**Student Name:** FARIHA FEHMIN U

**Register Number:** 613023104025

**Institution:** VIVEKANANDHA COLLEGE OF TECHNOLOGY

FOR WOMEN

**Department:** COMPUTER SCIENCE AND ENGINEERING

**Date of Submission:** 25-04-2025

# Problem Statement

Customer churn is a critical issue for businesses in subscription-based and service oriented industries. Retaining existing customers is often more cost-effective than acquiring new ones. This project aims to identify patterns in customer behavior that lead to churn, using machine learning models to predict which customers are most likely to leave. Addressing churn allows companies to implement targeted retention strategies and improve long-term profitability.

# Objectives of the Project

# Predict customer churn using historical customer data.

* Uncover hidden patterns and key factors contributing to churn.
* Build a machine learning model with high accuracy and interpretability.
* Provide actionable insights to support customer retention strategies.

# Scope of the Project

* Analyze customer behavior and demographic data.
* Focus on feature engineering, model development, and evaluation.
* Limitations: The project will use a publicly available dataset; real-time deployment is optional; only supervised learning models will be considered.
* Scope excludes live integration with business CRM systems.

# Data Sources

* Dataset: "Telco Customer Churn" dataset from Kaggle
* Source: Kaggle (Public)
* Nature: Static dataset (downloaded once)
* URL: https://www.kaggle.com/blastchar/telco-customer-churn

# High-Level Methodology

# Data Collection – Download dataset from Kaggle

* **Data Cleaning** –Handle missing values, remove duplicates, and correct inconsistent formats (e.g., encoding categorical variables).
* **Exploratory Data Analysis (EDA)** – Use visualizations like histograms, box plots, and heat maps to identify trends and relationships.
* **Feature Engineering** – Create binary variables, convert tenure into categorical ranges, and normalize numerical features.
* **Model Building** – Experiment with Logistic Regression, Decision Trees, Random Forest, and XGBoost.
* Models selected for balance between accuracy and interpretability
* **Model Evaluation** – Use accuracy, precision, recall, F1-score, and ROCAUC as evaluation metrics.
* Apply cross-validation to validate model performance.
* **Visualization & Interpretation** – Present findings using dashboards with bar charts, confusion matrices, and SHAP value plots.
* **Deployment** –Optional web app deployment using Streamlit or Flask for model demonstration

# Tools and Technologies

# Programming Language – Python

* **Notebook/IDE** –Google Colab / Jupyter Notebook
* **Libraries** – pandas, numpy, matplotlib, seaborn, scikit-learn, xgboost, shap
* **Optional Tools for Deployment** – Streamlit, Flask

# 7. Team Members and Roles

* **S**.**N.ABHI SHREE - Project lead:** Defineproject scope, manage team coordination, perform data cleaning, and Contribute to final reporting and presentation.
* **U. FARIHA FEHMIN -Data Scientist & EDA Expert:** Handle data preprocessing, conduct exploratory data analysis (EDA), build and compare machine learning models and contribute to documentation.
* **B. MONIKA -Feature Engineer & Model Evaluator:** Engineer features to improve model accuracy, evaluate models using precision, recall, F1-score, ROC-AUC, and interpret model results using SHAP and confusion matrices. Design dashboards for insight presentation.