**Phase-2**

**Student Name :** Vivega.G

**Register Number :** 613023104121

**Institution :**Vivekanandha college of technology for women

**Department :** B.E-Computer Science and Engineering

**Date of Submission :** 09.05.2025

**Github Repository Link:** https://github.com/vivega-21/nm-project-phase-2

### **1. Problem Statement**

The project addresses the critical challenge of early and accurate disease prediction using AI models based on patient data. In the healthcare sector, timely diagnosis is crucial to improving patient outcomes and reducing the burden on healthcare systems. The goal is to develop an AI-powered predictive model that can identify disease risks from structured patient data (such as demographics, medical history, lab results, etc.). This is a supervised learning problem, primarily classification, where the target is to predict disease presence or risk categories.Solving this problem enhances proactive medical intervention, reduces diagnostic delays, and contributes to more personalized healthcare.

### **2.Project Objectives**

* Build a robust machine learning model for disease prediction using patient data
* Achieve high accuracy, precision, and recall to minimize false negatives and false positives.
* Ensure the model is interpretable so healthcare professionals can trust and understand the predictions.
* Explore feature importance to identify key health indicators.

### **3.Flowchart of the Project Workflo****w**

Include a diagram showing the following steps: Data Collection → Data Preprocessing → Exploratory Data Analysis → Feature Engineering → Model Building → Model Evaluation → Results & Insights]

### **Data Description**

* Dataset: Patient Health Records Dataset (example source: UCI, Kaggle)
* Type of data: Structured data.
* Features: Demographics, medical history, lab results, symptoms.
* Number of records: [Insert number]
* Number of features: [Insert number]
* Dataset type: Static.
* Target variable: Disease diagnosis or risk category.

### **5.Data Preprocessing**

* Handled missing values using [imputation method].
* Removed duplicate records.
* Identified and treated outliers using [method].
* Converted data types for consistency.
* Applied one-hot encoding for categorical variables.
* Standardized numerical features to bring them to a common scale.

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

* Univariate: Analyzed distributions of age, gender, key lab metrics.
* Bivariate: Correlated health indicators with disease outcomes.
* Multivariate: Explored interactions between risk factors.
* Insights: Identified strong correlations between [e.g., high blood pressure] and disease occurrence.

**7.Feature Engineering**

* Created new features such as BMI category and risk scores.
* Extracted date-based features (if applicable).
* Applied feature selection to retain most impactful variables.

### **8.Model Building**

* Models used: Random Forest, XGBoost.
* Reason: Both are robust for classification tasks and handle feature importance well.
* Data split: 80% training, 20% testing.
* Metrics: Evaluated using accuracy, precision, recall, F1-score.

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

* Confusion matrix to evaluate classification performance.
* ROC curve for model comparison.
* Feature importance chart to highlight key predictors.
* Interpretation: The top features influencing predictions include [list key features].

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

* Programming Language: Python
* IDE/Notebook: Jupyter Notebook
* Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn, XGBoost
* Visualization Tools: Plotly, seaborn

**11. Team Members and Contributions**

Yalini.C,Sujitha.M,Vivega.G,Suganya.R,Sujitha.N and responsibilities such as

* Data Cleaning
* EDA
* Feature Engineering
* Model Development
* Reporting