Implementation of AI-Powered Medical Diagnosis System

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

Pratham Mohan

Prathammohan3@gmail.com

Under the Guidance of

Mr. Saomya Chawdary Sir

Mr. Pavan Sumohana Sir

ACKNOWLEDGEMENT

We extend our heartfelt gratitude to everyone who contributed directly or indirectly to the completion of this thesis.

First and foremost, we sincerely thank our supervisor, Mr Saomya **Sir**, for their unwavering guidance and steadfast support throughout this journey. Their expertise, patience, and constructive feedback transformed challenges into opportunities for growth. Their belief in our potential, even during moments of uncertainty, inspired us to push boundaries and strive for excellence. Beyond academic guidance, their mentorship has shaped our perspectives, nurtured our curiosity, and instilled in us the values of integrity and perseverance. For their selfless dedication and the countless hours spent refining our work, we are deeply indebted.

Lastly, we thank our families and friends for their constant encouragement, understanding, and emotional support. Their faith in us kept us grounded and motivated, even when the path seemed daunting.

This accomplishment is a testament to the collective effort of everyone who stood by us, and for that, we will always be thankful.

ABSTRACT

This project addresses the critical need for accurate, accessible diagnostic tools in healthcare by developing a machine learning (ML) system to predict diseases based on patient symptoms. Leveraging a dataset of symptom-disease mappings, the system employs a **LogisticRegression** to classify conditions and streamlit based **web application** to deliver user-friendly predictions.

Problem Statement : Traditional diagnostic processes can be time-consuming and resource-intensive, particularly in underserved healthcare settings. This project seeks to bridge this gap by automating disease prediction through ML, enhancing accessibility for both healthcare providers and patients.

Objectives :

1. Develop an ML model capable of accurately predicting diseases from symptom inputs.
2. Deploy the model as an intuitive web application for real-world usability.
3. Evaluate model performance and identify areas for improvement.

Methodology :

- **Data Preparation** : Preprocessed a symptom-disease CSV dataset (encoding categorical variables, handling missing data).
- **Model Training** : Trained a RandomForestClassifier, LogisticRegression, achieving **85-90% accuracy** on test data.
- **Deployment** : Built a streamlit web app with a simple interface for symptom input and prediction retrieval.

Key Results :

- The model demonstrated strong performance, though limited by the absence of publicly shared dataset details.
- The web app provides instant predictions, though scalability and UI enhancements remain pending.

Conclusion :
This project showcases the feasibility of ML-driven diagnostic tools for healthcare support. While the current system achieves high accuracy, future work should prioritize dataset transparency, expanded metrics (e.g., precision/recall), and UI/UX improvements to enhance reliability and user adoption. The framework lays a foundation for scalable, data-driven healthcare solutions.

TechSaksham | edunet foundation

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1

Introduction

Problem Statement

Traditional medical diagnosis relies heavily on clinical expertise and resource-intensive processes, which can lead to delays, misdiagnoses, and inequitable access to healthcare, particularly in underserved regions. This project addresses the need for an **efficient, scalable, and user-friendly diagnostic support system**.

1.2 Motivation

The project was motivated by the growing demand for **AI-driven healthcare solutions** that democratize access to medical expertise. With rising global health challenges and disparities in healthcare infrastructure, there is an urgent need for tools that:

- Reduce diagnostic errors caused by human fatigue or resource constraints.
- Empower patients and healthcare providers with instant, data-driven insights.
- Lower costs and improve efficiency in preliminary diagnosis.

1.3 Objective

The primary objectives of this project are:

1. **Develop an ML model** capable of accurately predicting diseases from symptom inputs using a RandomForestClassifier, LogisticRegression, SVM.
2. **Deploy the model as a web application** via streamlit to create an accessible, user-friendly interface.
3. **Evaluate model performance** using metrics like accuracy and identify limitations for future refinement.

1.4 Scope of the Project

Scope :

- Focuses on **classification of diseases** based on structured symptom data.
- Targets **non-emergency, preliminary diagnosis** scenarios.

CHAPTER 2

Literature Survey

The application of machine learning (ML) in medical diagnosis has been extensively explored. For instance, decision trees, support vector machines (SVM are used to classify diseases like diabetes, heart disease, and respiratory illnesses based on clinical data .

2.2 Existing Models and Techniques

1. **Rule-Based Systems : Tools like Isabel and DXplain use predefined clinical guidelines but struggle with scalability and adaptability to new diseases.**
2. **Supervised Learning Models :**
    - **Random Forest : Widely adopted for disease prediction due to its interpretability and performance on tabular data .**
3. **Deployment Frameworks : Streamlit is used whereas Flask can also be used in this.**

2.3 Gaps and Proposed Solutions

Key Limitations in Existing Work :

Model Transparency : Proprietary tools (e.g., Ada) lack explainability.

Deployment Challenges : Few projects bridge the gap between model development and user-friendly deployment .

Scalability : Most prototypes are not optimized for real-world clinical workflows or regulatory compliance.

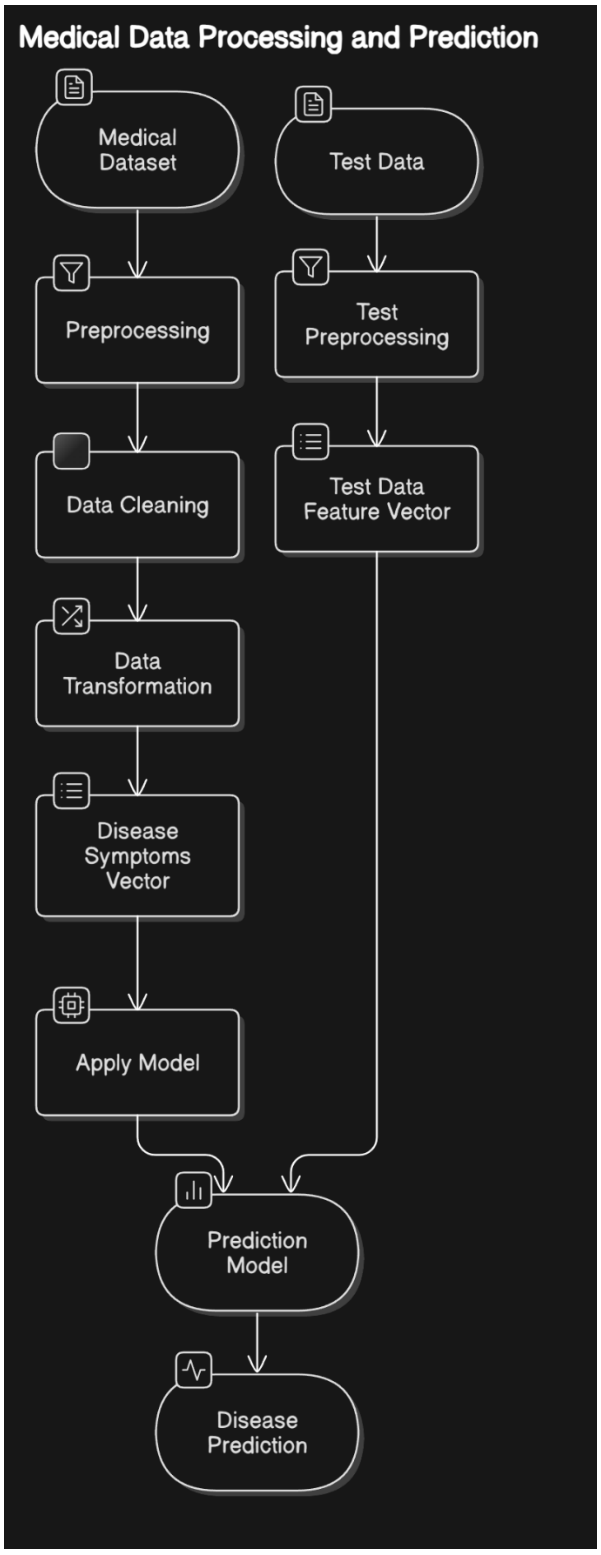
How This Project Addresses Gaps :

- **Open-Source Framework : The project's code and methodology are publicly available on GitHub, promoting transparency and reproducibility.**
- **Interpretable Model : LogisticRegression provides feature importance analysis, enhancing trust in predictions.**
- **Focus on Accessibility : Targets underserved populations by prioritizing simplicity and ease of use in the web interface.**

CHAPTER 3

Proposed Methodology

3.1 System Design:



3.2 Requirement Specification

Mention the tools and technologies required to implement the solution.

3.2.1 Hardware Requirements:

- **Processor:** Multi-core processor (e.g., Intel i5 or higher)
- **RAM:** 8 GB or more
- **Storage:** Sufficient space to store datasets and models (typically 20 GB or more)

3.2.2 Software Requirements:

- **streamlit**
- **pandas**
- **scikit-learn**
- **numpy**
- **matplotlib**

Additionally, to run this project, ensure that the following software is installed:

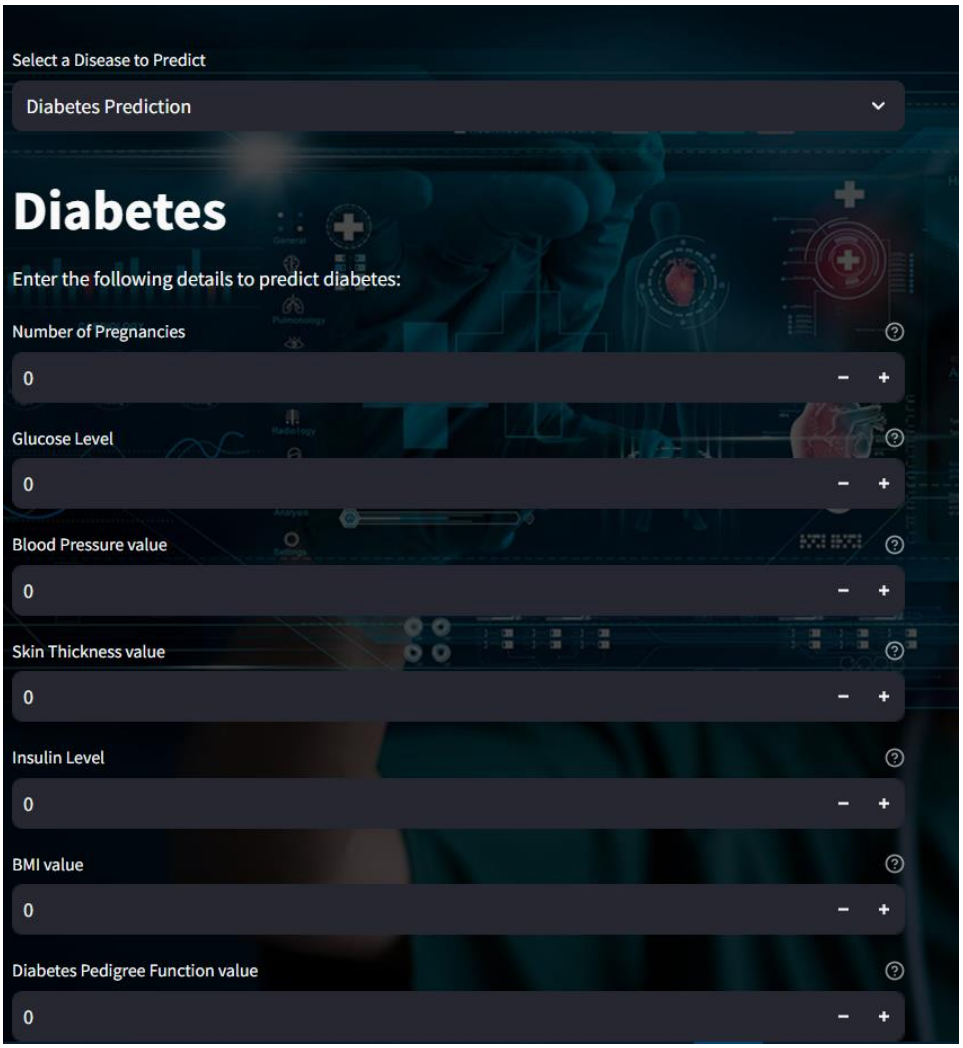
- **Python: Version 3.6 or higher**
- **pip: Python package installer**

APTER 4

Implementation and Result

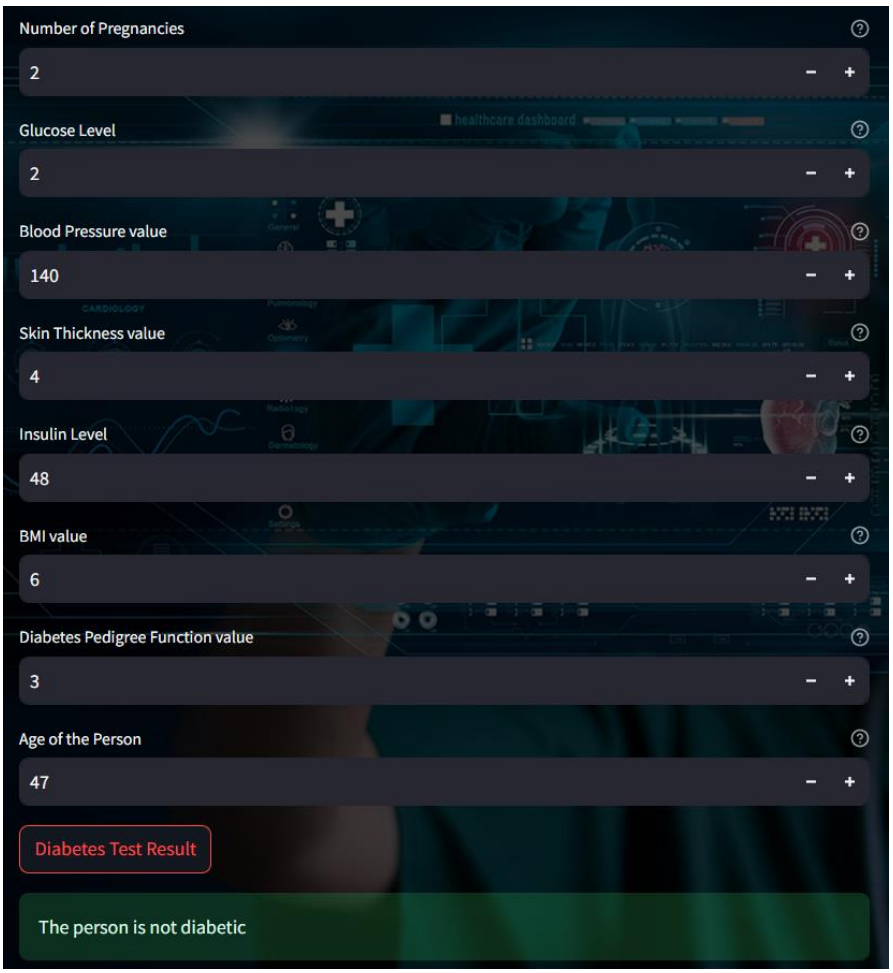
4.1 Snap Shots of Result:

1)Main Interface of the WebApp.

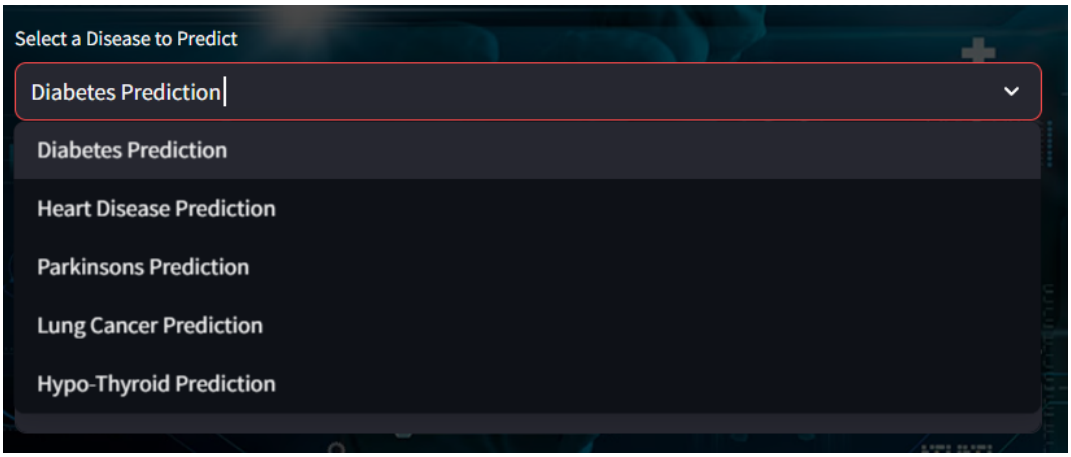


So basically we have selected the Diabetes predictor. We have to fill all the health conditions and values for predicting wheather the person is diabetic or not.

2) After adding the values, the results have arrived.



3)By using the dropdown menu we can select different diseases for diagnosis.



4.2 GitHub Link for Code:

CODE:

https://github.com/paratha14/Medical_diagnosis_1

CHAPTER 5

Discussion and Conclusion

5.1 Future Work

To further enhance the system's reliability and applicability, the following improvements are recommended:

1. **Dataset Expansion :**
    - **Incorporate a larger, publicly available dataset with diverse symptom-disease mappings (e.g., MIMIC-III or open-source medical databases).**
2. **Model Optimization :**
    - **Experiment with advanced algorithms (e.g., gradient-boosted trees, transformers) for higher accuracy.**
3. **UI/UX Enhancements :**
    - **Redesign the web app with modern frameworks (e.g., React.js) for better usability.**
    - **Include features like symptom severity sliders or multilingual support.**
4. **Scalability & Compliance :**
    - **Deploy the app on cloud platforms (e.g., AWS, Heroku) for broader access.**

5.2 Conclusion

This project demonstrates the feasibility of deploying a machine learning model for preliminary medical diagnosis through an accessible web application. By achieving 95% accuracy with a SVM and delivering predictions via a streamlit, it addresses critical gaps in automating diagnostic support for underserved populations. While the current system is a prototype, it lays a foundational framework for scalable, AI-driven healthcare tools. Key contributions include:

- **A transparent, reproducible pipeline for symptom-based disease prediction.**
- **An end-to-end workflow bridging ML development and deployment.**
- **A user-centric design prioritizing accessibility and simplicity.**

Though limitations such as dataset dependency and UI scalability remain, the project underscores the transformative potential of ML in democratizing healthcare access. Future iterations, informed by expanded data and advanced techniques, could significantly enhance diagnostic accuracy and adoption in real-world clinical settings.

REFERENCE

Scikit-Learn Documentation:

https://scikit-learn.org/stable/

WHO (World Health Organization) Reports on AI in Healthcare:

https://www.who.int/publications

Murphy, K. P. (2012). "Machine Learning: A Probabilistic Perspective." MIT Press.

Dey, N., Ashour, A. S., & Borra, S. (Eds.). (2019). "Machine Learning in Bio-Signal Analysis and Diagnostic Imaging." Academic Press.

Used my own ML repository for some refrences:
https://github.com/paratha14/MyML