PROCESO DE RECOLECCION DE DATOS

CAMPAÑA PLAZO FIJO

➢ *1.FUENTES*

IDENTIFICACIÓN DE FUENTES

- Base de dades interna del CRM

DESCRIPCIÓN DE FUENTES

- La base de datos CRM contiene información sobre los usuarios y las interacciones con ellos que en este caso son los contactos realizados y la frecuencia de los mismos , así también como datos financieros y si cuenta ya con un depósito a plazo o no.

➢ *2.MÉTODO DE RECOLECCIÓN DE DATOS*

PROCEDIMIENTOS Y HERRAMIENTAS

- Exportación programada en formato CSV, almacenada en un repositorio de GitHub diariamente. Esta tarea la realiza el departamento de IT.

FRECUENCIA DE RECOLECCIÓN

- Diaria.

SCRIPT DE DESCARGA:

```
import pandas as pd

csv_url = "https://raw.githubusercontent.com/ITACADEMYprojectes/projecteML/main/bank_dataset.CSV"

try:
    df = pd.read_csv(csv_url, on_bad_lines='skip')
    print(df.info())
except pd.errors.ParserError as e:
    print(f"Error al leer el archivo CSV: {e}")
```

[4] ✓ 0.6s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11162 entries, 0 to 11161
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        11152 non-null  float64
 1   job        11162 non-null  object
 2   marital    11157 non-null  object
 3   education  11155 non-null  object
 4   default    11162 non-null  object
 5   balance    11162 non-null  int64
 6   housing    11162 non-null  object
 7   loan       11162 non-null  object
 8   contact    11162 non-null  object
 9   day        11162 non-null  int64
 10  month      11162 non-null  object
 11  duration   11162 non-null  int64
 12  campaign   11162 non-null  int64
 13  pdays      11162 non-null  int64
 14  previous   11162 non-null  int64
 15  poutcome   11162 non-null  object
 16  deposit    11162 non-null  object
dtypes: float64(1), int64(6), object(10)
memory usage: 1.4+ MB
None
```

➢ *3.FORMATO Y ESTRUCTURA DE LOS DATOS*

TIPOD DE DATOS

- Numéricos: age, balance, day, duration, campaign, pdays, previous.

- Categóricos: job, marital, education, default, housing, loan, contact, month, poutcome, deposit.

FORMATO DE ALMACENAMIENTO

- Datos tabulares almacenados en ficheros csv.

➢ *4.LIMITACIONES DE LOS DATOS*

-Teniendo en cuenta la VALIOSA información detallada en:
https://archive.ics.uci.edu/dataset/222/bank+marketing

Y después de hacer un análisis a los datos en si, creo sería mejor para ML cambiar algunos datos categóricos por valores numéricos que son los que interpreta mejor el modelo.

En el campo pdays quizá cambiaría el valor -1 que representa que el cliente no ha sido contactado a por 0 para evitar errores en el análisis.

➢ *5.CONSIDERACIONES SOBRE DATOS SENSIBLES*

TIPOS DE DATOS SENSIBLES

- información Personal: Tales como edad, trabajo, estado civil, educación y método de contacto.

- Información Financiera: Si el cliente cuenta con un crédito, su balance económico, prestamos o si el cliente cuenta con un depósito a plazo.

- Información de Contactos: Datos sobre las campañas y los contactos efectuados al cliente.

MEDIDAS DE PROTECCIÓN

- Anonimato: Si bien es cierto que no tenemos nombres ni números de documentos en esta base de datos, podríamos crear un ID único para cada cliente ya que es una buena práctica además de una capa extra para mantener el anonimato sobre otros datos personales.

- Acceso Restringido: Se puede dar acceso a estos datos solo a personal autorizado que vaya a trabajar en el proyecto y si en con números de identificadores se refuerza la protección ya que sería un doble ciego sobre de quien son los datos.

- Cumplimiento de las Regulaciones: Se debe revisar constantemente de que no se incurra en un fallo que atente contra el incumplimiento de la regulación vigente poniendo en peligro el acceso y/o uso de los datos indebidamente.