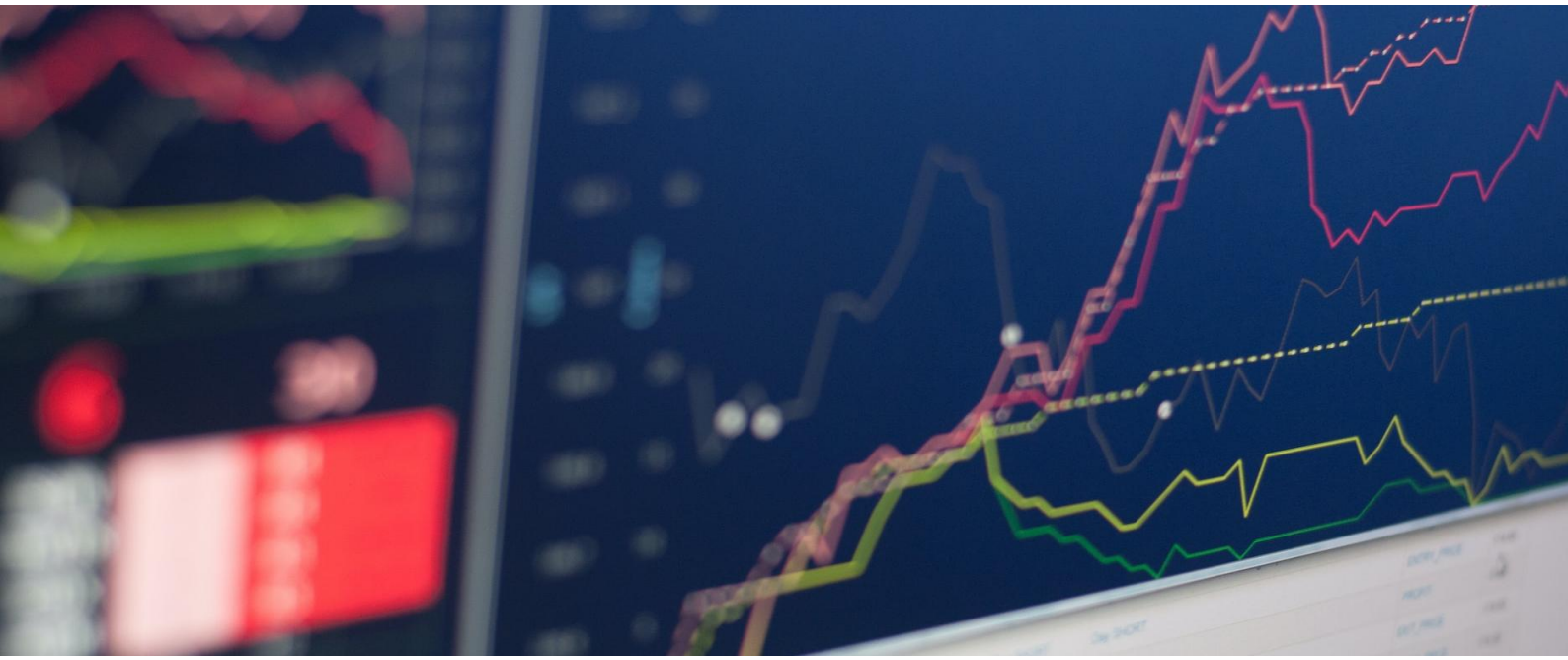
# Práctica 1

# Análisis y Preprocesamiento de Datos



## Objetivos

El objetivo de esta práctica es introducir al análisis y preprocesamiento de los datos.

## Temas

- Atributos. Identificación, clasificación y análisis.
- Representaciones Gráficas. Diagrama de Barra, de Caja, de Dispersión e Histograma.
- Correlación.
- Normalización y estandarización de Datos.

## Lectura

Cap. 4 del Libro Introducción a la Minería de Datos de Hernández Orallo.

**Dataset de Novel de Obesidad**

El dataset **"Estimation of Obesity Levels Based on Eating Habits and Physical Condition"** reúne información de 2.111 individuos de México, Perú y Colombia. El objetivo es predecir el **nivel de obesidad** de cada persona a partir de hábitos alimenticios, características físicas y estilo de vida.

El conjunto contiene **17 atributos** (16 predictivos y 1 variable objetivo) que abarcan aspectos relacionados con la edad, el peso, la estatura, los antecedentes familiares, el consumo de ciertos alimentos y bebidas, el uso de tecnología, la actividad física y los medios de transporte. El nivel de obesidad está clasificado en siete categorías, que van desde "peso insuficiente" hasta "obesidad tipo III".

| Atributo | Descripción |
|---|---|
| **Gender** | Género del individuo (Male/Female) |
| **Age** | Edad en años |
| **Height** | Altura en metros |
| **Weight** | Peso en kilogramos |
| **family_history_with_overweight** | Si tiene antecedentes familiares de sobrepeso |
| **FAVC** | Frecuencia de consumo de alimentos con alto contenido calórico |
| **FCVC** | Frecuencia de consumo de vegetales (1 a 3) |
| **NCP** | Número de comidas principales al día |
| **CAEC** | Consumo de alimentos entre comidas (no, Sometimes, Frequently, Always) |
| **SMOKE** | Si fuma o no |
| **CH2O** | Consumo de agua diario en litros |
| **SCC** | Si controla el consumo de calorías |
| **FAF** | Frecuencia de actividad física (horas/semana) |
| **TUE** | Tiempo de uso de dispositivos electrónicos (horas/día) |
| **CALC** | Consumo de alcohol (no,Sometimes,Frequently,Always) |
| **MTRANS** | Medio de transporte principal (Automobile,Bike,Motorbike,Public_Transportation,Walking) |
| **NObesity (Target)** | Nivel de obesidad (Insufficient_Weight, Normal_Weight, Overweight_Level_I, Overweight_Level_II, Obesity_Type_I, Obesity_Type_II, Obesity_Type_III) |

### *Ejercicio 1*

Complete la tabla indicando cuántos atributos corresponden a cada del dataset de obesidad.

| Tipo de atributo | | Cantidad |
|---|---|---|
| *Cuantitativo o numérico* | *Discreto* | |
| | *Continuo* | |
| *Cualitativo o categórico* | *Nominal* | |
| | *Ordinal* | |

### *Ejercicio 2*

Proponga una tarea de clasificación y una tarea de regresión que puedan realizarse a partir de los datos del dataset de obesidad.

### *Ejercicio 3*

Indique qué tipo de información brindan las siguientes representaciones gráficas:

a) Diagrama de Barras
b) Histograma
c) Diagrama de caja
d) Diagrama de dispersión

Luego, genere al menos un ejemplo de cada representación usando el dataset de obesidad y explique cómo interpretar cada uno.

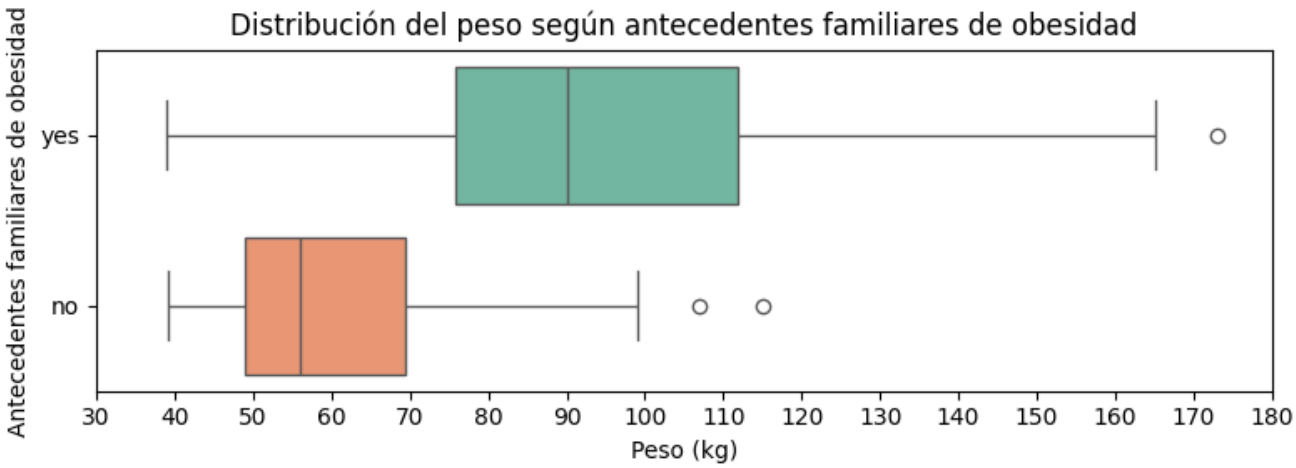
### *Ejercicio 4*

Complete el siguiente cuadro y dibuje el diagrama de caja del atributo **"weight"**

| Medida | Valor |
|---|---|
| Mínimo | |
| Máximo | |
| Q1 | |
| Q2 o mediana | |
| Q3 | |
| RIC | |
| Bigote superior | |
| Bigote inferior | |
| Intervalos de valores atípicos leves | |
| Valores atípicos leves | |
| Intervalos de valores atípicos extremos | |
| Valores atípicos extremos | |

### *Ejercicio 5*

Los valores del atributo peso (**"weight")** fueron agrupados según el atributo de antecedente de obesidad familiar ("**family_history_with_overweight**"). La figura muestra los diagramas de caja correspondientes.



Complete el siguiente cuadro y responda verdadero o falso justificando cada afirmación según los valores obtenidos:

| Medida | yes | no |
|---|---|---|
| Mínimo | | |
| Máximo | | |
| Q1 | | |
| Q2 | | |
| Q3 | | |
| RIC | | |
| Bigote Inferior | | |
| Bigote Superior | | |

a) Al menos el 25% de las personas con antecedentes familiares de obesidad pesan más de 100 kg.
b) Es atípico que una persona sin antecedentes familiares de obesidad (no) pese más de 115 kg.
c) La mediana del peso de las personas sin antecedentes familiares (no) es menor que 60 kg.
d) Todos los valores atípicos para personas con antecedentes familiares de obesidad son leves.

### *Ejercicio 6*

Discretice el atributo del consumo de agua diario en litros (**"CH2O"**) en tres intervalos: **Bajo, Medio y Alto**. Indique en la tabla los intervalos utilizados y la cantidad respectivas de ejemplos de cada uno al discretizar por rango y por intervalo. Luego, explique porque los ejemplos no quedaron divididos en intervalos con la misma cantidad de valores.

| | Rango | | | Intervalo | | |
|---|---|---|---|---|---|---|
| | Bajo | Medio | Alto | Bajo | Medio | Alto |
| **Intervalos** | | | | | | |
| **Cantidad de Valores** | | | | | | |

---

**Dataset Titanic**

El dataset del Titanic, es un conjunto de datos clásico utilizado para tareas de clasificación binaria. Se basa en información histórica de los pasajeros del RMS Titanic, que se hundió en 1912 tras chocar con un iceberg. El objetivo principal es predecir si un pasajero sobrevivió o no, basado en características como su clase social, edad, género y otros factores socioeconómicos.

El RMS Titanic se hundió el 15 de abril de 1912 durante su viaje inaugural. De las aproximadamente 2,224 personas a bordo, más de 1,500 murieron, convirtiendo este evento en uno de los naufragios más mortíferos en tiempos de paz.

| Atributo | Descripción |
|---|---|
| **PassengerId** | Identificador único del pasajero |
| **Survived** | Variable objetivo: Indica si el pasajero sobrevivió (1) o no (0) |
| **Pclass** | Clase del ticket (proxy de estatus socioeconómico) |
| **Name** | Nombre completo del pasajero, incluyendo títulos (ej. Mr., Mrs.) |
| **Sex** | Género del pasajero (male, female) |
| **Age** | Edad en años (puede ser fraccional para niños) |
| **SibSp** | Número de hermanos/esposos a bordo |

| | |
|---|---|
| **Parch** | Número de padres/hijos a bordo |
| **Ticket** | Número del ticket |
| **Fare** | Tarifa pagada por el ticket |
| **Cabin** | Número de cabina (muchos faltantes) |
| **Embarked** | Puerto de embarque (C, Q, S) |

### *Ejercicio 7*

Realice las siguientes tareas para preparar el dataset para que pueda ser utilizado para entrenar modelos de redes neuronales.

a) Visualice las primeras 5 filas y el resumen estadístico.
b) Identifique los atributos (columnas) con valores nulos y su porcentaje.
c) Analizar los valores faltantes y discutir cuales serían las alternativas posibles para tratarlos.
d) Los nombres de los pasajeros van acompañados de títulos que pueden ser importantes para la interpretación de los datos o para completar información faltante:
    i. Extrae el título (como Mr, Miss, Mrs, Master, etc.) del nombre de cada pasajero y crea una nueva columna llamada **Title**.
    ii. Unifica los valores para que queden **Mr**, **Miss** (Mlle, Ms), **Mrs** (Mme), **Master**, **Others** (resto).
    iii. Computa las edades faltantes utilizando la edad promedio por categoría.
e) Complete los valores faltantes utilizando las estrategias planteadas en los puntos anteriores.
f) Cree un nuevo atributo **FamilySize** que contabiliza los integrantes de familia a partir de los atributos **SibSp** (hermanos y esposo) y **Parch** (padres e hijos). No olvidar contar a la persona.
g) Numerice los atributos categóricos: **Sex**, **Embarked**, y **Title**.
h) Discuta y responda ¿Por qué one-hot encoding podría ser preferible a label encoding para el atributo **Pclass**?
i) Visualiza distribuciones: Histograma de **Age**, gráfico de barras para **Survived** por **Sex** y **Pclass.**

### *Ejercicio 8*

Calcule la correlación lineal entre los atributos **"Fare"** (Tarifa) y **"PClass"** (clase del ticket). Indique la intensidad de la correlación (no hay correlación/débil/fuerte) y el tipo (positiva/negativa). Explique el significado del valor de correlación obtenido.

| | Fare/PClass | PClass |
|---|---|---|
| **Valor** | | |
| **Intensidad** | | |
| **Tipo** | | |
| **Significado** | | |

### *Ejercicio 9*

Realice un análisis sobre los valores de los atributos del dataset automobile.csv. Para cada atributo que no pueda ser procesado directamente, indique que problema tiene (valores nulos o vacíos, valores categóricos, valores atípicos o outliers, etc.) y como solucionarlo.

### *Ejercicio 10*

Dada la siguiente tabla con mediciones de 2 características correspondientes a mediciones de altura y peso de personas:

| **Altura** | 1.65 | 1.81 | 1.70 | 1.62 | 1.74 | 1.70 | 1.80 | 1.73 | 1.68 |
|---|---|---|---|---|---|---|---|---|---|
| **Peso** | 75 | 86 | 82 | 78 | 77 | 87 | 90 | 83 | 80 |

a) Aplique las siguientes normalizaciones y grafíquelas con un diagrama de caja:

MinMax: $\frac{x_i - min(x)}{max(x) - min\,(x)}$ Standard: $\frac{x_i - media(x)}{stddev(x)}$ Robust: $\frac{x_i - Q1(x)}{Q3(x) - Q1(x)}$

b) Agregue la siguiente medición (2.20, 120) y repita el punto a)

c) Compare los diagramas de caja entre las normalizaciones de los puntos a) y b) y comente las diferencias.