Instituto Tecnológico y de Estudios Superiores de Monterrey

Escuela de Ingeniería y Ciencias

Ingeniería en Ciencias de Datos y Matemáticas

Agrupación de Vinos por Componentes Químicas

Reporte Examen Práctico

Uso de geometría y topología para ciencia de datos (Gpo 602) - MA2007B.602

Annette Pamela Ruiz Abreu A01423595

Profesor:

Dr. Alejandro Ucan-Puc

Monterrey, Nuevo León

03 de mayo de 2024

1. Introducción

La industria vinícola es una de las más importantes en el mundo, sin embargo, la calidad de un vino es subjetiva y depende de los gustos de cada persona. Existen diferentes tipos de vinos, entre ellos se encuentran los vinos tintos y los vinos blancos, los cuales tienen diferentes características que los hacen únicos. Algunas de las características que se buscan en un vino son: acidez, azúcar, cloruros, sulfatos, entre otros componentes químicos. Decidir la calidad de un vino es una tarea complicada, ya que depende de la percepción de cada persona, sin embargo, podemos tentar a resolver esta problemática al estudiar cúmulos de vinos con características similares.

La problemática es encontrar cúmulos de vinos con características similares y determinar si existe una relación entre las características de los vinos y su calidad.

Para poder resolver esto, se plantearon las siguientes preguntas de investigación que se intentaron resolver:

1. ¿Cuántos puntos conexos podemos identificar para darnos una idea de la cantidad de clústeres que hay?
2. ¿Qué características químicas son las que más influyen en el agrupamiento del vino?
3. ¿Qué agrupaciones naturales de vinos pueden identificarse basándose en sus características químicas?
4. ¿Cómo podemos nombrar o distinguir las agrupaciones? ¿Qué tipos de vinos están en la base de datos?
5. Teniendo las agrupaciones, ¿podemos determinar qué agrupación es "mejor.en términos económicos o de popularidad?
6. ¿Qué características químicas muestran la mayor variabilidad entre los diferentes grupos de vinos identificados?
7. ¿Existen combinaciones particulares de características químicas que tiendan a co-ocurrir en los vinos?

2. Metodología

Para resolver esta problemática, se empezó con la preparación de datos en donde se revisaron los datos nulos (no había), los tipos de datos (numéricos), si había datos duplicados y datos atípicos, se normalizaron para facilitar el proceso de análisis y finalmente se proyectaron los datos usando isomap. Isomap es una técnica de reducción de dimensionalidad no lineal que preserva las distancias geodésicas entre puntos en un espacio de alta dimensionalidad. Utiliza el enfoque de vecinos más cercanos para construir un grafo de vecindad y luego calcula las distancias geodésicas en este grafo para obtener una representación de baja dimensionalidad que conserve la estructura subyacente de los datos en el espacio original. Esta proyección facilita la visualización y el análisis de los datos en un espacio de menor dimensión, lo que puede ayudar a identificar patrones y relaciones entre las variables de interés.

En segundo lugar, se hizo un complejo de Rips para ver cuántos puntos conexos se podían identificar para dar una idea de la cantidad de clústeres que se pueden formar con los datos. Además, se usaron las visualizaciones de los complejos de Rips para determinar qué combinación de características químicas era la mejor

para poder diferenciar los tipos de vino de la mejor forma. Los complejos de Rips fueron complementados por visualizaciones de homología persistente para poder identificar la cantidad de componentes conexos que persistían en los datos. Para validar la cantidad de clústeres que se podían formar, se usó linkage clustering y PCA. Este método agrupa los puntos de datos según la distancia entre ellos, utilizando técnicas como el enlace simple, completo y promedio. Se comienza con cada punto de datos como un clúster individual y se fusionan gradualmente los clústeres más cercanos. Se realizó un análisis de agrupamiento jerárquico sobre datos escalados y reducidos a tres dimensiones mediante análisis de componentes principales (PCA).

Finalmente, se realizaron diversos mappers con métodos de agrupamiento como Kmeans y DBSCAN para formar los clústeres adecuados, visualizarlos en el espacio y poder analizar cada clúster y concluir sobre los tipos de vinos.

3. Resultados

Al hacer el complejo de Rips de la Figura 1 con los datos transformados utilizando isomap, es evidente que hay 3 componentes conexas diferentes, indicando que hay tres clústeres distintos en los datos. También se observa que hay dos clústeres que están mucho más cerca uno del otro que el tercero y pueden llegar a conectarse dependiendo del épsilon que se elija para graficar. En las visualizaciones de las Figuras 2 y 3 podemos ver que hay dos componentes conexas que persisten hasta 0.4 y tres en 0.3.

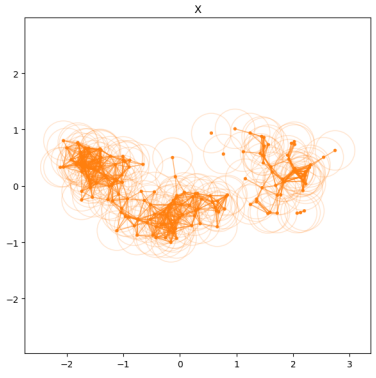

Figura 1: Complejo de Rips de los datos proyectados con Isomap

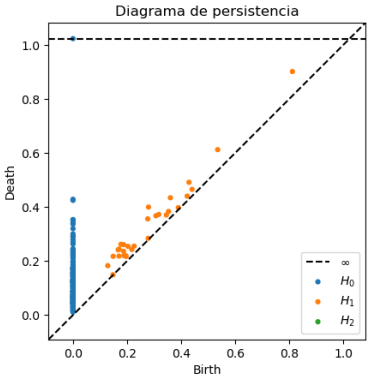

Figura 2: Diagrama de persistencia

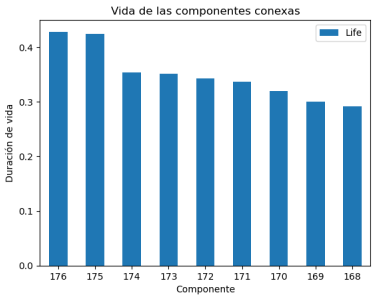

Figura 3: Barplot de persistencia de componentes conexas

Al hacer varias pruebas de complejos de Rips con diferentes columnas y valores de epsilon, se encontró que los mejores resultados se obtienen usando las siguientes columnas:

- Alcohol
- Malic Acid
- Flavanoids
- OD280
- Ash Alcanity
- Hue
- Magnesium
- Color Intensity
- Proline

Esto significa que estas columnas son las que más influyen en el agrupamiento de los vinos. Al hacer el complejo de Rips con estas columnas, se observa en la Figura 4 que hay 3 componentes conexas distintas, lo que nos indica que hay 3 clústeres diferentes en los datos.

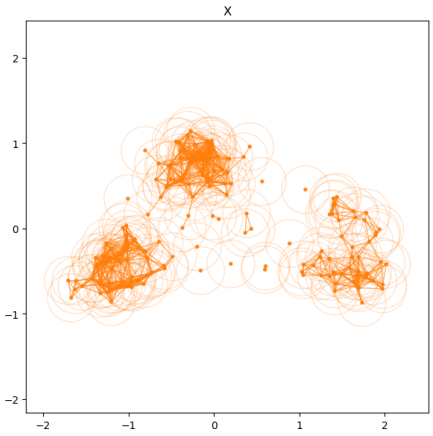

Figura 4: Complejo de Rips solo con los componentes químicos más influentes

Los datos que empeoraban las visualizaciones y hacían que la distinción de grupos no fuera clara eran:

- Nonflavanoid Phenols
- Total Phenols
- Ash
- Proanthocyanins

Esta cantidad de clústeres se confirmó con la metodología de linkage clustering y PCA. Al realizar este agrupamiento, se observa que cada método devuelve resultados de agrupaciones un poco distintas debido a que cada uno tiene un criterio diferente para unir los clústeres. Sin embargo, los tres métodos confirman la hipótesis inicial de que hay entre dos y tres clústeres, en donde dos clústeres están cercanos y pueden llegar a unirse (como se muestra en el Complete Linkage). Nuevamente, podemos concluir que hay 3 clusters distintos en los datos.

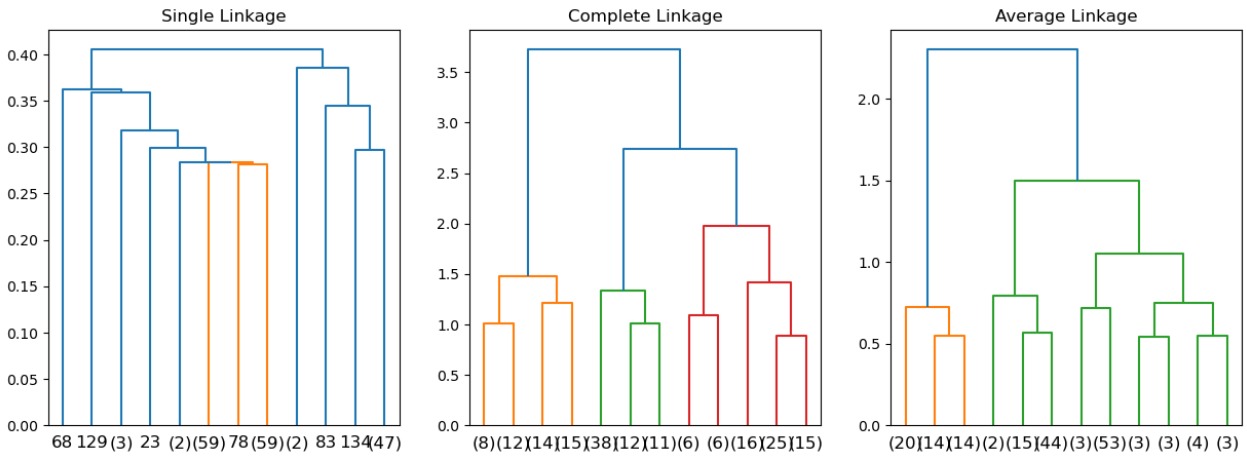

Figura 5: Linkage Clustering

Número de clústeres encontrados en Single Linkage: 2

Número de clústeres encontrados en Complete Linkage: 3

Número de clústeres encontrados en Average Linkage: 3

Los clústeres se pueden visualizar en un mapa bidimensional de factores usando Stochastic Neighbor Embedding como se ve en la Figura 6, o prediciendo los resultados, asignando un color a cada clúster y visualizando los datos originales en formato 2D como se ve en la Figura 7.

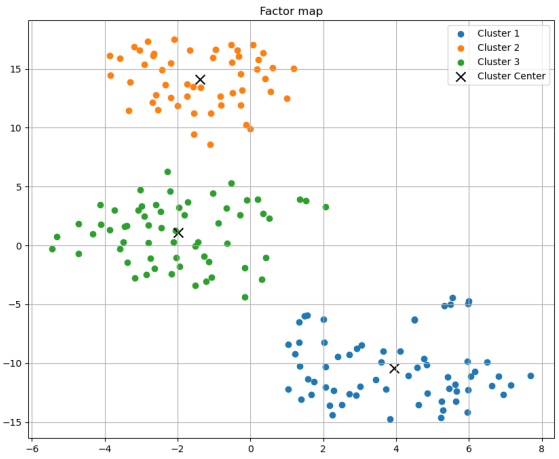

Figura 6: Mapa bidimensional de factores

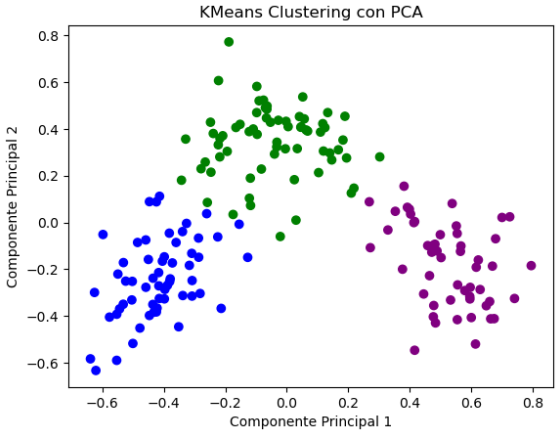

Figura 7: Datos proyectados con PCA y con color por clúster

Finalmente, se usó la librería Kmapper para realizar el mapeo de los datos usando Kmeans y DBSCAN, los cuales devolvieron la misma cantidad de clústeres con medias parecidas. Para clasificar los nodos por colores se usó la suma de alcohol, flavanoids, OD280 y proline. Estas cuatro columnas son las que mejor diferencian cada tipo de vino.

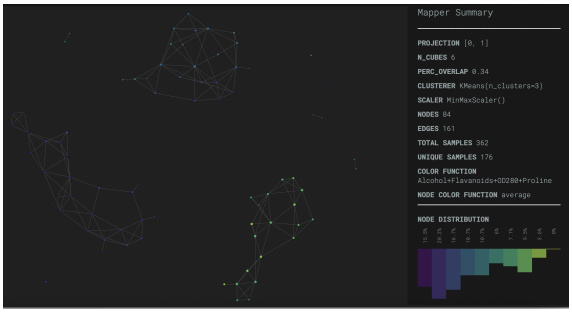

Figura 8: Mapper con kmeans

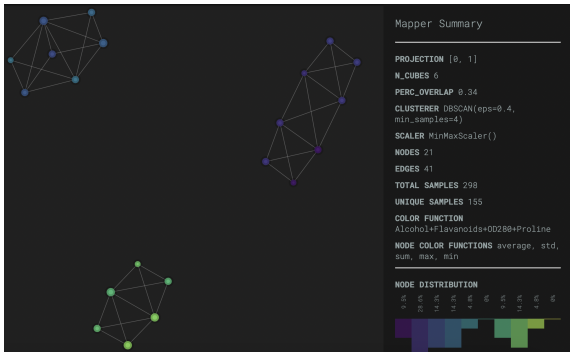

Figura 9: Mapper con DBSCAN

Ambos mappers devolvieron los siguientes resultados de los clústeres que se ven en la Figura 10.

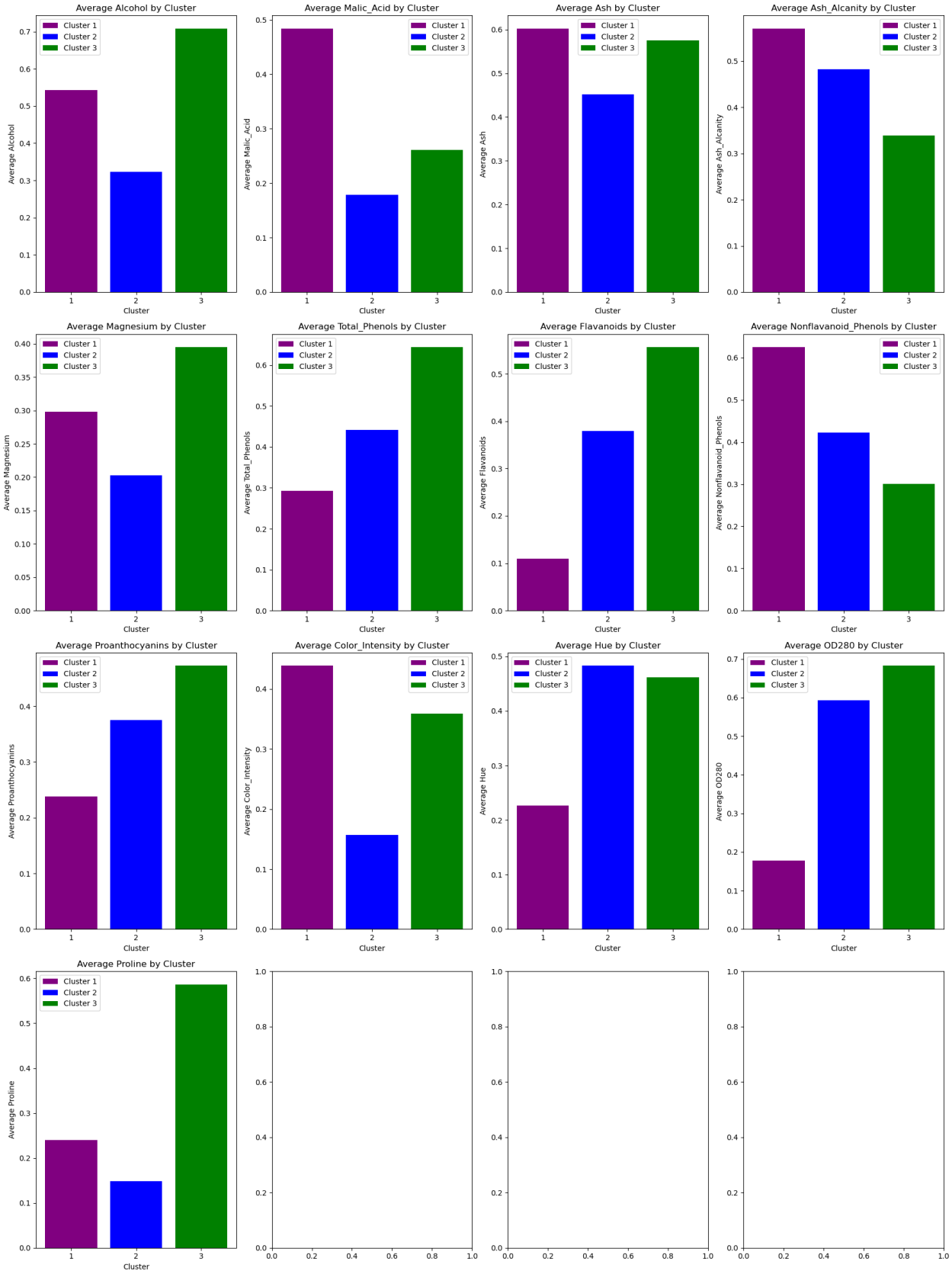

Figura 10: Características de los Clústeres de Vinos

Con base en las gráficas anteriores y una investigación realizada para las características de los vinos se concluye lo siguiente:

- El clúster 1 se caracteriza por tener un nivel medio de alcohol y un alto contenido de ácido málico. Además, muestra niveles bajos de total phenols, flavonoids y proantocianidinas. Estas características sugieren una posible composición de vinos blancos de clima frío, como Riesling o Sauvignon Blanc, que tienden a tener una acidez más alta y un perfil más ligero en compuestos fenólicos. Aunque el nivel medio de alcohol y la intensidad de color podría incluir una variedad de vinos tintos jóvenes y frescos (Acenología 2024).
- Dado que el clúster 2 exhibe características como bajo contenido de alcohol, malic acid y proantocianidinas, y niveles moderados de total phenols y flavonoids, junto con una baja intensidad de color y un bajo nivel de prolina, es probable que esté compuesto principalmente por vinos blancos. Este perfil sugiere vinos de cuerpo ligero con niveles más bajos de taninos y compuestos fenólicos asociados típicamente con vinos blancos como Pinot Noir, Riesling u otros blancos de características similares. La baja intensidad de color y la ausencia de prolina indican un perfil sensorial más fresco y menos astringente, lo que concuerda con las características típicas de muchos vinos blancos (FernÃ¡ndez et al. 2009).
- El clúster 3 está compuesto mayormente por vinos tintos como el Cabernet Sauvignon, el Merlot y el Syrah por las siguientes razones: Tiene vinos con un alto contenido de alcohol, lo que les confiere un cuerpo completo y robusto (TTB s. f.). Además, presentan niveles bajos de ácido málico, lo que resulta en una menor acidez, típica de vinos tintos de climas cálidos. Con un contenido alto de fenoles totales, estos incluyen vinos robustos y fuertes. Los flavonoides se presentan en niveles altos o medios, lo que añade intensidad de color y sabor a los vinos tintos audaces como el Cabernet Sauvignon o Syrah, así como a variedades más suaves como el Merlot y Zinfandel. Las proantocianidinas se encuentran en un nivel medio, común en muchos vinos tintos de este grupo. Tiene L-prolina alta, lo cual aumenta la dulzura, la viscosidad y el sabor a frutos rojos, y disminuye el amargor y la astringencia (Infowine s.f.).

La calidad del vino es algo subjetivo; sin embargo, con base en los vinos que se creen que pertenecen a cada grupo y las características que exhibe cada grupo, se puede decir que el cluster 3 probablemente tiene la mejor calidad, puesto que probablemente está compuesto de Cabernet Sauvignon, Syrah y Merlot (Wine Folly s. f.). Estos son de los vinos más populares y son vinos tintos robustos y envejecidos. Las características que se necesitan para asegurar la calidad del vino son (los que tienen un asterisco son los más importantes):

- Alcohol alto *
- Malic acid bajo
- Ash alto
- Ash alcanity bajo
- Magnesium medio
- Total phenols alto
- Flavanoids medio *
- Nonflavanoid bajo
- Proanthocyanins medio
- Color intensity medio
- Hue medio
- OD280 alto *
- Proline alto *

Para visualizar los procesos y más, se puede consultar el siguiente enlace que contiene el repositorio del proyecto.

Bibliografía

Acenología (feb. de 2024). *La química del color del vino*. URL: https://www.acenologia.com/quimica_color_vino_cienc1213/.

FernÃ¡ndez, V et al. (sep. de 2009). "CaracterizaciÃquÃmica y contenido mineral en vinos comerciales venezolanos". es. En: *Revista de la Facultad de AgronomÃa* 26, págs. 382-397. ISSN: 0378-7818. URL: http://ve.scielo.org/scielo.php?script=sci_arttext&pid=S0378-78182009000300005&nrm=iso.

Infowine (s.f.). *Importancia de los aminoácidos en el gusto del vino tinto*. URL: https://www.infowine.com/es/noticias/importancia_de_los_aminoacidos_en_el_gusto_del_vino_tinto_sc_20836.htm.

TTB (s. f.). *Wine Labeling: Alcohol Content*. URL: https://www.ttb.gov/regulated-commodities/beverage-alcohol/wine/labeling-wine/wine-labeling-alcohol-content.

Wine Folly (s. f.). *The 10 Most Popular Wines in the World*. URL: https://winefolly.com/deep-dive/the-10-most-popular-wines-in-the-world/.