Universidad Nacional Guillermo Brown

Trabajo Práctico Nº 1

Inferencia estadística y reconocimiento de patrones

Sosa Matias Luis
Barreto Nicolas

Fecha de entrega: 1/10/2025

Licenciatura en Ciencia de Datos
Tecnicatura en Programación

Profesor: Statti Florencia

Índice

1. INTRODUCCIÓN

1.1. Introducción

En este trabajo se analizará el conjunto de datos `winequality-red.csv`, el cual contiene información sobre características fisicoquímicas y sensoriales de vinos tintos de Portugal. El objetivo principal es predecir la variable `quality` (calidad del vino) utilizando distintos métodos de regresión y clasificación, y comparar su desempeño.

1.2. Objetivo

El objetivo es aplicar y comparar diferentes técnicas estadísticas y de machine learning para predecir la calidad del vino, evaluando tanto modelos de regresión como de clasificación, y seleccionar el método más adecuado para el conjunto de datos.

1.3. Descripción del dataset

El dataset winequality-red.csv contiene 12 variables: 11 predictoras fisicoquímicas y sensoriales, y una variable respuesta (quality) que representa la calidad del vino tinto en una escala discreta. Cabe aclarar explícitamente que **quality** es una variable ordinal de 6 categorías (3 a 8). El conjunto de datos incluye más de 1.500 observaciones.

El dataset contiene las siguientes variables:

- `***fixed.acidity***`
- `***volatile.acidity***`
- `***citric.acid***`
- `***residual.sugar***`
- `***chlorides***`
- `***free.sulfur.dioxide***`
- `***total.sulfur.dioxide***`
- `***density***`
- `***pH***`
- `***sulphates***`

- `alcohol`

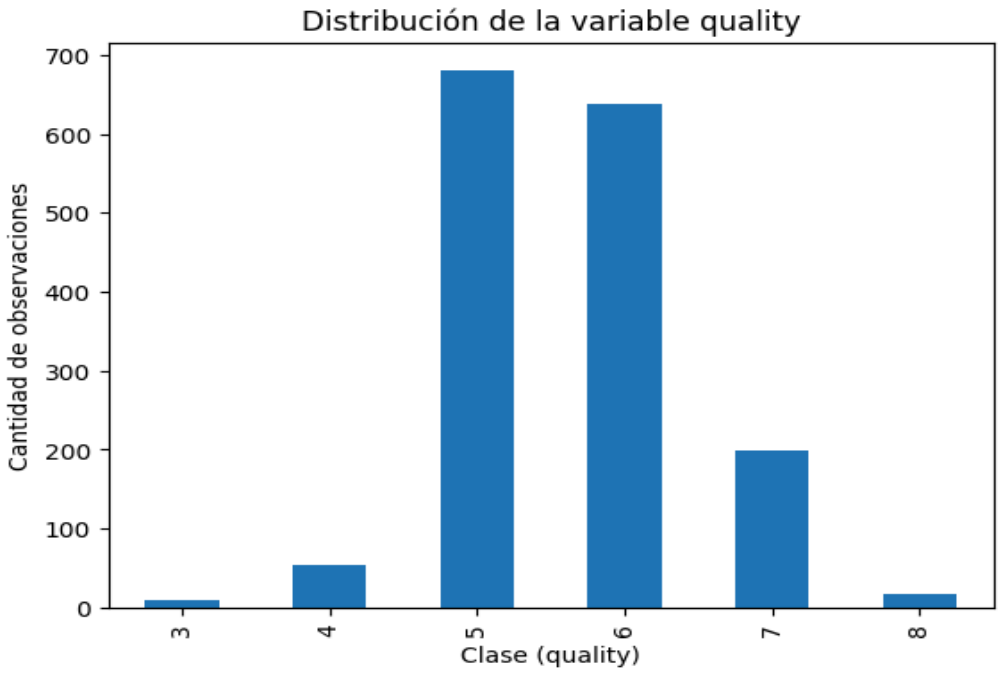
- `quality` “variable respuesta”

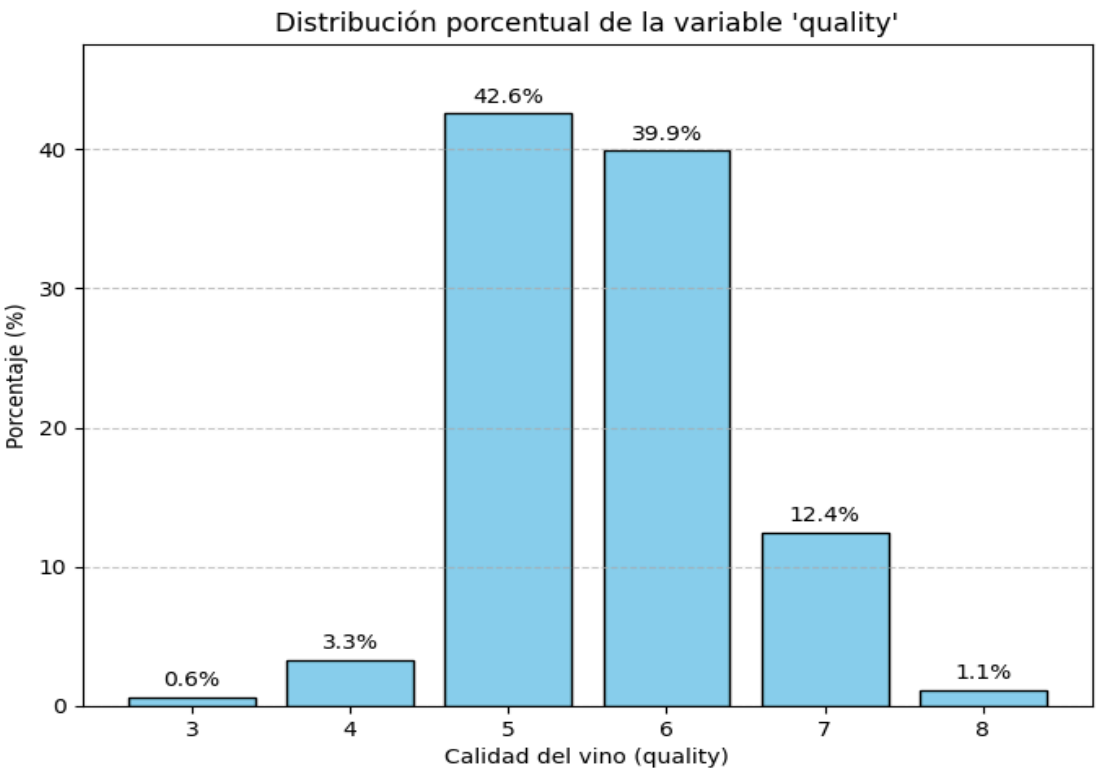
1.4. Distribución del dataset

El dataset contiene 1599 distribuidas de la siguiente manera:

Quality	Frecuencia
3	10
4	53
5	681
6	638
7	199
8	18

Como puede observarse, el dataset está **fuertemente desbalanceado**: las clases **5 (681 vinos)** y **6 (638 vinos)** concentran más del **82% de las observaciones**, mientras que las clases extremas (3, 4 y 8) son muy poco frecuentes, con menos del 5% en conjunto, más adelante será necesario tener en cuenta este desbalance.




Distribución porcentual de la variable 'quality'
Porcentaje (%)
0
10
20
30
40
0.6%
3.3%
42.6%
39.9%
12.4%
1.1%
3
4
5
6
7
8
Calidad del vino (quality)

2. Metodologías

En este trabajo se aplicarán y compararán distintos métodos de regresión y clasificación para predecir la variable ´**quality**´.

2.1. Métodos de regresión

- **Regresión lineal múltiple**: Permite modelar la relación lineal entre la variable respuesta y un conjunto de variables predictoras.
- **Ridge:** Variante de la regresión lineal que incorpora regularización L2 para evitar el sobreajuste.
- **LASSO:** Similar a Ridge, pero utiliza regularización L1, lo que permite realizar selección de variables.

2.2. Métodos de Clasificación

- **Regresión logística multinomial:** Extensión de la regresión logística utilizada cuando la variable respuesta puede tomar más de dos categorías.
- **K vecinos más cercanos (KNN):** Clasifica una observación según la mayoría de sus vecinos más cercanos en el espacio de variables.
- **Bayes ingenuo:** Clasificador probabilístico basado en el teorema de Bayes y la independencia entre variables.
- **LDA (Análisis discriminante lineal):** Busca una combinación lineal de variables que mejor separe las clases.
- **QDA (Análisis discriminante cuadrático):** Similar a LDA, pero permite matrices de covarianza diferentes para cada clase.

Para cada método se realizará:

- Ajuste del modelo y selección de parámetros (cuando corresponda, mediante validación cruzada).

- Evaluación de métricas de desempeño.

- Comparación de resultados para seleccionar el mejor modelo de regresión y el mejor de clasificación.

3. Aplicación de los distintos métodos

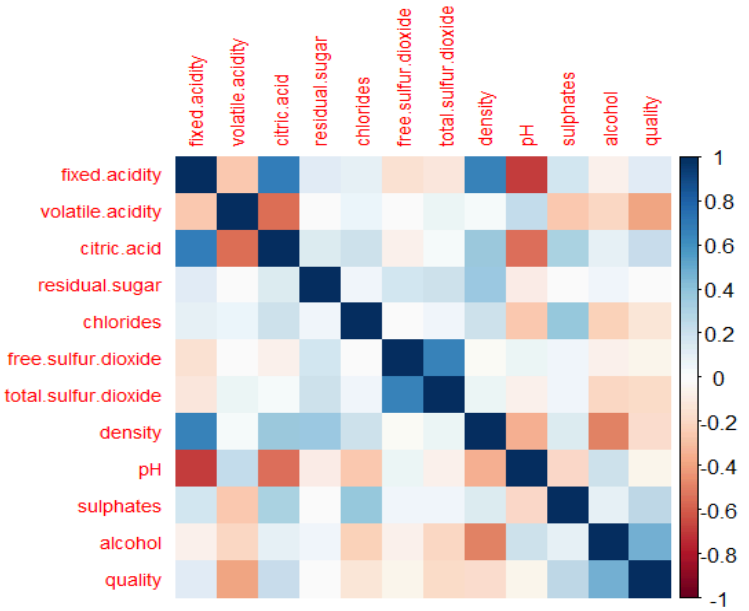
En esta sección se realizará una exploración del dataset para comprender la distribución de las variables, detectar posibles valores atípicos y analizar la relación entre las variables predictoras y la variable respuesta `quality`.

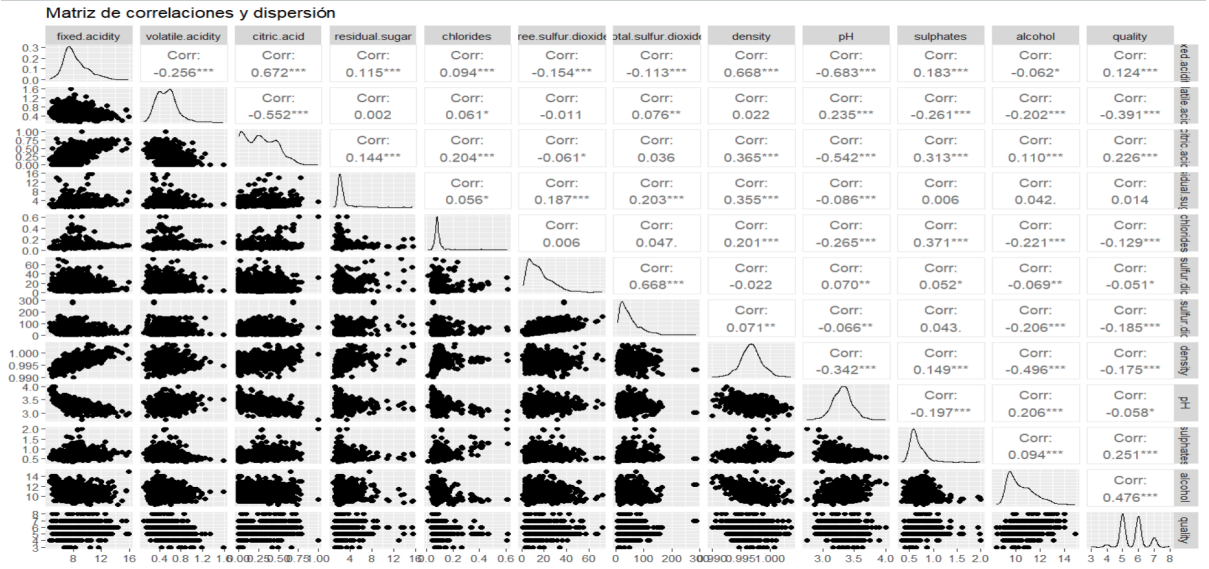
Se incluirán:

- Estadísticas descriptivas de las variables.

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956
##  Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037
##        pH          sulphates         alcohol         quality
##  Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000
##  1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.310   Median :0.6200   Median :10.20   Median :6.000
##  Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

- Visualizaciones como histogramas, diagramas de dispersión y gráficos de correlación.





El análisis de correlaciones muestra que la variable que más se asocia con la calidad del vino es el **alcohol**, con una correlación positiva relativamente fuerte: los vinos con mayor graduación suelen recibir mejores valoraciones. En segundo lugar aparecen los **sulphates** y el **citric acid**, también con correlaciones positivas, aunque más débiles, lo que indica que estos componentes pueden contribuir a percibir un vino de mayor calidad. Por el lado contrario, la **volatile acidity** tiene una correlación negativa bastante marcada: cuanto más alta es, peor tiende a ser la calidad.

El resto de las variables presentan relaciones muy débiles o casi nulas con la calidad (residual sugar, pH, chlorides, free sulfur dioxide), por lo que por sí solas no aportan mucho a la predicción.

3.1. Regresión

Utilizamos una división del dataset en un 80% para entrenamiento y un 20% para prueba, implementada con la función sample.split() del paquete caTools.

Para las validaciones cruzadas en Ridge y lasso utilizamos 10 folds con cv.glmnet() para seleccionar el mejor lambda.

Para evaluar los modelos utilizamos el error cuadrático medio (MSE) y R^2

3.1.1. Regresion Lineal Multiple

Luego de realizar el modelo de regresión lineal múltiple sobre el conjunto de entrenamiento, ejecutamos summary() y obtuvimos:

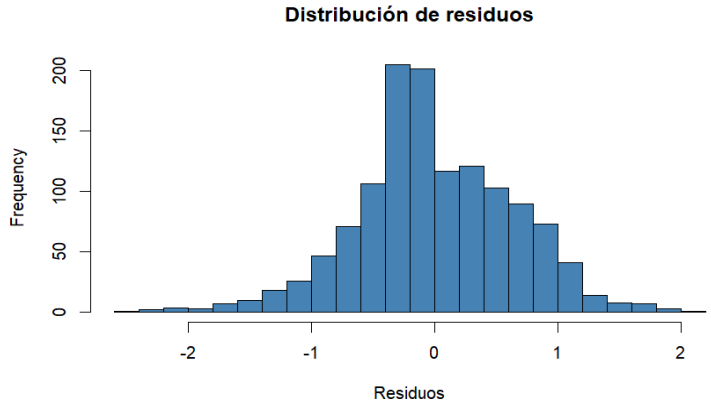
```
Call:
lm(formula = quality ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-2.52213 -0.35627 -0.04738  0.44215  2.00517

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.342e+01  2.323e+01   1.008 0.313441
fixed.acidity         2.725e-02  2.862e-02   0.952 0.341210
volatile.acidity     -1.032e+00  1.353e-01  -7.625 4.77e-14 ***
citric.acid          -1.857e-01  1.626e-01  -1.142 0.253523
residual.sugar        2.223e-02  1.623e-02   1.370 0.170964
chlorides            -1.456e+00  4.849e-01  -3.002 0.002731 **
free.sulfur.dioxide   4.483e-03  2.452e-03   1.828 0.067725 .
total.sulfur.dioxide -3.086e-03  8.191e-04  -3.767 0.000173 ***
density              -1.976e+01  2.372e+01  -0.833 0.404986
pH                   -3.457e-01  2.165e-01  -1.597 0.110515
sulphates             9.192e-01  1.257e-01   7.311 4.69e-13 ***
alcohol               2.834e-01  2.909e-02   9.741  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6474 on 1267 degrees of freedom
Multiple R-squared:  0.3536,    Adjusted R-squared:  0.348
F-statistic: 63.02 on 11 and 1267 DF,  p-value: < 2.2e-16
```

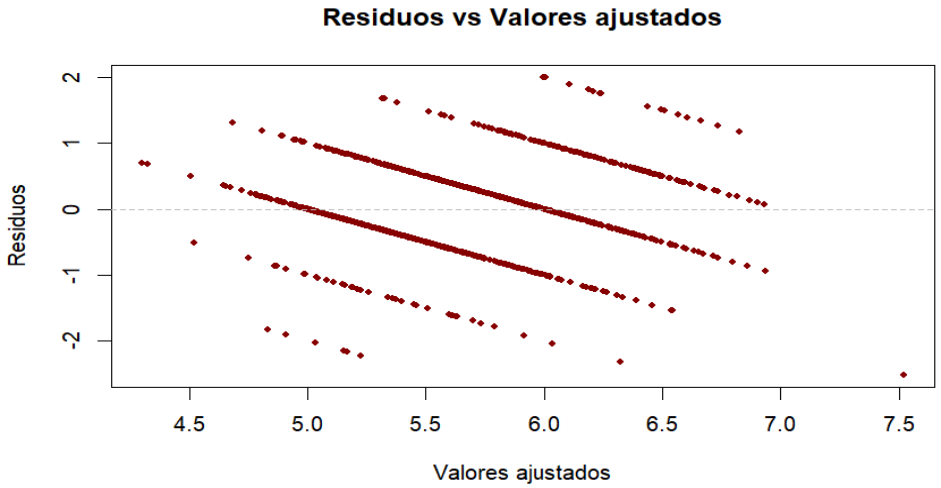
Se puede observar que el modelo es estadísticamente significativo ya que el p-valor es p < 2.2e-16, lo que indica que al menos una variable predictora tiene asociación con la calidad del vino. El R^2 es de 0.3536, lo que implica que el modelo explica aproximadamente el 35% de la variabilidad en la respuesta.

Además, analizando la columna Pr(>|t|), observamos que las variables más significativas son alcohol, volatile y sulphates y las menos significativas son: fixed.acidity, citric.acid, residual.sugar, density, pH y free.sulfur.dioxide

Para evaluar el modelo, averiguamos el error cuadrático medio (SME), y nos dió un valor de 0.4264839.

Por último, hicimos un histograma de los residuos e hicimos un gráfico de Residuos vs Valores Ajustados que se pueden ver a continuación:



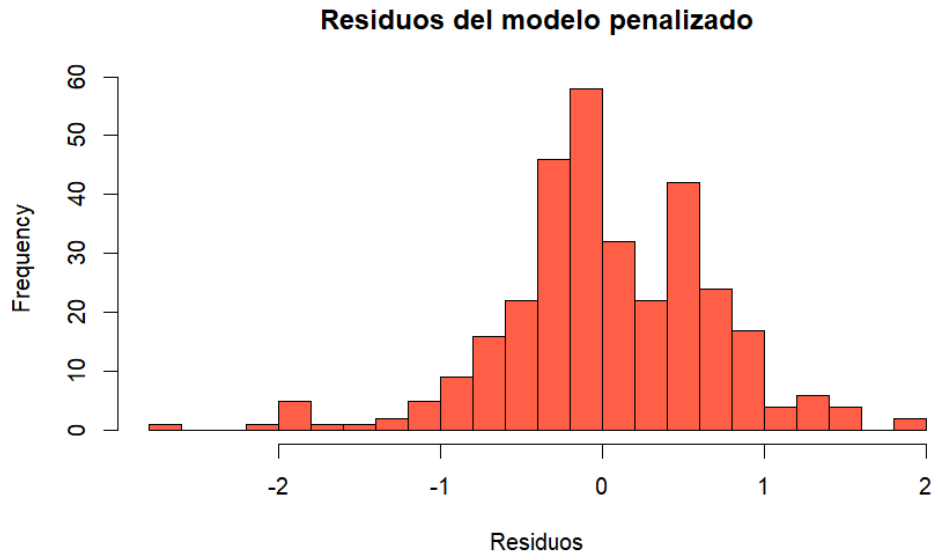


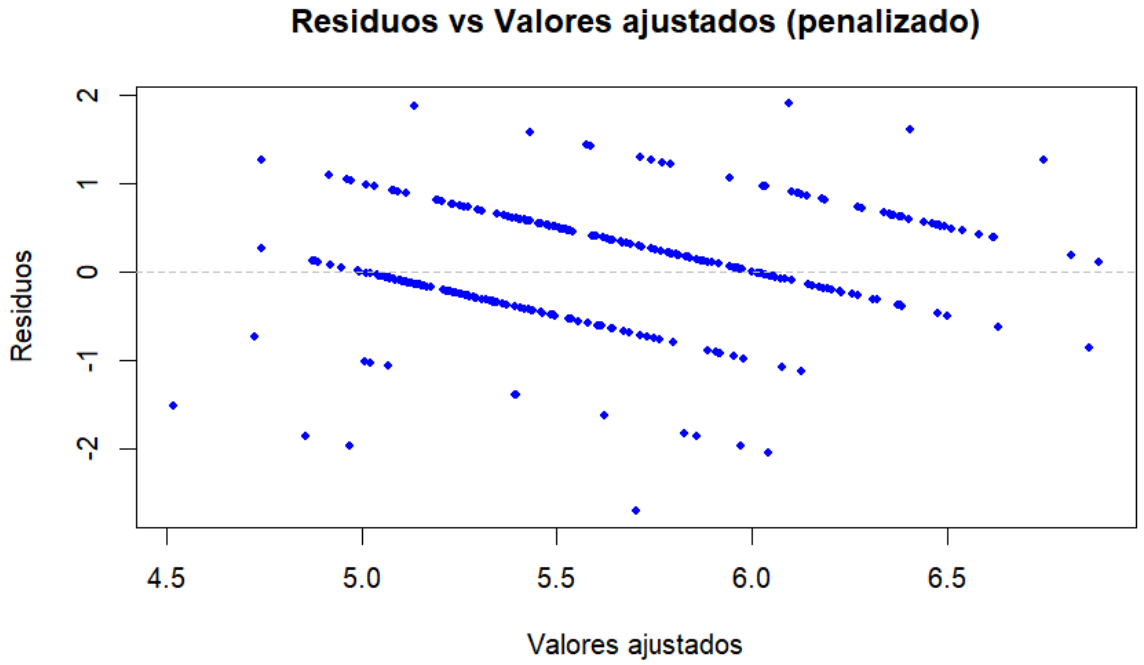
En el primer gráfico podemos ver que los residuos siguen una distribución aproximadamente normal, lo cual es bueno ya que es uno de los supuestos de este modelo y el segundo gráfico muestra una distribución aleatoria, sin patrones evidentes.

En el segundo gráfico podemos ver que hay cierta simetría de los residuos alrededor del 0. Además, se puede ver que los puntos se agrupan en forma de 6 rectas. Esto es porque la variable respuesta solo toma 6 valores enteros: 3,4,5,6,7 y 8.

3.1.2 Ridge

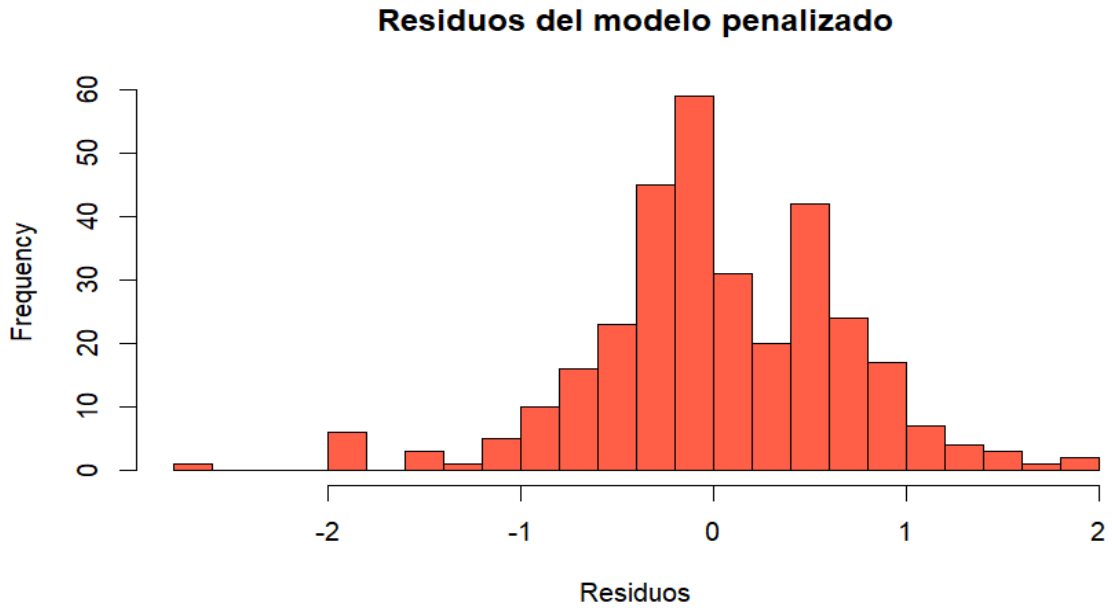
Luego de la validación cruzada, el lambda seleccionado fue `lambda.1se`, priorizando simplicidad sin perder rendimiento. El modelo obtuvo un MSE de 0.4278 y un R^2 de 0.3792.. Las variables más influyentes fueron density, chlorides y volatile.acidity. Los gráficos de los residuos presentan bastantes similitudes con los de cuadrados mínimos.

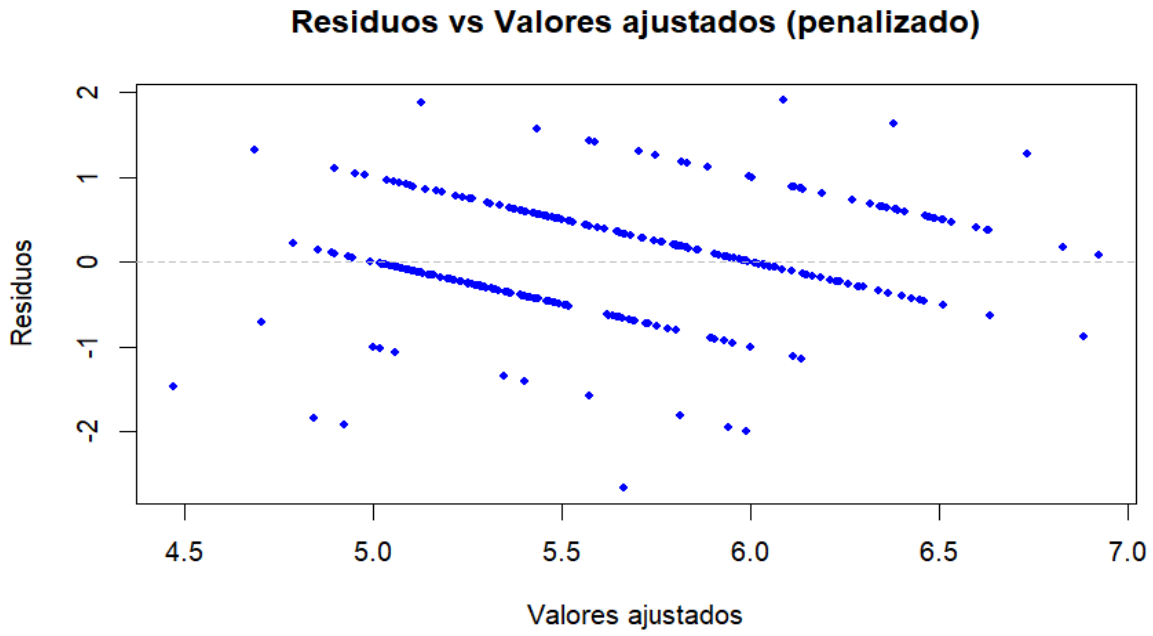




3.1.3. Lasso

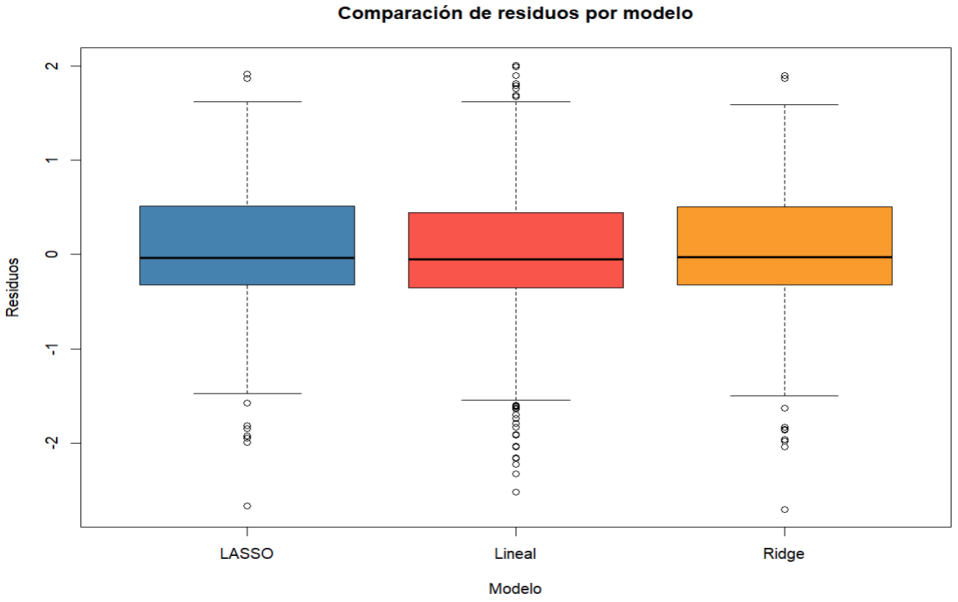
De igual forma que Ridge, se seleccionó `lambda.1`. El rendimiento sobre el conjunto de prueba fue MSE = 0.4304 y R^2 = 0.3754, similar al obtenido con Ridge. LASSO eliminó varias variables, conservando solo volatile.acidity (-0.79), sulphates (0.29) y alcohol (0.24) como predictoras relevantes. Los gráficos de los residuos siguen siendo similares.





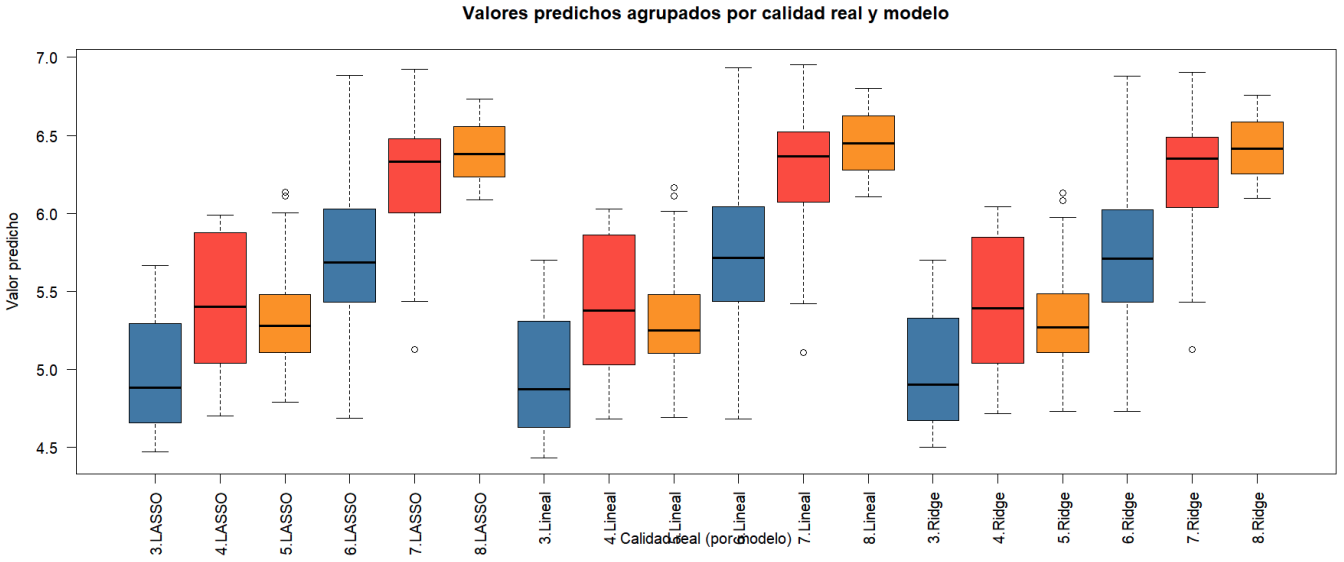
3.1.4 Boxplots

A continuación se presentan diferentes boxplots que permiten comparar visualmente el desempeño y las características de los modelos de regresión lineal múltiple, Ridge y LASSO. Estos gráficos facilitan la evaluación de la dispersión de los residuos, la distribución de los valores predichos y la magnitud de los errores, brindando una perspectiva complementaria a las métricas numéricas tradicionales.



Boxplot 1: Comparación de residuos por modelo

Este gráfico permite comparar la dispersión y simetría de los residuos de los modelos de regresión lineal múltiple, Ridge y LASSO. Se observa que los residuos de los tres modelos presentan una distribución similar, aunque los modelos penalizados (Ridge y LASSO) tienden a mostrar una leve reducción en la dispersión respecto al modelo lineal clásico. Esto sugiere que la regularización ayuda a controlar los errores extremos, aunque la diferencia no es drástica.



Boxplot 2: Valores predichos agrupados por calidad real y modelo

Este boxplot permite analizar cómo cada modelo predice la calidad para cada valor real observado. Se puede observar si algún modelo tiende a sobreestimar o subestimar la calidad en ciertos rangos, o si existe mayor dispersión en las predicciones para

determinados valores reales. En general, los tres modelos muestran una tendencia a concentrar las predicciones cerca de los valores reales, aunque puede haber mayor dispersión en los extremos de la escala de calidad. Esto sugiere que los modelos funcionan mejor en el rango central de la variable respuesta y pueden tener dificultades para predecir valores de calidad muy bajos o muy altos.

3.1.5 Conclusiones parciales

En resumen, los tres modelos presentan métricas similares. Se selecciona Ridge como el más adecuado, ya que alcanza el mayor valor de R^2. Aunque su ECM (0.4278) es levemente superior al de la regresión lineal múltiple, Ridge ofrece una ventaja adicional: penaliza la magnitud de los coeficientes, lo que contribuye a reducir el sobreajuste y mejorar la estabilidad del modelo ante nuevas observaciones.

Modelo	**R^2**	**ECM (MSE)**
Regresión lineal múltiple	0.3536	0.4265
Ridge	0.3792	0.4278
LASSO	0.3754	0.4304

3.2. Clasificación

El dataset utilizado para estos modelos se dividió en un 80% para entrenamiento y un 20% para prueba utilizando la función train_test_split().

También, se estandarizan todos los modelos mediante StandardScaler para mejorar el desempeño y la estabilidad de los mismos.

Para mitigar el fuerte desbalance de clases en la variable quality, aplicamos RandomOverSampler, que genera un conjunto de entrenamiento balanceado al aumentar aleatoriamente las observaciones de las clases minoritarias. Esto permitió mejorar la capacidad del modelo para reconocer vinos de calidad 3, 4 y 8, incrementando el recall y el F1-score macro. Sin embargo, esta técnica puede introducir cierto riesgo de sobreajuste, ya que los nuevos ejemplos son réplicas de observaciones existentes.

Para la evaluación del desempeño de los modelos se utilizaron las métricas **Accuracy, Precisión, Recall, F1-score macro** y el **AUC(área bajo la curva ROC)**. Estas métricas permiten analizar no sólo la proporción de aciertos globales, sino también la capacidad del modelo para identificar correctamente las clases minoritarias. En particular, el uso de la versión *macro* asegura que todas las clases tengan el mismo peso en la evaluación, lo que resulta especialmente relevante en contextos con fuerte desbalance de clases como en la variable *quality*.

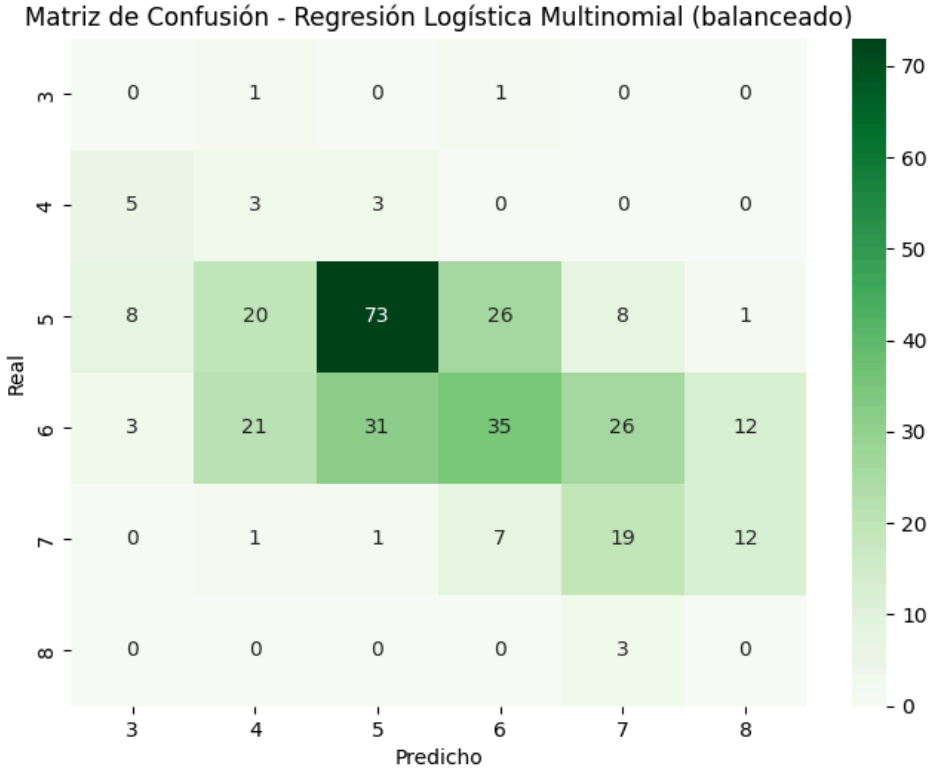
Utilizamos cross_val_Score para que divida los datos en varios “folds”(por defecto 5), entrenando el modelo en algunos folds y lo evalúa en los restantes, repitiendo el proceso para todos los folds. Así, obtienes una estimación más robusta del desempeño del modelo y detectar sobreajuste.

3.2.1 Regresión Logística Multinomial

Se ajustó un modelo de Regresión Logística Multinomial. Las métricas que obtuvimos al entrenarlo fueron las siguientes:

Métrica de evaluación	No balanceado	Balanceado
Accuracy	**0.590625**	**0.40625**
Accuracy medio CV	**0.5973192401960784**	**0.5923547400611622**
Precisión	**0.3116984215069676**	**0.26461256805459704**
Recall	**0.2728860294117647**	**0.2596549131016043**
F1-score	**0.27762808067026984**	**0.24246451603405894**
AUC	**0.7639904835106219**	**0.692943495071285**

Las matrices de confusión que se obtuvieron fueron las siguientes:



3.2.2 KNN(vecinos más cercanos)

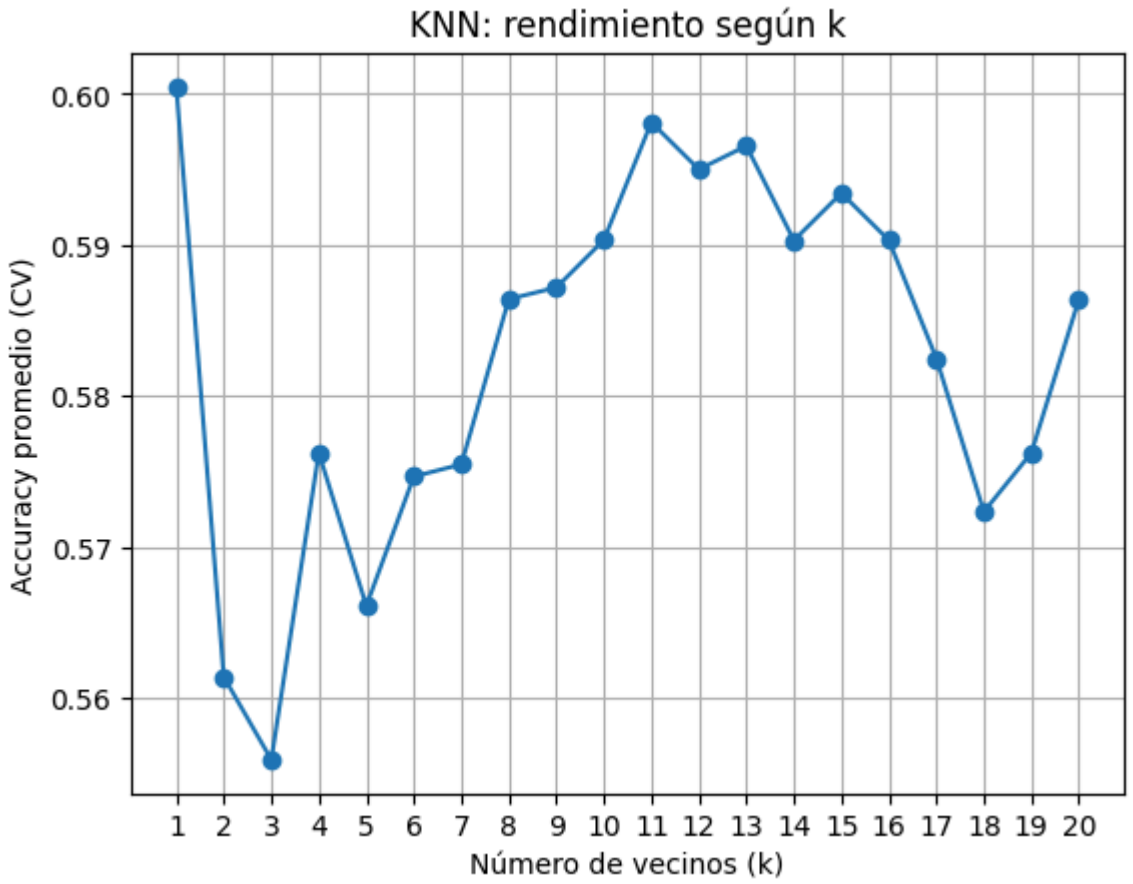
Explicación en profundidad de KNN:

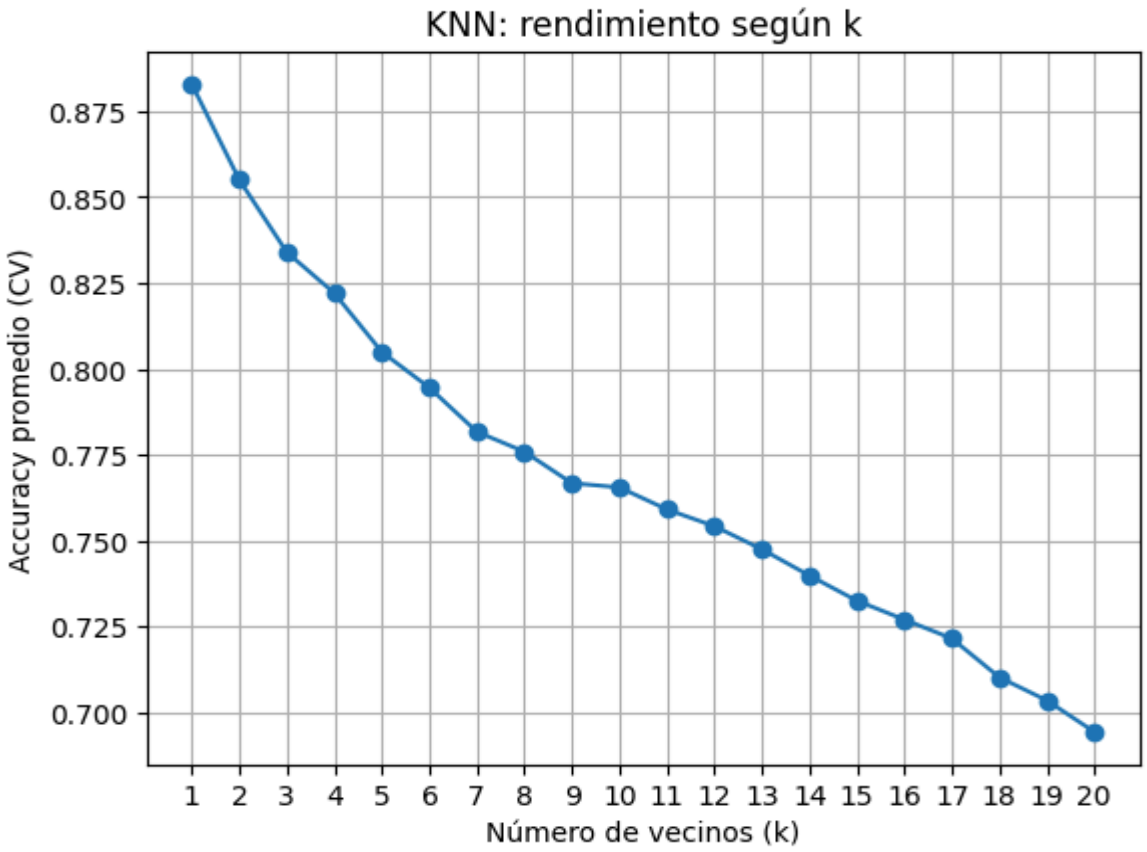
K-Nearest Neighbors (K vecinos más cercanos) es un método de clasificación. Lo que hace es calcular la distancia de cada observación nueva con cada una de las observaciones del conjunto de entrenamiento. Esta distancia suele ser la distancia euclídea aunque se puede realizar el algoritmo con otras distancias. Por último, se identifican los k vecinos más cercanos y se le asigna la clase que tiene la mayoría de sus vecinos más cercanos. Si esta observación tiene dos o más clases que empatan en cuánto a cantidad de vecinos más cercanos, se asigna al azar entre una de estas clases.

En cuánto al número K, es un hiper parámetro que hay que definir previamente. Generalmente, se eligen relativamente chicos (entre 1 y 10) y se evalúa el rendimiento mediante validación cruzada. Cuando la clasificación es binaria, se prefieren los k impares ya que evitan la posibilidad de empates.

Por último, en cuánto al conjunto de datos, no se hace ninguna suposición acerca de la distribución de las variables explicativas.

En este trabajo, se utilizó la clase **`KNeighborsClassifier`** de *scikit-learn*. Para seleccionar el valor de *k*, primero se analizó el rendimiento del modelo en términos de *accuracy* para valores de *k* entre 1 y 20, tanto en el conjunto balanceado como en el no balanceado. La elección final de **k=1** se validó empíricamente, ya que fue la que mejor desempeño mostró, permitiendo capturar con mayor precisión las diferencias locales entre vinos de distinta calidad. En cuánto a la distancia, utilizamos la euclídea, que viene por defecto en esta librería. A continuación, lo graficamos:

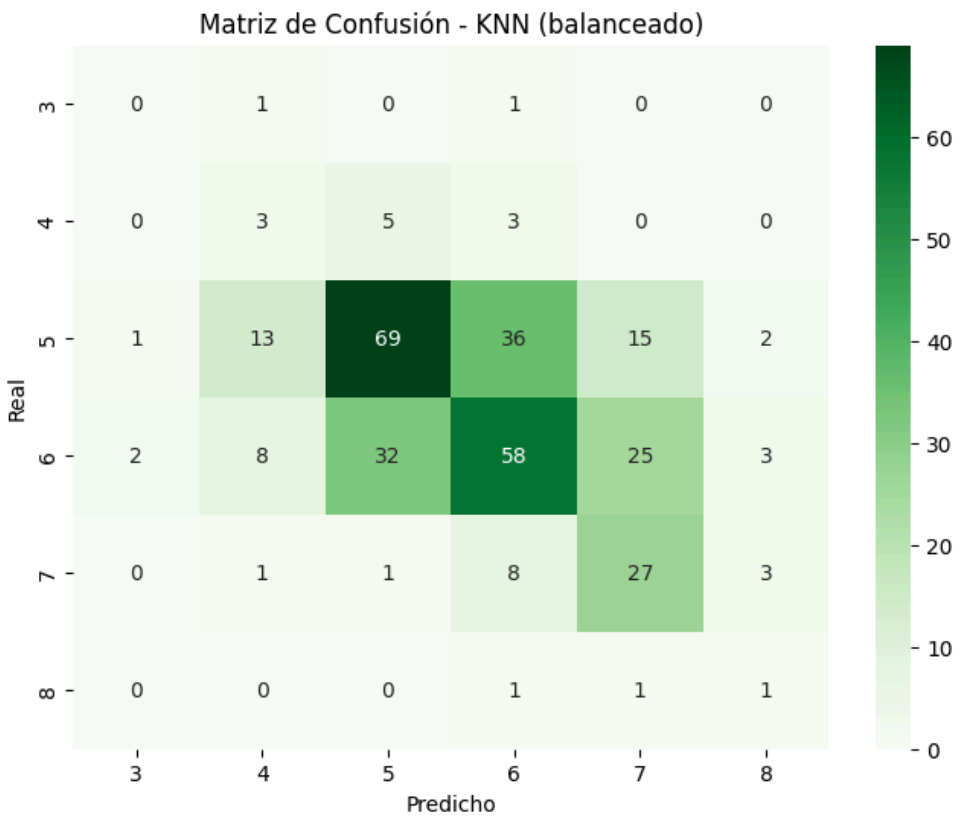




Con k=1 obtuvimos las siguientes métricas:

Métrica de evaluación	**Balanceado**
Accuracy	**0.625**
Accuracy medio CV	**0.882874617737003**
Precisión	**0.36296863395500956**
Recall	**0.401361036838978**
F1-score	**0.37621348694323836**
AUC	**0.6540949502546569**

La matriz de confusión para Knn con K=1 es:

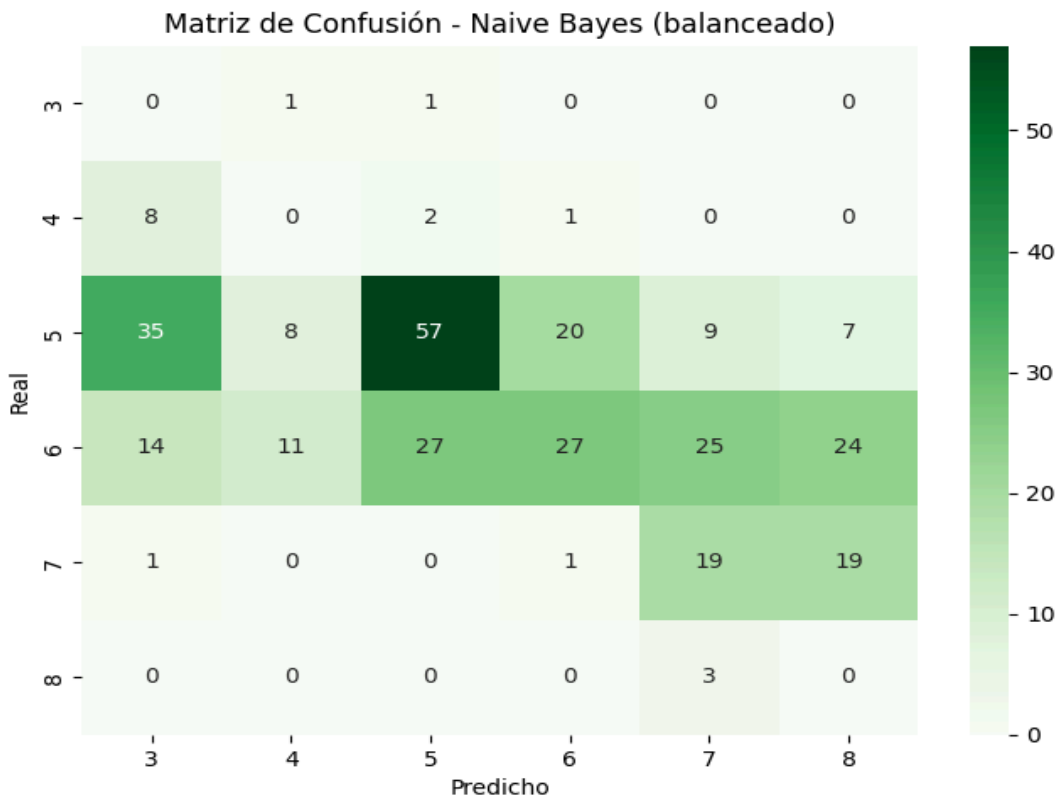


3.2.3 Bayes Ingenuo

En el caso de Bayes Ingenuo trabajamos con lo conjuntos balanceados y no balanceados y obtuvimos las siguientes métricas:

Métricas de evaluación	Naive Bayes	Naive Bayes (balanceado)
Accuracy	**0.5625**	**0.321875**
Accuracy medio CV	**0.5339950980392156**	**0.482262996941896**
Precisión	**0.3199545901932713**	**0.25757975604034716**
Recall	**0.3213722704991087**	**0.18417585784313725**
F1-score	**0.32035480859010274**	**0.20202147357131395**
AUC	**0.6837826791210272**	**0.6317021326419697**

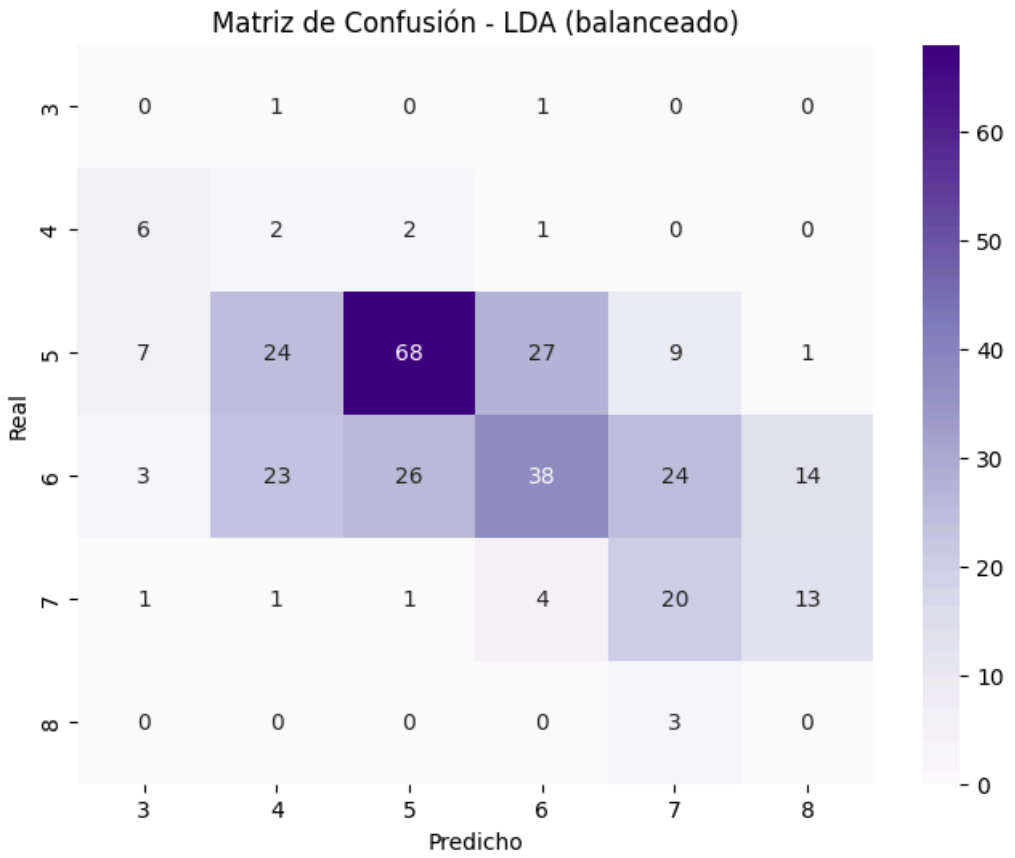
La matriz de confusión obtenida es:



3.2.4 LDA

Al aplicar LDA obtuvimos:

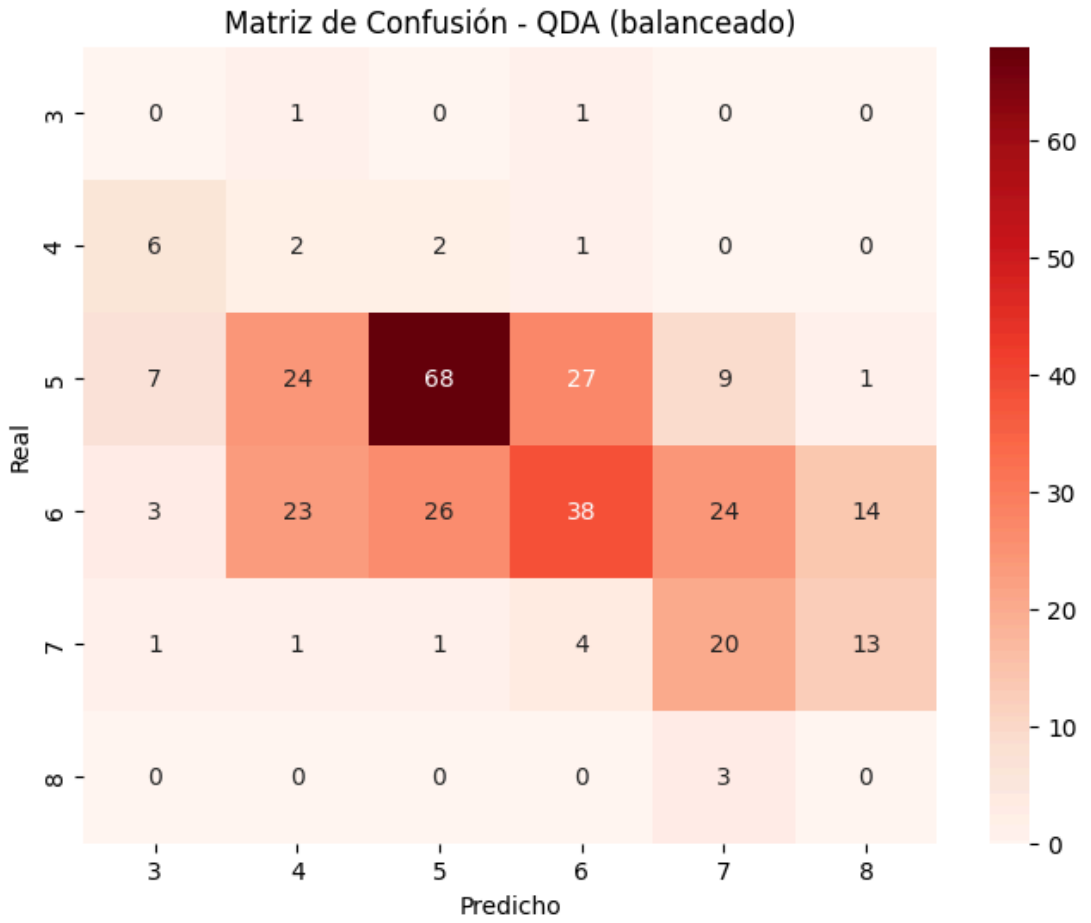
Métricas de evaluación	**LDA (BALANCEADO)**
Accuracy	**0.4**
Accuracy medio CV	**0.5553516819571864**
Precision	**0.2721001231430087**
Recall	**0.24644886363636365**
F1-score	**0.2411305550936803**
AUC	**0.7367112794947491**



3.2.5 QDA

Por último, al aplicar QDA obtuvimos:

Métricas de evaluación	**QDA (BALANCEADO)**
Accuracy	**0.4**
Accuracy medio CV	**0.6770642201834862**
Precision	**0.2721001231430087**
Recall	**0.24644886363636365**
F1-score	**0.2411305550936803**
AUC	**0.7172896298689038**

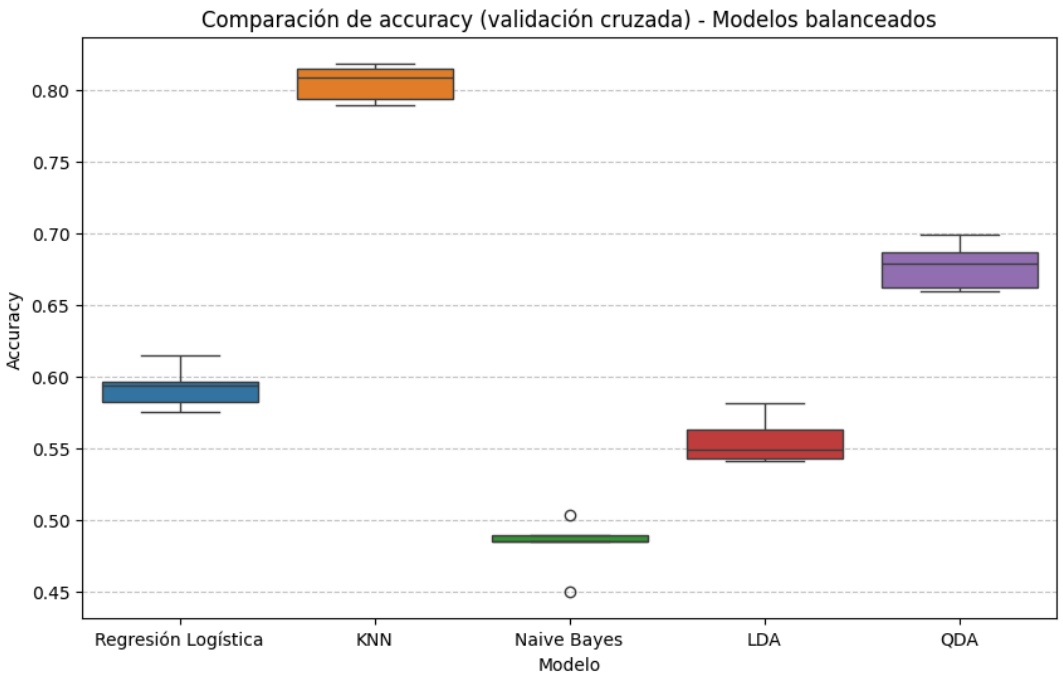


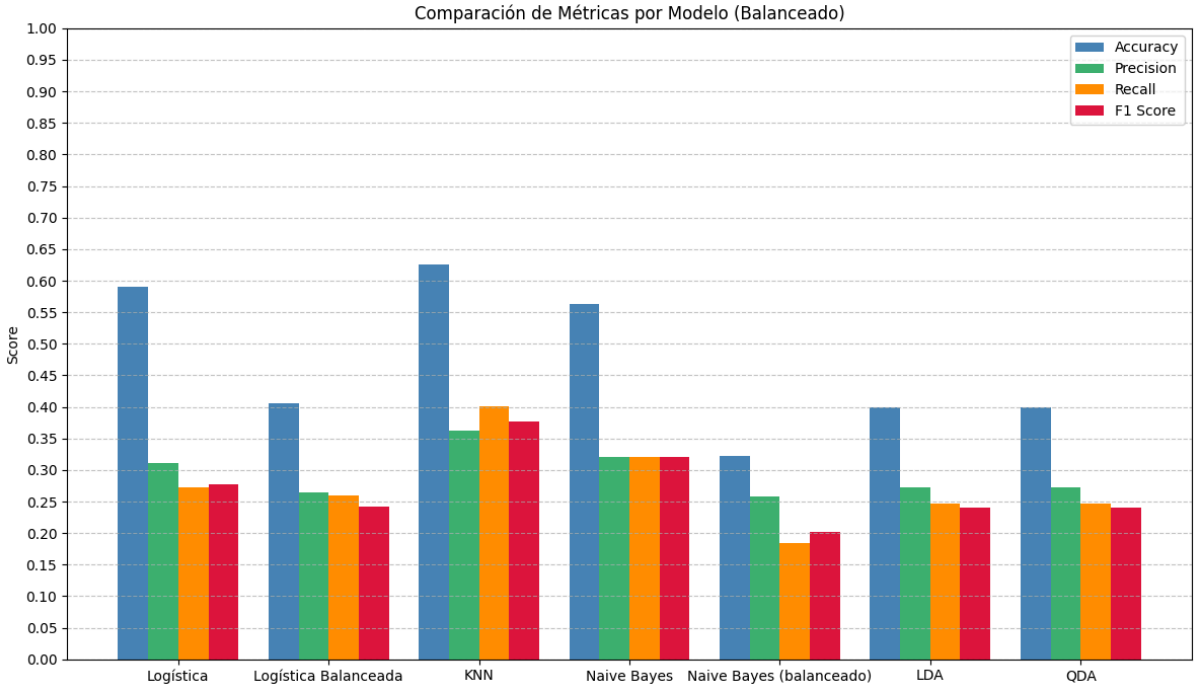
3.2.6 Conclusiones parciales

A partir de los resultados obtenidos en los modelos de clasificación, se observa que el desempeño varía según la técnica utilizada y el impacto del balanceo de clases. En general, las métricas muestran valores modestos de *Accuracy*, *Precisión*, *Recall* y *F1-score macro*, lo cual refleja la dificultad del dataset debido al fuerte desbalance de clases y a la naturaleza ordinal de la variable *quality*.

Los gráficos comparativos de barras permiten visualizar cómo cada modelo se desempeña en términos de las distintas métricas. Por ejemplo, en el caso de la regresión logística multinomial y el Bayes ingenuo, se aprecia una caída en el *Accuracy* al aplicar sobremuestreo, aunque mejora la capacidad de identificar clases minoritarias, lo que se refleja en los valores de *Recall* y *F1-score macro*. En LDA y QDA, en cambio, los resultados se mantienen más equilibrados, con valores de *AUC* relativamente altos que indican una aceptable capacidad de discriminación entre clases. Finalmente, KNN destaca como el modelo con mejor rendimiento global, especialmente con un número bajo de vecinos, lo que se evidencia en métricas superiores y en su comportamiento en los gráficos.

En conjunto, los gráficos muestran de manera clara la comparación entre modelos y métricas: permiten identificar fortalezas y debilidades de cada técnica.


Comparación de accuracy (validación cruzada) - Modelos balanceados
Accuracy
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
Regresión Logística
KNN
Naive Bayes
LDA
QDA
Modelo


Comparación de Métricas por Modelo (Balanceado)
Accuracy
Precision
Recall
F1 Score
Score
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Logística
Logística Balanceada
KNN
Naive Bayes
Naive Bayes (balanceado)
LDA
QDA

4. Conclusión

En este trabajo se evaluaron tanto métodos de regresión como de clasificación para predecir la calidad del vino. Los modelos de regresión explicaron una proporción limitada de la variabilidad, siendo Ridge el que obtuvo el mejor desempeño dentro de ese grupo. Sin embargo, dado que la variable respuesta representa categorías discretas de calidad (valores del 3 al 8), los enfoques de clasificación resultaron más adecuados.

Al comparar distintos clasificadores, se observó que el modelo KNN con un solo vecino alcanzó el mejor rendimiento, superando al resto de los métodos evaluados. Por lo tanto, se concluye que, dentro de las alternativas consideradas, el clasificador KNN con k = 1 es el más apropiado para este conjunto de datos.