**Phase-3 Submission Template**

**Student Name :**Vivega.G

**Register Number :** 613023104121

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

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

**Date of Submission :** 08**-**05-2025

**Github Repository Link**: https://github.com/vivega-21/nm-project-phase-3-/upload/main

# 1. Problem Statement

The healthcare industry faces significant challenges in early disease detection, leading to delayed diagnoses and increasedmortality. Manual diagnosis is often time-consuming and prone to error. This project aims to build an AI-powered system that predicts diseases using patient data, enabling early intervention. The problem is framed as a classification task where the model predicts the likelihood of a patient developing specific diseases based on health metrics. The healthcare industry faces significant challenges in early disease detection, leading to delayed diagnoses and increased

# 2. Abstract

This project develops an AI-driven system to transform healthcare by predicting diseases from patient data. Leveraging machine learning classification techniques, the system analyzes key patient attributes (e.g., age, medical history, lab results) to forecast disease risks. The workflow involves data preprocessing, exploratory analysis, model building, and deployment. The final model demonstrates strong predictive performance and is deployed via an interactive web app.

# 3. System Requirements

* **Hardware:** Minimum 8 GB RAM, i5 Processor (or equivalent) for model training.
* **Software:** Python 3.9+, libraries including pandas, scikit-learn, matplotlib, seaborn, TensorFlow/Keras, Streamlit or Gradio for deployment.
* **IDE:** Jupyter Notebook or Google Colab.

**4. Objectives**

* Predict the likelihood of various diseases based on patient health data.
* Enable faster, more accurate diagnosis support for healthcare professionals.
* Provide actionable insights to patients for early prevention.
* Demonstrate the business impact by reducing diagnostic errors and improving patient outcomes.

**5. Flowchart of Project Workflow**

Data Collection → Preprocessing → EDA → Feature Engineering → Modeling → Evaluation → Deployment

**6. Dataset Description**

* **Source:** Public dataset (e.g., Kaggle’s Disease Prediction or UCI Machine Learning Repository).
* **Type:** Public, anonymized patient data.
* **Size & Structure:** ~50,000 rows, ~15 columns (patient demographics, medical history, test results).

# 7. Data Preprocessing

* **Missing Values:** Imputed using mean/mode for numerical and categorical fields.
* **Duplicates:** Removed ~2% duplicate records.
* **Outliers:** Detected via boxplots and treated using capping methods.
* **Feature Encoding:** Categorical variables encoded using One-Hot Encoding.
* **Scaling:** Min-Max scaling applied to continuous variables.

# 8. Exploratory Data Analysis (EDA)

* **Visuals:** Histograms of age distribution, heatmap of feature correlations, boxplots for lab results vs. disease status.
* **Insights:**
* Older age groups show higher disease risk.
* Strong correlation between cholesterol levels and heart disease.
* Imbalanced dataset—disease prevalence ~15%.

# 9. Feature Engineering

* **New Features:** BMI calculated from height/weight; risk score combining age and lifestyle factors.
* **Feature Selection:** Used correlation matrix and feature importance ranking (e.g., Random Forest).
* **Transformations:** Applied log transformation to skewed lab results.
* **Rationale:** These steps improved model accuracy by ~5% by enhancing feature relevance.

# 10. Model Building

* **Models Tried:**
* Logistic Regression (baseline)
* Random Forest Classifier
* XGBoost Classifier (best performer)
* **Why Chosen:** Random Forest and XGBoost handle feature interactions well and are robust to overfitting

**11. Model Evaluation**

* **Metrics:**
* Accuracy: 89%
* Precision: 85%
* Recall: 82%
* F1-score: 83%
* **Visuals:**
* Confusion Matrix
* ROC Curve (AUC: 0.91)
* **Error Analysis:** Most errors stemmed from borderline cases with mild symptoms.

# 12. Deployment

* **Platform:** Streamlit Cloud.
* **Public Link:** [Insert your deployed app link]
* **UI:** Simple form where users enter patient details and get instant disease risk prediction.
* **Sample Output:** “Patient is at high risk of diabetes (85% probability).”

**13. Source code**

The full source code is available at: https://github.com/vivega-21/nm-project-phase-3-/upload/main

# 14. Future scope

* **Expand Disease Coverage:** Incorporate more diseases such as cancers and rare genetic disorders.
* **Real-Time Data Integration:** Enable real-time EHR data feeds for continuous monitoring.
* **Mobile App:** Develop a mobile version for easy access by patients and healthcare providers.

# 13. Team Members and Roles

* Yalini.C: Data preprocessing, model building,evaluation.
* Sujitha.M: EDA
* Suganya.R:Deployment setup, feature engineering
* Sujitha.N: Documentation, testing
* Vivega.G: UI design.