Valeria María Serna Salazar - A01284960
Tecnológico de Monterrey

May 4, 2024

EXAMEN PRÁCTICO — MA2007B

1. Planteamiento del Problema

El problema está enfocado en el análisis y la agrupación de un conjunto de datos de vinos. Para esto, se utilizarán diferentes técnicas de análisis y clustering con el objetivo de encontrar patrones significativos en los datos y poder juzgar la calidad de los vinos de cada cluster. El objetivo final es ver si se pueden rankear de mayor a menor calidad de vino de acuerdo a sus características químicas.

2. Preguntas de Investigación

- ¿Es posible encontrar grupos o clusters significativos de vinos basados en sus características químicas?
- ¿Es posible juzgar la calidad del vino de acuerdo al cluster al que pertenece?
- ¿Se pueden encontrar qué tipos de vino formarán parte de cada cluster que se realice?

3. Metodología

1. Preparación de los datos
2. Se importan los datos y se analiza su calidad (valores nulos, tipos de datos).
3. Se normalizan los datos para que las características tengan igual importancia en el análisis.
4. Análisis de Componentes Principales (PCA): Se aplica PCA para reducir la dimensionalidad de los datos y visualizar la distribución general.
5. KMeans Clustering: Se aplica KMeans para encontrar clusters de datos y se evalúa la calidad de los clusters obtenidos.
6. Dendrogramas: Se utilizan dendrogramas para entender las relaciones jerárquicas entre los datos y evaluar posibles agrupamientos.
7. Mapper y Análisis Topológico (TDA): Mapper se utiliza para encontrar patrones en los datos usando análisis topológico.
8. Se grafican los clusters y se juzgan los resultados.
9. Rankear de mayor a menor calidad de acuerdo al cluster al que pertenece el vino.

4. Dataset

Los datos utilizados durante este análisis fueron obtenidos de la base de datos UCI, específicamente del conjunto de datos de vinos italianos, que incluía una muestra de 178 elementos. Estos, aunque provienen de la misma región, son diferentes variedades. Cada variable contenía información derivada de un análisis químico, se tienen 13 variables numéricas que serán analizadas más adelante.

Para comprender a fondo el contexto del problema, es necesario indagar más a fondo sobre las variables que están presentes en la base de datos:

1. **Alcohol:** El contenido de alcohol afecta el cuerpo y el sabor del vino. Un vino con un alto contenido de alcohol puede resultar ardiente y desequilibrado si no tiene suficiente estructura, mientras que un vino con poco alcohol puede carecer de cuerpo.

2. **Ácido Málico:** Contribuye a la acidez del vino, similar a la de las manzanas verdes. Un exceso puede hacer que el vino sea demasiado ácido, mientras que una cantidad insuficiente puede hacer que sea plano.

3. **Ceniza:** Se refiere a los minerales presentes en el vino. Un contenido adecuado de ceniza puede mejorar la estructura y la complejidad del vino.

4. **Alcalinidad de la ceniza:** Indica la cantidad de compuestos alcalinos en el vino, lo que puede afectar el pH. Un nivel equilibrado es esencial para la estabilidad y longevidad del vino.

5. **Magnesio:** Es un nutriente importante para las uvas durante el crecimiento. Afecta la fermentación y el metabolismo de las levaduras. Un exceso puede dar sabores amargos, y una deficiencia puede reducir el rendimiento de la fermentación.

6. **Fenoles Totales:** Los fenoles son compuestos que afectan el color, el sabor y la estructura del vino. Un nivel adecuado es crucial para la longevidad y el sabor.

7. **Flavonoides:** Son un subgrupo de fenoles que afectan el color y la astringencia del vino. Los niveles altos pueden aportar un color intenso y estructura, mientras que los niveles bajos pueden resultar en vinos menos estructurados.

8. **Fenoles no flavonoides:** Aportan características como la oxidación y la acidez. Afectan el envejecimiento del vino y su estabilidad.

9. **Proantocianidinas:** Son taninos que contribuyen a la estructura y el envejecimiento del vino. Una cantidad alta puede resultar en vinos demasiado tánicos, mientras que cantidades bajas pueden hacer que el vino carezca de estructura.

10. **Intensidad del Color:** Es un indicador de la concentración de antocianinas y compuestos fenólicos. Un color intenso generalmente indica un vino joven y concentrado.

11. **Matiz:** Refleja el tipo de antocianinas presentes. Los tonos más jóvenes son púrpuras, mientras que los tonos más envejecidos tienden hacia el marrón.

12. **OD280/OD315:** Este índice mide el contenido de fenoles en el vino, lo que da una idea de su estructura y complejidad.

13. **Prolina:** Es un aminoácido que sirve como indicador del contenido de nitrógeno. Niveles altos pueden estar relacionados con la madurez y el contenido de azúcar en las uvas, afectando el sabor y el cuerpo del vino.

5. Procesamiento de Datos

El proceso para limpiar los datos fue bastante sencillo, pues se revisaron los valores nulos y no se encontró ninguno, además de que los rangos de valores eran adecuados, no habían outliers, todos eran de tipo numérico (float a excepción del magnesio que era int) y en general los valores numéricos eran correctos.

Al hacer un df.describe() fue posible observar que los rangos de valores entre las variables era muy diferente, pues algunos valores eran pequeños mientras que otros eran muy grandes, por lo que se optó por normalizar la base de datos al utilizar la librería sklearn y su clase StandardScaler. Esta, se usa para transformar los datos y lograr que tengan una media de 0 y una desviación estándar de 1. En otras palabras, se escalan los datos para que sigan una distribución normal estándar al aplicar la fórmula $X_{escalada} = \frac{X-\mu}{\sigma}$.

Después de la normalización, se aplicó un Análisis de Componentes Principales (PCA) para reducir la dimensionalidad del conjunto de datos. PCA ayuda a identificar los componentes principales que capturan la mayor parte de la variabilidad en los datos, reduciendo el número de características sin sacrificar demasiada información. En este caso, se redujeron los datos a tres componentes principales (de acuerdo a un Scree Plot que reveló que sería la cantidad correcta), logrando capturar el 67 % de la varianza, lo que es una porción significativa. Esto no solo simplifica la visualización, sino que también mejora la eficiencia de los algoritmos de clustering al permitirles trabajar con un espacio dimensional reducido.

6. Agrupamiento

El primer enfoque de agrupamiento utilizado fue un análisis jerárquico, cuyo propósito principal fue proporcionar una comprensión preliminar de la estructura de los datos mediante un dendrograma. Al hacer un Single Linkage Dendogram, se logró obtener una representación visual de las distancias entre los datos y así se logró empezar a explorar y a sugerir el número probable de clusters. Sin embargo, no proporciona una solución definitiva, sino más bien una orientación sobre cuántos clusters podrían ser razonables. Como es posible ver en la imagen 1, se empezaron a considerar 3 clusters.

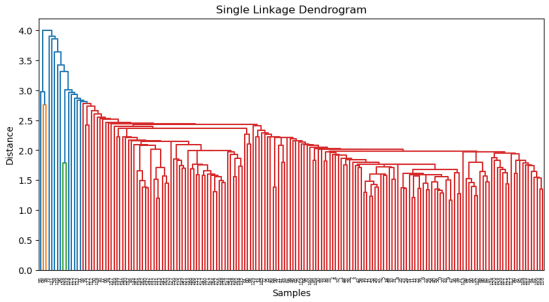

Figura 1: Dendograma

El siguiente paso fue aplicar un complejo de Rips para profundizar en la topología de los datos y comprender mejor su estructura. Se generó un diagrama de persistencia que mostró las características de los datos, revelando algunas estructuras persistentes que indican características significativas para el clustering que se busca hacer más adelante. En específico, en la figura 2 hay un 2 puntos rojos de H0 que representan características persistentes, mismos que podrían ser utilizadas más adelante.

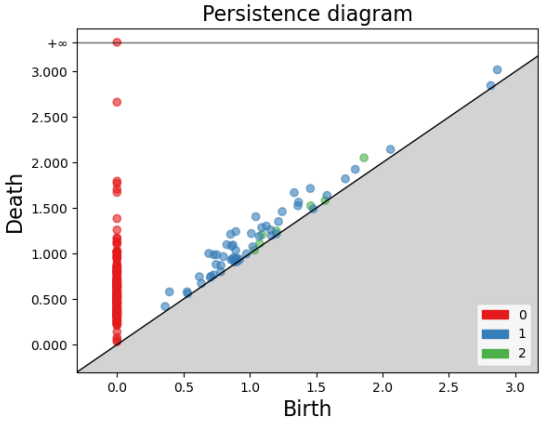

Figura 2: Persistence Diagram

7. Mapper / Visualización

KeplerMapper, conocido como kmapper, permite encontrar patrones y estructuras en conjuntos de datos complejos utilizando técnicas de topología algebraica. El enfoque de Mapper se basa en la idea de proyectar los datos en un espacio de menor dimensión, dividir ese espacio en particiones y luego identificar la conectividad entre los datos particionados para crear un grafo que represente la estructura topológica de los datos.

En el código se hace una partición del espacio de datos en 10 cubos con un 20 % de superposición, permitiendo la creación de regiones superpuestas para mejorar el descubrimiento de patrones. Después, se usa Kmeans para agrupar los datos en tres clusters diferentes, ese número fue encontrado anteriormente con el dendograma y que se consideró óptimo. Se genera la visualización en HTML y el resultado se muestra en la figura 3.

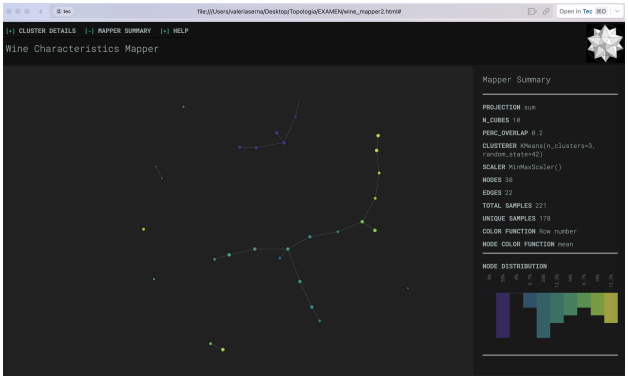

Figura 3: Mapper

Para continuar con el análisis, se enumeraron los clusters para hacer un análisis más profundo de lo que contiene cada uno. Cabe aclarar que los puntos solos no fueron utilizados, pues son outliers que no entraron en ningun cluster y, por lo tanto, no servirán para el análisis que se busca hacer. Al final, los clusters que se tomarán en cuenta son 4:

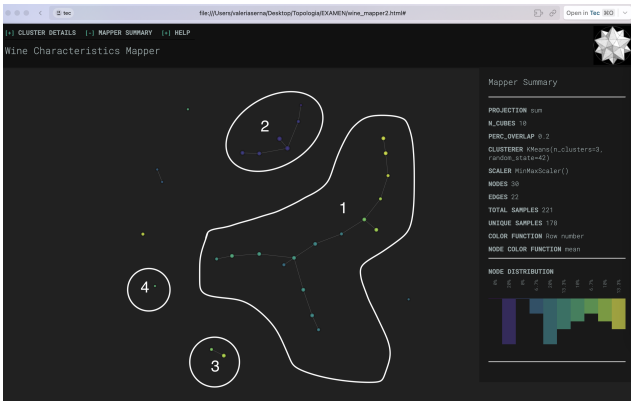

Figura 4: Mapper con Clusters Marcados

El siguiente paso fue hacer un dataframe con cada cluster seleccionado, obteniendo 4 que contenían diferentes agruapciones de los datos. Se obtuvo el promedio del valor de cada variable por cluster, para ver la manera en que estaban agrupados los datos y obtener de una manera más clara la importancia de las variables en cada grupo.

Con esto en mente, es posible empezar a ver como funcionan los clusters. Por ejemplo, el cluster 4 tiene, en promedio, valores bajos de Alcohol pero altos de matiz, el cluster 3 tiene valores altos de alcalinidad de la

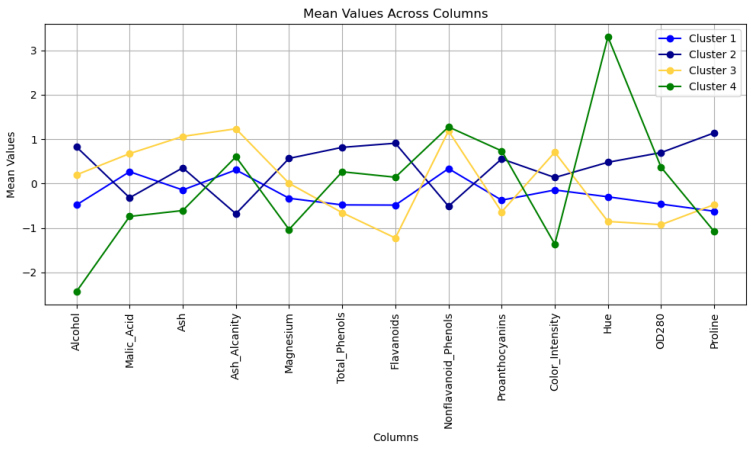

Figura 5: Promedio de Variable por Cluster

ceniza pero bajos de flavonoides, y así sucesivamente cada cluster se destaca en tener algunas variables altas y otras bajas. Entonces, ¿qué indican los clusters de la calidad del vino y por qué es relevante ver la manera en que están agrupados?

En términos generales, el conocer realmente que características conforma cada cluster es esencial para poder analizar los tipos de vino que lo constituyen y, específicamente en este caso, se realizó una investigación del tema y, de acuerdo a distintos artículos: Identification of red wine categories based on physicochemical properties [Bai et al., 2019], Physicochemical Properties Importance For Type Classification of Wines Using Machine Learning Techniques [Mor et al., 2022], Selection of important features and predicting wine quality using machine learning techniques [Gupta, 2018], se llegó a las siguientes conclusiones:

Clasificación de los Grupos de Vino

Cluster 1

Este grupo tiene un perfil equilibrado en cuanto a fenoles, flavonoides y OD280. Su intensidad de color es alta, pero el matiz más bajo indica cierta inestabilidad en el color. Esto, junto con el contenido fenólico moderado, lo coloca en una posición intermedia.

Cluster 2

Este grupo presenta el mayor contenido de fenoles y flavonoides, así como un alto OD280, lo que indica un vino con un perfil sólido y un buen potencial de envejecimiento. Su bajo ácido málico aporta menos acidez y el alto contenido de proantocianidinas mejora la estructura del vino, lo que lo hace mejor al paladar. Aunque la intensidad del color es baja, el equilibrio en su perfil general lo destaca.

Cluster 3

Aunque este grupo tiene un perfil equilibrado en cuanto a minerales y proantocianidinas, su contenido de fenoles y OD280 es el más bajo, lo que limita su potencial de envejecimiento y complejidad. El alto contenido de ácido málico puede dar una mayor acidez, afectando así el equilibrio del vino.

Cluster 4

Con un alto contenido de fenoles, flavonoides y OD280, este grupo tiene tambien un buen balance y un buen potencial antioxidante. Aunque tiene un menor contenido de alcohol (lo que lo hace un poco peor al cluster 2), su alto matiz sugiere estabilidad en el color, indicando su potencial para envejecer bien.

Clasificación Final

Organizados de mayor calidad a menor calidad de acuerdo a la literatura consultada y los valores de las variables, la clasificación resulta en:

1. Cluster 2: Mayor contenido fenólico, fuerte equilibrio general.
2. Cluster 4: Alto contenido fenólico y buen potencial de envejecimiento.
3. Cluster 1: Buen equilibrio, pero menor calidad que los anteriores.
4. Cluster 3: Limitado en contenido fenólico, con alta acidez.

Aunque es imposible saber qué tipo de vino específico se encuentra en cada cluster, pues la base de datos no menciona sus nombres como tal, de acuerdo a las características obtenidas de cada grupo se pueden hacer algunas conjeturas al respecto:

- El cluster 2 tiene niveles altos de fenoles totales, flavonoides y OD280, así como proantocianianidinas elevadas, esto sugiere un vino con estructura robusta y con un alto potencial de envejecimiento. Algunos vinos potenciales que podrían encontrarse adentro de estre gupo es el Cabarnet Sauvignon y el Syrah/Shiraz, ambos conocidos por sus sabores intensos, alto contenido fenólico y su potencial de envejecimiento.
- El cluster 4 tiene un alto contenido de fenoles totales, flavonoides y OD280, pero es una intensidad menor al anterior, por lo que se buscan sabores más ligeros, como el Merlot y el Sangiovese, que tienen frutalidad.
- El cluster 1 es moderado en fenoles totales, flavonoides y OD280, por lo que indica un perfil equilibrado pero aun menos intenso que los anterior, por lo tanto se podría encontrar un Pinot Noir o un Tempranillo, que son vinos mucho más ligeros.
- Por último, se encuentran vinos que no están tan equilibrados, con alta acidez y con menos potencial de envejecimiento, entre ellos podrian estar el Riesling y el Pinot Grigio.

Aunque los posibles vinos son conjeturas, es posible ver que la forma de los datos y su agrupación muestran un claro patrón, que lo hace sencillo de analizar y ver la manera en que podría utilizarse en aplicaciones prácticas. Esta agrupación permite identificar las características que definen la calidad y el estilo de los vinos, ofreciendo una valiosa herramienta para la clasificación y selección de vinos basada en datos objetivos. A través de técnicas como el Análisis de Componentes Principales y el mapeo topológico, es posible visualizar y entender la estructura subyacente de los datos, lo que facilita la comprensión de las características que influyen en la clasificación del vino.

Referencias

[Bai et al., 2019] Bai, X., Wang, L., and Li, H. (2019). Identification of red wine categories based on physicochemical properties. In *5th International Conference on Education Technology, Management and Humanities Science*, volume 19, page 2019.

[Gupta, 2018] Gupta, Y. (2018). Selection of important features and predicting wine quality using machine learning techniques. *Procedia Computer Science*, 125:305–312.

[Mor et al., 2022] Mor, N. S., Asras, T., Gal, E., Demasia, T., Tarab, E., Ezekiel, N., Nikapros, O., Semimufar, O., Gladky, E., Karpenko, M., et al. (2022). Physicochemical properties importance for type classification of wines using machine learning techniques. *agriRxiv*, (2022):20220051480.

Submitted by Valeria María Serna Salazar - A01284960 on 4 de mayo de 2024.