Desafío - Preprocesamiento de datos

En este desafío tendrás la oportunidad de poner a prueba los conceptos aprendidos durante la sesión. Los ejercicios están diseñados para reforzar practicar lo explicado en clases y poder implementar un caso real.

Lee todo el documento antes de comenzar el desarrollo individual, para asegurarte de tener el máximo de puntaje y enfocar bien los esfuerzos. Asegúrate de seguir las instrucciones específicas en cada ejercicio y de completar los requerimientos adicionales, si los hubiera. ¡A disfrutar aprendiendo!

Tiempo asociado: 4 horas cronológicas

Descripción

¡Bienvenidos, intrépidos data scientists y fanáticos del anime, a una tarea de preprocesamiento de datos llena de aventuras en el fascinante universo de las series de anime! En esta épica misión, te embarcarás en un viaje para convertir un conjunto de datos de episodios de anime en una fuente de conocimiento de alta calidad. Tu destino final es preparar estos datos para un análisis en profundidad y desbloquear los secretos que se ocultan en el mundo del anime.

La primera etapa de tu búsqueda implica la evaluación de la calidad de los datos. Como un verdadero cazador de datos, deberás rastrear y enfrentarte a los temibles valores atípicos, los datos faltantes que se esconden como camaleones y las inconsistencias que acechan en la oscuridad. Tu tarea es purificar los datos, como un héroe que elimina las impurezas en su camino. Documenta cada movimiento que haces para despejar estos obstáculos.

En tu próxima parada, te sumergirás en un mundo de análisis exploratorio. Armado con tu caja de herramientas de estadísticas y gráficos, desentrañarás los secretos de la distribución de variables clave. ¿Qué revelarán las gráficas y las estadísticas descriptivas? Estás destinado a descubrir patrones intrigantes y tendencias fascinantes. Estas revelaciones podrían convertirse en tu guía en este viaje.

Siguiendo tu viaje, ejercerás tu poder para crear nuevas características. Como un mago que conjura hechizos, concebirás al menos dos características nuevas que pueden enriquecer el análisis de episodios de anime. Comparte tus poderes de creación y explica el origen de estas nuevas características, así como la magia que utilizaste para calcularlas.

La Selección del Elegido

En un giro inesperado, utilizarás un método de selección de características para identificar las joyas más valiosas entre tus tesoros de datos. ¿Cuáles de estas características son las más importantes para predecir el destino de los episodios de anime? Tus elecciones deben ser sabias, como un rey que selecciona a sus caballeros. Esta etapa es crucial, y tu reino de conocimiento depende de ello.

Como un guerrero que se entrena constantemente, no te detendrás en la selección de características. Continuarás tu viaje iterativo, explorando cómo esta elección afecta tu comprensión y el rendimiento general del análisis. La adaptación es tu compañera constante.

Finalmente, compartirás tus hallazgos en una presentación épica. Crearás un informe o una presentación que resuma tus aventuras en el preprocesamiento de datos. Incluirás visualizaciones deslumbrantes, estadísticas sorprendentes y explicaciones que permitan a otros exploradores de datos seguir tus pasos. Destacarás los tesoros ocultos que descubriste en tu análisis exploratorio y cómo estas gemas pueden iluminar el camino hacia el entendimiento.

Recuerda, este viaje es un proceso épico y, como todo gran viaje, está lleno de desafíos y descubrimientos sorprendentes. Demuestra tu valía como un data scientist audaz y astuto, y que tu legado perdure en el vasto mundo del anime. ¡Buena suerte en tu búsqueda, aventurero de datos!

Para llevar a cabo todo esto, necesitarás el archivo de datos **imdb_anime.csv,** que contiene las siguientes columnas:

- Title: Nombre de la animación
- Genre: Género(s) bajo el cual cae la animación, por ejemplo, Acción, Aventura, etc.
- User Rating: IMDb calificación de usuarios sobre 10.
- Number of Votes: Total de usuarios de IMDb que han calificado la animación.
- Runtime: Duración de la animación en minutos.
- Year: Año en que se estrenó o comenzó a emitirse la animación.
- Summary: Un resumen breve o completo de la trama de la animación. Resúmenes completos se obtienen cuando están disponibles.
- Stars: Lista de actores principales o actores de voz involucrados en la animación.
- Certificate: Certificación de la animación, por ejemplo, PG, PG-13, etc.
- Metascore: Calificación de Metascore, si disponible, que es una puntuación agregada de varios críticos.
- Gross: Ganancias brutas o recaudación en taquilla de la animación.
- Episode: Indicador binario si la lista es para un episodio de una serie (1 para sí, 0 para no).
- Episode Title: Título del episodio si la lista es para un episodio; de lo contrario, será None (Ninguno).

Considerando estos datos:

1. Realiza un análisis de calidad de datos, revisando aspectos básicos y selecciona un primer conjunto de variables a eliminar. Luego de ello, realiza un análisis exploratorio inicial considerando gráficos de distribuciones de las diferentes variables, y concluye al respecto. Si observas algo raro respecto a los tipos de variables debes proponer algún tratamiento.

2. Transformación Inicial de Datos: las diferentes columnas que son datos de texto deben ser transformadas a numéricas para poder explorarlas de mejor forma por ejemplo:
    a. User Rating: Extraer el número correspondiente al rating
    b. Number of Votes: Convertir en número
    c. Year: Extraer el año de inicio del anime
    d. Otros. Aplica algún criterio para saber qué variables deben ser transformadas en primera instancia.

3. Revisión de outliers: ahora que tienes variables numéricas revisa la distribución y utiliza algún método para encontrar outliers, por ejemplo IQR o Z-score.

4. Transformación de variables finales: realiza un pequeño análisis de distribuciones y transforma las variables aplicando transformaciones como logaritmo o get_dummies para extraer las diferentes categorías. Genera una estrategia para lidiar con los valores nulos y crea las variables que te parezcan necesarias.

5. Análisis de Correlaciones: genera un análisis de correlaciones de las variables. No es necesario que apliques todos los métodos vistos en clases, basta que argumentes bien cuál utilizarás y por qué, y si necesitas algo más. La idea es generar gráficos para entender la relación entre las diferentes variables, poniendo foco en la variable objetivo.

6. Genera una función que resuma todo el procesamiento necesario para el dataset, que lea el dataset original y entregue un dataset ya tratado, con las columnas transformadas y creadas.

7. A partir de las columnas que obtuviste realiza una selección de variables según los siguientes métodos:
    a. Filtros basados en correlaciones
    b. Forward Selection.

    Compara ambos métodos y responde si coincide lo resultante con lo obtenido en el análisis exploratorio.

Requerimientos

1. Analiza datos y los prepara considerando datos nulos, faltantes o outliers, considerando el contexto dado y necesidades de transformación. **(3 puntos)**
2. Analiza correlaciones entre variables, justificando su selección desde el contexto e interpretando los indicadores obtenidos. **(3 puntos)**
3. Selecciona variables para un análisis, considerando diferentes métodos e interpretando sus resultados. **(4 puntos)**

¡Mucho éxito!

Consideraciones y recomendaciones

- Aprovecha las funciones que tiene la librería pandas para el tratamiento de datos.
- Sé ordenado al momento de trabajar y piensa en cómo iterar, es decir la misma función te sirve para revisar cómo queda un dataset antes y después de cierto tratamiento.
- Genera funciones para reutilizar código.
- Para recuperar la categorías de la columna 'Genre' puedes utilizar lo siguiente:
    - **data['Genre'].str.split(', ').str.join('|').str.get_dummies()**