AI-Powered Disease Predictor

A Project Report

submitted in partial fulfillment of the requirements

of

AICTE Internship on AI: Transformative Learning
with
TechSaksham – A joint CSR initiative of Microsoft & SAP

By

Anand Kumar

anandkashyap6048@gmail.com

Under the Guidance of

Pavan Sumohana

ACKNOWLEDGEMENT

I take this opportunity to express my greatest gratitude to my mentor, Pavan Sumohana, whose strong encouragement, sage advice, and motivational guidance were crucial during the course of this project. Their vast experience, deep knowledge, and precious suggestions played a critical role in making me understand AI-based medical forecasting systems. Not only did they lead me through difficult stages but also helped me finish the project with enhanced accuracy and efficiency.

In addition, I am deeply thankful to AICTE TechSaksham, Microsoft, and SAP for giving a shared learning platform. This internship has immensely developed my technical skills by exposing me to actual AI applications in the healthcare sector. This experience has enlarged my knowledge pool and made me stronger to design and deploy AI models.

Lastly, I would like to extend my sincere gratitude to my peers, relatives, and friends for their persistent encouragement and support. Their steadfast faith in me kept me going through the project, enabling me to successfully complete it.

ABSTRACT

Problem: Increased incidence of chronic diseases overburdens world healthcare.

Solution: Early disease prediction through machine learning.

Methodology:

Public repository data collection.

Data preprocessing and cleaning.

Exploratory Data Analysis (EDA).

Development of models (Random Forest, etc.) for 6 diseases.

Streamlit web application for live predictions.

Outcome: Validates the effectiveness of AI in disease prediction.

Future: Integration of real-time data, increased disease coverage, clinical proof.

Major Factors:

Data Quality: The accuracy of predictions relies significantly on the quality and availability of data that the model is trained on.

Model Selection and Tuning: Selection of machine learning algorithms as well as their parameters plays an important role in predictive performance.

Real-world Validation: Real-world use as well as clinical trials play a vital role in ensuring the effectiveness of the model.

Data Integration: Real-time integration of data from multiple sources becomes a key feature of a dynamic and accurate system.

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1

Introduction

1.1Problem Statement:

"Healthcare facilities find it difficult to process large volumes of patient data for prediction of early disease. Current tools are not accurate, leading to late diagnoses. Predictive systems based on AI are required to process data effectively, enhance diagnostic precision, allow timely interventions, cut costs, and improve patient outcomes while maintaining ethical data management."

1.2Motivation:

"Driven by the demand for quicker, more precise diagnoses, this project seeks to utilize machine learning to develop an automated disease prediction system. Through facilitating early interventions and lowering mortality, we hope to increase patient care efficiency and overall healthcare outcomes."

1.3Objective:

1 "Implement and deploy strong machine learning algorithms (Random Forest, Decision Tree, etc.) to precisely forecast diseases from patient data."

2 "Build an interactive Streamlit web app for live disease forecasting and visualization of AI-powered insights."

3 "Automate data processing and prediction workflows to streamline clinical decision-making and improve healthcare efficiency."

4 - "Give healthcare professionals AI-driven insights to improve diagnostic accuracy and aid in well-informed clinical decisions, ultimately enhancing patient outcomes."

1.4Scope of the Project:

This project will create a Python-based machine learning system to predict Diabetes, Heart Disease, Kidney Disease, Liver Disease, Breast Cancer, Lung Cancer.

The scope includes:

1 Build Python/ML models for 6 disease predictions, including data preprocessing and training.

2 Develop a Streamlit web application for visualizing results.

3 Recognize limitations of accuracy; real-world verification is required.

CHAPTER 2

Literature Survey

2.1 Scholarly Works that Already Exist

Recent studies highlight the expanding participation of computational learning methods in healthcare, specifically for predictive analysis and immediate detection of medical conditions. Random Forest, Decision Trees, Support Vector Machines (SVM), and Logistic Regression are among the algorithms that have repeatedly been used for disease classification and risk prediction. Methods including ensemble learning approaches and complex neural network architectures have demonstrated enhanced accuracy in medical diagnosis. Yet the enabling of real-time predictions and easy system deployment is another area that needs to be investigated further.

2.2 Deployed Systems and Methods

The project employs the Random Forest Classifier in combination with other machine learning algorithms and makes use of Python libraries like NumPy, Pandas, Scikit-learn, and Joblib. The models are conditioned on disease-specific data and involve feature processing and parameter adjustment to maximize predictive accuracy. The system scans patient health data for estimating probabilities of different diseases, such as Diabetes, Heart Disease, Kidney Disease, Liver Disease, Breast Cancer, Lung Cancer, and provides real-time predictive responses**.**

2.3 Shortcomings in Current Solutions

Existing solutions have a number of shortcomings, including:

 Predictive Accuracy and Reliability: Current models tend to lack accuracy due to imbalances in data or missing records.
 Live Data Incorporation: Most solutions fail to support real-time predictive features, which diminishes their utility in clinical applications.
 Confidentiality and Patient Data Protection: Securing sensitive medical data continues to be a challenge in the use of AI in healthcare environments.

CHAPTER 3

Proposed Methodology

3.1 System Design

☐Data Acquisition: Download disease datasets from Kaggle/UCI.

 Data Refinement: Clean data, convert categories, scale features.

Model Building: Build ML models (Random Forest, etc.) for each disease.

Web Deployment: Develop Streamlit application for real-time predictions; provide clear results.

3.2Requirement Specification

Listing down the tool and technologies needed to deploy the solution.

3.2.1 Hardware Requirements:

 Processor : AMD Rygen 5 / Intel i5 or greater

 Ram: 8 GB or greater

 Storage: 1GB or greater

Operating System : Windows10 / linux / macOs

3.2.2 Software Requirements:

 Programming Language: Python

Libraries:

 Data Processing : Pandas,NumPy

Machine Learning: Scikit-learn,Joblib

 Web Development: Streamlit

 Development Environment :VS Code, Jupyter Notebook

CHAPTER 4

Implementation and Result

4.1 Model Building for Disease Prediction

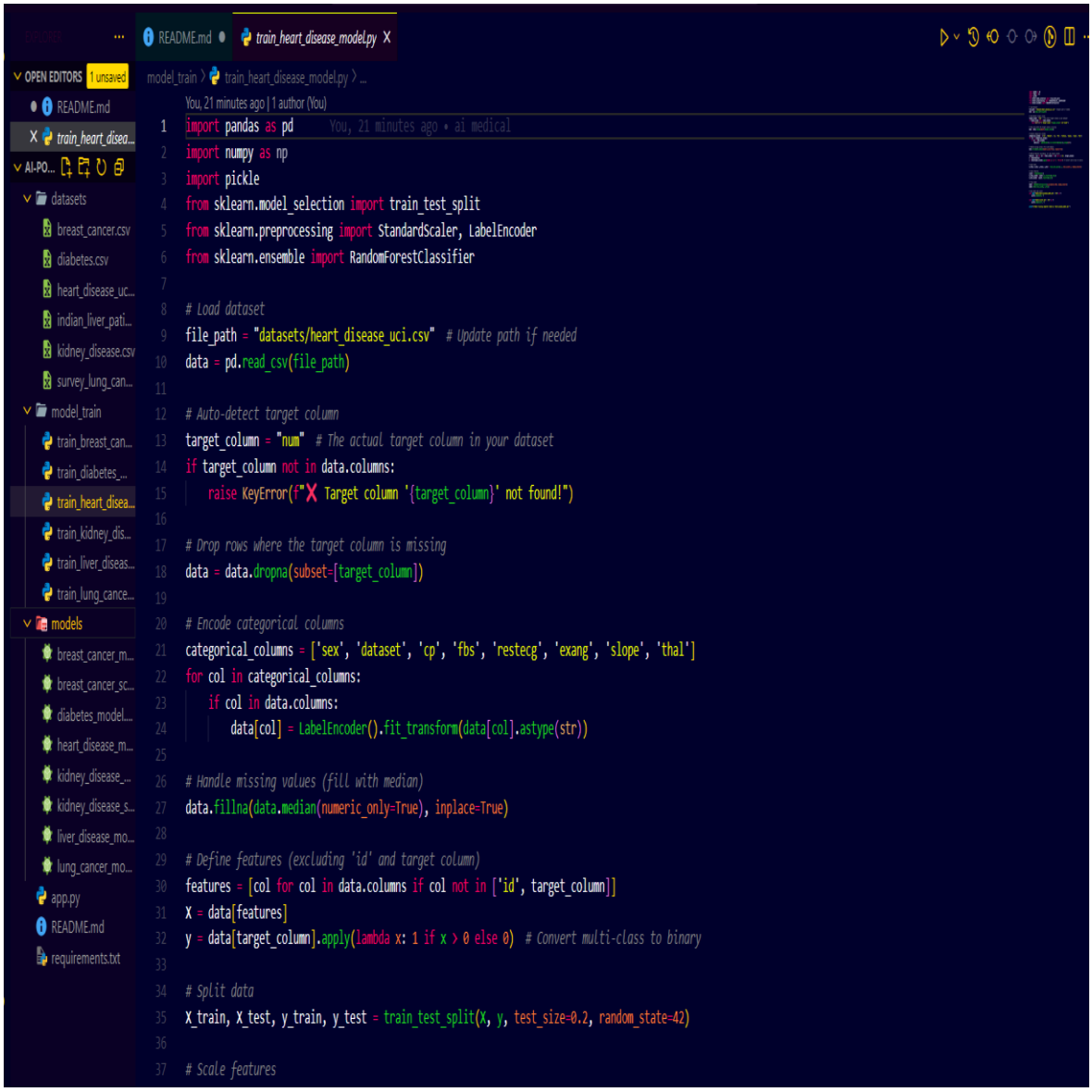
Developing several prediction models for Diabetes, Heart Disease, and Kidney Disease are part of the project using machine learning methods.

- The Heart Disease model predicts heart disease based on features such as age, blood pressure, cholesterol, and other medical factors.
- The Diabetes model predicts if the individual is diabetic or not by considering factors like glucose, insulin, BMI, and skin thickness.
- The Kidney Disease model predict kidney disease based on some factors like age, blood pressure, Specific Gravity, Albumin Level, Sugar Level

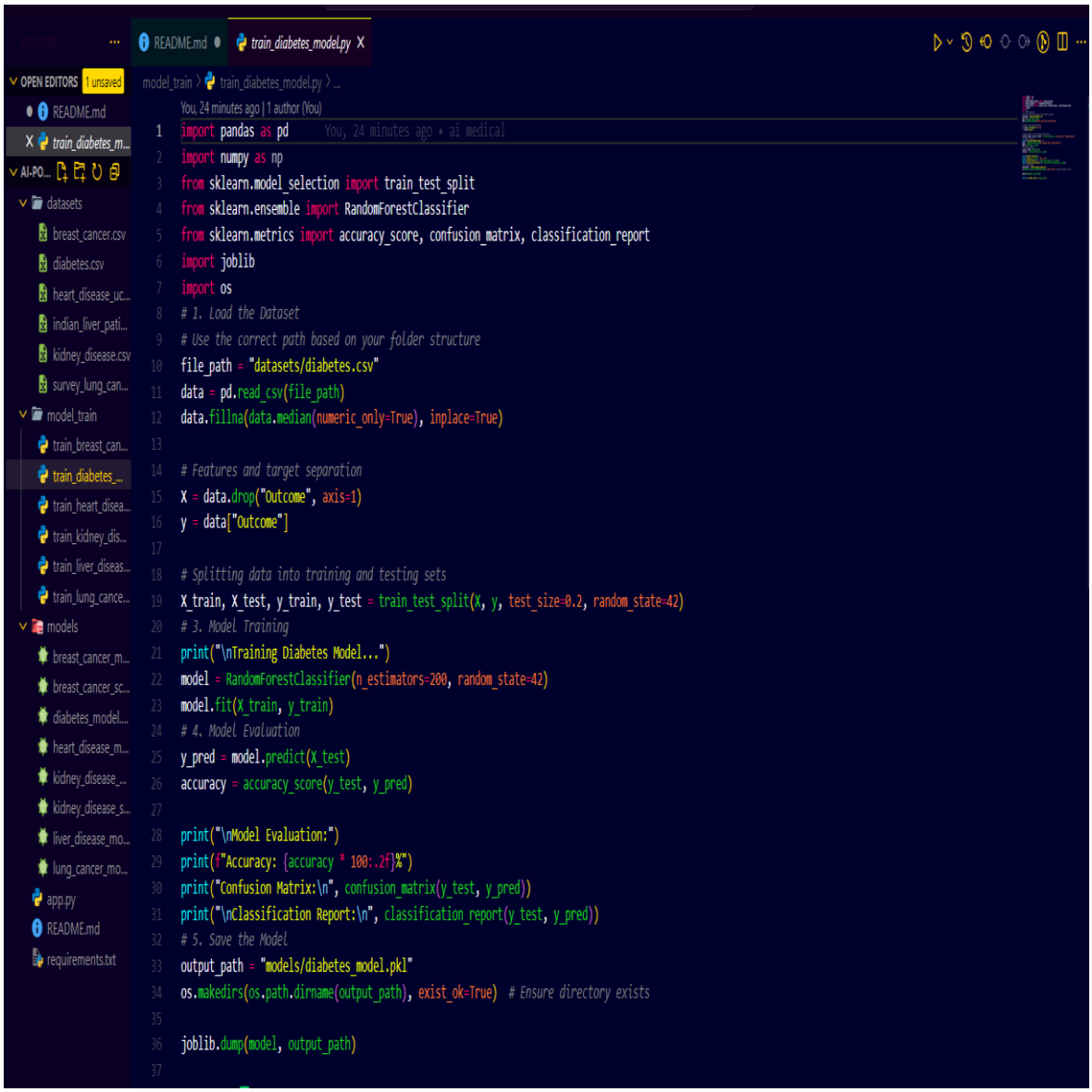
Scikit-Learn and Pandas libraries are used to train the models, and accuracy scores are used to measure the predictions.

Following are some code segments from the model development process:

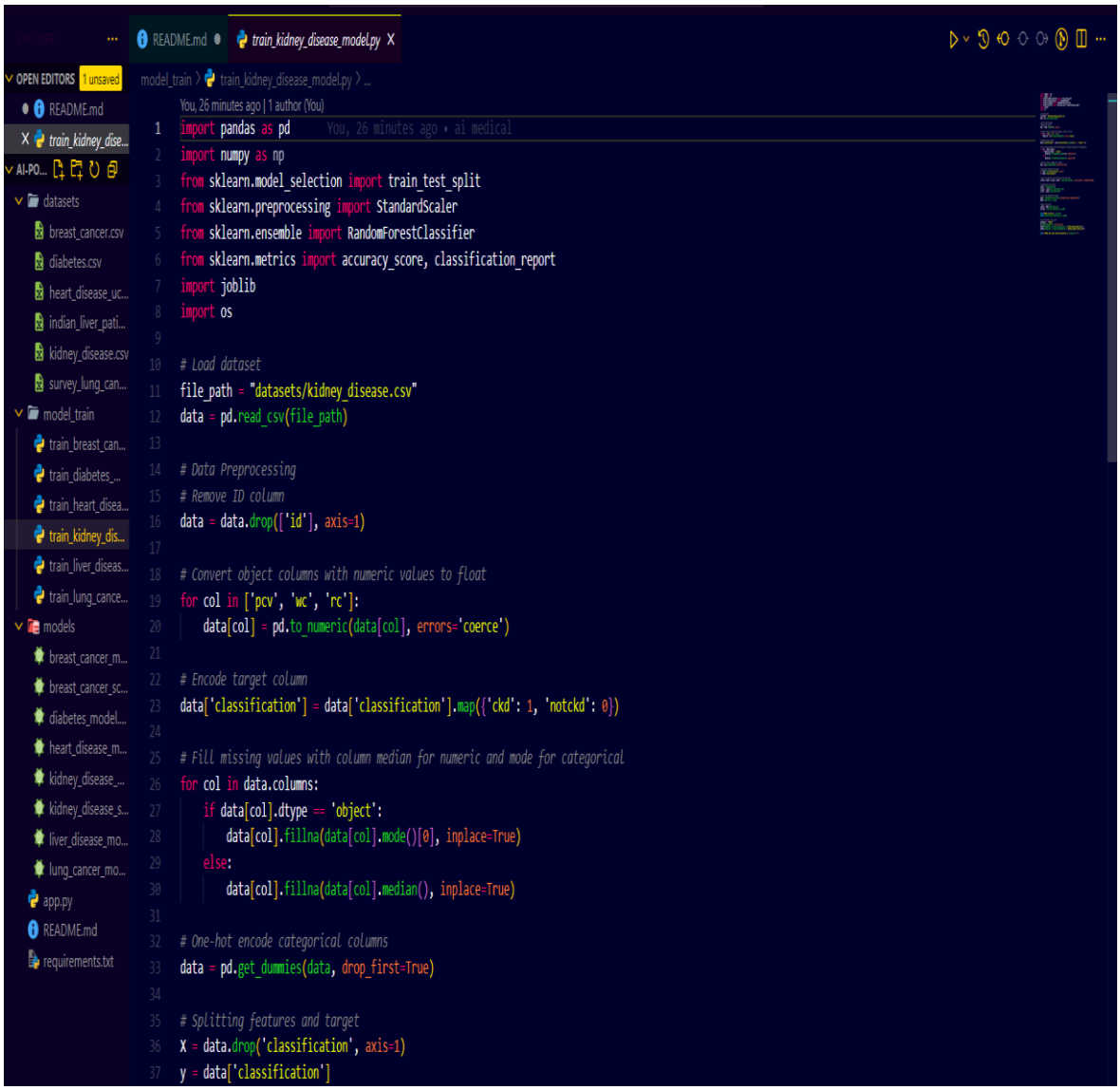
Heart Disease



Diabetes Disease



Kidney Disease

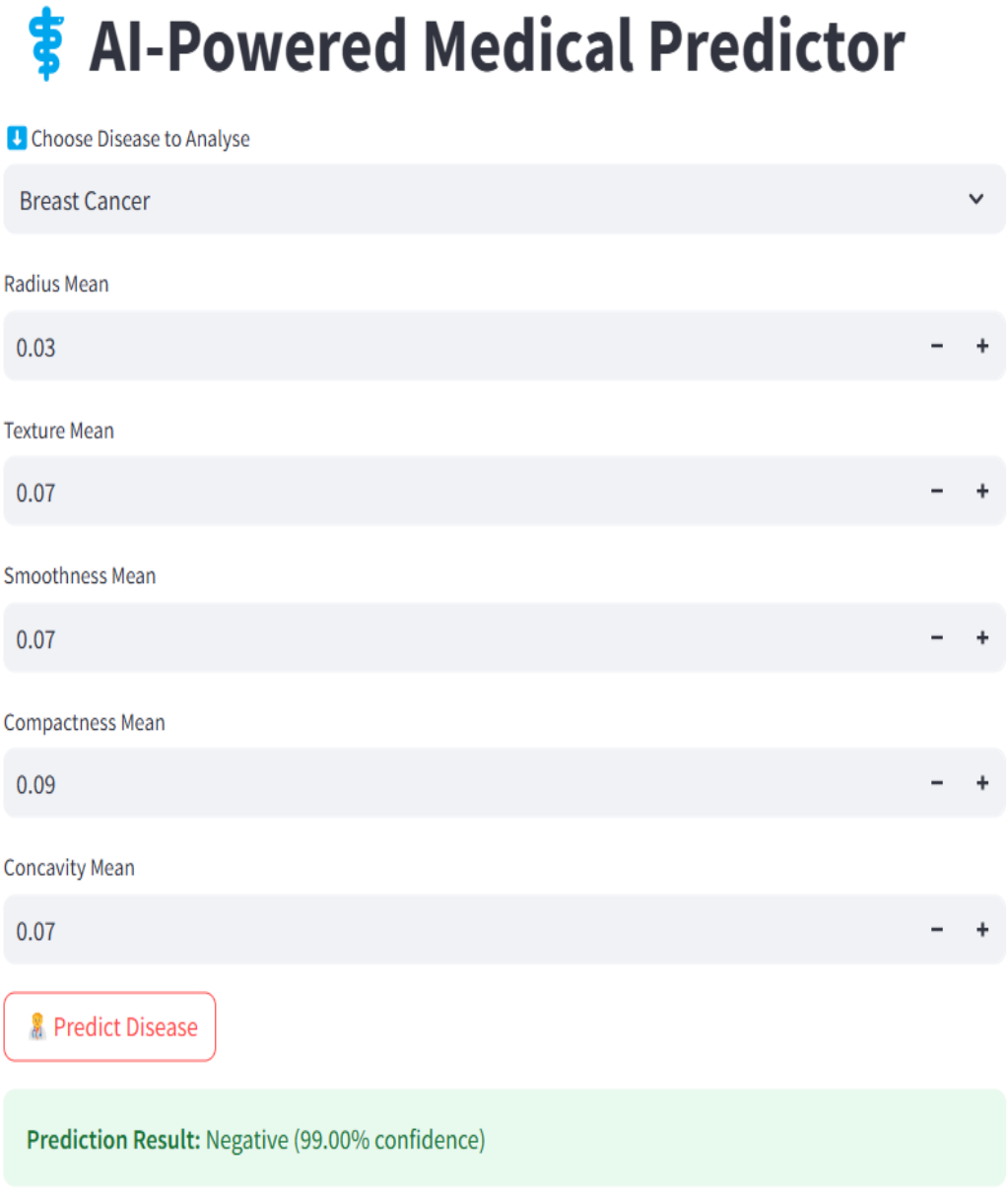


4.2 Web App

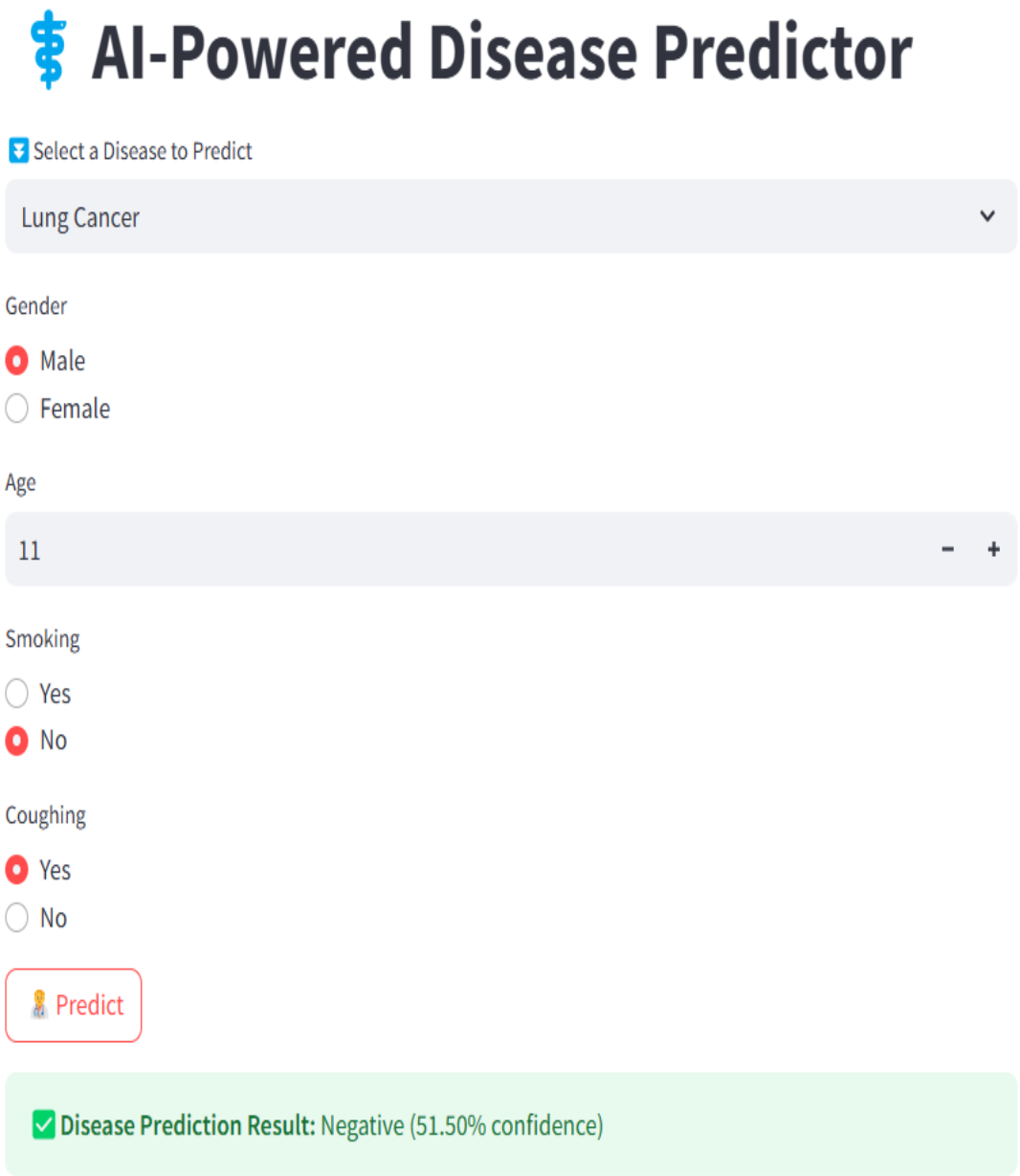
Web application is made by using app.py, allow user to enter health information and get prediction

Below I put some Screenshot that show how it predict disease

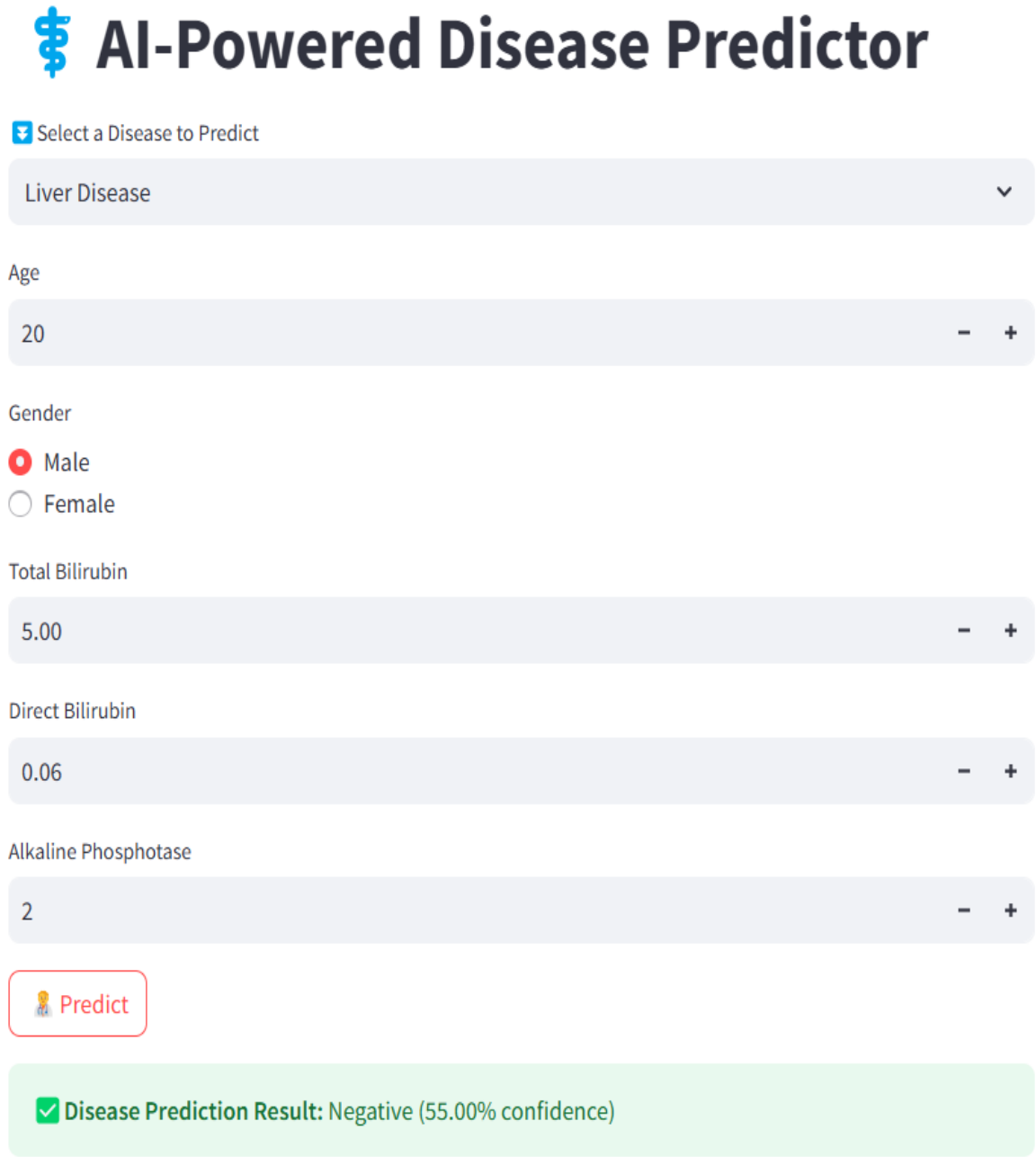
Breast Cancer: -



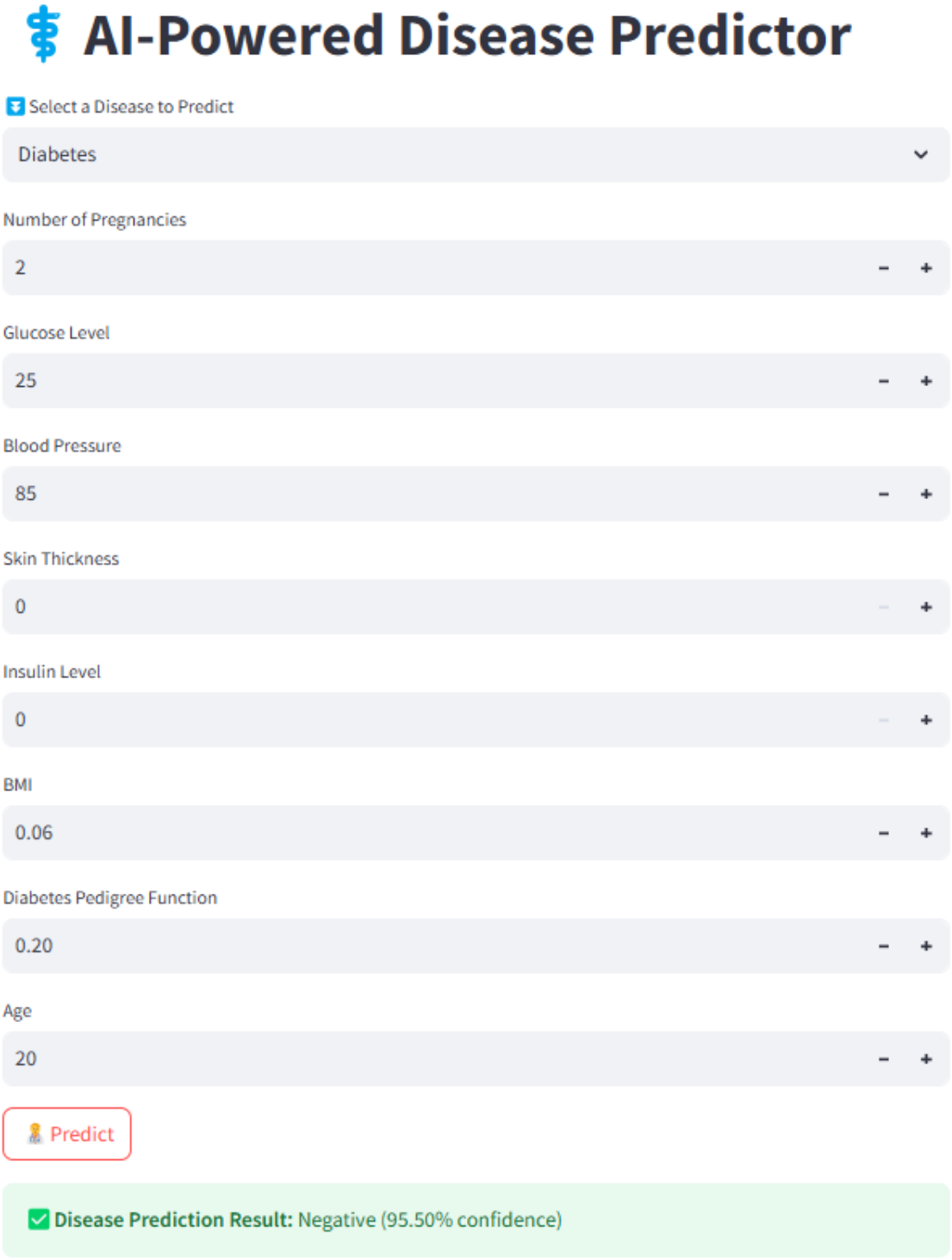
Lung Cancer: -


AI-Powered Disease Predictor
Select a Disease to Predict
Lung Cancer
Gender
Male
Female
Age
11
Smoking
Yes
No
Coughing
Yes
No
Predict
Disease Prediction Result: Negative (51.50% confidence)

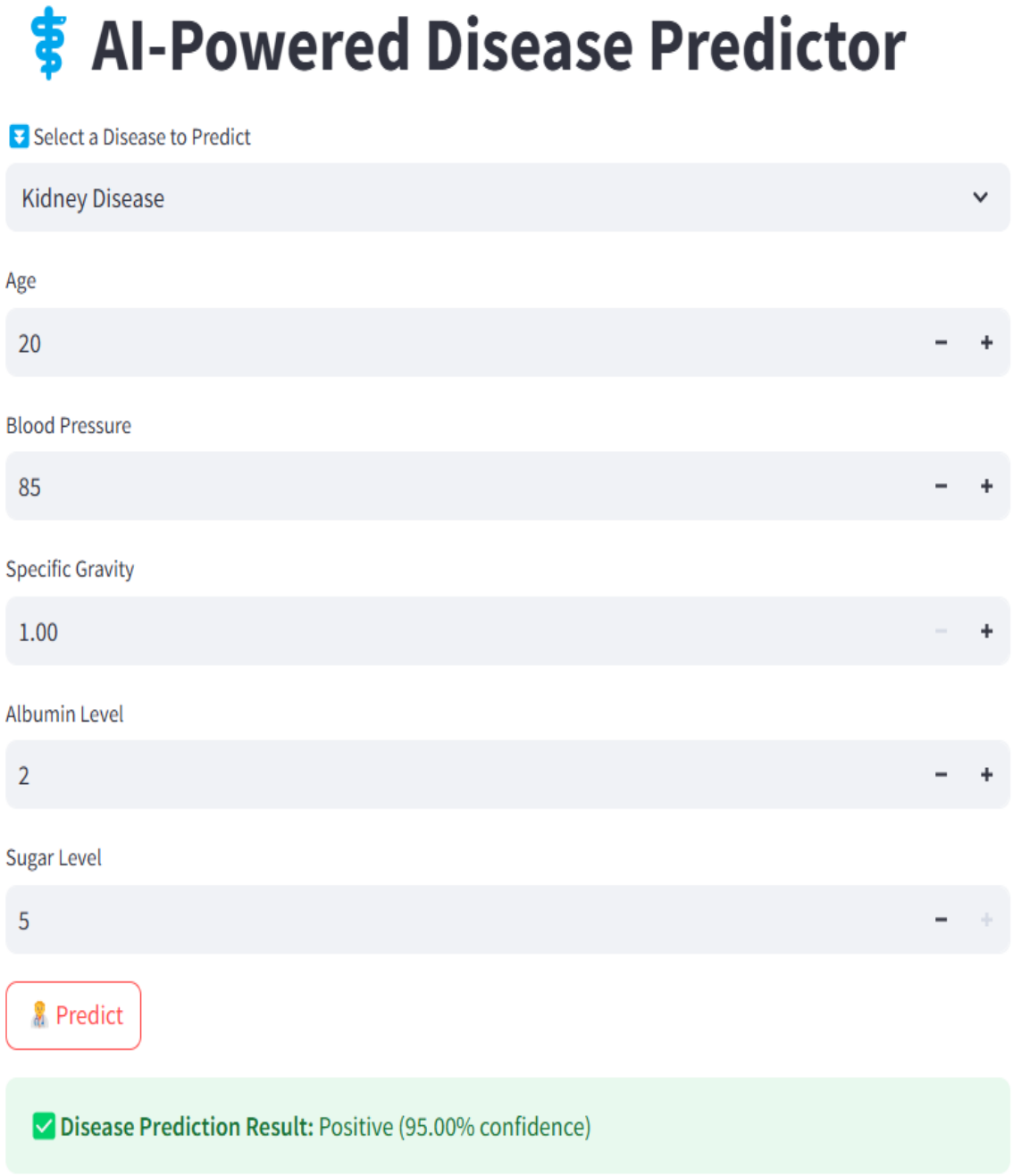
Liver Disease: -



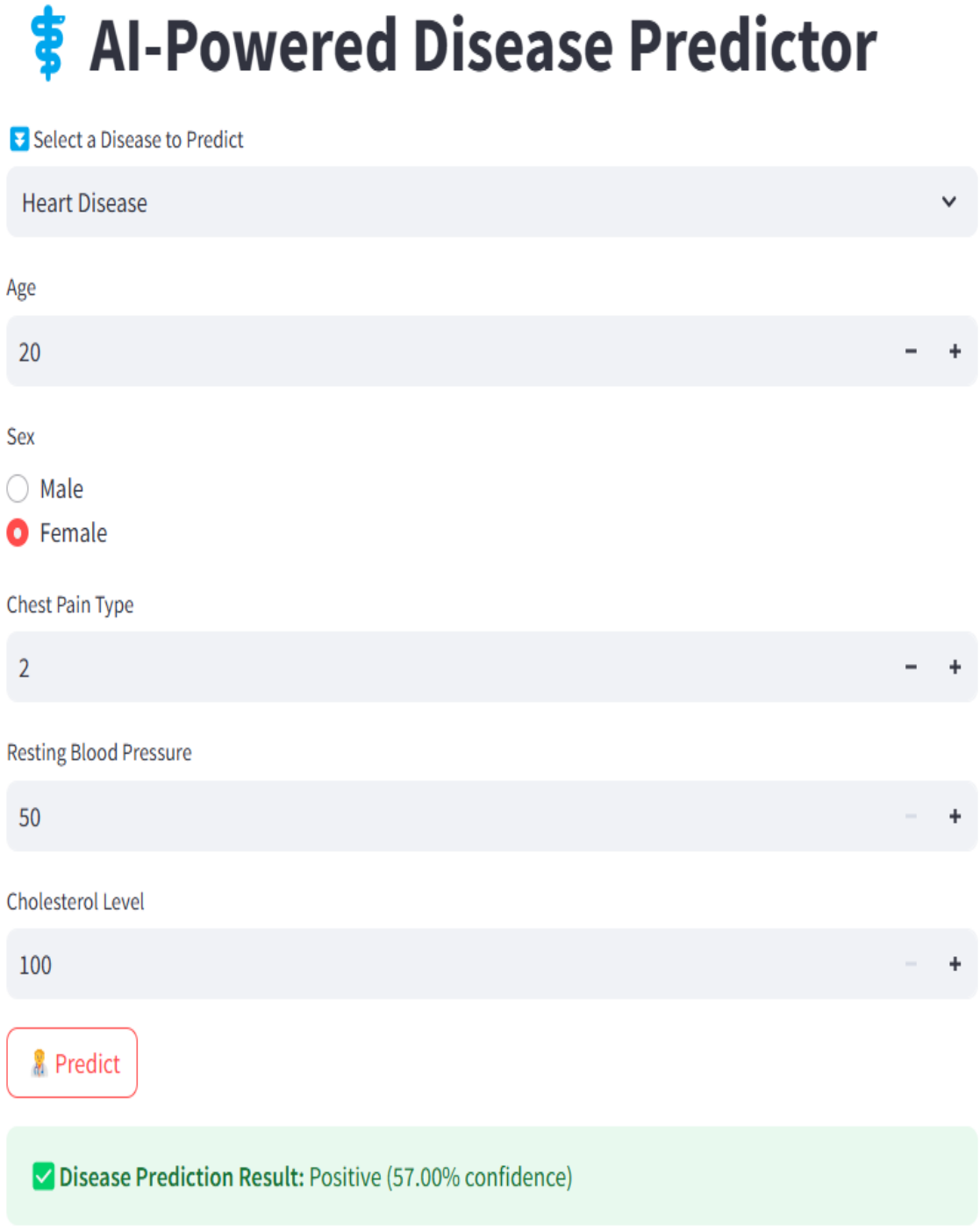
Diabetes: -

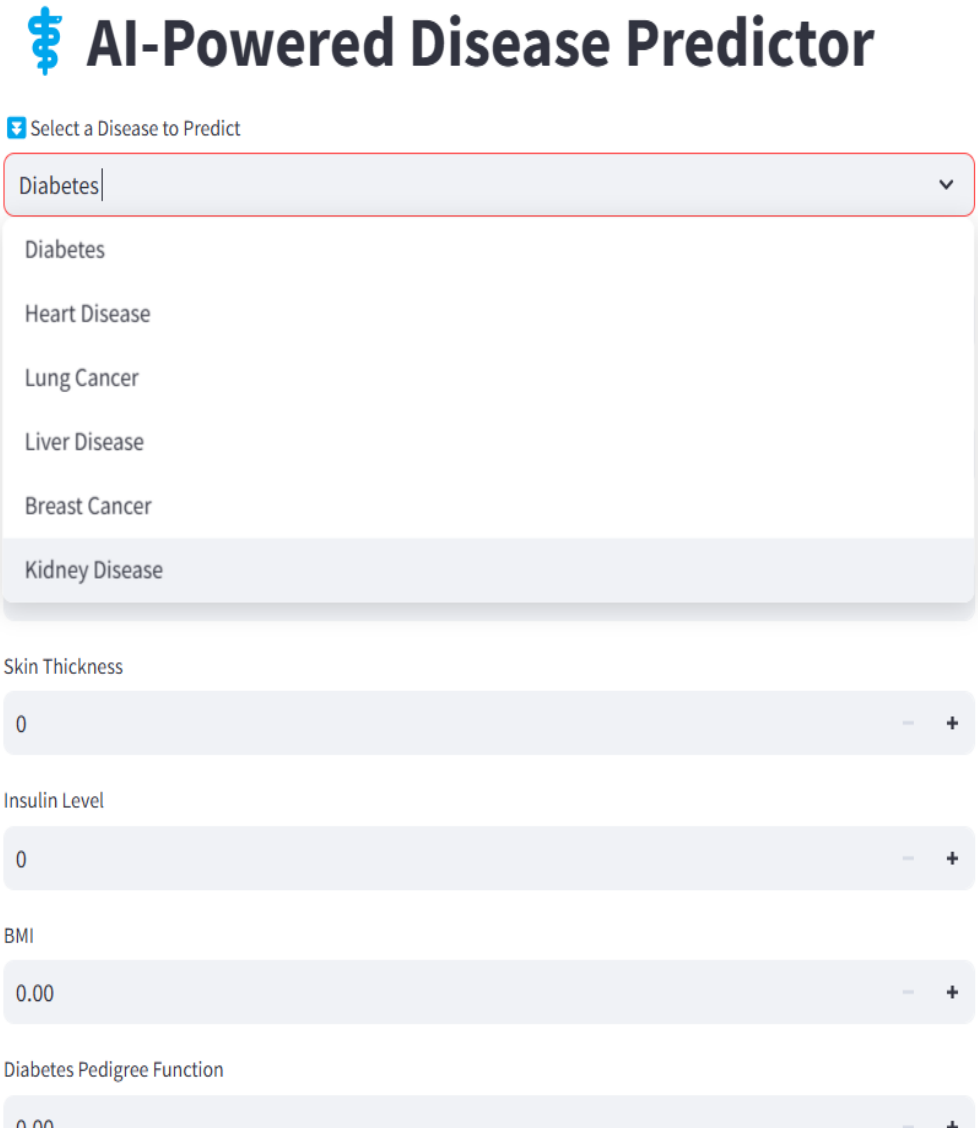

AI-Powered Disease Predictor
Select a Disease to Predict
Diabetes
Number of Pregnancies
2
Glucose Level
25
Blood Pressure
85
Skin Thickness
0
Insulin Level
0
BMI
0.06
Diabetes Pedigree Function
0.20
Age
20
Predict
Disease Prediction Result: Negative (95.50% confidence)

Kidney Disease: -



Heart Disease: -




AI-Powered Disease Predictor
Select a Disease to Predict
Diabetes
Diabetes
Heart Disease
Lung Cancer
Liver Disease
Breast Cancer
Kidney Disease
Skin Thickness
0
Insulin Level
0
BMI
0.00
Diabetes Pedigree Function

4.3 GitHub Link for code

With the help of this link anyone can see the project & code

GitHub Repository-

https://github.com/Anandkumarkashyap/AI-Powered-Disease-Predictor

CHAPTER 5

Discussion and Conclusion

5.1 Future Work:

For optimizing the efficiency and reliability of the system, the following are recommended:

Enhancing predictive precision by using state-of-the-art machine learning:

Using more advanced architectures like Deep Learning Networks (DLNs), Sequential Neural Networks (SNNs), and Image-based Neural Networks (IBNNs).

The use of such new paradigms for learning can greatly improve predictive results and identify more subtle relationships in patient physiological data.

Adding patient data in real-time for dynamic testing:

Facilitating ongoing patient data collection by networked medical devices, sensor technology, or health data interfaces.This would render the site more directly useful and relevant to healthcare professionals by allowing instant and responsive predictions. Expanding model coverage to encompass more varieties of medical conditions:

Creating specialized models for other diseases such as Renal Dysfunction, Hepatic Disorders, Neurodegenerative Diseases, and Pulmonary Neoplasms. This would offer flexibility and usability of the system to various forms of clinical settings.

5.2 Conclusion:

This project successfully demonstrates the application of machine intelligence algorithms for health condition onset prediction. By using trained models on clinical data repositories, the platform offers precise predictions for Glucose Imbalance, Cardiovascular Disorders, and Movement Impairment Syndromes.

Important Observations:

Streamlit graphical user interface provides a friendly interface, offering seamless interaction and ease of use for clinicians and scientific researchers.

These computer models of learning constructed in this project are beneficial for analytical insights by uncovering patterns in patient-provided information, thereby allowing medical personnel to formulate evidence-based treatment schedules.

The design of the system is scalable and allows immediate diagnosis and implementation of additional disease models in the future.

Lastly, the project shows the groundbreaking potential of machine learning to revolutionize medicine, specifically the capacity to predict disease before it happens and enhancing better clinical judgment.

REFERENCES

[1]. Ming-Hsuan Yang, David J. Kriegman, Narendra Ahuja, “Detecting Faces in Images: A Survey”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume. 24, No. 1, 2002.