MEMORIA PROYECTO
“WineMeApp!”

1. Introducción

Este proyecto tiene como objetivo desarrollar un motor de recomendación de vinos que aprovecha las notas de cata y maridaje. La verdadera innovación radica en la capacidad del cliente para personalizar las recomendaciones, permitiéndole ajustar la importancia relativa del precio y el rating de los vinos. Este enfoque único garantiza una experiencia de selección de vinos completamente adaptada a las preferencias individuales del usuario.

Además, al basar la selección del vino en las notas de cata y maridaje, aseguramos que el resultado de la recomendación no se limite a un solo tipo de vino. Por ejemplo, al solicitar una recomendación de un rosado, podría incluir tanto tintos suaves como blancos o espumosos. Este aspecto es especialmente interesante ya que evita que nos dejemos llevar por prejuicios o ideas preconcebidas sobre ciertos tipos de vinos, ofreciendo así resultados verdaderamente sorprendentes para el usuario o cliente.

1.1. Motivación

¿Quién no se ha visto en la situación, cuando pregunta el camarero por las bebidas, en una cena o comida, de no saber que pedir? Empezamos a buscar opciones en nuestra memoria, y finalmente desistimos pidiendo lo mismo de siempre. Como mucho, si nos vemos creativos y aventureros, nos conformamos con lo que pidió la persona de al lado. Por otro lado, algunos de los integrantes de este proyecto, cuando vivíamos como ‘expats’ en Reino Unido, observamos que muchas mujeres solían optar por vinos blancos de variedad *Chardonnay*. Nosotros, que somos originarios de zonas vinícolas tan reconocidas como Jerez de Frontera, con sus variedades de uva como el Palomino, Pedro Ximénez y Tintilla de rota, o de Zamora con vinos de Toro, y variedades como la Malvasía Castellana, Moscatel de Grano Menudo y Verdejo, nos sorprendía esta preferencia, ya que el abanico de opciones, tanto españolas como de otras regiones, era enorme.

Es en estos momentos cuando un buen sistema de recomendación de vinos, ya sea en la sección de vinos de grandes almacenes o supermercados gourmet, o bien cuando vas a pedir las bebidas en un restaurante, sería de gran utilidad. ¡Y es precisamente aquí donde nuestro recomendador “**Wine Me App!**" entra en juego!

2. Objetivos

2.1. Principal

El proyecto busca dar respuesta a la pregunta anteriormente desarrollada a través de un modelo híbrido de machine learning. Lo que produce este modelo es una selección de **vinos recomendados en base a uno que introduce el usuario/cliente**, teniendo en cuenta sus notas de cata y maridaje. Primero, el usuario verifica si el vino se encuentra en la base de datos, y luego introducirá el nombre del vino para conseguir la selección de vinos. Por último, se le pregunta al usuario el rango de precio

que desea gastarse, y como de importante es para él/ella el precio y el rating del vino (vasado en evaluaciones de los vinos por expertos y usuarios).

2.2. Secundarios

- **Visualización de las agrupaciones de los vinos** en función a las características de cata y maridaje en **2 dimensiones**. Es decir, el propósito de esto es ver cómo funciona el algoritmo de recomendación de forma visual cuando nosotros ejecutamos el modelo. Vinos con características similares de cata y maridaje se situarán cerca en el plot de 2 dimensiones, y aquellos con menos **similaridad** estarán más alejados.
- Visualización de una **nube de palabras** que represente las **notas de cata y maridaje** de los resultados de recomendación para un vino.
- Generación de un **plot visualizando los resultado del modelo TOPSIS**. Esto ayudara a elegir al usuario el mejor vino en base a sus gustos en cuanto a cata, maridaje y también en base a precio y ratings de los vinos. Lo que es importante y que añade el modelo TOPSIS al modelo de recomendación es que un alto porcentaje de decisión está en el usuario/humano. El modelo de recomendación es más bien una caja negra, que se basa en notas de cata y maridaje, pero no podemos elegir específicamente que notas de cata son más importantes para nosotros.
- Creación de una **web app** donde incluir el motor de recomendación de vinos incluyendo el sistema basado en notas de cata y maridaje y el sistema de pesos (modelo TOPSIS) junto con una descripción del proyecto, recursos e información de los creadores
- Inclusión de un **modelo extra**, que incluya **vinos internacionale**s basados en notas de cata como opción final para el cliente utilizando un dataset de Kaggle: https://www.kaggle.com/datasets/zynicide/wine-reviews
- El último objetivo, que engloba todo el proyecto es el **fomento de la cultura vinícola de vinos nacionales** y la puesta en valor de los mismos.

3. Datos

Los datos fueron obtenidos mediante *web scraping*, utilizando la librería `*BeautifulSoup*`, de la página web española www.bodeboca.com, recopilando información de 4792 vinos españoles. El conjunto de datos incluye las siguientes características: título o nombre del vino, link de la web de compra del vino, precio en euros, rating, volumen de la botella, bodega, tipo de vino con 33 categorías, grado alcohólico, añada, producción de botellas, subzona, variedad de uva que lleva el vino en %, origen, vista (características visuales del vino), nariz (características organolépticas), boca (características gustativas del vino), temperatura recomendada de servicio del vino, maridaje, nombre del viñedo, descripción, edad del viñedo, ,clima, suelo, rendimiento, cosecha, vinificación, envejecimiento y embotellado.

3.1. Preprocesamiento y limpieza

Una de las grandes labores de este proyecto ha sido la extracción, limpieza y preprocesamiento de datos. Debido a la falta de un dataset base, desarrollamos un código para extraer información de 33 vinos de cada una de las 170 páginas que contienen vinos españoles en bodeboca.com. Posteriormente, seleccionamos los datos relevantes para nuestro análisis. Dado que encontramos una cantidad significativa de datos faltantes, optamos por revisar alrededor de 800 vinos con la mayor cantidad de datos faltantes con énfasis en notas de cata y maridaje. Completamos estos datos verificando en otras webs similares como vinissimus.com, decantalo.com, vivino.es, entre otras. En el preprocesamiento hemos seguido las siguientes pautas: seleccionamos las columnas de

interés, convertido todo a minúsculas, eliminado acentos y símbolos raros, eliminado duplicados, las columnas numéricas las hemos revisado y cambiado comas por puntos, el grado alcohólico y en temperatura de servicio nos hemos quedado solo con los números, de las variedades de uva nos hemos quedado solo con el texto de las variedades, eliminando los símbolos de porcentaje. Por último, para la gestión de los datos faltantes 'NAs' hemos agrupado los vinos por tipo de vino como describía la página web (33 categorías) y hemos rellenado con la 'moda' de cada grupo. Consideramos que este enfoque era el más adecuado para nuestros datos. Finalmente, combinamos la información de las notas de cata (características visuales, olfativas y gustativas) junto con el maridaje de cada vino en una columna llamada 'descripcion'. Esta columna se preparó para aplicar Procesamiento del Lenguaje Natural (NLP), que será la columna utilizada en el modelo de recomendación. En esta fase de normalización pasamos a minúsculas, *tokenizamos*, eliminamos *stopwords* y *lematizamos* el texto de 'descripción'.

4. Análisis exploratorio de los datos (EDA)

Se realiza un análisis exploratorio en el que graficamos y evaluamos todas las variables numéricas y no numéricas de interés de los vinos obtenidos de bodeboca.com. A continuación, podemos ver los principales gráficos. La mayor parte de los vinos se encuentran entre 0 y 50 euros (Figs. 1), con una media de 33.80 euros. Los precios mínimo y máximo son 4,45 y 2800 euros. Las valoraciones o ratings de los vinos se centran en torno a 4.0 en puntuación y el grado alcohólico en torno a 12°.

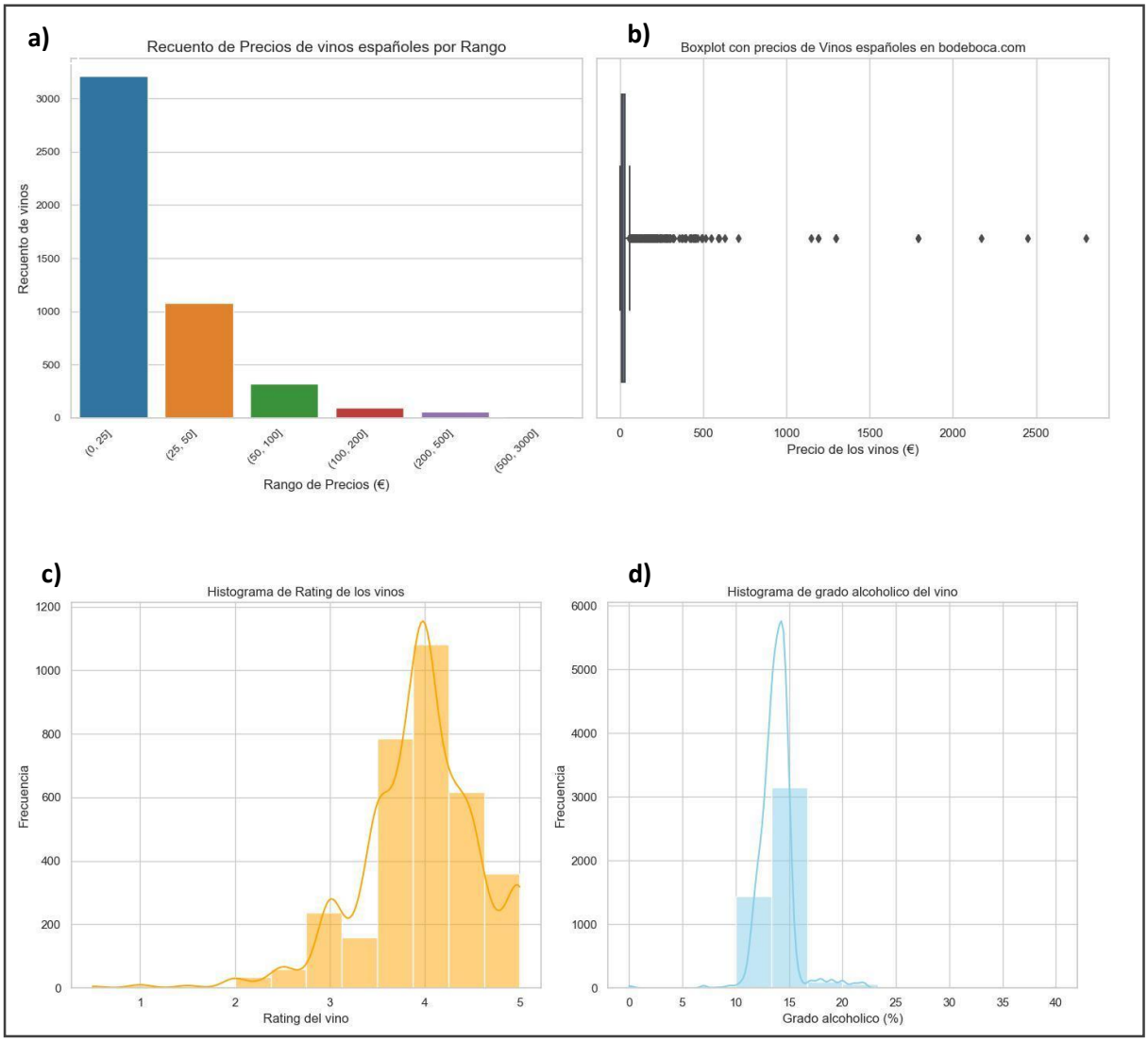

Fig. 1 Distribuciones de precios (a y b), ratings de vinos (c) y grado alcohólico (d) de los 4792 vinos.

Los tipos de vinos (Figs. 2) obtenidos de bodeboca proporcionaban 33 categorías de tipo de vino ('tinto', 'red vermouth', 'blanco', 'espumoso', 'amontillado', 'oloroso', 'tinto reserva', 'blanco fermentado en barrica', 'white vermouth', 'manzanilla', 'dulce px', 'palo cortado', 'palo cortado vors', 'fino', 'rosado', 'otro(s)', 'tinto joven', 'tinto crianza', 'amontillado vors', 'oloroso vors', 'aromatised wine', 'blanco naturalmente dulce', 'tinto dulce', 'blanco dulce', 'orange wine', 'tinto gran reserva', 'cava', 'sweet moscatel', 'oloroso dulce', 'dulce px vors', 'frizzante', 'dulce', 'rancio', 'rueda dorado'). Con esto, creamos una nueva columna `tipo2` donde reagrupamos los vinos en grupos mas generales, y pasamos de 33 categorias a 6: tinto, blanco, generoso, espumoso, rosado y vermouth. La categoria de vinos generosos es una amalgama de vinos dulces, manzanilla, fino, olorosos, aromatizados, palo cortado, vinos rancios…etc. Aquí priorizamos los vinos con alta graduacion alcoholica, basandonos en la recomendación de una experta enologa (Marta Benito Reyes). Los espumosos engloban los vinos frizzante, espumosos y cavas.

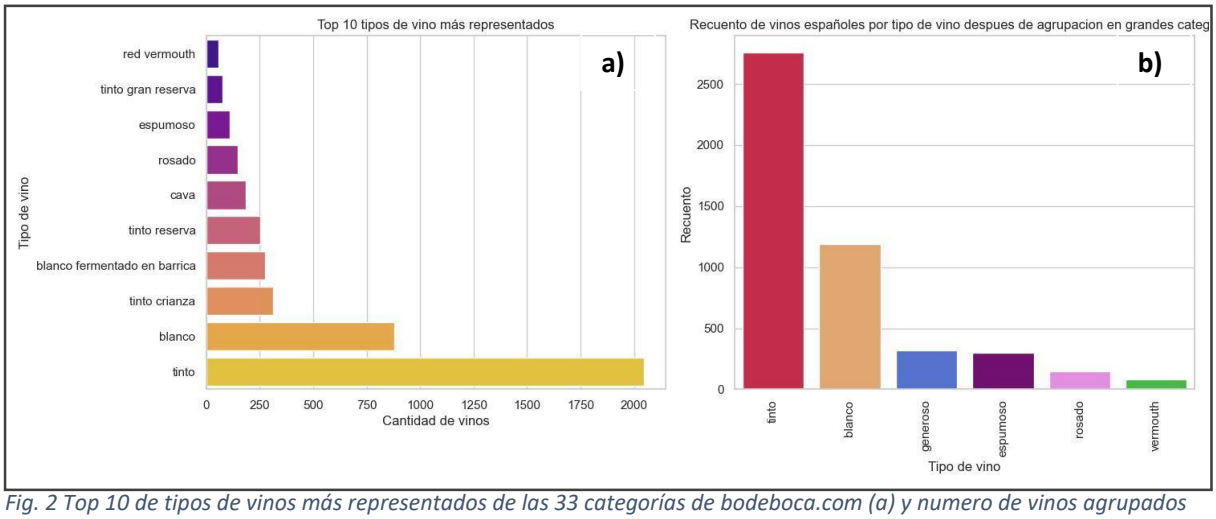

Fig. 2 Top 10 de tipos de vinos más representados de las 33 categorías de bodeboca.com (a) y numero de vinos agrupados en seis categorías generales (b)

Las columnas con información sobre bodegas y las denominaciones de origen de los vinos son categóricas con un alto número de variables únicas. En el caso de las bodegas, nuestro dataset tiene 896 bodegas, donde las bodegas Williams-Humbert y Landau de Jerez, familia Torres (Rias Baixas), y Bodegas Bilbainas (Rioja) son las que poseen el mayor número de vinos en nuestro dataset (Fig. 3a). Los vinos con mayor representación en nuestros datos son los de Rioja y Ribera de Duero (Fig. 3b).

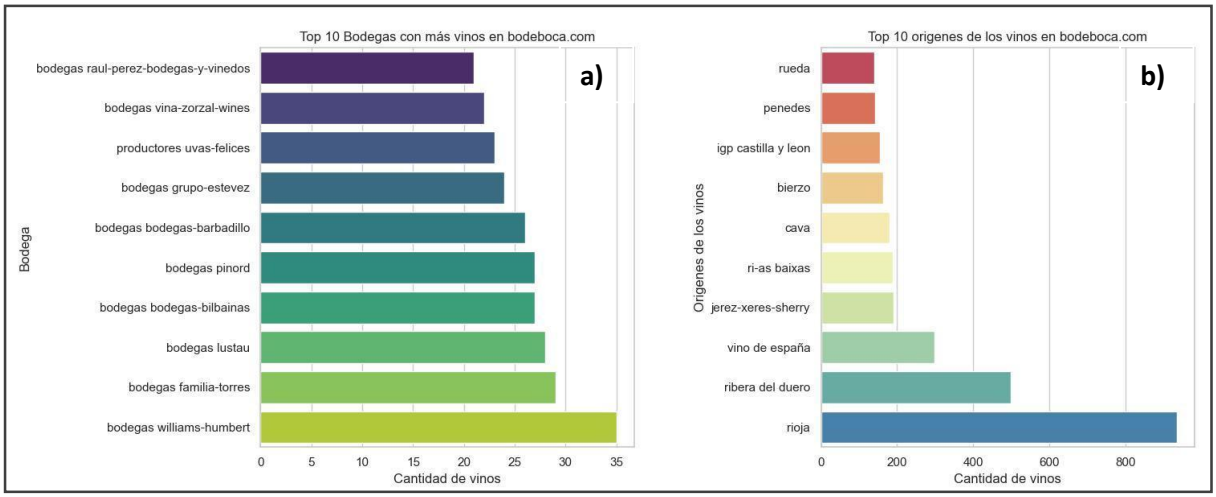

Fig. 3 Bodegas (a) y denominaciones de origen (b) con más vinos del dataset.

Finalmente, la columna con información sobre variedades de uva contiene información compleja ya que muestra la variedad o variedades de uva del vino en porcentajes. Para mayor funcionalidad en el análisis, hemos eliminado tanto los números como los porcentajes de la columna, para evaluar solo el texto de la columna. En la figura 4 podemos ver las variedades puras, es decir, vinos con un 100% de esa variedad de uva, con mayor representación en nuestros datos. Como vemos, los vinos con variedades 100% tempranillo, garnacha, palomino fino y albarino son las más representadas.

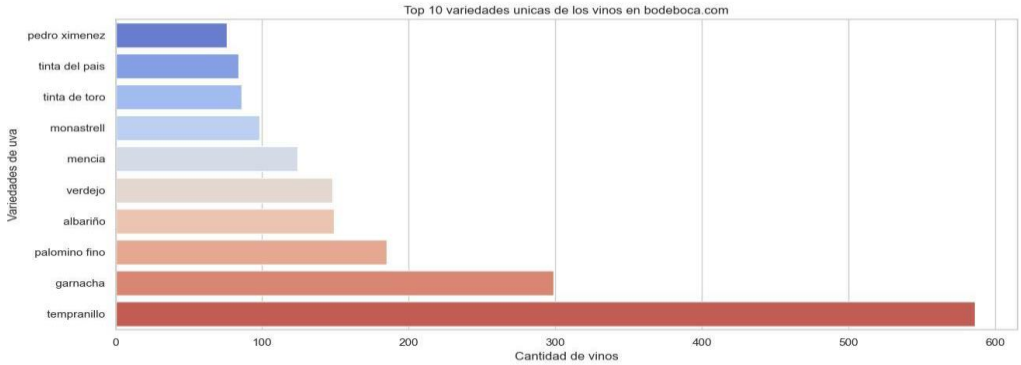

Fig. 4 Variedades de uva con mayor representación en el dataset de bodeboca.com

5. Modelos de recomendación y Web App en Streamlit

El modelo de recomendación creado en este proyecto es un **modelo híbrido** que combina dos enfoques poderosos. En primer lugar, implementamos un sistema de recomendación basado en **notas de cata y maridaje**, Procesamiento del Lenguaje Natural –NLP- al que nos referiremos como el **modelo NLP**. Para brindar aún más libertad de elección al usuario dentro de las recomendaciones, implementamos un modelo de *Technique for Order of Preference by Similarity to Ideal Solution-TOPSIS-*, que definimos como **modelo TOPSIS** para ponderar aspectos como el precio y el rating. Esto permite ajustar las recomendaciones según las preferencias individuales, mejorando así tanto la precisión como la relevancia de las sugerencias. Finalmente, integramos este modelo completo en una **aplicación web**: **WineMeApp!**, diseñada para mejorar la interfaz y la experiencia interactiva del usuario. En ella, el usuario puede no sólo ver el resultado de los modelos, sino interactuar con ellos.

5.1. Modelo de recomendación basado en notas de cata y maridaje

El **modelo NLP** se basa en el análisis de texto de notas de cata y maridaje, unificadas en la columna 'descripción' y procesadas mediante NLP. Después de la normalización realizada en el preprocesamiento de esta columna, realizamos la vectorización utilizando el *Google Universal Sentence Encoder versión 4*, disponible en *TensorFlow Hub*. Este modelo pre-entrenado, utiliza técnicas avanzadas de NLP y redes neuronales convirtiendo las descripciones de vinos en vectores de 512 dimensiones, capturando su significado semántico de manera eficiente. Este enfoque de vectorización es similar al CountVectorizer de sklearn y permite al modelo generar recomendaciones precisas utilizando características relevantes de cata y maridaje. De esta manera, almacenamos la información de los vectores en una matriz llamada '*embeddings*', la cual normalizamos, que representa un espacio de dimensiones relativamente bajas al que se pueden trasladar vectores de altas dimensiones. A continuación, una vez transformadas todas las notas de cata de cada vino, en vectores, podemos utilizar la similitud del coseno para obtener aquellos vectores más cercanos a uno dado usando la siguiente fórmula donde A y B serían los vinos a evaluar:

En este caso, nosotros lo transformaremos en una puntuación siguiendo las indicaciones para similitud textual del benchmark STS (Semantic Textual Similarity). En el contexto del benchmark STS, los puntajes de similitud del coseno los hemos medido en un rango de 0 a 1, donde:

$$Similitud\ (A,B) = \frac{A\ \cdot B}{||A||\cdot\ ||B||}$$

- 0: Indica ninguna similitud entre los dos vectores de texto. Esto significa que los vectores (vinos) son completamente diferentes y no comparten ninguna similitud semántica.
- 1: Indica una similitud perfecta entre los dos vectores de texto. Esto significa que los vectores (vinos) son idénticos o muy similares en términos de contenido semántico.

En resumen, los puntajes de similitud del coseno en el benchmark STS proporcionan una medida de la similitud semántica entre dos textos, en nuestro caso las notas de cata y maridaje del vino A y las mismas del vino B, donde un puntaje más alto indica una mayor similitud y un puntaje más bajo indica una menor similitud. Adicionalmente, hemos podido visualizar en 2 y 3 dimensiones, reduciendo de 512 a 2 y 3 dimensiones, cada vino en función de cata y maridaje de manera interactiva creando dos graficos con *plotly* usando un algoritmo de T-SNE (t-Distributed Stochastic Neighbor Embedding). Estas graficas dan una idea visual de lo que hace el algoritmo por detrás cuando le metemos un vino del dataset como consulta para buscar vinos recomendados en base a otros similares en notas de cata y maridaje. Finalmente, La función de recomendación de vinos basado en este algoritmo/modelo NLP introducirá como input el nombre de un vino de la base de datos, y dará como resultado 10 vinos similares semánticamente en base a las notas de cata y maridaje.

5.2. Modelo TOPSIS

El algoritmo TOPSIS es un método de toma de decisiones utilizado para evaluar la mejor opción entre un conjunto de alternativas basándose en múltiples criterios. Aquí tienes un resumen de cómo funciona:

- **1_Definición de criterios y alternativas:** En primer lugar, se identifican los criterios relevantes que se utilizarán para evaluar las alternativas. También se determinan las alternativas que se van a comparar.
- **2_Normalización de los datos**: Los valores de los criterios para cada alternativa se normalizan para que estén en una escala común y comparables. Esto puede implicar escalar los valores entre 0 y 1, o utilizar otros métodos de normalización como la estandarización.
- **3_Construcción de la matriz de decisión:** Se construye una matriz de decisión donde las filas representan las alternativas y las columnas representan los criterios normalizados. Con los datos normalizados, aplica pesos a cada atributo según los valores proporcionados y crea un DataFrame con características ponderadas. Los pesos van a poder personalizarse de manera que un factor u otro tenga mayor importancia o relevancia en la toma de decisiones (beneficio vs esfuerzo, precio vs ranking).
- **4_Identificación de las soluciones ideal y anti-ideal**: Se determinan dos soluciones de referencia: la solución ideal, que maximiza cada criterio, y la solución anti-ideal, que minimiza cada criterio.
- **5_Cálculo de las distancias:** Se calcula la distancia euclidiana entre cada alternativa y las soluciones ideal y anti-ideal. Esta distancia mide la proximidad de cada alternativa a cada una de estas soluciones de referencia.
- **6_Cálculo del índice de similitud a la solución ideal:** Se calcula el índice de similitud para cada alternativa dividiendo la distancia a la solución anti-ideal entre la suma de la distancia a la solución ideal y la distancia a la solución anti-ideal. Este índice cuantifica qué tan cerca está cada alternativa de ser la mejor o la peor solución posible.
- **7_Clasificación de las alternativas:** Finalmente, se clasifican las alternativas según su índice de similitud a la solución ideal. Cuanto más cercano a 1 sea este índice, mejor será la alternativa en comparación con las otras.

El algoritmo TOPSIS es particularmente útil cuando se enfrenta a decisiones multicriterio en las que se deben tener en cuenta diferentes aspectos o características para evaluar las alternativas. Proporciona una forma sistemática de comparar y clasificar estas alternativas en función de su desempeño en múltiples criterios.

5.3. Streamlit Web App: Wine Me App!

Hemos usado el framework Streamlit para implementar las funciones de los diferentes modelos, debido no solo a la versatilidad de éste, sino que el uso de Streamlit nos permitía seguir con el mismo lenguaje de programación (https://winemeup.streamlit.app/).

El modelo de web app que hemos realizado permite a los usuarios no solo ver el resultado de los modelos sino interactuar con estos.

Hemos diseñado una interacción simple para el usuario la cual incluye dos visualizaciones: una con los nombres de los vinos, enlaces de compra, precio, cata y maridaje entre otras características, y otra en la que el usuario puede ver en un plano euclídeo la posición de su vino con respecto a los demás que les muestra el modelo de recomendación

6. Conclusiones

En este proyecto hemos obtenido con eficacia el objetivo principal al crear un modelo híbrido que combina un enfoque basado en Procesamiento del Lenguaje Natural (NLP) con el modelo TOPSIS. Además, se ha implementado con éxito este modelo híbrido en una aplicación web en Streamlit, ofreciendo todas las funcionalidades y resultados del modelo, junto con visualizaciones de gráficos relevantes e interactivos.

Todos los objetivos secundarios fueron cumplidos de manera satisfactoria, lo que refleja el sólido progreso alcanzado en el proyecto. Si bien aún estamos en proceso de incluir el dataset de vinos internacionales encontrado en Kaggle, hemos alcanzado con éxito los demás objetivos.

El proceso de desarrollo de la aplicación web ha sido altamente enriquecedor, proporcionando una nueva perspectiva sobre la representación de modelos y datos, y ayudando a interiorizar conceptos de lógica computacional.

Es importante destacar que este proyecto es realista y está adaptado al mundo de la enología. Además, es escalable tanto para usuarios sin experiencia previa en enología como para expertos y aficionados al vino.

Habilidades demostradas:

- Recolección y preprocesamiento de datos a gran escala
- Modelado de NLP y aprendizaje automático
- Desarrollo de aplicaciones web interactivas
- Evaluación y validación de modelos

Impacto del proyecto:

- Mejora la experiencia de compra de vinos para usuarios de todos los niveles
- Potencia la exploración de nuevos vinos y maridajes
- Aporta valor a la industria del vino a través de la innovación tecnológica

7. Limitaciones y áreas de mejora

La información proporcionada por bodeboca.com no era uniforme para todos los vinos, lo que nos obligó a descartar columnas por no tener la suficiente información. Además, la información de rating, crucial para nuestro modelo TOPSIS y potencialmente relevante para futuros proyectos que generarian modelos predictivos de precios basados en su calificación, estaba frecuentemente ausente (40% NANs en el dataframe original de bodeboca.com –raw-) o presentaba valores muy altos por defecto. Esto podría llevar a sobreestimaciones de la calidad del vino. Por esta razón, optamos por realizar una verificación manual de esta información, junto con notas de cata y maridaje, en otras páginas web, un proceso que consumió una cantidad considerable de tiempo. Además, para garantizar una mayor precisión en nuestros datos, sería necesario realizar una verificación similar para todos los vinos en nuestro conjunto de datos.

Otro de los puntos de conflicto importante del proyecto fue a la hora de trasladar las funciones de los modelos de recomendación NLP y TOPSIS a la aplicación web en Streamlit. La implementación ha sido un verdadero desafío, ya que debido a ciertas peculiaridades del framework hemos tenido que modificar no solo las funciones del código sino la distribución de éstas y algunos parámetros definidos en ellas. Aun así, el mayor reto no ha sido el cambio en el código, sino la compatibilidad entre los módulos que han sido usado en los modelos y el framework, ya que para poder importar las funciones del archivo Python en el que teníamos los modelos, éstos debían tener las mismas versiones que el framework. Hemos tenido ciertos módulos los cuales una vez instalados causaban conflictos con los que había que instalar, y hemos tenido que poner un esfuerzo extra en lo que a resolver discrepancias se trataba.

Tras consultar con la enóloga Marta Benito Reyes, el rating global que le daría a la aplicación/algoritmo de Wine Me App! es de 4,5 sobre 5. Sin embargo, una de las áreas que ella considera que mejoraría el proyecto notablemente, para llegar al 5, sería añadir información de elaboración o vinificación para cada vino. Esto hace referencia a si ha fermentado en sus pieles, con los hollejos …etc. Esta información era poco consistente en bodeboca.com y en un caso hipotético de mejora de este proyecto para la comercialización de este producto sería aconsejable evaluar cada vino y contactar cada bodega para que facilitaran las fichas técnicas de sus vinos.

8. Posibles líneas de investigación para ampliar el proyecto

Este proyecto, como hemos visto a lo largo de la memoria, es muy versátil y podría ampliarse con numerosas propuestas. Entre ellas destacamos:

- Visualización de las ubicaciones de las bodegas de los vinos recomendados en un mapa utilizando librerías de geolocalización como *folium*
- Ampliación del modelo TOPSIS a mas categorías/drivers para ampliar la capacidad de decisión del usuario.
- Exploración de los datos de nuestros vinos y con ayuda experta crear columnas con valores numéricos para datos como niveles de cata (aromas a fruta y flores…), tipo…etc y crear gráficos de radar que muestren el vino origen y las recomendaciones.
- Incluir una tercera opción de recomendación, un segundo modelo NLP después del TOPSIS, con vinos internacionales basados en notas de cata, utilizando el conjunto de datos de Kaggle (+ 130K datos). La particularidad de este conjunto de datos radica en que está en inglés. En este caso, el conjunto de datos incluye vinos españoles, los cuales eliminaríamos. Posteriormente, tendríamos que traducir la información de la columna 'descripcion' del vino para el cual queremos recomendación al inglés utilizando un modelo preentrenado como *Hugging Face* y accederíamos a los vinos similares como anteriormente con similitud de cosenos.
- Como propuesta de una enóloga profesional, este proyecto podría llevarse a catas de expertos y probar para un vino de interés, hacer una cata con las 5 -10 recomendaciones

basadas en notas de cata y maridaje y después evaluar con un rating la recomendación de cada vino, si es verdaderamente una buena recomendación o no. Esto nos daría una valoración más exacta de la calidad del output del modelo de manera individualizada y con control en el experto.

9. Autores

- **Iván Pinto Grilo:** - Data Scientist
- **Maria Perez Sebastian:** - Data Scientist
- **Soraya Álvarez Codesal**: - Data Scientist – Wine Me App! Project: https://winemeup.streamlit.app/