Informe del Trabajo Práctico de Minería de Datos

1. Introducción

En este trabajo práctico, se realizó un análisis exploratorio de un conjunto de datos relacionado con recomendaciones de cultivos en función de distintas variables como la cantidad de nitrógeno, fósforo, potasio, humedad, temperatura, pH y lluvia. El objetivo principal fue aplicar diversas técnicas de análisis, normalización y reducción de dimensionalidad para extraer patrones significativos y evaluar el comportamiento de las variables.

2. Análisis Exploratorio de los Datos (EDA)

Se inició con una carga y visualización del conjunto de datos. A través de las primeras filas y un análisis de la distribución estadística de las variables mediante `describe()`, se observaron las características principales del dataset, como los valores mínimos, máximos y la desviación estándar de cada variable.

Tipos de datos: También se examinó el tipo de dato de cada variable, confirmando que la mayoría son variables numéricas, a excepción de la variable categórica `Cultivo`, que contiene los tipos de cultivos recomendados.

Distribución de las variables: Para entender mejor la distribución de los datos, se utilizaron diagramas de densidad (`kdeplot`) para variables como `Lluvia`, `Humedad`, `Temperatura`, `pH`, `Potasio`, `Fósforo` y `Nitrógeno`. Esto permitió observar cómo se distribuyen las observaciones dentro de cada una de estas variables.

Identificación de valores atípicos: Además, se aplicó un boxplot general para detectar la presencia de valores atípicos, aunque no se detallaron procedimientos adicionales para su manejo.

3. Normalización de los Datos

Dado que algunas variables mostraban varianzas muy altas, se optó por aplicar una **normalización mediante Z-score**, utilizando la librería `StandardScaler`. Este método estandariza los datos, asegurando que cada variable tenga una media de 0 y una desviación estándar de 1, permitiendo que las variables se comparen en la misma escala, lo cual es esencial antes de aplicar técnicas como el PCA o Isomap.

4. Reducción de Dimensionalidad

4.1. PCA (Análisis de Componentes Principales)

El **PCA** se utilizó para reducir la dimensionalidad del conjunto de datos y encontrar los componentes principales que capturen la mayor varianza en los datos. Los principales hallazgos fueron:

- **Proporción de varianza explicada**: Se seleccionaron 3 componentes principales que capturan el 77.05% de la varianza acumulada del dataset. Esta selección se realizó siguiendo el criterio de varianza acumulada y el criterio de Kaiser, que recomienda mantener los componentes con eigenvalues mayores que 1.
- **Componentes principales**: Se observó que las variables `Potasio` y `Fósforo` mostraron una alta correlación, sugiriendo que podrían tener un impacto significativo en las recomendaciones de cultivo.

4.2. ISOMAP

También se aplicó **Isomap**, un método de reducción de dimensionalidad no lineal, para proyectar los datos en un espacio de 2 componentes. Isomap preserva las distancias geodésicas entre los puntos en lugar de las distancias euclidianas, lo que puede capturar mejor la estructura intrínseca de los datos. La visualización mostró las diferencias entre los cultivos al proyectar los datos en un espacio bidimensional.

5. Visualización en 3D

Utilizando las tres primeras componentes principales obtenidas con PCA, se realizó una visualización en 3D para examinar cómo se distribuyen los cultivos en este espacio de componentes principales. La representación mostró que es posible distinguir algunos cultivos en función de estas tres componentes.

6. Análisis de Correlación

Se generó una matriz de correlación para evaluar las relaciones lineales entre las variables numéricas, lo que permitió identificar variables altamente correlacionadas. Las correlaciones más altas entre variables como `Potasio` y `Fósforo` sugieren que tienen un comportamiento similar dentro del conjunto de datos, lo cual puede ser útil en la toma de decisiones para las recomendaciones de cultivo.

7. Análisis de T-SNE

Finalmente, se planteó el uso de **t-SNE** como una herramienta para la visualización de datos de alta dimensionalidad. Este método es particularmente útil para captar relaciones complejas y patrones no lineales en los datos. Si bien no se incluyó un análisis detallado de t-SNE, se mencionó su potencial para futuros estudios sobre este conjunto de datos.

8. Conclusiones

- El análisis de las variables mostró que existe una alta correlación entre algunas de ellas, especialmente entre `Potasio` y `Fósforo`, lo cual podría implicar una fuerte relación en la recomendación de cultivos.
- La normalización de las variables fue esencial para la correcta aplicación de las técnicas de reducción de dimensionalidad.
- La aplicación de **PCA** y **Isomap** permitió visualizar los datos en espacios de menor dimensionalidad, identificando patrones útiles para la clasificación de cultivos.
- Aunque el gráfico del codo no fue concluyente, se optó por mantener 3 componentes principales, con base en la varianza acumulada y el criterio de Kaiser.

Este trabajo sienta las bases para un análisis más profundo de recomendaciones de cultivos, utilizando tanto técnicas lineales como no lineales para explorar la estructura subyacente de los datos.

Frutas de interes.

Las frutas de interés seleccionadas para este análisis son: Granada, Banana, Mango, Uva, Sandía, Melón, Manzana, Papaya y Coco.

Preprocesamiento de los Datos

1. **Selección de Frutas**: Se extrajo un subconjunto de datos basado en las frutas seleccionadas.
2. **Normalización de las Características**: Las características fueron normalizadas utilizando `StandardScaler` para garantizar que todas las variables se encuentren en la misma escala.

Reducción de Dimensionalidad con PCA

Se aplicó **Análisis de Componentes Principales (PCA)** para reducir la dimensionalidad de los datos a dos componentes principales. Esto permitió

visualizar los patrones principales que existen en los datos:

- Se graficaron las dos primeras componentes principales, mostrando cómo las frutas se agrupan en el nuevo espacio de características.
- El gráfico reveló que, aunque hay cierto solapamiento, las frutas parecen formar grupos definidos.

Clustering con K-Means

1. **Determinación del Número de Clusters**: Se utilizó el **Método del Codo** para determinar el número óptimo de clusters. El gráfico de la inercia mostró un "codo" en **4 clusters**, lo que sugiere que este número de clusters podría ser una buena elección.
2. **Gap Statistic**: También se calculó el **Gap Statistic**, que confirmó que el número óptimo de clusters podría estar entre 4 y 7.
3. **Entrenamiento del Modelo**: Se entrenó un modelo de **K-Means** con 4 clusters, y los datos fueron asignados a cada uno de estos clusters. Los centroides de cada cluster fueron calculados y mostraron las características promedio de los grupos, lo que facilitó la interpretación de las características de cada cluster.
4. **Visualización en 3D**: Se realizó una visualización tridimensional usando las variables más correlacionadas (Nitrógeno, Fósforo y Potasio), lo que permitió observar la separación de los clusters en el espacio 3D.

Clustering Jerárquico

1. **Dendrograma**: Se aplicó clustering jerárquico utilizando el método de enlace de Ward, y se generó un **dendrograma** que mostró cómo los diferentes puntos de datos se agrupan en distintos niveles.
2. **Silhouette Score**: Se utilizó el **Score de Silhouette** para evaluar la calidad de los clusters. El número óptimo de clusters basado en este método fue **4**, ya que proporcionó la mejor separación entre grupos.
3. **Método del Codo para Clustering Jerárquico**: El gráfico del codo mostró que entre **4 y 5 clusters** sería una buena elección, lo que concuerda con los resultados obtenidos en K-Means y Silhouette Score.
4. **Gap Statistic para Clustering Jerárquico**: Al aplicar el **Gap Statistic** al clustering jerárquico, se encontró que el número óptimo de clusters podría

estar entre **4 y 10**, reforzando la idea de que 4 clusters es una buena elección.

Conclusiones

- El análisis sugiere que **4 clusters** es una buena opción para capturar la variabilidad en los datos, tanto en el modelo de K-Means como en el clustering jerárquico.
- Los clusters identificados presentan patrones distintivos en los niveles de nutrientes, especialmente en las variables Nitrógeno, Fósforo y Potasio.
- Sin embargo, el número exacto de clusters óptimos no es definitivo, y podría estar entre **4 y 10**, dependiendo del método de evaluación utilizado.

Este trabajo demuestra la utilidad de combinar técnicas de reducción de dimensionalidad y agrupamiento para descubrir patrones y características clave en conjuntos de datos complejos.