Github link-[https://github.com/Ssathishkumar2325/disease-prediction](https://github.com/Ssathishkumar2325/disease-prediction%20)

Project Title: Transforming Healthcare with AI: Disease Prediction from Patient Data

PHASE-2

Student Name:Sathish kumar.S

Register Number:612823205056

Institution:Varuvan Vadivelan Institute of Technology

Department:B.Tech IT

Date of Submission:10-05-2025

1. Problem Statement

Accurate and early prediction of diseases is a growing challenge in the healthcare industry. The increasing volume of electronic health records (EHRs) and patient data provides a unique opportunity to harness machine learning (ML) and artificial intelligence (AI) for predictive diagnostics. The objective is to develop a model capable of forecasting the presence or risk of disease based on patient attributes like medical history, demographic profile, lifestyle indicators, and clinical parameters.

Solving this problem has the potential to revolutionize clinical decision-making, reduce healthcare costs, improve patient outcomes, and enable proactive intervention strategies.

---

2. Project Objectives

Develop an AI-driven model to predict disease likelihood using patient data.

Identify the most influential factors contributing to disease risk.

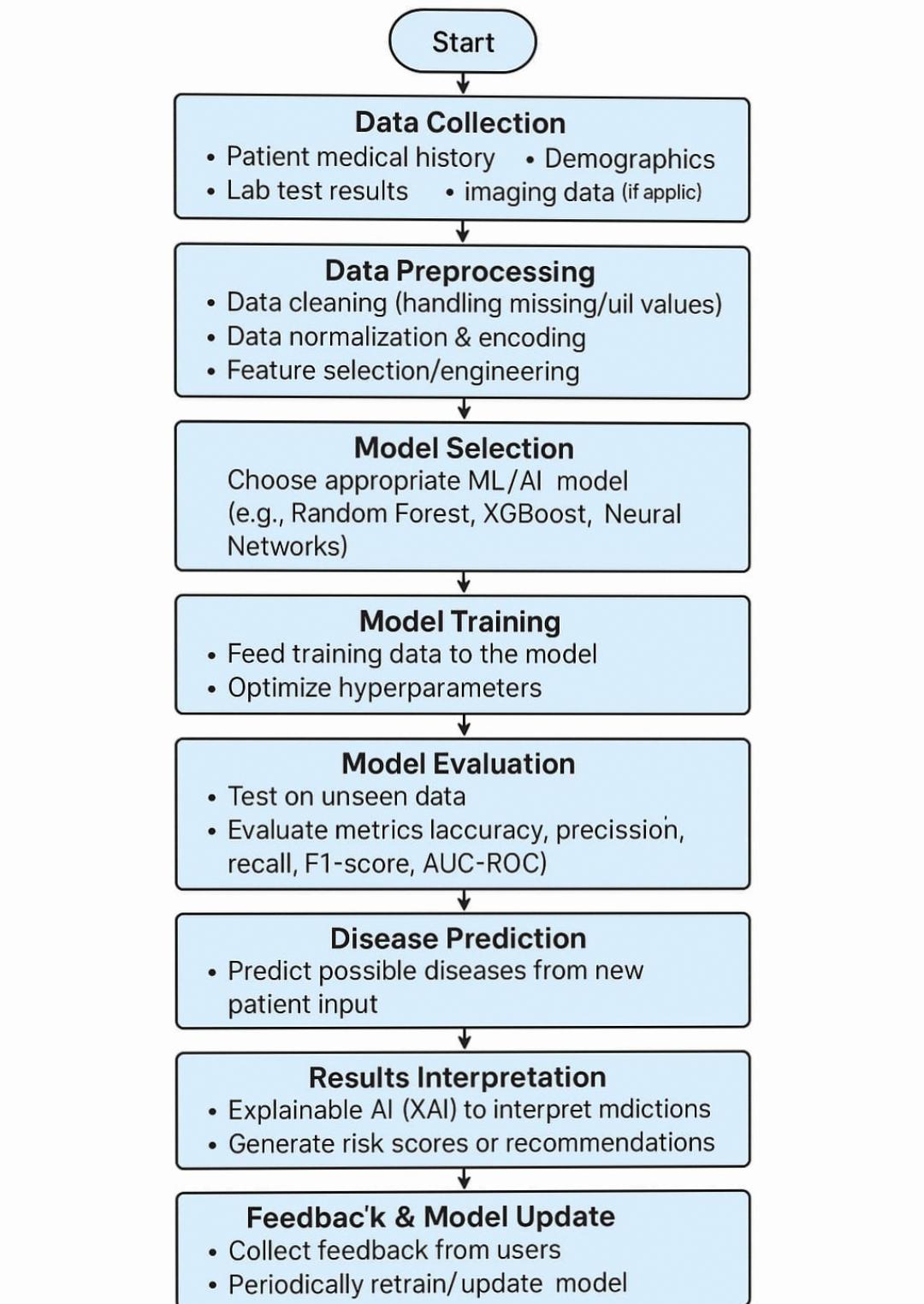
Interpret and visualize relationships among health metrics and outcomes.

Ensure model explainability and adaptability in clinical settings.

Provide an accessible interface (e.g., Gradio) for real-time disease risk prediction by healthcare professionals.

---

3. Flowchart of the Project Workflow



---

4. Data Description

Dataset Name: [Specify dataset, e.g., Heart Disease Dataset]

Source: [e.g., UCI Machine Learning Repository / Kaggle]

Type of Data: Structured, tabular

Records & Features: [e.g., 1000 records, 20+ features]

Target Variable: Disease status (binary: 0 = no disease, 1 = disease)

Attributes Covered: Demographics, clinical test results, lifestyle behaviors, family history.

Data source:

---

5. Data Preprocessing

Handled missing values through imputation or removal.

Converted categorical variables into numeric using one-hot encoding.

Standardized numerical features using StandardScaler.

Detected and addressed outliers using z-scores and visual methods.

Removed duplicate entries and low-variance features.

---

6. Exploratory Data Analysis (EDA)

1.Univariate Analysis: Histograms and boxplots of features like blood pressure, cholesterol.

2.Bivariate Analysis: Correlation matrix and scatter plots to examine feature relationships.

3.Key Insights:

Certain features like age, blood pressure, and cholesterol levels have strong correlations with disease presence.

Lifestyle factors such as smoking or physical activity play a significant role.

---

7. Feature Engineering

Created interaction terms (e.g., BMI × Age).

Introduced binary indicators (e.g., high\_risk = age > 50 and high BP).

Performed dimensionality reduction to remove redundancy.

Label encoded binary features.

---

8. Model Building

Algorithms Used:

Logistic Regression: for baseline and interpretability

Random Forest Classifier: to capture non-linear patterns and feature importance

XGBoost: for boosting performance on imbalanced data

Model Evaluation:

Accuracy, Precision, Recall, F1-Score

ROC-AUC for classification performance

Confusion Matrix for true/false predictions

---

9. Visualization of Results & Insights

Feature importance plots from tree-based models.

ROC curve comparisons across models.

SHAP or LIME plots to explain predictions.

Integrated the best-performing model into a Gradio interface for user interaction.

---

10. Tools and Technologies Used

Language: Python 3

Environment: Jupyter Notebook / Google Colab

Libraries:

pandas, numpy for data handling

matplotlib, seaborn, plotly for visualization

scikit-learn, XGBoost for modeling

Gradio for deployment

---

11. Team Members and Contributions

Sathick.U– Data Collection & Preprocessing

Siranjeevi A – Exploratory Data Analysis & Feature Engineering

Sathish Kumar.S– Model Development & Evaluation

---