Phase-2 Submission Template

Student Name: M. Mursitha

Register Number: 620123106070

Institution: AVS Engineering College

Department: ECE

Date of Submission: 10 – 05 - 2025

Github Repository Link: To be updated

1. Problem Statement

The rising complexity in diagnosing diseases due to massive and diverse patient data poses a significant challenge in healthcare. Early and accurate disease prediction can greatly enhance treatment success, patient care, and resource allocation. Using AI and machine learning techniques, we aim to build a predictive model that can analyze patient data and forecast potential health risks or diseases.

Type of Problem: Classification

Importance: Enhances healthcare efficiency, supports early diagnosis, reduces human error, and enables personalized treatment plans

2. Project Objectives

- Develop a machine learning model to predict diseases based on patient health data.
- Improve prediction accuracy using EDA and feature engineering.
- Prioritize model interpretability for real-world medical usage.
- Validate performance through appropriate classification metrics.

3. Flowchart of the Project Workflow

(Insert a flowchart diagram here: e.g., Data Collection -> Preprocessing ->

EDA -> Feature Engineering -> Model Building -> Evaluation -> Deployment)

4. Data Description

- Dataset Name: Disease Prediction Dataset
- Source: Kaggle
- Data Type: Structured
- Records: ~5000 patient records
- Features: ~20 features (e.g., age, gender, symptoms, medical history)
- Target Variable: Disease classification
- Nature: Static dataset

5. Data Preprocessing

- Missing values handled via median/mode imputation.
- Removed duplicate entries based on patient ID and symptom combination.
- Outliers treated using IQR for continuous variables.
- Label encoding for categorical variables like gender and symptoms.
- Features normalized using MinMaxScaler for model stability.

6. Exploratory Data Analysis (EDA)

- Univariate: Age and symptom distribution visualized with histograms and boxplots.
- Bivariate: Heatmaps revealed strong correlation between symptoms and target disease.
- Insights: Symptoms like chest pain, fatigue, and high sugar levels were strong indicators.

7. Feature Engineering

- Created symptom count feature.
- Transformed date of admission into weekday/weekend feature.
- Applied PCA to reduce dimensionality while preserving 95% variance.

8. Model Building

- Algorithms used: Random Forest, XGBoost
- Data split: 80% train, 20% test (stratified)
- Evaluation Metrics: Accuracy, Precision, Recall, F1-score
- Performance:
    - Random Forest: 91% Accuracy
    - XGBoost: 93% Accuracy

9. Visualization of Results & Model Insights

- Confusion matrix showed high precision for critical diseases.
- ROC curve: AUC > 0.9 for both models.
- Feature importance highlighted top symptoms like chest pain, blood pressure, sugar level.

10. Tools and Technologies Used

- Language: Python
- Notebook: Google Colab
- Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn, XGBoost
- Visualization Tools: matplotlib, seaborn, Plotly

11. Team Members and Contributions

- S.Thirulochine: Data Cleaning, Model Building
- V.Sandhiya: EDA, Feature Engineering
- M.Murshitha: Documentation, Visualization