**Phase-1 Submission**

**Student Name:** Natarajan R

**Register Number:** 410723104051

**Institution:** Dhanalakshmi College Of Engineering **Department:** Computer Science And Engineering **Date of Submission:** 27-04-2025



# Problem Statement

Healthcare systems worldwide face challenges in early detection and diagnosis of diseases, which often leads to delayed treatment and increased medical costs.

Traditional diagnostic methods can be time-consuming and resource-intensive. With the growing availability of patient data, there is a strong need for intelligent systems that can assist medical professionals by accurately predicting the onset of diseases at early stages. An AI - powered disease prediction model can help improve patient outcomes, optimize healthcare resources, and enable personalized medicine.

# Objectives of the Project

* + Develop an AI-based system capable of predicting the likelihood of diseases based on patient data.
  + Identify key medical indicators and risk factors contributing to various illnesses.
  + Assist health care professionals in early diagnosis and preventive care strategies.
  + Create visual and interpretable outputs to support decision-making in clinical settings.

# Scope of the Project

**Features to Build or Analyze:**

* Demographic data (age, gender, lifestyle habits)
* Clinical features (blood pressure, glucose levels, BMI, etc.)
* Historical medical data (past illnesses, genetic predispositions)
* Disease-specific markers (e.g., cholesterol levels for heart disease)

# Limitations/Constraints:

* Use only publicly available or synthetic datasets due to privacy regulations.
* Focus on a specific set of diseases (e.g., heart disease, diabetes ,liver disease) for model training.
* Predictions are intended for research and proof-of-concept purposes only — not for direct clinical use.
* Deployment limited to local or secure sandbox environments.

# Data Sources

Datasets:

* Heart Disease Dataset(UCI Machine Learning Repository)

Link: <https://archive.ics.uci.edu/dataset/45/heart+disease>

* Diabetes Dataset(UCI Machine Learning Repository)

Link: <https://archive.ics.uci.edu/dataset/34/diabetes>

* Liver Patient Dataset(UCI)

Link: https://archive.ics.uci.edu/dataset/225/ilpd+indian+liver+patient+dataset

Source: UCI

Type: Public datasets

Format: CSV(static data, downloaded once)

# High-Level Methodology

Data Collection:

* Datasets will be downloaded from UCI Machine Learning Repository.
* Additional synthetic data may be generated using libraries like Faker for expanded testing.

Data Cleaning:

* Check and handle missing or null values
* Normalize numerical features and encode categorical variables
* Remove duplicate records and resolve inconsistencies in formatting

Exploratory Data Analysis (EDA):

* Use histograms, boxplots, heatmaps, and correlation matrices to understand data distributions and relationships

* Perform univariate and multivariate analysis to uncover key patterns

Feature Engineering:

* Create new features such as risks core storage groups
* Apply transformations(log, scaling)to improve model input quality

Model Building:

* Train and compare classification models(LogisticRegression ,Random Forest, XG Boost, SVM, Neural Networks)
* Perform hyper parameter tuning using Grid Search CV or Randomized Search CV

Model Evaluation:

* Use metrics like accuracy, precision, recall, F1-score, ROC-AUC
* Apply K-fold cross-validation for reliable performance estimates

Visualization & Interpretation:

* Generate visual reports with confusion matrices, ROC curves, and SHAP values
* Use dashboards(e.g.,via Streamlit) to present interactive insights to users

Deployment:

* Develop a simple web application using Streamlit or Flask for demonstration
* Provide options for uploading patient data and viewing predictions with explanations

# Tools and Technologies

Programming Language:

* Python

Notebook/IDE:

* Google Colab / Jupyter Notebook

Libraries:

* pandas, numpy, matplotlib, seaborn – Data processing and visualization
* scikit-learn, xgboost, tensorflow or keras –Machine learning models
* SHAP, lime – Model interpretation
* streamlit, flask–Deployment(optional)

Optional Tools for Deployment:

* Streamlit for interactive app
* Docker for containerized deployment(if needed)
* GitHub for version control and collaboration

1. **Team Members and Roles**

| **Name** | **Role** | **Responsibilities** |
| --- | --- | --- |
| Natarajan R | Project Lead | Data preprocessing, model development, documentation |
| Naveenraj R | Data Analyst | EDA, visualization,and reporting |
| Tarun V | Machine Learning Engineer | Modeltuning, evaluation, and testing |
| Yokesh K  Suraj SK A | Web Developer  Organizer | Streamlit/Flask app development, deployment  Python developer |