Documentació del Procés de Recol·lecció de Dades per al Projecte Bank Dataset de ML amb accés a una Base de Dades SQL

1. Fonts

Identificació de Fonts:

- La base de dades on hi són les dades dels clients és una BD SQL privada del banc, accessible dins de la LAN corporativa mitjançant connexions segures.

Descripció de les Fonts:

- La base de dades conté informació detallada sobre els clients, les seves dades financeres i els resultats de campanyes de màrqueting prèvies.

2. Mètodes de Recol·lecció de Dades

Procediments i Eines:

- Connexió a la base de dades SQL amb credencials autoritzades des d'un script Python.
- Consulta SQL per extreure les dades necessàries.
- Guardar les dades extretes en un fitxer CSV per a posteriors anàlisis.

Freqüència de Recol·lecció:

- Dades extretes periòdicament amb periodicitat mensual

Exemple de Script Python per a la Connexió i Exportació:

```
import pandas as pd
import pyodbc  # Llibreria per connectar amb bases de dades SQL

# Configuració de la connexió SQL
conn = pyodbc.connect(
  'DRIVER={SQL Server};'
  'SERVER=192.168.1.100;'  # IP o nom del servidor SQL
  'DATABASE=BankDB;'      # Nom de la base de dades
  'UID=usuari;'                # Nom d'usuari autoritzat
  'PWD=contrasenya;'     # Contrasenya segura
)
```

```
# Consulta SQL per extreure les dades
query = """
SELECT
  age, job, marital, education, default, balance, housing, loan,
  contact, day, month, duration, campaign, pdays, previous, poutcome, deposit
FROM
  MarketingData
WHERE
  campaign_date >= '2024-12-31'; # afegir la data requerida per recollir les dades
"""

# Executar la consulta i carregar les dades en un DataFrame
df = pd.read_sql_query(query, conn)

# Guardar les dades en un fitxer CSV
csv_path = "bank_dataset.csv"
df.to_csv(csv_path, index=False)

print(f"Dades exportades correctament al fitxer: {csv_path}")
```

3. Format i Estructura de les Dades

Tipus de Dades:

- **Numèriques:** age, balance, day, duration, campaign, pdays, previous
- **Categòriques:** job, marital, education, default, housing, loan, contact, month, poutcome
- **Binàries:** default, housing, loan, deposit

Format d'Emmagatzematge:

- Fitxer CSV generat amb estructura tabular, una fila per client.

4. Limitacions de les Dades

- **Connexió restringida:** Accés només disponible dins de la LAN corporativa amb credencials adequades.
- **Dades incompletes:** Algunes columnes poden contenir valors desconeguts (unknown).

- **Actualització de les dades:** Les dades extretes reflecteixen només els registres disponibles fins al moment de l'execució.

5. Consideracions sobre Dades Sensibles

Tipus de Dades Sensibles:

- **Demogràfiques:** age, marital, education
- **Financeres:** balance, loan, housing
- **Comportamentals:** Històric de campanyes: campaign, pdays, previous

Mesures de Protecció:

- **Accés restringit:** Només personal autoritzat pot executar el script i accedir al fitxer CSV generat.
- **Seguretat de la connexió:** Comunicació encriptada amb el servidor SQL.
- **Anonimització:** Aplicar mecanismes per eliminar o transformar dades identificatives en dades pseudonimitzades abans de compartir-les (hashing,etc..)
- **Compliment normatiu:** Complir amb la GDPR i altres regulacions aplicables per protegir la privacitat dels clients.

Objectiu del Procés:

L'objectiu del procés descrit és realitzat la recollecció de dades directament des de la base de dades corporativa per garantir dades actualitzades i consistents per a l'entrenament i la validació de models predictius.