Customer Churn Prediction and Analysis Using Machine Learning

Problem Statement:

In today's competitive market, retaining existing customers is more cost-effective than acquiring new ones. Telecom companies face high churn rates due to factors like poor service, high costs, or contract issues. This project aims to predict customer churn using historical customer data, enabling companies to take proactive measures to reduce attrition.

Objectives:

- To explore and analyze customer behaviorl patterns through exploratory data analysis (EDA).
- To train and evaluate machine learning models to predict whether a customer is likely to churn.
- To compare model performance using key metrics (accuracy, precision, recall, F1-score).
- To identify the most influential features contributing to churn.
- To build a user-friendly Streamlit web application for interactive prediction and visualization.

Proposed Methodology:

1. **Data Collection:** Use the public *Telco Customer Churn* dataset from Kaggle, which includes customer demographics, service details, and churn labels.
2. **Data Pre-processing:** Handle missing values, encode categorical variables, and scale numerical features.
3. **Exploratory Data Analysis (EDA):** Visualize churn patterns by demographics, service types, and tenure using bar charts, heatmaps, and correlation matrices.
4. **Model Training and Evaluation:** Train and compare:
    - Logistic Regression
    - Random Forest Classifier
    - Gradient Boosting (XGBoost or LightGBM)

    Use stratified cross-validation and evaluate on accuracy, precision, recall, F1-score, and AUC.

5. **Feature Importance Analysis:** Visualize which features most influence churn using feature importance plots and SHAP values.
6. **Web Application Development:**

Build a multi-page Streamlit app:

- Page 1: EDA visualizations
- Page 2: Model comparison results
- Page 3: Live churn prediction via user input
- Page 4: About/project summary

Dataset Description:

- **Name:** Telco Customer Churn Dataset
- **Source:** Kaggle (https://www.kaggle.com/datasets/blastchar/telco-customer-churn)
- **Size:** 7,043 rows × 21 features
- **Target Variable:** Churn (Yes/No)
- Features include customer demographics, contract details, service usage, and billing information.

Expected Outcomes:

- Identification of key factors contributing to customer churn.
- An accurate and interpretable ML model for churn prediction.
- A functional, user-friendly Streamlit app for interactive use.
- Insights that can help businesses improve customer retention strategies.

Timeline of Activities:

Date	Activity
May 19-20	Finalize dataset, set up folder structure, explore data
May 22–23	Data preprocessing, handle missing values, encode variables
May 24–25	Perform EDA, generate visualizations, extract insights
May 26	Train and evaluate multiple ML models
May 27	Compare models, plot metrics, analyze feature importance
May 28	Build and connect Streamlit UI (EDA + Model pages)
May 29	Add prediction input page, clean UI, test full app
May 30	Final polish