**Análisis de Base de Datos**

**Objetivo:**  
Cada grupo deberá seleccionar una base de datos, realizar un análisis exploratorio básico, comprender su estructura y contenido, y presentar un informe con los hallazgos más relevantes.

Cada grupo debe realizar las siguientes tareas con la base de datos seleccionada:

1. **Selección de la Base de Datos**

* Nombre de la base de datos: bank-full
* Fuente (URL, repositorio, etc.): https://archive.ics.uci.edu/dataset/222/bank+marketing
* Formato del archivo (.csv, .xlsx.): csv

1. **Estructura de la Base de Datos**

* Número total de filas (registros): 45211
* Número total de columnas (atributos o variables):17

1. **Descripción de las Columnas**

Por cada columna se debe especificar:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Nombre de la columna** | **Tipo de dato** | **Descripción** | **Cantidad de datos** | **Valores únicos (si aplica)** |
| Age | numérico | Edad cliente | 45211 |  |
| job | Texto | tipo de trabajo (categórico: management,  technician,  entrepreneur,  blue-collar,  unknown,  retired,  admin.,  services,  self-employed,  unemployed,  housemaid,  student) | 45211 | management,  technician,  entrepreneur,  blue-collar,  unknown,  retired,  admin.,  services,  self-employed,  unemployed,  housemaid,  student |
| Marital | Texto | estado civil (categórico:  divorced, ' married) | 45211 | categórico:  divorced, ' married |
| Education | Texto | (categórico: ' primary', ' secondary', '  tertiary) | 45211 | primary', ' secondary', '  tertiary |
| Default | Bool | ¿Tiene crédito en mora? | 45211 | Yes, no |
| Balance | Numérico | saldo anual promedio | 45211 |  |
| Housing | Bool | ¿Tiene préstamo de vivienda? | 45211 | Yes, no |
| Loan | Bool | ¿tiene préstamo personal? | 45211 | Yes, no |
| Contact | Texto | tipo de comunicación de contacto (categórica: 'celular', 'teléfono', no sabe ) | 45211 | categórica: 'celular', 'teléfono', no sabe |
| Day | Fecha | último día de contacto de la semana | 45211 |  |
| Month | Fecha | último mes de contacto del año (categórico: 'ene', 'feb', 'mar', ..., 'nov', 'dic') | 45211 |  |
| Duration | Numérico | Duración del último contacto, en segundos (numérico). Nota importante: Este atributo afecta considerablemente el resultado objetivo (p. ej., si duración = 0, entonces y = 'no'). Sin embargo, la duración no se conoce antes de realizar una llamada. Además, después de finalizar la llamada, y se conoce obviamente. Por lo tanto, esta entrada solo debe incluirse para fines de referencia y debe descartarse si se pretende obtener un modelo predictivo realista. | 45211 |  |
| Campaign | Numérico | Número de contactos realizados durante esta campaña y para este cliente (numérico, incluye el último contacto) | 45211 |  |
| pdays | Numérico | Número de días transcurridos desde la última vez que se contactó al cliente de una campaña anterior (numérico; -1 significa que el cliente no fue contactado previamente) | 45211 |  |
| previous | Numérico | Número de contactos realizados antes de esta campaña y para este cliente | 45211 |  |
| poutcome | Texto | resultado de la campaña de marketing anterior (categórico: failure, other, success) | 45211 | failure, other, success |
| Y | Bool | ¿El cliente ha suscrito un depósito a plazo? | 45211 | Yes, no |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |

1. **Calidad de los Datos**

* ¿Hay valores faltantes en alguna columna? ¿Cuántos? No
* ¿Hay columnas con valores repetidos o inconsistentes? No
* ¿Existen datos fuera de rango o errores evidentes? No
* ¿Conoce la última fecha de actualización de los datos? 13/02/2012

1. **Análisis Cuantitativo Básico**

Para columnas numéricas:

* Mínimo, máximo, media, mediana, desviación estándar.
* (age)

|  |  |  |  |
| --- | --- | --- | --- |
| min | max | mediana | desvsta |
| 18 | 95 | 39 | 10,618762 |

* Balance

|  |  |  |  |
| --- | --- | --- | --- |
| min | max | mediana | desvsta |
| -8019 | 102127 | 448 | 3044,76583 |

* Duration

|  |  |  |  |
| --- | --- | --- | --- |
| min | max | mediana | desvsta |
| 0 | 4918 | 180 | 257,527812 |

* Pdays

|  |  |  |  |
| --- | --- | --- | --- |
| min | max | mediana | desvsta |
| -1 | 871 | -1 | 100,128746 |

Para columnas categóricas:

* Frecuencia de cada categoría.

|  |  |
| --- | --- |
| **job** | **cantidad** |
| admin. | 5171 |
| blue-collar | 9732 |
| entrepreneur | 1487 |
| housemaid | 1240 |
| management | 9458 |
| retired | 2264 |
| self-employed | 1579 |
| services | 4154 |
| student | 938 |
| technician | 7597 |
| unemployed | 1303 |
| unknown | 288 |
| **Total general** | **45211** |

|  |  |
| --- | --- |
| **education** | **Cuenta de education** |
| primary | 6851 |
| secondary | 23202 |
| tertiary | 13301 |
| unknown | 1857 |
| **Total general** | **45211** |

* Número de categorías únicas.

1. **Conclusiones Iniciales del Grupo**

* ¿Qué información relevante puede extraerse de esta base de datos?

Demográficos (edad, estado civil)

Académicos (nivel de estudio)

Contactabilidad (tiempo de conversación, fecha ult conversación)

* ¿Qué problemas de calidad o estructura podrían afectar un análisis más profundo?

Mucha información “unknown” desconicida

* ¿Qué preguntas podría ayudar a responder esta base de datos?

El objetivo de la clasificación es predecir si el cliente suscribirá un depósito.

* ¿Qué mejoras o limpiezas serían necesarias?

La base está muy limpia o estandarizada

1. **Entrega**

Cada grupo debe entregar este documento diligenciado y, si es posible, el archivo de base de datos con sus observaciones o anotaciones.