Pontificia Universidad Javeriana
Bogotá

Taller 3 – Procesamiento de datos a gran escala
Predicción de la calidad del vino

Fabian Andres Díaz
David Leon
Tomas Pinilla
Juan David Ramirez
Juan Diego Carreño

Pontificia Universidad Javeriana de Colombia
Procesamiento de datos a gran escala
Bogotá, Colombia
12 de noviembre de 2025

Índice general

Introducción y contexto

En este informe se presenta el desarrollo del Taller 3 de la asignatura *Procesamiento de datos a gran escala*. Se construyó un proceso completo de aprendizaje automático utilizando el conjunto de datos **Wine Quality** (vinos blancos y tintos de la región portuguesa *Vinho Verde*). El conjunto incluye 11 variables fisicoquímicas, una variable categórica (`type`) y una variable objetivo `quality` (valores 3 a 9). La descarga de referencia pública se encuentra en: `https://www.kaggle.com/datasets/rajyellow46/wine-quality`.

El archivo `winequalityN.csv` contiene 6 497 registros con 13 columnas. El desarrollo se realizó en Databricks y el presente documento evidencia EDA, preparación, modelado y validación.

Análisis exploratorio de datos (EDA)

Se verificó que 4 898 registros corresponden a vino blanco y 1 599 a vino tinto. La variable `quality` está sesgada hacia valores intermedios (5 y 6). Se detectaron valores ausentes marginales en columnas numéricas; se decidió imputarlos con la mediana.

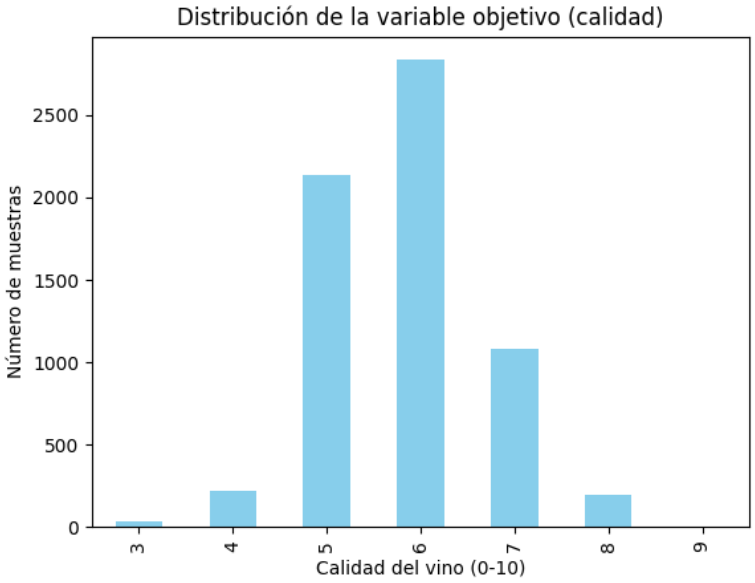

Figura 1: Distribución de `quality`. Las clases extremas (3 y 9) son poco frecuentes.

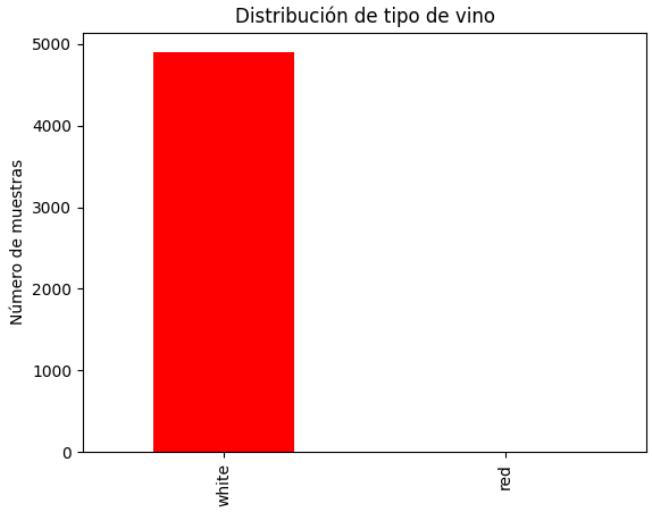

Figura 2: Proporción de vinos blancos y tintos.

Variables continuas

El contenido de alcohol oscila entre aproximadamente 8 % y 14 %, con ligera asimetría hacia valores bajos. El azúcar residual presenta cola larga; el pH se concentra entre 3,0 y 3,5.

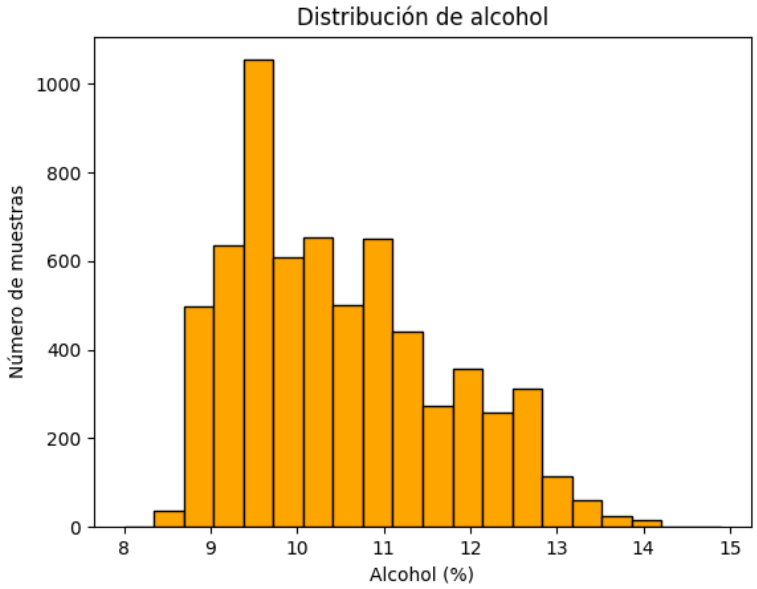

Figura 3: Distribución del contenido de alcohol.

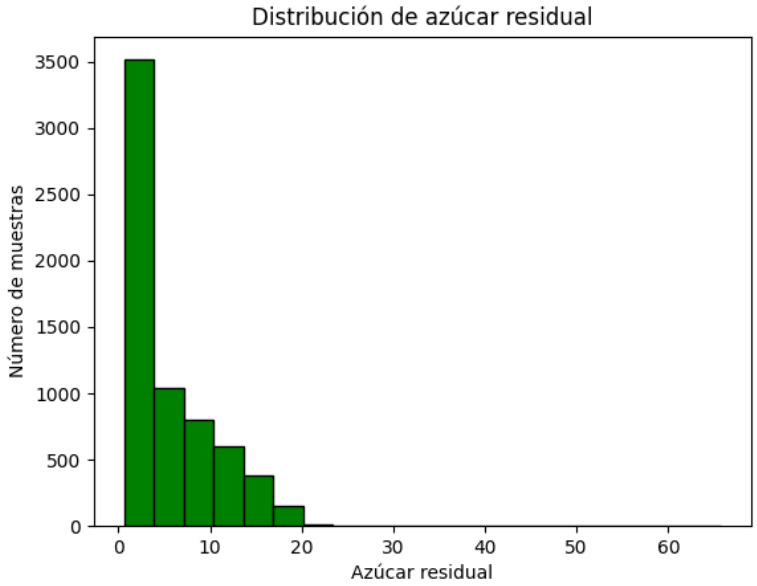

Figura 4: Distribución del azúcar residual.

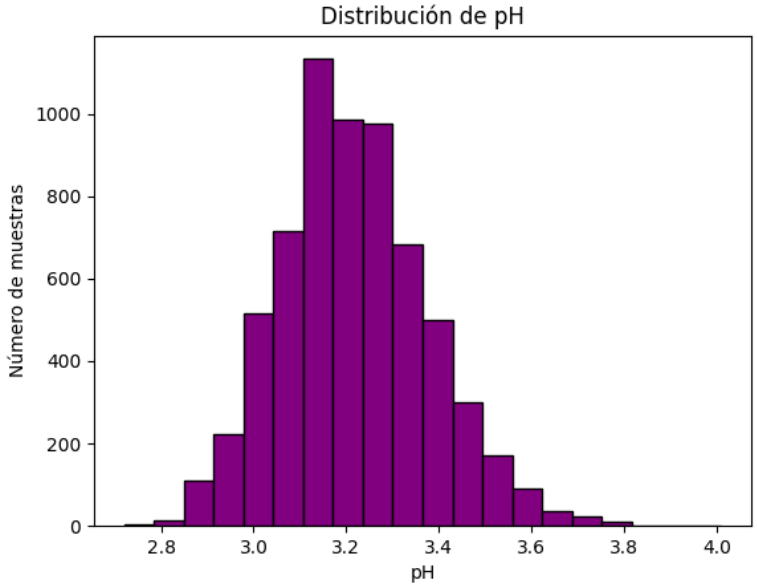

Figura 5: Distribución del pH.

Relaciones y correlaciones

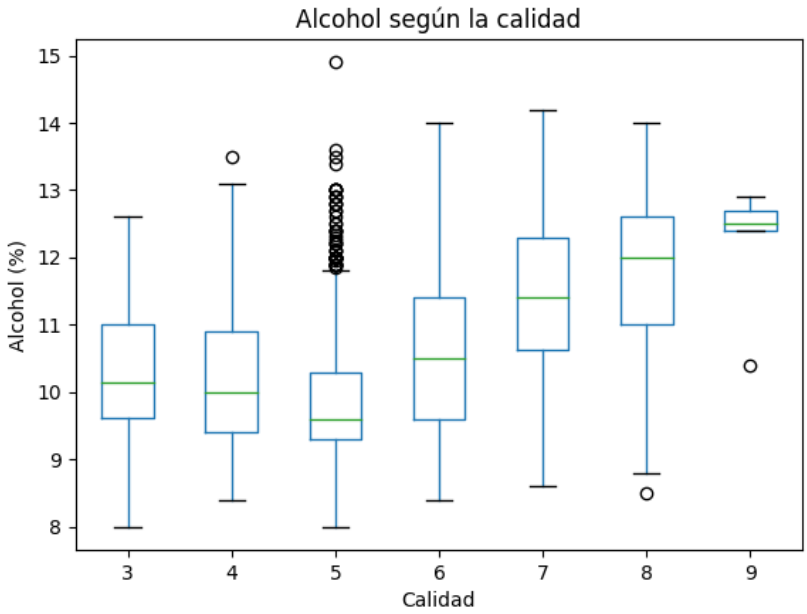

Figura 6: Alcohol por clase de `quality`. Tendencia ascendente para calidades más altas.

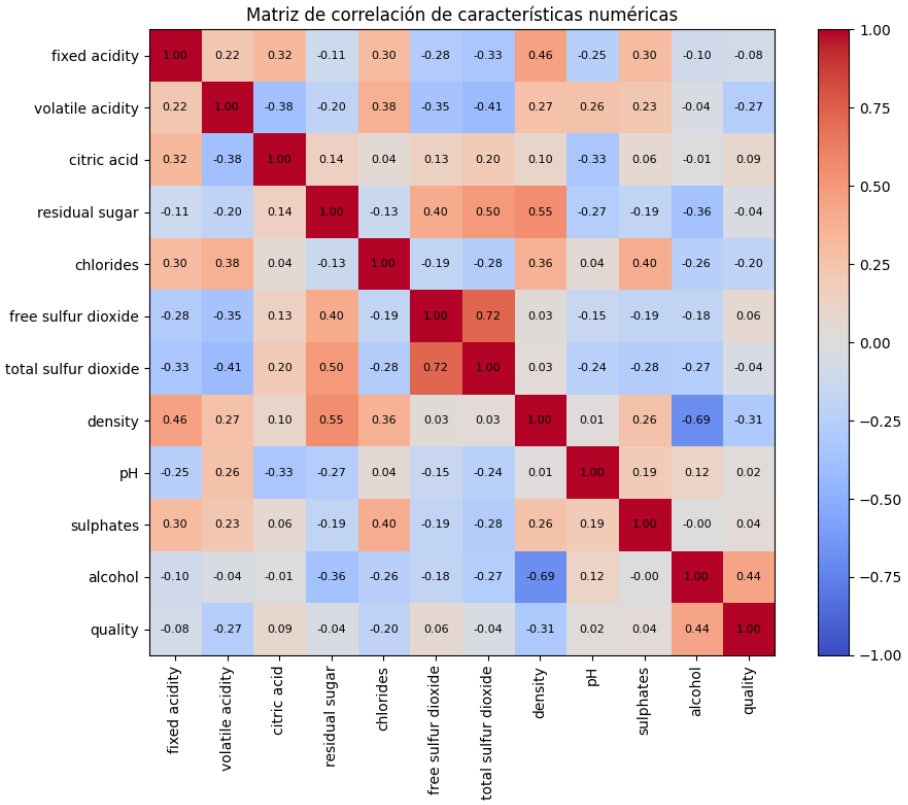

Figura 7: Matriz de correlación (variables numéricas).

Preparación de los datos

Se aplicó el siguiente *pipeline*:

1. **Imputación** de ausentes en variables numéricas con la mediana.
2. **Codificación** de `type` mediante *one-hot*.
3. **Estandarización** de atributos numéricos con `StandardScaler`.
4. **Split** estratificado 80/20 para `quality`.

Modelado y validación

Se evaluaron regresión logística, Bosque Aleatorio y Gradient Boosting. Métricas: exactitud y F1 ponderado. Se realizó búsqueda breve de hiperparámetros para los modelos basados en árboles.

Cuadro 1: Resumen de métricas para los modelos evaluados.

Modelo	**Hiperparámetros clave**	**Exactitud**	**F1 ponderado**
Regresión logística	Multinomial, `max_iter`=1000	0,537	0,507
Bosque Aleatorio (baseline)	$n = 100$, profundidad ilimitada	0,684	0,670
Gradient Boosting (baseline)	$n = 100$, tasa de aprendizaje 0,1	0,581	0,563
Bosque Aleatorio (ajustado)	$n = 400$, profundidad ilimitada	0,684	0,669
Gradient Boosting (ajustado)	$n = 200$, tasa 0,1, profundidad 3	0,595	0,581

El Bosque Aleatorio mostró el mejor desempeño global. Aumentar árboles de 100 a 400 no produjo mejoras relevantes; el F1 ponderado se mantuvo en torno a 0,67.

Informe de clasificación (resumen)

Las clases mayoritarias (5 y 6) alcanzaron F1 de 0,73 y 0,71. Las clases minoritarias (3 y 9) son difíciles de predecir por su baja frecuencia. Para mejorar estos casos podrían emplearse técnicas de *re-muestreo* o plantear el problema como regresión ordinal.

Conclusiones

- Se realizó un EDA que evidenció sesgo hacia calidades medias y asimetrías (azúcar residual).
- El *pipeline* maneja ausentes, codifica la variable categórica y normaliza numéricas.
- El Bosque Aleatorio fue el más robusto (exactitud $\approx 0{,}68$, F1 ponderado $\approx 0{,}67$).
- El desbalance limita el rendimiento en clases extremas; como trabajo futuro: balanceo, *cost-sensitive learning* o regresión ordinal.