Estrategias de aprendizaje NO supervisado para la clasificación de variedades de vinos.

Alejandra Aguirre, Daniel Manco, Esteban Arcila

Departamento de Ingeniería Industrial
Universidad de Antioquia
Medellín, Colombia

INTRODUCCIÓN

Las técnicas de clasificación y agrupación de datos tienen una función importante para la identificación de características específicas de estos, a través del reconocimiento de patrones comunes entre tal información. Esto tiene vital importancia cuando se cuenta con datos desconocidos y se necesita un análisis de ellos para comprender su naturaleza para la toma de decisiones.

En la industria de producción de vinos, es importante la determinación de las cantidades los componentes que contiene un vino, ya que permite:

- Garantizar la calidad del vino por parte del productor.
- Certificar la calidad del vino por parte de la asociación vitivinícola.
- Desarrollar datos del vino para aclarar las situaciones que han provocado anomalías en el sistema industrial.
- iniciar una nueva idea de negocio del vino según las demandas de diferentes cultivares.

Se debe tener en cuenta, que a pesar de que un tipo de vino provenga de una misma especie de uva, sus propiedades químicas pueden variar dependiendo de su proceso de vinificación, haciendo que su clasificación según sus características sea distinta, por tanto, varía su calidad y certificación.

Este caso de estudio pretende realizar una clasificación teniendo en cuenta los valores de los componentes químicos de cada muestra, con el fin de asignar una tipología acorde a lo anteriormente mencionado.

DISEÑO DE SOLUCIÓN PROPUESTO

Para este ítem se plantean una serie de objetivos que describen la metodología y procesos necesarios para generar resultados que permitan la toma de decisiones:

- Evaluar el rendimiento de diferentes técnicas de clasificación, como Hierarchical Clustering, K Means y DBSCAN en la clasificación de tipos de vino basada en datos químicos.
- Determinar cuál de las técnicas de clasificación ofrece la precisión más alta en la agrupación de los tipos de vino.
- Encontrar un algoritmo que permita la clasificación de los vinos que tienen una mejor calidad, identificando y prediciendo la calidad del vino en función de sus características químicas con el fin de que las bodegas y productores de vino tomen decisiones, como la comercialización de vinos específicos para satisfacer la demanda del mercado.

MATERIALES Y MÉTODOS

Para el desarrollo de este trabajo, se aplicó estadística descriptiva y modelos de machine learning a través de herramientas de programación. Se aplicó un tratamiento sistemático a las bases de datos con el objetivo de limpiar, transformar y visualizar bajo el siguiente proceso.

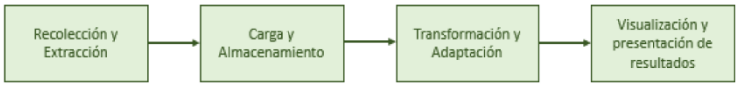

Imagen 1. Proceso de tratamiento a datos. Fuente: (Material de clase)

A. Bases de datos

La información necesaria para el proyecto se presenta en bases de datos. A continuación, se describe la base de datos trabajada y su contenido:

- **wine:** Contiene información química sobre vinos fabricados en tres cultivares diferentes de una misma región de Italia.

A continuación, se explica cada una de las variables que contiene este dataframe.

- **Alcohol (Grados):** Contenido de alcohol en el vino.
- **Malic Acid (g/L)**: El ácido málico es un ácido orgánico que se encuentra de forma natural en las uvas.
- **Ash (g/L):** La ceniza se refiere a las sustancias inorgánicas que quedan después de quemar una muestra de vino.
- **Alcalinity of ash (g/L de K2O):** Cantidad de alcalinidad en la ceniza.
- **Magnesium (mg/L):** Magnesio presente en el vino.
- **Total phenols (mg GAE/g de ES):** Los fenoles son compuestos químicos que se encuentran en las uvas y pueden aportar propiedades

antioxidantes y color al vino.

- **Flavanoids (mg CE/g de ES):** Los flavonoides son un subconjunto de fenoles que contribuyen a los colores y sabores del vino.
- **Nonflavanoids Phenols (mg GAE/g de ES):** Estos son otros tipos de fenoles que no son flavonoides.
- **Proanthocyanins (mg/L):** Las proantocianidinas son un tipo específico de flavonoides que pueden contribuir a la estructura y el sabor del vino.
- **Color intensity (cd/m²):** Profundidad y riqueza del color del vino.
- **Hue (grados):** Tonalidad del color del vino.
- **OD280/OD315 of diluted wines (unidades de absorbancia):** Estos valores se refieren a la absorbancia de luz del vino.
- **Proline (mg/L):** La prolina es un aminoácido que se encuentra en el vino.

B. Herramientas de programación

Para aplicar cada una de las fases de tratamiento de datos y modelado, se dispuso de una de las herramientas de programación y consulta más conocida. A continuación, se describirán las herramientas de programación y que funciones de estas se utilizaron.

- Python: Python es un lenguaje de programación ampliamente utilizado en las aplicaciones web, el desarrollo de software, la ciencia de datos y el machine learning (ML). [1]
- Una de las principales funciones o librerías utilizadas, además de las librerías para manipulación de datos numéricos y gráficos, de este lenguaje, fue *Pandas*. Es una librería de Python especializada en el manejo y análisis de estructuras de datos. [2]
- Otra de las librerías utilizadas fue *Sklearn* para todo el tema de machine learning, Esta librería proporciona una amplia variedad de algoritmos y herramientas para tareas como clasificación, regresión, agrupación, selección de características, reducción de dimensionalidad y más. Además, ofrece utilidades para la evaluación y validación de modelos. [3]

C. Imputación

Para realizar cualquier modelo de clasificación, primero se hace un análisis exploratorio de los datos donde lo que se pretende es hacer limpieza e imputación a su vez que se caracterizan las variables para que sean compatibles con todos los algoritmos que se van a utilizar posteriormente.

A partir del análisis anterior, se realiza una matriz de correlación, con el fin de encontrar variables que fueran colineales con otras variables. A partir de su lectura se decide eliminar la variable ***'Flavanoids'***, ya que presenta una alta correlación respecto a las demás variables. En la siguiente gráfica (véase imagen 1) se presenta la matriz de correlación del dataframe ***'wine'***.

Posteriormente se encuentran datos atípicos en algunas de las variables, los cuales se proceden a eliminar a través del rango intercuartil, con esto se pasa de tener 178 observaciones a 173. Este proceso se almacena en una nueva base de datos llamada ***'data_clean'.*** Esta nueva base de datos se estandariza a través de la función *StandarScaler* y se procede a almacenar en una nueva variable llamada ***'scaled_features'***.

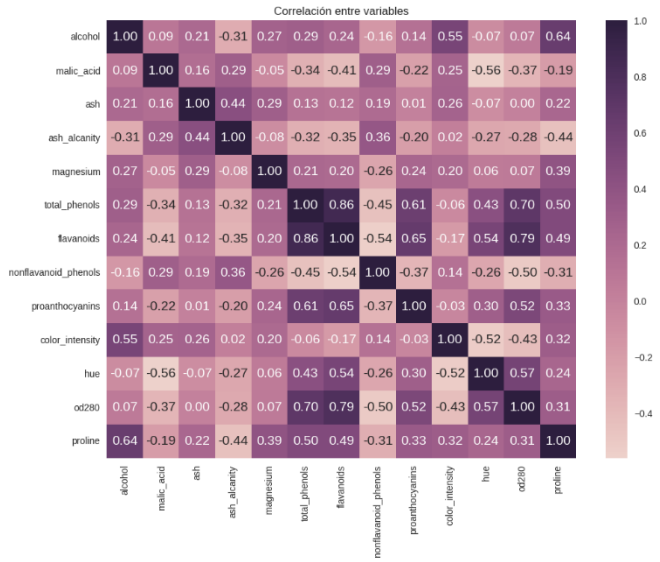

*Imagen 1. Matriz de correlación **'wine'** – Elaboración propia*

D. Reducción de la dimensionalidad

Como herramienta de reducción de dimensionalidad en los datos, se aplica la técnica de *Principal Component Analysis. (PCA)*. La función principal del *PCA* es reducir la dimensionalidad de un conjunto de datos mientras intenta retener la mayor cantidad de información posible en los datos originales. Esta herramienta se escogió ya que una de sus características es reducir la colinealidad de los datos, dado a la naturaleza ortogonal de los componentes que origina.

Para configurar esta herramienta, se le asigna a su parámetro ***'n_components'*** un valor del 85%, siendo este el valor de la varianza que se debe explicar dentro de los componentes. Posteriormente se grafican estos componentes. En la siguiente imagen se visualiza la gráfica anteriormente mencionada.

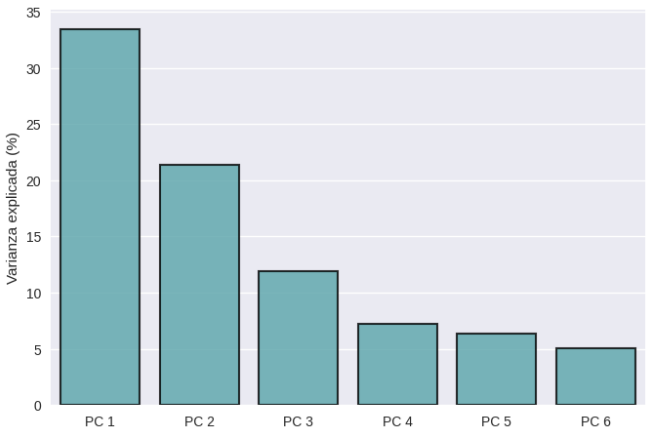

Imagen 2. varianza explicada por componente PCA – Elaboración propia

A continuación, se muestran los valores específicos que representa cada componente.

	Varianza explicada (%)
0	33.470550
1	21.330124
2	11.899387
3	7.254621
4	6.350213
5	5.025412

Imagen 3. Valores de varianza explicada por componente PCA – Elaboración propia

Según el análisis realizado, se decide seleccionar los primeros tres componentes, ya que contienen los valores mas altos de varianza. Estos componentes se anexan a la ultima base de datos creada, y se genera una nueva llamada ***'df_final'.***

RESULTADOS

A continuación, se mostrarán los resultados de las técnicas de clasificación realizadas. Se resalta que las métricas de desempeño que se calcularán son: *Calinski Harabasz Score* y *Silhouette Score*. Además, se tendrá en cuenta el comportamiento de la grafica de dispersión de cada modelo.

A. *K-Means*

Según los diagramas de siluetas, el kneelocator y el diagrama de codos, sugieren que el número de clusters en los cuales pueden estar clasificados los datos son tres. Se aplica esta sugerencia al modelo y se realiza un análisis de sensibilidad al modelo, el en cual se ejecuta el modelo con seis componentes PCA y con tres componentes.

Para mostrar los resultados, se muestran las métricas de desempeño de ambos modelos en el orden anteriormente mencionado.

-K-Means con 6 componentes PCA

```
### K-MEANS ###
Inertia:  1163.3036708277946
Silhouette Score:  0.28427481360421053
Calinski harabasz score:  66.68868148970328
```

Imagen 4. Métricas de desempeño K-Means con 6 componentes PCA – Elaboración propia

Estos resultados reflejan bajas metricas en el desempeño del modelo K-Means en la clasificacion. Posteriormente se analiza el modelo con 3 componentes PCA.

-K-Means con 3 componentes PCA

```
### K-MEANS ###
Inertia:  478.5475161192927
Silhouette Score:  0.4561077274802859
Calinski harabasz score:  160.95034442444768
```

Imagen 5. Métricas de desempeño K-Means con 3 componentes PCA – Elaboración propia

En este caso se nota una mejora considerable en las métricas de desempeño, mostrando la viabilidad del modelo bajo tres componentes que explican la varianza de los datos.

Por ultimo se grafica la distribución de los datos y la clasificación de los cluster según el numero asignado en la preparación del modelo. Este esquema se muestra a continuación mediante un diagrama de dispersión.

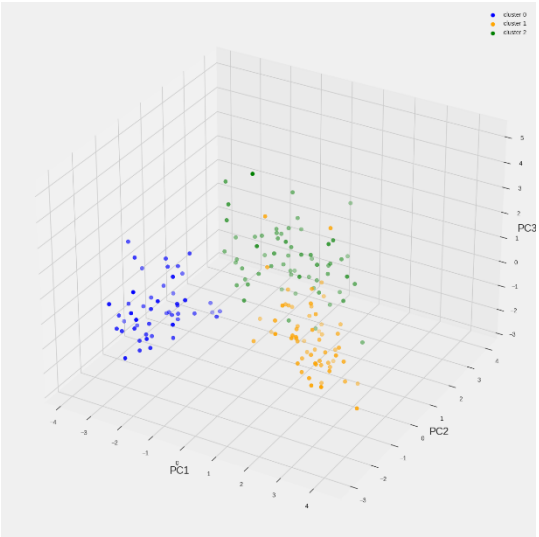

Imagen 6. Diagrama de dispersión K-Means – Elaboración propia

En este se observa la clasificación por cluster de los datos. Sin embargo, se evidencia algunos valores con solapamiento. Esto se da porque se tiene un valor de coeficiente de siluetas bajo a pesar de que esta métrica con tres componentes PCA mejoró. Por tanto, Estas métricas no son las ideales para un modelo de clustering, haciendo que algunos datos no se agrupen adecuadamente.

B. *Hierarchical Clustering*

Según el Dendograma utilizado para este modelo, sugiere el número de clusters apropiados para la clasificación de datos es de tres. También, a raíz de los resultados del modelo anterior, se decide que para este modelo también se aplicarán tres componentes PCA.

A continuación, se obtienen las métricas de desempeño para este modelo de clustering.

```
### Hierarchical Clustering ###
Índice de Calinski-Harabasz: 61.96948513908285
Puntuación de silueta: 0.2676034398174304
```

Imagen 7. Métricas de desempeño Hierarchical Clustering – Elaboración propia

Se observa que estas métricas son bajas, esto se evidenciará en el diagrama de dispersión que se mostrará a continuación.

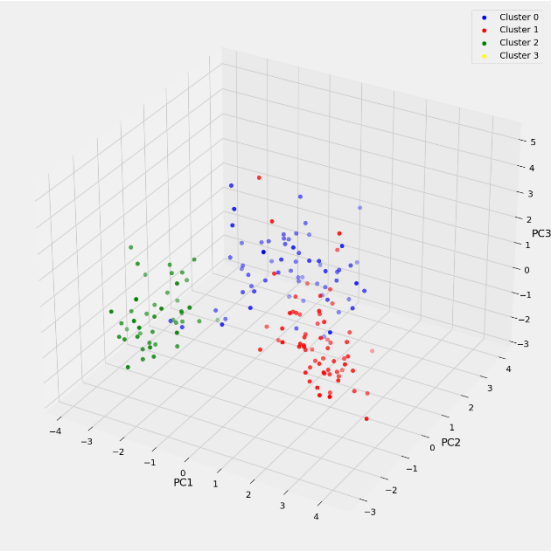

Imagen 8. Diagrama de dispersión Hierarchical Clustering – Elaboración propia

En esta grafica se evidencian las bajas métricas obtenidas en este modelo, ya que los clusters tienen solapamiento y no se presenta una separación evidente entre cada agrupación.

C. DBSCAN.

En este modelo de clasificación, uno de los parámetros principales es el *Epsilon* y el *min_samples*, siendo *Epsilon* la distancia máxima dentro de la cual se buscarán otros puntos para formar un cluster con el punto central, y el *min_samples* es el número mínimo de puntos requeridos para que un grupo de puntos se considere un cluster válido.

En la siguiente imagen, se mostrará los valores asignados a los parámetros anteriores.

```
DBSCAN
DBSCAN(eps=2.303597021291516, min_samples=10, n_jobs=-1)
```

Imagen 9. Parámetros modelo DBSCAN– Elaboración propia

Teniendo en cuenta esta información, se procede a calcular las métricas de desempeño para este modelo. Estos son sus resultados:

```
### DBSCAN ###
Silhouette Score:  0.25165786992874073
Calinski harabasz score:  2.5658413560268976
```

Imagen 10. Métricas de desempeño DBSCAN – Elaboración propia

Estos resultados muestran que el modelo tiene solapamientos y no tiene separaciones evidentes entre clusters. Esto se confirma en la siguiente gráfica.

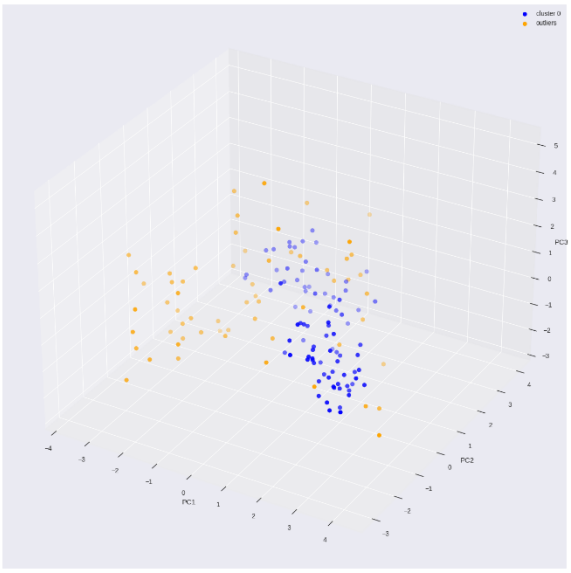

Imagen 11. Diagrama de dispersión DBSCAN – Elaboración propia

En este modelo se evidencia que con las metricas dadas al modelo, solo logra identificar un cluster, que son los datos de color azul, los puntos amarillos los identifica como ***'outlers'***, dando a mostrar que estos datos no cumplen con los valores establecidos de *Epsilon* y *min_samples*.

Este hallazgo muestra que este modelo de clustering no es el apropiado para este tipo de datos, ya que, para encontrar resultados considerables para el caso de estudio, se debe realizar un análisis de sensibilidad exhaustivo para encontrar un valor especifico para ambos componentes del modelo. Este análisis no hace parte del alcance del caso de estudio.

ANÁLISIS DE RESULTADOS

Observando las métricas de desempeño y los diagramas de dispersión de todos los modelos de clustering, se decide que el modelo que mejor realiza una clasificación de los datos es el K-Means, por tanto, este es el modelo al cuál se hará el análisis de cada uno de sus clusters.

INTERPRETACIÓN DE CLÚSTERS

Inicialmente, se observa la distribución de valores de los datos dentro de cada cluster que se conformó mediante el modelo K-Means..

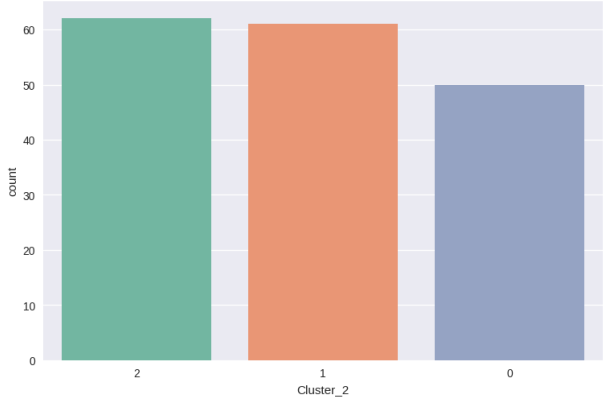

Imagen 12. Distribución de los valores dentro de cada cluster – Elaboración propia

Se evidencia que en el cluster (2) y (1) hay un equilibrio de datos, el desequilibrio de datos se observa en el cluster (0), pero es un desequilibrio moderado respecto a los observados en otros casos de estudio.

Después de esto, se debe considerar la importancia de cada uno de los componentes químicos en la composición del vino. Por eso se investiga sobre el aporte de cada variable al vino y se decide que las siguientes variables son las que más aportan a este. Estas variables son:

- **alcohol**
- **malic_acid**
- **magnesium**
- **total_phenols**
- **color_intensity**
- **proline**

Estas variables son a las que se le realizará el análisis dentro de cada cluster a través de histograma, las graficas de muestran a continuación.

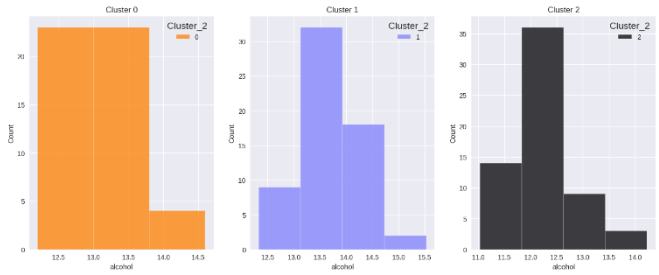

Imagen 13. Comportamiento de variable 'alcohol' dentro del cluster - Elaboración propia

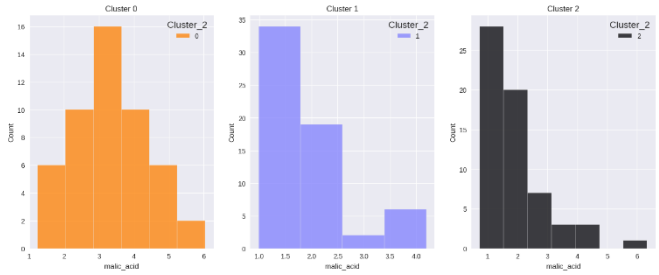

Imagen 14. Comportamiento de variable 'malic_acid' dentro del cluster – Elaboración propia

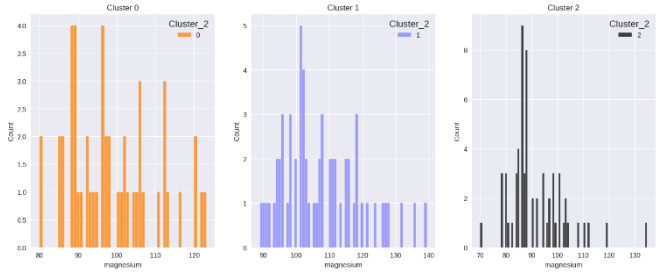

Imagen 15. Comportamiento de variable 'magnesium' dentro del cluster – Elaboración propia

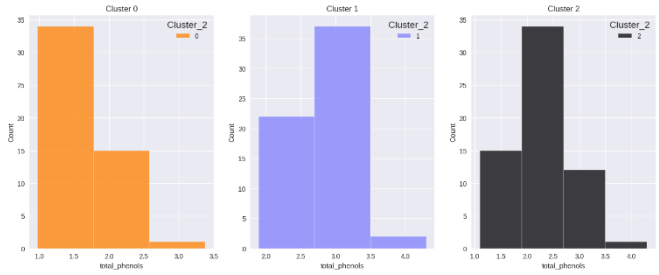

Imagen 16. Comportamiento de variable 'total_phenols' dentro del cluster – Elaboración propia

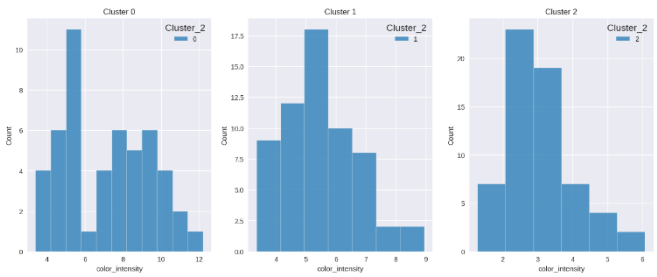

Imagen 17. Comportamiento de variable 'color_intensity' dentro del cluster – Elaboración propia

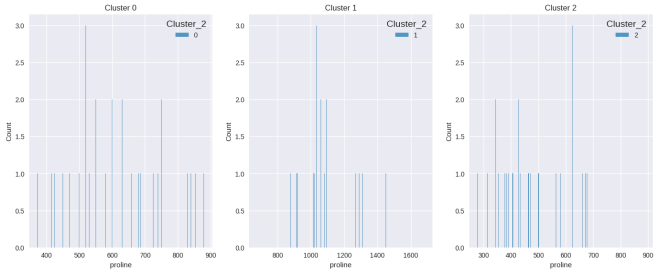

Imagen 18. Comportamiento de variable 'proline' dentro del cluster – Elaboración propia

A partir de los gráficos, se caracteriza cada uno de las variables y sus valores dentro de cada cluster.

A. *Cluster (0).*

Alcohol: Sus grados de alcohol se concentran entre el 12.5 y el 13.

Malic_acid: El acido málico se concentra entre el valor de 3 y 3.5.

Magnesium: Sus valores se concentran entre 90 y 98.

Total_phenols: Sus valores se concentran entre el 1 y 1.5.

Color_intensity: Selecciona los valores entre 4 y 12.

Proline: sus valores se concentran entre 400 y 900.

B. Cluster (1).

Alcohol: Sus grados de alcohol se concentran entre el 13 y el 13.5.

Malic_acid: El ácido malico se concentra entre el valor de 1 y 1.5.

Magnesium: Sus valores se concentran entre 100 y 103.

Total_phenols: Sus valores se concentran entre el 3 y 3.5.

Color_intensity: Selecciona los valores entre 4 y 9.

Proline: sus valores se concentran entre 800 y 1500.

C. Cluster (2).

Alcohol: Sus grados de alcohol se concentran entre el 12 y el 12.5.

Malic_acid: El ácido malico se concentra entre el valor de 1 y 2.

Magnesium: Sus valores se concentran entre 80 y 90.

Total_phenols: Sus valores se concentran entre el 2 y 2.5.

Color_intensity: Selecciona los valores entre 2 y 6.

Proline: sus valores se concentran entre 300 y 700.

CONCLUSIONES

A continuación, se postularán las conclusiones más importantes de este estudio.

- El método de Clustering K-Means podría definirse como el apropiado para el caso de estudio aplicado, ya que esta base de datos era relativamente pequeña, ni sus valores tenían altas oscilaciones.
- Teniendo en cuenta que se realizaron iteraciones de los modelos con bases de datos no estandarizadas, se decide por estandarizar los datos, ya que la varianza del PCA quedaría con un solo componente. Sin embargo, las iteraciones con el dataframe sin estandarizar arrojaron mejores resultados que los modelos con datos estandarizados, mostrando la opcionalidad de estandarizar en este caso de estudio.
- Se encontró que la clasificación de los vinos se puede explicar mediante los clusters, teniendo en cuenta que se podrían comercializar y producir vinos de gama económica, media y de lujo con las especificaciones que se encontraron en cada uno de ellos respectivamente para asegurar la calidad en cada una de las gamas.
- Se define que el Cluster (1) es el que contiene la mejor especificación químicas para asegurar una calidad superior en el vino.

REFERENCIAS

[1] Amazon, «AWS Amazon,» Amazon Web Services, 1 Enero 2022. [En línea]. Available: https://aws.amazon.com/es/what-is/python/. [Último acceso: 25 Septiembre 2022].

[2] A. S. Alberca, «Aprende con Alf,» Aprende con Alf, 14 Junio 2022. [En línea]. Available: https://aprendeconalf.es/docencia/python/manual/pandas/. [Último acceso: 25 Septiembre 2022]

[3] scikit-learn, « scikit-learn Machine Learning in Python» Diciembre 2022. [En línea]. Available: https://scikit-learn.org/stable