Tecnológico de Monterrey

Notas módulo 1 análisis de datos

Jorge Luis Tapia Peñaloza-A01793013

11 de oct del 2022

Ciencia de datos

Profesores: Jobish Vallikavungal/Julio César Galindo López

La variable que nos interesa predecir es la variable objetivo

Existen varios formatos para un conjunto de datos: .csv, .json, .xlsx, etc. El conjunto de datos se puede almacenar en diferentes lugares, en su máquina local o, a veces, en línea.

La biblioteca Pandas es una herramienta útil que nos permite leer varios conjuntos de datos en un marco de datos; nuestras plataformas de portátiles Jupyter tienen una biblioteca de Pandas incorporada, por lo que todo lo que tenemos que hacer es importar Pandas sin instalar.

Un dataframe es un objeto que representa una tabla.

Sentencias importantes

df.head() muestra los primeros 5 registros
df.tail() muestra los últimos 5 registros
df.describe() es un análisis estadístico de cada variable
df.columns columnas del dataframe
pd.read_csv leer un archivo de texto plano, separado por un carácter

La normalización es una forma de llevar todos los datos a un rango similar, para una comparación más útil.

El análisis de datos y, en esencia, la ciencia de datos, nos ayuda a desbloquear la información y los conocimientos.
a partir de datos sin procesar, para responder a nuestras preguntas.
Por lo tanto, el análisis de datos juega un papel importante al ayudarnos a descubrir información útil.