**Phase-1 Submission Template**

**Student Name:** [Abinaya S]

**Register Number:** [421623104005]

**Institution:[**Mailam Engineering College]

**Department:** [Computer Science and Engineering]

**Date of Submission:**[Insert Date]

**1.Problem Statement**

Despite advances in medical science, healthcare systems worldwide still struggle with timely and accurate disease diagnosis. Traditional diagnostic methods are often time-consuming, resource-intensive, and susceptible to human error, leading to delayed treatment and poorer patient outcomes.

**2.Objectives of the project:**

* The objective of this project is to design and develop an AI-powered system capable of predicting diseases based on patient data.
* such as medical records, lab results, demographic information, and clinical histories.

1. **Scope of the Project:**

The project will focus on developing a machine learning or deep learning model that can predict the likelihood of specific diseases (e.g., diabetes, heart disease, cancer).

**4**.**Data Source:**

* The project will utilize publicly available healthcare datasets containing patient records, clinical features, and diagnosis labels.
* Kaggle healthcare competitions and datasets (e.g., Disease Prediction Dataset, Patient Health Records).

**5.High-Level Methodology**:

**1. Problem Understanding and Requirement Analysis:**

Clearly define the diseases to predict.

Understand business goals, healthcare relevance, and project expectations.

**2. Data Collection:**

Acquire structured patient datasets from sources like Kaggle, UCI, or PhysioNet.

Ensure data is ethically sourced and complies with privacy laws.

**3. Data Preprocessing:**

Clean the dataset (handle missing values, outliers, duplicates).

Perform feature selection, encoding, scaling, and normalization.

Split data into training, validation, and testing sets.

**4. Exploratory Data Analysis (EDA):**

Analyze data patterns, correlations, and key features.

Visualize distributions and relationships among variables.

**5.Model Builiding:**

Select appropriate machine learning algorithms (Logistic Regression, Random Forest, XGBoost, etc.).Perform hyperparameter tuning using techniques like Grid Search or Random Search.

**6. Model Evaluation:**

Test the models using validation and testing datasets.

Evaluate performance using metrics such as Accuracy, Precision, Recall, F1-Score, and AUC-ROC.

**7. Model Deployment (Optional):**

Deploy the best model using a simple web framework like Streamlit or Flask for end-user interaction.

Create a basic dashboard for disease prediction output.

**8. Result Interpretation and Documentation:**

Analyze model outputs and interpret important features influencing disease prediction.

Document the process, results, challenges faced, and conclusion.

1. **Tools and Technologies:**

[List the tools, programming languages, and libraries you plan to use in your project. Include the following details]

* **Programming Language:** Python (main language for model development).
* **Data Processing and Analysis:** Pandas (data manipulation), NumPy (numerical computations) and Scikit-learn (preprocessing tools).
* **Exploratory Data Analysis (EDA) and Visualization:** Matplotlib (data visualization), Seaborn (advanced statistical plots) and Plotly (interactive charts, optional).
* **Machine Learning and Deep Learning Frameworks:** Scikit-learn (classification models like Logistic Regression, Random Forest, SVM), XGBoost or LightGBM (boosting algorithms for better performance), OrFlow or Keras (for deep learning models, optional if going deeper).
* **Model Evaluation:** Scikit-learn metrics module (confusion matrix, ROC curve, precision-recall).
* **Deployment (Optional):** Streamlit or Flask (for creating a simple web app to showcase your model).
* **Environment and Version Control:** Jupyter Notebook or Google Colab (for code development and experiments) and Git and GitHub (for version control and collaboration).
* **Dataset Sources:** Kaggle, UCI Machine Learning Repository, PhysioNet.