# Student Name: E.Aakash

# Register Number: 620523243001

# Institution: Cms college of engineering

# Department: Artificial intelligence and data science

# Date of Submission:

# Github Repository Link:

Problem Statement

The healthcare industry faces challenges in early and accurate diagnosis of diseases, especially in resource-constrained settings. Delayed or incorrect diagnoses can lead to severe health consequences. There is a need for a reliable, scalable, and intelligent system that can assist healthcare professionals by predicting diseases based on patient data.

# Abstract

This project leverages Artificial Intelligence (AI) to develop a disease prediction system using patient data. By analyzing health parameters such as symptoms, lab results, and demographics, the AI model predicts potential diseases, enabling faster diagnosis and better treatment. The system includes data preprocessing, feature engineering, machine learning model building, and deployment for real-world usage.

# System Requirements

* Hardware:
  + CPU: Intel i5 or higher
  + RAM: 8 GB or more
  + Storage: 100 GB minimum
* Software:
  + OS: Windows/Linux/Mac
  + Python 3.x
  + Libraries: Pandas, NumPy, Scikit-learn, TensorFlow/PyTorch, Flask/Streamlit
  + Jupyter Notebook or any IDE (e.g., VS Code)

# Objective

To develop an AI-based system that predicts diseases from patient data with high accuracy, aiming to support doctors in early diagnosis and reduce human error.

# Flowchart of the Project Workflow

1. Input Patient Data
2. Data Preprocessing
3. Feature Engineering
4. Train/Test Split
5. Model Training
6. Model Evaluation
7. Prediction Output
8. Deployment for Use

# Dataset Description

* + Data Source: [e.g., Kaggle, UCI Machine Learning Repository, Hospital Records]
  + Features:
    - Age, Gender
    - Symptoms
    - Test Results (e.g., BP, Sugar level)
    - Diagnosis (label)
  + Size: ~10,000 records, 25 features

# Data Preprocessing

* + Handling missing values
  + Encoding categorical variables
  + Normalizing numerical features
  + Removing duplicates and outliers

# Exploratory Data Analysis (EDA)

* + Visualization of distributions (histograms, boxplots)
  + Correlation matrix
  + Disease frequency chart
  + Feature importance analysis

# Feature Engineering

* + Creation of derived features
  + Selection of top features using statistical techniques
  + Dimensionality reduction (e.g., PCA)

# Model Building

* + Algorithms used:
    - Logistic Regression
    - Random Forest
    - XGBoost
    - Neural Networks
  + Hyperparameter tuning using GridSearchCV

# Model Evaluation

* + Metrics: Accuracy, Precision, Recall, F1 Score, ROC-AUC
  + Confusion matrix analysis
  + Cross-validation results

# Deployment

* + Web app using Flask or Streamlit
  + Model API using Flask REST API
  + Hosting: Heroku, AWS, or local server

# Source Code

Organized into modules:

* + data\_preprocessing.py
  + model\_training.py
  + predict.py
  + app.py (web interface)
  + requirements.txt

# Future Scope

* + Integration with Electronic Health Records (EHR)
  + Support for image-based diagnostics (e.g., X-rays)
  + Multilingual interface
  + Continuous learning from new data

# Team Members and Roles

* + Team Lead: Oversees development, final review
  + Data Scientist: Handles model development
  + Data Engineer: Manages data preprocessing
  + Frontend Developer: Builds web UI
  + Backend Developer: API integration and deployment