Proyecto Final, Estadística Inferencial, curso 2025.

Análisis de la calidad del Vino respecto a sus características fisicoquímicas.

Departamento de Ciencias Naturales y Exactas.

Equipo I

Avril Atahides, Alexia Aurrecochea, Milagros Cancela y Bruno Arias.

Universidad Católica del Uruguay.

Prof. Manuel Flores

2do Semestre, 2025

Análisis de la calidad del Vino respecto a sus características fisicoquímicas.

Problemática

¿Es posible identificar diferencias estadísticamente significativas entre los vinos de alta y baja calidad en función de sus propiedades físico-químicas?

En las últimas décadas, la vitivinicultura ha experimentado una transformación impulsada por la ciencia de datos y la ingeniería química. La posibilidad de modelar cuantitativamente la calidad del vino a partir de variables medibles, como la acidez, el pH o el contenido de alcohol, ha permitido mejorar la estandarización y optimizar procesos enológicos.
El contexto global refuerza la relevancia de este tipo de análisis: en 2023, el consumo mundial de vino se estimó en 221 millones de hectolitros, equivalentes a más de 29 mil millones de botellas, según la Organización Internacional de la Viña y el Vino (OIV). Este nivel de consumo refleja tanto la magnitud económica del sector como su impacto cultural a nivel planetario.

En Uruguay, nuestro país con una tradición vitivinícola consolidada, el consumo alcanzó aproximadamente 65 millones de litros en 2020, con un promedio per cápita de 22,2 litros por persona en 2021. Estas cifras ubican a Uruguay entre los países de mayor consumo relativo de América Latina, lo que subraya la importancia de estudios estadísticos que permitan comprender cómo los factores físico-químicos determinan la calidad percibida del vino nacional e importado.

El presente trabajo busca examinar, desde una perspectiva inferencial, si existen patrones estructurales y significativos en las propiedades del vino que permitan distinguir entre productos de diferente calidad. Esta pregunta tiene relevancia tanto para la industria vitivinícola (mejora de procesos y control de calidad) como para la investigación científica, al explorar la relación entre atributos físico-químicos y percepción sensorial.

Comprender cómo varía el nivel de calidad según las características fisicoquímicas del vino puede contribuir al desarrollo de modelos predictivos y sistemas de recomendación enológicos, reforzando la competitividad de los productores y la transparencia en la valoración de los vinos. Así como actúa como modelo de estudio real, para el proceso analítico a realizarse por los conformantes del equipo.

Dataset

La base de datos “Wine Quality” proviene del estudio realizado por Cortez et al. (2009), publicado en Decision Support Systems (Elsevier). Los datos fueron recolectados en 2009 a partir de organismos portugueses vinculados a la industria del vino, utilizando mediciones laboratoriales en muestras reales de vino “Vinho Verde” de las variantes tinto y blanco. El proceso de recolección fue estandarizado y contó con la evaluación sensorial de calidad por parte de expertos en cata.

El dataset, en formato CSV delimitado por punto y coma, está compuesto por 6,497 muestras independientes (1,599 de vino tinto y 4,898 de vino blanco), cada una representada por una fila. En la versión combinada de los archivos, las muestras incorporan la variable categórica “tipo” que identifica si el vino es tinto o blanco. Cada observación incluye 12 variables: 11 cuantitativas que describen características fisicoquímicas medidas en laboratorio (como alcohol, acidez volátil, sulfatos, pH, azúcar residual, entre otras), y la variable sensorial “quality” que recoge la calificación ordinal otorgada por los catadores (con valores enteros de 0 a 10). La estructura del mismo se puede visualizar en la Tabla 1.

El conjunto de datos no presenta valores faltantes, variables temporales ni agrupamientos espaciales, y cada registro corresponde a una muestra única. Si bien pueden existir duplicados, estos se consideran mediciones independientes de vinos con idénticas características.

La base resulta adecuada para el análisis estadístico inferencial, ya que permite explorar diferencias significativas en la calidad del vino en función de sus propiedades químicas, así como investigar la relación entre las características químicas y la percepción sensorial, discriminando entre niveles altos y bajos de calidad.

Variable	Tipo	Descripción	Categorías
quality	int	Calificación sensorial de la calidad.	Entero 0-10 (usual: 3-9)
type	chr	Tipo de vino.	red / white
alcohol	num	Contenido alcohólico	% vol
volatile acidity	num	Acidez volátil (principalmente ácido acético).	g/L
fixed acidity	num	Acidez fija (tartárico, málico, etc.).	g/L
citric acid	num	Ácido cítrico.	g/L
residual sugar	num	Azúcar residual.	g/L
chlorides	num	Cloruros (salinidad).	g/L
free sulfur dioxide	num	Dióxido de azufre libre.	mg/L

Variable	Tipo	Descripción	Categorías
quality	int	Calificación sensorial de la calidad.	Entero 0-10 (usual: 3-9)
total sulfur dioxide	num	Dióxido de azufre total	mg/L
density	num	Densidad del vino	g/cm³
pH	num	Acidez medida como pH.	Escala (≈2.5–4)
sulphates	num	Sulfatos (relacionados con SO_2).	g/L

Tabla 1: Descripción de Variables del Dataset Wine Quality (UCI)

Cabe mencionar que no posee dimensión temporal ni espacial: cada observación es independiente y no corresponde a una serie de tiempo ni a múltiples mediciones del mismo vino. Y el conjunto target, si bien se puede definir de forma discreta, lo podremos enfrentar como variable continua.

Hipótesis de Estudio

Para estructurar el análisis inferencial, se plantean las siguientes hipótesis estadísticas orientadas a contrastar la existencia de diferencias significativas entre vinos de alta y baja calidad:

Tipo de hipótesis	Formulación	Interpretación
Hipótesis nula (H_0)	No existen diferencias significativas en las medias de las variables fisicoquímicas (pH, alcohol, acidez, azúcares, etc.) entre los vinos de alta y baja calidad.	Las propiedades químicas no permiten discriminar estadísticamente entre categorías de calidad.
Hipótesis alternativa (H_1)	Existen diferencias significativas en las medias de al menos una de las variables físico-químicas entre vinos de alta y baja calidad.	Algunas propiedades químicas explican de manera significativa la calidad percibida y permiten su predicción.

Estas hipótesis permiten aplicar contrastes estadísticos (ANOVA, t-test, correlaciones y regresión lineal) para determinar qué factores influyen de forma significativa en la calidad, así como construir modelos predictivos basados en inferencia estadística.

Referencias

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4), 547–553. https://doi.org/10.1016/j.dss.2009.05.016

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: With applications in R (2nd ed.). Springer. https://www.statlearning.com/

Wooldridge, J. M. (2013). Introducción a la econometría: Un enfoque moderno (5ª ed.). Cengage Learning.

Organización Internacional de la Viña y el Vino (OIV). (2024). State of the World Vine and Wine Sector in 2023. Organisation Internationale de la Vigne et du Vin. https://www.oiv.int/sites/default/files/2024-04/OIV_STATE_OF_THE_WORLD_VINE_AND_WINE_SECTOR_IN_2023.pdf

Helgi Library. (2021). Wine consumption in Uruguay. Helgi Analytics. https://www.helgilibrary.com/charts/wine-consumption-total-rose-656-to-650-ml-in-uruguay-in-2020-43/

Dua, D., & Graff, C. (2019). Wine Quality Data Set [Data set]. UCI Machine Learning Repository. University of California, Irvine. https://archive.ics.uci.edu/dataset/186/wine+quality

Registro de IA

Para la elaboración del presente trabajo se utilizó la herramienta ChatGPT (modelo GPT-4o-mini, OpenAI) como apoyo en la redacción y refinamiento del planteo del problema, formulación de la pregunta de investigación, hipótesis y redacción de la relevancia aplicada.

Las ideas originales, el enfoque temático y las decisiones conceptuales sobre la problemática (concentración de empresas en horarios y zonas del transporte metropolitano de Montevideo focalizadas en el usuario) son del equipo. El uso de la herramienta se limitó a mejorar la claridad, coherencia y rigor técnico del texto, sin alterar el sentido ni los objetivos definidos por los autores.

El equipo asume plena responsabilidad por la comprensión, interpretación y validación del contenido final presentado.