MACHINE LEARNING

CIENCIAS DE LA COMPUTACIÓN
M.Sc. Richard F. Fernández Vásquez

¿Cuál es la importancia de realizar un preprocesamiento de datos?

Logro Unidad 1

Al finalizar la unidad, el alumno es capaz de aplicar adecuadamente técnicas de pre procesamiento de datos para posibilitar la implementación de una solución de Machine Learning para un problema del mundo real.

Contenido 3

- Identificación y tratamiento de valores faltantes
- Identificación y tratamiento de valores atípicos
- Transformación de datos
- Caso de Aplicación

Caso de Aplicación: Préstamos para viviendas

Housing Perú

La empresa Housing Perú es una empresa que proporciona préstamos hipotecarios.

La empresa valida la elegibilidad del préstamo después de que el cliente pasa por un proceso de elegibilidad basándose en su información que completa en un formulario en tiempo real.

Por lo tanto, cuanto más precisos seamos al predecir los clientes elegibles, más beneficioso será para la empresa.

Caso de Aplicación: Préstamos para viviendas

Variables:

-Loan_ID: ID de la solicitud de préstamo

-Gender: género del cliente (Female/Male)

-Married: estado civil del cliente, casado o no casado (Yes/No)

-Dependents: número de dependientes (0, 1, 2, 3+)

-Education: nivel de educación (Graduate/Under Graduate)

-Self_Employed: trabajador independiente (Yes/No)

-ApplicantIncome: ingresos del solicitante

-CoapplicantIncome: ingresos del cosolicitante

-LoanAmount: monto del préstamo expresado en miles

-Loan_Amount_Term: plazo del préstamos expresado en meses

-Credit_History: historial crediticio

-Property_Area: Urban/Semi Urban/ Rural

-Loan_Status: status del préstamo, aprobado o no aprobado (Y/N)

Caso de Aplicación: BUPA

BUPA Medical Research Ltd. Las primeras 5 variables son resultados de pruebas sanguíneas que se piensan pueden ser sensitivas (y posibles predictores) ante transtornos hepáticos producidos por un consumo excesivo de alcohol. Cada línea de la base de datos BUPA constituye un registro de un individuo de sexo masculino.

Información de atributos:

V1 volumen corpuscular

V2 fosfatosa alcalina

V3 alamine aminotransferase

V4 aspartate aminotransferase

V5 gamma-glutamyl transpeptidase

V6 numero de bebidas alcoholicas

V7 1(hígado enfermo) 2(hígado sano)

Trabajo en clase

- Buscar una base de datos en https://www.kaggle.com/datasets, y desarrollar los siguientes puntos en un notebook:
  - Link de la base de datos
  - Objetivo del caso
  - Análisis de valores faltantes
  - Análisis de valores atípicos
  - Conclusiones

Fuentes de datos

- https://www.datosabiertos.gob.pe/
- https://estadisticas.bcrp.gob.pe/estadisticas/series/
- http://iinei.inei.gob.pe/microdatos/
- https://www.inei.gob.pe/estadisticas/indice-tematico/economia/
- https://www.sbs.gob.pe/app/pp/seriesHistoricas2/paso1.aspx
- https://www.kaggle.com/datasets
- https://archive.ics.uci.edu/ml/datasets.php

CONSULTAS

pcsirife@upc.edu.pe