TAREA 4 – ALGORITMOS DE APRENDIZAJE NO SUPERVISADO

David Esteban Lasso Ordóñez

Universidad Nacional Abierta y a Distancia
Análisis de Datos (202016908_42)
Breyner Alexander Parra
2024

TABLA DE CONTENIDO

Anotación: El presente documento únicamente contiene los valores que no son abarcados en el trabajo colaborativo.

Gracias por su comprensión.

1 CUADRO SINÓPTICO

A continuación, se presenta el cuadro elaborado en razón a la identificación de modelos de aprendizaje no supervisado.

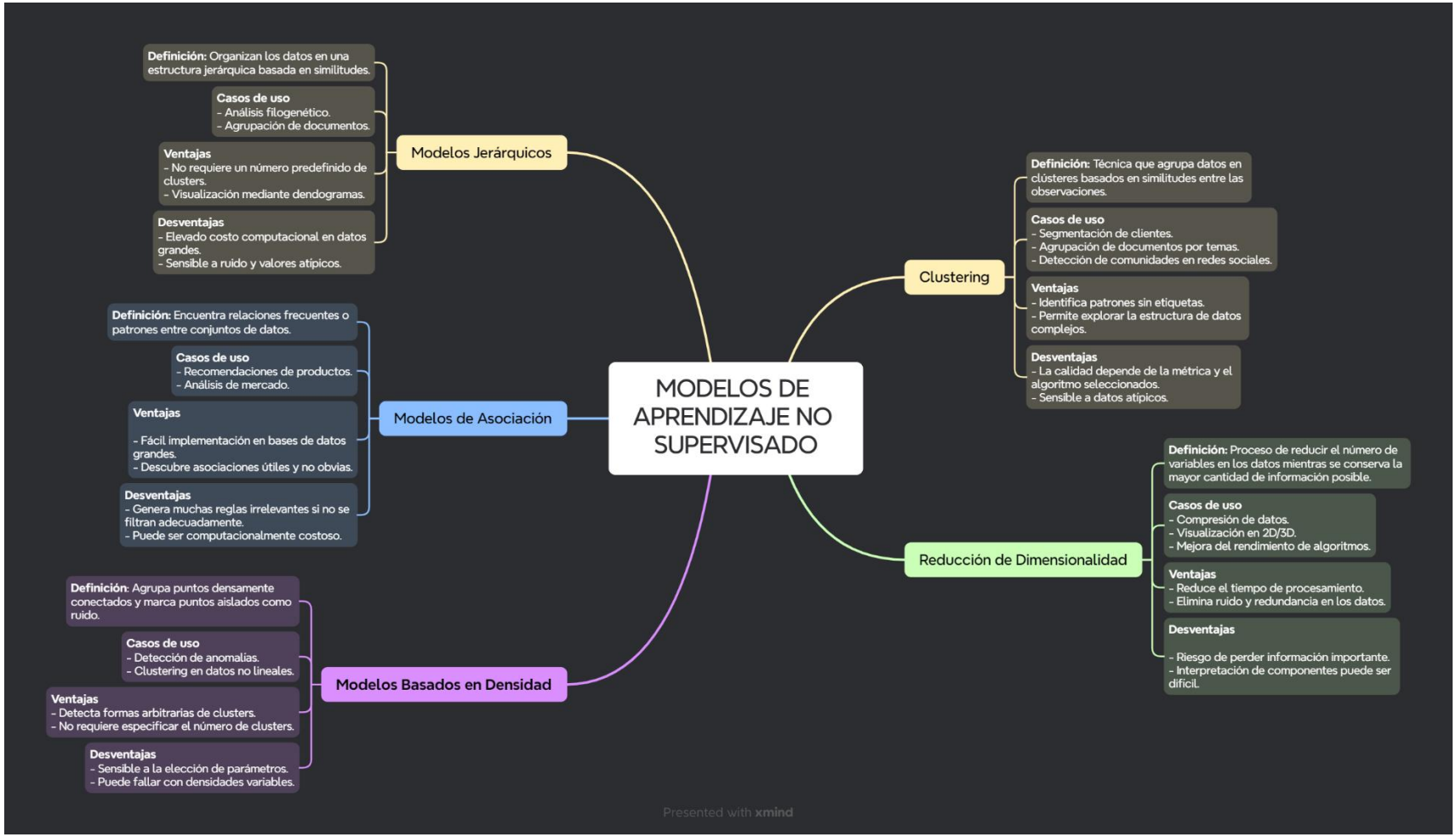

Figura 1. Cuadro sinóptico de modelos de aprendizaje no supervisados.

2 LISTADO DE DEFINICIONES

2.1 Coeficiente de Silhouette

Es una métrica utilizada para evaluar la calidad del clustering. Calcula la cohesión dentro de un cluster y la separación entre clusters, proporcionando un valor entre -1 y 1. Un valor cercano a 1 indica que los puntos están bien agrupados y separados de otros clusters.

2.2 Índice de Calinski-Harabasz

Métrica que evalúa la calidad del clustering al medir la dispersión intra-cluster frente a la dispersión inter-cluster. Un valor alto sugiere que los clusters son compactos y están bien separados.

2.3 Índice de Davies-Bouldin

Evalúa la separación y la compacidad de los clusters. Un valor más bajo indica una mejor calidad del clustering, ya que sugiere clusters compactos y bien separados.

2.4 Coeficiente de Correlación Cofenética

Métrica utilizada para medir qué tan bien un dendrograma preserva las distancias originales entre pares de puntos. Un valor cercano a 1 indica una alta correspondencia entre el dendrograma y las distancias originales.

2.5 Inertia

Medida utilizada en K-means para evaluar la suma de las distancias al cuadrado de cada punto respecto al centroide del cluster al que pertenece. Una inercia más baja indica clusters más compactos.

3 DISEÑO DE MODELO HIERARCHICAL CLUSTERING

3.1 Análisis exploratorio

3.1.1 Importacion de librerías

Debido a que anteriormente hemos usado jupyter notebooks, continuaremos desde este punto habiendo ya creado un documento con el Kernel Python 3.

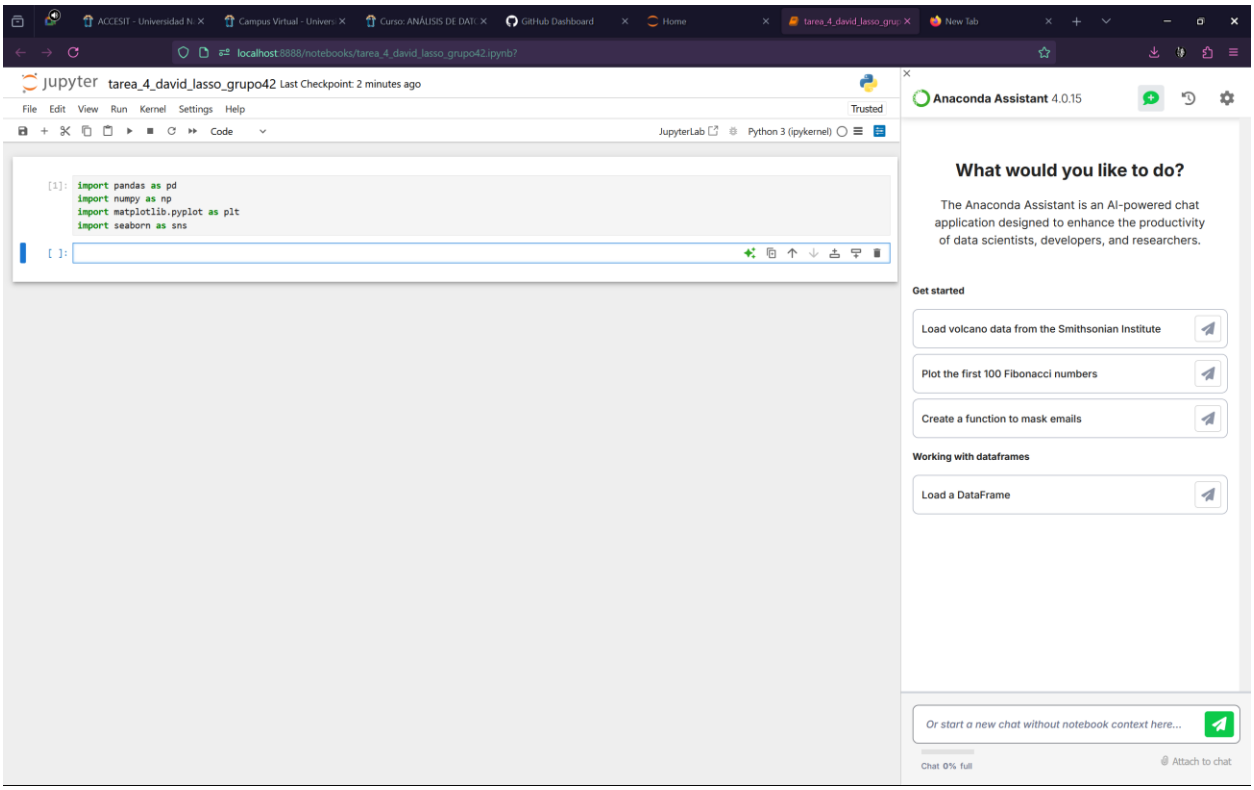

Figura 2. Importación de librerías.

3.1.2 Carga del dataset "Mall_customers" en un DataFrame.

Para que nuestro programa identifique el registro de datos, debemos de entregárselo en un formato compatible. Como ya contamos con este, únicamente debemos de realizar la conexión informando la ubicación de nuestro documento con una dirección absoluta.

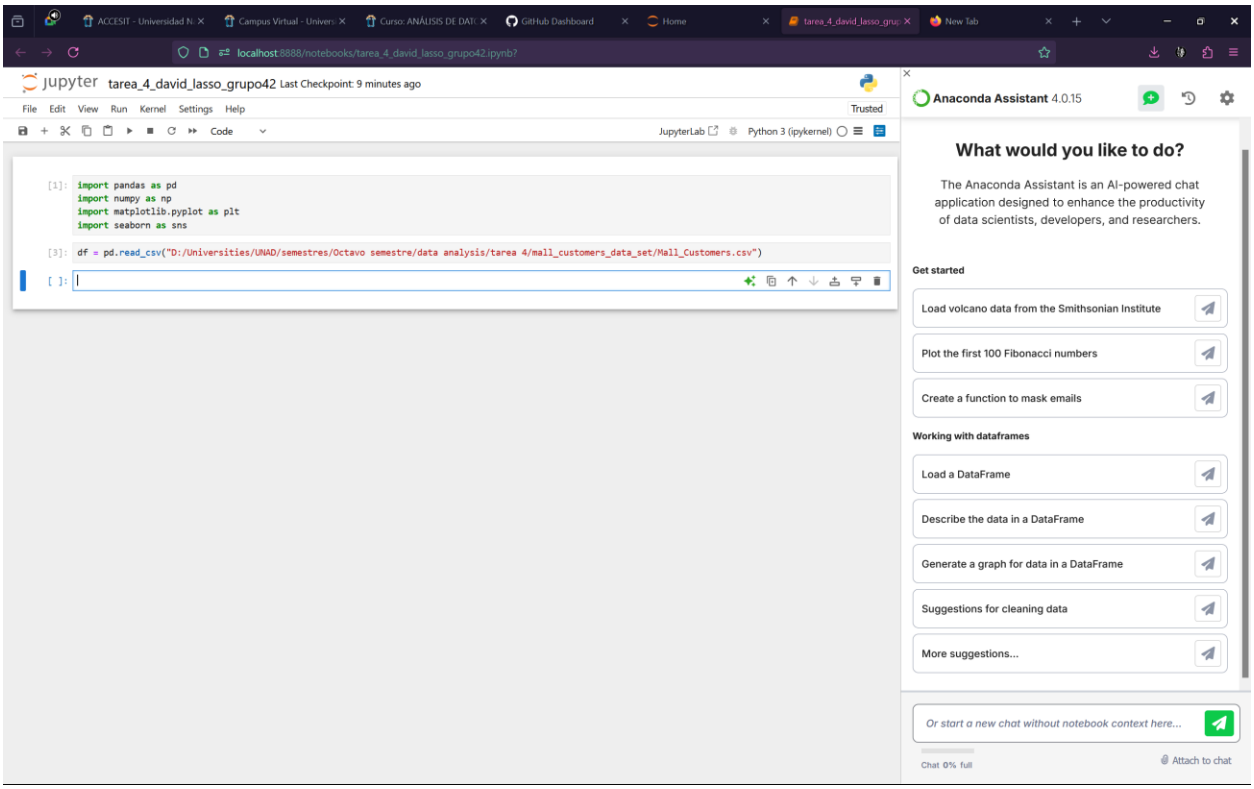

Figura 3. Carga del dataset.

3.1.3 Exploración del dataset

Una vez tenemos la información cargada, podremos observar algunas características como la primera fila del data set, obtener información de general de las columnas y la descripción estadística de los contenidos que se encuentran.

Todo esto lo podemos enviar en un único comando que contenga todas nuestras peticiones, según se observa en la figura 4.

3.2 Identificación de valores faltantes y atípicos

Ahora, como sabremos. Todos los conglomerados de datos pueden contener datos que están fuera de lugar, entonces, tendremos que encargarnos de eliminar esta información así que enviamos el código que se expone en la figura 5.

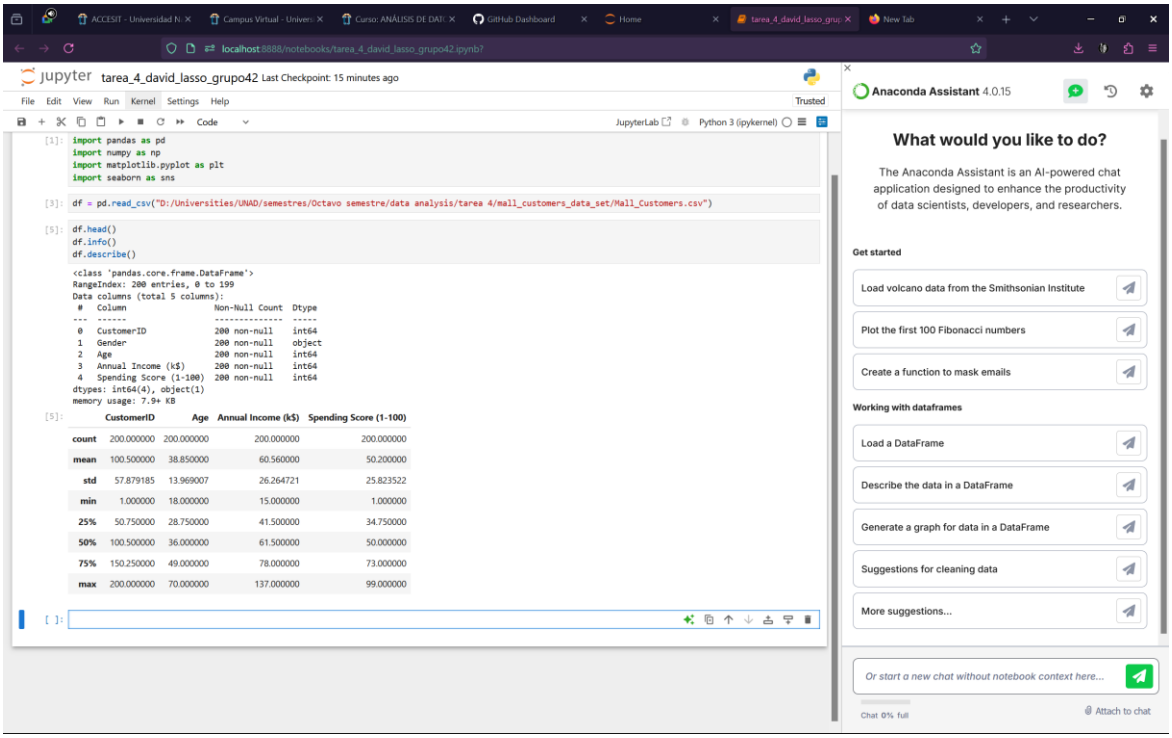

Figura 4. Análisis exploratorio del data set.

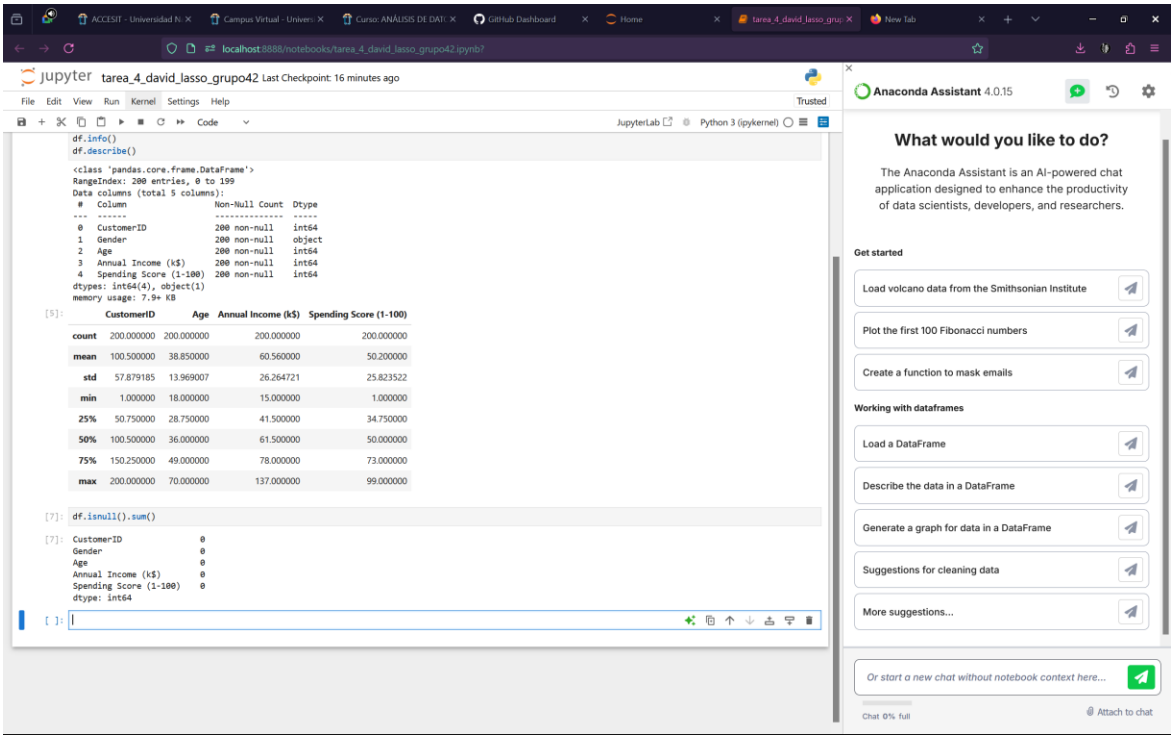

Figura 5. Identificación de valores faltantes.

Como se observa en la figura 5, no se encontraron valores faltantes. Aun así, se realiza otra visualización para asegurar los resultados.

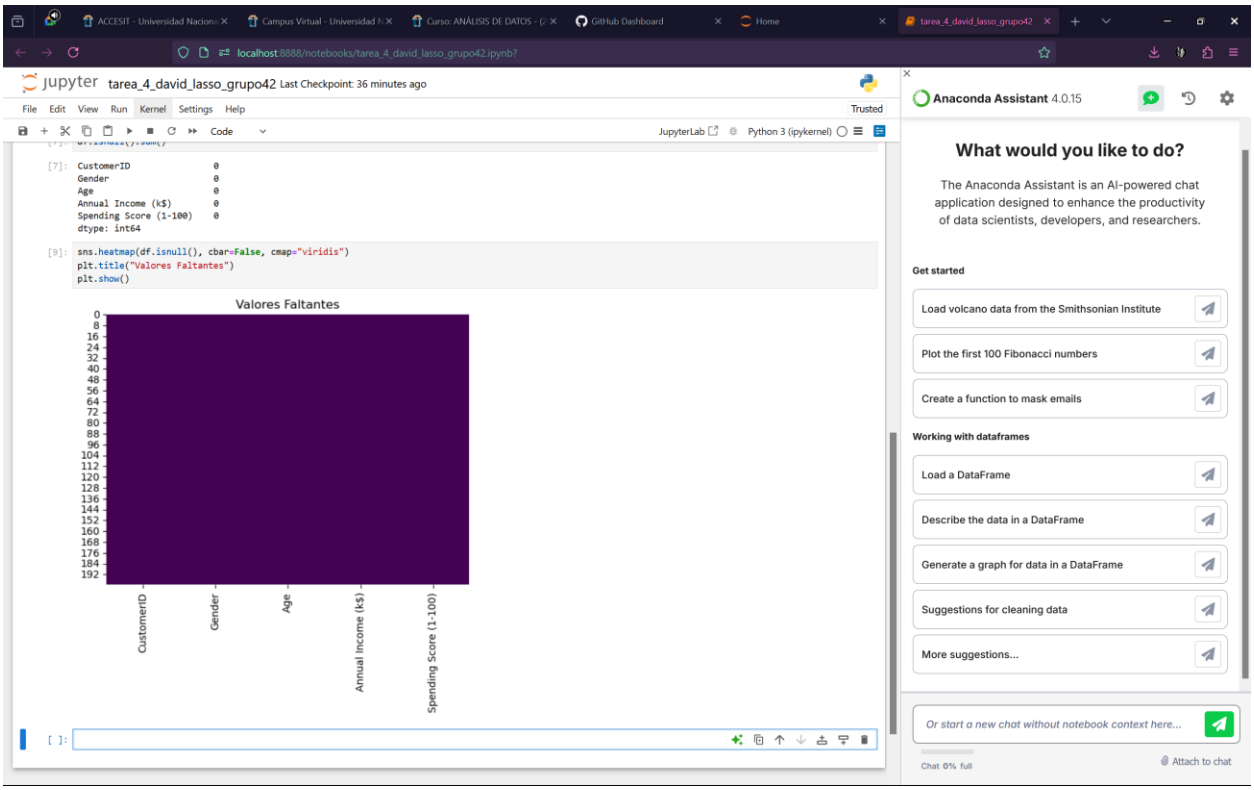

Figura 6. Visualización de datos faltantes para asegurar resultado.

Se puede observar en el resultado que no existen valores faltantes, concluyendo entonces que la pureza del data set es genuina y no es necesario realizar sustituciones o cambios adicionales para que cuente con la información completa.

3.3 Caracteristicas mas relevantes

3.3.1 Visualización de columnas numéricas

En este caso, observaremos inicialmente las variables que tienen un carácter numérico en su contenido.

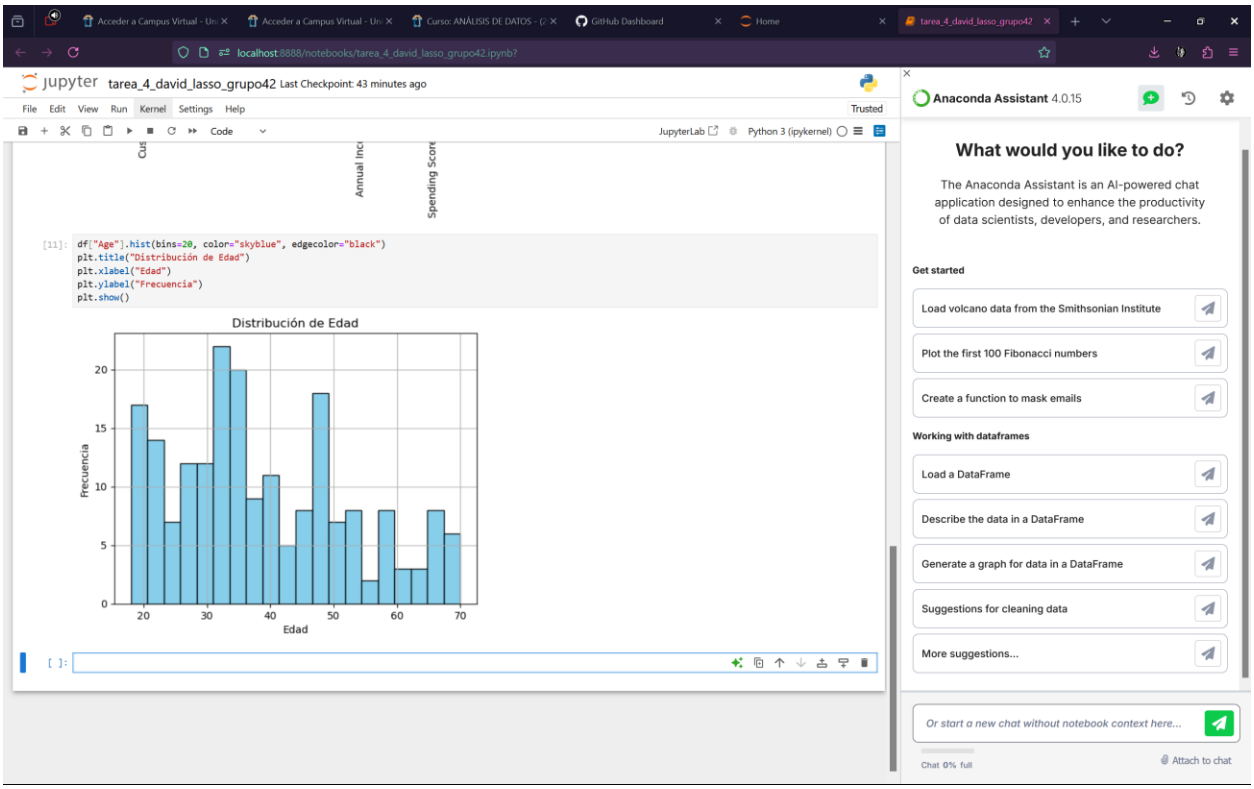

Figura 7. Visualización de variables individuales numéricas.

3.3.2 Visualización de columnas categóricas

Ahora que ya tenemos la salida de valores numéricos, deberemos hacer el mismo proceso, pero con valores categóricos, haciendo que podamos reconocer en este caso, la columna de género, el resultado se puede observar en la figura 8.

3.4 Entrenamiento del modelo

Una vez hemos realizado la visualización de características para permitirnos identificar cuales son las mas relevantes, definiremos entonces ahora el entrenamiento del modelo. Así podremos contar con el conjunto de datos escalado e identificando los valores mas importantes para el clúster que se está usando.

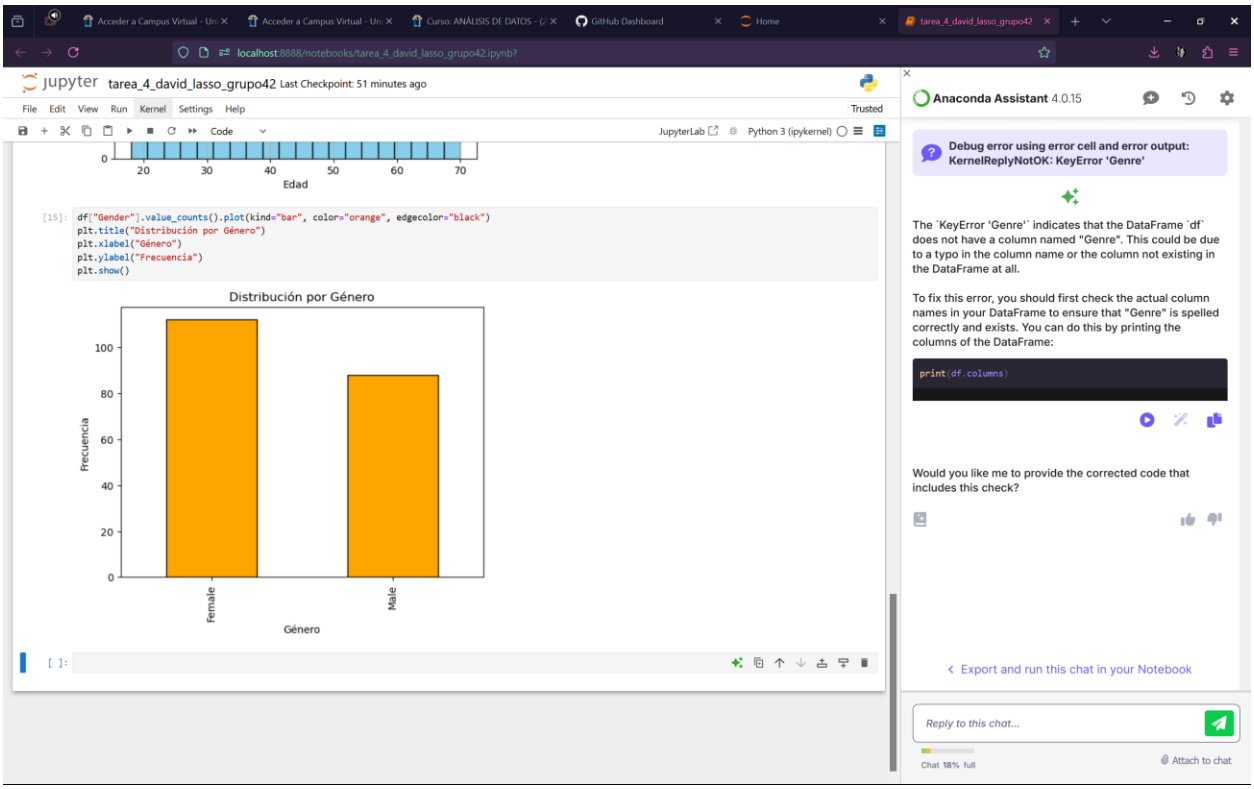

Figura 8. Visualización de característica por género.

3.4.1 Identificación de características más relevantes.

Una vez identificados los valores mediante todos los procesos realizados anteriormente, se clasifican como los mas importantes a los siguientes:

- Age (Edad): Información demográfica importante.
- Annual Income (Ingresos Anuales): Potencial poder adquisitivo.
- Spending Score (Puntuación de Gasto): Tendencia al consumo.

3.4.2 Normalizar las variables numéricas

Habiendo identificado que el proceso debe ser mediante un algoritmo de clustering, incluido el hierarchical clustering, son sensibles a las escalas de las variables. Primero realizamos una normalización de los datos para que cada variable tenga igual valor.

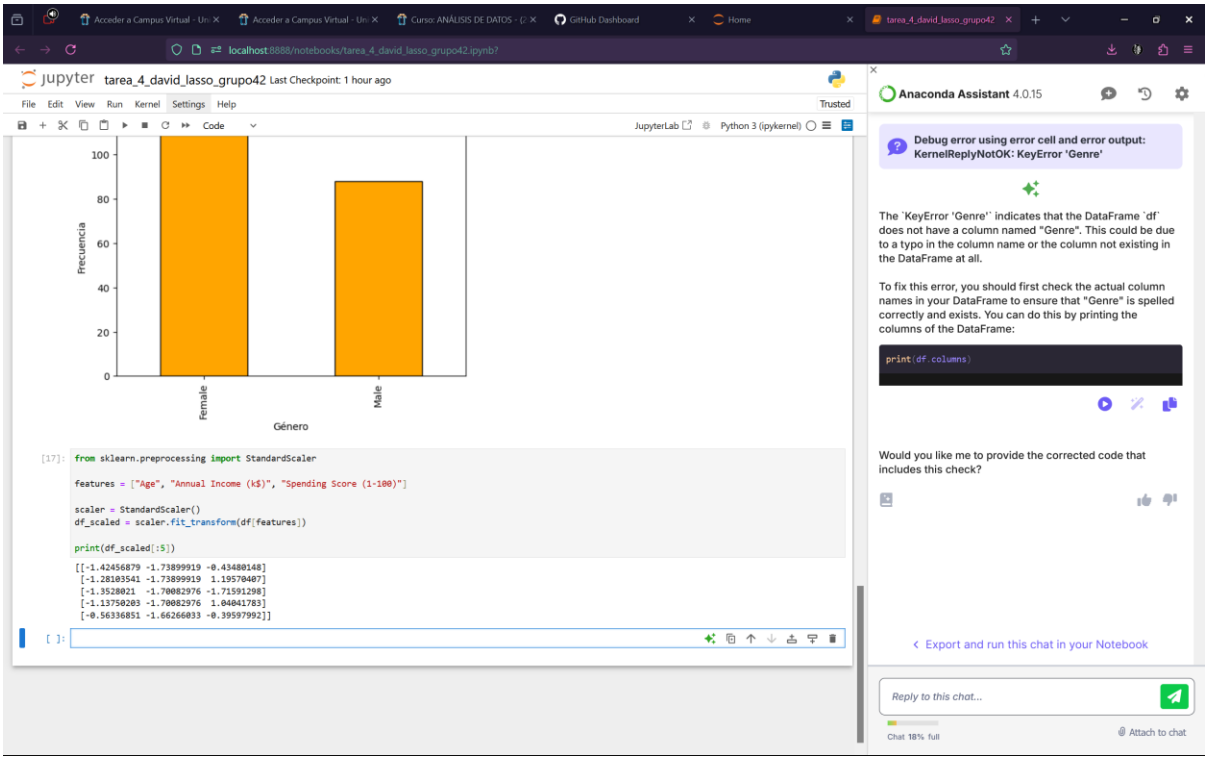

Figura 9. Normalización de variables.

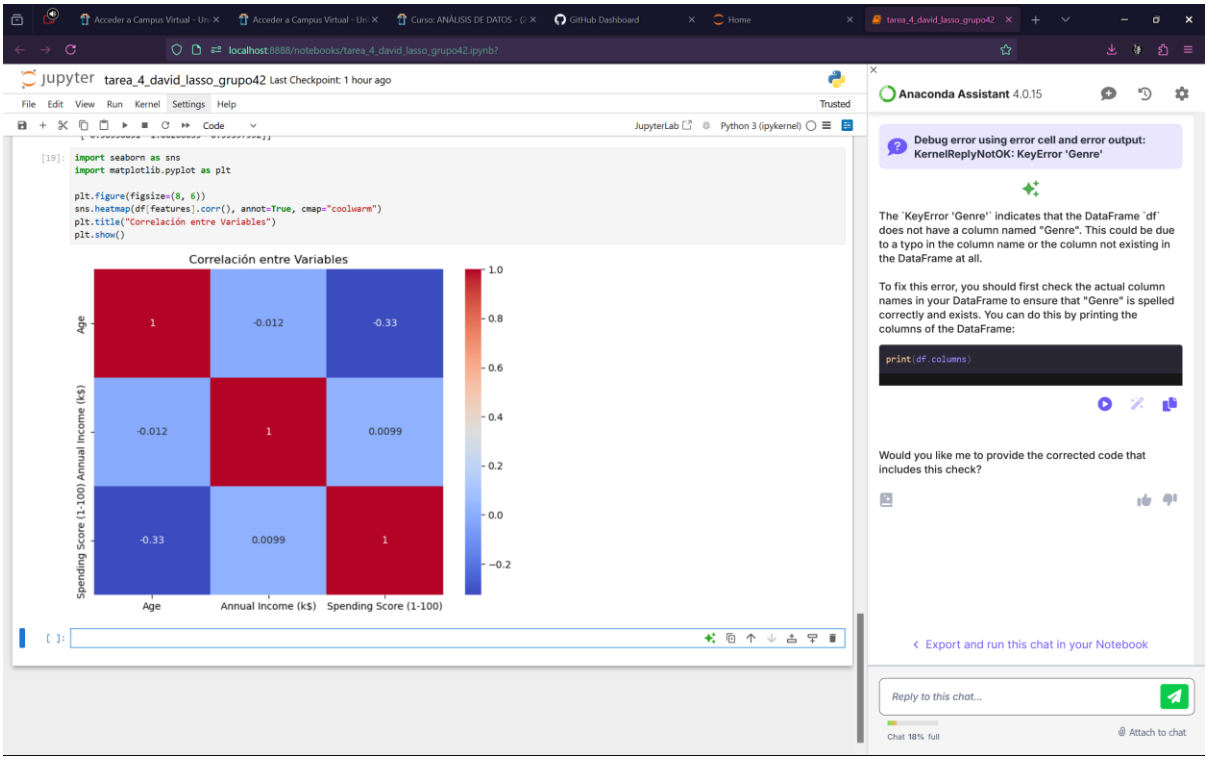

Figura 10. Análisis de relevancia entre características seleccionadas.

3.4.3 Generación de hiperparametros

A continuación, realizaremos la importación de librerías necesarias para el proceso, adicionalmente, se generamos un dendograma para calcular la distancia entre los clústeres.

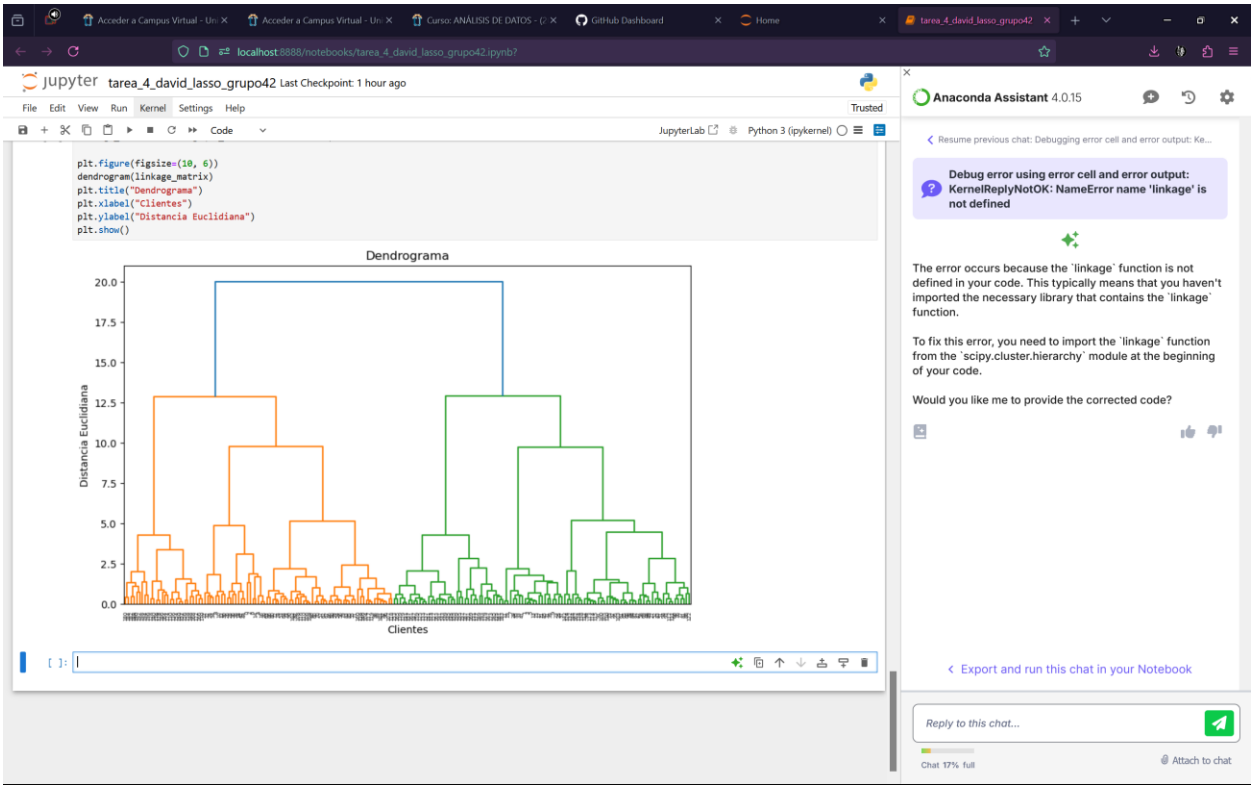

Figura 11. Dendograma generado para medición de clústeres.

Usando el numero de clústeres que vemos en el corte de la grafica que nos otorga, podremos decidir cuantas agrupaciones manejaremos, en nuestro diagrama vemos que la altura 5 obtendremos cuatro grupos grandes, así que intentaremos por este valor.

3.4.4 Configurar y entrenar modelo

Sabiendo el número de clústeres, configuramos el modelo y procedemos a integrarlo.

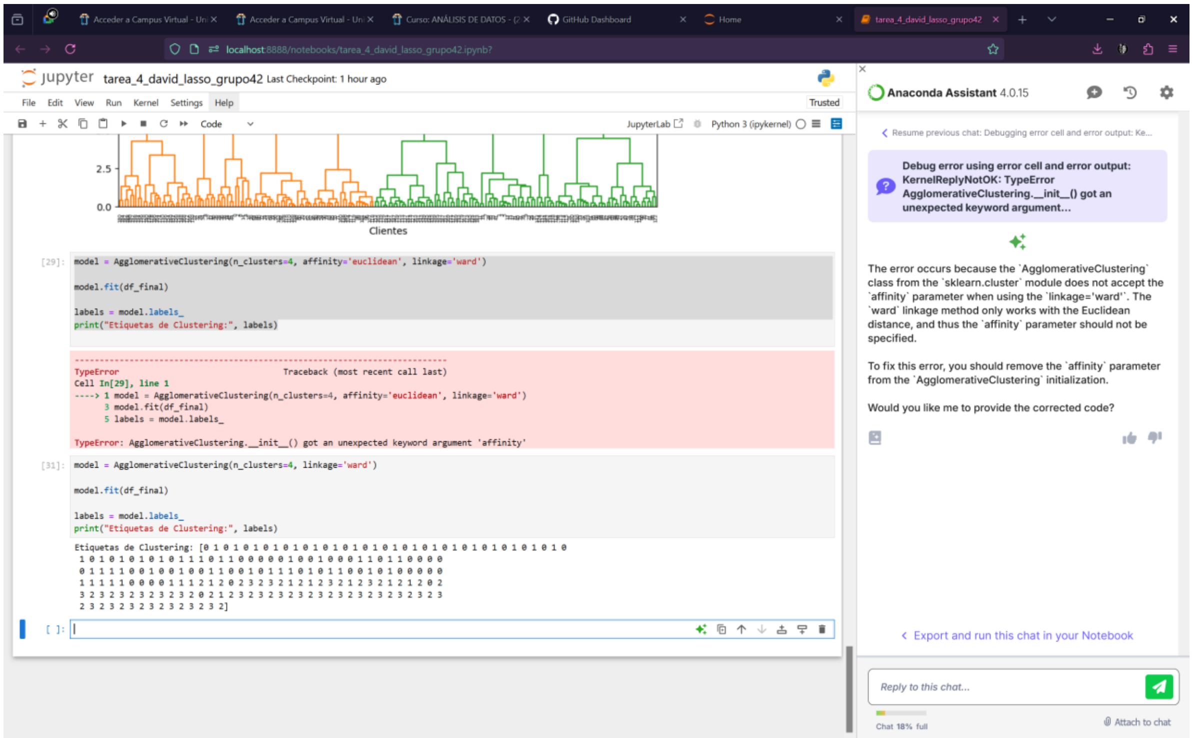

Figura 12. Entrenamiento de clúster.

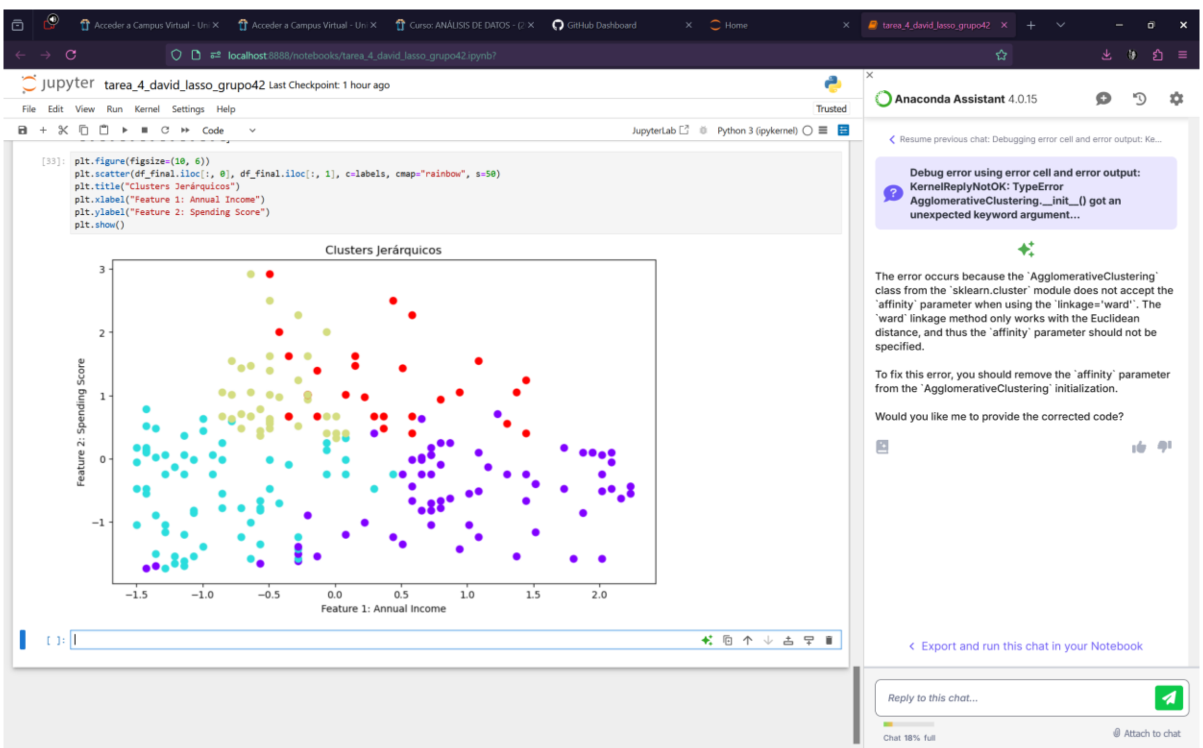

Figura 13. Visualización de clúster generado.

3.5 Evaluación del modelo de clustering jerárquico

Ahora que nuestro modelo ya fue entrenado, tendremos que evaluarlo con las métricas que establece la guía.

En este orden, y haciendo uso de la librería que contiene todas estas métricas de desempeño realizaremos lo siguiente:

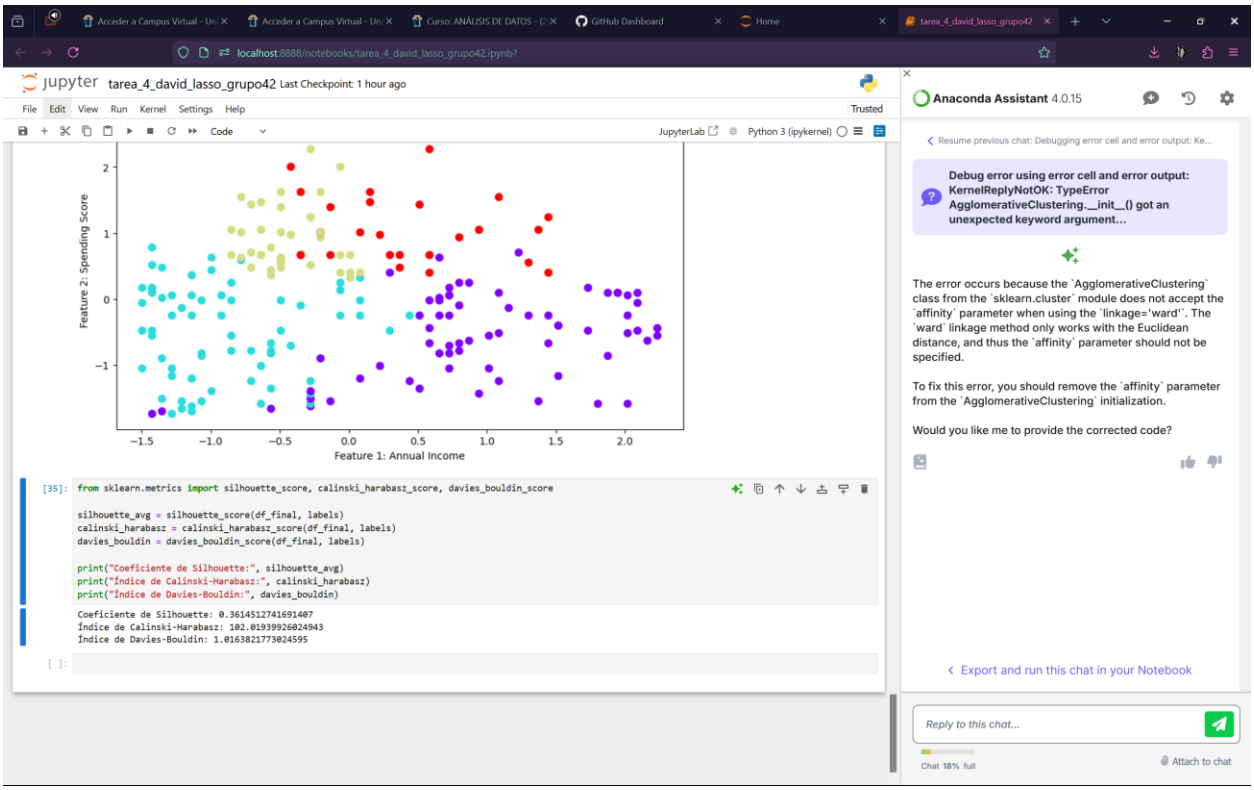

Figura 14. Resultado de evaluación de desempeño.

3.6 Graficas de resultados del modelo

Para evaluar los resultados, implementaremos un gráfico de dispersión 2D y el grafico de cajas para que así podamos definir los resultados como se evidencia en la figura 15 y figura 16.

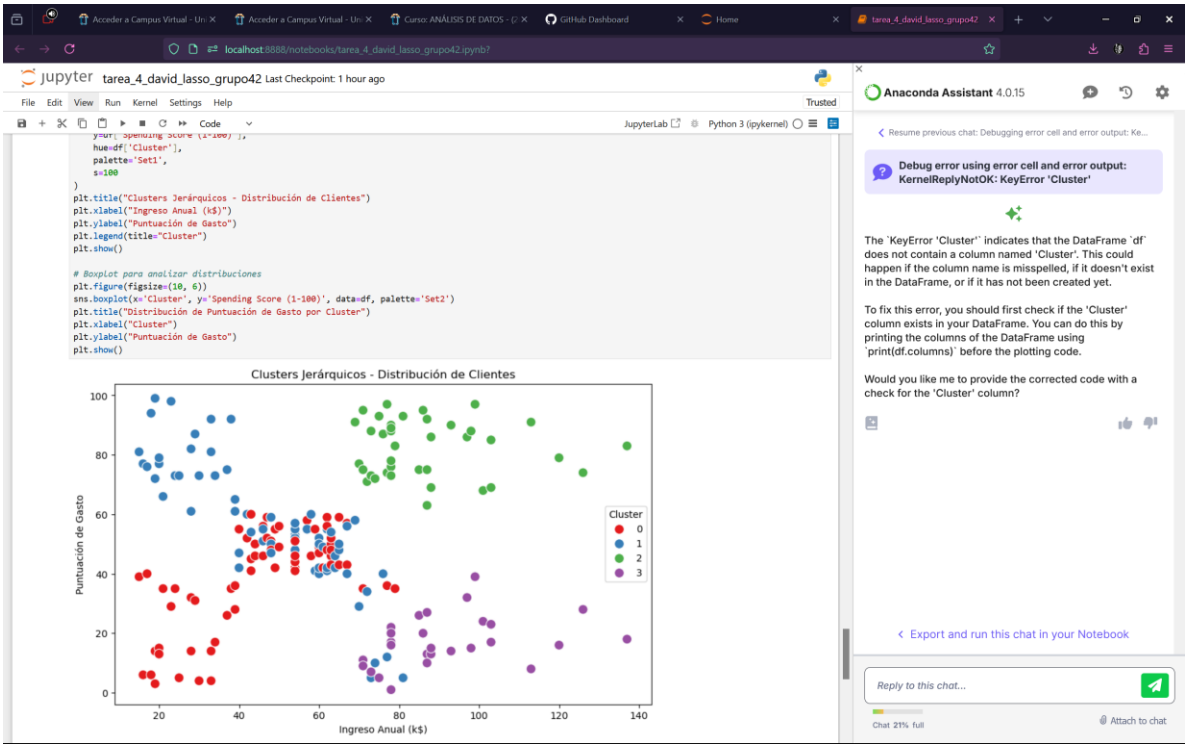

Figura 15. Gráfico de dispersión.

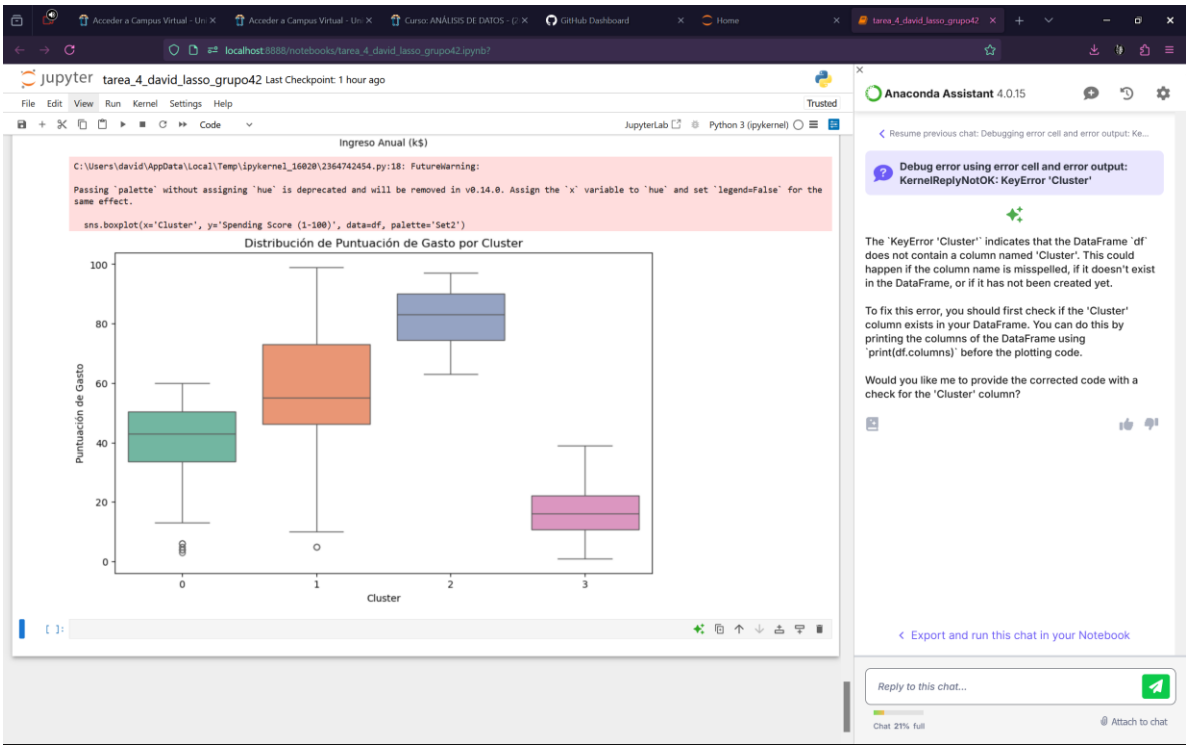

Figura 16. Gráfico de cajas.

3.7 Interpretación y análisis

El modelo de clustering jerárquico ha mostrado una segmentación interesante, con cuatro clústeres definidos. Los clústeres 2 y 3 tienen una separación clara, lo que permite diferenciar estos dos grupos de clientes con facilidad. Por otro lado, los clústeres 0 y 1 tienen una ligera superposición, lo que podría indicar que estos clientes comparten algunas características similares, pero sus perfiles aún pueden diferenciarse con más detalle en otras variables.

La evaluación del modelo ha mostrado resultados moderadamente buenos, con un coeficiente de Silhouette que sugiere que la separación entre los clústeres es aceptable, aunque con algunas áreas de mejora en la asignación de los puntos. El índice de Calinski-Harabasz indica que los clústeres están bien separados, mientras que el índice de Davies-Bouldin sugiere que la calidad de la segmentación podría mejorarse con ajustes en el preprocesamiento o la selección de características.

En términos de visualización, las gráficas de dispersión y el dendrograma proporcionan una representación clara de los resultados del modelo. Las gráficas de cajas también destacan la homogeneidad dentro de cada clúster, lo que sugiere que las segmentaciones de clientes son consistentes en términos de las variables utilizadas.

4 BIBLIOGRAFÍA

Carlos Véliz. (2020). Aprendizaje automático. Introducción al aprendizaje profundo. El Fondo Editorial de la Pontificia Universidad Católica del Perú. https://bibliotecavirtual.unad.edu.co/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=2600876&lang=es&site=eds-live&scope=site&ebv=EB&ppid=pp_I

Giuseppe Bonaccorso. (2018). Machine Learning Algorithms : Popular Algorithms for Data Science and Machine Learning, 2nd Edition: Vol. 2nd ed. Packt Publishing. https://bibliotecavirtual.unad.edu.co/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1881497&lang=es&site=eds-live&scope=site&ebv=EB&ppid=pp_Cover

Minguillón, J. Casas, J. y Minguillón, J. (2017). Minería de datos: modelos y algoritmos. Editorial UOC. https://elibro-net.bibliotecavirtual.unad.edu.co/es/ereader/unad/58656

Pratap Dangeti. (2017). Statistics for Machine Learning : Build Supervised, Unsupervised, and Reinforcement Learning Models Using Both Python and R. Packt Publishing. https://bibliotecavirtual.unad.edu.co/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1560931&lang=es&site=eds-live&scope=site&ebv=EB&ppid=pp_Cover