Phase-2 Submission Template

Student Name: DEEPAK
Register Number: 510623104017

Institution: C.Abdul Hakeem College Of Engineering And Technology

Department: B.E Computer science and engineering

Date of Submission: 09/05/2025
Github Repository Link:
https://github.com/Ismail23617/TRANSFORMING-HEALTHCARE-WITH-AI-POWERED-DISEASE-PREDICTION-BASED-ON-PATIENT-DATA

1. Problem Statement

The project focuses on early and accurate disease prediction using AI by leveraging patient health data.

This is a classification problem, where the goal is to predict whether a patient is likely to develop a particular disease based on health parameters.

Solving this problem helps healthcare systems proactively address risks, reduce diagnosis delays, and improve patient outcomes through timely interventions.

2. Project Objectives

Build and evaluate machine learning models for disease prediction using structured health datasets.
Improve model accuracy and interpretability for clinical usability.
Enable healthcare providers to use prediction outputs in real-time decision *making. After initial exploration, the focus also includes explainable AI using SHAP values.*

3. Flowchart of the Project Workflow

1. Data Collection

2. Data Preprocessing

3. Exploratory Data Analysis (EDA)

4. Feature Engineering

5. Model Building

6. Model Evaluation

7. Visualization & Interpretation

8. Deployment (optional)

4. Data Description

Dataset Used: PIMA Indian Diabetes Dataset

Source: UCI Machine Learning Repository

Type: Structured (CSV format)

Records & Features: 768 rows, 8 features

Target Variable: 'Outcome' (1 = Diabetic, 0 = Non-diabetic)

Nature: Static dataset

5. Data Preprocessing

Replaced missing values (e.g., zeros in glucose, BMI) with median imputation.

Removed duplicate rows.Converted data types for consistency.

Applied Min-Max scaling for normalization.Label encoding applied where necessar

6. Exploratory Data Analysis (EDA)

Univariate Analysis: Glucose and BMI had higher values in diabetic cases.

Bivariate Analysis: Correlation heatmap revealed glucose and insulin had the highest correlation with diabetes.

Insights: Glucose, BMI, and age are strong predictors.

Used visualizations like histograms, box plots, and pair plots.

7. Feature Engineering

Created new feature: BMI divided by Age (BMI/Age).

Removed features with very low variance.

Standardized continuous variables.

Selected top features using mutual information and domain knowledge.

8. Model Building

Models Used: Logistic Regression, Random Forest

Train-Test Split: 80:20

Random Forest Accuracy: ~81%

Logistic Regression Accuracy: ~76%

Metrics Used: Accuracy, Precision, Recall, F1-Score

Hyperparameter tuning via GridSearchCV improved model performance

9. Visualization of Results & Model Insights

Confusion Matrix: Displayed classification success.

ROC Curve: AUC for Random Forest was 0.84

Feature Importance: Glucose and BMI ranked highest.

Interpretation Tools: SHAP values for transparency and feature influence.

10. Tools and Technologies Used

Programming Language: Python

Notebook Environment: Google Colab

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

Visualization: matplotlib, seaborn

(Optional) Deployment: Streamlit + Flask

11. Team Members and Contributions

HarishKumar – Project Lead, Coordination, Communication

BalaMurugan – Data Preprocessing, EDA, Feature Engineering

Deepak – Model Building, Hyperparameter Tuning

AbineshGodwin – Web Interface, Streamlit Integration

MohammadAdnan – Report Writing, Testing, QA

12. Visualization of Results & Model Insights

Confusion Matrix: Displayed classification success.

ROC Curve: AUC for Random Forest was 0.84

Feature Importance: Glucose and BMI ranked highest.

Interpretation Tools: SHAP values for transparency and feature influence.

13. Tools and Technologies Used

Programming Language: Python

Notebook Environment: Google Colab

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

Visualization: matplotlib, seaborn

(Optional) Deployment: Streamlit + Flask

14. Team Members and Contributions

HarishKumar – Project Lead, Coordination, Communication

BalaMurugan – Data Preprocessing, EDA, Feature Engineering

Deepak – Model Building, Hyperparameter Tuning

AbineshGodwin – Web Interface, Streamlit Integration

MohammadAdnan – Report Writing, Testing, QA