Instituto Tecnológico y de Estudios Superiores de Monterrey

Uso de geometría y topología para ciencia de datos

Grupo 602

3 de mayo de 2024

EA2: Practico

Ana Paola Almeida Pérez A00833937

1 Planteamiento

En este proyecto se utiliza la base de datos "wine-clustering" que se puede encontrar en la plataforma para ciencia de datos "Kaggle". Esta base recolecta información de 178 muestras de diferentes tipos de vino (no se conoce la etiqueta), estos datos registran 13 características de los mismos: Alcohol, Malic Acid, Ash, Ash Alcanity, Magnesium, Total Phenols, Flavanoids, Nonflavanoid Phenols, Proanthocyanins, Color Intensity, Hue, OD280 y Proline que son propiedades importantes para inferir a partir de ellas cuestiones más específicas que nos puedan interesar.

Bajo este contexto, se busca generar una clusterización que nos permita encontrar aquellas propiedades relevantes para agrupar los registros y generar análisis que nos permitan relacionar esto con la calidad de los vinos.

Las preguntas de investigación de interés en este proyecto son:

- ¿De acuerdo con la clusterización, cuántos tipos de vinos hay en la base?
- ¿Qué características son importantes en la calidad del vino?
- ¿Qué tipo de relación hay entre las características según el tipo de vino?

2 Metodología

2.1 Exploración de los datos

- **Conocer la calidad de los datos registrados**: Se realizará una inspección inicial de los datos para comprender su calidad y fiabilidad. Esto implica verificar la integridad de los datos, identificar posibles errores o inconsistencias, y asegurarse de que los datos estén completos y correctamente etiquetados.
- **Identificar características de los datos que nos permitan elegir el modo de tratar los datos posteriormente (cantidad de datos, etc)**: Se analizará la distribución de los datos, la cantidad de muestras disponibles y la distribución de las características. Esto ayudará a determinar el enfoque adecuado para el análisis posterior y a seleccionar las técnicas más apropiadas para el procesamiento de datos.

2.2 Análisis estadístico

- **Conocer la normalidad y correlaciones de los datos**: Se realizará un análisis estadístico para comprender las relaciones entre las diferentes características. Esto proporcionará información sobre la normalidad de los datos y las posibles correlaciones entre las variables, lo que ayudará en la selección de técnicas de análisis adecuadas.
- **Identificar valores atípicos**: Se llevará a cabo una identificación de valores atípicos para detectar posibles datos anómalos que puedan afectar el análisis. Esto permitirá tomar decisiones informadas sobre cómo manejar estos valores atípicos durante el procesamiento de datos.

2.3 Preparación de los datos

- **Normalización de los datos**: Esto se realizará para garantizar que todas las características tengan la misma escala y contribuyan de manera equitativa al análisis. Esto ayudará a mejorar el rendimiento de los algoritmos de aprendizaje automático y a evitar sesgos en los resultados.
- **Reducción de la dimensión y el uso de PCA**: Uso de PCA (Análisis de Componentes Principales). Esto se basará en la necesidad de manejar eficientemente conjuntos de datos de alta dimensionalidad, mejorar la interpretación de los resultados y reducir el riesgo de sobreajuste en los modelos de aprendizaje automático.

2.4 Análisis topológico

El uso de K-means en la generación de un mapper es crucial para capturar las complejidades estructurales de los datos. Al aplicar K-means, se puede crear un mapper que visualice de manera efectiva las relaciones y agrupaciones latentes en los datos, proporcionando así una representación topológica comprensible. Los diagramas de persistencia derivados de este mapper permiten identificar estructuras significativas, como agujeros y ciclos, que revelan información importante sobre la distribución y la conectividad de los datos.

La proyección de los datos sobre la característica clave "Color_Intensity" es un paso estratégico que potencia la capacidad de identificar patrones y relaciones específicas relacionadas con la calidad del vino. Dado que "Color_Intensity" se ha identificado previamente como una característica relevante para la calidad del vino, proyectar los datos sobre esta variable permite centrar el análisis en aspectos específicos que podrían influir en la calidad del producto final. Esto facilita la identificación de correlaciones y tendencias relacionadas con la calidad del vino, lo que a su vez puede proporcionar información valiosa para la toma de decisiones y la optimización de procesos en la industria vitivinícola.

3 Resultados

Las medias y las desviaciones estándar nos dicen mucho de las distribuciones de los datos. Por ejemplo, el alcohol tiene una media más cercana al valor máximo del conjunto lo cual sugiere asimetría, además posee una desviación estándar que indica poca dispersión lo cual tiene sentido con el rango tan reducido que se obtiene (entre 11.03 y 14.83), esto habla de que la cantidad de alcohol no cambia tanto entre los tipos de vinos), mientras que otros componentes como proline muestran una enorme dispersión por lo que sería interesante revisar las posibles razones para esto. Por otro lado la variable "Ash Alcanity" tiene una media (19.49) que parece estar prácticamente a la mitad del rango de registro (10.6 - 30) o mediana, misma que se obtendrá más adelante.

Ya que algunas características sugieren distribuciones no-normales, no sería factible realizar modelos de regresión lineal o la realización de análisis de varianza ANOVA en grupos generados por un modelo de clusterización ya que no cumpliría los supuestos necesarios, en su lugar sería mejor optar por un proceso de clustering.

Dado que tenemos más de 50 datos podemos aplicar una prueba shapiro para determinar qué variables se acercan a una distribución normal y cuáles no, aunque es necesario aclarar que al no rechazar la hipótesis tampoco se puede asegurar que los datos sigan una distribución normal, sin embargo, es bastante probable. Además, se añadieron los histogramas y gráficas QQ-plot para visualizar mejor cómo se ajusta cada conjunto de datos perteneciente a cada compuesto presente en los vinos a lo que sería una distribución normal.

Hipótesis Nula (H_0): Los datos siguen una distribución normal.
Hipótesis Alternativa (H_1): Los datos no siguen una distribución normal.
Valor α: 0.05

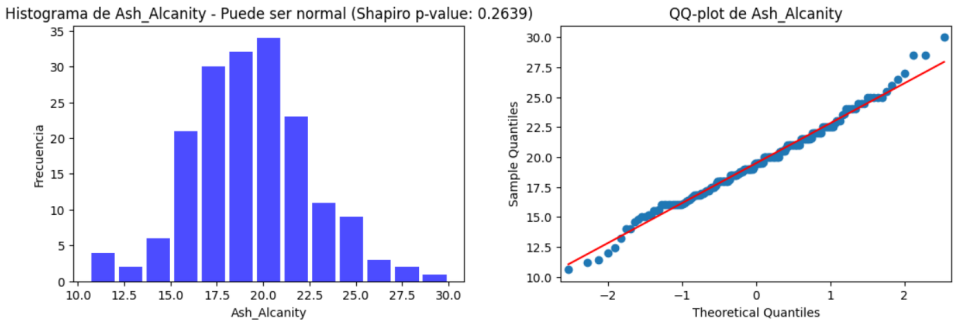

Figure 1: Prueba de normalidad en una variable ejemplo

Se realizaron las pruebas de las cuales se obtiene que la única característica posiblemente con distribución normal es, tal como se veía previniendo, Ash Alcalinity.

En cuanto a los histogramas, se observa que todos los registros de los compuestos, excepto Ash y OD280, tienen sesgos hacia la derecha, lo cual significa que los registros tienden a ser en cantidades menores ya que cuentan con un alargamiento en la cola correspondiente a los valores altos. En cuanto a interpretaciones en un contexto real de los vinos, se puede destacar lo siguiente:

- Vinos con valores extremadamente altos en las variables podrían considerarse excepcionales o únicos en términos de características sensoriales o composición química. Esto podría ser una oportunidad de estudio para los interesados en el vino, como la presencia de prolina que, a pesar de que en la mayoría de los vinos tiene un valor entre 500 y 700, hay una pequeña cantidad de vinos que llegan a tener un registro de más de 1600.
- Para los enólogos, el conocimiento de que la mayor parte de los registros para casi todas las características tienden a ser valores bajos comparados con los máximos que toman menor parte, podría influir en decisiones sobre prácticas de vinificación, selección de uvas, o métodos de producción para equilibrar o potenciar esas características.
- En cuanto a la revisión de estándares, dependiendo de las regulaciones y estándares de la industria del vino, los valores extremadamente altos podrían tener implicaciones para la conformidad con ciertos criterios de calidad o etiquetado.

Respecto a las gráficas QQ-plot, se observa que, a pesar de todo, la mayoría de las características tienen un buen ajuste a la de una distribución normal exceptuando las colas (por lo cual podría sugerir una regresión logarítmica). Los datos correspondientes a "OD80" y principalmente los de "Flavanoids" tienen una forma más distintiva que podría amoldarse mejor a una distribución logística.

La diferencia en la forma de estas distribuciones podría indicar que "OD80" y "Flavanoids" son características únicas o clave en el conjunto de datos de vinos. Podrían ser componentes críticos que contribuyen significativamente a la variabilidad y calidad de los vinos. Además, los flavonoides son conocidos por sus efectos antioxidantes; por lo tanto, variaciones en su concentración pueden ser de interés enológico.

Por otro lado, se realizó una matriz de correlaciones que nos brinda la siguiente información: Se identifica una alta correlación (88%) entre las variables *Total_Phenols* y *Flavanoids*. Entrando en contexto se obtienen las siguientes implicaciones:

1. **Perfil de Sabor y Aroma:** Ambas variables, los fenoles totales y los flavonoides, son compuestos químicos que contribuyen significativamente al perfil de sabor y aroma de los vinos. Un aumento en la concentración de uno podría estar asociado con un aumento similar en el otro, lo que podría influir en las características organolépticas del vino.
2. **Origen de la Uva:** La correlación puede sugerir que las condiciones del suelo, el clima o la cepa de uva específica en la región de cultivo podrían estar influyendo en ambas variables de manera similar. Esto podría ser de interés para los productores de vinos que buscan comprender las características distintivas de los vinos de una región en particular.
3. **Impacto en la Salud:** Tanto los fenoles como los flavonoides se han asociado con beneficios para la salud debido a sus propiedades antioxidantes. La correlación podría tener implicaciones para la percepción de los beneficios para la salud de consumir vinos ricos en estas sustancias.
4. **Procesos de Vinificación:** La correlación podría deberse a procesos específicos de vinificación utilizados en la producción de vinos. Por ejemplo, ciertas prácticas de maceración podrían influir en ambas variables de manera similar.

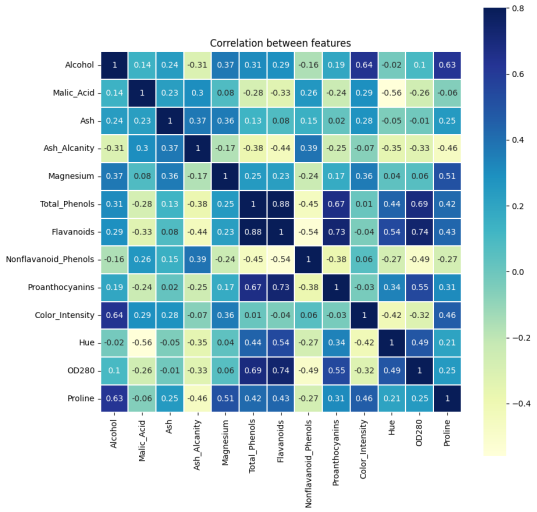

Figure 2: Tabla correlación

Se estandarizaron los datos ya que estandarizar los datos es una práctica fundamental en el proceso de análisis de datos, especialmente en contextos donde se aplican algoritmos sensibles a la escala de las variables, como el PCA y el K-means. En el contexto de este proyecto, donde se ha utilizado el PCA para reducir la dimensionalidad de los datos y el K-means para generar un mapper. De acuerdo a la gráfica obtenida se eligen 8 componentes que después de la proyección retienen el 92% de la varianza

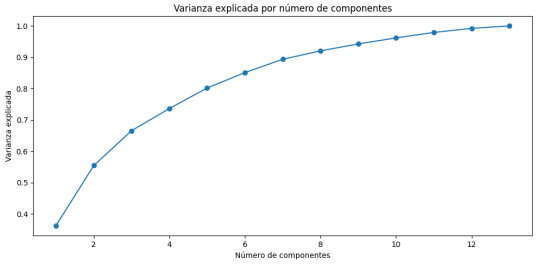

Figure 3: Varianza explicada

Para hacer el mapper se usó Kmeans, donde para elegir el mejor número de clusters se utilizó una gráfica del codo.

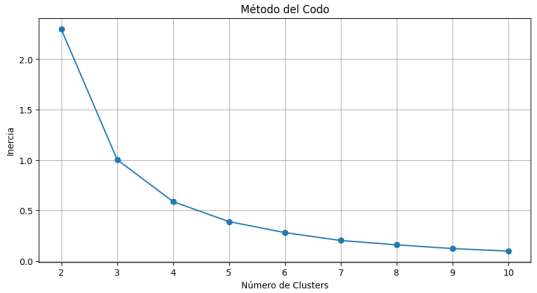

Figure 4: Codo

Para la realización del mapper, "Color Intensity" es una característica clave en la evaluación de la calidad y el carácter sensorial de un vino. Esta medida representa la intensidad del color del vino, que está influenciada por la concentración de pigmentos y compuestos fenólicos presentes en la bebida. La intensidad del color puede proporcionar información sobre la madurez de las uvas utilizadas en la elaboración del vino, así como sobre los métodos de vinificación empleados.

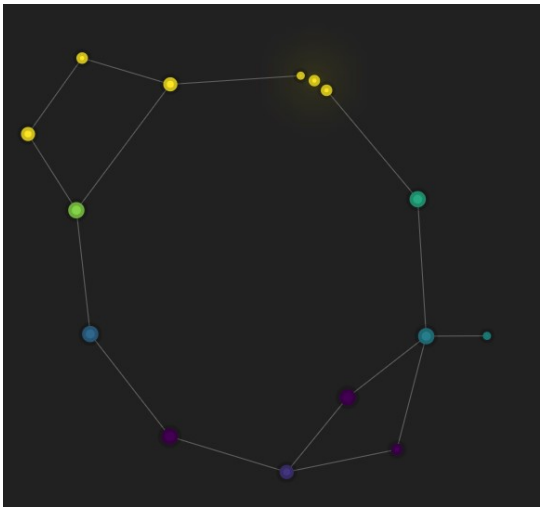

Figure 5: Mapper

En el contexto topológico del Mapper, la proyección sobre la característica "Color_Intensity" revela una estructura subyacente en el espacio de características de los vinos. La disposición de los nodos y conexiones en el grafo del Mapper sugiere la existencia de regiones o cúmulos de vinos con perfiles de intensidad de color similares. Estas regiones topológicas representan subconjuntos de vinos que comparten características sensoriales específicas en términos de color, lo que indica una coherencia en la composición química y las propiedades organolépticas de los vinos dentro de cada grupo.

La presencia de conexiones entre los nodos del Mapper indica que, aunque los vinos dentro de un mismo cúmulo comparten similitudes en términos de intensidad de color, también pueden exhibir variaciones en otras características, como el aroma, el sabor y la textura. Estas conexiones representan transiciones suaves o gradientes entre diferentes perfiles sensoriales, lo que sugiere la existencia de relaciones complejas y sutiles entre las características del vino.

En un contexto aplicado, la estructura topológica del Mapper proporciona información valiosa para la clasificación y la evaluación de la calidad del vino. Por ejemplo, los enólogos y catadores podrían utilizar esta información para identificar grupos de vinos con perfiles sensoriales similares y entender mejor las preferencias del consumidor. Además, el análisis topológico del Mapper podría ayudar a revelar relaciones ocultas entre las características del vino y factores externos, como la región vitivinícola de origen, las prácticas de cultivo y vinificación, y las preferencias culturales de los consumidores si se analizan a estas características mencionadas de los registros que pertenecen a cierto cúmulo, en este caso los de mayor intensidad que se diferencían mucho de los otros. En última instancia, esta comprensión más profunda de la estructura y la variabilidad del vino podría informar y mejorar la toma de decisiones en la producción y comercialización del vino. Además, como parte de la conclusión, se logró determinar topológicamente que se reconocen 3 tipos de vino en la base.

Puede encontrar el repositorio en GitHub en: https://github.com/anapaola03/ea2.