**Phase-2 Submission Template**

**Student Name:** C.Malavika

**Register Number:** 620123106061

**Institution:** AVS Engineering College

**Department:** ECE

**Date of Submission:** 10.05.2025

**Github Repository Link**  : https://github.com/Malavika09-04/Malavika.git

# 1. Problem Statement

The healthcare industry faces challenges in early and accurate disease diagnosis, leading to delayed treatments and increased healthcare costs. Traditional diagnostic methods are often time-consuming and prone to human error. The problem is to develop an AI-powered system that can predict potential diseases using patient data (like symptoms, medical history, demographics, and lab results) to enhance early diagnosis, treatment planning, and overall patient care.

# 2. Project Objectives

\* To collect and preprocess patient health data for predictive modeling.

\* To explore and identify patterns and correlations within the data.

\* To develop and evaluate AI/ML models for disease prediction.

\* To interpret model results and provide actionable insights.

\* To visualize predictions and trends for clinical decision support.

**3. Flowchart of the Project Workflow**

**+---------------------+**

**| Data Collection |**

**+---------+-----------+**

**+---------v-----------+**

**| Data Preprocessing |**

**+---------+-----------+**

**+---------v-----------+**

**| Exploratory Data |**

**| Analysis (EDA) |**

**+---------+-----------+**

**+---------v-----------+**

**| Feature Engineering |**

**+---------+-----------+**

**+---------v-----------+**

**| Model Building |**

**+---------+-----------+**

**+---------v-----------+**

**| Model Evaluation |**

**+---------+-----------+**

# 4. Data Description

The dataset includes the following fields

\* Patient ID

\* Age

\* Gender

\* Symptoms

\* Medical History

\* Lab Test Results

\* Diagnosed Diseases (Target Variable)

# 5. Data Preprocessing

\* Handling missing values.

\* Converting categorical variables into numerical (e.g., one-hot encoding).

\* Normalization/Standardization of numerical data.

\* Removing duplicates.

\* Balancing the dataset (e.g., using SMOTE for imbalanced classes).

# 6. Explor

# \* Visualizing age and gender distribution.

# \* Analyzing disease prevalence.

# \* Correlation heatmaps of features.

# \* Distribution plots of symptoms and lab results.atory Data Analaysis ( EDA)

# 7. Feature Engineering

\* Creating symptom clusters.

\* Aggregating lab test metrics.

\* Feature selection using correlation and importance scores.

\* Dimensionality reduction techniques (e.g., PCA)

8. Model Building\*

\* Algorithms used: Logistic Regression, Random Forest, XGBoost, Neural Networks.

\* Train-test split.

\* Hyperparameter tuning using Grid Search/Cross Validation.

\* Evaluation Metrics: Accuracy, Precision, Recall, F1 Score, AUC-ROC.

# 9. Visualization of Results & Model Insights

\* Confusion matrix heatmap.

\* ROC curve and AUC score.

\* Feature importance plots.

\* SHAP values for explainable AI.

# 

# 10. Tools and Technologies Used

\* Python

\* Pandas, NumPy

\* Scikit-learn, XGBoost

\* TensorFlow/Keras

\* Matplotlib, Seaborn

\* Jupyter Notebook/Google Colab

\* Git/GitHub

# 11. Team Members and Contributions

# \* \*M.Harini\* : Data collection, preprocessing, and EDA

# \* \*A.Iswarya\* : Feature engineering and model building

# \* \*S.Karthika\*: Model evaluation and visualization

# \* \*C.Malavika\*: Documentation and presentation

# 