# Phase-2 Submission Template

**Student Name:**EZHILARASAN.K  
**Register Number:** 422223243015  
**Institution:** SURYA GROUP OF INSTITUTIONS  
**Department:** B.TECH ARTIFICIAL INTELLIGENCE AND DATA SCIENCE  
**Date of Submission:** 13.05.2025  
**Github Repository Link:** <https://github.com/ezhillarasan/nm_phase2>

## 1. Problem Statement

Customer churn is a critical issue for businesses, as losing customers directly impacts revenue and growth. The goal of this project is to predict customer churn using machine learning techniques to identify patterns and factors that contribute to customers leaving a service. By analyzing historical customer data, we aim to build a predictive model that can help businesses take proactive measures to retain at-risk customers.

* **Problem Type:** Classification (predicting whether a customer will churn or not).
* **Impact:** Reducing churn improves customer retention, increases revenue, and enhances business sustainability.

## 2. Project Objectives

* **Technical Objectives:**
  + Preprocess and clean customer data to ensure quality input for modeling.
  + Perform exploratory data analysis (EDA) to uncover trends and correlations.
  + Engineer relevant features to improve model performance.
  + Build and compare multiple classification models (e.g., Logistic Regression, Random Forest, XGBoost).
  + Evaluate models using metrics like accuracy, precision, recall, and F1-score.
  + Identify key features influencing churn to derive actionable insights.
* **Evolution:** After initial data exploration, the focus may shift to addressing class imbalance or optimizing feature selection.

## 3. Flowchart of the Project Workflow

[Visual representation of the workflow:

1. Data Collection → 2. Data Preprocessing → 3. EDA → 4. Feature Engineering → 5. Model Building → 6. Evaluation → 7. Insights]

## 4. Data Description

* **Dataset:** Customer churn dataset (e.g., from Kaggle or telecom companies).
* **Type:** Structured tabular data.
* **Records:** [Number of records, e.g., 10,000].
* **Features:** Demographics, usage patterns, contract details, customer service interactions.
* **Target Variable:** Binary churn indicator (e.g., "Yes" or "No").

## 5. Data Preprocessing

* Handle missing values (impute or remove).
* Remove duplicates and justify retention if any.
* Detect and treat outliers (e.g., using IQR).
* Convert categorical variables (e.g., "Gender," "Contract Type") using one-hot encoding.
* Normalize numerical features (e.g., "Monthly Charges").
* Document all transformations in code and markdown.

## 

## 6. Exploratory Data Analysis (EDA)

* **Univariate Analysis:**
  + Histograms for feature distributions (e.g., "Tenure," "Monthly Charges").
  + Boxplots to identify outliers.
* **Bivariate/Multivariate Analysis:**
  + Correlation matrix to spot relationships (e.g., "Tenure vs. Churn").
  + Grouped bar plots (e.g., "Churn by Contract Type").
* **Insights Summary:**
  + High churn rates among month-to-month contract customers.
  + Features like "Total Charges" and "Tenure" likely influence churn.

## 7. Feature Engineering

* Create new features (e.g., "Average Usage per Month").
* Bin numerical variables (e.g., "Tenure" into groups: 0-12 months, 12-24 months, etc.).
* Apply PCA if dimensionality is high (optional).
* Justify feature additions/removals based on EDA.

## 8. Model Building

* **Models:** Logistic Regression, Random Forest, XGBoost.
* **Justification:**
  + Logistic Regression: Baseline interpretability.
  + Random Forest/XGBoost: Handle non-linear relationships and feature importance.
* **Data Splitting:** 70% training, 30% testing (stratified for class balance).
* **Metrics:** Accuracy, precision, recall, F1-score, ROC-AUC.

## 9. Visualization of Results & Model Insights

* Confusion matrices for each model.
* ROC curves to compare model performance.
* Feature importance plots (e.g., Random Forest).
* Interpret top features (e.g., "Tenure" is a strong churn predictor).

## 10. Tools and Technologies Used

* **Programming Language:** Python.
* **IDE:** Jupyter Notebook.
* **Libraries:** pandas, numpy, seaborn, matplotlib, scikit-learn, XGBoost.
* **Visualization:** Plotly, Matplotlib.

## 11. Team Members and Contributions

* **ABDUL RAHMAN .MS:** Data preprocessing, EDA.
* **AJAY .K :** Feature engineering, model development.
* **BALASUNDAR .R :** Documentation.
* **EZHILARASAN .K :** visualization.