TP5: Customer Churn Classification

Objective

The goal of this exercise is to predict whether a customer will churn (leave the service) based on various features such as their account length, call plans, usage patterns, and interaction with customer service.

Instructions

1. Read the Data

- Load the churn dataset into a pandas DataFrame
- Display the first few rows to understand the structure of the data

2. Explore the Data

- Print the shape of the dataset
- Identify missing values and summarize each column
- Analyze the distribution of the target variable (`Churn`)
- Visualize relationships between features using heatmaps and pair plots

3. Handle Missing Values

- Identify and handle missing values by either removing rows/columns or imputing values using statistical methods

4. Check and Handle Outliers

- Use box plots to detect outliers in numerical features
- Decide whether to remove, cap, or transform outliers

5. Clean the Data

- Check for and remove duplicate rows if present
- Drop unnecessary columns that do not add value to the classification task

6. Transform Categorical Variables

- Convert categorical features (e.g., `International plan`, `Voice mail plan`) into numerical values using one-hot encoding or label encoding

7. Scale the Data

- Normalize or scale numeric features to improve model performance

8. Split the Data

- Separate the features (X) from the target (y)
- Split the dataset into training and testing sets with an 80/20 ratio

9. Train Machine Learning Models

- Train at least two models from the following:
  - K-Nearest Neighbors (KNN)
  - Logistic Regression
  - Random Forest
- Use the training set to fit each model

10. Evaluate the Models

- Use the test set to predict the target variable
- Calculate and compare the following metrics for each model:
  - Accuracy
  - Precision
  - Recall
  - F1-Score

11. Compare the Models

- Summarize the results in a table or chart to compare model performance
- Identify the best-performing model and explain why it performed well

12. Conclusion

- Write a brief conclusion discussing the results, key insights, and potential ways to improve the analysis

Deliverables

- A Python notebook (`.ipynb`) or a Python file (.py) containing all the code, analysis, and results
- A summary report highlighting the key findings and the best-performing model