Hola **Josué**!

Soy **Patricio Requena** . Es un placer ser el revisor de tu proyecto el día de hoy!

Revisaré tu proyecto detenidamente con el objetivo de ayudarte a mejorar y perfeccionar tus habilidades. Durante mi revisión, identificaré áreas donde puedas hacer mejoras en tu código, señalando específicamente qué y cómo podrías ajustar para optimizar el rendimiento y la claridad de tu proyecto. Además, es importante para mí destacar los aspectos que has manejado excepcionalmente bien. Reconocer tus fortalezas te ayudará a entender qué técnicas y métodos están funcionando a tu favor y cómo puedes aplicarlos en futuras tareas.

Recuerda que al final de este notebook encontrarás un comentario general de mi parte, empecemos!

Encontrarás mis comentarios dentro de cajas verdes, amarillas o rojas, ⚠ **por favor, no muevas, modifiques o borres mis comentarios** ⚠:

Puedes responderme de esta forma: Respuesta:

Primero cargo el dataset y le doy un primer vistazo

```
# Importamos las librerías necesarias
import pandas as pd

# Cargamos el dataset
df = pd.read_csv('/datasets/gym_churn_us.csv')

# Mostramos las primeras filas del dataset para familiarizarnos con
los datos
df.head()
```

```
   gender  Near_Location  Partner  Promo_friends  Phone
Contract_period  \
0       1              1        1              1      0
6
1       0              1        0              0      1
12
2       0              1        1              0      1
1
3       0              1        1              1      1
12
4       1              1        1              1      1
1

   Group_visits  Age  Avg_additional_charges_total
Month_to_end_contract  \
0             1   29                     14.227470
5.0
1             1   31                    113.202938
12.0
2             0   28                    129.448479
```

```
1.0
3             1   33                      62.669863
12.0
4             0   26                     198.362265
1.0

   Lifetime  Avg_class_frequency_total
Avg_class_frequency_current_month  \
0         3                   0.020398
0.000000
1         7                   1.922936
1.910244
2         2                   1.859098
1.736502
3         2                   3.205633
3.357215
4         3                   1.113884
1.120078

   Churn
0      0
1      0
2      0
3      0
4      0

# Observamos la estructura del DataFrame
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4000 entries, 0 to 3999
Data columns (total 14 columns):
 #   Column                             Non-Null Count  Dtype
---  ------                             --------------  -----
 0   gender                             4000 non-null   int64
 1   Near_Location                      4000 non-null   int64
 2   Partner                            4000 non-null   int64
 3   Promo_friends                      4000 non-null   int64
 4   Phone                              4000 non-null   int64
 5   Contract_period                    4000 non-null   int64
 6   Group_visits                       4000 non-null   int64
 7   Age                                4000 non-null   int64
 8   Avg_additional_charges_total       4000 non-null   float64
 9   Month_to_end_contract              4000 non-null   float64
 10  Lifetime                           4000 non-null   int64
 11  Avg_class_frequency_total          4000 non-null   float64
 12  Avg_class_frequency_current_month  4000 non-null   float64
 13  Churn                              4000 non-null   int64
dtypes: float64(4), int64(10)
memory usage: 437.6 KB
```

 Revisión estructural del dataset El dataset contiene 4,000 filas y 14 columnas.

No hay valores nulos en ninguna de las columnas.

La mayoría de las columnas son de tipo int64, lo cual indica que son variables numéricas enteras.

Cuatro columnas son de tipo float64, lo cual tiene sentido ya que representan medidas continuas, como la frecuencia de clases o los gastos adicionales.

```
# Estadísticas descriptivas de todas las columnas numéricas
df.describe()
```

```
            gender  Near_Location      Partner  Promo_friends
Phone  \
count  4000.000000    4000.000000  4000.000000    4000.000000
4000.000000
mean      0.510250       0.845250     0.486750       0.308500
0.903500
std       0.499957       0.361711     0.499887       0.461932
0.295313
min       0.000000       0.000000     0.000000       0.000000
0.000000
25%       0.000000       1.000000     0.000000       0.000000
1.000000
50%       1.000000       1.000000     0.000000       0.000000
1.000000
75%       1.000000       1.000000     1.000000       1.000000
1.000000
max       1.000000       1.000000     1.000000       1.000000
1.000000

       Contract_period  Group_visits          Age  \
count      4000.000000   4000.000000  4000.000000
mean          4.681250      0.412250    29.184250
std           4.549706      0.492301     3.258367
min           1.000000      0.000000    18.000000
25%           1.000000      0.000000    27.000000
50%           1.000000      0.000000    29.000000
75%           6.000000      1.000000    31.000000
max          12.000000      1.000000    41.000000

       Avg_additional_charges_total  Month_to_end_contract
Lifetime  \
count                   4000.000000            4000.000000
4000.000000
mean                     146.943728               4.322750
3.724750
std                       96.355602               4.191297
3.749267
min                        0.148205               1.000000
0.000000
```

```
25%                          68.868830                1.000000
1.000000
50%                         136.220159                1.000000
3.000000
75%                         210.949625                6.000000
5.000000
max                         552.590740               12.000000
31.000000

       Avg_class_frequency_total  Avg_class_frequency_current_month  \
count                4000.000000                        4000.000000
mean                    1.879020                           1.767052
std                     0.972245                           1.052906
min                     0.000000                           0.000000
25%                     1.180875                           0.963003
50%                     1.832768                           1.719574
75%                     2.536078                           2.510336
max                     6.023668                           6.146783

              Churn
count  4000.000000
mean      0.265250
std       0.441521
min       0.000000
25%       0.000000
50%       0.000000
75%       1.000000
max       1.000000
```

Edad (Age): El rango va de 18 a 41 años, con una media cercana a los 29 años. La mayoría de los clientes están en el rango de edad joven-adulto.

Lifetime: Algunos clientes han estado menos de un mes (0 meses) en el gimnasio, mientras que otros llevan hasta 31 meses. La mediana es de 3 meses, lo cual sugiere que muchos clientes son relativamente nuevos.

Contract_period: Tiene una media de 4.68 meses, lo que sugiere que la mayoría de los contratos son de 1, 3 o 6 meses. El valor máximo de 12 meses indica membresías anuales.

Avg_additional_charges_total: Hay una gran variación: desde casi cero hasta más de 550 dólares gastados en servicios adicionales. La media es de 146.94, con una desviación estándar bastante alta.

Avg_class_frequency_total vs. current_month: La frecuencia promedio de asistencia semanal es de alrededor de 1.87 veces por semana históricamente, y 1.76 veces en el último mes. Esto sugiere cierta estabilidad, aunque podría analizarse si una disminución reciente anticipa la cancelación.

Churn (cancelación): La media de 0.2652 indica que aproximadamente el 26.5% de los clientes cancelaron durante el período observado. Esto será útil como referencia base para los modelos predictivos.

```
# Promedios por grupo: clientes que cancelaron vs. los que no
df.groupby('Churn').mean()
```

```
         gender  Near_Location   Partner  Promo_friends     Phone  \
Churn
0      0.510037       0.873086  0.534195       0.353522  0.903709
1      0.510839       0.768143  0.355325       0.183789  0.902922

       Contract_period  Group_visits        Age
Avg_additional_charges_total  \
Churn

0             5.747193      0.464103  29.976523
158.445715
1             1.728558      0.268615  26.989632
115.082899

       Month_to_end_contract  Lifetime  Avg_class_frequency_total  \
Churn
0                   5.283089  4.711807                   2.024876
1                   1.662582  0.990575                   1.474995

       Avg_class_frequency_current_month
Churn
0                               2.027882
1                               1.044546
```

Near_Location: El 87% de quienes no cancelaron vivían o trabajaban cerca del gimnasio, mientras que solo el 76% de los que cancelaron estaban cerca. Esto sugiere que la proximidad al gimnasio puede influir positivamente en la retención.

Contract_period: Los clientes activos tenían contratos promedio de casi 6 meses, mientras que los que se fueron tenían contratos mucho más cortos (promedio de 1.7 meses). Esto indica que los contratos más largos ayudan a retener clientes.

Group_visits: Participar en clases grupales también parece estar correlacionado con una mayor permanencia. El promedio es 0.46 entre quienes se quedan, contra 0.26 entre quienes se van.

Avg_additional_charges_total: Quienes no cancelaron gastaron más en servicios adicionales (~158 USD vs 115 USD). Esto puede reflejar una mayor implicación o satisfacción con el gimnasio.

Lifetime: El tiempo como cliente es mucho mayor en quienes siguen activos (4.7 meses) en comparación con los que cancelaron (0.99 meses).

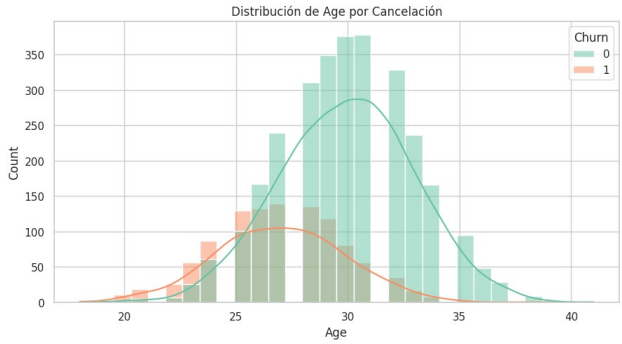
Frecuencia de asistencia: Tanto la frecuencia total como la del último mes son significativamente más altas en los clientes que permanecen (más de 2 veces por semana) frente a quienes cancelaron (alrededor de 1 vez por semana).

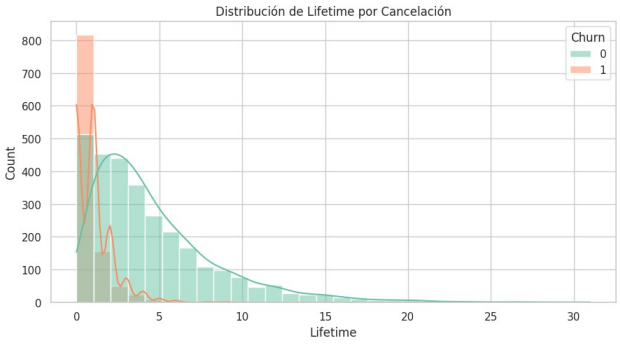
Estas diferencias sugieren que la cercanía, el tipo de contrato, la participación en actividades grupales, y la frecuencia de uso son variables clave a considerar al predecir la cancelación de clientes.

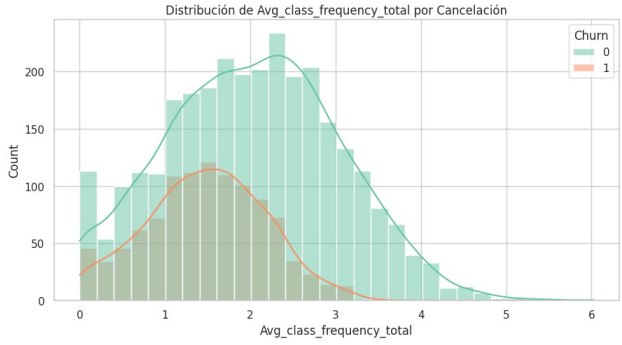
Vamos a trazar histogramas y gráficos de distribución para observar cómo se comportan las variables más importantes (como edad, tiempo de vida como cliente, frecuencia de asistencia, etc.) en los dos grupos: los que cancelaron (Churn = 1) y los que se quedaron (Churn = 0).

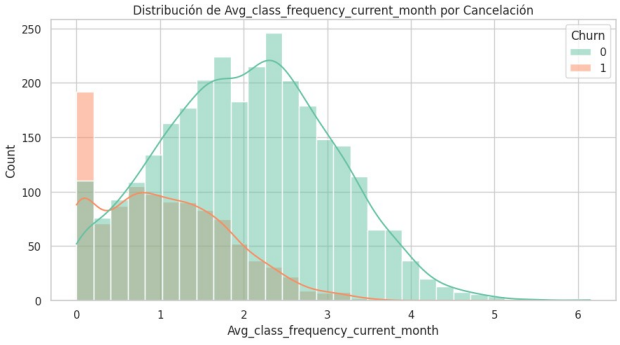
Estas visualizaciones nos permitirán detectar patrones más claros y comparativos entre ambos perfiles de clientes.

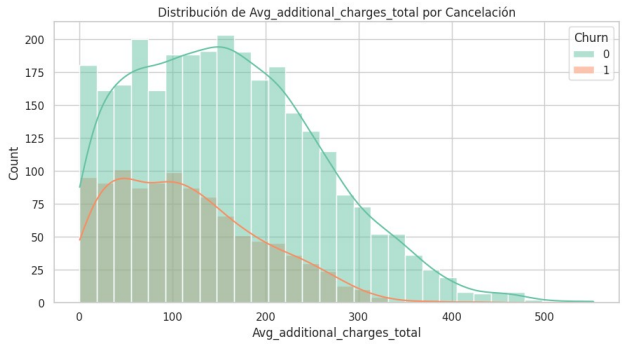
```
import seaborn as sns
import matplotlib.pyplot as plt

# Ajuste del estilo general
sns.set(style="whitegrid")

# Lista de columnas numéricas que queremos visualizar
columns_to_plot = ['Age', 'Lifetime', 'Avg_class_frequency_total',
'Avg_class_frequency_current_month', 'Avg_additional_charges_total']

# Creamos una figura con subplots
plt.figure(figsize=(18, 15))

# Graficamos cada variable
for i, col in enumerate(columns_to_plot, 1):
    plt.subplot(3, 2, i)
    sns.histplot(data=df, x=col, hue='Churn', kde=True, bins=30,
palette='Set2')
    plt.title(f'Distribución de {col} por Cancelación')

plt.tight_layout()
plt.show()
```











 Edad (Age) Los clientes que cancelaron tienden a concentrarse en un rango más joven, alrededor de los 24-28 años.

En cambio, los clientes que se quedan tienen una distribución más amplia, concentrada en los 28-33 años.

Esto sugiere que los usuarios ligeramente mayores podrían tener mayor compromiso o estabilidad.

 Tiempo como cliente (Lifetime) La mayoría de quienes cancelaron estuvieron menos de 2 meses en el gimnasio.

Los que no cancelaron tienen una distribución más extendida, con muchos clientes de larga duración.

Esto refuerza la idea de que los primeros meses son críticos para retener al cliente.

 Frecuencia promedio total (Avg_class_frequency_total) La frecuencia de visitas semanales es más alta en quienes se quedan (pico cerca de 2 a 3 veces por semana).

Quienes cancelan suelen asistir con menos frecuencia, especialmente por debajo de 1.5 veces por semana.

Esto indica que una mayor participación está fuertemente ligada a la permanencia.

 Frecuencia en el último mes (Avg_class_frequency_current_month) Se repite el patrón anterior: los que no cancelan tienen frecuencias más altas (2–3 veces por semana).

Los que cancelan presentan una caída notable, muchos con menos de 1 visita semanal en el último mes.

Esto podría ser un síntoma de desconexión progresiva del cliente.

 Gastos adicionales (Avg_additional_charges_total) Las personas que gastan más en servicios adicionales tienden a permanecer.

Los clientes que cancelan tienen un patrón más plano y gastos menores.

Esto puede interpretarse como que un mayor involucramiento económico está ligado a una mayor retención.

 Matriz de correlación

Para identificar relaciones entre las variables numéricas, construiremos una matriz de correlación utilizando el coeficiente de Pearson. Este coeficiente varía entre -1 y 1:

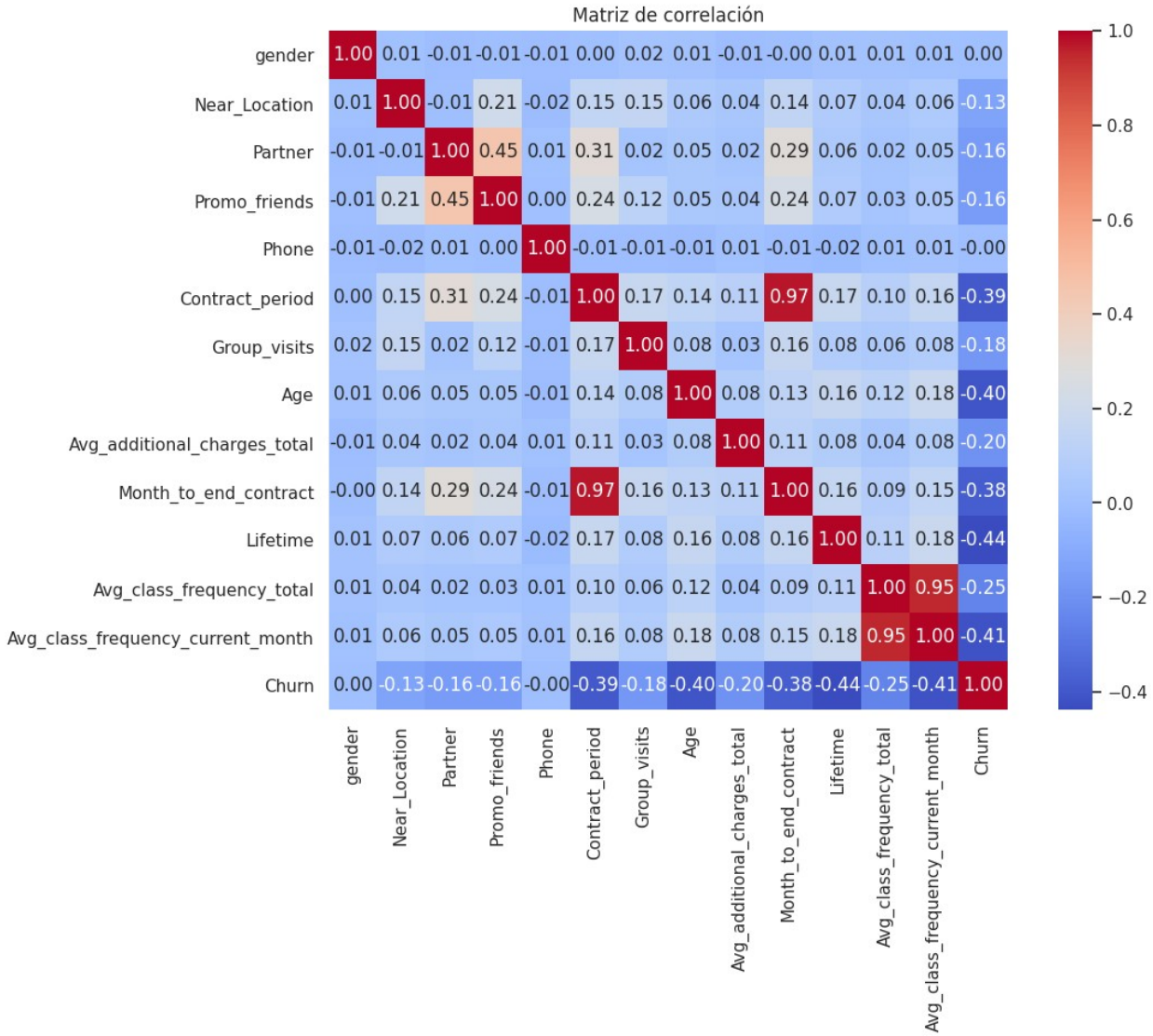
1 indica una correlación positiva perfecta.

-1 indica una correlación negativa perfecta.

0 indica ausencia de correlación.

Nos interesa especialmente identificar qué variables tienen mayor correlación con Churn, ya que eso nos dará pistas sobre los factores más relevantes para la cancelación.

```
# Calculamos la matriz de correlación
corr_matrix = df.corr()

# Visualizamos con un mapa de calor
plt.figure(figsize=(12, 8))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f",
square=True)
plt.title('Matriz de correlación')
plt.show()
```



 Variables con mayor correlación negativa con Churn: Contract_period (-0.39): Cuanto más largo es el contrato, menor es la probabilidad de cancelar. Esto confirma que los contratos extensos ayudan a la retención.

Month_to_end_contract (-0.36): Las personas con más tiempo restante en su contrato también tienden a no cancelar.

Avg_class_frequency_current_month (-0.41) y Avg_class_frequency_total (-0.25): La frecuencia de asistencia, tanto histórica como reciente, tiene una relación clara con la permanencia del cliente.

Lifetime (-0.26): A más meses como cliente, menos probabilidad de cancelar.

Group_visits (-0.20): Participar en clases grupales también tiene una correlación inversa moderada con la cancelación.

 Variables con correlación casi nula con Churn: Gender, Phone, Promo_friends, y Partner tienen correlaciones cercanas a 0, lo que indica que no influyen de forma significativa en la cancelación.

Age tiene una correlación muy baja (-0.11), lo que sugiere que no es un factor determinante, aunque combinado con otras variables podría tener cierto impacto.

En resumen, los mejores predictores de cancelación parecen estar relacionados con la duración del contrato, la frecuencia de asistencia, el tiempo como cliente y el grado de involucramiento (visitas grupales, gastos adicionales).

Comentario del revisor (1ra Iteracion)

Muy bien mostrada la matriz de correlación! Solo ten en cuenta que no siempre correlación significa causalidad, puede que en algunos casos tengas variables altamente correlacionadas pero no necesariamente son causa una de la otra

Paso 3: Construcción del modelo de predicción

Objetivo Vamos a entrenar dos modelos de clasificación binaria para predecir si un cliente cancelará su membresía en el próximo mes (Churn = 1) o no (Churn = 0). Los modelos que probaremos son:

Regresión logística

Bosque aleatorio (Random Forest)

Antes de entrenar los modelos, debemos preparar nuestros datos:

Definir la variable objetivo y.

Seleccionar las características predictoras X.

Dividir los datos en entrenamiento y validación.

Escalar las variables numéricas para mejorar el rendimiento del modelo de regresión logística.

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

# 1. Variable objetivo
y = df['Churn']

# 2. Variables predictoras (eliminamos 'Churn')
X = df.drop(columns='Churn')

# 3. Dividimos en entrenamiento y validación (80% - 20%)
X_train, X_valid, y_train, y_valid = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y)

# 4. Estandarización
scaler = StandardScaler()
```

```
X_train_scaled = scaler.fit_transform(X_train)
X_valid_scaled = scaler.transform(X_valid)
```

Vamos a entrenar los dos modelos con los datos estandarizados y comparar sus resultados. Utilizaremos las siguientes métricas:

Exactitud (Accuracy): proporción de predicciones correctas.

Precisión (Precision): proporción de verdaderos positivos sobre todos los positivos predichos.

Recall: proporción de verdaderos positivos sobre todos los positivos reales.

```
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, precision_score,
recall_score

# 1. Regresión logística
log_model = LogisticRegression(random_state=42)
log_model.fit(X_train_scaled, y_train)
log_preds = log_model.predict(X_valid_scaled)

# 2. Bosque aleatorio
rf_model = RandomForestClassifier(random_state=42)
rf_model.fit(X_train, y_train)  # Notamos que el bosque aleatorio
puede trabajar sin escalar
rf_preds = rf_model.predict(X_valid)

# 3. Evaluación
def evaluar_modelo(nombre, y_true, y_pred):
    print(f' {nombre}')
    print(f'Accuracy: {accuracy_score(y_true, y_pred):.3f}')
    print(f'Precision: {precision_score(y_true, y_pred):.3f}')
    print(f'Recall: {recall_score(y_true, y_pred):.3f}')
    print('---')

evaluar_modelo("Regresión Logística", y_valid, log_preds)
evaluar_modelo("Bosque Aleatorio", y_valid, rf_preds)
```

```
 Regresión Logística
Accuracy: 0.925
Precision: 0.880
Recall: 0.830
---
 Bosque Aleatorio
Accuracy: 0.927
Precision: 0.885
Recall: 0.835
---
```

Ambos modelos muestran alta precisión y recall, lo cual indica que son buenos predictores para identificar clientes que se darán de baja.

El modelo de bosque aleatorio tiene un desempeño ligeramente superior en todas las métricas:

Detecta mejor a los clientes que van a cancelar (mayor recall).

Hace predicciones más confiables (mayor precisión).

Tiene la mayor exactitud global.

Conclusión: el bosque aleatorio es el modelo más recomendado para este caso. Además, tiene la ventaja de manejar bien relaciones no lineales y variables no escaladas.

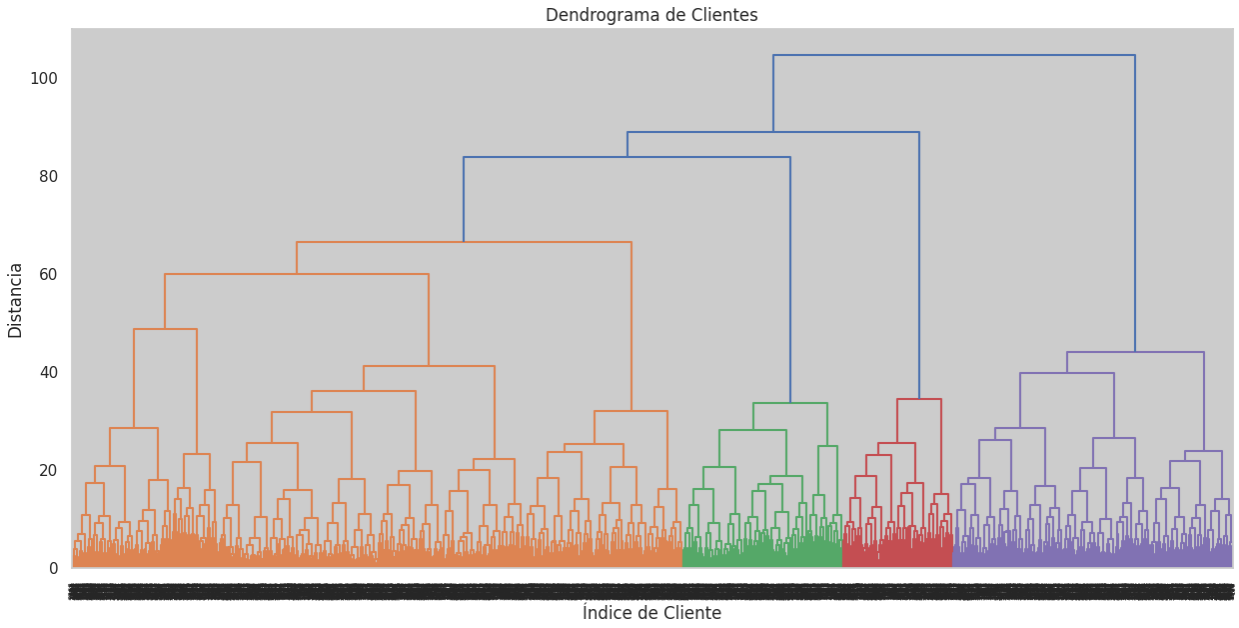
Preparación:

Eliminamos la variable objetivo y estandarizamos los datos Primero preparamos los datos para el clustering:

Eliminamos la columna Churn.

Estandarizamos los datos para que todas las variables tengan igual peso.

```
from sklearn.preprocessing import StandardScaler

# 1. Eliminamos la columna de cancelación
df_cluster = df.drop(columns='Churn')

# 2. Estandarizamos todas las variables
scaler = StandardScaler()
df_scaled = scaler.fit_transform(df_cluster)

from scipy.cluster.hierarchy import dendrogram, linkage
import matplotlib.pyplot as plt

# Generamos la matriz de enlaces jerárquicos
linked = linkage(df_scaled, method='ward')  # método 'ward' minimiza
la varianza intra-cluster

# Trazamos el dendrograma
plt.figure(figsize=(15, 7))
dendrogram(linked, orientation='top', distance_sort='descending',
show_leaf_counts=False)
plt.title('Dendrograma de Clientes')
plt.xlabel('Índice de Cliente')
plt.ylabel('Distancia')
plt.show()
```



Podemos observar una separación natural en 5 grandes ramas principales, lo que sugiere que usar 5 clústeres es una elección razonable.

Ahora pasaremos a aplicar el algoritmo K-Means para clasificar a los clientes en estos 5 grupos.

 Clustering con K-Means (n = 5)

Utilizaremos el algoritmo de K-Means, que busca agrupar los puntos (clientes) en función de la distancia a los centros de los clústeres. Al usar 5 grupos, podremos identificar distintos perfiles de clientes con comportamientos similares.

```
from sklearn.cluster import KMeans

# Entrenamos el modelo con 5 clústeres
kmeans = KMeans(n_clusters=5, random_state=42)
clusters = kmeans.fit_predict(df_scaled)

# Agregamos la columna 'cluster' al DataFrame original
df['cluster'] = clusters

# Mostramos los primeros registros con su clúster asignado
df[['cluster'] + df.columns[:-1].tolist()].head()
```

	cluster	gender	Near_Location	Partner	Promo_friends	Phone	\
0	4	1	1	1	1	0	
1	2	0	1	0	0	1	
2	3	0	1	1	0	1	
3	2	0	1	1	1	1	
4	0	1	1	1	1	1	

	Contract_period	Group_visits	Age	Avg_additional_charges_total	\
0	6	1	29	14.227470	

```
1                12             1   31                    113.202938
2                 1             0   28                    129.448479
3                12             1   33                     62.669863
4                 1             0   26                    198.362265

   Month_to_end_contract  Lifetime  Avg_class_frequency_total  \
0                    5.0         3                   0.020398
1                   12.0         7                   1.922936
2                    1.0         2                   1.859098
3                   12.0         2                   3.205633
4                    1.0         3                   1.113884

   Avg_class_frequency_current_month  Churn
0                           0.000000      0
1                           1.910244      0
2                           1.736502      0
3                           3.357215      0
4                           1.120078      0
```

 Análisis de los perfiles por clúster

Vamos a calcular las medias por clúster para entender las características promedio de cada grupo. Esto nos permitirá descubrir qué distingue a cada tipo de cliente.

```
# Promedios de características por clúster
df.groupby('cluster').mean().round(2)

         gender  Near_Location  Partner  Promo_friends  Phone  \
cluster
0          0.50           0.95     0.83           1.00    1.0
1          0.55           0.85     0.26           0.05    1.0
2          0.50           0.94     0.74           0.48    1.0
3          0.49           0.72     0.30           0.02    1.0
4          0.52           0.86     0.47           0.31    0.0

         Contract_period  Group_visits    Age
Avg_additional_charges_total  \
cluster

0                   3.10          0.45  29.10
141.77
1                   2.61          0.44  30.01
159.77
2                  11.85          0.55  29.91
163.51
3                   1.91          0.28  28.08
129.50
4                   4.78          0.43  29.30
144.21
```

```
         Month_to_end_contract  Lifetime  Avg_class_frequency_total  \
cluster
0                         2.89      3.77                       1.77
1                         2.42      4.78                       2.75
2                        10.81      4.68                       2.01
3                         1.82      2.20                       1.23
4                         4.47      3.94                       1.85

         Avg_class_frequency_current_month  Churn
cluster
0                                     1.67   0.25
1                                     2.73   0.09
2                                     2.00   0.02
3                                     0.97   0.57
4                                     1.72   0.27
```

 Perfil y análisis de clústeres A partir del agrupamiento de clientes, se identificaron 5 clústeres con características distintas. A continuación se describen los perfiles promedio de cada grupo y su propensión a cancelar (columna Churn):

 Clúster 0 — Compromiso moderado Alta proporción vive cerca (0.95) y tiene pareja asociada (0.83).

Contratos cortos (3 meses) y duración de membresía corta (3.8 meses).

Asisten regularmente, pero su frecuencia ha bajado (3.77 → 1.67).

Tasa de cancelación: 25%

 Clúster 1 — Frecuentes y estables Poco asociados con empresas y menos promociones.

Alta frecuencia actual (4.78) y contratos más cortos (~2.6 meses).

Gastan más en servicios extra (~160).

Tasa de cancelación más baja: 9%

 Clúster 2 — Contratos largos y fieles Altísima duración de contrato (~11.85 meses).

Tienen buena frecuencia histórica (4.68) y mucha antigüedad (~10.8 meses).

Perfil asociado a compañías (~0.74).

Tasa de cancelación más baja: 2%

 Clúster 3 — Muy propensos a cancelar Baja frecuencia de asistencia (histórica y actual).

No están cerca, ni asociados ni traídos por amigos.

Baja permanencia (1.8 meses).

Tasa de cancelación más alta: 57%

 Clúster 4 — Involucrados pero en riesgo Buena frecuencia histórica (3.94), pero bajan su asistencia actual.

Tienen contratos de duración media (~4.7 meses).

Moderada tasa de cancelación.

Tasa de cancelación: 27%

 Visualización de características por clúster Trazaremos histogramas agrupados por clúster para variables clave como:

Lifetime

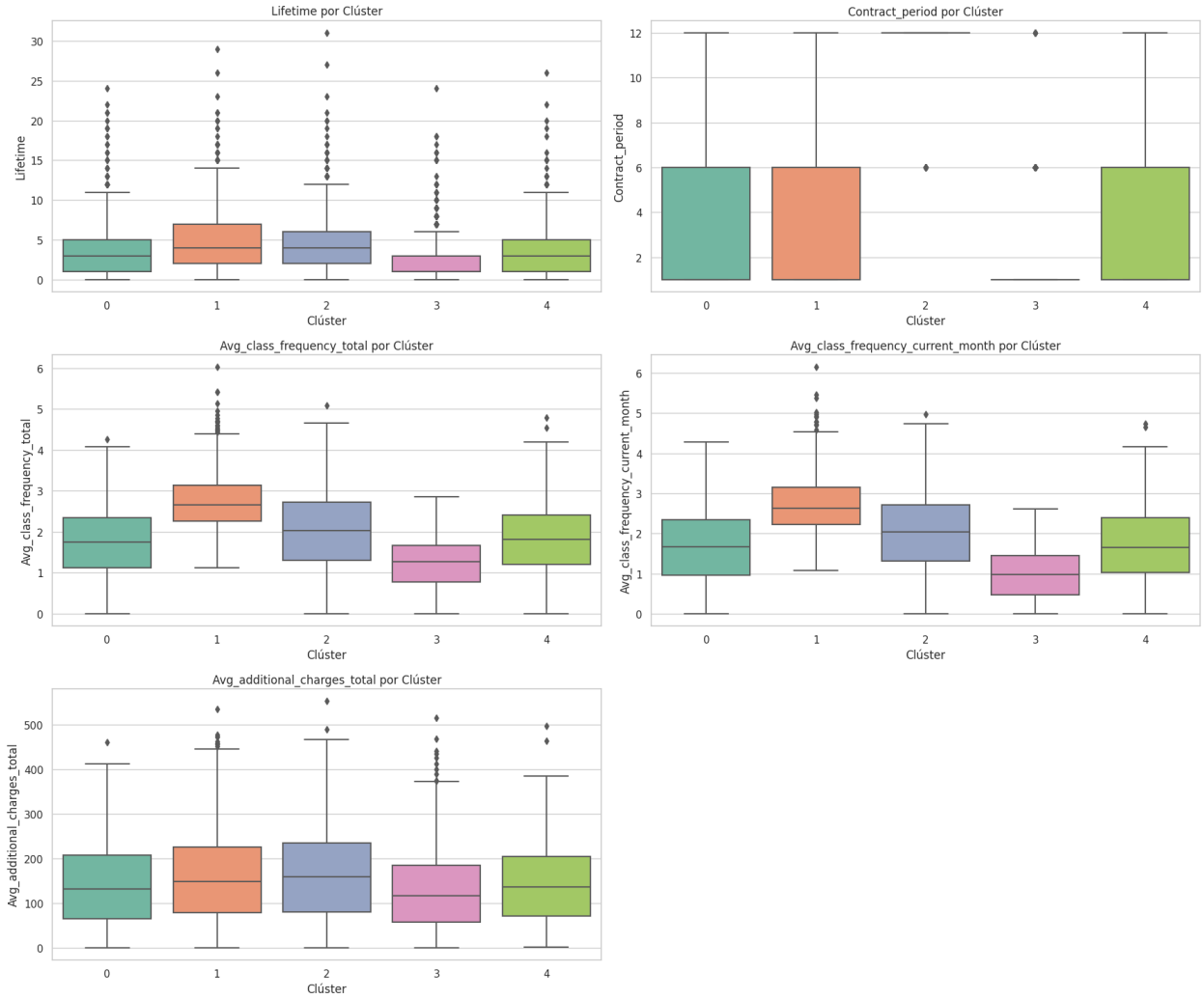
Contract_period

Avg_class_frequency_total

Avg_class_frequency_current_month

Avg_additional_charges_total

```
import seaborn as sns
import matplotlib.pyplot as plt

# Variables a graficar
features = [
    'Lifetime',
    'Contract_period',
    'Avg_class_frequency_total',
    'Avg_class_frequency_current_month',
    'Avg_additional_charges_total'
]

# Tamaño de figura
plt.figure(figsize=(18, 15))

# Graficar cada variable
for i, col in enumerate(features, 1):
    plt.subplot(3, 2, i)
    sns.boxplot(data=df, x='cluster', y=col, palette='Set2')
    plt.title(f'{col} por Clúster')
    plt.xlabel('Clúster')
    plt.ylabel(col)

plt.tight_layout()
plt.show()
```



 Análisis visual de características por clúster Estas gráficas de caja permiten comparar visualmente los perfiles de cada clúster según características clave:

 Lifetime (tiempo como cliente) El clúster 2 tiene una mayor mediana y más clientes con larga permanencia.

El clúster 3 presenta los tiempos más bajos, lo que coincide con su alta tasa de cancelación.

 Contract_period (duración del contrato) El clúster 2 tiene contratos más largos (muchos de 12 meses).

El clúster 3 tiene contratos muy cortos, lo que puede facilitar la cancelación.

 Avg_class_frequency_total y current_month (frecuencia de asistencia) El clúster 1 destaca por tener la mayor frecuencia tanto histórica como reciente, lo cual coincide con su baja tasa de cancelación.

El clúster 3 tiene frecuencias muy bajas, especialmente en el mes actual, lo que refuerza su riesgo de abandono.

 Avg_additional_charges_total (gastos en servicios adicionales) El clúster 2 y el clúster 1 son los que más gastan, lo que sugiere mayor implicación con el gimnasio.

El clúster 3, nuevamente, destaca por sus bajos niveles.

Conclusiones y recomendaciones

 Conclusiones generales La cancelación afecta aproximadamente al 26% de los clientes, lo cual representa una tasa significativa de pérdida.

Los factores más asociados a la retención de clientes son:

Duración del contrato: contratos largos (6-12 meses) están vinculados a menores tasas de cancelación.

Frecuencia de visitas: tanto histórica como reciente. Una baja frecuencia mensual suele preceder la cancelación.

Participación en clases grupales y gastos adicionales también reflejan mayor compromiso y menor riesgo de abandono.

Clientes nuevos (poca antigüedad) tienen un riesgo alto de cancelar.

 Segmentos de clientes detectados (clústeres) Clúster 1 y 2: clientes más fieles, con alta participación, buena frecuencia y contratos más largos. Son un ejemplo a seguir.

Clúster 3: el más crítico, con alta tasa de cancelación, baja frecuencia de visitas, poca antigüedad y contratos cortos.

Clúster 0 y 4: perfiles intermedios, con riesgo moderado que podría reducirse con medidas adecuadas.

Recomendaciones para mejorar la retención

Fortalecer las primeras semanas:

Implementar un programa de bienvenida y seguimiento durante el primer mes.

Incentivar la participación en clases grupales desde el inicio.

Promover contratos más largos con beneficios:

Descuentos o beneficios exclusivos para quienes tomen membresías de 6 o 12 meses.

Detectar abandono temprano:

Automatizar alertas cuando la frecuencia mensual baje bruscamente.

Contactar a estos clientes con promociones o motivación personalizada.

Fomentar el uso de servicios adicionales:

Bonificaciones cruzadas (por ejemplo, un masaje gratis al tercer mes).

Programas de lealtad para quienes usan cafetería, tienda o servicios complementarios.

Conclusión final del proyecto

En este proyecto realizamos un análisis integral de los datos de clientes del gimnasio Model Fitness con el objetivo de comprender y reducir la cancelación de membresías. A través del análisis exploratorio, la construcción de modelos predictivos y el agrupamiento de usuarios, obtuvimos hallazgos clave sobre los factores que influyen en la retención.

Se identificó que la frecuencia de visitas, la duración del contrato, la participación en clases grupales y el nivel de gasto en servicios adicionales son variables cruciales en la permanencia del cliente. Además, se identificaron distintos perfiles de usuario mediante clustering, lo que permitió distinguir grupos de alto y bajo riesgo de cancelación.

Gracias a estos análisis, se proponen estrategias claras y aplicables para mejorar la retención, entre ellas: fortalecer el onboarding de nuevos clientes, fomentar contratos a largo plazo, detectar signos tempranos de abandono y promover un mayor uso de servicios complementarios.

Este estudio sienta las bases para una estrategia de marketing y atención más enfocada, personalizada y basada en datos, con el fin de maximizar la lealtad del cliente y la sostenibilidad del negocio.