pragma

Prueba práctica Científico de Datos

1. Acabas de ingresar como científico de datos a una compañía que combina tecnología con analítica de datos para tomar mejores decisiones. Esta empresa está a punto de lanzar su marca a los Estados Unidos, y tiene pensado hacer uso de una de las redes sociales de mayor impacto en todo el mundo, YouTube.

   La estrategia consiste en posicionar su propio canal de YouTube, donde sean capaces de generar tráfico hacia la página web, y así aumentar la posibilidad de abrir negocios. Dentro de la estrategia, está contemplado entender mejor el comportamiento de esta red social: cuáles son los videos más populares, cuantos likes consiguen, cuantas visitas tienen, etc. Esto permitiría en el futuro decidir si hacer uso de sponsors, y si es así, escoger a aquellos con mayor potencial de retorno.

   Tu misión como científico de datos, es analizar los datos disponibles que se tienen sobre las interacciones en YouTube de los videos de mayor tendencia (no necesariamente de mayores visitas) en Estados Unidos. En el siguiente enlace a Kaggle podrás encontrar el dataset. Procura tomar los datos (en formato .JSON y .CSV) de los Estados Unidos, ya que es el mercado objetivo. Para este reto tendrás que ser capaz de procesar información estructurada y semiestructurada, de tal manera que puedas obtener información valiosa que ayude a tomar mejores decisiones.

   A continuación, algunas preguntas que podrían ser interesantes responder:

   - ¿Cuáles son las categorías de vídeos que reciben mayores vistas y Likes?
   - ¿Es posible encontrar agrupaciones o clasificaciones de videos?
   - ¿Cuál es la combinación de características o atributos más importantes que hacen de un video tendencia?
   - ¿La temporada o fecha en el que el video es publicado tiene alguna influencia?
   - ¿Es posible predecir cuantos likes o visitas tendrá un video? Si es así, crea un modelo que lo compruebe.

Ten presente que estas preguntas son sólo la base del análisis que la compañía espera recibir, pero existen muchas otras que podrían ser planteadas y de interés para la toma de decisiones (ya sean de análisis descriptivo, o predictivo). Se espera que realices todo un proceso de descubrimiento de información y obtención de conocimiento por medio de alguna metodología de data mining que conozcas. Es importante que definas con claridad cada una de las etapas y justifiques las tareas realizadas. Eres libre de escoger el modelo o modelos que creas pertinentes para extraer el mayor conocimiento posible, ejemplo: clustering, forecasting, regresión lineal, modelos de clasificación, modelos probabilísticos, diseño de experimentos, etc.

Imagina que el resultado del análisis lo vas a presentar a una audiencia multidisciplinar donde encontrarás perfiles como: CEO, CTO, Data Manager, y product managers.

Entregables:

- Jupyter Notebook con el paso a paso del modelo o modelos desarrollados (si lo haces con python, markdown si lo haces con R).
- Presentación de power point o el formato que prefieras de no más de 10 slides donde expongas:
    - Metodología usada para el proceso de descubrimiento de información.
    - Respuesta a las preguntas sugeridas con su apropiado análisis.
    - Modelo o modelos elegidos y la razón de su elección.
    - Insights y conclusiones propias a las que llegaste con el análisis (recuerda el público objetivo).