Universidad del Desarrollo
Magister en Data Science
Profesor: Víctor Navarro Aránguiz

Python para Data Science

Proyecto Final

15 de julio de 2024

Fecha de entrega: 04 de Agosto de 2024, 23:59:59
Objetivos:

- Aplicar los conceptos aprendidos en clases.
- Realizar el análisis de un dataset de su preferencia.
- Familiarizarse con el manejo de versiones a través de Git.

Se espera que a partir del dataset la/el estudiante defina un flujo de trabajo y sea capaz de limpiar el datset, crear nuevas variables a partir de las existentes, visualizar e interpretar los resultados de manera efectiva.

Puntaje total: 24 puntos

Pasos del Proyecto

1. Selección del Dataset

Como primera tarea se debe definir el dataset con el que se desea trabajar. Para ello, es importante poner énfasis en los siguientes puntos:

- ***¿Es de mi interés el dataset?***: De preferencia, elijan un dataset sobre un tema que les guste.
- ***¿Qué tipo de preguntas puedo responder con el dataset?***: Es importante definir qué es lo que se quiere hacer con el dataset.
- ***¿Puedo realizar visualizaciones con el dataset?***: Piensen en las posibles visualizaciones que quieran hacer, y revisen si son factibles con la información presente dentro del dataset.
- ***¿El tamaño del dataset es suficiente?***: Asegúrense de que el dataset tenga una cantidad de datos (filas) suficientes para realizar el análisis. Muy pocas filas (< 500) no es un trabajo muy desafiante, mientras que muchas filas (> 10 Millones) podría suponer tiempos muy largos a la hora de trabajar con el dataset.

El dataset puede ser cualquiera que estimen conveniente en base a los puntos anteriores. Si gustan pueden buscarlo en las siguientes páginas:

- https://www.kaggle.com/datasets: Repositorio de datasets de Kaggle ordenados por categoría
- https://github.com/imfd/awesome-data-chile: Lista de fuentes públicas de datos sobre Chile. Actualizado al 2023.
- https://www.ine.gob.cl/estadisticas/: Datos del Instituto Nacional de Estadísticas
- https://huggingface.co/datasets: Datasets alojados por Hugging Face, principalmente sobre procesamiento de lenguaje natural (NLP en inglés).

2. Control de versiones (3 puntos)

Se usará GitHub para el control de versiones. Será importante ir dejando registro de los cambios que se van realizando usando *commit messages* de manera pertinente.

En cuanto a la estructura del proyecto, se espera que se mantenga un orden separando las tareas a realizar en distintos notebooks. Por ejemplo, que exista la siguiente estructura:

```
- 01_Analisis_Exploratorio.ipynb # análisis inicial sobre los datos
- 02_Limpieza_de_Datos.ipynb     # limpieza en base al análisis anterior
- 03_Visualización.ipynb         # visualizaciones que desean mostrarse
- ...                            # Otros
```

Como referencia y recomendación para estructurar el proyecto pueden usar la plantilla Cookiecutter Data Science.

3. Inspección de Datos (5 puntos)

Se debe realizar una inspección inicial del dataset para entender su estructura, contenido y calidad.

Acciones a realizar

- Describir las columnas que tiene el dataset, mencionando qué representa y el tipo de dato.
- Examinar el dataset para identificar posibles problemas como valores faltantes o anomalías, documentando los hallazgos.

A modo de ejemplo, si el dataset es sobre una tienda y este tiene las columnas `producto` y `precio`, dentro del notebook se debería documentar algo del estilo:

```
- producto (string): Representa el producto de una tienda.
- precio (float): El precio de un producto.
...

Hemos encontrado que X productos se encuentran duplicados y que Y de ellos no tienen precio.
(mostrar código)
```

4. Limpieza de Datos (6 puntos)

En base a la inspección anterior, limpiar el dataset para garantizar que esté listo para el análisis.

Acciones a realizar

- Manejar casos de posibles duplicados, valores faltantes (y outliers, opcional). Si considera que parte del dataset ya está limpio, explicar su razón.
- Justificar cada paso del proceso de limpieza detalladamente

Un buen proceso de limpieza es aquel que permite obtener un conjunto de datos coherente y libre de errores, facilitando así un análisis preciso y confiable.

5. Creación de Variables a partir de las existentes (4 puntos)

Generar nuevas variables que puedan proporcionar información adicional o mejorar el análisis.

Acciones a realizar

- Identificar variables existentes que se pueden combinar o transformar para crear nuevas variables significativas.
- Crear y documentar nuevas variables, justificando su utilidad y explicando cómo mejoran el análisis o proporcionan información adicional.
- Proveer ejemplos concretos de cómo las nuevas variables pueden ser utilizadas en el análisis posterior.

Este enfoque permitirá enriquecer el dataset y descubrir relaciones y/o patrones que no son evidentes con las variables originales.

6. Visualización (4 puntos)

Crear y personalizar gráficos que representen adecuadamente los datos, utilizando tanto variables originales como creadas.

Acciones a realizar

- Generar gráficos variados que muestren diferentes aspectos del dataset, tales como distribuciones, relaciones y tendencias.
- Realizar agrupamientos y agregaciones para visualizar ciertas variables y sus interacciones.
- Justificar la creación de cada gráfico y la elección de las técnicas de visualización utilizadas, explicando cómo cada uno contribuye a una mejor comprensión del dataset.
- Asegurarse de que los gráficos sean claros, etiquetados adecuadamente y visualmente atractivos.

Este enfoque permitirá una interpretación más intuitiva y detallada de los datos, facilitando la comunicación de los hallazgos.

7. Aplicación Práctica del Dataset (4 puntos)

Proponer una aplicación práctica basada en los datos analizados.

Acciones a realizar

- Desarrollar una propuesta clara y detallada que utilice los hallazgos del análisis de datos.
- Describir cómo se implementaría la propuesta en un contexto real, incluyendo los pasos necesarios para su ejecución, así como los posibles desafíos y limitaciones.

Pauta de Evaluación

Control de Versiones (3 puntos)

- **3 puntos**: Uso consistente y detallado de *commit messages*; estructura del proyecto bien organizada con notebooks claramente separados.
- **2 puntos**: Uso regular de *commit messages*; estructura del proyecto organizada pero con algunas inconsistencias menores.
- **1 punto**: Uso ocasional de *commit messages*; estructura del proyecto algo desorganizada.
- **0 puntos**: No se utilizó control de versiones o la estructura del proyecto es completamente desorganizada.

Inspección de Datos (5 puntos)

- **5 puntos**: Descripción completa y clara de todas las columnas; identificación detallada y justificada de problemas como valores faltantes y anomalías.
- **4 puntos**: Descripción clara de la mayoría de las columnas; identificación de la mayoría de los problemas con algunas justificaciones.
- **3 puntos**: Descripción parcial de las columnas; identificación de algunos problemas pero con justificaciones insuficientes.
- **2 puntos**: Descripción incompleta de las columnas; pocos problemas identificados y con justificaciones débiles.
- **1 punto**: Descripción muy limitada de las columnas; problemas apenas identificados sin justificaciones.
- **0 puntos**: No se realizó la inspección de datos.

Limpieza de Datos (6 puntos)

- **6 puntos**: Limpieza completa y justificada de duplicados, valores faltantes y outliers; cada paso está claramente documentado y explicado.
- **5 puntos**: Limpieza completa de duplicados y valores faltantes; buena documentación y justificación de los pasos.
- **4 puntos**: Limpieza mayormente completa con algunas áreas sin justificar completamente; documentación adecuada.
- **3 puntos**: Limpieza parcial con varios pasos no justificados o documentados.
- **2 puntos**: Limpieza limitada con muchos errores o pasos faltantes; documentación insuficiente.
- **1 punto**: Limpieza muy superficial; apenas se realizó algún paso de limpieza.
- **0 puntos**: No se realizó la limpieza de datos.

Creación de Variables a partir de las existentes (4 puntos)

- **4 puntos**: Identificación creativa y justificada de nuevas variables; documentación clara de su utilidad y ejemplos concretos.
- **3 puntos**: Identificación de nuevas variables con buenas justificaciones; documentación adecuada.
- **2 puntos**: Algunas nuevas variables identificadas pero con justificaciones insuficientes o falta de ejemplos concretos.
- **1 punto**: Pocas o ninguna nueva variable identificada; justificaciones débiles.
- **0 puntos**: No se crearon nuevas variables.

Visualización (4 puntos)

- **4 puntos**: Gráficos variados y bien diseñados que muestran diferentes aspectos del dataset; claras justificaciones y etiquetado adecuado.
- **3 puntos**: Gráficos adecuados con buena variedad; justificaciones presentes pero con algunas áreas mejorables.
- **2 puntos**: Gráficos limitados en variedad o calidad; justificaciones insuficientes o etiquetado deficiente.
- **1 punto**: Muy pocos gráficos y de baja calidad; casi sin justificaciones ni etiquetado adecuado.
- **0 puntos**: No se realizaron visualizaciones.

Aplicación Práctica del Dataset (4 puntos)

- **4 puntos**: Propuesta clara, detallada y bien justificada; descripción completa de la implementación en un contexto real, con identificación de desafíos y beneficios.
- **3 puntos**: Propuesta clara y bien justificada; buena descripción de la implementación con algunos detalles menores faltantes.
- **2 puntos**: Propuesta adecuada pero con justificaciones y descripciones insuficientes.
- **1 punto**: Propuesta muy limitada con pocas justificaciones y descripciones vagas.
- **0 puntos**: No se propuso ninguna aplicación práctica.