

# Laboratorio 1: Agrupamiento de vinos según sus características físico-químicas usando Modelos de Mezcla Gaussiana

Franco Cornejo Gonzalez (4 horas)

*Departamento de Ingeniería Informática*

*Universidad de Santiago de Chile*, Santiago, Chile

franco.cornejo.g@usach.cl

**Resumen**—Este estudio investiga el agrupamiento de vinos utilizando Modelos de Mezcla Gaussiana (GMM) para segmentar los datos según sus características físico-químicas. Se empleó el conjunto de datos "Wine Quality", que abarca 1499 observaciones y 11 variables relevantes, con el objetivo de identificar patrones que permitan clasificar los vinos según su calidad.

**Palabras claves**—Modelos de Mezcla Gaussiana

## I. INTRODUCCIÓN

El vino, que antes era visto como un lujo, ahora lo disfruta un público mucho más amplio. La industria vitivinícola está invirtiendo en nuevas tecnologías para mejorar tanto la producción como la venta, donde la certificación y evaluación de calidad son clave. Estas prácticas ayudan a prevenir la adulteración y aseguran que se mantengan altos estándares de calidad, además de facilitar la producción y la fijación de precios (Cortez, Cerdeira, Almeida, Matos, and Reis, 2009).

La certificación incluye pruebas físico-químicas y sensoriales, aunque clasificar el vino es un desafío, ya que depende del trabajo de los enólogos. Pero gracias a los avances en tecnología de la información, se pueden analizar grandes volúmenes de datos, lo que mejora la toma de decisiones en la industria. Las técnicas de minería de datos, como los algoritmos de agrupamiento pueden ser de utilidad para reconocer patrones sobre la calidad de los vinos. Este estudio aplica el agrupamiento basado en modelos, junto a un análisis descriptivo, inferencia y exploratorio del conjunto de observaciones.

Se utiliza el dataset "Wine Quality" que cuenta con 1499 observaciones de distintos vinos con 11 variables numéricas sobre su composición físico-químicas. El dataset fue creado en la Universidad de Minho en Portugal para facilitar la investigación en el área de modelado de preferencias del vino basado en sus propiedades físico-químicas (Cortez et al., 2009).

Finalmente el artículo mostrará los resultados del agrupamiento con sus gráficos,

### I-A. Hipótesis de trabajo

Mediante modelos de mezclas gaussianas pueden segmentarse los vinos por grupos de diferentes calidades

Variable	Importancia
Alcohol	0.487728
Acidez volatil	0.391417
Sulfatos	0.306304
Dioxido de azufre total	0.260437
PH	0.236180
Cloruros	0.233378
Azucar residual	0.099555
Dioxido de azufre libre	0.098745
Acido citrico	0.087811

Tabla I: Importancia de las variables

usando como variables las características físico-químicas de los vinos.

## II. ESTADO DEL ARTE

### II-A. Agrupamiento

El agrupamiento o clustering es una técnica de aprendizaje automático no supervisado que tiene como objetivo organizar un conjunto de datos en grupos o "clústeres" basados en la similitud entre las observaciones de un mismo grupo. El uso de estos métodos es común en análisis genéticos y estudios de mercado (Miranda, 2017).

- **K-means:** Este es uno de los métodos más populares para el agrupamiento. Funciona dividiendo el conjunto de datos en  $k$  clústeres, donde  $k$  es un parámetro que se debe especificar. El algoritmo iterativamente asigna cada punto al clúster más cercano y recalcula los centroides de los clústeres hasta que la asignación de puntos no cambia (Slonim, Aharoni, and Crammer, 2013).
- **DBSCAN:** El algoritmo DBSCAN (Density-based Spatial Clustering of Applications with Noise, o Clustering espacial de aplicaciones con ruido basado en densidad) identifica áreas de alta densidad de puntos que se separan por regiones de baja densidad. A diferencia de K-means, DBSCAN no requiere especificar el número de grupos de antemano y es capaz de encontrar grupos de formas arbitrarias. Este algoritmo clasifica un punto como núcleo, borde o ruido, dependiendo de su densidad de puntos vecinos (Ester, Kriegel, Sander, and Xu, 1996).

- **HDBSCAN:** El algoritmo HDBSCAN (Hierarchical DBSCAN) es una extensión del algoritmo DBSCAN que permite identificar grupos de diferentes densidades. A través de un proceso jerárquico, HDBSCAN construye un árbol de grupos y luego extrae la estructura de grupos más estable en función de la densidad. Este enfoque mejora la identificación de grupos de diferentes densidades (Campello, Moulavi, Zimek, and Sander, 2013).

### III. METODO Y DATOS

#### III-A. Descripción de la base de datos

La base de datos "Wine Quality" fue creada para facilitar la investigación en modelado de preferencias del vino basado en sus propiedades físico-químicas. Este dataset busca servir para aplicar técnicas de minería de datos para predecir la calidad del vino utilizando pruebas analíticas fácilmente accesibles. Incluye información sobre vinos tinto, incorporando tanto pruebas físico-químicas como sensoriales, lo que la convierte en una herramienta útil en la investigación de la enología.

### IV. DESCRIPCIÓN DE LAS VARIABLES

A continuación se describen algunas de las variables más relevantes del dataset:

- **Acidez Fija (fixed acidity):** Ácidos naturales presentes en el vino por ejemplo tartárico, málico, cítrico, succínico y láctico, medidos en gramos por litro.
- **Acidez Volátil (volatile acidity):** Producida durante la fermentación, depende de la actividad de las bacterias lácticas y puede afectar el sabor y estabilidad del vino.
- **Ácido Cítrico (citric acid):** Ácido presente en las uvas.
- **Azúcar Residual (residual sugar):** Cantidad de azúcar que queda después de la fermentación.
- **Cloruros (chlorides):** Anión natural presente en diversas fuentes de agua.
- **Dióxido de Azufre Libre (free sulfur dioxide):** Actúa como conservante en el vino, protegiéndolo.
- **Dióxido de Azufre Total (total sulfur dioxide):** Suma del azufre libre y ligado, el segundo no disponible para actividad antimicrobiana o antioxidante.
- **Densidad (density)**
- **pH:** Indica el nivel de acidez o alcalinidad del vino. Un pH de 7 es neutro.
- **Sulfatos (sulphates):** Aditivo que contribuye a los niveles de dióxido de azufre ( $\text{SO}_2$ ), actuando como antimicrobiano y antioxidante.
- **Alcohol:** Grado alcohólico volumétrico del vino.
- **Calidad (quality):** Variable que mide la calidad del vino, con valores de 0 (muy mala calidad) a 10 (muy buena calidad).

Notar la Tabla III para comprender la frecuencia por calidad.

Nombre de Variable	Tipo	Valores
Fixed Acidity	Numérico	4.6 - 15.9 g/dm <sup>3</sup>
Volatile Acidity	Numérico	0.12 - 1.58 g/dm <sup>3</sup>
Citric Acid	Numérico	0 - 1 g/dm <sup>3</sup>
Residual Sugar	Numérico	0.9 - 15.5 g/dm <sup>3</sup>
Chlorides	Numérico	0.012 - 0.611 g/dm <sup>3</sup>
Free Sulfur Dioxide	Numérico	1 - 68 mg/dm <sup>3</sup>
Total Sulfur Dioxide	Numérico	6 - 289 mg/dm <sup>3</sup>
Density	Numérico	0.99007 - 1.00369 g/cm <sup>3</sup>
pH	Numérico	2.74 - 4.01
Sulphates	Numérico	0.33 - 2.00 g/dm <sup>3</sup>
Alcohol	Numérico	8.4 - 14.9 % vol
Quality	Categórico	0 - 10

Tabla II: Descripción de Variables del Dataset de Calidad de Vino

Calidad	Frecuencia
1	0
2	0
3	10
4	53
5	681
6	638
7	199
8	18
9	0
10	0
<b>Total:</b>	<b>1599</b>

Tabla III: Distribución de la variable por *calidad*

#### IV-A. Análisis estadístico descriptivo

#### IV-B. Método de agrupamiento basado Modelos de Mezcla Gaussiana

El método de agrupamiento basado en modelos destaca en el análisis estadístico y el procesamiento de datos al asumir que los datos provienen de una combinación de distribuciones probabilísticas. Aquí, aplicamos Modelos de Mezcla Gaussiana (GMM) para segmentar los vinos en grupos de calidad basados en sus características físico-químicas.

A diferencia de los métodos de agrupamiento tradicionales como K-means o DBSCAN, los GMM no solo asignan un punto de datos a un clúster específico, sino que estiman la probabilidad de que cada punto pertenezca a diferentes clústeres. Este enfoque asume que cada clúster sigue una distribución gaussiana, lo que le permite capturar formas de clústeres más complejas, como aquellas con estructuras elípticas o de diferentes orientaciones, lo que lo convierte en una opción ideal para datos distribuidos de manera no uniforme o con solapamiento entre grupos.

Selección del Modelo Óptimo Para determinar el número óptimo de clústeres en el conjunto de datos "Wine Quality", aplicamos el Criterio de Información Bayesiana (BIC), que selecciona el modelo con el mejor equilibrio entre ajuste y complejidad. Este criterio se complementa con el índice de Silhouette, que mide la calidad del agrupamiento basado en la cohesión interna de los clústeres y la separación entre ellos. También se analiza como se segmentan las calidades en los diferentes grupos.

Modelo	Volumen	Forma	Orientación	$S_j$	Parámetros
EII	Equal	S		$\lambda I$	I
VII	Variable	S		$\lambda_j I$	k
EEI	Equal	E	Axi-Alg	$\lambda \Lambda$	p
VEI	Variable	E	A-A	$\lambda_j \Lambda_j$	$P+k-1$
EVI	Equal	V	A-A	$\lambda \Lambda_j$	$pK+k+1$
VVI	Variable	V	A-A	$\lambda_j \Lambda_j$	$pK$
EEE	Equal	E	E	$\lambda V_i \Lambda V_j$	$p(p+1)/2$
EEV	Equal	E	V	$\lambda V_i \Lambda V_j$	$kp(p+1)/2-(k-1)p$
VEV	Equal	E	V	$\lambda_i V_i \Lambda V_j$	$kp(p+1)/2-(k-1)(p-1)$
VVV	Variable	V	V	$\lambda_i V_i \Lambda V_j$	$Kp(p+1)/2$

Figura 1: Tipos de modelos

El GMM se caracteriza por parametrizar la matriz de covarianza en tres aspectos geométricos [? ].

- Volumen:** Si es constante entre los grupos o varía.
- Forma:** Si los grupos son esféricos o elípticos.
- Orientación:** Si las orientaciones de los grupos están alineadas o pueden variar.

Estas características permiten que el modelo se ajuste de manera más precisa a la naturaleza de los datos. El GMM puede generar grupos de diferentes formas, tamaños y orientaciones, adaptándose de manera flexible a la complejidad inherente de los datos, esto sumado a las pocas suposiciones que se necesita hacer sobre la distribución de los datos lo hace ideal para trabajar con el dataset de vinos ya que no conocemos las densidades ni la forma de los grupos.

#### IV-C. Análisis estadístico

Primero, se llevó a cabo un análisis descriptivo de las variables del conjunto de datos, el cual se puede consultar en las Tablas X y III del anexo. Se puede notar que las dos variables con mayor dispersión son el dióxido de azufre total y la acidez fija, mientras que las menos dispersas son los cloruros y la densidad. Además, al observar el histograma de la Figura 7, se notan asimetrías en torno a las medias.

Con el objetivo de analizar la separabilidad de las calidades de los vinos, se elaboró un gráfico de cajas por cada calidad de vino para cada variable del conjunto de datos. Esto se puede ver en la Figura 2. Lo primero que destaca es el gran número de valores atípicos. Estos valores atípicos aparecen dentro de una misma calidad de vino, lo que indica que, incluso dentro de una misma categoría, las variables se comportan de forma impredecible. Por otro lado, las cajas están superpuestas entre las calidades, siendo las variables que menos se sobreponen por categoría la cantidad de alcohol, los sulfatos y las variables relativas a la acidez.

Esto último sugiere la posibilidad de algún nivel de correlación o multicolinealidad entre estas variables. Para determinar si existe correlación, se utilizó la matriz de correlación de Kendall, dado que las variables no siguen una distribución normal (véase la Tabla XI del anexo). En la matriz es evidente que las variables relativas a la acidez están correlacionadas entre sí, y las dos variables más correlacionadas son la cantidad de dióxido de azufre libre y total. Estas correlaciones, junto con el conocimiento previo

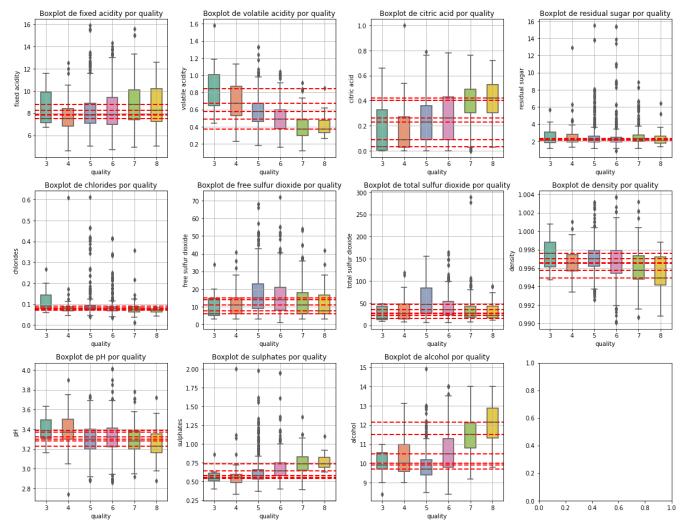


Figura 2: Gráficos de cajas de las variables por calidad.

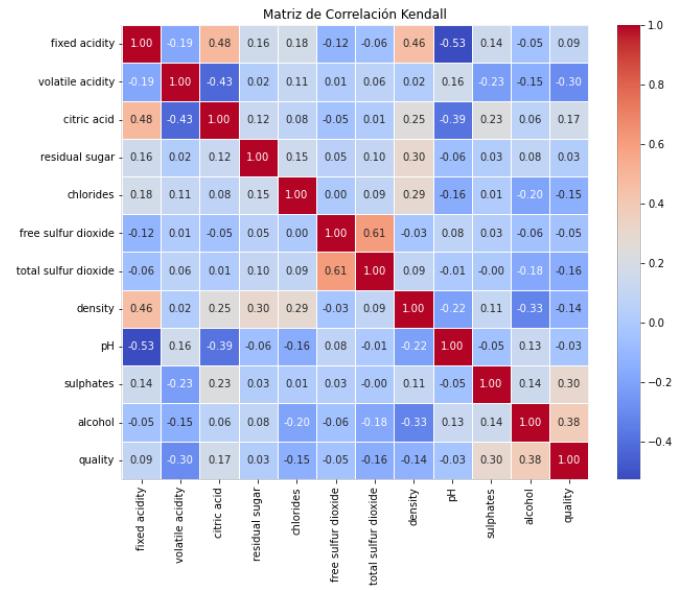


Figura 3: Matriz de correlación de Kendall.

sobre las variables correlacionadas, sugieren que probablemente existe redundancia entre las variables. Por lo tanto, es recomendable utilizar análisis de componentes principales para reducir la dimensionalidad del problema, y desde ya es factible pensar en reducir al menos dos dimensiones.

Para entender mejor la importancia de las variables y compararla con la importancia documentada en la literatura (Figura 9 del anexo), se realizó una Regresión Logística Multivariada (RLM). A diferencia de la regresión logística común, la multivariada no está limitada a clases binarias [6]. Antes de ejecutar la regresión, se detectaron y eliminaron las variables multicolineales mediante el test VIF (Figura 10 del anexo). Se determinó que la densidad y la acidez fija son variables multicolineales, es decir, pueden ser calculadas utilizando otras variables. Esto era esperable, ya que la densidad es la suma de las densidades de las demás

variables, y la acidez fija es un caso similar.

Los resultados obtenidos se muestran en la matriz de confusión de la Figura 8 del anexo, con una **precisión** del 53 %, un **recall** del 32 % y un **F1-Score** de 35 %. Con estos resultados, sabemos que la regresión no fue exitosa a la hora de clasificar, por lo que la Tabla IV no resulta de mucha utilidad a la hora de interpretar la importancia de cada variable.

Tabla IV: Importancia de las variables en la RLM.

Variable	Importancia
Acidez volátil	0.538367
Dióxido de azufre total	0.444212
pH	0.425318
Sulfatos	0.371024
Alcohol	0.357384
Cloruros	0.318430
Ácido cítrico	0.310520
Dióxido de azufre libre	0.221210
Azúcar residual	0.097192

También se hizo el gráfico de parejas (Figura 14, donde cada gráfico es una proyección de los datos en dos dimensiones y los colores representan las clases. Visualmente se puede notar la existencia de grupos por calidad, por ejemplo en la fila de acidez volátil se nota claramente un grupo con los vinos de mala calidad. En otros de los pares como los de la fila de alcohol se ve una especie de transición entre las calidades. El principal problema a la vista es que más que parecer grupos por calidad se nota una especie de gradiente por calidad, donde la mayoría de los datos se encuentran en una zona de trascisión entre calidades. Por ejemplo las calidades 5 y 7 están principalmente mezcladas, perdiéndose la calidad 6 entre estas dos y notando una coaptación clara de las calidades 5 y 7 solo en las zonas mas extremas de los gráficos.

Para finalizar esta sección, se realizó el análisis de componentes principales, cuyos resultados se observan en la Tabla V. Dicho análisis indica que, eliminando las dos variables menos significativas, se pierde menos del 3 % de la varianza. Por lo tanto, el problema se reduce ahora a 9 variables.

Tabla V: Varianza Explicada por CP.

Componente Principal	Varianza Explicada.
CP1	28.17 %
CP2	17.51 %
CP3	14.10 %
CP4	11.03 %
CP5	8.72 %
CP6	6.00 %
CP7	5.31 %
CP8	3.85 %
CP9	3.13 %
CP10	1.65 %
CP11	0.54 %

## V. PREPROCESAMIENTO DE DATOS

### V-A. Eliminar valores atípicos

La eliminación de valores atípicos fue complicada ya que al eliminarlos con métodos como el rango intercuartílico reduce

los calidades extremas en una proporción mucho mayor a las intermedias, dejando el problema aun mas desbalanceado. Para evitar esto el proceso se dividió en 2 etapas.

1. Eliminar datos que escapen de la cuarta desviación estándar, oseas que alguna de sus variables este por fuera del 99.99 %.
2. Usar rango intercuartílico para eliminar solo datos de las categorías 5, 6 y 7.

El resultado puede verse en la Tabla VI

Calidad	Frecuencia
1	0
2	0
3	10
4	51
5	532
6	545
7	155
8	18
9	0
10	0
<b>Total:</b>	<b>1311</b>

Tabla VI: Nueva distribución de la variable por *calidad*.

## VI. RESULTADOS

### VII. PRELIMINARES

Para evaluar la calidad del agrupamiento, en este trabajo se utilizó el GAP del que se hablo en la sección anterior y el **Índice Rand Ajustado (ARI)**, en lugar de otros métodos más comunes como el índice de silueta, debido a que el ARI mide la similitud entre las etiquetas verdaderas y las asignaciones de grupos. Un valor de ARI bajo indica un mal agrupamiento, mientras que un valor cercano a 1 sugiere un buen agrupamiento en relación a las clases (Millo Sánchez, Galpert Cañizares, Casa Cardoso, Grau Ábalos, Arco García, García Lorenzo, and Fernández Marín, 2014).

El uso de las clases originales del conjunto de datos de vinos no dio buenos resultados en los modelos de agrupamiento. Este desempeño pobre se atribuye principalmente al desbalance de clases, con una fuerte concentración de observaciones en las clases intermedias. Para abordar este problema, se realizaron diversas reclasificaciones y cuando una clase era demasiado grande se redujo su tamaño para no desbalancear, agrupando las clases originales de calidad de vino en varias configuraciones:

#### ■ Agrupación en 4 clases:

- Clase 1: 61 observaciones (3 y 4).
- Clase 2: 177 observaciones (5).
- Clase 3: 181 observaciones (6).
- Clase 4: 173 observaciones (7 y 8).

#### ■ Agrupación en 2 clases:

- Baja: 593 observaciones (3, 4 y 5).
- Alta: 718 observaciones (6, 7 y 8).

#### ■ Agrupación en 3 clases:

- Baja: 181 observaciones (3, 4 y 5).
- Media: 197 observaciones (6).
- Alta: 173 observaciones (7 y 8).

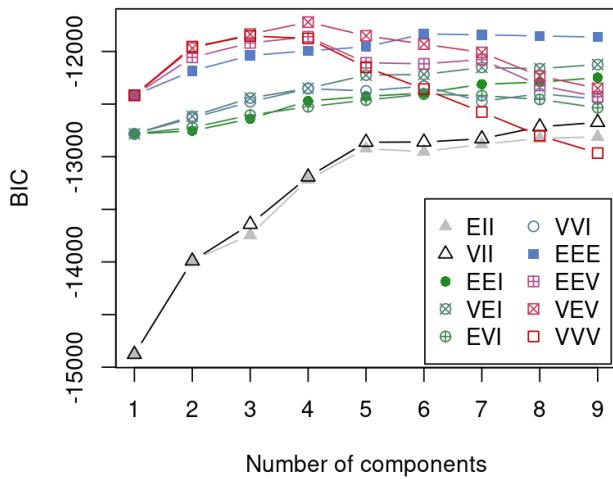


Figura 4: BIC.

Estas reclasificaciones se diseñaron para corregir el desbalance observado en las clases originales y mejorar la capacidad de los modelos de agrupamiento para identificar patrones significativos en los datos. También se separó los vinos en otros dos grupos para experimentar, el primer grupo echo con los mejores y peores vinos y el segundo grupo con los vinos de calidad intermedia.

Tabla VII: Mejores modelos generados.

Clases	Tipo de modelo	GAP	ARI
Originales	VEV	-34617.86	0.0701
4	VEV	-11721.65	0.0839
3	VEV	-23645.94	0.059
2	VEV	-33287.17	0.0735
Solo Buenos y malos	VEV	-6477.822	0.0922
Medios	VEV	-28406.72	0.0416

A partir de la comparación de la Tabla VII y experimentación se decidió continuar el trabajo con 4 clases. Notar la Figura 13 del Anexo.

## VIII. RESULTADOS

Para llevar a cabo el agrupamiento final se comenzó usando la función “mclustBIC()” de la librería “mclust”, con esto se obtuvieron las 3 mejores modelos basados en la forma, orientación y volumen. Estos 3 modelos son los con mejor BIC(criterio de información bayesiana) que se define según la complejidad de los modelos. En la Figura 4 vemos el BIC en el “eje y” y la cantidad de grupos en el “eje x”. Cada curva tiene un color que determina el tipo de modelo que se está usando.

El mejor de los modelos según el BIC puede verse en la Figura 11 y la Tabla XII del anexo. Este modelo presentó un ARI 0.0839 y como el objetivo es verificar la hipótesis se continuó probando los demás modelos hasta dar con el quinto mejor que fue el de tipo EVI con un ARI de 0.1424, siendo

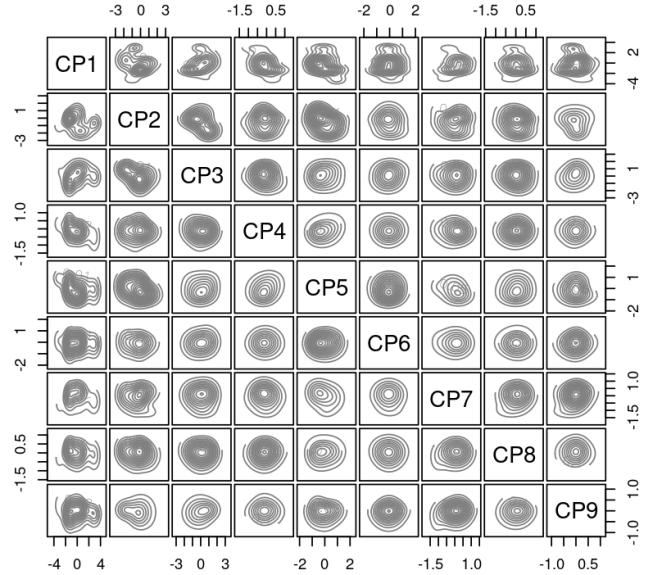


Figura 5: Gráfico asociado a los grupos y a la distribución de probabilidad respectiva en 2D

el mismo el modelo que mejor relacionó las clases con los grupos, la Tabla VIII y se aprecia una baja segmentación de las clases en los grupos a pesar de ser el mejor modelo encontrado. Sobre los grupos podemos ver la Figura 5 donde se ven las

Clase	G 1	G 2	G 3	G 4	G 5	G 6
Clase 1	4	22	12	6	1	3
Clase 2	5	56	90	4	10	11
Clase 3	7	50	52	9	34	32
Clase 4	11	21	12	12	110	23

Tabla VIII: Distribución de clases por categorías modelo EVI.

curvas de nivel asociadas a las funciones de probabilidad y la figura 6 donde se ven los grupos asociados a un color.

Finalmente se decide rechazar la hipótesis ya que no se fue posible segmentar las calidades en los grupos usando MMG.

## IX. CONCLUSIONES

Este estudio ha intentado segmentar los vinos en diferentes grupos de calidad mediante el uso GMM. Sin embargo, los resultados obtenidos no lograron alcanzar el objetivo de una segmentación efectiva. A pesar de los esfuerzos por aplicar diversas estrategias de reclasificación y análisis, se evidenció que la estructura de los datos presentaba desafíos significativos, como un fuerte desbalance en las clases y una notable superposición entre las diferentes categorías de calidad.

El análisis reveló que muchos vinos de calidades intermedias compartían características similares, dificultando la identificación clara de grupos diferenciados. Además igual se recentraron malos resultados al eliminar por completo los vinos de calidad intermedia dejando solo buenos y malos vinos. Todo esto se tradujo en una baja precisión en la clasificación y una baja correlación entre las clases originales y los grupos formados, lo que indica que los GMM, aunque

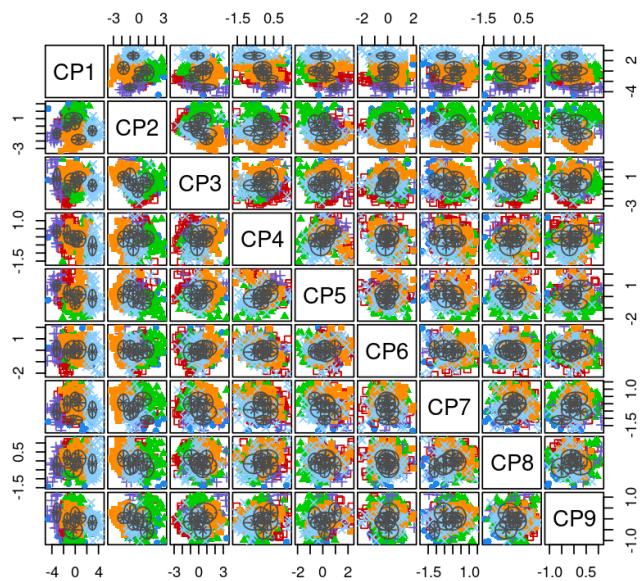


Figura 6: Gráfico agrupamiento del modelo EVI.

potentes en la captura de complejidades en la estructura de los datos, no fueron capaces de descomponer adecuadamente la calidad del vino en este contexto específico.

Dada la complejidad inherente a la calidad del vino, es crucial considerar que otros métodos de agrupamiento podrían ser más adecuados. En particular, técnicas basadas en densidad, como DBSCAN o HDBSCAN, o enfoques jerárquicos, podrían ofrecer una mejor adaptabilidad a la variabilidad y solapamiento en los datos. Además, el uso de técnicas de reducción de dimensionalidad podría ser explorado para simplificar la estructura de los datos y facilitar una mejor segmentación.

En conclusión, aunque se realizaron esfuerzos significativos en el análisis de agrupamiento, se recomienda la exploración de otros métodos que podrían ayudar a proporcionar una mejor comprensión de las calidades del vino y una segmentación más efectiva, abriendo así nuevas avenidas para investigaciones futuras en el campo de la enología.

#### REFERENCIAS

- [1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, “Modeling wine preferences by data mining from physicochemical properties,” *Department of Information Systems/R&D Centre Algoritmi, University of Minho*, 2009, viticulture Commission of the Vinho Verde Region (CVRVV), 4050-501 Porto, Portugal.
- [2] P. M. Miranda, “Uso de algoritmos de aprendizaje automático aplicados a bases de datos genéticos,” Master’s thesis, Máster en Estadística y Bioinformática, May 2017, programación para la Bioinformática, Pau Andrio Balado.
- [3] N. Slonim, E. Aharoni, and K. Crammer, “Hartigan’s k-means versus Lloyd’s k-means – is it time for a change?” in *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

- [4] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*. Portland, OR: AAAI Press, 1996, pp. 226–231.
- [5] R. J. G. B. Campello, D. Moulavi, A. Zimek, and J. Sander, “A framework for semi-supervised and unsupervised optimal extraction of clusters from hierarchies,” *Data Mining and Knowledge Discovery*, vol. 27, no. 3, pp. 344–371, April 2013.
- [6] J. Pérez, “La regresión logística como modelo de predicción del riesgo crediticio en las organizaciones de la economía social y solidaria,” *Universidad Internacional del Ecuador*, 2017, recibido: 06/01/2017. [Online]. Available: URLdelartculo
- [7] R. Millo Sánchez, D. Galpert Cañizares, G. Casa Cardoso, R. Grau Ábalos, L. Arco García, M. M. García Lorenzo, and M. Fernández Marin, “Agregación de medidas de similitud para la detección de ortólogos: validación con medidas basadas en la teoría de conjuntos aproximados,” *Computación y Sistemas*, vol. 18, no. 1, pp. 19–35, 2014.

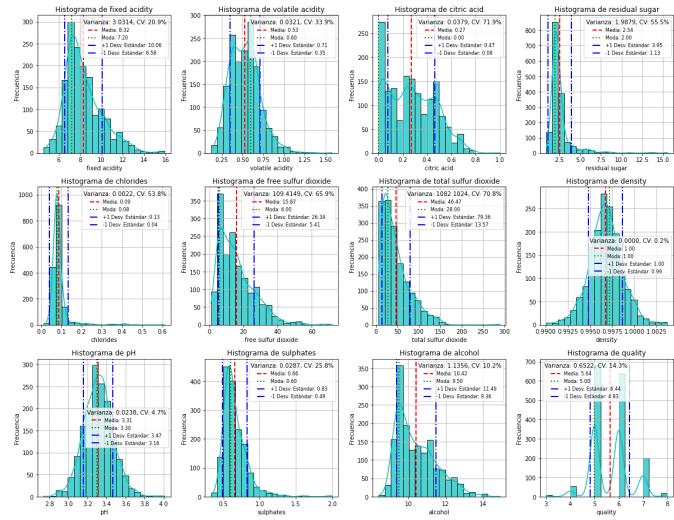


Figura 7: Histogramas.

## ANEXO

Tabla IX: Estadísticas descriptivas (Medidas de tendencia central).

Variable	Mínimo	Máximo	Q1	Mediana	Q3	Media
Fixed Acidity	4.6	15.9	7.1	7.9	9.1	8.311
Volatile Acidity	0.12	1.58	0.3925	0.52	0.64	0.5313
Citric Acid	0	1.0	0.09	0.25	0.42	0.268
Residual Sugar	0.9	15.5	1.9	2.2	2.6	2.532
Chlorides	0.012	0.611	0.07	0.079	0.09	0.0869
Free Sulfur Dioxide	1	72	13	21	32	15.615
Total Sulfur Dioxide	6	289	37	67	145	115.915
Density	0.99	1.004	0.996	0.996	0.997	0.9967
pH	2.74	4.01	3.205	3.31	3.4	3.311
Sulphates	0.33	2.00	0.55	0.62	0.73	0.658
Alcohol	8.4	14.9	9.5	10.2	11.1	10.442
Quality	3	8	5	6	6	5.657

Tabla X: Estadísticas descriptivas (Medidas de dispersión).

Variable	Desviación Estándar	Rango	Rango Intercuartilico.
Fixed Acidity	1.748	11.3	2
Volatile Acidity	0.1796	1.46	0.2475
Citric Acid	0.1967	1	0.33
Residual Sugar	1.356	14.6	0.7
Chlorides	0.0473	0.599	0.02
Free Sulfur Dioxide	10.250	67	14
Total Sulfur Dioxide	32.782	283	40
Density	0.0019	0.0136	0.0023
pH	0.1566	1.27	0.195
Sulphates	0.1704	1.67	0.18
Alcohol	1.082	6.5	1.6
Quality	0.8065	5	1

Tabla XI: Resultados del Test de Shapiro-Wilk.

Nombre de la Variable	p-valor	Distribución Normal
Fixed Acidity	1.5256e-24	No
Volatile Acidity	2.6868e-16	No
Citric Acid	1.0208e-21	No
Residual Sugar	0.000000e+00	No
Chlorides	0.000000e+00	No
Free Sulfur Dioxide	7.6974e-31	No
Total Sulfur Dioxide	3.5741e-34	No
Density	1.9401e-08	No
pH	1.7137e-06	No
Sulphates	5.8216e-38	No
Alcohol	6.6437e-27	No
Quality	9.5040e-36	No

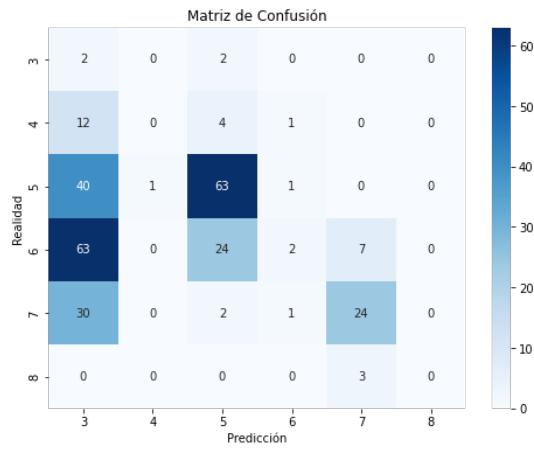


Figura 8: Matriz Confusión RLM.

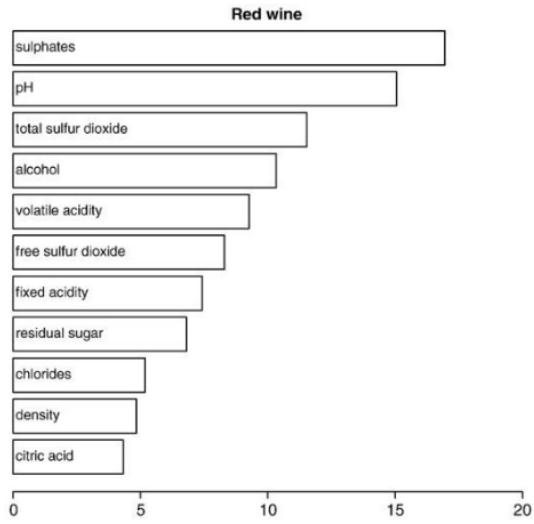


Figura 9: Importancia de las variables según la literatura[1].

```
VIF > 10
      feature          VIF
0   const  1.710538e+06
VIF > 5
      feature          VIF
0   const  1.710538e+06
1   fixed acidity 7.767512e+00
8   density  6.343760e+00
```

Figura 10: Test VIF.

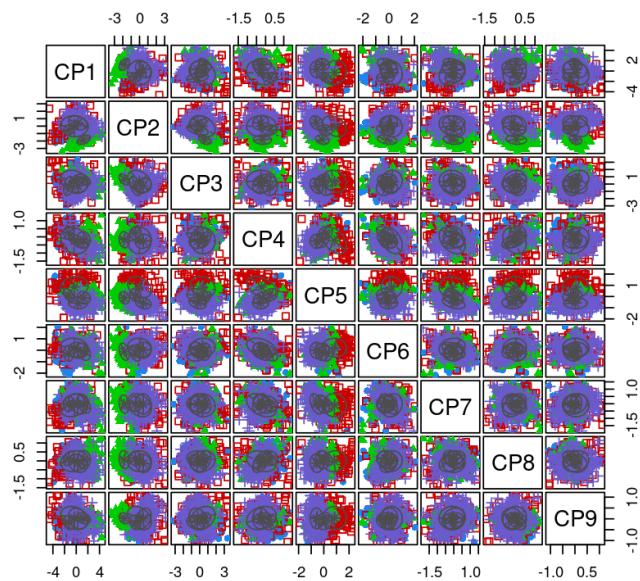


Figura 11: Modelo con mejor GAP.

Tabla XII: Distribución de clases por grupo mejor GAP.

Clase	Grupo 1	Grupo 2	Grupo 3	Grupo 4
Clase 1	15	15	1	17
Clase 2	49	20	12	95
Clase 3	61	24	34	65
Clase 4	49	35	86	19

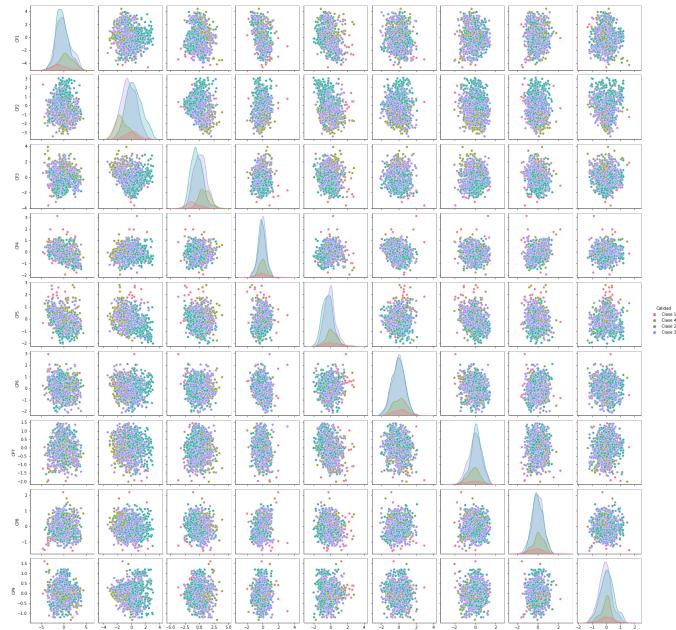


Figura 13: Gráfico de parejas, con 4 clases y componentes principales.

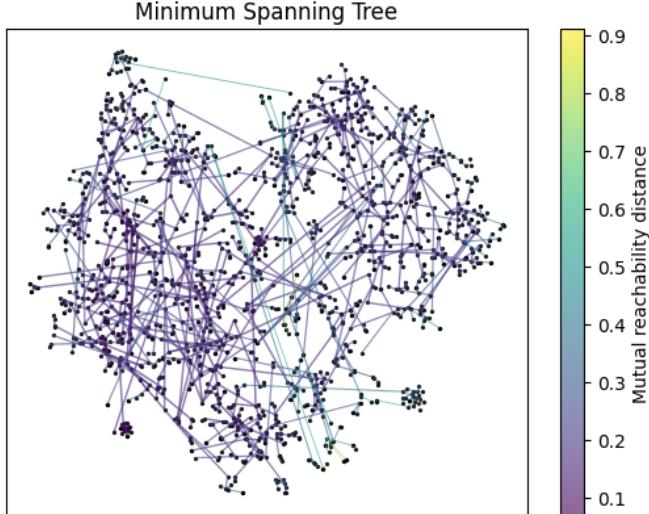


Figura 12: Agrupamiento vinos con HDBSCAN

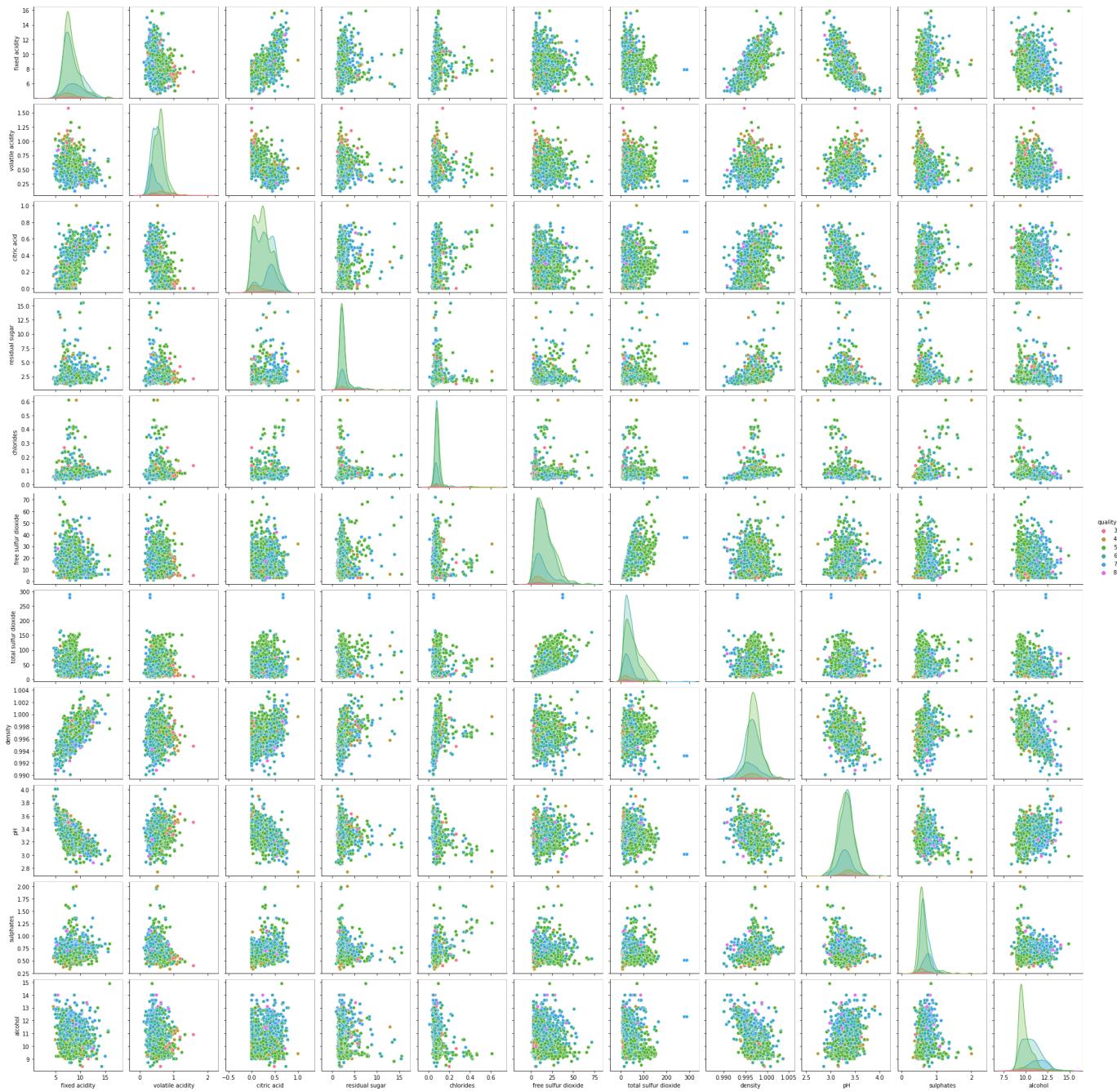


Figura 14: Gráfico de parejas conjunto de datos original con clases originales/