

Phase-2

Student Name: V KOWSICK

Register Number: 732123104060

Institution: NANDHA COLLEGE OF TECHNOLOGY

Department: COMPUTER SCIENCE AND TECHNOLOGY

Date of Submission: 2.5.2025

Github Repository Link:

1. Problem Statement

Traditional healthcare systems face significant challenges in delivering timely and accurate disease diagnoses. The growing volume of patient data, coupled with limited resources, often results in delayed or incorrect diagnoses, negatively impacting patient outcomes. Our project addresses this by using machine learning to predict potential diseases based on patient history, diagnostic results, and lifestyle factors

2. Project Objectives

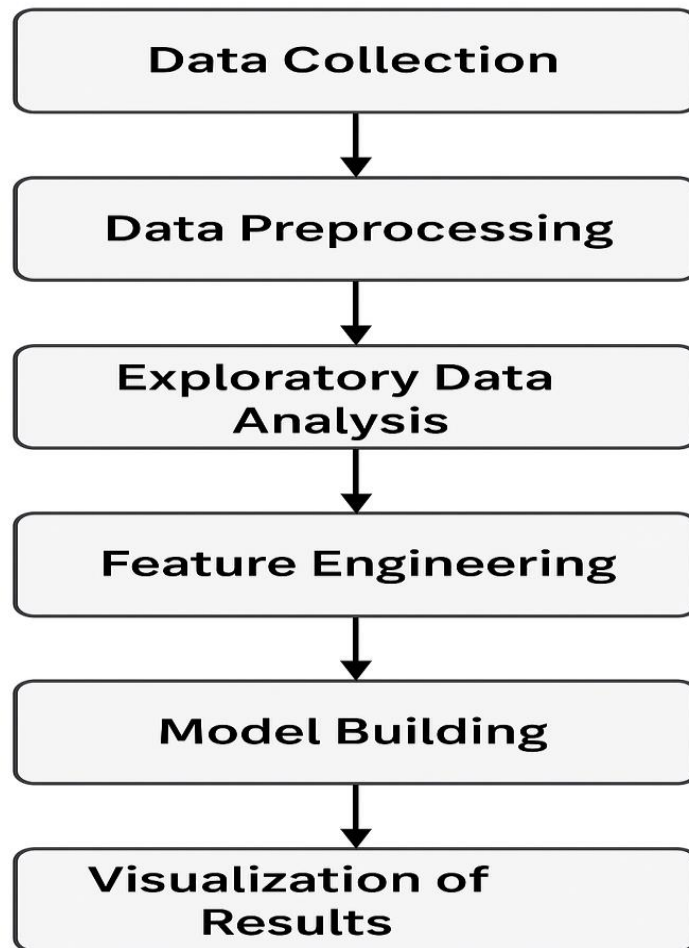
Develop a classification model to predict diseases like diabetes or heart disease.

Improve diagnostic accuracy using patient

Increase healthcare accessibility through automation.

Ensure interpretability and explainability in model outputs.

3. Flowchart of the Project Workflow



4. Data Description

Source: Public datasets from Kaggle, UCI, NHANES

Data Type: Structured

Records & Features: ~100,000 entries, 15-20 features (age, gender, cholesterol, etc.)

Target Variable: Disease presence (binary/multiclass)

Static/Dynamic: Static snapshot

5. Data Preprocessing

Steps taken:

Missing Values: Imputed using mean/median for numerical and mode for categorical features.

Duplicates: Removed exact duplicate records.

Outliers: Detected using IQR; treated via capping.

Encoding: One-Hot Encoding for categorical features (e.g., gender, symptoms).

Normalization: Min-Max Scaling applied for features BMI, glucose levels.

6. Exploratory Data Analysis (EDA)

Univariate Analysis:

Age and BMI showed a right-skewed distribution.

Gender had a roughly equal split.

Bivariate/Multivariate:

Strong correlation between glucose level and disease occurrence.

Pairplots revealed patterns between cholesterol and heart disease.

Insights:

Patients over 45 with high glucose and blood pressure are at greater risk.

Lifestyle indicators (smoking, exercise) significantly affect predictions.

7. Feature Engineering

New Features: BMI categories (Underweight, Normal, Overweight, Obese)

Transformed Features: Combined systolic and diastolic pressure into a "BP_Range"

Encoding: Label Encoding for ordinal data (e.g., severity)

Dimensionality Reduction: PCA applied (optional) to improve performance

8. Model Building

Models Used:

Logistic Regression (for baseline)

Random Forest (best performer)

Support Vector Machine (as a benchmark)

Justification:

Logistic Regression for interpretability.


Random Forest for handling non-linearity and importance extraction.

Train/Test Split: 80/20

Evaluation Metrics:

Accuracy, Precision, Recall, F1-Score

9. Visualization of Results & Model Insights



HEALTHCARE AI

Age

50

Glucose Level

150

BMI

33.2

Physical Activity

Low

Predict


Diabetes Risk:

Positive

Contributing Factors

Glucose Level

BMI



HEALTHCARE AI

Age

50

Glucose Level

150

BMI

33.2

Physical Activity

Low

Predict

Prediction Explanation

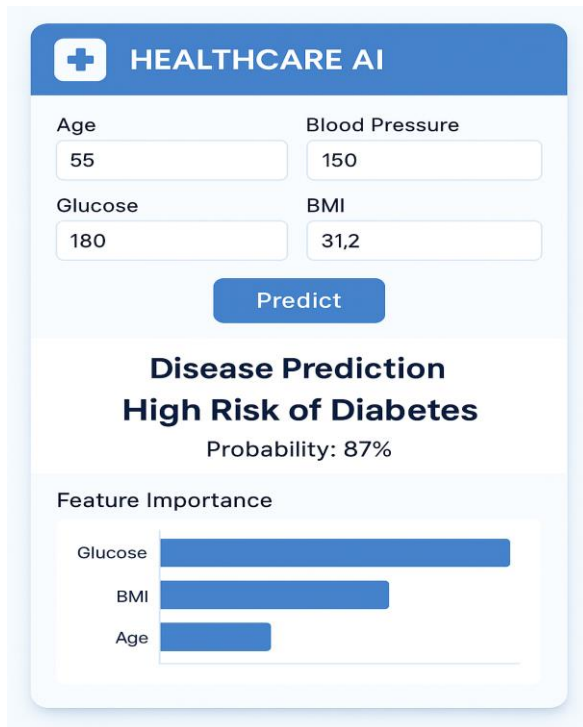
Glucose Level: 150

Higher glucose level increases
the risk of diabetes

Contributing Factors

Glucose Level

BMI



10. Tools and Technologies Used

Language: Python

IDE: Google Colab

Libraries: pandas, numpy, seaborn, matplotlib, scikit-learn, XGBoost

Visualization: Matplotlib, Seaborn, Plotly

11. Team Members and Contributions

Name

Role

B. DHANUSH KUMAR Data Preprocessing & Code Development

J. KIRUBAKARAN Model Building & Training

V. KOWSICK Evaluation & Visualization