Material Imprimible

Curso de Data Science

Módulo 2 - Exploración y procesamiento de datos

Contenidos:

- Librerias en Python
- DataFrame
- Tipos de datos y gestión de nulls
- LabelEncoder y OneHotEncoder
- Visualización estadística exploratoria

Librerías en Python

Pandas, NumPy, Scikit-Learn y Seaborn son algunas de las librerías de Python más populares utilizadas en el campo de la ciencia de datos. Cada una de estas librerías tiene un conjunto único de herramientas y funciones que son esenciales para el análisis y la visualización de datos.

Pandas es una librería de Python utilizada principalmente para la manipulación y análisis de datos en formato de tabla. Permite la carga de datos desde múltiples fuentes y formatos, incluyendo CSV, Excel, SQL y JSON. Con Pandas, los datos pueden ser transformados, filtrados, agrupados y unidos de manera eficiente, lo que lo convierte en una herramienta esencial para la limpieza y preparación de datos.

Link oficial: https://pandas.pydata.org/

NumPy, por otro lado, es una librería fundamental para el cálculo numérico en Python. Proporciona una estructura de matriz multidimensional eficiente para trabajar con datos numéricos, lo que permite la realización de cálculos matemáticos complejos de manera rápida y eficiente. NumPy es especialmente útil para el procesamiento de imágenes y el análisis de señales.

Link oficial: https://numpy.org/

Scikit-Learn es una librería de aprendizaje automático de Python que proporciona herramientas para la construcción y entrenamiento de modelos de clasificación, regresión y agrupamiento, entre otros. Es muy popular en la comunidad de ciencia de datos debido a su facilidad de uso, su compatibilidad con NumPy y Pandas, y su amplia selección de modelos de aprendizaje automático predefinidos.

Link: https://scikit-learn.org/stable/

Seaborn es una librería de visualización de datos que se basa en Matplotlib. Proporciona un conjunto de herramientas y funciones para la creación de gráficos estadísticos atractivos y eficaces, que permiten la visualización y comunicación de los resultados del análisis de datos de una manera clara y concisa.

Link: https://seaborn.pydata.org/

DataFrame

Un DataFrame de Pandas es una estructura de datos bidimensional, similar a una tabla de base de datos o una hoja de cálculo. Consiste en filas y columnas, donde cada columna puede tener un tipo de datos diferente (por ejemplo, cadenas, números o fechas). Los DataFrames de Pandas son muy útiles para el análisis de datos y se utilizan ampliamente en el campo de la ciencia de datos.

Estos objetos se pueden crear de diversas formas, como la lectura de datos desde un archivo CSV o Excel, o mediante el despliegue desde una lista o diccionario de Python. Una vez creado, se pueden manipular los datos de diversas formas, como la selección de filas y columnas, la agregación de datos, la unión de varios DataFrames y la creación de nuevas columnas a partir de operaciones con las columnas existentes.

Pandas también ofrece una amplia gama de funciones para la limpieza y preparación de datos, como la eliminación de valores nulos o duplicados, la transformación de datos, la eliminación de outliers y la manipulación de fechas y horas.

Tipos de datos y gestión de nulls

Dentro de un DataFrame los datos que se manipulan pueden ser:

- Object: son las columnas categoricas con variables alfanuméricas (Por ejemplo “Argentina”)
- Int: son columnas de valores enteros numéricos (Por ejemplo: [0 1 2 3 ...])
- Float: son columnas con valores numéricos decimales (Por ejemplo: [0.68 0.70 ..])
- Bool: son columnas True or False

Cada columna puede contener una cantidad de datos nulos que no son tolerados por los algoritmos de data science y siempre se recomienda hacer una gestión de los mismos. Se puede rellenar con algún valor fijo o hacer alguna técnica de reemplazo en específico (por ejemplo, definir el valor de la mediana en donde falten datos)

LabelEncoder y OneHotEncoder

En el caso de las variables de Texto (object), antes de cualquier análisis predictivo modelizado debemos llevarla a valores numéricos, para que el algoritmo correspondiente pueda interpretarla. Se utilizan dos técnicas para esto:

- Label Encoder es una técnica que asigna un número entero único a cada categoría en una variable categórica. Por ejemplo, si tenemos una variable "Color" con tres categorías: "Rojo", "Verde" y "Azul", el Label Encoder asignaría los números 0, 1 y 2 a cada una de estas categorías. De esta forma, se puede convertir la variable categórica en una variable numérica y se puede utilizar en modelos de aprendizaje automático. Sin embargo, el Label Encoder no es la mejor opción para todas las situaciones, ya que puede crear un orden numérico implícito entre las categorías que no existe en la realidad. En algunos casos, esto puede llevar a resultados erróneos en los modelos de aprendizaje automático.

- El OneHot Encoder es una técnica que crea una columna separada para cada categoría en una variable categórica y asigna un valor binario de 1 o 0 para indicar si la observación pertenece a esa categoría o no. Por ejemplo, si tenemos una variable "Color" con tres categorías: "Rojo", "Verde" y "Azul", el encoder crearía tres nuevas columnas: "Color_Rojo", "Color_Verde" y "Color_Azul". Si una observación pertenece a la categoría "Rojo", se le asignaría el valor 1 en la columna "Color_Rojo" y 0 en las otras dos columnas.

Visualización estadística exploratoria

La visualización exploratoria se puede realizar en múltiples herramientas, pero una de la más utilizadas, es mediante Seaborn en Python. Esta librería nos permite hacer diversos análisis para entender la calidad y distribución de los datos.

Los tipos de análisis más famosos son:

- Distribución univariada de cada variable mediante un histograma
- Matriz de correlación para analizar el vínculo de las variables entre si
- Gráficos de distribución y barra para comprender el peso de cada item
- Series de tiempo