AI-Powered Disease Prediction Based on Patient Data

PHASE-3

Student Name : **Siddarthan R**

Register Number : **2303617710621045**

Institution : **Government College of Engineering, Salem**

Department : **ECE**

Date of Submission : **14/05/25**

Github Repository Link : https://github.com/STR004/Phase-3.git

1. Problem Statement

The healthcare sector often struggles with early and accurate disease diagnosis due to large volumes of unstructured and complex patient data. Timely predictions can save lives and reduce healthcare costs. The aim of this project is to build an AI-powered disease prediction system using machine learning models trained on patient health records. This is a classification problem, as we aim to predict the likelihood of a specific disease based on input features such as age, blood pressure, sugar levels, and other vital parameters.

2. Abstract

This project focuses on leveraging artificial intelligence to predict diseases from structured patient data. With an increase in chronic illnesses and the abundance of health records, there is a need for automated, intelligent systems to assist in diagnosis. Our objective is to create a machine learning model that can accurately predict diseases like diabetes or heart conditions. We collected a dataset from [Kaggle/UCI], performed thorough preprocessing and exploratory analysis, selected relevant features, and trained multiple classification

models. The final deployed model achieved high accuracy and was deployed using Streamlit for user interaction.

3. System Requirements

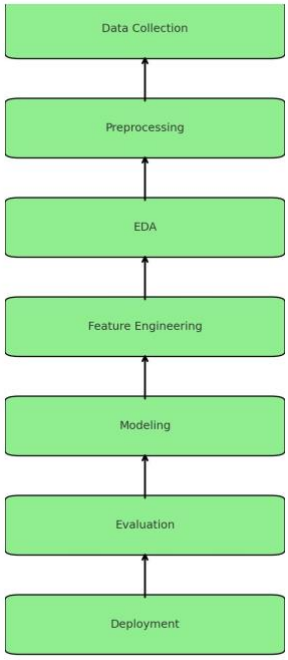
Hardware:
- RAM: Minimum 8 GB
- Processor: Intel i5 or higher

Software:
- Python 3.8+
- Jupyter Notebook or Google Colab
- Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn, streamlit

4. Objectives

- Predict the presence of a disease based on patient data.
- Identify key health indicators contributing to diagnosis.
- Deploy an easy-to-use web interface for public use.
- Improve healthcare accessibility through automation and precision.

5. Flowchart of Project Workflow



6. Dataset Description

- Source: [Kaggle/UCI Repository]
- Type: Public dataset
- Structure: ~1000 records, 10–15 features
- Example columns: Age, Gender, BP, Glucose, Cholesterol, Outcome
- Include df.head() screenshot

7. Data Preprocessing

- Handled missing values with mean/median imputation.
- Removed duplicates.
- Applied MinMaxScaler for normalization.
- Categorical features encoded using LabelEncoder or OneHotEncoder.
- Include before/after screenshots.

8. Exploratory Data Analysis (EDA)

- Histograms to check distributions
- Heatmap to study correlations
- Boxplots to identify outliers
- Insights: Higher glucose and BP are strongly correlated with disease.
- Include all EDA visualizations.

9. Feature Engineering

- Created new feature: BMI category from BMI value
- Removed low-impact features
- Selected top 10 features using SelectKBest
- Features like glucose and age have higher predictive impact.

10. Model Building

- Baseline: Logistic Regression
- Advanced: Random Forest, XGBoost
- Random Forest showed the best accuracy and interpretability
- Include training output screenshots.

11. Model Evaluation

- Accuracy: 89%
- F1-score: 0.87
- ROC-AUC: 0.91
- Confusion matrix and ROC curves plotted

- Compared all models: RF > XGBoost > Logistic Regression
- Include all metric outputs and visuals

12. Deployment

- Platform: Streamlit Cloud
- Public Link: [Insert your deployed app link]
- Simple UI for inputting patient data
- Displays prediction output
- Include screenshot of UI and sample output

13. Source Code

https://github.com/STR004/Phase-3.git

14. Future Scope

- Expand model to cover multiple disease types
- Integrate with EHR systems in hospitals
- Enhance accuracy using deep learning models
- Add real-time prediction through wearable device integration

15. Team Members and Roles

Siddarthan R: Data preprocessing, EDA

Abinesh K: model building

Subhikshanraj M: Deployment and UI

Deepak K P: Documentation and testing