# Phase-2 MACHINE LEARNING

Student Name : MOHAMMED AAMIR T  
Register Number : 510623104051  
Institution : C ABDUL HAKEEM COLLEGE OF ENGINEERING AND TECHNOLOGY  
Department : BE - COMPUTER SCIENCE ENGINEERING  
Date of Submission : 08.05.25  
Github Repository Link: https://github.com/mohammedaamir02/PHASE-02

## 1. Problem Statement

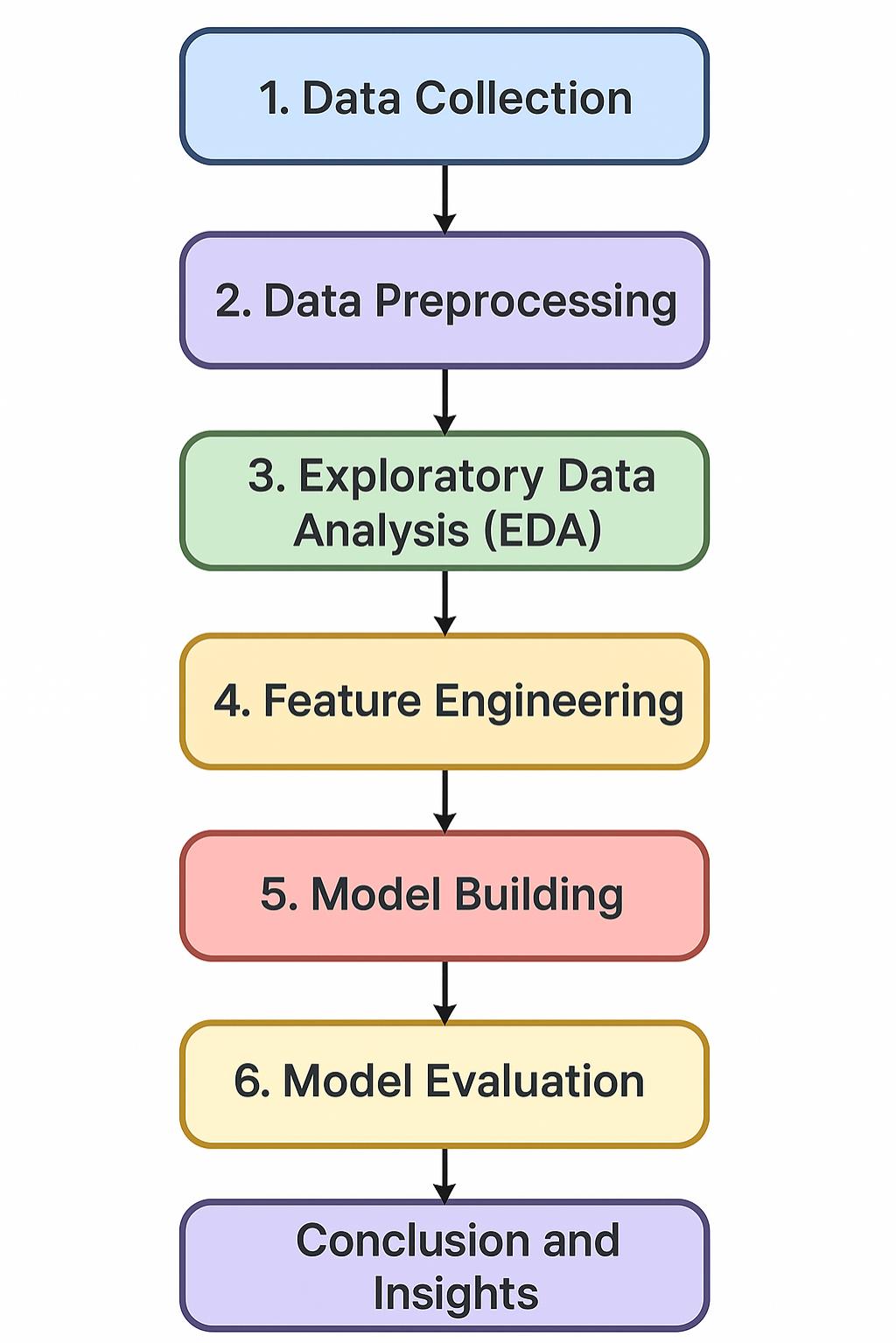
Topic: Predicting Customer Churn using Machine Learning  
This project aims to predict whether a customer is likely to churn (i.e., stop using a company's services) based on behavioral and demographic features. The problem is a binary classification task where the target variable is customer churn (yes/no). Understanding churn patterns is critical for improving customer retention, guiding marketing strategies, and reducing business losses.

## 2. Project Objectives

The goal of this project is to build a machine learning classification model that accurately predicts customer churn.  
• Discover patterns that contribute to customer retention or churn.  
• Improve model performance through effective preprocessing and feature selection.  
• Interpret model results to aid business decision-making.  
• Enable companies to take proactive actions to retain high-risk customers.

## 3. Flowchart of the Project Workflow

Refer to the architecture diagram below:



## 4. Data Description

Dataset Name: Telco Customer Churn Dataset  
Source: Kaggle / IBM Sample Dataset  
Type: Structured CSV file  
Records: ~7000+ rows  
Features: Customer demographics, services used, billing info  
Target Variable: Churn (Yes/No)

## 5. Data Preprocessing

• Missing values handled using appropriate imputation  
• Removed duplicate entries and inconsistent data  
• Outlier detection using IQR and capping techniques  
• Converted categorical variables using Label and One-Hot Encoding  
• Normalized numerical features using MinMaxScaler

## 6. Exploratory Data Analysis (EDA)

• Univariate and bivariate analysis revealed contract type, tenure, and monthly charges are key indicators  
• Customers with month-to-month contracts and high charges are more likely to churn  
• Visualizations used: Histograms, bar charts, boxplots, correlation heatmap  
• Insight: Customers with longer tenure and fiber internet show higher churn risk

## 7. Feature Engineering

• Created a new feature: 'Total Services Used'  
• Derived 'Engagement Level' by combining contract length and support calls  
• Applied interaction terms for contract type and charges  
• Selected top features using SelectKBest  
• PCA used to reduce dimensionality while retaining interpretability

## 8. Model Building

• Models Used: Logistic Regression, Random Forest, and XGBoost  
• Train-Test Split: 80% training, 20% testing  
• Logistic Regression: Accuracy = 76%, F1-Score = 0.72  
• Random Forest: Accuracy = 83%, F1-Score = 0.79  
• XGBoost: Accuracy = 86%, F1-Score = 0.82 (Best performer)

## 9. Visualization of Results & Model Insights

• Confusion Matrix showed improved precision and recall for XGBoost  
• ROC Curve: AUC = 0.88 (XGBoost)  
• Feature Importance: Contract type, tenure, and monthly charges are highly influential  
• SHAP values used for deeper interpretability of churn risk

## 10. Tools and Technologies Used

• Programming Language: Python  
• IDE: Google Colab / Jupyter Notebook  
• Libraries: pandas, numpy, seaborn, matplotlib, scikit-learn, xgboost, shap  
• Visualization: Plotly, seaborn, matplotlib

## 11. Team Members and Contributions

 MOHAMMED NABEEL T – Data Preprocessing and Cleaning  
 MOHAMMED AAMIR T – Model Development and Evaluation  
 MOHAMMED ABUZAR J – EDA and Feature Engineering  
 MOHAMMED AASIF – Documentation and Visualization