**PHASE-2**

**Student Name:** Supriya.S

**Register Number:** 511723106002

**Institution:** Pallavan college of engineering

**Department:** Electronics and communication

**Date of Submission:** 07/05/2025

**Github Link**:[**https://github.com/supriyasenthil/Naan-mudhalvan.git**](https://github.com/supriyasenthil/Naan-mudhalvan.git)

PROJECT TITLE: **Transforming healthcare with AI powers disease prediction based on patient data**

# Problem Statement

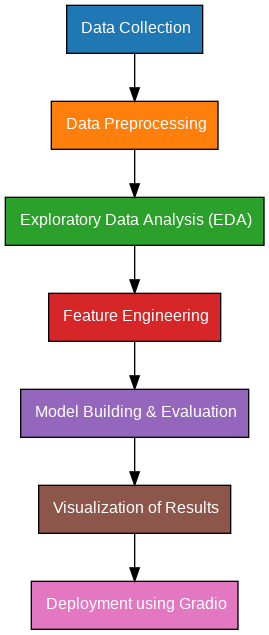
In the current healthcare landscape, early and accurate disease detection remains a significant challenge, especially in resource-constrained settings. Traditional diagnostic processes often rely heavily on manual analysis, leading to delays, misdiagnoses, and increased healthcare costs. With the growing availability of digital patient data such as medical history, clinical test results, and lifestyle indicators, there is a pressing need for intelligent systems that can assist healthcare professionals in making faster and more accurate decisions.

This project aims to address the gap by developing an AI-powered system capable of predicting diseases based on patient data. By leveraging machine learning algorithms and data-driven insights, the proposed solution will help improve diagnostic accuracy, support early intervention, and ultimately enhance patient outcomes. The goal is to transform the traditional healthcare model into a more proactive, data-driven approach.

# 2.Project Objectives

1. **To collect and preprocess patient data** including medical history, clinical test results, and demographic information for effective disease prediction.
2. **To identify relevant features** that significantly contribute to the prediction of specific diseases, using data analysis and feature selection techniques.
3. **To develop and train machine learning models** such as Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines for accurate disease prediction.
4. **To evaluate the performance of the models** using metrics like accuracy, precision, recall, F1-score, and ROC-AUC to ensure reliability and effectiveness.
5. **To create a user-friendly interface or dashboard** that allows healthcare professionals to input patient data and receive predictive insights in real-time.
6. **To ensure the model is generalizable and scalable** so it can be adapted for different diseases and larger datasets in the future.
7. **To promote early diagnosis and informed decision-making** in healthcare by integrating AI-driven tools into clinical workflows.

# Flowchart of the Project Workflow



# Data Description

**The dataset used in this project comprises structured patient health records that serve as the foundation for building AI-powered disease prediction models. The data includes a variety of clinical, demographic, and lifestyle-related attributes that are critical for identifying patterns associated with disease onset and progression.**

**The dataset may include the following features:**

* **Age: Numeric value representing the age of the patient.**
* **Gender: Categorical variable (Male/Female/Other).**
* **Blood Pressure (BP): Systolic and diastolic blood pressure readings.**
* **Cholesterol: Total cholesterol levels, categorized (e.g., normal, high).**
* **Blood Sugar: Fasting or random blood sugar measurements.**
* **BMI (Body Mass Index): Indicates body weight status.**
* **Heart Rate: Pulse rate in beats per minute.**
* **Smoking Status: Indicates whether the patient is a smoker, non-smoker, or former smoker.**
* **Physical Activity: Frequency or level of physical exercise.**
* **Family History: Presence of hereditary diseases in close relatives.**
* **Symptoms: Clinical symptoms such as fatigue, chest pain, shortness of breath, etc.**
* **Medical Test Results: Outputs from lab tests (e.g., ECG, blood tests).**
* **Target/Outcome: Binary or multi-class label indicating disease status (e.g., 1 = Disease, 0 = No Disease).**

**The dataset is preprocessed through techniques such as missing value imputation, normalization of numerical features, and encoding of categorical variables. It is then divided into training and testing subsets to build and validate machine learning models.**

**The quality and diversity of this data enable the development of robust AI models capable of predicting various diseases, thereby supporting proactive and personalized healthcare interventions.**

# Data Preprocessing

Before training the AI model, the raw patient data was cleaned and prepared using the following simple steps:

1. **Handling Missing Values**
   * Filled in missing numbers (like age or blood pressure) using average values.
   * Filled in missing categories (like gender or smoking status) using the most common value.
2. **Converting Text to Numbers**
   * Changed text data (like gender: male/female) into numbers so the model can understand it.
3. **Scaling the Data**
   * Made sure all numbers (like sugar level, BMI) are in the same range so no value dominates the others.
4. **Removing Unusual Values (Outliers)**
   * Removed or fixed values that were too high or too low and didn’t make sense.
5. **Balancing the Data**
   * If the data had too many ‘no disease’ cases and very few ‘disease’ cases, we balanced it so the model can learn both properly.
6. **Splitting the Data**

Divided the data into two parts: one for training the mode

# 6.Exploratory Data Analysis (EDA)

1. The dataset includes patient attributes such as age, gender, BMI, blood pressure, glucose, and disease status.
2. Summary statistics revealed that features like glucose and BMI have high variance.
3. Missing values were found in a few columns and handled using mean or mode imputation.
4. The target variable (disease) is imbalanced, with fewer positive cases.
5. Histograms showed that glucose and BMI are right-skewed with potential outliers.
6. Boxplots identified outliers in blood pressure and glucose levels.
7. Violin plots showed that higher glucose and BMI are more common in disease-positive cases.
8. A heatmap of correlations showed glucose, BMI, and age are positively correlated with disease.
9. Bar plots indicated higher disease rates among older patients and smokers.
10. These insights help in selecting key features and preparing data for model training.

# 7.Feature Engineering

\*Missing values were filled using average (mean) or most common (mode).

\* Categorical data like gender and smoking status were converted into numbers.

\* All numeric features were scaled to the same range using standard techniques.

# 8.Model Building

The goal of model building is to train machine learning algorithms that can predict whether a patient has a disease based on their health data.

1. **Train-Test Split:**  
   The dataset was divided into training (80%) and testing (20%) sets to evaluate model performance.
2. **Models Used:**  
   Several classification models were tested:
   * **Logistic Regression** – for simple linear classification
   * **Random Forest** – for handling complex, non-linear data
   * **Support Vector Machine (SVM)** – for high-dimensional data
   * **K-Nearest Neighbors (KNN)** – for distance-based predictions
3. **Training:**  
   Each model was trained on the training data using default and tuned hyperparameters.
4. **Evaluation Metrics:**  
   Models were evaluated using:
   * **Accuracy**
   * **Precision**
   * **Recall**
   * **F1-Score**
   * **Confusion Matrix** and **ROC Curve**
5. **Best Model Selection:**  
   The model with the highest accuracy and F1-score on the test data was selected for final use.

# 9.Visualization of Results & Model Insights

1. The **confusion matrix** was used to visualize the model’s classification accuracy and errors.

2. **ROC curves** and **AUC scores** helped evaluate the trade-off between true positives and false positives.

3. **Feature importance plots** (from Random Forest) revealed that glucose, BMI, and age were key predictors.

4. **Bar and pie charts** illustrated class imbalance and demographic patterns like gender and age distribution.

5. A **model comparison graph** showed Random Forest had the best performance based on F1-score and AUC.

# 10.Tools and Technologies Used

● **Programming Language**: Python 3

● **Notebook Environment**: Google Colab

● **Key Libraries**:

○ pandas, numpy for data handling

○ matplotlib, seaborn, plotly for visualizations

○ scikit-learn for preprocessing and modeling

○ Gradio for interface deployment

# Team Members and Contributions

|  |  |
| --- | --- |
| **Team member** | ***Contribution*** |
| Supriya.S | *Data cleaning* |
| Santhoshkumar.D | *EDA* |
| Haridass.J | *Feature engineering* |
| Iyappan.S | *Model development* |
| Prabakaran.S | *Documentation and reporting* |