CURSO: CC50 – ADMINISTRACIÓN DE LA INFORMACIÓN

CLASE: TF (TRABAJO FINAL GRUPAL)

TEMA: CREACION DE CONOCIMIENTO A PARTIR DE LOS DATOS EN PYTHON

Objetivo

Crear conocimiento a partir de los datos, al desarrollar un Análisis Exploratorio de Datos (EDA) e intentar resolver un problema básico de Modelización de Datos, en el marco de la ejecución de un proyecto de analítica.

Competencias

Acorde con el ABET Student Outcome (2) para la carrera de Ciencias de la Computación, serán evaluadas las competencias, según rúbrica **TF-RUBCC50**.

El estudiante demostrará lo aprendido sobre los Fundamentos de la Ciencia de Datos, aplicando de forma práctica: la teoría, técnicas y herramientas utilizadas para el análisis y analítica, a lo largo del ciclo de vida de los datos.

Descripción del Proyecto de Analítica

Alcance

Una consultora internacional, con sede en Lima, solicita desarrollar un proyecto de analítica con la finalidad de conocer las tendencias de los videos de YouTube en siete importantes países.

El proyecto responde a la necesidad de su cliente, una importante empresa de marketing digital, que desea obtener respuestas a varios requerimientos de información.

Requerimientos

A continuación, se detalla una lista de preguntas que al finalizar el proyecto deberá dar respuesta.

- **Por Categoría de Videos**

1. ¿Qué categorías de videos son las de mayor tendencia?
2. ¿Qué categorías de videos son los que más gustan? ¿Y las que menos gustan?
3. ¿Qué categorías de videos tienen la mejor proporción (ratio) de “Me gusta” / “No me gusta”?
4. ¿Qué categorías de videos tienen la mejor proporción (ratio) de “Vistas” / “Comentarios”?

- **Por el tiempo transcurrido**

5. ¿Cómo ha cambiado el volumen de los videos en tendencia a lo largo del tiempo?

- **Por Canales de YouTube**

6. ¿Qué **Canales de YouTube** son tendencia más frecuentemente? ¿Y cuáles con menos frecuencia?

- **Por la geografía del país**

7. ¿En qué Estados se presenta el mayor número de “Vistas”, “Me gusta” y “No me gusta”?

Adicionalmente, al cliente le gustaría conocer si:

- ¿Es factible predecir el número de “Vistas” o “Me gusta” o “No me gusta”?
- ¿Los videos en tendencia son los que mayor cantidad de comentarios positivos reciben?

Equipo de Trabajo

Se han conformado ocho equipos de trabajo, cada uno de ellos compuesto por los mejores estudiantes de la carrera de Ciencias de la Computación e Ingeniería de Software de la Universidad Peruana de Ciencias Aplicadas (UPC), divididos en dos secciones (CC51 y CC52).

Cada equipo será responsable de responder a cada uno de los requerimientos según el país que le sea asignado, utilizando el conocimiento y las herramientas aprendidas durante la asignatura de Administración de la Información, y respetando la estructura del entregable, así como su entrega y publicación en la fecha establecida.

Conjunto de Datos

El conjunto de datos motivo de análisis se denomina **Tendencias de las estadísticas de videos de YouTube** (**Trending YouTube Video Statistics)**. Este conjunto de datos es un registro diario de los videos de **YouTube** de mayor tendencia, cuyo contenido incluye varios meses sobre datos de tendencias diarias de videos en los siguientes países:

- EE. UU. (US)
- Gran Bretaña (GB)
- Alemania (DE)
- Canadá (CA)
- Francia (FR)
- Rusia (RU)
- México (MX)
- Corea del Sur (KR)
- Japón (JP)
- India (IN)

Los datos de cada país se encuentran en archivos individuales en formato CSV y la descripción de sus categorías en un archivo de tipo JSON.

Este conjunto de datos, en su versión original se obtiene desde el sitio web Kaggle, sin embargo, para este proyecto, ha sido modificado incorporándole cuatro nuevas columnas:

- **state:** nombre del Estado perteneciente al país (incorporado de forma aleatoria).
- **lat:** latitud geográfica de ubicación del Estado.
- **lon:** longitud geográfica de ubicación del Estado.
- **geometry:** (opcional) registra las coordenadas de las geometrías donde su ubica el Estado dentro del planeta. Es de utilidad si se decide utilizar la librería GeoPandas para la elaboración de mapas.

A cada equipo de estudiantes se le ha asignado un conjunto de datos por país de forma aleatoria, siendo este el resultado:

Nro. Equipo Sección CC51	País asignado	Archivos a descargar para el trabajo
1	Canadá (CA)	CAvideos_cc50.csv CA_category_id.json
2	EE. UU. (US)	USvideos_cc50.csv US_category_id
3	Japón (JP)	JPvideos_cc50.csv JP_category_id.json
4	Alemania (DE)	DEvideos_cc50.csv DE_category_id.json

Nro. Equipo Sección CC52	País asignado	Archivos a descargar para el trabajo
2	México (MX)	MXvideos_cc50.csv MX_category_id.json
3	India (IN)	INvideos_cc50.csv IN_category_id.json
4	Francia (FR)	FRvideos_cc50.csv FR_category_id.json
5	Gran Bretaña (GB)	GBvideos_cc50.csv GB_category_id.json

El conjunto de datos modificado con el que se trabajará se puede descargar desde AQUÍ.

Documento Entregable

Cada equipo de trabajo entregará en el Aula Virtual un único documento de tipo PDF, desarrollando los siguientes temas en el siguiente orden propuesto:

1. **OBJETIVOS DEL PROYECTO**
   Se debe describir cada uno de los objetivos o preguntas a resolver, así como el conocimiento que se pretende estimar o predecir mediante un modelo de datos.

2. **CASO DE ANÁLISIS**

   Explicación sobre el origen de los datos (procedencia de los datos, autor/autores, fecha, país, etc.)

   Casos de uso aplicables (describir, por ejemplo: ¿Para quién sería importante el análisis de estos datos?, ¿Quien o quienes se benefician?, ¿A quien va dirigido el desarrollo del proyecto?)

3. **CONJUNTO DE DATOS**

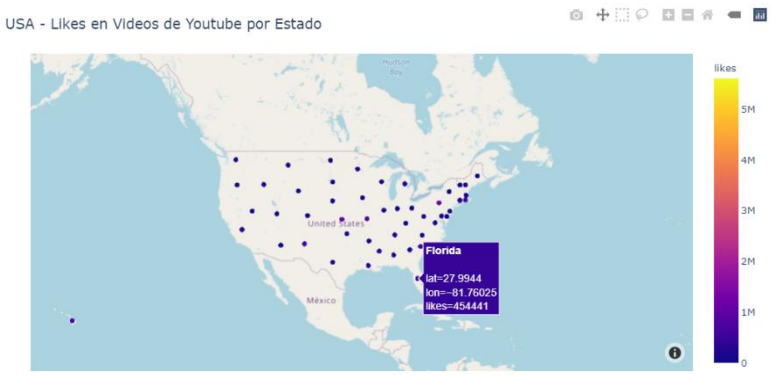
   Descripción de la estructura del conjunto de datos asignado.

4. **ANÁLISIS EXPLORATORIO DE LOS DATOS**

   Este análisis está compuesto por las tareas de carga, inspección, preprocesamiento y visualización de los datos.

   Producto del análisis exploratorio de los datos, con los datos limpios (y a partir del nuevo conjunto de datos generado), se deberá dar respuesta a los requerimientos solicitados.

   Tener en cuenta que cada respuesta deberá estar acompañada de una visualización. Como ejemplo, a continuación, una visualización para el requerimiento número siete:



- ❖ CARGAR LOS DATOS
- ❖ INSPECCIONAR LOS DATOS
  - Los equipos deberán explorar los datos del conjunto de datos asignado, verificando, por ejemplo, estructura, tipo, valores de los datos, nombre de columnas, etc.
- ❖ PRE-PROCESAR LOS DATOS
  - Considerar crear una nueva columna para la descripción de la categoría a partir del archivo json respectivo (si faltase alguna categoría, obtenerla desde el archivo Json de US).
  - Trabajar con las variables trending_date y publish_time como fechas y no como de tipo objeto. Ambas deben tener el mismo formato (yyyy-mm-dd o dd-mm-yyyy). Adicionalmente, se puede crear una columna para la hora de publicación.
  - Verificar datos faltantes. Si el dato faltante es el video_id, entonces, se removerán dichas observaciones. En otro caso, analizar la aplicación de alguna técnica para el tratamiento de los datos faltantes.
  - No olvidar explicar que técnica fue utilizada para eliminar o completar los datos faltantes.
  - Identificación de los datos atípicos u outliers (si los hubiera).
  - Explicación y aplicación de la(s) técnica(s) utilizada(s) para transformar los datos atípicos (si los hubiera).

- ❖ VISUALIZAR LOS DATOS
  - Toda visualización deberá tener un título, una leyenda y de ser necesario, una tabla de datos que complemente su entendimiento.

5. MODELIZAR Y EVALUAR LOS DATOS

- Identificar que variables en el conjunto de datos son susceptibles a ser modeladas.
- Describir el "conocimiento" que se intenta extraer a partir de la aplicación de un modelo de datos.
- Utilizar algún algoritmo innovador (según la pregunta responder) para crear un modelo de datos (e. algoritmo de regresión lineal).
- Obtener métricas y realizar una evaluación del resultado obtenido en el modelo.

6. CONCLUSIONES DEL PROYECTO

Las conclusiones resultan de las respuestas que cada equipo proporcionara por cada uno de los requerimientos del proyecto.

7. ARCHIVAR Y PUBLICAR

- Se deberá contemplar un repositorio en **Github.com** llamado EB-2022-1-CC50 conteniendo dos carpetas:
  - **data:** deberá contener el conjunto de datos inicial y el final resultante (limpio o preparado para análisis).

- **code:** deberá contener los notebooks Python utilizados para el proceso de carga, inspección, pre-procesado y visualización del conjunto de datos.
-

- El archivo. Readme, dentro de GitHub, deberá contemplar:
  - Objetivo del proyecto.
  - Nombre de los alumnos participantes.
  - Breve descripción del conjunto de datos (se puede adjuntar el archivo PDF).
  - Conclusiones.
  - Licencia de uso.

  Guiarse de estos ejemplos de publicaciones de trabajos en GitHub:

  https://github.com/fernandoabcampos/titanic-data-cleaning-and-validation

  https://github.com/navarroyepes/TCVDPRAC2

- Se ha creado un Foro en el Aula Virtual denominado FORO DEL PROYECTO DE ANALITICA, para que los equipos puedan compartir y/o encontrar respuestas a consultas/dudas durante la ejecución del proyecto.

- En el documento entregable, **se deberá incluir el enlace a la cuenta de Github.com** desde donde se accede a la publicación del proyecto.

Consideraciones adicionales

- Se evaluará el orden dentro de la organización del documento, así como la correcta redacción y gramática.
- Se valorarán las respuestas a preguntas que no hayan sido propuestas en la presente evaluación.
- La exposición de las conclusiones del presente trabajo final será en fecha y hora programada durante la semana 15.

Fecha límite de entrega: sábado 25/06/2022 a las 23:59h