**Phase-2**

**Student Name:** BHARATHI R

**Register Number:** 422623106009

**Institution:** UNIVERSITY COLLEGE OF ENGINEERING,PANRUTI

**Department:** ECE

**Date of Submission:**

**Github Repository Link:** [Update the project source code to your Github Repository]

# 1. Problem Statement

* Customer churn presents a significant challenge for subscription-based businesses. Understanding the behavioral and demographic patterns that lead to churn allows companies to implement effective retention strategies. This project aims to predict whether a customer is likely to discontinue a service, based on historical data. This is a binary classification problem where the target variable is whether the customer has churned or not. Solving this helps improve customer lifetime value and reduce acquisition costs

# 2. Project Objectives

* Develop a machine learning model to accurately predict customer churn
* Identify and interpret the key factors contributing to churn.
* Improve prediction metrics (accuracy, recall, F1-score) through data preprocessing and feature engineering.
* Translate model insights into actionable business strategies for retention.

**3. Flowchart of the Project Workflow**

[Insert a diagram showing: Data Collection -> Data Preprocessing -> EDA -> Feature Engineering -> Model Building -> Evaluation -> Interpretation & Reporting]

# 4. Data Description

* Dataset Source: [e.g., Kaggle - Telco Customer Churn Dataset]
* Type: Structured tabular data- Records: ~7,000 customer entries
* Features: 20+ including demographics, account details, and usage patterns
* Dataset Type: Static
* Target Variable: Churn (Yes/No)

# 5. Data Preprocessing

* Handled missing values using imputation (e.g., mode for categorical fields).
* Removed duplicates based on customer ID.
* Outliers identified in tenure and monthly charges were capped.
* Data type consistency ensured across categorical and numerical columns.
* Categorical features encoded using one-hot encoding.
* Normalized numerical features such as monthly charges and tenure for model compatibility

# 6. Exploratory Data Analysis (EDA)

* Univariate Analysis: Distribution plots for churn, tenure, contract type, etc.
* Bivariate Analysis: Correlation matrix; churn vs. contract type, payment method, etc.
* Higher churn among month-to-month contracts.
* Electronic check users show higher churn.
* Tenure inversely correlated with churn likelihood

# 7. Feature Engineering

* + Created TotalServices by summing subscribed services.
  + Binned tenure into categories (new, medium, long-term customers).
  + Derived interaction features like MonthlyCharges \* Tenure.
  + Removed redundant features post-EDA
  + Applied PCA (optional) for dimensionality reduction.

# 8. Model Building

* + Models Used: Logistic Regression, Random Forest, and XGBoost
  + Dataset split into 80:20 for training and testing (stratified).
  + Evaluation Metrics:
  + Accuracy, Precision, Recall, F1-score
  + ROC-AUC for threshold optimization

# 9. Visualization of Results & Model Insights

* + Confusion matrix and ROC curve for all models.
  + Feature importance plot from tree-based models.
  + Identified top drivers of churn: contract type, tenure, payment method, and total charges.
  + Visual comparison of models shows Random Forest slightly outperforming o

# 10. Tools and Technologies Used

* + Programming Language: Python
  + IDE: Google Colab
  + Libraries: pandas, numpy, seaborn, matplotlib, scikit-learn, XGBoost
  + Visualization: Plotly, seab

# 11. Team Members and Contributions

* **[BHARATHI R] : Data Cleaning, EDA**
* **[BHARATHIRAJA A] : Feature Engineering, Model Training**
* **[SAKTHIVEL T] : Evaluation**
* **[VETRIVEL S] : Reporting**
* **[ESAKKIANKEERTHIK S] : Documentation**