Phase-1 Submission

Student Name: Agnes Selestina S

Register Number: 410723104005

Institution: Dhanalakshmi college of engineering

Department: Computer Science and Engineering

Date of Submission: 30-04-2025

1.Problem Statement

AI – POWERED DISEASE PREDICTION

Transforming healthcare with AI-powered disease prediction based on patient data.

2.Objectives of the Project

Project Objective:

To develop an AI-powered system that predicts the risk of diseases using patient data, enabling early diagnosis, personalized care, and better clinical outcomes.

Key Outcomes:

- Train machine learning models to predict diseases based on patient demographics, medical history, and lifestyle factors.
- Classify patients into risk categories for targeted intervention.
- Identify key predictors influencing disease risk.
- Evaluate model accuracy using metrics like precision, recall, and AUC.
- Provide a decision-support tool for clinicians with risk insights and recommendations.

3.Scope of the Project

Features to Analyse/Build:

1. **Demographics:**
   - Age, gender, ethnicity, location
2. **Medical History:**
   - Past diagnoses, family history of diseases, previous hospitalizations
3. **Vital Signs & Clinical Metrics:**
   - Blood pressure, heart rate, BMI, cholesterol, glucose levels
4. **Lab Test Results:**
   - Blood work, urinalysis, liver/kidney function tests
5. **Lifestyle Factors:**
   - Smoking status, alcohol consumption, physical activity, diet
6. **Medication & Treatment History:**
   - Current/past prescriptions, treatment adherence
7. **Genetic or Genomic Data (if available)**
8. **Time-Series Data (for longitudinal analysis):**
   - Trends in vitals or lab values over time

Limitations & Constraints:

1. **Data Constraints:**
   - Use of publicly available or anonymized datasets (e.g., MIMIC-III, UCI Health datasets)
   - Missing or imbalanced data may affect model performance
2. **Model Constraints:**
   - Limited to interpretable models if required by clinical partners (e.g., logistic regression, decision trees)
   - Avoid black-box models unless explainability techniques (e.g., SHAP, LIME) are applied
3. **Deployment Constraints:**
   - Prototype may not be deployed in real clinical settings without regulatory approval (e.g., FDA clearance)
   - Compliance with data privacy laws (e.g., HIPAA, GDPR)
4. **Tool Constraints:**
   - Development limited to Python-based frameworks (e.g., scikit-learn, TensorFlow, PyTorch)
   - Visualization using tools like Streamlit, Dash, or Power BI

4.Data Sources

Dataset Description:

- **Dataset Name**: Healthcare Dataset
- **Source** : Kaggle (dataset)
- **Accessibility**: Public
- **Type** : Static

5.High-Level Methodology

- ***Data Collection*** – The dataset will be obtained through direct download from publicly available source **Kaggle** for disease diagnosis.

- ***Data Cleaning*** – Identify potential issues such as **missing values, duplicates,** or **inconsistent formats.**

- ***Exploratory Data Analysis (EDA)*** –

  Predictive Modeling & Risk Analysis •

  Techniques:

  1. **Logistic regression, decision trees, or random forests** – for predicting disease risk.
  2. **Survival analysis (e.g., Kaplan-Meier curves)** – for analysing time to event (e.g., time until readmission).

  • **Visualizations:**

  1. **ROC curves / AUC plots** – to evaluate model performance.
  2. **Survival curves** – to compare patient outcomes by treatment groups.

- ***Model Building:***

Supervised Learning

1. **Logistic Regression** – Simple, interpretable, great for binary outcomes.
2. **Random Forest** – Handles non-linear data, robust to noise.
3. **XGBoost/LightGBM** – High accuracy, handles complex patterns well.
4. **SVM** – Good for high-dimensional classification.
5. **Neural Networks** – Flexible, good for large and complex datasets.

Unsupervised Learning

6. **K-Means** – Fast and effective for patient clustering.
7. **Hierarchical Clustering** – Useful for exploring group hierarchies.

Deep Learning (for Images/Text)

CNNs – Best for medical image analysis.

Transformers (e.g., BERT) – Exellent for clinical text mining.

- ***Model Evaluation***

Metrics

1. **Classification:** Accuracy, Precision, Recall, F1 Score, ROC-AUC
2. **Regression:** MAE, RMSE, R^2
3. **Survival:** C-index, Log-rank Test
4. **Clustering:** Silhouette Score, Clinical relevance

Validation Strategies

5. **Train-Test Split** – Simple, quick check
6. **k-Fold CV / Stratified k-Fold** – Robust, keeps class balance
7. **Time Series Split** – For time-dependent data
8. **Bootstrapping** – Good for uncertainty estimation

- ***Visualization & Interpretation***

Visualization & Interpretation

1. **Charts:** Line, bar, scatter, boxplots, heatmaps
2. **Dashboards:** Interactive summaries (e.g., Power BI, Tableau)
3. **Model Explainers:** SHAP, LIME, feature importance
4. **Reports:** Clear visuals + insights for stakeholders

6.Tools and Technologies

- **Programming Language** – The main language we use is Python.
- **Notebook/IDE** –The platform we use is Google Colob.
- **Libraries** – The libraries we use is pandas, NumPy, seaborn, matplotlib.

7.Team Members and Roles

S.No	NAME	ROLE
1	Agnes Selestina S	Data Collection, Data Cleaning
2	Christina Ryka S	Visualization & Interpretation
3	Jeevikasri R	Exploratory Data Analysis (EDA), Feature Engineering
4	Keerthana R	Model Building, Model Evaluation