Student Name: Janani v
[**Register Number:** [712523121007]

Institution: [PPG INSTITUTE OF TECHNOLOGY]

Department: [BIOMEDICAL ENGINEERING]

Date of Submission: [25-04-2025]

Github Repository Link:[Github]

---

1. Problem Statement

Customer churn is a major business problem. This project uses machine learning classification techniques to predict churn based on demographic and behavioral data. The model helps businesses take proactive steps to retain customers by identifying key churn indicators, thus reducing revenue loss.

2. Project Objectives

- Classify customers as Churn or Non-Churn.
- Perform exploratory data analysis (EDA) and visualize key churn indicators.
- Compare machine learning models: Logistic Regression, Random Forest, XGBoost, SVM, and KNN.
- Use SHAP for explainability and interpretability of model decisions.
- Deploy the model using Streamlit to allow business users to make predictions interactively.

3. Flowchart of the Project Workflow

Upload Dataset
↓
Data Preprocessing
↓
Exploratory Data Analysis (EDA)
↓
Feature Engineering
↓
Model Building & Training
↓
Model Evaluation & Comparison
↓
SHAP Explainability
↓
Deployment using Streamlit

4. Data Description

- **Dataset:** churn.csv
- **Source:** Provided by the institution
- **Type:** Structured, tabular data
- **Size:** ~10,000 rows
- **Static or Dynamic:** Static dataset
- **Target Variable:** Exited (1 = Churn, 0 = Non-Churn)

5. Data Preprocessing

- Dropped unnecessary columns: RowNumber, CustomerId, Surname
- Encoded categorical variables: Geography, Gender using LabelEncoder

- Scaled numerical features using StandardScaler
- Verified and cleaned the data to ensure consistency

6. Exploratory Data Analysis (EDA)

- Plotted correlation heatmaps and churn distribution
- Identified strong predictors of churn such as:
  - Tenure
  - IsActiveMember
  - Geography
- Used visual tools such as count plots and heatmaps to understand feature interactions

7. Feature Engineering

- Applied label encoding for categorical variables
- Removed non-informative identifiers
- Standardized numerical features to improve model training

8. Model Building

- Trained five classification models:
  - Logistic Regression
  - Random Forest
  - XGBoost
  - SVM
  - KNN
- Used metrics like Accuracy, Precision, Recall, F1-Score, and ROC-AUC to compare models
- XGBoost performed best with highest accuracy and AUC

9. Visualization of Results & Model Insights

- Used SHAP to explain the most important features contributing to churn

- Visualized model results using:
  - Confusion matrix
  - SHAP beeswarm and bar plots
  - Feature importance plots

10. Tools and Technologies Used

- **Programming Language:** Python
- **Libraries:** Pandas, NumPy, Scikit-learn, XGBoost, SHAP, Matplotlib, Seaborn
- **Deployment & Visualization:** Streamlit
- **IDE:** VS Code

11. Team Members and Contributions

Name	Contributions
[Nidhidharshini. K]	Team Leader & Data Cleaning EDA
[Janani.V]	EDA (Exploratory Data Analysis)
[Mohana.A]	Feature Engineering
[Charumathi .K]	Model Evaluation
[Iniyavarshini.S]	Deployment