COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Ciencias Económicas y Empresariales

Modelos Predictivos y Sistemas de Recomendación en la Industria del Vino: Un Enfoque Basado en Machine Learning

Autor: Sofía Ferrer Bingoel
Clave: 202015635
Director: María de las Mercedes Barrachina

MADRID - abril 2025

Palabras clave:
Machine Learning, industria del vino, sistemas de recomendación, calidad del vino, análisis sensorial, preferencias del consumidor.

Resumen

Este trabajo de fin de grado aplica técnicas de *Machine Learning* al sector vinícola con un doble objetivo: por un lado, evaluar hasta qué punto las características fisicoquímicas del vino explican su calidad, y por otro, desarrollar un sistema de recomendación personalizado en función de las preferencias del consumidor.

Para ello, se han utilizado dos bases de datos distintas. La primera, centrada en variables como el nivel de alcohol, acidez o sulfatos, permitió construir modelos predictivos de calidad, entre los que destacó el algoritmo Random Forest por su precisión y capacidad explicativa. La segunda base de datos incluía características sensoriales (dulzura, acidez, cuerpo y taninos), a partir de las cuales se diseñó un sistema de recomendación basado en el algoritmo KNN, que ofrece al usuario sugerencias personalizadas según sus gustos.

Los resultados muestran que, si bien las propiedades químicas tienen un peso importante en la calidad del vino, no son suficientes para explicarla completamente, debido al papel de factores subjetivos como la marca, la variedad o la percepción individual. Asimismo, el sistema de recomendación propuesto demuestra ser una herramienta útil para guiar a consumidores con escaso conocimiento del vino, especialmente en un contexto de mercado joven que demanda mayor personalización.

Este estudio refleja el potencial del *Business Analytics* y la inteligencia artificial para apoyar tanto a productores como a consumidores en la toma de decisiones informadas dentro de un sector tradicional en plena evolución.

Índice

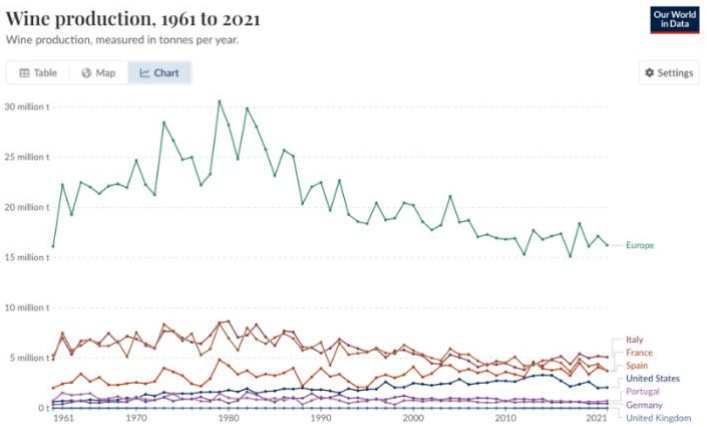
1. Introducción

1.1 Contexto y relevancia del estudio

El vino es un producto con una enorme importancia económica, social y cultural en muchas partes del mundo. En países como España, Francia e Italia, la viticultura representa no solo una fuente de ingresos clave para el sector agroalimentario, sino también un elemento distintivo de la identidad y el turismo gastronómico.

Según la Organización Internacional de la Viña y el Vino (OIV), la producción mundial de vino en 2022 fue de aproximadamente 258 millones de hectolitros, con la Unión Europea representando más del 60% de la producción global. España, en particular, es el país con la mayor superficie de viñedos a nivel mundial y el tercer mayor productor de vino, después de Italia y Francia (OIV, 2023).

Desde el punto de vista del consumidor, el vino es un producto que combina aspectos objetivos y subjetivos. Su calidad puede evaluarse a partir de parámetros fisicoquímicos como la acidez, el contenido de alcohol y la estructura de los taninos, pero también depende de percepciones subjetivas relacionadas con la marca, la región de producción y la experiencia sensorial del consumidor.



En la actualidad, la industria del vino enfrenta un mercado cada vez más competitivo, con consumidores que buscan productos adaptados a sus gustos personales. Tal como muestran los datos históricos de producción vinícola, la evolución del mercado ha estado marcada por una reducción progresiva de la producción en Europa y un crecimiento del interés por la diversificación de productos (Our World in Data, 2023). En este contexto, la capacidad de recomendar vinos en función de preferencias individuales se vuelve clave para fomentar el consumo y mejorar la experiencia del cliente.

Por otro lado, el concepto de calidad en el vino es complejo y multifactorial. No siempre los vinos con mejores características químicas son los mejor valorados, lo que abre la pregunta de si la calidad percibida está influenciada por elementos más subjetivos como la marca, el precio o el marketing.

1.2 Objetivos del estudio

La industria vinícola se encuentra en constante evolución, con consumidores cada vez más exigentes y como hemos dicho, un mercado altamente competitivo. La evaluación de la calidad del vino y la personalización de la experiencia del consumidor a través de sistemas de recomendación son áreas clave donde el Business Analytics y el Machine Learning pueden aportar un gran valor.

Este estudio tiene dos objetivos principales: analizar la relación entre las características fisicoquímicas y la calidad del vino, y desarrollar un modelo de recomendación que sugiera variedades de vino según las preferencias del consumidor.

Objetivo 1: Analizar la relación entre las características fisicoquímicas y la calidad del vino

El primer objetivo de este trabajo es evaluar hasta qué punto las propiedades químicas del vino pueden explicar su calidad. Para ello, se explorará una base de datos de vinos que incluye variables como acidez, dulzura, cuerpo, pH, contenido de alcohol, entre otras, y su relación con una calificación de calidad asignada.

Para alcanzar este objetivo, se plantea esta pregunta de investigación:

¿Las características fisicoquímicas del vino explican completamente su calidad?

Objetivo 2: Desarrollar un modelo de recomendación de vinos basado en las preferencias del usuario

El segundo objetivo es diseñar un sistema de recomendación que sugiera variedades de vino adecuadas para cada usuario, basado en sus preferencias personales. El modelo tomará en cuenta la dulzura, acidez, cuerpo, taninos y uso del vino.

Para alcanzar este objetivo, se plantea esta pregunta de investigación:

¿Es posible recomendar variedades de vino basándose en las preferencias del usuario?

1.3 Metodología del estudio

Este estudio aplica técnicas de Machine Learning y análisis de datos para abordar dos preguntas de investigación clave:

- Determinar si las características fisicoquímicas explican completamente la calidad del vino.

- Construir un sistema de recomendación que sugiera variedades de vino en función de las preferencias del consumidor.

Para lograr estos objetivos, se utilizarán dos enfoques metodológicos diferenciados, combinando análisis exploratorio de datos, modelado predictivo y técnicas de recomendación.

Fases de la metodología:

- Recopilación y preparación de datos

Se utilizarán dos bases de datos con información detallada sobre vinos:

Dataset 1: Contiene valores fisicoquímicos del vino y su calificación de calidad.

Dataset 2: Incluye información sobre variedades de vino y sus características sensoriales (dulzura, acidez, taninos, cuerpo).

Tareas para realizar:

- Eliminación de datos faltantes o erróneos.
- Normalización de variables para evitar sesgos en el modelado.
- Exploración y análisis estadístico preliminar.

- Análisis de la calidad del vino

Análisis exploratorio: Cálculo de estadísticas descriptivas para cada variable, identificación de correlaciones entre variables fisicoquímicas y calidad y visualización de tendencias mediante histogramas, gráficos de dispersión y heatmaps.

Modelos de predicción: Se entrenarán diferentes modelos de Machine Learning para predecir la calidad del vino.

- Modelo de recomendación de vinos

Se analizarán los atributos dulzura, acidez, taninos, cuerpo y uso del vino para establecer su influencia en la recomendación. Seguidamente, se calcularán distancias entre vinos basadas en similitudes de atributos (KNN), se determinará el número óptimo de vecinos (K) mediante validación cruzada y se realizarán pruebas para evaluar la precisión de las recomendaciones.

Por último, se medirá qué tan acertadas son las recomendaciones en función de las preferencias del usuario.

- Interpretación de Resultados y Conclusiones

- Reflexión sobre si las características fisicoquímicas son suficientes para explicar la calidad del vino.
- Evaluación de la efectividad del sistema de recomendación.
- Identificación de limitaciones del estudio y posibles mejoras futuras.

2. Marco teórico

2.1 El vino y cómo se clasifica

El vino es una bebida alcohólica obtenida de la fermentación del zumo de uva (Vitis vinífera), proceso en el cual los azúcares naturales de la fruta se transforman en alcohol gracias a la acción de levaduras. Esta transformación puede controlarse mediante distintos métodos para obtener una amplia variedad de estilos y sabores (Escuela Europea de Versalles, 2020).

La clasificación del vino es compleja y depende de múltiples factores, como el color, la cantidad de azúcar residual, la presencia de gas carbónico y el proceso de crianza.

Clasificación según el color

El color del vino es una de las formas más básicas de clasificación y depende del tipo de uva utilizada y del tiempo de contacto con los hollejos durante la fermentación. Existen tres categorías principales:

- **Vino tinto**: Elaborado con uvas tintas y sus hollejos, lo que le da su color rojo característico.
- **Vino blanco**: Se produce sin hollejos, utilizando solo el mosto de uvas blancas o tintas de pulpa clara.
- **Vino rosado**: Se obtiene de uvas tintas con un contacto corto con los hollejos para lograr una coloración ligera (Bodega Cortijo Moros Santos, 2023).

Clasificación según el contenido de azúcar

El contenido de azúcar residual en el vino tras la fermentación permite clasificarlo en distintas categorías:

- **Seco**: Menos de 5 g/L de azúcar.
- **Semiseco**: Entre 5 y 15 g/L de azúcar.
- **Semidulce**: Entre 15 y 50 g/L de azúcar.
- **Dulce**: Más de 50 g/L de azúcar (Bodega Cortijo Moros Santos, 2023).

Clasificación según la presencia de gas carbónico

Algunos vinos contienen gas carbónico natural o añadido, lo que influye en su textura y percepción en boca:

- **Vinos tranquilos**: No contienen gas carbónico perceptible.
- **Vinos de aguja**: Presentan una ligera efervescencia debido al gas natural.
- **Vinos espumosos**: Contienen burbujas generadas por una segunda fermentación en botella o en tanque, como el Champagne o el Cava (Escuela Europea de Versalles, 2020).

Clasificación según el envejecimiento

El tiempo que un vino pasa en barrica o botella modifica sus propiedades organolépticas y determina su categoría dentro de la crianza:

- **Vino joven**: Embotellado poco después de su fermentación, sin envejecimiento en barrica.
- **Vino de crianza**: Envejecido al menos 24 meses, con 6 meses en barrica para los tintos y 18 meses en total para los blancos y rosados.
- **Vino reserva**: Envejecido un mínimo de 36 meses, con 12 meses en barrica para tintos y 24 meses en total para blancos y rosados.
- **Vino gran reserva**: Envejecido 60 meses, con 18 meses en barrica para tintos y 48 meses en total para blancos y rosados (Escuela Europea de Versalles, 2020).

2.2 Características sensoriales

El vino es una bebida que ofrece una experiencia sensorial única, ya que involucra la vista, el olfato, el gusto y hasta el tacto y el oído en su apreciación. Las características sensoriales del vino permiten diferenciarlo según su estructura, estilo y calidad, siendo esenciales para su evaluación tanto por consumidores como por expertos enólogos y sumilleres.

Las principales propiedades organolépticas del vino incluyen su color, aroma, sabor, cuerpo, dulzura, acidez, taninos y graduación alcohólica.

Color y apariencia

El color del vino es el primer indicador de su naturaleza y estado. Está influenciado por diversos factores, como la variedad de la uva, la maduración, el proceso de vinificación y la edad del vino.

- Vinos tintos: Presentan una paleta de colores que va desde tonos púrpura en vinos jóvenes hasta tonos teja en vinos envejecidos.
- Vinos blancos: Pueden oscilar entre colores amarillo pálido y dorado, dependiendo de su tiempo de crianza.
- Vinos rosados: Su tonalidad varía desde rosa claro hasta tonos más intensos según la maceración con los hollejos de la uva.

El color del vino también puede revelar información sobre su evolución y almacenamiento. En términos generales, los vinos jóvenes presentan colores más

vibrantes, mientras que los vinos envejecidos pueden adquirir tonalidades más atenuadas debido a la oxidación (Bodegas Viñedos Amaró, 2022).

Aroma y bouquet

El olfato juega un papel crucial en la percepción del vino, ya que más del 80% de los sabores se perciben a través del sentido del olfato. En enología, se distingue entre:

- Aroma primario: Proviene directamente de la variedad de uva y puede incluir notas frutales, florales o herbáceas.
- Aroma secundario: Se desarrolla durante la fermentación y puede aportar matices de levadura, lácteos o especias.
- Bouquet: Es el conjunto de aromas terciarios adquiridos durante la crianza del vino en barrica o botella, que pueden recordar a vainilla, madera, tabaco o frutos secos (Bodegas Viñedos Amaró, 2022).

La complejidad aromática de un vino depende tanto de su elaboración como del tiempo de envejecimiento, siendo una característica clave en su evaluación sensorial.

Sabor y percepción en boca

El sabor del vino es el resultado de la combinación de percepciones del gusto y el olfato. Existen cinco características básicas que definen la estructura de un vino en boca (Carrera, 2019):

- Dulzor:

Determinado por la cantidad de azúcar residual presente en el vino. Los vinos secos tienen niveles mínimos de azúcar, mientras que los vinos dulces pueden contener más de 50 g/L. También puede percibirse un dulzor indirecto cuando el vino tiene una concentración elevada de alcohol o glicerina.

- Acidez:

Es la responsable de la frescura y equilibrio en un vino. Un vino con una acidez alta se percibe más vibrante y crujiente, mientras que una baja acidez da una sensación más plana. Es esencial en vinos blancos y espumosos, ya que potencia su frescura.

- Taninos:

Son compuestos fenólicos presentes en la piel, semillas y raspón de la uva. Aportan estructura y astringencia al vino, influyendo en su longevidad. Se encuentran en mayor concentración en vinos tintos y pueden suavizarse con la crianza.

- Cuerpo:

Se refiere a la densidad y peso del vino en boca. Depende de la concentración de alcohol, glicerina y extractos sólidos. Se clasifica en vinos de cuerpo ligero, medio o completo, siendo los vinos tintos envejecidos los que suelen tener mayor cuerpo.

- Alcohol:

Su presencia se percibe en la sensación de calidez en boca. Los vinos con una alta graduación alcohólica pueden percibirse más untuosos y redondos, mientras que los vinos de baja graduación pueden resultar más ligeros y refrescantes.

Estas cinco características permiten a los expertos clasificar los vinos y facilitar su elección a los consumidores, ya que cada combinación de dulzura, acidez, taninos, cuerpo y alcohol da lugar a perfiles sensoriales únicos (Carrera, 2019).

El estudio de estas propiedades resulta esencial para comprender la calidad del vino y desarrollar sistemas de recomendación personalizados, como el que se plantea en este estudio. Mediante el análisis de datos, se puede establecer cómo estas características influyen en la percepción del consumidor y en la predicción de la calidad del vino.

2.3 Factores que afectan el sabor

El sabor del vino es el resultado de una interacción compleja entre múltiples factores que van desde la composición química de la uva hasta el proceso de vinificación y las condiciones de envejecimiento. Estas variables determinan el perfil sensorial del vino, afectando su dulzura, acidez, cuerpo, taninos y aromas.

Entre los factores más influyentes en el sabor del vino se encuentran la variedad de uva, el clima y la región de cultivo, la añada, el proceso de vinificación y crianza, y la composición química del vino (Grupo Pago de Mar, 2021; Delgado, 2021).

La variedad de uva

Cada tipo de uva posee características químicas únicas que determinan el sabor del vino. Existen más de 600 variedades de uva utilizadas en la producción vinícola, aunque solo unas pocas son ampliamente cultivadas (Grupo Pago de Mar, 2021).

- Vinos mono-varietales: Se elaboran con una única variedad de uva, permitiendo resaltar su perfil de sabor distintivo.
- Vinos de coupage o ensamblaje: Mezclan varias variedades de uva para aportar mayor complejidad y equilibrio.

La edad del viñedo también es un factor relevante: las cepas más viejas producen menos uvas, pero de mayor concentración y calidad, intensificando el sabor del vino (Grupo Pago de Mar, 2021).

Clima y región de cultivo (Terroir)

El terroir hace referencia al conjunto de condiciones ambientales donde crece la vid, incluyendo el clima, la altitud, el tipo de suelo y la orientación del viñedo.

Efectos del clima:

- Climas fríos: Favorecen vinos con mayor acidez, menor graduación alcohólica y aromas más frescos (cítricos, florales).
- Climas cálidos: Generan vinos con mayor contenido de azúcar y alcohol, menor acidez y notas más maduras o afrutadas (UDLAP, 2023).

Efectos del suelo:

- Suelos ricos en minerales pueden aportar matices específicos al vino, como notas minerales o terrosas.
- Los suelos calcáreos pueden contribuir a vinos más elegantes y longevos, mientras que los suelos arcillosos producen vinos más estructurados.
- El terroir influye en la identidad del vino y es uno de los principales criterios en la clasificación de Denominaciones de Origen (DO) en países como España, Francia e Italia.

La añada

La añada indica el año de cosecha de las uvas utilizadas en la elaboración del vino y es un factor crucial en su calidad y sabor.

- En climas fríos, la variación entre añadas puede ser significativa debido a cambios en temperatura y precipitaciones.
- En climas templados, la diferencia entre añadas es menor y la calidad del vino suele ser más constante (Grupo Pago de Mar, 2021).

En regiones vinícolas reconocidas, ciertos años son considerados de excelente calidad, mientras que otras añadas pueden dar vinos menos expresivos debido a condiciones climáticas adversas.

El proceso de vinificación y crianza

El proceso de elaboración del vino influye directamente en su perfil de sabor.

Fermentación

La fermentación es la conversión de los azúcares de la uva en alcohol por acción de levaduras. Este proceso define características clave del vino, como su nivel de alcohol, cuerpo y estructura.

Crianza en barrica

El envejecimiento en barrica de roble aporta compuestos aromáticos al vino, incluyendo notas de vainilla, coco, especias o tabaco. Cuanto mayor sea el tiempo en barrica, más complejo será el vino (Delgado, 2021).

Crianza en botella

Después de la barrica, muchos vinos siguen evolucionando en botella, desarrollando matices terciarios como cuero, frutos secos o balsámicos.

Filtración y clarificación

Estos procesos eliminan partículas en suspensión en el vino, afectando su textura y limpidez. Algunos productores evitan filtrados agresivos para preservar la autenticidad del vino.

2.4 Concepto de calidad en el vino

2.4.1 Cómo se mide la calidad en la industria

Los vinos de alta calidad deben cumplir con una serie de características, entre las cuales destacan:

- ✓ Sustancia: Relacionada con la plenitud y el cuerpo del vino, determinada por la concentración de ingredientes.
- ✓ Intensidad aromática: Se refiere a la potencia de los olores y sabores percibidos en la cata.
- ✓ Complejidad: Hace referencia a la diversidad de aromas y sabores, lo que aporta mayor riqueza sensorial.
- ✓ Equilibrio: Relación armoniosa entre dulzor, acidez, alcohol y taninos.
- ✓ Densidad: Concentración de compuestos que aportan textura y profundidad al vino.
- ✓ Persistencia: Duración de la impresión sensorial en el paladar.
- ✓ Longevidad: Potencial de envejecimiento del vino sin que se deteriore su calidad (Calidad del Vino y Análisis Sensorial, 2023).

La calidad del vino comienza desde la producción en el viñedo y se controla en todas las etapas del proceso de vinificación. Los principales factores medibles incluyen:

Rendimiento del viñedo y calidad de la uva

Uno de los indicadores clave de calidad es la cantidad de uvas producidas por cada cepa. En términos generales, un menor rendimiento suele traducirse en uvas más concentradas en sabor y mejor calidad del vino.

- Peso del mosto: Mide la cantidad de sustancias disueltas en el jugo de la uva y es un indicador clave del potencial de calidad del vino.
- Edad de las cepas: Las vides más antiguas producen menos uvas, pero con mayor concentración de compuestos aromáticos.
- Selección de la uva: Se realiza manualmente o mediante procesos automatizados para eliminar uvas defectuosas antes de la vinificación (Calidad del Vino y Análisis Sensorial, 2023).

Análisis químico del vino

Para garantizar la calidad en la producción, los vinos son sometidos a análisis químicos que determinan parámetros clave:

- pH y acidez total: Un pH entre 3.0 y 3.6 es ideal para mantener la frescura y estabilidad del vino.
- Azúcar residual: Clasifica el vino en seco, semiseco o dulce.
- Nivel de alcohol: Determina el cuerpo y la sensación en boca del vino.
- Taninos y polifenoles: Son responsables de la estructura y la longevidad del vino, especialmente en tintos.
- Sulfuroso (SO_2): Regula la oxidación y evita la proliferación de bacterias (Calidad del Vino y Análisis Sensorial, 2023).

El análisis espectrofotométrico y la cromatografía de gases son métodos empleados en laboratorios enológicos para medir estos parámetros con precisión.

Influencia del Terroir

El terroir es un concepto clave en la calidad del vino, ya que determina el carácter y la identidad del producto. Este término agrupa la influencia del clima, suelo y la ubicación geográfica del viñedo:

- Clima: Afecta la maduración de la uva. Climas fríos producen vinos más ácidos y elegantes, mientras que climas cálidos generan vinos con mayor contenido de azúcar y alcohol.
- Suelo: La composición mineral del suelo influye en la estructura y complejidad del vino. Suelos calcáreos favorecen vinos con buena acidez, mientras que suelos arcillosos aportan más cuerpo.
- Métodos de cultivo: La agricultura ecológica y biodinámica han ganado popularidad, ya que se asocian con vinos más expresivos y sostenibles (Calidad del Vino y Análisis Sensorial, 2023).

2.4.2 Diferencias entre calidad objetiva (química) y calidad subjetiva (percepción del consumidor)

Si bien ambos enfoques son fundamentales en la industria vinícola, no siempre están alineados, ya que un vino que cumple con altos estándares químicos puede no ser percibido como agradable por el consumidor, y viceversa.

Calidad objetiva del vino: Evaluación científica y química

La calidad objetiva se mide a través de análisis fisicoquímicos que determinan la composición del vino, asegurando que sea estable, saludable y cumpla con los estándares de producción.

Parámetro	Importancia en la calidad
pH y acidez total	Determina la frescura y estabilidad del vino. Un pH entre 3.0 y 3.6 es ideal.
Azúcar residual	Influye en la percepción del dulzor. Clasifica los vinos en secos, semisecos o dulces.
Nivel de alcohol	Afecta la estructura y sensación en boca. La mayoría de los vinos tienen entre 10% y 15% vol.
Taninos	Definen la capacidad de envejecimiento del vino
Sulfuroso total y libre	Conservante que protege contra la oxidación y contaminación microbiológica
Color e intensidad	Evalúa la concentración de pigmentos
Aromas y compuestos volátiles	Se estudia con cromatografía de gases

Fuente: Elaboración propia a partir de Wein.plus (2023).

3. Técnicas de Machine Learning aplicadas al vino

3.1 Sistemas de recomendación: KNN

Los sistemas de recomendación son algoritmos diseñados para sugerir productos a los usuarios en función de sus preferencias. En este caso, el objetivo es desarrollar un modelo que recomiende variedades de vino a los consumidores según su perfil de sabor (dulzura, acidez, taninos, cuerpo) y el uso que desean darle (mesa, postre, aperitivo).

El K-Nearest Neighbors (KNN) es un algoritmo que clasifica nuevos datos en función de su similitud con datos existentes. En el caso del vino, se usará para encontrar los vinos más similares a las preferencias de un usuario. Para la recomendación de vinos, funciona así: cada vino está representado por un conjunto de características numéricas, y cuando un usuario introduce sus preferencias, el algoritmo busca los k vinos más cercanos en el espacio de características, recomendando aquellos más similares.

Este enfoque no es nuevo: el algoritmo KNN ha sido utilizado anteriormente en diversos contextos relacionados con la alimentación y el consumo, como en sistemas de recomendación de cervezas (Pazzani & Billsus, 2007) o incluso en plataformas para sugerir recetas y planes nutricionales personalizados (Singh et al., 2021). En el ámbito específico del vino, estudios como el de Cortez et al. (2009) han empleado KNN como uno de los modelos para clasificar la calidad de vinos en función de atributos químicos, demostrando su aplicabilidad tanto en predicción como en recomendación.

3.1.1 Modelos predictivos de calidad: Regresión y Random Forest.

Los modelos de regresión lineal y regresión múltiple buscan establecer una relación entre las variables explicativas (características químicas del vino) y la variable objetivo (calidad del vino). En el vino, se aplica de la siguiente manera:
Se recopilan datos sobre la composición química del vino, como pH, acidez, nivel de alcohol, contenido de taninos y azúcares residuales. Luego, se entrenan modelos de regresión para encontrar qué variables tienen mayor impacto en la calificación de calidad del vino, y así, se obtiene una ecuación matemática que predice la calidad del vino en función de sus características químicas.

4. Análisis 1: Modelo predictivo de calidad del vino

4.1 Exploración de datos

4.1.1 Descripción de la base de datos

El conjunto de datos utilizado en este estudio contiene información sobre 6,497 muestras de vino, con un total de 14 variables, incluyendo características fisicoquímicas del vino y una variable de calidad que representa la valoración del producto.

Las variables incluidas en el dataset son:

- botella_id: Identificador único de cada muestra de vino (eliminado en los análisis).
- acidez fija: Representa los ácidos no volátiles presentes en el vino.
- acidez volátil: Mide la cantidad de ácido acético en el vino, que en concentraciones elevadas puede producir sabores no deseados.
- ácido cítrico: Contribuye a la frescura y percepción ácida del vino.
- azúcar residual: Cantidad de azúcar que queda después de la fermentación.
- cloruros: Representan la cantidad de sal en el vino, lo que influye en su sabor.
- dióxido de azufre libre: Conservante que previene la oxidación y el deterioro microbiano.
- dióxido de azufre total: Suma del dióxido de azufre libre y combinado.
- densidad: Relación entre la masa y el volumen del vino, influye en su cuerpo.
- pH: Indica el nivel de acidez del vino.
- sulfatos: Compuestos que afectan la estabilidad del vino y su conservación.
- alcohol: Contenido alcohólico en el vino.
- color: Tipo de vino (rojo o blanco).
- calidad: Variable objetivo del estudio, que mide la calidad del vino en una escala del 3 al 9

4.1.2 Análisis estadísticos de variables

La tabla muestra un análisis estadístico descriptivo de las variables fisicoquímicas del vino presentes en el dataset. Para cada variable, se indican valores como el mínimo, máximo, media, mediana (50%), desviación estándar y los cuartiles (25% y 75%). Esta información permite entender la distribución y variabilidad de cada característica, y es un paso clave previo al modelado, ya que ayuda a identificar posibles outliers, rangos anómalos o la necesidad de normalización. Además, este análisis inicial ofrece una visión general de cómo se comportan las diferentes propiedades del vino antes de aplicar algoritmos predictivos.

	botella_id	acidez fija	acidez volatil	acido citrico	azucar residual	cloruros	dioxido de azufre libre	dioxido de azufre total	densidad	pH	sulfatos	calidad
count	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000
mean	3248.000000	7.215307	0.339666	0.318633	5.443235	0.056034	30.525319	115.744574	1.015451	3.218501	0.531268	5.818378
std	1875.666681	1.296434	0.164636	0.145318	4.757804	0.035034	17.749400	56.521855	1.248536	0.160787	0.148806	0.873255
min	0.000000	3.800000	0.080000	0.000000	0.600000	0.009000	1.000000	6.000000	0.987110	2.720000	0.220000	3.000000
25%	1624.000000	6.400000	0.230000	0.250000	1.800000	0.038000	17.000000	77.000000	0.992340	3.110000	0.430000	5.000000
50%	3248.000000	7.000000	0.290000	0.310000	3.000000	0.047000	29.000000	118.000000	0.994890	3.210000	0.510000	6.000000
75%	4872.000000	7.700000	0.400000	0.390000	8.100000	0.065000	41.000000	156.000000	0.996990	3.320000	0.600000	6.000000
max	6496.000000	15.900000	1.580000	1.660000	65.800000	0.611000	289.000000	440.000000	100.012000	4.010000	2.000000	9.000000

Algunas observaciones importantes:

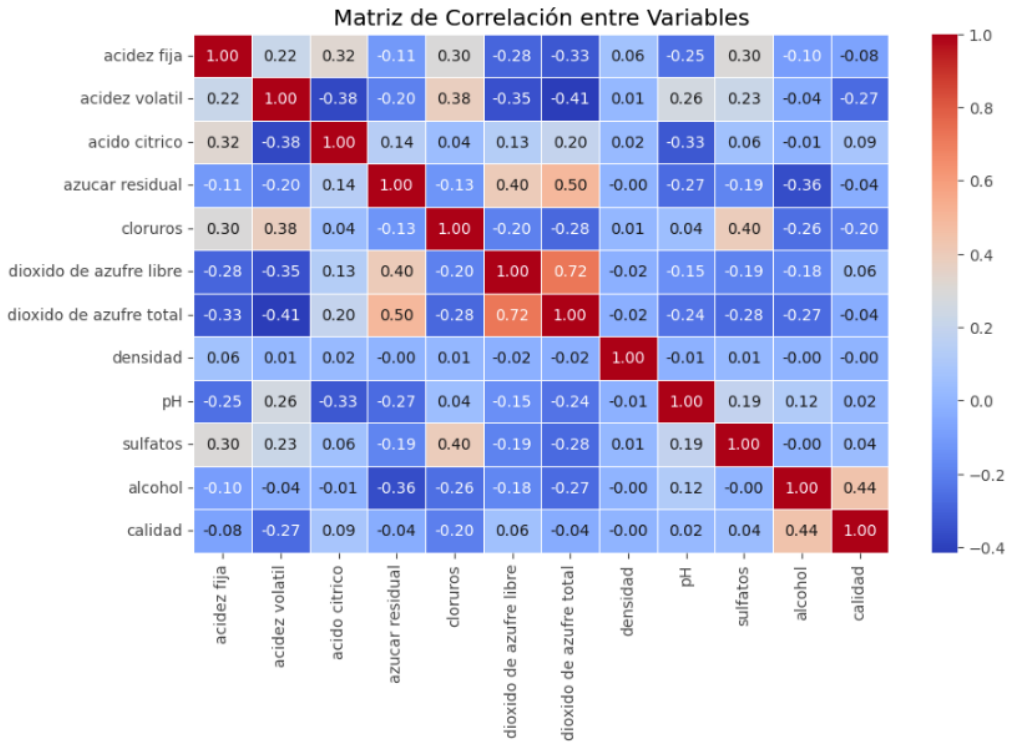
i. La calidad promedio del vino es 5.81, con una distribución que varía entre 3 y 9.

ii. El alcohol varía entre 8% y 15%, con una media de 10.5%.

iii. El nivel de acidez fija varía entre 3.8 y 15.9, lo que sugiere una amplia diversidad de perfiles de vino.

iv. El dióxido de azufre total y el azúcar residual presentan valores máximos muy elevados, lo que indica la presencia de algunos vinos con características muy distintas al promedio.

4.1.3 Identificación de patrones y correlaciones



La matriz de correlación muestra la relación lineal entre las distintas variables fisicoquímicas del vino y la variable objetivo "calidad". Los valores oscilan entre -1 y 1, donde valores positivos indican una correlación directa y valores negativos una correlación inversa. Esta representación permite identificar qué variables están más asociadas con la calidad del vino. Destacan el alcohol con una correlación positiva moderada (0.44) y la acidez volátil con una correlación negativa (-0.27), lo que sugiere que estas características influyen directamente en la percepción de calidad. Esta herramienta es clave para seleccionar variables relevantes en los modelos predictivos.

Principales hallazgos de la matriz de correlación:

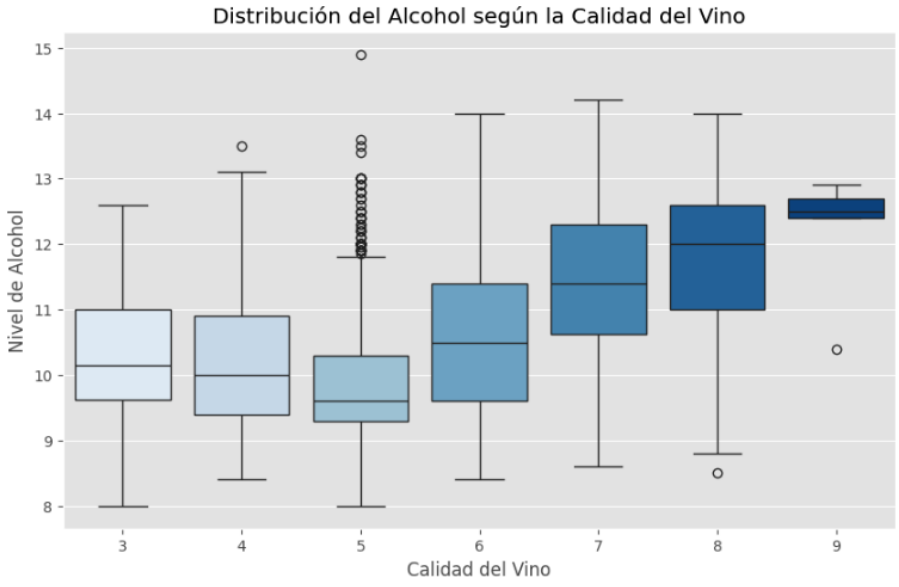
- El alcohol es la variable con mayor correlación positiva con la calidad del vino (0.44). Es decir, los vinos con mayor contenido alcohólico tienden a tener mejores puntuaciones de calidad.
- La acidez volátil tiene una correlación negativa con la calidad (-0.27), lo que indica que una acidez volátil elevada puede afectar negativamente la percepción del vino.
- El dióxido de azufre total y libre no tienen una relación fuerte con la calidad, aunque presentan una alta correlación entre sí (0.72).
- El pH tiene una relación débil con la calidad del vino, lo que sugiere que no es un factor determinante en la evaluación de los consumidores.
- Los sulfitos tienen una relación muy débil con la calidad, lo que indica que su uso en la conservación no afecta directamente la percepción de calidad.

Distribución de variables clave en función de la calidad

Se analizaron algunas características relevantes para observar cómo varía la calidad del vino en función de estas:

1- Alcohol vs. Calidad:

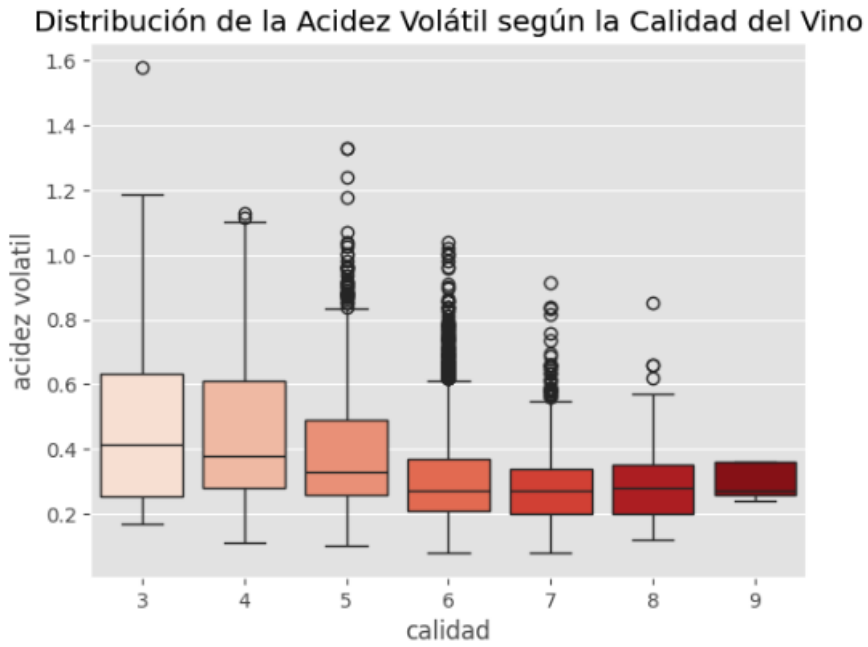
- Los vinos con mayor contenido de alcohol tienden a ser mejor valorados en términos de calidad.
- Los vinos con baja calidad presentan un menor nivel de alcohol, con una mediana en torno a 10%.
- Los vinos de mayor calidad (8-9) tienen una mediana superior a 12% de alcohol.



Este gráfico de caja representa la distribución del contenido de alcohol en función de la calidad del vino (escala del 3 al 9). Se observa una tendencia ascendente: los vinos con mayor puntuación de calidad tienden a tener niveles más altos de alcohol. Las medianas aumentan progresivamente, y los vinos mejor valorados (8 y 9) presentan una concentración alcohólica significativamente superior. Esto respalda la correlación positiva detectada previamente entre alcohol y calidad, sugiriendo que el contenido alcohólico puede ser un indicador relevante de la calidad percibida del vino.

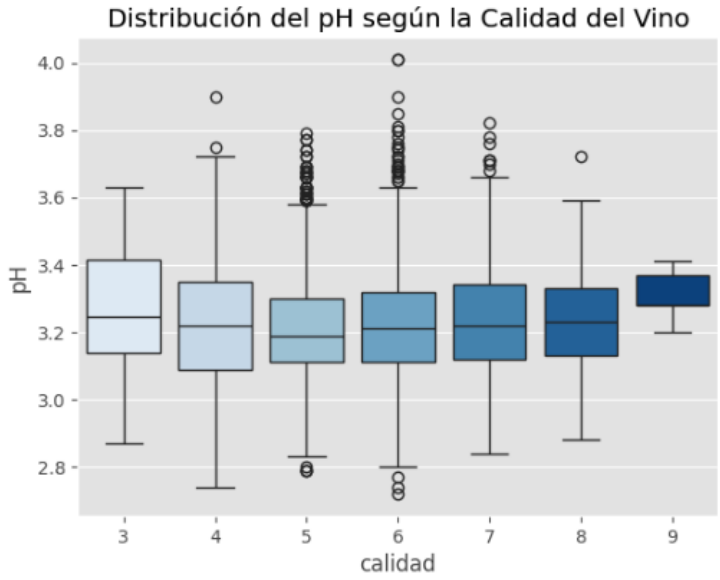
2- Acidez Volátil vs. Calidad:

- Los vinos de baja calidad presentan una acidez volátil más alta en comparación con los vinos de calidad superior.
- Los vinos con calidad entre 3 y 5 tienen una acidez volátil que varía entre 0.2 y 0.6.
- A medida que la calidad aumenta, la acidez volátil disminuye.



3- pH vs. Calidad:

- No hay una tendencia clara entre el pH y la calidad, lo que sugiere que el pH no es un determinante directo de la percepción de calidad.
- Sin embargo, se observa que los vinos con calidad superior a 7 suelen tener un pH más estable, sin valores extremos.



A partir del análisis de correlaciones y la distribución de variables, se puede concluir que el contenido de alcohol es el factor que más influye en la calidad del vino, mostrando una correlación positiva significativa. Esto sugiere que los vinos con un mayor porcentaje de alcohol tienden a ser mejor valorados en términos de calidad.

Por otro lado, la acidez volátil presenta una relación inversa con la calidad del vino, lo que indica que niveles elevados de este componente pueden afectar negativamente la percepción del producto. Dado que la acidez volátil está relacionada con la presencia de ácido acético, su exceso podría generar defectos sensoriales que disminuyan la valoración del vino.

Finalmente, se observa que tanto el pH como los sulfitos tienen una correlación muy débil con la calidad del vino. Esto sugiere que, si bien son parámetros importantes para la estabilidad y conservación del producto, no parecen ser determinantes en la evaluación sensorial que realizan los consumidores.

4.2 Construcción del modelo

4.2.1 Procesamiento de datos

El primer paso en la construcción del modelo predictivo de calidad del vino consistió en la preparación y limpieza de los datos para garantizar que fueran adecuados para el análisis y entrenamiento del modelo de Machine Learning.

1. Exploración y limpieza de datos: Se partió de un conjunto de datos compuesto por 6,497 observaciones y 12 variables, incluyendo características fisicoquímicas del vino, como acidez, azúcares, sulfitos, densidad, nivel de alcohol y pH, además de la

variable objetivo calidad, que es una clasificación de 3 a 9. Inicialmente, se detectaron dos problemas clave:

- Formato de datos incorrecto: La variable alcohol estaba codificada como un tipo object en lugar de un float, lo que impedía su uso en los modelos predictivos. Se realizó la conversión a formato numérico.
- Presencia de valores categóricos irrelevantes: La variable color (con valores como "rojo" o "blanco") fue eliminada del análisis, ya que no era una variable numérica directamente útil para los modelos.
- Variable identificadora: botella_id fue eliminada, ya que no aporta información útil para la predicción de la calidad del vino.

2. Normalización de datos: Dado que los modelos de Machine Learning pueden verse afectados por la escala de los datos, se aplicó una normalización utilizando StandardScaler, que estandariza cada variable restando la media y dividiendo por la desviación estándar. Esto fue necesario especialmente para variables como dióxido de azufre total, azúcar residual y acidez fija, que presentan diferentes escalas en comparación con otras características.
3. División del conjunto de datos Para evaluar correctamente el desempeño del modelo, el conjunto de datos fue dividido en:
   – 80% para entrenamiento (5,196 observaciones)
   – 20% para prueba (1,300 observaciones)

   Esta división permite evaluar la capacidad del modelo para generalizar a datos nuevos y evitar problemas de sobreajuste.
4. Tratamiento de desequilibrios en la variable objetivo: Se observó que la variable calidad no estaba distribuida uniformemente, con la mayoría de los vinos concentrados en valores de calidad 5, 6 y 7, mientras que categorías extremas como 3 y 9 tenían muy pocos ejemplos. Esto sugiere que el modelo podría tener dificultades para predecir correctamente los valores menos representados. Para mitigar este problema, se consideró la posibilidad de:
   - Aplicar técnicas de sobremuestreo (duplicar ejemplos de las clases menos representadas).
   - Aplicar reescalamiento de pesos en los modelos para que den más importancia a clases menos frecuentes.

4.2.2 Comparación de modelos de predicción

Para determinar el modelo más adecuado para predecir la calidad del vino en función de sus características químicas, se evaluaron tres enfoques distintos de Machine Learning: Regresión Logística, Random Forest y K-Nearest Neighbors (KNN). Cada modelo fue entrenado y probado utilizando el mismo conjunto de datos, dividiéndolo en un 80% para entrenamiento y un 20% para prueba.

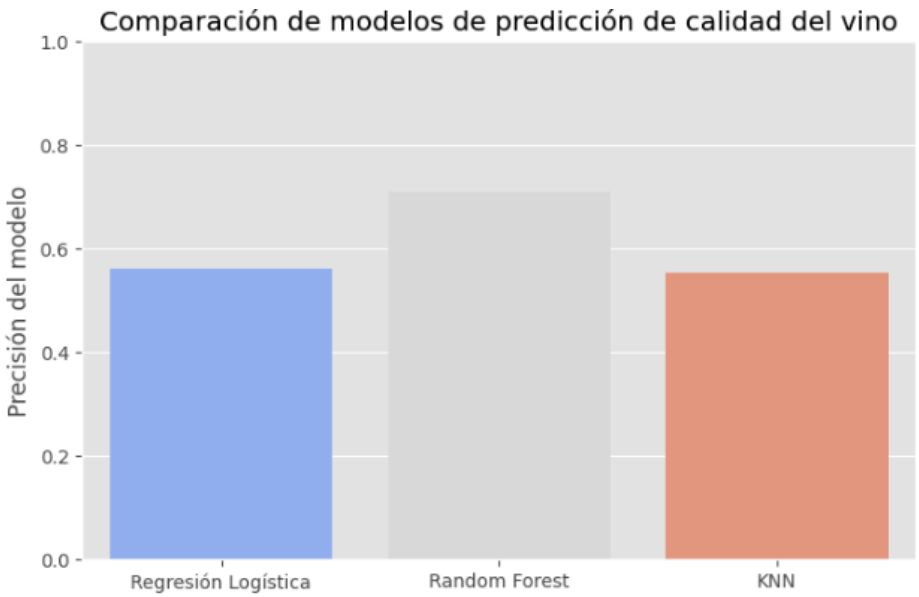
Los resultados obtenidos para cada modelo se presentan a continuación:

- Regresión Logística: Precisión de 0.56. Se observó un bajo rendimiento, especialmente para clases menos representadas.
- KNN (K-Nearest Neighbors): Precisión de 0.55. Presentó dificultades para clasificar correctamente las muestras de calidad extrema.
- Random Forest: Precisión de 0.71. Obtuvo la mejor performance general, con mayor capacidad de predicción para diferentes niveles de calidad.

Dado que el modelo de Random Forest fue el que obtuvo el mejor rendimiento en la fase comparativa inicial (precisión de 0.71), se decidió profundizar en su ajuste mediante la optimización de hiperparámetros utilizando la técnica GridSearchCV.

Esta herramienta permite realizar una búsqueda sistemática en un conjunto predefinido de combinaciones de hiperparámetros (como n_estimators, max_depth, min_samples_split, entre otros), evaluando cada combinación mediante validación cruzada.

El objetivo era mejorar el rendimiento del modelo y reducir el riesgo de sobreajuste, maximizando su capacidad de generalización en datos no vistos. La elección de este procedimiento se basa en su uso estándar y eficaz dentro del campo del Machine Learning para modelos de tipo ensamblado como Random Forest.



4.2.3 Evaluación del modelo y selección del mejor enfoque

Tras la optimización de hiperparámetros, el modelo Random Forest mejoró su rendimiento. Se ajustaron los siguientes hiperparámetros:

- n_estimators: 300
- max_depth: 30
- min_samples_split: 2
- min_samples_leaf: 1
- class_weight: 'balanced_subsample'

Estos cambios resultaron en un modelo con una precisión del 0.71, con mejoras en el recall y el f1-score para las clases menos representadas.

Para abordar el problema del desbalanceo en la distribución de clases, se utilizó SMOTE (Synthetic Minority Over-sampling Technique). Sin embargo, la aplicación de SMOTE resultó en una ligera disminución de la precisión (0.69), indicando que el sobremuestreo no fue necesario en este caso.

Finalmente, se analizó la importancia de las variables dentro del modelo Random Forest. Los resultados mostraron que el alcohol es la característica más determinante para la calidad del vino, seguido por los cloruros y la densidad. La acidez volátil tuvo una fuerte correlación negativa, lo que indica que valores elevados pueden perjudicar la percepción de calidad.

El modelo Random Forest optimizado fue seleccionado como el mejor enfoque para la predicción de la calidad del vino. Este modelo no solo obtuvo la mejor precisión global,

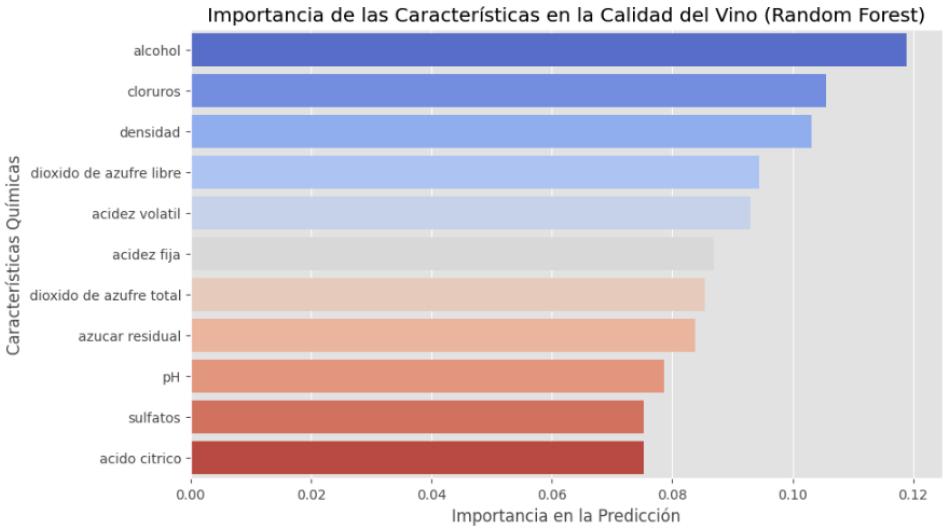
sino que también logró un mejor equilibrio entre las distintas clases, permitiendo una clasificación más fiable en función de las características químicas del vino.

4.3 Resultados y discusión

4.3.1 ¿Qué características impactan más en la calidad?

El análisis de importancia de variables en el modelo Random Forest reveló que el contenido de alcohol es el factor más determinante en la predicción de la calidad del vino, seguido de los cloruros, la densidad y el dióxido de azufre libre. Esto sugiere que los vinos con mayor graduación alcohólica tienden a recibir mejores puntuaciones de calidad, lo que puede estar relacionado con la percepción sensorial del cuerpo y el equilibrio del vino.

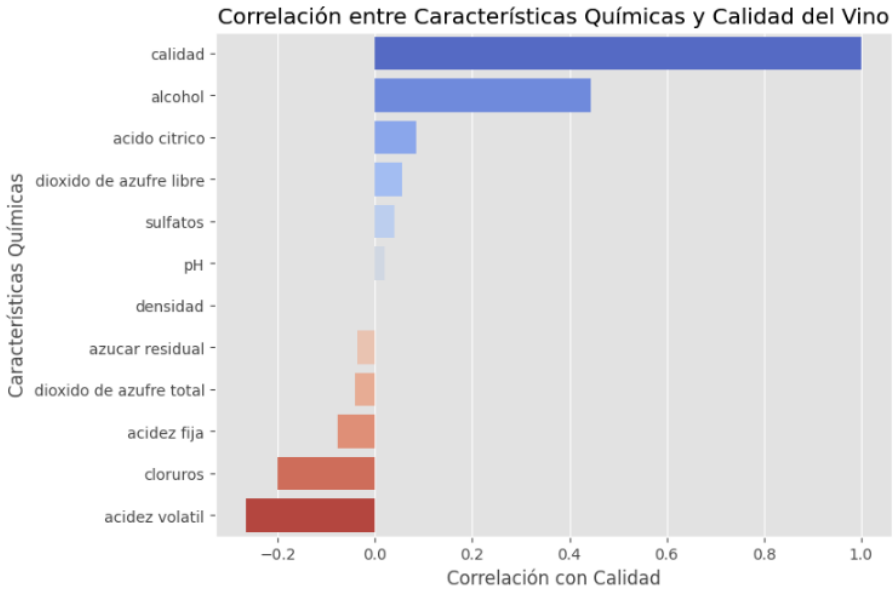
Por otro lado, la acidez volátil mostró una correlación negativa con la calidad, indicando que valores elevados de este componente pueden perjudicar la percepción del producto, posiblemente debido a la generación de aromas desagradables. Características como el pH y los sulfitos presentaron una relación más débil con la calidad, lo que indica que su impacto en la percepción global es menor.



4.3.2 ¿Es posible explicar completamente la calidad de un vino solo con datos químicos?

Si bien las características químicas permiten hacer una estimación bastante precisa de la calidad del vino, los resultados sugieren que la calidad percibida no puede ser explicada únicamente por estos factores. La precisión máxima obtenida en el modelo, incluso tras la optimización, fue del 72%, lo que indica que hay otros elementos que influyen en la percepción sensorial del consumidor y que no están reflejados en los datos químicos.

Factores externos como la variedad de la uva, las condiciones de fermentación, el envejecimiento en barrica y la percepción subjetiva del consumidor pueden jugar un papel clave en la calidad final del vino. Esto indica que, aunque el análisis químico es una herramienta valiosa para evaluar la calidad del vino de manera objetiva, no sustituye completamente la evaluación sensorial realizada por expertos y consumidores.



5. Análisis 2: Modelo de recomendación de vinos

5.1 Diseño del modelo de recomendación

5.1.1 Definición de inputs (dulzura, acidez y cuerpo

Para diseñar un sistema de recomendación personalizado, el modelo parte de las preferencias del usuario en relación con ciertas características clave del vino.
Los inputs definidos en el sistema son:

- Dulzura (sweetness): Nivel de dulzura del vino en una escala del 1 al 5 (1 = seco, 5 = muy dulce).
- Acidez (acidity): Percepción de frescura del vino en una escala del 1 al 5 (1 = baja acidez, 5 = muy ácido).
- Cuerpo (body): Intensidad y densidad del vino en una escala del 1 al 5 (1 = ligero, 5 = con mucho cuerpo).

Una vez que el usuario proporciona sus preferencias, el modelo buscará en la base de datos las variedades de vino que más se asemejen a ellas y recomendará las 5 variedades junto con 5 vinos más adecuados y su preferencia de 'uso'.

5.1.2 Explicación del algoritmo KNN y su lógica

El algoritmo KNN permite recomendar un vino basándose en los valores de dulzura, acidez y cuerpo que el usuario introduce. En lugar de entrenar un modelo con datos etiquetados, KNN simplemente compara las preferencias del usuario con los vinos en la base de datos y sugiere aquellos más similares.
Los pasos clave del modelo son:

1. Definir las características

Se consideran las variables: dulzura, acidez y cuerpo.
Cada vino en la base de datos tiene valores asociados a estas características.

2. Estandarización de datos

Como los valores de las características pueden tener diferentes escalas, se normalizan usando un StandardScaler para que tengan una media de 0 y una desviación estándar de 1.

3. Cálculo de la distancia entre el usuario y cada vino

Se calcula la distancia entre las preferencias del usuario y los vinos en la base de datos usando la distancia euclidiana, que mide la diferencia entre los puntos en un espacio multidimensional.

Fórmula de la distancia euclidiana: $d(A,B) = \sqrt{(A_1 - B_1)^2 + (A_2 - B_2)^2 + (A_3 - B_3)^2}$

Donde:

- AAA son las características del usuario (dulzura, acidez, cuerpo).
- BBB son las características de cada vino en la base de datos.

4. Selección del vecino más cercano

Se usa un modelo KNN con k=1, lo que significa que se selecciona el vino con la menor distancia al usuario. Esto permite encontrar la variedad de vino más similar a los gustos del usuario.

5. Recomendación de 5 vinos

Una vez identificada la variedad de vino más adecuada, se seleccionan 5 vinos específicos de esa variedad y se muestran al usuario junto con su uso recomendado (Table, Appetizer, Dessert, etc.).

5.2 Entrenamiento y pruebas del modelo

5.2.1 Procesamiento de datos y normalización

Para construir un modelo de recomendación de vinos eficiente y preciso, ha sido necesario transformar y preparar la base de datos de manera que sus variables sean interpretables por el algoritmo de Machine Learning. A continuación, se detallan los pasos llevados a cabo en este proceso.

La base de datos contenía información en diversos formatos, por lo que fue necesario estandarizar y convertir algunas variables en valores numéricos para su uso en el modelo.

Selección de una única variedad de uva

El dataset original contenía hasta cuatro columnas diferentes para describir las variedades de uva de cada vino (**varieties1, varieties2, varieties3 y varieties4**). Para simplificar esta información:

- Se mantuvo la **primera variedad disponible**, siguiendo el orden de importancia de las columnas.
- Si **varieties1 estaba vacía**, se tomó varieties2.
- Si **varieties2 también estaba vacía**, se tomó varieties3.
- Si **varieties3 estaba vacía**, se tomó varieties4.
- Si todas estaban vacías, se marcó como **NaN** (aunque esto ocurría en muy pocos casos).

De esta manera, cada vino quedó representado por una única variedad bajo la columna **variety_final**.

Conversión de Variables Categóricas a Números

Algunas características del vino estaban en forma de etiquetas textuales y debían convertirse a valores numéricos para su correcta interpretación en el modelo.

Dulzura (Sweet), Acidez (Acidity), Cuerpo (Body) y Taninos (Tannin)

Estas variables estaban categorizadas como:

- **SWEET1, SWEET2, ..., SWEET5**
- **ACIDITY1, ACIDITY2, ..., ACIDITY5**
- **BODY1, BODY2, ..., BODY5**
- **TANNIN1, TANNIN2, ..., TANNIN5**

Se realizó una **codificación ordinal**, transformando estas etiquetas en valores numéricos del **1 al 5**, preservando su jerarquía.

Transformación del Grado Alcohólico

La variable **degree** indicaba el nivel de alcohol de los vinos en distintos formatos, como:

- Rango de valores: **"14~15"**, **"17~19"**, **"10~12"**
- Valores únicos: **"14"**, **"12.5"**, **"13.8"**
- Algunos valores inconsistentes o atípicos (ej. **"40"**, **"60~70"**).

Para estandarizar esta variable:

- Se sacó el **promedio de cada rango**.
- Se eliminaron valores extremos o no relevantes para vinos estándar.
- Se creó una **nueva variable discreta (degree_category)** que agrupa los grados en 5 categorías:
    - **1** → Menos de 10°
    - **2** → Entre 10° y 12°
    - **3** → Entre 12° y 14°
    - **4** → Entre 14° y 16°
    - **5** → Más de 16°

Esto permitió utilizar esta variable como un factor de intensidad del alcohol en el modelo.

Simplificación del Uso del Vino (Use)

La variable **use** contenía múltiples combinaciones, como:

- "Table"
- "Dessert"
- "Appetizer"
- "Appetizer, Table"
- "Table, Dessert"
- "Appetizer, Table, Dessert, Etc"

Para hacerla más manejable, se reclasificó en cuatro categorías:

- Si solo tenía **Table, Dessert o Appetizer**, se mantuvo sin cambios.
- Si tenía combinaciones o valores nulos, se agrupó como **"all uses"**.

Esto permitió que el modelo pudiera diferenciar fácilmente los vinos en función de su uso principal.

Selección de Variables Relevantes

Para la construcción del modelo, solo se mantuvieron las variables necesarias:

- **name** → Nombre del vino
- **variety_final** → Variedad de uva predominante
- **type** → Tipo de vino (Red/White)
- **use** → Uso recomendado
- **degree_category** → Categoría de grado alcohólico
- **sweet** → Nivel de dulzura
- **acidity** → Nivel de acidez
- **body** → Nivel de cuerpo
- **tannin** → Nivel de taninos

El resto de las variables fueron eliminadas al no ser necesarias para el proceso de recomendación.

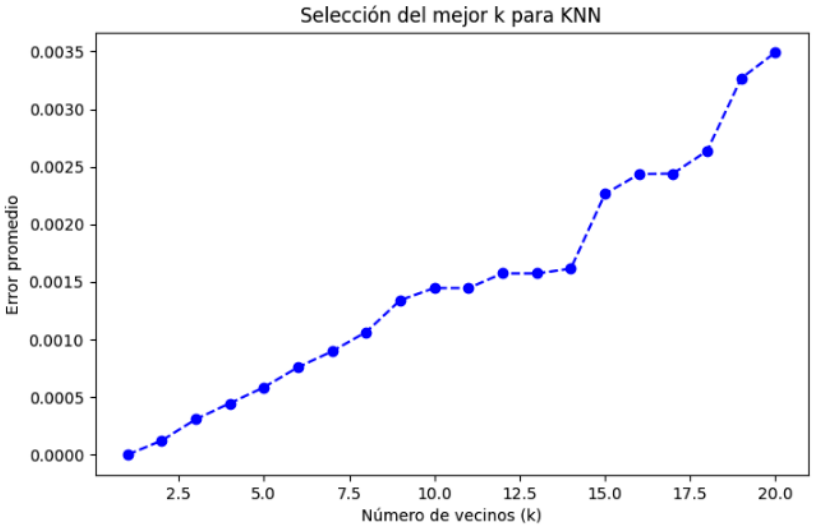
Escalado de Variables

Las variables **sweet, acidity, body, tannin** y **degree_category** estaban en una escala del 1 al 5. No obstante, el **KNN** se ve afectado por las diferencias de escala entre variables. Para evitar que una variable con valores más altos tenga mayor peso en la distancia, aplicamos **Min-Max Scaling** con la fórmula:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Así, todas las variables quedaron en un rango de 0 a 1 para que cada una contribuyera de manera equitativa a la recomendación.

5.2.2 Selección del número de vecinos (k)



- Error bajo con valores pequeños de k (1-5): Se observa que el error es prácticamente nulo con valores pequeños de k, lo que indica que el modelo se ajusta demasiado a los datos (overfitting).
- Aumento progresivo del error: A partir de k = 6 o 7, el error empieza a aumentar, lo que sugiere que la generalización mejora ligeramente, pero aún se mantiene dentro de un rango aceptable.
- Salto en el error después de k ≈ 15: A partir de este punto, el error empieza a aumentar más rápido, lo que indica que el modelo empieza a perder precisión.

Basándonos en la gráfica, un k entre 5 y 10 parece ser el punto óptimo. k = 5 o k = 7 sería una buena elección, ya que mantiene un bajo error sin caer en un sobreajuste excesivo.

5.2.3 Evaluación del rendimiento y mejoras posibles

El modelo fue probado con distintas preferencias de usuario para verificar su desempeño. Algunas de las mejoras identificadas incluyen:

- **Ajustar la ponderación de variables** si alguna característica influye demasiado en la recomendación.
- **Aplicar técnicas de clustering previas** para mejorar la segmentación de vinos antes de la recomendación.
- **Incluir más características del vino**, como la región de origen o método de producción.

5.3 Resultados y discusión

Tras la implementación del modelo de recomendación de vinos basado en KNN, se evaluó su desempeño a partir de distintos perfiles de consumidores con preferencias variadas en cuanto a dulzura, acidez y cuerpo. A continuación, se presentan tres ejemplos representativos y los vinos sugeridos por el sistema.

Ejemplo 1: Persona que prefiere un vino seco, poco ácido y con mucho cuerpo

Dulzura: 1 (muy seco)	Acidez: 2 (baja)	Cuerpo: 5 (muy estructurado)

name	variety_final	use
Chateau Ste. Michelle, Cold Creek Cabernet Sau...	Cabernet Sauvignon	Table
Raymond, Reserve Selection Cabernet Sauvignon	Cabernet Sauvignon	Table
Arciero, Zinfandel	Zinfandel	Table
Kenwood, Jack London Zinfandel	Zinfandel	Table
Terredora, Taurasi Campore Riserva	Aglianico	Table
Mastroberardino, Naturalis Historia	Aglianico	Table
Terredora, Taurasi Fatica Contatina	Aglianico	Table

Los vinos recomendados pertenecen en su mayoría a variedades con gran cuerpo, como Cabernet Sauvignon, que es reconocido por su estructura robusta y baja acidez. Además, estos vinos suelen envejecer bien y ser utilizados principalmente en comidas (Table), lo que encaja con el perfil del usuario.

Ejemplo 2: Persona que prefiere un vino muy dulce, con acidez media y poco cuerpo

Dulzura: 5 (muy dulce)	Acidez: 3 (media)	Cuerpo: 2 (ligero)

name	variety_final	use
Dr. ZenZen, Icewein	Silvaner	Dessert
Fritz Windisch, Silvaner Eiswein	Silvaner	Dessert
Fritz Windisch, Chardonnay Eiswein	Chardonnay	Dessert
Silvaner Icewine	Silvaner	Dessert
Windisch, Silvaner Eiswein	Silvaner	Dessert
Ginestet Sauternes	Semillon	Dessert
Castelnau de Suduiraut	Semillon	Dessert

En este caso, el modelo recomienda vinos de postre como Eiswein o Sauternes, que son conocidos por su elevada concentración de azúcar y su equilibrio con la acidez. Este tipo de vinos suelen servirse como acompañamiento de postres o quesos azules, realzando sus sabores con la combinación de dulzura y frescura.

Ejemplo 3: Persona que prefiere un vino semidulce, ácido y con cuerpo medio

Dulzura: 3 (dulce)	Acidez: 4 (alta)	Cuerpo: 3 (medio)

name	variety_final	use
Maxwell Frontignac Spatles	Riesling	Table
Ciel Cachirakgol Winery, RED	Muscat Bailey A , MBA	Table
Robert Weil, Rheingau Riesling Kiedrich Grafen...	Riesling	Table
Kendermann, Black Tower Red	Dornfelder	Table
Blue Nun, White	Rivaner	all uses
Blue Nun, Dornfelder	Dornfelder	Table
Kendermann, Black Tower Pink	Portugieser	all uses

Aquí, las recomendaciones incluyen Riesling, Dornfelder y Muscat Bailey A, variedades que tienden a presentar un balance entre dulzura y acidez, lo que las hace perfectas para platos de comida asiática, mariscos o incluso ensaladas con frutas.

El análisis de los resultados permite extraer varias conclusiones importantes sobre la capacidad del modelo para capturar las preferencias individuales y traducirlas en recomendaciones de vinos relevantes:

1. **Influencia de la dulzura en la recomendación:** Los usuarios que prefieren vinos muy dulces tienden a recibir recomendaciones de vinos de postre o fortificados, mientras que aquellos que buscan vinos secos obtienen principalmente sugerencias de vinos tintos estructurados.
2. **Acidez como factor de diferenciación:** La acidez del vino juega un papel clave en la recomendación, ya que los vinos con alta acidez se asocian con frescura y suelen recomendarse para platos más ligeros, mientras que los vinos con baja acidez son percibidos como más redondos y suaves.
3. **Importancia del cuerpo:** La preferencia por vinos con más cuerpo lleva a recomendaciones de vinos con mayor estructura y contenido tánico, mientras que los vinos ligeros tienden a sugerirse para un consumo más casual.
4. **Precisión del modelo:** El modelo de KNN basado en distancias euclidianas ha demostrado ser efectivo en capturar patrones de preferencia, proporcionando recomendaciones que son coherentes con las características esperadas de cada tipo de vino.

En conclusión, el sistema desarrollado ha demostrado su capacidad para adaptar las recomendaciones a los distintos perfiles de consumidores, facilitando la elección de vinos de acuerdo con sus preferencias individuales. Sin embargo, se podrían explorar mejoras adicionales en el modelo, como la incorporación de otras variables (ejemplo: precio o región de origen) para enriquecer la precisión de las recomendaciones.

6. Conclusiones

6.1 Conclusiones generales

Este trabajo ha permitido abordar de forma integral dos aspectos fundamentales de la industria del vino: la predicción de su calidad en función de características químicas y el desarrollo de un sistema de recomendación personalizado basado en preferencias sensoriales. Ambos enfoques, sustentados en técnicas de Machine Learning, han demostrado tener un alto potencial para mejorar la experiencia del consumidor y optimizar la toma de decisiones en la industria vitivinícola.

En primer lugar, se ha comprobado que la calidad del vino está parcialmente explicada por sus características fisicoquímicas. Variables como el contenido de alcohol, los cloruros y la densidad han resultado ser determinantes en el modelo predictivo, mientras que otras, como el pH o los sulfitos, muestran una influencia más limitada. Sin embargo, los resultados también indican que factores subjetivos o externos —como el precio, la marca o la percepción individual— siguen jugando un papel crucial en la valoración final del producto. Esto refuerza la idea de que la calidad del vino no puede ser entendida únicamente desde un punto de vista técnico, sino que está estrechamente ligada a la experiencia del consumidor.

En segundo lugar, el modelo de recomendación desarrollado se presenta como una herramienta de gran utilidad en el contexto actual, especialmente en una sociedad en la que los consumidores jóvenes muestran un menor conocimiento e interés inicial por el vino. Muchos de ellos no saben si realmente les gusta el vino ni qué estilo podría ajustarse mejor a sus preferencias. El sistema propuesto permite, a partir de inputs sencillos como dulzura, acidez y cuerpo, ofrecer recomendaciones personalizadas que faciliten la exploración de este producto de manera más accesible y atractiva.

Además, al vincular las recomendaciones con variables objetivas del vino, como su composición y estilo, se ofrece una experiencia más informada, que puede incluso

orientar decisiones de compra más racionales. Esto también puede ser útil a la hora de valorar si el precio de un vino está justificado por su calidad técnica, o si responde más bien a elementos subjetivos como el marketing, la región o la exclusividad.

En definitiva, el estudio demuestra que las herramientas de Business Analytics y Machine Learning pueden aportar un valor significativo tanto en el análisis como en la recomendación de vinos. Aplicaciones como esta pueden contribuir a acercar el mundo del vino a nuevos públicos, democratizando su consumo y fomentando una cultura más informada, accesible y personalizada.

6.2 Limitaciones del estudio

Una de las principales dificultades encontradas fue la búsqueda de una base de datos adecuada que permitiera dar respuesta a las dos grandes preguntas de investigación planteadas: por un lado, la relación entre las características químicas del vino y su calidad; y por otro, la posibilidad de recomendar vinos en función de las preferencias del consumidor. Encontrar conjuntos de datos fiables, completos y actualizados en el sector del vino no resultó sencillo, y en muchos casos fue necesario realizar un extenso trabajo de limpieza, transformación y unificación de información para adaptar los datos disponibles a los objetivos del estudio.

Además, una limitación destacada ha sido la ausencia de una base de datos que integrara simultáneamente características fisicoquímicas y sensoriales del vino. Las características sensoriales —como el aroma, el sabor, la persistencia o las notas organolépticas específicas— juegan un papel central en la percepción de calidad y en las decisiones de consumo. La falta de estas variables impidió evaluar con mayor profundidad la dimensión subjetiva del vino, que es justamente uno de los factores más diferenciadores en su valoración.

Asimismo, el modelo de recomendación se basa exclusivamente en variables como dulzura, acidez, cuerpo y taninos, dejando fuera elementos igualmente relevantes como el precio, la región de origen, el maridaje recomendado, la marca o incluso la ocasión de consumo. Esto puede limitar su aplicabilidad en situaciones reales, donde las preferencias del consumidor son mucho más complejas y multidimensionales.

Por último, hay que tener en cuenta que tanto los modelos predictivos como los sistemas de recomendación están sujetos a la calidad y representatividad de los datos empleados. Dado que algunos perfiles de vino o niveles de calidad estaban poco representados, los resultados podrían estar sesgados hacia los segmentos más frecuentes del dataset.

6.3 Líneas futuras de investigación

Este trabajo abre la puerta a múltiples líneas de investigación que podrían complementar y profundizar los hallazgos obtenidos, así como mejorar la utilidad y precisión de los modelos desarrollados.

1. Integración de características sensoriales y emocionales

Una línea prioritaria sería trabajar con bases de datos que incluyan descripciones sensoriales más detalladas del vino: aromas (frutales, florales, especiados), sabores predominantes, notas de envejecimiento o sensaciones en boca. Incluso variables subjetivas como emociones asociadas al consumo (relajación, celebración, sofisticación) podrían contribuir a construir sistemas de recomendación más personalizados y realistas.

2. Incorporación del precio como variable de análisis

Otra evolución lógica del estudio sería incorporar el precio del vino. Esto permitiría contrastar si vinos con mejores características químicas o sensoriales se corresponden realmente con precios más altos, o si hay vinos con buena composición que están infravalorados. Así, se podría recomendar no solo el vino más acorde a los gustos del consumidor, sino también el que ofrece mejor relación calidad-precio.

3. Análisis de percepción del consumidor y reseñas en línea

La combinación de datos estructurados (composición química, región, precio) con datos no estructurados, como opiniones de consumidores, puntuaciones en plataformas online o reseñas de sumilleres, permitiría enriquecer el análisis. Herramientas de procesamiento de lenguaje natural (NLP) podrían extraer patrones clave en las valoraciones subjetivas y compararlas con la calidad objetiva del vino.

4. Ampliación del sistema de recomendación

Actualmente, el sistema de recomendación se basa en las características básicas de dulzura, acidez y cuerpo. En el futuro, este sistema podría ampliarse incorporando elementos como tipo de ocasión (regalo, celebración, cena informal), tipo de comida, estación del año o incluso perfil demográfico del usuario. Esto permitiría generar recomendaciones mucho más sofisticadas, similares a las de una experiencia de compra guiada.

7. Bibliografía

Banco de España. (2022). *Informe sobre el sector agroalimentario y la viticultura en España*.

Bodega Cortijo Moros Santos. (2023). *Clasificación y tipos de vinos*. Recuperado de https://bodegasmorosanto.com/conoces-los-tipos-de-vino-que-existen/

Bodegas Viñedos Amaró. (2022). *Propiedades organolépticas del vino, un placer sensorial*. Recuperado de https://bodegasamaro.com/propiedades-organolepticas-del-vino-un-placer-sensorial/

Carrera, E. (2019). *Las 5 características básicas de un vino*. Recuperado de https://www.maset.com/es/blog/las-5-caracteristicas-basicas-de-un-vino

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.

Delgado, M. (2021). *Factores que influyen en el sabor del vino*. Recuperado de https://tauber.es/articulos/que-factores-influyen-en-el-sabor-y-aroma-del-vino/

Escuela Europea de Versalles. (2020). *¿Qué tipos de vino conoces?* Recuperado de https://escuelaversailles.com/tipos-de-vino/

Grupo Pago de Mar. (2021). *Los factores que influyen en el sabor de un vino*. Recuperado de https://www.pazodomar.com/los-factores-que-influyen-en-el-sabor-de-un-vino/

Organización Internacional de la Viña y el Vino (OIV). (2023). *State of the World Vitivinicultural Sector in 2022*. Recuperado de https://www.oiv.int

Our World in Data. (2023). *Wine production, 1961 to 2021*. Recuperado de https://ourworldindata.org/grapher/wine-production

Pazzani, M. J., & Billsus, D. (2007). Content-based recommendation systems. En P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The Adaptive Web* (pp. 325–341). Springer.

Singh, A., Bharti, A., & Kumar, A. (2021). A Personalized Recipe Recommendation System using KNN and Content-Based Filtering. *International Journal of Computer Applications, 183*(23), 1–5.

UDLAP. (2023). *Elementos que afectan la calidad y el sabor del vino*. Recuperado de https://cientisol.com/guia-definitiva-sobre-el-analisis-quimico-del-vino-que-parametros-hay-que-medir-y-como-hacerlo/

Wein.plus. (2023). *¿Qué factores influyen en la calidad del vino? Calidad del vino y análisis sensorial*. Recuperado de https://revista.wein.plus/faq/calidad-del-vino-y-analisis-sensorial/que-factores-influyen-en-la-calidad-del-vino

8. Anexos

Anexo 1. Código de Python

Código modelo de predicción de calidad de vino:

```
# -*- coding: utf-8 -*-
"""Copia de TFGBA.ipynb
Automatically generated by Colab.
Original file is located at
    https://colab.research.google.com/drive/1i5mElE4H1Jar1bHsOgFQzHCsHaosHZ0-
"""
from google.colab import drive
import pandas as pd
drive.mount('/content/drive')
file_path = '/content/drive/My Drive/calidad_de_vino_TFG.csv'
df = pd.read_csv(file_path)
df.head()
import pandas as pd
file_path = '/content/drive/My Drive/calidad_de_vino_TFG.csv'
df = pd.read_csv(file_path, sep=";")
df.head()
"""**EXPLORACIÓN DE LA BASE DE DATOS**"""
print(f"El dataset tiene {df.shape[0]} filas y {df.shape[1]} columnas.\n")
print("Tipos de datos en cada columna:\n")
print(df.dtypes)
# Resumen general del dataset
df.info()
# Identificar valores nulos
print("\nValores nulos por columna:\n")
print(df.isnull().sum())
# Estadísticas generales
df.describe()
"""**Análisis estadísticos de variables**"""
import matplotlib.pyplot as plt
import seaborn as sns
plt.style.use('ggplot')
# Histograma de todas las variables
df.hist(figsize=(12, 8), bins=20, edgecolor='black')
plt.suptitle("Distribución de las Variables del Dataset", fontsize=14)
plt.show()
"""# Cómo varía la calidad del vino en función de algunas características relevantes"""
df['alcohol'] = pd.to_numeric(df['alcohol'], errors='coerce')
# Boxplot de la calidad del vino según los niveles de alcohol
plt.figure(figsize=(10, 6))
sns.boxplot(x=df['calidad'], y=df['alcohol'], palette="Blues")
plt.title("Distribución del Alcohol según la Calidad del Vino")
plt.xlabel("Calidad del Vino")
plt.ylabel("Nivel de Alcohol")
plt.show()
```

```
sns.boxplot(x=df['calidad'], y=df['acidez volatil'], palette="Reds")
plt.title("Distribución de la Acidez Volátil según la Calidad del Vino")
plt.show()
sns.boxplot(x=df['calidad'], y=df['azucar residual'], palette="Purples")
plt.title("Distribución del Azúcar Residual según la Calidad del Vino")
plt.show()
sns.boxplot(x=df['calidad'], y=df['pH'], palette="Blues")
plt.title("Distribución del pH según la Calidad del Vino")
plt.show()

"""**Identificación de patrones y correlaciones**"""
import seaborn as sns
import matplotlib.pyplot as plt
# Selecciono solo variables numéricas y eliminar 'botella_id'
df_numeric = df.select_dtypes(include=['float64', 'int64']).drop(columns=['botella_id'],
errors='ignore')
# Matriz de correlación
plt.figure(figsize=(10, 6))
sns.heatmap(df_numeric.corr(), annot=True, cmap='coolwarm', fmt=".2f",
linewidths=0.5)
plt.title("Matriz de Correlación entre Variables")
plt.show()

"""**PROCESAMIENTO DE LOS DATOS**"""
# Verificar datos y estructura del dataset
print("Primeras filas del dataset:")
display(df.head())

print("\nInformación del dataset:")
display(df.info())

# ELIMINACIÓN DE VARIABLES IRRELEVANTES
df.drop(columns=['botella_id', 'color'], inplace=True, errors='ignore')  # Eliminar
columnas innecesarias

# SEPARACIÓN ENTRE VARIABLES INDEPENDIENTES (X) Y VARIABLE
OBJETIVO (y)
X = df.drop(columns=['calidad'])  # Variables a predecir
y = df['calidad']  # Variable objetivo

# Verificar tipos de datos
print("Tipos de datos en cada columna:")
print(df.dtypes)

# Verificar valores únicos en cada columna (para detectar problemas de formato)
for col in df.columns:
    print(f"\nValores únicos en {col}:")
    print(df[col].unique()[:10])  # Muestra solo los primeros 10 valores únicos

# Convertir la columna "alcohol" a tipo numérico
```

```
df['alcohol'] = pd.to_numeric(df['alcohol'], errors='coerce')

# Verificar si hay valores nulos después de la conversión
print(df.isnull().sum())  # Para ver si hay valores NaN
df.dropna(inplace=True)
print(df.isnull().sum())  # Para ver si hay valores NaN

# Importar librerías necesarias
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

# ELIMINACIÓN DE VARIABLES IRRELEVANTES
df.drop(columns=['botella_id', 'color'], inplace=True, errors='ignore')

# CONVERTIR TODAS LAS COLUMNAS A NUMÉRICO
for col in df.columns:
    df[col] = pd.to_numeric(df[col], errors='coerce')  # Convierte a numérico, si no puede
pone NaN

# DETECTAR Y ELIMINAR FILAS CON VALORES ERRÓNEOS
df.dropna(inplace=True)  # Eliminamos filas con valores no convertibles

# SEPARACIÓN ENTRE VARIABLES INDEPENDIENTES (X) Y VARIABLE
OBJETIVO (y)
X = df.drop(columns=['calidad'])  # Variables predictoras
y = df['calidad']  # Variable objetivo

# NORMALIZACIÓN DE VARIABLES NUMÉRICAS
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# DIVISIÓN DEL CONJUNTO DE DATOS EN TRAIN Y TEST (80% - 20%)
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2,
random_state=42, stratify=y)

# Verificar la forma de los conjuntos de datos
print("\nTamaño del conjunto de entrenamiento:", X_train.shape, y_train.shape)
print("Tamaño del conjunto de prueba:", X_test.shape, y_test.shape)

from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
import matplotlib.pyplot as plt
import seaborn as sns

# 1. REGRESIÓN LOGÍSTICA
log_reg = LogisticRegression(max_iter=1000)
```

```
log_reg.fit(X_train, y_train)
y_pred_log = log_reg.predict(X_test)
acc_log = accuracy_score(y_test, y_pred_log)

# 2. RANDOM FOREST
rf = RandomForestClassifier(n_estimators=100, random_state=42)
rf.fit(X_train, y_train)
y_pred_rf = rf.predict(X_test)
acc_rf = accuracy_score(y_test, y_pred_rf)

# 3. KNN (K-Vecinos más cercanos)
knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(X_train, y_train)
y_pred_knn = knn.predict(X_test)
acc_knn = accuracy_score(y_test, y_pred_knn)

# MOSTRAR RESULTADOS
print("Comparación de Modelos:")
print(f"Regresión Logística - Precisión: {acc_log:.4f}")
print(f"Random Forest - Precisión: {acc_rf:.4f}")
print(f"KNN - Precisión: {acc_knn:.4f}")

# VISUALIZACIÓN DE RESULTADOS
modelos = ['Regresión Logística', 'Random Forest', 'KNN']
precisiones = [acc_log, acc_rf, acc_knn]

plt.figure(figsize=(8,5))
sns.barplot(x=modelos, y=precisiones, palette='coolwarm')
plt.ylabel("Precisión del modelo")
plt.title("Comparación de modelos de predicción de calidad del vino")
plt.ylim(0, 1)
plt.show()

# INFORMES DETALLADOS
print("\Informe de clasificación - Regresión Logística")
print(classification_report(y_test, y_pred_log))

print("\ Informe de clasificación - Random Forest")
print(classification_report(y_test, y_pred_rf))

print("\ Informe de clasificación - KNN")
print(classification_report(y_test, y_pred_knn))

from imblearn.over_sampling import SMOTE
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, accuracy_score

# Aplicamos SMOTE SOLO en el conjunto de entrenamiento
smote = SMOTE(sampling_strategy='auto', k_neighbors=2, random_state=42)
```

```
X_train_resampled, y_train_resampled = smote.fit_resample(X_train, y_train)

# Verificamos la nueva distribución de clases después de SMOTE
import pandas as pd
print(pd.Series(y_train_resampled).value_counts())

# Entrenamos de nuevo Random Forest con los datos balanceados
rf_smote = RandomForestClassifier(n_estimators=150, max_depth=30,
min_samples_split=2,
                                  min_samples_leaf=1, max_features='sqrt', random_state=42)
rf_smote.fit(X_train_resampled, y_train_resampled)

# Predicción en el conjunto de prueba
y_pred_rf_smote = rf_smote.predict(X_test)

# Evaluación del modelo
print("\n **Informe de clasificación – Random Forest con SMOTE** \n")
print(classification_report(y_test, y_pred_rf_smote))

# Precisión global
accuracy_smote = accuracy_score(y_test, y_pred_rf_smote)
print(f"\n**Precisión con SMOTE: {accuracy_smote:.4f}**")

# Ajustar las etiquetas para que comiencen en 0
y_train_adjusted = y_train - y_train.min()
y_test_adjusted = y_test - y_test.min()

# Entrenar XGBoost con las clases corregidas
xgb = XGBClassifier(random_state=42)
xgb.fit(X_train, y_train_adjusted)

# Hacer predicciones y revertir la transformación
y_pred_xgb = xgb.predict(X_test) + y_train.min()

# Evaluar precisión
accuracy_xgb = accuracy_score(y_test, y_pred_xgb)
print(f"Precisión con XGBoost: {accuracy_xgb:.4f}")

from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import GridSearchCV

# Definir el modelo base
rf = RandomForestClassifier(random_state=42)

# Definir la rejilla de hiperparámetros
param_grid = {
    'n_estimators': [100, 200, 300],
    'max_depth': [10, 20, 30],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
```

```
    'class_weight': ['balanced', 'balanced_subsample']
}

# Aplicamos GridSearchCV con validación cruzada
grid_search = GridSearchCV(
    estimator=rf,
    param_grid=param_grid,
    scoring='recall_weighted',  # Optimizamos el recall general
    cv=5,
    n_jobs=-1,
    verbose=2
)

# Entrenamos el modelo con la búsqueda de hiperparámetros
grid_search.fit(X_train, y_train)

# Ver los mejores hiperparámetros encontrados
print("🔍 Mejores parámetros encontrados:", grid_search.best_params_)

# Evaluamos el modelo con los mejores parámetros
best_rf = grid_search.best_estimator_
y_pred_rf = best_rf.predict(X_test)

# Mostramos métricas actualizadas
from sklearn.metrics import classification_report

print("\n**Informe de clasificación – Random Forest Optimizado**")
print(classification_report(y_test, y_pred_rf))

# Importancia de las Variables en el Modelo Random Forest
importances = best_rf.feature_importances_
features = df.drop(columns=['calidad']).columns  # Recuperamos nombres de las
características

# Crear un dataframe con las importancias
feature_importance_df = pd.DataFrame({'Característica': features, 'Importancia':
importances})
feature_importance_df = feature_importance_df.sort_values(by='Importancia',
ascending=False)

# Graficar la importancia de las variables
plt.figure(figsize=(10, 6))
sns.barplot(x='Importancia', y='Característica', data=feature_importance_df,
palette='coolwarm')
plt.xlabel('Importancia en la Predicción')
plt.ylabel('Características Químicas')
plt.title('Importancia de las Características en la Calidad del Vino (Random Forest)')
plt.show()

# Calcular correlaciones con la variable objetivo (calidad)
```

```
correlaciones = df.corr()['calidad'].sort_values(ascending=False)
# Mostrar los resultados
plt.figure(figsize=(8, 6))
sns.barplot(x=correlaciones.values, y=correlaciones.index, palette='coolwarm')
plt.xlabel('Correlación con Calidad')
plt.ylabel('Características Químicas')
plt.title('Correlación entre Características Químicas y Calidad del Vino')
plt.show()
```

Código modelo de recomendación:

```
# -*- coding: utf-8 -*-
"""TFGBA2ºparte.ipynb
Automatically generated by Colab.
Original file is located at
    https://colab.research.google.com/drive/1c0NRMxGHCQw_4e9wzGxjHvi8ig0hbiz-
"""

from google.colab import drive
drive.mount('/content/drive')

import pandas as pd

# Ruta del archivo en Google Drive
file_path = "/content/drive/My Drive/Recomendacion_vino_tfg.csv"

# Intentar leer con diferentes codificaciones
try:
    df = pd.read_csv(file_path, delimiter=';', encoding='latin-1')  # Intento con Latin-1
except:
    try:
        df = pd.read_csv(file_path, delimiter=';', encoding='ISO-8859-1')  # Intento con
ISO-8859-1
    except:
        df = pd.read_csv(file_path, delimiter=';', encoding='windows-1252')  # Intento con
Windows-1252

# Mostrar las primeras filas
df.head()

# Ver tipos de datos
print(df.dtypes)

# Ver valores únicos en cada columna (para entender las categorías)
for column in df.columns:
    print(f"Valores únicos en {column}:")
    print(df[column].unique())
    print("\n")

#6. TRANSFORMAR LA VARIABLE "use"
```

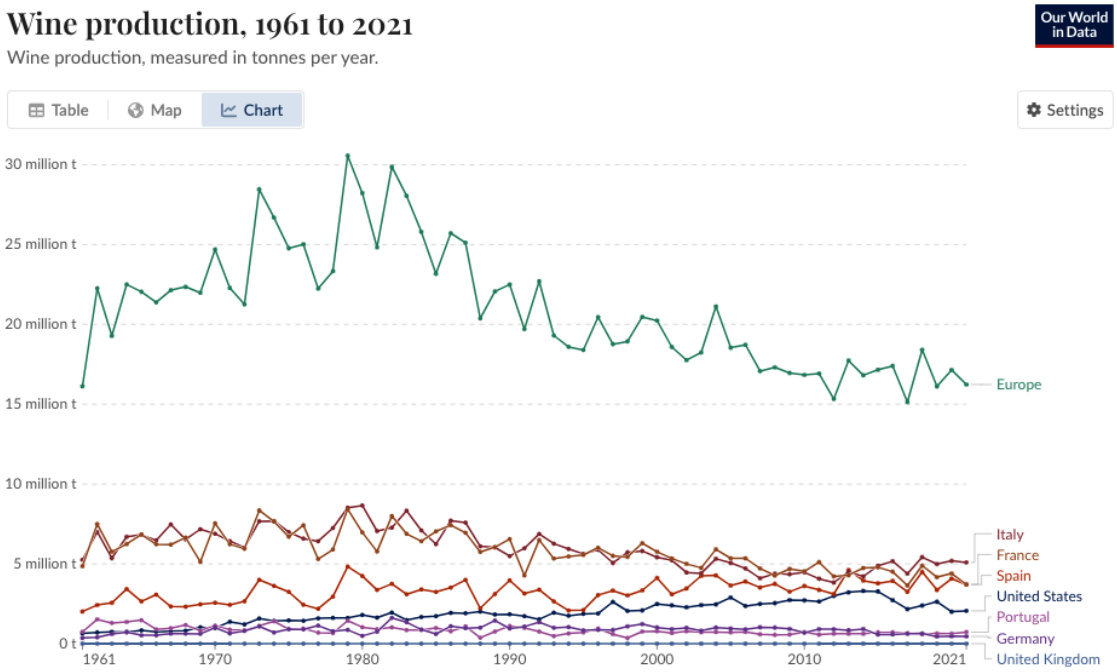
```
df['use'] = df['use'].apply(lambda x: x if x in ['Table', 'Dessert', 'Appetizer'] else 'all uses')

# MOSTRAR LOS NUEVOS VALORES ÚNICOS PARA VERIFICACIÓN
print("Valores únicos en 'use' después de la transformación:")
print(df['use'].unique())

# Mostrar las primeras filas para verificar cambios
df.head()

import numpy as np

# 1️⃣ Función para convertir 'degree' en valores numéricos
def convert_degree(value):
    if isinstance(value, str):
        value = value.replace(" ", "")  # Eliminar espacios en valores como "10 ~ 12"
        if "~" in value:  # Si es un rango, promediarlo
            parts = value.split("~")
            return (float(parts[0]) + float(parts[1])) / 2
        elif value.replace('.', '', 1).isdigit():  # Si es un número, convertirlo
            return float(value)
    return np.nan  # Si no es numérico, devolver NaN

# Aplicar la transformación a la columna
df['degree'] = df['degree'].apply(convert_degree)

# 2️⃣ Crear la escala de alcohol (1 a 5)
def categorize_degree(value):
    if np.isnan(value):
        return np.nan  # Dejar como NaN si no hay datos
    elif value < 10:
        return 1
    elif 10 <= value < 12:
        return 2
    elif 12 <= value < 14:
        return 3
    elif 14 <= value < 16:
        return 4
    else:
        return 5

df['degree_category'] = df['degree'].apply(categorize_degree)

# 3️⃣ Verificar la transformación
print("Valores únicos en degree_category después de la conversión:")
print(df['degree_category'].unique())

# Mostrar algunas filas transformadas
df[['degree', 'degree_category']].head(10)
```

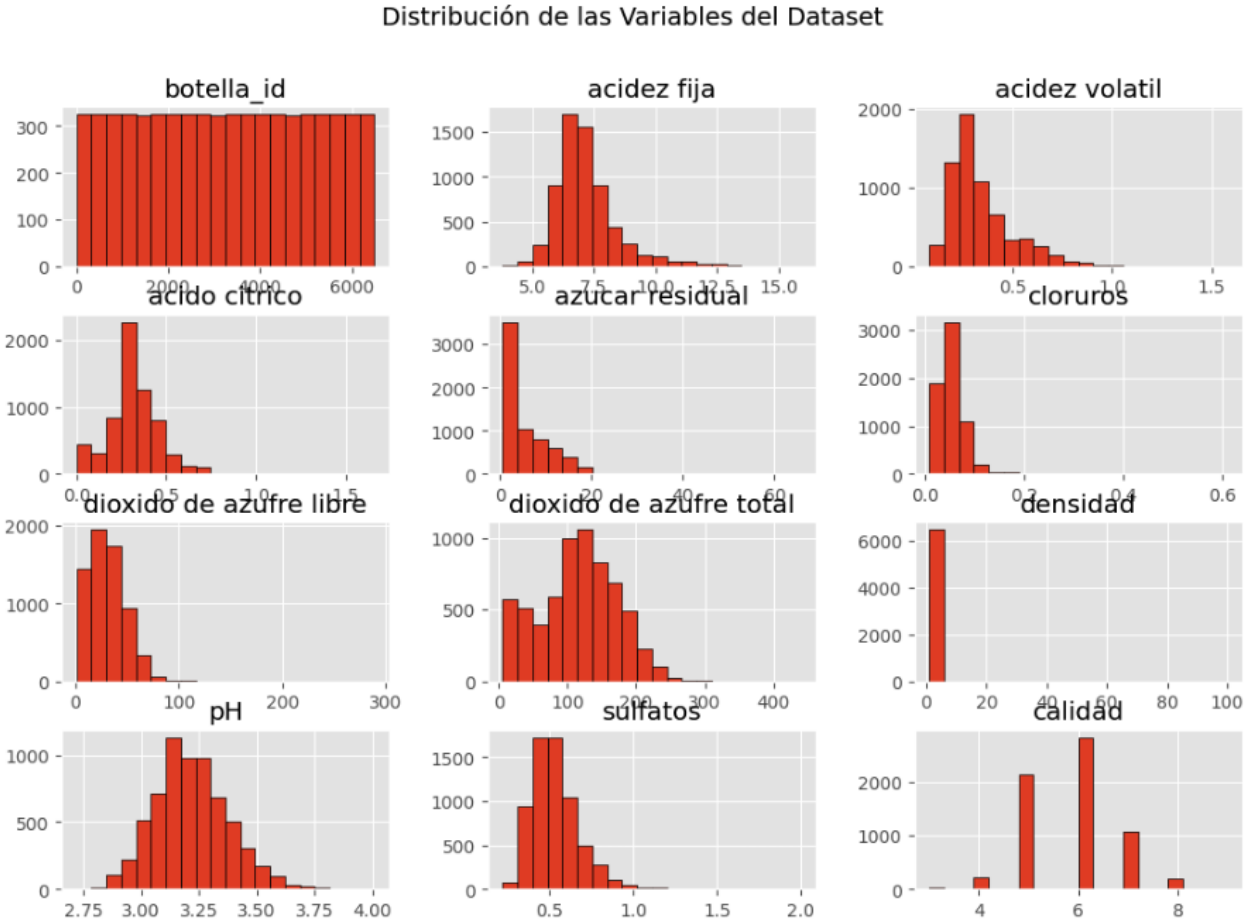
```
#  Diccionarios de mapeo para las variables
mapping_sweet = {'SWEET1': 1, 'SWEET2': 2, 'SWEET3': 3, 'SWEET4': 4, 'SWEET5':
5}
mapping_acidity = {'ACIDITY1': 1, 'ACIDITY2': 2, 'ACIDITY3': 3, 'ACIDITY4': 4,
'ACIDITY5': 5}
mapping_body = {'BODY1': 1, 'BODY2': 2, 'BODY3': 3, 'BODY4': 4, 'BODY5': 5}
mapping_tannin = {'TANNIN1': 1, 'TANNIN2': 2, 'TANNIN3': 3, 'TANNIN4': 4,
'TANNIN5': 5}

# 2. Aplicar la conversión en cada columna
df['sweet'] = df['sweet'].map(mapping_sweet)
df['acidity'] = df['acidity'].map(mapping_acidity)
df['body'] = df['body'].map(mapping_body)
df['tannin'] = df['tannin'].map(mapping_tannin)

# 3. Mostrar resultados
print("Valores únicos después de la transformación:")
print("Sweet:", df['sweet'].unique())
print("Acidity:", df['acidity'].unique())
print("Body:", df['body'].unique())
print("Tannin:", df['tannin'].unique())

# Mostrar algunas filas transformadas
df[['sweet', 'acidity', 'body', 'tannin']].head(10)

# Crear una nueva columna 'variety_final' con la mejor opción disponible
df['variety_final'] = df[['varieties1', 'varieties2', 'varieties3',
'varieties4']].bfill(axis=1).iloc[:, 0]

# Mostrar los valores únicos después de la transformación
print("Valores únicos en 'variety_final':", df['variety_final'].unique())

# Mostrar algunas filas para verificar
df[['varieties1', 'varieties2', 'varieties3', 'varieties4', 'variety_final']].head(10)

df.head()

# Seleccionar solo las columnas relevantes
df_modelo = df[['name', 'variety_final', 'type', 'use', 'degree_category', 'sweet', 'acidity',
'body', 'tannin']]

# Mostrar las primeras filas para verificar
df_modelo.head()
df.describe()
import pandas as pd
import numpy as np
from sklearn.preprocessing import MinMaxScaler
from sklearn.neighbors import NearestNeighbors
import matplotlib.pyplot as plt
```

```
# Variables relevantes para el modelo
features = ['sweet', 'acidity', 'body']
df_features = df[features].copy()

# Verificar si hay valores nulos
df_features.isnull().sum()

df_features.fillna(df_features.mean(), inplace=True)

# Aplicamos MinMaxScaler para escalar los datos entre 0 y 1
scaler = MinMaxScaler()
df_scaled = pd.DataFrame(scaler.fit_transform(df_features), columns=features)

# Verificar la normalización
df_scaled.describe()

# Probar distintos valores de k para encontrar el mejor
error_rates = []
k_values = range(1, 21)

for k in k_values:
    knn = NearestNeighbors(n_neighbors=k)
    knn.fit(df_scaled)
    distances, _ = knn.kneighbors(df_scaled)
    error_rates.append(np.mean(distances[:, -1]))

# Graficar error vs. k
plt.figure(figsize=(8, 5))
plt.plot(k_values, error_rates, marker='o', linestyle='dashed', color='blue')
plt.xlabel("Número de vecinos (k)")
plt.ylabel("Error promedio")
plt.title("Selección del mejor k para KNN")
plt.show()

best_k = 7  # Escoge el mejor valor de la gráfica

# Entrenar el modelo con el mejor valor de k
knn = NearestNeighbors(n_neighbors=best_k)
knn.fit(df_scaled)

def recomendar_vinos(sweet, acidity, body, df_original, df_scaled, knn_model, scaler):
    # Convertir inputs en array y normalizar
    user_input = np.array([[sweet, acidity, body]])
    user_input_scaled = scaler.transform(user_input)

    # Encontrar los k vecinos más cercanos
    distances, indices = knn_model.kneighbors(user_input_scaled)

    # Obtener los nombres de los vinos recomendados
    recomendados = df_original.iloc[indices[0]][['name', 'variety_final', 'use']]
```

```
    return recomendados

# Prueba con preferencias del usuario (ejemplo)
recomendaciones = recomendar_vinos(2, 2, 5, df, df_scaled, knn, scaler)
recomendaciones

def evaluar_modelo(df_original, df_scaled, knn_model, scaler):
    # Tomamos muestras de usuarios aleatorios y verificamos recomendaciones
    ejemplos = df_original.sample(5)[['sweet', 'acidity', 'body']]

    for _, row in ejemplos.iterrows():
        print(f"\nPreferencias del usuario: Dulzura={row['sweet']},
Acidez={row['acidity']}, Cuerpo={row['body']}")
        print(recomendar_vinos(row['sweet'], row['acidity'], row['body'], df_original,
df_scaled, knn_model, scaler))

# Ejecutar la evaluación
evaluar_modelo(df, df_scaled, knn, scaler)
```

Anexo 2. Visualizaciones detalladas





Wine consumption per person, 1960 to 2019

Average per capita consumption of wine, as measured in liters of pure alcohol per year. 1 liter of wine contains around 0.12 liters of pure alcohol.