Trabajo Práctico Grupal Obligatorio

Parte II

Codo a Codo 4.0 - Big Data / Data Analytics

Junio 2024

Guía para el proyecto en equipo

Pautas específicas para el proyecto

- Los enunciados, consignas, links a datasets y/o todo material necesario para el comienzo del TP, les será proporcionado en el presente documento.
- La versión final de esta etapa del TP deberá ser entregada el 12 de julio como fecha límite.
- El proyecto se calificará como Aprobado o No Aprobado, siguiendo las pautas establecidas en la rúbrica.

Calendario

12 de julio	15 al 19 de julio
Entrega	Publicación de resultados

Contexto, fuentes de datos y documentación recibida

Un análisis preliminar no arrojó irregularidades contables o financieras (gastos, deudas, pérdidas, desbalance de precios, etc.) o caídas significativas en la demanda en el mercado de Paraguay que pudieran explicar su baja tasa de retorno de inversión y, finalmente, se concluye que hubo un fallo de estrategia por parte de los directivos de CMM.

Recibimos un reporte y archivos del equipo paraguayo de auditoría, del que nos resulta pertinente lo siguiente:

- Resumen de ventas en Paraguay: **sales_in_Paraguay.xlsx** (ventas por distribuidor y producto).
- Perfil de los distribuidores en el pais: **distributors_profiles.csv.**
    - id
    - distributor: nombre de la empresa
    - distributor activities: actividades de la empresa
    - years in the contruction market: años activo en el mercado de la construcción

El equipo de comercialización de CMM nos proporciona los resúmenes y reporte de exportación de materiales a Paraguay, a partir de lo que podemos obtener:

- **exports_to_Paraguay.csv.**

Nuestro equipo de recolección datos proporciona información adicional para ampliar el contexto.
A partir de datos de la CNAEP (Clasificación Nacional de Actividades Económicas del Paraguay), Reportes de UNFPA (Fondo de Población de las Naciones Unidas), la DGEEC (Dirección General de Estadística, Encuestas y Censos de Paraguay), y el artículo: Mapeo de industrias del Paraguay registradas en el Ministerio de Industria y Comercio (https://revistacientifica.sudamericana.edu.py/index.php/scientiamericana/article/view/175/194), obtenemos el siguiente resumen:

- **locations_profiles.csv:** principales actividades económicas por ciudad
    - PYid: identificador de la ciudad
    - id: identificador de la ciudad (presente sólo si existe un distribuidor de la compañía en esa ciudad)
    - location: nombre de la ciudad
    - department: departamento
    - activities: principales actividades económicas de la ciudad

Una vez realizadas las tareas de limpieza, se asumen como certeros todos los datos proporcionados.

Datos

- **sales_in_Paraguay.xlsx**
- **distributors_profiles.csv**
- **exports_to_Paraguay.csv**
- **locations_profiles.csv**

No todos los archivos proporcionados son necesarios para la resolución analítica o visual. Debemos decidir cuáles son útiles para estas tareas, y cuáles nos proporcionan información sobre el contexto.
Luego de una exploración inicial, ya deberíamos ser capaces de identificar cuál fue el problema. Nuestra misión es demostrarlo de manera analítica y/o visual . Adicionalmente podemos investigar y explicar el motivo, analizando el contexto.

Pautas de entrega

- Debe entregar al menos un archivo de Google Colaboratory con los resultados obtenidos.
- Al importar las fuentes de datos a Google Colaboratory, debe utilizar el procedimiento explicado en el siguiente video.
- Se permitirá el uso de Google Sheets para una previsualización o control de los resultados obtenidos, pero todos los procesos deberán realizarse con librerías de Python:
    - Limpieza y modelado de datos (Numpy y/o Pandas).
    - Exploración (Numpy/Pandas y Matplotlib y/o Seaborn)
    - Visualización (Numpy/Matplotlib y/o Seaborn)

  Si utiliza Google Sheets para explorar el contexto, debe incluir en la justificación una captura de imagen de la tabla o resultados obtenidos.
- Los nuevos archivos generados deberán compartirse con permisos de edición con el instructor, y agregados a la página de links de archivo de Looker Studio de la Parte I del TP.
- Todos los archivos deben ser renombrados con el formato: **c240xx - Grupo xxx - TP Px - Título**. Por ejemplo:
    - Google Sheets: c24000 - Grupo 999 - TP P1 - Limpieza y modelado de datos CMM Latam - Ene-Abr 2024

- Looker: Colab: c24000 - Grupo 999 - TP P1 - Reporte CMM Latam - Ene-Abr 2024
- Google Sheets: c24000 - Grupo 999 - TP P2 - Auditoría CMM - Paraguay - Jun 2024
- Colab: c24000 - Grupo 999 - TP P2 - Analítica auditoría CMM - Paraguay - Jun 2024
- PDF: c24000 - Grupo 999 - TP P1 / PDF: c24000 - Grupo 999 - TP P2

- El TP deberá ser entregado en el siguiente formulario. En él se solicitará:
  - Datos de la comisión
  - Nº de Grupo
  - Nº de DNI de los integrantes del equipo **que llegaron a la instancia de entrega final.** Deberá ingresar al menos un Nº de DNI.
  - Subir una versión PDF del trabajo en Google Colaboratory y una versión PDF del trabajo en Looker Studio.
  - Se permitirá 1 sola respuesta, no editable.

ANEXO: Guía de actividades

- Importación de librerías y obtención de Datos
- Inspección preliminar y limpieza
  - Relevancia de columnas
  - Tipos de Datos
  - Detección y eliminación de duplicados y/o filas/columnas inapropiadas o innecesarias.
- Modelado
  - Obtención de los sets de datos apropiados
- Análisis
  - Obtener una respuesta al problema y analizar el contexto
- Visualización
  - Respaldar y mostrar los resultados de manera simple y amigable para un público sin grandes conocimientos técnicos.