

Analizando los datos de clientes de un supermercado

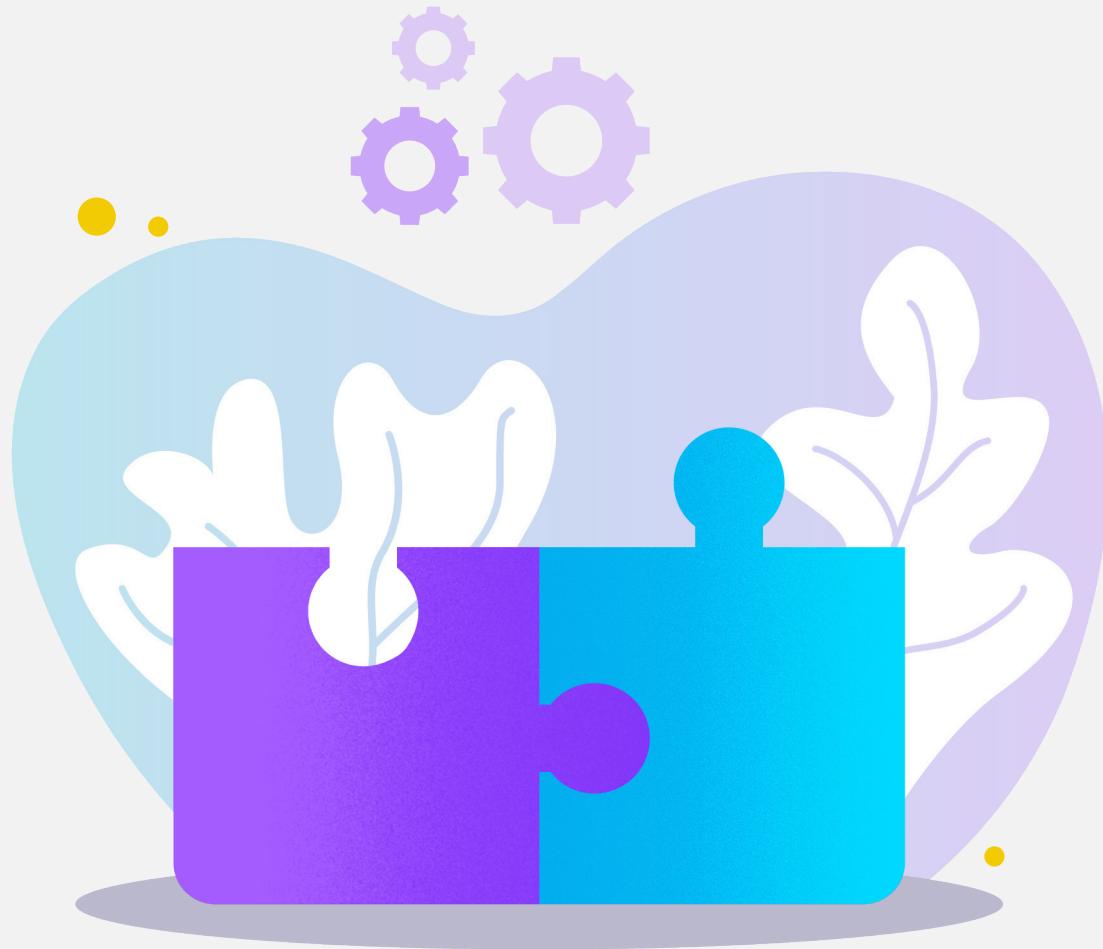




En este manual de ejercicios usarás lo aprendido sobre aprendizaje no supervisado para resolver cuatro misiones que pondrán a prueba tus conocimientos y habilidades adquiridas.

Responde cada uno de los ejercicios en este archivo interactivo, sigue las instrucciones de cada uno de ellos.

Una vez terminados tus ejercicios, ve al final del manual para revisar las respuestas correctas y compáralas con las tuyas. Puedes regresar a los ejercicios para aprovechar la retroalimentación.



Introducción

¿Recuerdas a Miqueas Hernández?

Le encomendaron la misión de hacer un modelo inteligente para determinar si los depósitos inusuales son sospechosos de lavado de dinero. Después de dos meses de trabajo arduo, logró encontrar anomalías en la base de datos de depósitos inusuales que le proporcionó Hugo Gallardo.

En el proceso desarrollo de la investigación, Miqueas realizó un programa script en Python para transformar los datos y agruparlos usando varios algoritmos de agrupamiento, así también implementó diferentes estrategias para seleccionar el número de grupos adecuados y visualizar los resultados para su interpretación.

Como resultado de este análisis aplicó métodos de transformación de datos, agrupamiento y estrategias de selección y comunicación por medio de un reporte, los resultados para segmentar datos de interés, a la vez que muestra información relevante para la toma de decisiones sobre casos de sospecha de lavado de dinero.

Autoevaluación

Como producto de su investigación, Miqueas obtuvo buenos resultados de encontrar suficiente base para sospechar que muchos depósitos inusuales son actividades ilícitas de lavado de dinero. Ahora es tu turno demostrar que seguirás el mismo proceso de análisis, para contribuir a la toma de decisión en la organización donde laboras.

Marca los elementos de la siguiente lista que describan cómo te percibes al final del módulo:

Reconozco los conceptos claves del aprendizaje no supervisado para tener claridad del modelo inteligente.	<input type="checkbox"/>
Describo los propósitos y áreas de aplicación del aprendizaje no supervisado para la toma de decisiones.	<input type="checkbox"/>
Analizo el tipo de transformación de datos numéricos que mejore los resultados.	<input type="checkbox"/>
Examo el tipo de transformación de datos categóricos necesaria para el algoritmo de agrupamiento en particular.	<input type="checkbox"/>
Reconozco la conceptualización del algoritmo de agrupamiento K-Means para comprender sus propiedades.	<input type="checkbox"/>

Valoro a partir de un ejemplo el algoritmo de agrupamiento K-Means identificando su procedimiento.	
Experimento una agrupación de datos usando K-Means para segmentar datos.	
Describo los diferentes índices de validación de los resultados para la selección de estrategias.	
Empleo gráficas K-Elbow para encontrar el valor óptimo de un hiperparámetro.	
Comparo las gráficas Silhouette Plot para cuantificar la eficiencia de un proceso comparado en los agrupamientos.	
Destaco la transformación de los datos basados en los resultados de agrupamiento para la toma de decisiones.	
Valoro la importancia relativa de las agrupaciones en un escenario de toma de decisiones.	
Selecciono las visualizaciones asociadas a los grupos obtenidos para una mejor interpretación de información.	

Si marcaste todos los elementos de la lista, ¡felicidades! Estás listo para practicar estos conocimientos a través de un caso. Si dejaste algún elemento sin marcar, te invitamos a que revises de nuevo el tema correspondiente y después retomes esta actividad.

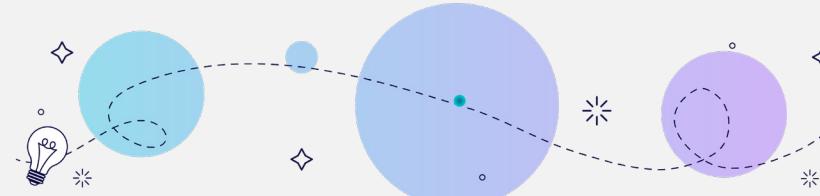


Ahora es tu turno, se te ha asignado la tarea de encontrar distintos grupos de clientes de acuerdo con sus ingresos y los gastos que realizan en un supermercado.

Ante las distintas formas en las que se almacena la información muchas veces resulta complicado poder organizarla, pues se puede recurrir desde utilizar archivos Excel o CSV y conjuntar a mano la información hasta utilizar algoritmos de agrupamiento de datos para obtener este tipo de información

Se requiere que dividas los datos en segmentos de clientes de acuerdo con sus características. Para completar esta tarea tienes a tu disposición una base de datos con información de los clientes como el género, la edad, los ingresos y el puntaje de gastos. Esta base de datos se llama “**Mall_Customers.csv**” y la puedes descargar desde el área de Archivos adjuntos. El sitio donde se obtuvieron estos datos fue el sitio oficial de Kaggle.

¿Tienes todo lo que necesitas? A continuación, se presentan cuatro misiones que deberás resolver para completar esta gran tarea que te asignaron. ¡Adelante!



Misión 1: Comienza por crear un *dataframe* con los datos

Propósito: Cargar y transformar los datos.

La información de los clientes del supermercado se encuentra en el archivo **Mall_Customers.csv**, el cuál es una base de datos obtenida de la página kaggle.com.

El archivo está estructurado en cinco columnas [CustomerID, Genre, Age, Annual Income (k\$), Spending Score (1-100)]

Antes de empezar a analizar los datos es importante seleccionar las columnas que sí aportarían valor, para que luego puedas transformar, de las que queden, las que no sean numéricas.

Por lo tanto, el objetivo es precisamente seleccionar las columnas adecuadas y transformar las no numéricas para tener los datos listos para agrupar. Además, debes transformar las variables numéricas para que todas tengan la misma magnitud.

¿Cuál es la manera correcta de cargar y transformar los datos?

Para resolver este ejercicio, usa el archivo **Mall_Customers.csv**

¿Ya tienes tu respuesta? ¡Vamos por esa primera misión!

Misión 1: Ejercicio 1

¿Cuál es la manera correcta de cargar y transformar los datos?

Resultado esperado:

Un `dataframe` con los datos cargados. Además, el uso de funciones de `Scikit_learn` para transformar los datos, en particular `OneHotEncoder()` o en su defecto una función de `pandas` para ese mismo propósito como `get_dummies()`.

a

```
from sklearn.  
preprocessing import OneHotEncoder,  
Normalizer  
  
import numpy as np  
  
import pandas as pd  
  
df2 = df.copy(deep=True)  
  
X_train = df2['Genre'].to_numpy()  
.reshape(-1,1)  
  
enc = OneHotEncoder()  
  
enc.fit(X_train)  
  
X_train = enc.transform(X_train).  
toarray()
```

b

```
from sklearn.  
preprocessing import OneHotEncoder,  
Normalizer  
  
import numpy as np  
  
import pandas as pd  
  
df2 = df.copy(deep=True)  
  
X_train = df2['Genre'].to_numpy()  
.reshape(-1,1)  
  
enc = OneHotEncoder()  
  
enc.fit(X_train)  
  
X_train = enc.transform(X_train).  
toarray()
```

c

```
from sklearn.  
preprocessing import OneHotEncoder,  
Normalizer  
  
import numpy as np  
  
import pandas as pd  
  
df2 = df.copy(deep=True)  
  
X_train = df2['Genre'].to_numpy()  
.reshape(-1,1)  
  
enc = OneHotEncoder()  
  
enc.fit(X_train)  
  
X_train = enc.transform(X_train).toarray()
```

```
values = df2['Genre'].unique()
nvalues = len(values)
i = 0
for v in values:
    df2['Genre_'+v] = X_train[:, i]
    i += 1
df2.drop('CustomerID', axis=1, inplace=True)
df2.drop('Genre', axis=1, inplace=True)
X_train = df2.to_numpy()
X = Normalizer().fit_transform(X_train)
```

```
values = df2['Genre'].unique()
nvalues = len(values)
i = 0
for v in values:
    df2['Genre_'+v] = X_train[:, i]
    i += 1
X_train = df2.to_numpy()
X = Normalizer().fit_transform(X_train)
```

```
values = df2['Genre'].unique()
nvalues = len(values)
i = 0
for v in values:
    df2['Genre_'+v] = X_train[:, i]
    i += 1
df2.drop('CustomerID', axis=1, inplace=True)
df2.drop('Genre', axis=1, inplace=True)
X_train = df2.to_numpy()
```

Para comprobar tu respuesta, conocer la retroalimentación y analizar paso a paso del proceso de resolución, pulsa el siguiente botón:

¡Vas en camino! Terminaste tu primera misión enfocada en cargar y transformar los datos, sin embargo, aún tienes tres más por resolver.

Continúa con el material para completar cada una de ellas.



Misión 2: Es momento de agrupar la información

Propósito: Agrupar y seleccionar el número de grupos adecuados.

Ahora que tienes una serie de datos transformados, lo siguiente es agrupar los datos de los clientes seleccionando el número de grupos adecuados luego de analizar la gráfica K-Elbow. **¿Qué código utilizarías para lograr esto?**

El resultado debe ser una variable con el vector de etiquetas adecuado para segmentar los datos de clientes.

Para este ejercicio, apóyate de los datos transformados listos para agrupar.

• ¡Continúa para completarlo!

Misión 2: Ejercicio 1

¿Qué código utilizarías para lograr esto?

El resultado esperado es una serie de datos agrupados con el número de grupos adecuados.

Selecciona la respuesta correcta.

a

```
results = {}

for k in range(2, nc):

    k_means = cluster.KMeans(n_
clusters=k).fit(X)

    y_pred = k_means.predict(X)

    results[k] = y_pred

clusters_counts = list(results.keys())
index_values = list(results.values())

plt.plot(clusters_counts, index_
values, 'o-')

plt.grid(True)
```

b

```
results = {}

for k in range(2, 11):

    k_means = cluster.KMeans(n_clusters=k) .
fit(X)

    y_pred = k_means.predict(X)

    results[k] = metrics.silhouette_-
score(X, y_pred)
```

c

```
results = {}

for k in range(2, nc):

    k_means = cluster.KMeans(n_clusters=k) .
fit(X)

    y_pred = k_means.predict(X)

    results[k] = metrics.silhouette_-
score(X, y_pred)

clusters_counts = list(results.keys())
index_values = list(results.values())

plt.plot(clusters_counts, index_values, 'o-')
```

```
plt.title('Maximize')

plt.xlabel('Num Clusters')
plt.xticks(clusters_counts)
plt.ylabel('Silhouette')
plt.show()

k = 5 # Debido principalmente a la gráfica K-Elbow

k_means = cluster.KMeans(n_clusters=k)
y_pred = k_means.fit_predict(X)
```

```
clusters_counts = list(results.keys())
index_values = list(results.values())
plt.plot(clusters_counts, index_values, 'o-')
plt.grid(True)

plt.title('Maximize')
plt.xlabel('Num Clusters')
plt.xticks(clusters_counts)
plt.ylabel('Silhouette')
plt.show() k = 5 # Debido principalmente a la gráfica K-Elbow

k_means = cluster.KMeans(n_clusters=k)
y_pred = k_means.fit_predict(X)
```

```
plt.grid(True)

plt.title('Maximize')
plt.xlabel('Num Clusters')
plt.xticks(clusters_counts)
plt.ylabel('Silhouette')
plt.show()

k = 5 # Debido principalmente a la gráfica K-Elbow

k_means = cluster.KMeans(n_clusters=k)
y_pred = k_means.fit_predict(X)
```

Comprueba tu resultado y los pasos necesarios para obtenerlo:

¡Lo estás haciendo increíble! Estás a la mitad de esta práctica, terminaste la segunda misión centrada en agrupar y seleccionar el número de grupos adecuados. Quedan sólo dos misiones más.

Continúa con el material para completar cada una de ellas.



Misión 3: Resumen de la segmentación

Propósito: Segmentar los datos y mostrar un resumen de la segmentación.

¡Vas muy bien! Llevas la mitad de esta práctica, cada vez estás más cerca de alcanzar tu objetivo. Lo que debes hacer ahora es segmentar los empleados generando una nueva tabla con los promedios de cada variable por grupo, además de dividir los datos creando nuevas bases de datos con los clientes repartidos por grupo.

El objetivo es generar una lista de *dataframes* con los clientes separados por grupo y un *dataframe* resumen de las variables por grupo.

¿Cuál código es el correcto para obtener los entregables?

Para ello, utiliza el *dataframe* con los datos de clientes y vector de etiquetas.

Misión 3: Ejercicio 1

¿Qué instrucciones consideras son las adecuadas para completar esta acción?

El resultado debe ser un *frame* llamado `jugadores2017` que contenga sólo las variables que pueden ser relevantes.

Selecciona la respuesta correcta.

a

```
clients = [df2[df2['groups']==g]
for g in range(k)]
grouped = pd.DataFrame()
grouped['Features'] = df2.columns[:-1]
for g in range(k):
    row = []
    for col in grouped['Features']:
```

b

```
clients = [df2[df2['groups']==g] for g
in range(k)]
grouped = pd.DataFrame()
grouped['Features'] = df2.columns[:-1]
for g in range(k):
    row = []
    for col in grouped['Features']:
        if clients[g][col].dtype != 'object':
```

c

```
clients = [df2[df2['groups']==g] for g in
range(k)]
grouped = pd.DataFrame()
grouped['Features'] = df2.columns[:-1]
for g in range(k):
    row = []
    for col in grouped['Features']:
        if clients[g][col].dtype != 'object':
```

```
    row.append(clients[g][col].  
value_counts().keys()[0])  
  
grouped['Group'+str(g)] = row
```

```
    row.append(np.round(clients[g]  
[col].mean(),2))  
else:  
    row.append(clients[g][col].  
value_counts().keys()[0])  
grouped['Group'+str(g)] = row
```

```
    row.append(np.round(clients[g]  
[col].mean(),2))  
else:  
    row.append(clients[g][col].value_  
counts().keys()[0])  
grouped['Group'+str(g)] = row
```

Recuerda confirmar tus resultados en el siguiente botón:

**¡Excelente trabajo! Terminaste
la tercera misión enfocada en
segmentar los datos y mostrar un
resumen de la segmentación.**

**Sólo queda una misión más. ¡Venga,
es la recta final!**



Misión 4: Gráficas y descripción de los resultados obtenidos

Propósito: Visualizar e interpretar los resultados.

Lo hiciste muy bien. Ahora estás en la última etapa de esta importante tarea, lo siguiente es crear una gráfica de barras con el promedio de variables por grupo y describe las características de los clientes en cada grupo.

Debes conseguir una gráfica del promedio de las variables por grupo y descripción de la segmentación.

Para lograrlo, utiliza los datos de clientes segmentados y descripción de las variables por grupo.

¿Qué características tienen los clientes que quedaron separados en grupos diferentes?

¡Vamos por esa última misión!

Misión 4: Ejercicio 1

¿Qué características tienen los clientes que quedaron separados en grupos diferentes?

El resultado que se espera es una gráfica de barras con la descripción de las variables e interpretación de los resultados para la toma de decisiones.

Selecciona la respuesta correcta.

a	b	c
<p>En el grupo 1 (de color naranja) vemos que las personas tienen altos ingresos y puntajes de gasto, este es el caso ideal para el centro comercial o las tiendas, ya que estas personas son las principales fuentes de ganancias. Estas personas pueden ser clientes habituales del centro comercial y están convencidas de las instalaciones del centro comercial.</p>	<p>En el grupo 1 (de color naranja) vemos que las personas tienen altos ingresos y puntajes de gasto, este es el caso ideal para el centro comercial o las tiendas, ya que estas personas son las principales fuentes de ganancias. Estas personas pueden ser clientes habituales del centro comercial y están convencidas de las instalaciones del centro comercial.</p>	<p>En el grupo 1 (de color naranja) vemos que las personas tienen altos ingresos y puntajes de gasto, este es el caso ideal para el centro comercial o las tiendas, ya que estas personas son las principales fuentes de ganancias. Estas personas pueden ser clientes habituales del centro comercial y están convencidas de las instalaciones del centro comercial.</p>

En el grupo 2 (de color amarillo) podemos ver que las personas tienen bajos ingresos anuales y puntajes de gasto bajos, esto es bastante razonable ya que las personas con bajos salarios prefieren comprar menos, de hecho, estas son las personas sabias que saben cómo gastar y ahorrar dinero. Las tiendas / centros comerciales estarán menos interesados en las personas que pertenecen a este grupo.

En el grupo 3 (de color verde) podemos ver que las personas tienen bajos ingresos, pero puntajes de gasto más altos, estas son aquellas personas que por alguna razón aman comprar productos con más frecuencia a pesar de que tienen bajos ingresos. Tal vez sea porque estas personas están más que satisfechas con los servicios del centro comercial. Es posible que las tiendas / centros comerciales no se dirijan a estas personas de manera efectiva, pero aun así no las perderán.

En el grupo 2 (de color amarillo) podemos ver que las personas tienen bajos ingresos anuales y puntajes de gasto bajos, esto es bastante razonable ya que las personas con bajos salarios prefieren comprar menos, de hecho, estas son las personas sabias que saben cómo gastar y ahorrar dinero. Las tiendas / centros comerciales estarán menos interesados en las personas que pertenecen a este grupo.

En el grupo 3 (de color verde) podemos ver que las personas tienen bajos ingresos, pero puntajes de gasto más altos, estas son aquellas personas que por alguna razón aman comprar productos con más frecuencia a pesar de que tienen bajos ingresos. Tal vez sea porque estas personas están más que satisfechas con los servicios del centro comercial. Es posible que las tiendas / centros comerciales no se dirijan a estas personas de manera efectiva, pero aun así no las perderán.

En el grupo 4 (de color azul) vemos que las personas tienen altos ingresos, pero puntajes de gasto bajos, esto es interesante.

En el grupo 2 (de color amarillo) podemos ver que las personas tienen bajos ingresos, pero puntajes de gasto más altos, estas son aquellas personas que por alguna razón aman comprar productos con más frecuencia a pesar de que tienen bajos ingresos. Tal vez sea porque estas personas están más que satisfechas con los servicios del centro comercial. Es posible que las tiendas / centros comerciales no se dirijan a estas personas de manera efectiva, pero aun así no las perderán.

En el grupo 3 (de color verde) podemos ver que las personas tienen bajos ingresos anuales y puntajes de gasto bajos, esto es bastante razonable ya que las personas con bajos salarios prefieren comprar menos, de hecho, estas son las personas sabias que saben cómo gastar y ahorrar dinero. Las tiendas / centros comerciales estarán menos interesados en las personas que pertenecen a este grupo.

En el grupo 4 (de color azul) vemos que las personas tienen ingresos promedio y una puntuación de gasto promedio,

En el grupo 4 (de color azul) vemos que las personas tienen ingresos promedio y una puntuación de gasto promedio, estas personas nuevamente no serán los principales objetivos de las tiendas o centros comerciales, pero nuevamente serán consideradas y se pueden usar otras técnicas de análisis de datos para aumentar su puntaje de gasto.

En el grupo 5 (de color rojo) vemos que las personas tienen altos ingresos, pero puntajes de gasto bajos, esto es interesante. Quizás estas son las personas que están insatisfechas o descontentas con los servicios del centro comercial. Estos pueden ser los principales objetivos del centro comercial, ya que tienen el potencial de gastar dinero. Entonces, las autoridades del centro comercial intentarán agregar nuevas instalaciones para que puedan atraer a estas personas y puedan satisfacer sus necesidades.

Quizás estas son las personas que están insatisfechas o descontentas con los servicios del centro comercial. Estos pueden ser los principales objetivos del centro comercial, ya que tienen el potencial de gastar dinero. Entonces, las autoridades del centro comercial intentarán agregar nuevas instalaciones para que puedan atraer a estas personas y puedan satisfacer sus necesidades.

En el grupo 5 (de color rojo) vemos que las personas tienen ingresos promedio y una puntuación de gasto promedio, estas personas nuevamente no serán los principales objetivos de las tiendas o centros comerciales, pero nuevamente serán consideradas y se pueden usar otras técnicas de análisis de datos para aumentar su puntaje de gasto.

estas personas nuevamente no serán los principales objetivos de las tiendas o centros comerciales, pero nuevamente serán consideradas y se pueden usar otras técnicas de análisis de datos para aumentar su puntaje de gasto.

En el grupo 5 (de color rojo) vemos que las personas tienen altos ingresos, pero puntajes de gasto bajos, esto es interesante. Quizás estas son las personas que están insatisfechas o descontentas con los servicios del centro comercial. Estos pueden ser los principales objetivos del centro comercial, ya que tienen el potencial de gastar dinero. Entonces, las autoridades del centro comercial intentarán agregar nuevas instalaciones para que puedan atraer a estas personas y puedan satisfacer sus necesidades.

No continúes sin corroborar si tu respuesta fue correcta. Usa el siguiente botón:

¡Felicidades, terminaste las cuatro misiones que tenías!

Ahora has agrupado datos de clientes de un supermercado y estás listo para describir los resultados y apoyar la toma de decisiones. Como parte de las conclusiones de la actividad, puedes deducir que, para aumentar las ganancias del supermercado, las autoridades deben apuntar a las personas que pertenecen al clúster 4 y al clúster 5 para lograr que aumenten sus puntajes de gasto y también deben mantener sus estándares para mantener a las personas que pertenecen al clúster 1 y al clúster 3 felices y satisfechos.



Cierre

Has resuelto de manera satisfactoria las cuatro principales acciones:

- Seleccionaste las columnas adecuadas y transformaste las no numéricas para tener los datos listos para agrupar. Además, transformaste las variables numéricas para que todas tengan la misma magnitud.
- Agrupaste los datos de clientes seleccionando el número de grupos adecuados luego de analizar la gráfica K-Elbow.
- Segmentaste los empleados generando una nueva tabla con los promedios de cada variable por grupo, además de dividir los datos creando nuevas bases de datos con los clientes repartidos por grupo.
- Creaste una gráfica de barras con el promedio de variables por grupo y describiste las características de los clientes en cada grupo.

Reflexión final

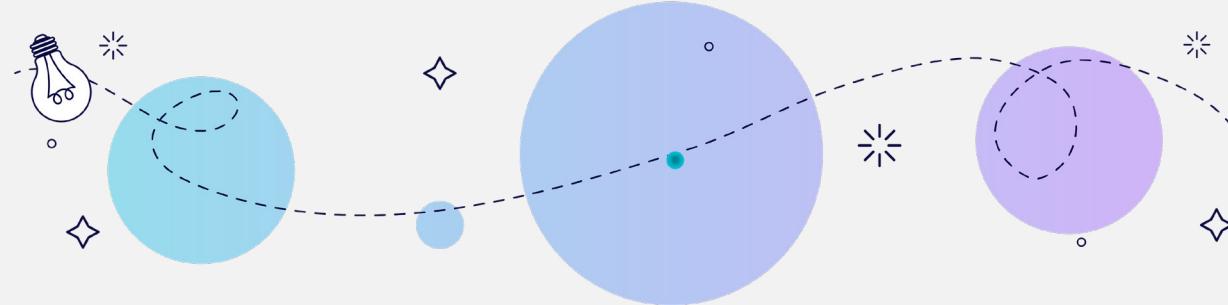
En esta práctica realizaste el proceso de agrupamiento de inicio a fin. Un detalle importante es que, a pesar de que en ocasiones el conjunto de datos es bastante completo y ya está limpio y en condiciones para su procesamiento, es indispensable analizarlo a profundidad y prepararlo acorde a los algoritmos de agrupamiento que serán usados. La razón principal es que estos algoritmos esperan un conjunto de datos en forma de matriz, por lo que este ajuste o preprocessamiento es prácticamente obligatorio.

Una vez que los datos están listos para ser la entrada de los algoritmos de agrupamiento, debes diseñar una estrategia para seleccionar el número de grupos en los que los datos serán agrupados. Este parámetro es necesario para la mayoría de los algoritmos de agrupamiento, por lo que este paso es muy importante. Una consideración es que, si el usuario final de este proceso, o el cliente que evalúa los resultados, tiene alguna noción de cuál sería un número de grupos adecuados, entonces puedes aprovechar esa información.

Seguidamente, la parte más importante es la segmentación de los datos. En la práctica solo usaste el algoritmo de agrupamiento más conocido, el K-means, pero debes saber que puedes usar otros como por ejemplo un algoritmo de agrupamiento jerárquico, si así lo requiere la solución. Es importante señalar que no se debe confundir esta parte con la selección del número de grupos adecuado. Ya en esta parte lo que estás logrando es dividir los datos en los grupos finales y dejar todo listo para explicarle al usuario lo encontrado en el proceso.

Finalmente, estás listo para interpretar los resultados obtenidos y para preparar conclusiones relevantes que ayuden en la toma de decisiones. Hay varias maneras de obtener distintas conclusiones, pero una estrategia muy extendida es revisar el comportamiento de las variables por grupo, incluso con la ayuda de visualizaciones. Esto permite caracterizar los grupos y redactar descripciones e interpretaciones de los resultados del agrupamiento. Note que todo este proceso es no supervisado, y solo en la interacción con el usuario final se puede analizar el impacto real de lo obtenido.

A continuación, realizarás un reto, el cual es una actividad evaluable en el que tendrás que aplicar lo aprendido en este curso y entregar en plataforma la evidencia lo solicitado.



• Respuestas y retroalimentación

Misión 1 / Ejercicio 1

Respuesta:

```
from sklearn.preprocessing import OneHotEncoder, Normalizer
import numpy as np
import pandas as pd
df2 = df.copy(deep=True)
X_train = df2['Genre'].to_numpy().reshape(-1,1)
enc = OneHotEncoder()
enc.fit(X_train)
X_train = enc.transform(X_train).toarray()
values = df2['Genre'].unique()
nvalues = len(values)
i = 0
for v in values:
    df2['Genre_'+v] = X_train[:, i]
    i += 1
df2.drop('CustomerID', axis=1, inplace=True)
df2.drop('Genre', axis=1, inplace=True)
X_train = df2.to_numpy()
X = Normalizer().fit_transform(X_train)
```

Retroalimentación: Transformas la columna Genre usando OneHotEncoder() ya que presenta valores categóricos, en particular Female y Male.

Además, eliminas las dos primeras columnas, ya que después de esta transformación ya la columna Genre no es relevante, y tampoco lo es CustomerID.

Veamos por qué las otras opciones no son correctas.

Opción	Retroalimentación
<pre>from sklearn.preprocessing import OneHotEncoder, Normalizer import numpy as np import pandas as pd df2 = df.copy(deep=True) X_train = df2['Genre'].to_numpy().reshape(-1,1) enc = OneHotEncoder() enc.fit(X_train) X_train = enc.transform(X_train).toarray() values = df2['Genre'].unique() nvalues = len(values) i = 0 for v in values:</pre>	<p>Esta opción es incorrecta</p> <p>Aunque sí transformas la columna Genre, no eliminas las columnas que ya no serán relevantes.</p>

```
df2['Genre_']+v] = X_train[:, i]  
i += 1  
  
X_train = df2.to_numpy()  
X = Normalizer().fit_transform(X_train)
```

```
from sklearn.preprocessing import OneHotEncoder,  
Normalizer  
  
import numpy as np  
import pandas as pd  
  
df2 = df.copy(deep=True)  
  
X_train = df2['Genre'].to_numpy().reshape(-1,1)  
enc = OneHotEncoder()  
  
enc.fit(X_train)  
  
X_train = enc.transform(X_train).toarray()  
values = df2['Genre'].unique()  
nvalues = len(values)  
  
i = 0
```

Esta opción no es correcta.

Falta transformar las columnas finales para que todas tengan los valores numéricos con la misma magnitud.

```
for v in values:  
    df2['Genre_'+v] = X_train[:, i]  
    i += 1  
  
df2.drop('CustomerID', axis=1, inplace=True)  
df2.drop('Genre', axis=1, inplace=True)  
X_train = df2.to_numpy()
```

Misión 2 / Ejercicio 1

Respuesta:

```
results = {}

for k in range(2, nc):
    k_means = cluster.KMeans(n_clusters=k).fit(X)
    y_pred = k_means.predict(X)
    results[k] = metrics.silhouette_score(X, y_pred)
clusters_counts = list(results.keys())
index_values = list(results.values())
plt.plot(clusters_counts, index_values, 'o-')
plt.grid(True)
plt.title('Maximize')
plt.xlabel('Num Clusters')
plt.xticks(clusters_counts)
plt.ylabel('Silhouette')
plt.show()

k = 5 # Debido principalmente a la gráfica K-Elbow
k_means = cluster.KMeans(n_clusters=k)
y_pred = k_means.fit_predict(X)
```

Retroalimentación: Es importante seleccionar el número de grupos adecuados, por lo que hay que utilizar alguna estrategia para ello.

En particular, este código plotea la gráfica K-Elbow donde se puede ver que 5 es un número de grupos adecuados, por lo que se deben agrupar los datos en 5 grupos.

Veamos por qué las otras opciones no son correctas.

Opción	Retroalimentación
<pre>results = {} for k in range(2, nc): k_means = cluster.KMeans(n_clusters=k).fit(X) y_pred = k_means.predict(X) results[k] = y_pred clusters_counts = list(results.keys()) index_values = list(results.values()) plt.plot(clusters_counts, index_values, 'o-') plt.grid(True) plt.title('Maximize') plt.xlabel('Num Clusters') plt.xticks(clusters_counts)</pre>	<p>Esta opción no es correcta.</p> <p>La variable results es un diccionario donde las llaves son el número de grupos en los que se pueden segmentar los datos y los valores, alguna medición de la calidad de esos agrupamientos, por ejemplo, valores del índice Silhouette.</p>

```
plt.ylabel('Silhouette')
plt.show()

k = 5 # Debido principalmente a la gráfica K-Elbow
k_means = cluster.KMeans(n_clusters=k)
y_pred = k_means.fit_predict(X)
```

```
results = {}

for k in range(2, 11):
    k_means = cluster.KMeans(n_clusters=k).fit(X)
    y_pred = k_means.predict(X)
    results[k] = metrics.silhouette_score(X, y_pred)

clusters_counts = list(results.keys())
index_values = list(results.values())

plt.plot(clusters_counts, index_values, 'o-')
plt.grid(True)
plt.title('"Maximize"')
plt.xlabel('Num Clusters')
plt.xticks(clusters_counts)
plt.ylabel('Silhouette')
plt.show()

k = 5 # Debido principalmente a la gráfica K-Elbow
k_means = cluster.KMeans(n_clusters=k)
```

Esta opción no es correcta

Según la idea del código, las etiquetas que se van almacenando en la variable `y_pred` se actualizan de una iteración a otra, por lo que es necesario volver a agrupar al final con el número de grupos seleccionado.

Misión 3 / Ejercicio 1

Respuesta:

```
df2['groups'] = y_pred1
clients = [df2[df2['groups']==g] for g in range(k)]
grouped = pd.DataFrame()
grouped['Features'] = df2.columns[:-1]
for g in range(k):
    row = []
    for col in grouped['Features']:
        if clients[g][col].dtype != 'object':
            row.append(np.round(clients[g][col].mean(),2))
        else:
            row.append(clients[g][col].value_counts().keys()[0])
    grouped['Group'+str(g)] = row
```

Retroalimentación: Primero, crea una lista de *dataframes* donde estén los clientes divididos por grupos.

Luego, crea un nuevo *dataframe* donde aparece la descripción de las variables por grupo, en particular el promedio de cada variable.

Pero, ¿por qué las otras opciones no son correctas?

Opción	Retroalimentación
<pre>df2['groups'] = y_pred1 clients = [df2[df2['groups']==g] for g in range(k)] grouped = pd.DataFrame() grouped['Features'] = df2.columns[:-1] for g in range(k): row = [] for col in grouped['Features']: if clients[g][col].dtype != 'object': row.append(np.round(clients[g][col].mean(), 2)) else: row.append(clients[g][col].value_counts().keys()[0]) grouped['Group'+str(g)] = row</pre>	<p>Esta opción no es correcta.</p> <p>Aunque creas un nuevo dataframe con el promedio de variables por grupo, no creas una lista de dataframes con los datos segmentados</p>

```
df2['groups'] = y_pred1  
  
clients = [df2[df2['groups']==g] for g in range(k)]  
  
grouped = pd.DataFrame()  
  
grouped['Features'] = df2.columns[:-1]  
  
for g in range(k):  
  
    row = []  
  
    for col in grouped['Features']:  
  
        row.append(clients[g][col].value_counts().keys()  
[0])  
  
    grouped['Group'+str(g)] = row
```

Esta opción no es correcta

Aunque segmentan los datos y describen las variables por grupo, descuidas que para las variables no numéricas no es posible calcular el promedio, por lo que es necesario otro estadístico para ello.

Misión 4 / Ejercicio 1

Respuesta:

En el grupo 1 (de color naranja) vemos que las personas tienen altos ingresos y puntajes de gasto, este es el caso ideal para el centro comercial o las tiendas, ya que estas personas son las principales fuentes de ganancias. Estas personas pueden ser clientes habituales del centro comercial y están convencidas de las instalaciones del centro comercial.

En el grupo 2 (de color amarillo) podemos ver que las personas tienen bajos ingresos anuales y puntajes de gasto bajos, esto es bastante razonable ya que las personas con bajos salarios prefieren comprar menos, de hecho, estas son las personas sabias que saben cómo gastar y ahorrar dinero. Las tiendas / centros comerciales estarán menos interesados en las personas que pertenecen a este grupo.

En el grupo 3 (de color verde) podemos ver que las personas tienen bajos ingresos, pero puntajes de gasto más altos, estas son aquellas personas que por alguna razón aman comprar productos con más frecuencia a pesar de que tienen bajos ingresos. Tal vez sea porque estas personas están más que satisfechas con los servicios del centro comercial. Es posible que las tiendas / centros comerciales no se dirijan a estas personas de manera efectiva, pero aun así no las perderán.

En el grupo 4 (de color azul) vemos que las personas tienen ingresos promedio y una puntuación de gasto promedio, estas personas nuevamente no serán los principales objetivos de las tiendas o centros comerciales, pero nuevamente serán consideradas y se pueden usar otras técnicas de análisis de datos para aumentar su puntaje de gasto.

En el grupo 5 (de color rojo) vemos que las personas tienen altos ingresos, pero puntajes de gasto bajos, esto es interesante. Quizás estas son las personas que están insatisfechas o descontentas con los servicios del centro comercial. Estos pueden ser los principales objetivos del centro comercial, ya que tienen el potencial de gastar dinero. Entonces, las autoridades del centro comercial intentarán agregar nuevas instalaciones para que puedan atraer a estas personas y puedan satisfacer sus necesidades.

Retroalimentación: Con base en nuestra técnica de aprendizaje automático, podemos deducir que, para aumentar las ganancias del centro comercial, las autoridades del centro comercial deben apuntar a las personas que pertenecen al clúster 4 y al clúster 5 para lograr que aumenten sus puntajes de gasto y también deben mantener sus estándares para mantener a las personas que pertenecen al clúster 1 y al clúster 3 felices y satisfechos.