Informe: Análisis de Clusterización con K-Means

Descripción del Conjunto de Datos

El conjunto de datos seleccionado para este proyecto es el "Dry Bean Dataset", el cual proporciona información sobre la clasificación de siete tipos de frijoles secos, basada en sus características físicas. La elección de este conjunto surge de la necesidad de agrupar los datos en clases o grupos, con el objetivo de identificar los rasgos distintivos de cada tipo de frijol y determinar qué variedades presentan similitudes o compatibilidad entre sí. Para ello, se utilizó un método de aprendizaje no supervisado, lo que hace del algoritmo K-Means una herramienta ideal para este análisis.

Proceso de Preprocesamiento y Aplicación de K-Means

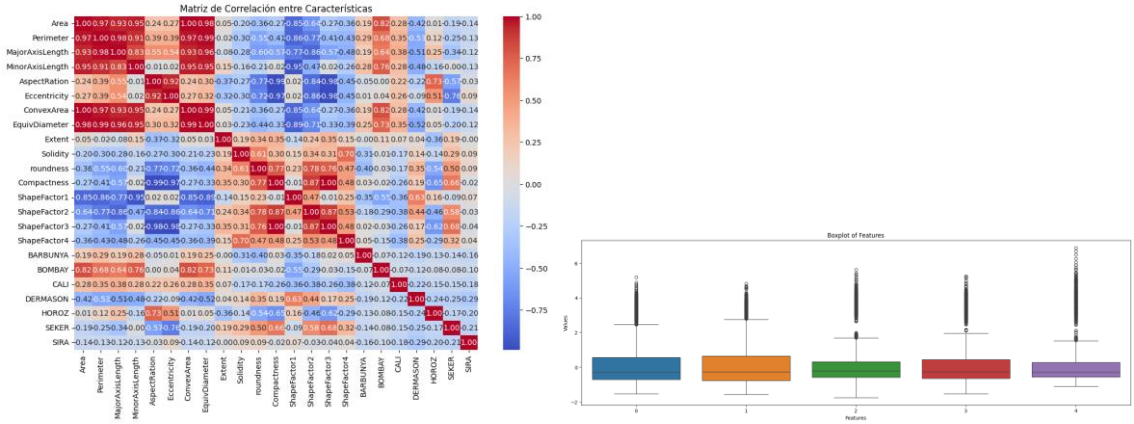
Análisis exploratorio

El proceso de preprocesamiento comenzó con un análisis detallado de las características del conjunto de datos, identificando las variables más relevantes y evaluando su comportamiento. Se inició con la conversión de los tipos de datos a formatos más adecuados, seguida de una interpretación basada en cálculos sencillos de análisis, como la varianza y la desviación estándar, lo que permitió obtener una comprensión más profunda de la información.

Transformación de los datos

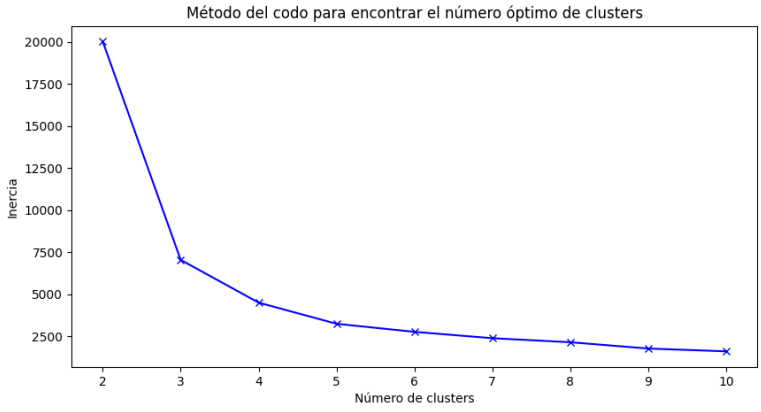
Una vez comprendido el contexto del dataset, se realizó una limpieza exhaustiva para evitar errores y analizar en mayor detalle los datos necesarios. Se eliminaron los valores nulos y se investigaron los siete tipos de muestras. En este proceso, se identificaron las variables más significativas utilizando un mapa de calor, entre las que se destacan: Área (Area), Perímetro (Perimeter), Longitud del Eje Mayor (MajorAxisLength), Longitud del Eje Menor (MinorAxisLength), Diámetro Equivalente (EquivDiameter) y Área Convexa (ConvexArea). Con estas variables seleccionadas, se preparó un nuevo conjunto de datos, que fue estandarizado para prevenir errores

futuros. Además, se llevó a cabo un análisis de los datos atípicos (outliers), concluyendo que, aunque existen en gran cantidad, eliminarlos comprometería la representatividad de la muestra. Por tanto, se decidió mantenerlos al considerarse relevantes.



Aplicación de K-Means

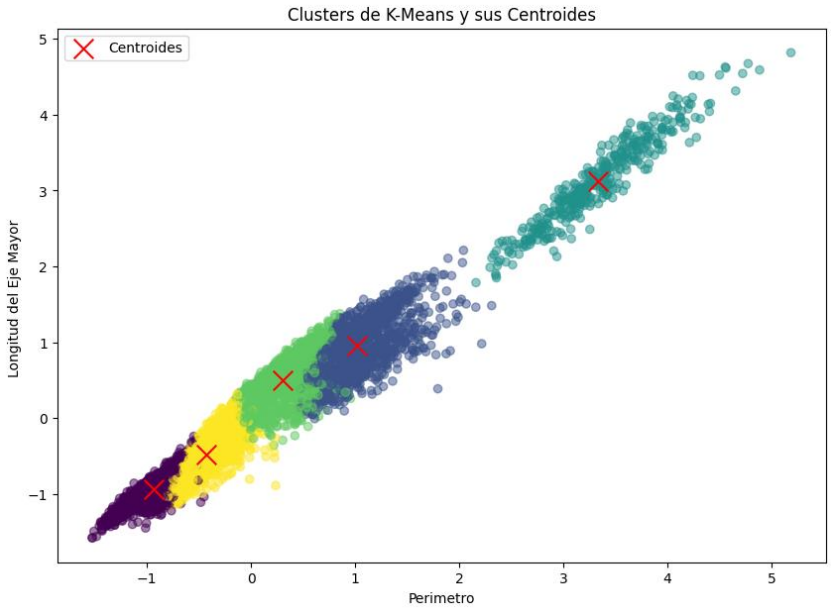
El algoritmo K-Means se aplicó para agrupar los datos en un número específico de clusters, determinado utilizando herramientas como el gráfico del codo, que indicó que el número ideal de clusters era 5, punto en el cual los datos comenzaban a estabilizarse. Posteriormente, el diagrama de silueta corroboró esta conclusión, sugiriendo que el número óptimo de clusters está entre 5 y 6. Tras definir la cantidad de clusters, se implementaron los centroides y se generó un gráfico que muestra claramente la separación entre las distintas agrupaciones.




```
KMeans
KMeans(n_clusters=5, random_state=0)
```

Interpretación de los Clusters

La gráfica de los clusters y sus centroides ofrece una visualización clara de cómo se agrupan los frijoles según sus características físicas. Al analizar los centroides y la distribución de los clusters, se puede observar que están bien separados, lo que demuestra la efectividad del modelo. Además, se aprecia la variabilidad en el tamaño de cada cluster, lo que refleja la diversidad dentro de las agrupaciones. Esta variabilidad puede deberse a la diferencia en el número de muestras entre los siete tipos de frijoles, ya que algunos contaban con más datos que otros. Finalmente, aunque los centroides no presentan diferencias extremadamente marcadas entre sí, se puede identificar una separación evidente entre los distintos clusters, lo que respalda la adecuada segmentación del conjunto de datos.



Conclusiones y Aplicaciones

El análisis de los datos mediante el algoritmo K-Means reveló patrones interesantes dentro del conjunto de datos 'Dry Bean'. Los clusters formados permiten segmentar los diferentes tipos de frijoles secos según sus características físicas, lo que puede tener aplicaciones en la industria alimentaria, ayudando a automatizar el proceso de clasificación de frijoles.

- Determinar la calidad de los frijoles y detectar posibles anomalías en las variables del grano.
- Identificar diferentes tipos de frijoles para mejorar la clasificación y el marketing.
- Analizar las características físicas de los frijoles para seleccionar las mejores variedades para el cultivo.
- Identificar los tipos de frijol que se acoplen mejor a cierto tipos de comidas
- Evidenciando las características de lo frijoles que mas se venden en cada región se puede segmentar la población para y tipos de frijoles para promocionarlos.