Student Name : Abinesh Godwin.D

Register Number : 510623104002

Institution : C.ABDUL HAKEEM COLLEGE OF ENGINEERING AND TECHNOLOGY

Department : COMPUTER SCIENCE AND ENGINEERING

Date of Submission : 09-05-2025

Github Repository Link:

https://github.com/Abineshgodwin/Transforming-healthcare-with-AI-powered-disease-prediction-based-on-patient-data.

Project Report

Project Title:

Transforming Healthcare with AI-Powered Disease Prediction Based on Patient Data

Problem Statement

Accurate and timely disease diagnosis remains a critical challenge in healthcare, especially in under-resourced areas. Traditional diagnostic methods are often manual, time-consuming, and prone to human error. The need for an automated, reliable, and intelligent system to predict diseases based on patient data is essential for improving early diagnosis and enabling proactive treatment.

Abstraction

This project proposes an AI-driven system capable of predicting the likelihood of various diseases based on patient data such as demographics, medical history, and clinical symptoms. By leveraging machine learning algorithms and healthcare datasets, the system learns patterns associated with different conditions and provides predictions with high accuracy. This solution supports medical practitioners by acting as a decision-support tool to enhance diagnostic confidence and healthcare outcomes.

System Requirements

Hardware Requirements:

- Processor: Intel i5 or higher
- RAM: 8 GB minimum
- Storage: 100 GB minimum
- GPU (optional for deep learning): NVIDIA GTX 1050 or above

Software Requirements:

- Python 3.x
- Jupyter Notebook / VS Code

- Libraries: pandas, NumPy, scikit-learn, matplotlib, seaborn, Flask/Streamlit
- Operating System: Windows/Linux/MacOS

Objectives

To build a predictive model using machine learning that identifies diseases based on patient health data.

To preprocess and analyze medical datasets for pattern discovery.

To design a user-friendly interface for doctors and patients.

To deploy the model using a web framework for real-time prediction.

To evaluate the model's performance using standard metrics like accuracy, precision, and recall.

Flowchart of the Project Workflow

Data Collection → Data Preprocessing → EDA → Feature Engineering → Model Training → Model Evaluation → Deployment

Dataset Description

The dataset contains anonymized patient data with attributes such as:

- Age, Gender
- Symptoms (e.g., fever, cough, fatigue)
- Medical history (e.g., diabetes, hypertension)
- Laboratory test results
- Diagnosis (target variable)

The data is sourced from publicly available healthcare datasets or hospital-provided anonymized records.

Data Preprocessing

- Handling missing values using mean/mode imputation
- Encoding categorical variables using Label Encoding / One-Hot Encoding
- Feature scaling using StandardScaler / MinMaxScaler
- Outlier detection and removal
- Balancing the dataset (e.g., using SMOTE)

Exploratory Data Analysis (EDA)

- Visualizing distributions of patient attributes
- Correlation heatmap to analyze relationships between features
- Identifying patterns and trends in disease occurrence
- Insights into high-risk groups

Feature Engineering

- Creating new features such as risk scores or symptom combinations
- Dimensionality reduction using PCA (if needed)
- Feature selection using mutual information, chi-square, or tree-based importance

Model Building

- Models used: Logistic Regression, Decision Tree, Random Forest, XGBoost, and Support Vector Machine
- Data split into training and testing (80-20)
- Hyperparameter tuning using GridSearchCV / RandomizedSearchCV
- Cross-validation (K-Fold)

Model Evaluation

- Accuracy
- Precision, Recall, F1-score
- Confusion Matrix
- ROC Curve and AUC score
- Comparing model performances and selecting the best one

Deployment

- The best-performing model is saved using `joblib` or `pickle`
- A web interface is created using Flask or Streamlit
- Users can input symptoms and other patient data to get disease predictions in real time
- Deployed locally or on platforms like Heroku or Render

Source Code

Includes:

- data_preprocessing.py

- eda_analysis.ipynb
- model_training.py
- app.py (for deployment)
- requirements.txt

Future Scope

- Integration with electronic health record (EHR) systems
- Real-time data streaming from wearable devices
- Expanding disease coverage to include rare diseases
- Enhancing prediction using deep learning models (e.g., LSTM for time-series health data)
- Multilingual voice interface for rural healthcare outreach

Team Members and Roles

Harish Kumar – Project Leader

- Coordinate the entire project timeline and deliverables.
- Allocate tasks to team members and ensure collaboration.
- Communicate with mentors/supervisors and manage final presentations.
- Oversee overall progress and maintain documentation quality.

Bala Murugan – Data Scientist

- Perform data cleaning, preprocessing, and feature engineering.
- Conduct Exploratory Data Analysis (EDA).
- Identify key variables and build initial ML pipelines.
- Collaborate with ML Engineer on model optimization.

Deepak – Machine Learning Engineer

- Develop, train, and fine-tune machine learning models.

- Evaluate model performance using appropriate metrics.
- Work on model explainability using SHAP/LIME.
- Optimize prediction accuracy and ensure robustness.

Abinesh Godwin – Software & Interface Developer

- Build a simple web-based user interface (using Streamlit, Flask, or Dash).
- Integrate the trained ML model into the application.
- Ensure usability and responsiveness of the dashboard.
- Assist with cloud deployment if needed.

Mohammad Adnan – Report Writer & Tester

- Prepare detailed project documentation, reports, and presentation slides.
- Ensure clarity in technical and non-technical writing.
- Test the web interface and validate model outputs.
- Provide feedback and assist in QA (Quality Assurance).