Trabajo práctico 1

Nombre: Juan Cruz Saldaño

Fecha: 21/08/2024

Índice

Índice **2**

Aplicación de la Ciencia de Datos en la Industria Vinícola **3**

Desafíos y Oportunidades de los Macrodatos **5**

Aplicación de la Inteligencia Artificial en la Industria Vinícola **6**

Proceso de Metodologico de la Ciencia de Datos **7**

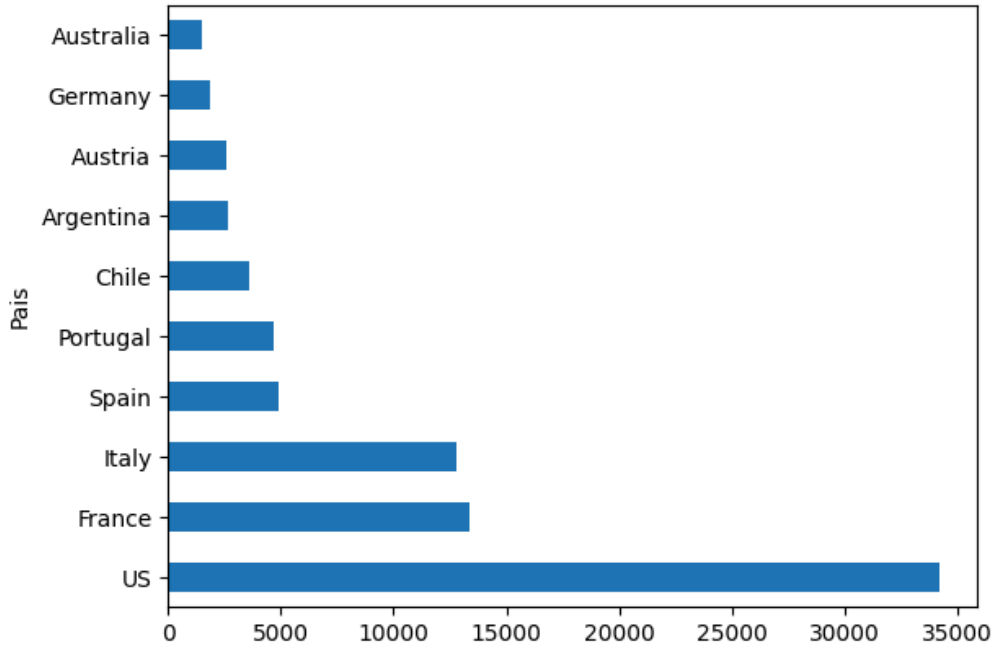
Aplicación de la Ciencia de Datos en la Industria Vinícola

La ciencia de datos es un campo interdisciplinario el cual combina diferentes áreas de estudio y tiene como objetivo la extracción de información relevante a través del procesamiento y análisis de distintos conjuntos de datos. Esta disciplina puede ayudar a los productores de vino a comprender los diferentes factores que influyen en el precio y la calidad del producto, como factores geográficos(país, provincia y/o región), las variedades de uvas, la antigüedad, entre otros. Además, la ciencia de datos puede intervenir a lo largo de todo el proceso, desde la extracción de reseñas sobre vino de múltiples fuentes, el preprocesamiento de los datos para eliminar el ruido o información errónea, hasta la creación de herramientas para sistemas de recomendación, predicción de precios y calidad del vino, e incluso para entrenar a nuevos catadores.

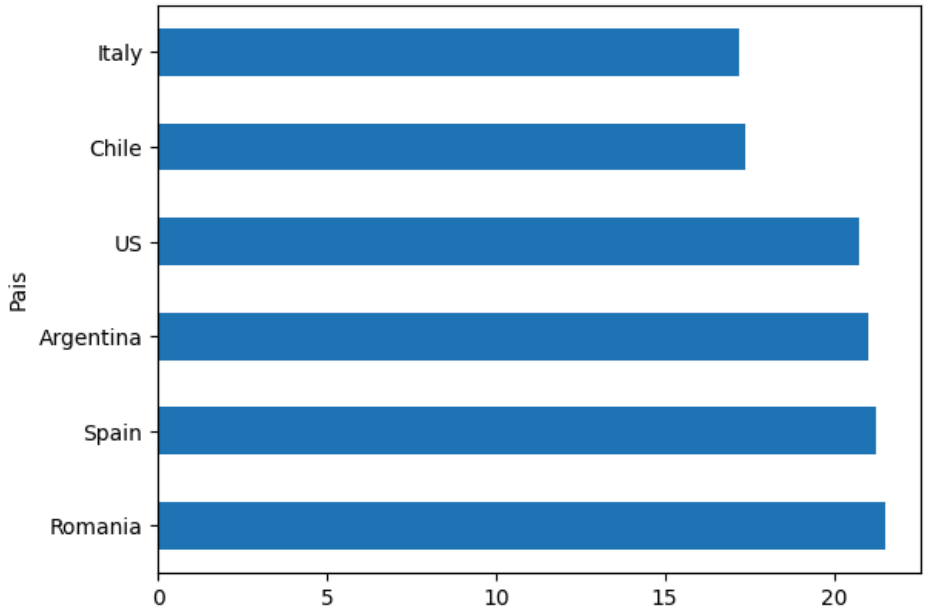
Supongamos la siguiente situación hipotética: somos una empresa multinacional la cual está buscando invertir en el sector vinícola y decidimos contratar una consultora externa para tomar una decisión informada. La consultora nos entrega un resumen exhaustivo de un conjunto de datos proporcionado por nosotros, y nos brinda información basada en nuestros requerimientos:

- En qué parte del mundo se producen vinos
- La relación calidad-precio, entre otros,

Podemos ver que los países con mayor producción de vino son:



Además, los vinos con mejor relación calidad precio son:



Basándonos en la información brindada, el dato de la relación calidad precio es particularmente interesante, ya que puede ser un indicador de regiones con alta producción de vino a un costo relativamente bajo. Además, la consultora nos informa que el precio promedio de una botella de vino a nivel mundial es de $27 USD, pero en algunos países, como Inglaterra, el precio promedio es de $51 USD (es decir, $24 USD más que el promedio). Como empresa, podríamos concluir que sería estratégico producir vinos de alta calidad en España y exportarlos a Inglaterra, donde podrían comercializarse a un precio superior, aprovechando también la proximidad geográfica entre ambos países.

Finalmente, la consultora nos entrega un modelo de inteligencia artificial que podríamos utilizar como predictor para estimar los posibles precios de venta si produjeramos vinos de determinadas características.

Desafíos y Oportunidades de los Macrodatos

Los macrodatos o Big data en la industria del vino presentan tanto desafíos como oportunidades que permitan mejorar aspectos como la producción, decisiones comerciales entre otros. Dentro de las oportunidades, se destaca la creación de

sistemas de recomendaciones personalizadas, donde los usuarios pueden acceder a aplicaciones que, en respuesta a determinados parámetros, les sugieren productos que mejor se ajustan a sus preferencias.

Sin embargo, para poder obtener beneficios de estos macrodatos, primero es necesario superar una serie de limitaciones::

- Calidad de los datos: En el dataset de reseñas de vinos, por ejemplo, encontramos una cantidad significativa de datos faltantes, duplicados o erróneos. Originalmente, el dataset contenía cerca de 130,000 filas, pero tras un proceso de limpieza y depuración, sólo 85,000 eran utilizables. Esto demuestra la importancia de contar con datos de alta calidad para obtener resultados precisos y útiles.
- Habilidades y conocimientos: Para analizar y extraer valor de los datos, es crucial tener un conocimiento profundo del sector vinícola. Además, dado que las descripciones de los vinos son subjetivas y varían mucho en estilo, es necesario contar con habilidades en procesamiento de lenguaje natural (NLP, por sus siglas en inglés) para poder analizar efectivamente las reseñas escritas por los catadores.

Para concluir, aunque el análisis de macrodatos en la industria del vino presenta grandes desafíos, como la calidad de datos y la necesidad de habilidades técnicas especializadas, es posible desarrollar y transformar la industria de manera en que los productores y comerciantes respondan mejor a las preferencias de los consumidores

Aplicación de la Inteligencia Artificial en la Industria Vinícola

La inteligencia artificial (IA), a través de técnicas como el aprendizaje automático (machine learning), el aprendizaje profundo (deep learning) y el procesamiento de lenguaje natural (NLP), ha abierto grandes posibilidades para todas las industrias, incluyendo la vinícola. A partir de este conjunto de datos, podríamos desarrollar distintos modelos predictivos, uno de los cuales podría ser la clasificación de variedades de vinos basándose en descripciones textuales.

Para ello, después de limpiar el conjunto de datos, es posible procesar las distintas descripciones para identificar palabras clave o tópicos comunes que se correspondan con ciertas variedades de vino. Técnicas de NLP, como el análisis de sentimiento, la tokenización y el modelado de tópicos, pueden ser aplicadas para extraer características relevantes de las descripciones textuales. Posteriormente, estas características pueden alimentar un modelo de clasificación, como un árbol de decisión, una máquina de vectores de soporte (SVM) o un modelo basado en redes neuronales, para predecir la variedad de vino en función de su descripción.

Dentro de los beneficios se pueden encontrar:

- Personalizar recomendaciones a los consumidores: Los modelos predictivos podrían personalizar las recomendaciones de vinos para los consumidores, basándose en sus preferencias de sabor descritas en texto.
- Entrenamiento de catadores: Un sistema de este tipo podría ser utilizado para entrenar a futuros catadores, ayudándolos a identificar y clasificar variedades de vino basándose en descripciones textuales, precios, localidad geográfica etc.

Para lograr un sistema útil, se deben superar ciertas limitaciones tales como:

- Calidad de los datos: La presencia de datos faltantes, errores o inconsistencias en las descripciones puede afectar negativamente la precisión del modelo.
- Subjetividad en las descripciones: Las descripciones de los catadores pueden ser subjetivas y variadas, lo que dificulta la creación de predictores consistentes y precisos.
- Complejidad del lenguaje: El lenguaje utilizado en las descripciones de vino puede ser muy específico debido a los modismos utilizados por los catadores, esto presenta un desafío adicional para los modelos de NLP en la extracción de características relevantes.

Es posible concluir que la inteligencia artificial, luego de superar las limitaciones prestadas por el conjunto de datos, presenta grandes oportunidades para poder mejorar la industria en general, e incluso generar un acercamiento con los clientes.

Proceso de Metodologico de la Ciencia de Datos

La metodología de la ciencia de datos proporciona un enfoque sistemático y riguroso para analizar grandes volúmenes de datos, como las reseñas de vinos, y extraer información valiosa que puede mejorar la comprensión de las preferencias y tendencias de consumo.

El primer paso es la limpieza de datos, que es crucial para garantizar que el análisis posterior sea preciso y relevante. Las reseñas de vinos suelen estar plagadas de datos faltantes, duplicados o inconsistentes. Mediante técnicas de limpieza, como la eliminación de duplicados, la imputación de valores faltantes, la normalización de los textos(eliminar mayúsculas y caracteres especiales) se puede preparar un conjunto de datos limpio que refleja de manera más precisa la realidad. Esto reduce el ruido en los datos, lo que permite un análisis más efectivo y confiable de las preferencias de los consumidores.

El segundo paso, es el análisis exploratorio de datos permite identificar patrones, tendencias y relaciones dentro del conjunto de datos de reseñas. Al analizar variables como las rese;as y el título de las rese;as, en estos podemos realizar distintos tipos de análisis:

- Análisis de sentimientos: Este consiste en descifrar los sentimientos y emociones a través de análisis de la gramática
- Topic modelling: Es un conjunto de técnicas para descubrir estructuras semánticas comunes, es decir, permite hallar el tema central
- Hallar palabras comunes: permite hallar las palabras comunes, y luego estas pueden ser asociadas a distintos atributos del producto.

Por último, la visualización de datos es esencial para comunicar los hallazgos del análisis de manera clara y comprensible. Herramientas de visualización, como gráficos de barras, gráficos de torta, histogramas, permiten a los analistas presentar información compleja de manera que sea fácilmente interpretable por los

interesados. Por ej, el siguiente gráfico podría ser una forma de representar palabras comunes en todas las reseñas:

Referencias

Referencias:

Juan Cruz Saldaño. (2024). *Gráficos de mi autoría*. GitHub. https://github.com/fowardelcac/Seminario-en-ciencia-de-datos/blob/main/Tp1/B_resultados.ipynb