### **Transforming healthcare with AI-powered disease prediction based on patient data**

**Student Name : M.k.Tharun**  
**Register Number :** 513523205301  
**Institution :** Annai Mira College of Engineering and Technology  
**Department :** B.Tech Information Technology  
**Date of Submission :** 26/04/2025  
**GitHub Repository :** <https://github.com/North-Abyss/NM-AI-powered-disease-prediction-based-on-patient-data>

### **1. Problem Statement**

* + The project addresses the growing need for early diagnosis in healthcare by leveraging AI to predict diseases based on patient data. Traditional diagnostics are often time-consuming and error-prone. By using machine learning models, this project aims to automate disease prediction, improve diagnostic accuracy, and enable timely interventions.

**Type of Problem :** Classification (predicting diseases like diabetes, heart disease, hypertension)

**Why It Matters :** Early prediction helps reduce mortality, prevent complications, optimize treatment plans, and reduce overall healthcare costs.

### **2. Project Objectives**

* + Build a classification model to predict disease risks based on health data.
  + Identify critical features influencing predictions (e.g., age, glucose level, BMI).
  + Enhance diagnostic speed and accuracy using AI/ML.
  + Develop a user-friendly dashboard or interface for healthcare use.
  + Post-EDA, the goal expanded to include feature importance insights and potential web deployment.

### **3. Flowchart of the Project Workflow**

**Data Collection**

**Data Preprocessing**

**Exploratory Data Analysis (EDA)**

**Feature Engineering**

**Model Building**

**Model Evaluation**

**Visualization & Interpretation**

**Deployment**

### **4. Data Description**

**Dataset Name & Source:**

* + **Heart Disease Dataset (Kaggle/UCI):** [Link for Heart Disease Data](https://github.com/sharmaroshan/Heart-UCI-Dataset/blob/master/heart.csv)
  + **Diabetes Health Indicators Dataset (Kaggle):** [Link for Diabetes Data](https://github.com/Helmy2/Diabetes-Health-Indicators/blob/main/diabetes_binary_health_indicators_BRFSS2015.csv)

**Type of Data:** Structured (CSV format)

**Records & Features:**

* + **Heart Disease:** ~300 records, 13+ features
  + **Diabetes:** ~25,000 records, 17 features

**Static or Dynamic:** Static

**Target Variable:** Presence/absence of disease (binary classification)

### **5. Data Preprocessing**

* **Missing Values:** Handled using mean/mode imputation
* **Duplicates:** Removed based on exact row matching
* **Outliers:** Detected via boxplots and treated with IQR method
* **Data Types:** Ensured consistent numerical/categorical types
* **Encoding:** One-hot encoding for categorical features (e.g., gender, smoking status)
* **Scaling:** Min-Max scaling used for numerical features (e.g., glucose, BMI)

All preprocessing steps are documented using code cells and markdown in Jupyter.

### **6. Exploratory Data Analysis (EDA)**

**Univariate Analysis:**

* Histograms for age, glucose levels
* Countplots for gender, disease status

**Bivariate Analysis:**

* Correlation heatmaps
* Scatterplots (e.g., BMI vs glucose)
* Grouped bar plots showing disease risk by age group

**Insights Summary:**

* High glucose, high BMI, and age > 45 are strong predictors.
* Strong correlation between hypertension and heart disease.

### **7. Feature Engineering**

**New Features Created:**

* Risk Score = Combination of glucose, BMI, and blood pressure
* Age Group = Binned feature (e.g., 20–30, 31–40...)
* **Dimensionality Reduction:** Not applied (features manageable)
* **Justification:** Domain knowledge + correlation insights supported creation of composite features

### **8. Model Building**

**Algorithms Used:**

* Logistic Regression
* Random Forest Classifier
* (Optional: XGBoost for performance comparison)

**Reason for Choice:**

* Logistic Regression: Simple baseline
* Random Forest: Handles non-linearity, robust with medical data
* Data Split: 80% training / 20% testing with stratification

**Metrics:**

* Accuracy, Precision, Recall, F1-Score, AUC-ROC
* Random Forest achieved the best performance (~87% accuracy)

### **9. Visualization of Results & Model Insights**

**Confusion Matrix:** Showed true positives and false negatives

* ROC Curve: AUC = 0.91 for Random Forest
* Feature Importance:
* Top features: Glucose, BMI, Age, Smoking
* Visualized via bar chart

**Interpretation:**

* The model is reliable in distinguishing high-risk patients
* Helps clinicians prioritize screenings and interventions

### **10. Tools and Technologies Used**

* **Programming Language:** Python
* **IDE/Notebook:** Jupyter Notebook, Google Colab
* **Libraries:** pandas, numpy, seaborn, matplotlib, plotly
* **Machine Learning Libraries:** scikit-learn, xgboost, tensorflow (optional for deep learning)
* **Deployment Tools (Optional):** Streamlit, Flask
* **Version Control:** Git and GitHub

### **11. Team Members and Contributions**

| ****Name**** | ****Contribution**** |
| --- | --- |
| **Vinayagaram D.** | Data Collection and Preprocessing |
| **Vishal Raj V.** | Exploratory Data Analysis and Visualization |
| **Yukesh M.** | Model Development (implemented Logistic Regression, Random Forest) |
| **Yuvanesh K.S.** | Model Evaluation and Testing (accuracy checks, confusion matrix, ROC curve analysis) |
| **Tharun M.K.** | Documentation and UI Development (prepared report, created basic demo UI in Streamlit) |