**Phase-1 Submission Template**

**Student Name:** S. MAHALAKSHMI

**Register Number:** 511523205033

**Institution:** P. T. Lee Chengalvaraya Naicker College of Engineering and Technology

**Department:** B. Tech Information Technology / 2 year

**Date of Submission**: 09.05.2025

1. **Problem Statement**

* Healthcare systems often struggle to predict diseases early due to lack of integrated data analysis. Manual diagnosis can be time-consuming and error-prone. By leveraging AI to predict diseases using patient data, healthcare providers can identify conditions early, improve treatment plans, and enhance patient outcomes.

1. **Objectives of the Project**

* Develop a predictive model that identifies potential diseases based on patient data.
* Improve diagnostic efficiency using machine learning algorithms.
* Deliver actionable insights to assist healthcare professionals in decision-making.
* Enhance early detection and prevention strategies.

1. **Scope of the Project**

* Analyze medical datasets with patient demographics, symptoms, test results, etc.
* Build and compare different machine learning models for disease prediction.
* Visualize results to provide interpretability.
* **Constraints:**

Limited to structured patient data; deployment may be restricted to local environments or web-based dashboards.

* **Model scope:**

**Focused on** predicition of chronic diseases(e.g., diabetes,heat disease).

**4.Data Sources**

* **Source:** Public datasets from platforms like Kaggle or UCI Machine Learning Repository.
* **Type:** Public, static datasets.
* **Example:** Disease symptom dataset, patient health records.

**5.High-Level Methodology**

* **Data Collection**:

Download from Kaggle or UCI repositories.

* **Data Cleaning**:

Handle missing values, remove duplicates, normalize formats.

* **EDA:**

Use histograms, correlation heatmaps, and pair plots to understand patterns.

* **Feature Engineering**:

Derive features such as symptom duration, age groups, etc.

* **Model Building**:

Use models like Logistic Regression, Decision Trees, Random Forests, and XGBoost.

* **Model Evaluation:**

Use accuracy, precision, recall, F1-score, and confusion matrix.

* **Visualization & Interpretation:**

Use graphs, bar charts, and confusion matrices for presenting results.

* **Deployment:**

Optionally deploy using Streamlit or Flask as a web app for demonstration.

**6.Tools and Technologies**

* Programming Language: Python
* Notebook/IDE: Google Colab
* Libraries: pandas, numpy, seaborn, matplotlib, scikit-learn, xgboost
* Optional Deployment Tools: Streamlit, Flask

**7.Team Members and Roles**

1. **S. MAHALAKSHMI:**

Data collection and cleaning

1. **S.K.LAKSHMIPRIYA:**

Exploratory data analysis and visualization

1. **C.NITHYAPRIYA:**

Model development and evaluation

1. **V.SANGEETHA:**

Report writing and presentation preparation