**Módulo: Machine Learning**

**Actividad: Desafio de Clustering**

**Nombre: Alan Giovanni Venegas Turrubiartes**

**Docente: Jesús Manuel Vázquez Nicolá**

**Fecha: 01/09/2024**

Introducción

En esta actividad, nos centramos en el análisis de clustering, una técnica de aprendizaje no supervisado que permite agrupar datos en clusters basados en similitudes. El objetivo fue practicar diferentes aspectos del clustering, como la exploración de datos, la selección de algoritmos y la evaluación de resultados. Utilizamos un conjunto de datos con características químicas de vinos para aplicar y comparar tres algoritmos de clustering: K-Means, DBSCAN y BIRCH. Además, evaluamos la calidad de los clusters generados mediante el índice de Silhouette y analizamos las visualizaciones obtenidas para entender mejor las agrupaciones resultantes.

Desarrollo

Exploración de Datos

Antes de aplicar los algoritmos de clustering, realizamos una exploración exhaustiva de los datos. El conjunto de datos de vinos contiene 178 instancias y múltiples características químicas, como el contenido de alcohol, la acidez y el color. Para simplificar la visualización y mejorar la interpretación, aplicamos Análisis de Componentes Principales (PCA), una técnica de reducción de dimensionalidad que transforma los datos a un espacio de menor dimensión.

Visualización con PCA: La reducción de dimensionalidad con PCA permitió representar los datos en un gráfico de dispersión bidimensional. Esta visualización ayudó a identificar patrones y posibles agrupaciones naturales en los datos. La distribución de los puntos en el gráfico proporcionó una visión inicial sobre cómo los datos podrían ser agrupados por los algoritmos de clustering.

Aplicación de Algoritmos de Clustering

K-Means: Este algoritmo divide los datos en un número específico de clusters predefinido. En nuestro caso, elegimos 3 clusters, basándonos en la visualización previa y el conocimiento previo sobre el conjunto de datos. K-Means intenta minimizar la variación dentro de cada cluster y maximizar la variación entre los clusters. La puntuación de Silhouette para K-Means fue de 0.56, lo que indica una buena separación entre los clusters, con un nivel razonable de coherencia dentro de cada grupo.

DBSCAN: A diferencia de K-Means, DBSCAN no requiere que especifiquemos el número de clusters. En su lugar, el algoritmo define clusters basados en la densidad de puntos. Utilizamos el parámetro eps para definir la distancia máxima entre puntos para que se consideren vecinos y el parámetro min\_samples para el número mínimo de puntos en un cluster. La puntuación de Silhouette para DBSCAN fue de 0.21, que es significativamente menor que la de K-Means. Esto sugiere que DBSCAN tuvo dificultades para encontrar una estructura de clusters clara en nuestros datos, posiblemente debido a una configuración inadecuada de los parámetros.

BIRCH: Este algoritmo es eficaz para grandes volúmenes de datos y es útil cuando se desea una estructura jerárquica de clusters. BIRCH también requiere especificar el número de clusters, por lo que usamos el mismo número que en K-Means. La puntuación de Silhouette para BIRCH fue de 0.56, similar a la de K-Means. Esto sugiere que BIRCH fue igualmente eficaz en la separación de los clusters, proporcionando una calidad de clustering comparable.

Evaluación de Resultados

La evaluación de los resultados de clustering se realizó utilizando el índice de Silhouette, que mide la calidad de los clusters. La puntuación de Silhouette varía de -1 a 1, donde un valor cercano a 1 indica clusters bien separados, mientras que un valor cercano a -1 indica que los datos están mal agrupados.

K-Means y BIRCH: Ambas técnicas mostraron una puntuación de Silhouette de 0.56, lo que sugiere que los clusters formados por estos algoritmos están bien definidos y separados. La similitud en las puntuaciones indica que ambos algoritmos son efectivos para el conjunto de datos utilizado.

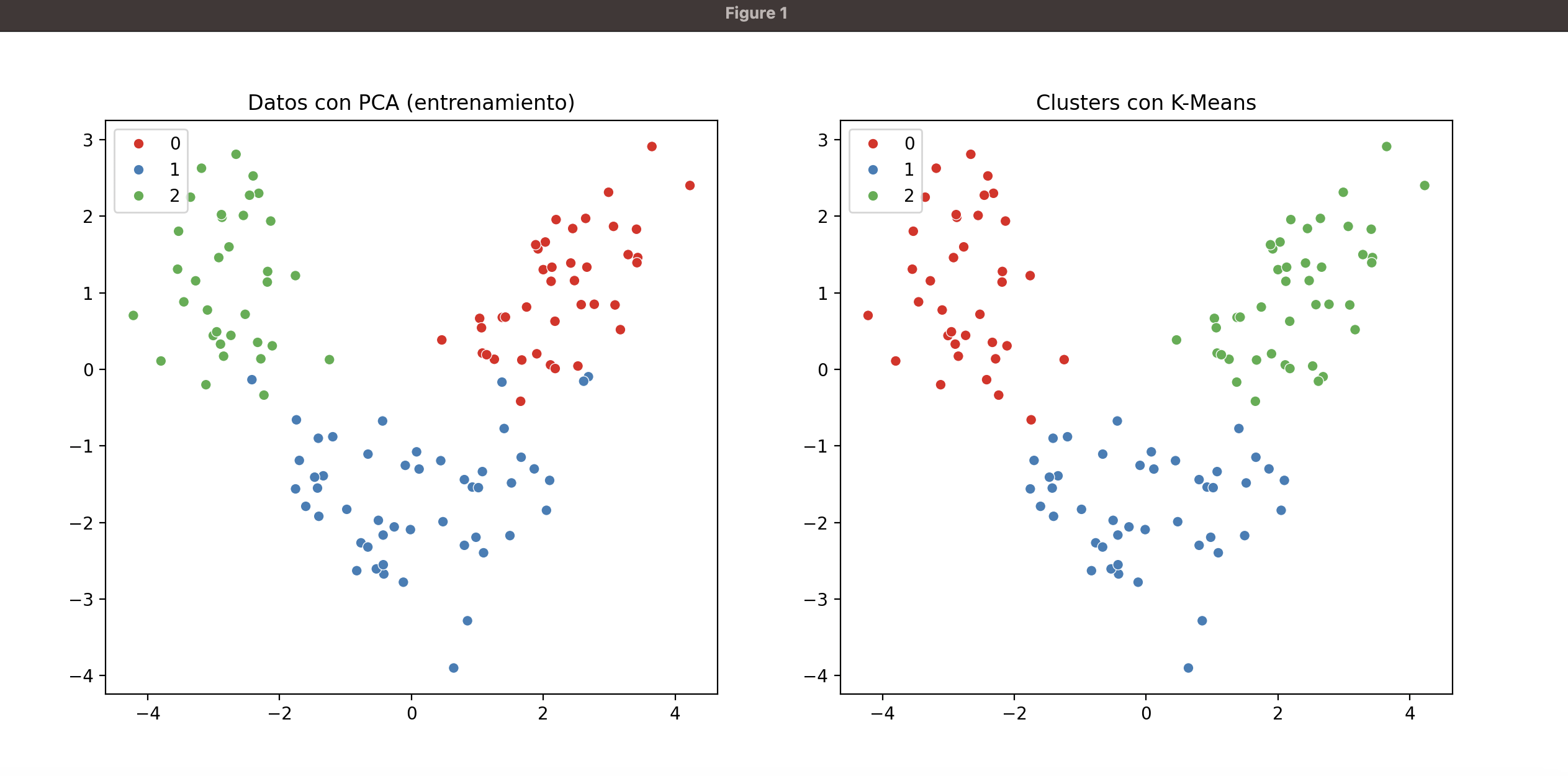
DBSCAN: La puntuación de 0.21 indica que los clusters generados por DBSCAN no están tan bien definidos. Esto puede deberse a la elección de parámetros, como el eps, que podrían no haber sido óptimos para la estructura de los datos.

Las visualizaciones proporcionaron dos gráficos clave:

Datos con PCA: Este gráfico mostró cómo los datos se distribuyen en un espacio reducido, ayudando a identificar posibles agrupaciones.

Clusters con K-Means: Este gráfico visualizó los clusters generados por el algoritmo K-Means, destacando la separación entre los grupos.

Graficas de Dispersión.



Conclusión

En conclusión, los resultados de esta actividad nos proporcionaron una visión clara de cómo diferentes algoritmos de clustering funcionan y se comparan. K-Means y BIRCH demostraron ser efectivos para el conjunto de datos, con puntuaciones de Silhouette similares que indican una buena separación entre clusters. DBSCAN, por otro lado, mostró una puntuación más baja, lo que sugiere que puede haber dificultades en la identificación de clusters claros con la configuración actual de parámetros.

Las visualizaciones y las métricas de evaluación ofrecieron una perspectiva útil sobre la calidad del clustering y ayudaron a comprender las características y limitaciones de cada algoritmo. En futuras aplicaciones, ajustar los parámetros de DBSCAN y explorar otros métodos de clustering podrían mejorar los resultados. La práctica y la evaluación continua son esenciales para aplicar clustering de manera efectiva en problemas reales.

Anexos

Codigo Utilizado en Python

Bibliografias