SEM309 – 10160 SEMINARIO DE PRÁCTICA EN CIENCIA DE DATOS

TRABAJO PRÁCTICO N° 1

BENEFICIOS DE LA CIENCIA DE DATOS EN LA INDUSTRIA VITIVINÍCOLA

MARIANO NATIELLO (VLDC000632)

SEPTIEMBRE 2024

ÍNDICE

1. CONTEXTO

El vino culturalmente ha sido parte de celebraciones durante miles de años, es un complemento gastronómico esencial que permite disfrutar de sabores y aromas de alimentos, en donde varios países dependen económicamente en gran medida de la industria vitivinícola, cuya producción y comercialización genera empleos y fomenta el turismo que beneficia a varios sectores.

El crecimiento constante de la industria del vino ha llevado a una mayor competencia entre bodegas y regiones vitivinícolas. Es por esto que en la actualidad, las reseñas y puntuaciones de vinos realizadas por expertos influyen considerablemente al momento de la compra en el comportamiento de los consumidores.

De acuerdo a lo anterior y a la importancia de analizar dichas reseñas, se ha considerado un conjunto de datos basado en 130.000 reseñas de vinos obtenidas del sitio Wine Enthusiast, cuyo análisis exhaustivo de los datos permitirá no solo descubrir información valiosa para sommeliers, enólogos, comerciantes y aficionados, sino que también ayudar a comprender con mayor detalle las características y calificaciones de los vinos.

2. OBJETIVOS

Para este Trabajo Práctico N° 1, se definen los siguientes objetivos:

a) Comprender principios básicos de la ciencia de datos y su aplicación en el mundo empresarial.
b) Identificar desafíos y oportunidades relacionados con los macrodatos.
c) Introducir los beneficios y limitaciones que implica la aplicación de la inteligencia artificial y el aprendizaje automático.

3. CASO PLANTEADO

A continuación y en base al cuestionario brindado, se detalla la reflexión sobre cada tema:

3.1. Relación de la ciencia de datos y la industria vitivinícola

¿De qué manera la ciencia de datos puede mejorar la comprensión de los factores que influyen en la calidad y características de los vinos, y cómo puede esta información ser utilizada por productores y comerciantes para tomar decisiones informadas?

Para responder esto, es necesario primero comprender conceptualmente la ciencia de datos. Por su lado, Laura Igual y Santi Seguí la definen en su libro *"Introducción a la Ciencia de datos"* como una metodología mediante la cual se pueden inferir conocimientos prácticos a partir de los datos, mientras que Walter Sosa Escudero inicia su libro "*Big Data*" presentando a la ciencia de datos como una nueva ciencia que involucra la estadística, la matemática, la computación, el diseño y todas las áreas de nuestra vida cotidiana.

Dicho autor, resalta que el poder de la ciencia de datos se encuentra en su capacidad para encontrar patrones útiles, hacer predicciones y ofrecer respuestas a preguntas complejas de diversas áreas. En su libro, Sosa Escudero relaciona la ciencia de datos con las tres "V" clave del Big Data: Volumen, Velocidad y Variedad.

Ahora bien, si nos focalizamos en el caso planteado del sitio Wine Enthusiast, podríamos establecer el **volumen** como el set de datos utilizado incluyendo 130.000 reseñas de vinos. **Velocidad**, considerando que las reseñas pueden ser generadas y actualizadas en todo momento, permitiendo realizar un análisis en tiempo real, y **variedad,** en cuanto a la variedad de datos que presenta la reseña de vinos mencionada, abarcando información detallada sobre su país de origen, descripciones, precios y puntuaciones, por citar algunos.

Mediante la ciencia de datos, es posible revelar información a ser utilizada por productores y comerciantes en la toma de decisiones estratégicas, considerando por ejemplo las preferencias de los consumidores, tendencias a lo largo de los años, accesibilidad del producto de acuerdo a su precio, entre otros.

Si bien el set de datos utilizado en nuestro caso se presenta de forma plana, si analizamos las columnas incluidas, podemos establecer a priori diferentes

relaciones que una vez trabajado el conjunto de datos y mediante la visualización de datos, podrían brindar información clave como:

- Cantidad de reseñas realizadas de una variedad.
- Bodegas, países, regiones y provincias de los vinos con mayor puntaje.
- Viñedo (y bodega) que logró mejores reseñas, por ejemplo aquellas mayores a 90 puntos.
- Relación entre variedad, vino, año de cosecha, precio, puntaje y catador que participó en la reseña.

Si bien lo anterior es un ejemplo de la información que se podría obtener, sin duda que la misma aportaría significativamente en la toma de decisiones de los productores vitivinícolas, considerando las reseñas realizadas y las características geográficas de la bodega en sí.

3.2. Desafíos y oportunidades de los macrodatos

Dado el gran volumen de datos disponible en el set de datos de reseñas de vinos, ¿qué desafíos y oportunidades presentan los macrodatos para el análisis y la visualización de información relevante en la industria del vino?

Tan solo con una exploración preliminar sobre el set de datos, utilizando Microsoft Excel, es posible detectar problemas con la calidad de datos en relación a que los mismos presentan inconsistencias (valores faltantes, duplicados, etc.) en varios atributos (*region_1, region_2, taster_name, taster_twitter_handle, variety*). Esto sin duda que se convierte en el primer desafío considerando el impacto que tendrá en la calidad y precisión de los resultados.

Contemplando las técnicas para la preparación de los datos, en este caso deberíamos tener en cuenta la limpieza y acondicionamiento de los datos a fin de asegurar que mismos sean precisos, consistentes y útiles.

En cuanto a las oportunidades, considerando desde ya herramientas para tal fin, podemos hacer uso de gráficos y visualizaciones interactivas (dashboard) para comunicar no solo los patrones y relaciones ocultas sino también para presentar la información que resulte del análisis exhaustivo de los datos.

3.3. Inteligencia artificial y aprendizaje automático

Aunque este caso de estudio se centra en el análisis exploratorio y no en la predicción, ¿cómo podría la inteligencia artificial y el aprendizaje automático ser utilizados para desarrollar modelos predictivos que identifiquen variedades de vino basándose en descripciones textuales? ¿Qué beneficios y limitaciones podrían surgir de esta aplicación?

Podemos mencionar la Inteligencia Artificial con un concepto amplio que abarca cualquier técnica que intenta imitar o replicar aspectos de la inteligencia humana en máquinas. Aprendizaje automático o machine learning (ML) como una de las técnicas utilizadas por la IA para aprender de la experiencia y nuevos datos y el Procesamiento de lenguaje natural (NLP) como una subárea de la IA y ML cuyo objetivo es interpretar y generar lenguaje de forma tal que tenga sentido para los humanos.

En base a lo mencionado en el párrafo anterior y puntualmente con nuestro caso, podríamos utilizar NLP para analizar “texto” en atributos particulares del dataset, por ejemplo *description* o *title,* con el objetivo de obtener y analizar el texto incluido. De esta manera, por ejemplo si utilizamos el texto especificado en “description” de la reseña 78.435 se destacan adjetivos como “charmer” (encanto), “well balanced” (equilibrada/o) y “delicious” (deliciosa/o), los cuales pueden interpretarse (desde lo humano) para comprender mejor las características del vino, utilizando esta información en modelos predictivos que contribuyan e impacten en gran medida en la comercialización de futuras producciones.

Si consideramos esto, las limitaciones podrían presentarse en descripciones ambiguas o subjetivas, en donde un sommelier podría percibir el aroma y sabor de un vino diferente a otro colega, lo cual dificulta al modelo para generalizar de una forma correcta.

Otras limitaciones se presentan por ejemplo en la falta de contexto. Interpretar “fuerte” o “suave” sin saber puntualmente cuál característica del vino se está referenciado, podría ser un ejemplo de esto.

Por último en cuanto a los beneficios, aplicar estos modelos predictivos permitirían proporcionar a los sommeliers, enólogos, comerciantes o aficionados, de una información precisa de la variedad del vino basado en las descripciones

que antes no se destacaban. Estos modelos predictivos, al contar con más reseñas, permiten establecer recomendaciones más fundamentadas y mejorar la toma de decisiones.

3.4. Aplicación de metodología de la ciencia de datos

¿Cómo puede la metodología de la ciencia de datos, incluyendo la limpieza de datos, el análisis exploratorio y la visualización, mejorar la comprensión de las preferencias y tendencias de consumo de vinos a partir de las reseñas disponibles?

Tanto la limpieza de datos, como el análisis exploratorio y la visualización de datos, permiten transformar la información de las reseñas de vinos en información valiosa en cuanto a tendencias de consumo u orientarlo a una nueva producción de vinos.

Si tomamos el dataset actual y tal como fue señalado, la limpieza de datos es necesaria para normalizar términos que permitirá al modelo identificar patrones de forma más precisa y por otro lado para evitar cualquier sesgo producto de datos faltantes o duplicados.

El análisis exploratorio nos brindará aquellos patrones o características significativas (determinar si los vinos que reciben mayor puntaje son aquellos más caros) o descubrir palabras que tengan relación con algún factor en particular (los vinos con “aroma a frutos rojos” resultan ser los de menor puntaje).

4. REFERENCIA BIBLIOGRÁFICA

Igual, L., & Seguí, S. (2017). *Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications*. Switzerland, CH: Springer International Publishing.

Sosa Escudero, W. (2019). *Big Data: breve manual para conocer la ciencia de datos que ya invadió nuestras vidas*. Bs As, AR: Siglo XXI Editores.

Enunciado del Trabajo Práctico N° 1 de la materia Seminario de práctica en Ciencia de Datos de la carrera Licenciatura en Ciencias de Datos.