Week 9 Deliverable
Group Name: Neural Nomad

Name: Nitish Jena
Email: nitishkumar.jena.2023@student.ism.de
Country: Germany
College: ISM Hochschule
Specialization: Data Science

Problem Description

The dataset "bank-additional-full.csv" contains various customer attributes used to predict whether a client will subscribe to a term deposit. However, the raw data contains missing values, outliers, and inconsistencies that must be addressed before using it for analysis or model training. The goal of this data cleansing and transformation process is to prepare a high-quality dataset by handling missing values, removing outliers, and improving data integrity.

GitHub Repository Link

https://github.com/nitishjena/Week-9/blob/main/Data_Cleaning.ipynb

Data Cleansing and Transformation Approaches

Handling Missing Values:

1. Dropping Missing Values

Rows where "marital" and "job" fields were 'unknown' were removed as these attributes play a significant role in customer profiling.

2. Imputing Missing Values

Missing values in "housing" and "loan" were replaced with the most frequent value (mode).

3. Predicting Missing "default" Values Using Machine Learning

- "default" contained missing values categorized as 'unknown'. Instead of dropping these, a Random Forest Classifier was used to predict them.
- One-hot encoding was applied to categorical features before training the model.

Handling Outliers:

Outliers in the "duration" column were removed using the Interquartile Range (IQR) method.

Final Dataset

After performing the cleansing operations:

- The dataset was free of missing values.
- Outliers were removed to maintain consistency.
- Machine Learning was used to impute categorical missing values, improving data completeness.
- The final dataset shape:

Results and Findings

- Dropped 'unknown' values for job and marital status.
- Replaced missing values in housing and loan with the mode.
- Used a Random Forest model to predict and replace missing "default" values.
- Applied IQR method to remove extreme outliers from the "duration" feature.
- The final dataset is cleaned, structured, and ready for further analysis or model training.