Phase-1 Submission

Student Name: SANJAY S

Register Number: 712523205050

Institution: PPG INSTITUTE OF TECHNOLOGY

Department: INFORMATION TECHNOLOGY

Date of Submission: 25.04.2025

1.Problem Statement

Modern healthcare systems generate vast amounts of patient data, yet much of it remains underutilized for proactive care. Delays in diagnosis can lead to complications, reduced treatment effectiveness, and higher healthcare costs. This project addresses the need for early and accurate disease prediction using AI-based models trained on patient data. By leveraging machine learning, we aim to empower healthcare providers to make data-driven decisions, reduce diagnostic time, and improve patient outcomes.

2.Objectives of the Project

- *Develop AI models capable of predicting diseases based on patient health records.*
- *Analyze key patterns and correlations in patient data relevant to disease prediction.*
- *Evaluate the performance of different machine learning models on medical datasets.*
- *Provide interpretable insights to support clinical decision-making.*

3.Scope of the Project

Features:

- **Early Detection:** *Identifies diseases at an early stage, improving treatment outcomes.*
- **Personalized Treatment:** *Tailors healthcare plans to individual patient profiles.*
- **Resource Optimization:** *Allocates medical resources efficiently based on predicted needs.*
- **Improved Patient Engagement:** *Empowers patients with information about their health risks*

Constraints:

- *Limited to structured patient data*
- *May use only publicly available datasets*
- *Focused on model prototyping (no real-time deployment)*

4.Data Sources

- **Dataset:** *Publicly available datasets from platforms such as Kaggle or UCI Machine Learning Repository (e.g., Heart Disease, Diabetes, or COVID-19 datasets).*
- **Type:** *Public and static*
- **Source:** *Downloaded datasets*

Kaggle heart disease:https://www.kaggle.com/code/desalegngeb/heart-disease-predictions

The AI system utilizes diverse data sources, including:

- **Electronic Health Records (EHRs):** *Medical history, diagnoses, medications, treatment plans.*
- **Laboratory Results:** *Blood tests, imaging reports, pathology findings*.

- **Genetic Information:** *Genomic data for hereditary disease risk assessment.*
- **Lifestyle Data:** *Diet, exercise, smoking habits, alcohol consumption.*
- **Wearable Devices:** *Heart rate, sleep patterns, physical activity levels*
- **Wearable Devices:** *Heart rate, sleep patterns, physical activity levels*

5.High-Level Methodology

- **Data Collection:** *Obtain patient datasets from Kaggle/UCI.*
- **Data Cleaning:** *Handle missing values, remove duplicates, standardize formats.*
- **Exploratory Data Analysis (EDA):** *Use seaborn/matplotlib for visualizing distributions and correlations.*
- **Feature Engineering:** *Generate features like risk scores, symptom categories, BMI, etc.*
- **Model Building:** *Apply models such as Logistic Regression, Random Forest, and XGBoost for prediction.*
- **Model Evaluation:** *Use accuracy, precision, recall, F1-score, and confusion matrix.*
- **Visualization & Interpretation:** *Create dashboards and charts to present key findings.*
- **Deployment:** *Optional deployment using Streamlit or Flask (for demo purposes).*

6.Tools and Technologies

List the tools, programming languages, and libraries you plan to use in your project. Include the following details

- **Programming Language** – *Python*

- **Notebook/IDE** – *Jupyter Notebook/Google Colab*
- **Libraries** – *Pandas, numpy, seaborn, matplotlib, scikit-learn, TensorFlow*
- **Optional Tools for Deployment** – *Stream lit or Flask,*

7.Team Members and Roles

- **Thamil selvan** – *Data Collection, Data Cleaning*
- **Tamilvanan K**– *Exploratory Data Analysis, Feature Engineering*
- **Rajeshwari S**– *Model Building and Evaluation*
- **Vignesh M** – *Visualization, Interpretation, and (optional) Deployment*