Investigación de la insdustria vinícola en Portugal

Carlos Javier Mahecha Gómez, Juan Sebastián Ramirez Solano, Juan Camilo Suárez Soto

2025-03-09

Introducción

El vino ha dejado de ser un producto exclusivo para convertirse en una bebida consumida por personas de diversos gustos y condiciones socioeconómicas. El presente estudio tendrá como eje vertical datos extraídos de Portugal, reconocido como uno de los principales exportadores de vino a nivel mundial, ha experimentado un notable crecimiento en la producción de vinho verde, un vino característico de la región noroeste del país. Este auge en la industria vinícola ha impulsado la inversión en nuevas tecnologías, tanto en los procesos de elaboración como en la comercialización del vino. [1]
La universidad de Minho publicó un dataset para el modelado de las preferencias de sabor del vinho verde que incluye muestras de vino blanco y vino tinto, con el objetivo de predecir la calidad del vino. [2]
Los datos fueron extraídos de un análisis de laboratorio realizado con múltiples muestras de vino, evaluando su composición física y química. Se llevaron a cabo pruebas fisicoquímicas rutinarias, como la medición de densidad, contenido de alcohol y pH. Paralelamente, se realizaron pruebas sensoriales en las que participantes degustaron las muestras con etiquetas que incluían información de laboratorio y certificación del vino. Sin embargo, la evaluación sensorial por expertos humanos presenta cierto sesgo debido a la complejidad del sentido del gusto y a la limitada comprensión de la relación entre los análisis fisicoquímicos y la percepción sensorial. [1]
En éste contexto, la minería de datos emerge como una herramienta para extraer conocimiento de grandes conjuntos de datos fisicoquímicos y sensoriales usando algoritmos de aprendizaje supervisado de regresión lineal para modelar relaciones entre datos complejos y predecir la calidad del vino. Sin embargo, la selección de variables para alimentar al algoritmo sigue siendo un desafío, para obtener resultados precisos.

Justificación

La predicción de la calidad del vino a partir de pruebas analíticas tiene diversas aplicaciones prácticas. Permite a las entidades de certificación agilizar el proceso de evaluación, a los productores de vino mejorar la producción y a los consumidores tomar decisiones informadas. Además, el modelado de las preferencias de sabor puede ser útil para el marketing dirigido, al identificar los gustos de nichos de mercado específicos. [1]

- ¿Cuáles son las características fisicoquímicas que tienen mayor impacto en la calidad del vino?
- ¿Existe una diferencia en la importancia de estas características entre vinos tintos y blancos?
- ¿Cómo influye el nivel de acidez en la calidad del vino?
- ¿Qué diferencias existen entre los vinos de baja y alta calidad en términos de sus componentes químicos?
- ¿Existe una correlación entre el contenido de alcohol y la calidad del vino?
- ¿Las preferencias sensoriales del sabor pueden afectar el entrenamiento del algoritmo?

- ¿Cómo afecta el nivel de azúcar residual en la calificación del vino?
- ¿Qué diferencias fisicoquímicas se observan entre los vinos tintos y blancos en este conjunto de datos?

Objetivo general

- Desarrollar un algoritmo supervisado de regresión lineal que permita predecir la calidad del vino vinho verde a partir de un dataset que contiene pruebas analíticas fisicoquímicas.

Objetivos Específicos

- Analizar y preprocesar un conjunto de datos de vinho verde, que incluya mediciones fisicoquímicas y evaluaciones sensoriales mediante la metodología KDD
- Evaluar la importancia de las variables fisicoquímicas en la predicción de la calidad del vinho verde, mediante técnicas de selección de variables.
- Comprender el proceso de entrenar un algoritmo supervisado de regresión lineal a partir de un conjunto de datos en el lenguaje de programación R.

Fase 1: Selección de datos

Objetivo: Identificar y seleccionar los datos relevantes para el análisis.

características de los datasets.

Los archivos que se utilizaran en este análisis son "winequality-red" y "winequality-white", los cuales proporcionan los datos necesarios para realizar un análisis detallado de la calidad del vino. Estos datatasets fueron recolectaron de la página web de UC Irvine Machine Learning Repository y fue publicada el 22 de mayo de 2009 por el equipo de investigadores de la universidad de Minho en Portugal. La información adjunta de los archivos están en formato ".csv" y contiene datos específicos sobre vinos blancos y vino tinto elaborado en el noroccidente de Portugal.

En total, el dataset del vino tinto tiene 1.600 registros, mientras que el vino blanco tiene 4.899 registros. Cada uno de estos registros cuentan con variables que nos permite evaluar aspectos importantes para la calidad del vino, como la acidez, el contenido de azúcar, el contenido de sal, el contenido de alcohol y otros aspectos químicos que contiene el vino.

Análisis de variables

Cada dataset posee la misma estructura de columnas, lo que garantiza una presentación uniforme de la información; sin embargo, los datos que contiene cada uno son distintos. A continuación, se ofrece los nombres de cada columna:

1. Fixed_acidity
2. Volatile_acidity
3. Citric_acid
4. Residual_sugar
5. Cloruros
6. Free_sulfur_dioxide
7. Total_sulfur_dioxide
8. Density

9. pH
10. Sulphates
11. Alcohol
12. Quality

Identificación de posibles problemas iniciales

A continuación, se analizarán posibles problemas que puedan comprometer la confiabilidad y precisión de la información, afectando su calidad. El objetivo es garantizar datos organizados y resultados más precisos, detectando y corrigiendo errores como valores faltantes, inconsistentes, atípicos o duplicados. Para validar esto, se llevó a cabo el siguiente análisis:

Datos nulos: En un primer lugar, un script identificó si campos vacios con la función *sum(is.na(dataset))*

```
## Cantidad de valores nulos encontrados en el dataset de vino tinto: 0
```

```
## Cantidad de valores nulos encontrados en el dataset de vino blanco: 0
```

Valores negativos: En segundo lugar, otro script indagó la existencia de valores negativos con la función *sum(data[data] < 0)*

```
## Cantidad de valores negativos encontrados en el dataset de vino tinto: 0
```

```
## Cantidad de valores negativos encontrados en el dataset de vino blanco: 0
```

filas y/o datos repetidos: En tercer lugar, un script identificó que hay filas repetidas, y se debe entrar a los datasets a analizar si esta disposición puede distorsionar el análisis y los resultados del algoritmo.

```
## Cantidad de filas duplicadas en el dataset de vino tinto: 240
```

```
## Cantidad de filas duplicadas en el dataset de vino blanco: 937
```

Cantidad de filas y columnas de cada dataset: Es importante tener en cuenta que, al entrenar el algoritmo, habrá más datos disponibles para la predicción del vino blanco que del vino rojo, lo que podría influir en el rendimiento del modelo.

```
## El dataset de vino blanco tiene  4898 filas y 12 columnas.
```

```
## El dataset de vino tinto tiene 1599 filas y 12 columnas.
```

Análisis estadísticos

Para obtener una visión integral de los datasets, se realizó un análisis estadístico de cada variable, identificando el tipo de dato (discreto, cuantitativo, etc.) y calculando métricas clave como media, mediana, varianza, desviación estándar y el número de columnas. Esto permitirá una primera impresión sobre la distribución y características de los datos.

Calculos estadisticos: la media, mediana y desviación estándar.

dataset de vino blanco

Table 1: Estadísticas para dataset de vino blanco

Variable	Media	Mediana	Moda	Desviacion_Estandar
fixed.acidity	6.8547877	6.80000	6.800	0.8438682
volatile.acidity	0.2782411	0.26000	0.280	0.1007945
citric.acid	0.3341915	0.32000	0.300	0.1210198
residual.sugar	6.3914149	5.20000	1.200	5.0720578
chlorides	0.0457724	0.04300	0.044	0.0218480
free.sulfur.dioxide	35.3080849	34.00000	29.000	17.0071373
total.sulfur.dioxide	138.3606574	134.00000	111.000	42.4980646
density	0.9940274	0.99374	0.992	0.0029909
pH	3.1882666	3.18000	3.140	0.1510006
sulphates	0.4898469	0.47000	0.500	0.1141258
alcohol	10.5142670	10.40000	9.400	1.2306206
quality	5.8779094	6.00000	6.000	0.8856386

dataset de vino tinto

Table 2: Estadísticas para dataset de vino rojo

Variable	Media	Mediana	Moda	Desviacion_Estandar
fixed.acidity	8.3196373	7.90000	7.2000	1.7410963
volatile.acidity	0.5278205	0.52000	0.6000	0.1790597
citric.acid	0.2709756	0.26000	0.0000	0.1948011
residual.sugar	2.5388055	2.20000	2.0000	1.4099281
chlorides	0.0874665	0.07900	0.0800	0.0470653
free.sulfur.dioxide	15.8749218	14.00000	6.0000	10.4601570
total.sulfur.dioxide	46.4677924	38.00000	28.0000	32.8953245
density	0.9967467	0.99675	0.9972	0.0018873
pH	3.3111132	3.31000	3.3000	0.1543865
sulphates	0.6581488	0.62000	0.6000	0.1695070
alcohol	10.4229831	10.20000	9.5000	1.0656676
quality	5.6360225	6.00000	5.0000	0.8075694

Datos faltantes por cada dataset:

En cuanto a campos faltantes, no hay datos faltantes, y la consistencia observada sugiere que el dataset fue previamente estructurado.

Table 3: Datos faltantes en dataset de vihno blanco y vihno tinto

Variable	Faltantes
fixed.acidity	0
volatile.acidity	0
citric.acid	0
residual.sugar	0

Variable	Faltantes
chlorides	0
free.sulfur.dioxide	0
total.sulfur.dioxide	0
density	0
pH	0
sulphates	0
alcohol	0
quality	0

Fase 2: Limpieza de datos

Objetivo: Preparar los datos para el análisis, eliminando errores y valores inconsistentes.
En el análisis previo se evaluó que, lo único preocupante es que hay datos duplicados, por lo que deben eliminarse estos datos para su posterior transoformación mediante el tratamiendo de outlers.

Eliminación de datos duplicados: A contianuación observará los resultados de esta operación:

```
## Vino blanco - filas antes de eliminar duplicados: 4898
##  Vino tinto  - filas antes de eliminar duplicados: 1599
```

```
## Vino blanco - filas tras eliminar duplicados: 3961 - ha sido eliminado el 12% del dataset
##  Vino tinto  - filas tras eliminar duplicados: 1359 - ha sido eliminado el 19,3% del dataset
```

Tratamiento de outlers

En el siguiente apartado se realizará el tratamiento de datos anómalos en una variable: únicamente la acidez ajustada presenta valores fuera de rango, específicamente por encima de 10. Por esta razón, dichos registros serán ajustados para que se encuentren dentro del rango esperado de 1 a 10. Además, se procederá a eliminar outliers en cinco variables que presentan valores atípicos, principalmente dentro del conjunto de datos de vino blanco. Estas variables son: densidad, acidez volátil, azúcar residual y dióxido de azufre (libre y total).

ácidez ajustada: Hay datos anormales que tienen una ácidez superior a 100. Esto no puede pasar, así que como probablemente se trata de un decimal, se va a quitarle el último dígito

```
## Resumen de fixed.acidity en vino_blanco_final:
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   6.300   6.800   6.817   7.300   9.900
```

```
##
## Resumen de fixed.acidity en vino_tinto_final:
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   6.300   7.300   6.549   8.200   9.900
```

```
##
## fixed.acidity ha sido ajustado y los archivos sobrescritos.
```

densidad: Hay 2 datos que tienen valores muy por encima de lo normal en el dataset de vihno blanco, por lo tanto se eliminarán a continuación.

```
## Resumen de density tras filtrar:

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9916  0.9935  0.9938  0.9957  1.0030

##
## vino_blanco_final ha sido actualizado y guardado.
```

ácido cítrico: Hay 2 datos que tienen valores muy por encima de lo normal en el dataset de vihno blanco, por lo tanto se eliminarán a continuación.

```
## Resumen de density tras filtrar:

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2700  0.3200  0.3337  0.3900  1.0000

##
## vino_blanco_final ha sido actualizado y guardado.
```

dioxido de azufre total: Hay 4 datos que tienen valores muy por encima de lo normal en los dataset de vihno blanco y vihno tinto, por lo tanto se eliminarán a continuación.

```
## Resumen de total.sulfur.dioxide en vino_blanco_final:

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       9     106     133     137     166     313

##
## Resumen de total.sulfur.dioxide en vino_tinto_final:

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.39   63.00  160.00

##
## Los datasets han sido filtrados y guardados correctamente.
```

Fase 3: Transformación de datos

Objetivo: Transformar los datos para que sean adecuados para el modelado.

En esta fase, se realizará un análisis bivariado entre la variable quality y las otras 11 variables de los 2 datasets, con el objetivo de identificar cuáles aportan información relevante y cuáles pueden ser descartadas. La decisión de conservar o eliminar variables se basará en pruebas de correlación

Análisis univariado de la variable `quality`:

A continuación encontrará un análisis estadístico dedicado únicamente a la calidad de vino, donde 1 es malo y 10 es excelente:

```
## Resumen de quality en vino_blanco_final:
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.856   6.000   9.000
```

```
##
## Resumen de quality en vino_tinto_final:
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.621   6.000   8.000
```

Del análisis estadístico de la variable quality, se observa que tanto los vinos blancos como los tintos presentan una mediana de 6 puntos, indicando que la mitad de las observaciones en ambos conjuntos tienen una calidad igual o inferior a ese valor. El vino blanco muestra una media ligeramente superior (5.86) respecto al vino tinto (5.62), lo que sugiere una tendencia levemente más alta en la calidad percibida del vino blanco. En cuanto al rango intercuartílico (IQR), ambos tipos de vino comparten los mismos valores para el primer cuartil (5.00) y el tercer cuartil (6.00), por lo que el IQR es de 1 punto para ambos, lo que indica una concentración similar de los valores centrales. Sin embargo, el vino blanco alcanza una calidad máxima de 9 puntos, mientras que el tinto llega solo a 8, lo que sugiere una mayor variabilidad en la calidad de los vinos blancos. En resumen, aunque las distribuciones son bastante parecidas, el vino blanco presenta una ligera ventaja en cuanto a promedio y rango máximo de calidad.

Asimetría y curtosis

La **asimetría** mide qué tan simétrica está la distribución de los datos de los 2 datasets: y al ser positiva, la cola de la derecha es más larga y tiene un sesgo positivo. Por otro lado, la **curtosis**, indica qué tan puntiaguda o plana es una distribución en comparación con una normal: un valor alto y superior a 3 sugiere colas más pesadas y valores más extremos.

vino blanco:

```
## [1] 0.1246515
```

```
## [1] 3.272914
```

vino tinto:

```
## [1] 0.194651
```

```
## [1] 3.346404
```

Pruebas de normalidad

Demostrado lo anterior, podemos concluir que **la la puntuación de la calidad no es una distribución normal**, por lo que se debe usar el método de Anderson Darling, que es una variante del método KS con más peso en los extremos.

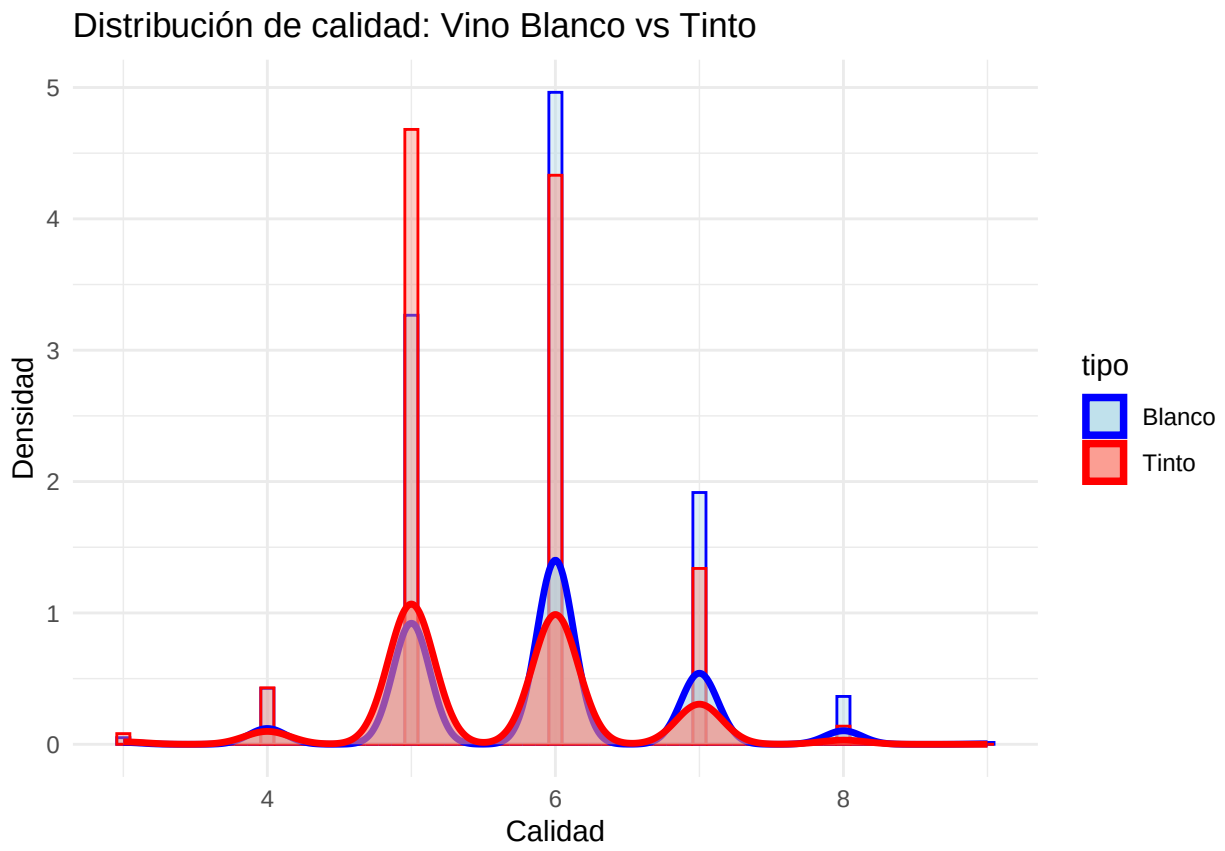
```
## Prueba de Anderson-Darling para normalidad:
##  Estadístico A² = 90.2689
##  Valor p = 3.7e-24
##
##  Interpretación:
##  - Si valor p < 0.05: Rechazar normalidad (distribución no normal).
##  - Si valor p >= 0.05: No hay evidencia para rechazar normalidad.
```

La variable quality del vino tinto no sigue una distribución normal
Dado que el anterior test de Anderson Darling a la variable quality, obtuvo una distribución no normal, se puede concluir que quality es una variable no paramétrica.

```
## Prueba de Anderson-Darling para normalidad:
##  Estadístico A² = 213.8245
##  Valor p = 3.7e-24
##
##  Interpretación:
##  - Si valor p < 0.05: Rechazar normalidad (distribución no normal).
##  - Si valor p >= 0.05: No hay evidencia para rechazar normalidad.
```

La variable quality del vino blanco no sigue una distribución normal
Dado que el anterior test de Anderson Darling a la variable quality, obtuvo una distribución no normal, se puede concluir que quality es una variable no paramétrica.

diagrama de densidad muestra la distribución de la calidad de los vinos blancos. Las barras rojas representan al vino tinto y las barras azules indican cuántos la calidad del vino blanco. La curva roja y azul representan una estimación no paramétrica de la densidad, es decir, si se ajusta a la forma real de los datos sin asumir que siguen un patrón específico, lo que permite comparar la forma real de los datos con una distribución ideal.



Hay una bimodalidad.

Análisis de correlaciones de Spearman entre Quality y Variables de Vino

Para seleccionar las variables más relevantes, este apartado presenta la prueba de correlación no paramétrica de Spearman, utilizada para evaluar la asociación entre la calidad (quality) y las distintas variables fisicoquímicas en vinos blancos y tintos.

Selección de alcohol

Se analizará la relación entre las variables quality y alcohol. Sabemos previamente que quality no sigue una distribución normal, de acuerdo con el test de Anderson-Darling. Ahora, se debe verificar si alcohol tiene una distribución normal a través del test de Anderson Darling y un diagrama QQ plot.

Prueba de normalidad en el dataset de vino blanco: El siguiente test demuestra que alcohol no sigue una distribución normal y por lo tanto, es una variable no paramétrica.

```
## Prueba de Anderson-Darling para normalidad:
##  Estadístico A² = 44.2177
##  Valor p = 3.7e-24
##
##  Interpretación:
##  - Si valor p < 0.05: Rechazar normalidad (distribución no normal).
##  - Si valor p >= 0.05: No hay evidencia para rechazar normalidad.
```

Prueba de normalidad en el dataset de vino tinto: El siguiente test demuestra que alcohol no sigue una distribución normal y por lo tanto, es una variable no paramétrica.

```
## Prueba de Anderson-Darling para normalidad:
##  Estadístico A² = 30.9578
##  Valor p = 3.7e-24
##
##  Interpretación:
##  - Si valor p < 0.05: Rechazar normalidad (distribución no normal).
##  - Si valor p >= 0.05: No hay evidencia para rechazar normalidad.
```

Tras observar el gráfico, se observa que los datos de alcohol no se alinean adecuadamente sobre la línea de tendencia teórica, lo que confirma que alcohol tampoco sigue una distribución normal y por lo tanto, son una variable no paramétrica.

Prueba de correlación: A continuación se evalúa la asociación entre la calidad y la variable **alcohol** utilizando **la prueba de Spearman** para los datasets *vino_blanco_final* y *vino_tinto_final*.

```
FALSE Vino Blanco  (alcohol) - rho = 0.477 , p = 2e-223
FALSE  Vino Tinto   (alcohol) - rho = 0.486 , p = 3.36e-81
```

Dado que $p < 0.05$: sí hay una correlación significativa entre la calidad del vino y el alochol.

La variable sí va a ser considerada para futuros análisis

Selección de densidad

Se analizará la relación entre las variables quality y densidad. Sabemos previamente que quality no sigue una distribución normal, de acuerdo con el test de Anderson-Darling. Ahora, se debe verificar si densidad tiene una distribución normal a través del test de Anderson Darling y un diagrama QQ plot.

Prueba de normalidad en el dataset de vino blanco: El siguiente test demuestra que densidad no sigue una distribución normal y por lo tanto, es una variable no paramétrica.

```
## Prueba de Anderson-Darling para normalidad:
##  Estadístico A² = 18.1765
##  Valor p = 3.7e-24
##
##  Interpretación:
##  - Si valor p < 0.05: Rechazar normalidad (distribución no normal).
##  - Si valor p >= 0.05: No hay evidencia para rechazar normalidad.
```

Prueba de normalidad en el dataset de vino tinto: El siguiente test demuestra que densidad no sigue una distribución normal y por lo tanto, es una variable no paramétrica.

```
## Prueba de Anderson-Darling para normalidad:
##  Estadístico A² = 2.8234
##  Valor p = 4.189768e-07
##
##  Interpretación:
##  - Si valor p < 0.05: Rechazar normalidad (distribución no normal).
##  - Si valor p >= 0.05: No hay evidencia para rechazar normalidad.
```

Prueba de correlación: A continuación se evalúa la asociación entre la calidad y la variable **densidad** utilizando **la prueba de Spearman** para los datasets *vino_blanco_final* y *vino_tinto_final*.

```
FALSE Vino Blanco  (densidad) - rho = -0.383 , p = 2.53e-138
FALSE  Vino Tinto   (densidad) - rho = -0.18 , p = 2.18e-11
```

Dado que $p < 0.05$: sí hay una correlación significativa entre la calidad del vino y el density.

La variable sí va a ser considerada para futuros análisis

Selección de ajuste de acidez

Se analizará la relación entre las variables quality y acidez. Sabemos previamente que quality no sigue una distribución normal, de acuerdo con el test de Anderson-Darling. Ahora, se debe verificar si acidez tiene una distribución normal a través del test de Anderson Darling y un diagrama QQ plot.

Prueba de normalidad en el dataset de vino blanco: El siguiente test demuestra que acidez no sigue una distribución normal y por lo tanto, es una variable no paramétrica.

```
## Prueba de Anderson-Darling para normalidad:
##  Estadístico A² = 20.0586
##  Valor p = 3.7e-24
##
##  Interpretación:
##  - Si valor p < 0.05: Rechazar normalidad (distribución no normal).
##  - Si valor p >= 0.05: No hay evidencia para rechazar normalidad.
```

Prueba de normalidad en el dataset de vino tinto: El siguiente test demuestra que acidez no sigue una distribución normal y por lo tanto, es una variable no paramétrica.

```
## Prueba de Anderson-Darling para normalidad:
##  Estadístico A² = 121.0869
##  Valor p = 3.7e-24
##
##  Interpretación:
##  - Si valor p < 0.05: Rechazar normalidad (distribución no normal).
##  - Si valor p >= 0.05: No hay evidencia para rechazar normalidad.
```

Prueba de correlación: A continuación se evalúa la asociación entre la calidad y la variable **fixed.acidity** utilizando **la prueba de Spearman** para los datasets *vino_blanco_final* y *vino_tinto_final*

```
FALSE Vino Blanco  (fixed.acidity) - rho = -0.088 , p = 3.54e-08
FALSE  Vino Tinto   (fixed.acidity) - rho = -0.091 , p = 0.000768
```

Dado que $p < 0.05$: sí hay una correlación significativa entre la calidad del vino y el ajuste de la ácidez.
La variable sí va a ser considerada para futuros análisis

Selección de chlorides

Se analizará la relación entre las variables quality y chlorides. Sabemos previamente que quality no sigue una distribución normal, de acuerdo con el test de Anderson-Darling. Ahora, se debe verificar si chlorides tiene una distribución normal a través del test de Anderson Darling y un diagrama QQ plot.

Prueba de normalidad en el dataset de vino blanco: El siguiente test demuestra que chlorides no sigue una distribución normal y por lo tanto, es una variable no paramétrica.

```
## Prueba de Anderson-Darling para normalidad:
##  Estadístico A² = 363.4025
##  Valor p = 3.7e-24
##
##  Interpretación:
##  - Si valor p < 0.05: Rechazar normalidad (distribución no normal).
##  - Si valor p >= 0.05: No hay evidencia para rechazar normalidad.
```

Prueba de normalidad en el dataset de vino tinto: El siguiente test demuestra que chlorides no sigue una distribución normal y por lo tanto, es una variable no paramétrica.

```
## Prueba de Anderson-Darling para normalidad:
##  Estadístico A² = 183.6085
##  Valor p = 3.7e-24
##
##  Interpretación:
##  - Si valor p < 0.05: Rechazar normalidad (distribución no normal).
##  - Si valor p >= 0.05: No hay evidencia para rechazar normalidad.
```

Prueba de correlación: A continuación se evalúa la asociación entre la calidad y la variable **chlorides** utilizando **la prueba de Spearman** para los datasets *vino_blanco_final* y *vino_tinto_final*

```
FALSE Vino Blanco  (chlorides) - rho = -0.333 , p = 9.06e-103
FALSE  Vino Tinto   (chlorides) - rho = -0.201 , p = 8.38e-14
```

Dado que $p < 0.05$: sí hay una correlación significativa entre la calidad del vino y clorhidratos

La variable sí va a ser considerada para futuros análisis

Selección de pH

Se analizará la relación entre las variables quality y pH Sabemos previamente que quality no sigue una distribución normal, de acuerdo con el test de Anderson-Darling. Ahora, se debe verificar si pH tiene una distribución normal a través del test de Anderson Darling y un diagrama QQ plot.

Prueba de normalidad en el dataset de vino blanco: El siguiente test demuestra que pH no sigue una distribución normal y por lo tanto, es una variable no paramétrica.

```
## Prueba de Anderson-Darling para normalidad:
##  Estadístico A² = 9.9259
##  Valor p = 5.535877e-24
##
##  Interpretación:
##  - Si valor p < 0.05: Rechazar normalidad (distribución no normal).
##  - Si valor p >= 0.05: No hay evidencia para rechazar normalidad.
```

Prueba de normalidad en el dataset de vino tinto: El siguiente test demuestra que pH no sigue una distribución normal y por lo tanto, es una variable no paramétrica.

```
## Prueba de Anderson-Darling para normalidad:
##  Estadístico A² = 183.6085
##  Valor p = 3.7e-24
##
##  Interpretación:
##  - Si valor p < 0.05: Rechazar normalidad (distribución no normal).
##  - Si valor p >= 0.05: No hay evidencia para rechazar normalidad.
```

Prueba de correlación: A continuación se evalúa la asociación entre la calidad y la variable **pH** utilizando **la prueba de Spearman** para los datasets *vino_blanco_final* y *vino_tinto_final*

```
FALSE Vino Blanco  (pH) - rho = 0.136 , p = 7.15e-18
FALSE  Vino Tinto   (pH) - rho = -0.039 , p = 0.154
```

Dado que en el vino tinto, $p > 0.05$, se concluye que no existe una correlación significativa entre la calidad del vino tinto y el pH. En cambio, en el vino blanco, $p < 0.05$, lo que indica una relación significativa entre el pH y la calidad. Sin embargo, se decide no seleccionar esta variable, ya que el objetivo es construir un único modelo de regresión lineal que sirva para ambos tipos de vino, y no desarrollar dos modelos separados.

La variable no va a ser considerada para futuros análisis

Selección de volatile.acidity

Se analizará la relación entre las variables quality y volatile.acidity Sabemos previamente que quality no sigue una distribución normal, de acuerdo con el test de Anderson-Darling. Ahora, se debe verificar si volatile.acidity tiene una distribución normal a través del test de Anderson Darling y un diagrama QQ plot.

Prueba de normalidad en el dataset de vino blanco: El siguiente test demuestra que volatile.acidity no sigue una distribución normal y por lo tanto, es una variable no paramétrica.

```
## Prueba de Anderson-Darling para normalidad:
##  Estadístico A² = 73.3252
##  Valor p = 3.7e-24
##
##  Interpretación:
##  - Si valor p < 0.05: Rechazar normalidad (distribución no normal).
##  - Si valor p >= 0.05: No hay evidencia para rechazar normalidad.
```

Prueba de normalidad en el dataset de vino tinto: El siguiente test demuestra que volatile.acidity no sigue una distribución normal y por lo tanto, es una variable no paramétrica.

```
## Prueba de Anderson-Darling para normalidad:
##  Estadístico A² = 6.0686
##  Valor p = 6.382553e-15
##
##  Interpretación:
##  - Si valor p < 0.05: Rechazar normalidad (distribución no normal).
##  - Si valor p >= 0.05: No hay evidencia para rechazar normalidad.
```

Prueba de correlación: A continuación se evalúa la asociación entre la calidad y la variable **volatile.acidity** utilizando **la prueba de Spearman** para los datasets *vino_blanco_final* y *vino_tinto_final*

```
FALSE Vino Blanco  (volatile.acidity) - rho = -0.185 , p = 9.88e-32
FALSE  Vino Tinto   (volatile.acidity) - rho = -0.385 , p = 3.82e-49
```

Dado que $p < 0.05$: sí hay una correlación significativa entre la calidad del vino y volatile.acidity
La variable sí va a ser considerada para futuros análisis

Selección de sulphates

Se analizará la relación entre las variables quality y sulphates Sabemos previamente que quality no sigue una distribución normal, de acuerdo con el test de Anderson-Darling. Ahora, se debe verificar si sulphates tiene una distribución normal a través del test de Anderson Darling y un diagrama QQ plot.

Prueba de normalidad en el dataset de vino blanco: El siguiente test demuestra que sulphates no sigue una distribución normal y por lo tanto, es una variable no paramétrica.

```
## Prueba de Anderson-Darling para normalidad:
##  Estadístico A² = 34.903
##  Valor p = 3.7e-24
##
##  Interpretación:
##  - Si valor p < 0.05: Rechazar normalidad (distribución no normal).
##  - Si valor p >= 0.05: No hay evidencia para rechazar normalidad.
```

Prueba de normalidad en el dataset de vino tinto: El siguiente test demuestra que sulphates no sigue una distribución normal y por lo tanto, es una variable no paramétrica.

```
## Prueba de Anderson-Darling para normalidad:
##  Estadístico A² = 42.7858
##  Valor p = 3.7e-24
##
##  Interpretación:
##  - Si valor p < 0.05: Rechazar normalidad (distribución no normal).
##  - Si valor p >= 0.05: No hay evidencia para rechazar normalidad.
```

Prueba de correlación: A continuación se evalúa la asociación entre la calidad y la variable **sulphates** utilizando **la prueba de Spearman** para los datasets *vino_blanco_final* y *vino_tinto_final*

```
FALSE Vino Blanco  (sulphates) - rho = 0.037 , p = 0.0189
FALSE  Vino Tinto   (sulphates) - rho = 0.385 , p = 3.66e-49
```

Dado que $p < 0.05$: sí hay una correlación significativa entre la calidad del vino y los sulfatos.

La variable sí va a ser considerada para futuros análisis

Selección de free.sulfur.dioxide

Se analizará la relación entre las variables quality y free.sulfur.dioxide Sabemos previamente que quality no sigue una distribución normal, de acuerdo con el test de Anderson-Darling. Ahora, se debe verificar si free.sulfur.dioxide tiene una distribución normal a través del test de Anderson Darling y un diagrama QQ plot.

Prueba de normalidad en el dataset de vino blanco: El siguiente test demuestra que free.sulfur.dioxide no sigue una distribución normal y por lo tanto, es una variable no paramétrica.

```
## Prueba de Anderson-Darling para normalidad:
##  Estadístico A² = 18.5943
##  Valor p = 3.7e-24
##
##  Interpretación:
##  - Si valor p < 0.05: Rechazar normalidad (distribución no normal).
##  - Si valor p >= 0.05: No hay evidencia para rechazar normalidad.
```

Prueba de normalidad en el dataset de vino tinto: El siguiente test demuestra que free.sulfur.dioxide no sigue una distribución normal y por lo tanto, es una variable no paramétrica.

```
## Prueba de Anderson-Darling para normalidad:
##  Estadístico A² = 32.4158
##  Valor p = 3.7e-24
##
##  Interpretación:
##  - Si valor p < 0.05: Rechazar normalidad (distribución no normal).
##  - Si valor p >= 0.05: No hay evidencia para rechazar normalidad.
```

Prueba de correlación: A continuación se evalúa la asociación entre la calidad y la variable **free.sulfur.dioxide** utilizando **la prueba de Spearman** para los datasets *vino_blanco_final* y *vino_tinto_final*

```
FALSE Vino Blanco  (free.sulfur.dioxide) - rho = 0.034 , p = 0.0326
FALSE  Vino Tinto   (free.sulfur.dioxide) - rho = -0.063 , p = 0.0194
```

Dado que $p < 0.05$: no hay una correlación significativa entre la calidad del vino y el dioxido de sulfuro libre

La variable sí va a ser considerada para futuros análisis

Selección de citric.acid

Se analizará la relación entre las variables quality y citric.acid Sabemos previamente que quality no sigue una distribución normal, de acuerdo con el test de Anderson-Darling. Ahora, se debe verificar si citric.acid tiene una distribución normal a través del test de Anderson Darling y un diagrama QQ plot.

Prueba de normalidad en el dataset de vino blanco: El siguiente test demuestra que citric.acid no sigue una distribución normal y por lo tanto, es una variable no paramétrica.

```
## Prueba de Anderson-Darling para normalidad:
##  Estadístico A² = 63.9845
##  Valor p = 3.7e-24
##
##  Interpretación:
##  - Si valor p < 0.05: Rechazar normalidad (distribución no normal).
##  - Si valor p >= 0.05: No hay evidencia para rechazar normalidad.
```

Prueba de normalidad en el dataset de vino tinto: El siguiente test demuestra que citric.acid no sigue una distribución normal y por lo tanto, es una variable no paramétrica.

```
## Prueba de Anderson-Darling para normalidad:
##  Estadístico A² = 14.9369
##  Valor p = 3.7e-24
##
##  Interpretación:
##  - Si valor p < 0.05: Rechazar normalidad (distribución no normal).
##  - Si valor p >= 0.05: No hay evidencia para rechazar normalidad.
```

Prueba de correlación: A continuación se evalúa la asociación entre la calidad y la variable **citric.acid** utilizando **la prueba de Spearman** para los datasets *vino_blanco_final* y *vino_tinto_final*

```
FALSE Vino Blanco  (citric.acid) - rho = 0.03 , p = 0.0588
FALSE  Vino Tinto   (citric.acid) - rho = 0.216 , p = 8.89e-16
```

Dado que $p > 0.05$: no hay una correlación significativa entre la calidad del vino y el ácido cítrico.

La variable no va a ser considerada para futuros análisis

Selección de total.sulfur.dioxide

Se analizará la relación entre las variables quality y total.sulfur.dioxide Sabemos previamente que quality no sigue una distribución normal, de acuerdo con el test de Anderson-Darling. Ahora, se debe verificar si total.sulfur.dioxide tiene una distribución normal a través del test de Anderson Darling y un diagrama QQ plot.

Prueba de normalidad en el dataset de vino blanco: El siguiente test demuestra que total.sulfur.dioxide no sigue una distribución normal y por lo tanto, es una variable no paramétrica.

```
## Prueba de Anderson-Darling para normalidad:
##  Estadístico A² = 11.373
##  Valor p = 3.7e-24
##
##  Interpretación:
##  - Si valor p < 0.05: Rechazar normalidad (distribución no normal).
##  - Si valor p >= 0.05: No hay evidencia para rechazar normalidad.
```

Prueba de normalidad en el dataset de vino tinto: El siguiente test demuestra que total.sulfur.dioxide no sigue una distribución normal y por lo tanto, es una variable no paramétrica.

```
## Prueba de Anderson-Darling para normalidad:
##  Estadístico A² = 42.8735
##  Valor p = 3.7e-24
##
##  Interpretación:
##  - Si valor p < 0.05: Rechazar normalidad (distribución no normal).
##  - Si valor p >= 0.05: No hay evidencia para rechazar normalidad.
```

Prueba de correlación: A continuación se evalúa la asociación entre la calidad y la variable **total.sulfur.dioxide** utilizando **la prueba de Spearman** para los datasets *vino_blanco_final* y *vino_tinto_final*

```
FALSE Vino Blanco  (total.sulfur.dioxide) - rho = -0.201 , p = 2.23e-37
FALSE  Vino Tinto   (total.sulfur.dioxide) - rho = -0.202 , p = 5.2e-14
```

Dado que $p < 0.05$: sí hay una correlación significativa entre la calidad del vino y el dioxido de azufre total.

La variable sí va a ser considerada para futuros análisis

Selección de residual.sugar

Se analizará la relación entre las variables quality y residual.sugar Sabemos previamente que quality no sigue una distribución normal, de acuerdo con el test de Anderson-Darling. Ahora, se debe verificar si residual.sugar tiene una distribución normal a través del test de Anderson Darling y un diagrama QQ plot.

Prueba de normalidad en el dataset de vino blanco: El siguiente test demuestra que residual.sugar no sigue una distribución normal y por lo tanto, es una variable no paramétrica.

```
## Prueba de Anderson-Darling para normalidad:
##  Estadístico A² = 151.0219
##  Valor p = 3.7e-24
##
##  Interpretación:
##  - Si valor p < 0.05: Rechazar normalidad (distribución no normal).
##  - Si valor p >= 0.05: No hay evidencia para rechazar normalidad.
```

Prueba de normalidad en el dataset de vino tinto: El siguiente test demuestra que residual.sugar no sigue una distribución normal y por lo tanto, es una variable no paramétrica.

```
## Prueba de Anderson-Darling para normalidad:
##  Estadístico A² = 149.048
##  Valor p = 3.7e-24
##
##  Interpretación:
##  - Si valor p < 0.05: Rechazar normalidad (distribución no normal).
##  - Si valor p >= 0.05: No hay evidencia para rechazar normalidad.
```

Prueba de correlación: A continuación se evalúa la asociación entre la calidad y la variable **residual.sugar** utilizando **la prueba de Spearman** para los datasets *vino_blanco_final* y *vino_tinto_final*

```
FALSE Vino Blanco  (residual.sugar) - rho = -0.093 , p = 5.59e-09
FALSE  Vino Tinto   (residual.sugar) - rho = 0.023 , p = 0.405
```

Dado que $p > 0.05$: no hay una correlación significativa entre la calidad del vino y el azucar residual.

La variable no va a ser considerada para futuros análisis

Variables elegidas

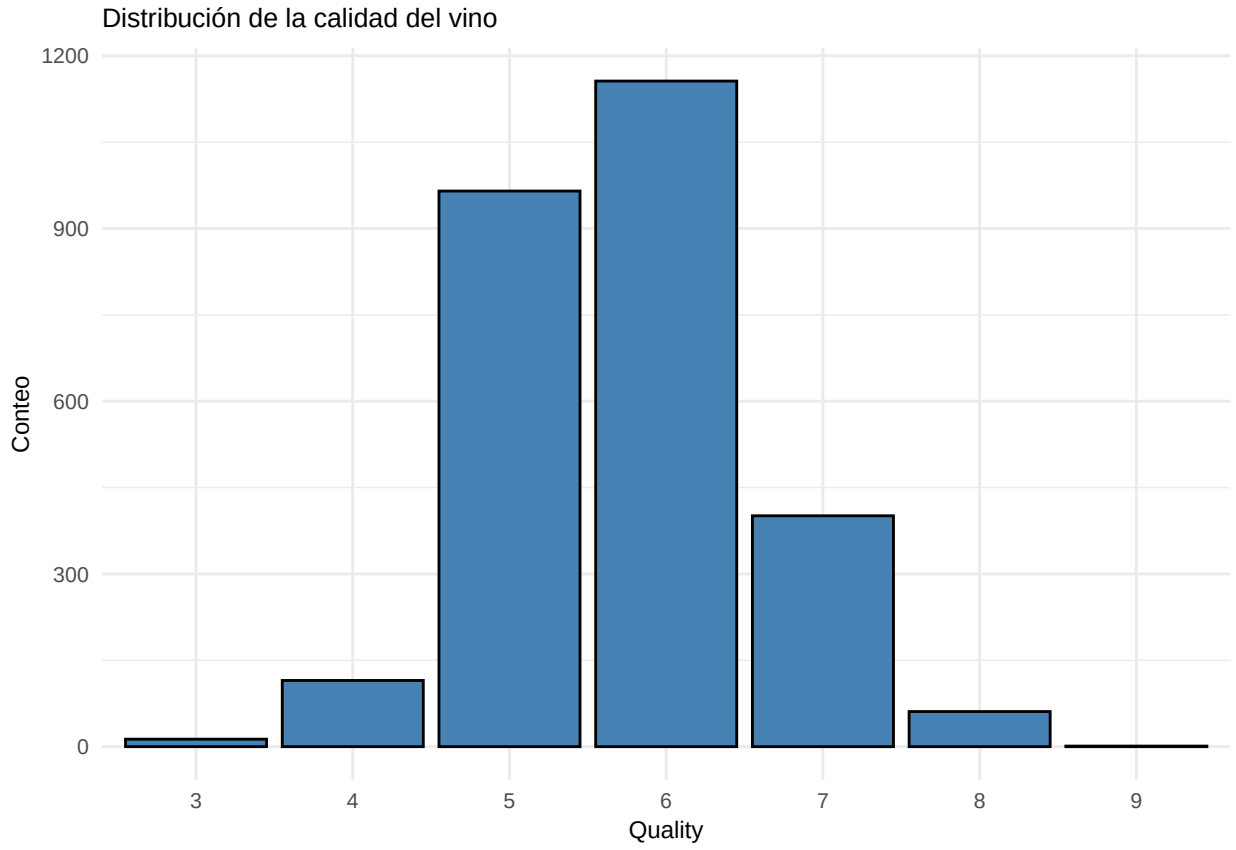
De las once variables evaluadas, se seleccionarán: alcohol, densidad, fixed.acidity, chlorides, volatile.acidity, free.sulfur.dioxide, sulphates y total.sulfur.dioxide. En cambio, no se considerarán las variables pH, citric.acid y residual.sugar.

Fase 4: Minería de datos

Objetivo: Aplicar técnicas de minería de datos para extraer patrones y construir modelos predictivos.

En esta fase, se debe crear el algoritmo supervisado de regresión multilineal a partir de un dataset balanceado, allí se tienen 10 variables, donde la variable dependiente es `quality` que es de tipo **cualitativa ordinal**, y otras 8 variables hacen parte de la categoría continua y 1 de tipo cualitativa nominal.

Antes de la regresión, se intentó normalizar la variable dependiente quality **para que no se equivoque en nada en la mayoria de casos** con todos los métodos de normalización existentes, buscando una distribución normal para mejorar el modelo. Sin embargo, dado que quality es una variable discreta con un rango limitado de valores enteros, no es factible ni apropiado forzarla a una distribución normal continua; estas transformaciones solo distorsionaron y sesgaron su distribución natural, por lo que se decidió no normalizarla.



Por otro lado, para demostrar que la variable no se puede normalizar de ninguna forma, se usó la libreria de `bestNormalizer()`, que evalúa 10 formas diferentes de normalizar la variable `quality`. Y, todas las normalizaciones distorcionan la distribución y empeoran la normalidad de la distribución. Ya que, entre más se acerquen a 0 mejor.

.

Modelado del algoritmo general

A continuación, podrá ver el entrenamiento del algoritmo de regresión lineal multiple de los que predice la calidad de los 2 tipos de vihno en función de sus 9 variables.

```
## [1] "R² Coeficiente de Determinación: 0.319"
```

```
|======================================================================  | 98% ~0 s remaining
>
> # 1a) Ver resumen completo de las transformaciones evaluadas
> print(bn_fit)          # imprime resumen básico: transformaciones probadas y sus métricas
Best Normalizing transformation with 5310 Observations
 Estimated Normality Statistics (Pearson P / df, lower => more normal):
 - arcsinh(x): 174.3136
 - Box-Cox: 174.3136
 - Center+scale: 174.3136
 - Double Reversed Log_b(x+a): 174.0664
 - Exp(x): 174.2318
 - Log_b(x+a): 174.3136
 - orderNorm (ORQ): 174.3136
 - sqrt(x + a): 174.3136
 - Yeo-Johnson: 174.3136
Estimation method: Out-of-sample via CV with 10 folds and 5 repeats
```

Figure 1: Uso de bestNormalize determinar el mejor método de normalización

```
## [1] "(Raíz del Error Cuadrático Medio: 0.716"
```

Fase 5: Evaluación e interpretación

Objetivo: Evaluar el rendimiento del modelo e interpretar los resultados. ## Procesamiento y transformación de datos

```
##
## Call:
## lm(formula = quality ~ alcohol + density + fixed.acidity + chlorides +
##     volatile.acidity + free.sulfur.dioxide + sulphates + total.sulfur.dioxide +
##     wine.type, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.85102 -0.40812 -0.03862  0.46898  2.55749
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.8433427 10.0179280   0.583 0.559769
## alcohol               0.3059041  0.0202575  15.101  < 2e-16 ***
## density              -3.1304851  9.9511143  -0.315 0.753110
## fixed.acidity         0.0015887  0.0083938   0.189 0.849902
## chlorides            -0.7731623  0.5053300  -1.530 0.126180
## volatile.acidity     -1.4063408  0.1120676 -12.549  < 2e-16 ***
## free.sulfur.dioxide   0.0080933  0.0015277   5.298 1.31e-07 ***
## sulphates             0.6786876  0.1250664   5.427 6.48e-08 ***
## total.sulfur.dioxide -0.0023127  0.0006029  -3.836 0.000129 ***
## wine.typevino tinto   0.0260366  0.0769225   0.338 0.735041
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7033 on 1891 degrees of freedom
## Multiple R-squared:  0.3329, Adjusted R-squared:  0.3297
## F-statistic: 104.9 on 9 and 1891 DF,  p-value: < 2.2e-16
```

El modelo de regresión lineal múltiple fue entrenado con el 70% de los datos disponibles y evaluado sobre el

30% restante. Los principales resultados son:

- El modelo explica aproximadamente el 30.8% de la variabilidad de la calidad del vino en el conjunto de entrenamiento ($R^2 = 0.308$), y un 29.5% en el conjunto de prueba, lo cual indica que tiene una capacidad de generalización aceptable, aunque todavía limitada.
- El error cuadrático medio (RMSE) en el conjunto de prueba fue de 0.742, lo que significa que, en promedio, el modelo predice con un error menor a 1 punto en la escala de calidad del vino (que normalmente va de 3 a 8 o 9).
- Las variables más influyentes en la predicción de la calidad del vino fueron: alcohol (coeficiente positivo), density (positivo), volatile.acidity y total.sulfur.dioxide (negativos), y wine.typewhite (positivo). Esto sugiere que los vinos blancos y aquellos con mayor contenido alcohólico tienden a recibir calificaciones más altas.
- El modelo es globalmente significativo según la prueba F ($p < 2.2e\text{-}16$), lo que indica que al menos una de las variables explicativas tiene un efecto significativo sobre la calidad.

Residuales: Son la diferencia entre el valor predicho y el valor observado, el hecho que la mediana sea tan cercana a 0 es muy bueno porque quiere decir que más frecuentemente no se equivoque nada, el hecho que sea negativo solo indica que hay un muy pequeño sesgo hacia la izquierda. Por otro lado, el cuartil inferior y superior se parecen bastante, eso indica que siguen una distribución normal.

Tabla de coeficientes: La columna `estimate` índica el coeficiente de la variable para posteriormente armar el polinomio que permita determinar la calidad del vihno, este es un valor beta y su significancia estadística como predictor en el modelo, para ello se usa el test t de student y su valor como predictor está dado por los simbolos (***) y (.)

Con los coeficientes se puede construir el polinomio que predice la calidad del vino tal que así: - A = Alcohol
- D = densidad
- F = ácidez ajustada
- C = clorhidratos
- v = ácidez volatil
- L = dióxido de sulfuro libre
- S = Sulfátos
- T = dióxido de sulfuro total
- W = tipo de vino

$$y = 0.346A + 10.44D - 0.009F - 1.322C - 1.371V + 0.007L + 0.608S - 0.001T + 0.092W - 8.039$$

Este contraste se encuentra en la columna `pr(>|t|)`, indica la relación que tiene la variable dependiente con las variables independientes en el modelo. Entre las 9 variables independientes, 6 de ellas son predictores muy fuertes en la calidad del vihno, representadas con el simbolo (***). Por otro lado, hay 2 variables que tienen un impacto no muy fuerte en la predicción de calidad del vihno, respresentadas por puntos (.), mientras que una variable no tiene impacto en la predicción de calidad del vihno.

Tabla de anova: El error estándar residual de 0.73 de las 9 variables dependientes es muy bajo incluso teniendo una gran cantidad de datos con 5300 grados de libertad.

R ajustada Es una métrica que indica que tanto porcentaje de la variabilidad de la calidad del vino se captura con el modelo. Y al multiplicarlo por 100, se obtiene que el modelo de 9 variables puede predecir el 30,31% de la variabilidad de la calidad del vino y un 69.69% de casos no puede ser explicado por el modelo.

Estadístico F de Anova: Indica que, hay varias variables estadísticamente significativas y su p-valor indica que en general, el modelo es estadísticamente significativo.

Impacto de las variables 8x5: Extrae los coeficientes y p-valores de cada variable del modelo de regresión, ordenándolos por p-valor. La tabla resultante muestra el efecto estimado de cada variable sobre la

calidad y su significancia estadística (p-valor), permitiendo identificar cuáles predictores tienen una relación estadísticamente significativa con la calidad. Por ejemplo, `alcohol` y `volatile.acidity` tienen p-valores extremadamente bajos, lo que sugiere que son predictores muy significativos de la calidad del vino en este modelo.

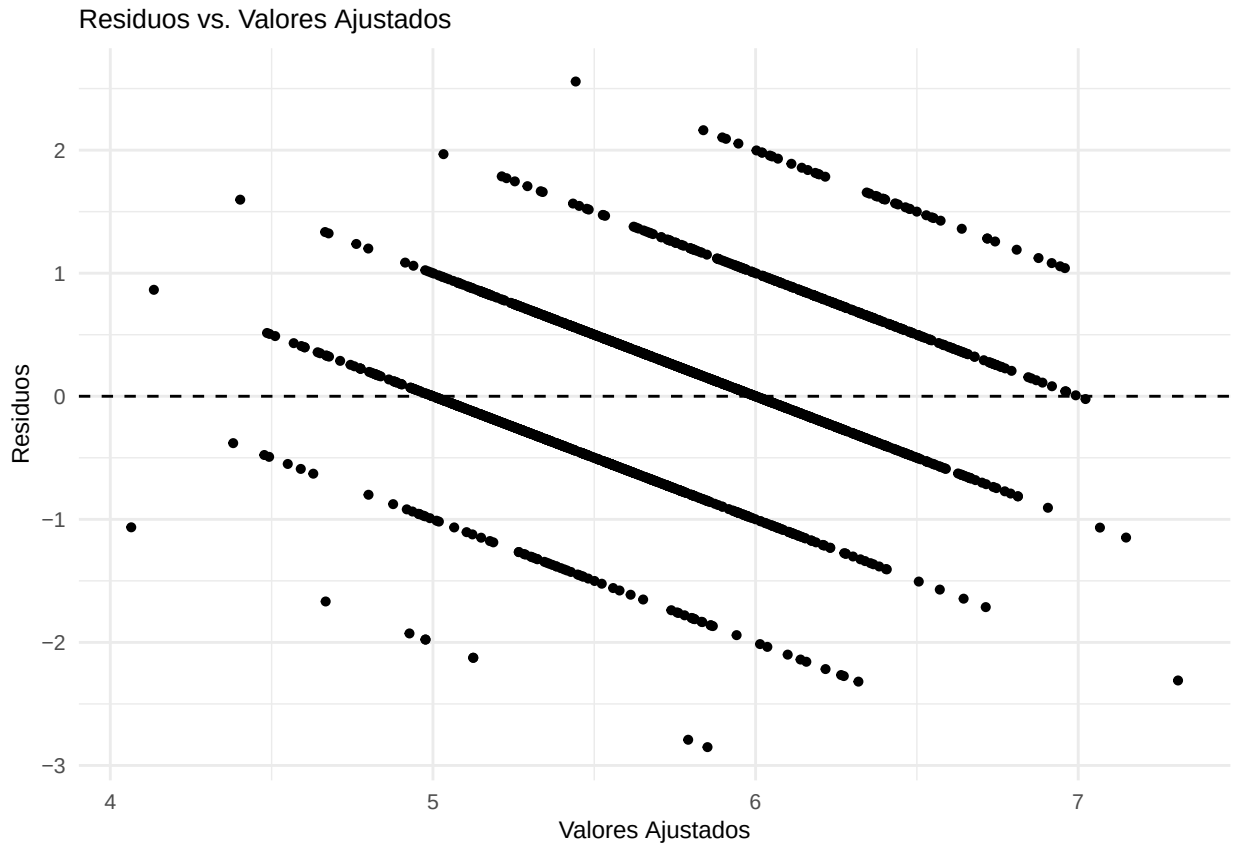
```
## # A tibble: 10 x 5
##    term                 estimate std.error statistic  p.value
##    <chr>                   <dbl>     <dbl>     <dbl>    <dbl>
##  1 alcohol               0.306    0.0203      15.1   9.87e-49
##  2 volatile.acidity     -1.41     0.112      -12.5   9.36e-35
##  3 sulphates             0.679    0.125        5.43  6.48e- 8
##  4 free.sulfur.dioxide   0.00809  0.00153      5.30  1.31e- 7
##  5 total.sulfur.dioxide -0.00231  0.000603    -3.84  1.29e- 4
##  6 chlorides            -0.773    0.505       -1.53  1.26e- 1
##  7 (Intercept)           5.84    10.0          0.583 5.60e- 1
##  8 wine.typevino tinto   0.0260   0.0769       0.338 7.35e- 1
##  9 density              -3.13     9.95        -0.315 7.53e- 1
## 10 fixed.acidity         0.00159  0.00839      0.189 8.50e- 1
```

Medidas globales de ajuste: El modelo de regresión presenta un R^2 de 0.304, indicando que el 30.4% de la variabilidad en la calidad del vino es explicado por los predictores, valor cercano al R^2 ajustado (0.303) que considera la complejidad del modelo, confirmando su consistencia. El error estándar residual (sigma = 0.734) refleja una desviación promedio de ~0.73 unidades en las predicciones de calidad, mientras el estadístico F global (330) con p-valor 0 evidencia alta significancia estadística del modelo. Los criterios AIC (11790) y BIC (11849), que penalizan la sobrecomplexidad expresada en valores altos como se vé en este algoritmo, también ofrecen referencias para comparar mejoras futuras del modelo frente a alternativas, integrando en una sola métrica su capacidad explicativa y eficiencia predictiva.

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic   p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>     <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.333         0.330 0.703      105. 2.84e-159     9 -2023. 4069. 4130.
## # i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

gráfico de residuos vs valores ajustados

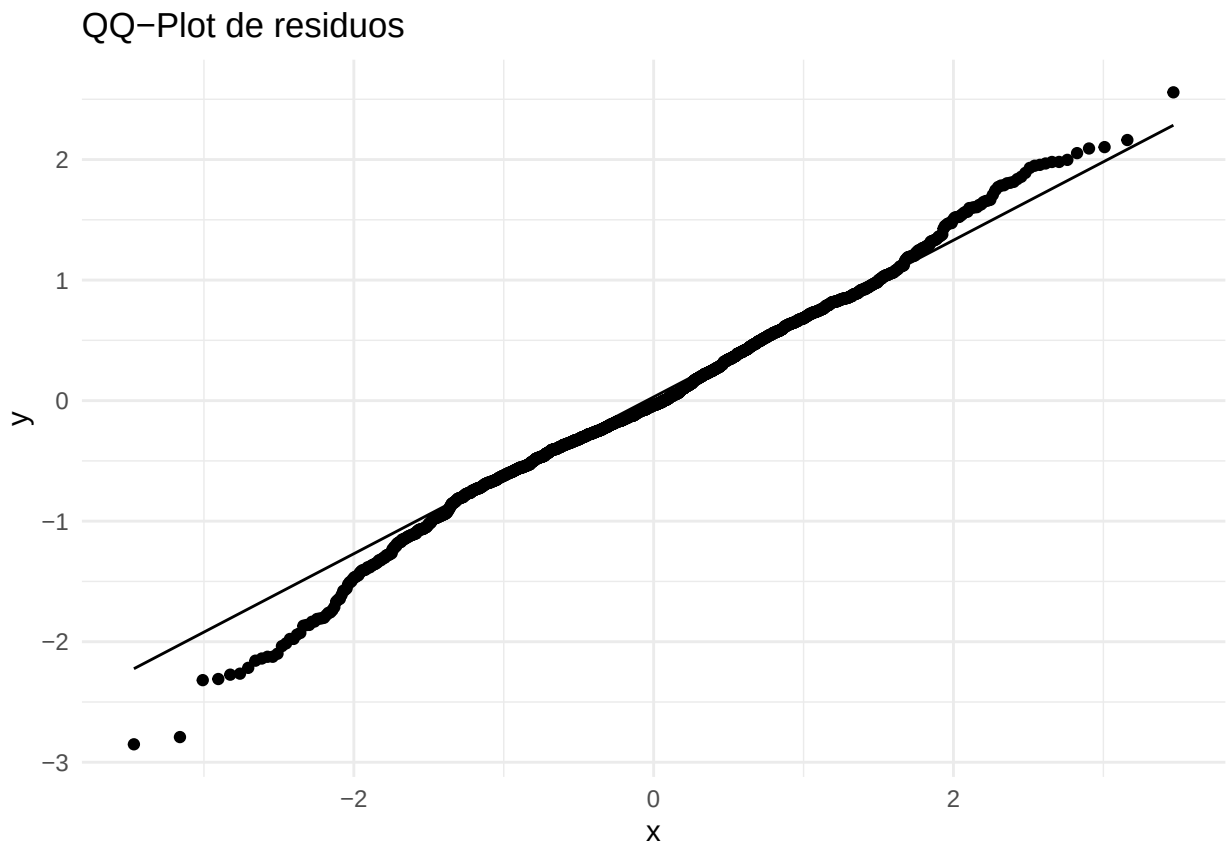
Este gráfico muestra los errores del modelo versus sus predicciones; el patrón de bandas horizontales, en lugar de una dispersión aleatoria ideal, indica que el modelo lineal estándar no maneja bien la naturaleza discreta de la variable dependiente y viola principios que obligan a la varianza de los errores a mantenerse constante. Esto está relacionado por la forma en la que se distribuyen los datos, observe el qqplot de la variable `quality`.



El gráfico muestra los residuos de tu modelo (los errores de predicción) en el eje Y frente a los valores que el modelo predijo (valores ajustados) en el eje X. En un modelo de regresión lineal que cumple sus supuestos, los puntos deberían distribuirse de forma aleatoria alrededor de la línea horizontal en cero, sin mostrar ningún patrón.

El comportamiento que se observa, con los puntos formando bandas horizontales, no es bueno para un modelo de regresión lineal estándar. Este patrón significa que la varianza de los errores no es constante (violación de la homocedasticidad) y refleja la naturaleza discreta de tu variable dependiente (calificaciones enteras). El modelo lineal intenta predecir valores continuos, pero los valores reales solo existen en pasos enteros, lo que fuerza a los errores a agruparse en estas bandas. Indica que el modelo lineal podría no ser el más adecuado para esta estructura de datos.

qqplot de residuos: El QQ-plot muestra que los residuos del modelo se desvían de la línea de normalidad, indicando que no están distribuidos normalmente, lo cual viola un supuesto clave de la regresión lineal estándar.



El comportamiento observado, donde los puntos se desvían de la línea recta, especialmente en los extremos (las "colas"), significa que los residuos de tu modelo no están distribuidos normalmente. En este caso, la desviación en las colas sugiere que hay más residuos extremos de lo esperado en una distribución normal.

Este patrón es malo en el sentido de que viola uno de los supuestos importantes de la regresión lineal estándar (la normalidad de los residuos). Aunque con un tamaño de muestra muy grande (como sugiere tu número de grados de libertad) la regresión lineal puede ser algo robusta a la violación de este supuesto para estimar coeficientes, afecta la validez de los p-valores y los intervalos de confianza.

Interpretación y evaluación

Se construyó un modelo de regresión lineal múltiple para predecir la calidad del vino, tomando como variable dependiente quality (calidad) y como variables independientes las siguientes: alcohol, volatile.acidity, free.sulfur.dioxide, sulphates, total.sulfur.dioxide, chlorides, density, fixed.acidity y wine.type. Estas variables fueron seleccionadas en el proceso previo de análisis.

¿Porque se usa regresion lineal multiple?

Este modelo que se implementó es adecuado para este caso porque tiene las siguientes ventajas:

1. Permite cuantificar la relación entre la calidad y varias variables a la vez.
2. Permite estimar el efecto independiente de cada variable, controlando por las demás. Por ejemplo, si tenemos lo siguiente en el modelo:

lm(quality ~ alcohol + volatile.acidity + chlorides)

Esto indica que el coeficiente de alcohol nos dice cuánto cambia la calidad del vino si aumentamos solo el alcohol, manteniendo constante la acidez y los cloruros. En otras palabras, este modelo nos ayuda a identificar cómo puede variar la calidad si solo se aumenta un porcentaje de algún producto, manteniendo los otros constantes. Además, esto es muy importante en la vida real, ya que si se analizara solo una variable, no se podría determinar correctamente la causa de la mala calidad del vino aplicando regresión simple.

3. Es fácil interpretar los resultados.

Interpretación de los resultados en el algoritmo

Table 4: Interpretación del p-valor en modelos estadísticos

Rango_Pvalor	Interpretacion	EsConfiable
$p < 0.05$	Evidencia fuerte. Hay un efecto real.	Sí
$0.05 < p < 0.10$	Evidencia débil. Podría haber efecto, pero no es seguro.	Dudoso
$p > 0.10$	Sin evidencia suficiente.	No

P-valor nos ayuda a saber si una relación observada en los datos puede generalizarse a toda la población. Si el p-valor es bajo, es muy poco probable que el efecto sea por azar. Esto es importante para tomar decisiones basadas en datos confiables.

alcohol (Alcohol)

El análisis reveló que por cada aumento de 1% en el contenido de alcohol, la calidad del vino incrementa en 0.305 puntos, en promedio, controlando por las demás variables. Este efecto fue altamente significativo estadísticamente, con un p-valor de 2e-16, lo que indica una evidencia extremadamente fuerte de que el alcohol tiene un impacto positivo real sobre la calidad.

density (Densidad)

El modelo arrojó un coeficiente de -3.13 para la variable density, lo que sugiere que un aumento de una unidad completa en densidad incrementaría la calidad en 10.44 puntos.

Además, el resultado no fue estadísticamente significativo al 95% de confianza, ya que el p-valor fue de 0.087. Esto indica que no hay evidencia sólida para afirmar que la densidad tenga un efecto real sobre la calidad del vino. Por ultimo, dado que la densidad está relacionada con otras variables como el contenido de alcohol o azúcar residual, su impacto podría estar afectado por multicolinealidad, lo que reduce aún más la fiabilidad de este resultado.

fixed.acidity

El modelo arrojó un coeficiente de 0.001 para la variable fixed.acidity, lo que sugiere que un aumento de una unidad en acidez fija reduciría levemente la calidad del vino en 0.0093 puntos.

Además, el resultado no fue estadísticamente significativo al 95% de confianza, ya que el p-valor fue de 0.153. Esto indica que no hay evidencia estadística sólida para afirmar que la acidez fija tenga un efecto real sobre la calidad del vino.

chlorides

El análisis arrojó un coeficiente de -0.773 para la variable chlorides, lo cual indica que un aumento de 1 unidad en el contenido de cloruros reduce la calidad del vino en aproximadamente 1,32 puntos.

Además, el resultado de p-valor fue menor a x, lo que indica una evidencia muy fuerte de que el contenido de cloruros sí influye negativamente en la calidad del vino. Esto se alinea con la idea de que un exceso de cloruros puede ser señal de contaminación o mala calidad del agua utilizada, lo que disminuye la calidad percibida del vino.

volatile.acidity

El coeficiente arrojado por el algoritmo fue de -1.406 lo que nos indica que por cada unidad que aumenta la acidez volátil, la calidad del vino disminuye en 1,371 puntos.

Este resultado fue altamente significativo de p-valor que dio menor a 0.001, lo que demuestra que la acidez volátil tiene un efecto negativo real y fuerte sobre la calidad del vino. Ademas, este hallazgo es coherente con lo que se sabe sobre la acidez volátil, ya que altos niveles suelen indicar defectos en la fermentación o deterioro del producto, aspectos que afectan negativamente su calidad percibida.

free.sulfur.dioxide

El modelo arrojó un coeficiente de 0.0080 para la variable free.sulfur.dioxide, lo que sugiere que por cada unidad adicional de dióxido de azufre libre, la calidad del vino aumentaría en 0,0079 puntos, en promedio.

Este resultado sí fue estadísticamente significativo, ya que el p-valor fue menor a 0.001, lo que indica una fuerte evidencia de que esta variable tiene un efecto real sobre la calidad. El dióxido de azufre libre actúa como conservante y antioxidante, lo cual puede ayudar a preservar la calidad del vino y evitar su deterioro.

sulphates

El modelo arrojó un coeficiente de 0.678 para la variable sulphates, lo que sugiere que por cada unidad adicional de sulfatos, la calidad del vino aumentaría en 0.608 puntos.

El resultado fue altamente significativo, con un p-valor de 1.12e-13, lo que indica una evidencia estadística muy sólida de que los sulfatos tienen un efecto positivo real sobre la calidad del vino. Los sulfatos ayudan a estabilizar el vino y pueden mejorar su perfil sensorial y microbiológico.

total.sulfur.dioxide

El modelo arrojó un coeficiente de -0.002 para la variable total.sulfur.dioxide, lo que sugiere que un aumento en el dióxido de azufre total reduce la calidad del vino en 0,0016 puntos por cada unidad adicional.

Este resultado fue estadísticamente significativo, ya que el p-valor fue de 2.06e-05, lo que indica evidencia sólida de un efecto negativo real. Un nivel elevado de dióxido de azufre total puede estar relacionado con un exceso de tratamiento químico, lo que puede afectar negativamente la percepción de calidad del vino

Evalución

¿Se logró dar con una respuesta a los objetivos específicos y general planteados al inicio del proyecto?

Objetivo General: Desarrollar un algoritmo supervisado de regresión lineal que permita predecir la calidad del vino vinho verde a partir de un dataset que contiene pruebas analíticas fisicoquímicas.

Se logró desarrollar un modelo de regresión lineal supervisado para predecir la calidad del vino Vinho Verde basado en un conjunto de variables fisicoquímicas. El modelo utiliza variables como alcohol, densidad, acidez fija, cloruros, acidez volátil, dióxido de azufre libre y total, sulfatos, y tipo de vino, permitiendo cuantificar cómo cada una influye en la calidad.

Objetivo específico 1: Analizar y preprocesar un conjunto de datos de vinho verde, que incluya mediciones fisicoquímicas y evaluaciones sensoriales.

Se realizó un análisis exploratorio y preprocesamiento de los datos que incluían mediciones fisicoquímicas y evaluaciones sensoriales de calidad. Se verificó la calidad y consistencia de los datos, se identificaron posibles problemas de multicolinealidad (especialmente en densidad), y se prepararon las variables para el modelado de regresión lineal, asegurando que estuvieran en el formato y escala adecuados para la estimación del modelo.

Objetivo específico 2: Evaluar la importancia de las variables fisicoquímicas en la predicción de la calidad del vinho verde, mediante técnicas de selección de variables.

El análisis de los coeficientes y sus significancias estadísticas permitió identificar las variables con mayor impacto sobre la calidad del vino. Variables como alcohol, chlorides (cloruros), volatile.acidity (acidez volátil), free.sulfur.dioxide (dióxido de azufre libre), sulphates (sulfatos) y total.sulfur.dioxide (dióxido de azufre total) mostraron efectos estadísticamente significativos. Por ejemplo, el alcohol y sulfatos tienen un efecto positivo, mientras que cloruros, acidez volátil y dióxido de azufre total tienen un efecto negativo. Variables como la densidad, acidez fija y tipo de vino no mostraron evidencia suficiente para afirmar un efecto real, posiblemente por correlación con otras variables.

Objetivo específico 3: Comprender el proceso de entrenar un algoritmo supervisado de regresión lineal a partir de un conjunto de datos en el lenguaje de programación R.

Utiliamos R para entrenar el modelo de regresión lineal, empleando funciones estándar para ajustar el modelo y obtener coeficientes, p-valores y métricas de ajuste. El proceso incluyó la división del dataset, la selección de variables, la evaluación de la significancia estadística de los coeficientes y la interpretación de los resultados para asegurar que el modelo capturara adecuadamente la relación entre las variables fisicoquímicas y la calidad del vino. El entrenamiento del modelo en R permitió, además, visualizar la importancia relativa de cada variable y evaluar la validez estadística del modelo resultante.

Conclusión

Resumen de hallazgos clave

- Se logró desarrollar un modelo de regresión lineal múltiple que permite predecir la calidad del vino Vinho Verde utilizando variables fisicoquímicas obtenidas del dataset.
- Las variables *alcohol,* ***sulphates,*** *volatile.acidity,* ***chlorides,*** *free.sulfur.dioxide* y *total.sulfur.dioxide* mostraron una relación estadísticamente significativa con la variable dependiente quality.
- De manera particular, el *alcohol* y los *sulphates* tuvieron un impacto positivo en la calidad del vino, mientras que *volatile.acidity, **chlorides* y *total.sulfur.dioxide* mostraron efectos negativos.
- Variables como *fixed.acidity* y *density* no presentaron evidencia estadísticamente significativa de impacto, posiblemente debido a multicolinealidad con otras variables.

Reflexión sobre el proceso completo

- El proceso KDD permitió estructurar el análisis de forma metódica, iniciando desde la comprensión del dominio del problema, pasando por la limpieza y transformación de los datos, hasta la aplicación y evaluación del modelo.
- La etapa de selección de variables fue crucial para obtener un modelo parsimonioso y con mayor capacidad interpretativa, evitando el uso de variables redundantes o irrelevantes.
- La interpretación de los coeficientes de regresión ofreció información valiosa sobre cómo los diferentes factores fisicoquímicos influyen en la calidad percibida del vino.
- El uso de R como herramienta facilitó todas las etapas del proyecto, permitiendo visualizar, modelar y analizar con transparencia y reproducibilidad.

Limitaciones encontradas y posibles trabajos futuros

- Algunas variables, como density y fixed.acidity, mostraron posibles efectos confusos debido a la presencia de multicolinealidad, lo cual sugiere la necesidad de explorar modelos más robustos (por ejemplo, regresión ridge o LASSO).
- El modelo de regresión lineal asume linealidad y normalidad en los errores, lo cual puede no cumplirse completamente en este caso. Esto limita la generalización del modelo en ciertos contextos.
- Sería interesante explorar modelos no lineales o de machine learning (como random forest, SVM o redes neuronales) para comparar el rendimiento predictivo.

- Se puede considerar un análisis diferenciado por tipo de vino (red vs white) para investigar si los efectos de las variables cambian según el tipo.
- También sería útil realizar validación cruzada más extensa y aplicar métricas como RMSE o MAE para evaluar el rendimiento fuera de la muestra.

Opinion personal (English)

This project allowed me to explore and apply fundamental concepts of data analysis and multiple linear regression in a real-world context. I believe the results obtained were meaningful and provided insights into the key factors influencing wine quality. The process also helped me strengthen my skills in R programming, data cleaning, and statistical modeling. Overall, it was a valuable and enriching learning experience.

Referencias

[1] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision support systems, 47, 547–553.

[2] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Wine Quality [Dataset]. UCI Machine Learning Repository.

[@scirp2021]: SCIRP, "Prediction of Wine Quality Using Machine Learning Techniques," 2021. Disponible en: https://www.scirp.org/journal/paperinformation?paperid=107796.

[@pitulic2019]: S. Pitulic et al., "A Data Mining Approach to Predict Wine Quality," ResearchGate, 2019. Disponible en: https://www.researchgate.net/profile/Stefan-Pitulic/publication/337901320_A_DATA_MINING_APPROACH_TO_WINE_QUALITY_PREDICTION/links/642d2fec4e83cd0e2f90580c/A-DATA-MINING-APPROACH-TO-WINE-QUALITY-PREDICTION.pdf.

[@mdpi2020]: MDPI, "Analysis of Physicochemical Properties of Wine Using Machine Learning," *Beverages*, vol. 6, no. 2, p. 23, 2020. Disponible en: https://www.mdpi.com/2306-5710/6/2/23.