

Curso Aprendizaje Automático (ML) con Python

Nelson López Centeno

Exploración y preparación de datos

- ☐ Análisis Exploratorio de Datos (EDA)

- ☐ Preparación de datos

 - ☐ Valores faltantes

 - ☐ Conversión de variables discretas

 - ☐ Escalado



Demos / Ejercicios

- ☐ Datasets

 - ☐ Titanic

 - ☐ Car Features and MSRP

 - ☐ California Housing Prices (1990)

- ☐ Librerías

 - ☐ Pandas

 - ☐ scikit-learn

 - ☐ DataPrep

 - ☐ YData Profiling

Análisis Exploratorio de Datos (EDA)

El análisis exploratorio de datos (EDA) es utilizado por los científicos de datos para analizar e investigar conjuntos de datos y resumir sus principales características, a menudo utilizando métodos de visualización de datos.

Proporciona una mejor comprensión de las variables del conjunto de datos y las relaciones entre ellas.

Puede ayudar a determinar si las técnicas estadísticas que se están considerando para el análisis de datos son apropiadas.

Análisis Exploratorio de Datos (EDA)

Tipos de EDA

Univariable

- no gráfico: estadística descriptiva
- gráfico: histogramas, box plots

Multivariable

- no gráfico: matriz de correlación
- gráfico: scatter plot, heat map

Preparación de datos

Valores faltantes

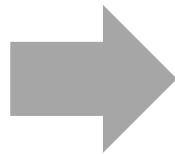
- Eliminar filas o columnas
- **Imputación**
 - Valor constante
 - Media
 - Mediana
 - Moda
 - Multivariable

Preparación de datos

Conversión de variables discretas

- Codificar el texto como valores numéricos
- Dummies (One-hot encoding)

ocean_proximity
1 H OCEAN
INLAND
NEAR OCEAN
ISLAND



1 H OCEAN	INLAND	NEAR OCEAN	ISLAND
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Preparación de datos

Escalado

La mayoría de los algoritmos de ML no funcionan bien si las diferentes características tienen diferentes escalas.

- Min-max (normalización)
- Estandarización



Ejercicios

EDA con la librería ydata-profiling

EDA con la librería DataPrep

Preparación de los datos del Titanic con Pandas

Preparación de los datos del Titanic con scikit-learn

