Multiple Disease Detection using Machine

Learning Algorithms

Project ID: 29242

B.Tech. Project Report submitted for fulfillment of

the requirements for the

Degree of Bachelor of Technology

Under Biju Pattnaik University of Technology

Submitted By

Shubham Shatabdi ROLL NO. CSE202040114 Vaibhav Kr. Choudhary ROLL NO. CSE202040035



2023 – 2024

Under the guidance of

Prof. Ashutosh Parida

NIST INSTITUTE OF SCIENCE & TECHNOLOGY (Autonomous) Palur Hills, Berhampur, Odisha – 761008, India

ABSTRACT

Our point is to anticipate the various sorts of illness in a single stage by utilizing the inbuilt python module

Streamlit. In this task we are utilizing Naïve Bayes algorithm, random forest, decision tree and svm classifier are utilized for prediction of a particular disease. The calculation which gives more accuracy is used to train the data set before implementation. To implement multiple disease analysis using machine learning algorithms, Streamlit and python pickling is utilized to save the model behavior. In this article we analyze Diabetes analysis, Heart disease and Parkinson’s disease by using some of the basic parameters such as Pulse Rate, Cholesterol, Blood Pressure, Heart Rate, etc., and also the risk factors associated with the disease can be found using prediction model with good accuracy and Precision. Further we can include other kinds of chronic diseases, skin diseases and many others. In this work, demonstrating that using only core health parameters many diseases can be predicted. The significance of this analysis is to analyze the maximum diseases to screen the patient's condition and caution the patients ahead of time to diminish mortality proportion. To implement multiple disease analysis used machine learning algorithms, Streamlit. We have considered three diseases for now that are Heart, Parkinson’s, and Diabetes and in the future, many more diseases can be added. The user has to enter various parameters of the disease and the system would display the output whether he/she has the disease or not. This project can help a lot of people as one can monitor the persons’ condition and take the necessary precautions thus increasing the life expectancy.

ACKNOWLEDGEMENT

We would like to take this opportunity to thank all those individuals whose invaluable contribution in a direct or indirect manner has gone into the making of this project a tremendous learning experience for me.

It is our proud privilege to epitomize my deepest sense of gratitude and indebtedness to our faculty guide, Prof. Ashutosh Parida for his valuable guidance, keen and sustained interest, intuitive ideas and persistent endeavour. His guidance and inspirations enabled us to complete our report work successfully.

We give our sincere thanks to Dr. Kunjabihari Swain, B.Tech. Project Coordinator and Dr. Susmita Mahato, CSE Project Coordinator, for giving us the opportunity and motivating us to complete the project within stipulated period of time and providing a helping environment.

We acknowledge with immense pleasure the sustained interest, encouraging attitude and constant inspiration rendered by Prof. (Dr.) Sukant K. Mohapatra (Chairman), Dr. Rajesh Kumar Panakala (Principal), Prof.(Dr.) Brojo Kishore Mishra N.I.S.T. Their continued drive for better quality in everything that happens at N.I.S.T. and selfless inspiration has always helped us to move ahead.

Shubham Shatabdi Vaibhav Kr. Choudhary

Roll- 202040114 Roll- 202040035

TABLE OF CONTENTS

ABSTRACT..............................................................................................................3

ACKNOWLEDGEMENT..........................................................................…………...4

CONTENTS……………….....................................................................……………..5

1.INTRODUCTION....................................................................................................7

1.1 Background

1.2 Motivation

1.3 Objectives

1. Literature Review
   1. Diabetes Research
   2. Chronic heart diseases
   3. Machine Learning in Healthcare
2. Methodology
   1. Data Collection
   2. Data Preprocessing
   3. Machine Learning Models
      1. KNN
      2. Random Forest
      3. XG Boost

3.5 System Architecture

1. Data Analysis and Model Training
   1. Exploratory Data Analysis
   2. Model Training and Validation
   3. Performance Metrics

1. User Interface Design
   1. Input Parameters
   2. Output Presentation
   3. User Interaction

1. Implementation
   1. Software and Tools
   2. Integration of Machine Learning Models

6.4 User Interface Implementation

1. Results and Discussion
   1. Model Accuracy and Performance

7.3 Discussion on Recommendations

1. System Study
   1. Feasibility Study

1. Conclusion

1. References

1. Appendices
   1. Code Listings
   2. Additional Figures and Tables

1. Introduction

Background:

Detecting multiple diseases using machine learning (ML) is a groundbreaking approach that combines the power of data analysis with the intricacies of medical diagnosis. In traditional healthcare systems, identifying various diseases often relies on manual examination, which can be timeconsuming and prone to human error. ML, however, offers a more efficient and accurate method for disease detection. Motivation:

Detecting multiple diseases using machine learning is like holding a torch in the darkness of uncertainty. This project has the power to revolutionize healthcare, making diagnoses faster, more accurate, and potentially saving lives. Imagine the positive ripple effect it could have on individuals and communities, providing early interventions and improved treatment outcomes.

Objectives:

In multiple disease prediction, it is possible to predict more than one disease at a time. So, the user doesn’t need to

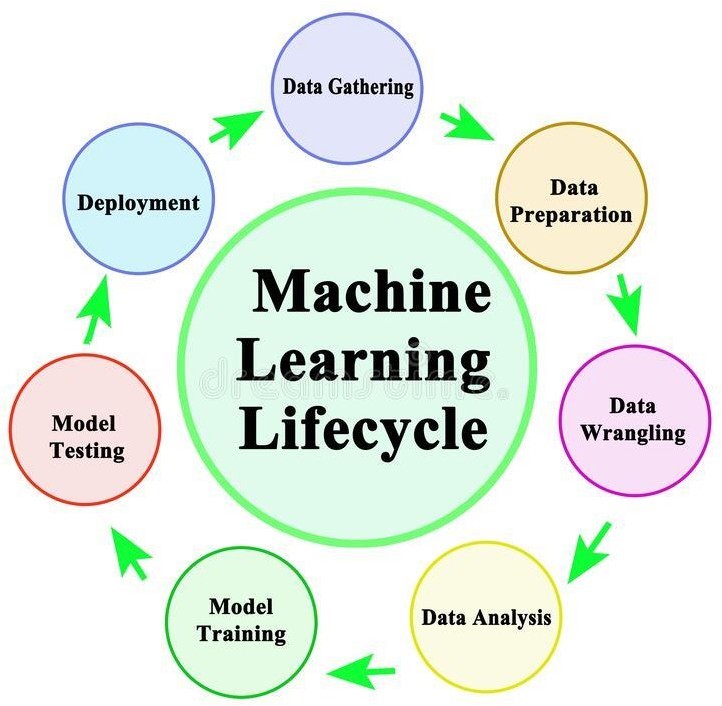
Traverse different sites in order to predict the diseases. We are taking three diseases that are Parkinson’s, Diabetes, and Heart. As all the three diseases are correlated to each other.

Data Sources:

Dataset from kaggle is selected as the experimental dataset to evaluate the Model performance. The Dataset consists of collection of Thousands of Medical equipments data through iot devices and medical reports. The main Aim is to train the Machine learning model is to get trained on the different types of data. So that it can be used for the prediction of Certain disease outcome whether the person is suffering from it or not on the basis of different conditions.

This dataset is evaluated with different machine learning models like Logistic Regression, Decision Tree and Random forest.

# Machine Learning Lifecycle



There are 7 stages in machine learning lifecycle. They are –

1. **Gathering data:** Here Dataset is collected from UCI machine learning repository.
2. **Data Preparation:** This step is further divided into 2 steps i.e
   1. **Data** exploration: It is used to understand the characteristics, format and quality of data. In this we find correlations, general trends and outliers.
   2. **Data** preprocessing: It includes the steps like column naming, feature selection and extraction etc.
3. **Data Wrangling:** It is the process of cleaning and converting raw data into useable format. Cleaning of data is required to address quality issues. In real world applications, collected data may have various issues, including
   1. Missing value
   2. Duplicate data
   3. Invalid data
   4. Outliers
4. **Analyse data:** Now the cleaned and prepared data is passed on to the analysis step. This step involves:
   1. Selection of analytical techniques
   2. Building models
   3. Review the result
5. **5.Train Model:** We use datasets to train the model using various machine learning algorithms. Training a model is required so that it can understand the various patterns, rules and features.
6. **Test Model:** In this step, we check the accuracy of model by providing test dataset to it. Testing the model determines the percentage accuracy of the model as per requirement of project or problem.
7. **Deployment:** In this step we deploy the model in real world system. Before deploying project, we will check whether it is improving its performance or not. The deployment phase is similar to making final report for a project.

# Methodology

Data Collection:

* Data is collected through electronic health records, medical imaging, or laboratory results.
* Addressing data diversity by including information from various demographics and regions.

Data Pre-processing:

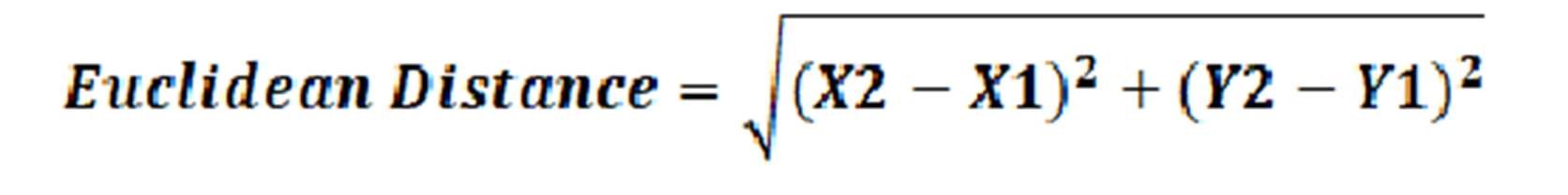
* Identifying and handling missing values in the dataset. Depending on the extent of missing data, options include imputation, removal of incomplete records, or using advanced techniques.
* Addressing any outliers in the data that might adversely affect model training.
* Standardizing or normalizing numerical features to ensure that they are on a similar scale. This helps prevent certain features from dominating the model training process due to their larger magnitudes.

Machine Learning Models:

The working of the K-NN algorithm is as followed:

Step-1: Start to select the K value for example k=5

Step-2: Then we will find the Euclidean distance between the points. It is calculated by the as:



Step-3: Then we will calculate the Euclidean distance of the nearest neighbour.

Step-4: Then count the number of the data points in each category .For example, find three values for Category A and two values for category B.

Step-5: Then assign the new point to the category having the maximum number of neighbours. For example, Category A has the highest number of neighbours so we will assign the new data point to category A.

Step-6: So finally, our KNN model is ready.

Random Forest Algorithm:

Random Forest working is possible in two phases, first is to create the random forest by merging N decision trees, and second is making predictions for each tree created in the first phase.

The working of the random forest is as follows:

Step-1: Firstly, it will select random K data points from the training set.

Step-2: After selecting k data points then building the decision trees associated with the selected data points (Subsets).

Step-3: Then choose the number N for decision trees that you want to build.

Step-4: Repeating steps 1 and 2.

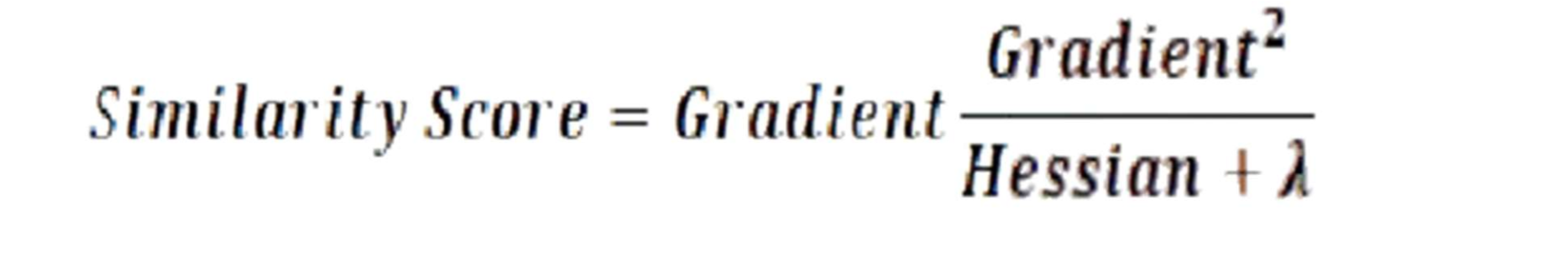
Step-5: Finding the predictions of each decision tree, and assigning the new data points to the category that wins the majority votes.

XG Boost Algorithm:

The working of XG Boost algorithm are as follows: Step 1: Firstly, create a single leaf tree.

Step 2: Then for the first tree, we must compute the average of the target variable as prediction and then calculate the residuals using the desired loss function and then for subsequent trees the residuals come from prediction that was there in the previous tree.

Step 3: Calculating the similarity score using formula:



where, Hessian is equal to the number of residuals; Gradient2 = squared sum of residuals; λ is a regularization hyper parameter.

Step 4: Applying a similarity score we select the appropriate node. The higher the similarity score the more homogeneity.

Step 5: Applying similarity scores we calculate Information gain. Information helps to find the difference between old similarity and new similarity and tells how much homogeneity is achieved by splitting the node at a given point. It is calculated by the formula:



Step 6: Creating the tree of desired length using the above method pruning and regularization can be done by playing with the regularization hyper parameter.

Step 7: Then we can predict the residual values using the Decision Tree you constructed.

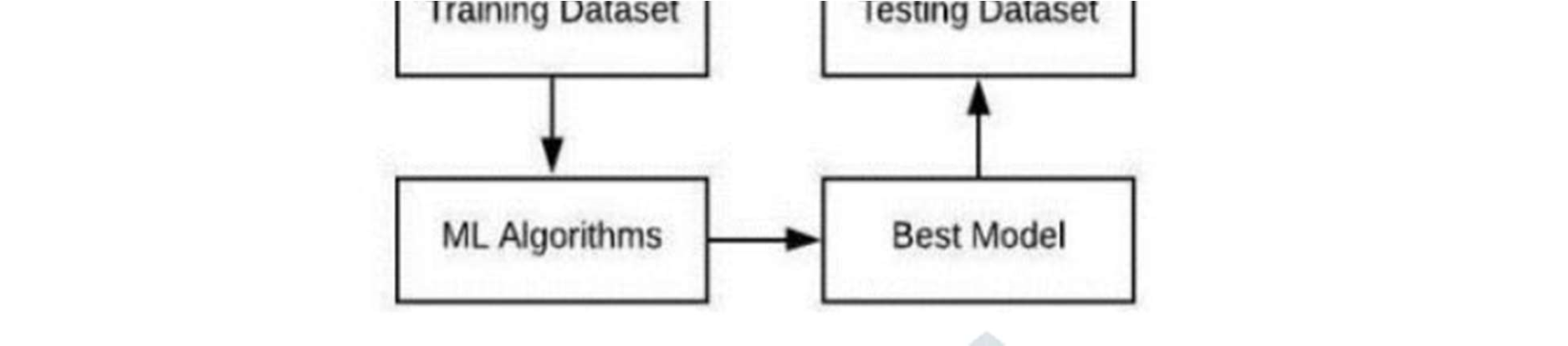
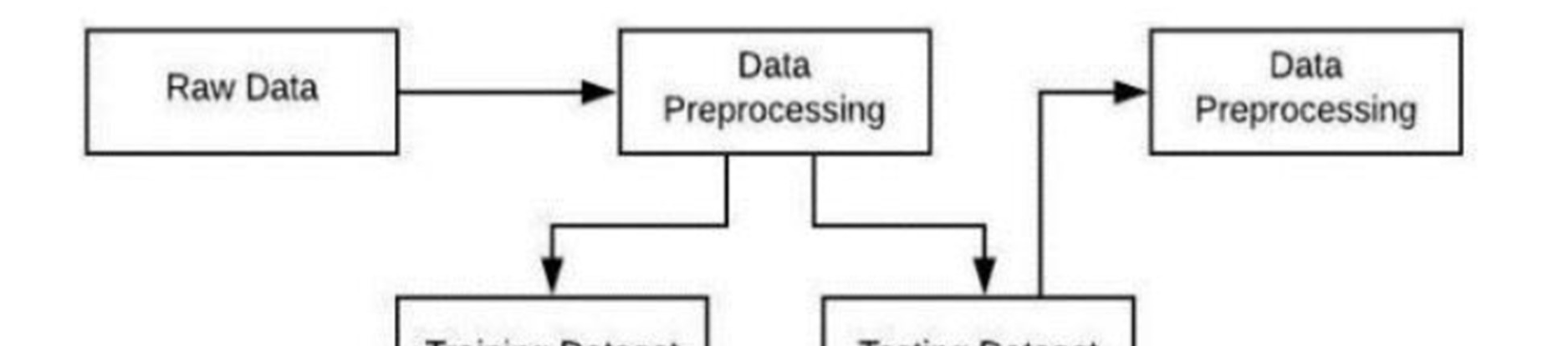
Step 8: The new set of residuals is calculated as:

where ρ is the learning rate.



Step 9: Then go back to step 1 and repeat the process for all the trees

System Architecture:



# Data Analysis and Model Training

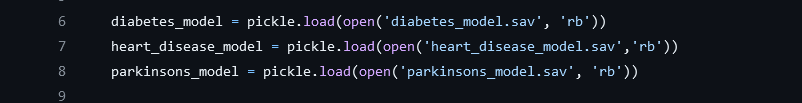
Exploratory Data Analysis:

In conducting the exploratory data analysis (EDA) for our multiple disease detection project using machine learning, we embarked on a comprehensive journey to understand the intricacies of our dataset. We began by computing descriptive statistics, unraveling the central tendencies and variability of each variable.Through these analyses, we gained a nuanced understanding of our dataset's characteristics, laying a solid foundation for informed preprocessing decisions and subsequent machine learning model development. The iterative nature of EDA remains integral, promising ongoing refinement as we progress in our quest for accurate and impactful disease detection.

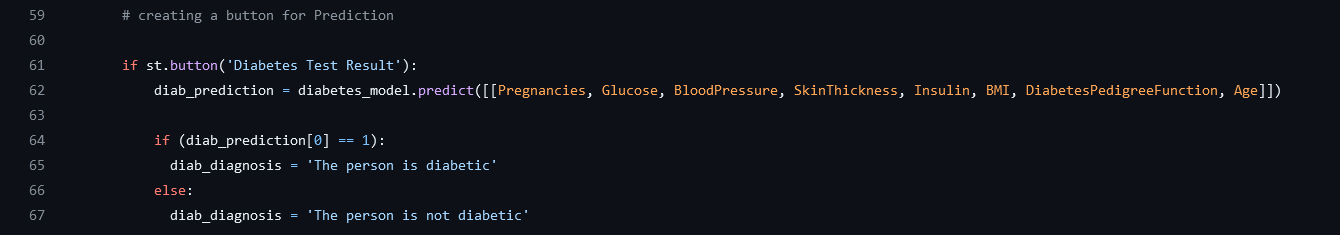
Model Training:

This step involves choosing the appropriate algorithm and representation of data in the form of the model. The cleaned data is split into two parts – train and test (proportion depending on the prerequisites) the first part (training data) is used for developing the model. The second part (test data), is used as a reference.

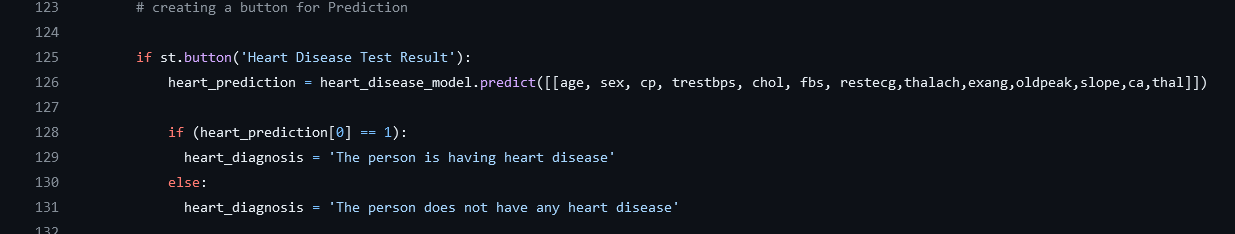
* Importing the dataset



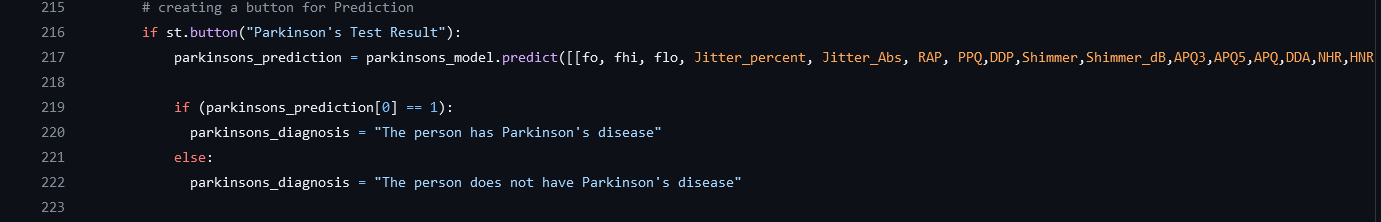
* Diabetes Prediction clause



* Heart Disease Prediction clause



* Parkinson’s Prediction clause



Performance Metrics:

The validation strategy encompasses rigorous testing against diverse datasets, including external validation and potentially clinical trials. Performance metrics such as accuracy, precision, recall, and F1 score are defined to quantitatively assess the effectiveness of our models.

# User Interface Design

Input Parameters:

It's essential to provide a clear description of the input parameters used for the multiple disease detection project. Here's a list of key input parameters to include age, insulin level, blood glucose level, blood pressure, etc.

Output Presentation:

Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements

User Interaction:

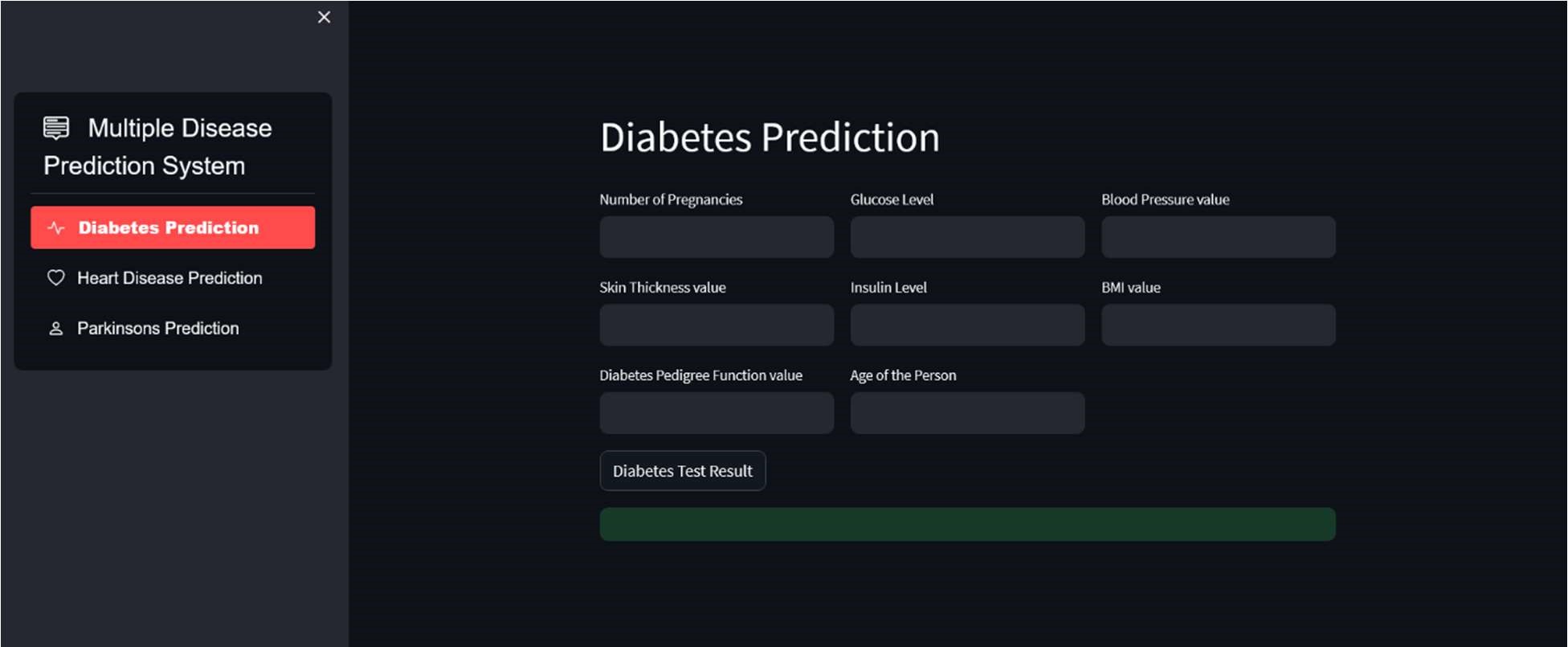


Fig 5.1: Diabetes Interface

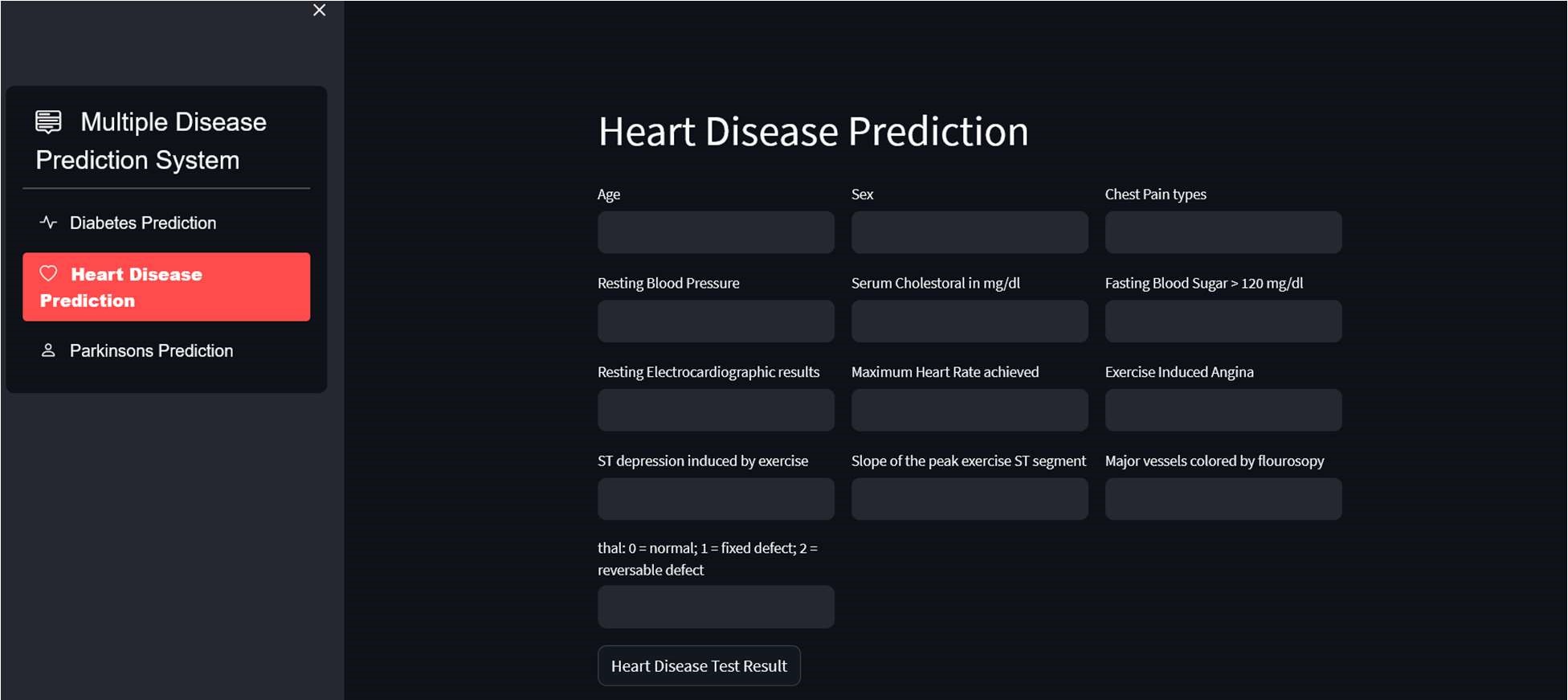


Fig 5.2: Heart Interface

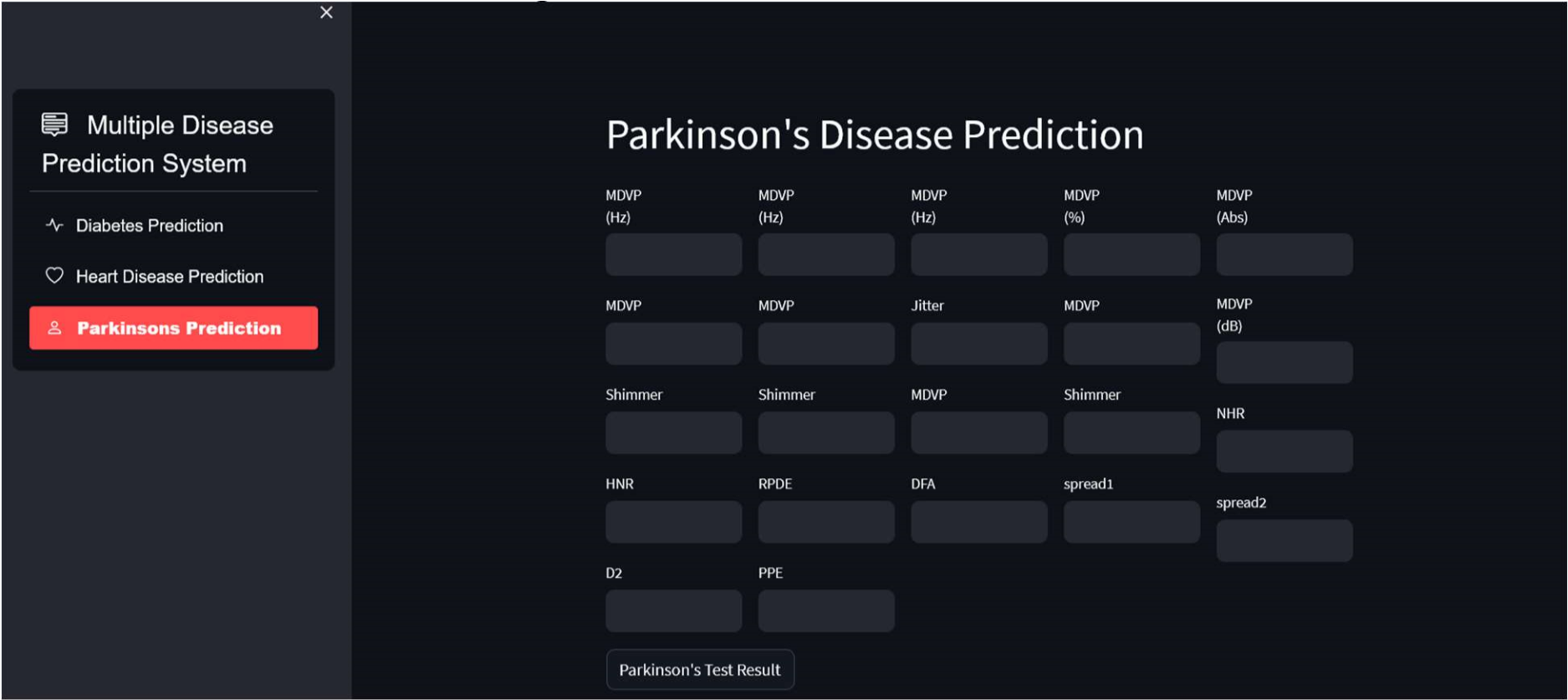


Fig 5.3: Parkinson’s Interface

# Implementation

Software and Tools:

Python used for model building including all the source codes.

Streamlit web framework used for hosting web app to deploy the models.

Html and CSS is used for designing the web page for the web app.

# Results and Discussion

In the system diabetes disease prediction model used KNN algorithm, heart disease uses the Random Forest algorithm and liver uses the random forest algorithm as these gave the best accuracy accordingly. There when the patient adds the parameter according to the disease it will show whether the patient has a disease or not according to the disease selected.

The parameters will show the range of the values needed and if the value is not between the range or is not valid or is empty it will show the warning sign that adds a correct value.

# Conclusion

In conclusion, our endeavor to develop a multiple disease detection system using machine learning represents a significant stride toward revolutionizing healthcare diagnostics. Through meticulous data preprocessing, exploratory data analysis, and model development, we have crafted a robust framework capable of identifying a diverse range of diseases from various data sources. The validation process, incorporating rigorous testing and ethical considerations, instills confidence in the reliability and real-world applicability of our models.

In essence, our multiple disease detection project stands at the intersection of cutting-edge technology and compassionate healthcare, embodying the promise of a future where early and accurate diagnoses contribute to improved patient well-being.

# References

1. Priyanka Sonar, Prof. K. Jaya Malini,” DIABETES PREDICTION USING DIFFERENT MACHINE

LEARNING APPROACHES”, 2019 IEEE,3rd International

Conference on Computing Methodologies and

Communication (ICCMC)

1. Archana Singh, Rakesh Kumar, “Heart Disease Prediction Using Machine Learning Algorithms”, 2020

IEEE, International Conference on Electrical and Electronics Engineering (ICE3)

1. Jadala Srilipi, Kesa Sruthi, Vodyati Vyshnavi, Dr Sreedhar

Bhukya, “Multiple Disease Prediction using Streamlit”, June 2022, IJIRT Volume 9 Issue 1

1. Laxmi Deepthi Gopisetti; Srinivas Karthik Lambavai

Kummera; Sai Rohan Pattamsetti; Sneha Kuna; Niharika

Parsi, “Multiple Disease Prediction System using Machine

Learning and Streamlit”, 2023 IEEE, 5th International

Conference on Smart System and Inventive Technology(ICSSIT)

# Appendices

Code Listings:

Include relevant sections of your code for transparency and reproducibility.

Additional Figures and Tables:

Attach any supplementary visual aids, figures, or tables that enhance the understanding of your project.

This detailed breakdown should give you a comprehensive overview of each section, ensuring a thorough understanding of your Intelligent Crop Recommendation System project.