Phase-2 Submission Template

Student Name: [Vasanthakumar.J]

Register Number: [422223243059] Institution: [Suryaq group of institutions]

Department: [Artificial intelligence and data science]

Date of Submission: [5.5.2025]

Github Repository Link: <https://github.com/vasanth-1605/nm-vasanth>

---

1. Problem Statement

Businesses often lose revenue when customers leave or stop using their services. Predicting customer churn--the likelihood that a customer will leave--helps businesses take preventive action.

This project uses supervised machine learning (classification) to predict customer churn based on historical customer data.

Why it matters: Accurate churn prediction can lead to better customer retention strategies and improved business growth.

2. Project Objectives

- Develop and compare machine learning models to classify customers as "churn" or "not churn".

- Improve accuracy, precision, and recall to ensure reliable predictions.

- Use insights to recommend actionable strategies for retention.

- The goal evolved to include feature importance analysis for better interpretability.

3. Flowchart of the Project Workflow

Data Collection -> Data Preprocessing -> EDA -> Feature Engineering -> Model Building -> Evaluation -> Insights

4. Data Description

- Dataset Source: Kaggle - Telco Customer Churn Dataset

- Type: Structured (CSV)

- Records: 7,043 rows, 21 features

- Dataset Type: Static

- Target Variable: "Churn" (Yes/No)

5. Data Preprocessing

- Handled missing values in "TotalCharges" via median imputation.

- Removed 10 duplicate records.

- Outliers identified using boxplots and treated with capping.

- Converted "SeniorCitizen" to categorical.

- Applied one-hot encoding to categorical columns.

- Standardized numerical columns using StandardScaler.

6. Exploratory Data Analysis (EDA)

Univariate:

- "Churn" rate ~27%.

- Most customers use electronic check payment.

Bivariate:

- "Contract type" and "tenure" strongly influence churn.

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

Insights:

- Long-tenure customers tend to stay.

- Fiber optic users show higher churn risk.

- These insights guided feature selection.

7. Feature Engineering

- Created "Tenure Group" by binning tenure into short, medium, long-term groups.

- Extracted number of services used by summing relevant binary columns.

- Removed redundant features like "customerID".

- No dimensionality reduction applied due to manageable feature size.

8. Model Building

Models Used:

- Logistic Regression

- Random Forest Classifier

Why:

- Logistic Regression for interpretability.

- Random Forest for handling complex interactions.

Data Split: 80% training, 20% testing (stratified on "Churn")

Metrics (on test data):

- Logistic Regression: Accuracy = 0.80, F1 = 0.65

- Random Forest: Accuracy = 0.83, F1 = 0.70

9. Visualization of Results & Model Insights

- Confusion Matrix: Random Forest showed fewer false negatives.

- ROC Curve: AUC = 0.86

- Feature Importance: "Contract", "MonthlyCharges", and "Tenure" are top features.

- Visualizations confirmed that model decisions align with business logic.

10. Tools and Technologies Used

- Language: Python

- IDE: Google Colab

- Libraries: pandas, numpy, seaborn, matplotlib, scikit-learn

- Visualization Tools: seaborn, matplotlib, Plotly

11. Team Members and Contributions

- ragul.s- Data cleaning, EDA

-vasanthakumar.j - Feature engineering, modeling

-gopinath.d- Documentation, visualizations,

-Kabilan.m- reporting