# Phase-2 Submission

**Student Name:** Natarajan R

**Register Number:** 410723104051

**Institution:** Dhanalakshmi College Of  Engineering

**Department:** B.E(Computer Science And Engineering)

**Date of Submission:** 07-05-2025

**Github Repository Link**:
https://github.com/Nattu3001/data-science.git

---

**Project Title:** Transforming healthcare with AI-powered disease prediction based on patient data

## 1. Problem Statement

- In today's healthcare landscape, early and accurate disease detection is essential for timely intervention and reducing mortality. Many patients suffer from undiagnosed or late-diagnosed conditions such as diabetes, heart disease, and liver disease, leading to costly and intensive treatments. This project aims to build AI-powered predictive models using patient health data to automatically assess the risk of these diseases.

- The problem is formulated as a classification task, where the target variable is a binary label indicating the presence or absence of a specific disease. By leveraging machine learning, we aim to transform static patient records into actionable clinical insights, helping healthcare providers make informed decisions more efficiently.

## 2. Project Objectives

- The primary goal is to develop AI-based predictive models for diabetes, heart disease, and liver disease using structured patient datasets. This supports the transformation of healthcare into a predictive and preventive model.

- Specific objectives:

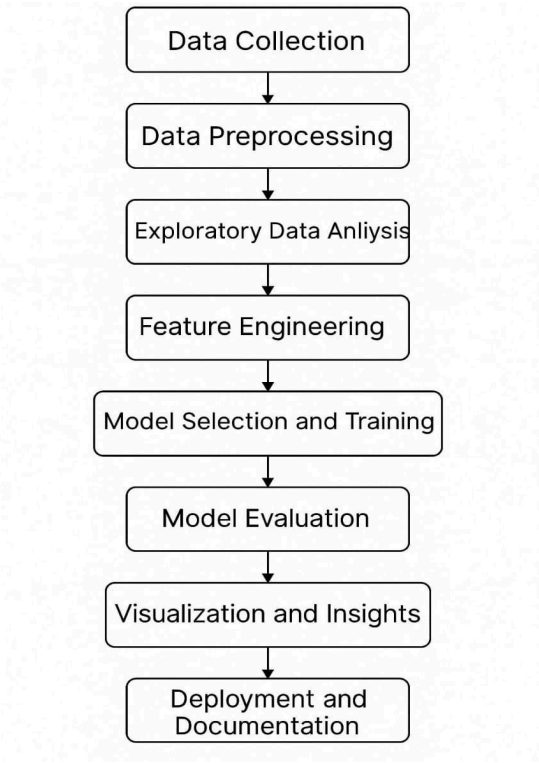
- To preprocess and analyze three datasets covering key patient health metrics.

- To build classification models using machine learning techniques like Logistic Regression, Random Forest, and XGBoost.

- To optimize models for accuracy, precision, recall, and F1-score.

- To provide interpretable insights into which factors contribute most to disease risk

## 3. Flowchart of the Project Workflow




## 4. Data Description

- Datasets Used:
- Diabetes Dataset – Pima Indians Diabetes (UCI)

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

- Heart Disease Dataset – UCI Cleveland Heart Disease

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

- Liver Disease Dataset – Indian Liver Patient Dataset (ILPD, UCI)

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

Type: Structured tabular data

- Records & Features:

Diabetes: 768 records, 9 features

Heart: 303 records, 14 features

Liver: 583 records, 10 features

- Target Variable:

Diabetes: Outcome

Heart: target

Liver: Dataset (1 for liver patient, 2 for healthy)

## 5. Data Preprocessing

- Handled missing values using imputation (e.g., mean/mode for liver dataset).
- Removed duplicates and verified data types.
- Applied outlier detection using IQR and visualizations.

- Label encoded categorical features (e.g., Gender in liver dataset).

- Normalized numerical features using MinMaxScaler to prepare for ML models.

## 6. Exploratory Data Analysis (EDA)

- Plotted histograms and boxplots for understanding distribution.

- Used pairplots and correlation heatmaps to analyze relationships.

- Found that features like glucose (diabetes), cholesterol (heart), and total bilirubin (liver) are strongly associated with disease outcomes.

- Noted class imbalances, particularly in the liver dataset.

## 7. Feature Engineering

- Created new ratios and bins (e.g., cholesterol-to-HDL ratio, BMI bins).

- Removed redundant features identified via correlation.

- Considered PCA for dimensionality reduction (optional, based on variance explained).

- Selected features with high relevance based on domain knowledge and model importance.

## 8. Model Building

Models used:

- Logistic Regression – Baseline linear model.
- Random Forest – Non-linear ensemble model.
- XGBoost – Gradient boosting model for accuracy.
- Data Split: 80% training, 20% testing
- Evaluation Metrics: Accuracy, Precision, Recall, F1-Score

Best models per disease:

- Diabetes: XGBoost
- Heart: Random Forest
- Liver: Logistic Regression (with balancing techniques)

## 9. Visualization of Results & Model Insights

- Plotted confusion matrices and ROC curves.
- Used feature importance graphs to interpret decision drivers.
- Visualized performance comparison using bar charts.
- Insights: Glucose and BMI for diabetes, chest pain type for heart disease, total bilirubin for liver.

## 10. Tools and Technologies Used

- Programming Language: Python
- IDE/Notebook: Google Colab, Jupyter Notebook
- Libraries: pandas, numpy, seaborn, matplotlib, scikit-learn, xgboost
- Visualization Tools: matplotlib, seaborn, plotly
- Version Control: GitHub

## 11. Team Members and Contributions

| Name | Contribution |
|---|---|
| Natarajan R | Data collection and cleaning |
| Naveen Raj R | Exploratory Data Analysis (EDA) |
| Tarun V | Feature engineering and model development |
| Yokesh K | Model evaluation and visualization |
| Suraj SK A | Documentation, report preparation, and GitHub |