# Phase-2 Submission Template

**Student Name:** P.AYYAPPAN

**Register Number:** 4127232430

**Institution:** TAGORE ENGINEERING COLLEGE

**Department:** B.TECH.ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

**Date of Submission:** 10-05-2025

**Github Repository Link:**
https://github.com/jivendran/Nm/blob/3a550da8867ce004f7f1f754ac96c2ab9903aa4c/Nm%20pr

---

## 1. Problem Statement

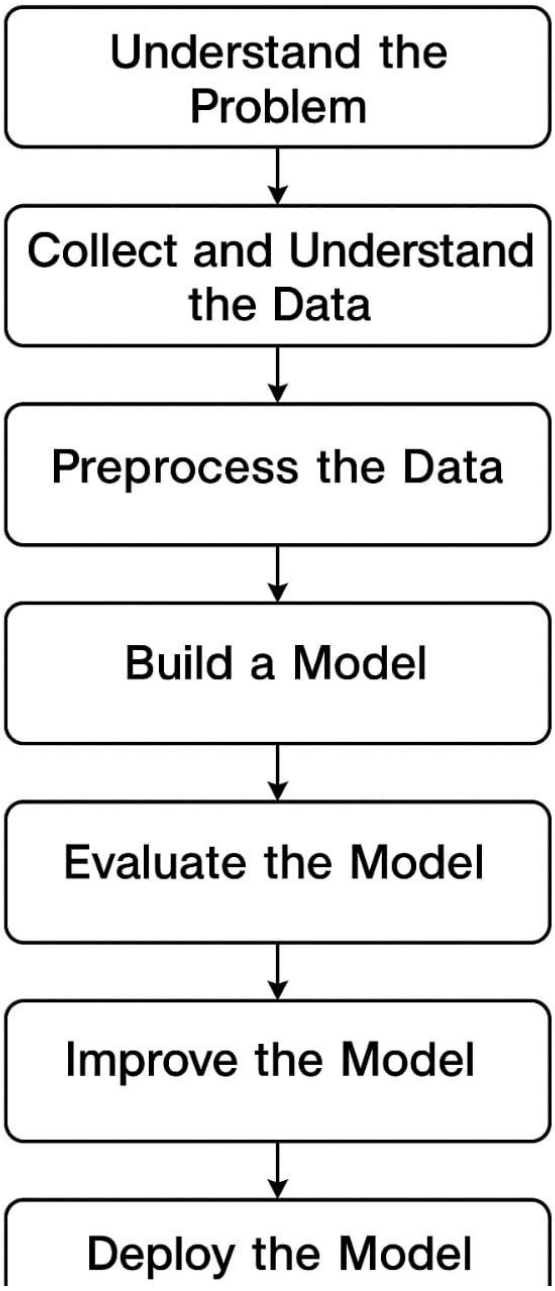
*Customer churn occurs when customers stop using a company's services. This project focuses on predicting churn using machine learning so that businesses can take action to retain high-risk customers. It is a binary classification problem with direct business value, especially in competitive industries like telecom and banking.*

## 2. Project Objectives

- *Develop a machine learning model that accurately predicts customer churn.*
- *Improve model interpretability to understand the reasons behind churn.*
- *Achieve high performance using classification metrics like accuracy, recall, and F1-score.*

- *Adjust objectives based on findings from data exploration.*

## 3. Flowchart of the Project Workflow



## 4. Data Description

- ✓ *Dataset Source: e.g., Kaggle – "Telco Customer Churn" dataset.*
- ✓ *Type: Structured data (CSV).*
- ✓ *Records & Features: ~7,000 rows and 20+ features.*
- ✓ *Static dataset*
- ✓ *Target variable: Churn (Yes/No)*

## 5. Data Preprocessing

- *Handled missing values (e.g., TotalCharges column imputed or dropped).*
- *Removed duplicate records.*
- *Converted categorical variables using one-hot encoding.*
- *Scaled numerical features using standardization.*
- *Transformed data types where necessary (e.g., from object to numeric).*

## 6. Exploratory Data Analysis (EDA)

- ❖ *Univariate Analysis: Plotted histograms, countplots for categorical variables.*
- ❖ *Bivariate Analysis: Checked churn rate across contract types, tenure, monthly charges.*

- *Correlation Analysis: Found features like tenure, Contract, and MonthlyCharges highly correlated with churn.aq*
- *Insights: Customers with month-to-month contracts and high charges are more*

## 7. Feature Engineering

- *Created tenure groups (e.g., new, mid, long-term customers).*
- *Extracted interaction terms between services (e.g., InternetService + Streaming).*
- *Removed redundant features like customerID.*
- *Used domain knowledge to simplify some categories.*

## 8. Model Building

- *Algorithms used: Logistic Regression and Random Forest.*
- *Why: Logistic Regression for interpretability; Random Forest for non-linear patterns.*
- *Train-Test Split: 80-20 stratified split.*
- *Metrics used: Accuracy, Recall, Precision, F1-Score, ROC-AUC.*
- *Initial performance: Random Forest outperformed Logistic Regression on F1-score.*

## 9. Visualization of Results & Model Insights

- *Confusion matrix: Visualized true vs predicted churn.*
- *ROC Curve: Compared classifier performances.*
- *Feature Importance (from RF): Contract, tenure, and MonthlyCharges most important.*
- *Interpretation: Month-to-month contracts are a major churn factor.*

## 10. Tools and Technologies Used

- ✓ *Programming Language: Python*
- ✓ *IDE: Google Colab*
- ✓ *Libraries: pandas, numpy, seaborn, matplotlib, scikit-learn, XGBoost*
- ✓ *Visualization Tools: seaborn, matplotlib, Plotly*

## 11. Team Members and Contributions

- **P. Ayyappa***n: Designed the model architecture and implemented the churn prediction algorithm.*
- **B. Mohamed Fahad***: Managed documentation and prepared the final project report.*
- **M. Jivendran***: Conducted testing, performance evaluation, and debugging of the system.*
- **K. Gopika***: Collected and analyzed churn-related features, supported feature engineering and data simulation*