Quiz 1

Ana Marchena anagabi0703@gmail.com
28309145

En este quiz, se desea que usted exprese su comprensión de los temas tratados hasta ahora (Tema 3). El examen consta de preguntas teóricas y prácticas, se espera que en cada caso argumente de manera apropiada sus respuestas. **Cualquier evidencia de plagio o uso indebido de LLMs, será penalizado con la mínima nota**.

Instrucciones:

- **Modifique las lineas de 61, 62 y 63 del código latex para colocar sus datos.**
- Modifique los archivos en la carpeta respuestas para incluír su respuesta a cada pregunta usando el mismo índice como nombre, ej: *respuestas/1.tex* será asumida cómo la respuesta a la pregunta 1.
- Envie el link de su proyecto en overleaf con permisos de edición en un mensaje directo.

Quiz

(2pts) Pregunta 1: Considerando una metodología de minería de datos moderna como CRISP-DM, ¿qué tipos de requerimientos de los datos se definen primero y por qué?

Respuesta: Analizando la metodología CRISP-DM, observamos que los requerimientos que primero se definen son los del negocio, ya que estos establecen el problema a resolver, los objetivos y los criterios de éxito. A partir de ellos se determinan posteriormente los requerimientos de los datos. Esto se hace de esta forma ya que los datos deben responder a un objetivo en específico del negocio, por lo cual, definirlos en un principio podría conllevar a obtener información irrelevante o insuficiente. En cambio, partir del negocio garantiza que el análisis realizado sea útil y esté alineado con la toma de decisiones.

(4pts) Pregunta 2: Usando los datasets disponibles en kaggle, seleccione 2 conjuntos de datos *(Incluya sus links en la respuesta)* y señale que problemas estructurales están presentes en dichos datasets. Indique también las transformaciones necesarias para cumplir los requerimientos estructurales vistos en clase.

Respuesta:

1. TMDB Top-Rated Movies Dataset

(https://www.kaggle.com/datasets/ranodipghosh/latest-top-rated-movies-from-tmdb)

- **Problemas estructurales presentes en el dataset:**

Se presentan varias problemáticas estructurales, en primer lugar, existe redundancia semántica entre atributos, ya que las columnas original title y title representan el mismo fenómeno (el nombre de la película). Ademas, estas mismas, en muchos casos, contienen valores idénticos lo que introduce ambigüedad y duplicación innecesaria de la información. Asimismo, el dataset carece de un identificador único para cada película, lo que dificulta la identificación unica de los registros y su integración con otras fuentes de datos. Otra problemática estructural es que algunas columnas contienen múltiples variables, como ocurre con release date, que agrupa información de día, mes y año en un solo atributo, limitando los análisis temporales.

- **Transformaciones necesarias:**

Para corregir estas problemáticas, en primer lugar, es necesario eliminar la redundancia, consolidando el atributo del nombre, para ello se debe realizar una reagrupación de columnas para mantener únicamente title como referencia principal. Esto simplifica el esquema y evita la redundancia semántica. El atributo original title puede ser eliminado o transpuesto a una tabla secundaria dependoendo de si aporta o no un valor diferenciador en el análisis realizado. También se deben de reagrupar las filas incorporando un ID para garantizar la unicidad. Finalmente, es requerido aplicar una descomposición estructural sobre release date para crear variables independientes como .ªño.º "década". Esta transformación es vital para que el modelo pueda identificar patrones estacionales o tendencias históricas que están ocultas en el formato de fecha completa.

2. Anime TV-Shows Dataset 2023
(https://www.kaggle.com/datasets/forgetabhi/anime-tv-shows-dataset-2023)

- **Problemas estructurales presentes en el dataset:**

En un principiop, la columna Lanzamiento combina la fecha de inicio y la fecha de fin en un solo campo (como "Oct 2016 - Dic 2016"), lo que impide filtrar por año, calcular duración de la serie o identificar si aún está en emisión. Además, algunos animes aparecen varias veces con diferentes temporadas o partes, y los nombres incluyen variaciones (como "Shingeki no Kyojin parte 2"), lo que complica agrupar los datos por serie principal.

- **Transformaciones necesarias:**

En primer lugar, la columna Lanzamiento necesita de una descomposición de columnas, separando este campo en atributos independientes como: Fecha de inicio, fecha de fin y estado de emision para las series en curso. Esta separación permitira que el modelo pueda realizar cálculos de duración y filtros cronológicos que antes estaban bloqueados por el formato libre.

Otro punto a tratar es la redundancia en los nombres de las temporadas y la falta de una jerarquía clara representan un problema de integridad y agrupación. Para corregirlo, se requiere la reagrupación de registros mediante la creación de un identificador de "Serie Principal"que consolide todas las temporadas bajo un mismo grupo. Además, es necesario incorporar un ID único por fila para garantizar la unicidad de cada entrada, facilitando así la aplicación de técnicas descriptivas como la definición de popularidad por franquicia.

(4pts) Pregunta 3: Usando los datasets disponibles en kaggle, seleccione 2 conjuntos de datos *(Incluya sus links en la respuesta)* y señale que problemas funcionales están presentes en dichos datasets. Indique también las transformaciones necesarias para cumplir los requerimientos funcionales vistos en clase.

Respuesta:

1. Harry Potter Dataset

(https://www.kaggle.com/datasets/gulsahdemiryurek/harry-potter-dataset)

- **Problemas funcionales presentes en el dataset:**

El dataset actual presenta problemas significativos de consistencia y completitud, ya que la gran mayoria de sus columnas contienen un alto porcentaje de valores nulos o desconocidos, lo que limita la posibilidad de un análisis estadístico. Además, algunas columnas combinan información compleja en un solo campo, como en las varitas, donde se incluye longitud, madera y núcleo, y trabajos, que puede listar múltiples roles. También hay valores no estandarizados en categorías como las especies, y algunas entradas incluyen información temporal o poco clara, lo cual dificulta la limpieza y agrupación de datos para un análisis.

- **Transformaciones necesarias:**

Para mejorar la calidad del dataset se deberia, en un inicio, normalizar las categorías de especies, estatus de sangre y color de cabello, y además, descomponer columnas complejas, como la de varitas, en atributos separados por: longitud, madera y núcleo. Es necesario además limpiar los valores nulos, designando datos o marcándolos como desconocidos. Asimismo, conviene estandarizar los nombres de trabajos y cargos, separando múltiples roles en columnas distintas. Finalmente, se podrían crear variables adicionales para facilitar análisis, como tipo de personaje, lo que nos daria un conjunto de datos mas completo y un mejor análisis.

2. Netflix popular movies dataset

(https://www.kaggle.com/datasets/narayan63/netflix-popular-movies-dataset)

- **Problemas funcionales presentes en el dataset:**

Existen vairas problematicas, en un inicio, existen columnas con valores faltantes o nulos, especialmente en certificado y duración. Además, algunas columnas contienen listas que deberían dividirse, como estrellas o género, ya que esto dificulta la agregación y filtración de los datos. Asimismo, la información temporal (año) incluye rangos de series (por ejemplo, 2018–) que necesitan un formato uniforme. Hay duplicados y categorías inconsistentes en campos como género y certificado, y finalmente, algunas columnas de texto, como descripción, requieren limpieza de caracteres especiales y etiquetas redundantes.

- **Transformaciones necesarias:**

En un inicio se debe de transponer filas y columnas según si se requiere un análisis por serie o película, también descomponer columnas como estrellas y género en múltiples columnas o en tablas relacionadas, normalizar año para manejar rangos y valores de series en curso, agrupar o categorizar los certificados y géneros para reducir inconsistencias, designar valores faltantes o definir valores desconocidos, y eliminar duplicados. Finalmente limpiar las columnas de texto de caracteres no deseados y etiquetas redundantes.

Asignación

(10pts) Usando los datos world-university-rankings-2023 , cree un notebook en python (jupyter) en donde realice el preprocesamiento que considere necesario para el conjunto de datos dados, considerando los principios discutidos en clase. Almacene dicho notebook en su perfil de github y coloque el link del repositorio en este archivo de asignacion

Respuesta: repo