Step 1: Exploratory Data Analysis (EDA)

University: Tek-Up
Group: ING-4-SDIA

Tutor: Haythem Ghazouani

1 Overview

Exploratory Data Analysis (EDA) is the most critical phase of the Machine Learning pipeline. It allows you to understand the "soul" of your data, uncover patterns, detect anomalies, and test hypotheses before modeling.

2 The Comprehensive EDA Workflow

To truly explore a dataset, you must follow these systematic steps:

Step 1: Data Inspection & Cleaning

Check data types (`df.info()`), detect missing values (`df.isnull().sum()`), and identify duplicates. Ensure the dataset matches the expected schema.

Step 2: Univariate Analysis

Analyze variables one by one.

- **Categorical:** Frequency counts and pie charts (e.g., Gender, Country).
- **Numerical:** Histograms to check distribution (Normal vs. Skewed) and Boxplots to detect outliers (e.g., Age, Balance).

Step 3: Bivariate Analysis (Variable vs. Target)

How do features relate to the target variable?

- **Categorical vs. Target:** Grouped bar charts (e.g., Is Churn higher for Active Members?).
- **Numerical vs. Target:** Boxplots or Violin plots (e.g., Does Churn correlate with Age?).

Step 4: Multivariate Analysis & Correlation

Use a **Heatmap** to check for multicollinearity (redundant features) and identify which variables have the strongest linear relationship with the target.

3 Implementation in Python

Below is the structured EDA logic used for the Bank Churn synopsis.

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

df = pd.read_csv("data/bank_churn.csv")

# 1. Target Distribution
sns.countplot(x='Exited', data=df)
plt.title('Target Variable Distribution (Imbalance Check)')
plt.show()

# 2. Distribution analysis using Histogram
sns.histplot(df['Age'], kde=True)
plt.title('Age Distribution')
plt.show()

# 3. Relationship between Age and Churn
sns.boxplot(x='Exited', y='Age', data=df)
plt.title('Age vs Churn (Bivariate)')
plt.show()

# 4. Correlation Matrix
plt.figure(figsize=(10,8))
sns.heatmap(df.corr(), annot=True, cmap='RdYlGn')
plt.title('Feature Correlation Map')
plt.show()
```

4 Conclusion

A good EDA should lead to clear **Feature Engineering** ideas. For example, if you notice that customers with 3+ products churn 80% of the time, you might create a new "High Risk Product Count" feature.