

## Phase-2 Submission

**Student Name:** Deeksha P

**Register Number:** 71252320235016

**Institution:** PPG Institute of Technology

**Department:** B.Tech Information Technology

**Date of Submission:** 09.05.2025

**GitHub Repository Link:** [Repo link](#)

---

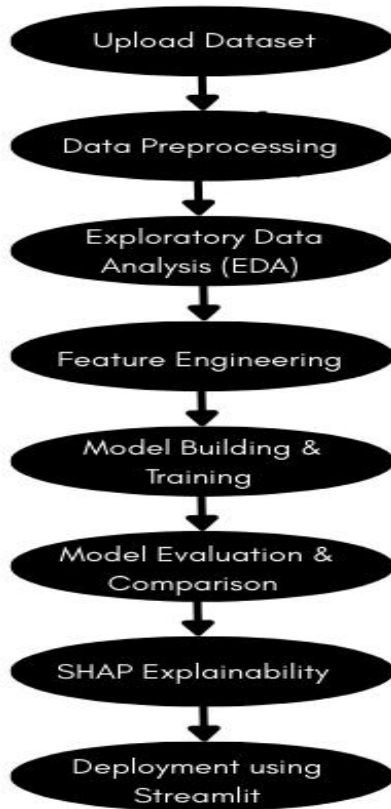
### 1. Problem Statement

- *Customer churn is a critical problem that affects long-term revenue and growth in industries like telecommute, banking, and subscription-based services. This project aims to solve a **binary classification** problem: predicting whether a customer will churn (i.e., leave the service) based on demographic and service usage features.*
- *Understanding churn behavior helps businesses reduce customer acquisition costs and improve retention strategies by proactively identifying high-risk customers.*

### 2. Project Objectives

- *Predict customer churn using machine learning classification models and important features that contribute to churn.*
- *Build interpret-able models that can be used for decision-making.*
- *Evaluate and compare different ML algorithms based on precision, recall, and F1 score.*
- *Deliver a functional model ready for integration with business dashboards.*

### 3. Flowchart of the Project Workflow



### 4. Data Description

- **Data set Name and origin:** Telco Customer Churn and kaggle
- **Dataset Link:**  
<https://www.kaggle.com/datasets/blastchar/telcocustomerchurn>
- **Type:** Structured data
- **Records:** 7043 (raw) → 7032 (preprocessed)
- **Features:** 21 original features → 31 engineered features
- **Target Variable:** Churn (Yes/No)
- **Data set Type:** Static

## 5. Data Preprocessing

- *Removed customer ID column as it does not influence churn.*
- *Handled missing/invalid entries in Total Charges by converting to numeric and removing invalid rows.*
- *Converted categorical variables using **one-hot encoding**.*
- *Normalized numeric fields (TotalCharges, MonthlyCharges) for better model learning.*
- *Ensured all features were numeric for compatibility with ML models.*

## 6. Exploratory Data Analysis (EDA)

### **Uni-variate Analysis:**

- *tenure and Monthly-charges showed diverse distribution.*
- *Most customers have a month-to-month contract and electronic check payment.*

### **Bivariate/Multivariate Analysis:**

- *High churn observed among customers with fiber optic internet and month-to-month contracts.*
- *Tenure is inversely related to churn likelihood.*

### **Insights Summary:**

- *Contract type, payment method, internet service, and tenure are strong churn indicators.*

## 7. Feature Engineering

- *One-hot encoding of categorical variables.*
- *Removed multicollinear columns and low-variance features.*
- *No PCA applied as models handled feature count well.*
- *No date features involved; no need for time-based extraction.*

## 8. Model Building

***Models Used:*** Logistic Regression

- *Random Forest Classifier*
- *XGBoost Classifier*

***Model Selection Justification:***

- *All selected models are interpretable and suitable for binary classification.*
- *Random Forest and XG Boost help in identifying feature importance.*

***Evaluation Metrics:***

- *Accuracy, Precision, Recall, F1-Score, ROC - AUC*

***Train-Test Split:*** 80/20 stratified split to maintain class balance.

## 9. Visualization of Results & Model Insights

- *Confusion Matrix: Evaluated true positives and false negatives.*
- *ROC-AUC Curve: Compared model discriminative power.*
- *Feature Importance: Identified top predictors like Contract, tenure, and Internet Service.*
- *SHAP Values (optional): To interpret individual predictions.*

## 10. Tools and Technologies Used

- *Programming Language: Python*
- *IDE/Notebook: Google Colab, Jupyter Notebook*
- *Libraries: pandas, numpy, matplotlib, seaborn, plotly, scikit-learn, xgboost*
- *Visualization Tools: Seaborn, Plot, SHAP*
- *Optional Deployment: Streamlit (not yet deployed)*

## 11. Team Members and Contributions

<i>NAME</i>	<i>ROLE</i>
<i>Thilshan S</i>	<i>Data cleaning</i>
<i>Dinesh D</i>	<i>EDA</i>
<i>Srimathi B</i>	<i>Feature engineering</i>
<i>Deeksha P</i>	<i>Model development</i>
<i>Tamilarasan B</i>	<i>Documentation reporting</i>