayoría son solteros, no tienen hijos y son empresarios, por lo que tienen un ingreso constante y considerable, por lo que el riesgo de que incumplan con los pagos es mínimo. En sumo a lo anterior, su ahorro promedio es considerable.

### Clúster 3:

1. Este clúster está conformado por **48 casos de una evaluación hipotecaria.**
2. Con un **ingreso promedio mensual de 3189.69 USD**
3. Con **gastos comunes promedios de 785.02 USD**
4. Tienen un **pago promedio mensual de su coche de 243.21 USD**
5. **Otros gastos en promedio de 548.27 USD**

Estos gastos en promedio representan el 49.42 % del ingreso promedio mensual.

1. Por otro lado, este grupo de usuarios tienen un **ahorro promedio de 23616.85 USD**
2. y un **valor promedio de vivienda (a comprar o hipotecar) de 277066.69 USD**
3. Además, su **estado civil en promedio es: 1.65** [ 0-soltero, 1-casado, 2-divorciado ]

En su mayoría son divorciados:

Imagen que contiene Texto

Descripción generada automáticamente

1. Tienen **en promedio, 1.98 hijos menores**

La mayoría tiene 3 hijos:

Un conjunto de letras blancas en un fondo blanco

Descripción generada automáticamente con confianza media

1. y tienen en **promedio un tipo de trabajo 6.21** [ 0-sin trabajo, 1-autónomo, 2-asalariado, 3-empresario, 4-autónomos, 5-asalariados, 6-autónomo y asalariado, 7-empresario y autónomo, 8-empresarios o empresario y autónomo ]

En su mayoría son empresarios y autónomos:

Texto

Descripción generada automáticamente con confianza media

Basándome en mi propio criterio, yo consideraría no darle el crédito a este grupo de usuarios, ya que sus gastos son bastante elevados en comparación a sus ingresos (gastan casi la mitad de sus ingresos), la mayoría son divorciados, con varios hijos y son empresarios. Considero que aunque puedan tener un ingreso considerable, sus obligaciones y gastos son elevados (mantenimiento de hijos + divorcio) por lo que el riesgo de que incumplan con los pagos es algo elevado. En sumo a lo anterior, su ahorro promedio es bastante bajo para sus responsabilidades que tienen.

### Clúster 4:

1. Este clúster está conformado por **20 casos de una evaluación hipotecaria**.
2. Con un **ingreso promedio mensual de 4843.75 USD**
3. Con **gastos comunes promedios de 1009.2 USD**
4. Tienen un **pago promedio mensual de su coche de 122.2 USD**
5. **Otros** **gastos en promedio de 572.85 USD**

Estos gastos en promedio representan el 35.18 % del ingreso promedio mensual.

1. Por otro lado, este grupo de usuarios tienen un **ahorro promedio de 36340.65 USD**
2. y un **valor promedio de vivienda (a comprar o hipotecar) de 337164.85 USD**
3. Además, su **estado civil en promedio es: 0.05** [ 0-soltero, 1-casado, 2-divorciado ]

En su mayoría son solteros:

Imagen que contiene Texto

Descripción generada automáticamente

1. Tienen en **promedio, 0.1 hijos menores**

La mayoría no tiene hijos:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

1. y tienen en **promedio un tipo de trabajo 1.9** [ 0-sin trabajo, 1-autónomo, 2-asalariado, 3-empresario, 4-autónomos, 5-asalariados, 6-autónomo y asalariado, 7-empresario y autónomo, 8-empresarios o empresario y autónomo ]

La mayoría de ellos no tiene trabajo.

Un conjunto de letras blancas en un fondo blanco

Descripción generada automáticamente con confianza media

Basándome en mi propio criterio, yo consideraría no darle el crédito a este grupo de usuarios, ya que sus gastos son algo elevados en comparación con sus ingresos. Aunque la mayoría sea soltera y sin hijos, se debe observar que o no tienen trabajo, o son autónomos (la mayoría), por lo que tienen un ingreso volátil o no lo tienen, por lo que el riesgo de que incumplan con los pagos es elevado.

### Clúster 5:

1. Este clúster está conformado por **12 casos de una evaluación hipotecaria**.
2. Con un **ingreso promedio mensual de 4466.42 USD**
3. Con **gastos comunes promedios de 1315.08 USD**
4. Tienen un **pago promedio mensual de su coche de 114.42 USD**
5. **Otros** **gastos en promedio de 502.75 USD**

Estos gastos en promedio representan el 43.26 % del ingreso promedio mensual.

1. Por otro lado, este grupo de usuarios tienen un **ahorro promedio de 23276.17 USD**
2. y un **valor promedio de vivienda (a comprar o hipotecar) de 269429.92 USD**
3. Además, su **estado civil en promedio es: 1.67** [ 0-soltero, 1-casado, 2-divorciado ]

La mayoría está divorciada:

Imagen que contiene Texto

Descripción generada automáticamente

1. Tienen en **promedio, 2.42 hijos menores**

La mayoría tiene un hijo:

Un conjunto de letras blancas en un fondo blanco

Descripción generada automáticamente con confianza media

1. y tienen en promedio un tipo de trabajo 6.75 [ 0-sin trabajo, 1-autónomo, 2-asalariado, 3-empresario, 4-autónomos, 5-asalariados, 6-autonomo y asalariado, 7-empresario y autónomo, 8-empresarios o empresario y autónomo ]

**La mayoría son empresarios o empresarios y autónomos:**

Un conjunto de letras blancas en un fondo blanco

Descripción generada automáticamente con confianza media

Basándome en mi propio criterio, yo consideraría no darle el crédito a este grupo de usuarios, ya que sus gastos son bastante elevados en comparación a sus ingresos (gastan casi la mitad de sus ingresos), la mayoría son divorciados, con uno o varios hijos y son empresarios. Considero que aunque puedan tener un ingreso considerable, sus obligaciones y gastos son elevados (mantenimiento de hijos + divorcio) por lo que el riesgo de que incumplan con los pagos es algo elevado. En sumo a lo anterior, su ahorro promedio es bastante bajo para sus responsabilidades que tienen.

### Clúster 6:

1. Este clúster está conformado por **6 casos de una evaluación hipotecaria**.
2. Con un **ingreso promedio mensual de 6404.5 USD**
3. Con **gastos comunes promedios de 1176.17 USD**
4. Tienen un **pago promedio mensual de su coche de 168.33 USD**
5. **Otros** **gastos en promedio de 769.33 USD**

Estos gastos en promedio representan el 33.01 % del ingreso promedio mensual.

1. Por otro lado, este grupo de usuarios tienen un **ahorro promedio de 61715.5 USD**
2. y un **valor promedio de vivienda (a comprar o hipotecar) de 625138.83 USD**
3. Además, su **estado civil en promedio es: 0.0** [ 0-soltero, 1-casado, 2-divorciado ]

Es decir, **todos son solteros.**

1. Tienen en promedio, 0.0 hijos menores, es decir, **no tienen hijos.**
2. y tienen en promedio un tipo de trabajo 1.17 [ 0-sin trabajo, 1-autónomo, 2-asalariado, 3-empresario, 4-autónomos, 5-asalariados, 6-autonomo y asalariado, 7-empresario y autónomo, 8-empresarios o empresario y autónomo ]

**La mayoría no tienen trabajo o son autónomos:**

Un conjunto de letras blancas en un fondo blanco

Descripción generada automáticamente con confianza media

Basándome en mi propio criterio, yo consideraría no darle el crédito a este grupo de usuarios, ya que siento que los datos son engañosos, puesto que se menciona que 2 de estos usuarios no tienen empleo, pero ganan lo mismo que el usuario que es empresario. Sin tomar el tipo de trabajo, si daría el crédito, pero considerando el tipo de trabajo y que son desempleados, no. En este caso tal vez se necesite más la opinión de un experto (aunque claro, en todos se necesitaría de un Actuario, por ejemplo)

plt.figure(figsize=(10, 7))

plt.scatter(MEstandarizada[:,0], MEstandarizada[:,1], c=MJerarquico.labels\_)

plt.grid()

plt.show()

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

# Conclusiones

En esta práctica pude aprender cómo generar clústeres a partir de una fuente de datos, con el objetivo de analizar si cada uno de estos grupos (en total fueron 7 grupos obtenidos) es apto para acceder a la adquisición de una casa a través de un crédito hipotecario con tasa fija a 30 años.

Se implementó el algoritmo de ascendente jerárquico para este análisis, pero, como se está trabajando con clústering (que entra dentro de la categoría de aprendizaje no supervisado), los cuales son algoritmos basado en distancias, se tuvo que escalar los datos, de tal manera que cada variable contribuyera de igual manera en el análisis, es decir, que ninguna variable pesara más que la otra.

Finalmente, se obtuvieron 7 clústeres o 7 grupos de usuarios, los cuales pude analizar y dar mis conclusiones sobre si es óptimo otorgarles el crédito o no, basándome en mis propios criterios.

En esta práctica pude aprender y visualizar de mejor manera la aplicación que tienen las distancias en el análisis de datos, de forma específica en el tema de clústering.

# Link de Google Colab

🔗 [OCG-Práctica4-Clústering.ipynb - Colaboratory (google.com)](https://colab.research.google.com/drive/1tSPW58j2cJeOiq3uN0SOk8okJu_mHw7-#scrollTo=AnL5_NLhSZ8v)