Descripción General del Proyecto - Project Overview.

Este proyecto tiene como objetivo analizar y preparar una base de datos bancaria para su posterior modelado, utilizando técnicas de limpieza, transformación y análisis de datos. A lo largo de este proceso, se aplican varias operaciones de manipulación de datos, tales como la fusión de datasets, eliminación de duplicados, limpieza de valores nulos, y conversión de variables categóricas a variables dummy para facilitar el uso en algoritmos de machine learning.

The objective of this project is to analyze and prepare a banking database for subsequent modeling, using various data cleaning, transformation, and analysis techniques. Throughout this process, several data manipulation operations are applied, such as merging datasets, removing duplicates, cleaning missing values, and converting categorical variables into dummy variables to facilitate use in machine learning algorithms.

1. Lectura de Archivos - Reading Files

El primer paso consiste en cargar los datos desde archivos locales al entorno de trabajo en Google Colab.

- **Archivo 1**: Leemos el archivo que contiene la información de las cuentas bancarias (archivo `account_info`).
- **Archivo 2**: Leemos el archivo que contiene la información de los clientes (archivo `customer_info`).

Ambos archivos son fundamentales para el análisis, ya que contienen la información principal sobre los clientes y sus cuentas bancarias.

The first step is to load the data from local files into the working environment in Google Colab.

- ***File 1: We read the file containing account information (`account_info`).***
- ***File 2: We read the file containing customer information (`customer_info`).***

Both files are essential for the analysis as they contain the primary information about the clients and their bank accounts.

2. Unión de Datasets - Merging Datasets

Una vez cargados los datasets, procedemos a unir los datos de `account_info` y `customer_info` utilizando la columna común `customer_ID`. La unión se realiza mediante un **left join** para asegurar que no perdemos información relevante de `account_info` en caso de que existan discrepancias entre los dos archivos.

Once the datasets are loaded, we proceed to merge the `account_info` and `customer_info` data using the common column `customer_ID`. The merge is performed via a left join to ensure that we do not lose relevant information from `account_info` in case of discrepancies between the two files.

3. Eliminación de Columnas Duplicadas - Removing Duplicated Columns

Después de la unión, observamos que se generan columnas duplicadas (por ejemplo, `ternure_x` y `ternure_y`). Para evitar redundancias, eliminamos una de las columnas duplicadas y revisamos que no haya duplicados en el dataset.

After the merge, we notice that duplicate columns (e.g., `ternure_x` and `ternure_y`) are generated. To avoid redundancy, we drop one of the duplicated columns and ensure there are no duplicated rows in the dataset

4. Limpieza del Dataset - Data Cleaning

a) Limpieza de la Columna Monetaria - Cleaning the Monetary Column

El dataset contiene columnas monetarias con el símbolo "€". Para hacer un análisis adecuado, es necesario retirar este símbolo y convertir estas columnas del tipo `Object` al tipo `float`.

The dataset contains monetary columns with the "€" symbol. To perform a proper analysis, it is necessary to remove this symbol and convert these columns from `Object` type to `float`.

b) Imputación de Valores Faltantes - Imputation of Missing Values

Las columnas categóricas y numéricas contienen valores faltantes. Para las columnas categóricas, rellenamos los valores faltantes con el modo (la categoría más frecuente), mientras que para las columnas numéricas, utilizamos la **media** para completar los valores vacíos.

Both categorical and numerical columns contain missing values. For the categorical columns, we fill missing values with the mode (the most frequent category), while for numerical columns, we use the mean to complete missing values.

5. Revisión y Corrección de Inconsistencias- Review and Correction of Inconsistencies

a) Valores Negativos en "EstimatedSalary" - Negative Values in "EstimatedSalary"

Detectamos que algunos registros de la columna `EstimatedSalary` tienen valores negativos, lo cual no es razonable para esta variable. Procedemos a reemplazar esos valores negativos.

We detected that some records in the `EstimatedSalary` column have negative values, which is not reasonable for this variable. Therefore, we proceed to replace those negative values.

b) Unificación de Categorías en "Geography" - Unifying Categories in "Geography"

La columna `Geography` presenta variaciones en los nombres de los países, como "FRA", "FRANCE", o "FRENCH". Normalizamos estos valores para que solo exista una descripción por país.

The `Geography` column presents variations in country names, such as "FRA", "FRANCE", or "FRENCH". We normalize these values so that only one description exists for each country.

6. Exploración de Datos - Data Exploration

Generamos gráficos para explorar la distribución y comportamiento de los datos. Esta exploración incluye la visualización de variables numéricas y categóricas relevantes para comprender mejor el dataset.

We generate graphs to explore the distribution and behavior of the data. This exploration includes visualizing relevant numerical and categorical variables to better understand the dataset.

7. Preparación de los Datos para el Modelado - Data Preparation for Modeling

Antes de proceder al modelado, eliminamos columnas que no tienen valor predictivo o que no son útiles en el análisis. Por ejemplo, eliminamos la columna `CustomerId`, ya que no aporta información relevante para el modelo.

Before proceeding with modeling, we remove columns that have no predictive value or are not useful for analysis. For example, we drop the `CustomerId` column, as it does not provide relevant information for the model.

8. Creación de Variables Ficticias (Dummy Variables)

Finalmente, convertimos las variables categóricas en variables numéricas mediante la creación de **variables ficticias**. Este paso es esencial para preparar los datos para modelos de machine learning que requieren que todas las entradas sean numéricas.

Finally, we convert categorical variables into numerical variables by creating dummy variables. This step is essential to prepare the data for machine learning models that require all inputs to be numerical.

9. Conclusión

Al completar el proceso de limpieza, imputación y transformación de datos, el dataset queda listo para aplicar modelos de machine learning o cualquier otro tipo de análisis avanzado. Los pasos detallados garantizan que los datos estén en un formato limpio y apropiado para el modelado, eliminando duplicados, valores inconsistentes y errores en las columnas.

After completing the data cleaning, imputation, and transformation process, the dataset is ready for machine learning models or any advanced analysis. The detailed steps ensure that the data is in a clean and appropriate format for modeling, eliminating duplicates, inconsistent values, and errors in the columns.