UNIDAD 3 APRENDIZAJE NO SUPERVISADO

APRENDIZAJE AUTOMÁTICO 0313_2361

German Andrés Álvarez López

Carlos Andrés Castro Marín

Jonathan Javier Montes Castro

Karla Victoria Torres Parra

PROFESOR

José Lisandro Aguilar Castro

UNIVERSIDAD EAFIT

MAESTRIA EN CIENCIA DE LOS DATOS Y ANALITICA

MEDELLÍN

2023

Introducción

Los algoritmos de aprendizaje supervisado son útiles cuando los conjuntos de datos a estudiar tienen etiquetas que los diferencian unos a otros, a esto también se le suele llamar variable dependiente. Cuando no se tienen etiquetas técnicas como el aprendizaje no supervisado (clusterización) se hacen necesarios para diferenciar los grupos de datos unos a otros así creando las etiquetas que los diferencian y poder hacer estimaciones a partir de estos conjuntos de datos previamente agrupados.

Objetivo de la iteración

Para probar el poder de los algoritmos de aprendizaje no supervisado utilizaremos un dataset previamente etiquetado y eliminaremos dicha etiqueta para comprobar el poder de agrupamiento de diferentes algoritmos de clusterización. Exploraremos los resultados con diferentes tipos de etiquetas para así visualizar cómo es que estos algoritmos agrupan correcta o incorrectamente dependiendo de la etiqueta seleccionada.

Contextualización del problema

Para este problema utilizaremos un dataset compuesto de dos conjuntos de diferentes tipos de vino (rojo y blanco) con sus respectivas variables que impactan la calidad, este dataset es descargado de la base de datos de Kaggle (Kaggle Datasets, 2017).

Tanto para el dataset vino rojo como el de vino blanco tenemos la siguiente lista de variables de entrada:

- Fixed Acidity
- Volatile Acidity
- Citric Acid
- Residual Sugar
- Chlorides
- Free Sulfur Dioxide
- Total Sulfur Dioxide

- Density
- pH
- Sulphates
- Alcohol

Y la variable de salida o etiqueta:

- Quality (score entre 0 y 10).

En este problema utilizaremos dos aproximaciones, la primera será utilizar la variable de sailda **Quality** como la etiqueta a buscar por medio de clusterización únicamente con el dataset de vinos blancos, a este le llamaremos **dataset 1**, la segunda es será utilizar como etiqueta si el vino es **rojo** o **blanco,** esto es, mezclar ambos datasets en uno solo y crear la etiqueta para el tipo de vino, a este le llamaremos **dataset 2**. Veremos que los resultados en ambos casos son totalmente diferentes.

Preparación de los datos e ingeniería de características

1. En el primer paso evaluamos si los datasets tienen algún valor nulo que deba ser eliminado o interpolado. Encontramos que ambos datasets están limpios.

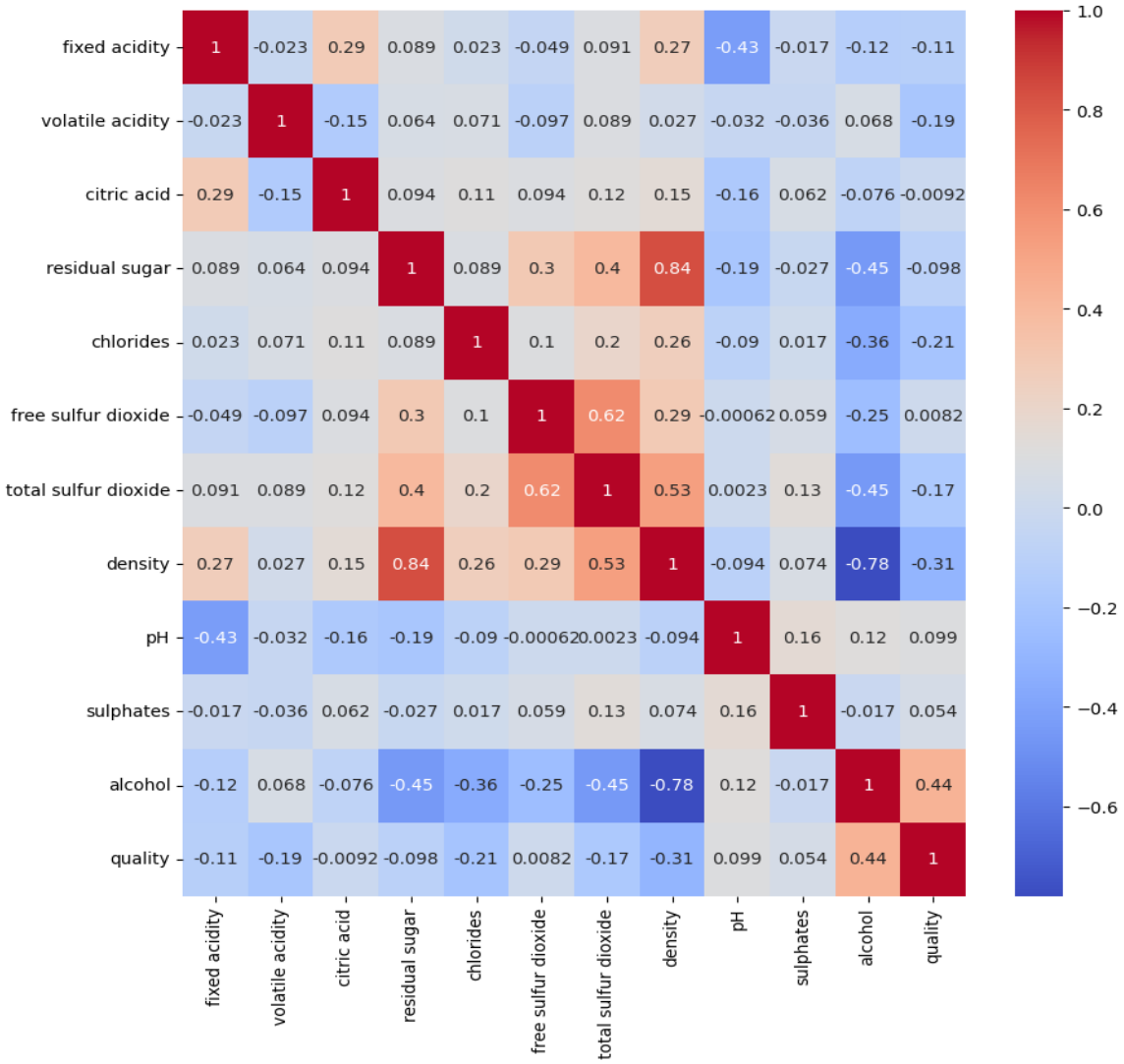
```
[7]  df.isnull().sum()

     fixed acidity           0
     volatile acidity        0
     citric acid             0
     residual sugar          0
     chlorides               0
     free sulfur dioxide     0
     total sulfur dioxide    0
     density                 0
     pH                      0
     sulphates               0
     alcohol                 0
     quality                 0
     dtype: int64
```

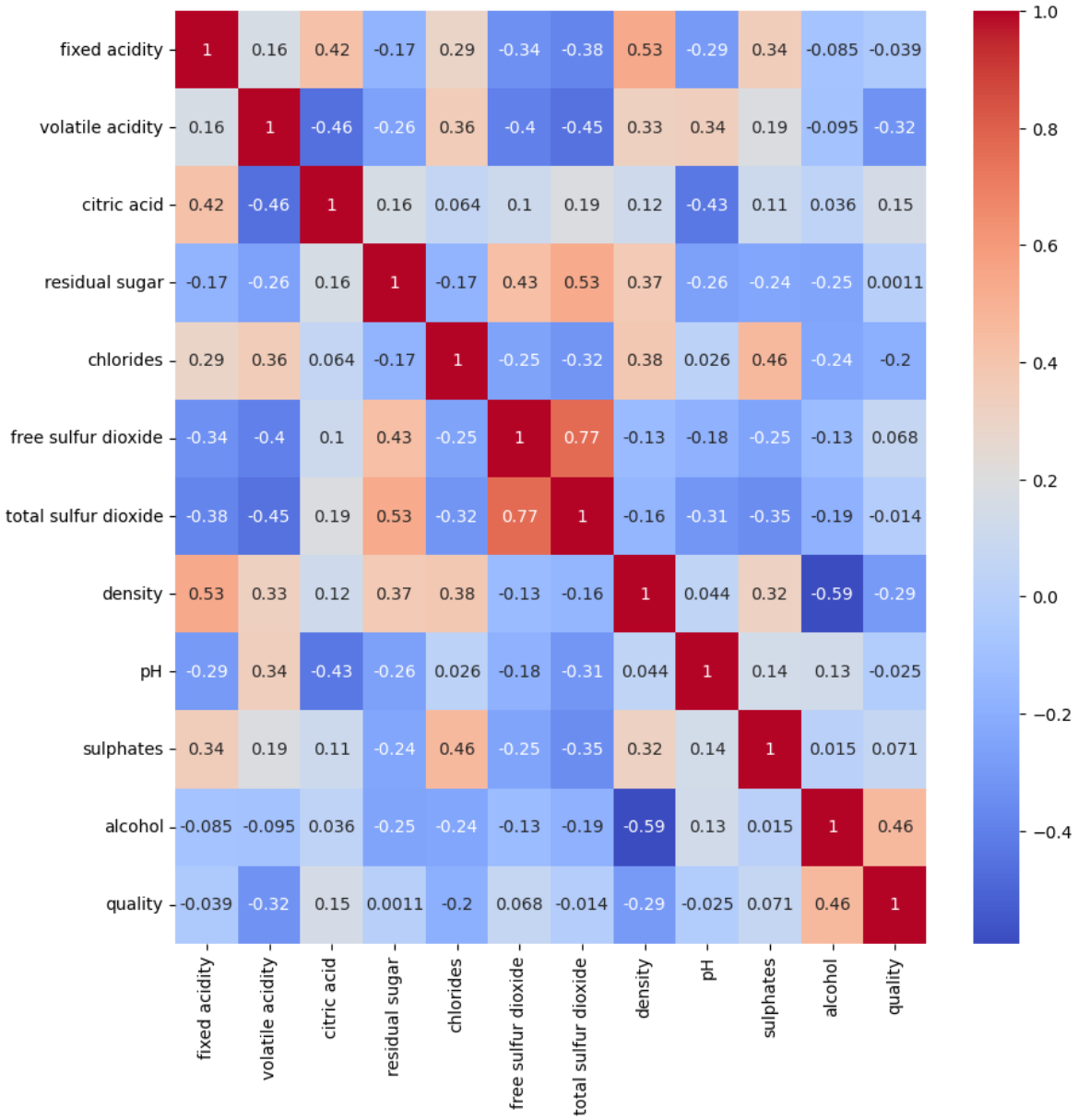
2. En el siguiente paso evaluamos las matrices de correlaciones de ambos datasets, observamos que para el dataset 1 no hay correlación clara entre la calidad que es la variable etiqueta, esto es de esperarse pues la variable calidad es una variable categórica que va de 0 a 10. Ahora bien, encontramos que existe correlación entre

algunas de las variables indicando posibles patrones en los datos, este es el mismo caso para el dataset 2.

Dataset 1 Matriz de Correlación



Dataset 2 Matriz de Correlación



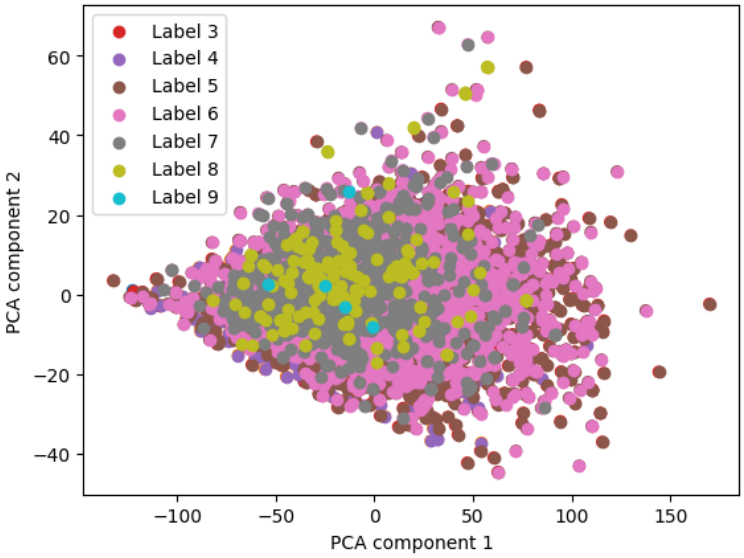
3. Posteriormente procedemos a remover outliers en los datos utilizando la distancia de mahalanobis para datos multivariados. Una vez conocidas las distancias nos quedamos únicamente con el cuantil 96 de los datos, es decir, eliminamos el 4% de los datos más alejados.

4. Para practicidad, eficiencia y visualización realizamos un análisis de componentes principales con el fin de reducir la dimensionalidad del problema a 2 dimensiones y así facilitar el proceso de clusterización.

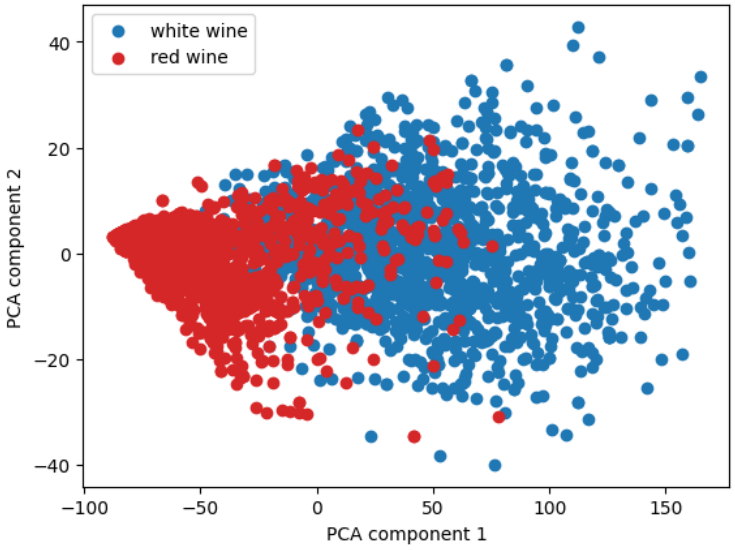
- Para el dataset 1 las 2 dimensiones explican el 98.8% de la varianza.
- Para el dataset 2 las 2 dimensiones explican el 99.5% de la varianza.

5. Una vez obtenidas ambas dimensiones, visualizamos las mismas para tratar de identificar los clústeres reales visualmente.

Clusters Dataset 1 en 2D



Clusters Dataset2 en 2D



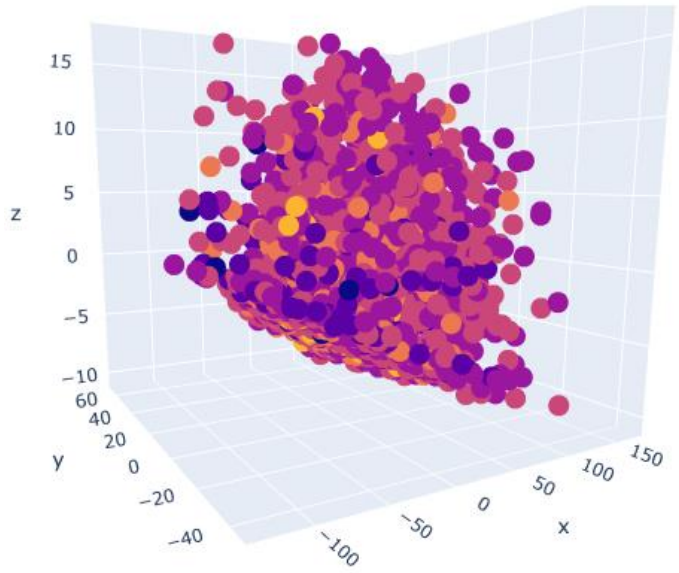
Podemos observar cómo en ambos dataset los clústeres hacen overlap, sin embargo, el clúster del dataset 2 tiene una diferenciación más clara que los clústeres del dataset 1. Esto pues, en el dataset 1 tenemos variables categóricas entre 0 y 10 mientras que en el dataset 2 las variables categóricas son binarias, vino rojo o blanco.

6. Para poder hacer una separación más clara de los clústeres procedemos a agregar una dimensión nueva, esto es, realizar el PCA para quedarnos con 3 dimensiones únicamente.

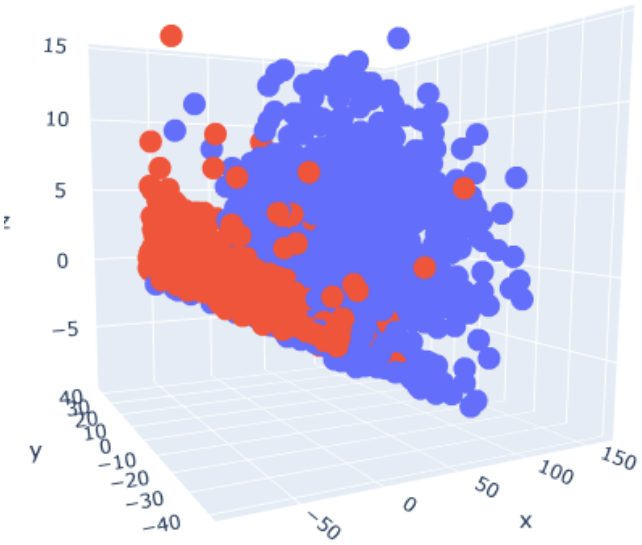
    - Para el dataset 1 las 3 dimensiones explican el 98.9% de la varianza.
    - Para el dataset 2 las 3 dimensiones explican el 99.8% de la varianza.

    Las siguientes son representaciones en 3D de los clústeres que pueden ser manipuladas con mayor facilidad dentro del notebook para interpretar la separación de los conjuntos.

Clusters en 3D del Dataset1



Clusters en 3D del Dataset 2



Observamos que agregar una dimensión nueva no nos da una representación visual clara de los clústeres del dataset 1, sugiriendo que probablemente la etiqueta de calidad no sea la agrupación apropiada de los datos. Por otro lado, vemos que los clústeres del dataset 2 están claramente diferenciados entre los grupos de vino blanco y rojo a excepción de contados outliers que pueden ser ruido inherente al dataset.

Modelamiento y Resultados

Para el modelamiento utilizaremos tres algoritmos de aprendizaje no supervisado.

- K-means: Es necesario definir previamente el número de clústeres.
- Hierarchical Clustering: Es necesario definir previamente el número de clústeres.
- DBSCAN: Hace la búsqueda de clústeres basado en la densidad de los puntos.

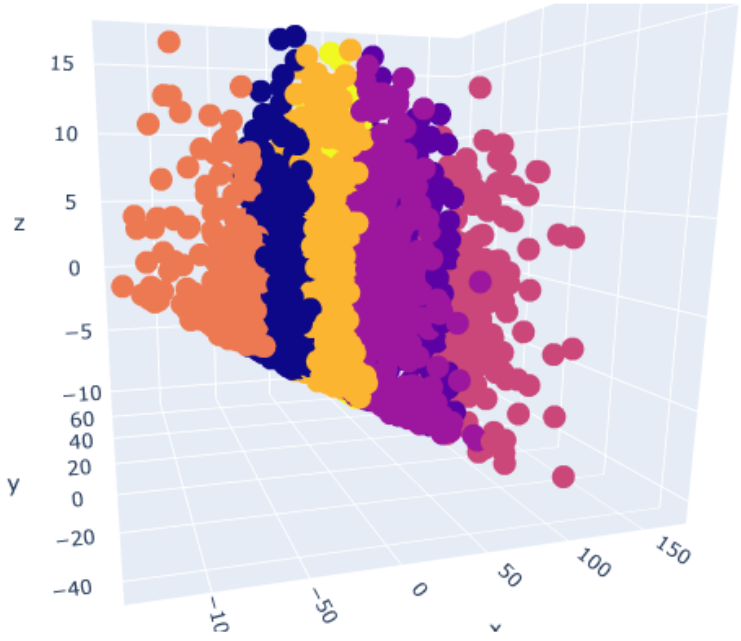
Para el dataset 1 el modelamiento se limitará a k-means y hierarchical clustering, el dataset 2 se realizará con los tres algoritmos.

Se utilizarán las siguientes métricas de rendimiento para todos los algoritmos:

- **Adjusted Rand Index:** Mide la similitud entre la estructura de agrupación obtenida por el algoritmo y una estructura de referencia conocida, ajustando por azar.

- **Silhouette Coefficient:** Mide la cohesión y separación de los clusters. Valores más altos indican clusters mejor definidos.
- **Homogeneity:** Mide qué tan puro es cada cluster, es decir, si todos sus elementos pertenecen a la misma clase.
- **Completeness:** Mide si todos los elementos de una misma clase están en el mismo cluster.
- **V-measure:** Combina homogeneidad y completitud en una sola medida, otorgando más peso al valor más bajo entre ambas.

1. **Modelamiento Dataset 1**
   - K-means

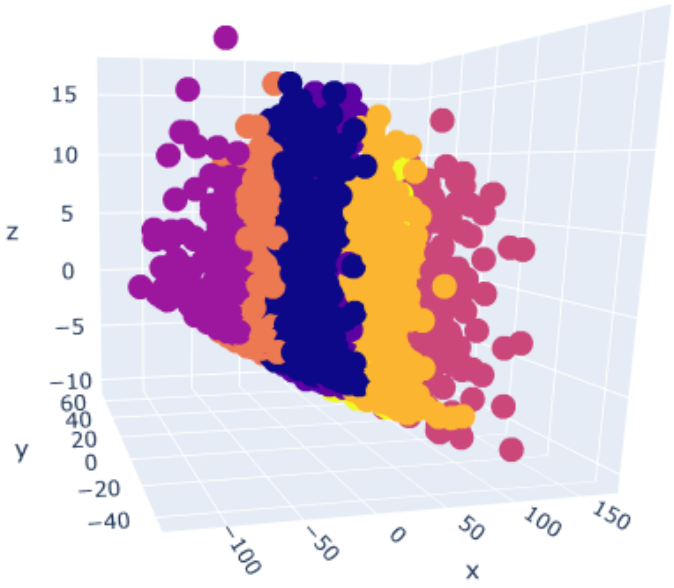


Adjusted Rand Index	0.012579
Silhouette Coefficient	0.191358
Homogeneity	0.978614
Completeness	0.095107
V-measure	0.173366

- Vemos que el ARI es aleatorio, el "agreement" entre los clusters reales y los predecidos.
- El coeficiente de Silhouette nos indica que la clusterización es moderada.
- El índice de homogeneidad, completitud y medida V son muy bajos. indicándonos la clusterización no hace buen fit a las clases reales.

Se concluye que la clusterización por K-means no es buena cuando la etiqueta evaluada es la calidad.

- Hierarchical Clustering

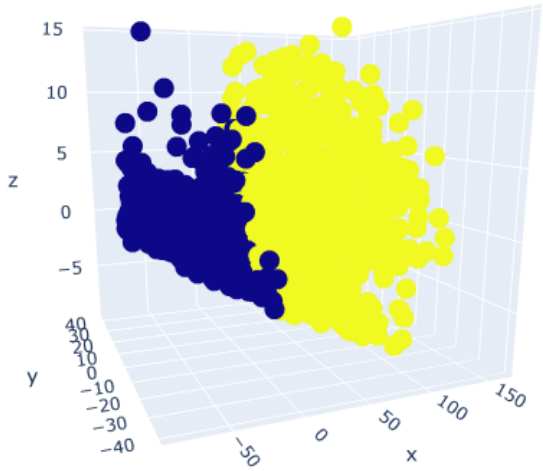


Adjusted Rand Index	0.012579
Silhouette Coefficient	0.191358
Homogeneity	0.978614
Completeness	0.095107
V-measure	0.173366

- Vemos que el ARI es aleatorio, el "agreement" entre los clusters reales y los predecidos.
- El coeficiente de Silhouette nos indica que la clusterización es moderada
- El índice de homogeneidad, completidad y medida V son muy bajos indicandonos la clusterización no hace buen fit a las clases reales.

Se concluye que la clusterización por Hierarchical Clustering no es buena cuando la etiqueta evaluada es la calidad. Los resultados son iguales a los de K-means.

2. **Modelamiento Dataset 2**
    - K-means

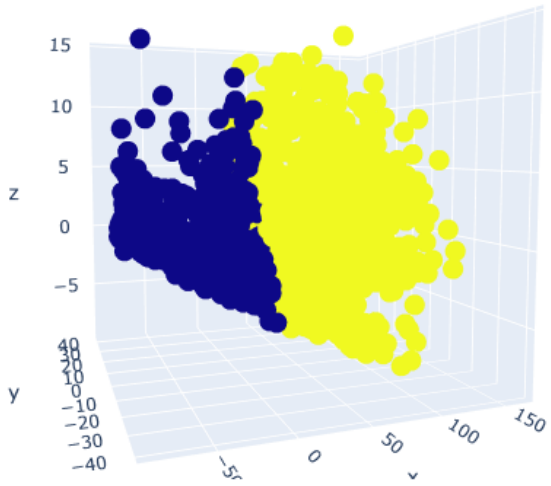


Adjusted Rand Index	0.012579
Silhouette Coefficient	0.191358
Homogeneity	0.978614
Completeness	0.095107
V-measure	0.173366

- El ARI nos indica un agreement relativamente bueno entre los clústeres reales y los predichos.
- El coeficiente de Silhouette nos indica que la clusterización es relativamente buena.
- El índice de homogeneidad, completitud y medida V son relativamente altos sugieren que los clústeres predichos hacen un match relativamente bueno respecto a los reales

En conclusión, observamos que el algoritmo de K-means se comporta relativamente bien para la creación de los clústeres cuando la etiqueta es vino blanco o rojo.

- Hierarchical Clustering

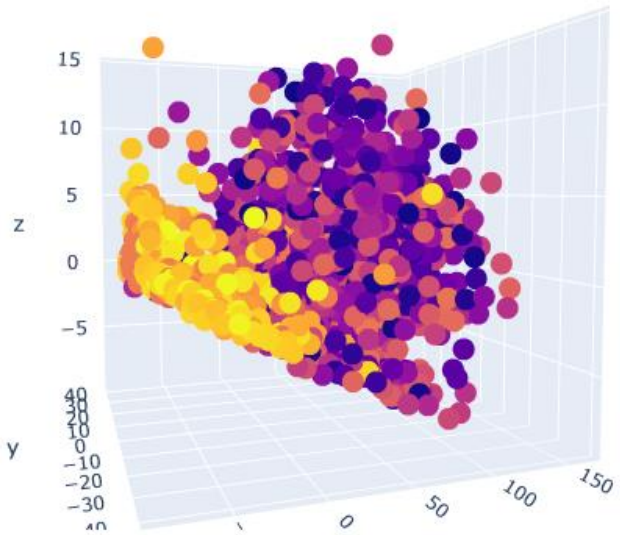


Adjusted Rand Index	0.012579
Silhouette Coefficient	0.191358
Homogeneity	0.978614
Completeness	0.095107
V-measure	0.173366

- El ARI nos indica un agreement algo bueno entre los clústeres reales y los predichos.
- El coeficiente de Silhouette nos indica que la clusterización es relativamente buena.
- El índice de homogeneidad, completitud y medida V son relativamente altos sugieren que los clústeres predichos hacen un match relativamente bueno respecto a los reales

En conclusión, observamos que el algoritmo de Hierarchical Clustering se comporta relativamente bien para la creación de los clusters. Los resultados son similares al K-means pero ligeramente inferiores.

- DBSCAN



Adjusted Rand Index	0.012579
Silhouette Coefficient	0.191358
Homogeneity	0.978614
Completeness	0.095107
V-measure	0.173366

Si bien DBSCAN nos separó los clústeres con una escala de calor donde observamos los clústeres más parecidos a los reales a un lado del espacio, los resultados no son suficientemente buenos comparados a K-means y Hierarchical Clustering. Finalmente, este creo una cantidad muy grande de etiquetas para los clústeres. Una aproximación final podría ser clusterizar nuevamente estas etiquetas con algún algoritmo para determinar si se pueden obtener mejores resultados.

Conclusiones

Las conclusiones de este proyecto son las mismas dadas en el modelamiento y resultados. Vemos que k-means superó al hierarchical clustering en el proceso de clutserización, si bien la diferencia es ligera puede estar dando a un nivel de complejidad

no tan alto en el dataset. Por otro lado, observamos que cuando la etiqueta utilizada es la variable categórica **quality**, el algoritmo de clusterización no da buenos resultados, es decir, no es capaz de encontrar grupos dependiendo de la calidad del vino.

Cuando la etiqueta es cambiada por el color del vino (rojo o blanco) vemos que los algoritmos son capaces de encontrar unos clústeres razonablemente buenos en comparación a los clústeres reales. Si bien DBSCAN fue capaz de separar los clústeres por escalas de calor, creo demasiadas etiquetas, una clusterización nueva podría hacer que este algoritmo arroje resultados interesantes.

Referencias

Kaggle Datasets. (2017). Retrieved from Kaggle: https://www.kaggle.com/datasets/maitree/wine-quality-selection