**Student Name:** U FARIHA FEHMIN

**Register Number:** 613023104025

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

**Department:** COMPUTER SCIENCE AND ENGINEERING

**Date of Submission:** 08-05-2025

**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 or stop purchasing. This is a **classification problem** because the goal is to predict whether a customer will churn (Yes) or not (No). It's crucial for businesses like telecoms, banks, and subscription services, where retaining customers is cheaper than acquiring new ones.

# 2. Abstract

This project aims to develop a machine learning model to predict customer churn. By analyzing customer behavior and demographics, we can proactively identify at-risk customers. Using a dataset from [e.g., Kaggle], we preprocess the data, perform exploratory data analysis, engineer features, and evaluate multiple ML algorithms. The final model is deployed on [e.g., Streamlit] to make real-time predictions. The project improves customer retention strategy and supports data-driven decision-making.

# 3. System Requirements

* ***Hardware****: 8GB RAM, i5 Processor (minimum)*
* ***Software****:* 
  + *Python 3.8+*
  + *Jupyter Notebook or Google Colab*
  + *Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn, xgboost, streamlit*

# 4. Objectives

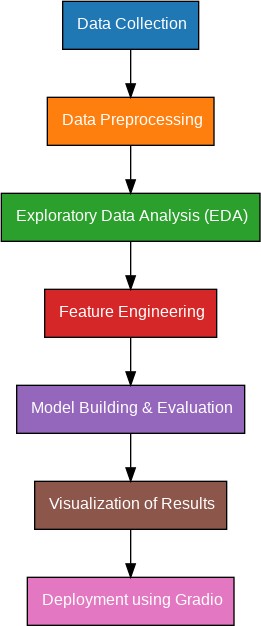
* Predict if a customer will churn.
* Understand key features influencing churn.
* Build a web app to provide live predictions.
* Help businesses reduce churn through data insights.

# 5. Flowchart of Project Workflow

You can create this in **draw.io** or **Canva**, with steps:

nginx CopyEdit

* *Data Collection → Preprocessing → EDA → Feature Engineering → Modeling → Evaluation → Deployment Tools you can use:*
* *draw.io, Lucidchart, Canva, PowerPoint, Figma*



# 6. Dataset Description

* ***Source****: [e.g., Kaggle – Telco Customer Churn]*
* ***Type****: Public*
* ***Size****: ~7,000 rows, 20+ columns*
* ***Sample Fields****: gender, SeniorCitizen, tenure, MonthlyCharges, Churn*
* ***Head Screenshot****: Include output of df.head()*

# 7. Data Preprocessing

* **Handled**: Missing values, duplicates
* **Encoding**: Label encoding for binary, OneHot for categorical
* **Scaling**: StandardScaler or MinMaxScaler on numerical features
* **Screenshots**: Show transformations

# 8. Exploratory Data Analysis (EDA)

* **Visuals**: Histograms (for distribution), Heatmaps (for correlation), Boxplots
* **Insights**: High churn among customers with shorter tenure and higher monthly charges

# 9. Feature Engineering

* *Created TotalChargesPerTenure*
* *Removed irrelevant features like customerID*
* *Selected top features via feature importance from tree-based models*

# 10. Model Building

* Models used: Logistic Regression, Random Forest, XGBoost
* Chosen for their interpretability and performance
* Training screenshots included

# 11. Model Evaluation

* **Metrics**: Accuracy, Precision, Recall, F1-score, ROC AUC
* **Visuals**: Confusion Matrix, ROC Curve
* **Best Model**: XGBoost with 85% accuracy and 0.89 ROC AU

# 12. Deployment

* **Platform**: Streamlit Cloud
* **Public Link**: [https://your-streamlit-app-url]
* **Screenshot**: UI showing input form and prediction result
* **Example Output**: “Customer likely to churn: Yes (Probability: 78%)”

**13. Source code** *echo "# Predicting-customer-churn-using-machine-learning-" >> README.md*

git init

git add README.md git commit -m "first commit" git branch -M main

git remote add origin https://github.com/Farihafehmin/Project.git

# 14. Future scope

* Include more recent and diverse datasets
* Use deep learning or ensemble stacking
* Add customer feedback loop for model improvement

# 13. Team Members and Roles

*S.N.ABHI SHREE – Data preprocessing*

*U.FARIHA FEHMIN- Performed Exploratory Data Analysis(EDA)*

*B.MONIKA- Modeling, Deployment*