Phase-2 Submission

Student Name: SANJAY S

Register Number:712523205050

Institution: PPG INSTITUTE OF TECHNOLOGY

Department: B TECH INFORMATION TECHNOLOGY

Date of Submission: 14.05.2025

Git hub Repository Link:
https://github.com/SanjayIT27/NM_Sanjay_DS

Transforming health care with AI-Powered disease prediction based on patient data

1. Problem Statement

- *In today's healthcare landscape, early and accurate disease prediction remains a critical challenge*
- *Using traditional diagnostic methods often results in delayed intervention, which can increase treatment costs and reduce patient survival rates.*
- *his project aims to leverage AI and machine learning to predict the likelihood of various diseases based on structured patient data, including*

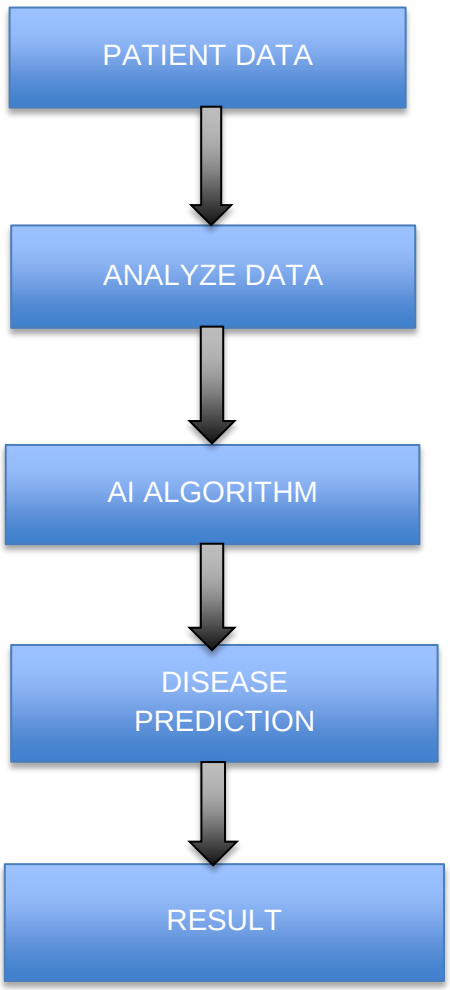
demographics, lifestyle factors, medical history, and test results.

2. Project Objectives

- *To build machine learning models that can accurately predict the presence or risk of diseases using patient data.*
- *To identify the most influential features (risk factors) associated with various health conditions.*

- *To create interpretable and scalable solutions that can be integrated into healthcare systems.*

3. Flowchart of the Project Workflow



4. Data Description

- ***Source**: [Specify dataset origin, e.g., UCI Heart Disease Dataset or Kaggle Patient Data]*
- ***Type**: Structured, tabular data*
- ***Number of Records and Features**: [Insert exact figures]*
- ***Dataset Nature**: Static*
- ***Target Variable**: Disease presence (binary or multiclass, depending on the dataset)*

  [

5. Data Preprocessing

- *Handled missing values using [mean imputation / removal / domain-specific methods].*
- *Removed duplicate records to maintain data integrity.*
- *Identified and treated outliers using IQR and z-score methods.*
- *Encoded categorical variables using One-Hot and Label Encoding.*
- *Scaled numerical features using Standard Scaler for algorithms sensitive to data distribution.*
- *Ensured data type consistency across all columns.*

6. Exploratory Data Analysis (EDA)

- ***Univariate**: Distribution of age, cholesterol, blood pressure using histograms and boxplots.*

- ***Bivariate****: Correlation heatmap revealed strong associations between age, blood pressure, and disease presence.*
- ***Multivariate****: Pair plots showed clusters indicating higher risk profiles.*
- ***Insights****:*
- *Age and cholesterol are strong predictors.*
- *Lifestyle variables like smoking and physical activity have notable influence.*

7. Feature Engineering

- *Derived new features such as BMI category and risk score index.*
- *Extracted time-based features (e.g., years since last check-up).*
- *Performed feature selection using mutual information and recursive feature elimination (RFE).*
- *Applied PCA (optional) for dimensionality reduction in experimentation.*

8. Model Building

- *Derived new features such as BMI category and risk score index.*
- *Extracted time-based features (e.g., years since last check-up).*
- *Performed feature selection using mutual information and recursive feature elimination (RFE).*
- *Applied PCA (optional) for dimensionality reduction in experimentation.*

9. Visualization of Results & Model Insights

- *Confusion Matrix: Displayed TP, FP, FN, TN for each class.*
- *ROC Curve: Compared AUC of models; Random Forest showed higher AUC.*
- *Feature Importance: Random Forest showed age, cholesterol, and smoking history as top predictors.*
- *Interpretation: Model confidently identifies high-risk patients with minimal false negatives.*

10. Tools and Technologies Used

- ***Programming Language**: Python*
- ***IDE**: JupYter Notebook*
- ***Libraries**: pandas, numpy, matplotlib, seaborn, scikit-learn, xgboost*
- ***Visualization**: matplotlib, seaborn, Plotly (optional)*

11. Team Members and Contributions

- *Clearly mention who worked on:*
    - *Rajeshwari S:Data cleaning*
    - *Sanjay S :EDA*
    - *Vignesh :Feature engineering*
    - *TamilVanan K :Model development*
    - *ThamilSelvan P :Documentation and reporting*