Trabajo Práctico N° 1

Minería de datos

Informe de análisis de datos y clustering

Tecnicatura Universitaria en Inteligencia Artificial

Facultad de Ciencias Exactas, Ingeniería y Agrimensura

Universidad Nacional de Rosario

2023

Integrantes:

Augusto Farias

Guido Lorenzetti

0. Introducción

El presente dataset tiene información de diferentes granos de cultivos. En el mismo se encuentran cantidades de nitrógeno(N), fósforo (P) y potasio (K), temperature, humidity, ph (grado de acidez o alcalinidad), rainfall (mm de lluvia caídos) y label (etiqueta del grano).

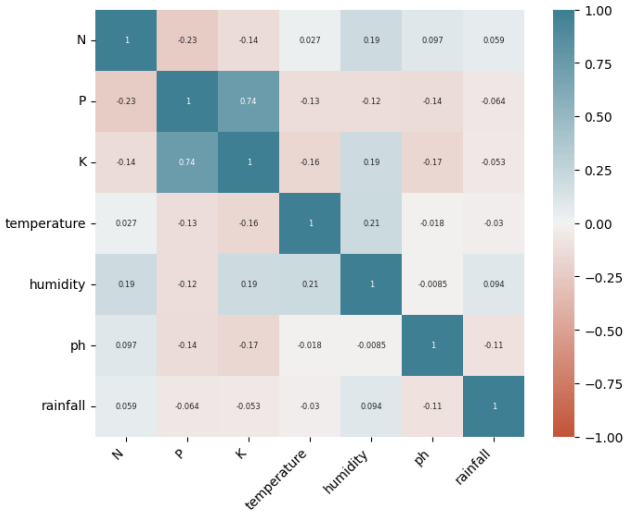
1. Análisis de Atributos:

No observamos valores nulos en el dataset, por lo cual podemos decir que está completo.

A la hora de verificar si existen outliers lo realizamos por tipo de grano. Es decir, se agrupan los datos por tipo de grano y se analizan los valores.
En este procedimiento no observamos outliers, es decir, valores atípicos o anormales.

Las columnas N, P, K, temperature, humidity, ph y rainfall son de tipo numérico, mientras que label, contiene un string que caracteriza al grano.

Para observar la correlación de los datos utilizamos la matriz de correlación de pearson:



Para trabajar con los datos procedimos a estandarizarlos para que las distintas escalas de los distintos datos no nos influyeran en el procedimiento. Para esto utilizamos StandarScaler() de la librería sklearn, la cual utiliza la media de la característica y la desviación estándar para normalizar los datos.

2. Análisis de PCA:

En este análisis, aplicamos el método de Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de nuestro conjunto de datos. A continuación, se detalla el proceso y los criterios de selección utilizados:

1. **Selección del número de componentes principales:**
   Para determinar el número adecuado de componentes principales, consideramos varios criterios:

a. **Proporción de varianza acumulada:** Observamos el gráfico de la varianza acumulada en función del número de componentes. Descubrimos que las cinco primeras componentes acumulan aproximadamente el 90% de la variabilidad total, cumpliendo así con el primer criterio (proporción de varianza acumulada de al menos el 75%-80%).

b. **Criterio de Kaiser:** Comprobamos los eigenvalues de las componentes. Notamos que algunos eigenvalues son cercanos a 1, lo que cumple con el segundo criterio (eigenvalues > 1).

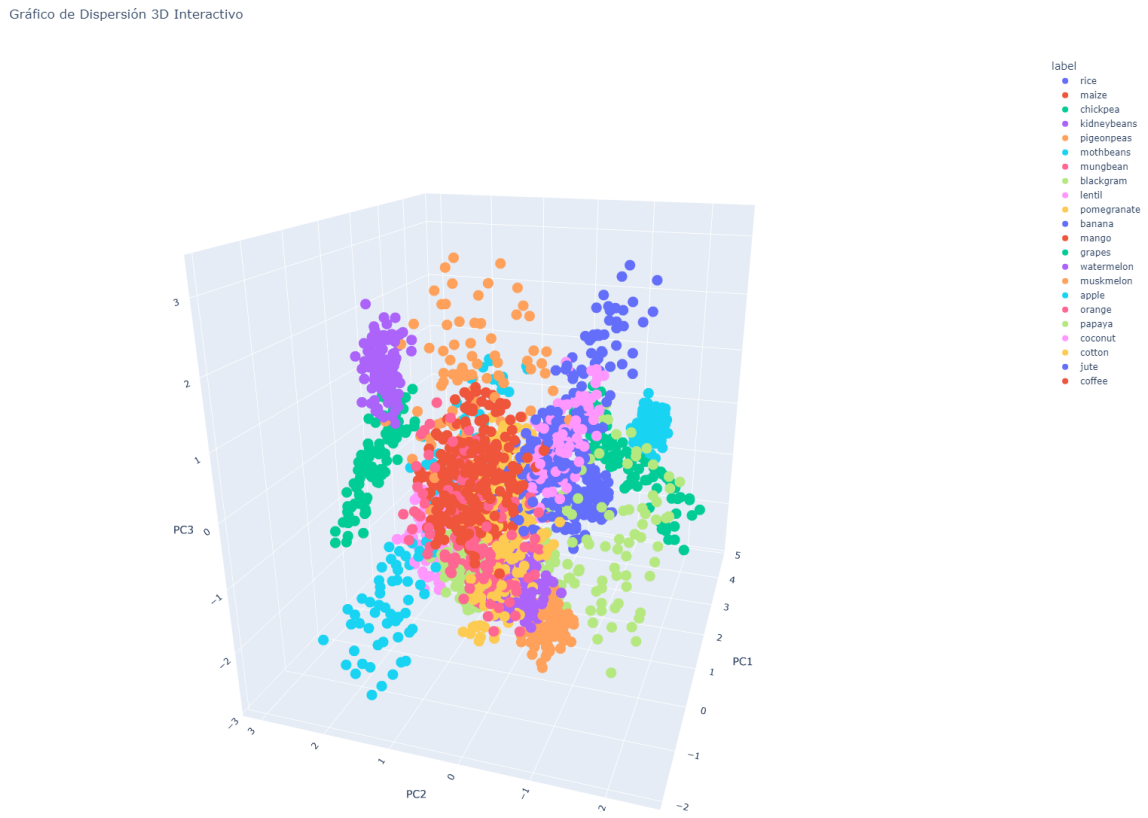
c. **Gráfico del codo (Scree plot):** Graficamos la proporción de varianza explicada por cada componente. Observamos que el quiebre en la gráfica parece producirse entre la cuarta y quinta componente. Tomando hasta la quinta componente, cubriríamos aproximadamente el 70% de la variabilidad total, lo que cumple con el tercer criterio.

En función de estos criterios, decidimos tomar cinco componentes principales, aunque con el propósito de visualización, elegimos utilizar solo tres.

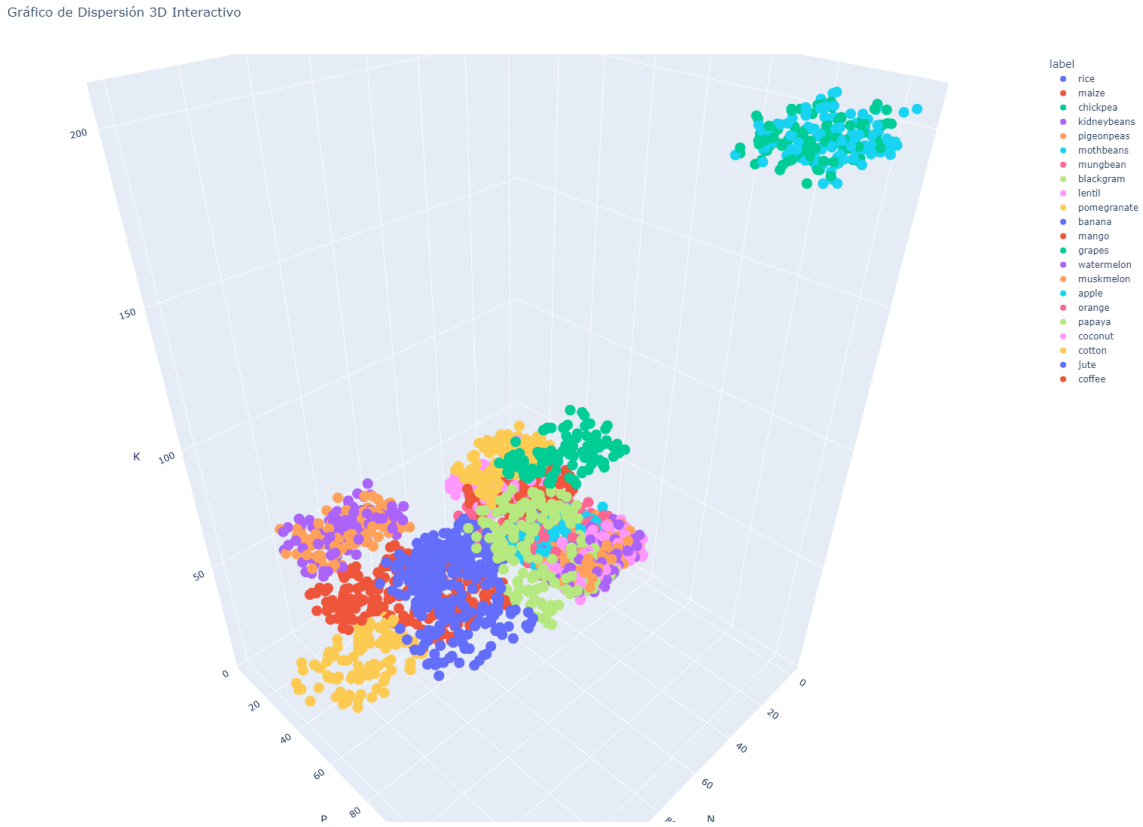
2. **Reducción de dimensionalidad y visualización:**

Aplicamos PCA con el número seleccionado de componentes principales (en este caso 3) y obtuvimos nuevas características transformadas.

Utilizamos un gráfico de dispersión 3D interactivo para visualizar los datos transformados en función de las tres componentes principales. Esto nos permite apreciar cómo las clases originales se distribuyen en el espacio de las componentes principales.



Además, comparamos esta representación con los gráficos de dispersión 3D de las características originales (sin escalar) y las características escaladas para evaluar la influencia de la estandarización en la dispersión de los datos.



Podemos observar la diferencia de dispersión entre el dataframe escalado y el mismo sin escalar.

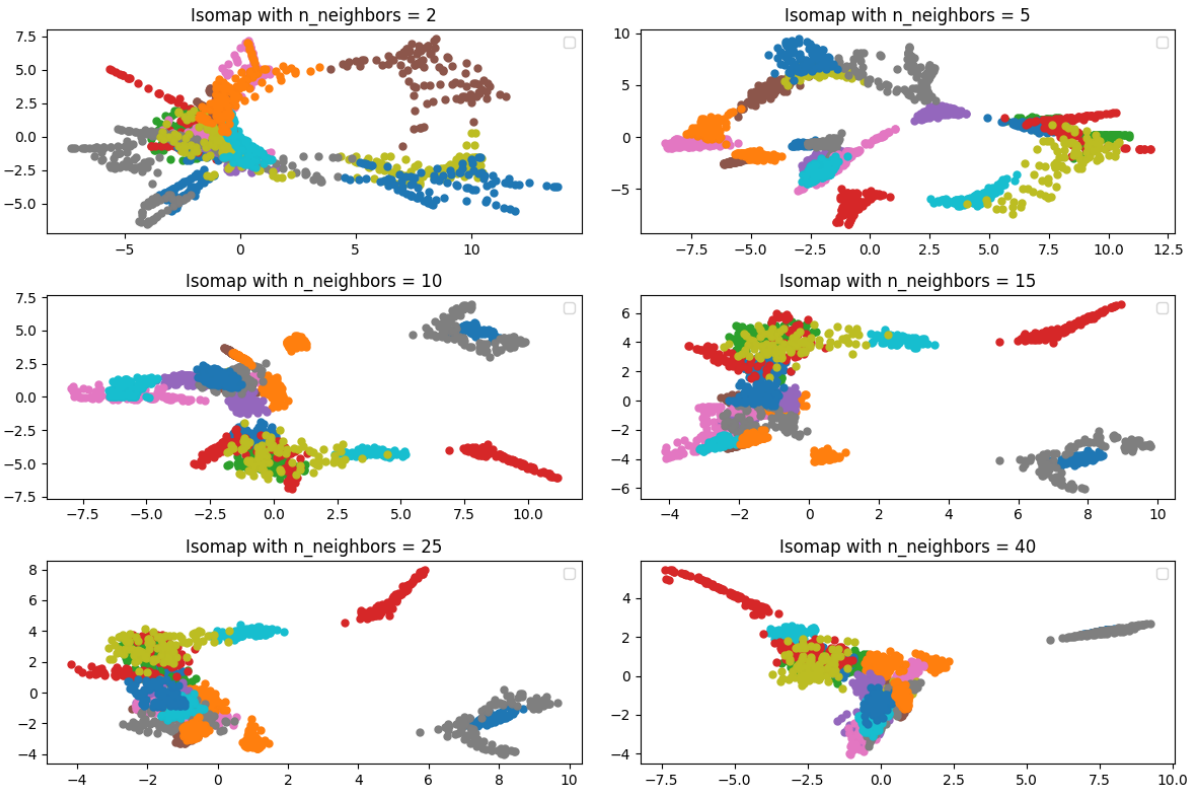
3. Análisis de Isomap:

En este análisis, aplicamos la técnica de Isomap para reducir la dimensionalidad de nuestro conjunto de datos. Isomap es un algoritmo de reducción de dimensionalidad que se basa en la preservación de distancias geodésicas (distancias a lo largo de caminos en el espacio de datos). Exploramos la influencia de diferentes valores de **`n_neighbors`** en el proceso de reducción y visualización de los datos.

Variación de n_neighbors:

Comenzamos el análisis evaluando el impacto de diferentes valores de **`n_neighbors`** en la representación Isomap. Creamos una serie de subplots, cada uno correspondiente a un valor diferente de **`n_neighbors`** (2, 5, 10, 15, 25 y 40). Para cada valor de **`n_neighbors`** , aplicamos Isomap y proyectamos los datos en un espacio bidimensional (2D).

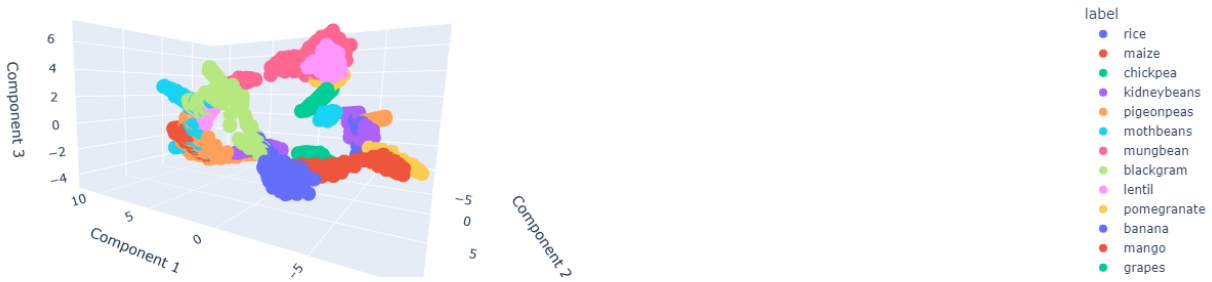
En los gráficos resultantes, observamos cómo las clases originales se distribuyen en función de las dos primeras componentes principales generadas por Isomap. Podemos apreciar cómo la elección de **`n_neighbors`** afecta la estructura y separabilidad de las clases en el espacio reducido.



Elección del número óptimo de componentes:

Para una representación tridimensional (3D), seleccionamos un valor de **`n_neighbors`** específico (en este caso, 5) y aplicamos Isomap para obtener tres componentes principales. Esto nos permite explorar la estructura tridimensional de los datos.

Utilizamos Plotly Express para visualizar los datos en un gráfico de dispersión 3D. En esta visualización, las clases se representan con diferentes colores, lo que facilita la identificación de patrones y relaciones en el espacio de tres dimensiones.



Resultados y Conclusiones:

El análisis de Isomap nos proporciona una representación efectiva de los datos en un espacio de menor dimensionalidad, lo que facilita la visualización y la identificación de patrones. La variación en **`n_neighbors`** permite explorar cómo la elección de vecinos cercanos influye en la estructura de los datos reducidos. La representación en tres dimensiones (3D) muestra cómo las clases se distribuyen en un espacio más rico.

En resumen, Isomap es una técnica valiosa para la reducción de dimensionalidad y la visualización de datos, y su elección de parámetros, como **`n_neighbors`** , puede influir en la interpretación de los resultados

4. Análisis de t-SNE:

El algoritmo t-SNE calcula una medida de similitud entre pares de instancias en el espacio de alta dimensión y en el espacio de baja dimensión. Luego trata de optimizar estas dos medidas de similitud usando una función de costo

Exploraremos cómo diferentes configuraciones de hiper parámetros afectan la proyección y la interpretación de los datos.

Variación de hiperparámetros:

Comenzamos probando diversas configuraciones de hiper parámetros, incluyendo el número de iteraciones, el número de componentes y la perplejidad. Cada configuración se representa en un subplot individual. La variación de estos hiper parámetros nos permite comprender cómo influyen en la disposición de las clases en el espacio de menor dimensión.

- **Número de Iteraciones:** Diferentes valores de iteraciones afectan la calidad y la convergencia de la reducción de dimensionalidad. A medida que aumentamos el número de iteraciones, t-SNE tiene más tiempo para ajustarse a las estructuras locales de los datos.
- **Número de Componentes:** Al elegir entre 2 o 3 componentes observamos cómo la elección influye en la visualización de los datos. Las representaciones tridimensionales pueden revelar estructuras más complejas y relaciones no capturadas en dos dimensiones.
- **Perplejidad:** La perplejidad controla la cantidad de información que t-SNE considera en el cálculo de las similitudes locales entre puntos. Variar la perplejidad permite explorar diferentes escalas en la disposición de los datos y puede afectar la separación de las clases.

Visualización y Conclusiones:

Para cada configuración, utilizamos t-SNE para obtener una proyección de menor dimensión de los datos y representamos estas proyecciones en subplots separados. En los gráficos, observamos cómo las clases originales se distribuyen en el espacio de menor dimensión. La paleta de colores utilizada permite identificar fácilmente las clases.

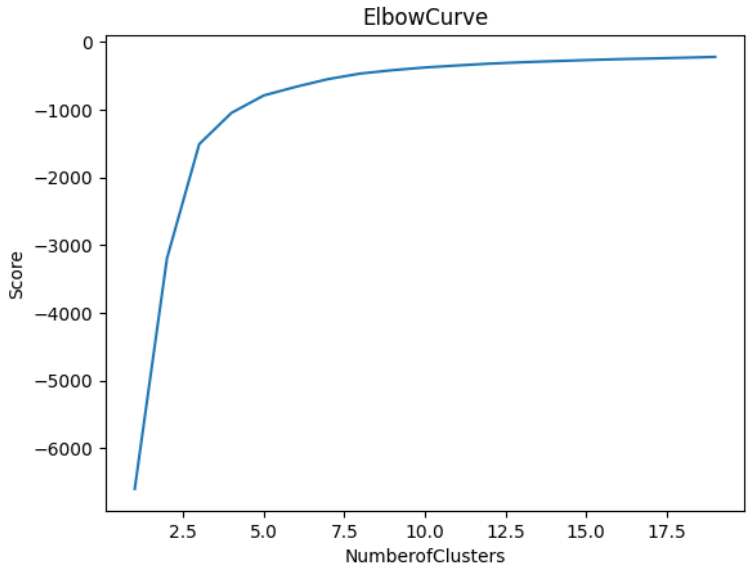
Los resultados sugieren que la elección de hiper parámetros tiene un impacto significativo en la disposición de las clases. La variación de la perplejidad, el número de iteraciones y el número de componentes puede revelar estructuras y relaciones diferentes en los datos.

5. Análisis de K-Means:

Aplicamos el algoritmo K-Means, una técnica de agrupamiento no supervisado, con el objetivo de clasificar los datos en grupos con similitudes entre sí. Este algoritmo divide los datos en clústeres al encontrar centros que minimizan las distancias entre los puntos y estos centros. El número de clústeres se selecciona considerando la estructura inherente de los datos.

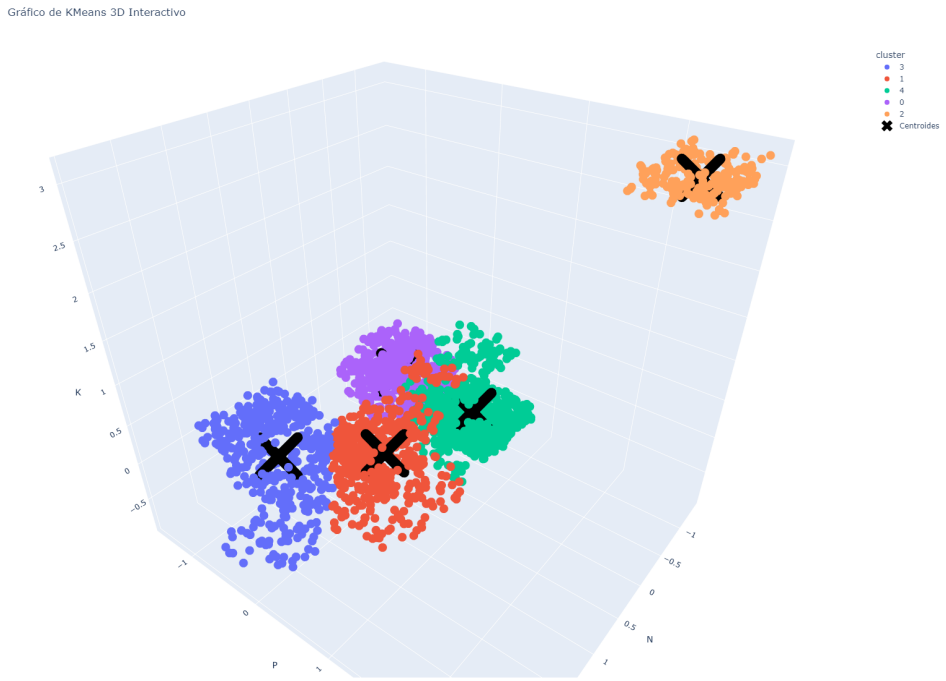
Selección del número óptimo de clústeres:

Comenzamos evaluando diferentes valores de **k** (el número de clústeres) para determinar cuál es el número óptimo. Utilizamos un enfoque de "Elbow Curve" o "Curva del Codo" para visualizar cómo varía la puntuación del algoritmo en función de **k**. La puntuación se calcula como la suma de los cuadrados de las distancias de cada punto a su centroide más cercano. Observamos la gráfica de la curva del codo y buscamos el punto en el que la disminución en la puntuación se estabiliza, lo que nos sugiere el número óptimo de clústeres.



Resultados del agrupamiento:

Después de determinar el número óptimo de clústeres, aplicamos K-Means con ese valor de **k** para agrupar los datos en clústeres. En el gráfico 3D interactivo, podemos observar cómo los puntos de datos se distribuyen en función de las tres características seleccionadas (en este caso, 'N', 'P', 'K'). Cada clúster se representa con un color distinto, lo que nos permite identificar las agrupaciones resultantes.



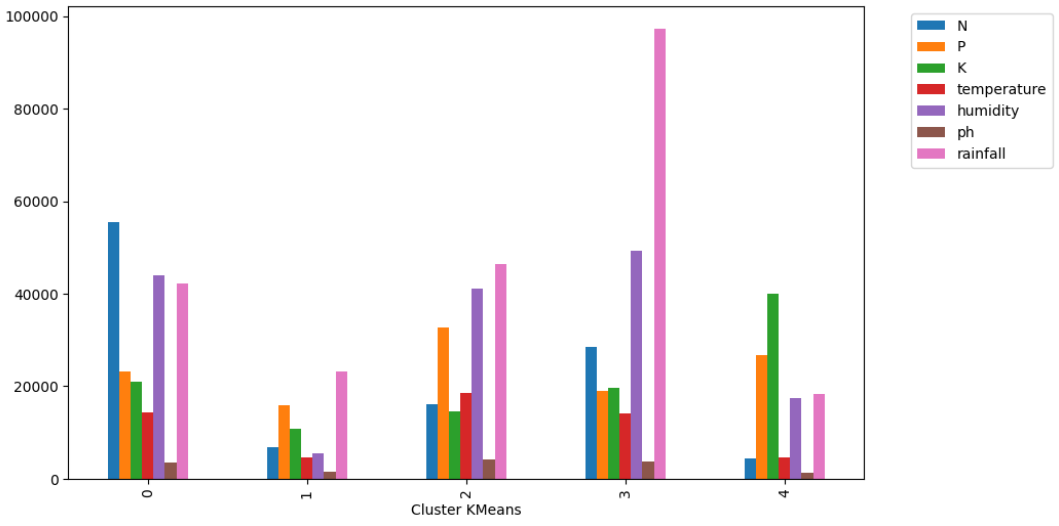
Si bien en el archivo .ipynb podemos analizarlo de forma interactiva, acá adjuntamos captura del mismo.

También representamos los centroides de cada clúster como símbolos de "x" de color negro en el gráfico 3D, lo que facilita la identificación de la ubicación central de cada grupo.

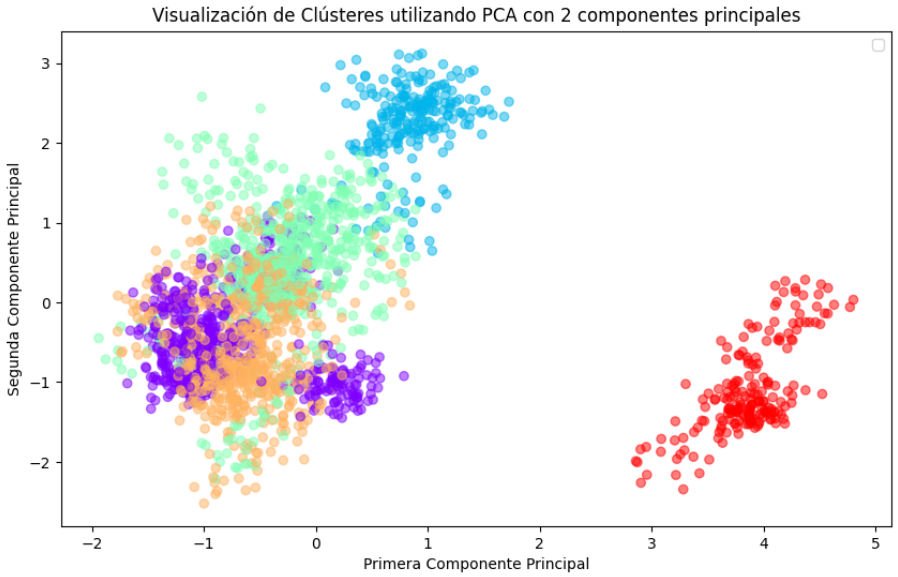
Interpretación de los clústeres:

Para comprender mejor la estructura de los clústeres, aplicamos K-Means a los datos escalados utilizando **`xWheat_scaled`** (`la matriz escalada`). Luego, asignamos las etiquetas de clústeres resultantes a nuestro conjunto de datos original. Podemos observar la distribución de observaciones en cada clúster y visualizarla mediante gráficos de barras

También observamos la distribución de las variables por grupos en el siguiente gráfico:



Además, utilizamos el Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de los datos a tres componentes principales. Esto nos permite visualizar los clústeres en un espacio tridimensional. Los puntos de datos se dispersan en función de las tres componentes principales, y se representan con colores que indican a qué clúster pertenecen.

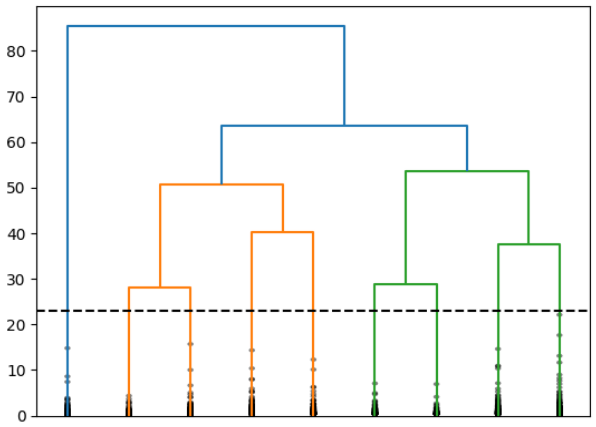


6. Análisis de Clustering Jerárquico:

En este análisis, exploramos la estructura de nuestros datos utilizando el algoritmo de Clustering Jerárquico. Este enfoque de agrupamiento busca crear una jerarquía de clústeres que pueden representar diferentes niveles de similitud en los datos. Empleamos una variedad de técnicas y métricas para evaluar y determinar el número óptimo de clústeres.

Dendrograma para la selección de clústeres:

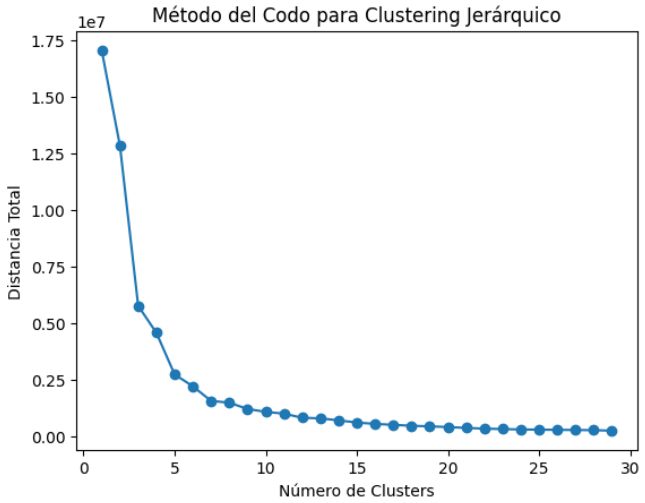
Comenzamos construyendo un dendrograma utilizando la técnica de enlace "ward". Este dendrograma nos proporciona una representación gráfica de cómo los clústeres se fusionan a medida que aumentamos el número de clústeres. Observamos el dendrograma y seleccionamos el número óptimo de clústeres basándonos en la altura de corte en el dendrograma.



Método del Codo y Gap Statistic:

Además del dendrograma, aplicamos el Método del Codo y la Gap Statistic para determinar el número óptimo de clústeres. El Método del Codo implica observar cómo la distancia total entre los puntos de datos y sus centroides disminuyen a medida que aumentamos el número de clústeres. Buscamos el punto en el que esta disminución se estabiliza, lo que nos proporciona una estimación del número óptimo de clústeres.

La Gap Statistic compara la inercia de nuestros datos reales con la inercia de datos aleatorios generados con la misma estructura. Buscamos el número de clústeres que maximiza la diferencia entre estas inercias, lo que indica el número óptimo de clústeres.

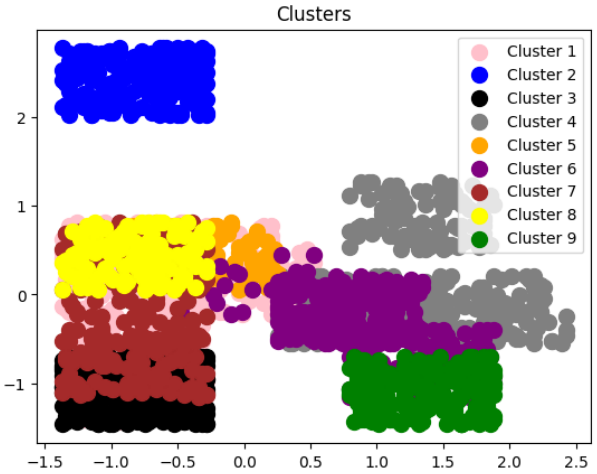


Coeficiente de Silhouette:

También evaluamos la calidad de la agrupación utilizando el Coeficiente de Silhouette. Este coeficiente mide cuán similares son los puntos de un clúster entre sí y cuán diferentes son de los puntos en otros clústeres. En general, el coeficiente de Silhouette varía en el rango de -1 a 1.

El valor obtenido del coeficiente de Silhouette es 0.3037, lo que indica que los clústeres tienen una separación moderada entre ellos, pero no están completamente bien definidos. Esto nos puede sugerir que la estructura de clústeres en el conjunto de datos es razonable, pero puede haber algunas áreas de superposición o ambigüedad entre los clústeres.

Finalmente, aplicamos el algoritmo de Clustering Jerárquico con el número óptimo de clústeres seleccionado. Visualizamos los resultados en un gráfico de dispersión 2D donde cada punto de datos se colorea según el clúster al que pertenece. Esto nos permite observar la estructura de los clústeres y cómo se agrupan los datos.



Conclusiones Generales:

1. El trabajo práctico proporcionó una sólida comprensión de las técnicas de análisis de datos y su aplicación en la exploración y visualización de datos.
2. A través del preprocesamiento y estandarización de datos, logramos preparar el conjunto de datos "Crop_recommendation.csv" para su análisis, lo que resultó fundamental para obtener resultados precisos y significativos.
3. El uso de técnicas como el Análisis de Componentes Principales (PCA), Isomap y t-Distributed Stochastic Neighbor Embedding (t-SNE) permitió una reducción efectiva de la dimensionalidad y proporcionó valiosas perspectivas sobre la estructura de los datos.
4. La aplicación de algoritmos de agrupamiento, como K-Means y Clustering Jerárquico, nos permitió identificar patrones y clústeres dentro de los datos, lo que puede ser útil para tareas de segmentación y análisis de similitud.
5. Los métodos utilizados para determinar el número óptimo de clústeres: Método del Codo y la Gap Statistic
6. Los resultados de las técnicas de visualización, como gráficos 2D y 3D, facilitaron la interpretación de la estructura de los datos y la presentación de hallazgos de manera efectiva.