**Phase-2**

**Student Name:** Keerthiga.A

**Register Number:** 510823205023

**Institution:** Ganadhipathy tulsis jain engineering college

**Department:** Information and technology

**Date of Submission:** 26/04/25

**Github Repository Link: https://github.com/keertz-04/ai-disease-prediction**

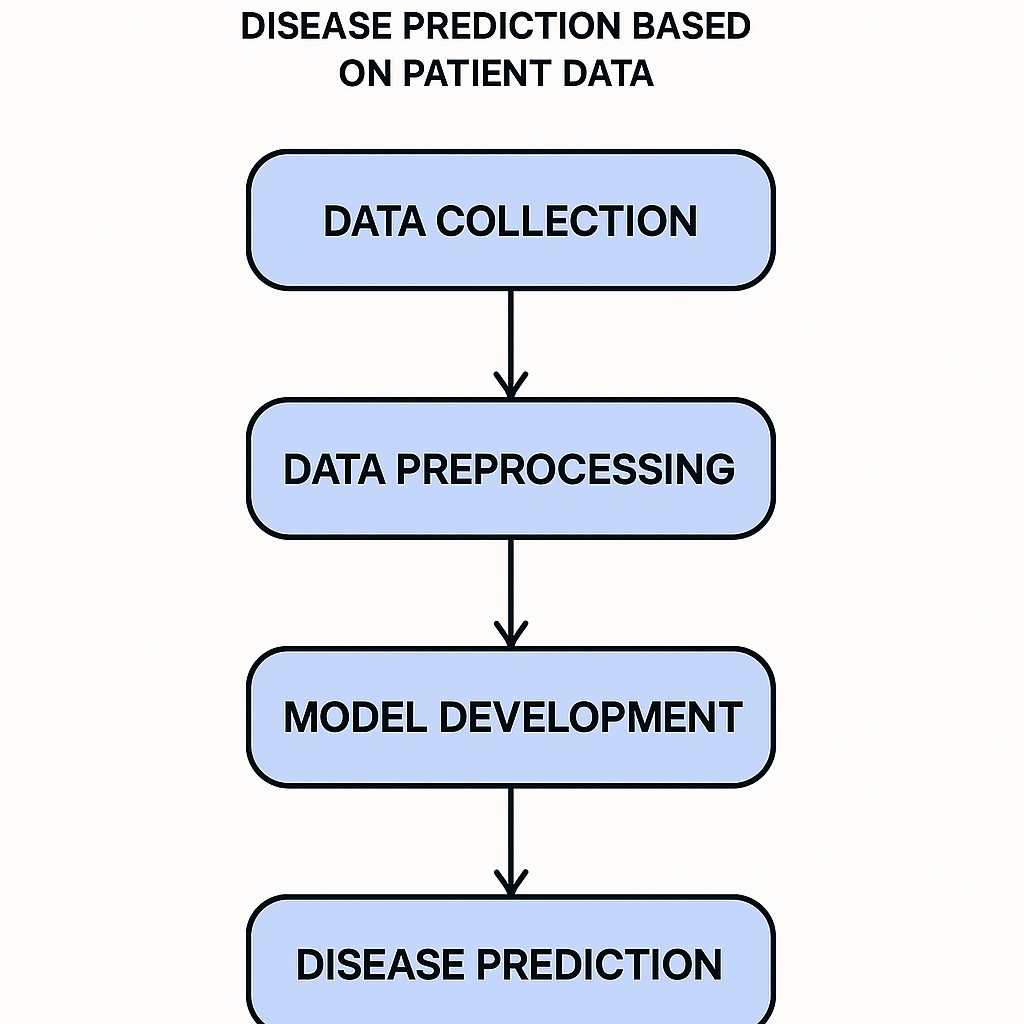
# 1. Problem Statement :

* + This project builds an AI model to predict disease risk (e.g., diabetes, heart disease) using patient data like age, BMI, and lab results, helping doctors make faster, data-driven decisions.

# 2. Project Objectives

* Predict disease risk using patient data with AI models.
* Preprocess data and train multiple classifiers.
* Evaluate performance (accuracy, recall, ROC-AUC).
* Ensure model interpretability (SHAP/LIME).
* (Optional) Build a basic user interface.
* Focus: Accurate, practical binary classification.

**3. Flowchart of the Project Workflow**



# 4. Data Description

* Dataset Name: Diabetes Prediction Dataset
* Source: Kaggle
* Type: Structured (tabular data)
* Records & Features: ~100,000 records, 18 features
* Nature: Static dataset
  + Target Variable: diabetes (0 = No, 1 = Yes)

# 5. Data Preprocessing

* Handled missing values via imputation and removed duplicates.
* Treated outliers using the IQR method.
* Ensured correct data types and consistency.
* Encoded categorical variables (label/one-hot encoding).
* Standardized numerical features for model input.

# 6. Exploratory Data Analysis (EDA)

* Univariate Analysis:Used histograms, boxplots, and countplots to explore feature distributions.
* Bivariate/Multivariate Analysis:Plotted correlation matrix, scatterplots, and pairplots to examine relationships.
* Analyzed feature impact on the target variable.
* Insights Summary:Identified key patterns—e.g., higher BMI and age linked to disease risk.Selected influential features for model training.

# 7. Feature Engineering

* Created features like BMI categories based on EDA insights.
* Extracted date parts (e.g., age) by splitting columns.
* Used binning and ratios to highlight key patterns.
* Applied PCA (optional) for dimensionality reduction.
* Each change was guided by domain knowledge and performance impact

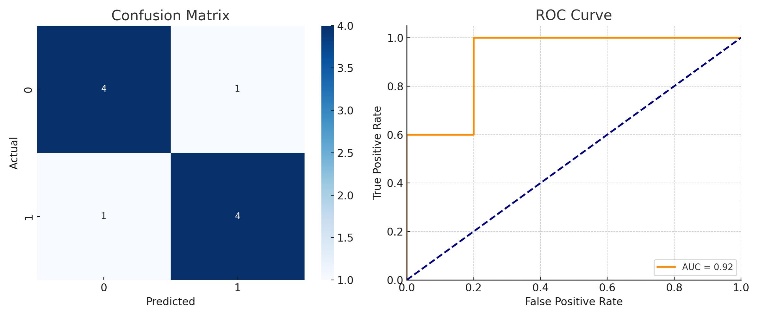
# 8. Model Building

* Implemented multiple models: Logistic Regression and Random Forest.
* Chosen for their effectiveness in binary classification and interpretability.
* Data split into training and testing sets (stratified).
* Trained models and evaluated using accuracy, precision, recall, and F1-score.

|  |  |  |
| --- | --- | --- |
| Metric | Logistic regression | Random Forest |
| Accuracy | 87.21% | 95.56% |
| Precisions | 87% | 96% |
| Recall | 87% | 96% |
| F-1 score | 87% | 96% |

# 9. Visualization of Results & Model Insights

* ROC & Confusion Matrix: Showed high accuracy and AUC.
* Top Features: BMI, age, glucose.
* Comparison: Random Forest outperformed others.
* Insight: Key features match medical risk factors



# 10. Tools and Technologies Used

* Language: Python
* IDE/Notebook: Google Colab
* Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn, XGBoost
* Visualization: Matplotlib, Seaborn, (optional: Plotly/Tableau)

# 11. Team Members and Contributions

 Keerthiga.A – Data collection, Development.

 Sakthivel.B- Model evaluation

 Gokul.V- Visualization, project co-ordination

 Santhiya.K-Documentation