Phase-2 Submission Template:

Student Name: [E.Balamurugan]

Register Number: [623323104006]

Institution: [vetri vinayaha college of engineering and technology]

Department: [CSE]

Date of Submission: [26.04.25]

Title :predicting customer churn using machine learning to uncover hidden pattern

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1.Problem statement:

* Customer churn—the phenomenon of customers leaving a service or product— is a critical issue for many businesses. Acquiring new customers is significantly more expensive than retaining existing ones. However, businesses often lack clarity on why customers churn and who is likely to churn.

* This project aims to develop a predictive machine learning model that can:

* 1. Accurately classify whether a customer is likely to churn.

* 1. Identify hidden behavioral patterns that correlate with churn.

* 1. Provide insights for targeted retention strategies.

2. Project Objectives:

* Collect and understand customer data relevant to churn.

* Preprocess and clean the data for model readiness.

* Explore the data to extract hidden trends and relationships.

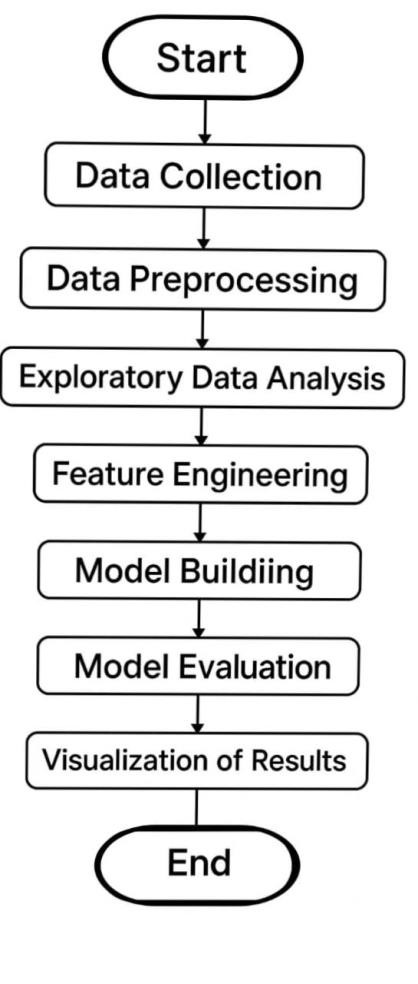
* Engineer meaningful features that enhance model performance.

* Build and evaluate multiple machine learning models.

* Visualize results to draw actionable business insights.

* Recommend strategies for reducing churn based on findings.

3 .Project Workflow (Flowchart):



4.Data Description:

A typical customer churn dataset includes the following types of features:

* CustomerID: Unique identifier
* Demographics: Gender, Age, Location
* Account Info: Tenure, Contract Type, Monthly Charges, Total Charges
* Service Usage: Internet Service, Phone Service, Streaming Services
* Customer Behavior: Number of Complaints, Customer Support Calls
* Churn (Target Variable): Yes / No

5. Data Preprocessing:

* Missing Values Handling:

Impute or remove missing TotalCharges entries (often blank strings).

* Data Type Conversion:

Convert TotalCharges to float.

* Categorical Encoding:

One-hot encode features like Contract, InternetService.

Label encode binary categories like Partner, Dependents.

* Scaling/Normalization:

Standardize MonthlyCharges, TotalCharges, and tenure using StandardScaler.

* Train-Test Split:

Split dataset (e.g., 80% training, 20% testing).

6.Exploratory Data Analysis (EDA):

* Churn Rate Visualization:

Pie charts and bar plots to show overall churn rate.

* Correlation Heatmap:

Show relationship between numerical variables and churn.

* Univariate Analysis:

Distribution of tenure, MonthlyCharges Bar charts for categorical features vs. churn

* Bivariate Analysis: tenure vs. churn Contract type vs. churn

TechSupport vs. churn

* Key Findings from EDA:

Customers with month-to-month contracts are more likely to churn.

Customers without tech support or online security have higher churn rates.

Higher MonthlyCharges often correlate with higher churn.

1. Feature Engineering:
   * New Features

AverageMonthlyCharges = TotalCharges / tenure

Binary feature for highspender based on MonthlyCharges threshold.

Group tenure into bins (e.g., new, intermediate, -term).

* + Interaction Terms:

Combine InternetService and TechSupport for better insights.

* + Remove Redundancies:

Drop features with low variance or multicollinearity.

1. Model Building:
   * Baseline Model:

Logistic Regression (simple, interpretable)

* + Advanced Models:

Decision Tree

Random Frosted

XGBoost

Vector Machine

LightGBM

Support Vector Machine Neural Networks (MLPClassifier)

* + Training and Evaluation:

Cross-validation (k-folord)Hyper parameter tuning using GridSearchCV / RandomizedSearchCV

* + Metrics:

Accurency,Precision, Recall, F1-score, ROC-AUC

1. Visualization of Results & Model Insights:

* + Confusion Matrix:

Evaluate true vs. predicted churn.

* + ROC Curve:

Plot True Positive Rate vs. False Positive Rate.

* + Precision-Recall Curve:

Useful for imbalanced datasets.

* + Feature Importance:

Use Random Forest or XGBoost to rank features. SHAP values for local and global explanations.

* + Customer Segmentation Visualization:

Segment customers based on risk levels.Cluster plots (e.g., K-Means for visualization).

10.Tools and Technologies Used:

* + Programming Languages: Python

* + Libraries: pandas, NumPy, scikit-learn, matplotlib, seaborn, XGBoost, SHAP

* + IDE/Notebook: Jupyter Notebook, Google Colab

* + Version Control: Git/GitHub

* + Optional Deployment: Flask, Streamlit

11. Team Members and Contributions:

* D.AJAI – project coordinator , machine learning modal developer ,web interface developer
* E.BALAMURUKAN -Data collector and preprocessing lead Documentation & presentation lead.
* M.KESAVAN – Data analyst, visualization & Report Design