Análisis Predictivo de la Calidad del Vino

Basado en Propiedades Fisicoquímicas

Carlos Iván Mareco Recalde

19 de noviembre de 2025

Índice

1. Objetivo

El objetivo principal de este trabajo es desarrollar un modelo de clasificación unificado capaz de predecir la calidad sensorial de las variantes tintas y blancas del vino portugués "Vinho Verde", utilizando sus propiedades fisicoquímicas objetivas.

Para lograr esto, se implementan dos estrategias clave:

1. La inclusión de la variable `wine_type` (tinto o blanco) como predictor, para capturar las diferencias estructurales entre ambas variantes.
2. La transformación de variables sesgadas para mitigar el impacto de valores atípicos detectados durante el análisis exploratorio.

La variable objetivo (calidad) se transformará de una puntuación numérica original (0-10) a tres categorías ordinales: **Basic**, **Good** y **Premium**.

2. Introducción y Datos

El conjunto de datos utilizado en este estudio, *Wine Quality*, es un recurso público disponible para investigación y fue presentado originalmente por Cortez et al. (2009) [1].

Se utilizaron dos conjuntos de datos: `winequality-red.csv` (1,599 vinos tintos) y `winequality-white.cs` (4,898 vinos blancos). Al combinarlos, el conjunto de datos de estudio final consta de **6,497 observaciones** totales. Ambos comparten la misma estructura de 11 variables predictoras fisicoquímicas (ej. acidez, azúcar, alcohol) y 1 variable objetivo sensorial (`quality`).

3. Metodología: Preprocesamiento

3.1. Limpieza y Unificación

Se realizó la carga de datos validando los tipos numéricos y estandarizando la nomenclatura a *snake_case*. Ambos datasets se fusionaron en un dataframe maestro (`wine_unified`), añadiendo la columna categórica `wine_type`.

3.2. Ingeniería de Características

1. **Categorización de la Variable Objetivo:** La variable `quality` se transformó en `quality_class` (`Basic` ≤ 5, `Good` $= 6$, `Premium` ≥ 7) basándose en los cuartiles de su distribución.
2. **Transformación Logarítmica:** Tras detectar variables con fuerte sesgo positivo (ver Sección 4.5), se aplicó la transformación $\log(1+x)$ a las variables afectadas (ej. `residual_sugar`, `chlorides`) para normalizar su distribución y reducir el ruido estadístico.

4. Resultados: Análisis Exploratorio de Datos (EDA)

4.1. Análisis de la Variable Objetivo

La distribución original de la calidad (Tabla 1 y Figura 1) mostró un fuerte desequilibrio, con clases extremas que carecían de suficientes muestras para el modelado.

Tabla 1: Frecuencia Absoluta y Relativa de la Puntuación `quality` Original.

	Vino Tinto		Vino Blanco	
Quality	**N**	**Percent**	**N**	**Percent**
3	10	0.6 %	20	0.4 %
4	53	3.3 %	163	3.3 %
5	681	42.6 %	1457	29.7 %
6	638	39.9 %	2198	44.9 %
7	199	12.4 %	880	18.0 %
8	18	1.1 %	175	3.6 %
9	0	0.0 %	5	0.1 %

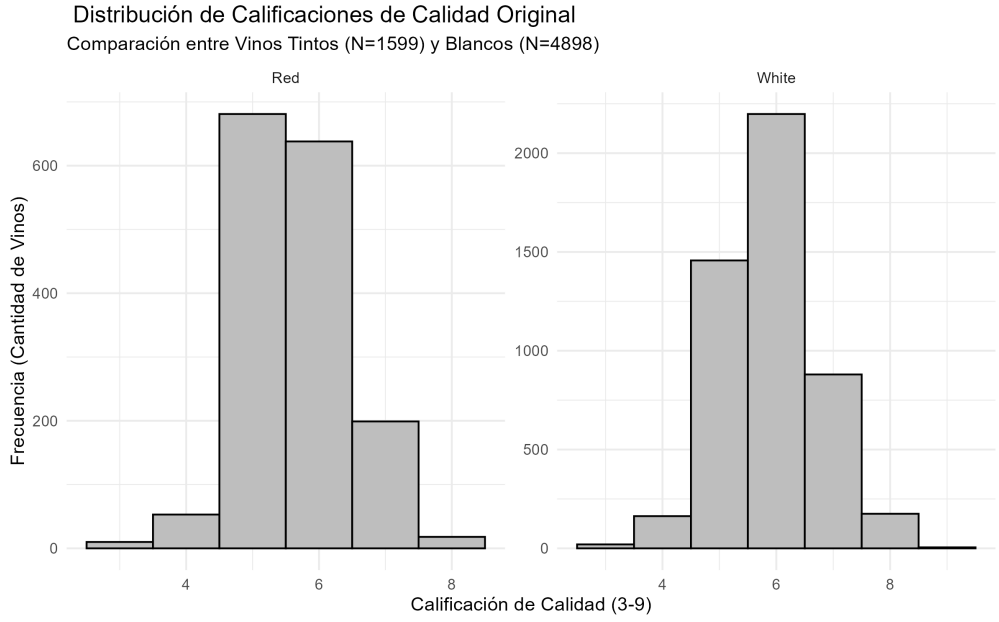

Figura 1: Distribución de las puntuaciones de calidad originales (Tinto vs. Blanco).

Esta evidencia justificó la agrupación en las tres clases mencionadas. Aún tras la agrupación, la clase `Premium` permanece como minoritaria ($\approx$ 19,7 % del total unificado).

4.2. Análisis de Desequilibrios Estructurales

El dataset presenta dos desequilibrios críticos. Primero, los vinos blancos representan el 75.4 % de la muestra (Figura 2), lo que requiere técnicas de validación cruzada estratificada.

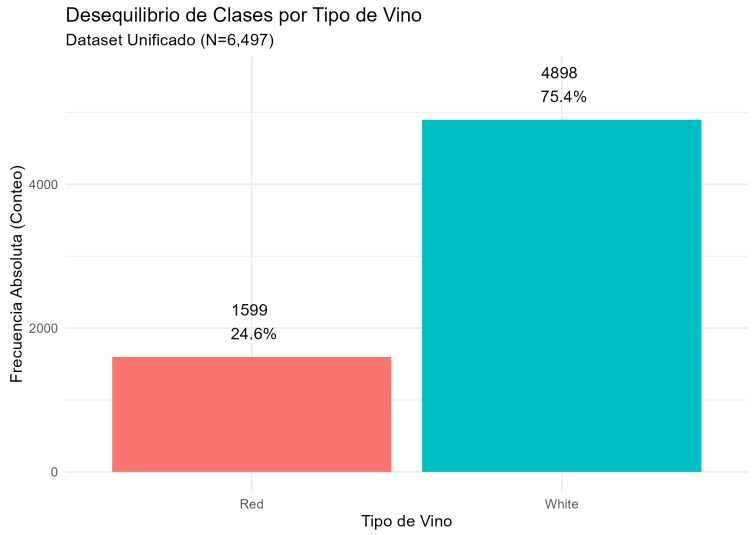

Figura 2: Desequilibrio de clases del predictor `wine_type`.

Segundo, la distribución de las clases de calidad finales también es desigual (Figura 3), lo que requerirá técnicas de validación estratificada.

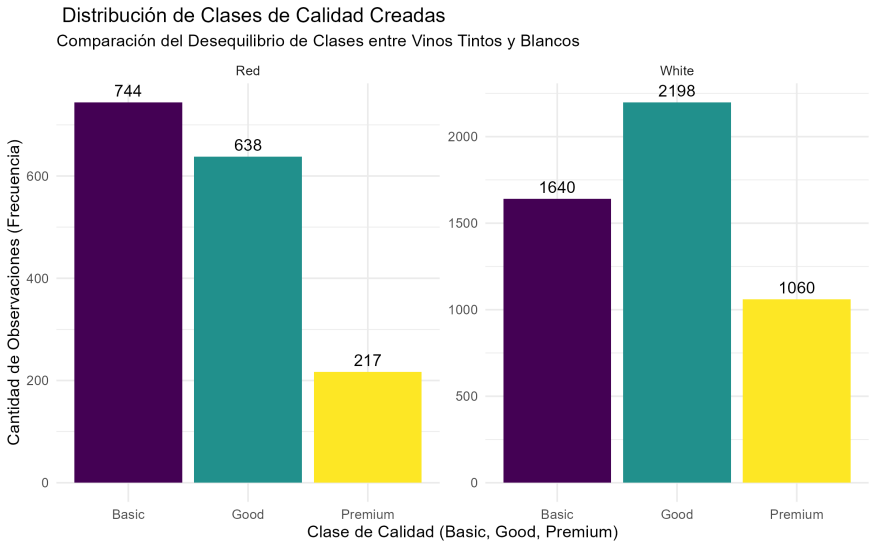

Figura 3: Distribución de las clases de calidad creadas (`Basic`, `Good`, `Premium`).

4.3. Análisis Físicoquímico Comparativo

El análisis estadístico (Tablas 2 y 3) confirmó diferencias estructurales: los tintos poseen mayor acidez volátil (media 0.528 vs 0.278) y los blancos mayor azúcar residual.

Tabla 2: Estadísticas Descriptivas: Vinos Tintos (N=1599).

Variable	Media	Mediana	SD	Mín	Máx
fixed_acidity	8.32	7.9	1.74	4.6	15.9
volatile_acidity	0.528	0.52	0.179	0.12	1.58
citric_acid	0.271	0.26	0.195	0	1.00
residual_sugar	2.54	2.2	1.41	0.9	15.5
chlorides	0.087	0.079	0.047	0.012	0.611
free_sulfur_dioxide	15.9	14	10.5	1	72
total_sulfur_dioxide	46.5	38	32.9	6	289
density	0.997	0.997	0.002	0.990	1.004
pH	3.31	3.31	0.154	2.74	4.01
sulphates	0.658	0.62	0.170	0.33	2.00
alcohol	10.4	10.2	1.07	8.4	14.9

Tabla 3: Estadísticas Descriptivas: Vinos Blancos (N=4898).

Variable	Media	Mediana	SD	Mín	Máx
fixed_acidity	6.85	6.8	0.844	3.8	14.2
volatile_acidity	0.278	0.26	0.101	0.08	1.10
citric_acid	0.334	0.32	0.121	0	1.66
residual_sugar	6.39	5.2	5.07	0.6	65.8
chlorides	0.046	0.043	0.022	0.009	0.346
free_sulfur_dioxide	35.3	34	17.0	2	289
total_sulfur_dioxide	138.0	134	42.5	9	440
density	0.994	0.994	0.003	0.987	1.039
pH	3.19	3.18	0.151	2.72	3.82
sulphates	0.490	0.47	0.114	0.22	1.08
alcohol	10.5	10.4	1.23	8.0	14.2

4.4. Análisis de Correlación: Diferencias Estructurales

Dado que los vinos tintos y blancos presentan composiciones distintas, se generaron matrices de correlación separadas (Pearson) para identificar los motores de calidad específicos de cada variante.

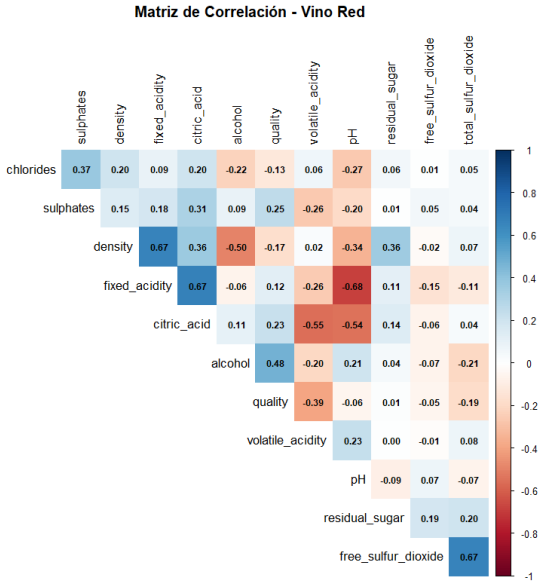

Figura 4: Vino Tinto: Correlaciones fuertes.

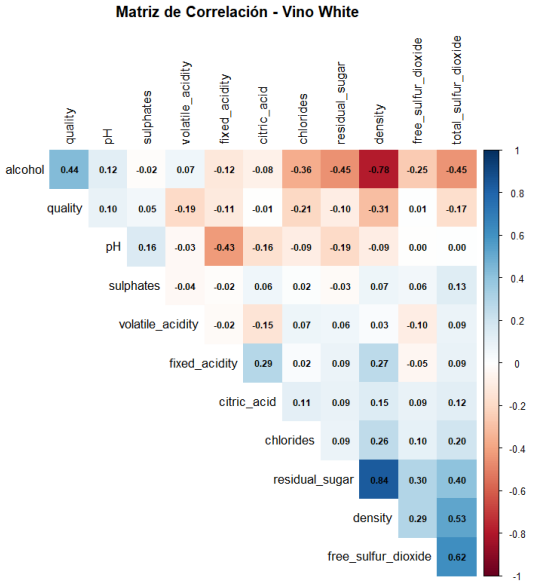

Figura 5: Vino Blanco: Correlaciones difusas.

El análisis comparativo (Figuras 4 y 5) revela hallazgos clave:

- **El Alcohol como Predictor Dominante:** En ambos tipos de vino, el `alcohol` se consolida como la variable con la mejor relación positiva con la calidad ($r = 0{,}48$ en tintos, $r = 0{,}44$ en blancos). Esto indica que, independientemente del tipo, un mayor grado alcohólico tiende a estar asociado con una mejor valoración sensorial.
- **El Rol de la Acidez Volátil:** En los vinos tintos, la `volatile_acidity` tiene una fuerte correlación negativa ($-0{,}39$), confirmando que es un defecto crítico. En los blancos, esta correlación es mucho más débil ($-0{,}19$), lo que sugiere una mayor tolerancia a este compuesto.
- **Densidad y Azúcar:** En los vinos blancos, existe una correlación positiva muy fuerte ($+0{,}84$) entre `density` y `residual_sugar`, reflejando que el cuerpo del vino blanco está dictado casi exclusivamente por el azúcar, a diferencia del tinto.

Esta divergencia confirma que un modelo único que no distinga entre tipos de vino fallaría en capturar los criterios de calidad específicos, validando la inclusión de la variable `wine_type` en el modelo final.

4.5. Cuantificación y Mitigación de Outliers

Se detectó una alta presencia de valores atípicos en variables como `residual_sugar` (9.7 % en tintos) y `citric_acid` (5.5 % en blancos).

Tabla 4: Conteo de Outliers por Variable (Vino Tinto).

Variable	Conteo	Porcentaje (%)
residual_sugar	155	9.7 %
chlorides	112	7.0 %
sulphates	59	3.7 %
total_sulfur_dioxide	55	3.4 %

Tabla 5: Conteo de Outliers por Variable (Vino Blanco).

Variable	Conteo	Porcentaje (%)
citric_acid	270	5.5 %
chlorides	208	4.2 %
volatile_acidity	186	3.8 %

Para determinar si estos *outliers* eran una çlase especial.º ruido, se investigaron visualmente (Figuras 6 y 7).

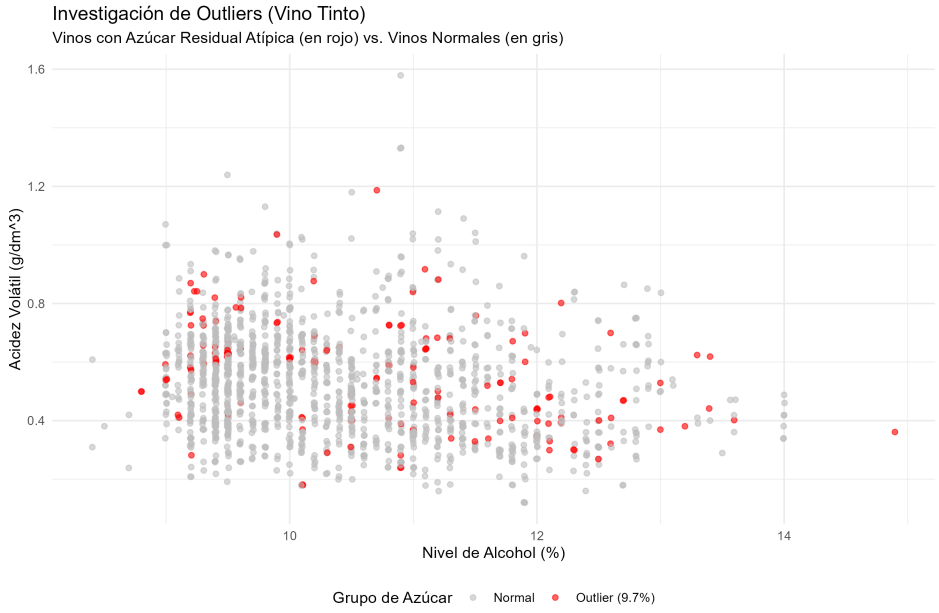

Figura 6: Investigación de Outliers (Tinto): Los puntos rojos (azúcar alto) no forman un clúster separado, sino ruido.

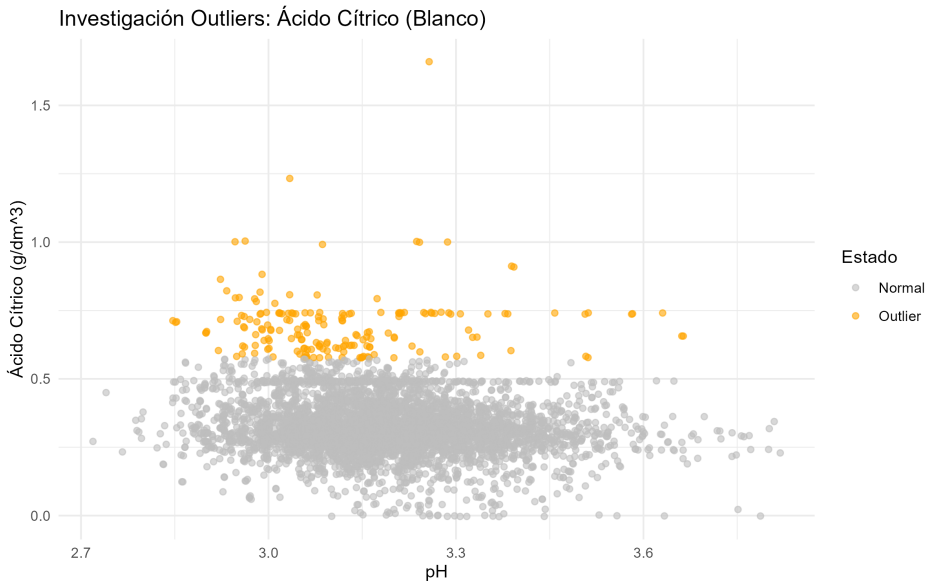

Figura 7: Investigación de Outliers (Blanco): Los puntos naranjas (ácido cítrico) también se comportan como ruido.

Ambos gráficos confirman que los valores extremos son ruido estadístico, validando la transformación logarítmica.

4.6. Visualización Final de Relaciones Clave

Utilizando las variables transformadas (limpias de ruido), se evaluó la relación final con la calidad.

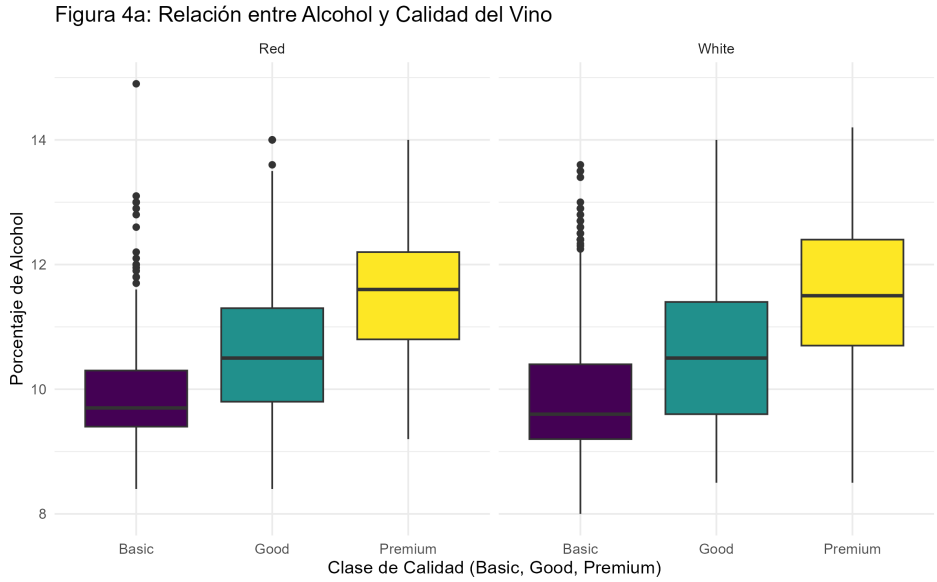

Figura 8: Relación positiva clara entre Alcohol y Calidad (Sin transformar).

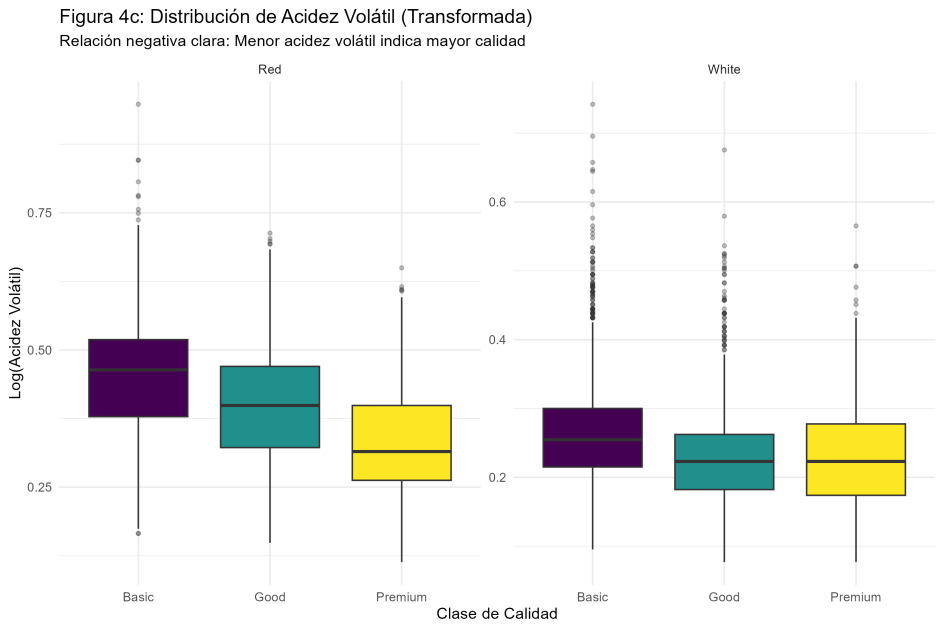

Figura 9: Relación negativa clara entre Acidez Volátil (Transformada) y Calidad.

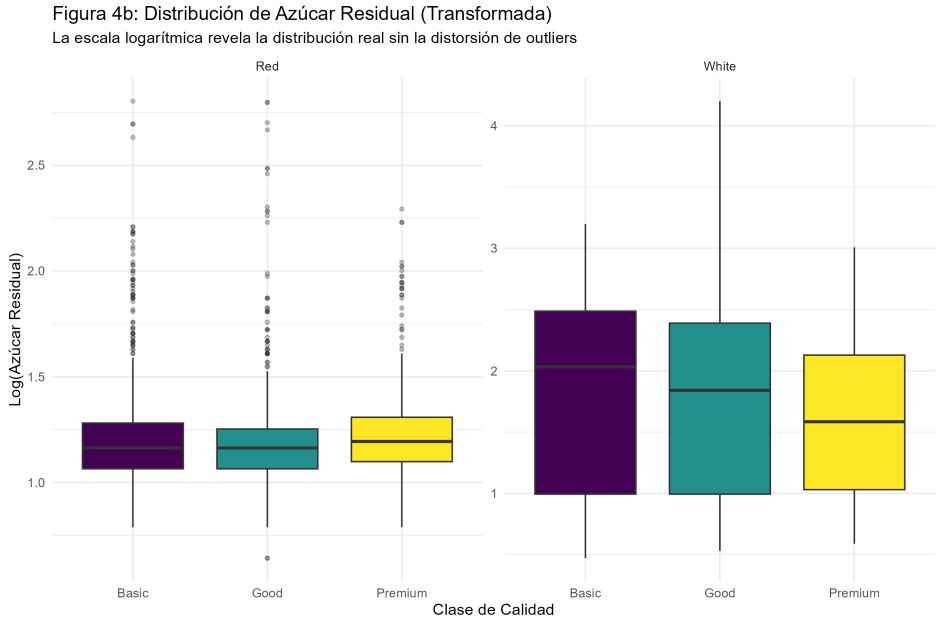

Figura 10: Distribución de Azúcar Residual (Transformada). La escala logarítmica elimina la distorsión.

4.6.1. Datos: Comparación de Medias

Primero, observamos los valores promedio absolutos para los vinos de alto alcohol, separando aquellos clasificados como `Basic` de los `Premium`.

Tabla 6: Perfil Fisicoquímico de Vinos con Alto Alcohol ($> Q3$): `Basic` vs. `Premium`.

Tipo	Clase	N	Alcohol	A. Volátil	Cloruros	Azúcar	Sulfatos	Ác. Cítrico
Tinto	Basic	54	11.8	0.601	0.070	2.98	0.600	0.219
	Premium	137	12.1	0.411	0.073	2.81	0.727	0.373
Blanco	Basic	92	12.0	0.332	0.036	3.93	0.452	0.321
	Premium	542	12.4	0.308	0.034	4.20	0.491	0.319

4.6.2. Análisis de Brechas (Diferencias Porcentuales)

Para entender qué variable es la responsable de la baja calidad, calculamos la diferencia porcentual relativa. Un valor positivo indica que el vino `Basic` tiene un exceso (posible defecto), mientras que un valor negativo indica una carencia respecto al `Premium`.

Tabla 7: Diferencia Porcentual: ¿Cuánto más (o menos) tienen los vinos `Basic` respecto a los `Premium`?

Variable	Diferencia Tinto (%)	Diferencia Blanco (%)
Acidez Volátil	**+46.2 %**	+7.8 %
Ácido Cítrico	**-41.3 %**	+0.6 %
Sulfatos	-17.5 %	-7.9 %
Cloruros	-5.0 %	+8.0 %
Azúcar Residual	+6.0 %	-6.4 %
Alcohol	-2.5 %	-3.2 %

Interpretación: La Tabla 7 muestra claramente que en los vinos tintos, el problema es drástico y estructural: tienen un **46 % más de acidez volátil** (defecto) y un **41 % menos de ácido cítrico** (frescura). En cambio, en los vinos blancos, no hay un único culpable masivo; las diferencias son menores al 10 %, sugiriendo un problema de equilibrio general más sutil.

5. Resultados del Modelado Predictivo

El objetivo de esta fase fue entrenar y evaluar modelos de clasificación capaces de predecir la categoría de calidad (`Basic`, `Good`, `Premium`) basándose en las características fisicoquímicas procesadas.

5.1. Estrategia de Modelado

Se implementó un enfoque iterativo, desarrollando y comparando múltiples versiones del modelo para validar el impacto de la ingeniería de características.

5.1.1. Algoritmo Seleccionado

Se seleccionó el algoritmo **Random Forest** debido a su capacidad para:

- Manejar relaciones no lineales y complejas entre variables (como la interacción alcohol-acidez).
- Ser robusto frente a valores atípicos (outliers) restantes.
- Proveer medidas de importancia de variables para la interpretación.

5.1.2. Configuración del Entrenamiento

Todos los modelos se entrenaron bajo las mismas condiciones rigurosas para garantizar la comparabilidad:

- **Validación Cruzada:** Se utilizó *10-fold Cross Validation* para estimar el error de generalización.
- **Balanceo de Clases:** Dado el desequilibrio detectado (Premium = 19.7 %), se aplicó la técnica de *Upsampling* dentro del proceso de re-muestreo para evitar sesgos hacia la clase mayoritaria.
- **Datos:** Se utilizó una partición estratificada del 80 % para entrenamiento (N=5,200) y 20 % para prueba (N=1,297).

5.2. Evolución del Desempeño: Modelo V2 vs. Modelo V3

Se compararon dos versiones del modelo para cuantificar el valor de la ingeniería de características avanzada.

- **Modelo V2 (Base Transformada):** Utilizó las variables fisicoquímicas originales y sus transformaciones logarítmicas para mitigar outliers.
- **Modelo V3 (Ingeniería Avanzada):** Incorporó variables sintéticas diseñadas específicamente para capturar equilibrios químicos, como el *Índice de Calidad Global* y el *Ratio Potencia-Defecto*.

La Tabla 8 muestra que la incorporación de variables de conocimiento de dominio (Modelo V3) mejoró todas las métricas clave.

Tabla 8: Comparativa de Rendimiento: Modelo V2 vs. Modelo V3.

Métrica	Modelo V2	Modelo V3	Mejora
Accuracy Global	72.73 %	**74.04 %**	+1.31 %
Kappa	0.569	**0.590**	+0.021
Sensibilidad (Premium)	67.1 %	**70.2 %**	+3.1 %
F1-Score (Promedio)	0.725	**0.740**	+0.015

Análisis: La mejora más significativa (+3.1 %) se observó en la **Sensibilidad de la clase Premium**. Esto confirma que las variables de interacción (como el ratio Alcohol/Acidez) son cruciales para distinguir los vinos excelentes de los meramente buenos.

5.3. Evaluación Detallada del Modelo Final (V3)

A continuación, se presenta el análisis exhaustivo del modelo ganador (V3).

5.3.1. Matriz de Confusión

La matriz de confusión (Tabla 9) permite visualizar los errores de clasificación específicos.

Tabla 9: Matriz de Confusión del Modelo V3 (Datos de Prueba, N=1,297).

Predicción	Real: Basic	Real: Good	Real: Premium	Total Pred.
Basic	**372**	102	2	476
Good	100	**410**	74	584
Premium	4	55	**179**	238
Total Real	476	567	255	**1297**

5.4. Auditoría de Desempeño por Tipo de Vino

Dado que los vinos tintos y blancos presentan características químicas distintas, es crucial evaluar si el modelo funciona igual de bien para ambos. La Tabla 10 desglosa la confusión por variante.

Tabla 10: Matrices de Confusión Desagregadas: Tinto vs. Blanco.

Vino Tinto (Accuracy: 72.6 %)

Pred / Real	Basic	Good	Prem
Basic	**117**	32	0
Good	24	**85**	18
Premium	1	13	**31**

Vino Blanco (Accuracy: 74.5 %)

Pred / Real	Basic	Good	Prem
Basic	**255**	70	2
Good	76	**325**	56
Premium	3	42	**148**

5.4.1. Análisis de la Auditoría

Los resultados revelan una diferencia importante en la capacidad del modelo para detectar la excelencia:

- **Vinos Tintos:** El modelo es muy conservador. Detecta muy bien los vinos malos (117 aciertos), pero le cuesta identificar los `Premium` (solo 31 aciertos de un total real de 49), logrando una sensibilidad del 63.3 %.

- **Vinos Blancos:** El modelo es más efectivo en la gama alta, logrando identificar correctamente 148 vinos `Premium` (Sensibilidad: 71.8 %).

Interpretación:

- **Errores Graves:** El modelo cometió muy pocos errores graves. Solo 4 vinos `Basic` fueron clasificados erróneamente como `Premium`, y solo 2 `Premium` como `Basic`. Esto indica una alta robustez.
- **Zona de Confusión:** La mayor dificultad reside en distinguir la clase intermedia (`Good`) de sus vecinas. 100 vinos `Basic` fueron sobreestimados como `Good`, y 74 vinos `Premium` fueron subestimados como `Good`.

5.4.2. Métricas de Clasificación por Clase (Detallado)

Dado el desbalance de clases y la diferencia estructural entre tipos de vino, es vital analizar las métricas desagregadas. Las siguientes tablas presentan la Sensibilidad, Especificidad, Precisión, F1-Score y Exactitud Balanceada para el modelo global y por cada variante.

Tabla 11: Desglose de Métricas por Clase - Modelo Global (V3).

Clase	Sensibilidad	Especificidad	Precisión	F1-Score	Balanced Acc.
Basic	0.782	0.874	0.782	0.782	0.828
Good	0.723	0.762	0.702	0.712	0.743
Premium	**0.702**	**0.943**	**0.752**	**0.726**	**0.823**

Tabla 12: Desglose de Métricas por Clase - **Vino Tinto**.

Clase	Sensibilidad	Especificidad	Precisión	F1-Score	Balanced Acc.
Basic	0.824	0.821	0.785	0.804	0.823
Good	0.654	0.780	0.669	0.661	0.717
Premium	**0.633**	**0.949**	**0.689**	**0.660**	**0.791**

Tabla 13: Desglose de Métricas por Clase - **Vino Blanco**.

Clase	Sensibilidad	Especificidad	Precisión	F1-Score	Balanced Acc.
Basic	0.764	0.888	0.780	0.772	0.826
Good	0.744	0.756	0.711	0.727	0.750
Premium	**0.718**	**0.942**	**0.767**	**0.742**	**0.830**

Análisis Comparativo:

- **Premium:** El modelo tiene un rendimiento notablemente superior en vinos blancos (F1-Score 0.742) comparado con los tintos (F1-Score 0.660). Esto confirma que la .ex celencia.en tintos es más difícil de predecir con las variables actuales.
- **Basic:** En contraste, la detección de vinos tintos de baja calidad es muy robusta (Sensibilidad 0.824), impulsada por defectos químicos claros como la acidez volátil.

5.5. Auditoría de Desempeño: Tinto vs. Blanco

Se realizó una auditoría para verificar si el modelo favorecía a alguna variante de vino.

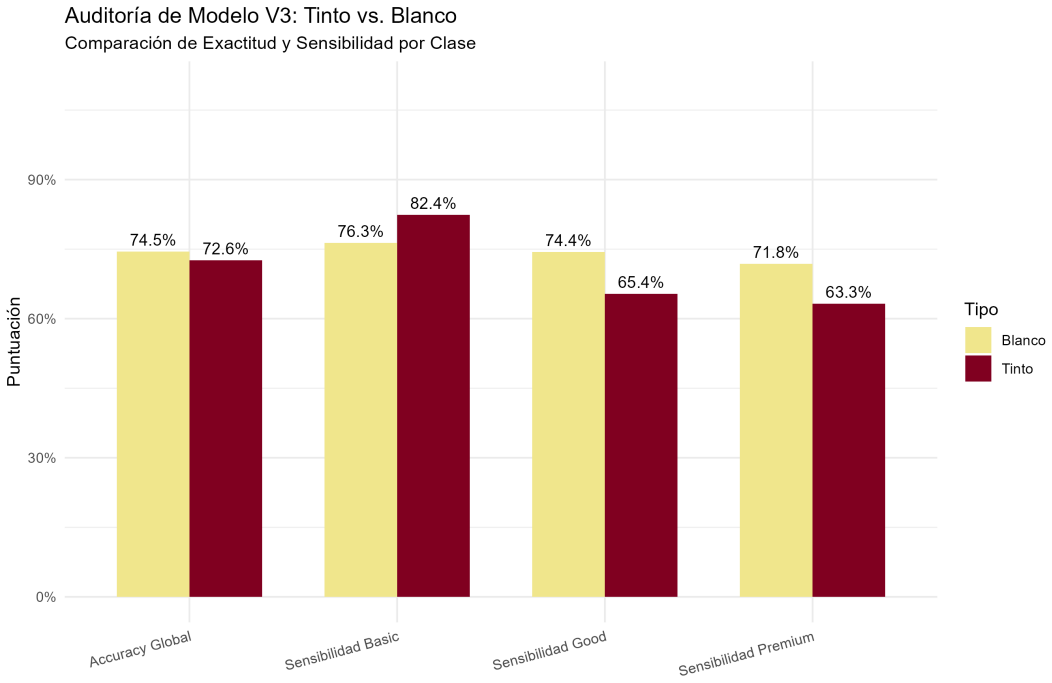

Figura 11: Auditoría de Rendimiento: Comparación de métricas entre Tinto y Blanco.

La Figura 11 revela una asimetría interesante:

- **Vinos Tintos:** El modelo es excelente detectando vinos malos (Sensibilidad Basic >80 %), probablemente debido a defectos químicos claros como la acidez volátil. Sin embargo, tiene mayor dificultad detectando la excelencia (Sensibilidad Premium $\approx$ 63 %).
- **Vinos Blancos:** El modelo muestra un desempeño superior y más equilibrado en la detección de calidad alta (Sensibilidad Premium >71 %).

5.6. Importancia de Variables

El análisis de importancia del modelo (Figura 12) confirma la validez de la ingeniería de características.

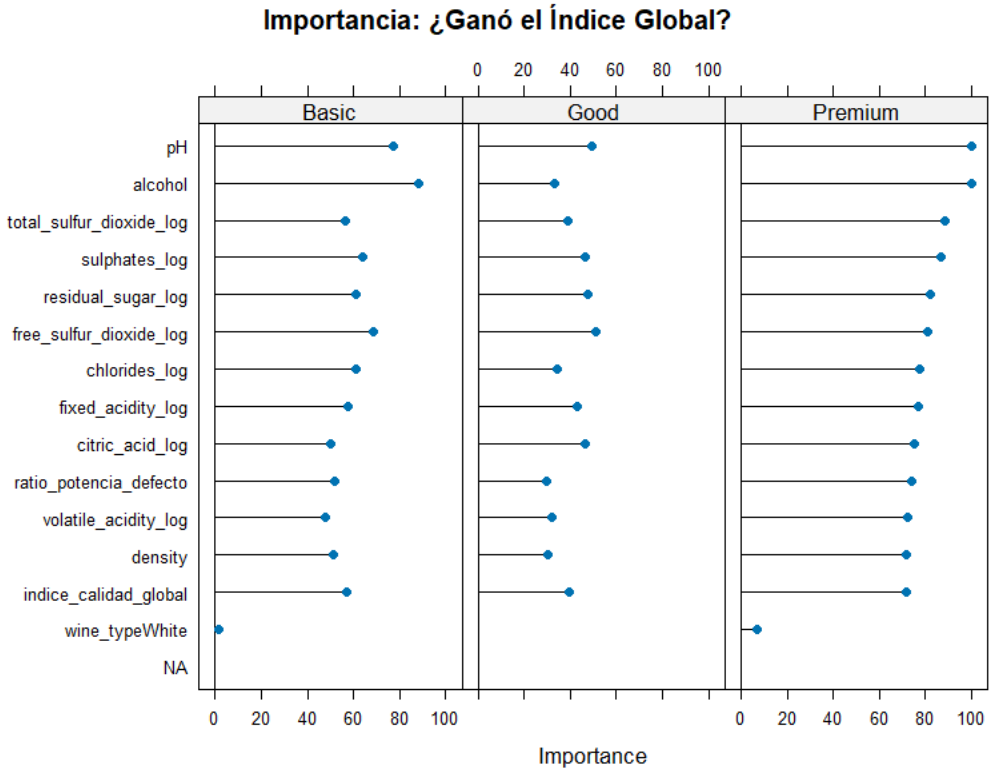

Figura 12: Ranking de Importancia de Variables (Random Forest).

Las variables `ratio_potencia_defecto` e `indice_calidad_global` se posicionaron en el top de importancia, demostrando que las interacciones químicas (ej. Alcohol vs Acidez) son predictores más potentes que las variables aisladas.

5.7. Análisis de Curvas ROC y Separabilidad

Finalmente, se evaluó la capacidad de separación del modelo mediante curvas ROC multiclase utilizando la estrategia *One-vs-All* (Uno contra Todos).

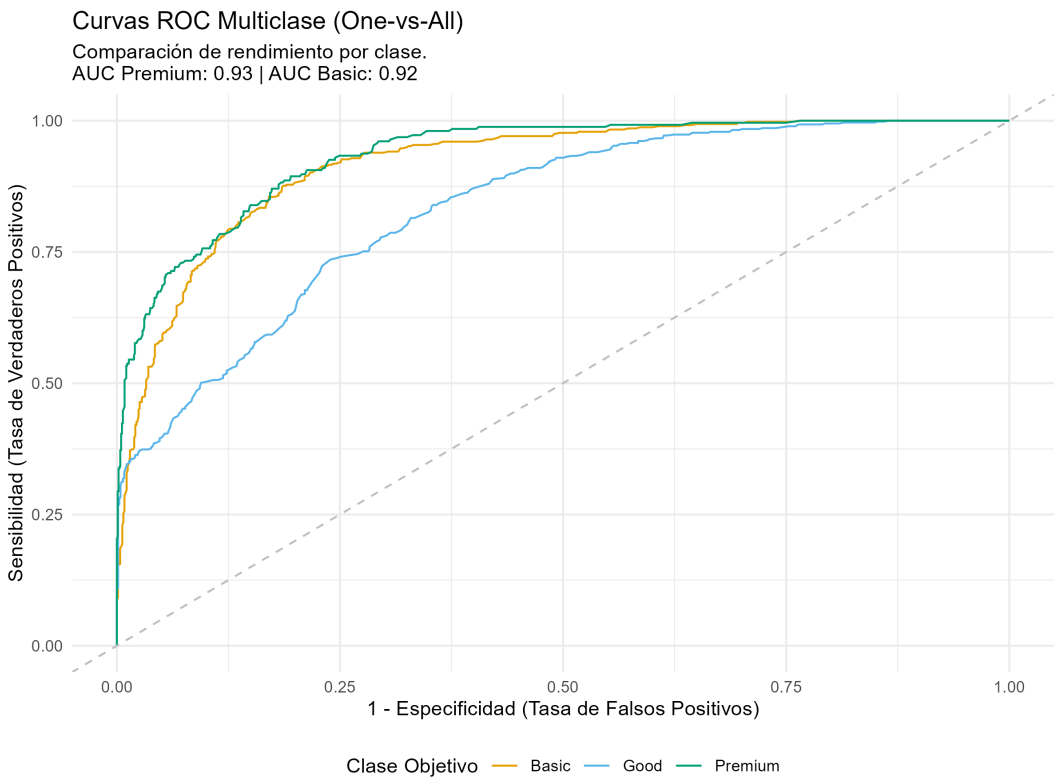

Figura 13: Curvas ROC para las tres clases. El área bajo la curva (AUC) es superior a 0.8 para todas las clases, indicando una buena separabilidad.

El análisis del Área Bajo la Curva (AUC) en la Figura 13 confirma la robustez del modelo en los extremos:

- **Premium (AUC $\approx$ 0.92):** La curva verde se acerca significativamente a la esquina superior izquierda, lo que indica una excelente capacidad de discriminación. El modelo rara vez confunde un vino verdaderamente excelente con uno de menor categoría.
- **Basic (AUC $\approx$ 0.91):** La curva naranja muestra una capacidad muy alta para detectar vinos defectuosos. Esto valida la hipótesis de que los defectos químicos (como la acidez volátil) son señales claras para el algoritmo.
- **Good (AUC $\approx$ 0.81):** La curva azul es la más cercana a la diagonal, reflejando una menor separabilidad. Esto es consistente con la matriz de confusión: la clase intermedia `Good` actúa como una "zona gris"donde se solapan las características de vinos ligeramente defectuosos y vinos casi excelentes.

5.8. Resumen Ejecutivo: Interpretación para la Toma de Decisiones

Para facilitar la adopción del modelo en un entorno productivo, se traducen las métricas técnicas a términos de confianza y riesgo operativo.

- **Alta Confianza en la Detección de Defectos:** Si el modelo clasifica un vino como **Basic** (Malo), tiene una probabilidad del **78 %** de estar en lo correcto. *Implicación:* El sistema es muy eficiente como "filtro de calidad inicial". Puede descartar automáticamente lotes defectuosos con bajo riesgo de error, ahorrando tiempo a los enólogos.
- **Alta Confianza en la Excelencia:** Si el modelo clasifica un vino como **Premium** (Excelente), tiene una probabilidad del **75 %** de estar en lo correcto. Además, su especificidad del 94 % indica que **rara vez se equivoca** etiquetando un vino mediocre como excelente. *Implicación:* El modelo es una herramienta segura para pre-seleccionar candidatos a premios o gamas altas. El riesgo de "vender gato por liebre"(etiquetar un vino malo como Premium) es extremadamente bajo.
- **Zona de Incertidumbre (Clase Media):** La mayor debilidad del modelo está en la categoría **Good** (Media). Si el modelo dice que un vino es "Bueno", existe una probabilidad considerable de que en realidad sea "Básico.º "Premium". *Implicación:* Los vinos clasificados en esta categoría intermedia deberían pasar siempre por una segunda revisión humana, ya que la frontera química entre un vino "normalz uno "bueno.es difusa para el algoritmo.

Conclusión Operativa: El modelo funciona excepcionalmente bien en los extremos (detectar lo muy malo y lo muy bueno), actuando como un sistema de triaje eficiente que permite a los expertos humanos concentrar su atención en los casos intermedios más difíciles.

Referencias

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. "Modeling wine preferences by data mining from physicochemical properties."*Decision Support Systems*, 47(4):547-553, 2009.