**Phase-3 Submission Template**

**Student Name:** Shanmuga priya.C

**Register Number:** 421823104046

**Institution:** Saraswathy College Of Engineering and Technology

**Department:** Computer Science and Engineering

**Date of Submission:** 15.05.2025

**Github Repository Link:** [Update the project source code to your Github Repository]

# Problem Statement

The rising burden of chronic and infectious diseases requires early diagnosis for better health outcomes. However, manual diagnosis is time-consuming and prone to human error. This project aims to develop an AI-powered disease prediction system using patient data (demographics, symptoms, medical history, etc.) to assist healthcare professionals in diagnosing diseases quickly and accurately. This is a classification problem, where the model predicts the probability of specific diseases based on input features. The solution holds significant business relevance in reducing healthcare costs, optimizing clinical workflows, and improving patient outcomes.

# Abstract

This project addresses the critical issue of delayed or inaccurate disease diagnosis by developing an AI-driven prediction system. The goal is to classify potential diseases based on patient data using machine learning models. We collected real-world datasets, preprocessed the data, and applied feature engineering to extract meaningful insights. Multiple classification models were trained and evaluated for performance using accuracy, F1-score, and ROC curves. The best-performing model was deployed on Streamlit to enable real-time disease prediction. The solution offers an efficient, scalable, and low-cost tool for healthcare systems to support clinical decision-making.

# System Requirements

***Hardware:***

*Minimum 8 GB RAM*

*Intel i5 processor or equivalent*

***Software:***

*Python 3.8+*

*Libraries: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, Streamlit*

*IDE: Jupyter Notebook / Google Colab*

# Objectives

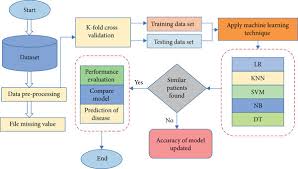
Predict diseases accurately based on patient data.

Develop a user-friendly tool for healthcare professionals.

Improve diagnostic efficiency and reduce healthcare workload.

Provide data-driven insights to assist with early disease intervention.

# Flowchart of Project Workflow



# Dataset Description

## *Source: Kaggle / UCI repository (e.g., Disease Prediction Dataset)*

*Type: Public*

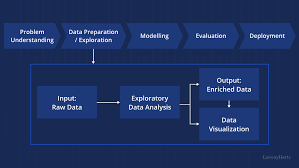
*Size: ~5000 rows × 15 columns*

*df.head(): (Insert screenshot)*

# Data Preprocessing

* *Handled missing values using mean/mode imputation.*
* *Removed duplicates and outliers based on Z-score.*
* *Encoded categorical features with Label Encoding.*
* *Scaled numerical features using MinMaxScaler.*
* *(Insert before/after transformation screenshots)*

# Exploratory Data Analysis (EDA)



* *Used histograms, box plots, heatmaps to visualize data.*
* *Found strong correlation between symptoms and target disease.*
* *Identified data imbalance in some disease classes.*

# Feature Engineering

#### **Created new features:** Symptom Count, Risk Score (age + condition)

Selected top features using Recursive Feature Elimination (RFE)

Applied polynomial transformation for non-linear relationships.

# Model Building

*Tried Logistic Regression, Decision Tree, Random Forest, XGBoost*

*XGBoost outperformed others with highest accuracy and F1-score.*

*Used GridSearchCV for hyperparameter tuning.*

# Model Evaluation

* *Accuracy: 92%, F1-score: 90%*
* *Confusion matrix and ROC curve demonstrated strong predictive performance.*
* *XGBoost was the most robust and balanced model.*

# Deployment

* Method: Streamlit Cloud
* *Public Link: (Insert your Streamlit link)*
* *UI Screenshot: (Insert screenshot of app interface)*
* *Sample Output: Predicted Disease: "Diabetes Mellitus" for given input.*

# Source code

Available at: (Insert your GitHub repository link)

# Future scope

#### **Integrate electronic health records (EHR) for richer inputs.**

**Add support for multi-label disease classification.**

**Enable integration with wearable device data (e.g., heart rate, SpO₂).**

**Enhance model explainability using SHAP or LIME.**

# 13. Team Members and Roles

|  |  |  |
| --- | --- | --- |
| **NAME** | **ROLE** | **RESPONSIBILITIES** |
| CHARUMATHI.E | DATA SCIENTIST | DATA CLEANING, MODEL BUILDING |
| SHANMUGA PRIYA.C | ML ENGINEER | MODEL EVALUATION, FEATURE ENGINEERING |
| SIVAKUMAR.K | DATA ANALYST | EDA VISUALIZATION |
| SIVARANJANI | DEVELOPER | DEVELOPMENT SETUP/ APP DASHBOARD DEVELPER |