UNIVERSIDAD DE LA REPÚBLICA URUGUAY

FACULTAD DE INGENIERÍA UDELAR

INTRODUCCIÓN A LA CIENCIA DE DATOS

Informe Tarea Final

Predicción de la calidad de vino tinto portugués a partir de sus propiedades fisicoquímicas

Diagnóstico de calidad de los datos, procesos de limpieza, y etapas de entrenamiento y validación de modelos de clasificación.

Ing. Isabella Buschiazzo

Ing. Lucía Gutierrez

Junio 2025

Índice

1. Introducción

En un contexto donde los datos son cada vez más accesibles, su verdadero valor depende de la capacidad para interpretarlos y comunicarlos de forma clara y rigurosa. Este informe se basa en el análisis de un conjunto de datos específico, con el propósito de responder preguntas relevantes a partir de técnicas de exploración, modelado y visualización. A través de este proceso, se busca no solo extraer información útil, sino también presentar hallazgos comprensibles que permitan una toma de decisiones informada.

En particular, se trabajará con un conjunto de datos que recoge las propiedades fisicoquímicas de vinos tintos producidos en Portugal. Este conjunto incluye 1599 muestras, cada una representada por una fila, mientras que las columnas corresponden a variables como acidez fija, acidez volátil, ácido cítrico, azúcar residual, cloruros, dióxido de azufre, densidad, pH, sulfatos y contenido de alcohol. La última columna asigna a cada muestra un valor numérico que representa su calidad sensorial (de 0 a 10), la cual se encuentra estrechamente vinculada a las propiedades fisicoquímicas previamente mencionadas.

El objetivo principal de este trabajo es desarrollar un modelo que permita predecir si un vino pertenece a la categoría de alta o baja calidad, en función de sus características fisicoquímicas. A tales efectos, se asume que un vino es de alta calidad cuando presenta una calificación superior a 7. Además, aunque este trabajo no se centra en dicho aspecto, es posible identificar las principales correlaciones entre las propiedades del vino y su calidad, lo que podría servir de base para ajustar el proceso de producción con el fin de mejorarla.

2. Desarrollo

2.1. Calidad de datos

Como primer paso del análisis, se evaluó la calidad del conjunto de datos. No se detectaron valores faltantes, lo cual constituye una ventaja inicial. No obstante, se identificaron algunas limitaciones importantes. En primer lugar, los datos presentan un desbalance significativo: la mayoría de las observaciones se concentran en las clases de calidad 5 y 6, lo que puede afectar negativamente el desempeño de los modelos de clasificación sesgandolos hacía las clases mayoritarias. En segundo lugar, mediante el análisis de distribución de cada variable, se detectaron ciertos valores atípicos (*outliers*), es decir, observaciones que se alejan considerablemente del rango esperado. [1, 2, 3, 4] Para mitigar este problema, se aplicó el método del rango intercuartílico (RIQ) con el fin de eliminarlos. [5, 6]

Adicionalmente, se observaron correlaciones elevadas, tanto positivas como negativas, entre determinadas variables, como la acidez fija, acidez volátil y pH, tal como se muestra en la Figura 1. Esta situación puede generar problemas de colinealidad, especialmente al utilizar clasificadores lineales, los cuales suponen independencia entre las variables predictoras. Una alta colinealidad puede derivar en estimaciones inestables de los coeficientes, dificultar la interpretación del modelo e incrementar su varianza y sensibilidad al ruido.

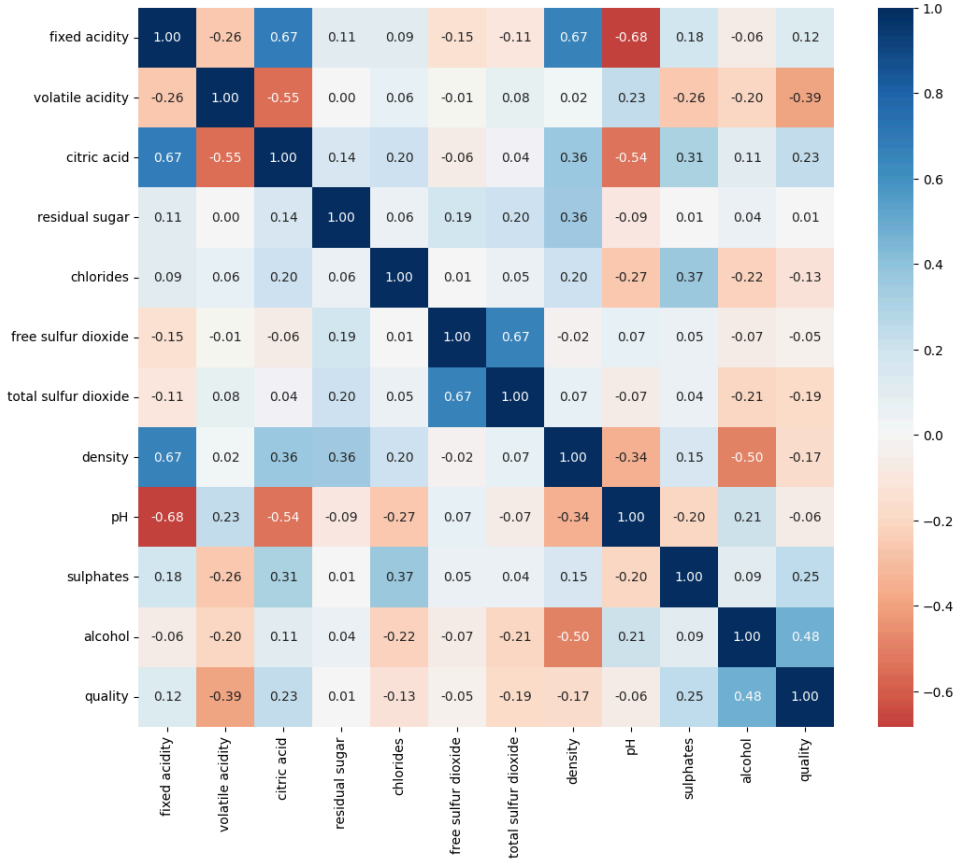

Figura 1: Matriz de correlaciones entre las variables para diferentes muestras de vino tinto.

Entre las estrategias posibles para abordar esta dificultad se encuentran la eliminación o combinación de variables correlacionadas. En este caso, se realizó el análisis de componentes principales (PCA), técnica que permite reducir la dimensionalidad del conjunto de datos conservando la mayor parte de la información relevante.

2.2. Procesamiento

Con el objetivo de desarrollar un modelo que permita predecir si un vino será de alta o baja calidad a partir de sus propiedades fisicoquímicas, se llevaron a cabo diversas etapas. En primer lugar, se generó una variable binaria denominada *alta_calidad*, que toma el valor 1 si la calificación del vino es mayor o igual a 7, y 0 en caso contrario. Esta variable se incorporó como una columna adicional en el conjunto de datos.

2.2.1. Muestreo y escalado

A continuación, se realiza un muestreo estratificado utilizando la función *train_test_split*, estableciendo que el 30 % de los datos conformen el conjunto de prueba y el 70 % restante, el conjunto de entrenamiento. Si bien las clases se encuentran balanceadas entre los subconjuntos (original, entrenamiento y prueba), se observa una marcada desproporción entre clases dentro de cada subconjunto; ya que la categoría de vinos de alta calidad representa únicamente el 13,5 % del total.

Este desbalance puede llevar a que muchos modelos tiendan a predecir sistemáticamente la clase mayoritaria, afectando negativamente la capacidad de detección de la clase minoritaria.

Para mitigar este problema, se recurre a la técnica Synthetic Minority Over-sampling Technique (*SMOTE*), aplicada exclusivamente sobre el conjunto de entrenamiento. Esta técnica genera ejemplos sintéticos de la clase minoritaria, aumentando su representación. El procedimiento consiste en identificar los vecinos más cercanos de cada muestra minoritaria y crear nuevos puntos de datos entre ellos, generando muestras artificiales pero coherentes.

Previamente a la aplicación de *SMOTE*, es necesario escalar los datos numéricos, ya que muchos algoritmos de aprendizaje automático, incluido *SMOTE*, son sensibles a la escala de las variables. Por ejemplo, si una variable como *residual sugar* tiene valores entre 0 y 65, y otra como *pH* varía entre 2,5 y 4 el modelo podría asignar mayor peso a la primera, aún cuando no sea más relevante desde el punto de vista predictivo. Para evitar esta distorsión, se emplea la función *StandardScaler*, que transforma cada variable para que tenga media cero y desviación estándar igual a uno, asegurando que todas las variables contribuyan equitativamente al cálculo de distancias.

2.2.2. Análisis de componentes principales

Una vez escalados los datos y aplicada la técnica *SMOTE* para balancear las clases, se llevó a cabo un Análisis de Componentes Principales (*PCA*) con el propósito de reducir la dimensionalidad del conjunto de datos, conservando la mayor cantidad posible de información y eliminando la multicolinealidad entre variables. Como primer paso, se representaron las observaciones en el plano definido por las dos primeras componentes principales, con el objetivo de visualizar la separación entre clases, que, como se observa en la Figura 2, no resultó satisfactoria.

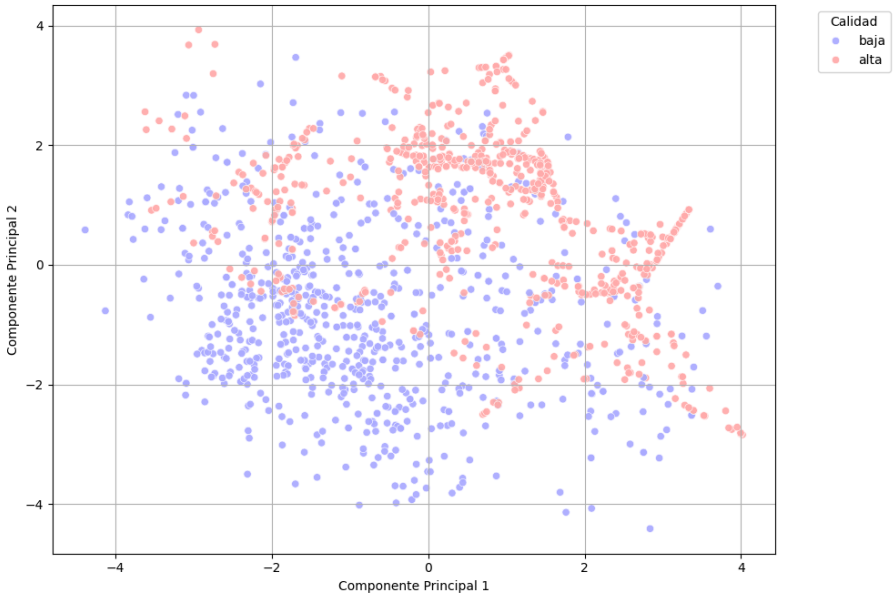

Figura 2: Proyección del PCA en dos componentes principales de vinos tintos según sus propiedades fisicoquímicas.

A su vez, para determinar el número óptimo de componentes a considerar, se analizó el número mínimo de componentes necesarios para alcanzar un umbral de varianza explicada acumulada. En este trabajo, se estableció un umbral del 90 %, por lo que el número óptimo de componentes resultante fue siete.

2.2.3. Modelos

Los modelos lineales, como la regresión logística (*Logistic Regression*) o el clasificador lineal de vectores de soporte (*LinearSVC*), suponen una relación lineal entre las variables independientes y la variable objetivo. Sin embargo, esta suposición podría no reflejar de manera adecuada la relación real entre la calidad del vino y sus propiedades fisicoquímicas. A pesar de ello, se evaluó el desempeño de ambos modelos tras aplicar el PCA.

El modelo de regresión logística (*Logistic Regression*) es un clasificador lineal que estima la probabilidad de pertenencia a una clase en función de una combinación lineal de las variables de entrada, a través de una función logística (sigmoidea). Esta característica lo hace interpretable y computacionalmente eficiente. En problemas multiclase, implementa estrategias como one-vs-rest, entrenando un clasificador binario para cada clase. Mediante este modelo, se analizó la repercusión de la técnica *SMOTE*, comparando el resultado aplicándola y sin aplicarla.

Adicionalmente, en los casos en que no se aplicó *SMOTE*, se realizó el entrenamiento evaluando los siguientes hiperparámetros:

- **C**: controla la fuerza de la regularización; valores más bajos implican una regularización más fuerte, penalizando coeficientes grandes para evitar el sobreajuste.
- **penalty**: define el tipo de regularización. Se consideraron penalizaciones L1, que pueden llevar ciertos coeficientes exactamente a cero, favoreciendo modelos más simples, y L2, que distribuye la penalización de manera más uniforme.
- **solver**: especifica el algoritmo de optimización utilizado para ajustar el modelo; su elección depende del tipo de penalización y del tamaño del conjunto de datos.

Lo mismo se realizó estudiando el modelo LinearSVC. Este modelo pertenece a la familia de los Support Vector Machines (SVM) y busca encontrar un hiperplano que separe las clases de manera lineal maximizando la distancia entre los puntos más cercanos de cada clase. Esta característica lo hace eficaz en problemas de alta dimensionalidad. En este caso, a la hora de analizar los hiperparametros se evaluó el modelo modificando el parámetro “C”.

Finalmente, considerando la complejidad del comportamiento de los datos, se evaluó el modelo Random Forest, el cual construye múltiples árboles de decisión y combina sus resultados para obtener una predicción más precisa y robusta. Cada árbol se construye dividiendo los datos en función de sus características, seleccionando en cada nodo la condición que mejor separa las clases, con el objetivo de maximizar la pureza de las ramas resultantes.

Este procedimiento se repite hasta que no es posible realizar más divisiones útiles o se alcanza una profundidad máxima predefinida. Como el modelo no es lineal, se aplicó sin la ejecución previa del PCA.

Con el objetivo de lograr una predicción más ajustada, se manipuló el umbral de clasificación. Este define la probabilidad a partir de la cual una observación se clasifica como perteneciente a una clase determinada. Al aumentar dicho umbral, se exige una mayor probabilidad para considerar un vino como de alta calidad, lo que incrementa la precisión del modelo y reduce la cantidad de falsos positivos (vinos de baja calidad clasificados erróneamente como de alta calidad). Sin embargo, es importante señalar que la elección del umbral depende del objetivo definido. Desde la perspectiva del consumidor, un umbral alto garantiza seleccionar vinos realmente destacados. En cambio, desde el punto de vista de una bodega interesada en identificar un conjunto preliminar amplio de vinos con potencial para una cata posterior, un umbral más bajo puede resultar más conveniente.

Más allá del umbral, existen otros hiperparámetros que pueden ajustarse durante el entrenamiento del modelo para mejorar su desempeño según las características del problema.

3. Consideraciones éticas

En relación con las consideraciones éticas y el manejo de datos, se analizaron aspectos vinculados a la privacidad, los posibles sesgos y limitaciones del modelo, así como los riesgos asociados y las medidas previstas para su mitigación.

En lo que respecta a la privacidad, si bien los datos utilizados no son de carácter personal, presentan sensibilidad desde el punto de vista industrial, ya que podrían revelar información confidencial sobre procesos internos de producción. Un error eventual en la elaboración podría dar lugar a vinos de calidad significativamente baja que, de hacerse públicos, podrían ocasionar problemas graves para la empresa, e incluso comprometer su continuidad. Por este motivo, resulta fundamental manejar con cautela la información divulgada.

Por otra parte, se identificaron posibles sesgos y limitaciones del modelo que pueden afectar su confiabilidad. La simulación empleada para entrenar el modelo no necesariamente contempla todos los escenarios reales, lo que podría conducir a sobreajuste y a un rendimiento deficiente frente a situaciones no previstas. Además, el desequilibrio entre clases puede generar errores de clasificación con consecuencias importantes. Los falsos negativos (vinos de alta calidad clasificados como de baja) podrían ocasionar pérdidas económicas y decisiones equivocadas en términos de producción o comercialización, mientras que los falsos positivos (vinos de baja calidad clasificados como de alta) podrían dañar la reputación y la percepción de calidad del producto final. Para mitigar estos riesgos, se recomienda aplicar técnicas de balanceo de clases, en este trabajo se utilizó *SMOTE*, y calibrar cuidadosamente los umbrales de decisión. Asimismo, es esencial validar el modelo con datos reales, evaluar su desempeño en escenarios adversos y establecer un monitoreo continuo que permita reentrenarlo con datos actualizados, garantizando su robustez y confiabilidad a largo plazo.

4. Bibliografía

[1] - **La Rioja.** (2021, 11 de junio). *Densidad de vino.* La Rioja. Recuperado el 28 de junio de 2025, de https://www.larioja.com/opinion/densidad-vino-20210611214244-nt.html

[2]- **Cientisol Soluciones Científicas.** (2023, 2 de agosto). *Guía definitiva sobre el análisis químico del vino: qué parámetros hay que medir y cómo hacerlo.* Recuperado el 28 de junio de 2025, de https://cientisol.com/guia-definitiva-sobre-el-analisis-quimico-del-vino-que-parametros-hay-que-medir-y-como-hacerlo/

[3]- **Yahya, S.** (2023, 9 de julio). *Analyzing Red Wine Quality.* Medium. Recuperado el 28 de junio de 2025, de https://medium.com/@spynyahya/analyzing-red-wine-quality-69aadb08a303

[4]- **Aprender de Vino.** (s. f.). *¿Qué es el pH del vino?* Recuperado el 28 de junio de 2025, de https://www.aprenderdevino.es/ph-y-vino/

[5]- Penn State Department of Statistics. (s. f.). *3.2 – Identifying outliers: IQR method*. *STAT 200: Elementary Statistics*.
Recuperado el 28 de junio de 2025, de https://online.stat.psu.edu/stat200/lesson/3/3.2

[6]- **Acharya, N.** (2024, febrero 13). *Understanding Outlier Removal Using Interquartile Range (IQR).* Medium. Recuperado el 28 de junio de 2025, de https://medium.com/@nirajan.acharya777/understanding-outlier-removal-using-interquartile-range-iqr-b55b9726363e