**Student Name:** U FARIHA FEHMIN

**Register Number:** 613023104025

**Institution:** VIVEKANANDHA COLLEGE OF TECHNOLOGY FOR WOMEN

**Department:** COMPUTER SCIENCE AND ENGINEERING

**Date of Submission:** 02-05-2025

**Github Repository Link:** https://github.com/Farihafehmin/Project.git

### **1. Problem Statement**

Customer churn prediction involves identifying customers who are likely to leave a service. This is a **classification problem** (typically binary: churn vs. no churn). Reducing churn helps businesses retain customers and increase revenue.

### **2. Project Objectives**

* Develop a machine learning model to predict customer churn.
* Improve prediction accuracy and recall.
* Gain insights into key churn indicators (e.g., tenure, service issues, usage patterns).
* Apply models like Logistic Regression, Random Forest, and XGBoost.
* Ensure model interpretability for business decisions.

### **3. Flowchart of the Project Workflow**

(Data Collection) → (Preprocessing) → (EDA) → (Feature Engineering) → (Model Building) → (Evaluation) → (Deployment)

### **4. Data Description**

* **Dataset:** Telco Customer Churn (Kaggle)
* **Type:** Structured, tabular
* **Size:** ~7,000 records, ~20 features
* **Target:** Churn (Yes/No)
* **Static dataset**

### **5. Data Preprocessing**

* Handled missing values in TotalCharges via imputation.
* Removed duplicates based on customerID.
* Converted categorical variables using one-hot encoding.
* Scaled numerical features using MinMaxScaler or StandardScaler.

### **6. Exploratory Data Analysis (EDA)**

* Customers with short tenure and month-to-month contracts showed higher churn.
* Services like tech support and internet type showed strong correlation with churn.
* Plotted churn distribution across demographics

### **7. Feature Engineering**

* Created TenureGroup (e.g., short-term, mid-term, long-term).
* Derived binary flags for services (e.g., HasTechSupport).
* Removed redundant columns like customerID.

### **8. Model Building**

* Models used: **Logistic Regression**, **Random Forest**, **XGBoost**
* Data split: 80% training, 20% testing (stratified)
* Evaluation metrics: Accuracy, Precision, Recall, F1-Score, ROC-AUC

### **9. Visualization of Results & Model Insights**

* **Confusion Matrix:** Shows TP, FP, FN, TN.
* **ROC Curve:** AUC > 0.85 indicates good model.
* **Feature Importance:** Contract Type, MonthlyCharges, Tenure most influential.
* Compared model performance in bar plots.

### **10. Tools and Technologies Used**

* **Language:** Python
* **IDE:** Jupyter Notebook
* **Libraries:** pandas, numpy, scikit-learn, seaborn, matplotlib, xgboost
* **Visualization:** Plotly, matplotlib

### **11. Team Members and Contributions**

* Data Cleaning: [S.N.ABHI SHREE]
* Model Development
* EDA: [U.FARIHA FEHMIN]
* Reporting & Documentation
* Feature Engineering: [B.MONIKA]